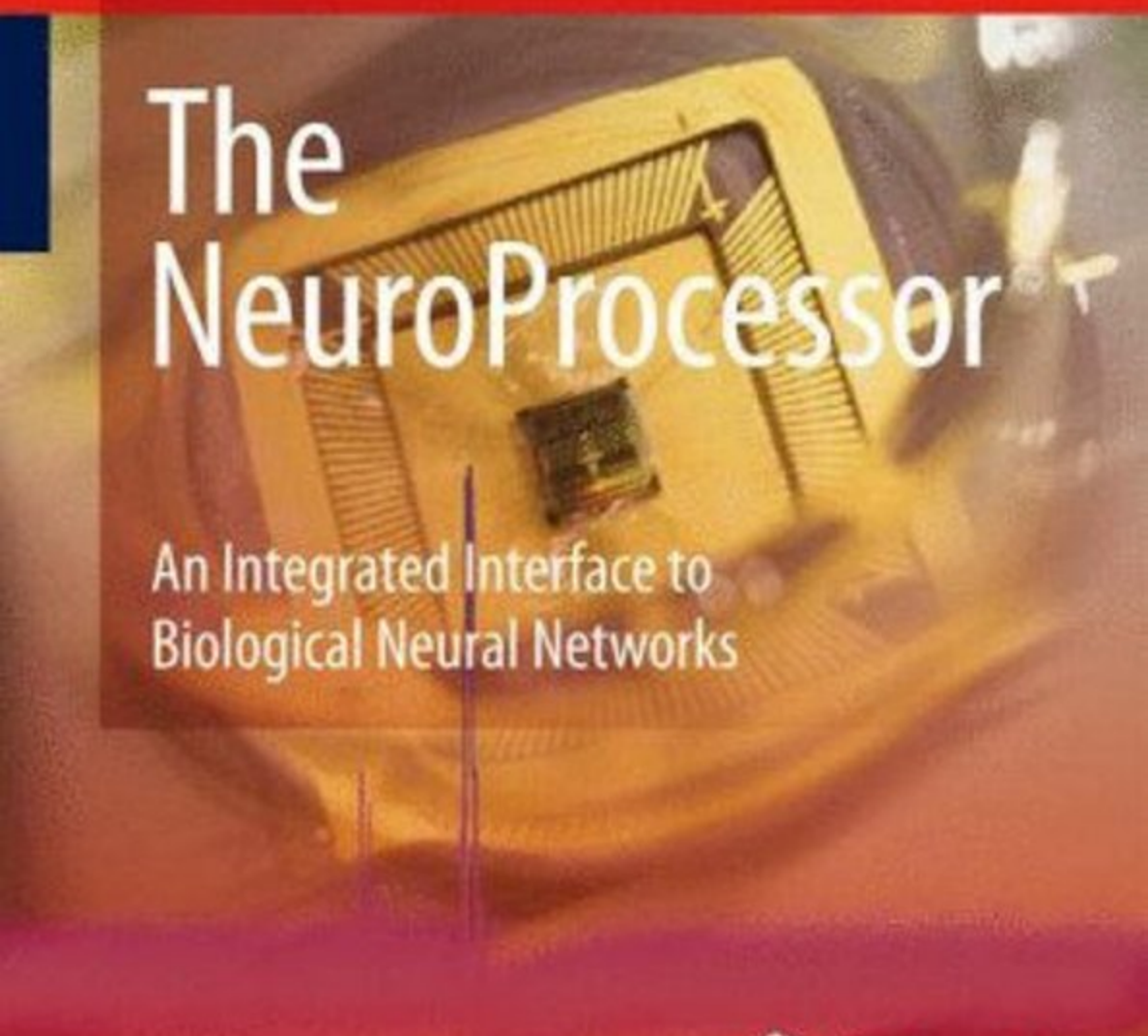
Yevgeny Perelman
Ran Ginosar

The NeuroProcessor

An Integrated Interface to Biological Neural Networks



Springer

The NeuroProcessor

Yevgeny Perelman · Ran Ginosar

The NeuroProcessor

An Integrated Interface to Biological Neural Networks

Springer

Dr. Yevgeny Perelman
Technion-Israel Institute of Technology
Dept. Electrical Engineering
32 000 Haifa
Israel
perelman@tx.technion.ac.il

Prof. Ran Ginosar
Technion - Israel Institute of Technology
Dept. Electrical Engineering
32000 Haifa
Israel
ran@ee.technion.ac.il

ISBN: 978-1-4020-8725-7 e-ISBN: 978-1-4020-8726-4

Library of Congress Control Number: 2008932564

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Contents

1

Introduction

Understanding brain structure and principles of operation is one of the major challenges of modern science. Since the experiments by Galvani on frog muscle contraction in 1792, it is known that electrical impulses lie at the core of the brain activity.

The technology of neuro-electronic interfacing, besides its importance for neurophysiological research, has also clinical potential, so called *neuroprosthetics*. Sensory prostheses are intended to feed sensory data into patient's brain by means of neurostimulation. Cochlear prostheses [1] are one example of sensory prostheses that are already used in patients. Retinal prostheses are currently under research [2].

Recent neurophysiological experiments [3, 4] show that brain signals recorded from motor cortex carry information regarding the movement of subject's limbs (Fig. 1.1). These signals can be further used to control external machines [4] that will replace missing limbs, opening the field of motor prosthetics, devices that will restore lost limbs or limb control.

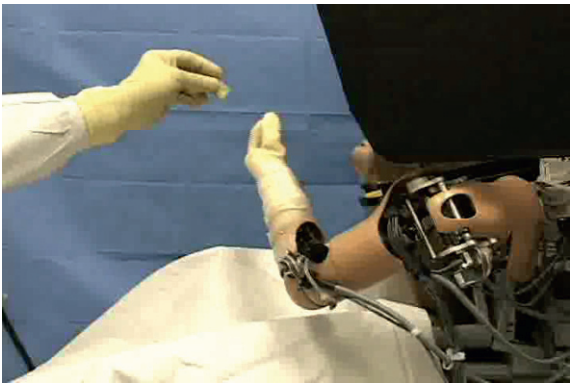


Fig. 1.1. Robotic arm controlled by monkey motor cortex signals. MotorLab, University of Pittsburgh. Prof Andy Schwartz, U. Pitt

Another group of prostheses would provide treatment for brain diseases, such as prevention of epileptic seizure or the control of tremor associated with Parkinson disease [5]. Brain implants for treatment of Epilepsy and Parkinson symptoms (Fig. 1.2) are already available commercially [6, 7].

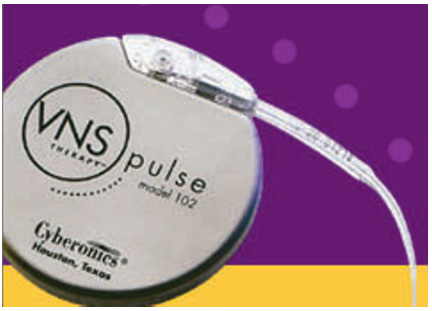


Fig. 1.2. Implantable device for Epilepsy seizures treatment [7]. Cyberonics, Inc. http://www.cyberonics.com/

The "far goal" of neural prosthetics is a device to replace higher-level cognitive functions of damaged brain. It will maintain bi-directional communication with neural tissue, decode, process and feed back neural data in order to replace lost functionality of damaged brain parts. Such devices are yet many years in the future, but even those are already mentioned in the literature [8].

Electronic devices for neuronal interfacing advance as new fabrication technologies have become available. Started as plain metal wires, neuronal interfaces gradually developed into complex micro-fabricated arrays of hundreds of three-dimensional sensing sites [9], some to be used in live animals (so called *in-vivo* experiments), others to sample data from cultured neural networks (*in-vitro* experiments). As neurophysiological research advances, increasing demands on the instrumentation push the interfacing devices towards tighter integration, larger numbers of sensing/stimulating points and wireless operation.

The number of recording sites involved in in-vivo experiments is expected to grow to thousands [10]. The devices for cultured networks interfacing, the Multi-Electrode Arrays, suffer currently from too low spatial resolution (hundreds of recording sites), which will probably grow manyfold. Latest reported state-of-the-art devices fabricated on silicon already include above ten thousand sensing points [11].

Increasing demands of neurophysiology on one hand and the growing complexity of neuro-electrical interfaces on the other hand pose new requirements for electronic devices supporting these interfaces. A very simple experiment can be conducted with a few electrodes connected with a shielded analog cable to an analog signal acquisition PC card. This approach becomes increasingly problematic when the number of electrodes grows larger; it is absolutely

impractical for wireless operation. In the latter case signals must be acquired, digitized and modulated for wireless transmission. Closer examination shows that mere signal acquisition and digitization is not sufficient for wireless operation of large-scale neuronal interfaces; it is simply impossible to transmit all the data acquired from the interface within a reasonable power budget.

It is therefore concluded that a new type of electronic device is needed for the emerging field of neuronal interfaces. This device, the *Neuroprocessor*, would allow *computational neuronal interfaces*. Beyond mere signal acquisition, the Neuroprocessor would perform computation on the acquired signals. At the early stages this computation would extract meaningful information out of raw recordings to minimize the required bandwidth for wireless communication. Later, the Neuroprocessor will interpret the signals and compute the required stimulation to feed back into the tissue and/or control external prosthetic devices.

1.1 Overview of the Book

This book focuses on computational interfaces with biological neural networks, with an emphasis on VLSI technology. Circuits for neuronal data acquisition and shaping are explored, together with algorithms for low-power integrated processing of neuronal data. An effort is also made in integrated in-vitro neuronal interfaces.

The book is organized as follows: A brief background on neuronal communication and microelectrode recording is presented in Chap. 2. An emphasis is placed on selected properties of extracellular microelectrodes. In Chap. 3 we argue that conventional, i.e. "non-computational" neuronal interfaces are insufficient for the evolving needs of neurophysiology research and of the emerging field of neuroprosthetics. We introduce the concept of a computational neuronal interface, the *Neuroprocessor* that performs significant computational tasks near the recording front-end without relying on an external host. The Neuroprocessor allows for significant reduction of the communication link bandwidth, enabling wireless operation of large-scale neuronal interfaces. It also enables autonomous operation, required by neuroprosthetic devices.

An important goal of this work was to develop an integrated, wireless-ready neuronal recording interface that can be incorporated into a multi-channel recording system. As part of this work, three front-end ICs, `NPR01 -NPR03` , were designed, fabricated and evaluated. Along with every IC, a suitable testing environment for electrical characterization was developed. Technical discussions regarding the circuit and architecture design of the first two generations are given in Chap. 4. The third generation of the front-end IC, `NPR03` , is a complete, fully-integrated, mixed-signal multi-channel recording interface. It was embedded into a miniature headstage, successfully tested in neuronal signal recording from a rat cortex. It was also successfully tested in recording

form neural tissue cultured in-vitro. NPR03 and accompanying systems and experiments are discussed in Chap. 5.

Chapters 6 and 7 present spike processing algorithms and in-vitro neuronal interfaces. The appendicies describe the detailed design of NPR01 -NPR03 .

2

Recording From Biological Neural Networks

The core functionality of neural networks is through electrical communication between neurons. Recording and stimulating electrical activity in neural networks is the enabling technology for most neurophysiology-related applications and research. This chapter presents a short description of mechanisms responsible for electrical activity in neurons, theoretical background for electrical transduction between biological medium and electronic circuits and some practical cases of such transducers, the neuronal probes. Finally, we describe a typical setup for multi-electrode recording and the informative content of the recorded signal.

2.1 The Neuron

During the second half of the nineteenth century it was largely understood that the brain consisted of a complex, interactive network of single cells (neurons) (Fig. 2.1) [12, 13].

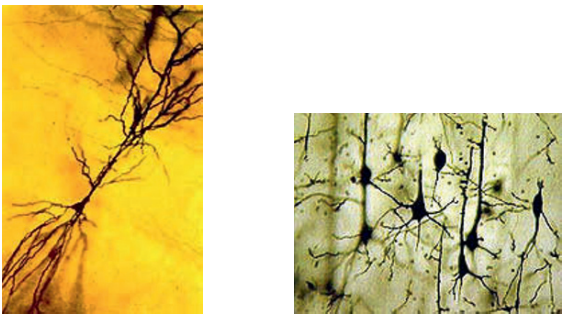


Fig. 2.1. A single neuron and a neural network [14]. Web: Neuroscience for kids. http://faculty.washington.edu/chudler/calpyr.html

Neurons are specialized, non-spherical cells consisting of a cell body (soma), many short dendritic processes, and one longer protrusion called the axon, enclosed by a thin double layer of molecules, the membrane (Fig. 2.2). An axon is a signal transmitter, it delivers the signals generated by the soma

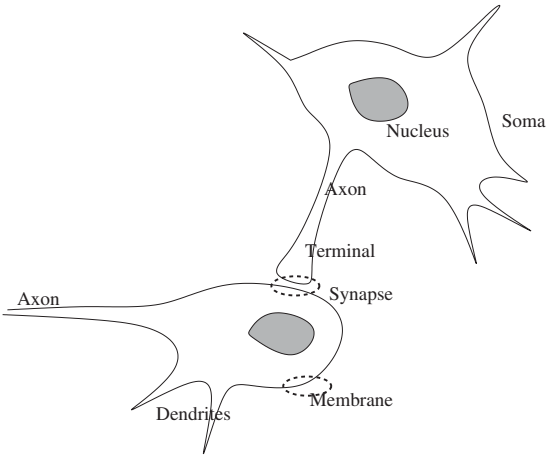


Fig. 2.2. Neurons

to its end terminal. Special chemicals, the neurotransmitters, are released from the terminal. They diffuse through the synapse towards the dendrite or the soma of a receiving (post-synaptic) neuron. Dendrites therefore are the "input terminals" of the neuron, they transduce the chemical synaptic inputs to electric potentials.

2.1.1 The Membrane and Resting Potential

The information is transferred among neurons via electrical potentials, called *action potentials*. These are short (order of 1 msec) deviations of the intracellular electrical potential from the *resting potential*. The neuron potential is controlled by the membrane, through the mechanism of *sodium-potassium pumps*. The mechanism of neuronal membrane operation was quantitatively described in [15], known as the Hodgkin-Huxley model.

The membrane isolates electrically the inside of the cell from the extracellular solution. Being a very thin (about 5 nm) layer of insulator, the membrane is capacitive from the electrical point of view. Sodium (Na+) and potassium (K+) ions can penetrate the membrane through special pores, sodium and potassium channels. The ions traverse the channels across the gradients of their electrochemical potentials. Both sodium and potassium channels are gated: they open or close according to the polarization of the membrane. In addition, a special channel exists: the sodium-potassium pump. It moves K+

and Na+ ions against the potential gradients by absorbing metabolic energy (ATP molecules). For each three Na+ ions moved out of the cell this pump pushes a pair of K+ ions into the cell, pulling the intracellular potential below the extracellular environment.

Due to the sodium-potassium pump operation the intracellular concentration of K+ is much larger than the extracellular concentration. The opposite holds for Na+. The Hodgkin-Huxley model treats the membrane permeability for each ion type as a non-linear conductance that is driven by the ion Nerst potential[1] (Fig. 2.3). g_{Na}, g_K and g_L are the membrane ion conductances

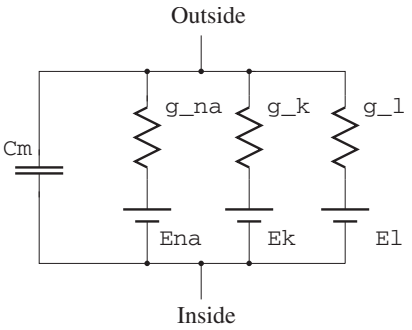


Fig. 2.3. Hodgkin-Huxley model of the neural membrane

for Na+, K+ and leakages. E_{Na}, E_K and E_L are the corresponding Nerst potentials. C_m is the capacitance of the membrane.

When resting, the permeability (the "ease of penetration" through the membrane) of potassium ions is about 100 times larger than that of sodium ions. Related to Fig. 2.3, g_K is 100 times larger then g_{Na}. Thus the intracellular resting potential is slightly larger than E_K. The actual resting cell potential value varies for different cell types between -50 mV and -90 mV [17], measured with respect to the potential of the extracellular solution.

2.1.2 Action Potential

The membrane potential is subject to change, due to the activity of presynaptic neurons: neurotransmitters absorbed by the dendrites perturb slightly the membrane potential. The perturbations are accumulated, resulting in a gradual depolarization of the membrane. The ion channels open gradually due to the membrane depolarization, until it reaches a certain threshold, about 20 mV above the resting potential. Beyond this point, Na channels open rapidly, avalanche-like. Sodium ions enter the membrane, making the

[1] Nerst equation gives the difference in ion potential across the membrane, as a function of an intra- and extracellular ion concentration ratio [16]

inside of the cell positive. The intracellular potential is pulled towards the Nerst potential of Na+ ions, typically some 100 mV above the resting value. With the rise of the cell potential Na conductance declines back to zero. At the same time the potassium conductance rises and K+ ions flow out of the cell. At the potential peak the inward flow of Na+ is exceeded by the outward K+ flow and the potential swings towards E_K, even below its resting value. At this point all the sodium channels are inactivated. The cell has gained some Na+ ions and has lost some K+ ions. The concentrations are restored by means of the sodium-potassium pump (energy consuming), during the *refractory period* that lasts typically a couple of milliseconds (Fig. 2.4).

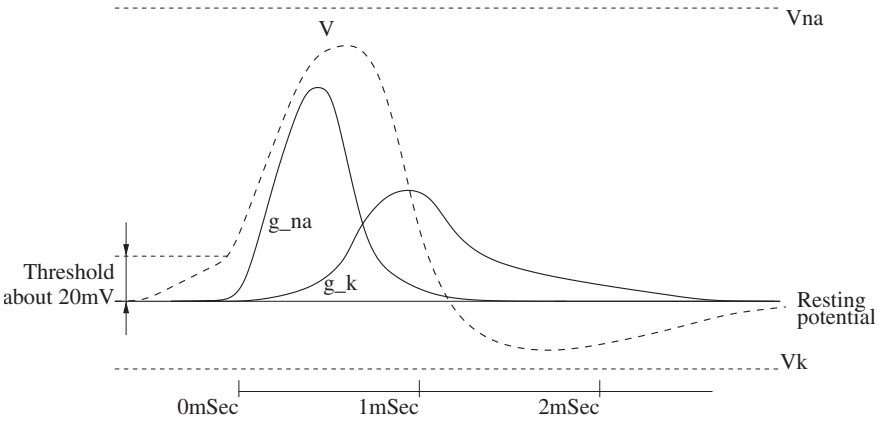


Fig. 2.4. Action potential

A special note about action potentials must be made, from the neuron behavior perspective: Firing of an action potential indicates that the membrane depolarization has gone beyond a particular limit. *The information in action potential is expressed in the bare fact of firing, and not in the shape of the pulse*. In digital communications this form of signalling is termed "pulse position modulation", or PPM.

2.1.3 Excitation Propagation

During an exhibition of action potential, there is a positive charge inside the cell (Na+ ions), while the extracellular volume adjacent to the soma is slightly negative. The excessive concentration of Na+ ions makes them flow out of the soma down the axon. Concurrently, outside the cell the Na+ ions flow towards the soma (due to the negative near the soma), Fig. 2.5. The current flow depolarizes an adjacent section of the membrane thus the excitation impulse travels along the axon. This form of propagation is called "uniform".

Another form of excitation propagation, the "saltatory propagation" happens when the axon membrane is covered by *myelin* cells, except for regularly

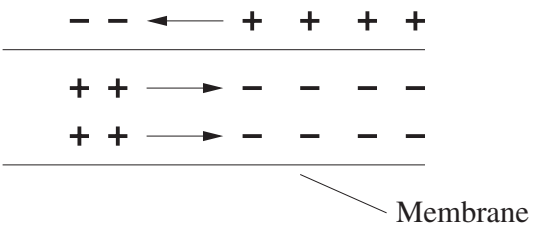


Fig. 2.5. Uniform excitation propagation

spaced points, the *nodes of Ranvier* Fig. 2.6. Since myelin is a good insulator,

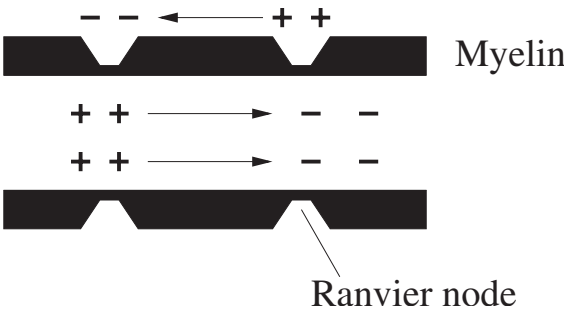


Fig. 2.6. Saltatory excitation propagation

excitation can not occur except in the places where the myelin cover is thinner, nodes of Ranvier. The excitation propagates in "jumps" between adjacent nodes and the impulse travels much faster: propagation speed in a myelinated nerve fiber is 80–120 m/S, while unmyelinated nerve conduction speed is 0.5–2 m/S [12, 17].

Table 2.1 summarizes some of the physical properties of neurons.

Table 2.1. Typical values of neuron physical properties [18]

Soma diameter	5–20 μm
Axon diameter	1–20 μm
Membrane thickness	5 nm
Dendrite length	up to 10 μm
Axon length	up to 1 m
Resting cell potential	-70 mV
Action potential duration	1 mSec
Action potential peak (above resting value)	100 mV

2.2 Interfacing Neurons Electrically

In the biologic environment the currents are carried in the electrolytic medium by means of ion conduction. Electronic circuits, which are commonly used for transduction and processing of neural signals, all use electronic conduction. An *electrode* (either recording or stimulating) provides transduction between these two media. Numerous textbooks treat the electrode-electrolyte interface electrochemically [16] and electronically [17, 19, 20]. First reviews on electrode properties can be found in [21] and [22]. Additional reviews are available in [23] and in [24].

It must be noted that ionic mobility in biological medium is typically six orders of magnitude below the electron/hole mobility in metals or semiconductors [16], thus the time constants of the two media differ significantly: aqueous electrodes operate typically in 10 kHz bandwidth [19].

As we are going to present in detail, the electrode transduction takes place either by capacitive coupling or by charge transfer, in which electrons are transferred to and from the solution ions. The transfer occurs by two types of chemical reactions: *oxidation* (electrons are donated) and *reduction* (electrons are absorbed). When voltages across the interface are low, voltage-driven charge transfers across the junction are negligible and the capacitive effect prevails. This is the common case for recording, which is usually done with a high-impedance preamplifier and no DC currents across the electrode. This mode of operation typically involves small-signal measurements and electrodes are viewed as networks of linear elements (mostly capacitive).

When it comes to neurostimulation (involving non-negligible DC currents), a current flow is conducted through an electrode by means of carrier exchange. Large-signal model of an electrode must be considered, which involves electrochemical mechanisms of charge transduction [19].

2.2.1 Double Layer Capacitance

When an electrode is placed into an electrolyte, a space charge layer builds up at the interface due to various chemical reactions [16]. The build up continues until a sufficiently strong electric field is formed to initiate a reverse reaction. At equilibrium forward and reverse reactions are equal and the net current across the junction is zero; the process resembles a PN semi-conductor junction.

The ion distribution in an electrolyte is modelled as a charge plane near the electrode (outer Helmholz plane, OHP), where the potential drops linearly, like in a common plate capacitor. The charge plane is followed by a cloud of mobile ions with approximately exponential potential drop (Fig. 2.7). The plate capacitance of the Helmholz layer can be calculated as:

$$C_H = \frac{\varepsilon_0 \varepsilon_r A}{d_{OHP}}$$

d_{OHP}, the OHP distance from the electrode, is extremely small, typically less than 10 Å. In [19], the worst-case capacitance of Helmholz layer is estimated at $0.11\,\mathrm{pF}/\mu\mathrm{m}^2$.

A model for the capacitance of the mobile ions was suggested by Gouy and Chapman and is reviewed in [16]. It is voltage-dependent, as the ion distribution depends on the potential applied across the junction:

$$C_D = \frac{\varepsilon_0 \varepsilon_r}{L_D} \cosh \frac{zV_0}{2V_t}$$

where V_0 is the potential over the junction, V_t is the thermal voltage, z is the ion charge, and L_D is the Debye length:

$$L_D = \sqrt{\frac{\varepsilon_0 \varepsilon_r V_t}{2z^2 q n^0}}$$

n^0 is the ion concentration in the solution (ions/liter).

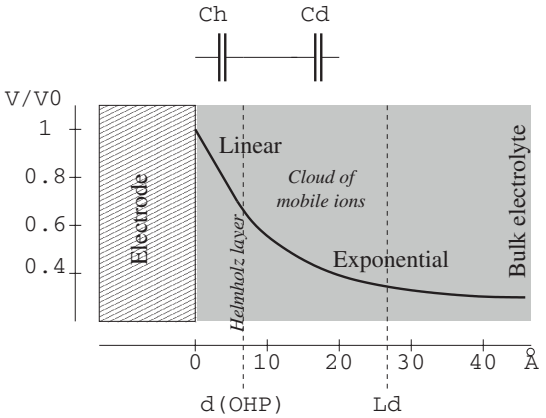


Fig. 2.7. Metal-electrolyte interface potential. Adopted from [24]

The effective capacitance of the electrode-electrolyte interface is the superposition of C_D and C_H:

$$C_I = \left(\frac{1}{C_H} + \frac{1}{C_D} \right)^{-1}$$

For most biological solutions, when a zero bias is applied, C_H and C_D are of the same order. A typical value of a net capacitance is about $0.05\,\mathrm{pf}/\mu\mathrm{m}^2$ [24].

2.2.2 Resistance at the Interface and Charge Transfer

To move charge into or out of the electrode a potential must be applied. Potential shift from the equilibrium value V_0 is called an *overpotential*:

$$\eta = V - V_0$$

There are four processes, each of them is associated with its overpotential. The total overpotential, η:

$$\eta = \eta_t + \eta_d + \eta_r + \eta_c$$

η_t is due to the charge transfer through the double layer, η_d is due to the diffusion of ions in the electrolyte towards the electrode, η_r and η_c are due to chemical reaction at the electrode and due to transfer of metal ions into electrolyte. The last two terms are usually insignificant in biological applications [20]. η_t dominates near the equilibrium. At higher currents, η_d becomes significant due to the limited rate of ion supply from the bulk solution.

At equilibrium, oxidation and reduction proceed at equal rates: $J_0 = J_{OX} = -J_{RED}$. This equilibrium current density, J_0, is called *exchange current density*. η_t can be related to the current density by the Butler-Volmer equation:

$$J = J_0(e^{(1-\beta)z\eta_t/V_t} - e^{-\beta z\eta_t/V_t})$$

β is the *symmetry factor* that reflects the differences in energy barriers of the two reactions. For small deviations from the equilibrium J can be linearized (assuming β of 0.5) as

$$J = J_0 \frac{z\eta_t}{V_t}$$

Thus the near-equilibrium charge transfer area conductance is

$$G_t = J_0 z/V_t$$

This value describes the charge transfer resistance in recording applications, where an electrode is coupled to a high-impedance preamplifier and the net current is zero. It is rather small: for 2×10^{-4}, the largest value of J_0 among those brought in [19], it is only $80\,\mathrm{p}\Omega^{-1}/\mu\mathrm{m}^2$. Thus an electrode with area of $1000\,\mu\mathrm{m}^2$ will have (at small bias) R_t of $12.5\,\mathrm{M}\Omega$.

In stimulation applications, where significant deviations from the equilibrium occur, R_t is significantly smaller. For instance, in a certain experiment [25], a $100\,\mu$A current was conducted through a $700\,\mu\mathrm{m}^2$ electrode by applying only a 1V potential.

2.2.3 Diffusion Resistance Near DC

When an electrode conducts a steady state current, an ion concentration is increased near the electrode with respect to the bulk solution. The concentration is due to the diffusion of ions from the solution towards the electrode. The diffusion process causes an overpotential to be developed, η_d.

For any electrode, there is some limiting rate at which ions can be supplied from the bulk. Let J_s be a corresponding limiting current density. Diffusion overpotential at current J is given by [20]:

$$J/J_s = 1 - e^{-|\eta_d|z/V_t}$$

This equation is valid for near-DC conditions.

2.2.4 AC Diffusion Resistance

Imagine a sinusoidal potential applied to the electrode. It would force a sinusoidally varying spatial concentration of ions. The variation will be largest at the interface and will decay deeper in the bulk. Damping of the variations will increase as the frequency increases. Thus the length to which the effect extends into the solution decreases with frequency and the concentration gradient at the interface increases with frequency, allowing more rapid supply of the ions to the interface. Therefore, effective impedance is smaller for larger frequencies.

The diffusion equations were solved by Warburg (the solution is reviewed in [16]). The solution is a frequency dependent parallel R-C impedance model, in series with a charge-transfer resistance R_t (Fig. 2.8).

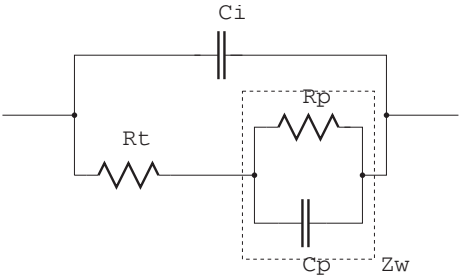


Fig. 2.8. Small-signal model of an electrode

Both R_p and C_p are frequency-dependent, and are given (for unit area) as:[2]:

$$R_p = \frac{2\sigma}{\omega^{1/2}}$$
$$C_p = \frac{1}{2\sigma\omega^{1/2}}$$
$$\sigma = \frac{V_t}{z^2 n^0 \sqrt{2D}}$$

where D is the diffusion coefficient of the ions. Warburg area impedance can be calculated as:

$$Z_w = (R_p + j\omega C_p)^{-1} = \frac{\sigma}{\omega^{1/2}}(1 - j)$$

[2] The presentation in [16] is preferred

Besides the $f^{-1/2}$ frequency dependence this impedance has a constant phase of 45°.

2.2.5 Electrode Noise

As we have seen above, the electrode impedance has an active (real) component along with a reactive (imaginary) component, therefore it must generate electronic noise. It has been shown in [22] and confirmed in [24] that the noise is thermal, and it is generated by the resistive part of the electrode impedance. To obtain an estimate of the electrode noise we shall consider two boundary cases for sample noise calculation: one where the electrode current is limited by charge transduction through the interface (R_t) and another when the current is diffusion limited (Z_W).

We shall make all the calculations for a $1000\,\mu\text{m}^2$ electrode. Using the typical values we have seen above, the interfacial capacitance C_I is about 50 pF and R_t (for charge transfer limited case) is 12.5 MΩ . If an amplifier

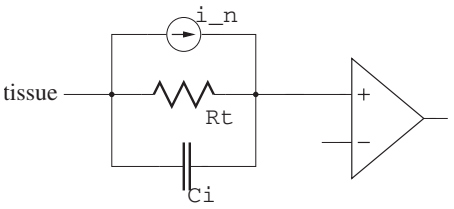


Fig. 2.9. R_t limited electrode

(Fig. 2.9) has an approximately infinite input impedance, then the noise power spectral density (PSD) at the amplifier input will be:

$$4kTG_t\left(\frac{R_t}{1+j\omega R_t C_I}\right)^2 = \frac{4kTR_t}{1+\omega^2 R_t^2 C_I^2}$$

Taking the integral with respect to f:

$$\overline{v_n{}^2} = \frac{2kT}{\pi C_I}\arctan 2\pi f R_t C_I$$

For measurement bandwidth from DC to 10 kHz, since $2\,\pi 10^4 R_t C_I$ is much larger than 1, we have:

$$\overline{v_n{}^2} = \frac{2kT}{\pi C_I} \times \frac{\pi}{2} = \frac{kT}{C_I}$$

which is classic kT/C noise, independent of R_t. The RMS value is about 9 μV .

For the diffusion-limited junction we shall take data presented in [17] (page 18, Table 1–4). It shows measurements of C_p of different electrodes in different solutions. The data was fit to

$$C_p = \frac{k}{f^\alpha}$$

by constants k and α. Among the presented cases, we select Pt/0.025N HCL, since its impedance is the closest match of Warburg model with α of -0.495 and k of $322\,\mu\mathrm{F/cm}^2$. For $1000\,\mu\mathrm{m}^2$ electrode k is 3.22 nF. The corresponding value of σ is about $1/16{\cdot}10^9\ \mathrm{sec}^{1/2}\mathrm{F}^{-1}$. For simplicity of calculation we neglect C_I with regard to C_p, although for frequencies around 10 kHz C_p falls below C_I (32 pF vs. 50 pF). The equivalent circuit is much like Fig. 2.9, but we use R_p and C_p instead of R_t and C_I.

$$PSD(v_n) = \frac{4kTR_p}{1+\omega^2 R_p^2 C_p^2} = \frac{4kT2\sigma\omega^{-1/2}}{1+1}$$

Integrating with respect to f:

$$\overline{{v_n}^2} = \frac{4kT\sigma}{\pi} \cdot \sqrt{f}$$

which is about $6\,\mu$V RMS. This value is somewhat overestimated, since we have neglected C_I.

A practical electrode will probably be somewhere in between the two cases, i.e. its behavior will be affected by both the charge-transfer process and the diffusion process. Note, that in both cases the noise is inverse proportional to the square root of the electrode area. Thus a larger (and less selective) electrode will generate less noise.

2.3 Neuronal Probes for Extracellular Recording

Neuronal probes (or neuronal electrodes) are used to measure the electrical activity of neural networks. Above we have briefly discussed the electrochemistry and electrical properties related to a generic metal electrode interfacing a living tissue. This section describes different types of such electrodes for *extracellular recording*, which means sensing the electrical current induced in the extra-cellular solution by the electrical activity of nearby neurons. Reviews of different types of electrodes can be found in [26], in [27] and in [18].

It is important to note that a signal picked up during an extracellular measurement can not usually be related to a particular unit. Moreover, extracellular electrode typically records activity from more than a single unit. The problem of identifying the active unit upon action potential discovery is usually referred to as "spike-sorting" (Chap. 6).

Techniques exist for *intra-cellular recording*, i.e. penetrating the soma by a special electrode and measuring the cell potential directly. Signals recorded this way are typically much cleaner and the originating neuron is known. However the complexity of fabricating, handling and placing the intracellular electrodes in the tissue is significantly higher, compared to extracellular electrodes.

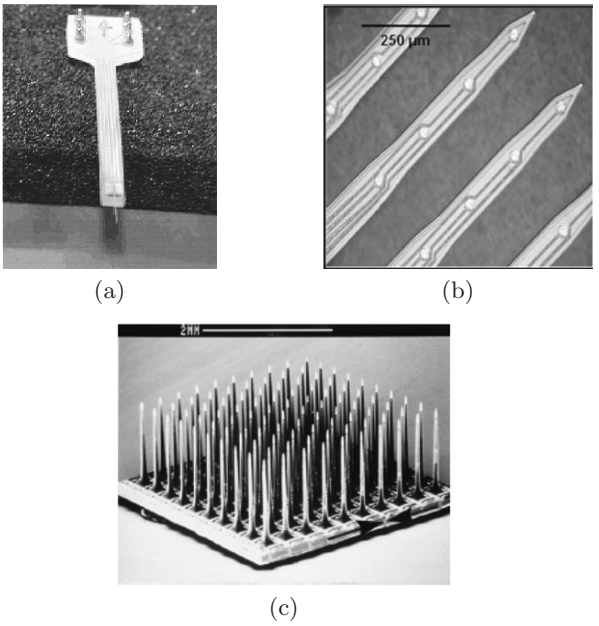
2.3.1 Penetrating Electrodes

Penetrating electrodes are usually thin needles, insulated along the entire length, with only the tip exposed. Traditionally, these are metal wires [21, 17]. The individual wires can be assembled into dense bundles for multi-site recording [28]. Such bundles are available commercially [29, 30].

Microfabrication techniques are used to produce multi-site electrode arrays on a silicon substrate (Fig. 2.10) [31, 32, 33], allowing for several potential advantages:

- Photolitography permits manufacturing precise recording site positions with uniform and repeatable characteristics.
- Thin film processing allows integrating multiple recording sites on a single silicon shaft, eliminating the need for work consuming assembly of discrete structures and reducing the overall device volume.
- Silicon substrate allows integrating electrode with on-chip circuitry, as was demonstrated in [33]. The recording and stimulating electronics was integrated with a multi-site probe of the "Michigan Probe" family.

The "Utah Microelectrode Array" [9] is another example of a micromachined multielectrode probe, consisting of a ten-by-ten array of 1mm silicon needles, glass isolated at the base.



(a) (b)

(c)

Fig. 2.10. Microfabricated probes. (a) [32]. (b) [31]. (c) [9]

2.3.2 Cuff Electrodes and Regenerating Sieve Electrodes

Cuff electrodes (reviewed in [26] and in [18]) are placed inside a tubular cuff warped around a nerve. Such electrodes may be used when inserting a penetrating electrode is inappropriate, for example when the nerve is too deep.

Regenerating sieve electrode [34, 35, 36] is a thin "holed" plate. During the implantation, the target nerve is cut, and the plate is placed inside the cut, in such way that nerve fibers (axons) regenerate through the holes in the array; the nerve "grows through" the plate. Sensing sites aligned near the holes sense only the fibers that pass through adjacent holes. Thus sieve electrodes (unlike cuff electrodes) are inherently selective to the different fibers in a nerve.

2.4 Recording from Cultured Neural Networks

Neuronal networks can be cultured out of the animal body on specialized devices, the *Multi-Electrode Arrays* (MEAs) [37, 38]. Recording from cultured networks has several advantages over in-vivo recording:

- Development of the network can be monitored under controlled and reproducible experimental conditions.
- Dense recording sites allow recording from a large number of neurons in small volumes, an impossible task to achieve by using microprobes and micromanipulators.
- Placement of neurons inside a cultured network can be forced, allowing development of patterned networks [39, 40], allowing studying the effects of network geometry on network behaviour.

Cultured networks are widely used in studies of neural network dynamics [41, 42]. They are also employed as biosensors for drug testing and environmental hazard detection [43, 44].

An MEA (first introduced in [37]), is a dish made of biocompatible material, such as glass, ceramic or silicon, with deposited sensing/stimulating sites, conducting wires and connection pads (Fig. 2.11). The entire device is insulated electrically, except for the electrode tips. The recording sites (usually several tens for an MEA) are typically of 10–20 μm diameter and 100–200 μm spacing. MEAs of various configurations in terms of electrode material, shapes and positioning have been fabricated. A review on MEA configurations and methods of fabrication is available in [46, 19].

2.4.1 MEAs on Silicon Substrate

As it is possible to grow neural networks upon glass substrate, it is possible to do that on silicon substrate as well, integrating recording electronics on the same die with the recording electrodes [11, 47, 48, 49, 50]. The electrical

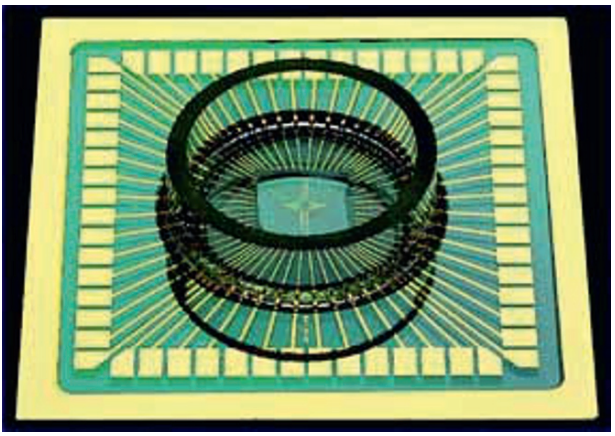


Fig. 2.11. An MEA from Multichannel Medical Systems [45]. Multichannel Systems, Germany. http://www.multichannelsystems.com

properties of neuron-silicon junctions are extensively treated in [51]. In silicon multi-electrode chips (MECs) neurons are capacitively coupled to gates of FET transistors integrated on the substrate. Neural activity is measured as action potentials affect the current flow through transistor channels. It was shown also [52] that individual neurons can be stimulated, (i.e. action potentials excited) by underlying electronic circuitry capacitively coupled to neural somata through a thin oxide layer.

There are two types of recording circuits: The first approach [53] utilizes a neuron placed on top of a thin oxide layer of a MOS transistor as a gate. Electrical activity of the neuron affects the electrical field across the transistor oxide and modulates the current through the channel. Another approach [49, 47, 11] uses a floating-gate MOS, with the gate capacitively coupled to a neuron via thin oxide layer. Action potentials modulate the gate potential which in turn affects the drain-source current.

Both methods require a voltage bias of Vth to exist between the transistor gate and the chip substrate in order for transistor to conduct current. This bias increases the effects of electrochemical corrosion, due to increased currents through oxide cracks. Shappir et al. [48] overcome this drawback by using a depletion MOS, that allows recording with zero bias voltage at the expense of an additional processing step.

2.5 Typical Multi-Electrode Recording Setup

A typical setup for multi-electrode neuronal recording experiment is presented in Fig. 2.12. The setup can be clearly separated into two major parts: the one that is mechanically attached to the subject (neuronal interface or the

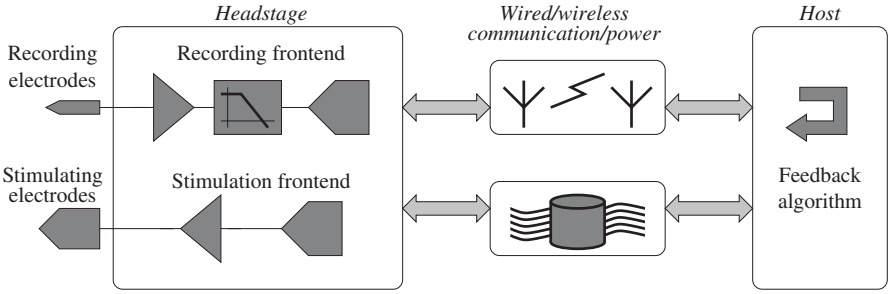


Fig. 2.12. Typical setup for multielectrode neuronal recording

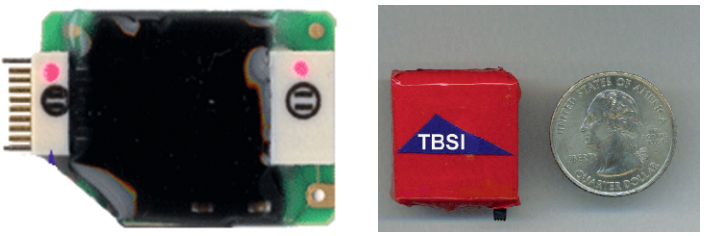
headstage) and the stationary part (the host). Inside the neuronal interface, signals acquired by the recording electrode arrangement are shaped (preamplified, filtered, possibly digitized) by the recording front-end. Either a wired or wireless communication link transfers the signals to the stationary host. In case of communication over wires, some sort of mechanical strain relief solution must be employed if the subject is to be let free. This is typically done by means of a "commutator", a mechanical device connecting two cables that allows both sides to be rotated freely with respect to each other (Fig. 2.13). The host performs the necessary computation and datalogging steps on the incoming input signals and calculates the stimulation feedback. Stimulation instructions are sent back into the interface where they are applied to the stimulation electrodes by the stimulating front-end.

Numerous implementations of such interfaces are available [54, 29, 45, 55]. The headstages are typically assembled of discrete components on miniaturized printed circuit boards (Fig. 2.13). A construction of such a headstage was described in [56].

Various headstage components, especially the recording front-end circuits have been implemented on VLSI chips, providing a higher level of integration. Various aspects of neuronal preamplifiers have been the subject of many studies: [58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68], including noise-power optimization and DC input stabilization (reviewed in greater detail in Chaps. 4 and 5).

Integration of processing electronics with the neuronal probes was also addressed. A micro-assembled device is presented in [69], with a micromachined electrode array mounted on top of the processing chip. In [64] a neuronal probe integrated on the same substrate with recording circuitry is described.

Wireless communication makes for another direction in neuronal interfacing research. A wireless headstage described in [57] is available at [55]. Analog signals from fifteen input channels pass intermediate modulation on different frequencies and then the cumulative signal is transmitted after another modulation of RF carrier. In [70] the digitized signal is transmitted from the recording device by passive telemetry. A commercial 2.4 GHz radio module (so



(a) (b)

(c)

Fig. 2.13. (a) 16-channel tethered headstage [54]. Plexon, Inc., US. http://plexoninc.com (b) Wireless headstage [57]. (c) 16-channel commutator [54]. Plexon, Inc., US. http://plexoninc.com

called *mote*) from [71] was used for wireless communication with a headstage in [72].

Power is yet another important aspect of neuronal interface operation. Successful attempts of remotely powering the front-end device by telemetry have been reported [70, 73]. Another report [69] describes an optically powered device with an integrated photo-voltaic cell.

A certain commonality among the existing devices is very relevant to our discussion: (almost) no computation is ever performed at the interface side. The front-end devices rely on the host "to be there" for any computation needed. In some exceptional cases, front-end circuit may measure some features of the recorded signal to assist the data processing on host. Two such cases (to the best of our knowledge) exist. In [62], the amplitude of a spike is measured. In [74] a threshold detection is applied, with the threshold level automatically calculated based on measurements of the input signal RMS.

2.6 Recorded Signal Information Content

Information exchange inside neural networks is carried out through action potential firing by individual cells, that inhibit or excite the action potentials of other cells. The shape of spikes generated by a neuron does not change over time (except for periods of bursts) [75]. The information is encoded by the positions of the spikes on the time axis, rather than by the features (e.g., height

or width) of the action potential waveforms. The times and the originating cells of the firing events therefore define the "informative content" in neuronal signals. After it is extracted from the recorded signal, higher level algorithms concerned with behavioral aspects of neuronal networks can be applied.

Extracting information out of the recorded signal can be divided into a pair of distinct tasks: detecting the firing events in the signal (so called *spike-detection*) and recognizing their sources (*spike-sorting*). Since firing events are associated with transient peaks in the measured potential, they can be detected by threshold crossing. Resolving the sources of these events is not straightforward, since an extracellular electrode will often sense activity from more than a single neuron. It is usually assumed that action potentials of different neurons will have different shapes on the recorded waveform. Shape-based classification techniques can be utilized for classification of originating units. We shall return to spike detection and sorting techniques in Chap. 6.

3

The Neuroprocessor

Wireless neuronal interfaces are in need in clinical practice, neuronal prosthetics and neurophysiology research. In the former, they will eliminate the transcutaneous wires, improving the quality of life for the patients and reducing contamination risk. In the latter, they will allow recording from freely behaving animals that are not constrained by the connecting wires. Needless to say, such interfaces have to be powered by miniature-size power cells, yet they are to provide sufficient battery life. For human patients, it has to be days if the battery is rechargeable or years if the battery is to be replaced.

Conventional neuronal interfaces such as described in Chap. 2 serve as mere transducers of the signal between the host and the tissue. As such, they transmit all the recorded data and rely on a permanently available host to perform the required computation/data logging. The communication bandwidth required for such operation can be easily calculated given the number of electrodes involved in the interface. There are indications that a good quality prediction of a limb movement may require recording from even thousands of cells [3, 76]. Experiments involving hundreds of cells were reported [3, 28]. The increase in the scale of neuronal interfaces is supported also by introducing microfabrication technologies into the development of neuronal probes, examples are the 100-electrode Utah array [9] and the Michigan probe available with up to 64 channels [77].

Let us consider an interface of a hundred of electrodes, each sampled at 25 Ksps with eight bits of precision, the cumulative datarate is 20 Mbps, far too high for a system powered by a miniature-size battery. This observation was already reported in [74, 78]. It uncovers a fundamental limitation of the non-computational paradigm, showing it inadequate for interfacing large numbers of neurons wirelessly.

We propose the *Neuroprocessor*, a computational neuronal interface. Unlike the conventional neuronal interfaces, the Neuroprocessor will perform significant amounts of computation close to the tissue, communicating only the (low-bandwidth) outcome. Eventually, the entire feedback algorithm that is currently executed on the host can be integrated into the Neuroprocessor,

eliminating the need of permanent host connection completely. This is an important advantage considering neuronal prosthetics, as a prosthetic device can hardly rely on the host to be constantly present.

3.1 Datarate Reduction in Neuronal Interfaces

Datarate explosion can be lowered to some extent, if we recall that information exchange inside neural networks is carried through neuronal firing. The times of neuronal firing events and their origins are the essential features in the recorded neuronal signals. Communicating the continuous signal from the recording electrode is a waste of bandwidth, knowing that neuronal firing events are relatively rare (up to tens of spikes per second) and do not last long (order of a millisecond); most of the time the electrodes record background noise. Preferrably, spike activity would be detected in the recorded signal and only signal portions containing such activity would be communicated. This approach for datarate relaxation was suggested in [74].

If an electrode is sampled at 25 KSps with eight bit precision, a hundred electrodes generate 20 Mbps. Assuming that a neuron fires twenty times a second on average and every electrode senses three to four neurons, the electrode would record close to a hundred spikes per second on average. Assuming also that a spike lasts 1–2 msec, the datarate can be reduced to 2–4 Mbps by detecting spikes in the recorded signal, and communicating only the active signal portions. Although about an order of magnitde bandwidth reduction can be achieved, the resulting datarates are still high.

Let us recall once again that what we seek in neuronal signals are the times of firing events and their sources. The times and sources of the firing events will be extracted from the recorded signal at the host by means of spike sorting (in a multi-unit experiment) or a mere spike detection (single-unit experiment). Communicating the analog waveforms of the signal, even clipped to the times of firing activity, is yet a waste of communication bandwidth. Detection and sorting would be preferrably done on the interface, limiting the communication to the mere indications of spikes and their sources. Using the figures as above, assuming that every spike results in 32-bit spike notification message, the cumulative datarate for a hundred of electrodes is only 320 Kbps, another order of magnitude datarate reduction. Such datarate can be communicated over low-power (tens of milliamps) wireless datalinks. Commercial products for such communication are already available: examples can be found at [71] (Zigbee standard [79]) or at [80] (MICS band).

3.2 Neuroprocessor Overview

The conceptual architecture of the Neuroporcessor is laid out in Fig. 3.1.

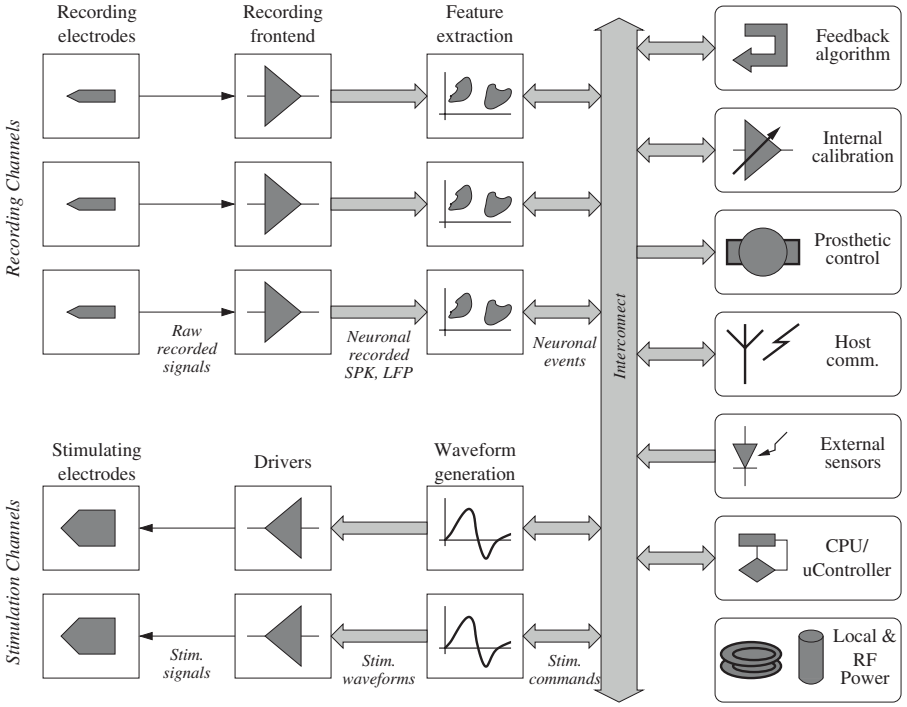


Fig. 3.1. Neuroprocessor conceptual architecture

Recording front-end brings the signals acquired by the recording electrodes into a form suitable for neuronal data extraction. This typically involves DC drifts removal, amplification and filtering (the front-end will be discussed in greater depth in Chaps. 4 and 5). Feature extraction may operate on digital or analog signals. Consequently, the front-end includes digitization.

The registered neuronal events may be used in different ways, depending on the particular applicaiton the Neurprocessor is used for: stimulative feedback calculation, prosthetic device control and/or indication to the host of the neuronal activity.

The stimulation path consists of waveform generation blocks driving stimulation electrodes. As the latter are typically large, their impedance tends to be significantly lower than that of the recording electrodes, potentially requiring special output drivers.

One important remark must be made regarding the Fig. 3.1: The data reduction is performed right after the front-end in every channel and only the event information is communicated on the bus. Doing otherwise (communicating raw signals on the internal bus to a central "feature extraction" unit) would cause the same communication load we have pointed to in previous sections to exist on the internal chip bus, i.e. the communication bottleneck would be "pushed" inside the chip. Intra-chip communications are far less

power consuming, and with the datarates aforementioned are probably manageable. However, there is no good reason to do so: as we are going to see in Chap. 6, the common extraction steps (spike detection and sorting) are typically performed on a channel with no regard to other channels. The extraction unit is better placed on every channel to save the bandwidth on the chip interconnect.

4

Integrated Front-End for Neuronal Recording

4.1 Background

A signal recorded by an extracellular microelectrode consists of several components in several frequency bands [81]. Neuronal firing activity occupies the 100–10.000 Hz frequency band. The amplitude of neuronal spikes picked up by an extracellular electrode is typically small, below 100 μV . Another component of neuronal signal is the Local Field Potential (LFP). The LFP carries cumulative information regarding the activity of large ensembles of cells [75]. It was shown to carry useful information with regard to sensory response [82] and motion [83, 84]. LFP occupies the low-frequency band, below 200 Hz and exhibits much larger amplitudes, of up to 5 mV. Large (hundreds of millivolts) slow drifts of electrode potential are the third and the most "annoying" component of a recorded signal. These drifts are associated with electrochemical reactions at electrode-tissue interface.

The electrode noise (Chap. 2) together with the background noise define the noise floor. A typical setup may provide signals with several microvolt noise floor [24, 85].

4.1.1 Blocking the DC Drifts

Blocking the DC drifts is one of the largest challenges facing integrated neuronal preamplifier design. Due to their large amplitudes, the drifts are to be blocked even before the first preamplifier stage to avoid saturation. The blocking circuit must therefore exhibit very low noise levels. Blocking must also occur at a very low frequency: several Herz, if the LFP is to be left intact, or several hundreds of Herz, if the LFP can be blocked. Such time constants are not readily available within an integrated circuit, making the blocking of DC drifts a challenging task. We would like to stress that due to a large number of experiments conducted with LFP signals, it seems advantageous not to block the LFP, but to make it available at the channel output.

Several approaches have been shown. Using off-chip capacitors in a feedback path of an input amplifier is suggested in [58]. The corner frequency is set so that the LFP is blocked as well. The convenience of using large capacitors comes at the expense of an increased pin-count (an additional pin per channel) and element count (an external capacitor per channel). This latter issue makes this approach impractical for implanted or minituarized head-stages serving hundreds of channels.

In this context, we would like to point out that the signal can be high-pass filtered by subtracting the low-frequency component from the input. Some of the presented works [68, 58, 59] take this approach placing a low-pass filter (LPF) in a feedback path of an amplifier (Fig. 4.1).

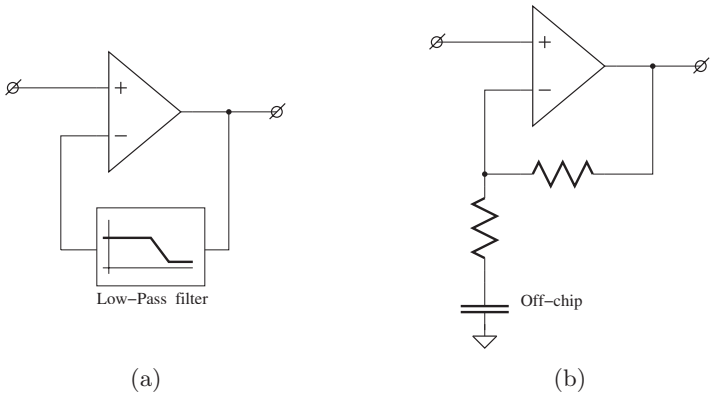


Fig. 4.1. (a) DC blocking with low-pass feedback. (b) Implementation in [58]

Fully integrated approaches were also demonstrated. One of the earliest fully integrated neuronal preamplifiers was published in [59]. A diode-capacitor feedback path was utilized for low-frequency filtering. A diode typically exhibits a very large small-signal impedance at near-zero current levels; this was used to achieve a large time constant. The drawback of the approach (as we see it) is that the input differential pair was placed outside the feedback loop. The DC drifts are blocked at the output of the first amplification stage. Large input offset may therefore drive the first stage far from the equilibrium point.

AC coupling the electrode to the preamplifier input seems therefore a better approach. This was demonstrated in [64, 66]. In both cases, the coupling capacitor was provided by the interface capacitance of the recording electrode. A diode was employed as a shunting element in [64]; A MOS transistor biased in subthreshold region was used in [66]. Relying on the electrode for the coupling capacitor has two disadvantages. First, the properties of the recording electrode must be known apriori, and the preamplifier must be designed with that particular electrode in mind. Second, the impedance of the recording

electrode is usually not purely reactive, it has also a resistive part, usually very large, but not infinite. Thus the DC gain of this scheme is not strictly zero, although it can be made very small [64].

In [68] AC coupling was implemented with an integrated capacitor and a diode-connected MOS transistor as a shunting element. The DC gain of this arrangement is strictly zero. It was also suggested to place the coupling capacitor underneath the bonding pads to save die area. AC coupling with a subthreshold MOS device for shunting was also employed in [57].

A different method was demonstrated in [63, 61, 86] (Fig. 4.2). The weak-inversion devices used in a feedback path provide for a very high small signal

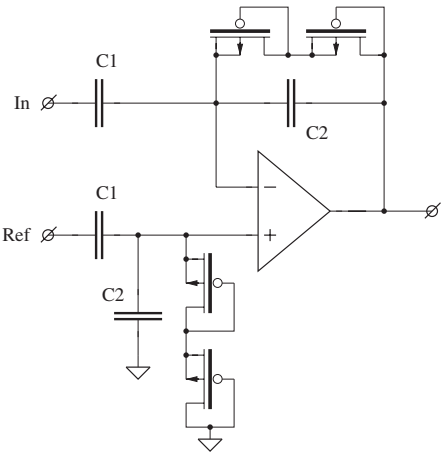


Fig. 4.2. Blocking DC with weak-inversion MOS devices

resistance at near-zero bias. When a higher voltage is applied across the device (in either direction) the current grows exponentially: either because of the opening of the MOS channel or because of the forward bias of the drain-substrate junction.[1] Thus the output voltage of this amplifier is forced within certain limits. In [62, 87] it was suggested to make the gate potential of the MOS devices adjustable. The corner frequency of the high-pass filter can thus be controlled.

[67] suggests a digital feedback for DC blocking: The output signal is to be digitized, processed and fed back to the amplifier negative input through a D/A converter. No implementation is published though, and the approach raises certain questions, regarding the implementation of a D/A converter with sub-millivolt accuracy and output noise at the microvolt level.

[1] Was used in [63, 61]. Olsson et al. [86] use a slightly different connection.

4.2 NPR01 : First Front-End Generation

The first recording front-end, fabricated also as technology and design platform validation step, included eight channels, each one consisting of a two-stage low noise single-ended preamplifier and a low pass filter. DC stabilization was achieved with input reset gates. Channel schematic is presented in Fig. 4.3. DC stabilization was achieved by periodically asserting ϕ_1 and ϕ_2

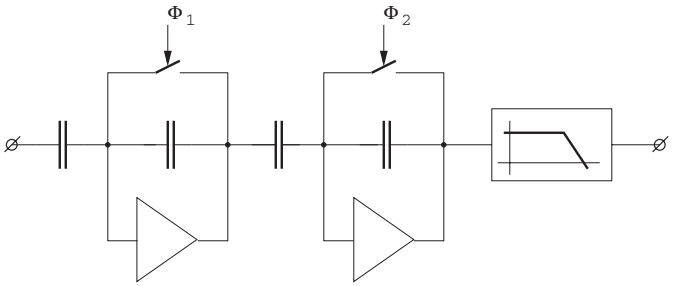
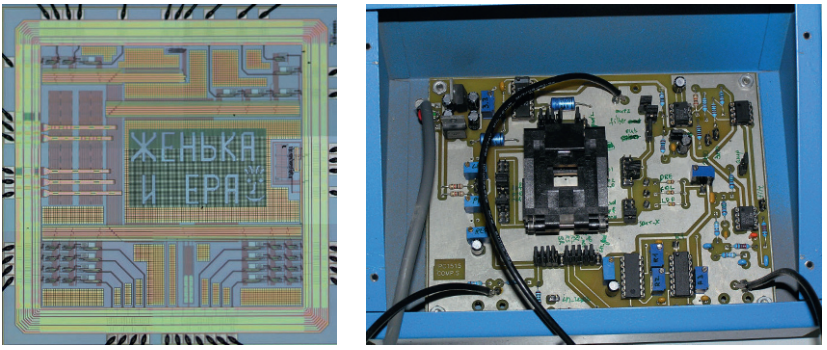


Fig. 4.3. NPR01 channel schematic

for short periods. Moreover, if ϕ_2 is deasserted after ϕ_1 is deasserted, then the output offset of the first stage is rejected at the second stage as well. A simple ring oscillator and the logic necessary for generation $\phi_{1,2}$ were also included on the chip.

The chip was fabricated in 0.35 μm , double-poly, triple-metal mixed-signal process (AustriaMicroSystems), with 3.3V power supply. It was tested electrically and found functional. It was also tested as a preamplifier on an MEA-interfacing board (Fig. 4.4). It was observed during the experiments that reset gates introduce too much switching noise into the input signal. Single ended



(a) (b)

Fig. 4.4. (a) NPR01 micrograph. (b) Test board

architecture of the amplifiers provided poor PSRR, allowing large supply interference. Both the supply interference and the switching noise completely obscured the neural activity. Two conclusions were made: The reset gate approach for stabilization of the input DC level was proved impractical and abandoned; differential input stages were employed in the following generations of sensing chips.

4.3 NPR02 : Analog Front-End With Spike/LFP Separation

The second version of the front-end chip [88], NPR02 , included twelve fully-differential recording channels each with a complete neuronal signal shaping chain. DC blocking was achieved with a first order high-pass filter at channel inputs employing integrated resistors and off-chip capacitors. NPR02 also introduces band-splitting of a neuronal signal into spike data and LFP. The chip was fabricated in 0.35 μm double-poly, quad metal mixed signal process by AustriaMicroSystems. NPR02 operates on a dual-rail supply of $+/-1.65$ V.

4.3.1 Splitting Spike and LFP

Cleared of the near-DC drifts, the neuronal signal has two components left: the spiking activity (occupying frequencies of 0.2–10 kHz) and the local field potential (below 100–200 Hz). Preferably, both are made available at the output. However, the combined signal is hardly usable, since the algorithms that operate on spike data require clearing the LFP and vice versa. Spikes and LFP must therefore be separated and provided on two separate outputs. The separation can be done in the digital domain, by digitizing the combined signal and applying digital filtering afterwards. It can also be done in the analog domain, potentially saving some power.

Since spikes are rare events, if one can detect (or even suspect) their presence in the signal by analog computation, then the digitizer can be activated only on the portions of the signal when a spike is suspected. Threshold detection, for example, is easily done in analog domain. Making digital computations, on the other hand, requires the digitizer and the splitting filters to operate continuously. Making separate analog outputs with spike and LFP information can therefore potentially lead to power saving on the digitizer and subsequent digital filters.

Splitting the combined signal can also relax the dynamic range required at the analog chains. We recall that the LFP amplitude can reach several millivolts, and the amplitude of the spikes is several hundreds of microvolts. The noise floor of spike recording is around several microvolts. Thus the dynamic range of the combined signal is defined by the amplitude of the LFP signal on one hand and the noise floor of the spikes on the other hand at levels of around 1000. The required resolution of the digitizer is 10 bit at least. If we

split the signal, the maximum dynamic range is at the spike part, which is now defined by the noise floor and the spike amplitude and is ten times lower. Thus the dynamic range of the parts of the analog chain following band splitting needs be only 100; and seven bits of resolution at the digitizer. The input preamplifier must provide a full dynamic range in both cases.

4.3.2 NPR02 Architecture

The architecture of a single NPR02 channel is shown in Fig. 4.5.

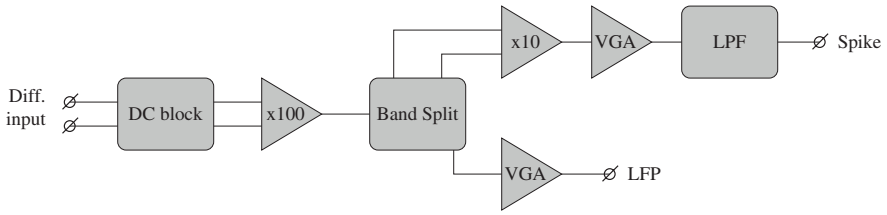


Fig. 4.5. NPR02 channel architecture

The signal is cleared of the DC component, amplified a hundred times and split into the spike and LFP parts. The spike part is then amplified by ten and amplified again by a variable gain amplifier. Spike band is limited by a second order Bessel filter with variable cutoff frequency. The LFP part is amplified by a variable-gain amplifier (VGA). Both spike and LFP outputs are buffered to chip pads.

Figure 4.6 shows the block diagram of an NPR02 channel. The input high-pass filter makes use of external capacitors. 8 MΩ resistors (high resistive poly) were placed on chip. To make a cutoff at about 1 Hz, 22 nF external capacitors can be used, available in miniature SMD packages. The band splitter was realized as a first order RC filter, with 5 MΩ resistor and 160 pF (gate oxide) capacitor.

Assuming the output LPF has a steep rolloff above some frequency f_b, the noise introduced by the splitter into the spike band is:

$$\sqrt{4kT \cdot R \cdot f_b}$$

which is about 30 μV for f_b of 10 kHz. The noise floor of the channel is aimed at a level of 2–3 μV . To suppress the splitter noise reliably, the preamp must provide gain of well above 20; the preamp gain was set to 100. Both VGAs provide digitally selectable gains of 2.5/5/7.5/10. The maximal total gain of the spike chain is therefore 10,000, and that of the LFP chain is 1,000.

The output LPF is a Sallen-Key biquad [89], realizing a second-order Bessel low-pass filter (Fig. 4.7). The cutoff frequency was made programmable through shorting resistor segments. The LPF provides buffered output that can drive chip pads.

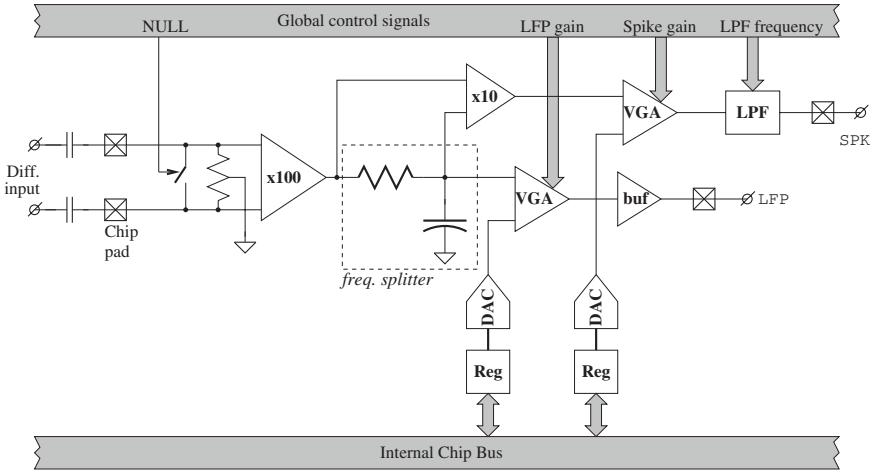


Fig. 4.6. NPR02 block diagram

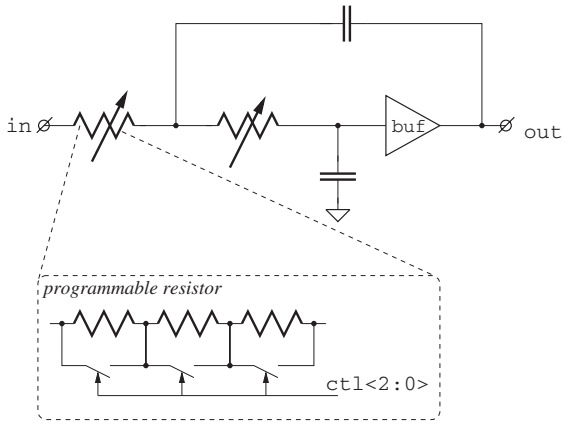


Fig. 4.7. Spike output LPF

DC offsets of both spike and LFP channels have to be compensated: The LFP channel amplifies the input preamp offset (hundreds of μV , typically) by up to 60 dB; unless compensated, it would limit the dynamic range severely or even saturate the VGA. The spike chain output offset is determined by the offset of the $\times$10 stage amplified by 40 dB, as the DC part of the preamp output signal is blocked by the band splitter. Smaller than LFP, spike offset is yet significant: the $\times$10 stage has larger input offset compared to the preamp, since the latter uses very large input devices due to the noise requirements.

Offset compensation is carried out by two calibration digital-to-analog converters (DACs), one for LFP and one for spike, applied to the last amplification stages (VGAs). The DACs are implemented as 5-stage R2R resistor

ladders, having 400 mV output swing. DAC values are stored in registers that can be individually accessed through a common bus with five address/data bits and three control bits.

4.3.3 Input Preamplifier

The circuit of the input preamplifier is shown in Fig. 4.8. A degenerated differential cascoded transconductor stage followed by a current amplifier loaded

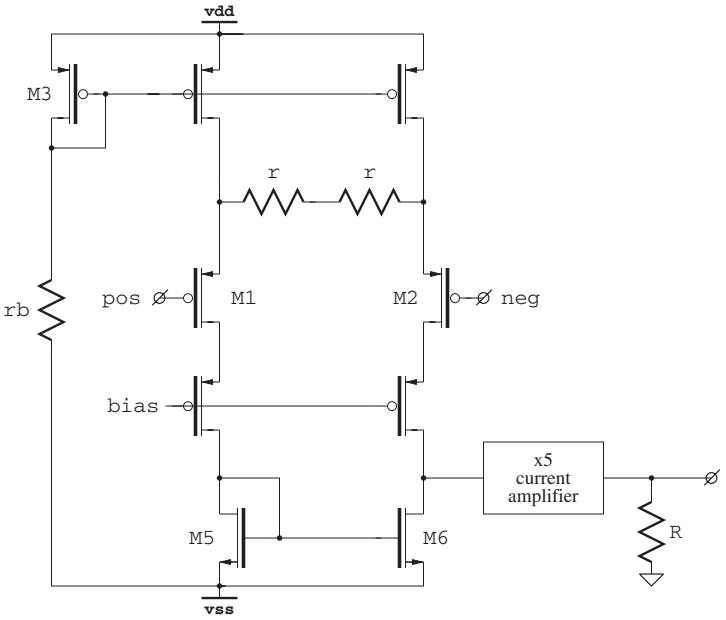


Fig. 4.8. NPR02 input preamplifier

with a resistor to convert the output current to voltage. The gain of the input stage is given by:

$$A = \frac{2R}{r + r_m} = \frac{2R/r}{1 + r_m/r}$$

where r_m is the transresistance of $M_{1,2}$. While r can be matched to R by using the same resistor types and employing appropriate layout techniques, there is no straightforward way of matching r to r_m. The sensitivity to r_m can be reduced by reducing the ratio r_m/r, but there is a limit on how high r can be due to the noise requirements (some 10 kΩ) and reducing r_m means more power.

Instead, we match r to r_m by appropriately controlling the bias currents through $M_{1,2}$. $M_{1,2}$ are operated in the subthreshold region (the smallest r_m for a given I_d) so that r_m is inversely proportional to I_d:

$$r_m = \frac{\eta V_{th}}{I_d}, \quad V_{th} = \frac{kT}{q}$$

I_d is given by:

$$I_d = \frac{V_{dd} - V_{gs3}}{r_b}$$

thus we can write the gain as:

$$A = \frac{2R}{r + r_b \frac{\eta V_{th}}{V_{dd} - V_{gs3}}}$$

We match r_b to r and keep V_{gs3} much lower than V_{dd}. Since the ratio of V_{th} and V_{dd} is small, the above expression becomes weakly dependent on process parameter η and on V_{gs3}. The chip is expected to work in constant temperatures (subject body), thus the dependence on Vth is not worrying. One sigma chip-to-chip channel gain variation of less than 2% was actually measured.

The procedure for sizing the preamp and choosing transistor current is described in Appendix A. Without going into too much detail here, the evident drawback of this circuit is the degeneration of the input differential pair. On one hand it is desirable to control the gain; on the other hand it reduces the effective input transconductance, degrading the power-noise performance (more noise for a given power). From this perspective, feedback circuits such as [63] and alike perform better: the gain is set by the feedback and no degeneration is needed.

However, there is also an advantage, a low input capacitance. Feedback circuits use a large input capacitor (10 pF typically). Input capacitance of the preamp is lower: gate capacitance of the input transistor is some 5 pF and it is reduced by degeneration. The effective input capacitance for the circuit shown is about 700 fF.

Because of the low input capacitance, the preamplifier designed for NPR02 was also used in a chip for in-vitro recording (Chap. 7). The recording sites in such a chip typically provide much lower interface capacitance (i.e. coupling capacitance between the tissue and the preamplifier), thus low capacitance at preamplifier input is essential.

4.3.4 NPR02 Measurements

Measurement setup and test board block diagram are presented in Fig. 4.9. The layout and the assembled test-board are shown in Fig. 4.10. The test-board includes a matrix of digitally controlled switches that connect inputs of NPR02 with a voltage divider driven by a waveform generator. The outputs can be connected through another switch matrix to a scope, via board output channels. Every channel is configurable, it can provide various gain levels and/or filtering (see Appendix A for details).

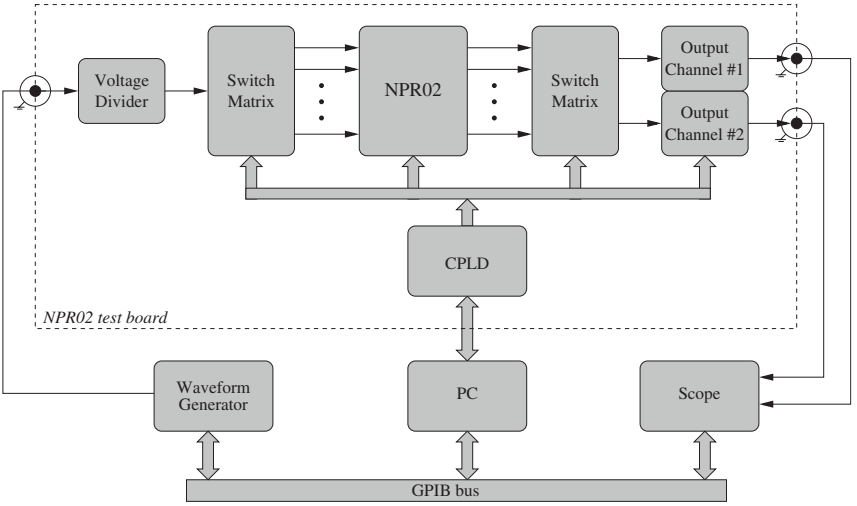
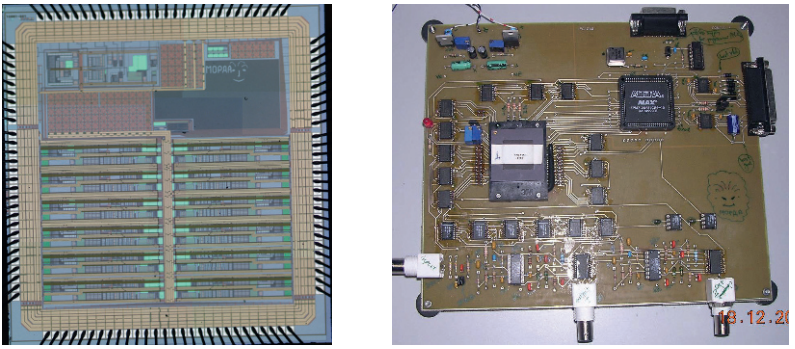


Fig. 4.9. NPR02 measurement setup

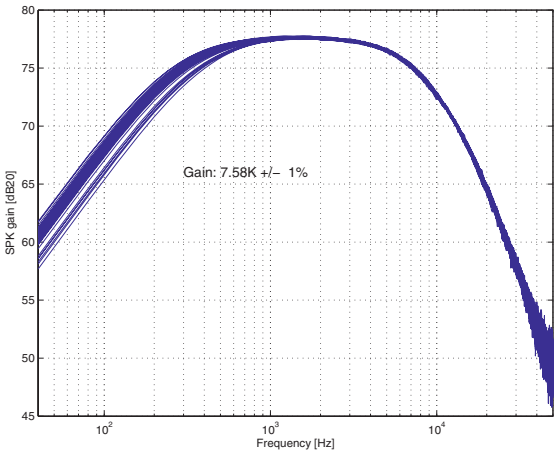
The switches and NPR02 are controlled by an on-board programmable logic device, that communicates with the computer host via a parallel port. The host also controls the waveform generator and the scope by means of a GPIB bus. This allows for developing fully automatic measurement procedures with MATLAB "Instrument Control" toolbox for NPR02 evaluation.



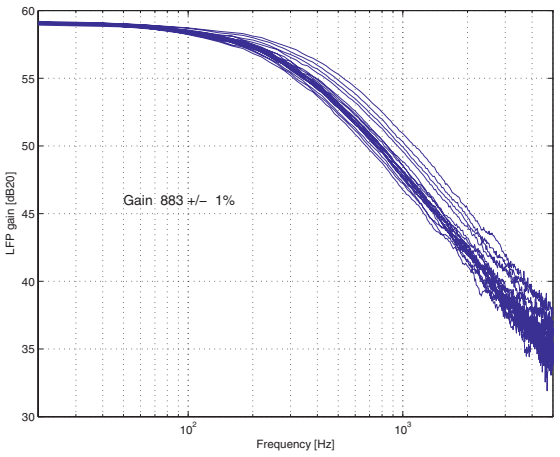
(a) (b)

Fig. 4.10. (a) NPR02 micrograph. (b) NPR02 test board

The measured step response for several channels from ten tested chips is shown in Fig. 4.11. The gains of the preamplifier, $\times 10$, and VGA stages measured to 0.92 of the intended values. Thus, gain error of 0.91^3 appears at the spike chain and gain error of 0.92^2 appears at the LFP chain. The variations are much smaller then expected; standard deviation of the channel gain is 1%.



(a)



(b)

Fig. 4.11. NPR02 frequency response: (a) spike. (b) LFP

The conclusion is that the statistical models are rather pessimistic. If the circuit is to be redesigned, we can allow much smaller degeneration and reduce power consumption. If we repeat the calculations for preamp sizing described in Appendix A using three times smaller r/r_m, the current consumption goes down to 40% of the present value.

A certain issue can be observed in Fig. 4.11. The splitter pole is displaced and exhibits rather large variation among the different curves. We relate this to a failure in a bias circuit in the splitter. The splitter uses a rather large gate capacitor to form the pole. The capacitor bulk has to be biased properly

to bring the MOS capacitance to its largest. This was done with a special bias block, which has failed.

Figure 4.12 shows the input-referred noise measured with `NPR02` , along with the simulated curve. The measurements agree with expectations. Total input referred noise (spike channel) is $3\,\mu$V .

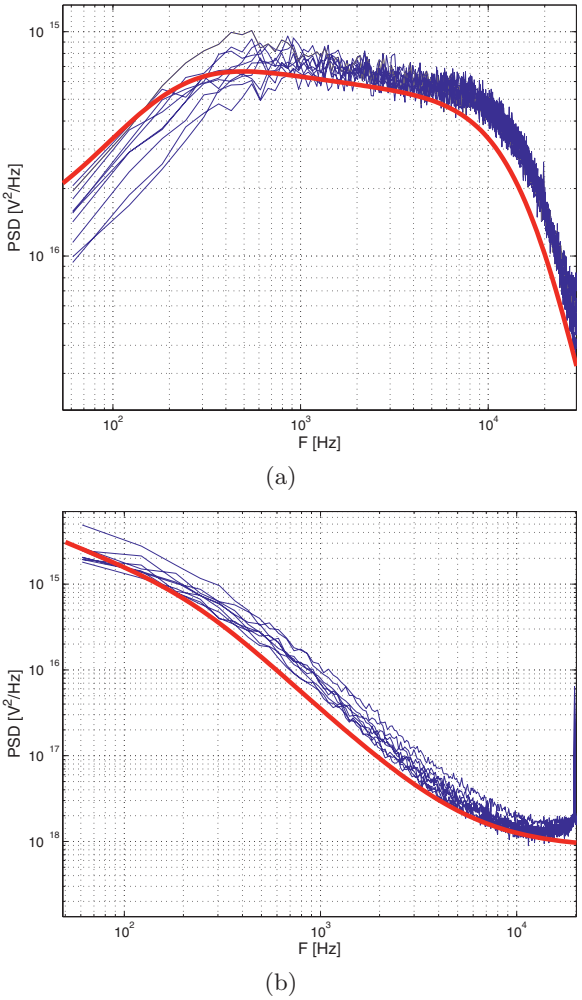


Fig. 4.12. Input-referred noise of the spike chain (a) and the LFP chain (b) `NPR02` . Thick lines denote the simulated curves

5

NPR03: Mixed-Signal Integrated Front-End for Neuronal Recording

5.1 Overview

NPR03 [90] is a fully integrated mixed-signal twelve-channel front-end. DC regulation is obtained with an input high-pass filter built of weak-inversion MOS devices and integrated capacitors. The corner frequency of this filter is digitally programmable with gate bias voltage setting by special DACs. NPR03 has differential inputs, spike/LFP band separation and digital offset calibration. Spike/LFP gains and output LPF cutoff frequencies are digitally programmable for each channel.

A 10-bit analog-to-digital converter (ADC) is integrated in every channel. A special "inverted dual-ladder" resistor DAC [91] was designed to be used in the ADCs. The channels communicate with the central controller over an internal synchronous bus. The controller takes care of channel readout, internal bus mastering and host communications over a five-wire bit serial synchronous line (McBSP [92]). All the channel parameters (offsets, gains, corner frequencies) and controller registers are accessible by the host through the McBSP interface. The same interface is also used for streaming the recorded data to the host. NPR03 can also be instructed to apply threshold detection on the recorded channels and stream only active portions of the recorded signals (signal clipping).

A special system incorporating an evaluation board and an embedded computer communication board was designed to interface the NPR03 with the host computer via Ethernet line. Host-side software with a Graphical User Interface was developed for NPR03 control and data display.

The improvements introduced in NPR03 compared to NPR02 are summarized below.

- A fully integrated preamplifier with digitally-tunable DC blocking filter was designed for the recording channel.

- All the parameters of the recording channels are stored in dedicated registers inside the channel. VGA gains and filter frequencies can be set for every channel individually (in contrast with NPR02).
- Ten bit channel-level ADCs allow for digitization of recorded signals at 40 KSps.
- The channels are accessed through an internal synchronous bus, both for register setting and for readout of the recorded signals.
- Chip operation is timed by an integrated controller. The controller is responsible for the internal bus management, register access, data readout, multiplexing and host communications.
- Host communication is carried out over a five-wire, synchronous, serial bus.
- The controller is capable of "signal clipping": it can apply threshold detection to the incoming signal and communicate a certain number of samples from a channel after a threshold crossing event has been detected on this channel.

Figure 5.1 shows the layout of NPR03 .

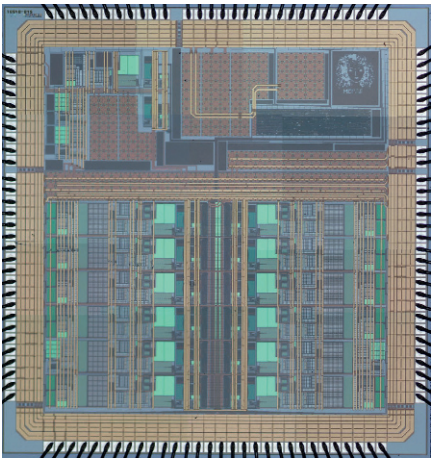


Fig. 5.1. NPR03 layout

5.2 NPR03 Architecture

The block diagram of NPR03 is shown in Fig. 5.2.

The controller has two modes of operation, *programming* and *streaming*. In the programming mode, contents of internal registers can be stored and fetched by the host. In the streaming mode, the controller continuously polls the channel ADCs, checks for threshold crossing events on every channel, and

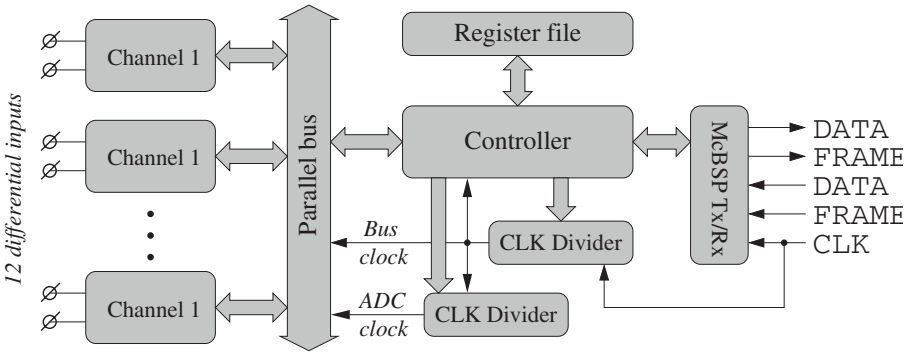


Fig. 5.2. NPR03 block diagram

transmits the active signal segments to the host. All twelve channels or an arbitrary subset thereof can be enabled for data streaming.

A threshold crossing event is triggered for a certain channel when the output of this channel falls below the low threshold or rises above the high threshold. A certain number of samples from that channel will be communicated to the host following the threshold crossing event. The threshold values and the number of samples to transmit after the threshold event are programmable. The entire data stream, without clipping, can be obtained from the chip by setting both thresholds identical.

5.2.1 Chip Communications

The chip communicates over a McBSP bus [93]. This is a five-wire, full-duplex, bit-serial synchronous bus; a synchronization clock signal is constantly supplied by the host. Communication is carried out in *frames*; the host sends 24-bit frames (we refer to this direction as *downwards*) and the chip replies with 16-bit frames (the *upwards* direction). The lengths of downward and upward frames were conveniently chosen to match the lengths of a single host instruction packet and a single reply packet respectively. NPR03 downward and upward frame formats are described in Appendix B.

The maximal data rate that is generated by the chip can be calculated as follows: A channel ADC sample is ten bit wide (although seven bits are sufficient, we have implemented ten bit ADCs for verification purposes). Together with a four bit channel number and a two bit control field, an ADC sample can be communicated in a single 16-bit frame. Sampled at 40 Ksps, a single channel would generate a 640 Kbps. Although there are only twelve channels in the current version of the chip, the bus interface was designed to support sixteen channels for future versions; the aggregate datarate is therefore 10.24 Mbps. The bus was set to operate on a slightly higher, 12.5 MHz clock signal.

5.2.2 Instruction Set and Register Access

The chip operation is controlled through instructions sent via the McBSP bus. Four instructions are available:

- *STORE reg val*: Store value in a register.
- *FETCH reg*: Fetch register contents.
- *RUN*: Start streaming data.
- *STOP*: Stop streaming data.

There are two kinds of parameters that control the chip, those affecting controller operation and those affecting the channels. The former include clock divider settings, threshold values, number of samples to communicate upon threshold detection and channel enabling bit mask; the registers for their storage reside in the controller and are accessed directly. The latter include offset calibration data, channel gains and filter frequencies; the registers are distributed over the channels and are accessed through the internal bus. NPR03 registers and their functionality are described in Sect. B.2.

The internal bus has eight data lines, two control lines and a clock. A register connected to the bus is identified by a distinct eight-bit address. Every bus access is carried out in two steps; during the address step (ADDR is high) the address is driven on DTA<7:0>. The register matching this address is *selected*. During the data step (ADDR is low) the contents of the selected register are driven on the bus by the channel (WR is low) or the register is updated with the value on the bus (WR is high). The bus can be accessed in three possible scenarios, SELECT-READ, SELECT-WRITE and SELECT-READ-MODIFY-WRITE. Figure 5.3 shows the bus signals during the last access scenario.

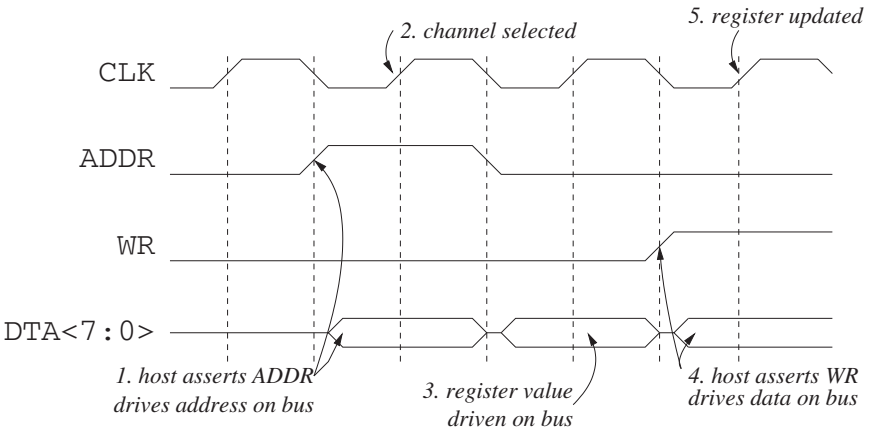


Fig. 5.3. SELECT-READ-MODIFY-WRITE bus access

A ten-bit ADC sample is read with two register accesses, since the registers are fixed eight-bit width. Another write access is needed to request the next

conversion at the channel ADC. Reading a single ADC sample out of a channel takes therefore five clock cycles. 16 channels sampled at 40 Ksps need bus clock rate of at least 3.2 MHz. Since the bus clock is derived form the McBSP clock by integer division, the most suitable division factor is 3, setting the bus clock to about 4.16 MHz.

5.3 Host Interface

A special interface provides for communication between a personal computer and the neuronal recording front-end. The basis of the interface is an Altera Nios II development kit board incorporating an Altera Cyclone II FPGA device, RAM and flash memory, and an integrated Ethernet physical interface/MAC (Fig. 5.4).

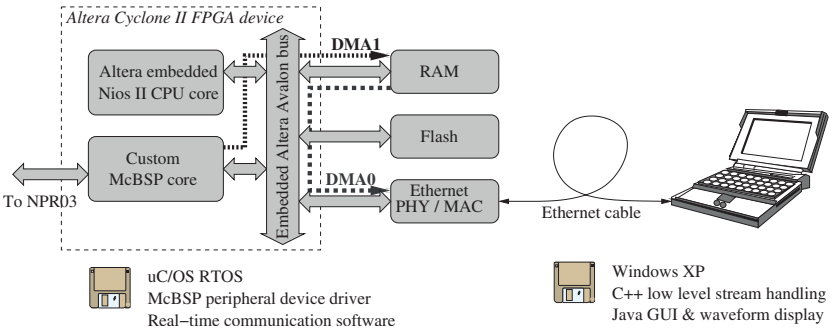
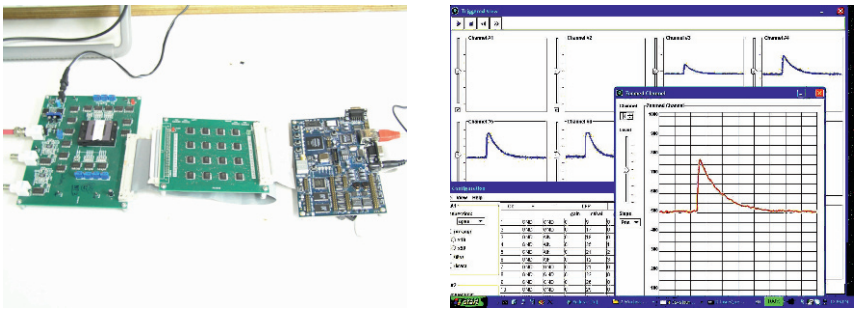


Fig. 5.4. `NPR03` host interface board

The FPGA incorporates an Altera Nios II embedded processor core, bus logic and custom-developed peripheral for McBSP communications with the neuronal recording front-end. The embedded processor executes the μC/OS real-time operating system (RTOS) and custom-developed real-time software for handling the neuronal data stream. The software reads the serial McBSP data, packetizes it and transmits the packets over Ethernet to a host computer using UDP/IP protocol. It also handles the incoming instructions from the host and communicates them to the chip.

The host side software consists of a low-level C++ module that handles the data stream in real time, dumps it onto the disk and performs the decimation necessary for an on-screen display. Displaying data on screen without some sort of decimation (i.e. downsampling) would result in too high screen refresh rates, imperceivable by the human eye. Data display and system control are performed by the top-level Java GUI module (Fig. 5.5).



(a) (b)

Fig. 5.5. (a) NPR03 evaluation board and host interface. (b) screenshot of the host-side software

5.4 NPR03 Channel

The block diagram of the NPR03 recording channel is shown in Fig. 5.6. Most of the analog circuitry was left unchanged from NPR02 . The input preamplifier, designed anew, is an exception and is discussed below.

5.5 Analog-to-Digital Converter

A 10-bit successive approximation ADC was added at every channel. It was intended to operate at 40 kSps sampling rate and provide input range close to the supply rails. The ADC employs a dual resistive ladder 10-bit DAC, comparator, digital controller and sample-and-hold.

Miller sample-and-hold was utilized (Fig. 5.7). The convenience of this scheme is that the opamp inputs are always kept close to ground potential, thus it does not require a rail-to-rail opamp input even when the whole circuit does operate rail-to-rail. A simple opamp with cascoded output stage was used in the SAH.

A rail-to-rail comparator was designed for the ADC (Fig. 5.8). Rail-to-rail operation was achieved with two complimentary input differential pairs; the pairs are followed by a summing stage and a latch.

A novel, dual resistive ladder DAC [91] was developed for in NPR03.[1] In a single resistive ladder DAC, the current flow is inversely proportional to the ladder resistance, while the output impedance (and thus, the settling time) is directly proportional to the ladder resistance. Therefore, current consumption multiplied by the settling time is a constant. Ten-bit DAC is hard to implement with a single ladder, due to the large number of taps to break out. Above eight

[1] For such low frequencies, a capacitive DAC would probably be much more power efficient. However, we could not have reliably designed one due to a severe bug in the design kit concerned with capacitance extraction.

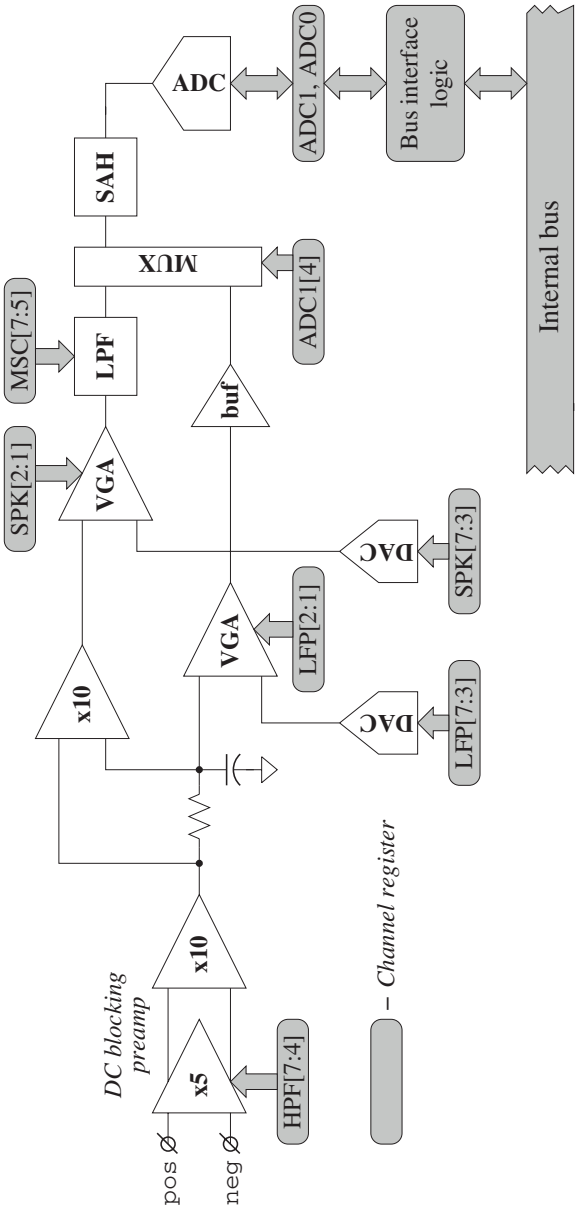


Fig. 5.6. NPR03 recording channel

bit, two serially connected ladders are typically used. To avoid loading the first ladder with the second ladder, the resistance of the latter is typically increased, raising the power-delay constant. In [91] we describe a dual-resistive ladder that provides no penalty associated with the usage of the second ladder; it

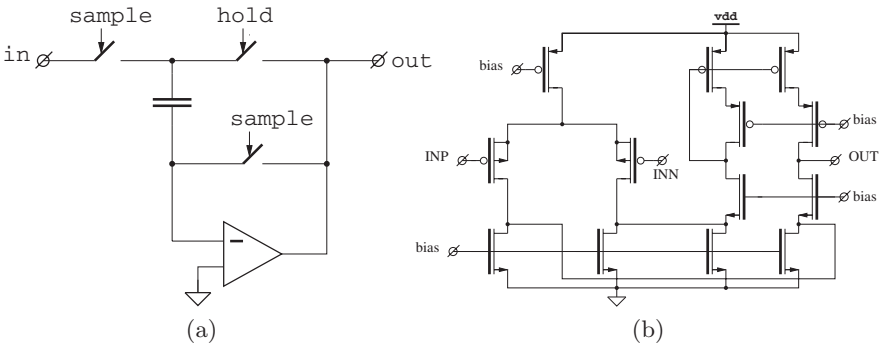


Fig. 5.7. (a) NPR03 SAH circuit. (b) Opamp schemtic

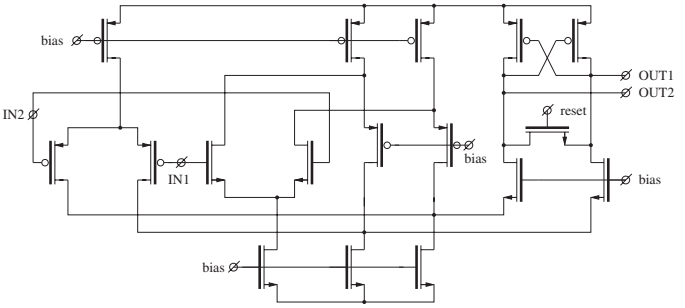


Fig. 5.8. NPR03 comparator

has a power-delay constant of a single ladder. It also reduces significantly the parasitic capacitance compared to existing dual-ladder schemes.

5.6 Integrated Preamplifier With DC Blocking

The preamplifier is shown in Fig. 5.9. A weak inversion MOS in parallel with C_f make for a first-order high pass filter for input DC stabilization. As the conductance provided by the feedback transistor does not belong to a set of controlled process parameters, the cutoff frequency can not be reliably set to a certain value by design. Hence, we have made it digitally programmable through gate bias voltage adjustment with a 5-bit DAC. We have indeed measured more than an order of magnitude variation in the cutoff frequency without calibration (Fig. 5.10), but with proper DAC setting we have managed to bring all the channels close to a target 1Hz cutoff.

5.6.1 Choosing C_i and C_f

Process documentation provided characterization of MOS transistors for currents of down to 1pA (for W/L of unity) (Fig. 5.10), corresponding to g_m

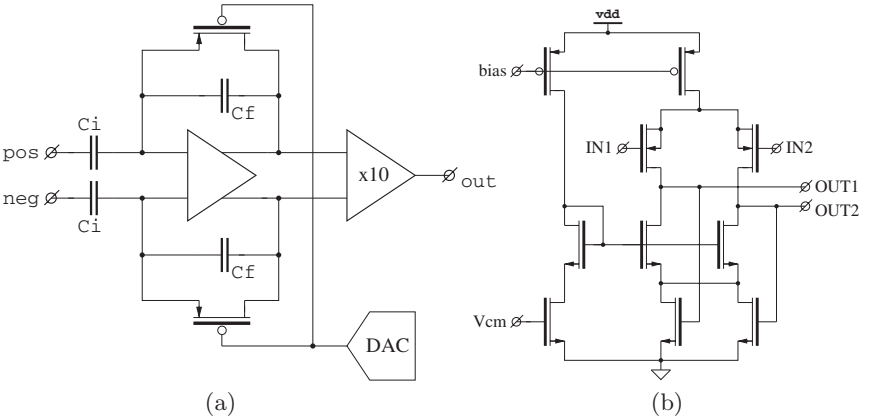


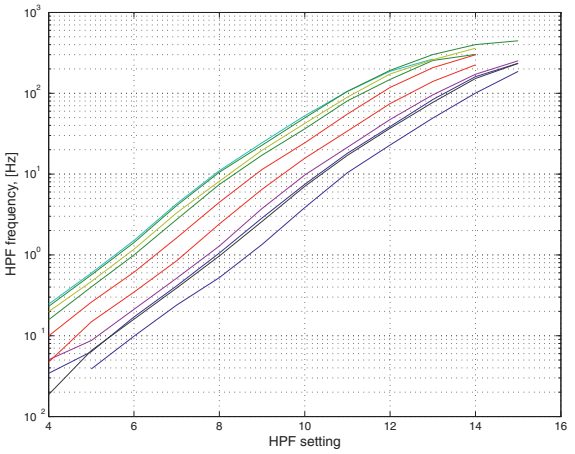
Fig. 5.9. (a) NPRO3 preamplifier. (b) Preamplifier opamp

values of about $40\,\mathrm{p}\Omega^{-1}$. As we did not wish to depart significantly of the characterized region in a first test-chip, we have chosen C_f of 1 pF. C_i defines the input capacitance of the preamplifier together with the parasitic capacitance of the bonding pads and ESD structures. When the impedance of the recording probe is known aprori, an optimal input capacitance can de determined to relax the tradeoff between noise contribution of the preamplifier and signal deterioration by the input capacitance. When the impedance of the electrode is not known, input capacitance of up to 10 pF is a common practice. We have chosen C_i of 5 pF, leaving up to another 5 pF for the parasitics and ESDs. This gives a modest gain value of five at the first stage; which can probably be increased with smaller C_f.

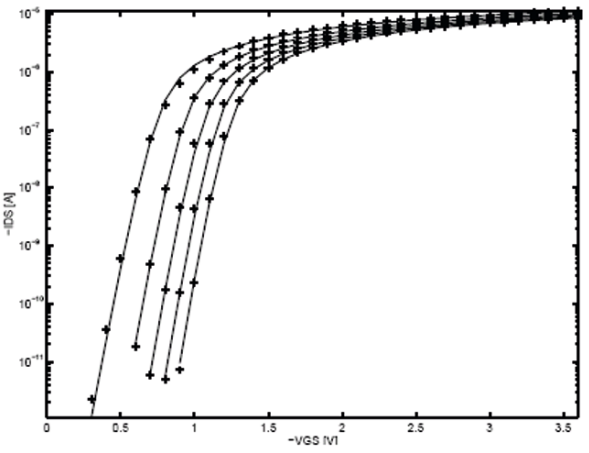
5.6.2 Noise Analysis

Several noise sources can be identified within the preamp circuit (Fig. 5.11): Shot noise due to leakage currents through pad ESD protection diodes (i_{esd}), shot noise due to leakage currents through diffusions of the feedback MOS (i_d), thermal noise of the feedback MOS (i_r) and the noise of the amplifier (i_a), represented by an equivalent current noise source at the output. To make further analysis easier we shall make several neglections and simplifying transformations on the circuit, bringing it to the form shown in Fig. 5.12:

- The circuit has an input current noise due to the ESD structures, thus its noise performance depends on the input impedance. The size of the ESD structures used is about $3500\,\mu\mathrm{m}^2$ and the typical value of diode leakage current density is $0.1\,\mathrm{fA}/\mu\mathrm{m}^2$, giving ESD leakage of 350 fA. Let us assume that the circuit is connected to a recording probe with characteristic impedance of Z_0 at 1 kHz. RMS noise voltage at the input in 10 kHz band is:



(a)



(b)

Fig. 5.10. (a) DC blocking HPF cutoff frequency vs. DAC input value, for several different dies. (b) Sub-threshold PMOS characterization curves

$$v_{n,esd} = \sqrt{2qI_{esd} \cdot Z_0^2 \cdot 10^4} = 3 \cdot 10^{-14} \cdot Z_0$$

For Z_0 of 10 MΩ (which is very large), $v_{n,esd}$ is 300 nV, which is negligible compared to contributions of other noise sources. The actual noise voltage due to ESD currents in the neuronal spike band (0.2–10 kHz) is even smaller, if we take into account that electrode impedance usually exhibits Warburg characteristics [17], i.e. decreases as $1/\sqrt{f}$.

- Since the electrode capacitance is typically much larger than C_i, we shall assume that the input is grounded.

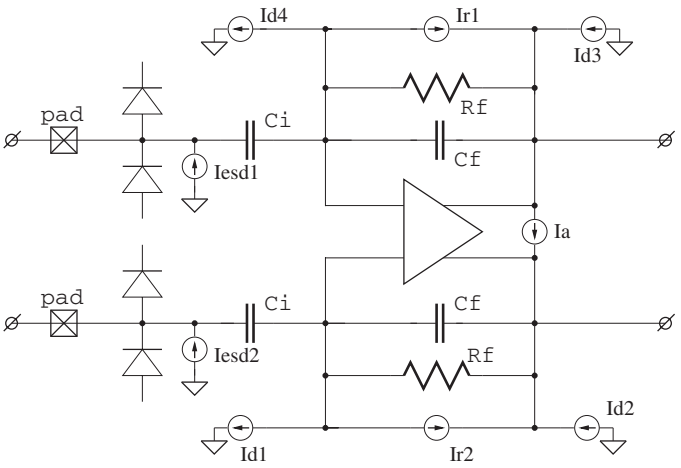


Fig. 5.11. Noise sources within the preamp

- The feedback transistors have diffusions of about $7\,\mu\mathrm{m}^2$, placing the leakage currents at about 0.7 fA. The associated shot noise density is about $2 \cdot 10^{-34}$ A^2/Hz. Assuming the cutoff frequency of 1 Hz, R_f is $(2\pi C_f)^{-1} = 160\,\mathrm{G}\Omega$, with the current noise density of $10^{-31}\,\mathrm{A}^2/\mathrm{Hz}$. Therefore, we can neglect the diffusion leakage noise sources, $i_{d,1-4}$.
- At last, we transform the circuit to a single-ended form, replacing the noise sources $i_{r,1}$ and $i_{r,2}$ with the equivalent i_r of twice the power spectral density.

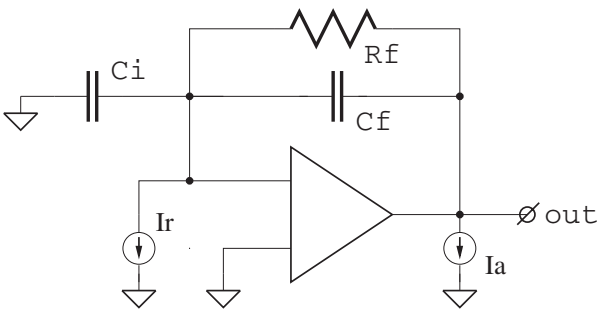


Fig. 5.12. Preamp noise sources, simplified

The transfer function of the current i_r is:

$$\frac{v_o}{i_r} = G_f^* \frac{g_m - sC_i^*}{g_m - G_f^*}$$

where C_i^* is C_i in parallel with the amplifier input capacitance, and G_f^* is G_f in parallel with C_f. If we bias the amplifier in the μA range, then g_m is hundreds of $\mu\Omega^{-1}$, much larger than sC_i^* and G_f^* at frequencies of interest. Thus,

$$\frac{v_o}{i_r} \cong G_f^*$$

We calculate the output voltage noise in the spike band due to i_r, assuming that a first-order high-pass filter (i.e. frequency splitter) limits the band below some f_1 and a low-pass filter with steep rolloff (output LPF in spike channel) limits the band above f_2.

$$\begin{aligned}
\overline{v_{n,r}{}^2} &= \int_0^{f_2} \frac{4kT2G_f}{\omega^2 C_f^2 + G_f^2} \cdot \frac{\omega^2}{\omega^2 + \omega_1^2}\, df \\
&= \frac{2kTG_f}{\pi^2 C_f^2} \frac{1}{f_0^2 - f_1^2} \int_0^{f_2} \frac{f_0^2}{f^2 + f_0^2} - \frac{f_1^2}{f^2 + f_1^2}\, df \\
&= \frac{2kTG_f}{\pi^2 C_f^2} \frac{1}{f_0^2 - f_1^2} \left[f_0 \arctan \frac{f}{f_0} - f_1 \arctan \frac{f}{f_1} \right]_0^{f_2} \\
&= \frac{2kTG_f}{\pi^2 C_f^2} \frac{1}{f_0^2 - f_1^2} \left[f_0 \arctan \frac{f_2}{f_0} - f_1 \arctan \frac{f_2}{f_1} \right]
\end{aligned}$$

f_0 is the cutoff frequency of the DC-blocking HPF, and is about 1 Hz. f_1 is the splitter pole location, and is about 200 Hz. f_2 is the spike band limit and is about 10 kHz. Thus we can approximate with $f_0 \ll f_1 \ll f_2$:

$$\begin{aligned}
\overline{v_{n,r}{}^2} &\cong \frac{2kTG_f}{\pi^2 C_f^2} \frac{1}{f_1^2} \left(\frac{\pi}{2} f_1 - \frac{\pi}{2} f_0 \right) \\
&\cong \frac{2kTG_f}{2\pi C_f^2} \frac{1}{f_1} = \frac{2kT}{2\pi C_f} \cdot \omega_0 \cdot \frac{1}{f_1} \\
&= \frac{2kT}{C_f} \cdot \frac{f_0}{f_1}
\end{aligned}$$

Reflecting $v_{n,r}$ to the input gives:

$$\overline{v_{n,r,in}{}^2} = \frac{2kT}{C_i} \cdot \frac{C_f}{C_i} \cdot \frac{f_0}{f_1} = \frac{2kT}{C_i} \cdot \frac{1}{A} \cdot \frac{f_0}{f_1}$$

For the values chosen, $v_{n,r,in}$ is 1.3 μV . It can be further decreased by choosing a smaller C_f. With adjustable cutoff frequency f_0, one gains control over the tradeoff between the amount of noise injected by R_f and low frequency input suppression.

There exists another interesting aspect of R_f implementation. In our design, the gate length of the feedback MOS is 40 μm , not so convenient to layout as a single gate. Thus we break the transistor into four serial gates of

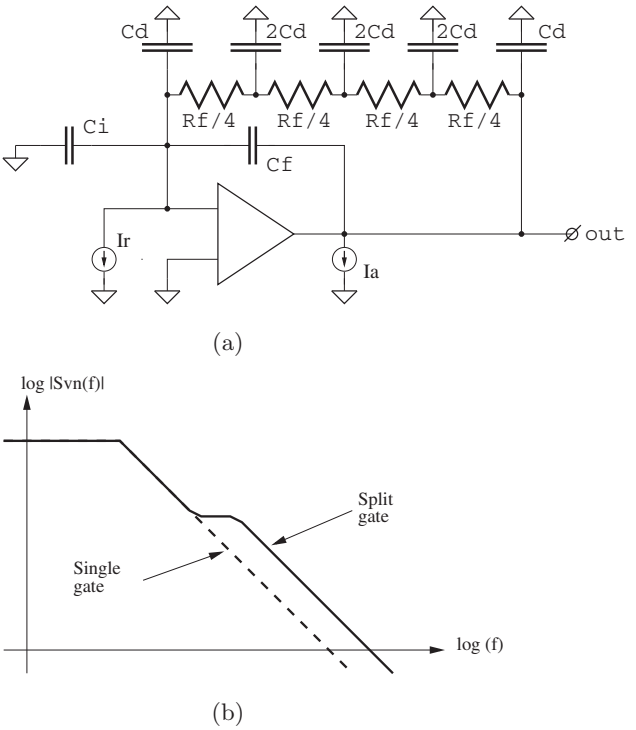


Fig. 5.13. Breaking feedback transistor into four gates: (a) equivalent circuit. (b) Increased R_f noise PSD at the output

$10\,\mu$m width (Fig. 5.13). By doing so, we insert parasitic diffusion capacitors into the feedback network. These capacitors shunt a part of the feedback current to ground and increase the contribution of R_f noise. The actual spike band noise due to feedback resistors is increased from $1.3\,\mu$V to $1.6\,\mu$V .

Targeting the total preamplifier spike band noise at $2\,\mu$V limits the contribution of i_a to $1.2\,\mu$V . The amplifier should be sized and biased accordingly. The sizing procedure is described in Sec. B.3.

5.6.3 Discussion

The expression for the input-reflected noise PSD of the integrated preamplifier used in `NPR03` is brought in Sec. B.3. It can be written in the form of

$$\left(\frac{C_i^*}{C_i}\right)^2 \cdot \left(\frac{1}{g_m}\right)^2 \cdot \left[\frac{16kT}{3} + \frac{1}{g_m}\sum_j i_{n,j}\right]$$

where g_m is the transconductance of the input differential pair, C_i^* is the total capacitance on the input node and $i_{n,j}$ are noise current power spectra of

other transistors in the signal path. It clearly shows that by making the ratios $i_{n,j}/g_m$ small enough, noise contributions of all but the input transistors can be suppressed. This requires bringing the transconductance of the bottom NMOS current sources below g_m, not easily done, remembering that NMOS sources and input differential pair exhibit the same current flow. NMOS transistors must have small W/L, typically resulting in extremely large lengths and large gate-source voltages. This was clearly demonstrated in [63] (albeit with a somewhat different circuit topology), where NMOS sources were made with L of $40\,\mu$m . The total area of the preamplifier was $0.16\,\text{mm}^2$. Such area requirements were considered unacceptable during NPR03 design. We have used shorter NMOS transistors resulting in a worse power-noise tradeoff and much smaller area, $0.075\,\text{mm}^2$. This noise-power-area tradeoff can be relaxed by a novel preamp circuit we describe in Sect. 5.9.

5.7 NPR03 Measurements

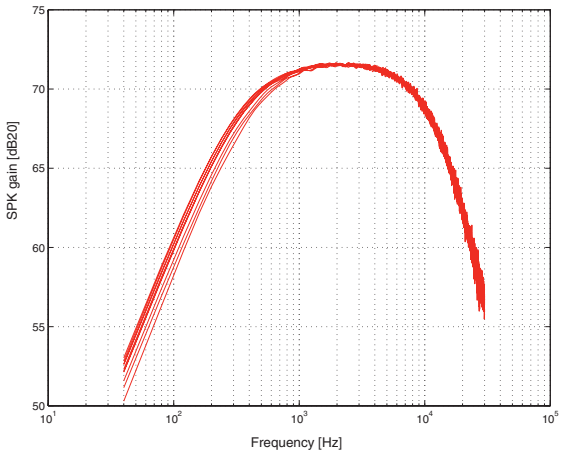
The measurements were carried out on the NPR03 chip automatically, using the evaluation kit (Fig. 5.5).

Figure 5.14 shows cumulative plots of the frequency response of spike and LFP channels from several dies. We have used the circuits from NPR02 , with an exception of re-designed preamplifier. Thus, splitter pole location deviation and gain errors that we have experienced with NPR02 are present. The gain is 3.8 K for spike and 430 for LFP channels (the new preamplifier has $\times 50$ gain, thus the target gains are two times smaller compared to NPR02).

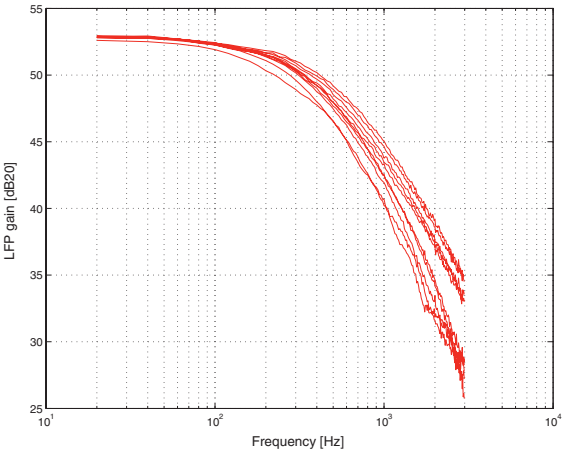
Figure 5.15 shows the behavior of the input HPF for several DAC settings, as measured on the LFP channel.

Spike and LFP channel noise PSD is shown in Fig. 5.16. Noise PSD was measured at the output and divided by the channel gain. Black solid lines represent the simulated curves. The noise measured is indeed close to the expectations with an exception for the displaced splitter frequency. RMS values are $2.9\,\mu$V for spike and $14\,\mu$V for LFP. On the LFP channel the $1/f^2$ curve of the feedback resistor is clearly visible. The LFP noise measurement is limited by the quantization noise of the sample-and-hold, not seen on the spike channel, as the LFP gain is almost ten times lower.

Another noise measurement was made in parallel on the integrated channel and a channel with external capacitors, such as were used in NPR02 (several such channels were included in NPR03 for testing) (Fig. 5.17). Both channels were measured with grounded input, in parallel. Heavy external interference is seen on the channel with extrnal cap, that is not present on the integrated channel. The interference is picked up on the external capacitors and PCB tracks between the capacitors and the chip inputs, that are not present with integrated channels. With some optimism about the improved external noise immunity of the integrated channels, it will not necessarily improve recording quality, as we expect most interference to be picked up on the electrodes.



(a)



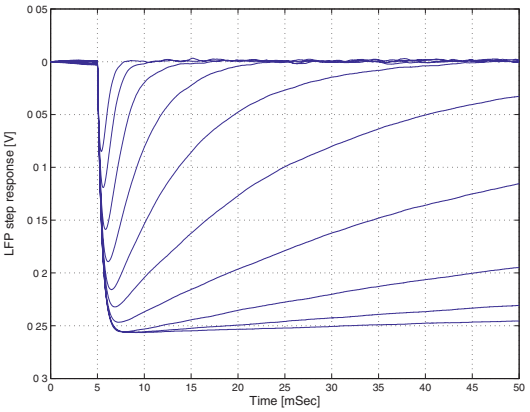
(b)

Fig. 5.14. Frequency response cumulative plots: (a) Spike. (b) LFP

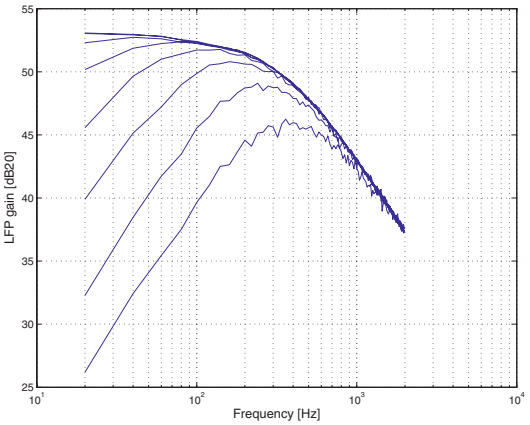
Measurements of DAC, ADC and SAH circuits are brought in Sect. B.4 and in [91].

5.8 An NPR03 -Based Miniature Headstage

A miniature headstage (Fig. 5.18) was designed using NPR03 . It was successfully tested in-vivo with Michigan probes implanted in rat cortex. Samples of recorded signals are shown in Fig. 5.19.

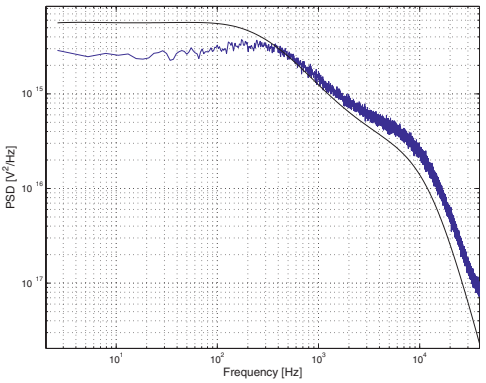


(a)

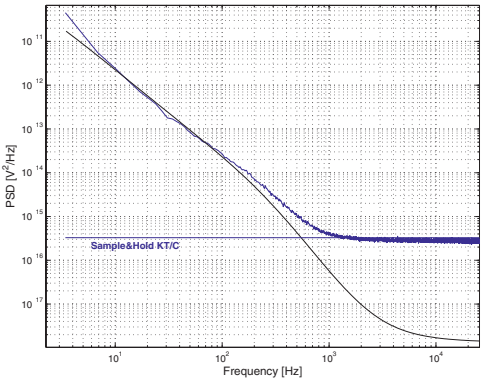


(b)

Fig. 5.15. Input HPF response for several DAC settings: (a) Time axis. (b) Frequency axis



(a)



(b)

Fig. 5.16. Channel noise reflected to input: (a) Spike. (b) LFP

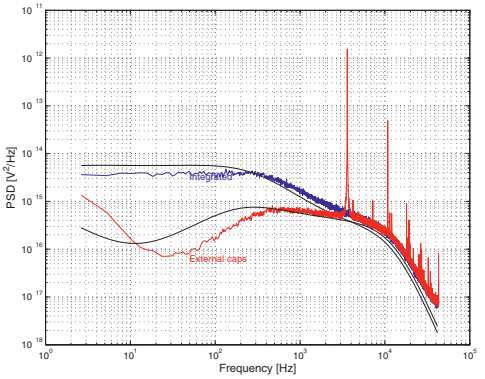
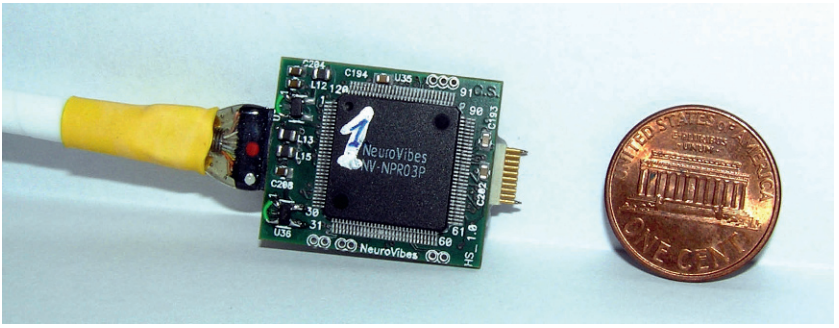
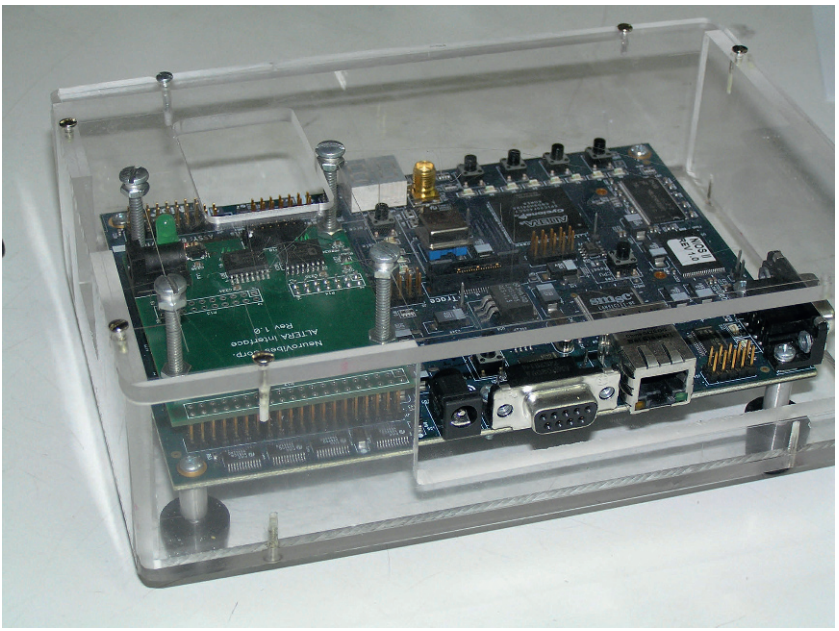


Fig. 5.17. Integrated vs. non-integrated channel noise PSD

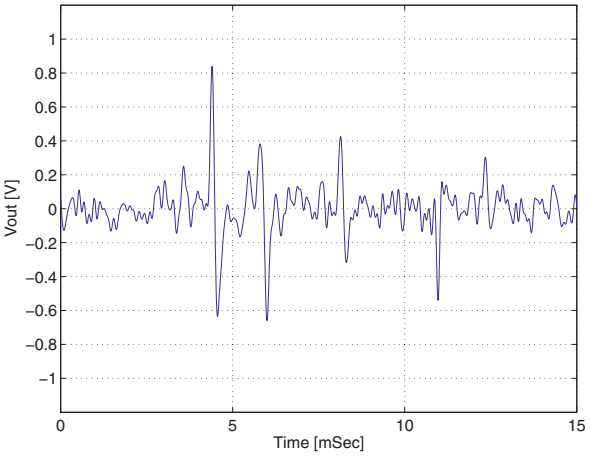


(a)

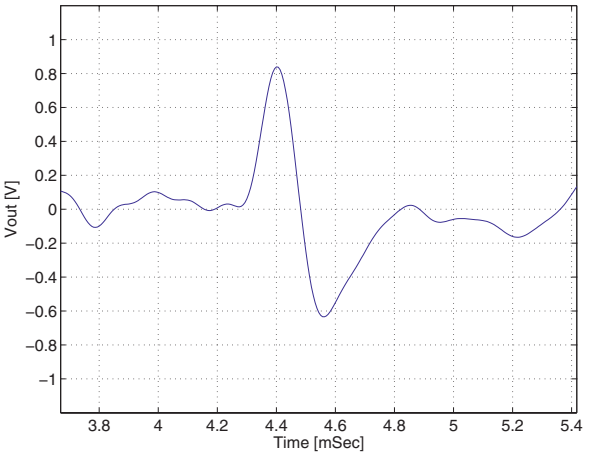


(b)

Fig. 5.18. (a) NPR03 headstage. (b) Encapsulated embedded interface board



(a)



(b)

Fig. 5.19. (a) Recorded signal with NPR03 headstage from rat cortex. (b) Closeup

The headstage interfaces a PC host via the same FPGA board used for NPR03 measurements. Headstage supply planning was done with special care, to avoid digital supply interference and ground loops. The headstage power supply system is shown in Fig. 5.20. Passing McBSP signals through magnetocouplers [94] allows for complete separation between digital (FPGA board) and analog (FPGA board daughter card) supplies. The headstage receives preregulated 5 V supply from the daughter card and regulates it down to 3.3 V and 1.65 V, for NPR03 power. To avoid passing the power lines through the flexible cable (potentially degrading the supply, introducing noise) the daughter card replicates the headstage regulator to power the magnetocouplers. Note

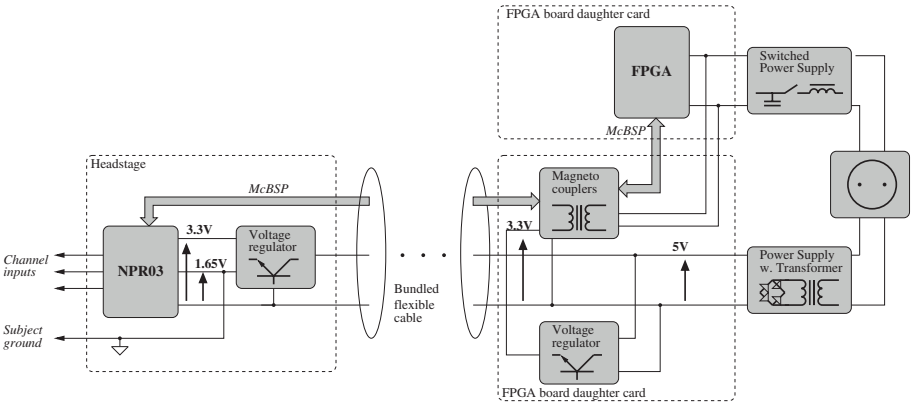


Fig. 5.20. NPR03 headstage supply wiring

that NPR03 analog ground, which should be the ground of the subject, is at 1.65 V potential with respect to the low point of the daughter card supply. For this reason, an isolated power supply (i.e. with transformer) must be used for powering the headstage, since it has no direct connection with wall ground. Otherwise, the headstage regulator can get shortened if the subject touches the wall ground (i.e. water pipes) or if one wishes to measure signals on the headstage and connects a scope to the headstage ground (scope ground is usually tied to wall ground).

The headstage was also tested with a glass multielectrode array (MEA). The setup is shown in Fig. 5.21.

Sample signals captured from MEA are shown in Fig. 5.22. The performance of NPR03 was also verified against a commercial sysem for MEA recording. Figure 5.23 shows segments of signals recorded by NPR03 (blue) and by the commercial system (red). The two signals recorded from the same MEA (but not simultaneously). Although the commercial system states somewhat larger noise level than NPR03 , the plot shows identical noise levels, because the limiting noise factor is the electrode, which is identical in both cases. Spike waveforms differ, due to different filter configurations implemented in both systems.

5.9 A Novel Opamp for The Front-End Preamplifier

A key issue with design of a front-end amplifier, such as the one in Fig. 5.9, or as the one brought in [63], is to bring down the transconductance of NMOS sources far below that of the input differential pair. Together with L^{-2} $1/f$ noise dependence, this results in very large NMOS channel lengths and overdrive voltages.

We suggest to improve this situation by replacing the bottom sources with resistors (Fig. 5.24). The immediate advantage is the lack of $1/f$ noise. The

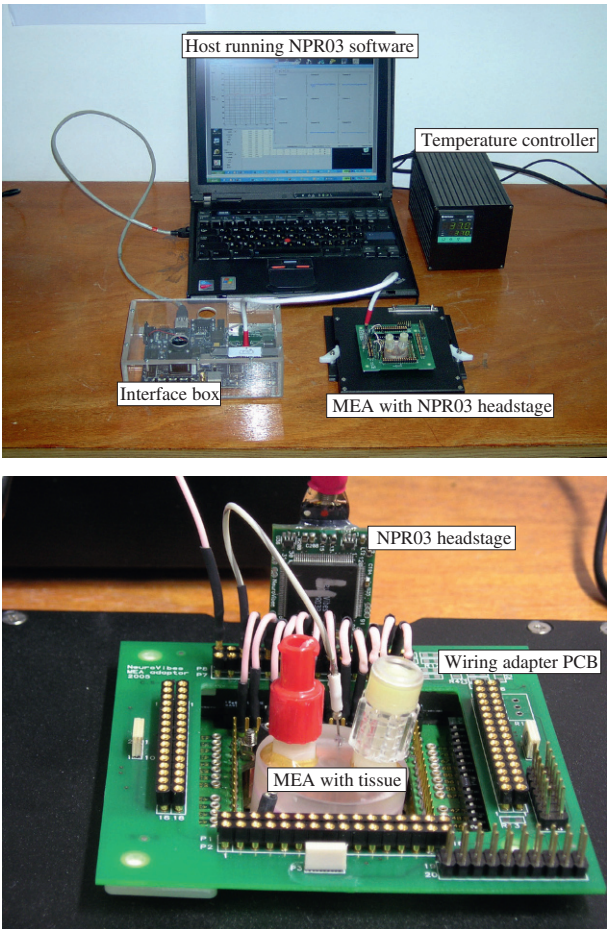


Fig. 5.21. Testing NPR03 with MEA

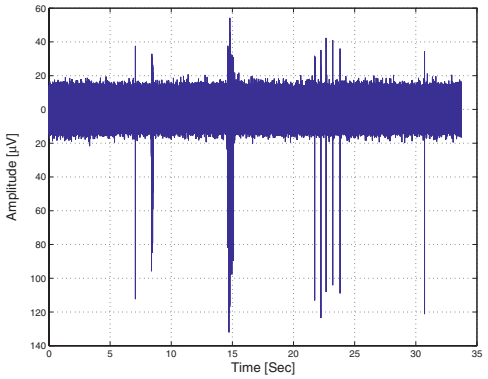
thermal noise of a saturated MOS transistor with overdrive voltage V_{ov} and channel current I can be written as:

$$S_{i,MOS} = \frac{8}{3} kT \left(\frac{2I}{V_{ov}} + g_{m,bs} \right)$$

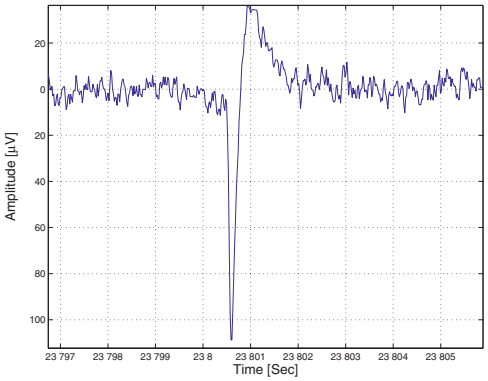
Following the notation of Sect. B.3, we replace $g_{m,bs}$ by ξg_m, and rewrite the above as:

$$S_{i,MOS} = \frac{8}{3} kT \left(1 + \xi\right) \frac{2I}{V_{ov}} = \frac{16}{3} kT \left(1 + \xi\right) \frac{I}{V_{gs} - V_T}$$

If we replace the transistor by a resistor, so that the operating point is unchanged, we must use $R = V_{ds}/I$. The noise generated by R is



(a)



(b)

Fig. 5.22. Signals recorded from MEA with NPR03 (a). A closeup on a single spike (b)

$$S_{i,R} = \frac{4kT}{R} = 4kT\frac{I}{V_{ds}}$$

One is not likely to place V_{ds} below V_{gs}, to maintain enough signal headroom. In [63], NMOS sources operate with $V_{ds} = V_{gs}$; in NPR03 preamp circuit, V_{ds} much higher than V_{gs} was used. Even when operating with $V_{ds} = V_{gs}$, the resistor is clearly generating less noise than MOS at the same operating point.

Replacing transistors with resistors deteriorates, however, the circuit gain. The maximal possible gain, assuming the input transistors are large enough to operate at a very weak inversion, is:

$$g_m R = \frac{I}{\kappa V_{th}} \cdot \frac{V_R}{I} = \frac{V_R}{\kappa V_{th}}$$

The gain is not likely to be above several tens; for feedback operation a larger value is desirable.

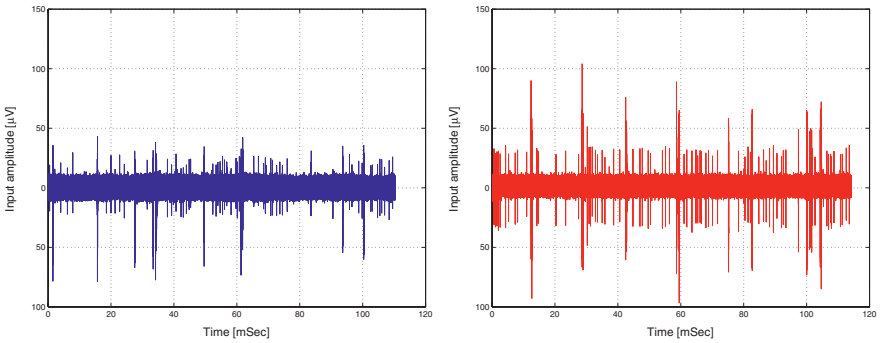


Fig. 5.23. Signal recorded by `NPR03` (left) and a commercial system (right)

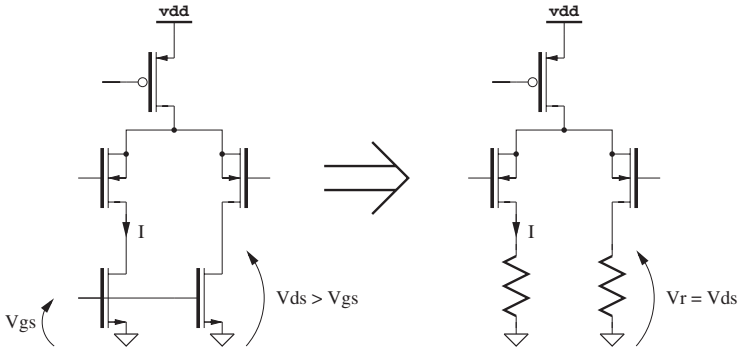


Fig. 5.24. Replacing NMOS current sources with resistors

The gain can be increased by introducing another stage; stability issues have to be addressed. In a typical two-stage amplifier the dominant pole is placed at the output of the first stage, taking advantage of its high output resistance. The second stage also typically carries a large current, so that the second pole (at amplifier output) is placed far above the dominant pole. In our case, neither of the above points holds: The first stage has a low output resistance of R. The second stage current must be significantly lower than that of the first stage, not to increase the current consumption. Hence, we place the dominant pole at the amplifier output and the second pole at the output of the first stage. The complete circuit is shown in Fig. 5.25.

5.9.1 Noise Analysis

By the following analysis we shall derive a ratio I_1/I_2 that yields the minimal noise under a given current consumption. For this approximate analysis we shall neglect the $1/f$ noise contribution. We expect it to be minor, as the

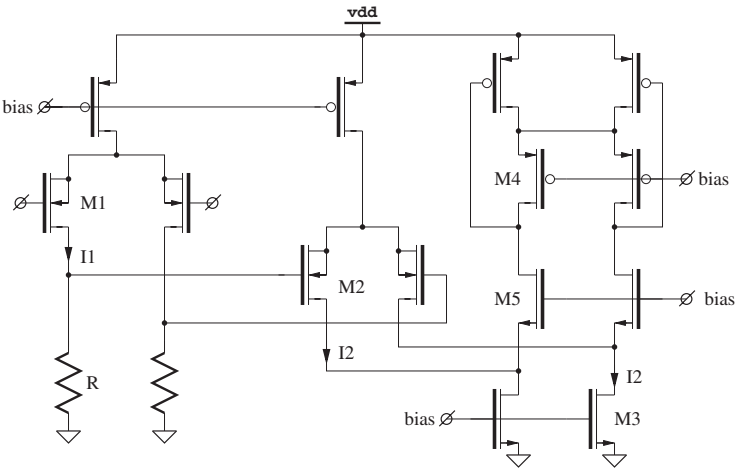


Fig. 5.25. Novel opamp circuit for input preamplifier

input PMOS devices have large gate areas and operate in subthreshold. The insignificant contribution of $1/f$ noise was also noted in [63].

The output thermal noise current PSD of the circuit (open loop) can be written as

$$S_{i_{o,n}} = 2 \cdot \frac{8}{3} kT \left[(1+\xi) g_{m1} R^2 g_{m2}^2 + \frac{3}{2} R g_{m2}^2 + (1+\xi)\left(g_{m2} + g_{m3} + g_{m4}\right) \right]$$

The output voltage noise of the closed loop circuit is:

$$v_{o,n} = \frac{C_i^*}{C_f} \cdot \frac{1}{sC_i^* + G_m} i_{o,n}$$

where C_i^* is the total capacitance at the input node, and G_m is the total transconductance of the opamp, $G_m = g_{m1} R g_{m2}$. We assume that for the frequencies of interest $G_m \gg sC_i^*$ and reflect $v_{o,n}$ to the input by circuit gain:

$$v_{i,n} = \frac{C_i^*}{C_i} \cdot \frac{1}{G_m} i_{o,n}$$

Substituting $S_{i_{o,n}}$, the input noise PSD can be written as:

$$S_{v_{i,n}} = \left(\frac{C_i^*}{C_i}\right)^2 \frac{16}{3} kT \frac{1+\xi}{g_{m1}} \left[1 + \frac{3}{2(1+\xi)} \frac{1}{g_{m1} R} + \frac{1}{g_{m1}} \cdot \frac{g_{m2} + g_{m3} + g_{m4}}{g_{m2}^2 R^2} \right]$$

It is clear that the noise decreases with g_{m2} and increases with g_{m4}, thus we shall size `M2` large, and `M4` small. We write the transconductance of a MOS transistor in subthreshold (accoridng to [95]) as

$$g_{m,j} = \frac{\kappa I_j}{V_{th}} \cdot G_j$$

where $G(I)$ is a function of the saturation current I_S:

$$G(I) = \frac{2}{1 + \sqrt{1 + 4I/I_S}}$$

$G(I)$ is a slowly changing function, ranging from 0.9 for very weak inversion ($I = 0.1I_S$) to 0.3 at the border of subthreshold region ($I = 10I_S$). For approximate calculations we shall assume G_4 of 0.3 and G_2 of 0.9. We substitute

$$g_{m1} = \frac{\kappa I_1}{V_{th}} G_1 \qquad g_{m2} = \frac{\kappa I_2}{V_{th}} G_2$$
$$g_{m3} = \frac{2 \cdot 2I_2}{V_{ov,3}} \qquad g_{m4} = \frac{\kappa I_2}{V_{th}} G_4$$
$$R = V_R / I_1$$

and rewrite $S_{v_{i,n}}$ as:

$$S_{v_{i,n}} = \left(\frac{C_i^*}{C_i}\right)^2 \frac{16}{3} kT(1+\xi) \frac{V_{th}}{\kappa I_1 G_1} \left[1 + \frac{3}{2(1+\xi)} \frac{V_{th}}{\kappa G_1 V_R} + \right.$$
$$\left. + \left(\frac{V_{th}}{\kappa V_R}\right)^2 \cdot \frac{I_1}{I_2} \cdot \frac{1}{G_1 G_2} \cdot \left(1 + \frac{4V_{th}}{\kappa V_{ov,3} G_2} + \frac{G_4}{G_2}\right)\right]$$

For any given I_1 and I_2, this equation yields an optimal width W for the input differential pair (the one that minimizes $S_{v_{i,n}}$). If we constrain the total current dissipation, $I = I_1 + 2I_2$, and rewrite the equation once again,

$$S_{v_{i,n}} = \left(\frac{C_i^*}{C_i}\right)^2 \frac{16}{3} kT(1+\xi) \frac{V_{th}}{\kappa G_1} \left[\frac{1}{I_1}\left(1 + \frac{3}{2(1+\xi)} \frac{V_{th}}{\kappa G_1 V_R}\right)\right.$$
$$\left. + \frac{2}{I - I_1} \left(\frac{V_{th}}{\kappa V_R}\right)^2 \left(\frac{1}{G_1 G_2}\right) \left(1 + \frac{4V_{th}}{\kappa V_{ov,3} G_2} + \frac{G_4}{G_2}\right)\right]$$

we can also find optimal ratio I_1/I. If we recall that $G(I)$ is a slowly changing function of I, this task reduces to minimizing $f(I_1)$,

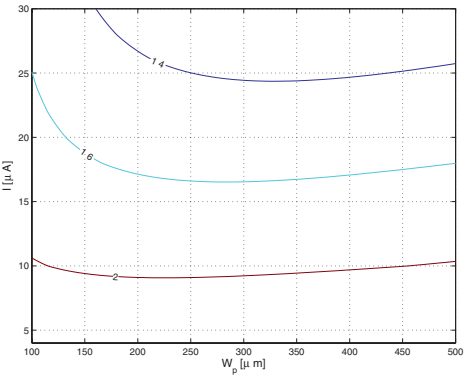
$$f(I_1) = \frac{1}{I_1}\left(1 + \frac{3}{2(1+\xi)} \frac{V_{th}}{\kappa G_1 V_R}\right)$$
$$+ \frac{2}{I - I_1} \left(\frac{V_{th}}{\kappa V_R}\right)^2 \left(\frac{1}{G_1 G_2}\right) \left(1 + \frac{4V_{th}}{\kappa V_{ov,3} G_2} + \frac{G_4}{G_2}\right)$$
$$= \frac{A}{I_1} + \frac{2B}{I - I_1}$$

which is easily found as

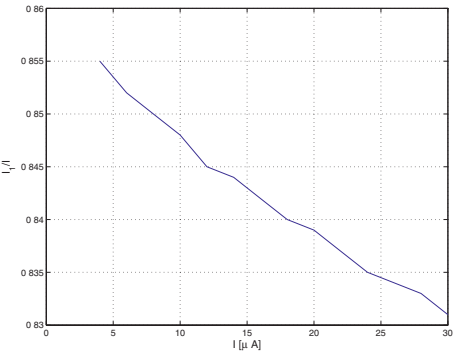
$$\frac{I1}{I} = \frac{1}{1 + \sqrt{2B/A}}$$

This can be calculated to $I_1/I = 0.87$, using 0.5 as an estimate of G_1. Thus we conclude that adding a second amplifier stage for a better gain incurs only a modest penalty in current consumption (13% of the total current).

Simple numerical model optimization (using MATLAB) for I_1, I_2 and W_1 (with L_1 of 1 μm) is shown in Fig. 5.26. The noise is calculated over 200 Hz-10 kHz bandwidth, corresponding to the bandwidth of neuronal spike signals; $1/f$ noise is included in the model. The ratio I_1/I is indeed very close to our



(a)



(b)

Fig. 5.26. Numerical model optimization: (a) Contour plots of input RMS noise vs. bias current and W_p. (b) Optimal ratio I_1/I vs. bias current (I)

prediction and exhibits only minor changes with I.

Figure 5.27 shows the input-referred RMS noise at optimal points as a function of I, obtained with MATLAB model and SPECTRE simulations.

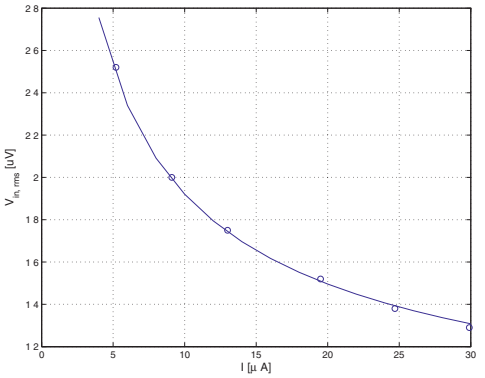


Fig. 5.27. RMS input noise voltage vs. bias current. MATLAB model and SPECTRE simulations

5.9.2 Stability

The loop transfer function of the circuit can be written as:

$$LT(s) = \frac{C_f}{C_i^*} \times \frac{sC_f - G_m(s)}{sC_o^*}$$

C_i^* and C_o^* are the total capacitances connected to the input and the output nodes of the amplifier. We have disregarded the effects of R_f, that donates a pole and a zero in very low frequencies (about 1 Hz) and does not affect stability. We have also assumed that the dominant pole is placed at the origin. In other words, the output impedance of the transconductor is very high. $G_m(s)$ has two poles: one is formed between R and the capacitance at the output of the first stage, the other is formed at the source of the output cascode transistor `M5`.

$$G_m(s) = \frac{G_{m0}p_1p_2}{(s+p_1)(s+p_2)}$$

Substituting this into LT yields:

$$LT(s) = \frac{C_f^2}{C_i^*C_o^*} \times \frac{s(s+p_1)(s+p_2) - G_{m0}p_1p_2/C_f}{s(s+p_1)(s+p_2)}$$

Our goal is to provide a 60° phase margin. We shall first find the unity gain frequency ω_0:

$$1 = |LT(j\omega_0)| = \frac{C_f^2}{C_i^*C_o^*} \left| \frac{j\omega_0(j\omega_0+p_1)(j\omega_0+p_2) - G_{m0}p_1p_2/C_f}{j\omega_0(j\omega_0+p_1)(j\omega_0+p_2)} \right|$$

Let's assume that the parasitic poles $p_{1,2}$ are well above this frequency:

$$\begin{aligned} 1 &= |LT(j\omega_0)| \\ &\cong \frac{C_f^2}{C_i^* C_o^*} \left| \frac{j\omega_0(p_1)(p_2) - G_{m0}p_1p_2/C_f}{j\omega_0(p_1)(p_2)} \right| \\ &= \frac{C_f^2}{C_i^* C_o^*} \frac{\sqrt{\omega_0^2 + G_{m0}^2/C_f^2}}{\omega_0} \end{aligned}$$

Solving for ω_0:

$$\omega_0 = \frac{C_f}{C_i^* C_o^*} G_{m0}$$

given that $C_i^* C_o^*/C_f^2 \gg 1$. Now we can check the phase at ω_0. Again, we assume that $\omega_0 \ll p_1, p_2$:

$$\begin{aligned} \angle LT(j\omega_0) &= \angle \frac{j\omega_0(j\omega_0 + p_1)(j\omega_0 + p_2) - G_{m0}p_1p_2/C_f}{j\omega_0(j\omega_0 + p_1)(j\omega_0 + p_2)} \\ &\cong \angle \frac{j - C_i^* C_o^*/C_f^2}{j(1 + j\omega_0/p_1)(1 + j\omega_0/p_2)} \\ &\cong -\frac{\pi}{2} - \frac{C_f^2}{C_i^* C_o^*} - \omega_0 \left(\frac{1}{p_1} + \frac{1}{p_2} \right) \end{aligned}$$

For phase margin of 60°, the following must be satisfied:

$$\begin{aligned} \frac{\pi}{6} &> \frac{C_f^2}{C_i^* C_o^*} + \omega_0 \left(\frac{1}{p_1} + \frac{1}{p_2} \right) \\ &= \frac{C_f}{C_i^* C_o^*} \left(C_f + G_{m0} \frac{p_1 + p_2}{p_1 p_2} \right) \end{aligned}$$

By substituting:

$$\begin{aligned} p_1 &= (RC_{p1})^{-1} \\ p_2 &= (g_{m5}C_{p2})^{-1} \end{aligned}$$

where C_{p1} and C_{p2} are the capacitances at the output of the first stage and at the source of M5, we obtain:

$$\frac{\pi}{6} > \frac{C_f}{C_i^* C_o^*} \left[C_f + \frac{\kappa V_R}{V_{th}} \frac{G_1 G_2}{G_5} \left(Cx2 + \frac{\kappa V_R G_5}{V_{th}} \frac{I_2}{I_1} C_{p1} \right) \right]$$

Using the ratio of I_2/I_1 for optimal noise performance, and substituting G_1, G_2 and G_5 of 0.5, 0.9 and 0.9, we have:

$$\frac{\pi}{6} > \frac{C_f}{C_i^* C_o^*} \left[C_f + 8.5 \cdot (C_{p2} + 1.4 \cdot C_{p1}) \right]$$

To satisfy the condition, we place a compensating load capacitor, C_L. C_o^* becomes $C_f + C_L$, and the condition can be satisfied by:

$$C_L > \frac{6}{\pi} \cdot \frac{C_f}{C_i^*} \left[C_f + 8.5 \cdot (C_{p2} + 1.4 \cdot C_{p1}) \right] - C_f$$

For C_f of 1pF, C_i of 5pF the condition is not difficult to hold: if we estimate C_{p2} as 100fF, and C_{p1} as 250fF (as it is loaded by diffusions of input transistors), C_L needs only be 400fF, which can be provided by a 200fF differential load.

5.9.3 Conclusions

Section 5.9 present the design of a power-optimized input preamplifier, achieving the optimum design point in power-noise tradeoff (Fig. 5.27). It features fully differential architecture, potentially allowing for high PSRR and CMRR. The usual NMOS active loads at the first stage were replaced by resistors, providing a better power-noise tradeoff. In particular, they significantly lower 1/f noise contribution. A second stage was added to provide high gain, that was deteriorated by active load replacement. It is shown that the second stage does not contribute significantly to power consumption (consumes 15% of the total current).

5.10 Conclusions

This concludes our discussion of the integrated front-end circuits for neural sensing. Three generations of such circuits were developed, with the latter being a system-on-a-chip, with twelve differential sensing channels with numerous programmable features and integrated, per-channel ADC. The channels accessed over an internal parallel bus by an integrated controller, capable of sampling all the channels and streaming the data over a digital serial communication channel. The front-end circuit was tested with both in-vivo and in-vitro recording and proved operational. Finally, we presented a novel design for the input stage to be included in future versions of the front-end.

6

Algorithms for Neuroprocessor Spike Sorting

6.1 Introduction

The first stage of processing the neural signals coming from a recording front-end may be the detection of time points of firing events. This task is typically referred to as *spike-detection*. Since an electrode might sense activity of several adjacent neurons, one may try to classify detected firing events to their originating neurons. This process is called *spike-sorting*. What makes spike-sorting possible is that action potentials generated by different neurons will have fairly different characteristic shapes. Sorting algorithms try to classify incoming spikes into shape-groups and treat every group as if originated by the same neuron.

Spike detection and sorting algorithms have been the subject of extensive work over many years [96]. Many different methods were proposed, including various clustering methods such as feature analysis [96] or PCA [97], template matching [98], wavelet-transform based methods [99, 100, 101] and artificial neural networks [102].

The purpose of the work presented in this chapter was to develop algorithms for integrated spike detection and sorting. The feasibility of on-chip sorting was shown in [78], which examined the power requirements of hardware implementation of some common sorting algorithms. We show below that spike-sorting power consumption can be reduced significantly by trading some classification accuracy in state-of-the-art clustering algorithms in return for considerable savings in power. Detailed description of the study, including algorithm validation, can be found in [103, 104] and in [105].

6.1.1 Clustering Methods

Clustering methods apply transformations on spike signals mapping them onto a shape-space to form scatter-plots. Under a successful mapping spikes of similar shapes fall close to each other, while spikes of different shapes lie

apart. The resulting scatter plot has a set of clusters, areas with high density of mapped spikes, each representing spikes of a different characteristic shape. Sorting is performed by assigning spikes to clusters, for example, by choosing the closest cluster. Figure 6.1 shows an example of a two-dimensional scatter-plot obtained by measuring two features of spikes. The process is successful if

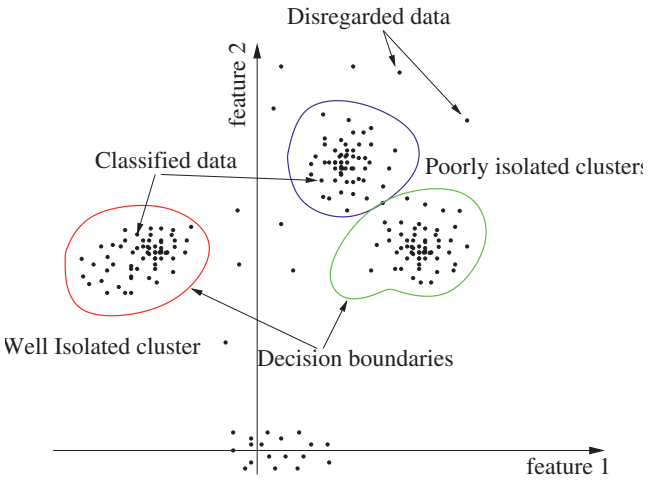


Fig. 6.1. Example scatter plot of two measured features

the clusters in the shape-space can be distinguished, and the distances between cluster centers (typical shapes) are larger than cluster spread (background noise influence).

Shape-space can be defined by selecting a set of spike shape features, such as the height of the first peak, the height of the second peak or spike "width", i.e. time interval between points at 50% maximal height. Measured features define the axes of this space. This is the method of *feature analysis*. To make it work reliably, one must select the most discriminative set of features. This technique was common in the past [96], mostly due to the very modest computation requirements: typically the selected features are measured on the signal and require none or very little computation. Today feature analysis is replaced by a more precise and significantly more computationally complex Principal Component Analysis [97].

Cluster boundaries can be defined on the scatter plot once it has been generated. Spikes that fall inside the boundaries of a particular cluster are assigned to that cluster. Those that fall outside of any cluster boundaries are disregarded. Often, decision boundaries are defined by manually drawing polygons on the shape-space [106]. Automatic approaches exist as well, such as K-means or Bayesian clustering [107, 108].

6.1.2 Spike Detection and Alignment

Both feature analysis and PCA do not operate on the continuous recorded signal. Instead, they process signal frames that contain suspected spikes, implying that initial detection must be performed prior to the spikes being sorted. This is typically done by monitoring threshold-crossing events of the input signal. The threshold is positioned somewhere between the background noise level and the level of spike peaks. Threshold placement trades off the amount of false-triggered spikes versus the amount of missed true-spikes. A circuit for automatic noise level estimation and subsequent threshold adjustment was proposed in [74].

Once the threshold crossing is detected, there still exists a question of frame alignment with respect to threshold crossing. For example, one might start the frame at the threshold crossing instant. However, since the threshold is placed high above the background noise level, a significant portion of spike (prior to the threshold crossing) is lost. Another possibility is to align the frames with respect to spike peaks (i.e. to position spike peaks at a certain offset from frame start). But the peak point can be affected by noise, since the signal has a zero derivative near the peak. Yet another approach was taken by the commercial software [29] (described in [109]): the frame is aligned to the point of minimal error between the spike and its representation by the principal components used for sorting. The rationale is clear, remembering that the whole purpose of the alignment is to assist sorting. The computational burden is immense: full principle component representation has to be calculated and subtracted from the signal for every possible frame offset.

6.1.3 Issues in Spike Sorting

Several factors make sorting spikes difficult. Simultaneous firing of two units (*overlapping spikes*) sensed by the same electrode may result in spike shape that is not similar to either of the units involved. Moreover, such firing does not have a characteristic shape, as its waveform depends on the relative position of firing events on the time axis. On the occasion of units firing in opposite directions, the spikes may cancel one another.

Action potential shapes tend to change when a "pulse train" is fired (sequence of pulses with a very short inter-spike interval). Towards the end of the train amplitudes are diminished and widths are increased. Instability in recording conditions, such as subject movement or electrode movement within the tissue can also influence the shapes of recorded spikes. Bayesian classification may perform better under non-stationary recording conditions as was shown in [107]. Moreover, thanks to classification certainty prediction of Bayesian method, it is possible to track recording problems based on drops in classification certainty.

6.2 Spike Sorting in a Neuroprocessor

Spike detection and sorting reduces the recorded data from raw analog waveforms to mere identifications of spike occurrences and origins. As discussed in Chap. 3, the Neuroprocessor must employ spike sorting following the recording front-end to allow low-power operation; otherwise it will spend too much power on communicating the unnecessary data.

Regardless of the underlying algorithms, spike sorting by clustering consists of two distinct tasks: mapping spikes onto a shape-space (scatter plot generation) and relating the points on the shape-space to clusters (definition of decision boundaries on the scatter plot). Both need certain computation parameters to be known before spikes can be actually sorted: for scatter plot generation, one has to set the mapping, i.e. determine the discriminative features or principal components; decision boundaries are required for clustering. With off-line experiments, where raw signals are recorded for a period of time for later processing, the sorting is performed on the entire set of recorded data. This is not possible with Neuroprocessors: real-time operation requires processing the spikes "on the fly". Neuroprocessor sorting algorithms must therefore operate in two phases: learning and production. During learning, production phase parameters are calculated. Ideally, learning is done once, during the initialization of the Neuroprocessor. In practice, however, the algorithms are to be retrained once in a while due to the instability of recorded signal shapes. Neuroprocessor must track the sorting process, and re-train itself when needed. With Bayesian sorting, for example, it is possible to track the classification certainty and re-train when a drop is detected.

Learning is based on knowledge gained from a cumulative picture of many recorded spikes, while production is performed on every spike separately. Consequently, one must strive to reduce the production phase computations to a minimum, while the learning can be very complex. Luckily, clustering algorithms follow this paradigm. For example, principal components are calculated as eigenvectors of the covariance matrix of the recorded spikes. Calculation of a covariance matrix requires a large number of spikes and finding eigenvectors of a large matrix is far from trivial. On the other hand, once we have calculated the principal components, signal mapping involves only computation of several dot-products. A similar approach is applied to clustering: determination of decision boundaries involves complex clustering algorithms; once the boundaries are defined, checking whether a given point lies inside a given polygon is trivial.

Figure 6.2 illustrates spike sorting in a Neuroprocessor. Production mapping is placed inside the channel, to avoid communicating the raw data even on the internal bus. The mapping stage also includes detection and alignment of the signal. Decision regarding the point assignment on the shape-space can be relayed to a global chip processor, as the amounts of data to be communicated are modest (point coordinates). The learning can be performed at any place in the path, including learning on the external host. If the mapping

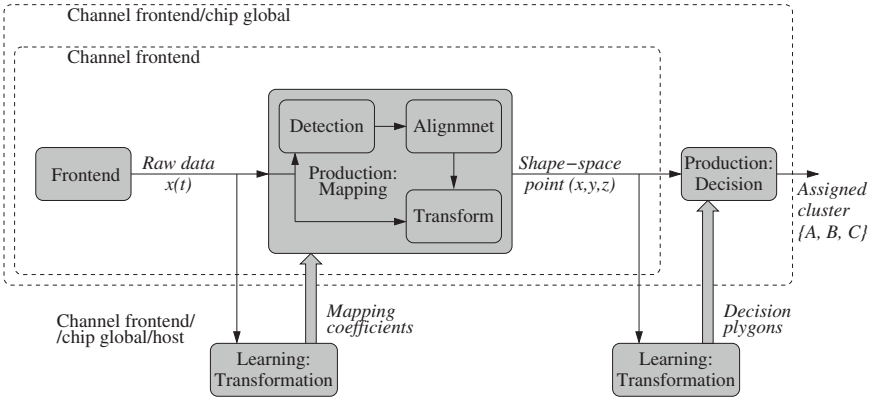


Fig. 6.2. Neuroprocessor spike sorting

training is to be done on a host, raw-signal communication is required. In that case, part of the chip bandwidth is allocated for raw-signal communication from the trained channel. Of course, only one or a small number of channels can be trained at a time, due to the limited bandwidth. With a large number of channels it also poses a constraint on how often a single channel can be retrained.

6.3 Spike Sorting Algorithms

This section presents several algorithms for spike sorting [103]. We start from a full-fledged Principal Component Analysis and gradually introduce changes to reduce its computation complexity (while also decreasing classification accuracy). Before that, we address the evaluation of these algorithms. An algorithm is characterized by two qualities: power consumption for involved computation and classification accuracy. Since there is no reliable way to determine power consumption prior to actual circuit design, we use computation complexity as an estimate for power consumption. This can be expressed as the total number of additions and multiplications required in a given algorithm. Multiplication is roughly counted as ten additions to generate a single algorithm "complexity index".

The accuracy of the algorithms was validated against commercial PCA-based spike-sorting reference software [29]. Spike data was taken from electrophysiological recordings, obtained from microelectrodes in various cortical regions. Spikes were detected and aligned off-line, by the reference software. The output of detection and alignment were passed to the reference spike-sorter and to the validated algorithm and the results were compared. Two types of errors emerge when a proposed algorithm is compared with the reference: A spike can be *unclassified*, i.e. it is not assigned to any cluster by

the algorithm, while it has been assigned to some cluster by the reference algorithm. A spike can also be *mis-classified*, i.e. assigned to a different cluster than expected. Both error types are combined into a cumulative error rate which is used as an accuracy measure. The algorithm validation scheme is shown in Fig. 6.3.

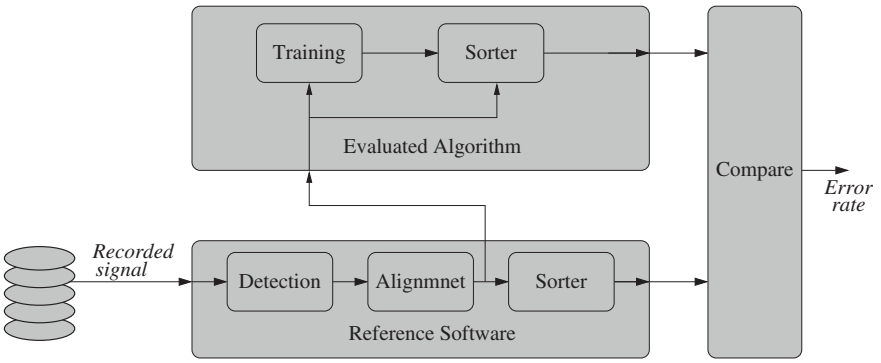


Fig. 6.3. Sorting algorithm accuracy validation

6.3.1 PCA Approximations

The reference sorting algorithm is based on PCA with two principal components. It operates on signal windows of N points. The transformation within the reference algorithm requires computing of two dot-products of N-point vectors, each requiring N multiplications and N additions. Hence, the overall computation burden is $2N$ multiplications and $2N$ additions, corresponding to a complexity index of $22N$.

The first approximation of the above algorithm, *Segmented PCA*, performs similar calculation with downsampled versions of the principle component vectors (Fig. 6.4). The signal is averaged over S time intervals T_i. The value of every integral is multiplied by the average of the principal component vector over interval T_i. A single multiplication is required for each segment, reducing the total number of multiplications for every vector downto S. The number of additions required for signal downsampling is N, and another S additions are needed for the dot-product of the downsampled signal. The total of $2S$ multiplications and $N + 2S$ additions make for complexity index of $22S + N$.

Another level of savings can be obtained by rounding the principal component values to the closest powers of 2. In such case multiplications can be replaced by bit shifting, which is counted as an addition. $N + 4S$ additions and zero multiplications yield a complexity index of $N + 4S$.

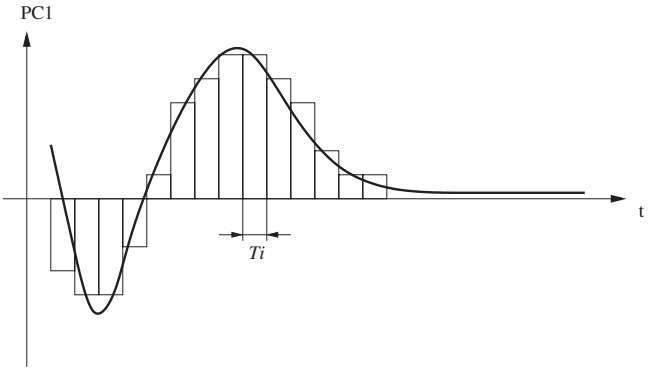


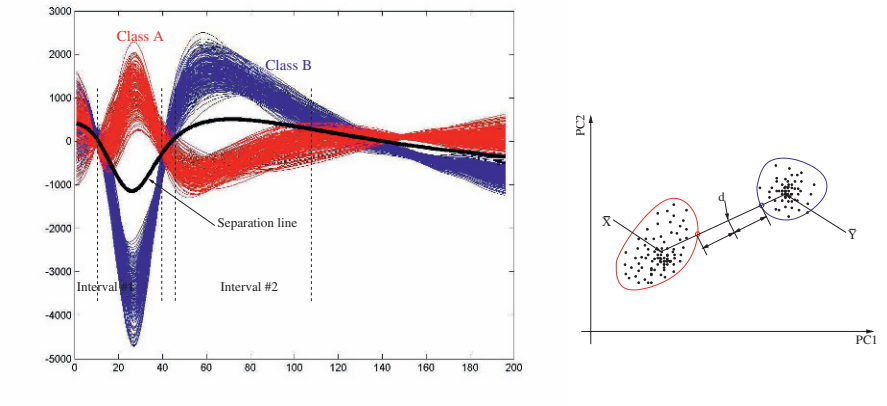
Fig. 6.4. Downsampling the principal component vector

6.3.2 Time Domain Classification

The *Hard Decision*, HD approach involved comparing the recorded signal with a certain set of points, or a "separation line" (Fig. 6.5)(a). Data points of spike of cluster A are likely to be above the separation line during interval #1 and below it during interval #2. Thus the comparisons would generate a high number of ones for interval #1 and a high number of zeroes for interval #2. Exactly the opposite holds for spikes of cluster B. The numbers of ones a signal generates over intervals #1 and #2 are summed into T_1 and T_2, respectively. These are compared against cluster threshold values Th_1 and Th_2. If $T_1 > Th_1$ and $T_2 < Th_2$ then the signal is assigned to cluster A. The procedure is described for two clusters, but can be extended to more clusters by adding more separation lines. The separation line can be obtained as an inverse-projection of a center-point between two clusters on the shape-space. Fig. 6.5(b) shows how the center point is placed between cluster boundaries on the line connecting the clusters' centers-of-mass. HD computation involves N comparisons of the signal window with the separation line, counted as N additions. Hence, the complexity index is N.

The *Soft-Decision* (SD) algorithm is similar to HD, except that the latter first compares the signal with the separation line and then sums up the results, while SD integrates the signal over intervals #1 and #2, directly obtaining T_1 and T_2, and then compares them against sorting thresholds. The number of sums to perform during integration depends on the length of time intervals #1 and #2. For our verification, we have chosen the total length of the intervals as 100 points, giving a complexity index of 100. Typically, $S < 100 < N$.

Looking at HD and SD sorting procedure on the shape-plane spanned by T_1 and T_2, we see that it divides the plane into four quadrants by two lines intersecting at the separation point (Th_1, Th_2) (Fig. 6.6). Two of the quadrants are not classified to either of the clusters. In case of closely-spaced clusters, this leads to a large number of unclassified spikes, as confirmed by verification.



(a) (b)

Fig. 6.5. Time-domain classification

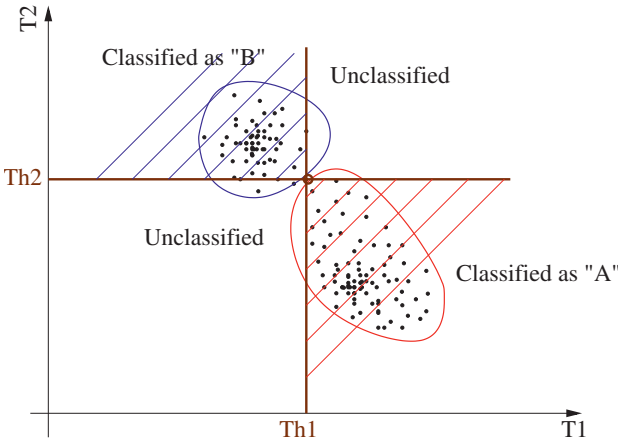


Fig. 6.6. Unclassified areas with HD/SD algorithms

6.3.3 Integral Transform

Time domain classification procedures described above are attractive due to their simplicity. Unfortunately, due to the presence of two unclassified quadrants in the shape space, the accuracy of the results is poor. The results can be improved even with the same shape-space mapping, by employing a better decision scheme. For example, in Fig. 6.6 the lines $x = Th_1$ and $y = Th_2$ can be rotated about (Th_1, Th_2) (Fig. 6.7). Half-plane left of $L1$ is assigned to B, and half-plane to the right of $L1$ to A. $L2$ in this case is not used.

We denote the above method of shape-space mapping as *Integral Transform* (IT); namely integrating the spikes over the two significant phases (i.e.

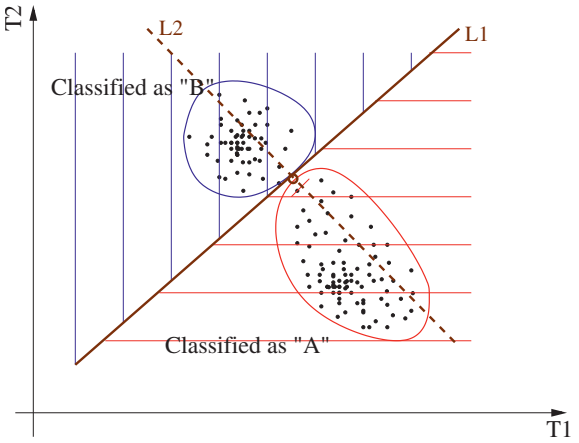


Fig. 6.7. IT sorting

periods when it reaches the peak amplitude). IT can be viewed as "PCA with degenerated PCs": the two vectors used for IT shape-space generation are rectangular windows over intervals #1 and #2. Our trials show that IT can provide accurate results when used with properly defined decision boundaries.

6.3.4 Decision Boundaries

So far, we have looked upon the creation of cluster-plots on the shape-space by various transformations of the signals. A separate problem is to define decision boundaries of the clusters after the plot has been created.

Linear classification has been selected to minimize the cost of hardware implementation. A single line divides the shape-space into two half-planes, which can be sufficient in simple cases for separation of two clusters (Fig. 6.7). In case of a larger number of clusters, or when "outliers" need be ruled out, more lines can be used. The decision regarding a position of a point with respect to every line requires a single multiplication and a single addition. Thus adding a decision line to the shape-space is relatively easy.

6.3.5 Validation

The algorithms were checked with N of 200, as dictated by the reference algorithm. Downsampled PCA versions were verified for seven and fifteen segments (PC7, PC15). Another complexity reduction (PC7R, PC15R) was attempted on downsampled PCA, as the principle component factors (downsampled) were approximated by the closest powers of 2, reducing the multiplication complexity to that of an arithmetic shift (counted as an addition). Validation results [105] are presented in Fig. 6.8 and in Table 6.1. The complexities

shown include the shape-space transformation and classification with a single straight line (which is the cause of a single multiplication in IT, PC7R and PC15R). Time domain algorithms yield high error rates due to large

Table 6.1. Summary of validation results

Alg.	Adds.	Muls.	Compl. Idx.	Uncl. (%)	Miscl. (%)	Err. rate (%)
HD	200	0	200	20	2.7	23
SD	100	0	100	19	2.2	21
IT	100	1	110	0.8	1.4	2.2
PC7R	230	1	240	0.5	1.2	1.7
PC15R	260	1	270	0.5	1.1	1.6
PC7	215	15	365	0.4	1.0	1.4
PC15	230	30	530	0.4	0.9	1.3
PCA	400	400	4,400	0.0	0.0	0.0

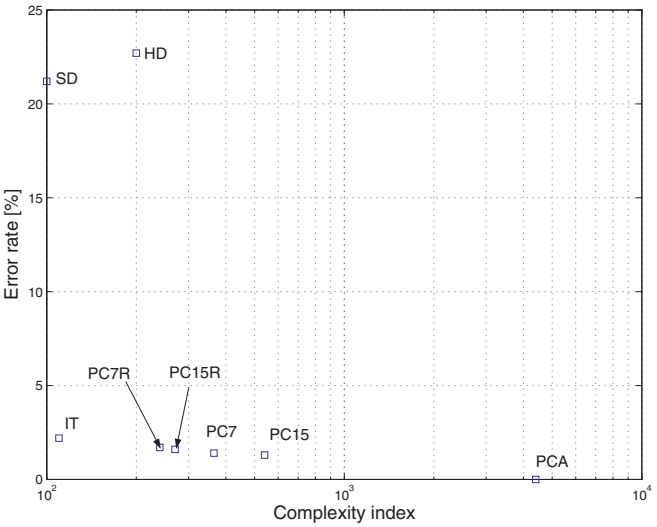


Fig. 6.8. Error rate vs. complexity

numbers of unclassified spikes. All versions of the segmented PCA are much more complex than IT yielding only marginal accuracy improvements. The IT algorithm clearly constitutes the knee point of the graph.

6.4 Detection and Alignment Algorithms

Spike detection and alignment (D&A) clips the signal around a suspected spike before it is sorted. Accurate D&A is essential for successful spike sorting, as sorting is sensitive to spike position in the frame (as are all the algorithms described above).

We verify several D&A algorithms [104, 110] with respect to their impact on sorting: a good algorithm aligns spikes so that only a small number of classification errors are caused at the subsequent sorter. Again, we compare our algorithms to the state-of-the-art commercial algorithm described in [109]. Figure 6.9 describes the validation scheme. All algorithms detect spikes by

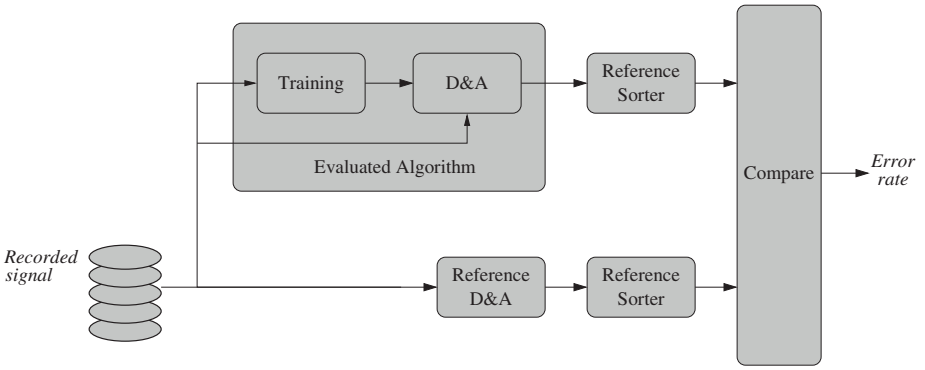


Fig. 6.9. D&A algorithms validation scheme

threshold crossing. The initial frame of M samples is generated, with K samples preceeding and N samples succeeding the threshold crossing event. The alignment is to determine the offset $i_0 \in \{1..K\}$ so that N samples starting from i_0 are passed to spike sorting.

6.4.1 Algorithms Verified

The reference algorithm (denoted as *PCA D&A*) computes principal component representation of the N-point frame for every $i \in \{0..K\}$, based on the first two principal components. The offset selected is the one that yields the smallest energy of error between the signal and its representation by the first two principal components. The required computation is lengthy: for every inspected offset, one must compute projections on the two PCs ($2N$ additions and $2N$ multiplications). Then compute the representation in PC space: $2N$ multiplications and N additions. Then subtract it from the signal (N additions). Then calculate the energy of the difference, which is another N additions and N multiplications. The whole computation takes $5N$ multiplications and $5N$ additions for every offset, giving complexity index of $K \cdot 55N$.

To reduce the computation burden, we have tried to align spikes with respect to the maximum correlation point of the incoming signal with the first principal component (*Maximum Projection Alignment, MPA*). We take advantage of the observation that this correlation typically exhibits a single maximum near a spike. The calculation is much simpler: for every offset inspected, one must compute only a projection of the signal on the first principal component, requiring N additions and N multiplications. The complexity index is therefore $K \cdot 11N$.

As in spike sorting PCA, we also tried to simplify PCA D&A by downsampling the signal and the principal components (i.e. *Segmented PCA D&A*). Downsampling the signal costs N additions. Assuming we use S segments, the computation with downsampled signal requires $5S$ additions and $5S$ multiplications, similarly to the regular PCA. The complexity index is therefore $K \cdot (55S + N)$.

Similarly to sorting algorithms, we tried to replace principal component shape-space generation (that requires heavy computation of PC projections) by a much lighter integral transform. Its major advantage is the lack of multiplications. In the same way MPA D&A seeks the maximal correlation with the first principal component, *Maximum Integral Transform Alignment* searches for the maximum correlation with the rectangular window over the first time interval. As in spike sorting MITA can be viewed as MPA with the first principle component degenerated to a rectangular window. The computation involves summation of samples in a moving window. When implemented efficiently, it can be done with $2N + K$ additions.

6.4.2 Validation Results

Error rates of the software sorter under various D&A algorithms are summarized in Table 6.2. We have also examined the performance of the sorter when no alignment is done, i.e. spikes are aligned to the point of threshold crossing. Simulations were performed with N of 200 and K of 50, in accordance with the reference algorithm. As with sorting, integral transform gives small accuracy

Table 6.2. Validation results summary

Algorithm	Additions	Multiplications	Complexity	Error rate (%)
Threshold	50	0	50	9.4
MITA	450	0	450	1.2
PC7	12,000	1750	30,000	0.7
MPA	10,000	10,000	110,000	0.3
PCA	50,000	50,000	550,000	0.0

penalty with a very strong reduction of computation complexity, constituting the knee point of the accuracy-complexity graph.

7

MEA on Chip: In-Vitro Neuronal Interfaces

Multielectrode arrays (MEAs) are the principal instruments for non-invasive interfaces with biological neural networks (Chap. 2). MEAs allow multichannel recording and stimulation with neuronal networks cultured for several months. Planar arrangement of electrodes upon the substrate defined by photolitographic methods potentially enables high spatial recording resolution with very large numbers of electrodes. Unlike penetrating probes, electrodes of MEAs need not be manually positioned inside the brain tissue one-by-one.

Similarly to penetrating probes, MEA must provide bio-compatible electrodes for tissue interfacing. To allow sensing of microvolt signal levels, it must be coupled to a low-noise signal conditioning front-end. A processing unit can acquire the signal and originate stimuli back to the tissue. A temperature controller must be included within a cultured network recording setup, to maintain the neurons at a/the constant temperature of 36°.

MEAs on glass substrate typically integrate recording electrodes connected by wires to contact pads on MEA perimeter Fig. 7.1. Contact pads are typ-

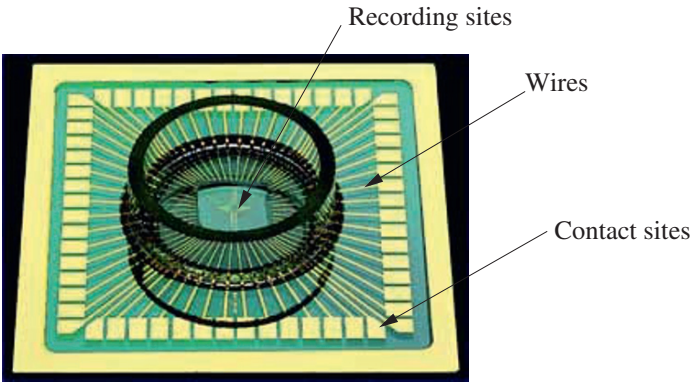


Fig. 7.1. Commercial MEA (Multichannel systems)

ically much larger than the recording sites, as they are clipped to recording front-end equipment. The breakout from the compact 2D structure of small recording sites into 1D structure of contact pads limits the number of channels the MEA can provide in practice.

Silicon multielectrode arrays, or *Multielectrode Chips* (MECs) allow integration of electronic circuits on the same substrate with electrodes (Fig. 7.2). The breakout towards the perimeter is relaxed, potentially allowing for a larger number of electrodes. In addition MEC provides a better system integration

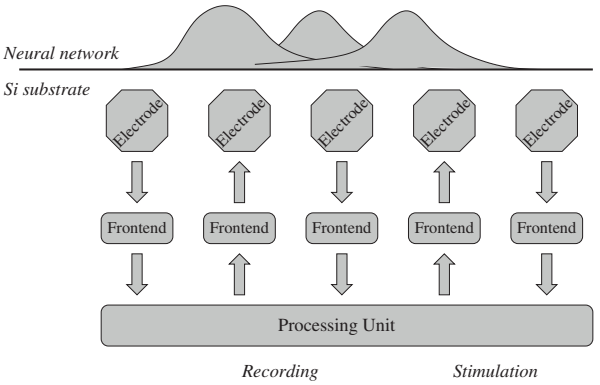


Fig. 7.2. Multielectrode chip concept

and enables on-die processing for data compression and stimulation feedback.

Integrated circuits typically use aluminum or copper for metal interconnects. Both are subject to fast corrosion when put into biological medium. Reported integrated circuits for multielectrode recording fabricated in standard CMOS processes [47, 11, 48, 50] involve various proprietary post-fabrication steps for electrode definition:

- [50]: gold deposited on top of aluminum contacts.
- [48]: areas of field oxide are etched to expose polysilicon electrodes.
- [11]: stack of Ti/Pt and dielectric layer deposited on top of the electrodes.
- [47]: Pt deposited on top of recording sites.

Such processing steps are hardly a commodity: they require special facilities and complicate MEC fabrication. Lithography is probably the most problematic, as it requires expensive mask sets.

In our work [111] we have addressed the development of standard CMOS MEC, involving no or only simple post-fabrication processing. The following summarizes the work:

- Design of a prototype sensor circuit
- Fabrication in standard CMOS process
- Testing of isolation and sealing techniques for MEC

- Testing and modeling of bare aluminum sensing sites
- Investigation of various post-processing steps

Post-processing steps under investigation included coating the MEC with thin-layers of various dielectrics. Such coating does not require lithography, as it can be applied on the whole die surface non-selectively (as long as the bondpads are protected).

7.1 Prototype Sensor

The prototype sensor addresses the following tasks:

- Interface a biological matter with aluminum microelectrodes
- Amplify extracellular potentials
- Allow comparison of different electrode sizes
- Allow characterization of different electrode coatings
- Provides basic temperature control

The chip includes thirty differential amplification channels coupled to electrodes of different sizes distributed over the chip surface. Amplifiers provide gain of 40 dB sufficient to drive the signals off the chip without degrading SNR. Unity gain stages are used to drive off-chip loads. A temperature sensor for detection of 37° level and a special resistor for tissue heating were also included on the chip (Fig. 7.3).

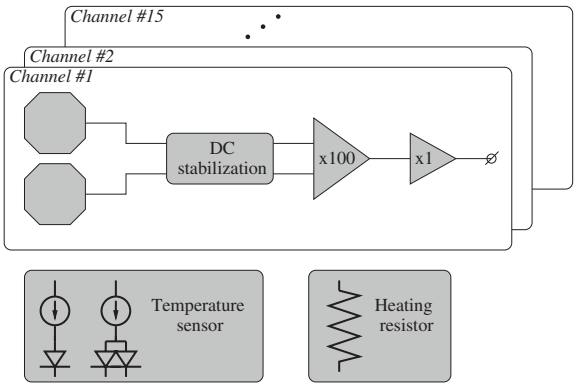


Fig. 7.3. MEC blocks

7.1.1 Electrode Design

Electrodes were designed as areas of exposed top metal interconnect layer. Passivation (thick oxide, about 2.5 μm width) above electrode sites was etched

by a standard process step, similarly as it is etched above bonding pad openings. Conceptional chip layer stack is shown in Fig. 7.4. Process design rules demand arrays of vias placed beneath exposed top metal, to improve mechanical stability of the bond pads. Vias contribute to the roughness of the top metal surface, degrading the quality of coating deposition. Since the electrodes are not bonded, one might consider removing via arrays from beneath the electrodes. We have decided to keep the vias, fearing fabrication reliability problems. Aluminum, being a non-noble metal, forms a thin (several nm)

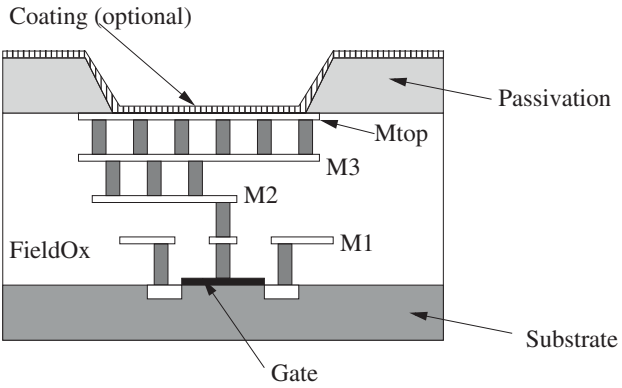


Fig. 7.4. Layer stack in MEC

layer of oxide when introduced into a biological medium. Due to the oxide barrier, a capacitive interface is formed. For 4 nm oxide thickness and dielectric coefficient of 9, area capacitance of $2\,\mu\mathrm{F/cm}^2$ is formed. For an electrode of $100\,\mu\mathrm{m}^2$, this would be 2 pF, quite sufficient for preamplifier coupling.

As shown in [111], spontaneous oxide barrier may fail to protect aluminum electrodes from corrosion. Additional dielectric layer (Al_2O_3, Ta_2O_5) can be deposited on the chip for electrode protection. Dielectric layer deposition is much less complicated than noble metal deposition, as it can be deposited over the entire chip area. Unlike noble metal, dielectric layer requires no high definition lithography and can be made with primitive mask sets for bonding pad protection. Dielectric layer, however, degrades the interface capacitance: for 40 nm deposition with dielectric constant of 20, area capacitance is only $0.44\,\mu/\mathrm{cm}^2$, 440 fF for $100\,\mu\mathrm{m}^2$ electrode, making amplifier design a challenge.

7.1.2 Low Noise Amplifier

Low noise preamplifier designed for `NPRO2` (Chap. 4) was employed for the prototype in-vitro sensor. It was found highly suitable for the in-vitro sensor due to its low input capacitance. Input capacitance of a general differential pair (Fig. 7.5) is defined by gate-source and gate-drain capacitors of the input

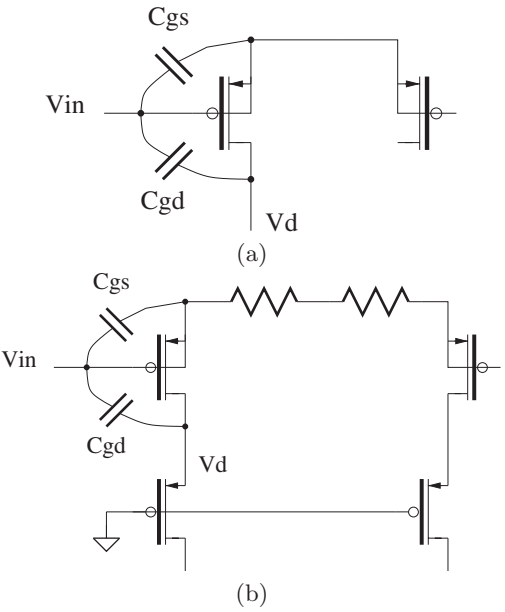


Fig. 7.5. (a) General differential pair. (b) Preamp circuit

transistor. The gate-drain capacitance exhibits Miller effect; it is effectively multiplied by the stage gain, i.e. V_d/V_{in}. In the preamplifier circuit, the small-signal drain voltage of the input transistors, V_d, is effectively zero due to the cascode transistor, thus the Miller effect is eliminated. Because of source degeneration, gate-source voltage is lower by a factor of $(g_m R_s + 1)^{-1}$, effectively lowering the gate-source capacitor by the same factor. Input capacitance of 850fF was simulated for the preamplifier. That can be compared, for example, to the circuit used in `NPR03`, having input capacitance of 5 pF.

7.1.3 Input DC stabilization

DC input stabilization was achieved with periodically shortened reset gates. Although continuous bias technique was successfully used in `NPR03`, reset gates were adopted for simplicity of modeling and implementation. Reset gate biasing suffers from several serious drawbacks: Periodic artifacts are introduced into the signal because of reset pulses. While the reset gates are opened, input potentials slew slowly due to leakage currents in reset transistor diffusions. Mismatches in these leakage currents generate potential differences in electrodes across the chip, contributing to corrosion. Continuous bias technique would enforce constant and equal potentials over all the electrodes potentially leading to better corrosion immunity.

High-impedance input node capacitance characterization was made possible by signal injection through an RC voltage divider. An array of capacitors

of known sizes can be selectively connected to the input node, to allow several independent measurements. Since there are several unknown capacitors connected to the input node, several equation have to be formed to find all the unknowns. The input node with characterization circuits is shown in Fig. 7.6.

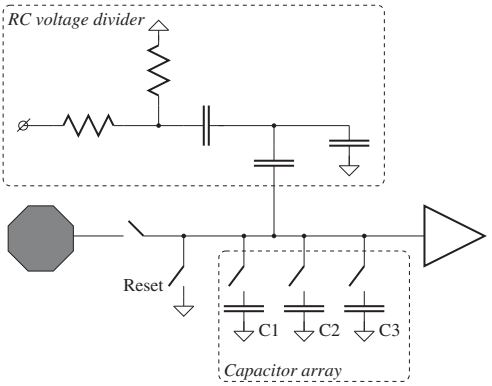


Fig. 7.6. Input node capacitance characterization

7.2 Temperature Sensor and Heater

A bandgap temperature sensor was implemented on chip, based on bipolar elements. The circuit is described in [111]. A 5 W heating element was implemented as a polysilicon resistor. As the process specifies poly resistor operation up to 16 V, the resistance of the heater must be about 50 Ω. To meet maximal current density specifications, the resistor must be made wide enough. Special care must also be taken to make sure the heat is distributed equally over the resistor length: poly turns must be avoided, and solid wire connection with large numbers of vias must be used. Maximal bond wire current specification must also be met: A total of twelve bond pads was allocated for heater supply.

7.3 Post-Processing and Bath Formation

Post processing involves coating the electrodes with dielectric layers to protect them from corrosion and prepare them for interaction with biological culture. A bath has to be formed on the packaged chip to contain the biological solution and protect the chip bonding pads and bond wires from interacting with it.

7.3.1 Post Processing

Deposition of thin dielectric layers was tested on unpackaged dies, using a crude metal mask for protection of pad areas. The layers were evaporated in an electronic beam evaporator. The dies obtained this way were not tested due to unavailability of a reliable bonding service. Evaporation was also made on factory-packaged chips, as the ceramic packages could withstand the process. Packaged chips do not require bonding mask. Figure 7.7 shows a processed packaged chip. Various technical aspects of the process are discussed in [111].

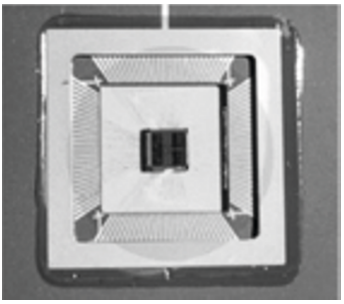


Fig. 7.7. Packaged chip with evaporated dielectric. The circle showing evaporation area is clearly visible

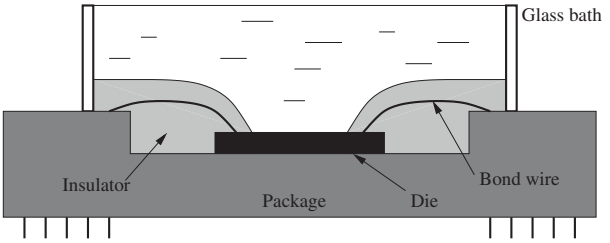
7.3.2 Culture Bath Formation

To contain the biological solution a glass bath was mounted on top of the chip package (Fig. 7.8). While the thick field oxide layer protects most of the chip area from contacting the solution, pad areas and bond wires remain open. A special isolation layer must be deposited for pads and bond wires protection.

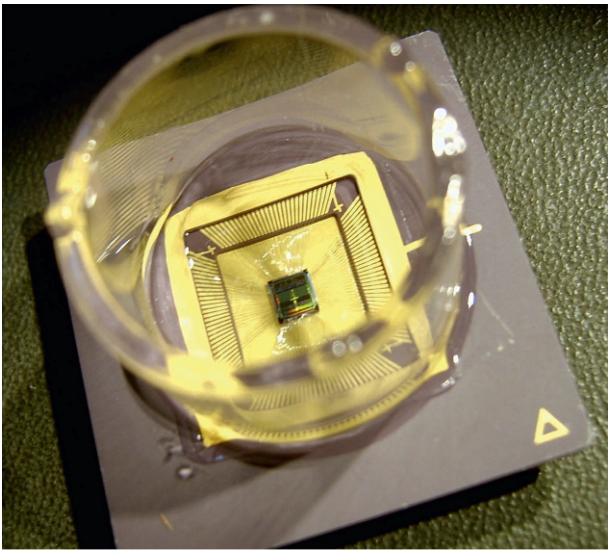
Two different epoxy types were tried for isolation. Epoxy-isolated chips failed, due to small cracks that appeared in the epoxy isolator causing opens at bonding sites (Fig. 7.9).

To prevent cracking, the epoxy was replaced by Sylgard-184 (Fig. 7.10). It is a low-viscosity material, widely used in bio-applications for electrical isolation, thanks to good biocompatibility. It was successfully used with six chips for up to six weeks of operation in a salty solution.

While Sylgard can provide the required isolation quality, the process of applying it to the chips is problematic, as it is fully manual. Extra care has to be exercised not to damage the wire-bonds. Isolation is, consequently, time-consuming and is not expected to have a good yield. One way to improve the situation is to change the chip layout: move all the pads to one side, far



(a)



(b)

Fig. 7.8. (a) Bath formation bond isolation. (b) Isolated chip with bath

from the active chip area (i.e. area with electrodes). This would make isolator application much easier.

Preferably, some automatic technique must be developed, especially if volume fabrication is desired. One existing technology to try is flip-chip bonding. The die has to be mounted to the PCB, provided that PCB has a hole, so that chip electrodes remain uncovered (Fig. 7.11).

7.3.3 Electrode Characterization

Electrode impedance was measured by injecting small-amplitude sine waves through resistors of known value (Fig. 7.12). Interface capacitance of 30–40 pF was measured with an aluminum electrode. 1.5 pF was measured with a coated electrode (Al_2O_3, TiO_2 stack of 110 nm). Both electrodes have a very large resistance (above 100 MΩ).

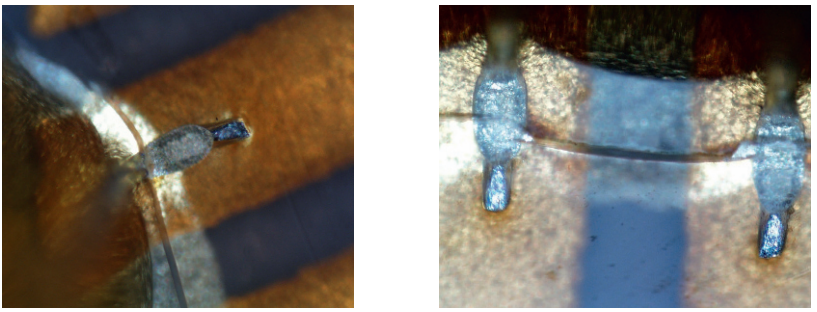


Fig. 7.9. Cracks in epoxy isolation

Fig. 7.10. Chip with applied Sylgard protection

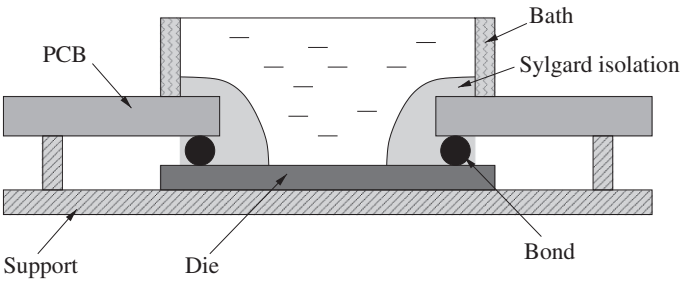


Fig. 7.11. Chip-on-board mounting of MEC

It was noted that electrodes suffer corrosion when placed in physiological medium and the chip supply is turned on. Corrosion was much slower when the supply was disconnected. The believed reason is in differences of electrode potential due to non-uniform leakage currents. Spontaneous oxide formed on the electrode surface does not provide sufficient protection: when an external potential of 1V is applied, the corrosion is fast enough to eliminate the electrode completely within seconds. Larger electrodes tend to degrade faster than the smaller ones.

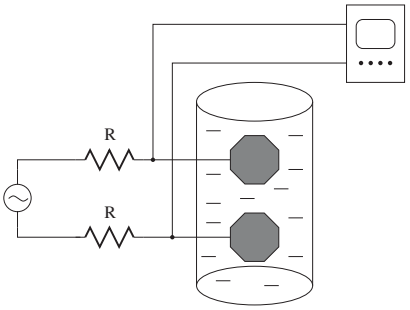


Fig. 7.12. Electrode impedance measurement

Coating with a dielectric layer improved chemical stability. Coated electrode could withstand potentials of volts. However, the success of protection varied significantly from one electrode to another (Fig. 7.13). As with plain Aluminum, larger electrodes degraded faster than smaller ones.

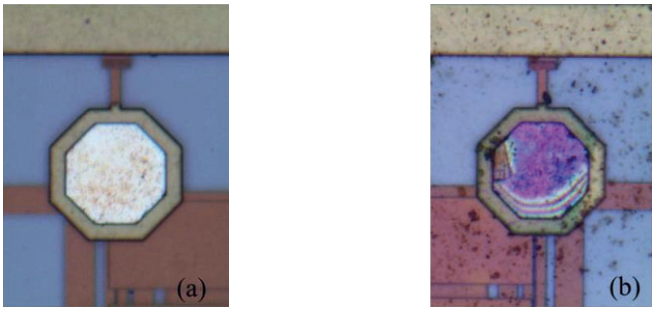


Fig. 7.13. (a) Coated electrode after 3V potential. (b) Identical electrode broken down at 2.5V potential

The non-uniformity of the results can be explained by roughness of the exposed top-metal area. It is caused in part by via arrays that are required under exposed top metal by the design rules (Fig. 7.14). Another probable cause are surface artifacts. This also explains why larger electrodes tend to degrade faster: the probability for a surface artifact grows linearly with area.

7.3.4 Culturing neural cells

Rat cells have been cultured on top of the isolated chips (Fig. 7.15). The same procedure that is routinely used for culturing cells on top of MEAs was employed with the chips. They were held inside an incubator for six weeks; the electrical activity was checked every two weeks. The recorded electrical

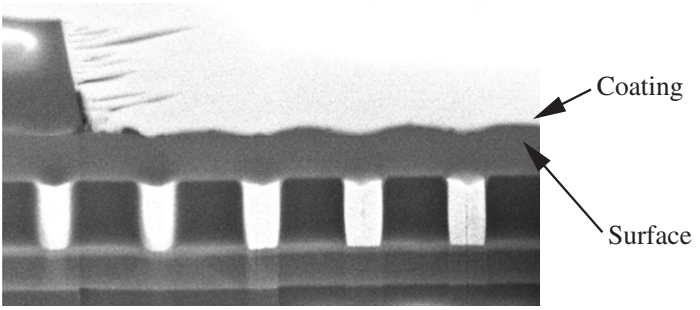


Fig. 7.14. SEM image of an electrode showing non-flat surface beneath the coating

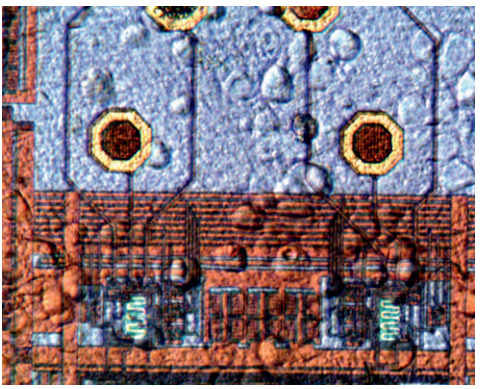


Fig. 7.15. Neuron cell cultured on top of MEC

activity remained stable, except for some degradation in spike amplitudes towards the sixth week. This can be related to degradation of the aluminum electrodes. Examples of the recorded spikes are shown in Fig. 7.16.

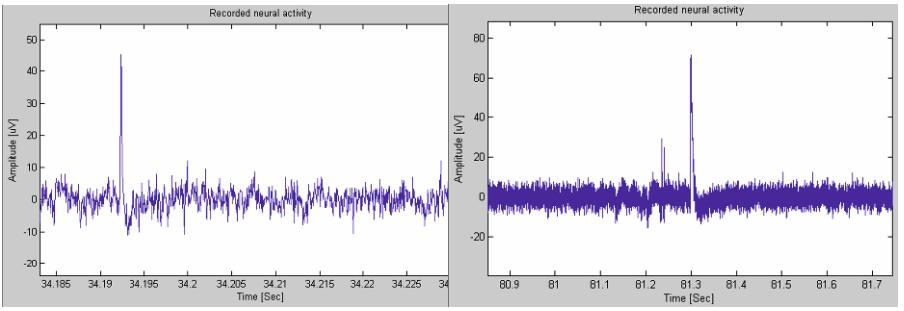


Fig. 7.16. Samples of recorded neural spikes

7.4 Conclusions and Future Work

A 0.35 μm prototype chip for in-vitro interfacing biological neural networks was designed and fabricated. The chip included aluminum electrodes, active circuitry for amplifying the recorded signals, circuits for electrode characterization and temperature control. A simple step for electrode protection by a thin isolator layer, requiring no lithography, was also considered. The chips were successfully isolated with Sylgard-184. Neurons were cultured on top of the chips for six weeks with stable electrical activity recordings.

Aluminum electrodes can potentially be used for neural sensing. They do not seem to have any toxic effects on the tissue. Corrosion prevention is of extreme importance. Thin-film isolation layer is an option, but work has to be done to improve the quality of isolation, without employing complicated post-processing steps, such as CMP. Perhaps eliminating vias under exposed top-metal areas can improve the surface roughness.

A robust, preferably automatic procedure for isolation of the bond pads and wires has to be developed. The current procedure of applying Sylgrad by hand yields working chips, but is extremely time-consuming. The effort can be substantially reduced by placing all the bond pads on one side of the chip, far from the active area. This, however will constrain the number of pins available for chip interfacing. Appropriate architectural solutions are to be developed for internal multiplexing of recorded signals to allow streaming over a narrow interface.

8

Conclusions

8.1 Research Contributions

The research presented in this book aimed at architectures for neuronal interfaces. The concept of the *Neuroprocessor — a computational neuronal interface*, the main contribution of this work, has been reviewed in detail. Unlike other existing neuronal interfaces providing for mere amplification, shaping and transduction of neuronal signals, the Neuroprocessor integrates computational capabilities at the recording (or stimulating) front-end. These capabilities are used for reduction of recorded data, so that only small portions of the raw datarate have to be communicated outside the Neuroprocessor. Thus, it enables low power operation of neuronal interfaces, such as required at miniature implants.

8.1.1 Integrated Neuronal Recording Front-End Circuits

Design considerations and test results of three generations of recording front-end integrated circuits, `NPR01` -`NPR03` , are shown. `NPR03` is a fully-integrated, 12-channel data acquisition system for neuronal signal recording, including low-noise amplification, band-splitting of neuronal signals, integrated per-channel ADC and numerous programmable features (including offset calibration). The channels communicate over an internal parallel bus with an integrated controller, capable of streaming the sampled data from all the channels on a serial communication link. `NPR03` included also a novel, robust scheme for DC stabilization of neuronal preamplifier input that was developed, implemented and successfully tested.

A miniature headstage employing `NPR03` is presented. The design of FPGA-based embedded computer board with real-time software accompanying the headstage is also described. The system was successfully tested in both in-vitro and in-vivo neuronal recording experiments, with the summary of the results presented in this book.

The design of a novel, fully differential neuronal preamplifier was presented, with detailed analytical derivation of the noise-power tradeoff optimum point, confirmed by numerical analysis and transistor-level circuit simulation. The preamplifier provides for an improved noise performance by replacing active loads at the input stage with resistors. Despite that, it suffers no gain deterioration thanks to a gain-boosting second stage, that is shown to draw only a small portion of the consumed current.

8.1.2 Low Power Algorithms for Spike Sorting and Detection

The development of several low-power algorithms for spike detection and sorting is shown. The algorithms were aimed at implementation on a special-purpose hardware. This is in contrast with the majority of known algorithms that are software-based. The algorithms were designed as modifications of algorithms known in the field with a goal of trading off a small percentage of computation accuracy for large savings in computational complexity. They were verified by numerical simulations on real neuronal recordings and compared with a state-of-the-art algorithm, chosen as a "golden-rule". It is shown that several percent of accuracy reduction can lead to orders of magnitude savings in computational complexity, and consequently in power consumed by the electronic circuits.

8.1.3 In-Vitro Neuronal Interfaces

The design of a Multielectrode array for in-vitro culturing of neuronal tissues on a CMOS chip is presented, together with encapsulation procedures for bond-pad protection. The fabrication requires no processing steps beyond the standard CMOS technology, enabling inexpensive and rapid mass-fabrication of such devices. The electrode sites are formed by removing the passivation above the top Aluminum layer. It was determined experimentally that neurons show no signs of intoxication when exposed to the Aluminum electrodes. It was also shown that Aluminum electrodes allow for several weeks of stable recording, but deteriorate rapidly when stimulation potentials are applied. A simple post-processing step of thin-film dielectric coating was applied for electrode protection. It was a non-selective coating, requiring no photo-lithography, the most expensive part in microfabrication. The electrodes demonstrated far better corrosion immunity, although poor quality of deposited thin film layers did not allow for an extensive study.

8.2 Future Work

8.2.1 Neuroprocessors

Further investigation of architecture aspects of the neuronal front-end chips should be made. In particular, there are several architectural questions left

out of this book. These points allow potential power optimizations in the recording front-end by learning the properties of the signals recorded:

- Analog vs. digital processing. Integrated circuits described in this book acquire the input signals continuously and convert them into digital streams. Data reduction algorithms are executed in the digital domain, requiring continuous operation of ADCs. This paradigm may be altered if spike detection is carried out in the analog domain (e.g., threshold detection by a comparator). The ADCs, thus, can be turned on only upon the detection of a spike, potentially saving power.
- Another potential for power saving lies with powering the LNA, the dominant consumer of power. It might be possible to lower its bias current (at the cost of higher noise) during inter-spike intervals, which can be adaptively determined.
- Yet more power can be saved within the LNA: if the spectrum of the spikes at every channel can be learned, a matching filter (analog or digital) can be constructed to emphasize the relevant frequency band and suppress the irrelevant bands. Thus, the noise seen within the input signal is reduced, and the LNA current can be lowered.

More research may be performed on implementation of the algorithms. Most importantly, implementations need to be evaluated regarding the power they require, raising among others the following questions: should the computations be performed by a central processor or distributed over the input channels; how the memory is to be implemented (central or distributed) and how the data get communicated within the Neuroprocessor (synchronous or asynchronous, serial or parallel, shared bus or dedicated interconnect, etc.).

A variety of circuit and architectural problems arise when stimulation is integrated with the Neuroprocessor, as it may require relatively high voltage or current levels. Fully-integrated, power-efficient voltage converters probably have to be designed for Neuroprocessors.

With Neuroprocessor dies available, the issue of bio-compatible packaging and implant encapsulation should also be considered. One of the most important problems is powering the implant: should it be done remotely (i.e., RF power), locally (battery cell) or combined (battery cell recharged by RF).

8.2.2 In-Vitro Recording

Studies of in-vitro recordings can be divided into processing, encapsulation and architecture research: Robust and reliable processing steps are to be developed for electrode protection. The current problem with thin-film dielectric deposition is chip surface roughness, caused in part by passivation openings above the electrodes. Perhaps partial thinning of the passivation layer can improve the situation.

The encapsulation procedure used in this work, although operational, required much handcraft and is hardly robust. Improvement can be achieved

with better layout configurations, where electrodes are placed on one side of the die while all the bonding pads are concentrated on the opposite side at a large distance.

When the channel count of an MEC increases, a problem of signal readout arises. Since it is infeasible to provide a wire to every electrode with large electrode counts, some sort of channel multiplexing is in order. With low-level signals involved, multiplexing at required sampling rates is non-trivial and has to be investigated. After that, the problem of data reduction (resembling the one in Neuroprocessors) has to be addressed: Although the MEC does not operate wirelessly, the number of channels it may incorporate is orders of magnitude larger than that of the in-vivo Neuroprocessor, making raw streaming of all the acquired signals impossible.

Appendix A

NPR02 Technical Details

A.1 NPR02 Preamp Sizing

A.1.1 Gain Deviation

First, we examine the expected preamp gain variation.

$$\begin{aligned} A &= \frac{R}{r + r_m} \\ \Rightarrow \frac{\Delta A}{A} &= -\frac{\Delta r_m}{r_m} \frac{r_m}{r + r_m} \\ &= \frac{r_m}{r_m + r} \left(\frac{\Delta I}{I} - \frac{\Delta \eta}{\eta} \right) \end{aligned}$$

Both $\Delta\eta/\eta$ is easily estimated with statistical simulations as a relative variation in transistor transconductance under constant current. Its standard deviation (for large transistors) was found to be around 2%.

$$I = \frac{Vdd - V_{gs3}}{R}$$

If we choose `M3` large enough, V_{gs3} is roughly equal to its threshold voltage. And relative current variation can be calculated as:

$$\frac{\Delta I}{I} = -\frac{\Delta V_T}{R} = -\frac{\Delta V_T}{V_T} \frac{V_T}{V_{dd} - V_T}$$

Thus,

$$\frac{\Delta A}{A} = -\frac{r_m}{r + r_m} \left(\frac{\Delta V_T}{V_T} \frac{V_T}{V_{dd} - V_T} + \frac{\Delta \eta}{\eta} \right)$$

Threshold voltage relative variation is about 10%. Substituting V_{dd} of 3.3 V and nominal threshold voltage of 0.7 V gives:

$$\frac{\sqrt{\overline{\Delta A^2}}}{A} = \frac{r_m}{r_m + r}\sqrt{\left(0.1\frac{0.7}{3.3-0.7}\right)^2 + (0.02)^2} = 0.03 \cdot \frac{r_m}{r + r_m}$$

Standard deviation of the gain is aimed at 1–1.5%. Thus a condition upon r and r_m can be derived, r_m must be placed between $r/2$ and r. r_m was roughly placed at $2r/3$.

A.1.2 Preamp Noise

The output current noise contribution of the thermal noise of M1 and M2 (Fig. 4.8), neglecting the contribution of the source-bulk transconductance $g_{m,bs}$ can be calculated as:

$$S_{i1,2} = 2 \cdot \frac{8}{3} g_m \left(\frac{r_m}{r + r_m}\right)^2$$

If we denote the ratio r/r_m by α, we can rewrite the above as:

$$S_{i1,2} = \frac{16}{3}\frac{g_m}{(1+\alpha)^2}$$

The contribution of degeneration resistors r is:

$$S_{ir} = 2 \cdot 4KT \cdot g \left(\frac{r}{r + r_m}\right)^2$$

can be rewritten as:

$$S_{ir} = 8KT\frac{\alpha g_m}{(1+\alpha)^2}$$

The contribution of M5 and M6 is:

$$S_{i5,6} = \frac{16}{3} KT g_{m5}$$

and the sum of the above components is:

$$S_i = \frac{16}{3} KT \left(g_m \frac{3/2\alpha + 1}{(1+\alpha)^2} + g_{m5}\right)$$

Knowing that the input transconductance is $(r + r_m)^{-1}$, we can reflect the above to input as

$$\begin{aligned} S_{Vin} &= \frac{16}{3} KT \left(g_m \frac{3/2\alpha + 1}{(1+\alpha)^2} + g_{m5}\right) \cdot (r + r_m)^2 \\ &= \frac{16}{3} KT \left(g_m \frac{3/2\alpha + 1}{(1+\alpha)^2} + g_{m5}\right) \cdot r_m^2 (1+\alpha)^2 \\ &= \frac{16}{3} KT \left(r_m(3/2\alpha + 1) + r_m^2 g_{m5}(1+\alpha)^2\right) \end{aligned}$$

Substituting r_m and $g_{m,5}$ we obtain:

$$S_{Vin} = \frac{16}{3} KT \frac{\eta V_{th}}{I} \left((1 + 3/2\alpha) + (1 + \alpha)^2 \frac{\eta V_{th}}{I} \sqrt{2k_5 I} \right)$$
$$= \frac{16}{3} KT \frac{\eta V_{th}}{I} \left((1 + 3/2\alpha) + (1 + \alpha)^2 \eta V_{th} \sqrt{\frac{2k_5}{I}} \right)$$

The associated input noise energy over bandwidth f_b can be calculated as:

$$\overline{v_n{}^2} = f_b \frac{16}{3} KT \frac{\eta V_{th}}{I} \left((1 + 3/2\alpha) + (1 + \alpha)^2 \eta V_{th} \sqrt{\frac{2k_5}{I}} \right)$$

For a given noise voltage and W_5/L_5, this equation can be solved iteratively for I. In Fig. A.1 the solution is shown for α of 3/2, f_b of 10 kHz and v_n of 2 μV. Limiting the overdrive voltage to some 200 mV, we obtain W_5/L_5 of

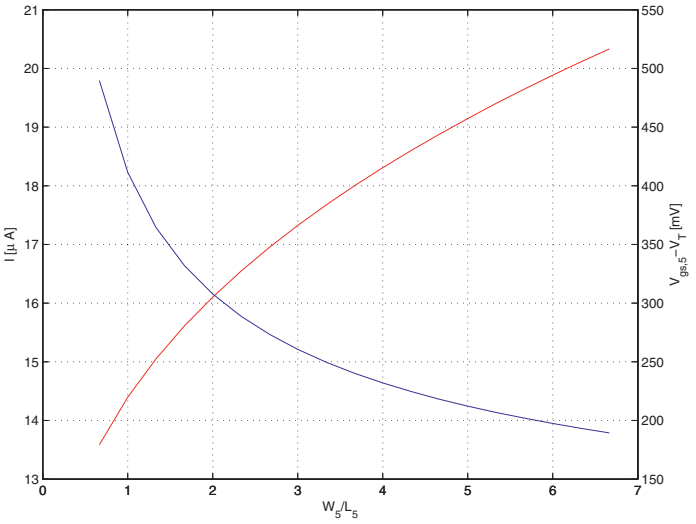


Fig. A.1. M5 current and overdrive voltage as function of W_5/L_5

5 and I of 20 μA. The actual value to be taken for I is higher, since this calculation did not take into account the effects of $g_{m,bs}$, noise contributions of PMOS current sources. Also, v_n must be set somewhat smaller than the goal 2 μV, to allocate some space to $1/f$ noise. I was set to 25 μA. This sets r_m to about 2 kΩ and r to about 3 kΩ. The input transconductance is therefore 200 $\mu\Omega^{-1}$.

$1/f$ noise of the circuit is dominated by M5 and M6. The output noise current density is:

$$S_{1/f} = \frac{k_F I^{A_F}}{C_{ox} L^2 \cdot f} = \frac{\beta_L}{f}$$

The spike band is limited by a single pole high pass filter at the low-frequency side with pole at f_1. At the high frequency side the band is limited at f_b and we assume that the rolloff is steep. Output current noise energy introduced into the spike band by $1/f$ is:

$$\begin{aligned}
\overline{i_{n,1/f}{}^2} &= \int_0^{f_b} S_{1/f} \frac{(f/f_1)^2}{1+(f/f_1)^2} df \\
&= \beta_L \int_0^{f_b} \frac{f/f_1^2}{1+(f/f_1)^2} df \\
&= \frac{\beta_L}{2} \left[\ln\left(1+\frac{f^2}{f_1^2}\right) \right]_0^{f_b} \\
&= \frac{\beta_L}{2} \ln\left(1+\frac{f_b^2}{f_1^2}\right)
\end{aligned}$$

Since f_b is about 10 kHz and f_1 is hundreds of Herz, this can be further simplified:

$$\overline{i_{n,1/f}{}^2} = \beta_L \ln \frac{f_b}{f_1}$$

The thermal input noise voltage is about $2\,\mu$V . Thus corresponding output current is $2\,\mu V \cdot 200\,\mu\Omega^{-1} = 400\,pA$. The goal is to keep thermal noise dominating over the spike band; we require the energy of $1/f$ noise to be much lower than that of thermal noise, for example:

$$\overline{i_{n,1/f}{}^2} = \frac{1}{4} \cdot \overline{i_{n,th}{}^2}$$

Thus we can write:

$$\beta_L \ln \frac{f_b}{f_1} = \left(\frac{400 \cdot 10^{-12}}{4} \right)^2$$

L_n can be calculated to satisfy this equation. The result for f_b of 10 kHz, f_1 of 200 Hz and $1/f$ constants for the process in use is $15\,\mu$m . Thus W_n was set to $80\,\mu$m and L_n to $15\,\mu$m .

The size of M1 and M2 can be chosen rather arbitrarily, as long as the W/L is kept high enough to keep the transistors in weak inversion. Choosing W, L too large is not desirable, as it increases the input capacitance. W was set to $800\,\mu$m and L to $2\,\mu$m.

A.2 NPR02 Testboard Output Channel

The channel schematic is presented in Fig. A.2. The channel consists of a two-stage amplifier and fourth-order Bessel filter. The filter is made of two

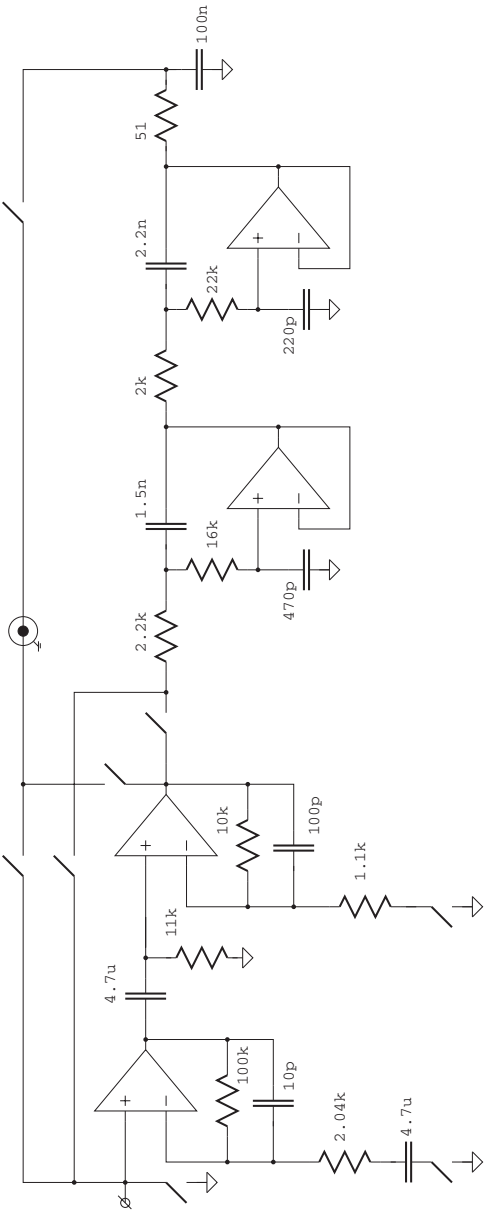


Fig. A.2. An output channel on NPR02 testboard

Sallen-Key biquads. The amplifier gain can be set to a unity, 10, 100 and 1000. Amplifier gain and channel configuration are set with eight digitally-controllable switches (MAX395). The channel can be configured so that the

amplifier, the filter or both may be connected or bypassed. TL074 operational amplifier was used within the channel.[1]

[1] High input impedance and low noise operation limited the operational amplifier choice to those with JFET input devices. It was also desired to find one available in quad packages.

Appendix B

NPR03 Technical Details

B.1 NPR03 Instruction Set

The format of the frames sent by the host to NPR03 (downward frames) is shown in Fig. B.1. The 24-bit frame includes eight bits of opcode mnemon-

Fig. B.1. NPR03 downward frame

ics, eight bits for register identification and eight data bits. The instructions are shown in Table B.1. A fetch/store command can address either a con-

Table B.1. NPR03 instructions

Opcode Mnemonic	Opcode	Address	Data	Action
Fetch	1	Register code	N/A	Fetch register content
Store	2	Register code	Value	Store register content
Start	3	N/A	N/A	Start streaming data
Stop	4	N/A	N/A	Stop streaming data

troller register or a register inside one of the channels. When addressing a channel register, AddrLo must be set to the required channel number (1..12) and AddrHi to the channel register number. To address a controller register, AddrHi must be set to the controller register number and AddrLo must be zero. Register codes are summarized in Table B.2.

The format of NPR03 upward frames is shown in Fig. B.2. The NP3 may send either of the two types of reply frames Table B.3. Replies sent in the idle

Table B.2. NPR03 register codes

Channel registers			Controller registers		
Register	AddrHi	AddrLo	Register	AddrHi	AddrLo
ADC0	1..12	0	CEMR0	13	0
ADC1	1..12	1	CEMR1	14	0
HPF	1..12	2	LTHR	15	0
MSC	1..12	3	HTHR	16	0
SPK	1..12	4	SCNR	17	0
LFP	1..12	5	CDIVCNT	18	0
			CDIVADC	19	0

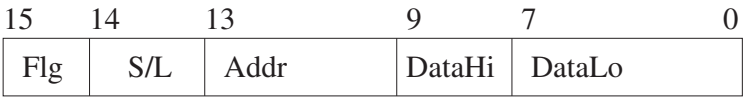


Fig. B.2. NPR03 upward frame

mode have Fl of "0". All idle-mode replies will have S/L field of "0", unless the chip has received an illegal instruction. When a register is fetched (idle mode), the 8-bit value is returned in DataLo. Replies sent in the streaming mode have Fl of "1". S/L of "1" means that a frame carries a sample of an LFP output; S/L of "0" means SPK output. The Addr field designates the channel number. DataHi and DataLo together make for a 10-bit data word.

Table B.3. NPR03 reply frame fields

Reply to	Fl	S/L	Addr	DataHi	DataLo
Fetch	0	0	n/a	n/a	Value
Store	0	0	n/a	n/a	n/a
Error	0	1	n/a	n/a	n/a
Stream	1	1=LFP, 0=SPK	1..12	D9-8	D7-0

B.2 NPR03 Registers

B.2.1 Channel Registers

The registers described below affect the operation of the recording channels. The whole set of these registers resides in every recording channel.

ADC0

RDY	D9	D8	D7	D6	D5	D4	D3

RDY (R/O)	Indicates that the channel ADC has finished a conversion
D9-3 (R/O)	Upper seven bits of the 10-bit conversion result

ADC1

D2	D1	D0	S/L	TST	REQ	n/a	n/a

D2-0 (R/O)	Lower three bits of the 10-bit conversion result
S/L (R/W)	Selects whether the ADC connects to the SPK ('0') or the LFP ('1')
TST (R/W)	Testing. Has no effect on channel operation.
REQ (W/O)	Conversion start request. When '1' is written into REQ, the ADC will start a conversion cycle as soon as it finishes the current conversion cycle

HPF

B3	B2	B1	B0	n/a	n/a	n/a	n/a

B3-0 (R/W)	Sets the cutoff frequency of the DC blocking HPF on the integrated channels

MSC

F2	F1	F0	CPL	NUL	TST	n/a	n/a

F2-0 (R/W)	LPF cutoff frequency at SPK output
CPL (R/W)	Connects the internal 8MΩ resistors to inputs (external-cap channels)
NUL (R/W)	Nulls the input
TST (R/W)	Testing. No effect on channel operation

SPK

OF4	OF3	OF2	OF1	OF0	G1	G0	n/a

OF4-0 (R/W)	SPK channel offset calibration (calibration DAC input)
G1-0 (R/W)	SPK channel gain (VGA gain setting)

LFP

OF4	OF3	OF2	OF1	OF0	G1	G0	n/a

OF4-0 (R/W)	LFP channel offset calibration (calibration DAC input)
G1-0 (R/W)	LFP channel gain (VGA gain setting)

B.2.2 Controller Registers

The registers below reside inside the controller, which can access them directly (not through the internal bus).

CEMR0

R8	R7	R6	R5	R4	R3	R2	R1

R7-0 (R/W)	Setting Ri to '1' enables data streaming from channel i

CEMR1

n/a	n/a	n/a	n/a	R12	R11	R10	R9

R12-9 (R/W)	Setting Ri to '1' enables data streaming from channel i

LTHR

LTHR (R/W) holds the eight most significant bits of the low threshold value. The two least significant bits are always taken as "00". Low threshold value is therefore LTHR*4.

HTHR

HTHR (R/W) holds the eight most significant bits of the high threshold value. The two least significant bits are always taken as "00". High threshold value is therefore HTHR*4.

SCNR

SCNR (R/W) holds the number of samples to be streamed out upon threshold crossing detection.

CDIVCNT

H3	H2	H1	H0	L3	L2	L1	L0

CDIVCNT (R/W) holds two four bit values, H3-0 and L3-0 that specify how the McBSP clock is divided to obtain the controller clock and the bus clock. H3-0 specifies for how many half-cycles of the McBSP clock the controller clock stays high. L3-0 specifies for how many half-cycles of McBSP clock the controller clock stays low. The count is one-based, i.e. a value of zero in H3-0 (or L3-0) means a single half-cycle.

CDIVADC

CDIVADC (R/W) has the same structure as CDIVCNT and similar interpretation. It defines how the controller clock is divided to obtain the clock for ADC operation.

B.3 NPR03 Preamp Sizing

Differential output current noise is contributed by the PMOS input pair and NMOS current sources (Fig. 5.9). Each transistor contributes channel shot noise of the form

$$i_d = \frac{8}{3} kT \left(g_m + g_{mbs} \right)$$

and $1/f$ noise

$$i_{f,pmos} = \frac{k_F I^{a_F}}{C_{ox} W L \cdot f}$$
$$i_{f,nmos} = \frac{k_F I^{a_F}}{C_{ox} L^2 \cdot f}$$

Bulk-source transconductance, g_{mbs} is given by:

$$g_{mbs} = \frac{\gamma}{2\sqrt{2|\Phi_F| - V_{bs}}} \cdot g_m = \xi g_m$$

The coefficient ξ can be calculated to about 0.25 for the process in use, for bulk-source voltages of 0–200 mV. i_d can be rewritten in terms of ξ as

$$i_d = \frac{8}{3} kT(1 + \xi) g_m$$

The total output current noise is therefore:

$$S_{i_{o,n}} = \frac{8}{3} kT(1 + \xi) \left(g_{m,n} + g_{m,p} \right) + \frac{k_{F,p} I^{a_{F,p}}}{C_{ox} W_p L_p \cdot f} + \frac{k_{F,n} I^{a_{F,n}}}{C_{ox} L_n^2 \cdot f}$$

The transfer function from the output current noise to output voltage noise can be shown as:

$$\frac{v_{o,n}}{i_{o,n}} = \frac{s(C_i + C_f + C_x) + G_f}{(sC_f + G_f)(g_m + sC_i)}$$

where C_x is the input capacitance of the amplifier. The noise is suppressed below f_1 by the splitter; unlike the noise injected by G_f, in this case there are no $1/f^2$ components, thus the noise below f_1 (order of 200 Hz) can be neglected. Above 200 Hz, we can neglect G_f with respect to sC_f. Also, g_m is much larger than sC_i for the frequencies of interest:

$$\frac{v_{o,n}}{i_{o,n}} = \frac{C_i + C_f + C_x}{C_f} \cdot \frac{1}{g_m}$$

This can be further reflected to input, multiplying by C_f/C_i:

$$v_{i,n} = i_{o,n} \cdot \frac{C_i + C_f + C_x}{C_i} \cdot \frac{1}{g_m}$$

The total input noise energy can be written after substituting $i_{o,n}$ and integrating from f_1 to f_2 (we approximate also with $f_2 \gg f_1$):

$$\overline{v_{i,n}{}^2} = \left(\frac{C_i + C_f + C_x}{C_i \cdot g_{m,p}}\right)^2 \times \\ \times \left[\frac{8}{3}kT(1+\xi)(g_{m,p} + g_{m,n})f_2 + \ln\frac{f_2}{f_1}\left(\frac{k_{F,p}I^{a_{F,p}}}{C_{ox}W_pL_p} + \frac{k_{F,n}I^{a_{F,n}}}{C_{ox}L_n^2}\right)\right]$$

As $v_{i,n}$ clearly decreases with $g_{m,p}$, we shall try to maximize the latter, thus the input PMOS pair will be in weak-inversion. $g_{m,p}$ dependence on bias current and transistor geometry is most conveniently described by EKV model [95]:

$$g_{m,p} = \frac{\kappa I}{V_{th}} \cdot G(I)$$

where $G(I)$ is given by

$$G(I) = \frac{1 - e^{-\sqrt{I/I_S}}}{\sqrt{I/I_S}}$$

κ is a process parameter, estimated to about 0.85. I_S is given by

$$I_S = \frac{2\mu_{0,p}C_{ox}V_{th}^2}{\kappa} \cdot \frac{W_p}{L_p} = 94n\mathrm{A} \cdot \frac{W_p}{L_p}$$

C_x is a gate capacitance of the input transistor, proportional to W_pL_p. Since $v_{i,n}$ increases with $g_{m,n}$, we shall place the NMOS sources in saturation, with large L_n and small W_n. With several simple substitutions we can write $g_{m,n}$ as a function of current and overdrive voltage:

$$g_{m,n} = \frac{2I}{V_{ovd,n}}$$

We limit the overdrive voltage on the NMOS transistors to 300 mV. With the above expressions for $g_{m,n}$, $g_{m,p}$ and C_x, $v_{i,n}$ is fully given with four circuit parameters: W_p, L_p, L_n and I. $v_{i,n}$ is clearly inversely dependent on L_n, and directly dependent on L_p, thus we shall place L_n at the largest possible value for layout convenience (10 μm), and we shall set L_n to a small value (1μm). Figure B.3 shows the contour plot of $v_{i,n}$ as a function of I and W_p. Figure B.3

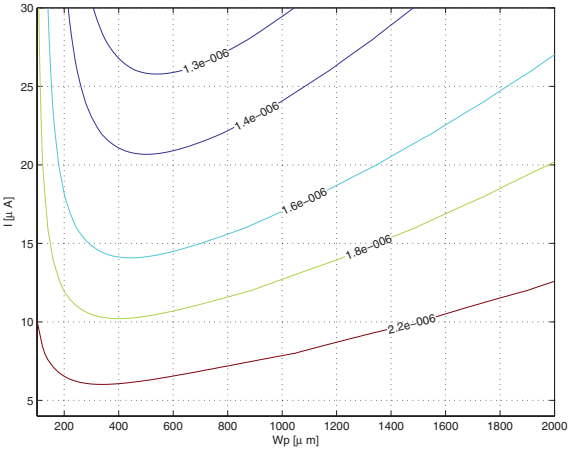


Fig. B.3. Amplifier noise contribution contour plot

shows that there is an optimal W_p for a given I. This is because $g_{m,p}$ increases with W_p, lowering the noise, but C_x increases with W_p making it worse. Trying lower L_n yields only insignificant savings in bias current, some 10% when L_n is moved from 1 μm to 0.5 μm .

The actual sizing chosen: W_p of 400 μm, L_p of 1 μm and I of 25 μA, making the amplifier noise contribution about 1.4 μV, a bit higher than the desired 1.2 μV. The total preamp noise is therefore 2.2 μV.

B.4 Measurements of Additional NPR03 Channel Circuits

B.4.1 SAH Measurements

Measurements of SAH circuits are shown in Fig. B.4. The measurements were performed on a stand-alone SAH circuit included in NPR03 for testing. Since the SAH was not intended to drive off-chip loads, the standalone circuit was measured at 10 kSps and not at the target 40 kSps. Direct measurement of the

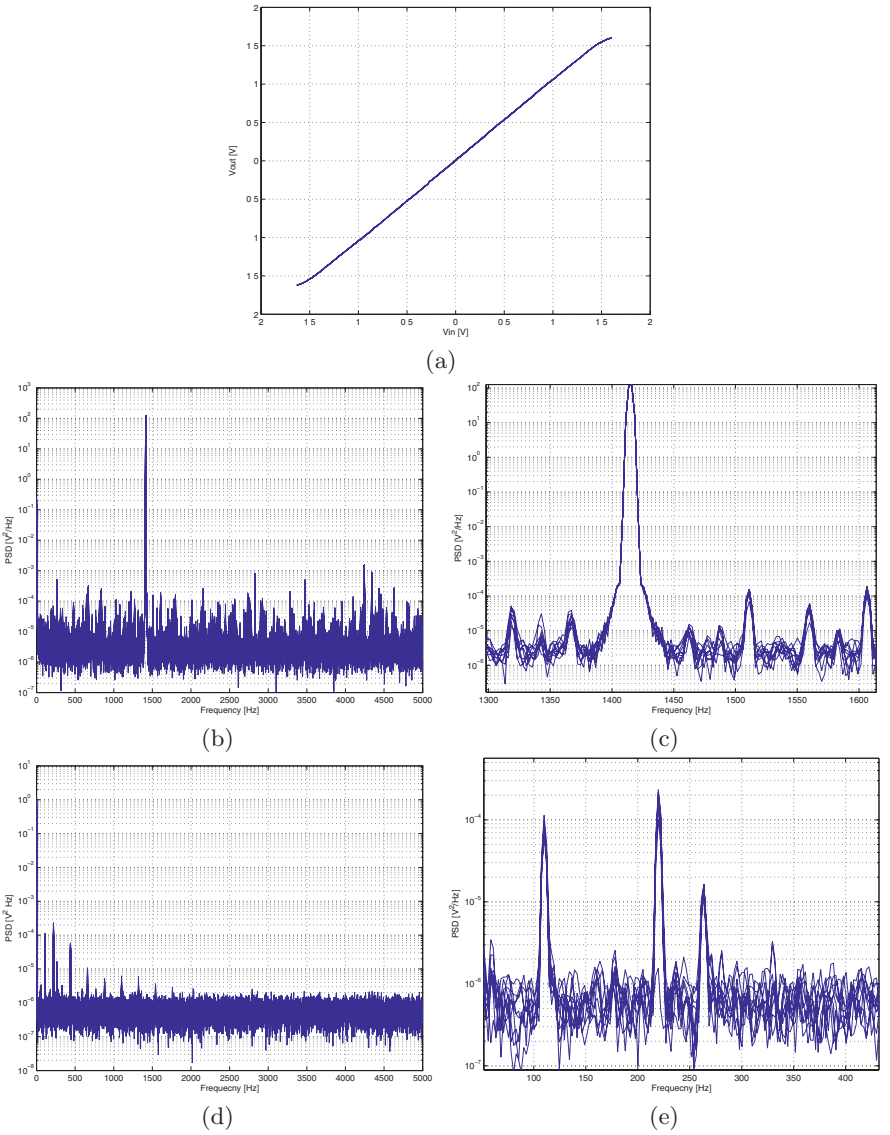


Fig. B.4. Cumulative plots of SAH measurements. (a) Transfer function. (b,c) Sine input with close-up. (d,e) Output noise PSD with close-up

transfer function proved somewhat problematic, due to the difficulty with supplying accurate input voltage (at least 10-bit accuracy was at need). Instead, a statistical measurement technique was employed, based on the following observation: We supply an input voltage uniformly distributed over the input range and measure the distribution of the output voltage. If we denote

the SAH transfer function by $TF(x)$, the output distribution density function $f_o(x)$ will be proportional to:

$$f_o(x) \propto \left(\frac{d}{dx} TF(x) \right)^{-1}$$

Therefore we can restore $TF(x)$ if we measure $f_o(x)$ by:

$$TF(x) = \int_0^x f_o^{-1}(\zeta) d\zeta$$

Input voltages of uniform distribution were supplied with a triangle-wave function generator, with output amplitude above the SAH full-scale. The input frequency was chosen to have a small greatest common divisor with the sampling frequency.

SAH was also measured with 1Vpp, 1.414 kHz sine wave at the input with THD of about 55 dB. Noise was measured with grounded input, with RMS value of 0.9 mV. All the measurements repeated on several dies with only small deviations.

B.4.2 ADC Measurements

Measurements were performed on a standalone ADC included in NPR03 for testing at 40 kSps rate (Fig. B.5). Transfer function was measured with statistical method similar to that used for SAH measurement. With 1Vpp input sine wave the ADC shows THD of about 55 dB. The ADC experiences several non-linearities at input voltages with 300 mV of supply rails, related to the input sample-and-hold. Unfortunately, test circuitry does not allow ADC measurements without the input SAH.

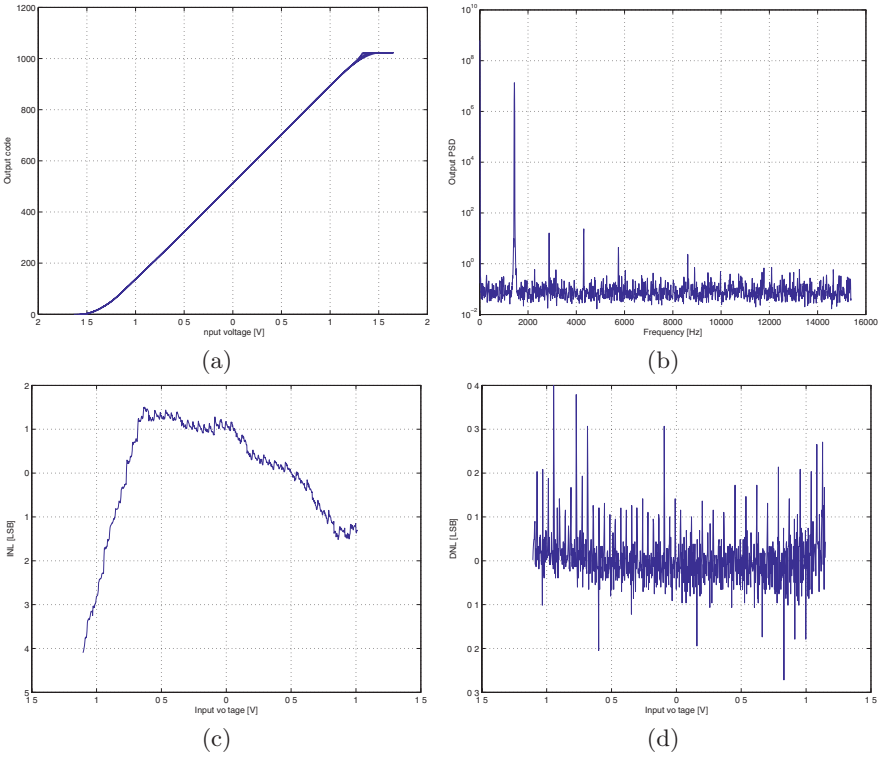


Fig. B.5. Cumulative plots of ADC measurements. (a) Transfer function. (b) Sine input. (c) INL. (d) DNL

References

1. F. Spelman, "The past, present and future of cochlear prostheses," *IEEE Eng. Med. Bio. Mag.*, vol. 18, p. 27–33, 1999.
2. G. Dagnelie and R. Massof, "Toward an artificial eye," *IEEE Spectr*, vol. 33, pp. 20–29, 1996.
3. J. Wessberg, C. Stambaugh, J. Kralik, P. Beck, M. Laubach, J. Chapin, J. Kim, S. Biggs, M. Srinivasan, and M. Nicolelis, "Real-time prediction of hand trajectory by ensembles of cortical neurons in primates," *Nature*, vol. 408, pp. 361–5, 2000.
4. A. Schwartz, "Cortical neuronal prosthetics," *Annu. Rev. Neurosci.*, vol. 27, pp. 487–507, 2004.
5. D. M. Durand and M. Bikson, "Suppression and control of epileptiform activity by electrical stimulation: A review," *Proc. IEEE*, vol. 89, no. 7, pp. 1065–82, 2001.
6. Medtronic, Inc., http://www.medtronic.com.
7. Cyberonics, Inc., http://www.cyberonics.com.
8. T. Berger, M. Baudry, R. Brinton, J. Liaw, V. Marmarelis, A. Park, B. Sjeu, and A. Tanguay, "Brain-implantable biomimetic electronics as the next era in neural prosthetics," *Proc. IEEE*, vol. 89, no. 7, pp. 993–1012, 2001.
9. R. Normann, "Microfabricated electrode arrays for restoring lost sensory and motor functions," in *Solid State Sensors, Actuators and Microsystems, Proc. 12th Int. Conf.*, 2003, pp. 959–962.
10. M. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, pp. 403–7, 2001.
11. B. Eversmann, M. Jenkner, F. Hofmann, C. Paulus, R. Brederlow, B. Holzapfl, P. Fromherz, M. Merz, M. Brenner, M. Schreiter, R. Gabl, K. Plehnert, M. Steinhauser, G. Eckstein, D. Schmitt-Landsiedel, and R. Thewes, "A 128×128 cmos biosensor array for extracellular recording of neural activity," *JSSC*, vol. 38, no. 12, pp. 2306–2317, 2003.
12. J. G. Nichols, A. R. Martin, and B. G. Wallace, *From Neuron To Brain*. Sinauer Associates, Inc., 1992.
13. R. R. Llinas, Ed., *The Biology Of The Brain. From Neurons to Networks*. W. H. Freeman & C., 1989.
14. E. H. Chudler, "Neuroscience for kids," http://faculty.washington.edu/chudler/neurok.html.

15. A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, pp. 500–544, 1952.
16. A. J. Bard and L. R. Faulkner, *Electrochemical Methods. Fundamentals and Applications*. John Wiley & Sons, 1980.
17. L. A. Geddes, *Electrodes and The Measurements of Bioelectric Events*. Whiley-Interscience, 1972.
18. W. L. C. Rutten, "Selective electrical interfaces with the nervous system," *Annu. Rev. Biomed. Eng.*, vol. 4, pp. 407–452, 2002.
19. G. T. A. Kovacs, *Enabling Technologies for Cultured Neural Networks*. Academic Press, 1994.
20. R. S. C. Cobbold, *Transducers for Biomedical Measurements: Principles and Applications*. Wiley, 1974.
21. D. A. Robinson, "The electrical properties of microelectrodes," *Trans. IEEE*, vol. 56, 1968.
22. R. C. Gesteland, B. Howland, J. Y. Lettvin, and W. H. Pitts, "Comments on microelectrodes," *Trans. RIE*, vol. 47, pp. 1856–1862, 1959.
23. M. Grattarola and S. Martinoia, "Modeling the neuron-microtransducer junction: From extracellular to patch recording," *IEEE Trans. Biomed. Eng.*, vol. 40, 1993.
24. D. A. Borkholder, "Cell based biosensor using microelectrodes," Ph.D. dissertation, Stanford University, Department of Electrical Engineering, 1998,
25. C. for Cell Biophysics Technion, personal communications with prof. Shimon Marom.
26. J. Struijk, M. Thornsen, J. Larsen, and T. Sinkjaer, "Cuff electrodes for long-term recording of natural sensory information," *IEEE Eng. Med. Bio. Mag.*, vol. 18, pp. 91–98, 1999.
27. K. Wise, D. Anderson, J. Hetke, D. Kipke, and K. Najafi, "Wireless implantable microsystems: High-density electronic interfaces to the nervous system," *Proc. IEEE*, vol. 92, no. 1, pp. 76–97, 2004.
28. M. Nicolelis, D. Dimitrov, J. Carmena, R. Crist, G. Lehew, J. Kralik, and S. Wise, "Chronic, multisite, multielectrode recordings in macaque monkeys," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 19, pp. 11 041–11 046, 2003.
29. Alpha-Omega Engineering, http://www.alphaomega-eng.com.
30. Tucker-Davis Technologies, http://www.tdt.com.
31. R. Vetter, J. Williams, J. Hetke, E. Nunamaker, and D. Kipke, "Chronic neural recording using silicon-substrate microelectrode arrays implanted in cerebral cortex," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 896–904, 2004.
32. T. H. Yoon, E. J. Hwang, D. Y. Shin, S. I. Park, S. J. Oh, S. C. Jung, H. C. Shin, and S. J. Kim, "A micromachined silicon depth probe for multichannel neural recording," *IEEE Trans. Biomed. Eng.*, vol. 47, pp. 1082–7, 2000.
33. C. Kim and K. Wise, "A 64-site multishank CMOS low-profile neural stimulating probe," *JSSC*, vol. 31, no. 9, pp. 1230–1238, 1996.
34. A. Blau, C. Ziegler, M. Heyer, F. Erndrest, G. Schwitzgebelt, T. Matthies, T. Stiglitz, J. U. Meyer, and W. Gopel, "Characterization and optimization of microelectrode arrays for in-vivo nerve signal recording and stimulation," *Biosens. Bioelectro.*, vol. 12, pp. 883–92, 1997.
35. G. T. Kovacs, C. W. Sorment, and J. M. Rosen, "Regeneration microelectrode array for peripherial nerve recording and stimulation," *IEEE Trans. Biomed. Eng.*, vol. 39, no. 9, pp. 893–902, 1992.

36. T. Akin, K. Najafi, R. R. Smoke, and R. M. Bradley, "A micromachined silicon sieve electrode for nerve regenerating applications," *IEEE Trnas. Biomed. Eng.*, vol. 41, no. 4, pp. 305–13, 1994.
37. C. A. Thomas, P. A. Springer, G. E. Loeb, Y. Berwald-Netter, and L. M. Okun, "A miniature microelectrode array to monitor the bioelectric activity of cultured cells," *Exp. Cell Res.*, vol. 74, pp. 61–6, 1972.
38. G. Gross, E. Rieske, G. Kreutzberg, and A. A. Meyer, "A new fixed-array multimicroelectrode system designed for longterm monitoring of extracellular single unit neuronal activity in vitro," *Neurosci. Lett.*, vol. 6, pp. 101–105, l977.
39. M. P. Maher, J. Pine, J. Wright, and Y.-C. Tai, "The neurochip: A new multielectrode device for stimulating and recording from cultured neurons," *J. Neurosci. Methods*, vol. 87, pp. 45–56, 1999.
40. P. Fromherz and G. Zeck, "Noninvasive neuroeletronic interfacing with synaptically connected snail neurons immobilized on a semiconductor chip," *Proc. Natl. Academy Sci. USA*, vol. 98, no. 18, pp. 10 457–62, 2001.
41. S. Marom and G. Shahaf, "Development, learning and memory in large random networks of cortical neurons: Lessons beyond anatomy," *Q. Rev. Biophys.*, vol. 35, pp. 63–87, 2002.
42. S. Potter, D. Wagenaar, R. Madhavan, and T. DeMarse, "Long-term bidirectional neuron interfaces for robotic control and in vitro learning studies," in *IEEE EMBS, 25th Ann. Int. conf.*, 2003, pp. 3690–3693.
43. G. Gross, B. Rhoades, H. Azzazy, and M. Wu, "The use of neuronal networks on multielectrode arrays as biosensors," *Biosen. Bioelectro.*, vol. 10, pp. 553–567, 1995.
44. A. Stett, U. Egert, E. Guenther, F. Hofmann, T. Meyer, W. Nisch, and H. Haemmerle, "Biological application of microelectrode arrays in drug discovery and basic research," *Anal. Bioanal. Chem.*, vol. 377, pp. 486–495, 2003.
45. MultiChannel Systems, http://www.multichannelsystems.com.
46. S. Potter, "Distributed processing in cultured neuronal networks," *Prog. Brain Res.*, vol. 130, pp. 49–62, 2001.
47. F. Heer, W. Franks, I. McKay, S. Taschini, A. Hierlemann, and H. Baltes, "CMOS microelectrode array for extracellular stimulation and recording of electrogenic cells," in *Circuits and Systems, 2004. ISCAS '04. Proc. 2004 Int. Symp. on*, vol. 4, 2004, pp. IV – 53–6.
48. A. Cohen, M. Spira, S. Yitshaik, G. Borghs, O. Shwartzglass, and J. Shappir, "Depletion type floating gate p-channel MOS transistor for recording action potentials generated by cultured neurons," *Biosen. bioelectro.*, vol. 19, pp. 1703–1709, 2004.
49. A. Offenhausser, J. Ruhe, and W. Knoll, "Neuronal cells cultured on modified microelectronic device surfaces," *J. Vac. Sci. Tech.*, vol. 13, no. 5, pp. 2606–2612, 1995.
50. L. Berdondini, T. Overstolz, N. de Rooij, M. Koudelka-Hep, M. Wany, and P. Seitz, "High-density microelectrode arrays for electrophysiological activity imaging of neuronal networks," in *Electronics, Circuits and Systems. Proc. of IEEE Int. Conf. on*, vol. 3, 2001, pp. 1239–1242.
51. P. Fromherz, "Electrical interfacing of nerve cells and semiconductor chips," *Chemphyschem*, vol. 3, pp. 276–284, 2002.
52. P. Fromherz and A. Stett, "Silicon-neuron junction: Capacitive stimulation of an individual neuron on a silicon chip," *Phys. Rev. Lett.*, vol. 75, no. 8, pp. 1670–4, 1995.

53. P. Fromherz, A. Stett, and B. Muller, "Two-way silicon-neuron interface by electrical induction," *Phys. Rev. Lett. E.*, vol. 55, no. 2, pp. 1779–82, 1997.
54. Plexon, Inc., http://www.plexoninc.com.
55. Triangle Biosystems, Inc., http://www.tbsi.biz.
56. I. Obeid, M. Nicolelis, and P. Wolf, "A low power multichannel analog front end for portable neural signal recordings," *J. Neurosci. Methods*, pp. 27–32, 2004.
57. J. Morizio, P. Irazoqui, V. Go, and J. Parmentier, "Wireless headstage for neural prosthetics," in *Neural Engineering, Proc. 2nd Int. IEEE EMBS Conf. on*, pp. 414–417, 2005.
58. J. Morizio, D. Won, I. Obeid, C. Bossetti, M. Nicolelis, and P. Wolf, "16-channel neural pre-conditioning device," in *Neural Engineering, Proc. of the 1st Int. IEEE EMBS Conf.*, pp. 104–107, 2003.
59. J. Ji and K. D. Wise, "An implantable CMOS circuit interface for multiplexed microelectrode recording arrays," *JSSC*, vol. 27, no. 3, pp. 433–443, 1992.
60. M. Dagtekin, W. Liu, and R. Bashirullah, "A multichannel chopper modulated neural recording system," *IEEE Eng. Med. Biol. Conf.*, 2001.
61. D. Chen, J. Harris, and J. Principe, "A bio-amplifier with pulse output," in *Proc. 26th Ann. Int. Conf. IEEE EMBS*, 2004, pp. 4071–4074.
62. T. Horiuchi, T. Swindell, D. Sander, and P. Abshire, "A low-power CMOS neural amplifier with amplitude measurement for spike sorting," in *Circuits and Systems, Proc. 2004 Int. Symp. on*, vol. 4, 2004, pp. IV – 29–32.
63. R. R. Harrison and C. Charles, "A low-power, low-noise CMOS amplifier for neural recording applications," *JSSC*, vol. 38, no. 6, pp. 958–965, 2003.
64. Q. Bai and K. Wise, "Single-unit neural recording with active microelectrode arrays," *IEEE Trans. biomed. Eng.*, vol. 48, no. 8, pp. 911–920, 2001.
65. K. Kim and S. Kim, "Noise performance design of CMOS preamplifier for the active semiconductor neural probe," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 8, pp. 1097–1105, 2000.
66. P. Mohseni and K. Najafi, "A fully integrated neural recording amplifier with dc input stabilization," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 5, pp. 832–837, 2004.
67. M. Mojarradi, D. Binkley, B. Blalock, R. Andersen, N. Ulshoefer, T. Johnson, and L. D. Castillo, "A miniaturized neuroprosthesis suitable for implantation into the brain," *IEEE Trans. Neural Syst. Rehabi. Eng.*, vol. 11, no. 1, pp. 38–42, 2003.
68. W. Patterson, Y. Song, C. Bull, I. Ozden, A. Deangellis, C. Lay, J. McKay, A. Nurmikko, J. Donoghue, and B. Connors, "A microelectrode/microelectronic hybrid device for brain implantable neuroprosthesis applications," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 10, pp. 1845–1853, 2004.
69. Y. Song, W. Patterson, C. W. Bull, J. Beals, N. Hwang, A. Deangelis, C. Lay, J. McKay, A. Nurmikko, M. Fellows, J. Simeral, J. Donoghue, and B. Connors, "Development of a chipscale integrated microelectrode/microelectronic device for brain implantable neuroengineering applications," *IEEE Transactions on Neural Syst. Rehabil. Eng.*, vol. 13, no. 2, pp. 220–226, 2005.
70. T. Akin, K. Najafi, and R. Bradley, "A wireless implantable multichannel digital neural recording system for a micromachined sieve electrode," *IEEE J. Solid State Circuits*, vol. 33, no. 1, pp. 109–18, 1998.
71. Chipcon, http://www.chipcon.com.

72. S. Farshchi, P. Nuyujukian, A. Pesterev, I. Mody, and J. Judy, "A TinyOS-based wireless neural sensing, archiving, and hosting system," in *Neural Engineering, Proc. 2nd Int. IEEE EMBS Int. Conf. on*, 2005.
73. R. Rangarajan, J. von Arx, and K. Najafi, "Fully integrated neural stimulation system (finess)," *Proc. 43rd IEEE Midwest Symp. Circuits Syst.*, pp. 1082–1085, 2002.
74. R. Harrison, "A low-power integrated circuit for adaptive detection of action potentials in noisy signals," in *IEEE EMBS, Proc. 25th Ann. Int. Conf. of*, 2003, pp. 3325–3328.
75. G. Kreiman, "Neural coding: computational and biophysical perspectives," *Phys. Life Rev.*, vol. 1, pp. 71–102, 2004.
76. A. Schwartz, R. Kettner, and A. Georgopoulos, "Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement," *J. Neurosci.*, vol. 8, pp. 2913–2927, 1988.
77. NeuroNexus Technologies, http://www.neuronexustech.com.
78. Z. Zumsteg, R. Ahmed, G. Santhanam, K. Shenoy, and T. Meng, "Power feasibility of implantable digital spike-sorting circuits for neural prosthetic systems," in *Proc. 26th Ann. Int. Conf. IEEE EMBS*, 2004, pp. 4237–4240.
79. ZigBee Alliance, http://www.zigbee.org.
80. Zarlink Semiconductor, http://www.zarlink.com.
81. J. Chapin and K. Moxon, Eds., *Neural Prostheses for Restoration of Sensory and Motor Functions*. CRC Press, 2001.
82. A. Arieli, A. Sterkin, A. Grinvald, and A. Aertsen, "Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses," *Science*, vol. 273, pp. 1868–1871, 1996.
83. J. Donoghue, J. Sanes, N. Hatsopoulos, and G. Gaal, "Neural discharge and local field potential oscillations in primate motor cortex during voluntary movements," *J. Neurophysiol.*, vol. 79, no. 1, pp. 159–173. 1998.
84. C. Mehring, J. Ricket, E. Vaadia, S. C. de Oliviera, Aersten, and S. Rotter, "Inference of hand movements from local field potentials in monkey motor cortex," *Nature Neurosci.*, vol. 6, no. 12, pp. 1253–1254, 2003.
85. K. Guillory and R. Normann, "A 100-channel system for real time detection and storage of extracellular spike waveforms," *J. Neurosci. Methods*, vol. 91, pp. 21–29, 1999.
86. R. Olsson, M. Gulari, and K. Wise, "A fully-integrated bandpass amplifier for extracellular neural recording," in *Neural Eng., 1st Int. IEEE EMBS Conf. on*, pp. 165–169, 2003.
87. Y. Perelman and R. Ginosar, "An integrated system for multichannel neuronal recording with spike / LFP separation and digital output," *2nd Int. IEEE EMBS Conf. Neural Engineering*, pp. 377–380, 2005.
88. Y. Perelman and R. Ginosar, "Analog frontend for multichannel neuronal recording system with spike and LFP separation," *Journal of Neuroscience Methods*, vol. 153, no. 1, pp. 21–26, 2006.
89. T. L. Deliyannis, Y. Sun, and J. K. Fidler, *Continuous-Time Active Filter Design*. CRC Press, 1999.
90. Y. Perelman and R. Ginosar, "An integrated system for multichannel neuronal recording with spike/LFP separation, integrated A/D conversion and threshold detection," *IEEE Trans. on Biomedical Engineering*, vol. 54, no. 1, pp. 130–137, 2007.

91. Y. Perelman and R. Ginosar, "A low power inverted ladder D/A converter," *IEEE Trans on Circuits Systems II*, vol. 53, no. 6, pp. 497–501, 2006.
92. Texas Instruments, http://www.ti.com.
93. T. Instruments. (2004) TMS320C6000 DSP multichannel buffered serial port (McBSP) reference guide. [Online]. Available: http://ti.com
94. Analog Devices, http://www.analog.com.
95. C. Enz, F. Krummenacher, and E. Vittoz, "An analytical MOS transistor model valid in all regions of operation and dedicated to low-voltage and low-current applications," *Analog Integrated Circuits and signal Process.*, vol. 8, pp. 83–114, 1995.
96. M. Lewicki, "A review of methods for spike sorting: The detection and classification of neural action potentials," *Network: Computation in Neural Syst.*, vol. 9, no. 4, pp. R53–R78, 1998.
97. M. Abeles and M. Goldstein, "Multispike train analysis," *Proc. IEEE*, vol. 65, no. 5, pp. 762–73, 1977.
98. X. Yang and S. Shamma, "A totally automated system for the detection and classification of neural spikes," *IEEE Trans. on Biomed. Eng.*, vol. 35, no. 10, pp. 806–816, 1988.
99. J. Letelier and P. Weber, "Spike sorting based on discrete wavelet transform coefficients," *J. Neurosci. Methods*, vol. 101, pp. 93–106, 2000.
100. K. Oweiss and D. Anderson, "A multiresolution generalized maximum likelyhood approach for the detection of uknown transient multichannel signals in colored noise with unknown covariance," in *Acoustics, Speech, and Signal Processing. Proc. IEEE Int. Conf.*, vol. 3, 2002, pp. 2993–2996.
101. K. Kim and S. Kim, "A wavelet-based method for action potential detection from extracellular neural signal recording with low signal-to-noise ratio," *IEEE Trans. on Biomed. Eng.*, vol. 50, no. 8, pp. 999–1011, 2003.
102. R. Chandra and L. Optican, "Detection, classification, and superposition resolution of action potentials in multiunit single-channel recordings by an online real-time neural network," *IEEE Trans. Biomed. Eng.*, vol. 44, no. 5, pp. 403–412, 1997.
103. Y. P. A. Zviagintsev and R. Ginosar, "A low-power spike detection and alignment algorithm," in *Neural Eng., 2005. Conference Proc. 2nd Int. IEEE EMBS Conference on.*, 2005, pp. 317–320.
104. Y. P. A Zviagintsev and R. Ginosar, "Low-power architectures for spike sorting," in *Neural Engineering, 2005. Conference Proc. 2nd Intl. IEEE EMBS Conference on.*, 2005, pp. 162–165.
105. A. Zviagintsev, "Hardware algorithms and architectures for low power spike detection and sorting," Master's thesis, Technion, Israel Institute of Technology, 2005.
106. F. Wood, M. Black, C. Vargas-Irwin, M. Fellows, and J. Donoghue, "On the variability of manual spike sorting," *IEEE Trans. on Biomed. Eng.*, vol. 51, no. 6, pp. 912–918, 2004.
107. M. Lewicki, "Bayesian modeling and classification of neural signals," *Neural Computation*, vol. 6, pp. 1005–30, 1994.
108. M. Sahani, "Latent variable models for neural data analysis," Ph.D. dissertation, California Institute of Technology, 1999.
109. I. Bar-Gad, Y. Ritov, E. Vaadia, and H. Bergman, "Failure in identification of overlapping spikes from multiple neuron activity causes artificial correlations," *J. Neurosci. Methods*, vol. 107, pp. 1–13, 2001.

110. Y. P. A. Zviagintsev and R. Ginosar, "Algorithms and architectures for low power spike detection and alignment," *J. Neural Eng.*, vol. 3, pp. 35–42, 2006.
111. A. Lyakhov, "VLSI sensor chip for in-vitro measurement of biological neural network activity," Master's thesis, Technion, Israel Institute of Technology 2005.

Index