Second Edition

ELECTRONIC IMAGING IN ASTRONOMY

Detectors and Instrumentation

Ian S. McLean

Springer

PRAXIS

Electronic Imaging in Astronomy

Detectors and Instrumentation (Second Edition)

Ian S. McLean

Electronic Imaging in Astronomy

Detectors and Instrumentation (Second Edition)

Published in association with
Praxis Publishing
Chichester, UK

Professor Ian S. McLean
Director, Infrared Laboratory
Department of Physics and Astronomy
University of California, Los Angeles
California
USA

The author gratefully acknowledges permission to use the images depicted on the cover.

Left image (radio):	NSF/VLA/Univ. Hertfordshire/M.J. Hardcastle
Middle image (X-ray):	NASA/CXC/CfA/R. Kraft et al.
Right image (visible):	WFI/ESO/MPG 2.2-m telescope/Benoît Vandame (ESO)
Keck telescopes (back cover):	WMKO/Ian McLean

SPRINGER–PRAXIS BOOKS IN ASTRONOMY AND PLANETARY SCIENCES
SUBJECT *ADVISORY EDITORS*: Philippe Blondel, C.Geol., F.G.S., Ph.D., M.Sc., Senior Scientist, Department of Physics, University of Bath, UK; John Mason, M.Sc., B.Sc., Ph.D.

ISBN 978-3-540-76582-0 Springer Berlin Heidelberg New York

Springer is part of Springer-Science + Business Media (springer.com)

Library of Congress Control Number: 2008926610

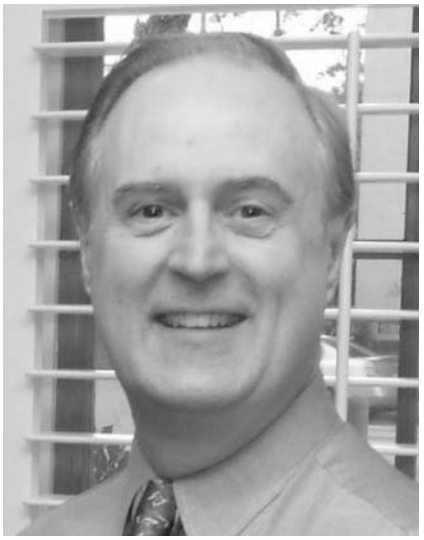
Cover design: Jim Wilkie
Project management: Originator Publishing Services Ltd, Gt Yarmouth, Norfolk, UK

Printed on acid-free paper

Contents

APPENDICES

About the author



Professor Ian S. McLean received his PhD in Astronomy from Glasgow University (U.K.) in 1974. While a member of staff at the Royal Observatory Edinburgh from 1979 to 1989 he developed the first CCD-based imaging spectro-polarimeter and the first facility-class infrared camera. He has been a member of the faculty of the Department of Physics and Astronomy at the University of California, Los Angeles and Director of the Infrared Laboratory since 1989. McLean is one of the world's leading authorities on the application of electronic imaging systems to advanced astronomical instrumentation. He served on the Science Steering Committee for the W. M. Keck Observatory for ten years and he has been team leader or co-investigator on several pioneering Keck instruments.

Preface

Throughout history astronomy has relied heavily on advances in technology. In some cases, astronomers have been the driving force behind those developments. Today's telescopes and observatories, whether on the ground or in space, are "high-tech" places where state-of-the-art electronic equipment is used to collect and analyze all forms of light. Driven by new technology, discoveries and revelations about the Universe have been coming at an incredible pace in recent years. Yet the demand for more sensitivity and better equipment is greater now than ever before. Modern astronomy is therefore as exciting and challenging for the professional engineer and applied physicist as it is for the astronomer. Moreover, would-be astronomers must reckon on acquiring a wide range of skills, or on working as a member of a multi-disciplinary team.

The first edition of *Electronic Imaging in Astronomy: Detectors and Instrumentation*, published in 1997, grew out of a precursor called *Electronic and Computer-Aided Astronomy: From Eyes to Electronic Sensors* which was stimulated by the desire to explain to others just how much applied physics and engineering goes into the seemingly "pure" science of astronomy. That earlier book, published in 1989, was also inspired by the remarkable impact which one small "silicon microchip", the CCD, had on astronomical imaging methods. For the 1997 book I added material on general principles and techniques, while maintaining a valuable historical perspective, and broadened the scope to include the remarkable growth of astronomical imaging across the entire electromagnetic spectrum, including my own area of infrared arrays. Since then, the rapid pace of technology and discovery has wrought many other fundamental changes in observational astronomy which have dated that account. For example, we have witnessed the success of numerous astronomical space missions, realized the importance of highly automated survey telescopes, and entered the era of very large telescopes. The CCD remains as ubiquitous as ever, but imaging is now possible at almost any wavelength from gamma rays to radio waves. Too often, the sophisticated and elegant instruments available at observatories are viewed by

students as a ''black box''. I would like my students to know what is inside the box, and how much effort it took to get it there. I hope this new edition will encourage more college courses on detectors and instrumentation for astronomy, but equally important, I hope it will encourage an even greater appreciation of the remarkable link between astronomy and technology.

This new edition builds on the successful format of the previous book. The text has been written on several levels in the hope that a wide range of people will be able to find something in it for them. In addition, the flow of the book has been arranged to be more useful as a college text on astronomical instruments and techniques. By using ''electronic imaging'' as a unifying theme, the aim is to provide a simplified, broad-based introduction to astronomical detectors and instrumentation that would be suitable for advanced undergraduate students and new graduate students who need a background in observational methods. Breadth rather than depth is more useful for students who have yet to choose their final career path. Inevitably there will be gaps in the coverage given the broad scope, but it is hoped that enough leads to other sources have been provided to help the reader follow any path of interest. By retaining some historical content and collecting diverse information in one volume, it is hoped that the book also serves as a useful reference for established professionals and anyone interested in this most important aspect of astronomy. Chapter 1 identifies the ''observables'' and provides a general introduction to electronic imaging. Chapter 2 treats the Earth's atmosphere and describes how Adaptive Optics is helping ground-based astronomy to compete with space. Chapter 3 contains an introduction to astronomical telescopes and the technologies behind the new very large telescopes. Chapter 4 illustrates the discovery power of astronomical detectors and instruments to explain what cameras and spectrometers do. Chapter 5 gives basic descriptions of the fundamental instruments of ''optical'' astronomy (i.e., the techniques that generally apply from the ultraviolet to the infrared). Detector classification is also explained, and semiconductors are introduced. Chapter 6 expands the discussion of instrumentation with an introductory tutorial for those interested in designing and building instruments. Chapters 7 and 8, respectively, give the basic principles of the CCD and then details on practical operation. Chapter 9 deals with a wide range of important calibration issues common to most electronic imaging devices and explains concepts such as flat-fielding and signal-to-noise ratios. An introduction to the techniques of image processing is given in Chapter 10. The remaining chapters expand the discussion to all the other wavebands, using the story of the CCD to link many of them together. Chapter 11 treats the revolutionary ''infrared array'' detector. Electronic imaging from UV to gamma rays is treated in Chapter 12. Chapter 13 explores electronic imaging from the sub-millimeter to the meter bands of radio astronomy. Finally, in Chapter 14, we look towards the future and speculate on the prospects for the development of new detectors and telescopes.

The approach is largely practical, with an emphasis on how things work. Historical perspectives are included because I have found that they are motivational, especially the role that modern astronomers and technologists have played in the development and exploitation of electronic imaging. Derivations of well-known material are not included, and mathematical expositions are at the undergraduate

level. Many college observatories have telescopes equipped with CCD cameras. This book can be used for a one-semester introductory course on modern astronomical detectors and instruments, and as a supplement for a practical or laboratory class. By supplementing this book with some more advanced material on optics and detector physics, this text also provides the core of an advanced course on astronomical instrumentation for new graduate (PhD) students. This book is also intended to be a useful reference for professionals in the scientific instrumentation field.

I would like to take this opportunity to thank all who helped and encouraged this work. I am particularly indebted to the Series Editor John Mason for his advice and to my publisher Clive Horwood of Praxis for his patience and long-time support. Many, many people kindly supplied me with information, reference materials, and photographs. In addition to all who helped with the earlier books, I especially wish to thank this time, Sean Adkins, Rachel Akeson, Taft Armandroff, Eric Becklin, Jim Beletic, Mike Bolte, Mark Casali, Eric Craine, George Djorgovski, Alex Filippenko, Bob Fischer, Neil Gehrels, James Graham, Don Hall, Martin Hardcastle, Wayne Holland, Mark Hunten, Derek Ives, Jim Janesick, Paul Jorden, James Larkin, David Leisawitz, Keith Matthews, Craig McCreight, Craig Mackay, Gerry Neugebauer, Rene Ong, Rene Racine, Chuck Steidel, Richard Stover, Jean Turner, Tony Tyson, John Vallerga, Greg Wirth, Peter Wizinowich, Erick Young, and Jonas Zmuidzinas.

It is a particular pleasure to acknowledge the many fine engineers and technologists that I have been privileged to work with over the years including my current team at the UCLA Infrared Lab in Los Angeles, as well as my former colleagues at the Royal Observatory Edinburgh (ROE), most especially the late Donald Pettie. During the writing of this second edition I was lucky to spend a sabbatical leave back in Edinburgh at ROE, now called the Astronomy Technology Centre (ATC). I am grateful to the Scottish Universities Physics Alliance (SUPA) for their support, and I want to thank Colin Cunningham and the entire ATC staff for their hospitality. Thanks too to my thesis adviser David Clarke (Glasgow University) who set me on the road to instrumentation. As a dedicated teacher, I also want to thank all of my students and postdocs over the years from whom I have learned so much. Finally, I am most appreciative of the unswerving support of my wife Jan and my family. I look forward to hearing from readers and teachers.

Ian McLean
Edinburgh and Los Angeles
March 2008

To my father,
for showing me the stars
and
to my family,
for being my stars

Figures

Tables

Abbreviations and acronyms

2dF	Two degree Field
2dFGRS	Two degree Field Galaxy Redshift Survey
2MASS	Two Micron All Sky Survey
6dF	Six degree Field
A/D	Analog to Digital
AAT	Anglo-Australian Telescope
ADC	Analog to Digital Converter; Atmospheric Dispersion Compensator; Atmospheric Dispersion Corrector
ADS	Astrophysics Data System
ADU	Analog-to-Digital Unit
AGN	Active Galactic Nuclei
AIPS	Astronomical Image Processing System
ALMA	Atacama Large Millimeter Array
ANSI	American National Standards Institute
AO	Adaptive Optics
AR	Anti-Reflection
ASCII	American Standard Code for Information Interchange
ASIC	Application Specific Integrated Circuit
ATRAN	Atmospheric TRANsmission code
BCCD	Buried Channel CCD
BIB	Blocked Impurity Band
BN	Becklin–Neugebauer
CADC	Canadian Astronomy Data Center
Caltech	California Institute of Technology
CARA	California Association for Research in Astronomy
CCD	Charge-Coupled Device
CCPS	California and Carnegie Planet Search
CCR	Closed Cycle Refrigerator

CDS	Correlated Double Sampling
CFHT	Canada–France–Hawaii Telescope
CGRO	Compton Gamma Ray Observatory
CHARA	Center for High Angular Resolution Astronomy
CID	Charge Injection Device
CMB	Cosmic Microwave Background
CMOS	Complementary Metal Oxide Semiconductor
COBE	COsmic Background Explorer
COTS	Commercial Off The Shelf
CRT	Cathode Ray Tube
CSO	Caltech Sub-millimeter Observatory
CTE	Charge Transfer Efficiency
CTIO	Cerro Tololo Inter-American Observatory
CW	Continuous Wave
CXO, CXRO	Chandra X-ray Observatory
CZT	Cadmium Zinc Telluride
D/H	Deuterium to Hydrogen
DCG	DiChromated Gelatin (layer)
DENIS	European DEep Near-Infrared Survey
DM	Deformable Mirror
DMA	Direct Memory Access
DN	Data Number
DQE	Detective Quantum Efficiency
DSB	Double SideBand
DSP	Digital Signal Processor
DSS	Digitized Sky Survey
E-ELT	European Extremely Large Telescope
EEV	English Electric Valve, now *e2v technologies*
EFL	Effective Focal Length
ELT	Extremely Large Telescope
EMCCD	Electron Multiplied CCD
EPROM	Erasable PROM
ESA	European Space Agency
ESO	European Southern Observatory
EUVE	Extreme UltraViolet Explorer
FCC	Federal Communications Commissions
FEA	Finite Element Analysis
FET	Field Effect Transistor
FITS	Flexible Image Transport System
FOV	Field Of View
FP	Fabry–Perot
FPA	Focal Plane Array
FPGA	Field Programmable Gate Array
FTS	Fourier Transform Spectrometer
FUSE	Far Ultraviolet Spectroscopic Explorer

FWHM	Full Width at Half Maximum
GALEX	GALaxy Evolution EXplorer
GBT	Green Bank Telescope
GLAST	Gamma-ray Large Area Space Telescope
GMT	Giant Magellan Telescope
GNAT	Global Network of Automated Telescopes
GPS	Global Positioning System
GRB	Gamma Ray Burster; Gamma Ray Burst
GTC	Gran Telescopio Canarias
GUI	Graphical User Interface
HD	Henry Draper
HDF	Hubble Deep Field
HESS	High Energy Stereoscopic System
HET	Hobby–Eberly Telescope
HR	Hertzsprung–Russell
HST	Hubble Space Telescope
IBC	Impurity Band Conduction
ICRF	International Celestial Reference Frame
IDL	Interactive Data Language
IEC	International Electrotechnical Commission
IF	Intermediate Frequency
IFU	Integral Field Unit
IPAC	Infrared Processing and Analysis Center
IRAF	Image Reduction and Analysis Facility
IRAS	InfraRed Astronomical Satellite
ISO	Infrared Space Observatory
IUE	International Ultraviolet Explorer
IVOA	International Virtual Observatory Alliance
JAXA	Japan Aerospace eXploration Agency
JCMT	James Clerk Maxwell Telescope
JFET	Junction Field Effect Transistor
JPL	Jet Propulsion Laboratory
JWST	James Webb Space Telescope
KAO	Kuiper Airborne Observatory
L3CCD	Low Light Level CCD
LED	Light Emitting Diode
LLNL	Lawrence Livermore National Labs
LO	Local Oscillator
LPE	Liquid Phase Epitaxy
LSB	Least Significant Bit
LSST	Large Synoptic Survey Telescope
LST	Local Sidereal Time
LUT	Look Up table
MAMA	Multi-Anode Microchannel Array
mas	milli-arcsecond

MBE	Molecular Beam Epitaxy
MCP	MicroChannel Plate
MCT	Mercury Cadmium Telluride (HgCdTe)
MEMS	Micro-Electro-Mechanical Systems
MKID	Microwave Kinetic Inductance Detector
MKS	Meter, Kilogram, Second (metric units)
MOS	Metal Oxide Semiconductor
MOVPE	Metal Organic Vapor Phase Epitaxy
MPPCCD	Multi-Pinned Phase CCD
MTF	Modulation Transfer Function
NASA	National Aeronautics and Space Administration
NEFD	Noise Equivalent Flux Density
NEP	Noise Equivalent Power
NIST	National Institute of Standards and Technology
NOAO	National Optical Astronomy Observatories
NRAO	National Radio Astronomy Observatory
NSF	National Science Foundation
OAP	Off Axis parabola
OFHC	Oxygen Free High Conductivity
op-amp	operational amplifier
OPD	Optical Path Difference
OTCCD	Orthogonal Transfer CCD
P-V	Peak-to-Valley
Pan-STARRS	Panoramic Survey Telescope And Rapid Response System
PCB	Printed Circuit Board
PCI	Peripheral Component Interconnect
PI	Principal Investigator
PMT	PhotoMultiplier Tube
POSS	Palomar Observatory Sky Survey
PROM	Programmable Read Only Memory
PSF	Point Spread Function
QE	Quantum Efficiency
rms	root mean square
ROE	Royal Observatory Edinburgh
ROIC	ReadOut Integrated Circuit
RVS	Raytheon Vision Systems
SALT	South African Large Telescope
SDSS	Sloan Digital Sky Survey
SIS	Superconductor–Insulator–Superconductor
SKA	Square Kilometer Array
SNR, S/N	Signal to Noise Ratio
SOFIA	Stratospheric Observatory For Infrared Astronomy
SQUID	Superconducting QUantum Interference Device
SR	Strehl Ratio
STScI	Space Telescope Science Institute

TDI	Time Delay and Integration; Time Delay Integration
TES	Transition Edge Sensor
TMA	Three Mirror Anastigmat
TMT	Thirty Meter Telescope
TTL	Transistor Transistor Logic
UBV	Ultraviolet, Blue, Visual (yellow)
UCO	University of California Observatories
UKIRT	U.K. InfraRed Telescope
UV/O/IR	UltraViolet, Optical, InfraRed
VERITAS	Very Energetic Radiation Imaging Telescope Array System
VLA	Very Large Array
VLBI	Very Long Baseline Interferometry
VLT	Very Large Telescope
VPH	Volume Phase Holographic
WF/PC, WFPC	Wide Field/Planetary Camera
WFE	WaveFront Error
WIYN	Wisconsin–Indiana–Yale and NOAO
WMAP	Wilkinson Microwave Anisotropy Probe
WMKO	W.M. Keck Observatory

Introduction

One summer when I was a small boy, my father told me the names of the brightest stars and pointed out the patterns of the major constellations. His knowledge of the night sky came from navigation experience as a sailing master on large cargo ships. I was fascinated, and wanted to go out every night to observe the stars. All astronomers talk about "going observing", but what does this mean? If you are an amateur enthusiast then it may mean going no farther than your backyard or your local astronomy club. For professional astronomers, however, the phrase means much more. Implicit is the fact that to understand the Universe we must observe it, and to do so we will need more than our human eyes. We will need all that modern electronic technology can offer. Today, the largest ground-based telescopes are located at relatively remote, pristine sites, high above sea level where the air is thin and the skies are astonishingly clear. So "going observing" can also mean going far away from home.

Access by professional astronomers to national ground-based observatories, as well as most university or privately owned facilities, is on a highly competitive basis. To obtain an allocation of "observing time" an astronomer must submit in writing a well-argued scientific case for his or her observational experiment. Deadlines are set typically twice or three times per year. Selection is done by peer review (i.e., by a committee formed from the body of scientists who actually use the facility). Unfortunately, all of the major telescopes are heavily oversubscribed, so disappointment is a fact of life. To maximize the progress of scientific experiments at each facility, and to make the optimum use of weather conditions, the astronomical community world-wide has expended considerable effort on technology. This means highly automated observatories with much reliance on well-engineered instrumentation and computers, and it also implies new cost-effective solutions for the design and management of telescopes and equipment. Observing time on large telescopes is therefore difficult to obtain and is very valuable; it is important that no time be wasted. Also, modern observatories are quite complex, and so guest astronomers who

may visit only twice per year cannot be expected to learn the myriad of operational details. To solve this problem, all large observatories provide one or more highly trained personnel to support the visitor. Usually a night assistant/ telescope operator will be provided; he or she will be responsible for control of the telescope and dome, ensuring efficient operation and keeping an observatory logbook. Other staff will be responsible for the preparation of observing equipment. The night assistant has the final word regarding safety matters, such as closing the telescope dome if the wind speed becomes too high. Sometimes a support scientist, who is a professional research astronomer on the observatory staff familiar with the instrumentation, will be available to assist first-time or irregular users of the telescope.

A guest observer (or GO) planning to use a modern, computer-controlled electronic imaging camera or spectrograph at one of these major facilities might encounter the following pattern of work. Visiting astronomers will probably arrive by air a few days before their allocated time to ensure that they are not travel-weary and to discuss their plans with observatory staff. They may have traveled from North America or Europe to Hawaii, or to Chile or Australia, or any of several other destinations. By mid-afternoon before the first night on the telescope, the observatory staff will be in the telescope dome making sure that the telescope and the instrument are functioning correctly. The visiting astronomer(s), often including graduate students receiving training in observational methods or seeking data for a thesis topic, may well elect to be present for these checks, and may wish to practice using the instrument. This may mean becoming familiar with a control panel, or with the operation of a computer console on which the observational modes can be displayed and changed by typing at a keyboard. To feel confident that they understand the operation of the instrument, the visiting astronomers will carry out some tests of their own such as a "noise check" on the detector, or a calibration image or spectrum. With everything ready for the evening, they return to the observatory residential lodge where a meal might be prepared for them or to make their own. This is usually a great chance to meet people from all over the world, and the dinner conversation is often buzzing with astronomical jargon! Just prior to sunset the "observers" go back to the telescope dome, usually in the company of the night assistant, to complete their preparations. A final list of objects and "coordinates" is supplied to the telescope control computer. As the twilight fades and the sky becomes dark enough to work, the night assistant will "call up" the first object on the target list and a computer will instruct electrically driven motors on the telescope's rotation axes to slew to that position. Using a special video camera at the focus of the telescope, the guest astronomer examines the field of view to confirm that the telescope is pointing at the object of interest by reference to an existing star chart. Sometimes nothing can be seen because the object(s) are too faint and require a long exposure. In that case, the field must be confirmed by checking the pattern of brighter non-target objects in the vicinity. When the object is correctly centered, the observation begins. Having configured the camera or spectrograph to the required settings by using the instrument's control computer, all that is required next is to issue a "start" command. The total time for which the measurement lasts is called the "integration time" and this may be anything from a fraction of a second to hours depending on the brightness of the object, the efficiency of the instrument, the

wavelength, and the nature of the experiment. If the integration time is long, then it is essential to ensure that the telescope continues to track the object very accurately. In principle, this can be done manually by viewing the object or a nearby star with the video camera, and pushing buttons on a "hand-set" connected to the telescope in such a way as to counteract any drift of the image. More likely, guiding will be performed automatically by the telescope control computer which will analyze the image of the guide star on the TV screen, compute any motion, and issue a correction to the drive motors of the telescope. When the exposure is complete, the image or spectrum will be displayed on a computer screen, an adjustment to the setting of the instrument might be made, and another exposure started. Meanwhile, some rapid analysis of the first result is carried out. This is crucial to the optimum use of telescope time.

The same pattern of work is repeated throughout the night. Nights can be long, from before dusk until well after dawn, typically 12 hours non-stop. Considerable concentration and often a degree of patience are required. Sometimes the latter is in relation to the other observers rather than with the experiment, and so some sustenance or "night lunch" might be taken on the job. Some look forward to opening up the little brown bag collected at dinner, others would just as soon not watch! Depending on how smoothly the experiment has progressed or on what has been found, tactical decisions may be required to optimize the use of the night. Certainly, as dawn approaches, an extra effort is made to get the most out of the remaining time. A golden rule of observing which many newcomers forget to follow is to assume that "every night is your last" and never leave a crucial measurement or calibration until tomorrow. Finally, with the last on-sky exposure complete, the night assistant will close the dome. A few more calibration frames are made and then the mirror covers are also closed and the telescope is returned to its parked position. At last the equipment is shut down or placed in standby mode, logbooks or fault reports are filled out, and the weary group rally round for the walk or drive back to the lodge. The observers sleep until early afternoon and then rise to prepare and review for the next night. Several days of this activity constitutes the "observing run". The visitors will then spend a few days at the offices of the observatory, perhaps to obtain electronic copies of their data and discuss their observations with local staff. Finally, they fly back to their home institute to analyze the astrophysical content of the data in detail and write a scientific paper.

If you plan to use single-dish radio telescope facilities then you may well follow the same pattern as optical and infrared astronomers, particularly for sub-millimeter observations. Interferometer and aperture synthesis telescopes such as the Very Large Array (VLA) near Socorro, New Mexico are too automated to have visiting astronomers present for interactive sessions. Usually, one submits an "observe" file and the observations are taken as part of a larger pre-programmed sequence. It may however be necessary for the astronomer to visit the facility during the data analysis phase to make use of computers and software not available at their home institute. Of course, the same process of competitive application for telescope time based on scientific merit is used. A similar situation is also encountered for telescopes located in space. While it may be necessary to visit the "home" institute of the satellite, such as the Space Telescope Science Institute in Baltimore (U.S.A.) or the Space Telescope European

Coordinating Facility near Munich (Germany), in the case of the Hubble Space Telescope, it is rarely necessary to visit the ground-station and operations center. Time on the Hubble Space Telescope and many other satellites is very competitive and is awarded in cycles on the basis of peer-reviewed scientific proposals. Once again, when time is awarded it is necessary to submit a detailed observing request that is programmed into a larger sequence by observatory staff. You will definitely not get to "play" with the telescope! Often, your data will simply arrive in the mail in electronic form, and usually in a form ready for you to begin scientific analysis.

Some ground-based observatories now operate a mode in which your observing request is carried out for you by observatory staff, and the execution is scheduled in such a way that your measurements are mixed in with those of other applicants. Your presence at the observatory is not needed. This approach can optimize the use of weather conditions, but tends to weaken the spontaneous response to a discovery that comes from being in control of the experiment. High-speed data links also make "remote observing" a real possibility, enabling the experiment to be controlled from a distant location. In fact, some telescopes are completely robotic (i.e., they are programmed to operate autonomously using weather stations and pre-programmed target lists). Of course, there are numerous privately owned observatories throughout the world, mostly associated with universities and research consortia. Some of these facilities are quite large, and others are fairly small. At many of these places there is a much greater degree of "do it yourself". Nevertheless, the pattern of preparation and work is essentially the same. Professional astronomers usually work in the comfort of a warm control room, rather than the dark and cold telescope dome, while electronic imaging devices and computer systems gather data. Occasionally, forays are made into the chilly mountain air outside to check on the weather or, as in my own case, simply to look up at the star-studded canopy of the night sky, marvel at its awe-inspiring beauty, and remind oneself that *this* is what it is all about!

1

The development of electronic imaging in astronomy

When viewed far from city lights the star-studded night time sky is an awe-inspiring sight that fires the imagination. Even the earliest civilizations realized that careful astronomical observations were important to survival because such observations enabled them to predict seasonal events, such as when to plant and when to harvest. Observational astronomy was also one of the earliest scientific activities.

1.1 OBSERVATIONAL ASTRONOMY

1.1.1 Historical development

The Greek astronomer Hipparchus (*c.* 127 BC) used astronomical observations to determine the lengths of the four seasons and the duration of the year to within 6.5 minutes. He also derived the distance to the Moon and the Sun, but his most amazing feat was to notice a small westward drift of the constellations which we now call the precession of the equinoxes. This effect causes the current Pole Star (Polaris) to move away from the North point and circle back after almost 26,000 years! Chinese astronomers recorded the appearance and fading of an exceptionally bright star in 1054 AD in the constellation we now call Taurus, but it was not until the 20th century that Edwin Hubble (1889–1953) associated this event with the supernova explosion which gave rise to the Crab Nebula, also known as Messier 1, the first entry in the list of nebulous objects studied by Charles Messier (1730–1817). Following the invention of the telescope in the early 1600s, Galileo Galilei (1564–1642) and others were finally able to enhance the sensitivity of the only light detector available to them, the human eye, and to resolve details such as craters and mountains on the Moon, the rings of Saturn, moons orbiting Jupiter, and the individual stars in the Milky Way. By making careful drawings (Figure 1.1) of what their eyes could detect during moments of minimum atmospheric turbulence, what astronomers today call moments of "good

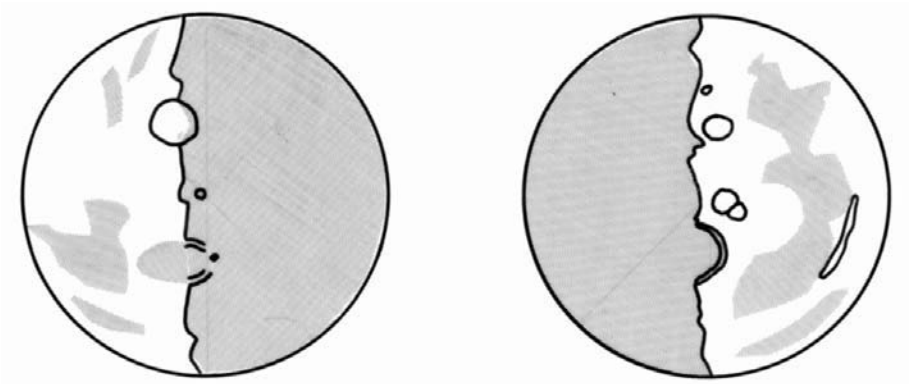


Figure 1.1. Hand-drawn sketches of features on the surface of the Moon as might have been made by the first astronomers to use telescopes to enhance the power of the eye.

seeing'', the early 17th-century scientists were able to convey pictorially those observations to others.

Better telescopes led to more astronomical discoveries, which in turn stimulated the development of even bigger and better telescopes. Opticians developed color-corrected lenses for telescopes and then, following Isaac Newton (1642–1727), telescopes using reflections from curved mirrors instead of transmission through lenses were gradually introduced. William Herschel (1738–1822), a prolific observer and discoverer of the planet Uranus, pioneered the construction of many reflecting telescopes with long "focal lengths" and large magnifications; in later years the emphasis would move to larger diameter mirrors rather than longer focal lengths. With the invention of the prism spectroscope by Joseph Fraunhofer (1787–1826), the chemical constitution of the Sun and stars became amenable to physical study. In Fraunhofer's early experiments a beam of sunlight was passed through a narrow rectangular slit in a mask and then through a glass prism to produce a colored spectrum in the manner similar to Newton and others (Figure 1.2). The critical addition made by Fraunhofer was a small telescope mounted on a movable arm which could be set to precise angles to view the spectrum. Initially, the light detector was still the human eye. Fraunhofer found that the normal band of colors from violet to red was crossed by numerous dark vertical lines. Eventually the pattern of these Fraunhofer absorption lines (actually images of the entrance slit partially devoid of light) was shown to be characteristic of individual chemical elements. The elements hydrogen, calcium, sodium, and iron were recognized in the spectra of the Sun, and later, the stars. Further spectroscopic observations of the Sun soon led to the discovery by Janssen and Lockyer (in 1868) of an unknown element which we now know to be a major constituent of the universe. This new element, helium, was named after the Greek word for the Sun *helios*; helium was not discovered on Earth until 1895.

When dry, gelatin-based photographic emulsions became routinely available in the late 19th century, astronomers such as Henry Draper (1837–1882) lost no time in

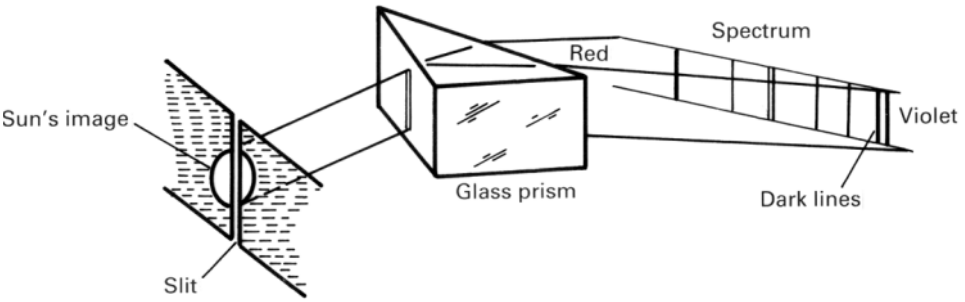


Figure 1.2. Joseph Fraunhofer's spectroscope and the dark lines in the spectrum of the Sun which now bear his name. This instrument combined with the photographic plate—rather than the human eye—opened the way to a physical understanding of the universe.

putting them to use to catalog the appearance and properties of a wide range of objects in the night sky. The photographic process was unarguably more accurate and more sensitive than the keenest human eye and the most artistic hand. From planets to stars to galaxies, the new observational tools were applied. Still larger telescopes were constructed, each a technical feat for its era, reaching a mirror diameter of 100 inches (2.54 meters) in 1917 with the completion of the Hooker Telescope on Mount Wilson by George Ellery Hale (1868–1938). Just one of the great discoveries which followed was the expansion of the universe by Edwin Hubble and Milton Humason in 1929.

The history of astronomy is marked by such sporadic progress. Each improvement in scientific apparatus, each new development in technology, helps to provide answers to old questions. Inevitably, the new observational methods uncover a host of new questions, which in turn drive the quest for even better measuring equipment! Progress in studying the universe has always been related to "deeper" surveys of the cosmos reaching to ever-fainter objects, or higher resolution yielding more and more fine detail, or larger statistical samples from which generalizations can be made, or broader spectral response to sample *all* the energy forms passively collected by the Earth. That trend has continued since the Renaissance of the 16th century to the present day in a kind of ever-increasing spiral, with new tools or technologies leading to new discoveries which in turn drive the development of better tools.

A key feature of observational astronomy has been record-keeping, maintaining archives of observations, usually in some pictorial form, for future investigators to compare and consider. In terms of its ability to convert light into a measurable quantity, the photographic plate is actually less sensitive than the human eye. The great advantage of the photographic plate, however, is that it can build up a picture of a faint object by accumulating light on its emulsion for a long period of time. It is therefore called an "integrating" detector. The eye cannot do this to any significant extent. Moreover, the plate provides a permanent record for future study by others.

By using a photographic plate as the recording device in a spectrometer, astronomers could extend their investigations effectively and efficiently into the domain of

quantitative astrophysics. Initially, of course, the flood of photographic material was analyzed by human eyes, and those eyes were mostly those of a dedicated group of female assistants hired by the director of the Harvard Observatory College, Edward Charles Pickering (1846–1919), toward the end of the last decade of the 19th century. Over forty women were employed by the observatory during the period of Pickering's tenure as director, and their efforts in handling the torrent of new astronomical data laid the foundations of modern astrophysics. Stellar spectral classifications led to the understanding that the colors of stars was largely a temperature sequence and that stars shine by the energy released in thermonuclear fusion reactions brought about spontaneously by the enormous temperatures and pressures at their centers. Among the most well-known of the Harvard ladies is Henrietta Leavitt (1868–1921) whose work on the class of stars called Cepheid variables, which pulsate in brightness with a period that is proportional to their true or absolute average brightness, led to a distance estimator and an appreciation of the true size and shape of our galaxy. During the first half of the 20th century, these tools inevitably resulted in more discoveries (per year) and a massive increase in the "data rate"; that is, the amount of information being collected, scrutinized, and archived for posterity. But these advances were only the beginning.

Even as the 100-inch (2.54 m) Hooker telescope was discovering the expansion of the universe, plans were being laid to build the great 200-inch (5.08 m) reflecting telescope on Mount Palomar in southern California. That telescope, named after George Ellery Hale, went into operation in 1949 and remained the largest telescope in the world until the construction of the Russian (then Soviet) 6 m Bol'shoi Teleskop Azimutal'ny (BTA) in 1976. Construction of both of these large telescopes was challenging. For the 200-inch, Hale secured a grant in 1928 from the Rockefeller Foundation, but optical figuring of the Pyrex mirror took from 1936 to 1947 with four years off for World War II. The telescope was dedicated in June 1948 ten years after Hale's death, but it was another 16 months before director Ira Bowen (1898–1973) opened the telescope for full-time use. Weighing about 1,000 tons, the dome of the Hale telescope stands 41 m (135 ft) high and is 42 m (137 ft) in diameter. Likewise, the BTA on Mount Pastukhov on the northern side of the Caucasus range has a dome that is 58 m high and a primary mirror of Pyrex weighing 42 tons with so much thermal inertia that it can only tolerate a 2°C change per day if it is to retain its optical figure. Thermal inertia, the large dome, and the site turn out to be limitations on the best image quality that can be delivered. In the years that followed, astronomers would apply those lessons learned.

Building telescopes larger than 5 meters in diameter was going to be difficult, but observational astronomy received multiple boosts in the 1960s partly by the construction of many new optical observatories with 4-meter class telescopes; that is, with mirror diameters of approximately 3 m to 4 m (metric dimensions are preferred, see Appendix B for conversions). Although the telescopes were slightly smaller, these new facilities were well-equipped and located on excellent but somewhat more remote mountain sites in different parts of the world including the Arizona desert, the mountains of northern Chile, and the summit of Mauna Kea on the Big Island of Hawaii. The story so far refers only to "optical" astronomy. Another part of the

1960s expansion was stimulated by the exciting new look at the universe which accompanied the rise of radio astronomy and the discovery of completely new phenomena such as the incredibly luminous and distant quasars, thought to be supermassive black holes at the center of large galaxies, and the remarkable pulsars, now understood to be spinning neutron "stars" embedded in the remnants of a supernova explosion. All of this occurred during the successful development of the Soviet and American space programs which led to satellite astronomy and the opening up of the X-ray, ultraviolet, and infrared regions in the 1960s and 1970s. History shows that the introduction of any new domain results in new discoveries (e.g., Harwit, 2003). Other, more subtle, transformations began to occur around this time too, through the introduction of electronic computing machines and electronic devices which could be used as detectors of light. Photocells and sensitive "night-vision" TV cameras came first, but the steep rise of consumer micro-electronic products through the 1970s was to accelerate the changes rippling through astronomy. Even the telescopes themselves could be improved by the use of electronically encoded computer-controlled drive systems, thereby enabling much faster setup times and more reliable tracking across the sky. The newest radio and optical telescopes were remotely controlled, and the concept of converting measurements into an electronic form readily acceptable to a computer became standard practice. Computer power expanded exponentially, and astronomers eagerly used those capabilities to the full.

Construction of larger telescopes stagnated until the mid-1980s when Jerry Nelson of the University of California broke the paradigm by suggesting the concept of a segmented mirror whose shape was controlled by a computer. Around the same time it was also realized that very large thin mirrors with low thermal inertia could be used if computer-controlled force-actuators maintained their shape throughout the night. Consequently, optical telescopes have now reached gigantic proportions with diameters around 10 m ($\sim$394 inches) for the twin telescopes of the W. M. Keck Observatory (WMKO) which began operations in 1993 and 1996, respectively. Moreover, there are now telescopes, both on the ground and in space, to cover far more than the visible light our human eyes are designed to see. Today, computers actively control the shape of optical surfaces in the telescope and in associated instruments, performing thousands of calculations per second to correct the image quality. Smaller, highly automated telescopes survey the entire sky to unprecedented depths and many of these images are immediately available in digital form to all astronomers. This flood of quantitative information is due to strides in the range and sensitivity of electronic detection devices. It is the impact of semiconductor electronic light-sensors attached to the new generation of telescopes (both on the ground and in space) which has had an effect as dramatic as the introduction of the photographic plate itself over one hundred years ago.

There can be little doubt that we are living in a time of rapid technology development. This is the Digital Age, the age of the "microchip". Semiconductor technology, of which the "silicon chip" found in computers is by far the most widely known example, has touched almost every aspect of our daily lives. The mass production of silicon chips has brought Personal Computers (PCs) of incredible

power, at relatively low cost, to almost every environment: homes, schools, offices, and industry. The Digital Age is also the age of global electronic communication. There can be few people left who haven't at least heard of the Internet and the World Wide Web! School kids can "download" images from the Hubble Space Telescope web site and "email" messages and pictures to friends half-way around the world almost instantaneously by typing at a computer keyboard.

What is a semiconductor? A semiconductor is a crystalline material with some of the properties of a good conductor of electricity (like copper metal), and some of the properties of an electrical insulator (like glass, for example). Because of its crystalline (solid-state) structure, a slab of such material behaves the same at all points. Semiconductor crystals can be "grown" in a controlled way from a melt, and moreover, the electrical properties can be tailored by introducing so-called impurity atoms into the crystal structure at the atomic level, so that by microscopic sculpting of the semiconductor material, all sorts of tiny electrical components and circuits can be constructed. The final piece, often not much larger than a thumbnail, is referred to as an "integrated circuit" or more commonly, as a "chip". Besides silicon, there is germanium, gallium arsenide, indium antimonide, and several other materials with these properties. Semiconductors can be used to manufacture a host of low-power micro-electronic components including amplifiers, all sorts of logic units, computer memory, very complex chips called microprocessors capable of many computational functions, and tiny imaging devices of remarkable sensitivity. Silicon is the most well-developed semiconductor so far, but even for silicon the potential for yet smaller and smaller microchips still exists. Astronomy has benefited in this semiconductor revolution because the apparatus needed for scientific experiments and for complex calculations, which were completely impossible before, are now viable with the aid of the latest electronic imaging devices and powerful high-speed electronic computers.

Almost all modern astronomical research is carried out with photo-electronic equipment, by which we mean instrumentation that converts radiant energy (such as light) into electrical signals which can be digitized; that is, converted into numerical form for immediate storage and manipulation in a computer. Usually highly automated and remotely controlled, these instruments, and telescopes to which they are attached, rely heavily on electronics and computers. Computers play an equally crucial role in helping astronomers assimilate, analyze, model, and archive the prodigious quantity of data from the new instruments. The ongoing miniaturization of computers and the ever-increasing availability of large amounts of relatively cheap computer memory means that astronomers can employ fairly complex electronic and computer systems at the telescope which speed up and automate data-gathering. As a result, those astronomical facilities, which may be costly initially, and the data they produce can be available to a much wider range of scientists than would otherwise be possible. Today, a large modern observatory requires an enormous breadth of engineering, scientific, and managerial skills to operate efficiently and produce the very best results.

Many readers will be familiar with sources of current and topical astronomical results, whether these are professional journals (e.g., *Nature*, the *Astrophysical Journal*) or popular magazines (e.g., *Sky & Telescope*) or any of the numerous

astronomical sites accessible on the World Wide Web. How are such remarkable observations obtained? Most press releases do not describe in detail the apparatus or the technology used in making the discovery. Of course, it would not be easy to do so because of the "jargon barrier" and the complexity of the technology itself. This is unfortunate, because it underemphasizes an important link between modern technology and the quest for fundamental knowledge embodied in astronomy, a search for answers to the most basic questions about our universe. Our theme throughout this book is to emphasize this link.

1.1.2 What are the observables?

Astronomy is truly an observational science. Unlike in a laboratory experiment, the conditions cannot be changed. That is, we on Earth are passive observers (so far) in almost all astronomical experiments, and we can do nothing other than intercept (observe) the various forms of energy which reach the Earth from the depths of space. Of course, there have been a few notable exceptions for solar system studies involving manned and unmanned spacecraft that have returned samples to Earth, and from time to time we can retrieve rocks from space which have survived passage through the Earth's atmosphere in the form of meteorites. Otherwise, the energy forms that we can intercept passively can be summarized as

— *electromagnetic radiation* (gamma rays through radio waves)
— *cosmic rays* (extremely energetic sub-atomic charged particles)
— *neutrinos* (tiny neutral particles with almost immeasurably small mass)
— *gravitational waves* (disturbances in a gravitational field).

Of these, the study of electromagnetic radiation which, as shown by the great Scottish mathematical physicist James Clerk Maxwell (1831–1879) in 1865, incorporates visible light, is still the most dominant. Gravitational waves, ripples in spacetime predicted by Einstein (1879–1955), have not yet been detected directly, but in the U.S.A. the Laser Interferometer Gravitational-Wave Observatory (LIGO), with sites in the states of Washington and Louisiana, went into operation in 2002; similar facilities exist in Germany, Italy, and Japan. Neutrino detectors and cosmic ray experiments have been developed successfully. Among the most well-known of the neutrino observatories are the Homestake Gold Mine in South Dakota (U.S.A.) where Ray Davis (1914–2006; Nobel Prize in Physics 2002) first uncovered the "solar neutrino problem" in which the Sun seemed to be emitting only one-third of the expected number of neutrinos based on the well-understood theory of nuclear hydrogen–helium fusion, and the Sudbury Neutrino Observatory in Ontario (Canada) which resolved the problem by detecting all three neutrino types when it was eventually realized that three kinds of neutrinos existed. The Kamiokande neutrino observatory in Japan was sufficiently sensitive that it detected neutrinos from the supernova explosion (SN1987A) of a star in the Large Magellanic Cloud about 170,000 lightyears away; a lightyear is about 9.5 trillion kilometers (about 5.9 trillion miles) and is the distance light travels in one year. The vast majority of cosmic

ray particles are protons, the positively charged nucleus of the hydrogen atom, although heavier nuclei are also observed. Low-energy cosmic rays must be detected from spacecraft, but higher energy rays generate an "air shower" when they impact the Earth's atmosphere resulting in faint flashes of blue light known as Cherenkov radiation which can be detected by a suitably designed large telescope on the ground. One of the first telescopes built to detect Cherenkov radiation was the Whipple telescope on Mt. Hopkins in Arizona (1968) but many newer facilities now exist.

Maxwell's equations are a set of four fundamental relationships that quantify experimental findings about electric and magnetic phenomena, especially those involving the magnetic field due to an electric current (Ampere's Law modified by Maxwell) and the electric field caused by a changing magnetic flux (Faraday's Law of electromagnetic induction). These two equations can be combined to show that both the electric and magnetic fields satisfy the known form for a wave equation. Maxwell's analysis revealed that light is essentially characterized by oscillations of electric and magnetic fields which give the radiant energy the property of a wave motion. Different regions of the electromagnetic "spectrum" correspond to different "wavelengths" (denoted by the Greek letter lambda, λ; see Appendix A for Greek alphabet), and the energy in the wave moves through empty space at a speed of 299,792,458 meters per second (m/s), which is of course the speed of light (usually denoted by the letter c); actually, Maxwell derived this number from two electrical constants. Useful approximate values for the speed of light are 300,000 km/s, 186,000 miles per second, and 670 million miles per hour. The frequency of the oscillations (denoted by the Greek letter nu, ν) is related to the wavelength by the very simple equation

$$\nu\lambda = c \tag{1.1}$$

In the simplest case of a monochromatic (single wavelength) wave traveling in the x-direction and vibrating in a fixed (x, y) plane, the oscillation can be described by a simple sinusoid (e.g., $y = a\sin(\omega t - kx + \phi)$ with $\omega = 2\pi\nu$ and $k = 2\pi/\lambda$) and the average intensity of light is proportional to the square of the amplitude (or swing) a^2 of the wave. The term ϕ is the phase, and can be set to zero by an appropriate choice of origin. A cosine works too because $\cos(\theta) = \sin(\theta + \pi/2)$, which is just a sine wave with a phase shift of $\pi/2$. The importance of Equation (1.1) is that it implies no restrictions on the frequencies or wavelengths themselves, only that their product must be the speed of light. Optical measurements show that normal visible light corresponds to wavelengths around 0.5 millionths of a meter and frequencies of 600 trillion cycles per second, but waves of much lower frequency (300 million cycles per second) with huge wavelengths of 1 meter or more should be possible. This result led to the prediction and subsequent discovery of radio waves. The unit of frequency (1 cycle per second) is now called the hertz (Hz) after Heinrich Hertz (1857–1894) who validated Maxwell's predictions by experiments with early radio antennas. Electromagnetic waves can bounce off certain surfaces (reflection), be transmitted through certain materials with a change of direction (refraction), curl around obstacles or through openings by diffraction, and "interfere" with one another to cause cancellation or amplification of the wave. Of these, the phenomenon of diffraction sets a

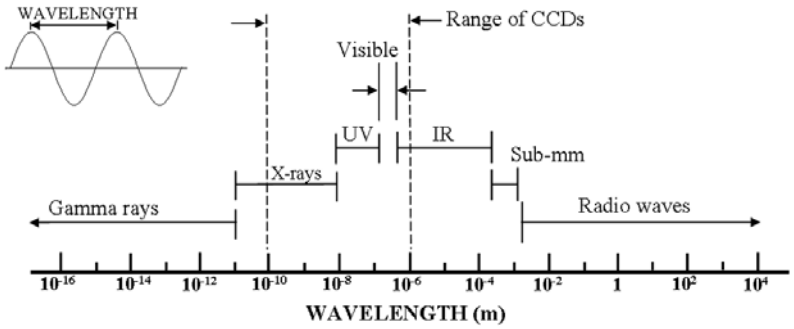


Figure 1.3. The electromagnetic spectrum: X-rays, light, and radio waves are all different forms of electromagnetic radiation. In the vacuum of empty space, each of these forms of radiant energy travel in straight lines with the same speed, the speed of light.

fundamental limit on measurements, and we will mention this limit many times in the quest for ultimate perfection in imaging. For now, we note only that the "angular resolution" or ability to separate two closely spaced stars a small angle apart on the sky, for a telescope of diameter D collecting light of wavelength λ, is given approximately by $57.296^{\circ}\lambda/D$ in the diffraction limit. Maxwell's equations, electromagnetic waves, and their interactions through interference, reflection, refraction, and scattering are described in any good college physics text. More details will be presented as needed in subsequent chapters. Because electromagnetic oscillations are transverse to the direction of propagation of the energy, these waves can be "polarized", which means they have an associated "plane of vibration".

As shown in Figure 1.3, all the well-known forms of radiant energy are part of this electromagnetic spectrum. The range in wavelengths is incredibly large. Radio waves are characterized by wavelengths of meters (m) to kilometers (km), whereas X-rays have wavelengths around 1 nanometer (nm) or one-billionth (10^{-9}) of a meter, comparable with the size of atoms. Other length units such as the micron (μm, 10^{-6} m) and the angstrom (Å, 10^{-10} m) are commonly used; scientific notation (powers of ten) and prefixes to standard units (such as nano- and micro-) are summarized in Appendix A. Visible light, with wavelengths from about 390 nm to 780 nm (or 0.39 μm–0.78 μm), occupies only a very small portion of this enormous radiant energy spectrum.

The rate at which the energy flows from a source is called the "radiance" or power, and the power emitted by the Sun, for example, is about 3.8×10^{26} watts (1 watt is equivalent to 1 joule per second). The power that is *received* by one square meter is the "irradiance" (measured in watts/m^2) and irradiance drops off inversely as the square of the distance from the source. Thus, at the average distance of the Earth from the Sun the solar irradiance is about 1366 watts per square meter above the Earth's atmosphere. Measurements that can be made on electromagnetic radiation are limited. Basically, we can determine

— the *direction* and *time of arrival* of the radiation
— the *intensity* at each wavelength or spectral energy distribution

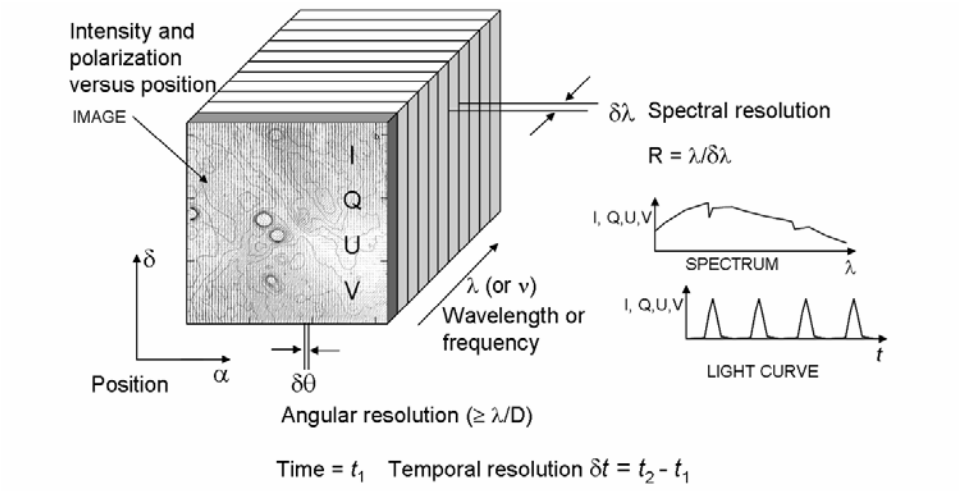


Figure 1.4. A pictorial summary of most of the observables for electromagnetic radiation. Polarization is represented by the Stokes intensities Q, U, and V; I is total intensity.

— the *polarization* or degree of alignment of the electric and magnetic fields in the radiation
— the *phase* or relation between waves.

Any of these quantities can vary with time and all can be observed with varying amounts of resolution (angular, spectral, or time) determined both by the limitations of measuring equipment and the wave nature of light. A map of the distribution of intensity over a given field of view is an "image" of that scene at the given wavelength (see Figure 1.4). All that we know about the universe must be extracted from measurements of these energy forms. Naturally, astronomy began as an optical science because human beings have built-in optical sensors, our eyes.

1.2 FROM EYES TO ELECTRONIC SENSORS

Looking through a telescope at the stars on a crisp, clear night is usually sufficient to get hooked on astronomy. It certainly was in my case. But the spectacular pictures from the Hubble Space Telescope so familiar to everyone since the early 1990s are nothing like what you see when you peer through a telescope with your eye. Why is that? Because, not only have electronic sensors been used to detect light that the eye simply cannot see, but also computers have processed the digital pictures to enhance the appearance of certain features for ease of study. If electronic sensors measure light that the eye cannot see then how can we represent such measurements, other than by a table of numbers? In practice, we assign colors that the eye can see to each of the invisible wavelengths in order to create a visualization of the scene. In this case the color is clearly false and does not represent what your eyes would see when looking at

this object. Visualization techniques will be explained in later chapters. For now, let's start by considering the detection of light and the features and limitations of the human eye.

1.2.1 The eye

In recent years vision scientists have found themselves working closely with astronomers because both share an interest in any technology that can overcome turbulence and improve resolution. One such collaboration is the Center for Adaptive Optics at the University of California, Santa Cruz (U.S.A.) where scientists involved in retinal imaging with live subjects are using the same technologies that astronomers use to overcome atmospheric turbulence (e.g., Carroll *et al.*, 2005).

The human eye is a remarkable "imaging system" (Figure 1.5). It contains a mechanism for forming an image (a lens), a means of detecting light (the retina), and a real time image-processing computer (the brain). It thus possesses all of the features required of an electronic imaging system. Some eye systems use a pinhole camera effect without any lens, but a wide range of eye types employ a lens, and also employ binocular vision. Human eyes have four kinds of detectors in the retina. Rod cells (2 μm diameter) designed to work mainly at low light levels over a wide range of wavelengths (panchromatic) through a process that actually depletes the light-sensitive chemical (rhodopsin), and three different types of cone cells. Cone cells (6 μm diameter) contain a different protein (iodopsin) and provide color vision because each type (S, M, or L in a 1 : 4 : 8 ratio) is sensitive to short, middle, and

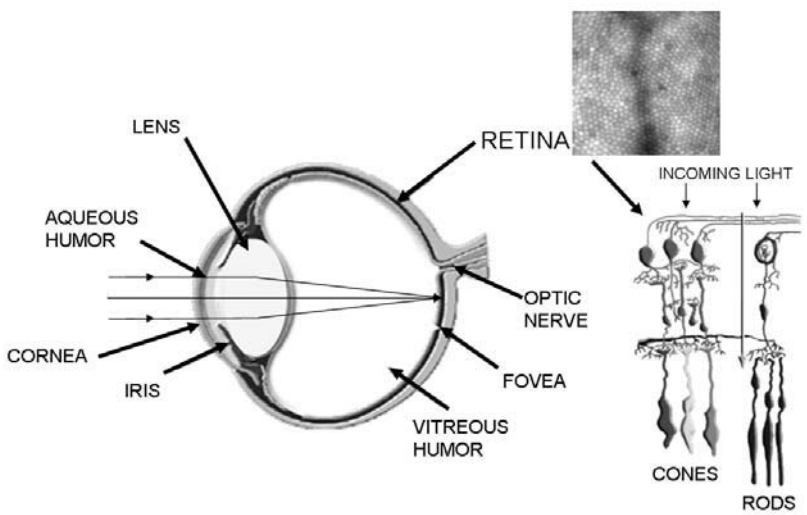


Figure 1.5. The basic construction of the human eye. The lens forms an upside-down image of the scene on the retina which detects and converts light into electrical signals that go to the brain. An adaptive-optics assisted view of part of the retina is shown in the insert. All the elements of an electronic camera are present in this biological system.

long wavelengths that correspond roughly to blue, green, and red light. The cones make up only 5% of the total number of photoreceptors (over 125 million) in the eye and are concentrated to the center of the retina. Although the human eye has only a single lens, its shape can be changed by the action of muscles, and the amount of light reaching the retina through the pupil, the colored part of the eye, is regulated by a variable-sized aperture called an iris. The eye's single lens forms an upside-down image on the retina which the brain inverts back to upright. Remarkably, the retina is a curved photo-detector following the eye's curved focal surface. The eye updates the brain very rapidly (about 30 times per second) and operates more like a video camera than a shutter camera. One remarkable feature of the eye–brain camera system is a high contrast ratio, meaning that it has a remarkable ability to detect fine structure in scenes that can range in brightness by over 1 billion to one. Over a period of time ranging up to one hour, the eye can adapt to low light level conditions. This process, called dark adaptation, is partly chemical in nature but also seems to involve longer exposure times and is certainly well-known to astronomers. The dark-adapted eye is least sensitive to red light, which is why a red flashlight is handy for visual observers. Another feature of vision that astronomers make use of when looking for faint objects is called "averted vision" in which the best sensitivity to faint objects is obtained by looking slightly away from the central field. This trick works because the density of rods is greatest away from the central field where the cones are concentrated. If you use this method on a dark night with no Moon and no city lights you can glimpse the Andromeda spiral galaxy which is about 2.5 million lightyears away. The eye–brain system can also record an "after-image" which is typically seen after exposure to a bright object that caused you to close your eyes. As we will see later, electronic cameras across the spectrum can also suffer from a similar artifact. An interesting phenomenon often discussed by visual astronomers is the Purkinje effect in which stars of the same brightness but different color appear to have different brightness, under low light level conditions. This effect is caused by the eye switching from photopic (cone) vision to scotopic (rod) vision. In the latter case, the peak sensitivity is bluer and so reddish objects appear fainter than bluish objects. Among the observables of intensity, color, and polarization, the eye has the hardest job with polarization. The phenomenon of Haidinger's brush allows the eye to detect polarization of light. It can be seen as a yellowish horizontal bar or bow-tie shape with fuzzy, brush-like ends, and sometimes with fainter bluish vertical areas, in the center of the visual field against the blue sky while facing away from the Sun; blue sky light is highly polarized 90° from the Sun. Haidinger's brush is usually attributed to the phenomena of dichroism and birefringence that we will encounter in later chapters on optical materials.

As mentioned previously, the ability to distinguish two closely spaced objects from a distance is called angular resolution, and the human eye has an angular resolution of about 1 minute of arc or 0.003 radians. A radian is a unit of angular measure; there are 2π radians in 360°, and therefore 1 radian $= 57.296^\circ = 206{,}265$ seconds of arc. Thus, the angular resolution in the diffraction limit is just λ/D using radians. For scale, the angular diameter of the Sun and Moon is about 30 minutes of arc. (Note that two objects of very different physical size, such as the Moon and the

Sun, can have the same angular diameter if located at appropriate distances from us. This is why the smaller, closer Moon can block out the larger, distant Sun during a solar eclipse.) So while the eye's resolution may seem quite good, it is not sufficient to discern the disks of the planets or resolve more distant astronomical objects. Under ideal conditions (i.e., when limited only by the wave nature of light) the dark adapted human eye has a pupil diameter of about 7 mm and therefore its diffraction limit (λ/D) for visible light is about 14 seconds of arc (almost one-quarter of a minute of arc). A modest telescope or a good pair of binoculars with a larger objective lens (e.g., 50 mm) can easily achieve a (diffraction-limited) resolution of about two seconds of arc. Moreover, large eyes collect more light in proportion to their area, hence creatures with big eyes can see better in the dark. Likewise, the larger the diameter of your telescope the more light it can collect. Given that the focal length (f) of the eye is about 22.3 mm and the aperture of the pupil is about 7 mm, then the ratio of these two ($F = f/D$), the focal ratio, is about 3.2. Finally, the average human eye is not sensitive to light with wavelengths smaller than about 390 nm (the near ultraviolet) or larger than about 780 nm (the near infrared). This range corresponds to the interval of wavelengths in which the Sun emits the most energy, and consequently this limitation gives us a biased view of the universe.

1.2.2 Photography

The blackening of silver halides when exposed to light gave birth to astronomical photography in the latter half of the 19th century. Basically, when light is absorbed by tiny crystals or "grains" of slightly soluble silver halide salts, like silver bromide (AgBr), there is a reaction of the form $Ag^+Br^- + \text{light energy} \rightarrow Ag^+ + Br + e^-$ in which a negatively charged electron (e^-) is released from the bromine ion (Br^-) and then combines with the positively charged silver ion to make a neutral silver atom ($Ag^+ + e^- \rightarrow Ag$). Grains containing free silver are said to contain a "latent image". During the chemical development process much more silver is produced around that initial deposit to amplify the effect and then any unexposed silver bromide is eliminated by the "fixing" process. Modern photographic emulsions are sensitive to a wider range of wavelengths than the human eye: from the ultraviolet to the near infrared. Because it is difficult to relate the sensitivity to illumination of one part of a plate to another, brightness cannot usually be measured to high accuracies, compared with the corresponding exercise for electronic imaging devices. In addition, the ever-present albeit faint light from the night sky eventually saturates the emulsion, rendering it insensitive. This behavior is illustrated by the "characteristic curve" of the emulsion which is a plot of the "density" of the developed plate vs. the logarithm of the "exposure". In this context, the density of a plate is a measure of how exposed (blackened) it is. Density (D) is just the logarithm (see Appendix A for information on logarithms) of the "opacity" (O) of the plate, which in turn is the reciprocal of the transmittance (T), defined as the ratio of the amount of light transmitted (I_{out}) by the

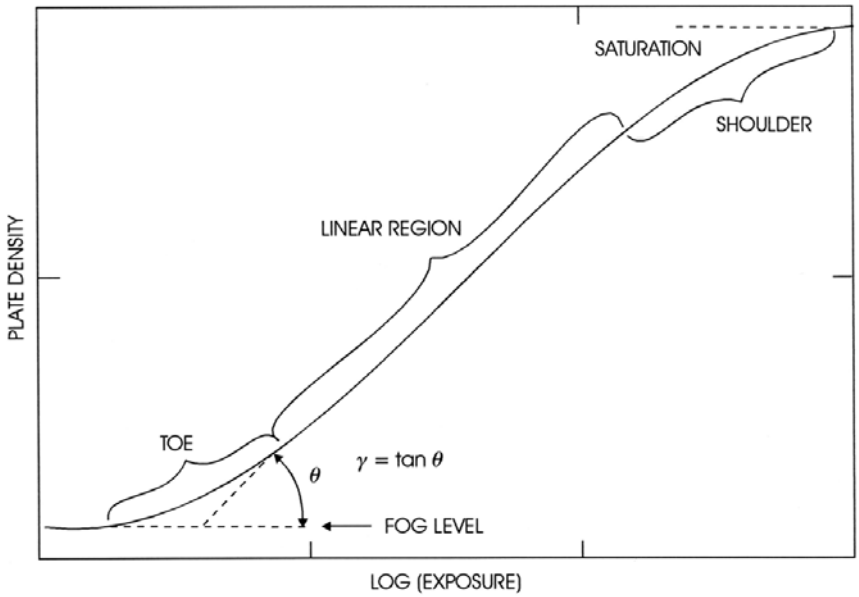


Figure 1.6. The characteristic curve of a photographic plate is a plot of density vs. exposure. The plate is linear over a limited range of exposure. Linearity at the lowest light levels is the goal of electronic imaging.

plate to the amount that is incident (I_{in})

$$D = \log(O) = -\log(T) = -\log\left(\frac{I_{out}}{I_{in}}\right) \tag{1.2}$$

The term "exposure" measures the *total* amount of energy falling on a unit area of the plate and is found by multiplying the irradiance (E) of the light (in watts/m^2) by the exposure time (t). A typical characteristic curve is shown in Figure 1.6. Note that a minimum density occurs even for no exposure; this is the "fog" level of the plate. Then there is a short curved region in which the response suffers the greatest reciprocity failure which means that the response of the plate is not linearly proportional to the total energy. As we will see, this non-linear behavior is in stark contrast to that of modern electronic imaging devices. Beyond a certain threshold in exposure the characteristic curve does become almost a straight line and the response of the plate to light is almost linear; the slope of this line defines the "gamma" for the plate ($\gamma = \tan\theta$). As gamma increases, the contrast improves for a properly developed plate. Once all the grains are blackened there can be no further increase in density and so the curve "turns over" at large exposures. Unfortunately, the characteristic curve can differ even between two emulsions of the same type. Clever techniques, such as hyper-sensitization, have been developed for treating photographic plates. For example, Kodak IIIaJ plates (Eastman Kodak, 1987) are "soaked" in nitrogen and hydrogen gas, in order to detect much fainter objects. All-sky photographic

surveys carried out from the 1950s to 1980s using special telescopes (in California, Australia, and Chile) have now been digitized and are available online. These important surveys and techniques will be reviewed in Section 4.1. Nevertheless, it is the very low efficiency with which light is converted to blackened grains that limits the sensitivity, if not the utility, of photographic plates in astronomy.

1.2.3 The photoelectric effect

From the 17th century, when Christiaan Huygens (1629–1695) and Isaac Newton (1642–1727) carried out fundamental experiments on the nature of light, evidence existed that light was a form of energy which tended to interact with matter in a way which suggested that the energy was transported as a "wave". For example, two beams of light could be made to "interfere" and produce a pattern of light and dark regions, very similar to the high crests and flat calms produced when two identical water waves meet. The effect is easily observed by dropping two pebbles simultaneously into a calm pond. Where two crests meet the amplitude of the wave is doubled, but where a crest and a trough meet there is cancellation and the water is calm. Eventually however, situations in which light behaved more like a stream of particles were encountered. It was not until the end of the 19th century with the work of Max Planck (1858–1947) that it was understood that light can carry energy only in specific amounts, as though it came in individual packets; Planck called these packets *quanta*. For light (or any electromagnetic wave) with a frequency of vibration ν, the energy of one quantum is given by

$$E = h\nu = \frac{hc}{\lambda} \tag{1.3}$$

where h is a constant of nature called Planck's constant ($= 6.626 \times 10^{-34}$ J s).

In the early 1900s Albert Einstein (1879–1955) worked on a phenomenon called the photoelectric effect and showed that a quantum of radiant energy (now called a "photon" of light) could eject a negatively charged electron from the atoms in certain materials, and that the photon of light behaves like a particle, with energy and momentum, and yet has a "wavelength" associated with it; the momentum of the photon is given by $p = h/\lambda$. It was also discovered that a beam of electrons impacting on the surface of certain materials could cause the ejection of photons of light (or phosphorescence), and these developments led to the invention of cathode ray tubes (CRTs), valve (tube) amplifiers, radio, and ultimately to television.

1.2.4 Photomultiplier tubes

In the late 1940s and early 1950s, a photoelectric device called a photomultiplier tube (PMT) became widely available for accurate measurements of starlight. In such a device, a photon or quantum of radiation strikes the surface of a certain type of "photoemissive" material which responds by emitting an electron, provided that the photon energy ($h\nu$) exceeds a minimum energy ($W = h\nu_c$) called the work function of

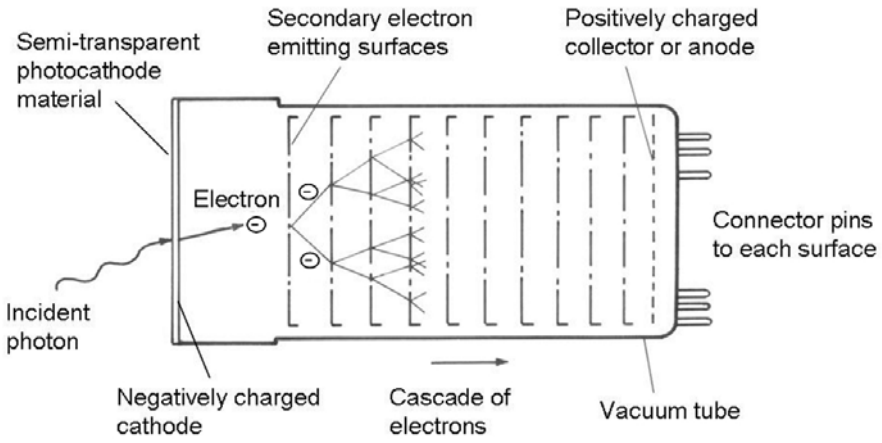


Figure 1.7. The photomultiplier tube produces a large cascade of electrical current when illuminated with faint light. The photoelectric effect in thin slabs of certain materials causes emission of a negatively charged electron when hit by a photon of sufficient energy. Electrons are drawn toward the positively charged anode by an electric field, emitting secondary electrons from intervening surfaces to produce a cascade.

the material. If the photon's energy is less than this critical value, then no emission occurs irrespective of how intense the light source is. Smaller than a certain cutoff frequency (ν_c) or larger than a certain cutoff wavelength ($\lambda_c = c/\nu_c$), the photon simply has insufficient energy to overcome the work function; adding more photons with the same limitation in energy does not help. Within the glass-encapsulated photomultiplier tube, the negatively charged electron ejected by a photon of sufficient energy is accelerated by an electric field and made to impact another photoemissive surface called a dynode (Figure 1.7). The electric field (E) is established by a voltage potential of about 150 V between succeeding dynode stages; note that this implies about 1,500 V (or 1.5 kV) across a tube with ten stages. On impacting the first dynode surface, the initial photoelectron generates three new electrons which are all accelerated and directed at yet another dynode surface where each electron generates three more. This cascading process is continued several more times until a final surface collects what is now a huge pulse of electrons (Q) generated by just one photon of light; in general, $Q = q^n$ where q is the number of electrons per emissive surface and n is the number of surfaces. All that needs to be done is to "count" these pulses for a given (arbitrary) time interval and determine the count rate; hence, the process is called pulse counting or photon-counting.

The beauty of this detector lies in the fact that the precision of the measurement depends only on the total number of photons counted. In fact, if N is the number of photons counted in a given time, then the random error or "uncertainty" in N is found to be the square root of this number $\sqrt{N}$; this kind of behavior is called Poisson statistics or photon-counting statistics and represents a fundamental limit (see Appendix F for more information on statistics). Random errors are also called

"noise" by analogy to acoustic sounds when completely jumbled together. In keeping with standard practice we will refer to all sources of random errors as noise. If you record 100 photons on average, then many repetitions of this measurement will yield 100 ± 10 photons or answers in the range 90–110 photons most of the time; this is a 10% range in uncertainty. For 1 million photons the random error is 1,000 which is only 0.1% of the average, a much smaller random error.

In some cases, the efficiency of photomultiplier tubes can be ten times that of a photographic emulsion, and indefinitely long measurements (equivalent to exposures) can be used to build up counts to detect extremely faint objects. The PMT was a major innovation because it cheaply and easily allowed astronomers to establish reproducible and accurate brightness measurements of stars. Initially, the most common photocathodes were the S-11 (Cs_3Sb on MnO) and the (Cs)Na_2KSb tri-alkali S-20 which responded from the UV to 600 nm and 700 nm, respectively. Although S-1 material (Ag–O–Cs) went beyond 1,000 nm into the near infrared, the response was very low (<1%). Most photomultipliers in use today have gallium arsenide (GaAs) photocathodes because these provide very good response from the UV to about 900 nm; these PMTs are also known as "negative electron affinity" devices. The real drawback, of course, is that only one tiny patch of the sky can be observed at a time; a photomultiplier tube has no "panoramic" or camera-like advantage. It is like a single cell on the retina of the eye.

1.2.5 Electron-scanning devices

Alongside the development of photomultiplier tubes came the TV tube and work on electronic image "intensifiers" for very low light level applications. The aim was to combine the attributes of accuracy and unlimited exposure time of the photomultiplier tube with the extended field of view of the TV camera. Military applications for ultra low light level camera systems helped to stimulate this technology and an astonishing variety of complex television-type image-intensified schemes were proposed and tested for astronomy in the 1960s and 1970s. Most of the devices begin with the light striking a photoemissive surface to release an electron, but a variety of ingenious methods were then employed to amplify the flow of electrons. Some image intensifier stages use electric fields (electrostatic amplification), others use magnetic fields, and some use an intermediate stage to generate brighter light by means of a phosphorescent screen. Detection of the final amplified stream of negatively charged electrons, or amplified light emission, is by some form of TV tube. The characteristic feature of all the various types of TV imaging devices is the use of a beam of electrons to "scan" across a region called the "target" on which there is an electrical charge pattern representing the optical image; the generic term for these electron-scanned imaging devices is vidicon (see Figure 1.8). In essence, the electron beam completes the circuit and provides a source of current to reset the target after exposure to light; the amount of current required for this action is the "video signal". TV tubes themselves were greatly improved during this time with the introduction of a range of variants called the SEC, SIT, and silicon vidicons and the Plumbicon or lead oxide vidicon, which was the primary "studio" camera for many years. All of the TV-based

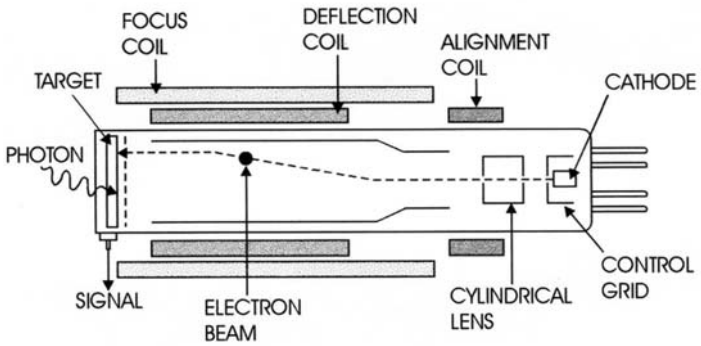


Figure 1.8. A schematic representation of a vidicon (TV tube) is shown to emphasize its electron-scanning beam. Although relatively compact, these imaging devices required high voltages and were difficult to use for astronomy.

image-intensified camera systems were complex, physically large, highvoltage (tens of kilovolts) systems. Some found success as faint-object TV track-and-guide cameras on large telescopes, but most turned out to be difficult to use for accurate brightness measurements in astronomical cameras and spectrometers. Problems with the stability and reproducibility of the detector, critical for scientific research, often seemed to preclude effective techniques for the removal of systematic errors in long exposures. The most successful of the image-tube devices were those which appealed to photon-counting rather than attempting to produce an amplified copy or "analog" image of the scene. Two of the best known systems at the time were the Wampler Scanner and the Boksenberg Image Photon Counting System (IPCS). In the Wampler Scanner, developed at the University of California's Lick Observatory by Joe Wampler, the amplified light emission from the intensifier stage was fed to a unit called an "image dissector" which has a photocathode and an electromagnetic focusing mechanism which directs photoelectrons to a small rectangular aperture behind which is an electron multiplier stage just as in a normal photomultiplier tube. The area of the photocathode seen by the electron multiplier section can be "scanned" by controlling the electromagnetic focusing coils. In the IPCS, developed at University College London by Alec Boksenberg and others, every fraction of a second a frame of the TV output was digitized and sent to a computer which immediately analyzed the pattern in "real time" to count photons. Saturation levels for the IPCS were around one event per pixel per second, which is a very low light level. The original Faint Object Camera (FOC) on the Hubble Space Telescope (launched 1990) was an IPCS system.

Ironically, perfection in TV tube construction was reached just about the time that a major breakthrough came in semiconductor devices. Things first began to take a different trend when in 1958 the first multitransistor integrated circuit (IC) was demonstrated only 10 years after the invention of the transistor itself at Bell Labs,[1]

[1] In 1996 Bell Labs and other AT&T research centers were incorporated into a new organization called Lucent Technologies, now the Alcatel-Lucent Bell Labs.

New Jersey, U.S.A., in 1948. In the first transistors a small crystal of germanium was used; by adding other materials into the crystal structure it was possible to create a small sandwich which amplified electrical currents just like a (much larger) glass-encapsulated tube or valve. Soon, silicon surpassed germanium as the best material to build transistors, the devices themselves got smaller as fabrication methods improved which led to the integrated circuit in 1958. During the 1960s better and better manufacturing processes led to smaller and smaller integrated circuit units still capable of electrical functions. Today, silicon chips smaller than a thumbnail can be constructed that contain millions of transistors (amplifiers).

One simple device easily manufactured from silicon technology is the photodiode, in which the strength of the electric current flowing through the device is proportional to the amount of light falling on it. Small, cheap, and low-powered, silicon photodiodes also had the immense benefit of very high quantum efficiency (QE); that is, 80%–90% of the incoming photons get converted to electrical charge. Compare this with1%–2% for photographic emulsions (perhaps 4% with hypersensitization) and 10%–20% efficiency for the tube-type systems. Moreover, by the nature of the properties of silicon, this kind of light sensor could be used far into the red, well beyond the limit of sensitivity of the human eye and much better than any existing detector. As imaging devices, silicon photodiodes were limited primarily by structural awkwardness in making a large densely packed two-dimensional (i.e., having width and length) assemblage or ''array'' of photodiodes with no gaps between them. Nevertheless, ''linear arrays'' of diodes (one or two lines wide), used directly (Reticons) and behind image intensifiers (Digicons), found applications in astronomical spectrometers and polarimeters (e.g., McLean *et al.*, 1979). Several systems like this came into operation in the late 1970s and some were used in first-generation instruments on the Hubble Telescope (Reticon and Digicon are trade names).

1.3 THE IMPACT OF SOLID-STATE IMAGING

In the late 1970s however, a new device challenged and virtually vanquished all contenders. It was the *charge-coupled device* or CCD. The CCD is an array of microscopic square-shaped light-sensitive regions arranged in a checkerboard pattern. Tiny light-sensitive squares, usually called ''pixels'' (for picture elements) are formed directly in a slab of silicon. These pixels are so small that they cannot be seen simply by examining the surface of the CCD by eye. The introduction of these devices into astronomy was revolutionary, and sent repercussions through not only optical astronomy, but also X-ray and infrared astronomy as well.

1.3.1 Invention of the CCD

Toward the end of 1969, two researchers, Willard S. Boyle and George E. Smith (Figure 1.9) at the same Bell Labs in Murray Hill, New Jersey, where the transistor was invented, were investigating new ways of imaging with solid-state, silicon

Figure 1.9. Willard S. Boyle (left) and George E. Smith, inventors of the charge-coupled device at the Bell Labs research center at Murray Hill, New Jersey in December 1974 when they received their patent for the CCD. Credit: Lucent Technologies Bell Labs.

methods in an effort to develop a Picturephone! This concept involved having a tiny, inexpensive solid-state camera built in to the telephone receiver to enable callers to see each other. Market research later failed to convince Bell Telephone of the worth of the Picturephone concept at that time; Picturephone is a registered trademark of Bell Labs AT&T. As luck would have it, the method devised by the two scientists turned out to be an incredible innovation, destined to change the whole philosophy of imaging away from vidicon-type TV tubes and even from celluloid movie film. Research on magnetic bubble memory, three-phase plasma display panels, and silicon vidicons was in full swing at Bell Labs when executive director of the semiconductor division, Bill Boyle, and his close friend George Smith, who was department head in charge of developing a silicon diode array camera tube for Picturephone applications, got together in front of a blackboard one afternoon and began musing about the idea, as George put it, of "friendly competition" from silicon technology. Bill asked George what about an "electric bubble". The obvious analogy to passing a magnetic domain from one site to another was to pass charge from one site to the next. It was already known that charge could be stored by insulating a small metal plate placed on the surface of a silicon crystal, but it was the concept of stringing these storage sites together and using voltage differences between them to pass the charge along that constituted the new idea. Devising the basic concept took only a couple of hours and within a few weeks George had a nine-element device under test! They began their paper, published in the *Bell System*

Technical Journal (Vol. 49 in April 1970), with the following words "A new semiconductor device concept has been devised which shows promise of having wide application." Their invention was of course the charge-coupled device, a name chosen by Bill Boyle. I still remember the excitement in Bill's voice during a 1986 telephone conversation from his home in Nova Scotia when I was working on *Electronic and Computer-aided Astronomy* as he recalled for me the tremendous reception given to the idea following its announcement. Mike Tompsett, who came to Bell Labs from English Electric Valve (EEV) in the U.K. where he had worked on the development of the pyroelectric vidicon and himself an author of a classical text on CCDs recalls, "CCDs were an idea of the time. Within a few months all sorts of applications were listed—and many were actually of relevance to the phone company." Bill and George recall reactions to their idea as ranging from "I should have thought of that" to "it will never work"! But it did work. Experimental verification of the charge-coupled device concept was published in the same issue of the *Bell System Technical Journal* by Gil Amelio, Mike Tompsett, and George Smith (Amelio *et al.*, 1970). Bill Boyle and George Smith have been honored many times for the innovation of the CCD concept, most recently in 2006 by the National Academy of Engineering's Draper Prize.

CCDs have a great many advantages over other electronic and photographic imaging devices. They are small, linear in their response to light, stable, low-power devices with excellent sensitivity over a very wide range in wavelengths and a huge range in light levels (see Figure 1.10a, b). Moreover, there was the likelihood of relatively low cost due to silicon manufacturing processes and mass production. Commercial applications of CCDs now range from home video/still cameras to professional broadcast cameras; remote-sensing; surveillance; robotics; medicine;

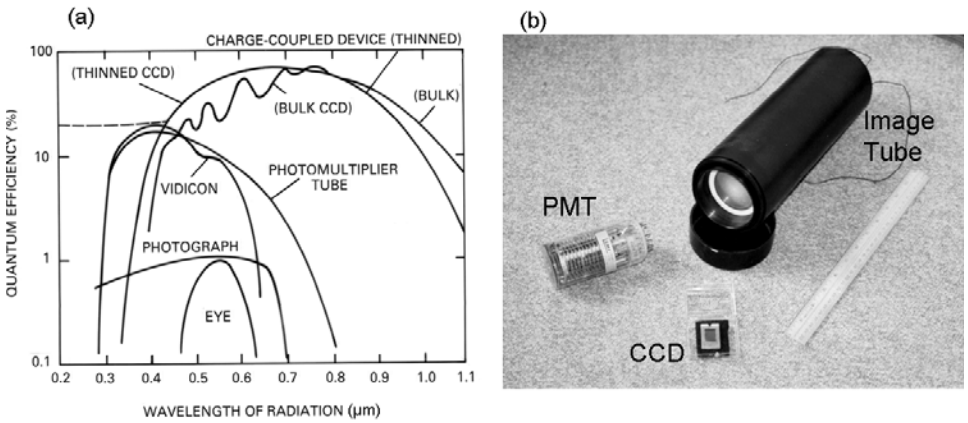


Figure 1.10. (a) An illustration of the sensitivity or quantum efficiency of certain CCDs to light of different wavelengths compared with other forms of detectors. Note that the scale on the left increases by factors of 10! Modern CCDs are even better than the ones illustrated here. One of the CCD's attributes is wavelength coverage. (b) A picture comparing the sizes of a CCD chip, a photomultiplier tube (PMT) and an image intensifier (black tube); the ruler is 12 inches long.

science, and many other applications. CCD-related terminology, like the word "megapixels" (Mpxl) has crept into daily speech. For astronomical applications CCDs represented such a dramatic improvement over all other imaging techniques that they were promptly embraced by astronomers. Indeed, it is fair to say that astronomers and space scientists were in the forefront of the pioneering days of CCDs. One of those very early developments was an effort by Texas Instruments to provide large CCDs under contract to the NASA Jet Propulsion Lab (Janesick, 2001) for inclusion in instruments being proposed for the Jupiter Orbiter Probe (later renamed the Galileo mission) and for the Space Telescope (later renamed the Hubble Space Telescope).

In an introductory paper to the conference on "Solid State Imagers for Astronomy" organized by Harvard astronomers John Geary and David Latham (June 10–11, 1981) at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts, Herbert Gursky of that institute opened the meeting with the following remark:

> "There is by now enough information, in the form of laboratory data on a variety of devices and published scientific results, to confirm the original assessment that solid state arrays will become a permanent entry in the repertory of astronomical instruments."

He was right! No major observatory in the world lacks a suite of CCD cameras, and CCD cameras are widespread among amateur astronomers too.

Astronomers now "design" their own CCDs and collaborate to organize runs at "silicon foundries" to make devices better suited for astronomy applications. Although the first CCDs used in astronomy had 10,000 pixels in an array of 100×100, devices are now available with 4 Mpxl–16 Mpxl, and astronomers try to package more than one CCD into an instrument by forming a "mosaic" of detectors, like this one in use at the Canada–France–Hawaii Telescope (CFHT) on Mauna Kea (Figure 1.11). Current projects have now reached the level of billions of pixels (gigapixels; Gpxl) in the focal plane. Later chapters will explain how CCDs work and what astronomers have to do in practice to achieve spectacular images, such as Figure 1.12 (see also Plate 1) and others throughout this book.

1.3.2 Electronic imaging across the spectrum

One remarkable feature of the CCD is that it can also be used to detect X-ray photons, and so you will find X-ray cameras using CCDs on the Chandra X-ray Observatory (CXO) in space, for example. While modern CCDs are also used for ultraviolet work, other competing imaging devices exist for this region and will be discussed in detail too. Stimulated in part by the impact of CCDs a decade earlier, an enormous revolution in infrared astronomy occurred with the advent of solid-state imaging devices for that wavelength region in the 1980s (e.g., McLean, 1988). In fact, the impact was perhaps even greater than CCDs because there was no precursor for infrared imaging, no equivalent of photography or TV tubes. Infrared "arrays" do

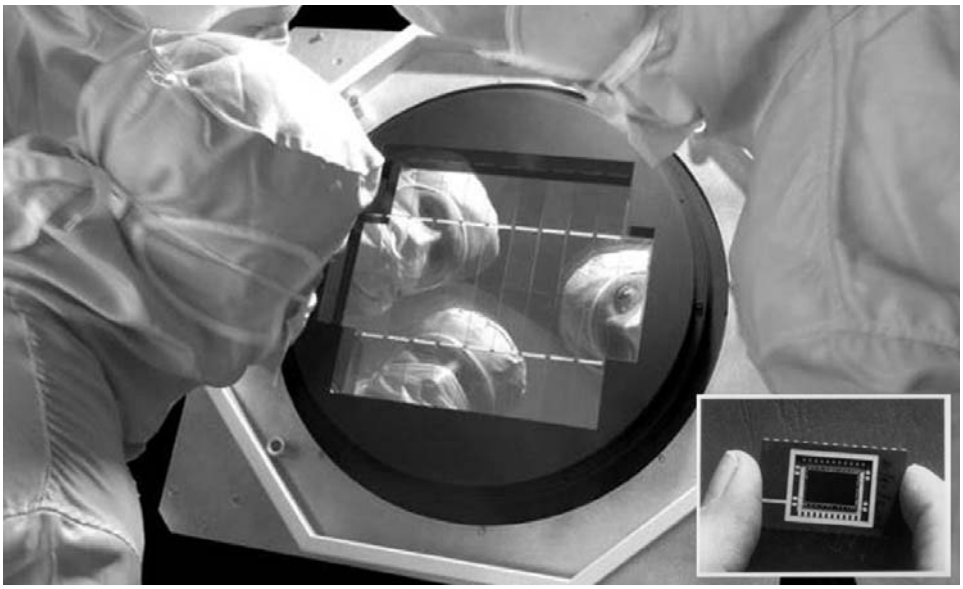


Figure 1.11. A large mosaic of CCDs is in use at the Canada–France–Hawaii Telescope (Mauna Kea). A total of 40 CCDs from *e2v* technologies (formerly EEV) are closely packed; the central 36 (4×9) yield 340 Mpxl. See Plate 2 for an image from this camera. Credit: the CFHT Corporation. Insert: An early CCD with about 250,000 pixels also made by *e2v*, Chelmsford, U.K. Credit: Paul Jorden (*e2v*).

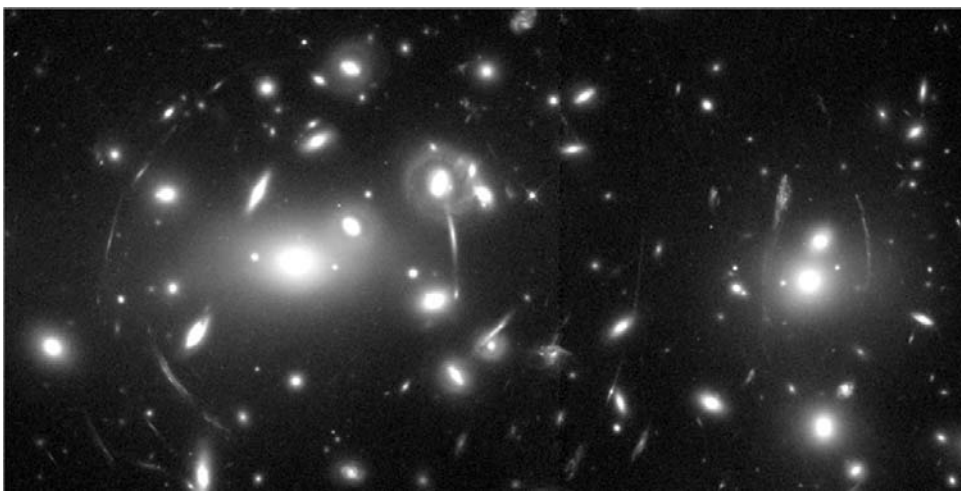


Figure 1.12. A deep CCD exposure obtained with the Hubble Space Telescope illustrates the curved arcs caused by Einstein gravitational lensing of the light from background galaxies by matter in the foreground cluster (Abell 2218). Credits: NASA, Andrew Fruchter, and the ERO Team [Sylvia Baggett (STScI), Richard Hook (ST-ECF), Zoltan Levay (STScI)] (STScI). See also Plate 1.

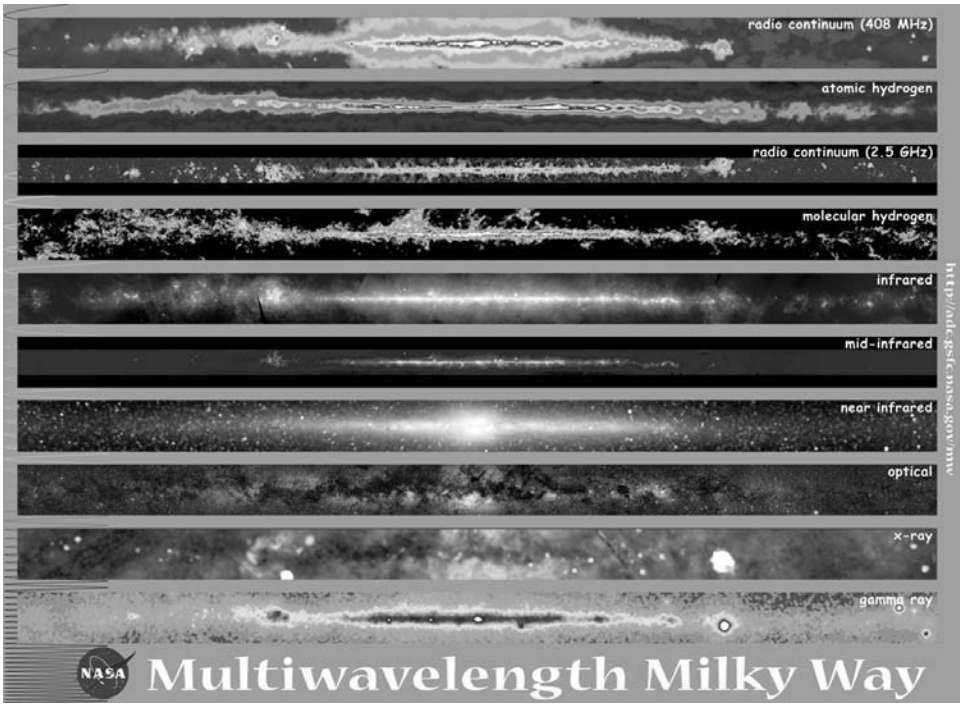


Figure 1.13. The multi-wavelength sky: views of the entire Milky Way are shown in this NASA mosaic using data from gamma rays to radio waves. A color version of this plot is included in the color plate section. Credit: NASA's Goddard Space Flight Center. See also Plate 3.

not employ the charge-coupling principle, and materials other than silicon are required to detect low-energy infrared photons, but high-density pixel arrays of infrared photon detectors are now available for the region from 1 μm to 150 μm. New devices relying on the phenomenon of "superconductivity" have been used to create very sensitive detector arrays to extend the range of wavelengths up to about one millimeter. Radio astronomers do not have cameras as such but they can still create pictures of the sky with exquisite detail using mapping techniques and sensitive receivers employing solid-state mixers. CCDs even re-appear in some radio systems that employ acousto-optic spectrometers. Astronomers have also obtained "images" of the cosmos by collecting high-energy gamma rays thanks to space missions like Integral and Swift. Consequently, we now have electronically generated views of the sky that cover the entire electromagnetic spectrum from gamma rays to radio waves (see Figure 1.13 for a compilation of views of the Milky Way) (see also Plate 3). All of the electronic detectors and methods mentioned only briefly here will be discussed in later chapters.

That the universe may look quite different at non-visible light wavelengths can be understood by remembering that the Sun emits a continuous spectrum of electromagnetic radiation. This emission is strongly peaked in the middle of the visual band

because the controlling factor is the "surface" temperature of the Sun which is about 5,800 degrees Kelvin (K). The Kelvin temperature scale (named for physicist Lord Kelvin, 1824–1907) has the same intervals as the Celsius scale but starts at absolute zero, the temperature at which particles have minimum energy (-273.15°C). Overall, the Sun's energy spectrum is close to that of an ideal emitter or blackbody and the energy emitted at any wavelength can be predicted using Planck's function (see Appendix C) if the absolute temperature of the body is known. The wavelength corresponding to the peak of the radiant emission is given by Wien's Displacement Law (Wilhelm Wien, 1864–1928) which shows directly the relationship with temperature ($\lambda_{\max} = \text{constant}/T$); expressing wavelengths in microns, then the "constant" in the equation is about 2,900 and so for $T = 5{,}800$ K the wavelength of peak emission is $2{,}900/5{,}800 = 0.5$ μm in the middle of the visible spectrum. Following this law, stars much hotter than the Sun emit most of their electromagnetic energy in the ultraviolet, whereas stars much cooler than the Sun emit mostly in the infrared. Exceedingly hot regions, like the accretion disk around a black hole, will emit X-rays while extremely cold interstellar clouds of gas are observed to emit millimeter waves.

1.4 THE CONTINUING GROWTH OF ASTRONOMICAL TECHNOLOGY

New discoveries about the universe have been coming at a remarkable pace in recent years, ranging from the very nature of matter and energy, to the discovery of other planetary systems beyond the solar system. These breakthroughs in no small way rely on the development of new technologies for astronomy. Among the technologies that have made a difference, both on the ground and in space, are larger and more sensitive UV/optical/IR detectors enabling more powerful cameras and spectrometers; improved methods for building larger telescopes; advances in optics and detectors for X-ray astronomy; the success of devices for the study of the cosmic microwave background; new digital signal-processing techniques and new receiver/antenna designs for radio astronomy. But perhaps the most impressive of all these new developments has been the advent of techniques that enable large ground-based telescopes to operate at their ultimate diffraction limit, unfettered by atmospheric turbulence, as if in outer space. These topics are mentioned briefly here to motivate further discussion in later chapters.

1.4.1 Bigger telescopes and better images

Today, carefully designed optical instruments have transmission factors and efficiencies that are virtually impossible to improve due to fundamental effects associated with the refractive and reflective properties of materials and the quantum efficiency of the detector itself. Once the quantum limits of sensitivity in detectors and instruments has been reached, the only way to gain large factors in efficiency is to construct even larger ground-based telescopes and to develop methods for counteracting the image-blurring effects of turbulence in the Earth's atmosphere.

With the combination of high-speed digital image processing and CCD technology, it is now possible to implement strategies which "optically correct in real time" the distortions in an image caused by passage of the light through the Earth's atmosphere. The promise of this new field of "adaptive optics" is that the very large telescopes now in operation around the world will perform as if they were located in outer space, instead of at the base of our turbulent atmosphere. Adaptive optics (or AO for short) relies on the principle that if the disturbance in the wave can be determined on a short timescale (milliseconds), then optical counteractive measures can be applied to return the wave to its ideal, flat condition before it reaches the science instrument. Suitable measurements can be obtained by observing bright stars near the object of interest; this is called natural guide star adaptive optics. Alternatively, the option now exists to create an "artificial star" with a laser beacon projected from the telescope; this method is called laser guide star adaptive optics. Provided that the signals are strong enough, the overall atmospheric conditions are fair, and that the correction system works fast enough, then the individual diffraction-limited images can be prevented from blurring out into a large seeing disk. Astronomers describe the efficiency of adaptive optics using the ratio of the measured peak intensity in the stellar image compared with the value expected for the pure diffraction-limited case. Values now approach 50% in the infrared with some AO systems. Compare this with values of less than 0.5% for atmospheric seeing-limited images. Something like 100 times as much light is being concentrated into an image that is 10–20 times smaller (sharper) than before.

1.4.2 Common-user instrumentation and automation

As telescopes have grown in size and instruments have become more costly, a strong trend toward shared or common-user facilities and large collaborations has occurred. Usually, significant effort is expended to ensure that the instrument is modular in design and reasonably flexible to allow innovative additions at a later time. To be competitive, each common-user instrument must be very sensitive, easy to use, and produce computer-compatible data in a form that can be readily handled by the guest astronomer at his/her home institute. The instrument must have a high degree of reliability not only because the original builders/designers may not be around to participate in every experiment done with it, but also because such instruments tend to become workhorses, and get used night after night on the cold mountaintops where the world's major observatories reside. Inevitably, therefore, these instruments must be well-engineered.

Modern astronomical instruments are highly automated. Observational sequences can be pre-programmed and linked to the motion of the telescope, and in some cases the entire system can be operated remotely from the other side of the world. Astronomers have found that automation pays dividends in scientific flexibility too. By providing a modular package of control programs and other software options, modern instruments can be reconfigured to optimally match the needs of the most ingenious researcher without recourse to manual adjustments. It is this feature of reprogrammability which ensures that the extra cost and extra effort

which must be expended to design and construct a common-user instrument is worthwhile.

1.4.3 Surveys and data archives; the virtual observatory

The last decade of the 20th century was remarkable in astronomy for many reasons. Not only did many large telescopes with apertures of 8 m–10 m appear, but also many smaller telescopes were built and designed to capitalize on the power of electronic imaging devices by performing deep astronomical surveys. Perhaps the best known of these is the Sloan Digital Sky Survey (SDSS). This remarkable project, spearheaded by Jim Gunn of Princeton University, uses a dedicated 2.5 m telescope located on Apache Peak, New Mexico to digitally map half of the northern sky to about 23 magnitude in several bands from the ultraviolet to the near-infrared. About one million galaxies and 100,000 quasars are then selected for spectroscopic determination of their distances using the same wide-field telescope, thus enabling a 3-D map of their distribution in space to be created. First light with the SDSS was obtained in May 1998 and routine operations began in April 2000. Two powerful instruments, incorporating 54 CCDs, are used for this survey: a large CCD camera to image 1.5 square degrees of sky at a time (8× area of full Moon) and a pair of spectrographs that can get spectra of over 600 galaxies in a single observation. Special software and computers keep pace with the enormous data flow. For example, in June 2007 the survey's sixth public data release (DR6) was announced with over 10 terabytes of images and spectra in accessible online catalogs. As Jim Gunn emphasizes, the scientific value of such a data set is vast, ranging from critical investigations of large-scale structure in three dimensions to the relationships of galaxies with their environments. Funded in part by the Sloan Foundation, the Digital Sky Survey is a joint venture involving Princeton University, the Institute for Advanced Study, Johns Hopkins University, University of Chicago, Fermi National Accelerator Lab, and the Japan Promotion Group. However, plans for next-generation surveys are already under way, such as the proposed Large Synoptic Survey Telescope (LSST) led by Tony Tyson (University of California, Davis) which will employ cameras with over 1 billion pixels and cover the observable sky every 3 days or so.

Consequently, there is a continuous worldwide effort to establish "astronomical data centers" where the digital data from a wide variety of astronomical sources, both ground-based and space, are not only safeguarded, but can be accessed electronically using facilities available on the World Wide Web. Several electronic astronomical data archives now exist. Among the earliest were the Canadian Astronomy Data Center (CADC), the Strasbourg Astronomical Data Archive, and the Astrophysics Data System (ADS) in the U.S.A. The latter is a distributed data system chartered to provide access to information gathered from NASA space missions. The roots of the development of ADS began in the late 1980s with the realization that the data volume from projected NASA missions was going to be enormous and that some action to gather and archive the data was going to be essential. The first operational release of ADS came out in June 1991 and was updated to a Graphical User Interface in 1992/1993, and in 1994 the service became available on the World Wide Web. The ADS

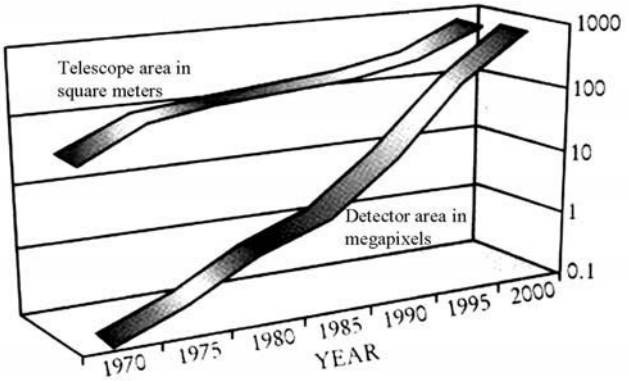


Figure 1.14. Graphs from the international virtual observatory alliance which illustrate the data explosion facing modern astronomy. Upper plot is the increase in telescope collecting area in square meters. Lower plot is the increase in number of pixels in the focal plane in megapixels. Credit: IVOA.

now provides access to several hundred catalogs, several data archives, and links to about 7 million records. MAST, the Multi-mission Archive at the Space Telescope Science Institute (STScI) is a NASA-funded project to support and provide a variety of astronomical data archives to the community with the primary focus on scientifically related data sets. This site has links to many other data centers within NASA and worldwide, including the infrared science archive (IRSA) and the high-energy astrophysics science archive center (HEASARC).

As pointed out in the report of the U.S. National Virtual Observatory Science Definition Team chaired by George Djorgovski of Caltech in April 2002 (Djorgovski, 2002) "the amount of data in astronomy is growing exponentially" and yet this is just a foretaste of things to come (Figure 1.14). Indeed, a major problem facing all of astronomy as a result of technological innovations is archiving and data management. Massive and growing data archives have given rise to the concept of the "virtual observatory" in which one goes first to the data base to "mine" the mountains of existing observations rather than to the telescope. Ground-based optical and near-optical observations alone now produce data in the form of images or spectra in such voluminous quantities as to make interactive analysis by human beings difficult. How much data? Typically, the information in an image is digitized to 16 bits or $2^{16} = 65{,}536$ levels of intensity for each pixel. Eight bits (0s or 1s) make a byte, and 2 bytes (16 bits) make a word; 32 bits is a double-word (corresponding to $2^{32} = 4{,}294{,}967{,}296$ pieces of information). A CCD with $1{,}024 \times 1{,}024$ pixels, also written as $1\,\mathrm{K} \times 1\,\mathrm{K}$ ($= 1\,\mathrm{Mpxl}$; here $\mathrm{K} = 2^{10}$), and 16-bit digitization per pixel requires 2 megabytes (MB) to store a single frame, and a typical $8\,\mathrm{K} \times 8\,\mathrm{K}$ camera needs 128 MB/frame. At 100 images per night (a lower limit at most observatories) the storage requirement becomes 12.8 gigabytes ($1{,}000\,\mathrm{MB} = 1\,\mathrm{GB}$) per night. In only 100 nights, that single observatory would have amassed over 1 terabyte (1,000 GB) of digital data from that camera alone.

Formed in June 2002 with over 16 member countries, the International Virtual Observatory Alliance (IVOA) has the goal of facilitating international coordination and collaboration for the development of the tools, systems, and organizational structures needed to enable an integrated, worldwide, virtual observatory with access to enormous data archives. As Robert Hanisch, project manager of the U.S. National Virtual Observatory, and Peter Quinn, director of the ESO Astrophysical Virtual Observatory, state in a 2003 introduction to the International Virtual Observatory Alliance (IVOA), what astronomy is facing is a "data avalanche". Moore's Law is driving astronomy as much as it is the semiconductor industry. Moore's Law is the empirical observation made in 1965 by Gordon E. Moore (co-founder of Intel) that the number of transistors on a chip seemed to be doubling annually. He later modified the estimate to a doubling rate of every two years in 1975. Although in no sense a physical law, it has become a guiding principle. The term "Moore's Law" was coined by Caltech professor and entrepreneur Carver Mead, himself a pioneer of micro-electronics. It is the advent of ever-cheaper computer storage technologies and high-speed networks, and the continued application of Moore's Law, that has led to planning for ambitious new survey telescopes like the LSST that will image the entire sky every few days and yield volumes of data measured in petabytes (PB) (i.e., thousands of terabytes). Changes like this will likely have a dramatic effect on the sociology of astronomy itself. As George Djorgovski (2002) points out, progress on information technology enables the exploration of data and effective knowledge extraction from the data. Modern data sets (and their fusion) have such high information content, that data-mining is not only profitable, but indeed necessary in order to exploit the data properly. This is what the VO framework is all about, not just archiving. It is also worth noting that the same situation now occurs in all sciences, not just astronomy, but that astronomy is one of the forefront scientific disciplines in making these developments.

1.5 TERMINOLOGY OF ELECTRONIC IMAGING

We conclude this introduction with a summary of some technical terms. Let's begin by asking what constitutes a perfect imaging system in astronomy? Simply put, image resolution across the field of view should be limited only by the fundamental diffraction of light, and the limit of detection for faint objects should be determined only by photon (Poisson) counting statistics. The limit of detection is described by the signal-to-noise ratio (SNR or S/N), and therefore the key to an ideal system is to maximize the amount of signal collected from the source and drive down the random errors (the noise) to the fundamental quantum (photon) limits.

1.5.1 Image formation

The image of a distant point source of light is never a point source. In practice, the image is described by a *point spread function* (PSF) that shows how the light from a point source (like a star) is spread over multiple pixels in the image; the PSF can vary

across the field of view and its shape depends on which factors are limiting image quality. In the absence of any other degrading effects, the image of an astronomical point source viewed by a circular objective lens (or primary mirror) of diameter D at a wavelength λ would be an Airy diffraction disk (Sir George Airy, 1801–1892). The intensity distribution for the Airy disk is given by $I(\theta) = I_0\{J_1(x)/x\}^2$ where $J_1(x)$ is a first-order Bessel function, $x = \pi D \sin\theta/\lambda$, and θ is the angular radius from the center of the aperture. This sharply peaked circularly symmetric function, illustrated in Figure 1.15, has the property that it first goes to zero when $x = 3.832$ and hence when $\sin\theta = (3.832/\pi)(\lambda/D)$, or for small angles when $\theta = 1.22\lambda/D$ (in radians). There is then a weak secondary maximum at $1.635\lambda/D$ followed by another dark minimum at $2.233\lambda/D$, and this is repeated with weaker peaks so that the central image appears surrounded by a fading set of rings. This effect is often seen in images taken with the Hubble Space Telescope and represents a fundamental limit. Note that most large telescopes also have a secondary mirror which is usually supported in the middle of the beam by four long struts or vanes. Diffraction of light at the vanes of the secondary mirror will produce the artifact of "spikes" in the recorded image of a bright star and the central obscuration will modify the PSF depending on the area of the secondary (Schroeder, 2000). Likewise, non-circular shapes produce non-circular diffraction patterns. For example, the hexagonal mirrors of the Keck telescopes produce a hexagonal variation of the Airy disk.

A common way to describe the observed PSF in astronomy is by the full width at half maximum intensity (*FWHM*) of its intensity profile. This is exactly what it

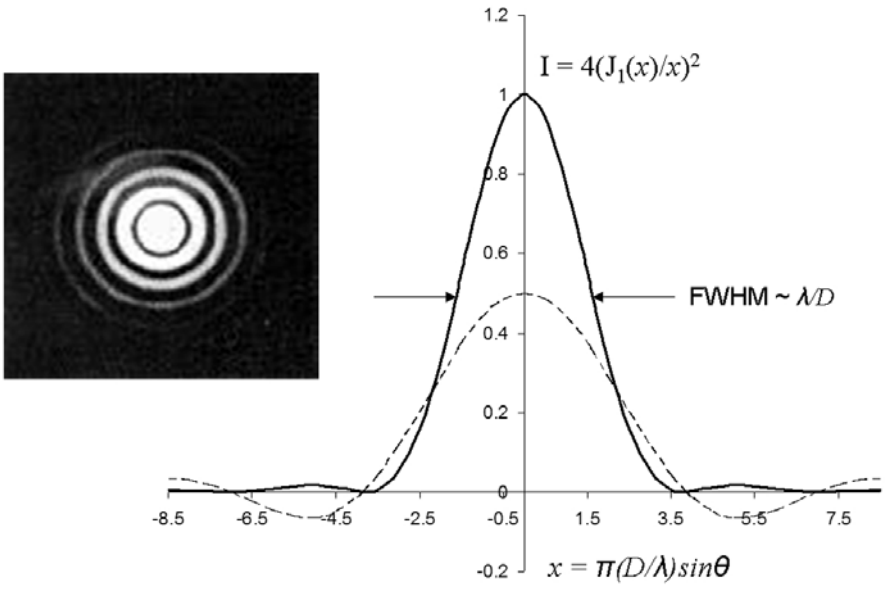


Figure 1.15. The point spread function (solid line) of the Airy diffraction pattern for a circular aperture of diameter D is illustrated both as an image and in cross-section.

sounds like. It is the angular width of the image in seconds of arc measured between the points where the intensity is one-half of the peak value. The FWHM can be related to the shape of the PSF especially when it is describable by a simple mathematical form like a Gaussian (bell-shaped) curve. One drawback of using the FWHM is that it may not describe the energy in the "wings" of the PSF; blurring due to seeing and light-scattering can redistribute energy into the wings of the profile in a way that is not consistent with a Gaussian curve. A better statement of performance is given by the diameter that contains 80% of the energy; this is called the encircled energy. For the diffraction-limited case this corresponds to $\sim 1.8\lambda/D$. For example, at the near-infrared wavelength of 1 μm on a 10 m telescope this image size is 1.8×10^{-7} radians or 0.037 seconds of arc. This is a remarkable concentration of light into a very small image size. In practice, many other factors affect image formation, such as atmospheric turbulence and geometric optics, making it difficult to achieve the diffraction limit.

For a real wave traveling through any optical system we can compare it with the best fitting spherical wave at any point, and the difference between these two wave fronts is called the "optical path difference" or OPD. The OPD can be expressed in units of length (like nanometers) or in "waves" as a fraction of a reference wavelength, such as the 632.8 nm wavelength of a red helium–neon laser. A wavefront error (WFE) can be specified in terms of either the "peak-to-valley" (P-V) optical path difference, which is the difference between the longest and the shortest paths over the beam, or as the "root mean square" (rms) wave front error over the given surface so that $\mathrm{WFE}_{\mathrm{rms}} = \sqrt{\Sigma\langle \mathrm{OPD}^2 \rangle}$. The rms wave front error is approximately one-fifth of the peak-to-valley OPD. The Rayleigh Criterion (Lord Rayleigh, 1842–1919) is an empirical resolution statement that defines two diffraction-limited PSFs to be distinguishable if the maximum of one Airy disk falls on the first minimum of the second; this is a separation of $1.22\lambda/D$ and is generally considered a conservative resolution estimate. It is found that a P-V wave front error of a quarter of a wave (0.25λ) is essentially diffraction-limited. In this case about 84% of the encircled energy lies within a diameter of $2.44\lambda/D$; about 50% of the energy falls within an angular diameter of λ/D which is why this estimate is usually the one given for the diffraction limit. Atmospheric turbulence will blur point-like images from an Airy disk with angular diameter $2.44\lambda/D$ to a much larger seeing disk.

Unfortunately, a spherical mirror or a spherical lens will not produce a point-like image of a point source because the rays from different annuli of the mirror (lens) focus at different distances. This geometrical effect is called "spherical aberration" and it blurs out the image to a size given in angular terms (diameter in seconds of arc) by:

$$\theta_{\mathrm{SA}}(\mathrm{mirror}) = 206{,}265/128(F)^3$$

where $F = f/D$ is the focal ratio, and

$$\theta_{\mathrm{SA}}(\mathrm{lens}) = 206{,}265n(4n-1)/[128(n+2)(n-1)^2F^3]$$

where $n =$ the refractive index of lens.

This aberration affects on-axis images. For an $f/15$ telescope ($F = 15$) the diameter of the blurred image is about 0.5 seconds of arc for the mirror and eight times that for the lens, assuming a normal refractive index of $n = 1.5$ for glass. There are "off-axis" aberrations that also affect image quality: namely, coma, astigmatism, and distortion (each of these image-degrading factors will be explained later). Of course, the telescope and instrument must be in focus, and optical components cannot be out of alignment.

It is important to distinguish the concept of "sampling" from that of angular resolution. *At leas*t two pixels are needed to properly sample the point spread function, otherwise two half overlapping images would blend into one. This requirement is called Nyquist sampling. The Nyquist frequency is the maximum spatial frequency in line pairs per millimeter which lines up the columns (or rows) of a CCD. For example, a CCD with 15 μm pixels implies 0.015 mm per pixel or 0.030 mm per pixel pair. Inverting this number (1/0.030) gives the Nyquist frequency of 33.3 line pairs per millimeter. Imaging performance can also be described using the modulation transfer function (MTF), where by modulation we mean the ratio $V = [I(\max) - I(\min)]/[I(\max) + I(\min)]$. This ratio is also known as the "visibility" and it is large for low spatial frequencies in which there are many pixels across the bright and dark regions of a test pattern. The MTF is just the ratio of the visibility in the image to that in the object.

1.5.2 Detector performance

A perfect detector is one with 100% efficiency in converting photons to electrons and converting electrons to measured counts. If there is no illumination there should be no signal. Repetitive measurements of the same constant light source should give the same average signal with a deviation about the average determined only by Poisson statistics. All pixels should have exactly the same characteristics. Unfortunately, there is no such device! However, modern imaging detectors come very close, as we shall see.

Detector classification schemes are discussed later. For now, we simply identify the most important properties of a photon detector as

— *Quantum efficiency*: the ratio of number of photoelectrons produced to the number of incident photons
— *Noise*: the irreducible electronic fluctuations in the output signal from an imaging device
— *Dynamic range*: the ratio of the maximum signal the device can detect without saturation compared with the noise level
— *Linearity*: the extent to which the output signal is proportional to the photon arrival rate
— *Spectral response*: the wavelength range over which the device is sensitive
— *Temporal response*: the time interval over which the device cannot respond to another incoming photon

— *Dark current*: the signal received from the device when it is not illuminated by light.

Earlier we introduced the terms radiance and irradiance to describe the power emitted by the source and that received by the detector. The spectral radiance per frequency interval (L_ν) is the power (in watts) leaving a unit-projected area of the surface of the source (in square meters) into a unit solid angle (in steradians; a sphere subtends 4π steradians at its center) towards the observer and in a unit frequency interval (in hertz). If dA is an element of the emitting surface, then $dA\cos\theta$ is the element of surface area perpendicular to the observer in direction θ relative to the outward normal to dA. The radiance L is therefore the integral of L_ν over all ν, and the radiant exitance M is the integral of L over all solid angles Ω. A Lambertian surface is one for which the radiance is constant, independent of direction of view. This implies that the emission is proportional to $L\cos\theta$ to compensate for the $dA\cos\theta$ projection element. The flux (or power) emitted by the source (Φ) is the radiant exitance multiplied by the total surface area. An optical system receives a portion of the power emitted by the source. The range of angles is limited by the field of view (FOV) of the imaging system. The solid angle subtended by the entrance aperture area (a) of the optical system at the source is given by $\Omega = a/r^2$. Note that for a distant point source (a star) this angle is exceedingly small, and so the practical value is determined either by the seeing disk or the diffraction width. Irradiance (E) is the power in watts/m^2 *received* at a unit surface element some distance from the source. The spectral irradiance (E_ν or E_λ) is the irradiance per unit frequency or wavelength interval.

While the terminology above is common in industry, somewhat different terms are used in astronomy, and this can be confusing. For the emitted radiation, L_ν and L_λ from above are usually called the Specific Intensity (I_ν and I_λ), and therefore the radiance L becomes simply the intensity (I). The total power (Φ) emitted by an astronomical source is called the luminosity (L, watts); note the unfortunate overlap in symbols. For the received radiation, astronomers call the irradiance (E), the "flux" (S, W/m^2), and the spectral quantities E_ν or E_λ become the flux density (S_ν or S_λ) and are usually given per unit solid angle (W m^{-2} ster^{-1}). The apparent magnitude of a source would then be $m = \text{constant} - 2.5\log_{10}(S)$. The radiometric unit, the lux, is such that 1 lux = photon flux of $\sim 3 \times 10^9$ photons/s/mm^2 for 550 nm photons and this corresponds to a visual magnitude of $m_V = -14$.

Finally, the electrical signal recorded by a photon detector can be predicted if the irradiance is known, together with the collecting area, the exposure time, the wavelength and spectral bandwidth, and the efficiency of the optical plus detector system. Ideally, if N is the signal rate and t the exposure time, the measured signal will be $Nt \pm \sqrt{(Nt)}$ and the signal-to-noise ratio will be $(Nt)/\sqrt{(Nt)}$, which is just $= \sqrt{(Nt)}$. In practice, the signal-to-noise ratio will be worse than this because the signal will need to be extracted from a "background" signal and there will probably be additional fixed noise sources. In later chapters we will develop this case in detail, for now it is sufficient to illustrate the problem by considering the case in which there is an additional photon background (the sky) and a fixed detector noise. Assuming all

of the light falls on a single detector pixel, and that the background contributes $N_B t$ photoelectrons while the detector adds a fixed noise R (in equivalent electrons), then the true signal is $(Nt) = (Nt + N_B t) - (N_B t)$, which requires a measurement with the source on the pixel and another with the source moved off the pixel to obtain the sky alone. The error in the difference measurement is found by *adding* the noise terms in quadrature (i.e., square, add, and take the square root). The noise is therefore $\sqrt{[(Nt + N_B t) + N_B t + R^2]}$ or $\sqrt{[(Nt + 2N_B t + R^2]}$ because, for the "signal + sky" the Poisson noise squared equals the signal itself, and for the sky measurement the same is true, while for the non-Poisson detector noise the value squared must be included. The signal-to-noise ratio in this simple case can now be written as $S/N = (Nt)/\sqrt{[(Nt + 2N_B t + R^2]}$ or, to facilitate comparison with the ideal case, we can write this as

$$S/N = \sqrt{(Nt)}/\sqrt{[(1 + 2N_B/N + R^2/Nt]} \tag{1.4}$$

If the background (N_B) and the detector noise (R) are negligible relative to N, then this expression for signal/noise becomes $\sqrt{(Nt)}$ as before. However, for a faint source (small N) these new terms could easily dominate, and therefore the signal-to-noise ratio will be much worse than before because the denominator is larger than one. This simple expression contains several insights. Longer exposures mitigate the fixed detector noise (R^2/Nt); more measurements of the sky would reduce the Poisson error by $\sqrt{n}$ where n is the number of times the sky is measured; when the background dominates then S/N scales as $\sqrt{t}$ and inversely as $\sqrt{N_B}$, while for the detector noise dominated case the S/N scales as t. When the $S/N = 1$ then the signal power (in watts) is often called the noise equivalent power (NEP). If the noise is distributed randomly then an S/N ratio of 5, also called a 5σ (five sigma) detection is a solid result; the noise is one-fifth (or 20%) of the signal. For accurate work the goal would be more like 1% or an S/N ratio of 100. Again, all of the topics introduced in this section will be developed further in later chapters.

1.6 SUMMARY

We have shown that progress in understanding the universe is intimately linked with developments in technology. Electronic sensors, larger telescopes, and advances in computers have enabled astronomers to observe and measure natural phenomena in a quantitative way. New technologies lead to new discoveries, and the quest for more knowledge drives the development and search for better techniques. Electronic imaging devices which convert photons to electrical charge are the primary tools of modern astronomers. Principal among these array detectors is the charge-coupled device or CCD which has now reached the billion pixel level. Breakthroughs in technology have enabled astronomical imaging across the entire electromagnetic spectrum, and the data collection rate is growing rapidly. Image quality is described in terms of a point spread function that describes how a point-like image is recorded. Detector performance is quantified by the signal-to-noise ratio, where the noise or

random error in the measurement depends both on the signals being measured and the detector itself.

1.7 EXERCISES

1 Why is astronomy called an "observational" science? List all the possible "observables" that can be used to study the cosmos.

2 (a) What is meant by the "characteristic curve" of a photographic plate? (b) Draw and label all parts of a characteristic curve with gamma = 2. (c) What does this curve imply for precise quantitative measurements? (d) How can photographic emulsions be made more sensitive?

3 (a) Calculate the frequency of a radio wave which has a wavelength of 1 mm. (b) Calculate the wavelength of a radio wave with a frequency of 1,427 MHz.

4 List all the properties of electromagnetic radiation which can be measured?

5 What is meant by a "quantum" of energy? Calculate the quantum energy associated with a photon of light of wavelength 500 nm. Given that 1 electron volt (eV) is 1.602×10^{-19} joules, convert the energy of a 500 nm photon to electron volts.

6 What is meant by an "image photon-counting system"? Why are they advantageous for very low light levels? What is their primary disadvantage?

7 In a given photomultiplier tube, each dynode surface generates three electrons for the impact of one. Show that the "gain" of a photomultiplier tube with 10 dynodes is 3^{10}. Assuming the charge on the electron $= 1.602 \times 10^{-19}$ coulombs, how much charge is this?

8 A photon-counting experiment collects 1 million counts in a given time. (a) What is the error in this measurement assuming Poisson statistics? (b) What is the fractional error expressed as a percentage? (c) If the measurement time is increased by a factor of 4, by how much does the percentage error improve (assuming a constant flux)?

9 Calculate the focal ratio (F) for a spherical mirror that gives about 1.0 seconds of arc blurring due to spherical aberration. Assuming this focal ratio and a refractive index of $n = 1.5$, what is the blurring (in seconds of arc) due to spherical aberration from a single objective lens? Does it make sense that the earliest telescopes all had long focal lengths?

10 Develop the simplified signal-to-noise ratio expressions for the two cases of background-limited and detector-noise-limited performance mentioned in the text and show explicitly the dependence on integration time. If the source and background each contribute 10 electrons/s and the readout noise is 10 electrons

rms, then which noise source dominates a 1 s exposure? Which noise source dominates an exposure of 100 s?

1.8 REFERENCES

Amelio, G.F.; Tompsett, M.F.; and Smith, G.E. (1970) Experimental verification of the charge coupled device concept, *The Bell System Technical Journal*, **49**(4), 593–600.

Boyle, W.S.; and Smith, G.E. (1970) Charge coupled semiconductor devices, *The Bell System Technical Journal*, **49**(4), 587–593.

Carroll, J.; Gray, D.C.; Roorda, A.; and Williams, D.R. (2005) Recent advances in retinal imaging with adaptive optics, *Optics and Photonics News*, January 2005, 36–42.

Clark, R.N. (1990) *Visual Astronomy of the Deep Sky*, Cambridge University Press and Sky Publishing, Cambridge, MA. Available at *http://www.clarkvision.com/rnc/index.html*

Csorba, I.P. (1985) *Image Tubes*, Howard Sams, Indianapolis, IN.

Djorgovski, S.G. (2002) Towards the National Virtual Observatory: A report prepared by the National Virtual Observatory Science Definition Team. Available at *http://www.us-vo.org/objectives.cfm*

Eastman Kodak (1987) *Scientific Imaging with Kodak Films and Plates*, Eastman Kodak, Rochester, NY.

Geary, J.; and Latham, D. (eds.) (1981) *Solid State Imagers for Astronomy*, SPIE, Bellingham, WA, Vol. 290.

Gunn, J.E.; Carr, M.; Rockosi, C.; Sekiguchi, M.; Berry, K.; Elms, B., de Haas, E., Ivezić, Ž.; Knapp, G., Lupton, R. *et al.* (1998) The Sloan Digital Sky Survey Photometric Camera, *Astron. J.*, **116**, 3040–3081.

Harwit, M. (2003) The growth of astrophysical understanding, *Physics Today*, **56**, 38.

Janesick, J.R. (2001) *Scientific Charge-Coupled Devices*, SPIE, Bellingham, WA.

McLean, I.S.; Coyne, G.V.; Frecker, J.E.; and Serkowski, K. (1979) High-resolution polarization structure of H-beta in Be shell stars measured with a new Digicon spectropolarimeter, *Astrophys. J.*, **228**, 802–808.

McLean, I.S. (1988) Infrared astronomy's new image, *Sky & Telescope*, **75**, 254.

Racine, R. (2004) Historical growth of telescope aperture, *Pub. Astron. Soc. Pacific*, **116**, 77–83.

Sandage, A. (1999) The first 50 years at Palomar: 1949–1999, *Annual Reviews of Astronomy & Astrophysics*, **37**, 445–486.

Schroeder, D.J. (2000) *Astronomical Optics*, second edition, Academic Press, San Diego, CA.

Tyson, J.A. (1986) Low-light level charge-coupled device imaging in astronomy, *J. Opt. Soc. Am.*, **3**(12), 2131–2138.

Web sites

European Virtual Observatory: *http://www.euro-vo.org/pub/*

International Virtual Observatory Alliance: *http://www.ivoa.net/pub/info/* The IVOA.pdf, an article by R.J. Hanisch and P.J. Quinn.

Multi-Wavelength Milky Way: a compilation of many surveys from radio to gamma rays (Credit: NASA's Goddard Space Flight Center): *http://mwmw.gsfc.nasa.gov/*

National Academy of Engineering's Draper Prize to Boyle and Smith for the CCD: *http://www.nae.edu/NAE/awardscom.nsf/weblinks/CGOZ-6K9L6P?OpenDocument*
U.S. National Virtual Observatory: *http://www.us-vo.org/*

1.9 SUGGESTIONS FOR ADDITIONAL READING

The following books provide additional material, either in terms of historical development, general astronomy, or more quantitative treatments.

Bode, M.F. (ed.) (1995) *Robotic Observatories*, Wiley/Praxis, Chichester, U.K.
Birney, D.S.; Gonzalez, G.; and Oesper, D. (2006) *Observational Astronomy*, second edition, Cambridge University Press, Cambridge, U.K.
Gingerich, O. (ed.) (1975) The nature of scientific discovery. *Proceedings of a Symposium in Honor of Copernicus*, Smithsonian Institution Press, Washington, D.C.
Hearnshaw, J.B. (1990) *The Analysis of Starlight: One Hundred and Fifty Years of Astronomical Spectroscopy*, Cambridge University Press, Cambridge, U.K.
Howell, S.B. (2006) *Handbook of CCD Astronomy*, second edition, Cambridge University Press, Cambridge, U.K.
Krisciunas, K. (1988) *Astronomical Centers of the World*, Cambridge University Press, Cambridge, U.K.
Learner, R. (1981) *Astronomy through the Telescope*, Van Nostrand Reinhold Company, New York.
Rieke, G.H. (2003) *Detection of Light from the Ultraviolet to the Submillimeter*, second edition, Cambridge University Press, Cambridge, U.K.
Verschuur, G.L. (1987) *The Invisible Universe Revealed: The Story of Radio Astronomy*, Springer-Verlag, New York.
Wall, J.V.; and Boksenberg, A. (eds.) (1990) *Modern Technology and Its Influence on Astronomy*, Cambridge University Press, Cambridge, U.K.
http://www.karger.com/gazette/64/fernald/art_1_0.htm This is an article on the evolution of eyes by Professor Russell D. Fernald, Stanford University.

2

Beating the atmosphere

Images of point-like astronomical sources formed on a CCD camera will be represented by a point spread function. In the absence of other degrading effects, the spreading of the image is determined by the diffraction of light. In practice, for ground-based observatories, the light must pass through the atmosphere, which has a major impact on image quality because of turbulence.

2.1 ATMOSPHERIC ABSORPTION AND TRANSMISSION

While the Earth's atmosphere provides the biosphere that we live in and protects life on the planet from harmful radiation from space, it is not so friendly to the pursuit of astronomy. The atmosphere absorbs and scatters incident electromagnetic radiation. It is the scattering of sunlight by air molecules that makes the sky seem blue; Lord Rayleigh (1842–1919) showed that the scattering is inversely proportional to wavelength to the fourth power (λ^{-4}) and so blue photons are scattered much more strongly and reach our eyes from all directions. Scattered sunlight is also highly polarized (90° from the Sun), a fact that is easily demonstrated with Polaroid sunglasses by tilting your head from side to side while looking at the blue sky to see that the intensity changes with the angle of your sunglasses. Under certain conditions the atmosphere also emits radiation. Of more concern is the fact that the atmosphere disturbs the incoming waves through turbulent air motion which in turn limits the ability of a telescope to achieve its ultimate angular resolution.

Figure 2.1 shows a simplified plot of just how opaque the Earth's atmosphere is to electromagnetic energy at each wavelength from gamma rays to radio waves; 100% opacity means that the transmission is zero. There are only two regions of the spectrum that are easily transmitted to the ground by the Earth's current oxygen-rich atmosphere. These regions are the Visible plus Near-Infrared, and the Radio, everything else is opaque. Even within the range of visible light the atmosphere still

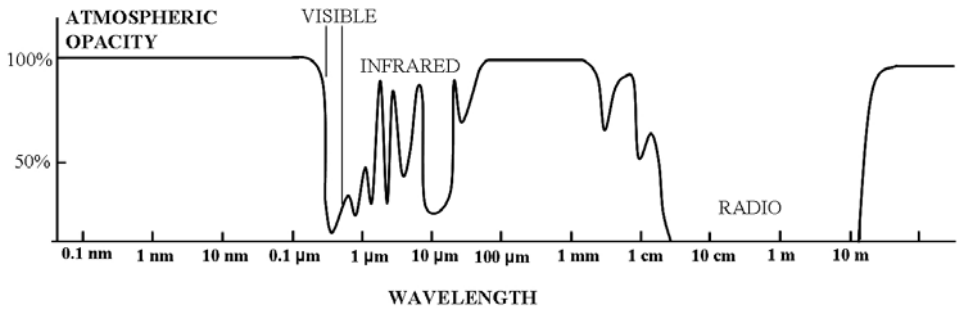


Figure 2.1. The transmission of the atmosphere at each wavelength from gamma rays to radio waves. Except for visible light, some near-infrared light, and radio waves, all other forms of electromagnetic radiation are blocked by the atmosphere.

absorbs in some very narrow bands of wavelengths, thus producing absorption effects of terrestrial origin when recording the spectra of astronomical sources. For example, oxygen absorbs strongly in two broad spectral bands called the A band (near 760 nm) and the B band (near 688 nm), well within the CCD spectral range. Each of these bands includes about 40 spectral line transitions with such strong absorption that the atmosphere is essentially opaque at those wavelengths. Water vapor absorption occurs weakly at 514 nm, 606 nm, 660 nm, 739 nm, and 836 nm and then more strongly in several bands from 970 nm to 1,940 nm. Ultraviolet waves, X-rays, and gamma rays are effectively blocked by the atmosphere, mainly by water vapor for X-rays, but also ozone (O_3), oxygen (O_2), and carbon dioxide (CO_2) for UV. Ultraviolet radiation from 400 nm to 320 nm (also called UV-A) can reach the surface and excellent UV observations can be made from sufficiently high mountaintop observatories. Between 320 nm and 290 nm is the range called UV-B in the terminology of biological damage from ozone depletion. Attenuation of these UV photons by the ozone layer is something like 350 billion to one compared with the top of the atmosphere. Below 290 nm (UV-C) electromagnetic radiation is completely blocked by ozone at about 35 km. Telescopes operating at very short wavelengths must be in space or carried very high into the atmosphere (above 50 km) by balloons or rockets.

Moving to the infrared, we find that the atmosphere is opaque at some wavelengths and transparent at others. The illustration in Figure 2.1 is only a guide. In detail there is much more absorption structure throughout the infrared, and we will return to this topic in Section 11.2.1 on infrared imaging. Again, the main culprits are the greenhouse gases, carbon dioxide (CO_2) and water vapor (H_2O), which create a series of "windows" for wavelengths out to 20 μm, about 40 times the wavelength of normal visible light. Beyond 20 μm, observations must be done from space (or from the stratosphere) until the wavelength reaches about 2 cm where the atmosphere again becomes transparent and radio astronomy is possible. From very high dry sites, however, there is the possibility of sub-millimeter wave observations under good conditions at 450 μm and 850 μm. At the longest wavelengths, radio waves longer than about 20 m are blocked by the ionosphere.

While water vapor and carbon dioxide do an efficient job of blocking out a lot of infrared and sub-millimeter radiation, the water vapor is sensitive to height in the atmosphere, and consequently high-altitude sites such as Mauna Kea, Hawaii at 4.2 km (13,796 ft) and the high deserts of the Chilean Andes are excellent places for ground-based astronomy. The realization that high-altitude sites might be better suited for astronomy is often credited to Sir Isaac Newton because of this famous statement in his 1730 treatise on *Opticks*:

> "For the Air through which we look upon the Stars, is in perpetual Tremor ... But, these Stars do not twinkle when viewed through ... large apertures. The only Remedy is a most serene and quiet Air, such as may perhaps be found on the tops of the highest Mountains above the grosser Clouds."

Charles Piazzi Smyth (1819–1900) was the Scottish Astronomer Royal (from 1846 to 1888) who experimented with "mountaintop" observing in Tenerife, Canary Islands in 1856 with a grant specifically to test Newton's idea. American astronomer Henry Draper (1837–1882) also suggested building observatories in mountainous areas and is known to have mentioned the Andes.

The relative concentrations of the permanent gases in a dry atmosphere are nearly constant: nitrogen (N_2) at 78.1% by volume, oxygen (O_2) at 20.9%, argon (A) at 0.9%, and carbon dioxide (CO_2) at 0.03%. All other permanent constituents are less than 0.002%. The two major variable constituents are ozone and water vapor. The maximum concentration of ozone occurs at high altitudes (10 km–30 km) and mainly affects UV transmission. Water vapor is a low-altitude phenomenon and varies strongly with temperature (and hence season) and altitude. Commonly, the amount of water vapor contained in the optical path is called the "precipitable water" and is measured in millimeters. Precipitable water is defined as the depth of the layer of water that would be formed if all the water vapor (H_2O molecules) along the line of sight was condensed in a container having the same cross-sectional area as the optical beam. It is not necessary to know the cross-sectional area of the beam because, if the area was larger and more water was condensed, it would be spread out over that larger area and its depth would be the same. The amount of precipitable water is usually expressed in millimeters per kilometer of path length or as the total millimeters in the airmass above the observatory. Values vary from 1 mm–15 mm at low astronomical sites (below 2 km) but 4 mm is typical for high dry locations like Mauna Kea.

Atmospheric pressure (P) below 120 km altitude is approximately given by an exponential decline $P(h) = P_0 e^{-h/H}$ where h is the altitude and H is called the scale height, and is the value of h where the pressure falls to $1/\mathrm{e} = 0.37$ (37%) of its value at sea level (P_0). The standard atmosphere (atm) has a pressure of 101.325 kilopascals (kPa) or 14.696 pounds per square inch (psi). Depending on temperature, the typical value for H is $\sim$8 km for the permanent constituents of the atmosphere. Water vapor content falls off much more rapidly with height because it is concentrated close to sea level. A typical value for a site like Mauna Kea is $H_{\mathrm{wv}} \sim 1.85$ km whereas the mountaintop is at 4.2 km. The site of the Atacama Large Millimeter Array (ALMA)

telescopes (Chajnantor, Chile) is at a height of 5.06 km with an expected 1 mm–2 mm of precipitable water vapor; for reference, the height of Mt. Everest is 8.848 km (29,029 ft).

The thickness of the atmosphere through which radiation has to pass is measured in terms of airmass, where one airmass is the optical thickness when looking straight up. By treating the atmosphere as plane-parallel slabs, airmass (X) is given fairly well by the secant ($=1/\text{cosine}$) of the zenith angle (z), also called the zenith distance, the angle between the zenith point overhead and the star; $X = \sec z$. For a star of known right ascension and declination (α, δ) on the sky, its zenith angle can be computed from the relation $\cos z = \sin\phi \sin\delta + \cos\phi \cos\delta \cos(\text{LST} - \alpha)$ where ϕ is the latitude of the site and LST is the local sidereal time (more information on coordinate systems and spherical trigonometry is given in Appendix E). The atmosphere absorbs preferentially in the blue and therefore both dims and reddens starlight. In fact, the attenuation or extinction as a function of wavelength can be written (approximately) as $E_\lambda = C_\lambda \sec z$ where C_λ is a constant over a limited band of wavelengths. Thus, if m_λ is the true magnitude of a star measured outside the Earth's atmosphere at this wavelength, then the observed magnitude at a given zenith angle becomes $m_\lambda(z) = m_\lambda + C_\lambda \sec z$ (Bouguer's Law). At the zenith, $z = 0^\circ$ and $\sec z = 1$, while at $z = 60^\circ$ (or an elevation angle of 30° above the horizon) $\sec z = 2$. Bouguer's Law is accurate for values of z up to about 60° after which the plane-parallel atmosphere assumption on which it is based breaks down. A more precise formula is then: $X = \sec z - 0.0018167(\sec z - 1) - 0.002875(\sec z - 1)^2 - 0.0008083(\sec z - 1)^3$. Values of the extinction coefficient C_λ have a large range from about 3.7 magnitudes per unit airmass in the UV (300 nm) to about 0.005 magnitudes per airmass in the near-infrared at the silicon CCD limit (1,100 nm). This "constant" is also variable from site to site and even during the night at a given site due to subtle changes in the atmosphere.

Stated another way, the measured photon arrival rate (S) is reduced by a transmission factor $T_{\text{atmos}}(\lambda)$ which depends on wavelength and is a number between 0 and 1. The transmittance is typically given by the exponential factor $T = e^{-\mu_a(\lambda)L}$ where $\mu_a(\lambda)$ is the absorption coefficient in units of cm^{-1} at a given λ, and L is the path length (in centimeters) for absorption. There will be additional transmission losses through the telescope and instrument, and then a final loss if the quantum efficiency of the detector (η) is less than one. Thus, the transmitted signal at a given wavelength will be $S = \tau\eta S_0$, where τ is the product of all transmission factors (atmosphere, telescope, instrument) and S_0 is the incident photon arrival rate.

Finally, the direction of light transmitted through the atmosphere changes because of refraction, and this causes a wavelength-dependent displacement of the star image from its true position. This effect makes a star move apparently towards the zenith from its true position by an amount $\Delta z = (n-1)\tan z$ where z is the observed zenith angle. For example, taking $n = 1.00029$ ($\sim$700 nm) then $\Delta z \sim 59.8'' \tan z$. At shorter wavelengths n is slightly larger. For example, taking $n = 1.000295$ ($\sim$480 nm) yields $\Delta z \sim 60.8'' \tan z$ which is a noticeable difference (1 arcsecond) and is known as atmospheric dispersion. This difference would be enough to cause considerable light loss at the slit of a spectrometer if the slit was less

than 1 arcsecond wide. One solution is to orient the slit at the parallactic angle (q), the angle between celestial north and zenith given by $\sin q = \cos \phi \, (\sin H / \sin \alpha)$, where ϕ is the latitude, H is the hour angle, and α is the right ascension. It is also possible to eliminate this form of image degradation using a device called an atmospheric dispersion compensator (ADC), as discussed in Section 3.2.5.

2.2 ATMOSPHERIC EMISSION, THERMAL AND NON-THERMAL

The atmosphere can also glow. One familiar example of this is the polar aurora. This non-thermal solar-induced emission comes mainly from air molecules impacted by ionized particles in the solar wind that are trapped by the Earth's magnetic field and forced to spiral into the atmosphere at the magnetic poles. Fluorescent emission from atomic oxygen at 557.7 nm results in a greenish glow, and there is a weaker effect from the red line at 630.0 nm. Blue and purple colors are emitted by atomic and molecular nitrogen. The effect on astronomical observing at mid-latitude sites is negligible.

Another solar-induced non-thermal process is generally called airglow but it has three components: OH molecules, O_2 molecules, and the near-infrared nightglow continuum. Of these, the strongest emission comes from the OH molecule which produces a dense "forest" of emission lines, especially in the 0.8 μm–2.5 μm region of the near-infrared. First identified astronomically in the optical red by Aden Meinel (1950), these emission bands are formed as OH molecules relax after absorption of ultraviolet photons from the Sun. Excitation is through the reaction $H + O_3 \rightarrow OH^* + O_2$, where the * indicates that the molecule is in an excited quantum state. Observed emission lines correspond to quantized vibrational and rotational energy transitions of the excited OH* molecule; vibration and rotation give successive levels of fine structure on top of electron transitions by emitting low-energy photons. OH emission comes from a relatively thin layer at an altitude of about 90 km, and the strength of the emission can vary by a factor of 2 or more in time periods of 30 minutes or less. Diurnal variations of the OH* emission are complex. Mean daytime and nighttime levels are about the same. There is usually a sharp dip in emission immediately after sunset, rising back before falling slowly to a pre-midnight dip recovering again around 2 : 00 AM and then falling slowly towards dawn. Temporal variations seem to be related to the motion of gravity-induced acoustic waves in the upper atmosphere. Observations suggest that the average integrated line flux in the atmospheric transmission band from 1.50 μm to 1.80 μm is about 3×10^4 photons s^{-1} m^{-2} $arcsec^{-2}$ μm^{-1} (Maihara *et al.*, 1993). A typical OH spectrum is shown in Figure 2.2 (Oliva and Origlia, 1992) for a small part of the near-infrared around 1.24 μm where atmospheric transmission is otherwise quite good; note that there is also emission from O_2 molecules. For a camera taking images in broad spectral bands from 0.8 μm to 2.5 μm, or for low-resolution spectroscopy, the OH lines are blended together to form a large and variable background that must be subtracted in order to detect faint astronomical sources. This subtraction process will add noise to the measurement, and thus the signal-to-noise ratio will be reduced. In a high-resolution spectrometer, however, OH lines can be separated and even used as

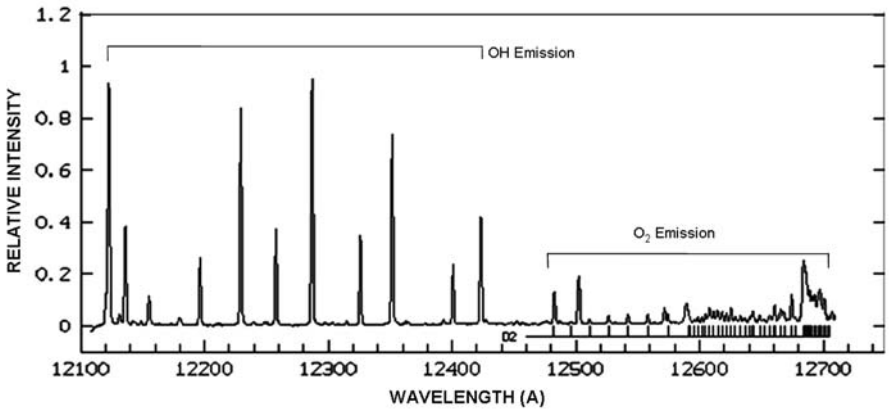


Figure 2.2. The presence of OH night-sky emission lines in part of the near-infrared window at 1.24 microns. Adapted from Oliva and Origlia (1992).

an accurate wavelength calibration source "built in" to the spectrum. Between the OH lines it is much darker, and sensitivity can be very high given a good enough detector. Technology is now available with special spectrometers to create a data cube made up of images at thousands of successive wavelengths which can then be "collapsed" into a broad band image after digitally removing those images that were affected by the presence of an OH emission line. When describing atmospheric radiation phenomena such as airglow, intensities are often given in Rayleighs (R), which is a unit describing photon emission from a volume of atmosphere as follows: if the surface brightness B is measured in units of 10^6 quanta per square centimeter per second per steradian, then in Rayleighs the surface brightness is $4\pi B$. To convert to photons s^{-1} m^{-2} $arcsec^{-2}$ μm^{-1} you can use $I = 0.0187R$.

Because of its finite temperature, the atmosphere also emits thermal (blackbody) radiation but with an emissivity factor (ε_λ) that depends on how opaque the atmosphere is at a given wavelength. Wavelengths where the atmosphere is only semi-transparent are therefore also wavelengths of strongest thermal emission. Water vapor is again the main culprit. Thermal emission is continuous with wavelength and is described by the modified Planck function $\varepsilon_\lambda B_\lambda(T)$. Expressing wavelength in microns (μm) and temperature T in kelvins, the spectral radiance $B_\lambda(T) = 2hc^2\lambda^{-5}/(e^{hc/\lambda kT} - 1)$ W m^{-2} μm^{-1} sr^{-1} which gives the following photon flux per square arcsecond at $T = 300$ K:

$$N_\lambda(T = 300) = 1.41 \times 10^{16}\lambda^{-4}/(e^{48/\lambda} - 1) \text{ photons } s^{-1}\ m^{-2}\ \mu m^{-1}\ arcsec^{-2} \quad (2.1)$$

Thus, for a temperature of 300 K (27°C, 80.6°F) thermal emission rises steeply from short wavelengths to a peak spectral radiance (B_λ) at 9.66 μm (given by $\lambda_{max} = 2{,}898/T$) of about 10 W m^{-2} m^{-1} sr^{-1}. Converting to a photon flux per square arcsecond at 300 K (Equation (2.1)) gives a peak rate of just over 10 billion photons per second per square meter per micron, and the maximum occurs near 12.2 μm. See Appendix C for other forms of the Planck function. On a cold high mountaintop

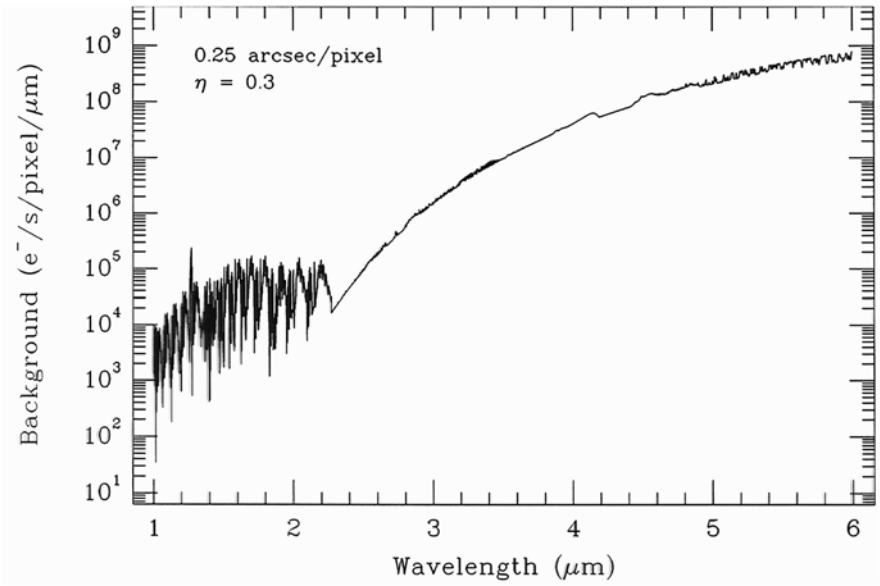


Figure 2.3. A plot of the combined OH night-sky emission and thermal emission from the telescope in the near-infrared from 1 μm to 6 μm showing a dramatic increase in the brightness of the night-time background at these wavelengths. Telescope: 76 m^2; 275 K.

where the nighttime temperature is 273 K (0°C, 32°F) the peak thermal emission (B_λ) shifts to 10.6 μm. Typically, the telescope's primary mirror will be in equilibrium with the temperature of the ambient air. Even in this case, thermal emission beyond 2.2 μm exceeds the average OH background emission and thus becomes the dominant emission source well beyond the peak value near 11 μm, at which point the background is some 10,000 times (ten magnitudes) brighter than at 2.2 μm. Figure 2.3 shows a typical plot of OH emission and thermal radiation from 1 μm to 6 μm; assumptions about the telescope area, temperature, pixel size on the sky and instrument transmission are shown. The intensity scale in Figure 2.3 is logarithmic, thus the night sky in the infrared is very bright compared with visible wavelengths. As the bulk of the atmosphere is generally ~20 K colder than the telescope mirrors, then for wavelengths less than ~13 μm it is the thermal emission from the telescope itself and any other warm optics in front of the detector that dominates the background. All but high-resolution spectroscopy in the infrared will be limited in signal-to-noise ratio by the enormous background emission from the telescope and atmosphere. Mid-infrared (beyond 8 μm) and far-infrared observations (beyond ~20 μm to 30 μm) are best done from space or the stratosphere.

Other sources of background light include the zodiacal light caused by sunlight scattered by tiny particles within the solar system; in the visible part of the spectrum zodiacal light is equivalent to about 22–23.5 magnitudes per square arcsecond and is concentrated in the ecliptic plane. Moonlight of course can be very bright and variable. For example, when the Moon is new the integrated sky background in

magnitudes per square arcsecond in the ultraviolet (U), blue (B), visual (V), red (R), and near-infrared (I) bands is 22.0, 22.7, 21.8, 20.9, and 19.9, approximately, whereas at full Moon these values become 17.0, 19.5, 20.0, 19.9, and 19.2.

2.3 TURBULENCE

Perhaps the biggest impediment for astronomy created by the atmosphere is turbulence. In the complete absence of turbulence, the image of a "point source" should be determined only by the quality of the telescope's optics and by the diffraction of light; that is, the apparent bending, spreading, or interference of the light wave due to disruption by an object in its path, even the mirrors and lenses of a telescope. Diffraction can be thought of as a more complicated example of interference, which in turn is a consequence of the wave nature of light. It is well-known that a plane wave passing through an aperture creates a disturbance that spreads out beyond the aperture; a familiar example would be water waves entering a walled harbor through a single opening and spreading into the protected areas. Long before it was understood that light was an electromagnetic wave, Christiaan Huygens (1629–1695) explained diffraction by the suggestion that new spherical waves (Huygens' wavelets) spread out from all points on the wavefront and this new "wavefront" is the shape that envelopes all of these wavelets. Interference between wavelets is now possible. Where the plane wave interacts with the edge of an aperture the symmetry is broken and the emergent wavefront can no longer be a plane wave. In a telescope, the aperture (lens or mirror) causes the waves to come to a focus, in effect converting the plane wave to a spherical wave, but interference effects will cause the final pattern to be blurred and complex. The general case leads to the Rayleigh–Sommerfeld diffraction integrals, but there are two well-known simpler cases. The first is known as Fraunhofer diffraction in which the light source and the location of the interference pattern are effectively at infinite distances from the aperture causing the diffraction. This condition, also called the "far field" limit, is achieved using lenses or mirrors. When either the source or the location of the diffraction pattern is at a finite distance from the aperture, the theory is called Fresnel diffraction (after French physicist Augustin-Jean Fresnel, 1788–1827, for his major contributions to wave optics). The diffraction pattern formed by plane waves from a point source (a very distant star) passing through the circular aperture of a telescope was solved by astronomer Sir George Airy (1801–1892) in 1883 who obtained the solution in terms of Bessel functions. Details of the derivation can be found in most classic texts on optics. As shown in Chapter 1 (Figure 1.15) a plot of the intensity across any diameter of the Airy disk reveals a bright central maximum surrounded by fainter rings separated by dark bands. The first dark ring occurs at an angular radius of $1.22\lambda/D$ radians from the center and the full width at half maximum (FWHM) is given approximately by λ/D radians (or $206{,}265\lambda/D$ seconds of arc) where λ is the wavelength of the light and D is the diameter of the telescope.

If you were wondering how interference and diffraction are related to photons and quantum theory, then perhaps a simple application of the Heisenberg Uncer-

tainty Principle will help (Jenkins and White, 1957). Consider a plane wave with wavelength λ incident on an aperture of width D. In the quantum picture the photon has a precise momentum given by $p = h/\lambda$; that is, the uncertainty in p is $\Delta p = 0$. Heisenberg's Uncertainty Principle requires that $\Delta x\, \Delta p \sim h$, and so the uncertainty in the position of the photon must be infinite $\Delta x = h/\Delta p$ where $\Delta p = 0$. This is completely consistent with a plane wave that extends indefinitely along its wavefront. At the aperture, however, the location must be limited to $\Delta x = D$ in order for the light to pass. Thus the photon's momentum is now uncertain by an amount $\Delta p = h/\Delta x = h/D$. Momentum is a vector quantity, and so an uncertainty in its value means that the direction of the emergent wave can vary by a small angular amount that we can estimate as $\theta = \Delta p/p$. But $\Delta p = h/D$ and $p = h/\lambda$, and so $\theta = \lambda/D$, roughly in accord with our expectations for the diffraction limit.

Let's take an example: for the infrared wavelength of 1 μm (10^{-6} m) on a 10-meter telescope, the value of λ/D is 10^{-7} radians or about 0.02 seconds of arc. This tiny angular resolution corresponds to the size of a small coin 1 cm (~0.4 inches) in diameter at a distance of 100 km (62 miles). In astronomical terms this is the same as the orbital radius of the Earth (about 150 million km) seen at a distance of about 160 lightyears. Alternatively, this angular resolution is only ten times the separation of the Earth and Moon when viewed from the distance of the nearest star system Alpha Centauri (4.2 lightyears away). These numbers are very intriguing if only they could be realized in practice! Unfortunately, as everyone who has looked through a telescope knows, star images are always much more blurred than this. Time-dependent turbulence blurs the tiny diffraction-limited image by rapid, random shifts of position resulting in a fuzzy "seeing" disk of light that can be 10 to 100 times larger in diameter depending on the site of the telescope and atmospheric conditions. Incoming waves are distorted by randomly moving cells of air with different densities, which in turn arise from temperature variations. American astronomer Horace W. Babcock (1912–2003) performed pioneering studies of this phenomenon from about 1936 onwards, making many hundreds of visual observations at sites in California (including Mt. Wilson) and Chile. Turbulence is characterized by the size of the typical atmospheric cell. It turns out that these cells, even at a very good site, are usually much less than 1 meter across (20 cm is typical), much less than the diameter of a modern large telescope, and it is thus *this* length that determines the size of the fuzzy image or seeing disk. Astronomers compare seeing-limited and diffraction-limited images using the *Strehl ratio* which is defined as the intensity at the peak of the actual seeing disk divided by the intensity at the peak of the true Airy diffraction pattern. The term comes from a closely related image sharpness criterion defined by Karl Strehl (*c*. 1895). Typically, the Strehl ratio is ~0.01. If this ratio could be increased to nearer unity, then most of the light would be in the central spike of the Airy diffraction pattern and the contrast against the sky background would be increased enormously. Smaller image sizes also mean that narrower slits can be used in spectrographs, which in turn implies that the whole spectrometer can be made more compact. To achieve such small images is the ultimate goal of *adaptive optics*, a ground-based method of achieving space-based image quality. Developments in adaptive optics (AO) can be traced to both the astronomical and the military

communities. Sustained efforts from the late 1960s and throughout the 1970s and 1980s by astronomy and non-astronomy groups has led to advanced AO systems, such as the pioneering system developed by Bob Fugate at the U.S. Air Force Starfire Optical Range in Albuquerque, New Mexico and the numerous astronomical AO systems on all sizes of telescopes from 3 m to 10 m in diameter located in all parts of the world. References to this already large and rapidly growing field are given at the end of this chapter, and the account given here draws on the work of Babcock, Beckers, Fugate, Hardy, Max, Roddier, Thompson, Tyson, and many others.

2.3.1 Kolmogorov theory and origin of seeing

What is the effect of the atmosphere on image quality? The Earth intercepts only a tiny fraction of the spherical wave emitted from a distant, point-like source such as a star. When these waves arrive at the Earth the wavefronts are essentially flat and parallel to each other. These plane wavefronts are distorted randomly by moving cells of air, each with a slightly different index of refraction. Variations in the refractive index are caused by variations in density, which in turn arise (mainly) from temperature variations in a fully developed turbulent atmosphere. Air has a refractive index very close to 1.0. For example, at a wavelength of 550 nm (yellow light) the refractive index of air is 1.000272663 at 20°C (68°F), 101.325 kPa (also called 1 atmosphere) and a relative humidity (RH) of 50%. At a high, dry observing site the numbers might be 2°C, 60 kPa, RH of 5%, and then the refractive index is 1.000172262. The National Institute of Standards and Technology (NIST) has a useful online calculator for the refractive index of air. The refractivity of air is given approximately by $n - 1 = 77.6 \times 10^{-6}(1 + 7.52 \times 10^{-3}\lambda^{-2})P/T$, with atmospheric pressure P in millibars and temperature T in kelvins; there is a weak but important dependence on wavelength (λ) which causes the atmospheric dispersion mentioned earlier. The pressure term contains a correction for water vapor pressure (P_{wv}) so that $P = (P_0 + 4{,}810P_{wv}/T)$ where P_0 is the normal atmospheric pressure, but this is a small effect for high-altitude sites. Variations in temperature are more important.

In a turbulent atmosphere each patch of air with a slightly different index of refraction acts like a lens, bending the incoming rays this way or that by small amounts. It is as if the atmosphere is full of drifting and merging lenses. When the light finally reaches the telescope, the rays are no longer parallel and the wavefront is no longer flat (plane). Consequently, the conditions needed to produce an image limited only by diffraction at the telescope aperture are violated. Atmospheric turbulence is usually visible to the naked eye in the form of twinkling of starlight, also called scintillation. This is a random intensity variation caused by interference, constructive and destructive, of coherent rays from the same source that have experienced phase retardations due to the continuously changing path lengths. For example, if some waves are retarded by a path length of half a wavelength, then they will cancel exactly with waves that have no delay when combined at the eye because wave crests fall on wave troughs. It is often more useful to deal with the phase delay (φ) rather than the path length (W), but they are simply related by $\varphi = (2\pi/\lambda)W$; thus, when $W = \lambda/2$, then $\varphi = \pi$ (radians) or 180°. It is well-known that twinkling is

significantly reduced or absent when using a telescope of modest size, and that the image seen with a large telescope is a blurred disk, the seeing disk. This implies that the larger aperture has averaged out the interference effects; it is now unlikely that most waves are exactly 180° ($\lambda/2$) out of phase with the rest. Likewise, if we look at an extended object, such as one of the solar system planets (Venus, Mars, Jupiter, or Saturn) we again see little or no scintillation because the light from different parts of a resolved object is not coherent (i.e., not in phase with other parts).

Air turbulence can occur in several different locations. Heat sources within the telescope dome cause local turbulence. For example, if the primary mirror is warmer than the nighttime air in the dome then convection cells arise above the mirror and image quality is degraded. Various measures such as ventilation and/or air-conditioning of the dome during the day are used to try to reduce the temperature differences. Wind patterns around the telescope dome can also cause turbulence. The wind speed must adjust from zero at the ground to the average wind velocity in the free atmosphere several hundred meters above the dome, which results in a turbulent "boundary" layer. During the day this layer can be thick (1 km) and very turbulent, but during nighttime the boundary layer collapses to a few hundred meters and becomes stratified. There is a strong temperature gradient up through the troposphere (the lowest portion of the atmosphere; *tropos* means turning or mixing) to the tropopause at 10-12 km where the stratosphere begins, and therefore wind shear will cause mixing and index of refraction fluctuations.

Image quality will be directly related to the statistics of the random perturbations on the incoming wavefront. To make any predictions about the turbulence requires a statistical model of the mixing. When the average velocity v_{av} of a viscous fluid with a characteristic length L is gradually increased the fluid motion is at first smooth and regular in space and time; this is called laminar flow. As the velocity increases the motion of the fluid becomes unstable and then random (turbulent) when v_{av} exceeds a critical value. The Reynolds number ($Re = v_{\text{av}}L/\nu$), where ν is the viscosity of the fluid, describes the balance between inertial and viscous forces. When Re exceeds the critical value ($Re > 100$) the flow transitions from laminar (viscous) to chaotic. For the atmosphere, $Re \sim 10^6$, which is high and therefore the fluid is turbulent in both space and time and exhibits considerable spatial structure. Although there is debate on how well it fits the real atmosphere, the properties of such a turbulent field are usually described statistically in terms of a structure function $D(r)$, and the most commonly adopted form is the Kolmogorov turbulence developed in the 1940s by Russian mathematician Andrey Kolmogorov (1903–1987) which gives the variation in refractive index (n) between two points on the wavefront separated by a distance $r = (r_1 - r_2)$

$$D_n(r) = \langle |n(r_1) - n(r_2)|^2 \rangle = C_n^2 r^{2/3} \tag{2.2}$$

In Equation (2.2) the symbols $\langle\,\rangle$ indicate the average value of the square of the difference in the refractive indices between the two points, and it turns out that this average is proportional to the separation r to the two-thirds power. The factor C_n^2 is a measure of the strength of turbulence and its value can range from $10^{-14}\,\text{m}^{-2/3}$ or

$10^{-15}\,\mathrm{m}^{-2/3}$ near the ground to about $10^{-18}\,\mathrm{m}^{-2/3}$ above 10 km. C_n^2 is not a constant and it varies with site and conditions, especially the local conditions right above the ground. It is usually written as $C_n^2(h)$ to show that it is a function of height (h).

Kolmogorov's work on the mechanical properties of turbulence was based on the idea that velocity fluctuations occur on a wide range of space and time scales which form a turbulent cascade from larger to smaller scales. Energy enters at low spatial frequencies. These large eddies are characterized by their "outer" scale length L_0 which corresponds to a spatial frequency of $k_{L_0} = 2\pi/L_0$ and might generally be at higher elevations, but will vary according to the local conditions. A mean value of $L_0 = 24\,\mathrm{m}$ was obtained from data taken at Cerro Paranal in Chile, for example (Conan *et al.*, 2000). The energy is transported to smaller and smaller eddies, and at $Re \sim 1$ the kinetic energy of the flow is converted into heat and the scale length drops rapidly to the "inner" value of l_0 which varies from a few millimeters near the ground to a centimeter high in the atmosphere. There is an interesting way to see that the $r^{2/3}$ law makes sense. The mean kinetic energy per unit mass in a turbulent cell is $v^2/2$ where v is the velocity of the cell, and the energy dissipation rate per unit mass (e) that gives rise to the turbulence is then $\sim v^2/\tau$ where $\tau = r/v$, because r is the local spatial scale of the cell. So, e is given by v^3/r, and solving for the velocity we get $v \sim (er)^{1/3}$ and the energy per unit mass is $\sim v^2$ or $e^{2/3}r^{2/3}$.

Between the inner and outer scale limits, small-scale fluctuations have a behavior that is independent of flow geometry. The statistical distribution of the size and number of these eddies is characterized by the spectral density function $\Phi(k)$, which is the Fourier transform of the covariance function $B_n(r)$. The covariance function is the average value of the product of two quantities, the value of the random variable at one point minus its mean value and the value of the variable at a nearby point minus the mean value. Often, the mean value of the random variable (f) is zero, and then $B(r) = \langle f(r_1)f(r_1 + r)\rangle$ which is also called the autocorrelation function, and the random variable in question is of course the refractive index. The variance $B(0) = \langle f \rangle - \langle f \rangle^2$ and the structure function is then related to these statistical terms by $D(r) = \langle [f(r_1) - f(r_1 + r)]^2 \rangle = 2[B(0) - B(r)]$. Stated differently, the structure function is a measure of the intensity of the fluctuations of a non-stationary random variable over a small enough time scale so that the difference function is essentially stationary. The power spectrum of the refractive index fluctuations obtained by Kolmogorov is given by:

$$\Phi_n(k) = 0.033 C_n^2(h)|k|^{-11/3} \tag{2.3}$$

The value of $|k|$ lies within $1/L_0$ and $1/l_0$. Kolmogorov's analysis allowed the inner scale to go to zero and the outer scale to infinity. In astronomical adaptive optics it is the outer scale that is of greater significance. The refractivity relation for air allows C_n to be related to the temperature structure C_T because $C_n = 77.6 \times 10^{-6}(P/T^2)C_T$ and hence $D_T(r) = C_T^2 r^{2/3}$. Thus, it is clear that the origin of seeing lies in the multitude of turbulent atmospheric cells acting like weak lenses, but we have not yet predicted the size of the seeing disk. To do that, we need to recognize that the lenses have the effect of reducing the coherence of the beam. Is there an aperture size

that is sufficiently small so as to maintain coherence all the way to the telescope? The answer is yes, and it was discovered by David Fried in 1965 (Fried, 1966).

2.3.2 Fried parameter

The stellar wavefront incident on the telescope has spatial variations in both amplitude and phase. Amplitude variations cause scintillation (twinkling) and contribute much less to image quality and seeing than phase variations. By finding the effect of a thin turbulent layer of thickness δh on an incident plane wave it is possible to derive a structure function for phase variations and then integrate this through the atmosphere. The phase structure function $D_\varphi(r)$ across the entrance of the telescope for Kolmogorov turbulence is given by

$$D_\varphi(r) = 6.88\left(\frac{r}{r_0}\right)^{5/3} rad^2 \tag{2.4}$$

where

$$r_0(\lambda, z) = 0.185\lambda^{6/5}\cos^{3/5} z\left(\int C_n^2\, dh\right)^{-3/5} \tag{2.5}$$

or

$$r(\lambda) = \left(\frac{\lambda}{\lambda_0}\right)^{6/5} r_0 \tag{2.6}$$

and λ is the wavelength, z is the zenith distance angle of the star, and C_n is the structure constant for refractive index variations and is integrated through the atmosphere. We can therefore characterize the size of the turbulence cells by a length known as the Fried parameter (r_0, pronounced *r*-naught) which is the length over which the wavefront is not significantly perturbed. The Fried parameter is a critically important quantity. It has two physical interpretations. First, it is the aperture over which there is approximately one radian ($1/2\pi$ waves) of rms phase error, and, second, it is the aperture which has the same "resolution" (as defined by Fried) as a diffraction-limited aperture in the absence of turbulence. Typically, $r_0 = 10$ cm at $\lambda_0 = 0.5$ μm in the visible part of the spectrum. For example, if the seeing is found to have an r_0 of 20 cm at the Keck 10 m telescope on Mauna Kea, then the image resolution will be no better than that given by a 20 cm (~8 inch) telescope! This is a serious problem.

Note that according to Kolmogorov theory, r_0 is larger at longer wavelengths. Seeing should be noticeably better at infrared wavelengths than at visible wavelengths for a large telescope. While this is generally observed to be true, it is important to remember that the expressions above are based on theory. Since the turbulent elements responsible for seeing generally last longer than the transit time across their diameters, it is the wind velocities at different heights in the atmosphere that determine the temporal variations in the wavefront. The speed of the turbulence varies with height, but is often described by an average velocity (v) which is about 10 m/s.

The temporal variations across the wavefront are given by

$$\tau_0 \approx 0.314 \frac{r_0}{v} \tag{2.7}$$

which is only a few milliseconds. The effects of turbulence happen on a fast time scale. The Greenwood frequency $f_G \sim 0.43v/r_0$ indicates the rate at which corrections are needed. For the values we have been discussing, $f_G \sim 43$Hz. Substituting the telescope diameter D for r_0 gives an estimate of the slower time scale for seeing image motion. Note that motion in the image plane is independent of wavelength.

Using a sharp "knife-edge" half-way across and precisely in the focal plane of the telescope to obscure part of the seeing disk of a star allows the observer to look back into the telescope and view the illumination of the primary mirror and to detect the direction, speed, and size of the changing pattern of turbulent cells. The image (sometimes called the Schlieren pattern) appears turbulent and is composed of changing light and dark patches. Horace W. Babcock performed pioneering studies of this phenomenon from about 1936 onwards, making many hundreds of visual knife-edge observations at Mt. Hamilton, Mt. Locke, Mt. Wilson, Mt. Palomar, and Cerro Las Campanas. The value of r_0 is observed to be site-dependent and also wavelength-dependent, being slightly better in the infrared. At most times at most sites, r_0 is a few centimeters. Values of r_0 as large as 100 cm, at a wavelength of 2.2 μm, have been reported for Mauna Kea. For the very large 8 m–10 m telescopes now in operation, even an r_0 of ~1 m is much too small.

Figure 2.4 shows what we could expect when taking a CCD exposure of a star image and then plotting the intensity profiles through the center of the image. If the exposure time is shorter than τ_0, the time taken for one cell of turbulence to move a

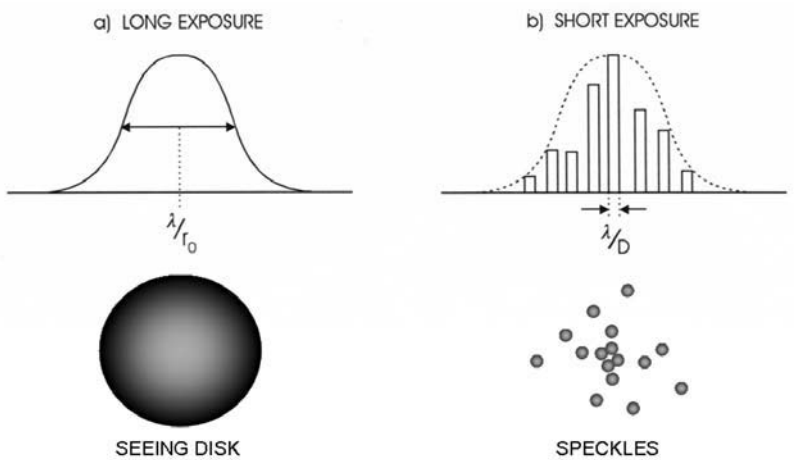


Figure 2.4. An illustration of what to expect when taking a CCD exposure of a star image and plotting the intensity profiles through the center of the image. If the exposures are short enough we see speckles of size determined by diffraction. For a long exposure the speckles blur into a broad seeing disk determined by atmospheric turbulence.

distance r_0 across the primary mirror diameter (D), then several small sharp images are seen. The total number of small images or "speckles" in the two-dimensional image is approximately $(D/r_0)^2$ and the width of each speckle is essentially the diffraction width $\sim\lambda_0/D$. Another short exposure will result in another set of small, sharp speckles, but each will be displaced randomly from the first set. When numerous short exposures are taken and added together, or an equivalent long exposure is taken, all of the displaced speckles blend together to form a broad seeing disk of width λ_0/r_0, instead of a sharp image with diffraction-limited width λ_0/D. This is consistent with the interpretation of the Fried parameter as defining the equivalent diffraction-limited aperture. Unless explicitly stated otherwise, r_0 usually corresponds to $\lambda_0 = 500\,\text{nm}$. The larger r-naught can be, the fewer the speckles, the better the long-exposure seeing, the longer the coherence time (τ_0), and the larger the Strehl ratio.

2.3.3 Speckle interferometry

One way around the seeing problem then, at least for sufficiently bright objects, is to take numerous snapshots, of such short duration that the image motion which leads to blurring is essentially "frozen". Using a camera with a high magnification in order to see the tiny diffraction-limited images, a movie of these snapshots shows a small compact image which changes its position randomly with time; this is the speckle pattern. Numerous short speckle images can be analyzed after the fact to reconstruct the diffraction-limited image using Fourier analysis techniques. This method is known as speckle interferometry and was pioneered by French astronomer Antoine Labeyrie in the 1970s. Apart from data handling, the primary drawback of speckle interferometry is that short exposures are required which in turn places a heavy demand on detector performance. Short exposures tend not to be limited by the fundamental photon noise of the background, but rather by the readout noise of the detector (unless this is exceptionally low). Detector noise restricts the application of the method to brighter stars. Figure 2.5 illustrates how impressive this technique can be, however. Observations of two close binary stars obtained by Andrea Ghez (UCLA) using an infrared speckle camera at a wavelength of $2.2\,\mu\text{m}$ are shown. Normal direct, seeing-limited imaging shows a single object, but speckle imaging reveals extremely close companions. The speckle pattern consists of many doubled images which can all be analyzed by converting to the spatial frequency domain with a Fourier transform. The fringe patterns on the left are the power spectra, or square of the Fourier amplitudes. Fringe spacing decreases as binary separation increases. Using an inverse Fourier transform of the amplitudes and phases of the speckle pattern allows an image to be constructed. More discussion of image reconstruction and Fourier techniques is given in Section 10.4.

2.3.4 Lucky imaging

A simpler strategy is to record each of the snapshots and use a computer process to "shift-and-add" each image so that the brightest speckles coincide with each other,

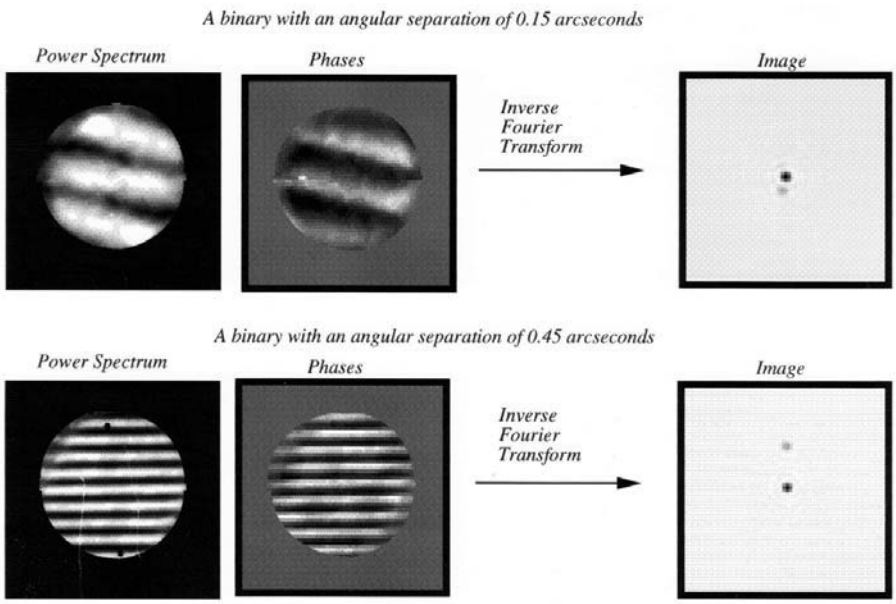


Figure 2.5. Fourier transform methods are used to analyze the spatial frequency spectra of both the amplitudes and phases of the speckle pattern. Two infrared observations of binary stars at a wavelength of 2.2 μm are shown. Credit: Andrea Ghez, UCLA.

but detector noise is still a problem. An extension of this idea is now available thanks to yet another innovation in technology involving CCDs. The concept is called "lucky imaging" and it requires that only the best of the many short exposures are selected for processing into the final image. This technique is not limited to bright objects because an extremely sensitive form of CCD is used in which the pixels act as "avalanche" photodiodes and produce a large "gain" (like a photomultiplier tube) so that a faint photon signal becomes a strong photo-electron signal which thus overcomes the intrinsic readout noise of the CCD. Advocated by Craig Mackay (Cambridge, U.K.), one of the pioneers of astronomical CCDs, the method uses devices made by *e2v technologies* in Chelmsford (U.K.) called "low light level" or L3CCDs. Another name for CCDs like this is electron-multiplied (EM). These devices offer a simple way to achieve excellent images by waiting for those lucky moments of excellent seeing, but at the expense of rejecting a large fraction of the snapshots. We will return to this class of CCDs in Section 7.5.1.

2.4 ADAPTIVE OPTICS

2.4.1 Measuring the wavefront

Is there a more general way to undo the effects of r_0 and get the angular resolution in long exposures to approach λ/D? The first step is to compensate for the random

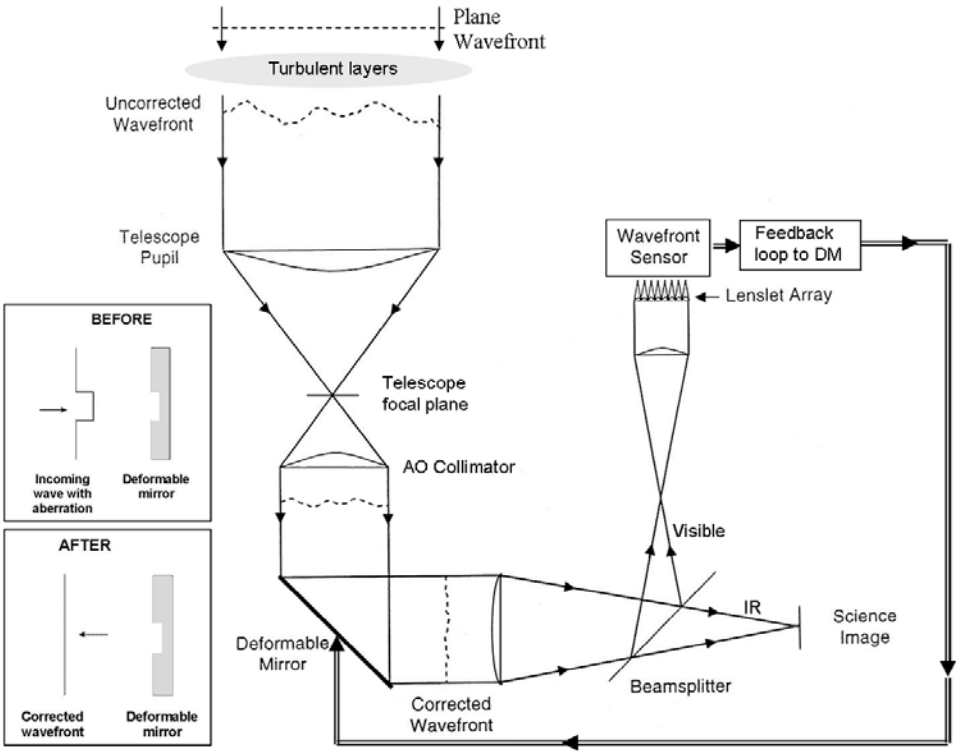


Figure 2.6. The basic layout of an AO system. Plane waves arriving at the Earth from a distant star become distorted by randomly moving cells of air with different indices of refraction. These distortions are sensed at high speed and corrections are sent to change the shape of a deformable mirror which corrects the wavefront at reflection.

wandering of the center (or centroid) of the seeing disk by means of a simple tip-and-tilt motion of a mirror, which could even be the secondary mirror of the telescope, and in this way redirect the overall wavefront to the same (fixed) place in the image. The second step is more complex. An optical arrangement is used to project a real, de-magnified image of the primary mirror of the telescope onto a much smaller mirror whose detailed "shape" can be changed or "deformed" by the forces applied from numerous small actuators behind it. To do this, a high-speed CCD camera acting as a wavefront sensor detects the changes in the slope of the coherent areas of turbulence, also referred to as sub-apertures. By analyzing the wavefront distortions very rapidly, it is possible to control the figure of the deformable mirror in such a way as to compensate for the wavefront distortions. The basic layout is illustrated in Figure 2.6 which shows the disturbed wavefront entering the telescope aperture and being re-imaged onto the deformable mirror (DM). If the otherwise plane mirror is deformed to have the same shape but *half* the wavefront error of the incoming wave, then the reflected wavefront will be very nearly a plane wave after leaving the DM. The reason for using half the distortion on the DM is because the corresponding part

of the wave must travel into and out of the depression in the DM, thus catching up with the rest of the wave as it reflects from the DM's surface. To determine what corrections are needed, part of the light is picked off by a beam-splitter and sent to the wavefront sensor. Correction signals generated by the wavefront sensor are fed back to the deformable mirror to "close the loop".

Wavefront sensors fall into two basic groups. Direct sensors split the pupil image plane into sub-apertures and use the intensity in each sub-aperture to deduce the phase of the wavefront. Within this group there are two sub-categories known as slope-sensing and curvature-sensing. In the slope-sensing group there are three methods known as Shack–Hartmann, lateral shear interferometer, and (more recently) pyramid sensors. Indirect sensors deduce wavefront properties from whole-aperture intensity measurements made at or near the focal plane. These methods include image sharpening, multi-dither, phase diversity, and phase retrieval. Techniques that use sub-apertures place greater demands on the CCD detector because less light is collected for each sub-aperture.

Among the several types of wavefront sensors available, there has been a general trend towards the Shack–Hartmann slope-sensing type with many sub-apertures. As illustrated in Figure 2.7, the basic construction is that of an array of tiny lenses or "lenslets" which is placed near a pupil image to produce a pattern of many star images on the detector, each corresponding to a different part of the primary aperture. By rapidly finding the centroids of each image it is possible to derive the slope of the wavefront at that instant. As an example, on the Keck 10 m telescope the pupil image on the lenslet array is reduced in size by a factor of 2,800 from 10 m. The lenslet array is approximately 18×18 lenses each about 200 microns in diameter for a

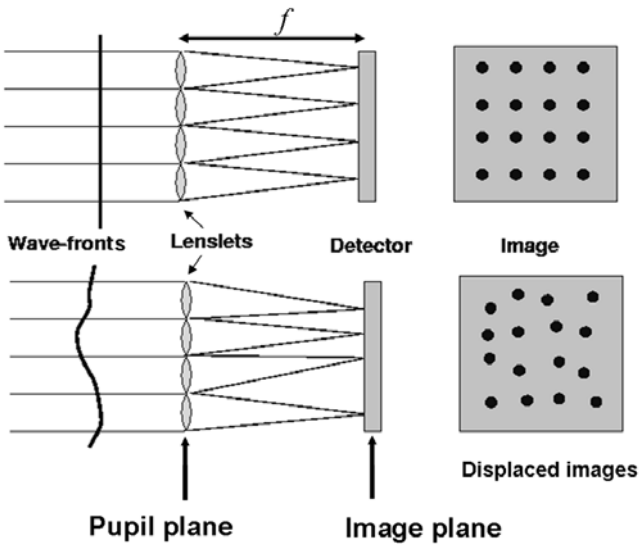


Figure 2.7. The operation of a Shack–Hartmann wavefront sensor system. A small format high-speed CCD is required to record the displacements of the images.

total size of about 3.6 mm on a side. To see how big the 200 μm sub-aperture corresponds to on the primary mirror, multiply the lenslet diameter by the magnification of 2,800 to get 56 cm. While this is completely consistent with the expected value of the Fried parameter r_0 from 1 μm to 2 μm, it clearly implies a much smaller light-gathering area for the wavefront sensor. In addition, as the lenslet size of 200 μm is much larger than a typical CCD pixel (~20 μm), then normal lenses will be needed in front of the CCD to reduce the scale. Typically, each spot might be matched to a 3×3 grid of pixels, with the inner 2×2 forming a "quad-cell" arrangement. A displacement in x can be found by comparing the difference in intensity between the left-hand and right-hand quadrant pairs, dividing by the total summed intensity in all quadrants and multiplying this ratio by the radius of the spot image. The same applies for y displacements. If the sub-aperture is about the same as r_0, then the wavefront error (WFE) is roughly given by 6.3/SNR (radians) where SNR is the signal-to-noise ratio of the measurement. If we require that the WFE be less than $\lambda/20$ or $\pi/10$ ($= 0.314$) radians, then SNR should be greater than 20. Typically, small high-speed CCDs with frame rates of a few thousand per second are needed with high QE (>80%) and low noise (<3 electrons).

In practice, the phase variations in the wavefront are described in terms of some algebraic quantities called Zernike polynomials $Z_j(n, m)$ in which n is the degree of a radial polynomial and m is the azimuthal frequency of a sinusoidal term. The value of this kind of description is that it allows aberrations in the wavefront to be visualized in almost the same way that the surface of the deformable mirror will be treated. Ray-tracing programs can provide theoretical wavefront patterns in terms of Zernike polynomials to help assess the effects of lenses. Noll (1976) gives normalized versions of the Zernike polynomials such that the rms value of each polynomial over the circle is unity. Table 2.1 lists the low-order terms and explains their meaning. The final column gives the mean square residual amplitude Δ_j in the phase variations at the telescope entrance caused by Kolmogorov turbulence *after* removal of the first j terms; the normalization factor is $S = (D/r_0)^{5/3}$. For large j the approximate value of Δ_j is given by

$$\Delta_j \approx 0.2944 j^{-0.866} \left(\frac{D}{r_0}\right)^{5/3} rad^2 \tag{2.8}$$

By substituting the appropriate terms and taking the square root, the last column of Table 2.1 yields the rms phase variation for each mode of correction. With no correction, the rms phase variation is $0.162(D/r_0)^{5/6}$ waves, while after tip/tilt correction the rms phase variation is reduced to $0.053(D/r_0)^{5/6}$ waves. If SR is the Strehl ratio, then the Maréchal approximations give

$$1 - SR \approx \Delta \approx 1 - e^{-\Delta} \tag{2.9}$$

for small wavefront disturbances. According to the above theory, it requires a huge number of terms to achieve a Strehl ratio as high as 80% under average seeing conditions on a very large telescope, which in turn implies a large number of actuators on the deformable mirror.

Table 2.1. Modified Zernike polynomials and the mean square residual amplitude for Kolmogorov turbulence after removal of the first j terms.

Z_j	n	m	*Expression*	*Description*	Δ_j/S
Z_1	0	0	1	Piston	1.030
Z_2	1	1	$2r\sin\varphi$	Tilt	0.582
Z_3	1	1	$2r\cos\varphi$	Tilt	0.134
Z_4	2	0	$\sqrt{3}(2r^2-1)$	Defocus	0.111
Z_5	2	2	$\sqrt{6}r^2\sin 2\varphi$	Astigmatism	0.0880
Z_6	2	2	$\sqrt{6}r^2\cos 2\varphi$	Astigmatism	0.0648
Z_7	3	1	$\sqrt{8}(3r^3-2r)\sin\varphi$	Coma	0.0587
Z_8	3	1	$\sqrt{8}(3r^3-2r)\cos\varphi$	Coma	0.0525
Z_9	3	3	$\sqrt{8}r^3\sin 3\varphi$	Trifoil	0.0463
Z_{10}	3	3	$\sqrt{8}r^3\cos 3\varphi$	Trifoil	0.0401
Z_{11}	4	0	$\sqrt{5}(6r^4-6r^2+1)$	Spherical	0.0377

2.4.2 The isoplanatic patch

Unfortunately, the correction process described above has several limitations. First, the star being used as a reference for the correction process must be sufficiently bright to provide a good signal-to-noise ratio in each of the sub-apertures across the primary mirror. Generally, the target object being observed will be much too faint to be used as its own reference. Second, the effectiveness of the compensation decreases as the angular distance from the reference star increases, simply because the wavefront distortions are not the same. Perfect compensation is limited to an angular patch of radius a few arcseconds around the reference star. This region is known as the isoplanatic region. The isoplanatic angle is usually defined as the radius of a circle over which the wavefront disturbance is essentially identical and is given approximately by

$$\theta_0 = 0.314\frac{r_0}{H} \tag{2.10}$$

where H is the average distance to the seeing layer. Sometimes the term is also used to refer to the angular distance over which image motions are practically the same as compared with seeing widths. In this case it is better to use the term "isokinetic patch", which is of order $0.3D/H$. As an example, consider an 8 m telescope at a site for which $r_0 = 13.3$ cm (corresponding to seeing of $0.75''$ at $\lambda = 500$ nm) and assume that the distance to the seeing layer is $H = 5{,}000$ m. Then, the radius of the

isoplanatic patch is only 1.7″, whereas the isokinetic patch is much larger at about 100″.

2.4.3 Deformable mirrors

Types of deformable mirrors (DMs) include segmented, continuous face-sheet, bimorph, Micro-Electro-Mechanical Systems (MEMS), and liquid crystal devices. It is also possible to make the secondary mirror of the telescope a deformable mirror. Most deformable mirrors currently use a thin ''face'' sheet of low-expansion glass supported on an array of many small, discrete actuators, which consist of several stacked and bonded layers of lead–magnesium–niobate (PMN) known as electrostrictive actuators. This material gets longer or shorter in response to an applied electric field. Devices with over 900 actuators are already available. The DM must respond faster than the coherence time of the atmosphere, and the range of motion, the stroke, needs to be several microns for a large telescope. Segmented DMs are just like the name implies. Each segment has 1–3 actuators and can correct in–out motion (piston) and tip–tilt. Bimorph devices consist of two piezoelectric wafers bonded together with an array of electrodes between them; the front surface acts as the deformable mirror. These devices are well-matched to curvature-sensing AO systems. Other technologies also hold promise and are being actively studied for astronomical systems including liquid crystal devices and MEMS devices. MEMS technology is fabricated using silicon micromachining, and actuation of the membrane mirror is by electrostatic forces. These DMs look like CCDs! Already there are formats with 1,000 actuators and a stroke of 2 μm, and 4,000 actuator devices are under development. Figure 2.8 shows a large continuous face-sheet DM in use at the Keck Observatory and a large MEMS DM, together with a bimorph mirror DM suitable for curvature sensing.

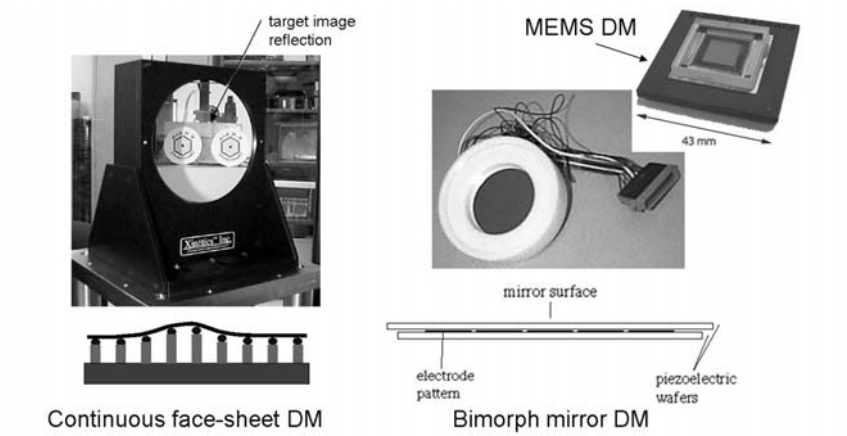


Figure 2.8. Three types of deformable mirrors. From left to right there is a continuous face-sheet driven by actuators, a bimorph mirror, and a large 43 mm MEMS device with a continuous face-sheet and about 1,000 elements.

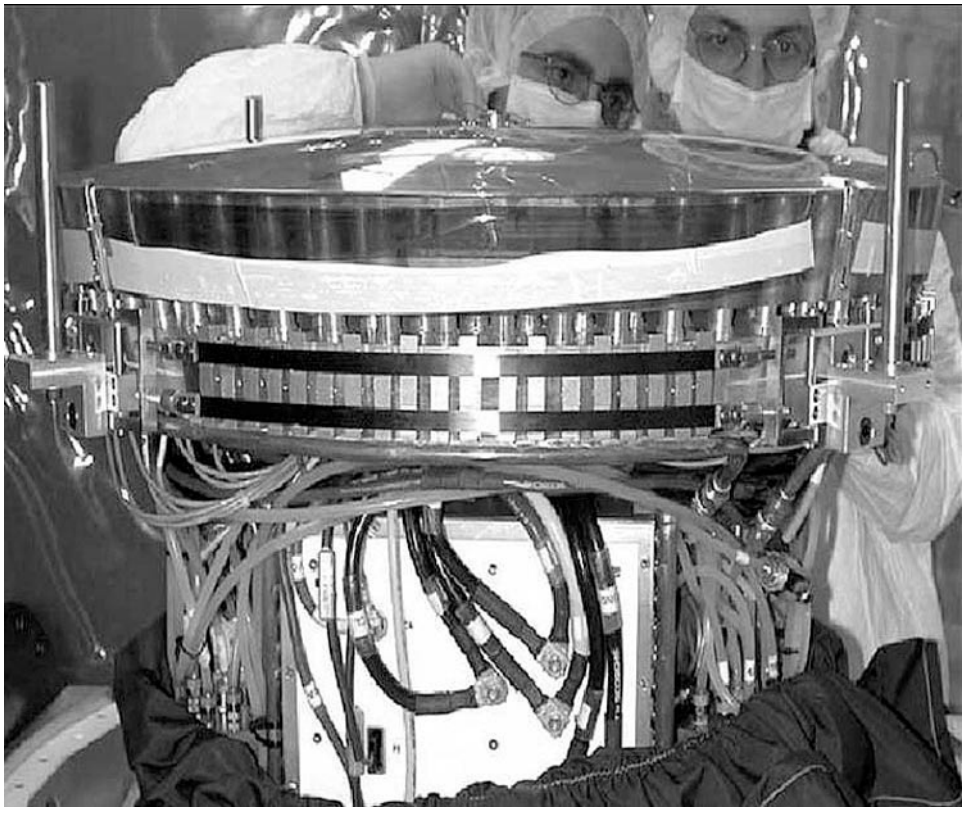


Figure 2.9. The deformable secondary mirror for the Monolithic Mirror Telescope (MMT) on Mount Hopkins. Credit: University of Arizona.

Pioneered by the University of Arizona (U.S.A.) and Arcetri Observatory (Italy), a different approach which eliminates lots of optical components places the deformable mirror as the secondary mirror of the telescope. Such a system was recently installed on the 6.5 m Monolithic Mirror Telescope (MMT) on Mt. Hopkins (Arizona) and a larger system is planned for the Large Binocular Telescope on Mt. Graham (Arizona). Figure 2.9 shows the MMT deformable secondary during assembly with over 300 magnetically activated force actuators to provide the bending force. In operation this system yields a factor-of-20 improvement of the Strehl ratio at the operating wavelength of 1.65 μm.

2.4.4 Natural guide star systems

In the late 1980s many observatories set out to develop adaptive optics systems. It was quickly recognized that corrections made in the visible part of the spectrum using special CCD-based wavefront sensors would provide good performance in the near-infrared for science instruments based on the new generation of infrared array

cameras. One of the earliest successes was the Come-On system (later upgraded and called Adonis) developed by the Observatoire de Meudon (France) and the European Southern Observatory (ESO) for the ESO 3.6 m telescope in Chile, but initially tested on the 1.52 m telescope at Haute Provence in France (Rousset *et al.*, 1990). Initially, this system had a deformable mirror with 19 piezoelectric actuators and a Shack–Hartmann wavefront sensor with 5×5 sub-apertures and a Reticon 100×100 photodiode array. The infrared science camera had 32×32 pixels. In 1993 the final system on the 3.6 m telescope employed 52 actuators and a correction rate of 400 Hz. At the time of writing it is still in operation using several different modern cameras. In 1997 ESO began a collaborative effort with a French consortium to develop the Nasmyth Adaptive Optics System (NAOS) for the Very Large Telescope observatory at Paranal in Chile. First light with this powerful system was obtained in November 2001. NAOS was unique because it had the option to use an infrared wavefront sensor. In the U.S.A., formerly classified experiments demonstrated 241 actuators and correction rates of 2 kHz. Spinning off from that work, an early astronomical system developed at the Lawrence Livermore National Labs (LLNL) under the leadership of Claire Max (now at the Center for Adaptive Optics, University of California, Santa Cruz and one of the early pioneers of AO) and Scott Olivier, was successfully deployed at the University of California's Lick Observatory in 1996 as a prototype for a system for the Keck telescope. That system is shown in Figure 2.10. The Shack–Hartmann wavefront sensor was configured with a triangular

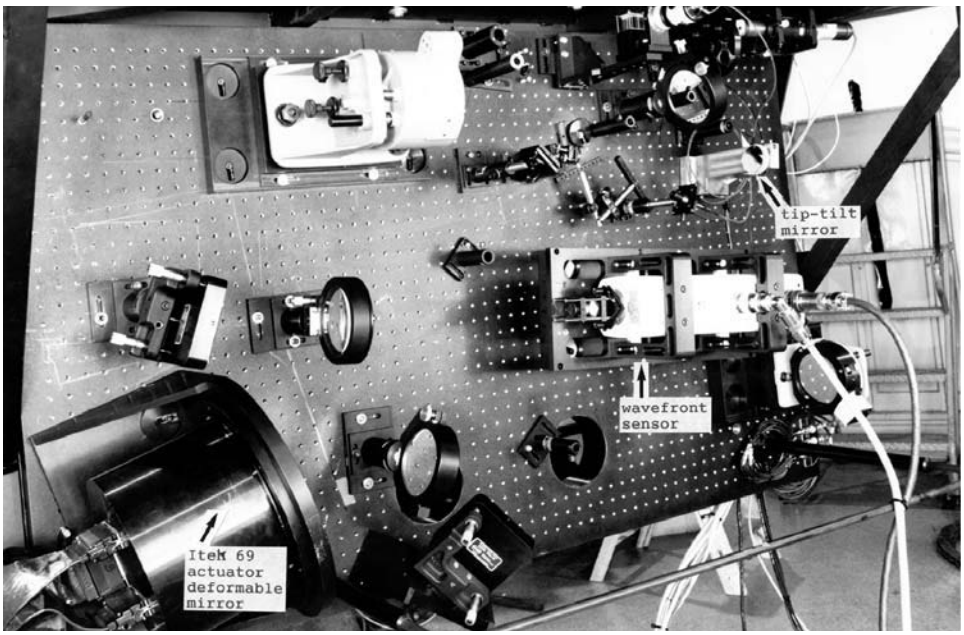


Figure 2.10. The Lick AO system optical bench (vertical) was one of the first in routine operation at a telescope. Credit: Claire Max.

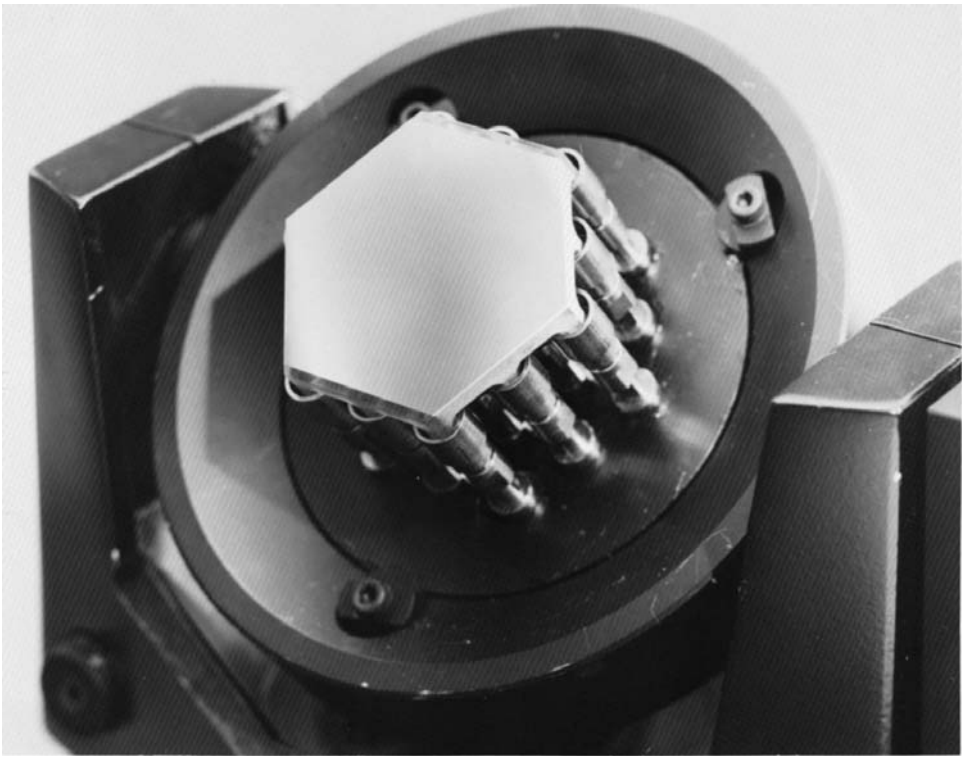


Figure 2.11. Close-up of the 37-element deformable mirror for the Lick 3 m telescope. Credit: Claire Max.

array of 37 sub-apertures each of which mapped to a diameter of 46 cm at the primary mirror of the 300 cm telescope. Images from the lenslets were recorded using a special high-QE, low-noise, very fast CCD camera built by Adaptive Optics Associates using a 64×64 pixel chip developed by Lincoln Laboratory. The wavefront re-constructor was based on four Intel i860 chips on a Mercury VME board controlled by a Unix workstation. The deformable mirror, shown in Figure 2.11, had a triangular array of 127 PMN actuators with a nominal stroke of 8 μm at 80 volts. The inter-actuator spacing of 11 mm at the DM corresponds to the 46 cm apertures mapped onto the telescope. A separate tip/tilt system conserved the range of the DM and additionally, provided tip/tilt information from a natural guide star during operation of the system with laser guide stars. The Lick AO system is used with an infrared camera developed by James Graham (UC Berkeley). A major redesign of the optics occurred from 1999 to 2001 and the system went into general use. Similar early efforts were carried out at the Canada–France–Hawaii Telescope by François Roddier, another pioneer of AO, with particular emphasis on curvature sensing, the use of bi-morph mirrors (Roddier, 1981, 1988) and a small number of avalanche photodiodes (APDs) to provide high sensitivity.

Since these early efforts to develop natural guide star AO systems, progress has been remarkable. Many of the basic building blocks are the same but computers and actuators have improved, and dozens of fully operational AO systems are now in use on telescopes of all sizes.

2.4.5 Laser guide star systems

To correct for atmospheric turbulence one must first measure the effects of turbulence, and this requires starlight. With Shack–Hartmann wavefront sensors the light collected by the telescope is divided among several sub-apertures of size comparable with r_0, and sampling rates of at least 100 Hz are needed. These requirements suggest that a star of visual magnitude around $m = 12$ or brighter will be needed. At visible wavelengths there are roughly 150,000 stars per steradian brighter than $m = 12$. This may seem a lot until we remember that the star must lie within the isoplanatic patch. Assuming that the isoplanatic patch is about 2 seconds of arc ($\sim$9.6 μrad) then there are 10^9 patches in the sky and only 150,000 suitable stars. In general, *no* natural guide star will be close enough to a particular target! Thus, natural guide star AO systems are said to have poor sky coverage. One way around this problem is to select the guide star and then perform scientific observations in that neighborhood. This works well when the targets are random galaxies at high redshifts because any direction in the sky is acceptable for finding such faint objects. Alternatively, one can use the curvature-sensing method which allows for much fainter guide stars by using all of the light collected by the telescope. But when no suitable reference star is found within the isoplanatic patch, then adaptive optics systems cannot function.

However, a remarkable alternative approach is now available. The idea is, quite simply, to create an "artificial star" using laser beams. First suggested in non-classified literature by Foy and Labeyrie in 1985, but already under extensive study by the U.S. military, the concept is as follows. A pulsed laser beam tuned to the wavelength of the orange-colored sodium D resonance line (589.0 nm and 589.6 nm) is projected through a telescope and focused on the so-called sodium (Na) layer in the upper atmosphere at an altitude of 92 km. Resonance fluorescence in this layer produces a glowing, artificial star. Outgoing laser pulses return from the Na layer in only 700 μs, retracing their path exactly, enabling the adaptive optics system to update the shape of the deformable mirror. The primary challenge is the fact that the required narrow-line laser source is not easy to construct. Several different laser technologies are possible including the use of pulsed dye lasers, sum frequency mixing of two lasers, and Raman lasers pumped with a frequency-doubled solid-state laser. There is another kind of laser-generated reference star which employs artificial stars excited by Rayleigh "back-scattering" of the laser beam by air molecules in the stratosphere. Since this method is not based on a narrow resonance, the wavelength is less critical. Sodium can still be used, but other wavelengths also work in principle. The main disadvantage of Rayleigh beacons comes from the lower altitude of the back-scattered light. Bob Fugate and others used lower atmosphere, Rayleigh scattering laser guide stars successfully for wavefront sensing at the USAF Phillips

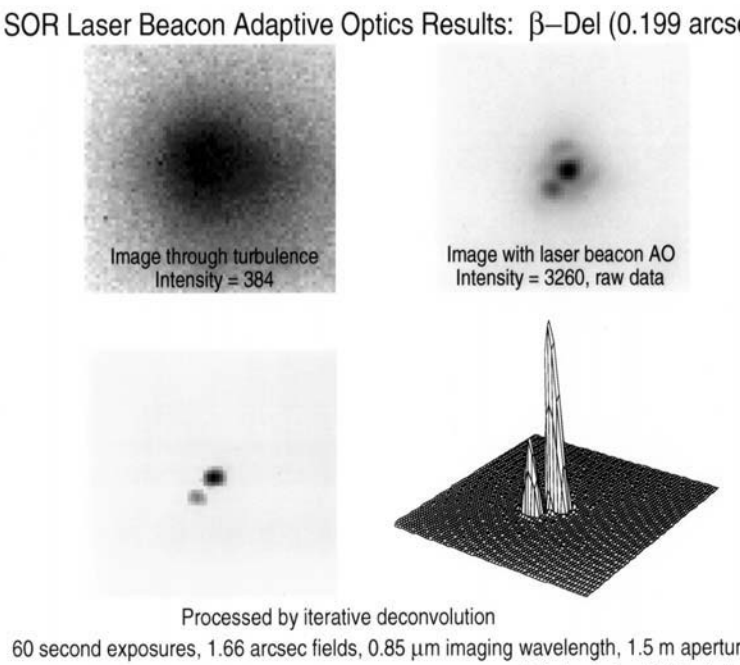


Figure 2.12. Results from the Starfire laser guide star AO system. Credit: Bob Fugate.

Laboratory 1.5 m telescope at the Starfire Optical Range in Albuquerque, New Mexico in the 1990s. One of those pioneering results is shown in Figure 2.12.

The sodium layer at 92 km is about 11.5 km thick and has a column density of neutral sodium atoms of $\sim 2 \times 10^9$ atoms per cm^2. It is believed to be a result of meteoric dust. When illuminated by the laser, Na atoms are excited to a higher energy level from which they return to the ground state either by spontaneous emission ($\sim 10^{-8}$ s) in all directions or by stimulated emission in which the photons are emitted in the same direction as the incoming beam. Some of the downward-emitted photons are collected by the telescope to form an artificial star. As the laser power increases, the brightness of the laser guide star increases too, up to a saturation point corresponding to the spontaneous emission becoming dominant (so that most of the light goes outward). For a luminous patch of 50 cm, corresponding to an angular size of

$$\frac{0.5\,\text{m}}{92{,}000\,\text{m}} \times 206{,}265 \approx 1'' \tag{2.11}$$

the pulsed laser power is about 5 kW for saturation. About 22 W of power will create a 9th-magnitude artificial star with this angular size.

When the laser transmitter is displaced from the telescope, the configuration is called "bistatic", whereas for the case in which the telescope itself transmits the laser beam the term applied is "monostatic". The positional offset in the bistatic configuration causes the laser guide star to be elongated, which decreases the

sensitivity of wavefront corrections. For example, if $d = 10\,\text{m}$ then the elongation is about 3 arcseconds. Most new telescopes install monostatic laser systems, but bistatic configurations can ensure that back-scattering effects in the lower atmosphere fall outside the isoplanatic patch. Jacques Beckers (1993, 1996) has proposed an ingenious scheme for removing the elongation effect. Essentially, a pulsed laser guide star is viewed with microsecond time resolution snapshots. In each snapshot the laser guide star appears as a round spot, but the spot moves across the sodium layer in about 60 μs. Rapid tracking of this motion in the Shack–Hartmann wavefront sensor, by moving the charges along the columns of a CCD at the same rate, should result in a sharp image and good sensitivity. Another new idea proposed by James Beletic (Teledyne Imaging Sensors) is to design a CCD with pixels that optimally match the elongated shape of the beam, so that all the light falls on a pixel anyway. In fact, CCDs are currently being designed with this idea in mind.

Initially, the laser guide stars had visual magnitudes in the range 12–14, which is bright enough for full adaptive optics at near-infrared wavelengths of 1 μm or longer. Claire Max and her team at the Lawrence Livermore National Labs (LLNL) in the U.S.A. were the first group to produce a bright (magnitude 9) laser guide star suitable for high-order adaptive optics corrections with many Zernike terms. The LLNL system was demonstrated on the University of California Observatories (UCO) 3 m Shane telescope at Lick Observatory in September 1996. It is based on a 127-actuator continuous-surface deformable mirror, a Shack–Hartmann wavefront sensor equipped with a fast-framing low-noise CCD camera, and a pulsed solid-state-pumped dye laser tuned to the atomic sodium resonance line at 589 nm. The laser typically generates about 22 W of power which produces a guide star with a visual (V) magnitude of about 8–9 and an image of about $1.0''$ in width with seeing conditions of $0.5''$; the diffraction-limited size of the guide star in the sodium layer is $0.6''$, and of course the beam traverses the atmosphere twice. It is possible to obtain Strehl ratios as high as 0.7 at a wavelength of 2.2 μm in good seeing.

One of the most productive laser guide star systems thus far is operated by the Keck Observatory (Hawaii). Beginning in 1994, the observatory staff under the leadership of Peter Wizinowich, who had already developed the natural guide star system (Wizinowich *et al.*, 2000), began working with Claire Max's group at LLNL to develop an artificial guide star system based on the successful laser demonstration at Lick Observatory. Development of a fully stable and maintainable laser system for the summit of Mauna Kea took several years, but the system was finally ready for astronomers in September 2003 when "lock" was achieved and infrared images of a 15th-magnitude science target were obtained. Wizinowich *et al.* (2006) give a detailed overview of the Keck AO system. The number of nights available to astronomers gradually increased and several infrared cameras and spectrometers (provided by UCLA and Caltech) became available to use behind the AO system. Almost half the time on the Keck II telescope each semester is devoted to AO observing. Figure 2.13 (see also Plate 9) shows the sodium laser beacon projecting from the Keck II dome. In February 2007 a new wavefront controller and new wavefront sensor, including a new CCD, was completed for both telescopes. The reason for providing AO systems on both telescopes is to enable the twin 10-meter apertures to be used as an inter-

Figure 2.13. A long-exposure picture of the sodium laser beacon projecting from the Keck II telescope. Credit: WMKO. See also Plate 9.

ferometer. It is also planned to equip Keck I with a new laser system in the near future. With the new wavefront controller, Strehl ratios peaked at about 70% under the best conditions in the near infrared. Figure 2.14 shows several examples of the diffraction-limited point spread function at near-infrared wavelengths ($J = 1.25$, $H = 1.65$, and $K = 2.2\,\mu$m) demonstrating Strehl ratios of 22%, 41%, and 63%, respectively, and image sizes of less than 0.05 arcseconds (upper panel). The lower panel shows images of Uranus plus Miranda (upper left, enlarged), Neptune, and Titan, respectively, at 2.2 μm demonstrating outstanding resolution even on extended objects. Uranus has an extent of $4''$ and the FWHM on Miranda was $0.053''$. Many images can be seen at the Keck Observatory web site.

2.5 OPTICAL AND IR INTERFEROMETERS

Having achieved the diffraction limit for a single telescope it is tempting to push further. By combining light collected by widely spaced telescopes one can overcome the diffraction limit of an individual aperture and achieve the angular resolution equivalent to the separation of the telescopes. This is the goal of interferometry. Whereas radio astronomers have been using multi-aperture interference techniques for decades, the realization of interferometers for the infrared and optical has been much harder because the wavelengths are about 100,000 times smaller. Nevertheless,

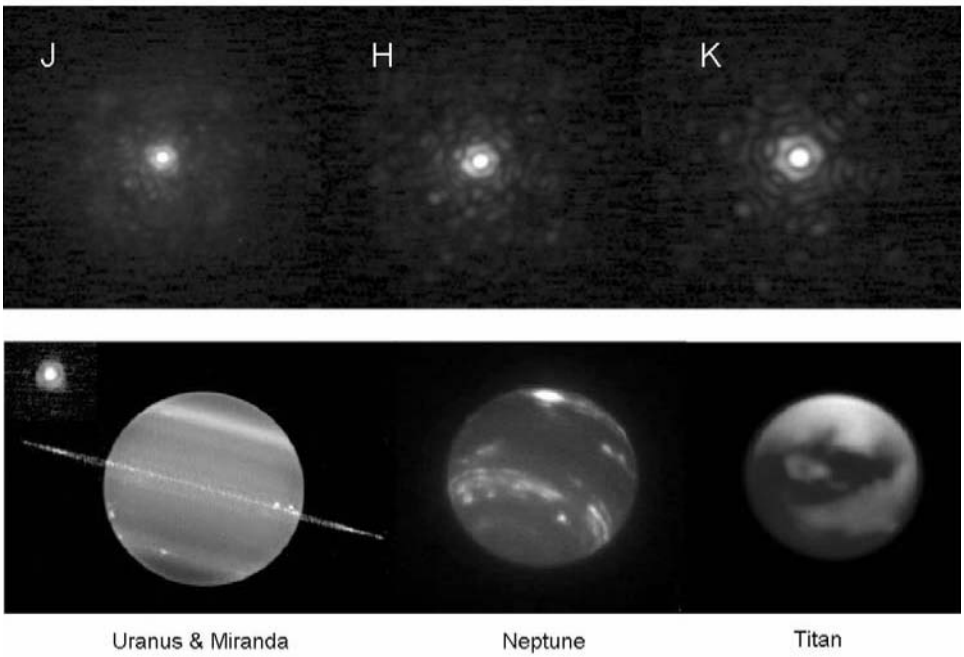


Figure 2.14. Images from the Keck laser guide star AO system. Upper panel, near-infrared images of the same star at three wavelengths (1.25 μm, 1.65 μm, and 2.2 μm) showing diffraction patterns. Lower panel, near-infrared AO images of Uranus (with Miranda insert at a different scale), Neptune, and Titan. Credit: Peter Wizinowich, Keck Observatory.

current optical/IR instruments are achieving the remarkable spatial resolution of 0.001 seconds of arc thanks to amazing engineering feats that allow sub-micron optical stability and coherence to be maintained over hundreds of meters of path length, while also controlling polarization and dispersion over broad wavelength bands. The history of stellar interferometers spans more than a century (Lawson, 2000), but it is only relatively recently that optical techniques have fully matured due to advances in computers, electronic imaging devices, sensitive positioning devices, and adaptive optics. This is a rapidly evolving field. A full discussion of interferometers is beyond the scope of this book. See Monnier (2003) for a comprehensive review and the web sites listed for current progress.

It was the English physician and physicist Thomas Young (1773–1829) who demonstrated the wave nature of light in 1802 and subsequently showed that the spatial structure of an object could be deduced *without* forming a direct image—but by analysis of "wave interference" effects. In this way, unresolved distant objects such as stars might be studied. In fact, Armand Hippolyte Fizeau (1819–1896) proposed an experiment to measure stellar diameters as early as 1868 and Edouard Stephan in 1872 attempted the experiment on Sirius (α CMa) using two apertures separated by 50 cm on the 80 cm refractor at Marseilles, France. He saw interference "fringes" and concluded that much larger separations were needed. In 1890 Albert Michelson

(1852–1931) suggested using two separate mirrors far apart and then combining the beams for visual inspection. The diameters of several red supergiant stars such as Betelgeuse (α Ori) were measured by Michelson and Pease from 1920 to 1930. Further advances had to wait for the technologies of lasers, solid-state detectors, high-speed electronics, and computers.

Maxwell's wave equation has a general solution of the form $y = f(x \pm ct)$ and the simplest type of electromagnetic wave is one in which the function f is a sine or cosine. As mentioned in Chapter 1, this is a plane-polarized monochromatic plane wave. For propagation along the x-axis and with the electric field only in the $x - y$ plane, the magnitude of the wave is given by

$$E_y = E_0 \sin(\omega t + kx + \varphi) \tag{2.12}$$

E_y represents the magnitude of the wave at time t and position x, while E_0 is the maximum amplitude of the wave and φ is the phase at $t = 0$ and $x = 0$. In this expression $k = 2\pi/\lambda$ is the wave number and $\omega = 2\pi\nu$ is the angular frequency. Note that $\omega/k = \nu\lambda = c$. The intensity of the radiation is proportional to E_0^2. What happens when two waves from the same source, but slightly out of phase, arrive at the same location? The answer was given in the famous double-slit experiment by Young. When two waves arrive at a common point P which is at the height h (above the center line) on a screen a large distance L from two small slits (S_1 and S_2), which have a separation of d, and tiny apertures that are very small compared to d, then an interference pattern is formed on the screen. That is, a pattern of light and dark bands known as interference "fringes" is observed. Consider two waves arriving at the point P having traveled different distances S_1P and S_2P and then superimpose the waves (add the electric fields) with a phase difference of $\delta = k\Delta = (2\pi/\lambda)[S_2P - S_1P]$. Assuming monochromatic waves, the intensity of the combined light (proportional to the square of the combined electric fields) on the distant screen is given by

$$I = 4E_0^2 \cos^2(\delta/2) \tag{2.13}$$

Because $\cos^2(\delta/2) = 1$ when $\delta = 0, 2\pi, \ldots, n2\pi$, where n is an integer, the brightness on the screen varies with phase (or position), i.e., a fringe pattern is formed. With $L \gg h$, angles are very small and then the extra path traveled by the light from the more distant slit is $d \sin\theta' \approx d \sin\theta \approx h/L$, therefore the path difference is $\Delta = hd/L$ and the phase difference is $\delta = (2\pi/\lambda)(hd/L)$. The intensity is a maximum when the phase difference is an integer multiple of 2π and this occurs when $\Delta = n\lambda$, hence we get $hd/L = 0, \lambda, 2\lambda, 3\lambda, \ldots, n\lambda$, or bright fringes occur when

$$h = n\lambda(L/d) \tag{2.14}$$

The minimum value of intensity is zero and this occurs when $\delta = \pi, 3\pi$, etc. (because $\cos 90^\circ = 0$) (i.e., when $hd/L = \lambda/2, 3\lambda/2, 5\lambda/2, \ldots, (n + \frac{1}{2})\lambda$), and so dark fringes occur when

$$h = (n + \tfrac{1}{2})\lambda(L/d) \tag{2.15}$$

Expressed in angular terms, the fringe spacing is $\Delta\theta = \lambda/d$ radians where d is the

projected slit separation or "baseline". The spatial frequency of the fringes (the number of fringes per radian) is just the inverse $u = d/\lambda$ rad^{-1}.

This fundamental result on the interference of two coherent waves is the basis for many techniques. One of the earliest applications to astronomy was the Michelson Stellar Interferometer used to measure binary star separations and the angular diameters of stars. Consider a telescope of aperture D. From Rayleigh's criterion for the angular separation of two point sources imaged by a circular mirror of diameter D, we have $\alpha = 1.22\lambda/D$. Now suppose that the telescope aperture is covered by a screen containing two parallel slits as in Young's experiment with a separation $d \sim D/1.22$. Interference fringes are observed. If this instrument is pointed to a double star and the slits are oriented to be perpendicular to a line joining the two stars, and if the angular separation of the two stars happens to be $\alpha = \lambda/2d$, then no fringes will be seen! This occurs because the fringe pattern of one star is exactly 180° out of phase with the other (crests land on troughs) and so the fringe pattern is canceled and the illumination is constant. Without ever taking a normal image, one could *infer* that the star was double from the fringe pattern alone and that the angular separation was $\lambda/2d$ or a multiple of it. In practice the distance d, called the baseline, is made adjustable and changed until the fringes first disappear. Albert Michelson (1852–1931) used this method in 1920 to measure the diameter of the disk of a single star. As very large values of d would be needed to resolve a star, Michelson realized that he could magnify the phase difference by increasing d. He placed two 6-inch mirrors on a steel strut in front of the 100-inch (2.5 m) Hooker telescope on Mt. Wilson and separated them by some 6.1 m (240 inches). These two mirrors fed a second pair about 100 inches apart to form the "double-slit" and to direct the beam to the primary mirror. If the distance between the outer pair is b then fringes will now disappear when $b\alpha = 1.22\lambda$ and so the sensitivity is magnified by b/d. Only seven stars, all red giants, were measured successfully including Betelgeuse and Arcturus. For the latter, Michelson found that the separation of 20 ft gave an angular diameter of 0.02 seconds of arc, which given the known distance implied a true diameter 27 times that of the Sun, very close to the modern value.

Advances in radar during WWII spurred the development of radio interferometry and led eventually to a rather different approach called the intensity interferometer developed by Hanbury-Brown and Twiss in 1956. Based on correlations of intensities (not electric fields) this concept led to the 188 m baseline of the Narrabri Interferometer and because it had blue sensitivity the instrument was able to measure the diameters of dozens of hot stars rather than red giants. This work led to long-baseline stellar interferometry by Antoine Labeyrie in 1975 at Nice (France) using a 12 m baseline by combining the electric fields directly before photon detection. The intensity interferometer relies on the post-detection correlation of the signals. Correlation is measured by the product of the two intensities. Uncorrelated time-dependent terms simply average to zero. Defining the fringe visibility in the same way as Michelson we have that $V = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$. The final integrated product or correlation function $K(b)$ where b is the baseline separation of the telescopes or antennas (equivalent of the double slits) is $K(b) \sim V(b)^2$. This is often called the "Vee-squared" mode. The angular resolution of the two-element

interferometer is just given by $1.22\lambda/b$, so that it is equivalent to that of a larger telescope whose diameter equals the separation of the pair.

Many diverse experiments are under way to develop and apply spatial interferometers. For example, the twin Keck telescopes are used as an interferometric pair with a baseline of 85 m, and the ESO VLT site has been set up to enable the four 8-meter and many smaller telescopes on a movable track system to perform as an interferometer array (the VLTI). Among the larger interferometers using small telescopes is the CHARA (Center for High Angular Resolution Astronomy) array of six 1 m telescopes located on Mt. Wilson (California). Distributed in a Y-shaped configuration providing 15 baselines ranging from 34 m to 330 m, this array is capable of resolving details ~200 micro-arcseconds (McAlister *et al.*, 2005). Thus far, very few "images" have been made by current optical/infrared interferometers because of the difficulty of obtaining enough baselines to cover all spatial frequencies, but new telescope arrays are now being commissioned which will make true imaging possible, in a manner similar to existing aperture synthesis with radio telescopes. In addition, the twin 8.4 m apertures of the Large Binocular Telescope (Mt. Graham, Arizona) are designed to sample all spatial frequencies up to 22.8 meters using interferometric imaging between the two telescopes, and provide a much larger field of view than other interferometers. Web site links for the major optical/infrared interferometers using arrays of smaller telescopes are provided at the end of the chapter.

Several advances in technology have played a major role in bringing optical/IR interferometers to fruition. Among these, and roughly in priority order, are systems integration methods, high readout speed low-noise detectors, laser metrology, AO (but only for big telescopes), fast path length and tip–tilt servo control, and single-mode optical fibers. Because of the Earth's rotation, the apparent position of an astronomical object is always changing. To track this motion in an interferometer requires a movable delay line to compensate for changing geometrical delays between wavefronts reaching any two telescopes. Precision at nanometer levels over distances as long as 100 m with optical elements moving at high speeds (>1 cm/s) make this a very challenging aspect of optical interferometry. Detector arrays need to be high-speed and low-noise, two parameters that do not usually go together. Commercially available intensified CCDs, avalanche silicon photodiodes, and near-infrared arrays using other semiconductors have all had a significant impact.

2.5.1 Phase closure

There are essentially two approaches to interferometry. One is aperture synthesis in which the distribution of brightness across a source is the Fourier transform of the complex visibility of the fringes from many baseline pairs observed sequentially in a Michelson stellar interferometer. This is essentially how the Very Large Array (VLA) of radio telescopes operates, and this method works because the interferometer is "phased". When the telescopes are not phased then atmospheric effects are problematic. Alternatively, there is the well-established radio technique of "closure phase" in which simultaneous phase measurements on a triangle of three baselines formed by three apertures yield a summed or closure phase which is independent of unknown

atmospheric effects. Baldwin *et al.* (1986, 1996) were the first to demonstrate optical closure phases, and this group has developed a prototype instrument known as the Cambridge Optical Aperture Synthesis Telescope (COAST). For optical detection, avalanche photodiodes are used, but these investigators use HgCdTe infrared array devices too. The same principles will be applied to the Magdalena Ridge Observatory (MRO) in New Mexico (see web site references).

2.6 SPACE TELESCOPES

Clearly, we must go to space to explore those parts of the electromagnetic spectrum that are simply not transmitted (or poorly transmitted) by the Earth's atmosphere. Much of the infrared and sub-millimeter, and all of the UV, X-ray, and gamma-ray regions require observations from space. In addition, going into space overcomes the degradation of image quality caused by the atmosphere, either directly through turbulence or indirectly because of reduced signal-to-noise ratios due to high backgrounds. One of the strongest proponents of space observatories was the eminent Princeton astronomer Lyman Spitzer Jr. (1914–1997). In 1962 he led a program to design an observatory that would orbit the Earth and study ultraviolet light. This observatory became NASA's successful Copernicus satellite which operated from 1972 to 1981. In 1965, the National Academy of Sciences established a committee to define the scientific objectives for a proposed "large" space telescope. Spitzer was chosen to head this group and expended considerable effort trying to convince the astronomical community and Congress. The first step in achieving Spitzer's dream came with the launch of the Orbiting Astronomical Observatory in 1968. He continued his efforts, and finally in 1975 NASA and the European Space Agency (ESA) began the joint development of what would eventually become the Hubble Space Telescope (HST). Launched in 1990, the Hubble Space Telescope formed one of NASA's four "great observatories". The other three were eventually known as the Compton Gamma Ray Observatory (CGRO) launched in 1991, the Chandra X-ray Observatory (CXRO) launched in 1999, and the Space Infrared Telescope Facility (SIRTF) which was launched in 2003 and then renamed in orbit as the Spitzer Space Telescope (SST) in honor of Lyman Spitzer. Of course, many other extremely successful satellites have greatly expanded our view of the universe across the entire electromagnetic spectrum. We will mention several important missions and their detectors in Sections 11.7.3, 12.2–12.4, 13.3, and 14.5.

At launch, the HST had the following instruments: Wide-Field/Planetary Camera (WF/PC), Faint Object Camera (FOC), Faint Object Spectrograph (FOS), Goddard High Resolution Spectrometer (GHRS), High Speed Photometer (HSP), and the Fine Guidance Sensors (FGSs). On the first Space Shuttle servicing mission (SM1) the HSP was removed and replaced with COSTAR to provide optical correction for the spherical aberration of the primary mirror discovered shortly after launch (Figure 2.15). Because of its location, the WF/PC could not be corrected using COSTAR and so it was replaced with WF/PC2. On SM2, the FOS and GHRS were replaced with the Space Telescope Imaging Spectrograph (STIS) an ultraviolet

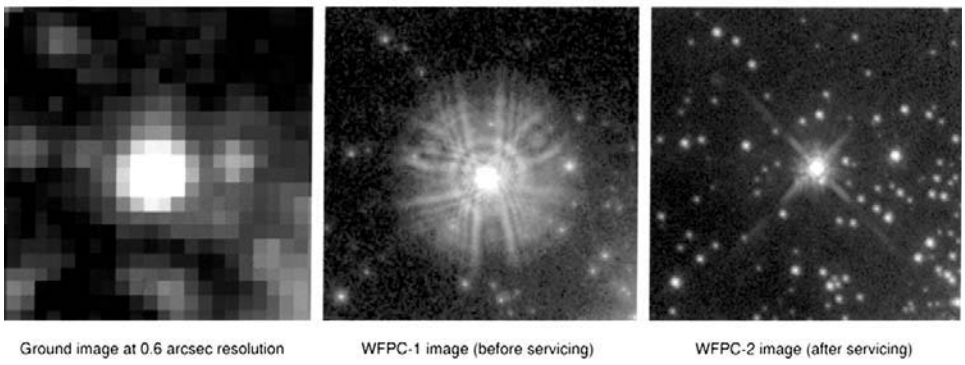


Figure 2.15. The panel shows a ground-based seeing-limited image of a star on the left, and then images from the 2.5 m Hubble Space Telescope before and after correction for the spherical aberration of the primary mirror. Credit: NASA.

spectrometer, and the Near-Infrared Camera and Multi-Object Spectrometer (NICMOS). On the third service mission, FOC was replaced with the Advanced Camera for Surveys (ACS), a new UV/visible camera. STIS, NICMOS, and ACS all had built-in correction for the primary mirror's spherical aberration, and thus COSTAR remained dormant after SM3. When the final Hubble service mission (SM4) occurs, WF/PC2 will be replaced with WF/PC3 and COSTAR will be replaced with COS, the Cosmic Origins Spectrograph. At the time of writing, STIS is not functioning and part of ACS is not working. It is hoped that these will be repaired during the final mission.

The HST is a 2.5 m diameter telescope, and therefore its diffraction limit (λ/D) is about $0.025''$ at 300 nm in the UV, but only $0.1''$ in the near infrared at 1.2 μm and $0.2''$ at 2.4 μm. Because the HST primary mirror is maintained at about 20°C it is too warm to use as a true infrared telescope. Also, the HST is in a relatively low orbit, about 320 nautical miles or 375 statute miles (600 km) above the surface, in order to be accessible to the Space Shuttle, which means that the Earth with a radius of more than ten times this distance provides an enormous nearby heat source and obstruction. Many future missions, such as the James Webb Space Telescope (JWST) will be placed at the L2 Lagrange point which is on the opposite side of Earth from the Sun at a distance of 1.5 million km.

2.7 SUMMARY

The atmosphere is only transparent to a fraction of the electromagnetic spectrum, and even in those windows of visibility there is both attenuation and emission. The best image quality requires the best observing sites, but even at the many high mountaintop observatories now in operation, turbulence in the Earth's atmosphere provides a stumbling block to obtaining diffraction-limited performance. The origin of seeing is explained using Kolmogorov turbulence theory, and the Fried parameter,

r-naught, is established to define seeing-limited resolution. To achieve the angular resolution improvement as well as the light-gathering boost from building very large telescopes we must either place observatories in space or develop technology to remove the effects of atmospheric turbulence. Both of these options are now available. Space telescopes and a range of electronic imaging devices have enabled us to picture the universe at all wavelengths from gamma rays to radio waves. Equally significant, the development of laser guide star adaptive optics has allowed the world's largest ground-based telescopes to reach their diffraction-limited resolution as if they were in space. Multiple telescopes in an optical/infrared interferometer array are now capable of even greater angular resolution.

2.8 EXERCISES

1 What is the diffraction-limited angular resolution of two 10 m telescopes separated by 85 m? What is the light-gathering power of the combination? Why is it more important to combine the two telescopes by interferometry rather than simply to "add" their light?

2 Estimate the angular resolution for a diffraction-limited telescope with a collecting aperture of 10 m diameter operating at a wavelength of 2 μm in the near infrared. What is the corresponding linear dimension in miles at the distance of Alpha Centauri (4.2 lightyears)? Suppose your telescope formed an interferometer with a baseline of 100 m, what would the resolution be now? Compare these numbers to distances within the solar system. Are you impressed? [Assume 1 lt-yr $\sim 6 \times 10^{12}$ miles.]

3 Explain the terms Fried parameter and isoplanatic patch? What is the order of magnitude of the Fried parameter for an 8 m telescope with seeing of 0.75″ at (a) visible wavelengths (5,000 Å) and (b) near-infrared wavelengths (2.2 μm)?

4 (a) Draw a simplified diagram of an adaptive optics system. Show the correct optical location of the deformable mirror. (b) Explain the role of a Shack–Hartmann wavefront sensor. (c) What is meant by the Strehl ratio?

5 What is the expected value of the Fried parameter at a wavelength of 500 nm if the observed seeing is 0.5 arcsec? What is the corresponding value of r_0 at 1.65 microns in the infrared assuming Kolmogorov turbulence? For a 10 m telescope, how many sub-apertures does this imply for infrared observations?

6 Determine the size of the isoplanatic patch and the isokinetic patch at the 0.5-micron wavelength for the case in Problem 5 assuming a turbulent layer at 5 km above the telescope.

2.9 REFERENCES

Babcock, H.W. (1953) The possibility of compensating astronomical seeing, *Publ. Astron. Soc. Pacific*, **65**, 229–236.

Babcock, H.W. (1990) Adaptive optics revisited, *Science*, **249**, 253–257.

Baldwin, J.E.; Haniff, C.A.; Mackay, C.D.; and Warner, P.J. (1986), Closure phase in high-resolution optical imaging, *Nature*, **320**, 595–597.

Baldwin, J.E.; Beckett, M.G.; Boysen, R.C.; Burns, D.; Buscher, D.F.; Cox, G.C.; Haniff, C.A.; Mackay, C.D.; Nightingale, N.S.; Rogers, J. *et al.* (1996), The first images from an optical aperture synthesis array: Mapping of Capella with COAST at two epochs, *Astron. Astrophys.*, **306**, L13–L16.

Beckers, J.M. (1993) Adaptive optics for astronomy: Principles, performance, and applications, *Ann. Rev. Astron. & Astrophys.*, **31**, 13–62.

Beckers, J.M. (1996) Techniques for high angular resolution astronomical imaging, in: J.M. Rodriguez-Espinoza (ed.), *Instrumentation for Large Telescopes*, Cambridge University Press, Cambridge, U.K.

Conan, R.; Ziad, A.; Borgnino, J.; Martin, F.; and Tokovinin, A.A. (2000) Measurements of the wavefront outer scale at Paranal: Influence of this parameter in interferometry, *Proc. SPIE*, **4006**, 963–973.

Foy, R.; and Labeyrie, A. (1985) Feasibility of adaptive telescope with laser probe, *Astron. & Astrophys*, **152**, L29–L31.

Fried, D.L. (1965) Statistics of a geometric representation of wavefront distortion, *J. Opt. Soc. Am.*, **55**, 1427–1435.

Fried, D.L. (1966) Limiting resolution looking down through the atmosphere, *J. Opt. Soc. Am.*, **56**, 1380–1384.

Gibson, J.H. (n.d.) UVB radiation definition and characteristics, downloaded from the UV-B Monitoring and Research Program web site at *http://uvb.nrel.colostate.edu/*

Hanbury Brown, R.; and Twiss, R.Q. (1956) A test of a new type of stellar interferometer on Sirius, *Nature*, **178**, 1046–1048.

Happer, W.; MacDonald, G.J.; Max, C.E.; and Dyson, F.J. (1994) Atmospheric-turbulence compensation by resonant optical backscattering from the sodium layer in the upper atmosphere, *J. Opt. Soc. Am. A*, **11**, 263–276.

Hardy, J.W. (1978) Adaptive optics a new technology for the control of light, *Proc. IEEE*, **66**, 651–697.

Hardy, J.W. (1998) *Adaptive Optics for Astronomical Telescopes*, Oxford University Press, Oxford, U.K.

Jenkins, F.A.; and White, H.E. (1957) *Fundamentals of Optics*, third edition, McGraw-Hill, New York.

Labeyrie, A. (1975) Interference fringes obtained on Vega with two optical telescopes, *Astrophys. J.*, **196**, L71–L75.

Lawson, P.R. (ed.) (2000) *Principles of Long Baseline Stellar Interferometry*, NASA-JPL Publication 00-009.

Maihara, T.; Iwamuro, F.; Yamashita, T.; Hall, D.N.B.; Cowie, L.L.; Tokunaga, A.T.; and Pickles, A. (1993) Observations of the OH airglow emission, *Publ. Astron. Soc. Pacific*, **105**, 940–944.

McAlister, H.A.; ten Brummelaar, T.A.; Gies, D.R.; Huang, W.; Bagnuolo, W.G., Jr.; Shure, M.A.; Sturmann, J.; Sturmann, L.; Turner, N.H.; Taylor, S.F. *et al.* (2005) First results

from the CHARA array. I. An interferometric and spectroscopic study of the fast rotator α Leonis (Regulus), *Astrophys. J.*, **628**, 439–452.

Meinel, A.B. (1950) Identification of the 6560 A emission in the spectrum of the night sky, *Astrophys. J.*, **111**, 433–434.

Monnier, J.D. (2003) Optical interferometry in astronomy, *Reports on Progress in Physics*, **66**, 789–857. (arXiv:astro-ph/0307036)

Noll, R.J. (1976) Zernike polynomials and atmospheric turbulence, *J. Opt. Soc. Am.*, **66**, 207–211.

Oliva, E.; and Origlia, L. (1992) The OH airglow spectrum: A calibration source for infrared spectrometers, *Astron. & Astrophys.*, **254**, 466–471.

Olivier, S.S.; An, J.; Avicola, K.; Bissinger, H.D.; Brase, J.M.; Friedman, H.W.; Gavel, D.T.; Max, C.E.; Salmon, T.; and Waltjen, K.E. (1995), Performance of laser guide star adaptive optics at Lick Observatory, in: Robert K. Tyson and Robert Q. Fugate (eds.), *Adaptive Optical Systems and Applications*, Proc. SPIE Vol. 2534, 26–37.

Roddier, F. (1981) The effects of atmospheric turbulence in optical astronomy, *Progress in Optics*, **19**, 281–377.

Roddier, F. (1988) Curvature sensing and compensation: A new concept in adaptive optics, *Applied Optics*, **27**, 1223–1225.

Rousset, G.; Fontanella, J.C.; Kern, P.; Gigan, P.; and Rigaut, F. (1990) First diffraction-limited astronomical images with adaptive optics, *Astron. & Astrophys.*, **230**, L29–L32.

Thompson, L.A. (1994) Adaptive optics in astronomy, *Physics Today*, **47**(12), 24–31.

Tyson, R.K. (1997) *Principles of Adaptive Optics*, second edition, Academic Press, Boston.

Wizinowich, P.; Acton. D.S.; Shelton, C.; Stomski, P.; Gathright, J.; Ho, K.; Lupton, W.; Tsubota, K.; Lai, O.; Max, C. *et al.* (2000), First light adaptive optics images from the Keck II telescope: A new era of high angular resolution imagery, *Pub. Astron. Soc. Pacific*, **112**, 315–319.

Wizinowich, P.; Le Mignant, D.; Bouchez, A.H.; Campbell, R.D.; Chin, J.C.Y.; Contos, A.R.; van Dam, M.A.; Hartman, S.K.; Johansson, E.M.; Lafton, R.E. *et al.* (2006), The W.M. Keck Observatory Laser Guide Star Adaptive Optics System: Overview, *Pub. Astron. Soc. Pacific*, **118**, 297–309.

Web sites

CHARA: *http://www.chara.gsu.edu/CHARA/*

Keck: *http://www.keckobservatory.org/*

KI: *http://planetquest.jpl.nasa.gov/Keck/keck_index.cfm*

Lick AO. Images of various Zernike modes as applied to the DM in the Lick AO system: *http://mthamilton.ucolick.org/techdocs/instruments/AO/ao_gallery.html*

Michelson Science Center: *http://msc.caltech.edu/*

MRO: *http://www.mro.nmt.edu/*

Optical Long Baseline Interferometry. This site is edited by Peter Lawson and has many links and references: *http://olbin.jpl.nasa.gov/intro/*

VLTI: *http://www.eso.org/projects/vlti/*

3

Telescopes

The first element of any astronomical imaging system is the telescope. Telescopes are of course the means by which the light from distant objects is collected and focused, but telescopes must do more than gather light. Providing excellent telescope optics, good tracking, and minimum air turbulence in the telescope dome are important steps in obtaining good images. Following a brief historical review of telescope development, we consider some basic optical properties and their applications to telescope design.

3.1 HISTORICAL DEVELOPMENT

By the end of the 13th century in Europe, artisans in glass-making centers like Venice and Florence had already found practical techniques for grinding and polishing glass relatively cheaply and easily. Moreover, people of that time were aware that the condition known today as *presbyopia*, in which the aging eye can no longer focus on something held at a comfortable distance, could be helped with a simple magnifying glass. But two smaller disks of glass, convex on both sides and supported in a frame were more convenient. Because these small disks were shaped like lentils, thicker in the middle, they became known as "lentils of glass" or (from the Latin) *lenses*. Concave lenses (inward-curving) that correct for *myopia* (near-sightedness) were made in Italy in the middle of the 15th century. It was not until the beginning of the 17th century, however, that these two types of lenses were combined to make a telescope. The earliest documented record is the 1608 patent application in the Netherlands by Hans Lippershey (*c*. 1570–1619), a German-born Dutch citizen, of a device with a convex and a concave lens in a tube with a magnification of about 3×. On learning of this device Galileo Galilei (1564–1642) made his own version of the telescope in the summer of 1609 and quickly increased the magnification to 8× and then 20×. The limiting factor of Galileo's telescope was its small field of view of about

15 arcminutes which meant that only a quarter of the full Moon could be seen. Galileo's telescope produced an upright image. From the early Galilean telescope of 1.52 m–1.83 m (5 ft–6 ft) in length, astronomical telescopes attained lengths of 4.57 m–6.10 m (15 ft–20 ft) by the middle of the 17th century. Typical of this time is the telescope made by Christiaan Huygens (1629–1695), in 1656. It was 7 meters (23 ft) long; its objective lens had an aperture of about 10 cm (~4 inches), it magnified about 100×, and its field of view was 17 arcminutes.

In a spherical-shaped lens, rays parallel to the optical axis (through the center) fail to converge at one point. Those farther from the optical axis come to a focus closer to the lens than those nearer the optical axis. This effect is called spherical aberration. To eliminate spherical aberration the lens curvatures must be either plane on one side and hyperbolic on the other, or spherical on one side and elliptical on the other. Fabrication of such shapes was beyond the technology of the time. In addition, Sir Isaac Newton (1672) showed that white light is a mixture of colored light and that every color had its own degree of refraction. Consequently, any curved lens will decompose white light into the colors of the spectrum, each of which will come to a focus at a different point on the optical axis. This effect, which became known as chromatic aberration, causes the image of a star to be surrounded by circles of different colors. Thus, lenses had limitations for astronomy, but telescopes with long focal lengths helped to reduce both of these effects.

In 1662, at the age of 24, the Scottish mathematician and astronomer James Gregory (1638–1675) wrote a treatise entitled *Optica Promota* describing the concept of a "reflecting" telescope made from two concave mirrors, one parabolic and the other ellipsoidal. Meanwhile, working independently in England, Newton had constructed the first reflecting telescope using a spherical mirror, a version of which was presented to the Royal Society in London in 1672. Gregory's more difficult design was eventually built successfully by Robert Hooke (1635–1703) and demonstrated in 1673. Around the same time Huygens was made aware of the idea of a similar reflective design to Gregory's, now attributed to an obscure monk named Laurent Cassegrain (1629–1693). Newton cast a 2-inch mirror blank of speculum metal (basically copper with some tin) and ground it into a spherical shape. He placed the mirror at the bottom of a tube and caught the reflected rays on a small flat secondary mirror placed at 45° near the top of the tube which reflected the image into a convex lens outside the tube for easy viewing by eye. When this instrument was shown to the Royal Society it caused a sensation; it was the first working reflecting telescope. Unfortunately, others were unable to grind mirrors of regular curvature, and to add to the problem the mirror tarnished easily and had to be re-polished every few months. Consequently, the reflecting telescope remained a curiosity for decades. By about 1723, however, John Hadley (1682–1744) the English inventor of the octant, a precursor to the sextant, and others had perfected better polishing techniques and the first parabolic version of the Newtonian telescope was made. By the middle of the 18th century many reflecting telescopes with primary mirrors up to six inches in diameter had been produced. James Short (1710–1769) is said to have made thousands of parabolic and elliptic mirrors around 1740. It was found that for large focal ratios, $f/10$ or more, the difference between spherical and paraboloidal mirrors

was negligible in the performance of the telescope. In the latter half of the 18th century large reflecting telescopes with parabolic ground mirrors came into their own. Sir William Herschel (1738–1822) built a reflector with a mirror diameter of 1.22 m (4 ft) and a 12.2 m (40 ft) focal length which he used to discover moons of Saturn, but by all accounts it was not as easy to use as his 6.1 m (20 ft) long 0.475 m (18.7-inch) reflector. Nevertheless, it remained the largest telescope for over 50 years until Lord Rosse's 1.83 m (72-inch) reflector was built in 1845. To tackle the serious problem of rapid tarnishing in metal mirrors, Herschel always had a spare ready to exchange when a mirror required re-polishing. Readers are referred to Henry King's book, *The History of the Telescope* (Dover, 1979) for more historical details.

In the 400 years since Galileo turned his telescope towards the sky and ushered in a revolution, telescope designs have evolved. In classical times the size of a telescope was characterized by its focal length, but modern astronomical telescopes are always identified by the diameter of the primary collecting aperture, or more generally by the diameter of the equivalent circle with the same collecting surface area. The largest telescope used by Galileo had a *diameter* of 4.4 cm (1.75 inches) but as of 2008 there were ten general-purpose optical telescopes with effective diameters greater than 800 cm (315 inches or 26.25 ft) and several others of special design. In a thorough compilation of astronomical telescopes through the ages, Racine (2004) finds that the doubling time for aperture size was about 50 years up until about 1950, but since then telescopes have been doubling in size at a rate that is at least twice as fast. Figure 3.1 is

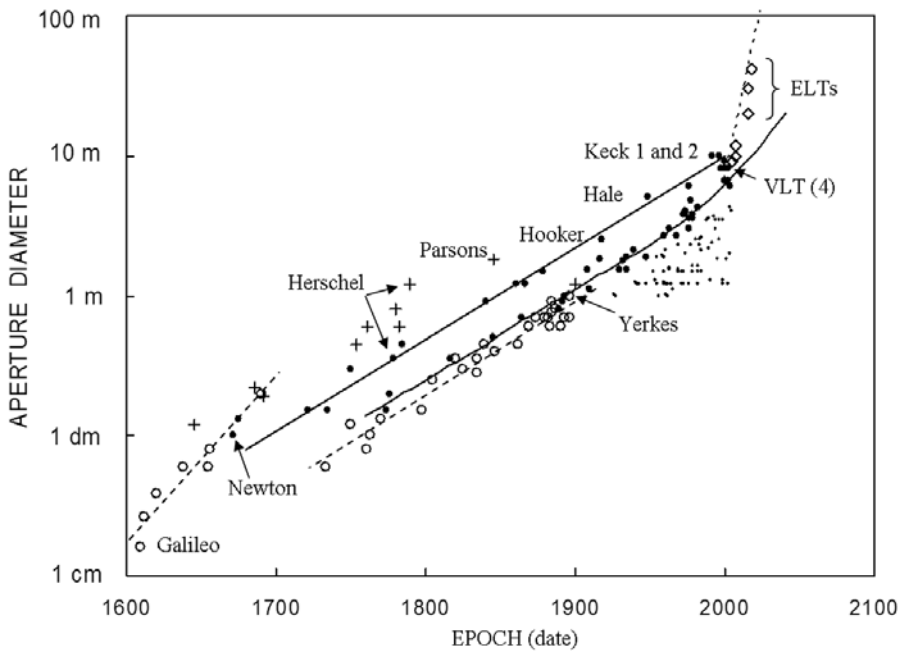


Figure 3.1. The growth of aperture size with time is plotted from the invention of the telescope to present day. Credit: René Racine.

Table 3.1. Telescopes larger than 3.5 m in the pre-Keck era.

Telescope name (date opened)/Site	*Primary* (m)	*f/ratio*	*Mounting*
Bol'shoi Teleskop Azimutal'nyi (1976) Mt. Pastukhov (Russia)	6.0	$f/4.0$	Altitude-azimuth
George Ellery Hale Telescope (1949) Mt. Palomar (California, U.S.A.)	5.08	$f/3.3$	Equatorial
Multiple Mirror Telescope (1977) Mt. Hopkins (Arizona, U.S.A.)	4.5	$f/2.7$	Altitude-azimuth
William Herschel Telescope (1982) Roque de los Muchachos (Canary Islands)	4.2	$f/2$	Altitude-azimuth
The Blanco Telescope (1968) Cerro Tololo (Chile)	4.0	$f/2.8$	Equatorial
Anglo-Australian Telescope (1974) Siding Spring (NSW, Australia)	3.9	$f/3.3$	Equatorial
Nicholas Mayall Telescope (1966) Kitt Peak (Arizona, U.S.A.)	3.8	$f/2.8$	Equatorial
United Kingdom Infrared Telescope (1979) Mauna Kea (Hawaii)	3.8 (thin)	$f/2.5$	Equatorial
Canada–France–Hawaii Telescope (1979) Mauna Kea (Hawaii)	3.6	$f/3.8$	Equatorial
European Southern Observatory (1976) Cerro La Silla (Chile)	3.6	$f/3.0$	Equatorial

an updated illustration from Racine's paper and shows the historical growth of all types of telescopes of modest size. Table 3.1 lists the world's largest telescopes (diameters > 3.5 m) prior to 1993.

Interestingly, after construction of the 5.08 m (200-inch) Hale Telescope on Palomar Mountain in California in 1949, the trend was for smaller telescopes in what has become known by the imprecise term "4-meter class"; in this definition all telescopes with primary mirror diameters between 3.6 m and 4.5 m are lumped together. Exceptions were the Soviet BTA and the Smithsonian MMT, both of which were unique in their design, and both of which may have been somewhat ahead of their time.

In March of 1993, the largest telescope in the world went into operation and a new era in astronomy was born. It was the first of a pair of 10-meter telescopes funded by the W. M. Keck Foundation for the California Institute of Technology (Caltech) and the University of California (UC) that employed the unique "segmented mirror"

Figure 3.2. The twin domes of the W.M. Keck Observatory on the summit of Mauna Kea. Each dome encloses a telescope with a segmented mirror having an effective aperture of about 10 meters. The domes are 85 m apart. Photo credit: author.

concept championed by Jerry Nelson of the University of California (Nelson, 1995). The second telescope was inaugurated in May 1996 (Figure 3.2). At that time, at least eight other optical/infrared telescopes with collecting apertures larger than 6.5 meters in diameter, and employing different technologies, were also under construction and several more were being contemplated. What drove this remarkable development?

Following the introduction and growth of CCDs, more and more area on the sky could be digitally imaged to deeper levels. In addition, the efficiency of spectroscopy had already been improved by the use of multi-slit devices and optical fibers to observe many objects simultaneously. Once the quantum limits of sensitivity in detectors and instruments has been reached, the only way to gain large factors in efficiency is to construct even larger ground-based telescopes and to develop methods for counteracting the image-blurring effects of turbulence in the Earth's atmosphere. There are essentially three fundamental issues:

(1) how to achieve a very large collecting aperture of the required optical performance
(2) how to support and control in the optimum way such a potentially very heavy mechanical structure
(3) how to enclose a very large telescope in a cost-effective way with negligible degradation on image quality due to vibration, air disturbance, or inadequate environmental protection (wind, dust).

Moreover, new telescopes must be designed to capitalize on the very best seeing conditions at the world's best sites, and must be designed with remote control in

mind. Roger Angel (University of Arizona), one of the pioneers in this field, summed up the situation (Angel, 1989),

> "the problems of building the new generation of telescopes are compounded because they must not only be bigger, but also must give sharper images than their predecessors."

Basically, it all comes down to how the mirrors are made and supported. There are three categories of new technology:

- *segmented mirrors*, smaller monolithic disks of thin polished glass are used to form the surface of an efficient rigid "backing" structure of steel or carbon fiber. Position actuators are still required to make the attachment and to correct for thermal and gravitational effects in the backing frame, but the time scale for such corrections is slow. Each segment is individually supported and global changes are sensed at the gaps between segments.
- *meniscus mirrors*, large monolithic disks of solid glass which are so thin that it must be accepted that they will be flexible and therefore they must be actively controlled to maintain the required shape during operation. Bending by unpredictable forces such as wind gusts requires a rapid servo system of precise force actuators.
- *honeycomb mirrors*, thick mirrors are constructed but large pockets of mass are removed from the back to make the mirror lightweight yet very stiff. The method involves a mold and a spinning furnace to form a concave parabolic front face while the ribs and back plate are formed by glass flowing down between the gaps in the mold.

Producing the mirror blank is only the first step. Larger and faster primary mirrors require new polishing methods to achieve their final figure. It is not the deep curvature itself that is the problem—but the asphericity that results in different curvature from place to place and between tangential and radial directions. A rigid pitch lap cannot accommodate the changes of curvature needed as it is stroked over a strongly aspheric surface, unless it is moved only in the tangential direction, but to do this will result in circular grooves or zones. Conventional polishing of 4-meter mirrors is typically limited to focal ratios of $f/2$ or slower. To polish the primary mirror blank for the 4.2 m William Herschel Telescope (Canary Islands), David Brown of Grubb Parsons, U.K. (Brown, 1986) used a lap which changed shape as it moved. His method was based on the fact that, when a full-sized lap is used to make polishing strokes across a paraboloid, the distortion required to maintain contact is that of coma. For the same reason, the off-axis aberration of a paraboloid is also coma. The general principle of making the lap change shape as it moves is called "stressed lap" polishing.

Table 3.2. The current generation of telescopes with $D > 6.5$ m.

Telescope and date	*Primary* (m)	*Mirror technology*	*Location*
Keck I, II (1993, 1996) (CARA)	10	Hexagonal segments of Zerodur	Mauna Kea, Hawaii
Hobby–Eberly Telescope (2000)	9.2	Hexagonal segments Spherical primary	Mt. Fowlkes, Texas
South African Large Telescope (2004)	9.2	Hexagonal segments Spherical primary	Sutherland, SA
LBT (2005, 2008) (former Columbus)	2×8.4	Borosilicate honeycomb	Mt. Graham, Arizona
Subaru (2000) (Japan)	8.2	Thin meniscus	Mauna Kea, Hawaii
VLT1, 2, 3, 4 (1998, 2000, 2001, 2002) (ESO)	4×8.2	Thin meniscus	Cerro Paranal, Chile
Gemini N, S (2000, 2002) (GTP)	2×8.0	Thin meniscus	Mauna Kea, Hawaii Cerro Pachón, Chile
Magellan I, II (2002, 2003) (OCIW)	6.5	Borosilicate honeycomb	Las Campanas, Chile
MMT upgrade (2002) (SI/UA)	6.5	Borosilicate honeycomb	Mt. Hopkins, Arizona

Each of the three methods described has been applied to build the current generation of very large telescopes for ground-based astronomy at optical and infrared wavelengths. Table 3.2 lists all telescopes operational or nearly so as of March 2008 with primary mirror diameters larger than about 6.5 meters.

The 6.5 m telescope called the MMT in this list is in the same enclosure as the original Multiple-Mirror Telescope and is known as the Monolithic Mirror Telescope in order to preserve the well-known acronym of the observatory. At the time of writing, a copy of the Keck Telescope with very slightly larger segments is nearing completion on La Palma for Spain and partners. It is called the Gran Telescopio Canarias (GTC) and will have an effective diameter of about 10.4 m when it begins science operations. The GTC will then top the list in Table 3.2. The unique Large Binocular Telescope (LBT) is specifically designed to combine the light from two side-by-side 8.4 m primaries giving it an effective aperture of almost 12 m. Both primary mirrors are installed, and work is continuing on combining the beams. Before looking at the technology associated with each of the three telescope types it is valuable to review the fundamental issues of telescope design in general.

3.2 TELESCOPE DESIGNS

Telescopes fall into one of three basic types: refractive (dioptric, using lenses); reflective (catoptric, using mirrors); hybrid (catadioptric, using a combination of mirrors and lenses). Hybrid designs are frequently the most popular for amateur astronomy because of their compact design, but all large professional telescopes are reflectors.

3.2.1 Basic optical properties

In this section we summarize and apply useful relationships that are derived in any good book on optics. It is well-known that when a ray of light strikes the boundary between two different transparent materials, such as air and glass, it is divided into a reflected ray (which is why you can see a faint reflection of yourself in a transparent window) and a transmitted ray whose direction in the glass is different from that in the air for oblique incidence; this ray is said to be refracted. A ray represents the direction of flow of the energy in an electromagnetic wave and is perpendicular to the wavefront. At the point of contact with the surface the ray makes an angle of incidence (θ) with a line that is perpendicular or "normal" to the surface at that point. The law of reflection states that the reflected ray lies in the same plane as the incident ray, and the angle of reflection (φ) relative to the normal is equal to the angle of incidence. Therefore

$$\theta = \varphi \tag{3.1}$$

On the other hand, the law of refraction states that the transmitted ray lies in the same plane as the incident ray, but moves in a different (deviated) direction such that the *sine* of the angle of refraction (θ') divided by the *sine* of the angle of incidence (relative to the normal at that point) is a constant equal to the ratio of the refractive indices (n/n') in the two materials. This result is commonly written as

$$n \sin\theta = n' \sin\theta' \tag{3.2}$$

The law of refraction is also called Snell's law after the Dutch scientist Willebrord Snell (1580–1626) who stated a version using the ratio of cosecants, but it was the French scientist René Descartes (1596–1650) who first used the ratio of sines.

These laws are consistent with the wave nature of light, and the meaning of the refractive index (n) is that it is a measure of the speed of light in the medium relative to that in vacuum (i.e., $n = c/v$). Also, $n = \sqrt{(\varepsilon_r \mu_r)}$ where ε_r and μ_r are the relative permittivity and permeability of the material; $\mu_r \sim 1$ for most materials at optical wavelengths. In most cases $n > 1$, but we will see in Section 12.1 that this is not true for X-ray telescopes. The change of direction of the ray (or wavefront) for oblique angles of incidence can be understood by realizing that the first part of the wavefront to intercept the surface is slowed down relative to an adjacent front which travels farther in the same time and so the entire wavefront seems to progressively change direction. The distance between crests inside the material is now λ/n. As shown by Newton, refraction separates white light into its component colors, which implies that

the refractive index varies with wavelength. The angular divergence of light of different wavelengths is called "dispersion" and the dispersive power of a given material is expressed as the difference in the refractive indices at two extreme wavelengths relative to the difference between a middle wavelength and vacuum; that is,

$$\text{dispersive power} = 1/V = (n_F - n_C)/(n_D - 1) \tag{3.3}$$

Here F, D, and C represent the wavelengths of the blue (486 nm), yellow (589 nm), and red (656 nm) Fraunhofer lines in the solar spectrum (now known to be due to Hβ, Na D, and Hα, respectively), but other values can be chosen for the range of interest. Take as a simple example the measured indices for crown glass, $n_F = 1.53303$, $n_D = 1.52704$, and $n_C = 1.52441$. Dispersion of the F and C rays is given by $(n_F - n_C) = 0.00862$ while the deviation of the D ray (relative to air) is given by $(n_D - 1) = 0.52704$, thus the reciprocal of the dispersive power is $V = 0.52704/0.00862 = 61.13$. Values of V are typically in the range 30–70 and a plot of V against n for a great many substances constitutes a glass table. A web site for Schott glass is given in the references.

When light travels a distance d in a medium of refractive index n the "optical path" is the product nd. Given the definition of n, the optical path is the distance that the light would travel in vacuum in the same time that it travels the distance d in the medium. There is a general principle involving the optical path which applies for any type of variation of refractive index. First stated in an incomplete form by Pierre Fermat (1608–1665), the modern version of Fermat's principle is that the path taken by a light ray in going from one point to another through any set of media is such as to render its optical path equal, in the first approximation, to other paths closely adjacent to the actual one. In the language of calculus, the optical path must be a minimum, maximum, or stationary point (point of inflection) for the actual path.

Paraxial optics is the approximation in which angles (of reflection or refraction) are sufficiently small so that $\sin\theta = \theta$ (in radians) to first order. The term "first order" comes from the fact that $\sin\theta$ can be expressed as a sum of terms with odd powers of θ, such as θ, $a_3\theta^3$, $a_5\theta^5$ and so on, with diminishing values for the coefficients (a_n). In Section 6.3.2 we will see the importance of expanding the approximation to "third order" (θ^3). To see the effect of setting $\sin\theta = \theta$, take, for example, $\theta = 10^\circ$. The angle in radians is found from $10^\circ\pi/180^\circ = 0.174533$ and the sine of 10° is 0.173648, a difference of only 0.5%. Even at 30° the error is just under 5%. A lens with a focal length of ten times its diameter ($f/10$) subtends an angle of about 1/10 of a radian or about 5.7°, thus systems with large focal ratios (slower systems) meet the paraxial approximation. With this assumption it is possible to derive a number of well-known formulas for lenses and mirrors. Such derivations are given in all standard textbooks on optics, thus only a few useful results are given here for completeness.

The "thin lens" equation in air or vacuum:

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{s'}, \qquad m = -\frac{s'}{s} = \frac{h'}{h}, \qquad f/\text{number} = \frac{f}{D} \tag{3.4}$$

where f is the focal length; s and s' are the object and image distances, respectively; h and h' are the object and image "heights"; m is the lateral or transverse

magnification; and D is the clear aperture diameter of the lens. A lens is considered thin if its thickness is small in relation to its focal length. The distances s and s' are called conjugate points. Note that if the object distance is infinitely large ($1/s = 0$) then the conjugate image distance is just the focal length of the lens ($s' = f$). This equation assumes that the lens is in vacuum ($n = 1$) or to a first approximation in air ($n \approx 1.0003$). If that is not so then the expression can be generalized to include the refractive indices of the mediums. There are situations in the design and calibration of spectrometers in which it is essential to differentiate between air and vacuum indices. The longitudinal magnification (along the optical axis) is m^2. With rays traveling left to right through the lens, the sign convention is that s is positive if on the left of the lens and s' is positive if on the right, and the negative sign for m indicates that the image is inverted when s and s' are both positive.

The spherical mirror equation:

An equation similar to Equation (3.4) holds for spherical mirrors in the paraxial approximation, and in this case it is easy to see that $f = R/2$ where R is the radius of curvature of the mirror

$$\frac{1}{f} = \frac{1}{s} + \frac{1}{s'} = \frac{2}{R}, \qquad m = -\frac{s'}{s} = \frac{h'}{h} \tag{3.5}$$

Optical "power" and the lensmaker's formula:

This useful form relates the surface curvature and properties of the lens medium into a single factor

$$P = \frac{1}{f} = (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{t(n-1)}{nR_1R_2}\right] \tag{3.6}$$

where P is the power of the lens; R_1 and R_2 are the radius of curvature of the front surface and the back surface, respectively; and t is the central thickness of the lens. If t is less than 1/6 of the lens diameter, then the third term can be neglected. The sign conventions are that R is positive if the center of curvature is behind the lens and negative if it is in front of the lens (light travels left to right). Power is usually stated in a unit called the diopter (1/meter) so that a power of 5 diopters means a focal length of $1/5 = 0.2$ m (or 20 cm). Thick lenses can be handled by defining two principal planes where refraction occurs or by point-to-point ray tracing using computer programs. For thick lenses, or two thin lenses separated by a distance d, the powers are additive so that the overall power, or effective focal length, is given by

$$P = P_1 + P_2 - (d/n)P_1P_2 \tag{3.7}$$

Rays traveling parallel to the axis are brought to a focus when they refract through a thin convex lens, whereas rays traveling parallel to the axis diverge after refracting through a concave lens and appear to come from a focal point on the same side of the lens as the incident light. Image formation and the resultant magnification depend on the relative positions of the object and the focal points. When the image formed by the lens cannot be made visible on a screen because no light actually travels to or from

that location, but only appears to do so, such images are said to be "virtual" and a simple example occurs when an object is placed closer to a convex lens than its focal point. The light appears to come from a magnified, upright virtual image on the same side of the lens as the object.

Newton's equation:

$$x'x = f^2 \tag{3.8}$$

where the distances are now measured relative to the focal points: $x = s - f$ and $x' = s' - f$. This form is most useful for calculating the amount of re-focus required.

Angular magnification:

This is defined in terms of the slope angles of the rays rather than the object and image heights. If the incident ray makes an angle u with respect to the optical axis (slope = tan u) and the refracted ray makes an angle u' with the optical axis, then the angular magnification is

$$M = \tan u'/\tan u = s/s' = h/h' \tag{3.9}$$

The Lagrange Invariant:

The relation for angular magnification given above can be generalized to account for different refractive indices in object and image space, respectively, which then yields the result $nh \tan u = n'h' \tan u'$, or in the paraxial limit ($\tan u = u$)

$$hnu = h'n'u' \tag{3.10}$$

where h and h' are the object and image heights; n and n' are the refractive indices in object and image space, respectively (usually $n = n' = 1$); and u and u' are the (small) angles with respect to the optical axis of the same ray in object space and in image space. The quantity ($nh \tan u$) is constant (invariant) in a system with any number of refracting or reflective surfaces. Astronomers use this in its area-solid angle ($A\Omega$) form, also known as *étendue* or throughput (see Section 6.9, Special topic). Because the total flux collected from a uniformly radiating source is proportional to ($A\Omega$), it follows that the Lagrange Invariant is a consequence of the conservation of energy.

3.2.2 The astronomical (lens) telescope

The principle of the classical astronomical telescope is illustrated in Figure 3.3 where the double-ended arrow marked O represents the long-focus (weak) objective lens and the smaller (strong) eyepiece lens is labeled E. For essentially infinite objects ($s \gg f$), Equation (3.4) shows that the objective forms an image at its focal point ($s' = f$), which in turn becomes the object for the eyepiece lens. If the eyepiece is moved along the axis such that the real image formed by the objective lens coincides with the focal point of the eyepiece lens (separation of the lenses is then the sum of the focal lengths), then the emergent rays are parallel and the final (virtual) image is at

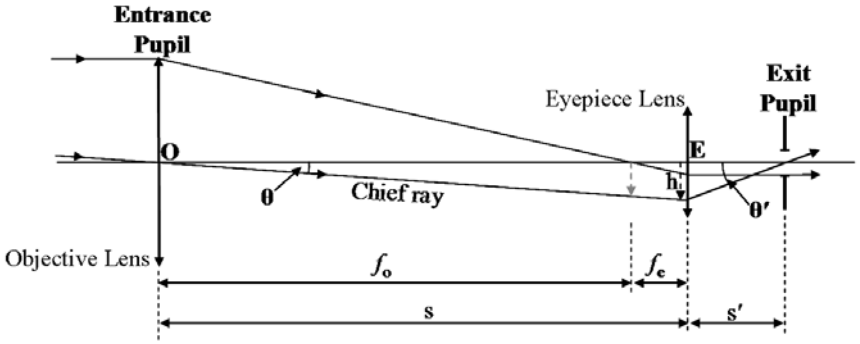


Figure 3.3. The principle of the astronomical telescope. The objective lens and the eyepiece lens are represented by vertical lines with double-ended arrows.

infinity. This form of the telescope is sometimes called Keplerian after Johannes Kepler (1571–1630). Note that the telescope is afocal (i.e., parallel rays in, parallel rays out) and the image is inverted. In Galileo's telescope the eyepiece lens is divergent (concave) and the lenses are separated by the difference (not the sum) of their focal lengths. Once again the system is afocal but the image in a Galilean telescope is upright, which is more convenient for terrestrial use. The objective lens defines the aperture or "stop" as no light beyond its edges is collected and it is said to form the "entrance pupil" for the system. Of course, the objective lens itself is a relatively nearby object for the eyepiece lens; therefore, another application of Equation (3.4) shows that an image of the entrance pupil is formed behind lens E, and this image becomes the "exit" pupil. These are useful terms that are employed frequently in telescope and instrument designs. One ray in the figure is labeled as a "chief ray" because it passed through the center of both the entrance and exit pupils. The distance from the eye lens to the exit pupil is called the "eye relief" and should be $\sim$8 mm.

To determine the magnifying power of a telescope of this type requires comparing the angle (θ) subtended by the distant object with that of the final image at the eye (θ'). By definition $M = \theta'/\theta$, but from the figure the height h is given by both $s \tan\theta$ and by $s' \tan\theta'$. Applying the thin lens equation (Equation (3.4)), and noting that $s = f_o + f_e$, yields expressions for $\tan\theta$ and $\tan\theta'$ which, using the small angle approximation ($\tan\theta \sim \theta$), gives $M = -f_o/f_e$. Therefore, the magnification of a classical astronomical telescope is just the ratio of the focal length of the objective lens to that of the eyepiece, with the negative sign implying that the image is inverted. Also, the diameters of the entrance and exit pupils are related to the focal lengths by the same ratio, $M = D_o/D_e$, where D_o is the diameter of the objective lens and D_e is the diameter of the exit pupil; the exit pupil is sometimes called the Ramsden circle for English instrument maker Jesse Ramsden (1735–1800). Many different forms of compound eyepiece lenses have been developed to correct aberrations and provide comfortable eye relief. With the advent of CCD imaging, most astronomical telescopes do not employ an eyepiece at all, but instead form images directly onto the detector pixels in the telescope's focal plane.

Because the refractive index changes with wavelength, a single lens forms a series of images, one for each color of light in the beam. Shorter wavelengths are bent more than longer wavelengths and therefore form a focus closer to the lens (i.e., the focal length f is a function of the wavelength). The horizontal distance along the optical axis between the different focal positions is called longitudinal chromatic aberration. But the lateral magnification (m) must vary as well because the image height (h') depends on s' which is controlled by f. The vertical difference in image height is called lateral chromatic aberration. The classical method for correcting both forms of chromatic aberration is to use two lenses of different materials (historically, crown glass and flint glass) in contact to make an "achromatic" doublet. Using the lens-maker's formula (Equation (3.6)) in the thin lens approximation ($t = 0$) for each lens and substituting for each P in the additive-power formula (Equation (3.7)) with the two lenses in contact ($d = 0$), we can write

$$P_D = P'_D + P''_D = (n'_D - 1)K' + (n''_D - 1)K'' \tag{3.11}$$

here the suffix is for the yellow D line of sodium; the prime and double prime refer to the crown and flint glasses, respectively; and the K symbols stand for $[1/R_1 - 1/R_2]$ for each lens. By writing a similar equation for the two extreme colors (F and C) and equating them to ensure that $P_F = P_C$, it follows that $K'/K'' = -(n''_F - n''_C)/(n'_F - n'_C)$, where the negative sign out front implies that one lens must be negative (concave). Taking the powers of each lens in Equation (3.11) separately, forming the ratio P'_D/P''_D, and substituting for K'/K'' yields the important result that $P'_D/P''_D = -V'/V''$ where these are the dispersion constants for the two glasses. A little more algebra allows the power of each lens in the achromatic doublet to be expressed in terms of the desired total power P_D or focal length $f_D(= 1/P_D)$:

$$P'_D = P_D[V'/(V' - V'')] \quad \text{and} \quad P''_D = -P_D[V''/(V' - V'')] \tag{3.12}$$

The steps required to create the achromatic doublet are therefore as follows: select the required focal length $f_D(= 1/P_D)$; select the best pair of glasses from the glass table on the basis of their dispersion constants V' and V''; use Equation (3.12) to calculate the powers of each lens; derive K' and K'', respectively, from each part of Equation (3.11) where P'_D and P''_D are defined; and, finally, the radii are found from the definition of K. To facilitate cementing of the convex and concave lenses, the second radius of the first lens should match the first radius of the second lens, $R''_1 = R'_2$, and it is convenient to use the same radius—but opposite curvature—on the entrance face of the convex lens, thus $R'_2 = -R'_1$. The radius of the last surface R''_2 is then adjusted to give the required power P''_D. Keeping the more curved crown glass element towards the incoming light reduces spherical aberration. Colors outside the corrected range can still cause a halo of color around a point source, referred to as the "secondary" spectrum. These problems, plus the sheer size and weight of large achromatic doublets reaching $\sim$1 m in diameter eventually caused the switch to mirrors.

3.2.3 Reflecting telescopes; the parabolic dish

When it became possible to make mirrors of high quality the reasons for their subsequent dominance in modern astronomy are clear. In addition to having a good surface finish a lens must be transparent and free of internal flaws that can cause scattering, while a mirror only needs to have a good surface finish. It is interesting to note, however, that a surface deviation of size d is doubled to $2d$ in the wavefront on reflection whereas in transmission the effect of the surface bump is only $(n-1)d$, which for $n = 1.5$ gives $0.5d$. In other words, mirrors have to be polished better in order not to disturb the wavefront. This is not a limitation today as illustrated by mirrors like that of the Subaru 8.2 m (27 ft) telescope which if scaled up by a factor of 10,000 to about 51 miles (85 km) in diameter would have no hills or valleys larger than 0.12 mm on average, which is about the thickness of an ordinary piece of paper. A lens can only be supported around its rim whereas a mirror can be supported across its back. Finally, a lens suffers chromatic aberration due to the wavelength dependence of its refractive index, and while it is possible to construct multiple lenses to reduce this effect, more lenses mean more light loss. A mirror is achromatic, but some metallic coatings absorb in the UV (Section 12.1).

Today, all large astronomical telescopes (from the UV to radio) are "reflectors" and use curved mirrors (Figure 3.4) to achieve light collection and focus. However, a careful examination of Figure 3.4 shows a potential problem with a spherical-shaped mirror as initially used by Newton. Unless the focal length is very long, so that the mirror's curvature is very slight, parallel rays from a distant (almost infinite) astronomical source will come to a different focus depending on how far they are from the center of the mirror. This is the familiar effect called spherical aberration already mentioned several times. Parallel incident rays at increasing lateral heights h cross the axis closer and closer to the mirror. At the paraxial focus point the image is blurred, but a minimum blurring, called the circle of least confusion, occurs slightly closer to the mirror. The diameter of the blur circle is given approximately in seconds of arc by $\beta = 206{,}265/128(F)^3$, where $F = f/D$; just drop the conversion factor 206,265 to go

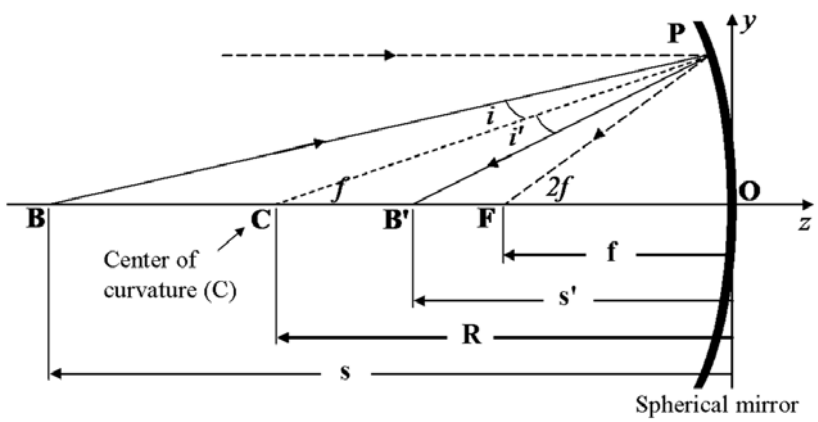


Figure 3.4. Reflection from a mirror with a spherical curvature.

back to radians. For a mirror with a focal ratio of $F = 15$, the blur is 0.48 seconds of arc, comparable with seeing, but for $F = 3$ the spherical aberration is one minute of arc!

Clearly, a spherical-shaped mirror is not the correct shape to compensate for the change of angle, but as shown by Gregory a conic section (ellipse, parabola, or hyperbola) will solve this problem. Conic sections are so-named because they are all derived from cross-sectional cuts through a cone. If you spin a bowl of water on a turntable, the natural shape taken up by the surface of the water is a parabola. We will return to the idea of spinning a liquid to create a parabolic surface. A fundamental property of a conic section shape is that any ray starting at one focus will form a perfect point image at the other. This behavior is called "stigmatism", not to be confused with "astigmatism" which is an aberration. For the parabola, one focus is at infinity and therefore rays from that focus are parallel to the axis and all such rays meet at the other focus.

Can a single mirror be used as a telescope? With some additional optics, many professional telescopes use the "prime" focus of the primary mirror for direct imaging. Really large telescopes like the 5 m Hale on Mt. Palomar have a prime focus "cage" that is so large and robust that observers once rode in it when doing photography before the days of automatic guiding. Among the current large telescopes the Subaru 8.2 m Telescope is unique with its large CCD camera at prime focus. Many 4 m class telescopes make excellent use of prime focus with both CCD cameras and infrared array cameras. In all cases, however, a "corrector" lens system is required in front of the CCD. We will discuss corrector lens systems later in this chapter.

In normal two-mirror telescopes, the "secondary" mirror can have a variety of surface shapes depending on the telescope design. A flat, plane mirror placed in the center of the tube to intercept the beam before the prime focus yields the well-known Newtonian design. Usually the plane mirror is at 45° to deflect the beam through a right angle and out the side of the tube, but it can also be perpendicular to the beam to send it back towards the primary; the 1.55 m (61-inch) U.S. Naval Observatory reflector at Flagstaff, Arizona is designed this way. Many large reflecting telescopes are of the Cassegrain design and contain a central hole in the primary mirror. The light reflected from the secondary passes through the Cassegrain hole, exits the telescope tube at the back of the primary mirror, and comes to a focus, the Cassegrain focus.

In the classical Cassegrain telescope the primary mirror is a paraboloid and the secondary mirror is a hyperboloid as illustrated in Figure 3.5. The hyperbolic secondary is placed to intercept the beam before it reaches the prime focus. Image quality is limited by the off-axis effect called "coma" in which images become more and more comet-like with small tails. An alternative design is the Gregorian (also shown in Figure 3.5) in which the converging rays from the primary are allowed to go through the prime focus and reflect from an elliptical secondary mirror back through the hole in the primary; the Gregorian has more field curvature than the classical Cassegrain. In either case, instruments can be attached at the focus behind the primary mirror. An important variation of this basic approach is called the

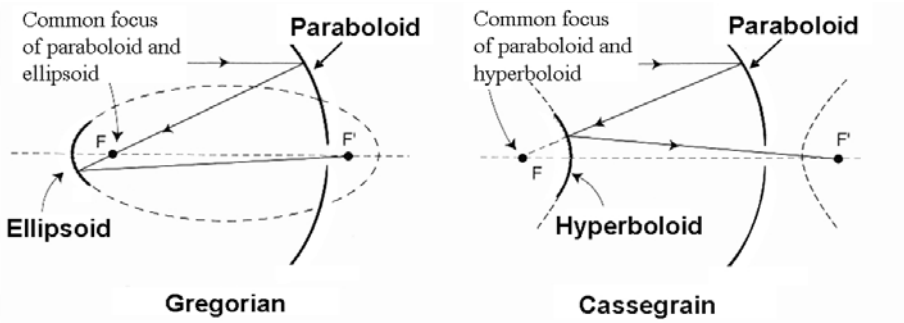


Figure 3.5. The principle behind the Gregorian and Cassegrain telescopes which use the properties of conic sections instead of spherical shapes.

Ritchey–Chrétien (or RC) design in which both the primary and secondary are hyberbolic surfaces. Invented by American astronomer George Willis Ritchey (1864–1945) and French astronomer Henri Chrétien (1879–1956), this design eliminates the effect of coma and thus gives good performance over a larger field of view.

The first successful Ritchey–Chrétien Telescope (or RCT) was a 0.5 m (20-inch) built in 1927. Hyperbolic curvatures are harder to test, but modern interference methods solve this concern. Ritchey was involved in the design of the 2.54 m (100-inch) Hooker Telescope and promoted the RC approach for the Palomar 5.08 m Telescope, but a disagreement between Hale and Ritchey, plus the higher risk, led to the choice of a paraboloid for the giant primary. RC designs have been used for the twin 10 m Keck Telescopes, the four 8.2 m VLTs, the twin 8 m Gemini telescopes, the 8.2 m Subaru, and the 10.4 m GTC. The Hubble Space Telescope is also a Ritchey–Chrétien design. Both of the 6.5 m Magellan telescopes employ the Gregorian design (Shectman, 2000). Two exceptions to the standard use of conic sections in large telescopes are the Hobby–Eberly Telescope (HET) and the South African Large Telescope (SALT) which both have spherical segmented primaries and are designed for a special application.

A third or "tertiary" mirror (usually a plane mirror) can be located to direct the light along an axis of rotation of the telescope. Depending on the type of telescope mounting used (see below), additional mirrors may be required, but the final focus is a "stationary" position, independent of where the telescope is pointing. For classical telescopes constructed on an equatorial mount, this focus is called the coudé (pronounced "koo-day"), and for telescopes using altitude-azimuth mounts, this is the Nasmyth (pronounced "nay-smith") focus, attributed to the Scottish engineer James Nasmyth (1808–1890) who also invented the steam-hammer! A change of secondary mirror is required with the coudé arrangement and the result is a very long focal length. In general, the Nasmyth focus does not require a change of secondary mirror, and thus the image scale is the same as at the Cassegrain focus.

Each focus is characterized by a "plate scale" which measures the number of seconds of arc on the sky corresponding to 1 mm at the focus of the telescope; the

longer the focal length (f), the smaller the number of seconds or arc per millimeter and the greater the magnification. The relationship between a length dx in the focal plane and the corresponding angle $d\theta$ on the sky is $dx = f \tan d\theta$, but for small $d\theta$ the tangent of the angle equals the angle itself measured in radians (i.e., $dx = f\, d\theta$). The plate scale is just $d\theta/dx = 1/f$ for $d\theta$ in radians. Converting from radians to seconds of arc ($''$) gives the relation:

$$\text{plate scale } (''/\text{mm}) = 206{,}265/f \tag{3.13}$$

where the focal length f is in millimeters. For example, if the focal length is 10 meters (10^4 mm) then the plate scale is $20.6265''$/mm. If the pixels of the CCD camera are only 0.01 mm in size then the pixel scale is $20.6 \times 0.01 = 0.206''$/pixel, or inverting this number we get $\sim$4.85 pixels per arcsecond.

The prime focus of the telescope always has the shortest focal length and there the scale is largest in seconds of arc per millimeter. The f/number essentially defines an angular field of view, the area of which determines how much light is concentrated into the focus. Small focal ratios (like $f/1$) mean large angles and more light, and hence a photographic exposure would take less time, it would be "faster". Also, the smaller the f/number is, the shorter the telescope for a given diameter, and hence the smaller the dome or enclosure. Modern telescopes tend to employ very fast primary mirrors to minimize the size and expense of the dome.

3.2.4 Correctors and hybrid telescopes

A well-known exception to the standard use of conic sections in telescope design is the Schmidt telescope developed by Bernhard Schmidt (1879–1935) in which the primary mirror has a spherical surface. A thin refracting "corrector plate" with a complex shape is placed at the entrance tube of the telescope to correct the spherical aberration of the spherical primary mirror. In this case, the focal surface is significantly curved and lies inside the telescope tube between the corrector plate and the primary mirror. Photographic plates or CCD detectors can be placed at this focus. Several famous Schmidt telescopes about 1.2 m (48 inches) in diameter, including the Palomar, U.K., and ESO Schmidt telescopes, have contributed much to astronomy by providing all-sky photographic surveys. Just as large refractors are limited to about 1 m in diameter, the corrector lens limits the size of Schmidt telescopes, with the largest being 1.34 m at Tautenberg (10 km northeast of Jena, Germany).

Other hybrid designs, such as the Maksutov and Schmidt–Cassegrain which also involve refractive corrector elements, are very popular among manufacturers of small telescopes, in the 4 inch–14 inch (10 cm–35 cm) range. The Schmidt–Cassegrain uses the same corrector plate as the Schmidt but has a small secondary mirror attached to the corrector to minimize diffraction which returns the beam towards the primary, where it passes through a hole as in a Cassegrain telescope. The Maksutov is similar except that the corrector plate is replaced with a more easily polished spherical correcting lens and the secondary mirror is an aluminized reflective patch on the inner surface of the lens.

As mentioned, to make use of the prime focus of a large telescope for wide-field imaging requires additional optics in front of the camera. Such prime-focus correctors are high-performance refractive elements. Charles G. Wynne (1911–1999) originated a relatively simple three-element corrector in 1979, and many modern correctors are derivatives of this design. As telescopes have got larger and primary mirrors faster, it has become challenging to produce a corrector with a sufficiently large field of view and as few lenses as possible to minimize light losses. One remarkable state-of-the-art example is the 170 kg prime-focus corrector of the Subaru 8.2 m telescope designed by Canon optical engineers to provide a field of view of 30 minutes of arc. There are seven lens elements in five groups beginning with a large spherical meniscus over 0.5 m in diameter. By translating one group of lenses across the optical path, the corrector also compensates for atmospheric dispersion caused by differences in the refractive index of air with wavelength.

3.2.5 Telescope mounts

Because the rotation of the Earth on its axis causes the daily rising and setting of the stars, all telescopes require a means to continually update the direction in which they are pointing. Otherwise, a star would simply drift through the field of view. No telescope points perfectly and no telescope can maintain its pointing accuracy from one part of the sky to another without numerous corrections. The sidereal rate is 360° in approximately 23 hours, 56 minutes, and 4.1 seconds, or about 15.041 arcseconds per second of time. The main problems are

- alignment of the rotation axes
- mechanical flexure of the telescope structure
- cyclic errors in the gear trains
- atmospheric refraction (which depends on wavelength, pressure, and temperature).

These effects are removed or minimized by a computer model of the telescope derived from hundreds of pointing measurements spread over the sky. Correction for the wavelength dependence in the refractive index of air requires an optical device called an Atmospheric Dispersion Compensator or ADC. Many ADCs use a pair of counter-rotating matched, compound prisms composed of two materials with different dispersive properties. The pair are arranged to give zero angular deviation at a selected wavelength, and their relative rotation provides the required color compensation. Another approach, more suitable for the very large range of wavelengths covered by CCDs (0.3 μm–1.1 μm), is to use prisms made of a single material, such as fused silica, and use the longitudinal separation of the pair to produce the color correction; the Cassegrain ADC of the Keck Telescope is of this design and can reduce the dispersion to $< 0.1''$ at $z = 60^\circ$ using prisms just over 1 m in diameter (Phillips *et al.*, 2006).

Telescope mounts fall into three basic categories depending on the motion of the axes (see Figure 3.6): (i) equatorial mounts, (ii) altitude-azimuth (alt-az) mounts, and (iii) transit mounts. There are several different implementations of the classical

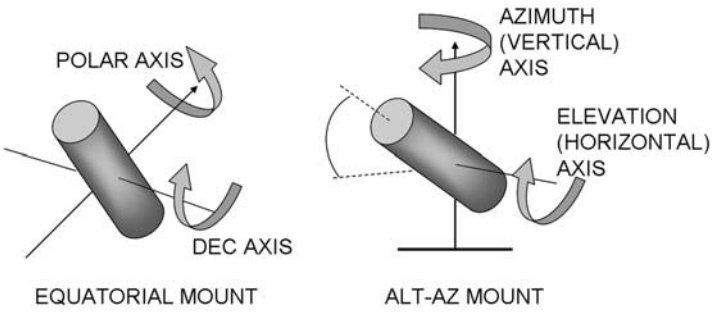


Figure 3.6. The axial movements of equatorial-mounted telescopes and altitude-azimuth mounted telescopes. The elevation axis of the alt-az is in a horizontal plane.

equatorial mount (German mount, English yoke, horseshoe mount), but all have the same fundamental purpose. In the equatorial mount, one axis of the telescope is parallel to the Earth's axis. This is called the polar axis, and it is rotated in the counter-direction from the Earth's motion at precisely the same rate. The other axis, the declination (dec) axis, moves at right angles to the first. Both axes must be moved to locate the star, but only the polar axis needs to be turned continuously to track the star. Note that seven out of ten of the older telescopes larger than 3.6 m listed in Table 3.1 have equatorial mounts, while none of the new generation listed in Table 3.2 have such mounts. Since the telescope is sweeping along the same arc in the sky as traversed by the star, the image at the telescope focus maintains the same orientation at all times. This is not so for the alt-az style of mounting which is identical to that used for radars. The telescope rotates about a vertical axis to point to any compass or azimuth direction, and it also rotates about a horizontal or elevation axis to point from the horizon to the point overhead (the zenith). To track the curved path of a rising and setting star both axes must be rotated, but the rate of rotation varies with the position of the object in the sky. In addition, the image at the focus rotates with time; this is called field rotation. The basic equation for field rotation rate is

$$\omega = \Omega \cos A \frac{\cos \phi}{\sin z} \tag{3.14}$$

where ω is the field rotation rate in radians per second; Ω is the sidereal rate (7.2925×10^{-5} radians/s); A and z are the azimuth and zenith distance, respectively, in degrees; and ϕ is the latitude of the telescope. Spherical trigonometry formulas are available to convert to (A, z) from right ascension (or hour angle) and declination (see Appendix E). Note that as z gets close to 0° (zenith) then $\sin z$ tends to zero and the rotation rate becomes infinitely fast. In practice, there is a dead zone of a few degrees near the zenith that cannot be used. This is in complete contrast to an equatorially mounted telescope which can access the zenith where airmass is least. Field rotation can be compensated by counter-rotating the entire instrument at a variable rate, or rotating an optical compensator such as a K-mirror. Without compensation there will be image smear. The amount of smear in a CCD image in seconds of arc is given by ω (rad/s) $\times$ exposure time (s) $\times$ plate scale ($''$/pixel) $\times$ radial distance from the

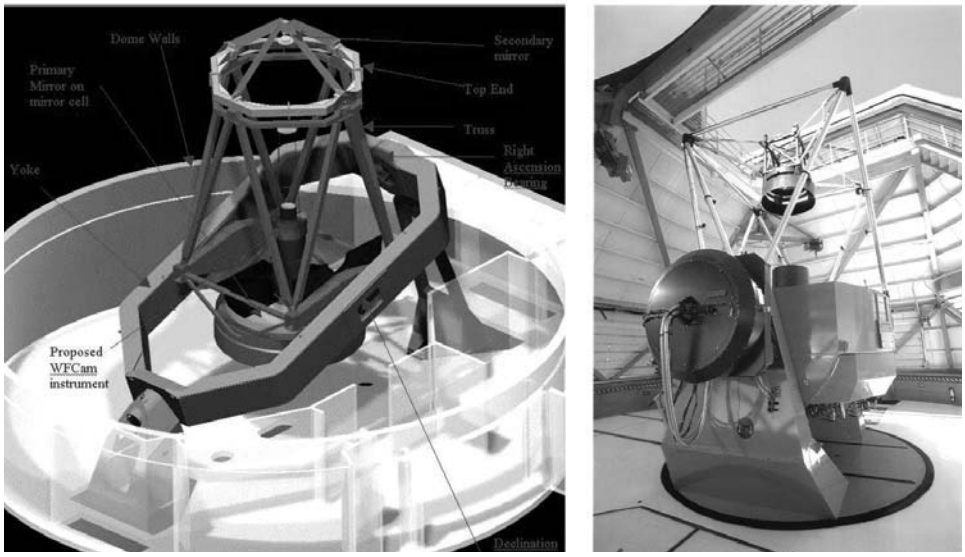


Figure 3.7. (a) Example of an equatorial telescope mount: the English yoke of the 3.8 m U.K. Infrared Telescope (UKIRT). (b) Example of an altitude-azimuth (alt-az) mount: the 3.5 m WIYN (Wisconsin–Indiana–Yale and NOAO) telescope (right). Credit: UKIRT and NOAO.

rotation axis (pixels). Despite these complications most modern telescopes are built this way because it aids the mechanical design and reduces the total weight and load on the bearings. Figure 3.7 shows the 3.8 m U.K. Infrared Telescope (UKIRT) with its English yoke equatorial mounting and the WIYN (Wisconsin–Indiana–Yale and NOAO) 3.5 m Telescope on Kitt Peak (Arizona) which has an alt-az mount, even though it is a smaller telescope; the azimuth turntable is in the floor and the telescope rotates in elevation about a horizontal axis just above the mirror cell.

The final mounting scheme is the transit telescope. In this case, the telescope is fixed to point only in the plane of the north–south meridian line, but it can select any point from zenith to horizon, and the stars are allowed to drift across the field of view. A modern example of this type of instrument is the Flagstaff Astrometric Scanning Transit Telescope (FASTT) at the U.S. Naval Observatory in Arizona. FASTT is a completely automated 20 cm (8-inch) refracting telescope using a 4 Mpxl scan-mode CCD camera to measure the positions of hundreds of thousands of stars to fainter than 17th magnitude.

There is also the question of the best way to support the mirrors of the telescope. Most modern reflectors use an open steel framework known as a truss, which is a structure with one or more triangular units constructed with straight slender members whose ends are connected at joints. The most favored design is the Serrurier truss developed by engineer Mark U. Serrurier (*c*. 1905–1988) in 1935 while he was working on the 5.08 m Hale Telescope. Two sets of opposing trusses support the primary and secondary mirrors before and after the declination (or elevation) pivot axis. The truss closest to the primary mirror is shorter because the primary mirror is

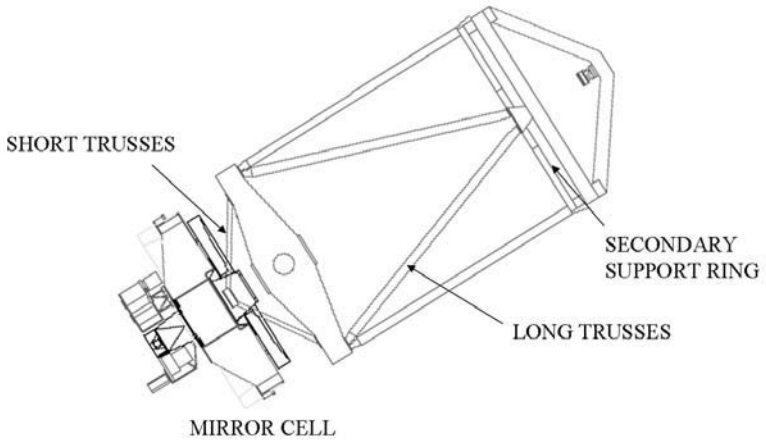


Figure 3.8. An example of the application of a Serrurier truss for mirror support. Note the short rear trusses to support the heavy mirror cell and the long trusses to support the lighter secondary mirror and support ring. Credit: UKIRT.

much heavier than the secondary. Figure 3.8 illustrates the Serrurier truss as used on the 3.8 m UKIRT. Each truss has an equal amount of flexure (bending under its own weight) such that the primary and secondary mirrors move, but remain in planes parallel to each other and therefore maintain their collimation (alignment of their optical axes). As an interesting aside, Mark Serrurier has a star on the Hollywood Walk of Fame in Los Angeles. This achievement was not for his telescope structural design—but for his later contributions (and those of his father) to mechanically complex motion picture editing equipment.

3.3 AUTOMATED TELESCOPES AND REMOTE OBSERVING

Many small telescopes are now in operation which are completely robotic with no human intervention, but scaling this up to huge telescopes and complex instruments is quite expensive. Despite the move towards automation and efficiency, there is a degree of reluctance on the part of many astronomers to give up the interactive part of observing which makes it so enjoyable and which often leads to exciting, last-minute changes of plan in response to a recent image or spectrum. How are these issues being tackled?

3.3.1 Remote observing

Global high-speed telecommunications means that remote observing at ground-based telescopes, in a manner analogous to remote observing using space facilities, is now practical. One of the earliest demonstrations of remote observing was carried out by the Royal Observatory Edinburgh (U.K.) in the early 1980s. Using satellite

telephone links, the U.K. Infrared Telescope in Hawaii and an infrared photometer were controlled and data returned to Edinburgh. Of course the data rate was low. Since then, many large observatories and numerous consortia of universities have continued to develop remote observing facilities with considerable success, in part because the rate of growth of digital communications technology has kept up with the bit rate from imaging technology. For example, astronomers in California can operate instruments on the Keck telescopes in Hawaii from one of several remote observing sites. In this case, a qualified telescope operator is onsite and the operation is not truly pre-programmed or robotic. There are two other forms of remote observing which are also used, remote "eaves-dropping" and "service" observing. These techniques are largely self-explanatory. With remote eaves-dropping, distant participants can take part in the observing session by viewing computer displays of the instrument status in real time and giving advice by means of a video-link. The actual observing is performed by a qualified team-member onsite. The service observing mode is similar except that those proposing the experiment are not expected to send a representative, nor are they expected to eaves-drop. The observations are carried out by a qualified staff scientist and the data are transferred to a digital storage medium in a standard form and then mailed to the team.

The next step in remote observing with optical/IR facilities is "queue scheduling" in a manner similar to radio telescopes. Indeed, this is one of the major goals of the International Gemini Telescopes Project for the 8 m telescopes on Mauna Kea and Cerro Pachón. The idea is that a suite of instruments are attached to the telescope and ready to go at all times, the one in use is selected by the position of a feed mirror, and the choice of instrument depends on the atmospheric conditions. For example, poorer seeing might be used by a spectrograph for relatively brighter objects, whereas excellent seeing and dark Moon conditions might call for a switch to a faint object spectrograph or imager, and a marked improvement in the dryness of the night might argue for a change to a long-wavelength infrared instrument. There is also the issue of rapid response to, for example, the discovery of a supernova or a gamma-ray burst. The issues are complex, but many smaller robotic optical telescopes are helping to lead the way (a few examples are mentioned below).

3.3.2 Automated imaging telescopes

There are now a number of sites where computers and weather-monitoring equipment make the decision to open the observatory and begin a pre-programmed set of observations. Progress in this area began with projects like GNAT (Global Network of Automated Telescopes) in the early 1990s (Crawford and Craine, 1994; Craine *et al.*, 2007). Initially, the only instrument used was a simple photoelectric photometer. Now, however, CCDs have led to Automated Imaging Telescopes (AITs). GNAT, a non-profit organization, is adopting the Moving Object and Transient Event Search System (MOTESS) of robotic imaging telescopes and deploying dozens of them around the world. The MOTESS system is cost-effective and very productive. By the end of its two-year trial period in 2003 it had discovered some comets and over 180 new asteroids, a few of which were on Earth orbit crossing trajectories. The

system consists of three identical Newtonian telescopes equipped with CCD cameras mounted side by side and no moving parts. These telescopes operate in "scan mode" by allowing the Earth's rotation to scan the field of view across each instrument. The cameras make one long uninterrupted exposure throughout the night without human assistance or intervention. There is no dead time for moving the telescope or stopping and starting the camera. Each image has a width nearly twice that of the full Moon and a length of 10–12 hours of time, typically resulting in about 200 square degrees per night per telescope. Also, each telescope contributes to a triplet of images of a given patch of sky. Images are aligned by a computer program and any moving objects, like asteroids, are immediately obvious when the set is displayed. About 1.5 million stars are also observed and tens of thousands of these are found to be variable. As Eric Craine of GNAT points out, analysts are needed to help sort out the huge number of variable stars identified so far.

Another innovative concept, ideal for teaching, is the "MicroObservatory" pioneered in the 1990s by a group at the Harvard-Smithsonian Center for Astrophysics in Cambridge, Massachusetts. The MicroObservatory consists of a network of five automated telescopes with CCD cameras that can be controlled over the Internet. Users are responsible for taking their own images by pointing and focusing the robotic telescopes, determining exposure times, and selecting filters. Each telescope is $\sim$1 m (40 inches) high and portable (135 lb), employing the Maksutov design, with a 15.2 cm (6-inch) spherical primary mirror and a 13.3 cm (5.25-inch) corrector. There is a 5.08 cm (2-inch) diagonal mirror which sends the light through a small, low power lens for focusing, then through a filter wheel with clear, blue (B), yellow–green (V), red (R), infra-red (IR), and neutral density (ND-4) filters. The filter wheel also has an opaque setting that allows for calibration and protection of the CCD. A 28 mm Minolta camera lens is used as the finder telescope. Currently, Kodak KAF0400 (768×512 pixels) and KAF1400 ($1{,}317 \times 1{,}035$ pixels) CCDs are used in the finder and main camera, respectively. Both are cooled by two-stage thermo-electric coolers to achieve a low dark current. Electronic binning gives a pixel scale of about 5 seconds of arc and a field of view larger than about two-thirds of a degree. In a 30 s exposure the main camera can detect 15th-magnitude stars. The concept is supported by the National Science Foundation, as well as Apple Computer Corporation and Eastman Kodak.

The Berkeley Automatic Imaging Telescope, pioneered by Richmond, Treffers, and Filippenko in 1993, was upgraded and moved to the Lick Observatory on Mt. Hamilton, where it is now known as the Katzman Automatic Imaging Telescope (KAIT) and is an integral part of the search for distant supernovae (used to study the acceleration of the universe). This 0.75 m (30-inch) fully robotic telescope has found about 50% of all the known nearby supernovae (about 700) in the past decade. It has also contributed to optical follow-up of gamma-ray bursts. Discoveries by KAIT are followed up by spectroscopy on larger telescopes like the Keck 10 m Telescope. KAIT is equipped with sophisticated encoders for accurate pointing, uninterruptible power supplies, absolute zero point switches for both right ascension and declination to enable the telescope to recover easily from a confused state, and of course CCD cameras for imaging and guiding. Currently, the main camera is a Proline PL77 from

Finger Lakes Instrumentation which incorporates a 512×512 *e2v technologies* back-illuminated CCD with 24 μm pixels. Equally important is the weather station hardware and software which monitors the temperature, relative humidity (RH), and wind speed. A raindrop sensor ensures that there is no precipitation, and an infrared sensor with a 12 μm–13 μm filter pointing at the sky serves as a "cloud" monitor. If the RH or wind speed exceed critical values, or if a single drop of rain hits the raindrop sensor, direct circuits from the weather station to the dome shutter motor override all computer control and close the dome shutter. In addition, there is a solar cell which prevents opening when the Sun is above a 10° altitude, and there is a "dead-man" timer which closes the observatory if no telescope activity occurs for 20 minutes. Highly efficient image-processing software is a major reason for the success of the supernova search program (Filippenko *et al.*, 2001).

The Liverpool Telescope (LT) is a fully robotic astronomical telescope of larger size which became operational in 2003 (Figure 3.9, left). It is a 2 m RC reflector on an alt-az mount housed in a clam shell enclosure on the summit of Roque de los Muchachos on the Canary Island of La Palma close to many other larger telescopes. Founded by British astronomer Mike Bode, the LT is owned and operated by the Astrophysics Research Institute of Liverpool John Moores University, and was designed and built by Telescope Technologies Limited, a spin-off company of the university. Performance and first results are described in Steele *et al.* (2004). Instruments on this telescope include an optical CCD camera with a 4.6 arcmin field of view using a $2\,\mathrm{K} \times 2\,\mathrm{K}$ chip from *e2v technologies*; a unique polarimeter (concept attrib-

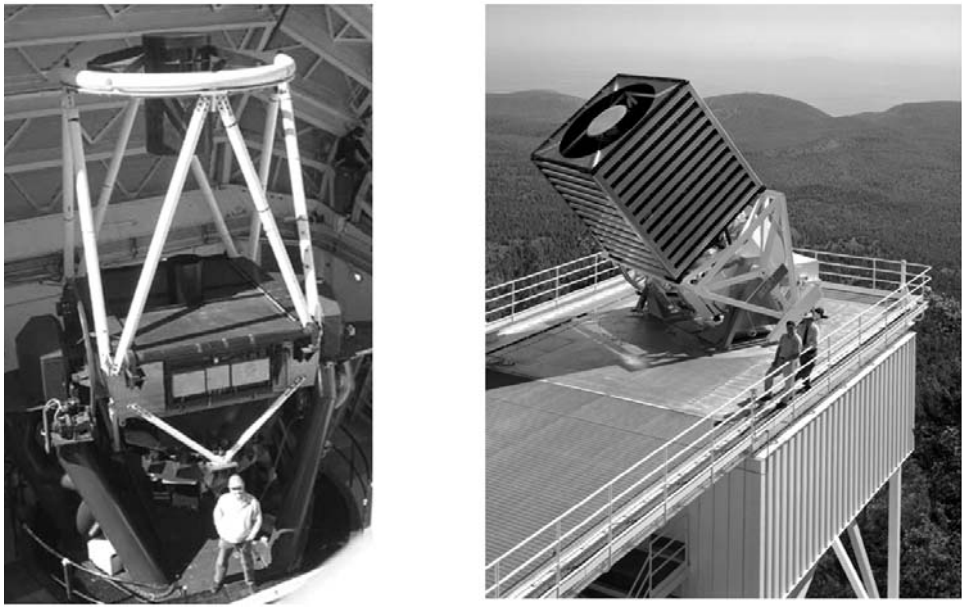


Figure 3.9. Left: the 2 m robotic Liverpool Telescope (alt-az) with founder Mike Bode in front. Right: the 2.5 m Sloan Digital Sky Survey Telescope (SDSS).

uted to David Clarke, Glasgow University) that creates small rings on a CCD for each source observed with the polarization signal mapped out as intensity variations along the circumference; and a near-infrared camera with a 256×256-pixel array. This relatively large telescope is truly robotic in that it is fully autonomous, using its control system to assess weather conditions, problems, and best observing strategies. Object-oriented software architectures are used, and the Java language is chosen for its wide usage and platform independence. Time on the LT is assigned competitively, and research concentrates on rapid response follow-up, small-scale surveys, monitoring of variable objects, and simultaneous coordinated observations with other ground/space telescopes. Two other robotic telescopes linked with the Liverpool Telescope in a global network for free educational access are those of the Faulkes Telescope Project: one telescope is in Hawaii and the other in Australia.

3.3.3 Survey telescopes

Many specialized survey telescopes, each optimized for a given task and always highly automated, have been built and the number is growing. Most of these telescopes have moderate apertures (up to about 3.5 m so far), and they are intended to perform surveys of the sky. The Sloan Digital Sky Survey (SDSS) is perhaps the most well-known of all the "survey" telescopes to date. This remarkable project uses a dedicated 2.5 m telescope located on Apache Peak, New Mexico (Figure 3.9, right) to digitally map half of the northern sky to about 23 magnitude in several bands and then to select about one million galaxies (about 1%) and 100,000 quasars (about 10%) for spectroscopy using the same wide-field telescope (see also Section 4.1.3). The SDSS Telescope is not robotic in that the nightly program is supervised by humans. Detectors for the photometric survey are an array of 30 $2{,}048 \times 2{,}048$ CCDs, and the u', g', r', i', and z' filters are all custom-designed (see also Sections 7.4.3, 9.6, and Plate 15).

Other examples of survey telescopes include the following incomplete list:

— MACHO, EROS, and OGLE Telescopes: these projects employ large-area CCD cameras on 1 m class telescopes to systematically survey particular regions of the sky in search of gravitational lensing effects.
— DENIS and 2MASS: again, 1 m class telescopes in the north and south used for a near-infrared all-sky survey using infrared array cameras (see Sections 4.1.4 and 11.7.1).
— Spectroscopic Survey Telescopes: existing telescopes such as the Anglo-Australian Telescope have been utilized for extensive spectroscopic surveys using special equipment at prime focus (Section 4.2.4), but very large telescopes have been designed with spectroscopic surveys in mind. The Hobby–Eberly Telescope (HET) is a 9 m segmented-mirror system with a spherical primary and a Gregorian secondary which is used at a fixed elevation angle for faint object imaging and spectroscopy. It is best suited for objects that are uniformly distributed across the sky and/or have low surface densities. Its southern equivalent is the South African Large Telescope (SALT).

Among new developments for the future are

— VISTA: an IR-optimized 4 m class survey telescope co-located with the VLT on Cerro Paranal in Chile. This telescope will produce the next-generation near-infrared survey (see Section 11.7.1).
— Pan-STARRS: a group of four 1.8 m telescopes being built by the University of Hawaii and intended for rapid surveys (Section 7.5.3).
— LAMOST: Chinese 4 m spectroscopic transit telescope for 4,000 objects.
— LSST: a very ambitious future project to build a 6.5 m wide-field telescope to rapidly survey the sky every few days (see Section 14.2).

3.4 VERY LARGE TELESCOPE DESIGN

3.4.1 Segmented primaries

In the vanguard of the wave of new large telescopes was the pair of 10 m collectors at the W.M. Keck Observatory (WMKO). The WMKO is managed by CARA, the California Association for Research in Astronomy, and comprises the California Institute of Technology (Caltech) and the University of California (UC), in collaboration with the University of Hawaii and NASA. The optical layout of each telescope is summarized in Figure 3.10. Each primary mirror is composed of 36 hexagonal "segments" of Zerodur, a ceramic glass made by the Schott glassworks in Germany. Each side of the hexagonal segment is 0.9 m (or 1.8 m across) and each segment is only 7.5 cm thick. There are 3×36, or 108, actuators that control the position of those segments to form a surface which is a perfect hyperboloid. Each mirror segment can be moved by the actuators at a rate of twice per second to maintain the global shape of the hyperboloid to within 50 nm; this is known as the Active Control System or ACS. Each segment is made using stress-lapped polishing, but trimming the mirror into its hexagonal form introduces small new strains which add distortions. These effects are removed by a technique called ion-figuring in which a beam of argon ions is used to essentially polish down individual regions. Finally, when each segment is located into the telescope mirror frame, fine-tuning of its shape is applied by means of a "warping harness" that exerts a permanent force on the hexagonal segment. Figure 3.11 shows the arrangement of the segments and their support structure. The mirror segments were made by Itek and Tinsley, and then ion-figured by Kodak in the U.S.A.

Development of the Keck Telescope actually dates back to 1977 when a committee of University of California (UC) astronomers met to consider plans for new telescopes and new sites. Jerry Nelson, who was at the Lawrence Berkeley Labs at that time, was struck by the fact that the mechanical flexure on the mirror due to gravity was proportional to the fourth power of its diameter (D^4). It seemed to him that this argued for smaller—not larger—mirrors! He began to look into ways of making a continuous mirror surface out of pieces or panels, just like radio telescopes. Jerry soon realized that he could not copy the radio telescope approach at optical

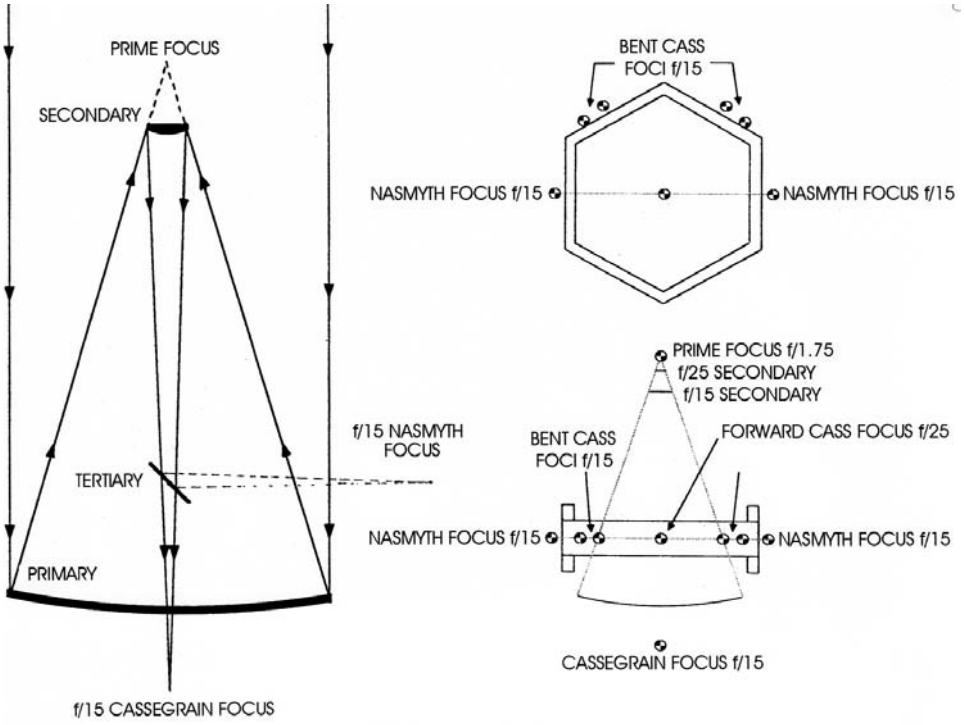


Figure 3.10. The optical layout of the 10 m Keck telescopes with the locations of all the $f/15$ foci. Credit: WMKO.

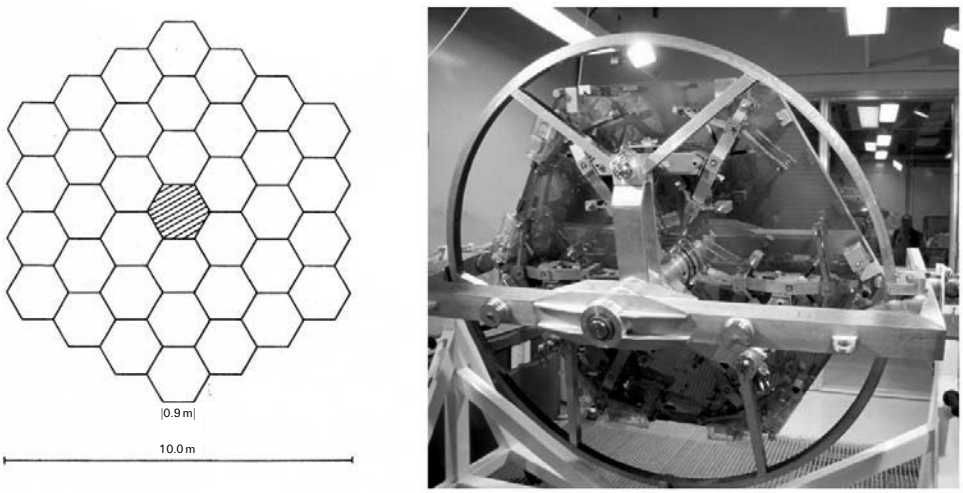


Figure 3.11. Arrangement of the 36 hexagonal segmented mirrors in the Keck primary is shown together with a close-up view of one of the segments. This single segment is mounted on its handling frame prior to aluminizing. The glass is Zerodur and the segment is 0.9 m along each edge or 1.8 m from opposite tips. Credit: Jerry Nelson and WMKO.

wavelengths, the surface accuracy is just too small. Deformation of the surface figure would have to be corrected with an "active" control system. Jerry recalled for me that he started out with the concept of a relatively large central monolith surrounded by two rings of mirrors, but within a year had dropped that idea in favor of hexagonal segments. Right from the start his vision was of a 10 m telescope. Together with Terry Mast, Jerry worked on all the technical issues, especially the active control system. What should be measured? What do you do with those measurements and how does this relate to the surface accuracy of the mirror? The breakthrough was the realization that they could perform a "global" best fit to the mirror shape without worrying in detail where one segment was with respect to another. The UC committee approved funds for a full-size prototype segment in 1979. During the period 1979–1984 details of the full-scale 36-segment structure were worked out leading to the famous "blue book" for a Ten Meter Telescope. The next hurdle was funding.

At first a grant to UC of $50 million was received in 1984 from the Hoffman Foundation, but this was not enough. The California Institute of Technology (Caltech) was approached as a partner who might be able to bring the balance of funding needed. Caltech found the W.M. Keck Foundation who wanted to fund the entire project—not just part of it. The apparent excess of funds actually stimulated the concept of twin 10 m telescopes working as an interferometer. In the end, however, difficulties in reaching legal agreements over the Hoffman gift led to the entire project being funded by the W.M. Keck Foundation, including a second telescope if the first was shown to be successful. An agreement with the University of Hawaii was successfully negotiated to construct the telescope on Mauna Kea, and all site work was done with two telescopes in mind. Balancing his work as project manager for the Infrared Astronomical Satellite (IRAS) at JPL, Gerald (Jerry) Smith supported the planning activities from 1981 to 1983 until IRAS was launched and then assumed the role of project manager full-time to make the Keck Observatory a reality. The project was officially funded by a $70 million grant to Caltech from the Keck Foundation in January 1985. By the end of 1991, with 18 segments installed and working beautifully to produce astronomical images, the Keck Foundation was convinced to begin the second telescope. The University of California provided funding for the first set of instruments and 25 years of operations. In 1996, NASA became an official partner in the Keck Observatory, and the first Scientific Director, Fred Chaffee, was appointed. My own involvement with the Keck Observatory began in 1989 when I joined the faculty at the University of California, Los Angeles (UCLA) and established a laboratory for infrared instrumentation with my colleague Eric Becklin (formerly University of Hawaii).

It is one thing to tip and tilt each segment to make all 36 images coincide at a single focus, but this is not the same as making the mirror behave as if it is a single piece of glass! The light from all the segments must be phased. What is needed is a set of adjusters attached to each mirror segment which, under computer control, can be manipulated remotely to bring all of the mirror segments into perfect alignment as if they were indeed part of a single giant surface. This is achieved using the Phasing Camera developed by Gary Channan and colleagues at the University of California, Irvine. Two position sensors are located along every inter-segment edge for a total of

168 sensors. A computer program uses the sensor readings to update 108 position actuators twice every second to keep the mirror in shape as the telescope tracks. Both telescopes are extremely productive scientifically and both are well-instrumented with optical and infrared spectrometers and cameras, multi-object spectrographs, and laser guide star adaptive optics. Keck I and II are 85 m apart with very large basements connected by a long tunnel to provide a beam-combining room for the fully operational Keck Interferometer (KI). It is clear that the segmented mirror approach works very well, which is an important conclusion because it paves the way to scale up this technology to even larger telescopes.

In 1994 contracts were let for the construction of the Spectroscopic Survey Telescope which was commissioned in 2000 as the Hobby–Eberly Telescope. A joint project involving Penn State University, University of Texas at Austin, Stanford University, Munich University, and Gottingen University, this telescope has a spherical primary made of 96 small segments, all identical, with an effective diameter of about 9 m and with a Gregorian corrector resulting in a small central obscuration. Moreover, the telescope remains stationary, aimed at a fixed elevation of 35°, rather like the huge 305 m (1,000 ft) wide Arecibo radio dish. Major instruments are coupled to the telescope by fiber optics. An almost identical version of the HET was built by another consortium in South Africa, the South African Large Telescope (SALT), which began operations in 2004. In 1996, Spanish astronomers announced plans for a 10 m class telescope almost identical to the Keck telescopes, but with slightly larger segments. The Gran Telescopio Canarias (GTC) achieved first light with 12 segments in July 2007. At the time of writing 24 segments are installed.

3.4.2 Thin-meniscus mirrors

In 1988 the European Southern Observatory (ESO) announced its intention to build an array of four independent 8 m telescopes (Figure 3.12) with the equivalent collecting area of a 16 m telescope on Cerro Paranal in Chile. Cerro Paranal is at an elevation of 2,635 m (8,645 ft) above sea level and is located in the Atacama Desert about 130 km (78 miles) south of the town of Antofagasta. This was a brand new site, quite distinct from the other locations occupied by ESO (La Silla) and by the Cerro

Figure 3.12. Layout of the ESO VLT: four 8 m telescopes with an equivalent collecting area of a 16 m telescope on Cerro Paranal in Chile. Credit: European Southern Observatory.

Tololo Inter-American Observatory. Each of the "Unit" Telescopes would be alt-az, and each would contain a large monolithic glass meniscus primary mirror ($f/1.8$) which would be so thin for its size that the mirror would require "active"—rather than "passive"—mechanical support. Hundreds of adjustable "pistons" placed against the back of the mirror would exert small forces that compensate for distortions of the figure of the mirror due to gravitational stress, thermal effects, windbuffeting, and slowly varying atmospheric blurring. Forces supplied by the mirror support need to be updated a few times per second under computer control, responding to an image analyzer observing a reference star. The secondary mirror, made of beryllium, is also computer-controlled for focus, centering, and tip/tilt. ESO pioneered many of the required techniques with their 3.5 m New Technology Telescope (NTT) and demonstrated the VLT concept with laboratory trials of a smaller prototype active control system. The VLT concept allows for individual operation of each of the Unit Telescopes, and incoherent combination of the apertures to achieve a 16 m light-bucket, as well as coherent combination to form an interferometer (the VLTI) with an angular resolution equivalent to a 100 m–200 m aperture. Each primary is oversized at 8.2 m, and the effective diameter of 8 m is set by the secondary mirrors. The focal ratios are fixed at $f/13.6$ at Cassegrain, $f/15$ at Nasmyth, and $f/50$ at coudé.

The first 8.6 m diameter mirror blank of 17.5 cm thick Zerodur (see Figure 3.13 which also shows the active control system) was successfully delivered by Schott in 1993 to the REOSC Optique factory near Paris where it was ground, figured, and polished to its final 8.2 m diameter size before shipping to the VLT site in Chile. The VLT Unit Telescope No. 1 went into operation in June 1998. Each of the other telescopes followed as shown in Table 3.2. An extensive suite of instruments (for a total of 12 foci) supports these telescopes. Details are given on the ESO web site.

The Japanese National Large Telescope (Subaru) is named after the old Japanese word for the young stellar association, the Pleiades (also known to many as the Seven Sisters). It too is an 8.2 m alt-azimuth telescope constructed with a meniscus primary

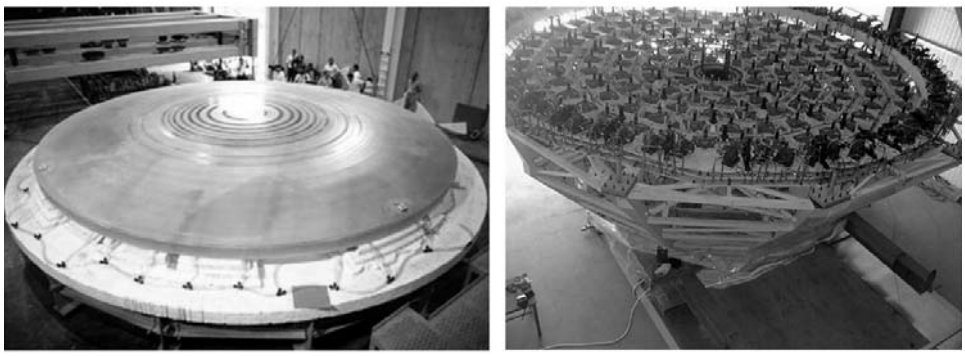


Figure 3.13. One of the thin-meniscus blanks for the 8.0 m VLT telescopes during construction. (Right): Computer-controlled pistons and actuators used for active control of the surface shape of the thin-meniscus mirror. Credit: European Southern Observatory.

mirror fabricated by Corning, U.S.A. Only 20 cm thick, this mirror is extremely flexible and could never maintain its surface figure against deformations due to gravity as the telescope points to different parts of the sky. By using 264 computer-controlled supporting structures equipped with highly sensitive force sensors, any deformations of the figure of the mirror are automatically compensated for by the control system. Subaru is located on Mauna Kea, not far from the twin Keck 10 m telescopes. Choosing the material for the mirror blank is an important issue. The glass must have good uniformity, high rigidity, and the smallest possible coefficient of thermal expansion, otherwise the surface figure will be deformed simply by changes in the ambient temperature. Ultralow expansion (ULE) glass developed by Corning Glass Works has a linear expansion coefficient of only 10^{-8} per degree Celsius. It required about 3 years to make approximately 40 boules of ULE glass, each one 1.5 m in diameter, and then fuse the boules into a blank of 8.3 m in outer diameter, which was then assembled so as to secure the same expansion properties throughout the blank. That stage of development was completed in mid-1993. Grinding and polishing (by Contraves, U.S.A.) required another 3 years, accompanied of course by numerous measurements of surface shape. The focal length of the primary is 15 m giving a focal ratio of $f/2$ (with a corrector) and a prime focus field of 30 arcminutes. At the Cassegrain focus the focal ratio is $f/12.2$ (or $f/35$ for infrared). High-precision load sensors are interrogated by the computer control system every few seconds. This system compensates mainly for gravity, but also for wind loading and residual thermal distortion.

An important way in which the Subaru differs from other large telescopes is its enclosure or "dome". Most commonly, the classical, semi-spherical dome is used to protect the telescope from the elements, and all domes must be carefully temperature-controlled to keep the mirror and structure in thermal equilibrium with the ambient air. Detailed studies of a "flushing-type" enclosure were made using wind-tunnel and water-tunnel modeling. In the end, the design adopted is a cylindrical-shaped enclosure 44 m high that co-rotates with the telescope and which is equipped with a number of active louvers, air filters, and air-flushing floor to maintain a fresh, gentle air stream under various outside wind conditions. Using computer models as a basis, the control computer arranges the louvers and air flow to get the best conditions. Subaru is well-equipped with instruments ranging from the ultraviolet cutoff around 0.3 μm all the way to 30 μm in the infrared. One example (SuprimeCam) is a mosaic of CCDs, comprising ten devices of $2\,\mathrm{K} \times 4\,\mathrm{K}$ pixels each, which is located at prime focus.

The Gemini Telescopes Project was an international collaboration between the United States, Great Britain, Canada, and the South American nations of Chile, Argentina, and Brazil. Two identical telescopes were built, one in the northern hemisphere and one in the southern hemisphere at Mauna Kea and Cerro Pachón, respectively. In the wide-field configuration, image performance goals called for 0.25 arcseconds (FWHM) over a 45 arcmin field lying within 20° from the zenith. At 2.2 μm in the infrared the goal was near diffraction-limited imaging using adaptive optics. One of the major drivers for these telescopes was optimized performance in the infrared. In fact, the goal for the emissivity of the Gemini North telescope on

Mauna Kea was a mere 2%, a goal which has now essentially been reached. This is achieved by using silver-coated mirrors which are cleaned regularly with jets of carbon dioxide (CO_2) snow and by using very narrow vanes to support the secondary mirror. Freshly coated silver yields an emissivity of about 2.5% for the combined effect of the primary, secondary, and secondary-support vanes. Degradation over several years causes the emissivity to rise to 3%–4% before re-coating is needed. Stringent image quality requirements led the Gemini Telescopes designers to opt for a Cassegrain-only arrangement. Because it was no longer required to bring a focus out along the elevation axis, the primary mirror and its cell could be raised so that the center of mass of the mirror, mirror cell, and Cassegrain instrument cluster could be balanced about the elevation axis. This approach results in a minimum mass above the primary mirror which is expected to reduce thermal seeing problems, reduce wind loading, make the primary more accessible for flushing to further improve seeing, and provide for a large cluster of four instruments at the Cassegrain focus and an adaptive optics facility.

After some controversy, in late 1993 a review panel selected thin-meniscus mirrors for the twin 8-meter telescopes over the competing technology of honeycombs of borosilicate (see below). Corning Inc. of Canton, New York was selected to provide the mirror blanks of ULE (ultra low expansion glass) and REOSC Optique (France) was chosen to do the polishing. The primary mirror support requires 120 actuators, and the alignment and figure is updated via onboard wavefront sensors every few minutes to correct for gravity and thermal deformation. A focal plane sensor controls fast guiding at >10 Hz via the tip–tilt secondary mirror. The secondary mirror itself is relatively large at a diameter of 1 m, but is made from silicon carbide (SiC) to reduce weight (50 kg), and it has two-axis articulation for fast tip–tilt motion up to 40 Hz. Large panels in the side of the telescope enclosure open to provide optimum flushing and ventilation, and the primary mirror is about 20 m above ground level.

3.4.3 Spin-cast honeycombs

The LBT, Magellan, and the MMT upgrade

Each of these telescopes are the beneficiaries of the other major innovation in telescope mirror design, that of honeycomb mirrors, pioneered by Roger Angel (University of Arizona and the Steward Observatory). To understand his idea we need to recall that the primary mirror must possess a curved parabolic surface. The deeper the curved surface the shorter the focal length of the primary mirror, and therefore the smaller the overall length and total weight of the telescope. Most classical telescopes have a ratio of focal length to diameter of about 3 to 4, which leads to large structures and large buildings (domes). It would be much more economical and provide a stiffer mechanical design if shorter focal lengths could be manufactured. The problem is the deep curve, and therefore the huge amount of glass which must be "ground and polished away" from the original flat-surface blank. Roger's idea was the following: melt the glass in a rotating furnace and allow it to

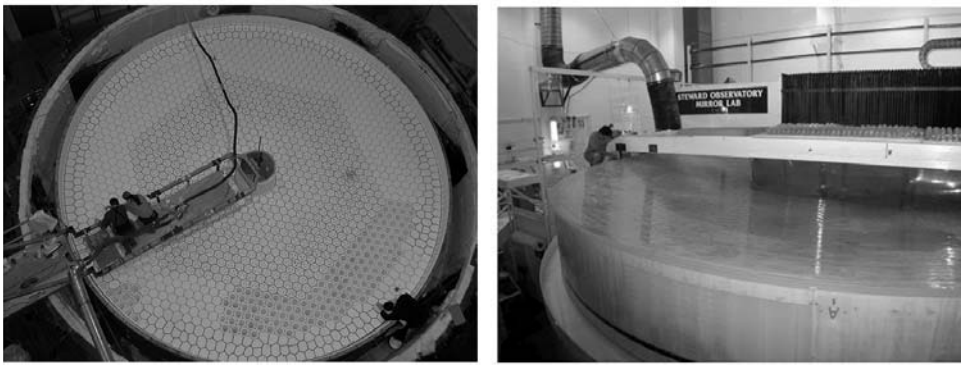


Figure 3.14. The largest spin-cast "honeycomb" mirror made to date is 8.4 m in diameter. It is shown here at the Steward Observatory Mirror Lab (SOML) at two different stages of preparation. Credit: University of Arizona.

cool while spinning. During the molten stage the liquid glass will form a deep curve in the shape of a parabola due to the centrifugal force produced by the rotating furnace, and will keep this shape as it cools while spinning. The term for this technology is "spin-casting". Roger and his team have cast mirror blanks from 3.5 m to 8.4 m in size (Figure 3.14). The mirror can be made very stiff and yet lightweight by using a "honeycomb" construction on the back surface. To do this a mold is made with a hexagonal block of ceramic fiber attached to the base of the furnace. As glass is melted into the mold it runs down the gaps between the blocks to form the ribs and backplate. After cooling, the mold is completely removed from the glass using a high-pressure spray of water to break up the ceramic fiber blocks. This casting method produces an internal honeycomb core with 11% of the solid density, with the ribs being 11 mm thick and the faces only 25 mm–28 mm thick. It takes about six weeks to anneal and cool the honeycomb blanks. The honeycomb structure also provides improved thermal performance compared with conventional solid blanks. Roger started this amazing endeavor in a backyard furnace in Tucson, Arizona, but it has since evolved to a large, sophisticated operation at the Steward Observatory Mirror Lab (SOML) under the football stadium at the University of Arizona. The web site for the SOML has many excellent pictures including some from inside the spinning furnace.

Borosilicate glass suitable for melting and casting may distort because of its larger coefficient of thermal expansion ($\sim 10^{-6}$ per $^{\circ}$C) compared with materials such as Zerodur and ULE which are highly specialized and partially crystalline. To overcome this effect, the honeycomb cells are continuously ventilated with air at ambient temperature. Laboratory tests on an earlier 1.8 m mirror showed negligible thermal distortion ($\lambda/50$), and a very rapid thermal time constant of about 35 minutes. Several 3.5 m honeycomb mirrors are installed in successful telescopes (e.g., the ARC telescope at Apache Point in New Mexico and the WIYN consortium telescope on Kitt Peak in Arizona). Three 6.5 m mirrors have been made, one of these was used to convert the Multiple-Mirror Telescope (MMT) to the Monolithic Mirror Telescope

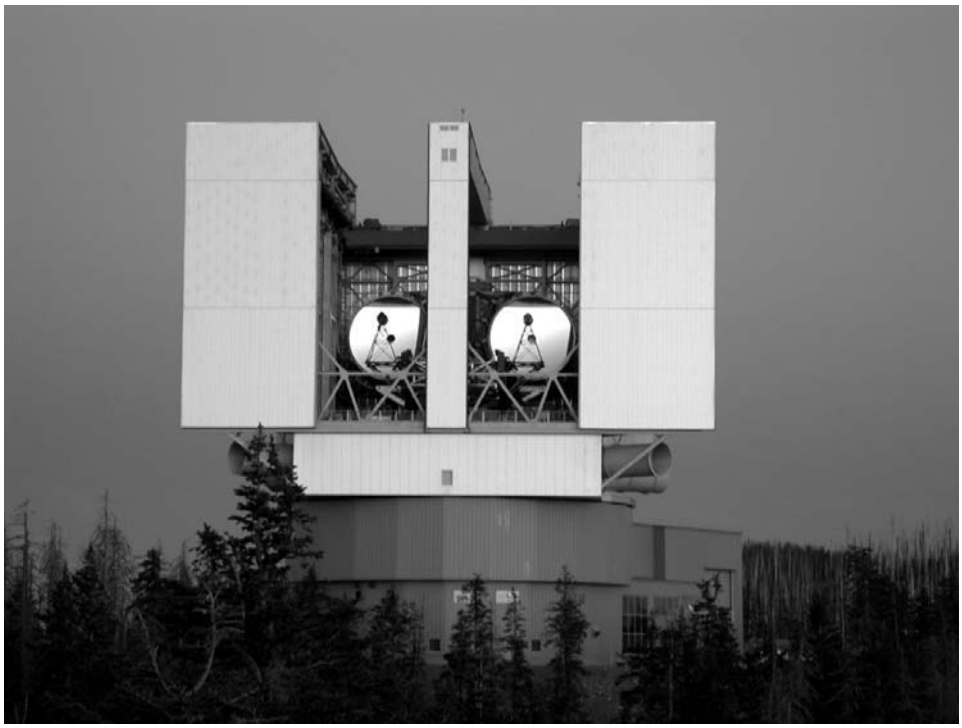


Figure 3.15. The Large Binocular Telescope on Mt. Graham with its pair of 8.4 m spin-cast primary mirrors installed. Credit: University of Arizona.

and the other pair became the twin Magellan telescopes of the Carnegie Institution of Washington, located at Las Campanas, Chile.

The Large Binocular Telescope project (formerly called the Columbus Project) is a unique telescope involving two $f/1.4$, 8.4 m borosilicate honeycomb mirrors on a common mount providing an interferometric baseline of 22.8 m which corresponds to $\lambda/D \sim 5$ milli-arcseconds at 500 nm. Located on Mt. Graham in Arizona, the LBT has both its 8.4 m primaries installed and work continues on combining the beams (Figure 3.15).

3.4.4 Prospects for Extremely Large Telescopes

With the successful completion of the segmented-mirror Keck telescopes in the mid-1990s it was natural for the astronomers at Caltech and UC to consider scaling up the idea to something larger. The report for the California Extremely Large Telescope (CELT) was issued in 2000, with a possible design for a 30 m telescope and a proposed construction timetable that would have led to an aperture doubling time of only 20 years after Keck I. However, the considerable cost of this "Extremely Large telescope" or ELT was prohibitive, not only to build but also to operate. After

a period of time the California universities merged their efforts with those of the U.S. and Canadian national observatories, in what is now termed a public–private partnership. Both of those organizations had carried out studies of their own for a Giant Segmented Mirror Telescope (GSMT) and a Very Large Optical Telescope (VLOT). The partnership was later renamed the Thirty Meter Telescope (TMT) project and Gary Sanders (Caltech) was appointed as the Project Manager. Due to conflict-of-interest issues, the U.S. National Observatory later withdrew its direct participation in the TMT project. Other large consortia also formed with the intention of building an ELT using one of the other technologies. The Giant Magellan Telescope (GMT) is a project to build a multiple mirror telescope with an effective aperture of 20 m using the borosilicate honeycomb technology and individual primaries of 8.4 m. Around the same time period, the European Southern Observatory consortium of nations began to consider a truly enormous telescope with a diameter of 100 m. Dubbed the OWL, for Over-Whelmingly Large telescope, many important studies were carried out to discover the problems and challenges of such an undertaking. At the time of writing, the European ELT is now scaled back to a proposed aperture of 42 m.

3.5 CASE STUDY: DESIGN OF A TWO-MIRROR TELESCOPE

This discussion follows the excellent exposition given by Dan Schroeder (2000) in *Astronomical Optics* and is intended to illustrate the basic rationale behind the design of modern large two-mirror telescopes. Spherical aberration is eliminated by the use of conic surfaces, and for a parabolic primary on-axis imaging is perfect. However, for a parabolic mirror receiving parallel light at an off-axis angle θ, with a vertex radius of curvature R, the angular aberrations to third order (AA3) (i.e., in the limit where $\sin\theta$ is approximated by $\theta - \theta^3/6$), may be expressed as

$$\mathrm{AA3} = 3a_1 y^2\theta/R^2 + 2a_2 y\theta^2/R + a_3\theta^3 \tag{3.15}$$

In this equation a_1, a_2, and a_3 are constants; and y is the ray height with respect to the optical axis. These three terms represent image blur due to coma, astigmatism, and distortion, respectively. In the absence of these aberrations, the foci for different θ lie on a curved surface, thus field curvature makes up the fifth aberration. Spherical aberration, coma, and astigmatism affect image quality; while distortion and field curvature affect image position, except that strong field curvature can cause defocus unless the detector is curved or the field is "flattened" with a suitable lens. An optical system in which both spherical aberration and coma are absent is called aplanatic.

The thick-lens power formula can be applied to a two-mirror system. Using the convention that $n = -1$ for a mirror, then the power of the combination is given by $P = 1/f = 1/f_1 + 1/f_2 - [(d/n)/(f_2 f_1)]$ or $f = (f_1 f_2)/[f_1 + f_2 - |d|]$ so that the net focal length (f) of the telescope is determined by the focal lengths of the two mirrors and their separation d. Figure 3.16 shows the Cassegrain configuration. Several

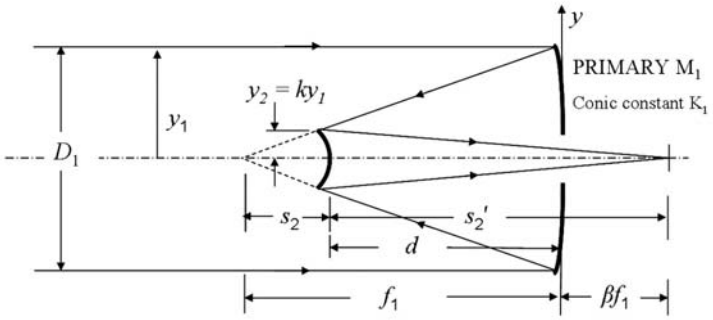


Figure 3.16. The main parameters involved in the design of a two-mirror telescope and their geometrical relationship to each other. Adapted from Schroeder (2000).

useful normalized parameters can be defined as follows:

$k = y_2/y_1$	where y_1 and y_2 are the heights of the rays at the edge of the primary and secondary, respectively;
$\rho = R_2/R_1$	where R_1 and R_2 are the vertex radii of the curvature of the primary and secondary mirror, respectively;
$m = -s_2'/s_2$	transverse magnification of secondary;
$f_1\beta = D_1\eta$	back focal distance: β and η are the back focal distance in units of the primary focal length f_1 and primary diameter D_1, respectively;
$F_1 = f_1/D_1$	primary focal ratio;
$F = f/D_1$	the system focal ratio, where f is the net focal length.

The parameters k and m are each positive for a Cassegrain and negative for a Gregorian telescope. Relationships between these terms can be found by differentiating the basic law of paraxial optics for mirrors given in Equation (3.5):

$$1/s' + 1/s = 2/R, \qquad \text{therefore, } ds' = -m^2\, ds$$

It is found that:

$$m = \rho/(\rho - k) \quad \rho = mk/(m-1) \quad k = \rho(m-1)/m \quad 1+\beta = k(m+1) \quad \text{and} \quad \eta = F_1\beta$$

The original object is always assumed to be at infinity for the primary mirror. The net power of a two-mirror telescope is found from

$$P = P_1[1 - k/\rho] = P_1/m \quad \text{(positive for Cassegrain, negative for Gregorian)}$$

The magnification m of the combination is $m = f/f_1 = F/F_1$.

If the secondary mirror is moved along the optical axis then both m and k are changed and so is the position of the focal surface, $ds_2' = -m^2\, ds_2$, where $ds_2 < 0$ when the secondary is moved closer to the primary. Relative to the fixed primary mirror, the focal surface has moved $ds_2' - ds_2 = -(m^2 + 1)\, ds_2$. Consequently, a very small change at the secondary leads to a large change in the location of the final focal plane. Motion of the secondary is limited in these telescopes by the onset of aberra-

tions because the separation d of the conic surfaces was chosen to make the on-axis spherical aberration zero; changing this separation significantly will reintroduce spherical aberration.

We are now almost in a position to calculate the first-order optical parameters of a two-mirror telescope. A minimum of five design parameters are required. These are, first, the size of the entrance pupil, y_1 (in effect, D_1), and, second, the system focal ratio, F, because, respectively, these determine the light-gathering power and resolution, and how closely the telescope and detector will match. The back focal distance ($f_1\beta$) is the third parameter because this determines the working distance for mounting instruments. The fourth significant parameter is the focal ratio of the primary mirror, F_1, because sensitivity of image quality to alignment is affected more strongly by the choice of F_1 than by any other factor; slower focal ratios are better. However, the focal length of the primary mirror $f_1 = F_1 D_1$ drives the size (and cost) of the telescope enclosure, so faster focal ratios are preferred. Hence, there is a trade-off between these two requirements. The fifth and final parameter is the field of view, which we have yet to discuss. As shown by Schroeder (2000) the full diameter of the secondary required to get a field of view of 2θ, where θ is the angular radius of the field in radians, is

$$D_2 = D_1[|k| + 2\theta F_1(1 - k)] \tag{3.16}$$

Because $k < 0$ for a Gregorian, the diameter D_2 of the Gregorian secondary is larger for the same θ and primary focal ratio F_1, hence it blocks a larger fraction of the light headed for the primary. A straightforward application of this relation, however, ignores the off-axis aberrations mentioned previously.

Coma can be eliminated by the Ritchey–Chrétien version of the Cassegrain in which both the primary and secondary are hyperbolic. Similarly, an aplanatic Gregorian can be formed if both mirrors are ellipsoids. The required conic constants for the primary and secondary mirrors can be derived from equations in Schroeder (2000):

$$K_1 = -1 - [2(1 + \beta)/m^2(m - \beta)]$$

and

$$K_2 = -[(m + 1)/(m - 1)]^2 - [2m(m + 1)/(m - \beta)(m - 1)^3]$$

The meaning of the conic constant (K) follows from the general equation for conic sections:

$$y^2 - 2Rz + (1 - e^2)z^2 = 0 \tag{3.17}$$

This equation gives the axial height y for any conic section with vertex radius of curvature R. The type of conic section depends on the eccentricity e. The term conic constant is defined as $K = -e^2$:

sphere	$e = 0$	$K = 0$
paraboloid	$e = 1$	$K = -1$
hyperboloid	$e > 1$	$K < -1$
oblate ellipsoid	$0 < e < 1$	$-1 < K < 0$
prolate ellipsoid	$e^2 < 0$	$K > 0$

The effects of astigmatism, distortion, and curvature can be estimated from a set of equations derived using third-order theory (Schroeder, 2000). These expressions give the angular blur for astigmatism (AAS) and distortion (ADI) in radians as a function of the off-axis angle θ, the final focal ratio (F), and the parameters m and β. The field curvature (κ) is expressed in terms of the vertex radius of the primary mirror (R_1):

$$\left.\begin{aligned} \text{AAS} &= (\theta^2/2F)[(m(2m+1)+\beta)/(2m(1+\beta))] \\ \text{ADI} &= \theta^3[(m-\beta)/(4m^2(1+\beta)^2)]\{m(m^2-2)+(3m^2-2)\} \\ \kappa R_1 &= 2[(m+1)/m^2(1+\beta)]\{m^2-\beta(m-1)\} \end{aligned}\right\} \tag{3.18}$$

If one takes the case in which $\beta = 0$ (i.e., the back focal length is reduced to zero), then the expressions simplify and enable an easier comparison with the classical Cassegrain in the same limit. For instance, $\text{AAS} = (\theta^2/2F)m$ for the classical Cassegrain and $(\theta^2/2F)(m+\frac{1}{2})$ for the RC. Thus, the RC has more astigmatism than the classical Cassegrain, whereas the Aplanatic Gregorian (AG) has less (because m is negative). Curvature is greater than normal for the RC, but less than this for the AG (and of opposite sign). However, the RC has twice the field size of a classical Cassegrain for the same image quality. If aberrations were the only discriminator, then the AG would be best because it has the smallest astigmatism and distortion, and the least curvature. But obstruction of light by the secondary in the Gregorian is larger than in the equivalent RC. Similarly, the primary–secondary separations are significantly larger for the Gregorian, as is the distance from the secondary to the final focal plane. Thus, for a given focal length, primary diameter, and final focal ratio, the physical length of the Gregorian is substantially greater, which has an impact on the size and cost of the telescope dome and even the telescope structure because it is longer and more massive. These factors notwithstanding, two excellent 6.5 m Aplanatic Gregorian telescopes exist in the form of the twin Magellan telescopes at Las Campanas, Chile.

With information now in hand on how to define image quality, Table 3.3 shows a possible flow for the design of a two-mirror telescope (see Exercises, Section 3.6).

All five variables are now defined. We have deduced the vertex radius of curvature for both mirrors, the conic constants for both mirrors, and the separation between the mirrors. For more details the interested reader should consult Schroeder (2000).

3.6 SUMMARY

Telescopes have doubled in size every 50 years since their invention in 1609. Reflecting telescopes with two mirrors, each a conic section, are the most effective solution for the control of aberrations. Technological advances in the way telescopes are made and controlled have led to a new generation of very large telescopes (8 m–10 m) in which the inevitable shape distortions, due to changes in gravitational loads as the telescopes move, are removed by active control of the mirror shape. Three new

Table 3.3. Design parameters for a two-mirror aplanatic telescope.

Chosen parameter	*Derived parameter*	*Comments*
Primary diameter D_1	Collecting area	Light grasp ($\sim D^2$) Diffraction limit (λ/D)
Primary focal ratio F_1	Primary focal length $f_1 = F_1 D_1$ Primary vertex radius $R_1 = 2f_1$	Size of dome Alignment tolerance
Back focal length (bfl)	$\beta = \text{bfl}/f_1$	Instrument volume Mirror support volume
Final focal ratio F	Magnification $m = F/F_1$ Effective focal length $f = mf_1$ Primary conic constant (K_1) Secondary conic constant (K_2) Aperture ratio k	Focus tolerance Plate scale From (m, β) From (m, β) From (m, β)
Field of view (2θ)	Secondary diameter D_2 Distance $s_2 = F_1 D_2$ Separation of mirrors $d = f_1 - s_2$ Image distance $s_2' = FD_2$ Secondary focal length f_2 from $f_2 = s_2 s_2'/(s_2 + s_2')$ Secondary vertex radius $R_2 = 2f_2$	For astigmatism (AAS) $< 0.5''$ Use image quality criterion to set field of view radius θ

technologies have been developed for large mirrors: segmented, meniscus, and honeycomb. Simultaneously, new generations of small robotic telescopes and a large number of semi-automated survey telescopes have been developed that also benefit from advances in computer technology. There are now many telescopes with apertures in the range 6 m–10 m, and the current technology looks like it can be extended by at least another factor of 3. Basic optical relationships for mirrors and lenses have been presented, and we have shown how to carry out a first-order design for a two-mirror telescope.

3.7 EXERCISES

1 Explain the underlying principle of the telescopes used by Galileo and Kepler. Why were telescopes from this era all very long? When Newton developed a reflecting telescope why was this design not adopted immediately?

2 Explain the difference between an equatorial mount and an alt-azimuth mount. Give an advantage and disadvantage of each.

3 The 3 m reflector at Lick Observatory has an $f/5$ prime focus and an $f/17$ Cassegrain (Cass) focus. (a) Which focus gives the fastest exposure on an

extended source and by what factor? (b) Which mode gives the fastest exposure on an unresolved point source?

4 You have a CCD camera with a large field of view attached to a typical small telescope and you see images with small "tails" toward the edges of the field, what could be wrong?

5 What are the arguments for constructing a new generation of very large telescopes? Apart from funding, why has it taken so long to consider building telescopes larger than the 5 m (200-inch) Hale Telescope on Mt. Palomar?

6 What are the three main methods for manufacturing very large mirrors for astronomical telescopes? Give an example of a telescope which uses each method.

7 Summarize the advantages and disadvantages of completely remote observing for ground-based astronomy. Why is it harder to come to a clear-cut decision for optical/IR telescopes than for radio telescopes and space telescopes?

8 Calculate the field of view covered by a mosaic of 6×5 CCDs with no overlaps if each CCD has $2{,}048 \times 2{,}048$ pixels mapped to $0.2''$ on the sky. Assuming that each exposure is 20 minutes, how long would it take to cover an area of sky of $10^\circ \times 10^\circ$? How long would it take to cover most of one hemisphere of sky?

9 Suppose you plan to construct a robotic observatory that will take observations while you are fast asleep at home, or curled up in front of your television set watching re-runs of old science fiction movies. Make a list of the equipment you would need and the steps that the computer program would have to execute from start to finish to ensure both good quality of data and safe operating conditions.

10 Using the information in Section 3.5 develop the RC design for the Keck 10 m telescopes assuming a primary mirror diameter of 10 m, a primary focal ratio of 1.75, a back focal length of 24% of the primary focal length, a final focal ratio of $f/15$, and a field of view determined by $<0.5''$ astigmatism. Calculate the radii of curvature and conic constants of the primary and secondary mirrors and find the distance between them.

3.8 REFERENCES

Angel, J.R.P. (1989) The revolution in ground-based telescopes, *Quart. J. Roy. Astron. Soc.*, **31**, 141–152.

Bode, M.F. (ed.) (1995) *Robotic Observatories*, Wiley/Praxis, Chichester, U.K.

Brown, D.S. (1986) *Technical Digest of OSA Workshop on Optical Fabrication and Testing, Seattle*, Optical Society of America, Washington, D.C.

Craine, E.R.; Culver, R.B.; Kraus, A.L.; Tucker, R.A.; Walker, D.; and Wing, R.F. (2007) A new program to search for flare events in long period variable light curves: Archived GNAT data, in B.D. Warner, J. Foote, D.A. Kenyon, and D. Mais (eds.), *Proceedings 26th Annual Conference of the Society for Astronomical Sciences.*

Crawford, D.; and Craine, E.R. (eds.) (1994) *Instrumentation in Astronomy VIII*, Proc. SPIE, Vol. 2198, Bellingham, WA.

Filippenko, A.V.; Li, W.D.; Treffers, R.R.; and Modjaz, M. (2001) The Lick Observatory Supernova Search with the Katzman Automatic Imaging Telescope, in W.P. Chen, C. Lemme and B. Paczynski (eds.), *Small-Telescope Astronomy on Global Scales*, Astron. Soc. Pacific, San Francisco, Conf. Ser. Vol. 246, 121–130.

Fischer, R.E.; and Tadic-Galeb, B. (2000) *Optical System Design*, McGraw-Hill, New York [Excellent modern introduction to optical design.]

Jenkins, F.A.; and White, H.E. (1976) *Fundamentals of Optics*, fourth edition, McGraw-Hill, New York. [Classic text on optics.]

King, H.C. (1979) *The History of the Telescope*, Dover, New York.

Kitchin, C.R. (1995) *Telescopes and Techniques*, Springer-Verlag, London.

Nelson, J. (1995) The Keck Telescopes and scientific instruments, in M. Iye and T. Nishimura (eds.), *Scientific and Engineering Frontiers for 8–10 m Telescopes*, Universal Academy Press, Tokyo, pp. 43–50.

Phillips, A.C.; Miller, J.; Cowley, D.; and Wallace, V. (2006) The Keck-I Cassegrain Atmospheric Dispersion Corrector, in I.S. McLean and M. Iye (eds.), *Ground-based and Airborne Instrumentation for Astronomy*, Proc. SPIE, Vol. 6269.

Racine (2004) The historical growth of telescope aperture, *Publ. Astron. Soc. Pacific*, **116**, 77–83.

Richmond, M.W.; Treffers, R.R.; and Filippenko, A.V. (1993) The Berkeley Automated Imaging Telescope, *Pub. Astron. Soc. of the Pacific*, **105**, 1164–1173.

Schroeder, D.J. (2000) *Astronomical Optics*, second edition, Academic Press, San Diego.

Shectman, S.A. (2000) Magellan project, *Proc. SPIE*, **4004**, 47–56. Available at *http://www.ociw.edu/Magellan*

Steele, I.A.; Smith, R.J.; Rees, P.C.; Baker, I.P.; Bates, S.D.; Bode, M.F.; Bowman, M.K.; Carter, D.; Etherton, J.; Ford, M.J. *et al.* (2004) The Liverpool Telescope: Performance and first results, in J.M. Oschmann Jr. (ed.), *Ground-based Telescopes*, Proc. of the SPIE, Vol. 5489, pp. 679–692.

Wynne, C.G. (1979) Field correctors for telescopes at better observing sites, *Monthly Notices Royal Astron. Society*, **189**, 279–285.

Web sites

Gemini Observatory: *http://www.gemini.edu/*
Glass: *http://www.schott.com/advanced_optics/english/download/abbe_vdnd_e.pdf*
KAIT: *http://astro.berkeley.edu/~bait/kait.html*
Keck: *http://www.keckobservatory.org/support/magazine/2007/june/07june_5.htm*
Liverpool Telescope: *http://telescope.livjm.ac.uk/*
Mauna Kea Observatory: *http://www.ifa.hawaii.edu/mko/*
Pan-STARRS: *http://pan-starrs.ifa.hawaii.edu/public/home.html*
Steward Observatory Mirror Lab: *http://mirrorlab.as.arizona.edu/index.php*
Subaru telescope: *http://www.naoj.org/Introduction/telescope.html*
VLT/ESO: *http://www.eso.org/public/*
WIYN telescope: *http://www.noao.edu/wiyn/*

4

The discovery power of modern astronomical instruments

Once the light has been collected by the telescope, and perhaps corrected by the adaptive optics system, it then goes through an instrument to a detector. To motivate a more detailed study of detectors and instrumentation, this chapter provides a limited review of the kinds of measurements that can be made. In the space available only a small and incomplete sampling is possible. Subsequent chapters cover the underlying principles of instruments and detectors, and expand the discussion to other wavelength regimes.

4.1 IMAGING THE SKY; MORE THAN PICTURES

Mapping the distribution of celestial sources on the sky at the wavelength of interest serves not only to locate the position of the source precisely, a practice called astrometry, but also to provide information on its form and that of its local environment. Positional changes of a faint nearby source against the stellar background might locate comets or asteroids or trans-Neptunian objects lying beyond the orbit of Pluto. With sufficient angular resolution the orbit of one star around another can be observed directly for some binary systems. Statistical properties of all kinds of stars and galaxies become practical given large-scale photometric surveys of the sky to great depth.

In 1838 Friedrich Bessel (1784–1846) published the first measurement of stellar parallax, the tiny back-and-forth angular shift in position of a foreground object against the more distant stars, which is caused by the Earth's motion (and hence changing viewpoint) in its orbit around the Sun; the average radius of the Earth's orbit is 1 AU. He detected a motion of 0.3 seconds of arc for the binary star 61 Cygni. For such small angles the distance to the object in AU is just $d = 206{,}265/p''$ where p is the parallax angle in seconds of arc ($''$). When $p = 1''$ the distance is 206,265 AU, and so it is convenient to introduce a new unit of distance called the parsec (pc) where

1 pc = 206,265 AU. A smaller parallax gives a larger distance in parsecs ($d = 1/p$). For 61 Cygni $p = 0.3''$, and therefore $d = 4.33$ pc; the modern value is $p = 0.287''$ and $d = 4.48$ pc. Until the advent of automated plate-measuring machines, computers, and then CCDs, astrometric measurements were limited in number. In 1989 the European Space Agency satellite Hipparcos (an acronym chosen to sound like Hipparchus, the discoverer of precession) was launched to perform astrometry from space. By 1993 Hipparcos had obtained the positions, parallaxes, and proper motions (non-periodic displacements due to the star's motion within the galaxy) of 118,218 stars with milli-arcsecond (mas) accuracy. Later, additional catalogs called "Tycho" were issued with 1,058,332 stars to positional accuracies of 20 mas–30 mas, and then 2,539,913 stars covering 99% of all stars brighter than 11th magnitude. The detector for these measurements was an image dissector tube scanned at 1,200 Hz and photomultiplier tubes, no CCDs were used in this case. The U.S. Naval Observatory B1.0 catalog is a tabulation from digitized plates that gives positions for over 1 billion stellar objects to ~0.2 seconds of arc. There have been major changes in the field of astrometry in recent years (Seidelmann and Kovalevsky, 2002). One driving force has been the increased accuracy of measurements using very long baseline interferometers (VLBIs) to achieve positional accuracies for extragalactic radio sources to much less than 1 mas, which has now led to a new "space-fixed" reference system. The new International Celestial Reference Frame (ICRF) is defined to be close to the previous dynamical reference frame (the FK5) at J2000.0. In addition, the Global Positioning System (GPS) of satellites has yielded accurate and continuous time transfer and geodesy observations of the polar motion and Earth rotation to the same sub-mas levels. At these levels of accuracy the definitions of the reference systems and the methods of reduction must be based on the theory of relativity. The fundamentals of astrometry are described by Kovalevsky and Seidelman (2004).

Measuring the brightness of a source over relatively broad wavelength bands is called photometry; the term radiometry is used at longer wavelengths. It is also possible to use narrow bands that isolate specific spectral features such as emission from hydrogen gas or other ionized atoms. Narrow-band images can be extremely effective in delineating nebulae, supernova remnants, and shock fronts. Photometric methods will be discussed in Sections 5.1 and 9.6. Here we introduce only a few basic terms to maintain the flow and connect radiometric measurements with the familiar astronomical magnitude system. The magnitude system for astronomical brightness measurements dates back about 2,000 years to the work of Hipparchus (*c.* 127 BC) and Ptolemy (*c.* 137 AD) who listed the brightest stars visible to the naked eye and called them stars of the "1st magnitude" while those just barely discernible to the eye were designated as stars of "6th magnitude". After the invention of the telescope, efforts were made by the Herschels to develop a better system, but the scale in use today stems from the work of English astronomer Norman Pogson (1829–1891) who in 1856 proposed a mathematically precise logarithmic scale that was in approximate agreement with the ancient system. He noticed that a 6th-magnitude star was about one hundred times fainter than a first magnitude star and so he proposed that a difference of five magnitudes be set exactly equal to 100. This means that the ratio between successive magnitudes is given by the fifth root of 100, which is equal to

2.5119. If m_1 and m_2 are the magnitudes of two stars whose fluxes have been measured to be S_1 and S_2, respectively, then $S_1/S_2 = 2.5119^{(m_2-m_1)}$. Note that m_2 comes before m_1 in the exponent because fainter objects (smaller S) means larger magnitude values. A more convenient way of writing Pogson's result is to use logarithms to the base 10 as follows:

$$\log(S_1/S_2) = \log(2.5119^{(m_2-m_1)}) = (m_2 - m_1)\log(2.5119) = (m_2 - m_1)(0.400002)$$

Then, to very good accuracy, we have that $2.5\log(S_1/S_2) = (m_2 - m_1)$ or reversing the order we get the traditional expression:

$$m_1 - m_2 = -2.5\log(S_1/S_2)$$

Note that this well-known equation does not allow either S_1 or S_2 to be determined but only their ratio. In other words, the magnitude of one star is found relative to another and the photometry is relative. Calibration of magnitude systems is discussed in Section 9.6. Brightness ratios (equivalent to magnitude differences) between two wavelength bands (e.g., blue and yellow/visual) define "colors" and can also yield basic physical information, such as the temperature of the object. A plot of magnitude vs. color for stars of known distance yields the important Hertzsprung–Russell (HR) diagram of stellar luminosity as a function of temperature. Likewise, plotting one color against another color, especially over the broad optical/IR regime, is a powerful method for separating different classes of objects. For distant galaxies, color information can even provide an estimate of the distance (called a photometric redshift), which would otherwise require a spectrum of the object. Brightness measurements can be made at all wavelengths, and comparisons among the X-ray, optical, infrared, and radio fluxes are often diagnostic of the physical process. In fact, the opening up of these other wavelength regimes produced many surprises in terms of which objects were brightest. For example, the front cover of this book shows how different one object can appear (Centaurus A) in radio, X-ray, and visible light (Hardcastle *et al.*, 2008). Figure 4.1 shows these images again plus an infrared view from the Spitzer Space Telescope. Plate 4 shows a false-color composite.

Likewise, variations of brightness with time can be extremely important. Periodic variations can reveal an eclipsing binary system, for instance, or a pulsating star like a Cepheid variable which in turn, through the period-luminosity law, allows a distance estimate. Non-periodic variations in brightness might be used to reveal unusual stellar activity like a nova, or an active galactic nucleus harboring a black hole and accretion disk, or the slow decay in brightness of a distant supernova whose light-curve shape over many weeks is characteristic of the type of explosion.

4.1.1 Early surveys of the sky

Imaging the entire sky is a daunting task no matter what technology is used. The number of square degrees in the whole sky is about 41,254 while for comparison the area covered by the Moon is only $\sim$0.2 square degrees. Most telescopes give high magnification and hence a small field of view (typically $<$1 degree across). The exception is the Schmidt telescope which can provide $\sim$42.25 square degrees of field.

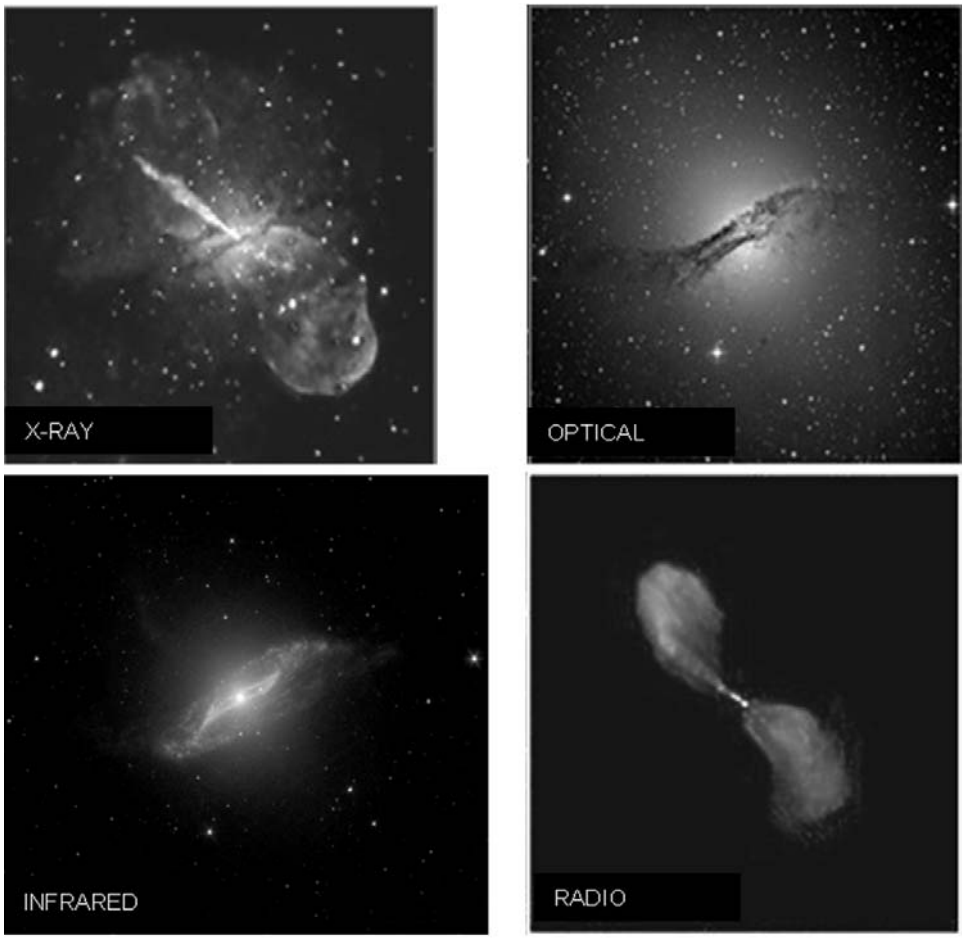


Figure 4.1. Images of the galaxy Centaurus A (Cen A) in X-rays, visible light, infrared, and radio (see also Plate 4) illustrate a dramatic change in appearance with wavelength. Credits: NASA/NSF/NRAO/ESO. See book cover credits.

Even so, this still implies at least 977 images to cover the entire sky. Exposure times need to be long enough to reach faint magnitudes (e.g., 20th magnitude in visible light) and the image quality must be uniformly good. For an all-sky survey two telescopes are needed: one in the northern hemisphere, the other in the south; and we also need a detector with a large number of pixels to cover this enormous field of view. The advent of Schmidt telescopes and large 14-inch photographic plates in the 1930s made the concept of whole-sky photography possible. Funded by the National Geographic Society, the Palomar Observatory Sky Survey (POSS) of the northern sky was carried out from 1950 to 1957 with the 1.2 m (48-inch) Schmidt on Mt. Palomar (California) using Kodak 103aO and 103aF plates. The Palomar Schmidt telescope has an $f/4.5$ 1.83 m (72-inch) primary and a 1.2 m aperture covered by a

glass corrector. In 1987 this telescope was renamed the Samuel Oschin Schmidt. A second epoch survey POSSII was started in 1985 and ended in 2000, stimulated in part by the availability of finer grain emulsions such as Kodak IIIaJ (blue) and IIIaF (red) in the 1970s, and by the need for guide stars for the future Hubble Space Telescope. Meanwhile, similar telescopes were built in the southern hemisphere and new surveys were established by the U.K. Schmidt Telescope (then part of the Royal Observatory Edinburgh but now operated by the Anglo-Australian Observatory) at Siding Springs in Australia, and the ESO Schmidt Telescope in Chile. The U.K. Schmidt had an achromatic corrector and was supported by the COSMOS (later the Super-COSMOS) plate-measuring and archive facility in Edinburgh. In the early 1980s the Palomar telescope was upgraded with an achromatic corrector plate, and POSSII was then designed to be the northern complement of the U.K. survey. Blue, red, and near-infrared (IVN) emulsions were used to obtain images in three colors, hence tripling the numbers of photographs to be taken. While each plate covered $6.5^\circ \times 6.5^\circ$ of sky and a typical plate might easily contain 4.5 billion (potential) picture elements, the grains of emulsion, only 1 out of 50 incident photons was detected (quantum efficiency $= 0.02$). The product of the number of pixels and the quantum efficiency of a typical (hyper-sensitized) Schmidt plate such as Kodak IIIaJ emulsion is about 60 times better than that of a single $1{,}024 \times 1{,}024$ (one megapixel) silicon CCD with 80% quantum efficiency. However, the construction of large mosaics of CCDs with hundreds of megapixels has now overtaken even this advantage of large-area plates. Moreover, the CCD provides very large gains in the ultraviolet and far-red parts of the spectrum. Higher quantum efficiency implies that fainter limits are reached in the same time. Alternatively, the same detection limit can be reached in a much shorter time, hence allowing several other patches of sky to be measured.

4.1.2 Digitized surveys

Using plate-measuring machines (e.g., Figure 4.2), the all-sky photographic surveys produced by the Palomar (first and second epoch), U.K., and ESO Schmidt Telescopes are now in digital form and available online as the Digitized Sky Survey (DSS). You can go to the DSS web site and type in the coordinates (right ascension and declination) of the region of interest. If the object has a name, then you can use that instead and the facility will find the coordinates for you. For example, I typed in M1 (the Crab Nebula) and the coordinates were returned. You can then select the field size to display (up to 15×15 arcmin) and the type of image format, either GIF or FITS. The GIF format is good if you want a picture for a finding chart or illustration, but the FITS format, the standard among professional astronomers, is better if you want access to the digital data to change the contrast or make measurements on the image. It is easy to obtain display programs for FITS files, including an add-on for the well-known Adobe Photoshop program (see Section 10.1.2). A plate-scanning device and the result of my search of the DSS are shown in Figure 4.2.

New sky surveys are now digital from the outset. In fact, the Palomar Oschin Schmidt was converted in 2000 to create an all-digital partial sky survey (DPOSS:

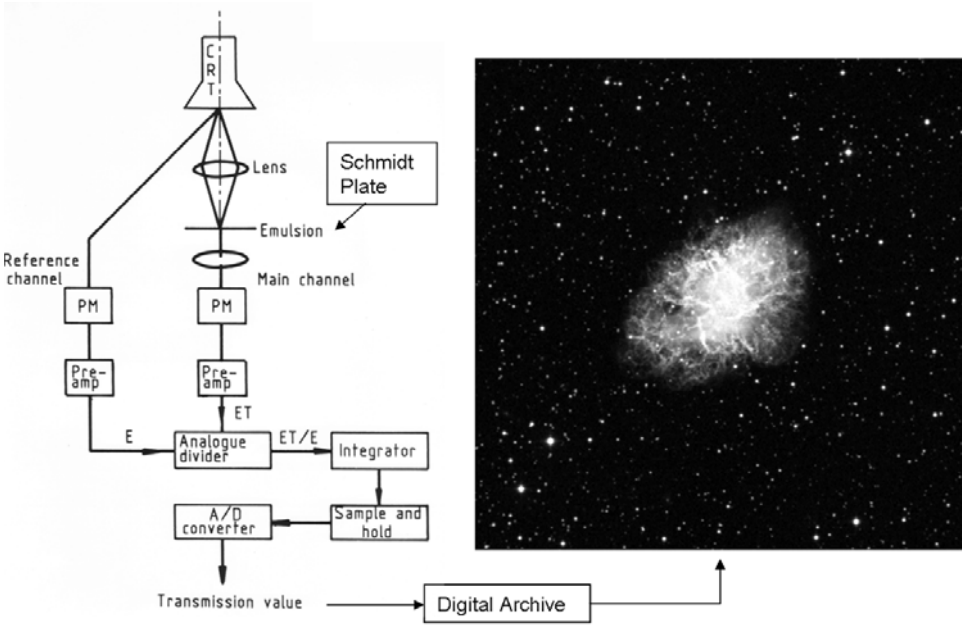


Figure 4.2. A plate-measuring machine used to convert photographic images to digital form and an image of the Crab Nebula (M1) obtained over the Internet from the Digitized Sky Survey. PM = photomultiplier; T = transmission; E = incident illumination. Credit: Royal Observatory Edinburgh.

Djorgovski *et al.*, 1999) using a mosaic of 112 CCDs instead of the original 14-inch plates (Figure 4.3). An enormous mural from part of this digital survey can be seen on a wall at the public Griffith Observatory in Los Angeles. This successful digital survey led to the discovery of many new Kuiper Belt Objects including Quaoar, Sedna, Orcus and Eris (e.g., Brown *et al.*, 2004) and many new distant quasars. DPOSS contains about 3 terabytes of images plus catalogs of extracted sources. Software using artificial intelligence methods was developed to automate classification and measurement of the 50 million galaxies and 500 million stars.

Electronic imaging devices allow astronomers to go well beyond the boundaries of visible light photography. Among the advantages of CCDs over photographic plates in this context are

(1) Very high sensitivity, a much larger fraction of the light falling on the CCD is converted to a measurable quantity (in this case, an electrical voltage).
(2) Greater coverage of the spectrum, the silicon material of the CCD is sensitive to light from the atmospheric cut-off in the UV (0.3 μm) to the near-infrared around 1 μm.
(3) Immediate compatibility with computers, no developing or processing of film, the "output" from a CCD is suitable for transfer to a computer.
(4) Instant display of the image on a computer screen.

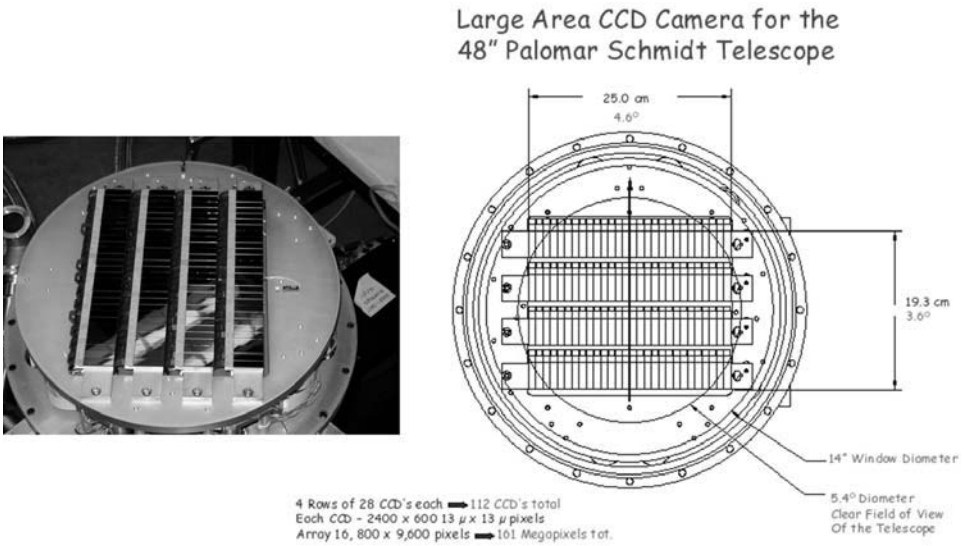


Figure 4.3. Mosaic of 112 CCDs used to make the Digital Palomar Observatory Sky Survey using the same Schmidt telescope as the original photographic survey. Credit: California Institute of Technology.

CCDs offer the additional very important advantages of precision and stability which are essential for quantitative brightness measurements. Because they are readily calibrated, CCDs are therefore excellent photometers.

Smaller telescopes are astonishingly revitalized by the addition of a sensitive CCD camera or other CCD-based instruments. For example, since the initial discovery of a planet orbiting another star in 1995, hundreds are now known and some of these are edge-on systems. Therefore, many small telescopes with CCD cameras have been able to witness the tiny changes in brightness (usually a fraction of 1%) of the parent star when the planet transits across the face of the star. One example is shown in Figure 4.4. The light curve at the top is from a 286 mm $f/2.9$ Schmidt ($\sim$4-inch aperture) ground-based telescope with a CCD camera, while the one below is from the Hubble Space Telescope STIS instrument. This object is a 7.65-magnitude Sun-like star (HD209458) about 150 lightyears away in the constellation Pegasus and contains a large Jupiter-like object in a very close (3.5-day) orbit that lies within about 3° of our line of sight to the star (Charbonneau *et al.*, 2000). Measurement accuracy with a CCD camera is typically $\sim$1.5 milli-magnitudes or 0.0015 magnitudes, a factor of 10 smaller than the depth of the light curve. Transit photometry and radial velocity spectroscopy together allow stellar and planetary radii to be deduced. It is even possible to probe the planetary atmosphere because at some wavelengths more light will be absorbed than others and the planet's radius will appear different and thus the transit depth will be different; the multiple light curves in Figure 4.4 (bottom) are different wavelength regions. In March 2005 the Spitzer infrared space telescope detected direct thermal emission from HD209458b by subtracting out the

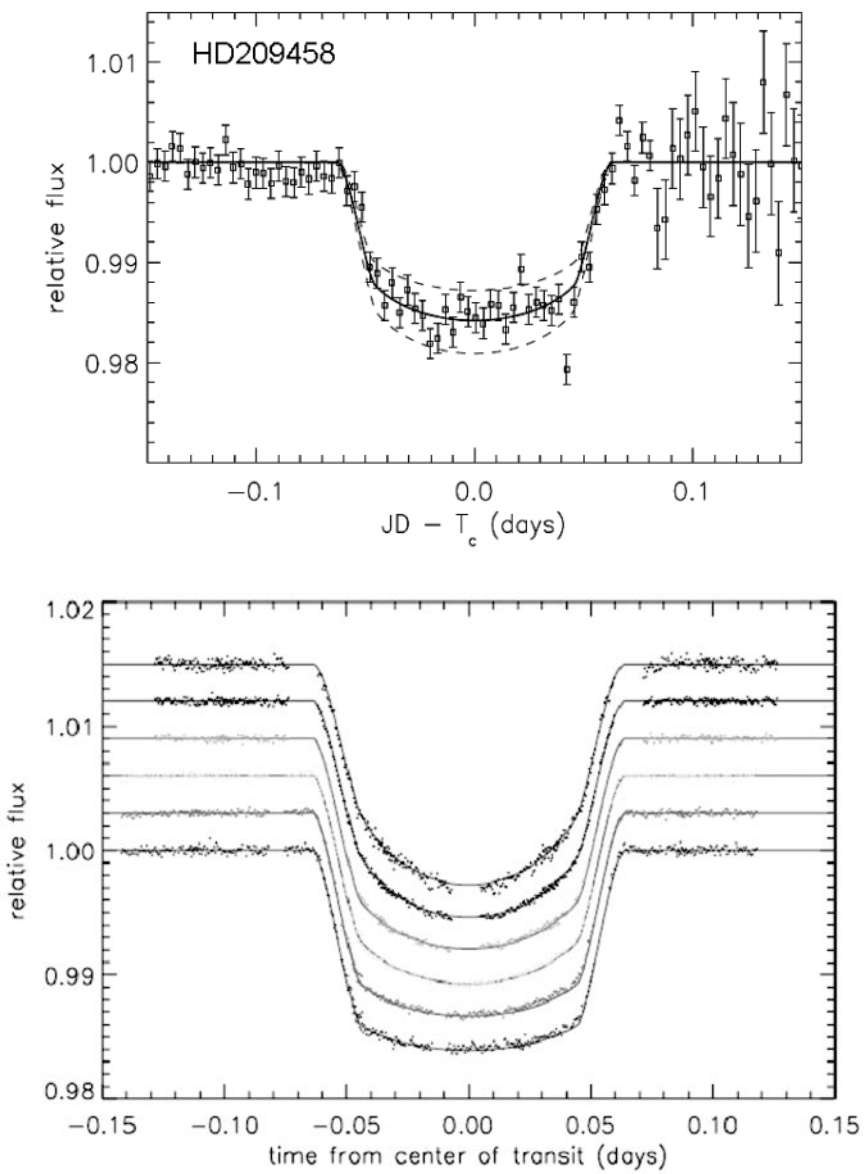


Figure 4.4. The light curve of the transit of a planet across its star in the system HD209458 obtained with a sensitive CCD camera. Credit: David Charbonneau.

light of the star measured while the planet was *behind* the star—rather than in front of it.

In the previous chapter we also mentioned CCD-equipped telescopes that are used in automated fashion to find supernova explosions in distant galaxies. The decaying light curve can then be followed by many telescopes over a period of weeks

to determine if the event was a Type Ia, the so-called "standard candle" explosion that results from detonation of a white dwarf star in a close binary system. Because the luminosity (emitted power) of such objects is known, their apparent brightness (watts per square meter arriving at the Earth) can be used to determine distance from the inverse square law. Combined with spectroscopy to get the redshift, observations of distant supernovae have led to new insights about the expansion of the Universe and support the idea that the expansion is accelerating due to the presence of dark energy (Riess *et al.*, 1998; Perlmutter *et al.*, 1999). In the current standard model of cosmology, dark energy accounts for almost 75% of the total mass-energy of the Universe. The nature of the physics driving cosmic acceleration is perhaps the biggest question facing astrophysics today. Is it Einstein's "cosmological constant" that associates an energy density with the vacuum of space itself or something else? Dark energy causes expansion because it has strong negative pressure, like tension in an extended rubber band. In general relativity, negative pressure contributes to gravitational acceleration (the bending of spacetime) by causing repulsion. Huge resources and large collaborations are being amassed to tackle this problem in an effort to determine the equation-of-state parameter $w = p/\rho$ which measures the pressure-to-energy-density ratio; $w = -1$ for a cosmological constant but not for other explanations. In addition to supernova searches from space-based and ground-based facilities with large CCD/IR cameras, other avenues include weak gravitational lensing, baryon acoustic oscillations, and cluster counts.

Still other telescopes use CCD cameras to track non-sidereal moving objects such as asteroids, including the near-Earth asteroids whose orbits cross that of the Earth. Pioneering efforts to introduce CCDs for this purpose were first demonstrated in the early 1990s at the University of Arizona (Gehrels, 1991), and there are now a great many projects of this nature. Perhaps the most ambitious of all of these is the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS) at the University of Hawaii. Using an array of four telescopes each equipped with CCD cameras having billions of pixels, Pan-STARRS will be able to survey large areas of sky rapidly for transient objects, from asteroids to supernovae. A link to the web site is given in the references.

What do CCD cameras look like on the outside? In most CCD cameras the chip is mounted in a small vacuum chamber because it must be cooled to a low temperature using either liquid nitrogen or an electrically operated cooler. Every CCD camera requires a "box of electronics" with circuits to operate the CCD. One of these circuits, the analog-to-digital converter, translates the voltage signal from the CCD into a digital number which can be stored by a computer. Thus, despite the relatively small size of a single CCD chip, the entire vacuum-cryogenic camera can be quite large, especially if many CCDs are used to form a mosaic. An external view of the CFHT MegaCam which has a 40-CCD mosaic is shown in Figure 4.5 together with a remarkable image of a nebula obtained by astronomer Jean-Charles Cuillandre (see also Plate 2).

In all cameras there will be some small amount of light which gets into the detector after being scattered from the slender struts or vanes which support the secondary mirror of the telescope, or from the edges of the primary mirror. Scattered

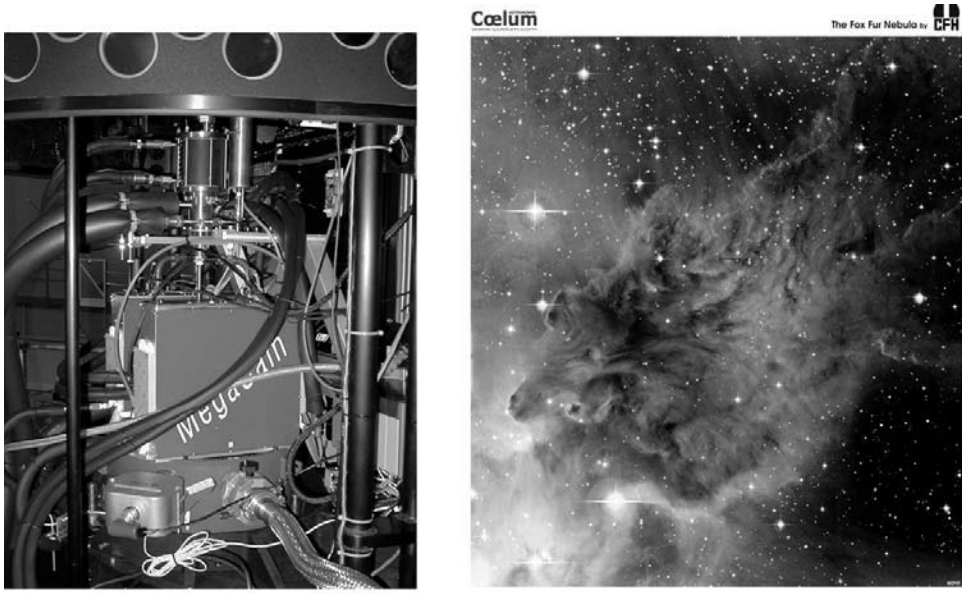


Figure 4.5. An exterior view of the cryogenic CFHT MegaCam and an image obtained with this 40-CCD camera (see also Plate 2). Credit: Jean-Charles Cuillandre.

light is particularly serious when looking for very faint sources close to a very bright source. One way to overcome this is to place a small opaque finger or disk in the focal plane to blot out or occult the bright object, in much the same way as the Moon covers up the Sun during a total eclipse to reveal the outer layer or corona of the Sun's atmosphere. To cut down scattering from the secondary mirror support vanes, a blackened mask is constructed with a shape which imitates the size and orientation of the vanes, and this mask is placed very accurately inside the camera at a position where there is an optical image of the secondary mirror. If there is no such image then one can be created by means of a "field lens" (i.e., a weak lens placed at the focus of the telescope). This arrangement is called a coronagraph. Figure 4.6 shows a remarkable result from an instrument using this technique. The thin diagonal streak emerging from behind the mask is a faint 400 AU long disk of protoplanetary material around the 4th-magnitude star Beta Pictoris. First detected by the Infrared Astronomical Satellite in 1983, this optical discovery image was obtained in 1984 with an early CCD camera (Smith and Terrile, 1984). New coronagraph images with even higher resolution taken from 1995 to 1997 with the Hubble Space Telescope are also shown. Many such objects are now known, and measurements like this have opened up the study of young proto-planetary disks. Stringent efforts are also being made to include sophisticated coronagraphs in advanced near-infrared adaptive optics systems for very large telescopes to enable the direct imaging of warm planets around nearby stars (Macintosh *et al.*, 2006). These high-resolution imaging systems will also obtain spectra sufficient to identify the presence of methane, as seen in Jupiter.

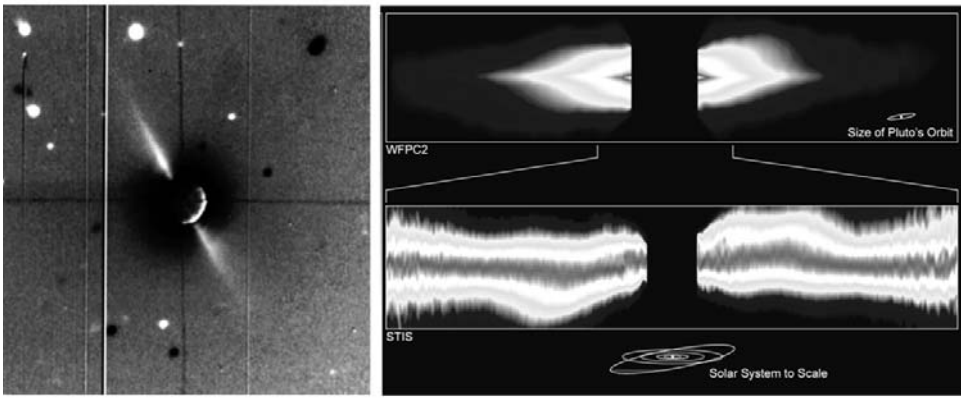


Figure 4.6. Left: the first optical image by Smith and Terrile of the disk around the star Beta Pictoris obtained with an early CCD camera and an occulting spot (a stellar coronagraph). Credit: Brad Smith. Right: Hubble Space Telescope images of Beta Pic showing the full 1,500 AU extent with the WF/PC2 camera and a close-up view of the warped inner regions down to 15 AU using the STIS cameras. Credit: Al Schultz (CSC/STScI and NASA) and Sally Heap (GSFC/NASA).

4.1.3 Drift scanning and the Sloan Digital Sky Survey

A remarkable property of the charge-coupled device that gives it its name is that the electronic charge packet representing the image on a group of pixels can be moved (along a column of pixels) one row at a time at *any* desired rate. Charge-coupling is explained in Section 7.2.2. Here, we just need to know that the CCD can be operated at the "sidereal rate" (i.e., the rate at which the Earth turns on its axis relative to the most distant stars). Suppose a sensitive CCD camera is placed at the focus of a moderate-sized telescope which is pointing at a field on the celestial equator that is just crossing the meridian, but the telescope drive motors are switched off. Stars will drift across the CCD pixels, at the sidereal rate of about 15.4 arcseconds per second of time, producing trails. Now, initiate the electronic process of reading out the CCD charge pattern at the sidereal rate and in the same direction as the star moves, and then open the shutter. There will no longer be star trails. Instead, the charge image from previous pixels is added to the next one and the current position of the charge pattern will move along the column so as to keep up with the current optical image position (see Figure 4.7). More and more photons will be collected and, ultimately, the entire column of pixels will be read out and will have contributed to the detection process. Each read-out adds a row of pixels to a stored frame of data whose width is that of the CCD but whose length is arbitrary. Carrying on in this way, a huge strip of sky can be surveyed systematically to a deep level without actually moving anything except electrons! This extremely important technique is called "drift scanning" and it is used extensively to perform the Sloan Digital Sky Survey (SDSS). Similar methods are used by many other digital sky surveys.

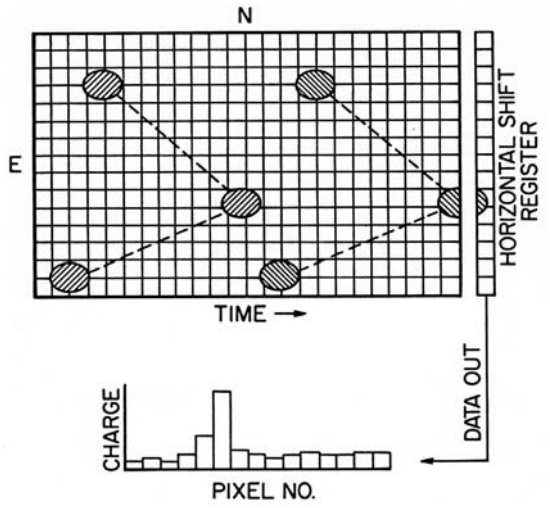


Figure 4.7. The principle of "drift-scanning" in which the unique charge-coupling property of a CCD allows the image charge to be moved along columns from pixel to pixel at any rate. This technique is used to produce the Sloan Digital Sky Survey.

The SDSS, already in its second phase, has provided a very deep, all-digital survey of 50% of the northern sky (and about 25% of the whole sky). In what ways is this survey better than digitized photographic plates? First, the survey is deeper (around 21–22 magnitude) because a larger telescope was used (2.5 m vs. 1.2 m) and, second, observations were made in five well-defined color bands spanning the visible and near-infrared, which implies more powerful diagnostics for extracting different types of sources. All-digital from the beginning, the SDSS benefits from the linearity of CCDs which makes for good photometry and from advances in computing power to provide an automated data reduction "pipeline" that can extract identified sources and produce catalogs for incremental release. Finally, follow-up spectroscopy is built in to the SDSS (Section 4.2.4). At the time of SDSS Release 6 (June 2007) over 9,583 square degrees of sky had been surveyed and over 1,271,680 spectra had been obtained. Sample images and spectra are available at the SDSS web site. Equally fascinating was the announcement in August 2007 of Google Sky, a project to visualize the SDSS and other digital imagery (e.g., DPOSS) in the same way as satellite maps of the Earth are used in Google Earth.

4.1.4 The Two Micron All Sky Survey

In the 1960s a survey of about 70% of the sky at 2.2 μm was carried out by Caltech scientists Gerry Neugebauer and Bob Leighton. Using only a small telescope and single-element detectors, that survey revealed about 5,700 sources. Thirty years later, following the advent of digital infrared array cameras in the mid-1980s, a pair of highly automated 1.3 m telescopes, one on Mt. Hopkins, Arizona in the north and the

other at Cerro Tololo, Chile in the south, provided a much deeper Two Micron All Sky Survey (2MASS: Skrutskie *et al.*, 2006). Each telescope was equipped with a three-channel infrared camera to observe simultaneously at three infrared bands designated J (1.25 μm), H (1.65 μm), and K_s (2.17 μm), respectively. Infrared techniques will be discussed in detail in Chapter 11. The cameras imaged the sky with a pixel resolution of about 2 seconds of arc while the telescopes scanned smoothly in declination at a rate of about 1 minute of arc per second. Thus, the 2MASS data "tiles" are 6° long in declination and one camera frame wide (8.5 arcminutes). As the telescope scanned in declination, its secondary mirror was tilted in the opposite direction at the same rate to momentarily "freeze" the image on the detector. Then, in less than one-tenth of a second, at the end of a 1.3-second exposure the secondary mirror flipped back to its start position and froze a new piece of sky slightly displaced from the previous frame. This mode of operation using infrared arrays that do not operate using the charge-coupling principle is called Time Delay and Integration (TDI) and is the equivalent of drift-scanning with a CCD. The 2MASS scanning arrangement causes a progressive shift of about 1/6 of a frame in the camera's field of view along the declination axis, and final images have a total integration time of 7.8 seconds. The northern survey began in June 1997 and the southern one in March 1998 with both operations complete by February 2001. The University of Massachusetts (UMass) was responsible for the overall management of the project, and for developing the infrared cameras and onsite computing systems at both facilities. The Infrared Processing and Analysis Center (IPAC) at Caltech had responsibility for all data-processing through the production pipeline, and for construction and distribution of the data products. This project involved the participation of several different institutions and was funded by the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF). The All-Sky Release data products include a Point Source Catalog containing positions and brightness information for over 470 million objects, an Extended Source Catalog containing positions, magnitudes, and basic shape information for nearly 1.6 million resolved sources, and an Image Atlas containing nearly 5 million J, H, and K_s images covering the survey area. Sources with J, H, K_s magnitudes fainter than 15.8, 15.1, and 14.3 were obtained over most of the sky and positional accuracies of about 0.1 seconds of arc were achieved.

2MASS was extremely important because the longer near-infrared wavelengths yielded an unprecedented view of the Milky Way, free from the obscuring effects of interstellar dust, which dims and reddens stars at visible wavelengths. This infrared survey thus revealed the true distribution of luminous mass over the entire span of our Galaxy. It was also the first all-sky photometric census of galaxies brighter than $K_s = 14.5$ mag, including galaxies in the 60°-wide "Zone of Avoidance", where dust within the Milky Way renders optical galaxy surveys incomplete. The catalog of >1,000,000 galaxies provides a rich statistical data base, including photometric measurements and structural parameters for large samples of galaxies in differing environments, measured at wavelengths which are sensitive to the stellar populations dominating the luminous mass. 2MASS also provided the statistical basis to search for rare—but important—objects, which are either low-temperature and thus

extremely red (e.g., extremely low luminosity stars and brown dwarfs), or heavily obscured at optical wavelengths (e.g., dust-obscured active galactic nuclei and globular clusters located in the Galactic plane). There is an impressive showcase of 2MASS images available online, especially the three-color composite visualizations, and one needs to remember that although the colors used are blue, green, and red, the wavelengths being rendered are over three times longer and well beyond human vision. Among the many discoveries from 2MASS are hundreds of very cool objects of sub-stellar mass (<78 Jupiter masses) called "brown dwarfs" with some so cold as to have atmospheres rich in methane gas like Jupiter and Saturn. Two new spectral types were added to the long-standing list of stellar spectral classes to accommodate these objects. In addition, new star-forming regions and previously unknown galaxies were revealed by the survey's ability to probe the dusty plane of the Milky Way, and for that same reason we now have a more accurate census of the number of galaxies in our local Universe. From 1996 to 2001 a similar survey was carried out independently and entirely in the southern hemisphere by the European Deep Near-Infrared Survey (DENIS) team. For this survey two near-infrared bands (J and K) were used together with a red optical band (I) at 0.8 μm, and the telescope was a 1.0 m reflector at La Silla, Chile. The DENIS catalog consists of over 355 million point sources detected in 3,662 strips each covering 30 degrees in declination and 12 arcminutes in right ascension, or about 16,700 square degrees. This survey combines features of SDSS and 2MASS by using two infrared bands and one optical band and the additional leverage in wavelength is powerful in detecting the reddest objects.

At longer infrared wavelengths, observations from space are required. In the 1980s, the Infrared Astronomical Satellite (IRAS) provided a low-resolution all-sky survey at 12 μm, 20 μm, 60 μm, and 100 μm which still remains useful. This survey led to the realization that faint emitting clouds, called interstellar cirrus, covered much of the Galaxy, and that some distant galaxies were extremely luminous in the far infrared but dimmed by dust at shorter wavelengths. As illustrated by the "Multi-Wavelength Sky" images compiled by Astronomical Picture of the Day co-guru Jerry Bonnell (web site reference given), we now have all-sky surveys from radio (73.5 cm) to gamma rays ($<1.2 \times 10^{-12}$ cm).

4.1.5 Deep imaging in selected fields

The well-known Hubble Deep Field (HDF) is an example of a digital image of a very small patch of sky (near the Big Dipper) observed with an extremely long exposure (10 consecutive days or about 150 orbits) to reach the faintest possible magnitudes. As it is so well known, no reproduction is given here. Most of the galaxies in the HDF are ~30th magnitude in the visible, or about four billion times fainter than the detection limits of the human eye. Conceived by former Space Telescope director Bob Williams and using most of his own discretionary time, this famous Hubble Space Telescope (HST) image was assembled from 342 separate CCD exposures using the Wide Field and Planetary Camera 2 (WFPC2) from December 18–28, 1995. A patch of sky was chosen well away from the plane of the Milky Way, with almost no foreground stars and very little foreground dust extinction (based on IRAS

results). The location had to lie in HST's northern continuous viewing zone at a declination near 62° (12 h 36 min 49.40 s, +62° 12′ 58.00″ J2000 Equinox) to probe into the depths of the extra-galactic (outside of our galaxy) Universe. Images of the Hubble Deep Field (HDF) are published in color because they were obtained using different filters. Four broad filters were used to cover the ultraviolet to near-infrared, F300W, F450W, F606W, and F814W, where the numbers are the central wavelength of the filter in nanometers. To make the three-color representation the two red filters were combined. About 35 hours was spent on each band except for the 300 nm filter which required 50 hours. This narrow pencil beam should be representative of the Universe as a whole. The HDF represents a time tunnel because some of the 1,500 galaxies are relatively nearby while others are at immense distances (billions of lightyears), and therefore we see them not as they are now but as they were when the light first left them billions of years ago. Many of the objects in the HDF can be followed up using spectrographs on much larger ground-based telescopes such as the Keck telescopes. Other "ultra" deep fields followed the HDF, and every new facility across the entire electromagnetic spectrum on the ground or in space, with the ability to contribute an image or a spectrum, attacked the same regions. Thus, Williams' foresight resulted in an unprecedented wealth of knowledge surrounding this one field.

4.1.6 Diffraction-limited imaging

Powerful telescopes in space can provide diffraction-limited images from the near-UV to the far-infrared, but the advent of adaptive optics on large ground-based telescopes has changed this advantage because these telescopes are much bigger than those in space, and angular resolution goes as λ/D. Ground-based diffraction-limited imaging in the near-infrared has had an enormous impact on many areas of study ranging widely from non-spacecraft studies of the outer planets and moons of the solar system to the orbital motions of stars close to the black hole at the center of the Milky Way. One of the most impressive AO results to date is the astrometry carried out by two separate teams of the motion of stars located close to the physical center of our Galaxy.

Two separate groups, one led by Andrea Ghez at the University of California, Los Angeles and the other by Reinhard Genzel at the Max Planck Institute for Extraterrestrial Physics, Germany have been using separate adaptive optics facilities at the Keck and ESO/VLT observatories, respectively, since the mid-1990s to follow the motions of individual stars near the Galactic Center. The center of the Milky Way is approximately 8,000 parsecs or 26,000 lightyears from the Sun and shrouded from visible sight by 30 magnitudes of extinction from interstellar dust and gas clouds. At near-infrared wavelengths, however, the extinction is much less, only a few magnitudes, allowing infrared cameras to "see" all the way to the center of the Galaxy. Of course, the region is crowded and it is difficult to distinguish individual stars in seeing-limited images. Using adaptive optics to overcome atmospheric turbulence and the largest possible telescopes to get the smallest angular resolution, the picture changes and it becomes possible to resolve regions that are only about five

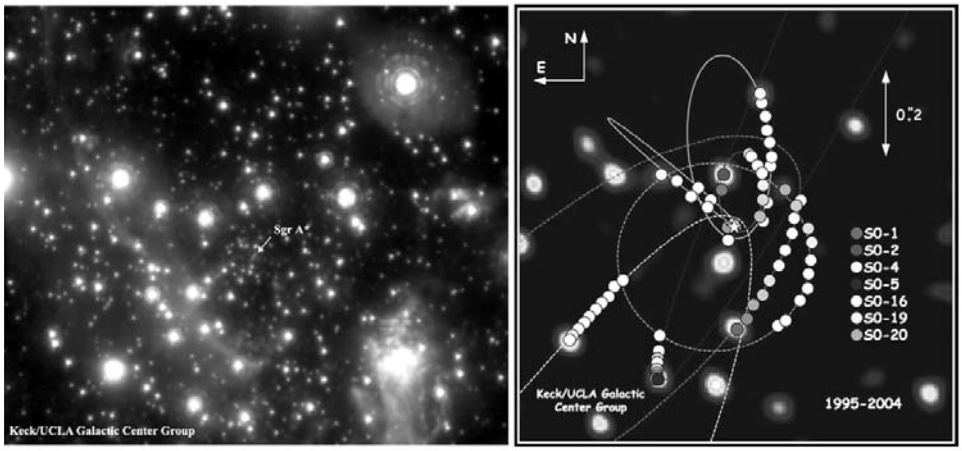


Figure 4.8. Left: a diffraction-limited infrared image of the Galactic Center using laser guide star adaptive optics on the Keck II telescope (see also Plate 10); the image is 10×10 arcsec. Right: the orbits of stars revolving around the central black hole; scale 1×1 arcsec centered on Sgr A*. Credit: Andrea Ghez.

times the size of the solar system despite the enormous distance to the Galactic Center. There is no bright object at what appears to be the geometric center of the galaxy, but there is a point-like radio source called Sagittarius A-star (Sgr A*) and numerous fast-moving stars with tight orbits around some unseen mass at that location. Stunning AO images like the one shown in Figure 4.8 (see also Plate 10) can be used to track the motions of these stars, and then the classical orbital mechanics of Newton can be used to derive the enclosed mass; remember that this is an infrared image and that no visible light reaches us from the Galactic Center. Both groups agree that the mass driving those motions appears to be $\sim$4 million times that of the Sun and yet only a very faint and occasionally variable source is seen there, leading to the conclusion that a black hole resides at the center of our galaxy (Genzel *et al.*, 2003; Ghez *et al.*, 2005).

4.1.7 Interferometers; expanding the baseline

Optical/infrared interferometers have made contributions to the study of stellar diameters and binary star orbits, but the new generation of instruments in the optical and infrared, especially those associated with large telescopes, are now contributing to non-stellar science by mapping the pre-planetary accretion disks around young stars and probing the structure of the black hole accretion disks in nearby active galaxies. These techniques are in their infancy and rapidly improving. A review of the state of the art up through 2003 is given by Monnier (2003). At radio wavelengths, however, interferometry is a technique going back to the early years of the subject, and there are today many different arrays of antennas in operation, including trans-oceanic links to establish Very Long Baseline Interferometry (VLBI). This technique

achieves incredible angular resolution ($\sim \lambda/b$ where b is the baseline), with the consequent impact on astrometry already mentioned (see Chapter 13 for further discussion of radio imaging).

4.2 SPECTROSCOPY; ATOMIC FINGERPRINTS

4.2.1 Introduction

Spectrometers provide astronomers with fundamental physical information on the chemical composition, temperatures, densities, and velocities of remote objects in the Universe. To record an image of the spectrum, almost all astronomical spectrometers use CCD detectors. Usually, the image of the spectrum itself is not as useful as the "extracted" plot of intensity vs. position (wavelength) in the spectrum. The term "spectrometer" is a generic term for any device which "measures" the spectrum. Astronomers always use spectrographs (i.e., spectrometers which contain an imaging device to make a record of the spectrum). Note that this is not the same as an image of the scene at the telescope focus. Instruments which can record both the normal two-dimensional image of the scene *and* the one-dimensional wavelength spectrum are called imaging spectrometers or sometimes "3-D spectrometers". When a non-imaging detector such as a photomultiplier tube is used, the spectrum can only be recorded by either scanning the detector along the spectrum in small steps or scanning the image of the spectrum over the detector. Such instruments are called scanning monochromators. Spectrographs fall into several classes depending on the amount of fine detail or spectral "resolution" achieved. The resolving power (R) is defined by the ratio of the wavelength (λ) divided by the smallest discernible change in wavelength $\Delta\lambda$ (delta lambda); $R = \lambda/\Delta\lambda$. Values of R range from a few hundred to a few hundred thousand. These classes are roughly

- faint object spectrographs, low resolution ($R \sim 500$)
- intermediate dispersion spectrographs ($R \sim 5{,}000$)
- high-resolution spectrographs ($R > 25{,}000$)
- imaging spectrometers (depends on technique).

To "disperse" or spread the incident light beam into a spectrum some instruments employ a prism, but most astronomical spectrographs use a diffraction grating. The principle of the diffraction grating relies on the wave nature of light. Thousands of very fine parallel lines or grooves are cut in an optical substrate, and reflected (or transmitted) light waves "interfere" to produce a spectrum. As the beam of light is spread out into a spectrum the less bright any part of the spectrum becomes, and the more difficult it is for a detector to make an accurate measurement. If the signal-to-noise ratio goes as $\sqrt{N}$ but the number of photoelectrons recorded (N) scales roughly as the width ($\Delta\lambda$) of the spectral band admitted, then the difference between a broadband filter for imaging where $\lambda/\Delta\lambda = 5$ and a spectrometer where $\lambda/\Delta\lambda = 5{,}000$ implies a large degradation ($\sim\sqrt{1{,}000} = 31.6$) unless compensated

by much longer exposures or larger collecting area. Consequently, for very faint sources the performance of the detector often becomes the limiting factor.

Astronomical spectra are typically of three types: continuous, emission, or absorption. Hot incandescent lamps such as a tungsten filament produce a continuous spectrum (all wavelengths) with a spectral energy distribution like that of a black body (thermal) source. The highly ionized interior of the Sun is a continuum source. The wavelength of peak emission is determined by the temperature. When this radiation passes through less dense and cooler plasma, then photons with specific energies (wavelengths) can be absorbed depending on the chemical constitution of the gas. The outer layers of the Sun produce such an effect so that the emergent solar spectrum is continuous, but with numerous missing wavelengths due to absorption by hydrogen, helium, and a host of heavier elements, typically called "metals" by astronomers. When passed through the rectangular entrance slit of a spectrometer and dispersed into an image, the observed spectrum will be seen to have dark lines. The photon energy absorbed by the intervening gas will usually be re-emitted as photons of the same wavelength but traveling in random directions. An emitting cloud viewed from the side without projection against the continuous source will therefore show emission lines at these wavelengths, the identical wavelengths that were absorbed. An emission-line spectrum can be seen in extended objects like the solar corona, planetary nebulae, and reflection nebulae. Each species of atom produces a distinct pattern of lines, whether seen in emission or absorption. The natural width of a spectral line is very narrow, but many processes can broaden spectral lines, such as random thermal motions of the emitting atoms, rotation of the entire star, pressure (collisions) in the stellar atmosphere, and effects due to strong magnetic fields. If the spectral resolution (R) is low, then the line will be "unresolved" and its apparent width will be determined by instrumental effects. At visible wavelengths, many lines in the solar spectrum can be resolved when $R = 10{,}000$.

When photographic spectrometers were first introduced into astronomy by Henry Draper and others (Chapter 1), and the spectra of stars first examined by eye, the meaning of the dark lines was not fully understood. Stellar spectra were classified with a letter based on the appearance or non-appearance of certain lines and on the apparent width of the lines. When it was realized that the spectral appearance depended on the temperature and ionization state, then the letters were simplified and rearranged into the familiar spectral types OBAFGKM, from hottest to coolest, with stars of Type M being around 3,000 K and stars of Type O closer to 50,000 K. The Sun is a Type G with a Sub-type 2 (G2) corresponding to its surface temperature of about 5,800 K. The conclusive discovery of brown dwarfs in 1995 led to two further letters (L and T) in recent years (Kirkpatrick, 2005) to describe much cooler objects. Brown dwarfs are characterized as shown in Figure 4.9 by strong *infrared* absorption bands due to superheated steam at 1.35 μm and 1.8 μm. As the temperature decreases further (L5–T1), the steam bands get deeper. Then, additional absorption due to methane occurs at 1.67 μm and 2.2 μm and the objects become more like the gas giant planets of our solar system (McLean *et al.*, 2001). Brown dwarfs are extremely faint at visible wavelengths and required infrared surveys to find them.

The measured wavelength of a given spectral line depends on the velocity of the

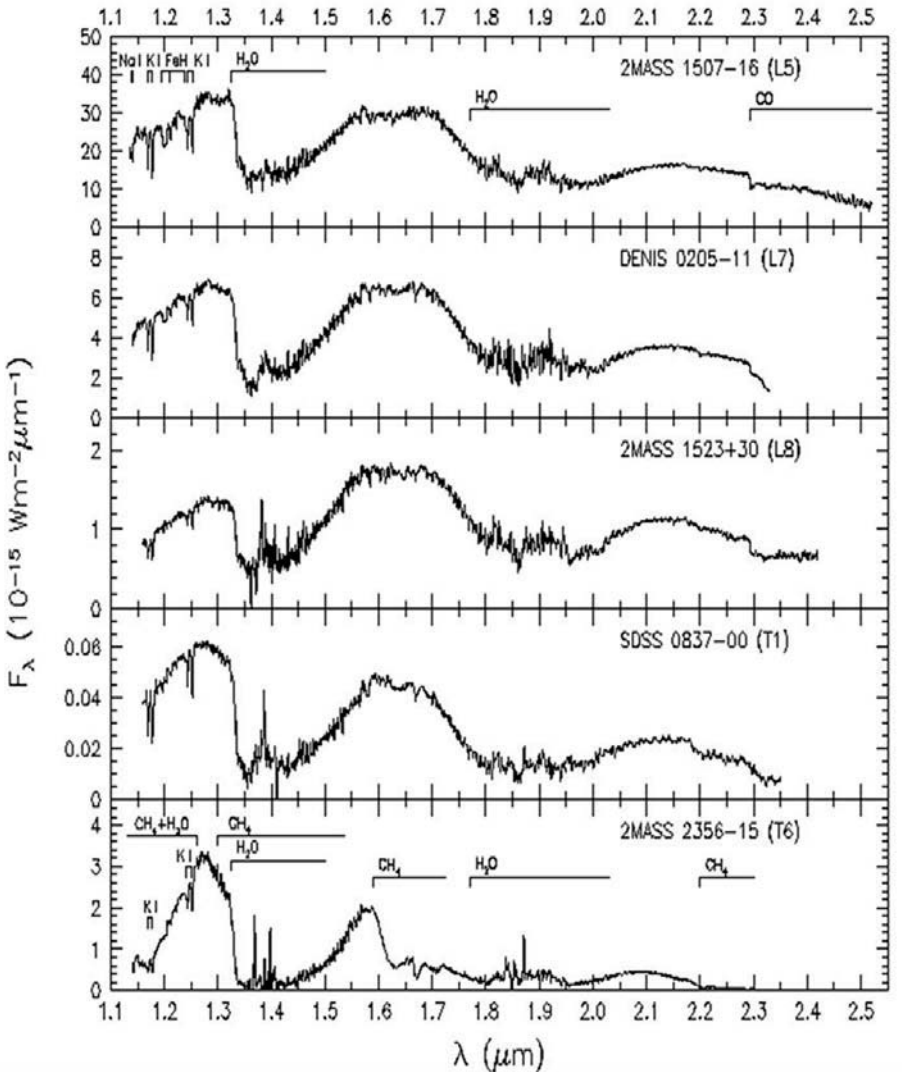


Figure 4.9. A series of infrared spectra from 1.1 μm to 2.5 μm obtained with the NIRSPEC instrument developed by the author for the Keck II telescope, showing the properties of recently discovered brown dwarfs with spectral types L and T.

emitting atoms. If the emitting source is moving towards the observer then the line appears shifted to the blue (higher frequency), and if the source is moving away then the shift is to the red (lower frequency). This is the well-known Doppler effect, which also applies to sound waves. For atomic speeds much less than the speed of light, the component of the velocity (V) of the emitting object along the line of sight (the radial velocity) is given by $V/c = [\lambda_{obs} - \lambda_{em}]/\lambda_{em}$, where the suffixes *obs* and *em* indicate

the observed and emitted wavelengths. Basically, the smallest velocity that can be detected corresponds to matching the Doppler change in wavelength to the resolution element of the spectrometer (i.e., $V/c = \Delta\lambda/\lambda = 1/R$); in practice $\sim 0.1\,\Delta\lambda$ is possible. Taking $c = 300{,}000$ km/s then $R = 10{,}000$ is sufficient to detect a motion of 30 km/s (67,000 miles per hour). Interestingly, this is the orbital speed of the Earth in its motion around the Sun, whereas the Sun's motion around the Galactic Center is closer to 250 km/s. The explosive stellar winds from some hot stars exceed 1,000 km/s, and the apparent redshifted spectral lines of *all* distant galaxies caused by the expansion of the Universe results in even greater velocities. For cosmological studies it is customary to define the "redshift" factor $z = [\lambda_{\rm obs} - \lambda_{\rm em}]/\lambda_{\rm em}$, or $(1 + z) = \lambda_{\rm obs}/\lambda_{\rm em}$, but this is not a Doppler effect. In an expanding Universe, distant objects appear redshifted because of the expansion of space itself, and $z = H_0(d/c)$ for small distances (d), as discovered long ago by Edwin Hubble. In this expression H_0 (H-naught) is the Hubble constant measured at the present epoch. As the redshift is due to the expansion of space, values of $z > 1$ are permitted. Astronomers prefer to work with z because $1/(1 + z)$ is proportional to the size of the Universe at that epoch; an object with $z = 3$, for instance, is being observed when the Universe was one-fourth its present size. The actual recession velocity of the distant source depends on the cosmological model of the Universe (based on general relativity). For a quasar with $z = 6.3$, for example, a flat Universe with the fraction of normal and dark matter at 30% yields the correct faster-than-light velocity of 585,611 km/s. Using the special relativistic Doppler formula which gives $1 + z = \sqrt{[(1 + V/c)/(1 - V/c)]}$ would result in 285,254 km/s, which is sub-light but incorrect in this application because the large redshift is not caused by the emitting atoms moving at speeds near that of light. For detailed insight into all sorts of cosmological facts, including the cosmic microwave background, I recommend the web site of my colleague Ned Wright at UCLA. Note that spectral resolving powers $\sim$1,000 are sufficient to measure z to an accuracy of one part in a thousand or 0.1%.

Thus, spectrometers have three basic functions; namely, to identify the chemical constituents and measure their abundance, to determine the temperature (and other physical parameters), and if possible to determine the radial velocity of the object. What forms do these instruments take?

4.2.2 High resolution, from cosmic abundances to planet hunting

The classical spectrograph has a single, long, narrow slit which is ideal for studies of nebulae and galaxies, and which gives good sky subtraction for single objects. Of course, if the slit is long enough and the field dense enough then there may be more than one star contained in the slit length. This was the standard configuration for many years until the schemes described in Section 4.2.4 were devised to observe many more objects simultaneously. High-resolution spectrographs tend to be very large instruments. Consequently, they are usually located at a stationary focus of the telescope, where the image position remains fixed and independent of the telescope axis (e.g., coudé and Nasmyth; Chapter 3). Some of these instruments are so large that they effectively occupy a small room, and you can literally walk "inside" the

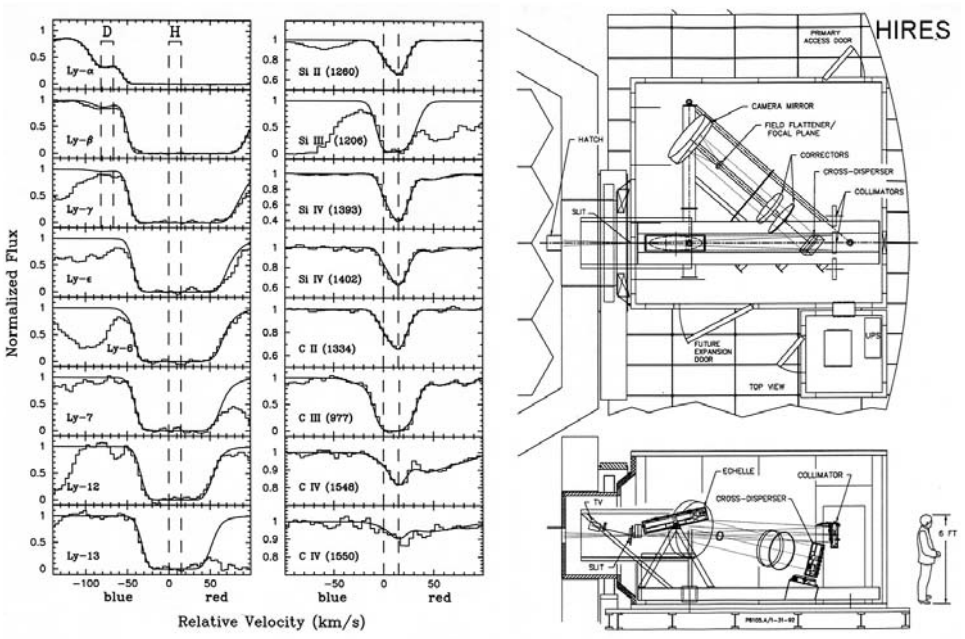


Figure 4.10. Left: a high-resolution spectrum of a quasar revealing absorption due to deuterium (D). Credit: David Tytler. Right: the layout of the High Resolution Echelle Spectrograph (HIRES) on the Keck telescope. This instrument is contained in a large room. Credit: Steve Vogt.

instrument, although you would not normally be allowed to do so in order to keep the optics safe and clean. One example of this class is the HIRES (High Resolution Spectrograph) on the Keck I telescope (Figure 4.10, right), which was developed by Steve Vogt and colleagues of the University of California, Santa Cruz.

Among the many applications of such an instrument is the study of the subtle spectral differences between chemical isotopes, in particular hydrogen and deuterium, in distant galaxies in order to determine their primordial abundance (a typical spectrum is also shown in Figure 4.10). Standard Big Bang nucleosynthesis predicts that a measurement of the primordial deuterium-to-hydrogen (D/H) abundance ratio will give a sensitive estimate of the ratio of baryons (ordinary matter) to photons in the Universe. Using the cosmic microwave background to derive the density of photons then yields the density of baryonic matter. Distant quasars can provide the luminous sources to probe the hydrogen gas clouds in the outer halos of unseen galaxies in the high-redshift Universe. The strongest spectral line is usually the ultraviolet Lyman-alpha line of normal hydrogen, but a weak line of deuterium should appear 82 km/s to the blue. For redshifts $z > 2.5$ the Lyman series moves from the UV to the visible. To properly resolve the weak deuterium line requires high resolution ($R = 30{,}000$–$60{,}000$) and a very big telescope. The first observations of this kind using the Keck telescope were made by Tytler, Fan, and Burles (1996). By 2003

D/H had been measured to better than 10% and the baryon density was also confirmed by WMAP (Section 13.3).

A contrasting application of high-resolution spectroscopy came to everyone's attention on October 5th, 1995 when the first planet to be found orbiting another star was announced. The planet is in orbit around the 5.5-magnitude G5 V star 51 Pegasi which lies about 15.4 pc from Earth. Using the ELODIE high-resolution echelle spectrograph at the Geneva Observatory, Swiss astronomers Michele Mayor and Didier Queloz made the announcement at the *Ninth Cambridge Workshop on Cool Stars, Stellar Systems and the Sun* held in Florence, Italy (Mayor and Queloz, 1995). One week later California astronomers Geoff Marcy and Paul Butler added 51 Peg to their own planet search program using the high-resolution (Hamilton) echelle spectrometer at Lick Observatory (California), also built by Steve Vogt (UCSC), and quickly confirmed the detection of an object with about half the mass of Jupiter. What was remarkable about this find was not the large reflex motion of the star in response to the planet of ± 57 m/s, but the exceedingly short period of 4.2 days and therefore the small orbital radius of 0.05 AU, ten times closer than Mercury is to the Sun (Figure 4.11, left). It was not long before Marcy and Butler added many more "extrasolar" planets to 51 Peg, and other groups developed similar instruments to join the hunt. The Swiss team began a southern hemisphere program using an instrument called CORALIE, and a similar study was initiated at the Anglo-Australian Telescope by Chris Tinney and colleagues. In the mean time other groups initiated or re-vamped programs, and Geoff Marcy and Paul Butler extended their work at the 3 m telescope by moving to the 10 m Keck telescope and the HIRES echelle spectrometer already mentioned. By 2007 well over 200 extrasolar planetary systems were known. The California and Carnegie Planet Search (CCPS) web site

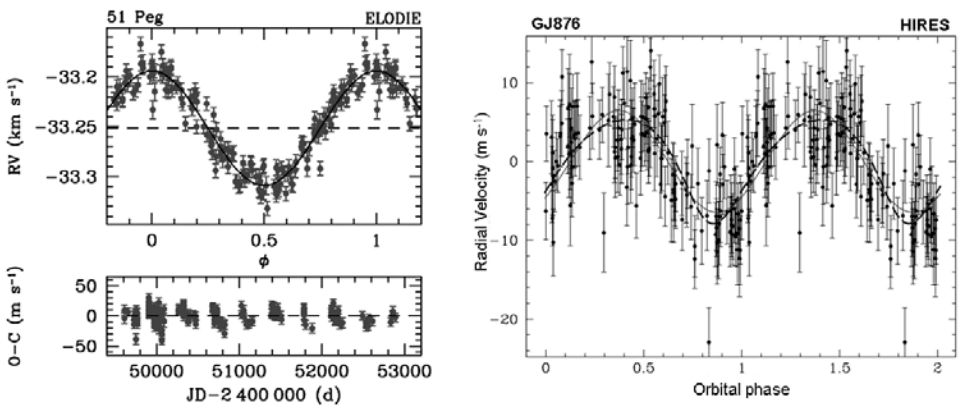


Figure 4.11. Left: the radial velocity curve of 51 Peg obtained using the ELODIE spectrograph by Mayor and Queloz. Right: a radial velocity curve obtained with HIRES on Keck I, showing the Doppler reflex motion of GJ 876 due to a planet with a mass of at least 7.5 Earth masses. Credit: California and Carnegie Planet Search.

developed by Geoff and Paul is an excellent resource for information on the latest results.

The basis of these measurements is the reflex motion on the star being orbited; the star itself is forced to orbit about the common center of mass, called the barycenter, of the star–planet system. From Kepler's Third Law in Newton's form we have that

$$a^3 = (GM_{\text{star}}/4\pi^2)P^2 \tag{4.1}$$

where M_{star} is the mass of the star; a and P are the semi-major axis and period of the planet's orbit, respectively; and G is Newton's gravitational constant. If a is in AU, P in years, and M_{star} in solar masses then $G/4\pi^2 = 1$. By observing the Doppler shifts of the spectral lines in the star caused by its motion around the barycenter we can analyze the shifts for periodicity and find P. Because the type of star is well-known (mostly solar type and carefully selected to have a known parallax from the Hipparcos satellite) then M_{star} is given. Thus, the size of the orbit (a) is found. The average velocity of the planet in its orbit can then be determined from $V_{\text{pl}} = \sqrt{(GM_{\text{star}}/a)}$. Next, applying conservation of angular momentum to this two-body system requires that $M_{\text{pl}}V_{\text{pl}} = M_{\text{star}}V_{\text{star}}$ where M_{pl} is the mass of the planet and V_{star} the average velocity of the star in its small orbit around the common center of mass. It is V_{star} that is responsible for the observed Doppler shift because for part of the orbit the star is moving towards us and for part of the orbit it is moving away from us. A plot of the Doppler shifts converted to velocities ($= c\,\Delta\lambda/\lambda$) does not give V_{star}, however—but its projection onto the sky $K = V_{\text{star}} \sin i$ where i is the (unknown) angle of inclination of the orbit pole with respect to our line of sight. If $i = 90^\circ (\sin 90^\circ = 1)$ then the orbital plane is edge-on to us (making transit observations possible) and $K = V_{\text{star}}$. Substituting $V_{\text{star}} = K/\sin i$ into the momentum equation gives

$$M_{\text{pl}} \sin i = (M_{\text{star}}K)/V_{\text{pl}} = 11.2M_{\text{earth}}[M_{\text{star}}/M_{\text{sun}}][K/1\ \text{m s}^{-1}][30\ \text{km s}^{-1}/V_{\text{pl}}] \tag{4.2}$$

Thus, the derived quantity is $M_{\text{pl}} \sin i$, and assuming $i = 90^\circ$ gives the "minimum" mass for the planet. If the orbital pole is only 30° tilted to our line of sight then the true mass is doubled (because $\sin 30^\circ = 0.5$). The extreme right side of the equation evaluates the expression in terms of "earth masses" when the mass of the star is $1M_{\text{sun}}$, the limiting velocity precision is 1 m/s, and the planet's velocity is 30 km/s; to detect the analog of Earth requires about 0.1 m/s.

Spectroscopic measurements like this are challenging. The Sun orbits its center of mass in common with Jupiter with a speed of only 14.7 m/s. This speed is many times smaller than the errors of conventional velocity measurements of the Doppler shift. For instance, even for a spectrograph with a resolving power of 100,000 then $\Delta V = 3$ km/s—not 3 m/s as required to detect a Jupiter-mass planet at 5 AU. How can such a small effect be observed? Precise calibration of the wavelength scale is obtained by allowing the incoming starlight to pass through a chamber containing an almost translucent gas (usually iodine) which absorbs a small amount of light at a few very specific places in the spectrum. Positional accuracies can be determined to a tiny fraction of a spectral resolution element ($\sim 0.1\ \Delta\lambda$). In addition, thousands of lines are

utilized in a cross-correlation technique to improve accuracy even further. The HIRES instrument on Keck routinely achieves residual errors of ± 1 m/s.

The discovery of hundreds of other planetary systems, many of them completely different from our own with giant planets orbiting close to the star, is certainly one of the greatest achievements of the late 20th century, and one which is directly attributable to advances in technology. One example of radial velocity planet detection is GJ876 (Rivera *et al.*, 2005) which has a 7.5 Earth-mass planet discovered with the HIRES spectrometer on the Keck telescope (Figure 4.11, right). The three large CCDs in this spectrometer lie behind a field-flattening lens which is required because the "camera" in the HIRES instrument is a Schmidt telescope.

4.2.3 Medium-resolution and low-resolution spectroscopy

Intermediate-resolution spectrographs are the workhorses of astronomy and are in use all over the world. Most telescopes larger than about 2 m will have an intermediate-resolution spectrograph ($R \sim 5{,}000$). A major fraction of the "dark time" (i.e., the period of the month around new Moon when the sky is darkest) is devoted to obtaining spectra of faint objects with these instruments. Many objects are of such low brightness that the light of the darkest night sky still dominates the raw picture. Computer processing is required to measure and remove the signature of the background sky from the spectrum of the faint object.

In 1982 John Beverley (Bev) Oke (1928–2004) pioneered the concept of a "double" spectrograph for the 5 m Hale Telescope at Palomar, driven in large part by the huge spectral range of the CCD. Many Cassegrain spectrographs are now "double" systems. One example is the LRIS (Low Resolution Imaging Spectrograph) on the Keck telescope which was developed initially by Oke and his Caltech colleague Judith Cohen (Oke *et al.*, 1995). Jim McCarthy and Chuck Steidel of Caltech later took responsibility for adding the blue channel (Steidel *et al.*, 2004). The layout of the double spectrograph is shown in Figure 4.12. Because the spectral range of a CCD spans 0.3 μm to 1.1 μm it is more efficient to use a special beam-splitter to separate the blue and red parts of this range and then use optimized cameras and CCDs to record the blue and red spectra. Instruments like this are often used to study the most distant objects in the Universe, such as galaxies and quasars at high redshift, supernovae, and the after-glow of gamma-ray burst counterparts. Figure 4.13 shows the LRIS-B spectrum of Q1307−BM1163, a 22nd-magnitude galaxy at $z = 1.411$, obtained with this instrument. The plot gives raw counts vs. wavelength, specifically to show the relative count rate as a function of wavelength from 300 nm to 710 nm (3,000 Å–7,100 Å), and is the sum of three 1,800 s exposures with a UV-sensitive CCD, actually a pair of 2 K × 4 K devices from *e2v technologies*. A trace like this one is produced easily from the digital CCD image of the spectrum by a computer program which plots the signal along a row of pixels. Because the redshift is $z = 1.411$, the identified absorption features along the base are actually observed at 2.411 times their true wavelength. For example, the C II line at the left observed at 321.6 nm is really the UV line at 133.4 nm, and the Mg II doublet observed at 674.1 nm at the right is really redshifted from 279.6 nm. These features are caused by interstellar absorption in the

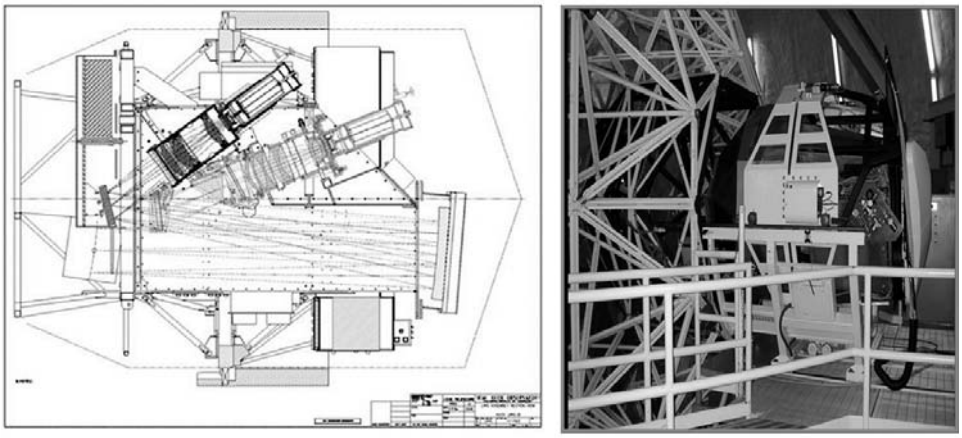


Figure 4.12. The layout of the LRIS double-spectrograph for the Keck I telescope. CCD cameras are attached at two places. The instrument is shown on the right during installation. Credit: California Institute of Technology and Keck Observatory.

distant host galaxy. It can be deduced that this galaxy is forming stars at the rate of $\sim$30 solar masses per year and that the elements heavier than hydrogen and helium have close to the same abundance as our Sun. Viewed at a look-back time of $\sim$10 billion years, this galaxy is turning gas into stars and enriching its interstellar medium at a much faster rate than ever experienced by the Milky Way. Its fate is likely to be an elliptical galaxy devoid of gas or perhaps the bulge of a massive spiral galaxy.

The "faint object" class of spectrographs is basically a CCD camera capable of imaging a spectrum of a faint object with low spectral resolution so as not to "spread out" the available light too much. The less the spectrum is spread out, the more light there will be on any given pixel of the CCD and the fainter the source that can be

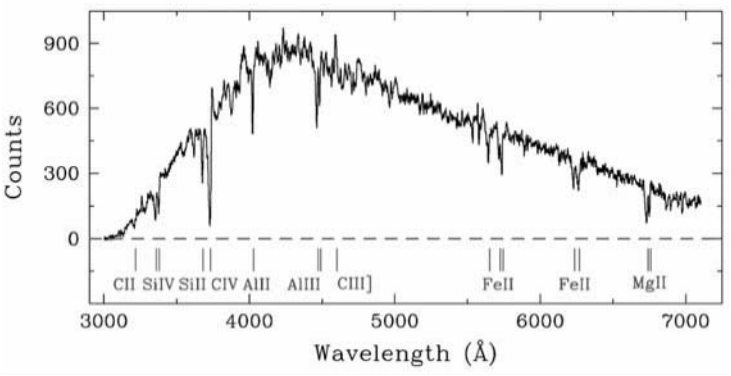


Figure 4.13. A spectrum of the ultraviolet absorption features in the galaxy Q1307−BM1163 shifted into the visible by its redshift of $z = 1.411$. The observations were made using the blue arm of the LRIS spectrograph on Keck I. Credit: Chuck Steidel.

detected. Obviously a compromise is needed which ensures that enough spectral resolution is retained, otherwise features or lines in the spectrum which could yield the velocity of the object, for example, will be hopelessly smeared out.

4.2.4 Multiobject spectroscopic surveys; 3-D maps of the Universe

Spectroscopic measurements can be painfully slow, even with very sensitive CCD detectors. Typically, the signal at any point in the spectrum of an astronomical source will be at least several hundred times fainter than the signal in a direct image. Thus, the possibility of recording spectra from several objects at once is very attractive.

Slit-less spectroscopy and objective prism spectroscopy are two similar ways to observe the spectra of many objects at the same time. Removing the slit is okay for point sources, the spectral resolution then being determined by the seeing disk rather than the slit width, but this method only works at wavelengths where the sky background is very dark and the field not too dense to avoid overlapping spectra. Because the sky is an extended source it contributes an integrated (broadband) background which is superimposed on the spectrum of the point source, thus reducing contrast compared with a slit spectrograph. The simplest slit-less methodology is to place a thin prism over the entrance aperture to a small telescope so that all the star images become little spectra, an idea that dates back to Alex-Marie Rochon around 1770. One famous objective prism survey is the Henry Draper (HD) Catalog of 225,300 sources produced by Annie Jump Cannon (1863–1941) and Edward C. Pickering at Harvard from about 1893 to 1924 using photographic plates inspected by eye. Modern applications include objective prism surveys with large Schmidt telescopes, like the U.K. Schmidt and others, to look for emission-line objects such as quasars. The Hubble Space Telescope's Advanced Camera for Surveys (ACS) and Near-Infrared Camera and Multi-Object Spectrometer (NICMOS) do not have entrance slits but can still capitalize on the low backgrounds in space by making slit-less spectral images using internal dispersing elements called grisms, which combine the properties of a grating and a prism (see Chapter 5).

Another way to solve the problem is the multi-object spectrograph which employs an entrance slit composed of multiple sub-sections which can be positioned by computer to pick up many different objects in the field of view. Today, many observatories have a spectrograph with a "multi-slit" system. Figure 4.14 shows the raw CCD image obtained with the DEIMOS multi-object spectrograph on the Keck II telescope, together with part of a "slit-mask" and a close-up of one of the machined slits. Slits are cut at different angles and in different locations on the mask and therefore the spectra do not line up perfectly on the detector. Emission lines that fill the entire length of a slit are due to night-sky OH lines. DEIMOS carries up to 11 masks in a cassette and each mask can have about 400 slitlets spanning a field of view of about 16.7×5.0 arcminutes on the sky. The spectral resolution is typically around 5,000–10,000 and a mosaic of eight CCDs provides a camera with 64 megapixels fed by $f/1.29$ optics. Among the many studies carried out by this instrument is the DEEP2 redshift survey to obtain spectra of $\sim$50,000 faint galaxies with $z > 0.7$ in order to study the evolution of galaxy properties and their tendency to cluster,

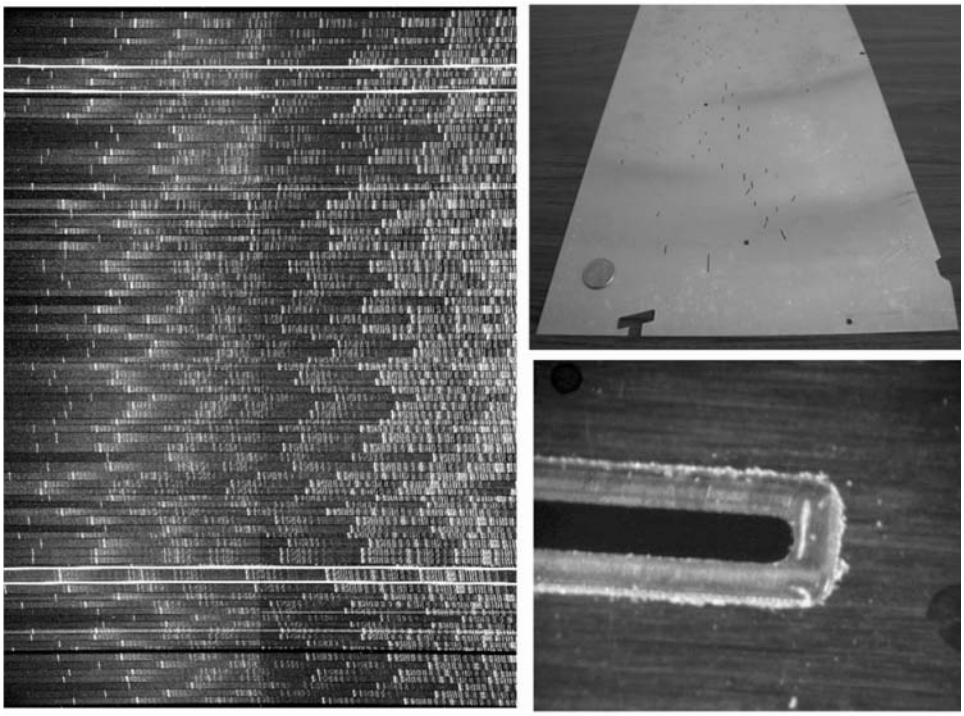


Figure 4.14. Multi-slit spectroscopy is illustrated here by results from the DEIMOS spectrograph on the Keck II telescope. Left: raw images of the spectra. Right: the upper figure shows a typical slit-mask and the lower image gives a close-up of a milled slit. Credit: Greg Wirth, Keck Observatory.

compared with lower redshift samples. This project is complementary to other wide-field large redshift surveys like SDSS and 2dF mentioned below by using higher spectral resolution to measure line widths and rotation curves in addition to redshift, and to the VLT/VIRMOS Deep Survey which covers a larger area but with lower spectral resolution and fewer objects at high redshift.

Slit-mask technology has also been extended to the near-infrared part of the spectrum where the challenge is increased because the mask must be cold ($<$150 K) to prevent the infrared detectors from seeing its thermal glow. Several different technologies are possible including exchangeable slit-masks that are loaded through a vacuum-cryogenic airlock system, or movable opposing slit bars that quantize the y-axis but allow any slit location in the x-direction, or micro-shutters which is an outgrowth of the general area of micro-electro-mechanical systems (MEMS) technology. MEMS devices represent the integration of mechanical elements, sensors, actuators, and electronics on a common silicon substrate. The micro-mechanical components are made using "micromachining" methods that selectively etch away parts of the silicon wafer, or add new structural layers to form the required device. Examples that have already made their way into astronomy are MEMS-based

deformable mirrors (DMs) for wavefront control in adaptive optics systems. Boston Micromachines Corporation makes a deformable mirror only 10.5 mm in size with an actuator array of 1,024 elements, each capable of delivering a stroke of 1.5 μm; apertures of twice this size with over 4,000 actuators are likely to be part of future AO systems such as the Planet Imager for the Gemini South 8 m telescope. Another MEMS technology is the "microshutter" which is being developed at the Goddard Space Flight Center for use in the Near-Infrared Spectrograph for the James Webb Space Telescope. Microshutters are tiny cells 100×200 μm in size arranged in a grid of 250×250 (62,500). All shutters are slightly magnetized and are opened by scanning a permanent magnet over the array. Selected apertures are held open electrostatically by applying a voltage difference between the shutter and an electrode on the wall. After the magnet has passed by, the resilience of the hinges flips the remaining shutters closed.

Of even greater impact on multi-object spectroscopy has been the technology of transparent optical fibers. Developed mainly for the telecommunications industry, these slim, flexible glass conduits, which resemble normal electrical cables on the outside, can be used as "light-pipes" to transmit light over very long distances with only slight losses or attenuation. The basic idea is to position one end of each of several optical fibers in the focal plane of the telescope at points corresponding to interesting objects, such as a cluster of distant galaxies, and to stack up the other ends along the entrance slit of the spectrograph; the spectrum of each source is therefore recorded simultaneously.

The first of the fiber optic coupled systems, appropriately named MEDUSA, was developed at the Steward Observatory in Arizona by John Hill, Roger Angel, and co-workers in 1979. Such systems are now fairly common on telescopes all over the world and have drastically transformed many branches of astronomy where statistical properties on many sources are required. Examples include 2dF (2-degree field), the 400-fiber system on the 3.9 m Anglo-Australian Telescope (AAT), shown in Figure 4.15, which uses robotic arms arranged around the perimeter of the field to place the fibers in the focal plane, and the 600-fiber spectrographs on the SDSS 2.5 m telescope which employ manual "plug-plates" to insert fibers into a pre-drilled mask (see Figure 4.16, top).

Optical fibers have also revitalized the photographic Schmidt telescopes in a novel way. In the early 1980s, at the Anglo-Australian (formerly U.K.) Schmidt Telescope in Australia, Fred Watson (then of the Royal Observatory, Edinburgh) pioneered an instrument called FLAIR (and FLAIR II) which was a fiber optic system (with over 90 fibers) allowing simultaneous spectroscopic observations of objects scattered over the huge $6.5^\circ \times 6.5^\circ$ field of view. Locating the fiber ends with the target objects was achieved with a specially made warped-to-shape 1 mm thick positive glass plate, rather than the usual negative plate, of the field. In a process that took 6–7 hours, fibers were glued to this plate with ultraviolet-curing optical cement and the assembled unit loaded into the position normally occupied by the photographic plate in the telescope. The bundle of fiber cables was led away and fed into the slit of a standard CCD grating spectrometer set up on an optical bench on the dome floor. Today, FLAIR has been replaced with a new system called 6dF (for 6-degree

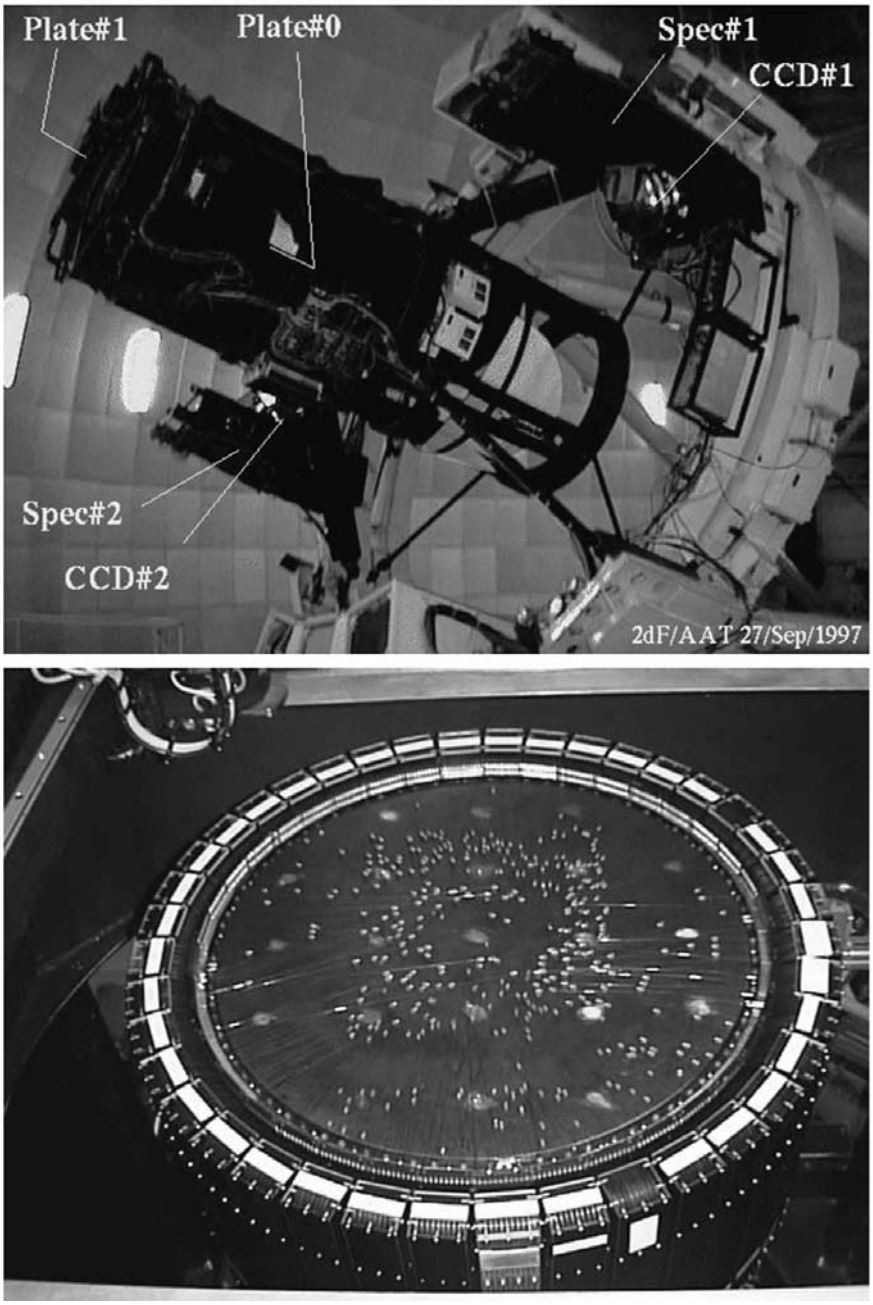


Figure 4.15. The multi-fiber system of the 2dF instrument on the Anglo-Australian Telescope (AAT). While observations are under way, a robotic mechanism configures hundreds of fibers on a second plate. When ready, the two massive units exchange positions by mechanically tumbling over. Credit: Anglo-Australian Observatory.

Figure 4.16. Two fiber systems: at the top is the "plug-plate" for 600 fibers in the Sloan Digital Sky Survey. Credit: SDSS. Bottom: a fiber-optical system called 6dF that robotically positions small prisms and fibers to a plate that is inserted into the wide-field U.K. Schmidt Telescope. Credit: AAO.

field) which is a fully automated, pick–place, magnetic button fiber-positioning system (Figure 4.15, bottom). The *r*-theta positioning robot can reconfigure up to 150 target fibers of 6.7″ (100 μm) diameter in less than one hour, and two interchangeable plates keeps one in use while the other is being configured. Each button carries a light-collecting prism attached to an optical fiber in a 10 m cable that directly feeds a spectrograph on the dome floor to give spectra with a resolution of about 0.1 nm/pixel–0.4 nm/pixel and a limiting magnitude around 18 in the blue. Correlated with the 2MASS and the DENIS infrared-imaging surveys, this instrument can provide over 100,000 galaxy spectra in the southern hemisphere (Jones *et al.*, 2004).

Using Hubble's Law, the SDSS and other redshift surveys provide the distance to over one million of the nearest galaxies, giving us a three-dimensional picture of the Universe through a volume about one hundred times larger than explored previously. With techniques like this, a complete redshift (distance) limited survey, of quasars or faint galaxies, becomes viable thus affording us the opportunity to visualize accurately the Universe in three dimensions. One example is shown in Figure 4.17.

Using spectroscopy, the SDSS has found small faint companion galaxies to the Milky Way that had gone unnoticed, and it has seen long streams of stars left behind by satellite galaxies merging with our own. These discoveries are important because one of the leading models for galaxy formation predicts that dwarf galaxy companions to the Milky Way should be more numerous than currently observed. Also,

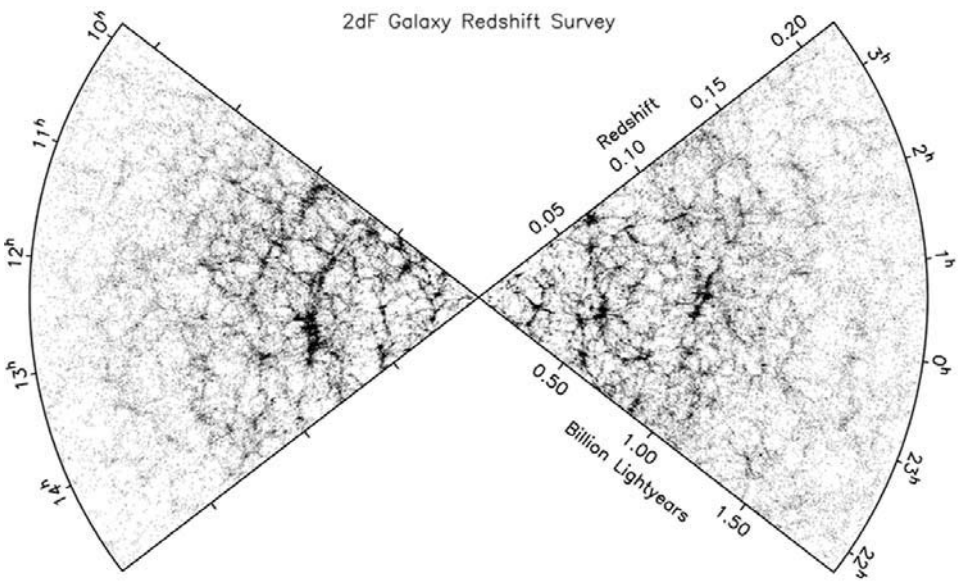


Figure 4.17. Mapping the large-scale structure of the Universe; an example of a three-dimensional plot of galaxy positions from the 2dF redshift survey. Credit: 2dF project.

using 200,000 quasars and 13 million galaxies, the Survey has revealed the expected large-scale gravitational lensing (or cosmic magnification) of distant background sources (quasars) by foreground matter, as required by Einstein's General Relativity and the standard model in which much of the mass is in the form of dark matter. From initial discoveries in the 1970s (Gregory and Thompson, 1978), it has been known that the Universe has large-scale structure in the form of enormous clusters of galaxies spread out in filaments of ever larger sizes. Using a three-dimensional map built up with more than 600,000 galaxies covering about one-tenth of the sky, the SDSS-II discovered galactic structures spanning a billion light years. These observations are consistent with dark matter and dark energy models, and with the idea of galactic structure imprinted by cosmic sound waves in the early Universe (baryon acoustic oscillations). These sound waves have a length scale of about 450 million lightyears and were first reported in 2005 by independent teams from the SDSS and the Australian-U.K. Two Degree Field Galaxy Redshift Survey (2dFGRS).

4.2.5 Imaging spectroscopy; x, y, and λ

Even the multi-object spectrograph suffers from the fact that in order to record the dispersed spectrum of any object in the field of view of the telescope the remainder of the field, which is most of it, must be hidden from the spectrometer by the slit-mask. An image of the scene cannot be recorded with high spectral resolution unless the measurements are repeated many times with slight displacements of the telescope. Scanning or stepping the telescope in a sufficiently smooth uniform way to produce a two-dimensional image of the scene composed of numerous strips, with every pixel in every strip providing a spectrum of that point in the image, places stringent requirements on the telescope motion control. Although many experiments have obtained three-dimensional data cubes (x, y, λ) this way, other technologies offer more advantages. There are two basic approaches in trying to achieve both spatial and spectral information simultaneously. One is to use interference-based techniques such as the Fabry–Perot interferometer or an imaging Michelson interferometer. The other method involves innovative optical designs called integral field units.

Developed initially by Charles Fabry (1867–1945) and Alfred Perot (1863–1925) as long ago as 1896, the technique that bears their name remained dormant until the 1950s because it was considered too difficult to manufacture and use. The Fabry–Perot "etalon" is essentially a cavity formed by two face-to-face circular plates of high reflectivity and low absorption which are held parallel and flat to a tiny fraction (typically 1/200) of the wavelength of light (i.e., a mere 4.5 nm in the mid-visible). This device has a much greater efficiency or throughput for a given spectral resolution than the normally used diffraction grating. (The throughput of a spectrometer is the "solid angle" it can accept multiplied by its "aperture" often written as $A\Omega$.) As the spacing or gap between the pair of reflective surfaces is changed, so the wavelength transmitted by the etalon changes. The beauty of the Fabry–Perot etalon lies in the fact that it behaves essentially like an ordinary wavelength selection filter and can be

placed directly into the beam of an imaging system to yield a picture of the entire field of view, with the spectral purity and detail that could otherwise only be obtained by a fairly powerful spectrometer. Fabry–Perot etalons are most useful for studies of specific spectral lines and have very limited wavelength span compared with a classical spectrograph. The Fabry–Perot (or FP for short) can be "tuned" to different wavelengths by careful adjustment of the gap between the reflective plates. This can be accomplished under computer control, and a whole series of images can be stacked up in computer memory corresponding to slightly different wavelengths or, equivalently, slightly different velocities. The resulting collection of images is known as a "data cube" because it has three "dimensions"; namely, the ordinary two spatial dimensions of a picture, and a third dimension which is the wavelength at which the picture was recorded. A plot of the signal values along the wavelength dimension for any spatial position yields the spectrum at that point in the source.

The Fabry–Perot allows only a very small range of wavelength space to be covered, but has the advantage of producing an excellent image of the field. Slit stepping or scanning can provide large wavelength coverage, but it is hard to obtain good–quality images along the scanning direction for most telescopes. Integral field spectroscopy combines some of the advantages of each of these and is the most common approach today.

One way to implement an integral field mode is to employ an image slicer. This technique was pioneered by Ira Bowen for Palomar in the 1940s, but modern machine technology makes it even more practical today. One example of this method is employed in the "3-D" integral field spectrometer developed at the Max-Planck Institute (Krabbe *et al.*, 1995). This near-infrared instrument uses a complex mirror with many tilted facets to subdivide the image in the focal plane of the telescope into narrow strips and then another similar mirror to "stack" these parts one beside the other along the length of a spectrograph slit (Figure 4.18). Thus, every region of the image produces a spectrum. The field of view that can be covered is relatively small ($8'' \times 8''$).

A completely different approach is to subdivide the focal plane image into numerous very small segments using an array of tiny lenses. This is the basic idea behind the TIGER instrument introduced at the CFHT in 1995 by Roland Bacon. First, the normal image is greatly magnified and fed to a "microlens" array, which effectively slices up the image before it goes into the spectrograph. The pattern on the CCD detector is complicated, representing both spatial and spectral information for every small segment in the field. Variations of the microlens approach have been developed for other instruments too, and extended to the near-infrared in cryogenic instruments. An alternative strategy pioneered in the ARGUS instrument at CFHT is to subdivide the focal plane with numerous, closely packed optical fibers in a two-dimensional pattern, but collect all the fibers into a one-dimensional stack which can be fed directly to the long slit of a conventional spectrograph. These techniques are now being extended to "deployable" IFUs, which is an approach that allows small integral field units to be positioned anywhere over a much larger field of view, and to the infrared.

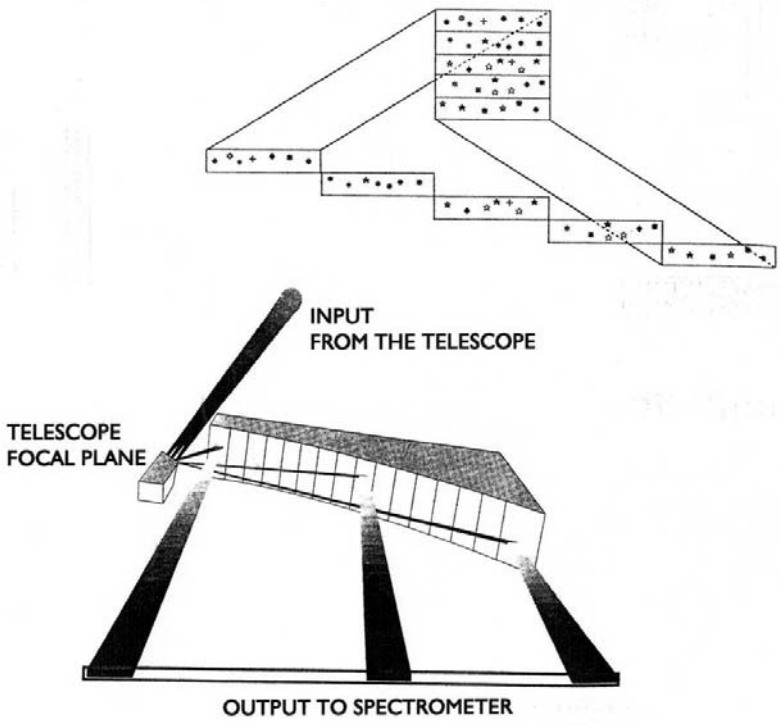


Figure 4.18. The "image-slicing" concept used in the first "3-D" near-infrared imaging spectrograph developed at the Max Planck Institute to obtain images and spectra of a field simultaneously. Credit: Alfred Krabbe.

4.3 POLARIZATION; TRANSVERSE WAVES

4.3.1 Introduction

Polarization is the term used to describe the phenomenon in which a fraction (or all) of the electromagnetic waves in a beam of light are vibrating in the same plane, such as the vertical or horizontal planes (or any orientation in between) along the line of sight to the source. If there is no preferred orientation and the beam contains electromagnetic waves having a random jumble of all planes of vibration then the beam is said to be "unpolarized". When there is a preferred plane of vibration which does not change no matter at which point in the wave cycle we sample, the beam is said to be "linearly" polarized. On the other hand, if the plane of vibration rotates by 360° through a wave cycle, then the beam is said to be "circularly" polarized (Clarke and Grainger, 1971). The light emitted by an ordinary tungsten lamp is essentially unpolarized, whereas the scattered sunlight which constitutes "blue sky" is very strongly linearly polarized (Rayleigh scattering).

Many important astrophysical phenomena produce polarization either through the interaction of unpolarized light with matter, or by the generation of polarized

light by the atoms themselves. For example, reflection from solid surfaces, scattering of photons by electrons, molecules, and small grains, and absorption by certain materials in the interstellar medium all cause polarization; atoms emitting light in the presence of a magnetic field (as in sunspots) suffer the Zeeman effect and the emitted radiation is polarized; very high energy (relativistic) electrons spiraling in a magnetic field around a neutron star will emit polarized light called synchrotron radiation. Polarization spectra and images therefore contain additional information about physical processes and source geometry which cannot be discerned from brightness measurements alone.

Determining the polarization state requires relative brightness measurements of the source at different "settings" of a polarization-sensitive element called a polarization modulator. The simplest modulator is a retardation plate made from a crystal exhibiting birefringence or double refraction; that is, the material has different refractive indices (n_o, n_e) for polarization perpendicular (ordinary) and parallel (extraordinary) to the axis of anisotropy, respectively. Birefringence was first described in calcite crystals by the Danish scientist Rasmus Bartholin (1625–1698) in 1669, but many other materials exhibit this effect. In calcite, n_e is smaller than n_o, and thus light polarized along the extraordinary axis (parallel to the axis of anisotropy) propagates faster through the crystal. Light with polarization components along both crystal axes will in general emerge in a different polarization state depending on the thickness of the crystal. If $\Delta n = n_o - n_e$ is the birefringence, then the relative phase shift between orthogonally polarized waves will be $\gamma = 2\pi\,\Delta n\,L/\lambda$, where L is the thickness of the crystal. If $L = \lambda/4$ (a quarter-wave plate), then linearly polarized light emerges as circular polarization when the fast axis makes an angle of 45° with the incident plane of polarization. A half-wave plate has $L = \lambda/2$ and thus rotates the direction of the polarization of the emergent light relative to the incoming light. Other devices that create retardations in crystals are the electro-optic effect (Pockels cell) and the photo-elastic effect (Kemp modulator). Polarization is discussed in more detail in Chapter 5. In all these examples, however, retardation is exact (quarter-wave or half-wave) for only one wavelength, but just as achromatic lenses can be made by combining materials with different dispersions so achromatic wave plates can also be formed. In fact, super-achromatic wave plates are available that span the entire wavelength range of a CCD. Because of the systematic—but unpredictable—intensity variations in the light due to poor seeing or poor tracking, it is essential that the rate at which one moves from one setting of the modulator to another be fairly rapid. Alternatively, some means must be found to measure two polarization positions simultaneously, so that both are affected in the same way by any systematic errors. Mechanical rotation of wave plates is usually fairly slow for applications with CCD cameras (<0.01 Hz), and therefore two polarization states are observed simultaneously to correct for any drifts.

4.3.2 Polarization maps and spectra

One remarkable technical solution is to use the unique charge-coupling attributes of the CCD to build up the alternating (polarization) signals *on the chip* until enough

counts have accumulated to swamp any electronic read-out noise. This idea was first suggested to me by Pete Stockman (now at the Space Telescope Science Institute) in an imaging context using a Pockels cell device when we were both at Steward Observatory in Arizona in 1978/1979. It led me to the idea of developing an instrument called the Imaging SpectroPolarimeter (or ISP) based on the properties of a three-phase CCD (McLean, 1981). Although basically a CCD camera, the ISP could be converted to a spectrometer by placing a "grism", a small right-angled prism with a transmission grating deposited on the hypotenuse face, in one of the filter positions, and converted to a polarimeter (imaging or spectro) by inserting a polarization modulator in front of the entire optical system. In spectropolarimetry mode, a special polarizer made of calcite was placed under the slit of the spectrograph which produced two oppositely polarized spectra (the E and O rays) on the CCD; actually, two tiny slits were used, one for the object and one for the sky, and so four spectra appeared on the CCD. The structure of the CCD allows "bi-directional" charge transfer, therefore alternating images (or spectra) corresponding to orthogonal polarization states of the modulator (an achromatic half-wave plate) could be stored in the top and bottom thirds of the CCD array while only the central part was used for actual light collection (see Figure 4.19). The collected "charge image" in the center is transferred by charge coupling to the appropriate (upper or lower) storage area during the dead-time required to move to the other modulator setting. No other detector can perform this trick!

One scientific result from this instrument was the first very high resolution images of the synchrotron polarization from the Crab Nebula (M1), and the consequent disentangling of nebular and pulsar polarizations (Figure 4.20). Polarization measurements have been important in revealing the nature of active galactic nuclei (AGN), powered by the accretion disk around a supermassive black hole. Although shrouded by gas and dust clouds that obscure the AGN, information from the core comes indirectly via scattered light which is polarized.

Since the charge-shifting concept was introduced, the charge-coupling efficiencies of CCDs have improved enormously and other variations of this idea have made their way into astronomical spectroscopy, most notably the "nod and shuffle" method for removing sky background (see Glazebrook and Bland-Hawthorn, 2001 for a detailed explanation). In this method the telescope is moved back and forth (nodding) between the object position and the sky reference position while also shifting (shuffling) the charge on the CCD pixels between illuminated and storage regions to ensure that the same pixels are used for both object and sky.

Although it is much harder to measure than most other properties, polarization can now be detected across the electromagnetic spectrum from X-rays to radio waves. There is considerable current interest in trying to determine the polarization of the cosmic microwave background (CMB) because it may contain important information on the formation of the early Universe. The CMB comes from the "surface of last scattering" when the Universe has expanded enough to become neutral (age $\sim$380,000 years), thus the CMB can only become partially polarized if there are free electrons around for photons to interact with via Thompson scattering. Consequently, the Universe must have been re-ionized by the first generation of stars to

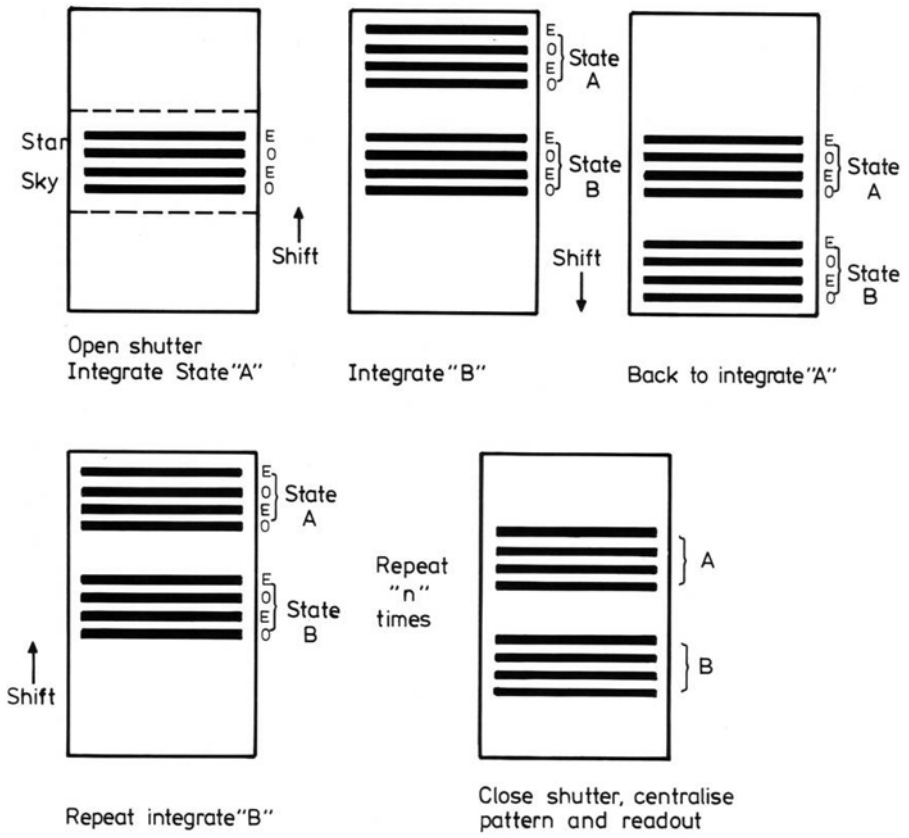


Figure 4.19. The unique bi-directional charge-shifting principle employed in the CCD Imaging Spectropolarimeter (or ISP) developed at the Royal Observatory Edinburgh by the author. A calcite crystal below the slit produces two orthogonally polarized spectra: the O (ordinary) and E (extraordinary) rays. Charge-shifting is used in general for differential measurements such as polarimetry and nod-and-shuffle spectroscopy.

form. Measurements from the Wilkinson Microwave Anisotropy Probe (WMAP) suggest that ~10% of the CMB photons have scattered in this way and that this re-ionization event happened about 400 million years after the Big Bang. Clearly, when the full panoply of technologies for the entire electromagnetic spectrum is considered, the power of modern astronomical instrumentation is very impressive.

4.4 SUMMARY

In addition to detectors and new telescopes, many other technologies, such as lasers, fiber optics, micromachining, diamond-polishing, and advanced metrology, have

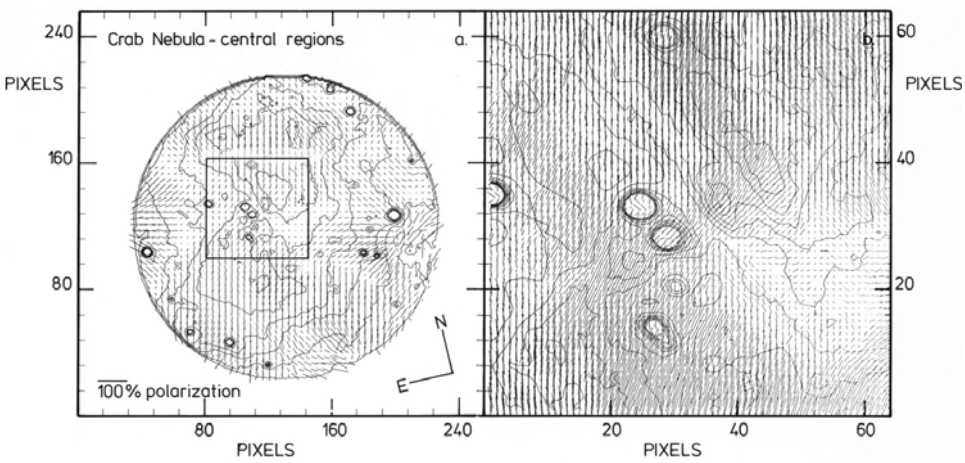


Figure 4.20. A contour map obtained by the author of the bright emission associated with the Crab Nebula supernova remnant overlaid with tiny line segments which represent the amount of polarization and the orientation of the magnetic field in the nebula. Credit: McLean *et al.* (1983).

contributed to astronomical instruments being more efficient. Instruments have been developed that can measure all of the observables. The enormous computing power applied to surveys of the sky like SDSS and 2MASS, and the enormous light-gathering power of modern ground-based telescopes supported by space missions to span the unobservable regions, has led to a flood of discoveries since the 1990s. Among the many revelations that could be listed is the discovery of Pluto-sized objects in the outer solar system, the discovery of hundreds of other planetary systems, the discovery of brown dwarfs (the missing link between small stars and gas-giant planets), the best evidence yet for a black hole at the center of the Milky Way, conclusive evidence for dark matter throughout the Universe, and the remarkable implication from supernovae studies that the expansion of the Universe seems to be accelerating. There is no doubt that these discoveries are linked directly to enabling technologies.

4.5 EXERCISES

1 How are the images that are recorded on large photographic plates handled and made available in digital form for analysis?

2 Explain the difference between a Faint Object Spectrograph and a High Resolution Spectrograph. Give an example of a project that would require one or other of these instruments.

3 What techniques are used to produce spectra of many objects simultaneously?

Discuss the advantages of multi-object spectroscopy over conventional spectroscopy.

4 Cite some examples of modern instruments and telescopes and the discoveries made with them.

5 Suppose you wish to search the entire sky for relatively faint, point-like objects called quasars whose spectra show strong emission lines at high redshifts. What technique would you use?

6 Suppose you wish to study the gravitational lensing of a distant quasar caused by a massive fore-ground galaxy. What kind of instruments would you need?

7 What instrument could you use to prove that the light from the Crab Nebula supernova remnant was due to non-thermal emission from the process called synchrotron radiation?

8 What is meant by integral field spectroscopy? Explain the different approaches and their advantages and disadvantages.

9 Explain the technique of drift-scanning and compare it with Time Delay and Integration.

10 What methods and instruments are required to detect the Doppler reflex motion of a star due to the presence of an orbiting planet?

4.6 REFERENCES

Bacon R.; Adam, G.; Baranne, A.; Courtes, G.; Dubet, D.; Dubois, J.P.; Emsellem, E.; Ferruit, P.; Georgelin, Y.; Monnet, G. *et al.* (1995) 3D spectroscopy at high spatial resolution, I: Concept and realization of the integral field spectrograph TIGER, *Astronomy and Astrophysics*, **113**, 347.

Brown, M.E.; Trujillo, C.; and Rabinowitz, D. (2004) Discovery of a candidate inner Oort cloud planetoid, *Astrophys. J.*, **617**, 645–649.

Charbonneau, D.; Brown, T.M.; Latham, D.W.; and Mayor, M. (2000) Detection of planetary transits across a Sun-like star, *Astrophys. J.*, **529**, L45–L48.

Clarke, D.; and Grainger, J.F. (1971) *Polarized Light and Its Measurement*, Pergamon, London.

Coyne, G.V.; Magalhaes, A.M.; Moffat, A.F.J.; Schulte-Ladbeck, R.E.; Tapia, S.; and Wickramasinghe, D.T. (eds.) (1988) *Polarized Radiation of Circumstellar Origin*, University of Arizona Press, Tucson, U.S.A.

Djorgovski, S.G.; Gal, R.R.; Odewahn, S.C.; de Carvalho, R.R.; Brunner, R.; Longo, G.; and Scaramella, R. (1999) The Palomar Digital Sky Survey (DPOSS), in S. Colombi, Y. Meillier, and B. Raban (eds.), *Wide Field Surveys in Cosmology*, Editions Frontières, Gif sur Yvette, p. 89; available as astro-ph/9809187.

Gehrels, T. (1991) Scanning with charge-coupled devices, *Space Sci. Rev.*, **58**, 347–375.

Genzel, R.; Schödel, R.; Ott, T.; Eckart, A.; Alexander, T.; Lacombe, F.; Rouan, D.; and Aschenbach, B. (2003) Near-infrared flares from accreting gas around the supermassive black hole at the Galactic Centre, *Nature*, **425**, 934.

Ghez, A.M.; Salim, S.; Hornstein, S.D.; Tanner, A.; Lu, J.R.; Morris, M.; Becklin, E.E.; and Duchêne, G. (2005) Stellar orbits around the Galactic Center Black Hole, *Astrophys. J.*, **620**, 744–757.

Glazebrook, K.; and Bland-Hawthorn, J. (2001) Microslit Nod-Shuffle Spectroscopy: A technique for achieving very high densities of spectra, *Publ. Astron. Soc. Pacific*, **113**, 197–214.

Gregory, S.A.; and Thompson, L.A. (1978) The Coma/A1367 supercluster and its environs, *Astrophys. J.*, **222**, 784.

Hardcastle, M.J.; Kraft, R.P.; Sivakoff, G.R.; Goodger, J.L.; Croston, J.H.; Jordán, A.; Evans, D.A.; Worrall, D.M.; Birkinshaw, M.; Raychaudhury, S. *et al.* (2008) New results on particle acceleration in the Centaurus A jet and counterjet from a deep CHANDRA observation, *Astrophys. J.*, **670**, L81–L84.

Jones, D.H.; Saunders, W.; Colless, M.; Read, M.A.; Parker, Q.A.; Watson, F.G.; Campbell, L.A.; Burkey, D.; Mauch, T.; Moore, L. *et al.* (2004) The 6dF Galaxy Survey: Samples, observational techniques and the first data release, *Mon. Not. Royal Astron. Soc.*, **355**, 747–763.

Kirkpatrick, J.D. (2005) New Spectral Types L and T, *Ann. Rev. of Astron. and Astrophys.*, **43**, 195–245.

Kovalevsky, J.; and Seidelman, P.K. (2004) *Fundamentals of Astrometry*, Cambridge University Press, Cambridge, U.K.

Krabbe, A.; Weitzel, L,; Kroker, H.; Tacconi-Garman, L.E.; Cameron, M.; Thatte, N.A.; Samann, G.; Boeker, T.; Genzel, R.; and Dropatz, S. (1995) 3D: A new generation imaging spectrometer, *SPIE*, **2475**, 172–183.

Macintosh, B.; Graham, J.R.; Palmer, D.; Doyon, R.; Gavel, D.; Larkin, J.; Oppenheimer, B.; Saddlemeyer, L.; Wallace, J.K.; Bauman, B. *et al.* (2006) The Gemini Planet Imager, in B. Ellerbroek and D.B. Calia (eds.), *Advances in Adaptive Optics II*, *Proc. SPIE*, **6272**, 12 pp.

Mayor, M.; and Queloz, D. (1995) A Jupiter-mass companion to a solar-type star, *Nature*, **378**, 355–359.

McLean, I.S.; Cormack, W.A.; Herd, J.T.; and Aspin, C. (1981) The Royal Observatory Edinburgh (ROE) Chare-coupled Device Camera System, *Proc. SPIE*, **290**, 155.

McLean, I.S.; Aspin, C.; and Reitsema, H. (1983) High-resolution polarization images of the Crab Nebula with a charge-coupled device, *Nature*, **304**, 243–246.

McLean, I.S.; Prato, L.; Kim, S.S.; Wilcox, M.K.; Kirkpatrick, J.D.; and Burgasser, A. (2001) Near-infrared spectroscopy of brown dwarfs: Methane and the transition between the L and T spectral types, *Astrophys. J.*, **561**, L115–L118.

Monnier, J.D. (2003) Optical interferometry in astronomy, *Reports on Progress in Physics*, **66**, 789–857.

Oke, J.B.; Cohen, J.G.; Carr, M.; Cromer, J.; Dingizian, A.; Harris, F.H.; Labreque, S.; Lucinio, R.; Schaal, W.; Epps, H.; and Miller, J. (1995) The Keck Low-Resolution Imaging Spectrometer, *Pub. Astron. Soc. Pacific*, **107**, 375–385.

Perlmutter, S.; Aldering, G.; Goldhaber, G.; Knop, R.A.; Nugent, P.; Castro, P.G.; Deustua, S.; Fabbro, S.; Goobar, A.; Groom, D.E. *et al.* (1999) Measurements of omega and lambda from 42 high redshift supernovae, *Astrophys. J.*, **517**, 565–586.

Riess, A.G.; Filippenko, A.V.; Challis, P.; Clocchiatti, A.; Diercks, A.; Garnavich, P.M.; Gilliland, R.L.; Hogan, C.J.; Jha, S.; Kirshner, R.P. *et al.* (1998) Observational evidence

from supernovae for an accelerating universe and a cosmological constant, *Astron. J.*, **116**, 1009–1038.

Rivera, E.; Lissauer, J.J.; Butler, R.P.; Marcy, G.W.; Vogt, S.S.; Fischer, D.A.; Brown, T.M.; Laughlin, G.; and Henry, G.W. (2005) A $\sim$7.5 $M_\oplus$ planet orbiting the nearby star, GJ 876, *Astrophys. J.*, **634**, 625–640.

Seidelmann, P.K.; and Kovalevsky, J. (2002) Application of the new concepts and definitions (ICRS, CIP and CEO) in fundamental astronomy, *Astron. & Astrophys.*, **392**, 341–351.

Skrutskie, M.F. *et al.* (2006) The Two Micron All Sky Survey (2MASS), *Astron. J.*, **131**, 1163–1183.

Smith, B.A.; and Terrile, R.J. (1984) A circumstellar disk around Beta Pictoris, *Science*, **226**, 1421–1424.

Steidel, C.C.; Shapley, A.E.; Pettini, M.; Adelberger, K.L.; Erb, D.K.; Reddy, N.A.; and Hunt, M.P. (2004) A survey of star-forming galaxies in the $1.4 \leq z \leq 2.5$ redshift desert: Overview, *Astrophys. J.*, **605**, 534–550.

Tegmark, M.; Blanton, M.R.; Straus, M.A.; Hoyle, F.; Schlegel, D.; Scoccimarro, R.; Vogeley, M.S.; Weinberg, D.H.; Zehavi, I.; Berlind, A. *et al.* (2004) The three-dimensional power spectrum of galaxies from the Sloan Digital Sky Survey, *Astrophys. J.*, **606**, 702–740.

Tytler, D.; Fan, X-M.; and Burles, S. (1996) Cosmological baryon density derived from the deuterium abundance at redshift $z = 3.57$, *Nature*, **381**, 207–209.

General astronomy texts referenced

Bennett, J.O.; Donahue, M.; Schneider, N.; and Voit, M. (2007) *The Cosmic Perspective*, fifth edition, Benjamin Cummings, San Francisco.

Karttunen, H.; Kröger, P.; Oja, H.; Poutanen, M.; and Donner, K.J. (eds.) (2007) *Fundamental Astronomy*, fifth edition, Springer-Verlag, Berlin.

Web sites

2MASS at IPAC: *http://www.ipac.caltech.edu/2mass/*
CCPS: *http://exoplanets.org/*
DPOSS: *http://www.astro.caltech.edu/~george/dposs/*
Gemini Planet Imager: *http://gpi.berkeley.edu/index.html*
Multiw_sky: *http://antwrp.gsfc.nasa.gov/htmltest/jbonnell/www/multiw_sky.html*
Pan-STARRS: *http://pan-starrs.ifa.hawaii.edu/public/*
Sloan Digital Sky Survey: *http://www.sdss.org/*
STScI Digitized Sky Survey: *http://archive.stsci.edu/cgi-bin/dss_form*
WMAP: *http://map.gsfc.nasa.gov/*

5

Instrumentation and detectors

Thus far, many different astronomical instruments and techniques have been introduced without much explanation of the underlying physical principles. In this chapter each class of instrument is examined in more detail. For optical instruments, typical layouts are shown and the basic relationships involving spatial and spectral resolution are given. Each major type of detector is introduced and the basic properties of semiconductors are also presented.

5.1 PHOTOMETERS AND CAMERAS

Broadly speaking, there are four classes of instruments used in astronomy: namely, (1) *photometers/cameras* which measure the brightness and direction of radiation, also sometimes called radiometers; (2) *spectrometers* which measure the distribution of brightness (or energy) as a function of wavelength; (3) *polarimeters* which determine the degree of alignment of wave vibrations in a beam; and (4) *interferometers* which rely on coherent phase relationships to achieve interference effects before performing imaging or spectroscopy. Variations of these instruments exist from X-ray wavelengths to radio wavelengths, although the methods of implementation differ considerably. In general, the descriptions which follow are applicable for UV, visible, and infrared wavelengths (UV/O/IR).

5.1.1 Photoelectric photometers

A photometer is a device for measuring the apparent brightness of a source, one of the most fundamental observables. Measurements are usually made after the light has been collected by a telescope and after transmission through the atmosphere. Ideally we would like to measure the power received per square meter integrated over all wavelengths (i.e., the irradiance E or astronomical flux S). Instead, measurements of

brightness are usually limited to a band of wavelengths selected by means of an optical "filter". Initially, colored glass filters and the detector's own wavelength-dependent response to light determined the wavebands used, but it is now possible to design and make an optical filter to pass any specific band of wavelengths desired. These filters are known as "interference filters" because they utilize destructive interference in multiple, very thin, dielectric (non-conducting) layers deposited on the glass substrate. We will return to their construction in Section 5.4.3. For photometry of individual stars, a detector with a single cell can be used, such as a photomultiplier tube (PMT). Several systems of brightness measurements have been in use since the introduction of PMTs. The most familiar of these is the UBV system (U = ultraviolet, B = blue, and V = visual or yellow). The original UBV system of Johnson and Morgan (1953) was defined by glass filters approximately centered at wavelengths of 360 nm, 440 nm, and 550 nm and the photoelectric response of a CsSb (S-4) photocathode typical of the RCA 1P21 PMT available at that time. Since then many photometric systems have been developed. Careful work is required to relate one filter system to another, and this task forms part of the calibration of the instrument.

Essential features of a simple photoelectric photometer are shown in Figure 5.1. The instrument is a light-tight box attached to the telescope by means of a flange. At the telescope focus, which falls inside the box, a circular aperture (or diaphragm) isolates a given star. Usually, these apertures will be interchangeable by constructing a series of them in a wheel or slide. The size of the aperture needs to be larger than the image of the star (the "seeing" disk), but not excessively so, otherwise too much "sky" background is included. Another wheel or slide carries a selection of filters such as UBV. The detector is usually a PMT (e.g., a thermoelectrically cooled GaAs tube). A Fabry lens produces an image of the telescope primary mirror, the collecting aperture, onto the detector—and *not* an image of the star. From the Thin Lens Equation, the distance (s) between the Fabry lens and primary mirror is $\sim f_{\text{tel}}$ which is $\gg f_{\text{lens}}$, thus the distance between the Fabry lens and the PMT is $s' \sim f_{\text{lens}}$. All light rays from the star *must* pass through this pupil image no matter where the star is located in the aperture. This design prevents movement of the illuminated image on

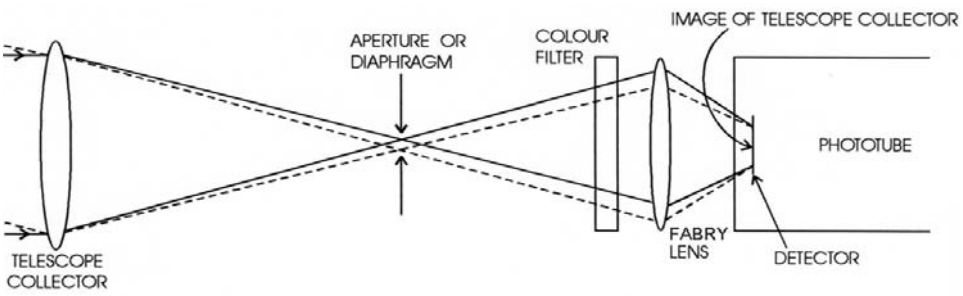


Figure 5.1. The basic layout of a photometer. Note that the image of the star is not focused onto the detector. By imaging the telescope's collecting aperture on the detector the signal strength is independent of the position of the star within the diaphragm.

the detector which might occur due to drifting of the star image across the diaphragm due to poor tracking. Consequently, the signal is stable and insensitive to variations in detector response over the photocathode surface. As shown in the figure, solid lines trace rays when the star is centered in the focal plane aperture and the dashed lines indicate the light path when the star is at the edge of the aperture. In principle, photometry can be carried out by measuring the current that appears at the anode of the tube as a result of the electron multiplication process, but this is a noisy method at low light levels because of the wide variation in pulse strength (often called pulse height) from identical events at the photocathode. A better solution is to count the pulses emitted by the anode irrespective of their height. The output of the PMT is fed to a pulse amplifier which gives out rectangular voltage pulses of a standard width and with a height proportional to the original signal from the anode. These signals go into a pulse height discriminator which is set to reject the many small pulses associated with amplifier noise. Those pulses that are passed can be counted with digital electronics and supplied to a computer. The PMT sits in a base socket, and a simple resistor chain between the base pins of the PMT can establish the inter-dynode voltages. Usually, the photocathode is at a negative potential of about 1,600 volts and the anode is at ground potential. Thermoelectric cooling to -20°C is often sufficient but some PMTs with high dark currents can be cooled with dry ice to -78°C. Additional information on photoelectric photometers is given in Henden and Kaitchuck (1998).

Until the advent of CCDs, many telescopes were equipped with photoelectric photometers. One of the great advantages of the photomultiplier tube is its speed of response to a change in brightness; typically, one-thousandth of a second. There are many useful and important applications of high-speed photometry. For example, objects such as cataclysmic variables and pulsars suffer rapid changes in brightness on short time scales. Also, when stars are occulted by the Moon (or a planet) passing in front of them, or satellites of planets are occulted by the planet itself, there is a very rapid dimming which yields the physical dimensions of the sources. PMTs are also ideal for polarization measurements which require very accurate differential photometry.

5.1.2 Camera systems

Figure 5.2 shows the basic layout of a camera. In the simplest design, the detector (CCD or other array detector) is placed directly in the focal plane of the telescope behind a light-tight shutter. Filters are therefore located in a wheel or slide in the converging beam from the telescope. Care is required to ensure that all filters have the same "optical path" (i.e., the product of refractive index and thickness), in order to avoid refocusing the telescope after each filter change. This method works well when the image scale of the telescope is well-matched to the pixel size of the detector, but tends to become challenging for very large telescopes, as we will see shortly. An alternative approach, shown in the figure, is to collimate the beam by placing a lens after the focal plane at a distance equal to its focal length ($s = f_{\mathrm{coll}}$) and to re-image the field onto the detector with a camera lens (or mirror) with $s' = f_{\mathrm{cam}}$. This design

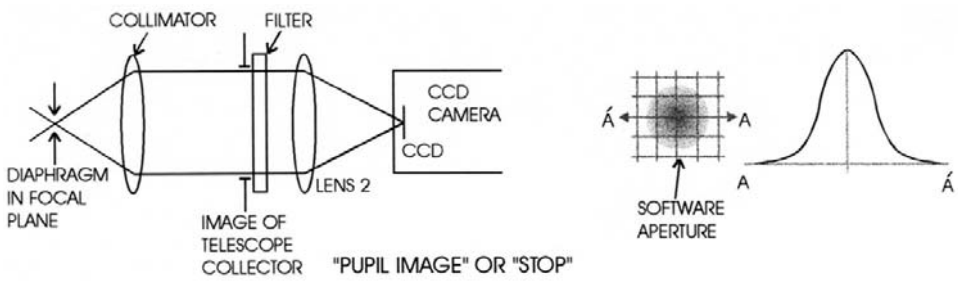


Figure 5.2. The basic layout of a camera system in which optics are used to collimate the diverging beam from the telescope focus and re-image the field at a different magnification. Filters can be placed in the collimated beam.

has many advantages. First, by selecting the focal lengths of the collimator and camera sections one can either magnify or reduce the plate scale; $m = f_{\rm cam}/f_{\rm coll}$. Filters of arbitrary thickness can be located in the parallel (collimated) beam, and the filters can be placed near the "pupil" image of the primary mirror. In addition, a circular aperture or "stop" can be placed at the pupil image to reject stray light from outside the beam. A stop is extremely important in infrared cameras where the pupil is at cryogenic temperatures and so it becomes a "cold" stop. Of course, in this design, the image of the star may drift by a small amount due to tracking or pointing errors, but photometry is performed "after the fact" on the digital image by selecting an appropriately sized "software aperture" and summing up all the signals. An annulus around the summed region is used to construct an estimate of the sky flux contained in the summed aperture. Thus, no separate measurement of the sky is required. Because the star image is spread over many pixels, and as different pixels are used for the sky image, it is essential to have a good procedure to normalize all the pixels to the same sensitivity or gain. This is a general requirement with array detectors, which is covered in considerable detail in Chapter 9.

5.1.3 Pixel sampling and matching to the plate scale

There are two issues to be considered when matching the spatial or spectral resolution element to the physical size of the detector pixels: (1) maximizing observing efficiency, meaning more light onto a pixel and therefore keeping the required integration time to a minimum; and (2) accomplishing this task without compromising the ability of the camera system to obtain very accurate brightness measurements (photometry). The spatial resolution element may be determined by seeing conditions or by optical constraints. In general, the image is either critically sampled, meaning that there will be about 2 pixels (also known as the Nyquist limit) across the resolution element, or it will be oversampled, which implies that there may be about five pixels across the resolution element. It is very rare to design a system which is undersampled deliberately. In a spectrometer, the width of the entrance slit is usually the determining factor. A narrow slit implies higher spectral resolution, but the highest efficiency is achieved when the slit is wide enough to accept the full image diameter.

Using the discussion of Chapter 3, consider first the plate scale of the telescope which is given in seconds of arc per mm (″/mm) by:

$$(ps)_{\text{tel}} = \frac{206{,}265}{f_{\text{tel}}} \tag{5.1}$$

Here, f_{tel} is the focal length of the telescope in millimeters ($f_{\text{tel}} = D_{\text{tel}} \times F$ where F is the focal ratio or f/number); and the numerical factor is the number of seconds of arc in 1 radian. Plate scales vary considerably. For instance, at the prime focus of the 3.6 m Canada–France–Hawaii Telescope (CFHT) the scale is 13.70″/mm, whereas at the Cassegrain focus the scale is 7.33″/mm. With an infrared telescope, however, the focal ratio is usually larger (slower) so that at the Cassegrain focus of the 3.8 m U.K. Infrared Telescope (UKIRT) the scale is only 1.52″/mm. Our f/16 24-inch reflector at UCLA, which might be typical of many campus telescopes, gives 21.1″/mm. For direct imaging, the angle on the sky subtended by the detector pixel is

$$\theta = (ps)_{\text{tel}} d_{\text{pix}} \tag{5.2}$$

where d_{pix} is the physical pixel size in mm; the pixels are usually square. For CCDs and near-infrared array detectors, values range from about 0.009 mm (9 μm) up to about 0.030 mm (30 μm); detector pixels on mid-infrared arrays may be significantly larger. For 20 μm detector pixels we would get 0.27″/pixel and 0.15″/pixel at the prime and Cass foci of the CFHT, respectively; 0.42″/pixel on the 24-inch at UCLA; and only 0.03″/pixel on UKIRT. We need to compare these values with the image quality to determine whether or not some optical magnification is required. For example, for our "rooftop" conditions on the UCLA campus we use 3″ for the average seeing disk, whereas for the instruments on the CFHT and other telescopes on Mauna Kea, Hawaii one might adopt 0.3″–0.5″ for the seeing! Calculating the required magnification factor can proceed as follows:

- choose a value for the diameter of the seeing in seconds of arc;
- decide on the sampling (p = 2–5 pixels);
- divide seeing diameter by sampling factor to get angular size of 1 pixel, $\theta_{\text{pix}} = \theta_{\text{see}}/p$ in arcseconds;
- given the size of the detector pixels, derive the plate scale at the detector from $(ps)_{\text{det}} = \theta_{\text{pix}}/d_{\text{pix}}$;
- the required magnification (m) is then

$$m = \frac{(ps)_{\text{tel}}}{(ps)_{\text{det}}} \tag{5.3}$$

  where $m = f_{\text{cam}}/f_{\text{coll}}$ as before.

Note that m also defines an Effective Focal Length (EFL = mf_{tel}) for the entire optical system. If $m > 1$, then the optical components are a magnifier, whereas if $m < 1$ (the usual case), then the optics are called a "focal reducer". We can also relate the pixel size in seconds of arc to the f-number of the focal reducer optics (or simply, "the

camera") by

$$\theta_{\text{pix}} = 206{,}265 \frac{d_{\text{pix}}}{D_{\text{tel}}(f/\text{number})_{\text{cam}}} \tag{5.4}$$

where $(f/\text{number})_{\text{cam}} = f_{\text{cam}}/D_{\text{cam}} = F_{\text{cam}}$.

Example. If $d_{\text{pix}} = 27\,\mu\text{m}$ and $D_{\text{tel}} = 10\,\text{m}$ (Keck telescope), then $\theta_{\text{pix}} = 0.56''/(f/\text{number})_{\text{cam}}$. Assuming seeing of $0.5''$ (on Mauna Kea) and 2-pixel sampling, this implies $\theta_{\text{pix}} = 0.25''$ which leads to $F_{\text{cam}} = 2.2$. Remembering that the optics must be well-corrected, this is quite a fast camera (more on this issue in Chapter 6). For $18.5\,\mu\text{m}$ pixels, however, we would need an $f/1.5$ camera! As CCD pixels get smaller and telescope mirrors get larger, it becomes more challenging to invent an optical reimaging (or matching) system. Oversampling to 3 or 4 pixels, or creating a smaller image size in the first place (with an adaptive optics system to eliminate atmospheric turbulence effects), will makes things much easier.

The image of a distant point-source object produced by a perfect telescope with a circular entrance aperture (D_{tel}) should have a bright core surrounded by fainter rings, the Airy diffraction disk, and the first dark minimum between the bright core and the first faint ring corresponds to an angular radius (in radians) of

$$\theta = 1.22 \frac{\lambda}{D_{\text{tel}}} \text{ radians} \tag{5.5}$$

where λ is the wavelength. Again, to convert this angle to seconds of arc ($''$) multiply by 206,265, the number of seconds of arc in one radian.

Example. If $\lambda = 0.5\,\mu\text{m}$, $D_{\text{tel}} = 0.5\,\text{m}$, then $\theta = 0.25''$. The same limit is reached at $\lambda = 1\,\mu\text{m}$ for a 1 m telescope and again at $10\,\mu\text{m}$ for a 10 m telescope. Alternatively, at $0.5\,\mu\text{m}$ on a 10 m telescope the diffraction limit is only $0.0125''$, which is much smaller than the typical seeing disk at any ground-based observatory. As shown in Chapter 2, the size of the seeing disk can be predicted as λ/r_0, where the Fried parameter r_0 is the length over which the incoming wavefront is not significantly disturbed by motions in the Earth's atmosphere. For $r_0 = 20\,\text{cm}$ at a wavelength of $0.5\,\mu\text{m}$ the seeing would be at least $0.5''$.

The angular radius of the diffraction-limited image can be related to a physical size by multiplying by the appropriate focal length (f_{tel} in this case) and using the definition of focal ratio,

$$r_{\text{diff}} = 1.22\lambda(f/\text{number})_{\text{tel}} \tag{5.6}$$

which shows that the *diameter* of the diffraction-limited image from an $f/15$ telescope is $2.44 \times 15 = 36.6$ microns at $\lambda = 1\,\mu\text{m}$. The same formula can also be used to find the physical size of the image spot in a camera system. For example, an $f/3$ camera

has a diffraction spot of only 7.3 microns at a wavelength of 1 μm, half this size in the mid-visible, and double this in the near-infrared bands.

Optical matching to a CCD in spectroscopic mode is similar to normal camera mode except that it will usually be relevant to choose the pixel size to correspond to the spectral resolution of the spectrometer, and this is partially determined by how narrow the entrance slit is made. Sometimes this will be smaller than the seeing disk, although designers try to ensure that all of the light is accepted without loss of spectral resolution; the consequence of this is that the spectrograph has to be physically very large (see Section 5.2). As in the case of spatial resolution, it is best to oversample the spectral resolution with 2, 3, or more pixels.

5.2 SPECTROMETERS

All spectrometers have essentially the same basic design, but many different implementations are possible depending on the constraints and choice of spectral disperser. Figure 5.3 shows the essential features of the layout. Instead of a wide-field aperture at the focal plane of the telescope, a mask with a narrow slit is used. The slit width must be matched to either the seeing conditions or the diffraction disk depending on the design and application. As the beam diverges from the slit it is first collimated and then directed to the dispersing system (one or more prisms or diffraction gratings) after which the spectrally dispersed beam is finally collected by the camera optics and re-imaged onto the CCD or other array detector. The important quantities to determine when beginning the design are (1) the resolving power (R), (2) the slit width, (3) the diameter of the collimated beam, (4) the sampling or matching of the slit width to the detector pixels, and (5) the resulting f/number of the camera system. Several key terms and properties follow.

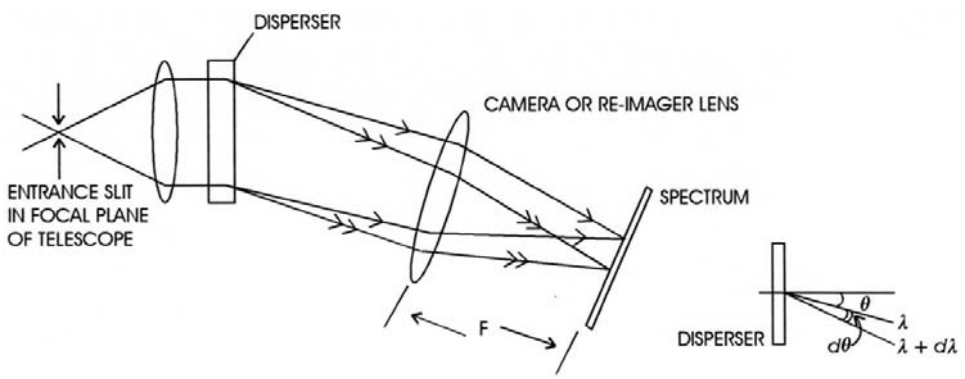


Figure 5.3. The essential features in the optical layout of a spectrometer. The beam is collimated before intersecting the dispersive element and then the spectrum is re-imaged with camera optics onto the detector.

5.2.1 Resolution and dispersion

Angular dispersion (AD): the rate of change of the dispersed angle (θ) of the beam with respect to wavelength (λ)

$$AD = \frac{d\theta}{d\lambda} \tag{5.7}$$

(Although expressed as a derivative, $d\theta$ and $d\lambda$ can also be taken to be small intervals.)

Linear dispersion: relates an interval of length (dx in millimeters) along the spectrum to a wavelength interval ($d\lambda$ in Å or nanometers)

$$LD = \frac{dx}{d\lambda} = \frac{dx}{d\theta}\frac{d\theta}{d\lambda} = f_{\text{cam}}\frac{d\theta}{d\lambda} \tag{5.8}$$

where f_{cam} is the focal length of the spectrograph camera and the units are usually expressed as mm/Å. A more useful form is the *Reciprocal Linear Dispersion* which is simply the inverse of the above expression in Å/millimeters.

Resolving power: is the ability to distinguish two wavelengths separated by a small amount $\Delta\lambda = \lambda_2 - \lambda_1$

$$R = \frac{\lambda}{\Delta\lambda} \tag{5.9}$$

Note that the "resolution" is often stated as

$$\frac{1}{R} = \frac{\Delta\lambda}{\lambda} = \frac{V}{c} \tag{5.10}$$

where the non-relativistic Doppler formula is used to relate the smallest detectable wavelength shift ($\Delta\lambda$) to the velocity V of a source that could cause this shift. For example, for $R = 10{,}000$ or 0.01% resolution ($\Delta\lambda = 0.0001\lambda$), then $V = 0.0001c = 30\,\text{km/s}$.

5.2.2 Diffraction gratings

The usual dispersing element is a diffraction grating and the general grating equation is

$$m\lambda = d(\sin i + \sin\theta)\cos\gamma \tag{5.11}$$

where d is the spacing of adjacent grooves or slits; i is the angle of incidence of the collimated beam; θ is the angle of the emergent diffracted beam; γ is the angle out of the normal plane of incidence (usually 0°, hence cos = 1); and m is an integer called the "order" of interference. For zero order ($m = 0$), $\sin\theta = -\sin i$ or $\theta = -i$. The negative sign comes from the fact that we have chosen to call i and θ positive when on the *same* side of the normal. Whenever the rays cross over the normal the angle of diffraction is taken to be negative. With this sign convention, Equation (5.11) applies when the grating is used in transmission and when the grating is used in reflection.

There is an alternative form of the equation that uses a negative sign between the terms to describe a reflection grating. In that case, the angles are positive if they are on *opposite* sides of the normal. If the medium on either side of the grating is not a vacuum then a more general form of Equation (5.11) would have ($n_1 \sin i + n_2 \sin \theta$). For simplicity we will use Equation (5.11) as it stands. The angular dispersion of a grating is therefore given by

$$\frac{d\theta}{d\lambda} = \frac{m}{d \cos \theta \cos \gamma} \tag{5.12}$$

Substituting for m/d gives

$$\frac{d\theta}{d\lambda} = \frac{\sin i + \sin \theta}{\lambda \cos \theta \cos \gamma} \tag{5.13}$$

Usually, $\cos \gamma \sim 1$, and therefore angular dispersion is determined entirely by i and θ for a given λ. Many combinations of m and d yield the same AD provided the grating angles remain unchanged. Typical "first-order gratings" ($m \sim 1$) have 300–2,400 grooves or lines/millimeter; the number of lines per millimeter is given by $T = 1/d$. In addition, $\cos \theta \sim 1$ and slowly varying, thus the angular dispersion is almost constant and the relationship between position and wavelength on the detector (Equation (5.8)) is approximately linear. Coarse-ruled reflection gratings (large d) can achieve high angular dispersion by making i and θ very large, typically 60°. Such gratings are called "echelles" and have groove densities from 20 to 200 lines/millimeter with values of m in the range 10–100. As we will see below, this results in severe overlap of orders unless a second disperser of lower resolving power at right angles to the first is used to "separate" the orders.

In a standard astronomical spectrograph, the light emerging from the slit (of width w) is collimated into a parallel beam of diameter D_{coll} and directed onto a reflection grating at an angle of incidence i so that the illuminated length is $W = D_{\text{coll}}/\cos i$. The diffracted beam returning from the grating is collected by a camera system and focused onto an array detector. The magnification between the slit and the detector is determined by $f_{\text{cam}}/f_{\text{coll}}$. For practical reasons, it is much more convenient to keep the collimator and camera optics in a fixed position and allow some limited motion of the grating, than to fix the grating and require the camera optics to move in an arc to pick up different parts of the diffracted beam. Thus, there is usually an optical axis defined by the collimator and another defined by the camera. These axes intersect on the reflection grating and the angle between these beams is called the "spectrograph angle" (φ). Figure 5.4 is a simple schematic of this fixed relationship which also illustrates spectral displacements (Δx) due to a shift in the slit position as occurs in multi-slit spectrometers.

If the grating is ruled on a concave spherical mirror instead of a plane surface it will diffract and focus the light at the same time. In addition to eliminating chromatic aberration in the re-focusing optics, this approach is ideally suited for the far-ultraviolet where transmission lenses are difficult to obtain. It is found that if R_g is the radius of curvature of the concave grating, a circle of diameter R_g (or radius $R_g/2$) can be drawn tangent to the grating at its mid-point which defines the locus of points where the spectrum is in focus, providing that the entrance slit also lies on this

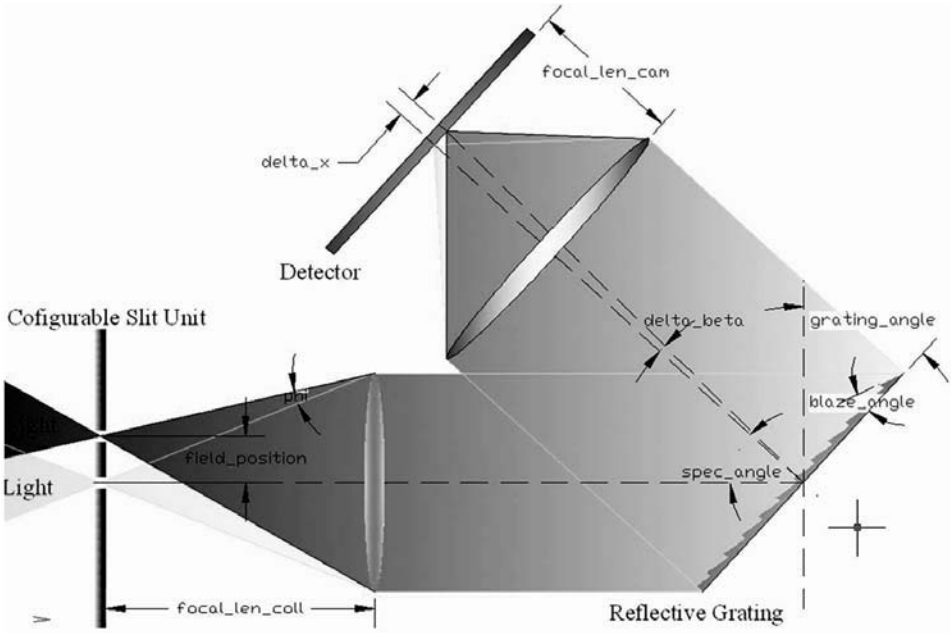


Figure 5.4. Typical arrangement for an astronomical spectrograph with a fixed angle between the collimator and the camera axes. Credit: David Guerrera.

circle. The circle is called the Rowland Circle (H.A. Roland 1848–1901) and was used, for example, in the ultraviolet spectrographs on the Far Ultraviolet Spectroscopic Explorer (FUSE) satellite.

A grating produces a different magnification in the dispersion direction than at right angles to the dispersion. The "anamorphic" magnification factor describes this effect and is found by determining the change in θ for a change in i:

$$\frac{d\theta}{di} = \left|\frac{\cos i}{\cos \theta}\right| \tag{5.14}$$

and the size of the slit image (Δx) at the detector becomes

$$\Delta x = w \frac{\cos i}{\cos \theta} \frac{f_{\text{cam}}}{f_{\text{coll}}} \tag{5.15}$$

where w is the true slit width; and f_{cam} and f_{coll} are the focal lengths of the camera and collimator optics; the ratio of focal lengths is the normal magnification factor. For example, if $i < \theta$ the grating normal is more nearly pointed at the collimator, the image of the slit is wider in the spectral direction, pixel sampling is better, but resolution is reduced. The opposite is true when the grating normal is pointing more nearly at the camera.

If the grating is to accept all the light from the collimator then it follows that the ruled width of the grating (W) must be $W = D_{\text{coll}}/\cos i$. In the diffraction-limited case

the resolving power of a grating spectrometer is then

$$R = mN = \frac{mW}{d} = \frac{W(\sin i + \sin \theta)}{\lambda} \tag{5.16}$$

where N is the total number of grooves illuminated. In practice, spectrometers are usually slit width limited or seeing-limited. If the slit is matched to the angular size (θ_{see}) of the seeing disk, then $\theta_{\text{see}} = \lambda/D_{\text{tel}}$ and

$$R = \frac{W(\sin i + \sin \theta)}{\theta_{\text{see}} D_{\text{tel}}} \tag{5.17}$$

Note that as the telescope diameter increases the resolving power R decreases, unless W, the grating size, gets larger too! Hence the effort to produce larger and larger reflection gratings and echelle gratings as the size of telescopes has increased. Taking $\theta_{\text{see}} = p \times \theta_{\text{pix}}$, where p is the number of pixels across the slit image, and converting to seconds of arc, gives the form which shows explicitly the tradeoffs of size vs. resolution:

$$R = \left(\frac{\sin i + \sin \theta}{\cos i}\right) \frac{D_{\text{coll}}}{D_{\text{tel}}} \frac{206{,}265}{p\theta_{\text{pix}}} \tag{5.18}$$

This important formula makes it clear that as telescopes get larger the spectrograph (defined by the beam size D_{coll}) gets larger too, all else being equal.

The intensity distribution (I) from an ideal grating can be derived by expanding the analysis of wave interference from single and double slits to N slits (Chapter 2). The result is

$$I = A_0^2 \frac{\sin^2 \beta}{\beta^2} \frac{\sin^2 N\gamma}{\sin^2 \gamma} \tag{5.19}$$

Here $\gamma = (\pi d \sin \theta)/\lambda$ is the phase difference between adjacent slits of separation d; and $\beta = (\pi b \sin \theta)/\lambda$ is the phase difference from the center of one slit (of width b) to its edge. Thus, the $\sin^2 \beta/\beta^2$ term represents single-slit diffraction while the second term is the result of interference from N slits. The slit width b is sometimes called the active width of the ruling. It can be shown that the $(\sin^2 N\gamma)/(\sin^2 \gamma)$ term has strong maximum values equal to N^2 for $\gamma = 0, \pi, 2\pi, \ldots$ and these maxima correspond to $d \sin \theta = 0, \lambda, 2\lambda, 3\lambda, \ldots = m\lambda$. Secondary maxima between orders are strongly suppressed by the slowly varying denominator. The relative intensities of the different orders $m = 0, 1, 2, 3, \ldots$ are governed by the broader single-slit diffraction pattern which creates an envelope whose first minimum occurs when $b \sin \theta = \lambda$, thus drastically reducing the intensity at $m = \pm 1$ and beyond as shown in Figure 5.5. For any given order of diffraction, except $m = 0$, different wavelengths are diffracted at different angles, thus producing a spectrum. But the single-slit diffraction pattern maximizes the diffracted intensity at zero order, where no dispersion occurs, and thus the first-order and second-order spectra are very faint. We need to be able to "shift" the peak of this envelope to $m = 1$. This is possible for reflection gratings.

Blaze angle: By tilting the facets of a reflection grating through an angle θ_B with respect to the plane of the grating surface it is possible to maximize the grating

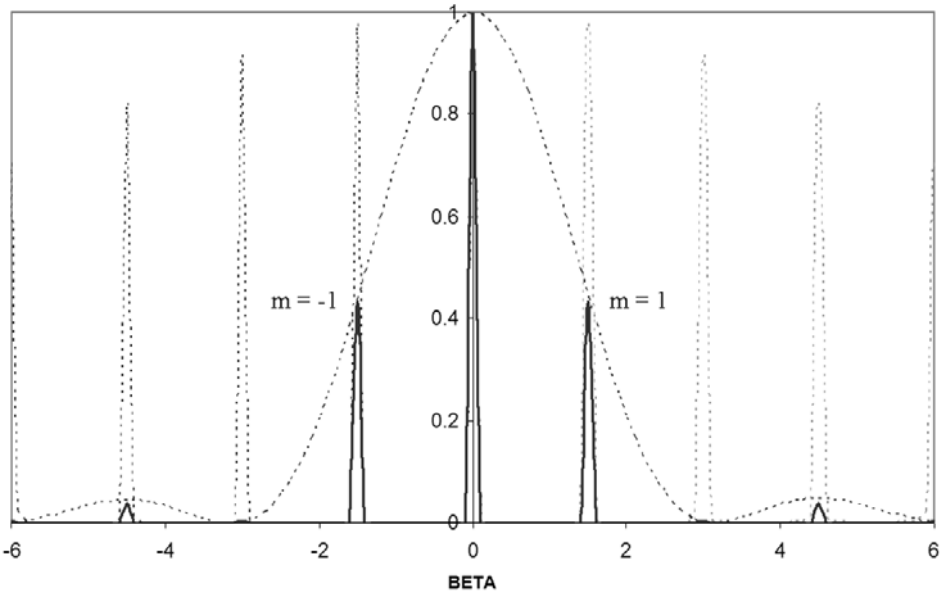


Figure 5.5. The interference effects of a diffraction grating. The larger the number of grooves or slits the sharper each order. The dotted line shows the overall blaze function. By tilting the facets the blaze function can be moved to coincide with $m = 1$.

efficiency in the direction in which light would have been reflected in the absence of diffraction. Grating efficiency is a maximum when the angle of incidence and angle of diffraction are related by $(i + \theta) = 2\theta_B$. The separation between the beams $(i - \theta)$ is just the spectrograph angle φ. Thus,

$$m\lambda_B = 2d \sin\theta_B \cos(\varphi/2) \tag{5.20}$$

There is a special case when $\varphi = 0$, then the incident ray enters along the normal to the facet and the diffracted ray leaves along the same direction. This is the "Littrow" condition, and the incident and diffracted angles measured relative to the grating normal are now equal to each other and to the blaze angle. The grating equation simplifies to $m\lambda_B = 2d \sin\theta_B$ and the resolving power is given by

$$R = \frac{2D_{\text{coll}} \tan\theta_B}{p\theta_{\text{pix}} D_{\text{tel}}} \tag{5.21}$$

The only way to work in the Littrow condition is with a central obscuration in the optics. Alternatively, one can use the "near" Littrow condition by moving off by 10°–20° or the "quasi" Littrow condition by going out of the plane ($\gamma > 0°$). Grating efficiency drops rapidly as the angle away from Littrow grows, whereas the drop is very slow for the quasi-Littrow mode but the adverse effect is that the slit images are

now tilted. The tilt angle χ is given by

$$\tan \chi = \tan \gamma \frac{(\sin i + \sin \theta)}{\cos \theta} \tag{5.22}$$

which for a $\tan \theta_B = 2$ echelle grating gives $\tan \chi = 4 \tan \gamma$. For example, for $\gamma = 5^\circ$, $\chi = 19.3^\circ$ and there is also a change $\Delta\chi$ in this angle across an order; the higher the order the smaller the change.

Free Spectral Range: For a given pair of incident and diffraction angles the grating equation is satisfied for all λ for which m is an integer. There are two wavelengths in successive orders, λ and λ', for which $m\lambda' = (m+1)\lambda$. The wavelength difference $\lambda' - \lambda$ is called the free spectral range (FSP); therefore

$$\Delta\lambda_{FSP} = \frac{\lambda}{m} \tag{5.23}$$

The two wavelengths are diffracted in the same direction and require either an "order sorter" filter or a cross-disperser element (when m is large) which is another grating or a prism at right angles to the first one. Order sorting filters must be carefully chosen to cut on and off sharply to prevent order overlap (see discussion of interference filters in Section 5.4.3). Grating efficiency is difficult to calculate. Peak efficiency should occur at λ_B when $m = 1$ and then declining peaks should occur at λ_B/m in subsequent orders. Away from the peak, efficiency drops and a good guide is that efficiency is down to 40% of the peak value at $\lambda_- = 2m\lambda_B/(2m+1)$ on the blue side and $\lambda_+ = 2m\lambda_B/(2m-1)$ on the red side.

Most astronomical gratings are of the surface relief kind in which the grooves are formed on the surface of the substrate (direct ruling) or as a replicated grating in a material bonded to the substrate. Reflection gratings can be coated with a reflective surface such as silver or gold, where the latter is particularly useful in the infrared. A relatively new alternative technology for grating fabrication is the Volume Phase Holographic or VPH grating. Not to be confused with a normal holographic grating which is another method for creating a surface relief grating, a VPH grating is an optical substrate in which the refractive index varies periodically throughout the body of the grating (Barden *et al.* 2000; Baldry *et al.* 2004). The grating body is made from a thin (3 μm–30 μm) slab of dichromated gelatine (DCG) trapped between glass plates (see Figure 5.6). Light passing through a VPH transmission grating obeys the following grating equation:

$$m\lambda = n_i \Lambda_g (\sin \alpha_i + \sin \beta_i) \tag{5.24}$$

where m is an integer representing the order; n_i is the refractive index of the medium; Λ_g is the grating period (equivalent to groove spacing) and is the projected separation between the fringes; $\Lambda_g = \Lambda/\cos \varphi$ where φ is the slant angle between the grating normal and the plane of the fringes. The angles of incidence (α_i) and diffraction (β_i) are relative to the grating normal with the convention that zero order (no diffraction) corresponds to $\beta_i = -\alpha_i$. The equation applies to each layer, where $i = 0$ is the air, $i = 1$ is the glass substrate, and $i = 2$ is the DCG layer. High-diffraction efficiency can

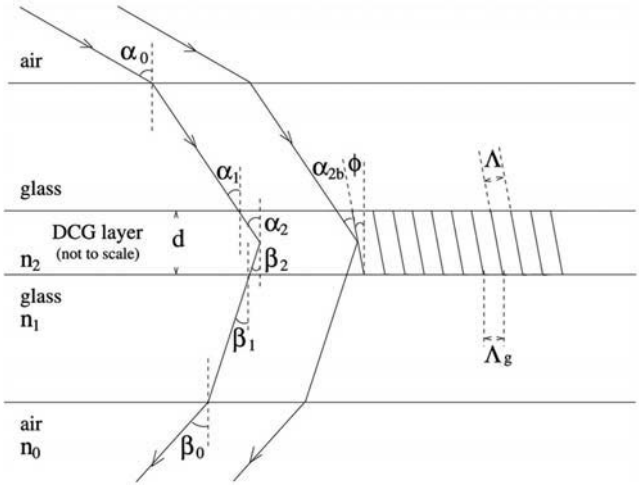


Figure 5.6. The basic parameters for a volume phase holographic grating. Bragg diffraction occurs at the planes of constant refractive index in the dichromated gelatin (DCG) layer. Credit: Ivan Baldry.

occur when light is effectively reflected from the plane of the fringes (i.e., when $\beta_2 + \varphi = \alpha_2 - \varphi$ in the DCG layer). This behavior is the same as Bragg diffraction of X-rays from the atomic layers in a crystal lattice. In both cases, because the thickness of the medium is much greater than the wavelength, constructive interference occurs for radiation scattered in that direction. The Bragg condition implies

$$m\lambda = 2n_2\Lambda \sin \alpha_{2b} \tag{5.25}$$

where n_2 is the refractive index of the DCG layer; and α_{2b} is the "Bragg angle" or angle of incidence with respect to the plane of the fringes $\alpha_{2b} = \alpha_2 - \varphi$. At wavelengths sufficiently displaced from the Bragg condition there is no diffraction. Diffraction efficiency also depends on the semi-amplitude of the refractive index modulation (Δn_2) and the grating thickness (d). The DCG holds a fringe pattern generated by holography which provides planes of constant refractive index separated by a length $\Lambda = 1/v_g$. Index variations are the result of density variations which are trapped into the material by exposure to light (the fringe pattern) because those regions collapse to a different density in the process of swelling the gelatin with water and then drying it rapidly. One form for the refractive index is $n_2(x, z) = n_2 + \Delta n_2 \cos[2\pi v_g(x \sin \gamma + z \cos \gamma)]$, which gives the variation in the x, z plane where z is the optical axis through the VPH, and γ is the angle between the normal to the planes and the z-axis. Line densities can range from 300 to 6,000 lines/millimeter and index modulations of 0.02–0.1 are typical. Because of the Bragg condition, it is necessary to articulate the camera to a new angle to tune to a new wavelength. This area of technology is receiving a great deal of research attention, in part because of the possibility of making VPH gratings in large sizes. A VPH is used in the 6dF spectrograph mentioned in the previous chapter. Regular updates can be

found in conference proceedings. Some suggested links are included in the reference list.

Another important new area receiving attention is the use of "immersion" gratings in which the grating surface is coupled to or embedded in a prism so that the factor n is returned to the grating equation, thus $m\lambda = 2dn\sin\theta$. A smaller collimated beam can be used to achieve the same resolution, or higher resolution for the same-sized beam. In infrared applications where the substrate can be a high-index material like silicon ($n \sim 4$), the gain is considerable.

5.2.3 Prisms

Prisms find many applications in modern spectrographs both in the role of primary disperser in (usually) low-resolution instruments and as a cross-disperser in high-resolution echelle spectrographs. In principle, prisms are cheaper and easier to make than diffraction gratings. Also, as there are no interference effects there are no overlapping orders to handle. However, obtaining high resolving power is very difficult, and unlike a diffraction grating the spectral resolution changes strongly with wavelength. The basic layout of a prism disperser is shown in Figure 5.7. From the definition of angular dispersion in Eqution (5.7) we can write:

$$\frac{d\theta}{d\lambda} = \frac{d\theta}{dn}\frac{dn}{d\lambda} = \frac{B}{D_{\text{cam}}}\frac{dn}{d\lambda} \tag{5.26}$$

In this expression $dn/d\lambda$ describes the wavelength dependence of the refractive index n. The purely geometric term $d\theta/dn$ can be derived by differentiating Snell's law applied to the second surface and then doubling the rate to account for both surfaces. At minimum deviation the angle $\varphi = \alpha/2$, giving $d\theta/dn = [2s\sin(\alpha/2)/s\cos\theta]$. However, $2s\sin(\alpha/2) = B$, the base length of the prism, and $s\cos\theta = D_{\text{cam}}$, the emergent collimated beam width towards the camera. The resolving power of a prism

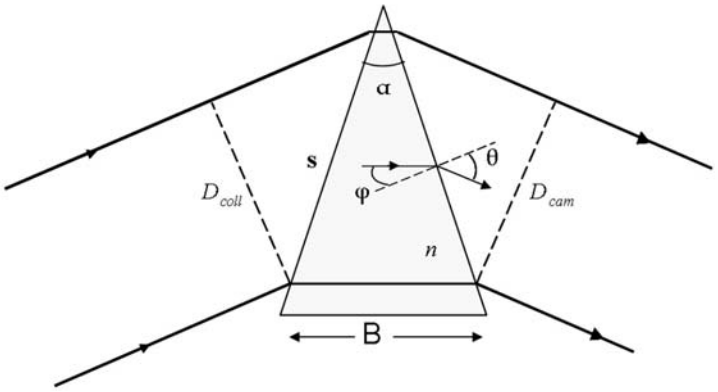


Figure 5.7. The relationship of angles and lengths in a prism at minimum deviation used to derive the resolving power.

is $R = B(dn/d\lambda)$ and for a slit-limited instrument the resolving power is given by:

$$R = \frac{\lambda}{\theta_{\text{res}} D_{\text{tel}}} B \frac{dn}{d\lambda} \tag{5.27}$$

Plots of the dispersion for many different glasses can be found in optical texts. The refractive index usually increases steeply towards shorter wavelengths, and therefore the blue end of a prism spectrum is more spread out than the red end.

5.2.4 Grisms

A very popular way to convert a camera into a spectrograph is to deposit a transmission grating on the hypotenuse of a right-angled prism and use the deviation of the prism to bring the first order of diffraction on axis. Such a device is called a "grism", and Figure 5.8 shows the basic geometry (not to scale). The advantage of a grism is that it can be placed in a filter wheel and treated like another filter. The basic relationships required to design a grism are

$$m\lambda_c T = (n-1)\sin\phi \tag{5.28}$$

and

$$R = \frac{EFL}{2d_{\text{pix}}}(n-1)\tan\varphi \tag{5.29}$$

where λ_c is the central wavelength; $T(= 1/d)$ is the number of lines per millimeter of the grating; n is the refractive index of the prism material; ϕ is the prism apex angle; EFL is the effective focal length of the camera system (see Equation (5.3)); and d_{pix} is the pixel size. The factor of 2 assumes that two pixels are matched to the slit width. In practice, the number of free parameters is constrained by available materials and grating rulings, and given conditions within the camera system. Resolving powers (two pixels) of $R \sim 500$–$2{,}000$ are practical.

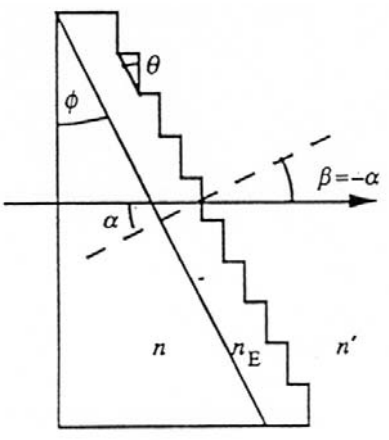


Figure 5.8. A simplified schematic giving the basic geometry of a grism, a transmission diffraction grating deposited on the hypotenuse face of a right-angled prism.

5.3 POLARIMETERS

With the exception of a few important cases (reflection nebulae, synchrotron emission from supernova remnants, and cyclotron emission from magnetic white-dwarf systems), the levels of polarization are quite small in astronomy. Typically only a few percent of starlight is linearly polarized by passage through the interstellar medium and much less is circularly polarized. Nevertheless, polarization is a powerful tool because it provides unique "geometric" information which measurement of intensity alone cannot do. To measure the polarization properties of light (fraction polarized, direction of vibration, and handedness of rotation) all polarimeters "convert" the polarization information into brightness modulations which are directly measurable with an electronic detector. Polarimeters therefore benefit from photometer, camera, and spectrometer designs. In essence, a photopolarimeter, an imaging polarimeter, and a spectropolarimeter are all created by adding a "polarization modulator" to the design.

5.3.1 Modulators and polarizers

The modulator can be a rotatable plate of a special optical material (or composite material such as quartz and magnesium fluoride) which exhibits a crystal structure property called birefringence which makes the material sensitive to the orientation or plane of vibration of the incident light wave. Essentially, the crystal will distinguish between electromagnetic waves which are vibrating at right angles to one another, and will "slow down" one of them. Alternatively, the modulator can be made of a material in which birefringence can be introduced by external means. For example, the application of an alternating electric field to certain crystals is one method. This is called a Pockels cell. Another way of inducing birefringence is by a variable mechanical strain. This is the photoelastic or Kemp modulator. In each case the modulator is followed by a polarizer (like Polaroid but usually a glass component like a Glan–Thompson or Wollaston prism rather than a plastic component to improve transmission). Operation of the modulator/polarizer combination results in a controlled, periodic variation of brightness of the transmitted light provided the incident light was polarized. These brightness variations are recorded by the detector. Conceptually, all polarimeters fall into this basic design structure: the light from the telescope must first pass through the polarimeter section comprising the modulator and polarizer, then through the camera or spectrometer part to arrive at the detector.

Although the simplest approach is to rotate a polarizer (such as Polaroid) in the beam, this is rarely done unless the remainder of the optical train is completely insensitive to polarization. A more robust approach is to construct the polarization modulator in two parts. First is a retardation device which introduces a known and controllable phase shift into the beam, and second is a fixed polarizer (or analyzer) that only allows one plane of polarization to pass unhindered and reduces others by the factor $\cos^2 \theta$, where θ is the angle between the polarizer's axis and the plane of polarization in the beam. The intensity transmitted by the analyzer is therefore modulated by the action of the phase retardation device. Unlike the Pockels cell

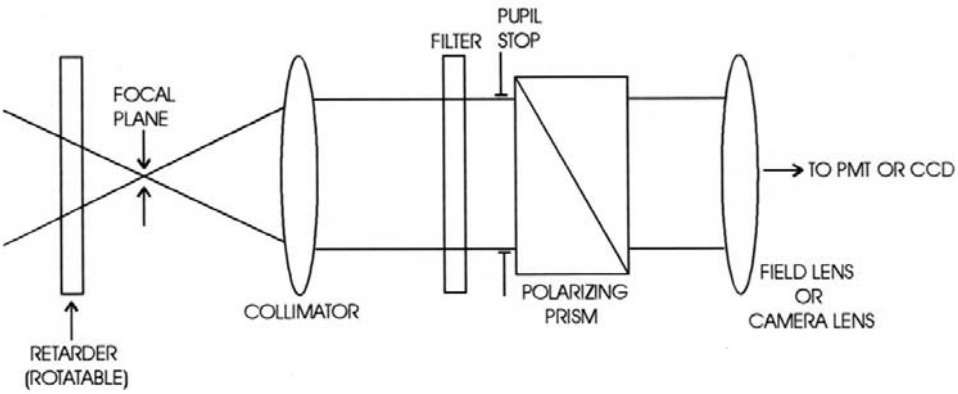


Figure 5.9. A typical layout for a camera or a photometer used as a polarimeter. The retardation device and the polarizing prism can be one of many types.

and Kemp modulator, in which perfect half-wave or quarter-wave retardation can only occur at a single wavelength, retardation plates can be made achromatic by combining two birefringent materials of different refractive indices. In general, the "retarder" is the first optical component in the system (apart from the telescope mirrors) to minimize spurious or instrumental sources of polarization. The analyzer must also be carefully placed so as to feed a constant direction of vibration to the detector and not be "crossed" inadvertently with any partial polarizer in the system (such as a diffraction grating) which would reduce throughput.

A typical polarimeter layout for a camera or a photometer is shown in Figure 5.9. The polarizing prism can be a "double-beam" device (like a Wollaston prism) which produces two polarized images on the detector known as the *o* (ordinary) and the *e* (extraordinary) images. A spectropolarimeter is constructed in a similar way, with the analyzer in front of the dispersing element. To overcome brightness modulations caused by the atmosphere (seeing, transparency), polarimetry must either be done rapidly, or by a ratio method. Most photopolarimeters using PMTs operate using the rapid modulation approach, but this is not suitable for CCDs and array cameras because of readout noise. A partial solution is to "charge-shift" back and forth on the CCD as described in Chapter 4, but the best solution is to use a double-beam instrument and take ratios of images and/or spectra in the *e* and *o* channels.

5.3.2 The Stokes parameters

Linear polarization is described by three parameters: intensity (I), degree (or fraction) of linear polarization (p), and the direction of the (fixed) plane of vibration projected on the sky (θ). Circular polarization is similarly described by three parameters: intensity (I), degree of circular polarization (q), and handedness of the rotation of the electric vector ($+$ or $-$). A more convenient way to express polarization information is to use the four Stokes parameters (I, Q, U, V). These quantities are phenomenological (i.e., they are more directly related to actual measurements). The

Stokes parameters are easily related to the amplitudes (E_x, E_y) of the electric vector in two orthogonal directions and to the phase difference (δ) between the two components (Clarke and Grainger, 1971). The degree of linear and circular polarization is given by

$$p = \frac{[Q^2 + U^2]^{1/2}}{I}, \qquad q = \pm \frac{V}{I} \tag{5.30}$$

and the direction of vibration of the linearly polarized part is given by

$$\tan 2\theta = \frac{U}{Q} \tag{5.31}$$

and it follows that

$$\left.\begin{aligned} Q &= Ip \cos 2\theta \\ U &= Ip \sin 2\theta \\ V &= Iq \end{aligned}\right\} \tag{5.32}$$

The intensity of light transmitted by a retarder of retardance τ at angle ψ followed by a perfect polarizer with principal plane at $\varphi = 0^\circ$ or $\varphi = 90^\circ$ (upper/lower signs, respectively) is given by

$$I' = \tfrac{1}{2}[I \pm Q(G + H \cos 4\psi) \pm UH \sin 4\psi \mp V \sin \tau \sin 2\psi] \tag{5.33}$$

where

$$G = \tfrac{1}{2}(1 + \cos \tau), \qquad H = \tfrac{1}{2}(1 - \cos \tau), \qquad \tau = \frac{2\pi}{\lambda}\delta \tag{5.34}$$

There are two special cases of particular interest:

(1) *The Quarter-Wave Retarder*: $\delta = \lambda/4$, $\tau = 90^\circ$, $G = H = \frac{1}{2}$ which gives

$$I' = \tfrac{1}{2}[I \pm \tfrac{1}{2}Q \cos 4\psi \pm \tfrac{1}{2}U \sin 4\psi \mp V \sin 2\psi] \tag{5.35}$$

(2) *The Half-Wave Retarder*: $\delta = \lambda/2$, $\tau = 180^\circ$, $G = 0$, $H = 1$ which gives

$$I' = \tfrac{1}{2}[I \pm Q \cos 4\psi \pm U \sin 4\psi] \tag{5.36}$$

The modulation is at four times the rotation angle ψ. Method (2) does not allow the circular component (V) to be determined, but is more efficient in modulating the intensity to derive Q and U, and is the method most often used for stellar polarimetry. Solar magnetographs, on the other hand, must determine the circular component, and Method (1) is the basis for those instruments.

There are many ways to solve these equations for the Stokes parameters. As an example, consider the case of linear polarization (Method 2). The simplest solution, which also emphasizes the direct relation between the Stokes parameters and measured quantities, is to set the angle ψ to four discrete values (0°, 22.5°, 45°, and 67.5°)

which yields

$$\left.\begin{aligned} I'(0^\circ) &= \tfrac{1}{2}(I+Q) & I'(45^\circ) &= \tfrac{1}{2}(I-Q) \\ I'(22.5^\circ) &= \tfrac{1}{2}(I+U) & I'(67.5^\circ) &= \tfrac{1}{2}(I-U) \end{aligned}\right\} \tag{5.37}$$

and solving for I, Q, and U gives

$$\left.\begin{aligned} Q &= I'(0^\circ) - I'(45^\circ) & U &= I'(22.5^\circ) - I'(67.5^\circ) \\ I &= I'(0^\circ) + I'(45^\circ) & I &= I'(22.5^\circ) + I'(67.5^\circ) \end{aligned}\right\} \tag{5.38}$$

Note that I is redundantly determined, and that Q and U have the same units as I. It is also common practice to form the normalized Stokes parameters by taking the ratio of Q/I and U/I and V/I; these ratios are sometimes referred to as the Stokes parameters in the literature and even given the same symbol.

5.3.3 Mueller matrices

To derive some of the equations given, or to explore the impact of optical elements that cause or change polarization, a method of handling all four Stokes parameters at the same time is needed. Mueller calculus is a matrix method for the representation of incoherent, partially polarized light developed in 1943 by Hans Mueller at the Massachusetts Institute of Technology. Each optical element is represented by a 4×4 matrix $[M]$ and the incident light is represented by a 1×4 Stokes vector $\{I, Q, U, V\}$. If the beam of light passes in turn through elements represented by Mueller matrices $M_1, M_2, \ldots, M_n$, then the Stokes vector of the emergent beam (S') is found from

$$S' = M_n M_{n-1} \cdots M_2 M_1 S$$

Because matrix multiplication is not commutative in general, then this order cannot be reversed. Matrix multiplication proceeds by first computing the Stokes vector for the combination $M_1 S = S_1$ and then applying $M_2 S_1$ to find S_2 and so on until $S' = M_n S_{n-1}$.

To simplify matrices for well-known elements such as a perfect polarizer or a half-wave plate, it is necessary to adopt a specific coordinate system intrinsic to the element, such as the polarization axis of a polarizer or the fast axis of a wave plate. In a typical polarimeter like the one in Figure 5.9 the reference axes will be defined by the final polarizer in the system (also called the analyzer). If other elements (partial polarizers, retarders) are oriented with their principal planes at different angles (ψ), then the simplified matrix must be transformed to that coordinate system using the relationship

$$M' = R(-\psi) M R(\psi) \tag{5.39}$$

The rotation matrix $R(\psi)$ is given by

$$R(\psi) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos 2\psi & \sin 2\psi & 0 \\ 0 & -\sin 2\psi & \cos 2\psi & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.40}$$

The rotation matrix is applied twice, once to rotate to the principal axis where the simplified matrix M applies and again to rotate back to the real coordinate system. A set of matrices referred to their principle planes can be found in Clarke and Grainger (1971) and Tinbergen (1996). The latter reference also shows practical instrument designs.

5.4 INTERFEROMETERS

Interferometer techniques are applied in astronomy in two different ways: one as a collection method and the other as a detection method. By combining the light collected by widely separated telescopes one can overcome the diffraction limit of an individual telescope. This powerful technique has long been applied to radio astronomy, but optical and infrared interferometers of remarkable resolution have now been developed. Single-aperture telescopes have also been equipped with interferometer equipment for specific detection purposes. Several types of detection interferometers have been used for spectroscopy, such as the Fourier Transform Spectrometer (FTS) which is a scanning Michelson interferometer (not to be confused with the Michelson Stellar Interferometer), and the Fabry–Perot interferometer which is an imaging spectrometer. Here we summarize the principles of these instruments.

5.4.1 The Fourier Transform Spectrometer (FTS)

The FTS is a scanning Michelson interferometer with collimated light as an input. A typical scheme is shown in Figure 5.10. For a collimated monochromatic beam, the intensity at the detector is determined by the "path difference" $\Delta x = 2(x_b - x_a)$, where x_a refers to the arm containing the fixed mirror A, and x_b is the distance to the scanning mirror B. The phase difference is given by $k\Delta x$ where $k = 2\pi/\lambda$. The fraction of the incident beam in the output is given by

$$T(k, \Delta x) = \tfrac{1}{2}[I + \cos(2k\Delta x)] \tag{5.41}$$

from which it follows that $T = 1$ when the combining beams are in phase and $T = 0$ when they are 180° out of phase. Given an incident beam whose spectrum is $I(k)$, the signal F measured in the output is

$$F(\Delta x) = c \int I(k) T(k, \Delta x)\, dk = \text{constant} + \frac{c}{2} \int I(k) \cos(2k\Delta x)\, dk \tag{5.42}$$

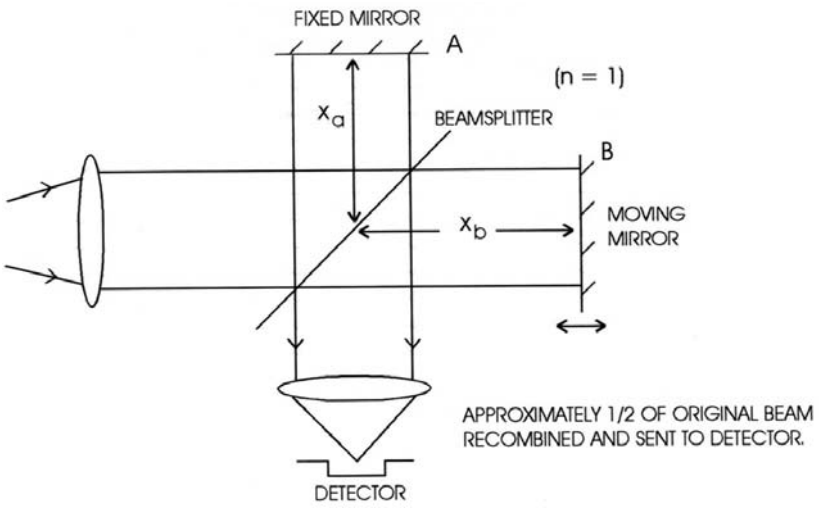


Figure 5.10. The principle of the scanning Michelson interferometer. As the mirror is scanned the intensity recorded by the detector is modulated to produce an interferogram. The spectrum can be extracted by an inverse Fourier transform.

where c is a constant. The measured signal $F(\Delta x)$ is called the interferogram and the last integral is the Fourier cosine transform of the spectrum. Therefore, the transform of the interferogram is $I(k)$.

An FTS can have very large resolving power. For example, since R is given by $4\Delta x_{\max}/\lambda$ and with $\Delta x_{\max} = 10$ cm we get $R = 400{,}000$ at 1 μm wavelength. Moreover, *all* the light falls on the detector so, in principle, the signal-to-noise ratio is high. The primary disadvantage for astronomical work is the fact that the measurements require a time sequence to determine the spectrum, during which atmospheric conditions may vary.

5.4.2 The Fabry-Perot etalon

The Fabry-Perot interferometer is an imaging spectrometer which is formed by placing a device called an "etalon" in the collimated beam of a typical camera system. A typical arrangement is shown in Figure 5.11. The etalon consists of two plane-parallel plates with thin, highly reflective coatings on their inner faces. The plates are in near contact but separated by a distance d. Assuming that the refractive index of the medium in the gap is n (usually $n = 1$) and θ is the angle of incidence of a ray on the etalon (usually very small), then multiple reflections and destructive interference within the gap occurs and the wavelengths transmitted with maximum intensity obey the relation

$$m\lambda = 2nd\cos\theta \tag{5.43}$$

For monochromatic light, the image is a set of concentric rings. To ensure that a sufficiently narrow band of light passes through the system, it is necessary to "pre-

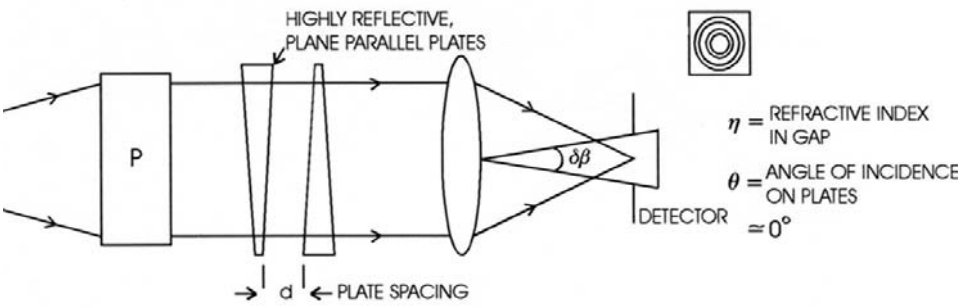


Figure 5.11. A typical arrangement for a Fabry–Perot interferometer. The device *P* is used to narrow the range of wavelengths fed to the etalon.

filter'' the light. This can be done with a very narrow band interference filter. Usually, a circular aperture isolates the central order which has an angular diameter $\delta\beta = \sqrt{(8/R)}$ and the free spectral range is given by

$$\Delta\lambda_{FSP} = \frac{\lambda}{m} = \frac{\lambda^2}{2nd} \tag{5.44}$$

The resolving power ($R = \lambda/\delta\lambda$) is

$$R = \frac{2Fnd}{\lambda} \tag{5.45}$$

where $F(= \Delta\lambda_{FSP}/\delta\lambda)$ is called the ''finesse'' of the etalon, which is a measure of the plate quality and the reflectance (r) of the coatings; $F = \pi\sqrt{r}/(1-r)$ and typical values are 30–50. Defining $\delta = (2\pi/\lambda)(2nd\cos\theta)$, the transmitted intensity is $I(\delta) = I(0)/[1 + (2F/\pi)^2 \sin^2(\delta/2)]$. Among the many applications of Fabry–Perot etalons is the Taurus Tunable Filter (Bland-Hawthorn and Kedziora-Chudczer, 2003) which allows for wide-field (10 arcmin) narrow-band imaging in the CCD range from 370 nm to 1,000 nm with resolving power of 100–1,000. Charge shuffling (see Chapter 4) synchronized to frequency (band) switching is used to suppress systematic errors compared with conventional imaging.

5.4.3 Interference filters

Interference filters are multi-layer thin-film devices that operate with the same principle as the Fabry–Perot etalon just described. As shown in Figure 5.12, incident light undergoes multiple reflections between coated surfaces which define a cavity. Whenever there is no phase difference between emerging wavefronts, constructive interference produces a transmission maximum as given by Equation (5.43). The simplest bandpass filter is a very thin Fabry–Perot etalon in which the air gap is replaced by a thin dielectric layer with a half-wave optical thickness for the desired wavelength. The high reflectors of the cavity are normal quarter-wave stacks with a broad-band reflectance peaking at the design wavelength. This entire assembly of two quarter-wave stacks, separated by a half-wave spacer, can be applied to a single

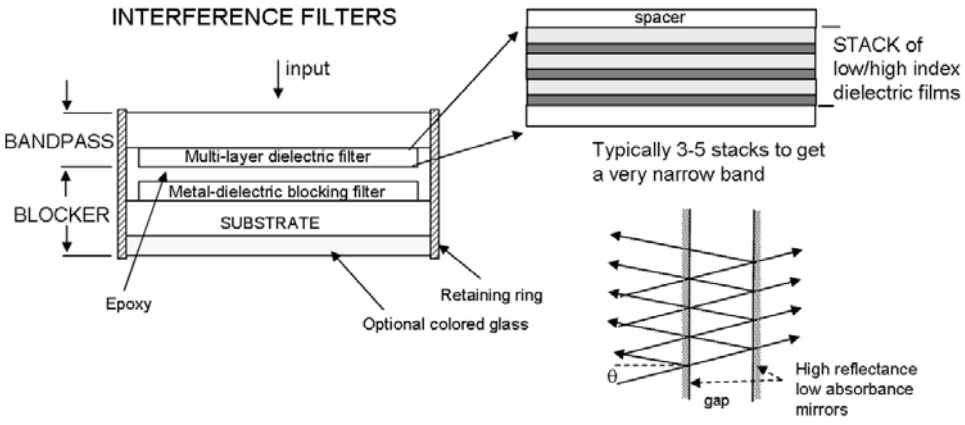


Figure 5.12. Simplified design of a multi-layer interference filter.

surface in one continuous vacuum deposition run. Two or more such filters can be deposited on top of each other separated by an absentee layer, to form a multi-layer filter. With 3–4 layers the band slopes are much steeper and the filters are close to being ''top hat'' or square. The basic construction of the entire filter including the blocking layers is shown in Figure 5.12. Typically, a three-cavity filter would have a base width (where the transmission is 1% of peak) of 1.9–2.2 times the FWHM. In terms of the external angle of incidence (φ), it can be shown that the wavelength of peak transmission at small angles is given by

$$\lambda = \lambda_0 \sqrt{[1 - (n_o/n_e)^2 \sin^2 \varphi]} \tag{5.46}$$

where n_0 is the refractive index of the external medium ($\sim$1); and n_e is the effective refractive index of the spacer. As the angle of incidence increases from zero the center wavelength shifts to the blue, thus ''tilt-scanning'' can be used to provide a short-wavelength scan.

5.5 DETECTORS

5.5.1 Classification

We have tried to group instruments into broad classes, and it is possible to do the same with detectors. Detectors of electromagnetic radiation are generally grouped into three broad groups:

(1) *Photon detectors* in which individual photons release one or more electrons (or other charge carriers) on interacting with the detector material; photon detectors have wide application from gamma rays to the far-infrared.
(2) *Thermal detectors* in which the photon energy goes into heat within the material, resulting in a change to a measurable property of the device, such as its electrical

conductivity; thermal detectors have a broad spectral response but are often used for infrared and sub-millimeter detection.

(3) *Coherent detectors* in which the electric field of the wave is sensed directly and phase information can be preserved. The most common form of coherent detection takes advantage of wave interference with a locally produced field, either before or after conversion of the electromagnetic radiation to an electrical signal. Coherent detectors are used from the far-infrared to the radio.

To distinguish between photon and thermal detectors consider the following. The response of an ideal thermal detector is independent of the spectral distribution of the photons and depends only on the total power (P) absorbed, and therefore its output per watt per unit wavelength interval is flat, independent of wavelength. On the other hand, a photon detector measures the rate of arrival of photons ($N = P/h\nu$), and as the number per second per watt of incident power ($N/P = \lambda/hc$) is proportional to wavelength, its response increases with wavelength up to some maximum wavelength where the photon energy is no longer sufficient to produce a photoelectric event. Photon detectors can be subdivided into (1) photoemission devices employing the external photoelectric effect in which the photon causes a charge carrier (electron) to be ejected from the material and (2) photoabsorption devices that use the internal photoelectric effect in a semiconductor to free a charge carrier within the material.

The most well-known detector in the photoemission category is the photocathode of a photomultiplier tube (PMT) already described, in which an electron is emitted from the photocathode surface and subsequently amplified by a cascade of impacts with secondary surfaces before being detected as a charge pulse. Photoemissive materials with large work functions can provide excellent detectors far into the ultraviolet. Most importantly, it is possible to create ultraviolet imaging devices based on this process. For example, long, narrow curved tubes or "microchannels" of lead oxide (PbO) can perform the same function as the secondary surfaces in a PMT resulting in a large pulse of electrons emerging from the end provided there is a potential gradient. Such channels can be packaged very close together (like straws in a box) to make a two-dimensional array of microchannels. We will discuss this important class of devices later under UV imaging methods. Finally in this category, some materials emit other lower energy photons (fluoresce) rather than electrons which enables detection of the more energetic photon by a process called "down-conversion"; we will see that this has applications to CCDs in Section 8.8.

Photo-absorption is the largest category, with many possible outcomes, including chemical change (as in photography). In this book we are concerned mainly with absorption processes in semiconductor devices, and there are essentially two basic types of interactions: the photoconduction effect and the photovoltaic (or photodiode) effect. The photoconductor is composed of a single uniform semiconductor material in which the conductance is changed by the creation of free charge carriers in the material when photons are absorbed. There is usually always an external applied electric field. In the photodiode (or photo-junction), internal electric fields and potential barriers are created by suitable junctions between different materials or deliberate variations in the electrical properties of the material so that photo-generated carriers

in these regions respond to those fields. Before proceeding further it is important to review the properties of semiconductors in general.

5.5.2 Semiconductors

The properties of any solid material depend on both the atomic structure of the atoms of the material and the way the atoms are arranged within the solid (i.e., the crystal structure). Electrons can exist in stable orbits near the nucleus of an atom only for certain definite values of their energy. When individual atoms come close together to form a solid crystal, electrons in the outermost orbits, or upper energy levels, of adjacent atoms interact to bind the atoms together. Because of the very strong interaction between these outer or "valence" electrons, the outer orbits and therefore the upper energy levels are drastically altered. The result is that the outer electrons are shared between the different atomic nuclei. A simple diagram depicting the "energy levels" of the electrons for a combination of two atoms would therefore have two permitted levels near the core of each atom. A combination of three atoms would have three levels near the core because the outer electrons of all three atoms can be shared. The higher, unoccupied orbits would also split, indicating that they too can in principle take two or three electrons. Even the tiniest sliver of a real crystal will contain many hundreds of millions of atoms, and so there are a huge number of split levels associated with each atom in the crystal because of the sharing of outer electrons. In other words, the energy levels or orbits are spread out into a "band". The lowest band of energies, corresponding to all the innermost orbits of the electrons, is filled with electrons because there is one electron for each atom. This band of allowed, filled energy levels is called the "valence band". Conversely, the upper energy band is empty of electrons because it is composed of the combined unoccupied higher energy levels or orbits of the individual atoms in the crystal. It is called the "conduction band" for reasons that will become apparent. Thus, the individual atoms have a gap between the permitted inner and outer orbits (i.e., a gap in energy between the inner filled levels and the outer unoccupied levels). The energy region between the valence band and the conduction band in the crystal must be a "forbidden energy gap" (E_G). Figure 5.13 summarizes this description. Note that the crystal must be pure and contain atoms of only one kind, otherwise additional energy levels corresponding to those atoms will be formed. More importantly, the periodic or repetitive crystalline structure must be unbroken to avoid distortions in the energy levels caused by abnormal sharing of electrons. Of course, in practice both of these conditions are violated in real crystals, and departures from the simplified model presented here contribute to degraded performance of devices such as transistors and CCDs. In metals, the valence and conduction bands overlap, and so any of the many valence electrons are free to roam throughout the solid to conduct electricity and heat, and to move in response to the force of an electric field; an electric field could be produced by attaching a battery to both ends of the piece of metal. An insulating material, on the other hand, has a highly ordered structure and a very wide forbidden energy gap. The conduction band is totally empty of electrons and so cannot contribute to an electrical current flow. Electrons in the completely

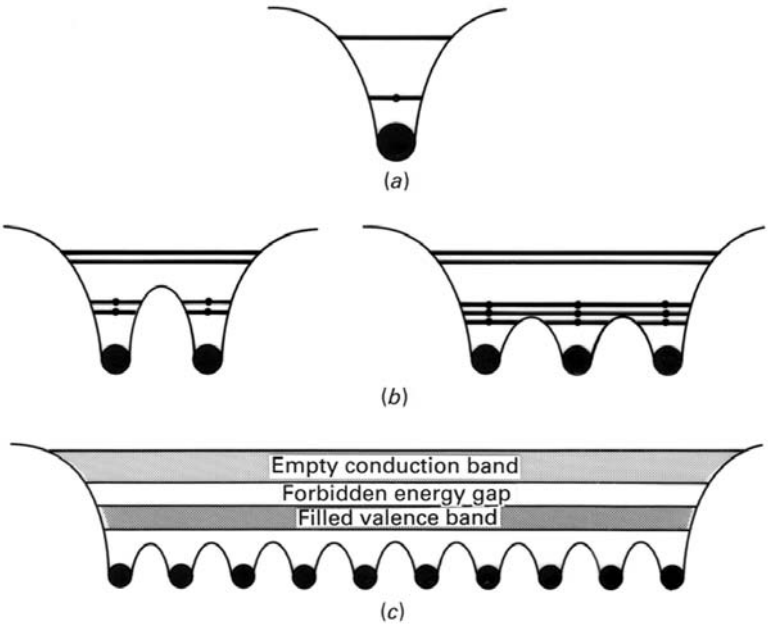


Figure 5.13. Simplified schematic of the formation of a bandgap in a semiconductor crystal.

filled valence band cannot move in response to an electric field because every nearby orbit is occupied.

In a semiconductor, a few electrons can be elevated from the valence band to the conduction band across the forbidden gap merely by absorbing heat energy from the random, microscopic, jostling motions of the crystal structure at normal "room" temperature. Thermal energy is given approximately by

$$E_{\text{th}}(\text{eV}) = kT = 0.026(T/300)\,\text{eV} \tag{5.47}$$

where k is Boltzmann's constant; and T is absolute temperature. At room temperature ($T = 300\,\text{K}$) the thermal energy is quite small at 0.026 electron volts. Electrons promoted to the conduction band can then conduct electricity (i.e., they are free to move under the influence of an electric force field). Interestingly, the corresponding vacancies or "holes" left in the valence band allow it to contribute to electrical conductivity as well because there is now somewhere for electrons in adjacent atoms to go; descriptions of solid-state devices therefore refer to "electron–hole" pairs.

Because the number of electrical charge carriers (electrons in the conduction band, holes in the valence band) is much less than in the case of a metal, semiconductors are poorer conductors than metals but better than insulators. The width of the forbidden energy gap in semiconductors is an important quantity. Most semiconductor crystals have bandgap energies around 1 eV, but the range is from almost 0 eV to about 3.5 eV. As shown above, 1 eV is roughly 38 times larger than the thermal or heat energy in the crystal atoms at room temperature. Remember also

Table 5.1. Part of the periodic table of elements showing the location of semiconductors.

IB	*IIB*	*IIIA*	*IVA*	*VA*	*VIA*	*VIIA*
		5 B Boron	6 C Carbon	7 N Nitrogen	8 O Oxygen	
		13 Al Aluminum	14 Si *Silicon*	15 P Phosphorus	16 S Sulfur	17 Cl Chlorine
29 Cu Copper	30 Zn Zinc	31 Ga Gallium	32 Ge *Germanium*	33 As Arsenic	34 Se Selenium	35 Br Bromine
47 Ag Silver	48 Cd Cadmium	49 In Indium	50 Sn Tin	51 Sb Antimony	52 Te Tellurium	53 I Iodine
79 Au Gold	80 Hg Mercury	81 Tl Thallium	82 Pb Lead	83 Bi Bismuth		

http://www.webelements.com/

that visible light photons have energies around 2.25 eV (for 550 nm). As the number of electrons which can be promoted to the conduction band by absorbing heat will vary with the temperature of the crystal, typically as $\exp(-E_G/2kT)$, those semiconductors with larger bandgaps are preferred because transistors and other devices made from them will be less sensitive to environmental changes. For this reason silicon is preferred to germanium. If the semiconductor is cooled to a low temperature, random elevation of valence electrons to the conduction band can be virtually eliminated.

Table 5.1 is a section of the periodic table of the elements showing that the primary semiconductors like silicon and germanium belong to the "fourth column" elements, which also includes carbon. Each of these elements has four valence electrons. Compounds of elements on either side of the fourth column can be formed, and these alloys will also have semiconductor properties; gallium arsenide (GaAs) and indium antimonide (InSb) are III–IV (or "three–four") compounds and mercury–cadmium–telluride (HgCdTe) is one possible II–VI (or "two–six") compound. Column numbers indicate the number of valence electrons. Small numbers with the symbols are the atomic numbers (number of protons or electrons).

When a photon is absorbed in the crystalline structure of silicon, its energy is transferred to a negatively charged electron, the photoelectron, which is then displaced from its normal location in the valence band into the conduction band. When the electron reaches the conduction band it can migrate through the crystal. Migration can be stimulated and controlled by applying an electric field to the silicon crystal by means of small metal plates called "electrodes" or "gates" connected to a voltage source. Absorption of photons in silicon is a function of the photon energy

(and hence wavelength). The photon flux at depth z in the material is given by

$$I(z) = I(0)e^{-\alpha z} \tag{5.48}$$

where α is the absorption coefficient. At a temperature of 300 K (27°C), $\alpha \sim 5\,\mu m^{-1}$ at $\lambda = 400$ nm in the blue, but only $0.1\,\mu m^{-1}$ at 800 nm in the far-red; the lower the value of α the greater the depth reached for the same absorption ratio. At 77 K (liquid nitrogen temperature) α reduces to $4.0\,\mu m^{-1}$, $0.25\,\mu m^{-1}$, and $0.005\,\mu m^{-1}$ at 400 nm, 600 nm, and 800 nm, respectively. Clearly, red (low-energy) photons pass deeper into the silicon before being absorbed. Eventually, for the reddest light (i.e., the lowest energy photons), there is simply not enough energy to elevate a valence electron to the conduction band. In other words, for each semiconductor there is a wavelength of light beyond (redder than) which the material is insensitive to light because the photons are not energetic enough to overcome the forbidden energy gap (E_G) in the crystal. The cutoff wavelength is given by

$$\lambda_c = \frac{hc}{E_G} \tag{5.49}$$

where h is Planck's constant; and c is the speed of light; $hc = 1.24$ for wavelengths in microns and energy in electron volts. These wavelengths are given in Table 5.2. Some of the materials listed are sensitive well into the infrared region and will be discussed again in Chapter 11.

Table 5.2. Forbidden energy gaps for some common semiconductors.

Name	*Symbol*	T (K)	E_G (eV)	λ_c (μm)
Gallium nitride	GaN	295	3.45	0.36
Silicon carbide	SiC	295	2.86	0.43
Cadmium sulfide	CdS	295	2.4	0.5
Cadmium selenide	CdSe	295	1.8	0.7
Gallium arsenide	GaAs	295	1.35	0.92
Silicon	Si	295	1.12	1.11
Germanium	Ge	295	0.67	1.85
Lead sulfide	PbS	295	0.42	2.95
Indium antimonide	InSb	295 77	0.18 0.23	6.9 5.4
Mercury cadmium telluride	$Hg_xCd_{1-x}Te$	77	0.1 ($x = 0.8$) 0.5 ($x = 0.554$)	12.4 2.5

See: *http://www.semiconductorsdirect.com*

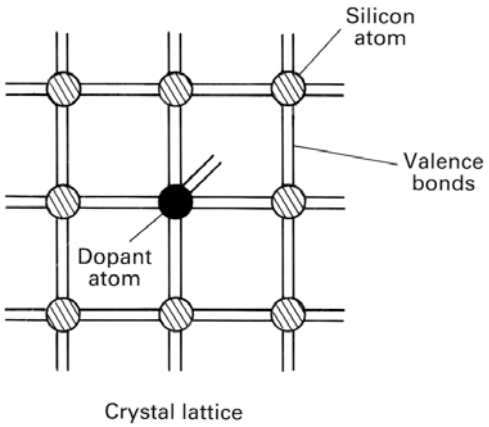


Figure 5.14. Impurity atoms in a semiconductor (silicon) crystal lattice have a "doping" effect in terms of donating or accepting electrons.

All of the materials listed in Table 5.2 are considered "intrinsic" semiconductors because each has a well-defined bandgap intrinsic to the material. Of course, real band structures are more complex, and not all materials have bandgaps that are "direct" in the sense discussed here, including silicon, due to the need to conserve momentum as well as energy in the transition, but the overall effect is a more complicated behavior of absorption coefficients. It is also possible to create a different kind of semiconductor known as an "extrinsic" semiconductor in which impurity atoms produce intermediate energy levels within the forbidden gap. For example, when silicon atoms in the crystal structure are deliberately replaced with other atoms the semiconductor is said to be "doped" (Figure 5.14). If the impurity atom has more valence electrons than the semiconductor then it will donate these negative charges to the conduction band; such a material is called *n-type*. Conversely, if the impurity atom has fewer valence electrons than the semiconductor then a positively charged hole is left in the valence band ready to accept any available electrons; this material is a *p-type* semiconductor. In p-type material there is an excess of holes, and so electrons

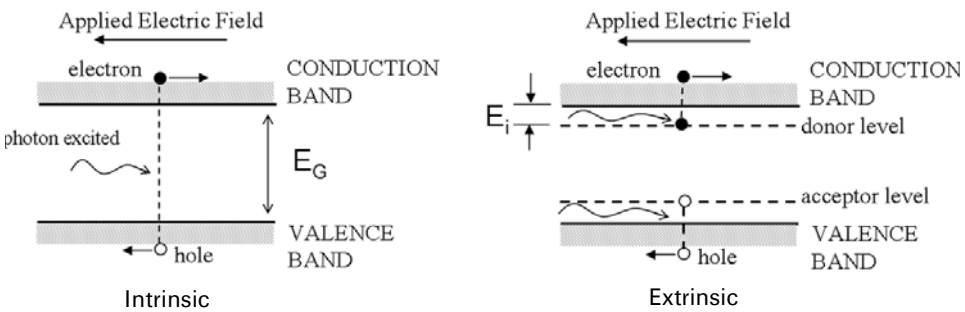


Figure 5.15. An intrinsic bandgap and the location of energy levels within the bandgap due to doping to form an extrinsic semiconductor.

Table 5.3. Extrinsic semiconductors, doping material, and long-wavelength cutoff.

Base	*: Impurity*	λ_c (μm)	*Base*	*: Impurity*	λ_c (μm)
Silicon (Si)	: In	8.0	Germanium (Ge)	: Au	8.27
	: Ga	17.1		: Hg	13.8
	: Bi	17.6		: Cd	20.7
	: Al	18.1		: Cu	30.2
	: As	23.1		: Zn	37.6
	: P	27.6		: Ga	115
	: B	28.2		: B	119.6
	: Sb	28.8		: Sb	129

are said to be the minority carriers of charge, whereas the opposite is true for n-type material.

Because of the much lower transition energies, extrinsic semiconductors are used in infrared photon detection. A photon with insufficient energy to excite an electron–hole pair directly can still be the cause of an excitation from an energy level associated with the impurity atom (see Figure 5.15). Several extrinsic semiconductors are listed in Table 5.3 together with approximate values of their long wavelength cutoff points. It is customary to write the symbol for the semiconductor first followed by a colon and then the doping material (e.g., Si : As).

5.5.3 Photoconductors and photodiodes

Photoconductor: This is the simplest application of a semiconductor for detection of photons. A typical photoconductor arrangement is shown in Figure 5.16. Photons are absorbed and create electron–hole pairs. If the material is extrinsic rather than intrinsic, then E_i must be substituted for E_G. Also, for extrinsic materials there are limits on solubility of the dopants, and high concentrations introduce unwanted conductivity modes such as "hopping" which involves conduction between neighboring impurity atoms without raising an electron to the conduction band. In the discussion below we assume that the semiconductor has been cooled to eliminate thermally generated charges. In practice, both electrons and holes contribute to the photocurrent, but it is usually the electrons that dominate. The average photocurrent (I) between the terminals that is generated by an incident flux with power P (watts) is given by

$$I = (e\eta P/h\nu)(v\tau/l) \tag{5.50}$$

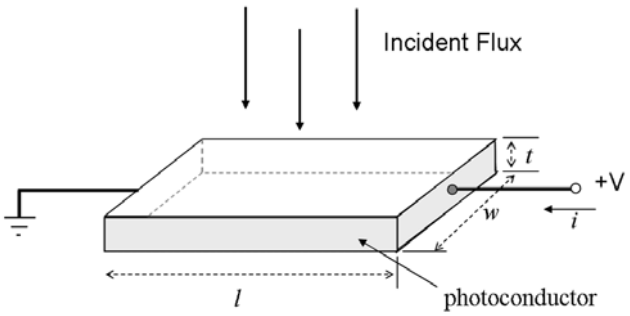


Figure 5.16. The basic construction and operation of a semiconductor used in photoconduction mode.

In this expression η is the quantum efficiency; and $P/h\nu$ is just the photon arrival rate. The quantity τ is called the mean carrier lifetime and measures how long the photogenerated charge exists before recombination. Values are usually less than to much less than a few milliseconds but depend on doping and temperature. The average charge carrier velocity is v, which is related to the applied electric field across the photoconductor $E = V/l$ by $v = \mu E$, where μ is called the mobility of the charge carrier. Thus, l/v is the transit time across the device from one terminal to the other, and the quantity $G = v\tau/l$ is just the ratio of mean carrier lifetime to transit time. It is known as the "photoconductive gain". The response of the detector (in amps per watt or volts per watt) is just I/P or V/RP where V is the bias voltage across the photoconductor, and the resistance R due to the photocurrent is $l/\sigma A$ and the conductivity $\sigma = ne\mu$, where n is the average density of carriers. It follows that $S = (e\eta G/hc)\lambda$. Finally, the root-mean-square noise for a photoconductor is given by $\sqrt{(4eGIB)}$ where B is the electrical bandwidth of the measurement.

Photodiodes: Junctions between p-type and n-type regions are used many times in semiconductor structures to produce different devices. One such device is the photodiode. When a p–n junction is formed, electrons from the n region tend to diffuse into the p region near the junction and fill up some of the positively ionized states or holes in the valence band thus making that p-type region more negative than it was. Similarly, the diffusion of holes from the p-side to the n-side leads to an increasingly more positive electrical potential. A narrow region forms on either side of the junction in which the majority charge carriers are "depleted" relative to their concentrations well away from the junction. As the concentration of electrons in the n-type material is usually very much larger than in the p-type material, the flow of electrons would tend to be one way were it not for the fact that the diffusion process itself begins to build up an electrostatic potential barrier which restrains the flow of electrons from the n-type region; the build-up of electrons on the p-side makes it negatively charged which starts to repel further diffusion. The magnitude of this potential barrier (V_0) depends on impurity concentrations (i.e., on the number of donor electrons at the junction that are available for transfer to nearby acceptor

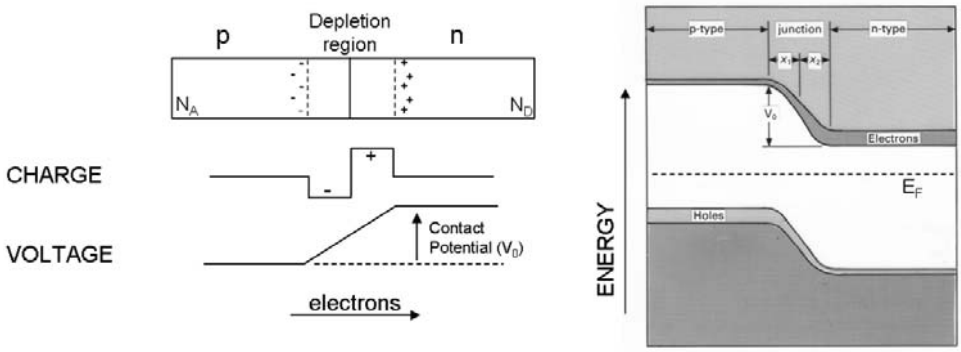


Figure 5.17. The formation of a p–n junction between p-doped and n-doped materials results in a region depleted of carriers and the creation of a potential barrier.

levels) and is just equal to the required shift of the energy bands needed to ensure that the Fermi level (E_F) remains constant throughout the crystal. The Fermi level is the energy at which there is a 50/50 chance of the corresponding electron energy state or orbit being occupied by an electron. For an intrinsic semiconductor, E_F lies half-way between the valence and conduction bands whereas for an n-type doped semiconductor the Fermi level moves up toward the conduction band, and conversely p-type doping lowers the Fermi level.

The potential drop across the "depletion layer" is about 0.6 V for silicon and 0.3 V for germanium, half of the forbidden energy gap. Figure 5.17 shows how this condition is represented by an energy band model of voltage or potential vs. distance from the junction. The width of the junction region ($X_1 + X_2$) increases with increasing voltage (V_0) and decreasing number density of acceptor atoms.

If a positive voltage is applied to the p-side of the junction it will tend to counteract or reduce the built-in potential barrier and attract more electrons across the junction, whereas a negative voltage on the p-side will enhance the internal barrier and increase the width of the depletion region; these conditions are called "forward" and "reversed" bias, respectively. Therefore, on one side of a p–n junction there is a region which is more negative and on the other side there is a region which is more positive than elsewhere in the crystal. When light of the correct wavelength is absorbed near the junction an electron–hole pair is created and the potential difference across the junction sweeps the pair apart before they can recombine. Electrons are drawn toward the region of greatest positive potential buried in the n-type layer which therefore behaves like a charge storage capacitor. Of course, as more electrons accumulate, the positive potential is progressively weakened. In the photodiode, an electron–hole pair is created within the depletion region by the absorption of a photon, and the charge carriers are immediately separated by the electric field across the junction. The current due to an incident photon flux (signal and background) of power P is just $I = e\eta P/h\nu$ and the root-mean-square noise is given by $\sqrt{(2eIB)}$, where B is the electrical frequency bandwidth of the measurement. Comparing this with the photoconductor we see that $G = 1$ for the photodiode and that the

noise is less by a factor of $\sqrt{2}$ because recombination does not occur in the depletion region.

5.5.4 Thermal detectors

In the class of thermal detectors there is only one type that is used widely in astronomy and that is the "bolometer" (see Figure 5.18). Semiconductor bolometers, based on silicon or germanium, are well-developed for far-infrared and sub-millimeter astronomy. Bolometers are difficult to construct in general and are used mainly under very high background conditions. Essentially, a bolometer consists of a sensitive thermometer and a high cross-section absorber that absorbs almost all of the incident radiation falling on it (i.e., $QE(\eta) \sim 100\%$). The absorber has a heat capacity of C joules per kelvin. The thermometer and absorber are connected by a weak thermal link to a heat sink at a low temperature, and the conductivity of this link is G watts per kelvin.

Suppose the detector element of the bolometer absorbs an amount of energy $E = \eta P \Delta t / h\nu$ in a time interval Δt from a source with power P. The energy is converted to heat which raises the temperature by an amount $\Delta T = T - T_0 = E/C$ above that of a heat sink at T_0. The temperature rise decays exponentially as power in

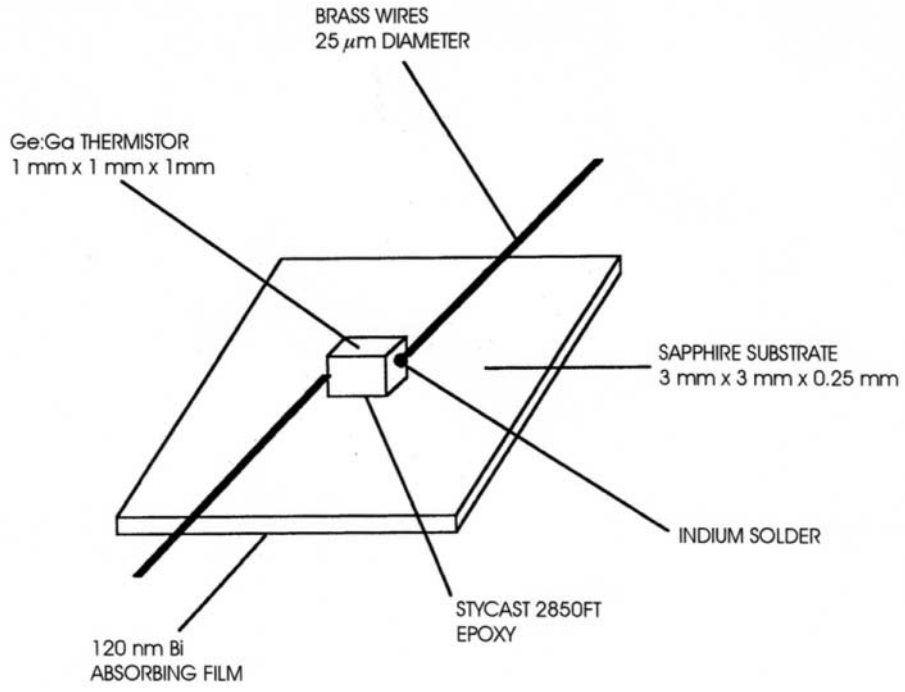


Figure 5.18. Basic layout of a classical, single-element bolometer using a 1 mm cube of germanium doped with gallium as the heat-sensitive device.

the absorber flows out to the heat sink via the weak link, and the time constant is $\tau = C/G$. The temperature rise is proportional to the absorbed energy. In the classical circuit, a constant bias current, generated by the bias supply and a load resistor, flows through the bolometer. Provided that the bias power remains constant

$$T = T_0 + (P + P_{\text{bias}})/G \tag{5.51}$$

and the temperature rise causes a change in bolometer resistance, and consequently a change in the voltage across it which can be measured. Differential on/off source measurements are normally required to remove bias levels. A much more detailed account is given in Rieke (2003). Some instruments now have bolometer arrays and these will be discussed with sub-millimeter techniques in Section 13.4.

5.5.5 Coherent detectors

Any device that directly responds to the electric field of the wave or photon can be considered a coherent detector or, more traditionally, a coherent receiver. The most important form of receiver is the heterodyne or super-heterodyne which functions by mixing signals of different frequencies and detecting a signal at the difference or "beat" frequency between the original two frequencies. Depending on the frequency of the incoming wave, the electric field may be converted to an electrical signal which is then amplified before being mixed with a Local Oscillator. For frequencies below 1 GHz, cryogenic transistor pre-amplifiers are used. From 1 GHz to 40 GHz, FET, parametric, and maser amplifiers are employed, but above 40 GHz the mixer must precede the pre-amplifier in order to reduce the frequency before amplification. The Local Oscillator (LO) produces a strong signal with a frequency that is close to but different from the signal frequency. The beat frequency, or intermediate frequency (IF), is $\nu_{\text{IF}} = \nu_{\text{S}} - \nu_{\text{LO}}$. Subsequent amplification and filtering of the intermediate frequency signal by a large factor then follows and the resultant signal is rectified by a diode and integrated. The key element in this entire process is the mixer. This device must be a non-linear device that converts power from the original frequencies to the beat frequency, and is typically a diode because it is a good approximation to a square-law device in terms of its current–voltage (I–V) behavior. If $I = V^2$, then the output current is proportional to the input power P, because V^2 is proportional to the electric field E^2 which is a measure of P. Within the mixer the electric fields sum as vectors and the resultant power is the square of the amplitude of the electric fields. One of the best modern mixers is a junction made by separating two *superconductors* (not semiconductors) with a thin insulator; this is the SIS mixer, and will be described in more detail in Chapter 13.

5.6 SUMMARY

Consideration of the basic layout of an imaging system reveals the importance of matching the physical pixel size of the detector to the appropriate angular scale on the sky. The basic dispersion relationships for prisms, gratings, and grisms are the

fundamental starting point for selecting a suitable spectrometer design. It was shown that as telescopes get larger, gratings and spectrographs must get larger too. The concept of the grism provides an extremely compact way to convert a camera into a spectrometer. Polarization modules can be added to cameras and spectrographs to convert vibration information into measurable intensity modulations. Interference effects provide the basis of instruments such as the Fabry–Perot interferometer or the scanning Michelson interferometer (FTS). Most of the concepts discussed apply to UV and IR instruments too. Detectors are grouped into one of three categories: photon, thermal, or coherent. Photon detectors either use the external photoelectric effect like a photocathode, or the internal photoelectric effect. CCDs, photodiodes, and photoconductors are all based on the internal photoelectric effect in a suitably doped semiconductor. There are many other useful semiconductors in addition to silicon.

5.7 EXERCISES

1 Describe a basic photoelectric photometer. What precautions would you take to ensure that the signal remained constant even if the star drifted off-center in the aperture? How would your design change if this were an "imaging" system? How would you extract a measurement of the magnitude of the star in this case?

2 Calculate the f/number of a camera lens system required to match 24 μm pixels to $0.2''$ on the sky for a 10 m telescope. Comment on whether or not this would be challenging. What is the field of view for a $1{,}024 \times 1{,}024$ pixel detector?

3 Consider the design of a diffraction grating spectrometer for a 10 m telescope. The 2-pixel resolution element is $0.5''$ and the required resolving power is $R = 20{,}000$. Assume that the configuration is Littrow. Two gratings are available, a first-order grating blazed at 17.5° ($2 \tan \theta_B = 0.63$) and an echelle grating blazed at 63.5° ($2 \tan \theta_B = 4$). Determine the collimator size D_{coll} in both cases. Which is more practical? Assuming the telescope has an f/ratio of 15, what is the focal length of these two collimators?

4 Describe what is meant by a grism. Design a grism with an index of refraction of 2.4 and an apex angle of 30° which will have a central wavelength of 2.2 microns in the near-infrared and a resolving power of $R = 500$ for 2 pixels. Assume the pixel size is 27 microns.

5 Explain the terms linear and circular polarization. How can a camera or a spectrometer be converted to measure polarization? Describe three kinds of polarization modulator and state one advantage and one disadvantage of each.

6 A polarization experiment provides counts at four angular settings each 22.5° apart of a half-wave plate $N(0) = 2{,}000$, $N(22.5) = 1{,}800$, $N(45) = 1{,}000$, $N(67.5) = 1{,}200$. Determine the normalized Stokes parameters, the degree of polarization, and the position angle of the direction of vibration. Assuming that

the errors of measurement follow a Poisson distribution so that $\sigma(N) = \sqrt{N}$, estimate the error in the normalized Stokes parameters.

7 Describe the design of a Fabry–Perot interferometer. For a resolving power of $R = 20{,}000$ at $\lambda = 0.5\,\mu m$ with an air-spaced etalon of Finesse 40, what is the gap d and the free spectral range $\Delta\lambda_{FSP}$?

8 What is the required scan length of an FTS working at a wavelength of $10\,\mu m$ in the mid-infrared if the required resolving power is $R = 100{,}000$?

9 What are the three classes of detectors? Distinguish between the photoelectric effect as it occurs in a photomultiplier tube and the creation of electron–hole pairs in a photoconductor.

10 Explain the basic process in a p–n junction that leads to charge separation and the creation of an internal electric field. What happens if a positive voltage is applied to the p-side of the junction?

5.8 REFERENCES AND SUGGESTED FURTHER READING

Baldry, I.K.; Bland-Hawthorn, J.; and Robertson, J.G. (2004) Volume Phase Holographic Gratings: Polarization properties and diffraction efficiency, *Pub. Astron. Soc. Pacific*, **116**, 403–414.

Barden, S.C.; Arns, J.A.; Colburn, W.S.; and Williams, J.B. (2000) Volume-Phase Holographic Gratings and the efficiency of three simple Volume-Phase Holographic Gratings, *Pub. Astron. Soc. Pacific*, **112**, 809–820.

Bianco, A.; Bertarelli, C.; Conconi, P.; Molinari, E.; Quaglia, C.; Toso, G.; Zerbi, F.M.; and Zerbi, G. (2006) New developments in photochromic materials for volume phase holographic gratings, in E. Atad-Ettedgui, J. Antebi, D. Lemke (eds.), *Optomechanical Technologies for Astronomy*, SPIE, Bellingham, WA [*Proc. SPIE*, **6273**, 120].

Bland-Hawthorn, J.; and Kedziora-Chudczer, L. (2003) Taurus Tunable Filter: Seven years of observing, *Pub. Astron. Soc. Australia*, **20**, 242–251.

Born, M.; and Wolf, E. (1999) *Principles of Optics*, seventh edition, Cambridge University Press, Cambridge, U.K.

Clarke, D.; and Grainger, J. (1971) *Polarized Light and Optical Measurement*, Pergamon Press, Oxford, U.K. [An introduction to polarimetry.]

Henden, A.; and Kaitchuck, R. (1990) *Astronomical Photometry*, Wilmann-Bell, Richmond, VA.

Oke, J.B. (1974) Absolute spectral energy distributions for white dwarfs, *Astrophys. J. Supplement*, **27**, 21.

Rieke, G.H. (2003) *The Measurement of Light from the UV to the Sub-millimeter*, Cambridge University Press, Cambridge, U.K.

Schroeder, D.J. (2000) *Astronomical Optics*, Academic Press, San Diego, CA.

Tinbergen, J. (1996) *Astronomical Polarimetry*, Cambridge University Press, Cambridge, U.K.

6

Designing and building astronomical instruments

There are many important factors and constraints to be aware of when developing new astronomical instrumentation, whether for small or large telescopes. Of course, complete engineering details are beyond the scope of this book, or any one book, but the following sections will at least provide an appreciation for what is involved and serve as a starting point for newcomers to instrument building.

6.1 BASIC REQUIREMENTS

Understanding the application is the very first step. What are the science goals? Sometimes the goals are fairly general, such as "provide the most sensitive camera with the widest possible field of view consistent with median seeing conditions". This approach is reasonable, on the grounds that the uses of such an instrument are so numerous. On the other hand, the science goals may be quite specific, such as "provide an instrument to search for planets around other stars via Doppler velocities in the 3 m/s range", or "provide an instrument to carry out a survey of redshifts ($z > 1$) of a very large statistical sample of faint galaxies". In these cases the spectrographs involved would be quite different from each other and different from a conventional "workhorse" spectrograph. For the planetary search the spectral resolution needs to be very high, and this instrument must provide exceptional long-term stability, whereas for the faint galaxy survey the resolution is much lower, but numerous slit-masks and/or optical fibers are needed to provide a large multi-object advantage. Clearly, the choice of instrument and the details of the design will depend on the kind of science to be done. If it is imaging science, what field of view is required? What are the angular resolution and the wavelength range requirements? Is temporal resolution an issue? Are the measurements going to be read–noise limited or background-limited? The basic requirements must come from the science goals, but be aware of creating a monster with many heads! Too many scientific options in

one instrument can be disastrous. In turn, the science requirements are used to generate a "specification" for the instrument. Candidate designs can then be analyzed in a Conceptual Design phase (also called a System Design phase) and the best or most appropriate design selected. There is always more than one way to achieve the desired goals, and changes in technology result in fresh approaches and new ways to improve older methods.

6.2 OVERALL SYSTEM LAYOUT

Laying out all the essential "building blocks" and their interconnections in pictorial form is the next step in the design process. Invariably a modular approach to the instrument works well. A very simple "block" diagram adapted from the author's first infrared camera is shown in Figure 6.1. The technology is old but the principle is not, and unlike complex modern instruments this diagram is readable on a small page. In its most basic features this infrared camera system is similar to all other imaging systems irrespective of the detector. Although it is hard to generalize, certain building blocks are almost always present in an astronomical instrument. These include

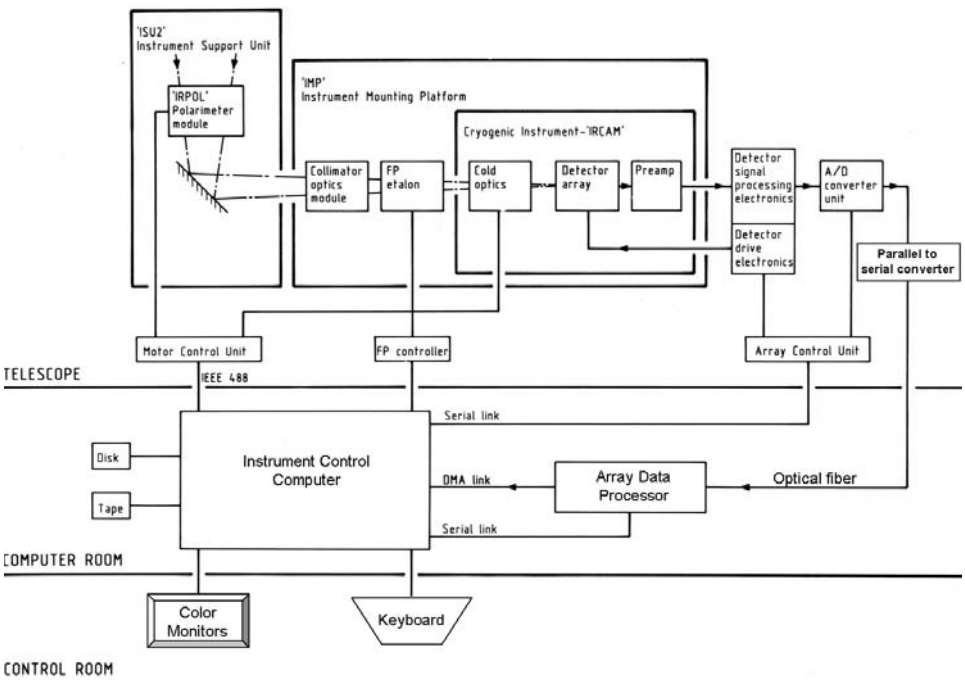


Figure 6.1. A block diagram layout of an entire camera system for a large telescope. The illustration is for IRCAM, the first infrared camera system developed for the 3.8 m U.K. Infrared Telescope.

(1) the detector (the photon sensor itself and circuitry packaged close by)
(2) an opto-mechanical system (lenses, mirrors, filters, gratings, fibers, mounts)
(3) an enclosure and cooling system (for the detector and other parts of the instrument)
(4) signal-processing hardware (e.g., amplifiers and noise suppression circuits) and the analog-to-digital converter (ADC or A/D)
(5) detector "drive" electronics (pulsed and dc bias circuits)
(6) timing logic and synchronization circuits
(7) a "motion control" system and "housekeeping" system (e.g., temperature control)
(8) an electronic interface to a computer (e.g., ethernet, telemetry)
(9) a host computer and peripherals
(10) an image display system and image processing/restoration software.

These ten items form the basis of a great many astronomical instruments employing some form of electronic imaging device. In fact, the above list could apply to almost any form of detector system used in astronomy if the items are understood in their most general sense. At the heart of all instruments is the detector. Usually, it is the performance of the detector system that determines whether the instrument is "state of the art" or not.

6.3 OPTICAL DESIGN

In practice, the optical design of a modern astronomical instrument is likely to be carried out by a professional designer with considerable experience of the application. However, some training in optical design for all young observational astronomers and for anyone interested in building astronomical instruments is very valuable. Understanding the issues, recognizing problems, and being able to communicate requirements is important. Courses are available which teach practical methods of ray tracing and optical design. Excellent texts on optics include Born and Wolf (1999), Fischer and Tadic (2000), Hecht (2001), Kingslake (1978), Schroeder (2000), and Smith (2000).

6.3.1 First order to ray tracing

A good strategy is to break down the optical design into the following steps and stages:

(1) "first-order" requirements
(2) constraints
(3) performance specification
(4) ray tracing and optimization
(5) tolerance analysis.

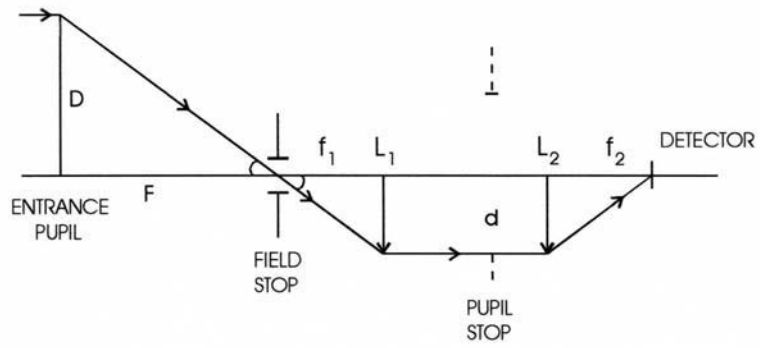


Figure 6.2. A "first-order" optical layout to collimate the beam from the telescope and then re-image the field of view onto a detector. Pupil size and location is determined together with magnification and field-of-view constraints.

By "first-order" requirements is meant a simple design, using "thin-lens" formulae, and known facts about the system, such as the f/number and plate scale of the telescope, the object/image distances and the location of any pupils formed, the required field of view, and the desired magnification or image scale at the detector. A typical "first-order" layout is shown in Figure 6.2. An image of the sky is formed at the telescope focal plane, and this "object" is then re-imaged by the instrument optics onto the detector. Depending on the ratio of the focal lengths of the transfer optics, there will be a change in magnification of the final image. In combination with all optics upstream, the transfer optics will form an image of the primary mirror of the telescope at some location. This image is the entrance pupil. At this stage, the simple equations of elementary optics given in Chapter 3 can be employed. The "thin-lens" equation and the equivalent form for single mirrors can now be applied. It is not likely that simple lenses will be adequate when aberrations are considered, but the required power, beam sizes, and detector scales can be estimated. Likewise, the following simple relationships on displacements and deviations illustrated in Figure 6.3 can prove useful in laying out the initial design:

$$\left.\begin{aligned} \alpha &\approx (n-1)A && \text{thin wedge} \\ z &= \frac{(n-1)t}{n} && \text{parallel plate in converging beam} \\ d &= t\sin\theta\left(1-\frac{\cos\theta}{n\cos\theta'}\right) && \text{displacement by parallel plate} \end{aligned}\right\} \tag{6.1}$$

where α is the angular displacement caused by a wedge of small angle A (angles in radians); z is the longitudinal (focus) displacement caused by a plane-parallel plate of thickness t perpendicular to a converging (or diverging) beam; and d is the lateral displacement caused by a plane-parallel plate at angle θ in a parallel (collimated) beam. These formulae are useful when considering the effects of filters,

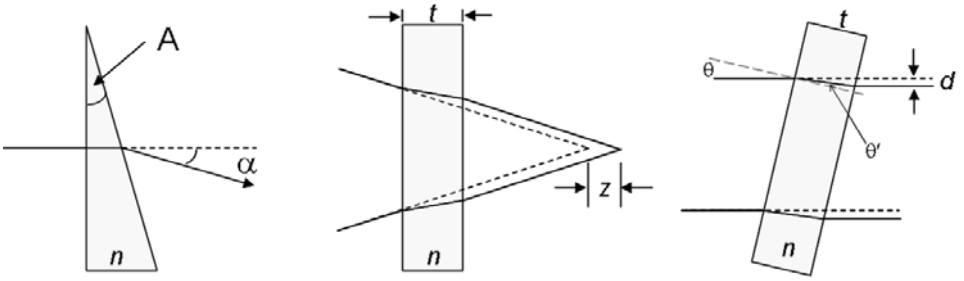


Figure 6.3. The effect of wedges and tilted plane-parallel plates on the optical beam.

entrance windows to vacuum enclosures, dichroic beam-splitters, and polarizing beam-splitters.

Next, identify and list all the known constraints on the design. For example, the wavelength range, the transmittance goals, restrictions or limits on scattered light which are probably driven by the signal-to-noise calculations and the science goals, the desired back focal length, and other optical-mechanical packaging issues (size, weight, thermal mass), polarization effects (due to gratings, or birefringence in crystals or boundary conditions), environmental concerns (thermal changes, shock, and vibration), ability to test and align the optics, and finally the cost of fabrication.

Except in a few cases, it will not be possible to complete the design of the instrument by purely analytic means. The final step is therefore to enter the prescription into a "ray-tracing" program and develop a more sophisticated model. Many excellent programs are available. One of the older packages, and an industry standard, is called Code V ("code five") from Optical Research Associates (ORA). Other very popular packages include OSLO from Sinclair Optics and ZEMAX by Zemax Development Corp.; the latter was the first specifically written for a Windows user interface. Illumination packages include ASAP from Breault Research Organization (BRO), LightTools from ORA, and ZEMAX (engineering edition). Beware that a ray-tracing program cannot design a system for you, it can only trace what you enter, so first-order analysis is very important, and it often helps to begin with an existing design and modify it. Most ray-tracing programs will provide an algorithm which attempts to optimize a given design or search for different designs within constraints which you can control. In this way you can "explore" some options, but be prepared to use up a lot of computing time. A ray-tracing program can assist the designer in studying what the effect of these variations might be, and what compensation techniques (such as refocus) can be applied. It is important to understand the limitations of a given optical design, in order to assess the impact on astronomical goals, as well as the impact on cost and manufacturability.

6.3.2 Aberrations

We have already alluded to optical imperfections in earlier chapters. A perfect optical system would obey the paraxial equations irrespective of the value of θ in Snell's Law.

Imperfect images caused by geometric factors are called aberrations. To see the impact of larger values of θ we can expand the sine functions in Snell's Law in a Taylor series.

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \frac{\theta^9}{9!} - \cdots \tag{6.2}$$

Retaining only the terms in theta to the first power (i.e., $\sin\theta$ is replaced by θ), we arrive at the familiar "first-order" or "paraxial" equations shown previously (Chapter 3). However, by including third-order terms (i.e., replacing $\sin\theta$ with $\theta - \theta^3/6$) we are led to a useful set of equations for describing lens aberrations as departures from paraxial theory; these equations are called Seidel or "third-order" aberrations. For monochromatic light, German mathematician and astronomer Philipp von Seidel (1821–1896) classified aberrations as spherical aberration, coma, and astigmatism which all affect image quality, and distortion and field curvature which affect the image position. In multi-color (polychromatic) light there is also chromatic aberration and lateral color. For completeness and for their value in recognizing potential problems with astronomical instruments, each of these well-known effects is summarized here.

Spherical aberration is caused by the fact that a spherical surface, whether on a lens or a mirror, is geometrically speaking the wrong shape to ensure that all light rays converge to a focus at the same point. Spherical aberration is an "axial" aberration because rays at greater and greater radii from the center of a positive lens (the marginal rays) focus closer and closer to the lens; this is positive spherical aberration. Negative lenses have negative spherical aberration. The difference between the marginal and the paraxial focal points along the axis is called the longitudinal spherical aberration, while the difference between the paraxial focus and the marginal ray intercept at the paraxial focal plane is called the lateral or transverse spherical aberration. A focal plane or detector placed on the axis will see a large blurry image instead of a point source. (This effect occurred on the Hubble Space Telescope because the primary mirror was over-polished toward the edge by a mere 2 μm from its designed hyperboloid.) The circular image obtained has a minimum size called the "circle of least confusion" which is located slightly closer to the lens (for a positive lens) than the paraxial focus and roughly half-way between the paraxial and marginal focal points. For three simple cases the angular diameter β (in radians) of the blur circle is given by:

$$\left.\begin{aligned} \text{spherical mirror:} \quad & \beta = 1/128F^3 \\ \text{parabolic mirror:} \quad & \beta = 0 \\ \text{simple lens:} \quad & \beta = n(4n-1)/128(n+2)(n-1)^2F^3 \end{aligned}\right\} \tag{6.3}$$

where F is the focal ratio; and n is the index of refraction.

Spherical aberration in a lens can be minimized by varying the shape or "bending" of the pair of surfaces, because many different combinations of curvature

produce the same focal length. The shape factor of a lens is defined in terms of its radii of curvature (R_1, R_2) by

$$q = (R_2 + R_1)/(R_2 - R_1) \tag{6.4}$$

and it can be shown that,

$$q = -2(n^2 - 1)p/(n + 2) \tag{6.5}$$

gives the shape factor to produce minimum spherical aberration where p is the position factor $(s' - s)/(s' + s)$ or in terms of focal length $f, p = (2f/s) - 1$; if the lens is used in parallel light then $p = -1$. Alternatively, the lens power $(P = 1/f)$ can be "split" between two or more "slower" lenses (larger f/number). Since the angular diameter of the blur circle is inversely proportional to the cube of the focal ratio (Equation (6.3)), then splitting an $f/2$ lens into a pair of $f/4$ lenses reduces the spherical aberration of each by a factor of 8, and the combination has about 0.5 of the original spherical aberration. A doublet with a positive and a negative element can further neutralize spherical aberration because, varying as the *cube* of focal length, the spherical aberration changes sign with the sign of the focal length. If the curvature of the lens or mirror surface departed from that of a sphere in such a way as to compensate for the difference between sin θ and θ, then spherical aberration would be eliminated! Both marginal and paraxial rays would focus at the same point (for an on-axis object placed at infinity). A parabolic (conic) surface achieves this ideal, and, although more expensive, non-spherical shapes can now be polished into lens surfaces.

Coma is an off-axis aberration. The ray which passes through the center of the entrance pupil from any field point is called the "chief" or "principal" ray. This ray defines the image height (see Figure 6.4). Now consider parallel rays from the off-axis point passing through concentric points on the lens which lie on the circumference of a circle. This annular zone on the lens produces a different magnification from the chief ray, which results in a displaced image in the form of a ring called a "comatic circle" with a diameter and offset that increases with the distance of the transmitting zone from the optical axis. The final result is an overall image shaped like a comet with a conical flare. Tangential coma (the tail) is about three times the radius of the comatic circle which defines the sagittal coma (C_s); the triangular head of size C_s has ~55% of the light. Third-order theory predicts that C_s is zero when $q = -Gp/W$, where $G = 3(2n + 1)/4n$ and $W = 3(n + 1)/4n(n - 1)$, and q and p have the same meaning as before. A doublet designed to correct for spherical aberration can at the same time be corrected for coma. Coma scales linearly with off-axis angle (θ) and inversely as the square of the f/number.

The importance of stops (apertures) is illustrated by the expression for the angular size of the tangential coma for a single spherical mirror:

$$\text{spherical mirror:} \qquad \beta = (3/16)(\theta/F^2)[1 - z/2f\,] \tag{6.6}$$

The angular size of the coma flare is in the same units as the off-axis field angle θ; F is the focal ratio; $f = R/2$ is the paraxial focal length of the mirror; and z is the distance

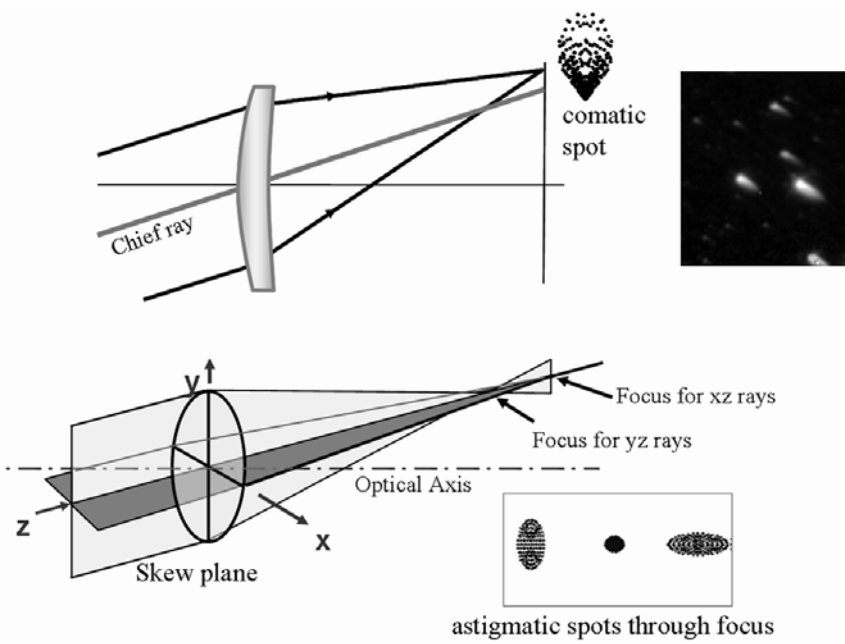


Figure 6.4. Two figures illustrating coma and astigmatism in lenses. The inset shows star images affected by coma. Based on Fischer and Tadic (2000).

from the vertex to the point where the entrance pupil crosses the axis. Coma is zero when the stop is set at a distance $z = 2f$ from the vertex (i.e., at the center of curvature). This is the principle of the Schmidt telescope. Because a spherical mirror has no particular axis of symmetry, a stop placed at the center of curvature ensures that every parallel ray bundle reaching the mirror is symmetrical about its chief ray, and coma is zero.

For a system to be free of coma *and* spherical aberration, it must satisfy a relation known as the "Sine Condition" discovered by Ernst Abbe (1840–1905) and also known as the Abbe condition:

$$\sin\theta/\sin\theta' = \theta_p/\theta'_p = \text{constant} \tag{6.7}$$

Here θ and θ' are the slope angles of the rays in object and image space. As an aside, Abbe was a university professor who became research director of the Carl Zeiss optical works in Jena, Germany in 1866, becoming a partner in 1876 and also director of the Jena Observatory in 1878. He is credited with introducing modern work practices such as the 8-hour work day, holiday pay, sick pay, and pensions.

Astigmatism is also an off-axis aberration and results from a cylindrical-shaped departure of the wave from its ideal spherical shape. All off-axis points lying in a plane which is skew with respect to the plane containing the optical axis form a focus at a point, which is displaced from the focus produced by rays lying along a line in the

plane perpendicular to the first. A focal "line" is seen at one position, and a similar line rotated through an angle of 90° is seen at the other point. A blurred circular image is seen in between. Astigmatism scales as the square of the off-axis angle and inversely as the f/number; $\beta = \theta^2/2F$ for lens or mirror. A tilted plate such as a dichroic beam-splitter or filter placed in a converging beam will produce astigmatism. Astigmatism is also illustrated in Figure 6.4.

Field curvature describes the natural tendency for optical systems to image better on curved surfaces than on flat planes. In the absence of astigmatism, the image is formed on a curved surface known as the "Petzval" surface. Positive elements usually have inward-curving Petzval surfaces (i.e., the surface bends back towards the lens) and negative elements have outward-curving fields, thus some measure of correction can be accomplished by combining positive and negative elements. In Seidel theory the field curvature is spherical of radius R, where $1/R = \sum(1/n_i f_i) = \sum P_z$, the sum being over all elements of the system. Only the refractive indexes (n) and focal lengths (f) are involved in the "Petzval sum", so that a field-flattening lens may be placed at the image point, where it makes no contribution to the power of the system, and its focal length is chosen to make the Petzval sum zero. Field curvature scales like astigmatism.

Distortion occurs when the image of an off-axis point, even though it is sharply defined, does not form at the position predicted by paraxial theory. The image may be well-corrected for coma and astigmatism, but is simply not at the correct location. Typical distortion patterns are "pincushion", "barrel", and "keystone" in which the displacements are toward the center, away from center, and a different stretch from top to bottom of the image. Distortion does not lower the resolution of the imaging system and could be removed by computer processing if required; it is usually described as a percentage shift from the paraxial position.

Chromatic aberration is of course due to the fact that the refractive index of a material is a function of wavelength. Longitudinal chromatic aberration is the difference in focus as a function of color, and lateral color is the difference in image height (or magnification) with color. The blur circle diameter is inversely proportional to f/number and the Abbe V-factor or its equivalent ($V_{\text{eq}} = n - 1/\Delta n$), which is a measure of the dispersion of the material. Color-corrected lenses can be designed by using two or more components, one positive and one negative, but materials with two different dispersion characteristics are also needed. Of course, there is no chromatic aberration with mirror systems.

6.3.3 Wavefront errors

In any optical system there will be a real distorted wavefront traveling through the optics. At each location one can consider the difference between the real wavefront and the best fitting spherical surface. This difference is called the "optical path difference" or OPD. The OPD can be given in microns or in "waves" as a fraction

of a reference wavelength such as 632.8 nm for a HeNe laser. The deviation of the wavefront from perfect, or the wavefront error (WFE), can be specified in terms of the "peak-to-valley" (P-V) optical path difference, which is the difference between the longest and shortest paths. Visualization of the OPD is observed with an interferogram image. If the OPD $= \lambda/4$ peak to valley, over the beam, then the system meets the Rayleigh criterion for diffraction-limited systems and is almost perfect. Alternatively, one can form the "root-mean-square" (rms) OPD summed over the entire wavefront, with the caveat that the rms wavefront error clearly depends on the distribution of the deviations or bumps over the wavefront. Two wavefronts could have the same P-V wavefront error—but very different rms values—depending on the number of bumps and dips over the wavefront; as a rule of thumb, (P-V) OPD is about 3–5 times the rms.

If σ is the rms amplitude of the surface roughness, then the surface is "smooth" if $4\pi\sigma \ll \lambda$, and in this case the fraction of light within any angular radius θ compared with the light enclosed in the diffraction pattern of a perfect surface is the *Strehl ratio* $S = \exp[-(4\pi\sigma/\lambda)^2]$. For example, if $\sigma = \lambda/20$, then $S = 0.67$, and 33% of the light is scattered out of the diffraction spot. This is the total integrated scatter or TIS. Equally important is the area of the optical component over which it is important to maintain the desired OPD. This is called the "footprint" of the beam and is generally the same size as the optical component in the vicinity of a pupil image, but may significantly "underfill" a component near a focal plane. It is generally assumed that all the wavefront errors add in quadrature (i.e., square, add, and take the square root). A tabulation of the contributions to the wavefront error for each optical component derived from the ray-tracing results, together with an allowance or "tolerance" for manufacturing, for changes due to temperature and for alignment errors, is called a wavefront error budget.

Ultimately, performance of the optical design will have to be quantified. This can be done using encircled energy, or ensquared energy as pixels are usually square (e.g., 80% within 2×2 pixels), the modulation transfer function (MTF) across the field, a distortion map, limits on scattering, and a wavefront error (WFE) budget. Note that the linear blur diameter due to diffraction by a circular aperture is $2.44\lambda F$ (where F is the focal ratio or f/number) and a diffraction-limited system would have 84% of the energy within this diameter; 50% of the energy falls within an angular diameter of λ/D. Ray-tracing programs provide several tools for analyzing and displaying the performance of an optical system:

(1) Spot diagrams are produced by tracing a large number of random rays through the system to produce a cluster of impact points on the focal plane to give a visual perception of the primary aberrations, such as coma or astigmatism. The rms spot sizes are geometric only, and do not include diffraction effects.
(2) Encircled energy plots give the total amount of energy within a circle of a given radius as a function of that radius, including diffraction effects.
(3) Tangential ray fans show the variation of an aberration (i.e., where a ray lands in the focal plane) as a function of the axial height of the ray at the optical component (usually normalized to the diameter of the component). For example,

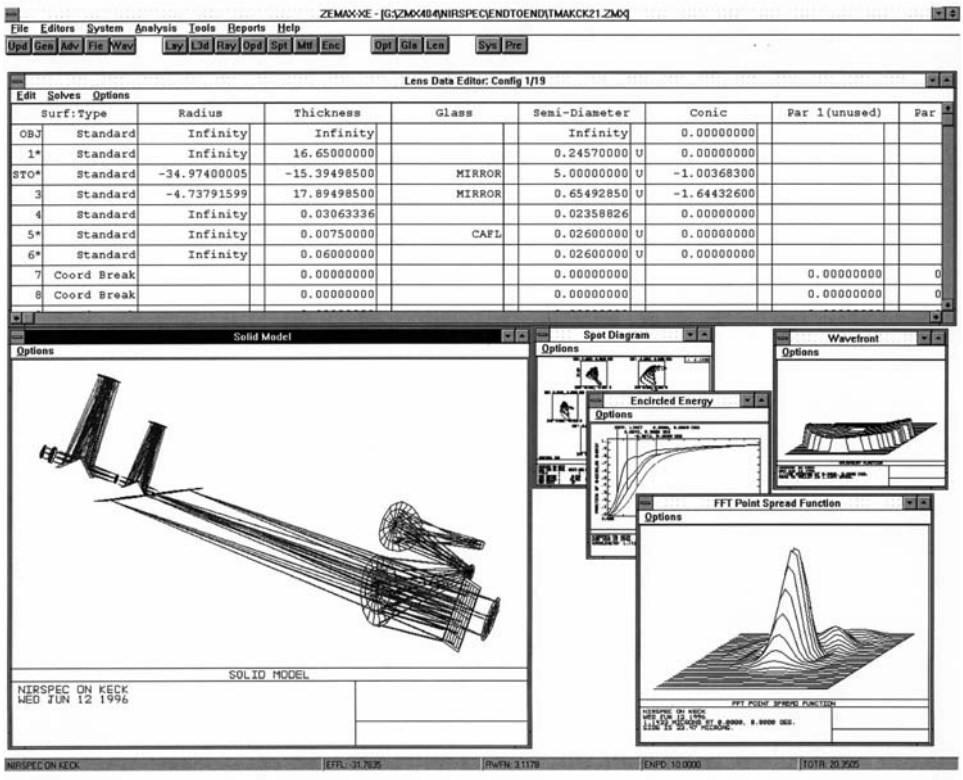


Figure 6.5. The screen of a ray-tracing program (ZEMAX) illustrating the input data, the layout, and several diagnostics such as spot diagrams and encircled energy plots. Credit: Don Figer.

the x-axis would correspond to $\pm d/2$, where d is the lens diameter and the y-axis would give the height in the image plane corresponding to a given x-value or height at the optic. The paraxial focus is at $(0, 0)$. Spherical aberration gives a characteristic ''s-shape'' lying on its side. Coma shows a typical ''u-shape''.

(4) The modulation transfer function (MTF) is essentially the modulation in the image as a fraction of the modulation in the original object, where by modulation we mean $[I(\max) - I(\min)]/[I(\max) + I(\min)]$. This ''contrast'' or visibility ratio is large for low spatial frequencies (many pixels across the bright and dark regions of a test pattern) and decreases quickly as the spatial frequency increases. Spatial frequency is given in cycles per millimeter or in line pairs per millimeter, with the highest possible frequency being determined by diffraction ($= 1/\lambda F$). Many other options are available. An example output from a ray-tracing program is shown in Figure 6.5. This is a trace of the NIRSPEC spectrometer for the Keck telescope.

In principle, aberrations can be controlled by making the surface(s) of the primary optical components depart from their normal spherical (or conic) shapes

in such a way that they are actually compensating for the misbehavior or aberrations which the normal spherical surface would cause. The "sag" or z-coordinate (with z along the optical axis) of the standard axially symmetric surface is

$$z = \frac{cr^2}{1 + \sqrt{1 - (1 + K)c^2r^2}} + \alpha_1 r^2 + \alpha_2 r^4 + \alpha_3 r^6 + \cdots \tag{6.8}$$

where $c(= 1/R)$ is the curvature; r is the radial distance from the axis; and K is the conic constant. For spheres $K = 0$, and for a parabola $K = -1$, hyperbolas have $K < -1$, and K lies between -1 and 0 for ellipses; $K > 0$ corresponds to an oblate ellipsoid. The higher order terms are the "aspheric" components which appear in systems like corrector plates. A different way of describing the higher order terms which allows departures from rotational symmetry is to use the sum

$$\sum_{i=0}^{N} A_i Z_i(\rho, \varphi) \tag{6.9}$$

where the Z_i are called Zernike polynomials in the normalized radial coordinate ρ (rho) and the angular coordinate φ (phi). This is the terminology used to study a system of multiple mirrors like the Keck telescope, or the deformable mirror in an AO system (see Chapter 2).

Manufacture of aspherical optics is only practical with the aid of modern, precision computer-controlled milling and turning machines using diamond-tipped cutters. The final surface depends greatly on the choice of material and the nature of the asphere. Aspheric surfaces directly machined using diamond-tipped cutters are a particularly attractive technological innovation for infrared optical components which must be cooled to low temperatures. For example, it is possible to make an aspheric mirror by direct diamond machining of a solid block of aluminum which has already been machined by normal lathes to provide its own mirror support and attachments! The differential thermal contraction between metal supports and glass components is eliminated because mirror and support are made of the same material.

There are several useful optical designs that have been adapted and used many times in astronomical cameras and spectrometers, including the off-axis parabola, the Offner Relay and the Three-Mirror Anastigmat (TMA). The off-axis parabola (OAP) is a section of a paraboloid that does not contain the vertex. Light rays parallel to the original optical axis through the missing vertex are brought to a focus by the OAP at the original focal point of the paraboloid. Conversely, rays diverging from that focal point become collimated after reflection from the off-axis section. This arrangement is used in cameras and spectrometers to produce a collimated beam with the least number of optical surfaces, just one in this case! On-axis images are almost perfect. Off-axis images are limited by coma. A second off-axis parabola taken from the diametrically opposite section of the paraboloid will have the opposite sign for coma. Thus, two identical, opposing off-axis parabolas can create a 1 : 1 relay with no coma and no spherical aberration. This is sometimes called a Mersenne relay and is the basis for many adaptive optics systems. If the OAPs have different focal lengths then the system can provide either magnification or reduction, but it will not be perfectly

corrected for coma; such a scheme is used in the NIRSPEC instrument shown in Figure 6.5. A common three-mirror relay is the Offner (1975) which consists of a single large reflector and a small secondary reflector. Only a part of the large reflector is used. The input beam is collimated by part of the large mirror to produce a pupil on the small mirror which reflects light back to the other section of the large mirror to create a converging beam. The mirrors are usually spherical. A more sophisticated system is the Three-Mirror Anastigmat (TMA) by Korsch (1972) in which all three surfaces are conic sections. A TMA can be used as a fast, all-reflecting camera or a collimator, or a telescope. Performance is excellent although distortion (keystone) is usually considerable. A TMA is the last optical element in the design shown in Figure 6.5.

Lens designs are harder to group, but most camera and collimator systems must make use of achromatic doublets for the wavelengths of interest. To reduce spherical aberration it may be necessary to split the power and double the number of components. The Cooke triplet composed of two positive lenses with a negative lens in between can produce a well-corrected camera or collimator. Harland Epps of the University of California, Santa Cruz has been involved in a remarkable number of astronomical optical designs and recalls many of them in a short review (Epps, 2002).

6.3.4 Coatings and interference filters

Of considerable importance in commercial refractive optics is the *anti-reflection coating*. Since most high-quality lenses are actually made from multiple simpler components, it becomes crucial to eliminate reflections between these components. The reflectance (R) of a surface is a function of the index of refraction (n) on both sides and the angle of incidence (θ). At normal incidence ($\theta = 0^\circ$) in air ($n_1 = 1$) we have

$$R = \left(\frac{n-1}{n+1}\right)^2 \tag{6.10}$$

For glass with $n = 1.5$ the reflectance is 0.04 or 4%. Coating the lens with a very thin layer of material which has a refractive index intermediate between that of air and the original lens material can reduce the reflection losses by destructive interference at both boundaries. The basic relationship is given by

$$2nd = (m + \tfrac{1}{2})\lambda \tag{6.11}$$

where m is an integer; and d is the thickness of the film. Thus, the minimum thickness for an anti-reflection coating occurs when $m = 0$ and $d = \lambda/4n$. Note that since λ is the wavelength in air and λ/n is the wavelength inside the thin film (λ_n), then $d = \lambda_n/4$ which is why such simple coatings are called a "quarter-wave" layer. The full solution of this interference problem using Maxwell's equations takes into account all of the other multiple reflections inside the thin film and shows that in addition to achieving quarter-wave thickness, the ideal value for n is $\sqrt{(n_g)}$, where n_g is the index of refraction of the substrate glass (and the surrounding medium is assumed to be air). A widely used material which almost meets this requirement

for optical glass is magnesium fluoride (MgF_2) with $n = 1.38$. Clearly, the simple coating described above can only function perfectly at one wavelength, whereas modern astronomical instruments are designed to perform over a wide range of wavelengths (to take advantage of the response of a silicon CCD, for instance). It is possible to develop "multi-layer coatings" using several materials which provide a reasonable average reduction in reflectance over a wider wavelength range. Such coatings are called "broad-band" anti-reflection coatings. Significant improvements in the production of durable broad-band AR coatings have been very beneficial to astronomers, and such coatings have even been deposited on the surfaces of CCDs and infrared arrays; the large refractive index ($n = 4$) of silicon results in considerable reflection losses (about 36%), and therefore sets an upper limit of 64% on the quantum efficiency *unless* an AR coating is used.

As mentioned in Chapter 5, similar principles are employed to develop multi-layer interference filters for narrow-band work, such as imaging in nebular lines. These filters can be thought of as a "solid" Fabry–Perot etalon ($m\lambda = 2n_e d \cos\theta$), where n_e is an effective refractive index. Note that as θ increases from 0° the central wavelength is "scanned" to shorter wavelengths, so tilting these filters shifts the central wavelength.

6.4 MECHANICAL DESIGN

Converting the optical design into a mechanical layout which ensures that all the optical components are properly mounted, aligned, and stable is often much harder than it seems, and can therefore lead to poor images. Several issues should be considered:

- choice of lens and mirror-mounting schemes
- mechanical and thermal stress on the optics
- alignment of the optics
- flexure and stability under gravity if the instrument moves with the telescope
- method of attaching to the telescope
- stray light baffles and light-tight enclosures
- moving parts such as focus drives, filter wheel, and shutter mechanisms
- ease of handling, assembly, and disassembly
- integration of electrical wiring
- cooling systems, thermal paths, and thermal mass

Because astronomical instruments exhibit an enormous range in size, weight, and complexity it is hard to generalize. Typically, the first step is to draw a rough layout around the optical components in order to assign real thicknesses and dimensions to the support structure (lens barrels, mirror mounts, filter wheels, etc.) and look for "first-order" problems such as collisions between components and vignetting. Estimate the location of the center of mass of the whole instrument. Look for poor design arrangements such as a heavy component supported by a horizontal rod which

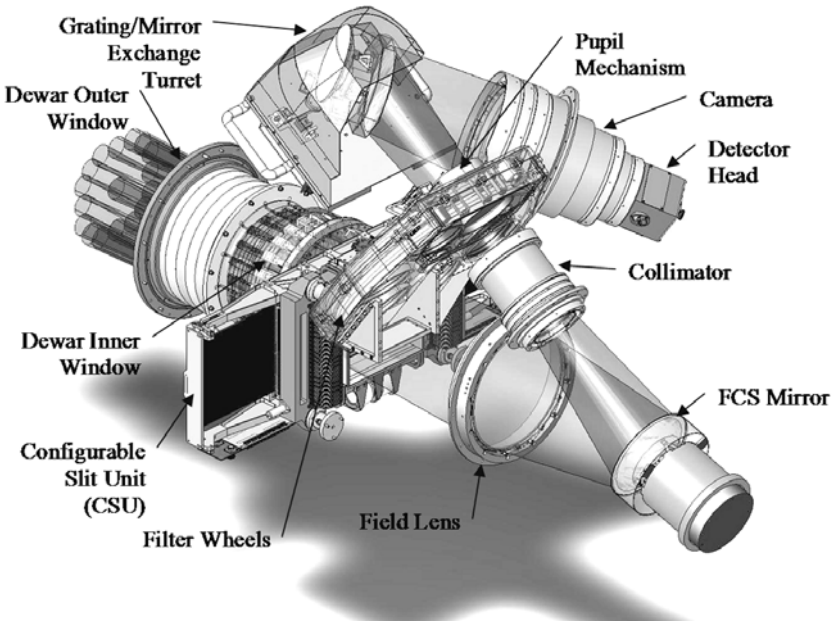


Figure 6.6. An engineering drawing prepared using a computer-aided drawing package (Solidworks). The instrument is a multi-object spectrometer called MOSFIRE. Credit: Bob Weber.

is either too thin or lacks a triangular support strut. Try to understand what might bend (flex) as the telescope carries the instrument around. Three-point (ball/groove/flat) "kinematic" mounting schemes may be required to minimize stress and to enable units to be removed and put back in exactly the same place. After this first look, a detailed mechanical design can be developed based on a more careful assessment of the dimensions and properties of walls, struts, wheels, bearings, and other support components required to meet specifications on strength, flexure, load bearing, heat flow, and so on.

A major asset in mechanical design is the availability of excellent computer-aided drawing tools such as AutoCad, Solidworks, and others. These powerful 3-D drawing programs run on many platforms, including PCs and workstations, and allow one to import files from ray-tracing packages such as ZEMAX, and then build the mechanical design around the optical rays and surfaces. Figure 6.6 shows this technique applied to a complex instrument called MOSFIRE for the Keck I telescope.

6.4.1 Materials and properties

Most instruments are constructed from aluminum (Al) alloys to save weight, although handling frames and large rotating bearing structures are always made from steel. Some use is made of copper, brass, and stainless steel, fiberglass materials

like G-10, and occasionally some plastics like Delrin. In general, there will be a total weight limit and a "moment" limit (such as 227 kg or 500 lb at 1 m from the mounting flange) for instruments attached to a telescope. It is very important to know if and how the instrument will bend or flex as the telescope points to different parts of the sky. For instruments which are not fully enclosed in a vacuum chamber, a "space frame" structure of rods is often ideal for strength, weight, and rigidity. Box-like structures tongued and grooved for light-tight fitting and triangular gussets and buttresses for additional stiffness are usually adequate for small photometers and cameras. External surfaces are usually anodized, and internal surfaces are black anodized and/or painted with matt black paint.

All the basic properties of the materials to be used must be collected together, including the density, coefficient of thermal expansion, Young's modulus, yield strength, compression strength, shear strength, heat treatment, hardness, specific heat, and emissivity. Many properties of materials are a function of temperature, so the range of applicability needs to be known or the "integrated" effect over the required temperature range should be used. Elementary physics can be used at this stage to estimate the tensile (stretching) and shear (tangential) stresses (force per unit area, F/A) on various rods, struts, and plates in the instrument for comparison with tabulated limits on "yield" strength and "breaking" strength. The yield strength defines the stress beyond which the object will not return to its original shape when the forces are removed. Stress has the same units as pressure and is measured in newtons/m^2 (or pascal, Pa) in the SI system, but it is also commonly quoted in pounds per square inch (psi), especially in the United States. From knowledge of the Young's modulus (E) and the shear modulus (S) of the material one can estimate the strain ($\Delta L/L$) and displacement since

$$F = EA\frac{\Delta L}{L}, \qquad F = SA\frac{\Delta x}{h} \tag{6.12}$$

where A is the area over which the force is acting; L is the original length; h is the separation between the two planes which have sheared by Δx. Advanced reference works on mechanical engineering provide equations for stress and strain on many complicated shapes, but ultimately it may be necessary to use a computer program to perform a Finite Element Analysis (FEA) in which the structure is replaced by numerous much smaller identical mechanical blocks or elements whose loads and interface (or boundary) conditions are specified.

Vibrations are often a concern in sensitive scientific equipment on a telescope. Hooke's Law states that for elastic bodies stress is linearly proportional to strain. For an elastic spring stretched or compressed by a force F to a distance x, Hooke's Law becomes $F = kx$ where k is the spring constant. For a mass m vibrating on a spring with constant k the frequency of the simple harmonic motion is given by $\sqrt{(k/m)}$. Many sub-systems in an instrument can be approximated by a spring system. In order to avoid resonance effects, in which certain vibration frequencies excite large displacements, it is important to ensure that the fundamental frequency is not close to that of natural vibrations due to closed-cycle refrigerators or other vibration-producing systems in the observatory.

Aluminum is commonly used for small instruments. It is one-third as dense as steel and has 60% of the electrical conductivity of copper. Alloys are specified by four digits plus a suffix beginning with T (for heat treatment) or H (for work hardness): T0 is no heat treatment (dead soft) and T10 is fully tempered (hard). For example, 1100-T0 is a very soft, nearly pure aluminum alloy, whereas 6061-T6 is a relatively hard yet easily machined alloy and readily available. Aluminum can be joined by a process called heliarc welding in which an electric arc is used as the heating (fusing) element in an inert (argon) atmosphere, or it can be electron-beam welded. The metal is weakened near the weld, and this should be taken into account when estimating the strength of a joint. Like other alloys, aluminum can be heat-treated in a variety of ways, such as annealed, quenched, or tempered. Annealing is the process of heating a metal to above a transition temperature and then slowly cooling it. Quenching is similar, except the cooling is rapid. Tempering is an intermediate treatment in which previously hardened metals are reheated to below the transition point in order to relieve stresses and then cooled at a rate that preserves the desired property. Parts to be used at cryogenic temperatures in scientific instruments, especially metal optics, must be "thermally cycled" several times to release stresses and eliminate small changes in dimensions over time known as "micro-creep".

Other well-used materials include the following. Stainless steel, which is an alloy of iron, chromium, and nickel, is also used in astronomical instruments. The 300 series have a high nickel content and are all fairly non-magnetic, with Type 304 being the best and most common. Invar is an iron–nickel alloy with a coefficient of thermal expansion that is about ten times smaller than steel. Copper is soft, but provides the best electrical and thermal conductor and is often used as the thermal heat sink or cold block for the chip in cooled CCD cameras. Oxygen-free high conductivity (OFHC) copper is preferred for its purity, excellent conductivity, and resistance to hydrogen embrittlement. Brass, an alloy of copper and zinc, was once very common in scientific instruments, but its use is more limited now. Fiberglass materials like G-10 are often used for thermal isolation in vacuum-cryogenic applications, and the Dupont plastic known as Vespel is very useful because it is impregnated with molybdenum disulphide (MoS_2).

Powerful 3-D modeling packages (as used in Figure 6.6) now make it easy to visualize how parts fit together. However, for an instrument which has never been built before, there is still merit in constructing a "space model" (i.e., a full-size replica made from the preliminary drawings but using light-weight materials such as foamboard, cardboard, thin metal sheeting, tubing, and light woods). Figure 6.7 shows a space model of the NIRSPEC instrument for the Keck telescope. Although time-consuming to do well, the act of building the space model to match the preliminary drawings will almost certainly reveal design flaws and focus attention on areas of difficulty in the real instrument. Potential problem areas are (i) the order of assembly; (ii) optical alignment and verification; (iii) location and installation of baffles; (iv) electrical wiring and connectors; (v) cooling paths; (vi) general handling fixtures. On a much grander scale, a fantastic full-size model of the 6.5 m James Webb Space Telescope has been displayed to astronomers at various meetings.

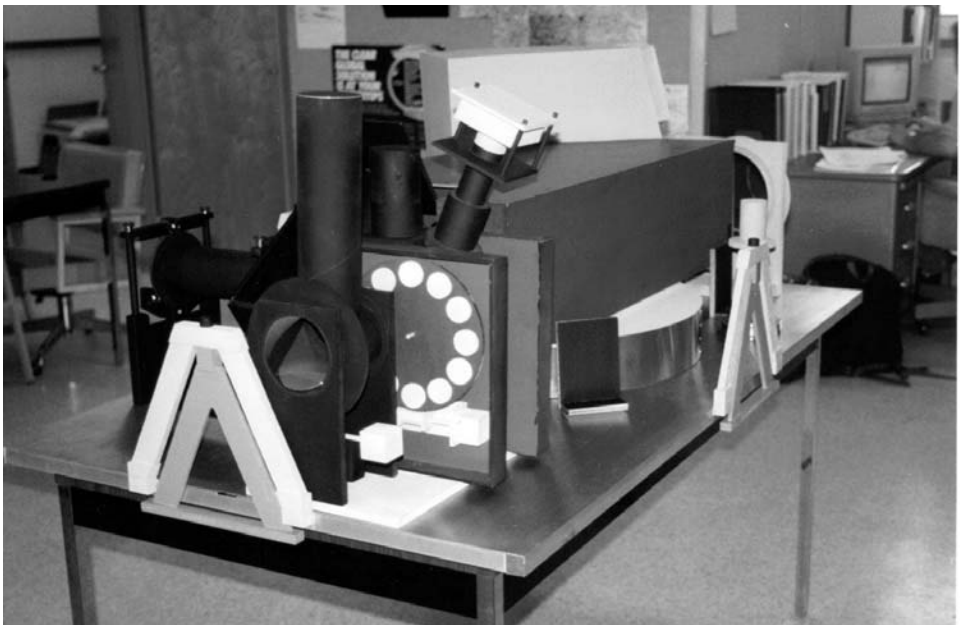


Figure 6.7. A "space" model of an infrared spectrograph constructed at the IR Lab at UCLA.

6.4.2 Mechanisms

A common need in many instruments is a filter wheel. Whenever possible, filters are located in a parallel beam close to the position of the pupil image. Doing so minimizes the size of the filter, reduces the influence of any dust or scratches on the filters, and eliminates any focus shift when filters are changed. Filters placed in divergent beams may have to be larger, and the optical path (*nd*) through each filter will need to be balanced to ensure that there is no change in focus when one filter is changed for another of different thickness and refractive index. If N filters are mounted in a wheel, the minimum diameter of the filter wheel is approximately ND_{pupil}/π, where D_{pupil} is the diameter of the pupil image. There are two common methods of driving a filter wheel: a direct drive shaft through the center of the wheel, or an edge drive using gear teeth around the perimeter. Edge-driven wheels usually employ a worm gear and achieve a very large mechanical advantage (or gear ratio). Direct drives may require intermediate gearing to yield some mechanical advantage. The rotating shaft can be driven by a DC servo motor or one of the many small stepper motors on the market. A DC servo motor runs continuously until stopped, whereas a stepper motor only moves when commanded to do so by receiving a pulse. Stepper motors are attractive because the mechanism can be operated "open loop" and the position of the wheel can be determined simply by counting pulses. Mechanical drive trains tend to suffer "backlash", which means that the amount of motion required is not the same in both directions of rotation. To counteract this effect the mechanism must be moved in one direction only or be mechanically spring-loaded (pre-loaded).

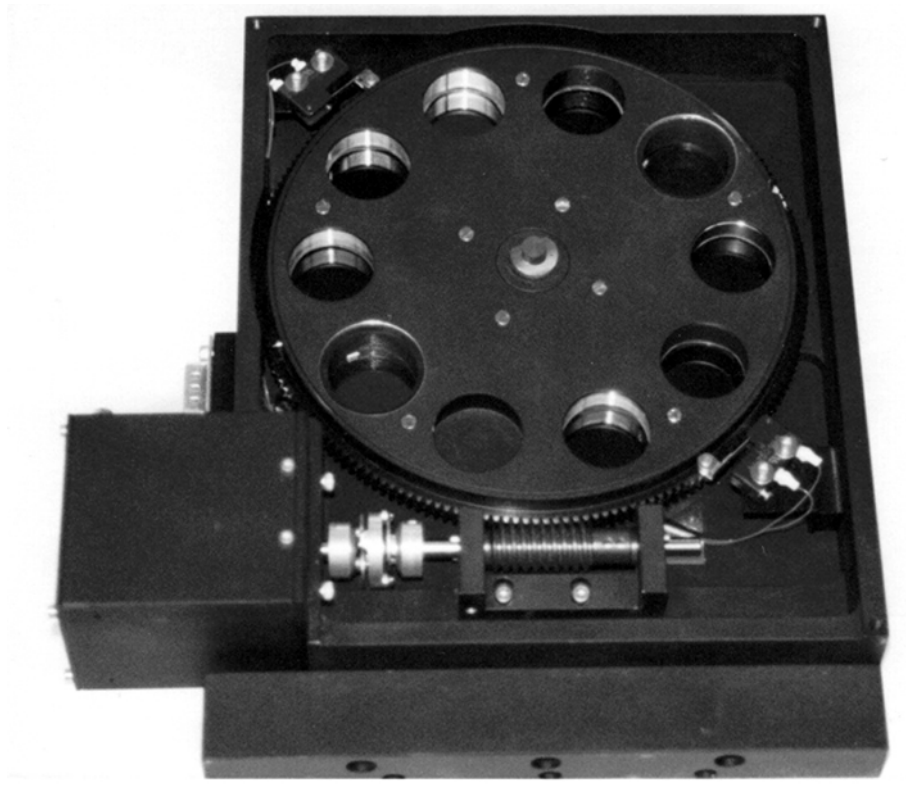


Figure 6.8. A typical worm-driven mechanism, in this case a simple filter wheel, is shown. A cryogenic motor is used to drive the wheel.

Figure 6.8 shows a typical, small worm-driven wheel. The stepper motor shaft is attached to the worm shaft by a "flexible" wafer coupling which compensates for lack of precise alignment and any differential thermal contraction. Since the pitch of the worm threads is small, the worm cannot be rotated by turning the wheel, and therefore it is safe to remove the holding torque by switching off the power to the motor. Typical stepper motors have 200 full steps per turn (1.8° per step) but are usually driven in "half-step" mode which gives 400 steps per turn (0.9° per step). If this motor was attached directly to the hub of the filter wheel then one step would displace the center of the filter with respect to the optical axis by $\sim 0.016R$, where R is the radius measured to the center of the filter. For example, if $R = 50$ mm then the displacement is 0.8 mm or about 3% of the diameter of a 1-inch filter. When the wheel is worm-driven there is a huge gear ratio and resolution advantage. If there are 180 teeth around the circumference then it takes 72,000 steps (in half-step mode) to execute a complete turn, and one step now corresponds to a displacement of $\sim 87\,\mu$rad per millimeter, or about 4 microns in the above example. All parts of the filter wheel (the rotor, coupler, wormshaft, worm, and the wheel) can be considered as disks with moment of inertia $I = \frac{1}{2}MR^2$, where M is the mass and R is the appropriate radius,

except that the rotational inertia of the wheel is reduced by the square of the gear ratio. If the angular acceleration is α radian/s^2 then the corresponding torque is $\tau = I\alpha$.

Positional data can be obtained in one of two ways. In a "closed-loop" approach, the filter position is sensed using a continuous absolute position encoder or at least a discrete encoding scheme (such as a micro-switch) at each filter position. Various forms of absolute encoders are commercially available. The encoder is attached to the drive shaft or edge of the gear and gives a direct reading of its position. An "open-loop" system has a single "datum" or reference position and relies on counting steps to know where the mechanism is and which filter is installed. The latter method is simpler to install and works surprisingly well. It is often used in infrared instruments because of the difficulty of getting an absolute encoder for cryogenic applications. If this approach is adopted then it is wise to include at least two "datum" or "home" positions. Miniature lever-type switches, Hall effect switches, and light-emitting diode (LED) switches are common in many instruments.

Infrared instruments pose special problems because everything is cooled to cryogenic temperatures—not just the detector. Plastic mounted switches contract differently from their metal attachments, stainless steel ball and roller bearings must be degreased, and a dry lubricant such as molybdenum disulphide (MoS_2) burnished on. Degreasing can be done by popping the seal and placing the bearing in a beaker of research-grade ethyl alcohol for 5 minutes to loosen up the grease and then transferring the beaker to an ultrasound bath for 30 minutes to dissolve the grease. Lastly, the bearings should be individually rinsed in alcohol, dried with a stream of dry nitrogen gas, and stored in sealed plastic bags; cleanliness is essential.

Finally, bear in mind physical handling of the instrument and its components. Think about safety issues and wherever possible build in handles and feet. Decide whether or not a lifting mechanism is required and think about storage containers.

6.5 CRYOGENICS AND VACUUM METHODS

Most state-of-the-art detectors require cooling for optimum performance. There are several categories of cooling systems which might be required in cameras and spectrographs for ground-based astronomy depending on the application.

1. Thermoelectric coolers and liquid circulation coolers: these normally operate over the range $-20°$C to $-50°$C and are suitable for photomultiplier tubes, certain CCDs which have low dark currents, and high-speed applications such as telescope-guiding cameras. This class of cooler is very popular in CCD cameras used on small telescopes because it is much easier and simpler to use than the options below.

2. Liquid and solid cryogens

— Dry ice (solid CO_2): cheap and readily available. Coming in the form of a "snow" it is most often used as the coolant for GaAs PMTs. Temperatures around $-78°$C (or 195 K) can be achieved with dry ice.

— Liquid nitrogen (LN_2): relatively cheap. Cools detectors (and other components if required) to -196°C (or 77 K), which is the normal boiling point of liquid nitrogen. Almost all professional CCD cameras employ LN_2 cooling systems, and certain near-infrared devices such as HgCdTe arrays are also cooled with LN_2.
— Liquid helium (LHe): more expensive, but will cool detectors to -269°C or 4 K and is required for low-bandgap semiconductor materials and bolometers used in infrared instruments. Liquid helium requires a special double-walled evacuated "transfer" tube and some experience in handling the transfer to recognize when liquid is being collected and transfer is complete. The LHe reservoir is usually pre-cooled with LN_2 which is then blown out with LN_2 gas and the LHe transfer begun immediately.

  For liquid cryogens the cooling ability is expressed in terms of the product of the density (ρ) and the latent heat of vaporization (L_V):

$$(\rho L_V)_{\mathrm{LHe}} = 0.74 \text{ W h L}^{-1}$$

$$(\rho L_V)_{\mathrm{LN_2}} = 44.7 \text{ W h L}^{-1}$$

  For example, a 10 W heat load boils away 1 liter (l) of LHe in 0.07 hours whereas 1 liter of LN_2 lasts 4.5 hours. By attaching a vacuum pump to the LN_2 vent and reducing the pressure above the liquid it is possible to solidify the nitrogen and achieve $\sim$65 K.
— Liquid neon (LNe) and liquid hydrogen (LH): these provide an intermediate solution between LHe and LN_2, but are rarely used in astronomy applications.

3. Electrical heat engines or closed-cycle refrigerators: because LHe is fairly expensive compared with liquid nitrogen, and requires special care in handling, many infrared instruments and sub-millimeter/radio receivers employ multi-stage closed-cycle refrigerators (CCRs). Typical two-stage Gifford–McMahon systems using 99.999% pure helium gas at about 300 psi as the working fluid, such as the 350 model from CTI Inc. (now part of Brooks Automation Inc., U.S.A.), can provide two cold levels, usually 65 K and 10 K and extract heat at a rate of about 20 watts and 7 watts from each stage (see Figure 6.9 which also shows the thermal cycle). Both single-stage and triple-stage versions are available, and the larger units provide about 100 watts of cooling power. If the mass to be cooled is very small (e.g., a single CCD), then much simpler systems such as small Sterling cycle coolers can be used. Vibration damping is needed, especially for CCR-cooled instruments in sensitive AO systems, and counter-balance weights may also be required.

4. ^{3}He systems: even lower temperatures are obtained using ^{3}He systems in sub-millimeter and far-infrared bolometers. The basic principle of a ^{3}He cryostat is to condense helium-3 gas by bringing it in contact with a pumped helium-4 reservoir (yielding $\sim$1.2 K). Low temperatures below 300 mK are then obtained (for small samples) by reducing the vapor pressure on top of the liquid helium-3 using an internal sorption pump (charcoal).

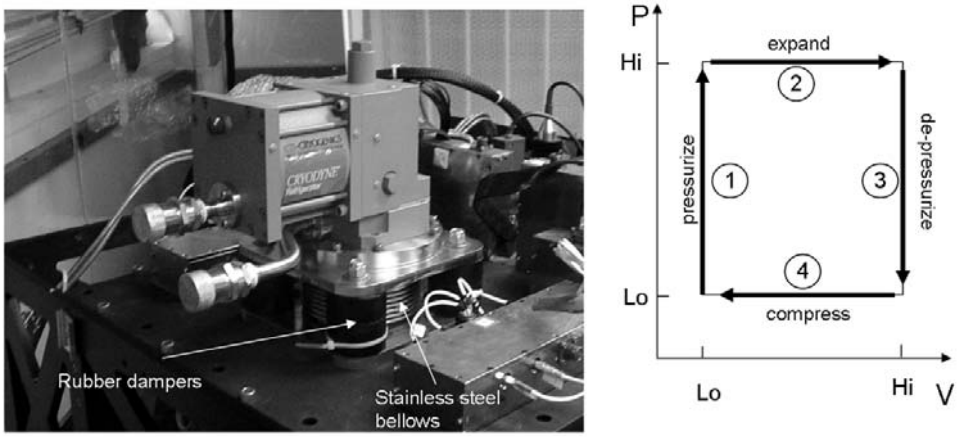


Figure 6.9. The cold head of a two-stage closed-cycle refrigerator (CCR) attached to the UCLA twin-channel camera for Lick Observatory. This unit is the CTI model 350. Also shown is the thermal cycle.

The cooling system can be the dominant issue in the case of a large-infrared instrument in which *everything* must be cooled. Choosing the detector temperature will depend on factors like dark current and noise, whereas the temperature of the surrounding opto-mechanical structure may be determined by its thermal emission. Typically, steps in the thermal analysis of an instrument would be as follows:

- determine what is to be actively cooled (e.g., detector only, whole instrument)
- tabulate the required operating temperatures (e.g., detector, optics)
- calculate the thermal energy removed from the mass to be cooled
- determine the heat loads due to conduction and radiation
- select the appropriate cooling system
- estimate the cool-down time
- estimate the "hold" time (for liquid cryogens).

The heat H (joules) removed from a mass m (kg) which is cooled from a temperature T_h to T_c is given by

$$H = mC(T_h - T_c) \tag{6.13}$$

where C is the specific heat of the material in joule/kg per K ($\mathrm{J\,kg^{-1}\,K^{-1}}$). The specific heat of a substance usually changes with temperature and it depends on the conditions under which the heat is applied (i.e., constant volume C_v or constant pressure C_p), but for solids the difference is generally small. For aluminum, C is about 900 J/kg · K, copper is 385 J/kg · K, steel is about 450 J/kg · K and water is 4,190 J/kg · K. Specific heat is also tabulated in cal/g · K; using the conversion 4.19 joules per calorie gives 1 cal/g · K for water. As an example, to cool a mass of 1 kg (2.2 lb) of copper from 290 K to 80 K requires the removal of $1 \times 385 \times 210 = 80{,}850$ joules and 2.3

times that amount for aluminum. If heat removal is to be accomplished in $t = 1$ hour (3,600 s) then the average power is $H/t = 22.5$ W.

The rate of transfer of heat Q_H (in watts) by conduction along a rod of uniform cross-sectional area (A) and temperature gradient dT/dx is given by

$$Q_H = -kA\frac{dT}{dx} \tag{6.14}$$

where k is called the thermal conductivity (W m^{-1} K^{-1}) and is about 240 for Al, 400 for Cu, and only 0.9 for glass. In steady-state conditions we can write dT/dx as $\Delta T/L$, where L is the length of the conductor. If we think of heat flow as "current" and temperature difference as "voltage" then by analogy with Ohm's Law ($V = IR$) for electrical circuits, we can define a thermal resistance such that

$$\Delta T = Q_H R, \qquad R = \frac{1}{k}\frac{1}{A/L} \tag{6.15}$$

It is often the case that the optimum operating temperature of a detector does not coincide with that of the cooling stage. For example, liquid nitrogen is actually much colder than the typical CCD operating temperature of 135 K–150 K, and liquid helium is too cold for typical near-infrared detectors. What is needed is a temperature intermediate between the cryogenic bath and the warm (ambient temperature) body of the vacuum chamber. To achieve this it is usual to set up a thermal analogy to an electrical potentiometer circuit. One solution is to attach a copper block to the outer walls of the cryostat by means of pillars of fiberglass tube, which is a poor thermal conductor, and also to connect it by a copper wire to the copper face of the liquid-nitrogen reservoir. A simple design is shown in Figure 6.10. If $R_{\rm fg}$ and $R_{\rm Cu}$ are the thermal resistances of the fiberglass and copper braid, respectively, then the

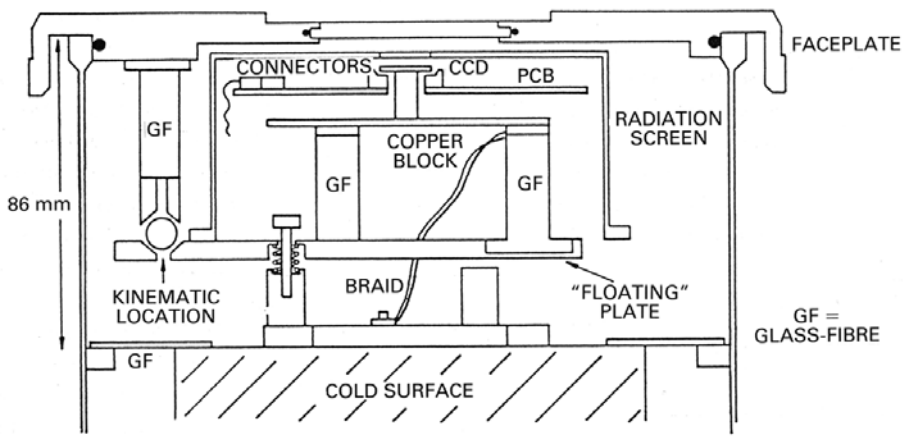


Figure 6.10. A typical mounting scheme for a CCD detector in a liquid-nitrogen cryostat. The arrangement allows the CCD to be operated at a temperature between 77 K and ambient. Credit: Paul Jorden.

intermediate temperature T_{ccd} is given by

$$\frac{T_{\mathrm{ccd}} - T_c}{T_h - T_c} = \frac{R_{\mathrm{Cu}}}{R_{\mathrm{fg}} + R_{\mathrm{Cu}}} \tag{6.16}$$

Alternatively, if the CCD temperature is chosen, then the ratio of R_{fg} to R_{Cu} can be found. The sum of the thermal resistances determines the total heat flow, which one tries to keep reasonably small. By placing limits on the heat flow then both resistances are fixed. For a given thermal conductivity (k), each thermal resistance can be achieved by the appropriate choice of A/L. Trimming the length and cross-section (number of wires) of the braid and the length and cross-section of the fiberglass tube enables a "balance" to be reached between the tendency of the LN_2 to cool down the copper block (and hence the CCD) and the tendency of the block to warm up to room temperature via the fiberglass pillars. In addition, a heavy-duty resistor can be attached to the copper block to enable the block to be heated (by passing a current through the resistor) at any time. A temperature sensor, usually a diode, is also threaded into the copper block, and a circuit controls the current to the heating resistor depending on the voltage measured across the diode which has a constant current flowing through it. This allows some measure of control over the actual operating temperature. Temperature control units which handle these functions are available commercially (e.g., Lakeshore Cryotronics Model 340).

Since k is a function of temperature, it is often more convenient to integrate over the required temperature range and give Q_H in the form

$$Q_H = \frac{A}{L}[I_{T_h} - I_{T_c}] \tag{6.17}$$

where L is the total length of the conductor; and I is a tabulated property for many materials called the thermal conductivity integral which accounts for the variation of thermal conductivity (k) with temperature. In this expression, T_h and T_c represent the hot and cold temperatures between which the heat is flowing. If A and L are measured in square centimeters and centimeters, respectively, then I is in watts/centimeter. Typical values are shown in Table 6.1.

Note the poor conduction of G-10 compared with aluminum. Optical glass is similarly poor, and therefore a large lens or mirror cools much more slowly than the metal supporting it. Although only an approximation, Equation (6.17) is extremely useful and can be used in a number of ways. For example, given A/L and the temperatures (T_h, T_c) we can derive the cooling rate Q_H. Alternatively, given the cooling rate, A/L, and the initial temperature we can derive I_{T_c} and hence T_c.

In large cryogenic infrared instruments, heat transfer by radiation is the dominant mechanism. The power radiated from a body of area A at an absolute temperature T is given by

$$Q_H = \varepsilon \sigma A T^4 \tag{6.18}$$

where $\sigma = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$ is the Stefan–Boltzmann constant; and ε is the emissivity of the surface; $\varepsilon = 1$ for a perfectly black surface and is less than 0.1 for a shiny metallic surface. Polished aluminum foil yields an emissivity of about 0.05 and

Table 6.1. Values of the thermal conductivity integrals in watts/centimeter for four materials.

Temperature (K)	*OFHC Copper* (watts/cm)	*6061 Aluminum* (watts/cm)	*Stainless steel* (watts/cm)	*G-10 Fiberglass* (watts/cm)
300	1,520	613	30.6	1.00
250	1,320	513	23.4	0.78
200	1,120	413	16.6	0.55
150	915	313	10.6	0.37
100	700	211	6.3	0.19
77	586	158	3.2	0.11
50	426	89.5	1.4	0.07
10	25	3.6	0.03	0.005

goldized Kapton gives ~0.04; anodizing increases the emissivity by a factor of 10 or more. The net rate of heat transfer by radiation from a body at temperature T_h onto a body at temperature T_c is given by

$$Q_H = \sigma A_c F_{hc}[T_h^4 - T_c^4] \tag{6.19}$$

Here F_{hc} is an "effective" emissivity which also depends on the geometry of the cryostat (or dewar) configuration, such as concentric cylinders or plane-parallel plates, both of which are quite realistic. In both cases, when the emissivities of the two surfaces are small (<5%) and equal, then $F_{hc} \sim \varepsilon/2$.

For a given temperature differential, radiation load is minimized by reducing the surface area and achieving the lowest emissivity (shiniest) surfaces. It is also possible to add n "floating" shields which reduce the radiated heat load on the innermost body by a factor of $(n + 1)$; but careful application is critical. Various forms of floating shields are available. One form is called Multiple Layer Insulation (MLI). Typical emissivities range from 0.03 for polished aluminum and gold foil to 0.32 for polished anodized aluminum to 0.95 for a matt black surface. A useful rule of thumb is that if $\varepsilon = 0.05$ and the hot and cold temperatures are 290 K and 80 K, respectively, then the radiation load is ~10 W/m^2–11 W/m^2. Note that radiation load is very sensitive to T_h.

The cooling time for a mass m from initial temperature T_i to T_{i+1} ($T_i > T_{i+1}$) is found by dividing the heat removed by the effective rate of cooling,

$$\Delta t_{T_i, T_{i+1}} = \frac{mC_p(T_i)[T_i - T_{i+1}]}{(A/L)k(T_i)[T_i - T_{\text{sink}}] - Q(T_i)} \tag{6.20}$$

where T_{sink} is the temperature of the cold source (LN_2 or CCR stage), and the rate at which heat is removed by conduction to this reservoir is governed by the A/L and

conductivity of the connecting path. The term $Q(T_i)$ is the rate at which heat comes back into the mass at temperature T_i (e.g., via radiation from the walls and conduction along shafts and other mounting structures). The total cooling time to an equilibrium temperature T_{eq}, given by $(A/L)k[T_{eq} - T_{sink}] = Q(T_{eq})$, is just the summation of Equation (6.20) over all the individual cooling intervals. Cool-down is a non-linear function of time. At first the cooling is very rapid and then it slows down as the heat coming into the system (by conduction and radiation) begins to balance the heat removed.

Dewars or cryostats come in various forms but are usually either single reservoirs for LN_2 or double reservoirs for LN_2 and LHe. Access to the interior is obtained by removing a plate which maintains a vacuum seal by using a rubber o-ring. The o-ring sits in a shallow groove but projects above the level of the surface to be sealed and is compressed by the opposing surface. Among the other factors in cryostat design are mechanical flexure (bending), hold-time or how long the liquid cryogens last, outgassing which is the slow, natural escape of gases trapped on interior surfaces, electrical and optical access, and the ease with which the cryostat can be attached to other instruments. For infrared work there is the additional problem of shielding the sensitive array detector from the flux of infrared photons emitted copiously by all warm objects, such as the outside walls of the cryostat. This is accomplished by a double radiation shield, with the inner shield being as cold and as light-tight as possible.

The larger the surface area of a cryostat, the greater the deformation due to the atmospheric pressure of 14.7 psi at sea level (but only 61% of this value at the summit of Mauna Kea, Hawaii) on the walls, and the greater the stress on the joints. Metal vacuum chambers are most often cylindrical because a cylinder is an ideal shape for resisting pressure. Flat plates are usually employed to seal the ends of the chamber. The deflection at the center of a flat end plate clamped at its edge (e.g., by welding) is

$$\delta = \frac{3(1-\mu^2)R^4}{16Ed^3}P \tag{6.21}$$

and the maximum tensile stress, which occurs at the edge, is given by

$$S_{max} = \frac{3}{4}\left(\frac{R}{d}\right)^2 P \tag{6.22}$$

where μ is Poisson's ratio (typically 0.3 for metals); E is the elasticity or Young's modulus; R is the radius of the plate; d the thickness; and P the external pressure. For 304 series stainless steel $E = 2.8 \times 10^7$ psi and for 6061-T6 aluminum $E = 10 \times 10^7$ psi. An acceptable value of R/d for steel is 30 which yields a relative deflection of $\delta/R = 0.002$. To achieve the same result an aluminum plate must be thicker with $R/d = 20$. In the case of aluminum, a 0.5-inch thick plate with a 10-inch radius ($R/d = 20$) deflects 0.02 inch and the maximum tensile stress for $P = 14.7$ psi is 4.41×10^3 psi. Compare this with the yield strength of 6061-T6 material of 40×10^3 psi and 5×10^3 psi for almost pure aluminum. Since welding will weaken the strength of the material, the true yield strength is around 20,000 psi, which is still

safe. The deflection of a circular plate with unclamped edges (e.g., an o-ring seal with relatively few bolts) is larger and requires up to a 50% increase in thickness for a given radius to achieve a comparable δ/R. Specifically,

$$\delta = \frac{3(1-\mu)(5+\mu)R^4}{16Ed^3}P \tag{6.23}$$

and

$$S_{\max} = \frac{3}{8}\left(\frac{R}{d}\right)^2(3+\mu)P \tag{6.24}$$

where the maximum tensile stress now occurs at the center.

Design objectives for a good cooling system should include the following:

(1) Minimum detector movement, either along the optical axis (defocus) or at right angles to the optical axis (image smear), to yield spatial stability on the CCD relative to the telescope focal plane.
(2) Minimum effort to keep the system uniformly cold. When using liquid nitrogen as the coolant, this requirement simply means that the LN_2 should not run out too quickly (i.e., it should have a good "hold-time"). The hold-time of a cryostat depends on its thermal and mechanical design, the amount of power generated by the device being cooled, and the quality of the vacuum. Typically, the hold-time in hours is about $45 \times$ volume of LN_2 in liters divided by the input power in watts. A hold-time of less than 12–14 hours is not acceptable for astronomy since a refill would be required during the night. For a 1-liter capacity cryostat to last 15 hours the heat input should be less than 3 W.
(3) Good accessibility. When it is necessary to work on components inside the cryostat it is important to have enough "finger-room", good handling arrangements, and adequate safety protection for delicate electronic parts including the CCD itself. Actual handles on the cryostat or on a large internal sub-assembly are often overlooked.
(4) Cost and manufacturing time. To achieve the above objectives usually means that a cryostat purchased "commercially off the shelf" (COTS) will require some modifications or additions, which will cost time and money. Nevertheless,this can be the best approach when a good dialogue between the company and the observatory is established. The alternative approach is to design your own cryostat. This is a difficult route and may not save time and money. More often, researchers find it cost-effective to construct a hybrid cryostat by separating the cooling function from the function of supporting the detector and associated components. In the case of a CCD camera, a small "camera head" vacuum chamber can be attached to a proven, commercially available cryostat. A typical LN_2 cryostat used with a CCD is shown schematically in Figure 6.11.

After initial evacuation the pressure in a cryostat may tend to rise slowly again due to "out-gassing". A simple, widely used class of vacuum pumps which remove unwanted trace gases in the vacuum chamber by physically tying up molecules on a

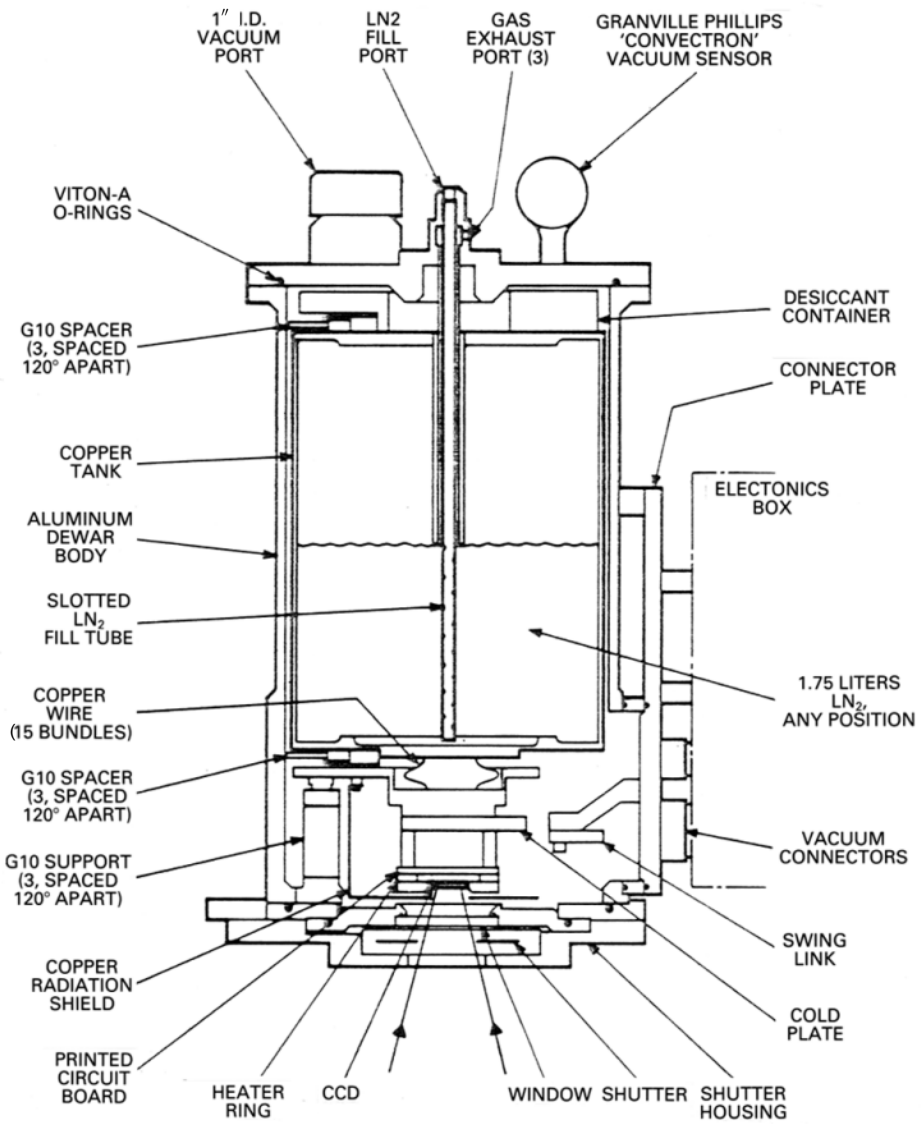


Figure 6.11. A cross-sectional view of a typical liquid-nitrogen (LN_2) cryostat illustrating all of the components needed in its construction. Credit: NOAO.

surface are called ''sorption pumps''. These can be constructed as small chambers about one inch in diameter filled with a sorbent material such as ''activated charcoal'' or one of the synthetic zeolite materials known as molecular sieves. These materials have a huge effective surface area of thousands of square meters per gram. At liquid-nitrogen temperatures these materials will absorb air, and at very low temperatures of about 10 K, zeolite can be used to partially absorb helium.

Air at 20°C and 1 atmosphere of pressure (1 atmosphere = 760 torr and 1 torr = 132 Pa) contains about 2.7×10^{19} molecules per cubic centimeter and the average distance traveled between collisions, called the mean free path, is about 7×10^{-6} cm. In general,

$$\lambda_{\text{mfp}} = \frac{1}{\sqrt{2} n \pi d^2} \tag{6.25}$$

where n is the number density of molecules; and d is the diameter of the molecule. A "rough" vacuum is about 10^{-3} torr (mean free path = 5 cm) and a "high" vacuum would be 10^{-6} torr (mean free path 5×10^3 cm). The capacity of a vacuum pump is specified in terms of the pumping speed at the inlet, which is just the volume rate of flow $S = dV/dt$ liter/s. The throughput of the flow is $Q_p = PS$ (torr-liter/s) and the throughput of the pumping line is given by $Q_p = C\,\Delta P$, where ΔP is the pressure gradient and C is called the conductance which depends on gas pressure and viscosity. Since this equation is also analogous to Ohm's Law ($V = IR$) for electrical circuits, the net pumping speed of a pump and a system of pumping lines is given by

$$\frac{1}{S} = \frac{1}{S_{\text{pump}}} + \frac{1}{C_{\text{lines}}} \tag{6.26}$$

where the net conductance is found by adding the individual conductances like their electrical counterparts. Two equations are given for C (in liter/s). The first corresponds to viscous flow when the mean free path is small, and the other to molecular flow when the mean free path is large compared with tube dimensions and C is independent of pressure. Both apply to air at 20°C:

$$C = 180 \frac{D^4}{L} P_{\text{av}} \quad \text{or} \quad C = 12 \frac{D^3}{L} \tag{6.27}$$

It is assumed that the tube is circular with diameter D and length L in centimeters and the pressure is in torr. Finally, the pump down time (in seconds) of a system with volume V from pressure P_0 to P assuming a constant net pumping speed S and no out-gassing is

$$t = 2.3 \frac{V}{S} \ln\left(\frac{P_0}{P}\right) \tag{6.28}$$

Typically, the chamber is rough-pumped to about 5×10^{-2} torr with a mechanical pump and then pumped to a lower pressure with a diffusion pump or turbo-molecular pump. Typical pump speeds are 100 liter/s at the inlet. It is important to know how long the rough pumping will require because the pressure in the diffusion pump section cannot be allowed to rise significantly. Out-gassing from metals can be about 10^{-7} torr liter/s per cm^2 after pumping for 1 hour, but this can be reduced an order of magnitude after 24 hours of pumping. Plastics outgas at 100 times this rate and should be avoided. Pressure within the vacuum vessel can be monitored with several types of gauges. During rough pumping the most common gauge is a Pirani gauge in which the temperature of a filament depends on the rate of heat loss to the surrounding gas; heat loss decreases as the gas pressure falls. For lower pressures a Penning gauge is used. The principle of this device is the production of a small electrical

Figure 6.12. A pumping station with a diffusion pump (vertical stack) and a roughing (rotary) pump in the lower right.

current when molecules are ionized in a magnetically confined cold-cathode discharge between two electrodes at room temperature; the lower the pressure the smaller the current. Figure 6.12 shows a two-stage pumping station with a diffusion stack and LN_2 trap on the left for high vacuum work and a rotary roughing pump below.

6.6 ELECTRONICS DESIGN

If the optics provide good images, but the electronics system is noisy, then you will end up with poor signal-to-noise ratios! Needs in this category generally split into analog and digital electronics. As before, a block diagram layout using a "top-down" approach is best because it usually results in identification of much smaller units,

some of which can be duplicated many times and others can be purchased commercially. The primary analog circuits are amplifiers. Depending on the detector, there may be a pre-amplifier and a post-amplifier, either dc or ac coupling may be required, operational amplifiers (op-amps) may suffice or designs based on individual transistors (typically MOSFETs and JFETs) may be needed. Digital circuitry can be very compact and much effort is saved by using some form of microprocessor and programmable array logic and commercial data acquisition boards. One of the key components is the analog-to-digital converter (ADC). A comparison of the voltage resolution of the ADC and the output signal strength from the detector will determine the amplifier gain. Careful attention to grounding schemes and isolation of the digital and analog sections is very important for control of electronic noise. It is also important to consider how the electronics will be packaged and integrated with the mechanical design.

Apart from the basic rules of direct-current circuits such as Ohm's Law, $V = IR$, where I is the current and V the potential difference across an electrical resistance R, and the usual rules for adding resistances in series ($R_T = R_1 + R_2 + \cdots + R_n$) and in parallel ($1/R_T = 1/R_1 + 1/R_2 + \cdots + 1/R_n$), it is important to remember issues such as power dissipation. For example, the power (watts) dissipated as heat in a wire of resistance R is $W = IV = I^2R = V^2/R$ and the resistance is given by $R = \rho L/A$ where ρ is the resistivity (the inverse of conductance), L is the length, and A is the cross-sectional area of the conductor. Wires of the correct gauge must be chosen to carry the current needed. In vacuum systems there is no conduction with the air to help carry heat away and therefore electrical wires and components can over-heat if not properly clamped down (heat-sunk) to a large, cold mechanical plate.

The relationship between electric field, electric potential, and charge distribution is important for understanding CCDs, infrared arrays, and other detectors. A capacitor is defined as any two conductors separated by an insulator. The capacitance, C, of a system is given by $C = Q/V$, where Q is the charge on either conductor and V is the potential difference between them. An important special case is the parallel plate capacitor composed of two closely spaced plates of area A and separation d. The electric field magnitude (E) between the plates is given by $E = V/d = Q/\varepsilon A$ and the capacitance is given by

$$C = \frac{\varepsilon A}{d} \tag{6.29}$$

where ε is the permittivity of the insulator between the plates. It is often convenient to use the capacitance per unit area C/A which is just ε/d. More generally, the electric field is the gradient of the electric potential (i.e., $E_x = -dV/dx$). When capacitors are connected in series (same charge on all) the total capacitance is given by the sum of the reciprocals so that $1/C_T = 1/C_1 + 1/C_2 + \cdots + 1/C_n$. If n capacitors are connected in parallel (common voltage across all) then the total capacitance is given by $C_T = C_1 + C_2 + \cdots + C_n$. This is *opposite* to the rule for resistors. When a voltage is applied to an uncharged capacitor, the charge will build up to the value $Q = CV$ exponentially with a time constant determined by resistance in the circuit. The time constant is $\tau = RC$ and the variation in charge with time is given by

$q = Q(1 - e^{-t/RC})$. Likewise, when the voltage is removed the charge will decay exponentially with the same time constant. These effects are fundamental to the operation of the output of a CCD and will be used many times in subsequent chapters.

A capacitor can also be constructed from a metal plate on one side and a semiconductor substrate on the other with an insulator in between. If the semiconductor is silicon, the insulator is usually silicon dioxide (usually called the "oxide" layer). This construction is known as a metal oxide semiconductor or MOS capacitor and is the basis for the CCD. Because there are two types of semiconductor, p-type and n-type, then there can be p-MOS and n-MOS capacitors. Suppose we have a p-MOS capacitor so that the semiconductor has an excess of acceptor sites (i.e., the majority charge carriers are holes), and we apply a positive voltage to the metal plate relative to the semiconductor substrate. Holes are driven away from the surface of the semiconductor by the positive electric field, leaving behind uncompensated negatively charged acceptor atoms, thus creating a region depleted of mobile carriers (holes in this case). The number of holes driven away equals the number of positive charges on the metal gate, thus $Q_i = eN_A x_d$, where e is the numerical value of the charge on the electron, Q_i is the ionized acceptor charge concentration (in coulombs/cm^2) in the depletion region of the semiconductor, x_d is the depletion depth (cm), and N_A is the acceptor-doping concentration (atoms/cm^3). The depletion region is non-conductive and acts as an insulator with a capacitance per unit area (and hence an ability to store charge) of $C_{\text{dep}} = \varepsilon_s / x_d$, where ε_s is the permittivity of silicon. Finally, because the gradient of the electric field depends on the charge distribution we have that $dE/dx = \rho/\varepsilon_s$ where in this case the symbol ρ is the total charge density (particles/cm^3) due to p free holes, n free electrons, N_A fixed dopant acceptors (atoms/cm^3), and N_D fixed dopant donors (atoms/cm^3). Most free carriers are swept out of the depletion region by the electric field, and for a p-type semiconductor the only term is N_A and therefore $dE/dx = d^2V/dx^2 = eN_A/\varepsilon_s$. From these equations and the boundary conditions we can solve for the electric field and electric potential in the depletion region.

We have already made use of an important electrical result, the voltage divider rule, when studying the analogous situation in heat flow. In a circuit with resistors R_1 and R_2 in series, the current $I = V/(R_1 + R_2)$ flows through both due to the potential V. The voltage drop across R_2 is given by $V_{\text{out}} = IR^2$ or

$$V_{\text{out}} = \left(\frac{R_2}{R_1 + R_2}\right) V \tag{6.30}$$

Most solid-state imaging devices (CCDs, infrared arrays) require a pre-amplifier fairly close to the output pin before the signals can be sent down long lengths of cable. A typical design for a CCD preamp is shown in Figure 6.13. Note the capacitor between the detector and preamp itself, indicating that this circuit is "ac-coupled". For higher speed operation, as encountered with infrared detectors, it is necessary to eliminate this capacitor and directly "dc-couple" the detector and preamp. Under these circumstances there will have to be an additional circuit component to "offset" the dc level prior to amplification. In some cases the preamp is placed inside the

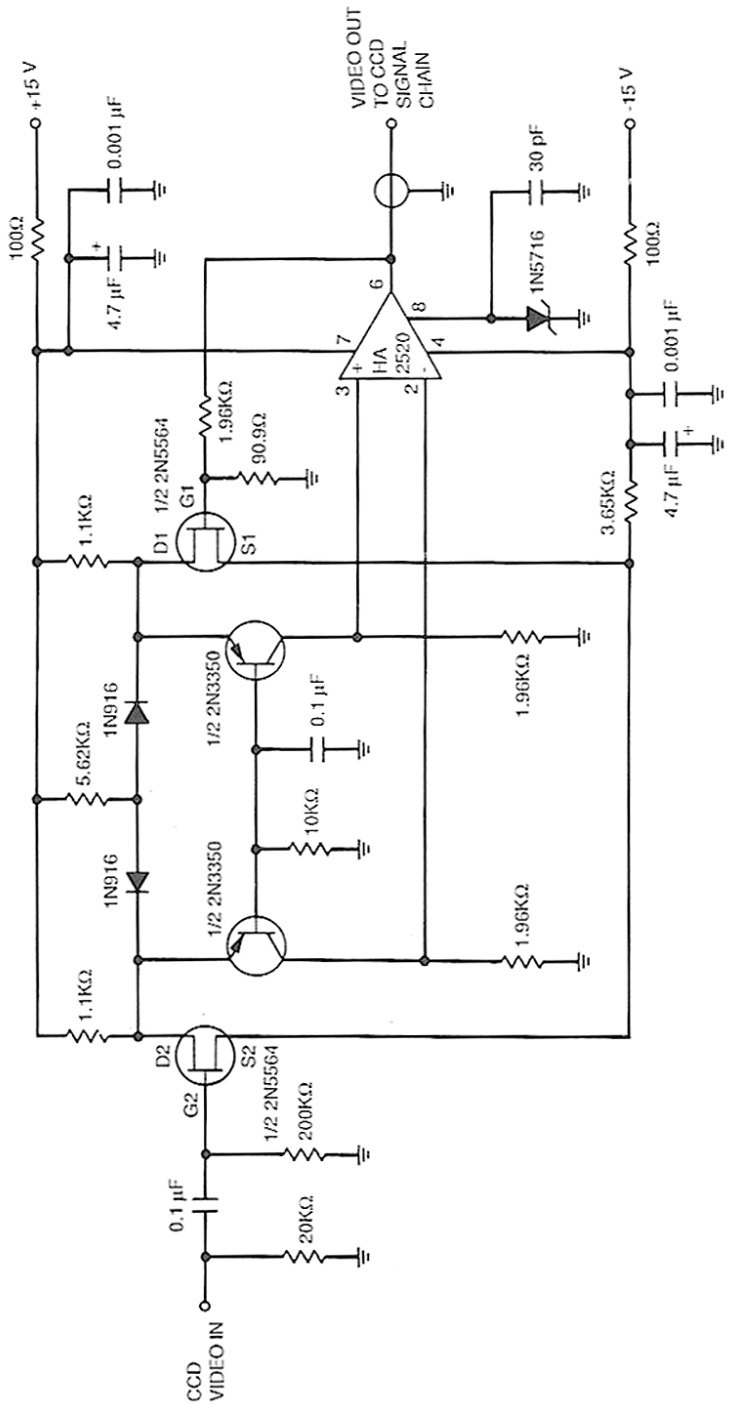


Figure 6.13. A CCD pre-amplifier based on an original concept used at the NASA Jet Propulsion Lab (JPL). Credit: Lloyd Robinson.

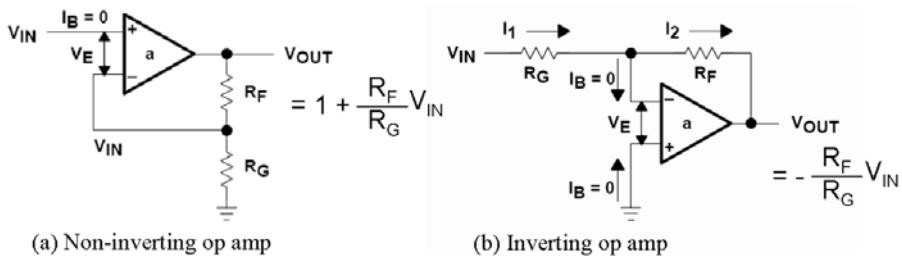


Figure 6.14. The use of op-amps for inversion and non-inverted gain.

cryostat, as close as possible to the CCD. Other workers have had good success with preamps placed immediately outside of the cooled chamber provided the lines are well-shielded and low-capacitance (e.g., twisted pair wires). This approach is obviously attractive from the point of view of accessibility. Circuit simulation packages are available for workstations and PCs to help the designer, of which the best known is probably SPICE (Simulation Program with Integrated Circuit Emphasis) which was originally developed at the University of California in 1976. Later commercial implementations employ component libraries to store the characteristics of each device. Software such as PCAD enables designers to layout the circuit for manufacture on a printed circuit board (PCB) and the file (called a Gerber file) can be downloaded directly to the company who will fabricate the board.

Preamps can be designed with discrete components or with operational amplifiers (op-amps) which are constructed on a single integrated circuit chip and are "differential" amplifiers having two inputs labeled "+" and "−" for non-inverting and inverting, respectively. A wide range of op-amps are available, and therefore it is necessary to examine the data sheets carefully to select the best device for the application. Parameters to check are input impedance, output impedance, output voltage range, supply voltage range, common-mode rejection ratio, which is the ability of the device to reject noise common to both inputs, input bias current, input offset voltage and its temperature dependence, frequency response, and noise. Manufacturers normally report the root-mean-square (rms) noise voltage divided by the square root of the frequency bandwidth of measurement (e.g., 10 nanovolts per root hertz), but watch out for the difference in voltage noise and current noise. Figure 6.14 illustrates two useful arrangements of op-amps that appear in CCD circuits.

Field Effect Transistors: FETs are encountered frequently, both in detectors themselves and in external circuitry. A FET is a three-pin device in which the output current is determined by the voltage applied to the input terminal. The name derives from the fact that the electric field produced by the input voltage controls current flow. Two basic types are MOSFETs and JFETs (see Figure 6.15). The MOS prefix indicates that the input terminal has a metal oxide semiconductor structure, whereas the J prefix indicates that the input gate of the device forms a pn junction with the substrate. In both cases, two n-type regions called the "drain" and the "source" are formed in a p-type substrate. In the MOSFET these two regions are linked by a thin

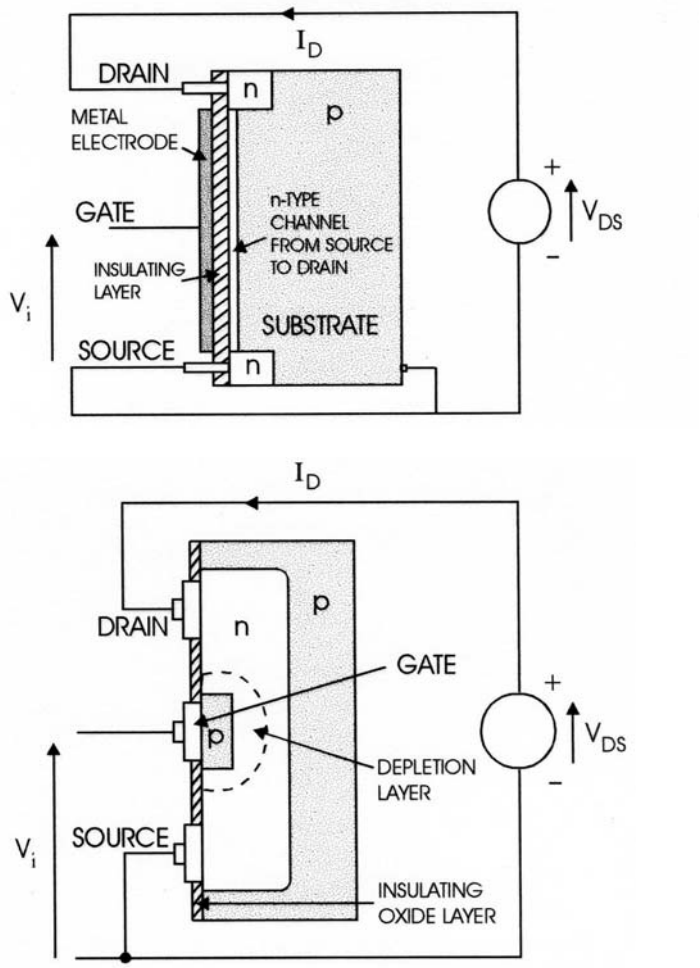


Figure 6.15. Schematic representations of the basic structures of a MOSFET (top) and a JFET (bottom). Both rely on controlling the width of a conducting channel by means of an internal electric field.

channel of n-type silicon at the surface which forms the "S" part of the MOS input or "gate". The voltage applied to the gate controls the effective depth of the thin n-type channel and hence the amount of current that can flow from drain to source. If the gate is positive with respect to the source then the current increases. In the JFET, the MOS gate is replaced with a small region of p-type material diffused into the n-type channel to form a reverse-biased pn junction with a depletion region which effectively controls the flow of current from the drain to the source. As the gate voltage is made more negative the depletion region increases and conductivity drops.

A common circuit found in CCD and IR array detectors is the "source follower" or common-drain amplifier in which the source voltage (the output) tends to follow

the gate voltage (the input). When the product of the transconductance (g_m) and the source load resistance (R_s) is large, then $1/g_m R_s \ll 1$ and $V_{\text{out}} \sim V_{\text{in}}$; in practice there may be a loss of gain by about 20% but the advantage of these circuits is very high input impedance and fairly low output impedance. FETs are also used as simple switches. A sufficiently large negative gate voltage effectively cuts off current flow completely. Another use of FETs is as a constant current source when a constant voltage is applied to the gate. A simple constant current arrangement can be obtained by connecting the gate and source and placing the load between the drain and the drain supply voltage.

Digital circuitry includes timing-pattern generators, analog-to-digital converters, interface circuits or buffers, and memory circuits. Here there is a lot more commercial hardware available to the designer. While it is certainly possible to use discrete TTL (Transistor Transistor Logic) and CMOS (Complementary Metal Oxide Semiconductor) logic chips for small "hardwired" systems, more flexibility is achieved by using one of the several classes of microprocessors together with some devices called programmable logic arrays (PLAs; also called PALs and PLDs).

Electrical interference and noise can be a serious problem in scientific apparatus with low signal levels. An excellent treatment of basic principles is given by Ott (1988). Ground loops must be eliminated, signal wires must be screened or shielded, separate power supplies should be used for digital and analog sections, switching-type power supplies should be avoided unless very well shielded, and opto-isolation should be used. In the opto-isolator digital signals are turned into pulses of light from a light-emitting diode which are then detected by a photo-transistor and converted back into electrical pulses. A useful summary of good procedures is

- establish a single point ground for each system and do not allow the analog and digital parts of the system to share the same ground
- use large diameter multi-strand cable to reduce the impedance for ground connectors
- shield all the signal cables completely, but ground the shield at the power source end only
- use optical fibers or shielded twisted pair cables where possible
- group cables by function and sensitivity to noise, avoid long parallel runs, separate dissimilar types such as analog signal cables, digital signal cables, and power lines
- use shielded cabinets.

Electric motors fall into three categories: ac motors, dc motors, and stepper motors. DC motors are used extensively in precision control applications requiring low power, whereas for high-power roles and less demanding precision ac motors are selected. In dc motors there is an almost linear relationship between speed and applied voltage, and between torque and current. A stepper motor consists of a central rotor surrounded by a number of coils (the stator) with diametrically opposite coils connected together so that energizing that pair of coils will cause the rotor to align itself with that pair. Applying power to each set of windings in turn causes the

rotor to jump or step to the next position. To minimize the number of wires involved, the coils are grouped in sequence with, for example, the fifth coil connected to the first and similarly for the ninth, thirteenth, etc., so that when the rotor is at the fourth coil and the first and fifth are energized together, the rotor steps to the fifth because it is closest. The speed of rotation is controlled by the pulse repetition rate. Since the frequency can be changed easily by modifying the pulse waveform period, then it is possible to build in acceleration and deceleration ramps.

6.7 SOFTWARE

If the electronic imaging detector is the "heart" of an astronomical instrument, then software must be the "muscles"! Good software can make an immense difference to the ease with which an astronomer can learn to use an instrument efficiently and productively. Many levels of software may be required depending on the instrument. For example, software will be required for the following tasks:

1. Interface to the astronomer
   - Graphical User Interface (GUI, often pronounced "goo-ee")
   - Command Line Interface (CLI)
   - scripting or macro interface to allow sequences to be pre-programmed
2. Data acquisition
   - capture and storage of digitized data
   - manipulation of the data stream and attachment of "header" information
3. Data display and analysis
   - display of the data
   - "quick-look" facilities for analyzing the data
4. Instrument control
   - conversion of high-level commands to hardware signals
   - handling of status signals from the hardware.

Software is an area in which many astronomy graduate students can make a contribution, but it is very important to take the opportunity to develop good programming and documentation habits. More than one level of programming may be needed. For example, it is frequently the case that a "high-level" language is used for the user interface (such as C) and "low-level" coding of Digital Signal Processors (DSPs) and other microprocessors is done in the appropriate assembly language. Image-processing software and data reduction "pipelines" are now very common. Some of the most important systems will be discussed in Chapter 10.

Again, the top-down approach works best. Produce a flow chart of all the data and communication paths and develop an iterative approach to improve the model of the software as illustrated in Figure 6.16. Keep the code modular rather than monolithic. Try to make it easy for the first time user by employing a Graphical User Interface, while also catering to the expert user by providing shortcuts and hot keys. Figure 6.17 shows a GUI developed at UCLA for the NIRSPEC instrument at the

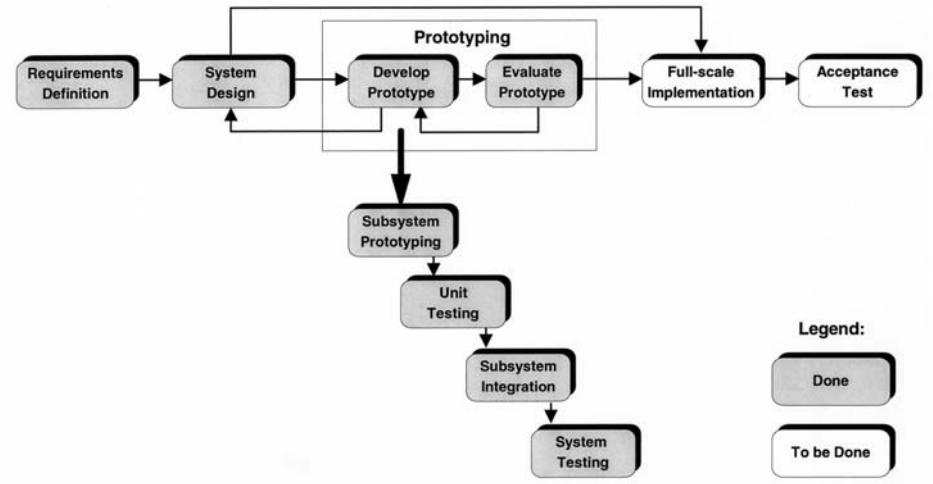


Figure 6.16. A flow chart illustrating a systematic method of software development for astronomical instruments.

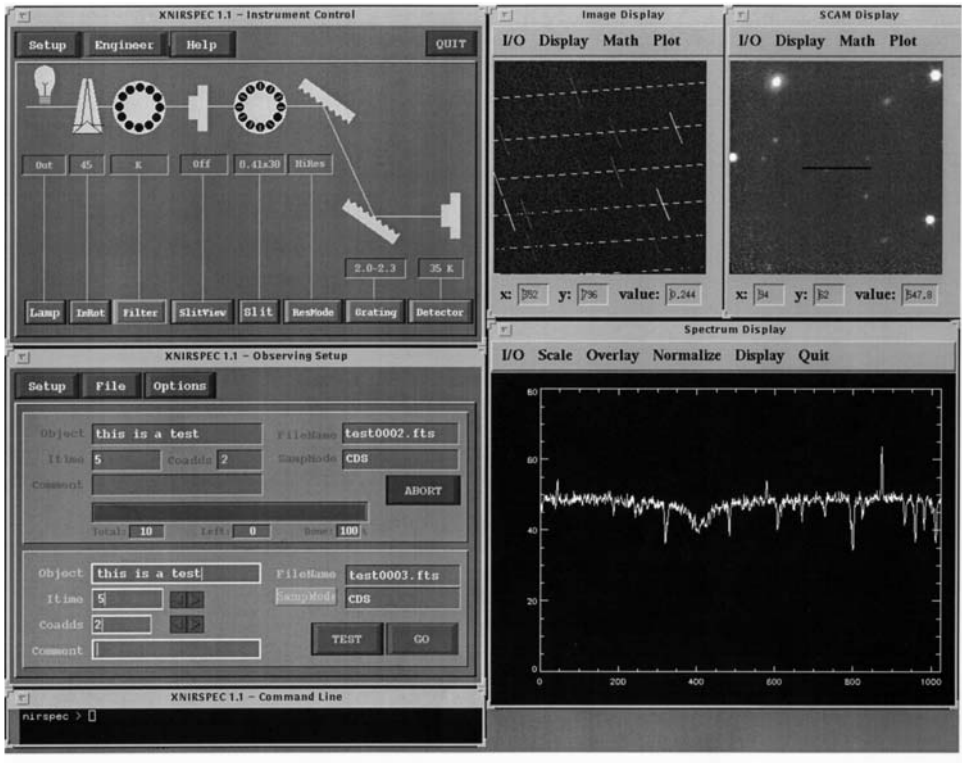


Figure 6.17. An example of a Graphical User Interface (GUI) used to control a spectrograph and imaging system. This screen was developed for the Keck NIRSPEC project using DataViews and IDL.

Keck telescope. Stick with a programming language that is widely accepted and supported. In practice this probably means ANSI C (ANSI is the American National Standards Institute) and establish a set of documentation standards for coding the software and adhere to them, even if it takes more time; other people will be very grateful later. Finally, it is worth remembering that it is almost always the software and user interface to an instrument that people see, rather than the physical apparatus in the telescope dome (or in orbit). Therefore, it is worthwhile to expend considerable effort to get the user interface and software performance as good as possible.

6.8 SUMMARY

Designing and developing modern astronomical instruments requires a broad knowledge of basic physics, engineering practice, and software. This chapter has touched on some of the background knowledge needed to really appreciate what goes into the designing and building of such instruments. A modular approach is generally advisable. The design specification must flow down from the scientific requirements. Elementary optical and mechanical calculations will need to be followed by detailed ray tracing and computer-aided drawing in three dimensions. Thermal analysis for vacuum cryogenic systems will be required, and electronics to operate the instrument and detector will need to be designed. Software to control the entire system must be robust and easy to use. It is important to recognize both the limits and opportunities of the technologies selected.

6.9 EXERCISES

1 Draw a block diagram layout of a cooled CCD camera system with a filter wheel which is to be used on a small telescope and remotely controlled from a "warm room" 15 m away using a PC computer.

2 What is meant by an anti-reflection coating and why are they important in astronomy applications? What is the optimum refractive index of the coating to be applied to a silicon surface ($n = 4$)? What thickness of coating is required at a wavelength of 2.2 μm?

3 Match the following three detectors to a 0.2 m telescope and then to an 8 m telescope: a Kodak KAF-4200 CCD with 9 μm pixels in a 2,048 × 2,048 format, a SITe CCD with 22 μm pixels in a 1,024 × 1,024 format and the Raytheon InSb array with 27 μm pixels in a 1,024 × 1,024 format. Assume Nyquist sampled seeing of 2″ and 0.5″, respectively.

4 Given the relation shown below for the angular blur size diameter (radians) due to spherical aberration in a single lens of refractive index $n = 1.5$ and focal ratio

$f/2$, determine the linear diameter of the image of a point source if the focal length of the camera is 50 mm:

$$\beta = \frac{n(4n-1)}{128(n+2)(n-1)^2}\,\frac{1}{(f/\text{number})^3}$$

How does this compare with the size of a typical CCD pixel? How can this limitation be overcome in practice?

5 The diameter of the blur circle for an on-axis point source image in seconds of arc for a spherical mirror is given by

$$\beta = \frac{206{,}265}{128(F)^3}$$

Determine the focal ratio (F) that yields an image *diameter* of 1 arcsecond.

6 Spherical aberration is zero for a paraboloidal mirror, but sagittal coma is given by

$$\beta = \frac{1}{16(f/\#)^2}\theta$$

where β is the blur circle diameter in seconds of arc and θ is the off-axis field angle in seconds of arc; the head-to-tail extent is 3β. Determine the image blur due to coma $1'$ off axis for an $f/3$ mirror. How would this change if the primary was $f/1.5$?

7 (a) Calculate the diffraction limit for an $f/2$ lens with a focal length of 50 mm at a wavelength of 500 nm in the mid-visible. (b) The depth of focus is the amount of defocus which introduces $\pm\lambda/4$ wavefront error; $\Delta f = \pm 2\lambda(f/\#)^2$. Determine the depth of focus for the lens in part (a).

8 Calculate the thermal stress induced in an aluminum strut which is fixed at both ends if the temperature decreases by 210 K and compare this with the yield strength. Is there a problem? Ignore the temperature dependence of the constants. Assume $\alpha = 24 \times 10^{-6}\ \text{K}^{-1}$, Young's modulus $E = 10 \times 10^6$ psi, and the yield strength is 40,000 psi. [Solution: $F/A = -\alpha E\,\Delta T = 50{,}400$ psi, because $\Delta L/L = \alpha\,\Delta T$ in the definition of Young's modulus. Yes, the strut will buckle and yield.]

9 An infrared cryostat has a surface area of $5\,\text{m}^2$. Assuming that the geometric factor is one-half the emissivity of 5%, calculate the radiation load on a 77 K interior from (a) laboratory temperature of 300 K and (b) mountain observatory temperature of 275 K. What could you do to reduce the load on the internal cold components?

10 Derive expressions for the étendue of a spectrometer with an entrance aperture defined by the slit, and derive a similar expression for a seeing-limited camera. See theory below.

Special topic

Étendue: also called luminosité or throughput. Consider a surface element of area dS and brightness B radiating into a solid angle $d\Omega$. The flux (energy rate in J/s) dF is

$$dF = B \cos\theta \, dS \, d\Omega$$

dS is called the entrance pupil and could be the seeing disk, the Airy diffraction disk, or the entrance slit of the spectrometer depending on the conditions. Integrating from θ to $\theta + d\theta$ the value of $d\Omega = 2\pi \sin\theta \, d\theta$ yields

$$F = 2\pi B \, dS \int_0^{\theta_m} \cos\theta \sin\theta \, d\theta = BU$$

where θ_m is the maximum value of the half-angle of the cone of rays from element dS, and the quantity $U = \pi \sin^2 \theta_m \, dS$ is called the étendue. When the étendue is evaluated for different systems (cameras and spectrometers) it is always the product of the area (A) of the entrance aperture and the solid angle (Ω) subtended by the entrance pupil; it is therefore also known as the $A\Omega$ (pronounced ay-omega) product. Assuming no losses due to absorption or reflection, then conservation of energy demands that $U(= A\Omega)$ is conserved.

6.10 REFERENCES AND SUGGESTED FURTHER READING

Avallone, E.A.; and Baumeister III, T. (eds.) (1996) *Mark's Standard Handbook for Mechanical Engineers*, tenth edition, McGraw Hill, New York. [Very extensive treatment of mechanics.]

Born, M.; and Wolf, E. (1999) *Principles of Optics*, seventh edition, Cambridge University Press, Cambridge, U.K. [Classic text.]

Epps, H.W. (2002) Astronomical optics: Back then ... right now ... soon to follow, in F. Bash and C. Sneden (eds.), *Astronomical Instrumentation and the Birth and Growth of Astrophysics*, Astronomical Society of the Pacific, San Francisco, ASP Conference Series Vol. 270, pp. 102–122.

Fischer, R.F.; and Tadic, B. (2000) *Optical System Design*, McGaw-Hill, New York. [Excellent introductory text.]

Hecht, E. (2001) *Optics*, fourth edition, Addison-Wesley, Reading, MA. [Excellent introduction]

Korsch, D. (1972) Closed form solution for three-mirror telescopes, corrected for spherical aberration, coma, astigmatism and field curvature, *Applied Optics*, **11**, 2986–2987.

Kingslake, R. (1978) *Lens Design Fundamentals*, Academic Press, New York. [Classic text; mainly lenses.]

Moore, J.H.; Davis, C.C.; and Copland, M.A. (1983) *Building Scientific Apparatus*, Addison-Wesley. [Dated, but outstanding introduction to laboratory practice.]

Offner, A. (1975) New concepts in projection mask aligners, *Opt. Eng.*, **14**, 130–132.

Ott, H.W. (1988) *Noise Reduction Techniques in Electronic Systems*, second edition, John Wiley & Sons, New York. [Classic text; excellent.]

Rieke, G.H. (2003) *Detection of Light from the Ultraviolet to the Submillimeter*, second edition, Cambridge University Press, Cambridge, U.K. [Excellent unified treatment of detectors.]

Schroeder, D.J. (2000) *Astronomical Optics*, second edition, Academic Press, San Diego, CA. [Covers telescopes and spectrographs in detail; essential reading.]
Smith, W.J. (2000) *Modern Optical Engineering*, third edition, McGraw-Hill, New York. [Classic text; excellent.]
Storey, N. (2006) *Electronics: A Systems Approach*, third edition. Prentice Hall, Harlow, U.K. [Comprehensive and very readable.]
Wolfe, W.L.; and Zissis, G.J. (eds.) (1989) *The Infrared Handbook*, third edition, Information Analysis Center for the Office of Naval Research. [Classic reference.] NIST: *http://www.ncnr.nist.gov/equipment/displex.html*

Web sites for optical software

http://www.opticalres.com
http://www.breault.com
http://www.zemax.com

7

Charge-coupled devices

At the heart of all astronomical instruments is some form of detector to convert electromagnetic energy into an electrical signal. Having indicated that the dominant detector in modern astronomy is the charge-coupled device (CCD), it is important to consider these remarkable detectors in more detail. We begin with a brief historical review of CCDs from invention to present day, an amazing story by itself, and then cover the basic principles of how CCDs work. Practical details will follow in the next chapter.

7.1 THE EARLY YEARS

7.1.1 Invention and development

As mentioned in Chapter 1, the charge-coupling principle was invented in 1969 by Willard Boyle and George Smith and demonstrated in a simple one-line nine-electrode device by Gil Amelio, Mike Tompsett, and George Smith at the Bell Laboratories in New Jersey, U.S.A. Larger image-forming devices of 100×100 pixels were not introduced until 1973, and Boyle and Smith received the basic patent at the end of 1974. From the original small arrays available around 1973, CCDs have come a long way. Formats of $2{,}048 \times 4{,}096$ pixels[1] with no wires or structure on three sides are readily available with most observatories using CCD "mosaics" composed of many such "close-butted" devices. Figure 7.1 shows an historical collection of CCDs including some large-format devices.

[1] These strange numbers are simply powers of 2; $2^{10} = 1{,}024$, $2^{12} = 4{,}096$. There is no fundamental reason to use these numbers.

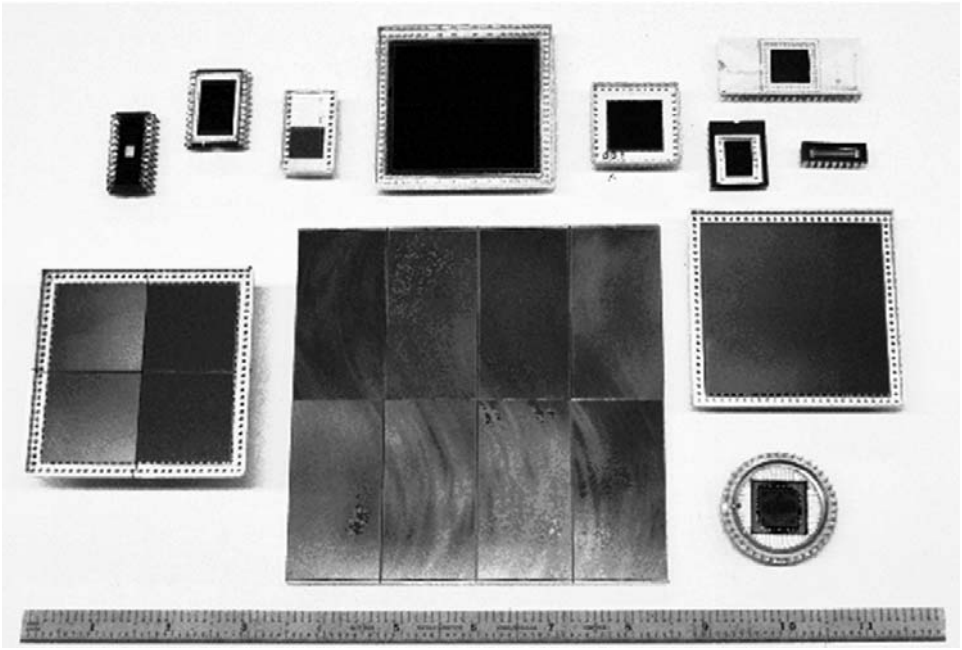


Figure 7.1. A collection of CCDs including eight large-format ($2\,\mathrm{K} \times 4\,\mathrm{K}$) devices butted together to form a 64-megapixel mosaic. Credit: Gerry Luppino.

7.1.2 The astronomical push

Many astronomy-related groups familiar with imaging technology, usually with vidicon-type systems, were alert to the potential of CCDs in the early 1970s. Gerald M. Smith, Frederick P. Landauer, and James R. Janesick of the Advanced Imaging Development Group at the NASA Jet Propulsion Laboratory operated by the California Institute of Technology (Caltech) in Pasadena, and Caltech scientist James Westphal (1930–2004) were among the first to recognize the potential advantages of such an imaging device for astronomy and space applications. In 1973, JPL joined with the National Aeronautics and Space Administration (NASA) and with Texas Instruments (TI) Incorporated (Dallas) to initiate a program for the development of large-area CCD imagers for space astronomy, in particular for the proposed Galileo mission to Jupiter. Originally scheduled for 1981, the Galileo spacecraft was finally launched in 1989 and arrived at Jupiter in December 1995. This incredibly successful mission ended on September 21, 2003 when the satellite was directed to plunge into Jupiter's atmosphere.

During the period 1973 to 1979 Texas Instruments (TI) developed CCD arrays of 100×160 pixels and 400×400 pixels, then 500×500 pixels and finally an 800×800-pixel array. Testing and evaluation of these devices was carried out at JPL by Fred Landauer and by a young engineer named Jim Janesick, who just happened to be an amateur astronomer. In 1974, Jim attached a 100×100 Fairchild CCD to his small

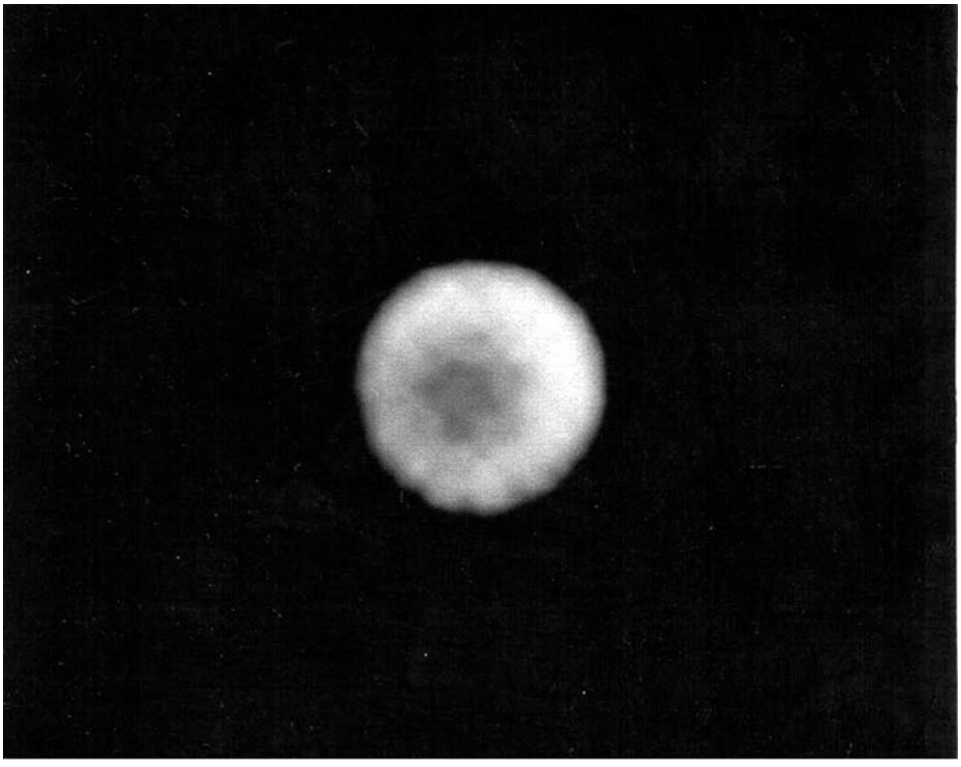


Figure 7.2. An image of the planet Uranus obtained in 1976 through a methane-band filter with a Texas Instruments CCD. This was the first astronomical CCD image on a professional telescope. Photo courtesy of Brad Smith.

20.3 cm (8-inch) telescope and succeeded in imaging the Moon (Janesick, 2001). Having already approached one astronomer at a national institute about testing a CCD on a professional telescope, and having been turned down, Jim luckily met and teamed up with Dr. Bradford Smith, a planetary scientist at the University of Arizona's Lunar and Planetary Laboratory. In early 1976 they obtained the first astronomical imagery with a charge-coupled device on a professional telescope. Using the 61-inch telescope designed for planetary imaging on Mt. Bigelow in the Santa Catalina Mountains near Tucson (Arizona), CCD images of Jupiter and Saturn were obtained using a special filter to pick out methane gas in the atmospheres of those giant planets. When the team turned the telescope to Uranus (Figure 7.2), they immediately thought something had gone wrong. It looked like a "donut"! After checking focus and everything else Brad realized that it must be correct, they were observing "limb-brightening" of Uranus in the methane band for the first time. As Brad Smith recalled vividly to me, "all who participated and who saw those images agreed that the potential of the CCD was superior to other imaging equipment of the time." I joined the Lunar and Planetary Lab in 1977 and got a chance to use one of

those early systems with Brad's post-doctoral researcher, Harold Reitsema (now at Ball Aerospace and Technologies Corp., Boulder, CO).

At about the same time, NASA awarded contracts for the procurement of instruments for an Earth-orbiting Space Telescope. One of the people awaiting NASA's decision was John Lowrance at Princeton University. John was working with SEC vidicon technology in the 1970s when CCDs were invented. SEC (or secondary electron conduction) vidicons seemed like the most appropriate imaging devices, at least in terms of their wavelength of response, for space ultraviolet/optical experiments such as the Orbiting Astronomical Observatory (OAO). While NASA and various advisory bodies deliberated, John continued to pursue the development of SEC vidicon systems. Finally, after many delays, NASA abandoned the OAO series in favor of the Space Telescope. By that time CCDs had been invented, and JPL had begun their studies. Following a crucial meeting at which CCD results were demonstrated, the initial plan to utilize SEC vidicons for the Space Telescope was dropped and the concept of Principal Investigator instrument teams was introduced. A proposal from a team led by planetary scientist James Westphal (1930–2004) of Caltech in collaboration with JPL was accepted for the inclusion of CCD cameras on the Space Telescope. Westphal had heard about CCDs at a committee meeting a couple of years earlier. When he insisted on knowing more about the JPL results, the chairman (Bob O'Dell) sent him along to JPL to see for himself! There he met Fred Landauer and learned that CCDs were indeed very low noise devices; 100 electrons (e) noise had been observed and 30e was predicted. On his return to Caltech he mentioned these numbers to colleague Jim Gunn of Princeton University, who was getting ready for a major project involving vidicon technology, and whose instant reaction was "that will revolutionize astronomy if it is true!"

John Lowrance, like Jim Westphal, moved away from vidicon technology and began working with CCDs. Luckily there was a key player in the game right on his doorstep, the Electro-optics Division of the Radio Corporation of America (RCA) in nearby Lancaster, Pennsylvania. At the head of the RCA group working on CCDs was Dick Savoye, and Dick was enthusiastic about the astronomical applications, and moreover he believed that the technology possessed by RCA would yield devices extremely sensitive to blue light, as later demonstrated by the superb blue response of the thinned, backside-illuminated 512×320 RCA CCD. John Lowrance at Princeton and John Geary at Harvard each established good relations with RCA and began testing these devices in the late 1970s. John Geary, having first tried an unthinned device on the 1.5 m and 60 cm telescopes on Mt. Hopkins in April 1980, visited the RCA facility in Lancaster shortly thereafter to show them the splendid results obtained so far, and urge them to provide him with a thinned CCD. He received one of the very first thinned backside-illuminated CCDs manufactured by RCA; this device was considered a reject and was lying in the desk drawer of RCA engineer Don Battson. John put it on the telescope on Mount Hopkins in September 1980, and there it remained for almost a decade!

Meanwhile, the Texas Instruments (TI) chips evolved through a program of systematic development toward the eventual goal of an 800×800 array. One of the key figures on that program from the outset was Morley Blouke. Several

approaches to the design and fabrication of CCDs were tried. A major constraint was that the device must be able to survive the harsh radiation environment around Jupiter. Therefore, two different constructions evolved called the "buried-channel" and the "virtual-phase" CCD. Tens of thousands of CCDs were being manufactured under contract to NASA (the final number was 75,000), and JPL realized, as Jim Janesick stated in a proposal to the Director of JPL in October 1976, that there was "... a need to expose and familiarize astronomers and scientists to the capabilities of the CCD for use in planetary observation and stellar studies."

Around this period (1974–1977) other companies were also beginning to develop CCDs. The first company to actually market a high-quality device was a division of Fairchild Semiconductor (Milpitas, CA) which produced a 100×100 CCD in 1974; this is the chip Jim Janesick attached to his own small telescope in 1974 to image the Moon. James Early of Fairchild was a strong advocate for the new technology, and Gil Amelio had moved from Bell Labs to Fairchild. At the Kitt Peak National Observatory (KPNO) in Tucson (Arizona), Richard Aikens and Roger Lynds had been working on low light level imaging systems for astronomy for many years. Soon the KPNO (now NOAO) began a program of development of CCDs. With Steve Marcus, this team began working on the Fairchild device. The Fairchild CCD201 and CCD202 image sensors were designed for TV applications and, although capable of high performance, they had a serious impediment for astronomical work due to the interline transfer construction which meant that they had columns of picture elements which were alternately light-sensitive and totally insensitive due to a cover layer of opaque metal strips; in terms of the image falling on the CCD these devices were half-blind! Richard Aikens left the KPNO to set up his own company in 1978, called Photometrics, which played an important role in stimulating the manufacturing of CCDs and the development of scientific camera systems in general. Photometrics became part of Roper Scientific in 1998. Fairchild also changed names several times through successive owner companies including Schlumberger, Weston, Loral, Lockheed-Martin, and BAE Systems; it is now Fairchild Imaging.

There was a time of great frustration in the late 1970s about the lack of access to CCD technology by the mainstream astronomical community. Development of the Wide-Field/Planetary Camera, abbreviated WFPC—but pronounced "wiff pick"—was going well and many people were now aware of the sensitivity and the scientific potential of CCDs. Industry too was embracing the new technology, but commercially available products were scarce. During this interlude other forms of less suitable solid-state imagers were tried such as the Charge Injection Device (or CID) from General Electric (America), or the interline transfer device from Fairchild already mentioned. When 512×320 RCA CCDs appeared in the late-1970s it was a welcome relief.

The first RCA CCDs were frontside-illuminated which meant they had a poor response to blue light. Soon, however, the thinned backside-illuminated CCDs appeared. Clearly, RCA had "the secret" for treating or passivating the thinned backside surface, and these CCDs displayed outstanding sensitivity over a huge spectral range, better even than the TI chips. Unfortunately, there was one weakness. The design of the on-chip output amplifier was poor, and so the CCD was 5–10 times

"noisier" in electrical terms than the TI CCD. Later RCA CCDs were much better. Sadly, in 1985, RCA withdrew from the CCD market for commercial reasons. Detector development work has continued, however, at the David Sarnoff Labs (Princeton, NJ).

In early 1980 a somewhat unexpected source of astronomical CCDs appeared. Craig Mackay of the Institute for Astronomy in Cambridge, U.K. had been working on silicon vidicons. Progress was slow due to lack of funds. He had met Jim Westphal on Palomar Mountain in 1975 and was aware of the TI work on charge-coupled devices, but he learned of a British source of CCDs by a curious coincidence. Silicon vidicons had good spectral response, but they were "noisy". Craig needed a very low noise amplifier design. He contacted an eminent designer named Ken Kandiah at the British Atomic Energy Authority at Harwell and asked him to visit Cambridge. Kandiah offered a design based on a Junction Field Effect Transistor (JFET) but recommended Craig to David Burt at the GEC Hirst Research Centre in Chelmsford, U.K. for a design based on the more readily available metal oxide semiconductor (MOS) transistors. When Craig met David Burt he learned that GEC had a very advanced CCD program. The following year, Craig and his then PhD student Jonathan Wright, put together a CCD drive system based on an existing vidicon camera. The noise associated with typical GEC CCDs was reported as 7 electrons in March 1982 while selected devices gave a mere 3 electrons. Remembering Jim Gunn's excitement on hearing that devices with better than 100 electrons noise were possible, then 3 electrons was a truly amazing result.

By June 1981, the date of the Harvard–Smithsonian conference on solid-state imagers, the number of independent astronomy groups working on CCD systems had already grown from 5 to 20. Devices in use came exclusively from TI, RCA, and GEC (U.K.). Astronomers were clearly pushing the technology as hard as they could in a direction that was good for scientific imaging, yet with only three manufacturers one of which had low-noise devices (GEC), one of which had high quantum efficiency devices (RCA), and the other of which should have had devices with both properties, but was (a) having problems with blue sensitivity and (b) not available for sale anyway, it was understandable that people began to worry. When production of the GEC devices moved to English Electric Valve (now *e2v*) there was the inevitable hiccup in supply, and when RCA withdrew from the field it seemed like the dream had become a nightmare. Eventually, TI CCDs "excess to NASA requirements" started to become available in the U.S.A. Exceptionally detailed studies of the TI chips by Jim Janesick and colleagues at JPL advanced the understanding of CCDs and their optimization, and new devices by companies such as Thomson–CSF (Europe) had been studied in detail by a team at the Royal Greenwich Observatory (RGO) including Paul Jorden who later joined *e2v technologies*. In 1985 astronomers learned of a most exciting prospect. It was the formation of a team at Tektronix Inc. led by Morley Blouke to produce scientific-grade CCDs with large formats and outstanding performance. The initial goal would be a 512×512 array with good-sized pixels (0.027 mm) leading to a chip with $2{,}048 \times 2{,}048$ pixels. Unfortunately, by mid-1986 it became clear that some sort of unexpected fabrication or processing problem was resulting in large numbers of defects called "pockets", thus rendering

otherwise excellent low-noise devices unusable and hopes were again dashed. This time the situation for many astronomical groups was serious because new instrument designs and funding for instrument developments had been tied to the expected Tektronix chips.

Morley and his team did not give up and, in collaboration with several interested parties, they valiantly followed every lead in an effort to understand, model, and eliminate such problems. The research at Tektronix and at JPL led to an in-depth understanding of the solid-state physics of CCDs which, as Morley remarks, "ought to be much easier to understand than a transistor." Around mid-1988 Tektronix began to ship CCDs to customers with long-standing orders. Later, the Tektronix CCD group was spun off into a company called Silicon Imaging Technologies, Inc. (SITe) which, for example, supplied all the CCDs for the highly successful Sloan Digital Sky Survey.

Among the many people frustrated by the dry spell in CCD supplies during that era was Richard Aikens, founder and president of Photometrics Ltd. In an unprecedented move he contracted with a so-called "silicon foundry" (a division of Ford Aerospace later taken over by Loral Corporation) to produce a custom CCD with 516×516 pixels (marketed as 512×512 pixels), and this turned out to be an outstanding success. The advantage of this approach is that the silicon foundry can quote for device production without having to consider the "end use" product. By the early 1990s Dick Bredthauer and his team at Ford Aerospace (later Loral and then Lockheed-Martin) had made a $4{,}000 \times 4{,}000$ CCD with 15-micron pixels; Dick is now president of Semiconductor Technology Associates (STA). In addition, Photometrics announced the availability of a chemical phosphor coating called Metachrome II which can be applied safely to any CCD by vacuum sublimation, and thereby improve its response to blue light. In August of 1988 Lloyd Robinson reported excellent initial experimental results for another brand-new device. As the result of a National Science Foundation grant to Lick Observatory, a contract was placed with EG&G Reticon Corporation to construct a large CCD suitable for spectroscopic applications; the format chosen was $400 \times 1{,}200$. Finally, in the same year, new initiatives at EEV (later called Marconi and now called *e2v technologies*) in the U.K. and at Thomson–CSF in France were announced. A thinning program and a mosaic construction program had begun at EEV. Meanwhile, funded by the European Southern Observatory and the French agency INSU, Thomson-CSF in Grenoble had developed a "buttable" version of their excellent low-noise front-illuminated device, and a 2×2 mosaic had been constructed at the European Southern Observatory near Munich.

These approaches set a trend that has continued until the present day. Astronomers now work directly with a silicon foundry to obtain customized CCDs. Companies like *e2v* in the U.K. have become major suppliers to astronomical facilities, providing chips for most of the large-mosaic cameras. Other manufacturers such as Kodak cater to a mass market and provide CCDs for companies that manufacture complete CCD camera systems. In addition, some U.S. government-funded labs, such as Barry Burke's group at the MIT/Lincoln Labs (Lexington, MA) where the orthogonal transfer devices have been developed or the group at the Lawrence

Berkeley Lab (Berkeley, CA)) working on deep-depletion CCDs, can provide special devices. Most astronomical developments concentrate on forming mosaics of high-yield formats like the $2{,}048 \times 4{,}096$ chips, and optimizing the response at both long and short wavelengths. The largest single scientific CCD manufactured so far is a $9{,}216 \times 9{,}216$ device from Fairchild Imaging (Milpitas, CA).

7.2 BASIC PRINCIPLES OF CCDs

7.2.1 Charge storage

A CCD is essentially an array or grid (Figure 7.3) of numerous individual picture elements (pixels), each one of which can absorb photons of light and utilize the energy to release an electron within the semiconductor. If we are intent on making an imaging device, then we do not want photon-generated electrons to migrate away from the site of impact of the original photons. To confine the electron within a pixel requires a special electrostatic field to attract the charged electron to a specific spot. What happens to the next photon? Clearly we need to create a storage region capable of holding many charges. This can be done by applying metal electrodes to the semiconductor silicon together with a thin ($100\,\text{nm} = 0.1\,\mu\text{m}$) separation layer made from silicon dioxide, which is an electrical insulator. The resulting structure behaves

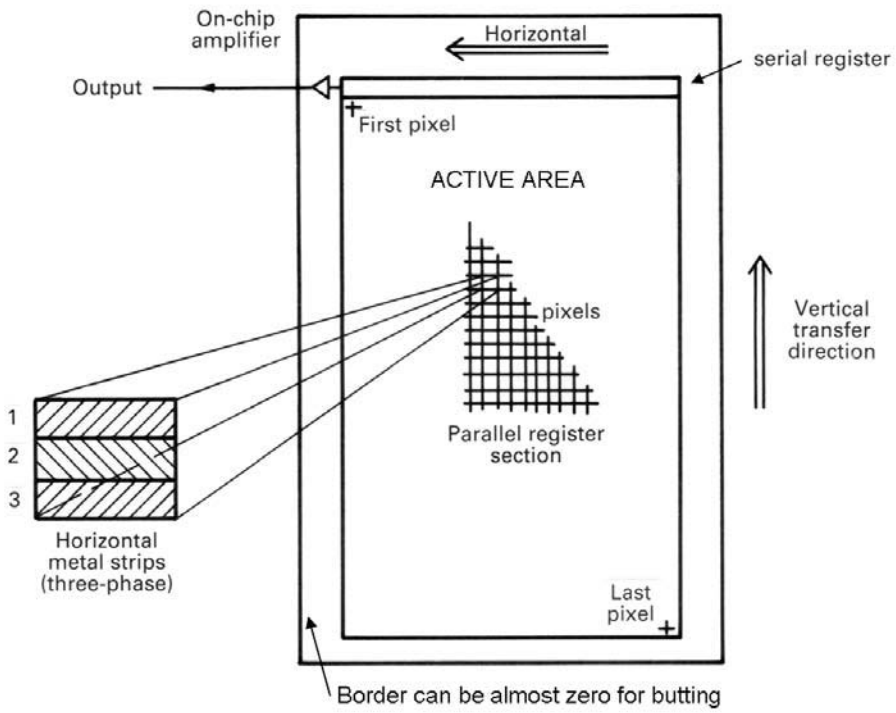


Figure 7.3. The general layout of a CCD as a grid of pixels on a slab of silicon.

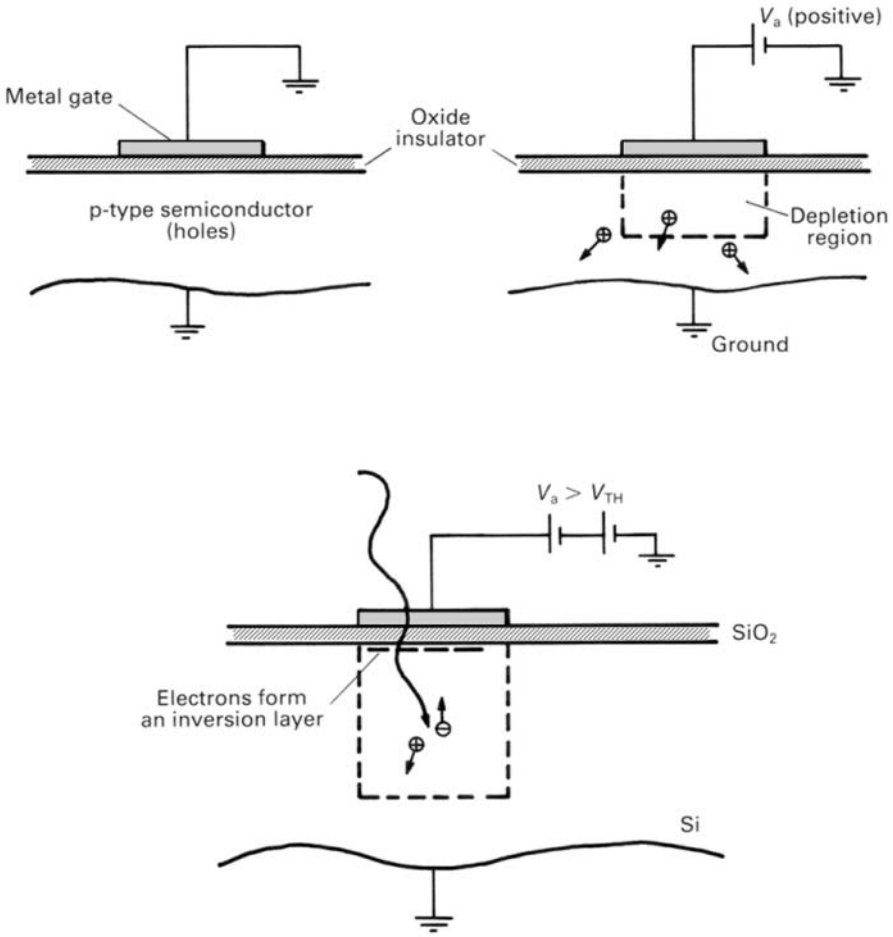


Figure 7.4. The development of a single metal oxide semiconductor (MOS) storage well, the basic element in a CCD, for different applied gate voltages.

like a parallel-plate capacitor which can therefore store electrical charge. It is called an MOS (metal oxide semiconductor) structure. An electric field is generated inside the silicon slab by the voltage applied to the metal electrode. If the material is p-type (the usual case) then a positive voltage on the metal gate will repel the holes which are in the majority and sweep out a region depleted of charge carriers. These conditions are illustrated in Figure 7.4. When a photon is absorbed in this region it produces an electron–hole pair, but the hole is driven out of the depletion region, and the electron is attracted towards the positively charged electrode. The MOS capacitor is the combination of two parallel-plate capacitors (namely, the oxide capacitor and the silicon depletion region capacitor), and therefore the capacitance is proportional to the area of the plates (electrodes) and inversely proportional to their separation. As the voltage on the plate can be controlled, then the depletion width can be increased

or decreased, and so the capacity to store charge can also be controlled. The depletion region shown in Figure 7.4 is an electrostatic "potential well" or "bucket" into which many photo-generated charges can be collected. Typically, the number of electrons stored is just $Q = CV/e$, where e is the charge on the electron ($1.6 \times 10^{-19} C$), V is the effective voltage, and the capacitance C is given by the "parallel-plate" formula $C = A\kappa\varepsilon_0/d$ in which A is the area of the pixel or gate electrode, d is the thickness of the region, κ is the dielectric constant of the SiO_2 insulator ($\sim$3.9), and ε_0 is the permittivity of free space (8.85×10^{-12} farad/m). As the voltage on the electrode increases, the "depth" of the well increases; other ways are needed to create sidewalls to the well. Eventually, at a certain "threshold" voltage, even the minority charge carriers, electrons for a p-type semiconductor, will be drawn to the electrode where they form an inversion layer.

Following Janesick (2001) we can provide a more quantitative description. A p-MOS capacitor consists typically of p-type (boron-doped) silicon (Si), a thermally grown layer of silicon dioxide (SiO_2) about 100 nm thick to act as a dielectric insulator and a conductive (metallic) gate usually made of deposited poly-silicon (silicon with randomly oriented crystal grains). In the p-type material, holes are the majority carriers. If a negative voltage is applied to the gate while the silicon substrate is at ground potential, then a highly conductive layer of holes will accumulate in a few nanoseconds at the Si–SiO_2 interface. This is called accumulation mode. If d is the thickness of the oxide insulator and $\varepsilon_{ox} = \kappa\varepsilon_0$ is the permittivity of SiO_2 (3.45×10^{-11} F/m), then the capacitance per unit area (F/m^2) of the oxide is just $C_{ox} = \varepsilon_{ox}/d$. Assuming $d = 100$ nm, then $C_{ox} = 3.45 \times 10^{-4}$ F/m^2. Sometimes there are two insulating layers, one of SiO_2 and the other of silicon nitride ($\varepsilon_{nit} = 6.63 \times 10^{-11}$ F/m), and so these capacitors must add in series so that $C_T = C_{ox}C_{nit}/(C_{ox} + C_{nit})$. Now, if a positive voltage is applied to the gate, holes are driven away from the surface leaving behind negatively charged boron ions, and thus creating a depletion region devoid of mobile charge carriers. The number of holes driven away in depletion mode equals the number of positive charges on the gate electrode, thus $Q_i = eN_A x_d$, where Q_i is the ionized acceptor charge concentration (C/m^2) beneath the depleted gate and x_d is the depth (m) of the depletion region. N_A is the concentration of boron (acceptors) in atoms/m^3 and e is the numerical value of the charge on the electron. As the depletion region is non-conductive it acts like an insulator with a capacitance per unit area of $C_{dep} = \varepsilon_{Si}/x_d$ where $\varepsilon_{Si} = 1.04 \times 10^{-10}$ F/m; the dielectric constant of silicon is $\sim$11.7. Thus, the net gate capacitance in depletion mode is the series combination of C_{ox} and C_{dep}, and because C_{dep} is the smaller capacitance then it dominates the series combination.

Gate voltage is constant throughout its thickness because it is a conductor. There is a voltage drop across the oxide layer and then the voltage in the depleted p-type silicon will depend on the charge distribution, but it must eventually drop to the ground potential of the substrate (i.e., $V = E = 0$ at $x = x_d$). The variation of voltage (V) with depth (x) is determined by Poisson's equation:

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\varepsilon_{Si}} \tag{7.1}$$

where ρ is the charge density. The origin ($x = 0$) is taken as the Si–SiO_2 interface. In principle ρ is given by $e[p + n + N_A + N_D]$, where p is the number density of free holes, n the number density of free electrons, N_A the number density of localized fixed ionized acceptors (atoms/m^3), and N_D the number density of fixed ionized donors (atoms/m^3). Most free carriers in the depletion region are swept away by the electric field, and thus for a p-channel device we expect $\rho = -eN_A$, where the sign is negative due to the absence of holes. Therefore, $d^2V/dx^2 = eN_A/\varepsilon_{Si}$. Integrating Equation (7.1) with this expression for ρ and applying the condition that $dV/dx = 0$ at $x = x_d$ gives

$$\frac{dV}{dx} \equiv -E_x = \frac{eN_A}{\varepsilon_{Si}}(x - x_d) \tag{7.2}$$

Integrating again using the fact that $V = 0$ when $x = x_d$ gives

$$V = \frac{eN_A}{2\varepsilon_{Si}}(x - x_d)^2 \tag{7.3}$$

Equation 7.3 implies that the most positive voltage relative to the substrate occurs at the Si–SiO_2 interface where $x = 0$ and that the surface voltage V_S is

$$V_S = \frac{eN_A}{2\varepsilon_{Si}}(x_d)^2 \tag{7.4}$$

and the electric field at the surface is

$$E_S = \frac{eN_A}{\varepsilon_{Si}}x_d \tag{7.5}$$

Note that the electric field is zero in the metallic gate, constant across the oxide (dielectric layer) at V_{ox}/d, and decreases linearly to zero at $x = x_d$ inside the depletion region. The potential is constant across the gate, drops by the amount $V_{ox} = E_S d$ across the oxide layer, and then decays quadratically from the value V_S to $V = 0$ at $x = x_d$. Electrons generated by the internal photoelectric effect in the depletion region will be attracted to the most positive potential at the Si–SiO_2 interface. The MOS capacitor in depletion mode is the basic element of a surface channel CCD. If V_G is the voltage applied to the gate then

$$V_G = V_{ox} + V_s = E_S d + V_S \tag{7.6}$$

Therefore,

$$V_G = \left(\frac{eN_A x_d}{\varepsilon_{Si}}\right)d + \left(\frac{eN_A}{2\varepsilon_{Si}}\right)x_d^2 \tag{7.7}$$

As photo-generated electrons collect at the surface, the number of ionized acceptor atoms decreases and the depletion region becomes slightly smaller. For a fixed gate potential of V_G the effective voltage drop V_Q at the gate caused by the collected charge is given by

$$V_Q = V_G - (eN_e/C_{ox}) \tag{7.8}$$

where N_e is the signal charge measured in electrons.

The charge storage capacity of the MOS capacitor is the amount of charge required to bring the surface potential back to zero volts. Because the gate and substrate potentials are fixed, signal electrons at the Si–SiO_2 interface are shared between the oxide and depletion capacitances, therefore the capacitors are in parallel. A change of surface potential of V_S due to a charge Q is then

$$\Delta V_S = -Q/(C_{\text{ox}} + C_{\text{dep}}) \tag{7.9}$$

Because C_{ox} usually dominates over C_{dep}, then $Q \sim C_{\text{ox}}\,\Delta V_S$, and therefore the "full-well" capacity (N_{FW}) of a MOS capacitor expressed in electrons is approximately given by

$$N_{\text{FW}} = C_{\text{ox}} V_S / e \tag{7.10}$$

Example. Using numbers from Janesick (2001) for a typical doping concentration of $N_A = 10^{15}$ atoms/cm^3 and a depletion depth $x_d = 1.5\,\mu$m for a 10 V gate voltage, the surface potential for an empty well is $V_S = 3.44$ V. Assuming $C_{\text{ox}} = 3.45 \times 10^{-8}$ F/cm^2, then the corresponding charge is $Q = 1.19 \times 10^{-7} C/\text{cm}^2$ and for an electrode area of 32 μm^2 the full-well charge is about 240,000 electrons.

This is the basic concept of the CCD MOS charge storage well. We will see in Section 7.3.3 that there are disadvantages to collecting charge at the Si–SiO_2 surface, and therefore a modified design is needed. For now, we focus on the next feature of the CCD: charge transfer.

7.2.2 Charge-coupling and clocking

It was Jerry Kristian and Morley Blouke who first pointed out that taking a picture with a CCD is a bit like measuring the rainfall over a rather large plantation! Let's suppose you distribute a large number of buckets in a rectangular pattern of rows and columns over a field. After it has stopped raining, measure the amount of water collected in each bucket by shifting the entire array of buckets towards a conveyor belt located at one end of the field, loading the buckets one whole row at a time onto the belt and then allowing the belt to convey each bucket from that row to a metering station where the amount of water in each is recorded. Then transfer another row onto the conveyor belt and keep repeating the sequence until all rows have been transferred to the metering station. By knowing the amount of water in each bucket the pattern of rainfall over the field can be generated. To visualize this pattern as a picture we convert the amount of rainfall in the bucket to a shade of grey (black for no rain and white for a full bucket) and mark this color on our picture at the original location of the bucket. If instead of rain we think of photons of light, then our buckets are actually the pixels of the CCD detector.

Two views of a simplified CCD construction called a "three-phase" structure are shown in Figures 7.5 and 7.6. Semiconductor silicon (in this case p-type) is covered with a thin electrical insulating layer of silicon oxide on top of which are placed three sets of metal electrode strips. One of the three strips is set to a more positive voltage than the other two, and so it is under this one that the depletion region or bucket

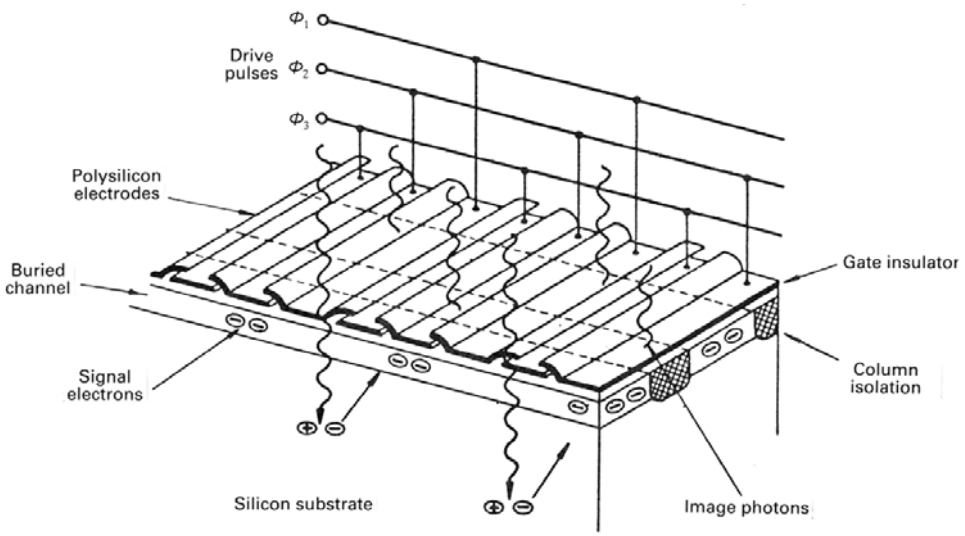


Figure 7.5. A perspective view showing the structure of a three-phase CCD. Credit: *e2v technologies*.

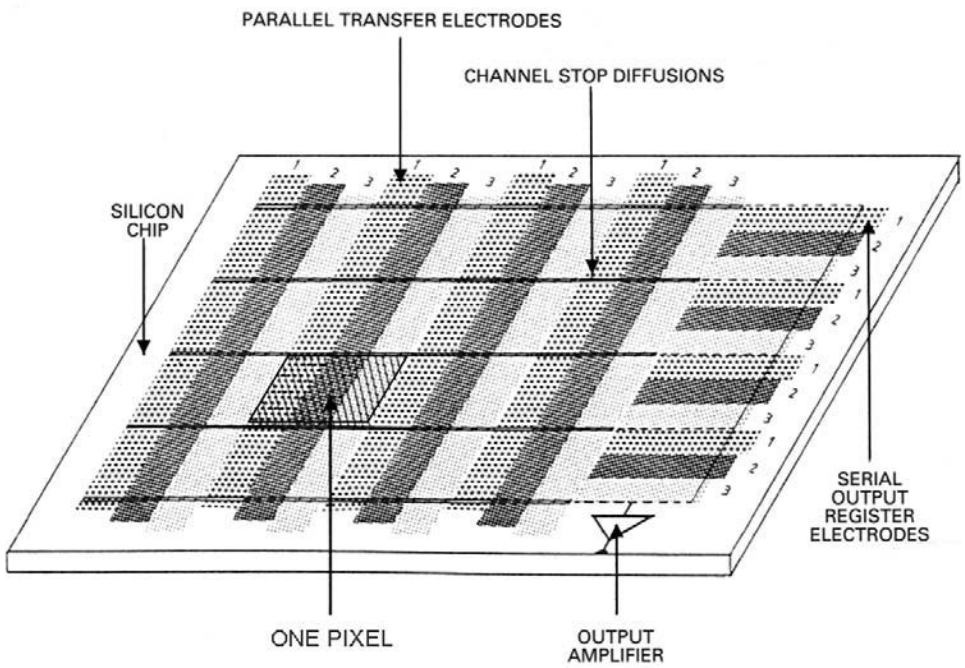


Figure 7.6. A schematic view from above showing the interface between the parallel and serial registers of a CCD. Credit: *e2v technologies*.

forms, and where the electrons accumulate. We have created two walls of the well. By heavily doping the silicon crystal structure with a certain impurity, it is possible to create a very narrow channel which totally obstructs any movement of charge along the length of the electrode. These "channel stops" and the triplet of electrodes define the pixel, with photon-generated charges being drawn to, and collected under the middle (most positive) electrode, like water in a well.

The unique feature of the CCD, which gives it its name, is the way in which the photo-generated charge, and hence the image of the scene, is extracted from the storage/detection site. It is called "charge-coupling". To transfer charge from under one electrode to the area below an adjacent electrode, raise the voltage on the adjacent electrode to the same value as the first one. This is like lowering the floor of the well. The charges will now flow, like water, and be shared between both buckets (refer to Figure 7.7). Transfer can be in either direction, and by connecting sets of electrodes together the entire charge stored on the two-dimensional imaging area can be moved simultaneously in that direction. When the voltage on the original electrode is reduced to zero, the transfer is complete because the collapse of the storage well pushes any remaining charges across to the new electrode. Because it takes three electrodes to define one pixel, three of the above transfers are required to move the two-dimensional charge pattern by one whole pixel step along the direction at right angles to the electrode strips. The process of raising and lowering the voltage can be repeated over and over and is known as clocking. These drive or clock pulses can be described in a diagram called a timing waveform (also shown in Figure 7.7) and are not difficult to produce electronically. As can be seen in Figure 7.6, there is another set of electrodes at right angles to the first to enable charges to be moved along that row.

7.3 CCD CONSTRUCTIONS

7.3.1 Interline and frame transfer CCDs

Most CCDs are manufactured for use in video cameras, and therefore the charge-coupling transfer must be done at very high speed (TV frame rates) with light falling continuously onto the CCD. To achieve this, several manufacturers adopt an approach called "interline" transfer. In this scheme charges are moved sideways at high speed by one pixel to be relocated in a pixel which is shielded from light by an extra overlying strip of opaque metal. The charges are then coupled lengthways to transfer down the shielded column to the output row of the CCD. Another approach is that of "frame" transfer. In a frame transfer CCD there are no blind spots in the imaging area, instead a duplicate imaging area, contiguous with the first, is covered by an opaque mask. The entire charge distribution representing the scene is shifted very rapidly lengthways until it disappears under this mask. It is then transferred lengthways once more, but at a slower rate, to be finally read out. The transfer rate for readout is a little faster than the permitted "dwell time" on the image scene,

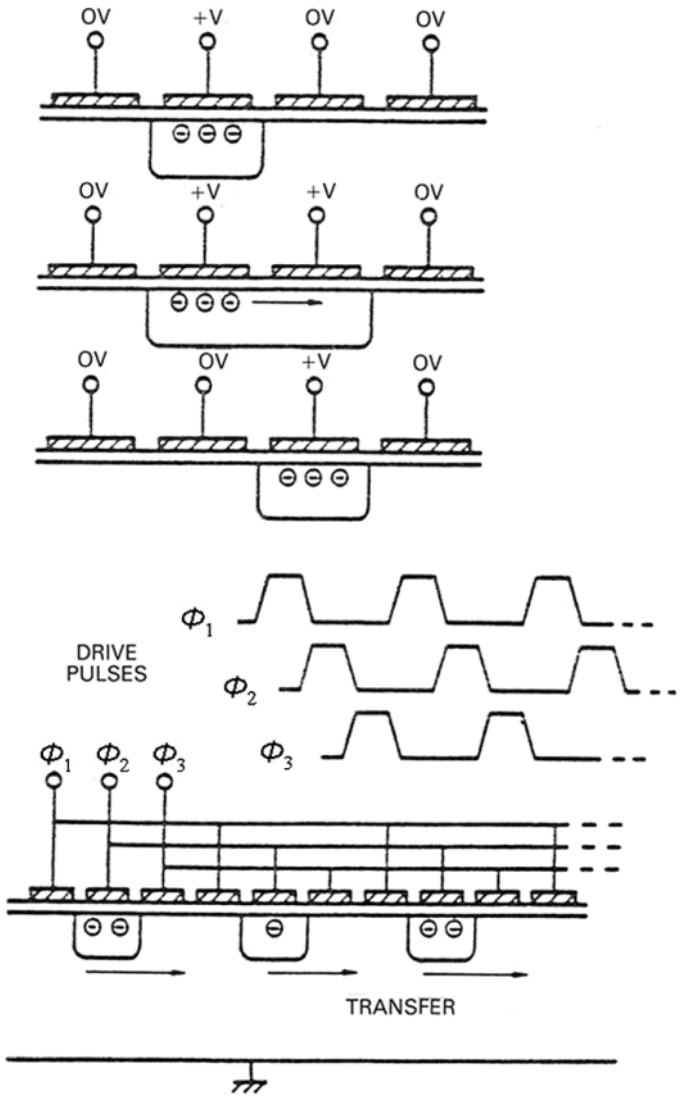


Figure 7.7. The basic charge-coupling principle in a three-phase CCD and the associated timing or clock pattern.

whereas the image-to-storage transfer rate is extremely high to minimize blurring due to the illumination which is still present.

Frame transfer CCDs are best suited to astronomical applications, especially as the mask covering the storage area can be removed to yield a CCD of twice the original size! Blurring of the optical scene during readout is not a problem in astronomical applications because a light-tight electronically controlled shutter can be used to block off the incoming illumination. However, as CCD mosaics have become

larger, mechanical shutters have become more challenging. Very large slit-type shutters ($\sim$480 mm $\times$ 480 mm for Pan-STARRS) have been developed by the Argelander Institute for Astronomy, University of Bonn, Germany; see web address at end of chapter.

7.3.2 CCD outputs

So far we have indicated that the charge distribution representing the image scene can be coupled or transferred lengthways by indefinitely sending clocking pulses to the three electrodes. In practice, of course, the transfer direction is terminated (as already shown in Figure 7.6) by a special row of CCD pixels called an "output register" which has electrodes arranged at right angles to the main imaging area of the CCD, like the collection conveyor belt in the rainfall model, so that it can transfer charge horizontally rather than vertically. Since the output register is a single row it is usually called the serial register, whereas the main area of the CCD is called the parallel register. At the end of the output register is a single output amplifier. Figure 7.8 displays an electrical schematic of the circuit at the end of the serial register. A complete clocking sequence consists of the following:

(1) A vertical shift of the entire image scene by one pixel. This delivers a row of charge to the output register.
(2) A horizontal shift through all the pixels in the output register. This delivers each charge in that row to the output amplifier, one pixel at a time, where the charge can be detected and converted to a voltage which can be measured and recorded.

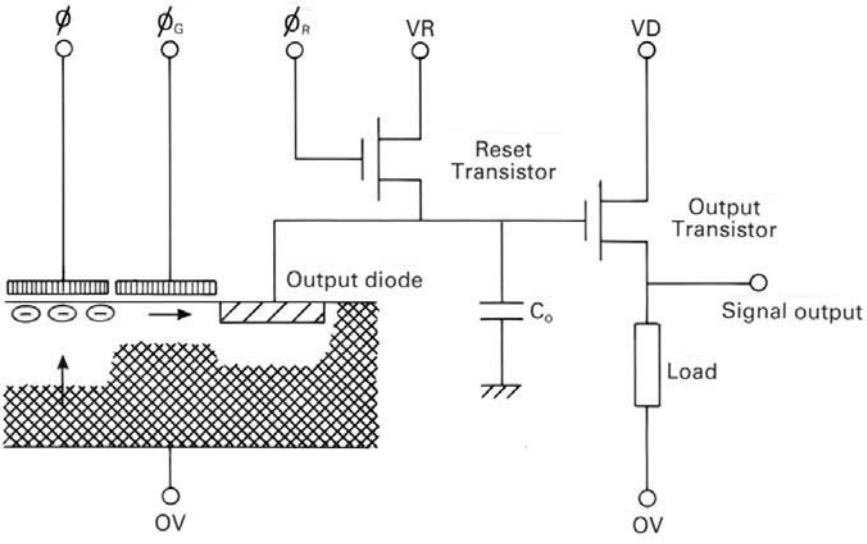


Figure 7.8. Output stage of a CCD showing how charge is extracted in pulses by reading and resetting the output node periodically.

(3) Another vertical transfer to deliver the next row in the image to the serial register.
(4) Another horizontal transfer to move all the charges in the new row to the output.

This process is repeated until all the rows in the CCD have been delivered to the output register and out to the output amplifier.

In general, the pixels (electrodes) associated with the output register are larger than those in the imaging area to ensure that they have more storage capacity and are therefore much less prone to saturation. We will discuss the effect of saturation later. Referring to Figure 7.8, a packet of electrons with charge Q is allowed through the final output gate onto an effective storage capacitance C (essentially a reverse-biased diode) which causes an instantaneous change $V = Q/C$ in the voltage of the input line of the on-chip transistor used as a source follower, which in turn yields a voltage change at the output line. Before the next pulse arrives the storage capacitor is recharged. The readout is destructive (i.e., it can only be done once), but this is of little consequence. Unwanted electrical noise due to the periodic resetting of the output diode can feed through to the output transistor and degrade the sensitivity of the CCD to faint signals, but fortunately this type of "reset noise" can be totally eliminated by subsequent electronic signal processing (see Chapter 8).

7.3.3 Buried-channel CCDs

While it is true that the simple CCD structures described so far can be constructed (i.e., metal electrode strips deposited directly onto an insulating layer on the top of a slice of uniformly doped silicon), they are far from ideal. In these CCDs the most positive potential lies at the silicon surface immediately under the insulating oxide layer; electrons are therefore stored and transferred at the surface of the silicon semiconductor and hence they are called "surface-channel" CCDs. Unfortunately, the surface layer has many crystal irregularities and defects in the crystal lattice which can readily "trap" charge, but not so readily release it! The result is very poor charge-coupling and severe image smear.

To avoid the surface-trapping phenomenon it is advantageous to "grow" another layer of silicon onto the existing p-type substrate to separate it from the insulating layer. If this is a more highly doped n-type layer (typically phosphorus-doped) then a more complex depletion region is created with a potential minimum (or collecting layer) which is substantially "buried" inside the bulk silicon. This is the "buried-channel" CCD (or BCCD). All modern CCDs are of this construction.

The unit cell of a BCCD is shown schematically in Figure 7.9, and the distribution of potential is given in Figure 7.10. Recall the discussion on p. 193 (Figure 5.17) of the pn junction in which the most positive potential occurred on the n-side but within the depletion region. All that has been added on the n-type layer is an insulating oxide and a metal control gate. The most positive potential still lies well away from any surface. If a voltage more positive than any of the gate voltages is applied to this n-type layer, a depletion region will form at the pn junction. An additional depletion under the gate is generated by a difference in gate potentials, and the exact depth of the buried channel can therefore be controlled by the electrode

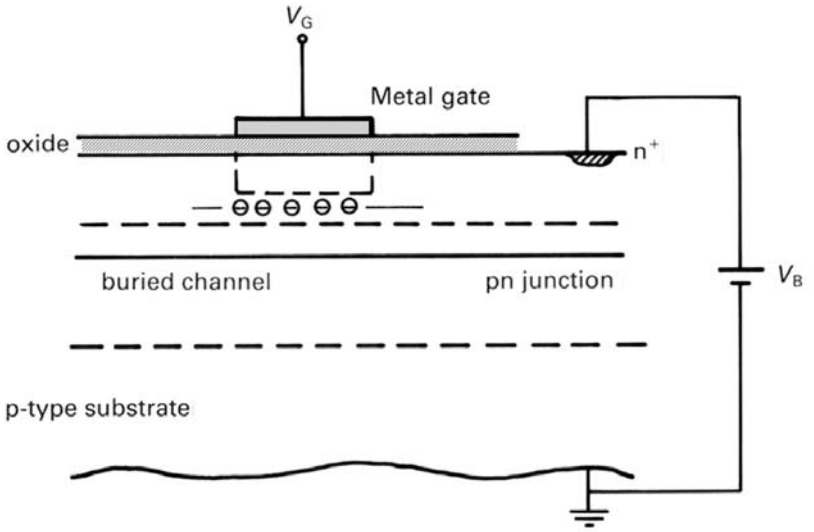


Figure 7.9. A single storage site in the buried channel CCD. The charge collection layer lies well below the surface.

potentials as before. It is important that the n-type region be completely depleted of majority carriers (electrons) in order to distinguish electrons created by absorption of photons. If the applied gate voltage is negative relative to the n-channel potential then we have an n-MOS capacitor in which the majority carriers (electrons) are repelled away from the surface. Charge transfer within the bulk silicon is very efficient because the number of trapping sites is considerably fewer and they are much less noisy. The

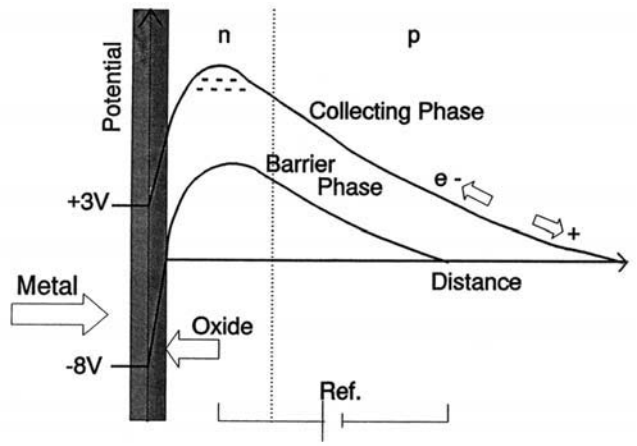


Figure 7.10. The distribution of voltage for a buried channel CCD pixel in two cases. Credit: Jim Janesick.

analogy with buckets of water still holds, and charges stored at the embedded collection layer can still be charge-coupled from one pixel to the next by pulsing or clocking the metal electrodes to remove/replace the adjacent potential barriers. A penalty is paid in terms of a somewhat reduced storage capacity, and a buried-channel CCD will saturate before an equivalent surface-channel CCD. Incidentally, similar benefits are accrued if the output amplifier is also constructed in buried-channel form. A full analysis of the buried channel construction using Poisson's equation can be used to derive the location of the potential minimum (see, for example, Janesick, 2001; Rieke, 2003). Briefly, in the fully depleted n-type layer we assume $\rho = eN_D$ positive charges and in the p-type layer we have $\rho = -eN_A$ negative charges as before. Letting t represent the depth of the n-type layer then the location of the potential maximum relative to the pn junction is given by

$$x_n = t - x_p(N_A/N_D) \tag{7.11}$$

where x_p is the depletion depth of the p-type region; and the variation of potential (V_2) in the n-type region from $x = 0$ to $x = t$ is given by

$$V_2 = V_{\max} - \frac{eN_D}{2\varepsilon_{\mathrm{Si}}}(x - x_n)^2 \tag{7.12}$$

where $V_{\max}$ is the channel's maximum potential which is determined by the junction potential and the ratio of acceptor-to-donor concentrations. The effective capacitance per unit area for an empty potential well is the series combination of the oxide and depletion capacitance of width $t - x_n$; that is,

$$(C_{\mathrm{eff}})^{-1} = (d/\varepsilon_{\mathrm{ox}}) + (1/\varepsilon_{\mathrm{Si}})[t - x_n] \tag{7.13}$$

When electrons are generated by photon absorption they will move to the potential maximum and remain there, which causes a portion of the storage well to become un-depleted. The effective capacitance per unit area becomes

$$(C_{\mathrm{eff}})^{-1} = (d/\varepsilon_{\mathrm{ox}}) + (1/\varepsilon_{\mathrm{Si}})[t - x_n - (Q/2N_D)] \tag{7.14}$$

where Q is the electron surface density (electrons/cm^2 assuming N_D in atoms/cm^2). The interested reader is referred to Janesick (2001) for more details. With typical values for the parameters the full-well capacity is lower than the surface-channel CCD, but charge transfer efficiency is greatly improved.

In summary, to construct a practical high performance CCD one starts with the buried channel, a region where electrons are confined in depth (z) below the gate but still free to move from side to side (in x and y); this is usually n-type on a p-type substrate. Next, narrow columns of heavily doped n-type material are diffused into the normal n-type region to produce channel stops which subdivide the x-direction and constrain electrons to move only in the y-direction along columns. A serial register is formed by leaving a gap at one end between the endpoints of all the channel stops and the edge of the CCD. Electrons emerging into this narrow strip could move in x once more. Finally a silicon dioxide insulation layer is grown on top, and the basic three-phase electrode gate structure is added as rows or strips across the

entire parallel area to define subdivisions (pixels) in the y-direction; the serial register receives similar treatment but the gates are at right angles to the parallel section.

7.3.4 Two-phase, four-phase, and virtual-phase CCDs

There are alternative structures to the three-phase CCD outlined above. If bi-directional charge motion is not required then the CCD can be doped along the direction of vertical transfer in such a way as to provide a barrier to backward-flowing charges. This is a two-phase CCD. The principle is illustrated in the upper panel of Figure 7.11. An "implant" is diffused into one-half of the substrate below each of the two electrodes. The presence of this layer affects the depth of the depletion region immediately beneath it in such a way that the depletion is always greater under the implant. With one electrode at a low voltage and the other at some "fixed" voltage level, the result is a staircase-shaped well. When the low electrode is raised to a higher voltage than the fixed electrode, charges are forced to move left to right as shown in the figure.

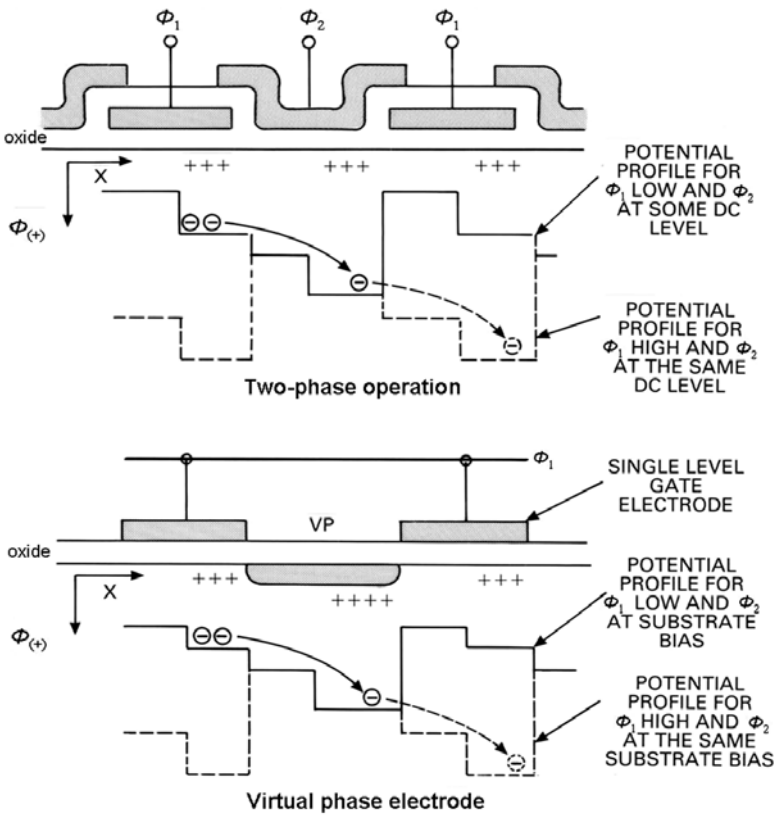


Figure 7.11. The operation of a two-phase CCD (top) and the Texas Instruments virtual-phase CCD (bottom). Credit: Jim Janesick.

Texas Instruments Corp. developed the first single-clock CCD known as the "virtual-phase" CCD. The principle of this device is very similar to a two-phase CCD as shown in the lower panel of Figure 7.11. Consider again the two-phase CCD. One electrode was left at a constant voltage to produce an intermediate depletion region. There is really no necessity to do this with a semi-opaque overlying metal gate as the same effect can be achieved by diffusion directly into the silicon substrate. Only one electrode need be physically present and the other half of the pixel is left clear and uncovered except for its implant. This is the "virtual" electrode or virtual phase. As in the two-phase device, the roles of other electrodes are taken by a series of potential steps created by different levels of surface doping. These potential steps direct the flow of charge as the voltage on the single electrode is driven above and below the potentials of the uncovered or virtual electrode. Apart from a considerable simplification in the number of clocks, this kind of CCD structure was chosen to permit an improved response to blue light by minimizing the amount of absorption due to the polysilicon (conducting) electrodes when the chip is frontside-illuminated. These CCDs were used on the Galileo Mission to Jupiter because with implants instead of polysilicon electrodes there would be less likelihood of gate shorts caused by ionizing radiation. A few CCDs have also been made with four physical electrodes per pixel which would correspond to the four potential steps apparent in the two-phase and virtual-phase devices. This complexity gives complete control over all combinations of phases and can be used for special applications involving two alternating image scenes, so-called "chopped" signals, or to increase charge storage capacity.

7.3.5 Backside-illuminated CCDs

So far we have been describing CCDs in which the illumination or photon flux was compelled to pass through the overlying electrode structure in order to reach the depletion (or storage) region in the silicon. In practice, this simple approach results in severe absorption of blue light in the electrodes. Virtual-phase CCDs are one approach to solving this problem. Alternatively, the CCD can be turned over and illuminated from the backside! Before this becomes effective, however, the thick silicon substrate must be reduced in thickness (either mechanically or chemically) to only 10 μm or so. Figure 7.12 compares both constructions. Such thinned, backside-illuminated CCDs have excellent response in the blue and violet. If thinned too much they lose their red response because the red photons need more absorption length and if this is not there, they will pass right through the silicon!

One disadvantage of the thinned CCDs is that they are more mechanically fragile and prone to warping. Some thinned CCDs are mounted to a supporting substrate. Also, interference "fringing" can occur due to multiple reflections internal to the CCD substrate or between the silicon and the supporting substrate in a manner similar to fringes which can be seen when two flat glass plates are placed almost in contact at a small angle. For imaging, this problem can be eliminated by computer processing and calibration techniques discussed in more detail in Section 9.4, but it is a serious drawback for spectroscopy.

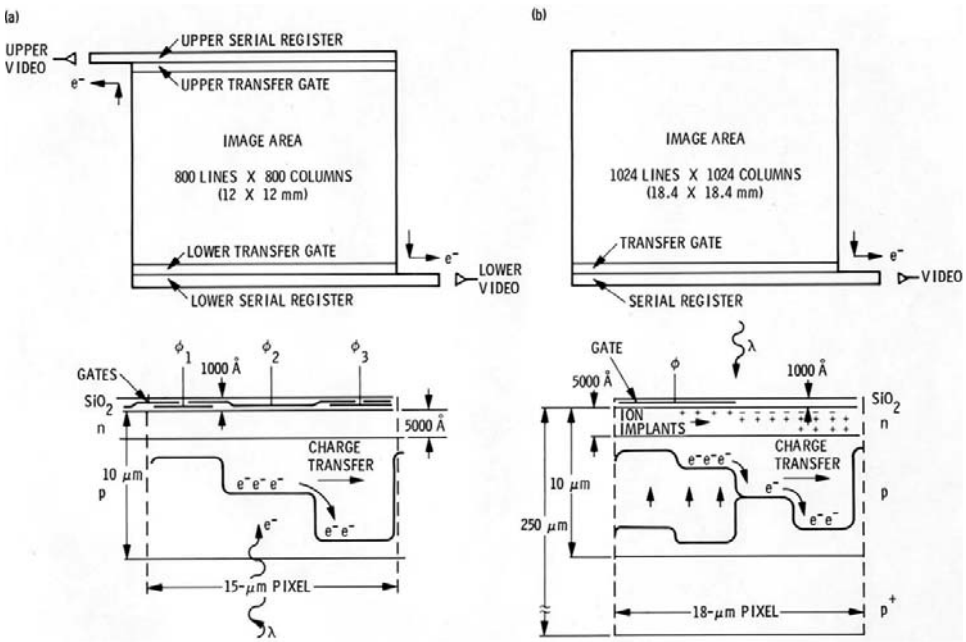


Figure 7.12. The structure and dimensions of a thinned, backside-illuminated three-phase CCD and a thick, front-illuminated virtual-phase device are compared. Credit: Jim Janesick.

7.3.6 Multi-pinned phase CCDs

Several years ago it was realized that the dominant contributor to the dark current in CCDs was thermal generation due to surface "states" at the Si–SiO_2 interface. In fact, surface dark current is 10^2–10^3 greater than dark current generated by the bulk of the CCD. Two factors control the dark current at the silicon–silicon dioxide "interface": the density of interface states and the density of free carriers (holes or electrons) that populate the interface. Electrons that thermally "hop" from the valence band to an interface state and then to the conduction band will produce an electron–hole pair that will be collected in the potential well of the CCD pixel. Free carriers could "fill" the interface states and inhibit the hopping mechanism and hence drastically reduce the dark current, but standard operation of CCDs does just the opposite. In a CCD where the voltage on the gate electrode is such that the surface potential at the Si–SiO_2 interface is greater than the silicon substrate potential, the resulting depletion drives away all the free carriers and therefore *maximizes* dark current from surface states. Dark current is now controlled solely by the density of interface states and is thus dependent on fabrication processes. If, however, the CCD can be operated in "inversion" mode then holes from the channel stop regions migrate to populate the interface states below the inverted gate and thereby eliminate surface dark current (by filling the hopping sites). If all three phases of a three-phase CCD are driven into inversion (i.e., all gate electrodes set very negative relative to the substrate) then there is no potential well in which to collect charges. Partial inversion

can be accomplished by biasing a collecting phase higher than the substrate (say, $\varphi_1 = 3$ V) and biasing the two barrier phases into inversion ($\varphi_2 = \varphi_3 = -8$ V). Inverting two barrier phases reduces the dark current generation by 2/3. Multi-pinned phase (MPP) CCDs are designed in a special way to allow operation in a totally inverted mode. Clearly, to obtain any charge storage capacity while totally inverted the potential of one or more phases must be offset from the others. For a three-phase CCD this can be done by doping the silicon beneath Phase 3 with boron. In effect, the boron implant cancels the effect of the phosphorus doping used in the buried channel and reduces the potential in the region. When biasing all phases into inversion, charge will collect under Phases 1 and 2, while Phase 3 will now act as the barrier phase. As the clocks are driven negative, Phase 3 will attain inversion before Phases 1 and 2. Three-phase MPP CCDs could also be fabricated by implanting more phosphorus under Phase 3, forcing charge to collect there and resulting in Phases 1 and 2 becoming the barriers. By offsetting the positive clock level of Phase 3 by the amount of the built-in MPP potential, the CCD can be operated only partially inverted (to gain back some well depth). The well depth of an MPP CCD is determined by the potential which is built in by the implant and can thus be 2–3 times lower than the partially inverted case.

Care must be exercised with CCDs using MPP architecture if the ultimate in dark current performance is required. If the CCD becomes saturated because of very substantial light overload or because the power to the CCD was turned off while the device was cold, then trapped charges will raise the dark current (by a factor of 2–3 typically) for hours until it gradually decays back to its normal low state. There are other pitfalls too. As explained above, when a CCD phase is inverted, holes from the channel stops migrate and collect beneath the inverted gate thereby pinning the surface-to-substrate potential. Unfortunately, some of these holes become trapped at the Si–SiO_2 interface, and when the clock is switched back to the non-inverted state to transfer charge these trapped holes are accelerated out of the interface with sufficient energy in some cases to cause impact ionization, which results in a "spurious charge". This is potentially serious in large arrays because spurious charge is produced during each pixel transfer and has a shot noise behavior which can overwhelm the on-chip amplifier noise and dominate the observed readout noise of the chip. For example, suppose on average that 1 electron of spurious charge is produced for 10-pixel transfers. After 1,024 transfers this process would have resulted in about 102 electrons and an associated noise of 10 electrons rms, whereas the on-chip amplifier might be capable of achieving 3–7 electrons rms of read noise. To overcome the shot noise produced by spurious charge production, Jim Janesick and other investigators suggest three strategies. First, slow down the rise time of the drive clocks (falling edges are not important in this context) by adding an RC network at the output of the clock driver board to allow the holes to return to the channel stops slowly. Limit the clock voltage swing to the smallest value possible consistent with good charge transfer efficiency and hence reduce the driving electric fields and the acceleration of the charges; less spurious charge will be generated. Finally, a "tri-state" clocking scheme can be tried in which an intermediate clock level is established which is just above the inverted state but not enough for complete charge transfer. A slow

transition from the inverted phase to the intermediate phase followed later by the complete swing allows the trapped holes to be released slowly from the oxide. This approach was employed by the JPL team on the 800×800-pixel virtual-phase CCDs on the Galileo probe to Jupiter.

The vertical registers of an MPP CCD are always inverted to produce low dark current, and the suppression of spurious charge in these registers is helped by the high capacitive load to the imaging pixels which slows down the clock edges. On the other hand, the horizontal register is clocked much more rapidly and would produce significant amounts of charge if it was being continuously brought in and out of inversion. Consequently, the horizontal registers of an MPP CCD do not receive the MPP implant. Unfortunately, this is not the case for virtual-phase (VP) CCDs which require that the horizontal register be switched in and out of inversion. Spurious charge produced by this effect actually limited the system noise on the Galileo chips to three times the intrinsic noise of the on-chip amplifier, even with tri-state clocking. Recall that the big advantage of VPCCDs, apart from simplicity of clocking, is the fact that they have significant blue response for a front-illuminated chip because of the absence of a polysilicon gate over the virtual phase.

MPP technology offers benefits other than dark current. For example, surface residual image charge and pixel non-uniformity are improved, and "anti-blooming" features can be employed. Pinning also increases the CCD's tolerance to high-energy radiation and allows the CCD to be erased rapidly because trapped charges are neutralized by the preponderance of holes in the inverted phase. The anti-blooming mode works as follows. During integration, suppose φ_3 remains inverted at all times to create the barrier phase. Photo-generated charges therefore build up under φ_1 and φ_2. Phases 1 and 2, however, are slowly *switched* between the inverted state and just above the optimum full-well voltage (typically $-8\,\mathrm{V}$ to $+5\,\mathrm{V}$ according to Jim Janesick for Loral, Lincoln Labs, and SITe CCDs). The required switching rate obviously depends on the rate of charge build-up, but about 50 Hz is typical. What happens is that one of the two phases reaches saturation and charges begin to get trapped in the Si–SiO_2 interface. At this point the phases are switched. Holes generated by the phase going into inversion neutralize the trapped electrons. Simultaneously, charge builds up under the other phase starting at surface full well (because of the transfer) and again electrons begin to enter the oxide and get trapped. Phases 1 and 2 are switched back, and the inversion eliminates the trapped charge under the second electrode. This process of back-and-forth switching continues during the entire integration period and inhibits the saturating pixel from blooming and bleeding charge up and down the column.

7.4 ASTRONOMICAL REQUIREMENTS

7.4.1 Slow-scanning, cooling, and optimization

For astronomy applications, CCDs cannot be used like normal TV cameras taking frames every 1/30th of a second. Instead, the CCD must be used as if it were a

photographic emulsion in a camera, in other words we need to take long exposures to build up a charge image from a faint source. Even when the charge image is removed during the readout process, we cannot do that rapidly either, because the charge transfer efficiency will be impaired and the electronic noise is greater at higher readout rates. In fact, compared with TV frame rates, astronomical CCDs must be read out very slowly and hence this mode is called "slow scan". Typically, rates of about 50,000 pixels per second (50 kHz) are used which implies that it will take 20 seconds to read out an array with one million pixels. Clearly, the CCD must not be exposed to light during this time period otherwise there would be considerable smear. Although the readout time may seem long, remember that the exposures are typically much longer still, perhaps 1,000 seconds or more.

With such long exposures, however, the second problem arises: dark current. To permit long exposures, astronomical CCDs must be cooled to temperatures well below the freezing point of water, which implies using a vacuum chamber to avoid frosting. The advent of MPP CCDs has helped considerably in reducing the cooling requirements, but for the most stringent applications in astronomical spectrographs, more cooling is required and most CCD cameras at professional observatories use modified liquid-nitrogen cooling systems.

7.4.2 CCD mosaics

To obtain cameras with much larger numbers of pixels the general approach has been to construct "mosaics" of CCDs. As shown in Figure 7.13, CCDs that have been specially manufactured to ensure that three sides have no connections, and can therefore be butted very closely to one another, enable one to construct a larger mosaic. In this case, eight $2\,\mathrm{K} \times 4\,\mathrm{K}$ chips were used to create a mosaic of $8\,\mathrm{K} \times 8\,\mathrm{K}$. Even on the fourth side, connection space is kept as small as possible. The first operational wide-field CCD camera was the MegaCam instrument on the Canada–France–Hawaii Telescope (CFHT) which started science operations in 2002. Using 40 $2\,\mathrm{K} \times 4\,\mathrm{K}$ CCDs from *e2v*, this camera can image a field of view of 0.9 square degrees on the sky (Figure 7.14). Many other wide-field cameras are now in operation and more are planned. A useful metric to gauge the efficiency of a large-area survey is the product $\Omega D^2 q$, where Ω is the solid angle of the field of view, D is the telescope diameter, and q is the total throughput/quantum efficiency of the instrument assuming that the seeing disk is resolved by the sampling. The time required to complete a survey to a given limiting magnitude is inversely proportional to this figure of merit; the larger this $A\Omega$ product the shorter the survey time. At the time of writing the largest anticipated mosaic is the 3-gigapixel camera for the proposed Large Synoptic Survey Telescope, while the largest operational array is one of a planned set of four 1.4-gigapixel cameras for Pan-STARRS (Figure 7.14; see also Plates 15 and 16).

Construction of multi-CCD mosaics is not trivial and considerable effort has gone into both the packaging of the CCD itself and the design of the camera head in order to make these systems practical. For example, in some wide-field imagers there is sufficient curvature of the focal surface that either the CCDs must be laid out on a

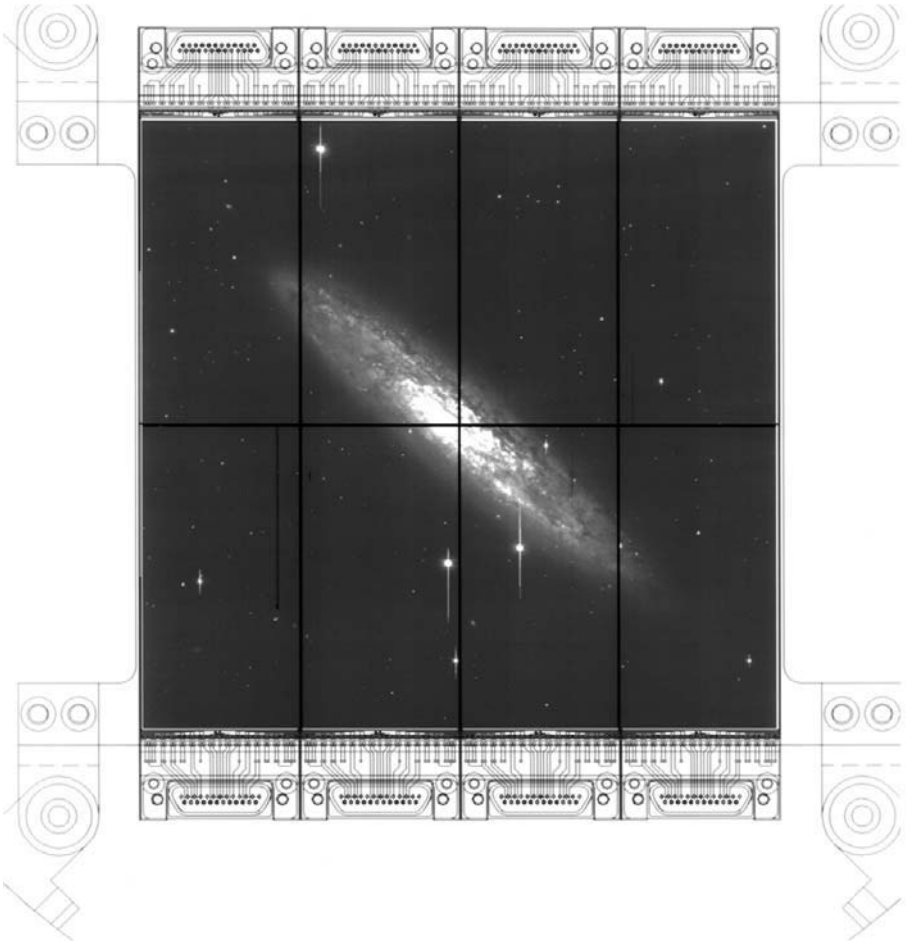


Figure 7.13. The layout of eight custom CCDs packaged into a mosaic overlaid on an image obtained with this camera system. Credit: Gerry Luppino.

curved surface or a large field-flattening lens must be introduced. Even if the focal surface is flat, depth of field is another factor. For fast optical systems the depth of field is shallow and the CCDs need to be in a common plane to better than $\sim$25 μm or one-thousandth of an inch. Every chip needs to be cooled efficiently, and there must be an easy way to remove and replace individual devices. Thus, packaging is a critical issue (Burke, Jorden, and Vu, 2005).

7.4.3 Drift scanning

As mentioned in Chapter 4 (Figure 4.7), one alternative to a large mosaic of detectors is to use either drift scanning or time-delay integration to obtain large swaths of sky.

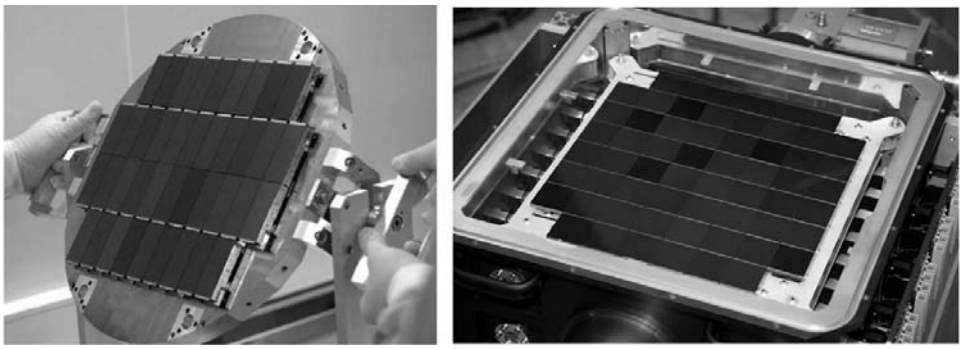


Figure 7.14. Two very large mosaics: the CFHT MegaCam (left) and the Pan-STARRS 1.4-gigapixel camera (right); see Plate 16. Credit: CFHT and University of Hawaii, respectively.

This technique is especially suitable for non-tracking telescopes designed to point at one location in the sky. In its original concept, drift scanning requires that the CCD is read out along its columns at normal slow scan rates during an exposure while a mechanical platform moves the detector at exactly the same rate in the opposite direction to ensure that the photon image always falls on the moving charge image. Each object in the image is sampled by every pixel in a column which therefore eliminates pixel-to-pixel variations along the column direction. The exposure time per object depends on the scan rate and the size of the CCD.

Time-delay integration or TDI is similar to drift scanning except that the CCD does not move, the telescope drive is switched off, and the charge is read out at exactly the sidereal rate as the sky drifts by. Both methods have the same advantage in terms of flat fielding because every pixel in a column contributes to the detection, but the integration time per object is limited in TDI mode by the size of the CCD chip. For a CCD with N pixels in the scan direction each subtending θ arcseconds on the sky, the integration time in seconds is given by

$$t_{\text{int}} = (N\theta/15.04)\cos\delta \tag{7.15}$$

where δ is the declination. For example, with $N = 2{,}048$ and $\theta = 0.5$ arcsec the integration time is only 68 seconds at the celestial equator and 34 seconds at $\delta = 60^\circ$. Images obtained using drift scanning or TDI tend to have slightly elongated PSFs in the east–west direction because the sky motion is continuous whereas the CCD charges are moved at discrete time intervals by the clocking. Objects separated by quite small declination differences, such as one field of view (17 arcminutes in the example above), will have different motion rates. At a declination of 60° this results in a difference of 0.29 s in the integration time, which is quite noticeable in terms of image smear.

The Sloan Digital Sky Survey uses a camera with 30 $2{,}048 \times 2{,}048$ CCDs arranged in a pattern of 6 columns $\times$ 5 rows (Plate 15) and makes TDI scans along great circles to image a swath of sky 2.5° wide. Scans are made at the sidereal rate in a direction perpendicular to the rows of CCDs. Each row of six CCDs has a different

photometric filter. The effective integration time per color is 54 seconds with a time delay of 72 seconds between colors caused by the sum of the CCD crossing time and the spacing of the chips which are not close-buttable.

Even more challenging is the imager for the Oschin Schmidt telescope (Mt. Palomar) which uses a close-butted mosaic of $28 \times 4 = 112$ CCDs, each with $600 \times 2{,}400$ pixels (161 megapixels in total) with one pixel subtending $0.85''$ on the sky and the entire array covering $4.6^\circ \times 3.6^\circ$. In operation since 2003, on a clear night a 4.6° wide by 120° long strip can be covered in four colors. The CCD array is oriented with the columns in the clocking direction lined up precisely east–west and TDI occurs at the sidereal rate. Each of the four rows of CCDs can have a different filter. The telescope does not track during the observations. At declinations other than the equator, stars follow arcs of circles and stars at different north–south positions move at different rates. The first effect smears the image in the north–south direction, and the second effect causes a smear in the east–west direction. To keep the smearing of the point spread function below about 1 arcsecond in any direction, each CCD is physically rotated by an amount dependent on the declination being scanned in such a way that the clocking direction of each CCD is tangential to the arcs that the stars are moving in at that location in the array. This remarkable concept is accomplished by mounting each of the four CCDs in a north–south row on an Invar finger. Each of the four fingers can be rotated by a different amount by cams which are driven by external, computer-controlled stepper motors. In addition, each column of CCDs in the mosaic is scanned along a slightly different declination, and therefore the parallel clocks reading out the CCDs are synchronized at slightly different rates. At the equator, the clock rate is approximately 17 lines/second which gives a CCD crossing time or exposure of about 141 seconds.

7.5 SPECIAL DEVICES

7.5.1 High-speed CCDs with on-chip gain

Despite remarkable improvements in CCD technology over the years a CCD is not a photon-counting detector because of its intrinsic (small) readout noise. However, a new technology developed by *e2v technologies* provides a way to reduce the readout significantly and increase the readout rate, which normally causes the noise to rise. Known as the L3CCD, for low light level CCD, the device consists of a normal CCD to which has been added an extended output register as shown in Figure 7.15. If the extended register is clocked with a 10 V swing in the usual way, it will behave like the normal serial register. However, one electrode is constructed in such a way that if it is clocked with a much higher voltage (typically 40 V), then there is a 1%–2% probability that an electron transferred into that electrode will create a second electron by avalanche multiplication. If the extended register has several hundred elements in which this can happen then a large gain is possible, similar to the cascade in a photomultiplier tube. For example, for a 1% probability of an extra electron released and 600 elements the gain is $1.01^{600} = 392$. At the end of the extended

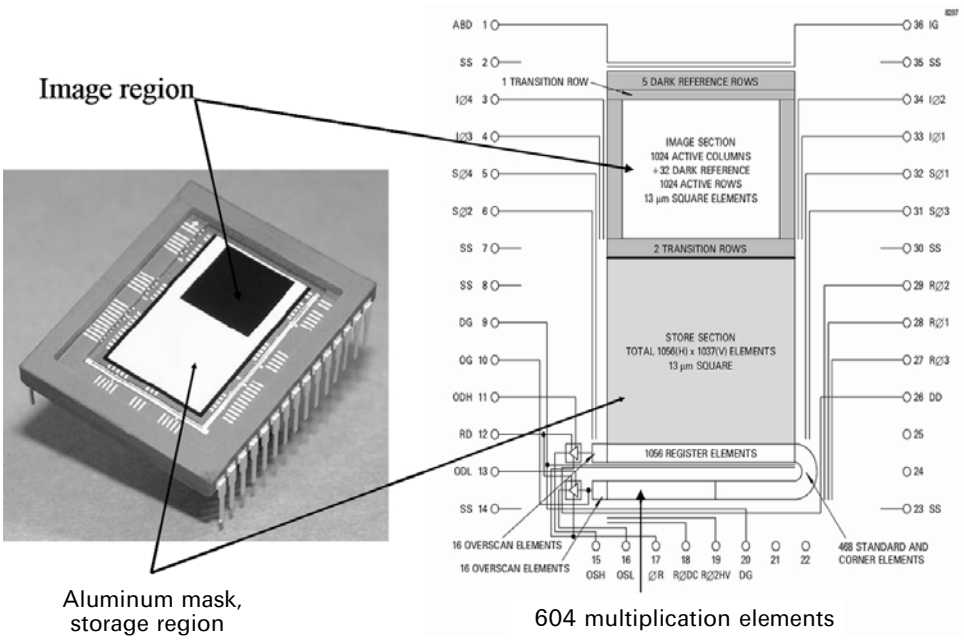


Figure 7.15. The layout of the extra serial register with avalanche gain in a low light level CCD (L3CCD) or electron-multiplied CCD (EMCCD). Credit: *e2v technologies*.

register the same kind of output amplifier as normally found in a CCD can be used. Readout noise is a voltage noise within the transistor of the output amplifier, but it is usually converted to an equivalent charge using the output capacitance and expressed in terms of a number of electrons (R). Because of the gain factor (G) in the extended register, the equivalent readout noise is reduced by a large amount ($R_{\text{eff}} = R/G$). When the CCD is read out at high pixel rates the readout noise can increase to perhaps 100 electrons rms, but the large gain factor makes this seem like the equivalent of a normal slow-scanned CCD with a noise of 1 electron rms or less. At the time of writing *e2v* have several formats up to 512×512 equipped with the electron multiplication register, and Craig Mackay (Cambridge, U.K.) and colleagues have used several of these for astronomy (Mackay *et al.*, 2001). More recently Texas Instruments have developed their own version of this technology, called electron-multiplied CCD or EMCCD. In principle, the extra register can be added to any existing CCD format.

Because of the reduction of the readout noise these CCDs can be used in photon-counting mode, provided that the dark current is eliminated by sufficient cooling. The signal-to-noise ratio (SNR) would be

$$\text{SNR} = S/\sqrt{[2S + (R/G)^2]} \tag{7.16}$$

The multiplication effect is a stochastic process that adds to the photon shot noise as well, hence the $2S$ term. If a pixel recorded S photoelectrons, then after multiplication

the uncertainty would be $\sqrt{2S}$ instead of $\sqrt{S}$, which is equivalent to a reduction by a factor of 2 in the quantum efficiency for high signal applications. This problem can be solved by using the CCD as a true image photon-counting system, so that each pixel event is recognized as a single photon. Because L3CCDs can be read out at hundreds of frames per second the actual signal rate in terms of photons per pixel per second can still be substantial, unlike the image photon counting systems mentioned in Chapter 1 where problems arise at signal rates above about 1 photon/s/pixel.

Different CCD controllers are needed for L3CCDs and EMCCDs because of the high pixel rates. If slower speeds are used, however, then a 40 V clock line can be added to a standard controller. The new controller structure must allow both high pixel clocking rates and precise timing control of the clock edges, especially edges used for analog-to-digital conversion, to a small fraction of the pixel period. Very fast clean waveforms are needed, and the signal-processing system must allow the full dynamic range of the CCD even at the highest pixel clocking rates (e.g., 20 MHz). Current devices tend to show "clock-induced charge" (CIC) events across the chip which may be related to the sharp clock edges. Another example of the use of the L3CCD is in ULTRASPEC, an instrument for high-speed spectrophotometry (Ives, 2008). This technology may also be very promising for next-generation AO systems.

7.5.2 Deep-depletion CCDs for no fringing

The depth at which an incident photon generates an electron–hole pair in a silicon CCD depends on wavelength. Longer wavelength photons penetrate deeper into the silicon before charge generation occurs. In front-illuminated CCDs at visible wavelengths the charge is mainly generated within the depletion region beneath the electrode gates as required. The electric field in the depletion region drives charge to the buried channel for charge storage. At longer wavelengths near the silicon cutoff, however, much of the charge generation occurs deep in the essentially field-free silicon, below the depletion region, and therefore this charge can diffuse away in a random manner. If the signal charge is collected in the wrong pixel as a result of diffusion then there is a loss of spatial resolution. For backside-illuminated devices the total thickness of silicon is usually less than 20 microns and the longest wavelengths ($\lambda > 800$ nm) can pass right through without absorption. Apart from the large drop in quantum efficiency, some of these photons can reflect back into the substrate from the front side and cause interference fringes. To eliminate these problems and get the best far-red performance one can adjust the anti-reflection coatings and increase the active thickness layer of the CCD.

The actual depth of the depletion is proportional to V/N_A, where V is the channel potential, and N_A is the doping concentration of the silicon below the buried channel. By optimizing the deposition conditions it is possible to produce a silicon layer with much lower doping concentrations and hence higher resistivity (usually given the Greek letter, rho, ρ), which in turn means a deeper depletion region, by a factor of 3 or more (Figure 7.16). These devices are called "high-rho" or "deep-depletion" CCDs. Devices with this property were developed at the Lawrence Berkeley National Lab (e.g., Holland *et al.*, 2003), but are now also available from *e2v*

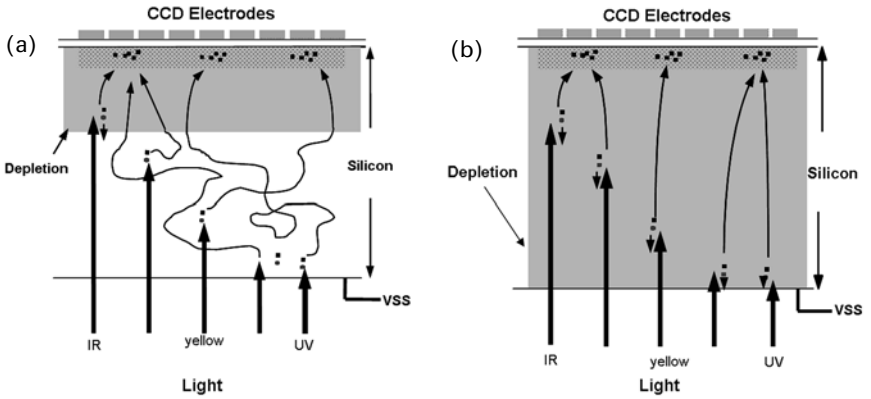


Figure 7.16. (a) Thick device, back-illuminated with normal depletion (absorption of photons beyond 800 nm requires a thicker device which leads to poor collection at shorter wavelengths). (b) Thick device, back-illuminated with deep depletion (a deep depletion region ensures that photons of all wavelengths are collected correctly). Credit: Paul Jorden.

technologies and MIT/Lincoln Labs. Such devices have a much improved red response which in turn results in much less fringing. Quantum efficiency plots are shown in Chapter 9 (Figure 9.3).

7.5.3 Orthogonal transfer CCDs

Orthogonal transfer (OT) devices are a unique kind of CCD technology developed at MIT/Lincoln Labs for Pan-STARRS and were first described by John Tonry (University of Hawaii) and others in 1997. The layout of the transfer gates is symmetric to 90° rotations as shown in Figure 7.17. Thus, the charge can be shifted using a three-phase parallel clock in the vertical direction (as usual), or in the horizontal direction, by setting the appropriate vertical or horizontal gates to a negative potential to act as a barrier. With this architecture, the collected charge can be moved up, down, left, or right. An optical image dancing around on a CCD detector due to atmospheric turbulence results in a blurred image in a long exposure. However, if you knew which direction the image was moving, then you could move the charge already stored from the previous moment in the same direction. With an OTCCD, you can keep pace with the motion. In effect, you would be performing the tip/tilt correction in real time on the CCD itself using the electronic charges. Of course, moving the electrons around while the photons are still falling can cause blurring, but as the offset between the optical image and the collected charge never needs to be larger than 0.5 pixel, the blurring is equivalent to convolution with a 1-pixel point spread function. Non-perfect charge transfer efficiency can also cause blurring, but a charge transfer efficiency of more than 0.99999 and a few thousand shifts per exposure creates a very small effect that is lost in the optical PSF. Orthogonal transfer clocking can also create spurious charge for large clock voltages, but as a normal readout involves thousands of shifts anyway, the effect is not significant. Because the Pan-STARRS

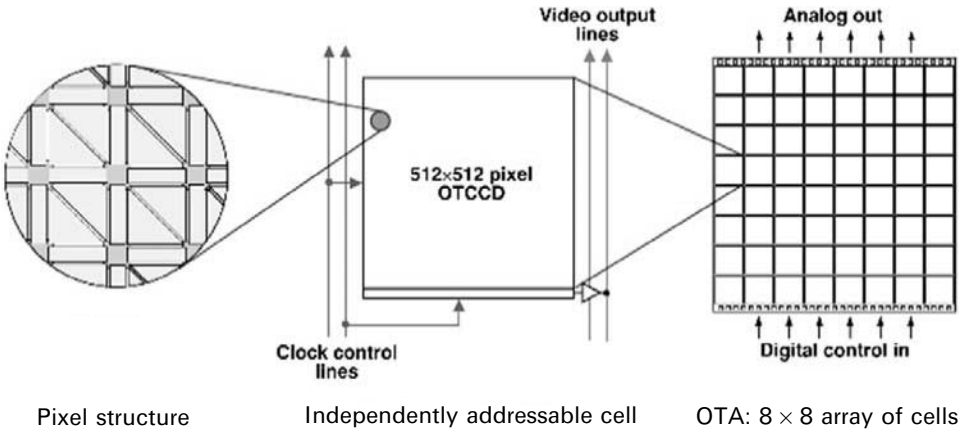


Figure 7.17. The layout of the orthogonal transfer CCD. Credit: John Tonry

focal plane is made up of many individual OTCCDs, some units can be read out to track suitable guide stars and the appropriate shifts applied to the adjacent OTCCDs. Because the field of view is very large for Pan-STARRS compared with the isokinetic patch, the same correction would not work across the entire field. However, with numerous groups of OTCCDs in the focal plane there is no need to apply the same correction to each region. First light with one of the Pan-STARRS cameras was obtained in August 2007. In a related development effort, an array of 64 (8×8) OTCCD cells to yield a device with $3{,}840 \times 3{,}952$ pixels is being developed by Semiconductor Technology Associates (STA) under contract with the WIYN Consortium and licensed through MIT/Lincoln Labs for a 1-degree imager for the 3.5 m telescope that will eventually require 64 of these $\sim 4\,\mathrm{K} \times 4\,\mathrm{K}$ units.

7.5.4 Customized CCDs

Orthogonal transfer CCDs are only one way in which this technology can be customized for special applications. Other developments include laying out radial strips so that the charge can be clocked along a radius vector from the center and at different rates as a function of annulus (Beletic *et al.*, 2005). Application of devices like this include handling the observed elongation of a laser guide star spot caused by the large separation between the launch axis and the edge of the primary mirror, and the finite thickness of the sodium layer. The primary limitation with all of these custom designs is cost and yield. Many more devices must be fabricated at higher cost than standard chips to get the required performance. CCDs can also be constructed with optimized anti-reflection coatings, even as a function of position across the detector, and CCDs have now been developed with internal "electronic" shutters that stop them acting as detectors. Barry Burke (MIT/Lincoln Labs), Paul Jorden (*e2v technologies*), and Paul Vu (Fairchild Imaging) provide a good review of CCDs as of 2005; see Burke, Jorden, and Vu (2005).

7.6 SUMMARY

The basic structure of a CCD is a two-dimensional grid of metal oxide semiconductor pixels controlled by overlying metallic electrodes arranged in strips. Charges accumulate where the light falls on the CCD and build up an image. The pattern of charge can be "read out" by systematically pulsing or "clocking" the electrode strips to cause the charge pattern to "couple" from one pixel to the next and move along the columns of the CCD to an output register. Modern CCDs are of the buried-channel construction in which the charges collect beneath the surface. CCDs can be front-illuminated, thinned and backside-illuminated, or thick but with a deep depletion. For astronomy applications, CCDs must be cooled and operated in "slow-scan" readout mode. However, CCDs with on-chip gain circuits (electron-multiplied) are now available that allow low-noise performance at very high speeds. In addition, different types of CCDs can be designed for special applications such as electronic tip/tilt control or with variable-sized pixels to match laser guide stars. Although individual CCDs with 4 Mpxl–16 Mpxl are readily available, many observatories combine CCDs to create much larger mosaics, with several hundred megapixels now common, and gigapixel cameras now feasible.

7.7 EXERCISES

1 Describe with the aid of diagrams a three-phase surface-channel CCD, and show how photo-generated charges are collected, stored, and transferred to the output.

2 Describe with the aid of a diagram what is meant by a "virtual-phase" CCD.

3 Draw a diagram of a "buried-channel" CCD to illustrate where charges are collected. Cite its advantages over a surface-channel CCD.

4 Explain what is meant by "inverted" operation. Why is inverted operation of a CCD advantageous?

5 Compare and contrast "interline" and "frame transfer" CCDs for astronomy applications. Why is the interline transfer approach attractive for standard TV and video rate applications?

6 What is meant by a "thinned" CCD? What are the advantages and disadvantages of thinned CCDs?

7 Briefly explain the operation of a multi-pinned phase or MPPCCD.

8 What is an electron-multiplied (EM) CCD? Give an example of the gain achieved.

9 What is meant by a "high-rho" CCD, and what problem does it solve?

10 Describe the orthogonal transfer (OT) CCD and explain its application to provide tip/tilt corrections over a large field of view.

7.8 REFERENCES AND SUGGESTIONS FOR FURTHER READING

Amelio, G.F. (1974) Charge-coupled devices, *Scientific American*, February.

Beletic, J.E.; Beletic, J.W.; and Amico, P. (2006) *Scientific Detectors for Astronomy 2005: Explorers of the Photon Odyssey*, Astrophysics & Space Science Library, Springer-Verlag, Dordrecht, The Netherlands.

Benyon, J.D.E.; and Lamb, D.R. (1980) *Charge-coupled Devices and Their Applications*, McGraw-Hill, New York.

Boulade, O.; Vigroux, L.; Charlot, X.; Borgeaud, P.; Carton, P-H.; de Kat, J.; Rousse, J.Y.; Mellier, Y.; Gigan, P.; Crampton, D.; and Morbey, C.L. (1998) Megacam: The next-generation wide-field imaging camera for CFHT, in S. D'Odorico (ed.), *Optical Astronomical Instrumentation*, SPIE, Bellingham, WA [*Proc. SPIE*, **3355**, 614–625].

Boyle, W.S.; and Smith, G.E. (1971) Charge-coupled devices: A new approach to MIS device structures, *IEEE Spectrum*, **8**(7), 18–27.

Burke, B.; Jorden, P.; and Vu, P. (2005) CCD technology, *Experimental Astronomy*, **19**, 69–102. [Compact review from MIT/Lincoln Labs, *e2v technologies*, and Fairchild Imaging.]

Gorjian, V.; Wright, E.; and McLean, I.S. (1997) Drift scanning using infrared arrays: A powerful method for background-limited imaging. *Publ. Astron. Soc. Pac.*, **109**, 821–826.

Gunn, J.E.; Carr, M.; Rockosi, C.; Sekiguchi, M.; Berry, K.; Elms, B.; de Haas, E.; Ivezić, Ž.; Knapp, G.; Lupton, R. *et al.* (1998) The Sloan Digital Sky Survey Photometric Camera, *Astron. J.*, **116**, 3040–3081.

Holland, S.E.; Groom, D.E.; Palaio, N.P.; Stover, R.J.; and Wei, M. (2003) Fully depleted, back-illuminated charge-coupled devices fabricated on high-resistivity silicon, *IEEE Trans. Electron Devices*, **50**, 225–238. See also *http://www-ccd.lbl.gov/*

Howell, S.B. (2006) *Handbook of CCD Astronomy*, second edition, Cambridge University Press, Cambridge, U.K.

Janesick, J.R. (2001) *Scientific Charge-Coupled Devices*, SPIE, Bellingham, WA. [Excellent detailed book on all aspects of CCDs.]

Janesick, J.; and Blouke, M. (1987) Sky on a chip: The fabulous CCD, *Sky & Telescope*, September, **74**, 238–242.

Janesick, J.R.; Hynecek, J.; and Blouke, M.M. (1981) A virtual phase imager for Galileo, in *Solid State Imagers for Astronomy: Proc. Soc. Photo-Opt. Instr. Eng. (SPIE)*, **290**, 165–173.

Jorden, P.R.; Pool, P.; and Tulloch, S.M. (2004) Secrets of E2V Technologies CCDs, in P. Amico, J.W. Beletic, and J.E. Beletic (eds.), *Scientific Detectors for Astronomy*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Kristian, J.; and Blouke, M. (1982) Charge-coupled devices in astronomy, *Scientific American*, October, **247**, 66–74.

Mackay, C.D. (1986) Charge-coupled devices in astronomy, *Annual Reviews of Astronomy and Astrophysics*, **24**, 255–283.

Mackay, C.D. (1992) *The Role of Charge Coupled Devices in Low Light Level Imaging*, AstroMed Press & Publications, Cambridge, U.K.

Mackay, C.D.; Tubbs, R.N.; Bell, R.; Burt, D.J.; Jerram, P.; and Moody, I. (2001) Subelectron read noise at MHz pixel rates, *Proc. SPIE*, **4306**, 289–298.

McLean, I.S. (1989) *Electronic and Computer-Aided Astronomy: From Eyes to Electronic Sensors*, Ellis Horwood/John Wiley & Sons, Chichester, U.K.

Rieke, G.H. (2003) *Detection of Light from the Ultraviolet to the Submillimeter*, second edition, Cambridge University Press, Cambridge, U.K.

Sequin, C.H.; and Tompsett, M.F. (1975) *Charge Transfer Devices*, Academic Press, New York.

Tonry, J.L.; Burke, B.E.; and Schechter, P.L. (1997), The orthogonal transfer CCD, *Publ. Astron. Soc. Pac.*, **109**, 1154.

Web sites

Bonn shutters: *http://www.astro.uni-bonn.de/~ccd/shutters/*

For low light level CCDs: *http://www.e2v.com/products/ccd-and-cmos-imaging-and-semiconductors/imaging-scientific-sensors.cfm*

For orthogonal transfer CCDs: *http://pan-starrs.ifa.hawaii.edu/public/* and *http://www.noao.edu/wiyn/ODI/index.htm*

CCDs for college teaching

Complete thermoelectrically cooled CCD cameras packaged into small units ready for attaching to modest-sized telescopes have flooded onto the market and opened up a great opportunity for amateur astronomers and for teaching uses at schools, colleges, and universities. If you are reading this book it is likely that you have access to a CCD camera. A wide variety of cameras are available with small CCDs of only 192×165 pixels such as the TC211, up to those with $2{,}048 \times 2{,}048$ pixels like the Kodak KAF4200. Some companies sell only the camera heads and electronics while others provide PC-based image-processing software ready to support popular cameras. A few companies provide complete turn-key packages containing camera, computer, and software ready to control your small telescope. One of the best resources is to look at the ads placed in magazines such as *Sky & Telescope* and then go to the web sites of the companies involved.

8

Practical operation of CCDs

Modern CCDs are fairly predictable in their operation and characteristics, but there are still many subtleties to successful operation, especially for devices customized for astronomy. Issues include maximizing the ratio of signal to noise, obtaining stability and repeatability in performance, and finding suitable methods of control. These topics are treated by recounting some of the developments which uncovered problems and led to today's solutions. There is no intention here to provide a "constructor's manual", only to alert the potential user to a host of practical issues. Many of the same practical issues apply for other detectors and other wavelength regimes.

8.1 CLOCK VOLTAGES AND BASIC ELECTRICAL FACTORS

CCD manufacturers provide a data sheet which gives the electrical pin connection diagram for the device (Figure 8.1), the names and symbols for each pin, the voltages or range of voltages to be applied to each pin, and the timing diagram (i.e., a diagram showing the time sequence of the CCD drive signals and the relationship between them). Terminology varies, but certain basic functions are common to all. Voltages applied to CCDs are of two types: fixed voltages, referred to as "dc bias" levels, which remain unchanged after switch-on; and pulsed or "clock" voltages which can be switched back and forth between two voltage levels known as high and low levels. As described in Chapter 7, clock voltages are applied in a precise order and time sequence to charge-couple the electrons from one storage well to the next. Although the type of mounting package used (its size, number of pins, and their names) differs from device to device, certain functions are required in all CCDs, and in particular in the frame transfer image sensors used widely in astronomy, namely:

1. *Serial (horizontal) register clocks*. One pin for each phase or electrode used to define the pixels in the horizontal register is required (i.e., three for a three-phase

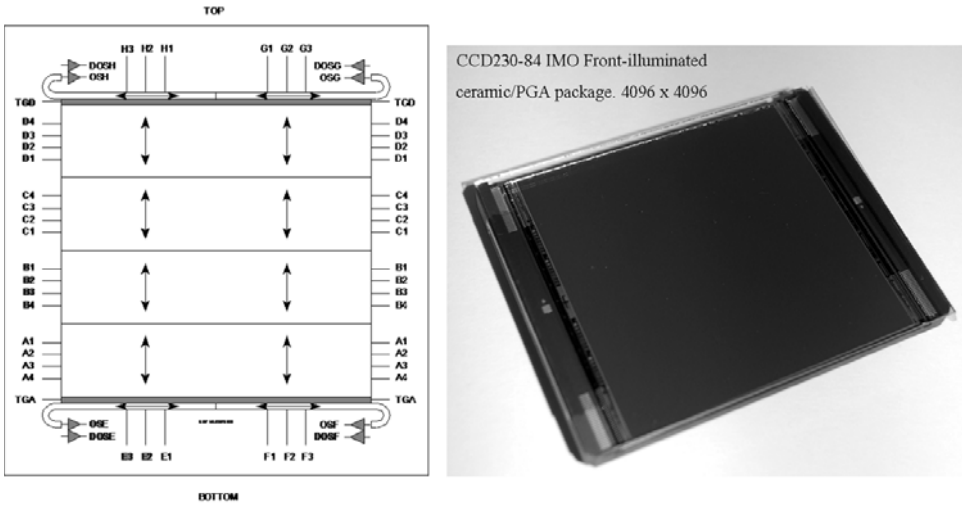


Figure 8.1. The device schematic for a 4 K × 4 K frame transfer CCD to illustrate typical pin assignments and terminology. Each manufacturer will provide information of this kind on a data sheet. Credit: Paul Jorden and *e2v technologies*.

CCD). In many larger CCDs two completely separate serial output registers are provided on opposite sides of the active area. Sometimes each serial register has outputs at both ends giving four outputs in total.

2. *Parallel (vertical) register clocks*. Again one voltage line is needed for each phase. In frame transfer CCDs the two-dimensional vertical or parallel register is split into two identical sections, the "image" section and the "store" section, which can be controlled separately and therefore two sets of pins are needed.
3. *Reset transistor clock*. A single, periodically recurring voltage pulse is required to reset the CCD output amplifier, or more accurately the output charge-collecting capacitor during the readout process. Each of these clock voltages will have a specified high and low level, and since voltage difference or "swing" (usually in the range 5 V to 10 V) can have important consequences on performance, it is often arranged to select the levels on demand from the CCD controller. The total range of voltage required to operate a CCD is generally less than 20 volts, except for electron-multiplied CCDs that require one clock at about 40 V.

The most important dc bias voltages are as follows:

1. *Substrate voltage (V_{sub})*. This is the reference for all other voltages. This voltage is usually, but not always, kept at ground or zero volts.
2. *Reset drain voltage (V_{RD})*. This voltage is applied to the "drain" terminal of the on-chip "reset" field effect transistor (FET) at the CCD output to establish the level to which the output node (capacitor) must return after each charge packet is read out.

3. *Output drain voltage (V_{OD})*. This voltage applied to the drain terminal of the onchip output amplifier determines the operating point of that transistor.
4. *Output gate voltage (V_{OG})*. The output gate is essentially an extra (last) electrode in the serial output register.

Nomenclature varies slightly from one manufacturer to another, but details are always given on the data sheet.

Pulsed voltage signals, corresponding to individual charge packets, emerge from the output transistor (OS), also called the video output, which is connected to ground via another transistor external to the CCD to provide a constant current load. The output transistor source current (I_{OS}) is important for most CCDs. Some CCDs have additional options. For example, the vertical register can be clocked up or down to independent serial registers. One serial register may terminate with an on-chip amplifier designed for low-noise, slow-scan operation whereas the other register might have an output amplifier optimized for TV video rates. Some chips have a separately clocked gate known as a "summing well" which has the storage capacity of two serial pixels. Almost all CCDs have electrical input connections and test points; the manufacturer will specify whether these pins should be fixed at a high or low voltage. Remember also that voltages are relative to the substrate voltage which may not always be zero volts. Normally, a CCD manufacturer will provide an initial set of operating voltages for a particular chip, but will not necessarily optimize these voltages for cooled slow-scan astronomical work. Builders of commercial astronomical CCD camera systems can provide this service. For some CCDs, small changes of one-tenth of a volt to clock swings or dc bias values can often yield substantial improvements in low light level behavior.

Safe handling: CCDs, including thinned devices, can withstand substantial illumination overloads without permanent damage. Mechanically, CCDs are also quite robust unless heavily thinned. However, they are integrated circuits of the CMOS type and their tiny gate connections can be short-circuited by static electricity discharges. Precautions against static must be taken when handling CCDs. For example, CCDs should be stored in electrically conducting containers, earthing straps should be tied to the wrist during handling operations, no nylon clothing worn, and the work performed in a clean, ionized (electrically conducting) airflow. Finally, some kind of protection on the drive signals is essential if power supplies are used which have a rating even slightly above the maximum recommended voltages for the CCD. Usually this is achieved by the use of Zener diodes on the drive outputs. With these precautions, a CCD camera can last a long time.

8.1.1 The analog signal chain

The analog signal chain includes the pre-amplifier, post-amplifier, noise removal circuits, and analog-to-digital converter (ADC). In addition, low-noise dc power supplies are required to provide the bias voltages, and a "level shifter" circuit is

usually needed to convert simple 0 V–5 V TTL clock pulses to the levels required for the chip. Figure 8.2 shows a typical signal chain after the pre-amplifier.

The heart of the electronic system is a signal-processing unit designed to sample and filter noise, but proper grounding, good power supply decoupling, and optimum control timing are also very important factors in achieving a low-noise system. Undoubtedly, "ground-loops" are the most common cause of noisy CCD systems. A ground-loop forms when two interconnecting parts of an electronic system are separately connected to ground via small—but different—impedance paths. As a result, a voltage difference can exist between "grounds", and currents can flow. Ground-loops between the telescope and camera body and driver electronics, within the driver electronics, or between the driver electronics and the computer system can cause interference with the readout electronics. The observable effect is a clearly visible pattern of diagonal stripes in the image; these patterns are synchronized with the mains frequency. The solution to this problem is to have only a single ground point in the system, to which all the zero reference points and shields are connected; because many wires may radiate from such a single ground point it is often called a "star ground" (see Figure 8.3). All connections to the star ground should be made as short as possible (< 1 m for frequencies up to 10 MHz), with the lowest resistance electrical wires available. Wherever possible, circuit boards should use copper ground planes. Many designers carefully isolate the entire instrument from the telescope structure, even if the telescope is known to have an excellent earth ground, which means that extreme care is required when attaching any other piece of electrically powered apparatus to the instrument in case a ground-loop is formed.

Similarly, electrical noise from motors, light dimmers, or computer parts can be "picked up" by capacitive coupling, inductive coupling, or radiative coupling if inadequately shielded wires or components are used in the CCD system. Remember, the CCD system is capable of detecting signals of only a few millionths of 1 volt! Co-axial cables which have a surrounding braided copper shield are quite effective (90%), and "twisted-pair" wire is often used in preference to a single wire for carrying critical signals to the CCD. Signals transmitted over long distances, such as from telescope to control room, usually use optical fibers if possible. Most astronomical CCD systems have a low-noise pre-amplifier inside or very close to the cryostat to boost the CCD signal, and most systems convert from the weak analog output to healthy digital signals before transmitting the data over long cables. The physical environment at cold, high-altitude mountaintop observatories can cause "drifts" in the operating points of many electronic components; this effect can be overcome by careful component selection and packaging.

8.1.2 CCD sequencers and clock drivers

Much of the complexity in a CCD camera, or other electronic imaging systems, revolves around the critically important timing diagram (see Figure 8.4 for an example). Numerous events must occur in strict time order and at precise time intervals. For example, in a typical CCD system the following events must occur

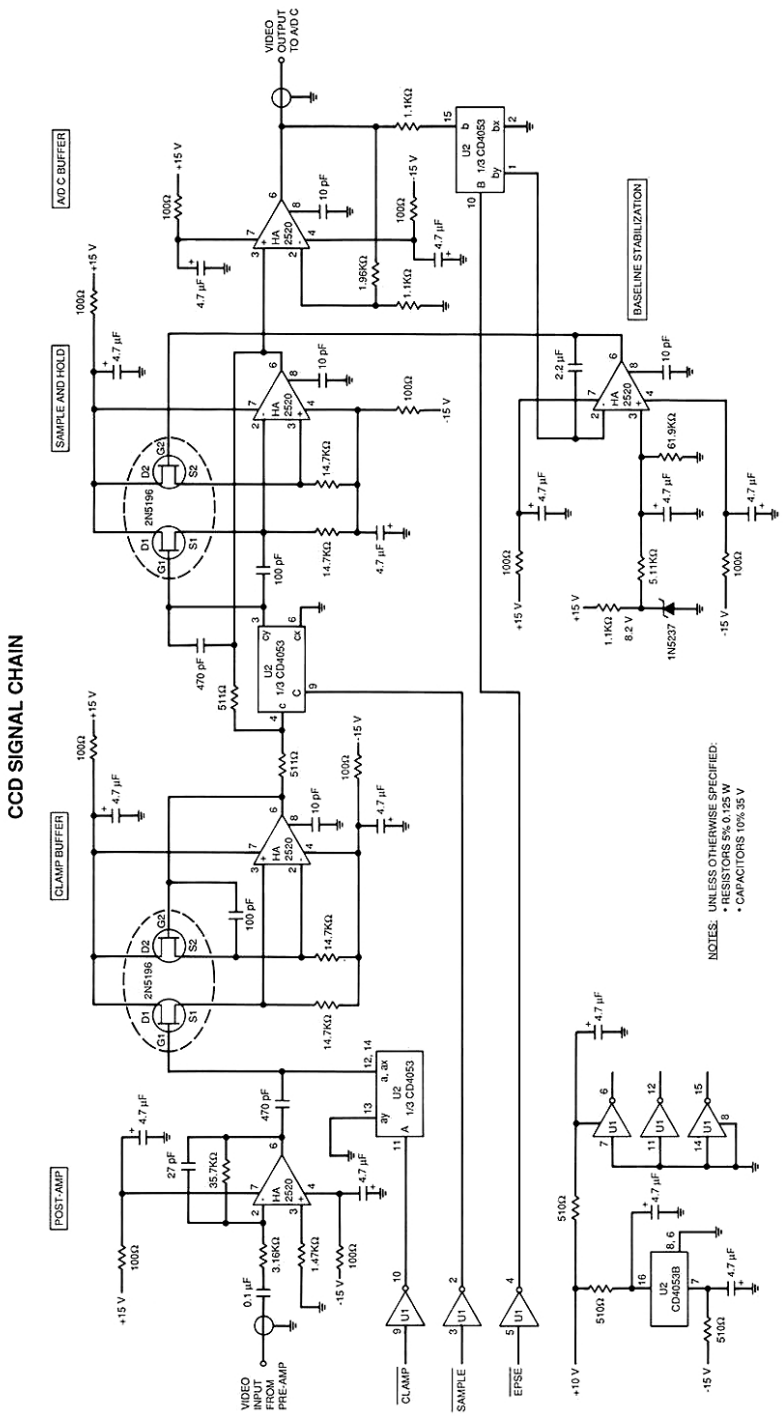


Figure 8.2. Part of the analog signal chain. This is a practical correlated double-sampling circuit used at JPL. Credit: Jim Janesick.

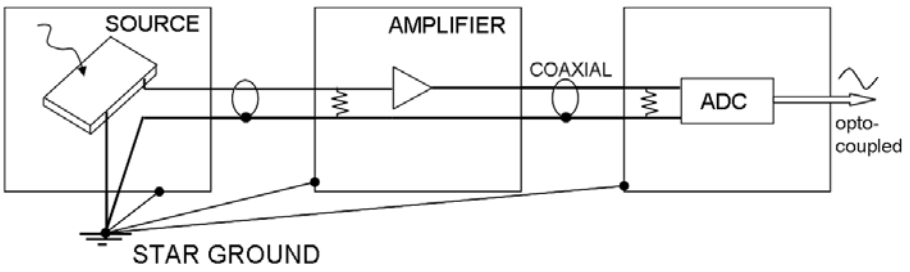


Figure 8.3. The concept of a star ground.

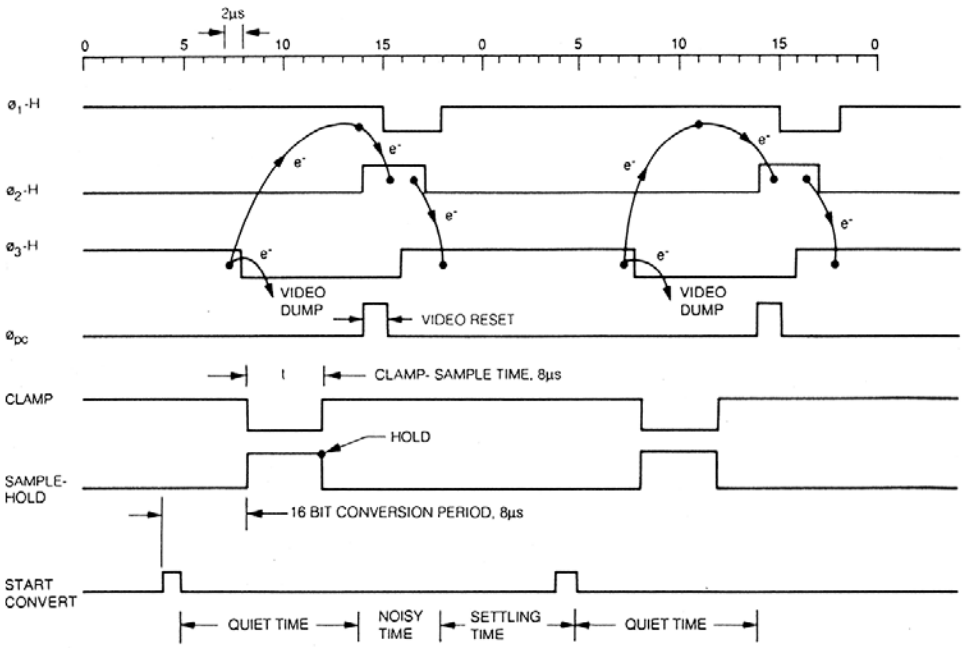


Figure 8.4. A typical timing diagram. The waveforms are produced digitally (usually as 0 V–5 V TTL signals) and then converted to the required voltages by a "level shifter" circuit. Credit: Jim Janesick.

in time sequence:

- check instruction status from operator
- prepare the CCD (e.g., unrecorded readouts to flush the CCD of trapped charge)
- open the shutter for a timed period
- close the shutter
- read out the CCD according to a precise pattern
- digitize the signal from each pixel
- store the data in a computer
- return the CCD to "standby" mode, if appropriate.

That part of the system responsible for such timing activities is often called the sequencer. Almost all CCD systems fall into one of the following design categories, either "hard-wired" or "programmable".

Hard-wired designs: By "hard-wired" we mean that the electronic functions are carried out by circuitry, and cannot be altered by typing instructions on a computer keyboard. This approach results in a very compact and rugged system, and is favored by those with traveling systems or with severe constraints on space and/or weight, such as small telescopes. Obviously, the hard-wired approach is more restrictive in the sense that the circuitry has to be altered, although sometimes only by re-programming a special logic chip, in order to effect a change such as switching to another type of CCD, or to an alternative clocking sequence.

In a typical hard-wired CCD camera, the digital controller for the camera is built with commercially available logic chips such as TTL or low-power, high-performance CMOS. Readout of the CCD is in a fixed form, and all clocks and signal-processing pulses are generated on a single board, usually called the "clock logic" board. This circuit board derives its time-base from a small device known as a quartz crystal oscillator. The clock logic board in turn feeds both the CCD clock "driver" board, which has more current-handling capacity than logic chips for pulsing the CCD electrodes, and the "analog processor and digitization" board. A fourth board, containing several low-noise well-regulated dc power supplies for setting the fixed voltage levels on the CCD as well as providing auxiliary functions such as temperature control, completes the so-called "camera head" electronics. Figure 8.5 shows an LN_2-cooled CCD cryostat with electronics box attached. The whole four-board controller could fit into a small box of typically $13 \times 15 \times 20$ cm ($5.1 \times 5.9 \times 7.9$ inch) or smaller.

Programmable designs: By "programmable" we mean that the electronic functions can be changed by computer commands. This capability becomes important when many different kinds of CCDs are to be used, tested, or developed. There are a great many ways to provide programmable systems, especially since the advent of digital signal processor (or DSP) chips and programmable read-only memories (PROMs). If the microprogram is stored in PROM, the system can be operated without the main data-gathering computer and the program will always be available. There is the disadvantage that one must replace the PROMs with new ones to effect a timing or sequence change. This process is called "burning" PROMs and requires special—although very compact—equipment and a computer to develop the new PROM codes on. Erasable PROMs (EPROMs) are available which at least save on the purchase of these microchips. PROMs provide a simple and reliable solution when the camera is fully commissioned and performing in a scientific instrument. For development engineers, however, it can be much faster and more convenient to replace the PROMs with random access memory (RAM) chips. These devices must have the microprogram sent to them from the main computer, which has overall control, each time the sequencer is switched on; this process is known as "downloading". Several different programs can be stored in the host computer.

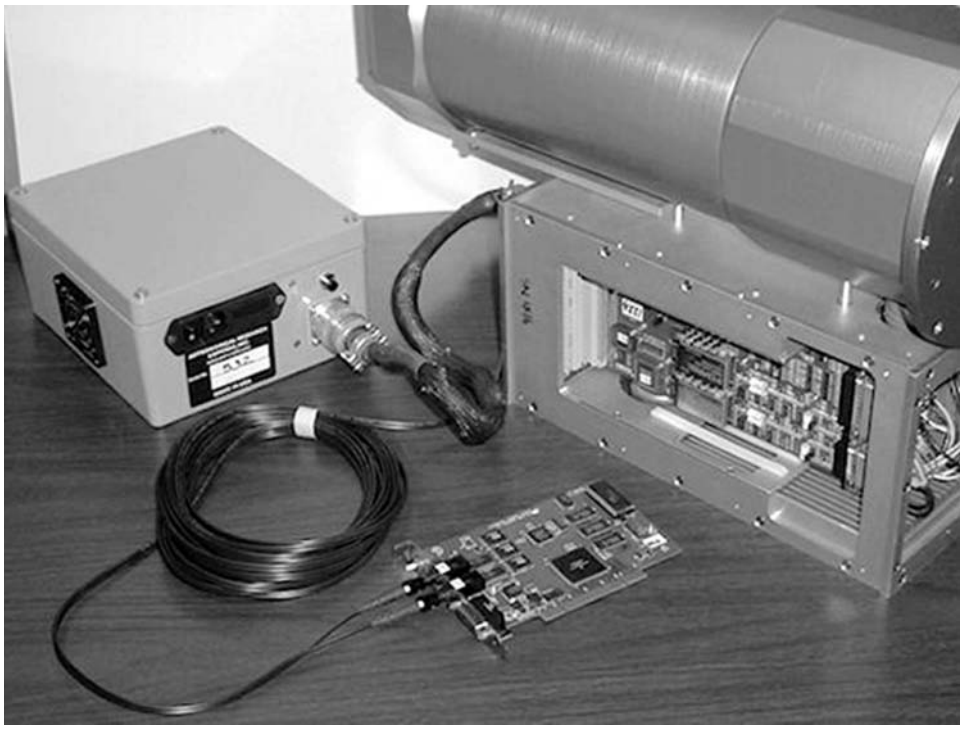


Figure 8.5. A typical camera "head" electronics box and CCD controller. This one shows the system developed by Astronomical Research Cameras. Credit: Bob Leach.

A practical approach to controller design requires room for expansion and change as new devices come along. It therefore makes sense to use a modular design. For example, one might design a system to allow the following range of configurations:

(a) slow or fast readout
(b) on-chip pixel binning (two axes)
(c) digitization of only a subset of the image area
(d) bi-directional charge shifting
(e) overscanning.

These modes can be built up from a small set of primitive waveforms. The longest of these involves the digitization of a pixel value which typically takes 20 μs–50 μs due to the integrating correlated double-sampling method to be described in Section 8.6. Also, operational modes require groups of consecutive sequences of the various waveforms. For instance, suppose the waveform sequencer is based on a programmable read-only memory (PROM) designed to accept 16-bit commands with two parts from the microprocessor. The first part specifies which waveform is required,

while the second part specifies the number of waveforms to be generated (i.e., a "repeat" count). To smooth the flow of commands and eliminate timing problems a FIFO (first-in-first-out) buffer can be placed in the 16-bit data path between the microprocessor and the waveform sequencer. This method provides a clean split between the hardware and the software with the microprocessor responsible for breaking down high-level commands (e.g., "take an image") into a stream of sequencer-compatible 16-bit commands. The PROM-based nature of the sequencer implies that it can be easily updated to provide any desired format of driving/readout waveform.

Many programmable controllers now incorporate powerful 32-bit processors, such as DSPs, which can provide significant intelligence and computing power within the controller itself, and yet are easily programmable with all the desired waveforms. Motorola 56000 series DSPs are used in the controllers developed by Bob Leach of Astronomical Research Cameras (ARC), San Diego, CA. Inmos T805 transputers were used in general-purpose systems developed at Cerro Tololo Inter-American Observatory (CTIO) by Roger Smith, Alistair Walker, and colleagues. These and similar systems were used at major observatories in Hawaii, La Palma, Arizona, Chile, South Africa, Australia, and in my own work for many years. Unfortunately, Inmos transputer chips are no longer available.

A survey of instruments and detectors in 2005 (Simons *et al.*, 2005) showed that the controllers made by ARC (also known as Leach controllers and SDSU controllers for their origin at San Diego State University) are the most widely used by astronomers, followed by the European Southern Observatory (ESO) controllers called IRACE and FIERA. However, about 30% of the total could not be grouped with the major observatories or suppliers, thus representing a large number of custom-made systems. Many of the large astronomical institutes continue to develop their own in-house systems.

The current version of the ARC controller is Gen IV. These CCD/IR array controllers are modular systems intended to be tailored to individual requirements. A general-purpose timing generator built around a DSP can be programmed to generate waveforms for readout and control sequences. The voltage levels required to control the detector are all programmable by the user, as is the way the video output gets processed. A variable number of circuits can be included in a system depending on the number of devices to be operated and the number of distinct video signals to be processed. All software source code and almost all electronic circuit diagrams are made available at no extra cost. A typical system has an eight-channel CCD video processor and DC bias supplies on a single board, a 12.5 Mpxl/s fiber optics data link into PCI (peripheral component interconnect) and PMC (PCI mezzanine card) interface boards, controller programming in C rather than DSP assembler language, and software support for Sun Solaris, Linux, and Windows 2000/XP operating systems. The CCD video board can process eight channels simultaneously with speeds up to 3 million pixels per second at 16-bit resolution using a conventional dual-slop integrator (Section 8.6), and it provides 32 dc bias outputs so each of the eight readout circuits can be driven independently. The timing board combines the functions of waveform generation, fiber optic communication, utility,

and power control on one board. An Altera FPGA (field programmable gate array) is the main processor. An FPGA is a small chip containing programmable sections called "logic blocks" and programmable interconnections between blocks. Logic blocks can be programmed to perform logic such as AND, XOR (exclusive OR), or more complex combinations of functions, like decoders, FIFOs, or simple mathematical functions. In most FPGAs, the logic blocks also include memory elements, which may be simple flip-flops or more complete blocks of memory. Ethernet interfaces are also included. In the ARC controller the FPGA is used for generating waveforms from tables with a resolution of 4 nanoseconds, and it also contains a C-programmable RISC (reduced instruction set computer) processor that manages the waveform generator. The clock driver board can handle both CCDs and IR arrays with a programmable ± 12 V range with 15 ns rise times.

Another example of a controller is the system produced at the NOAO in Tucson (Arizona) which is called MONSOON. This system is designed to be scalable, handle many channels, and operate IR arrays as well as CCDs. Four electronic modules are supported by core firmware and software suites, and upgradable 1 Gbit/s fast optical fiber links. The names of the four units illustrate the functions required: a Master Control Board (MCB) controls communication to the Pixel Acquisition Node (PAN) computer, the Detector Head electronics crate and the detector clock sequencer; the Clock and Bias Board (CLK) produces low-voltage biases and clock signals; the IR Acquisition board (IRACQ) acquires up to 36 channels of dc-coupled video signal for infrared cameras; and the CCD Acquisition Board (CCDACQ) handles up to eight channels of ac-coupled video signal and provides high-voltage biases. All control functions are performed by a generic pixel server program written in C. The sequencer is handled by an application-specific micro-controller in an FPGA.

Yet another example of an advanced controller is the one for Pan-STARRS developed by Peter Onaka and colleagues at the University of Hawaii. That system has been specifically developed to operate multiple orthogonal transfer CCDs in a huge mosaic camera. The controller technology is also based on FPGAs, gigabit ethernet links, and LINUX PCs. Lastly, developments in Europe are under way to produce a superfast controller specifically for use with Electron Multiplication CCDs in which the horizontal clocking in particular needs to run at >15 MHz per pixel with a 45-volt swing. Electronics for this application of CCDs is closer to radio frequency (RF) circuitry where detailed attention must be paid to cable lengths and impedances. Next-generation AO cameras using EMCCD technology will be looking for sub-electron noise at kilohertz frame rates. In fact, Craig Mackay (Cambridge, U.K.) has already developed his own controller for EMCCDs. His system uses a high-speed clock sequencer on a chip made by Kodak (the KSC1000) for clock waveform generation, the ARM processor as used in cell/mobile phones for system control, and USB links for command/control and data acquisition. ESO is also developing a next-generation controller that is also FPGA-based.

Future CCD controllers are likely to be reduced to a single application-specific integrated circuit (ASIC) not much different in size from the CCD itself. Devices like this are already becoming available for large-format infrared arrays where the complexity of the electronics is increased by the need for multiple (typically 32) readout

channels per detector. This development is particularly important for large cameras on future space missions. For example, a group at the Rutherford Appleton Lab (U.K.) has developed a space-qualified four-channel CCD controller on an ASIC. Likewise, the group working on SNAP (the SuperNova Acceleration Probe) at the Lawrence Berkeley Lab (U.S.A.) is also developing ASIC-based controllers for use with the large CCD focal plane that is planned for that mission. On an ASIC, mixed analog and digital signals can be combined, but at the cost of expensive fabrication processes. FPGAs at present can only be used for digital signals and are much cheaper. Very compact space-qualified CCD controllers have been developed for the GAIA focal plane (Laborie *et al.*, 2007) that consists of over 100 CCDs. Each controller sits very close to its CCD, occupying a space of only $3 \times 6 \times 11$ cm, generates all the clocks and biases, and processes all the video signals for that chip. Teledyne Imaging Sensors has developed an ASIC (the SIDECAR) to control the *near-infrared* arrays for the planned James Webb Space Telescope, and ground-based versions are now available. At present, the cost is comparable with a standard ARC/SDSU controller. ASIC systems have more risk compared with FPGA systems, unless large partnerships are formed and the vendor assumes more responsibility for getting things right. I am indebted to Derek Ives (U.K. ATC) and Mark Hunten (NOAO) for their input on these topics.

8.2 DARK CURRENT AND COOLING

Random (Brownian) motions of atoms at normal room temperatures within the silicon lattice will release sufficient energy to give rise to a continuous stream of electron–hole pairs in the absence of light. This process for electron–hole pair production is called "thermal" because the energy source is heat and the "dark current" produced can be very substantial. At room temperature the dark current of a typical, non-inverted CCD is about 100,000 electrons per second per pixel (or equivalently, 1.8×10^{-9} A/cm^2 for 30 μm pixels) which means that the CCD storage wells will fill up, saturate, and spill over on dark current alone in just a few seconds. If the CCD is read out very rapidly and continuously at a high rate (frequency) such as 5 MHz, then the dark current is cleared after only a brief accumulation time (about 1/60th of a second) and so the CCD can be used for ordinary TV applications. Most astronomical sources are much too faint to yield a good image in such short snapshots, especially as there is a small but definite penalty of increased noise for every read-out. Fortunately, the solution is straightforward. The CCD needs to be cooled to a low temperature. Or at least that solution seems straightforward now, but when it was first suggested to Bell Labs engineers by Jim Westphal back in 1974 it was considered radical!

Thermal dark current undergoes a very rapid decay with decreasing temperature, and therefore a drastic reduction in dark current can be achieved by cooling the CCD to a temperature somewhere between those of "dry ice" (solid carbon dioxide or CO_2) and liquid nitrogen (LN_2). For charge-coupled devices there are three main sources of dark current. These are

- thermal generation and diffusion in the neutral bulk silicon
- thermal generation in the depletion region
- thermal generation due to regions called "surface states" at the junction between the silicon and the silicon dioxide insulation layer.

Of these sources, the contribution from surface states is the dominant contributor for multi-phase CCDs. Dark current at the $Si–SiO_2$ interface depends on two factors: the density of interface states and the density of free carriers (electrons or holes) that populate the interface. Electrons can more easily thermally "hop" from the valence band to an interface state (also called a mid-band state because the energy level associated with this disturbance in the crystal lattice lies in the forbidden energy gap of the normal semiconductor crystal) and then to the conduction band, thus producing an electron–hole pair. The presence of free carriers will fill interface states and, if the states are completely filled, will suppress thermal hopping (valence band electrons no longer have a stepping stone to the conduction band) and therefore substantially reduce dark current. Classical CCD operation as described in Section 7.2.1 actually depletes the potential well and the interface of free carriers, thus maximizing dark-current generation. Under depleted conditions, dark current is determined by the quality of the silicon–silicon dioxide interface. As explained in Section 7.3.6, dark current can be significantly reduced by operating the CCD in "inversion" mode. With this arrangement, holes are attracted up to the $Si–SiO_2$ interface at these electrodes, thus populating the troublesome states with holes. Not only does inverted operation improve dark current, but the neutralization of the interface states eliminates "residual" images associated with saturation and problems of charge capture. The inversion phenomenon was noticed by Tony Tyson of Bell Labs (now at University of California, Davis) while investigating RCA CCDs, but its real significance was not appreciated until a similar effect was uncovered later by Jim Janesick and colleagues at JPL for the Texas Instrument devices. Virtually all CCDs which are capable of inverted operation are now run in this manner, and new CCDs are manufactured with multi-pinned phase (MPP) technology which allows the CCD to operate totally inverted at all times while maintaining other performance characteristics. MPP CCDs have achieved dark levels less than $0.025\,\text{nA/cm}^2$ at room temperature and thinned backside-illuminated devices, with a substantial part of the silicon etched away, have attained dark currents of $0.010\,\text{nA/cm}^2$. MPP technology also aids charge transfer efficiency by allowing higher operating temperatures to be used.

Roughly speaking, in silicon the dark current $N_{dk}(T)$ falls to about one-third of its previous value each time the temperature (T) is reduced by about 10°. More precisely, the relation can be expressed as

$$N_{dk} = 2.55 \times 10^{15} N_0 d_{\text{pix}}^2 T^{1.5} e^{-\frac{E_G}{2kT}} \quad \text{electron/s/pixel} \tag{8.1}$$

where N_{dk} is in electron/s/pixel; and N_0 is the dark current in nanoamps per cm^2 at room temperature (T_0); d_{pix} is the pixel size in centimeters; T is the operating temperature in K; E_G is the bandgap energy in electronvolts; and k is the Boltzmann

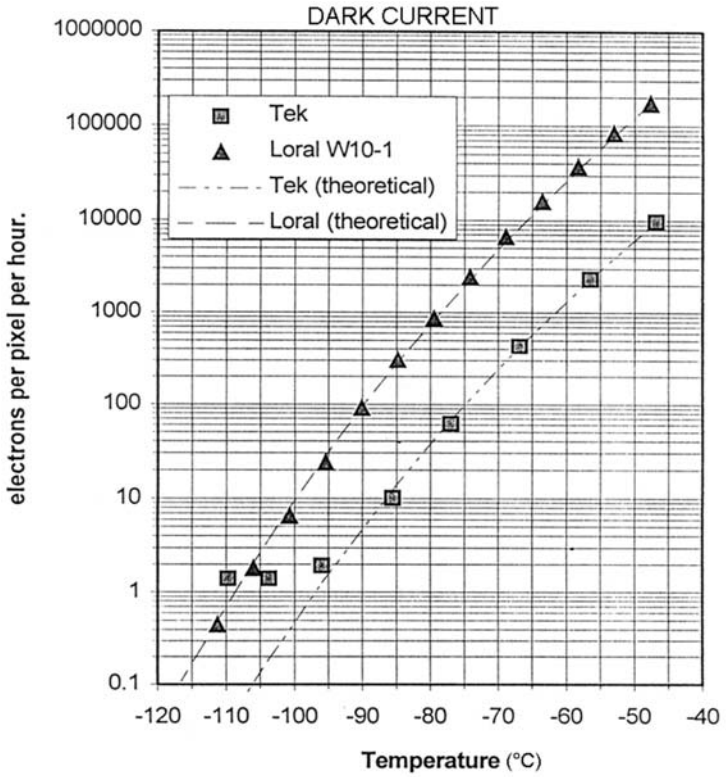


Figure 8.6. A plot of dark current vs. temperature for two CCDs. Credit: Jim Janesick.

constant (8.62×10^{-5} eV/K). The bandgap energy (in electronvolts) varies with temperature and is given by

$$E_G = 1.1557 - \frac{7.021 \times 10^{-4} T^2}{1{,}108 + T} \tag{8.2}$$

Typical operating temperatures for astronomical CCDs range from about 223 K to 150 K (or -50°C to -123°C), depending on the application and whether or not the CCD employs MPP. It is possible to end up with a dark current of only a few electrons per pixel per *hour* instead of 100,000 electrons per second! Two devices are compared in Figure 8.6.

Example. The room temperature (300 K) dark current of a MPP CCD with 12 μm pixels is found to be 0.01 nA/cm^2. Predict the dark current at -50°C.

Solution. First find the bandgap for this temperature ($T = 223$ K) using Equation (8.2), $E_G(223) = 1.1295$ eV. The pixel area is 1.44×10^{-6} cm^2. Substitute

in Equation (8.1) to find that $N_{dk} = 2.12 \times 10^{-2} e$/s/pixel, or about 76 electrons/pixel per hour.

8.3 COSMIC RAYS AND LUMINESCENCE

With the elimination of dark current, cooled CCDs become capable of extremely long exposures. However, there is now another effect which comes into play. High-energy sub-atomic particles (protons) entering the Earth's atmosphere from outer space will generate a shower of secondary particles called muons which can be "stopped" by a sufficiently thick layer of silicon. The energy released can generate around 80 electrons per pixel per micron of thickness in the silicon. With a collection depth of 20 μm, a muon event is seen on a CCD image as a concentrated bright spot a few thousand electrons strong. Thinned backside-illuminated CCDs and those CCDs with epitaxial (i.e., thin) collection layers (like the *e2v* CCDs) are therefore much less prone to cosmic rays than the thicker frontside-illuminated or deep-depletion arrays. High-altitude observatories may also encounter more problems than sea-level laboratories from cosmic-ray hits.

For CCDs used in space the problem is even more severe due to the high flux of primary ionizing particles themselves. For example, the Galileo CCD camera would have arrived at Jupiter in a saturated state if neutron testing had not been carried out before launch to reveal very large dark spikes in the array. The problem was solved by operating the CCD at -120°C instead of the planned -40°C. Likewise, the operating temperature of the Cassini and Hubble CCDs was selected to be -90°C to lessen the effect of traps induced by proton impacts. Unfortunately, the CCDs on the Chandra X-ray Observatory experienced a proton-related radiation problem that seriously degraded charge transfer efficiency soon after launch. These topics are discussed in detail in Janesick (2001). Briefly, there are two major types of damage called ionization and bulk damage. Permanent damage can occur in the gate dielectric (e.g., SiO_2) layer when electron–hole pairs are formed there by ionizing radiation. Holes generated in the dielectric layer become trapped generating a voltage shift to the clock and output amplifier bias levels. A second problem occurs at the Si–SiO_2 interface where ionizing radiation can break the weak bonds and create interface states, which in turn can dramatically increase the dark current. Bulk damage occurs when energetic particles collide with silicon nuclei. Such events displace atoms from their positions in the crystal lattice and create vacancy–interstitial pairs. This kind of damage causes a permanent degradation of charge transfer efficiency, increased dark current, dark-current spikes, and effects on the output amplifier.

Because cosmic-ray events are so conspicuous and concentrated to a few pixels in ground-based cameras, it is generally straightforward to remove them from images of non-crowded fields and non-extended objects. Otherwise, the only approach is to take multiple frames of the same region and compare them, as there is very little chance of a cosmic-ray event hitting the same pixel two times in quick succession. Figure 8.7 is the "difference" of two long exposures showing multiple hits in each exposure; when a difference image is formed the bright spots in the second frame appear black.

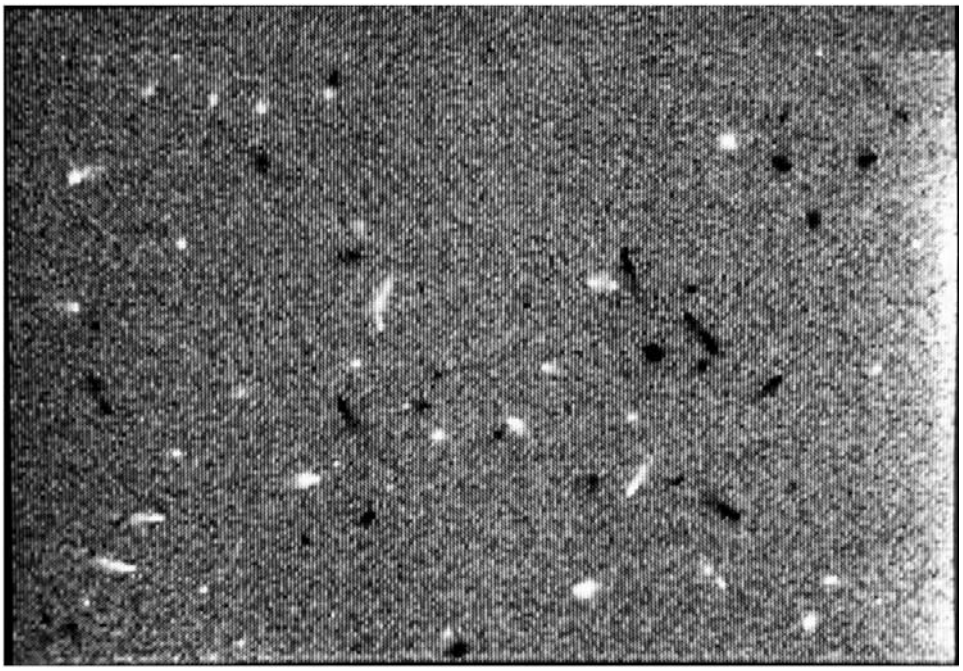


Figure 8.7. The difference of two 90-minute dark exposures with a CCD reveals numerous events even with no illumination; these are ionizing particles or high-energy photons. The difference image shows the events in the second exposure in black. Credit: NOAO.

Early experiments conducted by Craig Mackay of the Institute of Astronomy in Cambridge (U.K.) revealed that the events attributed to cosmic-ray hits were actually a combination of cosmic rays and low-level X-ray emission from several forms of UV and blue-transmitting glass (e.g., Schott UBK7 and GG385). In fact, the X-ray emission dominated by 5 : 1. This discovery alleviated the concern over the fact that the muon event rate was predicted by cosmic-ray physicists to be several times lower than the observed rate of events being recorded by CCDs. With fused silica windows the actual event rate was only 1.5–2 events/cm^2/min, or 90–120 events/CCD frame/hour exposure.

Another source of unwanted signal in all CCDs and similar devices, such as solid-state infrared imagers, is luminescence or glow. An electrical path with lower-than-normal resistance in the semiconductor caused by, for example, partial shorts between electrodes in the parallel or serial registers, can act like a light-emitting diode (LED). Incorrect voltages applied to certain parts of the device, in particular the output transistor amplifier, can cause similar problems. This effect can be a serious limitation in some CCDs because it will limit the integration time to only a few minutes. Device selection or, alternatively, active (computerized) electronic control of the applied voltages can circumvent this limitation. The simplest example of the latter approach is to electronically switch off the output transistor during a

long integration and to reduce the clock level (or voltage potential) on the pixels. Just prior to the end of the exposure the transistor is switched on again, the clock swing increased to improve charge transfer efficiency, and the horizontal register "purged" several times to clear it before the CCD is read out.

8.4 BAD PIXELS AND CHARGE TRANSFER EFFICIENCY

Charges are accumulated under the CCD electrodes until the end of the timed exposure and then moved out by charge-coupling of one pixel to the next through repetitive clocking (or pulsing) voltages applied to the electrodes. Unfortunately, as recognized immediately by the inventors Bill Boyle and George Smith, there are several possible impediments to the success of the transfer process. Problems occur when the wells are (a) virtually empty, (b) almost full, or (c) when a defective pixel gets in the way.

As the CCD wells fill up, the total stored charge can distort the electric potential of the applied clock voltage, and thus charge transfer or coupling is degraded. Moreover, when the wells are completely full, a condition called "saturation", there is nothing to stop the charge spilling over into the adjacent pair of pixels in that column; in general, there is no serious sideways (row) spread because of the heavily doped channel stops. If the source of light causing the saturation is very strong then the adjacent pixels may also fill up, and so the spreading continues and a streak appears up and down that column; this is also referred to as "charge bleeding". An example is shown in Figure 8.8.

Global charge transfer efficiency (CTE) drops when a fixed amount of charge is "trapped" for *each* pixel transfer. Visually, the result is the appearance of a tail on point-like images; the tail will be parallel to the transfer direction. Trapped charge is not normally lost, but more often it is "skimmed" to be released later and is therefore related to a general phenomenon called "deferred charge" (as illustrated in Figure 8.9). Deferred charge can have several origins including spurious potential pockets, design faults leading to specific potential pockets at certain locations, and charge traps associated with impurities. For most modern CCDs, poor charge transfer efficiency occurs at a level of $\sim$10 electrons/pixel and is mainly of concern only in very short exposures or in spectroscopic applications in which the dispersed background in the exposure is below 100 electrons. One way around deferred charge effects is to "pre-flash" the CCD with a short flood of light from an LED positioned so as to uniformly illuminate the array before each exposure. This action adds a constant amount of charge, sometimes called a pedestal, to each pixel which raises the base level to above 100 electrons (say) back into the acceptable region of good transfer efficiency; note that a penalty of $\sqrt{100} = 10$ electrons noise is added to the readout noise. The technique of introducing a substantial charge pedestal was termed (by someone with a sense of humor) as adding a "fat zero"; inevitably, someone nicknamed the smaller charge pedestal needed to offset the deferred charge problem the "skinny zero". In astronomical circles the most common description is pre-flash.

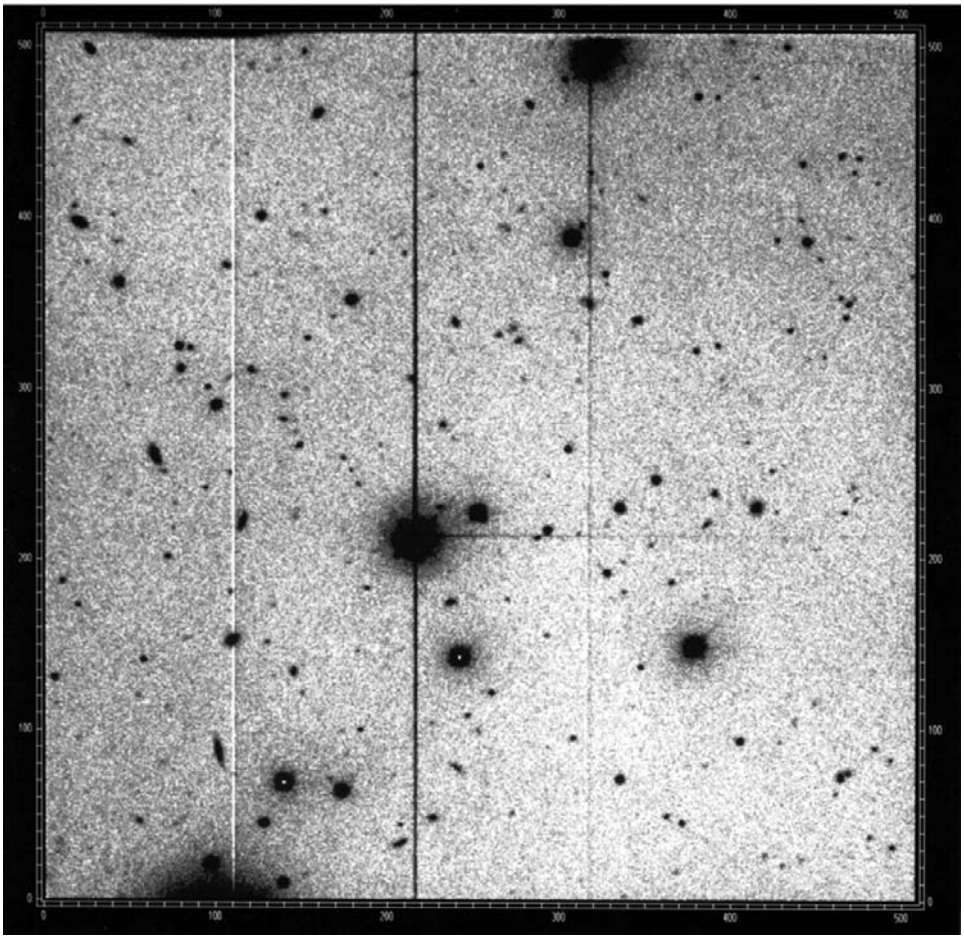


Figure 8.8. An astronomical CCD image (displayed as a ''negative'') in which a bright star has saturated. When the CCD pixel saturates it spills over and ''bleeds'' up and down the column. The horizontal trail is evidence of poor charge transfer efficiency in the serial register. The white line is a non-working or ''dead'' column. Credit: NOAO.

Apart from these general problems of charge transfer inefficiency there are other difficulties related to specific faults on a given CCD. A break in one of the polysilicon electrode strips due to a fault in the manufacturing process can effectively stop or seriously delay charge from being transferred past that point in that column; this defect is called a ''blocked column''. Crystal defects in the bulk substrate silicon can sometimes spread into the depletion region during manufacture. Such defects disrupt the semiconductor nature of the silicon, and in general behave like sponges which can soak up any charges in their area. Delayed release of charge can have a wide range of time constants (from milliseconds to hours) and columns with this behavior are

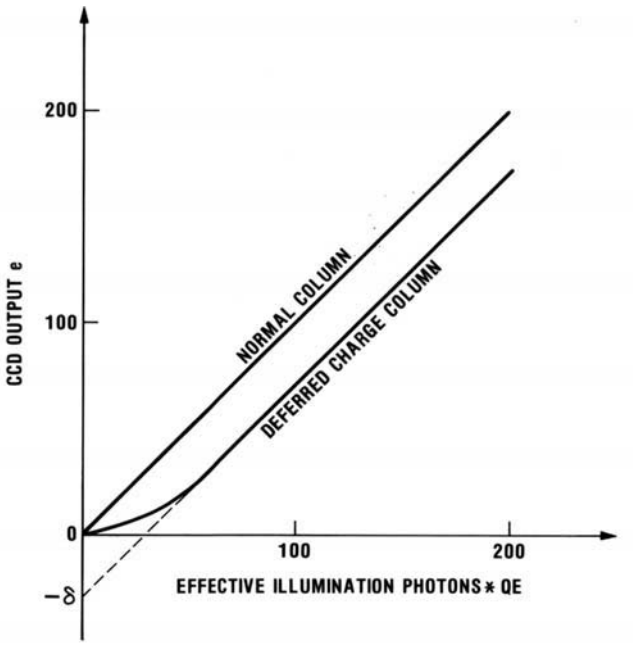


Figure 8.9. A plot of the number of electrons detected at the output of a CCD vs. the number of electrons present in the original pixel. If there is no trapped charge anywhere during the readout, then the graph should follow that for a normal column. It does not, indicating trapped charge whose readout has been "deferred". Credit: Tony Tyson.

usually irretrievable by image processing. Needless to say, a charge trap near the beginning of the horizontal (serial) register is a disaster! Even the design of the overlapping electrode structure can create pockets of resistance in the channel potential which trap charge. Trapping sites are particularly serious for surface-channel CCDs. Using our water-in-a-well analogy, traps would behave like small potholes at the base of the well; such pockets fill up easily but do not empty their contents merely by raising and lowering the floor of the well. Today, the "yield" of commercial buried-channel devices free from all these problems is high.

CTE is quoted in the form 0.99999 per transfer or, in CCD jargon, people speak of CCDs with "five nines of CTE". What this means is that after one transfer 99.999% of the original charge will have made it successfully, after the second transfer 99.999% of that charge will be moved on so that 99.998% of the original charge is still intact, and finally, after 1,000 transfers, 99.00% (i.e., 0.999991000) of the original charge on the initial pixel will have been transferred to its new location (the output, say), while 1% of the original signal will be spread over the intervening pixels. In practice most of that 1% is in the immediately adjacent (following) pixel. We have already mentioned that CTE is a function of temperature. It also depends on the clock frequency (or "readout rate"), the rise-and-fall times of clock voltage pulses, and their degree of overlap. The CTE in the parallel (column) direction is

usually different from that in the serial (horizontal) register and there can be large variations from column to column.

Measuring the level of charge transfer efficiency is not as simple as it might seem at first sight. A relatively straightforward method, which is both qualitatively and quantitatively useful, is to flood the CCD with light, read out the device, but "overscan" both registers to produce a resulting image which is several pixels larger, along both axes, than the real CCD. Most CCD controllers permit overscanning, but some CCDs have special additional pixels in the serial register which confuse the overscan experiment. This technique is also called "extended pixel edge response" or EPER. When the signals occurring in the imaginary pixels following the last row and column of the real CCD are compared with the dark level and with the illumination level then any charge transfer problems leading to deferred charge will be obvious. Some investigators lump all rows or columns together to give a mean measure of CTE, but careful examination of an overscanned uniformly illuminated image reveals variations from column to column. Some averaging will be needed to reduce the effect of readout noise. The CTE is then given by

$$CTE(e^-) = 1 - \frac{Q_d}{NQ_0} \tag{8.3}$$

where Q_d is the net deferred charge in the overscan; Q_0 is the charge on the last real pixel; and N is the number of pixels transferred. An example of EPER at extremely low light levels is shown in Figure 8.10.

The spurious potential pocket problem plagued most manufacturers early on, but it is now well-enough understood. One example is the isolated single-pixel trap or "pocket" which affects CTE randomly on a local rather than global level. Some CCDs were literally riddled with pockets. Detailed tests showed that these traps are localized to a single level of the polysilicon electrode within a pixel. This was the phenomenon which put the Tektronix (SITe) devices on hold for a time. Two clues finally emerged. The TI family of 800×800 CCDs showed no large single-pixel traps and it was discovered that the impurity boron, used to dope the p-type substrate, had diffused through the thin (epitaxial) n-type layer of the buried channel construction during a high-temperature phase of manufacture. This effect was called the "p^+ epitaxial diffusion tail". The Tek CCDs, on the other hand, exhibited very little diffusion of boron. The second clue was the realization that the change from the older practice of using silicon dioxide as the insulating layer on the semiconductor surface to an oxide nitride double-layer could be relevant; the new material had significantly reduced short-circuits between the polysilicon electrodes and therefore improved manufacturing yield. Pinholes created by surface contamination during the oxide-growing stage could allow nitride to diffuse through the oxide to the n-type semiconductor. Amazingly, at that tiny location the different insulator material would cause a slight variation in electric potential within the semiconductor.

Ultimately, with all other problems solved, it is traps in the bulk silicon crystal that limit charge transfer efficiency. Fortunately, the quality of manufactured semiconductor silicon has greatly improved over the years and will probably continue to do so. Global CTE performance of selected CCDs are believed to be limited by "bulk

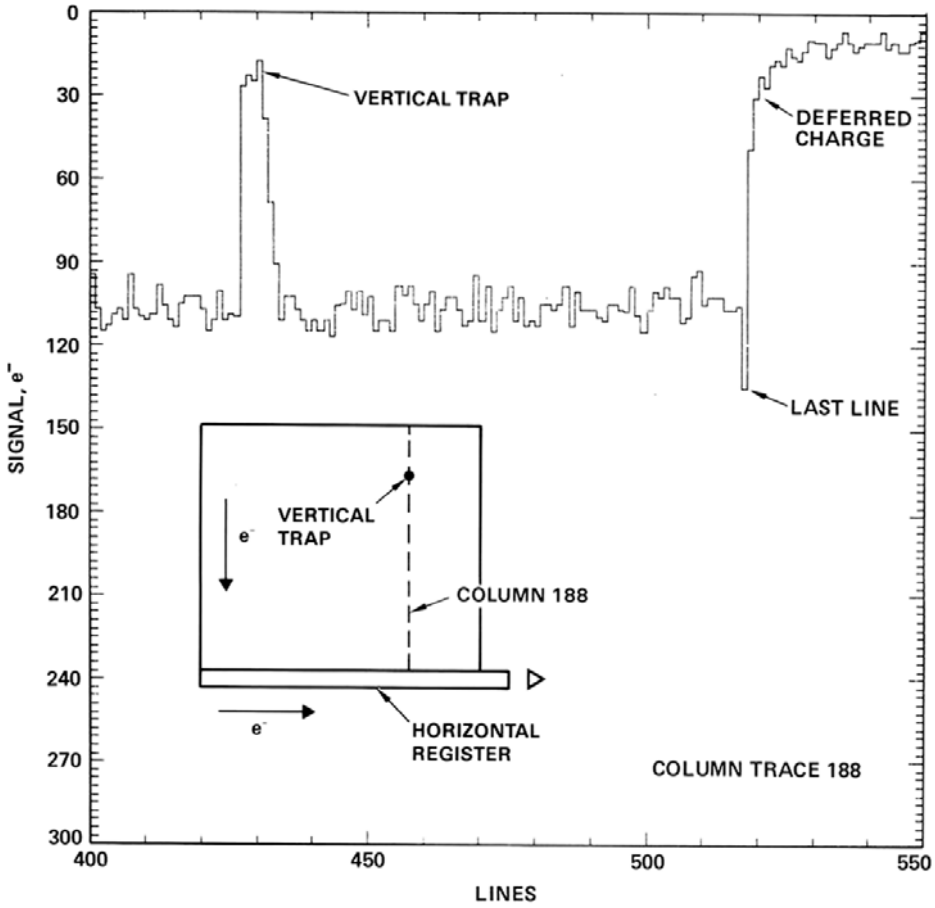


Figure 8.10. A plot of the signal (in electrons) along a particular column of a CCD showing the loss of charge in a "trap", the curved response following that trap, and the last pixel indicating deferred charge. Credit: Jim Janesick.

traps" at the level of 0.9999995 (or "six nines five") per pixel transfer. In other words, only 3 electrons out of a total of 10,000 electrons are deferred for 512 pixel transfers.

8.5 NOISE SOURCES

A CCD is by its very nature a digital imaging device. The clocking procedure described earlier delivers a stream of charge "packets" from pixels in the image area all the way to the output amplifier; the charge (Q) in each packet is proportional to the amount of light in that part of the original image scene. As each charge packet arrives at the output field effect transistor it causes a change in voltage to occur (of

amount $V = Q/C$, where C is the capacitance at the output node); the smaller the node capacitance, the larger the voltage change for the same size of charge packet. For the earliest CCDs the output capacitance was fairly high (e.g., $C \approx 0.6$ picofarads or pF which yields about 0.25 microvolts or μV per electron in the charge packet), whereas for modern CCDs the node capacitance is <0.1 pF which gives a healthy >16 μV per electron. Much larger values are possible; some Kodak CCDs using an extremely small output MOSFET give 15 μV/electron, yet the overall noise under slow-scan conditions is greater than 10 electrons because other noise sources become larger as the MOSFET gets smaller.

It is desirable that the noise performance of a CCD camera system be limited only by the output transistor of the CCD and not by any other part of the electronic system. To achieve this goal, one must understand the noise sources associated with the CCD and take steps to get them to an irreducible minimum; this minimum is the ultimate "readout noise" (R), usually quoted as the root-mean-square noise in electrons. There are several potential sources of unwanted electronic noise. These include

- background charge associated with fat-zero offsets
- transfer loss fluctuations
- reset or kTC noise
- MOSFET noise
- fast interface state noise.

When a pre-flash is used to introduce a fat-zero charge to aid transfer efficiency or eliminate charge skimming, the consequence is a noise equal to the square root of the total number of charges in a pixel.

During charge transfer a fraction of the charges are left behind. However, this fraction is not constant but may fluctuate, and so an additional noise component is added to the signal noise. This "transfer noise" is given by

$$\sigma_{\text{tr}} = \sqrt{2\varepsilon n N_0} \tag{8.4}$$

where $\varepsilon = 1 - \text{CTE}$ is the fraction of charges not transferred; n is the number of transfers; and N_0 is the original charge. The factor of 2 occurs because the Poisson-distributed noise happens twice, once for trapping and once for release. This effect can be of order 70 electrons for surface-channel CCDs, but is typically ten times smaller or better for buried-channel CCDs and astronomical light levels. For very large CCDs this effect implies that exceptionally good charge transfer efficiency must be achieved.

Noise associated with the recharging of the output node is given by $\sqrt{(kTC)/e}$ or about $284\sqrt{(C)}$ at a CCD temperature of 150 K, where the capacitance, C, is in picofarads. This effect is called "reset noise", and it is the dominant source in most cases.

Other noise sources associated with the output MOSFET, such as one-over-f noise (i.e., noise that varies roughly as 1/frequency), can generally be made quite

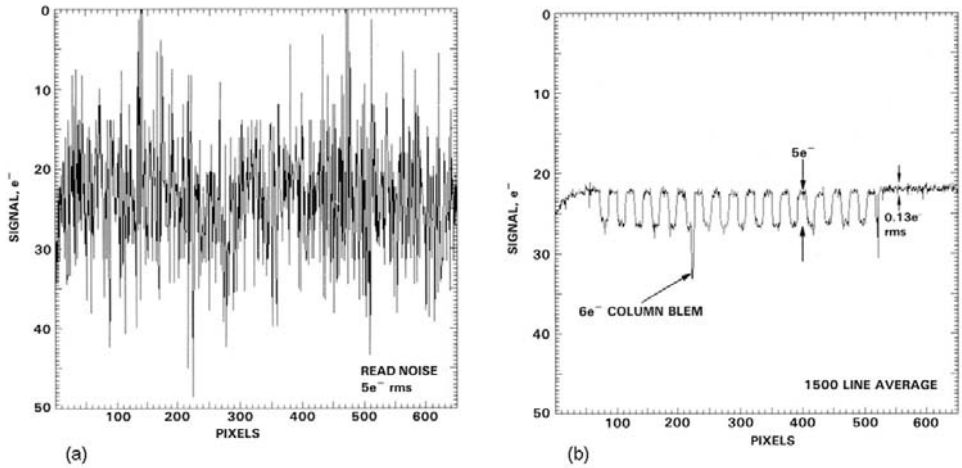


Figure 8.11. (a) A single raw trace of the signal from a Loral CCD at the five-electron level. Embedded in the noisy trace is an unseen 5*e* peak-to-peak square-wave pattern. (b) After 1,500 lines have been averaged the random noise is only 0.13*e* as seen in the overscan region. A 6*e* column blemish (BLEM) has also emerged. Credit: Jim Janesick.

small by good manufacture, typically a few electrons. Traps that absorb and release charges on very short time scales, thereby causing a fluctuation in the charge in any pixel, are called "fast interface states". In this case the noise is given by

$$\sigma_{ss} = \sqrt{2kTnN_{ss}A} \tag{8.5}$$

where k is Boltzmann's constant; T is the absolute temperature; n is the total number of transfers (not the number of pixels); N_{ss} is the surface density of traps; and A is the surface area. This effect is very serious for surface-channel CCDs but is normally quite small (of order 5 electrons or less) for good buried-channel devices. It is a remarkable tribute to the foresight of Boyle and Smith, and the skills of every scientist and engineer who has worked hard on CCD technology, to realize that devices are now routinely made with a readout noise of less than 5 electrons (see Figure 8.11a, b). The lowest noises that have been consistently measured (without electron multiplication) are in the range 2–3 electrons.

8.6 SIGNAL PROCESSING AND DIGITIZATION

Here, we consider CCD signal processing in more detail, and in particular explain the crucially important technique known as "correlated double-sampling" or CDS. As each charge packet arrives at the output node it produces a voltage change which must first be amplified and then digitized by an analog-to-digital (A/D) converter. This process is not instantaneous; it requires a finite amount of time and hence the term "slow-scan". A high degree of accuracy is required, such as can be achieved with

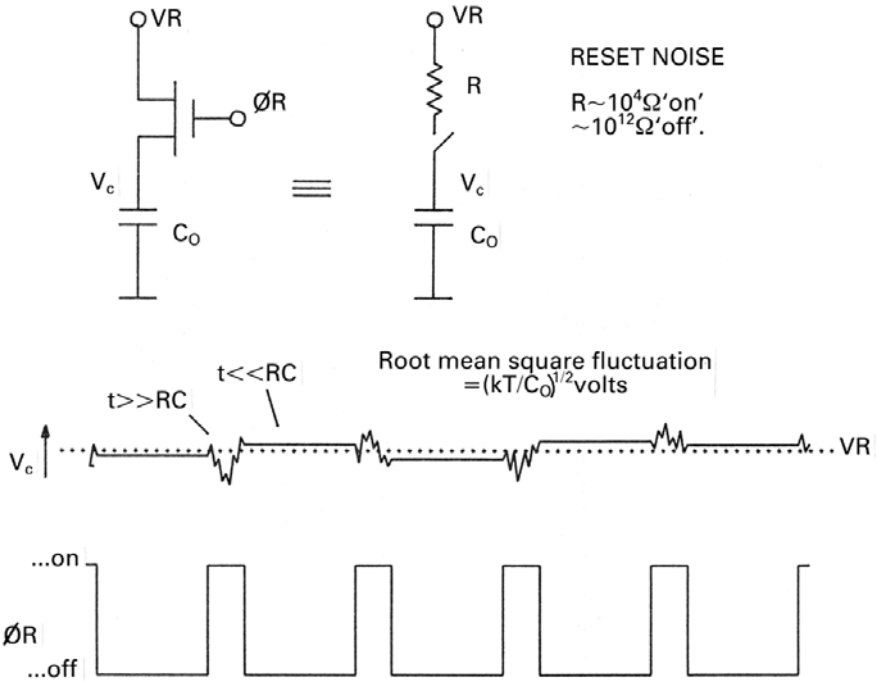


Figure 8.12. An equivalent "switch" circuit to explain the operation of the reset transistor. The resistance, and therefore the RC time constant, is very different between off and on states.

a 16-bit A/D converter; a 16-bit A/D converter divides a specified voltage range, typically 10 volts, into 65,536 (2^{16}) parts and therefore each voltage interval is 152.5 μV in size. The A/D circuit matches up the actual voltage to the nearest number on the scale of 0–65,535.

To measure the voltage of each charge packet we need a "reference" voltage. We could use ground, but it is important to reset the output capacitance back to some nominal value on each readout cycle, otherwise we would be forming the difference between one charge packet and the previous one while drifting away (in voltage) from the ideal operating point of the MOSFET. There is another way. The output capacitor can be recharged to a fixed voltage by briefly pulsing the gate of another transistor, called the reset transistor (see Figure 8.12), to briefly turn that transistor on (like closing a switch) so that current can flow from a power supply to charge up the node to the supply level. When the reset pulse disappears the reset transistor is turned off (like opening a switch) and the output becomes isolated to await the next charge-transfer from the horizontal register. As a capacitor is charged to a certain voltage level (V_{RESET} or V_{RD}) it does so exponentially, rising steeply at first and then leveling off to approach its final value as shown in Figure 8.13. Again, due to random thermal agitation of electrons in the material, there is "noise" or uncertainty on the mean value and so the final voltage can lie anywhere within a small range,

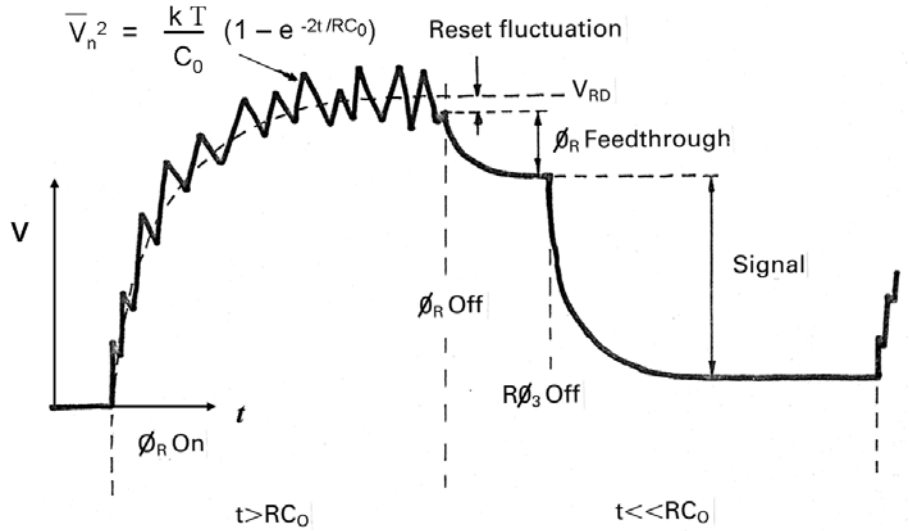


Figure 8.13. The charging profile of the output of a CCD when reset. The curve is "noisy" but when the reset pulse disappears the last value of the signal becomes frozen.

given by

$$\text{reset noise} = \sqrt{\frac{kT}{C}} \text{ volts or } \frac{\sqrt{kTC}}{e} \text{ electrons} \tag{8.6}$$

In this expression, k is Boltzmann's constant (1.38×10^{-23} joules/K); e is the charge on the electron; T is the absolute temperature of the output node in degrees Kelvin (K); and C is the node capacitance. If C is expressed in picofarads (i.e., 10^{-12} coulombs/volt), then this noise uncertainty (called the "reset noise" or from the formula "kTC noise", pronounced kayteecee) is simply $400\sqrt{(C)}$ electrons at room temperature and about $250\sqrt{(C)}$ at 120 K (−153°C); for a typical modern device this would yield <80 electrons noise, which greatly exceeds the readout noise of the MOSFET alone and so some means must be found to remove it.

8.6.1 Correlated double-sampling

Fortunately, removal of reset noise is quite straightforward due to the fact that whatever the final reset voltage actually is, and it must be in the range $(V_{\text{RESET}} - \sqrt{(kT/C)})$ to $(V_{\text{RESET}} + \sqrt{(kT/C)})$, it will get "frozen" at that value because the leakage of current through the switched-off reset transistor is exceedingly slow (its "RC time constant" is seconds compared with the microseconds between arrivals of discrete CCD charge packets). Hence, if this uncertain reset level is sampled by the A/D just prior to a charge packet being dumped at the output, and then again after the charge has been added, it will have *exactly* the same value in each sample. Forming the difference of these "before and after" signals automatically eliminates this voltage level without ever knowing exactly what it was. This

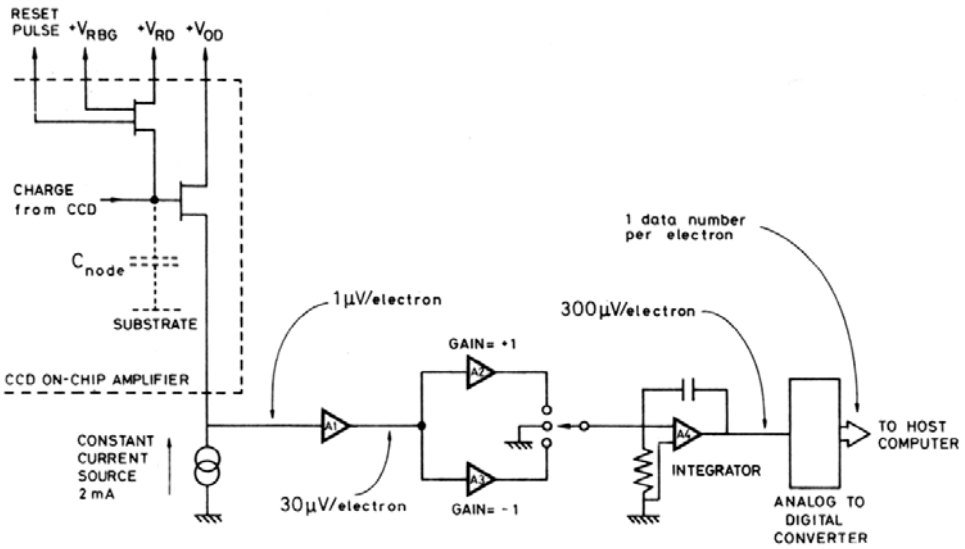


Figure 8.14. A block diagram of the principle of the correlated double-sampling (CDS) method of removing reset noise before the signal is digitized and sent to a computer. Credit: Craig Mackay.

technique is known as correlated double-sampling (or CDS) because two samples of the CCD output voltage are taken per pixel and the offset due to the reset noise in each sample is the same (in other words, it is correlated).

In practice it is advantageous to integrate or sum the signals rather than merely spot-sample them. Integration smoothes out high-frequency noise and limits the effects of $1/f$ noise for frequencies lower than the overall signal-processing time. A schematic version of how a full, integrating CDS circuit works is depicted in Figure 8.14. Signals from the CCD output transistor are amplified by a low-noise solid-state operational amplifier (A1 in the diagram) and fed to the integrating amplifier A4 (with the capacitor in its "feedback loop" from the output to the input) by way of a three-position switch. One of the three positions is ground, and the other two contain a *non-inverting* and an *inverting* amplifier, respectively, each with a gain of $\times 1$ (in other words, in one case the signal voltage is multiplied by 1 and in the other case by -1). Amplifier A1 is connected to A4 via A3 (times -1 gain) for a precise period of time (e.g., 20 μs), during which the reset voltage level is being integrated (negatively) onto the capacitor on A4. Next, the switch is connected to ground for a few microseconds which isolates the integrator and allows the signal charge packet to be transferred from the serial horizontal register onto the CCD output node. Then the switch is thrown to the non-inverting amplifier A2 ($\times 1$ gain) and the "reset plus voltage" level is integrated (positively) onto the same integrating capacitor on A4 for exactly the same period of time. As A2 and A3 give identical—but opposite—signals in identical times, the resulting output of A4 is the real *difference* signal (i.e., the signal due to the charge packet itself). Reset noise is eliminated! The output of the

integrator is fed to a sample-and-hold circuit which "freezes" the value and then to a fast analog-to-digital converter (often in the same commercial package) to convert the voltage to a number. The parallel output (typically 16 wires) of the A/D converter goes directly to the data input part of a computer, or is converted to a very high speed serial signal sent over an optical fiber transmission line to the computer.

Choosing the correct amplifier gains can be important. To work at very low signal levels, the total system gain should be such that the root-mean-square (rms) readout noise corresponds to a few "bits" (also called data numbers or DNs, and ADUs) of the A/D converter. The maximum or saturation signal from a CCD is less than 500,000 electrons, but some devices have a readout noise of only 5 electrons; the ratio of these two numbers is called the "dynamic range". In this case the dynamic range is 100,000 : 1 whereas a 16-bit A/D can only give a range of 65,536 : 1. If we choose a gain to give 1 electron per DN then the readout noise will be 5 DN and the A/D will saturate at 65,536 DN, which is only 13% of full well. If we use 5 electrons per DN then the readout noise will be 1 DN (and not well-determined due to digitization noise) but we can utilize 65% of the well. A more realistic situation is that the full-well capacity is less than 250,000 electrons, which gives a maximum dynamic range of 50,000 : 1 which is within the 16-bit A/D range of 65,536 : 1. If the application is such that the photon flux nearly fills the CCD wells, then the dynamic range will be limited by photon noise—and not readout noise. For example, on a signal of 250,000 photoelectrons the photon noise is $\sqrt{N}$ or 500 electrons and the dynamic range is $250{,}000/500 = 500$—and not 50,000. As most CCD applications in astronomy are for low-light levels then it is better to use a high gain and work only in the lower part of the wells (e.g., 5 electrons/DN in the last example yields a readout noise of 2 DN and the A/D saturates at 65% of full well). A 10-volt 16-bit A/D has a resolution of 152.6 μV per DN or 30.5 μV per electron if 5 electrons/DN are required. A typical CCD output yields only 2 μV per electron, and so an amplifier gain of 15 is needed to match up to the A/D. It is possible to install variable amplifier gain settings, but any method of changing the gain introduces another calibration step for astronomers.

Digitization noise can be shown to be about $0.289g$, where g is the gain in electrons/DN (Merline and Howell, 1995). In the above example, with $g = 5$, this corresponds to a noise of 1.45 electrons to be added in quadrature with the readout noise. Thus, the readout noise term should become $[R^2 + (0.289g)^2]$.

The CDS integration and digitization time will determine the pixel readout rate, typically this is 10 kHz–50 kHz. In principle, the longer the processing time spent on each pixel the better the final noise performance. The CCD output noise from the CDS is inversely proportional to S_V, the sensitivity of the CCD output node (μV/e), and proportional to the CDS-filtered noise voltage spectrum $V_n(f)$ of the on-chip amplifier; $V_n(f)$ has two components: a white noise (independent of frequency) and a $1/f$ term. Readout noise will decrease approximately as the square root of the CDS integration period until correlation is lost by $1/f$ noise. A plot of readout noise vs. CDS integration time for three CCD types is shown in Figure 8.15 from Jim Janesick. With a CDS integration time of 8 μs the Lincoln Laboratory CCD gave 2 electrons noise. Typical pixel times range from about 20 μs for many CCDs to about 80 μs for

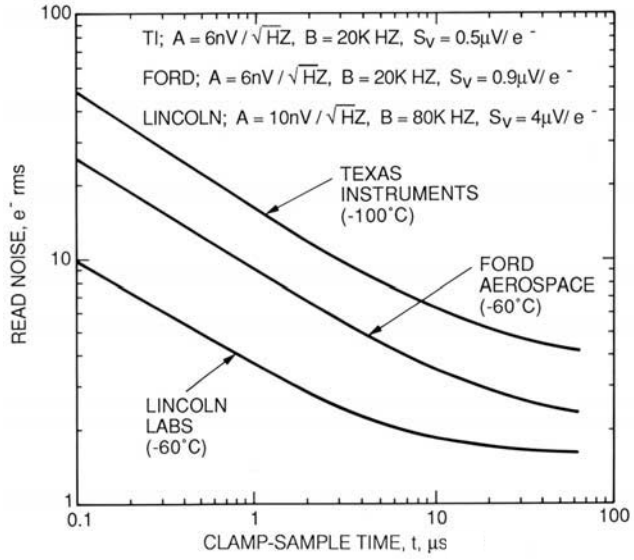


Figure 8.15. Theoretical readout noise performance of three CCD types as a function of CDS integration or clamp-to-sample time. The plots were derived from an equation but were found to be in good agreement with experimental data. Credit: Jim Janesick.

devices from *e2v technologies* which have very low $1/f$ noise. For a $2{,}048 \times 2{,}048$ CCD with a 20 μs pixel period the total readout time is 80 s and for 50 μs it takes 3 minutes and 20 seconds to read the entire chip!

Low-noise performance depends on the use of very low noise transistors such as the 2N5564 JFET in the JPL-style pre-amp and on good operational amplifiers in the signal-processing chain (e.g., Harris 2520 or National Semiconductor Type LF358). The pixel clock timing should be synchronized with the crystal-controlled microprocessor clock and with a non-varying (in time) instruction sequence in the microprocessor to avoid noise due to time "jitter" or changes in the amount of signal induced by accidental coupling from the microprocessor circuits. It is essential to maintain a precisely uniform readout rate since, due to the settling times of various components, temporal jitter can produce small variations in the steady or "baseline" voltage of the amplifiers which the CDS cannot cancel and so the noise will increase.

Janesick and Elliott (1992) gives a very useful noise relationship for the total read noise floor when off-chip amplifier noise is included:

$$\sigma(e^-) = \frac{\sqrt{\Delta f}}{S_V}\left(\sigma_{\text{ccd}}^2 + \sigma_{A_1}^2 + \frac{\sigma_{A_2}^2}{A_1}\right)^{1/2} \tag{8.7}$$

where σ is the "total read noise floor"; A_1 is the pre-amplifier gain; Δf is the equivalent noise bandwidth (Hz); S_V is the node sensitivity (V/e^-); and σ_{ccd}, σ_{A_1}, and σ_{A_2} are the frequency-dependent white-noise figures for the CCD output

amplifier, pre-amplifier, and post-amplifier, respectively (V/root-Hz); an on-chip amplifier gain of unity is assumed.

Example:. Take σ_{ccd}, σ_{A_1}, and σ_{A_2} to be 10, 2.5, and 20 nV/root-Hz, respectively, and assume that the bandwidth is 250 kHz and the node sensitivity is 1 μV/electron. For a pre-amplifier gain of $A_1 = 22$, find the noise generated by each component and the total system.

Solution. Substituting the numbers in each term in the equation gives $5e^-$, $1.25e^-$, and $0.45e^-$, respectively, for the CCD, pre-amplifier, and post-amplifier, and $5.17e^-$ for the quadrature sum. Note that although the post-amplifier has the highest voltage noise, this is suppressed by the pre-amplifier gain.

8.6.2 On-chip binning

Although CCD architectures vary from one manufacturer to another, all frame transfer devices incorporate a structure which permits another type of signal processing called "on-chip binning". Onchip binning of charges is the process of adding or combining charge packets from rectangular groups of pixels at a "summing point", which is the output node capacitor of the CCD. These groups of pixels effectively become single larger pixels, sometimes called "superpixels". On-chip binning is achieved as follows.

Any number of rows can be co-added together in the serial register simply by omitting the horizontal clocking sequence which usually occurs after each vertical transfer. Likewise, when the horizontal register is clocked, the charges from any number of pixels can be combined on the output capacitor by simply *omitting* the reset pulse which usually occurs between each horizontal transfer. By not running the horizontal (serial) clocks after a vertical transfer, the next vertical transfer will co-add the next row of charges into the horizontal register. Similarly, by not resetting the output transistor after a pixel has been read out, the charge associated with the following pixel will be co-added with the first. In this way rows and columns can be "binned" together (Figure 8.16). The final picture has smaller dimensions and less resolution, but it may have higher signal to noise because more signal electrons were combined "on-chip" before the noise of the readout process was added. Of course, care must be taken to avoid saturation. A $1{,}024 \times 1{,}024$ CCD using 4×4-pixel on-chip binning (omit 4 horizontal transfers and 4 reset pulses) will produce an image with only 256×256 pixels. This feature is most useful in high-resolution echelle spectrometers where binning is usually slight ($\times 2$, say) or not at all along the spectrum so as not to lose spectral resolution, but substantial ($\times 15$ or more) at right angles to the spectrum since it allows more signal per readout and so a much better signal-to-noise ratio is obtained in a faint spectrum. This technique is also used for optimizing the resolution in variable seeing conditions or to obtain a preliminary look at a field containing very faint sources. A clock sequence which has these properties can be "called up" when required if the sequencer is based on a programmable design.

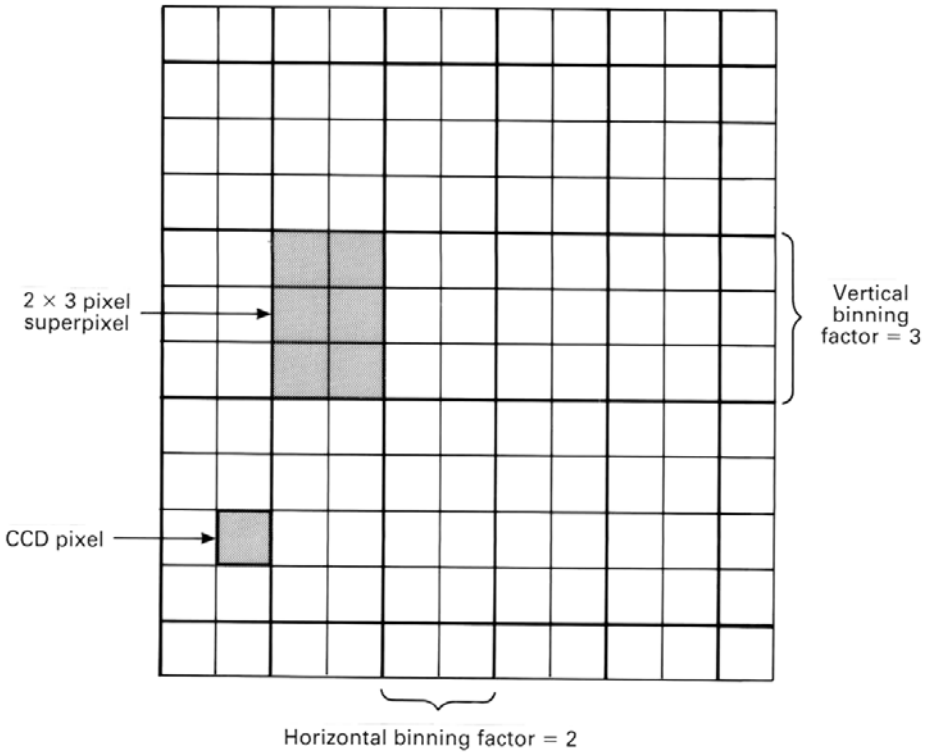


Figure 8.16. The principle of on-chip charge binning.

8.6.3 Overscanning and sub-arrays

When a CCD is read out rapidly with essentially zero exposure time, a picture appears on the screen which, although random, has an average value which is not zero. This value, called the "bias" value, must be subtracted from every image before any other arithmetic manipulation of the data can be performed. A bias frame can be obtained by closing the shutter of the CCD camera and setting the exposure time to zero. Alternatively, the bias level can be obtained by pretending that the CCD array is larger than it really is (say, 10 pixels in each direction), so that the electronics continues to "clock" even though there is no longer any physical correspondence with points on the CCD; ten more vertical transfers are requested and ten more horizontal transfers are included each time the serial register is activated. Again, a programmable controller can deal with this request by simply calling up a different "repeat count" for each type of clocking sequence. When the resulting image is displayed, it will now have a "border" on two sides which contains no true signal, only the voltage corresponding to the bias level (Figure 8.17). Many pixel values in this "overscan" area can be averaged to get a good estimate of the bias level. Thus, no extra frames are required. Bias frames will be discussed again later when considering the detailed calibration of CCDs.

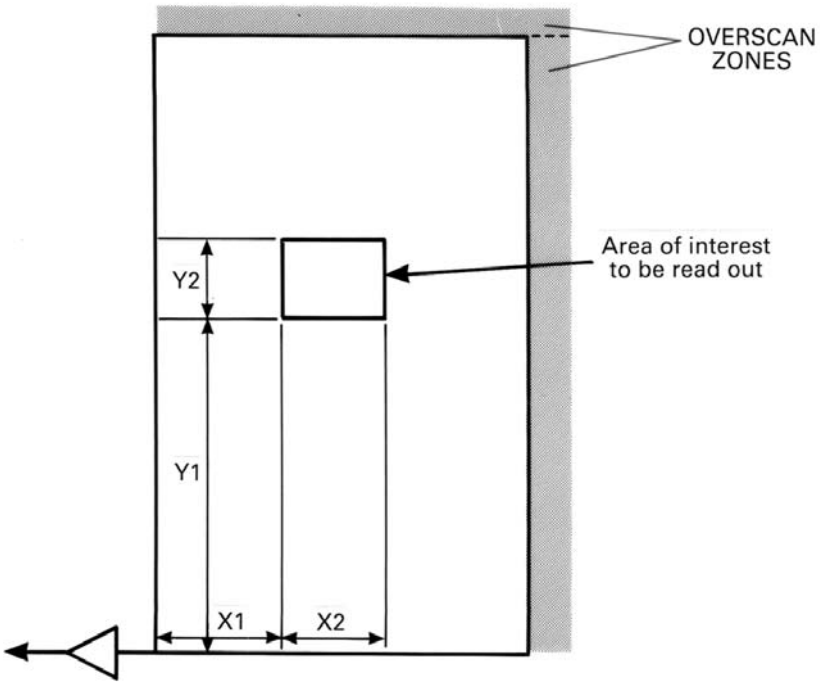


Figure 8.17. The overscan region of an image as well as an arbitrary sub-array.

Finally, just as you can send more pixels than are required, you can also opt to read out a sub-section or "sub-array" of the CCD (as shown in Figure 8.17). This is usually done by controlling the A/D unit with a (pulse) command to inhibit digitization except within the sub-area specified. There are many applications where this is useful. For example, CCDs used for guiding on a single-star need not read out the entire array. Speed is increased and thus exposure times for the sub-array can be very short. This is highly advantageous for infrared arrays which must operate under high background conditions.

8.7 UNIFORMITY OF RESPONSE

As Craig Mackay (Cambridge, U.K.) has aptly put it, "the only uniform CCD is a dead CCD"! When a CCD is exposed to a scene in which the brightness is absolutely constant or uniform across the entire chip (this is called a "flat-field"), the actual output image from the CCD is generally not "flat". The image will be non-uniform in several ways. Some pixels will be more sensitive and some will be less sensitive than the average, and these pixels may be grouped in odd-shaped patterns or there may be a gradient in sensitivity from one corner of the CCD to the other. Bad pixels, blocked columns, and out-of-focus dust specks on the window will stand out dramatically.

Some effects will even change with different light levels; this kind of behavior is said to be "non-linear". In addition, the flat-field response of a CCD is a function of color (i.e., of the wavelength and passband being used to illuminate the detector). This dependence is due mainly to the fact that photons of different wavelengths are absorbed at different depths in the silicon and that the material is not homogeneous. For example, at 500 nm the absorption depth is 1 μm, but at 300 nm it is less than 0.01 μm and at 1,100 nm it is > 300 μm.

Modern CCDs, especially smaller format devices which are mass-produced, can be made with very good uniformity. For large, thinned CCDs specially tailored for astronomy, there may still be problems. The important point is that these non-uniformities are a small price to pay for other performance features because they can be systematically eliminated by a pixel-by-pixel division of the observed image by the corresponding pixel values in the flat-field calibration frame. Suppose a_{ij} is the signal value on the pixel in the ith row and jth column of the image frame, and b_{ij} is the corresponding value in the flat-field picture. For each pixel, the computer program forms the normalized ratio

$$r_{ij} = \langle b \rangle \frac{a_{ij}}{b_{ij}} \tag{8.8}$$

where $\langle b \rangle$ is the mean signal value averaged over the entire flat-field frame.

This arithmetic task, tedious for humans, is trivially accomplished by a computer. Pixel-to-adjacent-pixel variations in detector sensitivity of perhaps a few percent of the average are normal, and across the entire CCD variations as large as 20% might occur. Reducing such non-uniformities in response is a major pursuit, but their occurrence is not a serious limitation provided they are stable. One very big advantage of CCDs is that the pixels themselves are absolutely fixed in position, unlike in a vidicon-type TV tube where the pixel is defined by where the electron readout beam happens to impact, and so stability in the uniformity of the CCD's response to light is a natural consequence. It is built in.

Because photons of different wavelengths are absorbed at different depths there are several consequences. Front-illuminated CCDs will have wavelength-dependent absorption due to the electrodes. In thinned, backside-illuminated devices, surface effects and multiple reflections from the back surface of the CCD cause color-dependent non-uniformity. At the longest wavelengths, interference fringes due to multiple internal reflections are seen. This fringing phenomenon can also be removed, and will be discussed in Section 9.4. Obtaining an acceptable and appropriate flat-field is 90% of the solution to high-quality, very deep imaging with CCDs. Practical techniques for flat-fielding at the telescope are reviewed in Chapter 9 on characterization and calibration of instruments.

8.8 UV FLASHING AND QE PINNING

Early tests revealed that the 800×800 CCDs to be used for the original Wide-Field/Planetary Camera (WF/PC) on the Hubble Space Telescope showed very poor

response to blue and ultraviolet light, which was very bad news for a telescope expected to have great ultraviolet performance. As tens of thousands of chips had been made, and since the WF/PC was a key instrument, a solution *had* to be found. Of all the CCDs produced for astronomy at the time, only the TI and the RCA CCDs were thinned backside-illuminated devices. The RCA chip showed superior blue sensitivity and had useful sensitivity at ultraviolet wavelengths well below 0.40 μm (or 4,000 A), but had a higher readout noise than was desirable. The thinned TI detectors had little or no response to blue light even though the exposed silicon surface was free of any electronic circuitry or other covering. Later, a second problem became apparent with the thinned TI CCDs. The effective quantum efficiency actually *increased* after exposure to light, so that the use of these CCDs as accurate, linear, brightness-measuring detectors was badly compromised. This effect was called quantum efficiency hysteresis or QEH.

Initially, chemical coatings were deposited on the TI CCDs to act as ultraviolet-to-optical converters. The idea for such a coating was due to Jim Westphal (Caltech) who told me in an interview that, in an effort to learn more about ultraviolet behavior, he was browsing through the Caltech Library and was surprised to find a 1968 volume by Samson which was highly relevant. He was even more surprised to learn that not only had knowledge of UV fluorescence been around for over 50 years and that a salicylate, like Aspirin, was a phosphor which if smeared on a photocell would give it UV sensitivity, but the book also listed two little-used materials which fluoresced under UV illumination at the relatively long wavelength of 0.5 μm, ideal for CCDs; these substances were lumogen and coronene. Jim chose coronene because it could be obtained easily, and he and Fred Landauer coated one-half of a JPL/TI CCD and tested it at Palomar Observatory (CA) during an hour or so of twilight. (This use of a CCD on a telescope was shortly after the successful run by Brad Smith and Jim Janesick.) Incoming ultraviolet photons caused the coronene to glow in visible light, where the CCD was sensitive, and produced an effective efficiency to blue and ultraviolet light of perhaps 14% (i.e., one in seven photons at these wavelengths were actually detected). The coronene also seemed to act as an anti-reflection coating and so helped to improve normal visible sensitivity.

Later, Martin Cullum and colleagues at the European Southern Observatory (ESO) headquarters in Munich, Germany reported the use of a fluorescent laser dye coating painted onto front-illuminated EEV (now *e2v*) CCDs to obtain quantum efficiencies as high as 25% at a wavelength of 0.35 μm (350 nm) in the ultraviolet. The laser dye is reported as being more uniform in response in the blue (0.40 μm or 400 nm) where coronene actually has a sharp dip in performance. The importance of achieving good UV, as well as good optical/red, response in CCDs can be judged by the international efforts to devise better "down-converters" like coronene. At Photometrics, Richard Aikens felt this issue was sufficiently important to team up with a chemist named Gary Simms. Thanks to Gary's efforts, Photometrics soon offered a phosphor coating called Metachrome II which can be applied to any CCD by vacuum sublimation. Subsequently, many other CCD vendors produced other coatings. In fact, *e2v technologies* now has a wide range of coatings, including down-converters for UV work and a "graded" anti-reflection

(AR) coating that can optimize the CCD for spectroscopic work. The most popular AR coating for silicon is hafnium oxide ($n = 1.9$).

Workers at JPL also noticed a peculiar effect in a high-quality, thinned CCD which had been kept in a dry, sealed container. The device was found to have a much higher dark current than it had a few months earlier. Checking if it might be the coronene which had somehow affected the CCD, the JPL workers illuminated the device with an ultraviolet lamp to look for fluorescence. Soon after, the dark current was found to have fallen to normal! This clue led to a fix for both problems (i.e., poor UV response and QEH).

What was the explanation? It seems that imperfections in the thin "native-oxide" (i.e., silicon dioxide or SiO_2) layer grown in air on the exposed backside surface of the CCD had created traps for electrons and holes. A net positive charge at the surface can develop, causing a shallow bucket or potential well which can trap electrons. This unwanted potential well must prevent photo-induced electrons from moving to the depletion region where the charge should have been collected. Since blue and UV photons are absorbed closest to the surface the result is a loss of sensitivity in the blue and UV. Traps can also absorb electrons from the silicon itself, electrons produced perhaps by photons; the electrons trapped in the unwanted potential well at the surface will affect the depth of that well and so the UV quantum efficiency will appear to change with illumination level. Time constants for the release of trapped charge can vary from seconds to days depending on the energy level and the CCD temperature. If the energy level of surface traps is mid-way between the valence and conduction levels, thermally induced dark current could be quite large.

The "fix", exposure to UV light, works because energetic UV photons in the wavelength range 2,000 Å to 2,500 Å (which is not transmitted by the Earth's atmosphere) preferentially fills the most energetic or active trapping sites, creating a net *negative* charge on the CCD surface. Since like charges repel, the negatively charged surface generated by the "UV flood" eliminates the unwanted potential well and encourages electrons formed near the surface to move away from the surface toward the depletion region. Thus, there is no loss of quantum efficiency and no QEH. The energy level of surface defects is forced away from the middle of the forbidden band and so the dark current is also reduced.

Fortunately, the negative charging procedure can be done near room temperature, and when the CCD is cooled to below $-100^{\circ}C$ or 173 K the charge is "immobilized" and provides a permanent field until the CCD is next warmed up. The quantum efficiency of the CCD is driven to its theoretical maximum and held or "pinned" there; this is the QE-pinned condition. UV flooding requires wavelengths between 2,000 Å and 2,550 Å, and about 10^9 photons per pixel (at about 2,139 Å) which can be provided by a zinc lamp. It also appears to require the presence of oxygen during the flood. At room temperature, the improvement disappears in a few hours, especially if humid air is present. In a constant vacuum at $-120^{\circ}C$ the improved UV response can last for at least six months. UV flooding is not a perfect solution. Apparently, it works better for some CCDs than others (in particular, CCDs which are insufficiently thinned); thinning should be such as to reach all the

way to the active layer where depletion occurs. For some CCDs the process will only work if oxygen (O_2) gas is present. The best results occur if the flooding is done warm and then the CCD cooled immediately afterward. Further, it appears that if the UV radiation is too hard (i.e., too short a wavelength, for example, 1,850 Å) the process is reversed!

Negative surface charging can also be chemically induced. Exposure of a thinned CCD to nitric oxide gas (which is poisonous) for a few minutes, dramatically improved the performance at 400 nm in the blue. The effect was stronger than for UV flooding and somewhat more permanent. The original success of RCA chips in achieving superlative blue response was, in retrospect, not given sufficient attention. Having produced scanned electron beam silicon vidicon targets for several years, RCA knew the importance of surface charging vs. electron–hole pair recombination. Their 1976 patent application describes doping of the thinned backside layer to create a potential whose effect is to send photo-generated electrons inward toward the charge collection layer!

A better method of obtaining control of the backside CCD surface is to deposit a thin electrically conducting coating on the CCD that allows an adjustable electrical voltage at the surface. Problems with good insulation between the conductor and the CCD limited the success of this proposed electrically controlled layer, but in the attempt it was discovered that simply the act of depositing an ultra thin layer just 4 Å thick, of the right kind of metal (namely, gold or platinum), produced an astonishing and permanent improvement in the UV, the extreme UV (XUV), and even the soft X-ray region. The new surface coating was termed the "flash gate" by Jim Janesick (Figure 8.18).

The highest possible quantum efficiency is achieved with a thinned backside-illuminated CCD. There are several possible techniques for thinning CCDs, including etching with solutions of hydrofluoric, nitric, and acetic acids and mechanical lapping. For a considerable time the only excellent thinned CCDs which were not highly warped were those from RCA. Some thinned TI devices intended for the Hubble Space Telescope were so badly warped that custom corrective optics were required if the chip was to be used in a relatively fast camera system. For the most part, all such difficulties have been eliminated. All major CCD manufacturers have developed thinning processes. Thinning and backside passivation has also been developed by Mike Lesser at Steward Observatory (Arizona). Mike is able to take thick devices and convert them to thinned devices or devices with UV-enhanced properties (Lesser, 1994). The primary problem with thinned devices is still the interference-fringing behavior in the far-red, but as shown in Section 7.5.2, deep-depletion CCDs are solving this problem..

CCD fabrication has come a long way. Today, very high quality, low-noise scientific imagers can be produced with high yield. Standard and custom CCDs can be designed to cover imaging in the X-ray, UV, visible, and near-infrared regions. Special designs for CCD packaging have allowed large mosaics to be constructed with pixel densities now at gigapixel levels, and availability to devices is no longer a problem.

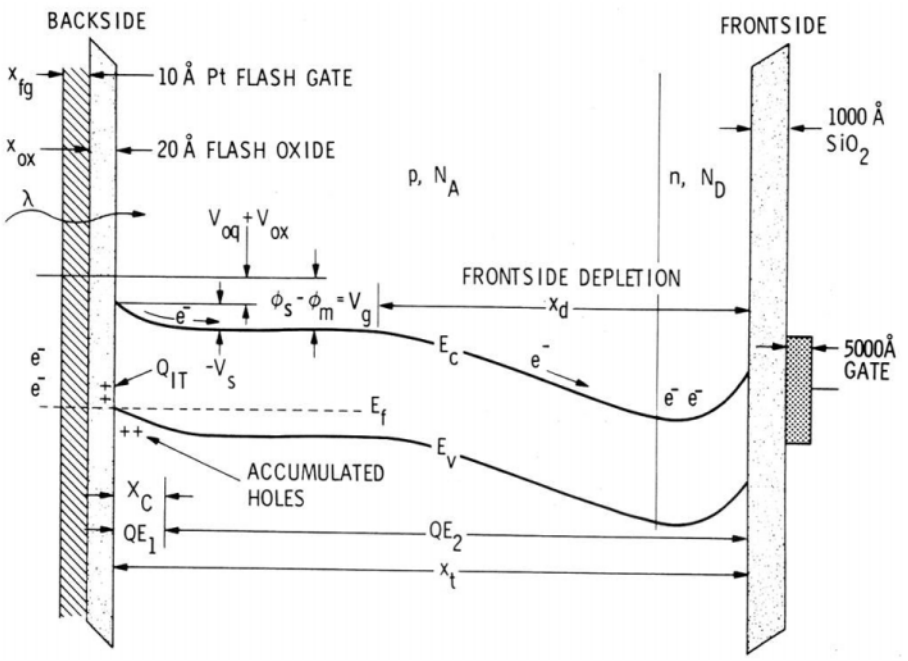


Figure 8.18. A cross-sectional view of a thinned, buried channel, backside-illuminated CCD with an ultrathin coating of platinum called a "flash gate". The sketch is not to scale. The thickness of the flash gate is only 0.001 microns. Credit: Jim Janesick, JPL.

8.9 SUMMARY

This chapter has introduced all of the technical terms that astronomers are likely to encounter when trying to assess different CCD systems. Slow-scanning, cooling, correlated double-sampling, binning, and overscanning are some of the key terms. We have also reviewed CCD controllers, clocking, cosmic-ray events, non-uniformity, and deferred charge. Although it may seem that there are a great many problems in getting CCDs to excel, much of this has to do with the exceptionally stringent conditions required by astronomy applications. CCDs have improved enormously in recent years. Excellent scientific CCD cameras can now be purchased commercially, but it is still necessary to "build your own" when you try something adventurous such as a large mosaic of CCDs.

8.10 EXERCISES

1 Why must CCDs be cooled to low temperatures for astronomical use?

2 Explain what is meant by "inverted" operation and why is it advantageous?

3 What is meant by CTE or charge transfer efficiency? If a 2,048 × 2,048 CCD is quoted as having a CTE of "five nines", what fraction of the original charge in the last pixel will remain when it reaches the output?

4 What are the main sources of "noise" in a buried-channel CCD with a standard floating diffusion output?

5 Calculate the value of kTC noise, in electrons, for a CCD at 150 K if it has an output capacitance of 0.5 picofarads.

6 A certain CCD requires a "fat-zero" or "pre-flash" of 400 electrons. If it has a readout noise of 15 electrons what will be the final noise when the pre-flash is used?

7 A CCD camera seems to be showing a faint—but regular—pattern of sloping lines on bias frames (i.e., very short exposures with the shutter closed). What could be wrong?

8 What is meant by "overscanning" a CCD? How could you use this technique to quantify the charge transfer efficiency (CTE) of a CCD?

9 What causes "fringing" in CCD systems? Why is it worse at longer wavelengths?

10 What is meant by the term "down-converter"? Name two such materials and explain how they are used?

8.11 REFERENCES AND SUGGESTED FURTHER READING

Beletic, J.E.; Beletic, J.W.; and Amico, P. (2005) *Scientific Detectors for Astronomy 2005: Explorers of the Photon Odyssey*, Astrophysics & Space Science Library, Springer-Verlag, Dordrecht, The Netherlands. [Multiple technical articles.]

Gunn, J.E.; Emory, E.B.; Harris, F.H.; and Oke, J.B. (1987) The Palomar Observatory CCD camera, *Publ. Astron. Soc. Pacific*, **99**, 518–534. [An original work.]

Holst, G.C. (1996) *CCD Arrays, Cameras and Displays*, SPIE Press, Bellingham, WA, Vol. PM30.

Howell, S.B. (ed.) (1992) *Astronomical CCD Observing and Reduction Techniques*, Conf. Series Vol. 23, Astronomical Society of the Pacific, San Francisco, CA.

Janesick, J.R.; and Elliott, T. (1992) "History and advancement of large area array scientific CCD imagers", in S.B. Howell (ed.), *Astronomical CCD Observing and Reduction Techniques*, Conf. Series Vol. 23, pp. 1–67, Astronomical Society of the Pacific, San Francisco, CA.

Janesick, J.R.; Elliott, T.; Daud, T.; and Campbell, D. (1986) The CCD flash gate, in D.L. Crawford (ed.), *Instrumentation in Astronomy VI*, *Proc. SPIE*, **627**, 543–582.

Janesick, J.R. (2001) *Scientific Charge-Coupled Devices*, SPIE, Bellingham, WA. [Excellent book on all aspects of CCDs.]

Laborie, A.; Davancens, R.; Pouny, P.; Vétel, C.; Chassat, F.; Charvet, P.; Garé, P.; and Sarri, G. (2007) The GAIA focal plane, *Focal Plane Arrays for Space Telescopes III*, *Proc. SPIE*, **6690,** 66900A-1.

Lesser, M.P. (1994) Improving CCD quantum efficiency, in D.L. Crawford and E.R. Craine (eds.), *Instrumentation in Astronomy VIII*, *Proc. SPIE*, **2198**, 782–791.

Loose, M.; Beletic, J.; Garnett J.; and Xu, M. (2007) High performance focal plane arrays based on Hawaii-2/4RG and SIDECAR ASIC, *Focal Plane Arrays for Space Telescopes III*, *Proc. SPIE*, **6690**, 66900C.

Merline, W.; and Howell, S.B. (1995) A realistic model for point-sources imaged on array detectors: The model and initial results, *Experimental Astronomy*, **6**, 163–210.

Rieke, G.H. (2003) *Detection of Light from the Ultraviolet to the Submillimeter*, Cambridge University Press, Cambridge, U.K.

Robinson, L. (ed.) (1988) Instrumentation for ground-based optical astronomy: Present and future, *Ninth Santa Cruz Summer Workshop in Astronomy and Astrophysics, 1987*, Springer-Verlag, New York. [Early work.]

Simons, D.A.; Amico, P.; Baade, D.; Barden, S.; Campbell, R.; Finger, G.; Gilmore, K.; Gredel, R.; Hickson, P.; Howell, S. *et al.* (2005) Instruments, detectors and the future of astronomy with large ground-based telescopes, in J.E. Beletic, J.W. Beletic, and P. Amico (eds.), *Scientific Detectors for Astronomy 2005*, *ASSL*, **336**, 13–43 (2006); also appeared in *Experimental Astronomy*, **19**, 15–44 (2005).

White, M.H.; Lampe, D.R.; Blaha, F.C.; and Mack, I.A. (1974) Characterization of surface channel CCD image arrays at low light levels, *IEEE J. Solid State Circuits*, **SC-9**, 1–12. [An original work that introduces correlated double-sampling.]

Web sites

ARC/SDSU Controller: *http://www.astro-cam.com/*
Monsoon NOAO: *http://www.noao.edu/ets/new_monsoon/home/news/index.htm*

9

Characterization and calibration of array instruments

All electronic imaging devices require calibration in order to be used for quantitative work in photometry and spectroscopy. It is important to understand how the properties of the detector can be measured and how the behavior of the detector affects photometric and spectroscopic analyses. This chapter describes important steps in these calibrations in terms of CCDs, but the same considerations apply to other array detectors. Signal-to-noise expressions for array instruments are also developed.

9.1 FROM PHOTONS TO MICROVOLTS

The observed quantity in an experiment is the stream of photons, but the detected quantity is a small voltage (V_o) which is amplified and digitized. If N_p photons are absorbed in the integration time (t), then $\eta G N_p$ electrons will be detected. Here $\eta(< 1)$ is the quantum efficiency, and $G \sim 1$ is called the photoconductive gain and allows for intrinsic amplification within the detector in some cases ($G = 1$ for a CCD). Multiplying by the charge on the electron (e) gives the total number of coulombs of charge detected, and the resulting voltage at the output pin of the array detector is

$$V_o = \frac{A_{\mathrm{SF}} \eta G N_p e}{C} \tag{9.1}$$

In this expression, C is the capacitance of the output node of the detector (CCD or infrared array); and A_{SF} is the amplification or "gain" of the output amplifier which is usually a source follower (typically $A_{\mathrm{SF}} \sim 0.8$); the suffix SF stands for "source follower". In practical terms, we need only know the combined quantity $A_{\mathrm{SF}} \eta G / C$, but it is desirable to know these quantities individually too. Therefore, the first step is usually to determine the quantum efficiency or QE.

9.1.1 Quantum efficiency and DQE

In principle, quantum efficiency can be determined in the laboratory with a stable and well-designed calibration system constructed to properly illuminate the detector through a known spectral passband with the minimum of other optics in the beam. One example of an experimental setup is shown in Figure 9.1, and a practical realization is illustrated in Figure 9.2 which shows the UCO/Lick Observatory automated QE measurement system. Either an incandescent lamp or a grating spectrometer can be used as a source of illumination. After passing through a device called an "integrating sphere" which randomizes the light rays and produces a uniformly illuminated source, the light passes through a shutter and a filter holder. At longer wavelengths a stable blackbody source (commercially available) can be used and the integrating sphere is not needed. The light is then split by mirrors: part is directed towards the detector cryostat and part toward a calibrated photodiode. Exposures are taken at the desired wavelengths and recorded along with the signal from the calibrated photodiode. Mounting the camera unit on an optical bench allows it to be moved closer or farther from the light source in a controlled manner. This enables the experimenter to use the inverse square law for light as a way of changing the illumination on the detector. For some wavelengths and passbands it is also possible to use non-wavelength-dependent attenuating filters called "neutral density" filters since their attenuation can be determined fairly accurately with the calibrated photodiode. With this setup it is easy to obtain the relative QE as a function of wavelength, but to convert this to absolute quantum efficiency at any given wavelength requires a precise calibration of the illumination. The exact transmission or "profile" of the filter passband at each wavelength and an accurate determination of the solid angle on the source subtended by a pixel are also required. It is usually easier to obtain the solid angle with a well-defined geometry controlled by

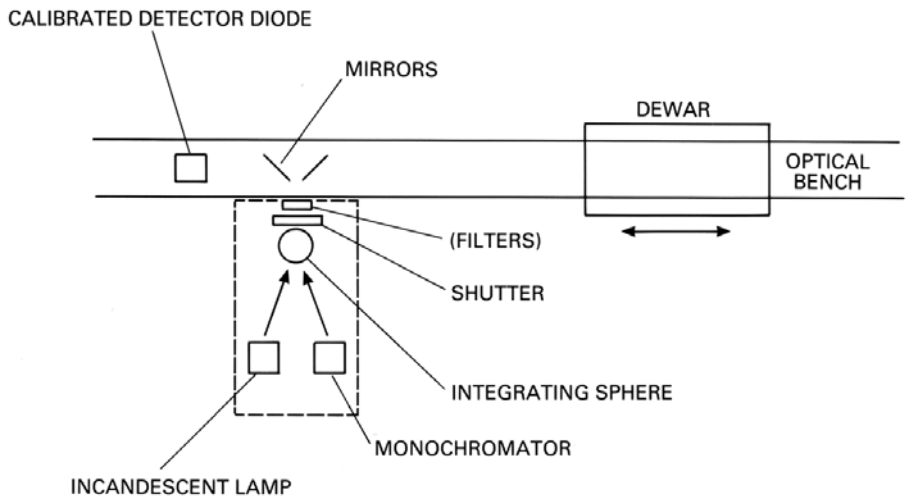


Figure 9.1. A possible laboratory arrangement for calibration and characterization of CCDs.

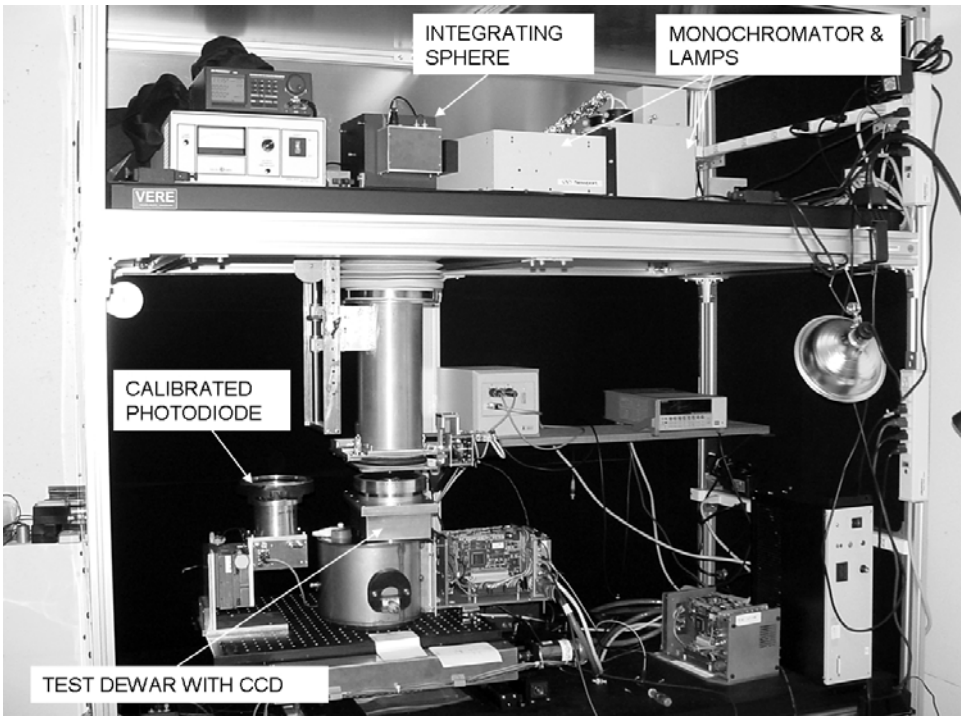


Figure 9.2. The UCO/Lick Observatory automated QE measurement facility. Credit: Richard Stover and UCO.

baffles rather than optics because adding refractive or reflective elements to the setup just introduces other unknown quantities into the experiment. Filter profiles are measured in commercial spectrophotometers. Note that in the case of an infrared filter the scan needs to be done at the operating temperature (e.g., 77 K) because the passband broadens and shifts to shorter wavelengths as temperature decreases. At infrared wavelengths it is easier to be sure of the illumination level by using a blackbody source at a known temperature because the energy spectrum is given by the Planck function $B_\lambda(T)$ which is determined only by the absolute temperature (T). Good laboratory setups can yield both the relative quantum efficiency as a function of wavelength and the absolute QE. The quantum efficiency of a deep-depletion CCD measured with the UCO/Lick Observatory equipment by Richard Stover is shown in Figure 9.3. The reflectance (R) is also measured, and there is good agreement between $1 - R$ and QE(η) except at the short and long ends of the wavelength range. At the shortest wavelengths the absorbed photons create electron–hole pairs too far from the depletion region in these thick devices, and for the longest wavelengths the absorption lengths are too long and no electron–hole pairs are created.

Electrical measurements can be used to determine $A_{\rm SF}$ independently. For $A_{\rm SF}$, the simplest approach is to change the output drain voltage and observe the change in

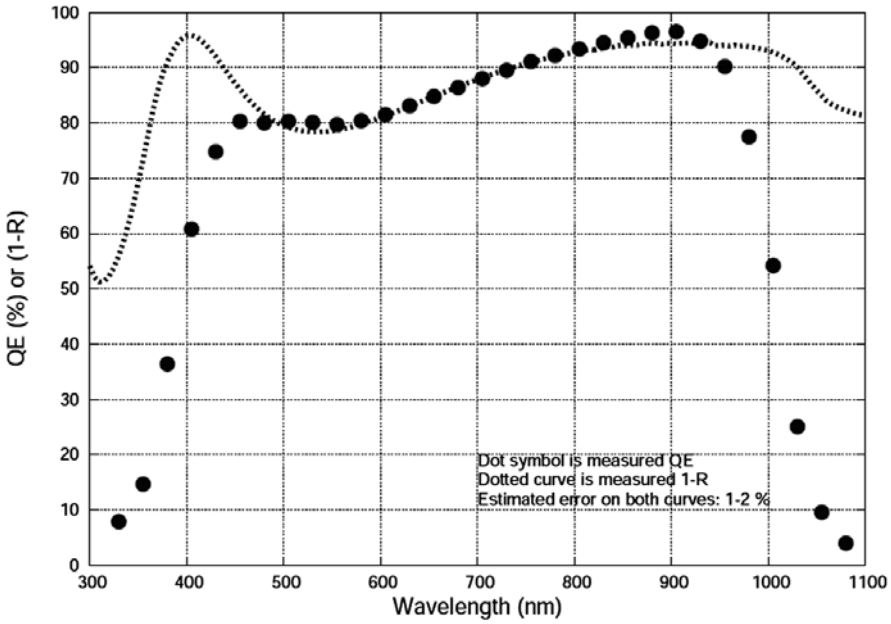


Figure 9.3. Curves of the measured QE and reflectance of a deep-depletion CCD using the UCO/Lick automated system. Credit: Richard Stover.

the output source voltage; the ratio will yield A_{SF}. To measure C a controlled charge Q can be injected and the voltage V measured, then $C = Q/V$. Alternatively, one can expose the detector to a substantial light level to yield a large output signal in which the dominant noise is photon noise. If N is the total number of charges collected, then the measured voltage is $V = eN/C$, and the noise is $\sigma_V = e\sqrt{N}/C$. By squaring the noise and forming the ratio we get

$$C = \frac{eV}{\sigma_V^2} \tag{9.2}$$

thus allowing C to be determined from the *mean* signal V and the *variance* σ_V^2 of the measured voltage noise on the signal.

By observing the signal from a star of known brightness and energy distribution, one can take advantage of the fact that the solid angle (on the sky) of a telescope is very well-defined. Unfortunately, the product (τ) of all the unknown optical transmissions is now included, and so the derived quantity (assuming $G = 1$ and A_{SF} and C are known from electrical measurements) is $\tau\eta$. While this is all that is needed for calibration, it is still very helpful to know where light is being lost so that improvements can be made.

When discussing systems which exhibit readout noise, as opposed to systems with pure photon-counting detectors (PCDs), it is useful to introduce the concept of detective quantum efficiency or DQE. DQE is defined as the quantum efficiency of

Table 9.1. Detective quantum efficiency (DQE) as a function of readout noise R (electrons rms) and number of incident photons N_p for two values of the true QE (30% and 60%).

Read noise $R(e^-)$	*Incident number of photons* (N_p)					
	1	*10*	*100*	*1,000*	*10,000*	*100,000*
1	6.9 (22.5)	22.5 (51.4)	29.0 (59.0)	29.9 (59.9)	30.0 (60.0)	30.0 (60.0)
10	0.1 (0.4)	0.9 (3.4)	6.9 (22.5)	22.5 (51.4)	29.0 (59.0)	29.9 (59.9)
100	0.001 (0.004)	0.009 (0.215)	0.1 (0.4)	0.9 (3.4)	6.9 (22.5)	22.5 (51.4)

an idealized imaging system with no readout noise but which produces the same signal-to-noise ratio as the actual CCD system in question.

In the ideal case, the signal-to-noise ratio for a CCD pixel is given by

$$\frac{S}{N} = \frac{\eta N_p}{\sqrt{(\eta N_p + R^2)}} \tag{9.3}$$

where η is the quantum efficiency at the wavelength of concern; N_p is the total number of photons incident on the pixel in the exposure time; and R is the root-mean-square (or rms) value of the readout noise. Here for simplicity we ignore any correction to R due to digitization noise. An idealized detector with *no* readout noise would have QE equal to η' and a noise, given by Poisson statistics, of the square root of $\eta' N_p$. Equating the two signal-to-noise ratios yields an expression for the DQE(η') of

$$\eta' = \eta \frac{1}{\left(1 + \dfrac{R^2}{\eta N_p}\right)} \tag{9.4}$$

which shows that η' is less than η, and that the DQE of a CCD is dependent on the signal level (N_p) (Table 9.1). To keep the DQE within 10% of the QE, N_p must be greater than $10 \times R^2/\eta$.

For a good CCD with 5 electrons noise and a QE of 0.5 (50%), N_p must be greater than 500 photons, which corresponds to a star of visual (V) magnitude 22.0 in a 10-second integration time for direct imaging on a 4 m telescope. On the other hand, a spectrometer might require an integration time 1,000× longer than this. Clearly, this equation puts stringent demands on CCDs for low noise and high quantum efficiency.

9.1.2 Photon transfer function

The digital signals actually recorded by the CCD system, usually called data numbers (or DNs) or sometimes analog-to-digital units (ADUs), must be turned back into

microvolts, then into electrons, and finally to photons in order to calibrate the system. The relation between DN and microvolts at the CCD output depends on the "gain" of the amplifiers in the system, and the conversion between microvolts at the CCD output and an equivalent charge in electrons requires knowledge of the capacitance (C) of the output node of the on-chip amplifier. The actual data counts or DN recorded in a given time by the CCD camera system are linearly related to the numbers of electrons in the charge packets by the following expression

$$S = \frac{(N_e + N_d)}{g} + b \tag{9.5}$$

where S is the recorded output signal in data numbers (or counts); N_e is the number of electrons in the charge packet (ηN_p); and the system photon transfer gain factor is g electrons/DN; b is the (small) electronic offset or bias level (in DN) for an empty charge packet; and N_d is the (small) residual dark current signal still present after cooling the device. The value of b can be controlled by the designer and is usually substantially above zero, by ten to 100 times the readout noise (expressed in counts), to ensure that there are no problems with the A/D unit when it receives zero or slightly negative signals by chance. (Bipolar A/D units can be used which will properly record negative and positive signals.) Typical values of b would be in the range 50–500 DN depending on the particular CCD and its associated readout noise. Note: Both the bias (b) and the dark current (N_d) can be determined from measurements without illumination and can therefore be subtracted. There are two ways to derive the transfer factor g in electrons/DN: either by calculation, knowing the overall amplifier gain and the capacitance of the CCD, or by a series of observations of a uniformly illuminated scene at different brightness levels. We will describe each method in turn.

Let V_{fs} be the full-scale voltage swing allowed on the A/D unit, and n be the number of bits to which the A/D can digitize. The full-scale range is therefore subdivided into 2^n parts, the smallest part (the least significant bit or LSB) is simply 1 DN. Thus, the voltage corresponding to 1 DN at the A/D unit is $V_{\text{fs}}/2^n$; as an example, suppose the full-scale voltage is 10 volts and the A/D is 16 bits, then 2^n is 65,536, and so the ratio 0.0001525 V, or 152.5 μV, at the A/D is equivalent to 1 DN. Similarly, for a 14-bit A/D the range is 16,384 and 1 DN corresponds to 610 μV. We need the number of microvolts corresponding to 1 DN at the CCD rather than at the A/D, so we must divide the number derived above by the total gain product A_g of all the amplifiers in the system; usually this entails the on-chip amplifier (A_{SF}), the pre-amplifier (A_{pre}), and a post-amplifier (A_{post}). To convert this number of microvolts to an equivalent charge of electrons we must then multiply by the CCD capacitance (C) and divide by the value of the charge on the electron (e). Therefore,

$$g = \frac{V_{\text{fs}} C}{2^n A_g e} \tag{9.6}$$

where $e = 1.6 \times 10^{-19}$ coulombs.

The value of C varies with the type of CCD, but is typically about 0.1 pF. In this case, we get $g = 95.3/A_g$. If the value of g is too small then large signals produced by the CCD will be above the maximum input level of the A/D unit. On the other hand, if g is too large then errors due to the process of conversion of analog-to-digital signals (called quantization errors) will significantly add to system noise. A rule of thumb is to arrange for g to be about 0.5 to 0.25 times the CCD readout noise expressed in electrons. So, for a 10-electron device we might chose g to be $5e^-$/DN implying an external gain of about 27, if $A_{SF} = 0.7$. In this case, the digitization error will be $0.289g$ or $\sim 1.5e^-$ which when added in quadrature with R is only a small correction. Note that a full-scale reading of the A/D corresponds to 65,535 DN $(2^{16} - 1)$ which in this case is equivalent to 327,675 electrons; this is smaller than the physical full well of some CCDs but larger than the well depth of others. The transfer gain factor g can also be determined experimentally.

There are essentially two approaches to measuring g. Obtain several exposures of a flat-field and examine the mean signal and noise from each and every pixel independently before adjusting the illumination level, which must be absolutely constant during the measurements to avoid adding additional noise. Alternatively, take fewer exposures and examine the mean signal from a small array of pixels (in a good clean area of the CCD or IR array) which have first been flat-fielded to remove the dispersion (or noise) in the mean due to their individual variations in sensitivity. The latter method is the best and most widely used, but it will not work well unless pixel-to-pixel variations in sensitivity are removed by good field flattening.

From either data stream X_i, one obtains a mean, dark/bias-subtracted signal (S_M) in *counts* and a variance (V_M), which is the square of the standard deviation of a single observation from that mean.

$$S_M = \frac{1}{n}\sum X_i \quad \text{and} \quad V_M = \frac{\sum (X_i - S_M)^2}{(n-1)} \tag{9.7}$$

where the summation indicated by $\sum$ is over all n pixels in the data set. Now, if all is well, there should be only two noise sources, namely,

(1) photon noise on the signal photoelectrons (p)
(2) readout noise in electrons from the CCD output amplifier (R).

These two noise sources are independent and random, and therefore add together in quadrature, so that the total noise is given by

$$(\text{noise})^2 = p^2 + R^2 \tag{9.8}$$

It is important to realize that this expression (Equation (9.8)) applies to photo-electrons—and not to counts (DN). The measured quantities, the mean signal (S_M) and its variance (V_M), are in DN. To convert from electrons to DN in Equation

(9.8) we need to divide each noise term by g (electrons/DN), to give

$$\left(\frac{\text{noise}}{g}\right)^2 = \left(\frac{p}{g}\right)^2 + \left(\frac{R}{g}\right)^2 \tag{9.9}$$

This equation is simpler than it looks because the left-hand side is now exactly V_M, the observed variance in DN. Also, the mean number of photoelectrons is $g(e^-/\text{DN})S_M(\text{DN})$ or gS_M and the photoelectron noise (p) on this number is simply the $\sqrt{(gS_M)}$ for Poisson statistics, so $p^2 = gS_M$. Hence, Equation (9.9) becomes

$$V_M = \frac{1}{g}S_M + \left(\frac{R}{g}\right)^2 \tag{9.10}$$

Equation (9.10) is just the equation of a straight line in a signal–variance plot of $y = V_M$ and $x = S_M$. Plotting these "observed" quantities (noise-squared and signal) as the illumination changes will yield a straight line of gradient (slope) $m = 1/g$ with the value of the intercept on the V_M axis when $S_M = 0$ giving $(R/g)^2$, which yields R as g is known from the slope (see Figure 9.4 for an illustration). Graphs of noise vs. signal on a logarithmic scale (Figure 9.5) can also be used to display the same result, especially to illustrate behavior over a large dynamic range. Each data point on the line should be determined several times to improve the straight-line fit. This method of deriving g and R is known as the "variance method" or the "photon transfer method" and is extremely powerful. The transfer plot not only provides a completely independent way to derive R and g, but also shows where the CCD or IR array begins to become non-linear and saturate. At the lowest signal levels the noise (variance) is dominated by the fixed readout noise, whereas for larger signals it is dominated by

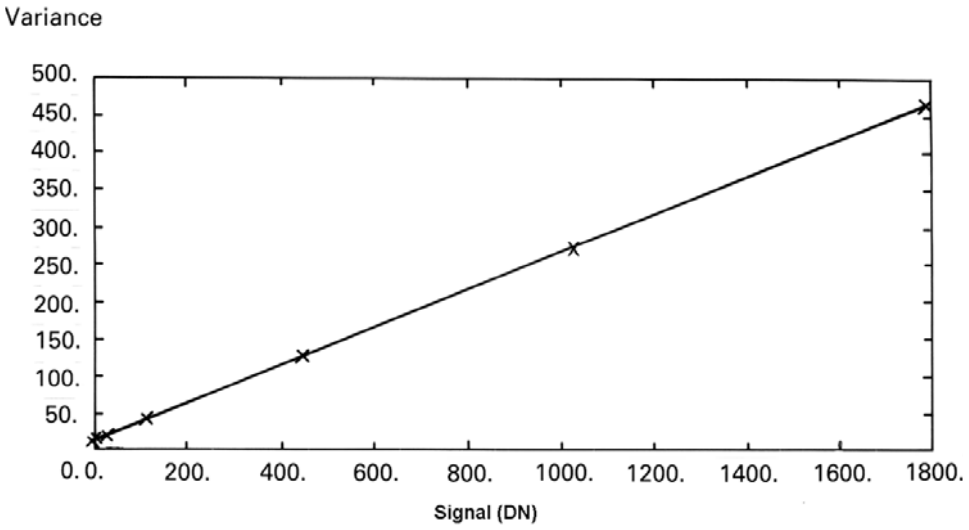


Figure 9.4. A plot of variance (noise-squared) vs. signal in data numbers showing the expected linear graph.

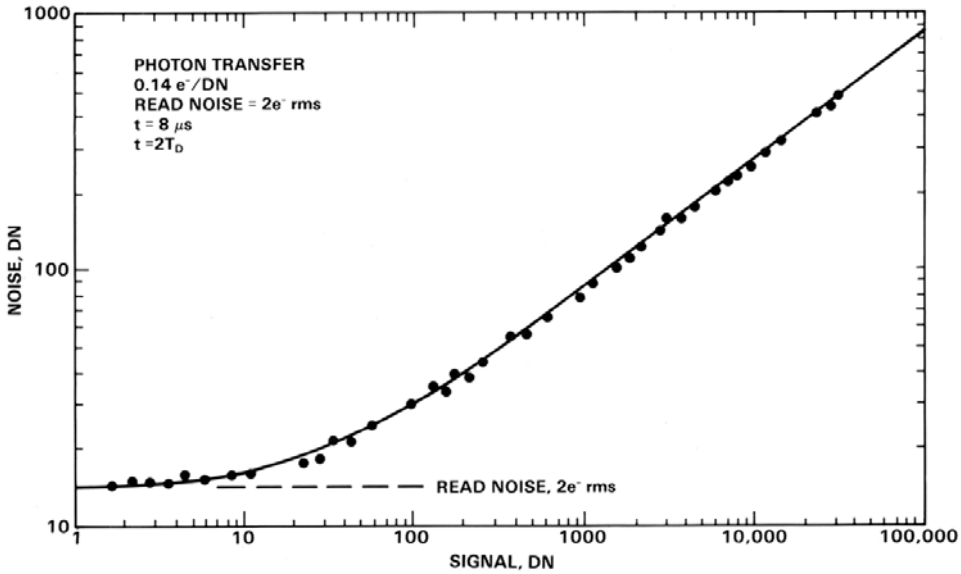


Figure 9.5. A plot of noise vs. signal on a logarithmic scale.

photon noise, the statistical fluctuations in the photon detection rate itself. Finally, at the largest signals, the CCD (or the ADC unit) begins to saturate and the noise actually falls because many pixels have identical values. In principle, the photon transfer curve can be derived for every pixel on the detector so that the behavior of each can be examined independently.

In practice this technique requires some care. Several (perhaps five or seven) flat-field exposures are taken at each of many illumination levels, usually increasing by factors of 2, and a dark frame is taken at each of the exposure levels. Sub-areas of the CCD or infrared array that are free from artifacts and show minimal non-uniformity are selected. For each illumination level the mean picture is computed, and this is subtracted from the individual images to obtain pixel deviations. Alternatively, the mean, dark-subtracted image can be used to flat-field (divide) the individual, dark-subtracted images. The square of the standard deviation of each pixel from the mean value (variance) in the clean sub-array is calculated and averaged over several frames, and the mean bias-corrected signal value is derived from the difference between the averaged picture and an average of several dark frames of the same exposure. A least-squares straight-line fit can then be made to the data set to derive the slope and intercept.

9.2 NOISE, BIAS, AND DARK CURRENT

Typically, in the "no-signal" condition, the CCD (or IR array) electronics system will always produce a small positive reading for each pixel. This electronic signature is

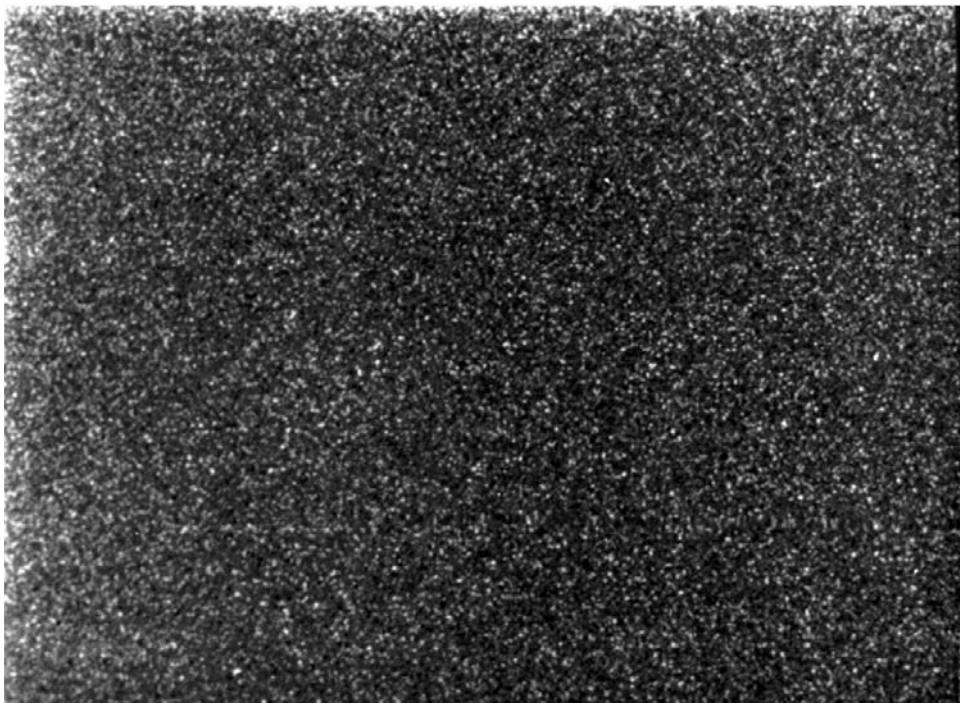


Figure 9.6. A clean bias frame showing no serious amplifier fixed pattern noise or faint diagonal bars due to ground-loop interference.

therefore known as the bias level (given the symbol b in Equation (9.5)) and can be determined easily by taking a very short unexposed frame (zero exposure time, shutter closed), also alluded to as an "erase" frame. An example of a bias frame is shown in Figure 9.6. Several bias frames can be averaged to reduce the random readout noise, while preserving any spatially coherent noise and the pixel-to-pixel bias levels. A mean bias frame must be subtracted from each exposed frame as the first step in data reduction. For some applications it may be sufficient to subtract a single, mean bias *number* from each pixel rather than a bias frame (pixel-by-pixel subtraction), but the normal practice in astronomy is to take many bias frames and use the median (middle) of that set to perform a frame subtraction. Detector arrays that show significant fixed pattern noise on bias frames cannot use a single value for bias subtraction.

Bias information can also be conveniently obtained in many CCD systems by using an overscan. That is, the sequencer is told to send more clock pulses than are actually required to vertically and horizontally read out the real CCD (e.g. $1{,}034 \times 1{,}034$ instead of $1{,}024 \times 1{,}024$). The entries in the image so obtained which lie outside the physical area of the CCD should only contain bias level signals, provided the CTE of the device is good. This latter point is important, because overscanning a uniformly illuminated CCD is one way of probing charge transfer

efficiency or CTE. Charges which are trapped during the primary transfer will slowly leak out and will therefore appear at the output as a result of the overscanning. Usually, the median bias frame is computed and subtracted from the object frame. Next, the overscan in the bias frame is compared with the overscan in the astronomical object frame to derive any small offset which is averaged and subtracted as a single number (scalar) from the object frame. For older CCDs requiring a pre-flash, the pre-flash must also be applied to the bias frame.

If the system is working perfectly then a bias frame should contain very little fixed pattern structure; it should be dominated by random readout noise variations. The standard deviation of a good-sized patch of the array detector therefore gives an immediate estimate of the readout noise (R/g) in data numbers without recourse to the full "photon transfer curve experiment described previously. If there is some unavoidable fixed pattern in the bias frame then it is straightforward to take the difference between two bias frames to eliminate this effect. The measured noise distribution in the differenced image is simply $\sigma = (\sqrt{2})(R/g)$.

In several CCDs it is necessary to turn off, or reduce to a very low level, the current through the output transistor and the clock voltage swing in the horizontal register during long exposures. If this is not done then an immense accumulation of charge can occur in the corner of the chip and along the rows nearest to the horizontal register. The horizontal register and output transistor need to be switched on again in sufficient time to allow them to be swept clean and restabilize before the integration time ends. Similarly, in infrared arrays the output drain voltage needs to be kept to a minimum value. Dark-current levels are determined by long exposures with the CCD shutter kept closed. Exposure times of 1 hour might be needed to determine the dark-current levels with sufficient accuracy. Several identical exposures of this duration will also enable cosmic-ray and radioactivity events to be isolated and counted. Thinned CCDs or those with an epitaxial (or thin) layer are less prone to cosmic-ray events.

Dark current is more significant in infrared arrays, and it may not be linear and scalable from different exposures. It is therefore prudent to either take a suite of exposures that encompass those actually used or simply use the same exposure times and repeat the dark measurements many times to get a good mean value.

9.3 FLAT-FIELDING STRATEGIES

Pixel-to-pixel variations in sensitivity (QE) arise due to physical differences between pixels as the result of fabrication processes and due to optical attenuation effects such as microscopic dust particles on the surface of the CCD. Although these variations can be as small as a few percent of the mean sensitivity, it is important to reduce pixel-to-pixel differences much further for astronomical observations because such variations result in a "noisy" image at a level corresponding to a few percent of the sky brightness.

A common practice is to observe the inside of the telescope dome (if it is matt white) or place a huge white card on the dome. In both cases the screens are so close that the telescope is completely out of focus, which helps to ensure that the field is

uniformly illuminated (i.e., flat). Dome illumination is usually done with tungsten lamps that unfortunately do not remotely mimic the spectrum of the night sky which is a complex composite of a thermal continuum, scattered light, plus a number of practically monochromatic (and variable) night-sky emission lines. Optical filters can sometimes exhibit "red leaks" (i.e., an unwanted transmission to light at much longer wavelengths). Although the light leak is at a very low level (less than 1%), the sensitivity of the CCD is so good that even this is a problem. The solution is to "block" the red light with an additional filter. Tungsten lamps are much redder than the sky, so dome flat fields with such filters would be a poor match. Also, the dome illumination passing through the instrument fills the beam in almost the same way as when observing objects in the sky, but not quite, and that could make a difference. For faint objects it is the light of the sky that dominates, and so it is better to try to use the sky itself as a flat-field. For brighter objects it is their own intrinsic color which matters, and that neither matches sky nor dome. Most workers have found it desirable to establish a set of narrow—rather than broad—passbands for imaging so as to limit the effect of color-dependent non-uniformity.

To first order, CCD detectors are quite linear, so only a simple arithmetic division pixel by pixel, using an image of a uniformly illuminated scene, the flat field, of the appropriate color, followed by re-scaling, is required to remove pixel-to-pixel variations in sensitivity. Let I_{FF} be the uniform illumination of the flat-field source on a pixel in row i, column j whose quantum efficiency is η_{ij}. The observed signal from that pixel in DN is

$$(X_{ij})_{\mathrm{FF}} = \frac{1}{g}\eta_{ij} I_{\mathrm{FF}} \tag{9.11}$$

where g is the conversion in electrons/DN.

The mean signal in the flat-field is obtained by averaging X_{ij} over all the rows and columns; let us call this S_{FF}:

$$S_{\mathrm{FF}} = \frac{1}{g}\eta_M I_{\mathrm{FF}} \tag{9.12}$$

where η_M is now the mean QE averaged over the entire array.

For the true image scene on the sky (I), we have a similar expression for the recorded signal on the pixel on the ith row and jth column:

$$X_{ij} = \frac{1}{g}\eta_{ij} I_{ij} \tag{9.13}$$

To eliminate the position-dependent QE response (η_{ij}) we form the *ratio* of image scene to flat-field

$$\frac{X_{ij}}{(X_{ij})_{\mathrm{FF}}} = \frac{I_{ij}}{I_{\mathrm{FF}}} \tag{9.14}$$

and finally we rescale this ratio to the mean of the flat-field to give

$$\frac{X_{ij}}{(X_{ij})_{\mathrm{FF}}} S_{\mathrm{FF}} = \left(\frac{\eta_M}{g}\right) I_{ij} \tag{9.15}$$

Thus the flat-fielded, re-scaled image, constructed by the operations on the left-hand side of this equation, differs from the true image scene I_{ij} by a single, constant scale factor η_M/g, which can be determined by observations of a source of known brightness if required.

Many flat-field exposures are averaged to increase the accuracy of the flat-field itself and to remove from the flat-field various artifacts such as cosmic-ray events. Also, as already mentioned, the color of the flat-field should be as good a match as possible to that of the image scene because the quantum efficiency is a function of wavelength and the cancellation described above may not be perfect if the two scenes are significantly different in color. To detect the very weakest signals, particularly for extended images, it is necessary to correct for residual non-linearity by more complex means, such as obtaining a series of flat-fields at different exposure levels and determining the response of each pixel individually by means of a polynomial-fitting routine.

Dome flats require lamps that are usually attached to the secondary mirror support ring of the telescope. In some cases it is also possible to use an independent projector. For very high accuracy work or for observations on objects fainter than the night sky itself, various systematic errors tend to dominate over the expected random errors from photon arrival statistics. The first advance in counteracting low-level systematic errors in the detector was the drift scan technique and the concept of time delay and integration (TDI) as already mentioned in Sections 4.1.4 and 7.4.3. To recap briefly, the CCD charge pattern is transferred slowly along columns while the image from the telescope is physically "scanned" along with it in precise synchronization. In this way, the image keeps up with the charge pattern and every pixel along a column contributes to the final image of the object. Figure 9.7 illustrates how the signal ramps up to a steady level and then ramps down when the scan is stopped, and how the efficiency improves after multiple scans (Hall and Mackay, 1984). The drift scan technique averages over some of the fixed pattern noise on the CCD (at least in one dimension) and is effective in reducing systematic errors to below 1% of the night-sky background level due mainly to the fact that the "color match" in deriving the flat-field from the sky itself is much better than from an artificial lamp illuminating the inside of the telescope dome. There are also variations on this technique, such as the "short drift scan" in which the drift is over a much smaller number of pixels than in a whole column.

There are a number of steps which can be taken to reduce systematic errors to a level of at least 0.1% of the night-sky background (assumed to be very uniform over the small areas imaged by CCDs) without the need for a mechanical stage. Tony Tyson (University of California Davis, and formerly Bell Labs) was one of the first to demonstrate that these methods would enable a large 4 m class telescope to reach its theoretical limit of 27th magnitude in a 6-hour exposure. The dominant sources of error are the mismatch in color between calibration flat-fields and the actual night-sky background and, for some thinned CCDs, interference fringing due to unblocked night-sky emission lines. Generation of a master flat-field and sky frame from the object frames themselves has been shown to remove systematic effects to better than 0.03% of night sky. What this means is that after flat-fielding, the apparent variations

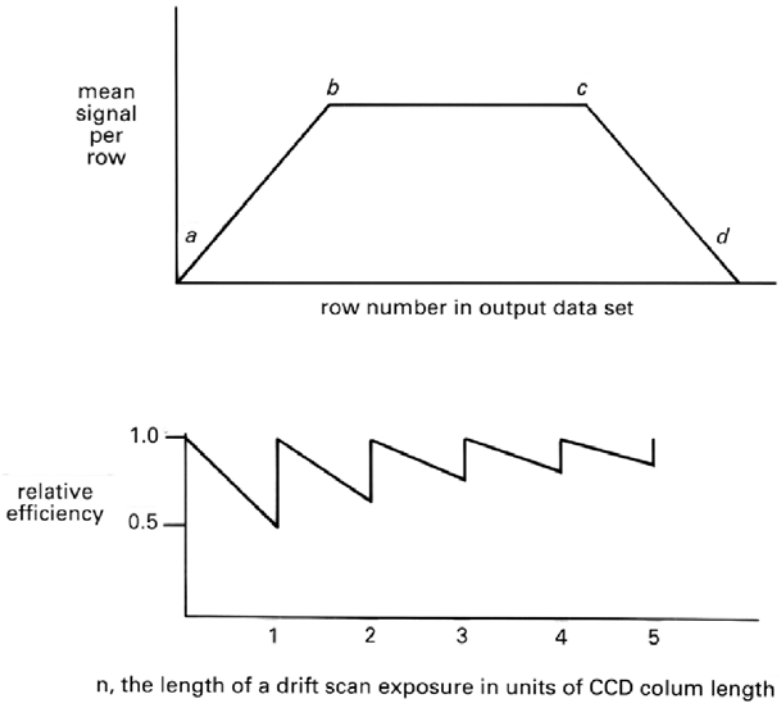


Figure 9.7. The ramp-up and ramp-down of the mean signal from a given column with the drift scan technique for flat-fielding and the efficiency of the drift scan as the scanned area is increased. Credit: Craig Mackay.

in brightness from pixel to pixel are, on average, only 3/10,000 of the mean sky brightness, assumed constant across the image. Similar results have been obtained using this approach with infrared array detectors. Essentially, this powerful technique involves numerous observations of a piece of relatively blank sky with the telescope pointing to a slightly different position on the sky (displaced by, say, 5 arcsec–10 arcsec) for *each* CCD exposure. Positions can be chosen randomly or in a simple pattern, but it is best not to repeat the pattern exactly. When the sequence of "dithered" exposures is examined later, and the frequency histogram examined, or the number of times a certain value or signal is found in a given pixel plotted against those signal values, then one signal value (or a small range) will turn out to be most favored. This value is just the signal corresponding to pure background night sky. An occasional faint object will fall on the given pixel very infrequently because of the semi-random displacement between exposures. The mathematical statistic used to determine the most likely sky value from such a data set is usually the median (or middle value) since it is extremely tolerant of a value discrepant from all the rest, including values due to bad pixels. If sufficient frames are taken, it is possible to calculate the mode, which literally is the most frequently occurring value. Clearly, the disadvantage of this "dithering" technique is that it will not work on object frames

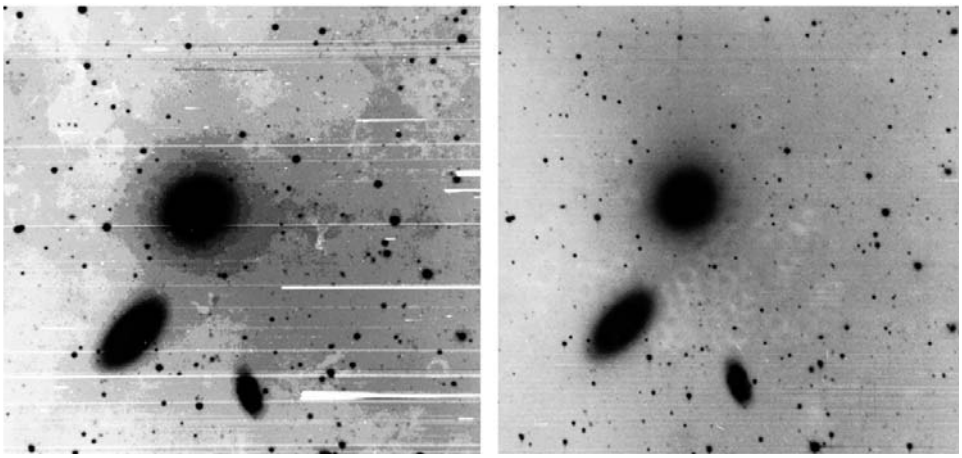


Figure 9.8. (a) A raw CCD image with many defective pixels; (b) same image flattened by using "median sky flats" by shifting the images in a "dither" pattern. Credit: Harold Ables, U.S. Naval Observatory.

which are too crowded, such as the image of a large galaxy or nebula, or a centrally condensed cluster of stars or galaxies unless much larger moves are made. A combination of dome flats and sky flats is often recommended in that instance. However, with a telescope of the right scale and a 2,048 × 2,048 CCD, astronomers at the U.S. Naval Observatory in Flagstaff, Arizona managed to apply this technique to the field shown in Figure 9.8; the chip was an early "setup" device not intended for scientific use, but the powerful dithering technique shows that the images clean up very nicely!

As a matter of practicality, it is essential to "normalize" the various flat-fields before applying a median-filtering algorithm, because a drift in the mean level of the illumination, while not disastrous for the overall technique, will affect the calculation of the median value. It is very important that any additive effects which do not vary with the sky brightness should be removed before scaling. CCDs and infrared arrays exhibit both additive and multiplicative effects which require calibration. In the "additive" class we have

(1) electronics pattern noise ("bias effects")
(2) charge skimming or trapping
(3) interference fringes
(4) LED activity

and in the "multiplicative" class there are the effects of

(1) quantum efficiency variations across the array
(2) transmission of optics and coatings
(3) thickness variations of thinned arrays and CCDs.

Clearly, it is important to devise steps to remove these effects and to perform the corrections in an optimum sequence. In general, a good recipe for reduction and calibration of raw CCD images is the following pattern:

(1) *Subtract bias and bias structure:* a "bias structure" image, obtained by averaging many bias frames is subtracted, and a further small scalar subtraction to eliminate any offset in bias between program and bias frames is also made using overscan data. For CCDs using pre-flash the bias must also use pre-flash.
(2) *Subtract dark:* most CCDs exhibit some dark current or a low-level light-emitting diode (LED) activity, leading to "electronic pollution" during long exposures. To correct this, the "median image", rather than the "mean image", of many long (bias-subtracted) dark exposures, obtained as recently as possible, must be subtracted from the object frames. If these frames do not correspond exactly to the exposure time used on the illuminated frame, then careful experiments of dark current vs. exposure time must be done to prove whether or not the dark current is sufficiently linear to be scaled from one exposure time to another.
(3) *Divide by flat-field:* many exposures on a diffusely illuminated screen are normalized (adjusted to the same mean) and averaged for each filter or spectrograph setting. Exposure levels in each flat-field (FF) and the illumination colors must be similar to the astronomical exposures; this means filtering the dome lights. For spectroscopic applications the color of the FF is less important. The bias-corrected, dark-subtracted, normalized flat-field "master" averages are then divided into the dark-corrected object frames to calibrate for QE variations (and other effects such as out-of-focus dust spots) from pixel to pixel. Alternatively, the master flat-field is derived from a series of images of sparsely populated regions of the sky, each of which has been displaced (semi-randomly) from the other. Since a star will not fall on the same pixel twice, then the median (middle value) or better still, the mode (the most frequently occurring value) of this set of "dithered" images, will be a frame composed purely of uniformly illuminated sky. In practice, it is best to use a limited number of sky frames (e.g., 7–13) close in time to the frame being flat-fielded (e.g., some before and after), and form a "running median" by stepping through the data set.
(4) *Subtract fringe frame (sky subtraction):* night-sky emission lines limited to a very narrow band of wavelengths can cause interference fringes on some back-illuminated CCDs, in sky-limited exposures. A "fringe frame" derived from an adaptive modal filter routine (see Section 9.4) must be scaled and subtracted from the already flat-fielded object frame derived from Step 3. This step effectively performs a sky subtraction.
(5) *Interpolate over bad pixels:* either a bad-pixel map for the particular CCD in use is supplied as input to a median interpolation routine or, if the two-dimensional displacement or dithering technique is used during observing, then bad pixels at low levels will automatically get filtered out of the final image. Remaining bad pixels can be replaced by good data from other frames for that part of the scene.

(6) *Remove cosmic-ray events:* a computer algorithm can be used to identify non-star-like point sources, or the random offset technique can be used to eliminate (by median filtering) cosmic radiation events greater than several times the normal, random fluctuations in the sky background level.
(7) *Registration of frames and median filtering:* if the random offsetting technique is used, re-registration of data frames to a fraction of a pixel using interpolative software routines, followed by median filtering and edge trimming, is used to produce the final cleaned image.

Note: all of these basic steps also work for IR arrays, but extreme care is required because the backgrounds are larger and more sensitive to variations.

9.4 FRINGES AND SKY EMISSION

Interference fringes produced by backside-illuminated thinned CCDs in the far-red were mentioned in Section 7.3.5. Fringing is not a problem over most of a CCD's wavelength range. In narrow-band work, near-infrared arrays can also show significant fringe patterns caused by variable OH emission lines in the night sky in the far-red and near-infrared range. Because difference frames are routinely obtained with IR arrays, these "additive" effects are usually removed easily, except when the OH emission is extremely variable. Obviously enough bias, dark, and flat-field frames must be taken to ensure sufficient accuracy in the median-processed images. Although CCDs and infrared arrays have improved dramatically over the years, it is nevertheless essential to invest effort in calibrations if one is to extract the faintest signals and push the instrument/telescope combination to the limit of its capability. There is usually a sufficient opportunity to take good calibration data during daylight or twilight hours.

Fringe removal can be performed by a technique called adaptive modal filtering. This routine computes the absolute difference between the mean and the median of values associated with a pixel over all the images in a set and rejects deviant values until this difference falls below a certain value or a maximum number of values have been rejected. A given pixel is then median-filtered over all the images. Of course, this technique fails with large, extended objects such as nearby galaxies and nebulae, but it is very suitable for relatively uncrowded fields in which two-thirds of the actual CCD area is occupied by sky. This fringe frame must then be scaled by trial and error and subtracted from the object frame until a patch of sky on the object frame is entirely flat. Figure 9.9 shows a 500 s CCD exposure in a far-red band on a 4 m telescope before and after fringe removal; this particular thinned CCD has strong fringing. The correction effect is remarkable! The introduction of deep-depletion CCDs and improved anti-reflection coatings has drastically reduced problems of this kind.

9.5 LINEARITY

If operated properly, CCDs are linear detectors over an immense dynamic range. That is, the output voltage signal from a CCD is exactly proportional to the amount

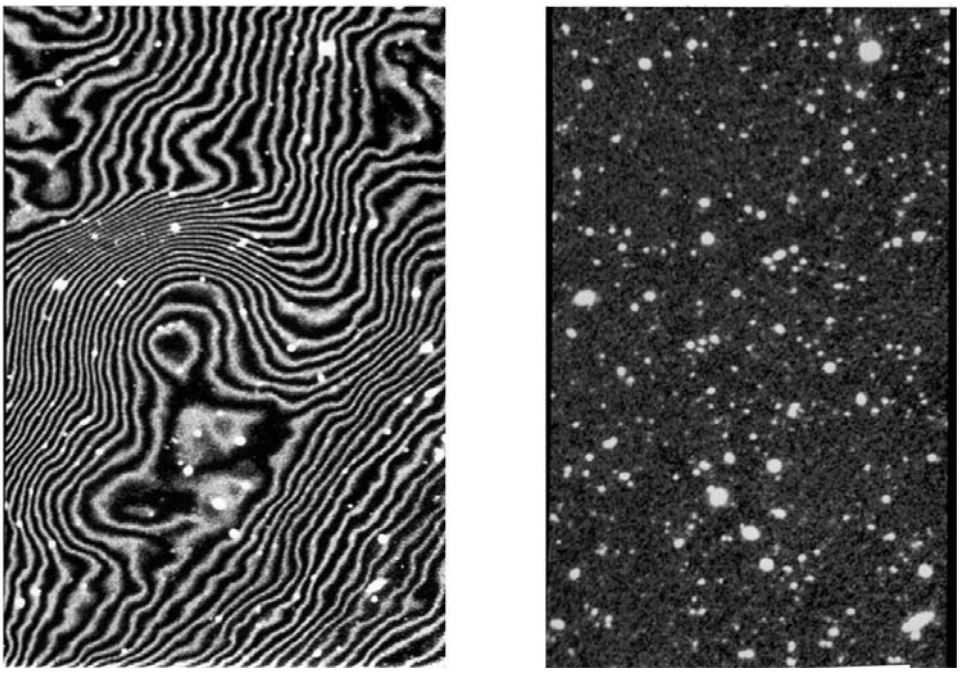


Figure 9.9. (a) A severe fringe pattern due to night-sky emission lines on a deep 4 m telescope exposure with a thinned, back-illuminated CCD. (b) The same field after processing to remove the fringes.

of light falling on the CCD to very high accuracy, often better than 0.1% of the signal. The good linearity makes it possible to calibrate observations of very faint objects by using shorter—but accurately timed—exposures on much brighter photoelectric standard stars. Linearity curves are usually derived by observing a constant source with various exposure times. This method assumes that exposure times can be very accurately controlled, which is generally a good assumption, and that $1/f$ noise sources are negligible. Simply put, if the exposure time is doubled the (dark + bias)-corrected signal should also double. Non-linear behavior from CCDs can occur if incorrect voltages are applied. Care must be taken to ensure that the output transistor is operating in its normal linear regime and it is essential to use the correct clock voltages to ensure the CCD pixel is fully inverted, or use a CCD with MPP built in.

9.6 PHOTOMETRY

Photometry is the process of obtaining accurate numerical values for the brightness of objects. Sometimes relative brightness is adequate and this may be in relation to other objects in the same field of view or nearby in the sky. For other studies it is imperative to derive the true or absolute amount of radiant energy reaching the

Earth; this is much harder to do, but it is essential if we are to understand the distribution of mass and energy in the Universe. The term photometry, of course, derives from measurements of visual light, but it can be applied generally to the measurement of the energy transported by electromagnetic radiation. Some of the terminology of this field is confusing, so it may be useful to review a few basic concepts. An excellent introductory text on photoelectric photometry is the book by Henden and Kaitchuk (1990). Howell (2006) has a very good discussion of CCD photometry with many references.

Monochromatic flux is the power received (in $\mathrm{W\,m^{-2}\,Hz^{-1}}$) from integrating the specific intensity over the angular size of the source. Unfortunately, this use of the term flux is in conflict with its official radiometric definition in which the flux is the total power integrated over all frequencies. We can also use wavelength units: $\mathrm{W\,m^{-2}\,\mu m^{-1}}$.

Magnitudes are relative measures of the monochromatic flux of a source. If F is the monochromatic flux due to a source, measured outside the Earth's atmosphere, and F_0 a reference monochromatic flux at the same wavelength, then the corresponding magnitude (m) of the source is defined by

$$m = m_0 - 2.5 \log F + 2.5 \log F_0 \tag{9.16}$$

If $m_0 = 0$, representing a reference star, then $2.5 \log F_0$ defines the reference or zeropoint of the magnitude scale. In practice, measurements are not monochromatic, but are made over a finite spectral band defined by a transmission filter. Therefore, there are different magnitude systems for different sets of spectral bands. There are two well-known magnitude systems: one is the Vega system and the other is the AB system.

Vega system: in the Vega system, Vega is assigned 0 magnitudes in every bandpass, and therefore F_0 must be determined by measuring Vega (or an equivalent star of the same type) in the given band and with the same conditions. Unfortunately, as the effective wavelength shifts to redder values for stars with a cooler spectral type than Vega, extensive "color-dependent" corrections need to be developed.

AB system: in the AB system, which is not based on Vega, it is assumed that the flux constant F_0 is the *same* for all wavelengths and passbands. Based on the work of Bev Oke (1974) using the Hale 5 m telescope, if the monochromatic flux F is in frequency units then

$$m_\nu = -2.5 \log F - 48.6,$$

whereas if F is in wavelength units then

$$m_\lambda = -2.5 \log F - 21.1$$

The latter scale is sometimes called the STMAG system because it has been adopted by the Hubble Space Telescope. The implication of these two magnitude systems is

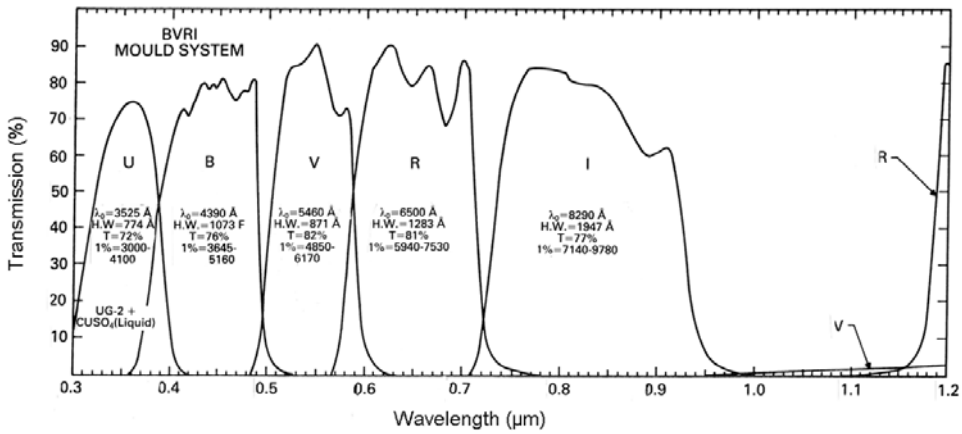


Figure 9.10. Standard filter bandpasses used with CCDs: the Mould system.

that only a source with a constant flux would have the same magnitude at all wavelengths (or frequencies). All three systems are pinned to give the same magnitude and photon flux for Vega in the Johnson V-band.

Bolometric magnitudes: this gives a magnitude corresponding to the total flux integrated over all wavelengths; the zeropoint is $F_b = 2.52 \times 10^8\ \mathrm{W\,m^{-2}}$.

Color indices: this is the difference between magnitudes at two separate wavelengths (e.g. BV, UB) in the conventional UBV system; the letters designating the filters are used for the magnitudes themselves, although m_U, m_B, and m_V are also employed.

The original photometric system is the UBV system of Johnson and Morgan first introduced in 1953 ($\mathrm{U} = 365\,\mathrm{nm}$, $\mathrm{B} = 440\,\mathrm{nm}$, and $\mathrm{V} = 550\,\mathrm{nm}$) and subsequently extended to the red ($\mathrm{R} \sim 700\,\mathrm{nm}$, $\mathrm{I} \sim 900\,\mathrm{nm}$). A modified UBVRI system known as the Kron–Cousins System is more frequently encountered in which R and I are re-defined to 650 nm and 800 nm, respectively; a version known as the Mould System is shown in Figure 9.10. In addition, an important narrower band system was developed by Trinh Thuan and Jim Gunn (the uvgri system) which avoids night-sky lines and the serious overlap of wide-band systems, and gives better flat-fields for CCDs (see Table 9.2).

Figure 9.11 shows the effect of a "red leak" in the B (blue) filter using a logarithmic scale to enhance the lowest levels at 0.01% of the peak transmission. The consequences of red leaks depend on the "color" of the illumination and the sensitivity of the detector at long wavelengths. CCDs are sensitive at wavelengths beyond the original PMTs used to establish the UBVRI system, so it is important to calibrate this effect.

For the Sloan Digital Sky Survey (SDSS) a specific set of five non-overlapping filters was evolved from the Thuan and Gunn System. The SDSS system uses wider bands to ensure high efficiency for faint-object detection, and it spans the entire range

Table 9.2. A summary of the major photometric systems.

Kron–Cousins System		*Thuan–Gunn System*	
Wavelength (Å)	*Width* (Å)	*Wavelength* (Å)	*Width* (Å)
U 3,600	700	u 3,530	400
B 4,400	1,000	v 3,980	400
V 5,500	900	g 4,930	700
R 6,500	1,000	r 6,550	900
I 8,000	1,500	i 8,200	1,300

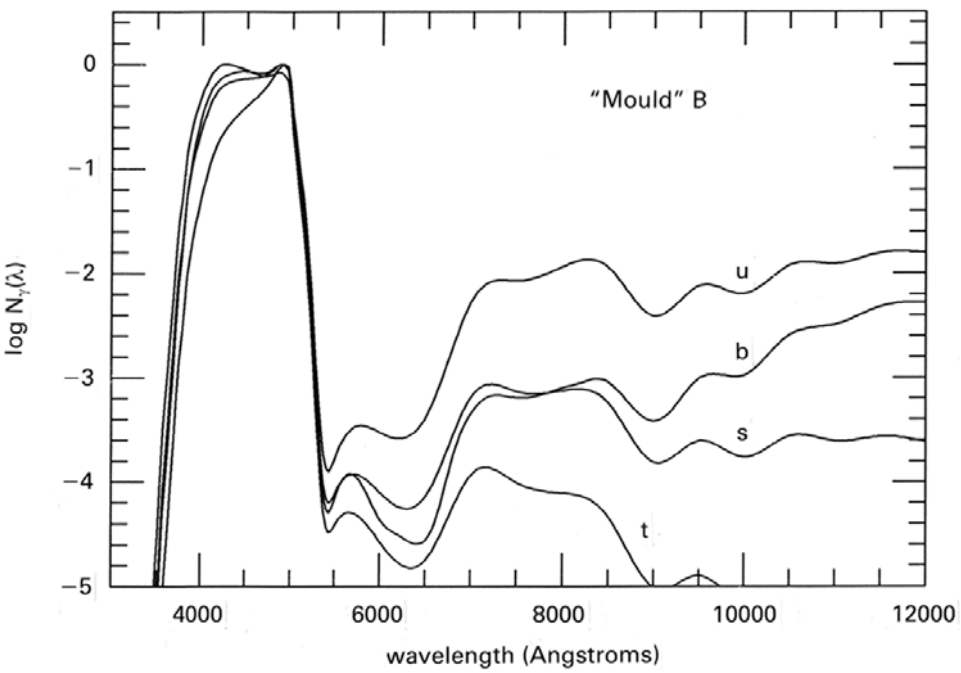


Figure 9.11. The effect of (accidental) imperfect blocking is a "leak" of red photons to which the CCD is very sensitive. The consequence depends on the spectrum or color of the source; *b* and *u* are the balanced (filtered) and unbalanced artificial lamps illuminating the telescope dome, *s* is a typical solar spectrum, and *t* is the twilight sky which is quite blue.

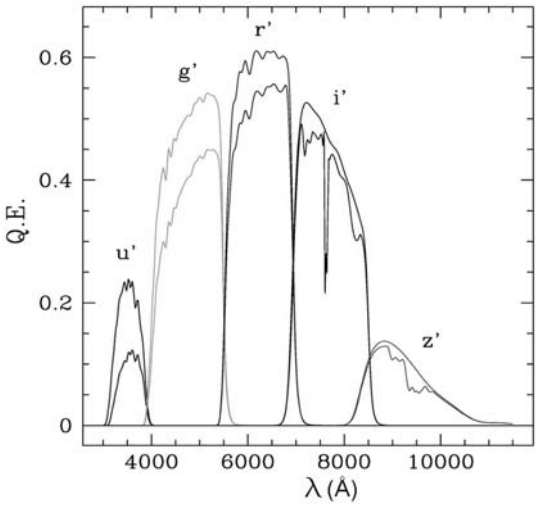


Figure 9.12. The Sloan Digital Sky Survey filter set.

of sensitivity of the CCD from the UV atmospheric cutoff at 0.3 μm to the silicon bandgap limit at 1.1 μm. Figure 9.12 shows the SDSS filter set, and Table 9.3 summarizes the filter properties; the figure also shows the expected impact of atmospheric absorption in each band for a zenith angle of about 30° at the Apache Point Observatory in New Mexico. See Gunn *et al*. (1998) for details of the SDSS photometric camera and *www.sdss.org* for general information. The effective wavelengths and detection limits were taken from the sixth data release information.

In general, the accuracy with which brightness can be measured on CCD frames is extremely high, and repeatability from one CCD image to another is excellent. Differential photometry in which one object is compared with many in the same CCD field has enabled observers to measure magnitudes at the 0.001 or "milli-magnitude" level routinely (Everett and Howell, 2001). Limitations on accuracy are really only introduced when comparing CCD photometry with the results of classical photoelectric photometers which employed photomultiplier tubes. Calibration of CCDs against the old phototube systems to high accuracy (0.5% or 0.005 magnitudes or better) is difficult due mainly to the gross mismatch between CCDs and PMTs in

Table 9.3. Sloan Digital Sky Survey passbands and sensitivity limits.

Property	u'	g'	r'	i'	z'
λ_{eff}	355.1 nm	468.6 nm	616.5 nm	748.1 nm	893.1 nm
Width	56.0 nm	137.7 nm	137.1 nm	151.0 nm	94.0 nm
Limits	22.0	22.2	22.2	21.3	20.5

terms of the huge span in wavelengths to which CCDs are sensitive, and the effect which this has on the filter profile. To a lesser extent, small systematic errors can be introduced by the measuring method during CCD data reduction and analysis. In general, CCDs perform better, and can be calibrated more easily, if the spectral bandwidth is limited to a few hundred angstroms (i.e., a few hundredths of microns). This is not "standard", although the Thuan–Gunn System approaches this ideal, and it does not easily permit a comparison with classical photomultiplier tube photometry. It has been said that classical photoelectric photometry was "a great idea at the time" but now that we have CCDs we should do everything over again! This is already happening at some level as a result of the Sloan Digital Sky Survey.

There are two basic procedures for obtaining photometric information from CCD (and infrared) images:

(a) aperture photometry
(b) profile fitting.

Each is discussed in turn and some additional detail is given in Chapter 10 on computers and image processing.

(a) Aperture photometry: conceptually this is analogous to photoelectric photometry. After the image has been obtained, a computer program (e.g., *apphot* in IRAF) is used to reconstruct the signal which *would* have been obtained from an object in the field of view if the light had passed through a physical aperture (usually circular) of a certain diameter in arcseconds. The imaginary aperture is called a "software aperture". Although this approach sounds quite straightforward there are subtle practical difficulties which can influence the results. For example, procedures used to estimate the background level, the center of the optical image, as well as the shape of the image are all important. Typically, the background sky brightness will be taken as the average in an annular ring with inner radius just beyond the limit of the object and extending a few pixels out. The algorithm should be capable of recognizing when one of the pixels in the sky annulus contains a star (higher signal than the rest) and it should eliminate this value from the average. There is also the difficult question of whether or not to include pixels that are only partially within the annulus. The best procedure is to try a series of increasing values for the radius of the star aperture until the signal becomes reasonably constant. A plot of the derived "instrumental" magnitude ($-2.5\log\{$counts/s within aperture$\}$) vs. the radius of the aperture yields a "curve of growth" from which "aperture corrections" can be deduced. Faint objects are best measured with a small aperture, but bright standard stars usually need a larger aperture.

(b) Profile fitting: also called point-spread-function (PSF) fitting, where by PSF is meant the actual recorded shape on the detector of an unresolved, point source, this method relies on modeling the image rather than summing over the image. Mathematical curves are "fitted" to the real data using computer programs until a

good match is obtained. The stellar image is usually compared with a Gaussian profile:

$$I(r) = I(0)\, e^{-r^2/2\sigma^2} \tag{9.17}$$

where $e = 2.718$ is the base of natural logarithms; $I(0)$ is the peak intensity; and r is the radial distance from the center of the image. The quantity σ measures the width of the distribution; 68% of the light lies within $\pm 1\sigma$, and 98.7% of the light lies within $\pm 2.5\sigma$. Provided the PSF (i.e., σ) is constant across the image, programs like DAOPHOT (Stetson, 1987) will identify the bright stars, deduce their Gaussian profiles and subtract those profiles away, thereby revealing fainter stars. Other mathematical functions can also be used. It is often more convenient to describe the point spread function in terms of its Full Width at Half of the Maximum intensity; the FWHM (in pixels or arcseconds) is related to σ of a Gaussian by the simple equation

$$\text{FWHM} = 2.35\sigma \tag{9.18}$$

The use of CCDs and other array detectors for photometric measurements is founded on two basic assumptions, namely,

(1) The response of each pixel is a well-defined function of exposure level, optical bandpass, and device architecture and control. Considerable effort may be required to fully optimize and stabilize detector control for precision photometry.
(2) The incident signal from the astronomical source can be calibrated or transferred to the desired "standard" system.

For all techniques, whether profile-fitting or apertures, an understanding of the image profile and the centroid of the image is important because centering errors and inappropriate apertures or fitting parameters can lead to systematic effects. There are four other important issues: (1) passband mismatch of the filters, including narrow-band filters; (2) red/infrared "leaks" in the filters which complicate flat-fielding; this can be very serious in the infrared where the backgrounds rise so steeply; (3) the finite opening and closing times of electromechanical shutters; and (4) changes in the atmospheric attenuation or "airmass" for long on-chip integrations. Items (1) and (2) need to be eliminated by design. It is usually assumed that the basic CCD calibration procedures of bias subtraction, dark subtraction, flat-fielding and defringing have been applied correctly, that the effects are small or "infinitely well-known"! This is not always true. Extreme care and patience at the telescope is required to ensure that all necessary calibration data are obtained with sufficient accuracy so as not to limit the photometric determinations, and that nothing has happened "on the night" which might render reductions difficult (e.g., an accidental ground-loop, detector temperature instability, electronic failures, or inadvertent saturation which may result in latent images in some detectors. Shutter timing errors (Item 3) are mainly relevant for large iris-type shutters that require a finite time (δt) to close, and thus the pixels at the center experience a longer exposure time of $T + \delta t$. For large arrays the error could be $\sim$0.1 s. However, the double-screen Bonn shutter

(web site given in references) alleviates this problem and provides a way to cut off light to very large mosaics for which iris shutters are not available anyway. The shutter has two opaque screens that follow each other in either direction. In a long exposure the first screen moves through and the aperture remains open for a while until the second screen moves to close it. Although CCD pixels are covered and uncovered at different times, every pixel experiences the same *interval* of time, to an accuracy of a millisecond or better. To handle a very short exposure, the closing blade starts moving while the opening blade is still in the aperture uncovering pixels. The effect is a moving slit of light across the CCD that ensures every pixel is exposed for the same time interval. Even for the largest shutter (480 mm) the blades move in less than 1 s across the aperture.

Once photometric values are obtained they must be compared or calibrated against well-measured "standard" sources, so that different data sets can be compared, or the values related to theoretical predictions. With hindsight, the classical photometric systems such as the UBV system may not be the most rational for use with CCDs, but such comparisons are inevitable. Photoelectric standard stars must be observed over a wide range of airmasses; this was Item 4 in our list above. The airmass (X) is given essentially by the secant ($= 1/\cos$) of the zenith angle of the object since this quantity is proportional to the thickness of atmosphere (treated as plane-parallel slabs) through which the light passes. Straight overhead the airmass is 1.0 (because $\cos 0^\circ = 1$), and at an altitude of 30 degrees the airmass is 2.0 (because $\cos 60^\circ = \frac{1}{2}$), thus the loss in magnitude from the top to the bottom of the atmosphere is doubled. In general, then, if ζ is the zenith angle of a star and α is the difference between the true magnitude m and the magnitude that would be observed at the zenith, the true magnitude is related to the observed magnitude by

$$m = m(\zeta) - \alpha_\lambda \sec(\zeta) \tag{9.19}$$

where the suffix λ has been added to emphasize that α is a function of wavelength. Plotting the measured magnitudes against the airmass, ($X = \sec(\zeta)$) gives the slope of the line (α), and the intercept on the magnitude axis corresponding to $X = 0$ gives the true magnitude above the atmosphere (Figure 9.13). Standard stars must also cover an appropriate range of spectral types (or temperatures) to obtain color coefficients since the CCD magnitude derived will depend somewhat on how blue or how red the star is.

The instrumental magnitudes are calculated from

$$\mathrm{IM} = -2.5 \log(\mathrm{counts/s}) \tag{9.20}$$

and these can be compared on excellent nights with standard stars to produce a reference set. The parameters to be determined are

(1) the "zeropoint" (i.e., the magnitude corresponding to 1 count/s for a star of zero color above the Earth's atmosphere);
(2) the "color equation" relating the CCD photometric system to the older photoelectric systems;
(3) the "extinction" factor or light-loss through the Earth's atmosphere per unit

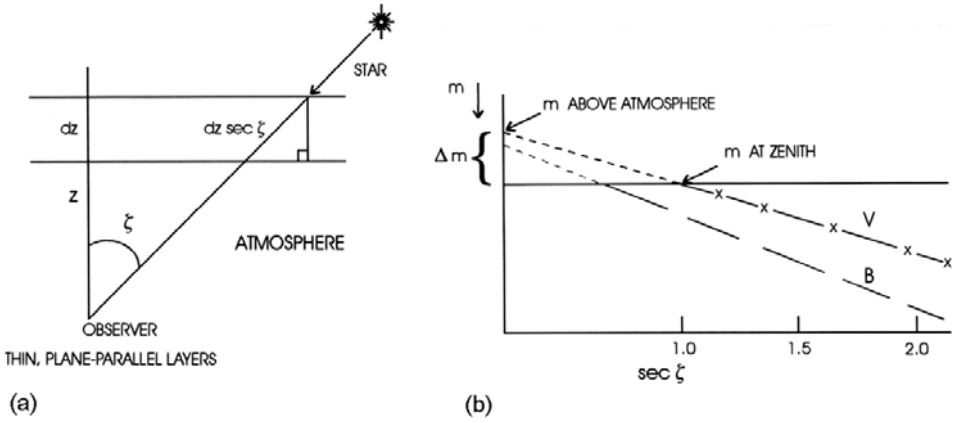


Figure 9.13. A plot of observed magnitude against airmass to determine the extinction coefficients and zeropoints.

airmass; at the zenith, airmass $(X) = 1$, whereas at an altitude of 30° above the horizon $X = 2$.

The resulting equation for the magnitude (m) of a star is given by:

$$m = -2.5\log(\text{counts/s}) - \alpha * (\text{airmass}) + \beta * (\text{color}) + ZP \quad (9.21)$$

Peter Stetson (Herzberg Institute of Astrophysics, Canada), who developed the DAOPHOT package, has shown that, during a long CCD exposure when the airmass has changed from X_0 at the beginning to X_1 at the end, the best estimate is a "weighted" mean given by

$$\bar{X} = \frac{(X_0 + 4X_{1/2} + X_1)}{6} + O(e) \quad (9.22)$$

where $X_{1/2}$ is the airmass at the midpoint of the integration; and $O(e)$ is a small error of about 1 part in 10,000. Color correction is still tricky because even two stars at the same airmass will undergo different extinctions if they are not identical in color.

9.7 SPECTROSCOPY

Spectroscopic applications put the most stringent requirements on detectors and drive the quest for the lowest possible noise performance. A majority of large astronomical spectrographs employ diffraction gratings to disperse light, although prisms with transmission gratings applied to them are also frequently used. Figure 9.14 shows the "image" of a spectrum on the detector. A cross-dispersed echelle spectrum that fills the detector is shown in Figure 9.15. Spectroscopic calibrations proceed in much the same way as with imaging. A flat-field is required to remove optical interference effects caused by the near-monochromatic light and variations in the

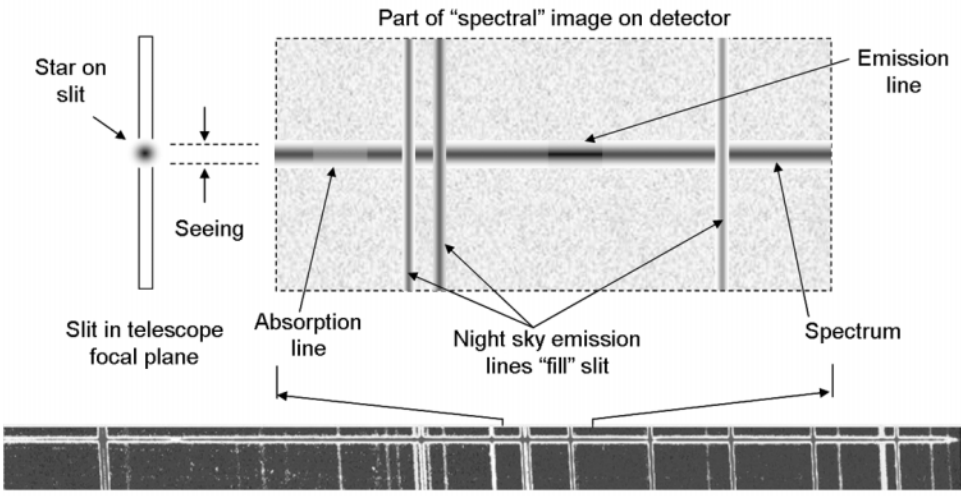


Figure 9.14. An illustration of the typical appearance of a spectrum on an array detector, including the presence of night-sky lines. The width is determined by the seeing.

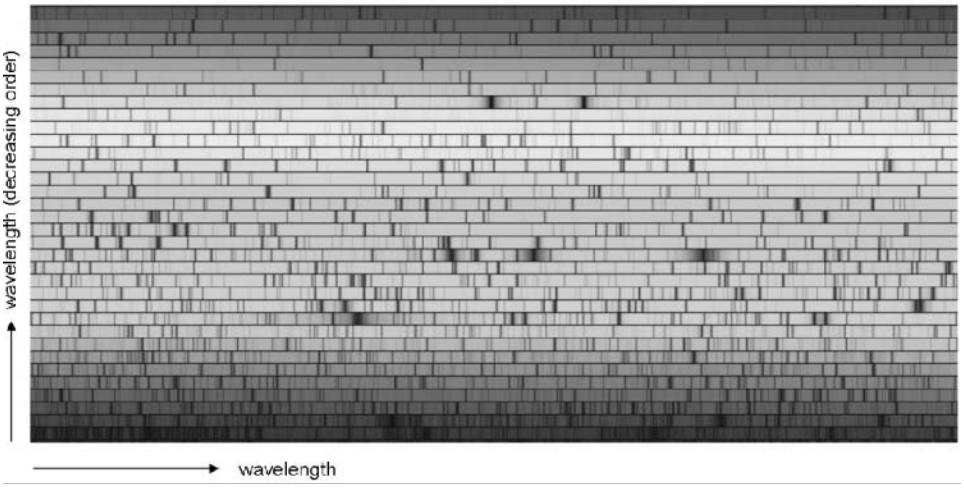


Figure 9.15. A cross-dispersed echelle spectrometer fills the detector array with many spectral segments.

thickness of thinned backside–illuminated CCDs. Bias frames must precede the derivation of the flat-field as before, but now, because of the much weaker signals and longer integrations, dark current may be more serious. Many dark frames may need to be averaged and subtracted from the object frame. The flat-fielded spectra must be ''sky-subtracted''. To do this it is normal to collapse or sum together all rows of the spectrum containing source flux and all rows containing sky spectra. After allowing for the fact that there will probably be more rows of sky spectrum than

source spectrum, the pair of flat-fielded, summed spectra are subtracted. Next, the relationship between pixel number and wavelength is determined using arc lamp (emission-line) spectra containing numerous lines with accurately known wavelengths. In the near-infrared it is often convenient to use the numerous, sharp OH night-sky lines for wavelength calibration as they are already "built in" to the observed spectrum. Atmospheric extinction corrections are applied to the intensity, and correction to absolute flux levels is accomplished by forming the ratio of the observed spectrum to that of a flux standard. Stars of Type A0 V that are not too distant and therefore not reddened by interstellar dust are used. Fainter objects, such as white dwarfs which have almost featureless spectra, are also utilized as flux standards. A summary of the key steps is as follows:

(1) Identify the direction of dispersion (is increasing wavelength the same as increasing pixel numbers?)
(2) Interpolate over dead pixels or columns to prevent these extreme deviant values from ruining the subsequent steps, but keep a "map" of these locations so as not to forget that there was no real data in those pixels.
(3) Sum up and normalize to unity the flat-fields. Flat-fields are sometimes taken with the spectrograph slit wide open if that is an option. In this way, the orders overlap considerably giving a uniform illumination on the CCD when viewing a quartz lamp illuminating a white screen on the inside of the dome.
(4) Divide the observed stellar spectra by the flat-field to remove pixel-to-pixel sensitivity variations.
(5) Some software packages (e.g., IRAF) require that you define the positions of the orders across the CCD. This may require observations of a brighter star if the program star is too faint. The program (e.g., "aptrace") will then know where to find spectra.
(6) Extract the rectangular subsets of CCD pixels corresponding to the stellar spectrum. Do the same for the arc lamp using the normal slit width and quartz lamp exposures. In IRAF this is done with a program called "apsum".
(7) Divide the flat-fielded stellar spectra by the "white-light" spectrum obtained with the normal slit and quartz lamp. This white light spectrum must itself be flat-fielded with the open-slit flat-field before this division. The purpose of this step is to remove interference fringe effects rather than pixel-to-pixel variations in sensitivity.
(8) Identify emission lines in the arc lamp used as a wavelength calibration device. The most commonly used lamp is a thorium–argon lamp spectrum. This is a painstaking step, but an important one if accurate transformations from pixels to wavelengths are required.

One feature of CCDs of special use to spectroscopy is on-chip binning. In this mode the extension of the spectrum along the length of the entrance slit, and therefore at right angles to the dispersed spectrum itself, can be summed into one pixel while still on the CCD by a suitable clocking sequence. The readout noise associated with this new binned pixel will be approximately the same as before even though the

charges from several pixels have been summed together. Thus, apart from compressing the size of the image frame to be handled by the computer, a gain in signal-to-noise ratio is obtained. Care must be taken not to saturate the CCD by summing too many charges.

9.8 POLARIMETRY

Array devices used in polarimeters and spectropolarimeters are reduced in a similar way to achieve a basic set of clean frames before any other operations (e.g., determining Stokes parameters) are carried out. In the case of polarimeters, a set of flat-fields corresponding to different positions of the polarimeter's waveplate are needed; as this element rotates in the beam then dust spots and other artifacts cannot be flat-fielded with just one waveplate orientation. After flat-fielding, the images must be very carefully registered to a tiny fraction of a pixel before forming the difference between pairs of frames corresponding to orthogonal polarization states. For example, if a half-waveplate is used to determine the linear polarization components $Q/I = p\cos 2\Theta$ and $U/I = p\sin 2\Theta$, where I is the total intensity, p is the percentage or fractional linear polarization, and Θ is the direction of vibration of the electric vector of the linearly polarized component, then Q/I and U/I can be measured by the difference in counts at four waveplate rotations from an arbitrary starting point.

Stars or objects of known polarization (e.g., the Crab Nebula) should be observed to verify the efficiency of the measurement and determine any scale factor. In addition, completely unpolarized sources should be observed to discover the "instrumental polarization" inherent in the system. The instrumental values of Q and U should be subtracted before p and Θ are calculated. More details can be found in Tinbergen (1996). See also the web reference for polarization.

9.9 SIGNAL-TO-NOISE CALCULATIONS

The calculation of the signal-to-noise ratio (S/N) of a CCD-type detector can be complicated depending on how closely one wishes to model the CCD's behavior. Here, we will assume that the CCD electronics system is perfect, and so there are no other sources of unwanted noise save the irreducible minimum readout noise. The simplest approach to analyzing a CCD image is to construct a "final frame" by performing the following steps to subtract dark current and normalize with a flat-field:

$$\text{FINAL FRAME} = \frac{\text{OBJECT FRAME} - \text{DARK FRAME}}{\text{FLAT FRAME} - \text{DARK FRAME}}$$

The dark-subtracted flat-field frame in the denominator is usually "normalized" beforehand by dividing each pixel value with a constant equal to the mean or median of the entire frame; this step is included in the noise formulation below. For infrared arrays, the (brighter) night sky is often used as the flat-field source and so another

version of this equation would substitute SKY FRAME for FLAT FRAME because the "sky–flat" is derived from many dithered frames of a relatively blank field of sky.

We assume that the only sources of "noise" are as follows:

(1) readout noise, R electrons
(2) photon (Poisson) noise on the signal (S) from the object
(3) photon (Poisson) noise on the signal (B) from the sky background
(4) shot noise on the dark-current signal (D).

The signals S, B, and D are to be given in electrons per second and will be derived from observed counts in data numbers (DNs) by multiplying by the transfer factor g electrons/DN. In addition to the photon-generated signals S, B we assume that the dark current D behaves according to Poisson statistics, so that the noise is equal to the square root of the number of electrons recorded. To find the original photon arrival rate for S or B divide by the quantum efficiency (η) (e.g., $S = \eta N_S$).

For random independent noise sources, the noise terms can be added in quadrature (i.e., by square and add), and we must allow for the noise sources in each of the two calibration frames, the SKY/FLAT and DARK. We can further assume that the number of OBJECT, SKY/FLAT, and DARK frames combined or "co-added" to form the FINAL FRAME are not necessarily equal, but for simplicity we assume that the normal practice pertains of keeping the same exposure time (t) for each image. This approach and terminology will work for CCDs and infrared arrays. Suppose we have n_O OBJECT FRAMES, n_B SKY BACKGROUND or FLAT FRAMES, and n_D DARK FRAMES. Let

$T = tn_O$ be the TOTAL integration time accumulated on the OBJECT FRAMES;
$f =$ the ratio of the source signal to that of the "background" signal per pixel;
$\varepsilon_B = n_O/n_B$ is the ratio of the number of object frames to background frames;
$\varepsilon_D = n_O/n_D$ is the ratio of the number of object to dark-current frames;

then, for a source covering n pixels on the CCD, we can estimate a total signal-to-noise ratio (S/N) of

$$\frac{S}{N} = S\sqrt{T}\left[u_r^2 + S + \sum_{i=1}^{n}\left\{\begin{array}{c}\left(B + D + \dfrac{R^2}{t}\right) + \varepsilon_D\left(D + \dfrac{R^2}{t}\right) + \\ + (1+f)^2\varepsilon_B\left(B + D + \dfrac{R^2}{t}\right) + (1+f)^2\varepsilon_D\left(D + \dfrac{R^2}{t}\right)\end{array}\right\}\right]^{-1/2} \tag{9.23}$$

where we now define S as the *total* object signal summed over n pixels (i.e., $S = \sum(S_i)$); and the term u_r represents the average, over n pixels, of any residual

error due to failure in the flat-field. A derivation of this equation is given by McCaughrean (1988). Note that B, D, and f are not the same from pixel to pixel. The terms in the denominator of this equation can be understood as follows. The first term is due to the residual non-uniformity (u_r), and the second term is the Poisson noise in the source signal itself. The first term following the summation sign, $(B + D + R^2/t)$, is the background and dark-current Poisson noise in quadrature with the readout noise in the *raw* source frame, and is always present. The next term contains only dark current and readout noise and is the error due to subtracting a dark frame from the object frame; the more dark frames that are averaged, the smaller is ε_D and the less significant this term. The third term comes from the application of the SKY/FLAT correction and the fourth term is the result of dark subtraction for the SKY/FLAT term. Again, the more SKY/FLATS that are used the smaller ε_B. Note that the additional scaling factor of $(1 + f)^2$ arises when the object frame is divided by the flat-field and the ratio renormalized by multiplying by the mean of the flat-field. If the flat-field signal is not large enough ($\varepsilon_B \ll 1$), then this term dominates for very bright sources, and for very faint sources it contributes as much noise as the raw source frame.

We need to distinguish the cases of point sources and extended objects. A stellar object has a seeing disk of diameter θ_{FWHM} arcseconds, corresponding to the full width at half maximum intensity (FWHM), whereas each CCD pixel is a square θ_{pix} arcseconds on a side. The number of pixels covered by the star's image is approximately

$$n_{\mathrm{pix}} = \frac{\pi}{4}\left(\frac{\theta_{\mathrm{FWHM}}}{\theta_{\mathrm{pix}}}\right)^2$$

and the summations ($\sum$) shown in Equation (9.23) must be taken over this number of pixels to estimate the S/N ratio. Over a small number of pixels it is usual to assume that the background is uniform (i.e., $\sum B_i = n_{\mathrm{pix}} B$). If the angular size of the object is much larger than the seeing diameter, then it is more convenient to deal with "surface brightness" in magnitudes per square arcsecond (as is done for the sky), and so $n_{\mathrm{pix}} = 1$ and each pixel is treated separately.

In the ideal case with accurate calibration data so that ε_B and ε_D are very small indeed, and if u_r, the residual non-flatness, is negligible then the equation is simpler:

$$\frac{S}{N} = \frac{S\sqrt{n_o t}}{\sqrt{S + n_{\mathrm{pix}}\left(S + B + \dfrac{R^2}{t}\right)}} \tag{9.24}$$

This is the basic form used to estimate observing times and limiting magnitudes, and is sometimes called the "CCD equation". If only one exposure is used, then $n_o = 1$. If digitization noise is a concern then the "effective" readout noise should be used as discussed earlier (p. 302). There are two further simplifying cases:

1. Background-limited or "sky-limited" case: B very much larger than $(D + R^2/t)$, and S is much weaker than B:

$$\frac{S}{N} = S\sqrt{T}[(n_{\text{pix}}B)]^{-1/2} \tag{9.25}$$

In this case, the accuracy of the measurement scales only as the square root of the integration time, an improvement in S/N by a factor of 5 implies 25 times longer in integration time. For a given signal level S, the S/N decreases as the square root of the increasing background. This is why it is much harder to get deep images at infrared wavelengths where the backgrounds are higher. Also, the signal-to-noise ratio scales only as the *diameter* (D_{tel}) of the telescope (because $S/\sqrt{B} \sim D^2/D = D$). Similarly, the signal-to-noise ratio scales only as the square root of the detector quantum efficiency.

Example. Suppose that the total source (S) is only 1% of the brightness of the sky in a single pixel and the source is spread over 4 pixels. To just barely detect this source requires $S/N = 1$. Therefore, $BT = 40{,}000$ photoelectrons. If the sky background (B) gives about 400 electrons/s per pixel, then this observation will take 100 s. Note that for R^2/t to be negligible (1% of B, say ~4 electrons/s), then R^2 must be less than $4t$ or 400 in this case. That is, $R < 20$ electrons rms if the entire observation is made in one exposure ($t = T$).

2. Detector noise-limited case: (R^2/t) very much larger than $(B + D)$, and a weak signal S:

$$\frac{S}{N} = \frac{S\sqrt{T}}{\left[n_{\text{pix}}\left(\dfrac{R^2}{t}\right)\right]^{1/2}} = \frac{St}{R}\sqrt{\frac{n_o}{n_{\text{pix}}}} \tag{9.26}$$

which shows that the signal-to-noise ratio for a given source S, and detector noise R, scales linearly with the *on-chip* integration time and only as the square root of the number of repetitions. The signal-to-noise ratio also scales linearly with quantum efficiency (η) and with telescope collecting area (A) because the signal is proportional to area and quantum efficiency. One should try to select t to achieve background-limited operation, but if this is not possible then it is best to maximize t, and minimize oversampling by reducing n_{pix}.

Example. In a high-resolution spectrograph the background plus dark current is 0.1 electrons/s/pixel, and so for $t = 1{,}000$ s we are formally readnoise-limited if $R^2/t > 0.1$ or $R^2 > 100$ (i.e., if R is larger than 10 electrons).

The above treatment shows that the highest signal-to-noise ratios are obtained when

(a) sufficiently accurate calibration frames are obtained
(b) the readout noise and dark current are as small as possible
(c) the on-chip integration time is as long as possible
(d) quantum efficiency and telescope area are as large as possible.

Table 9.4. Absolute flux from a zero-magnitude star like Vega.

Symbol	λ (μm)	ν (Hz)	F_λ ($W\,cm^{-2}\,\mu m^{-1}$)	F_ν (Jy)*
U	0.36	8.3×10^{14}	4.35×10^{-12}	1,880
B	0.43	7.0×10^{14}	7.20×10^{-12}	4,440
V	0.54	5.6×10^{14}	3.92×10^{-12}	3,810
R	0.70	4.3×10^{14}	1.76×10^{-12}	2,880
I	0.80	3.7×10^{14}	1.20×10^{-12}	2,500
J	1.25	2.4×10^{14}	2.90×10^{-13}	1,520
H	1.65	1.8×10^{14}	1.08×10^{-13}	980
K	2.2	1.36×10^{14}	3.8×10^{-14}	620
L	3.5	8.6×10^{13}	6.9×10^{-15}	280
M	4.8	6.3×10^{13}	2.0×10^{-15}	153
N	9.1	3.0×10^{13}	1.09×10^{-16}	37

* These units are called jansky. One jansky equals $10^{-26}\,W\,m^{-2}\,Hz^{-1}$, or $(3 \times 10^{-16}/\lambda^2)\,W\,cm^{-2}\,\mu m^{-1}$ if λ is expressed in microns.

The power (in watts) collected by a telescope of area A_{tel} (cm^2) in a wavelength interval of $\Delta\lambda$ (μm) from a source of apparent magnitude m (below the Earth's atmosphere), transmitted by an optical system of efficiency τ (<1) onto a CCD detector of quantum efficiency η (<1) can be simply estimated by

$$P(\lambda) = \tau(\lambda)\eta(\lambda)A_{\text{tel}}\,\Delta\lambda\,F_\lambda(0) \times 10^{-0.4m} \quad \text{W} \tag{9.27}$$

where $F_\lambda(0)$ is the flux in $W\,cm^{-2}\,\mu m^{-1}$ (see Table 9.4) from a zeroth magnitude standard star above the atmosphere. The transmission factor τ is the product of all the transmission or reflectance factors in the system. For instance,

$$\tau = \tau_{\text{tel}}\tau_{\text{optics}}\tau_{\text{filters}}$$

Example. Two telescope mirrors with a 95% reflectance each, six lenses with 96% transmission each, and a filter with 80% transmission. The total transmission is $(0.95)^2(0.96)^6(0.8) = 0.57$.

Remember to subtract out the area of the central hole in a Cassegrain telescope when evaluating A_{tel} and allow for any stops in the system which reduce the effective aperture of the primary mirror; one example would be an undersized secondary on an infrared telescope. Note that this equation assumes that the filter which defines $\Delta\lambda$ has a "box-car" profile of exactly this width and that the spectrum of the source scales

with that of the standard star, neither of which are true in practice. A more rigorous treatment requires that each profile be defined as a function of wavelength and the total power is then found by integration.

Since the energy of a single photon of wavelength λ is just hc/λ joules (1 W = 1 J/s) per photon, then the photoelectron detection rate is

$$S(\lambda) = (hc)^{-1}\tau(\lambda)\eta(\lambda)A_{\text{tel}}\lambda\,\Delta\lambda\,F_\lambda(0) \times 10^{-0.4m} \quad \text{electrons/s} \tag{9.28}$$

where h is Planck's constant; c is the speed of light and therefore $(hc)^{-1} = 5.03 \times 10^{18}\,\text{J}^{-1}\,\mu\text{m}^{-1}$. Dividing by g electrons per DN gives the observed signal rate in counts per second (or DN/s). If we set the count rate at 1 DN/s then we can derive the corresponding magnitude m_{zp} which is the "zeropoint" of the instrumental scale,

$$m_{\text{zp}} = 2.5 \log\left\{\frac{\tau\eta\lambda\,\Delta\lambda\,A_{\text{tel}}F_\lambda(0)}{hcg}\right\} \tag{9.29}$$

The zeropoint m_{zp} can also be derived from observations of a standard star of known magnitude. Usually, one plots observed instrumental magnitudes, $-2.5\log\{S(\text{DN/s})\}$ against airmass ($X = \sec\zeta$) and extrapolates the line to $X = 0$ (above the atmosphere) to get m_0. The zeropoint is $m_{\text{zp}} = m_0 + m_{\text{true}}$. Note: the zeropoint is positive and a numerically larger number. Having obtained m_{zp} by observations we can then derive the product $\tau\eta$ which describes the system efficiency in the given passband:

$$2.5\log(\tau\eta) = m_{\text{zp}} - 2.5\log\left\{\frac{\lambda\,\Delta\lambda\,A_{\text{tel}}F_\lambda(0)}{hcg}\right\} \tag{9.30}$$

Example. In the K′-band ($\lambda = 2.125\,\mu\text{m}$, $\Delta\lambda = 0.35\,\mu\text{m}$) the zeropoint is observed to be 20.4 on a telescope with an area of 72,236 cm^2 using a camera with a gain of 25 electrons/DN. Taking $F_\lambda(0) = 4.34 \times 10^{-14}\,\text{W cm}^{-2}\,\mu\text{m}^{-1}$ gives

$$2.5\log(\tau\eta) = 20.4 - 2.5\log\left\{\frac{2.125 \times 0.35 \times 72{,}236 \times 4.34 \times 10^{-14}}{1.99 \times 10^{-19} \times 25}\right\} = -1.28$$

which corresponds to $\tau\eta = 0.31$ (or 31%).

The signal from the sky or background (which includes thermal emission from the telescope at infrared wavelengths) depends on many things, including local conditions (temperature) and the amount of moonlight (phase of the moon and angular separation from the moon). A simple approach is to use the same form as before

$$B(\lambda) = (hc)^{-1}\tau(\lambda)\eta(\lambda)A_{\text{tel}}\lambda\,\Delta\lambda\,F_\lambda(0) \times 10^{-0.4m_{\text{sky}}}\theta_{\text{pix}}^2 \quad \text{electrons/s} \tag{9.31}$$

but with m_{sky} representing the sky brightness empirically as a magnitude per square arcsecond relative to a star of zero magnitude above the atmosphere, and introducing the pixel area on the sky θ_{pix}^2. In the thermal infrared, F_λ can be replaced with the Planck function ($B(\lambda)$) and an emissivity factor (ε).

Since B is proportional to $A_{\text{tel}}\theta^2$ then,

$$A_{\text{tel}}\theta_{\text{pix}}^2 = \frac{\pi}{4} D_{\text{tel}}^2 \left(206{,}265 \frac{d_{\text{pix}}}{D_{\text{tel}}(f/\#)_{\text{cam}}} \right)^2 \propto \frac{1}{(f/\#)_{\text{cam}}^2}$$

where d_{pix} is the pixel size of the detector and $(f/\#)_{\text{cam}}$ is the focal ratio of the camera system, which differs only by the magnification factor m from the telescope focal ratio. Many optical telescopes have focal ratios of about $f/8$ whereas infrared telescopes use $f/36$ or larger to reduce the background. For a given camera system, the background is independent of telescope diameter and depends only on focal ratio; two telescopes of the same focal ratio will provide the same background, although the plate scales will be different. At a good site with no moonlight, m_{sky} is about 22–23 in the blue, 21 in the visible and red, about 20 or brighter (and variable) near 1 μm, and about 13 at 2.2 μm, assuming the standard passbands.

Example. Consider a CCD camera with $\tau = 0.5$ and $\eta = 0.5$ at a wavelength of 0.54 μm in the middle of the visible spectrum with a filter passband of $\Delta\lambda = 0.1$ μm on a 4 m class telescope (i.e., a 4 m diameter primary mirror with a 1 m Cassegrain hole in its center). With the appropriate value of $F_\lambda(0)$ from Table 9.4 we get

$$S = 3.136 \times 10^{10} \times 10^{-0.4m} \quad \text{electrons/s}$$

so that for $m = 17$ we obtain 5,000 electrons/s whereas for the extremely faint sources seen on deep CCD images at $m = 27$, we detect only 0.5 electrons per second.

A useful expression for predicting the "limiting magnitude" of an array camera in the background-limited or sky-limited case is to solve for S in terms of S/N, convert to counts by dividing by g (electrons/DN), and substituting this equation into the definition of the instrumental magnitude to give

$$m = m_{\text{zp}} - 2.5 \log \left\{ \frac{1}{g} \frac{S}{N} \sqrt{\frac{n_{\text{pix}} B}{T}} \right\} \tag{9.32}$$

The "limit" corresponds to $S/N = 1$.

Example. A camera with a gain of 10 electrons/DN and a zeropoint of $m_{\text{zp}} = 18$ forms a star image across 5 pixels with an average background of 200 electron/s/pixel. What is the limiting magnitude for a 1-hour exposure?

As we are assuming the background-limited case, we are not concerned with the way the exposure has been broken up; all that matters is $T = t = 3{,}600$ s and $S/N = 1$:

$$m = 18 - 2.5 \log\{0.1 \times 1 \times \sqrt{(1{,}000/3{,}600)}\} = 18 + 3.2 = 21.2$$

In reality, signal-to-noise calculations are often more complex than presented here because star images are not sharp-edged, filter transmissions are not simply represented by a wavelength interval, and the stellar energy spectrum is not the same as that of Vega.

Example. Twenty magnitudes is a factor of 10^8; therefore, $m = 20$ corresponds to 38 μJy at V, 6.2 μJy at K, and 0.37 μJy at N.

9.10 SUMMARY

The important parameters for a CCD system include the detected quantum efficiency, readout noise, uniformity (flat-field), linearity, and gain. We have shown how to determine all of these important parameters in a CCD, or similar detector array, by following the steps from quantum detection of photons to evaluation of fluxes, magnitudes, and signal-to-noise ratios by processing data frames. Although the steps may seem complicated at first sight, the procedures are routine. Good correction of the inherent pixel-to-pixel variations in a CCD by a carefully constructed flat-field can enable astronomers to detect extremely faint sources that are not apparent in the raw image. Relative photometry to thousandths of one-magnitude accuracy is possible. We have also reviewed how to calculate signals and backgrounds for a CCD camera and presented a formula for the signal-to-noise ratio often called the "CCD equation". The results of this chapter are directly applicable to infrared arrays too. When all these steps are followed, CCDs and infrared detectors become extremely powerful for quantitative astrophysics in photometers and spectrographs.

9.11 EXERCISES

1 Explain the difference between QE and DQE. Produce additional entries in Table 9.1 for CCDs with 5 and 15 electrons of readout noise.

2 A sequence of flat-field observations is obtained at different exposure levels. The following signal and variance values were found: 200 DN, 75; 400 DN, 125; 800 DN, 200; 1,800 DN, 475. Plot the data and derive the gain factor g (electrons/DN) and the readout noise R.

3 Derive the amplifier gain A_g for a detector with a capacitance of 0.1 pF and a source follower gain of 0.75, if a 16-bit A/D with a 10-volt swing is used and the value of $g = 25$ electrons/DN.

4 Explain the concept of drift scanning. Why does it produce a good flat-field along the scan direction?

5 What is the difference between flat-fielding and fringe correction?

6 Explain the two approaches of extracting magnitudes from CCD and infrared images.

7 What is meant by the "zeropoint" of a magnitude scale?

8 Describe the major steps needed to calibrate a high-resolution spectrograph with a CCD or IR array detector.

9 What additional procedures are required to calibrate a CCD-based polarimeter camera?

10 Calculate the photon arrival rate for a 24th-magnitude star in the V-band on a 4 m telescope with a camera having an efficiency of 30%. Assuming that the pixels are 0.3 arcseconds, and the readout noise is 10 electrons, and dark current is negligible, is the measurement background-limited? What integration time is required to achieve a signal-to-noise ratio of 10?

9.12 REFERENCES

Craine, E.; and Crawford, D. (eds.) (1999) *Precision CCD Photometry*, ASP Conference Series, Vol. 189, Astronomical Society of the Pacific, San Francisco. [Many papers.]

Davis, L.E. (1994), *A Reference Guide to the IRAF/DAOPHOT Package*, NOAO, Tucson, AZ.

Everett, M.E.; and Howell, S.B. (2001), A technique for ultrahigh-precision CCD photometry, *Publ. Astron. Soc. Pacific*, **113**, 1428–1435.

Gunn, J.E.; Carr, M.; Rockosi, C.; Sekiguchi, M.; Berry, K.; Elms, B., de Haas, E., Ivezić, Ž.; Knapp, G., Lupton, R. *et al.* (1998) The Sloan Digital Sky Survey Photometric Camera, *Astron. J.*, **116**, 3040–3081.

Hall, P.; and Mackay, C.D. (1984) Faint galaxy number–magnitude counts at high galactic latitude, *Mon. Not. R. Astron. Soc.*, **210**, 979–992.

Henden, A.A.; and Kaitchuk, R.H. (1990) *Astronomical Photometry*, Wilmann-Bell, Richmond, VA.

Howell, S.B. (ed.) (1992) *Astronomical CCD Observing and Reduction Techniques*, ASP Conference Series, Vol. 23, Astronomical Society of the Pacific, San Francisco.

Howell, S.B. (2006) *Handbook of CCD Astronomy*, second edition, Cambridge University Press, Cambridge, U.K., ch. 5, pp. 102–134.

Janesick, J.R.; Hynecek, J.; and Blouke, M.M. (1981) A virtual phase imager for Galileo, in *Solid State Imagers for Astronomy*, *SPIE*, **290**, 165–173.

Jaschek, C.; and Murtagh, F. (eds.) (1989) *Errors, Bias and Uncertainties*, Cambridge University Press, Cambridge, U.K.

Johnson, H.L.; and Morgan, W.W. (1953) Fundamental stellar photometry for standards of spectral type on the revised system of the Yerkes spectral ATLAS, *Astrophys. J.*, **117**, 313–352.

Léna, P. (1988) *Observational Astrophysics*, Springer-Verlag, Berlin.

McCaughrean, M.J. (1988) The astronomical application of infrared array detectors, PhD thesis, University of Edinburgh, U.K.

Oke, J.B. (1974) Absolute spectral energy distributions for white dwards, *Astrophys. J. Supplement*, **27**, 21–35.

Stetson, P.B. (1987) DAOPHOT: A computer program for crowded-field stellar photometry, *Publ. Astron. Soc. Pacific*, **99**, 191–222.

Tinbergen, J. (1996) *Astronomical Polarimetry*, Cambridge University Press, Cambridge, U.K.

Tyson, J.A.; and Seitzer, P. (1988) A deep CCD survey of 12 high-latitude fields, *Astrophys. J.*, **335**, 552–583.

Web sites

SDSS home page: *http://www.sdss.org/*
Bonn Shutters home page: *http://www.astro.uni-bonn.de/~ccd/shutters/index.php*
Polarization: *http://www.jach.hawaii.edu/UKIRT/instruments/irpol/irpol-stds.html*

10

Image processing and analysis

Computers are used in the control of telescopes and instruments, for acquisition of digital data from electronic detectors, for image display, analysis, numerical simulations, and more. In this chapter we present some important terminology and explain the standard astronomical data file format known as FITS. Astronomical software is a vast subject beyond the scope of a single chapter. Our goal here is to introduce some well-known software packages but focus more on basic concepts including image visualization, high-pass and low-pass filtering, false color and image restoration by deconvolution.

10.1 COMPUTERS

The growth rate of computer technology is enormous. Keeping up to date is best done by watching for reviews and surveys in both trade and popular magazines. Our emphasis here is on those aspects of computer technology likely to be of most interest to astronomers, such as acquisition of digital data from a detector, image analysis software, and data storage facilities. Sections 10.1.1 and 10.1.2 introduce some terminology.

10.1.1 Data acquisition and data transfer

The "data rate" is the number of digitized pixel values transferred per second to the host computer. For an array detector, this rate is largely determined by the electronic configuration, although the actual application and device physics may also be factors. Recall that the integrating correlated double-sampling (CDS) method used with CCDs requires an interval of time to digitize a pixel value, typically about 20 μs to 100 μs, corresponding to pixel readout rates of 50 kHz and 10 kHz, respectively. For a 20 μs pixel time and digitization to 16 bits (65,536 voltage levels, or 0 to 65,535), the

maximum output data rate is 800,000 bits/second. For detector systems running at 5 MHz however, the rate would be 100 times larger. Note that this level of digitization is mandatory if one is to capitalize on the large dynamic range of CCDs. So this data rate is 100 kilobytes/s. Serial data rates are usually specified in bits per second (bps) and parallel rates in Mbytes/s. Driven by diverse applications ranging from audio reproduction to military radar, analog-to-digital converters (ADCs) are getting faster each year. When selecting an ADC it is important to look carefully at performance details. A key parameter is the signal-to-noise ratio in decibels (db), more is better. Today's low-noise CCDs can reveal limitations in ADCs. For example, as the digitization process is discrete then signals will be rounded up or down by 0.5 DN, leading to digitization noise. Also, if the conversion "accuracy" or signal-to-noise ratio is equivalent to several bits (e.g., 2 bits = 4 DN) then at some ADC step the error will be $\pm 4g$ electrons, where g is the number of electrons per DN, but which step has this error will be unknown. For a 16-bit ADC with a 1-bit error (>90 db signal-to-noise ratio) then the effect is only 0.003%. Transfer of data from the ADC to the computer can be accomplished in two ways. Either by a high-speed serial link which is a single line along which signals are sent in a sequential pattern, or by a multi-way cable called a parallel link which sends all 16 bits at once.

One of the most commonly encountered serial standards is RS-232-C in which a "logic 1" level is represented by a voltage in the range -5 V to -15 V, and "logic 0" by a voltage between $+5$ V and $+15$ V; up to 25 lines are specified, although seldom used. In the RS-423-C standard, only two wires are used and the logic voltages are 0 V and 5 V. The RS-422-C is similar but the lines are "balanced" to achieve much higher data rates. In personal computers, RS-232 has been superseded by USB (Universal Serial Bus) which is faster, has lower voltages and smaller connectors, and can also handle data transfer. The current version (USB 2.0) runs at 480 Mbit/s.

A rather common method of data transfer today is the use of a high-speed, single-cable serial link called Ethernet which is also used as a general means of interlinking several computers in a Local Area Network or LAN. In the original Ethernet approach (due originally to Xerox, DEC, and Intel) messages and data are broadcast along a coaxial cable from the sender station to the receiver station. To send a message, the sending station senses the cable to see if it is free. If it is free, the station transmits but it also continues to sense in case some other station on the net transmitted at the same time. If this were to happen a "collision" would have occurred and the transmission would be unsuccessful, so both stations stop sending, wait a random time interval, and then try again. This is called "carrier sense multiple access with collision detection" or CSMACD. Originally, coaxial cable was used but it was replaced by thin-wire Ethernet based on unshielded twisted pairs, with a system of linked hubs ultimately replacing the CSMA-CD method for a full duplex system in which devices at both ends can receive and transmit at the same time. Rates improve steadily, but 100 Mbits/s to 1 Gbits/s is common at present.

Of extreme importance has been the introduction of optical fibers. Optical transmission, which is insensitive to electrical interference, has rapidly emerged as the best high-speed, high-capacity, low-error communications link for data transmission. In an optical fiber, the electrical pulse which would have been sent down a

copper wire is converted to a pulse of light by a transmitter at one end and then back into an electrical pulse by a receiver at the other end. Typically, the optical fiber itself has a very small diameter of only 50 μm–75 μm.

For parallel data transfers, many computers handle this flow by using Direct Memory Access (DMA). Parallel data highways within a computer system are called "buses" and there are different buses for data transfer to memory, for control, and for addressing. The number of lines in the address bus determines the number of memory locations that can be specified. A 32-bit bus gives over 4 billion locations. Two commonly encountered systems are SCSI (pronounced "skuh-zee") and PCI. In general, digital data from an astronomical detector are saved initially on a magnetic disk drive (the "hard drive") and later transferred to another medium such as an optical disk (CD or DVD) for longer term storage; this process is called "archiving".

For compatibility between two devices, several parameters must match besides the rate: in particular, the voltage levels, the timing, the format, the "code" by which numbers and characters are digitized, and the rules or "protocols" by which data transfers are acknowledged, called "handshaking", and by which computer operations are tracked, a process called "housekeeping". One instrument control philosophy/environment known as EPICS, developed at the Los Alamos labs in the U.S.A., is being used at several large astronomical centers. EPICS is based on a transport layer called Channel Access, a message system protocol over Ethernet hardware, which connects Unix host computers to smaller systems running VxWorks (Wind River Systems, Alameda, CA), typically in a VME crate. VME is a bus system used in mid-size computers such as the early Sun Microsystems machines. Many different CPUs are available to run in a VME bus system including Motorola MC68000 and Sun Sparc families. VME is capable of very high speeds, extremely compact hardware, and receives widespread commercial support. The VME bus is designed for input–output paths which are 32 bits wide.

Another type of data acquisition system uses the IEEE-488 standard which is based on the Hewlett-Packard (Palo Alto, CA) general-purpose interface bus (GPIB). Input–output data paths are 8 bits wide (parallel), and so transfer of a 16-bit word consists of two serial bytes. Each unit on the bus is independent, and connections to the bus are by means of a standard 24-pin connector which allows other connectors to be added on to it to create a "stack" of several devices on the same bus. Instruments connected to the bus are designated talkers, listeners, and controllers. Again, a computer program is required to control the modules on the bus, but the support for this system is very widespread and many commercially available pieces of lab apparatus (such as oscilloscopes, temperature monitors, voltage–frequency converters) come with an IEEE-488 interface.

Many astronomical instruments employ powerful digital signal processors (DSPs) for detector and instrument control. DSP chips are made by several manufacturers (e.g., Motorola and Texas Instruments) and are widely used in equipment of all kinds, from cellular phones to automobiles. Several observatory-built CCD and infrared camera systems employ DSP chips as the "intelligence" in a controller/sequencer.

10.1.2 Data file formats

In a CCD system the "photon image" on the detector is converted to a "charge image" in the pixels. The (x, y) location of the charge image accurately mimics the location of the arriving photons and the amount of charge is (usually) linearly proportional to the number of photons at that location in the image. During the readout process each charge packet is converted to a voltage and then digitized with an analog-to-digital converter (ADC) in order to associate a number with each pixel location. If the ADC is a 16-bit device then the numbers representing the intensity in the image will range from 0 to 65,535. Those numbers are stored in a two-dimensional array or table in a computer such that the correct intensity is associated with the appropriate pixel (x, y). This is the "digital image". It may be stored on a disk or held in memory, but the CCD image is now an array of numbers arranged just like the columns and rows of the detector itself. Usually, the image on disk is a master copy that is protected from change. Copies can be manipulated by operations on the individual digital pixel values.

There are numerous formats for image files such as bitmap (BMP), GIF, TIFF, JPEG, and descriptions of these can be found by a simple web search. However, the recognized standard format among professional astronomers is the Flexible Image Transport System (FITS) developed by Don Wells, Eric Greisen, and Ron Harten (1981). Sometimes instruments save data in a "native" form consistent with the software environment at the observatory. Nevertheless, files are always converted to FITS for transport away from the observatory. FITS is used throughout astronomy. FITS files consist of three parts: a "header", the image "data" in binary form, and a "tailer". Unlike JPEG and GIF images, FITS images cannot be viewed in web browsers. Special display software is required.

A FITS "header" comprises an integer multiple of 36 lines of 80 bytes (the 80 bytes is a relic of 80-character punched cards) giving 2,880 bytes, or 5,760 bytes for 72 lines, and so on. If fewer than 36 lines are used, then the remainder must be filled out with the ASCII (American Standard Code for Information Interchange) character for a blank space (hexadecimal value of 20). Each line, also called a "card image", begins with a "keyword" in bytes 1 through 8, which identifies the information type for that line. The construction of the keywords is very specific. Each word must be left-justified and consist of only eight valid ASCII characters with no blank spaces except at the end, to pad out the keyword to eight characters if necessary. Longer keywords such as TELESCOPE are contracted to TELESCOP to remain within the eight-character limit. Only uppercase letters, the digits 0 through 9, periods, and hyphens are all that is allowed. Bytes 9 and 10 may contain an equal sign and a space if the keyword has an associated numerical or text value. Numerical values are always right-justified between bytes 11 through 30, whereas text strings begin with a single quote at byte 11 and must end (with a single quote) by byte 80. An optional "comment" can be added after the value if separated by a space followed by a slash (/). When a keyword has no associated value, then bytes 9 through 80 can contain any ASCII text characters. The following order of keywords is required: SIMPLE, BITPIX, NAXIS, NAXIS1, NAXIS2, . . ., NAXIS*n*, and END.

Table 10.1. FITS keywords and their meanings.

SIMPLE	Has the value in byte 30 of either T (true) of F (false). It is simply a statement of whether or not the file conforms to the FITS standard.
BITPIX	An integer describing the number of bits in the data values. Options are 8, 16, 32 for 8-bit, 16-bit, and 32-bit unsigned integers. Floating-point data can be represented (e.g., −32 and −64 for 32-bit and 64-bit, respectively).
NAXIS	The dimension of the data array. If value is zero, no data follows. Value of 1 for 1-D data such as intensity values in a spectrum. For image data, NAXIS = 2 (e.g. rows and columns of CCD), and NAXIS = 3 would be used for a data cube of spatial coordinates vs. velocity. The maximum value of NAXIS is 999.
NAXIS1, NAXIS2, NAXIS*n*	Each specify number of elements along that axis, with the convention that NAXIS1 is the axis whose index changes most rapidly and NAXIS*n* is the axis whose index changes the slowest. For example, in a CCD image the number of columns would go in NAXIS1 and the number of rows in NAXIS2.

Completing the header is the keyword END which is located in bytes 1–3 and the remaining fields (to 80) are filled with ASCII blanks. Several optional keywords may be inserted after NAXIS*n* and before END. For instance, BSCALE and BZERO relate the array values and the true values through the relation:

$$\text{true value} = \text{BSCALE} \times \text{array value} + \text{BZERO}$$

and can be used to convert signed 16-bit array values (−32,768 to +32,767) into unsigned 16-bit pixel values (0 to 65,535) by setting BZERO to 32,768.0 and BSCALE to 1.0. Standard self-explanatory additional keywords with associated character strings are OBJECT, TELESCOP, INSTRUME, and OBSERVER. DATE-OBS and DATE have character string values and are intended to record the date on which the observations were obtained and the date on which the header was written, respectively. The usual format for the date is dd/mm/yy and Universal Time is preferred. The keyword ORIGIN is used with a character string to identify the institution creating the FITS file. COMMENT and HISTORY are two keywords which do not have associated values, and any valid ASCII text can be inserted in bytes 9 through 80. Any number of COMMENT or HISTORY lines is allowed, consistent with the 36-line header. If more than 36 lines are required then the keyword EXTEND should be inserted before line 36, and unused lines up to 72 will need to be padded with the ASCII blank character.

Immediately following the header (at byte 2,881) begins the "data" in a continuous sequence according to the NAXIS parameters already declared. According to

the FITS standard, 8-bit integer data must be represented by unsigned binary integers contained in one byte, and 16-bit data values must be stored as signed binary integers in two bytes with the most significant byte first. This convention is not followed by many computers (especially PCs) or programs, and consequently "byte-swapping" may be needed to import FITS data into another system. Also, although 16-bit digitization is standard, many imaging systems use 12 bits, 14 bits, or 15 bits, and some use 24 bits. The BZERO keyword can be helpful in offsetting the zero point of the stored values and the 24-bit numbers would require to be handled using the 32-bit convention of four bytes with the most significant byte first. Finally, the "tailer" of a FITS file is ASCII null (00) characters used to pad out the final 2,880-byte record.

10.2 DATA REDUCTION AND ANALYSIS SYSTEMS

For most astronomers, image processing means simply "data reduction", and is largely associated with the simplest aspects of visualization and mathematical manipulation of the grid of numbers stored in the computer to represent the intensity $I(x, y)$ (photon counts) at each (x, y)-pixel location on the array detector during the given exposure time. Raw electronic imaging data may be uneven in appearance until it is correctly calibrated. Usually, the two-point correction of dark/bias subtraction and flat-field division is adequate. Most processing steps must be carried out pixel by pixel. For example, we will need to remove the pattern of dark current which was accumulated during the CCD integration. To do this we subtract a data array containing the "dark" frame from the data array containing the "raw" frame using a vector arithmetic algorithm which moves from pixel to pixel calculating the difference and entering the answer in a new data array. The process is illustrated by the simple graphic in Figure 10.1. Next, the dark-subtracted frame can be divided by another digital image representing the (normalized) flat-field. Again, using a pixel-by-pixel division, the number in the first data array is divided by the corresponding pixel value in the second data array, and the quotient is entered into that pixel location in the output data array. Of course, simple scalar arithmetic is also possible. For example, each and every pixel value can be multiplied (or divided) by a constant number and the product is entered into that pixel location in the output array. One

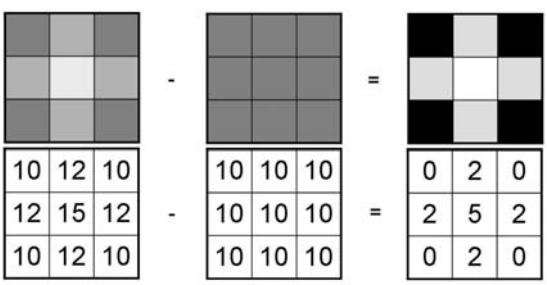


Figure 10.1. A simple operation on digital image data to subtract one frame from another.

example might be division by the exposure time so that the numbers now represent counts per second. Software is also required for a number of other steps needed to "reduce" the data to a form suitable for further analysis.

There are several major suites of computer programs which have been developed specifically to support the reduction of astronomical data, and especially for CCD-type imaging and spectroscopic systems. Most of these programs have been "packaged" within an environment which allows the users to select the appropriate task and even set up a sequence of tasks to be performed without writing and compiling computer code from scratch. The most well-known packages are

- AIPS
- IRAF
- STSDAS
- STARLINK
- MIDAS
- IDL

With the exception of IDL, which is a commercial package not specifically designed for astronomy but nonetheless suitable, all the others were produced by astronomers and are maintained by the astronomical community.

The Astronomical Image Processing System (AIPS) is a software package for interactive (and, optionally, batch) calibration and editing of radio interferometric data and for the calibration, construction, display, and analysis of astronomical images made from those data using Fourier synthesis methods. Design and development of the package began at the National Radio Astronomy Observatory (NRAO) in Charlottesville, Virginia in 1978. It presently consists of over 4,300 files containing 1.46 million lines of text. AIPS contains over 350 distinct application tasks. In 1997, considerable effort was expended to ensure that AIPS would be capable of handling the data from orbiting Very Long Baseline Interferometer satellites. AIPS has been the principal tool for display and analysis of radio images (both continuum "maps" and spectral-line "cubes") from the Very Large Array (VLA) since early in 1981. AIPS contains facilities for display and editing of data in the aperture, or u–v plane; for image construction by Fourier inversion; for deconvolution of the point source response by CLEAN and by maximum entropy methods; for image combination, filtering, and parameter estimation; and for a wide variety of image and graphical displays. In 1992, the NRAO joined a consortium of institutions seeking to replace all of the functionality of AIPS using modern coding techniques and languages. AIPS++ as it was initially called has been renamed CASA which is now (December 2007) in beta release. (More details can be found at: *http://casa.nrao.edu/*)

IRAF is the Image Reduction and Analysis Facility, a general-purpose software system for the reduction and analysis of astronomical data. Used extensively in the United States, IRAF goes beyond classic AIPS in the sense that every effort has been made to make this software system as "portable" and as device-independent as

possible. After an interruption, IRAF is again supported by the National Optical Astronomy Observatories (NOAO) in Tucson, Arizona. The main IRAF distribution includes a good selection of programs for general image processing and graphics, plus a large number of programs for the reduction and analysis of optical and IR astronomy data (the ''noao'' packages). Other external or layered packages are available for applications such as data handling from the Hubble Space Telescope (optical), EUVE (extreme ultra-violet), or ROSAT and Chandra (X-ray). These external packages are distributed separately from the main IRAF distribution but can be easily installed. The IRAF system also includes a complete programming environment for scientific applications, which includes a programmable Command Language scripting facility, the IMFORT Fortran/C programming interface, and the full programming environment in which the portable IRAF system and all applications are written.

The Space Telescope Science Data Analysis System (STSDAS) is a software package for reducing and analyzing astronomical data. It is layered on top of IRAF and provides general-purpose tools for astronomical data analysis as well as routines specifically designed for HST data. In particular, STSDAS contains all the programs used for the calibration and reduction of HST data in the post-observation processing pipelines. The concept of ''pipeline'' processing is now widespread, and both ground-based and space-based instruments tend to be delivered with software packages designed to reduce the raw data from the given instrument. Some of these software investments are enormous, and often there is much duplication of effort. As the number, complexity, and data rates of telescope/instrument combinations increases it becomes challenging to develop and maintain the associated pipelines. Some effort towards the development of a ''common pipeline library'' has been considered and developed by, for example, the European Southern Observatory.

The ESO-MIDAS system is built to allow easy integration of complex analysis algorithms as well as flexibility in interactive use and in the creation of user-specific procedures from the basic building blocks. The first design proposal for ESO-MIDAS, made late 1980, used some ideas from the U.K. STARLINK project for software interface definitions. The present version follows a similar philosophy in its application program interfaces, but has been expanded to the new Standard Interfaces which have a broader base than previously. Basic MIDAS was designed to work on a single computer with appropriate peripherals. It consists of the following three parts: the monitor, the applications, and the interfaces. MIDAS is command-line driven and the monitor functions as a command interpreter, keeps logs, and displays on-line help. Applications perform the actual data operations such as image display, image processing, fitting software, and data input/output management. The MIDAS Command Language (MCL) provides tools to construct complex command strings. For more details see the MIDAS web page and Warmels (1991).

STARLINK was set up in 1980 in the U.K. to help astronomers use computers to analyze their observations. The main objectives of the project were to coordinate the provision of interactive data reduction and analysis software tools, and to encourage sharing and standardization to minimize duplication of effort. The STARLINK software collection contains hundreds of items, including external software like IRAF. The main support facility was closed down in 2005 but continuing support

exists at the U.K. Joint Astronomy Centre in Hawaii. More details can be obtained from the STARLINK web site.

IDL (Interactive Data Language) is a complete visualization package for scientific, engineering, and medical data which incorporates a powerful array-oriented language to enable users to create complex functions, procedures, or applications without any conventional programming in FORTRAN or C. Operators and functions work on entire data arrays, and there is immediate compilation and execution of the IDL commands. Image display and graphics capabilities are fully integrated with the computations and a superb tool called IDL Widgets enables the user to create a graphical user interface (GUI) to IDL programs. This latter feature is particularly attractive to instrument builders. IDL is supported on UNIX, VMS, Microsoft Windows, and Macintosh systems. Many astronomers now use IDL for data simulation and modeling, and for the standard data reduction tasks (like flat-fielding) normally handled by IRAF, but they still tend to return to packages such as DAOPHOT for astronomical photometry.

10.2.1 The IRAF package

As the use of IRAF is very widespread it is worth providing a brief introduction to this system. The IRAF Command Language (CL) organizes the many system and application programs (called tasks) into a logical hierarchy of packages. A package is a collection of related tasks, and the major packages are

- system (system utilities)
- language (the Command Language itself)
- dataio (data input and output)
- images (general image processing)
- noao (astronomy packages from NOAO)
- plot (general graphics utilities).

The first two packages (system and language) include tasks that are available as soon as you enter IRAF. For tasks within the other packages you must first "enter" those packages by typing its name. Sub-packages exist within the parent package and can only be reached through their parent package. Below is a "tree" of some of the more commonly used tasks. Sub-packages are italicized. You will need to consult your local IRAF documentation for more options.

system	*language*	*dataio*	*images*	*noao*	*plot*
allocate	bye	rfits	hedit	*digiphot*	contour
deallocate	ehistory		imarith	*apphot*	
devstatus	eparam		imcopy	qphot	
directory	logout		imdelete	*imred*	
help	lparam		imheader	*ccdred*	
lprint	print		imhistogram	combine	
			imrename	*proto*	
			imstatistics	fixpix	
			tv		
			display		

Assuming that you have logged in to a computer running IRAF and changed to the IRAF directory (probably by typing cd iraf), you enter IRAF's command language by typing "cl" (without the quotes) and hitting the return key. The computer should return with a prompt cl>. Typing a "?" (note: we mean type the symbol without the quotes. This will apply to all further typing instructions in this section) will list the IRAF packages, and you can get help on any one of them by typing "help packagename" at the prompt. Similarly, typing "help taskname" will give information on the parameters that each task requires. If you execute a task unintentionally, then use the "Ctrl C" combination (the Ctrl key and the C key together) to terminate the task. To leave the sub-package and return to the parent package, type "bye", or alternatively use "Ctrl D".

Most likely, your first action will be to import your data into IRAF from a data tape or disk using the dataio package and rfits task. For instance,

```
cl >dataio
da>devstatus mycomputer!mta   [This checks whether or not the tape/disk drive is
                               already allocated.]
da>allocate mycomputer!mta
da>help rfits                 [Hit spacebar to advance a page; "q" to quit.]
da>epar rfits                 [epar or eparam is shorthand for edit parameter
                               settings associated with this task.]
```

Simply edit the displayed list by typing in the values you want and press "Ctrl D" to exit. For example, set

```
fits_file=mycomputer!mta
fits_list=1-999               [Or whatever you want.]
iraf_file=raw
```

This creates IRAF image files for every FITS file from 1 to 999 or as many files as are on the disk if it is less than this number, and name the files raw001.imh, raw002.imh, and so on:

```
da>rfits
FITS data source (mycomputer!mta):⟨return⟩
File list (1–999):⟨return⟩
Iraf filename (raw):⟨return⟩   [As you already set the parameters you can hit the
                               "Return" key at each prompt.]
```

When all the files have been read, there should be an indication of this and the prompt will return. Be sure to deallocate the device for the next user:

```
da>deallocate mycomputer!mta
da>bye
cl >
```

and remove the data tape or disk from the device.

The next step is to display the raw images on a screen. In IRAF this is done with a facility called SAOimage (pronounced "S-A-O image") which creates a display window separate from the IRAF command window. There are many SAOimage alternatives. For example, SAOImage DS9 is an astronomical data visualization application that supports FITS images, multiple frame buffers, and many scale algorithms and color maps). It is a stand-alone application and requires no installation or support files. Versions of DS9 currently exist for Sun Solaris, Sun Solaris64, Linux, Linux64, MacOSX, Darwin Intel and PPC, and Windows XP. Other display packages are

Skycat: ESO's next-generation image display program
Jsky: ESO's library of reusable Java components for astronomy
Fv: a self-contained interactive FITS file editor for Windows, Unix, and Mac OS-X.

From outside the IRAF window pull down a menu and select SAOimage; a new display window will be created. You can drag and re-size this window in the usual way. Go back to the IRAF window, click in it, and type the following in response to each prompt (remember to end with a carriage return):

```
cl >images
im>tv
tv >display imagename     [Where imagename might be raw012(.imh); try dev$pix
                          for a built-in image of M51.]
```

Note that there are actually four frame buffers available and you can specify one of these by adding the numbers 1–4 after imagename; no entry defaults to buffer 1. Also, it is not necessary to type the extension ".imh" after imagename.

SAOimage provides mouse-controlled buttons to select gray or false color (click the left mouse button on "color"), then choose a "color mapping" (click on "cmap"), choose any of the options "Gray", "BB", "HE", "18", "A", "B" by clicking on it, and finally change the appearance of the display (contrast and stretch) by holding down the left mouse button and dragging the mouse around. Watch the color bar at the bottom of the frame and reset at any time by clicking on "Gray" or "BB" or one of the others. You can magnify or reduce the display and you can "pan" around within the displayed image. All of these features are fairly typical of any image-processing package. Click back in the IRAF window when you wish to leave the display window.

IRAF can be used to perform all the steps necessary to reduce CCD images, such as dark subtraction, flat-fielding, registration, and photometry. Suppose you want to average a few dark frames and then subtract them from each of your image frames. You need to enter the ccd reduction sub-package "ccdred" from the major package "noao" and then execute the task called "combine". For example,

```
cc>help combine           [Read the on-line help.]
cc>epar combine           [Edit the parameter file. You will be concerned with the
```

parameters: images (the input data), output (name for final frame), combine (either an average or a median), exposure, and scale. The last pair determine how to scale the frame: by exposure time or by the mode. Ctrl D to exit.]

cc>combine

To subtract two images of equal exposure time, such as the average dark frame from the individual image frames, the steps would go something like

cl>images	[Enter the major package. Read the help file and edit the epar file. The key parameters are the operand 1 (the first file), op (+, −, *, /), operand 2 (the second file), result (the output file).]
im>imarith file1–file2 outputfile	[You can also type the filenames directly on the command line.]

Be sure to think about a logical naming scheme before you start! The output file should be distinguishable from the raw data frame, and it should be fairly obvious that it has been dark-subtracted (e.g. f012d, for the dark-subtracted version of file number 12). The new data array, "outputfile" (or f012d in this case), can then be displayed. Do the same to a set of flat-field images to subtract the dark from each and then use "combine" again to put all the dark-subtracted flats together. This time you will probably choose the "median" option within combine and use the "scale" option to scale by the mode instead of exposure time. Eventually you will have a final flat-field frame, and you can use "imarith" again with "/" as the "op" and the flat-field frame as "operand 2" to divide each dark-subtracted data frame by the flat-field. A lot of the above procedures can be "scripted" using the IRAF command language to alleviate you of doing repetitive tasks.

IRAF also allows users to create their own tasks by writing scripts. An IRAF script is basically a series of tasks grouped together under a single umbrella name. A user-defined task will have a "param" file like other tasks, and it may have both required and optional parameters. Scripts are written to files with a .cl suffix. They are defined at the command line with

cl>task mytask=filename.cl

where mytask is the name of the task defined in the script, and filename.cl is the file in which it is located. The scripting language in IRAF is very simple. A script consists of a task declaration, followed by a definition of the parameters. The body of the script is started with the word "begin" and ends with the word "end". In between, IRAF commands are entered, one on each line. Variables may also be employed, but must be declared by type at the beginning of the body of the script. The usual programming statements are allowed as well (such as "while" loops, "if–then" statements). Comments start with the symbol "#".

Example. As a simple example, consider a script that will add two images together to make a third image, which will then be divided by a constant if that option is chosen. Further, assume the constant will most often be 1,000 and will not usually be changed. The script to do this is provided below courtesy of Harry Teplitz (formerly UCLA, currently at the Spitzer Science Center). Explanatory comments begin with #:

```
# IRAF SCRIPT addimages.cl
# script to add two images and divide by a constant
# first declare the task name and the required parameters
  procedure addimages (image1, image2, image3)

# next list all the parameters, and give the prompt for them that will appear in the
# param file

  string image1          {prompt="first image"}
  string image2          {prompt="second image"}
  string image3          {prompt="output image"}
  real number=1000       {prompt="number to divide by}
  bool divbool           {prompt="divide by number?"}
# notice that the optional parameter is given a default value
# start the main program
  begin

# add image1,image2 to make image3
# notice that the required parameters are entered in parenthesis and separated by
# commas
  imarith (image1, "+", image2, image3)

# check to see if division should be done
# notice that the commands inside {} are only execute if divbool is "yes"
  if (divbool==yes) {
     imarith (image3, "/", number, image3)
        }
# end program
  end
```

This script could be executed at the command line as follows:

```
first define the task
   cl>task addimages=addimages.cl
then give command
   cl>addimages image1 image2 image3
```

For simple aperture photometry of isolated stars you can use the "qphot" task in the "apphot" sub-package of the "noao" package. The parameter file contains options to allow you to vary the size of the annulus around the star from which

the sky background will be deduced; "annulus" is the inner radius in pixels and "dannulus" is the width of the annulus. For standard stars, specify a list of several "apertures" (e.g., 2, 3, 4, 5, 6 pixels) for the star image up to the value of "annulus" to provide a "curve of growth" from which you can later make aperture corrections for other stars if required. You can execute qphot and move about the SAOimage window placing the cursor on star after star. Hitting the f key produces a tabulation of centroid positions, and magnitudes appear in the xterm window, while pressing the space bar sends the result to a printable file, with the same root name as the image but with the extension *.mag.1*.

For more sophisticated photometry and crowded-field photometry you will need to go to packages which invoke point spread function (PSF) fitting such as DAOPHOT which can be run as part of IRAF. DAOPHOT is a very large program developed over many years by Peter Stetson of the Dominion Astrophysical Observatory in Canada. Most PSF-fitting programs start with a fit to a mathematical function and then modify the PSF with corrections based on the mean differences between the observed profiles and the functional form. In very crowded fields, an iterative process is used in which the best estimate of the PSF is scaled to subtract out the effects of nearest neighbor stars on the PSF. A new PSF is derived from the cleaned-up frame and the process is repeated.

Another useful package for astronomers is SExtractor (for Source Extractor) (Bertin and Arnouts, 1996) which is designed mainly for classifying galaxies from stars in deep images and providing initial photometric and structural (shape) parameters for large-field surveys.

10.2.2 The IDL package

IDL handles images by loading them into its memory. While in memory, images are stored in two parts. First is a string that contains the FITS header and then follows a matrix of real numbers which are the pixel values. The Space Telescope Science Institute has provided IDL scripts for reading from FITS files, manipulating the header information, and writing to FITS files. Simple display routines are available as well (tv, tvscl). For example, to read in a display a FITS file called "myfile.fits", one might enter the following at the command line:

```
IDL>image = readfits("myfile.fits", header)
IDL>tvscl, image
```

IDL commands are either procedures or functions. A procedure is a command that executes a set of operations without creating a new variable. For example, "tvscl" is a procedure that will plot a matrix of numbers with auto-scaling. Procedure commands have two kinds of parameters: arguments and keywords. Arguments follow the procedure name and are separated by commas. They are identified by IDL according to their order on the command line. Arguments may be required or optional. Keyword parameters are also separated by commas, but they are identified by their names. For example, to write the data contained in the matrix "image" to the

file "outfile.fits" with the header information in "hdr_str" (including the header is an optional argument), with undefined pixels set to zero (NaN = Not a Number is the keyword for undefined pixels):

```
IDL>writefits, "outfile.fits", image, hdr_str, NaNvalue=0
```

Functions are very similar except that they create new variables, as in the readfits example given above. Functions also have the form

```
IDL>new_variable=function(parameters)
```

IDL allows users to create their own procedures and functions using the IDL programming language. This language is a full programming language like FORTRAN or C++, even though it is a higher level interface for routines that are actually written in C. All IDL commands are available in programs, as are a variety of common programming statements. The language is constructed to make guessing the right syntax very easy. For example, the command to print to the screen is "print". IDL is not optimized specifically for astronomical use, but a large number of useful programs are already available to perform complicated functions like reading and writing FITS data.

Programs are usually written in files with ".pro" extensions. They can be loaded into IDL's memory with

```
IDL>.run myprogram
```

Once this is done, the program may be executed just as any other function or procedure. Variables may be declared explicitly in IDL (a good programming practice), but the language is very forgiving. Comments begin with ";". Large blocks of programs, such as the body of a loop, are started with "begin" and finished with "end". The main body of the program begins with the function or procedure declaration and ends with "end".

Example. As a simple example, consider a script that will take a list of FITS files with a common "rootname" and subtract a bias image from each, and optionally divide by a flat-field. The output will be written to a new series of FITS files:

```
; IDL program "subflat.pro"
; This program will subtract the bias from a series of images that have a common root to
; their name followed by a number. It will then optionally divide by a flat-field image.
; Output will be written to fits files with a different root name

; first define the procedure
; the "$" allows the definition to continue on the next line
; capitalization and indentation are stylistic, not required by the IDL language
 PRO SUBFLAT, root, out, first, last, $
```

```
  BIASNAME = biasname, FLATNAME = flatname, DIVFLAT = divflat
; first we check to see if bias and flat are defined. If not, they will have the default names
; of "bias" and "flat"
  IF NOT(KEYWORD_SET(BIASNAME)) THEN biasname="bias"
  IF NOT(KEYWORD_SET(FLATNAME)) THEN flatname="flat"
; now define the loop variable to start with first
  I = first
; now read in the bias data
  biasdata=READFITS(biasname+".fts")
; if flat-fielding will be necessary read in the flat-field notice that the conditional lines are
; inside begin/end
  IF KEYWORD_SET(DIVFLAT) THEN BEGIN
     flatdata=READFITS(flatname+".fts")
     END
; next loop through the images. Note that the body of the loop begins with "begin" and
; ends with "end"
  WHILE (I le last) DO BEGIN
     infits=root+STRTRIM(i,1)+".fts" ; make name for of input fits file
     data=READFITS(infits, hdr)      ; read that file into data and hdr
     outdata=data biasdata           ; subtract off the bias from data
  IF KEYWORD_SET(DIVFLAT) THEN BEGIN
     outdata=outdata / flatdata      ; divide by flat if requested
     END
  outfits=out+STRTRIM(i,1)+".fts"    ; construct output filename
  WRITEFITS, outfits, outdata, hdr   ; write data, w/ same header
  I=I+1                              ; increment the loop variable
  END
; end the procedure
END
```

This program could then be loaded at the command line with

```
IDL>.run subflat                    (the .pro is assumed).
```

Then it could be executed on im1.fts, im2.fts, im3.fts, with bias.fts and flat.fts, to make out1.fts, out2.fts, and out3.fts, with

```
IDL>subflat, "im", "out", 1, 3, bi="bias", fl="flat", /div
```

Note that only enough of the keyword names must be given to unambiguously identify each one. Also note that a keyword may be set using /keyword_name instead

of explicitly giving it a value. This is very useful for yes/no types of keywords. The IDL Astronomy User's Library, at the Goddard Space Flight Center, contains over 500 astronomically oriented routines (see *http://idlastro.gsfc.nasa.gov/homepage.html* for more information). There is also a handy IDL-based interactive display tool called ATV (ATV.pro) written and maintained by Aaron Barth at University of California, Irvine (see *http://www.physics.uci.edu/~barth/atv/*)

10.2.3 FITS Liberator

Thanks to a joint ESA/ESO/NASA effort, anyone with a desktop computer running Adobe® Photoshop® software can now download astronomical images in FITS format and successfully import them into this program. FITS Liberator is free software that allows Photoshop to accept FITS files. Photoshop can then be used to create colorized "layers" in order to combine images at different wavelengths or different spectral bands to produce a whole range of sophisticated multi-color images. The web site is given in the references. Apart from the interactive advantage, FITS Liberator enables the creation of images with improved quality because the many grayscale levels in the original observations are preserved. The current version already provides advanced image display options. In addition, the web site provides step-by-step guidance on how to make Hubble-like images.

10.3 PRINCIPLES OF IMAGE ANALYSIS AND PROCESSING

10.3.1 Displaying images

Clearly this is the first step. We will always want to "see" the image, either in a raw form or in various stages of reduction to final form. Software is required to "map" the image onto the computer screen as shown in Figure 10.2; the display on the right is using SAOimage from IRAF.

Initially, this display is a linear mapping of true intensity to values in the display range, which might be 0–255, for instance, to give 256 "levels". The conversion is stored in a LUT (or look-up table). If the weakest intensity is set to correspond to 0 and the brightest signal assigned 255, then all intermediate signals are binned into the intermediate levels as shown in Figure 10.3a. It is generally advantageous if the display software has a cursor which can be moved over the displayed image to "read back" to the screen the (x, y) pixel coordinates and the true intensity at that point in the image—not the scaled value from 0 to 255. If the dynamic range in the image is large, $I_{\max} \gg I_{\min}$, then the resulting linear mapping does not have good contrast. One way to bring up the faint end is to significantly reduce the intensity level assigned to 255 (white). For instance, if we set the white level to 10% of the peak signal, then the remaining range of signals is mapped into 255 resulting in a display in which all the brighter objects are white, but all the fainter signals are now visible. The display is said to have been "stretched" and it is clear that the "transfer function" is steeper. In fact, the 0 (black level) can also be moved and need not correspond to the weakest

Figure 10.2. A computer screen and a display package (SAOimage) to allow digital images to be displayed and enhanced.

signal. Thus, any "window" of signal levels can be stretched from 0 to 255 display levels (see Figure 10.3b).

A variation on this linear stretching approach is to add a point in between the black and white levels and use a different steepness of transfer function in each part. For instance, a steep transfer function could be applied to display data with signal levels of 0%–10% of the peak value and these could be mapped into 0–200 levels, with the remaining signal levels (10%–100% of peak) being mapped to the 201–255 levels as in Figure 10.3c. Such a plot gives a good stretch to the faint end without grossly "overexposing" the bright end. Because the transfer function is now non-linear, although composed of two straight-line segments, we might as well consider *any* non-linear mapping of signal-to-display levels. Enhancing the contrast of faint objects near the sky brightness level can be done with a non-linear transformation such as a "logarithmic curve" (Figure 10.3d) which rises steeply at first to increase the contrast of faint objects, but levels off more slowly to compress the bright end of the map. Repetitive, linear (sawtooth) ramps can also be used to "wraparound" all the gray levels several times (Figure 10.3f). Finally, one of the most powerful non-linear distortions is a transformation which "equalizes" the histogram of signal values vs. the number of pixels with that signal. This condition is illustrated in Figure 10.4. Histogram equalization is very good at bringing out faint objects near the background level.

To see the effect of changing the stretch consider the left-hand image in Figure 10.5 which shows the result of stretching the input levels linearly from 0 to 255. Compare this with the result on the right where "white" is set at 25 instead of 255. This figure was made using FITS Liberator. Note that a histogram of the distribution of signal values is provided.

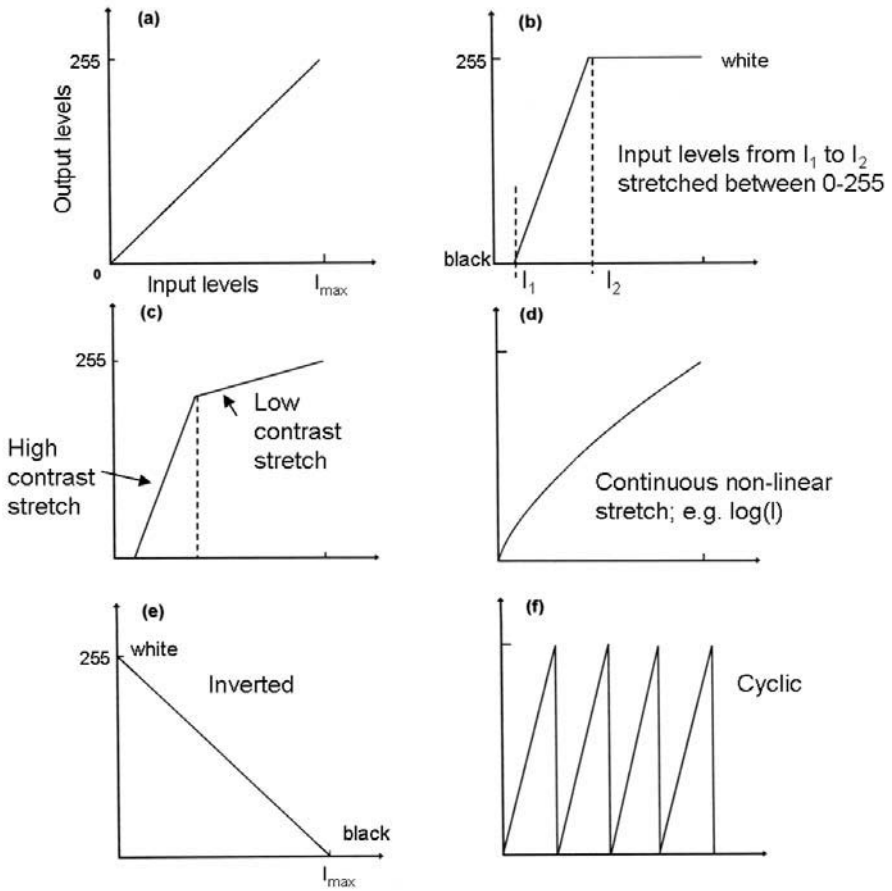


Figure 10.3. Examples of six look-up tables (LUTs) or display-stretching transformations: (a) linear; (b) linear between two intensities; (c) two-step linear; (d) logarithmic; (e) inverse; and (f) sawtooth or wraparound.

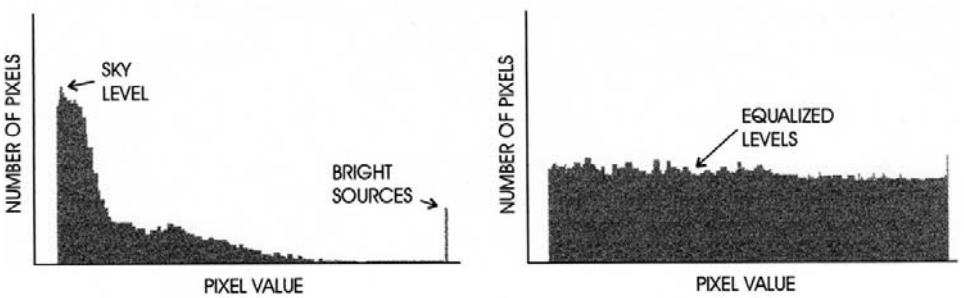


Figure 10.4. (Left): histogram of the distribution of signal values in the image. In a sparse field most of the pixels record the sky value. (Right): a display transformation that results in equalization of the histogram bringing up faint objects.

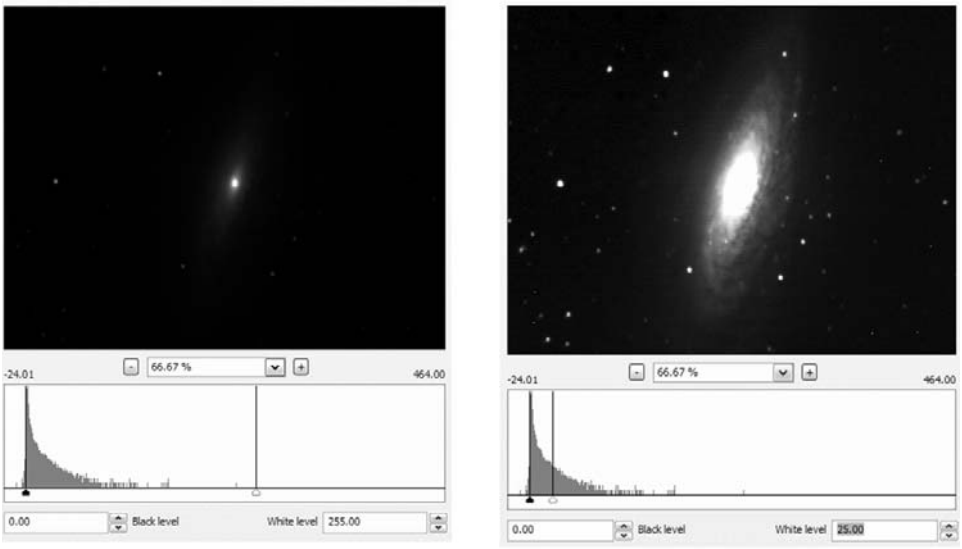


Figure 10.5. Two different linear stretches on the same image using FITS Liberator.

In the above discussion we have assumed that the 256 display levels represent shades of "gray", but it is also possible to match each interval to a particular shade of color. When the distribution of brightness in an image is represented by arbitrary colors it is known as a "false-color" representation. False-color display is really a simple form of image enhancement. Compare Figure 10.6 in shades of gray with its false-color version in Plate 12.

Even in visible light, astronomical images are usually obtained with a combination of optical filters and detector sensitivities that don't remotely match that of the human eye. Therefore, the term "true color" must be used very cautiously. Many spectacular color images are available to the general public, and they are often represented implicitly as natural color, but if you could really see these objects with your eyes they would simply not appear as colorful as shown. Rudy Schild at the Smithsonian Astrophysical Observatory is among the many people who have strived to get these issues right, and I follow his terminology here. In most circumstances three images are used to balance the colors. By adjusting the contrast of the three originals until they are equal and then changing their individual brightness levels, it is usually possible to balance the image so that the color of average stars like the Sun is white. Distinctions that result from using different filter combinations are as follows:

- spectrally augmented color (e.g., replacing the "red" filter with an "infrared" filter which makes red stars look redder and blue stars look even bluer);
- substitute-filter color (e.g., the use of "narrow-band" filters to isolate specific features, such as the methane absorption band in Jupiter to enhance cloud formations);

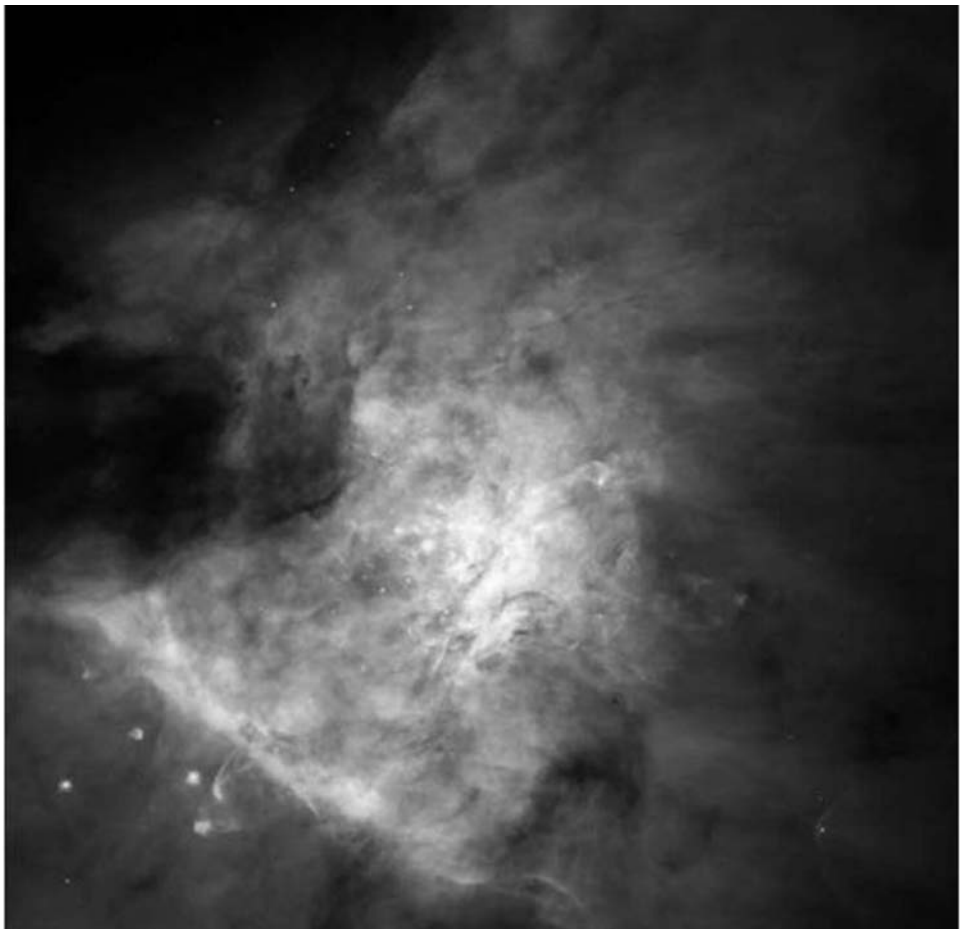


Figure 10.6. The Orion Nebula in certain narrow-band filters rendered here in gray but in false color in Plate 12 demonstrates the effectiveness of adding color for emission highlighting.

- emission highlighting (e.g., the use of narrow-band filters to enhance nebular emission light, such as pink H-alpha and green oxygen, against starlight);
- color translates (e.g., images made at wavelengths completely invisible to the human eye, such as infrared, radio, X-ray);
- enhanced color (e.g., increasing the color saturation to improve color contrast, such as Voyager images of Jupiter)
- two-filter color (e.g., when images are only available at two wavelengths), then the third must be interpolated to produce an image which looks natural but really has no information at the third wavelength.

There are several examples of most of these cases in the Color Plates. For example, the GALEX ultra-violet image of the tail of gas behind the star Mira (Plate 5) is

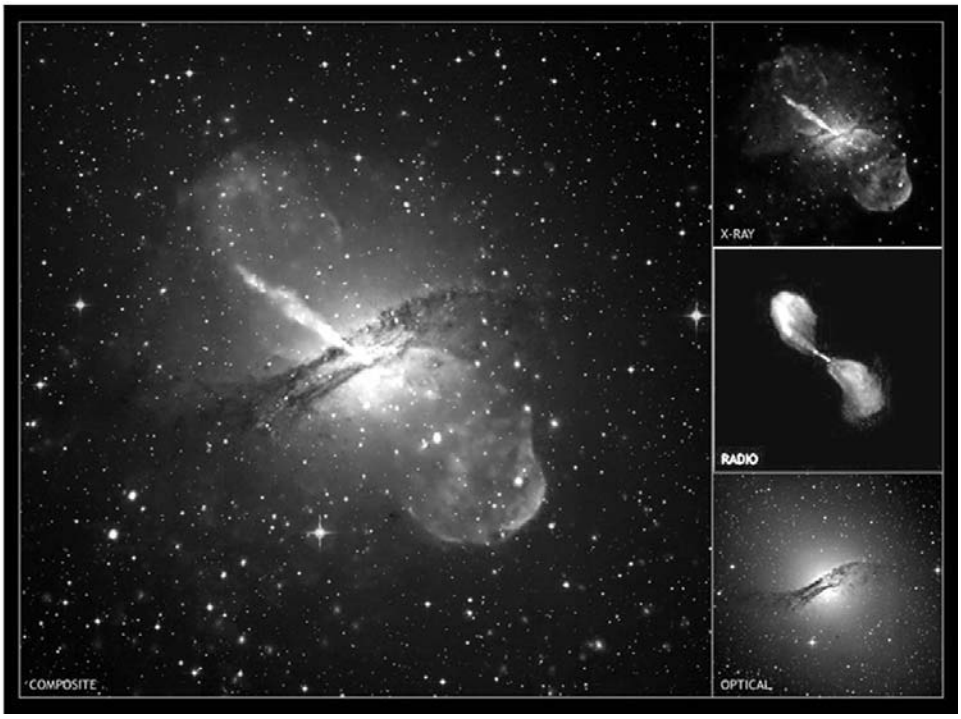


Figure 10.7. The active galaxy Centaurus A in X-ray, radio, and visible light combined into a composite false-color composite is shown here in grayscale. The color version is shown in Plate 4 and the book cover. Credit: see color section.

optically invisible, but is represented as bluish. Near-infrared images are color-translated by making the J, H, and K wavelengths correspond to blue, green, and red colors (Plate 13). Plates 6 and 7 represent the sky as it appears in the mid- and far-infrared by employing shdes of red. As shown by the image of Cen A on the book cover, remarkable combinations are possible when ordinary colors are used to depict X-rays, radio waves, and visible light in the same image; Cen A is shown in Figure 10.7 and in color in Plate 4.

10.3.2 Image enhancement

Images which have had vector or scalar arithmetic processing can be displayed as normal images or they can be displayed as

- pseudo three-dimensional images
- two-dimensional contours or isophotes
- one-dimensional cuts or cross-sections.

In addition, the reduced image can also be further treated with one of several processes:

- low-pass filter removes the finest spatial details
- high-pass filter enhances fine spatial details
- block smoothing replaces each pixel with the average of those in an $n \times n$ block
- Gaussian smoothing replaces each pixel with a weighted average depending on the width of a bell-shaped "Gaussian" profile of width σ pixels
- unsharp masking, enhances fine details by subtracting a low-pass filtered (fuzzy) image from the original.

Other processes which the computer must be able to supply include

- Integer and non-integer pixel shifts in which the numbers in the original data array are mapped to new locations in a new data array and, if the shift is not an integral number of pixels, the new intensity value is "interpolated".
- Bad-pixel corrections by interpolation across the incorrect pixel value in both x and y to estimate a replacement value.
- Deconvolution and image restoration. The observed image is a convolution of the initial image and a point spread function which is the instrumental response to a perfect point source. Deconvolution is achieved using a Fourier transform.

A "low-pass filter" is a "smoothing" process—rather than a stretching process—which reduces the contribution of high spatial frequencies, or fine detail, in an image while maintaining the lower, more slowly changing spatial patterns. Simply replacing the value associated with each pixel in an image with the average of its immediate neighbors constitutes an elementary low-pass filter. Noise often contains high spatial frequencies (i.e., relatively large and random changes from pixel to pixel), therefore a low-pass filter can reduce noise. The appearance of a low-pass filtered image is "softer" or more blurry than the original, but it is useful for bringing up faint nebulosity in noisy images.

In a "high-pass filter" the process is one of "sharpening" which emphasizes fine details (high spatial frequencies) by suppressing the lower spatial frequencies or more slowly changing spatial patterns. The image looks "sharper" but the noise pattern will also get enhanced unless the original image is just fuzzy rather than noisy.

Both of the filters just discussed are simple convolutions. A convolution is described by the following process:

$$G = F * H$$

where

$$G(x, y) = \iint_{-\infty}^{+\infty} F(x - x', y - y')H(x', y')\, dx'\, dy' \tag{10.1}$$

The way to visualize this action is to think of x, x', y, y' as pixel coordinates in an image $F(x, y)$, and $H(x', y')$ as a weighting function or "kernel". At the point $(x - x', y - y')$ the image has a certain value (the brightness at that point) and we

multiply that brightness by the value of the weighting function at the point (x', y') and integrate (or in practice *sum*) over all the values of x' and y'. Usually, the weighting function is restricted in its range and is zero beyond a certain range of pixels. The next pixel in the image is selected and the whole process is repeated until the entire image has been covered. A simple way to implement these filters is to establish a small matrix with an odd number of elements (e.g., 3×3, 5×5, 7×7) and "scan" this matrix (or kernel) over the image. The new intensity corresponding to the current location of the center of the matrix is found by summing all the products of each weighting value in the matrix with the underlying intensity at that point in the image; if parts of the matrix overhang the edge of the image then those elements contribute zero to the sum. Two simple examples of low-pass and high-pass filters are shown below using only 3×3 matrices:

low pass

1 1 1	0 1 0
1 1 1	1 10 1
1 1 1	0 1 0
straight average of nearest neighbors, strong blurring	more weight given to central pixel, modest blurring

high pass

0 −1 0	−1 −1 −1
−1 16 −1	−1 8 −1
0 −1 0	−1 −1 −1
more sharpening	less sharpening

In the low-pass filter, brightness values surrounding the central pixel are "added in" which smoothes or blurs fine detail, whereas in the high-pass filter the adjacent signals are "subtracted out" to give emphasis to the central pixel.

Specific two-dimensional functions can be used to define H, such as a radially symmetric Gaussian profile of the form $\exp[-r^2/2\sigma^2]$, where σ is the radius at which the function has only 61% of its initial value, and hence at 3σ the value is down to 1%.

Median filtering is a powerful—yet simple—example of non-linear filtering which provides a way of eliminating single-pixel "spikes" with highly deviant brightness (due to cosmic rays, for instance). In terms of the matrix approach mentioned above, each element is a 1, but the replacement value is not the sum of the products. Instead, the products are sorted into a "list" with increasing intensity ($I_{\min}$ to $I_{\max}$) and the intensity which lies in the "middle" of this odd-numbered sequence (i.e. the median intensity) is selected to become the replacement value. Note that even when the filter is centered on the pixel with the spurious value, that value will be at one end of the ordered list and will never be selected as the replacement value. To see this, just

consider a random group of numbers with a mean value around 10 and a range, say, from 7 to 14. Now add one number with a value of 100. Sort the list from minimum to maximum and the high value will end up at one end. This method works on pixels that are too bright or pixels that are too dark. It is possible to design the algorithm so that it will only act on single-pixel events and/or exceptionally deviant values and leave well-sampled star images alone, otherwise the median will remove the stars too, which is in fact what you want when forming a median sky flat.

Median filtering is deliberately used to remove stars and faint compact sources in deep-sky imaging so that the sky itself can be used as a flat-field. The ordered list of intensities is obtained by displacing or "dithering" the telescope numerous times in a random pattern between individual background-limited exposures so the faint point-like sources never fall on the same pixel twice.

There are several enhancement filters which are based on the "gradient" or rate of change of brightness in the image. For example, taking the difference between an image and a copy of itself in which the data values have been "shifted" over by a small number of pixels produces an estimate of the slope and enhances edges where the gradient is large. A more sophisticated form of this filter produces "side-illuminated" or "bas-relief" images which have a 3-D appearance (Figure 10.8). Other filters in this category are the Laplacian and the Sobel filter.

One of the most well-known filtering methods is the "unsharp mask". This technique has been used very successfully by astro-photographer David Malin to enhance prime-focus images from the Anglo-Australian Telescope, and can certainly be used with electronic images from CCD cameras and infrared arrays. Basically, the image is first smoothed with a low-pass filter to create the blurry, "unsharp" mask, and then this frame is subtracted from the original image. Features with low spatial

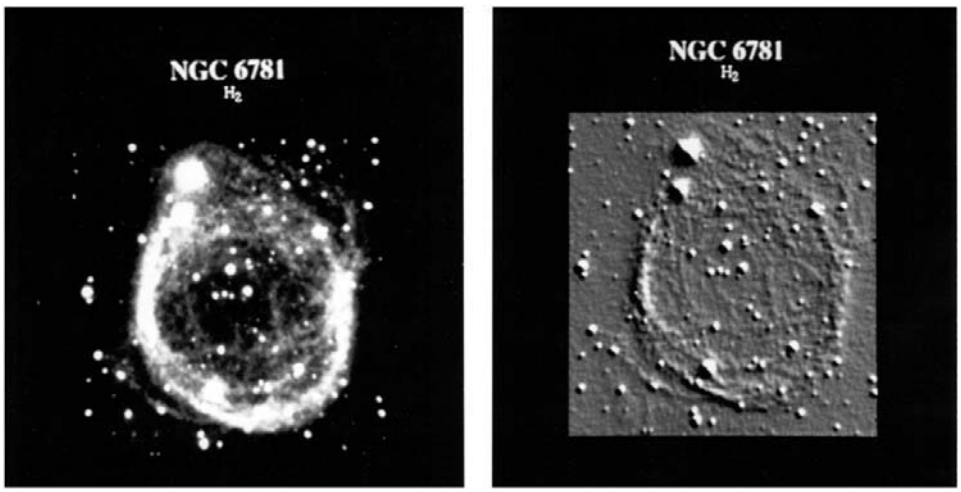


Figure 10.8. A display that has been (left) stretched with a logarithmic scale, and (right) treated to enhance sharp features using bas-relief.

frequencies are removed leaving an image dominated by high spatial frequencies. To control the harshness of the effect, the original image and the unsharp mask can be scaled differently to give more weight to the original image.

10.4 IMAGE RESTORATION

Image restoration is the recovery of images from raw, "image-like" data, whereas image reconstruction implies the production of an image derived from data with a more complex form of encoding so that an image is not readily discernible. Image restoration is the common form of image analysis used in UV–visible–IR astronomy because these disciplines have direct imaging detectors such as CCDs, infrared arrays, and microchannel plates. Radio astronomy, on the other hand, is often faced with reconstruction of images from interferometric (Fourier) data. Although this field had been around for a long time, it received an immense boost following the discovery of the spherical aberration in the primary mirror of the Hubble Space Telescope which resulted in images with a large halo of light around a sharp core. Luckily, the behavior or point spread function of the image was measurable and various techniques could be applied to retrieve the true image until the problem was corrected by the next service mission. Among these algorithms and methods were

- *Richardson–Lucy* deconvolution
- *maximum entropy* deconvolution
- *pixon*-based image reconstruction.

Some techniques such as the Richardson–Lucy method are distributed with popular image-processing packages such as the Space Telescope Science Data Analysis System (STSDAS) which runs under IRAF. This package also includes Maximum Entropy and Clean algorithms. Commercial packages such as MemSys5 from Maximum Entropy Data Consultants Ltd., U.K., are also widely used. Nick Weir of Caltech has developed the MEM front-end program for MemSys5, and this combination is also available through the Space Telescope Science Institute. Pixon image reconstruction software is available by arrangement from Dr. Richard C. Puetter at University of California, San Diego.

The mathematical basis for most of these methods is a theorem known as Bayes Theorem, and the methods are said to be Bayesian. Image reconstruction in its most general form is an "inverse" problem in which the data, $D(x)$, is related to the true signal, $I(x)$ through a relationship like

$$D(x) = \int dV_y H(x, y) I(y) + N(x) \tag{10.2}$$

where x and y can be n-dimensional vectors; and the quantity $H(x, y)$ is called a "kernel" function which describes how the act of measurement distorts or corrupts the true signal; and $N(x)$ is the "noise" or error associated with the measurement either due to instrumental effects or signal strength (counting statistics) or a combina-

tion. $I(x)$ is also called the "hypothesis" of the data or simply the "image". Usually, $H(x, y)$ is the simple "blurring" function that we have been calling the point spread function or PSF, and the equation can be reduced to a convolution integral:

$$D(x) = \int dV_y H(x - y) I(y) + N(x) \tag{10.3}$$

In the more general case of image reconstruction—rather than image restoration—then $H(x, y)$ is more complex.

Linear inversion methods, such as Fourier deconvolution, offer a simple closed-form solution for $I(x)$,

$$I(x) = F^{-1}\left[\frac{F(D)}{F(H)}\right] \tag{10.4}$$

where $F(f(x))$ is the Fourier transform of the function $f(x)$ and $F^{-1}(f(x))$ is the inverse Fourier transform of $f(x)$. Although computationally expedient, linear inversion methods tend to propagate noise and generate unwanted artifacts in the solution. Non-linear methods are of course more complex, but most can be interpreted in terms of a Bayesian Estimation scheme in which the hypothesis sought is judged to be the most probable. Bayesians use conditional probabilities to factor the joint probability distribution $p(D, I, M)$, where D, I, and M are the data, the un-blurred image, and the "model" of the image. The mathematical expression of the model reflects knowledge of the physics of the imaging process, the instrumental properties, and the method of modeling the data (e.g., as a discrete sum). Two Bayesian formulations are expressed below:

$$p(I \mid D, M) = \frac{p(D \mid I, M) p(I \mid M)}{p(D \mid M)} \tag{10.5}$$

$$p(I, M \mid D) = \frac{p(D \mid I, M) p(I \mid M) p(M)}{p(D)} \tag{10.6}$$

where $p(X|Y)$ is the probability of X given that Y is known. In the first equation, the model is fixed, and this is the basis of maximum likelihood and maximum entropy methods, whereas in the second formulation the un-blurred image (I) and the model (M) are taken together and the goal is to maximize the probability of the image/model pair, given the data. This is the basis for the more recent pixon-based algorithms. Each term on the right-hand side of these equations has a physical significance. The term $p(D|I, M)$ measures the likelihood of obtaining the data (D) given a particular un-blurred image (I) and model (M) (i.e., it is a "goodness-of-fit" or GOF parameter). The term $p(I|M)$ is particularly interesting since it makes no reference to the data, and therefore this term can be assigned a value *prior* to making any measurements. Consequently, $p(I|M)$ is known as the image "prior". In maximum likelihood methods, the image prior is assumed to be a constant and the goodness-of-fit term would be the standard chi-square distribution

$p(D|I, M) = \exp(-\chi^2/2)$. Typically, images restored by these methods show residual errors associated with the brightest sources and often produce spurious sources sprinkled across the field. In maximum entropy (ME) methods, the image prior is given by $p(I|M) = \exp(\alpha S)$ where S is called the "entropy" and is a measure of the degree to which the uncertainty in the information has been eliminated ($S = -\sum p_i \log p_i$ for $i = 1 \cdots n$, where p_i is the probability that the random variable X takes the value x_i, and α is the "weighting" factor between the GOF and the image prior). Maximizing the entropy minimizes the information uncertainty. Since many of the spurious sources caused by pure goodness-of-fit methods are the result of "overfitting" of the data (i.e., using too many free parameters), maximum entropy overcomes this by asserting that the best image prior is one in which the image is completely flat! This is both the strength and weakness of the MEM method. It is a strong point because it tends to reduce the number of free parameters and the tendency to overfit the data, but it is a weakness because a "flat" image is not the most likely image prior. A real (astronomical) image will have regions which are indeed very flat (e.g., the sky background), and regions which vary slowly in brightness (e.g., reflection nebulosity, elliptical galaxy starlight), and regions with significant structure on the finest spatial scales (e.g., stars). In regions of the image devoid of any changing structure we could use much larger cell sizes (coarser resolution) to record our data because we are not losing any information, whereas we will need small cells to sample the finest structure in the image. Maximizing the image prior $p(I|M)$ could therefore be achieved by decreasing the total number of variable-sized cells used and packing as much signal as possible into each of these cells while still maintaining an adequate goodness of fit. This is the basis of the remarkably powerful new image reconstruction method developed by Rick Puetter and Robert Peña at UC San Diego known as "pixons", the term they coined to describe variable-sized cells. Pixon sizes follow the information density in the image (small when needed, large when not) and in fact each pixon represents a degree of freedom. Practical implications of the method require some choice of pixon shape, usually a radial symmetric cell with a "fuzzy" boundary, but the best shape should be related to the image properties. Pixons naturally eliminate signal-correlated residual errors and spurious sources because the technique eliminates all of the great many unconstrained degrees of freedom inherent in other approaches. The degrees of freedom represented by the pixons are in fact, the only ones constrained by the data. Reviews of the pixon technique with numerous examples are given in Puetter (1995, 1996). Figure 10.9 shows an example of the pixon method, and Figure 10.10 is the result of the maximum entropy method.

The Richardson–Lucy iterative inversion scheme was first devised in 1972 by William Hadley Richardson and independently discovered for astronomical applications by Leon Lucy in 1974. It has been used extensively for Hubble Space Telescope image restoration and is available within the STSDAS/IRAF environment. Typical RL restorations require less computer time than maximum entropy methods, but more than simpler filtering. The Richardson–Lucy algorithm is a non-linear iterative algorithm based on a linear data-gathering model which estimates the prior probability of the image value. Among the most important characteristics of the algorithm

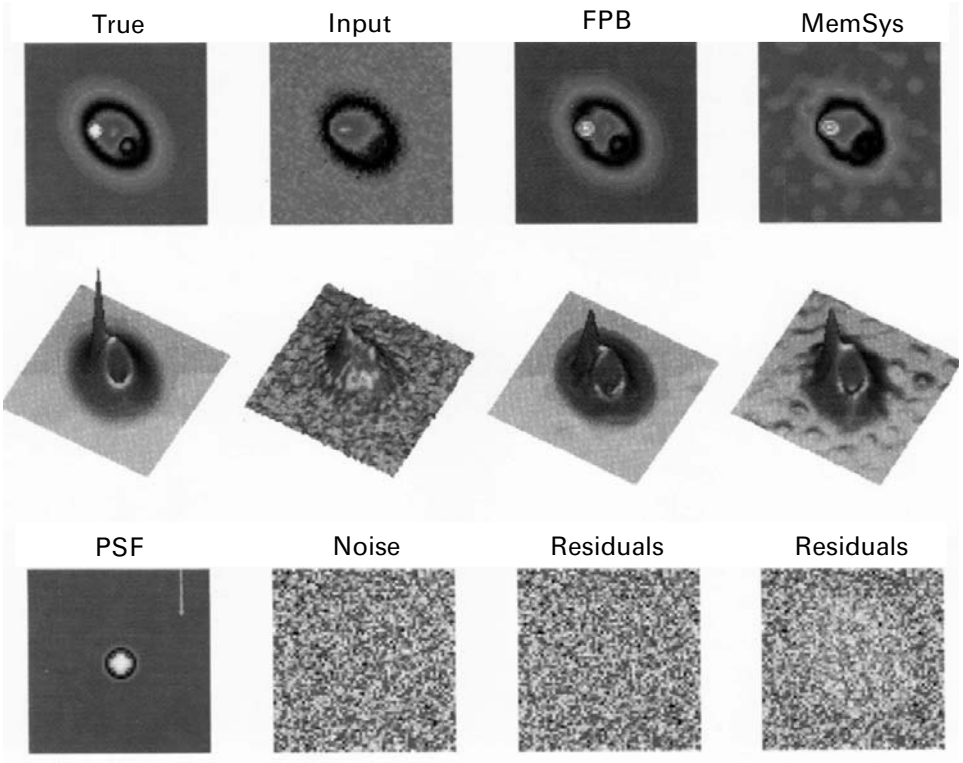


Figure 10.9. Image restoration using the pixon method developed by Rick Puetter and Robert Peña at UCSD, and a comparison with other methods. Credit: Rick Puetter.

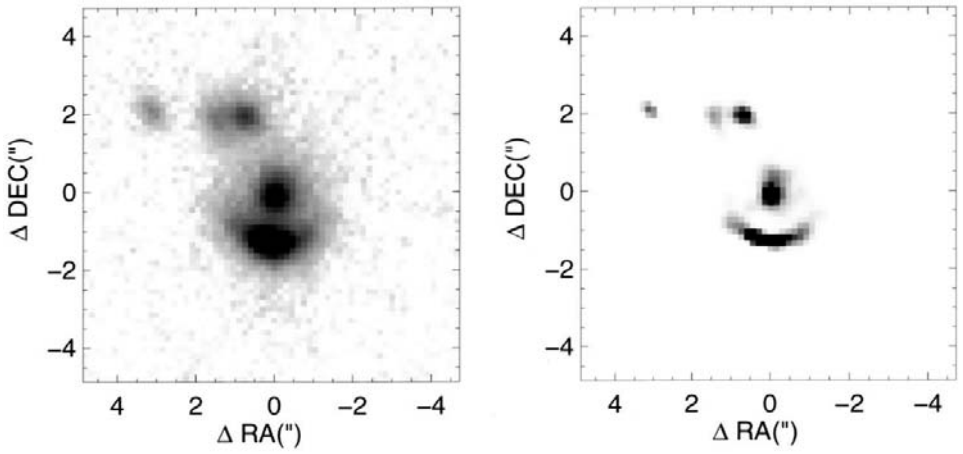


Figure 10.10. An infrared image (left) from the Keck telescope under excellent seeing conditions showing a peculiar arc near the galaxy FSC 10214+47241. Image restoration using MEM reveals that the arc is part of an ''Einstein ring'' due to gravitational lensing of light from a more distant object by the galaxy in the foreground. Credit: James Graham.

are the following:

- Image flux is preserved giving the restored image good photometric linearity, even if the iteration is stopped before convergence.
- RL iteration converges to the maximum likelihood solution for Poisson statistics in the data (Shepp and Vardi, 1982) and is therefore suited to astronomical detectors with photon-counting detectors. Readout noise from CCDs is a complication.
- Non-negativity is preserved in the restored image provided that the PSF and the original image are both non-negative. There is no tendency to "dig" into a zero background.
- Small errors in the PSF are not significant.

Despite these advantages, the RL method suffers from the same generic problem of all maximum likelihood techniques which attempt to fit the data as closely as possible (noise amplification). After many iterations (>50) on an object with an extended light distribution such as a galaxy or nebula, the restored image develops a "speckled" appearance as a result of fitting the noise in the data too closely. Another problem is "ringing", the appearance of dark and light rings at sharp boundaries such as a bright star. Ringing is due to the blurring by the PSF and the loss of information due to the lack of higher spatial frequencies.

A useful program for HST users is Tiny Tim, a program written by John Krist (Space Telescope Science Institute) which generates simulated Hubble Space Telescope PSFs. Mirror zonal errors, time-dependent aberrations for the pre-repair cameras, and field-dependent obscuration patterns for WF/PC-1 and WF/PC-2, as well as filter passband effects are all included. The output FITS files can be normally sampled or subsampled. This program has many applications including subtracting unsaturated PSFs to reveal faint objects, convolving models with the PSFs to predict actual observations, and deconvolving the PSF from the observations to produce cleaner images.

10.5 SUMMARY

Astronomers use computers for a wide range of purposes. Modern instrumentation is usually controlled by a combination of fast hardware and some form of digital signal processor at the front end and a powerful image display workstation or fast PC at the back end. Large software packages, such as IRAF, AIPS, STSDAS, STARLINK, and MIDAS, have been developed by the astronomical community for image analysis and data processing. Commercially, packages such as IDL are also widely used. The standard protocol for image files in astronomy is FITS (Flexible Image Transport System) but a program called FITS Liberator is now available from ESA/ESO/NASA to allow Adobe Photoshop to import such images. Digital images can be stretched, filtered, and false-colored for image enhancement. Image restoration techniques based on deconvolution theory and Bayesian probabilities are used to recover

images from known aberrations or in support of diffraction-limited imaging with known point spread functions.

10.6 EXERCISES

1 Calculate the data rate for a $1{,}024 \times 1{,}024$-pixel CCD camera system which digitized each pixel to 16 bits in a time interval of 100 microseconds per pixel. Compare this with an infrared camera system of the same detector format but a pixel rate of only 5 microseconds per pixel. Suppose the IR detector provided 32 simultaneous outputs, what would the data rate be then?

2 Explain the difference between false-color used to enhance faint optical images and false-color used to illustrate the appearance of a nebula at three near-infrared (invisible) wavelengths.

3 Describe the difference between a high-pass and a low-pass filter.

4 Explain how a median filter can remove cosmic-ray hits or other deviant pixel values.

5 How is an "unsharp" mask produced?

6 Explain the difference between image enhancement and image restoration.

7 Compare and contrast IRAF and IDL for analysis of CCD data.

10.7 REFERENCES

Bertin, E.; and Arnouts, S. (1996) SExtractor: Software for source extraction, *Astronomy & Astrophysics Supplement*, **117**, 393–404.

Eichhorn, G. (1994) An overview of the Astrophysics Data System, *Experimental Astronomy*, **5**, 205.

Gull, S.F.; and Skilling, J. (1999) *Quantified Maximum Entropy MemSys5 User's Manual V1.2*, Maximum Entropy Data Consultants, Bury St. Edmunds, U.K.

Lucy, L.B. (1974) An iterative technique for the verification of observed distributions, *Astronomical Journal*, **79**, 745–754.

Puetter, R.C. (1995) Pixon-based multiresolution image reconstruction and the quantification of picture information content, *Journal of Imaging Systems and Technology*, **6**, 314–331.

Richardson, W.H. (1972) Bayesian-based iterative method of image restoration, *J. Opt. Soc. Am.*, **62**, 55–59.

Schild, R.E. (1994) The many hues of astronomical color imaging, *CCD Astronomy*, **1**(2).

Shepp, L.A.; and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography, *IEEE Trans. Medical Imaging*, **MI-1**, 113–122.

Starlink Bulletin, Issue No. 2 (1988) Rutherford Appleton Laboratory, Chilton, U.K.

Stetson, P.B. (1993) DAOPHOT: A computer program for crowded-field stellar photometry, *Pub. Astron. Soc. Pacific*, **105**, 527–537.

Warmels, R.H. (1991) *The ESO-MIDAS System: Astronomical Data Analysis Software and Systems I*, PASP Conf. Series Vol. 25, Astronomical Society of the Pacific, San Francisco, p. 115.

Wells, D.C.; Greisen, E.W.; and Harten, R.H. (1981) FITS: A flexible image transport system, *Astronomy & Astrophysics Supplement*, **363**.

Web sites

ESO-MIDAS: *http://www.eso.org/sci/data-processing/software/esomidas/*
FITS Liberator: *http://www.spacetelescope.org/projects/fits_liberator/index.html*
IRAF Home page: *http://iraf.noao.edu/*
NRAO AIPS: *http://www.aips.nrao.edu/*
Space Telescope SDAS: *http://www.stsci.edu/resources/software_hardware/stsdas*
Tiny Tim: *http://www.stsci.edu/software/tinytim/tinytim.html*
UK STARLINK (closed 2005): *http://www.starlink.rl.ac.uk/*

11

Electronic imaging at infrared wavelengths

Because of the intrinsic bandgap of silicon, CCDs do not respond beyond 1.1 μm. To cover the huge infrared range out to at least 150 μm requires different materials and techniques. In this chapter we describe the remarkable development and impact of infrared "array" detectors and the enormous explosion of infrared facilities. In contrast to UV, X-ray, and gamma-ray astronomy, there are extensive opportunities to make infrared observations from ground-based sites, as well as from the stratosphere and from space.

11.1 INTRODUCTION

So great was the impact of the CCD that it is frequently said to have "revolutionized" astronomy. That same pronouncement would seem like an understatement for the advent of the "infrared array" which occurred about a decade later. Infrared observations are extremely important in astrophysics for many reasons. For example, because of the Hubble expansion of the Universe, the visible light from distant galaxies is stretched, effectively moving the observed spectrum into the infrared for the most distant objects.[1] Equally important, infrared wavelengths are much more penetrating than visible light, and can therefore reveal the processes at work in star-forming regions which are typically enshrouded in clouds of gas and dust. Similarly, infrared observations can allow us to "see" all the way to the center of the Milky Way and reveal the nature of the central mass of the Galaxy. Cold interstellar material emits no visible light but it does emit in the far-infrared, which provides a

[1] The wavelength shift is measured by the scale factor $(1 + z)$ where z is called the "redshift" (i.e., $\lambda = \lambda_0(1 + z)$). For example, for a galaxy at redshift $z = 2.5$, when the Universe was only $1/(1 + 2.5) = 0.29$ its present size, the familiar red H-alpha line from hot hydrogen gas at 656.3 nm is found at the infrared wavelength of 2,297 nm (≈2.30 μm).

means to study the dust itself. Lastly, energy transitions in molecules that involve quantized rotation and vibration states result in the emission of low-energy infrared photons, and thus infrared spectroscopy is a powerful diagnostic tool to probe the chemistry of the interstellar medium and the coolest stars.

For many years the infrared part of the spectrum was considered to be the region just beyond the red limit of sensitivity of the human eye, at a wavelength of about 720 nm (or 0.72 μm). With the advent of CCDs, "optical" astronomy extended its territorial claims to about 1.1 μm, the cutoff wavelength for detection of light imposed by the fundamental bandgap of silicon (recall that $\lambda_c = 1.24/E_G$; for $E_G = 1.13\,\text{eV}$, $\lambda_c = 1.1\,\mu\text{m}$). So where is the "real" optical–IR boundary for ground-based astronomy? A reasonable response is that it occurs at 2.2 μm–2.4 μm because, as shown in Chapter 2, at these wavelengths there is a marked and fundamental change in the nature of the "background" light entering the telescope/detector system. Consequently, there is a practical change in observing methods and instrument design. For wavelengths shorter than ~2.2 μm the background light comes mainly from OH emission in the Earth's upper atmosphere, whereas at longer wavelengths the dominant source of background radiation is the thermal (heat) emission from the atmosphere and telescope optical components. The domain of infrared astronomy is typically subdivided as follows. Near-infrared (NIR) is now taken to be the interval from about 0.9 μm to 5.5 μm, although the term short-wave infrared (SWIR) is used specifically for 0.9 μm–2.5 μm, and the *thermal* near-infrared refers to the part from 2.5 μm to 5.5 μm. Current NIR detectors already overlap with CCDs for wavelengths less than 1.1 μm and new devices will perform down to ~0.5 μm. Large-format IR arrays are available, and thus NIR merges smoothly with the classical optical regime. Mid-infrared (MIR) extends from ~5 μm to 30 μm and far-infrared (FIR) stretches from ~30 μm to ~200 μm. Observations at these longer wavelengths are more challenging from the ground hence the interest in observations from the stratosphere. Wavelengths longer than about 200 μm (or 0.2 mm) are now referred to as the sub-millimeter, and although sub-millimeter astronomy is closely allied with infrared wavelengths in terms of the objects and regions of space which are studied, some of its techniques are more akin to those of radio astronomy. To appreciate the remarkable transformations in IR astronomy in recent years and the impact of technology, it is worthwhile to consider briefly the historical development.

11.1.1 Early history of infrared astronomy

Infrared astronomy had an early—albeit somewhat accidental—origin when, in 1800 in a series of papers, Sir William Herschel (1738–1822) discoverer of Uranus, noted that a thermometer placed just beyond the reddest end of a spectrum of sunlight not only increased its temperature compared with two other thermometers set off to the side, but also showed a greater heating than any other location within the spectrum. Herschel called these unseen radiations "calorific rays" and proved that they were refracted and reflected just like ordinary light. Herschel's discovery occurred about 65 years before Maxwell showed that light was only one form of electromagnetic

radiation. The prolific Sir William also made another observational discovery which he called "holes in the sky", by which he meant irregularly shaped dark regions where the dense distribution of Milky Way stars, so visible in the eyepiece of his large telescope, seemed simply to vanish. These dark blobs, absent of stars, would be cataloged by pioneering American astrophotographer E.E. Barnard (1857–1923) almost a century later, and recognized as dense clouds of gas and dust. But it would not be until the Swiss-born U.S. astronomer Robert J. Trumpler (1886–1956) proposed in 1930 that the general interstellar medium was filled with interstellar "dust", which affected distance measurements by dimming and reddening blue light much more than red light, that a motivation for infrared studies would slowly emerge. Herschel had no way of knowing that the "calorific rays" he had discovered would one day provide the means to penetrate and explore his "holes in the sky."

11.1.2 The beginning of modern infrared astronomy

Despite some additional development of infrared radiometers, and a variety of infrared observations of stars and solar system bodies up through the 1920s (see Martin Harwit's chapter in *The Century of Space Science*, 2001 and Harwit, 1999 for more details), the major breakthroughs did not come until after World War II. Rapid developments in infrared detector technology were stimulated not by any commercial market but by military requirements. Lead sulfide (also known as galena, PbS) is a semiconductor with a fundamental (direct) bandgap of 0.41 eV at room temperature dropping to 0.286 eV (4.3 μm) at 4.2 K. PbS was used in the classical photoconductor mode at 77 K (Chapter 5) with a fixed voltage across the detector.

In the early 1960s two physicists Gerry Neugebauer and Bob Leighton at the California Institute of Technology (Caltech) began a "two-micron sky survey" (TMSS) with an angular resolution of 4 minutes of arc. The pair constructed their own survey telescope, a 1.57 m (62-inch) $f/1$ parabolic dish, by machining the primary mirror from aluminum metal on a lathe. Then, to improve the surface finish, they poured on a layer of epoxy and spun the paraboloid about a vertical axis until the material set; this technique is similar to the principle used today in the spinning furnace to shape molten glass into a deep parabolic curve. After applying a reflective coating the telescope produced images of about 2 arc minutes, which was sufficient considering the survey had to cover about 30,000 square degrees of the sky above a declination of -30°. Their detector comprised eight separate PbS photoconductors used in a pair-wise fashion to alternate between detectors in order to remove background radiation. In 1965 Neugebauer, Martz, and Leighton announced the discovery of incredibly bright, "1st-magnitude" infrared sources with extremely faint optical counterparts. This was not expected. Then, based on observations made in 1966 by graduate student Eric Becklin, he and Neugebauer (1968) announced the infrared detection of the objects at the center of the Milky Way, not seen in visible light due to 30 magnitudes of extinction. Another Becklin and Neugebauer find lay in the heart of the well-known Orion Nebula where they discovered a very bright yet optically invisible young star (now named the BN object).

Many of the new sources showed a trend of increasing brightness at longer wavelengths where PbS was not sensitive. A crucial step forward to exploring longer wavelengths was the invention of the liquid helium cooled gallium-doped germanium (Ge : Ga) bolometer in 1961 by Frank Low, another physicist who was at that time working for the Texas Instruments Corporation. Frank later moved to the National Radio Astronomy Observatory and then to the University of Arizona in Tucson, where he not only established a formidable infrared program, but also set up a company, called IR Labs, to provide cryogenic detector systems to other researchers. Using a cryogenic instrument at the telescope was a challenging prospect in those days. However, because of its wavelength-independent response, the gallium-doped germanium detector opened up much longer wavelengths to astronomers. Frank built detectors for 10 μm and 21 μm. He also developed a telescope and bolometer system that could be mounted on a NASA Learjet and flown above most of the terrestrial water vapor, thus enabling him to observe at 70 μm. Infrared observations from balloons and rockets also began in the mid-1960s. In 1967 Doug Kleinmann and Frank Low reported observations at 22 μm that led to the discovery of an extended infrared-glowing cloud near the BN object in Orion, now known as the Kleinmann–Low Nebula, and by 1970 it was realized that some distant galaxies emitted far more infrared radiation than all other wavelengths combined (Kleinmann and Low, 1970).

Still driven by military requirements for heat-seeking devices, the lead sulfide cell was replaced by a more sensitive photodiode made from indium antimonide (InSb). Don Hall (now at the University of Hawaii) played a key part in the introduction of InSb to astronomy (Hall *et al.*, 1975) and has continued to push the frontiers of infrared detector developments ever since. Stimulated by the discovery rate, there was a push for telescopes that were "optimized" for infrared work. In 1970 the Mount Lemon Infrared Observatory was established in the Catalina Mountains near Tucson, Arizona, while U.K. astronomers built a 1.5 m infrared "flux collector" on Tenerife in the Canary Islands in 1971. Harry Hyland pioneered the study of southern hemisphere skies at Mount Stromlo using a series of detectors (Hyland, 1971). By 1978 infrared observations with a single-element detector were being made on the Anglo-Australian Telescope by David Allen (1946–1994) and colleagues, and Allen had produced the first book on the "new astronomy". The need for better far-infrared observations led to the development of the Kuiper Airborne Observatory (KAO), a modified C-141A jet transport aircraft with a 91.5 cm (36-inch) Cassegrain telescope, capable of operating at altitudes of up to 14 km (45,000 ft). The KAO made many significant discoveries including the first detection of faint rings around Herschel's planet, Uranus, in 1977. The KAO operated out of the NASA Ames Research Center at Moffett Field, California from 1974 to 1995.

By 1979 a new generation of 3 m–4 m class telescopes dedicated to infrared astronomy had come into operation including the U.K. 3.8 m Infrared Telescope (UKIRT) and the NASA 3 m Infrared Telescope Facility (IRTF) whose first director was Eric Becklin. Both of these telescopes were located on the 4.2 km (14,000 ft) summit of Mauna Kea, Hawaii which was rapidly becoming recognized as an exceptional site. Other telescopes optimized for infrared astronomy soon followed, and some "optical" telescopes began to provide optional configurations for infrared

work, including the 5 m Hale telescope. At wavelengths longer than 2.4 μm, moonlight is almost undetectable, and so infrared became the "bright time" or full Moon option. The period up to the launch of IRAS has been reviewed by Low, Rieke, and Gehrz (2007).

11.1.3 The launch of IRAS

The Anglo-American-Dutch Infrared Astronomical Satellite (IRAS) mission gave astronomers their first deep all-sky survey in the infrared. With the launch of IRAS on January 25, 1983, infrared astronomy took a quantum leap. The IRAS mission mapped the entire sky at wavelengths of 12 μm, 25 μm, 60 μm, and 100 μm, produced a point source catalog of over 245,000 sources (more than 100 times the number known previously), and made numerous unexpected discoveries, including a dust shell around the "standard star" α Lyrae (Vega) and about 75,000 galaxies believed to be in a "starburst" state. IRAS had a lifetime of only 10 months in operation until the onboard supply of 475 liters of superfluid helium coolant, which held the detectors at 1.8 K and everything else at about 10 K, was finally exhausted. The 60 cm telescope and its detectors then warmed up and lost their sensitivity. IRAS was so successful that follow-up missions involving "observatory class" cryogenic satellites were planned by both ESA and NASA. The European project, called ISO (Infrared Space Observatory), was launched successfully in late 1995 and operated until 1998, while the American project, initially called SIRTF (Space Infrared Telescope Facility), was delayed and then finally launched in 2003, at which time it was renamed the Spitzer Space Telescope. In the interim period an infrared instrument (NICMOS) was placed into service on the Hubble Space Telescope.

11.2 INFRARED WAVEBANDS

11.2.1 Atmospheric windows

Water vapor (H_2O) and carbon dioxide (CO_2) block out a lot of infrared radiation from space. Figure 11.1 shows a more detailed transmission spectrum of the Earth's atmosphere in the near-infrared and mid-infrared than given in Chapter 2. These plots are derived from models using the ATRAN software developed by Steve Lord (1992) and are available on the Gemini Observatory web site. Water vapor absorption is sensitive to altitude and occurs in certain wavelength intervals, between which the atmosphere is remarkably transparent. These atmospheric "windows" of transparency allow astronomers to define photometric bands. The standard windows are listed by central wavelength and the full width at half maximum (FWHM) intensity in Table 11.1.

There is also a relatively poor window (designated X) from about 30 μm to 35 μm which is accessible from dry high-altitude sites or from Antarctica. There are variations of some bands (such as K_{short} 2.0 μm–2.3 μm, K′ 1.95 μm–2.30 μm, and

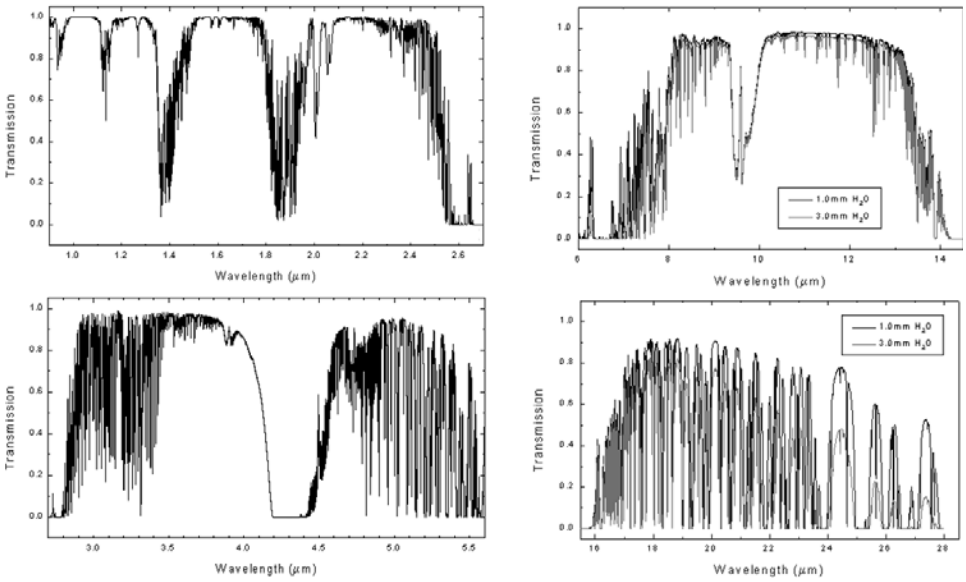


Figure 11.1. Details of the near-infrared transmission profile of the atmosphere above Mauna Kea (14,000 ft) for a typical water vapor level. Plots created using the ATRAN code developed by Steve Lord. Credit: Gemini Observatory web site.

L′ 3.5 μm–4.1 μm) which have been developed to improve performance at given sites. Interference filters can be manufactured to match these windows. Several different filter sets are in use, and therefore care must be taken when comparing photometric observations with those of others.

Table 11.1. Infrared windows in the Earth's atmosphere.

Center wavelength (μm)	*Designation of the bandwidth*	*Width (FWHM)* (μm)
1.25	J	0.3
1.65	H	0.35
2.2	K	0.4
3.5	L	1.0
4.8	M	0.6
10.6	N	5.0
21	Q	11.0

11.2.2 The high-background problem

There are two major sources of unwanted background photons. One component is OH emission lines and the other is black-body thermal emission from the telescope which, even at an ambient temperature close to 0°C (273 K), emits prodigiously in the infrared. Thermal emission from any warm optics in the beam can be predicted from two quantities: the absolute temperature T (K) which determines the spectrum of the radiation from the Planck function $B_\lambda(T)$, and the emissivity $\varepsilon(\lambda)$ of each component which determines the fraction of black-body radiation added to the beam. Objects which "appear" black to our visual senses may not be black at longer wavelengths (i.e., they may reflect some infrared light). To estimate the emissivity (ε) of telescope mirrors (due to absorption) we can apply Kirchhoff′s Law $\varepsilon = 1 - R$ where R is the measured spectral reflectivity. For example, if the reflectivity is measured to be 96% then the emissivity is 4% and is additive for the train of warm optics. There will be an additional component of emissivity from dust on the mirror surface. To emphasize how bright the night sky is at infrared wavelengths we can compare the brightness in magnitudes of one square arcsecond in the blue ($\lambda = 0.43\,\mu$m) $m \approx 24$ (no moonlight), with that at 2.2 μm in the near IR $m \approx 13.5$, and also at 10 μm where the sky and telescope combined are brighter than $m \approx 0.0$ (depending on emissivity and temperature)! The most effective way of eliminating telescope background is to cool the entire telescope. On Mauna Kea at 14,000 ft above sea level the temperature hangs around 1°C. Temperatures in Antarctica near the South Pole are lower still, ranging from −13.6°C to −82.8°C, and hence have stimulated the development of Antarctic astronomy. In the stratosphere the average temperature is about −50°C and the residual level of water vapor is extremely low. Consequently, flying a telescope on an airplane to this altitude can be a very effective alternative to space missions.

11.2.3 Chopping

Early infrared astronomers found a solution to the problem of a bright sky: a technique called "chopping". The infrared beam is rapidly switched between the source position on the sky and a nearby reference position, by the use of an oscillating or "wobbling" secondary mirror in the telescope itself. Typical wobbling secondary mirrors on IR-optimized 3 m–4 m telescopes such as the IRTF and UKIRT are small, 0.24 m–0.31 m in diameter, and have a slow f/ratio ($\sim f/35$). On the larger 8 m–10 m telescopes these secondary mirrors are bigger and more massive (e.g., 1.0 m for Gemini). Moreover, on alt-az telescopes the chop direction must be variable. Chopping typically takes place at a frequency of ~10 Hz–20 Hz. In a photometer, this method involves isolating the astronomical object in a small aperture and first measuring the total brightness of "object plus sky" included in the aperture. Chopping changes the location of the image in the focal plane quickly so as to record the signal from a nearby piece of sky containing no objects in view. By forming the difference, the sky signal is eliminated provided it has remained constant. In addition, it is usually necessary to move the entire telescope every minute or so to enable the sky

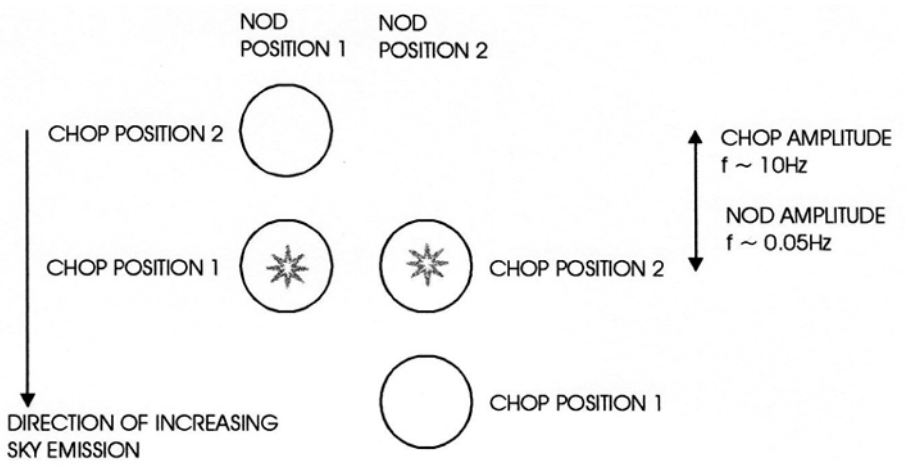


Figure 11.2. Chopping and nodding remove background flux and gradients at infrared wavelengths.

on the "other side" of the object to be measured and thereby eliminate any systematic trend or gradient; this step is called "nodding", and the amount of the nod is usually the same as the "throw" of the chop for symmetry (Figure 11.2). The difference between the pair of chopped signals for nod position 1 is given by

$$C_1(x) = S + B_{\text{tel},1} - B_{\text{tel},2} + \left(\frac{d}{dx} B_{\text{sky}}\right) \Delta x \tag{11.1}$$

where B_{tel} and B_{sky} are the telescope and sky backgrounds at the two chop positions separated by Δx. These terms are usually always much larger than the source flux, S. For the second nod position the signs are reversed and the difference signal is

$$C_2(x) = S - B_{\text{tel},1} + B_{\text{tel},2} - \left(\frac{d}{dx} B_{\text{sky}}\right) \Delta x \tag{11.2}$$

and adding these two results gives the required source signal

$$S = \tfrac{1}{2}(C_1(x) + C_2(x)) \tag{11.3}$$

Chopping and nodding are generally required at wavelengths longer than about 3.5 μm, and nodding alone is required for good background subtraction at shorter wavelengths too.

Another reason for using a secondary mirror with a slow f/ratio is to significantly reduce the background on a given detector pixel by stretching the plate scale. For example, going from an $f/9$ secondary to an $f/36$ gives a smaller scale in arcseconds per mm by a factor of 4, and reduces the flux falling on each square millimeter by a factor of $4^2 = 16$. An infrared secondary mirror is "undersized" to permit chopping, and therefore it is overfilled by the beam from the primary mirror. This means that the primary mirror no longer defines the entrance pupil of the system; it is now defined by the size of the secondary mirror. In general, the secondary mirror is

not surrounded by a black baffle tube in the normal way, because it is important to ensure that any subsequent image of the secondary formed inside the instrument is surrounded by sky, which produces a lower background than a warm, black baffle. Often, the secondary will be gold-coated for best infrared performance since gold is more reflective than aluminum in the IR. In addition, there will be either a small deflecting mirror or a hole at the center of the secondary with access to the sky. Such precautions eliminate thermal photons from the central Cassegrain hole in the primary mirror. Telescopes built this way are said to be infrared-optimized.

11.3 INFRARED ARRAY DETECTORS

11.3.1 The infrared "array" revolution, *déjà vu*

Lacking the long pre-CCD history of photographic imaging enjoyed by optical astronomy, it is easy to appreciate the staggering boost to infrared astronomy that occurred when the first true array detectors were introduced (e.g., McLean, 1988, 1995). Recent reviews of infrared detectors and materials are given by Paul Richards and Craig McCreight in *Physics Today*, February 2005 and by George Rieke (2007) in the *Annual Reviews*.

Many forms of infrared array devices with closely packed pixels were constructed during the period 1974–1984 by several different companies, due mainly to the extreme importance of the infrared for military applications. Both CCD and CID (charge injection device) readouts were used, and different materials were evaluated. For example, Koch *et al.* (1981) described work on InSb monolithic charge-coupled infrared arrays at Santa Barbara Research Center; Kosonocky *et al.* (1981) reported on the development of a 256-element PtSi Schottky Barrier IR CCD line sensor at RCA Labs; Baker *et al.* (1981) described the work at Mullard in the U.K. to make a 32×32 HgCdTe photovoltaic array hybridized to silicon circuitry; and Rode *et al.* (1981) reported on Rockwell's hybrid arrays fabricated in HgCdTe or InAsSb and multiplexed to a Si CCD via direct injection. Hybrid array development was reviewed by D.H. Alexander of Hughes Aircraft Co. (1980). However, despite a rich technical literature, even by 1982 when I carried out a detailed survey at the suggestion of Malcolm Longair, Director of the Royal Observatory Edinburgh (ROE) and the U.K. Infrared Telescope, pixel formats for available devices that had made it outside the classified arena to astronomers were very small (32×32 or less); very few devices were actually for sale and none had the performance needed for low-background astronomical applications. Prospects seemed bleak.

An early champion of astronomical infrared array devices was Craig McCreight of the NASA Ames Research Center who led in-house tests and coordinated a major NASA-funded program involving a number of other groups (e.g., McCreight, 1981). As early as 1979, John (Eric) Arens and co-workers at the Goddard Space Flight Center tested a 32×32-pixel bismuth-doped silicon CID array made by Aeroject ElectroSystems at a wavelength of 10 μm (Arens *et al.*, 1981, 1983). Astronomical

observations were published in Arens *et al.* (1984) and in the thesis of Richard Tresch-Fienberg. Elsewhere, a 32×64 platinum silicide (PtSi) Schottky Barrier array was evaluated by the Kitt Peak National Observatory (Dereniak *et al.* 1984) and at the NASA Infrared Telescope Facility (Rich Capps) in collaboration with the U.S. Air Force.

A lot of military funding had gone into the development of mercury–cadmium–telluride (HgCdTe) devices for mid-IR work, but much of that work was classified. Long linear photodiode arrays using individual, switched MOSFET multiplexers were tested and described in the technical literature; one of the best of these arrays was a 32-element linear array of indium antimonide (InSb) developed by Jim Wimmers, Dave Smith, and Kurt Niblack at Cincinnati Electronics Corporation which was used successfully by astronomers in near-infrared spectrographs (Niblack, 1985). Each of the early devices always had some drawback for astronomy, such as poor quantum efficiency or high readout noise ($\sim$1,000 electrons) and, of course, just not enough pixels! The most hopeful sign came from tests of a 32×32 array of InSb detectors reported in 1983 by Judith Pipher and Bill Forrest of the University of Rochester (U.S.A.) at a NASA Ames detector workshop organized by Craig McCreight. This device was a ''reject'' loaned to them by Alan Hoffman, a former colleague who was now employed by Santa Barbara Research Center (SBRC) in California (now Raytheon Vision Systems). Astronomical results from this camera were published in Forrest *et al.* (1985).

Having visited Judy and Bill in Rochester and made them aware of UKIRT's interest in this kind of array from an, as yet, undisclosed source, they passed on the information to Alan Hoffman. Alan and I had already met, but I was not aware that he was the source of Judy and Bill's detector. We made contact again and Alan approached his senior management with the prospect that a major observatory was interested in helping to develop a device optimized for astronomy. Wisely, SBRC also approached the Kitt Peak National Observatory (now the National Optical Astronomy Observatories, NOAO) and found interest there too. From 1982 to 1984, I worked closely with senior SBRC engineers (Alan Hoffman and Jim West) and marketing personnel (Dick Brodie and Carol Oania) to define the technical specification and costs of a new infrared array device suitable for astronomy (McLean and Wade, 1984). Our goal was to image using sub-arcsecond pixels just like CCDs. Smaller beams on the sky meant much lower background levels per pixel, an unusual condition compared with the high-background strategic applications. Unfortunately, the original readout design proposed by SBRC had to be abandoned as too complex and subject to amplifier glow. Luckily, an alternative readout scheme was suggested and already tested by Al Fowler at NOAO, but with platinum silicide (PtSi) rather than InSb as the detector material. This was the basic source follower per detector. We all agreed that the alternative device was going to be noisier than desired but should still work, and fabrication began on the well-known 58×62 InSb array in 1984. Meanwhile, Judy Pipher, Bill Forrest, Giovanni Fazio (Harvard), and others were working towards an instrument definition for NASA's Space Infrared Telescope Facility and were also negotiating for detector development work at SBRC; this instrument would become the Infrared Array Camera (IRAC) on Spitzer.

At about the same time, contracts were being developed for second-generation Hubble Space Telescope instrumentation. Several groups of U.S. astronomers (in Arizona, Chicago, and Hawaii) had obtained access to new arrays made from mercury–cadmium–telluride (HgCdTe or MCT) from Jon Rode of the Rockwell International Science Center in Thousand Oaks, California. By a circuitous route I had already met Jon in 1982 when carrying out the detector survey for the Royal Observatory and UKIRT. Having learned that the U.K. had a program of HgCdTe array development at Mullard Ltd., I visited Ian Baker there, but he suggested that if our interests were only 1 μm–5 μm then I should try Rockwell in California. At our meeting in 1982, however, Jon and I concluded that the development of a low-background, near-infrared MCT array would be too expensive for ground-based astronomy at that time. In addition, this material was still largely classified. By the mid-1980s, however, things were beginning to change. Stimulus for MCT array development for astronomy was generated by funding for a proposed new instrument for the Hubble Space Telescope. The instrument was called NICMOS and the Principal Investigator was Rodger Thompson at the University of Arizona. Marcia Rieke and Rodger Thompson at the University of Arizona, and Mark Hereld at the University of Chicago were among the first to successfully demonstrate the new Rockwell arrays at the telescope. Subsequently, Jon Rode moved up in the company at Rockwell and his role was taken over by Kadri Vural. Both Jon and Kadri have been strong supporters of the astronomy detector programs ever since, and over the years Kadri has played a very significant role in encouraging and promoting the development of (MCT) arrays for wide use. This development would prove to be very important, because these array devices can be customized to the shortest IR wavelengths (~2.5 μm) and could be run at 77 K using liquid nitrogen and operated by existing CCD controllers.

Several European sources of infrared array technology, for space applications in particular, were also recognized, and French astronomers were already using an InSb array of 32×32 pixels with the charge injection principle (Sibille *et al.*, 1982). Preparations for both ISO and SIRTF (Spitzer) had stimulated work on longer wavelength devices in both Europe and America, again much of the NASA development being supported by Craig McCreight's program. In addition, a very important development that had occurred in 1979 would soon make its way to astronomy arrays. Mike Petroff and Dutch Stapelbroek, at what was then another division of Rockwell (now DRS Technologies), perfected a new way to construct extrinsic silicon photoconductors that significantly improved their performance. The method was called blocked impurity band (BIB), and this array technology would be developed later for Spitzer and ground-based applications. In addition, work was under way at the University of Arizona and elsewhere to develop doped germanium detectors into a viable form of array device. Thus, excitement was high!

The first project to develop a true, user-friendly, facility-class infrared array camera based on the new 58×62 InSb array began at the Royal Observatory Edinburgh (ROE) in June 1984. I was the principal investigator for that development, and two years later in September 1986 we delivered IRCAM to the 3.8 m UKIRT in Hawaii (Figure 11.3). I was fortunate to have the enthusiastic support of Eric Becklin

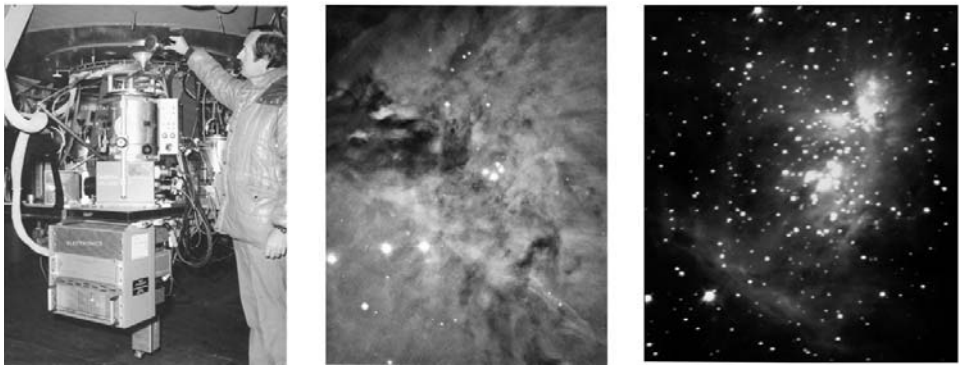


Figure 11.3. (Left) The author with IRCAM (1986), the first common-user camera system on the UKIRT 3.8 m infrared telescope to employ the 58×62 InSb arrays from SBRC (Raytheon). (Center) a visible light image of the Trapezium region of the Orion Nebula. (Right) An infrared image of the same region obtained with IRCAM at a wavelength of 2.2 microns. The bright source above the Trapezium is the Becklin–Neugebauer (BN) object. See also Plate 13.

who was at ROE on sabbatical leave from the University of Hawaii. Al Fowler received his device at NOAO at about the same time, and so he and I were frequently in touch across continents, trying to compare results as we went along. Indeed, the development was not trouble-free. For example, the batch of InSb used to make the astronomy devices suffered a loss of quantum efficiency when cooled to operating temperature ($\sim$30 K). This was a trying time for SBRC manager Dave Randall and scientists like Alan Hoffman and Geoff Orias. Fortunately, InSb material with completely different doping had been developed simultaneously for the SIRTF (Spitzer) project, and so the ground-based program was able to acquire some of that material which performed very well. So little was known about how these "astronomy arrays" would work that Al Fowler and I, as well as Alan Hoffman's team at SBRC, wanted to be very cautious until we had "first light" on the telescope. In my case that event occurred at 8 AM on the morning of October 23, 1986 in broad daylight! Together with my long-time colleague Colin Aspin, who had written much of the software for the camera, we obtained the first infrared image of a cosmic source with IRCAM, it was the Orion Nebula (far right in Figure 11.3). Also present at the telescope were our colleagues Gillian Wright (now PI of the MIRI instrument for JWST) and UKIRT operator Dolores Walther, who were both more than a little surprised by the quiet manner in which Colin and I accepted this momentous event. Of course, we had taken thousands of test images and knew that the camera would work, and certainly on something as bright as the BN object. We became much more jubilant in the months ahead as more and more challenging targets were observed, and it became clear that the flood gates had opened (McLean, 1987a, b; 1988). Two graduate students working on that project with me were Mark McCaughrean (1988) and John Rayner (1988), and they were responsible for producing remarkable images of OMC-1 and OMC-2 with IRCAM. Although the SBRC detector had an array of only 3,596 pixels, that was 3,595 more pixels than we had before!

By March 1987 the first astronomical results from several of the new infrared arrays had begun to appear. A key moment in infrared astronomy was a "workshop" on infrared array detectors in Hilo, Hawaii in March 1987. The meeting was organized by Eric Becklin and Gareth Wynn-Williams of the University of Hawaii, Honolulu, with local support from David Beattie and me on behalf of UKIRT. Don Hall gave the summary at the end of the meeting. The first true images from the new arrays were very encouraging, and we all realized that infrared astronomy had changed. To those of us who had straddled the apparent divide between optical and infrared astronomy, it was like history repeating itself. For me, the euphoria was similar to the Harvard–Smithsonian meeting on optical CCDs in 1981.

Eric Becklin and I joined forces in 1989 and moved to the University of California, Los Angeles to build infrared instruments for the new Keck 10 m telescopes. Six years later, in July 1993, we hosted a meeting at UCLA entitled "Infrared Astronomy with Arrays: The Next Generation". By then, everyone was already using 256×256 detectors for near-infrared work and 128×128 devices at mid-IR wavelengths; and plans for $1{,}024 \times 1{,}024$ arrays were announced at that conference (McLean, 1994; 1995). Short-wavelength mer-cad-tel was emerging as a powerful means for optical telescopes to extend their capability to 2.5 μm, while the longer wavelength arrays of InSb and extrinsic silicon (Si : As) were proving better than anyone had dared hope. Moreover, it was clear from the papers presented by 300 participants from all over the world that the new detectors had been quickly assimilated into the subject and that a wide range of new astrophysics was being produced. At the time of writing (2008), near-infrared arrays (1 μm–5 μm) with formats of $2{,}048 \times 2{,}048$ pixels and mid-infrared arrays (5 μm–30 μm) with $1{,}024 \times 1{,}024$ pixels are standard. Gallium-doped germanium detectors for 70 μm and 160 μm have been made into arrays of 32×32 pixels for space applications. Moreover, just like CCDs before them, these devices are being built into larger mosaics for both cameras and spectrometers, or becoming the heart of more complex instruments such as diffraction-limited cameras and integral field spectrometers.

11.3.2 The hybrid structure

To generate an infrared image an IR array must convert radiation into electrical charge by the internal photoelectric effect or it must absorb the energy with a bolometer and

- store the electrical charge at the site of generation (i.e., in a pixel)
- transfer the charge on each pixel to a single (or a small number of) outlets (the multiplexing task)
- enable the charges to be removed sequentially as a voltage which can then be digitized.

As each of the steps and stages in infrared detection are so similar to those employed already in a silicon CCD, it was natural to attempt to manufacture an entire CCD from some of the other known semiconductor materials. The problems of

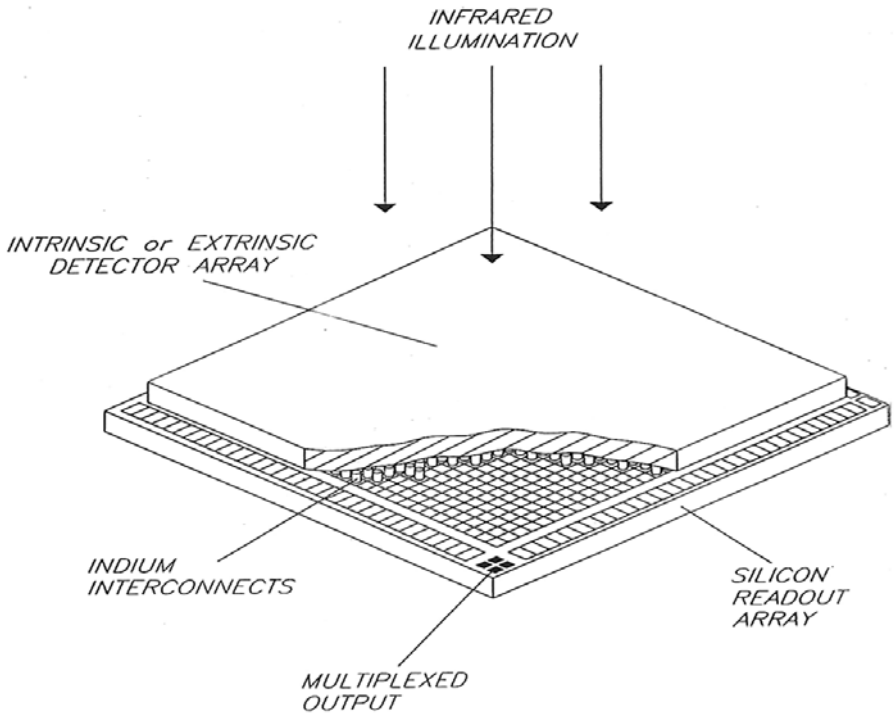


Figure 11.4. The "hybrid" structure of infrared array devices. The two slabs are separated by a grid of tiny indium bumps that remain soft at cryogenic temperatures.

manufacture turned out to be very difficult due to the limited experience in processing and purifying such materials compared with the extensive technology base for silicon. Finally, a different approach was taken, the "hybrid" array.

In the hybrid infrared array, the functions of detecting infrared radiation and then multiplexing the resulting electrical signal are separated, with the latter task going, of course, to silicon. At present then, infrared arrays are like "sandwiches" (Figure 11.4) in which the upper slab is the IR sensor (e.g., InSb, HgCdTe; Si : As, Ge : Ga), and the lower slab is a silicon multiplexer of some sort. The infrared part is really a tightly packed grid of individual IR pixels with minimum dead space between them. Initially, pixel sizes were quite large, 76 μm in the case of the 58×62 InSb array, but the current generation of IR arrays have pixels in the range 18 μm–27 μm. Both slabs are provided with a grid of electrical connections in the form of tiny raised sections, referred to as "bumps", of an electrical conductor called indium; indium remains soft at low temperatures. The slabs are literally pressed together to enable the indium bumps to mate. Gaps between the bumps are usually backfilled with an epoxy to help maintain the integrity of the sandwich, especially through thermal cycles to low temperatures. A microscopic array of "switches" made from metal oxide semiconductor field effect transistors (MOSFETs) are used to access the signal from each

IR detector (whether photodiode or photoconductor). Charge storage may occur on the junction capacitance of the IR sensor itself (in the case of a photodiode) or on a separate storage capacitor associated with the silicon circuitry. The entire structure is often called a focal plane array (FPA) or a sensor chip assembly (SCA), and the silicon "readout-integrated circuit" chip by itself is called a ROIC. Initially, two approaches were used in multiplexing the signal outputs from each unit cell. In one approach, each pixel is read sequentially by connecting its signal to an output bus, whereas in the second approach each pixel can be accessed randomly for connection to the output amplifier. The latter method, called direct readout (DRO) is very attractive but requires more circuitry; the SBRC 58×62 InSb array had this feature. Sequential readout is more easily implemented in the multiplexer using CMOS shift registers, and this approach is now standard.

It should be clear by now that infrared array detectors are not based on the charge-coupling principle of the silicon CCD. This is an important distinction with some practical implications. For example, IR arrays do not "bleed" along columns when a pixel saturates, and bad pixels do not block off others in the same column. Also, "non-destructive" readout schemes are possible and very effective. On the other hand, on-chip charge binning and charge-shifting are not possible. The "unit cell" of an infrared array contains a silicon field effect transistor (FET) used as a source follower amplifier, essentially providing a buffer for the accumulated charge in the infrared pixel. The term used to describe this structure is SFD (source follower per detector). Either the upper slab is constructed on an IR-transparent substrate (e.g., sapphire) or the bulk semiconductor is physically thinned to enable photons to penetrate to the pixel locations on the underside. A typical IR array unit cell with photodiode pixel elements is shown in Figure 11.5 and the details of the detection process are summarized below:

- the internal photoelectric effect produces electron–hole pairs
- an electric field separates the electrons and holes
- migration of electrons across the junction decreases the reverse bias, discharging a capacitor
- the amount of charge is $Q = CV/e$ electrons where $e = 1.6 \times 10^{-19} C/e^-$, V is the voltage across the detector, and C is the effective capacitance (which is a function of pixel geometry, doping, and bias conditions)
- each detector is connected to a source follower (SF) amplifier whose output voltage follows the input voltage with a small loss of gain (i.e., $V_{\text{out}} = A_{\text{SF}} V_{\text{in}} \sim 0.7 V_{\text{in}}$)
- the output voltage of the source follower can be sampled or "read" (with an analog-to-digital converter) without affecting the input
- after sampling, the voltage across the diode can be *reset* to the full reverse bias in readiness for the next integration.

The reset action is accomplished with another FET acting as a simple on/off switch in response to a pulse applied to its gate. Different forms of infrared arrays are summarized below.

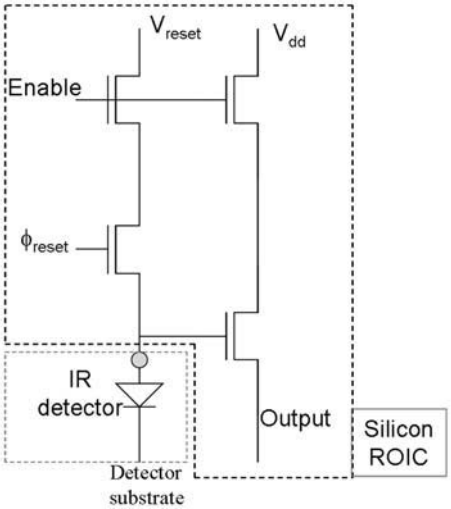


Figure 11.5. The "unit cell" of a typical near-infrared (photodiode) array; this is a four-transistor unit cell. Each detector has an output source follower and a reset switch, and there are two other FETs used for addressing the pixel.

11.3.3 Photovoltaic devices

Teledyne Imaging Sensors (formerly Rockwell Scientific Imaging, Camarillo, CA) has led the development of large-format HgCdTe arrays for astronomy (see Vural, 1994; Garnett and Beletic, 2004). A process known as PACE-I (Producible Alternative to CdTe for Epitaxy) was used initially to fabricate arrays on 76.2 mm (3-inch) diameter wafers. Briefly, metal-organic vapor phase epitaxy (MOVPE) CdTe is deposited onto a polished sapphire substrate, and then HgCdTe is grown via liquid phase epitaxy (LPE) from a Te-rich solution to a thickness of about 13 μm. The detector junctions are formed by boron implantation and then passivated by ZnS. Light enters from the backside through the sapphire which can transmit out to 6.5 μm. The percentage of Hg to Cd determines the cutoff wavelength. For example, with $Hg_{1-x}Cd_xTe$, then for $x = 0.196$ $E_G = 0.09\,\text{eV}$ and $\lambda_c = 14\,\mu\text{m}$. Similarly, $x = 0.295$ yields $E_G = 0.25\,\text{eV}$ and $\lambda_c = 5\,\mu\text{m}$, and $x = 0.55$ gives $E_G = 0.73\,\text{eV}$ and $\lambda_c = 1.7\,\mu\text{m}$. Newer devices are made by molecular beam epitaxy (MBE) and are "substrate-removed" somewhat like thinned, back-illuminated CCDs, and it is this process that not only provides response down to visible wavelengths but also improves resistance to particle damage for space applications. Devices with these properties spanning the range 0.5 μm–2.5 μm are being developed for NASA's James Webb Space Telescope. Raytheon Vision Systems (formerly Santa Barbara Research Center, Goleta, CA) has also produced a family of HgCdTe arrays for astronomy (see below), and 16 of these are featured in what is currently the largest IR camera.

Indium antimonide photodiode arrays specifically designed for astronomy are available from Raytheon Vision Systems (RVS). The pn junctions are diffused into

the InSb substrate. In the earliest SBRC arrays the photox front-side passivation was hygroscopic and required a gate electrode to control the front-surface potential and hence the dark current. With the introduction of the 256×256-pixel array, a new gateless detector-side passivation was used which produced a marked improvement in performance. As the indium bumps are on the surface of the InSb above the pn junctions, illumination must be from the backside. As the substrate is not transparent, InSb arrays must be thinned to about 10 μm by diamond machining or etching, again similar to the production of back-illuminated silicon CCDs. A thick wafer of infrared material connected to a silicon readout using indium bumps and epoxy will tend to pull apart as the device is cooled due to the difference in thermal expansion between the materials. If the layer is very thin, however (as in the InSb case), it is more compliant, like a rubber sheet attached at many points to a rigid block. HgCdTe arrays, however, require a balanced composite substrate (BCS), or the substrate must be removed.

11.3.4 Impurity band conduction devices

Initially, most long-wavelength arrays used extrinsic silicon photoconductors, but these have been replaced by the technology called impurity band conduction (IBC). IBC is the generic name for the blocked impurity band (BIB) devices invented by Mike Petroff and Dutch Stapelbroek in 1979 at DRS Technologies (formerly Boeing and originally Rockwell, Anaheim, CA). Raytheon Vision Systems make similar devices. Although readily available, extrinsic silicon photoconductors doped with As, Ga, or Sb suffer from many drawbacks which seriously degrade performance. For example, the quantum efficiency is dependent on the concentration of the selected impurity atoms, but there are also unwanted impurities such as boron which can nullify the effect of the given dopant. Usually the concentration of the impurity must be kept low to prevent tunneling or "hopping", but then the only way to increase QE is to make the infrared active layer very thick (several hundred microns), which leads to many operational problems. IBC detectors overcome these difficulties.

In an IBC device, a heavily doped infrared-active layer is placed in contact with a pure (undoped) epitaxial layer (the blocking layer) so that the overall thickness of the device can be greatly reduced (see Figure 11.6). The blocking layer is isolated by an oxide layer from metal contact pads and the device is usually back-illuminated. Dark current due to "hopping" is prevented by the blocking layer which enables much higher doping levels to be used in the active layer. Although it is ten times thinner than the equivalent photoconductor, the device can achieve high quantum efficiency. Because of the high donor density, the applied (bias) electric field causes a migration of holes toward the (negative) metal contact forming a depletion region similar to that in a photovoltaic device. As a result, IBC devices do not exhibit generation–recombination noise since the collected electrons are transported over a region devoid of holes. Under normal bias conditions IBC devices will not exhibit any "photoconductive gain", but gain is possible at higher bias values (see Rieke, 2003 for a more extensive discussion). These devices are constructed in the following

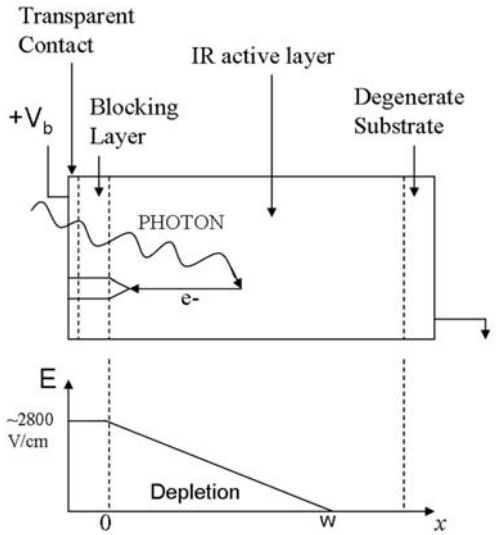


Figure 11.6. Illustration of the detection principle and operation of a typical blocked impurity band (BIB) or impurity band conduction (IBC) device.

manner. The process starts by placing a (to be buried) contact on a relatively thick (300 μm–500 μm) intrinsic Si substrate. The infrared-active (doped) layer is then grown, followed by the blocking layer, and the final contact is masked to define the individual pixels and establish the connection to the multiplexer via indium bumps.

Mid-IR arrays include the 256×256, 320×240, and $1{,}024 \times 1{,}024$ extrinsic silicon devices from Raytheon and the 256×256 and $1{,}024 \times 1{,}024$ extrinsic silicon BIB detectors from DRS Technologies. The largest Raytheon device is the 1 K × 1 K Aquarius Si : As IBC detector for 5 μm–28 μm, which has 30 μm pixels, outputs in blocks of 128 columns or 32 columns (i.e., either 16 or 64 outputs and all input/output connections on two sides) to allow close-butting on the other two adjacent sides. There are two reference columns and the chip can be reset row by row or globally, and the integration capacity is programmable at either 1 million or 15 million electrons. Centered windowing is also possible and frame rates of 150 Hz can be achieved with 64 outputs and 3 MHz pixel rates. Input-referred noise in the high-gain (smaller well) state is $\leq 1{,}000e^-$ rms. With anti-reflection coating, the QE > 40% and the device is operated at 8 K–10 K. Likewise, DRS Technologies has developed a $1{,}024 \times 1{,}024$ Si : As high-flux mid-IR focal plane array using their silicon blocked impurity band (BIB) technology. The pixels are 18 μm square, with a charge capacity of 5 million electrons, and the readout noise at maximum frame rates (100 Hz) is $\leq 1{,}000e^-$ rms. Gain is selectable for high and low flux. Operating temperature range is 2 K–12 K, and at 6 K the dark current is $< 10e^-/\text{s}$. The array is two-side buttable. This device has windowing capability for the central 256×256 pixels, integration time control, and 16 outputs.

11.3.5 Far-infrared arrays based on germanium

For wavelengths longer than 40 μm there are no appropriate shallow dopants for silicon, and therefore extrinsic germanium (Ge) must be used. There are a number of problems with the use of germanium. For example, to control dark current the material must be relatively lightly doped, and therefore absorption lengths become long (3 mm–5 mm). Because the diffusion lengths are also large (250 μm–300 μm) then pixel dimensions of 500 μm–700 μm are required to minimize crosstalk. Large pixels imply higher hit rates for cosmic rays, especially in space applications, and this in turn implies that the readout device must have very low noise so that the background limit is reached in the shortest possible exposure time. But a large detector pixel means a large capacitance and more noise. Also, the photoconductive gain is inversely proportional to the inter-electrode spacing resulting in poor QE, unless side-illuminated detectors with transverse contacts are used. Finally, because of the very small energy bandgaps, these detectors must operate at liquid helium temperatures well below the silicon "freeze-out" range. Despite all these challenges, a 32×32-pixel Ge : Ga array was successfully developed for the 70 μm band of the MIPS instrument on the Spitzer infrared space telescope as shown in Figure 11.7, together with a smaller array of stressed Ge : Ga detectors to extend the coverage to longer wavelengths (Young *et al.* 1998). By illuminating the detector pixels edge-wise the required long absorption path is provided and the readout circuitry is stacked behind the detectors in what is called *Z*-stack or *Z*-plane construction. The silicon readout circuitry itself was specially developed but was based on previous designs by RVS. Because the detectors are 2 mm long they offer a large cross-sectional area for ionizing particle strikes in orbit, and the MIPS devices experience hits every 10–15 seconds. However, the electrical circuit uses feedback on the readout amplifier to quickly return the bias to normal, thus preventing crosstalk to neighboring pixels, and due to the sampling technique used (ramp sampling) it is possible to recover good data from before and after the hit. The PACS instrument on the Herschel far-infrared space telescope will have a similar 70 μm array like MIPS, and a similar array of stressed Ge : Ga detectors for a band at 160 μm was used on the AKARI satellite.

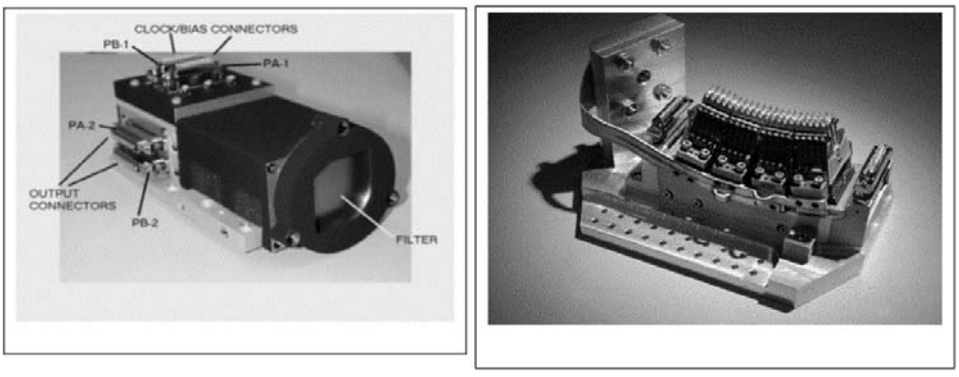


Figure 11.7. The Ge : Ga arrays used in the MIPS instrument in Spitzer. Credit: Erick Young.

11.3.6 Other forms of infrared arrays

There is a fundamentally different infrared array device which is used in industrial surveillance applications. This is the platinum silicide (PtSi) array. A PtSi array is based on the Schottky Barrier principle shown schematically in Figure 11.8. When a metal such as platinum silicide is brought into contact with p-type silicon, electrons flow across the junction until stopped by the electric field created by the additional negative charges in the silicon. At this point the Fermi levels are equal and the result is an asymmetric potential barrier for conduction holes. Holes in the metal must now overcome a slightly higher barrier to enter the valence band of the semiconductor. The barrier height ψ is determined by the contact potential and can be less than the semiconductor bandgap. For PtSi, $\psi = 0.22\,\text{eV}$ and the cutoff wavelength is $\lambda_c = 5.6\,\mu\text{m}$. Quantum efficiency is given by the product of the QE for absorption in the silicide layer (<10%) and the probability of producing a hole that will escape the barrier, $\frac{1}{2}[1 - \sqrt{(\psi/h\nu)}]^2$, so the QE falls steadily towards the long-wavelength cutoff; the intrinsic absorption in silicon sets the short-wavelength cutoff. Usually, the silicide layer is very thin and covered by a dielectric (insulating) layer of SiO_2 which has a metal (Al) surface layer. Together with an anti-reflection coating on the p-type silicon substrate, the entire device becomes an "optical cavity" which can be "tuned" in thickness so that interference effects increase the QE for certain wavelengths. Typical QEs are less than 2%–3%. This is far too low to be competitive with HgCdTe or InSb for astronomy. However, these devices can be made large and uniform, because it is an all-silicon process and they are not as expensive as the InSb and HgCdTe hybrid arrays. Glass, Sekiguchi, and Nakada (1995) describe a Japanese-made PtSi array of 1,040 × 520 pixels with QE of 6% at J and 2% at K, operated at 60 K, in an infrared camera.

Very narrow bandgap semiconductor crystals are harder to grow, so is tempting to seek a method of "engineering" small bandgaps in wide-bandgap material such as gallium arsenide (GaAs). Using GaAs and AlGaAs junctions it is possible to create a "quantum well" associated with either the conduction or the valence band. The quantum well is equivalent to the particle-in-a-box problem in quantum mechanics,

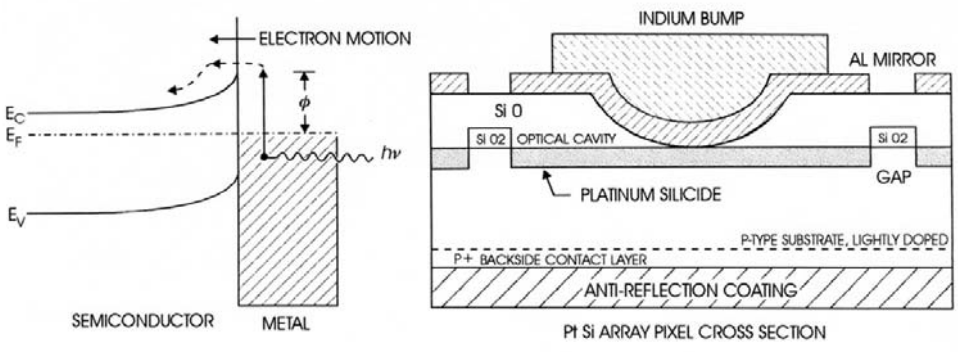


Figure 11.8. Principle of the platinum silicide (PtSi) device and the formation of a Schottky Barrier.

and so the well will contain energy levels or sub-bands. The energy difference between these sub-bands or levels ($h\nu$) is very small compared with the normal bandgap. The transition between sub-bands provides the required infrared photon detection process in these devices known as QWIPS (Gunapala *et al.*, 1994; Rieke, 2003).

11.4 PRACTICAL OPERATION OF INFRARED ARRAYS

Although charge-coupling is not involved, infrared arrays still require a set of clock voltages to collect and remove charges. Output signals (a few $\mu V/e^-$) must be amplified and digitized into data numbers, pixel-to-pixel variations must be corrected with a flat-field, kTC-reset noise must be eliminated with correlated double-sampling, and dark current must be minimized by cooling. Most importantly, the same image-processing methods can be used.

11.4.1 Linearity

In IBC devices the storage capacitor is external to the pixel and fixed. Typical near-infrared arrays have at least two sources of capacitance, however, the pn detector junction (C_{det}) and C_{FET} (the source follower FET) plus other stray capacitance. In both cases there are two sources of current to drive the discharge of the capacitors, namely photoelectrons and "dark current" electrons. There is an obvious potential for non-linearity between photon flux and output voltage in the photovoltaic detectors which is absent in CCDs, simply because the detector capacitance is *not* fixed, but does in fact depend on the width of the pn depletion region, which in turn depends on the value of the reverse bias voltage. (Variations in the size of the depletion region do occur in a CCD, but this is a small effect.) The bias voltage changes continuously as the cell integrates, irrespective of whether it is storing photogenerated charges or dark current charges. Using $Q = CV$ and taking both C and V as functions of V then

$$dQ = \left(C + \frac{\partial C}{\partial V} V \right) dV \equiv I_{\text{det}}\, dt \tag{11.4}$$

The rate of change of voltage with time (dV/dt) is not linear with detector current I_{det} because of the term $\partial C/\partial V$. Representing all fixed capacitances as C_{fix} then the denominator can be shown to be given by

$$C + \frac{\partial C}{\partial V} V = C_{\text{fix}} + C_0 \left(\left(1 - \frac{V}{V_{\text{bi}}} \right)^{-1/2} + \frac{1}{2} \frac{V}{V_{\text{bi}}} \left(1 - \frac{V}{V_{\text{bi}}} \right)^{-3/2} \right) \tag{11.5}$$

where C_0 is the junction capacitance at zero bias

$$C_0 = A_{\text{det}} \left[\frac{e\varepsilon_s}{2V_{\text{bi}} \left(\dfrac{1}{N_A} + \dfrac{1}{N_D} \right)} \right]^{1/2} \tag{11.6}$$

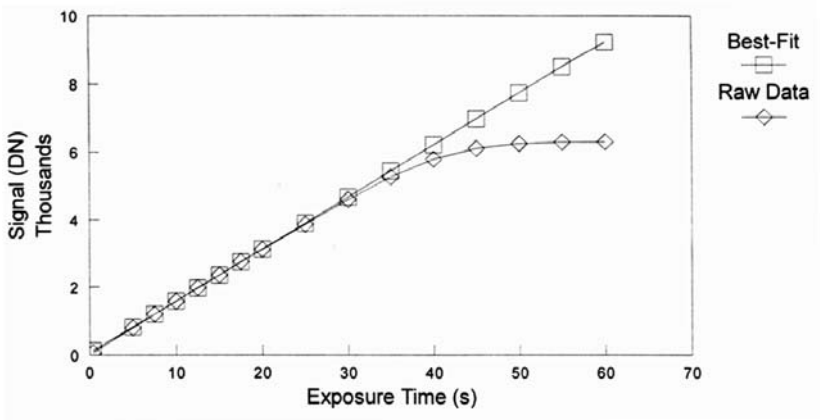


Figure 11.9. The non-linearity of an NIR array due to the voltage dependence of junction capacitance.

and A_{det} is the detector area; N_A and N_D are the acceptor and donor doping concentrations (atoms/cm^3), with $N_A \gg N_D$ to form an abrupt junction; ε_s is the permittivity of the material in the junction (e.g. $17.7\varepsilon_0$ for InSb); and V_{bi} is the built-in potential of the diode given by

$$V_{\text{bi}} = \frac{kT}{e} \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{11.7}$$

where n_i is the intrinsic carrier concentration; T is the absolute temperature; and the constants k and e have their usual meaning. As shown in Figure 11.9 the effect is relatively weak, rather slowly varying (<10% worst case) and easily calibrated to high precision with a series of flat-fields at different exposure levels.

11.4.2 Dark current and cooling

There are three main sources of dark current: diffusion, thermal generation–recombination (G–R) of charges within the semiconductor, and leakage currents. The latter are determined mainly by manufacturing processes and applied voltages, but diffusion currents and G–R currents are both very strong (exponential) functions of temperature and can be dramatically reduced by cooling the detector. For a photodiode, the sum of the diffusion, G–R, and leakage currents is given by

$$I_{\text{dark}} = \frac{kT}{eR_{0_{\text{diff}}}}(\exp^{eV/kT} - 1) + \frac{2kT}{eR_{0_{\text{GR}}}}\left(1 - \frac{V}{V_{\text{bi}}}\right)^{1/2}(\exp^{eV/2kT} - 1) + I_{\text{leak}} \tag{11.8}$$

where V is the voltage across the detector; and $R_{0_{\text{diff}}}$ and $R_{0_{\text{GR}}}$ are the detector impedances at zero bias for diffusion and generation–recombination. Typical values for InSb detectors are about $10^5\,\Omega$ at 140 K and $10^{10}\,\Omega$ at 77 K, respectively. Note that the "thermal voltage" $kT/e = 4.3$ mV at 50 K, and therefore $\exp[eV/kT] \gg 1$ for typical bias values ($V > 100$ mV) and temperatures ($T < 77$ K). Diffusion

dominates at high temperatures, but below 100 K dark current is mainly G–R and falls almost exponentially until limited by small leakage currents.

Because the bandgaps are much smaller than for silicon, infrared arrays must be cooled to lower temperatures than CCDs. Dark currents below 1 electron/minute/pixel have been achieved in the most recent HgCdTe arrays at 77 K (LN_2), and about 0.1 electron/s/pixel in InSb arrays with low bias and temperatures about 30 K. For IBC arrays with even smaller effective bandgaps for operation at ~20 μm, the detector must be cooled to ~4 K. A rough rule of thumb (Hoffman, Love, and Rosbeck, 2004) for the maximum operating temperature ($T_{\max}$) for a detector with cutoff λ_c is

$$T_{\max} = \frac{200\ \text{K}}{\lambda_c\ (\mu\text{m})} \tag{11.9}$$

This result implies temperatures below 80 K for a 2.5 μm cutoff, below 40 K for a 5 μm cutoff, and less than 10 K for material with a 20 μm cutoff.

11.4.3 Noise sources

As with CCDs, the readout noise describes random fluctuations in voltage which are added to the true signal during readout. Similarly, readout noise is always converted from a voltage (V_{noise}) to an equivalent number of electrons (R) at the detector by using the effective capacitance (C) (i.e., $R = CV_{\text{noise}}/e$ electrons). Reset or "kTC noise" occurs when the detector capacitance is recharged (i.e., reverse bias is re-established). When the reset transistor is on, the voltage across the detector increases exponentially to the reset value V_{RD} with a time constant of $R_{\text{on}}C$, where R_{on} is the "on" resistance of the transistor; this time constant is very short. Random noise in the reset charge at times very much longer than this time constant has a root-mean-square (rms) voltage noise of $\sqrt{(kT/C)}$ or an equivalent charge noise of $\sqrt{(kTC)}$, where k is the Boltzmann constant, and T is the absolute temperature. Like a CCD output, $R_{\text{off}}C$ is very long, and thus the reset noise is frozen.

Example. Estimate the kTC noise in electrons for a capacitance of 0.07 pF at 77 K. We use the form $\sqrt{(kTC)}$ to obtain a charge of $\sqrt{(1.38 \times 10^{-23} \times 77 \times 7 \times 10^{-14})} = 8.62 \times 10^{-18} C$ and divide by $e = 1.6 \times 10^{-19} C$ to convert to electrons. Thus, kTC noise is ~54 electrons at 77 K.

The unknown offset in voltage can be eliminated by taking the difference between the output voltages before and after reset. This procedure is just correlated double-sampling. Typical noise values for NIR arrays are ~15–25 electrons in CDS mode, but ~5 electrons has been obtained using "multiple reads" with the non-destructive readout feature of these arrays. MIR devices are usually noisier but often are used in background-limited applications in any case. Background signals in broad-band imaging applications can fill the pixels in fractions of a second depending on wavelength. Thus, the CCD equation still applies.

11.4.4 Quantum efficiency

InSb and HgCdTe arrays used in the near-IR have very good QE, typically >80% with suitable anti-reflection coatings. Mid-IR arrays of doped silicon have lower QE in the 30%–40% range depending on operation. For the near-infrared detectors, non-uniformities in thinning, the quality of the passivation applied to make the detector surface chemically inert, variations in doping density through a substrate or impurity centers in liquid phase epitaxy are all important in controlling QE. Since the absorption cross-section changes with wavelength, the quantum efficiency also varies with wavelength, usually decreasing towards shorter wavelengths.

11.4.5 Multiple outputs

One big practical distinction between IR arrays and CCDs is immediately apparent at the telescope, especially for longer wavelengths. When an infrared image of a relatively faint object is obtained and displayed, all that is visible is the flat-field pattern of the device and any bad pixels! The reason is the huge infrared background flux (16 magnitudes per square arcsecond at 1.25 μm rising to around 0th magnitude per square arcsecond at 10 μm). To be able to see anything you must subtract the background. This is most conveniently done by moving off the target object to a relatively blank area of sky nearby and repeating the exposure. Now, if the *difference* of these two images is formed and displayed, then the background flux will be eliminated (to first order) and faint objects will be visible; this is "nodding" if accomplished by moving the whole telescope and "chopping" if using the wobbling secondary mirror. Infrared cameras must provide this feature of displaying a difference image, otherwise it is very hard to know what is happening at the telescope! If you wait too long to observe the sky value then it might change, and even a small change will cause the difference to fail, and remnants of the flat-field structure will just reappear. Because the background flux is so very large, a variation of ~1% may swamp your target. Sky subtractions do not correct for pixel-to-pixel variations in QE and so flat-field corrections must still be applied for photometry. Higher backgrounds imply shorter saturation times. To deal with this condition infrared arrays have multiple outputs. The minimum number is 4, most have 32 outputs, and some have 64 or more. Readout arrangements vary from device to device. Figure 11.10 shows the patterns for two particular arrays.

11.4.6 Array controllers

Some arrays, such as the HgCdTe devices developed by Teledyne, use simple CMOS shift registers with 0 V–5 V clocks, whereas the InSb arrays from Raytheon require clocks which operate from −3 V to −7 V, and hence a "level shifter" circuit is needed. Infrared array controllers are nevertheless very similar to those for CCD systems. Many observatories employ almost identical architectures with modifications to handle speed and special requirements (e.g., the ARC and MONSOON controllers

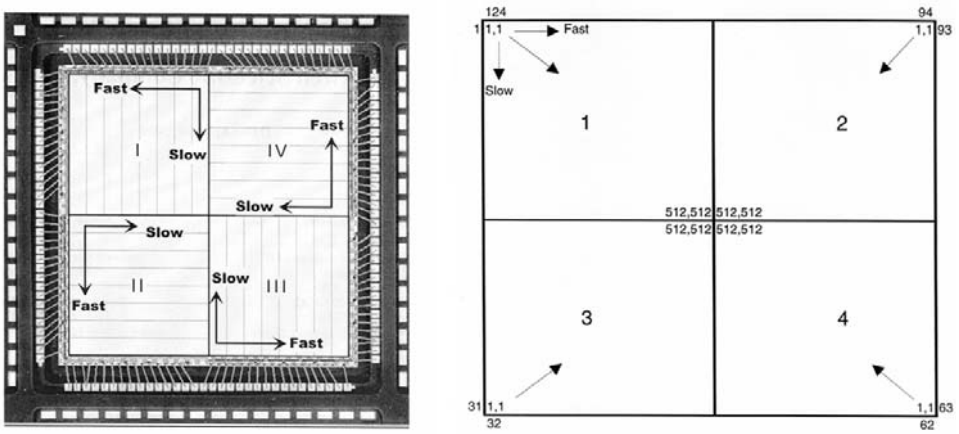


Figure 11.10. (Left) The HgCdTe 1,024 × 1,024 HAWAII array showing a readout scheme in each quadrant. (Right) The InSb 1,024 × 1,024 ALADDIN array showing the edge-to-center readout scheme.

previously mentioned operate both CCDs and IR arrays). Although details vary, any controller will need to provide the following basic features:

1. *Fast (column) register clocks*. When a given row is addressed, the fast register is clocked from column to column to enable every pixel in that row to be connected in turn to the output bus. Typical shift registers require one or two clock phases. A start pulse is also required in many cases.
2. *Slow (row) register clocks*. Each row is addressed in sequence by pulsing the row register clocks. This is a slow clock because between each transition sequence the column shift register is operated to scan a row. Again, one-phase or two-phase clocks and a start pulse are common.
3. *Reset clock*. A single voltage pulse is required to reset each pixel. Unlike a CCD, it is the charge collection node at each pixel site that is being reset rather than the output amplifier node. Some detectors cannot reset each pixel individually but must reset one row at a time, and some detectors have a "global" reset which can reset all the pixels at once. If the reset pulse is not sent after the pixel is addressed and the signal digitized, then the IR detector simply keeps on integrating. This is a "non-destructive" readout. This feature is absent from most CCDs unless the output has been specially designed with a floating gate (such as the "skipper" CCDs).

In addition to power supply lines, the most important dc bias voltages are

1. The substrate voltage (usually ground) and the detector substrate voltage or detector "common" supply, which may not be the same.

2. The drain voltage on the output transistor which should be minimized to reduce amplifier glow.
3. The unit cell drain voltage. In some arrays the "bias" across the detector which establishes the well depth is determined by the *difference* between the detector substrate (or common) and the unit cell drain voltage (i.e., $V_{\text{detbias}} = V_{\text{detsub}} - V_{\text{dduc}}$).

In some devices (e.g., the Aladdin 1,024 × 1,024 InSb arrays) dc bias lines that switch between two levels are required for the clamp lines (V_{ddCl} and V_{ggCl}) and pulse shaping with an RC time constant of ~1 μs is needed on the reset lines (V_{rstG} and V_{rstR}). In general, the detector outputs can be operated with either load resistors or a constant current source (~100 μA–200 μA). Finally, because most arrays have multiple outputs, many parallel channels of digitization are required.

Another significant difference between CCDs and IR arrays is the readout rate and the operational speed of the electronics. Even at the shortest infrared wavelengths the pixels can fill up in a few tens of seconds in broad-band imaging applications and the maximum integration time drops milliseconds in the thermal IR. In fact, it is customary to sum many short on-chip exposures in a "co-adder" before writing a data frame to disk. Thus, the integration time is the product of the on-chip exposure time and the number of "co-adds" used. Consequently, readout times for a full frame are much faster than with CCDs, and there is no time to employ integrating dual slopes when doing correlated double-sampling. The pre-amplifier must be dc-coupled to the output of the chip, and higher speed electronics must be used in general, which has implications for system noise. The pre-amp circuit shown in Figure 11.11 was developed at UCLA to operate a twin-channel IR camera using both InSb and HgCdTe arrays. It contains both a load resistor and a current source, based on the low-noise 2N4393 FET (Motorola, Inc.). Each channel has a gain op-amp. These op-amps are arranged in a non-inverting configuration and are offset by an adjustable voltage level from another op-amp in a buffer configuration with a potentiometer at the input.

An innovation for both space and ground-based astronomy is the use of "application-specific integrated circuits" or ASICs, to achieve considerable simplification of detector electronics and reduce power consumption. At the time of writing, only Teledyne provides a commercial ASIC to operate its ground-based H1-RG and H2-RG devices. The original small ASIC chip was developed for the James Webb Space Telescope (JWST) and adapted for ground-based use by designing a custom cryogenic card. This ASIC (known as SIDECAR) is capable of supplying all the clock and bias voltages to the devices, and digitizing the outputs to 16 bits before transferring the data stream to the host computer. The interface to the ASIC is via a USB link to a custom card called the Jade-2.

11.5 READOUT MODES

Unlike silicon CCDs, there is usually no shutter in an infrared camera to determine the exposure time. This is partially due to the requirement that the shutter be cold

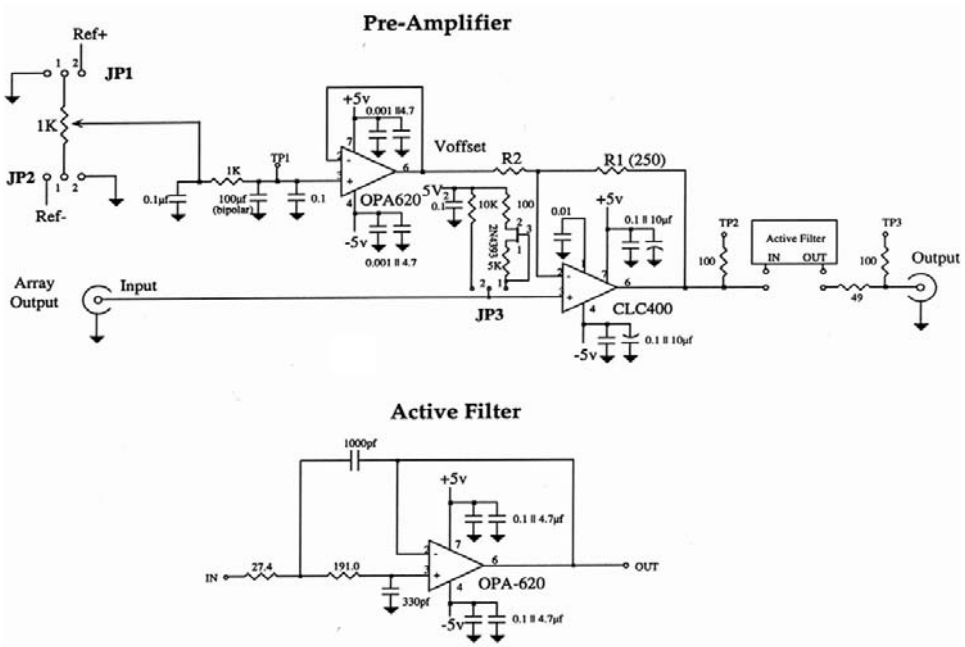


Figure 11.11. A high-speed, dc-coupled pre-amplifier design for IR arrays.

(inside the dewar), and that it operate very rapidly to accommodate the short exposures demanded by the high backgrounds. This does not mean that the light path cannot be blocked for dark-current measurements. Most IR cameras contain a position in the filter wheel which is opaque, usually a blank metal disk or plug, at the appropriate cryogenic temperature to eliminate thermal emission. In general, however, IR arrays are continuously exposed to light, and therefore the exposure time is controlled by the sequence of reset and read pulses. In practice, the time required to read out an infrared array depends on the settling time of the on-chip circuits.

Example. Suppose the device requires 4 µs for the output to settle when each new pixel is selected. The array is $256 \times 256 = 65{,}536$ pixels, but there are four independent outputs operating in parallel so the effective number of pixels is only one-fourth as many, or 16,384. The readout time is therefore $16{,}384 \times 4\,\mu s = 65.5$ ms per frame. This is called the "frame time" (T_F), and the "frame rate" is the reciprocal of this number ($1/T_F$), or 15.3 frames per second, usually stated as 15.3 Hz. Similarly, the pixel time of 4 µs corresponds to a pixel rate of $1/(4 \times 10^{-6})$ pixels/s or 250 kHz. We will use this example for the discussion on readout modes below.

A critically important feature of IR arrays is the ability to perform a nondestructive readout in which the charge on the detector side is not altered by sampling its present value. Figure 11.12 shows a schematic representation of the output voltage from an IR array. After reset, the signal shifts to the "pedestal" level, and then the detector begins to discharge due to photocurrent and/or dark current. Note that the

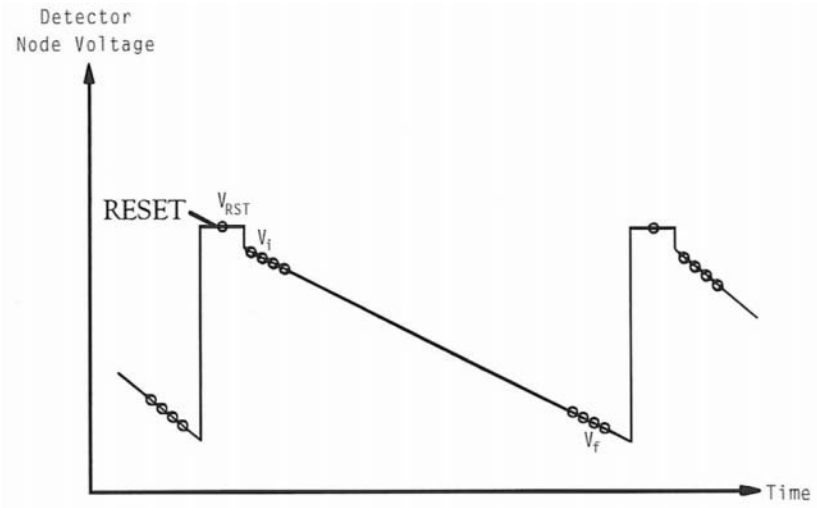


Figure 11.12. The schematic variation of the output voltage as a function of time for a typical pixel in an infrared array detector. Associated readout modes are described in the text.

sign of this plot is arbitrary. Some detectors discharge from more positive to less positive voltages, others discharge from negative voltages towards zero volts.

11.5.1 Single-sampling

To clear the detector of charge and define the beginning of the integration period the array must be reset. Usually, resets occur on a pixel-by-pixel basis as each pixel is selected (addressed) in turn. Some detectors have a "global reset" capability, and others can only be reset one row at a time. The process is generally performed at the same rate as the normal readout (e.g., ~66 ms in our example). Once a pixel has been reset it immediately begins to accumulate light. By the time the last pixel has been reset the first pixel has been integrating (for almost 66 ms in our example). No digitization occurs during the reset cycle; no data is stored in memory. If the required exposure time is T_i then readout must begin T_i milliseconds after the first pixel has been reset or $T_i - 65.5$ ms after the last pixel is reset. Relative timing between the "reset" waveforms and the "readout" waveforms must be precise and stable in order for the integration time to be well-defined and consistent. When the readout waveform is sent, each pixel address is accompanied by a "command-to-convert" pulse rather than by a reset pulse. Thus, each pixel value is digitized but not reset. Single sampling is incapable of removing kTC noise and drift, because only one sample is taken, but it is a useful mode that directly measures the signal relative to the reset or bias level and unambiguously detects saturation. The bias/reset level can be determined by using the shortest possible exposure (the readout time) and no illumination (achieved by blanking off the beam).

11.5.2 Correlated double-sampling (CDS)

There are several ways to perform CDS. One possibility is to digitize the reset level in the sampling scheme already described and subtract the two results. A second approach is to reset the pixel and then immediately digitize the level after the reset has been removed (the "pedestal" level) but before moving on to address the next pixel. Both of these approaches were common with the first IR arrays, but they turned out to be poor because of a fairly strong time-dependent response to the reset action which requires milliseconds—not microseconds—to settle out. In addition, the act of de-addressing the current pixel in order to address the next pixel adds noise.

11.5.3 Reset–read–read or Fowler sampling

The best scheme is to ripple through and reset the entire array pixel by pixel and then *immediately* read out the entire array again but non-destructively (i.e., no reset pulse sent), digitizing the signal in each pixel and saving this frame in memory. After waiting the required integration time (T_{int}), non-destructively read out and digitize the whole frame again (Figure 11.13). This method is called the "reset–read–read" mode or the Fowler sampling mode (after Al Fowler at NOAO who was the first to promote its use) and is less noisy than other methods. The effective signal is the second read minus the first read (i.e., $S_2 - S_1$):

$$S_2 - S_1 = [(T_{\text{ro}} + T_{\text{int}})\dot{N}_e + b + c] - [T_{\text{ro}}\dot{N}_e + b + c] = T_{\text{int}}\dot{N}_e \qquad (11.10)$$

The amplifier bias (b) and the unknown but correlated offset from the reset level (c) subtract out; and the dot in $\dot{N}_e$ represents that it is a count rate in electrons/s. This

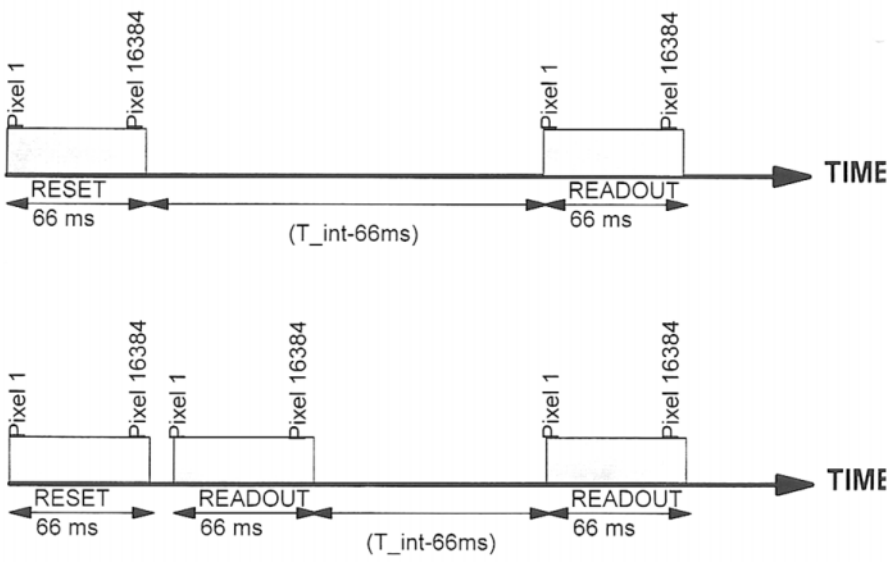


Figure 11.13. Graphical summary of the reset–read–read or Fowler sampling mode.

differencing process is usually accomplished in a front-end co-adder by storing the first read as a negative number and *adding* the second number to the stored value (rather than overwriting the value). The unknown kTC-noise and actual reset level are eliminated. If the exposure is very long, there may still be drifts which are not removed by this method. One way to remove those effects is to take a "dark" frame in the same sampling mode with the same exposure time and subtract it. This of course also removes the dark signal.

Using the reset–read–read mode makes it harder to know when the array has saturated. For instance, suppose the readout time is 66 ms, the pixel saturation level is 8,000 DN, but the flux rate from a certain star is 60,000 DN/s. In an exposure time of 0.1 s the flux is 6,000 DN in the pixel containing the star. Between reset and the first non-destructive read each pixel integrates for 66 ms, so the pixel with the bright star has a signal level of 3,960 DN when it is recorded. When it is read again for the second sample, 0.1 s have elapsed and another 6,000 DN of charge should have accumulated giving a total of 9,960 DN. But the saturation level is 8,000 DN, and additional flux above this level will not significantly affect the detector which is now fully de-biased. So, the actual value of the second sample will be 8,000 DN, and the difference, which is all that is saved, will be 4,040 DN instead of 6,000 DN. The detector is hopelessly saturated, but the displayed counts do not indicate this. The way to check this is to use single-sampling mode and determine the true count rate on the star. A line plot across the star image will show a dip in the center.

The reset–read–read process just described can be carried out multiple times to improve noise performance. The technique is the same as just described except that the initial (first) readout of the entire array following reset is repeated m times instead of just once. This was shown in Figure 11.12 for four samples. One cycle is often called Fowler-1, and so m cycles would be Fowler-m. At the completion of the required integration time the final full frame readout is also done m times. Of course, the exposure time T_{int} must be long enough to accommodate m frame times. This is the usual case for narrow-band imaging and for near-IR spectroscopy where this technique is very effective in reducing noise. The sum of the initial values is given by S_1 where

$$S_1 = \sum_{n=1}^{m} (nT_{\text{ro}}\dot{N}_e + b) \tag{11.11}$$

and T_{ro} is the readout time for one frame; $\dot{N}_e$ is the flux in photoelectron/s; and b is the bias offset. The sum of all the final readouts S_2 is given by

$$S_2 = \sum_{n=1}^{m} [(nT_{\text{ro}} + T_{\text{int}})\dot{N}_e + b] \tag{11.12}$$

and the difference is

$$S_2 - S_1 = mT_{\text{int}}\dot{N}_e \tag{11.13}$$

so that by dividing by m we get the desired integrated flux. The signal $S_2 - S_1$ is m times as large, but the readout noise adds in quadrature and is only $\sqrt{m}$ as large, so the effective readout noise in the final integrated flux is $R_{\text{eff}} = R(\sqrt{m})/m$ or $R/\sqrt{m}$. For

16 reads ($m = 16$), the effective readout noise is only a quarter of that obtained from a single pair of samples (Fowler-1). This technique, not available on a CCD, has enabled near-IR arrays to get down to below $5e^-$ rms noise despite their initially higher readout noise. There can be a penalty for some devices, however, because the amplifier glow associated with each readout cycle is cumulative.

11.5.4 Sampling up the ramp (UTR)

In this approach, the signal is sampled many times at regular intervals throughout the duration of the exposure, rather than multiple times at the beginning and at the end. Therefore, the signal can be seen to "ramp" up (Chapman *et al.*, 1990; Garnett and Forrest, 1993). This approach is very useful if some pixels will saturate before the end of the exposure time. It also helps for space applications where the backgrounds are low, and cosmic-ray hits are frequent during long exposures. Likewise, for low-background AO applications and for low-background high-resolution spectroscopy, this mode is practical. Finally, it is a good readout mode for establishing steady-state thermal conditions in the detector which reduces drift. A best-fitting straight line can be applied to the data points to get the mean flux rate. A complete mathematical analysis of multiple-sampled schemes was given by Garnett and Forrest (1993). These authors showed that both multiple Fowler sampling and line-fitting or "up-the-ramp" (UTR) sampling are superior to correlated double-sampling in readnoise-limited conditions (and when the readnoise is itself dominated by white noise). Both sampling methods provide the expected $\sqrt{n}$ improvement, where n is the number of samples, unless or until $1/f$ noise dominates. Fowler sampling depends on duty cycle and is best when sampling the pedestal and signal levels each for 1/3 of the total observing time: a 2/3 duty cycle. Under these conditions Fowler sampling is approximately 6% inferior to line-fitting. For background-limited performance the difficulty is that for any non-destructive multiple-sampling scheme, successive signal measurements are correlated in their noise.

11.6 INFRARED INSTRUMENTS

11.6.1 General issues

Every class of visible light instrument has its infrared counterpart. Building the IR version is significantly more challenging for several reasons, not least of which is the fact that everything, not just the chip, must be reduced to cryogenic temperatures. Many of the more robust IR optical materials (e.g., zinc sulfide, zinc selenide) don't transmit well in the visible, which hampers alignment and setup. On the other hand, crystalline materials like calcium fluoride and barium fluoride which do transmit both optical and IR light are fragile and harder to work optically. Cryogenic refractive indexes are also needed. Elimination of diffusely scattered light using blackened baffles requires care because anything truly black has almost 100% emissivity and will therefore be a strong infrared emitter unless very cold. There is a rule which says *if it's black it must be cold, if it's white then make it gold.* Black anodized aluminum is

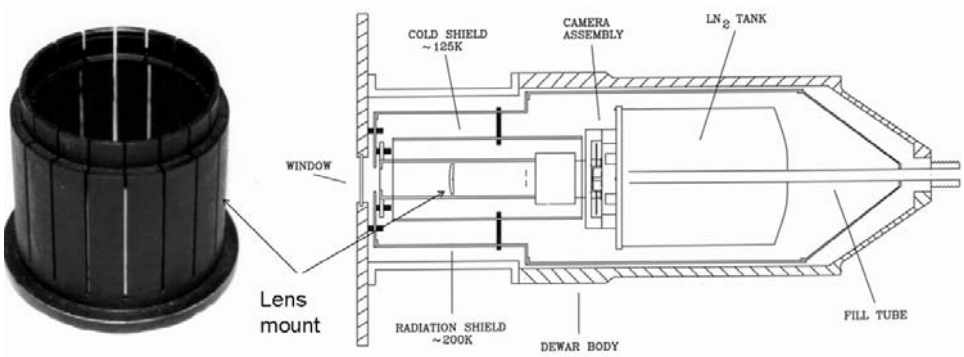


Figure 11.14. A lens-holder design made in aluminum but with slits cut into the barrel to allow the tube to relax outwards as it contracts around a lens. Right: a simple LN_2-cooled cryogenic IR camera.

not adequate. Special infrared black paints such as Parsons black, Aeroglaze, and Nextel are required. Care is required in applying these paints, otherwise they will eventually flake off due to cryogenic cycling. All dimensions will change during cool-down of the instrument and, worse, parts not made from the same materials will shrink by different amounts due to dissimilar coefficients of expansion. Lens holders could crush their optical components, optical separations will change, and materials may experience stress. All these things must be calculated beforehand, and each component must be constructed in such a way as to achieve the correct dimensions after it is cold. One way to avoid damaging a lens is to use a holder which is "springy". This can be achieved by cutting slits in the aluminum barrel as shown in Figure 11.14, which also shows a simple IR camera with the camera lens inside the cryostat. A pupil image is an ideal location to place a circular hole, called a "cold stop". By matching the size of the cold stop to the size of the pupil image, all unwanted off-axis rays from warm structure in the telescope are eliminated. The best cold stop is a "Lyot stop" which has a central disk (of metal) with four tiny supports designed to mask scattered light and thermal emission from the secondary mirror. Cold filters to define the wavelength interval are placed close to the pupil image, preferably in a collimated beam. Cooling is performed with either liquid cryogens or closed-cycle refrigerators.

11.6.2 IR cameras

Infrared cameras are now widespread. Short-wave mer-cad-tel with a cutoff at 2.5 μm is very popular, and LN_2-cooled cameras with 256×256 pixels or $1{,}024 \times 1{,}024$ pixels are easily operated with CCD controllers. Just like CCDs, larger format cameras use multiple arrays. One of the first of these was a mosaic of four $2{,}048 \times 2{,}048$ HgCdTe (H2-RG) arrays from Teledyne in the University of Hawaii's ULBCam used on their 2.2 m (88-inch) telescope on Mauna Kea. Developed by Don Hall and colleagues, this 16 Mpxl camera uses detectors being developed for the JWST near-infrared camera

Figure 11.15. The Wide-Field Camera (WFCAM) for the 3.8 m UKIRT undergoing tests in a lab prior to delivery. This large instrument has four $2\,\mathrm{K} \times 2\,\mathrm{K}$ HgCdTe arrays. Credit: ROE/ATC.

(NIRCam). Another similar arrangement is HAWK-I, which is a powerful new camera for the Nasmyth focus of UT4, one of the four 8 m VLTs at ESO in Chile. Using four $2\,\mathrm{K} \times 2\,\mathrm{K}$ HgCdTe arrays of a slightly older type (H2) from Teledyne to cover the 0.9 μm–2.5 μm region, HAWK-I has high-throughput optics and a field of view of $7.5' \times 7.5'$ with $0.1''$ pixels. Another camera with a mosaic of four $2\,\mathrm{K} \times 2\,\mathrm{K}$ HgCdTe Teledyne arrays is the Wide Field Camera (WFCAM) for the 3.8 m UKIRT, currently performing the deep survey known as UKIDSS. This enormous camera, shown in Figure 11.15, is physically very large due to the field of view and the plate scale of the larger telescope. It is located in a forward Cassegrain position and uses a fully enclosed cryogenic Schmidt camera, with corrector lens and large primary mirror to focus the beam from the UKIRT secondary mirror onto the detector array. WFCAM was delivered to UKIRT from ROE (Project Scientist Mark Casali now at ESO, and Project Manager David Lunney) in late 2004. Currently, the largest camera with sensitivity out to 5 μm is NEWFIRM, a New Extremely Wide-Field IR Mosaic for the NOAO 4 m telescope on Kitt Peak (Arizona) which employs four close-butted $2\,\mathrm{K} \times 2\,\mathrm{K}$ InSb (Orion) arrays from Raytheon.

11.6.3 Infrared spectrometers

Just like its optical counterpart, an infrared spectrometer relies on a diffraction grating, but unlike optical spectrographs, the grating and the entrance slit and all

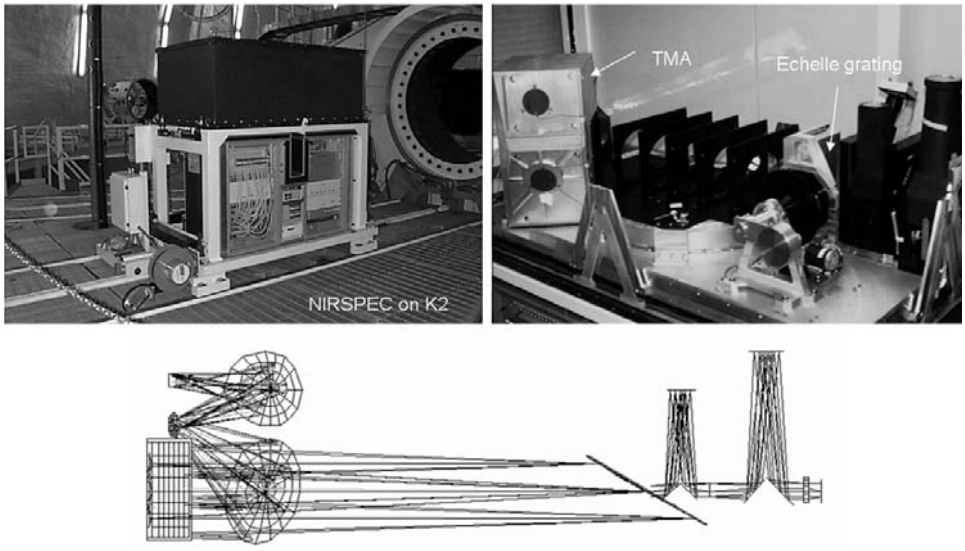


Figure 11.16. The optical layout inside the high-resolution, cross-dispersed infrared echelle spectrograph (NIRSPEC) developed by the author and James Graham (UCB) for Keck II.

surrounding optics and metal must be cooled to cryogenic temperatures. In the absence of a source on the slit, the spectrum will be dominated by the infrared OH emission lines from the night sky. Figure 11.16 shows the physical and optical layout of a high-resolution, cross-dispersed infrared echelle spectrograph for the Keck Telescope called NIRSPEC. This instrument uses a $1{,}024 \times 1{,}024$ InSb array from Raytheon, and is comparable in complexity with most large, cryogenic instruments being developed for large telescopes today. Taking advantage of the fact that NIRSPEC is stationary on the Nasmyth platform, so that the gravity vector is fixed, the optical design is laid out on a flat bench mounted on the lower plate of a rectangular vacuum enclosure. The other five sides form a shell that can be lowered over the optics to seal with an o-ring on the bottom plate. This makes the entire instrument easily accessible. Care must be taken to prevent thermal gradients from the bench to the top of tall components. There are two basic ways to achieve this: either by ensuring that the cross-sections of aluminum parts are large so that A/L is maximized, and/or attaching copper straps from the cold bench to the areas that need additional cooling. The entire optical path needs to be enclosed and baffled, and everything needs to be shrouded with a cold shield to prevent thermal photons from the warm walls reaching the detector. NIRSPEC remains the only cross-dispersed near-infrared (1 μm–5 μm) spectrometer on a large telescope. The CRIRES spectrometer on the VLT has even higher spectral resolution—but is not cross-dispersed. Instead, the long spectrum is spread across four $512 \times 1{,}024$ InSb arrays from Raytheon to provide a $512 \times 4{,}096$ spectrum. In the mid-infrared, MICHELLE is a 7 μm–26 μm imager and spectrometer for the Gemini-N 8 m telescope which employs a Raytheon Si : As IBC array with 320×240 pixels (PI: Alistair Glasse,

ROE/ATC), and the T-ReCS (Thermal-Region Camera Spectrograph) on Gemini-S uses the same type of array (PI: Charlie Telesco, University of Florida).

11.6.4 AO cameras and integral field spectroscopy

Many NIR cameras have now been used with adaptive optics (AO) systems, and results for targets such as the Galactic Center have already been mentioned. Recently, many integral field techniques for the IR have emerged and have been successfully demonstrated on telescopes (Figure 11.17). One such instrument for the Keck telescope, called OSIRIS, is optimized for AO. Developed at UCLA by James Larkin, OSIRIS employs an array of tiny lenses (called lenslets) in the focal plane of the telescope to dissect the image into hundreds of small spatial patches. The light from each patch is then collimated by the microlens so that a tiny image of the telescope entrance pupil (the primary mirror) is formed behind each lens. Each of these images is then dispersed by a grating spectrometer onto a $2\,\mathrm{K} \times 2\,\mathrm{K}$ HgCdTe (H2) detector. The image is a set of interleaved spectra, one for each patch of the sky dissected by the lenslet array. Thus, like image slicers and fiber-fed integral field instruments, OSIRIS obtains both spatial and spectral information simultaneously. As it is fed by the adaptive optics system of the Keck 10 m telescope, the spatial resolution is diffraction-limited at $<0.053''$, the spectral resolution is $R \sim 4{,}000$. Figure 11.17 also shows an image of Saturn's moon, Titan, obtained with OSIRIS. On the top is the raw

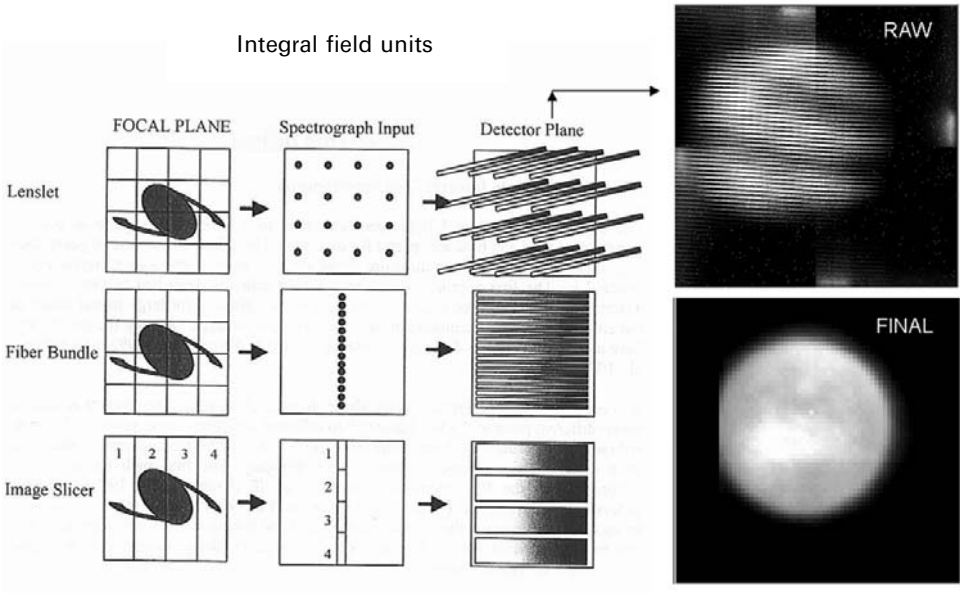


Figure 11.17. Integral field spectroscopy options, and before and after views of Titan obtained with OSIRIS, a lenslet-based AO infrared spectrograph on the Keck Telescope. Credit: OSIRIS Team.

image on the detector. Although the vague outline of an extended object is apparent, it is hard to recognize it because this image is really composed of the spectra of all the tiny spatial locations. A computer program extracts each spectrum and stacks it behind its spatial location in a data cube, rather like straws in a box. If you "extract" the straw then you have the spectrum represented by its length. However, if you cut the box at some intermediate position and view it end-on, then you "see" the image of the field at that wavelength. If you sum up all the signals along each straw then you "collapse" the spectrum into a broad-band image. Some of the images in the stack will be badly affected by the presence of strong OH lines at that wavelength/position. These noisy images can be removed, like removing a particular card from a deck of cards, and then the remaining images can be collapsed to form a deeper and less noisy image in which the strongest OH emission has been suppressed. By scanning through the K-band data cube of Titan, for example, one finds some wavelengths that penetrate through the smog layer to the ground, but other wavelengths correspond to methane absorption in the atmosphere and show only high-altitude haze.

11.7 THE IMPACT OF INFRARED ARRAYS

11.7.1 Ground-based observing

By 1994, near-infrared detector technology had made such huge advances that the only existing near-infrared sky survey, the TMSS by Neugebauer and Leighton (1969), no longer served as a useful basis for interpreting observations or selecting sources. A new and deeper survey was needed. With array detectors the entire sky could be covered with 50,000 times the sensitivity of the original survey and with a resolution of 2 arcsec—rather than 2 arcmin. This new survey was called the Two Micron All Sky Survey (2MASS), and it was led by Mike Strutskie for the University of Massachussetts (now at University of Virginia) with contributions from many other groups. Two matching 1.3 m telescopes were placed in each hemisphere, one on Mt. Hopkins in Arizona and the other on Cerro Tololo, in Chile. The northern hemisphere 2MASS facility began routine survey observations in June 1997, and the southern instrument began routine survey operations in March 1998. The 2MASS cameras included three HgCdTe (2.5 μm cutoff) infrared arrays with 256×256 pixels, one for each of the three bands J (1.25 μm), H (1.65 μm), and K_s (2.17 μm). A form of time delay integration was used by means of a tilting secondary mirror to freeze a frame during the sky-scanning process. With each source getting an integration time of 7.8 s, and mapping at a rate of 70 square degrees per band per night, the survey reached limiting magnitudes, meaning a signal-to-noise ratio of 10 ($\sim\pm0.1$ magnitudes), of at least $J = 15.8$, $H = 15.1$, and $K_s = 14.3$. Operations ended in 2001 and the 2MASS All-Sky Catalog was released in 2003 by the Infrared Processing and Analysis Center (IPAC) in Pasadena, California. Data products include a digital atlas of the sky comprising approximately 4 million 8×16 arcmin atlas images, having about 4 arcsec spatial resolution in each of the three wavelength bands; a point source catalog containing accurate positions and fluxes for ~300 million stars and other

unresolved objects; an extended source catalog containing positions and total magnitudes for more than 1,000,000 galaxies and other nebulae. Another survey also enabled by infrared array technology is DENIS, which is a deep near-infrared survey of the southern sky in two near-infrared bands (J at 1.25 μm and K at 2.16 μm) and one optical band (I at 0.82 μm) simultaneously. This survey was conducted by a European consortium, using a 1 m telescope at the ESO facility at La Silla, Chile. DENIS started in 1996 and operations at La Silla were completed in 2001. The third data release provided 365 million sources. The limiting magnitudes of the DENIS survey are: Gunn-$i = 18.5$, $J = 16.5$ and $K_s = 14.0$, respectively.

A decade later, the successor to 2MASS is already under way. UKIDSS, the UKIRT Infrared Deep Sky Survey, began in May 2005 and will survey 7,500 square degrees of the northern sky, extending over both high and low Galactic latitudes, in JHK to $K = 18.3$. This depth is at least three magnitudes deeper than 2MASS. UKIDSS will be the true near-infrared counterpart to the Sloan Digital Sky Survey, and will also produce a panoramic atlas of the Galactic plane. UKIDSS is made up of five separate surveys and includes two deep extra-Galactic elements, one covering 35 square degrees to $K = 21$, and the other reaching $K = 23$ over 0.77 square degrees. The survey instrument is WFCAM (see Figure 11.15), and the pixel scale of 0.4 arcsec gives an exposed solid angle of 0.21 square degrees. Data Release 1 was announced in January 2008 comprising about one-fifth of the final data volume when the survey completes in 2012. WFCAM has taken 30 times the amount of data taken in the entire 25-year history of the telescope before its arrival. Having started the ball rolling with the first infrared images from UKIRT back in 1986, I am delighted by this remarkable achievement.

A next-generation southern-hemisphere survey is also being developed. When it becomes available in 2008, VISTA will be a 4 m wide-field survey telescope, equipped with a near-infrared camera (1.65-degree diameter field of view) containing 67 million pixels of mean size 0.34 arcsec and broad-band filters at Z, Y, J, H, K_s and a narrow band filter at 1.18 micron. The VISTA camera, developed at Rutherford Appleton Labs and the Astronomy Technology Center (U.K.), contains 16 HgCdTe VIRGO detectors from Raytheon Vision Systems and will be the largest IR mosaic thus far (Figure 11.18). The telescope has an az-alt mount and quasi-Ritchey–Chrétien optics with a fast $f/1$ primary mirror giving an $f/3.25$ focus to the instrument at the Cassegrain focus. The VISTA camera includes a wide-field corrector lens system (three Infrasil lenses), autoguider, and active optics sensors. The telescope is located at ESO's Cerro Paranal Observatory in Chile about 1,500 m from the four VLTs.

11.7.2 The Stratospheric Observatory for Infrared Astronomy

Airborne astronomy will get a boost when the Stratospheric Observatory for Infrared Astronomy (SOFIA) begins preliminary operations in 2009. SOFIA's telescope consists of a parabolic 2.7 m primary mirror and a hyperbolic (chopping) secondary mirror in a bent Cassegrain configuration with two Nasmyth foci, a nominal IR focus, and an additional visible light focus for guiding. The telescope weighs almost 19 tons and looks through a door that is 4.8 m (16 ft) high cut into the fuselage. A flat

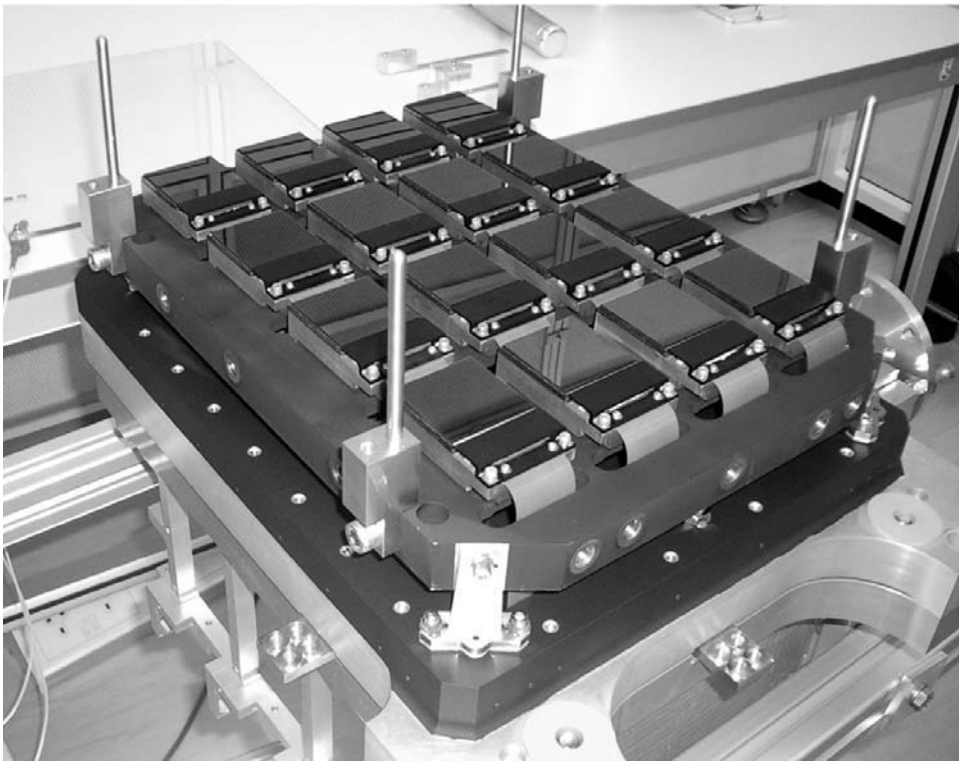


Figure 11.18. An array of 16 2 K × 2 K HgCdTe (Raytheon) detectors for the VISTA telescope on Cerro Paranal (Chile). Credit: STFC.

tertiary mirror reflects the IR beam into the infrared Nasmyth focus, which lies 0.30 m behind the instrument flange inside the pressurized main cabin on the other side of a large bulkhead. If the fully reflecting tertiary is replaced with a dichroic mirror, the transmitted optical light is reflected by a second tertiary ~289 mm behind the dichroic and sent to the visible Nasmyth focus. There it is fed into the Focal Plane Imager (FPI), an optical guiding camera system. Independent of the FPI there are two other imaging and guiding cameras available: the Wide Field Imager (WFI) and the Fine Field Imager (FFI). Both of these cameras are attached to the front ring of the telescope. Several interchangeable science instruments will be available for use on different flights. More information is available on the SOFIA web pages. The instruments are

— FORCAST (T. Herter, Cornell): Faint Object infraRed CAmera for the SOFIA Telescope (5 μm–40 μm)
— GREAT (R. Guesten, MPIfR): German REceiver for Astronomy at Terahertz frequencies (60 μm–200 μm)

— HAWC (D.A. Harper): High-resolution Airborne Wideband Camera (50 μm–240 μm)
— FIFI-LS (A. Poglitsch): Field Imaging Far-Infrared Line Spectrometer (42 μm–210 μm)
— EXES (J. Lacy): Echelon-Cross-Echelle Spectrograph (5 μm–28 μm)
— CASIMIR: CAltech Submillimeter Interstellar Medium Investigations Receiver (250 μm–600 μm)
— FLITECAM (I. McLean): First Light Infrared Test Experiment CAMera (1 μm–5 μm)
— HIPO: High-speed Imaging Photometer for Occultation (0.3 μm–1.1 μm)
— SAFIRE (H. Moseley): Sub-millimeter And Far InfraRed Experiment (145 μm–655 μm)

11.7.3 IR astronomy in space

In 1997 the NICMOS (Near-Infrared Camera and Multi-Object Spectrograph) replaced the Faint Object Spectrograph (FOS) during the second Hubble servicing mission. Rodger Thompson of the University of Arizona is the Principal Investigator of the NICMOS collaboration. The instrument has three adjacent—but not contiguous—cameras, designed to operate independently and simultaneously. Initially designed to use solid nitrogen as a cryogen, the dewar was built with a design lifetime of 4.5 ± 0.5 years. A thermal short developed after installation which caused Camera 3 to be no longer parfocal with the other two cameras, and the lifetime was shortened. NICMOS was unavailable for science from January 1999 to March 2002, but during the third Hubble servicing mission a new NICMOS Cooling System (NCS) was connected to the bayonet fittings of the NICMOS interface plate. This allowed the NCS to circulate cryogenic neon gas through the cooling coils in the existing dewar, thus providing the cooling power to bring temperature down to 75 K–86 K. The NCS consists of a cryocooler which provides the mechanical cooling; a capillary-pumped loop (CPL) which transports the heat dissipated by the cryocooler to an external radiator; and a circulator loop which transports heat from the inside of the NICMOS dewar to the cryocooler via a heat exchanger. NICMOS employs three low-noise, high-QE, 256×256-pixel HgCdTe (2.5 μm) arrays from Teledyne that were specially developed for this project, and thus the detectors are known by the instrument's name. Because the HST primary mirror is maintained at a temperature of $\sim 20^{\circ}$C to preserve its accurate shape, thermal background emission limits the use of Hubble in the infrared.

The "great observatory" counterpart to the Hubble Telescope for the infrared is NASA's Spitzer Space Telescope (formerly SIRTF, the Space Infrared Telescope Facility). Named after Princeton Astrophysicist Lyman Spitzer Jr. (1914–1997), widely considered the father of space astronomy, this cryogenic observatory was launched on August 25, 2003. Spitzer consists of a 0.85 m light-weight $f/12$ beryllium mirror telescope cooled to about 5.5 K and three cryogenically cooled science instruments covering the wavelength range from 3 μm–180 μm: IRAC, IRS, and MIPS. Image resolution is set by diffraction beyond 6.5 μm. In addition to imaging and

photometry from 3 μm to 180 μm, Spitzer also has a spectrometer for 5 μm–40 μm and a spectrophotometer for 50 μm–100 μm. Instruments on Spitzer are summarized briefly below:

— *IRAC* (Principal Investigator Giovanni Fazio, Harvard) is a four-channel camera that provides simultaneous images at 3.6 μm, 4.5 μm, 5.8 μm, and 8 μm over a 5.12×5.12 arcmin field of view. Each of the four detector arrays in the camera are 256×256 pixels in size. IRAC uses two sets of detector arrays. The two short-wavelength channels use InSb arrays and the long-wavelength channels use Si : As arrays (both detector types were provided by Raytheon Vision Systems).
— *IRS* (Principal Investigator Jim Houck, Cornell University) has four separate modules: a low-resolution, short-wavelength mode covering 5.3 μm–14 μm; a high-resolution, short-wavelength mode covering 10 μm-19.5 μm; a low-resolution, long-wavelength mode for 14 μm-40 μm; and a high-resolution, long-wavelength mode for 19 μm-37 μm. Each module has its own entrance slit. The detectors are 128×128 BIB arrays of Si : As for the shorter wavelengths and Si : Sb for the longer wavelengths (these detectors were made by DRS Technologies).
— *MIPS* (Principal Investigator George Rieke, University of Arizona) has three detector arrays: a 128×128 Si : As array identical to the IRS arrays, operating at 24 μm and with a 5 arcmin field of view; a 32×32 Ge : Ga array for 70 μm, also with a field of view of 5 arcmin; and a 2×20 Ge : Ga array, mechanically stressed to extend its photoconductive response to ~200 μm, and with a field of view of 0.5×5 arcmin. A scan mirror is used to provide mapping. MIPS is cooled by superfluid liquid helium to a temperature of about 1.5 K (−457°F).

Launched in May 2006, AKARI is an infrared astronomical satellite from the Japan Aerospace Exploration Agency (JAXA). It was the first new infrared survey satellite since IRAS. AKARI has a 68.5 cm telescope cooled to 6 K, and observes in the wavelength range from 1.7 μm to 180 μm. Originally known as ASTRO-F, AKARI has two instruments: the FIS (Far-Infrared Surveyor) and the IRC (Infra-Red Camera). The FIS has two detectors using both normal and stressed Ge : Ga and four observational bands; the stressed devices receive a pressure equivalent to 40 kg–60 kg weight on each square millimeter. FIS has four bands with the following properties and detectors: 50 μm–80 μm, Ge : Ga 20×2 pixels, each 26.8 arcsec; 60 μm–110 μm, Ge : Ga 20×3 pixels, each 26.8 arcsec; 110 μm–180 μm stressed Ge : Ga 15×3 pixels, each 44.2 arcsec and 140 μm–180 μm, stressed Ge : Ga 15×2 pixels, each 44.2 arcsec. The IRC is composed of three independent camera systems: NIR covers 1.7 μm–5.5 μm with a 512×412 InSb array; MIR-S covers 5.8 μm–14.1 μm with a 256×256 Si : As array; and the MIR-L camera is assigned to longer mid-infrared wavelengths of 12.4 μm–26.5 μm using the same detector type. AKARI ran out of its onboard supply of liquid helium on August 26, 2007, after the successful completion of its mission.

The next all-sky infrared survey will be WISE, NASA's Wide-field Infrared Survey Explorer (Principal Investigator Edward Wright, University of California,

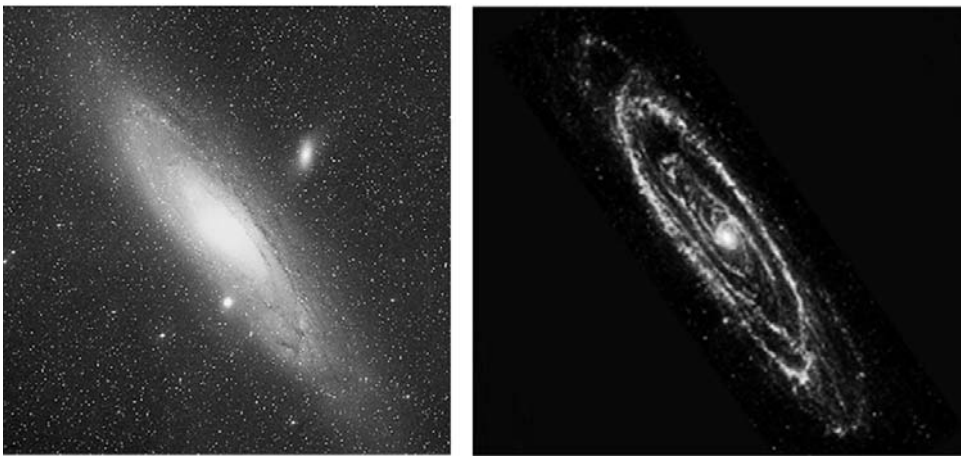


Figure 11.19. Two contrasting views of M31 (Andromeda galaxy) in visible and IR light. Credit: Spitzer Science Center

Los Angeles) which is expected to launch in late 2009. With a short 7-month cryogenic mission and large-array detectors, WISE will observe the entire sky in four bands: 3.3 μm, 4.7 μm, 12 μm, and 23 μm. The last two bands are like the IRAS bands, but WISE will be 500× more sensitive than IRAS. Using a 40 cm beryllium telescope and a scan mirror to stabilize the line of sight while the spacecraft continually scans the sky, WISE will integrate for 11 s during an exposure cycle and make eight or more exposures at each position for over more than 99% of the sky. The FOV = 47 arcmin wide and the pixel scale is 2.75″/pixel. HgCdTe and Si : As arrays of 1,024 × 1,024 pixels from Teledyne and DRS provide the cameras for the four channels.

Figure 11.19 shows the dramatic contrast between the Spitzer infrared image of M31 (Andromeda) and its well-known visual appearance. Color Plates 6, 10, 11, and 13 illustrate a few of the many remarkable results obtained with infrared arrays. Remember, when you look at these images, you are not looking at visible light. In many cases the objects are optically invisible! Normal blue/green/red colors are used to "translate" the infrared images into a visual representation. All of the data reduction and data-handling techniques developed for CCDs are immediately applicable. Thus, the impact of infrared array technology has indeed revolutionized this field.

11.8 SUMMARY

Infrared array detectors are solid-state imaging devices made from semiconductors with smaller bandgaps than silicon. Although similar in concept to CCDs in terms of the conversion of photons to electrons, they do not use an MOS unit cell to store charge and they do not use the charge-coupling principle to read out the electronic image. Infrared arrays are constructed in two steps. First, a two-dimensional grid of

infrared detectors is formed on a slab of IR-sensitive material, and then a matching grid of field effect transistors is produced on a slab of silicon, together with additional circuits to "address" each pixel. These pieces are mated by "bump-bonding" them together with tiny columns of indium. Thus, the functions of IR detection and multiplexing the signals are separated. Infrared arrays are at—or exceed—the 1-megapixel level for the range 1 μm–30 μm, and arrays of ~1,000 elements are already possible in the far infrared.

11.9 EXERCISES

1 Explain the terms "atmospheric window" and "thermal background" as they apply to ground-based infrared astronomy.

2 Use the Planck function to calculate the monochromatic flux from a black body at a temperature of 300 K. Convert this to photons at a wavelength of 2.2 microns.

3 Describe with the aid of diagrams how an infrared array is constructed and explain the basic principle of its operation. Assume that the detector is a photodiode of indium antimonide (InSb).

4 Explain the charge storage method used in infrared arrays and how there can be a source of non-linearity between photon flux and output voltage.

5 Why are CCDs not used as the silicon multiplexers for IR arrays?

6 What is meant by non-destructive readout?

7 Describe the construction and principle of the blocked-impurity-band (BIB) detector. Why is this device better than a photoconductor?

8 How is a gallium-doped germanium detector array constructed?

9 Compare and contrast the construction, operation, and performance of CCDs and IR arrays.

10 What is the source of the non-linearity in near-infrared array detectors?

11 Sketch the output signal as a function of time for an infrared array, and use the graph to illustrate what is meant by (a) single-sampling, (b) correlated double-sampling, and (c) multiple-sampling.

12 Describe the concept of up-the-ramp sampling.

11.10 REFERENCES

Alexander, D.H. (1980) Hybrid focal-plane array development, in *Infrared Image Sensor Technology: Proc. SPIE Meeting, Washington, April 8/9*, Vol. 225, pp. 2–9.

Allen, D.A. (1975) *Infrared: The New Astronomy*, Keith Reid, Shaldon, UK.

Arens, J.F.; Lamb, G.M.; and Peck, M.C. (1981) Infrared camera for 10 μm astronomy, *Proc. SPIE*, **280**, 61.

Arens, J.F.; Lamb, G.M.; and Peck, M.C. (1983) Infrared camera for ten micrometer astronomy, *Opt. Eng.*, **22**, 267–268.

Arens, J.F.; Lamb, G.M.; Peck, M.C.; Moseley, H.; Hoffmann, W.F.; Tresch-Fienberg, R.; and Fazio, G.G. (1984) High spatial resolution observations of NGC 7027 with a 10 micron array camera, *Astrophys. J.*, **279**, 685–693.

Baker, I.; Wilcock, S.; and King, R.E.J. (1981) Staring arrays for use in the 8–14 micrometer band, in *Advanced Infrared Detectors and Systems: Proc. IEE, London, October 1981*, pp. 76–81.

Becklin, E.E.; and Neugebauer, G. (1966), Observations of an infrared star in the Orion Nebula, *Astrophys. J.*, **147**, 799–802.

Becklin, E.E.; and Neugebauer, G. (1968) Infrared observations of the Galactic Center, *Astrophys. J.*, **151**, 145–161.

Chapman, R.; Beard, S.; Mountain, M.; Pettie, D.; and Pickup, A. (1990) Implementation of a charge integration system in a low background application, *Instrumentation in Astronomy VII: Proc. SPIE*, Vol. 1235, pp. 34–42.

Dereniak, E.L.; Brit, J.P.; Fowler, A.M.; Joyce, R.R.; Boreman, G.; and Ewing, W.S. (1984) Evaluation of a PtSi Schottky infrared CCD for astronomy, *Applied Optics*, **23**, 889–895.

Epchtein, N.; de Batz, B.; Capoani, L.; Chevallier, L.; Copet, E.; Fouqué, P.; Lacombe, P.; Le Bertre, T.; Pau, S.; Rouan, D. *et al.* (1997), The deep near-infrared southern sky survey (DENIS), in *ESO Messenger*, **87**, 27.

Fellgett, P. (1951) Theory of infrared sensitivities and its application to investigations of stellar radiation in the near-infrared, PhD thesis, University of Cambridge, Cambridge, U.K.

Forrest, W.J.; Moneti, A.; Woodward, C.E.; Pipher, J.L.; and Hoffman, A. (1985) The new near infrared array camera at the University of Rochester, *Pub. Astron. Soc. Pacific*, **97**, 183.

Fowler, A.M.; and Gatley, I. (1990) Demonstration of an algorithm for read-noise reduction in infrared arrays, *Astrophys. J. (Letters)*, **335**, L33–L34.

Garnett, J.D.; and Beletic, J.W. (eds.) (2004) *Optical and Infrared Detectors for Astronomy*, SPIE, Bellingham, WA [*Proc. SPIE*, **5499**].

Garnett, J.D.; and Forrest, W.J. (1993) Multiply sampled read limited and background limited noise performance, in A.M. Fowler (ed.), *Infrared Detectors and Instrumentation: Proc. SPIE*, Vol. 1946, pp. 395–404.

Glass, I.S.; Sekiguchi, K.; and Nakada, Y. (1995) An infrared camera based on a large PtSi array, in A.G.D. Philip, K.A. Janes, and A.R. Upgren (eds.), *New Developments in Array Technologies and Applications*, IAU Symposium No. 167, Kluwer Academic, Dordrecht, The Netherlands.

Gunapala, S.; Sarusi, G.; Park, J.; Lin, T-L.; and Levine B. (1994) Infrared detectors reach new lengths, *Physics World*, December issue, 35–40.

Hall, D.N.B.; Aikens, R.S.; Joyce, R.; and McCurnin, T.W. (1975) Johnson noise limited operation of photovoltaic InSb detectors, *Applied Optics*, **14**, 450–453.

Harwit, M. (1999) Neugebauer, Martz & Leighton's observations of extremely cool stars, *Astrophys. J.*, **525**, 1063–1064.

Harwit, M. (2001) The early days of infrared space astronomy, in J.A.M. Bleeker, J. Geiss, and M. Huber (eds.), *The Century of Space Science*, Kluwer Academic, Dordrecht, The Netherlands, pp. 301–330.

Hoffman, A.W.; Love, P.J.; and Rosbeck, J.P. (2004) Megapixel detector arrays: visible to 28 μm, *Proc. SPIE*, **5167**, 194–203.

Hyland, A.R. (1971) Galactic infrared astronomy, *Proc. Astron. Soc. Australia*, **2**, 14.

King, H.C. (1955) *The History of the Telescope*, Charles Griffin & Co. [republished by Dover in 1979].

Kleinmann, D.E.; and Low, F.J. (1967) Discovery of an infrared nebula in Orion, *Astrophys. J.*, **149**, L1–L4.

Kleinmann, D.E.; and Low, F.J. (1970) Observations of infrared galaxies, *Astrophys. J.*, **159**, L165–L172.

Koch, T.L.; Merilainen, C.A.; and Thom, R.D. (1981) *Advanced Indium Antimonide Monolithic Charge Coupled Infrared Imaging Arrays*, Santa Barbara Research Center Final Report, NASA, Washington, D.C. Available from STI.

Kosonocky, W.F.; Erhardt, H.G.; and Shallcross, F.V. (1981) *Optimization Study of 256-element IR-CCD*, RCA Labs Final Report, NASA, Washington, D.C. Available from STI.

Lord, S.D. (1992) *A New Software Tool for Computing Earth's Atmospheric Transmission of Near- and Far-infrared Radiation*, NASA Ames Research Center, Technical Memorandum 103957.

Low, F.J. (1961) Low-temperature germanium bolometer, *Journal of the Optical Society of America*, **51**, 1300–1304.

Low, F.J.; Rieke, G.H.; and Gehrz, R.D. (2007) The beginning of modern infrared astronomy—1960 to 1983, *Ann. Rev. Astr. & Astrophys.*, **45**, 43–75.

Maihara, T.; Iwamuro, F.; Yamashita, T.; Hall, D.N.B.; Cowie, L.L.; Tokunaga, A.T.; and Pickles, A.J. (1993) Observations of the Ott airglow emission, *Publ. Astron. Soc. Pacific*, **105**, 940.

McCaughrean, M.J. (1988) The astronomical application of infrared array detectors, PhD thesis, University of Edinburgh, U.K.

McCreight, C.R. (1981) Development and evaluation of integrated infrared arrays for astronomical applications, *Proc. SPIE*, **304**, 187–195.

McLean, I.S. (1987a) Results with the UKIRT infrared camera, in C.G. Wynn-Williams and E.E. Becklin (eds.), *Infrared Astronomy with Arrays*, Institute for Astronomy, University of Hawaii, Honolulu.

McLean, I.S. (1987b) Results with the UKIRT infrared camera, in *Infrared Sensors and Sensor Fusion: Proc. SPIE*, Vol. 782, pp. 138–141.

McLean, I.S. (1988) Infrared astronomy's new image, *Sky & Telescope*, **75**, 254–258.

McLean, I.S. (ed.) (1994) *Infrared Astronomy with Arrays: The Next Generation*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

McLean, I.S. (1995) Infrared arrays: The next generation, *Sky & Telescope*, **89**, 18–24.

McLean, I.S.; and Wade, R. (1984) Infrared array detectors for the United Kingdom Infrared Telescope (UKIRT), in *Instrumentation in Astronomy V: Proc. SPIE, London, 1983*, pp. 158–161.

Neugebauer, G.; and Leighton, R.B. (1969) *The Two-Micron Sky Survey*, NASA SP-3047, NASA, Washington, D.C.

Neugebauer, G.; Martz, D.E.; and Leighton, R.B. (1965) Observations of extremely cool stars, *Astrophys. J.*, **142**, 399–401.

Niblack, C.A. (1985) Improved performance characteristics for indium antimonide photovoltaic detector arrays using a FET-switched multiplexing technique, in *Solid State Imaging Arrays: Proc. SPIE*, Vol. 570, pp. 127–136.

Rayner, J.T. (1988) Aspects of the development of a 1–5 micron infrared camera for astronomy, PhD thesis, University of Edinburgh, U.K.

Richards, P.L.; and McCreight, C.R. (2005) Infrared detectors for astrophysics, *Physics Today*, **58**, 41–47.

Rieke, G.H. (2003) *Detection of Light from the Ultraviolet to the Submillimeter*, second edition, Cambridge University Press, Cambridge, U.K.

Rieke, G.H. (2007) Infrared detector arrays for astronomy, *Ann. Rev. Astron. & Astrophys.*, **45**, 77–115.

Rode, J.P.; Blackwell, J.D.; Linebarger, R.S.; and Seib, D.H. (1981) Staring hybrid arrays for IR imaging, in *Advanced Infrared Detectors and Systems: Proc. IEE, London, October 1981*, pp. 65–69.

Sibille, F.; Chelli, A.; Lena, P.; and Stefanovitch, D. (1982) Two-dimensional infrared speckle interferometry with a 32×32 InSb charge-injection device (CID) array, in *Instrumentation in Astronomy IV: Proc. SPIE*, Vol. 331, pp. 26–28.

Skrutskie, M.F.; Schneider, S.E.; Stiening, R.; Strom, S.E.; Weinberg, M.D.; Beichman, C.; Chester, T.; Cutri, R.; Lonsdale, C.; Elias, J.; *et al.* (1997) The Two Micron All Sky Survey (2MASS): Overview and status, in F. Garzón *et al.* (eds.), *The Impact of Large Scale Near-IR Sky Surveys*, Kluwer, Dordrecht, The Netherlands, p. 25.

Vural, K (1994) The future of large format HgCdTe arrays for astronomy, in I.S. McLean (ed.), *Infrared Astronomy with Arrays: The Next Generation*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Wynn-Williams, C.G.; and Becklin, E.E. (eds.) (1987) *Infrared Astronomy with Arrays*, The Institute for Astronomy, University of Hawaii, Honolulu.

Young, E.T.; Davis, J.T.; Thompson, C.L.; Rieke, G.H.; Rivlis, G.; Schnurr, R.; Cadien, J.; Davidson, L.; Winters, G.S.; and Kormos, K.A. (1998), Far-infrared imaging array for SIRTF, *Proc. SPIE*, **3354**, 57–65.

Web sites

ATRAN: *http://6747sp.sofia.usra.edu/cgi-bin/atran/atran.cgi*

2MASS: *http://www.ipac.caltech.edu/2mass/* and *http://pegasus.phast.umass.edu/index.html*

DENIS: *http://cdsweb.u-strasbg.fr/denis.html*

Home page for SOFIA: *http://www.sofia.usra.edu/*

Home page for Spitzer Space Telescope: *http://www.spitzer.caltech.edu/*

Home page for UKIDSS: *http://www.ukidss.org/*

Infrared Astronomy Timeline: *http://coolcosmos.ipac.caltech.edu/cosmic_classroom/timeline/timeline_onepage.html* [Caution: this timeline does not properly detail the introduction of infrared array detectors.]

12

Electronic imaging at ultraviolet, X-ray, and gamma-ray wavelengths

Astronomy in the ultraviolet, X-ray, and gamma-ray parts of the spectrum can only be carried out from above the Earth's atmosphere using satellites or rockets. In the UV and X-ray regimes we again come across the CCD, but other important electronic imaging technologies are also used for reasons that will be explained. High-energy photons with extremely small (sub-atomic) wavelengths require different kinds of detectors, and even the telescope design must change. Within the scope of this text, we can only illustrate some of the most important innovations and advances that have made imaging possible in these exciting fields.

12.1 INTRODUCTION

In terms of photon wavelengths, the ultraviolet region spans roughly 300 nm to ~10 nm, X-rays from ~10 nm to ~0.01 nm, and gamma rays the regime below that down to and below nuclear dimensions (one-millionth of 1 nm). This also represents an enormous range in photon energies, more than a factor of a billion. As described in Chapter 2, photons with wavelengths shorter than 300 nm (3,000 Å) are not transmitted by the Earth's atmosphere. Consequently, observations in the ultraviolet (UV), X-ray, and gamma-ray regimes must be carried out from space. If normal optical telescopes with parabolic mirrors can be used then the diffraction-limited formula for angular resolution with a circular aperture ($\theta \sim \lambda/D$) suggests that ultraviolet observations should have better resolution than visible light images for a given telescope aperture because of the smaller wavelength. However, the surface quality of the mirror becomes very important as the wavelength is decreased. Recall that the Strehl Ratio is related to the rms amplitude of the surface roughness by $S = \exp[-(4\pi\sigma/\lambda)^2]$, and a $\sigma = \lambda/20$ surface will scatter 33% of the light out of the diffraction spot. At 500 nm this corresponds to a 25 nm rms surface smoothness, but at 100 nm in the UV it would imply a surface smooth to 5 nm. The Hubble Space

Telescope is a conventional Ritchey–Chrétien telescope possessing one of the smoothest primary mirrors ever polished with a surface roughness of ~2 nm–3 nm which allows it to perform into the UV. Clearly, for even smaller wavelengths, achieving diffraction-limited performance from reflecting telescopes becomes challenging. But there is another problem too. In general, highly polished metallic surfaces have a higher reflectance than dielectric materials like glass. At normal incidence, silver and aluminum reflect over 90% of all visible light, which is why metallic coatings are applied to glass telescope mirrors. The amount reflected increases to 100% at grazing angles of incidence. Metallic reflection is a function of wavelength. For example, silver (Ag) has a strong minimum reflectance (<10%) around 320 nm. The optical properties of a dielectric are specified by its refractive index (n), but for a metal we need another property, the absorption index ($\kappa_0 = a\lambda/4\pi$) to measure the attenuation of photons due to interactions with the high density of free electrons in metals. In this expression a is the normal absorption coefficient (e.g., units of μm^{-1}) used in the standard exponential law of absorption $I = I_0\, e^{-ax}$, and thus the intensity drops to $1/e^{4\pi\kappa_0}$ in going the distance $x = \lambda$ into the medium. The complex refractive index is described by $n' = n - i\kappa_0$, where $i = \sqrt{(-1)}$, and the ratio of reflected-to-incident intensities is given by

$$r = [(n-1)^2 + \kappa_0^2]/[(n+1)^2 + \kappa_0^2] \tag{12.1}$$

which in the absence of absorption ($\kappa_0 = a = 0$) reduces to the well-known relationship for dielectrics. In practice n' can be expressed as $1 - c - id$ at very short wavelengths, where c and d are values that depend on the wavelength and the material. Extreme UV photons ($\lambda < 90$ nm) and X-rays ($\lambda < 10$ nm) are either completely absorbed (large a) or pass right through the mirror material (small a, low cross-section for interactions) at normal incidence. Fortunately, these photons will reflect off the surface of certain materials if grazing incidence angles are used due to the fact that the refractive index of metals at short wavelengths is less than 1 ($c > 1$ in expression for n'), which creates a situation similar to light going from a denser medium (glass, water) into air and leads to a critical angle for total internal reflection. In this case, the incident photons suffer total *external* reflection at some critical angle. An approximate empirical formula for the critical angle expressed in minutes of arc is

$$\theta_{\text{crit}} = 2.2(\sqrt{\rho})/E \tag{12.2}$$

where ρ is the density of the mirror material in kg/m^3; and $E = hc/\lambda = 1.24\,\text{keV}/\lambda$ (nm) is the photon energy expressed in kilo-electronvolts (keV); $1\,\text{eV} = 1.602 \times 10^{-19}$ J. For energies of a few kilo-electronvolts the critical angle is of order 1° for materials like gold and iridium.

12.1.1 Grazing incidence telescopes

Employing grazing incidence implies nearly parallel orientation of the reflecting surfaces, which leads to very long focal length systems. The simplest focusing system is the parabolic dish used extensively from the near-ultraviolet to the radio to eliminate spherical aberration, but a single parabola experiences coma for off-axis

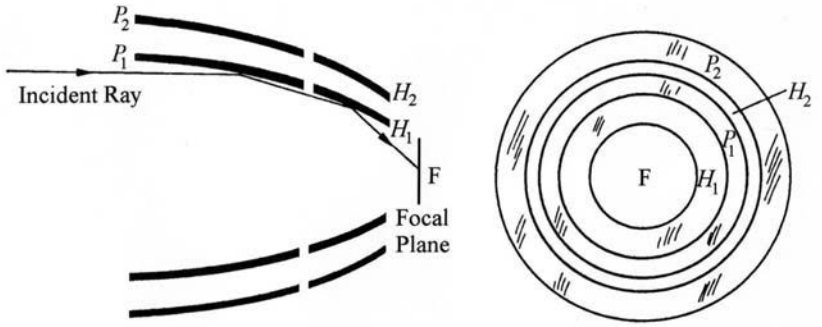


Figure 12.1. The basic concept of a Wolter-1 grazing incidence X-ray telescope.

rays. Abbe's sine condition requires that, to be free of coma, we need $r = h/\sin\theta$, where θ is the angle of incidence, h is the height of the ray from the optical axis, and r is the radius of a circle around the focal point. If the radius of curvature is very large, as it is for a typical optical telescope, then coma is minimized. But X-ray telescopes require grazing incidence and so cannot have large radii of curvature; the parabola must be highly curved. This problem was solved in 1952 by the German scientist Hans Wolter (1911–1978) while working on the development of an X-ray microscope. He used a coaxial hyperboloid as a secondary mirror which was properly confocal with the initial paraboloid, just as is done at optical wavelengths, but in this case both mirrors are used at grazing incidence (Figure 12.1). This arrangement very nearly meets the Abbe sine condition. The focal point is found from the relation

$$Y_0 = Z_0 \tan(4\theta) \tag{12.3}$$

where Y_0 is the distance from the optical axis to the point where the two conic surfaces intersect; and Z_0 is the distance along the optical axis from the focal point to the intersection plane of the two surfaces. The angle θ is the sum of the two grazing angles on each mirror. In practice, the curvatures and angles are much less than shown and the focal point is far behind the mirrors (Figure 12.2). Wolter described three different imaging configurations, now known as Wolter Types I, II, and III, but the design most commonly used by X-ray astronomers is the Type I because it has the simplest mechanical configuration and offers the possibility of nesting several reflecting surfaces inside one another, thereby increasing the useful collecting area. Wolter Type II uses a convex secondary hyperboloid and has a longer focal length but narrower field of view. In the Type III system the initial paraboloid is convex and the secondary is a concave ellipsoid. Almost all past space-borne X-ray and extreme UV missions have used stacked Wolter-I telescopes. By stacked we mean that the paraboloids and hyperboloids are replaced with conical off-axis sections of the solids of revolution, and many such sections are nested inside one another to increase the effective collecting area (Figure 12.2).

Example. For a grazing incidence angle of 0.86 degrees and an aperture height of 0.6 m the focal length is approximately 10 m. Thus, the plate scale is 206,265/10,000

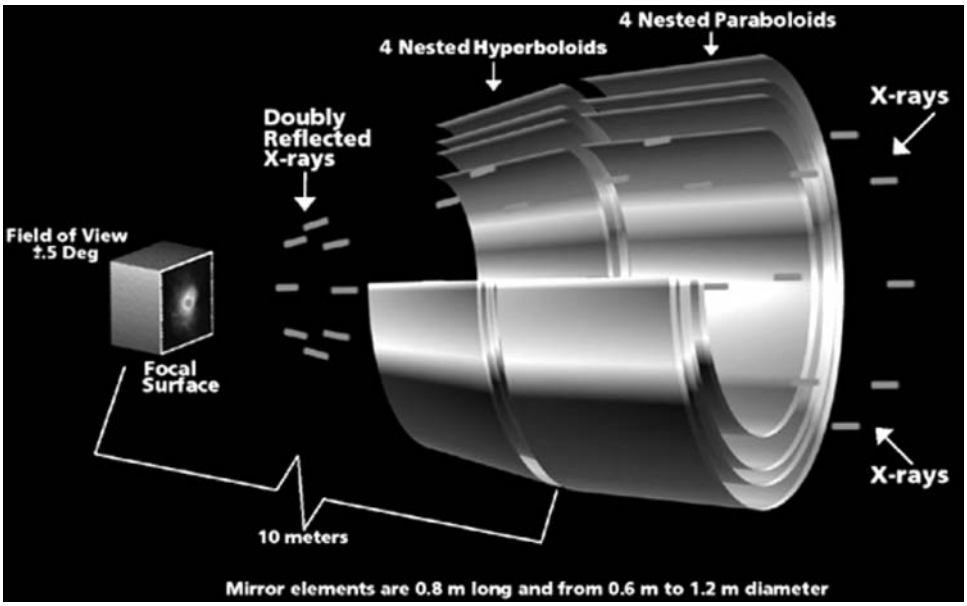


Figure 12.2. The arrangement of grazing incidence optics in the Chandra X-ray Telescope. Credit: NASA/CXC.

$= 20.63$ arcsec/mm. If we want to match this telescope directly to a CCD with 24 μm pixels, then 20.63×0.024 yields a pixel resolution of about 0.5 arcsec. With conical sections about 0.8 m long formed as four concentric stacks, this setup describes the Chandra X-ray Telescope.

NASA's Chandra X-ray Observatory has four stacked Zerodur reflectors with an outer mirror diameter of 120 cm and a geometrical sensitive area of 1,145 cm^2, corresponding to an aperture filling factor of only 0.1. In ESA's XMM-Newton observatory, the resolution is 5 arcsec using 3 modules of 58 stacked reflectors with an outer foil diameter of only 70 cm. Each module has a geometrical area of 1,750 cm^2 and an aperture filling factor of 0.45. The multi-foil approach gives more collecting area but the difference in resolution between Chandra and XMM is due to the difficulties in fabricating thin X-ray mirror shells with high optical performance.

Detectors of UV and X-rays are many and varied. In the UV, a major problem is the requirement for the rejection of long-wavelength photons, in other words visible light. There are essentially two classes of detectors that cover a wide range of photon energies: photo-emissive devices and solid-state devices. The latter category includes the silicon CCD. Among the photo-emissive devices are other kinds of panoramic detectors using photocathodes with large work functions like CsI and various methods of achieving pixel definition. Together with the change from classical to Wolter telescopes for the extreme UV and X-ray wavelengths, imaging is therefore possible in the UV and X-ray bands, but what about even higher energy photons?

12.1.2 Coded mask telescopes

High-energy X-ray and gamma-ray photons with sub-atomic sized wavelengths cannot be focused easily because they are too penetrating, but some degree of directionality can be imparted using a "collimator" to restrict the angle of acceptance. The simplest kind of collimator is a honeycomb of long, closely packed tubes. The angular field of view is determined by the width divided by half the length of the tube ($w/2L$). Of course, the walls have to be of dense enough material to stop the high-energy photon from reaching the detector by passing through other tube walls. A variation of this approach is the "lobster eye" collimator which is a honeycomb collimator curved into a spherical shape to help direct rays to a common location. An alternative approach used on recent missions including ESA's INTEGRAL satellite and NASA's SWIFT spacecraft is called the "coded mask" telescope. Figure 12.3 shows the coded masks used on SWIFT and INTEGRAL.

A generic coded mask telescope is sketched in Figure 12.4. Basically, the mask contains both transparent and opaque patches, and each detector pixel records the sum of the signals from a different combination of incident directions. The point source function of a coded mask telescope is not just a slightly blurred image at one location, like an Airy function for an optical telescope, but is in fact spread over the entire detector plane. However, all is not lost. The principle of the coded mask is as follows. Photons from a certain direction in the sky project the mask pattern (shadow) onto the detector; this projection has the same coding as the mask pattern, but is shifted relative to the central position by a distance that corresponds uniquely to the direction of the photons. A two-dimensional detector, which needs to be well-matched to the mask elements, accumulates the summation signals from a number of shifted mask patterns. Each shift encodes the position, and the signal strength encodes the intensity of the sky at that position. Clearly, each part of the detector array may receive photons from any position within the observed sky. After a certain illumination period, the accumulated detector image can be decoded to an image of

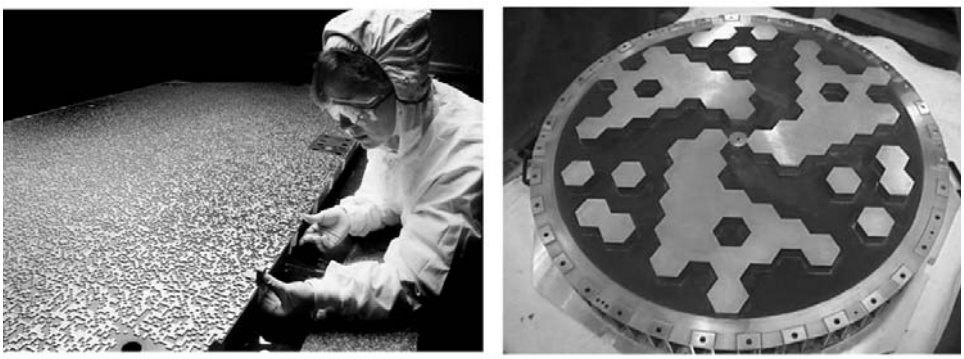


Figure 12.3. (Left) The random coded mask telescope for the SWIFT gamma-ray satellite; and (right) the cyclic coded mask for the INTEGRAL gamma-ray satellite. Credit: SWIFT and INTEGRAL Teams.

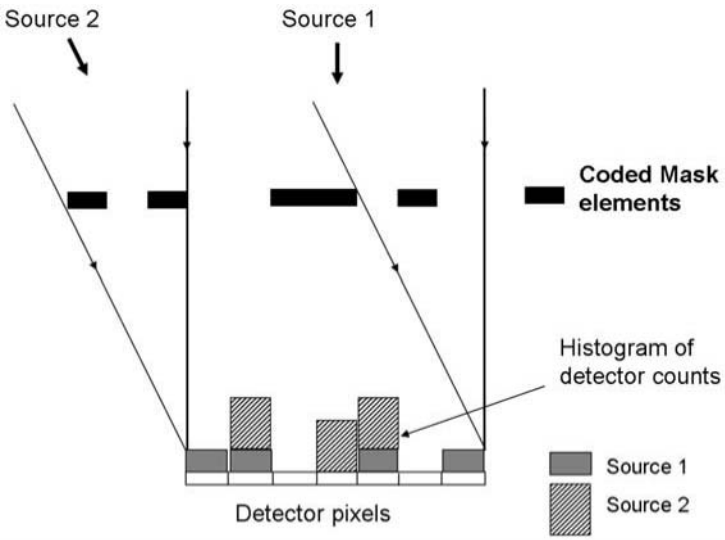


Figure 12.4. The basis of the coded mask technique.

the sky by determining the strength of every possible shifted mask pattern using an autocorrelation algorithm. Proper performance of a coded-mask camera requires that every sky position is encoded on the detector in a unique way. Stated in terms of the autocorrelation function (ACF) of the mask pattern, this means that the ACF should consist of a single peak and flat sidelobes (a delta function), which therefore puts constraints on the type of mask pattern and on the way its (displaced) projections are detected. An important difference compared with direct-imaging systems is the fact that Poisson ($\sqrt{N}$) noise from any source in the observed sky is, in principle, induced at any other position in the reconstructed image. Thus, the imaging quality of the camera is determined by the type of mask pattern, the spatial response of the detector, and the reconstruction method.

Two types of mask patterns were proposed initially: a pattern of Fresnel zones (Mertz and Young, 1961) and the random pinhole pattern (Dicke, 1968; Ables, 1968). A camera with a Fresnel zone plate has not yet been applied to extra-solar X-ray and gamma-ray astronomy, but the concept of the random pinhole pattern has. The random pinhole pattern is an extension of the pinhole camera. A pinhole camera has ideal imaging properties but delivers a poor signal-to-noise ratio because the sensitive area is severely restricted by the size of the pinhole. Sensitivity can be increased by enlarging the pinhole, but at the expense of angular resolution. However, the open area can be increased while still preserving angular resolution by using many duplicate pinholes at random. The random character of the pinholes is necessary in order to meet the condition that the auto-correlation function be as close to a delta function as possible. Non-random patterns also exist, based on "cyclic difference sets" or "uniformly redundant arrays", which approach this ideal. The angular resolution limit in seconds of arc is the quadratic sum ($\theta^2 = \theta_m^2 + \theta_d^2$) of the

mask element size term $\theta_m = 206(m/L)$ and the detector resolution element $\theta_d = 206(d/L)$ where the mask-to-detector separation L is given in meters and the detector pixel size and mask element size are expressed in millimeters; the angular resolution is degraded by $\cos^2 \varphi$ for off-axis sources. The image reconstruction algorithm must collect photons from all over the detector, but noise events will also be collected from all over the detector, so all the noise contributes to each pixel. On SWIFT the coded mask telescope has a mask size of $2{,}400 \times 1{,}200$ mm with 52,000 lead (Pb) elements each $5 \times 5 \times 2$ mm in a random pattern, and the separation length results in a resolution of about 17 minutes of arc. INTEGRAL uses a cyclic difference mask 1,200 mm in diameter with 72 large tungsten (W) elements each $60 \times 60 \times 50$ mm in size. Representing the mask with an array M of 1 (open) and O (opaque) elements, the detector array D will be given by the convolution of the sky image S by M, plus an un-modulated background array term B; thus, $D = S * M + B$. We need to find a special array M for which there exists a correlation inverse G such that $M * G = \delta$-function. In which case we have that

$$S' = D * G = S * M * G + B * G = S * \delta + B * G = S + B * G \tag{12.4}$$

S' differs from the real sky image S only by the term $B * G$, which for a flat array B is a constant term that can be measured and removed (Goldwurm *et al.*, 2001). The implication of this discussion about imaging telescopes even for high-energy photons is that there are corresponding "array" detectors—and there are. We will now pursue the detector technology by reviewing each wavelength regime consecutively.

12.2 ULTRAVIOLET DETECTORS AND INSTRUMENTS

For practical reasons associated with the technologies used, the UV is often subdivided into four regions; the Extreme UV (EUV) from ~5 nm to 90 nm (~50 Å to 900 Å), the Far UV (FUV) from 90 nm to 120 nm (~900 Å to 1,200 Å), the UV from 120 nm to 200 nm (1,200 Å to 2,000 Å), and the Near UV (NUV) from 200 nm to 300 nm (2,000 Å to 3,000 Å). Unlike the well-defined infrared windows, however, there are no sharp boundaries and sometimes three divisions are used instead: namely, the near-UV, mid-UV, and far-UV. Departures from visible light techniques become greater as the wavelength becomes shorter. For example, not many materials have good transmission in the UV (magnesium fluoride MgF_2 is one that does transmit down to ~115 nm); any contamination that settles on the optics will be opaque in the UV; it is hard to make UV filters that don't leak slightly at longer wavelengths; and, as we have seen, as the EUV is approached the design of telescopes and other optical systems must change to use grazing incidence angles.

While several ultraviolet satellites (in particular, Copernicus, TD-1, and ANS) laid the early groundwork, it was the launch of the highly successful IUE (International Ultraviolet Explorer) satellite in 1978 that really opened up this vast field to all astronomers. Originally suggested by Bob Wilson as early as 1964, this remarkable NASA/ESA/U.K. satellite was one of the longest running space operations (1978–1996). Located in geosynchronous orbit 36,000 km (22,700 miles) from Earth, IUE

carried a telescope with a diameter of 45 cm (18 inch) and was equipped with both high-dispersion and low-dispersion ultraviolet spectrographs covering the wavelength interval from 1,250 Å to 3,200 Å. With the launch of the Roentgen Satellite (ROSAT) in 1990 and the Extreme Ultraviolet Explorer (EUVE) in 1992, ultraviolet astronomy pushed its boundaries into the 100 Å–1,000 Å region. The EUVE satellite (1992–2001) supplied images in four wavelength regions across the whole EUV band and carried three EUV spectrometers. Its all-sky survey found 801 objects including the first extra-galactic detection at these wavelengths (Bowyer and Malina, 1994). The majority of the sources discovered by EUVE lay within a few hundred lightyears of the Sun, and included such hot, young luminous stars as Epsilon Canis Majoris, white dwarf stars, and cataclysmic variable stars like SS Cygni. Interstellar hydrogen atoms absorb EUV radiation so efficiently that if the density around the solar system was about 100 atoms per cubic centimeter then there would be enough absorption to limit our view to within 10 lightyears of the Sun. The fact that EUV stars were discovered at all implies that the density is much lower. The solar neighborhood lies in a low-density bubble, but EUVE's electronic imaging systems revealed that there are "tunnels" through the neutral gas in some directions.

Several UV experiments also utilized the Space Shuttle, such as WUPPE (the Wisconsin Ultraviolet PhotoPolarimeter Experiment) and ORFEUS (Orbiting and Retrievable Far and Extreme Ultraviolet Spectrometer), and, of course, the Hubble Space Telescope (HST) is itself a superb ultraviolet collector. Past UV-sensitive instruments on HST include FOC and GHRS and current instruments include STIS and ACS (see Chapter 2 for more on HST). After Service Mission 4 (SM4) the no-longer-needed COSTAR corrector will be replaced by the UV spectrometer COS (Cosmic Origins Spectrograph).

The Far Ultraviolet Spectroscopic Explorer (FUSE), which operated from 1999 to 2007 was another successful mission. FUSE observed from about 90 nm to 120 nm and employed several unique design features. Instead of a single mirror, four separate mirror segments (off-axis parabolas) were used, two of which were coated with silicon carbide to enhance reflectivity in the FUV, and the other two were coated with lithium fluoride over aluminum which performs better at longer UV wavelengths. Light from the four optical channels was dispersed by four spherical, aberration-corrected holographic diffraction gratings on a large Rowland circle with a resolving power of $\lambda/\Delta\lambda = 24{,}000$–$30{,}000$. Imaging devices using photocathodes as described below were used.

Historically, the emphasis of UV missions has been spectroscopy of point sources rather than imaging or studies of extended sources such as galaxies. Therefore, until 2003, there were only shallow all-sky surveys in the UV. GALEX (the Galaxy Evolution Explorer Mission) launched in April 2003 is an all-sky survey down to AB magnitudes $\sim$21 using a 50 cm (20-inch) telescope imaging in two broad UV bands centered at 150 nm and 230 nm. Understanding how galaxies were formed is the primary mission of GALEX, led by the California Institute of Technology (Martin *et al.*, 2005), but its state-of-the-art UV cameras have provided some remarkable images of all kinds of astronomical sources. On-orbit performance is summarized in the paper by Morrissey *et al.* (2005). Figure 12.5a (and Plate 5) shows

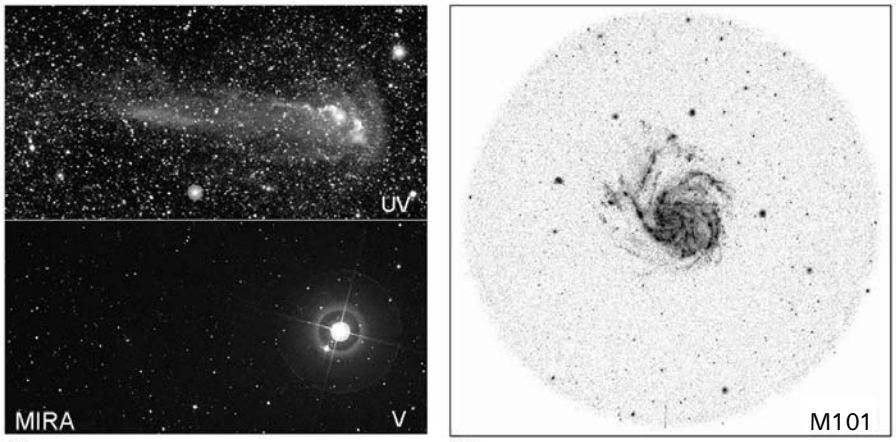


Figure 12.5. (Left) The upper panel is an ultraviolet image of the star Mira (*o* Ceti) that reveals a long tail not seen in visible light (lower panel); and (right) is the galaxy M101, illustrating GALEX's field of view. Credit: NASA/JPL-Caltech/C. Martin (Caltech)/M. Seibert (OCIW).

a remarkable "tail" behind the famous pulsating star Mira; Figure 12.5b is a "negative" image of M101 showing the detector's field of view.

Unlike the visible region, however, where the silicon CCD has become the unquestioned detector of choice, there are several kinds of UV detector systems. CCDs and other photoconductive devices are used, but so too are a wide range of photoemissive detectors with various photocathodes tailored to suit certain wavelength ranges. Photoemissive detectors are inherently good UV, FUV, and EUV detectors, but some form of intensification is almost always needed. Large electron gains are usually achieved using micro-channel plates which together with a variety of two-dimensional readout schemes provide pixel-like imaging performance. See papers in Siegmund (2007) for all UV, X-ray, and gamma-ray detectors and instruments, and Joseph (1995), Carruthers (1994), and Timothy (1991) for earlier accounts. I am indebted to John Vallerga (UC Berkeley) for a recent unpublished review of UV detectors in space.

One of the major concerns for UV work already mentioned is that the detector should be "solar-blind" (i.e., it should *not* be sensitive to visible light photons). The reason for this restriction is simply that many astronomical sources emit 10^4 to 10^8 visible photons for every UV photon in the 1,000 Å to 2,000 Å wavelength range, and if the detector actually has its maximum sensitivity at visible light wavelengths, then there will be an enormous background unless the visible light is heavily filtered out. Unfortunately, even blocking to reduce visible light by a factor of 10^5 still isn't sufficient in many cases and such filters (e.g., the alkali metal or Wood's filter) also absorb UV photons. A CCD with a 20% QE in the ultraviolet would end up with a detective quantum efficiency (DQE) of only 1%–4% when it was made solar-blind; the DQE takes into account all losses. Another issue arises from the fact that many

UV scenes have most of the light concentrated in a few pixels, and therefore the local dynamic range or contrast can be more important than the global dynamic range over the entire detector (i.e., the response to a flat-field).

Detectors which are operated cold, such as CCDs, will form traces of condensable material on their surfaces, even in the vacuum environment of space, which will destroy UV performance. Using a warm window into a vacuum-sealed CCD camera is an option, but only for UV wavelengths longer than 1,200 Å where suitable window materials exist, and even then the transmission is only about 60%–70%.

12.2.1 UV-sensitive CCDs

In principle, silicon charge-coupled devices are capable of responding to photons over a huge wavelength range, from 0.1 nm to 1,000 nm (1 Å–10,000 Å). Frontside-illuminated CCDs are almost completely insensitive to wavelengths less than 400 nm (4,000 Å), and even very thin (~15 μm) backside-illuminated CCDs can have severe UV sensitivity problems unless considerable effort and expense is devoted to backside surface control. The toughest wavelength for CCDs is around 2,500 Å because the absorption depth is only about 3 nm (30 Å), much less than the wavelength of the UV light, and so about 70% of the incident radiation is simply reflected! One solution to this problem is to apply a UV phosphor coating to the CCD. An early example of such a coating is the organic phosphor, Lumogen, which is also used in yellow "highlighting" pens. This material absorbs all ultraviolet photons with wavelengths shorter than 4,200 Å and re-emits the energy at yellow (longer) wavelengths around 5,200 Å. This process is called "down-converting". A layer of Lumogen about 6,000 Å thick can be thermally vacuum deposited onto the CCD and will yield about 15% quantum efficiency from 500 Å to 4,000 Å in wavelength; this is a very big improvement, but it is still only a 3% detected quantum efficiency (DQE) after the device is rendered "solar-blind". Lumogen is very inexpensive compared with backside treatments and can be applied to either the front or the backside of the CCD. Because it is transparent at longer wavelengths it causes no harm and actually improves the quantum efficiency a bit by acting as an anti-reflection coating. The original backside-illuminated TI 800×800 chips for the Hubble Space Telescope WFPC1 instrument were meant to be intrinsically UV-sensitive, but it turned out that they too needed a coating (Coronene). Other coatings are also possible, and some companies have proprietary coatings. While large-format CCDs are available with Lumogen coatings from companies like *e2v*, outgassing of the phosphor in sealed systems used in space without re-pumping is a potential problem, both because of contamination and because of changes in the coating's response. Intrinsic QEs of ~30% have been achieved with suitable coatings.

Alternative UV CCD technologies include the use of very thin gate structures (known as "thin-poly" gates) to minimize absorption or the use of virtual phase technology which provides an "open" phase with no overlying gate metal at all. This is a good approach which requires no post-fabrication treatment whatsoever. For thinned CCDs, one can apply a backside charge to drive the photogenerated electrons

towards the wells by either creating an electrostatic charge in the native oxide by soaking the CCD in a strong oxidizer such as nitric oxide gas, or by use of a thin conducting bias gate on top of a thin insulating layer, or by very shallow ion implants to "build in" an electric field.

Due to the considerable progress on thinning, coating, and QE-pinning of CCDs, and with readout noise around $1e^-$ rms, CCD detectors have become very important for UV and NUV astronomy, and several space missions include UV-sensitive CCDs. Around 2,000 Å, however, the problems for CCDs are formidable.

12.2.2 Microchannel plates

There are a variety of observing modes, such as high-dispersion spectroscopy in the violet and ultraviolet, for which the finite readout noise of currently available CCDs, the limited UV response, and the slower data access time of CCDs, makes them less satisfactory than photon-counting detectors (or PCDs) employing good UV photo-emissive photocathodes and some form of light intensification. One of the most widely used UV detectors is the microchannel plate (MCP), which is shown schematically in Figure 12.6. MCPs are used across most of the UV and have been the detector of choice for almost all major UV missions. For example, GALEX employs two MCPs fabricated by the experimental astrophysics group at the

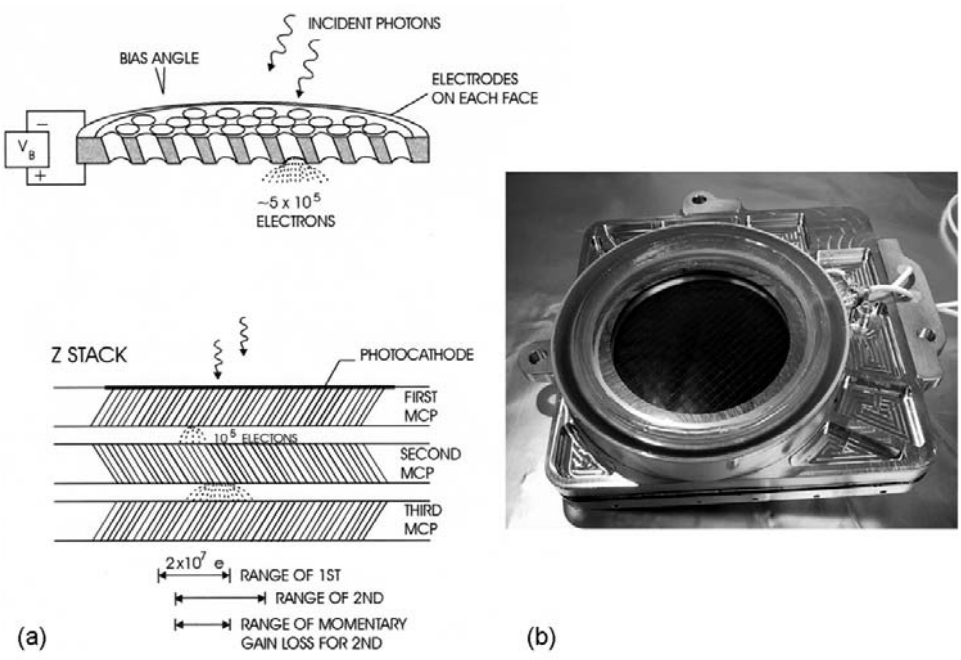


Figure 12.6. (a) Structure of a microchannel plate device; and (b) the GALEX MCP detector. Credit: Experimental Astrophysics Group, UC Berkeley.

University of California Berkeley Space Sciences Laboratory. This group has been responsible for the development of most of the UV detectors in space. For example, there were seven on EUVE, four in two different instruments on SOHO, two in FUSE, and a custom-designed MCP for COS (Hubble), to give just a partial list (Vallerga, 2007). An MCP is essentially a small, thin disk of lead oxide glass with numerous microscopic channels running parallel to each other from one face of the disk to the other. When an electric potential is applied between the two faces, the MCP becomes an image intensifier. Each tiny channel or pore acts like a photomultiplier tube since electrons hitting the walls eject additional electrons resulting in a cascade of electrons. For one incident photon a charge cloud of 500,000 electrons can emerge from the channel and can therefore be detected or counted. The length of the microchannel is typically 50–100 times the diameter of the channel, which implies a large surface-to-volume ratio and the tendency to trap residual gas unless exceptional measures on cleanliness and plate conditioning are employed. Because MCPs are operated at potentials of a few thousand volts, residual gases can lead to destructive discharges. The channels have diameters ranging from 5 μm to 25 μm on 10 μm–40 μm centers, and plates with active areas as large as 100×100 mm^2 are available; the GALEX detectors are 75 mm in diameter with an active area of 68 mm. Note that the response of the MCP is a strong function of the angle of incidence of the photons, which is not the case for a CCD detector. Photocathodes can be placed on the top face or on a window in proximity focus immediately above the MCP. Materials with large work functions such as CsI, CsTe, and KBr have good UV quantum efficiency but very low response to visible photons. More recently, gallium nitride (GaN), which has a bandgap of 3.4 eV, has been added to the list of photocathodes available for UV astronomy. The material can have a high QE and be effectively solar-blind with additional filters (Ulmer, 2006). Various anode structures or solid-state devices can be placed below the MCP to act as readouts. One possibility is to use a CCD to form a microchannel plate intensified CCD (ICCD). DQEs of 10%–20% are routinely obtained with MCPs.

Each conversion between light and electrons causes degradation in spatial resolution which can be overcome by employing photon counting and centroiding techniques. Photon-counting MCP detectors, which are inherently solar-blind, currently have a DQE advantage over direct CCDs at wavelengths shorter than 2,000 Å, but lack the dynamic range of a large-pixel CCD. The disadvantage of these photon-counting approaches is the framing rate of the CCD and the phosphor decay effects which can limit the local dynamic range to fewer than 5 counts/pixel/second. In the extreme UV this is not always a problem.

A large-format MCP intensified CCD for the X-ray Multi-Mirror Mission (XMM) is illustrated in Figure 12.7. Another line of development, pursued at the Goddard Space Flight Center, is to replace the CCD with a charge injection device (CID) because the addressable readout capability of the CID can be used to enhance the local dynamic range in selected regions according to the brightness distribution in a given image. CIDs are MOS unit cells, but instead of being charge-coupled, each cell is individually addressable and each has an additional storage well to permit non-destructive reads. This chip architecture results in high capacitance and high noise

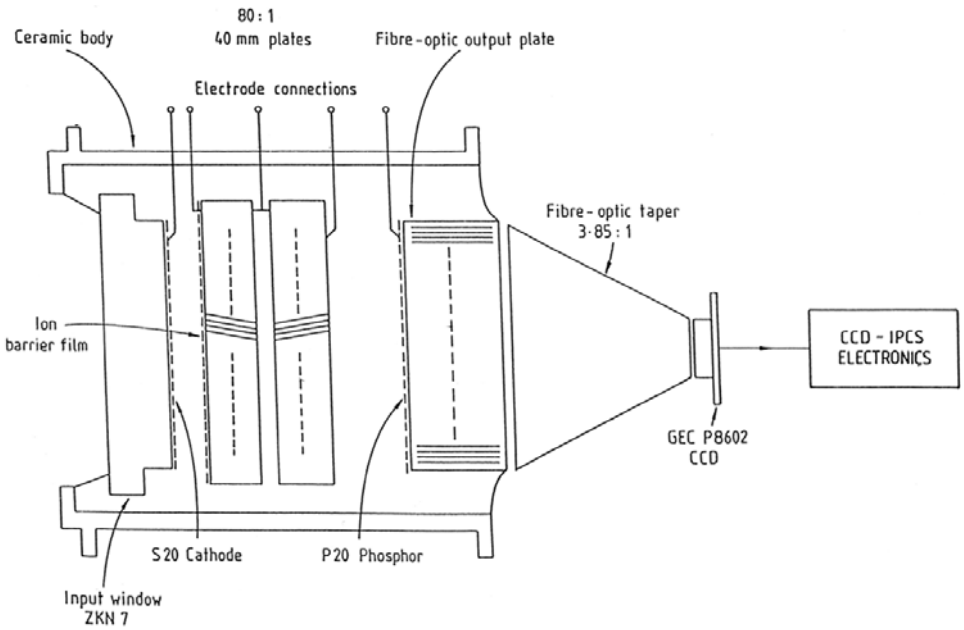


Figure 12.7. The microchannel plate intensified CCD (MIC) for XMM schematic. Credit: John Fordham.

compared with a CCD, but this disadvantage is eliminated when there is sufficient gain in the intensifier.

Microchannel plate detectors also use a variety of other anode structures. One of the simplest is a single resistive anode in which the location of the event is determined by the amount of charge or current "divided" between amplifiers attached to the corners. Other anode structures include the wedge and strip anode, the spiral anode, and the delay line, each of which are described as "continuous" anodes. It is also possible to utilize "discrete" anode structures at the expense of many more amplifiers and encode the event location through direct detection. One such system is called the Multi-Anode Microchannel Array (MAMA) and another is the capacitive readout system called Coded Anode Converter (CODACON).

The wedge and strip anode (Figure 12.8) is relatively inexpensive to fabricate (it was used on the EUVE satellite, for example) despite some fixed pattern noise. Delay line anodes suffer much less from the fixed pattern noise caused by stretched and compressed mapping regions of the detector. In a delay line the charge cloud from the MCP strikes an anode structure and two pulses begin to propagate in opposite directions and the event location is deduced from the difference in arrival times at the ends, in a manner similar to gas proportional counters. Several delay line technologies have been developed by the experimental astrophysics group at University of California, Berkeley. Two-dimensional planar delay lines on a flat substrate were developed for the Far Ultraviolet Spectroscopic Explorer (FUSE) mission

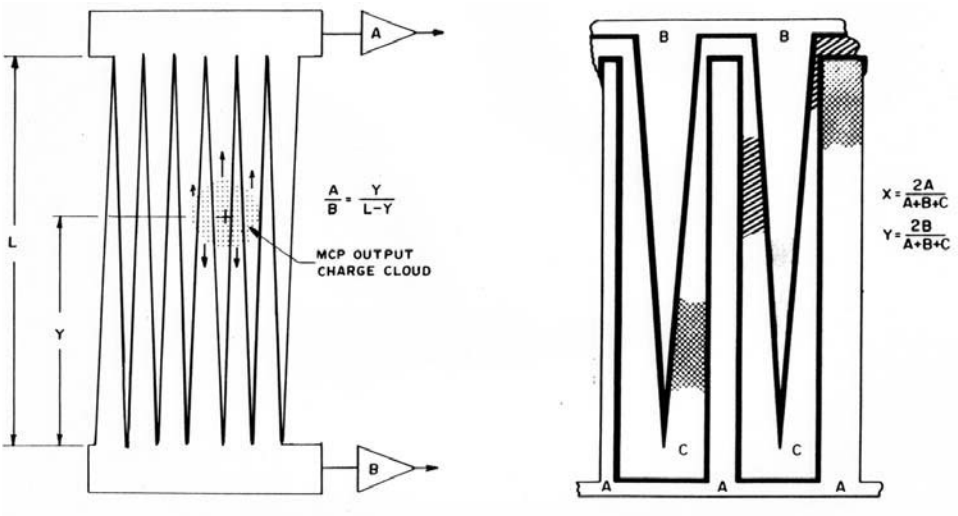


Figure 12.8. The wedge and strip readout scheme for an MCP.

(1999–2007), and delay lines have been used on the ORFEUS (Orbiting and Retrievable Far and Extreme Ultraviolet Spectrometer) and SOHO (Solar and Heliospheric Observatory) satellites. The joint ESA/NASA SOHO satellite has 12 instruments onboard covering a wide range of EUV and UV imagers, spectrographs, and coronagraphs, and helioseismology and solar wind experiments. Large-format delay lines are readily possible and are well-suited to instruments with large curved focal planes such as the Rowland circle spectrograph on FUSE. A double delay line with outputs at each corner is used with the GALEX MCPs. For COS (Cosmic Origins Spectrograph) to be installed during the fourth and presumably final Hubble service mission, the Berkeley Group developed a long, curved MCP (178×10 mm) with $32{,}768 \times 1{,}024$ pixels, a CsI photocathode and a resolution of 25 μm.

All readout structures which require very large gain usually employ not one—but several—MCPs in a stack called a "*Z*-stack" (see Figure 12.6 again). A stack of three MCPs can provide a gain of 2×10^7 electrons, which is needed for a delay line to give a spatial resolution of about 25 microns. MCPs suffer from "gain sag" due to the inability of the channels to replenish the charge extracted in time for the next pulse. Consequently, a second event arriving within the charge cloud radius while some of the lower channels are still depleted of charge will experience less gain. Whether this is a disadvantage or not depends on the application. For example, the Space Telescope Imaging Spectrograph (STIS) project selected the MAMA detector largely for this reason. MAMA detectors are depicted schematically in Figure 12.9 and one of the devices developed by Ball Aerospace for STIS is shown in Figure 12.10. There is also a MAMA detector (actually the flight spare from STIS) in the Advanced Camera for Surveys (ACS) on HST.

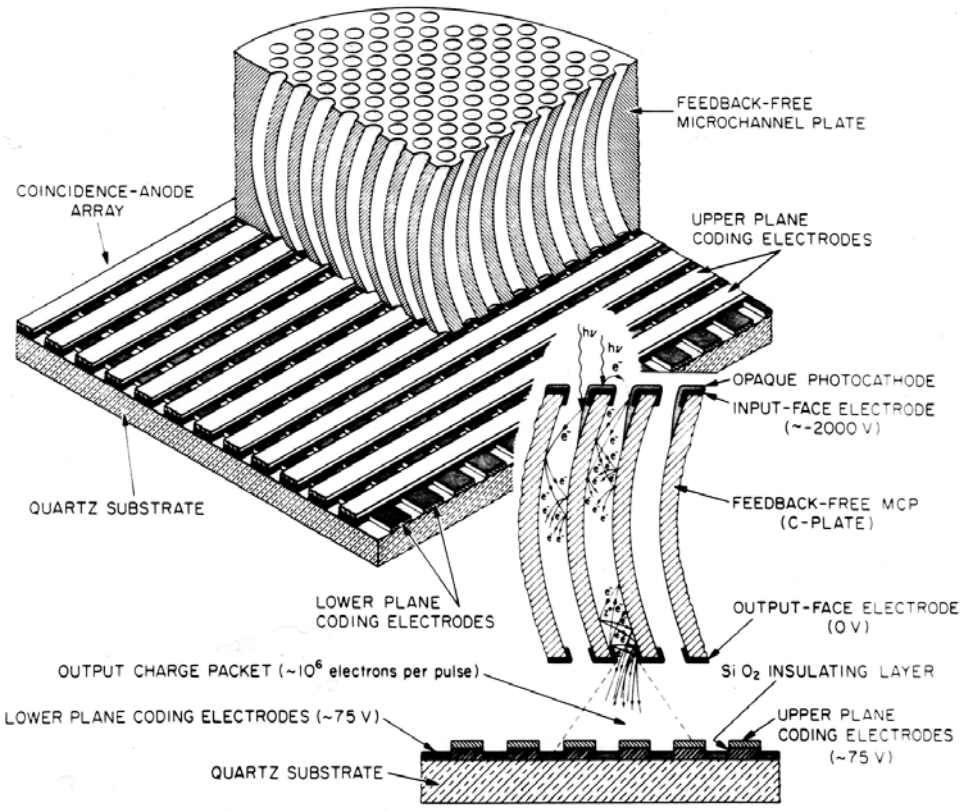


Figure 12.9. The construction of a microchannel plate detector with a MAMA readout. Credit: Ball Aerospace.

12.2.3 Electron-bombarded CCDs

An alternative approach to a UV imaging detector is to accelerate the electrons released from the UV-sensitive photocathode to high energies and allow them to impact a solid-state detector. The Digicon detectors used in the Hubble Space Telescope's Faint Object Spectrograph (FOS) and Goddard High-Resolution Spectrograph (GHRS) are in this class. Their disadvantage is the bulky and heavy electromagnetic focusing systems. New magnet technology, the availability of CCDs instead of linear arrays of diodes, and the fact that certain opaque photocathodes can actually have detective quantum efficiencies of 70% in the far ultraviolet, has kept this approach competitive. The structure of an EBCCD which has flown on sounding rockets and on the ORFEUS orbital mission is shown in Figure 12.11. Photoelectrons from the photocathode are accelerated to several thousand volts before bombarding the CCD. Each impact on the CCD produces 1 secondary electron for every 3.6 eV of incident energy. This result turns out to open up an opportunity for CCDs in the X-ray regime.

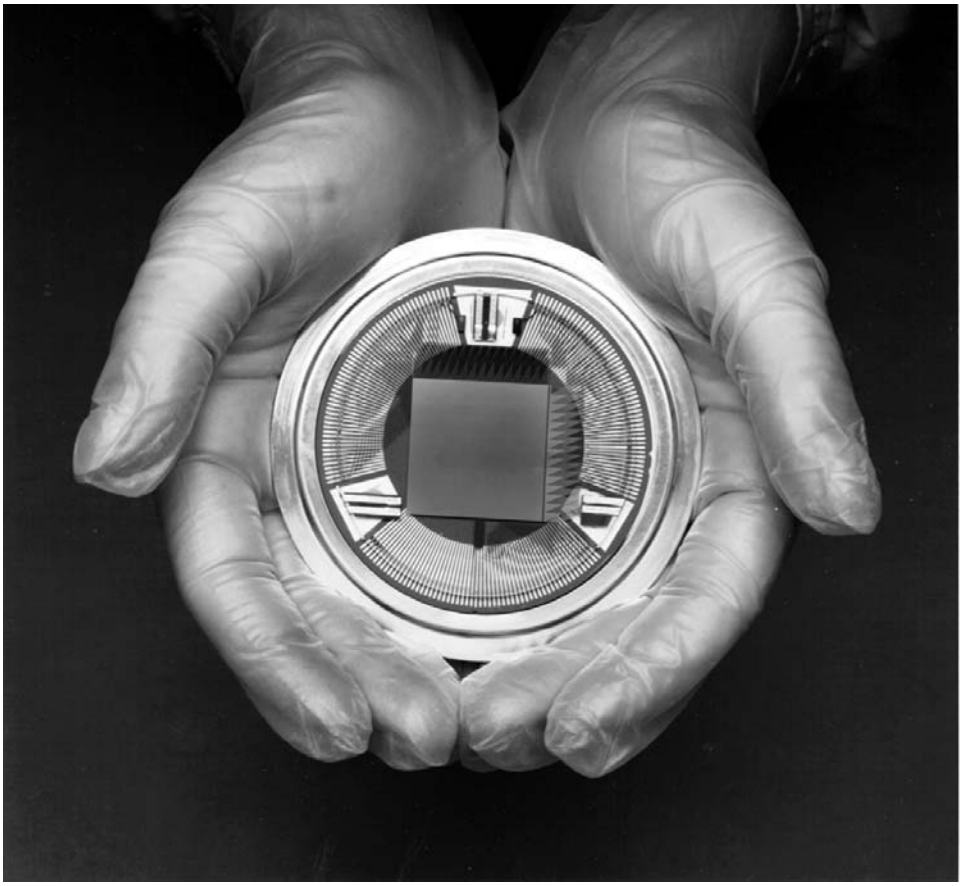


Figure 12.10. One of the MAMA detectors constructed for the Hubble STIS instrument. This device has the equivalent of 1,000 × 1,000 pixels. The same device is used in the ACS. Credit: Ball Aerospace.

12.3 X-RAY DETECTORS AND INSTRUMENTS

12.3.1 Introduction

One of the major successes of the space era for astronomy has been the ability to study X-ray emissions from cosmic sources. Beginning with brief rocket flights in the late 1940s, X-ray astronomy has flourished through several small satellites in the early 1970s which produced the first all-sky surveys and revealed hundreds of previously unknown sources, including black-hole candidates. This pioneering work led to three large satellites called the High Energy Astrophysical Observatories (HEAOs) launched between 1977 and 1979 which contained several X-ray and gamma-ray experiments. Especially successful among the HEAO missions was the second satellite which was renamed the Einstein Observatory (1978–1981). A big advance came with

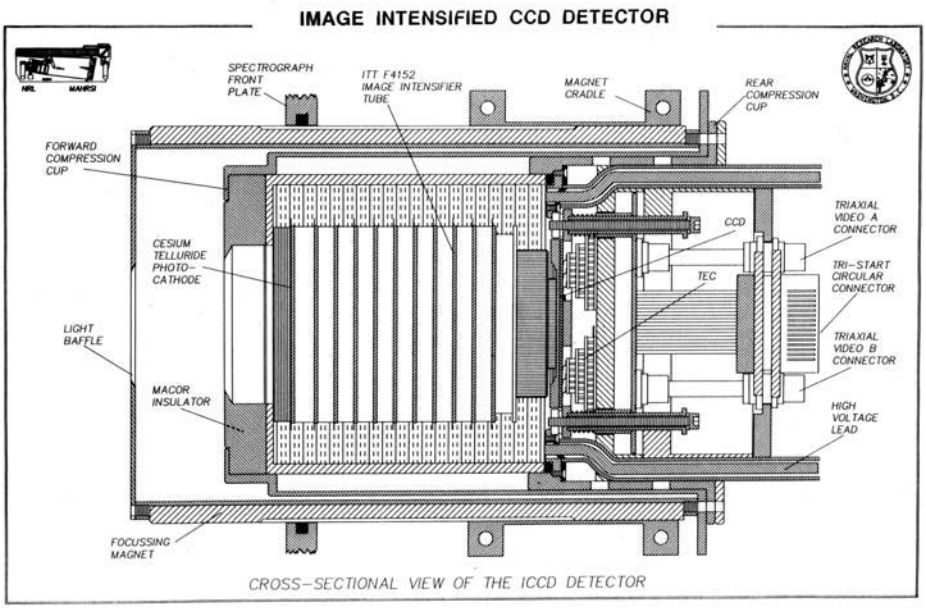


Figure 12.11. Electron-bombarded CCD (EBCCD) camera for ultraviolet imaging. Ultraviolet light is converted to electrons which are focused and accelerated to impact a CCD. Credit: George Carruthers.

the launch of ROSAT (1990–1999) because this satellite carried an X-ray telescope and a wide-field X-ray camera. Finally, with ESA's X-ray Multi-Mirror (XMM)-Newton mission (launched 1999) and NASA's Chandra X-ray Observatory (launched 1999), imaging at X-ray wavelengths became routine. The latter mission is named in honor of the late Indian American Nobel Laureate, Subrahmanyan Chandrasekhar, generally known as "Chandra" (which means "moon" or "luminous" in Sanskrit). Widely regarded as one of the greatest astrophysicists of the 20th century, Chandra moved from India to the United States in 1937, where he joined the faculty of the University of Chicago and he remained there until his death in 1995. He was awarded the Nobel Prize in 1983 for theoretical studies of the physical processes important to the structure and evolution of stars.

As a reminder, around a wavelength of 7 nm–10 nm (or 70 Å–100 Å), the extreme ultraviolet merges into the X-ray region which extends to wavelengths as short as 0.01 nm (0.1 Å) beyond which the photons are called gamma rays. At such short wavelengths it becomes more convenient to describe the photon by its energy $E = h\nu = hc/\lambda$ and convert this to electronvolts. A useful relation to remember is that the energy in keV is given by

$$E = 1.24\,\text{keV}/\lambda\,(\text{nm}) \tag{12.5}$$

for wavelengths in nm. The range of energies is from 0.1 keV to 125 keV. Lower energies (<10 keV) are called "soft" X-rays and the high-energy end is the "hard" X-ray region.

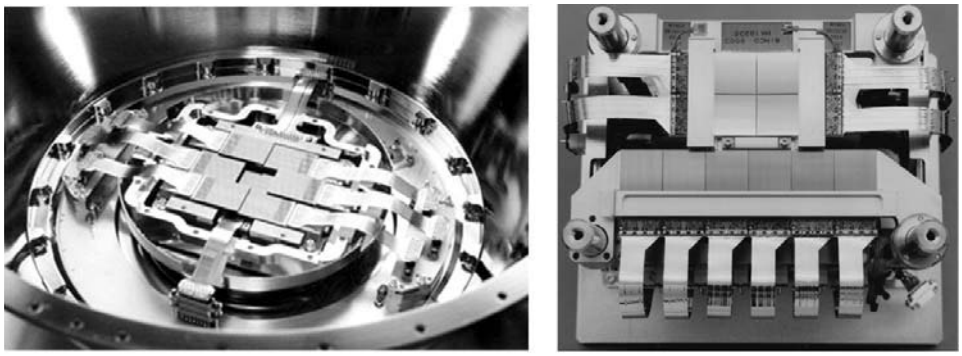


Figure 12.12. (Left) The EPICS CCD X-ray camera on XMM-Newton. (Right) The ACIS CCD X-ray camera on the Chandra X-ray Observatory. Credit: ESA and NASA.

The European Photon Imaging Camera (EPIC) onboard the XMM-Newton satellite is an X-ray imaging and spectroscopy instrument based on CCD technology (Figure 12.12). EPIC consists of three CCD cameras at the focus of three grazing incidence telescopes allowing medium spatial resolution of about 30 arcsec, but a point source location ability of 5 arcsec–10 arcsec, and a broad energy range from 0.1 keV to 10 keV with spectroscopic resolving power $E/\Delta E$ from 5 to 60. Two of the cameras contain a mosaic of seven 600×600-pixel devices, and the other is a variation which uses 12 back-illuminated 64×200-pixel CCDs with p–n junctions instead of MOS unit cells.

Formerly known as AXAF, the Chandra X-ray Observatory is the most powerful X-ray telescope yet. One of the major instruments onboard is the MIT/Penn State Advanced CCD Imaging Spectrometer (ACIS). It is also shown in Figure 12.12. This X-ray camera can generate high-resolution images of active galaxies, supernovas, and quasars, and simultaneously measure the energy of the detected X-rays. The ACIS instrument uses ten $1{,}024 \times 1{,}024$ CCDs to obtain images of astronomical sources in the X-ray region of the spectrum. Four CCDs are arranged in a 2×2 mosaic used for imaging (ACIS-I), and six CCDs are arranged in a linear strip which can be used either for imaging or spectroscopy. Two of the CCDs are back-illuminated and the others are front-illuminated. The idea at these wavelengths is to provide images which are spectrally resolved or, in other words, X-ray photons of different energies are separately identified.

12.3.2 CCDs in the X-ray regime

Early evaluations of CCDs by several groups led to considerable insights in device design which have enabled CCDs to be used for UV and X-ray imaging. Figure 12.13 shows the approximate efficiency with which photons are converted to electronic charges over the immense spectral range from 0.1 nm to 1,000 nm (1 Å–10,000 Å) for a backside-illuminated and UV-flooded CCD. The measured data points agree closely with theoretical expectations.

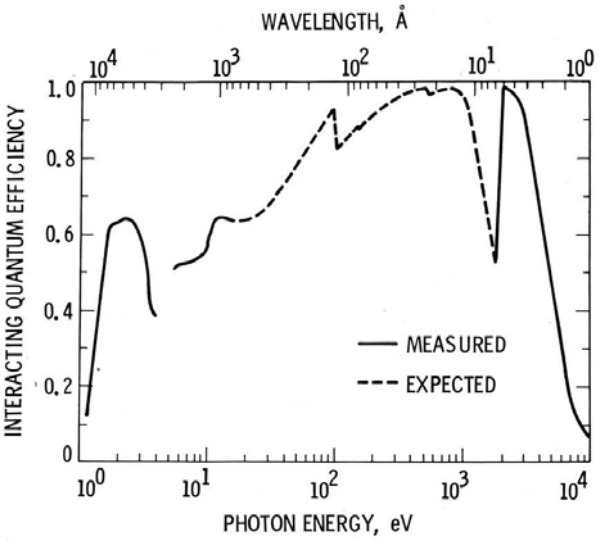


Figure 12.13. The quantum efficiency of a thinned backside-illuminated CCD over the entire X-ray, ultraviolet, and optical region of the spectrum. Credit: Jim Janesick.

The photoelectric absorption of a "soft" X-ray photon in silicon ejects an electron with energy $E - b$, where $E = h\nu$ is the energy of the X-ray, and b is the energy with which it is bound to the atom (typically 1,780 electronvolts or 1.78 keV). When free, this energetic electron produces a trail of electron–hole (e–h) pairs by collisions with orbital electrons. On average it takes 3.65 eV of energy to produce a single e–h pair in silicon. If all the energy of an X-ray photon was used to produce e–h pairs directly, there would be no statistical variation in the amount of charge generated by the X-ray event. However, a small but significant amount of energy is transferred into the silicon crystal lattice structure, giving a small statistical difference in the number of e–h pairs actually formed. This uncertainty is characterized by the "Fano factor". The Fano factor (F) reflects the ultimate limit of detector energy resolution and relates the measured width or spread in energy (δE) and the average energy (E) needed to produce e–h pairs in silicon by

$$\delta E = 2.355\left(\frac{FE}{3.65}\right)^{1/2} \tag{12.6}$$

where δE is the full-width-at-half-maximum energy of the observed spread (in electrons); E is the photon energy in electronvolts; and F is the Fano factor. The Fano factor is approximately 0.1 for silicon, and Fano noise-limited CCDs are now a reality over the entire soft X-ray regime from 100 eV to 10,000 eV. To achieve this performance the CCD readout noise must be less than 2 electrons (rms), and any "fat zero" or "pre-flash" level needed to eliminate any form of charge-trapping must be less than 1 electron! Several CCD manufacturers now produce devices with such properties. Besides X-ray astronomy, there are several potential applications in

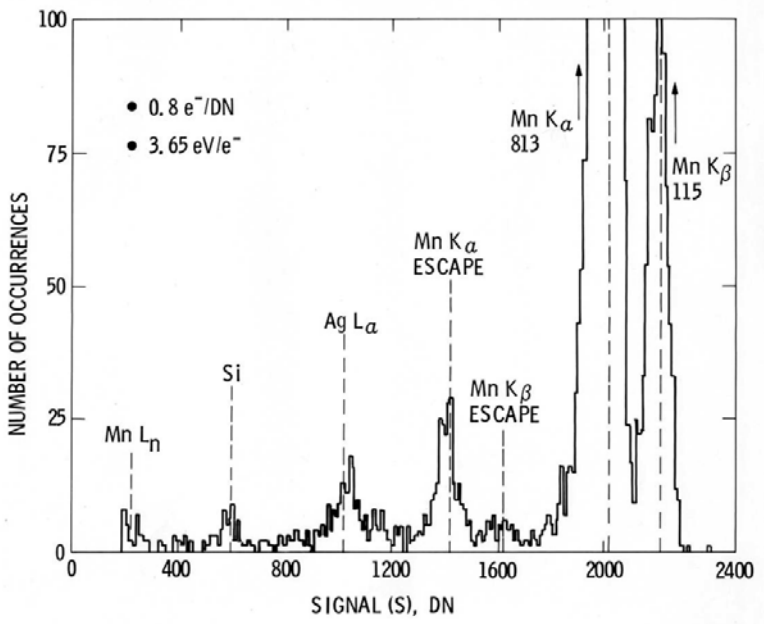


Figure 12.14. A histogram of X-ray events recorded from a CCD camera revealing its ability to distinguish X-rays of different energies from different atomic species. Credit: Jim Janesick.

high-temperature plasma diagnostic instrumentation and X-ray experiments on synchrotron sources. An example of the CCD acting as an X-ray detector is shown in Figure 12.14 for a laboratory energy source of Fe55 (an isotope of iron); note that Ag (silver) and Mn (magnesium) are also detected in the source. Jim Janesick promoted the use of the X-ray sensitivity of CCDs both as an analyzer of X-ray emission and as a diagnostic of CCD performance. In particular, X-rays can be used to determine the charge transfer efficiency of CCDs because a known charge is deposited in a pixel by the X-ray event. For example, the Fe55 source yields 5.9 keV X-ray photons (the K-alpha line) which deposit 1,620 electrons each in the CCD. Several references to the technical literature are given at the end of this chapter, and two examples of the usefulness of X-rays for testing CCDs are shown in Figure 12.15.

12.3.3 X-ray spectroscopy

X-ray wavelengths are of the same order as the spacing between atoms in crystal structures, and the regularity of these structures can lead to strong reflections due to interference effects. Consider a crystal in which the atoms are aligned in planes separated by a distance d. As the X-ray wave enters the crystal at an angle θ to the surface, a small portion is reflected at the same angle on the opposite of the normal. The rest of the wave continues on to the second layer of atoms where the process occurs again. The wavefront from the second layer is retarded by a distance of $2d \sin\theta$ relative to the first reflection. The summation of all the reflected waves

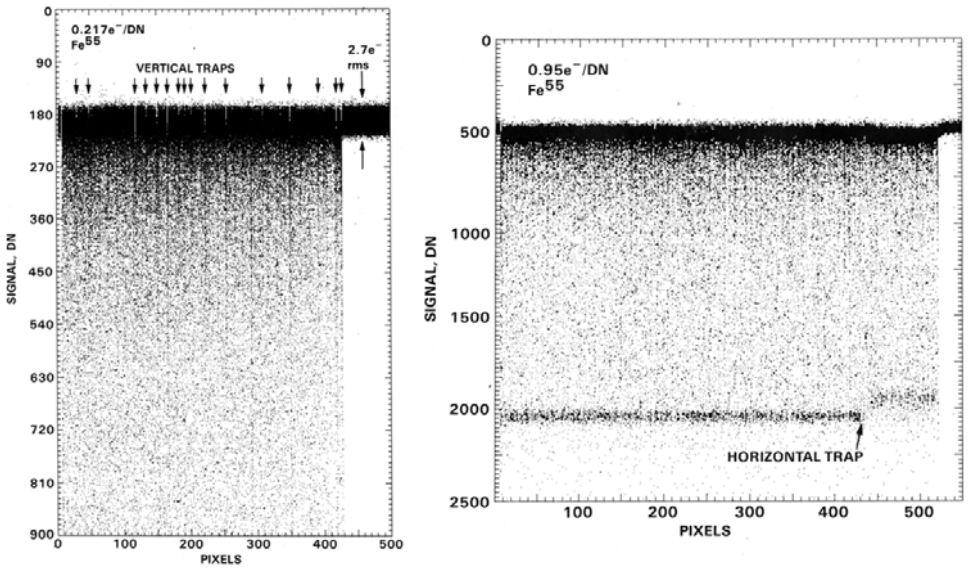


Figure 12.15. (Left) An Fe55 X-ray line trace response of a CCD which exhibits Fano noise-limited performance at the $2.7e^-$ readnoise level. (Right) An Fe55 X-ray line trace response for a CCD containing a small single-pixel trap in its horizontal register. The trap displaces about $100e^-$ into deferred charge. The dark line at about 2,100 DN is due to the $1{,}620e^-$ single-pixel events from the Fe55 source. Partial events and events split between pixels result in the grayscale gradation of signal values.

from subsequent layers will result in an intense beam only if they remain in phase with each other. This condition for constructive interference is given by Bragg's Law (W.H. Bragg, 1862–1942 and W.L. Bragg, 1890–1971):

$$2d \sin \theta = n\lambda \tag{12.7}$$

where n is an integer; and λ is the wavelength. The behavior is similar to an optical diffraction grating except that it is the regular atomic spacing that provides the coherence at X-ray wavelengths. Bragg spectrometers of enormous size are used on the Chandra X-ray telescope to provide a spectroscopy mode of operation.

12.3.4 X-ray instruments: the Chandra X-ray Observatory (CXO)

To illustrate the nature of cameras and spectrometers at these wavelengths a little further, we look more closely at the instrument complement on the CXO (Figure 12.16). Two of the four instruments in the focal plane are designed for imaging: HRC and ACIS. The other two instruments, LETG and HETG, are grating arrays which can be inserted into the path to enable either HRC or ACIS to obtain X-ray spectra. The High Resolution Camera (HRC) uses a pair of micro-channel plates. They each consist of a 10 cm (4-inch) square cluster of 69 million tiny lead oxide glass tubes that

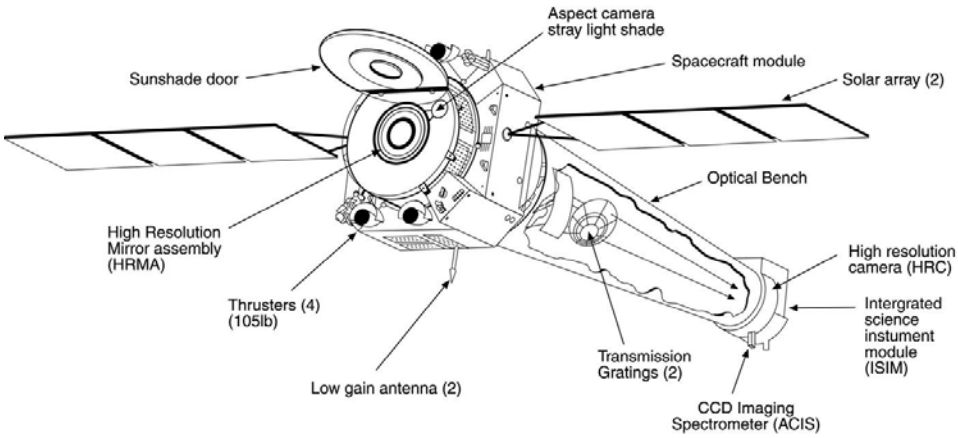


Figure 12.16. Layout of the Chandra X-ray Telescope and instrument package.

are about 10 μm in diameter (1/8 the thickness of a human hair) and 1.2 mm long. X-rays that impact the photocathode cause a cascade of about 30 million electrons to land on the crossed-grid of wires which allows the position of the original X-ray to be determined with high precision. The HRC's resolution can be ~0.5 seconds of arc, and it is especially useful for imaging supernova remnants and the hot tenuous matter in distant galaxies and clusters of galaxies. The Advanced CCD Imaging Spectrometer (ACIS), as the name suggests, is an array of CCDs. One group of CCDs is a closely spaced 2×2 array of four chips, and the other is a line of six devices. This instrument is especially powerful because, as a result of the way X-rays interact with the silicon CCD, it can form an image and simultaneously measure the energy of each incoming X-ray. Pictures of objects can be obtained using only X-rays produced by a single chemical element, enabling one to compare the appearance of a supernova remnant using radiation emitted by either oxygen, neon, or iron ions.

The other two instruments are dedicated to high-resolution spectroscopy: the High Energy Transmission Grating Spectrometer (HETGS) and the Low Energy Transmission Grating Spectrometer (LETGS). Each spectrometer is activated by swinging a large assembly into position behind the grazing incidence mirrors. This assembly holds hundreds of gold transmission gratings. Bragg diffraction of the incoming X-rays changes their direction by an amount that depends on the X-ray energy (or wavelength). One of the focal-plane cameras, either HRC or ACIS, detects the location of the diffracted X-ray, enabling precise determination of that energy. The resolving power of ~1,000 is sufficient to distinguish individual X-ray lines. For the LETG the grating is made of fine wires or bars with a regular spacing of 1 μm. The fine gold wires are held by two different support structures: a linear grid with 25.4 μm and a coarse triangular mesh with 2 mm spacing. The gratings are mounted onto a toroidal ring structure matched to the Chandra mirrors. These gratings are designed to cover an energy range of 0.08 keV to 2 keV. The HETG gratings, on the other hand, have a much finer period: 0.2 μm or 2,000 Å for the highest energy and 0.4 μm

or 4,000 Å for the medium-energy X-rays. To distinguish between them the two types of gratings are oriented at slightly different angles, so that the X-rays are diffracted in an "X" pattern at the focal plane. The size of the gold grating bars is smaller than a wavelength of visible light! These gratings take advantage of the fact that the gold bars are partially transparent to X-rays, so that diffraction is more efficient, and more X-rays are captured in the high-resolution spectrum. HETG gratings are designed for the range of 0.4 keV to 10 keV.

All of the focal-plane instruments are mounted on the Science Instrument Module or SIM. The SIM contains mechanisms to move the science instruments in and out of the beam, insulation for thermal control, and electronics to handle the operation of the instruments via the communication, command, and data management systems of the spacecraft. The CXO science instruments are controlled by commands transmitted from the Operations Control Center in Cambridge, Massachusetts. As with HST and other observatory-class satellites, a preplanned sequence of observations is radioed to the spacecraft and stored in the onboard computer for later execution. Data are stored on a recorder for later transmission to the ground every eight hours during regularly scheduled contacts with the Deep Space Network. The stored measurements are transmitted to the Jet Propulsion Laboratory and then to Operations Control at the Chandra X-ray Center (CXC) in Cambridge, MA for processing and analysis. Figure 12.17 is an X-ray image of Cen A, also shown in color on the book cover and in Plate 4, which illustrates the remarkable success of the CXO.

12.4 GAMMA RAYS

Cosmic processes thought to be sufficiently energetic to produce gamma rays include the decay of radioactive nuclei, accelerated particles traveling in strong magnetic fields (synchrotron radiation), electrons colliding with lower energy photons and boosting their energies to gamma-ray levels (inverse Compton scattering), very energetic nuclei colliding with other nuclei, energetic electrons "braking" (Bremsstrahlung) as they traverse matter because the deceleration of charged particles results in emission of radiation, and, finally, the annihilation of matter in the presence of antimatter. An example of the latter process is the conversion of an electron and a positron into a pair of 511 keV gamma rays ($e^+ + e^- \rightarrow \gamma + \gamma$). Localized sources of gamma rays include supernovas, quasars, the central regions of active galaxies, pulsars and neutron stars, gamma-ray bursters (GRBs), and black holes. Soft gamma rays have energies in the 120 keV–1.2 MeV range corresponding to photon wavelengths of ~0.01 nm to 0.001 nm; photon energies above 1.2 MeV are generally called gamma rays even up into the very high energy (VHE) regimes of GeV and TeV photons. In fact, photons from the Crab Nebula (M1) have been detected up to 100 TeV. Thus, gamma rays span nine orders of magnitude in energy ($10^{14}/10^5$), which is a huge dynamic range unsurpassed by any other waveband. Consequently, a single-detector technology will not cover this entire range. The regime from 100 GeV to 100 TeV is covered by atmospheric Cherenkov telescopes and air shower arrays.

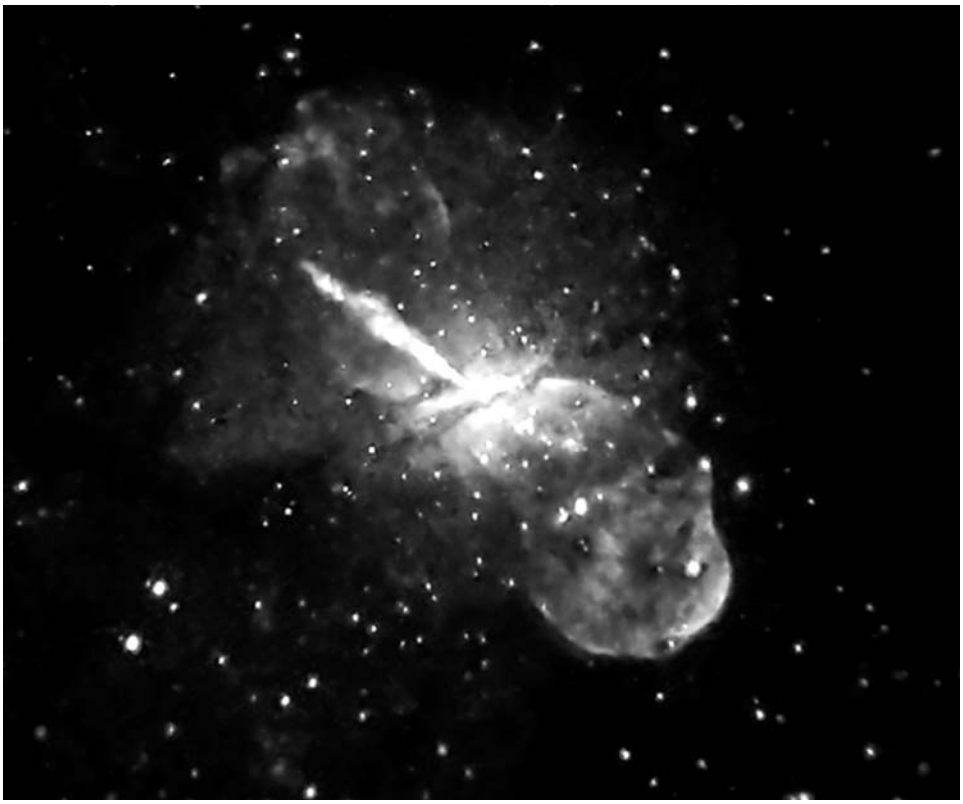


Figure 12.17. A multi-band X-ray image of Centaurus A from the CXO (see also Plate 4). Credit: NASA/Kraft *et al.*

Gamma rays cannot be readily focused even using grazing incidence mirrors. Moreover, the primary detectors are different from those considered so far. Gamma rays are also different from other photons in that they may not be fully absorbed. In fact, there are three regimes: total absorption by the photoelectric effect, Compton scattering, and pair production. The first two regimes are only important for a relatively narrow band of energies between 100 keV and 10 MeV. Above 10 MeV it is pair production that dominates. Pair conversion telescopes use devices like spark chambers and silicon strip detectors to track the particles produced.

12.4.1 Detectors for high energy

Geiger (or Geiger–Müller) tube and proportional counters

These devices are collectively known as gas-filled ionization detectors, and typically consist of a gas-filled chamber subdivided into a number of low and high electric

field regions by some arrangement of electrodes. A typical arrangement might use argon as the gas and create the potential difference between a central anode wire and a cylindrical wall as the cathode. The electrodes are maintained at a potential difference such that the entry of an ionizing (high-energy) photon will trigger a discharge in the medium filling the cell, which in turn creates an avalanche of electrons and a pulse of current which can be detected or counted at the anode. In a Geiger counter the avalanche quickly saturates and the detected pulse is independent of the original photon energy. Moreover, the device requires time to recover before another discharge can be triggered. A proportional counter (PC) employs the same basic principle but operates below the trigger voltage so that pulse saturation is avoided and the strength of the pulse is proportional to the energy of the original interaction. The avalanche gain will be much less (10^5 instead of 10^8) but still sufficient for pulse detection. At low photon energies (X-rays) the PC requires a window, typically of plastic, mica, or beryllium. Proportional counters can be packaged together to form a basic two-dimensional grid which allows imaging.

Scintillation counters

Scintillation devices work when the high-energy photon stimulates the emission of visible light which can then be detected with a photomultiplier tube. Sodium iodide (NaI) and cesium iodide (CsI) can be used as scintillators for energies up to several hundred kilo-electronvolts, and organic materials such as stilbene ($C_{14}H_{14}N_2$) and bismuth germinate ($Bi_4Ge_3O_{12}$ or BGO) can be used for 10 MeV or higher energies. The decay constant for optical emission lies in the microsecond range for most inorganic scintillators, and in the nanosecond range for plastics. Thus, by surrounding an alkali halide crystal with a plastic shield, and observing the sandwich or "phoswich" that these two create using a single photomultiplier tube, the pulse *shape* can be used as a discrimination tool to determine whether the energy loss occurred in the shield or in the main detector. This is an excellent method of background discrimination.

Spark chambers

A spark chamber consists of a series of electrically charged plates separated by small gaps with progressively higher voltages on each plate. An ionizing particle moving through the chamber causes a series of sparks (discharges) which mark the path through the chamber. For the detection of gamma rays at energies above 20 MeV, a plate of material with high atomic weight (e.g., tungsten, tantalum) is used to cause the photon to create an electron–positron pair, and the path of each particle is then tracked. Other devices are typically used around the spark chamber to act as triggering devices or as anti-coincidence detectors to help distinguish between cosmic-ray particles and gamma-ray photons.

Silicon strip detectors

A reverse biased p–n junction, with sufficient bias voltage to cause complete depletion of the carriers, forms a kind of semiconductor version of the "ionization chamber". If a p-type silicon layer is formed on top of an n-type substrate, narrow insulating tracks of silicon dioxide can be used to separate regions of the p-type layer into pads, each pad behaving as an independent detector. The passage of a charged particle creates thousands of electron–hole pairs, and the holes drift to the negative p-type pad, giving an observable signal. Many narrow strips $\sim$50 μm wide are used in silicon strip detectors, thus providing good position resolution. Sheets of silicon detectors can be stacked in towers to produce a solid-state equivalent of the spark chamber. An incident gamma-ray photon is forced to pair-produce by a plate of high atomic weight material, and the electrons and positrons from the conversion cause ionization in the Si strip.

Germanium solid proportional counter:

A liquid-nitrogen cooled cylinder of solid germanium (Ge) with a central axial anode and surrounded by a cylindrical cathode can be used like a proportional counter. The incoming gamma ray scatters off Ge atoms until its energy is lost in ion–electron pair production. Electrons are attracted to the anode and the number of released electrons is proportional to the energy of the gamma ray. The spectral resolution is quite high ($\delta E/E \sim 0.2\%$) at 1 MeV. This unit can form a pixel in a large package of similar units to create an imaging device. This is indeed the approach used by the INTEGRAL satellite which has a stack of 19 Ge detectors to match its coded mask.

Cadmium zinc telluride detector arrays

CdZnTe or CZT is a large-bandgap semiconductor with high stopping power. This material can provide room temperature detection (no cooling) of 10 keV to 1 MeV photons with good position and energy resolution ($\Delta E/E \sim 3\%$ at 122 keV). CZT detectors are fabricated with very thin metal electrode geometries deposited on the detector surfaces. These electrodes are electrically biased creating a difference in electrical potential within the detector volume. When ionizing radiation interacts with the CZT crystal, electron–hole pairs are created in proportion to the energy of the incoming radiation. The negatively charged electrons and positively charged holes migrate to the oppositely charged electrode where they are collected. The resulting charge pulse is detected by the pre-amplifier, which produces a voltage pulse whose height is proportional to the incident energy of the incoming photon. The signal from the pre-amplifier is then fed into a shaping amplifier that converts the signal into a Gaussian pulse and amplifies it. The signal can then be fed into a standard counting system. The importance of this technology is that CZT elements can be tightly packaged to make an "array" detector for gamma rays (Figure 12.18). A large array of CZT elements is used as pixels to create a focal plane imager for gamma-ray observations in the Burst Alert Telescope (BAT) instrument onboard the SWIFT satellite.

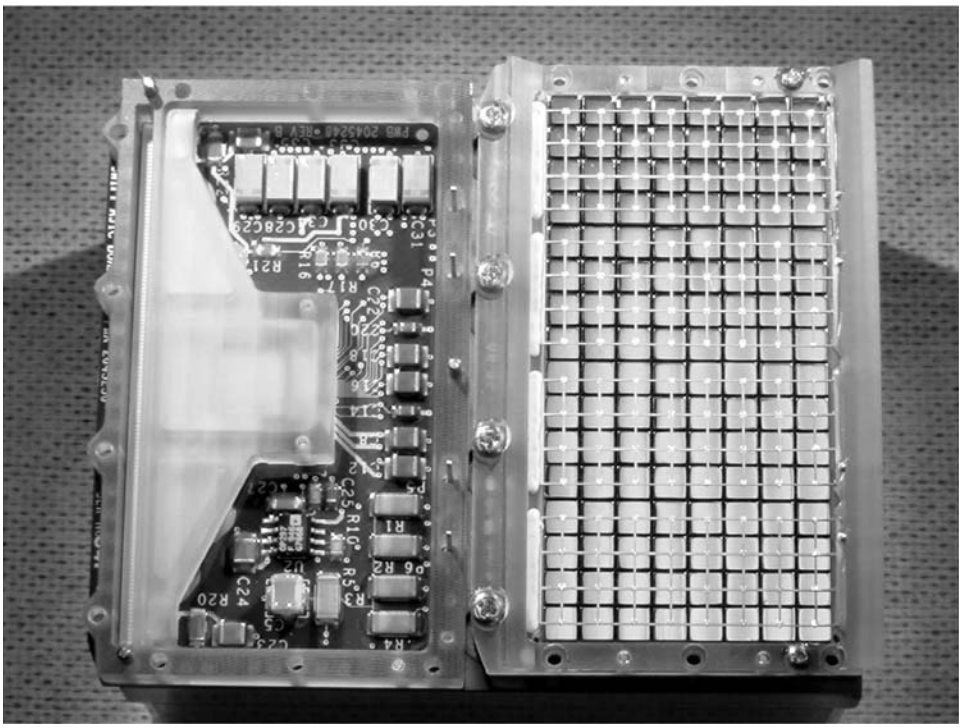


Figure 12.18. One block of CZT detectors from the BAT focal plane on the SWIFT gamma-ray satellite. Credit: Neil Gehrels.

Atmospheric Cherenkov telescopes and air shower arrays

Here the Earth's atmosphere becomes the medium that forces high-energy gamma rays to pair-convert. The particles or the Cherenkov light produced in the shower from that initial conversion are then detected at the ground by air shower particle arrays, or in the atmosphere by wide-field telescopes looking for the Cherenkov light.

12.4.2 Gamma-ray observatories in space

CGRO, the Compton Gamma Ray Observatory, was launched in 1990 and began operations in 1991 as one of NASA's four "great observatories", and was an extremely successful mission that opened up the field and paved the way for follow-up missions. Among its discoveries was that GRBs (gamma-ray bursters) were uniformly distributed over the sky suggesting an extra-galactic origin. The satellite was eventually allowed to burn up in the atmosphere at the end of its life. CGRO had a complement of four instruments which illustrate the detector technologies just mentioned and lay the ground work for those on more recent satellites: the Burst and Transient Source Experiment (BATSE), the Oriented Scintillation

Spectrometer Experiment (OSSE), the Imaging Compton Telescope (COMPTEL), and the Energetic Gamma Ray Experiment Telescope (EGRET).

Size is crucial for gamma-ray astronomy because the number of gamma-ray events recorded is directly related to the mass of the detector material used to intercept them. As the flux of gamma rays is small, large instruments are needed to detect a significant number in a reasonable time. CGRO detected photons with energies from the high end of the X-ray spectrum at 20 keV up to 30,000 million electronvolts (30 GeV).

Each of the CGRO instruments involved scintillators. Because cosmic rays, particles traveling with very high energies, also cause scintillators to produce light, and because these particles are much more numerous than gamma rays, there is a "background problem" very similar to trying to see normal stars during the daytime or infrared stars against the thermal (heat) emission from the Earth's atmosphere and warm telescope. Almost all astronomical measurements must be made against some kind of unwanted background. To reduce the background of cosmic rays, each CGRO instrument distinguished gamma rays by using a plastic outer detector called an "anti-coincidence counter" which detects cosmic rays and other charged particles by giving off a signal different from gamma rays. Computer programs compare the signals from the main instrument and the anti-coincidence counter to identify true gamma rays. The CGRO instruments are summarized briefly below.

OSSE: Energy range 0.1 MeV–10 MeV. Four identical detectors with a main scintillation crystal made of sodium iodide (NaI) surrounded by other scintillators that absorb gamma radiation arriving from the sides and back. A very massive collimator in front defined a 3.8×11.4 degree field of view. Each detector could be rotated to point on and off a source.

BATSE: Energy range 20 eV–600 keV. Observed the entire sky (not blocked by the Earth) in the field of view of its eight scintillation detectors and looked for changes in gamma-ray intensity in time intervals as short as a fraction of a millisecond.

COMPTEL: Energy range 1 MeV–30 MeV. Gamma rays in this energy range were scattered by electrons in the liquid scintillators in COMPTEL's upper detectors. This process, called Compton scattering, occurs only between electrons and high-energy photons—not between electrons and other particles. The scattered gamma rays were detected by a lower set of crystal scintillation detectors. COMPTEL had a 60° field of view.

EGRET: Energy range >10 MeV. When a gamma ray with energy greater than 10 MeV encountered the metal layers of the EGRET spark chamber, it produced an electron–positron pair. The track of the pair and the angle between them is related to the direction and energy of the original gamma ray. Large electric fields between the plates cause breakdown of the inert gas in the chamber, but only along the track of the charged particles producing a trail of sparks that was imaged. A massive

scintillation detector below the spark chambers helped to determine the energy. EGRET had a 30° field of view.

Launched in 2003, SWIFT was designed to study gamma-ray bursts (GRBs) at a high rate (>100 per year) and obtain excellent positions (0.3 arcsec–5 arcsec) for each GRB using a multi-wavelength imaging approach (gamma ray, X-ray, UV, and optical). At least three times more sensitive than the BATSE on CGRO, SWIFT can also provide spectroscopy from 180 nm to 600 nm and from 0.2 keV to 150 keV, thus enabling one to determine the redshift of the source directly. The payload is comprised of three instruments which work in tandem. Within 20 to 75 seconds of a detected GRB, the telescope slews autonomously so that the fields-of-view of the pointed instruments overlap the location of the burst. The afterglows are monitored over their durations, and the data rapidly disseminated to other observatories for follow-up action. The role of each instrument is briefly summarized.

BAT (15 keV–150 keV): With its large field of view (FOV) and high sensitivity, the Burst Alert Telescope detects about 100 GRBs per year, and computes burst positions onboard with arcminute positional accuracy. BAT's 32,768 pieces of $4 \times 4 \times 2$ mm CdZnTe (CZT) form a huge pixelated gamma-ray camera of size 1.2×0.6 m. Groups of 128-detector elements are assembled into 8×16 arrays, each connected to 128-channel readout application-specific integrated circuits (ASICs). Detector modules, each containing two such arrays, are further grouped by eights into blocks to build up the large focal-plane gamma-ray camera (Figure 12.19). This hierarchical structure, along with the forgiving nature of the coded aperture technique, means that BAT can tolerate the loss of individual pixels, individual detector modules, and even whole blocks without losing the ability to detect bursts and determine locations. The CZT array camera has a nominal operating temperature of 20°C, and its thermal gradients (temporal and spatial) are kept to within 1°C. A typical bias voltage is −200 V, with a maximum of −300 V. An important part of BAT is its D-shaped coded aperture mask, made of ~52,000 lead tiles ($5 \times 5 \times 2$ mm) mounted on a 5 cm thick composite honeycomb panel, which is supported by composite fiber struts 1 meter above the detector plane. Because the large FOV requires the aperture to be much larger than the detector plane and the detector plane is not uniform due to gaps between the detector modules, the BAT coded aperture uses a completely random, 50% open–50% closed pattern, rather than the commonly used uniformly redundant array pattern. The mask has an area of 2.7 m^2, yielding a half-coded FOV of 100×60 degrees, or 1.4 steradians. A graded *Z*-fringe shield located both under the detector plane and surrounding the mask and detector plane reduces background from the isotropic cosmic diffuse flux and the anisotropic Earth albedo flux by ~95%. This shield is composed of layers of Pb, Ta, Sn, and Cu, which are thicker nearest the detector plane and thinner near the mask. The combination of the 4 mm square CZT pieces, plus the 5 mm square mask cells and the 1 m detector-to-mask separation gives an instrumental angular resolution of ~17 minutes of arc FWHM, yielding a conservative positional accuracy of 4 arcminutes for bursts, and steady-state sources given a signal-to-noise ratio of 8 for the burst threshold.

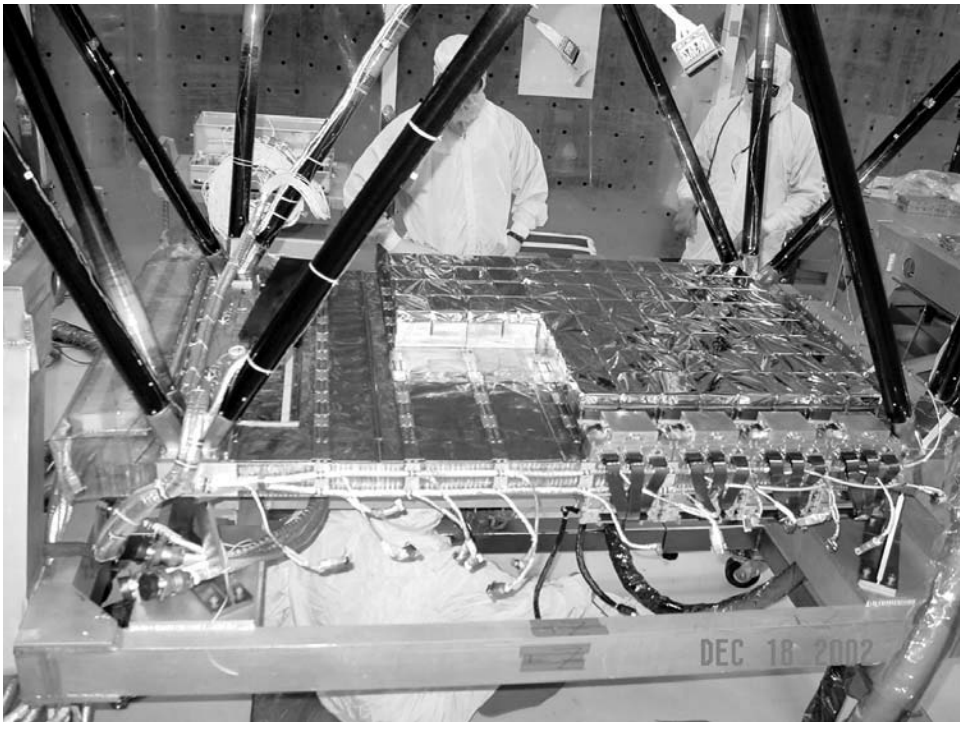


Figure 12.19. The large focal plane of individual CZT blocks on SWIFT being assembled beneath the coded mask telescope. Credit: Neil Gehrels.

XRT (0.3 keV–10 keV): The X-Ray Telescope takes images using a CCD detector and is able to obtain spectra of GRB afterglows during pointed follow-up observations. The images are used for higher accuracy position localizations, while the spectra are used to determine redshifts from X-ray absorption lines. The XRT focal plane camera is based on the *XMM*/*EPIC* cryostat design. Cooling of the CCD is done using a Peltier cooler connected to a cold finger and radiator at the back of the telescope. A thin filter is placed in front of the CCD to block optical light. The CCD22 device from *e2v* was designed for the *EPIC* MOS instruments on *XMM*. It is a three-phase frame transfer device which utilizes high-resistivity silicon and an open-electrode structure to achieve a useful bandpass of 0.2 keV to 10 keV. The image section of each CCD is a 600×602 array of 40 mm pixels, each pixel corresponding to 2.36 arcsec in the SWIFT focal plane, whereas the storage region is a 600×602 array of 39×12 μm pixels. The readout register is split into two sections, and may be read out using either output node, or may be split and read out using both nodes simultaneously. This CCD may also be operated in timing mode or window mode, which allow faster readout of fewer pixels so that bright sources may be observed without saturation.

UVOT (170 nm–650 nm): The UV Optical Telescope is essentially a copy of the XMM Optical Monitor (OM). The UVOT takes images and obtains spectra (via a grism filter) of GRB afterglows during pointed follow-up observations. These images are used for position localizations to within 0.3 arcsec–2.5 arcsec, while the spectra are used to determine redshifts and Lyman-alpha cutoffs. The UVOT is a 30 cm modified Ritchey–Chrétien UV/optical telescope co-aligned with the X-ray Telescope and mounted on the telescope platform common to all instruments. Photons register on a photon-counting microchannel plate intensified CCD (MIC) with $2{,}048 \times 2{,}048$ pixels covering a 17×17-arcminute FOV and a wavelength range of 170 nm–650 nm. The camera is a copy of the MIC detectors mentioned earlier for XMM-Newton, and the optics are the flight spares from XMM. SWIFT is a remarkable illustration of the fact that astronomers can now obtain images all across the electromagnetic spectrum, in this case simultaneous for gamma rays, X-rays, UV, and visible light.

As SWIFT Principal Investigator Neil Gehrels points out, gamma-ray astronomy has a serious background problem even from space, and thus imaging observations are background-limited. Therefore, the instrument sensitivity (signal-to-noise ratio) scales as the square root of observing time and the detector (telescope) area. The next gamma-ray space mission will be GLAST (Gamma-ray Large Area Space Telescope).

12.4.3 Ground-based gamma-ray observatories

Very high energy photons can be detected from the ground by using the atmosphere as a detector. Cosmic-ray particles (hadrons), stripped nuclei from protons all the way up to iron, and very high energy photons create secondary particles in the upper atmosphere which are moving at speeds slightly in excess of the local speed of light in the medium (air), causing an abrupt change in the electric field near atoms that they pass, rather like a shock front, causing the atoms to radiate. The emitted light is known as Cherenkov radiation. This radiation is concentrated in a cone spreading outwards from the direction of motion of the particle. For energies $E > 100$ GeV, the emission is at visible wavelengths, and the shower can be imaged directly with an "array of pixels" each of which is a discrete photomultiplier tube spanning $\sim 0.1^\circ$–0.2° on the sky. This imaging array is at the prime focus of a large optical reflector. It is the imaging quality that makes the Cherenkov telescope so powerful at rejecting 99.99% of hadronic background. Showers caused by gamma-ray photons develop differently from those due to hadrons. Due to the smaller transverse momentum of electromagnetic interactions, the Cherenkov radiating particles in the gamma-ray shower are, on average, closer to the direction of the primary. In addition, as there is no penetrating component in electromagnetic showers, then the local contribution from individual particles to the light is small and the fluctuations in the shower image are less. Because gamma-ray images are better defined it is easier to characterize their direction and discriminate them from the isotropic background of hadron showers.

The power of the imaging technique was first demonstrated using an array of PMTs in the focal plane of the Whipple Observatory 10 m optical reflector in

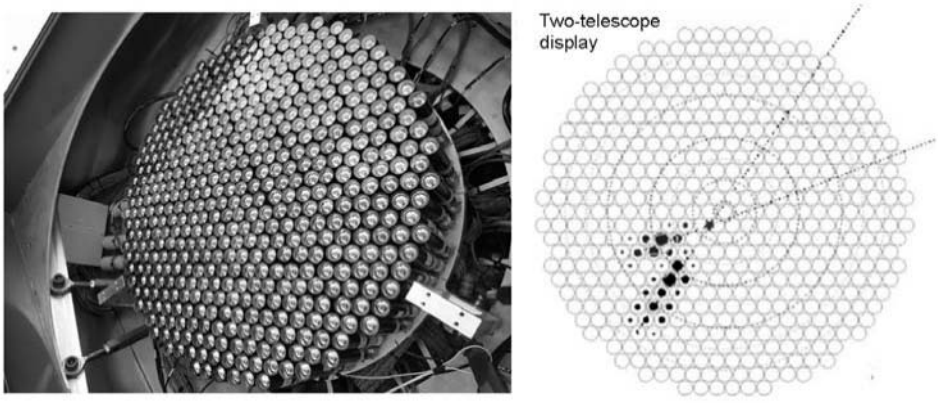


Figure 12.20. (a) An array of 499 PMTs form the pixels in the VERITAS camera designed to observe Cherenkov light produced when very high energy gamma rays impact the upper atmosphere; and (b) air Cherenkov light images from two telescopes intersect to define arrival direction. Credit: Luis Valcarel.

Arizona, and the technique has now been extended significantly by the use of multiple telescopes with separations of the order of the lateral spread of the light from the shower (Weekes, 2001). Known as VERITAS (Very Energetic Radiation Imaging Telescope Array System) this new major ground-based gamma-ray observatory is an array of four 12 m optical reflectors for gamma-ray detection in the 50 GeV–50 TeV energy range. Four VERITAS telescopes are in operation. Using a prime-focus "camera" composed of hundreds of individual PMTs, each PMT acting as an individual pixel, the telescopes observe a 3.5° field of view within which the Cherenkov shower usually appears as a small, narrow elliptical illumination patch pointing towards the optical axis. Figure 12.20 shows the VERITAS PMT camera. Another project of similar scale, called HESS (High Energy Stereoscopic System), is operating in Namibia. HESS also consists of 4×12 m telescopes with each telescope having a 4.8 degree field of view. HESS began operations in the southern hemisphere about 3 years before VERITAS. A third system is MAGIC (Major Atmospheric Gamma Ray Imaging Cherenkov) which has two 17 m diameter telescopes with a threshold for detection of 50 GeV gamma rays. It is located on El Roque de los Muchachos, La Palma, Canary Islands. All of these facilities require state-of-the-art, high-speed PMTs and electronics (amplifiers, triggers, data acquisition), which in large part are a derivative of particle physics experiments.

12.5 SUMMARY

Imaging at wavelengths shorter than visible light is now possible all across the UV, X-ray, and gamma-ray domain. Charge-coupled devices optimized for UV work can be used directly or as readouts in intensified photon-counting cameras. The most

common UV detector is the microchannel plate with a variety of photocathodes and readout schemes to provide imaging to the extreme ultraviolet. In the X-ray domain, silicon CCDs are again found to be useful imaging devices, but telescope design becomes harder as grazing incidence mirrors are needed. At gamma-ray energies, collimators and coded mask telescopes are the only way to direct gamma-ray photons to a focal plane, but large arrays of CdZnTe detectors have permitted imaging even at these wavelengths. Very high energy gamma-rays hitting the Earth's atmosphere can be studied indirectly by the resulting Cherenkov radiation using cameras with photomultiplier tubes as picture elements.

12.6 EXERCISES

1 Why do CCDs exhibit good response in the X-ray region and in the visible, but poor response in the ultraviolet?

2 Describe the principle of the microchannel plate (MCP), and explain what is meant by a "Z-stack". Under what circumstances does an MCP with a Z-stack exhibit "gain sag"?

3 Why must UV detectors be "solar-blind"?

4 What is meant by a Fano noise-limited CCD?

5 How are EUV and X-ray telescopes constructed to provide focal-plane images?

6 Summarize the primary detector technologies used for the detection of gamma rays.

7 What is meant by atmospheric Cherenkov radiation, and what can be learned from ground-based observations?

12.7 REFERENCES

Ables, J.G. (1968) Fourier transform photography: A new method for X-ray astronomy, *Proc. Astron. Soc. Australia*, **1**, 172–173.

Bowyer, S.; and Malina, R. (1994) Extreme ultraviolet images of the Cosmos, *Physics World*, **7**(1), January.

Carruthers, G.R. (1994) Ultraviolet and X-ray detectors, in R.W. Waynant and M.N. Ediger (eds.), *Electro-Optics Handbook*, McGraw-Hill, New York.

Dicke, R.H. (1968) Scatter-hole cameras for X-rays and gamma rays, *Astrophys J. Letters*, **153**, L101.

Elvis, M. (ed.) (1990) *Imaging X-ray Astronomy: A Decade of Einstein Observatory Achievements*, Cambridge University Press, Cambridge, U.K.

Fleck, B.; Domingo, V.; and Poland, A. (eds.) (1995) *The SOHO Mission*, Kluwer Academic, Dordrecht, The Netherlands.

Gehrels, N.; Chincarini, G., Giommi, P.; Mason, K.O.; Nousek, J.A.; Wells, A.A.; White, N.E.; Barthelmy, S.D.; Burrows, D.N.; Cominsky, L.R. *et al.* (2004) The Swift Gamma-Ray Burst Mission, *Astrophy. J.*, **611**, 1005–1020.

Goldwurm, A.; Goldoni, P.; Gros, A.; Stephen, J.; Foschini, L.; Gianotti, F.; Natalucci, L.; de Cesare, G.; del Santo, M. (2001) Gamma-ray imaging with the coded mask IBIS telescope, in B. Battrick, A. Gimenez, V. Reglero, and C. Winkler (eds.), *Exploring the Gamma-ray Universe: Proc. of Fourth INTEGRAL Workshop, September 2000, Alicante, Spain*, ESA SP-459, ESA, Noordwijk, The Netherlands, pp. 497–500. Available at *http://arxiv.org/abs/astro-ph/0102386*

Heap, S.R.; Brandt, J.C.; Randall, C.E.; Carpenter, K.G.; Leckrone, D.S.; Maran, S.P.; Smith, A.M.; Beaver, E.A.; Boggess, A.; Ebbets, D.C. *et al.* (1995) The Goddard High-Resolution Spectrograph: In-orbit performance, *Pub. Astron. Soc. of the Pacific*, **107**, 871–887.

Janesick, J.; Elliott, T.; Collins, S.; Daud, T.; Campbell, D.; and Garmire, G. (1987) Charge-coupled device advances for X-ray scientific applications in 1986, *Optical Engineering*, **26**(2).

Janesick, J.; Elliott, T.; Bredthauer, R.; Chandler, C.; and Burke, B. (1988) Fano-noise-limited CCDs, *Optical and Optoelectronic Applied Science and Engineering Symposium: X-ray Instrumentation in Astronomy*, SPIE, Bellingham, WA.

Joseph, C.L. (1995) UV image sensors and associated technologies, *Experimental Astronomy*, **6**, 97–127.

Konda, Y. (ed.) (1987) *Exploring the Universe with the IUE Satellite*, Kluwer Academic, Dordrecht, The Netherlands.

Martin, D.C.; Fanson, J.; Schiminovich, D.; Morrissey, P.; Friedman, P.G.; Barlow, T.A.; Conrow, T.; Grange, R.; Jelinsky, P.N.; Millliard, B. *et al.* (2005) The Galaxy Evolution Explorer: A space ultraviolet survey mission, *Astrophys. J.*, **619**, L1–L6.

Mertz, L.; and Young, N.O. (1961) Fresnel transformations of images, in K.J. Habell (ed.), *Proc. Int. Conf. on Opt. Instrum. Techniques*, Chapman & Hall, London, p. 305.

Morrissey, P.; Schiminovich, D.; Barlow, T.A.; Martin, D.C.; Blakkolb, B.; Conrow, T.; Cooke, B.; Erickson, K.; Fanson, J.; Friedman, P.G. *et al.* (2005) The on-orbit performance of the Galaxy Evolution Explorer, *Astrophys. J.*, **619**, L7–L10.

Musso, C.; Chiappetti, L.; and Bignami, G.F. (1995) Optimizing mission science: The read-out modes of the EPIC for X-ray astronomy, *Experimental Astronomy*, **6**, 235–248.

Siegmund, O.H.W. (ed.) (2007) *UV, X-Ray, and Gamma-Ray Space Instrumentation for Astronomy XV: Proc. of SPIE*, Vol. 6686.

Stern, R.A.; Liewer, K.; and Janesick, J.R. (1983) Evaluation of a virtual phase charge-coupled device as an imaging x-ray spectrometer, *Review of Scientific Instruments*, **54**, 198–205.

Timothy, J.G. (1991) Photon-counting detector systems: Current status and future prospects, in B.L. Morgan (ed.), *Photoelectronic Image Devices*, Institute of Physics Conference Series No. 121, Institute of Physics, Bristol, U.K.

Ulmer, M.P. (2006) Future UV detectors for space applications, *Proc. SPIE*, **6189**, 24.

Vallerga, J. (2007) Presentation given at *UV Detectors in Space: WSO Detector Workshop, Leicester University, December 2007*, private communication.

Weekes, T.C. (2003) *Very High Energy Gamma Ray Astronomy*, Institute of Physics, Bristol, U.K.

Web sites

FUSE home page: *http://fuse.pha.jhu.edu/*

GALEX home page: *http://www.galex.caltech.edu/index.html*
HESS home page: *http://www.mpi-hd.mpg.de/hfm/HESS/HESS.html*
HST instruments: *http://hubblesite.org/the_telescope/nuts_.and._bolts/instruments/*
SOHO home page: *http://sohowww.nascom.nasa.gov/*
CXRO Center: *http://chandra.harvard.edu/*
INTEGRAL home page: *http://www.esa.int/esaMI/Integral/*
SWIFT home page: *http://heasarc.nasa.gov/docs/swift/swiftsc.html*
VERITAS home page: *http://veritas.sao.arizona.edu/*

Suggested further reading

Barstow, M.A.; and Holberg, J.B. (2003) *Extreme Ultraviolet Astronomy*, Cambridge University Press, Cambridge, U.K.

Charles, P.A.; and Seward, F.D. (1995) *Exploring the X-ray Universe*, Cambridge University Press, Cambridge, U.K.

Chown, M. (1996) *Afterglow of Creation*, University Science Books, Sausalito, CA.

Drew, J.E. (ed.) (1995) *New Developments in X-ray and Ultraviolet Astronomy*, Pergamon Press, Oxford, U.K.

Fraser, G.W. (1989) *Detectors for X-ray Astronomy*, Cambridge University Press, New York.

Gomez de Castro, A.I.; and Wamstekker, W. (eds.) (2006) *Fundamental Questions in Astrophysics: Guidelines for Future UV Observatories*, Springer-Verlag, Dordrecht, The Netherlands.

13

Electronic imaging at sub-millimeter and radio wavelengths

Photon detectors can operate into the far-infrared, but a point is reached where no suitable shallow-doped materials exist and a transition is required to thermal detectors for the sub-millimeter and coherent detectors for the radio. Even so, the creation of two-dimensional arrays of detectors is still possible in principle. Recent technology developments in the sub-millimeter and millimeter bands have led to cameras with moderately high pixel densities. Aperture synthesis methods at longer wavelengths allow high-resolution mapping using multiple telescopes. Microwave images of the entire sky by the COBE and WMAP satellites have revealed the cosmic background in great detail. In this chapter we review the devices and techniques for creating images at these wavelengths.

13.1 INTRODUCTION TO RADIO ASTRONOMY

In 1932, after about four years of work studying background "static" or noise in ship-to-shore communications at a radio wavelength of 15 meters, a young radio engineer named Karl Jansky working at the Bell Telephone Laboratories in Holmdel, New Jersey (U.S.A.) (the same laboratories from which would later come the invention of the transistor, the CCD, and the discovery of the cosmic microwave background) realized that a certain kind of radio noise developed a peak approximately once every 23 hours 56 minutes. The signal seemed strongest when the constellation of Sagittarius was high in the sky. As the center of the Milky Way Galaxy lies in the direction of Sagittarius, Jansky correctly concluded that he was detecting radio waves from outer space. Unfortunately, Jansky's work went unnoticed by professional astronomers, but not by an engineer in Illinois, named Grote Reber. During the period 1936–1944 Reber completed a map of the radio emission from the Milky Way using a "backyard" concave dish 9.1 m (about 30 ft) in diameter and "tuned" to a wavelength of 1.87 m. The antennas used by Jansky and Reber are on display at the

National Radio Astronomy Observatories (NRAO) in Green Bank, West Virginia (U.S.A.). Of course, the development of radar during World War II (1939–1945) stimulated the technology and very soon afterwards "radio observatories" began to appear all over the world.

Reber's radio dish may seem quite large by comparison with "backyard" optical telescopes (Jansky used a large rotating assembly of linear antennas or aerials, rather than a dish) but despite its physical size, the angular resolution of this and many radio telescopes is worse than the human eye. Angular resolution for a radio telescope is usually controlled by the wave phenomenon of diffraction for which the Rayleigh criterion gives

$$\theta = 1.22\frac{\lambda}{D}\text{radians} \approx 70^\circ\frac{\lambda}{D} \tag{13.1}$$

Example. For a radio wavelength of $\lambda = 1$ m and a telescope diameter $D = 10$ m, the angular resolution is only 7° on the sky. At a wavelength of 1 mm the resolution of a 10 m telescope improves by a factor of 1,000 to 25″, if the surface of the dish is smooth enough.

Radio astronomy began with equipment for detecting electromagnetic waves with wavelengths of about 1 meter. Large macroscopic wavelengths (from ~0.3 mm to ~30 m) enable groups of charged particles to produce coherent emission with fixed-phase relationships between waves, which accounts for the extraordinary brightness of pulsars at radio wavelengths. And because radio wavelengths are much larger than interstellar dust grains, scattering is negligible and so the radio sky is dark both day and night, and the neutral interstellar medium is transparent. Stimulated by the huge military and civilian demand for communications, radio-receiving equipment improved and observations were extended to the centimeter band and most recently to millimeter and sub-millimeter wavelengths. The sub-millimeter regime lagged behind mainstream radio astronomy largely because of lack of technology development outside of astronomy. For example, the communications industry abandoned millimeter waveguides in favor of fiber optics in the 1970s. All that has changed now, and sub-millimeter and millimeter astronomy is one of the fastest growing areas in astronomy. For example, one of the largest astronomical projects to date is the international development of the Atacama Large Millimeter Array (ALMA). The immense importance of this (sub-millimeter and millimeter) waveband lies in the fact that numerous molecules have strong emission lines in this region of the spectrum, making it ideal for mapping the cold molecular gas clouds from which new stars are born.

The "window" for ground-based radio observations is quite large. For wavelengths less than 2 cm, atmospheric water vapor begins to attenuate radio signals and high-altitude sites are obligatory. The warm (~300 K) attenuating atmosphere also emits radio noise that degrades sensitivity. For example, emission by water vapor above Green Bank precludes summer observations at this national facility for wavelengths below 3 cm. On the other hand, wavelengths longer than 10 m–20 m

suffer absorption and scattering in the Earth's ionosphere. As the radio waveband is used extensively for a wide range of communication purposes, it has become necessary to regulate that certain wavebands are allocated purely for radio astronomy; a list can be obtained from the Federal Communications Commissions (FCC) or from the National Radio Astronomy Observatories (NRAO). For example, the band from 1,400 MHz to 1,427 MHz includes the 21 cm line of hydrogen, a critical diagnostic for mapping the distribution of this otherwise invisible gas.

It is customary in radio astronomy to use frequency (ν) rather than wavelength (λ), but the two are of course easily interchanged using the relationship

$$\nu\lambda = c \equiv 2.9979 \times 10^8 \text{ m/s} \tag{13.2}$$

where c is the speed of light. The meter waveband corresponds to frequencies lower than 300 megahertz (MHz), the band from 1 cm to 10 cm corresponds to frequencies from 30 gigahertz (GHz) to 3 GHz, and the millimeter and sub-millimeter waveband corresponds to frequencies above 300 GHz. The far-infrared wavelength of 100 μm (0.1 mm) corresponds to 300 terahertz (THz). The range from 30 MHz to 300 MHz is also called VHF (Very High Frequency) and the range from 300 MHz to 3 GHz is called UHF (Ultra High Frequency). Most classical radio astronomy occurs in the 1 GHz–30 GHz range and this is also called the microwave region. There are named bands within this region including the L-band at 1.5 GHz (20 cm), the S-band at 3 GHz (10 cm), the X-band at 10 GHz (3 cm), the Ku-band (Ku = K underband) at 15 GHz (2 cm), and the K-band at 30 GHz (1 cm).

Example. Using the rule of thumb that 1 mm is equivalent to 300 GHz (as opposed to 299.79 GHz), what is the frequency of the 6 cm band used by the Very Large Array (Socorro, New Mexico)? As 6 cm = 60 mm, it is 60 times larger than 1 mm and therefore the frequency will be 60 times smaller: $\nu(\text{GHz}) = 300\text{ GHz}/\lambda\text{ (mm)} = 300/60 = 5\text{ GHz}$.

The power received at a unit surface element per unit frequency (or wavelength) interval is called the flux density S_ν (or S_λ) and is usually measured in $\text{W m}^{-2}\text{ Hz}^{-1}$ (or W m-3); an alternative terminology is spectral irradiance. Radio astronomers also use a flux unit called the jansky; $1\text{ Jy} = 1\text{ flux unit} = 10^{-26}\text{ W m}^2\text{ Hz}^{-1}$. In the 1940s, a sensitivity of a few jansky was considered good. Nowadays, signal strengths are measured in millionths of a jansky (μJy) in some cases. For an extended source the brightness (b_ν) is the flux density per unit solid angle measured in jansky per steradian. The total power (P) collected by the antenna (ignoring efficiency factors and polarization state) is given by

$$P = b_\nu A \Omega_b \, \Delta\nu \tag{13.3}$$

where A is the collecting area; Ω_b is the solid angle of the telescope beam on the source; and $\Delta\nu$ is the frequency bandwidth. For a radio telescope, the $A\Omega$ product or étendue is given by $A\Omega = \lambda^2$ because of diffraction.

In the radio region we always have $hc \ll \lambda kT$ which leads to the Rayleigh–Jeans form of the Planck function $B_\nu = 2kT/\lambda^2$, where the constants have their usual

meaning. The temperature (T) in this equation is the true thermal temperature of the source in this case. However, it is convenient to use this relationship to define a "brightness temperature" T_B for *any* source given by

$$T_B = \frac{L_\nu \lambda^2}{2k} = \frac{P}{2k\,\Delta\nu} \tag{13.4}$$

The quantity $T_A = P/k\,\Delta\nu$ is called the "antenna temperature" corresponding to the observed power and $T_A = T_B(\Omega_s/\Omega_b)$ where Ω_s is the source size. Receiver performance is usually specified in terms of a noise temperature T_N which is the thermal temperature of a resistor that generates the same detected power as the system. The power (P_N) generated by these fluctuations is

$$P_N = kT_N\,\Delta\nu \tag{13.5}$$

The minimum possible noise temperature is given by the quantum limit

$$T_N(\min) = h\nu/(2k) \tag{13.6}$$

At 30 GHz (1 cm) this number is only 0.72 K. The ratio of photon energy $h\nu$ to thermal energy kT is always very low at radio frequencies.

At optical and near-infrared wavelengths electronic imaging of astronomical sources can be obtained directly using a two-dimensional grid of detector pixels in the telescope focal plane. Each pixel absorbs photons and converts the energy to electrical charge. Because this photon detection process completely ignores the wave nature of light, it is said to be "incoherent". The lack of photon detectors for wavelengths beyond about 150 μm leads to the use of a thermal detector (bolometer) which measures total absorbed power. At sub-millimeter wavelengths it is now possible to use very sensitive discrete bolometers arranged in a tightly stacked array for imaging. At even longer wavelengths the detection and measurement of radio signals relies primarily on coherent processes in which devices respond directly to the electric field strength in the wave and preserve phase information. Coherent detectors work by interference of two electric fields: one in the incoming wave and another from a man-made source called a local oscillator. The radio wave is collected by the "antenna" and the device that produces the signal is the "receiver". Radio receivers that function by mixing signals of different frequency are called "heterodyne" receivers. The word heterodyne is derived from the Greek roots *hetero*, meaning "different", and *dyne*, meaning "power", and was pioneered by Canadian inventor-engineer Reginald Fessenden (1866–1932) who made the first audio transmission by radio in 1900. The key is the mixing device, and even here there is an opportunity to use many elements as we will see. In addition, it is common for well-separated radio telescopes to link together using radio communications to form a large array, even a transcontinental array, and thereby achieve a much higher effective angular resolution than an individual telescope.

Obviously, radio astronomy is a vast subject with an enormous amount of literature. Most universities teaching astronomy will have separate courses for fields like radio astronomy. Our focus here is a basic look at some of the latest detector systems that are enabling astronomers to produce electronic images (and spectra) at

these wavelengths. More detail can be obtained from the web pages of the national radio observatories and classic texts like Kraus (1986). For an excellent up-to-date textbook see *Tools of Radio Astronomy* by Rohlfs and Wilson (2004).

13.2 RADIO TELESCOPES

13.2.1 Antennas

The simplest form of antenna (also called an aerial) is the classical half-wave dipole which consists of two conducting rods (usually copper), each one-quarter of the desired wavelength and separated by a small gap with coaxial cable going from the closer spaced ends to a receiver. Free electrons in the conductors are set into motion by the incoming electromagnetic wave, generating an alternating electric current at the same frequency as the incoming wave. Although it may seem rather pointless to use a dipole antenna, because it is sensitive to only a rather narrow range of frequencies, some specific frequencies are of great interest such as the 1,427 MHz (21 cm) line of hydrogen which has enabled astronomers to map the structure of the Milky Way. More important is the fact that the half-wave dipole has a broad point source function, which radio astronomers call the ''antenna response'' function or pattern. The strongest response is for a plane-polarized wave with a direction of propagation perpendicular (broadside) to the conductors because the electric field in the wave is parallel to the rods in that case. Let's take this to be the optical axis. There is no response to waves coming along the axis of the rod (which makes the electric field of the wave perpendicular to the rod and so no current can flow along the rod). The response function is represented by a polar diagram which plots the response as a function of angle on the sky, and there is a reciprocity theorem that says that the transmitting and receiving pattern should be the same. It can be shown that the power pattern of a dipole depends on the field angle θ as $P \sim \sin^2 \theta$. Here the $\sim$ sign means proportional to in this case. Therefore, the antenna gain is given by $G = G_0 \sin^2 \theta$ and for a half-wave dipole $G_0 = 3/2$; thus:

$$G(\theta, \varphi) = (3/2) \sin^2 \theta \tag{13.7}$$

In Equation (13.7), φ is the azimuthal angle about the optical axis and in the plane perpendicular to the optical axis. In general, an antenna having a peak gain $G_{\max}$ must beam most of its power into a solid angle $\Omega \sim 4\pi/G_{\max}$. The beam width at the half-power (BWHP) points represents the equivalent term to full width at half maximum (FWHM) for the Airy diffraction pattern or the seeing disk at visible wavelengths. Note that the Airy function for an optical telescope is usually represented by an (x, y) plot, but it could also be drawn as an (r, θ) plot, in which case it would be a polar diagram with sidelobes just like the standard radio antenna pattern; the sidelobes are tiny and very close to the main beam because λ/D is so small at optical frequencies.

In general, a dipole has rather poor directional sensitivity. By combining the signals from a stacked array of dipoles spaced at half-wavelength intervals, directionality is much improved. However, as this combination behaves like the slits of a

diffraction grating, interference causes additional secondary peaks or sidelobes to appear in the polar pattern. Sidelobes are the equivalent of the secondary maxima in the Airy diffraction pattern, and they imply that the antenna will detect radiation at high angles of incidence to the main beam. The angular width of the main beam between the first minima (nulls) for n end-to-end elements is given by $\sin a = 1/n$, whereas for a broadside array of elements it is $\sin a = 2/n$. Even more directionality can be obtained by adding extra rods which are coplanar with the dipole (usually several in front and one behind acting as a pure reflector), but *not* electrically connected to the dipole or the receiver, and hence are said to be "parasitic". This is called a parasitic antenna or Yagi–Uda antenna (after its inventors Shintaro Uda and Hidetsugu Yagi of Tohoku Imperial University, Japan in 1926). The reflector element concentrates the power into the forward direction.

The effective collecting area (A_e) of an antenna is defined in terms of the collected power ($P_1 = P_\nu \, \Delta\nu$) compared with the flux density S_1 of that part of the radio wave whose polarization coincides with the antenna. Thus, $P_1 = A_e S_1 \, \Delta\nu$ and for random polarization $S_1 = (1/2)S$, and therefore the power extracted is $P = (1/2)A_e S \, \Delta\nu$. It can be shown that the effective collecting area of all lossless antennas is given by:

$$A_e = \lambda^2/4\pi \tag{13.8}$$

This remarkable result applies to the simple dipole antenna or to a large parabolic dish antenna. Effective collecting area and gain are related by

$$A_e(\theta, \varphi) = \lambda^2 G(\theta, \varphi)/4\pi \tag{13.9}$$

An obvious disadvantage of the directional antennas just described is the fact that the wavelength is at once fixed by the choice of the length of dipole. A much more versatile approach is to use a large parabolic dish of metal panels as the reflector. The dish will collect more energy and bring the radio waves to a focus, effectively narrowing the power pattern so that $\theta_{\mathrm{HPBW}} \sim \lambda/D$. In practice, $1.2\lambda/D$ is a good rule of thumb. A dish can increase the maximum gain and effective area, which can in fact approach its normal geometric area of $\pi D^2/4$. As long as the diameter of the dish is large compared with the wavelength, then the rules of geometric optics still apply and the parabolic shape is best. Selectable dipole feeds (or, better, a collecting horn and waveguide that channels the wave to a probe at the focus) can be used. Prime focus can be used, but so too can the Cassegrain and Gregorian foci by using convex and concave secondary reflectors, respectively. Most large radio telescopes must have a very fast focal ratio $f/D \sim 0.4$ to ensure that the support structure for the secondary or "sub-reflector" is not unreasonably large. As mentioned previously, to obtain an angular resolution on the sky of $\sim 0.7^\circ$ requires that the ratio of the diameter of the dish to the observing wavelength must be at least 100 : 1. To ensure diffraction-limited performance the surface must be smooth to $\sim\lambda/20$ but, as the wavelength is now much larger, the actual physical deviations from a smooth parabolic surface can be correspondingly bigger, ranging from 50 μm rms at 1 mm to 5 mm rms at 10 cm. Therefore, the world's largest steerable parabolic dishes are very large. For example, the largest single-dish radio telescope in the world is the partially steerable RATAN-600 built in 1977 in present-day Russia which has a 576 m diameter circle of

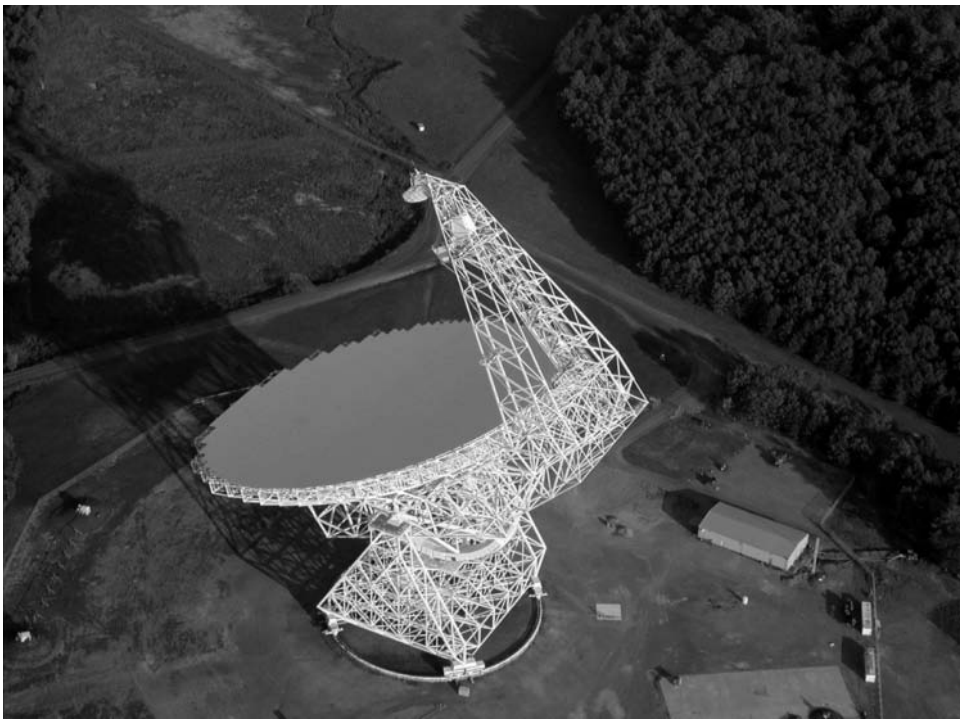


Figure 13.1. The 100 m Green Bank Telescope (GBT) of the U.S. National Radio Astronomy Observatories (NRAO) in West Virginia. Credit: NRAO/AUI.

rectangular radio reflectors. The Effelsberg Radio Telescope located 40 km south of Bonn in Germany is 100 m in diameter. It has a high-quality polished surface of metal plates and typically works at wavelengths around 6 cm (4,996 MHz). At shorter wavelengths the surface accuracy degrades the collecting efficiency. The Robert C. Byrd Green Bank Telescope (GBT) in West Virginia (U.S.A.), which became operational in 2000, has a surface area of 100×110 m. Its 2,004 panels are made from aluminum with a surface accuracy of better than 76 μm rms, and 2,209 actuators adjust the panel to correct for distortions due to gravity as the telescope moves. The dish of the GBT is an off-axis parabola which means that instruments at the prime focus do not obscure the beam (Figure 13.1). The IRAM Millimeter Radiotelescope at Pico Veleta in Spain is a carbon fiber structure 30 m in diameter with panels machined to an average surface precision of 100 μm, which permits diffraction-limited performance at a wavelength of about 2 mm. Significantly better surface accuracy (30 μm) is obtained by the 15 m James Clerk Maxwell Telescope (Figure 13.2) on Mauna Kea, Hawaii, and better still (15 μm) with the 10 m dish and panels of the Heinrich Hertz Sub-Millimeter Telescope Observatory on Mount Graham, Arizona, a joint project between the University of Arizona and the Max Planck Institute for Radio Astronomy in Bonn. Made from carbon fiber reinforced plastic

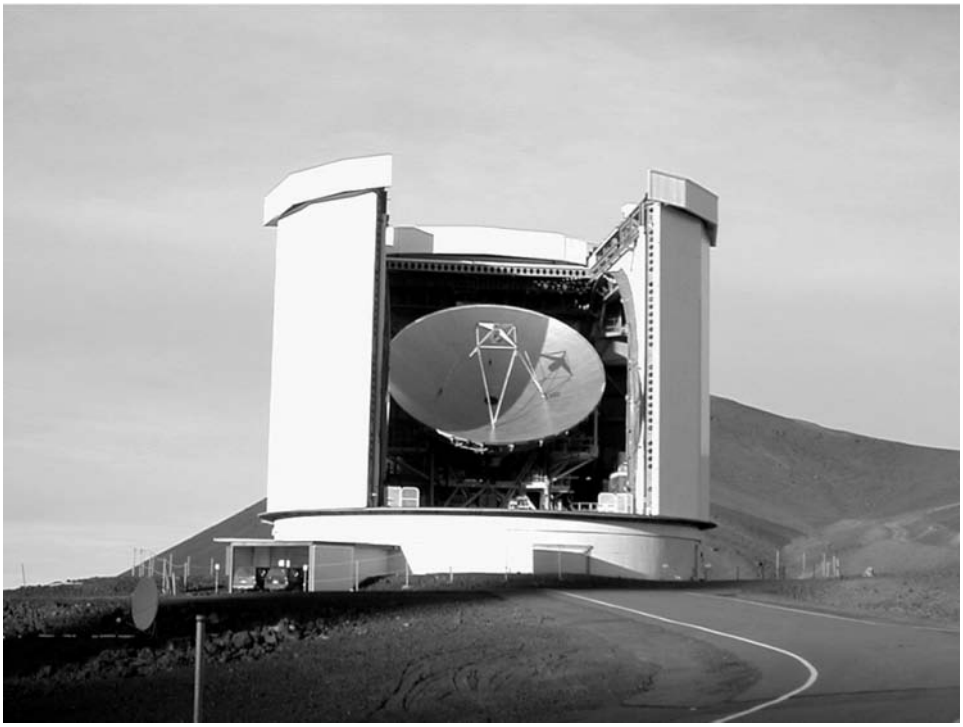


Figure 13.2. The 15 m James Clerk Maxwell sub-millimeter telescope (JCMT). Credit: JCMT.

that is 20 times less sensitive to thermal change than polished metal panels, this dish permits the telescope to work at submillimeter wavelengths as short as 0.350 mm. The largest non-steerable radio telescope is the 300 m dish at Arecibo, Puerto Rico, and the first very large steerable dish was the 76.3 m (250 ft) Lovell Telescope at Jodrell Bank, U.K. built in 1957 and famous for detecting Sputnik 1.

13.2.2 Receivers

Typically, a wave received at the focus of a radio telescope enters a "feed" which may be a direct dipole or a "corrugated horn" which matches a waveguide and feeds a resonant cavity, which in turn defines a frequency interval or bandpass near the frequency of the wave. The horn suppresses sidebands and couples the radiation onto the much smaller detection system. At the base of the horn is a dipole or a pair of crossed dipoles (to detect both polarizations). Enormous amplification of the detected signal is needed (10^{14}–10^{19}), which therefore implies a cascade of amplifiers, but the initial one, the "pre-amplifier" is critical. An example of a horn is shown in Figure 13.3 and the typical arrangement of components (for centimeter wavelengths) is shown in Figure 13.4.

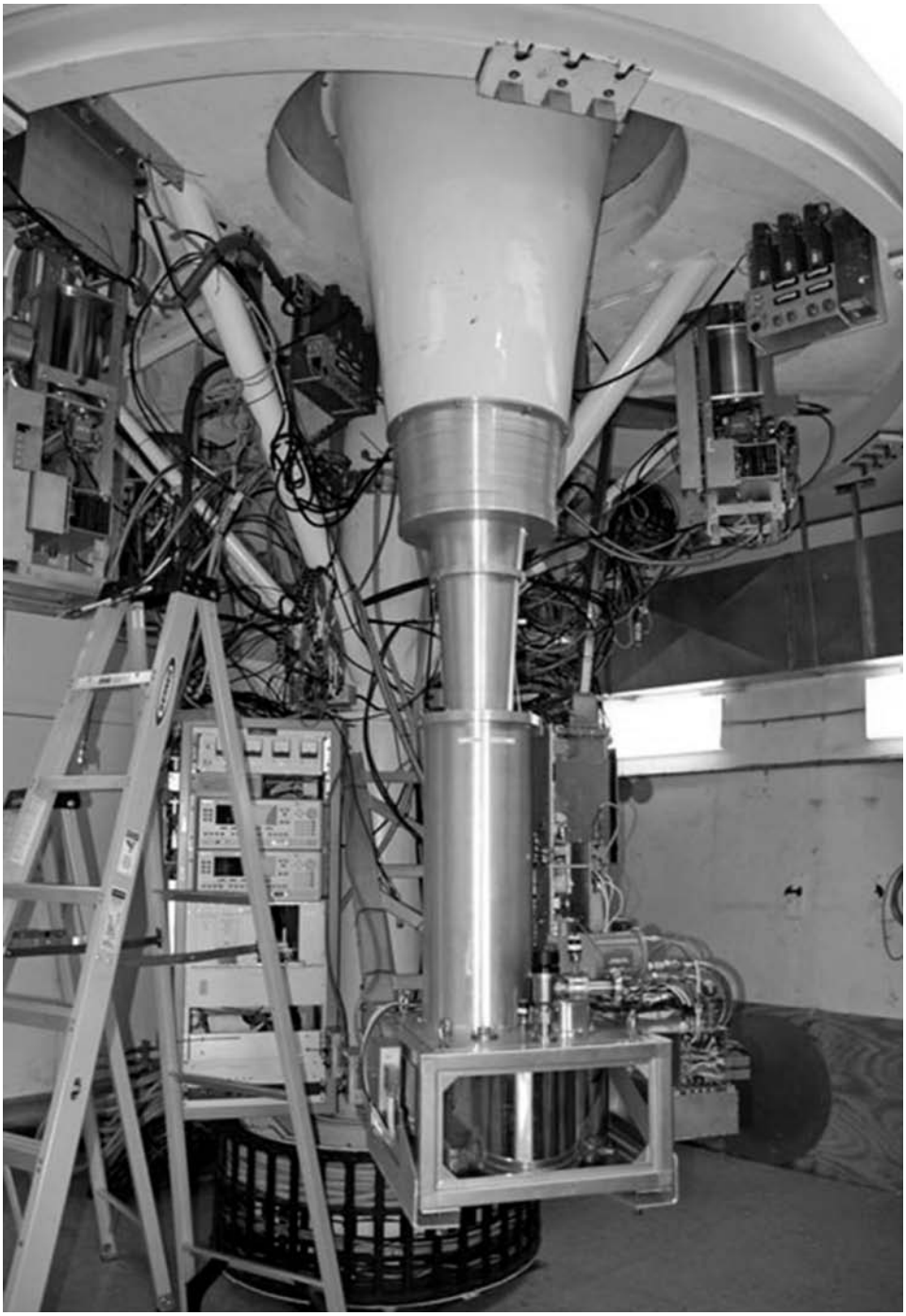


Figure 13.3. A feed horn at the Gregorian focus of the GBT. Credit: NRAO/AUI.

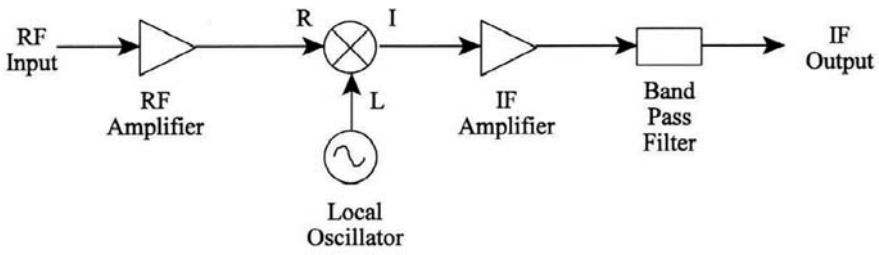


Figure 13.4. Basic layout of a heterodyne radio detection system showing antenna, mixer, local oscillator, IF amplifier, and detector.

The simplest radiometer measures the average total power received over a well-defined radio frequency bandwidth $\Delta\nu$ and over a time interval τ. Just as for optical and infrared wavelengths, the weak astronomical source is measured against a background of many other radio signals such as the cosmic microwave background, the atmosphere, and the noise in the receiver itself. Power is usually expressed as a temperature, and the total system power is T_{sys}. The total noise in the measurement is given by the practical form of the radiometer equation:

$$\sigma_T = T_{\text{sys}}[(1/\Delta\nu_{\text{RF}}\tau) + (\Delta G/G)^2]^{1/2} \tag{13.10}$$

where ΔG represents possible fluctuations in gain. If those fluctuations are negligible ($< 10^{-3}$) then the equation simplifies to its ideal form ($T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$). In a manner similar to chopping in the infrared, one way to minimize fluctuations in receiver gain and atmospheric emission is to perform differential measurements by switching rapidly between two adjacent feeds, as first suggested by Robert Dicke (1916–1997) in the 1940s. The main drawback of Dicke switching is that the measured noise is doubled to $2T_{\text{sys}}/\sqrt{(\Delta\nu_{\text{RF}}\tau)}$ as a result of the difference measurement.

Nearly all practical radiometers are more complex. As is well-known from elementary physics, if two signals of different—but similar—frequencies are added they produce a signal at the "beat" frequency, which is the difference between the two original frequencies and is therefore a much lower frequency. While the resulting "mixed" signal contains frequencies only from the original two signals, its amplitude is modulated at the difference, or beat frequency. Heterodyne receivers measure this amplitude. Effectively, the mixer device multiplies the RF signal by a sine wave of frequency ν_{LO} generated by a local oscillator (LO). To illustrate the effect consider the product of two sine waves where t is time:

$$2\sin(2\pi\nu_{\text{LO}}t) \times \sin(2\pi\nu_{\text{RF}}t) = \cos[2\pi(\nu_{\text{LO}} - \nu_{\text{RF}})t] - \cos[2\pi(\nu_{\text{LO}} + \nu_{\text{RF}})t] \tag{13.11}$$

The difference frequency $\nu_{\text{LO}} - \nu_{\text{RF}}$ is called the intermediate frequency (IF). At microwave frequencies the mixer comes first because low-noise amplifiers are difficult to design and the LO signal is fed directly into the horn (waveguide) along with the antenna signal. Otherwise, as shown in Figure 13.4, the mixing is done after the signal becomes an electrical current. The advantage of heterodyne receivers lies in the "down-conversion" of the frequency from the large (GHz or MHz) radio frequencies

to the much lower intermediate frequency (IF) range (kHz), where conventional electronics can be used. In addition, control over the RF range being covered depends only on tuning the oscillator, and therefore back-end devices following the un-tuned IF amplifier can operate over fixed frequency ranges. An ambiguity exists in the sign of the difference signal. It is not possible to tell whether the true frequency was larger or smaller than the local oscillator frequency. Because the IF signal can arise from a combination of two possible inputs, it is called a "double-sideband" or DSB signal. This is a serious problem for observations of spectral lines at radio frequencies, and therefore it is usually desirable to operate in a "single-sideband" or SSB configuration if possible by using a narrow bandpass rejection filter in front of the receiver or by tuning the mixer.

Again, even if there are no sources in the beam, a non-zero current will be measured. This signal arises from various causes such as residual thermal radio emission from the atmosphere, telescope, or waveguide; thermal radio emission from the ground detected in the sidelobe pattern of the antenna; thermal noise in the detector itself. The noise is equivalent to a small amount of power and can be given an effective noise temperature as mentioned earlier which represents the background power against which the signal is to be detected. If the receiver operates in the double-sideband (DSB) mode then the bandwidth for the noise measurement is generally $\Delta\nu = 2\Delta\nu_{\mathrm{IF}}$. In the best case the limit is imposed by the detector itself and cannot be lower than the physical temperature of the device. Noise temperatures in the range 10 K–200 K are typical, with 50 K being typical at wavelengths of 6 cm to 21 cm. Figure 13.5 shows a four-beam cryogenic (20 K) receiver and feed horns for the GBT Q-band receiver (40 GHz–52 GHz).

The mixer has to be a non-linear circuit element that converts power from the original frequencies to the beat frequency. This element is usually a diode for the following reason. The power (P) in an electromagnetic wave is proportional to the square of the electric field strength (E^2), and the amplitude of the signal voltage (V) is proportional to E. Therefore, the power is proportional to V^2 and the ideal characteristic of a mixer would give an output current which is proportional to input power (in other words, I proportional to V^2). Such a device is called a "square law" mixer. In a practical diode mixer, the I–V curve can be represented by a dc term plus a linear term (which gives no net response) and higher order terms which are quadratic and cubic in the voltage difference ΔV. Over a suitable range of the curve I is proportional to V^2. The output of the mixer is a signal with a frequency of $\nu_{\mathrm{IF}} = \pm(\nu - \nu_{\mathrm{LO}})$, where ν_{LO} is the frequency of the local oscillator, which is passed to the intermediate frequency (IF) amplifier and then to a rectifying and smoothing circuit called a "detector". This terminology seems quite strange compared with our previous usage where the detector is the device that receives the photons and converts them to an electrical signal. The detector part of the receiver is usually called the "back-end" and can be quite complicated with a number of options, even for the same "front-end" system.

Rather than a single detector, a "back-end" spectrometer often receives the output from the IF amplifier. The spectrometer could consist of a number of electrical "filters" tuned to different frequencies with detectors on their outputs, a digital

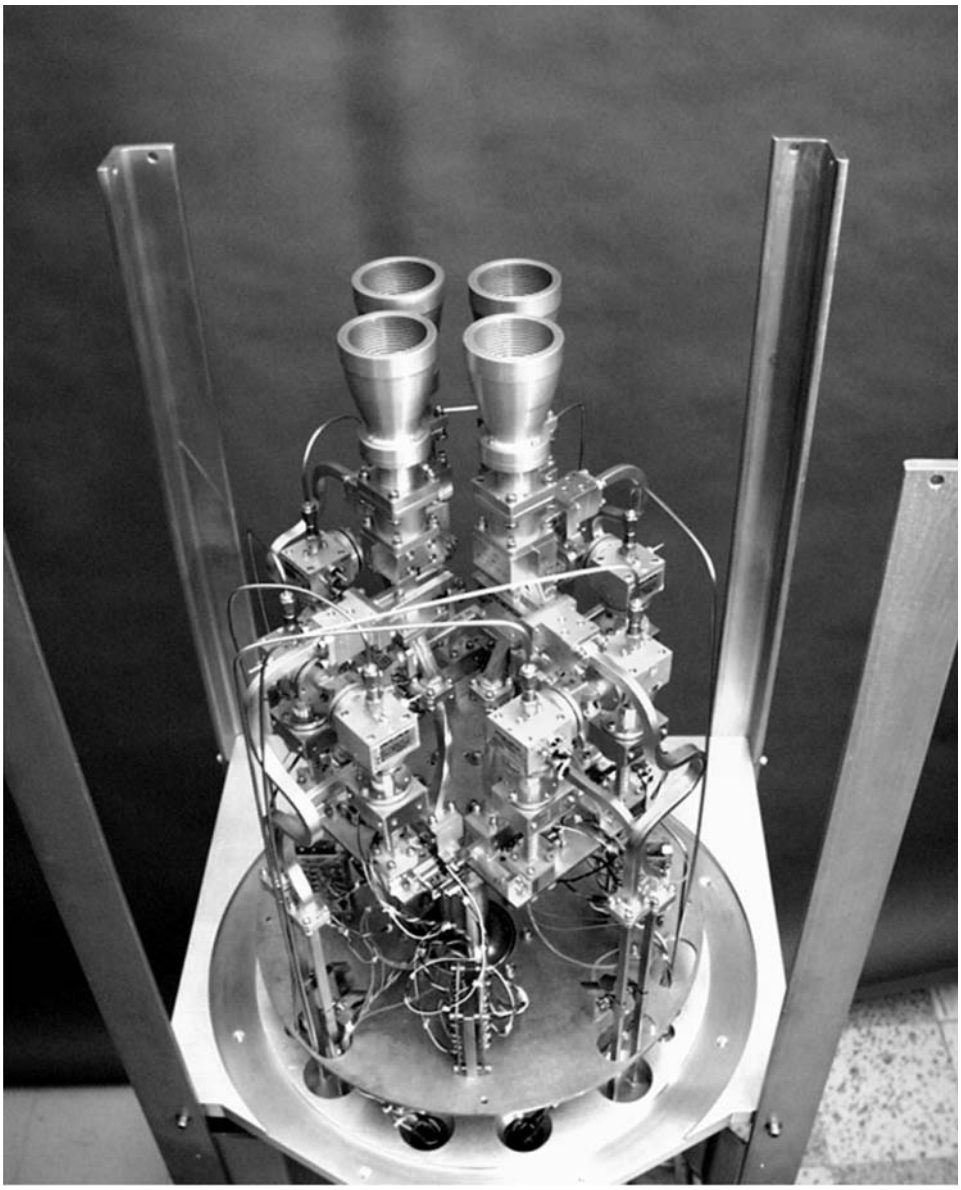


Figure 13.5. A four-beam cryogenic receiver for the Q-band (40 GHz–52 GHz) at the Green Bank 100 m telescope. Credit: NRAO/AUI.

correlation computer, or an acousto-optic spectrometer (AOS) such as a Bragg cell, where the frequencies are converted to ultrasonic waves that disperse a monochromatic light beam (e.g., a laser) onto an array of visible light detectors. For example, the Green Bank Telescope Spectrometer provides the observer with a

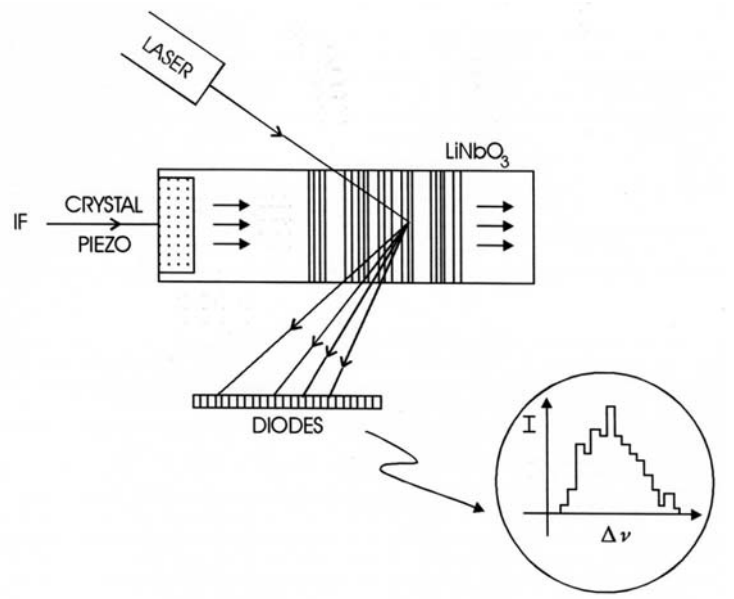


Figure 13.6. The principle of the acousto-optic spectrometer.

variety of spectral line observing modes. This spectrometer is a modular system, with four quadrants. Quadrants may be used independently or grouped together into banks of 1, 2, or 4 quadrants to provide the observer with 1 to 3 different levels of spectral resolutions for each observing mode. When the four quadrants are independently operated, they can be configured to acquire data at up to eight different frequencies. The spectrometer performs digital autocorrelations of the input signals. These input signals may be both polarizations in a selected bandwidth centered on a specified spectral line, or both polarization inputs from different feeds of multi-feed receivers, or combinations of these in different spectral windows. Spectrometer modes are divided into two major types: wide bandwidth, low resolution and narrow bandwidth, high resolution. Thus, at radio wavelengths the "spectrometer" is often a cabinet full of electronics performing fast calculations. Large single-dish radio telescopes are often used to study spectral line emissions from extended and point sources.

The principle of the acousto-optic spectrometer (AOS) is illustrated in Figure 13.6. Here, the IF signal is applied to a piezo-optic device which sets up an ultrasonic wave in a lithium niobate crystal that is illuminated by a helium–neon (He–Ne) laser beam. If the IF signal is monochromatic then all the energy is diffracted into one pixel of the CCD or diode array. For a more complex distribution of frequencies in the IF signal the intensity pattern spread over the diodes gives the spectrum.

In most receivers the non-linear element is a significant noise source, and amplification of the signal or the IF signal is needed. Various types of amplifiers are possible. Parametric amplifiers are essentially *RLC* circuits with a resistor (R),

inductor (L), and capacitor (C). Typical noise temperatures for a cooled parametric amplifier are in the range 10 K–20 K. Maser amplifiers are based on the use of an external power source to fill the excited electron energy levels of a solid or a gas above the normal population levels expected when the material is in thermal equilibrium. The stimulated emission produced by the incident wave is in phase with it and in the same direction, and therefore we have amplification. One common form of maser is the ruby maser which is a single crystal of Al_2O_3 with 0.05% of the Al^{3+} ions replaced by chromium ions Cr^{3+}. The typical noise temperatures of these amplifiers are 15 K–50 K.

As the function of the local oscillator is to provide a coherent, noise-free signal at a frequency very close to the signal frequency it must provide sufficient power. Local oscillators employ either klystrons, which are wave generators for frequencies below 180 GHz (developed mainly for telecommunications and radar); or the carcinotron, a wave generator which works up to 1,000 GHz but is relatively heavy and power-consuming; harmonic generators are frequency multipliers which use a non-linear element like a Schottky diode which radiates harmonics at twice, three times (and so on) the frequency when pumped by a klystron. The power obtained reaches several milliwatts up to about 500 GHz, and this is sufficient for modern mixers.

13.2.3 Detectors/Mixers

No quantum detector of the kind we have discussed repeatedly in this book (which uses the liberation of a (photo)electron by an incident photon) can work at wavelengths longer than about 0.2 mm. Across the entire radio spectrum, the electromagnetic field or the current which it induces in an antenna is applied to a non-linear element (diode) or mixer. The mixer either measures the total power or changes the signal frequency to one which is more easily measured. One of the most frequently used devices is the Schottky diode, and another very exciting development, especially at millimeter wavelengths, is the superconducting junction. We will consider each in turn.

When a metal and a semiconductor are brought into contact, the majority charge carriers of the semiconductor leave the contact zone until the Fermi levels of the metal and semiconductor are equalized. As a result, a "barrier" or "depletion region" empty of majority carriers appears in the semiconductor; typical barrier widths are $\sim$0.3 μm. Even without any voltage across the junction, a current can flow through it. In practice the semiconductor used is heavily doped gallium arsenide (GaAs) with a very thin lightly doped layer on the surface to reduce quantum-mechanical tunneling and a contact layer of metal on top. Cooling the detector to a temperature of about 20 K yields a low-noise detector. Above 300 GHz (wavelengths shorter than 1 mm) the capacitance of the diode creates an RC filter which reduces its response.

There has been considerable excitement in the scientific literature in recent years about high-temperature superconductors, but other superconductors have made a big impact on detectors for radio astronomy. When a thin insulating barrier is created between a normal metal and a superconducting metal—note, not a semiconductor—or between two superconductors, the structures are called SIN and SIS junctions.

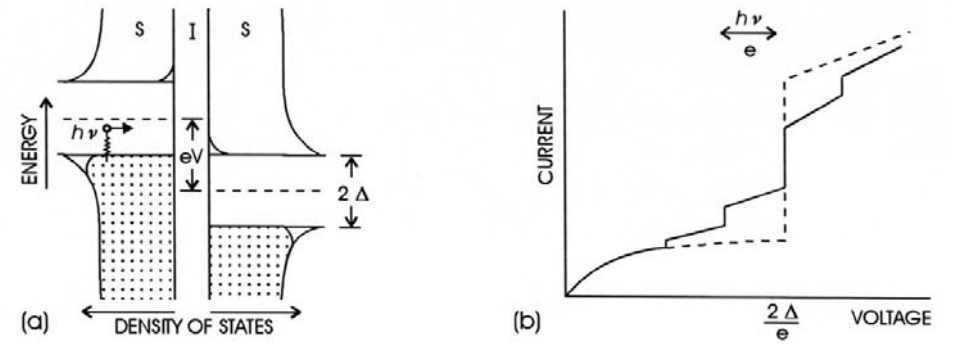


Figure 13.7. (a) The basic structure of the energy levels in an SIS detector. No current flows if $eV < 2\Delta$, but absorption of a photon can excite a charge carrier to the energy where the tunnel effect occurs. (b) The current–voltage behavior without illumination (solid line) and with photons present (dashed line).

Non-linear current can flow by quantum-mechanical tunneling, and hence the device can be used as a mixer. Figure 13.7 shows the basic structure of the energy levels on each side of the insulator. When the voltage V is large enough that occupied states on one side are opposite vacant states on the other side then a tunneling current can flow. Conversely, no current flows if $eV < 2\Delta$. Absorption of a photon can excite a charge carrier to the energy where the tunnel effect occurs. Part (b) of the figure shows the current–voltage behavior without illumination (solid line) and with photons present (dashed line). The high-frequency limit is about the same as the Schottky diode. The Nobeyama Radio Observatory in Japan has a focal-plane array of 5×5 SIS receivers (shown in Figure 13.8).

Lastly, in very pure (low-doped) InSb at liquid helium temperatures, a fraction of the conduction electrons stay free and interact only weakly with the crystalline lattice. The absorption of photons by the electrons raises their temperature and rapidly changes the resistance of the material. Since this change in conductivity is proportional to incident energy which in turn is proportional to the square of the electric field in the wave, we again have a mixer. This device is known as the "hot electron bolometer" or HEB mixer. Its bandpass is limited to about 1 MHz, but it can work at frequencies up to at least 500 GHz. Detailed information about receiver systems and back-end options is available at the web sites of all the major radio observatories.

13.2.4 Interferometers and aperture synthesis

Radio astronomy faces two serious hurdles. First, the antenna plus detector is a "single-pixel" device which makes imaging difficult, and, second, because of the much larger wavelengths involved, high angular resolution can only be achieved by making very large antenna complexes or very large single dish type collectors so that the ratio of wavelength-to-telescope diameter again becomes small. Single-dish telescopes have performed extensive low-resolution surveys of the entire sky in specific wavelength

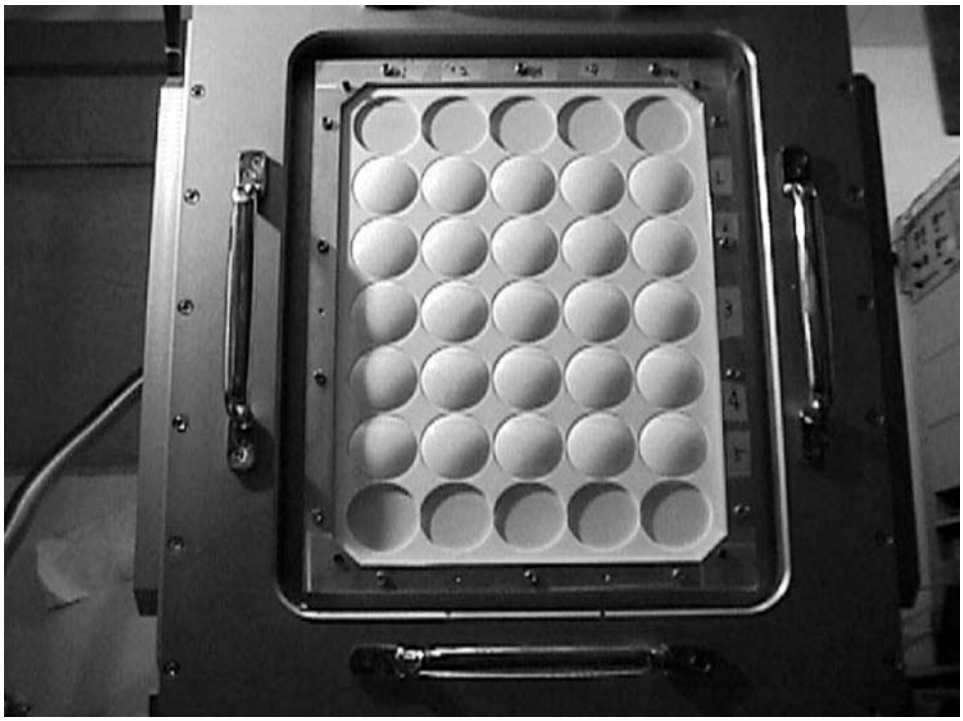


Figure 13.8. An array of 25 SIS receivers located in the focal plane of the Nobeyama Radio Telescope (NRT) in Japan. Credit: NRT.

bands such as the 21 cm line of neutral hydrogen. Nevertheless, even the largest radio dish in existence cannot come close to the resolution of the best optical telescopes.

To solve this problem, radio astronomers introduced the technique of interferometry between multiple telescopes in the late 1940s. In this approach, radio signals are carried over wires between two well-separated radio telescopes and made to "interfere" or mix together in such a way as to provide the angular resolution (though not the collecting area) of a huge telescope with a diameter equal to the separation of the two smaller telescopes. The concept is the same as the "double-slit" experiment and the famous 1920s' experiment carried out by Michelson (Chapter 2) with an optical interferometer having variable spacings between the pair of telescopes to measure the diameter of a star. Figure 13.9 illustrates the principle again for two radio telescopes. A plane wavefront (from a very distant source) arrives at an angle θ to the vertical and intercepts telescope A. At that moment it has an extra distance d to travel before reaching telescope B, and therefore there is a "phase lag" between the two detections. When the two signals are combined in the receiver they will interfere according to the phase lag ($= 2\pi d/\lambda$). If the distance d is an exact multiple of the wavelength, $d = n\lambda$, then the waves will be in phase and interfere constructively to reinforce each other; if d is an integer number of half-wavelengths then the waves will

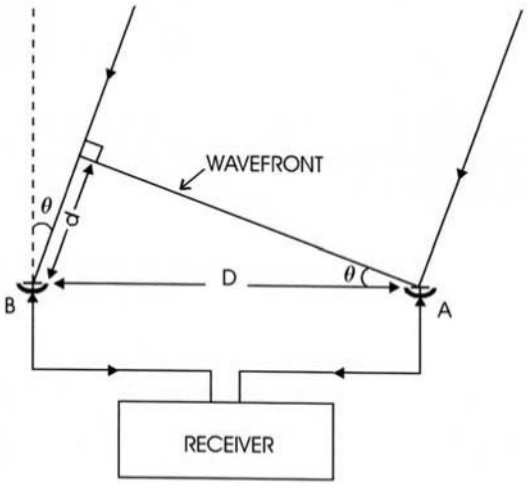


Figure 13.9. The principle of radio interferometry is illustrated by the two antennas observing the same source. There is a phase difference between the arriving wavefronts.

be completely out of phase and cancel each other. If D is the separation of the radio telescopes, then

$$\sin\theta = n\frac{\lambda}{D} \tag{13.12}$$

and a large signal is recorded each time a source moving across the sky above a fixed interferometer satisfies the above constraint. Small-diameter sources such as the Sun, Cygnus A, and Cassiopeia A respond to the interference pattern of an interferometer pair, whereas the diffuse galactic background does not. Radio telescopes separated by entire continents have now been linked together successfully in a technique known as very long baseline interferometry or VLBI. In principle, two radio telescopes on opposite sides of the Earth could resolve celestial sources as small as 0.00001 arcseconds apart!

Radio-linked rather than cable-linked interferometers were developed at both Jodrell Bank, U.K. and in Australia. Among the pioneers were Hanbury-Brown, Mills, and Palmer. Their success led eventually to the isolation of a class of very compact, starlike (quasi-stellar) radio sources, the quasars. Synthesizing the brightness distribution from interference fringes alone cannot be done without ambiguity. Problems of source identification arise because of the strong secondary sidelobes in the simple two-telescope arrangement mentioned above. By using multi-element interferometers *and* by combining two interferometers at right angles, with special electronic switching and phase-shifting techniques, it is possible to create a system with a single narrow beam. This is the principle of the original 450 m (1,500 ft) Mills Cross Telescope built in 1954 about 40 km west of Sydney, Australia. This design is restricted in wavelength coverage and area of sky.

Another approach to high angular resolution with radio telescopes is the method known as "aperture synthesis" which requires a system of two radio telescopes: one

at a fixed location and the other "movable" on a railroad track over a large range in positions. As conceived by Sir Martin Ryle (1918–1984) and his colleagues at Cambridge, the idea was to observe all the interferometer positional pairs which existed within the equivalent large aperture radio telescope, to effectively recreate or synthesize a uniformly filled aperture. The original "Tee synthesis" telescope at the Mullard Observatory, Cambridge, U.K. had a multi-element interferometer along an E–W line as the fixed antenna and a small single-dish telescope which moved along a N–S line. The intention was to sample the whole range of spacings; radio astronomers call this "filling the *u–v* plane". The *u–v* plane is the equivalent in the Fourier transform spatial frequency domain to angular distribution of the source on the sky; spatial separations of antennas are measured in units of wavelength. As the aperture is not filled there will always be some missing spatial frequencies. A clever step was to allow the Earth's rotation to sweep out a range of telescope spacings, simply by the projection effect, without moving either pair of detectors! Ryle shared the Nobel Prize in physics in 1974. Other telescopes began to follow, such as the Westerbork Synthesis Radio Telescope in the northeast of the Netherlands with 14 antennas completed in 1970; this synthesis array underwent a major upgrade in 2000. By 1980 one of the finest systems of radio telescopes in the world began operating in the desert near Socorro, New Mexico. Called the Very Large Array (VLA), it consists of 27 concave dishes, each 26 m (85 ft) in diameter. The 27 telescopes are arranged along the arms of a gigantic Y-shape covering an area 27 km (17 miles) in diameter (Figure 13.10).

It soon became clear that it was not always going to be possible to devote the time to filling in the *u–v* plane. A significant breakthrough was to employ a computational

Figure 13.10. The Very Large Array (VLA) of 27 antennas in New Mexico (U.S.A.). Credit: NRAO/AUI.

technique called the CLEAN process which was introduced by Swedish astronomer Jan Högbom (1974). The principle goes like this. An ideal point source observed by an unfilled aperture (not enough interferometer pairs) will appear as a point of radio emission surrounded by large and widespread sidelobes. If it is the only point source, the pattern will match the theoretical pattern of the unfilled aperture; therefore, subtract the ideal pattern and look at what is left. If another weaker radio source is present, then repeat the process. The result will be a list of point sources and a residual map of random noise. The key to the CLEAN process is that for the majority of radio sources, most of the u–v map and consequently the field of view on the sky is *empty* of sources.

When the CLEAN algorithm started running on radio telescope computers it worked better than anyone expected, even if the source was somewhat extended! The VLA immediately began to produce maps (i.e., radio images) as though it had sampled the u–v plane fully. Finally, the method of "adaptive calibration" was introduced by Readhead (Caltech) and Wilkinson (Jodrell Bank) in 1978 in which the amplitude and phase of a diverse set of telescopes was corrected by examining the response to a point source and iterating until it matched the theoretically expected value.

Aperture synthesis arrays have now been extended to the sub-millimeter band also, with the first being the Smithsonian Submillimeter Array (SSA) consisting of six 6 m telescopes on Mauna Kea. Earlier, the JCMT and the CSO on Mauna Kea were linked to form the first sub-millimeter interferometer. These telescopes are separated by 164 m which yields a fringe-spacing of 1.1 arcsec at 345 GHz. The next major step will be the Atacama Large Millimeter Array in Chile discussed in Chapter 14.

With the introduction of aperture synthesis and very long baseline interferometry, together with antenna and receiver developments and the advent of very fast computers, radio astronomers have been able to completely overcome the initial drawbacks to imaging mentioned earlier, and, indeed, they have been able to "map" areas of the sky with even higher angular resolution than any optical ground-based telescope. The 22 cm radio image of the active galaxy Cygnus A in Figure 13.11 (left) is an example. The optical source corresponds to the small object in the very center of this 2×1 arcmin field. Enormous jets interact with cold material far outside the galaxy to produce radio emission. The right-hand image of Figure 13.11 shows another active galaxy, M87 (see also Plate 8); this time the field of view is 15×15 arcmin and the wavelength is 90 cm.

13.3 THE MICROWAVE BACKGROUND

In recent years a wealth of data, much of it from studies of the cosmic microwave background (CMB), has led cosmologists to accept a model of the Universe that has revolutionized physics. While the geometry of spacetime is flat, the expansion of the Universe is accelerating, driven by "dark energy" that currently comprises about 2/3 of the total mass–energy budget. Baryons make up only 5% of the total, the remainder being in some exotic form of "dark" matter.

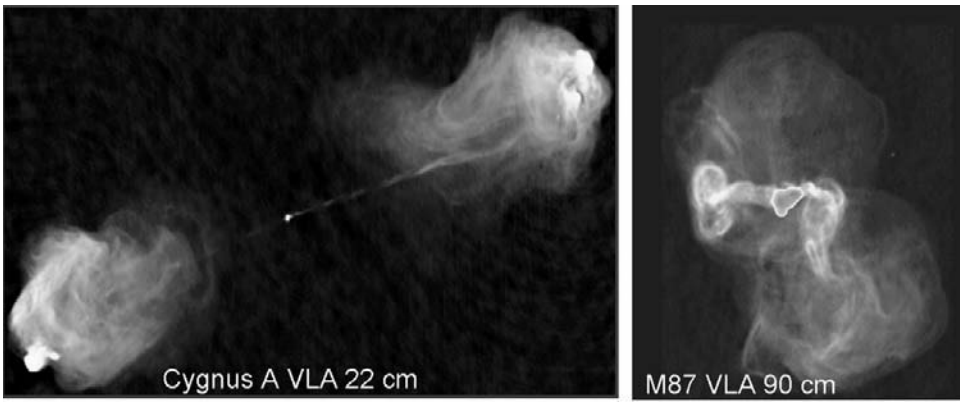


Figure 13.11. Radio images obtained by the VLA: (a) Cygnus A (2×1 arcmin field of view); the double lobes have no optical counterpart; and (b) M87 (15×15 arcmin field of view) (see also Plate 8). Credit: NRAO/AUI.

The peak in the cosmic microwave background occurs around 1.4 mm (213 GHz) and is therefore difficult to study from the ground except at high-altitude sites. Many balloon-borne experiments have been used to overcome the atmosphere, such as the 1998 and 2003 missions of the BOOMERANG instrument developed by an international team and flown over the Antarctic. A 1.2 m mirror focuses the radiation from the sky onto eight horns located in the focal plane of the telescope. Four of those horns couple the radiation to polarization-sensitive bolometers, which operate at 2.1 mm and detect both components of linear polarization, while the other four horns feed photometers which split a single polarization into two bands centered at wavelengths of 1.2 mm and 1 mm. All of the sensors are cryogenic bolometers. See the web site for more details.

Observations of the CMB from space have proved extremely successful, especially those from the pioneering experiments onboard the Cosmic Background Explorer (COBE) satellite launched in 1989. With later confirmation by the 2001-launched Wilkinson Microwave Anisotropy Probe (WMAP), the discoveries from COBE led to the award of the 2006 Nobel Prize in Physics for Dr. John Mather (Goddard Space Flight Center) and Dr. George Smoot (University of California Berkeley), the lead scientists of the COBE mission.

13.3.1 COBE and WMAP

To understand the importance of the COBE satellite (1989–1993) consider the following brief discussion. The Sun consists of about 74% hydrogen, 25% helium, and 1% of all other (heavier) elements combined. It is easy to explain that the trace of heavier elements must have been formed by nuclear reactions inside stars which have

long ago exploded and enriched the interstellar medium, but it is not so easy to understand the amount of helium in this way. Ralph Alpher and George Gamow were the first to realize that the helium must be primordial, and that the Universe must at one time have been at least as hot as the center of the Sun, so that helium could form. During this period the Universe would have been filled with many high-energy, short-wavelength photons. However, the Universe has expanded so much since that era that all those short-wavelength photons have been stretched to become low-energy, long-wavelength photons. As a result, the temperature of this cosmic radiation field is now only a few degrees above absolute zero and its spectrum is that of a black body with a corresponding peak emission at a wavelength of a few millimeters. In the early 1960s Robert Dicke and Jim Peebles at Princeton began designing an antenna to look for this radiation. In 1964, just a few miles away however, at the Bell Telephone Laboratories in Holmdel, New Jersey, Arno Penzias and Robert Wilson were working on a new microwave (7.35 cm) horn antenna designed to relay telephone calls to Earth-orbiting communications satellites. To their surprise, no matter where they pointed their antenna in the sky, they detected a faint background noise. Eventually, Penzias and Wilson were able to show that the microwave radiation was coming from space and that it was isotropic or uniform in all directions. Moreover, the signal strength was matched by that of a black body at a temperature of only 3 K. Through a colleague they learned of the work of Dicke and Peebles and came to realize that they had in fact detected the remnant radiation left over from the Big Bang.

The study of the microwave background has continued since 1965 and reached a peak shortly after 1989 with the launch of the COBE satellite. COBE was designed to measure the spectrum and angular distribution of the cosmic microwave background over the wavelength range from 1 μm to 1 cm. The spacecraft contained three major instruments: the Differential Microwave Radiometer (DMR), the Diffuse Infrared Background Experiment (DIRBE), and the Far Infrared Absolute Spectrometer (FIRAS). In addition to electronics and solar panels, COBE consisted of a large central structure containing a superfluid helium dewar operating at 1.5 K within which were FIRAS and DIRBE. The DMR experiment was composed of three pairs of radiometers around the outside of the dewar.

DIRBE was designed to search for isotropy in the cosmic microwave background over the wavelength range 1 μm–300 μm by making absolute brightness maps in 10 bands: JKLM in the near-infrared (see Chapter 11); the four IRAS bands at 12 μm, 25 μm, 60 μm, and 100 μm; and 120 μm–200 μm and 200 μm–300 μm. A small off-axis Gregorian telescope with a 19 cm primary mirror was used. Mike Hauser was the principal investigator for DIRBE. Comparing the spectrum of the CMB to that of a precise black body onboard was the task for the FIRAS experiment led by John Mather. Two spectral ranges were covered: a low-frequency range from about 1 cm to 0.5 mm (wavenumbers from $1\,\mathrm{cm}^{-1}$ to $20\,\mathrm{cm}^{-1}$) and a high-frequency range from 0.5 mm to 0.1 mm (wavenumbers $20\,\mathrm{cm}^{-1}$–$100\,\mathrm{cm}^{-1}$). A non-imaging parabolic concentrator with a flared aperture to reduce diffraction was used to collect radiation from a 7° diameter beam, and spectral resolving power was obtained with a Michelson interferometer. Six differential microwave radiometers formed the

DMR experiment led by George Smoot. Two independent radiometers were used at each of three frequencies 31.5 GHz, 53 GHz, and 90 GHz (9.5 mm, 5.7 mm, and 3.3 mm). Each pair measured the difference between the CMB emission from two parts of the sky separated by 60°, but the combination of spacecraft spin (75 s), orbit (103 min), and orbital precession (1° per day) allowed each sky position to be compared with all others redundantly. Heterodyne receivers were used with the inputs being switched at 100 Hz between two identical corrugated horn antennas.

FIRAS measured the spectrum of the cosmic background radiation over the huge wavelength range indicated above and obtained a near-perfect fit to a black-body curve with a temperature of 2.735 K. The DMR experiment was used to study the uniformity of the microwave background which at first sight appears to be the same in all directions (isotropic). Closer examination shows that it is slighter greater (hotter) in the direction of the constellation Leo and slightly cooler in the opposite direction (Aquarius). This small "dipole" effect is due to the motion of the Earth and solar system through the cosmos; it is simply a Doppler effect. Our speed toward Leo is about 390 km/s, and taking into account the velocity of the Sun around the center of the Milky Way galaxy, it is found that our entire Galaxy is moving in the general direction of the Centaurus Cluster (of galaxies) at 600 km/s due perhaps to the gravitational pull of a large mass dubbed the Great Attractor. Once the dipole asymmetry in the microwave background due to our local motion is eliminated it is possible to begin to study how smooth the true cosmic background is. Just when it was thought that no variations in the incredibly smooth background would be seen, COBE data finally revealed tiny fluctuations of about 1 part in 100,000. The fluctuations were seen after some extremely complex and careful computer processing and modeling of the data by Edward Wright at UCLA (Data Team leader) and his colleagues on the DMR experiment (see Figure 13.12a).

WMAP (led by Charles Bennett, Johns Hopkins University) was designed to improve on the angular resolution achieved by COBE. Specifically, the goal was to map the relative CMB temperature over the full sky with an angular resolution of at least 0.3° (compared with COBE's 7°), a sensitivity of 20 μK per 0.3° square pixel, and with systematic errors reduced to 5 μK per pixel. To achieve these goals, WMAP used back-to-back differential microwave radiometers that measure temperature differences between two points on the sky. WMAP's microwave system consists of ten four-channel "differencing assemblies", each of which receives signals from a pair of feeds. The output from each assembly is proportional to the temperature (brightness) difference between the two lines of sight on the sky. One assembly operates at 22 GHz, one at 30 GHz, two at 40 GHz, two at 60 GHz, and four at 90 GHz. The remarkable increase in resolution is shown in Figure 13.12b. One result of analyzing these maps is the angular spectrum of the fluctuations which shows the relative brightness of the patches in the map vs. the angular size of those patches. Ripples in such a plot can be compared with the predictions of cosmological models. WMAP was named for David Todd Wilkinson (1935–2002) who contributed to both COBE and WMAP.

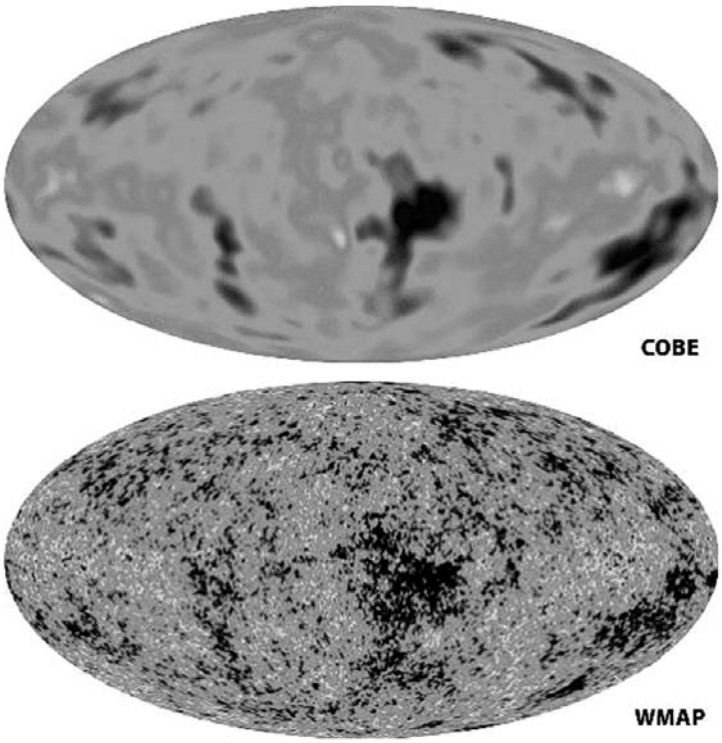


Figure 13.12. (a) An all-sky image of the fluctuations of the microwave background measured by the DMR experiment on the COBE satellite. (b) The same image obtained at higher resolution by WMAP. Credit: NASA.

13.4 SUB-MILLIMETER DETECTORS AND INSTRUMENTS

The sub-millimeter regime is at the interface between the radio and optical/IR regions, and techniques from both camps are combined. For example, the use of mirrors, lenses, and filters is essentially the same as for the optical/IR, whereas the application of antennas, feed horns, and waveguides is clearly from radio technology. Cold interstellar clouds with a temperature of 10 K will emit thermal radiation with a peak at about 0.3 mm, and this is the material from which stars form. Unlike many other parts of the electromagnetic spectrum, all-sky mapping has not occurred yet. One difficulty already mentioned is that the atmosphere is almost totally opaque at these wavelengths except for some weak windows that partially open at high, dry observing sites like Mauna Kea. In addition, detector development has been slower than optical/IR or radio because there are no commercial or military applications. Moreover, extremely low detector temperatures are required, typically well below liquid helium (4 K) and into the millikelvin range. Detectors for the sub-millimeter fall into two main classes: bolometers or continuum detectors for broad-band imaging, and heterodyne systems for spectroscopy and narrow-band observations.

Here we will focus on bolometers. The development of two-dimensional arrays of bolometer pixels has improved rapidly in recent years, especially for devices known as thermal edge sensors, leading to the possibility of CCD-like observing in the sub-millimeter. Many other promising schemes are also under active development (see Zmuidzinas and Richards, 2004 for a review).

13.4.1 Bolometers

As mentioned in Section 5.5.4, a bolometer measures the energy (power) from a radiation field by measuring the change in electrical conductivity of a device as it is heated by the radiation. It is an incoherent detector. Bolometers have a spectral response that is nearly flat over a wide bandwidth, but specific bands can be defined with filters or waveguides, although this area in itself is challenging. Typically, a metal-coated dielectric material is used as the absorber and a semiconductor is used as the thermometer. Each pixel is illuminated by a feed horn, and a small focal plane array can be built up by stacking individual bolometers side by side. A schematic layout of a bolometer is shown in Figure 13.13. The key features are an absorber of heat capacity C, a heat sink held at a fixed temperature T_0, and a weak link of thermal conductance G between them. The thermometer has a resistance R and the output circuit draws a constant current I through a load resistor R_L which allows voltage differences to be measured. The incoming power P_{sig} causes the temperature to change

$$T = T_0 + (P_{\text{bias}} + P_{\text{sig}})/G \tag{13.13}$$

The "thermometer" should be made of a material that gives a large change in resistivity for a small change in temperature. Most bolometers incorporate a semiconductor resistance element such as silicon or neutron transmutation doped germanium (NTD Ge).

These principles are employed in a monolithic silicon bolometer array developed for the SHARC II instrument on the 10 m Caltech Sub-millimeter Observatory (CSO) on Mauna Kea. The array, which has $12 \times 32 = 384$ "pop-up" silicon pixels (see Figure 13.14), was developed at the NASA Goddard Space Flight Center and delivered to Caltech in March 2002. Each silicon pixel is 1×1 mm, coated with a bismuth film, and located $\lambda/4$ above a reflective "backshort" to maximize radiation

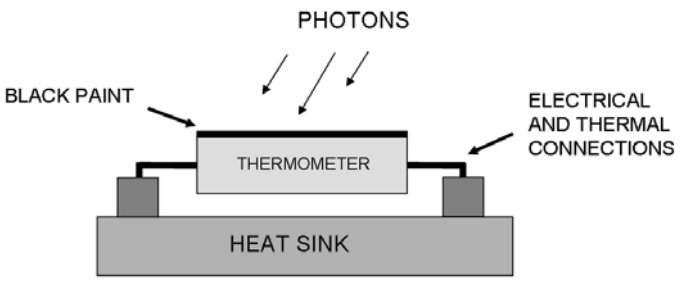


Figure 13.13. Schematic of a simple bolometer element.

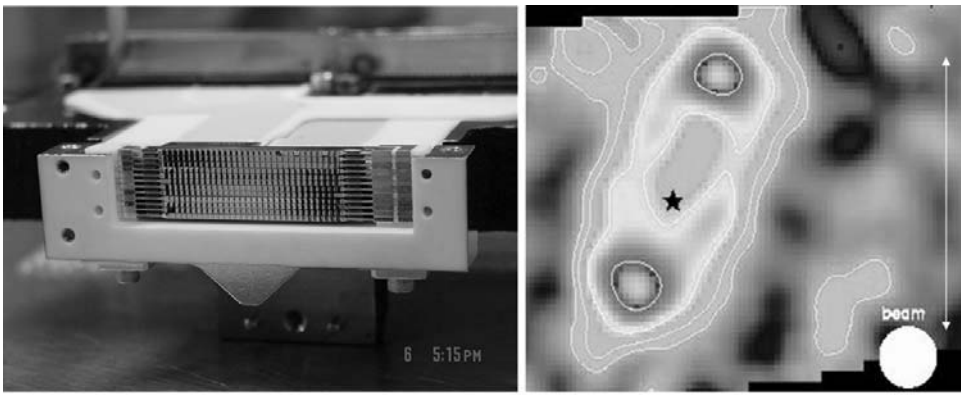


Figure 13.14. (Left) The SHARC II pop-up bolometer array on the 10 m CSO. (Right) A cold debris disk around a young star. Credit: CSO.

absorption. The pixels cover the focal plane with greater than 95% filling factor. Each doped thermistor occupies nearly the full area of the pixel to minimize $1/f$ noise. Operating in the 350 μm and 450 μm atmospheric windows, SHARC II is cooled to 300 mK (0.3 degrees above absolute zero) by a single-shot ^{3}He refrigerator. The CSO web site has more details. Figure 13.14 also shows why the sub-millimeter cameras are so powerful. At these wavelengths one can detect emission from very cold and otherwise invisible material lying in a disk around a young star.

A different bolometer system was developed for the 15 m James Clerk Maxwell Telescope (JCMT) which is also on Mauna Kea. That instrument is called SCUBA, the Sub-millimeter Common-User Bolometer Array (Holland *et al.*, 1999). SCUBA has two arrays of NTD Ge detectors: one of 37 pixels optimized for 850 μm and one of 91 pixels optimized for 450 μm. Both arrays view the same area of sky simultaneously and provide a field of view of approximately 2.3 arcminutes. Each pixel is diffraction-limited, which gives 6″ at 450 μm and 12″ at 850 μm. There is a filter mechanism to allow the 850 μm array to be used at 750 μm or 600 μm and the 450 μm array can be used at 350 μm. In addition to the arrays, there are three separate pixels individually optimized for 1,100 μm, 1,400 μm, and 2,000 μm. The focal plane is slightly under-sampled because of the use of detector feed horns. Dithering (changing the pointing) of the telescope between exposures allows the observer to make a fully sampled image. These bolometers are cooled to 0.1 K. Figure 13.15 (left) shows one of the bolometer arrays, and Figure 13.15 (right) shows image contours of sub-millimeter emitting galaxies discovered by SCUBA on top of an optical picture of the same field. These peculiar, optically faint, highly active galaxies, now known as "SCUBA galaxies" were not expected. They turn out to be at high redshifts corresponding to an era of active star formation and galaxy merging in the early Universe (age $\sim$3 Gyr). Recent observations taken with IRAC and MIPS on *Spitzer* have helped to correlate the appearance of SCUBA galaxies with their emission at shorter wavelengths.

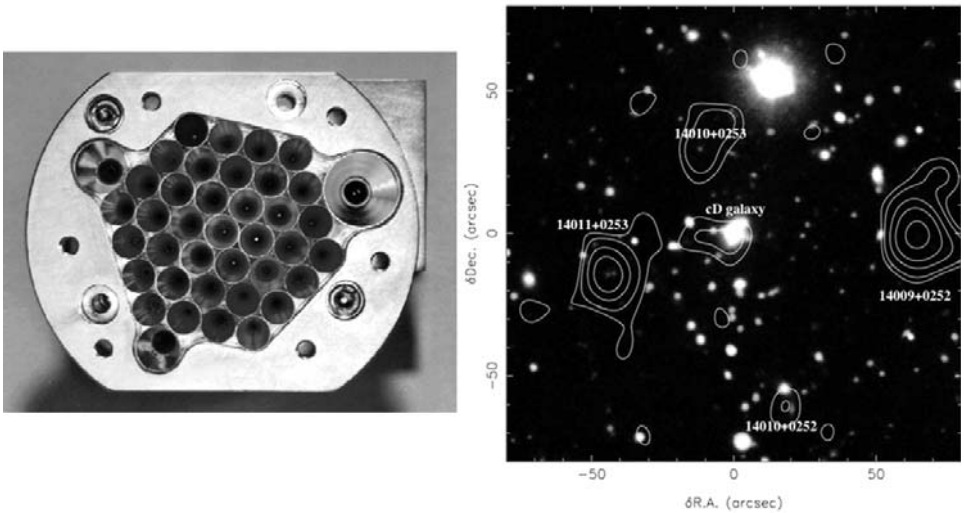


Figure 13.15. (Left) The SCUBA bolometer array with feed horns on the 15 m JCMT. (Right) Image contours of sub-millimeter emitting galaxies on an optical image. Credit: JCMT.

There are several reasons for the exceptionally cold operating temperatures of the sub-millimeter array detectors. First, the detectors are very sensitive, and it is essential to eliminate all thermal (black-body) radiation to avoid swamping the weak astronomical signal. Second, to make the detectors sensitive it is necessary to reduce noise in the detectors and cooling is effective in doing this. Also, the lower the heat capacity the larger the temperature rise for a given signal power input. Most sub-millimeter bolometer arrays run at ~300 mK, and cooling is often provided by self-contained "sorption fridges" using separate ^{3}He and ^{4}He stages, but some operate at 100 mK using "dilution" refrigerators. Self-contained units that combine both approaches are now available.

Bolometer performance is usually described in terms of noise equivalent power (NEP) and noise equivalent flux density (NEFD). Recall that the definition of NEP is that it equals the signal power that gives an $S/N = 1$ for an integration time (τ) of 0.5 seconds. The S/N ratio is proportional to $\tau^{1/2}$, so if the result of a measurement of power (P) produces a signal-to-noise ratio (SNR) in a time τ, then

$$\mathrm{NEP} = [P(2\tau)^{1/2}]/\mathrm{SNR} \tag{13.14}$$

An integration time of 0.5 seconds is chosen because, for an ideal integrator, the bandwidth $\Delta\nu$ is related to the integration time by $\Delta\nu = 1/(2\tau)$, and hence an integration time of 0.5 seconds corresponds to a post-detection bandwidth of 1 Hz and the units of NEP are W Hz$^{-1/2}$ (watts per root hertz). Noise equivalent flux density depends on the NEP of the detection system and also the properties of the telescope, how good the optical coupling is, and other efficiency factors:

$$\mathrm{NEFD} = \mathrm{NEP}/[A_{\mathrm{tel}}\eta_{\mathrm{inst}}\eta_{\mathrm{couple}}\eta_{\mathrm{chop}}\,\Delta\nu] \tag{13.15}$$

where NEP is the total measured NEP (detector + background); $A_{\rm tel}$ is the effective area of the telescope; $\eta_{\rm inst}$ is the transmission of the instrument; $\eta_{\rm couple}$ is the coupling efficiency of instrument to telescope; $\eta_{\rm chop}$ takes into account the chopping efficiency (if appropriate); and $\Delta\nu$ is the spectral passband. The units of NEFD are $\mathrm{W\,Hz^{-1/2}\,m^{-2}\ Hz^{-1} = Jy/Hz^{1/2}}$. Therefore, the SNR for a source of flux density (S_ν) after an integration time (τ) is given by

$$\mathrm{SNR} = [S_\nu(2\tau)^{1/2}]/\mathrm{NEFD} \tag{13.16}$$

13.4.2 Transition edge sensors

A new approach to thermal detectors uses the voltage-biased superconducting Transition Edge Sensor (TES) and Superconducting Quantum Interference Device (SQUID) readout amplifier. These devices can be made entirely by thin film deposition and optical lithography. Negative electrothermal feedback reduces the response time, improves the linearity, and isolates the bolometer response from changes in infrared loading or heat sink temperature. The benefit in linearity comes at the cost of sudden saturation. There is also some suppression of Johnson noise, and the SQUID amplifiers, which operate at the bolometer temperature, dissipate very little power and have significant noise margin.

What is a TES? A bilayer of a thin superconducting metal and a normal metal acts like a single superconductor with a tunable critical temperature T_C. Molybdenum and copper are the usual materials. The superconducting transition is sharp at $\sim$5 mK. As this film warms up slightly its resistance increases, and provided it is held with a constant voltage (V) across it, then its self-heating (Joule heating) produces a power $P = V^2/R$ that decreases, thus providing negative electro-thermal feedback and maintaining stability. The bolometer remains in the narrow transition region around the critical temperature. Many TES pixels can be fabricated on a single silicon wafer substrate. Moreover, a multiplexed readout scheme can be implemented on an independent wafer, and the two wafers can be bump-bonded together similar to hybrid IR arrays. However, unlike in the IR array hybrid where a simple source follower FET handles the signal current from the photodiode or IBC detector element, the extremely small currents in a TES require a Superconducting Quantum Interference Device or SQUID. A SQUID is the union of a superconducting loop and an SIS Josephson junction. Current flows around the loop to make the flux within the loop quantized. As SQUID voltages are very small ($<$1 mV), then one solution is to put 100 of them in series! A row and column biasing scheme is used just like an IR array to connect to each SQUID in turn; this is called time division multiplexing (TDM) and a typical layout is shown in Figure 13.16.

SCUBA-2 is a new $\sim$10,000-pixel sub-millimeter camera for the James Clark Maxwell Telescope based on a large mosaic of TES bolometers hybridized to a SQUID multiplexer (Figure 13.17). Operating in two channels simultaneously at 450 μm and 850 μm, each channel will have an array of $\sim$5,000 TES bolometer pixels arranged by butting four 40×32 sub-arrays on a ceramic printed circuit board.

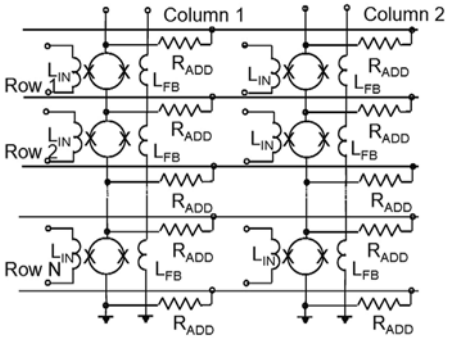


Figure 13.16. The row and column multiplexing scheme of SQUIDS used to read out an array of transition edge sensors. Credit: NIST.

Niobium flex cables fold over to SQUID array amplifiers on the side and ribbon cables take the multiplexed signal to the outside of the large dewar. One of the four TES arrays is shown installed in the focal plane in Figure 13.18; in fact, SCUBA-2 was delivered to JCMT in March 2008 with only one array installed, the others will be installed later. The SCUBA-2 detector is cooled with a ''dry'' dilution fridge that

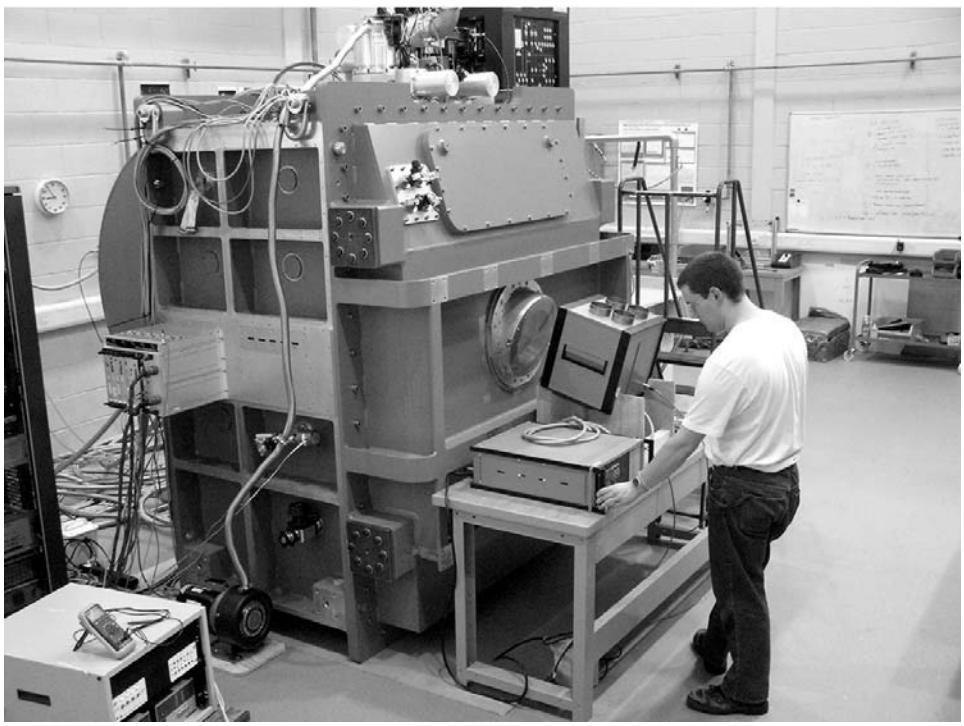


Figure 13.17. SCUBA-2 under test at the ROE/ATC. Credit: Wayne Holland and the ATC.

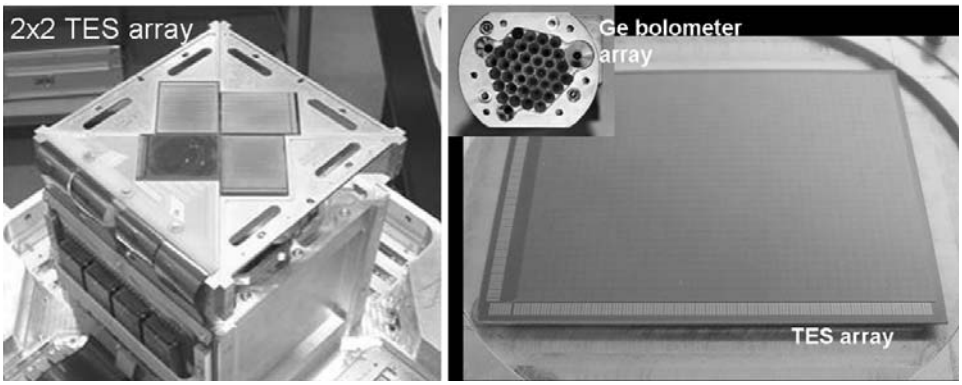


Figure 13.18. (Left) The SCUBA-2 focal plane with one TES array installed (~5,000 pixels). (Right) The relative pixel density of the original SCUBA bolometer compared with the SCUBA-2 camera made with a large array of transition edge sensors. Credit: Wayne Holland.

provides about 500 μW at 120 mK, but the detectors require a heat sink at 60 mK. A device that looks like a hairbrush with bristles or tines of high-conductivity beryllium copper alloy is attached to the underside of each detector to supply the bolometer's heat sink. SCUBA-2 will be background-limited and capable of imaging large areas of sky efficiently. Figure 13.18 (right) compares the TES array with the original SCUBA bolometer array (Holland *et al.*, 2006) to illustrate the remarkable growth in the number of pixels.

13.4.3 Microwave kinetic induction detectors

Existing sub-millimeter detector systems are difficult to multiplex, leading to very complex instruments. The Microwave Kinetic Inductance Detector (MKID) is a new detector design based upon a superconducting resonance circuit that can be used in combination with a suitable radiation absorber or antenna to detect radiation from around 1 mm all the way to X-rays with photon energies of several kilo-electronvolts (Day *et al.*, 2003; Zmuidzinas and Richards, 2004). The main advantage of the MKID principle is that it is very easy to do frequency division multiplexing, and this should enable the realization of large arrays of MKIDs. The underlying principle of the MKID is superconducting pair-breaking, in which the detector senses the change in the number of Cooper pairs due to the absorption of radiation incident on the superconductor. This is done by means of a measurement of the complex surface impedance of the superconductor. Cooper pairs are represented as paired particles at the Fermi energy E_f, the quasiparticles are represented as single particles at energies $E = E_f + \Delta$, where Δ is the energy gap of the superconductor. A photon with an energy $h\nu > 2\Delta$ incident on the superconducting film can be absorbed by breaking up Cooper pairs and creating a number of quasiparticle excitations. As a consequence, both the induction and resistance (L_s and R_s) change due to photon absorption. The resulting change in surface impedance is read out by making the

superconducting film part of a resonant circuit that is capacitively coupled to a through line. The change in surface impedance associated with the change in quasiparticle number can be indicated by a variable inductor, because at low temperatures $\omega L_s \gg R_s$. If the circuit is at resonance it short-circuits the through line with the result that we observe a strong decrease in transmitted power between contacts. The absorption of a photon will cause the resonance center frequency to shift to lower values and to make the measured dip in the transmission smaller. Both effects together produce a phase change at the original resonance frequency. This phase change as a function of quasiparticle number is the signal that is read out to measure the absorption of a photon. For use in a radiation detector, each pixel of the MKID would consist of a suitable antenna or absorber, coupled to a superconducting resonator. The resonator consists of a superconducting coplanar waveguide (a CPW is a planar structure consisting of a central conductor and a ground plane at both sides) with a length of $\lambda/4$, where λ is the wavelength corresponding to the resonance frequency. In the $\lambda/4$ resonator the central conductor is shorted at the far end to the ground plane and floating at the coupler end, where it is coupled capacitively to the through line by letting it run parallel with it. The length of the coupler (and the distance between the through line and the part of the resonator running parallel with it) determines the coupling Q of the resonator. The entire structure acts as a capacitively coupled resonator. A close-up of part of the device is shown in Figure 13.19. Responsivity is defined as the phase change per quasiparticle added to the central line of the resonator. The key to the operation of the MKID is the fact that the Q factor of the superconductor is determined by the ratio of the imaginary and real part of the complex surface inductance, which increases exponentially with decreasing temperature. Hence, the response will also increase exponentially with decreasing temperature. A readout using a very high Q resonator will in principle enable a sensitivity for the whole detector limited by the intrinsic quasiparticle generation–recombination noise, which can be as low as $\mathrm{NEP} \sim 10^{-20}\ \mathrm{W}/\sqrt{\mathrm{Hz}}$ at temperatures below 1/10 of the transition temperature of the superconductor.

The main advantage of MKIDs is that they are much easier to fabricate and system integration is simplified considerably because only a few coaxes are needed for the readout. On the other hand, others have come up with an adaptation of SQUID multiplexing for a TES array that uses microwave frequency multiplexing and MKID-type resonators instead of the time domain scheme being used for SCUBA-2. So, the future looks very interesting for detectors for sub-millimeter and millimeter astronomy.

13.5 SUMMARY

Electronic imaging of astronomical sources at wavelengths from the sub-millimeter to $\sim$1 m is possible. Single-dish radio telescopes have relatively large beams and perform large-area surveys or spectroscopy, but interferometers like the VLA can use aperture synthesis at centimeter wavelengths to make images with the same detail as optical

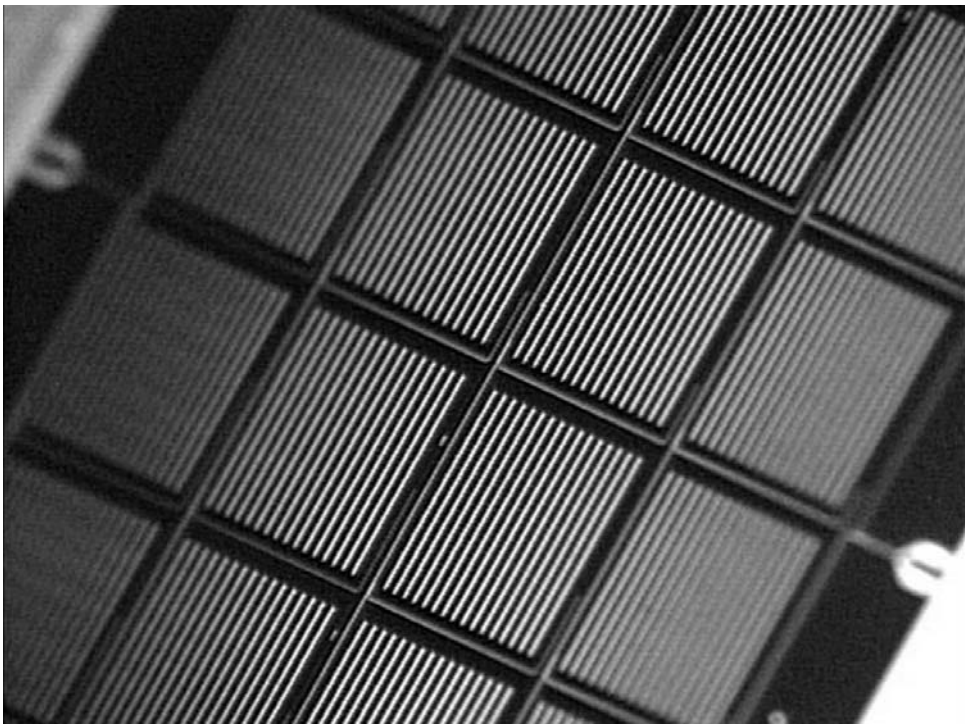


Figure 13.19. A close-up view of the slots in a microwave kinetic induction device (MKID). Credit: Jonas Zmuidzinas, CSO.

pictures. Superconducting SIS junctions are used as sensitive receivers in the radio, sometimes packaged into small arrays with feed horns. Radiometers have mapped the cosmic microwave background at millimeter wavelengths in detail. Small arrays of Si and Ge bolometers have already provided new science at sub-millimeter wavelengths, but large bolometer arrays of superconducting transition edge sensors (TES) will bring CCD-like images to sub-millimeter wavelengths. Large millimeter interferometers are under construction (ALMA) and new detector technologies (MKIDs) are in development.

13.6 EXERCISES

1 Why must radio telescopes be much larger than optical telescopes to achieve moderate angular resolution on the sky? Calculate the approximate angular resolution of a radio telescope of diameter 26 m (a single dish in the VLA) working at a wavelength of 6 cm.

2 Show that $h\nu/kT \ll 1$ for most astronomical sources by using the temperature of the solar photosphere ($\sim$5,800 K) as a typical example. The result is $\sim 8 \times 10^{-6}$.

3 Replace $\exp(h\nu/kT)$ with its Taylor series approximation for small $h\nu/kT$, namely $(1 + h\nu/kT)$, in the Planck function and derive the Rayleigh–Jeans approximation $2kT\nu^2/c^2$.

4 What is the flux density in microjansky of the 1 GHz radio emission from a solar-type star ($T = 5{,}800$) at a distance of 3.26 lightyears ($= 1$ parsec or $\sim 3 \times 10^{16}$ m)? Take the solar radius as 7×10^8 m. The answer is 0.3 μJy.

5 Explain the fundamental principle of radio detection and introduce the concepts of a local oscillator, mixer, and "detector". Why is it convenient to describe the performance of radio receivers in terms of a noise temperature?

6 What is the advantage of combining the signals from two well-separated radio telescopes observing the same source at the same time?

7 Explain the term "aperture synthesis" and the principle behind the CLEAN algorithm.

8 Describe the basic operation of a bolometer.

9 Explain the principle of the transition edge sensor. What advantage(s) does it have over a traditional bolometer?

10 Describe the three main experiments on the COBE satellite and what was learned from them about the cosmic microwave background.

13.7 REFERENCES AND SUGESTED FURTHER READING

Condon, J.J.; and Ransom, S.M. (n.d.) *Essential Radio Astronomy*, NRAO. Available at *http://www.cv. nrao.edu/course/astr534/ERA.shtml*

Day, P.K.; LeDuc, H.G.; Mazin, B.A.; Vayonakis, A.; and Zmuidzinas, J. (2003) A broadband superconducting detector suitable for use in large arrays, *Nature*, **425**, 817–821.

Högbom, J.A. (1974) Aperture synthesis with a non-regular distribution of interferometer baselines, *Astron. Astrophys. Suppl.*, **15**, 417–426.

Holland, W.S.; Robson, E.I.; Gear, W.K.; Cunningham, C.R.; Lightfoot, J.F.; Jenness, T.; Ivison, R.J.; Stevens, J.A.; Ade, P.A.R.; Griffin, M.J.; Duncan, W.D.; Murphy, J.A.; and Naylor, D.A. (1999) SCUBA: A common-user submillimetre camera operating on the James Clerk Maxwell Telescope, *Monthly Notices of the Royal Astronomical Society*, **303**, 659.

Holland, W.S.; Duncan, W.; Kelly, B.D.; Irwin, K.D.; Walton, A.J.; Ade, P.A.R.; and Robson, E.I. (2003) SCUBA-2: A new generation sub-millimeter imager for the James Clerk Maxwell Telescope, in T.G. Phillips and J. Zmuidzinas (eds.), *Proc. SPIE: Millimeter and Submillimeter Detectors for Astronomy*, Vol. 4855, pp. 1–18.

Holland, W. S. *et al.* (2006) SCUBA-2: A 10,000 pixel submillimeter camera for the James Clerk Maxwell Telescope, in J. Zmuidzinas, W.S. Holland, S. Withington, and W.D. Duncan (eds.), *Proc. SPIE: Millimeter and Submillimeter Detectors and Instrumentation for Astronomy III*, Vol. 6275.

Kraus, J.D. (1986) *Radio Astronomy*, second edition, Cygnus-Quasar Books, Durham, NH.

Readhead, A.C.S.; and Wilkinson, P.N. (1978) The mapping of compact radio sources from VLBI data, *Astrophys. J.*, **223**, 25–36.
Signore, M.; and Dupraz, C. (eds.) (1992) *The Infrared and Submillimeter Sky after COBE*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
Rohlfs, K.; and Wilson, T.L. (2004) *Tools of Radio Astronomy*, 4th edition, Springer-Verlag, Berlin.
Verschuur, G.L. (1987) *The Invisible Universe Revealed: The Story of Radio Astronomy*, Springer-Verlag, New York.
Wolstencroft, R.D.; and Burton, W.B. (eds.) (1987) *Millimetre and Submillimetre Astronomy*, Kluwer Academic Publishers, Dordrecht, The Netherlands.
Zmuidzinas, J.; and Richards, P.L. (2004) Superconducting detectors and mixers for millimeter and submillimeter astrophysics, *Proc. IEEE*, **92**(10), 1597–1616.

Web sites

BOOMERANG: *http://cmb.phys.cwru.edu/boomerang/*
COBE home page: *http://lambda.gsfc.nasa.gov/product/cobe/*
CSO home page: *http://www.submm.caltech.edu/cso/*
NRAO home page: *http://www.nrao.edu/*
SCUBA-2 home page: *http://www.roe.ac.uk/ukatc/projects/scubatwo/index.html*
WMAP home page: *http://wmap.gsfc.nasa.gov/*

14

Future developments

In this concluding chapter we briefly review some new technologies, discuss the current trends in astronomical detectors and instrumentation, and summarize plans for some future new facilities. Despite all the advances since the invention of the CCD, many important questions about the Universe remain to be answered, and the experiments needed to study these questions will spur the development of new technology. Larger telescopes, better detectors, and more efficient instruments are a few of the driving factors.

14.1 SCIENTIFIC CHALLENGES

Among the many intriguing puzzles in astronomy still unsolved are several overarching topics that are typically repeated in decadal reviews and proposals for future new facilities. This list of key topics includes the following:

- the discovery of nearby Earth-like worlds, the statistics of planetary systems, and the evidence for biological activity elsewhere
- the detection and tracking of near-Earth asteroids
- a deeper understanding of the origin and formation of stars and planetary systems
- the Black Hole at the center of the Milky Way and tests of General Relativity
- the origin and evolution of the supermassive black holes in quasars
- the origin of cosmic gamma-ray bursts
- the nature and distribution of dark matter in the cosmos
- the nature and distribution of dark energy in the cosmos
- detection of the first starlight in the early Universe
- formation of the first galaxies.

To be sure, some of these problems can be tackled with lengthy programs using existing facilities or upgraded ones, but others will require new and sometimes radically different approaches. Electronic imaging across the electromagnetic spectrum, with even larger telescopes and with finer detail than today, is likely to be at the heart of most of these new developments.

14.2 NEW GROUND-BASED TELESCOPES

It is now about four centuries since Galileo Galilei turned his tiny telescope to the sky and made his first sketch of the Moon's surface. As shown at the beginning and throughout this book, the rate of progress in telescope development has been steady, until recently when the pace seems to have picked up due to rapid advances in technology. In 1998 when the previous edition of this book appeared the only optical telescopes over 6 m in diameter were the twin 10 m segmented mirror telescopes of the W.M. Keck Observatory operated by the California Association for Research in Astronomy. A decade later, in 2008, the number of telescopes in the Very Large Telescope (VLT) category had grown to 15, and plans to develop "Extremely" Large Telescopes (ELTs) with diameters in the range 20 m–40 m had surfaced. There has even been one proposal for a 100 m telescope known as OWL (for Over-Whelmingly Large telescope)! In addition, because large telescopes tend to be used for relatively narrow fields of view, several radically new ideas for large-aperture—but wide-field—telescopes for surveys have been proposed, including the Large Synoptic Survey Telescope (LSST) and Pan-STARRS. All of these new facilities would serve ground-based visible and infrared astronomy, but new ground-based facilities for the radio regime also got under way during this period. ALMA, the Atacama Large Millimeter Array, will dramatically change the field of sub-millimeter and millimeter astronomy, and radio astronomy at centimeter wavelengths will be boosted by projects like the Extended VLA (E-VLA) and the Square Kilometer Array (SKA). Below is a brief summary of each project as of 2008. More information and status reports can be found at the web sites provided.

GMT: The Giant Magellan Telescope

The GMT is a multi-mirror telescope that employs today's largest stiff monolith mirrors as segments. Six off-axis 8.4 m segments surround a central on-axis segment, forming a single optical surface with a collecting area equivalent to a filled aperture 21.4 m in diameter, and the resolving power of a 24.5 m (80 ft) primary mirror. The focal length of this primary mirror combination is 18 m, and the focal ratio is $f/0.7$. Each mirror will be made using the same honeycomb borosilicate mirrors that have already been deployed successfully on the Magellan telescopes in Chile, and the Multiple Mirror Telescope (MMT) and Large Binocular Telescope (LBT) in Arizona. The secondary mirror is composed of seven thin adaptive shells, with each segment mapping to a single primary mirror segment. By making the secondary an adaptive optics component the telescope should get diffraction-limited performance over

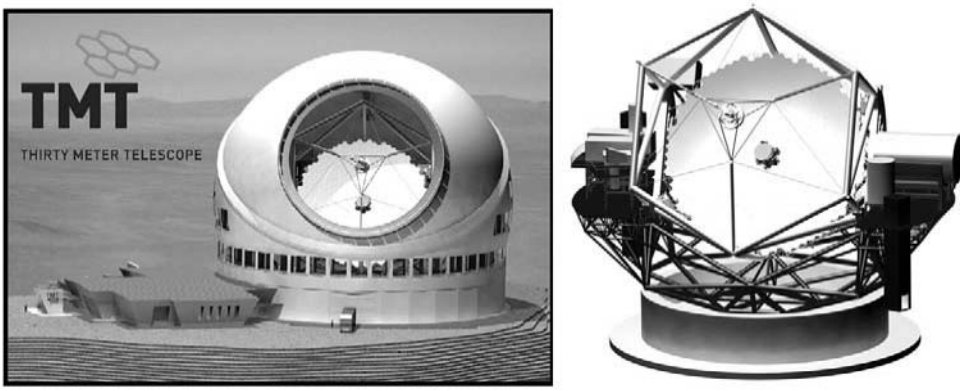


Figure 14.1. Left: Artist's concept of the Thirty Meter Telescope (TMT). Note the unusual dome and the size scale as judged by the human figures and vehicles. Right: Close-up view of the telescope and structure. Credit: TMT Corporation.

modest fields of view, and with the addition of ground-layer adaptive optics at the focus, the corrected field will be 10 arcmin–20 arcmin. The final focal ratio at the straight Gregorian focus will be $f/8.4$ with an image scale of $1.0''$/mm.

TMT: The Thirty Meter Telescope

Based on lessons learned from the Keck telescopes, the TMT (Figure 14.1) is a wide-field, Ritchey–Chrétien telescope with a 30 m (98 ft) diameter $f/1$ hyperboloidal primary mirror composed of 492 hexagonal segments, a fully active 3.1 m secondary mirror, and an articulated tertiary mirror. The optical beam of the telescope feeds a suite of adaptive optics (AO) systems and science instruments mounted on very large Nasmyth platforms surrounding the telescope azimuth structure. These platforms will be large enough to support at least eight different AO/instrument combinations covering a broad range of spatial and spectral resolution. The final focal ratio at the Nasmyth foci is $f/15$ with an image scale of $0.46''$/mm. At the time of writing, TMT is in a very advanced design phase and has received partial funding from the Gordon and Betty Moore Foundation to begin the construction phase following site selection.

E-ELT: The European Extremely Large Telescope

This telescope is a five-mirror concept based on a 42 m (140 ft) $f/1$ segmented primary mirror composed of 906 segments each 1.45 m wide, and a secondary mirror of diameter 6 m. A tertiary mirror 4.2 m in diameter will relay the light to the adaptive optics system which is composed of two large mirrors: a 2.5 m mirror supported by 5,000 or more actuators able to distort its shape a thousand times per second, and one 2.7 m mirror that gives accurate image stabilization.

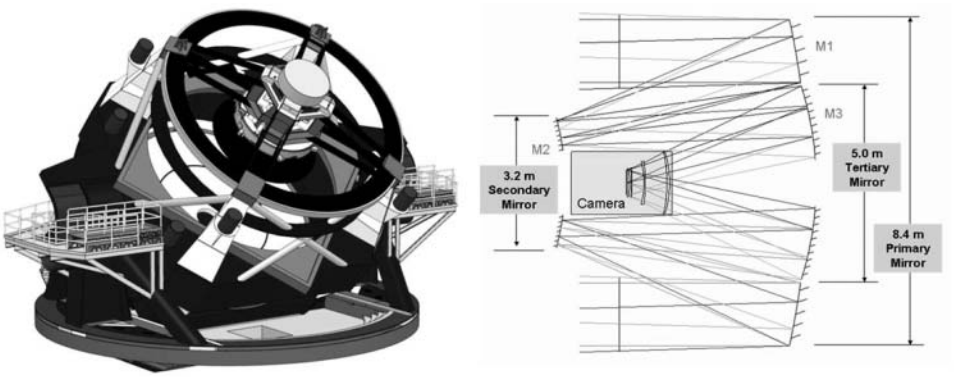


Figure 14.2. Mechanical design and light path for the Large Synoptic Survey Telescope (LSST). Credit: Tony Tyson.

LSST: The Large Synoptic Survey Telescope

The optical design for the LSST is illustrated in Figure 14.2. LSST is an optimized, compact three-mirror system feeding a large refractive corrector producing a 3.5° field of view covering a flat focal plane 64 cm in diameter and paved with pixels. The LSST covers an on-sky area of 9.62 square degrees, nearly 50 times the area of the full Moon. The telescope consists of three aspheric mirrors: an 8.4 m primary (M1), a 3.4 m convex secondary (M2), and a 5.0 m tertiary (M3). Light enters the LSST and reflects off M1 back up toward the front of the telescope where it is reflected by M2 back down to the large M3 mirror. The light reflected by this tertiary mirror then goes to a camera with over three billion pixels. To cover the enormous field of view there will be over 200 CCDs in the focal plane (Figure 14.3), grouped in "rafts" of nine (3×3). For speed, each $4\,\mathrm{K} \times 4\,\mathrm{K}$ chip needs to have 16 outputs. Thus, with 200 CCDs there will be 3,200 digital channels. In addition to the latest generation of CCDs, LSST is considering CMOS (Complimentary Metal Oxide Semiconductor) arrays. These low-noise self-shuttering devices clock the photo-electrons down through several transistors under each pixel, rather than the bucket brigade used in CCDs. As stated on their web site, the LSST will carry out a relentless campaign of 15-second exposures, covering the available sky every three nights, and therefore providing an almost movie-like window on objects that change or move on rapid timescales: exploding supernovae, potentially hazardous near-Earth asteroids, and distant Kuiper Belt Objects. The superb images from the LSST will also be used to trace billions of remote galaxies, measure the distortions in their shapes produced by lumps of dark matter, and provide multiple tests of dark-energy models. LSST is expected to produce 6 GB of raw data every 15 seconds!

Pan-STARRS: The Panoramic Survey Telescope And Rapid Response System

This is another innovative design for a wide-field imaging facility which has already been mentioned. By combining relatively small mirrors with very large digital

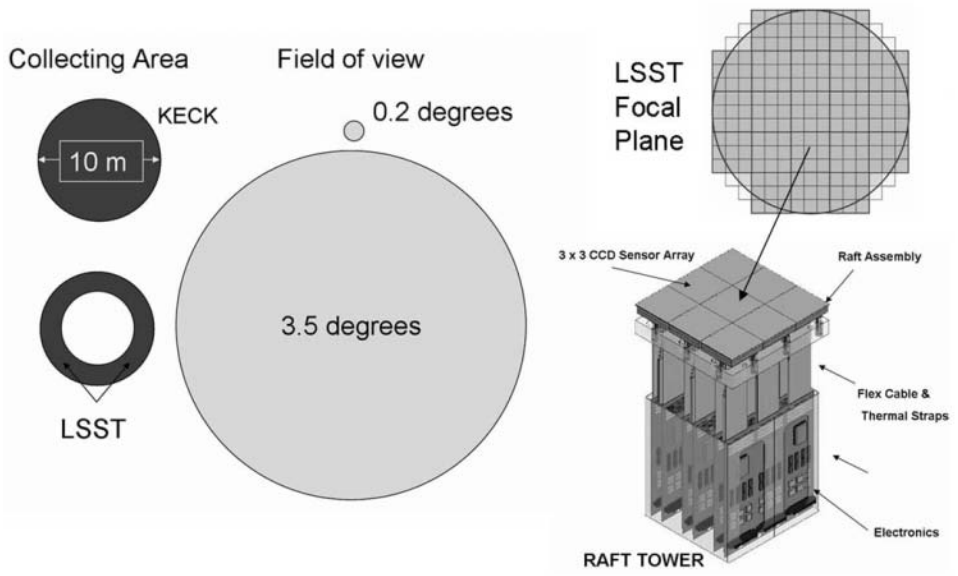


Figure 14.3. An illustration of the field of view of the LSST and the density of CCD pixels in the focal plane. Credit: Tony Tyson.

cameras, an economical observing system can be deployed that will be able to observe the entire available sky several times each month. Pan-STARRS will be composed of four individual optical systems, each with a 1.8 m diameter mirror observing the same region of sky simultaneously. Each mirror will have a 3° field of view and be equipped with a 1.4 billion pixel CCD camera. Spatial sampling on the sky will be about 0.3″. While searching for potential killer asteroids in survey mode, Pan-STARRS will cover 6,000 $\deg^2$ per night. With exposure times of 30–60 seconds, Pan-STARRS will reach a limiting magnitude of 24. Orthogonal Transfer CCDs (OTCCDs) that allow the shifting of charge along both rows and columns will enable on-chip image motion compensation. The first Pan-STARRS 1.4-gigapixel camera achieved first light in August 2007.

ALMA: The Atacama Large Millimeter/submillimeter Array

Illustrated in Figure 14.4, ALMA will be composed of 64 high-precision antennas with diameters of 12 m (40 ft) and will operate at wavelengths of 0.3 mm to 9.6 mm. The ALMA site is on the Chajnantor Plain of the Chilean Andes in the District of San Pedro de Atacama, 5,000 m (15,500 ft) above sea level. Providing astronomers a new window on celestial origins, ALMA will probe the first stars and galaxies, and directly image the formation of planets. The 64-element synthesis array will have reconfigurable baselines ranging from 150 m to 18 km. Resolutions as fine as 0.005″

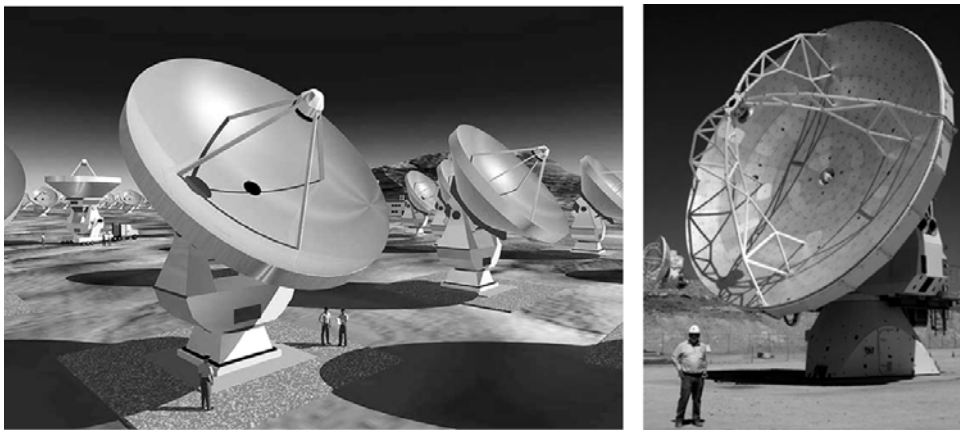


Figure 14.4. Left: Artist's concept of the array of antennas planned for the Atacama Large Millimeter and sub-millimeter Array (ALMA). Right: One of the first antennas to arrive on site. Credit: ALMA Project.

will be achieved at the highest frequencies, a factor of 10 better than the Hubble Space Telescope. Shown beside the artist's conception of the array is the first antenna, VertexRSI unit-1 on site in Chile. During 2007 seven antennas were delivered to the site, and in early 2008 the large antenna transporters arrived. Current information on status can be obtained from the ALMA web site and the NRAO Newsletter. A discussion of how ALMA will impact astronomy can be found at *http://www.cv.nrao.edu/naasc/disk07/*

SKA: The Square Kilometer Array

This unique facility will be an aperture synthesis radio interferometer operating over the frequency (wavelength) range 100 MHz (3 m) to 25 GHz (1.2 cm). A number of configurations are under study to provide the 1 million square meters of collecting area trading off number of stations vs. antenna size. The equivalent aperture will have a diameter of about 3,000 km yielding an angular resolution of $<0.1''$ at 1.4 GHz. About 50% of the collecting area will be contained within an inner array of 5 km diameter to provide very high sensitivity at arcsecond resolutions for studies of the faint spectral-line signatures in the early Universe. It is hoped that the development of new radio and computer technologies will lead to a 50-fold gain in sensitivity compared with the VLA in New Mexico. The design goal for the imaging processor computation rate is 10^{15} operations/s which should be possible by 2010 if Moore's Law continues to hold.

Of course, the above list of new facilities is incomplete, and it also ignores the numerous upgrades and special additions being made to smaller telescopes to make them more useful in the ELT era. Furthermore, it does not include the full-blown operation of optical/IR interferometers such as the VLTI in Chile.

14.3 NEXT-GENERATION ADAPTIVE OPTICS

Most current AO systems use a single laser guide star, but future systems are likely to employ multiple lasers to provide many guide stars in order to overcome the "cone effect" for the largest existing telescopes and the new Extremely Large Telescopes (ELTs), and to provide a larger corrected field of view. By looking through the atmosphere at many angles it is possible to make a "tomographic" reconstruction of turbulence. Currently, all AO systems perform science measurements in the infrared, but it may soon be possible to achieve good corrections well into the optical regime, down to at least H-alpha at 656.3 nm. A combination of low-order AO corrections and fast, low-noise electron-multiplied CCDs can extend this even farther, although with less efficiency due to the low percentage of CCD frames with the highest resolution (lucky imaging).

AO systems for the next generation of extremely large telescopes will face many challenges. For example, many systems will try to operate at temperatures $\sim 35^{\circ}$C below ambient to reduce thermal infrared backgrounds. Multiple sodium laser beacons ($\sim$6) with 17 W–25 W per beacon will be needed. In addition to the laser, other critical components include the laser launch telescope and the beam transfer optics. The finite thickness of the sodium layer ($\sim$10 km) results in a perceptible radial extension of the image in the Shack–Hartmann camera. For TMT, for instance, the effect is an elongation of almost $4''$ at the edge of the enormous 30 m aperture. Using a pulsed laser instead of a continuous wave (CW) laser would result in a series of spots instead of a fainter elliptical image. A polar coordinate CCD containing separate patches of pixels for each wavefront sensor subaperture would provide significant advantages, with or without a pulsed laser, by reducing the total number of pixels which must be read out and increasing the signal-to-noise ratio. If a pulsed laser is available then charge shifting could be synchronized with the laser's pulse repetition rate to compensate for radial motion of the spot in a highly specialized application of time delay integration. Low-noise, high-speed infrared detector arrays will be needed for the natural guide star tip/tilt wavefront sensors in order to achieve the sky coverage demanded by the extremely large telescopes. These detectors will need to have a readnoise level of 5–10 electrons, at least 256×256 pixels, and operate at frame rates of at least 500 frames per second. Achieving these goals will require developments, perhaps along the lines of using arrays of infrared-sensitive avalanche photodiodes. Finally, one of the most challenging and critical components of future AO systems will be the real-time controller. The typical controller, in the case of the TMT, for instance, will need to (i) compute and apply commands to two high-order deformable mirrors with a total of $\sim$7,500 actuators and one tip/tilt platform on the basis of multiple wavefront sensor inputs; (ii) interface with telescope and AO sub-systems; and (iii) compute wavefront structure functions to estimate the AO PSF. These real time controller tasks are at least 100 times greater than the current systems on 8 m–10 m telescopes. Algorithms will be implemented using state-of-the-art digital signal processor (DSP) chips and field programmable gate arrays (FPGA) to perform the reconstructions at a rate of over 1,000 per second.

Cheaper and more powerful solid-state lasers, faster and more sensitive wavefront sensors, larger deformable mirrors with more elements, new kinds of curvature sensors, and even faster computer processing power are needed, but all of these technologies are almost available now.

14.4 NEXT-GENERATION INSTRUMENTATION

For optical and IR instruments there is a move towards a massively parallel approach. This is already evident in the use of hundreds of fibers or slits on a mask for multi-object spectroscopy, but the multiplex advantage is likely to grow into the thousands and be extended to include diffraction-limited systems.

The concept of smart focal planes (Cunningham and Evans, 2008) has been advanced and refers to devices that enable the efficient sampling of a telescope's focal plane to feed spectroscopic and imaging instruments. One example is integral field units (whether fiber, image slicer, or lenslet). Although more complex to build and use than conventional spectrometers, integral field units have a unique advantage because they yield spectral and spatial information simultaneously. Technology improvements will make these systems cleaner and easier to use. Other examples are robotic beam manipulator units, especially cryogenic modules for infrared applications, and the use of micro-optical–electromechanical systems (MOEMS) such as miniature slit shutters. One of the drivers for these kinds of considerations is the following. Consider a 30 m telescope working at its diffraction limit at a wavelength of 1 μm. The FWHM in this case is $0.0069''$ given as usual by λ/D, but to Nyquist-sample this image we need at least two pixels and so the pixel size would be $0.00345''$. However, the typical AO-corrected field of view is likely to be about $120''$, and thus to cover this field we would need an (infrared) detector with almost 35,000 pixels on a side or $35\,\mathrm{K} \times 35\,\mathrm{K} = 1.2$ billion pixels in total. While this has been achieved already for silicon CCDs working in the visible (Pan-STARRS), it remains very challenging for near-infrared arrays whose current cost is ~10 U.S. cents per pixel implying a focal-plane mosaic costing over $120 million. The question arises as to whether it is necessary to have a contiguous field of view. Sometimes the density of targets within the 2-arcminute field will be low, and therefore many pixels are being wasted looking at blank sky. Why not concentrate a smaller number of camera pixels where they are needed and use robotic manipulators to redeploy them to other areas within the larger field of regard. Deployable integral field units can then be duplicated identically to enable tens of small areas (e.g., galaxies) within the AO-corrected field to be studied simultaneously. Systems like this are already under development. One example is the K-band Multi-Object Spectrometer (KMOS), which is a new infrared multi-integral field spectrograph being built as a second-generation instrument for the ESO VLT on Cerro Paranal (Chile).The baseline design (Sharples *et al.*, 2006) employs 24 configurable arms that position fold mirrors in the focal plane, each of which selects a sub-field of 2.8×2.8 arcseconds. These fields are magnified onto 24 image slicers that partition each field into 14 parts with 14 pixels along each slice at 0.2 arcsec spatial sampling. Light from the slicers is dispersed by

three identical cryogenic spectrometers which generate 14×14 spectra each with $\sim$1,000 Nyquist-sampled spectral resolution elements for all of the 24 independent sub-fields.

14.5 NEW SPACE TELESCOPES

The assault on other parts of the electromagnetic spectrum will continue from new space-based telescopes. Succeeding the Hubble Space Telescope will be the James Webb Space Telescope (JWST) with an aperture of 6.5 m ($\sim$21 ft). Figure 14.5 shows the conception of the telescope with its segmented mirrors and large Sun shield. Located about 1.5 million km (almost 1 million miles) from Earth at the L2 point, JWST will be infrared-optimized and operate from the visible to the mid-infrared (0.6 μm–30 μm). JWST's near-infrared camera (NIRCam) has 10 mercury–cadmium–telluride (HgCdTe) arrays and operates from 0.6 μm to 5.0 μm. Likewise the Near Infrared Spectrograph (NIRSpec) covers the same wavelength range but employs MEMS microshutter arrays to provide a multi-object spectroscopy mode for up to 100 objects at a time. MIRI, the Mid-IR Instrument, is both an imager and spectrometer for the 5 μm–27 μm using three arsenic-doped silicon (Si : As) arrays.

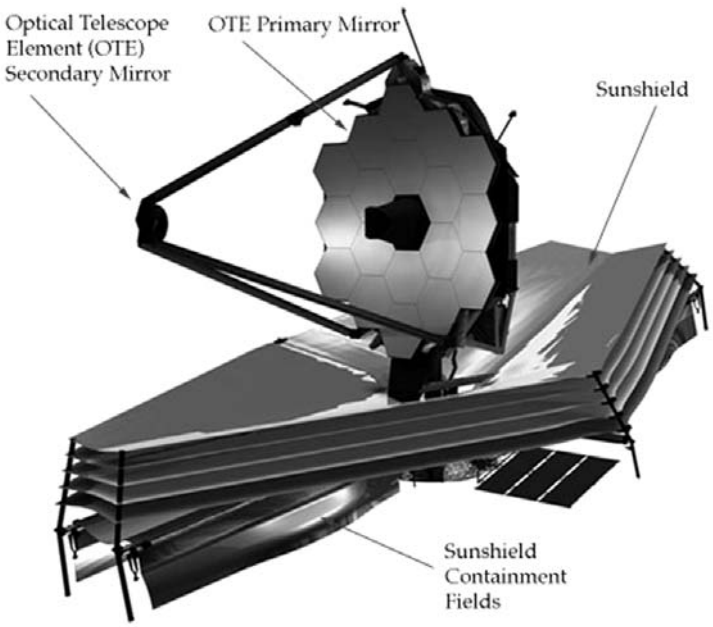


Figure 14.5. Artist's view of the James Webb Space Telescope as it will be deployed in space. The effective aperture is 6.5 m, and JWST will be sent to the L2 point about 1 million miles from Earth. Credit: NASA.

Other space missions expected to launch in the near future include ESA's Herschel, previously known as the Far-Infrared and Sub-millimeter Telescope (FIRST), a 3.5 m telescope to study the region from 60 μm to 670 μm. Herschel has three onboard instruments: PACS and SPIRE provide imaging and spectroscopy from 60 μm to 210 μm and from 200 μm to 670 μm, respectively, using bolometer arrays and Ge : Ga photoconductor arrays, and HIFI is a very high resolution heterodyne spectrometer operating in seven bands from 480 GHz to 1,910 GHz with SIS or HEB mixers. Herschel will be co-launched with Planck, an ESA mission to study the cosmic microwave background (CMB). Planck carries a 1.5 m telescope that feeds microwave radiation into two instruments: the Low Frequency Instrument (LFI) is an array of 22 tuned radio receivers in the focal plane to image the sky at three frequencies between 30 GHz and 70 GHz and the High Frequency Instrument (HFI) is an array of 52 bolometers to image at six frequencies between 100 GHz and 857 GHz. Planck has 50 times better angular resolution than the Cosmic Background Explorer (COBE) discussed in Chapter 13 and will be sensitive to CMB variations in temperature at the level of one part in a million. Both of these satellites will operate at the L2 point.

In the high-energy regime the next major mission will be NASA's Gamma-ray Large Area Telescope (GLAST) mentioned in Chapter 12. Unlike SWIFT and INTEGRAL, the telescope is not a coded mask but more like a large spark chamber similar to the EGRET experiment on the earlier Compton Gamma Ray Observatory that operated from 1991 to 1999, but using new technology. The Large Area Telescope (LAT) consists of a four-by-four array of tower modules. Each tower module consists of 19 interleaved planes of silicon strip detectors and lead converter sheets. Silicon strip detectors (SSDs) are better able to track the electron or positron produced from the initial gamma ray than previous types of detectors as it moves down toward the calorimeter at the base. The LAT has a field of view of about 145 degrees and is 50 times more sensitive than EGRET at 100 MeV and even more sensitive at higher energies. It will be able to locate sources to within 0.5 arcmin–5.0 arcmin.

The search for possible edge-on planetary systems that transit their star will be tackled by NASA's Kepler mission. Using a 0.95 m telescope designed to have an unusually large field of view (105 square degrees), Kepler will monitor the brightness of over 100,000 stars simultaneously and continuously in visible light (430 nm–890 nm) for 3.5 years by means of a large mosaic of 42 CCDs, each one with $2{,}200 \times 1{,}024$ pixels yielding 95 megapixels in total (see Figure 14.6). This will be the first very large CCD mosaic in space. Kepler will be in an Earth-trailing heliocentric orbit (like Spitzer).

Another mission with a large CCD mosaic in space will be ESA's GAIA, a follow-on mission to HIPPARCOS to chart a 3-D map of the nearest billion stars. GAIA is based on a dual telescope with a common structure and a large shared focal-plane CCD camera. Both telescopes are three-mirror anastigmats (TMAs) which have been mentioned earlier in Chapter 6. The mirrors and structure are made from silicon carbide (SiC), and the common focal plane has 106 CCDs (almost 1 Gpxl) in a format spread over 0.5×1.0 m. The CCDs, also from *e2v technologies*, all have the

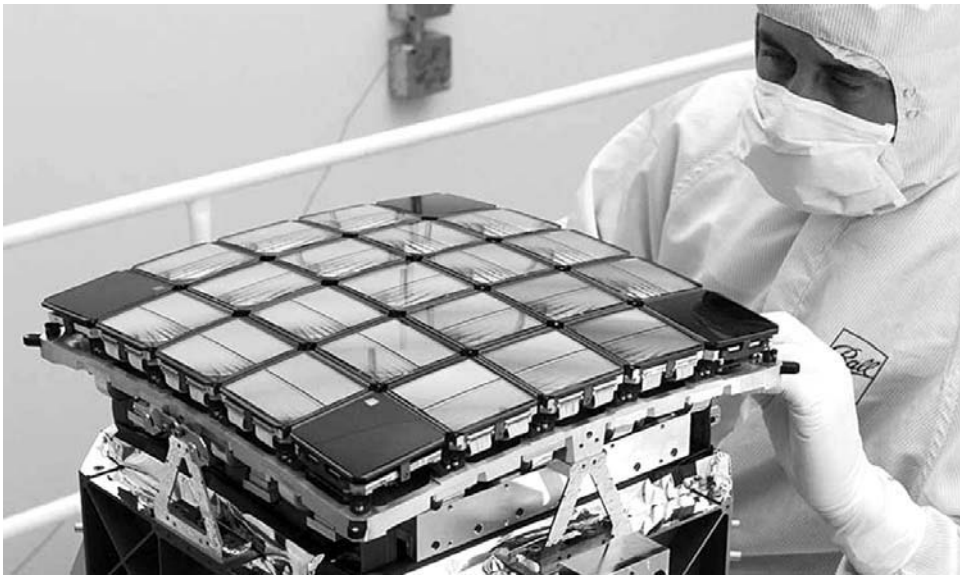


Figure 14.6. The Kepler mission will be involved in looking at millions of stars for evidence of planetary transits. Shown here is the large CCD focal plane already under construction. This will be one of the largest detector mosaics to fly in space. Credit: Ball Aerospace.

same format and are large-area, back-illuminated, full-frame devices operated in time delay integration (TDI) mode with a TDI period of 982.8 μs. GAIA is intended for an L2 orbit.

There will also be a space mission to study the dark matter and dark energy in the Universe. Currently called the Joint Dark Energy Mission (JDEM), this space telescope will feature a very large array of both CCDs and infrared arrays. Although less certain at this time, the space interferometer mission (SIM) and related missions with very high angular resolution are likely to occur eventually.

14.6 THE FUTURE OF ELECTRONIC IMAGING

CCD technology has now evolved to a stable and near-perfect state for scientific instrumentation. The emphasis in the near term is likely to be in areas of packaging and electronic control in order to make the planned very large gigapixel mosaics for LSST, Pan-STARRS, and JDEM practical and cost-effective. Larger format electron-multiplied CCDs with improved performance and near-perfect photon-counting characteristics is another area of ongoing development, and the equivalent device for the near-infrared is already under consideration. Custom-designed CCDs with special features, such as graded anti-reflection coatings and in-chip electronic shutters, or special shaped pixels, or special readout modes are all possible now, and this trend will probably continue. CMOS imaging devices now common in phones

and cameras have not yet been used extensively in scientific cameras, but this and other "smart-pixel" technology, such as a multi-band pixel (a pixel that can detect and discriminate between two wavelengths at the same time) is a growth area.

Infrared arrays will soon reach the $4\,\mathrm{K} \times 4\,\mathrm{K}$ size in the NIR, and buttable designs will allow these to be built into large mosaics, in the same way as silicon CCDs. Mid-IR detectors are likely to grow beyond the current $1\,\mathrm{K} \times 1\,\mathrm{K}$ pixel level and be packaged into mosaics too, most likely for space missions. In the far-infrared and sub-millimeter regimes, the advent of superconducting transition edge sensor (TES) arrays with 1,000 elements was a major advance, and this area is likely to grow and improve. The newer technology of microwave kinetic inductance devices (MKIDs) is poised for rapid growth, stimulated by new telescope concepts such as CCAT, the Cornell Caltech Atacama Telescope, and others.

In almost all of these array detector systems there is an increasing need to simplify the electronics and make the system more compact. This requirement for JWST drove the development of application-specific integrated circuits (ASICs) designed to replace all of the usual electronics needed to operate an array and digitize the output signal. Some ground-based NIR arrays are now using ASICs too, and hopefully this technology will extend to CCDs and most other array devices.

Still under development is the superconducting tunnel junction (STJ) array which is essentially a photon-counting detector with high quantum efficiency from the visible to the near-IR *and* with the ability to discriminate the energy of individual photons. An array of STJ devices would constitute a combined camera and spectrograph (i.e., a three-dimensional camera that has no gratings or prisms to disperse the light and therefore very little optical light-loss). A superconducting tunnel junction consists of two films of superconductor sandwiching a thin insulating layer. When operated at temperatures below the critical temperature of 1 K (i.e., below about one degree above absolute zero), the equilibrium state is easily disturbed by an incident photon. By applying a magnetic field and a small bias voltage across the junction, an electrical charge proportional to the energy of the perturbing photon can be extracted. Although STJ devices have been considered for X-ray applications, progress has led to the possibility of using the method for UV and visible photons too. Rando *et al.* (2000) describe a 6×6 array of TaAl STJs in a cryogenic camera used at a ground-based observatory.

14.7 CONCLUSION

The years since the invention of the CCD in 1970 have been a remarkable period of growth and development for astronomy. This time period matches my own career quite well as I graduated with my first degree in 1971, and it is sobering to look back at all that has transpired in astronomy since then. New technologies have led to new discoveries, and the promise of even greater discoveries has stimulated the drive for even better instrumentation. As expected, access to space has meant that gamma-ray, Xray, and ultraviolet astronomy have all blossomed, and CCD-like imaging is now available even at the shortest wavelengths. Infrared astronomy underwent a

tremendous surge with the advent of InSb, HgCdTe, and Si : As array detectors in the mid-1980s, and yet another boost with the introduction of adaptive optics techniques for the elimination of atmospheric turbulence. New telescopes and new detectors for far-infrared to millimeter wavelengths have finally opened up that part of the spectrum to CCD-like imaging too. Conventional ground-based optical astronomy, far from becoming entrenched, has continued to expand with new telescopes and better CCDs. Even the very method of building optical reflecting telescopes underwent a radical change during this period. Adaptive optics and laser guide stars have enabled the suite of very large telescopes now in operation to achieve their diffraction-limited performance, at least in the near-infrared. Of course, the amount of data being obtained, studied, and archived is enormous, but Moore's Law has enticed computer manufacturers to keep pace, and hopefully concepts like the Virtual Observatory will be fully realized.

Throughout this book I have tried to show that the underlying reason for these advances has been a willingness of astronomers to grasp the very latest technologies and push them to their limits. From Galileo Galilei's eyes (1609) to gigapixel CCD cameras (2009), from sketches of the uneven surface of the Moon to the detection of the cosmic microwave background and ripples in the fabric of spacetime, we have come a long way. Our understanding of the Universe has grown in leaps and bounds, hand in hand with developments in technology, and the cycle continues. I hope it will always be so, and I think Galileo would be pleased.

14.8 REFERENCES

Cunningham, C.; and Evans, C. (2008) Smart focal plane technologies for VLT instruments, *Proc. ESO Workshop on Science with the VLT in the ELT Era.*

McLean, I.S.; and Iye, M. (eds.) (2006) *Ground-based and Airborne Instrumentation for Astronomy*, SPIE, Bellingham, WA [*Proc. SPIE*, **6269**].

Rando, N.; Verveer, J.; Verhoeve, P.; Peacock, A.J.; Andersson, S.; Reynolds, A.; Favata, F.; Perryman, M.A.; and Goldie, D.J. (2000) S-Cam 2: Performance and initial astronomical results, *Proc. SPIE*, **4008**, 646–656

Sharples, R.M.; Bender, R.; Bennett, R.; Burch, K.; Carter, P.; Casali, M.; Clark, P.; Content, R.; Davies, Richard; Davies, Roger *et al.* (2006) Design of the KMOS Multiple-object Integral Field Spectrograph, *Proc. SPIE*, **6269**, 49.

Web sites

ALMA home page: *http://www.alma.nrao.edu/*
CCAT home page: *http://www.submm.org/ccat.html*
E-ELT at ESO: *http://www.eso.org/sci/facilities/eelt/*
GAIA home page: *http://sci.esa.int/science-e/www/area/index.cfm?fareaid=26*
GLAST home page: *http://glast.gsfc.nasa.gov/*
GMT home page: *http://www.gmto.org/*
HERSCHEL home page: *http://www.esa.int/science/herschel*
JDEM reference: *http://universe.nasa.gov/program/probes/jdem.html*

JWST home page: *http://www.jwst.nasa.gov/*
KEPLER home page: *http://kepler.nasa.gov/*
KMOS reference: *http://www.cfai.dur.ac.uk/new/projects/kmos/kmos.html*
LSST home page: *http://www.lsst.org/lsst_home.shtml*
Pan-STARRS home page: *http://pan-starrs.ifa.hawaii.edu/public/*
PLANCK home page: *http://www.esa.int/science/planck*
SIM reference: *http://planetquest.jpl.nasa.gov/SIM/sim_index.cfm*
SKA home page: *http://www.skatelescope.org/*
SPIE Astronomical Instrumentation: *http://spie.org/x13667.xml*
TMT home page: *http://www.tmt.org/*

Answers to selected exercises

CHAPTER 1

3. (a) 300 GHz;
   (b) 21 cm.
5. $E = 3.98 \times 10^{-19}$ J = 2.48 eV.
7. Number of electrons at dynodes $1, 2, 3 \cdots = (3), (9), (27), \cdots . = 3^1, 3^2, 3^3 \cdots$ and after 10 dynodes it is 3^{10}. Thus, $Q = 9.46 \times 10^{-15}$ C.
8. (a) ±1,000 counts;
   (b) 0.1%;
   (c) a factor of $\sqrt{4} = 2$ (to 0.05%).
9. $F^3 = 206{,}265/128 \times 1.0 = 1{,}611.45$; $F \sim 11.725$. Lens = 8× the mirror ∼8″. Doubling focal length implies aberrations are $2^3 = 8$ times smaller; earliest telescopes had long focal lengths.
10. Readout noise dominates for 1 s; background limits S/N for 100 s.

CHAPTER 2

1. Diffraction-limited angular resolution corresponds to $D = 85$ m:

$$(206, 265)1.22\lambda/D$$

   Angular resolution is 0.003 seconds of arc per micron of wavelength. Light-gathering power is only equivalent to $\sqrt{2} \times$ 10-m = 14.1 m telescope (not 85 m). Interferometry gives huge gain in resolution (85/10 = 8.5).
2. Using $1.22\lambda/D$, then $\theta = 0.05''$ at 2 μm. Linear dimension = (0.05″/206,265″ per rad) × (4.2 lt-yr × 6×10^{12} miles per light-year): linear size = 6.15 million miles.

Baseline of 100 m = 10× better resolution: linear size ≈ 615,000 miles or about 2.5× the distance to the Moon.

3. Seeing $\sim \lambda_0/r_0$:
   (a) $r_0 = (0.5 \times 10^{-6}\,\text{m}/0.75'')206{,}265 = 0.14\,\text{m}$;
   (b) $r/r_0 = (2.2/0.5)^{6/5} = 5.9$, thus at 2.2 μm, $r = 0.83\,\text{m}$.
5. $r_0 = (0.5 \times 10^{-6}\,\text{m}/0.5'')206{,}265 = 0.21\,\text{m}$; $r/r_0 = (1.65/0.5)^{6/5} = 4.2$, $r = 0.86\,\text{m}$; number of sub-apertures $(D/r)^2 = (10/0.86)^2 = 134$.
6. From 5, $r_0 = 0.21\,\text{m}$; isoplanatic angle $= 0.314 r_0/H = (0.314 \times 0.21\text{m}/5{,}000\,\text{m}) = 13.2\,\mu\text{rad}$ or 2.7 arcsec. The isokinetic angle is larger by $D/r_0 = 47.6$, thus isokinetic angle = 129.5″.

CHAPTER 3

3. (a) prime focus is faster by $(17/5)^2 = 11.56$.
   (b) Cassegrain is better in practice because the larger plate scale reduces the level of sky background.
4. Tails toward the edges of the field implies coma in the optics. For a simple telescope/camera this could mean a misaligned secondary mirror.
8. FOV = 40.96 × 34.13 arcminutes or 0.39 square degrees. Number of frames for 100 square degrees is ~258. At 20 minutes each this takes ~86 hours of observing. One hemisphere is 20,627 square degrees, so to cover this takes 17,720 hours of observing (over 2 years of time). This field of view is too small or the integration time is too long to be useful.
10. Limit on astigmatism of <0.5″ implies FOV ~20′. Parameters are $R_1 = 35\,\text{m}$, $R_2 = 4.85\,\text{m}$, $K_1 = -1.00404$, $K_2 = -1.64344$, and $d = 15.204\,\text{m}$. Answers may vary slightly.

CHAPTER 4

No entries.

CHAPTER 5

2. The $f/\# = (206{,}265) d_{\text{pix}}/D_{\text{tel}}\theta_{\text{pix}} = (206{,}265''/\text{rad}) \times (24\,\mu\text{m})/(10\,\text{m} \times 10^6\,\mu\text{m/m})(0.2'') = 2.48$. Yes, this camera would be challenging. 3.4′ × 3.4′.
3. $D_{\text{coll}} = RD_{\text{tel}}(\varphi/206{,}265)/2\tan\theta_B$, $R = 20{,}000$, $D_{\text{tel}} = 10\,\text{m}$, $\varphi = 0.5$:
   (a) $(2\tan\theta_B = 0.63) \Rightarrow D_{\text{coll}} = 77\,\text{cm}$;
   (b) $(2\tan\theta_B = 4) \Rightarrow D_{\text{coll}} = 12\,\text{cm}$. The echelle is more practical.
   Focal length = 15 × D_{coll}:
   (a) 11.55 m.
   (b) 1.8 m.

4. $n = 2.4$, $A = 30°$, $\lambda_c = 2.2\,\mu\text{m}$, and $R = 500$ for 2 pixels. Pixel size $= 27\,\mu\text{m}$. Assuming $m = 1$ (first order), $T = (n-1)\sin A/\lambda_c = 318$ lines per millimeter and EFL $= 2d_{\text{pix}}R/(n-1)\tan A = 33.4\,\text{mm}$.

6. $$Q = N(0) - N(45)/N(0) + N(45) = 2{,}000 - 1{,}000/3{,}000 = +0.33,$$
$$U = N(22.5) - N(67.5)/N(22.5) + N(67.5)$$
$$= 1{,}800 - 1{,}200/3{,}000 = +0.20,$$
$$p = \sqrt{(0.33^2 + 0.20^2)} = 0.39(39\%),$$
$$\theta = \tfrac{1}{2}\tan^{-1}(U/Q) = 15.6°,$$
$$\sigma(N)/N = 1/\sqrt{N} = 1/\sqrt{3000} = 1.8\%.$$

7. $R = 20{,}000$ at $\lambda = 0.5\,\mu\text{m}$; an air-spaced etalon of finesse 40; find gap d and free spectral range $\Delta\lambda_{\text{FSP}}$? $\Delta\lambda_{\text{FSP}} = \lambda^2/2nd$, but $2nd = R\lambda/F = 250\,\mu\text{m}$, giving $\Delta\lambda_{\text{FSP}} = 0.001\,\mu\text{m}$.
8. Scan length $= \frac{1}{4}\lambda R = (10\,\mu\text{m})(100{,}000)/4 = 25\,\text{cm}$.

CHAPTER 6

2. $n = \sqrt{n_g} = \sqrt{4} = 2$. Thickness $= \lambda/4n = 2.2\,\mu\text{m}/8 = 0.28\,\mu\text{m}$ (or 2,750 Å).
3. Match the following three detectors to a 0.2 m telescope and then to an 8 m telescope: for the small telescope seeing $= 2''$ and 1 pixel $= 1''$, for the large telescope seeing $= 0.5''$ and we have $0.25''$/pixel. Kodak KAF-4200 CCD with $9\,\mu\text{m}$ pixels in a $2{,}048 \times 2{,}048$ format:
   (a) $f/\# = 9.3$, FOV $= 34'$
   (b) $f\# = 0.93$, FOV $= 8.5'$.
   SITe CCD with $22\,\mu\text{m}$ pixels in a $1{,}024 \times 1{,}024$ format:
   (a) $f\# = 22.7$, FOV $= 17'$
   (b) $f/\# = 2.27$, FOV $= 4.27'$.
   Hughes-SBRC InSb array with $27\,\mu\text{m}$ pixels in a $1{,}024 \times 1{,}024$ format:
   (a) $f\# = 27.8$, FOV $= 17'$
   (b) $f/\# = 2.78$, FOV $= 4.27'$.
   The Kodak CCD cannot be matched to a large telescope unless the scale is reduced to less than $0.1''$/pixel (FOV $= 3.4'$).
4. Linear size $= \beta f$ and $f = 50\,\text{mm}$ given
$$\beta = [1.5(4(1.5) - 1)]/[(128)(1.5 + 2)(1.5 - 1)^2(2)^3] = 0.00837 \text{ radians}$$
   Linear size $= 418.5\,\mu\text{m} \gg$ CCD pixel, therefore must use a multi-element lens system.
5. For a 1-second-of-arc diameter blur the focal ratio is given by
$$F^3 = 206{,}265/128(\beta) = 1{,}611, \qquad \text{hence } F = 11.724$$

6. Sagittal image blur due to coma 60″ off axis (2.9×10^{-4} rad) for an $f/3$ mirror is $\beta = 60''/16(3)^2 = 0.42''$.
   The tail is 3× longer or 1.25″. For $f/1.5$, the blur is $(3/1.5)^2 = 4$ times greater = 1.67″.
7. (a) Diameter of diffraction blur = 2.44 (0.5 μm)(2) = 2.44 μm for $f/2$ lens at 500 nm.
   (b) Depth of focus: $\Delta f = \pm 2\lambda(F)^2 = \pm 4$ μm.
8. Assuming $\alpha = 24 \times 10^{-6}\ \mathrm{K}^{-1}$, Young's modulus $E = 10 \times 10^6$ psi and the yield strength of the aluminum strut = 40,000 psi.
   Since $F = EA(\Delta L/L)$ and $\Delta L/L = \alpha \Delta T$, therefore $F/A = -\alpha E \Delta T = 50{,}400$ psi.
   The strut will buckle because the stress exceeds the yield strength.
9. $A = 5\,\mathrm{m}^2$, $\varepsilon = 5\% = 2F_{hc}$, $T_c = 77$ K:
   (a) for $T_h = 300$ K, $Q_H = 57$ W
   (b) for $T_h = 275$ K, $Q_h = 40$ W.
   Add floating shields or multi-layer insulation.
10. Spectrometer: slit width = w, slit height = h

$$A\Omega = (\tfrac{1}{4})\pi D_{\mathrm{coll}}^2 (w/f_{\mathrm{coll}})(h/f_{\mathrm{coll}})$$

Seeing-limited camera: seeing disk diameter = θ

$$A\Omega = (\tfrac{1}{4})\pi D_{\mathrm{coll}}^2 (w/f_{\mathrm{coll}}(\tfrac{1}{4}\pi\theta^2/f_{\mathrm{coll}}^2)$$

CHAPTER 7

1. See Figures 7.7 and 7.8. Draw the timing diagram very carefully.
4. Inverted operation attracts minority carriers from the channel stops which fill surface traps and eliminate dark current.
8. EMCCD gain: with a 40 V clock there is a 1% chance per transfer of creating a second electron by avalanche multiplication. For 600 elements in the register the average gain is $G = (1.01)^{600} = 392$.

CHAPTER 8

3. Last pixel has $n = 2{,}048 + 2{,}048 = 4{,}096$ transfers. CTE = 0.99999, fraction of original charge (Q_0) left after n transfers is

$$Q' = (\mathrm{CTE})^n Q_0 = (0.99999)^{4096} Q_0 = 0.9599$$

or 96%. One can also use

$$Q' \approx \{1 - n(1 - \mathrm{CTE})\} Q_0 = \{1 - 4096(1 \times 10^{-5})\} Q_0$$
$$= \{1 - 0.04096\} Q_0 = 0.9590$$

or 96%.

5. "kTC-noise" $= (1/e)\sqrt{(kTC)}$ electrons. For $T = 150\,\mathrm{K}$, $C = 0.5 \times 10^{-12} F$, noise $= 201e^-$.
6. Shot noise on "pre-flash" $= \sqrt{400} = 20$ electrons. This adds in quadrature with readout noise of 15 electrons. Final noise $= \sqrt{\{(20)^2 + (15)^2\}} = 25e^-$.
7. Picking up 50-cycle or 60-cycle harmonics from mains; probably due to a ground-loop.

CHAPTER 9

2. $V = (0.25)S + 18.75$; linear regression fit. Gain factor $g = 1/0.25 = 4.0e^-/DN$. Readout noise $R = g\sqrt{18.75} = 17.3e^-$.
3. $C = 0.1\,\mathrm{pF}$, $A_{sfd} = 0.75$, $n = 16$, $V_{fs} = 10$ volt, $g = 25$ electrons/DN.
   The total gain is given by $A_g = V_{fs}C/2^n ge$, therefore

$$A_g = [10(0.1 \times 10^{-12})]/[2^{16}(25)(1.6 \times 10^{-19})] = 3.81$$

   Dividing A_g by the source follower gain (0.75) gives the gain of the preamp/postamp combination as 5.09.
5. Flat-fielding is a multiplicative or gain correction. Fringing is an additive term.
10. Given $m = 24$, V band, $D_{\mathrm{tel}} = 4\,\mathrm{m}$, $\tau\eta = 0.30$, $\theta_{\mathrm{pix}} = 0.3''$, $R = 10e^-$, and $I_d \sim 0$

$$\begin{aligned} F &= (3.92 \times 10^{-12} \times 10^{-0.4(24)}\ \mathrm{W\,cm^{-2}\,\mu m^{-1}}) \\ &\quad \times (0.55\,\mu\mathrm{m})(0.09\,\mu\mathrm{m})(0.3)\frac{(125{,}600\,\mathrm{cm}^2)}{(1.99 \times 10^{-19}\ \mathrm{J\,\mu m})} \\ &= 9.2e^-/\mathrm{s} \end{aligned}$$

   Assuming $1''$ seeing then star flux is diluted over $n \sim 9$ pixels. Average flux $= 1e^-$/s/pixel. If sky $= 22$ mag per square arcsecond then $F_{\mathrm{sky}} = 6.3 \times F \times (0.3)^2 = 5.2e^-$/s/pixel. Background dominates over signal by about 5 : 1 on average (ignores seeing profile). Readout noise $= 10e^-$, so equivalent flux $= 100e^-$. Must integrate for $t > R^2/F_{\mathrm{sky}} = 100/5.2 = 19.2\,\mathrm{s}$ to become background-limited; easily achieved, therefore sky-limited. Signal-to-noise ratio $\approx F\sqrt{t}/\sqrt{(nF_{\mathrm{sky}})}$ ignores signal and readout noise; assumes perfect flats. For S/N $= 10$ then $10 = 9.2\sqrt{t}/\sqrt{(9 \times 5.2)}$, or $1.34\sqrt{t} = 10$ and hence $t = 55.3\,\mathrm{s}$.

CHAPTER 10

1. Data rate = (number pixels digitized × number of bytes/pixel)/(frame time)

$$= 1M \times 2/1{,}024^2 \times 10^{-4} = 0.02\,\mathrm{Mbyte/s}$$

   IR detector data rate is $100/5 = 20$ times faster or ~0.4 Mbyte/s. For 32 outputs the data rate is 20/32 times slower again (63%) or 0.24 Mbyte/s.

CHAPTER 11

No entries

CHAPTER 12

No entries

CHAPTER 13

1. Wavelengths are much larger. Assuming diffraction-limited
   angular resolution $= 70^{\circ}\lambda/D = 70^{\circ}(6\,\text{cm})/(26 \times 100\,\text{cm}) = 9.7'$

CHAPTER 14

No entries

Appendix A

Powers-of-10 notation

When writing numbers which are very large or very small it is useful to introduce a shorthand notation called "powers of 10" or scientific notation. Thus, instead of writing 1,000,000 for 1 million, we write 1×10^6 and understand this to mean 1 followed by six zeros. Similarly, instead of writing 0.000001 for one-millionth, we write 1×10^{-6} where the minus sign tells us that this number is 1 divided by 1 million. The leading number (mantissa) should be a value between 1 and 10 (e.g., 1.496). Names and symbols are given to the most frequently used powers of 10 as shown below. These names and symbols can be prefixed to any of the units of measurements given below to infer that the basic unit is to be multiplied by that power of 10. Note the case of the symbol!

Name	*Symbol*	*Power*	*Name*	*Symbol*	*Power*
yocto	y	10^{-24}	yotta	Y	10^{24}
zepto	z	10^{-21}	zeta	Z	10^{21}
atto	a	10^{-18}	exa	E	10^{18}
femto	f	10^{-15}	peta	P	10^{15}
pico	p	10^{-12}	tera	T	10^{12}
nano	n	10^{-9}	giga	G	10^{9}
micro	μ	10^{-6}	mega	M	10^{6}
milli	m	10^{-3}	kilo	k	10^{3}
centi	c	10^{-2}	hecto	h	10^{2}
deci	d	10^{-1}	deka	da	10^{1}

Notes:

(i) $10^0 = 1$ and in fact $N^0 = 1$ where N is any number.
(ii) $N^a \times N^b = N^{a+b}$ and $N^a \div N^b = N^{a-b}$ (e.g., $10^8 \times 10^2 = 10^{10}$).
(iii) $(N^a)^b = N^{ab}$ (e.g., $(10^6)^2 = 10^{12}$).
(iv) We use the definition of 1 billion as 1,000,000,000 (10^9) and 1 trillion is 10^{12}.

If $x = a^y$, then the index y is called the *logarithm* of x to the base a; the logarithm is the number to which the base must be raised to produce the value x. The logarithm (log) of a product is the sum of the individual logs; $\log(cd) = \log(c) + \log(d)$. Similarly, the log of a ratio is the difference of the logs: $\log(c/d) = \log(c) - \log(d)$. Another useful result is that $\log(c^b) = b\log(c)$. These rules are needed to understand and manipulate astronomical magnitudes. There are two commonly used bases, the base-10 system ($a = 10$) and the natural or Naperian system ($a = e \approx 2.718$); the latter is usually given the symbol *ln* instead of *log*. The $\log(10) = 1$, the $\log(1) = 0$, and the $\log(0.1) = -1$.

Decibels (dB) are a logarithmic scale for noise power $\Delta P/P$ and $1\,\text{dB} = 10\log(\Delta P/P)$ or, in terms of voltage noise, $1\,\text{dB} = 20\log(\Delta V/V)$. When $\Delta P/P = 0.5$ then $\text{dB} = -3.0$, and thus the level at which the signal has dropped by 50% is called the "3 dB point".

The Greek alphabet

alpha α	beta β	gamma γ	delta δ, Δ	epsilon ε	zeta ζ	eta η
theta θ	iota ι	kappa κ	lambda λ	mu μ	nu ν	xi ξ
omicron o	pi π	rho ρ	sigma σ, Σ	tau τ	upsilon υ	phi φ
chi χ	psi ψ	omega ω, Ω				

Appendix B

Units of measurement and useful conversions

Quantity	*SI unit*	*Symbol*
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

SI is the Système International; sometimes also called the MKS (for meter, kilogram, second) system. See NIST reference: *http://physics.nist.gov/cuu/units/*

Quantity	*Symbol*	*Derived unit*	*Equivalent*
Force	F	newton, N	$1\,\mathrm{N} = 1\,\mathrm{kg\,m/s^2}$
Energy	E	joule, J	$1\,\mathrm{J} = 1\,\mathrm{N\,m} = \mathrm{m^2\,kg\,s^{-2}}$
Power	P	watt, W	$1\,\mathrm{W} = 1\,\mathrm{J/s} = \mathrm{m^2\,kg\,s^{-3}}$
Electric charge	Q	coulomb, C	$1\,\mathrm{C} = 1\,\mathrm{A\,s}$
Potential	V	volt, V	$1\,\mathrm{V} = 1\,\mathrm{J/C} = 1\,\mathrm{W/A} = \mathrm{m^2\,kg\,s^{-3}\,A^{-1}}$
Capacitance	C	farad, F	$1\,\mathrm{F} = 1\,\mathrm{C/V} = \mathrm{m^{-2}\,kg^{-1}\,s^4\,A^2}$
Resistance	R	ohm, Ω	$1\,\Omega = 1\,\mathrm{V/A} = \mathrm{m^2\,kg\,s^{-3}\,A^{-2}}$
Frequency	ν	hertz, Hz	$1\,\mathrm{Hz} = 1\,\mathrm{s^{-1}}$
Pressure	P	pascal, Pa	$1\,\mathrm{Pa} = 1\,\mathrm{N/m^2} = \mathrm{m^{-1}\,kg\,s^{-2}}$

Notes: other units and conversions:

$$
\begin{aligned}
1\text{ micron} &= 1\text{ micrometer }(\mu\mathrm{m}) = 10^{-6}\,\mathrm{m}\\
1\text{ angstrom (A)} &= 10^{-10}\,\mathrm{m} = 10^{-8}\,\mathrm{cm} = 0.1\,\mathrm{nm}\\
1\text{ inch (in)} &= 2.54\,\mathrm{cm} = 25.4\,\mathrm{mm}\text{ (exact)}\\
1\,\mathrm{cm} &= 0.3937\,\mathrm{in}\\
1\,\mathrm{m} &= 39.37\,\mathrm{in}\\
12\,\mathrm{in} &= 1\,\mathrm{ft}\\
5{,}280\mathrm{ft} &= 1\text{ mile (mi)}\\
1\,\mathrm{mi} &= 1.6093\,\mathrm{km} \approx 8/5\,\mathrm{km}\\
1\,\mathrm{km} &= 0.6214\,\mathrm{mi} \approx 5/8\,\mathrm{mi}\\
1\text{ nautical mile} &= 1{,}852\,\mathrm{m}\\
1\text{ ounce (oz)} &= 28.3\,\mathrm{g}\\
1\,\mathrm{g} &= 0.0353\,\mathrm{oz}\\
1\,\mathrm{kg} &= 2.2046\text{ pound (lb)}\\
1\,\mathrm{lb} &= 0.4536\,\mathrm{kg}\\
1\text{ metric ton (tonne, t)} &= 1{,}000\,\mathrm{kg} = 2{,}204.6\,\mathrm{lb}\\
1\text{ liter (L)} &= 10^{-3}\,\mathrm{m^3} = 1{,}000\,\mathrm{cm^3} = 1{,}000\,\mathrm{mL}\\
\text{A temperature change of 1 degree Kelvin (1 K)} &= 1^{\circ}\mathrm{C}\text{ (Celsius)}\\
&= 1.8^{\circ}\mathrm{F}\text{ (Fahrenheit)}\\
\text{Freezing point of water} &= 0^{\circ}\mathrm{C} = 273.15\,\mathrm{K} = 32^{\circ}\mathrm{F}
\end{aligned}
$$

$$
\begin{aligned}
1\ \text{day (d)} &= 24\ \text{h} = 86{,}400\ \text{s} \\
1\ \text{sidereal year (yr)} &= 365.256\ \text{d} = 3.156 \times 10^{7}\ \text{s} \\
1\ \text{erg} &= 10^{-7}\ \text{joule (J)} \\
1\ \text{J} &= 10^{7}\ \text{erg} \\
1\ \text{calorie (cal)} &= 4.186\ \text{J} \\
1\ \text{kWh} &= 3.600 \times 10^{6}\ \text{J} \\
1\ \text{horsepower (hp)} &= 550\ \text{ft lb/s} = 745.7\ \text{W} \\
1\ \text{atmosphere (atm)} &= 760\ \text{mm Hg} = 14.70\ \text{lb/in}^2 \\
&= 1.013 \times 10^{5}\ \text{Pa} \\
1\ \text{bar} &= 10^{5}\ \text{Pa} = 0.9870\ \text{atm} \\
1\ \text{torr} &= 1\ \text{mm Hg} = 133.3\ \text{Pa} = 1.33\ \text{mbar} \\
1\ \text{radian} &= 180/\pi = 57.29583^{\circ} = 3437.75' \\
&= 206265'' \\
1'' &= 4.848\ \mu\text{rad}
\end{aligned}
$$

Notes:

(i) One *electronvolt* (eV) is the energy gained by an electron of charge e after being accelerated by a potential (V) of 1 volt: $1\,\text{eV} = e \times 1\,\text{V} = 1.602 \times 10^{-19}$ J (approximately). Alternatively, 1 joule $= 6.242 \times 10^{18}$ eV.
(ii) The product RC of a resistance R and a capacitance C has units of time, and is known as the "RC time constant".

Prefixes for binary multiples

In December 1998 the International Electrotechnical Commission (IEC) approved standard names and symbols for prefixes for binary multiples.

Factor	*Name*	*Symbol*	*Origin*	*Derivation*
2^{10}	kibi	Ki	kilobinary $(2^{10})^1$	kilo $(10^3)^1$
2^{20}	mebi	Mi	megabinary $(2^{10})^2$	mega $(10^3)^2$
2^{30}	gibi	Gi	gigabinary $(2^{10})^3$	giga $(10^3)^3$
2^{40}	tebi	Ti	terabinary $(2^{10})^4$	tera $(10^3)^4$
2^{50}	pebi	Pi	petabinary $(2^{10})^5$	peta $(10^3)^5$
2^{60}	exbi	Ei	exabinary $(2^{10})^6$	exa $(10^3)^6$

Examples:

One kibibit, $1\,\mathrm{Kibit} = 2^{10}\,\mathrm{bit} = 1024\,\mathrm{bit}$, but one kilobit, $1\,\mathrm{kbit} = 10^{3}\,\mathrm{bit} = 1{,}000\,\mathrm{bit}$. One mebibyte, $1\,\mathrm{MiB} = 2^{20}\,\mathrm{B} = 1{,}048{,}576\,\mathrm{B}$, but one megabyte, $1\,\mathrm{MB} = 10^{6}\,\mathrm{B} = 1{,}000{,}000\,\mathrm{B}$.

Appendix C

Physical and astronomical constants

Physical constants

Constant	*Symbol*	*Value*
Speed of light in vacuum	c	$2.99792458 \times 10^{8}\,\mathrm{m\,s^{-1}}$ (exact)
Charge on the electron	e	$1.6022 \times 10^{-19}\,\mathrm{C}$
Planck constant	h	$6.6261 \times 10^{-34}\,\mathrm{J\,s}$
Boltzmann constant	k	$1.3807 \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Electron mass	m_e	$9.1094 \times 10^{-31}\,\mathrm{kg}$
Proton mass	m_p	$1.6726 \times 10^{-27}\,\mathrm{kg}$
Avogadro's number	N_A	$6.0221 \times 10^{23}\,\mathrm{mol^{-1}}$
Gas constant	R	$8.3145\,\mathrm{J\,mol^{-1}\,K^{-1}}$
Gravitational constant	G	$6.6743 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Permeability constant	μ_0	$4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}$ (exact)
Permittivity constant	ε_0	$8.8542 \times 10^{-12}\,\mathrm{F\,m^{-1}}$
Stefan–Boltzmann constant	σ	$5.6704 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$
Pi	π	3.14159
Base of natural logs	e	2.71828

The constant hc has the value 1.99×10^{-25} J m, or expressing 1 m as 10^{6} μm then $hc = 1.99 \times 10^{-19}$ J μm, or $hc = 1.2422$ eV μm.

Astronomical data

Mean radius of the Earth	$R_\oplus$	6.37×10^6 m
Mass of the Earth	$M_\oplus$	5.98×10^{24} kg
Mean Earth–Sun distance = 1 AU*		1.49598×10^{11} m (≈93 million miles)
Mean radius of the Sun	$R_\odot$	6.96×10^8 m
Mass of the Sun	$M_\odot$	1.989×10^{30} kg
Luminosity of the Sun	$L_\odot$	3.827×10^{26} W
1 Lightyear	lt-yr	9.4605×10^{15} m (≈5.9 trillion miles)
1 Parsec	pc	3.0857×10^{16} m = 3.2616 lt-yr = 206,265 AU*

* AU = Astronomical Unit.

Blackbody radiation; the Planck function

$$B_\lambda = \frac{2hc^2}{\lambda^5}\frac{1}{(e^{hc/k\lambda T}-1)}\,\mathrm{W\,m^{-2}\,m^{-1}\,sr^{-1}}$$

The maximum of B_λ occurs at $\lambda_{\max} = 2{,}898/T$ for $\lambda_{\max}$ in μm (also called Wien's Displacement Law), or

$$B_\nu = \frac{2h\nu^3}{c^2}\frac{1}{(e^{h\nu/kT}-1)}\,\mathrm{W\,m^{-2}\,Hz^{-1}\,sr^{-1}}$$

The maximum of B_ν occurs at $\nu_{\max} = 58.78T$ for $\nu_{\max}$ in GHz. Note: this is *not* the same as $\lambda_{\max}/c$.

Combinations of constants:

$$2hc^2 = 1.191 \times 10^{-16}\,\mathrm{W\,m^2} \quad \text{and} \quad hc/k = 1.439 \times 10^{-2}\,\mathrm{m\,K}$$

$$2h/c^2 = 1.4745 \times 10^{-50}\,\mathrm{J\,s^3\,m^{-2}} \quad \text{and} \quad h/k = 4.799 \times 10^{-11}\,\mathrm{s\,K}$$

Photon form of Planck Function

Divide B_λ by the photon energy hc/λ:

$$N_\lambda = \frac{2c}{\lambda^4}\frac{1}{(e^{hc/k\lambda T}-1)}\,\mathrm{photons\,s^{-1}\,m^{-2}\,m^{-1}\,sr^{-1}}$$

To convert to square arcseconds note that 1 steradian = 4.255×10^{10} arcsec^2. The maximum in N_λ occurs at $\lambda_{\max} = 3{,}670/T$ for λ in μm.

Appendix D

Astronomical magnitude scale and relation to lux

In the electro-optical industry the term illuminance is used to describe the amount of light received per unit surface area, and it is measured in a unit called *lux*. One lux is the illuminance produced by a standard light source of 1 candela at a distance of 1 meter, and 60 candelas is the luminous intensity of a $1\,\mathrm{cm}^2$ *black body* at the temperature of melting platinum (2,042 K). One lux is also equivalent to 1 lumen per square meter. Such units are never used in astronomy, but are found in reference to CCD-based cameras intended for low light level television applications. An illuminance of 1 lux is a photon flux of about 3×10^9 photons/s/mm^2 at $\lambda = 550\,\mathrm{nm}$, or approximately equivalent to a star of visual magnitude -14. This is roughly equivalent to the full Moon! The brightness in lux of any source in the "visual" waveband is

$$B(\mathrm{lux}) = 10^{-0.4(m_v+14)}$$

The illuminance of Vega ($m_v = 0$) is only 0.000002 lux. Ordinary hand-held commercial TV "camcorders", even those with CCDs, operate at frame rates of 1/60th second and are typically designed with 1-inch lenses to image scenes with light levels greater than a few lux (bright moonlight). Intensified CCD systems are capable of imaging with illumination levels as weak as 0.02 lux. While a 2 lux CCD camcorder cannot take images in a room where the illumination level is equivalent to the star Vega, it could actually image the star itself if the lens was removed and the detector attached to a modest telescope. The telescope collects more light and focuses all of the available light onto a single (or small number) of pixels so that the illuminance is increased by several hundred thousand. If the CCD is cooled to allow integration times of many seconds, instead of 1/60th second, then objects can be recorded which are several hundred times fainter still.

Appendix E

Basic observational astronomy facts

Coordinates: Positions on the sky are given with reference to the Celestial Equator and the Celestial Poles as shown in Figure A5.1. The coordinates are analogous to longitude and latitude on the Earth's surface and are called Right Ascension (RA or α) and Declination (Dec or δ). RA is measured eastwards from the First Point of Aries, an arbitrary marker like the Greenwich meridian, and Dec is measured $\pm 90^{\circ}$ from the Celestial Equator.

Luni-solar precession: The First Point of Aries moves backwards along the Celestial Equator at the rate of 50.2 arcseconds per year. Correcting for this effect yields the mean equator and mean equinox.

Nutation: The wobble of the Earth's axis as it precesses. Correction for this effect gives the true equator and true equinox.

Often, all these corrections are done at the observatory by the telescope control program, but formulas are included in the Astronomical Almanac which is published annually and jointly by the U.K. and the U.S.A. Published coordinates are usually referred to a specific equinox such as J2000. For the nearest objects, an additional correction for proper motion (the component of the object's true motion in space at right angles to our line of sight) must be made; the rates are tabulated in surveys of proper motion. The object's true position at the present epoch will depend on the interval of time since the epoch of the tabulated coordinates.

To find out when an object is observable you need to convert from RA and Dec to hour angle and zenith distance (or zenith angle) for your latitude. Several basic relationships are given below; it is useful to be familiar with the celestial sphere and basic spherical trigonometry (Green, 1985).

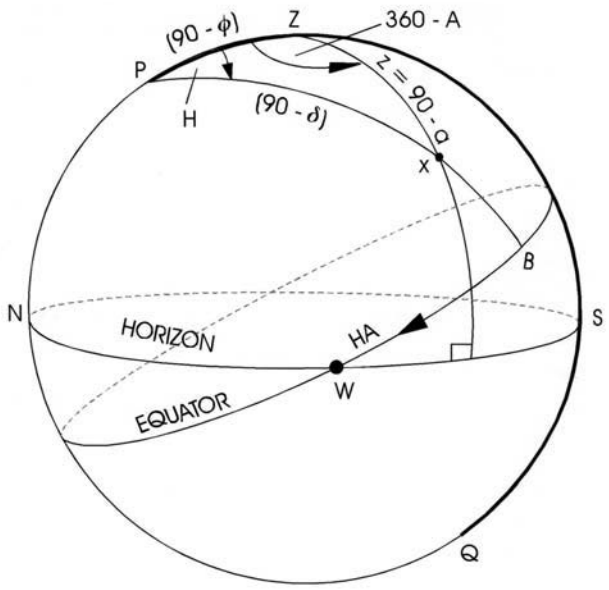


Figure E5.1. The celestial sphere and spherical triangle relations. The north celestial pole is at P, and ϕ is the observer's latitude.

Hour angle: The Hour Angle (H) is measured westwards from 0 on the meridian; east is therefore negative and west is positive. In practice, H is given in hours of time—and not in degrees—because this gives a more intuitive indication of when the object can be observed; for example, 3 hours east of the meridian ($H = -3$ h) means that it will be three hours until that object crosses the meridian. The usual conversion between degree and hours applies:

$$360^\circ = 24\,\text{h}, \quad 15^\circ = 1\,\text{h}, \quad 1^\circ = 4\,\text{min},$$
$$15' = 1\,\text{min}, \quad 1' = 4\,\text{s},$$
$$15'' = 1\,\text{s}, \quad 1'' = 0.067\,\text{s}$$

Note: Conversions from angle to time on the sky depend on declination (δ). The interval of time parallel to the RA direction is shortened by the factor $\cos\delta$. On the equator $\delta = 0$ and $\cos\delta = 1$. At the poles $\delta = 90^\circ$ and $\cos\delta = 0$.

Zenith distance: ZD or $z = 90^\circ - a$, where a is called the altitude or elevation angle above the horizon; a ZD $> 90^\circ$ implies that the object is below the horizon.

Azimuth: Azimuth (A) is the compass bearing in degrees measured north through east.

Local sidereal time: LST is the hour of right ascension on the meridian at that moment. Most observatories have a sidereal clock. There are approximately 23

hours, 56 minutes, and 4.1 seconds in one sidereal day. The sidereal time at midnight on March 22 is 12 hours, and the LST advances about 2 hours per month. By September 21 the LST is 24 (or 0) hours, and thus this is the RA on the meridian at midnight. In general: $LST = RA + HA$. LST is tabulated at 0 hours Universal Time for every day of the year in the Astronomical Almanac.

Universal Time is the time on the Greenwich Meridian. Most observatories have a clock keeping UT. Exact UT can be obtained from radio signals issued by the WWV service.

Julian Date (no connection to the Julian Calendar) is a numerical day count from an arbitrary zero point. Julian Day number is the number of days that have elapsed since noon Greenwich Mean Time on January 1, 4713 BC plus the decimal fraction of a day since the preceding noon up to the event being recorded. Julian Dates are given in the Astronomical Almanac.

Coordinate transformations: Relationship between Hour Angle (H) and Dec (δ) given azimuth (A) and elevation (a):

$$\sin(\delta) = \sin(a)\sin(\varphi) + \cos(a)\cos(\varphi)\cos(A)$$

$$\cos(H) = [\sin(a) - \sin(\delta)\sin(\varphi)]/\cos(\delta)\cos(\varphi)$$

and $RA = LST - H$. Relationship between azimuth (A) and altitude/elevation (a) given the Hour Angle (H) and Dec (δ):

$$\sin(a) = \sin(\delta)\sin(\varphi) + \cos(\delta)\cos(\varphi)\cos(H) \quad \text{where } H = LST - RA$$

$$\cos(A) = [\sin(\delta) - \sin(\varphi)\sin(a)]/\cos(\varphi)\cos(a)$$

$$\text{or} \quad \tan(A) = \sin(H)/[\sin(\varphi)\cos(H) - \cos(\varphi)\tan(\delta)],$$

and the variation of the parallactic angle (q = angle PXZ) with time is given by

$$\tan(q) = \sin(H)/[\tan(\varphi)\cos(\delta) - \sin(\delta)\cos(H)]$$

$$\text{or} \quad \sin(q) = \cos(\varphi)\sin(H)/\sin(\alpha)$$

The elevation (or altitude) of the celestial pole above the horizon in degrees is equal to the latitude (φ) of the site. A plot of airmass (i.e., sec z) vs. time (or HA) will tell you the best time to observe your source. The times of sunset, sunrise, and the position of the Moon are tabulated in the Astronomical Almanac. The times of sunset, sunrise, and twilight are tabulated at 4-day intervals of time and 10-degree intervals of latitude.

References

Green, R.M. (1985) *Spherical Astronomy*, Cambridge University Press, Cambridge, U.K.

Roy, A.E.; and Clarke, D. (2003) *Astronomy: Principles and Practice*, fourth edition, Institute of Physics, Bristol, U.K.

Appendix F

Useful statistics

Mean and weighted mean:

$$\bar{x} = \frac{\sum(x_i/\sigma_i^2)}{\sum(1/\sigma_i^2)} \rightarrow \frac{1}{N}\sum x_i, \quad \sigma_i = \sigma$$

where σ is the standard deviation of a single observation from the mean; and N is the number of observations. The standard error of the mean is $\sigma/\sqrt{N}$.

Variance:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N-1}$$

is determined from the sum of the squares of the "residuals" or differences from the mean. Error in a sum (or difference) of two random variables (A and B):

$$A \pm \sigma_A, \quad B \pm \sigma_B$$
$$C = (A \pm B) \pm \sqrt{\sigma_A^2 + \sigma_B^2}$$

The errors add in quadrature.

Error in a ratio of two random variables (A and B):

$$A \pm \sigma_A, \quad B \pm \sigma_B$$
$$C = \frac{A}{B} \pm C\sqrt{\left(\frac{\sigma_A}{A}\right)^2 + \left(\frac{\sigma_B}{B}\right)^2}$$

The fractional errors add in quadrature.

The Chi-squared test

Compares the observed frequency distribution $f(x_i)$ of possible measurements x_i with the predicted distribution $NP(x_i)$, where N is the number of data points and $P(x_i)$ is

the theoretical probability distribution:

$$\chi^2 = \sum_{j=1}^{n} \frac{[f(x_j) - NP(x_j)]^2}{NP(x_i)}$$

The "reduced" chi-squared is χ^2/v, where v is called the "degrees of freedom" and is given by $N - p$, where N is the number of data points and p is the number of parameters determined from those data points.

The Gaussian distribution

For a very large number of observations (n) of a random variable, the probability of obtaining the value x is

$$P(x, \bar{x}, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \bar{x}}{\sigma}\right)^2\right]$$

where the standard deviation σ is related to the full width at half maximum (FWHM) of the distribution by $\text{FWHM} = 2.354\sigma$, and to the probable error by $\text{PE} = 0.6745\sigma$. The probable error corresponds to the range of the variable which contains 50% of the measurements. Error bars giving estimates of σ should be applied to all measurements. Assuming that the errors follow a Gaussian distribution, the probability of finding the variable in the interval $\pm 1\sigma$ is 68%. For $\pm 2.5\sigma$ it is 98.7%.

The Poisson distribution

This is a discrete probability distribution that expresses the probability of a number of events occurring in a fixed period of time if these events occur with a known average rate and independently of the time since the last event. If the expected number of occurrences in this time interval is λ, then the probability that there are exactly k occurrences (k being a non-negative integer, $k = 0, 1, 2, \ldots$) is equal to

$$f(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Here, e is the base of natural logarithms ($e = 2.71828\ldots$); k is the number of occurrences of an event, the probability of which is given by the function f; and $k!$ is the factorial of k. The number of observed occurrences fluctuates about its mean λ with a standard deviation $\sigma_k = \sqrt{\lambda}$. These fluctuations are called Poisson noise.

Index

Printing: Mercedes-Druck, Berlin
Binding: Stein+Lehmann, Berlin