

Srinjoy Mitra · David R.S. Cumming
Editors

CMOS Circuits for Biological Sensing and Processing

 Springer

CMOS Circuits for Biological Sensing and Processing

Srinjoy Mitra • David R.S. Cumming
Editors

CMOS Circuits for Biological Sensing and Processing

 Springer

Editors

Srinjoy Mitra
School of Engineering
University of Glasgow
Glasgow, UK

David R.S. Cumming
School of Engineering
University of Glasgow
Glasgow, UK

ISBN 978-3-319-67722-4

ISBN 978-3-319-67723-1 (eBook)

DOI 10.1007/978-3-319-67723-1

Library of Congress Control Number: 2017956345

© Springer International Publishing AG 2018, corrected publication 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Silicon chips consisting of CMOS circuits form the basic building block of all modern computing technologies from which we benefit so greatly. They also provide the foundation for Moore's Law, which drives the technology to pack even more transistors in the same area. However, the rate of growth in circuit density has already started to slow, and fundamental limits are looming on the horizon. Furthermore, this extreme scaling works well for memories and microprocessors in the digital world, but not for interfacing with the actual physical world, which is analog in nature. CMOS-based microelectronics, therefore, seeks to develop in new ways, not only to continue delivering better performance in traditional computing but also to provide novel functions on integrated circuits. Interface to biosensors is one such domain where CMOS circuits have shown great promise over the past decade. There are a wide variety of applications where sensing and processing biological information close to the signal source could significantly improve the quality of measurement and enhance the usability of the device. Most advanced bioelectronic devices intend to design CMOS chips with direct interface to biology, compared to various traditional setups consisting of complex wiring and massive instrumentation. Such applications (of CMOS ICs) go beyond the realm of traditional computing devices and consumer electronic gadgets. They, in fact, track the More-than-Moore axis defined in the ITRS roadmap. However, all these custom applications come with their own challenges and post-processing techniques are often necessary to make use of the cheap versatile silicon platform. Nevertheless, researchers around the world are now heavily engaged in the integration of biosensors and systems using CMOS.

With these advances, it is increasingly possible to create highly heterogeneous systems using a great variety of technologies. Furthermore, it is also becoming possible to miniaturize these integrated biosensors for point-of-care, wearable, or implantable applications. Given the diverse range of biosensors and what constitutes a meaningful signal for a particular application, the necessary IC design specifications can vary widely. While access to such integrated silicon sensors can drive the need for even better ICs (more signal sources, smarter on-chip processing, multimodal sensing, etc.), new transducers can make ICs designed to interface

with older ones redundant. In spite of the plethora of techniques developed to interface widely varying sensors, there is indeed a certain thread of commonality in this domain. This is the need for understanding and innovating in analog and mixed-signal CMOS circuits. In order to provide an appropriate solution, it is often necessary for the IC designer to have some knowledge of the transducer technology (including its limitations) and how it interacts with biology. This can also lead to extracting more meaningful signal from custom ICs rather than depending on generic methodology.

This book provides a state-of-the-art overview of such circuits and transducers for biological sensing and processing in a multitude of applications, written by leaders in the field. The book highlights the techniques in analog and mixed-signal circuit design that potentially can cross boundaries and benefit the very wide community of biomedical engineers. The first five chapters describe techniques related to biomolecular sensing that could be used for either *in vivo*, *ex vivo* or *in vitro* applications. Chapters 6, 7, 8, 9, 10, and 11 deal with specific needs for implantable or wearable electronics and consider the necessary system specifications. The last two chapters emphasize signal processing for highly scaled-up systems that will be essential to future biosensing integrated circuit technologies.

This book would not have been possible without the patient persistence of Springer Editorial Director Charles Glaser. We are particularly thankful to all the contributing authors who have given an up-to-date account of their own work and their field of research in general. We also thank Samadah Patil, Mohammed Al-Rawhani, Chunxiao Hu, Srinivas Velugotla and Claudio Accarino from the MST Group in the University of Glasgow for kindly reviewing the articles.

Glasgow, UK
23 August 2017

Srinjoy Mitra
David R.S. Cumming

Contents

1	CMOS Nano-Pore Technology	1
	Sina Parsnejad and Andrew J. Mason	
2	Metabolomics on CMOS for Personalised Medicine	23
	Boon Chong Cheah and David R. S. Cumming	
3	Flexible Single-Photon Image Sensors	47
	Pengfei Sun, Ryoichi Ishihara, and Edoardo Charbon	
4	CMOS Multimodal Sensor Array for Biomedical Sensing	77
	Kazuo Nakazato	
5	Micro-NMR on CMOS for Biomolecular Sensing	101
	Ka-Meng Lei, Nan Sun, Pui-In Mak, Rui Paulo Martins, and Donhee Ham	
6	Microelectronics for Muscle Fatigue Monitoring Through Surface EMG	133
	Pantelis Georgiou and Ermis Koutsos	
7	Design and Optimization of ICs for Wearable EEG Sensors	163
	Jiawei Xu, Rachit Mohan, Nick Van Helleputte, and Srinjoy Mitra	
8	Circuits and Systems for Biosensing with Microultrasound	187
	Holly Susan Lay and Sandy Cochran	
9	High-Density CMOS Neural Probes	211
	Bogdan Raducanu, Carolina Mora Lopez, and Srinjoy Mitra	
10	Photonic Interaction with the Nervous System	233
	Patrick Degenaar	

11 Implantable Microsystems for Personalised Anticancer Therapy 259
Jamie R. K. Marland, Ewen O. Blair, Brian W. Flynn,
Eva González-Fernández, Liyu Huang, Ian H. Kunkler,
Stewart Smith, Matteo Staderini, Andreas Tsiamis, Carol Ward,
and Alan F. Murray

12 Compressed Sensing for High Density Neural Recording 287
Jie Zhang, Tao Xiong, Srinjoy Mitra, and Ralph Etienne-Cummings

**13 Very Large-Scale Neuromorphic Systems for Biological Signal
Processing** 315
Francky Catthoor, Srinjoy Mitra, Anup Das, and Siebren Schaafsma

Erratum E1

Index..... 341

Chapter 1

CMOS Nano-Pore Technology



Sina Parsnejad and Andrew J. Mason

1 Nano-Pore-Based Sensing

1.1 Introduction to Nano-Pore Sensing

Recently great research effort has been put into chemistry-enabled sensing systems such as air pollutant detectors, biosensors for antibody detection, as well as more complex emerging areas such as silicon nanowires [1] and carbon nanotubes [2]. All these methods are subsets of a greater family called electrochemical sensors. Electrochemical sensors utilize tracking and exploit charge transfer in chemical reactions for discovering a sensing target. In recent years, new opportunities for mainstream utilization of these techniques have emerged throughout the literature. For example, electrochemical sensing has received a great deal of attention for developing biosensors in recent years [3, 4]. They are also being used in fields such as environmental sensing, safety monitoring, and clinical diagnosis in both stationary and portable systems. The popularity of electrochemical sensors is due to reliability, simplicity, low cost, and portability of these systems [5]. Electrochemical biosensors are a subset that associate with and detect a biological event or a marker using charge transfer. Amperometric electrochemical biosensors detect this event by receiving the said charge through an electrode and transducing it into a current signal.

The original version of this chapter was revised.

An erratum to this chapter can be found at https://doi.org/10.1007/978-3-319-67723-1_14

S. Parsnejad • A.J. Mason (✉)

Michigan State University College of Electrical and Computer Engineering,
East Lansing, MI, USA

e-mail: parsneja@msu.edu; mason@msu.edu

The information retrieved from electrochemical biosensors is usually integrated over a large number of charge transfers and integrated over time from individual electrodes [6]. This is, however, contrary to the growing demand for tracking the charge transfer of individual molecules in a quantitative manner. This is not possible with traditional electrochemical systems due to the fact that electrochemical electrodes form a layer of charge known as double-layer capacitance on their surface shared with the solution. This capacitance is noisy and thus masks the passing of low doses of charge created by a single-molecule electrochemical event [6].

To enable single-molecule electrochemical detection, several methods have been proposed including redox cycling, nanoparticles, and nano-pores. A nano-pore is basically a nanoscale-sized path for charged or neutral particle transfer with well-defined flow characteristics [7]. The operation principles of nano-pores revolve around the fact that an open pore can demonstrate constant ion flow under the right circumstances and any disruption to this flow would be detectable. One way to utilize this configuration for detecting particles is to observe the changes in charge flow through the nano-pore due to the presence of particles with certain sizes. Furthermore, this method can also be utilized to observe the behavior of the nano-pore under certain environmental and internal stimulation conditions, enabling characterization of the pore. Nano-pores have been reported for use in fields including single-molecule biosensing, protein detection, and ultrafast label-free DNA sequencing [8, 9] as well as more specific applications such as detection of microRNAs [10], measurement of molecular forces [11], and identifying anthrax toxins [12]. As the demand to bring these technologies to the mainstream advances, there is a growing need for better understanding of nano-pores and their interfacing circuitry. The next sections of this chapter discuss the structure of nano-pores, especially ion channel proteins, and overview some of the challenges in developing electronic interfaces for nano-pores.

1.2 Nano-Pore Types and Implementations

Two main types of nano-pores are currently being used for amperometric measurements: ion channels and solid-state nano-pores. The first practical implementation of nano-pores was detection of single-stranded DNA molecules by passing them through a staphylococcal α -hemolysin ion channel protein acting as a nano-pore gate [13]. An ion channel is a pore-forming membrane protein that forms on cell's bilayer lipid membrane (BLM). A BLM is a continuous sheet that forms a barrier around the cell and is populated with membrane proteins that are responsible for communication with the outside world. A typical BLM populated by various membrane proteins where each is performing a specific function is depicted in Fig. 1.1.

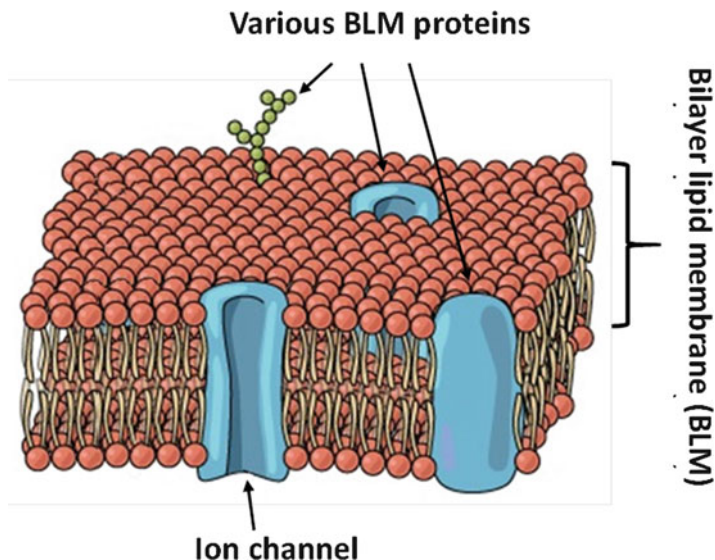
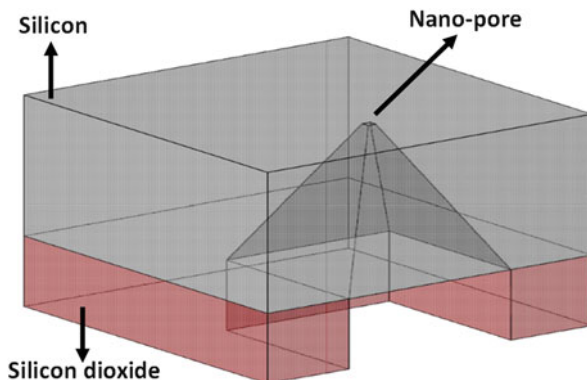


Fig. 1.1 Typical BLM populated with various membrane proteins source

The bulk of BLM is composed of amphiphilic phospholipids that are hydrophilic on one tail and hydrophobic on the other side. Accompanying these cells are the membrane proteins that are anchored to the surface and perform various tasks. Ion channels are one class of membrane protein, and they are the fundamental excitable elements in most cell membranes [14]. Their jobs include establishing a resting membrane potential and enabling action potentials by gating ions through the BLM. Thus, they are responsible for regulating cell volume and controlling ion flow across secretory and epithelial cells.

Solid-state nano-pores are another approach to nano-pore sensing. Due to recent advancements in microfabrication techniques, it is possible to implement microfabricated solid-state pores with nanometer-sized openings. An example of a nano-pore created on a silicon nitride surface is discussed in [15] and shown in Fig. 1.2. These pores are typically implemented in dielectric materials such as glass, silicon dioxide (SiO_2), silicon nitride (Si_3N_4), and polymers [16]. Two example methods for creating these openings are ion-beam sculpting for fabricating pores on a Si_3N_4 substrate [17] and implementation of precise nano-pores in a SiO_2 substrate by shrinking a large hole ($\sim 20\text{--}200$ nm) to single nanometer precision using the surface tension created by high-energy electron beam together with a visual feedback [18]. In terms of electrical properties and performance, there is not much difference between an ion channel and a microfabricated pore. Thus, throughout this chapter, the term nano-pores will generally refer to either of these structures.

Fig. 1.2 Solid-state nano-pore fabricated on a silicon substrate



1.3 Challenges in Nano-Pore-Based Sensing

Nano-pores transduce bio/chemical events into variations in ionic current flow that can vary in magnitude depending on factors such as the pore dimensions, viscosity of the analytes, capacitances stemming from the isolating layers or the electrodes, etc. Typically these ionic currents have an amplitude in pA region with timing characteristics, e.g., pulse width, as short as tens of microseconds. This imposes stringent limitations on the nano-pore interface system since the noise floor should at least be 10 dB lower than the signal level for it to be detectable. Hence a typical noise floor of fA levels is required for effective nano-pore sensing. For example, sodium ion channels typically have a conductance of 100 pS resulting in a current in the order of 10 pA when excited with a biasing voltage of 100 mV. As a result, they require a root-mean-square (RMS) noise level of less than 100 fA in a 1 kHz bandwidth [19]. Another example is a microfabricated glass and polyethylene terephthalate membrane that demonstrate an RMS noise ranging from 5 to 25 pA, depending on membrane resistance, at a bandwidth of 40 kHz [16]. The benchmark instrument for current sensing in electrophysiology is the Axon Axopatch 200B, which provides a noise floor of 25 fA RMS at 1 kHz frequency ($0.7 \text{ fA}/\sqrt{\text{Hz}}$ until 1 kHz) [19]. However, this device is bulky, expensive, and designed for specific lab environments and experienced users. The limited existing instrumentation has created an obstacle for the development of nano-pore sensor that exploits their capabilities for precision measurement and single-molecule detection. However, as discussed in the next section of this chapter, recent advancements in CMOS technology and microfabrication techniques have enabled the opportunity to utilize custom CMOS circuitry for high-speed, low-noise, and cost-effective customizable alternatives to interfacing with nano-pore sensors.

2 CMOS Approaches for Nano-Pore Sensing

This section reviews solutions for characterization of nano-pores and detecting nano-pore events with a focus on addressing the challenge of interfacing these nano-pores through CMOS design. First, design considerations related to nano-pore working principles and methods of utilizing nano-pores as sensors from a physical and structural point of view will be discussed. Then, the most effective CMOS interfacing methods will be presented.

2.1 Electrical Model of Nano-Pores

Throughout this section, the electrical model of a conducting ion channel will be discussed. It is entirely possible to estimate the electrical behavior of a conducting ion channel as electrical components even though the component values would vary based on the overall nano-pore working conditions. The electrical model is valuable for understanding the basic behavior of ion channels under a given input voltage. The circuit equivalent of an ion channel implanted on a membrane is shown in Fig. 1.3 (left). The parallel capacitance and resistance depict a low-pass filter-like operation with a time constant of $R_{ch}C_M$, where R_{ch} represents the voltage-current behavior of the channel and C_M represents the capacitance of the exposed cell membrane area. Based on empirical evidence, C_M has a typical capacitance of $1 \mu\text{F}/\text{cm}^2$ if the target nano-pore is an ion channel with a typical BLM structure [14].

A model for the voltage-current relationship of the conducting ion channel shown in Fig. 1.3 under equilibrium conditions is given by [14]:

$$I_{\text{pore}} = g_{\text{ch}} (E - E_{\text{pore}}) \quad (1.1)$$

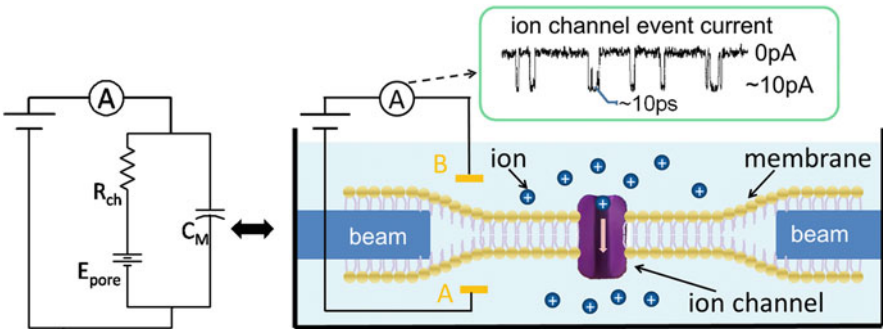


Fig. 1.3 Ion channel physical structure (*right*) with an equivalent circuit model (*left*) and current response during an amperometric test (*top*)

where I_{pore} is the total ionic current passing through the nano-pore under a given input voltage of E , g_{ion} is the total channel conductance, and E_{pore} is the equilibrium potential of the conducting nano-pore. By definition, if the nano-pore is conducting only a certain type of ion, E_{pore} would be the total voltage gradient created inside the nano-pore due to the presence of ions passing through it, i.e., if the net voltage over the nano-pore is E_{pore} , then the flux of ions through the pore will be zero. Notice that (1.1) only represents the equilibrium state of nano-pore operation and does not factor in the time constant $R_{\text{ch}}C_{\text{M}}$. Furthermore, the voltage-current relationship of nano-pores is nonlinear. For example, in case of an ion channel, the relationship is affected by variables such as diffusion currents due to higher concentration of permanent ions on one side of the pore.

The current response of a typical ion channel on a synthetic BLM layer formed on a silicon structure is depicted in Fig. 1.3 (top). Depending on the membrane size and type as well as the electrolyte conditions, ion channel event responses may vary in amplitude from ~ 1 to 200 pA with an offset depending on the aperture size and other environmental variables. For ion channels, the incident pulse length can be anywhere between 1 μs and 1 ms [20], and the incidents generally occur in a pulse-like format that indicates the opening and closing events of the ion channel or particle passing through. The analog interface for ion channel incidents must have proper performance characteristics to accurately measure such fast-changing and small amplitude signals. Solid-state nano-pores exhibit similar characteristics and analog interface requirements.

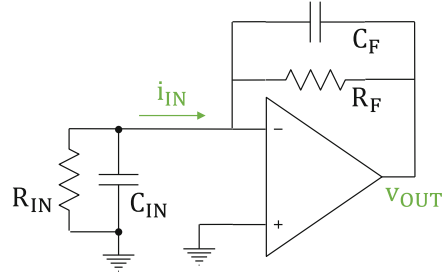
2.2 Amperometric Sensing with Nano-Pores

One of the most prevalent electrochemical detection methods is amperometry. First used in ion chromatography, amperometry is accomplished in the presence of a target analyte by applying a voltage between a working and auxiliary (reference) electrode in a two-electrode system or a working and reference electrode in a three-electrode system. The response is signified by a current that passes through the working electrode. This current is directly proportional to the mole concentration of the analyte oxidized or reduced on the electrode surface at a given input voltage, as described by Faraday's law [21]:

$$i_t = \frac{dQ}{dt} = nF \frac{dN}{dt} \quad (1.2)$$

where i_t is the current generated at the working electrode surface at the time t , Q is the charge stored at the electrode surface in form of a double-layer capacitance, t is time, n is the number of electrons transferred per mole of analyte, N is the number of moles of analyte oxidized or reduced, and F is the Faraday constant (96,485.33 C/mol). In a normal electrochemical reaction, selectivity is achieved through selection of an input voltage that causes oxidation or reduction of different target analytes.

Fig. 1.4 A standard TIA structure with a resistive and capacitive feedback interfacing a standard electrical model for an ion channel



Performing amperometry in nano-pores is slightly different from the normal process in that the input voltage between the working and reference electrodes does not aim to oxidize or reduce analytes. Rather, this voltage is used to facilitate the electrostatic transportation of ions through the nano-pore. For some ion channels, this voltage may also impact the behavior of the pore, for example, the open or closed state of a voltage-gated channel. Regardless, the basic amperometry function is the same: measure a response current at a given working electrode potential. Thus, the working electrode is usually connected to a buffer state that performs two fundamental actions: it controls the voltage on working electrode through a feedback loop and converts the Faraday response current into an electrically viable signal.

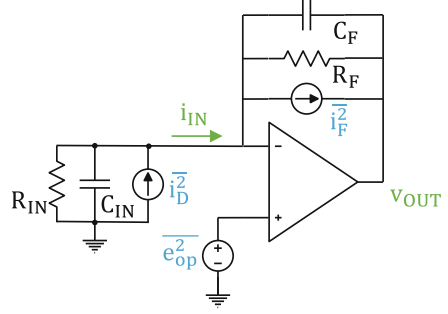
The most widely used amperometric readout circuit is a trans-impedance amplifier (TIA). A TIA consists of a standard operational amplifier with a negative feedback element that regulates the negative input of the operational amplifier without saturating the op-amp output voltage. The most commonly used feedback element is a resistor in parallel with a capacitor for stability. Thus, the output voltage to input current ratio would be equal to $(-R)$. TIAs are widely used as amperometric sensor interfaces due to their simplicity, linearity, and input-output voltage isolation [22]. Figure 1.4 depicts a standard TIA structure with a resistive and capacitive feedback interfacing to an equivalent nano-pore circuit model. This model will be used as the basis for noise discussion below.

2.2.1 Noise Limitations in Amperometric Interfacing of Nano-Pores

The amperometric current produced by a nano-pore is small and has a short duration. The bandwidth required for the interface circuit to be able to capture these fast-transitioning signals spans from DC to hundreds of kHz. Thus, the interface circuit is susceptible to both low-frequency noise element such as flicker noise and high-frequency noise elements. Consequently, the interface circuit must demonstrate exemplary noise performance. The ideal input-output relationship of the resistive and capacitive feedback shown in Fig. 1.4 can be expressed as:

$$v_{\text{OUT}} = -\frac{R_{\text{F}}}{1 + j2\pi f R_{\text{F}} C_{\text{F}}} \times i_{\text{IN}} \quad (1.3)$$

Fig. 1.5 A classic TIA amplifier with resistive feedback and all possible noise sources. The terms i_{IN} and v_{OUT} represent the input-referred and output total noise



where R_F and C_F are the feedback elements, v_{OUT} and i_{IN} are the output and input elements of the TIA, and f is the system frequency. The feedback capacitance acts as a low-pass filter, ensuring interface circuit stability in higher frequencies given the presence of an input capacitance C_{IN} [22]. A detailed noise model of the Fig. 1.4 circuit is shown in Fig. 1.5 where C_{IN} represents the sum of the equivalent nano-pore capacitance (C_M from Fig. 1.3) and any parasitic capacitance at the input.

The total noise observed at the input, i_{IN} , is composed of three main elements:

$$\overline{i_{IN}^2} = \overline{i_D^2} + \overline{i_F^2} + \overline{i_{op}^2} \quad (1.4)$$

where i_D^2 represents noise created by environmental elements and the sensor itself, i_F^2 represents the noise created by the feedback loop, and i_{op}^2 is the input-referred noise created by the op-amp circuitry e_{op}^2 . For resistive feedback, the term i_F^2 can be written as:

$$\overline{i_F^2} = \frac{4kT}{R_F} \quad (1.5)$$

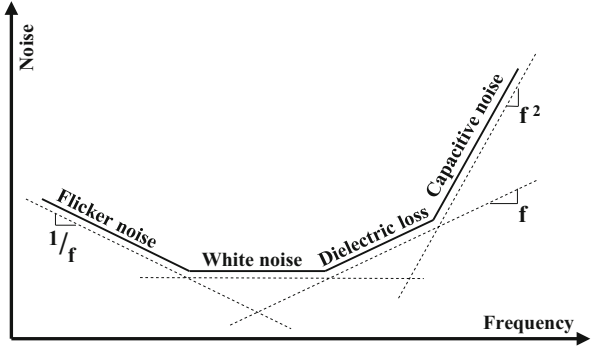
where k is the Boltzmann constant and T is temperature in Kelvin. If $\overline{v_{op-out}^2}$ is the total output voltage noise created by e_{op}^2 , it can be inferred that:

$$\overline{v_{n-out}^2} = \overline{e_{op}^2} \times \left| 1 + \frac{Z_F}{Z_{IN}} \right| = \overline{e_{op}^2} \times \frac{(R_{IN} + R_F)^2 + 4\pi^2 f^2 R_F^2 R_{IN}^2 (C_F + C_{IN})^2}{R_{IN}^2 (1 + 4\pi^2 f^2 R_F^2 R_{IN}^2)} \quad (1.6)$$

where Z_F and Z_{IN} represent the total feedback and input impedance seen on the feedback loop and input node, respectively. (1.6) can be simplified to:

$$\overline{v_{n-out}^2} = \overline{e_{op}^2} \times \frac{\left(1 + \frac{R_F}{R_{IN}} \right)^2 + 4\pi^2 f^2 R_F^2 (C_F + C_{IN})^2}{1 + 4\pi^2 f^2 R_F^2 R_{IN}^2} \quad (1.7)$$

Fig. 1.6 Noise spectrum and the contributing elements observed with nano-pore interface circuits (Adapted from Rosenstein and Shepard [23])



where the term with R_F/R_{IN} dictates the interface circuit noise performance at lower frequencies. Hence, a high input resistance is very beneficial to the interface circuit at lower frequencies. For nano-pore interfaces, R_{IN} can generally be assumed to be much greater than R_F , which permits the input-referred current noise generated by the op-amp to be expressed as:

$$\overline{i_{op}^2} = \frac{\overline{v_{op-out}^2}}{|Z_F|^2} = \overline{e_{op}^2} \times \left[\frac{1}{R_F^2} + (2\pi f)^2 (C_F + C_{IN})^2 \right] \quad (1.8)$$

Thus, the total input-referred noise at input i_{IN} is:

$$\overline{i_{IN}^2} = \overline{i_D^2} + \frac{4kT}{R_F} + \overline{e_{op}^2} \times \left[\frac{1}{R_F^2} + (2\pi f)^2 (C_F + C_{IN})^2 \right] \quad (1.9)$$

The noise spectrum seen at the output of a typical TIA system interfacing with a nano-pore is depicted in Fig. 1.6 [23]. The flicker noise and dielectric loss are produced by the input capacitance. The dielectric loss noise is due to thermal dissipation in lossy dielectric materials [23] which has a direct correlation with frequency. The white noise and the capacitive noise are, however, caused by $\overline{e_{op}^2}$. The interface circuit bandwidth should be on the order of hundreds of kHz, which means that the capacitive noise created by the terms containing f^2 will definitely hinder the sensitivity of the interface circuit and is a common concern in the nano-pore interface circuit literature [19].

Based on (1.9), the two best ways to solve the capacitive noise issue using a TIA are to decrease $\overline{e_{op}^2}$ or to decrease the overall capacitance seen at the TIA input. The term $\overline{e_{op}^2}$ can be improved by dedicating more circuit area and/or power to the TIA or using novel op-amp techniques. Decreasing the capacitance can usually be achieved by system miniaturization. Generally, miniaturization of sensor area and electrode dimensions is desirable because it reduces the parasitic capacitances as well as the double-layer capacitance formed on the electrode [19]. Furthermore, reducing sensor dimensions decreases the impact of electrolyte resistance on amperometric instrument noise performance [24].

2.2.2 Other Limitations for Amperometric Sensing

Noise performance is one of the main challenges to interfacing nano-pores, but other major issues must also be addressed. One of the main issues is input offset currents. As mentioned before, nano-pores monitor ionic current flow and variations in current due to the presence of target molecules. The background ionic current can have constant or variable amplitude and can be as much as $10\times$ the amplitude of sensing event currents. This background current creates an input offset that can challenge coherent sensing of the nano-pore event. Most modern amperometric interface circuits utilize a TIA with capacitive feedback due to its simplicity, low use of hardware resources, and good noise performance [19]. However, a large input offset can saturate a basic TIA interface circuit. For example, a large input offset will force a switched-capacitor interface circuit to dramatically increase its clock rate, compromising the SNR performance of the system.

Recently, there has been a push to implement arrays of nano-pores in the form of self-forming ion channels with a microfluidic electrolyte delivery system [25–27]. Interfacing multiple nano-pores at the same time creates restrictions on the area and power consumption of each readout circuit channel in the array [20]. However, this creates a critical design trade-off because one of the best methods to improve noise performance, as described above, is to decrease $\overline{e_{\text{op}}^2}$, which can be achieved by increasing the TIA op-amp's area and/or power rather than reducing it. Hence, addressing this trade-off for array implementations is one of the ongoing challenges in the field of nano-pore sensing.

2.3 Nano-Pore Interface Circuits

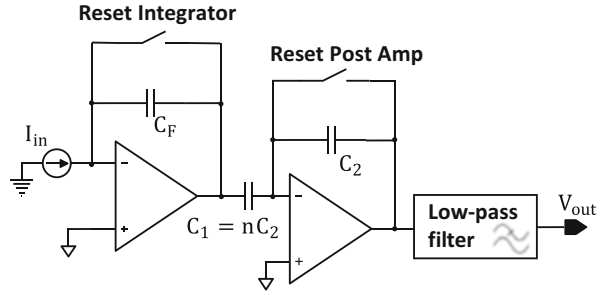
Although the development of interface circuits for both solid-state and ion channel nano-pores is a relatively new field of study [19], several prominent works have already been reported in the literature with a focus on three main issues: offset cancellation, noise performance, and array implementation. Based on (1.9), there are only a few options for improving noise performance in a classic TIA structure. However, many approaches have been reported to improve noise by stepping away from the classic TIA structure. To begin analyzing these, let's express (1.9) in a simplified and more generalized form as:

$$\overline{i_{\text{IN}}^2} = \overline{i_{\text{D}}^2} + \overline{i_{\text{F}}^2} + \overline{e_{\text{op}}^2} \times \left[\frac{1}{R_{\text{F}}^2} + (2\pi f)^2 (C_{\text{IN}})^2 \right] \quad (1.10)$$

Based on (1.10), the methods for improving nano-pore interfacing noise can be classified as:

- Improving $\overline{i_{\text{F}}^2}$ and increasing R_{F}^2 by enhancing feedback constructs such as capacitive feedback, enhanced feedback, and active feedback amplifiers

Fig. 1.7 Capacitive interfacing of amperometric current using resetting switches (Adapted from Goldstein et al. [28])



- Decreasing total input-referred noise of the op-am ($\overline{e_{op}^2}$)
- Decreasing total input capacitance, C_{IN} , and $\overline{i_{ID}^2}$ by using a lab-on-CMOS approach

The capacitive feedback approach forgoes the feedback resistor to avoid resistive feedback noise, effectively making the value of R_F approach infinity at DC. Assuming that R_{IN} also has a large value, capacitive feedback would decrease the noise floor significantly and save the circuit area of a large feedback resistor. Note that the feedback capacitance should be significantly lower than the total input capacitance to avoid introducing another noise performance factor. The area saved by eliminating the feedback resistor can be utilized for noise reduction on the CMOS circuitry. However, implementing capacitive feedback will leave the interface vulnerable to input DC offset signals that could saturate the op-amp if left unchecked [19]. One solution to this issue is to incorporate an input current DC offset-canceling mechanism as shown in Fig. 1.7 [28]. This is an integrator structure with periodic reset switching to stop the integrator from saturating, followed by a gain stage and a low-pass filter. The circuit was designed for measuring DNA strands passing through a silicon-based nano-pore implemented in a 0.5 μm bulk CMOS technology. While this circuit displays exemplary performance for its intended purpose, the periodic reset limits its bandwidth to a maximum of 10 kHz, which is not ideal for nano-pores with faster transition times. This integrating structure consumes 1.5 mW of power and 0.1386 mm^2 for each channel.

A popular method for enhancing the basic TIA noise performance is to use a feedback with a relatively small resistance coupled with an active current attenuator as shown in Fig. 1.8. The feedback current i_{out} is produced according to:

$$i_{out} = \frac{i_{in}}{M} = \frac{v_{out}}{MR_{base}} \quad (1.11)$$

where M is the attenuation factor shown in Fig. 1.8, i_{out} is the feedback current, v_{out} is the TIA output voltage, R_{base} is the small feedback resistance, and i_{in} is the current passing through R_{base} . By using the attenuation factor M , the total feedback resistance seen by the op-amp is amplified by M , while its input-referred noise power is decreased by a factor of M^2 . Hence, a 10 $\text{k}\Omega$ resistance with $M = 100$

Fig. 1.8 Nano-pore current interface with enhanced feedback resistors (Adapted from Ferrari et al. [29])

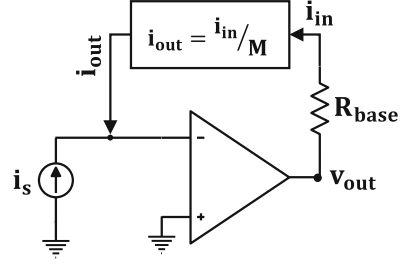
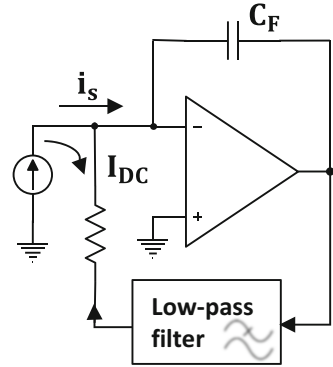


Fig. 1.9 Capacitive feedback with resistive DC input offset cancellation (Adapted from Ferrari et al. [29])



would act as a $1 \text{ M}\Omega$ feedback resistance while demonstrating an input-referred current noise (where noise $\propto 4KT/R$) of a $100 \text{ M}\Omega$ resistor [29]. However, the amplified resistance lacks the accuracy required to sense nano-pore events [29] and is thus usually used in conjunction with a feedback capacitor, which makes it susceptible to input offset currents.

Figure 1.9 shows an alternative feedback loop with an extremely low bandwidth low-pass filter that absorbs any incoming DC signal and offset, leaving the higher-frequency signal currents for the capacitance that has no noise contribution to the interface circuit [29]. The resistance in the feedback is small, but it does not have noise contributions at higher frequencies since it is deactivated at higher frequencies through the low-pass filter.

An advanced version of the circuit in Fig. 1.9 is shown in Fig. 1.10 [29]. The block dubbed $H(s)$ filters out all the AC components leaving the R_{dc} portion of the circuit with only the DC offset input signal. R_{dc} is composed of R_{base} and an attenuator circuit described above. The attenuation factor is chosen by the size ratio of the transistors T_M and T_I and can be controlled with a high accuracy. Hence, this resistive feedback has minimal effect on interface circuit noise performance and effectively removes the R_F and $\overline{i_F^2}$ portions of (1.10) while negating the effects of input DC offset.

The enhanced resistive feedback structure has significant benefits in terms of noise performance. The capacitive feedback does not require continuous resetting, which means that the only limit on interface circuit bandwidth is imposed by the

Fig. 1.10 An enhanced capacitive feedback nano-pore interface circuit (Adapted from Ferrari et al. [29])

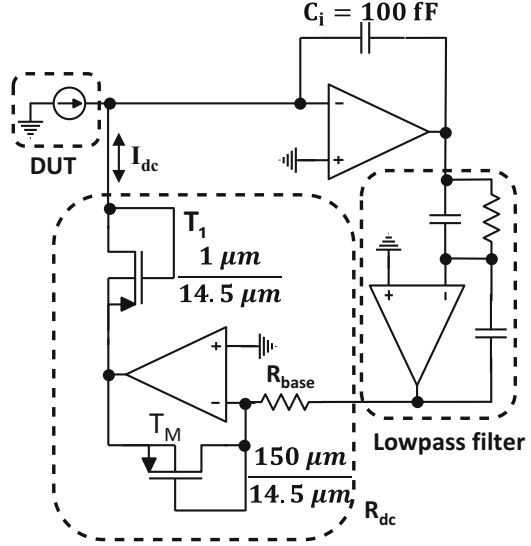
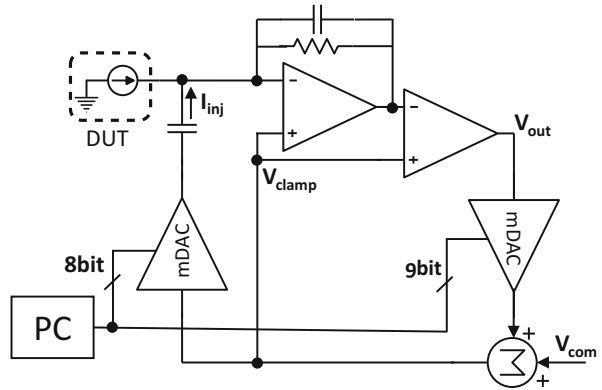


Fig. 1.11 Interface circuit with enhanced capacitive feedback for DC offset compensation (Adapted from Weerakoon et al. [30])



noise floor. This is evident by a bandwidth of up to few MHz [29]. However, this circuit requires 45 mW to operate a single channel and consumes an area of 0.35 mm^2 in a $0.35 \mu\text{m}$ CMOS technology.

An alternative approach to enhance resistive feedback with capacitive feedback is depicted in Fig. 1.11 [30]. This circuit was designed to interface an array of patch-clamp bundles together in a massive package. The input capacitance is driven by a DAC and generates an injection current, I_{inj} , that mitigates the input DC current. The implemented feedback loop consists of multiple DACs and two amplifier stages for stability and thus requires significant hardware resources, area, and power.

Fully active feedback control that eliminates the need for spacious passive components is another approach to solving noise performance and input offset cancelation issues while also reducing effective interface circuit area. The fully active feedback presented in [31] is depicted in Fig. 1.12. Ignoring the external

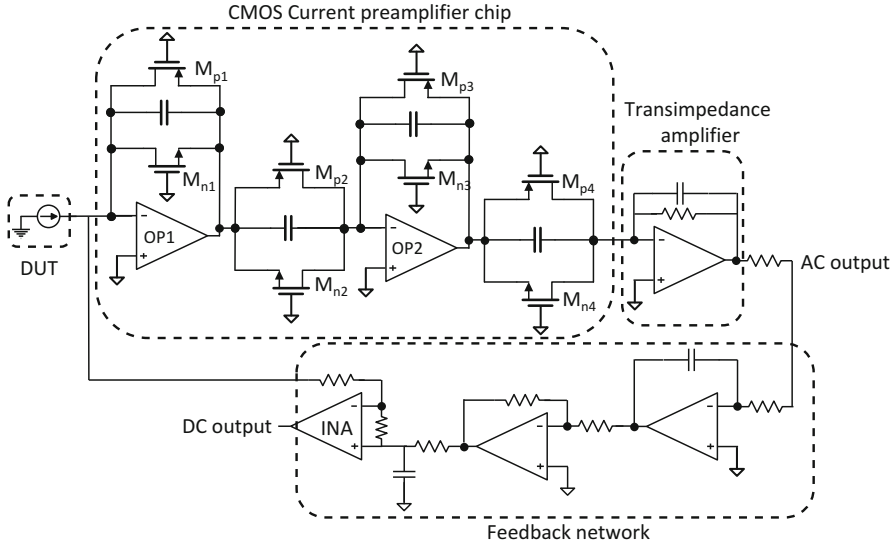


Fig. 1.12 Fully active feedback structure with external feedback structure for DC offset cancellation (Adapted from Ciccarella et al. [31])

feedback network for now, one can see that OP1 has a fully active feedback loop consisting of M_{p1} , M_{n1} , and a very small capacitance. Together with M_{p2} and M_{n2} , this architecture acts as a current buffer. The input feedback transistors M_{p1} and M_{n1} have their gates set to ground, so they are usually off and do not produce any noise. The slightest input current would force the OP1 output voltage to react and pass this current through one of the feedback transistors operating in subthreshold region. The same voltage is also applied at M_{p2} and M_{n2} , which have channel widths M times larger than the feedback pair and thus would conduct a current M times greater than the input current. Because these transistors never fully turn on, they have very little effect on the noise floor. The capacitances are included to ensure stable gain at higher frequencies. Due to CMOS implementation, this feedback structure consumes less area than a resistive feedback or a capacitive feedback with a switching mechanism. The external feedback loop is a similar to that presented in [29] and removes the DC input component. However, the structure in Fig. 1.12 also produces an output, DC_{out} , equal to the input DC current.

In other work, a TIA was implemented using transistors in feedback loop that mimics a tunable resistance and achieves a large resistive feedback while reducing size significantly [32]. The downside of this structure is that, for an array implementation, it would require manual tuning to set the desired feedback resistance in each channel. Furthermore, this architecture requires special deep N-well CMOS technology to enable control of the resistance through the transistor's bulk voltage that has an adverse impact on circuit noise. As a result, this interface circuit has a 3pA RMS noise at 5 kHz frequency.

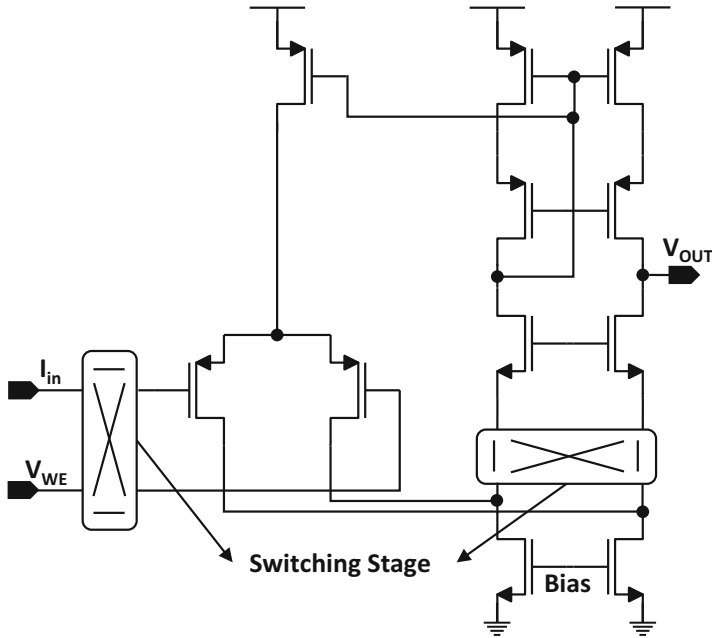


Fig. 1.13 Op-amp with internal chopper circuit for flicker noise cancellation (Adapted from Jafari and Genov [33])

The discussion above has focused on enhancing performance of the feedback loop rather than that of the op-amp itself. Despite the fact that op-amp performance has received significant attention in the literature, many op-amp designs have been reported for specific bioelectronic applications. Among these, an op-amp with internal chopper stabilization was reported to mitigate flicker noise at the op-amp's input stage [33]. As shown in Fig. 1.13, the input stage was separated from the rest of the op-amp using a switching stage that alternates the input positive and negative poles. As a result, the flicker noise is chopped and mitigated leaving the interface circuit with an RMS noise of 0.07 pA at 1 kHz bandwidth with a small area and power consumption. The main limitation with this structure is that it cannot achieve high bandwidths due to the intense switching required. Furthermore, the switching impairs the input capacitive noise by increasing the overall op-amp input-referred white noise which, in turn, yields capacitive noise at higher frequencies.

Another effort to improve op-amp performance utilizes a matched pair of ultra-low noise JFETs in the input stage [34]. Together with a massive bias current, the interface circuit is able to tolerate input capacitances as large as 100 pF with an RMS noise of 5 pA at a 25 kHz bandwidth. This interface circuit was implemented with commercial off-the-shelf (COTS) components and would produce extra parasitic capacitances when integrated with a nano-pore interface. Fortunately, the interface circuit has a very robust noise performance and should be able to tolerate the extra capacitance.

2.4 *Multichannel Nano-Pore Arrays*

Due to recent advances in microfabrication and biological science techniques, a trend has developed to use microarrays for parallel sensing [35, 36]. These structures accelerate the development time by quickly extracting valuable experimental data. Similarly, combining nano-pores with microfluidic delivery systems offers great potential for enabling parallel recording of biological signals. Integrated arrays are especially suitable for ion channels formed on a BLM structure due to their self-assembling nature that enables them to automatically form in designated spots within a microfluidic system [37]. Parallel array implementation is also appealing to overcome the inherently low yield in forming functional nano-pore interfaces, especially ion channels. The patch-clamp extraction method for ion channels is very laborious and time consuming with low odds of extracting the correct target ion channels. Even using synthetic BLMs, the chances of successfully forming an interface with the desired number of functional ion channels are very low [38]. Hence, a logical use for self-forming ion channel nano-pores would be in the form of an array where a large number of nano-pore are implemented simultaneously, increasing the odds of successful formation of ion channels.

Electrically interfacing with arrays of nano-pore introduces a new set of design challenges that need to be addressed. The works presented in [26, 27] demonstrate parallel recording functionality with very limited number of array channels and sub kHz bandwidth. These systems are not able to detect single-molecule events due to the plethora of parasitic capacitances and lack of noise fine-tuning. These challenges could be overcome by integrating the sensing circuitry within a microfluidic system containing the nano-pores through an approach dubbed lab-on-CMOS [39]. Implementing array sensing by integrating CMOS interface circuitry in close proximity of a nano-pore array system [20] can greatly reduce environmental and wiring noise. However, this high level of integration sets severe limitations in terms of the power and area consumed by the CMOS nano-pore interface circuitry. These limitations are thoroughly discussed in [20, 40] which seek to implement arrays of hundreds of ion channel in a microfluidic system implemented on the surface of a CMOS interface chip. In both these works, each nano-pore is assigned to a pixel amplifier stage that is in very close physical proximity to mitigate the effects of excessive input capacitance. As a result, the pixel amplifiers have very strict limitations in terms of the area they consume. Furthermore, because this approach enables hundreds of pixel amplifier stages right beneath (microns away from) the ion channel sensors, the potential for the CMOS circuitry to heat or even thermally denature the nano-pores must be considered. As a result, strict power consumption limitations must be met by the CMOS pixel circuitry.

To realize a nano-pore array, the op-amps should work in parallel to simultaneously interface each channel of the array as depicted in Fig. 1.14a. A problem with this approach, however, is that the op-amps require a much larger surface area than the nano-pores and thus limit the density of the array that can be achieved. Looking at Fig. 1.14, it can be observed that all the op-amps have the same voltage

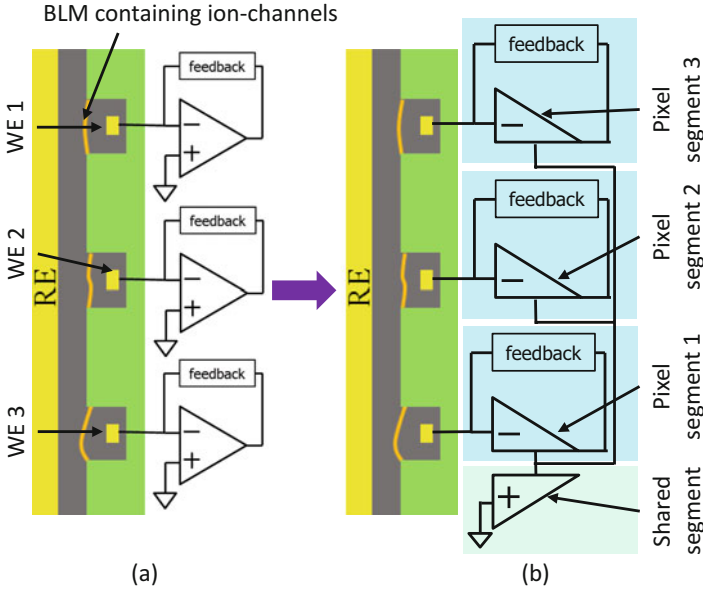


Fig. 1.14 Pixel interfaces for nano-pore current sensing with (a) typical TIA structure and (b) shared TIA structure. In (b), the circuit area shaded green is the shared segment of the op-amp with the positive input, and the blue-shaded areas are the gain stages with the negative inputs of the op-amp that are used for amplifying the signal

on their positive inputs. This voltage is the analog ground voltage that is shared globally throughout the whole CMOS interface chip. Stimulation of the ion channel containing BLMs is done through the RE electrode. Assuming all op-amps are the same, we can view each op-amp as being composed of a positive input half and a negative input half. Furthermore, because all of the op-amps have the same positive inputs, it is possible to share all of the positive op-amp halves so that one shared positive segment can serve an array of negative input halves. Defining the negative input halves for each input channel as a pixel segment results in the shared op-amp structure shown in Fig. 1.14b and utilized in [20, 40]. Many benefits arising from such configuration including reduced pixel area, reduced pixel power consumption, and the opportunity to direct the resources saved by this structure to enhancing noise performance.

The shared op-amp concept was first presented for biosensing in [41] where multiple op-amps in a TIA configuration were assigned to the same positive terminal input, as depicted in Fig. 1.15. The shared segment in Fig. 1.15 is essentially the positive side of n op-amps. The shared segment provides the pixel segments with a voltage, V_C , that helps maintain the semi-differential status of the interface circuit. Notice that the shared segment has an internal feedback loop, where the current source from M_{pb-1} can be considered to be the input of the loop and the output node is V_C . Hence, the transistors M_{ps} , M_3 , M_4 , M_5 , and M_6 are considered to be

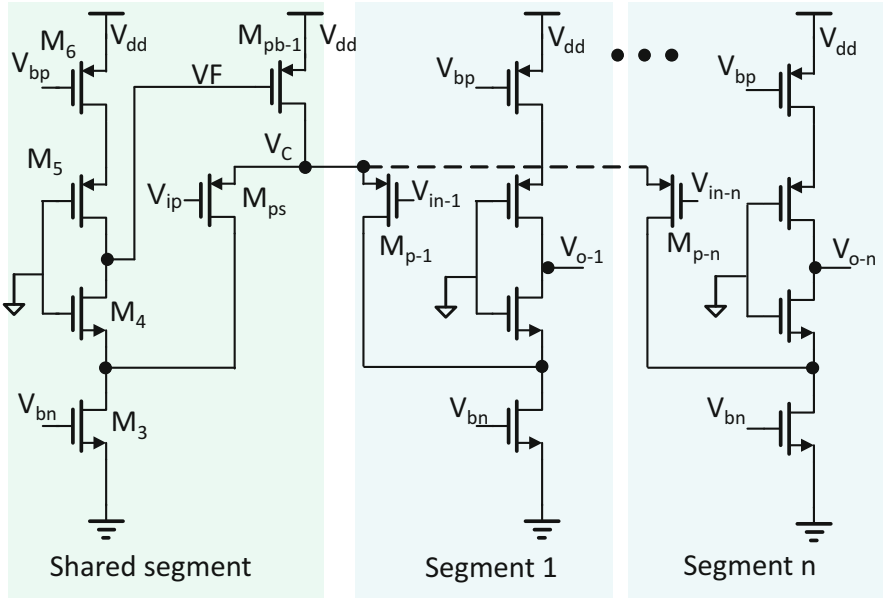


Fig. 1.15 Shared op-amp structure circuitry

the feedback components. The feedback type can be considered as a parallel output, parallel input with $\beta = g_c$. This means that the resistance seen at the common node is extremely small due to the fact that g_c and R_{in} have a large transconductance and resistance, respectively. Thus the voltage V_C is held constantly by feedback loop present at the shared segment. This is because V_C is a node with a very low resistance connected to a voltage source, and the resistance at V_C is so small that every pixel segment can act as a fully differential stage without affecting the other segments. As a result, each pixel segment sees a very small resistance at the source of their input transistors and operates as a single op-amp with their positive input fixed to a certain voltage.

This section has outlined the challenges regarding amperometric interfacing of nano-pores from a circuit prospective. It was suggested that the greatest challenge in interfacing nano-pores is achieving the noise performance and accuracy needed to detect very small variations in ionic currents through a nano-pore. Methods reported in the literature to address these issues were reviewed and discussed. Finally, new approaches for the array implementation of nano-pores were discussed, and the resulting area and power limitations on interface circuitry were also reviewed.

3 Future Prospects

As their name implies, nano-pores are naturally very small, taking an area less than $1 \mu\text{m}^2$. Furthermore, they operate within an electrochemical framework, which makes them suitable for portable sensing and personalized healthcare. However, operating nano-pore interfaces in real-world conditions would require new approaches in terms of establishing microfluidic systems, setting up a proper environment, and achieving demanding interface circuit requirements. Hence, advances in material, fabrication technology, and interfacing would put nano-pores in demand for emerging markets. A better understanding of nano-pore functionality would result in products that enable targeted drug release, point-of-care diagnosis, or toxin and heavy metal detection in nonmedical fields. Array-enabled molecular mass sensing is another emerging field for nano-pores. Better mass production techniques could result in nano-pore array solutions for DNA sequencing and rare bacteria detection. Now that the feasibility of nano-pore systems has been established, there is a need to break the nano-pore technology out of the academic realm and into emerging markets for electrochemical sensing.

4 Conclusions

In this chapter, the operation principles of nano-pores and their classifications were described. Challenges for interfacing individual nano-pores as well as arrays of nano-pore were explored, and solutions from literature were identified. Section 1 introduced the concept of nano-pore-enabled electrochemical sensing with a focus on molecule detection. Different nano-pore categories, namely, ion channels and solid-state nano-pores, were discussed and presented with a brief literature review. Section 2 introduced an equivalent circuit model for the behavior of a generic nano-pore device and introduced the limitations faced by interfacing nano-pores due to their noisy nature and challenging signal response. The main interface circuit challenges were identified as noise and input offset performance, and solutions available in literature were reviewed. Section 2 also discussed the challenges faced by interfacing arrays of nano-pores, namely, interface circuit area and power consumption, and some new approaches to overcome these challenges were reviewed.

References

1. G. Zheng, C.M. Lieber, Nanowire biosensors for label-free, real-time, ultrasensitive protein detection, in *Nanoproteomics: Methods and Protocols*, ed. by S. A. Toms, R. J. Weil (Humana Press, Totowa, 2011), pp. 223–237
2. J. Wang, Carbon-nanotube based electrochemical biosensors: a review. *Electroanalysis* **17**(1), 7–14 (2005)
3. J. Wang, Electrochemical biosensors: towards point-of-care cancer diagnostics. *Biosens. Bioelectron.* **21**(10), 1887–1892 (2006)
4. D. Zhang, Q. Liu, Biosensors and bioelectronics on smartphone for portable biochemical detection. *Biosens. Bioelectron.* **75**, 273–284 (2016)
5. H. Li et al., CMOS electrochemical instrumentation for biosensor microsystems: a review. *Sensors* **17**(1), 74 (2016)
6. X.-S. Zhou, E. Maisonhaute, Electrochemistry to record single events, in *Electrochemistry: Volume 11 - Nanosystems Electrochemistry*, vol. 11, (The Royal Society of Chemistry, UK, 2013), pp. 1–33
7. K. Zhou, J.M. Perry, S.C. Jacobson, Transport and sensing in nanofluidic devices. *Annu. Rev. Anal. Chem.* **4**(1), 321–341 (2011)
8. C. Dekker, Solid-state nanopores. *Nat. Nanotechnol.* **2**(4), 209–215 (2007)
9. J.J. Kasianowicz et al., Nanoscopic porous sensors. *Annu. Rev. Anal. Chem.* **1**(1), 737–766 (2008)
10. M. Wanunu et al., Rapid electronic detection of probe-specific microRNAs using thin nanopore sensors. *Nat. Nanotechnol.* **5**(11), 807–814 (2010)
11. O.K. Dudko et al., Extracting kinetics from single-molecule force spectroscopy: nanopore unzipping of DNA hairpins. *Biophys. J.* **92**(12), 4188–4195 (2007)
12. K.M. Halverson et al., Anthrax biosensor, protective antigen ion channel asymmetric blockade. *J. Biol. Chem.* **280**(40), 34056–34062 (2005)
13. D.W. Deamer, D. Branton, Characterization of nucleic acids by nanopore analysis. *Acc. Chem. Res.* **35**(10), 817–825 (2002)
14. B. Hille et al., *Ion Channels of Excitable Membranes* (Sinauer, Sunderland, 2001)
15. W. Asghar et al., Shrinking of solid-state nanopores by direct thermal heating. *Nanoscale Res. Lett.* **6**(1), 372 (2011)
16. J.D. Uram, K. Ke, M. Mayer, Noise and bandwidth of current recordings from Submicrometer pores and nanopores. *ACS Nano* **2**(5), 857–872 (2008)
17. J. Li et al., Ion-beam sculpting at nanometre length scales. *Nature* **412**(6843), 166–169 (2001)
18. A.J. Storm et al., Fabrication of solid-state nanopores with single-nanometre precision. *Nat. Mater.* **2**, 537–540 (2003)
19. M. Crescentini et al., Noise limits of CMOS current interfaces for biosensors: a review. *IEEE Trans Biomed Circuits Syst* **8**(2), 278–292 (2014)
20. H. Li et al., Ultracompact microwatt CMOS current readout with picoampere noise and kilohertz bandwidth for biosensor arrays. *IEEE Trans Biomed Circuits Syst.* **12**(1), 35–46 (2018)
21. R.S. Martin et al., Recent developments in amperometric detection for microchip capillary electrophoresis. *Electrophoresis* **23**, 3667–3677 (2002)
22. A. Bhat, Stabilized TIAs key to reliable performance. *Electronic Design* (2011), p. 2. Available at: <http://electronicdesign.com/analog/stabilized-tias-key-reliable-performance>
23. J.K. Rosenstein, K.L. Shepard, Temporal resolution of nanopore sensor recordings. in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2013), pp. 4110–4113
24. D. Wei et al., Electrochemical biosensors at the nanoscale. *Lab Chip* **9**(15), 2123–2131 (2009)
25. X. Liu, L. Li, A.J. Mason, High throughput single-ion-channel array microsystem with CMOS instrumentation. in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), pp. 2765–2768

26. B. Le Pioufle et al., Lipid bilayer microarray for parallel recording of transmembrane ion currents. *Anal. Chem.* **80**(1), 328–332 (2008)
27. F. Thei et al., Parallel recording of single ion channels: a heterogeneous system approach. *IEEE Trans. Nanotechnol.* **9**(3), 295–302 (2010)
28. B. Goldstein et al., CMOS low current measurement system for biomedical applications. *IEEE Trans Biomed Circuits Syst* **6**(2), 111–119 (2012)
29. G. Ferrari et al., Transimpedance amplifier for high sensitivity current measurements on Nanodevices. *IEEE J. Solid State Circuits* **44**(5), 1609–1616 (2009)
30. P. Weerakoon et al., An integrated patch-clamp potentiostat with electrode compensation. *IEEE Trans Biomed Circuits Syst* **3**(2), 117–125 (2009)
31. P. Ciccarella et al., Integrated low-noise current amplifier for glass-based nanopore sensing. in *10th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)* (2014), pp. 1–4
32. J. Kim, K. Pedrotti, W.B. Dunbar, An area-efficient low-noise CMOS DNA detection sensor for multichannel nanopore applications. *Sensors Actuators B Chem.* **176**, 1051–1055 (2013)
33. H.M. Jafari, R. Genov, Chopper-stabilized bidirectional current acquisition circuits for electrochemical Amperometric biosensors. *IEEE Trans Circuits Syst I Regul Pap* **60**(5), 1149–1157 (2013)
34. M. Carminati et al., Design and characterization of a current sensing platform for silicon-based nanopores with integrated tunneling nanoelectrodes. *Analog Integr. Circ. Sig. Process* **77**(3), 333–343 (2013)
35. B.B. Haab, M.J. Dunham, P.O. Brown, Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol.* **2**(2), research0004.1 (2001)
36. M.F. Lopez, M.G. Pluskal, Protein micro- and macroarrays: digitizing the proteome. *J. Chromatogr. B* **787**(1), 19–27 (2003)
37. X. Liu, *CMOS Instrumentation for Electrochemical Biosensor Array Microsystems* (Michigan State University. Electrical Engineering 2014)
38. L. Li, A. Mason, Development of an integrated CMOS-microfluidic instrumentation array for high throughput membrane protein studies. in *IEEE International Symposium on Circuits and Systems* (2014), pp. 638–641
39. Y. Huang, A.J. Mason, Lab-on-CMOS integration of microfluidics and electrochemical sensors. *Lab Chip* **13**(19), 3929–3934 (2013)
40. S. Parsnejad, H. Li, A.J. Mason, Compact CMOS amperometric readout for nanopore arrays in high throughput lab-on-CMOS. in *Proceedings – IEEE International Symposium on Circuits and Systems* (2016) July, pp. 2851–2854
41. S. Ayers et al., Design of a CMOS Potentiostat circuit for electrochemical detector arrays. *IEEE Trans Circuits Syst I Regul Pap* **54**(4), 736–744 (2007)

Chapter 2

Metabolomics on CMOS for Personalised Medicine

Boon Chong Cheah and David R. S. Cumming

1 Introduction: Personalised Medicine

Conventional healthcare approaches rely on the diagnosis of a certain disease and provide identical treatment to a large population of people, also known as “one drug fits all” [1]. However, each individual is unique, and the progression of any disease varies from patient to patient, which leads to different responses to treatment for each person. The genetic makeup of all humanity is 99.1% identical, and only a small fraction of an individual’s genetic variation (of 0.9%) accounts for the large variation we see in human beings [2]. In the context of medicine, this is seen in heritable diseases in certain families or ethnicities and variation in the manifestation of diseases. Medical practice has therefore undergone a shift towards tailoring treatment to suit an individual’s needs, taking into consideration both genetic and environmental factors, which is known as personalised medicine. Figure 2.1 shows a diagram comparing conventional medicine and personalised medicine. The concept of personalised medicine was introduced by the development of genetic medicine with the observation by Garrod [4] in 1931 that interindividual variations to drugs are due to each person’s genetic constitution. The subsequent discovery of the double-stranded structure of the DNA by Watson and Crick [5] in 1953 has made it possible to provide detailed descriptions of how an individual’s genetics is involved with disease. The discovery of polymerase chain reaction by Mullis [6] in 1986 later facilitated the study of the genome. In the middle of the twentieth century, Brenner described the correlation between DNA and proteins and identified the first gene-disease association, known as Huntington’s disease that is now classified as Mendelian inheritance [2, 7]. However, many common diseases

B.C. Cheah • D.R.S. Cumming (✉)
School of Engineering, University of Glasgow, Glasgow, UK
e-mail: BoonChong.Cheah@glasgow.ac.uk; David.Cumming.2@glasgow.ac.uk

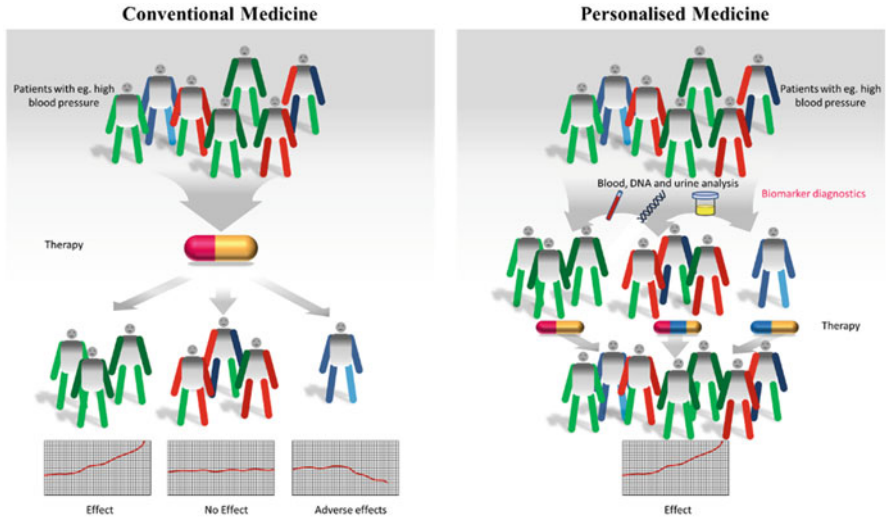


Fig. 2.1 A diagram showing the difference between conventional medicine and personalised medicine. The same treatment offers to a heterogeneous population may have no effect or even adverse effects in some patients. Therefore, an accurate diagnosis is required for everyone to provide effective treatments to all (Adapted from Bayer Healthcare [3])

involve multiple genes and environmental interaction, called multifactorial diseases. Hence, modern medicine has now entered into the era of genomics, made possible by the advancement of parallel and high-throughput sequencing technologies that enabled the sequencing of human genome in 2000 [8]. This offers medicine the possibility of tailoring optimal treatment based on an individual patient's genome.

Even with the tremendous progress made by genomics, it alone is inadequate to diagnose specific disease in an individual with absolute certainty; hence other omics such as transcriptomics, proteomics and metabolomics have spawned that will further facilitate the development of personalised medicine [9]. Metabolomics is of particular interest since it bridges the gap between the genotype and phenotype. The metabolome contains many biochemical reactions [10], and the metabolic chain of biochemical events provides a window on human health determined by the genes, proteins and environmental factors. Metabolomics is the study of the metabolome, which is the set of all metabolites in human cells, tissues or organs. It is a powerful tool that indicates the underlying disease process for an individual's pathology at any given time and can potentially provide an early indication of disease. The analytical platforms for metabolome detection today rely on large and expensive equipment such as nuclear magnetic resonance (NMR) spectroscopy, gas chromatography-mass spectroscopy (GC-MS) and liquid chromatography-mass spectroscopy (LC-MS) [11, 12]. All these analytical instruments have their specific strength in detecting certain groups of metabolic reactions. For example, NMR is used in toxicology and GC-MS is used in lipidomics. Currently, there is no single

analytical platform that can detect the entire metabolome as efficiently as it is now possible to determine an individual's genome. Gene sequencing is simplified by the homogeneity of the chemical reactions required to collect data. Whilst much more heterogeneous, enzyme-based reactions offer an attractive vehicle for measuring a broad range of metabolites.

2 Enzymes: Metabolite Detection

Enzymes are biological catalysts that are found in all life forms, from viruses to mammals. Their role is to maintain the many chemical reactions required to sustain life [13], and as a consequence they are involved in many metabolic pathways. These chemical reactions involve building, breaking and modifying chemical compounds, and enzymes are used to accelerate the rate of these reactions, a process called catalysis. Furthermore, enzymes have a unique characteristic called specificity, where they are able to recognise precise chemical compounds and their binding sites to catalyse chemical reaction, while the enzyme remains intact. As shown in Fig. 2.2, an enzyme is biologically shaped as a “lock” (active sites), and only the correct “keys” (substrates) can be attached for a reaction to occur [14]. Since the discovery of enzymes, they have been heavily used in industrial processes for foods [15], beverages [16] and detergents [17]. More importantly, enzymes are exploited as biosensing recognition elements, in conjunction with different analytical methods such as spectrophotometry, electrochemistry or fluorescence, to quantify metabolite levels. Therefore, enzymes can be used as a high-throughput metabolite quantification platform for metabolomic applications.

Kinetic studies are normally used to characterise all chemical reactions, both non-catalysed and catalysed. The most commonly asked question is the variation of both the substrate and product over time. The chemical reaction for converting substrate to product goes through different chemical states, and the highest energy

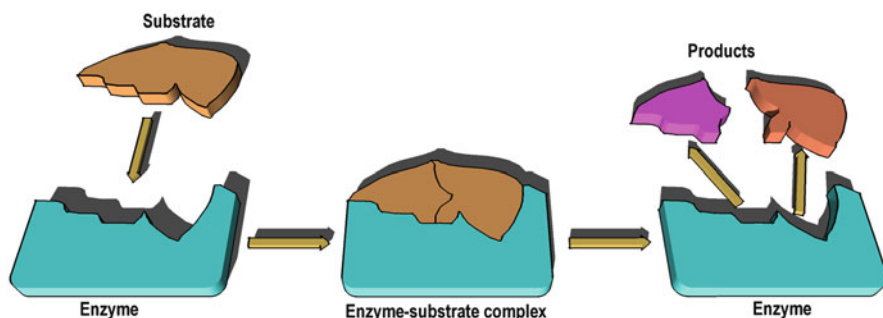


Fig. 2.2 The process of breaking a chemical compound during an enzyme reaction, described as a “lock and key” model

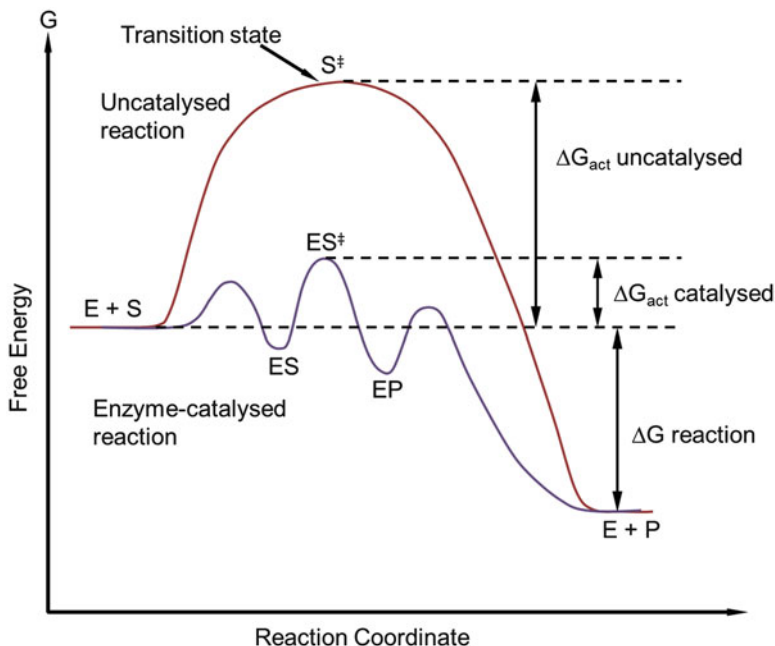


Fig. 2.3 The free energy profile of a non-catalysed (*red*) and an enzyme-catalysed reaction (*purple*), known as a reaction coordinate diagram (Reproduced from Bugg [18])

in between this transformation is called transition state, which is the limiting factor of a chemical reaction. The energy difference between the substrate and the product shows which direction of the reaction is thermodynamically favourable. Furthermore, the free energy between the substrate and the transition state, which is free energy of activation, determines the speed of the reaction. In an enzyme reaction, the enzyme acts as a catalyst to stabilise the transition state or find a lower energy pathway, to speed up the reaction. As can be seen in Fig. 2.3, an enzyme reaction involves multistep progression with intermediates and a transition state. An intermediate is a stable or semi-stable chemical species known as local energy minimum, whereas a transition state is known as local energy maximum. In an enzyme reaction, enzyme and substrate will come together and form a reversible enzyme-substrate complex. The enzyme will provide functional groups in the active sites to bind the substrate more tightly, forming a stabilised enzyme-transition state complex [18]. The more stable the enzyme-substrate complex, the faster the reaction. When the reaction is completed, an enzyme-product complex is finally formed and dissociated into product molecules and free enzyme.

All enzymes have different catalytic power, which is defined as the maximum number of substrate molecules converted to product by one enzyme molecule per second, under the conditions at which the substrate is saturating. Hence, all enzymes have different rate of conversion of substrate to product. In order to

describe the kinetics of the complex enzyme reaction, a simple chemical kinetic equation is insufficient to fully describe the reaction. In 1913, Michaelis and Menten [19] analysed the enzyme reactions with a mathematical model that is known as Michaelis-Menten equation, and it is given as below:

$$v = v_{\max} \frac{[S]}{K_m + [S]} \quad (2.1)$$

where v is the velocity of product formation or substrate depletion, v_{\max} is the maximal reaction rate, K_m is the Michaelis-Menten constant and $[S]$ is the substrate concentration. K_m is normally used as a relative measure of substrate-binding affinity for a specific enzyme:

$$[E \bullet S] = [E] \times \frac{[S]}{K_m} \quad (2.2)$$

where $[E \bullet S]$, $[E]$ and $[S]$ are enzyme-substrate complex, enzyme and substrate concentration, respectively.

From Eqs. (2.1) and (2.2), it becomes clear that when the substrate concentration is equal to K_m , the concentration of enzyme-substrate complex is equivalent to the concentration of enzyme. This therefore shows that K_m is the substrate concentration at which the enzyme is half saturated. In addition, K_m also shows the substrate concentration at which the reaction rate is at half the maximum velocity of the enzyme turnover of the substrate. When the substrate concentration is a lot higher than enzyme concentration, the upper limit of v is reached, and all enzyme molecules are present in the form of enzyme-substrate complexes; hence v_{\max} can be given by:

$$v_{\max} = k_{\text{cat}} \times [E_T] \quad (2.3)$$

where k_{cat} is the enzyme turnover number or catalytic power and $[E_T]$ is the total enzyme concentration.

Not all enzyme reactions follow the Michaelis-Menten model because there are a few assumptions that have to be made [20]:

- There is only one substrate change in the enzyme reaction.
- The substrate concentration is a lot higher than enzyme concentration.
- Only the initial reaction rate is taken in account when the product concentration is still low.
- The course of observation of the enzyme reaction is very short and only occurs when the change of substrate and product concentration is very low.

Nevertheless, Michaelis-Menten model provided a useful model to understand a wide range of enzymes.

Using Michaelis-Menten equation, it is clearly noted that at zero substrate concentration, the reaction rate is zero. When the substrate concentration doubles,

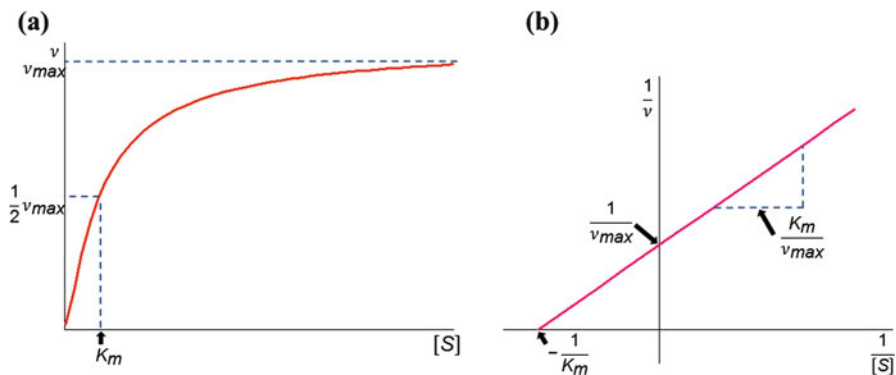


Fig. 2.4 (a) The typical reaction rate versus substrate concentration plot. (b) The double reciprocal Lineweaver-Burk plot. Both plots are a graphical representation of the relationship between enzyme reaction rate and substrate concentration (Reproduced from Meisenberg and Simmons [20])

the reaction rate doubles in proportion to the substrate concentration. When the substrate concentration reaches K_m , the velocity is half of the maximal reaction rate. At a high substrate concentration, the enzyme will be saturated with substrate, and the reaction rate will approach v_{max} . The relationship between reaction rate and substrate concentration can be seen in Fig. 2.4. By understanding the enzyme kinetics of enzymes, the rate, sensitivity and dynamic range of biosensing and diagnostic devices can be considerably improved, providing a powerful tool for high-throughput detection of many metabolites.

3 Microarray Technologies

Microarray technologies provided the solution for high-throughput assay techniques for diagnostic technologies. Microarrays are multiplexing technologies that enable parallel analysis of different substances using a very small amount of sample material in one single measurement. Normally, a microarray contains microspots of biological probing molecules that are functionalised on a surface in rows and columns. Specific target molecules will bind to the corresponding probing molecules and produce a signal that could be detected by various sensing technologies. The first microarray was implemented by Roger Ekins [21] in 1989 to perform immunoassays, but it was not taken up by users. With the advent of genomics, a rapid, low-cost and high-throughput sequencing technique was required to increase the number of genomes being sequenced a year. With the advancement in large-scale synthesis of oligonucleotide probes using polymerase chain reaction (PCR), a microarray of oligonucleotides was invented to sequence large pieces of DNA via hybridisation. DNA is a uniform and stable molecule, and its primary nucleotide

base sequence binds to its complementary target DNA with very good specificity [22]. Hence, the microarray also known as the “DNA chip” has made sequencing on a large scale possible. The number of genomes sequenced per annum is now expected to reach into the hundreds of millions in the near future [23].

However, microarray technologies for complex molecules such as enzymes are not well established. The challenge of developing a microarray of enzymes is due to the fact that the primary amino acid sequence of an enzyme does not determine the affinity of an enzyme to target molecules. It depends on the molecular structure of an enzyme, its “folding”, and the possibility of several different ways of interaction with the target molecules. Similar to how microarray technologies are used to revolutionise genomics, it is now expected that microarrays will be used to produce large-scale assessment of different metabolites. The following sections will review the technologies that are currently used for developing planar DNA microarrays and also review on the adoption of some of these technologies to create an enzyme microarray for metabolomics. The focus will be on printed microarrays owing to low-cost, flexible and mass-manufacturing capability that is well suited for integration with CMOS technology.

3.1 Printed Microarrays

Printed microarrays were the first to be used in research laboratories [24]. Printing normally uses either a contact or non-contact method to produce microspots of probing molecules on top of a planar solid surface. In contact printing, the print head reaches the surface to deliver the droplets. Non-contact printing uses the same technique as a computer printer (bubble-jet or inkjet) to disperse droplets with a gap between the print head and the surface. Both printing techniques use only nanolitre volumes of probe molecule solution. Figure 2.5 shows the difference between contact and non-contact printing. These microarrays are very low cost and extremely simple to produce. Furthermore, they are very flexible in how they can be used to create microarrays of different probing molecules. The only disadvantage of a printed microarray is cross-contamination, where a cleanroom environment is required to maintain the integrity of the biological samples while performing the printing.

There are two types of printed DNA microarrays:

- (a) *Double-stranded DNA (dsDNA) microarrays*: A dsDNA from a known genomic sequence is printed on the surface and followed by a step to denature the double-stranded DNA into a single-stranded DNA that can be readily used to hybridise with its complementary single-stranded DNA. This method normally produces 200–800 base pairs of DNA. Generally, the hybridisation signal strength and sensitivity increases with longer DNA probes but suffers from random hybridisation to nontargeted sequence that leads to poor specificity. Hence, this DNA microarray has high sensitivity but suffers in specificity.

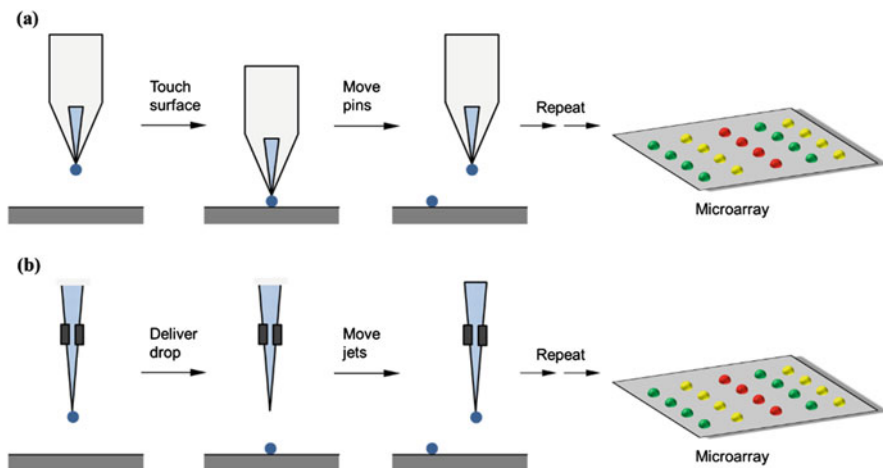


Fig. 2.5 A diagram showing the working mechanism of a (a) contact printing and (b) non-contact printing (Reproduced from Schena et al. [25])

(b) *Oligonucleotide microarrays*: A DNA microarray made up from chemically synthesised oligonucleotides, which are very short. Typically, it produces 25–80 base pairs of DNA. Shorter DNA probes reduce the errors in synthesis, and more specific genomic sequence can be investigated with smaller mismatch tolerance, leading to a very specific DNA microarray.

Before an entire DNA microarray is printed, the solid surface (typically a glass slide) has to be treated to allow immobilisation of DNA at the particular position. The glass slide can be coated with a layer of positively charged polymer to enable electrostatic interaction with the negative charge of the phosphate group in the DNA. Another method is to salinise the glass slide to produce positively charged amine groups and enable covalent bonding via UV cross-linking with the thymidine bases in the DNA. In addition, a glass slide often can be treated with aldehyde to produce carboxylic groups for covalent binding of amino groups at the modified ends of oligonucleotides.

By adopting the printed DNA microarray technology, MacBeath et al. [26] demonstrated the first protein microarray using specific enzymes to detect different small molecules. They used a high-precision contact-printing robot to produce an enzyme microarray. The printing enzyme solution was mixed with 40% glycerol to prevent evaporation when the enzymes were being functionalised on the surface. The immobilisation technique for the enzyme was similar to the coupling of oligonucleotide to a glass slide surface, where they used an aldehyde-treated surface to react with the amine groups in the enzyme. The method allows the enzymes to attach to the surface with a variety of orientations and without disrupting the ability of the enzymes to interact with the targeted small molecules. Bovine serum albumin (BSA) was also used to quench the unreacted aldehyde-treated surface to

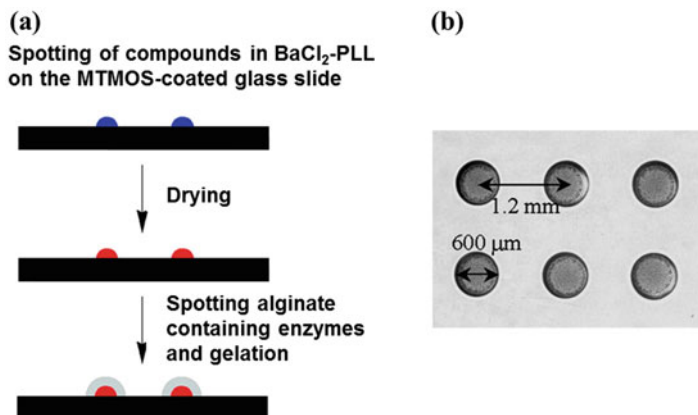


Fig. 2.6 (a) A schematic diagram showing the formation process of alginate gel enzyme on a solid surface. (b) A micrograph of the entrapped enzyme in alginate gel forming an array (Adapted the images from Lee et al. [27] with kind permission from Proceedings of the National Academy of Sciences (PNAS), Copyright (2005) National Academy of Sciences, USA)

prevent any non-specific binding that could give rise to false readings. A more recent enzyme printed microarray was presented by Lee et al. [27] as a metabolic enzyme microarray (MetaChip) in 2004. Rather than using covalent binding for enzyme immobilisation, they used a physical trapping technique for putting enzymes on a glass slide surface by using alginate gels. They treated the glass slide surface with methyltrimethoxysilane (MTMOS) to allow the printed microspots to form a hemispherical shape on the surface. They first printed all the required compounds (different substrates that are specific to an enzyme reaction in a suitable buffer), together with 0.1 M BaCl₂ and 0.01% poly-L-lysine for alginate formation, and let it dry. Next, they printed an enzyme solution with 0.5% alginate, causing instantaneous gelation of the alginate matrix and producing an enzyme microarray. The printing process was done within a microarray spotter chamber under 95% relative humidity to maintain enzyme microspots hydrated. The process of alginate gel enzyme entrapment and a micrograph of the enzyme alginate microspot are presented in Fig. 2.6.

3.2 *In Situ Synthesised Microarrays*

In situ synthesised microarrays are extremely high-density microarrays and were first to develop and patent by Affymetrix and are now widely known as GeneChips [24]. Using photolithography from semiconductor manufacturing technology, oligonucleotide microarrays were fabricated step by step by cycling photolithographic masks to reach the required sequence length. In a later

invention from Roche called NimbleGen, they replaced the photographic masks with programmable micromirrors to write the pattern on the surface. A further improvement was made by Agilent Technologies, where they used inkjet printing that was adopted from printed microarrays, to deposit four different nucleotide bases and stack them into the required sequence length. However, synthesising enzymes on top of a surface is more complicated than synthesising a DNA sequence. In order to produce an enzyme-synthesised microarray, the specific DNA sequence of the enzymes has to be first produced on the surface. Normally these genes, which are used for protein synthesis, are more than 200 base pairs that are difficult, expensive and time-consuming to synthesise. Furthermore, after the DNA sequence of the specific gene is synthesised, ribosomes and complex environmental conditions mimicking the human body are required to create the required enzymes. Such difficult chemical synthesis increases the margin of error. Nevertheless, He et al. successfully attempted to produce an enzyme (luciferase) microarray, named PISA (protein in situ array) [28, 29].

3.3 High-Density Bead Arrays

The bead array is also a highly packed microarray that was first developed by Illumina [24] and is shown in Fig. 2.7. In contrast to printed and in situ synthesised microarray that immobilises probing molecules directly on top of the surface, a bead array functionalises probing molecules on different types of beads. For the placement of the beads on top of a solid surface, a chemical or a micro-electro-mechanical etching technique has to be used to form microwells that are a good fit to the beads. Unlike printed and in situ synthesised microarrays in which the location of the specific probing biomolecules are known, all the beads with probing molecules are randomly arranged into the microwells. Typically, bead arrays employ coloured or size-coded microspheres to identify the location of the specific probing molecules on the solid surface. DNA bead arrays have been successfully used in the Ion Torrent low-cost semiconductor integrated circuit-based technology for large-scale genome sequencing [31]. Bead arrays have also been adopted by Luminex and BD Biosciences to perform protein assays. However, multiplexing enzyme-based bead arrays for metabolite detection is not widely available.

4 Sensor System-on-Chip

The emergence of Human Genome Project has impacted modern medicine significantly. Massively parallel DNA sequencing has substantially reduced the cost and resulted in various applications such as targeted genomics for diagnostics, bacterial genomics for infectious disease and the study of personal genomes for inherited disease. Furthermore, sequencing technologies are progressing from large-scale

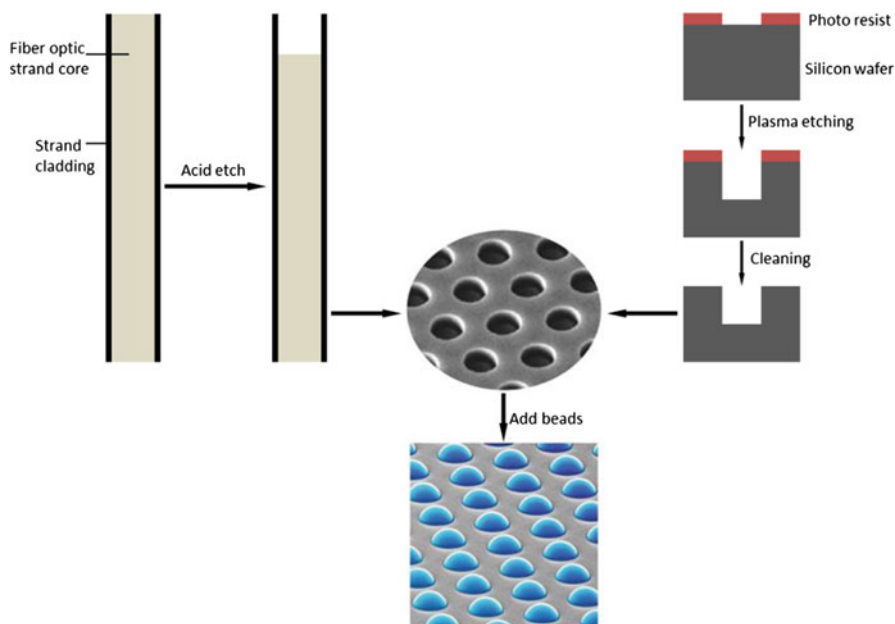


Fig. 2.7 A schematic diagram showing the process of creating Illumina bead array using (*left*) fibre optic bundles and (*right*) a silicon wafer (Reproduced from Fan et al. [30] with kind permission from BioTechniques)

biomedical research efforts to clinical laboratories or portable devices, with the aim of one day reaching the hands of the patient. Further advancement requires the number of genomes being sequenced a year to increase and implies the need of a low-cost mass-manufacturing process to achieve a price point of \$100/genome [32].

Concurrently with the advancement of personalised medicine, the electronic industry is also progressing. In particular CMOS technology, which is a low-cost, large-scale and high-quality mass-manufacturing technique for producing integrated circuits, continues to be developed according to Moore's law that describes its exponential improvement. However, even by using only legacy technology, considerable scale-up of microarray technology becomes possible. This has led to the introduction of so-called More-than-Moore technologies, whereby CMOS is adapted to work in new, and nontraditional, markets for electronic devices. CMOS was introduced in the 1970s, and it is currently widely used in modern computing with the invention of digital logic chips. More recently, CMOS has also impacted imaging technologies and communication systems with mixed-signal digital-analogue chips. Over the decades that CMOS has been under development, it is estimated that there has been a cumulative investment of 3 trillion dollars [23]. This investment has created a global CMOS chip manufacturing base that produces billions of chips per year, with unmatched quality for digital and analogue

electronics [23]. Because of the phenomenal success of the microelectronic industry in low-cost, mass manufacture, it is therefore very attractive to use a CMOS chip as a core component to enable large-scale sensing technology. For instance, the CMOS-based image sensor has displaced the use of chemical film in the photographic industry, rendering once giant companies obsolete. In addition, this also allows the new technological CMOS-based device to scale down in price following Moore's law. It is therefore possible to anticipate that CMOS will become integral platform for low-cost disposable sensors in personalised medicine. In this section we will review the technological advancements of CMOS technology for chemical sensor systems and present the development of CMOS-based devices to improve personalised healthcare.

4.1 CMOS-Based Chemical Sensor Systems

In order to improve sensing technologies, a lot of attempts have been made to fabricate chemical sensor using silicon substrate for the incorporation into CMOS process [33]. The potential of developing sensor system using silicon has been demonstrated with the emergence of microelectromechanical systems (MEMS) technology. The micromachining method has enabled the development of three-dimensional integrated chemical sensor systems on CMOS. This sensor system technology uses bulk micromachining to wet or dry etch silicon to create sensing structures by a process of backside etching of CMOS wafers [34] or surface micromachining using a sacrificial layer to create sensing structures on the front side of CMOS wafers [35]. Frontside processing to deposit active materials and form sensor devices has also been studied extensively [36]. CMOS technology therefore provides a platform for sensor integration with good device attributes such as miniaturisation, low-power consumption and rapid response. As a result, various sensors with different physical characteristics have been realised using CMOS-based chips:

- *Chemomechanical sensors*: a change in mass is collected on the sensing area. The detection can be monitored from the deflection of a micromechanical structure or the frequency shift of a resonating structure or travelling acoustic wave.
- *Thermal sensors*: a change in temperature is obtained using a thermocouple or bolometer as a sensing material. This can be measured from the resistance of a bolometer or the voltage difference between two different thermocouples.
- *Optical sensors*: a change in light intensity that can be detected using silicon-based photodiodes or other semiconductor materials.
- *Electrochemical sensors*: a charge transfer from a chemical reaction can be measured in voltage, current or resistance.

System-on-chip (SOC) is a term used to describe the creation of whole electronic systems including analogue mixed-signal and digital systems on the same chip. Using the aforementioned methods of sensor integration on CMOS, the concept

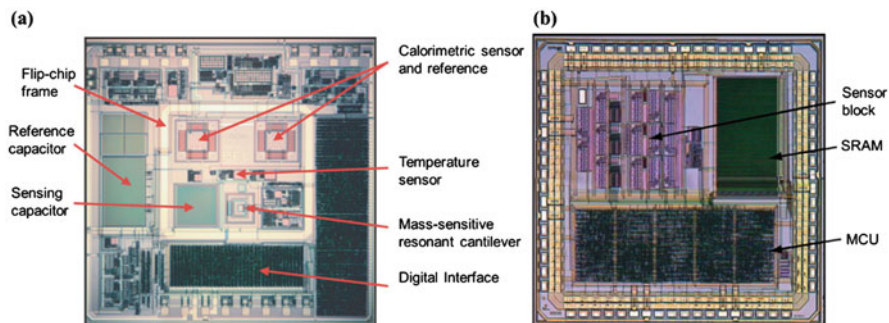


Fig. 2.8 Micrographs of the two earliest sensor systems on CMOS chip, (a) gas sensor (Image from Hagleitner et al. [38] with kind permission from IEEE) and (b) capsule-based pH sensor (Image from Hammond et al. [39] with kind permission from IEEE)

of a sensor system-on-chip (SSOC) has emerged. One of the earliest examples of a SSOC was demonstrated by Yeow et al. [37]. Their sensor system consisted of 15×16 array of pH-ISFET (ion-sensitive field-effect transistor) with integrated on-chip control and readout circuitry. The sensor system was developed using standard CMOS process but required four new masks and four extra processing steps to complete the device. A considerably more sophisticated sensor system was developed by Hagleitner et al. [38] using multiple sensing technologies. They reported a CMOS smart gas sensor that consisted of four different transducers: a mass-sensitive cantilever, a capacitive sensor, a calorimetric sensor and a temperature sensor. All of these sensors were integrated with readout and signal conditioning circuits. An analogue-to-digital conversion (ADC) was also performed on chip to transfer digital signal off-chip due to better signal-to-noise performance. Using unmodified CMOS process, a few extra micromachining steps to create the cantilever and calorimetric sensor, the sensor system was successfully fabricated. Figure 2.8a shows the chip layout of the sensor system.

The Yeow and Hagleitner devices both required considerable additional processing steps, deviating substantially from a typical CMOS process flow. In 2005, Hammond et al. demonstrated a capsule-based pH sensor system-on-chip for in vivo diagnostic applications, as shown in Fig. 2.8b [39]. The SSOC had an on-chip ion-sensitive field-effect transistor (ISFET) to measure pH, analogue circuits, data conversion and a programmable microcontroller unit with static random access memory (SRAM) for programme memory to control the procedure of collecting and processing pH data. The data could be stored in the on-chip SRAM or transmitted off-chip as a serial bitstream. The whole system including the pH sensor was fabricated using unaltered CMOS process. The most recent development in SSOC was reported by Huang et al. [40]. The system consisted of four different types of sensors: MEMS with modified hydrogel, nanowire, ISFET and bipolar transistor, which could be used to detect glucose, protein, pH and temperature, respectively. The system was designed to monitor all of these physiological parameters of the

human body in real time. A digital processor was implemented in the system to control all the functions of the chip, including signal processing, ADC and signal amplification. The chip also incorporated a wireless data transmitter. In addition, this system had self-powering capabilities, where an on-chip circuitry was designed to receive two different off-chip harvesting energy platforms: solar cell and RF coil. This system including the post-processing that was required is illustrated in Fig. 2.9.

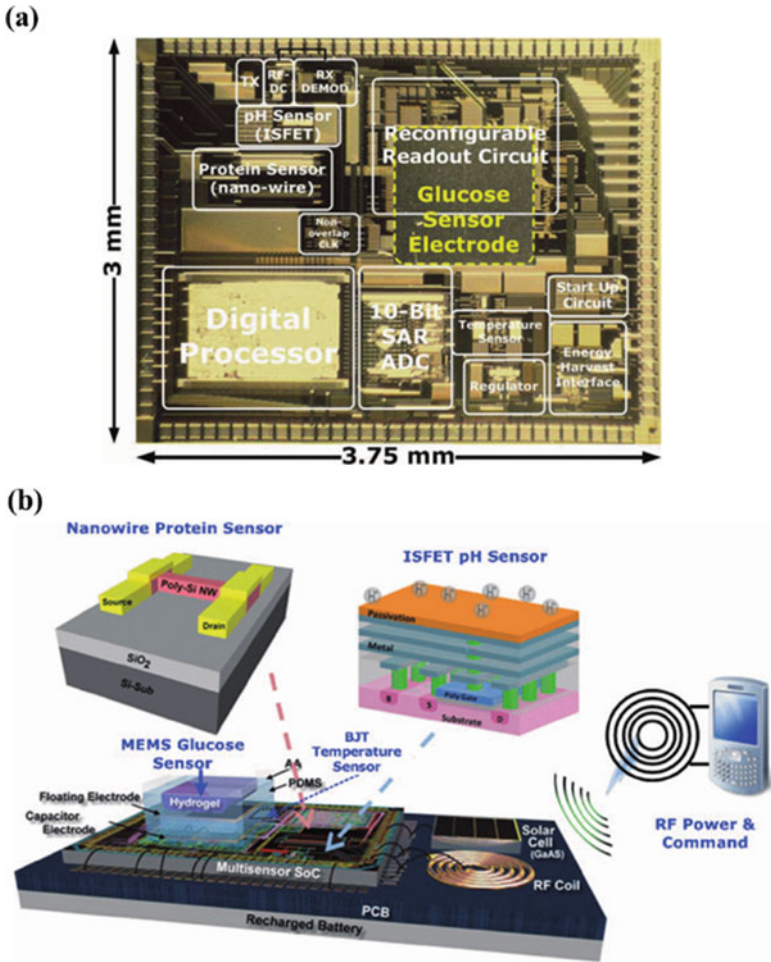


Fig. 2.9 (a) Micrograph of the smart self-powering system for monitoring physiological parameters of the human body in real time, showing the positions of all functionalities of the chip. (b) A schematic diagram describing the overall system architecture of the sensor system (Images from Huang et al. [40] with kind permission from IEEE)

4.2 Chemical Sensors on Foundry CMOS

As we have seen in the preceding section, it is possible to make a wide range of sophisticated sensors on CMOS to make a SSOC provided that it is possible to carry out several additional processing steps that are not normally executed in a CMOS foundry. The exception was the pH metering chip that employed design techniques that made it possible to create the complete system using only foundry-processed CMOS. The sensor technology on the pH metering chip was an ISFET that is a type of electrochemical sensor.

Electrochemical sensors are one of the largest and oldest groups of chemical sensors, thus they have inevitably been one of the many sensor types integrated on to CMOS. The Beckman glass electrode for pH sensing was introduced in the early twentieth century [41]. However, a glass electrode cannot be integrated on to CMOS technology. A solid-state device is therefore required to replace glass electrode as electrochemical sensor to measure pH. Consequently, Bergveld [42] introduced the ISFET as the first miniaturised silicon-based chemical sensor in 1970. A modified CMOS process was required to fabricate Bergveld ISFET. Bausells et al. [43] successfully demonstrated the possibility of integrating an ISFET on to CMOS using an unmodified two-metal process in 1999. A key step in this advancement was to exploit the standard passivation layers used to finish foundry CMOS. These are typically Si_3N_4 or a silicon oxynitride, and they were already known to work as pH-sensing membranes on ISFETs. The use of CMOS to make the ISFET made it possible to exploit the electronic functionality and the chemical functionality on a monolithic integrated circuit. Using this method it is also possible to use the technical libraries of a foundry process to rapidly prototype relatively advanced circuits and systems quickly. The operation and physical characteristics of the first ISFET were similar to the conventional MOSFET, with the exception that the metal gate was not deposited so that the gate oxide was exposed as a membrane to aqueous solution. In solution different ions (e.g. H^+ or OH^-) adsorb on to the membrane according to their concentration in solution, thus modifying the surface charge. The ISFET therefore detected the ion concentration of the solution as a gate threshold voltage shift. In equilibrium, the threshold voltage is given as below [44]:

$$V_T = V_{\text{FB}} - \frac{Q_B}{C_{\text{OX}}} + 2\phi_F \quad (2.4)$$

where V_{FB} is the flat band voltage, Q_B is the depletion charge in the silicon substrate, C_{OX} is the capacitance of the gate oxide and ϕ_F is the Fermi potential. The ISFET sensing function arises from the influence of the solution on the flat band voltage:

$$V_{\text{FB}} = E_{\text{REF}} - \Delta\phi + \chi_{\text{SOL}} - \frac{\phi_S}{q} - \frac{Q_{\text{SS}} + Q_{\text{OX}}}{C_{\text{OX}}} \quad (2.5)$$

where E_{REF} is the reference electrode's potential, $\Delta\phi + \chi_{\text{SOL}}$ is the potential at the gate oxide/electrolyte interface, ϕ_S is the work function of the semiconductor

substrate, q is the electron charge, Q_{SS} is the surface charge density and Q_{OX} is the total fixed oxide charge. χ_{SOL} is the dipole potential of the electrode that is a constant, and $\Delta\phi$ is dependent of the concentration of ion species in the electrolyte. The change in ion concentration is therefore detectable by the ISFET threshold voltage shift.

The first ISFETs were in fact sensitive to different ion species (e.g. metal and halide ions), since the oxide membrane surface was not particularly selective for H^+ or OH^- ions. In the intervening years until now, there has been a great deal of work on making ISFET membranes selective for hydrogen, potassium [45], sodium [46] and copper ions [47], amongst others. Direct sensing of biomolecules has also been demonstrated. Caras and Janata [48] reported the first use of an ISFET to detect penicillin directly via the modification on the sensing area with an enzyme. ISFETs have since been used in conjunction with various bio-recognition materials such as DNA and enzymes to detect biological substances. With the development of genomics, various efforts have been made to use the ISFET as a DNA sensing platform. Purushothaman et al. demonstrated a procedure for single nucleotide mismatch detection using an ISFET [49]. Well-established GeneChip technology uses single-stranded DNA with a DNA polymerase to sequentially add complementary nucleotides. The application of this method to an ISFET successfully demonstrated an electrical signal as a consequence of proton release in the polymerisation reaction [50]. In a contemporaneous investigation, Milgrew et al. [51, 52] used an unmodified CMOS process to develop a large transistor-based sensor array chip for direct extracellular imaging in tissue culture. The chip consisted of 16×16 pixel array of ISFETs with independent addressing, and each ISFET was used as a pH sensor with a sensitivity of 46 mV/pH. Since the chip relied on an unmodified foundry CMOS process, the design technique was immediately suitable for scale-up to large arrays. Ion Torrent demonstrated and commercialised a CMOS-integrated ISFET array (165 M ISFETs) that had the capability to collect large genome sequences [31]. The architecture of Ion Torrent genome sequencer is shown in Fig. 2.10. Using bead array technologies for DNA fragments and DNA polymerase functionalisation on the surface of integrated CMOS ISFET array, the genome sequencing of bacterial and a detailed study of a human genome were presented. This has provided a key technology for personalised medicine using genomics, reliant on a CMOS-based sensor system.

More recently, many researchers have been focussing their efforts on developing a sensor system on a CMOS platform for metabolome assessment. The aim is to deliver metabolomic phenotype detection for personalised medicine with the same success that has been achieved using genomics. However, metabolite quantification relies on many enzymes to deliver the required bio-recognition that is very complex. Unlike genome sequencing that only relies on the production of hydrogen ions to detect nucleotide attachment, not all enzyme reactions that are needed produce protons as a reaction product. For instance, the enzymatic reaction equations below show commonly used enzymes to quantify glucose and cholesterol levels, respectively:

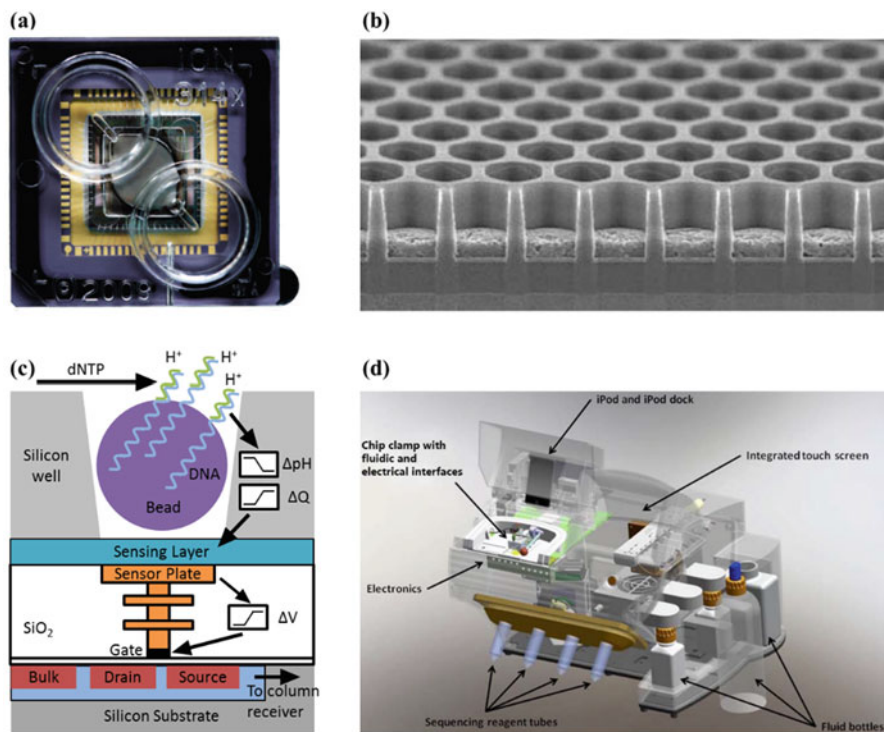
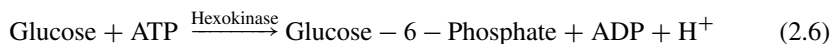


Fig. 2.10 (a) Ion Torrent sequencing chip packaged and wire bonded in ceramic carrier, with moulded fluidic lid to allow addition of sequencing reagents. (b) A scanning electron micrograph of a large array of microwells that is fabricated on Ion Torrent CMOS chip, in order to accommodate beads with DNA template. (c) A schematic diagram showing the working principle of the sequencing technique, containing a bead with DNA template in a microwell and the underlying circuitry of the ISFET sensor. Single nucleotides are added sequentially and give a signal change in an ISFET if there is a binding event occurrence that produces a hydrogen ion. (d) The complete instrument that is required for an automated sequencing process (Reproduced from Rothberg et al. [31] with kind permission from Nature)



The hexokinase reaction produces H^+ ions that can be readily detected by an ISFET, but the cholesterol oxidase reaction releases H_2O_2 . The H_2O_2 can be detected using light absorbance at 240 nm or alternatively via a secondary reaction in which a peroxidase enables oxidation of a colour dye to produce a more routinely detected colour change. Cheah et al. [53] investigated the use of CMOS ISFET sensor for label-free quantification of metabolites via hydrogen ions detection.

Using a hexokinase-based assay (Eq. 2.6) on an ISFET chip, it was demonstrated that a device with sufficient sensitivity to measure glucose at concentrations within the physiological range that occurs in the human blood was feasible. Detailed evaluation of the enzyme kinetics (see Eq. 2.1) was also possible. Furthermore, since the device consisted of a large array of independent measurement channels that could be averaged to suppress the noise floor, it was shown that measurement of lower glucose concentrations down to the level seen in human tears was possible. More recently, Al-Rawhani et al. [54] developed a disposable sensing CMOS platform for total cholesterol quantification in the human blood serum. They used an off-the-shelf LED and a CMOS photodiode sensor to quantify cholesterol levels mediated by colour changes based on the reaction of Eq. (2.7). The results that they presented were comparable to that obtained from a benchtop spectrophotometer. This optical diagnostic system is shown in Fig. 2.11.

It is clear that a single detector is not adequate to unravel the entire human metabolome. However, it is clear that CMOS provides more than one method of measurement, opening up the possibility of making multiple measurements using diverse sensing methods simultaneously on a single chip. CMOS technology has the potential to provide a suitable monolithic platform to integrate multiple detectors such as photodiodes and ISFETs to facilitate a range of assays that will be able to assess a large portion of the human metabolome.

4.3 Future Prospects for CMOS in Omics Technology

The advancement of omics technologies has provided invaluable information to understand the molecular pathology of an individual patient's disease that is at the heart of the development of personalised medicine. Several different omics technologies may be deployed. For instance, DNA sequencing technologies can be used to profile gene expression of RNA to collect cellular information; miniaturised and highly parallel immunoassays can be used to acquire information of different protein levels. Recently, Shakoor et al. [55] developed a plasmonic sensor that is monolithically integrated on a CMOS photodiode. The device is illustrated in Fig. 2.12. Detailed sensitivity analysis using refractive index adjustment with glycerol in water and preliminary results on the detection of Protein A using a surface-binding assay on the chip was presented. The work potentially opens up a pathway to integrate label-free immunoassays on to a CMOS platform for protein quantification, hence providing a new tool for proteomics.

The widespread availability of low-cost CMOS and its original intended use as an electronic technology means that there is great potential for co-integration of sensors to electronics – the SSOC. Using communication systems, an individual's physiological status can be stored into the data cloud and accessed by clinicians when necessary [56]. As a consequence, research is now developing new technologies to facilitate the required integration to create sophisticated diagnostic platforms [57, 58].

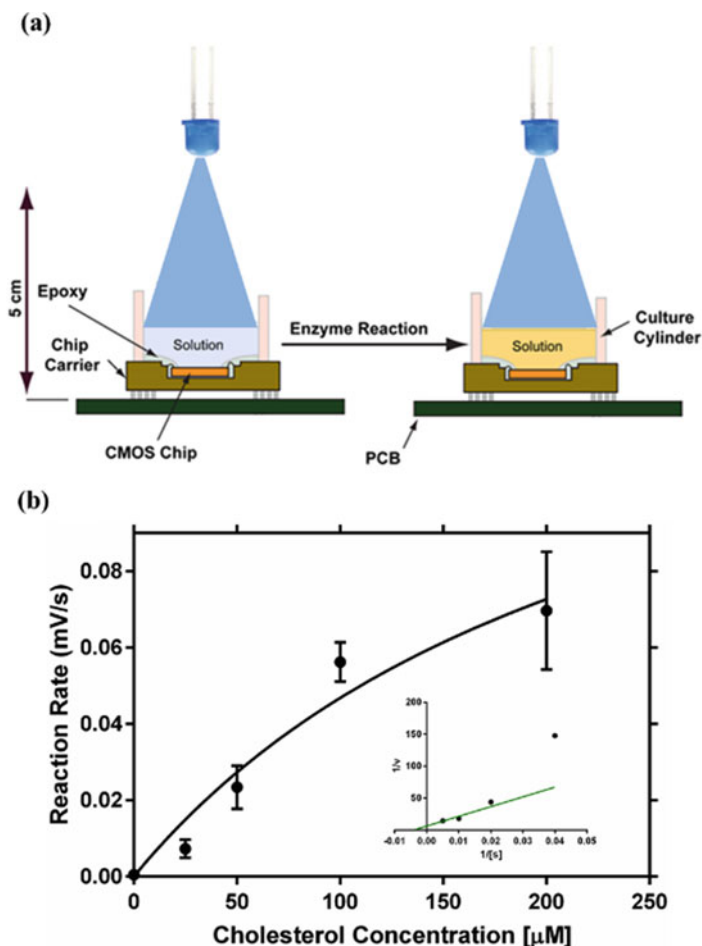


Fig. 2.11 (a) A schematic diagram showing the experimental setup of the CMOS photodiode and a green LED, to quantify cholesterol level in the human blood serum. (b) The enzyme kinetics information (reaction rate versus cholesterol concentration plot and Lineweaver-Burk plot) of cholesterol oxidase was successfully obtained using the disposable diagnostic sensing platform

5 Conclusion

To conclude, the background of personalised medicine has been reviewed, and the importance of genomics and metabolomics for future healthcare delivery has been discussed. The theory of enzyme processing and sensing has been presented to show how enzymes can be used to quantify metabolites. Different microarray technologies have been described to provide low-cost and high-throughput DNA sequencing. These technologies can be adapted to produce enzyme microarrays for the advancement of metabolomic measurement. The development of CMOS sensor systems was described, and it has been illustrated how CMOS technology

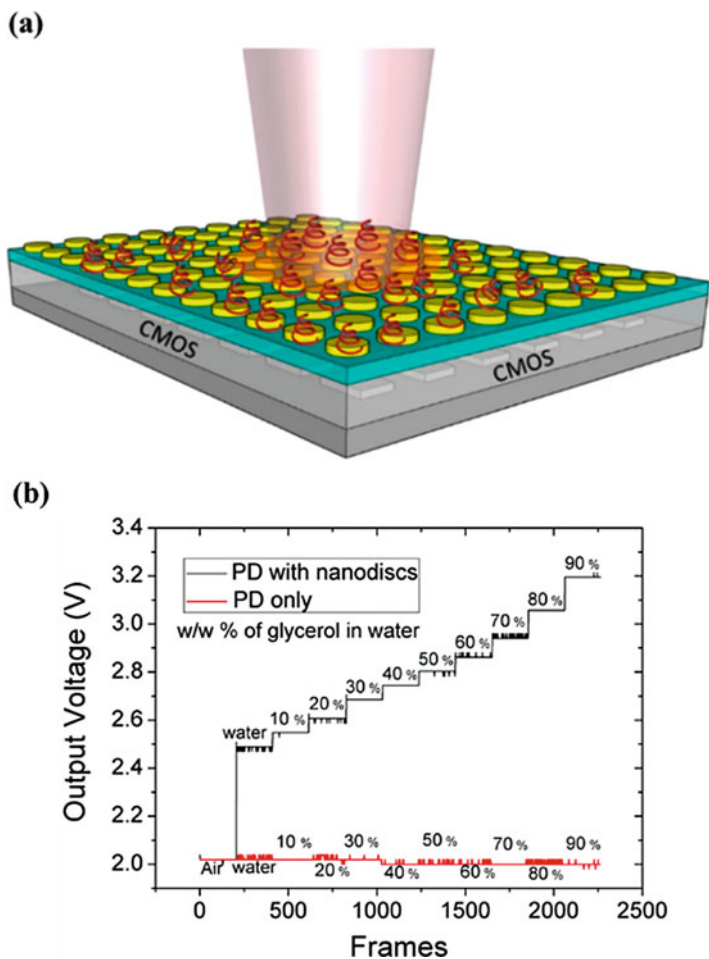


Fig. 2.12 (a) A schematic diagram showing the integration of plasmonic sensor on top of a CMOS chip for protein detection. (b) The change in light intensity that is detected by photodiode, due to the resonance wavelength shift of the plasmonic sensor for increasing concentrations of glycerol [55]

is a suitable platform for advanced sensor systems. These sensor systems are now being investigated as devices for use in personalised medicine. At present, metabolomics still relies on individual, large-scale and expensive instruments such as NMR spectroscopy and mass spectroscopy for metabolite detection. Similar to the success that has been seen in the use of advanced array technologies, including the ISFET technology of Ion Torrent for genomics, metabolomics is undergoing a transformation reliant on the use of CMOS sensor systems to develop a personal metabolome machine. CMOS technology could one day be used as a single analytical platform for the all the families of “omics” technologies to provide precision and personalised healthcare.

References

1. K.K. Jain, Basics of personalized medicine, in *Textbook of Personalized Medicine*, 1st edn., (Springer, Cham, 2009), pp. 1–27
2. G. Novelli, Personalized genomic medicine. *Intern. Emerg. Med.* **5**(Suppl 1), 81–90 (2010)
3. Personalized Medicine. Bayer Healthcare. Available <http://pharma.bayer.com/en/innovation-partnering/research-focus/oncology/personalized-medicine/>. Accessed 1 Jan 2017 (2006)
4. C.R. Scriver, The salience of Garrod's 'molecular groupings' and 'inborn factors in disease'. *J. Inherit. Metab. Dis.* **12**(1), 9–24
5. J.D. Watson, F.H.C. Crick, Molecular structure of nucleic acids. *Nature* **171**, 737–738 (1953)
6. K.B. Mullis, F.A. Faloon, Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**(C), 335–350 (1987)
7. G.P. Bates, The molecular genetics of Huntington disease – a history. *Nat. Rev. Genet.* **6**(10), 766–773 (2005)
8. E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczkzy, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R.A. Gibbs, D.M. Muzny, S.E. Scherer, J.B. Bouck, E.J. Sodergren, K.C. Worley, C.M. Rives, J.H. Gorrell, M.L. Metzker, S.L. Naylor, R.S. Kucherlapati, D.L. Nelson, G.M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D.R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H.M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R.W. Davis, N.A. Federspiel, A.P. Abola, M.J. Proctor, R.M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D.R. Cox, M.V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G.A. Evans, M. Athanasiou, R. Schultz, B.A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W.R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J.A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D.G. Brown, C.B. Burge, L. Cerutti, H.C. Chen, D. Church, M. Clamp, R.R. Copley, T. Doerks, S.R. Eddy, E.E. Eichler, T.S. Furey, J. Galagan, J.G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L.S. Johnson, T.A. Jones, S. Kasif, A. Kasprzyk, S. Kennedy, W.J. Kent, P. Kitts, E.V. Koonin, I. Korf, D. Kulp, D. Lancet, T.M. Lowe, A. McLysaght, T. Mikkelsen, J.V. Moran, N. Mulder, V.J. Pollara, C.P. Ponting, G. Schuler, J. Schultz, G. Slater, A.F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y.I. Wolf, K.H. Wolfe, S.P. Yang, R.F. Yeh, F. Collins, M.S. Guyer, J. Peterson, A. Felsenfeld, K.A. Wetterstrand, A. Patrinos, M.J. Morgan, P. de Jong, J.J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y.J. Chen, C. International Human Genome Sequencing, Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921 (2001)
9. A. Harvey, A. Brand, S.T. Holgate, L.V. Kristiansen, H. Lehrach, A. Palotie, B. Prainsack, The future of technologies for personalised medicine. *New Biotechnol.* **29**(6), 625–633 (2012)

10. D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D.D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G.E. Duggan, G.D. MacInnis, A.M. Weljie, R. Dowlatbadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B.D. Sykes, H.J. Vogel, L. Querengesser, HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, 521–526 (2007)
11. R. Kaddurah-Daouk, B.S. Kristal, R.M. Weinsilboum, *Metabolomics: a global biochemical approach to drug response and disease.* *Annu. Rev. Pharmacol. Toxicol.* **48**, 653–683 (2008)
12. W.B. Dunn, D.I. Ellis, *Metabolomics: Current analytical platforms and methodologies.* *TrAC – Trends Anal. Chem.* **24**(4), 285–294 (2005)
13. R.A. Copeland, A brief history of enzymology, in *Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis*, 2nd edn., (Wiley, New Delhi, 2000), pp. 1–10
14. H.F. Gilbert, Enzyme mechanism, in *Basic Concepts in Biochemistry*, 2nd edn., (McGraw-Hill, New York, 2000), pp. 80–93
15. R.A. Baker, L. Wicker, Current and potential applications of enzyme infusion in the food industry. *Trends Food Sci. Technol.* **7**(9), 279–284 (1996)
16. O. Kirk, T.V. Borchert, C.C. Fuglsang, Industrial enzyme applications. *Curr. Opin. Biotechnol.* **13**(4), 345–351 (2002)
17. F. Hasan, A.A. Shah, S. Javed, A. Hameed, Enzymes used in detergents: lipases. *Afr. J. Biotechnol.* **9**(31), 4836–4844 (2010)
18. T.D.H. Bugg, *Introduction to Enzyme and Coenzyme Chemistry*, 3rd edn. (Wiley, Hoboken, 2012)
19. L. Michaelis, M.L. Menten, The kinetics of invertase action. *Biochem. Z.* **49**, 333–369 (1913)
20. G. Meisenberg, W.H. Simmons, Enzymatic reactions, in *Principles of Medical Biochemistry*, 3rd edn., (Elsevier, Amsterdam, 2012), pp. 39–53
21. R.P. Ekins, Multi-analyte immunoassay. *J. Pharm. Biomed. Anal.* **7**(2), 155–168 (1989)
22. D. Stoll, J. Bachmann, M.F. Templin, T.O. Joos, Microarray technology: an increasing variety of screening tools for proteomic research. *Drug Discov. Today TARGETS* **3**(1), 24–31 (2004)
23. B. Merriman, I. Torrent, J.M. Rothberg, Progress in ion torrent semiconductor chip based sequencing. *Electrophoresis* **33**(23), 3397–3417 (2012)
24. M.B. Miller, Y.W. Tang, Basic concepts of microarrays and potential applications in clinical microbiology. *Clin. Microbiol. Rev.* **22**(4), 611–633 (2009)
25. M. Schena, R.A. Heller, T.P. Theriault, K. Konrad, E. Lachenmeier, R.W. Davis, Microarrays: Biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**(7), 301–306 (1998)
26. G. Macbeath, S.L. Schreiber, Printing proteins as microarrays for high-throughput function determination. *Science* (80-.). **289**(5485), 1760–1763 (2000)
27. M.-Y. Lee, C.B. Park, J.S. Dordick, B.P. Puc, D.S. Clark, Metabolizing enzyme toxicology assay chip (MetaChip) for high-throughput microscale toxicity analyses. *Proc. Natl. Acad. Sci.* **102**(4), 983–987 (2005)
28. M. He, M.J. Taussig, Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). *Nucleic Acids Res.* **29**(15), 73–78 (2001)
29. M. He, O. Stoevesandt, M.J. Taussig, In situ synthesis of protein arrays. *Curr. Opin. Biotechnol.* **19**(1), 4–9 (2008)
30. J.B. Fan, S.X. Hu, W.C. Craumer, D.L. Barker, BeadArray-based solutions for enabling the promise of pharmacogenomics. *BioTechniques* **39**, 583–588 (2005)
31. J.M. Rothberg, W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, M.J. Milgrew, M. Edwards, J. Hoon, J.F. Simons, D. Marran, J.W. Myers, J.F. Davidson, A. Branting, J.R. Nobile, B.P. Puc, D. Light, T.A. Clark, M. Huber, J.T. Branciforte, I.B. Stoner, S.E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J.A. Fianza, E. Namsaraev, K.J. McKernan, A. Williams, G.T. Roth, J. Bustillo, An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011)

32. S. Goodwin, J.D. McPherson, W.R. McCombie, Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**(6), 333–351 (2016)
33. A. Hierlemann, H. Baltes, CMOS-based chemical microsensors. *Analyst* **128**(1), 15–28 (2003)
34. G.T.A. Kovacs, N.I. Maluf, K.E. Petersen, Bulk micromachining of silicon. *Proc. IEEE* **86**(8), 1536–1551 (1998)
35. J. Bustillo, R. Howe, R. Muller, Surface micromachining for microelectromechanical systems. *Proc. IEEE* **86**(8), 1552–1574 (1998)
36. Q. Chen, D. Chitnis, K. Walls, T.D. Drysdale, S. Collins, D.R.S. Cumming, CMOS photodetectors integrated with plasmonic color filters. *IEEE Photon. Technol. Lett.* **24**(3), 197–199 (2012)
37. T.C.W. Yeow, M.R. Haskard, D.E. Mulcahy, H.I. Seo, D.H. Kwon, A very large integrated pH-ISFET sensor array chip compatible with standard CMOS processes. *Sensors Actuators B Chem.* **44**(1–3), 434–440 (1997)
38. C. Hagleitner, A. Hierlemann, D. Lange, A. Kummer, N. Kerness, O. Brand, H. Baltes, Smart single-chip gas sensor microsystem. *Nature* **414**, 293–296 (2001)
39. P.A. Hammond, D. Ali, D.R.S. Cumming, A system-on-chip digital pH meter for use in a wireless diagnostic capsule. *IEEE Trans. Biomed. Eng.* **52**(4), 687–694 (2005)
40. Y. Huang, T. Tzeng, T. Lin, C. Huang, P. Yen, P. Kuo, C. Lin, S. Lu, A self-powered CMOS reconfigurable multi-sensor SoC for biomedical applications. *IEEE J. Solid State Circuits* **49**(4), 851–866 (2014)
41. A.J. Bard, L.R. Faulkner, Potentials and thermodynamics of cells, in *Electrochemical Methods: Fundamentals and Applications*, 2nd edn., (Wiley, Hoboken, 2001), pp. 44–86
42. P. Bergveld, Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Trans. Biomed. Eng.* **17**(1), 70–71 (1970)
43. J. Bausells, J. Carrabina, A. Errachid, A. Merlos, Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. *Sensors Actuators B Chem.* **57**(1–3), 56–62 (1999)
44. S.M. Sze, K.K. Ng, MOSFETs, in *Physics of Semiconductor Devices*, 3rd edn., (Wiley, Chichester, 2007), pp. 293–360
45. S.D. Moss, J. Janata, C.C. Johnson, Potassium ion-sensitive field effect transistor. *Anal. Chem.* **47**(13), 2238–2243 (1975)
46. M. Wipf, R.L. Stoop, A. Tarasov, K. Bedner, W. Fu, I.a. Wright, C.J. Martin, E.C. Constable, M. Calame, C. Scho, Selective sodium sensing with gold-coated silicon nanowire field-effect transistors in a differential setup. *ACS Nano* **7**(7), 5978–5983 (2013)
47. S. Wakida, Y. Kohigashi, K. Higashi, Y. Ujihira, Chemically modified copper ion-selective field-effect transistor with 7,7,8,8-tetracyanoquinodimethane. in *Proceedings of IEEE International Conference on Solid-State Sensors and Actuators (TRANSDUCERS '91)* (1995), pp. 925–927
48. S. Caras, J. Janata, Field effect transistor sensitive to penicillin. *Anal. Chem.* **52**(8), 1935–1937 (1980)
49. S. Purushothaman, C. Toumazou, C.P. Ou, Protons and single nucleotide polymorphism detection: a simple use for the ion sensitive field effect transistor. *Sensors Actuators B Chem.* **114**, 964–968 (2006)
50. T. Sakurai, Y. Husimi, Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. *Anal. Chem.* **64**(17), 1996–1997 (1992)
51. M.J. Milgrew, M.O. Riehle, D.R.S. Cumming, A large transistor-based sensor array chip for direct extracellular imaging. *Sensors Actuators B Chem.* **111–112**, 347–353 (2005)
52. M.J. Milgrew, M.O. Riehle, D.R.S. Cumming, A 16x16 CMOS proton camera array for direct extracellular imaging of hydrogen-ion activity. in *2008 IEEE International Solid-State Circuits Conference – Digest of Technical Papers* (2008), pp. 590–592
53. B.C. Cheah, A.I. Macdonald, C. Martin, A.J. Streklas, M.A. Al-rawhani, B. Nemeth, J.P. Grant, M.P. Barrett, R.S. Cumming, An integrated circuit for chip-based analysis of enzyme kinetics and metabolite quantification. *IEEE. Trans. Biomed. Circuits Syst.* **10**(3), 721–730 (2015)

54. M. Al-Rawhani, B.C. Cheah, A. MacDonald, C. Martin, C. Hu, J. Beeley, L. Gouveia, J. Grant, G. Campbell, M. Barrett, D. Cumming, A colorimetric CMOS-based platform for rapid total serum cholesterol quantification. *IEEE Sensors J.* **17**(2), 240–247 (2016)
55. A. Shakoor, B.C. Cheah, D. Hao, M. Al-Rawhani, B. Nagy, J. Grant, C. Dale, N. Keegan, C. McNeil, D.R.S. Cumming, Plasmonic sensor monolithically integrated with a CMOS photodiode. *ACS Photon.* **3**(10), 1926–1933 (2016)
56. J.D. Piette, H. Datwani, S. Gaudioso, S.M. Foster, J. Westphal, W. Perry, J. Rodríguez-Saldaña, M.O. Mendoza-Avelares, N. Marinec, Hypertension management using mobile technology and home blood pressure monitoring: Results of a randomized trial in two low/middle-income countries. *Telemed. e-Health* **18**(8), 613–620 (2012)
57. Y. Huang, A.J. Mason, Lab-on-CMOS integration of microfluidics and electrochemical sensors. *Lab Chip* **13**(19), 3929–3934 (2013)
58. M. Punjiya, C.H. Moon, S.S. Nanolab, Multi-analyte paper-analytical-devices (PAD) with CMOS integration for point-of-care diagnostics. *Proc. IEEE. Int. Symp. Circuits Syst.* **2016**, 2883–2886 (2016)

Chapter 3

Flexible Single-Photon Image Sensors

Pengfei Sun, Ryoichi Ishihara, and Edoardo Charbon

1 Photon Counting for Biomedical Imaging Applications

Biomedical imaging [1] is receiving more and more attention as a collection of technologies to create visual representations of the interior of a body for clinical analysis and medical intervention. It could be classified into several branches, including radiography, magnetic resonance imaging (MRI), nuclear medicine, ultrasound, elastography, photoacoustic imaging, tomography, and so on [2–5].

Advances in imaging technologies could enable us to create new ways to monitor and treat disease, often using noninvasive methods. Thanks to imaging technology, more and more biomedical instruments and facilities for clinical analysis and scientific research have been developed, as shown in Fig. 3.1 [6, 7].

Just as photon counting is being developed in biomedical imaging applications, many other fields such as time-resolved Raman spectroscopy, 3D imaging, and even space communications are also being investigated [8–10].

Advanced biomedical imaging, such as the microenvironment study of fluorescent molecules, requires precise measurement of subnanosecond lifetimes based on very weak optical signals [8]. The requirement to detect single photons, which is the ultimate level of sensitivity in optical radiation sensor, makes it very challenging for conventional imaging techniques, which mainly rely on sufficient light intensity and thus multiple photons. Photon-counting techniques have become

P. Sun • R. Ishihara
TU Delft, Delft, The Netherlands

E. Charbon (✉)
Advanced Quantum Architecture Laboratory (AQUA), EPFL, Rue de la Maladière 71b,
2000 Neuchâtel, Switzerland
e-mail: edoardo.charbon@epfl.ch

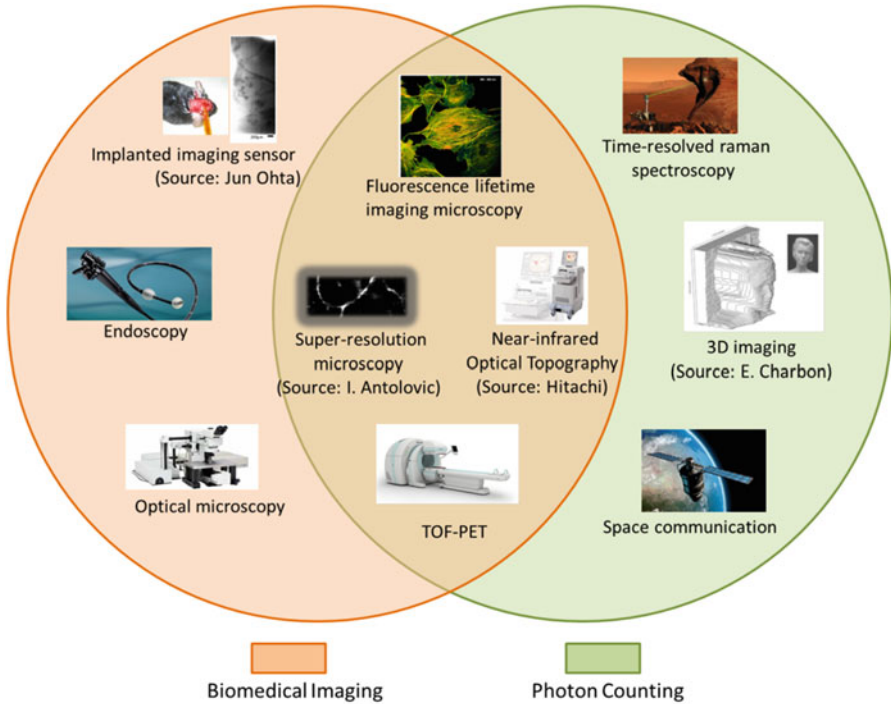


Fig. 3.1 Image sensor for biomedical applications

more and more influential in biomedical imaging fields, such as time-correlated single-photon counting (TCSPC), which offers both single-photon sensitivity and picoseconds temporal resolution. Whether it is for accurately locating the position of a tumor, providing an early diagnosis of a disease, achieving unique insight into the fluorescence lifetime imaging, or even getting a super-resolution image of a biomolecule, the development and application of photon-counting technologies provide exciting opportunities for microelectrical researchers to collaborate with life scientists and clinicians.

Over the last decade, the push toward new applications, such as pill cameras, healthcare chips, capsule endoscopy, retinal prosthesis, edible probes, and implantable sensors, has continued [6, 11, 12].

Different from conventional applications, as shown in Table 3.1, these new configurations contain novel CMOS image sensor chips, which allow biomaterials or living tissues to be in direct contact with the surface, thus enabling more compact biomedical imaging systems.

Such a compact system can also be implanted into a living body. The implantation of an imaging system can also enable new applications of clinical devices. A typical application in the field of biomedical sensing is the retinal prosthesis [13–15], as a nonliving, electronic substitute for the retina. It aims to restore vision to

Table 3.1 Trend of biomedical imaging technology development

Hospital (conventional)	Personal use (new trend)
X-ray camera	Pill camera
MRI, PET	Healthcare chip
NIRS	Mobile NIRS
Endoscopy	Capsule endoscopy
DNA sequencer	DNA chip
Stimulation needle	Retinal prosthesis

Source: Jun Ohta

someone blinded by retinal eye disease. Another application is chronic biomedical monitoring [16], where a wearable or implantable miniaturized image sensor could be left in situ to continuously monitor a person's health status, providing more accurate information about the progression of diseases such as cancer and other inflammatory or chronic ailments.

2 A Novel Photon Counter: Flexible Ultrathin-Body Single-Photon Avalanche Diode

2.1 Flexible Single-Photon Sensor Solution

To meet the requirements of implantable or wearable biomedical applications, novel photodetector solutions need to be developed, in which backside illumination and new substrate post-processing are core technologies, while inherent CMOS compatibility is a prerequisite [17, 18], as shown in Fig. 3.2.

The ideal solution should address two challenges, one of which is that the photon detector needs to be highly sensitive to detect single photons with low noise while at the same time being CMOS compatible. For these reasons it was decided that single-photon avalanche diodes would be the ideal photodiode in this work because of good timing resolution and CMOS integration compatibility. The other challenge is that the photon detector should have high fill factor to let a high enough photon flux impinging from both front and backsides through. Furthermore, the substrate material should be biocompatible and bendable to fit in the curved imaging plane in the human eyeball or other body surface.

By comparing typical flexible electronics solutions, SOI substrate transfer technology is the preferred solution to realize CMOS large-scale single-photon imaging sensor systems, when compared with other flexible electronics solutions as listed in Table 3.2.

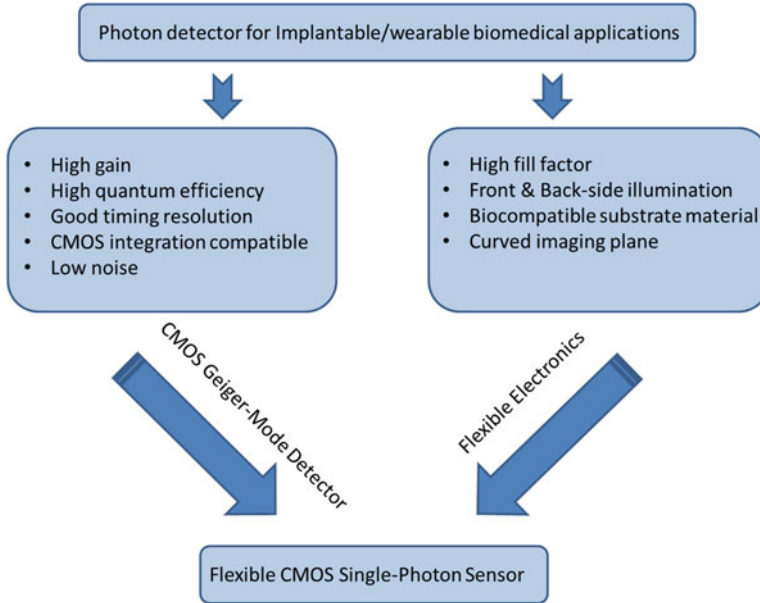


Fig. 3.2 Flexible single-photon sensor solution

Table 3.2 Comparison of flexible electronic solutions

	Substrate thinning	Thin film processing	Device transfer	SOI substrate transfer
Device performance	Good	Poor	Good	Good
Integration scale	High	Low	Middle	High
Mechanical flexibility	Middle	High	High	Good
Process complexity cost	Low	Low	High	Middle

2.2 First Flexible Ultrathin-Body SOI Single-Photon Avalanche Diode

2.2.1 SOI SPAD Fabrication

Different from traditional CMOS-compatible SPADs, which are generally implemented in bulk silicon wafer, SPADs using SOI technology are proposed in this work, providing a promising solution to realize BSI and flexible image sensor further involving substrate transfer technology.

The proposed SOI SPAD cross section and model are shown in Fig. 3.3 (a) and (b), respectively. Note that the introduction of epitaxy on SOI occupies the whole SPAD device. Since the thickness of the epitaxy layer is a design parameter, the body doping needs to be carefully optimized to avoid premature edge breakdown (PEB),

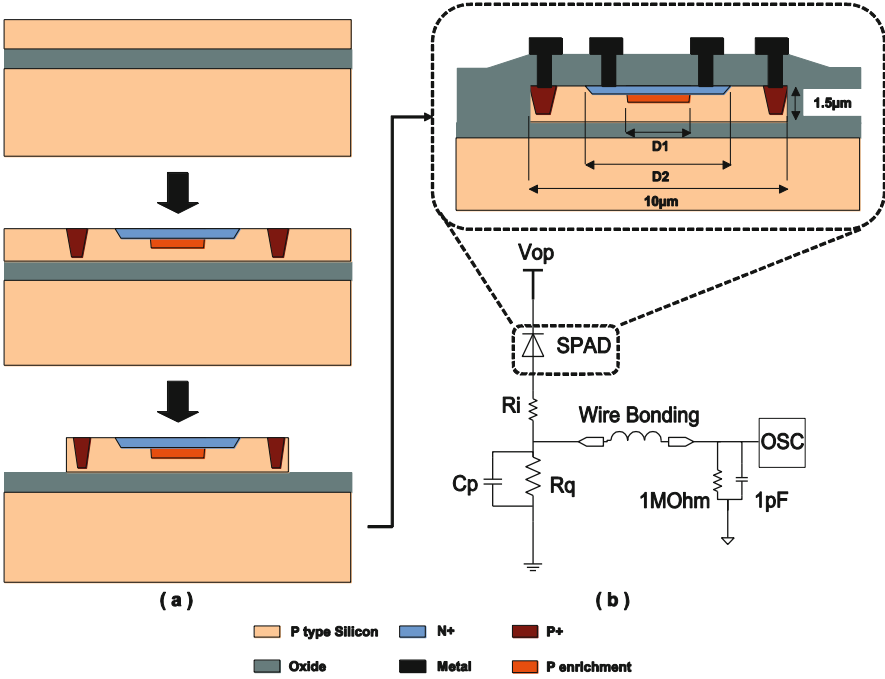


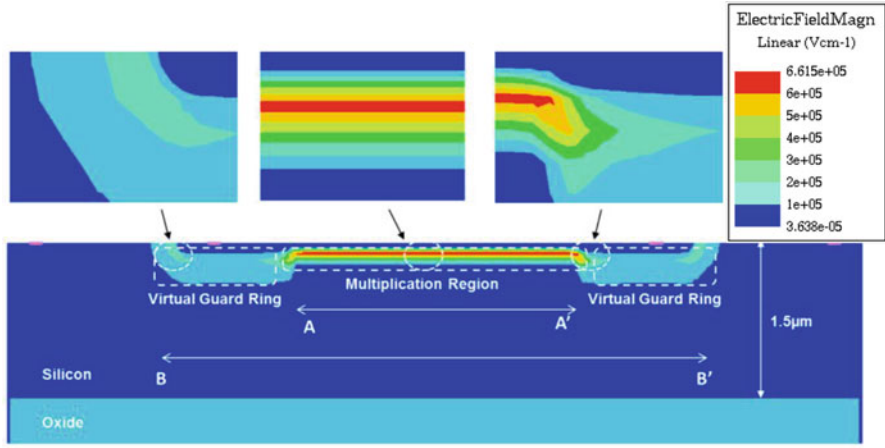
Fig. 3.3 (a) Device fabrication and cross section of device structure: $D1$ is the diameter of the enhancement layer and $D2$ is the diameter of the implicit guard ring. (b) Schematic of passive quenching circuit and simplified SPAD model

while the silicon body under the virtual guard ring is fully depleted. Enrichment doping also needs to be optimized to control band-to-band tunneling noise and/or PEB.

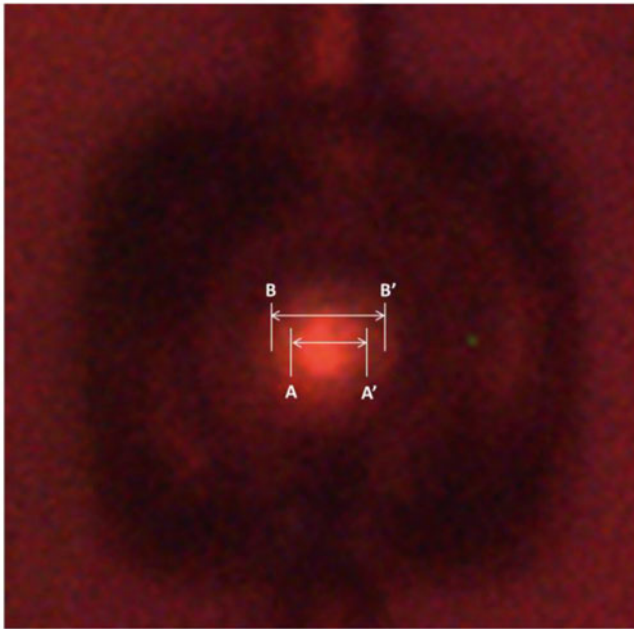
Fabrication begins with a p-type SOI wafer prepared by epitaxy technology. To ensure good mechanical compatibility and the application of dual-side illumination, the thickness of the top silicon layer was reduced to 1.5 μm. To the best of our knowledge, the SPAD proposed in this work has the thinnest body to date, if compared with conventional SPAD structures, including reach-through SPADs and planar SPADs [19–22].

The N+P junction, with depth of approximately 100 nm, is formed by implantation. A P+ enhancement region is made to form the multiplication region, and an all-around virtual guard ring is defined implicitly by the existing doping difference. Guard rings are used to isolate active regions, so as to prevent electrical cross talk and to prevent PEB, so as to minimize DCR [23].

The Medici-simulated electric field contours shown in Fig. 3.4a demonstrate the electric field reduction at the guard ring for PEB prevention. By optimizing the implantation and junction profiles, the electric field can be constrained in the multiplication region, and PEB can be suppressed effectively. It is also proved by the ionization light emission image in Fig. 3.4b.



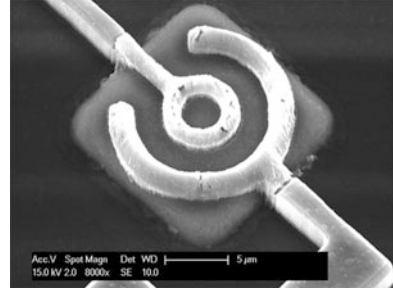
(a)



(b)

Fig. 3.4 (a) 10- μm -diameter SPAD device overlaid on a Medici-simulated electric field plot. In silicon, impact ionization occurs with electric fields higher than 2.5×10^5 V/cm. The multiplication region is in correspondence with the enhancement layer, where the absolute value of breakdown voltage, $|V_{BD}|$, is lower but uniform. In the virtual guard ring areas, the electric field is reduced, thus preventing premature edge breakdown. (b) Light emission by impact ionization test indicating the high electric field across the central multiplication region

Fig. 3.5 SEM image of an individual SOI SPAD



After junction implantation, device islands are formed by dry etching. A highly doped surrounding ring along the periphery of the island is made to separate the device from defect and trap centers at the island edges. 1.5 μm TEOS PECVD oxide is deposited as insulator by two-time etch-back to form a spacer at the silicon island step. Contact holes are opened by dry etching. Then, a 3 μm physical vapor-deposited (PVD) Al/Si (1%) is used to form anode and cathode contacts. The typical device diameter is 10 μm , with different pitch available (see D1 and D2 in Fig. 3.3).

A passive quenching circuit, which is simple but effective, was used as shown in Fig. 3.3b. The anode of the SPAD with internal series resistance R_i is connected to a quenching resistance R_q in parallel with parasitic capacitance C_p . Due to the presence of R_q between the anode of the diode and ground, the avalanche current upon photon arrival produces a voltage rise across the quenching resistor and brings the voltage across the diode below breakdown, thus quenching the avalanche. The output signal in Geiger mode is probed by an oscilloscope. A SEM micrograph of a single device is shown in Fig. 3.5.

2.2.2 Substrate Transfer Process

The substrate transfer process is summarized in Fig. 3.6. After SOI SPAD devices are processed, PECVD oxide is deposited at the backside of the SOI wafer and patterned as mask as shown in Fig. 3.6a for the following deep reactive ion etching (DRIE). The polyimide is coated and cured on the top of the device at 400 $^{\circ}\text{C}$. Then, the polyimide is patterned to expose the metal contact and light absorption area on the frontside of SPADs in Fig. 3.6b. The silicon substrate under the buried oxide layer is etched away by DRIE. It is however well known that the etch rate of silicon strongly depends on both the area of silicon exposed and the aspect ratio [24, 25]. In order to minimize the backside silicon substrate etching variation, the process is designed to include endpoint detection and optimized over-etching time. The over-etching stops at 1- μm -thick buried oxide layer as shown in Fig. 3.6c. Hence, the SPAD device layer on polyimide can be easily released by means of reasonable mechanical stress, as shown in Fig. 3.6d. As a result, the SPAD's layer has been successfully transferred to flexible polyimide layer and was further mounted to PEN (polyethylene naphthalate) or other flexible substrates for package.

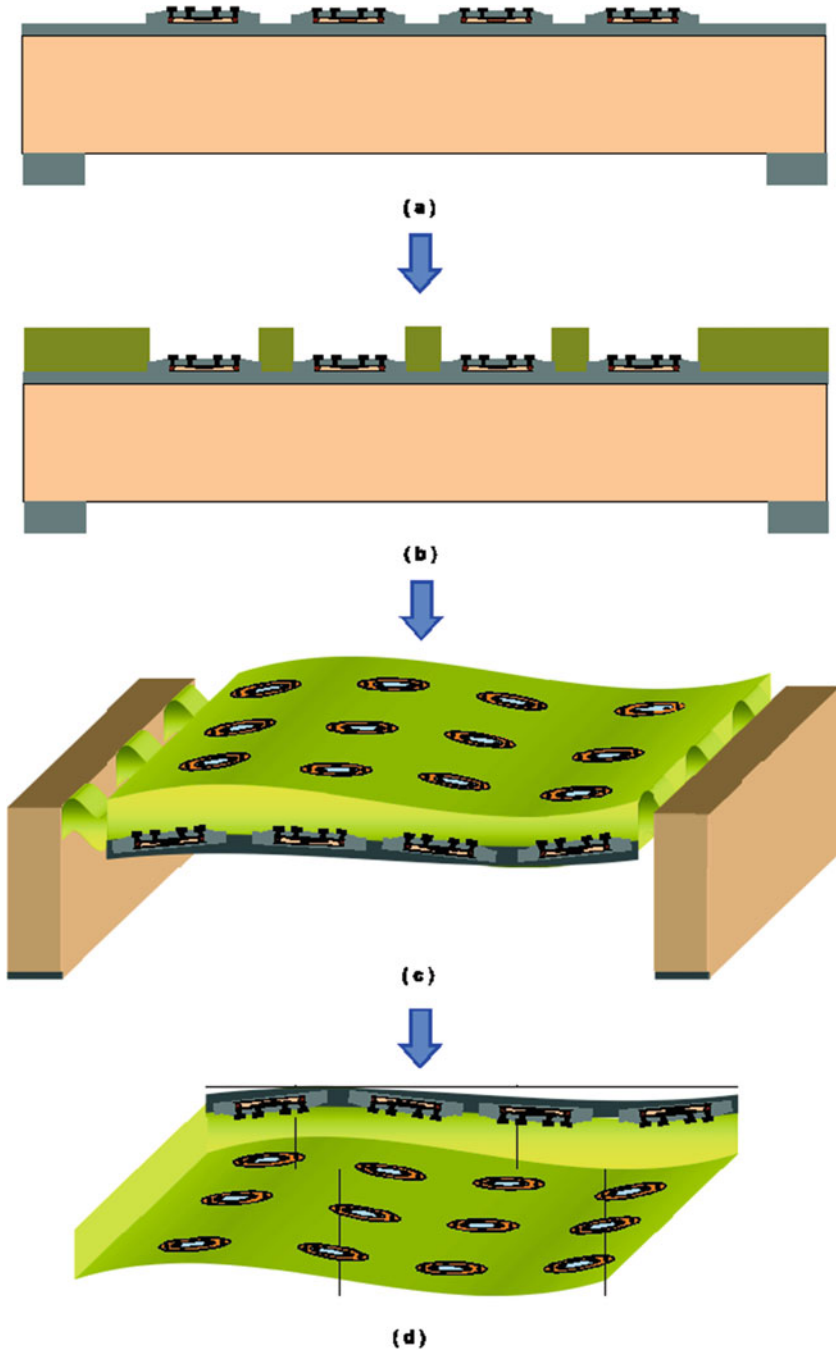


Fig. 3.6 SPAD layer transfer process. (a) Backside mask patterning. (b) Polyimide coating and patterning. (c) Substrate etching by DRIE. (d) Flexible device layer releasing

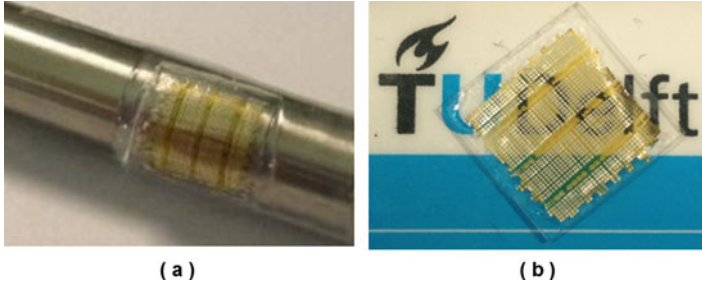


Fig. 3.7 (a) Flexible SPADs on PEN substrate bent onto 10-mm-diameter cylinder. (b) Flexible SPADs mounted onto quartz

As shown in Fig. 3.6d, SPADs can operate in dual-side illumination (DSI) since the light absorption area on both the front- and backsides is exposed.

The maximum temperature of the transfer process is 400 °C (polyimide curing process in oven with temperature monitor); thus, there is little thermal impact on the fabricated devices during transfer process.

Figure 3.7 shows photographs of SPADs after being transferred onto a flexible polyimide layer.

2.3 Comprehensive Characterizations on Flexible Ultrathin-Body Single-Photon Avalanche Diode

To the best of our knowledge, the reported device is the first flexible SPAD and also the first ultrathin-body SPAD with dual-side illumination at the time of its publication [26, 27]. DCR, PDP, afterpulsing, and time jitter performance of the device on flexible substrate is consistent with that of a device before transfer. The performance of the proposed device compares favorably with that of CMOS SPADs, while it can operate in dual-side illumination, as shown in Fig. 3.8.

The peak PDP can reach 11% in FSI mode and 6% in BSI mode, and the minimum DCR was less than 20 kHz with negligible afterpulsing probability. In BSI, the sensitive spectrum may be wider, and PDP, especially at long wavelengths, could be enhanced. Furthermore, fill factor could be improved significantly, due to a wider carrier collection region and the absence of blocking metal layers in BSI mode. It could be improved further by designing larger active area SPADs and by optimizing the doping profiles, so as to widen the ratio of active to guard ring areas. This technology is CMOS compatible, enabling CMOS circuit monolithic integration and large-scale SPAD array with readout circuit. A further improvement can be introduced with a buried layer at the backside interface or some other material to solve the superficial carrier generation and charge collection issue [28]. Such a dual-side illumination SPAD provides a novel methodology to overcome the limits of BSI applications in CMOS technology, simultaneously enhancing the fill factor, while pixel pitch keeps scaling down.

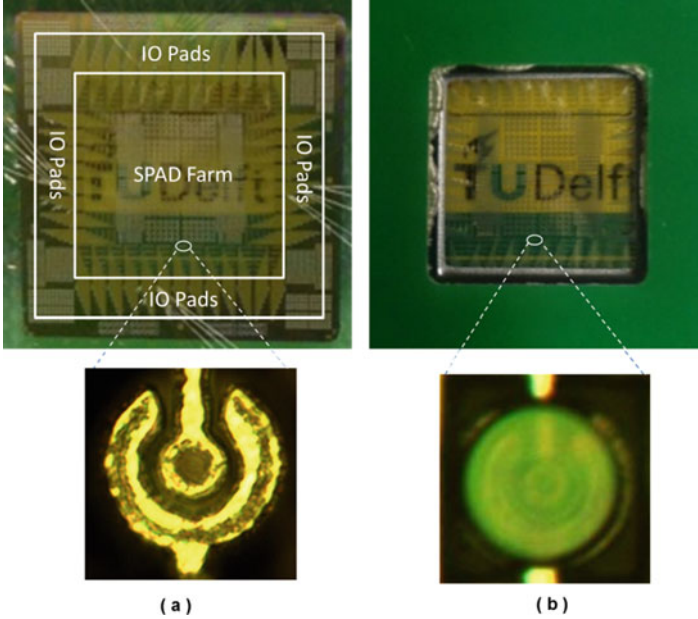


Fig. 3.8 (a) Front of flexible SPADs bonded on PCB. (b) Back of flexible SPADs bonded on PCB

Table 3.3 Summary of APD performance before and after transfer [27]

Performance	SOI SPAD			Flexible SPAD (FSI)			Flexible SPAD (BSI)			Unit
	Min.	Typ.	Max.	Min.	Typ.	Max.	Min.	Typ.	Max.	
Avalanche breakdown voltage		12.7			13.4			13.4		V
Excess bias	0		1.2	0		1.2	0		1.2	V
DCR	20		120	20		177	20		177	kHz
PDP			11.3			11.0			6.1	%
Timing jitter (405 nm wavelength laser)		500			380			1400		ps
Timing jitter (637 nm wavelength laser)					900			680		ps

The characterization of APDs operating in proportional and Geiger modes before and after substrate transfer, in FSI and BSI, respectively, is summarized in Table 3.3.

Further analysis of the Geiger-mode performances of this flexible SPAD configuration is investigated comprehensively in [29]: dark count rate (DCR), V_{BD} , and photon detection probability (PDP) are studied based on different junction parameters, operation temperature, and device structures.

Experimental results show that dark count rate (DCR) by band-to-band tunneling can be reduced by optimizing multiplication doping. DCR by trap-assisted avalanche, which is believed to be originated from the trench etching process,

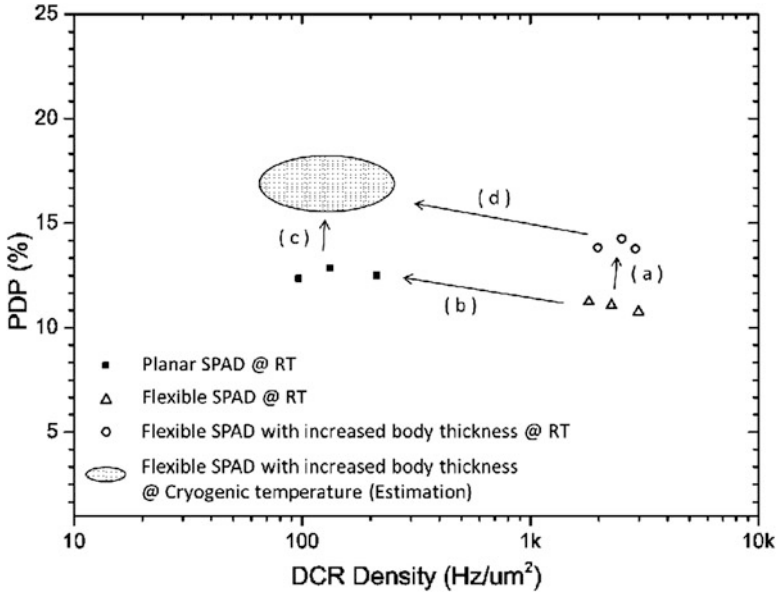


Fig. 3.9 Plot of DCR density vs. PDP in recent developments, each introducing a technological innovation or optimization with a consequent performance improvement. Note: PDP refers to FSI. (a) Increase of body thickness; (b) isolation by LOCOS; (c) isolation by trench and operation at cryogenic temperature; (d) operation at cryogenic temperature [29]

was further reduced, resulting in a DCR density of tens to hundreds of Hertz per square micrometer at cryogenic temperature. The influence of the trench etching process onto DCR is also proved by comparison with planar ultrathin-body SPAD structures without trench. Photon detection probability (PDP) can be achieved by wider depletion and drift regions and by carefully optimizing body thickness. The comparison of DCR density and PDP based on trench-isolated SPAD and planar SPAD at different temperatures is summarized in Fig. 3.9.

3 A Flexible Dual-Side Single-Photon Image Sensor

3.1 Pixel Structure and Fabrication Flowchart

As shown in Fig. 3.10, each of the two neighboring pixels contains a SPAD, a quenching resistor, and CMOS buffer circuits powered by supply voltage V_{DD} . All the components are based on a trench-isolated silicon island structure to achieve high levels of flexibility. Polyimide and polymer are used as flexible substrates and also act as a microlens to increase fill factor [30].

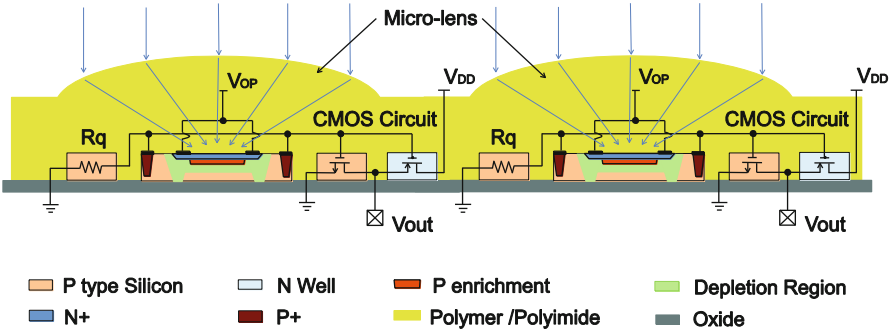


Fig. 3.10 Flexible CMOS SPAD sensor pixel schematic and SPAD cross section

For the flexible CMOS SPAD sensor, fabrication begins with a p-type SOI wafer prepared by epitaxy technology. The N-well is formed by implantation and followed by a thermal drive-in process. LOCOS is used to isolate transistor channels. After gate oxide growth, polysilicon is deposited and doped by phosphorous diffusion as gate. Source and drain are formed by implantations.

After CMOS transistors are built, a N+P junction is formed by implantation. The manufacture flow is further optimized by variable process parameters such as enrichment doping dose and epitaxy layer thickness, which are analyzed comprehensively in Sect. 3. After junction implantation, the device islands are formed and isolated with trenches by dry etching. Then, a passivation layer is deposited by two-time etch-back to form a spacer at the trench step. Contact holes are opened by dry etching. Then, a metallization layer is sputtered and patterned to realize the first metal interconnection. After the second metallization, which is similar to the first one, the wafer is sent to alloy. The device fabrication flow is summarized in Fig. 3.11a.

The flexible substrate transfer and microlens imprint process are summarized in Fig. 3.11b. Followed by the backside oxide deposition in Fig. 3.11a, the oxide is patterned as mask, as shown in step 1. The sol-gel polymer is coated on top of the device after the backside oxide mask is patterned. Then, the polymer is patterned and the quartz mold is brought into contact with the polymer. Pressure must be applied to form the microlens array on top of the SPAD sensor [31] with lateral alignment accuracy of less than 1 μm . The imprinting process was being completed at the time of the writing of the thesis. The silicon substrate under the buried oxide layer is then etched away. The etching stops at the buried oxide layer and the SPAD image sensor layer on new flexible substrate can easily be released.

Thanks to the function of the polymer layer as both flexible substrate and microlenses, the SPAD sensors with CMOS circuit systems can act as flexible imager with higher fill factor. According to our current design, the fill factor is theoretically expected to be higher than 10% with microlenses.

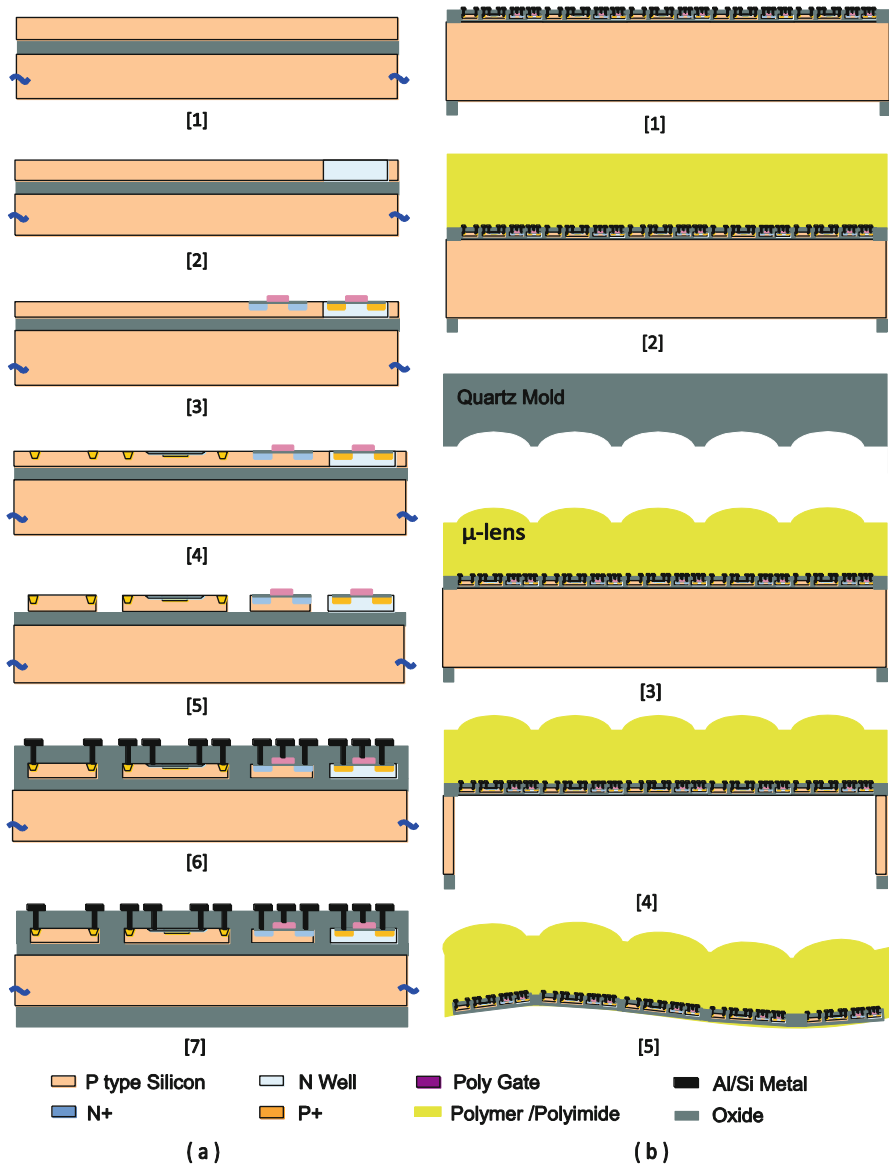


Fig. 3.11 (a) Pixel device fabrication flow chart. (1) Silicon epitaxy process on SOI wafer; (2) N-well implantation and driving in; (3) CMOS transistor fabrication; (4) SPAD junction implantation; (5) trench etching process; (6) metallization (note: only first metallization is shown in this figure); (7) backside oxide deposition. (b) Flexible substrate transfer and microlens fabrication. (1) Oxide mask fabrication on backside; (2) sol-gel polymer coating and curing; (3) microlens imprinting by quartz mold; (4) substrate etching; (5) layer releasing

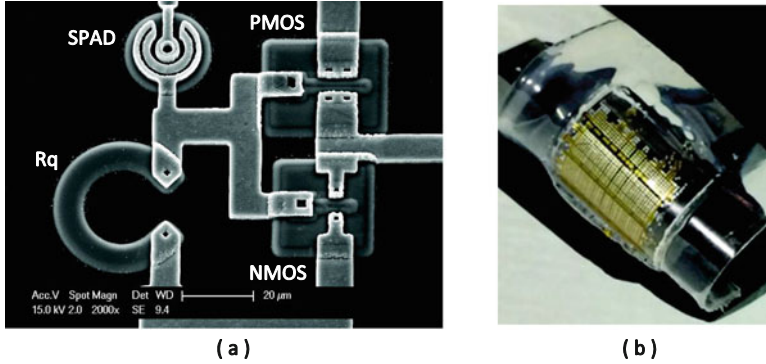


Fig. 3.12 (a) SEM microphotograph of a CMOS SPAD sensor pixel; (b) photo of bent flexible SPAD pixel farm sample

3.2 Pixel Characterizations

Following the fabrication in the proposed technology of Fig. 3.11, the first SPAD image sensor integrated with CMOS buffering circuit on flexible substrate is demonstrated as shown in the SEM microphotograph of Fig. 3.12a. The photo of bent flexible SPAD pixel farm sample, which has been released and mounted to PDMS piece, is shown in Fig. 3.12b.

For the flexible device characterization, the chip was packaged through wire bonding and then measured. Because the SPADs have been already equipped with quenching resistors and CMOS buffering circuits, which were also integrated on flexible substrate, the operation voltage V_{OP} and the power for circuit V_{DD} need to be provided through the pad. The SPADs were operated in Geiger mode and the output was monitored by a high-speed oscilloscope (LeCroy Wavemaster 8600A). Figure 3.13 shows the flexible SPAD package on PCB and the whole measurement setup.

Thanks to reduced parasitic capacitance, obtained through the integration of the output buffer, the dynamic range of SPADs could be enhanced. Afterpulsing and cross talk are negligible at the dead time. Intrinsic silicon layer could be doped by diffusion from epitaxy layer during a large amount of thermal budget during CMOS process, as shown in Fig. 3.14, resulting comparable PDP and timing jitter in FSI and BSI, as shown in Figs. 3.15 and 3.16, respectively.

The comparison between flexible trench-isolated SPAD, planar SPAD, and flexible CMOS-buffering SPAD pixel is summarized in Table 3.4.

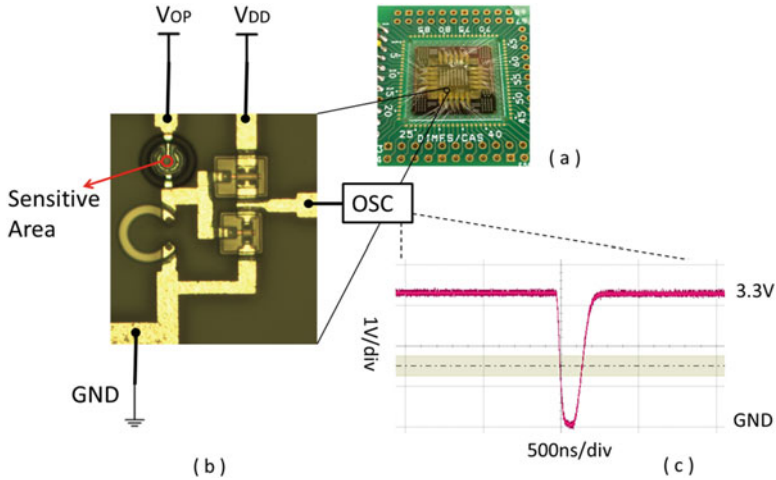


Fig. 3.13 (a) Flexible CMOS SPAD pixel farm sample mounted onto PCB. (b) Microscopic image of flexible CMOS SPAD pixel. (c) Oscilloscope image of dark count quenching pulse of pixel output

Fig. 3.14 SOI body doping profile comparison before and after high thermal budget CMOS process simulated by TSUPRM4

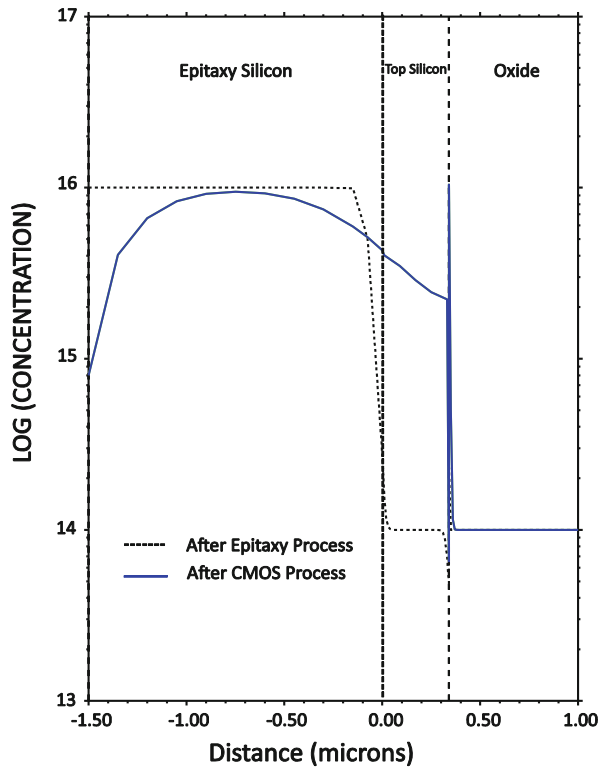


Fig. 3.15 PDP of frontside and backside illumination comparison between devices with/without co-integrated CMOS process [29]

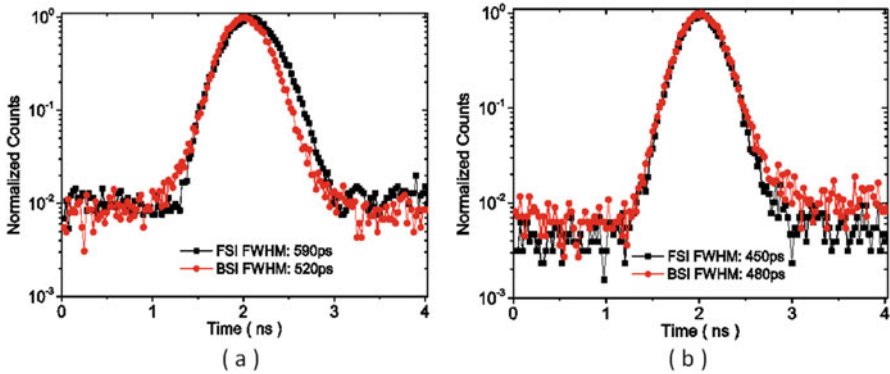
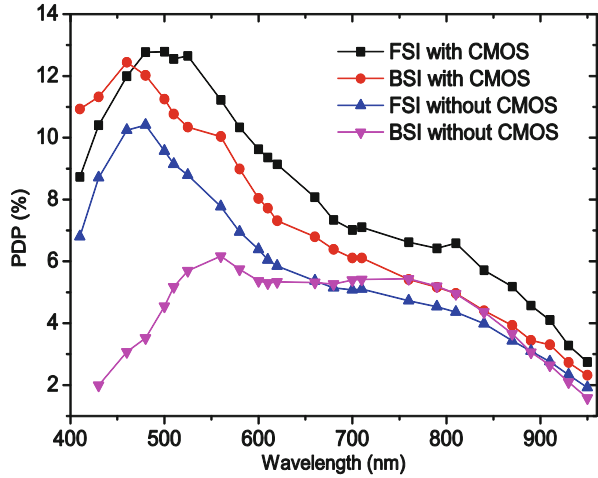


Fig. 3.16 Timing jitter measurements of flexible CMOS SPAD sensor pixel in FSI and BSI (a) using 405 nm laser and (b) using 637 nm laser [29]

3.3 A Flexible 32 × 32 Single-Photon Avalanche Diode Image Sensor

In this section, we extend the flexible CMOS SPAD sensor to array format, in which the first flexible 32 × 32 SPAD image sensor with in-pixel and off-pixel electronics integrated in CMOS is presented.

3.3.1 Sensor Architecture

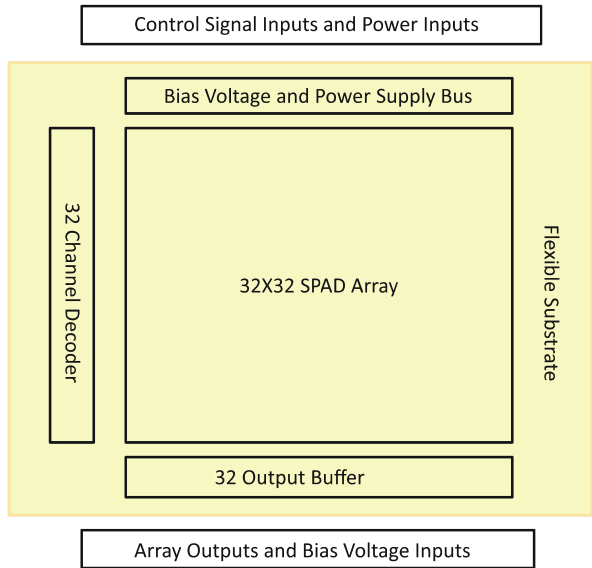
The functional diagram of the flexible CMOS SPAD sensor comprising 32 × 32 pixels is shown in Fig. 3.17. The sensor uses two power supplies: operational voltage

Table 3.4 Performance comparison between flexible SPAD, planar SPAD, and pixel [29]

Performance	Flexible trench-isolated SPAD ^a			Planar SPAD			Flexible CMOS SPAD pixel			Unit
	Min.	Typ.	Max.	Min.	Typ.	Max.	Min.	Typ.	Max.	
V_{bd}		17.2			25.5			26.5		V
DCR density at RT ($V_E = 1.5$ V)	3.5		6.3	0.1		0.3		4.3		$\text{kHz}/\mu\text{m}^2$
DCR density at 80 K ($V_E = 1.5$ V)		0.4						0.15		$\text{kHz}/\mu\text{m}^2$
PDP at FSI			14.3			13.0			13.0	%
PDP at BSI			6						12.5	%
Afterpulsing probability	0.5						0.15		1.9	%
Timing jitter at FSI (637 nm wavelength)		900						450		ps
Timing jitter at BSI (637 nm wavelength)		680						480		ps

^aConfiguration includes only SPAD and R_q . No CMOS transistor fabrication process

Fig. 3.17 Functional architecture of the flexible CMOS SPAD sensor



V_{OP} and digital CMOS circuit power V_{DD} of 3.3 V. The readout circuitry consists of a 5-to-32 channel decoder for row addressing. Each column has an output buffer. The main sensor array is implemented on flexible substrate, while the input and output pads are designed on the peripheral silicon substrate frame.

As shown in Fig. 3.18a, the pixel consists of a SPAD, a quenching resistor, and 4-transistor circuit which functions as buffer and switch. The N^+ cathode is biased to a high operation voltage V_{OP} , around 25 V, which is common to all the pixels

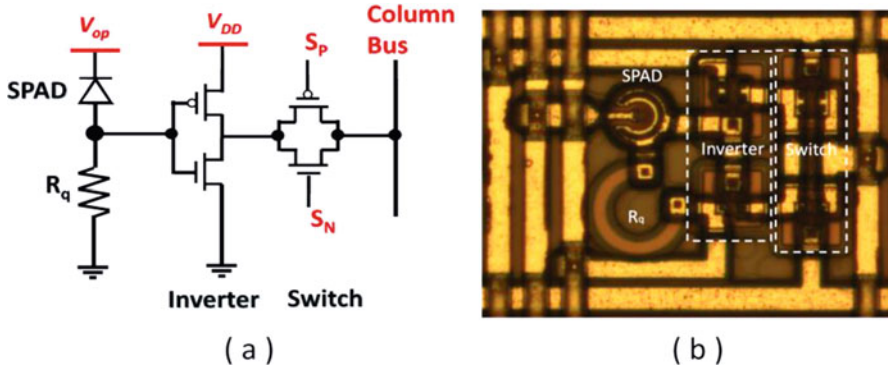


Fig. 3.18 (a) Functional architecture of the flexible CMOS SPAD sensor. (b) Micrograph of a pixel

in the array. The size of single pixel is $162 \times 125 \mu\text{m}$, and the microscopic image is shown in Fig. 3.18b. The inverter stage converts the Geiger-mode voltage pulse to a digital pulse. The switch stage, which is a transmission gate, feeds the digital pulse to output bus when the row is addressed. All the functional circuitries are implemented on flexible substrate.

3.3.2 Fabrication Flow

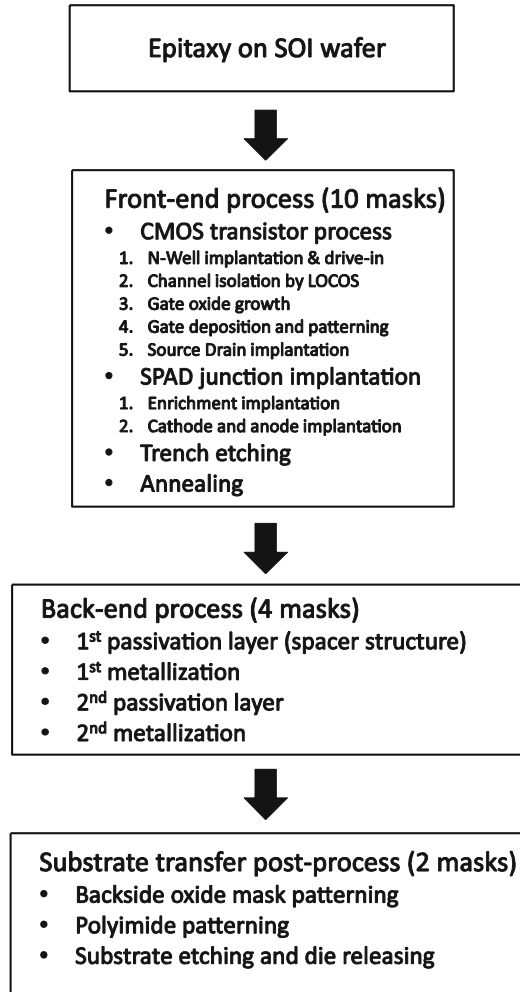
Based on the fabrication in Fig. 3.11, the flowchart is developed as shown in Fig. 3.19. It starts from silicon epitaxy on SOI wafer lightly doped with Boron. The front-end process includes CMOS transistor fabrication, which consists of N-well implantation and thermal drive-in, LOCOS isolation, gate oxide growth, gate patterning, and source and drain implantations. SPAD junction is formed by three implantations. Trenches are etched for device isolation and wafer are sent to anneal. It takes 10 lithography masks in total to implement the front-end process.

Different from fabrication in Fig. 3.11, the back-end process is much more complicated because two-layer metallization needs to be implemented on top of the device array with a topography step of $1.85 \mu\text{m}$. Etch-back and redeposition of passivation layers is used to form spacers during the topography steps. According to the layout, back-end process parameters, such as spacer dimension, thicknesses of different passivation layers, and metal layers, are optimized. Overall, there are four lithography masks in back-end process.

The images of metallization and cross section are shown in Fig. 3.20, (a) and (b), respectively.

After the back-end process, the chip is functional as SOI and followed by substrate transfer post-process, which would take two lithography masks. After oxide deposition on backside, the first lithography, which is also implemented on backside, forms the patterns for substrate etching. Polyimide is then patterned to

Fig. 3.19 Flowchart diagram of flexible CMOS SPAD sensor fabrication



expose SPAD sensitive area and pad area. Substrate etching would be implemented by using DRIE as shown in Fig. 3.21a, and die was released by reasonable mechanical force, which is shown in Fig. 3.21b.

3.3.3 Array Characterization

As shown in Fig. 3.22a, the SOI chip was mounted onto CPG-144 package by wire bonding. After substrate transfer post-processing, presented in Chap. 10, the flexible chip was released as shown in Fig. 3.22b. To characterize the flexible chip, it also needed to be mounted onto special designed chip board and wire bonded to pads. The front- and backside images are shown in Fig. 3.22, (c) and (d), respectively [32].

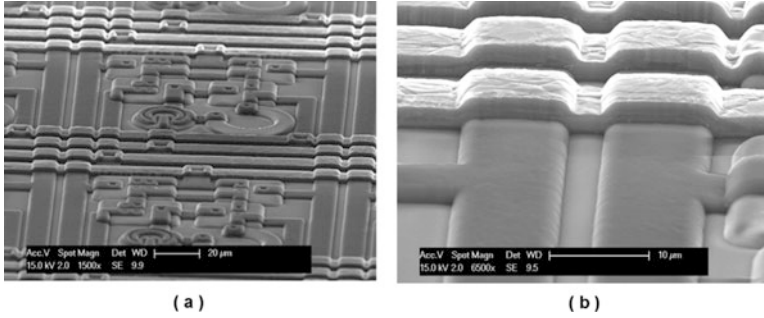


Fig. 3.20 (a) SEM image of second metallization. (b) SEM image of the second metal layer at topography step

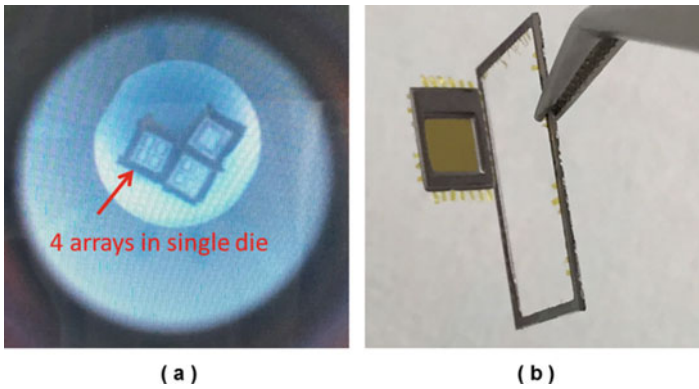


Fig. 3.21 (a) Image during substrate etching in plasma. (b) Last one of the four arrays releasing from a die after substrate etching

The DCR measurements were performed under different V_{OP} , which ranges from 22 to 24 V. DCR distribution histograms over an array under different operation voltages are measured and compared as shown in Fig. 3.23. In general, the DCR distribution was modeled and fitted with a Gaussian distribution, excluding the hot spots. The mean value of DCR increases as the operation voltage increases, and the FWHM value of distribution becomes larger [33].

By cooling down the flexible chip from 22.7 °C (room temperature) to -70 °C, mean value of DCR with V_{eb} around 4 V was reduced from 200 to 30 kHz exponentially with the temperature. FWHM of distribution histogram became smaller as the temperature went down, as shown in Fig. 3.24.

A contour image of PDP nonuniformity across the 32×32 array is shown in Fig. 3.25a, when the sensor was illuminated with a uniform light. DCR has been measured prior to the illumination and subtracted from every pixel across the whole chip. A contour image of V_{BD} across the chip is shown in Fig. 3.25b, and V_{BD} of SPAD in each pixel is extracted by “DCR fit” method [33].

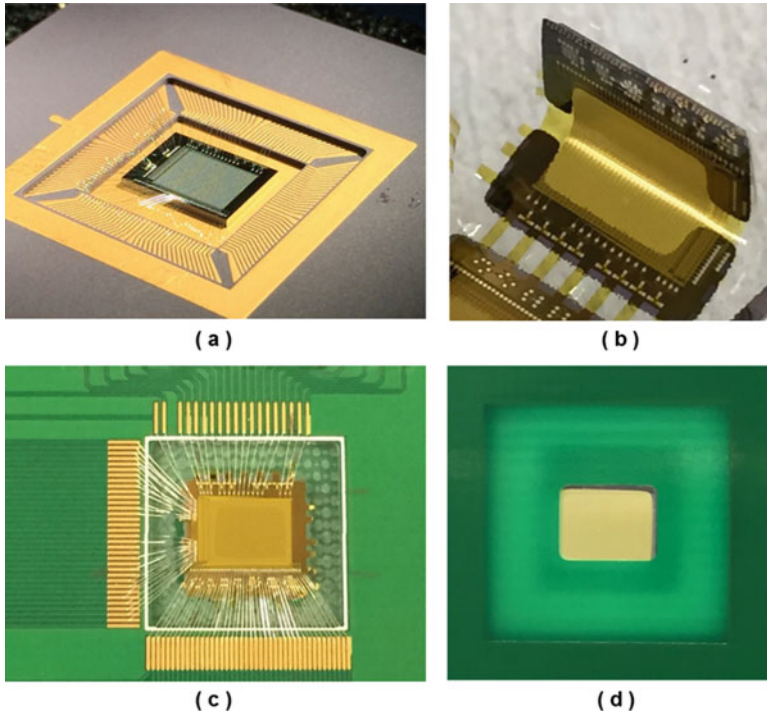


Fig. 3.22 (a) SOI chip mounted in CPG-144; (b) bendable chip released after substrate etching; (c) frontside of flexible chip mounted onto PCB; (d) backside of flexible chip mounted onto PCB

Cross talk is negligible due to the fact that both electrical and optical isolations are suppressed in the flexible chip. Electrical isolation is ensured by the fact that SPADs and ancillary circuitry are SOI trench-isolated structures. Optical isolation is ensured by large pixel pitch, when compared with the multiplication region.

It is further proved by gathering histogram information on inter-arrival times between the counts of two adjacent pixels, as shown in Fig. 3.26a. The inter-arrival time histogram shows a distribution fitted well with an exponential, proving negligible cross talk probability. The inter-arrival times between the counts of two nonadjacent pixels also show negligible cross talk probability, as shown in Fig. 3.26b.

3.3.4 Dual-Side Imaging

A compact signal processing module was developed containing chip board, USB communication board, and FPGA board. Camera modules are mounted on both

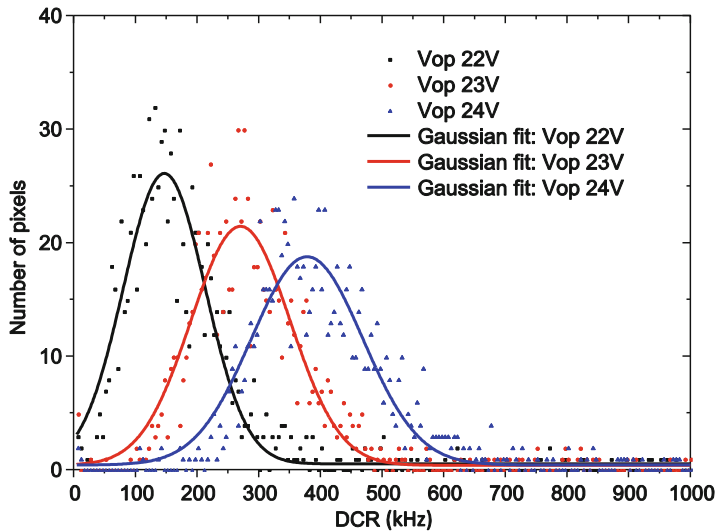


Fig. 3.23 DCR distribution histograms at different V_{OP} [32]

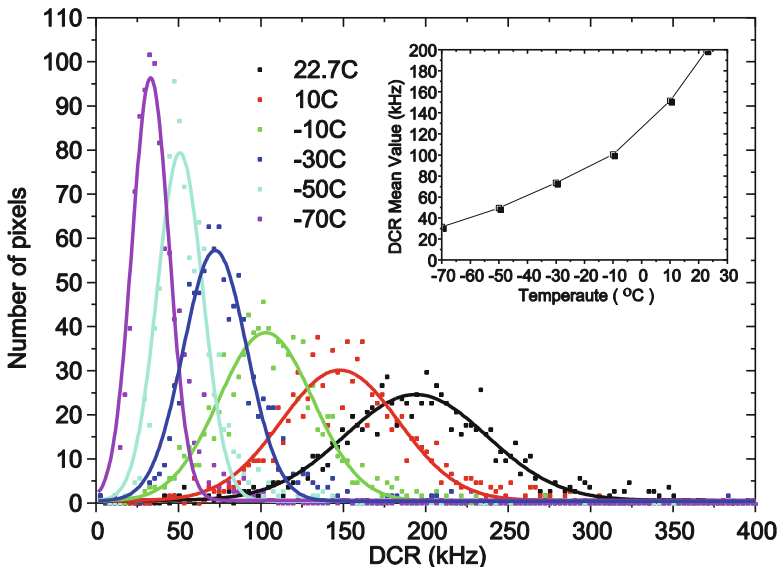


Fig. 3.24 DCR distribution histograms at different temperatures (*inset*: DCR mean value at different temperatures) [32]

front- and backsides. The detail setups of dual-side imaging camera are shown in Fig. 3.27. Light source is composed of surface light together with mask (TUD logo). It is mounted on a moving stage during imaging and resolution enhancement experiments.

Fig. 3.25 (a) PDP nonuniformity across the sensor; (b) V_{BD} across the sensor [32]

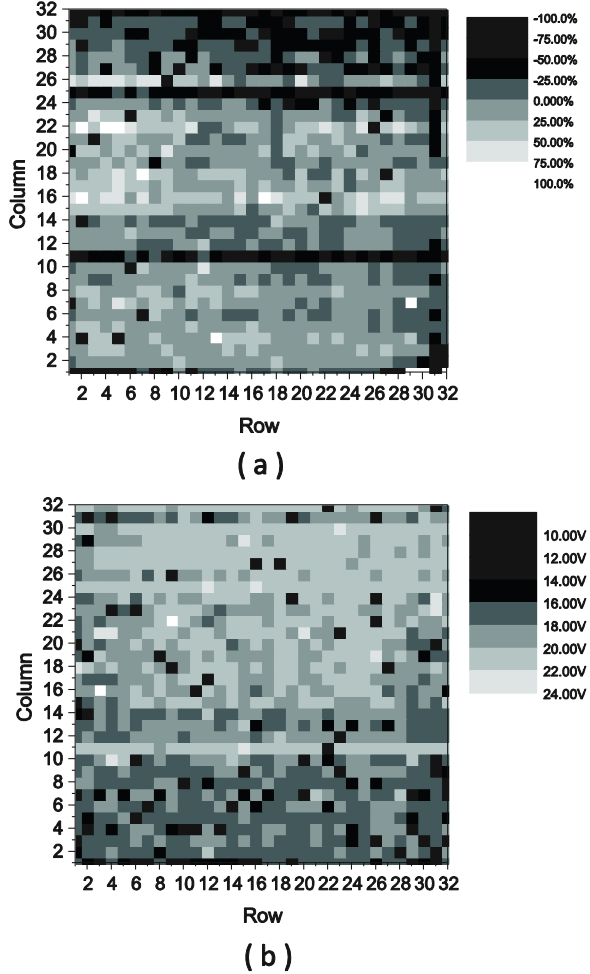


Figure 3.28 shows the imaging results of FSI and BSI, respectively. The 32×32 chip was scanned at frame rate of 32 ms/frame and readout by 32 column buffers parallel. By optimizing the optic setup, the imaging from both sides is comparable.

3.3.5 Resolution Enhancement

As shown in Fig. 3.29, by shifting the imaging object by half pixel dimension in X , Y , and both X and Y directions, multiple frames could be integrated together to get an image with resolution enhanced by two times in both X and Y directions [34].

This method is applied in the imaging experiments with flexible CMOS SPAD sensor chip, and different resolutions such as 64×64 and 128×128 imaging

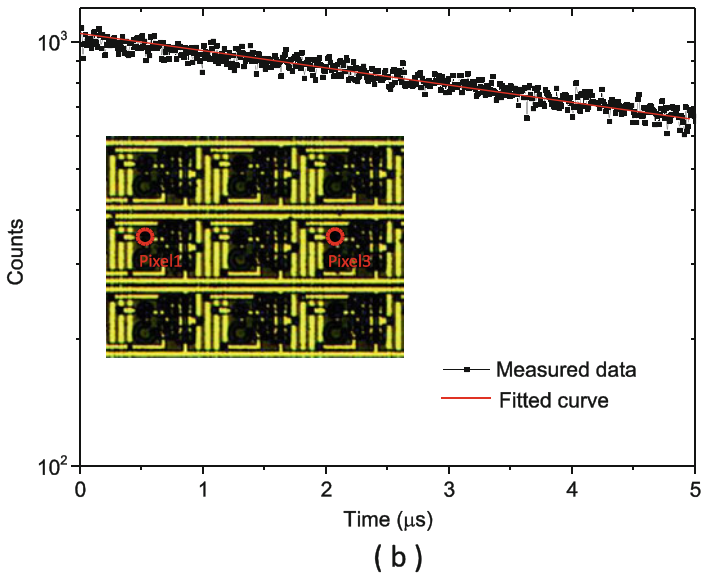
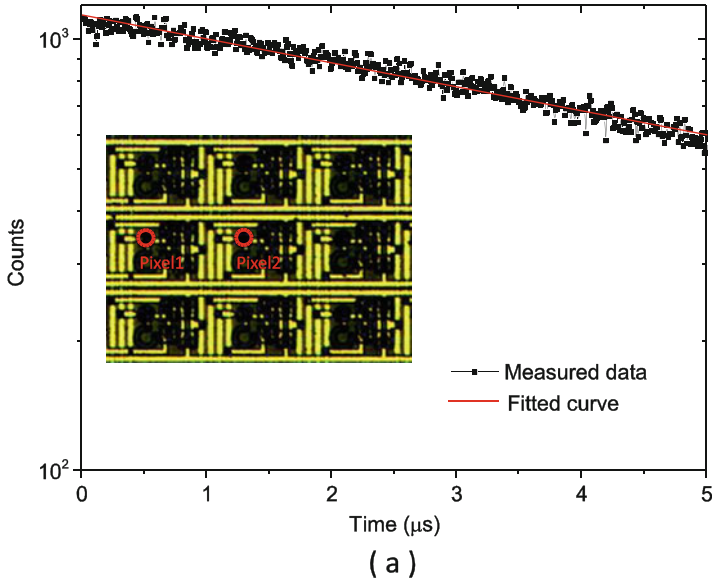


Fig. 3.26 Cross-inter-arrival time histogram measured from (a) two adjacent pixels; (b) two nonadjacent pixels

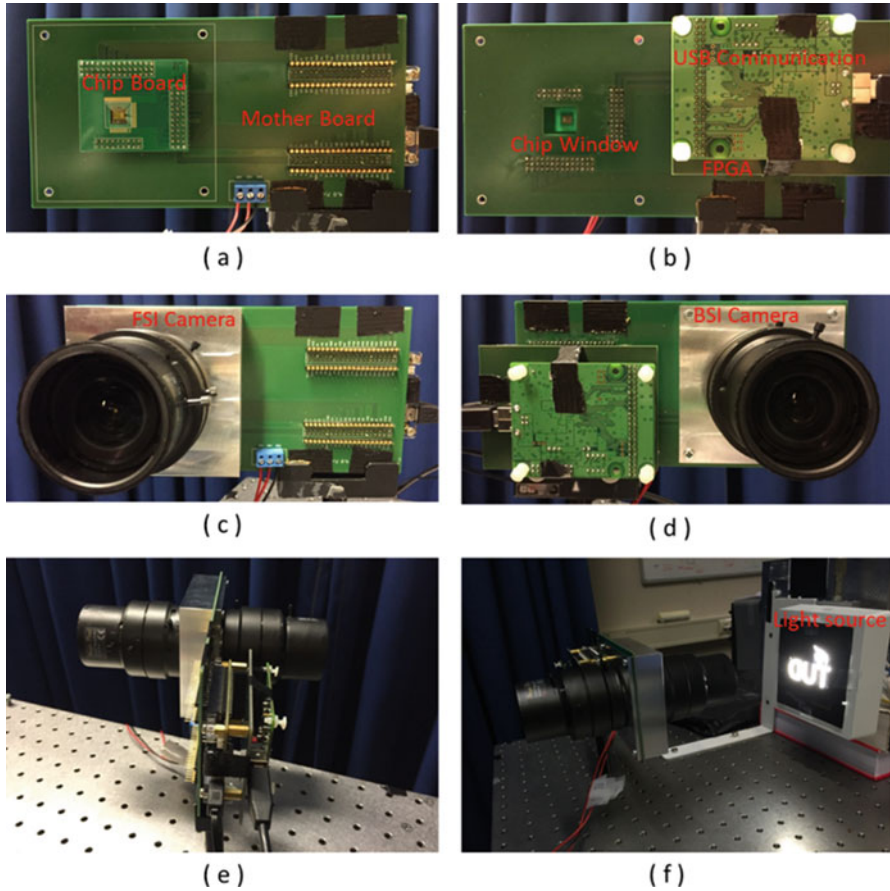


Fig. 3.27 (a) Frontside imaging signal processing module. (b) Backside of signal processing module. (c) Frontside imaging camera module. (d) Backside imaging camera module. (e) Dual-side imaging compact module. (f) Imaging setup

Fig. 3.28 “TUD” logo imaging on flexible CMOS SPAD image sensor. (a) Frontside imaging. (b) Backside imaging



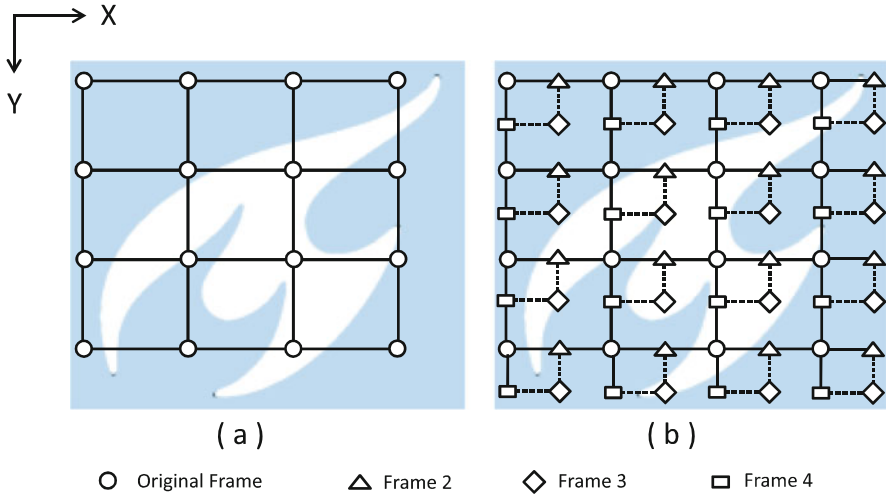


Fig. 3.29 Resolution enhancement method. *Frame 2*: shifting half pixel dimension in X direction. *Frame 3*: shifting half pixel dimension in both X and Y directions. *Frame 4*: shifting half pixel dimension in Y direction

results are achieved, which is shown in Fig. 3.30. Imaging results based on different exposure time are also compared. Higher resolution and longer exposure time enable higher contrast and imaging quality.

The electrical and optical performance is summarized in Table 3.5.

4 Conclusions

The primary goal of this work was to explore and develop a novel single-photon avalanche diode technology in applications of flexible and implantable biomedical electronics for which we defined four contributions. In order to meet the requirements, focus was placed on the investigation of the potential features of the system, such as ultrathin body, front- and backside illumination modes, high fill factor and sensitivity, flexible substrate post-processing, and CMOS integration compatibility. The challenges of this investigation have been successfully addressed. The world's first flexible ultrathin-body SPAD was demonstrated and extended to flexible 32×32 CMOS SPAD image sensor level. The state of the art in SPAD technology, as well as in flexible biomedical sensors, has been considerably enhanced.

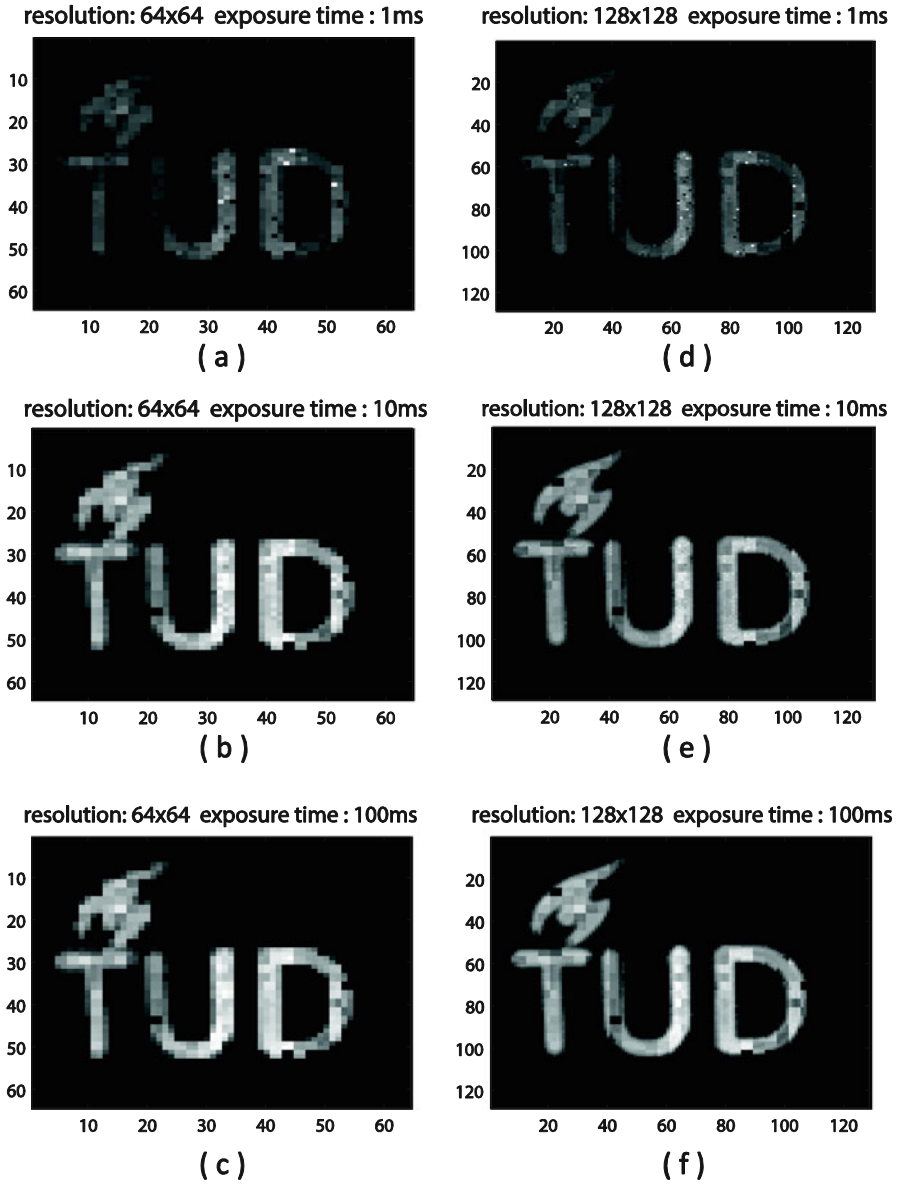


Fig. 3.30 Comparing of imaging based on different resolution and exposure time

Table 3.5 Performance summary of flexible CMOS SPAD image sensor chip

Item	Min.	Typ.	Max.	Unit	Comments
Resolution	32 × 32		128 × 128		
Row rate		1 k		fps	
Frame rate		30		fps	
Digital supply voltage		3.3		V	
SPAD V_{OP}	22		25	V	
Power dissipation			30	mW	Entire flexible chip
DCR		120		kHz	$V_{OP} = 22$ V
DCR temperature dependence	1.5X/20 °C				
V_{BD} nonuniformity		11%			
PDP nonuniformity		22.1%			At 560 nm; $V_{OP} = 23$ V
PRNU		19.7%			
Signal-noise ratio		36		dB	
Dynamic range		25		dB	Estimated up to 52 dB by active quenching at low temperature
Detectable light intensity	3.5			μ W/cm ²	

References

1. A. Webb, G.C. Kagadis, Introduction to biomedical imaging. *Med. Phys.* **30**(8), 2267 (2003)
2. L.V. Wang, H.-I. Wu, *Biomedical Optics: Principles and Imaging* (Wiley, Chicester, 2012)
3. A.F. Laine, In the spotlight: biomedical imaging. *IEEE Rev. Biomed. Eng.* **1**, 4–7 (2008)
4. Acikel, V., & E. Atalar, Intravascular magnetic resonance imaging (MRI). In *Biomedical Imaging: Applications and Advances* (Elsevier Inc., 2014), pp. 186–213
5. S. Chua, A. Groves, *Biomedical Imaging*. (Elsevier, 2014)
6. Z. Gorocs, A. Ozcan, On-chip biomedical imaging. *IEEE Rev. Biomed. Eng.* **6**, 29–46 (2013)
7. C.M. Tempny, B.J. McNeil, Advances in biomedical imaging. *JAMA* **285**(5), 562–567 (2001)
8. W. Becker, A. Bergmann, G. Biscotti, A. Rueck, Advanced time-correlated single photon counting technique for spectroscopy and imaging of biological systems. *Proc. SPIE Commer. Biomed. Appl. Ultrafast Lasers IV* **5340**, 1–9 (2004)
9. N. Bertone, M. Wabuyele, A. Kapanidis, H. Dautet, M. Davies, “Single photon counting with a focus on biomedical applications”, Application Note PerkinElmer Optoelectronics. (2003)
10. S. Lebid, R. O’Neill, C. Markham, T. Ward, S. Coyle, Functional brain signals: a photon counting system for brain activity monitoring, in *IEEE Conference Proceedings of the Irish Signals and Systems Conference* (Ireland, 2004), pp. 469–474
11. J. Ohta, Implantable CMOS imaging devices for bio-medical applications, in *IEEE 54th International Midwest Symposium on Circuits and Systems* (Seoul, Korea, 2011), pp. 1–4
12. J. Ohta, T. Tokuda, K. Sasagawa, T. Noda, Implantable CMOS biomedical devices. *Sensors* **9**(11), 9073–9093 (2009)
13. M. S. Humayun, J. D. Weiland, G. Chader, E. Greenbaum, *Artificial Sight*. (Springer, 2007)
14. T. Tokuda, M. Takahashi, K. Uejima, K. Masuda, T. Kawamura, Y. Ohta, M. Motoyama, T. Noda, K. Sasagawa, T. Okitsu, S. Takeuchi, J. Ohta, CMOS image sensor-based implantable glucose sensor using glucose-responsive fluorescent hydrogel. *Biomed. Opt. Express* **5**(11), 3859–3870 (2014)

15. T. Noda, K. Sasagawa, T. Tokuda, Fabrication of fork-shaped retinal stimulator integrated with CMOS microchips for extension of viewing angle. *Sensors Mater.* **26**(8), 637–648 (2014)
16. G. Park, H.J. Chung, K. Kim, S.A. Lim, J. Kim, Y.S. Kim, Y. Liu, W.H. Yeo, R.H. Kim, S.S. Kim, J.S. Kim, Y.H. Jung, T.-i. Kim, C. Yee, J.A. Rogers, K.M. Lee, Immunologic and tissue biocompatibility of flexible/stretchable electronics and optoelectronics. *Adv. Healthc. Mater.* **3**(4), 515–525 (2014)
17. J. Yoon, S.M. Lee, D. Kang, M.A. Meitl, C.A. Bower, J.A. Rogers, Heterogeneously integrated optoelectronic devices enabled by micro-transfer printing. *Adv. Opt. Mater.* **3**(10), 1313–1335 (2015)
18. S.G. Wu, C.C. Wang, B.C. Hseih, Y.L. Tu, C.H. Tseng, T.H. Hs R.S. Hsiao, S. Takahashi, R.J. Lin, C.S. Tsai, Y.P. Chao, K.Y. Chou, P.S. Chou, H.Y. Tu, F.L. Hsueh, L. Tran, A leading-edge 0.9 μ m pixel CMOS image sensor technology with backside illumination: future challenges for pixel scaling, in *Technical Digest – International Electron Devices Meeting, IEDM* (San Francisco, USA, 2010)
19. P. Webb, A. Jones, Large area reach-through avalanche diodes for radiation monitoring. *IEEE Trans. Nucl. Sci.* **21**(1), 151–158 (1974)
20. M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, E. Charbon, A low-noise single-photon detector implemented in a 130 nm CMOS imaging process. *Solid State Electron.* **53**(7), 803–808 (2009)
21. M.A. Karami, M. Gersbach, H.-J. Yoon, E. Charbon, A new single-photon avalanche diode in 90nm standard CMOS technology. *Opt. Express* **18**(21), 22158–22166 (2010)
22. R.K. Henderson, E.A. G. Webster, R. Walker, J.A. Richardson, L.A. Grant, A 3 \times 3, 5 μ m pitch, 3-transistor single photon avalanche diode array with integrated 11V bias generation in 90nm CMOS technology, in *Technical Digest – International Electron Devices Meeting, IEDM* (San Francisco, USA, 2010)
23. F. Zappa, S. Tisa, A. Tosi, S. Cova, Principles and features of single-photon avalanche diode arrays. *Sensors Actuators A Phys.* **140**(1), 103–112 (2007)
24. H. Jansen, H. Gardeniers, M. De Boer, M. Elwenspoek, J. Fluitman, A survey on the reactive ion etching of silicon in microtechnology. *J. Micromech. Microeng.* **6**, 14–28 (1996)
25. K.R. Williams, K. Gupta, M. Wasilik, Etch rates for micromachining processing – Part II. *J. Microelectromech. Syst.* **12**(6), 761–778 (2003)
26. P. Sun, B. Mimoun, E. Charbon, R. Ishihara, A flexible ultra-thin-body SOI single-photon avalanche diode. *Int. Electron Devices Meet.* **11**(1), 284–287 (2013)
27. P. Sun, E. Charbon, R. Ishihara, A flexible ultrathin-body single-photon avalanche diode with dual-side illumination. *IEEE J. Sel. Top. Quantum Electron.* **20**(6), 276–283 (2014)
28. S. Nikzad, T.J. Cunningham, M.E. Hoenk, R.P. Ruiz, D.M. Soules, S.E. Holland, Direct detection of 0.1–20 keV electrons with delta doped, fully depleted, high purity silicon p-i-n diode arrays. *Appl. Phys. Lett.* **89**(18) (2006)
29. P. Sun, R. Ishihara, E. Charbon, Flexible ultrathin-body single-photon avalanche diode sensors and CMOS integration. *Opt. Express* **24**(4), 3734–3748 (2016)
30. P. Sun, R. Ishihara, E. Charbon, A flexible 32 \times 32 SPAD image sensor with integrated microlenses, in *International Image Sensor Workshop (IISW)*, Session 11, paper 3 (2015)
31. J. Pavia, M. Wolf, E. Charbon, Measurement and modeling of microlenses fabricated on single-photon avalanche diode arrays for fill factor recovery. *Opt. Express* **22**(4), 4203–4213 (2014)
32. P. Sun, J. Weng, R. Ishihara, E. Charbon, A flexible 32 \times 32 dual-side single-photon image sensor, in *International Image Sensor Workshop (IISW)*, Session 5, paper 6 (2017)
33. I.M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, E. Charbon, Nonuniformity analysis of a 65-kpixel CMOS SPAD imager. *IEEE Trans. Electron Devices* **63**(1), 1–8 (2015)
34. B. Sajadi, D. Qoc-Lai, A.T. Ihler, M. Gopi, A. Majumder, Image enhancement in projectors via optical pixel shift and overlay, in *IEEE International Conference on Computational Photography* (Cambridge, USA, 2013), pp. 1–10

Chapter 4

CMOS Multimodal Sensor Array for Biomedical Sensing

Kazuo Nakazato

1 Introduction

The integration of biochemistry on a chip, where several biochemical reactions are controlled and detected electrically, may find a lot of biomedical applications, such as home healthcare, tailor-made medicine, food security, evidence-based care, blocking infectious disease at immigration, drug discovery, and so on (Fig. 4.1). Electrical detection using complementary metal-oxide semiconductor (CMOS) integrated circuits has great potential since it eliminates the labeling process and achieves high-accuracy and real-time detection. It also offers the important advantages of low-cost and compact equipment, which leads to a portable diagnostic inspection system that anyone can operate anywhere, to obtain immediate results. Such portable diagnostic inspection system becomes very powerful through internet connection with a main database.

Particularly, a configuration of matrix sensor array realizes the parallel detection of different kinds of biomolecules, real-time two-dimensional images of biomolecular interactions, and the improvement in accuracy by analyzing statistical distribution and combining different kinds of sensing methods.

Electrochemical biosensors are devices that combine a biological component (a recognition part) and a physicochemical detector component (a transducer). The recognition part can be constructed using enzymes, antibodies, cells, tissues, nucleic acids, and peptide nucleic acids. The transducer consists of potentiometric, amperometric, and impedimetric sensor circuits. Electric potential, electric current, and impedance have their own advantages and disadvantages in sensing and give complementary information of molecules. Electric potential is not influenced by

K. Nakazato (✉)

Department of Electronic Engineering, Graduate School of Engineering, Nagoya University,
Nagoya, Japan

e-mail: nakazato@nuee.nagoya-u.ac.jp

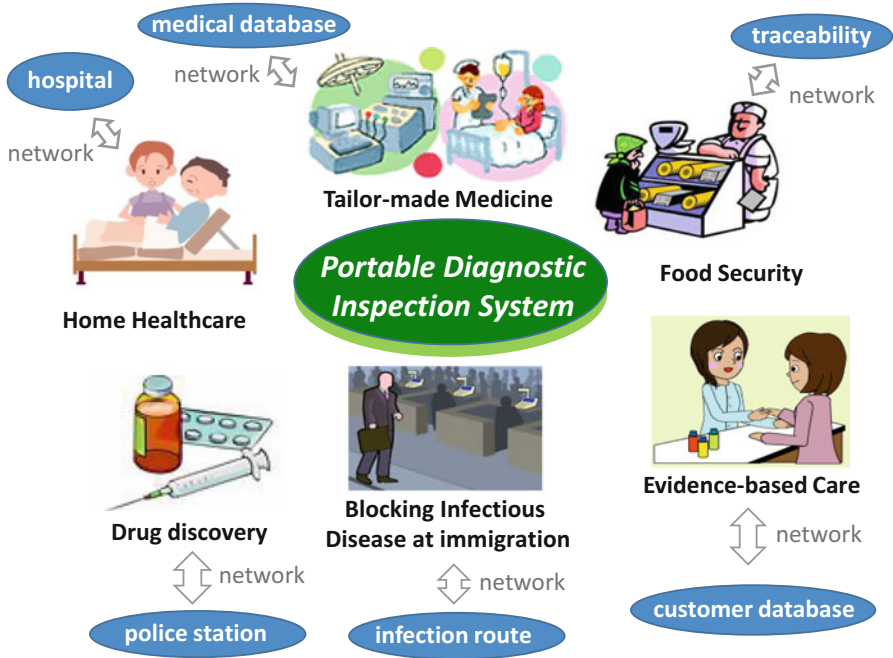


Fig. 4.1 Portable diagnostic inspection system and its applications

electrode size, which means that the sensor cell is scalable. However, the signal has logarithmic dependence on the concentration of molecules, which makes it difficult to detect DNA sequence in homopolymer region. The electric current has linear dependence but is influenced by electrode size, so strongly affected by contaminants on the electrode. Impedance is suitable to the detection of volume, and the electrode does not need to be in contact with electrolyte solution. However, the system becomes rather complicated to choose input frequency and amplitude and analyze output amplitude and phase.

On the other hand, a photonic sensor is still a powerful tool to detect biochemical interactions. CMOS multimodal sensor arrays can choose the advantages of each sensing scheme and improve the reliability by the combination.

2 Design of CMOS Multimodal Sensor Array

One of the design principles of our sensor array is that the structure must be compatible with standard CMOS integrated circuits in order to supply stable, uniform, and low-price chips. We use the 6-in., 0.6- μ m, two-polysilicon, three-metal mixed signal CMOS process of the Taiwan Semiconductor Manufacturing Company (TSMC).

Gold is a standard electrode material used in electrolyte solutions because of the low ionization tendency and the formation of self-assembled monolayer through thiols. However, gold cannot be introduced into the standard CMOS process since it creates deep-level electron traps in silicon. Post-CMOS processes to form gold electrodes, positive light-sensitive polyimide protection layers, and SU-8 (an epoxy-based negative resist) microfluidics are added by MEMS CORE Co., Ltd.

Another one of our design principles is that the cell circuits should not influence the sensed system. To sense natural chemical reactions, not disturbed by the sensor circuit, we have developed a source-drain follower for the potentiometric sensor to decrease the capacitive disturbances [1] and dual switch for the amperometric sensor to keep the electrode potential always constant [2].

Since the signal from biological interaction is quite different from the signal handled in information and technology, new circuit technologies must be developed. For the biological sensing circuits, we set three principles: operation in the subthreshold region, current-mode circuit, and time-domain signal representation. Operation in the subthreshold region reduces the power consumption to avoid heating the chip itself, since biochemical interactions are sensitive to temperature. Current-mode circuits have lower noise and lower power consumption compared to voltage-mode circuits since the charging and discharging of interconnection are substantially reduced. Time-domain signal has wider dynamic range compared to the amplitude-domain signal, which becomes lower and lower according to the scale-down of transistors. Several new circuits have been developed such as cascode source-drain follower operated in the subthreshold region [1], current-mode analog-to-digital converter [3–6], current-mode analog-to-time converter [7, 8], and so on.

For the detection of large numbers of molecules, averaging of signals is very effective to increase the signal-to-noise ratio, based on the law of large numbers in probability and statistics. Current signal can be easily time integrated by charging a capacitor. On the other hand, event-driven circuits are suitable for single-molecule detection. A neuromorphic circuit using AER (Address Event Representation) is one of the candidates [9].

2.1 Potentiometric and Photometric Sensor Array

The detection of electric potential change based on a field-effect transistor (FET) [10] has shown excellent sensitivity for ion concentration [11, 12] and specific DNA sequences, including single-nucleotide polymorphisms (SNPs) [13]. Rothberg et al. demonstrated a genome sequencing chip that contains 13 million pH sensors on a $17.5 \times 17.5 \text{ mm}^2$ die [14]. On the other hand, photonic sensors have been used as a standard method of biochemical sensing such as immunoassay and bioluminescent assay [15]. The combination of potentiometric and photometric sensors could increase the accuracy and applications.

Since biological interaction is strongly affected by temperature, the power consumption of the sensor chip should be lower than 10 mW to avoid heating

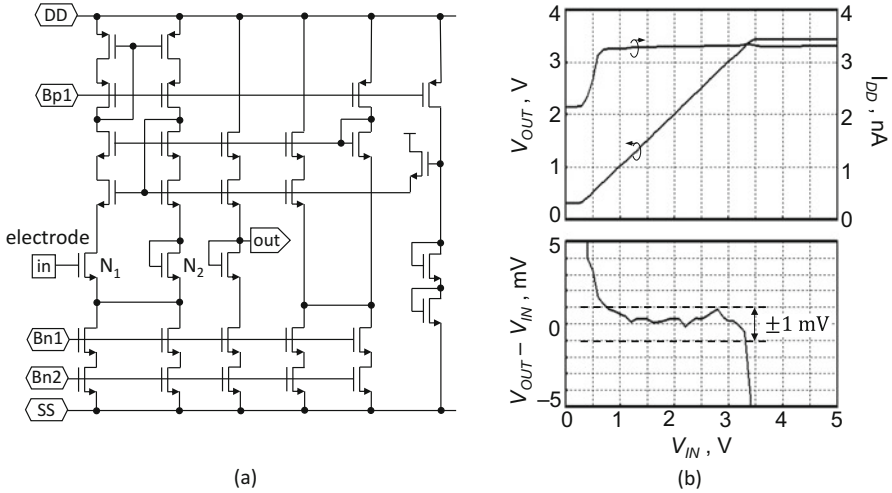


Fig. 4.2 Cascode source-drain follower circuit for potentiometric sensor cell; (a) circuit schematic and (b) measured characteristics

the chip itself. One of the methods to reduce the power consumption is to design circuit operated in the subthreshold region. Figure 4.2 shows a cascode source-drain follower as a potentiometric sensor cell [1]. In the front-end transistor, N_1 , V_{GD} (voltage between gate and drain), and V_{GS} (voltage between gate and source) are kept constant, so the capacitances C_{GD} and C_{GS} become negligible, resulting in less influence to the sensed system. Since source, drain, body voltages, and drain current in transistors N_1 and N_2 are the same, the gate voltages become the same, resulting in $V_{out} = V_{in}$. In this voltage follower configuration, electrode voltage can be processed in the following circuits without disturbing the sensed system, and the output is less influenced by the threshold voltage and temperature. Figure 4.2b shows the transfer characteristics. The measured circuit includes a common bias circuit shown in Fig. 4.3b. The power consumption of the part of Fig. 4.2a is as low as 10 nW. The trade-off of the low power consumption is the slow response time, which is inversely proportional to the power consumption and around 1 ms at 10 nW. The potentiometric sensor cell in 64×64 sensor array is designed at 250 nW, so the total power is 1 mW, and one frame can be acquired in 0.2 s.

Figure 4.3 shows the photonic sensor part. Transistor N_1 is used to fix the voltage in photodiode P. Transistor N_2 keeps N_1 in the saturation region. The photocurrent i is copied and outputted by a wide-swing cascode current mirror. The bias circuit is shown in Fig. 4.3b. Current signal is useful for low-noise and low-power circuits since the charging and discharging of capacitors and capacitive-coupled cross talk can be reduced. Furthermore, current can be time integrated by a capacitor as charge. Time integration is one of the simplest and most powerful methods to reduce noise since it works as a low-pass filter (averaging filter). Furthermore, $\Delta-\Sigma$ analog-to-digital converter (ADC) can be constructed as shown in Fig. 4.4. When switch C is

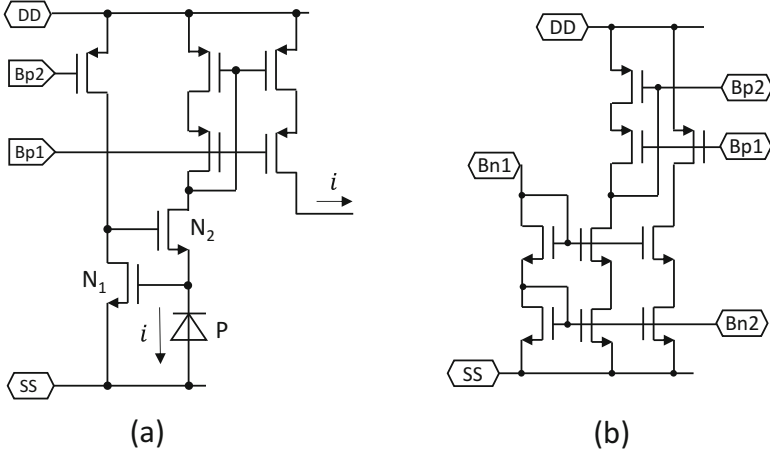


Fig. 4.3 Circuit schematic of (a) photonic sensor cell and (b) wide-swing cascode current mirror to supply bias voltages

high, the input current is time integrated by a capacitor C_i . Since the input voltage is fixed at V_G by virtual short, the output voltage of operational amplifier V_o is given by

$$V_o = V_i - \frac{1}{C_i} \int_0^t i(\tau) d\tau,$$

where V_i is the initial V_o as shown in Fig. 4.4c. When V_o becomes lower than V_L , output of comparator, V_C , becomes high, and a finite amount of charge is injected into the input node [5]. Figure 4.4b shows the part of 1b DAC, constructed by a D-type flip-flop followed by a nonoverlapping clock generator and a charge pump: when V_C becomes high, nonoverlapping clock is generated, and charge pump injects the following charge to the input node:

$$\Delta Q = C_r (V_p - V_G - V_{os}),$$

where V_{os} is the offset voltage of op-amp on the input side. Input current is time integrated without input-signal loss, even during charge injection. Coarse digital signal is obtained by counting the number of charge injections, n . After a defined time, switch C is set to low, and time integration is stopped. The initial and final voltages, V_i and V_f , are detected by increasing V_L by a counter and a DAC. Fine digital signal is obtained by counting the number of steps, m , as shown in Fig. 4.4d. The range of current measurements can be adjusted by the integration time. When $C_i = 1$ pF and 10 bits counter, current between 10 pA and 1 μ A can be converted with 10 pA resolution by 1 ms integration time. Sub-pA resolution can be obtained by increasing the integration time [16].

In the current-mode ADC, there are three modes, C (coarse), S (send), and F (fine). C mode is the sampling state where switch C is set to high. S is the state that

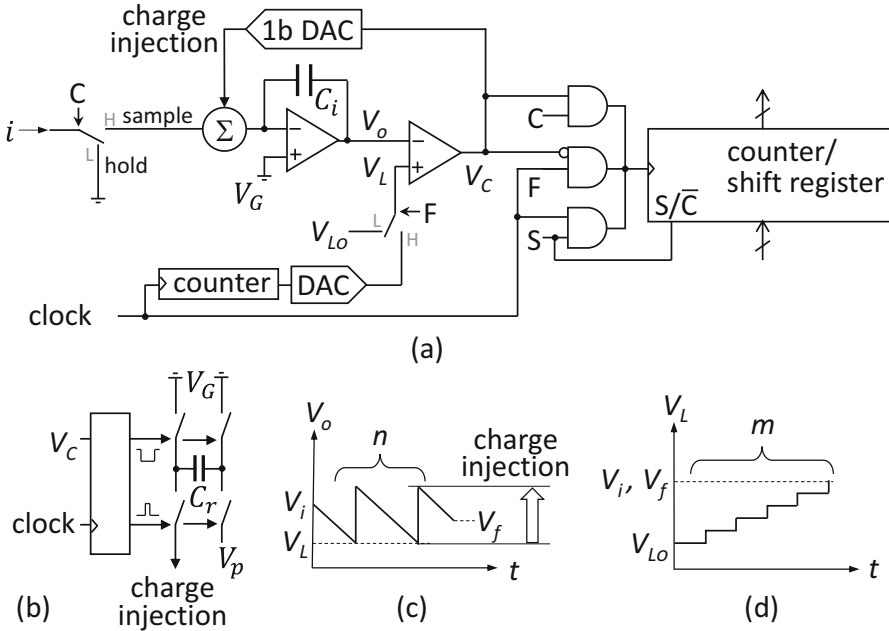


Fig. 4.4 Current-mode $\Delta-\Sigma$ ADC. (a) Schematic diagram, (b) 1b DAC, (c) coarse digital conversion, and (d) fine digital conversion

counter-values are outputted by the shift register. F is the state to detect V_i or V_f . The current-mode ADC takes the following sequence: F S C S F S. Since the difference between V_f and V_i is used, an undesired offset voltage caused by switching can be canceled, a kind of correlated double sampling (CDS).

Figure 4.5 shows the 64×64 potentiometric and photonic sensor array where 64 current-mode ADCs are integrated, one per one column. One selected row signals are converted to digital signals in parallel and outputted in serial. The power consumption was 10 mW, mainly consumed by current-mode ADCs. Potential and photonic two-dimensional images are simultaneously obtained at 3 fps (frames per second).

2.2 Potentiometric, Amperometric, Impedimetric, and Photometric Sensor Array ASSP (Application-Specific Standard Product)

In large-scale integration circuit fabrication, the initial cost for making a set of photomasks is quite high. On the other hand, the chip cost is extremely low

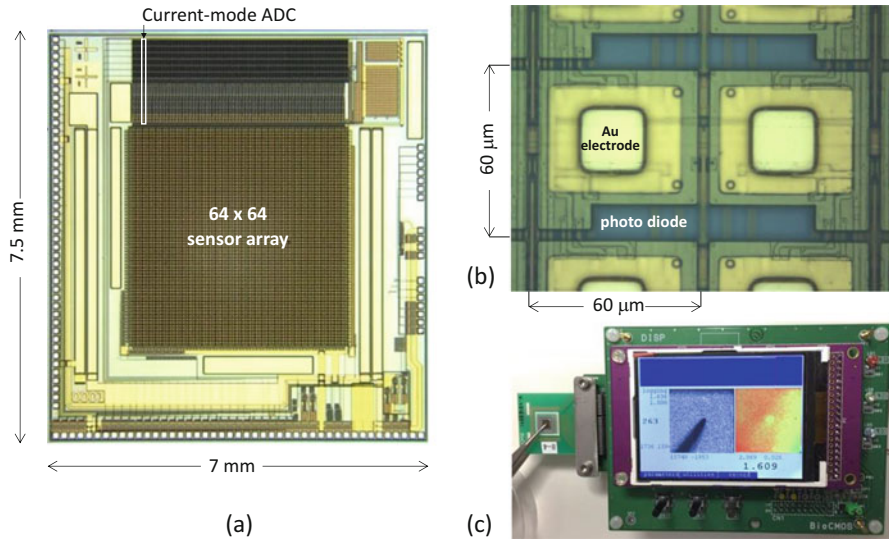


Fig. 4.5 A 64×64 potentiometric and photometric sensor array. Photomicrograph of (a) chip and (b) sensor cell. (c) Operation of the chip. RMS noise of the potentiometric sensor is 0.5 mV. Photocurrent between 10 pA and 1 μ A with 10 pA resolution can be detected by the photometric sensor. Power supply and total current are 5 V and 2 mA, respectively. One frame is acquired in 0.3 s

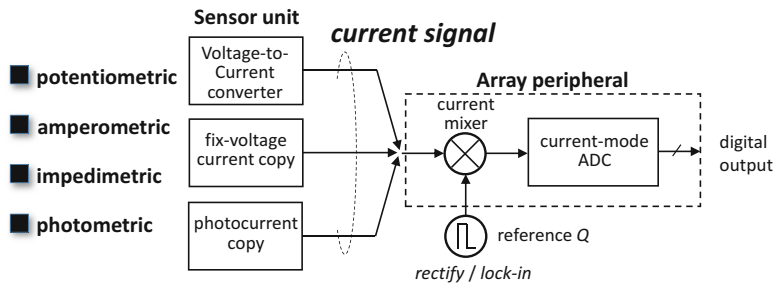


Fig. 4.6 An ASSP (application-specific standard product) integrating multimodal sensors. All sensor units output current signal, which are rectified and converted to digital signals in current mode

if a large number of chips are produced. This means that standardization and general-purpose sensor chips are important. Customization can be achieved by post-CMOS processing using contact photolithography.

One of the examples of application-specific standard product (ASSP) is shown in Fig. 4.6. Potentiometric, amperometric, impedimetric, and photometric sensor units transfer the current signal to the array peripheral circuit, where the current signal is rectified or locked in by a current mixer and time integrated by a current-mode ADC.

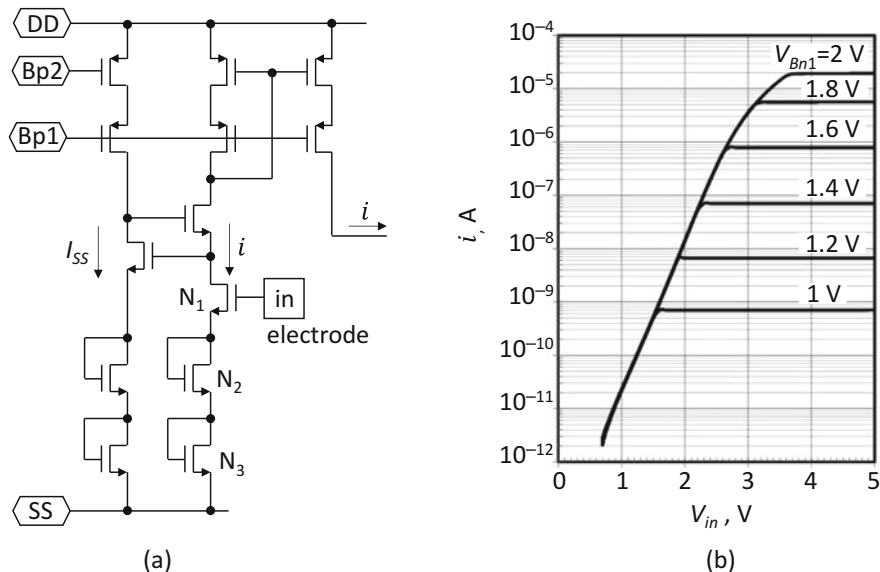


Fig. 4.7 Voltage to current converter for potentiometric sensor unit; (a) circuit schematic and (b) measured characteristic. Current I_{SS} is supplied by common bias circuit shown in Fig. 4.3b. Two source degeneration transistors N_2 and N_3 expand the input voltage range triple, and the limit of total current is controlled by voltage V_{Bn1}

First, voltage to the current converter for the potentiometric sensor cell is discussed. MOSFET is a transconductor which converts voltage on the input port to the current on the output port. However, there are two problems when using MOSFET alone as a voltage to current converter. One problem is that the input voltage range is rather small if it is operated in the subthreshold region where large current change can be obtained. The second problem is that a large current can flow and the power consumption cannot be controlled when the gate voltage accidentally becomes extremely high. Voltage to current converter shown in Fig. 4.7 solves these problems. Two source degeneration transistors, N_2 and N_3 , connected to the source of the front-end transistor N_1 expand the input voltage range triple, and the current flowing N_1 , i , is controlled by I_{SS} since N_1 enters into the triode region when $i \gg I_{SS}$. The measured current voltage characteristic is shown in Fig. 4.7b where I_{SS} is supplied through the bias circuit in Fig. 4.3b. The maximum current of i is controlled by V_{Bn1} . In the case of $V_{Bn1} = 1.6$ V, the input voltage range is 2 V, and the current limit is 800 nA. The measured total input-referred noise was 0.45 mV with current integration at every 1 ms [17]. One of the problems of this circuit is that the sensed signal is influenced by the threshold voltage variation. A simple method to correct the threshold voltage variation is calibration by attaching a switch transistor to the electrode. However, in such configurations, the electrode becomes non-floating. In the situation to detect electronic charge around the electrode, only the transient change can be detected since the leakage current of the switch transistor

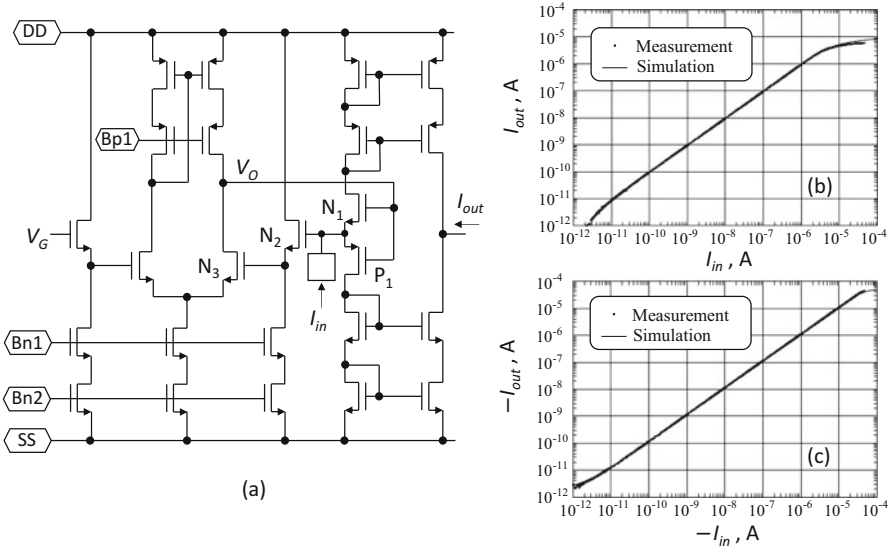


Fig. 4.8 (a) Circuit schematic of amperometric/impedimetric sensor unit. Electrode voltage is fixed at V_G , and the current flowing into electrode is copied to output. Measured characteristics for (b) $I_{in} > 0$ and (c) $I_{in} < 0$

will bring electronic charge to the electrode so as to compensate for external charge. Redox potential, discussed in Sect. 3.1, can be obtained as an absolute value, since the chemical equilibrium keeps the electrode voltage against the leakage.

Figure 4.8 shows amperometric and impedimetric sensor circuit where electrode voltage is fixed at V_G , and the current flowing into electrode I_{in} is copied and transferred to the followed circuit. When $I_{in} > 0$, the current flows P_1 . When $I_{in} < 0$, the current flows N_1 . Transistors N_1 and P_1 work as source followers to keep electrode voltage at V_G . Level shifter N_2 is needed to keep N_3 in the saturation region. Figure 4.8b, c shows the measured transfer characteristics when (b) $I_{in} > 0$ and (c) $I_{in} < 0$ [18]. A wide range of current was detected by this circuit. One of the problems of this circuit is that the response becomes extremely slow when the input current changes the polarity since V_o is changed by the sum of the absolute values of NMOSFET and PMOSFET threshold voltages, nearly 2 V.

For the mixer circuit, a Gilbert circuit is usually used as shown in Fig. 4.9a [19]. In this circuit, currents i_1 and i_2 are converted to the voltages by resistors, and the DC part is obtained by the followed low-pass filter. Instead of conversion to voltage, mixed current can be obtained by Gilbert circuit and folded cascode as shown in Fig. 4.9b where

$$\begin{aligned}
 q &= +1 && \text{when } Q = \text{high} \\
 q &= -1 && \text{when } Q = \text{low}
 \end{aligned}$$

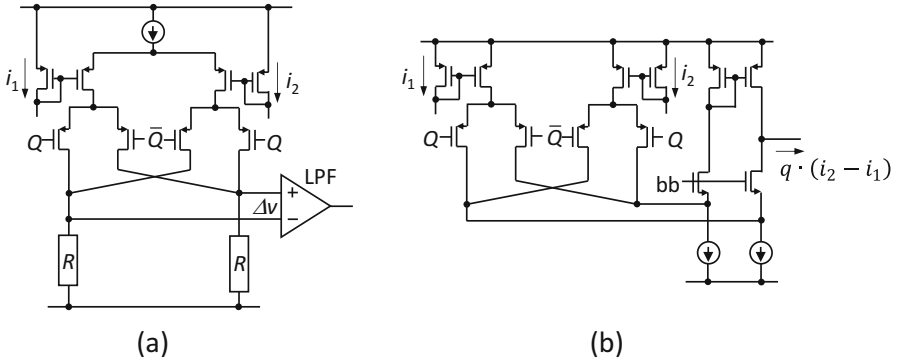


Fig. 4.9 Mixer circuitry. (a) voltage-mode Gilbert mixer, (b) current-mode mixer

The mixed current $q(i_2 - i_1)$ is averaged by the followed current-mode ADC. By introducing complex voltage v and current i ,

$$v = \text{Re}[\mathbf{v}e^{j\omega t}], \quad i = \text{Re}[\mathbf{i}e^{j\omega t}],$$

where v is the input RF voltage and $i = i_2 - i_1$, and complex admittance is defined by

$$\mathbf{i} = \mathbf{Y} \cdot \mathbf{v}.$$

After current mixer, averaged current is obtained as

$$\langle i \rangle(\varphi) = \frac{1}{2\pi} \oint d(\omega t) q \cdot i = \frac{2}{\pi} [\text{Re}[\mathbf{Y}] \cos(\varphi) + \text{Im}[\mathbf{Y}] \sin(\varphi)] v,$$

where φ is the phase of Q relative to the input RF voltage v . Measurements are performed at three different phases of φ , $-\pi/4$, $\pi/4$, and $3\pi/4$. The real and imaginary parts of admittance are obtained by taking the differences between two values,

$$\text{Re}[\mathbf{Y}] = \frac{\sqrt{2}\pi}{4v} \left[\langle i \rangle\left(\frac{\pi}{4}\right) - \langle i \rangle\left(\frac{3\pi}{4}\right) \right],$$

$$\text{Im}[\mathbf{Y}] = \frac{\sqrt{2}\pi}{4v} \left[\langle i \rangle\left(\frac{\pi}{4}\right) - \langle i \rangle\left(-\frac{\pi}{4}\right) \right].$$

The difference makes it possible to eliminate undesired offset effects. The elimination is a kind of CDS.

Figure 4.10 shows the ASSP chips formed for a microtiter plate which consist of 8×12 wells separated by 9 mm. After the fabrication of a 6-in. wafer by standard CMOS process, interconnections among $9 \times 9 \text{ mm}^2$ chips are formed by contact photolithography in post-CMOS process. Each chip has input and output registers

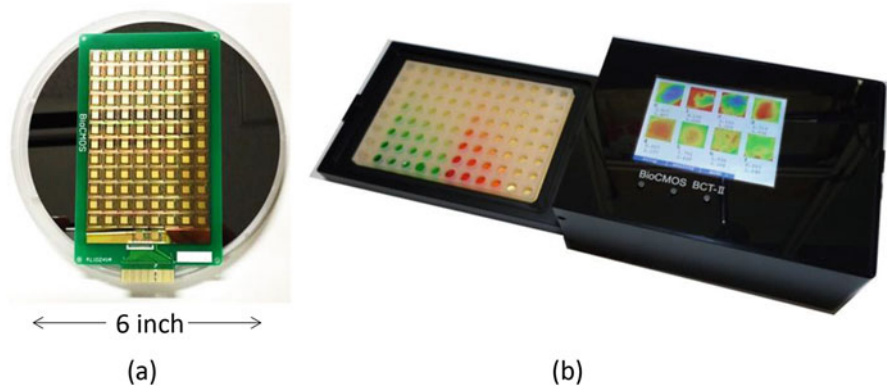


Fig. 4.10 Wafer-scale chip for 8×12 microtiter plate. Interconnection among 96 chips are formed by contact photolithography. Each chip has registers to select input and output

which determine the acceptance of input signals and output of the sensed signals, respectively. Microtiter plate is a standard for biological experiments, and wide varieties of equipment are available.

2.3 High-Density Potentiometric, Amperometric, and Impedimetric Sensor Array

For the high-density sensor array, two transistor cells are designed. Figure 4.11a shows amperometric/impedimetric sensor cells. Dual-switch configuration [2] is applied to keep electrode voltage always constant, in this case, 2.2 V, since, when the electrode voltage is changed, it takes a few, or tens of, seconds before reaching a steady state of redox current; rapid multipoint measurement cannot be achieved with a simple switching scheme. Figure 4.11b shows potentiometric sensor cells. In both sensors the cell size and electrode size are $4.15 \times 4.85 \mu\text{m}^2$ and $1 \times 1 \mu\text{m}^2$ on mask, respectively, and the sensed signals are transferred as current signals to the array peripheral circuit where a set of current mixer and current-mode ADCs are formed as 1 per 16 columns as shown in Fig. 4.11c. One frame can be acquired in 2 s at $250 \mu\text{s}$ integration time.

Gold electrodes are commonly used in electrolyte solutions because gold has the lowest ionization tendency. However, since gold creates deep levels in silicon, gold cannot be introduced into standard CMOS processes. Thus, gold electrodes must be formed by the post-CMOS process. However, since the post-CMOS process is usually constructed based on contact photolithography with line/space of several microns, self-aligned processes are required for the formation of submicron gold electrodes.

Electroless plating is one such self-aligned process. Figure 4.12 shows (a) the flowchart, (b) photomicrograph, and (c) scanning electron micrograph of the

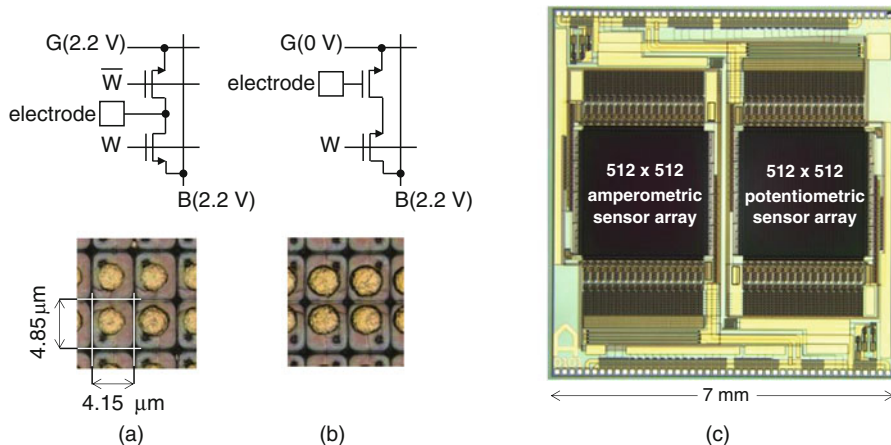


Fig. 4.11 High-density sensor array, (a) amperometric/impedimetric sensor cell, (b) potentiometric sensor cell, (c) experimental chip of 512×512 amperometric sensor array and 512×512 potentiometric sensor array

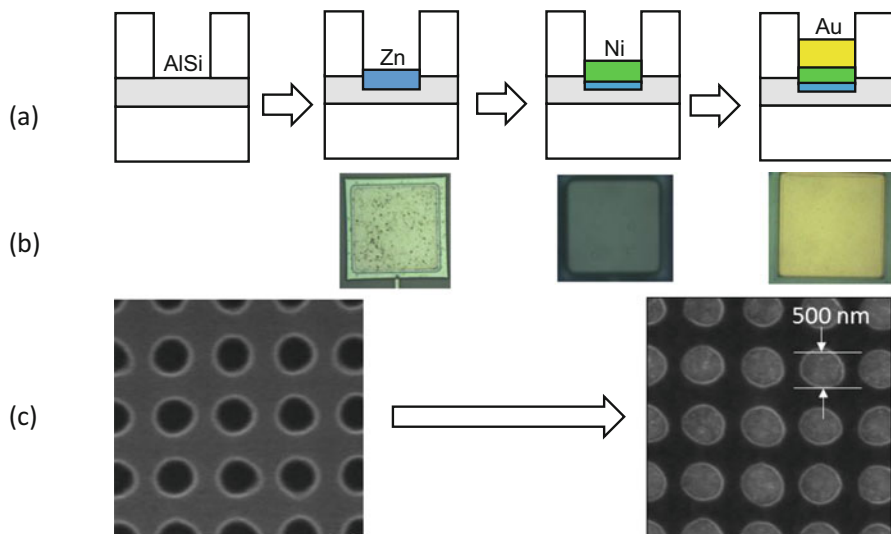


Fig. 4.12 Electroless plating to form submicron gold electrodes. (a) Top interconnection Al-Si layer formed by standard CMOS process is converted by Au by three electroless plating steps, from AlSi to Zn, from Zn to Ni, and from Ni to Au. Conversions are performed by the difference of ionization tendency without any electric field nor light excitation. (b) Photomicrographs after each steps. (c) Scanning electron micrograph before and after electroless plating

electroless gold plating to realize 500 nm microelectrode array based on [20], combining with preprocessing [21] and de-smutting [22]. The average of absolute values of surface roughness R_a was 21 nm using mild acidity zincate solution.

3 Applications for Biomedical Sensing

3.1 Enzyme Sensor with Redox Mediator

There are two principles of the potentiometric sensor. One principle is the detection of electronic charge around an electrode with no electron transfer to the electrode. We call this method “charge detection.” The gate potential is determined by Poisson’s equation. First, a probe layer is formed on an FET. Then, target molecules are supplied. Specific molecules are selectively taken into the probe layer on the FET channel, which detects the molecular charge in the probe layer. In the case of DNA detection, the probe is single-stranded (ss) DNA with a known sequence, immobilized on the substrate. When the target ssDNA is supplied, specific hybridization occurs if the target DNA is complementary to the probe DNA. Occurrence, or nonoccurrence, of specific hybridization can be detected by the difference in charge, since a nucleotide has a negative charge on the phosphate group.

The other principle is the detection of chemical equilibrium potential, i.e., redox potential, accomplished by electron exchange between the electrolyte/molecule and the electrode [23–26]. In this case, the gate potential is determined by the Nernst equation.

The charge detection method has a number of serious problems with the stability to detect charge in solutions. First, the molecular charge is screened by ions in a solution. Screening length is around 3 nm in the case of an ion concentration of 10 mM. This can be extended if a low ion concentration is used; however, in this case, a very high impedance environment is produced, and the relaxation of electric potential becomes long. Second, the charge distribution is influenced by the shape of the molecule. It is generally understood that single-stranded DNA takes a Gaussian shape and double-stranded DNA takes a rodlike shape. It is unclear whether the detected signal is caused by a change in charge or a change in structure. Especially in a flow system, the molecular shape fluctuates, which leads to unstable electric potential. Third, the electrode enters a floating state. Although UV irradiation reduces the threshold voltage variation [27], embedded charge causes a large threshold voltage variation.

Instead of using the charge detection method, a redox potential detection method was developed using a ferrocenyl-alkanethiol modified gold electrode [23], as shown in Fig. 4.13. This redox potential sensor detects the ratio of reductant to oxidant concentration from the following Nernst equation:

$$V = V_0 + \frac{k_B T}{nq} \ln \left(\frac{[\text{Ox}]}{[\text{Red}]} \right),$$

where V is the electrode potential, V_0 is the standard electrode potential, k_B is the Boltzmann constant, T is the absolute temperature, n is the number of transition electrons, q is the elementary charge, $[\text{Ox}]$ is the oxidant concentration, and $[\text{Red}]$ is the reductant concentration.

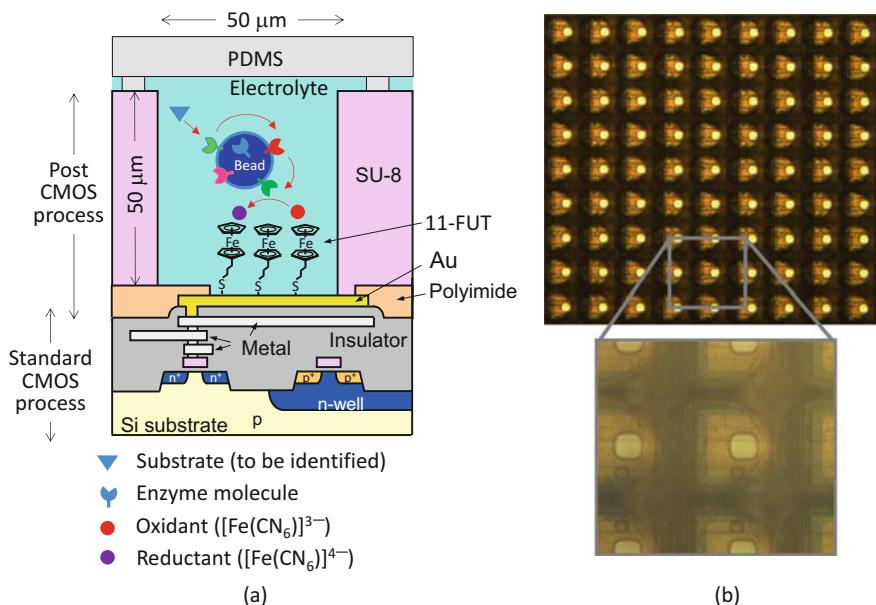


Fig. 4.13 Enzyme sensor with a redox mediator; (a) schematic cross section and (b) photomicrograph of redox potential sensor array. Substrate concentration is projected to the ratio of oxidant and reductant concentrations by enzyme reaction. 11-FUT SAM (self-assembled monolayer) is a good electron mediator and protects electrode from electrolyte solution

We fabricated a chip that integrates a 32×32 redox potential sensor array [25]. The sensor chip was dipped in $500 \mu\text{M}$ 11-ferrocenyl-1-undecanethiol (11-FUT) ethanol for 24 h. A hexacyanoferrate mixture totaling 10 mM was used for the oxidant and reductant. Six orders of concentration ratios of reductant to oxidant were detected by this sensor array. The sensitivity at a temperature of 25°C was 59 mV/decade which is exactly equal to the value predicted by the Nernst equation. Experiments showed that the potential drift can be drastically reduced to less than 0.5 mV/h by the redox potential detection method, compared to 30 mV/h in the charge detection method.

The redox potential sensor array was applied to an enzyme sensor with a redox mediator as shown in Fig. 4.13a. The enzyme sensor was composed of enzyme reaction part and reaction detection part. When a substrate (target) exists in solution, the substrate and oxidoreductase will react in the solution. Therefore, oxidant will change to reductant. In the present case, the oxidant is $[\text{Fe}(\text{CN}_6)]^{3-}$, and the reductant is $[\text{Fe}(\text{CN}_6)]^{4-}$.

In Fig. 4.13, biotin-immobilized enzyme molecules are bound to avidin-immobilized microbeads with a diameter around $30 \mu\text{m}$. A SU-8 trench was developed for the purpose of improving the portability of the microbeads. In addition to high-accuracy transportation, the trench enables the simultaneous

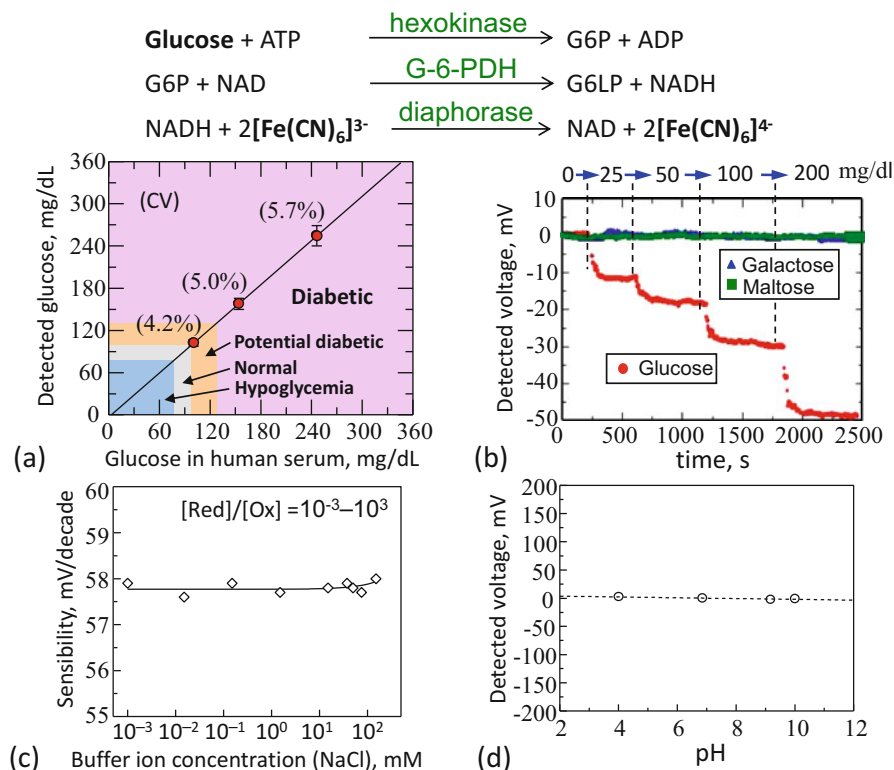


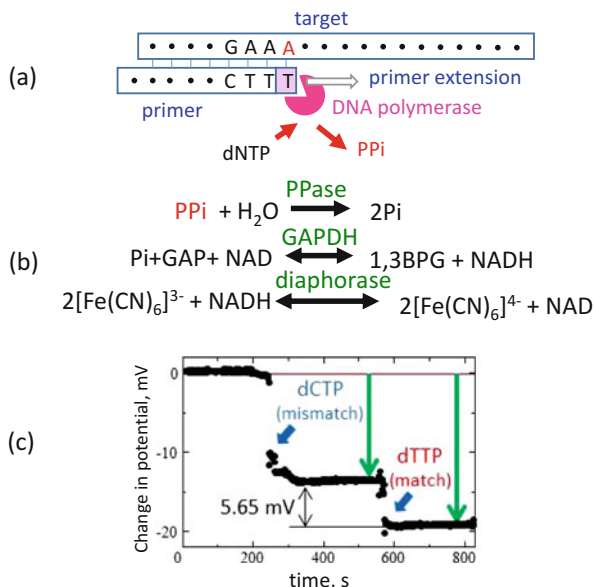
Fig. 4.14 Glucose sensor. (a) Detected glucose vs. given glucose in human serum (JCCRM 521-11) obtained from the Reference Material Institute for Clinical Chemistry Standards, Japan. (b) Change in gate potential when glucose, galactose, and maltose samples are supplied. (c) Dependence of buffer ion concentration. (d) pH sensitivity

detection of different objects and rapid detection in an array. This configuration reduces the amount of enzyme molecules and achieves multipurpose use of the sensor cell.

Figure 4.14 shows a glucose sensor as one of the examples. Experiments were performed by mixing glucose in enzyme solution on the chip without microbeads [25]. This redox potential sensor array successfully detected a human serum glucose level with coefficient of variation around 5%. The detection limit of glucose concentration was less than $1 \mu\text{M}$.

It is important to confirm that the detected signal is glucose specific, and not caused by other factors. Figure 4.14b shows the detection signals when glucose, galactose, and maltose samples are supplied, showing no sensitivity to galactose and maltose, which are usually contained in human blood, and cause interference in conventional glucose sensors. The redox potential method was not affected by buffer ion concentration (Fig. 4.14c) or pH (Fig. 4.14d), so this glucose sensor is applicable to a wide range of sample conditions.

Fig. 4.15 (a) Principle of DNA polymerization. (b) Enzyme-catalyzed reaction. When DNA polymerization occurs and PPI is released, it is converted into redox agents with the help of three enzymes: PPase, GAP, and diaphorase. (c) Measured time course of potential change in DNA single-base polymerization. Tested base sequences are 5'-CACAC TCACA GTTTT CACTT-3' for primer and 3'-GTGTG AGTGT CAAAA GTAA ATGAG ATAGC-5' for target



Next, experiments were performed using microbeads [28]. By introducing enzyme-immobilized microbeads, the amount of enzyme can be dramatically reduced. Three enzymes, hexokinase, G-6-PDH, and diaphorase, are immobilized on a bead by biotin-avidin binding. During the measurements using different glucose concentrations for longer than 3 h, the reaction rate did not degrade; thus, the activity of the enzyme was maintained. The reusability of enzyme-immobilized microbeads and sensor cells was demonstrated throughout the 11 measurements in interval of every 30 min.

As for the other example, the redox potential method was applied to the DNA sequencer by detecting DNA single-base polymerization [29] as shown in Fig. 4.15. After hybridization of primer and target DNAs, dNTP and Taq DNA polymerase are introduced for primer extension. When dNTP matches the sequence of the target, pyrophosphate (PPI) is released (Fig. 4.15a). Released PPI is detected by enzyme reaction (Fig. 4.15b) using three enzymes, PPase, GAPDH, and diaphorase [30]. Figure 4.15c shows the measured time course of potential change in DNA single-base polymerization. First, the reference solution (100 mM Tris-HCl (pH 8.8), 9.9 mM $\text{K}_3[\text{Fe}(\text{CN})_6]$, and 0.1 mM $\text{K}_4[\text{Fe}(\text{CN})_6]$) was introduced to the solution flow cell. Next, after about 250 s from introducing the reference solution, the measurement solution (reference solution plus 1.6 mM MgCl_2 , 2 mM NAD, 2 mM GAP, 10 U/ml PPase, 10 U/ml GAPDH, and 10 U/ml diaphorase) with mismatched base (0.4 mM dCTP and 20 U/ml Taq DNA polymerase) was introduced. The interfacial potential changed in spite of mismatching between the template DNA probe and dCTP. Using molybdenum blue method, we found that 2 mM GAP contains about 70 μM Pi which is the cause of the interfacial potential change after

introducing the measurement solution. Finally, about 300 s after introducing the measurement solution with mismatch base, the measurement solution with matched base (dTTP) was introduced. There was a 5.65 mV difference between the reaction solutions with a mismatched dNTP and those with a matched dNTP, which achieves a sufficient signal-to-noise ratio of more than 20 dB. The dialysis of reagents is effective for eliminating the effect of Pi in the measurement solution, and the actual signal of DNA polymerization can be detected.

3.2 Counting Bacteria/Viruses One by One

One of the applications of the amperometric sensor array is the direct counting of viruses and bacteria [31]. When an electrode captures a virus or bacterium, the amount of redox current decreases, and the existence of the virus or bacterium can be detected. To achieve the single detection rule that electrode and target are the same size, microelectrodes are formed by electroless plating.

Figure 4.16 shows the measured CV using 10 mM $K_4[Fe(CN)_6]$ and Na_2SO_4 [32]. Figure 4.16a shows the results of the median of the 16×16 array with 6 μm square microelectrodes when the current was measured seven times. The measured current is almost identical during all seven measurements, which showed that the electroless-plated microelectrode is robust enough. The measured peak currents among the first, third, and seventh measurements are 5.3 ± 0.2 nA, which is close to 7.55 nA of the theoretical value of microdisk electrodes. The difference may be explained by trench structure. Figure 4.16b shows the histogram of the measured peak currents in the first measurement. All 16×16 electrodes functioned well. The histogram in all 16×16 electrodes of the limiting current forms a normal distribution, and its standard deviation is 0.3 nA.

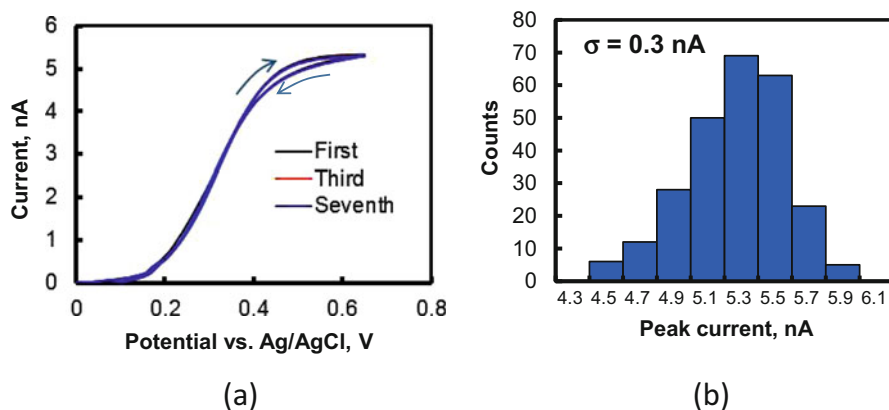


Fig. 4.16 Measured results of cyclic voltammetry (CV). (a) Results of seven iterations. (b) Histogram of the measured peak current

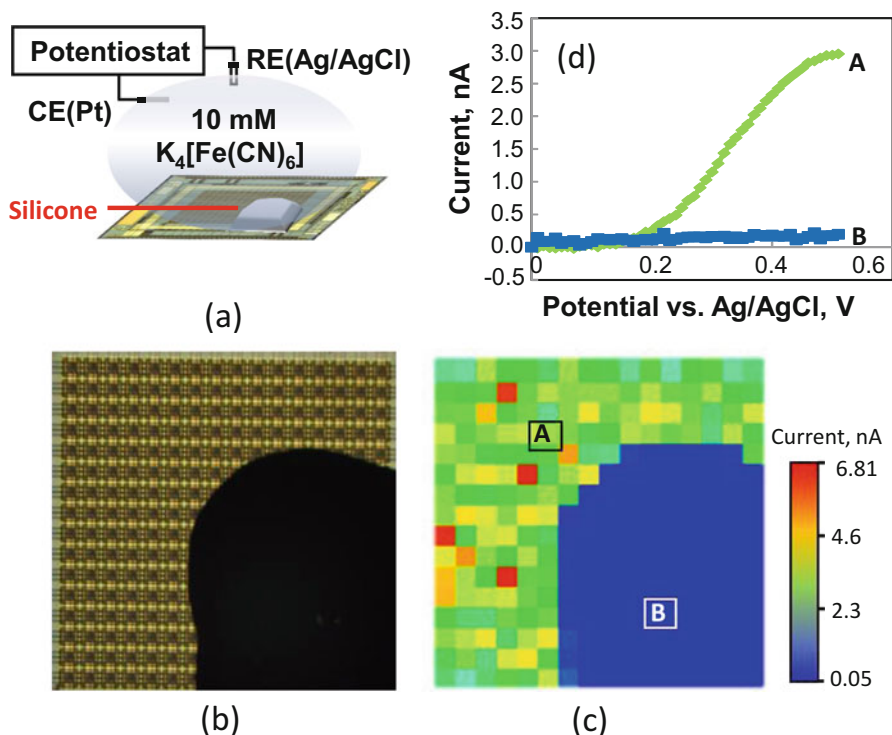


Fig. 4.17 Experiment to confirm the principle of 2D counting: (a) setup, (b) photomicrograph of chip covered partially by silicone paste, (c) sensor output, (d) measured CV at electrodes A and B not covered and covered by silicone, respectively

We confirmed the principle of the 2D imaging by using the plated electrode array. Figure 4.17a illustrates the experimental setup to confirm the principle. Silicone coated a part of the electrode array, as shown in Fig. 4.17b. The measured currents are shown in Fig. 4.17c. Figure 4.17d shows that the difference of the electrode with and without silicone has clearly come out of the current curve because the electrode coated by silicone does not flow current.

To verify the effectiveness of the proposed platform, direct counting of HeLa cells was demonstrated by using the 32×32 sensor chip. $40 \mu\text{l}$ of HeLa cells (8.1×10^5 cells/ml) solution in PBS was dripped and incubated for 2 h at room temperature to allow the cells to settle on the microelectrodes. Measurements were taken after drawing out the unnecessary solution (PBS with redundant cells that did not settle on the microelectrodes). Because the size of one HeLa cell is about $10 \mu\text{m}$, it can cover the entire microelectrode. When a cell is attached to the microelectrode, the current is blocked by the cell, reducing the amount of current. By detecting this change, we tried to determine whether a cell is on the microelectrode or not.

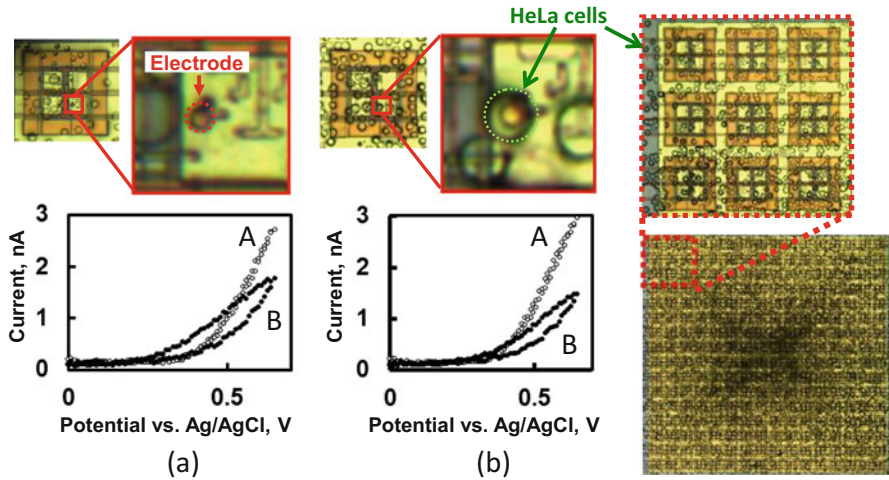


Fig. 4.18 Measured current as a function of counter electrode potential (a) without and (b) with HeLa cells

Figure 4.18 demonstrates the photomicrograph of the CMOS sensor surface with well-distributed HeLa cells. Although HeLa cells were gathered at the center of the chip, it does not affect the measurement. Figure 4.18a, b shows the measured current as a function of the reference electrode's potential with and without HeLa cells. In Fig. 4.18a, the peak current without HeLa cells decreases by 35%, because the PBS utilized for cells to settle on the microelectrode reduced the concentration of $K_4[Fe(CN)_6]$. On the other hand, in Fig. 4.18b, the peak currents with HeLa cells decrease by 50%. The difference in decrement (35% vs. 50%) arose from the presence of HeLa cells. Thus, the measured results indicate that the proposed sensor platform can detect HeLa cells directly.

To evaluate the feasibility of direct bacteria counting, bacteria-sized microbeads were counted using $1.2 \times 2.05 \mu\text{m}^2$ 4×4 microelectrode array. To imitate the bacteria, $1 \mu\text{m}$ -diameter magnetic microbeads were used. By attaching a neodymium magnet to the bottom of the chip, magnetic microbeads are immobilized at the surface of the chip. Figure 4.19 shows the measured results. The CV's peak current was reduced from 2.54 nA to 1.83 nA. The difference of 0.71 nA is sufficient to maintain signal-to-noise ratio under drift (less than 0.1 nA) and the input-referred noise (21.8 pA). Furthermore, a large-scale microelectrode array (1024×1024) chip was fabricated, and the basic operation was confirmed [16]. The chip operates from a 5 V supply at 1.9 mA.

Fig. 4.19 Measured CV of $1.2 \times 2.05 \mu\text{m}^2$ 4×4 microelectrode array without and with $1 \mu\text{m}$ microbeads

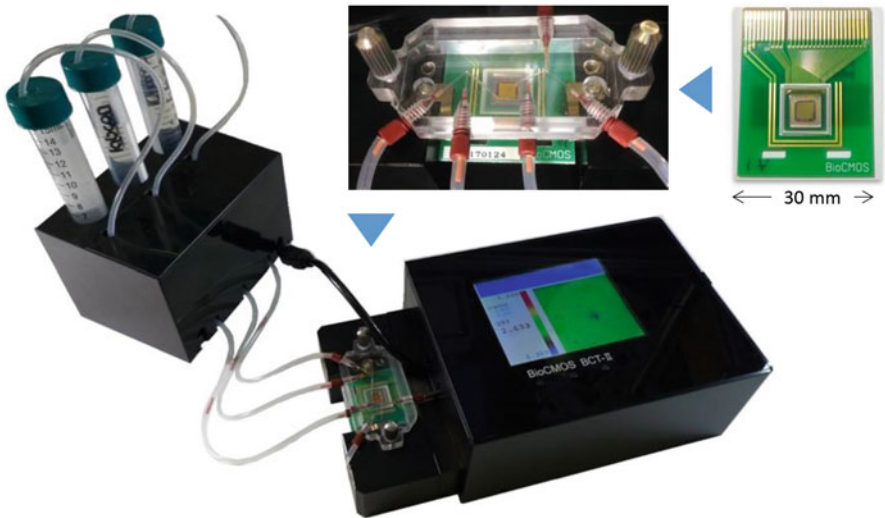
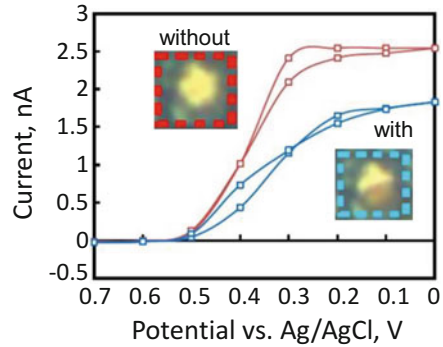


Fig. 4.20 Stand-alone portable diagnostic inspection equipment. Chip is bonded on a PCB and covered by holder with fluidic connection

4 Stand-Alone Portable Diagnostic Inspection System

A stand-alone portable diagnostic inspection system was constructed with a size of 18 cm by 10 cm by 5.5 cm and a weight of 850 g (Fig. 4.20). The equipment contains three printed circuit boards (PCBs). One is a microcomputer board with 32b MCU (microcontroller unit), a high-speed USB communication unit, and a USB file system unit. One is an interface board to communicate chip with 0.2 Hz–37.5 MHz DSS (direct digital synthesizer), 16b ADC, and 4 channel 12b DAC. The other one is a display board with QVGA color LCD (liquid crystal display), Wi-Fi module, and pump control units. Power is supplied by USB adpoter. The total power consumption is 5 V 130 mA when LCD is off and 210 mA when LCD is on.

This equipment is used in three configurations, stand-alone, PC control through USB, and smartphone control through Wi-Fi.

Chip power voltage is 4.5 V through voltage regulator to isolate external equipment. Potentiometric, amperometric, impedimetric, and photometric sensor array chip is bonded on a PCB; two silicone rubber frames are attached, one on the chip and one on the PCB; and silicone paste is filled between them to protect bonding wires. Several biomedical sensor systems can be constructed based on this equipment, including glucose sensors, DNA sequencers, bacteria or viruses counting, ion chromatography using capacitance-coupled contactless conductivity detection and UV-VIS absorption spectroscopy, and on-chip electrophoresis.

5 Future Prospect

For biomedical applications, it is important to detect the small amount of biomolecules. The present redox potential sensor detects 1 μM molecules in a $50 \times 50 \times 50 \mu\text{m}^3$ well, which corresponds to 125 amol/assay. The present bioluminescent assay achieves 0.01 amol/assay. To achieve such sensitivity, some biochemical amplification methods must be employed.

Single-molecule detections have been reported by substrate recycling [33] and redox recycling [34]. The detection during the amplification of biomolecules is also important such as qPCR (quantitative polymerase chain reaction) [35, 36], qLAMP (quantitative loop-mediated isothermal amplification) [36], and qNASBA (quantitative nucleic acid sequence-based amplification). To realize such amplification, the next step is the integrations of microfluidics and controlling scheme of biomolecular interactions on CMOS integrated circuits. One of the candidates is droplet digital amplification and detection.

6 Conclusions

CMOS multimodal biosensor arrays and their applications were described using standard CMOS processes. Silicon semiconductor technology is highly sophisticated and conservative. It is very hard to introduce new materials and new fabrication process if not applied to huge amount of products, such as memory, otherwise, it is difficult to maintain the fabrication line. The applications of biosensors are very wide and cannot be focused on one application. We must design circuits under the constriction of the standard CMOS process. However, there is a lot of room in circuit-based design.

In the early stage, electrochemical CMOS sensors were designed by the combination of operational amplifiers (op amps). Op amp is an almighty analog circuit. However, it is not ideal with respect to size and power consumption. Furthermore, bio-signal is quite different from signal treated in conventional information and

technology. Biomedical interaction occurs in ms time scale, whereas information technology treats less than 1 ns signal. In biosensors, low power, low noise, and high accuracy are more important than high speed. New circuit design technology must be developed specific to CMOS biosensors.

Acknowledgments This research is financially supported by a Grant-in-Aid for Scientific Research (No. 25220906, 20226009) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by Adaptable and Seamless Technology Transfer Program through Target-Driven R&D (No. AS272S001b) from the Japan Science and Technology Agency, Japan.

References

1. K. Nakazato, M. Ohura, S. Uno, CMOS cascode source-drain follower for monolithically integrated biosensor Array. *IEICE Trans. Electron.* **E91-C(9)**, 1505–1515 (2008). <https://doi.org/10.1093/iete/e91-c.9.1505>
2. J. Hasegawa, S. Uno, K. Nakazato, Amperometric electrochemical sensor Array for on-Chip simultaneous imaging: circuit and microelectrode design considerations. *Jpn. J. Appl. Phys.* **50**, 04DL03 (2011). <https://doi.org/10.1143/JJAP.50.04DL03>
3. M. Bennati, F. Thei, M. Rossi, M. Crescentini, G. D'Avino, A. Baschiroto, M. Tartagni, A sub-pA $\Delta\Sigma$ current amplifier for single-molecule nanosensors. *ISSCC Dig. Tech. Papers*, 348–349 (2009). <https://doi.org/10.1109/ISSCC.2009.4977451>
4. M.H. Nazari, H. Mazhab-Jafari, L. Leng, A. Guenther, R. Genov, CMOS neurotransmitter microarray: 96-channel integrated potentiostat with on-die microsensors. *IEEE Trans. Biomed. Circuits Syst.* **7(3)**, 338–348 (2013). <https://doi.org/10.1109/TBCAS.2012.2203597>
5. J. Rothe, O. Frey, A. Stettler, Y. Chen, A. Hierlemann, Fully integrated CMOS microsystems for electrochemical measurements on 32×32 working electrodes at 90 frames per second. *Anal. Chem.* **86**, 6425–6432 (2014). <https://doi.org/10.1021/ac500862v>
6. M. Yang, S.C. Liu, T. Delbruck, A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. *IEEE J. Solid State Circuits* **50(9)**, 2149–2160 (2015). <https://doi.org/10.1109/JSSC.2015.2425886>
7. M. Takihi, K. Niitsu, K. Nakazato, Charge-conserved analog-to-time converter for a large-scale CMOS biosensor array. *IEEE Int. Symp. Circuits Syst. (ISCAS 2014)*, 33–36 (2014). <https://doi.org/10.1109/ISCAS.2014.6865058>
8. K. Ikeda, A. Kobayashi, K. Nakazato, K. Niitsu, Design and analysis of scalability in current-mode analog-to-time converter for an energy-efficient and high-resolution CMOS biosensor array. *IEICE Trans. Electron* (2017) (in press)
9. P. Georgiou, C. Toumazou, An adaptive ISFET chemical imager chip. *IEEE Int. Symp. Circuits Syst. (ISCAS 2008)*, 2078–2081 (2008). <https://doi.org/10.1109/ISCAS.2008.4541858>
10. P. Bergveld, Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Trans. Biomed. Eng.* **BME-17**, 70–71 (1970)
11. K. Sawada, S. Mimura, K. Tomita, T. Nakanishi, H. Tanabe, M. Ishida, T. Ando, Novel CCD-based pH imaging sensor. *IEEE Trans. Electron Devices* **46(9)**, 1846–1849 (1999)
12. D.M. Garner, H. Bai, P. Georgiou, T.G. Constantinou, S. Reed, L.M. Shepherd, W. Wong Jr., K.T. Lim, C. Toumazou, A multichannel DNA SoC for rapid point-of-care gene detection. *ISSCC Dig. Tech. Papers*, 492–493 (2010). <https://doi.org/10.1109/ISSCC.2010.5433834>
13. T. Sakata, Y. Miyahara, Direct transduction of allele-specific primer extension into electrical signal using genetic field effect transistor. *Biosens. Bioelectron.* **22**, 1311–1316 (2007). <https://doi.org/10.1016/j.bios.2006.05.031>
14. J.M. Rothberg et al., An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011). <https://doi.org/10.1038/nature10242>

15. M. Kamahori, K. Harada, H. Kambara, A new single nucleotide polymorphisms typing method and device by bioluminometric assay coupled with a photodiode array. *Meas. Sci. Technol.* **13**, 1779–1785 (2002)
16. K. Gamo, K. Nakazato, K. Niitsu, A Current-integration-based CMOS amperometric sensor with 1024×1024 bacteria-sized microelectrode array for high-sensitivity bacteria counting. *IEICE Trans. Electron* (2017) (in press)
17. K. Niitsu, K. Ikeda, K. Muto, K. Nakazato, Design, experimental verification, and analysis of a 1.8-V-input-range voltage-to-current converter using source degeneration for low-noise multimodal CMOS biosensor array. *Jpn. J. Appl. Phys* **56**, 01AH06 (2017). <https://doi.org/10.7567/JJAP.56.01AH06>
18. T. Kuno, K. Niitsu, K. Nakazato, Amperometric electrochemical sensor array for on-chip simultaneous imaging. *Jpn. J. Appl. Phys* **53**, 04EL01 (2014). <https://doi.org/10.7567/JJAP.53.04EL01>
19. A. Manickam, A. Chevalier, M. McDermott, A.D. Ellington, A. Hassibi, A CMOS electrochemical impedance spectroscopy (EIS) biosensor array. *IEEE Trans. Biomed. Circuits Syst.* **4**(6), 379–390 (2010). <https://doi.org/10.1109/TBCAS.2010.2081669>
20. S. Hwang, C.N. LaFratta, V. Agarwal, X.J. Yu, D.R. Walt, S. Sonkusale, CMOS microelectrode array for electrochemical lab-on-a-chip applications. *IEEE Sensors J.* **9**(6), 609–615 (2009). <https://doi.org/10.1109/JSEN.2009.2020193>
21. M. Datta, S.A. Merritt, M. Dagenais, Electroless remetalization of aluminum bond pads on CMOS driver chip for flip-chip attachment to vertical cavity emitting lasers (VCSEL's). *IEEE Trans. Compon. Pack. Technol.* **22**(2), 299–306 (1999)
22. Kanigen Technical Report No. 9. <http://www.kanigen.co.jp/file/report9.pdf> (in Japanese)
23. M. Kamahori, Y. Ishige, M. Shimoda, Enzyme immunoassay using a reusable extended-gate field-effect-transistor sensor with a ferrocenylalkanethiol-modified gold electrode. *Anal. Sci.* **24**, 1073–1079 (2008)
24. Y. Ishige, M. Shimoda, M. Kamahori, Extended-gate FET-based enzyme sensor with ferrocenyl-alkanethiol modified gold sensing electrode. *Biosens. Bioelectron.* **24**, 1096–1102 (2009). <https://doi.org/10.1016/j.bios.2008.06.012>
25. H. Anan, M. Kamahori, Y. Ishige, K. Nakazato, Redox-potential sensor array based on extended-gate field-effect transistors with ω -ferrocenylalkanethiol-modified gold electrodes. *Sensors Actuators B Chem.* **187**, 254–261 (2013). <https://doi.org/10.1016/j.snb.2012.11.016>
26. W. Guan, X. Duan, M.A. Reed, Highly specific and sensitive non-enzymatic determination of uric acid in serum and urine by extended gate field effect transistor sensors. *Biosens. Bioelectron.* **51**, 225–231 (2014). <https://doi.org/10.1016/j.bios.2013.07.061>
27. M.J. Milgrew, D.R.S. Cumming, Matching the transconductance characteristics of CMOS ISFET arrays by removing trapped charge. *IEEE Trans. Electron Devices* **55**(4), 1074–1079 (2008). <https://doi.org/10.1109/TED.2008.916680>
28. H. Komori, K. Niitsu, J. Tanaka, Y. Ishige, M. Kamahori, K. Nakazato, An extended-gate CMOS sensor Array with enzyme-immobilized microbeads for redox-potential glucose detection. *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS 2014)*, 464–467 (2014). <https://doi.org/10.1109/BioCAS.2014.6981763>
29. H. Ishihara, K. Niitsu, K. Nakazato, Analysis and experimental verification of DNA single-base polymerization detection using CMOS FET-based redox potential sensor array. *Jpn. J. Appl. Phys.* **54**, 04DL05 (2015). <https://doi.org/10.7567/JJAP.54.04DL05>
30. H. Tanaka, P. Fiorini, S. Peeters, B. Majeed, T. Sterken, M. Op de Beeck, M. Hayashi, H. Yaku, I. Yamashita, Sub-micro-liter electrochemical single-nucleotide-polymorphism detector for lab-on-a-chip system. *Jpn. J. Appl. Phys.* **51**, 04DL02 (2012). <https://doi.org/10.1143/JJAP.51.04DL02>
31. Y. Ishige, Y. Goto, I. Yanagi, T. Ishida, N. Itabashi, M. Kamahori, Feasibility study on direct counting of viruses and bacteria by using microelectrode array. *Electroanalysis* **24**(1), 131–139 (2012). <https://doi.org/10.1002/elan.201100482>

32. K. Niitsu, S. Ota, K. Gamo, H. Kondo, M. Hori, K. Nakazato, Development of microelectrode arrays using electroless plating for CMOS-based direct counting of bacterial and HeLa cells. *IEEE Trans. Biomed. Circuits Syst.* **9**(5), 607–619 (2015). <https://doi.org/10.1109/TBCAS.2015.2479656>
33. T. Satoh, J. Kato, N. Takiguchi, H. Ohtake, A. Kuroda, ATP amplification for ultrasensitive bioluminescence assay: detection of a single bacterial cell. *Biosci. Biotechnol. Biochem.* **68**(6), 1216–1220 (2004). <https://doi.org/10.1271/bbb.68.1216>
34. M.A.G. Zevenbergen, P.S. Singh, E.D. Goluch, B.L. Wolfrum, S.G. Lemay, Stochastic sensing of single molecules in a nanofluidic electrochemical device. *Nano Lett.* **11**, 2881–2886 (2011). <https://doi.org/10.1021/nl2013423>
35. C. Toumazou et al., Simultaneous DNA amplification and detection using a pH-sensing semiconductor system. *Nat. Methods* **10**(7), 641–646 (2013). <https://doi.org/10.1038/NMETH.2520>
36. H. Norian, R.M. Field, I. Kymissis, K.L. Shepard, An integrated CMOS quantitative-polymerase-chain-reaction lab-on-chip for point-of-care diagnostics. *Lab Chip* **14**, 4076–4084 (2014). <https://doi.org/10.1039/c4lc00443d>

Chapter 5

Micro-NMR on CMOS for Biomolecular Sensing

Ka-Meng Lei, Nan Sun, Pui-In Mak, Rui Paulo Martins, and Donhee Ham

1 Introduction of NMR and NMR-Based Biosensing

Nuclear magnetic resonance (NMR) is the resonant energy exchange between RF magnetic fields and nuclear spins subjected to static magnetic fields. It has offered a powerful tool to examine the properties of atomic nuclei, to sense molecules, to determine molecular structures ranging from small organic molecules to proteins, and to image biological tissues [1–6], making profound impact in a broad range of areas in physical and life sciences. Traditional NMR instruments are bulky, heavy, and expensive. Their miniaturization can greatly expand the scope of NMR applications. For example, small, low-cost NMR biomolecular sensors would be relevant to medical diagnostics in the context of personalized medicine. Over the past several years, by combining custom-designed semiconductor integrated circuits and small permanent magnets, we have developed several miniature NMR

K.-M. Lei • P.-I. Mak (✉)

State Key Laboratory of Analog and Mixed-Signal VLSI and FST-ECE, University of Macau, Macau, China

e-mail: pimak@umac.mo

N. Sun (✉)

Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, 78756, USA

e-mail: nansun@mail.utexas.edu

R.P. Martins

State Key Laboratory of Analog and Mixed-Signal VLSI and FST-ECE, University of Macau, Macau, China

Instituto Superior Técnico, Universidade de Lisboa, 1649-004, Lisboa, Portugal

D. Ham (✉)

School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, 02138, USA

e-mail: donhee@seas.harvard.edu

relaxometry systems, aimed in particular at biomolecular sensing applications. We originally reported them in [7–15]. The present chapter reviews these developments. A large amount of material in this chapter are borrowed from our papers cited above. The goal of this chapter is to coherently synthesize and integrate the series of the developments.

Magnet, coil, and RF transceiver (TRX) are three key components of NMR system. The magnet produces static magnetic fields that polarize nuclear spins. The coil and RF TRX are used to generate excitation RF magnetic fields and also to monitor the dynamics of nuclear spins. By far the largest part is the magnet, which is typically implemented as a superconducting magnet, which can produce a strong static magnetic field on the order of tens of Tesla. Such strong static field, which leads to a stronger NMR signal and a finer NMR spectral resolution, is necessary in advanced NMR spectroscopy, especially in applications to determine the structures of biological macromolecules. At the same time, the superconducting magnet is large, leading to the traditionally bulky NMR instrumentation.

In creating our miniaturized NMR systems, we used many orders of magnitude smaller-sized permanent magnets with much weaker static magnetic fields (<1 T). This is fundamentally possible because these systems are aimed to perform NMR relaxometry (as opposed to NMR spectroscopy), in which the demands for the NMR signal strength and spectral resolution are greatly relaxed. To detect the NMR signal, significantly weakened due to the use of the small-sized permanent magnets with the weak static fields, we developed highly sensitive RF TRXs with the complementary metal-oxide-semiconductor (CMOS) technology. The CMOS RF TRX integrated circuits (IC) replace the traditional discrete RF TRXs, further contributing to the system miniaturization.

Although NMR is a too well-established subject [2, 6], we here review its basics and its application to biomolecular sensing to the extent germane to our works, in order to make this chapter self-contained. NMR experiments can be done with various atomic nuclei. We will consider protons, the nuclei of hydrogen atoms (^1H), a most common subject of NMR experiments.

Due to spin, the proton has an intrinsic magnetic moment $\vec{m} = \gamma \vec{S}$ where $\gamma \sim 42.6$ MHz/T is the gyromagnetic ratio of proton and \vec{S} is the spin angular momentum operator. The single-proton Hamiltonian in an external static magnetic field $\vec{B}_0 = B_0 \hat{z}$ (\hat{z} : unit vector along z-axis) is then given by $H = -\vec{m} \cdot \vec{B}_0 = -\gamma \vec{S} \cdot B_0 \hat{z} = -\gamma B_0 S_z = -\omega_0 S_z$, where $\omega_0 \equiv \gamma B_0$ is the Larmor frequency. S_z has two eigenstates: the “spin-up” state with eigenvalue $\hbar/2$ and the “spin-down” state with eigenvalue $-\hbar/2$, where $\hbar \equiv h/(2\pi)$ (h : Planck constant). Therefore, the spin-up and spin-down states are the eigenstates of the Hamiltonian as well, with energies $-\hbar\omega_0/2$ and $\hbar\omega_0/2$, respectively.

A large number of ^1H proton magnetic moments in a macroscopic sample, such as water, in $B_0 \hat{z}$ and at ambient temperature T will reach thermal equilibrium, where the proton population ratio between the spin-up and spin-down states will be $e^{\hbar\omega_0/(2kT)} / e^{-\hbar\omega_0/(2kT)}$ (k : Boltzmann constant). Since $\hbar\omega_0 \ll kT$ at room temperature and with typical B_0 values, this ratio is only very slightly larger than 1. Nonetheless, with a typical macroscopic sample containing a large number of protons, the

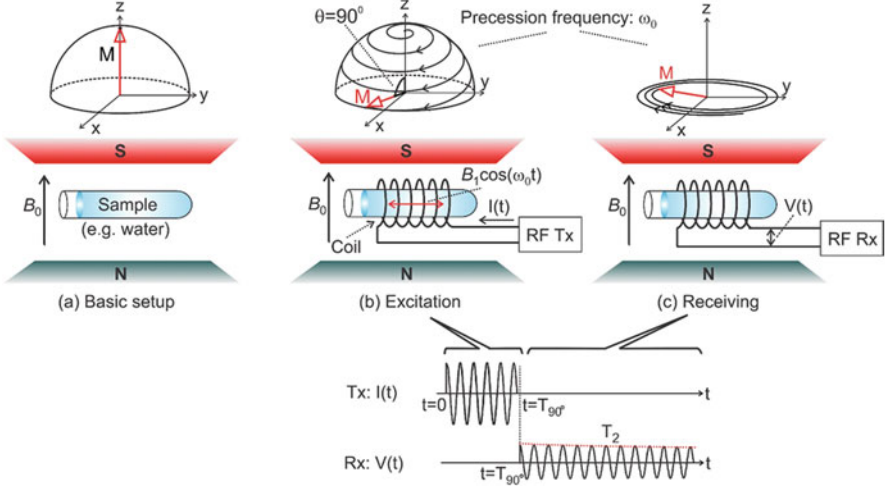


Fig. 5.1 Physics of NMR. (a) Equilibrium state of the magnetic moments with B_0 . (b) Excitation with a dynamic B_1 . (c) Relaxation to equilibrium with characteristic time T_2

absolute difference in populations is still considerable, producing an appreciable net macroscopic magnetic moment \vec{M} along the positive z-axis (Fig. 5.1a). Specifically, $\vec{M} = M_0 \hat{z}$ with:

$$M_0 = \gamma \sum_j S_{j,z} = N\gamma \frac{\hbar/2 \cdot e^{\hbar\omega_0/2kT} - \hbar/2 \cdot e^{-\hbar\omega_0/2kT}}{e^{\hbar\omega_0/2kT} + e^{-\hbar\omega_0/2kT}} \cong \frac{N\gamma^2 \hbar^2 B_0}{4kT}, \quad (5.1)$$

where N is the total number of protons and the summation runs over all protons. Thus, a stronger B_0 and a larger sample size (larger N) yield a stronger M_0 .

Once the sample attains the equilibrium macroscopic magnetic moment $M_0 \hat{z}$ in the static field $B_0 \hat{z}$, let a perturbing RF magnetic field of angular frequency ω and amplitude B_1 be applied to the sample perpendicularly to $B_0 \hat{z}$. Without loss of generality, x-axis may be assigned to the direction of the RF field, which then can be written as $\vec{B}_1(t) = \hat{x} B_1 \cos(\omega t)$. This RF field can be generated by wrapping a coil around the sample and transmitting an RF current into the coil (Fig. 5.1b). The resulting time evolution of the macroscopic magnetic moment $\vec{M}(t)$ can be described by the following equation of motion:

$$\frac{d\vec{M}(t)}{dt} = \gamma \vec{M}(t) \times [B_0 \hat{z} + B_1 \hat{x} \cos(\omega t)], \quad (5.2)$$

which states that the torque $\vec{M}(t) \times [B_0 \hat{z} + \vec{B}_1(t)]$ exerted on $\vec{M}(t)$ is equal to the time derivative of the macroscopic angular momentum $\vec{M}(t)/\gamma$.

$\vec{M}(t)$ remains to be $M_0\hat{z}$ if the excitation frequency ω is not tuned closely into ω_0 . If ω closes onto ω_0 , $\vec{M}(t)$ absorbs energy from the RF magnetic field and uses it to increase its potential energy by increasing the angle θ it makes with the static field $B_0\hat{z}$ (Fig. 5.1b). If the RF perturbation at $\omega = \omega_0$ continues, θ keeps increasing to the maximum, π , with $\vec{M}(t)$ attaining the maximum potential energy, then, θ decreases back to 0 with $\vec{M}(t)$ releasing its potential energy back to the RF field, and then, θ grows again, so on and so forth. This resonant energy exchange between the RF field and $\vec{M}(t)$ is NMR. The oscillation of θ occurs at frequency $\omega_1 = \gamma B_1/2$ (Rabi oscillation).

In addition to the Rabi oscillation that periodically modulates the z -component of $\vec{M}(t)$ at frequency ω_1 , $\vec{M}(t)$ also undergoes precession about the z -axis due to the torque of the static field $B_0\hat{z}$ exerts; thus its x and y components are also periodically modulated. This precession motion, which does not alter the potential energy with respect to the static $B_0\hat{z}$ field, occurs at ω_0 . Overall, $\vec{M}(t)$ will exhibit a spiral downward (or upward) motion (Fig. 5.1b). The downward/upward motion of $\vec{M}(t)$ (change of θ) is much slower than the spiral motion (precession), as $B_1 \ll B_0$ thus $\omega_1 \ll \omega_0$ virtually always holds true.

After a certain duration of the RF excitation ($\omega = \omega_0$), which tips \vec{M} away from the z -axis to a certain angle $\theta = \theta_0$, the coil is switched from the RF transmitter to the RF receiver (Fig. 5.1c). This switchover terminates the resonant energy exchange process, and θ is now kept at θ_0 for a while (\vec{M} will eventually relax back to the equilibrium position of $\theta = 0$, but this energy relaxation is a leisurely process taking several seconds, which is far slower than the faster dynamics considered here), but the ω_0 precession persists. The consequent period variation of the magnetic flux across the coil induces a sinusoidal voltage $V(t)$ with frequency ω_0 across the coil, which registers at the RF receiver. We call $V(t)$ NMR signal (although NMR actually occurs during the excitation mode). Typically, to maximize the amplitude of $V(t)$ in the reception mode, the excitation mode is terminated when $\theta_0 = 90^\circ$ with \vec{M} on the xy plane; as the Rabi oscillation frequency is $\omega_1 = \gamma B_1/2$, the duration of the excitation mode to change θ from 0° to 90° is given by

$$T_{90^\circ} = \frac{\pi}{\gamma B_1}. \quad (5.3)$$

$V(t)$ decays exponentially with a characteristic time called T_2 (Fig. 5.1c). This damping is not due to, and occurs faster than, the energy relaxation parenthetically mentioned shortly before. It is caused by random interactions among the proton spins (spin-spin interactions), which perturb the ω_0 precession of each proton spin, causing its phase to undergo a random walk process [16–18]. Consequently, the precessions of a large number of protons grow out of phase with time, rendering

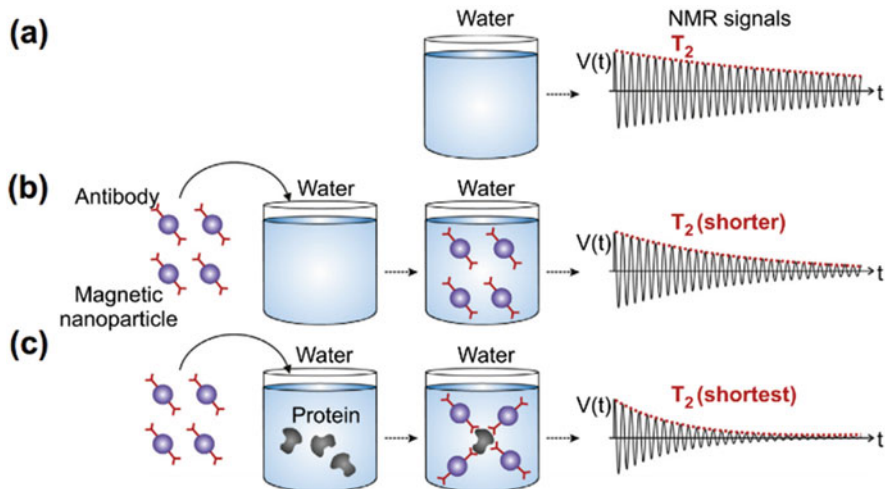


Fig. 5.2 NMR-based biomolecular detection. (a) Without MNP and target. (b) With only probe-decorated MNP. (c) With target and probe-decorated MNP link together

the vector sum \vec{M} to exponentially decay, leading to the damped $V(t)$. $T_2 \sim 1$ s for pure water at $T = 300$ K. This damping is typically very slow as compared to the spin precession; for example, for $\omega_0 \approx 21$ MHz ($B_0 \approx 0.5$ T), 10^7 precession cycles occur before $V(t)$ decays appreciably.

To detect particular proteins in a sample, magnetic nanoparticles (MNPs) coated with antibodies that can specifically bind to the target proteins are introduced into the sample (Fig. 5.2) [5]. Consider a biological sample containing a large number of water molecules (thus a large number of ^1H protons). In the absence of the target proteins (Fig. 5.2b), the MNPs stay monodispersed. These MNPs incessantly move around due to Brownian motion, producing fluctuating magnetic fields. These disturb precessions of the proton spins, increasing their phase noise beyond that due to the basic spin-spin interactions. Therefore, the phase coherence is lost at a higher rate, reducing T_2 . In the presence of target (Fig. 5.2c), the MNPs assemble into local clusters, which are even more efficient in accelerating the dephasing of ^1H proton precessions, yielding an even smaller T_2 . In summary, by monitoring T_2 of ^1H NMR signal, target proteins can be detected [5].

2 Miniature NMR System with CMOS IC

Before we present the detailed design of the miniaturized NMR systems, let us first discuss a few general system-level considerations.

2.1 Magnet Miniaturization

The signal-to-noise ratio (SNR) of NMR is proportional to B_0^2 [19]. Thus, from the SNR point of view, it is desirable to increase the magnetic field B_0 as much as possible. Nowadays, magnetic fields as strong as 20 T have been used. In order to produce such strong magnetic field, the only viable option is to use a superconducting electromagnet. Yet, they are big, expensive, power hungry, and have high maintenance costs. Thus, it is not suitable for a small low-cost NMR system. To significantly reduce magnet size and cost, the only feasible solution is to use a small permanent magnet. There are several challenges associated with this. First, a permanent magnet produces a relatively weak magnetic field B_0 . For a typical permanent magnet, B_0 is usually ~ 1 T or below, which leads to a lower SNR. Furthermore, the other major challenge for using a small magnet is its B_0 inhomogeneity. For an inhomogeneous B_0 , nuclear spins at different locations in the sample precess at different Larmor frequencies. As a result, the spectral lines of the NMR signal are smeared, making it impossible to decipher the fine frequency shifts needed for high-resolution NMR spectroscopy applications. To solve this problem, a brute-force method is used before to increase the dimension of the permanent magnet. For instance, large permanent magnets that occupy 1 m^3 and weigh several hundreds of kilograms have been used; however, this method completely defeats the purpose of miniaturization.

Several shimming mechanisms including electrical shimming coils [20] and arranging with additional smaller shimming magnets [21–23] have been reported. These magnetic field shimming techniques are necessary for high-resolution NMR spectroscopy, but it may not be needed for NMR relaxometry applications, whose parameters of interests are relaxation times T_1 and T_2 . Instead, the relaxation times can be accurately measured in the presence of magnetic field inhomogeneity by using NMR spin-echo techniques, such as the CPMG pulse sequence [24, 25]. This greatly reduces the burden on the magnet design. In our NMR miniaturization works, since our intended application is biomolecular sensing whose key information is T_2 , we chose not to shim the magnet in order to reduce the magnet design complexity and, instead, rely on the use of spin-echo techniques to address the magnetic field inhomogeneity problem.

Although there are many challenges with using a small permanent magnet, it does bring the key benefits of being lightweight with small dimensions and low cost. Another advantage of using a permanent magnet is that it allows the use of solenoidal coils whose coil sensitivity B_1/I_1 (I_1 is the induced current on the coil by B_1) is three times higher than the saddle-shaped coils that are usually adopted for electromagnets [19].

2.2 Coil Miniaturization

The size of the coil does not pose a challenge for NMR system miniaturization, as it is typically much smaller than the magnet and the TRX. The major motivation to develop a small coil is to increase the coil sensitivity B_1/I_1 , as it directly affects SNR. For a solenoidal coil, it is easy to derive that its B_1/I_1 is inversely proportional to the coil length and diameter for the fixed number of turns. Thus, by shrinking the coil dimension, we can increase B_1/I_1 and the SNR. Another benefit is that an increased B_1/I_1 means that for the same B_1 , the required coil current I_1 during NMR excitation is smaller, which lowers the requirement on the transmitter power.

So far, most research works have focused on solenoidal microcoils, which offer higher B_1/I_1 [26–31]. Nevertheless, as they are hand-wound, they cannot be batch fabricated and encounter severe fabrication difficulties at small dimensions (<1 mm) [32, 33]. Planar microcoils can be better candidates than solenoidal ones. Although the sensitivity of planar microcoils is lower, they have the following key advantages due to their standard photolithography-based fabrication process: (1) they can be easily batch fabricated into a large array with high (μm) resolution [4, 32–37]. (2) Their planar structure allows them to be integrated with a microfluidic system [34–37]. As a result, the sample introduction can be greatly simplified and the sensitivity can also be improved [32–34]. (3) Because their fabrication process is compatible with the silicon IC process, it is possible to integrate the planar microcoil array with a multichannel TRX on the same IC chip [32, 33, 38, 39]. In this way, the wiring between the microcoil array and the TRX becomes very simple.

2.3 Transceiver Miniaturization

Once the magnet size is reduced, the next biggest component to miniaturize is the TRX. Existing commercial NMR TRXs are built using discrete electronics, leading to their bulky and expensive setup. Moreover, their performance is not optimal, especially in terms of sensitivity and power consumption. The conventional NMR TRX, due to its bulky size, has to be placed far from the coil and connects to the coil through a long transmission line [40]. The use of this long wire not only causes signal loss but also requires the TRX to be matched to $50\ \Omega$ in order to avoid impedance mismatch. However, $50\ \Omega$ -matching provides neither efficient power delivery for the power amplifier (PA) nor noise matching for the receiver.

With the fast progress of RF IC technology over the past two decades, it is now feasible to integrate the entire RF TRX onto a small millimeter-sized silicon chip. The TRX integration brings many advantages. It not only reduces the TRX

size and cost but also increases SNR and reduces power consumption. With the significant size reduction, the TRX IC can be placed in close proximity to the coil to completely obviate the need for the transmission line, which allows any desirable impedance to be used. Thus, during the RF transmission, the coil can be matched to a low impedance for efficient power delivery. In the receiving mode, receiver noise matching can also be achieved by having a different matching network to match the coil to a high impedance. This high impedance transformation can passively amplify the NMR signal, leading to a substantially reduced receiver noise figure (details are provided in Sect. 5.3). For example, our research shows that a good noise figure of less than 1 dB can be achieved by using a low-power integrated TRX. Compared to the discrete single-ended TRX, the integrated TRX can also be made fully differential, leading to reduced sensitivity to common-mode noise and perturbations from power supply, bias circuit, and silicon substrate. Also, the integration can greatly reduce parasitic inductances and capacitances that cause performance degradation.

In the conventional discrete NMR transmitter, the PA usually operates under a large power supply (100 V) in order to deliver a large power of 100 W or beyond. Such a high-power PA is undesirable as it is big and expensive. It is hard to integrate on chip because standard silicon CMOS technology cannot operate under such high voltage, as it can cause transistor gate breakdown. Fortunately, the increase in the microcoil sensitivity B_1/I_1 helps to lower the power. For the same B_1 , a 10 times increase in B_1/I_1 means a 10 times reduction in I_1 and a 100 times reduction in the transmitter power. If the sample has a relatively long relaxation time, we can also decrease B_1 , which further reduces the power. For example, for the small NMR systems that we developed, since they are used to analyze fluidic samples that have long T_2 time (>10 ms), we can use relatively long excitation pulses (40 μ s). Under this scenario, the required PA power can be as small as 1 mW. This makes it very easy to integrate the PA on chip with other TRX electronics. The detailed hardware design of two CMOS NMR TRXs will be revealed next.

3 Basics of Miniature CMOS NMR System

Figure 5.3 shows a miniature NMR system [12, 15], which weighs 2 kg, and employs a $B_0 \approx 0.5$ T, 1.25 kg magnet. The Larmor frequency $f_L \equiv \omega_0/(2\pi)$ is ~ 21.3 MHz. The field inhomogeneity $\Delta B_0/B_0$ at the center of the magnet is about 50 ppm over a 5 μ L volume. The 500 nH planar copper microcoil is in-house fabricated on a glass substrate [4], where the electroplated copper is ~ 15 μ m thick, with which $Q \approx 16$ at f_L . A 5 μ L sample is held on the microcoil with a 5 μ m-thick passivation layer in between, inside a microfluidic container fabricated on top [4].

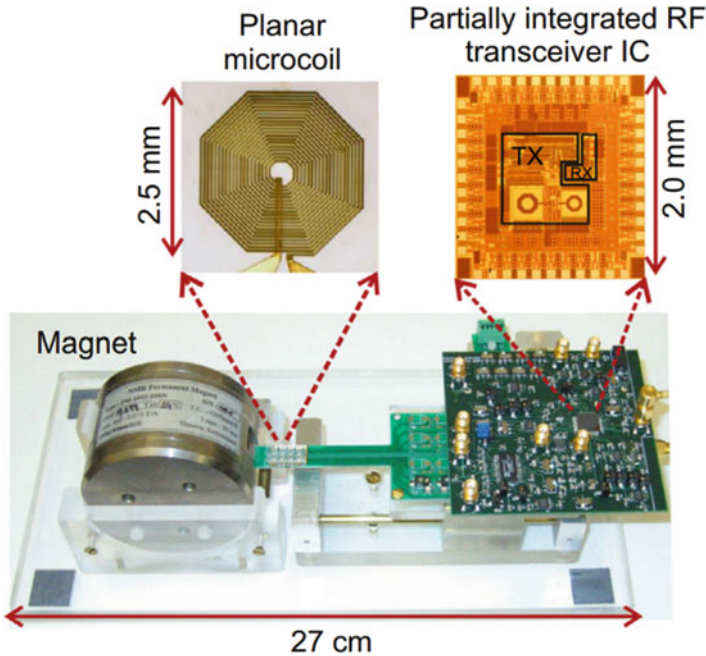


Fig. 5.3 The prototype of the first miniature NMR system with customized CMOS IC [12, 15]

3.1 Transceiver IC Architecture

Figure 5.4 shows the architecture of the RF TRX used in this NMR system. The receiver path is in the lower portion. A weak spin-echo signal whose maximum available power is 0.5 fW appears at the front-end node [“2” in Fig. 5.4]. In the frequency domain, the spin-echo signal is centered at the f_L with the bandwidth of $\gamma\Delta B_0/(2\pi) \approx 1.1$ kHz. The signal is amplified by a low-noise amplifier (LNA) and a variable-gain amplifier (VGA) and then is down-converted using mixers and quadrature local oscillator (LO) signals (“I” and “Q”) with frequency $\omega_0 + \delta$. We select $\delta/(2\pi) = 3$ kHz, which is high enough to prevent swamping by $1/f$ noise and is low enough to facilitate the rejection of the out-of-band noise with a bandpass filter with a moderate quality factor. The image noise is rejected by a digital-domain algorithm employing a Hilbert transformer.

The transmitter path is in the upper portion of Fig. 5.4. The excitation RF magnetic field is produced by the same quadrature LO signals used in the receiver. Their frequency, $\omega_0 + \delta$, deviates from ω_0 , but the transmitted 90° and 180° pulses are windowed sinusoids with bandwidths of 15 kHz and 7.5 kHz; thus, they can still excite the entire sample, across which the f_L has a variation $\gamma\Delta B_0/(2\pi) = 1.1$ kHz. By gating the quadrature oscillator signals with the digital pulse generator, we produce the CPMG pulse sequence. The PA is implemented off chip.

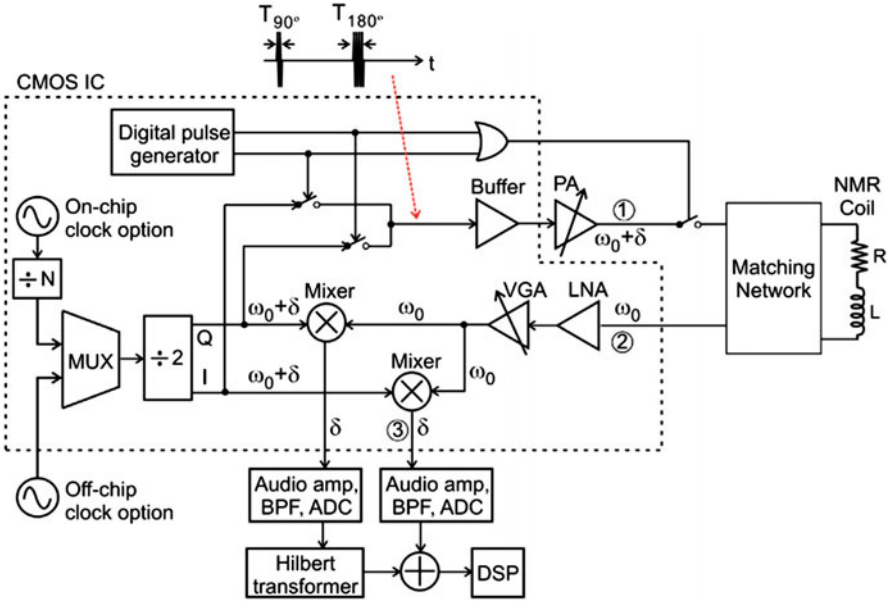


Fig. 5.4 CMOS RF transceiver architecture of the first NMR prototype

3.2 Receiver LNA and Noise Matching

The LNA is of common-source configuration. Since the f_L is smaller than the $1/f$ noise corner (~ 50 MHz) of NMOS transistors in the $0.18\ \mu\text{m}$ CMOS technology used, PMOS devices M_1 and M_2 with smaller $1/f$ noise corner (~ 1 MHz) are used as input transistors. These PMOS transistors are built in separate N-wells; thus, they also help isolate the LNA from substrate noise produced by the transmitter. The coupling of the LO signal into the LNA input is suppressed by cascading (transistors M_3, M_4). With active load M_5 and M_6 , the LNA achieves a high voltage gain of 41 dB. A common-mode feedback circuit (CMFB) ensures a correct output common mode in the LNA; the CMFB compares the output common-mode voltage V_{cmo} to $V_{\text{bias},3}$, and its output V_{cmfb} is used to drive the gates of transistors M_5 and M_6 . The Miller capacitor C_3 and resistor R_3 are for frequency compensation.

Channel thermal noise of transistors is the dominant noise source. The calculated input referred noise (IRN) of the LNA is:

$$-v_{n,LNA}^2 \Delta f \approx \frac{8kT\gamma_n}{g_{m1}} \left(1 + \frac{g_{m5}}{g_{m1}} \right), \quad (5.4)$$

where g_m denotes transistor transconductance and γ_n denotes transistor channel thermal noise coefficient. Here we have ignored the noise of transistors M_3, M_4 , and the CMFB circuit (and the input-referred current noise of the LNA is negligible

at f_L). To minimize the $1/g_{m1}$ factor, a large tail current (4 mA) and wide input transistors (900 μm) are used. To minimize the g_{m5}/g_{m1} term, transistors M_5 and M_6 are made much narrower than transistors M_1 and M_2 . Subsequently, M_5 and M_6 need a large V_{GS} , which is provided by stacking transistors M_7 and M_8 in the CMFB. In this way the IRN of the LNA is minimized. The measured IRN of the entire receiver is 1.8 nV/ $\sqrt{\text{Hz}}$, which is primarily contributed by the LNA.

Ultimately, we seek to minimize the receiver noise figure F , for which both the IRN minimization and the optimal LNA-coil matching play crucial roles. Consider a general passive network between the coil and the LNA, whose voltage transfer function from the coil side to the amplifier side assumes a value of α at f_L . Assuming the NMR signal $V(t)$'s rms value = V_{rms} , coil's thermal noise $\overline{v_{n,R}^2} = 4kTR\Delta f$ over the signal bandwidth Δf , and the LNA's input referred voltage noise $\overline{v_{n,LNA}^2}$ also over the signal bandwidth, and voltage gain α in the matching network, noise figure F of the LNA is given by:

$$F = \frac{SNR_{\text{in}}}{SNR_{\text{out}}} \cdot \frac{V_{\text{rms}}/\overline{(v_{n,R}^2)}^{1/2}}{\alpha V_{\text{rms}}/(\alpha^2 \overline{v_{n,R}^2} + \overline{v_{n,LNA}^2})^{1/2}} = \left(1 + \frac{\overline{v_{n,LNA}^2}/\Delta f}{\alpha^2 \cdot 4kTR}\right)^{1/2}, \quad (5.5)$$

where we have neglected the noise of the passive network, which we will shortly justify. From Eq. (5.5), it is clear that in addition to the small IRN of the LNA, a larger voltage gain α helps lowering F . To attain a large value of α , we use a single shunt capacitor C as the passive network, and the C is chosen in such a way that it resonates with the coil inductor at f_L . In this way, a large value of α is obtained, specifically: $\alpha = \sqrt{Q^2 + 1} \approx 16$. Note that this resonant noise matching leads to impedance mismatch between the LNA and the coil. At f_L , capacitors are far less lossy than the coil, which justifies the omission of the noise of the passive network in writing Eq. (5.5). The large value of α due to the resonance noise matching together with the minimized IRN of the LNA significantly reduces F , whose measured value is 0.7 dB.

3.3 Digital Pulse Generator

The CPMG pulse sequence is obtained by using one of the two quadrature LO signals for the 90° pulse and the other quadrature LO signal for the 180° pulses. The axes of rotation for \vec{M} in the rotating frame corresponding to the two quadrature signals are perpendicular; thus the CPMG sequence is obtained. The arrangement to gate the two quadrature LO signals in the transmitter to obtain the CPMG pulse sequence is shown in Fig. 5.4 and is detailed in Fig. 5.5. A finite-state machine (FSM) and a counter are used. The counter output, t_s , is reset to 0 whenever the FSM transits between its internal states. Thus, t_s represents how long the FSM stays at its present state and is used to control its state transition. The FSM's two outputs D_1 and

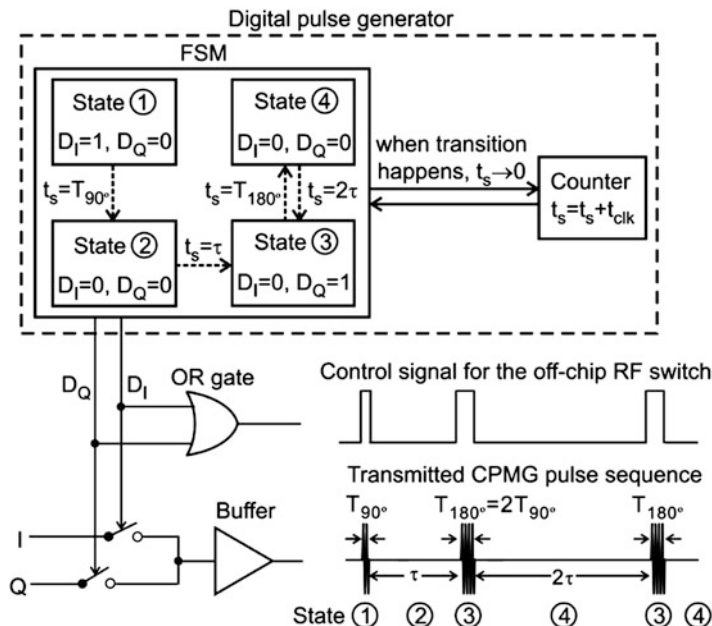


Fig. 5.5 The schematic of the digital pulse generator with the control signal and CPMG pulse sequence

D_Q gate the quadrature signals “I” and “Q” to produce the CPMG pulse sequence. D_I and D_Q also feed an OR gate, which controls the off-chip switch between the coil and the PA (Fig. 5.4). The FSM deals with 4 states, “1,” “2,” “3,” and “4” in Fig. 5.5, which represent the time duration of the 90° pulse, the time interval between the 90° pulse and the first 180° pulse, the time duration corresponding to any 180° pulse, and the time duration between any two adjacent 180° pulses, respectively. The FSM starts from state “1” in Fig. 5.5 and sequentially moves to states “2,” “3,” and “4” when t_s accumulates to the corresponding values. Then the FSM goes back and forth between states “3” and “4” to repeat the 180° pulses.

3.4 On-Chip NMR Coil Integration

It is also feasible to integrate the planar microcoil in the same TRX IC for lab-on-a-chip (LOC) operation. As illustrated in Fig. 5.6 [13, 14], this “1-chip” NMR system integrates the TRX together with the planar microcoil on the same IC. Due to thin metals, the 430 nH integrated microcoil has a Q of 1.9, even after connecting five metal layers in parallel; this low Q is due mainly to the coil’s dc resistance, while the substrate and skin effect are negligible at f_L . The chip is packaged with the coil part exposed and the rest encapsulated. This open package provides an effective container on top of the coil to hold a 5 μL sample, while the sample is separated

Fig. 5.6 The prototype of the “1-chip” NMR system with integrated TRX and sensing coil [13, 14]

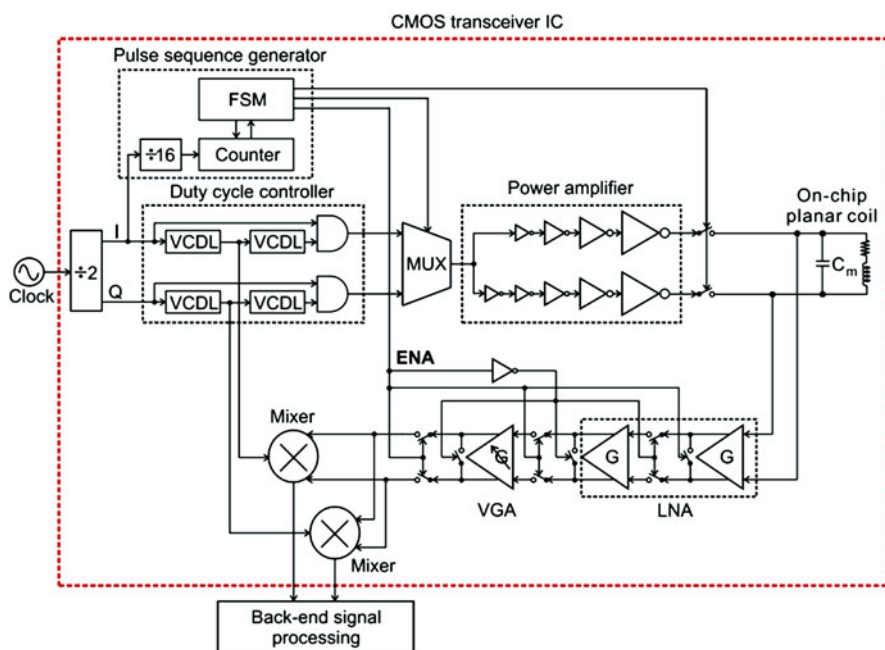
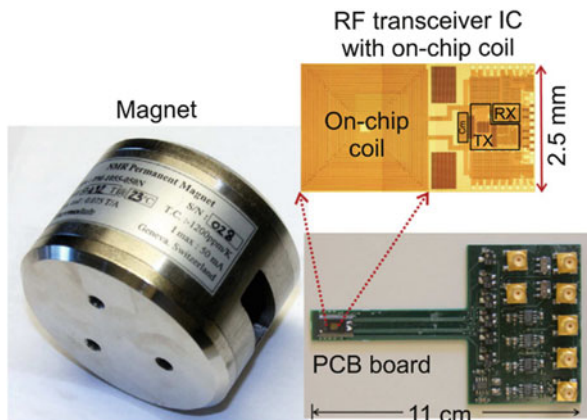


Fig. 5.7 CMOS RF transceiver for the “1-chip” NMR system

from the coil by the passivation layer native to the CMOS process. To cope with the SNR reduction due to the lossy coil, the TRX was redesigned. It also incorporates a PA on chip. The TRX architecture is shown in Fig. 5.7. The dashed lines show the integration boundary, which includes the coil, capacitor C_m for resonant noise matching, and TRX front end except the LO signal source. The mixer outputs feed the off-chip back-end signal processing unit.

The integrated PA is a differential chain of inverters that are quadrupled in size to amplify power and ensure the output drivability (Fig. 5.7). The output amplitude of this class-D PA is fixed at V_{DD} , while tuning the output power is necessary to control the 90° and 180° pulse durations (e.g., see Eq. (5.3) for the 90° pulse). For power tuning, we vary the duty cycle of the transmitted signal: the signal power at f_L is maximal for the duty cycle of 50% and is 0 for the duty cycle of 0%. Other harmonic components of the output signal will be irrelevant, as they cannot excite the proton spins, that is, this scheme exploits the natural high- Q filtering ability of the proton spin in the static magnetic field. The duty cycle is tuned by using voltage-controlled delay lines (VCDLs) and logic gates. The receiver is a revision from the receiver for the first prototype to further reduce the IRN, i.e., the active loads are replaced by resistive loads, the CMFB thus is removed, and two stages are cascaded to compensate for the reduced gain due to the resistive loads. The measured IRN of the entire receiver is $1.26 \text{ nV}/\sqrt{\text{Hz}}$, which is 30% less than that of reported for the first prototype. The measured noise figure is 2.2 dB, which is larger than that of the first prototype; this is because the voltage gain $\alpha = \sqrt{Q^2 + 1}$ of the noise matching network has been substantially reduced in this second prototype. The measured output impedance of the differential PA is 27Ω . The duty cycle can be tuned from 0% to 45%, which translates to the output power tuning from 0 to 80 mW.

3.5 NMR and Biomolecular Sensing Experiments

Figure 5.8 shows the measurement of the down-converted ^1H NMR signal with a water sample [12, 15]. The repeated ringings are spin-echo responses to the CPMG pulse sequence. $T_2 = 523 \text{ ms}$ is extracted from the exponentially decaying envelope of the spin echoes. The repeated spikes between the echoes are due to the coupling of the large excitation pulses.

The miniature NMR systems detected a range of biological molecules and compounds [12–15]. For example, avidin (if any) and MNPs ($\varnothing \sim 30 \text{ nm}$) coated with biotins are introduced into a water sample. In the absence of avidin, the particles stay monodispersed, yielding a T_2 of 73 ms. In the presence of avidin, the biotinylated MNPs bind to avidin, forming clusters [5] and reducing T_2 to 31 ms. The measured detection threshold is 20 fmol avidin (in $\sim 5 \mu\text{L}$, i.e., 4 nM).

Yet another example, now with the prototype with on-chip NMR sensing coil, is the detection of human bladder cancer cells (Fig. 5.9) [13, 14]. MNPs (\varnothing : 40 nm) coated with monoclonal antibody to bladder cancer cell surface markers are introduced into a water sample. In the absence of cancer cells, the MNPs are monodispersed, but in the presence of the cancer cells, MNPs bind to the cell surface. In the latter sample, centrifugation [41] separates the cells and unattached MNPs, and the unattached particles are washed away. The T_2 time difference is evident (Fig. 5.9), and system detects down to ~ 18 cells per μL .

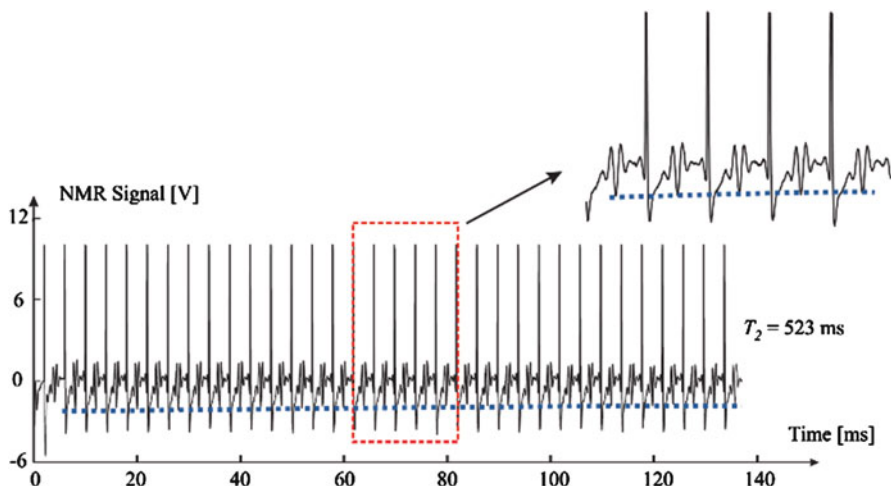


Fig. 5.8 Measured ^1H NMR signals from pure water [12, 15]

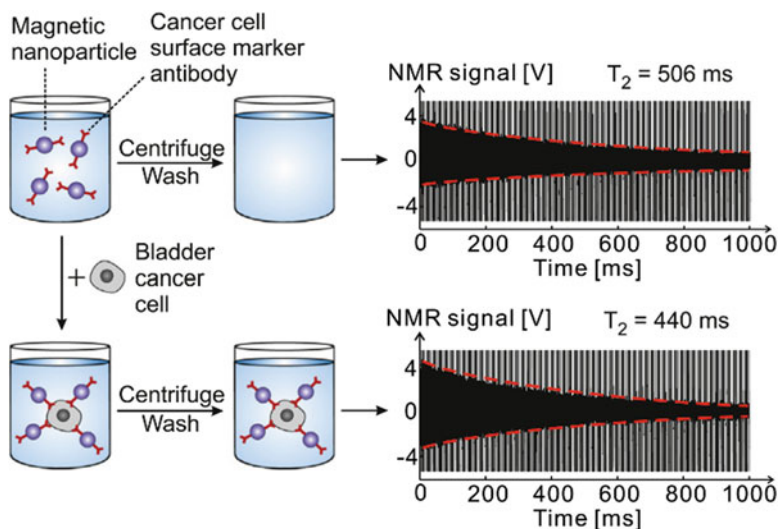


Fig. 5.9 Human bladder cancer cell detection with the “1-chip” NMR prototype [21, 22]

4 Electronic-Automated Sample Management for NMR

The previous works demonstrated the ability of the CMOS IC to miniaturize the size and footprint of the NMR system. Regrettably, it is inflexible to pipeline multiple samples to the NMR sensing region for higher throughput and real-time result comparison (e.g., concentration of the analytes) due to the limited NMR sensing region of the portable magnet. The operation of tiny samples beforehand,

which can involve multistep multisite treatments, relies heavily on the human efforts, degrading the throughput and consistency of diagnostic results while raising the chance of sample contamination. To address this issue, certain efforts have been undertaken to facilitate sample manipulation in NMR systems like capillary electrophoresis [42] and microfluidic channels [4, 34]. Still, these methods involve several laboratory accessories (e.g., pumps and pressure generators) and fixed fluidic paths/pipes that have low portability and reconfigurability and are inadequate for point-of-care (PoC) application.

Unlike conventional channel microfluidics, digital microfluidics (DMF) is highly amenable for co-integration, electronic automation, and reconfiguration [43–49]. This biocompatible platform has been adopted in a wide variety of biological applications, including cell culturing [43, 50, 51], DNA amplification [52–54], and single protein molecule capturing [55]. Microdroplets (e.g., $<10 \mu\text{L}$) in the DMF device can be transported over an electrode array by modifying the surface tension of the electrode utilizing the principle of electrowetting-on-dielectric (EWOD). Such distinct microdroplet controllability renders the DMF a promising droplet management platform for PoC devices. In addition, as the DMF device is planar, all droplets can be preloaded in the device before routinely executing the reaction or screening, enhancing the consistency of the experiments.

Herein we disclose the design of a NMR platform for chemical/biological assay (Fig. 5.10) [9–11]. It mainly equips with a CMOS TRX for performing NMR experiment and a DMF device and its electronics for electronic-automated sample management. The CMOS TRX has a similar structure as described in Sect. 1.3. This section will focus on the design of the DMF device and its co-integration with the NMR electronics.

4.1 Portable Magnet and RF Coil Codesign

This platform shares the same portable magnet (Fig. 5.11a) with the works described in Sect. 1.3. As discussed above, the planar coil on a low-cost, two-sided printed circuit board (PCB) is appealing for its consistency of parameters and disposability under volume production. Yet, typically the planar coil is a circular spiral as shown in Fig. 5.11b. The dominant magnetic field of this coil is in its axial direction (z -direction). This spiral coil is common due to its high sensitivity. However, since B_1 -field should be orthogonal to B_0 -field (z -direction) for NMR experiment, B_1 -field has to be in either the x - or the y -direction. For solenoids and spiral coils, the circular planes of the coils need to be in the x - z plane in order to generate B_1 -field in the y -direction. This restricts the usable space and thus the number of DMF electrodes inside the magnet since the width is $2.3 \times$ longer than the height. To resolve this, a PCB butterfly-coil containing two square spiral loops connected in series with different rotation (i.e., clockwise and counterclockwise) entails the generation of the plane-parallel B_1 -field (x -direction) (Fig. 5.11c). Routed with square loops, the butterfly-coil can effectively concentrate the magnetic field between the centers

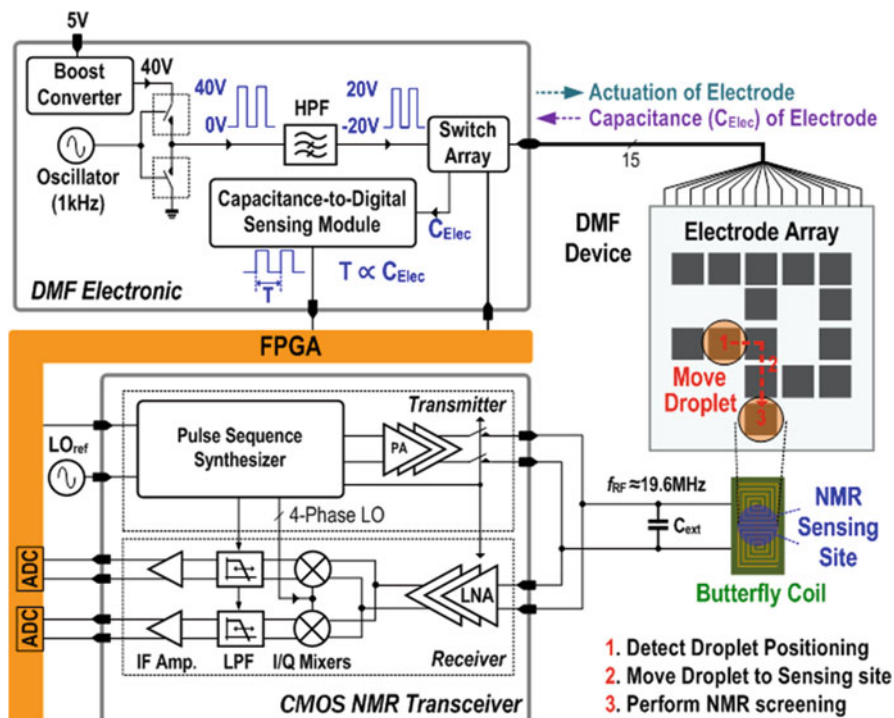


Fig. 5.10 Block diagram of the NMR TRX cooperated with the DMF device. It includes a CMOS NMR TRX with a butterfly-coil input, a DMF device, and a DMF electronic [9–11]

of the two loops ($\sim 40\%$ stronger than its circular loop counterpart). By inserting this butterfly-coil in the portable magnet (x - y plane), B_1 -field travels orthogonally to B_0 -field, easing the integration of the DMF device with higher number of electrodes inside the portable magnet. Yet, as the amplitude of the NMR signals is commensurate with B_1 , the system will have a lower SNR ascribed to the lower B_1 of the butterfly-coil when compared with the spiral counterpart ($0.5\times$). Nevertheless, it has been proven that the butterfly-coil is less susceptible to environmental couplings (e.g., powerline cables, equipment, and RF interference) since they appear as a common-mode noise [56–58].

4.2 DMF Device Fabrication and Droplet Actuation

DMF is a LOC technology based on the principle of EWOD. Droplets inside the DMF device are manipulated via surface-tension modulation induced by an electric field, enabling electronic-automated movement with low sample volumes. When compared with the conventional channel microfluidic devices, DMF can avoid

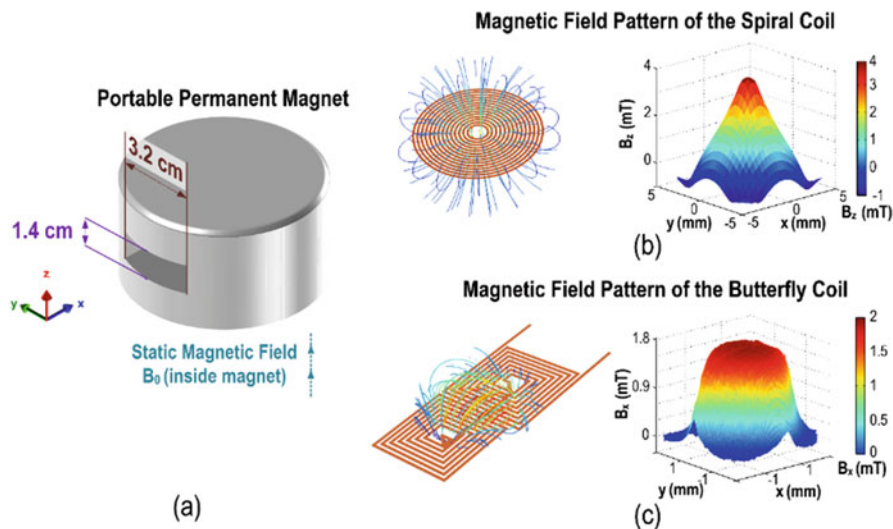


Fig. 5.11 (a) Geometry and limitation from the opening gap of the portable magnet. (b) and (c) The EM simulation of the magnetic field direction and strength from a spiral coil (with 14 turns) and a butterfly-coil (with 7 turns on each spiral), respectively

separate and complex networks of connections and laboratory gadgets such as pumps and valves, and the device is reconfigurable in such a way that the droplets can move freely over a surface of electrode matrix. The contact angle of the droplet is altered according to V [59]. By creating an unbalance contact angle on the droplet, there exists a force causing the movement of the droplets, and thus the droplets can be controlled electronically by driving the electrodes with a desired voltage signal.

The fabricated DMF device has 15 electrodes, each with $3.5 \times 3.5 \text{ mm}^2$. The one with chromium plating is patterned by lithography and wet etching to achieve customized electrode array, followed by Ta_2O_5 and Parylene C deposition to enhance the EWOD force. Indium tin oxide (ITO) coated on another glass plate (thickness, 0.5 mm) serves as a ground plane for all of the electrodes. A hydrophobic Teflon[®] layer covers both plates. To manipulate the droplets, the DMF electronics (Fig. 5.10) control the overall DMF device. For the actuation of droplets, a high voltage inverter (0–40 V) driven by an oscillator ($\sim 1 \text{ kHz}$) is utilized to generate a square wave for driving the electrodes, with the high voltage generated by a DC-DC boost converter. A switch array controls the on/off patterns of the electrodes. Further, a capacitance-to-digital module is included to sense the position of the droplets in real time, as the high-permittivity water droplets ($80\times$ of air) affect the capacitance of the electrodes (C_{Elec}) [60]. These electronics are integrated on the same PCB with the NMR system to enhance the integration level and facilitate the operations. A field-programmable gate array (FPGA) masters the operations including sample locating and transporting. This entire DMF module forms a closed-loop control for complete automated sample management, where the FPGA

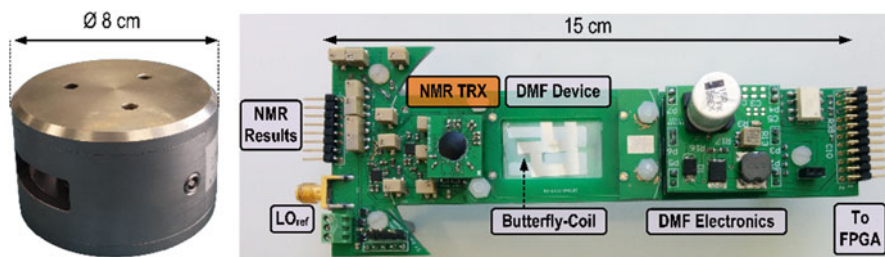


Fig. 5.12 The system hardware of the NMR system with a digital microfluidic device [9, 11]

controls the path of the droplets together with the operations of the NMR assays. To prevent crosstalk from appearing on the NMR results, during the NMR assay the DMF module switches off.

4.3 Experimental Results

Figure 5.12 illustrates the NMR system, including the DMF device and its control circuit, integrated on a single PCB for compactness and better reproducibility. An FPGA (DE0-Nano) monitors the schedule of sample movement and NMR signal acquisition and digitization. A software visualizes the NMR assay results and shows the execution status of the experimental protocol in real time. Samples are preloaded inside the DMF device before the experiment, with their transportations and mixing together with the triggering of NMR experiments executed electronically. The platform firstly starts with detecting the location of the samples inside the DMF device. The average pulses counted on the occupied and vacant electrodes are 277.5 and 757.7, respectively. This $2.73\times$ difference is sufficient to identify whether the electrodes are vacant. Moreover, with this sensing module adopted, the system is under a closed-loop control to attain an efficient droplet management scheme. The operations of the droplets including the trigger of the NMR experiment, stabilization of the ^1H nuclei before the NMR experiment, and antimerging droplets paths can be manipulated and optimized by the software within the shortest time to boost the efficiency and throughput of the system.

After the identification of the droplet location, the program starts to transport the droplets to the NMR sensing site. The droplets are guided to the destination gradually with their positions tracked in real time to ensure successful movement. To visualize these movements, the movement of the droplets was recorded outside the magnet. Figure 5.13a exhibits an example using a water droplet ($8\ \mu\text{L}$) routed to the NMR sensing site. Figure 5.13b shows the progressive movement of the droplet. The DMF platform guides the droplet to the corresponding electrode through the application of a voltage signal progressively to achieve sample transportation with an average velocity of 1.17 mm/s. The elevation of the actuation voltage improves

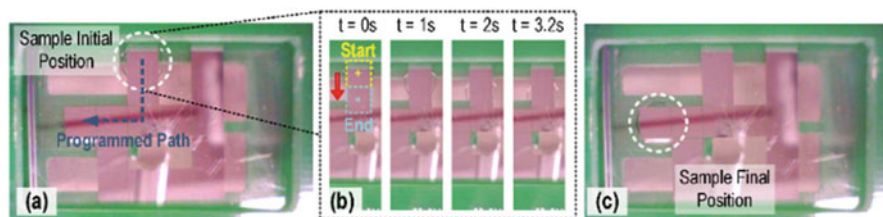


Fig. 5.13 Operation of the NMR system. (a) Initial position of the sample and its projected path. (b) Droplet moves to the adjacent electrode. (c) Final position of the droplet [9, 11]

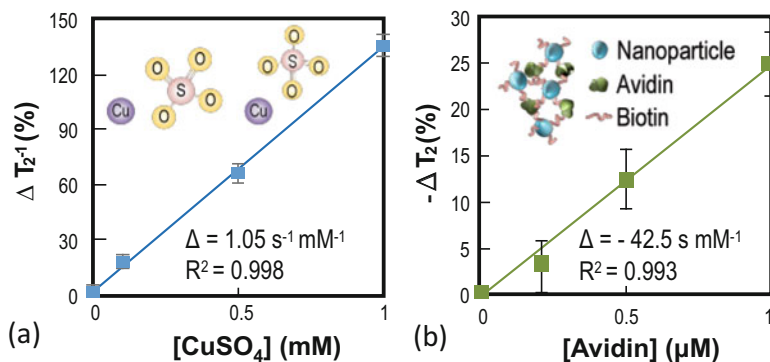


Fig. 5.14 (a) The correlation of ΔT_2^{-1} (with reference to 0 mM of CuSO_4) with the concentration of CuSO_4 . (b) The correlation of ΔT_2 (with reference to 0 μM of avidin) with the concentration of avidin [2, 52]

the velocity. Yet, this burdens the electric fields on the Ta_2O_5 layer and deteriorates the reliability of the DMF device. Thus, a moderate voltage of 40 V_{pp} was chosen, as the velocity is not a critical issue for the application.

Upon the arrival on the NMR sensing site, the system triggers the NMR assay automatically (Fig. 5.13c). After removing the image noise for the quadrature signals, the acquired NMR results are analyzed and displayed in the program. The T_2 from the water droplet ($1.16 \pm 0.03 \text{ s}$) can be derived, similar to the works described in Sect. 1.3.

According to the study, CuSO_4 will affect the T_2 of the water [61, 62]. We prepared CuSO_4 with different concentrations for the experiment. As shown in Fig. 5.14a, the NMR relaxometer can detect the CuSO_4 concentration with respect to T_2^{-1} . The second experiment demonstrates the capability of the system to pinpoint specific biological targets with predesigned probe-decorated MNPs similar to the works described in Sect. 1.3. Fig. 5.14b depicts the experimental results and shows that the T_2 value decreased proportionally to the concentration of avidin with an achieved sensitivity of 0.2 μM .

One unique feature of this NMR relaxometer is the capability to handle distinct samples and perform NMR assays on them sequentially. This feature is demon-

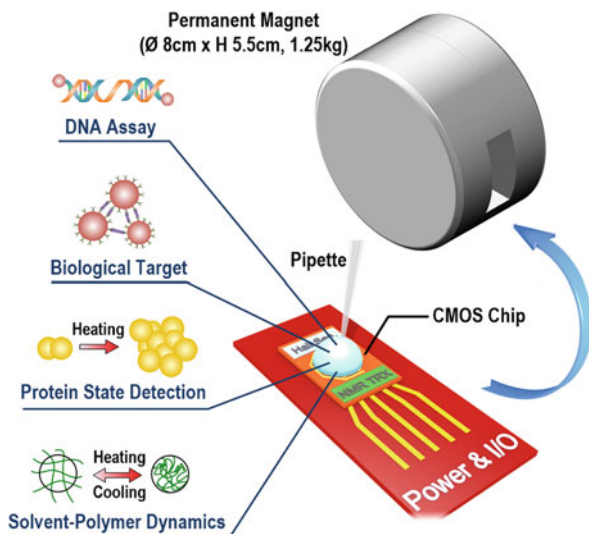
strated by placing two stationary targets and two identical probe-decorated MNPs droplets inside the DMF device at the same time. As the capacitance-to-digital module can track the location of the droplets, their individual paths can be procured at the software level. The 1st probe-decorated MNP droplet (7.5 μL) is firstly guided to a still target (2.5 μL) for mixing and then to the NMR sensing site to extract the T_2 . Concurrently, the 2nd probe-decorated MNP droplet is guided to another target for mixing. The 1st mixture is led away from the sensing site after finishing the diagnosis in 48 s, allowing the 2nd mixture to enter. The assays are completed after the 2nd mixture finished screening and the raw data are processed using the PC for concentration quantification (T_2 for the water sample, 256 ms; for avidin, 211 ms). Two or more probes and target pairs can be placed inside the DMF device for enhancing the throughput, depending on the geometry of the DMF device. With the droplet movement controlled by the program automatically and their positions tracked in real time, the optimization of the route and timing management can be done at the software level. This 2.2-min experiment validates the entire system as being capable to transport, mix, and analyze multiple distinct samples in real time while reducing the labor work (error) and the risks of defilement.

5 Magnetic Field Calibration for Portable NMR System

Another challenge for the CMOS NMR system with the portable magnet is the shifting of B_0 . The ambient temperature severely affects the f_L of the protons since f_L shifts with the temperature-dependent B_0 (temperature coefficient, -1000 ppm/K). Without calibration, LO frequency deviation from f_L will paralyze the system due to improper excitation frequency on the nuclei. This instability calls for a calibration scheme to enhance the robustness of the system for PoC application, especially at outdoor. Conventional frequency stabilization techniques are based on the measured NMR signals [63–65]. Yet, if the B_0 -field fluctuates large enough such that the excitation pulses cannot excite the nuclei effectively, those calibration schemes may not work properly.

To circumvent the above challenges, a trailblazing B_0 -field stabilization scheme for portable magnet is proposed here. As the primary influence on the operation is B_0 , sensing the B_0 directly can provide information for calibration immediately. Herein a handheld high-sensitivity NMR CMOS platform utilizing a portable magnet is reported, as illustrated in Fig. 5.15 [7, 8]. A Hall sensor with low-noise readout circuit is embedded with a CMOS NMR TRX to achieve better robustness. The current driver stabilizes the B_0 -field of the magnet against ambient variation. The stabilized B_0 avoids the need of a frequency synthesizer to tune the LO, and an untuned LO can be generated directly by the crystal oscillator. To minimize the B_0 -field offset error between the sample and the Hall sensor, the samples under assay are loaded on the on-chip planar coil by a handheld pipette. The sensing coil also serves as a sample heater for thermal profiling. Further, benefitting from the versatility of the NMR assays, this handheld tool unifies multi-type assays such as

Fig. 5.15 Conceptual diagram of the proposed NMR platform for PoU applications. Different samples such as protein and DNA can be put directly atop the CMOS chip for assays. A portable magnet is entailed to magnetize the nuclei inside the samples [7, 8]



target detection, protein state analysis, and solvent-polymer dynamics in a platform, rendering it suitable for healthcare, food industry, and colloidal applications.

Figure 5.16 shows the schematic of the proposed NMR platform. It includes B_0 -field stabilization to enhance the robustness and simplify the hardware. The Hall sensor and the readout circuit, together with an off-chip current driver, manage the lateral B_0 -field and stabilize the bulk magnetization and f_L of the nuclei. The dynamic B_1 -field transduction between the nuclei and the electronics is based on an on-chip planar coil driven by a TRX together with the matching capacitor C_M , to excite/obtain the magnetic signal to/from the droplet samples ($2.5 \mu\text{L}$) normal to the chip surface. The design of the CMOS TRX is similar to those in Sects. 1.3 and 1.4. Furthermore, the on-chip NMR sensing coil can act as a sample heater for thermal profiling. A BJT-based temperature sensor aids both the Hall sensor thermal correction and thermal profiling of the samples. The design of the Hall sensor and the readout circuit will be revealed in detail.

5.1 Hall Sensor, Readout Circuit, and Current Driver

The NMR TRX integrates with a Hall sensor to sense the B_0 -field variation of the permanent magnet, which must be appeared orthogonally to the B_1 -field of the planar coil. The detected B_0 -field variation modulates the current passing through the auxiliary coil of the magnet to stabilize the resultant B_0 -field, which allows system-level calibration.

Hall sensors can detect the magnetic field normal [66–68] or parallel [69, 70] to the chip surface. As the B_1 -field generated by the planar coil is normal to the

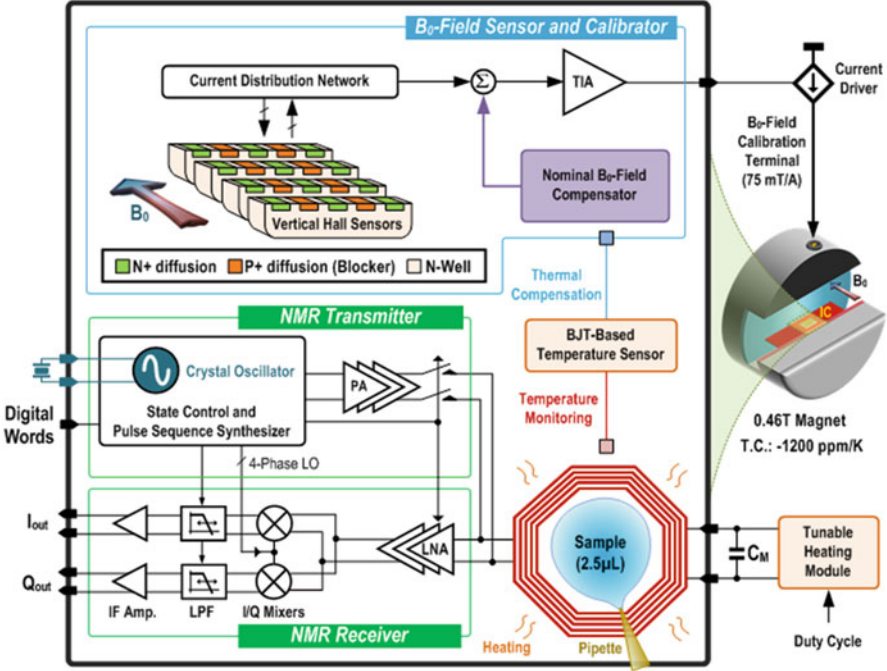


Fig. 5.16 System block diagram. The TX and RX transduce between magnetic and electrical signals with a thermal-controlled spiral coil. The B_0 -field sensor and calibrator automatically stabilize the bulk magnetization on the μL sample [7, 8]

chip surface and thus the B_0 -field has to be in lateral direction, the latter solution is required, which can be achieved by a vertical Hall sensor (VHS). Each VHS sub-element contains an n-well as the substrate and three n-diffusions as contacts. P-diffusions are embedded between the n-diffusions to avert the current flowing at the surface, soothing the $1/f$ noise. This architecture renders the device with full standard CMOS compatibility. Working in the current mode implies the input is a constant current source (I_{Bias}) and the outputs are two current terminals (I_P and I_N). When no external magnetic field exists, there is an equal split of I_{Bias} between I_P and I_N , which should be equal with no current difference. When considering B_0 in the lateral direction (i.e., normal to the cross section of the VHS device), the charge carriers from the input terminal will experience a Lorentz force deflecting the charges, leading to nonidentical magnitudes of I_P and I_N expressed by:

$$I_P = \frac{I_{\text{Bias}}}{2} + I_{\text{Hall}}(B_0), \quad (5.6)$$

$$I_N = \frac{I_{\text{Bias}}}{2} - I_{\text{Hall}}(B_0), \quad (5.7)$$

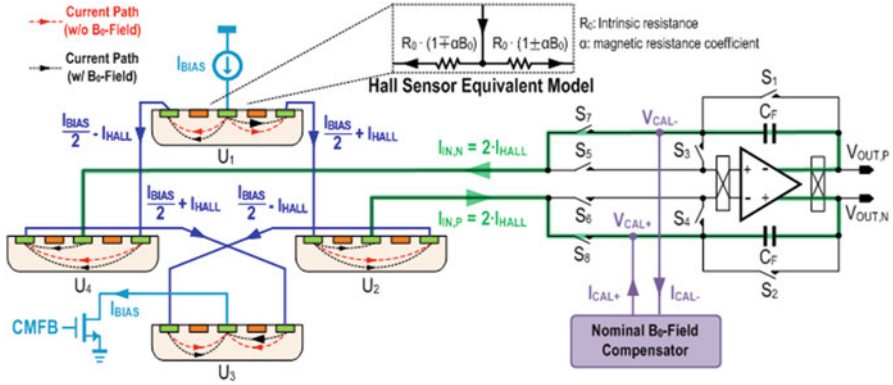


Fig. 5.17 Proposed current-mode fourfolded VHS arranged in Wheatstone bridge to sense the lateral B_0 -field and its readout circuit (spinning circuitry is omitted for simplicity). The green arrows highlight the current paths of I_{Hall}

where $I_{Hall}(B_0)$ is the induced Hall current on each output terminal commensurate with B_0 . Thus, the measurement of I_{Hall} can determine B_0 . Yet, I_{Hall} can be much smaller than I_{Bias} . Such prodigious bias component stiffens the measurement on I_P and I_N . To circumvent this measurement barrier, four identical VHS sub-elements U_{1-4} are arranged to form a Wheatstone bridge (Fig. 5.17). The Wheatstone bridge prunes I_{Bias} and only the B_0 -dependent term I_{Hall} appears at the Wheatstone bridge's output. Furthermore, this configuration not only features a fully differential architecture but also doubles the output Hall current improving the sensitivity of the VHS. A common-mode feedback circuit regulates the tail of the Wheatstone bridge, where two-phase spinning eliminates the effect of mismatch between the VHS sub-elements by periodically interchanging the output and supply terminals of the Wheatstone bridge.

The induced differential currents from the Wheatstone bridge ($I_{IN,P}$ and $I_{IN,N}$) are then converted to voltages for recording. Among multifarious options of transimpedance amplifier (TIA), the current integrator formed by a high-gain amplifier with shunt integrating capacitor (C_F) appears as a promising solution, since it inherently offers low-pass filtering on the outputs without passive noise sources (e.g., feedback resistors for resistive feedback TIA), leading to a better noise performance [71–73]. Furthermore, the variation of the integration time T_{INT} can alter the gain of the fully differential TIA, providing conversion flexibility. The core of this TIA is a two-stage amplifier with a telescopic first stage, where the DC gain (A_{DC} , 100 dB) guarantees accurate and stable operation (GBW = 100 MHz, PM = 50° at 15-pF loads), at a power budget of 2 mW excluding the bias circuit. A chopper deals with the offset and $1/f$ noise of the amplifier with a chopping frequency of 1 MHz. This chopping technique reduces the $1/f$ noise corner by 10,000× (from 200 kHz to 20 Hz) in an open-loop configuration. During the reset phase, S_{1-4} nullify the residual voltages on C_F , while opening S_{5-8} prevents the current to flow from the Wheatstone bridge into the TIA. In the measuring phase,

$I_{IN,P}$ and $I_{IN,N}$ flow through S_{7-8} and charge C_F , causing complementary voltage ramps at the outputs of the TIA, expressed by (assuming an ideal TIA):

$$V_{OUT} = V_{OUT,P} - V_{OUT,N} = \frac{4I_{Hall}T_{INT}}{C_F}. \quad (5.8)$$

After the measuring phase, the Hall currents from the Wheatstone bridge cease by switching off S_{5-8} , while turning on S_{3-4} allows the reading of the voltages on C_F . The three-phase operation can be repeated, and the results can be averaged to reduce the background noise.

5.2 Prototype and Experimental Results

Fabricated in the 0.18 μm CMOS process, the chip occupies an area of $2.0 \times 3.8 \text{ mm}^2$, dominated by the planar coil (dimension, $2.0 \times 2.0 \text{ mm}^2$; L , 506 nH; Q , 1.84) as shown in Fig. 5.18a. The dimension ($14 \times 6 \times 11 \text{ cm}^3$) and weight (1.4 kg) of the system are dominated by the 0.46-T portable magnet. In addition to the CMOS chip, there are a system PCB, a commercial FPGA (DE0-Nano), and a current driver. A customized program in the PC controls the platform, simplifying the operation and visualization of the assay results (Fig. 5.18b).

Before each NMR assay, the Hall sensor is turned on first with the VHS sensing the B_0 -field. As the B_0 -field may shift away from its nominal value due to environmental variation (e.g., temperature and sample-to-magnet position), an untracked f_L can be easily off-centered from the excitation frequency f_{EXC} (BW: 16.7 kHz, equivalent to 0.5 mT in terms of B_0 -field), affecting operation of the platform. With the proposed calibration scheme, the VHS and readout circuit track the B_0 -field variation; they show a sensitivity of 4.12 V/T. The eventual B_0 -field is then balanced by modulating the auxiliary coil of the magnet with a particular magnitude of DC current, according to the result from the VHS. Thus, f_L can be reset

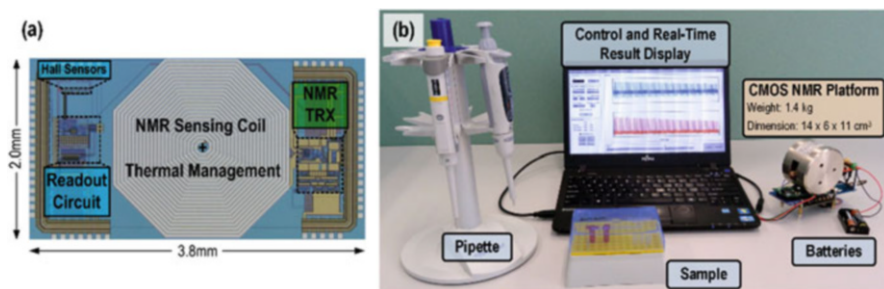


Fig. 5.18 (a) Chip photo; (b) experimental setup. The results are visualized in the PC, and the CMOS NMR platform is powered by two batteries for portability [7, 8]

to match with f_{EXC} to proceed the NMR assay. Associated with signal averaging performed in the frequency domain to suppress the background noise, the proposed calibration improves the B_0 -field stability by $13\times$ (from 2 to 0.15 mT) at 0.46 T nominal B_0 -field ($f_L = 19.6$ MHz), corresponding to a variation on f_L of 6.9 kHz.

One of the crucial aims of this platform is to detect and quantify the biological target inside the samples for disease screening. For instance, consider human immunoglobulin G (IgG), which protects the body from infections by binding themselves to different pathogens. Protein A, which specifically binds human IgG, is used as a probe to detect the existence of IgG inside the samples by functionalizing them on the water-soluble MNPs ($[\text{Fe}_2\text{O}_3]$, $10 \mu\text{g/mL}$; \emptyset , 25–30 nm). As illustrated in the previous works, these MNPs have significant impacts on T_2 of the samples according to the target concentration inside the samples. When IgG is absent in the sample, the MNPs stay monodispersed inside the solution, yielding a T_2 of 258 ms. Consequently, T_2 of the sample drops to 232.2 ms when the concentration of IgG is 12.5 nM inside the sample. To evince the selectivity of the NMR-based biomolecule screening, chicken immunoglobulin Y (IgY), which does not conjugate with the protein A, is also tested with the platform. T_2 from the samples has negligible change ($<2\%$) from varying concentration of chicken IgY (5–50 nM), thus validating the selectivity of the assay.

Alternatively, by selecting corresponding probe functionalized MNP, this versatile NMR-based screening scheme is capable of sensing widespread biomolecules. This is manifested by the detection of DNA for life-threatening bacteria screening. With a similar sensing mechanism, the platform quantifies the synthesized DNA derived from *Enterococcus faecalis* by pair of probe-decorated MNPs (\emptyset : 100 nm). The limit of detection of DNA for this platform, which is defined as $\Delta T_2 = 3\sigma$ above the blank sample, is estimated to be <100 pM. Additionally, the dynamic range of the detection can be impelled to 125 nM of DNA by varying the concentration of the MNP (from 6.25 to $10 \mu\text{g/mL}$). The response of the assay to single-nucleotide polymorphism is indistinguishable to T_2 baseline ($<4\%$), substantiating the possibility of differentiation of a single-base mismatch DNA.

6 Summary of Reported CMOS NMR Systems

Table 5.1 summarized the reported CMOS NMR system in this chapter. All of them are implemented with CMOS IC and portable magnet for footprint miniaturization. This was made possible by using small-sized, low-quality magnets and by designing high-performance integrated RF TRXs that counter the effect of the low-quality magnets. The works involved a repertoire of technical areas, ranging from IC to biomolecular detection to the science and technology of NMR.

7 Prospects for CMOS NMR Systems

There are continuous efforts to develop miniature CMOS IC for in vitro diagnosis as it enables attractive features such as low cost, chip-scale operation, massive parallelism, and accessibility. Compared to other transducing mechanisms such as electrical-based [74, 75] or fluorescent-based [76, 77] detection, the NMR-based biomolecule sensing mechanism here obviates complex hardware and sample processing steps (i.e., etching and immobilization of probe on the CMOS IC, labeling on the target biomolecule) before the assays [78]. It shows as an alternative for in vitro diagnostic platform.

Yet, the CMOS NMR systems and their biomolecular detection applications we reviewed in this paper are focused on the measurements of relaxation times, limited by the field inhomogeneity of the specific small magnets. The application scope of the small NMR systems will be significantly broadened, if their capability can be expanded into the realm of high-resolution spectroscopy. For instance, Ha et al. reported a CMOS NMR system for spectroscopy purpose with a larger permanent magnet (size, $12.6 \times 11.7 \times 11.9 \text{ cm}^3$) [63]. With both hardware-level inhomogeneity calibration and software-level drifting correction for the magnetic field, they demonstrated the detection of chemical structures from one-dimensional and two-dimensional ^1H NMR spectroscopy as well as two-dimensional heteronuclear ($^1\text{H}/^{13}\text{C}$) NMR spectroscopy. This potentially broadens the applicability of portable NMR from merely biosensing to drug discovery, chemical structure analysis, protein analysis, etc.

Beyond in vitro diagnosis applications, these CMOS NMR IC can also be utilized for biomedical imaging purposes with conventional superconducting magnet [38, 79, 80], benefitting from the versatility of NMR. Further, the concepts of integrating the RF IC for high-frequency atomic sensing are not only applicable to NMR but also appropriate to other spin-based magnetic interactions such as electron spin resonance (ESR). Focusing on detecting the spin of unpaired electrons instead of atomic nuclei for NMR, the operating frequency of ESR is 660 times of NMR counterpart under the same magnetic field (i.e., $\sim 14 \text{ GHz}$ with 0.5 T magnet), and thus it intrinsically has a higher sensitivity than NMR spectroscopy [81]. ESR is particularly useful for measuring oxidative stress of cells, which associates with the development of a variety of chronic and degenerative diseases such as cancer and Alzheimer's disease. Several works have reported the ESR spectrometer implemented with CMOS IC, which gear toward smaller footprint and lower costs for the overall platforms and demonstrated the measurement of DPPH from sample of volume down to 27 nL with the ESR spectrometer [82–84]. Thus there is a huge potential for spin-based magnetic sensing with CMOS IC and suggests an interesting research path.

Acknowledgments K.-M. Lei, P.-I. Mak, and R. Martins acknowledge the support from Macau FDCT – 047/2014/A1. N. Sun acknowledges NSF Grant No. 1254459. D. Ham acknowledges the STC Center for Integrated Quantum Materials, NSF Grant No. DMR-1231319.

References

1. M.A. Brown, R.C. Semelka, *MRI: Basic Principles and Applications*, 4th edn. (Wiley-Blackwell, 2010)
2. D. Canet, *Nuclear Magnetic Resonance: Concepts and Methods* (Wiley, New York, 1996)
3. H. Gunther, *NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry* (Weinheim, Germany, 1995)
4. H. Lee, E. Sun, D. Ham, R. Weissleder, Chip-NMR biosensor for detection and molecular analysis of cells. *Nat. Med.* **14**(8), 869–874 (2008)
5. J.M. Perez, L. Josephson, T. O’Loughlin, D. Hogemann, R. Weissleder, Magnetic relaxation switches capable of sensing molecular interactions. *Nat. Biotechnol.* **20**(8), 816–820 (2002)
6. J.A. Slichter, *Principles of Magnetic Resonance* (Springer, Heidelberg, 1990)
7. K.-M. Lei, H. Heidari, P.-I. Mak, M.-K. Law, F. Maloberti, R.P. Martins, A handheld high-sensitivity micro-NMR CMOS platform with B-field stabilization for multi-type biological/chemical assays. *IEEE J. Solid State Circuits* **52**(1), 284–297 (2017)
8. K.-M. Lei, H. Heidari, P.-I. Mak, M.-K. Law, F. Maloberti, R.P. Martins, A handheld 50pM-sensitivity micro-NMR CMOS platform with B-field stabilization for multi-type biological/chemical assays, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (2016b), pp. 474–475, Feb 2016
9. K.-M. Lei, P.-I. Mak, M.-K. Law, R.P. Martins, A μ NMR CMOS transceiver using a butterfly-coil input for integration with a digital microfluidic device inside a portable magnet. *IEEE J. Solid State Circuits* **51**(10), 2274–2286 (2016c)
10. K.-M. Lei, P.-I. Mak, M.-K. Law, R.P. Martins, A palm-size μ NMR relaxometer using a digital microfluidic device and a semiconductor transceiver for chemical/biological diagnosis. *Analyst* **140**(15), 5129–5137 (2015b)
11. K.-M. Lei, P.-I. Mak, M.-K. Law, R.P. Martins, A μ NMR CMOS transceiver using a butterfly-coil input for integration with a digital microfluidic device inside a portable magnet, in *2015 IEEE Asian Solid-State Circuits Conference (A-SSCC)* (2015c), pp. 1–4, 9–11 Nov 2015
12. Y. Liu, N. Sun, H. Lee, R. Weissleder, D. Ham, CMOS mini nuclear magnetic resonance system and its application for biomolecular sensing, in *2008 IEEE International Solid-State Circuits Conference (ISSCC)* (2008), pp. 140–602, 3–7 Feb 2008
13. N. Sun, T.J. Yoon, H. Lee, W. Andress, R. Weissleder, D. Ham, Palm NMR and 1-Chip NMR. *IEEE J. Solid State Circuits* **46**(1), 342–352 (2011)
14. N. Sun, T.J. Yoon, H. Lee, W. Andress, V. Demas, P. Prado, R. Weissleder, D. Ham, Palm NMR and one-chip NMR, in *2010 IEEE International Solid-State Circuits Conference – (ISSCC)* (2010), pp. 488–489, 7–11 Feb 2010
15. N. Sun, Y. Liu, H. Lee, R. Weissleder, D. Ham, CMOS RF biosensor utilizing nuclear magnetic resonance. *IEEE J. Solid State Circuits* **44**(5), 1629–1643 (2009)
16. D. Ham, A. Hajimiri, Virtual damping and Einstein relation in oscillators. *IEEE J. Solid State Circuits* **38**(3), 407–418 (2003)
17. X. Li, W. Zhu, D. Ham, Phase diffusion and lamb-shift-like spectrum shift in classical oscillators. (2010a). arXiv:0908.2214v3
18. X.F. Li, O.O. Yildirim, W.J. Zhu, D. Ham, Phase noise of distributed oscillators. *IEEE Trans. Microwave Theory Tech.* **58**(8), 2105–2117 (2010b)
19. D.I. Hoult, R.E. Richards, The signal-to-noise ratio of the nuclear magnetic resonance experiment. *J. Magn. Reson.* **24**(1), 71–85 (1976)
20. R. Gruetter, Automatic, localized in vivo adjustment of all 1st-order and 2nd-order shim coils. *Magn. Reson. Med.* **29**(6), 804–811 (1993)
21. E. Danielli, J. Perlo, B. Blümich, F. Casanova, Small magnets for portable NMR spectrometers. *Angew. Chem. Int. Ed.* **49**(24), 4133–4135 (2010)
22. E. Danielli, J. Perlo, F. Casanova, B. Blümich, High-performance shimming with permanent magnets, in *Magnetic Resonance Microscopy*, 1st edn., (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009a), pp. 485–499

23. E. Danieli, J. Mauler, J. Perlo, B. Blumich, F. Casanova, Mobile sensor for high resolution NMR spectroscopy and imaging. *J. Magn. Reson.* **198**(1), 80–87 (2009b)
24. H.Y. Carr, E.M. Purcell, Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Phys. Rev.* **94**(3), 630–638 (1954)
25. S. Meiboom, D. Gill, Modified spin-echo method for measuring nuclear relaxation times. *Rev. Sci. Instrum.* **29**(8), 688–691 (1958)
26. B. Behnia, A.G. Webb, Limited-sample NMR using solenoidal microcoils perfluorocarbon plugs, and capillary spinning. *Anal. Chem.* **70**(24), 5326–5331 (1998)
27. D.L. Olson, T.L. Peck, A.G. Webb, R.L. Magin, J.V. Sweedler, High-resolution microcoil ¹H-NMR for mass-limited, nanoliter-volume samples. *Science* **270**(5244), 1967–1970 (1995)
28. F.D. Doty, Probe design and construction, in *Encyclopedia of Magnetic Resonance*, (Wiley, New York, 2007)
29. A.P.M. Kentgens, J. Bart, P.J.M. van Bentum, A. Brinkmann, E.R.H. Van Eck, J.G.E. Gardeniers, J.W.G. Janssen, P. Knijn, S. Vasa, M.H.W. Verkuijlen, High-resolution liquid- and solid-state nuclear magnetic resonance of nanoliter sample volumes using microcoil detectors. *J. Chem. Phys.* **128**(5) (2008)
30. K.R. Minard, R.A. Wind, Solenoidal microcoil design – part II: optimizing winding parameters for maximum signal-to-noise performance. *Concepts Magn. Reson.* **13**(3), 190–210 (2001a)
31. K.R. Minard, R.A. Wind, Solenoidal microcoil design. Part I: optimizing RF homogeneity and coil dimensions. *Concepts Magn. Reson.* **13**(2), 128–142 (2001b)
32. R.M. Fratila, A.H. Velders, Small-volume nuclear magnetic resonance spectroscopy. *Annu. Rev. Anal. Chem.* **4**(1), 227–249 (2011)
33. C.J. Jones, C.K. Larive, Could smaller really be better? Current and future trends in high-resolution microcoil NMR spectroscopy. *Anal. Bioanal. Chem.* **402**(1), 61–68 (2012)
34. C. Massin, F. Vincent, A. Homsy, K. Ehrmann, G. Boero, P.A. Besse, A. Daridon, E. Verpoorte, N.F. de Rooij, R.S. Popovic, Planar microcoil-based microfluidic NMR probes. *J. Magn. Reson.* **164**(2), 242–255 (2003)
35. C. Massin, C. Azevedo, N. Beckmann, P.A. Besse, R.S. Popovic, Magnetic resonance imaging using microfabricated planar coils, in *2nd Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology* (2002), pp. 199–204
36. C. Massin, A. Daridon, F. Vincent, G. Boero, P.-A. Besse, E. Verpoorte, N.F. de Rooij, R.S. Popovic, A microfabricated probe with integrated coils and channels for on-chip NMR spectroscopy, in *Micro Total Analysis Systems 2001: Proceedings of the μ TAS 2001 Symposium, Held in Monterey*, 1st edn., ed. by J. M. Ramsey, A. van den Berg (Springer, Dordrecht, 2001), pp. 438–440, 21–25 Oct 2001
37. H. Ryan, S.H. Song, A. Zass, J. Korvink, M. Utz, Contactless NMR spectroscopy on a chip. *Anal. Chem.* **84**(8), 3696–3702 (2012)
38. J. Anders, G. Chiamonte, P. SanGiorgio, G. Boero, A single-chip array of NMR receivers. *J. Magn. Reson.* **201**(2), 239–249 (2009)
39. V. Badilita, K. Kratt, N. Baxan, J. Anders, D. Elverfeldt, G. Boero, J. Hennig, J.G. Korvink, U. Wallrabe, 3D solenoidal microcoil arrays with CMOS integrated amplifiers for parallel MR imaging and spectroscopy, in *2011 IEEE 24th International Conference on Micro Electro Mechanical Systems* (2011), pp. 809–812, 23–27 Jan 2011
40. D.I. Hoult, The NMR receiver: a description and analysis of design. *Prog. Nucl. Magn. Reson. Spectrosc.* **12**(1), 41–77 (1978)
41. H. Lee, T.J. Yoon, J.L. Figueiredo, F.K. Swirski, R. Weissleder, Rapid detection and profiling of cancer cells in fine-needle aspirates. *Proc. Natl. Acad. Sci.* **106**(30), 12459–12464 (2009)
42. J.D. Trumbull, I.K. Glasgow, D.J. Beebe, R.L. Magin, Integrating microfabricated fluidic systems and NMR spectroscopy. *I.E.E.E. Trans. Biomed. Eng.* **47**(1), 3–7 (2000)
43. I. Barbulovic-Nad, H. Yang, P.S. Park, A.R. Wheeler, Digital microfluidics for cell-based assays. *Lab Chip* **8**(4), 519–526 (2008)
44. J. Gao, X.M. Liu, T.L. Chen, P.I. Mak, Y.G. Du, M.I. Vai, B.C. Lin, R.P. Martins, An intelligent digital microfluidic system with fuzzy-enhanced feedback for multi-droplet manipulation. *Lab Chip* **13**(3), 443–451 (2013)

45. F. Lapierre, M. Harnois, Y. Coffinier, R. Boukherroub, V. Thomy, Split and flow: reconfigurable capillary connection for digital microfluidic devices. *Lab Chip* **14**(18), 3589–3593 (2014)
46. M.G. Pollack, A.D. Shenderov, R.B. Fair, Electrowetting-based actuation of droplets for integrated microfluidics. *Lab Chip* **2**(2), 96–101 (2002)
47. M.H. Shamsi, K. Choi, A.H.C. Ng, A.R. Wheeler, A digital microfluidic electrochemical immunoassay. *Lab Chip* **14**(3), 547–554 (2014)
48. V. Srinivasan, V.K. Pamula, R.B. Fair, An integrated digital microfluidic lab-on-a-chip for clinical diagnostics on human physiological fluids. *Lab Chip* **4**(4), 310–315 (2004)
49. A.R. Wheeler, Chemistry – putting electrowetting to work. *Science* **322**(5901), 539–540 (2008)
50. I. Barbulovic-Nad, S.H. Au, A.R. Wheeler, A microfluidic platform for complete mammalian cell culture. *Lab Chip* **10**(12), 1536–1542 (2010)
51. G.J. Shah, A.T. Ohta, E.P.Y. Chiou, M.C. Wu, C.-J. Kim, EWOD-driven droplet microfluidic device integrated with optoelectronic tweezers as an automated platform for cellular isolation and analysis. *Lab Chip* **9**(12), 1732–1739 (2009)
52. Y.-H. Chang, G.-B. Lee, F.-C. Huang, Y.-Y. Chen, J.-L. Lin, Integrated polymerase chain reaction chips utilizing digital microfluidics. *Biomed. Microdevices* **8**(3), 215–225 (2006)
53. Z. Hua, J.L. Rouse, A.E. Eckhardt, V. Srinivasan, V.K. Pamula, W.A. Schell, J.L. Benton, T.G. Mitchell, M.G. Pollack, Multiplexed real-time polymerase chain reaction on a digital microfluidic platform. *Anal. Chem.* **82**(6), 2310–2316 (2010)
54. R. Sista, Z. Hua, P. Thwar, A. Sudarsan, V. Srinivasan, A. Eckhardt, M. Pollack, V. Pamula, Development of a digital microfluidic platform for point of care testing. *Lab Chip* **8**(12), 2091–2104 (2008)
55. D. Witters, K. Knez, F. Ceysens, R. Puers, J. Lammertyn, Digital microfluidics-enabled single-molecule detection by printing and sealing single magnetic beads in femtoliter droplets. *Lab Chip* **13**(11), 2047–2054 (2013)
56. P. Andreani, K. Kozmin, P. Sandrup, M. Nilsson, T. Mattsson, A TX VCO for WCDMA/EDGE in 90 nm RF CMOS. *IEEE J. Solid State Circuits* **46**(7), 1618–1626 (2011)
57. T. Mattsson, Method of and inductor layout for reduced VCO coupling. U.S. Patent 7,151,430, 19 Dec 2006
58. M. Nagata, H. Masuoka, S.I. Fukase, M. Kikuta, M. Morita, N. Itoh, 5.8 GHz RF transceiver LSI including on-chip matching circuits, in *Bipolar/BiCMOS Circuits and Technology Meeting* (2006), pp. 263–266, Oct 2006
59. F. Mugele, J.C. Baret, Electrowetting: from basics to applications. *J. Phys. Condens. Matter* **17**(28), R705–R774 (2005)
60. J. Gong, C.J. Kim, All-electronic droplet generation on-chip with real-time feedback control for EWOD digital microfluidics. *Lab Chip* **8**(6), 898–906 (2008)
61. W.K. Peng, L. Chen, J. Han, Development of miniaturized, portable magnetic resonance relaxometry system for point-of-care medical diagnosis. *Rev. Sci. Instrum.* **83**(9) (2012)
62. J.M. Pope, N. Repin, A simple approach to T2 imaging in MRI. *Magn. Reson. Imaging* **6**(6), 641–646 (1988)
63. D. Ha, J. Paulsen, N. Sun, Y.Q. Song, D. Ham, Scalable NMR spectroscopy with semiconductor chips. *Proc. Natl. Acad. Sci.* **111**(33), 11955–11960 (2014)
64. E. Kupce, R. Freeman, Molecular structure from a single NMR sequence (fast-PANACEA). *J. Magn. Reson.* **206**(1), 147–153 (2010)
65. G.A. Morris, H. Barjat, T.J. Horne, Reference deconvolution methods. *Prog. Nucl. Magn. Reson. Spectrosc.* **31**(1), 197–257 (1997)
66. H. Heidari, E. Bonizzoni, U. Gatti, F. Maloberti, A CMOS current-mode magnetic hall sensor with integrated front-end. *IEEE Trans. Circuits Syst. I Regul. Pap.* **62**(5), 1270–1278 (2015)
67. J. Jiang, K. Makinwa, A hybrid multipath CMOS magnetic sensor with 210 μ m resolution and 3MHz bandwidth for contactless current sensing, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (2016), pp. 204–205, Feb 2016
68. J.F. Jiang, W.J. Kindt, K.A.A. Makinwa, A continuous-time ripple reduction technique for spinning-current hall sensors. *IEEE J. Solid State Circuits* **49**(7), 1525–1534 (2014)

69. C. Sander, M.C. Vecchi, M. Cornils, O. Paul, From three-contact vertical hall elements to symmetrized vertical hall sensors with low offset. *Sens Actuators, A* **240**, 92–102 (2016)
70. G.M. Sung, C.P. Yu, 2-D differential folded vertical hall device fabricated on a p-type substrate using CMOS technology. *IEEE Sensors J.* **13**(6), 2253–2262 (2013)
71. M. Crescentini, M. Bennati, M. Carminati, M. Tartagni, Noise limits of CMOS current interfaces for biosensors: a review. *IEEE Trans. Biomed. Circuits Syst.* **8**(2), 278–292 (2014)
72. D. Kim, B. Goldstein, W. Tang, F.J. Sigworth, E. Culurciello, Noise analysis and performance comparison of low current measurement systems for biomedical applications. *IEEE Trans. Biomed. Circuits Syst.* **7**(1), 52–62 (2013)
73. K.-M. Lei, H. Heidari, P.-I. Mak, M.-K. Law, F. Maloberti, Exploring the noise limits of fully-differential micro-watt transimpedance amplifiers for Sub-pA/ $\sqrt{\text{Hz}}$ sensitivity, in *2015 11th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)* (2015a), pp. 290–293, June 2015
74. L.K. Lee, S. Choi, J.O. Lee, Yoon JB, Cho GH (2012) CMOS capacitive biosensor with enhanced sensitivity for label-free DNA detection, in *2012 IEEE International Solid-State Circuits Conference*, pp. 120–122, Feb 2012
75. A. Manickam, A. Chevalier, M. McDermott, A.D. Ellington, A. Hassibi, A CMOS electrochemical impedance spectroscopy (EIS) biosensor array. *IEEE Trans. Biomed. Circuits Syst.* **4**(6), 379–390 (2010)
76. T.C.D. Huang, S. Sorgenfrei, P. Gong, R. Levicky, K.L. Shepard, A 0.18- μm CMOS array sensor for integrated time-resolved fluorescence detection. *IEEE J. Solid State Circuits* **44**(5), 1644–1654 (2009)
77. H. Norian, R.M. Field, I. Kymissis, K.L. Shepard, An integrated CMOS quantitative-polymerase-chain-reaction lab-on-chip for point-of-care diagnostics. *Lab Chip* **14**(20), 4076–4084 (2014)
78. K.-M. Lei, P.-I. Mak, M.-K. Law, R.P. Martins, CMOS biosensors for in vitro diagnosis – transducing mechanisms and applications. *Lab Chip* **16**, 3664–3681 (2016a)
79. J. Anders, P. SanGiorgio, G. Boero, A fully integrated IQ-receiver for NMR microscopy. *J. Magn. Reson.* **209**(1), 1–7 (2011)
80. J. Handwerker, M. Eder, M. Tibiletti, V. Rasche, K. Scheffler, J. Becker, M. Ortmanns, J. Anders, An array of fully-integrated quadrature TX/RX NMR field probes for MRI trajectory mapping, in *42nd European Solid-State Circuits Conference* (2016b), pp. 217–220, 12–15 Sept 2016
81. M.J.N. Junk, Electron paramagnetic resonance theory, in *Assessing the Functional Structure of Molecular Transporters by EPR Spectroscopy*, (Springer, Berlin/Heidelberg, 2012), pp. 7–52
82. J. Handwerker, B. Schlecker, U. Wachter, P. Radermacher, M. Ortmanns, J. Anders, A 14GHz battery-operated point-of-care ESR spectrometer based on a 0.13 μm CMOS ASIC, in *2016 IEEE International Solid-State Circuits Conference (ISSCC)* (2016a), pp. 476–477, 31 Jan–4 Feb 2016
83. X. Yang, A. Babakhani, A full-duplex single-chip transceiver with self-interference cancellation in 0.13 μm SiGe BiCMOS for electron paramagnetic resonance spectroscopy. *IEEE J. Solid State Circuits* **51**(10), 2408–2419 (2016)
84. X. Yang, A. Babakhani, A single-chip electron paramagnetic resonance transceiver in 0.13- μm SiGe BiCMOS. *IEEE Trans. Microwave Theory Tech.* **63**(11), 3727–3735 (2015)

Chapter 6

Microelectronics for Muscle Fatigue Monitoring Through Surface EMG

Pantelis Georgiou and Ermis Koutsos

1 Introduction

Electromyography (EMG) is a technique used to evaluate the electrical activity of muscles. Etienne-Jules Marey was the first to record the electrical activity of muscles in 1890, introducing the term electromyography [1]. Muscle activity is controlled by the nervous system and the resulting muscle contractions give rise to electrical currents which in turn form the EMG signal. The EMG signal appears to be random, yet it is a rich source of information which can facilitate such an insight into our muscles and especially their activation and fatigue level.

Muscles are vital for strength, balance, heat, posture and movement. Feedback driven from the way our muscles behave and fatigue has the potential to improve the quality of life of millions of people suffering from muscular related disorders and set the foundations for ground-breaking rehabilitation technologies. To date, there are no reliable metrics of fatigue and point-of-care autonomous systems are still a dream. In this chapter we present work which addresses the challenge of reliably and efficiently estimating a muscle's fatigue state and the implementation of integrated systems to detect it using CMOS microelectronic technology.

2 Muscle Fatigue

There exist three types of fatigue: localised muscle fatigue, central fatigue (nervous system) and peripheral fatigue. Localised muscle fatigue analysis has numerous

P. Georgiou (✉) • E. Koutsos

Centre for Bio-Inspired Technology, Institute of Biomedical Engineering, Imperial College London, London SW7 2AZ, UK

e-mail: pantelis@imperial.ac.uk; ekoutsos@imperial.ac.uk

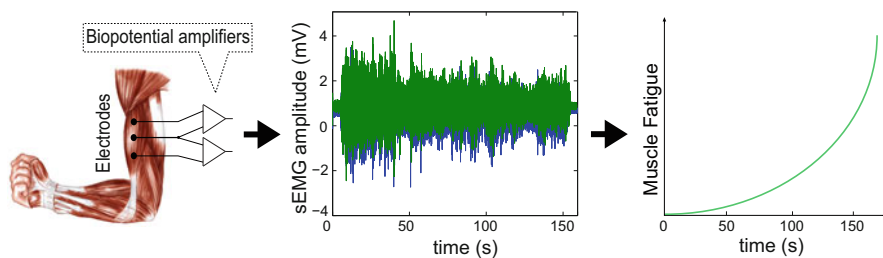


Fig. 6.1 EMG biofeedback cycle, where the EMG is amplified, visualised and processed to take advantage of the resulting muscle fatigue estimate

applications. The demand for EMG analysis and biofeedback in rehabilitation (surgery recovery, stroke recovery, spinal cord injuries, brain injuries) and its application in conjunction with Electroencephalogram (EEG) for identifying neuropathies, brain disorders and assisting Brain Machine Interfaces (BMIs) is rapidly growing in translational healthcare. EMG analysis is advantageous in these areas by using measured results from the body to change the way we behave, improve our performance and achieve better compliance to rehabilitation through the process of biofeedback. Consider Fig. 6.1 for a visual representation of the EMG biofeedback cycle.

EMG analysis has been widely applied in ergonomics to isolate, understand and treat muscle pain; fatigue analysis has been leveraged to study the effects of exposures at work. In heavy work environments, mechanical overstress (fatigue) from prolonged low-level contractions can be harmful, causing injuries and leading to muscular disorders [2]. Another example involves long distance truck drivers, where excessive fatigue in the shoulder muscles can lead to driver loss of focus. In pursuit of identifying the underlying causes of back and neck pain, a condition affecting almost 80% of all people, EMG has been used to assess changes in muscle fatigue, recruitment and coordination [3, 4]. EMG and fatigue analysis plays a vital role in biofeedback rehabilitation. For example, knee rehabilitation after surgery due to osteoarthritis focuses on maintaining a balance between the two large muscles that hold the patella (knee cap) in place, Vastus Lateralis and Vastus Medialis. Careful tracking of muscle fatigue of these two muscles can provide essential adjustments to the rehabilitation procedure and avoid patellar dysfunction [5]. Finally, strength training and movement analysis in sports science have greatly benefited from EMG feedback. Monitoring muscle activity and adjusting training regimes can maximise effectiveness and avoid injuries due to excessive fatigue.

Modern fatigue analysis tools involve collecting large amounts of EMG data, using either wireless or wired approaches. Decoding the EMG and extracting specific information is computationally demanding and is performed in workstations, also involving advanced signal processing techniques and dedicated personnel continuously monitoring the process. Using wires introduces noise sources, which deteriorate the quality of the EMG signal and limit the mobility of the user in various exercises. Moreover, these techniques limit considerably the possible applications

based on muscle fatigue, because they reduce the user’s degree of freedom and the number of available measuring sites. On the other hand, wireless approaches continuously digitise and transmit the raw EMG signal. To increase effectiveness of biofeedback, numerous sensors are needed to monitor several muscle groups. However, the sheer amount of transmitted data and processing requirements form a bottleneck and challenge biofeedback feasibility.

Increasing the number of channels while reducing power consumption can be achieved by introducing localised processing techniques and thus reducing the data rate. Efforts to create a portable fatigue monitor date back over three decades. At that time, most of the designs involved either custom analogue circuits or micro-controller based, digital implementations [6–12]. Due to the limited resources of the time, the designs offered limited tracking capabilities and involved bulky equipment. Even though modern microcontrollers are becoming more power efficient and more capable, microcontroller based systems are still limited by the number of EMG channels that can be processed in parallel and range of processing methods that can allow efficient computation. Whilst the current approaches offer invaluable insight, real-time biofeedback outside a controlled environment remains a critical barrier. Consequently, there is a clear need for a portable, compact and wearable system, capable of unsupervised EMG fatigue analysis. Figure 6.2 depicts the limitations of the current approaches for EMG muscle fatigue analysis along with the application areas that will benefit in the future from real-time biofeedback systems. Most of the EMG applications shown in Fig. 6.2 are limited to clinical/laboratory environments and have not made their way in our daily lives due to the lack of real-time biofeedback systems. However, with the use of CMOS technology, EMG detection and processing can be efficiently combined to create a miniaturised EMG system which can be easily scaled to monitor several muscles concurrently and efficiently, allowing for real-time, continuous operation. Thus, CMOS technology can be exploited to unlock the feasibility of biofeedback in the future.








Current Technologies	EMG Fatigue Analysis Applications	Future of Biofeedback
<div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>Wireless</p>  </div> <div style="text-align: center;"> <p>Wired</p>  </div> </div>  <p>www.delsys.com</p> <ul style="list-style-type: none"> + Wired systems are accurate, ideal for clinical use + Portable systems extend user mobility - Post processing of EMG - Gap between recording systems and analysis tools - Limited application range - Lack of real-time feedback 	<div style="border: 1px solid blue; border-radius: 10px; padding: 5px; margin-bottom: 10px;"> <p>Biomechanics Strength training Active therapy training Surgery rehabilitation</p> </div> <p>Orthopedic research Posture analysis Risk prevention Design improvements product certification</p> <div style="text-align: center; margin-top: 10px;">  <p>CMOS Technology</p> </div> <ul style="list-style-type: none"> + Low power and small form factor + Embedded custom algorithms + Combined EMG detection and processing + Real-time biofeedback feasibility 	<div style="display: flex; justify-content: space-around;">   </div>  <p>www.athos.com</p> <ul style="list-style-type: none"> + Local detection and processing + Information driven system + Scalable, integrated electronics + Multimodal, real-time monitoring + Continuous, multi-site monitoring + Real-time biofeedback

Fig. 6.2 CMOS technology can unlock the potential of future EMG biofeedback systems by overcoming the limitations current technologies are facing. Integration of detection and processing stages will significantly reduce data transmission and allow several EMG channels to be processed concurrently, thus increasing the insight and effectiveness of biofeedback in the future

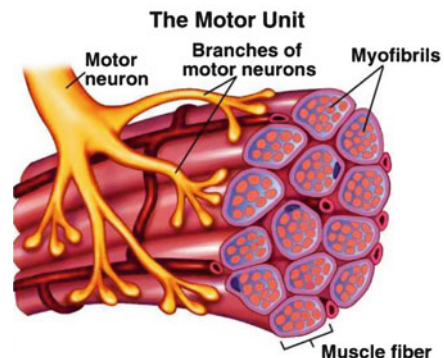
3 Fundamentals of EMG and Muscle Fatigue Analysis

3.1 Muscle Contraction and EMG Formation

The muscle consists of cells called muscle fibres. Muscle fibres group together into forming motor units (MU) and many MUs make up a muscle. Every MU has a dedicated motor neuron that excites it. Hence, all the muscle fibres comprising the MU act as one unit, flexing or relaxing together. A diagram of the MU can be seen in Fig. 6.3. Along the fibre axis there is a membrane called the sarcoplasmic reticulum, which stores and releases the calcium ions (Ca^{2+}) that trigger the muscle contraction [12].

To contract a muscle, an electrical impulse from the motor neuron travels to the muscle and when it reaches the neuron synapse, it releases a chemical message, which in turn causes an action potential in the muscle cell (AP). Since all muscle fibres inside a MU contract simultaneously, each MU gives a single AP (MUAP), formed by the superposition of all the muscle fibre APs. As more force is needed, more motor units are recruited and the firing rate is increased. The superposition of many MUAPs form a bigger wave and as this wave travels, an increasing potential difference is measured between the surface electrodes. Muscles get energy from adenosine triphosphate (ATP). ATP is made in the mitochondria inside the muscle fibre cells. During an anaerobic respiration, glucose is broken down to lactic acid and ATP. During an aerobic respiration, glucose, glycogen, fats and amino acids are broken down with oxygen and ATP is produced. Muscles consist of two basic types of fibres; fast twitch and slow twitch fibres. Fast-twitch fibres are capable of developing greater forces, contracting faster and have greater anaerobic capacity. In contrast, slow-twitch fibres develop force slowly, they can maintain contractions longer and they have higher aerobic capacity [12].

Fig. 6.3 Muscle anatomy showing motor unit, motor neuron and muscle fibre. Figure from [13]



3.2 *EMG Detection*

There are two ways to detect and record EMG signals; by using invasive needle-electrodes inserted in the muscle or non-invasive surface electrodes. With needle electrodes one can identify individual MUAPs. The recorded sEMG signal is a superposition of all MUAPs from the neighbouring MUs close to the skin electrode. The sEMG signal appears random in nature and has an amplitude distribution in the range of 0–5 mV prior to amplification. Thus, a signal amplifier is required. Typically, a differential amplifier is used as a first stage amplifier [1]. As any biomedical signal, sEMG is affected by noise. Noise accumulates from several sources; inherent noise of the electronics used, ambient noise (large offset) and motion artefacts—arising from electrode interference and electrode cables [1]. The bandwidth of the sEMG signal is between 10 and 500 Hz [14]. The limited bandwidth of the signal makes it easier to remove motion artefacts, DC offsets and high frequency noise. The Surface Electromyography for Noninvasive Assessment of Muscles (SENIAM), a European initiative to standardise sEMG detection procedures, offers the following recommendations for the front-end amplifier:

High-pass filter corner frequency $\Rightarrow \leq 10$ Hz

Low-pass filter corner frequency $\Rightarrow \sim 500$ Hz

Input referred noise $\Rightarrow \leq 1 \mu\text{V}_{\text{RMS}}$ (10–500 Hz)

Input impedance (gel electrodes) $\Rightarrow \geq 100 \text{ M}\Omega$

Input impedance (dry electrodes) $\Rightarrow \geq 1000 \text{ M}\Omega$

Common Mode Rejection Ratio (CMRR) $\Rightarrow \geq 80$ dB

Gain \Rightarrow Suitable for adequate Analogue to Digital Converter (ADC) resolution (≥ 40 dB)

The recorded sEMG signal is a spatially localised recording of the muscles electrical activity using skin attached electrodes. Thus, sEMG amplitude, power spectrum and quality are highly dependent upon muscle geometry, electrode position, force levels and skin-electrode contact [4, 14–17]. As a result, electrode configuration and location can affect the reproducibility of measurements and is still a major barrier on clinical sEMG monitoring [4, 18–21]. Moreover, the electrode location with respect to the lateral edge of the muscle and detection area will determine the amount of crosstalk from nearby muscles that may be detected by the electrode. As the muscle flexes and extends, the distance between the MUs and the skin electrode constantly changes, while the muscle fibres also change length. Since the tissue between MUs and the skin acts as a volume conductor, the detected sEMG potential is inversely proportional to this distance (MU—skin distance). The volume conductor between the surface electrode and the excited MU acts as a spatial low-pass filter. Hence, MUs closer to the surface would generate a steeper response at higher frequencies than that of MUs located deeper in the muscle.

One or more electrodes can be used to record the sEMG signal at a single muscle location. A monopolar configuration uses a single electrode and a reference signal, usually the body ground, for detection and amplification. Although this

configuration contains the entire information available from the detected signal and does not have any impact on the recorded sEMG frequency, it is very sensitive to common mode signals, thus it is not suitable for real-life applications. A bipolar configuration uses two electrodes and is the most widely used configuration as it offers increased robustness to common mode variations [4]. Higher order electrode structures, such as the double differential configuration (three electrodes) or four electrode configurations (normal double differentiating configuration) can be used to limit the detection volume, reduce crosstalk and increase spatial selectivity. Furthermore, literature has shown that a reduction in the Interelectrode Distance (IED) can limit the detection volume of the electrodes [4].

4 Myoelectric Manifestations of Muscle Fatigue

Although we experience fatigue everyday, defining or quantifying muscle fatigue is not a simple task. Fatigue is interpreted as a feeling of weakening, muscle pain or performance decrease. However, these feelings are not suitable for measurement or quantification. Fatigue itself is not a physical value. Hence, quantifying muscle fatigue is proven to be difficult, as there is no universal index for it. The inability to maintain a certain muscle force, to perform a certain task or to generate the same level of Maximum Voluntary Contraction (MVC) force, sometimes associated with pain, serves as the clinical definition of localised muscle fatigue [22].

As a result, muscle fatigue is defined by measuring physical variables such as force, level of MVC, power produced upon contraction or angular velocity of a joint. Moreover, it can be defined by myoelectric (ME) variables such as MU firing rates, conduction velocity, muscle activation, sEMG amplitude and sEMG spectral estimates. For non-invasive, real-time and realistic applications measurement of fatigue is based on the analysis of sEMG and not on physical variables. Fatigue is better evaluated in time, showing progressive changes in the muscle during a contraction. Thus, fatigue is assessed from the beginning to the end of a muscle contraction [4].

In 1912 Piper was the first to observe a progressive “slowing” of the sEMG during voluntary static contractions [4, 23]. Due to the random nature of the sEMG, it is easier to quantify this spectral compression in the frequency domain. In addition, Cobb and Forbes noted an increase of the sEMG amplitude during static, sub-maximal, fatiguing contractions [4, 24]. These myoelectric changes precede the mechanical manifestations of muscle fatigue. Although there are several myoelectric manifestations of muscle fatigue, there are many difficulties for establishing a universal fatigue index. Firstly, the EMG signal varies in terms of amplitude and frequency between muscles (size and number of muscle fibres) and people. Secondly, some sEMG attributes depend on the force levels exerted by the muscles. For example, in sub-maximal effort, sEMG signal shifts to higher amplitudes and lower frequencies as the muscle fatigues. In maximal effort the EMG signal shifts to lower amplitude and again lower frequencies. Due to the force-sEMG relationship [25–28], a high sEMG amplitude is attributed to a strong contraction.

4.1 Muscle Fibre Conduction Velocity

As the muscle fatigues lactic acid and K^+ accumulate in the extracellular muscle space, impairing the conduction of action potentials across the muscle membrane, thus slowing down MUAPs [29]. Furthermore, fast-twitch muscle fibres dominate high force production but get easily fatigued and drop-out. Following that, fatigue resistant, slow-twitch muscle fibres take over. These have smaller amplitudes since they are located deeper inside the muscle and exhibit slower conduction velocities, associated with the observed spectral compression. MFCV is a measure of the travelling speed of MUAPs along their propagation in muscle tissue. MUAPs originate from the innervation zone of the muscle and will propagate to the tendon region. As the muscle fatigues, MFCV decreases while the sEMG power spectrum compresses to the left [30].

Lindstrom et al. developed a mathematical model shown in Eq. (6.1), linking MFCV with the observed EMG power spectrum [31]. In this equation, $P(f)$ is the EMG power spectrum, v is conduction velocity and $G\left(\frac{f}{v}\right)$ is the shape of the spectrum of the detected surface action potential. However, changes in the spectral content of the sEMG signal are disproportionately larger than decreases in MFCV. Furthermore, recoveries in frequency are more rapid than lactate removal in the muscle. Thus, MFCV can provide a more detailed insight on muscle fatigue and muscle recovery [32, 33]. One of the advantages of MFCV is that it is reliable under static and dynamic contractions [32].

$$P(f) = \frac{1}{v^2} G\left(\frac{f}{v}\right) \quad (6.1)$$

5 A Wearable MFM System Based on the Muscle Fibre Conduction Velocity

5.1 MFCV Tracking Algorithm

Cross-correlation is a robust and accurate measurement of similarity between signals, with a high degree of noise immunity [31]. Cross-correlation uses two signals and estimates the delay between them. The sEMG detection points must lie in the muscle fibre path, from the innervation zone to the myotendinous part of the muscle. “The innervation zone is a small region or band of muscle tissue wherein MUAPs originate and then propagate bidirectionally toward each tendon” [34]. Figure 6.4 shows the position of the innervation zone of the short head of the biceps brachii. It can be seen that the zone is close to the belly of the muscle. However, this observation does not hold true for every muscle in our body, hence the innervation point should be located for each muscle separately [35]. Figure 6.4 further illustrates the propagation of the MUAPs along the muscle fibre in two directions (up and down) [35].

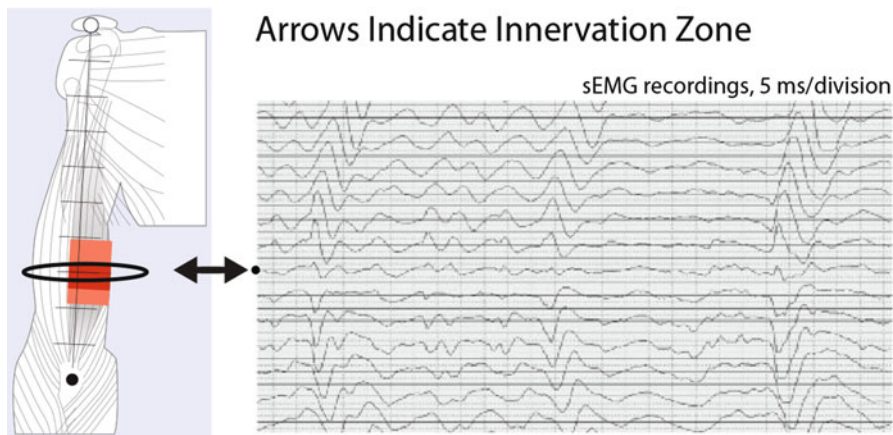


Fig. 6.4 Propagation of MUAPs along the muscle fibre in two directions, originating from the innervation zone. Figure from [35]

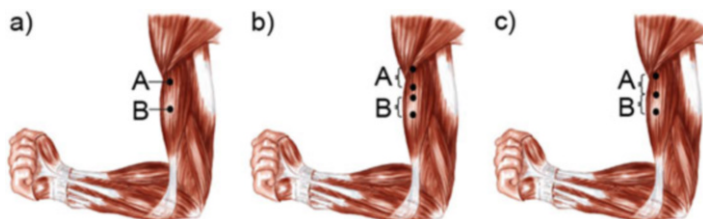


Fig. 6.5 Cross correlation detection points on biceps brachii muscle. (a) single electrode, monopolar detection, (b)–(c) double electrode, bipolar detection. Figure from [36]

The detection points **A**, **B** for the short head of the biceps brachii are shown in Fig. 6.5a. In this example the detection points are located above the innervation zone. Placing the detection points below the innervation zone would result in the same observations. Single differential (bipolar) detection was used. The resulting SD detection points are shown in Fig. 6.5b. It was observed experimentally that bringing the two electrode sets closer and sharing the middle electrode, as in Fig. 6.5c, would reduce the DC offset between the two recordings and at the same time improve the detection of the delay between the two points. A better choice may have been the use of double differential detection. Double differential detection would form a spatial filter and would provide a signal that better isolates the peaks and valleys of the superimposed MUAPs, which form the sEMG [37–39]. However, this method would require the use of two triplets of electrodes and more complicated front end amplification topologies. Single differential detection was selected in an effort to reduce electrode footprint, to maximise muscle applicability, and maintain system simplicity.

5.2 Design Considerations for the CMOS Cross-Correlator

Adding to the list of advantages in MFCV estimation, cross-correlation is a straight-forward operation which can be implemented using custom CMOS logic in Application Specific Integrated Circuits (ASICs). However, cross-correlation is a computationally intensive time-domain operation because of the large amount of multiplications required. Cross-correlation is described by the following equation, Eq. (6.2):

$$r_l = \sum_{i=1}^n x(i)y(i-l) \quad (6.2)$$

where l is the time shift between the two signals being correlated. For every point in time, all the samples (n) in the correlation window must be multiplied and accumulated. In order to evaluate the accuracy of the cross-correlator to be designed in CMOS, the MATLAB[®] `xcorr()` function was used. Simulations with a varying correlation window and sEMG sampling frequency were used to find the optimum parameters for the modelled cross-correlator. Simulations were conducted using real retrospective sEMG data from the biceps brachii muscle of a test subject, under static, non-fatiguing, contractions. All sEMG recording were made with the following parameters, unless stated otherwise; for the recording of data an Octal BioAmp ML138 by ADInstruments was used, along with pre-gelled electrodes provided by AMBU Neuroline (72000-S/25). The signal is bandpass filtered 10–500 Hz. The sampling frequency was set to a high value of 10 k samples/s to allow further post-processing (filtering, under-sampling) of the recordings. The Interelectrode Distance (IED) is 2 cm and the active electrode diameter is smaller than 10 mm.

Since the muscle contraction was a non-fatiguing one, the time delay between the two recorded channels was constant and subsequently removed. Following that, an artificial delay of 10 ms was introduced to one of the recordings to simulate an increase of the delay between the two signals, similar to a fatiguing contraction. The delay was introduced artificially in order to obtain a linear increase in the signal's 65 s duration, thus creating a highly controlled test signal. The sEMG signal has a bandwidth of 10–500 Hz. In theory, a 1 kHz sampling rate with a 200 ms correlation window would suffice for an accurate estimation of the time delay between the two channels.

Implementation of a cross-correlator that utilises all the bits of the digitised sEMG signal in an ASIC dedicated for muscle fatigue monitor would be a waste of resources, power and silicon area. The sEMG signals can be converted to a discrete signal with the aid of a single threshold. This work demonstrates that single threshold is an adequate alternative to digitising the complete sEMG signal, while retaining the necessary information for cross correlation and delay estimation. The novelty of this research work lies in eliminating the need of cross-correlating the whole sEMG signal, by cross-correlating a single bit approximation of it. Thus,

complicated cross-correlator architectures can be replaced by simple, bit-stream cross-correlator designs. Since a bit-stream consists of signal values of 1 (High) or 0 (Low), the complicated multiplications required for cross-correlation are replaced by AND gates.

A system based on this approach seems very promising. In the case of cross correlating two sEMG signals, the reference EMG signal (point **A**) is stored in one buffer and the delayed sEMG signal (point **B**) in another buffer. A single bit converter is used to digitise the analogue sEMG signals and convert them to two digital bit-streams. The type of converter used is a comparator. The signal from the second buffer is shifted one sample at the time, applying the XOR function over the selected window. The result is the sum of all the zeros computed by the logic function. In order to provide a comprehensible and representative error value between the two datasets, the Mean Absolute Relative Difference metric was used (MARD, in %). Table 6.1 presents the resulting MARD between the *xcorr()* function and the linear fit under different correlation windows and sEMG sampling frequencies. A MARD value of 5% or less was deemed satisfactory for the intended application. The optimum operating region of a cross-correlator used in MFCV estimation is highlighted in green. Relaxing the operating parameters would result to a higher MARD value, highlighted in orange. As seen in Table 6.1, the accuracy of estimation is increasing with sampling frequency. This is to be expected, since there are more data samples available to locate an accurate time delay. With a sampling

Table 6.1 Error results between *xcorr()* and the linear fit of the added time delay between the two resulting bit-streams

Correlation window (s)	Sampling frequency (kHz)							
	1	2	3	4	5	6	7	8
0.1	347.15	266.14	109.34	47.26	21.38	10.71	14.72	6.38
0.2	441.93	112.31	28.91	16.38	9.46	7.45	6.51	5.29
0.3	250.85	43.76	24.03	15.04	8.63	8.40	5.87	3.74
0.4	180.13	39.26	21.55	14.66	8.66	6.17	4.80	4.09
0.5	144.55	34.47	18.36	12.86	9.35	5.98	4.22	3.23
0.6	122.71	33.39	18.52	12.61	8.70	5.81	3.41	2.72
0.7	106.43	34.25	19.18	12.12	8.84	6.04	2.73	2.27
0.8	86.58	32.65	18.06	12.77	9.18	3.68	2.83	2.64
0.9	88.20	33.27	18.82	12.62	8.01	4.13	2.94	2.22
1	85.34	31.50	17.85	12.08	7.17	3.83	2.11	1.57
1.1	80.12	29.69	18.74	12.83	7.66	3.77	1.97	2.36
1.2	83.68	31.66	18.68	10.92	7.20	2.85	2.16	1.76
1.3	87.49	32.12	18.22	10.70	4.20	3.16	1.89	2.56
1.4	86.42	30.04	16.79	10.99	4.86	3.39	2.43	2.02
1.5	78.31	32.37	17.79	10.01	4.90	2.56	1.78	5.02
	Minimum absolute relative difference (%)							

Correlation windows and sEMG sampling frequency are varied to obtain optimum working region

frequency of 1 kHz and a maximum detectable delay of 10 ms, there are only 10 available data samples for the estimation. However, this number increases to 80 using a sampling frequency of 8 kHz. Moreover, accuracy is increasing with the correlation window up to the value of 1.3 s and then decreases again, resulting to an optimum window around 1 s.

5.3 Proposed Bit-Stream Cross-Correlator

The aim of the MFCV cross-correlator is to estimate the time delay between the two bit-stream signals $x(n)$ and $x_d(n)$. $x(n)$ is the digitised EMG signal detected at point **A** (see Fig. 6.5b) and $x_d(n)$ is a delayed version of $x(n)$ detected at point **B**. Thus, there is no need to buffer $x_d(n)$ for cross-correlation since it is already shifted in time. Instead of storing and shifting data to find a time lag, it is possible to continuously return the similarity of the two signals for a given time lag by counting all the time instances where the two signals are the same. The proposed bit-stream architecture is similar to the operation of correlator banks [40]. Several discrete time lags for the cross-correlation output are obtained by continuously delaying the input signal and repeating the aforementioned process. Thus, obtaining a similarity (correlation) result for every discrete time delay. Finally, the time lag between the two signals is returned by the counter with the larger value. This approach greatly simplifies the number of transistors required. However, the time lag resolution is limited. The time lag dynamic range depends on the number of time delays introduced to the system. Furthermore, the resolution of the system can be varied according to the sampling frequency ($T = \text{Delay}$) and the cross-correlation time window. Consider Fig. 6.6 for the proposed bit-stream correlator which comprises of several correlation stages, where each stage has a delay block, a counter and a gate responsible for the cross-correlation. The delay line is constructed by connecting serially several “unit delays” (discrete time-lags). The two bit-streams are cross-correlated using an XNOR function and an accumulator at the output of every “unit delay”. Each accumulator increases its current value by 1 when $x_d(n)$ matches the corresponding time-lag D_n . When the reference signal $x(n)$ reaches the incremental delay which matches the time delay between $x(n)$ and $x_d(n)$, then the accumulator value corresponding to that incremental delay will be the maximum out of every accumulator along the delay path. Thus, the computed delay as a result of cross correlating the two signals is given by Eq. (6.3) where n is the number of discrete time-lag, C is the accumulator value and i is the time of evaluation.

$$D_{n,i} = \max(C_1, C_2 \dots C_n) \Big|_{t=i} \quad (6.3)$$

It is observed that in the proposed architecture the buffer size of $x(n)$ in [41] is reduced to the time delay between the two signals. In this approach, buffering of

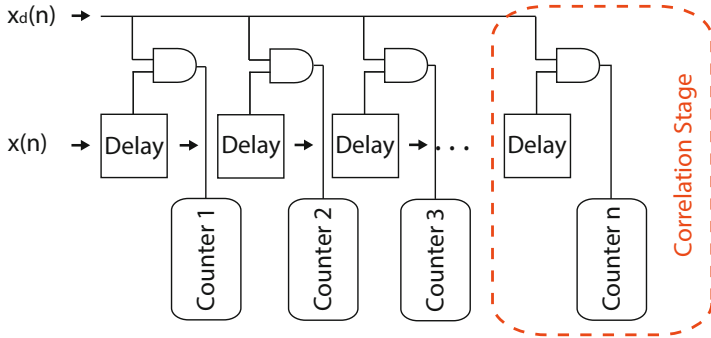


Fig. 6.6 Proposed architecture of the bit-stream cross correlator. The buffering blocks have been replaced with delay blocks

the two signals is perceived in time. Instead of saving the incoming data, they are processed in real-time for the duration of a specified time window. Periodic reset of the accumulators is necessary in order to obtain a new cross correlation result. The disadvantage of this architecture is the loss of temporal accuracy. However, there is a trade-off between temporal and delay-estimation accuracy. The longer the reset period, the more data points the system accumulates. Thus, time-averaging is performed, resulting to a more indicative estimation of the constantly varying delay between the two signals. The advantages of these approach are the following. There is a significant reduction of the number of transistors needed. This will have a positive impact on power consumption and silicon area. Furthermore, the computational complexity of the correlator is reduced to a minimum.

In order to decide the number of correlation stages to be added to the system, the following assumptions were made. The longest delay would occur at the end of a fatiguing contraction. Physiological values reported in literature for a fatiguing isometric contraction indicate that MFCV very rarely is measured to be less than 3 m/s [42–47]. Referring back to Table 6.1, a sampling (shifting) frequency of 6 kHz was selected in order to achieve a good accuracy (low MARD values) while keeping the required delay buffer length to a minimum. Thus, the delay buffer was designed with a length of 40 correlation-stages. The correlation window is not influenced by the buffer length. Hence, the trade-off between buffer length and correlation window is swapped with the trade-off between maximum detection window and sampling frequency. However, increasing the detection window in the proposed bit-stream cross-correlator is less costly than increasing the correlation window in the architecture proposed in [41].

The dynamic range of the system can be varied according to the sampling frequency and the delay estimation accuracy according to the cross-correlation time window. In order to characterise the system, it was modelled in SIMULINK and compared with MATLAB[®] *xcorr()* function using real sEMG data from the biceps brachii muscle of a test subject during a static, fatiguing contraction. The sEMG signal was amplified to a peak-to-peak voltage of 300 mV. Figure 6.7a shows the

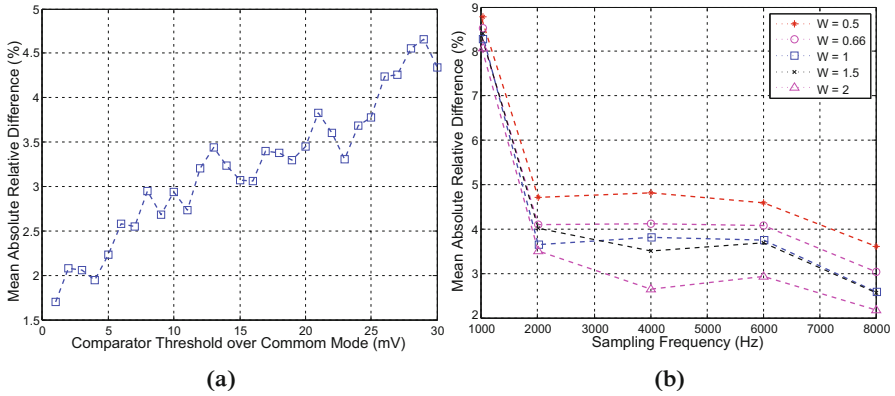


Fig. 6.7 Performance of the proposed bit-stream cross-correlator with varying threshold, correlation window and sampling frequency: (a) MARD between MATLAB[®] *xcorr()* function and modelled system with a varying threshold, (b) MARD between MATLAB[®] *xcorr()* function and modelled system, where W is the correlation window (s)

MARD between the modelled system and MATLAB[®] simulations for a varying comparator threshold. It can be seen that the relative error has a small sensitivity to the chosen threshold level. The modelled system's output is compared to the results of the *xcorr()* function. By looking into Table 6.1, a window of 1 s was used in the *xcorr()* function as it exhibited good accuracy across a wide range of sampling frequencies. Figure 6.7b shows a parametric simulation with the sampling frequency and correlation window as variables. Increasing the sampling frequency results to higher accuracy but limits the dynamic range of the system. The results show that sEMG bit-stream cross-correlation is an accurate representation of cross-correlation which allows it to use for tracking MFCV. A longer correlation window would yield better delay estimation accuracy but would result to a slowly tracking system. The proposed approach of this bit-stream cross-correlator resembles the basic architecture of a time-to-digital converter. According to [48], the measurement interval ΔT in a time-to-digital converter can be expressed as in Eq. (6.4), where T_c is the period of the clock. The error in the measured time interval ΔT is ϵ_r and can be equal to twice the clock period. Thus, a delay chain length of 40 yields a relative error of 5%, which is considered acceptable for the intended application.

$$\begin{aligned} \Delta T &= NT_c + \epsilon_r \\ \epsilon_r &\in [-T_c; +T_c] \end{aligned} \quad (6.4)$$

5.3.1 Correlation Stage

As mentioned in the section above, the bit-stream correlator is made of several correlation stages connected in series. Each stage comprises of a delay block,

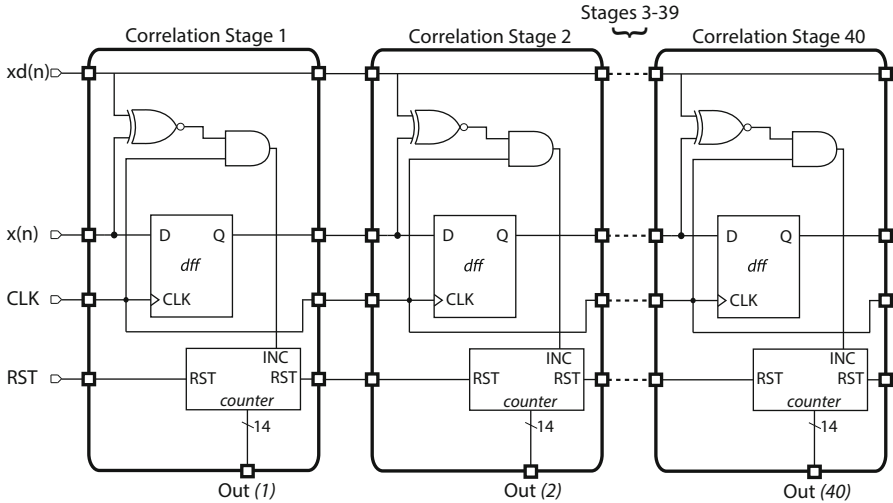


Fig. 6.8 Correlation stage with delay block, counter and correlator (XNOR)

a counter and a gate responsible for the cross-correlation. The cross-correlator proposed by Lande et al. utilised an inverter chain as a buffer. However, it was designed to operate at a much higher frequency than the one required for this application. As such, an inverter chain would be too long, increasing silicon area and power consumption. A simple D-type Flip Flop (DFF) acts as a delay block, and the delay time is controlled by the sampling frequency of the system (CLK). A 14 bit ripple counter was found to be enough to meet the requirements of the system. The use of XNOR gates as a bit correlator improves the original AND gate design by taking all possible digital cases into consideration. The circuit schematic is shown in Fig. 6.8. A second clock (RST) with a much lower frequency than CLK is used to reset the counters and thus define the cross-correlation time window. The dynamic range of the system can be varied according to the sampling frequency CLK and the delay estimation accuracy according to the cross-correlation time window (RST).

5.3.2 Maximum Detector Stage

At the end of the correlation time window (RST), all the counters of the system are read. The correlation stage (i.e. delay) of the counter with the maximum value best represents the time lag between the two input signals. A maximum function compares all the counters in cycles. The maximum function comprises of smaller blocks, each comparing two 14 bit numbers, and operates using sequential logic. It starts by comparing all the results in pairs and then proceeds with evaluating the results of the last comparison. Every maximum operation returns a binary flag, which passes down to the next comparison and indicates which of the two compared

numbers is the maximum. Thus, a binary one means the first of two numbers is bigger. This way, the counter position number (hence delay number) and not the counter value is returned when the operation is finished. Consider Eq. (6.5) as an example for calculating the counter path. The *max* function is enabled only prior to the counter evaluation stage, thus saving power. All the counters and delay blocks stop processing, in order to allow the *max* function to settle to one output (sequential logic). Then, the system output is ready to read.

$$\begin{aligned}
 \text{out}(1) &= f_{38} \wedge f_{37} \wedge f_{35} \wedge f_{30} \wedge f_{20} \wedge f_0 \\
 \text{out}(2) &= f_{38} \wedge f_{37} \wedge f_{35} \wedge f_{30} \wedge f_{20} \wedge \bar{f}_0 \\
 \text{out}(3) &= f_{38} \wedge f_{37} \wedge f_{35} \wedge f_{30} \wedge \bar{f}_{20} \wedge f_1 \\
 &\vdots \\
 \text{out}(8) &= f_{38} \wedge f_{37} \wedge f_{35} \wedge \bar{f}_{30} \wedge \bar{f}_{21} \wedge \bar{f}_3 \\
 &\vdots
 \end{aligned}
 \tag{6.5}$$

6 MFCV Tracking ASIC Architecture

The developed MFCV ASIC shown in Fig. 6.9 consists of four major building blocks: (a) a sEMG dual channel Instrumentation Amplifier (IA), (b) two Sallen Key low-pass filters, (c) a bit-stream converter comprised by two analogue comparators and (d) a digital bit-stream cross-correlator. In addition, the MFCV ASIC includes a bias circuit generator, a digital timing control circuit and an SPI to communicate with any receiving circuit.

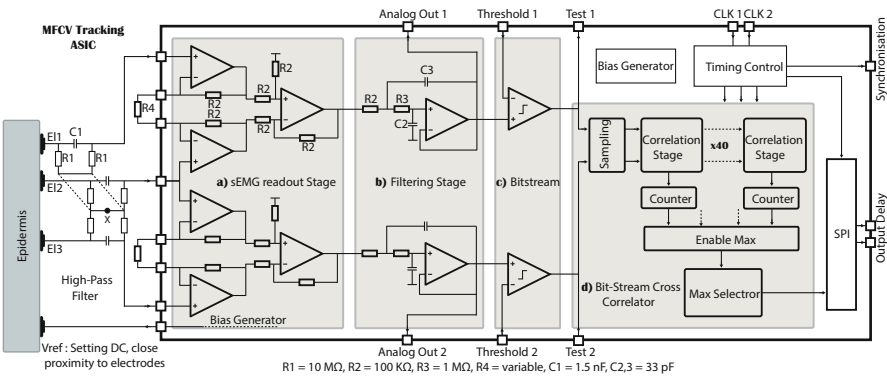


Fig. 6.9 System block architecture of the muscle fibre conduction velocity tracking ASIC

The operation of the MFCV ASIC can be described as follows: the dual IA amplifies the detected sEMG signals which are then low-pass filtered to extract the signal attributes in the required frequency band (10–500 Hz). Following that, the bit-stream converter digitises the sEMG signals and feeds them to the bit-stream cross-correlator that computes the time delay between them.

6.1 sEMG Signal Processing

The circuit has been implemented in a commercially available 0.35 μm CMOS technology provided by AMS (C35B4). Standard cell libraries were used for the implementation of the operational amplifiers and comparators in stages **a**, **b** and **c** (OP05B, COMP).

A common front-end architecture in biopotential measurements is a DC-coupled fully differential amplifier followed by a difference amplifier, as in the classical three opamp instrumentation amplifier [49–54]. The equivalent input noise only depends on the two opamps constituting the fully differential amplifier. Instrumentation amplifiers should be capable of rejecting up to 300 mV DC Polarisation Voltage (PV) from the biopotential electrodes [54, 55]. Thus, AC coupling is needed to avoid saturating the output due to input DC offsets. Existing architectures that can implement rejection of such high DC voltages without using off-the-shelf components lead to lower performance IAs [56, 57] or suffer from trade-off's resulting in reduced PV [58–60]. On the other hand, the use of conventional, simple, off-chip high-pass filters significantly reduces the input impedance, where especially the common mode input impedance is very important for achieving high Common Mode Rejection Ratio (CMRR).

Thus, an external floating high-pass filter is used, as shown in Fig. 6.9. The intended application of the ASIC as a wearable device allows the use of external components. The main advantage of floating high-pass filter compared to conventional passive high-pass filters is the elimination of the grounded resistor, implementing very large common mode input impedance [54, 55]. With $R1 = 10 \text{ M}\Omega$ and $C1 = 1.5 \text{ nF}$ the resulting cutoff is 10.6 Hz to best filter out motion artefacts [61]. However, this filter structure requires a fourth electrode to bias the input filter structure and set the DC voltage of the body around the electrodes. This reference DC level must satisfy the amplifier's input common mode range requirements and is set at a voltage $V_{\text{ref}} = V_{\text{supply}}/2$. The common mode voltage of the detected sEMG signal feeding the amplifiers is the averaged DC electrodes (1–3) voltage at node **X** (Fig. 6.9). Since the input network is not grounded, when a common mode input voltage is applied, no currents flow through the network.

The low-pass filter of Sallen Key topology is implemented with a cutoff of 2.5 kHz, with $R2 = 100 \text{ k}\Omega$, $R3 = 1 \text{ M}\Omega$ and $C2, C3 = 33 \text{ pF}$. The structure is duplicated to allow two channel operation. The reference voltages of the two comparators are kept separate to allow offset mismatch compensation. Furthermore, the feedback resistors responsible for the amplifier gain ($R4$) were not implemented on the IC, but left to be completed with external components to allow gain flexibility.

7 MFCV Tracking ASIC Experimental Results

The MFCV tracking ASIC micro-photograph is shown in Fig. 6.10. The layout area for the digital core is $1122\ \mu\text{m}$ by $494\ \mu\text{m}$ and for the analogue front end is $800\ \mu\text{m}$ by $500\ \mu\text{m}$, with a total area of $1\ \text{mm}^2$. The total static and dynamic power consumption of the digital core (clocked at $10\ \text{kHz}$) is $1\ \mu\text{W}$ and of the analogue front end is $2.071\ \text{mW}$ from a $3.3\ \text{V}$ supply. The power consumption of the analogue front end includes the power of two IAs, two high-pass filters and two bit-stream converters (Table 6.2).

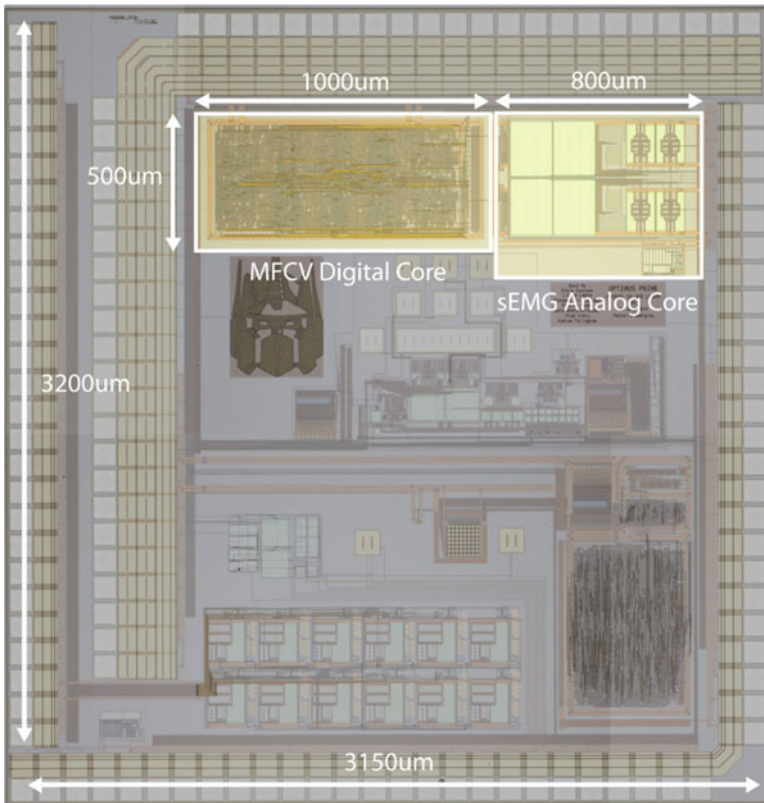


Fig. 6.10 Die micro-photograph of the MFCV tracking ASIC

Table 6.2 MFCV ASIC parameters, size and power breakdown (C35, $3.3\ \text{V}$)

Core	Measured current	Size
Analogue	$616\ \mu\text{A}$	$0.4\ \text{mm}^2$
Digital	$290\ \text{nA}$	$0.554\ \text{mm}^2$
Biasing	$11.5\ \mu\text{A}$	–
Total power	$2.071\ \text{mW}$	–

7.1 System Validation

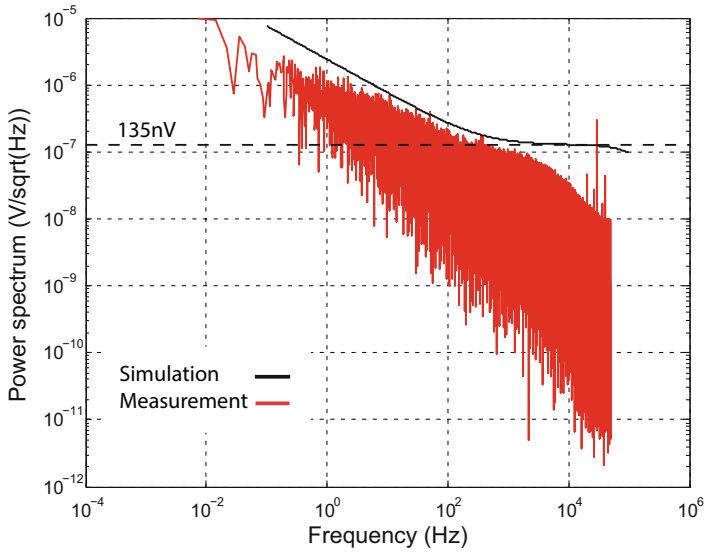
The characterisation of the IA is performed at the output of the sEMG readout stage. Figure 6.11a shows that the measured integrated input-referred noise of the IA (0.1–2.5 kHz) is $3.579 \mu\text{V}$ with a noise floor at $135 \text{ nV}/\sqrt{\text{Hz}}$. The sEMG has a range of $50 \mu\text{V}$ – 5 mV . Thus, the presented front end biopotential amplifier is more than adequate to amplify sEMG signals for the proposed application. Figure 6.11b shows the measured differential and common-mode gain of the IA followed by the high-pass filter. The external floating low pass filter was included during measurements. The resulting CMRR is 85.24 dB in the bandwidth of interest (83.1 dB at 50 Hz). A segment of the detected and amplified sEMG signals is shown in Fig. 6.12. The delay between the two signals can be observed visually. However, the cross-correlator will perform this task more accurately.

7.2 Validation on Human Subjects

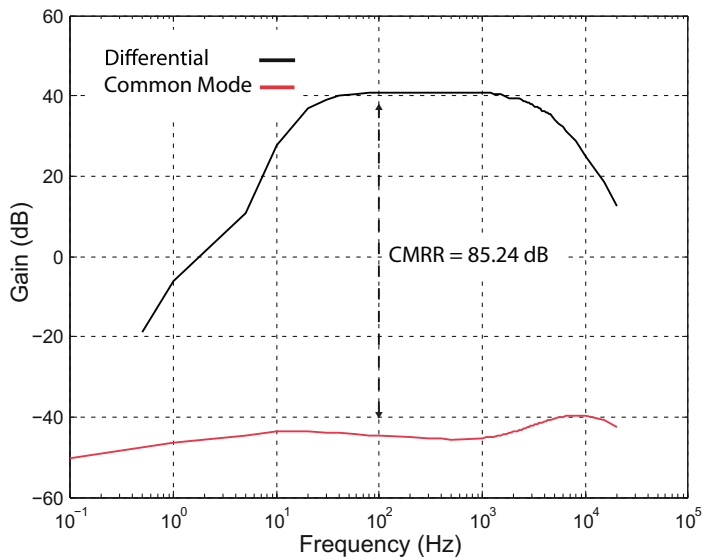
Twenty (20) volunteers (4 female, 16 male) were used to evaluate the performance of the MFCV tracking ASIC. This work has been approved by the Imperial College Joint Research Compliance Office ethics committee (ICREC ref: 15IC2481).

7.2.1 Experimental Protocol

Subjects were seated or standing, depending on their preference, with the back and elbow fixed against a wall in order to minimise compensatory movements. Subjects were required to keep in contact with the wall throughout the whole duration of the testing. Consider Fig. 6.13 for a graphical representation of the experimental setup. The subject specific calibration of EMG sensors involved the performance of five Maximum Voluntary Contractions (MVCs). Following that, the subjects were asked to perform isometric contractions by pulling against a handle attached to an electronic scale. The bottom end of the scale was attached to the ground. Subjects were required to sustain a constant force, pre-set at 70% of their MVC force, for as long as they deemed possible. After the force reading dropped more than 10% of the pre-set level, the experiment would stop. A normal luggage scale was used as a force monitor. The scale was modified by adding an instrumentation amplifier to the scale's resistor bridge network that monitors force. The voltage output of the modified scale was sampled by the same ADC that sampled the two amplified sEMG outputs of the MFCV ASIC. A visual representation of the bit-stream allowed for correct threshold positioning. Furthermore, a detailed and real-time plot of the force output with visible limits helped the subjects monitor their force output and keep it steady between the acceptable limits. The force boundaries were adjusted for each subject according to their MVC levels. Four disposable surface EMG electrodes were placed on the skin of the participant's arm to monitor activity of the biceps



(a)



(b)

Fig. 6.11 (a) Input-referred noise measurement (*red*) and simulation (*black*) results of the IA channel 1, (b) Common-mode rejection ratio measurement of the IA (chan. 1) using a 500 mV_{pp} input signal

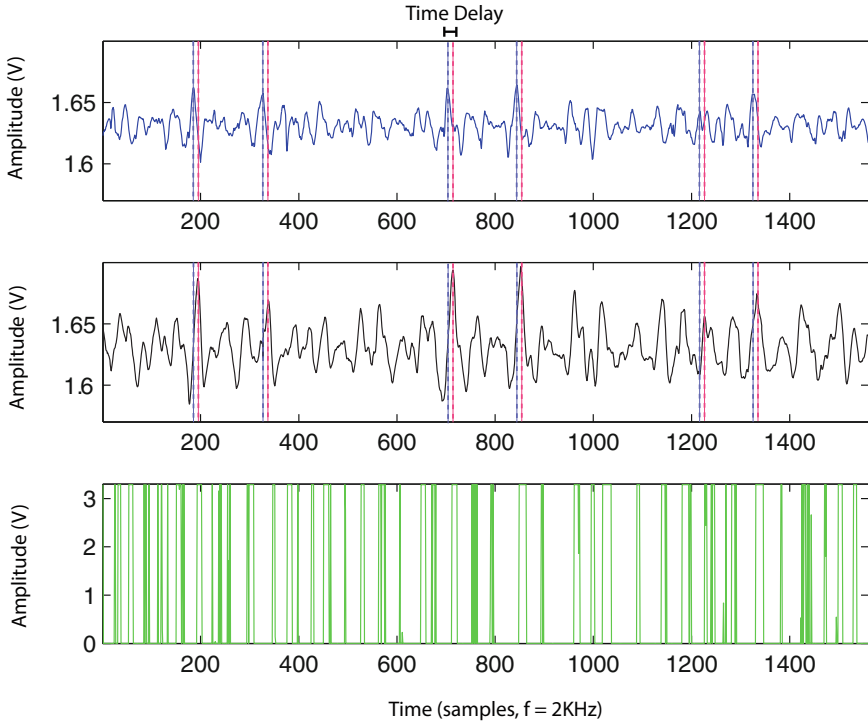


Fig. 6.12 *Top*: sMEG segment from Channel 1, *middle*: sMEG segment from Channel 2. Delay between the two channels is easily detectable by eye, *bottom*: Comparator output from channel 2 for reference

brachii muscle, according to Fig. 6.14. Prior to electrode placement, the skin was prepared by cleaning it with medical alcohol. A detailed architectural diagram and a photograph of the experimental setup are also shown in Fig. 6.14.

A requirement of the system is the positioning of the two independent thresholds for the bit-stream converters. The thresholds could be placed at the common-mode voltage, as in the analysis of the modelled system. However, in order to avoid noise in the absence of sEMG to contribute to MFCV estimation, the thresholds are placed just above the noise floor. The threshold levels should be low enough to be closer to the base of the sEMG MUAP spike and at the same time be above the peak noise amplitude. This could also be achieved with the use of a comparator with hysteresis.

7.2.2 Results from Human Trials

The accuracy of the ASIC is established by a direct comparison between the ASIC and MATLAB[®]. The amplified sEMG signal from the ASIC is recorded using a 16-bit ADC by ADInstruments. Following that, the MFCV is computed in MATLAB[®] using the *xcorr()* function and then compared to the ASIC estimate.

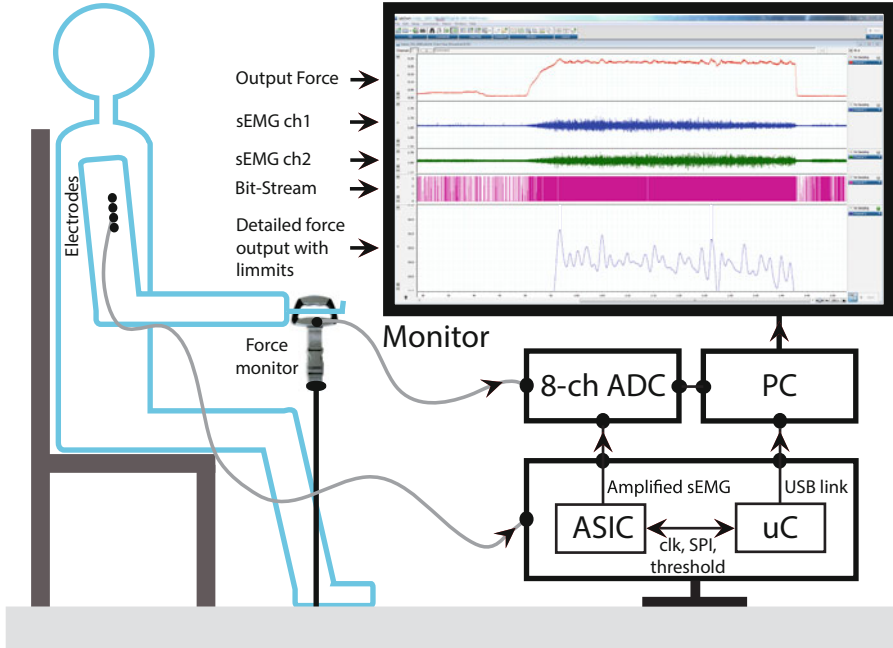


Fig. 6.13 Experimental setup protocol. The subjects keep the force output constant and within limits with the aid of a visual, real-time representation of the applied force

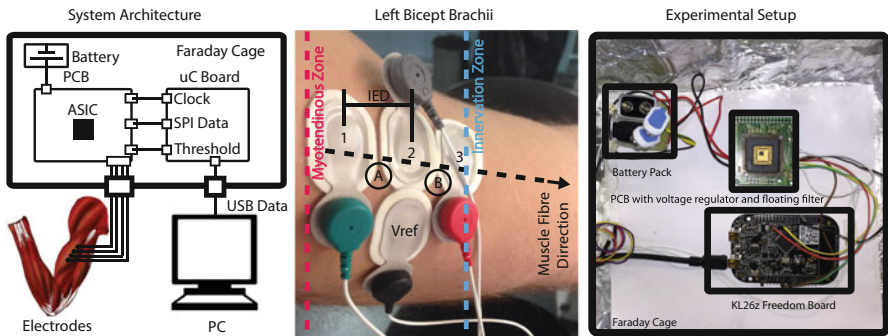
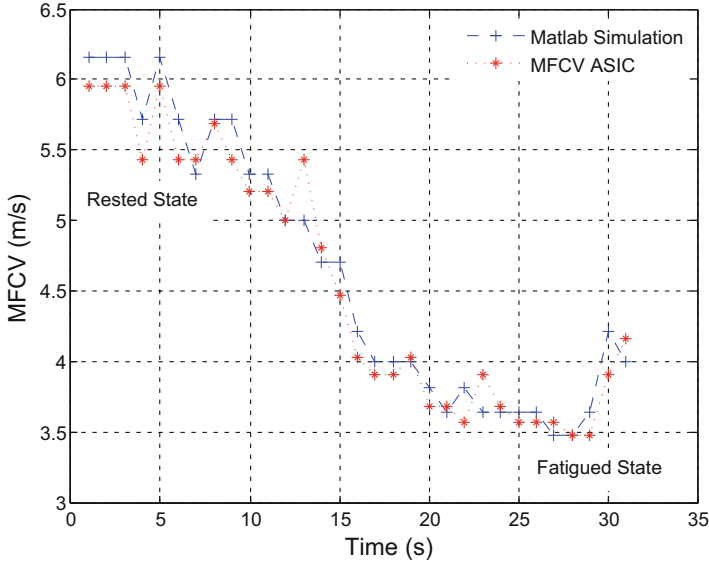
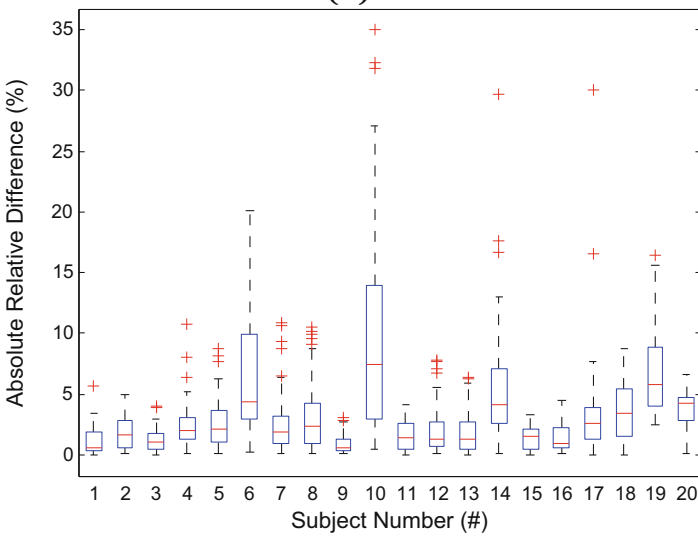


Fig. 6.14 *Left*: Architectural diagram of experimental setup. *Middle*: Electrode configuration for bipolar single differential sEMG amplification. Muscle zones are displayed. *Right*: Experimental setup including microcontroller for data transmission, clock generation and threshold position. Faraday cage used in noise measurements only

An example of the estimated MFCV by Matlab and that measured by the ASIC from the same subjects is presented in Fig. 6.15a. The subject starts the static contraction from a rested state, with an MFCV of 6 m/s. As the subject gets tired, the



(a)



(b)

Fig. 6.15 (a) MFCV trial results from subject 11. The rested state MFCV is 6 m/s and progresses to 4 m/s as the muscle fatigues, (b) Fatigue trial results comparison between MFCV tracking ASIC and MATLAB®. The MARD is presented in percentile

Table 6.3 Subject trial results: (A) subject number, (B) change in MFCV from chip, (C) change in MFCV from MATLAB®, (D) MARD and (E) MVC force

(A) #	1	2	3	4	5	6	7	8	9	10
(B) (m/s)	1.13	1.40	2.38	2.13	1.59	3.01	1.64	4.11	0.88	4.80
(C) (m/s)	1.22	1.71	2.66	3.28	1.69	3.48	1.32	2.93	1.01	5.47
(D) %	1.12	1.77	1.26	2.56	2.67	6.47	2.37	3.04	0.93	9.64
(E) Kg	19.5	23.0	23.5	23.5	13.2	23.5	15.0	13.5	20.5	16.0
(A) #	11	12	13	14	15	16	17	18	19	20
(B) (m/s)	1.78	5.72	2.6	20.67	1.32	2.13	0.48	1.09	3.35	2.28
(C) (m/s)	2.14	5.73	2.85	18.41	1.44	2.27	2.15	0.99	3.54	2.35
(D) %	1.60	2.08	1.86	5.38	1.34	1.39	3.11	3.63	6.78	3.70
(E) Kg	28.5	27.6	25.4	24.6	36.0	32.0	19.0	34.0	35.0	17.6

Table 6.4 Comparison between MFCV physiological values

Literature	[42]	[43]	[44]	[45]	[46]	[47]	This work
Mean/range	5.1	4.2	2.6	4.64	2.53	4.4	5.4
MFCV (m/s)			-5.3		-4.87		

MFCV drops to lower values, until the point that the subject is completely fatigued and stops the experiment. The MFCV has reached a new lower value of 4 m/s.

Table 6.3 presents the change in MFCV from a rested state to a complete fatigue state of the muscle for all 20 subjects. The estimated values of MFCV decrease in all subjects during fatigue, as seen in Table 6.3 (B), (C) by the relative change in MFCV from the ASIC and MATLAB®, respectively. Although MFCV is not always monotonically decreasing, the end point is at a lower value than the starting point in all subjects. The relative error between MATLAB® and the ASIC is estimated across the complete MFCV trends and not only on the relative change of MFCV and is shown in Table 6.3 (D). Time lag is converted to velocity using Eq. (6.6), where $CLK1$ is the sampling frequency, $OutputDelay_{\#}$ is the ASIC delay estimate between 1–40 and IED is the Inter Electrode Distance (Fig. 6.14, IED = 2 cm). The estimated MFCV values match the physiological values reported in literature for a fatiguing isometric contraction [42–47]. The reported MFCV values are compared with the MFCV values seen in this study in Table 6.4.

The correlation window was set to 1 s ($CLK2 = 1$ Hz) to maximise the accuracy of the bit-stream cross-correlator. The sampling/delay clock was set to 6.25 kHz to establish a wide MFCV dynamic range. The same parameters were imported in MATLAB® for analysis. The ASIC is found to have a Mean Absolute Relative Error (MARD) of 3.2%. Subject specific MARD results are presented in Table 6.3. The two MFCV trends (MFCV: MATLAB®, MFCV: ASIC) are compared point by point and the distribution of comparison data is displayed in a box plot in Fig. 6.15b.

$$CV \text{ (m/s)} = \frac{IED \cdot CLK1}{Output \text{ Delay}_{\#}} \quad (6.6)$$

Table 6.5 Summary of MFCV tracking ASIC

Technology	0.35 μm 2P4M
Circuit size	0.954 mm^2
Voltage supply	3.3 V
Current consumption	Total: 628 μA
	Bit-stream cross-correlator: 290 nA
	Biasing circuits: 11.5 μA
	IA: 180 μA
	Low-pass filter: 60 μA
	Bit-stream converter: 68 μA
CMRR	85.24 dB
Gain	Adjustable (set to 111)
Cut-off frequencies	10.6 Hz, 2.5 kHz internal
Input referred noise density	135 nV/ $\sqrt{\text{Hz}}$
Correlation window	Adjustable by $CLK2$
Estimation accuracy	Dependable on $CLK1$
MFCV dynamic range	Dependable on IED and $CLK1$
Number of delays	40
MARD error	3.2%

7.3 Discussion

The ASIC is capable of estimating MFCV with a mean relative error of 3.2% compared to MATLAB[®] analysis. A large proportion of the error can be attributed to the nature of the bit-stream cross-correlator, as it resembles a time-to-digital converter. Hence, some error arises by quantising in time with limited resolution, as the bit stream cross-correlator has a total of 40 quantisation values. Furthermore, misalignment between ASIC output data and MATLAB[®] estimations gives rise to unexpected error. Nevertheless, the system provides good estimation accuracy which is suitable for the purposes of muscle fatigue monitoring. Increasing the number of delays (above 40) will yield a wider dynamic range and allow higher sampling clock rate thus achieving higher estimation accuracy. This however will have a direct impact on the ASIC size. The systems performance specifications are summarised in Table 6.5.

8 Wearable Muscle Fatigue Monitor

The MFCV tracking ASIC was integrated into a miniaturised PCB bearing a microcontroller in an effort to design a complete monitoring platform for muscle fatigue. The board featured BLE communication and was controlled by a custom designed Android application. Following that, a custom 3D printed case was

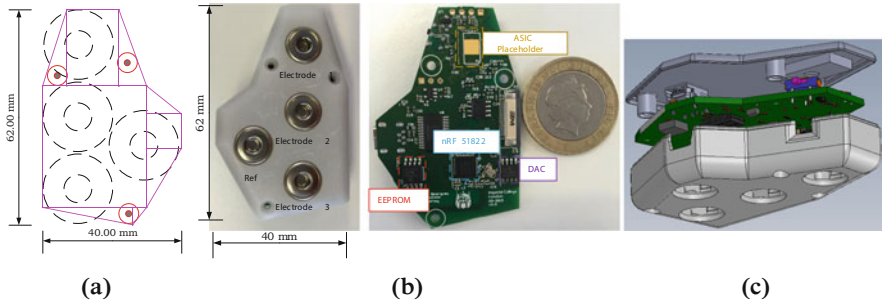


Fig. 6.16 (a) Minimum area for 4 electrode configuration. (b) Bottom side of the case with electrode snap sockets and completed PCB with soldered components. (c) Exploded view of the case with the PCB inside. The device is attached on the user by the electrode adhesive

developed to hold the electronics and interface them with the electrodes. This effort manifested to the creation of the first wearable device dedicated to muscle fatigue tracking.

The wearable node was designed to interface with the commercially available, low cost AMBU Neuroline (72000-S/25) self-adhesive, gel electrodes. The chosen electrodes are widely available and since they are self-adhesive, they would be responsible for attaching and holding the wearable device in place. First, the four electrodes (3 for sensing, 1 for biasing) were arranged in a pattern that minimises their footprint. This area would define the maximum dimensions for the wearable device, as seen in Fig. 6.16a. In order for the electrodes to be interchangeable, an electrode attaching mechanism with snap sockets was designed and 3D printed. The distance between each electrode is 2 cm. As the ASIC requires one electrode to be the furthest from the innervation zone, the wearable contains a slide switch that swaps the internal connections of the two furthest electrodes. Consequently, the wearable can be used in any orientation on the human body, allowing it to be worn in a more comfortable position. The developed PCB holds the following components: the MFCV tracking ASIC, power managements circuits (regulator, battery charger), USB UART communication controller, on-board Electrically Erasable Programmable Read-Only Memory (EEPROM) and a Nordic Semiconductor nRF51822 2.4 GHz Bluetooth Low Energy System-on-Chip (SoC). If the wearable is not connected to a smartphone, the MFCV values from the ASIC are stored in the EEPROM. The wearable is powered by a 3.7 V Lithium-Polymer rechargeable battery, with a capacity of 300 mAh and has a battery life of 27 h. The miniaturized PCB with the components soldered on top can be seen in Fig. 6.16b. The Android application provides the user interface to the wearable (Fig. 6.17). Table 6.6 summarises the specifications of the developed muscle fatigue tracking wearable device.



Fig. 6.17 (a) Wearable muscle fatigue tracking system, showing the Android application and the wearable device, (b) Complete muscle fatigue tracking system, showing the Android application and the wearable device attached on a user

Table 6.6 Summary of muscle fatigue tracking wearable device

Power consumption during EMG/fatigue stream	41.48 mW
Battery life for continuous EMG/fatigue streaming	27 h
Power consumption with BLE continuously inactive	17.43 mW
Battery life with BLE inactive	64 h
10-bit ADC LSB	1.172 mV
EEPROM fatigue data capacity	130,816 values

9 Conclusion

This chapter provided an overview of the mechanics of muscle contraction, sEMG formation and their defining role in the interpretation and analysis of muscle fatigue. Although the sEMG manifestations of muscle fatigue have been thoroughly explored, the physiological changes the muscle undergoes during a fatiguing effort are highly complex and not fully understood yet. Part of the difficulty lies on the nature of sEMG, as it is the superposition of all the underlying motor units. Interestingly, the sEMG signal carries a lot of information about the muscle. Several sEMG attributes change progressively as the muscle fatigues. Currently, a gap exists between sEMG signal processing techniques for fatigue analysis and methods of sEMG collection and processing. Custom signal processing methods adapted for fatigue monitoring and implemented in low power CMOS design would bridge that gap.

Furthermore, a first of its kind MFCV tracking ASIC was presented along with its detailed implementation and operation. The MFCV ASIC utilises a novel bit stream approach that greatly simplifies the sEMG signal without any loss of information, thus reducing computational complexity and minimising power consumption. The developed ASIC is capable of tracking MFCV from surface EMG signals using the method of cross-correlation. The system consists of a sEMG instrumentation

amplifier, a filtering stage, a bit-stream converter and a bit-stream cross-correlator. A test group of 20 people was used to demonstrate the ability of the proposed bit-stream correlator to accurately estimate MFCV in real time during static fatiguing contractions. The ASIC has a MARD error of 3.2% compared to MATLAB[®] analysis for the same dataset. The system draws 628 μA from a 3.3 V power supply and is implemented in a commercially available 0.35 μm CMOS technology. Hence this is the first working and validated System-on-Chip for muscle fatigue monitoring. The ASIC was embedded in a Muscle Fatigue Monitoring wearable device capable of streaming live sEMG and muscle fatigue data wirelessly to a user controlled smart-phone application using low power Bluetooth technology. The wearable node is lightweight and attached to the skin through the self-adhesive, pre-gelled, commercial surface electrodes.

In conclusion, a novel bit stream approach was developed that greatly simplifies the sEMG signal without any loss of information, thus reducing computational complexity and minimising power consumption. The developed technology was evaluated in a clinical study with a set of healthy individuals, which confirmed the accuracy and efficiency of the developed technology. Finally, the microchip was embedded in a wireless node operating with commercially available wet electrodes, introducing the use of custom Application Specific Integrated Circuits (ASICs) in wearable electronics for unsupervised muscle fatigue monitoring. There is an increasing trend to develop wearable electronics, however, these lack integrated processing and consist of sensing components such as electrodes. In addition, to date biosensors are regarded as stand-alone measuring devices. However, distributed and coordinated body sensing can substantially increase the amount of available information and in turn, biofeedback effectiveness. Thus, integrating multiple “smart” (autonomous processing), centrally controlled EMG monitoring nodes in a wearable yet unobtrusive platform (such as a “smart” clothing), has the potential to revolutionise patient monitoring process and expand lab based testing into everyday life situations. The presented technology can significantly advance wearable sensors, as it offers integrated, unobstructive, multi-node sensing.

References

1. M. Reaz, M. Hussain, F. Mohd-Yasin, Techniques of EMG signal analysis: detection, processing, classification and applications. *Biol. Proced. online* **8**(1), 11–35 (2006)
2. G. Caffier, D. Heinecke, R. Hinterthan, Surface EMG and load level during long-lasting static contractions of low intensity. *Int. J. Ind. Ergon.* **12**(1–2), 77–83 (1993)
3. S.H. Roy, C.J. De Luca, L. Snyder-Mackler, M.S. Emley, R.L. Crenshaw, J.P. Lyons, Fatigue, recovery, and low back pain in varsity rowers. *Med. Sci. Sports Exerc.* **22**(4), 463–469 (1990)
4. R. Merletti, P.A. Parker, *Electromyography: Physiology, Engineering, and Non-invasive Applications*, vol. 11 (Wiley, New York, 2004)
5. M. Cifrek, S. Tonković, V. Medved, Measurement and analysis of surface myoelectric signals during fatigued cyclic dynamic contractions. *Measurement* **27**(2), 85–92 (2000)
6. J. Petrofsky, Filter bank analyser for automatic analysis of the EMG. *Med. Biol. Eng. Comput.* **18**(5), 585–590 (1980)

7. L.D. Gilmore, C.J. De Luca, Muscle fatigue monitor (MFM): second generation. *IEEE Trans. Biomed. Eng.* **1**, 75–78 (1985)
8. F.B. Stulen, C.J. De Luca, Muscle fatigue monitor: a noninvasive device for observing localized muscular fatigue. *IEEE Trans. Biomed. Eng.* **12**, 760–768 (1982)
9. R. Merletti, D. Biey, M. Biey, G. Prato, A. Orusa, On-line monitoring of the median frequency of the surface EMG power spectrum. *IEEE Trans. Biomed. Eng.* **1**, 1–7 (1985)
10. A. Peyton, Circuit for monitoring the median frequency of the spectrum of the surface EMG signal. *IEEE Trans. Biomed. Eng.* **5**(BME-34), 391–394 (1987)
11. J. Duchêne, F. Goubel, Acquisition and processing of surface EMG signals with a low-cost microprocessor based system. *J. Biomech.* **15**(10), 791–793 (1982)
12. G. Inbar, O. Paiss, J. Allin, H. Kranz, Monitoring surface EMG spectral changes by the zero crossing rate. *Med. Biol. Eng. Comput.* **24**(1), 10–18 (1986)
13. M.Z. Jamal, Signal acquisition using surface EMG and circuit design considerations for robotic prosthesis in *Computational Intelligence in Electromyography Analysis-A Perspective on Current Applications and Future Challenges* (InTech, Rijeka, 2012)
14. C.J. De Luca, The use of surface electromyography in biomechanics. *J. Appl. Biomech.* **13**, 135–163 (1997)
15. A. Cechetto, P. Parker, R. Scott, The effects of four time-varying factors on the mean frequency of a myoelectric signal. *J. Electromyogr. Kinesiol.* **11**(5), 347–354 (2001)
16. D. Farina, W. Jensen, M. Akay, *Introduction to Neural Engineering for Motor Rehabilitation*, vol. 40 (Wiley, New York, 2013)
17. Y. Blanc, U. Dimanico, Electrode placement in surface electromyography (sEMG) minimal crosstalk area (MCA). *Open Rehabil. J.* **3**, 110–126 (2010)
18. E. Zuniga, X. Truong, D. Simons, Effects of skin electrode position on averaged electromyographic potentials. *Arch. Phys. Med. Rehabil.* **51**(5), 264–272 (1970)
19. J.H. Viitasalo, P.V. Komi, Signal characteristics of EMG with special reference to reproducibility of measurements. *Acta Physiol. Scand.* **93**(4), 531–539 (1975)
20. A. Rainoldi, M. Nazzaro, R. Merletti, D. Farina, I. Caruso, S. Gaudenti, Geometrical factors in surface EMG of the vastus medialis and lateralis muscles. *J. Electromyogr. Kinesiol.* **10**(5), 327–336 (2000)
21. D. Farina, R. Merletti, M. Nazzaro, I. Caruso, Effect of joint angle on EMG variables in leg and thigh muscles. *IEEE Eng. Med. Biol. Mag.* **20**(6), 62–71 (2001)
22. D.B. Chaffin, Localized muscle fatigue-definition and measurement. *J. Occup. Environ. Med.* **15**(4), 346–354 (1973)
23. H. Piper, *Elektrophysiologie menschlicher muskeln* (Springer, Berlin, 1912)
24. S. Cobb, A. Forbes, Electromyographic studies of muscular fatigue in man. *Am. J. Physiol.–Legacy Content* **65**(2), 234–251 (1923)
25. E. Kuroda, V. Klissouras, J. Milsum, Electrical and metabolic activities and fatigue in human isometric contraction. *J. Appl. Physiol.* **29**(3), 358–367 (1970)
26. B. Bigland, O. Lippold, The relation between force, velocity and integrated electrical activity in human muscles. *J. Physiol.* **123**(1), 214 (1954)
27. R. Harding, R. Sen, Evaluation of total muscular activity by quantification of electromyograms through a summing amplifier. *Med. Biol. Eng.* **8**(4), 343–356 (1970)
28. P. Komi, Relationship between muscle tension, EMG and velocity of contraction under concentric and eccentric work, in *New Concepts of the Motor Unit, Neuromuscular Disorders, Electromyographic Kinesiology* (Karger Publishers, Basel, 1973), pp. 596–606
29. A. Fuglsang-Frederiksen, The utility of interference pattern analysis. *Muscle Nerve* **23**(1), 18–36 (2000)
30. D.A. Gabriel, J.R. Basford, K.-N. An, Assessing fatigue with electromyographic spike parameters. *IEEE Eng. Med. Biol. Mag.* **20**(6), 90–96 (2001)
31. D. Farina, R. Merletti, Methods for estimating muscle fibre conduction velocity from surface electromyographic signals. *Med. Biol. Eng. Comput.* **42**(4), 432–445 (2004)
32. K. Masuda, T. Masuda, T. Sadoyama, M. Inaki, S. Katsuta, Changes in surface EMG parameters during static and dynamic fatiguing contractions. *J. Electromyogr. Kinesiol.* **9**(1), 39–46 (1999)

33. J. Potvin, L. Bent, A validation of techniques using surface EMG signals from dynamic contractions to quantify muscle fatigue during repetitive tasks. *J. Electromyogr. Kinesiol.* **7**(2), 131–139 (1997)
34. G. Kamen, D. Gabriel, *Essentials of Electromyography* (Human Kinetics, Champaign, 2010)
35. M. Barbero, R. Merletti, A. Rainoldi, *Atlas of Muscle Innervation Zones* (Springer, Berlin, 2011)
36. Anatomy diagram, Arm muscles (2015)
37. R. Merletti, L.L. Conte, Advances in processing of surface myoelectric signals: part 1. *Med. Biol. Eng. Comput.* **33**(3), 362–372 (1995)
38. H. Broman, G. Bilotto, C.J. De Luca, A note on the noninvasive estimation of muscle fiber conduction velocity. *IEEE Trans. Biomed. Eng.* **5**(BME-32), 341–344 (1985)
39. R. Merletti, L.R.L. Conte, Surface EMG signal processing during isometric contractions. *J. Electromyogr. Kinesiol.* **7**(4), 241–250 (1997)
40. T.-D. Chiueh, P.-Y. Tsai, I.-W. Lai, *Baseband Receiver Design for Wireless MIMO-OFDM Communications* (Wiley, New York, 2012)
41. T.S. Lande, T.G. Constandinou, A. Burdett, C. Toumazou, Running cross-correlation using bitstream processing. *Electron. Lett.* **43**(22), 1181–1183 (2007)
42. M. Zwarts, T. Van Weerden, H. Haenen, Relationship between average muscle fibre conduction velocity and EMG power spectra during isometric contraction, recovery and applied ischemia. *Eur. J. Appl. Physiol. Occup. Physiol.* **56**(2), 212–216 (1987)
43. T. Sadoyama, T. Masuda, H. Miyano, Relationships between muscle fibre conduction velocity and frequency parameters of surface EMG during sustained contraction. *Eur. J. Appl. Physiol. Occup. Physiol.* **51**(2), 247–256 (1983)
44. S. Andreassen, L. Arendt-Nielsen, Muscle fibre conduction velocity in motor units of the human anterior tibial muscle: a new size principle parameter. *J. Physiol.* **391**(1), 561–571 (1987)
45. T. Sadoyama, T. Masuda, H. Miyata, S. Katsuta, Fibre conduction velocity and fibre composition in human vastus lateralis. *Eur. J. Appl. Physiol. Occup. Physiol.* **57**(6), 767–771 (1988)
46. X. Ye, T. Beck, N. Wages, Relationship between innervation zone width and mean muscle fiber conduction velocity during a sustained isometric contraction. *J. Musculoskelet. Neuronal Interact.* **15**(1), 95–102 (2015)
47. M. Naeije, Estimation of the action potential conduction velocity in human skeletal muscle using the surface EMG cross-correlation technique. *Electromyogr. Clin. Neurophysiol.* **23**, 73–80 (1983)
48. S. Henzler, Time-to-digital converter basics, in *Time-to-Digital Converters* (Springer, Berlin, 2010), pp. 5–18
49. R. Pallas-Areny, Interference-rejection characteristics of biopotential amplifiers: a comparative analysis. *IEEE Trans. Biomed. Eng.* **35**(11), 953–959 (1988)
50. J.H. Nagel, Biopotential amplifiers. *Biomed. Eng. Handb.* 1185–1195 (1995)
51. M.J. Burke, D.T. Gleeson, A micropower dry-electrode eeg preamplifier. *IEEE Trans. Biomed. Eng.* **47**(2), 155–162 (2000)
52. E.M. Spinelli, N.H. Martinez, M.A. Mayosky, A single supply biopotential amplifier. *Med. Eng. Phys.* **23**(3), 235–238 (2001)
53. E.M. Spinelli, N. Martínez, M.A. Mayosky, R. Pallàs-Areny, A novel fully differential biopotential amplifier with dc suppression. *IEEE Trans. Biomed. Eng.* **51**(8), 1444–1448 (2004)
54. E.M. Spinelli, R. Pallàs-Areny, M.A. Mayosky, Ac-coupled front-end for biopotential measurements. *IEEE Trans. Biomed. Eng.* **50**(3), 391–395 (2003)
55. R.F. Yazicioglu, S. Kim, T. Torfs, H. Kim, C. Van Hoof, A 30 w analog signal processor ASIC for portable biopotential signal monitoring. *IEEE J. Solid State Circuits* **46**(1), 209–223 (2011)
56. N. Verma, A. Shoeb, J. Bohorquez, J. Dawson, J. Guttag, A.P. Chandrakasan, A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system. *IEEE J. Solid State Circuits* **45**(4), 804–816 (2010)

57. R.R. Harrison, C. Charles, A low-power low-noise cmos amplifier for neural recording applications. *IEEE J. Solid State Circuits* **38**(6), 958–965 (2003)
58. R. Yazicioglu, P. Merken, R. Puers, C. Van Hoof, A 60 uw 60 nv/ hz readout front-end for portable biopotential acquisition systems. *IEEE J. Solid State Circuits* **42**(5), 1100–1110 (2007)
59. R.F. Yazicioglu, P. Merken, R. Puers, C. Van Hoof, A 200 weight-channel EEG acquisition ASIC for ambulatory EEG systems. *IEEE J. Solid State Circuits* **43**(12), 3025–3038 (2008)
60. T. Denison, K. Consoer, W. Santa, A.-T. Avestruz, J. Cooley, A. Kelly, A 2 uw 100 nv/rthz chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials. *IEEE J. Solid State Circuits* **42**(12), 2934–2945 (2007)
61. C.J. De Luca, L. Donald Gilmore, M. Kuznetsov, S.H. Roy, Filtering the surface EMG signal: movement artifact and line noise contamination. *J. Biomech.* **43**(8), 1573–1579 (2010)

Chapter 7

Design and Optimization of ICs for Wearable EEG Sensors

Jiawei Xu, Rachit Mohan, Nick Van Helleputte, and Srinjoy Mitra

1 Basics of EEG Scaling

In modern clinical practice, scalp electroencephalography (EEG) recording is one of the most important noninvasive procedures to measure the electrical activity of the human brain. EEG has a wide range of applications from brain disorder diagnosis [1], stroke rehabilitation [2], brain-computer interface (BCI) [3], and gaming [4]. Conventionally, EEG signal is obtained by placing electrodes on the scalp (Fig. 7.1) [5] along with conductive gel to reduce the electrode-tissue contact impedance. The recorded EEG signal between two electrodes is a differential voltage representing the average intensity and spontaneous activities of a group of neurons underlying the skull. In time domain, EEG response with peaks and valleys (Fig. 7.2) indicates the power spectrum associate with brain activities. In frequency domain, most of the signal falls within a narrow band of 0.5–100 Hz. Some of the prominent frequency bands are called alpha (7–14 Hz), beta (15–30 Hz), theta (4–7 Hz), and delta (1–4 Hz), each having characteristic neurophysiological traits.

As a noninvasive method, EEG plays an important role in both clinical applications and cognitive research. Some clinical use cases include epileptic seizure diagnosis, stroke detection, brain coma patterns, and brain injury evaluation. Other conditions such as dizziness, headache, dementia, and sleeping disorder problems may also be visible in EEG recording. In recent years, EEG is also widely used

J. Xu (✉)
imec/Holst Centre, Eindhoven, The Netherlands
e-mail: jiawei.xu@imec-nl.nl

R. Mohan • N. Van Helleputte
imec, Leuven, Belgium

S. Mitra
School of Engineering, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK

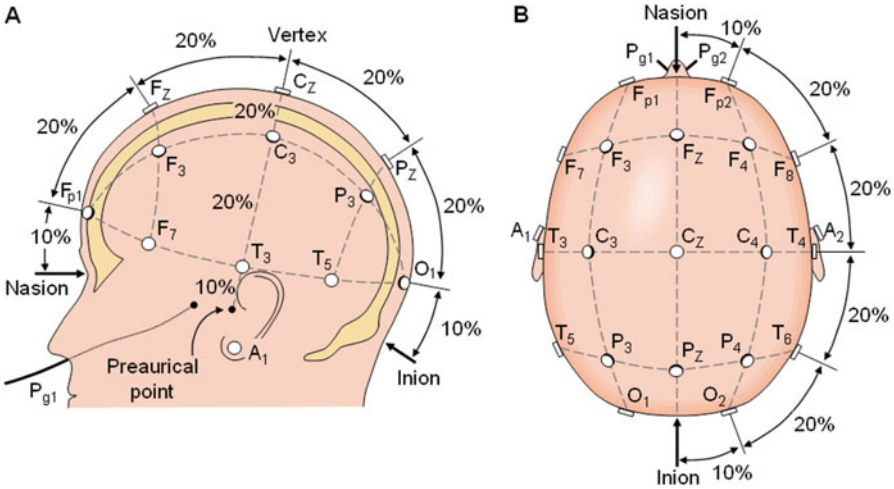


Fig. 7.1 The typical 10–20 EEG recording placement [5] suitable for wearable EEG measurement. Seen from (a) left and (b) above the head: A ear lobe, C central, Pg nasopharyngeal, P parietal, F frontal, Fp frontal polar, O occipital

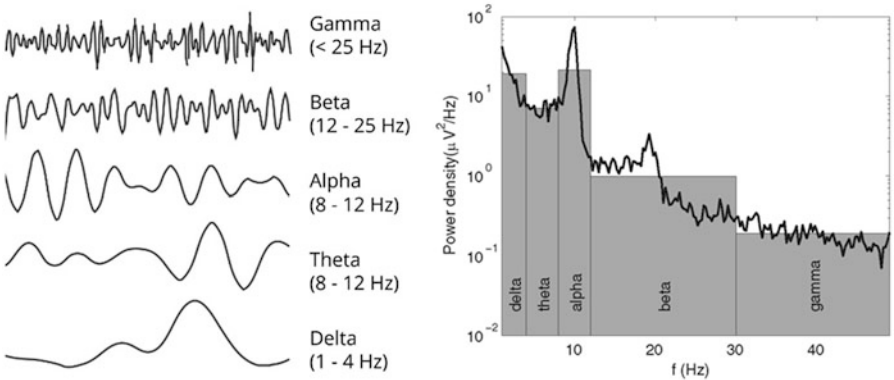


Fig. 7.2 Example of typical EEG signals [6] and an EEG spectrum [7] with its qEEG approximation in terms of band powers, given by the areas of the gray bars

for nonclinical applications such as neurofeedback and brain-computer interface (BCI) study. Neurofeedback records the brain waves which can be used as feedback to the user for self-regulation of brain function training. BCI allows the users to communicate and control a computer system with their brain activities for intended task; typical use cases include user mobility rehabilitation, smart environment, neuromarketing, games, and security [8].

Since the first EEG measurement instrument was invented in 1924, modern EEG acquisition systems have been developed toward smaller size, lower power, and increased user comfort. Due to the growing needs of ambulatory brain activity



Fig. 7.3 Evolution of EEG acquisition systems toward small size and low power. This trend reflects people’s growing needs for care, cure, and prevention of chronic diseases



Fig. 7.4 State-of-the-art wireless and wearable EEG systems

monitoring, further improvement in convenient, long-term EEG recording device is an important research domain (Fig. 7.3). Most of the conventional EEG recording instruments are typically bulky and power hungry and require expert assistance. While some of commercial EEG devices are marketed as portable and for ambulatory usage [9–11] (Fig. 7.4), they do not provide the same data quality for medical grade applications. Thanks to the advances in electrode material, low-power electronics, and integration technology, some state-of-the-art EEG devices have been demonstrated with the necessary analog performance, power efficiency and form factor for long-term and comfortable monitoring of various brain activities [12].

1.1 Introduction to EEG Measurement

Scalp EEG recording requires two types of components: an electrode and an EEG readout circuit (Fig. 7.5). When the electrode is in contact with the skin, chemical reaction occurs between the electrode (metal conductor) and the electrolyte (body fluid); this converts ionic current into electric current. An instrumentation amplifier (IA) is generally the first electronic component in the readout electronics. With inputs connected to a pair of EEG electrodes, the IA amplifies the microvolt-level signal. Apart from two recording electrodes, a third electrode, i.e., bias electrode, is used to bias the subject with a predefined DC voltage to avoid floating inputs. Scalp EEG has a low amplitude ($<100 \mu\text{V}_{\text{pp}}$) and a low bandwidth ($<100 \text{ Hz}$), because EEG measures the voltage fluctuations of a group of neurons in the brain; when neural ionic current reaches to the scalp, its magnitude is attenuated by the skull. The high-frequency components of the signal are heavily filtered by cortical tissues.

As the first block in the signal chain, the characteristics of electrode-tissue interface often become a performance-limiting factor. Practical concerns related to the interface are electrode impedance ($Z_{\text{elec}1,2,3}$ in Fig. 7.5), electrode polarization voltages ($V_{\text{DC}1,2,3}$ in Fig. 7.5), susceptibility to motion artifact, and user comfort. The EEG electrodes are typically made of gold, silver, and silver chloride (AgCl) for good signal quality. Gold electrodes have low impedance but are expensive and not suitable for DC signal measurement since they are polarized electrodes; thus, DC signals cannot pass through skin-electrode interface. Ag/AgCl electrodes are nonpolarized, and they have been considered as the clinical reference by exhibiting the best stability while still being DC compliant and inexpensive.

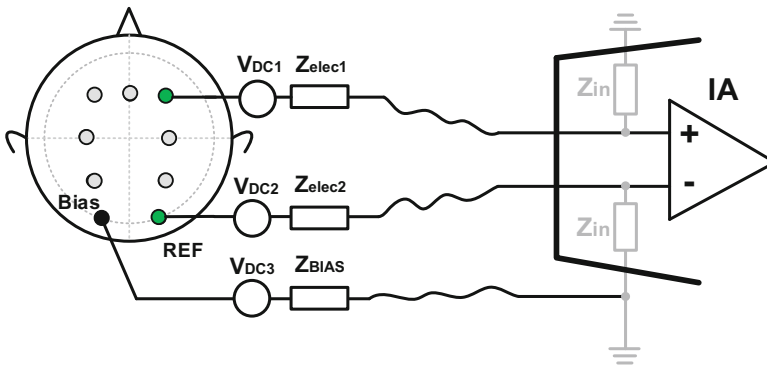


Fig. 7.5 Simplified diagram of bipolar scalp EEG measurement

1.2 Types of EEG Electrodes

Generally, noninvasive biopotential electrodes can be divided into three categories: wet, dry contact, and dry noncontact [13] (Fig. 7.6). Depending on materials and the electrode concepts, a significant difference in skin-electrode impedance can be found between various interfaces. Conventional wet electrodes utilize Ag/AgCl and hydrogel to ensure easy conversion of current, leading to low electrode impedance. Besides, wet electrodes mitigate motion artifact by means of adhesive gel and low variance of electrode impedance. Although gel exhibits low resistance of a few hundred ohm, the equivalent contact impedance between the gel and the skin surface dominates the overall electrode-skin impedance, which is often modeled with a complex impedance of $51\text{ k}\Omega//47\text{ nF}$ [14]. On the other hand, dry contact electrodes eliminate the use of gel to facilitate long-term and convenient EEG recording at the cost of higher skin-electrode contact impedance, ranging from several hundreds of $\text{k}\Omega$ [15] to a few tens of $\text{M}\Omega$ [16]. Dry noncontact electrodes do not even require direct scalp coupling as the electrode and skin are isolated by an air gap. However, this increases the difficulty of ionic current conversion and thus results in even higher electrode impedance (capacitive) and makes it more susceptible to gain attenuation and motion artifact.

While considering user comfort, different types of electrodes have their own advantages and disadvantages. Gel-based electrodes require long preparation time but are acceptable for short (30 min) recording cycles. However, the need to repeatedly apply the gel, itchiness as gel dries up, and the need to wash hair after long-term recording are all significant barriers for user acceptability. Though dry electrodes can mitigate many of these problems, the common rigid metal pin electrodes (used to penetrate the hair and for low contact impedance) can cause discomfort and pain when tightly fitted. Alternatively, electrodes with different

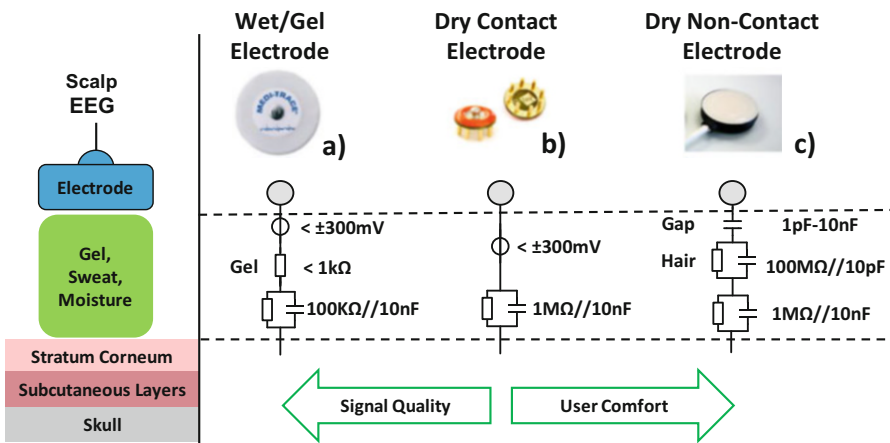


Fig. 7.6 Equivalent electrical models of various electrode-tissue interfaces

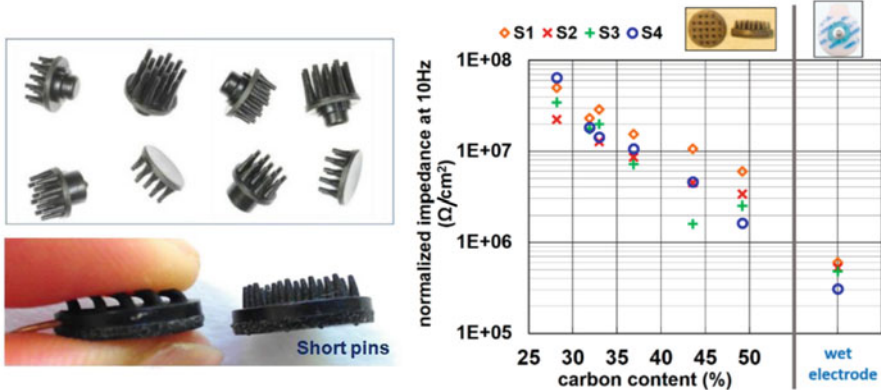


Fig. 7.7 Soft polymer EEG electrode and normalized impedance [18] at 10 Hz measured on the forearm skin of four subjects (S1–S4). For both types of electrodes, the electrode impedance is dominated by the contact impedance between the skin-electrode interfaces

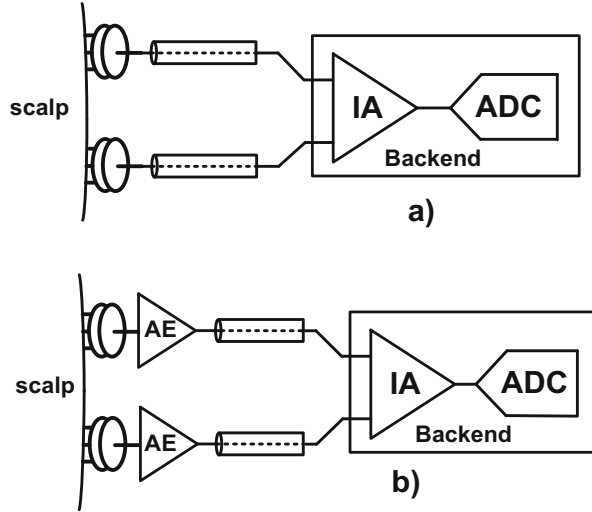
materials such as foam [17] or rubber polymer (Fig. 7.7) [18] are used to provide soft contacts while still exhibiting relatively low impedance. The contact impedance could still be approximately ten times higher than wet electrodes when measured on the skin. Noncontact electrodes might provide the best user comfort, but their reliability in recording medical grade EEG is yet to be tested.

Electrode polarization voltage, or half-cell potential, is an important parameter of EEG sensors. The polarization voltage develops across the electrolyte-electrode interface because of the unbalanced distribution of anions and cations [20]. Some extra electrons move from electrode to electrolyte or vice versa, charging the electrode to a voltage known as the half-cell potential. This voltage appears as a DC offset superimposed on EEG signals for each electrode, and its magnitude depends on the metal type, the ion concentration of body fluid, and the temperature. Ag/AgCl electrodes have been used in many clinical applications because of the lowest polarization voltage of 220 mV and a low baseline drift of 0.13 mV/°C at 25 °C.

2 General Requirements of EEG System

The most important requirements for a EEG readout system are (i) amplify signals of few microvolts while tolerating up to ± 300 mV electrode offset (IEC standards [14]), (ii) suppress capacitively coupled main interference, and (iii) achieve high input impedance to mitigate electrode-tissue impedance variation due to motion/mismatch. The detailed system specification will be discussed in the following section.

Fig. 7.8 Illustration of EEG readout circuits: (a) a conventional solution based on an IA, and (b) a proposed solution based on active electrodes for wearable EEG



As the first stage of the EEG acquisition chain, the readout circuit plays a very important role. High-quality EEG recording heavily relies on low-noise circuits to amplify the μV input signals while minimizing the interference from the environment. The amplified EEG output signal is further conditioned in the analog domain before digitization. This section presents the general design challenge and specifications of EEG readout circuits and gives several examples of state-of-the-art CMOS ICs for brain activity monitoring.

Depending on the interface type, an EEG readout circuit may be implemented with two architectures: passive electrodes with a differential instrumentation amplifier (IA) suitable for low-impedance wet electrodes and local active circuits followed by IAs for dry electrodes. The active circuits placed close to dry electrodes are often called active electrodes (AE). Both architectures (Fig. 7.8a and b) have pros and cons on their analog performance matrix.

2.1 Passive Electrode EEG Acquisition

EEG signals acquired by electrodes need to be amplified by an instrumentation amplifier (IA) (Fig. 7.8a); the key advantage of a differential IA is its low-power feature and the capability to reject common-mode interference. However, for a dry electrode interface, when the IA's inputs are connected to two electrodes with high impedance via unshielded lead wires, the pickup of noise and cable motion significantly reduces the signal quality. Thus, shielding on both lead wires and electrodes is required, which may increase the system complexity and cost.

2.2 Active Electrode EEG Acquisition

Alternatively, an electrode with a co-integrated electrode amplifier (Fig. 7.8b), i.e., an *active electrode (AE)*, reduces noise pickup by means of minimizing the routing between electrode and the amplifier. Furthermore, the AE's low output impedance mitigates cable motion artifacts, thus eliminating the need of shielded cables. Nevertheless, it should be noted that an EEG channel always require two AEs. This not only increases power consumption but also the number of conductor wires (e.g., analog output, power supply and reference, etc.) connecting to the back-end analog signal processor. In practice, many wires can significantly increase the system volume and complexity, especially when additional functions (e.g., analog-to-digital conversion, electrode-tissue impedance measurement) and clocks are integrated in the AEs.

2.3 EEG System Specifications

The specification of a clinical EEG device must be compliant with medical standards, while the requirements for nonclinical research on cognitive neuroscience or BCI applications can be relaxed. Table 7.1 summarizes the key electrical parameters defined in the IEC standard and the IFCN standard, along with the proposed generic specification matrix of a wearable EEG system. This parameter matrix can be

Table 7.1 Parameter matrix defined in medical standards and the proposed electrical specifications for wearable EEG acquisition

	IEC 60601-2-26 [14]	IFCN [21]	Target specifications ^a
Applications	Clinical EEG	Clinical EEG	Wearable EEG
Input voltage range	0.5 mV _{pp}	–	1–10 mV _{pp}
Input referred noise (per channel)	6 μV _{pp}	0.5 μV _{rms} (0.5–100 Hz)	1 μV _{rms} (0.5–100 Hz)
Bandwidth	0.5–50 Hz	0.16–70 Hz	0.5–100 Hz
Electrode offset tolerance	±300 mV	–	±300 mV
Input impedance at 50/60 Hz	–	>100 MΩ	>100 MΩ
CMRR at 50/60 Hz	–	110 dB	100 dB 80 dB (with 51 kΩ//47 nF)
THD (1–10 mV _{pp} input)	–	–	1%
Power dissipation	–	–	100 μW/channel ^b
Safe DC current	50 μA	–	50 μA

^aFor both passive and active electrode-based systems

^bConsidering standard available batteries and a device with 16 electrodes

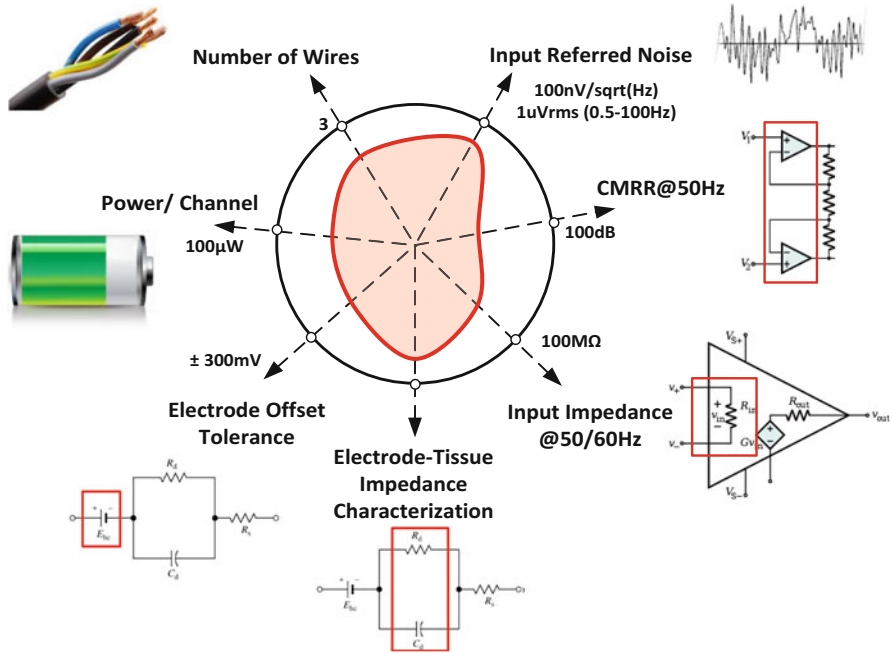


Fig. 7.9 A spider chart illustration of major electrical specifications and the overall performance of an EEG system

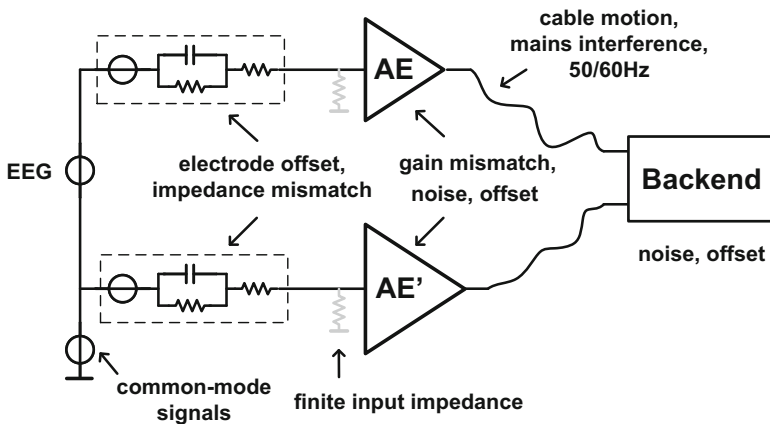


Fig. 7.10 Illustration of major aggressors of one-channel AE-based readout

visualized in a spider chart (Fig. 7.9), where the total area of the chart reflects the overall performance of an EEG system. The following section will discuss each parameter and system aggressors (Fig. 7.10) in detail.

2.3.1 Power Dissipation and Supply Voltage

Battery size and capacity are the major determinants of system size and form factor. Therefore, the power dissipation of an EEG acquisition IC is an important design consideration. To realize a continuous operation up to 24 h with a 3.6 V coin cell lithium battery, the whole system (without wireless transmission) must consume an average supply current of less than 5 mA. On the other hand, the output voltage of a battery also determines the maximum supply voltage of the IC.

2.3.2 Noise

According to IEC standard [14], an EEG system should exhibit maximum input-referred noise of $6 \mu\text{V}_{\text{pp}}$ to detect low-frequency microvolt-level EEG signals. This corresponds to an integrated noise of approximately $0.91 \mu\text{V}_{\text{rms}}$ (assuming 10% of the time, noise will exceed nominal peak-to-peak value), which is equivalent to an average noise density of $91 \text{ nV}/\sqrt{\text{Hz}}$ over 100 Hz bandwidth. State-of-the-art EEG amplifiers usually achieve an input-referred noise of less than $1 \mu\text{V}_{\text{rms}}$ over 0.5–100 Hz bandwidth. Furthermore, $1/f$ noise of the CMOS readout IC usually dominates the low-frequency noise behavior and therefore must be mitigated by dynamic circuit techniques. One should also note that the noise power of an AE system is two times with respect to the noise of single AE.

In practice, however, the noise contribution of electrode-tissue contact impedance can be a dominating factor. Figure 7.11 presents the noise of a parallel RC circuit which mimics the contact impedance model (see Fig. 7.6). As a rule of thumb, the

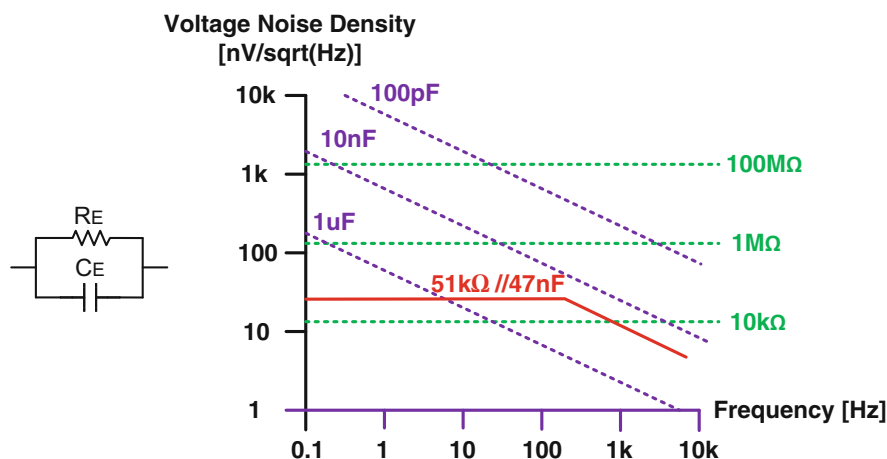


Fig. 7.11 Noise limitation of parallel connected RC electrode impedance [19]. Thermal noise from the source resistance is shown in *green*, and the equivalent electrode noise shaped by the capacitive component is shown in *red*

noise of an IA should be approximately equal or slightly less compared to that of the electrode impedance. This design strategy helps to maximize the noise-power efficiency of the EEG system.

2.3.3 Electrode Offset Tolerance

Electrode offset can easily saturate the IA or at least significantly reduce its output dynamic range. Per IEC standard, scalp EEG systems should be able to accommodate up to ± 300 mV electrode offset while still maintaining its noise performance. As a conventional approach, a high-pass filter with passive capacitive coupling realizes rail-to-rail offset rejection at low power; however, the large size and mismatch of capacitors and resistors reduce the input impedance and the CMRR. Alternative solution is to use an active DC servo loop in either analog or digitally assisted manner. For AEs, the electrode offset is the polarization voltage superimposed on top of subject's body bias voltage.

2.3.4 Input Impedance

Input impedance of the EEG amplifier is very important for recording with dry electrodes. This is because the voltage divider effect, formed by electrode-tissue impedance (ETI) and the circuit input impedance, reduces the effective voltage gain of the AE/IA. To minimize such gain attenuation, electronics connected directly to the electrode should aim for a very high input impedance of at least $100\text{ M}\Omega$, at least an order of magnitude higher than the ETI value. This requirement is especially necessary for dry electrodes, in which the ETI can be even up to a few $\text{M}\Omega$ over EEG bandwidth.

Furthermore, if the AE input impedance is not significantly high, the ETI mismatch between two electrodes can lead to a drastic reduction in systematic CMRR (between 50 dB and 80 dB), even if the two AEs are perfectly matched.

2.3.5 Common-Mode Rejection Ratio (CMRR)

To reject common-mode interference, e.g., 50/60 Hz powerline interference, an IA with 110 dB CMRR is required for clinical EEG; otherwise the input common-mode signal will be converted into a noticeable differential error, polluting the output visibility and reducing amplifier's output dynamic range. For dry electrodes, the CMRR is typically limited by the electrode mismatch and the finite input impedance of the AEs; thus the requirements of high-CMRR IAs can be relaxed.

For a simple AE-based system, the CMRR can be as low as less 60 dB due to the mismatch of the two AEs. However, several CMRR enhancement techniques, such as driven right leg (DRL) [22], common-mode feedback (CMFB) [23], and common-mode feedforward (CMFF) [24], can help to improve the CMRR to more than 80 dB.

2.3.6 Number of Connecting Wires

In its simplest form, an EEG IA has its two input terminals connected to two electrodes. In this case, cabling is not an issue. For AE-based systems, each AE will be connected to a back-end signal processor via multiple wires for power supply and data transfer. Minimizing the number of wires becomes an important consideration to reduce the system cost and complexity, especially when tens of AEs are used for multichannel EEG acquisition or when multiparameter brain activity recording beyond EEG acquisition is required.

3 State-of-the-Art IA Architectures

This section introduces several differential IA architectures which can be interfaced with passive electrodes for EEG measurement. These IAs are not very suitable for dry electrodes.

3.1 Capacitively Coupled Chopper IA

Figure 7.12 depicts a micropower capacitively coupled IA [25] that utilizes chopper modulation to eliminate its intrinsic $1/f$ noise. It achieves $0.98 \mu\text{V}_{\text{rms}}$ input-referred noise over a bandwidth of 100 Hz. In addition, placing the input chopper before the coupling capacitors mitigates their mismatch and ensures a CMRR of more than 100 dB. This IA utilizes a low-power fully integrated DC servo loop (DSL) to suppress electrode offset. The DSL works by extracting the DC component of the output signal and feeding it back to the IA's current summing nodes (i.e., virtual ground) via two capacitors. A large giga-ohm resistor in the DSL (that helps extract the DC) is realized by cascading switched-capacitor resistors. Nevertheless, the maximum electrode offset tolerance is subject to IA's noise and power consumption and is limited to 50 mV in this design. Another limitation of this IA is low input impedance determined by switched-capacitor resistors.

Fig. 7.12 Simplified schematic of instrumentation amplifier [25], illustrating the multiloop feedback paths around the amplifier

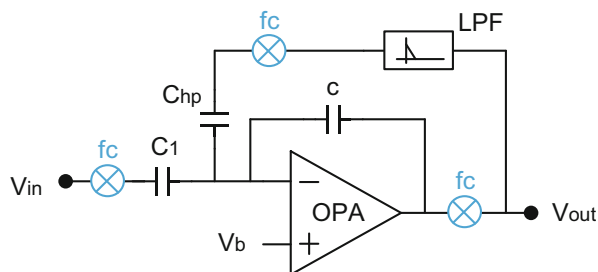


Fig. 7.13 Simplified schematic of a capacitively coupled instrumentation amplifier [26] with the positive feedback loop for input impedance boosting

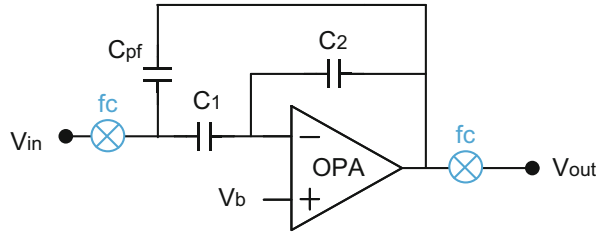
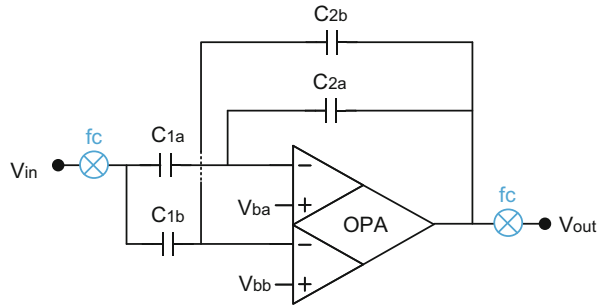


Fig. 7.14 Simplified schematic of the dual-path capacitively coupled amplifier [29] with current-reuse opamp and DC servo loop



Several succeeding works use new circuit design techniques to improve the IA's overall performance. For example, [26] (see Fig. 7.13) utilizes an impedance boosting loop configured in positive feedback. Thus, a portion of the input current is provided by the feedback, therefore reducing the current drawing from source and increasing the input impedance. Another IA [27] improves the electrode offset tolerance up to 250 mV by using large feedback capacitors in the DSL. A power-efficient IA in [28] employs a current-reuse input stage, which takes advantage of the low input swing required at virtual ground. Alternatively, the IA presented in [29] (Fig. 7.14) uses a dual-path capacitively coupled amplifier; the current of core amplifier is reused by both NMOS and PMOS input pairs, which leads to a low noise-power efficiency factor (NEF) of only 1.74.

It should be noted that chopper modulation can be performed at different locations of a capacitively coupled IA for high input impedance. Figure 7.15 presents a low-power IA [30] for EEG acquisition, where the chopper modulation is performed inside the capacitive feedback loop, i.e., at the virtual ground, to facilitate high input impedance and rail-to-rail rejection of electrode offset. Thus, the input impedance is determined only by input capacitance and independent of chopping frequency. On the other hand, chopping after the coupling capacitors at virtual ground also poses issues on noise performance, as input current noise induced by input chopper switches will be converted into significant residual $1/f^2$ noise through high impedance node [32]. Therefore, the IA should utilize a very large input coupling capacitor of 1 nF to suppress residual noise, which again reduces input impedance, as well as the CMRR of the IA to approximately 60 dB.

Fig. 7.15 Simplified schematic of the chopper-stabilized capacitively coupled IA [30] with offset canceling servo loop (C_{INT} and C_{IN} are off-chip capacitors)

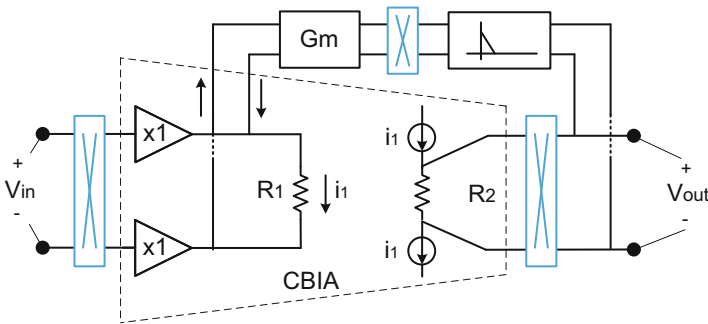
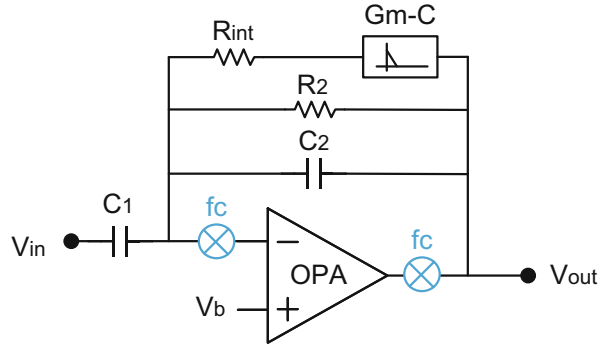


Fig. 7.16 Current-balancing IA [34] with resistive feedback. The DC servo loops for both electrode offset and IA's intrinsic offset are implemented with gm-C-based integrators

3.2 Current-Balancing Chopper IA

Current-balancing IAs [34] using resistive feedback topology (Fig. 7.16) are also widely used for EEG acquisition. In this architecture, the input voltage is first buffered and then converted into current through an input balancing resistor R_1 , realizing high input impedance. The current is mirrored and converted back to voltage via an output resistor R_2 . Therefore, IA's voltage gain is equal to the ratio between two resistors and the ratio of current mirrors. Chopper modulation is typically included to further reduce the noise and improve the CMRR. In [34], the DC servo loop for electrode offset compensation is realized with a gm-C-based integrator followed by a second gm stage (Fig. 7.16). The maximum electrode offset tolerance is about 50 mV determined by the output current of the second gm stage and the input balancing resistor. To overcome this limitation, another IA [35] utilized a DC servo loop configured in voltage feedback (Fig. 7.17), where the integrated DC output voltage is directly fed back to the input of IA, leading to a much larger electrode offset tolerance of at least 300 mV while only consuming tens of nano-ampere current.

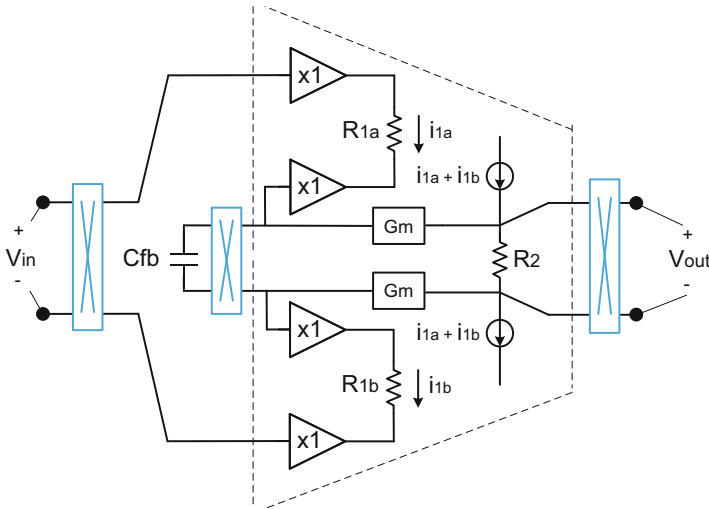


Fig. 7.17 Current-balancing chopper amplifier [35] with a DSL based on voltage feedback

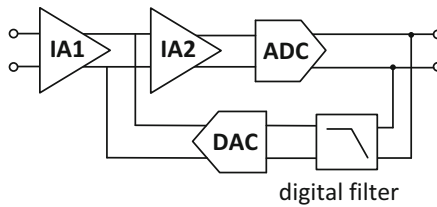


Fig. 7.18 Simplified diagram of a digitally assisted IA [36] local field potential (LFP) DAC to compensate electrode offset and LFP signal, respectively. Low-pass filter (LPF) is realized in digital domain for area efficiency

3.3 Digitally Assisted IA

Conventional analog DC servo loops implemented with large external resistors or capacitors are not applicable to implementable devices. To solve this issue, an IA [36] utilizes an area-efficient digitally assisted architecture (Fig. 7.18) that moves the low-pass filter (LPF) into digital domain to reduce chip area and enable a very low supply voltage of 0.5 V. The IA replaces traditional analog filters with a mixed-signal servo loop, which allows simultaneous measurement of the action potentials (AP) (300 Hz–10 kHz) and local field potentials (LFP) (1 Hz–300 Hz). This DSL consists of a 7b noise-efficient DAC, a digital filter, and a compact 8b ADC can cancel the input offset up to 50 mV. Quantization noise of the DAC is mitigated by feeding the LFP signal back to the output of the first stage IA. With the help of digital filter, a small chip area of only 0.013 mm^2 is achieved while consuming $5 \text{ }\mu\text{W}$ from a 0.5 V supply.

4 State-of-the-Art AE Architectures

Unlike IAs, the amplifiers (or buffers) used in the AE should be single ended. However, active electrodes can use one of the differential IA architectures described in the previous section. When an IA is used as an AE, one input terminal is connected to a reference voltage.

4.1 *Capacitively Coupled Inverting AE*

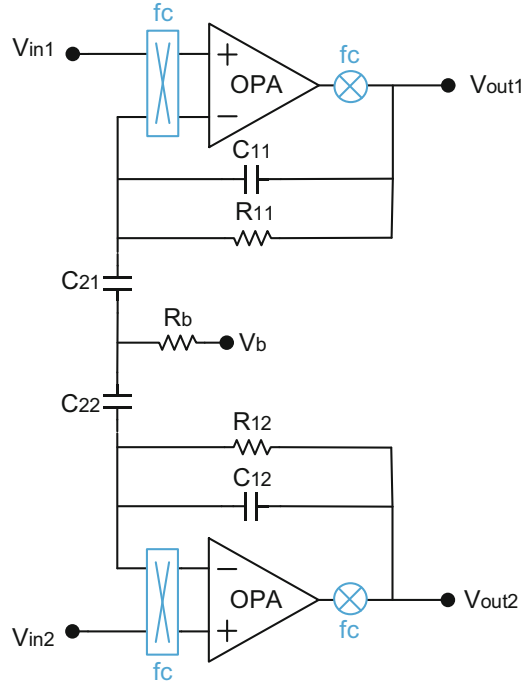
A capacitively coupled AE [23] is in principal similar as the capacitively coupled IA architecture in [30] (Fig. 7.15) but added techniques for improved performance. Apart from chopping and impedance boosting methods previously described, the AE system utilizes an inter-chip common-mode feedback (CMFB) technique between two AEs to improve their CMRR to more than 80 dB. This CMFB loop also excludes the electrode impedance to achieve enhanced stability. As a major limitation, the proposed CMFB needs an extra summing amplifier for common-mode signal extraction, increasing system overall power and complexity. The $1/f^2$ noise induced by high impedance virtual ground and chopping current noise dominates the low-frequency noise of the IA.

In [31], the analog output of AE is combined with positive supply V_{dd} and a current driver to reduce the total number of cables; however, this scheme is at the expense of a reduced output dynamic range.

4.2 *Capacitive Non-inverting AE*

To realize high impedance AEs for dry electrodes, non-inverting amplifiers (instead of inverting ones) have been used in [24] (Fig. 7.19) at the cost of slightly low input dynamic range. In both cases, $1/f$ noise is mitigated by chopping, and electrode offset is compensated by DC servo loops implemented with active integrators. The noise in [24] is reduced compared to [23] by using a very large on-chip capacitor of 5 nF at virtual ground at the cost of reduced input impedance. At system level, a common-mode feedforward (CMFF) technique (Fig. 7.19), connecting all capacitors C_{2X} together, improves the system CMRR to 80 dB. This CMFF technique is specifically suitable for non-inverting amplifier architectures.

Fig. 7.19 Non-inverting AEs [24] with CMFF technique for CMRR enhancement and the equivalent circuit of a pair of AEs consisting of one EEG channel



4.3 Digitally Interfaced AE

Conventional solutions of active electrodes require two types of ICs per EEG channel, e.g., two AE ICs and one back-end analog signal processing IC. To reduce system complexity and cabling, a digital active electrode (DAE) ASIC [33] was proposed for wearable EEG (Fig. 7.20). The DAE ASIC is built to perform analog signal processing and digitization by fully integrating amplifiers, a 12b ADC and a digital I²C interface. With a standard two-wire I²C bus configured with 4b addresses, up to 16 DAEs ICs (e.g., 15 channels) can be connected to a generic microcontroller for simultaneous recording of EEG and electrode-tissue impedance (ETI). It can also be used for multiparameter biopotential signal recording, such as ECG, EMG, and EOG. This DAE-based system significantly reduces its complexity and cost by using such modular DAE sensors. On circuit level, the DAE core amplifier utilizes a “functionally DC-coupled” architecture, i.e., a standard IA with a voltage-based DC servo loop, to enable DC measurement while still achieving input-referred noise of $0.65 \mu\text{V}_{\text{rms}}$ (0.5–100 Hz) and electrode offset tolerance of $\pm 350 \text{ mV}$. In system level, a common-mode feedforward (CMFF) technique (Fig. 7.21), generally applicable to any AE-based system, improves the CMRR of an AE pair from 40 to 102 dB without inducing any instability.

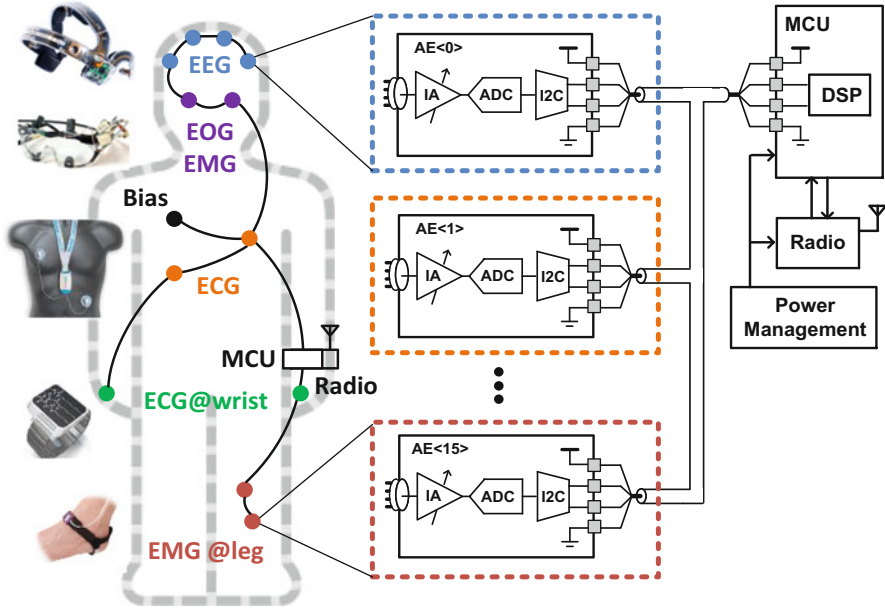


Fig. 7.20 Wearable digital active electrode (DAE) system [33] for multiparameter physiological measurement

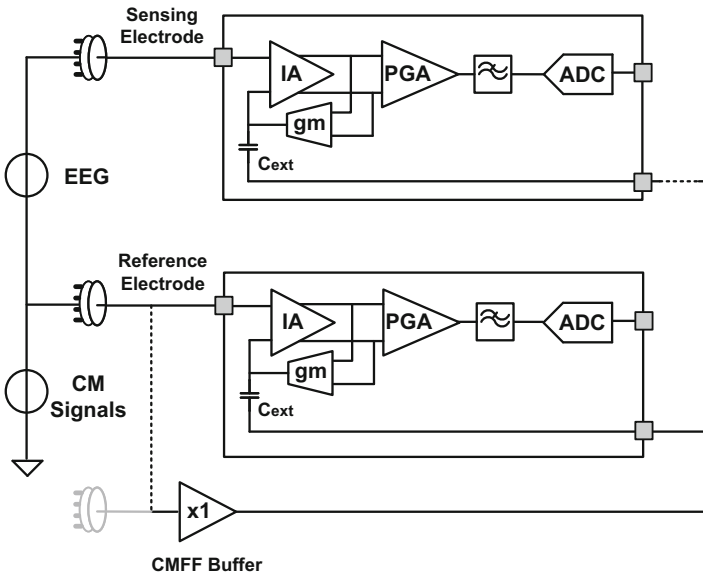


Fig. 7.21 "Functionally" DC-coupled AE with large electrode offset tolerance and the general common-mode feedforward (CMFF) technique for CMRR improvement

4.4 Digitally Assisted AE (or IA)

AEs can be conceived as a useful circuit block not only for standard EEG monitoring but also for any remotely powered sensor node that are placed stand-alone at different parts of the body or even implanted. For such a device, co-integrating state-of-the-art digital and RF circuits together with biomedical analog front-end circuits pose design challenges for conventional low-voltage AFE design. The biggest challenge comes from the trade-off between even lower supply voltage (typically <1 V) and a large dynamic range of the input signal (typically >80 dB).

The amplifier in [38] (Fig. 7.22) explores the inherent spectrum of most biopotential signals to its advantage. Since most of the signal power resides at low frequencies, a spectrum equalization technique is proposed by means of using an analog differentiator and a digital integrator. This reduces the dynamic range requirements of ADC and eliminates the need for a PGA. Alternatively, a time-domain amplifier [40] (Fig. 7.23) is proposed to work at 0.35 V low supply voltage. This amplifier converts the input voltage into time domain (e.g., pulse width) and then converts the pulse width modulated signal back to analog output.

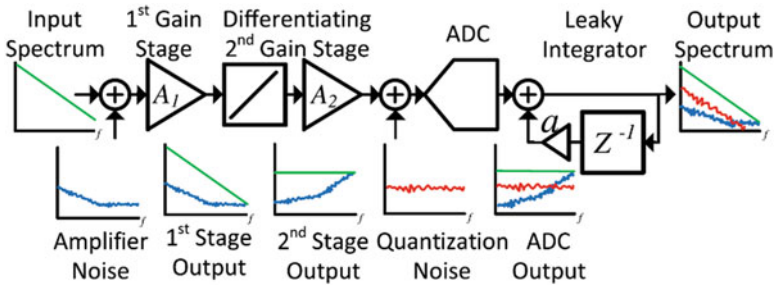
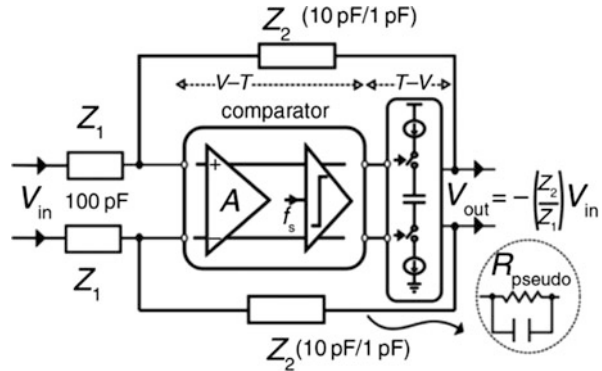


Fig. 7.22 Schematic of low-power equalization AFE [38] taking use of the spectrum feature of ECoG signals

Fig. 7.23 Schematic of time-domain instrumentation amplifier [40]



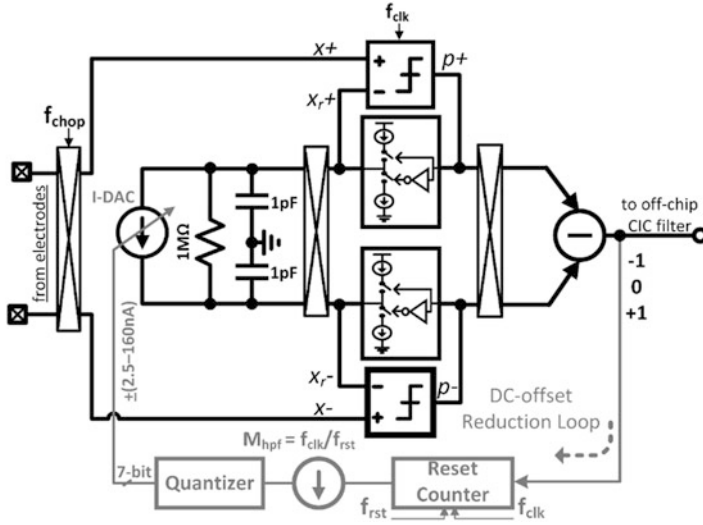


Fig. 7.24 Time-domain AFE in 40 nm CMOS for ambulatory applications [41]

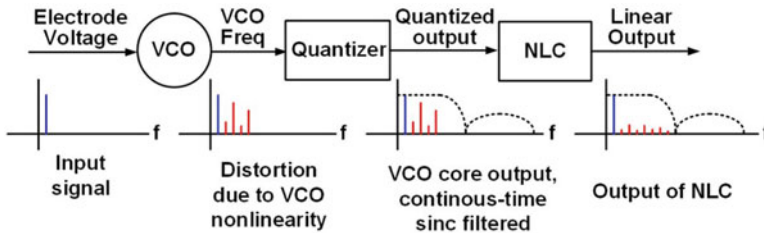


Fig. 7.25 Schematic of VCO-based AFE for EEG acquisition in 40 nm CMOS [37]

The schematic in Fig. 7.24 extends this concept further to implement a 0.6 V time-based AFE in 40 nm CMOS. The circuit comprises of pseudo-differential stages, with each stage implementing an asynchronous delta modulator. The modulator converts the input signal from the electrodes into a PWM signal, similar to the previous design. Chopping is used to reduce the flicker noise while keeping the area of the transistors small. The entire AFE consumes a silicon area of 0.015 mm² and 5 μW power consumption while being able to accommodate up to a 40 mV_{pp} input signal and 300 mV electrode offset – thereby making it an especially good candidate for the large-array EEG/ECOG acquisition.

Alternatively, the implementation presented in the schematic in Fig. 7.25 implements a phase-domain AFE in 40 nm CMOS. The input signal from the electrodes is converted to a phase value using a VCO which is then converted into a digital value, once again using a counter-based implementation. One of the challenges of using a VCO-based design is that they are highly nonlinear in nature. To overcome this, a nonlinearity correction algorithm is implemented in the digital domain using

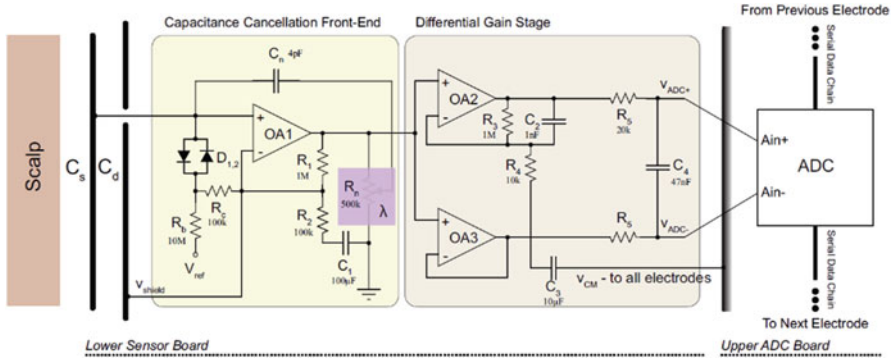


Fig. 7.26 Schematic of the capacitive noncontact AE [39]. The first stage (OA1) performs impedance bootstrap via lead bias resistor R_G . The second stage (OA2, OA3) performs single-end to differential conversion. Data output from 16-bit ADC is an output to a common serial daisy chain

a LUT-based approach. Both designs take the advantages of flexibility and power efficiency of digital circuits, leading to low-power and ultra-compact ASICs.

4.5 AEs for Noncontact Sensing

Apart from resistive contact AEs, capacitive noncontact AEs [41] featured with ultrahigh input impedance further improve the user comfort. Since the AE is capacitively coupled to the skin, the amplifier must provide its own input bias through a lead bias resistor, which determines AE’s input impedance. To achieve a very high input impedance and to minimize the voltage gain attenuation, non-inverting amplifier [41] (Fig. 7.26) with input impedance bootstrap technique is used. Impedance bootstrap is a positive feedback realized by feeding a portion of the output AC signal back to the reference terminal of the lead bias resistor, and so the voltage swing across lead bias resistor is close to zero.

5 Conclusion

This chapter reviews state-of-the-art wearable EEG recording ICs and systems, discusses the general design challenges of scalp EEG interfaces, proposes the generic system specifications of a wearable EEG acquisition IC, and presents two major architectures for EEG readout circuits, namely, passive electrode EEG and active electrode (AE) EEG. Several design examples of IA and AE architectures and key IC design techniques are reviewed and compared.

References

1. M. Teplan, Fundamentals of EEG measurement. *Meas. Sci. Rev.* **2**, 1–11 (2002)
2. A. Ramos-Murguialday, D. Broetz, M. Rea, et al., Brain–machine interface in chronic stroke rehabilitation: a controlled study. *Ann. Neurol.* **74**(1), 100–108 (2013)
3. L.F. Nicolas-Alonso, J. Gomez-Gil, Brain computer interfaces, a review. *Sensors* **12**, 1211–1279 (2012)
4. L.D. Liao et al., Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors. *J. Neuroeng. Rehabil.* **9**(1), 5 (2012)
5. J. Malmivuo, R. Plonsey, *Bioelectromagnetism: Principles and Applications of Bioelectric and Biomagnetic Fields* (Oxford University Press, New York, 1995)
6. NeuroSky [Online]. <http://neurosky.com/2015/05/greek-alphabet-soup-making-sense-of-eeb-bands/>
7. S.J. Van Albada, P.A. Robinson, Relationships between Electroencephalographic Spectral Peaks Across Frequency Bands. *Front. Hum. Neurosci* **7**, 56 (2013.) PMC. Web. 10 Sept. 2017
8. S.N. Abdulkader et al., Brain computer interfacing: applications and challenges. *Egypt. Inf. J.* **16**(2), 213–230 (2015)
9. NeuroSky [online]. <http://neurosky.com/biosensors/eeb-sensor/biosensors/>
10. Emotive [online]. <https://www.emotiv.com/epoc/>
11. Cognionics [online]. <http://www.cognionics.com/index.php/products/hd-eeb-systems/quick-20-dry-headset>
12. S. Patki et al., Wireless EEG system with real time impedance monitoring and active electrodes, in *IEEE Biomedical Circuits and Systems Conference* (BioCAS), Hsinchu, pp. 108–111, November 2012
13. Y.M. Chi, T.P. Jung, G. Cauwenberghs, Dry-contact and noncontact biopotential electrodes: Methodological review. *IEEE Rev. Biomed. Eng.* **3**, 106–119 (2010)
14. IEC 60601-2-26: Medical electrical equipment – Part 2-26: Particular requirements for the basic safety and essential performance of electroencephalographs (2012)
15. Y.-H. Chen et al., Comfortable polymer dry electrodes for high quality ECG and EEG recording. *Sensors* **12**, 23758–23780 (2014)
16. R. Dozio, A. Baba, et al., Time based measurement of the impedance of the skin-electrode interface for dry electrode ECG recording. *Proc. IEEE EMBC*, 5001–5004 (2007)
17. C.T. Lin, L.D. Liao, et al., Novel dry polymer foam electrodes for long-term EEG measurement. *IEEE Trans. Biomed. Eng.* **58**(5), 1200–1207 (2010)
18. Y.H. Chen, M. Op de Beeck, et al., Comb-shaped polymer-based dry electrodes for EEG/ECG measurements with high user comfort. *IEEE EMBC*, 551–554 (2013)
19. Biosemi [online]. https://www.biosemi.com/faq/without_paste.htm
20. S. Lee, J. Kruse, Biopotential electrode sensors in ECG/EEG/EMG systems, ADI, 2008
21. M.R. Nuwer et al., IFCN standards for digital recording of clinical EEG. *EEG. Clin. Neuro.* **106**(3), 259–261 (1998)
22. B.B. Winter, J.G. Webster, Driven-right-leg circuit design. *IEEE Trans. Biomed. Eng.* **1**, 62–66 (1983)
23. J. Xu, R.F. Yazicioglu, et al., A 160 μ W 8-channel active electrode system for EEG monitoring. *IEEE Trans. Biomed. Circuits Syst.* **6**, 555–567 (2011)
24. J. Xu, S. Mitra, et al., A wearable 8-channel active-electrode EEG/ETI acquisition system for body area networks. *IEEE J. Solid State Circuits* **49**(9), 2005–2016 (2014)
25. T. Denison, K. Consoer, A. Kelly, et al., A 2.2 μ W 100 nV/ $\sqrt{\text{Hz}}$ chopper-stabilized instrumentation amplifier for chronic measurement of neural field potentials. *IEEE J. Solid State Circuits* **42**(12), 2934–2945 (2007)
26. Q. Fan et al., A 1.8 μ W 60 nV/ $\sqrt{\text{Hz}}$ capacitively-coupled chopper instrumentation amplifier in 65 nm CMOS for wireless sensor nodes. *IEEE J. Solid-State Circuits* **46**(7), 1534–1543 (2011)
27. M. Altaf, C. Zhang, J. Yoo, A 16-channel patient-specific seizure onset and termination detection SoC with impedance-adaptive transcranial electrical stimulator. *IEEE J. Solid-State Circuits* **50**(11), 2728–2740 (2015)

28. F.M. Yaul, A.P. Chandrakasan, A sub- μ W 36 nV/ $\sqrt{\text{Hz}}$ chopper amplifier for sensors using a noise-efficient inverter-based 0.2 V-supply input stage. *Dig. ISSCC*, 94–96 (2016)
29. S. Song, M. Rooijackers, P. Harpe, et al., A low-voltage chopper-stabilized amplifier for fetal ECG monitoring with a 1.41 power efficiency factor. *IEEE Trans. Biomed. Circuits Syst.* **9**(2), 237–247 (2015)
30. N. Verma, A. Shoeb, J. Bohorquez, et al., A micro-power EEG acquisition SoC with integrated feature extraction processor for a chronic seizure detection system. *IEEE J. Solid State Circuits* **45**(4), 804–816 (2010)
31. M. Guermendi et al., Active electrode IC combining EEG, electrical impedance tomography, continuous contact impedance measurement and power supply on a single wire. *Proc ESS-CIRC*, 335–338 (2011)
32. J. Xu et al., Measurement and analysis of current noise in chopper amplifiers. *IEEE J. Solid State Circuits* **48**(7), 1575–1584 (2013)
33. J. Xu, B. Busze, et al., A 15-channel digital active electrode system for multi-parameter biopotential measurement. *IEEE J. Solid State Circuits* **50**(9), 2090–2100 (2015)
34. R.F. Yazicioglu, P. Merken, R. Puers, et al., A 60 μ W 60 nV/ $\sqrt{\text{Hz}}$ readout front-end for portable biopotential acquisition systems. *IEEE J. Solid-State Circuits* **42**(5), 1100–1110 (2007)
35. N. Van Helleputte et al., A multi-parameter signal-acquisition SoC for connected personal health applications. *Dig. ISSCC* **57**, 314–315 (2014)
36. R. Muller et al., A 0.013 mm² 2.5 μ W, DC-coupled neural signal acquisition IC with 0.5 V supply. *IEEE J. Solid-State Circuits* **1**, 232–243 (2012)
37. E.D. Kondylis et al., Detection of high-frequency oscillations by hybrid depth electrodes in standard clinical intracranial EEG recordings. *Front. Neurol.* **5**, 1–10 (2014)
38. W. Smith, B. Mogen, E. Fetz, B. Otis, A spectrum-equalizing analog front end for low-power electrocorticography recording. *Dig. ESSCIRC*, 107–110 (2014)
39. Y.M. Chi, G. Cauwenberghs, Micropower non-contact EEG electrode with active common-mode noise suppression and input capacitance cancellation. *Proc. EMBC*, 4218–4222 (2009)
40. R. Mohan, L. Yan, G. Gielen, et al., 0.35 V time-domain-based instrumentation amplifier. *Electron. Lett.* **50**(21), 1511–1513 (2014)
41. R. Mohan, S. Zaliasl, G.G.E. Gielen, C. Van Hoof, R.F. Yazicioglu, N. Van Helleputte, A 0.6-V, 0.015-mm², time-based ECG readout for ambulatory applications in 40-nm CMOS. *IEEE J. Solid State Circuits* **52**(1), 298–308 (2017)
42. W. Jiang, V. Hokhikyan, H. Chandrakumar, V. Karkare, D. Markovic, 28.6 A +50 mV linear-input-range VCO-based neural-recording front-end with digital nonlinearity correction. *Dig. ISSCC*, 484–485 (2016)

Chapter 8

Circuits and Systems for Biosensing with Microultrasound

Holly Susan Lay and Sandy Cochran

1 Introduction

Ultrasound (US) imaging has a long history in medical applications, with documented use dating back to the 1940s [52]. As it is noninvasive, inexpensive and relatively portable, it has become a recognised technique in the clinical sector. More recent research has expanded the range of frequencies used for medical imaging above 20 MHz, allowing the creation of higher-resolution microultrasound (μ US) images. μ US is an imaging modality with resolution in the range 10s–100s of microns, allowing imaging of thin layers of tissue and tissue structure which cannot be resolved using traditional frequencies. This has led to its adoption in small animal imaging, intravascular imaging, optical and ophthalmic applications and other situations which require fine detail for diagnosis and interpretation.

The higher resolution of μ US comes at a cost in penetration depth because of the frequency-dependent attenuation present in biological tissue. While the general system structure of μ US electronics is the same as in the standard US, the higher loss factors, higher frequencies and other application-specific considerations introduce specific challenges to the development of the necessary support electronics.

The first section of this chapter reviews the basics of US imaging and places it in context in biosensing. The second section explores the current uses of μ US in biosensing, and the third section establishes the electronic system specifications for a medical US system, reviewing existing hardware and describing the necessary IP blocks. The fourth section describes an application-specific integrated circuit (ASIC) implementing these blocks, and the final section reviews existing applications of similar circuits.

H.S. Lay (✉) • S. Cochran
School of Engineering, University of Glasgow, Glasgow, UK
e-mail: holly.lay@glasgow.ac.uk

1.1 The Nature of Ultrasound

US imaging uses sound waves above the range of human hearing to map the mechanical properties of a target object and displays them in the form of an image. The strength of the resultant echoes is dependent on the intensity of the initial pressure wave, the attenuation of the propagating media (caused by both absorption and scattering) and the changes in acoustic impedance encountered. Acoustic impedance, Z , is related to the density, ρ , and the phase speed of sound in the material, v :

$$Z = \rho v \quad (8.1)$$

Z allows calculation of transmission and reflection coefficients with the same equations as electrical transmission lines [45].

Conventional medical US operates in the range 1–20 MHz, with μ US systems operating at frequencies $f_c > 20$ MHz. While acoustic propagation in biological tissue is nonlinear to various degrees, this can be neglected when performing conventional US imaging, so received US signals can be assumed to have the same bandwidth characteristics as the original transmitted signal [26], with amplitude and phase modulation introduced by the propagation path and the reflection coefficient:

$$V(t) = A(t) \sin(2\pi ft + \theta) \quad (8.2)$$

In Eq. (8.2), $V(t)$ is the voltage output from the system, $A(t)$ and θ are, respectively, the amplitude and phase characteristics combining the original signal and its passage through tissue, and $\sin(2\pi ft)$ is the original signal.

US waves are generated using ultrasonic transducers operating under the direct and converse piezoelectric effects. In medical systems, they are predominantly operated in the pulse-echo mode in which the same transducer transmits and receives the pressure signals [10]. From a system's point of view, the most important parameter of these devices is their capacitance. Most devices use parallel plate electrode configurations to induce and sense the internal electric fields, resulting in a dominant capacitive term which can have a large impact on system loading and must be modelled in any electronics design [25].

In US imaging, the main incentive for increasing the transducer's centre frequency, f_c , is to achieve a corresponding increase in spatial resolution of the final image. This can be separated into two components: the axial resolution, measured on the axis normal to the face of the US transducer, and the lateral resolution, measured on the axis parallel to the face of the transducer. The axial resolution is primarily a function of the temporal length of the initial ultrasonic pressure wave, which is related both to f_c and the impulse response of the device. The lateral resolution is a function of the width of the active area of the transducer, its aperture, the US wavelength, λ , and the depth of the focal point of the beam. The ideal lateral resolution, L , can be expressed as:

$$L = \lambda \frac{\text{focal depth}}{\text{aperture}} = \lambda f_{\#} \quad (8.3)$$

where the ratio of the focal depth to the aperture size is called the f -number, $f_{\#}$. Increasing f_c and the bandwidth of the transducer will thus improve both the axial and lateral resolutions, resulting in the ability to distinguish smaller features [33].

λ can be calculated from the speed of sound, v , divided by the frequency of the signal, f . The speed of sound in most biological tissues lies between $v = 1484$ m/s (water) and $v = 1590$ m/s (liver) [3]. When calculating values, it is common to use $v = 1500$ m/s as a convenient approximation. Based on this, $\lambda = 1$ mm at $f = 1.5$ MHz and $\lambda = 0.1$ mm at $f = 15$ MHz. This tenfold increase in resolution allows the detection of structures of interest to radiologists, corresponding with the limit of viewing with the naked eye [62]. Further increases in f provide resolutions of 10s of microns, which allow imaging of discrete layers with thicknesses < 1 mm and identification of small changes in tissue structure characteristic of early-stage disease progression [13].

1.2 Relevance to Biosensing

1.2.1 Sensing in the Body

The most familiar uses of medical US feature relatively large external probes intended for imaging large fields of view, e.g. in obstetrics, but there is also extensive application of US imaging in minimally invasive devices in current clinical use and as subjects of active research and development. Current clinical standards recognise and encourage the use of US in endoscopic imaging to supplement optical instruments for imaging subsurface tissues and previously identified areas of disease which have not yet penetrated to the surface. This can allow earlier diagnosis and treatment, improving patient outcomes [22]. While some work has been done in the μ US domain, standard probes for oesophageal imaging operate in the range $7.5 < f < 12$ MHz, balancing resolution and penetration depth (“EUS Imaging” [12]). Lower-frequency probes can also be used for transoesophageal echocardiography (TEE) imaging of the heart (“Transoesophageal Echocardiography (TEE)” [57]).

Another established clinical application of minimally invasive US imaging is in catheter probes for intravascular imaging (IVUS). Because of the small size of the probes and the shallow imaging depth, μ US in the range 20–30 MHz is commonly used in these catheters [5].

Research into further applications of US in biosensing is currently under way, including the use of US transducers integrated into needles to assist surgical vision [21], biomonitoring of implants with Doppler US [60] and integration of μ US and other biosensing technologies into capsule endoscopy [29].

1.2.2 Need for CMOS Circuits and Systems

Reviewing current and future applications of US and μ US in biosensing, it is clear that circuit and system approaches are heavily influenced by the form factor and limited access to the probes. Any relevant probe must be hermetically sealed, with all required power and signal wires/antennas within packages small enough to introduce into narrow access channels such as the throat or rectum or through surgical channels less than 5 mm in diameter. These tight physical restraints mandate miniaturisation and integration, particularly the use of custom CMOS circuits, and careful system partitioning to achieve adequate system performance within a viable device envelope. Hence, later sections of this chapter explore current developments in μ US CMOS and interconnection technologies.

2 Microultrasound for Biosensing

2.1 *Ultrasound in SAW Chemical Sensors*

Current research into the manipulation, testing and analysis of chemical samples has led to multiple developments in the lab-on-a-chip (LOAC) domain. LOACs are favoured for their ability to allow miniaturisation and automation of many of the tasks required for the development of new proteins and peptides, amongst other chemical research [18]. While there has been significant research into the use of optical tweezers for sample manipulation, optical devices are limited by the need for expensive and bulky optics and lasers and by relatively small working volumes. A proposed alternative is the use of ultrasonic surface acoustic waves (SAW) to manipulate samples noninvasively on devices which can be easily miniaturised and offer a path to large-scale manufacturing [31].

SAW transducers manipulate microfluidic samples by inducing travelling surface waves which couple into droplets placed on the surface of the device. Microstreaming is induced in the droplets, causing internal mixing and patterning, translation of the droplets or, at very high power, jetting and atomisation [31]. This allows LOAC sensing without the use of external fluid manipulation devices, opening a pathway to full system miniaturisation.

2.2 *Biosensing with Ultrasonic Biomarkers*

An interesting issue in US imaging for the detection of biomarkers is that it is not a true imaging modality but rather a mapping of relative changes in the mechanical properties of the material and its response to pressure waves. Because of this, multiple layers of information about the tissue properties are encoded into the returned echoes. These can be extracted with appropriate analytical techniques

and reference materials in a technique known as quantitative ultrasound (QUS). Depending on the homogeneity of the tissue and the degree to which the imaging parameters are understood, properties such as the acoustic impedance, attenuation and backscattering coefficient can be calculated and compared against reference values to determine tissue health and disease progression [27, 61].

QUS processing is particularly valuable in biological environments where clinicians look for differences in subsurface tissues which are difficult to determine from greyscale US images. IVUS imaging has proven a natural focus for this approach as the properties of plaques developing in the circulatory system have important impacts on long-term patient outcomes. Research with commercial 20 MHz IVUS catheters (Volcano Corporation, San Diego, USA) used QUS to explore the relationship between various biomarkers and tissue characteristics in patients with coronary disease and showed statistical significance between various pharmaceuticals and associated biomarkers [16].

US sensors have also found applications in biometrics due to their mechanical imaging properties. Fingerprints can be detected with US either through capacitive detection of loading by the fingerprint ridges [47] or by pulse-echo imaging of the surface layer of the skin [53]. As both types of detection depend on the fingertip ridges having the correct material properties and physical contours, these sensors are resistant to many spoofing approaches possible with optical sensors. 3D approaches have also been used to image skin pores [35], allowing multiple biometrics to be collected with a single device.

2.3 Parameters of Microultrasound

The difference in f_c between conventional US and μ US is small when compared with contemporary radiofrequency (RF) standards in the 10s–100s of GHz range. However, the strong influence of frequency on the behaviour of US in biological tissues leads to key differences in the design and performance of μ US systems which must be considered, but which also make μ US useful for biosensing and health monitoring.

One major consequence of moving to μ US is the increased attenuation in all biological tissues, noted previously. Attenuation in dB can be assumed to be linear with respect to frequency, as a first approximation, with variations between different tissue types. As an example, the human liver has an attenuation coefficient $\alpha \approx 0.8$ dB/cm at 1 MHz, 15 dB/cm at 10 MHz and 200 dB/cm at 100 MHz [10]. This relationship is well understood, and the loss of penetration depth with frequency is appreciated; there remains a need for an increased emphasis on system gain and noise performance in μ US systems to maximise imaging performance within the unavoidable physical limitations of the approach.

Recalling Eq. 8.3, $f_{\#}$ for a transducer is a function of the ratio between the focal depth and the aperture. Hence, as the focal length is reduced in correspondence with the increased attenuation, the aperture must also be reduced to maintain $f_{\#}$, resulting

in comparatively smaller devices. For array systems, the individual elements must also be reduced, as US transducer elements are constrained by the same grating lobe considerations as in antenna design [10]. For linear arrays, this mandates elements no more than λ in width and approximately $\lambda/2$ for arrays steered off-axis. This results in much more difficult manufacturing and can impact electrical connectivity, a topic discussed further in Sect. 3.

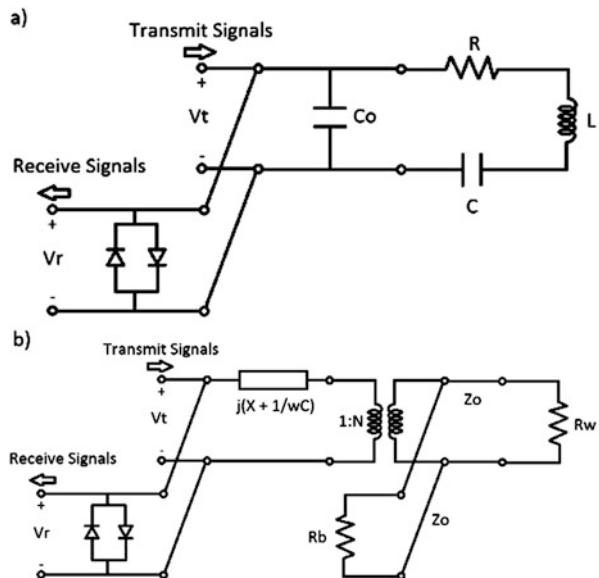
3 Electronic Systems for Ultrasound

3.1 Overall System Architecture

Full electronic modelling of ultrasound transducers can be quite difficult, due to the mechanical loading effects of the acoustic elements of the system, and different models have been developed depending on the aspects of the transducer being assessed in a given engineering situation. Two commonly used models are Mason's model and the KLM model [20]. Mason's model (Fig. 8.1a) in its simpler form models the transducer as a modified tank circuit with a dominant capacitive term and is well suited for analysis centred on the fundamental resonance frequencies and impedance loading effects. However, the component terms can be difficult to calculate for a given configuration without analytical measurements.

The KLM model (Fig. 8.1b) models the electrical and mechanical components more explicitly and can be more easily related intuitively to material properties. It is a better model for calculating loss factors and the full electrical impedance curve of

Fig. 8.1 Simplified electrical circuit approximations for ultrasound transducers. (a) Mason's model is a modified tank circuit suitable for SPICE-based modelling at or near a resonance. (b) The KLM model is conceptually closer to the physical components of the system and provides both loss and acoustical data and electrical impedance data



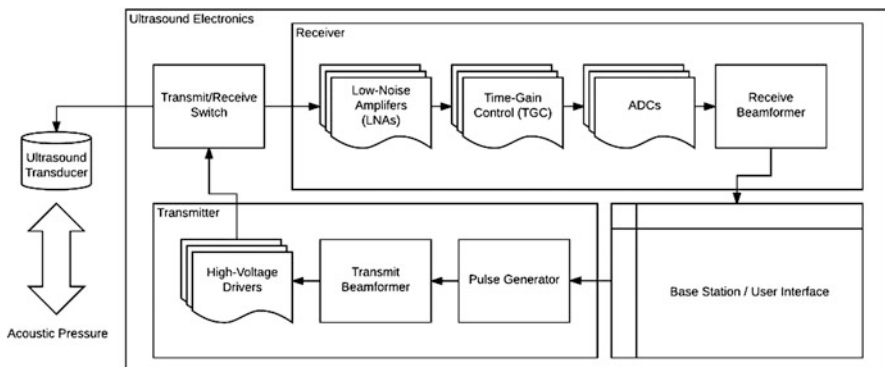


Fig. 8.2 Top-level system diagram of a typical US transmit and receive system. Additional stages may be included depending on system requirements, particularly filtering, partial beamforming and multiplexing

a device. However, the loss terms appear in the form of complex impedances [25], which are incompatible with most SPICE-based circuit models, making it poorly suited to interfacing with extended circuit models. Full system analysis commonly uses a combination of these two models and/or finite element analysis to obtain all the required specifications. Both circuit models in Fig. 8.1 also include a simple diode-based amplifier protection circuit used to prevent overvoltage at the input to the receive amplifiers, but other transmit/receive switching configurations are possible [32].

The standard architecture used in US and μ US systems is well established, with most innovation occurring in the specifics of individual system blocks or in the addition of compatible functionality on top of the accepted standards. Base US systems can be separated into transmit and receive subsystems, with some functional overlap but significant differences in specification which require separate treatment. A typical block diagram of a clinical US system is shown in Fig. 8.2.

The first fundamental decomposition of the electronics is into the transmitting (TX) and receiving (RX) circuits. This separation owes its existence to the significant difference in operating voltage between the two halves of the system, caused by the efficiency loss inherent in the converse and direct piezoelectric conversions in the ultrasonic transducer, ultrasonic attenuation within the imaged medium and electrical, mechanical and ultrasonic losses within the physical components of the system [10].

The TX circuit is designed to generate a sequence of high-voltage pulses of given duration and frequency with intervening time delays (Fig. 8.3). These pulses are applied successively to the ultrasonic array to set up a corresponding pressure wave propagating into the medium. The pulse length can vary from one to multiple cycles at f_c , depending on the specification for power delivery and spatial resolution. The pulse amplitude is on the order of 50–200 V at frequencies >50 MHz for μ US [49]. While such frequencies are no longer considered difficult in electronics design, the

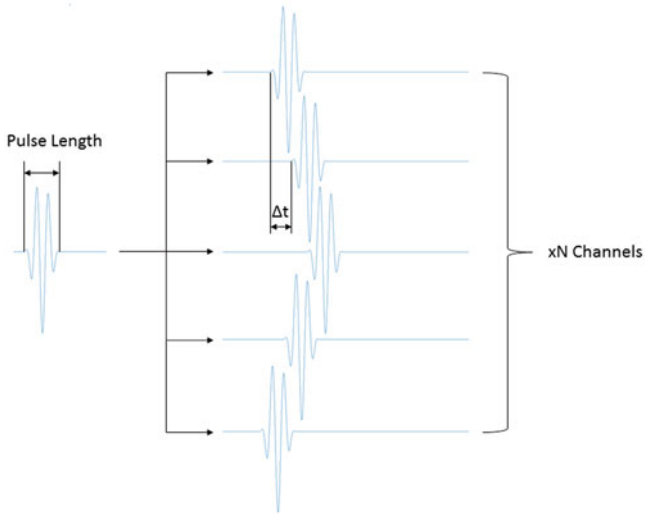


Fig. 8.3 Initial TX pulse and beamforming operation. The beamformer takes the output pulse from the TX driver and outputs one pulse per channel, with delays programmed for the desired ultrasonic focusing

combination of frequency and amplitude driving highly capacitive devices presents a significant design challenge with limited off-the-shelf solutions. Commercially available solutions will be addressed in more detail in the next section.

The complexity of the TX electronics in a particular system is dependent on not only the operating frequency and voltage but also the focusing requirements and the number of elements in the array. In a system with a single element either unfocused or physically focused, it is sufficient to supply a single drive signal. For array systems with electronic focusing, however, a separate drive signal is needed for each array element, requiring replication of the drive circuitry by a factor equal to the number of elements operated together in the array. Commercial systems can have element counts greater than 512 [2], making reduction of the electronic footprint critical to overall system size.

For electronic focusing, the TX pulses must also be delayed with respect to each other to allow shaping of the output acoustic pulse. These electronic delays mimic the path length variance seen in a mechanical lens and are small with respect to the period of the pulses [55]. As exact implementation of the required delays would often demand a system clock frequency many times greater than f_c , many systems either compromise focal accuracy or use analogue delay components to achieve the final delay resolution [7, 54].

US RX electronics share many of the same challenges as the TX electronics with respect to the frequency, channel count and need for fine focussing delays if hardware beamforming is performed. However, the >60 dB reduction in voltage compared with the TX side requires several additional components in the signal chain (Fig. 8.2).

The standard RX signal chain is identical across all channels and features protection circuitry preceding a series of amplification stages, beginning with a low-noise amplifier (LNA) followed by one or more variable- (VGA) and fixed-gain amplifiers. The protection circuitry is extremely important to protect the vulnerable LNA inputs from the high-voltage TX pulses [32]. The specification of the amplification stages varies between systems, depending on the attenuation and desired penetration depth of the image, and additional filtering and matching networks may be included to maximise signal integrity. The full signal chain up to the digitisation phase is referred to as the analogue front end.

The analogue components of the receive network then lead into the analogue-to-digital converters (ADCs), whose specification should be based not only on the Nyquist conditions of the received signals and the desired voltage resolution but should also consider the intended focal delay resolution. While post-sampling interpolation is often used to increase delay resolution, its accuracy is dependent on the linearity of ultrasonic propagation, so higher sampling rates are needed when imaging highly nonlinear media such as biological tissue.

There are multiple approaches to handling the digital processing required to create greyscale images from the RF RX signals. Most systems deal with the initial focal delays and summation (delay-and-sum architecture) [55] in the same circuits as the analogue processing before passing data to the main processor. This can be achieved using digital signal processing (DSP) chips, but it is far more efficient to use either field-programmable gate arrays (FPGAs) or ASICs due to the natural parallelism of the signal chain. Performing the first stage beamforming on the system boards is time efficient and also allows a reduction in the bandwidth of the data connection to the processor on the order of the number of channels in the active focal aperture. With data sampling rates greater than 20 megasamples per second (MSPS) per channel, this can make the difference between the need for a 100 MB/s data channel and a 1–10 GB/s channel.

Final image formation, i.e. scan conversion, is conventionally performed in the main processor of the US scanner, though there has been an increase in GPU-based implementations in recent years with the rise of affordable chipsets driven by the PC gaming market.

3.2 Chipset Implementation

The commonality in the basic signal flow in the US RX system has been addressed by commercial chips integrating the main analogue stages with suitable ADCs in a single package. Featuring up to eight channels/chip, these can be effective in developing a full US system rapidly. However, they are less space and power efficient than the ASIC solution and do not always address DSP requirements, so they are not suitable for all applications. Complementary TX chips are also sometimes available. There are currently three major suppliers of chipsets for the medical US market: Analog Devices Inc. (Norwood, USA), Texas Instruments Inc. (TI, Dallas, USA) and Maxim Integrated (San Jose, USA).

Analog Devices (“Ultrasound | Analog Devices” [58]) currently offers only analogue RX front-end chips, but parts are available integrating different components, allowing better design control and power consumption through the absence of unnecessary features. All Analog Devices chipsets offer eight parallel channels and sampling rates up to 125 MSPS with all chips supplied in $10 \times 10 \text{ mm}^2$ ball-grid array (BGA) packages.

TI offers both TX and RX chipsets (“AFE5807 | Ultrasound | Medical Analog Front End | Description & parametrics” [1]), with up to eight channels on RX. The sampling frequency is limited to 80 MSPS, with most chips functioning only up to 65 MSPS. Chips are packaged in $15 \times 9 \text{ mm}^2$ and $9 \times 9 \text{ mm}^2$ FBGA (fine-pitch BGA) packages. Compromise may be required on TX as eight-channel TI chips can produce only 15 MHz transmit pulses, and the higher frequency, 50 MHz chips are available only with two channels. The 15 MHz chip comes in a $12 \times 12 \text{ mm}^2$ very thin quad flat no-lead (WQFN) package and the 50 MHz one in a $13 \times 13 \text{ mm}^2$ new FBGA (nFBGA) package.

Maxim Integrated offers TX, RX and TX-RX chips for US (“Ultrasound Imaging – Maxim” [59]), but with a lower frequency range than other suppliers. Their eight-channel RX chips allow up to 50 MSPS in a $10 \times 10 \text{ mm}^2$ package. The eight-channel TX chip has a 20 MHz analogue signal bandwidth in the same package. The TX-RX chip integrates TX and RX chips with the same performance in a single die, but it is packaged in a relatively large $10 \times 23 \text{ mm}^2$ chip-scale BGA (CSBGA) package, the size potentially presenting a barrier to integration into smaller, more portable applications.

There has been substantial improvement in the specifications offered by all three companies in the last two decades, so a system designer must reassess them regularly when developing analogue front-end circuits over multiyear timescales.

3.3 CMOS Mismatch with μUS Frequencies and Form Factors

By reviewing the specifications of the current commercial chipsets, certain common trends become apparent which make US electronics design at high frequencies particularly difficult. The highest sampling frequency is 125 MSPS, with 80 MSPS or less more common. Given the Nyquist sampling criterion, the highest value results in a system bandwidth limit of 62 MHz. However, US system design typically allows up to 100% bandwidth on all signals, based on high-performance transducer impulse responses. This reduces f_c to a maximum of 40 MHz. The figure is still lower in many systems as the amplitude-modulated nature of US signal interpretation means that quadrature sampling is desirable for envelope detection without the use of Hilbert transforms or other calculations which are difficult to implement with reduced instruction set computing (RISC) hardware [44]. With quadrature sampling, 125 MSPS limits the f_c to just over 30 MHz, accommodating only the lower end of μUS applications.

The other major factor to consider with commercial chipsets is the size when packaged. The chips discussed in the previous section range from 9 to 15 mm on a side, discounting the large TX-RX chip, and have a maximum of eight channels. Assuming a 128-channel system with both TX and RX capabilities, and using $10 \times 10 \text{ mm}^2$ chips as an average size, this results in 3200 mm^2 of circuit board space just to mount the packages, with a much larger area required to house other components including decoupling, routing, FPGA/DSP chips and supporting power and ground planes. While μUS systems have been successfully developed based on the use of chipsets, they are by necessity either limited in channel count [28] or large enough that they are not functionally portable [8]. With this in mind, the next logical step is to consider the development of ASICs integrating the necessary analogue and digital components into a single chip.

In academic research, there has been only limited development of ASICs for US systems at any frequency. Work has focused on ASICs to integrate with micromachined US transducers, rather than the bulk ceramic used in current clinical systems. This follows from the original design intent of micromachined transducers, which is to maximise system integration by using foundry processes for sensor fabrication, and aligns with the expertise of the research groups involved, which lies in the domain of silicon-level device development.

Most research to date has focused on transducer designs which are compatible with existing foundry processes, requiring only additional layers of standard materials. Examples have been reported at standard US frequencies [17, 46] and for μUS [48]; there are many other examples for $f_c \leq 100 \text{ MHz}$. However, many approaches are limited to the main analogue components, requiring additional hardware to support the digital circuitry. Hybrid approaches introducing thin piezoelectric films into the micromachined device stack have had some success in research and development but are far less common, with only a handful at 8–9 MHz [34, 56] and $>50 \text{ MHz}$ [24, 37]. In this case, the implementations at μUS frequencies include integrated digital circuits, further reducing overall circuit complexity.

Apart from ASICs reported in the public domain, it can be assumed that major commercial clinical system suppliers develop ASICs in-house. However, the specifications and technical details of these chips are not publically available, so their characteristics can be inferred only from the overall system performance, and they are thus not discussed further here.

4 Integrated Circuits for Microultrasound

The specific performance metrics for a μUS system make it particularly difficult to implement using existing hardware, as outlined in the previous section. While discrete components are available which can meet some of the requirements, the consequences of their use on the system as a whole, particularly the large physical dimensions and interconnections, make them impractical, and the development of ASICs is thus effectively mandatory for systems intended for clinical use.

A fully integrated system is, by necessity, a mixed signal implementation, with high-voltage CMOS required for the drive circuits for most types of μ US transducer. The particular impact of μ US system requirements on integrated circuit (IC) design is reviewed in this section.

4.1 Key Factors

4.1.1 Physical Dimensions

As discussed previously, many potential uses for μ US feature miniaturised probes inserted through surgical channels or natural orifices. These probes usually measure 4–10 mm in outer diameter, placing strict limits on the size of the ASIC floor plan. There is also a strong preference for bare die solutions, as the extra area taken up by packaging is often prohibitive.

When calculating estimated die area for a given application and system partitioning, it is important to keep in mind the large number of channels/interconnects required for a system with large array element count. For a 128-channel system with full TX and analogue circuitry integrated in the ASIC, a minimum of 256 pins are needed just for input and output signals, with control, bias and power pins requiring, e.g. 50 more. Not only can this place limits on the minimum perimeter of the die, but the resulting 300+ pin fan-out is demanding of space, expanding the footprint of the supporting printed circuit board (PCB).

Difficulties with pinout and fan-out caused by the high channel counts in μ US applications are an example of a more general interconnection problem. Connections to US transducers are usually made with soldered cable connections or wire bonds like those used in IC packaging [6]. With the smaller pitch required by μ US wavelengths, current research into US and μ US interconnection emphasises flip-chip technologies and high-density flexible circuits to maximise density and robustness [4, 14].

4.1.2 Frequency

f_c for an US system affects both the analogue and digital parts of an ASIC design. The RX amplifier chain must have sufficient bandwidth to avoid signal distortion prior to digitisation. To simulate the final bandwidth correctly during the design phase, it is critical to model the large capacitance presented by the transducer load. While impedance matching is possible externally to reduce loading effects, integration of matching inductors for high channel count systems is not practical in most μ US applications.

μ US frequency requirements have two major impacts on digital design, one on the TX and RX beamforming and the other on digitisation. Both require higher internal clock rates than conventional US for any on-chip processing as well as

input and output pins, with resulting stricter design criteria for trace lengths, jitter and rise/fall times. In particular, the higher sampling rate required by the Nyquist criterion can result in an inability to use pre-existing ADC blocks and other intellectual properties (IP), adding complexity and design time. Additionally, it is usually impractical to implement high enough clock rates to achieve full μ US beamforming delays, so additional circuitry and layout work is required to implement either fixed delay paths or to calculate delays via interpolation [7, 9].

4.1.3 Gain to Overcome Attenuation

To generate images suitable for clinical diagnosis, μ US systems require multiple gain stages to overcome the frequency-dependent attenuation discussed previously. As signals become more attenuated over the course of a complete signal acquisition cycle due to the additional travel time through tissue, one stage of the amplifier chain is a voltage-controlled amplifier used for time gain compensation (TGC). The gain of this amplifier is increased through the course of a single acquisition and then reset to a low gain for each new acquisition. The entire gain chain will normally achieve 60–90 dB of gain, depending on application [49].

4.1.4 Power Consumption

The importance of total power consumption is heavily dependent on the particular application and probe design. For probes with a fixed connection to an external base station such as catheters or endoscopes, power consumption is not critical, so long as any waste heat produced within the internal components can be dissipated via cooling systems. For battery-powered autonomous systems such as implants and capsules, however, power consumption must be strictly controlled, both to minimise heat production and to extend battery life.

4.1.5 Noise Performance

The high frequencies, mixed signal design and high gain issues previously discussed can lead to concerns about noise performance both within the system and coupled from external sources. The noise induced by the ultrasound transducer itself is dominated by thermal noise [39] which can be calculated from the equivalent circuits (Fig. 8.1) using Nyquist's relation. The RX chain, in particular, must be shielded from parasitic interference from the high-voltage transmitter, the ADC and the digital logic blocks as the input signals to the amplifier chain usually have low signal-to-noise ratio (SNR) due to acoustic attenuation. The beamforming process increases SNR as most of the noise is decorrelated between channels, but standard design procedure is to use a low-noise amplifier (LNA) at the head of each amplifier chain to control the overall noise figure and maximise SNR.

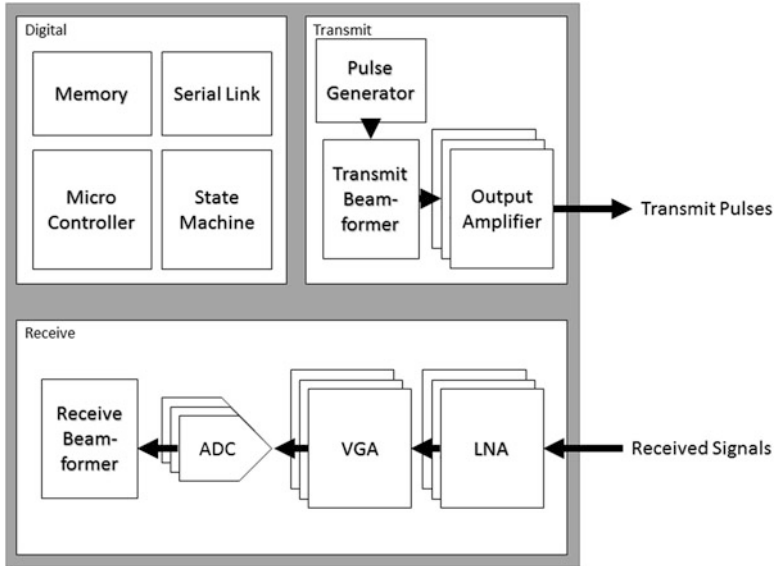


Fig. 8.4 Basic IP blocks for full TX and RX processing of μ US. The digital and TX blocks shown in the top half of the floor plan should be well isolated from the analogue RX blocks to avoid cross-coupling interference

4.1.6 Summary of IP Blocks

Reviewing the specifications discussed in this section as well as the overall system design discussed in Sect. 2, a generalised set of IP blocks necessary for a fully integrated μ US IC can be defined, as shown in Fig. 8.4.

5 Implementation and Demonstration

Previous sections have focused on a general discussion of the application of μ US to biosensing and the required circuit and systems to implement a functional system suitable for clinical purposes. While these principles can be applied in any application, specific applications and particularly the types of μ US transducers that are used may require modifications to the general system design and will affect which of the specifications dominates the design process. This section will discuss the various types of μ US transducer currently in use in clinical and research environments and their interactions with IC architectures.

Examples of the three main configurations of piezoelectric transducers are given in Fig. 8.5. The first configuration, Fig. 8.5a, is based on bulk piezoceramic or piezocomposite material operating in the thickness mode and can also be used for thick films such as polyvinylidene fluoride (PVDF). The second configuration,

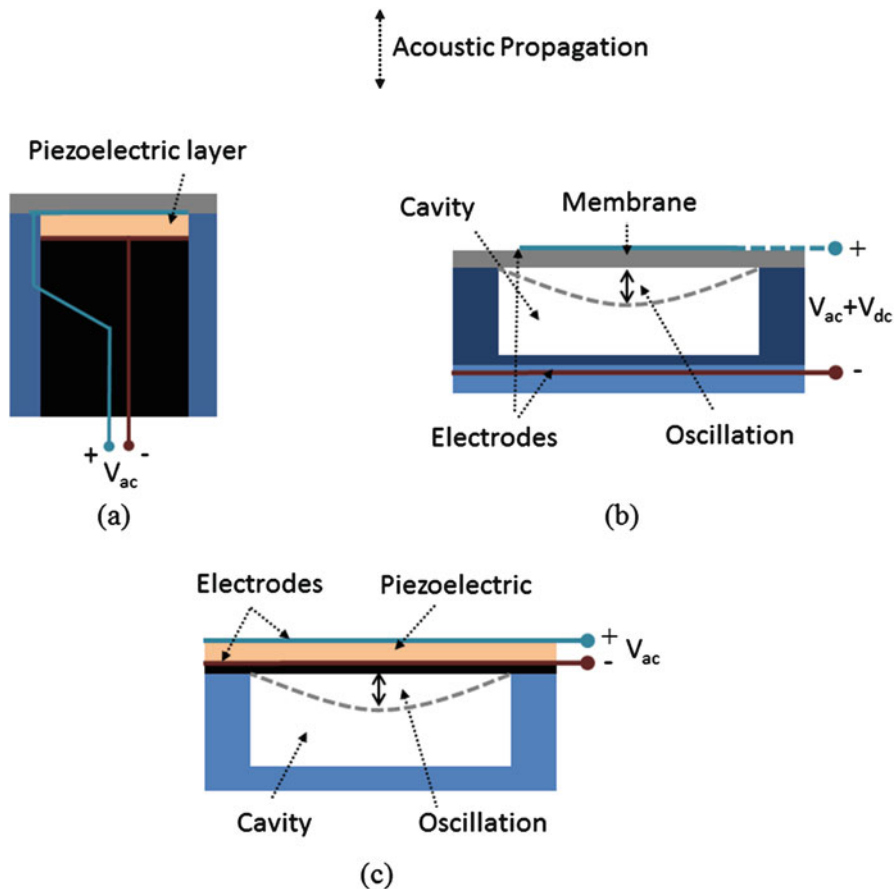


Fig. 8.5 Mechanical configurations of ultrasonic transducers and their electrical input/output ports. **(a)** Standard bulk ceramic or thick film configuration for thickness-mode operation. **(b)** Capacitive micromachined ultrasonic transducer (CMUT) configuration with flexural mode and capacitive readout. **(c)** Piezoelectric micromachined ultrasonic transducer (PMUT) configuration with voltage readout

Fig. 8.5b, is an example of a CMUT which operates in a flexural mode, with acoustic signals transduced into changes in capacitance rather than changes in electric field. The final configuration, Fig. 8.5c, is an example of a PMUT. PMUTs represent a compromise between the integration possible with CMUTs and the traditional piezoelectric materials as they operate in the same flexural mode as CMUTs but with a thin piezoelectric film providing electromechanical transduction, allowing the use of more conventional electronics.

5.1 *Alternative Implementations*

5.1.1 **Bulk Piezoelectric Components**

The most common type of US transducer is manufactured using bulk lead zirconate titanate (PZT), a piezoelectric material with properties that make it suitable for operation in both the TX and RX modes of US and μ US imaging. Other materials such as crystalline lead magnesium niobate doped with lead titanate (PMN-PT) and lithium niobate (LiNbO_3) have also been used with very similar manufacturing principles, so for the purposes of the electronics, they can be treated as a single category.

Despite being the most common implementation, there is minimal reporting in the literature on the design of ASICs for bulk piezoelectrics, with most applications using commercial chip solutions [9]. Integrated flip-chip solutions are particularly difficult to implement with bulk PZT due to the temperature sensitivity of PZT and other similar materials and acoustic interaction of the physical IC structure with the transducer. Flip-chip bonding as implemented conventionally uses relatively high temperatures which can degrade the piezoelectric properties of bulk ceramics. Some groups have had some success working with alternative epoxies [40, 43], but the acoustic loading effects make optimisation and testing difficult, and work in this area is still at an early stage. This has led to approaches which achieve a compromise between integration and acoustic optimisation by direct bonding to an intermediate connector which is then interfaced with the supporting electronics [50].

5.1.2 **PVDF Piezoelectric Film**

PVDF is a piezoelectric polymer which comes in the form of a thick film and can be used in place of bulk ceramic in the same transducer configuration. Its material properties make it inefficient in TX mode compared to PZT, but its RX properties are excellent. Its acoustic impedance is also a much better match to tissue, making it a contender in applications where a good acoustic impedance match is critical.

PVDF transducers typically have higher capacitance than PZT because of their reduced thickness, requiring additional work on the electronics. However, the much lower permittivity allows the film to be coupled to substrate electrodes without the use of conductive epoxies or adhesives, leading to the development of direct-bonded solutions impossible with bulk materials [11]. Capacitive coupling through non-conductive epoxies also allows the electrodes to be patterned on the interconnecting flexible circuit rather than the PVDF [23]. Array element patterns can thus be designed and manufactured through flexible circuit processes, simplifying prototyping and allowing an easier pathway to mass manufacturing.

5.1.3 CMUTs

CMUTs are an alternative approach to US transducers, able to be manufactured on the same semiconductor fabrication line as ASICs and other ICs. This brings the potential to integrate a complete US imaging system in a single IC. As mentioned previously, conventional transducers operate in the thickness mode when generating and detecting US. CMUTs operate in a flexural mode and generate charge rather than voltage output. Additionally, they need a high-voltage DC bias, so circuits designed to support them do not port well to other transducer types.

The integration-led aspects of CMUT design have led to the development of a relatively large number of IC designs, with varying approaches to system partitioning between the base station and ASICs integrated with or close to the transducer. Because of the physical properties of membranes, individual CMUTs are often small relative to the ultrasonic array elements required for imaging. It is thus standard to use them in sub-arrays, with switching networks integrated in the system design to allow the correct connectivity [15]. The combined complexity of these networks and the necessary drive circuitry has led most implementations to integrate only the networks and preliminary gain stages with the probe, with the remainder of the receive system in the base station [30, 51].

5.1.4 PMUTs

As the manufacturing approach to PMUTs is, in some ways, a compromise between bulk ceramic approaches and CMUTs, it is logical that the supporting electronic systems are also a compromise. As with CMUTs, individual PMUTs are often quite small relative to the area of each array element, so they are usually used in sets connected electrically in parallel. However, a major advantage is that their piezoelectric film layer allows the same TX and RX electronics to be used as in conventional bulk ceramic arrays, though operating at lower voltages. The very limited thickness of the piezoelectric films in the PMUT membranes reduces the drive voltage to <10 V, simplifying the analogue TX circuitry and reducing electrical coupling levels [42].

The need for access to piezoelectric thin film deposition and IC fabrication facilities has led to the development of PMUTs and their associated electronics in only a few key centres, but μ US systems have nevertheless been developed with both TX and RX capabilities [24]. Systems have also been implemented successfully to integrate PMUT arrays with a TX and RX IC for fingerprint sensing [19]. In this implementation, the flexural mode operation of the PMUTs allowed the ASIC to be integrated directly under the PMUT array without affecting the ultrasonic performance of the transducers, allowing very high electronic density.

6 Conclusions and Future Work

MicroUS is already being applied in key areas of biosensing and has great potential for broader applications as the field matures. The inherently small dimensions of μ US devices and their suitability for use in minimally invasive medical devices lead directly to the need for supporting circuitry which can be fully integrated with millimetre-scale devices. Research in this field is already exploring various applications of high-density circuit design as well as IC-oriented interconnections. Related research and development in the mobile phone, medical device and LOAC sectors are also providing new solutions which may lead to further development in the μ US domain.

The current state of the art in electronic systems for μ US has begun a migration to IC technologies. However, the unusual combination of specifications for high voltage, mid frequency, low noise and high SNR has led to a piecemeal approach aimed at particular applications, and true generalisation has not yet been achieved, either in the research domain or in commercial offerings. Partly because of this, μ US is still a niche application, limiting resources available to fund ASIC development.

6.1 Future Work

There has been a significant progress in the past few years in terms of μ US imaging in general and in the development of ICs. μ US scanners are now available commercially from VisualSonics (Fujifilm VisualSonics Inc., Toronto, Canada) with f_c up to 70 MHz, with established capabilities in small animal imaging and early work in humans. Vermon (Vermon S.A., Tours, France) currently offers transducers up to 18 MHz, and they have published research with $f_c = 30$ MHz [36], suggesting the intention to enter the μ US market at an appropriate time. With the wider availability of μ US imaging, new applications within the imaging and biosensing domains are likely to appear with increasing frequency over the next decade.

Current research is investigating the integration of US sensors and circuitry into minimally invasive medical devices such as the surgical needles previously mentioned as well as endoscopy capsules [29], an application currently dominated by optical techniques. These and other similar approaches require particularly compact form factors which emphasise the need for miniaturised electronics compatible with direct connection to the active sensors. Work in this area will need to address questions of power management, electronic space efficiency and system partitioning. This will provide the potential for creation of stand-alone μ US devices which are small enough and versatile enough to find many applications within the biosensing domain, as well as for non-destructive testing in confined spaces and other potential industrial applications.

Outside traditional biomedical imaging, continuous wave and contact measurement modes such as Doppler, SAW and fingerprint detection often feature reduced

electronic complexity which allows further miniaturisation, creating the potential for sensing devices which can be implanted at the point of interest [38, 41].

μ US has consistently enjoyed attention from research and commercial interests as it has an appealing balance between imaging resolution and tissue penetration without the need for ionising radiation. However, the necessary integrated electronics to fully realise the biosensing potential of this modality have lagged somewhat behind the ultrasonic and mechanical developments, presenting a worthy engineering challenge. Through the existing approaches and the adoption of new interconnect and integration technologies driven by the microelectronics sector, practical, miniaturised μ US systems are clearly feasible in the near future.

References

1. AFE5807 | Ultrasound | Medical Analog Front End | Description & parameters [WWW Document], (2017), URL <http://www.ti.com/product/afe5807/description?keyMatch=AFE5807&tisearch=Search-EN>. Accessed 10 Jan 2017
2. M. Analoui, J.D. Bronzino, D.R. Peterson, *Medical Imaging: Principles and Practices* (CRC Press, Boca Raton, 2012)
3. H. Azhari, *Basics of Biomedical Ultrasound for Engineers* (Wiley-IEEE, Hoboken, 2009)
4. A.L. Bernassau, D. Flynn, F. Amalou, M.P.Y. Desmulliez, S. Cochran, Techniques for wirebond free interconnection of piezoelectric ultrasound arrays operating above 50 MHz, in *Ultrasonics Symposium (IUS), 2009 IEEE International. Presented at the Ultrasonics Symposium (IUS), 2009 IEEE International*, (2009), pp. 1–4. <https://doi.org/10.1109/ULTSYM.2009.5441700>
5. E. Berry, S. Kelly, J. Hutton, H. Lindsey, J. Blaxill, J. Evans, J. Connelly, J. Tisch, G. Walker, U. Sivananthan, M. Smith, Intravascular ultrasound-guided interventions in coronary artery disease: a systematic literature review, with decision-analytic modelling, of outcomes and cost-effectiveness. *Health Technol. Assess.* **4**, 1–117 (2000)
6. J.A. Brown, F.S. Foster, A. Needles, E. Cherin, G.R. Lockwood, Fabrication and performance of a 40-MHz linear Array based on a 1-3 composite with geometric elevation focusing. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **54**, 1888–1894 (2007). <https://doi.org/10.1109/TUFFC.2007.473>
7. J.A. Brown, G.R. Lockwood, A digital beamformer for high-frequency annular arrays. *Ultrason. Ferroelectr. Freq. Control IEEE Trans. On* **52**, 1262–1269 (2005). <https://doi.org/10.1109/TUFFC.2005.1509785>
8. Chang-Hong Hu, Fan Zheng, Yi Huang, J.M. Cannata, K.K. Shung, Ping Sun, Design of a 64-channel digital high frequency linear array ultrasound imaging beamformer on a massively parallel processor array, in *Ultrasonics Symposium, 2008. IUS 2008. IEEE. Presented at the Ultrasonics Symposium, 2008. IUS 2008. IEEE*, (2008), pp. 1266–1269. <https://doi.org/10.1109/ULTSYM.2008.0306>
9. Chang-Hong Hu, K.A. Snook, P.-J. Cao, K. Kirk Shung, High-frequency ultrasound annular array imaging. Part II: Digital beamformer design and imaging. *Ultrason. Ferroelectr. Freq. Control IEEE Trans. On* **53**, 309–316 (2006). <https://doi.org/10.1109/TUFFC.2006.1593369>
10. R.S.C. Cobbold, *Foundations of Biomedical Ultrasound* (Oxford University Press, Oxford, United Kingdom, 2006)
11. V. Daeichin, C. Chen, Q. Ding, M. Wu, R. Beurskens, G. Springeling, E. Noothout, M.D. Verweij, K.W.A. van Dongen, J.G. Bosch, A.F.W. van der Steen, N. de Jong, M. Pertijs, G. van Soest, A broadband Polyvinylidene difluoride-based hydrophone with integrated readout circuit for intravascular photoacoustic imaging. *Ultrasound Med. Biol.* **42**, 1239–1243 (2016). <https://doi.org/10.1016/j.ultrasmedbio.2015.12.016>

12. EUS Imaging [WWW Document], (2017), URL http://www.eusimaging.com/reference/papers/artifacts/artifacts_print.html. Accessed 18 Jan 2017
13. A. Fatehullah, S. Sharma, I.P. Newton, A.J. Langlands, H. Lay, S.A. Nelson, R.K. McMahon, N. McIlvenny, P.L. Appleton, S. Cochran, I.S. N athke, Increased variability in ApcMin/+ intestinal tissue can be measured with microultrasound. *Sci Rep* **6**, 29570 (2016). <https://doi.org/10.1038/srep29570>
14. J.O. Fiering, P. Hultman, W. Lee, E.D. Light, S.W. Smith, High-density flexible interconnect for two-dimensional ultrasound arrays. *Ultrason. Ferroelectr. Freq. Control IEEE Trans. On* **47**, 764–770 (2000). <https://doi.org/10.1109/58.842067>
15. R. Fisher, K. Thomenius, R. Wodnicki, R. Thomas, S. Cogan, C. Hazard, W. Lee, D. Mills, B. Khuri-Yakub, A. Ergun, et al, Reconfigurable arrays for portable ultrasound, in *Proceedings of IEEE Ultrasonics Symposium*, (2005), pp. 495–499
16. H.M. Garc a-Garc a, V. Klauss, N. Gonzalo, S. Garg, Y. Onuma, C.W. Hamm, W. Wijns, J. Shannon, P.W. Serruys, Relationship between cardiovascular risk factors and biomarkers with necrotic core and atheroma size: a serial intravascular ultrasound radiofrequency data analysis. *Int. J. Card. Imaging* **28**, 695–703 (2012). <https://doi.org/10.1007/s10554-011-9882-6>
17. U. Guler, A. Bozkurt, 5G-3 A low-noise front-end circuit for 2D cMUT arrays, in *2006 IEEE Ultrasonics Symposium. Presented at the 2006 IEEE Ultrasonics Symposium*, (2006), pp. 689–692. <https://doi.org/10.1109/ULTSYM.2006.186>
18. S.R. Heron, R. Wilson, S.A. Shaffer, D.R. Goodlett, J.M. Cooper, Surface acoustic wave nebulization of peptides as a microfluidic Interface for mass spectrometry. *Anal. Chem.* **82**, 3985–3989 (2010). <https://doi.org/10.1021/ac100372c>
19. D.A. Horsley, Y. Lu, H.Y. Tang, X. Jiang, B.E. Boser, J.M. Tsai, E.J. Ng, M.J. Daneman, Ultrasonic fingerprint sensor based on a PMUT array bonded to CMOS circuitry, in *2016 IEEE International Ultrasonics Symposium (IUS). Presented at the 2016 IEEE International Ultrasonics Symposium (IUS)*, (2016), pp. 1–4. <https://doi.org/10.1109/ULTSYM.2016.7728817>
20. J.W. Hunt, M. Arditi, F.S. Foster, Ultrasound transducers for pulse-Echo medical imaging. *IEEE Trans. Biomed. Eng. BME* **30**, 453–481 (1983)
21. Y. Jiang, C. Meggs, T. Button, G. Schiavone, M.P.Y. Desmulliez, Z. Qiu, S. Mahboob, R. McPhillips, C.E.M. D emor e, G. Casey, S. Eljamel, S. Cochran, D.R. Sanmartin, 15 MHz single element ultrasound needle transducers for neurosurgical applications, in *2014 IEEE International Ultrasonics Symposium. Presented at the 2014 IEEE International Ultrasonics Symposium*, (2014), pp. 687–690. <https://doi.org/10.1109/ULTSYM.2014.0169>
22. S. Kelly, K.M. Harris, E. Berry, J. Hutton, P. Roderick, J. Cullingworth, L. Gathercole, M.A. Smith, A systematic review of the staging performance of endoscopic ultrasound in gastro-oesophageal carcinoma. *Gut* **49**, 534–539 (2001). <https://doi.org/10.1136/gut.49.4.534>
23. J.A. Ketterling, O. Aristizabal, D.H. Turnbull, F.L. Lizzi, Design and fabrication of a 40-MHz annular array transducer. *Ultrason. Ferroelectr. Freq. Control IEEE Trans. On* **52**, 672–681 (2005). <https://doi.org/10.1109/TUFFC.2005.1428050>
24. I. Kim, H. Kim, F. Griggio, R.L. Tutwiler, T.N. Jackson, S. Trolier-McKinstry, K. Choi, CMOS ultrasound transceiver Chip for high-resolution ultrasonic imaging systems. *IEEE Trans. Biomed. Circuits Syst.* **3**, 293–303 (2009). <https://doi.org/10.1109/TBCAS.2009.2023912>
25. R. Krimholtz, D.A. Leedom, G.L. Matthaei, New equivalent circuits for elementary piezoelectric transducers. *Electron. Lett.* **6**, 398–399 (1970)
26. W.K. Law, L.A. Frizzell, F. Dunn, Determination of the nonlinearity parameter B/a of biological media. *Ultrasound Med. Biol.* **11**, 307–318 (1985)
27. H.S. Lay, B.F. Cox, M. Sunoqrot, C.E.M. D emor e, I. N athke, T. Gomez, S. Cochran, Microultrasound characterisation of *ex vivo* porcine tissue for ultrasound capsule endoscopy. *J. Phys. Conf. Ser.* **797**, 012003 (2017). <https://doi.org/10.1088/1742-6596/797/1/012003>
28. H.S. Lay, G.R. Lockwood, A low cost receive beamformer for a high frequency annular array, in *Ultrasonics Symposium (IUS), 2011 IEEE International. Presented at the Ultrasonics Symposium (IUS), 2011 IEEE International*, (2011), pp. 462–465. <https://doi.org/10.1109/ULTSYM.2011.0111>

29. H.S. Lay, Y. Qiu, M. Al-Rawhani, J. Beeley, R. Poltarjonoks, V. Seetohul, D. Cumming, S. Cochran, G. Cummins, M.P.Y. Desmulliez, M. Wallace, S. Trolrier-McKinstry, R. McPhillips, B.F. Cox, C.E.M. Demore, Progress towards a multi-modal capsule endoscopy device featuring microultrasound imaging, in *2016 IEEE International Ultrasonics Symposium (IUS)*. Presented at the 2016 IEEE International Ultrasonics Symposium (IUS), (2016), pp. 1–4. <https://doi.org/10.1109/ULTSYM.2016.7728692>
30. J. Lim, C. Tekes, F.L. Degertekin, M. Ghovanloo, Towards a reduced-wire Interface for CMUT-based intravascular ultrasound imaging systems. *IEEE Trans. Biomed. Circuits Syst.*, 1–11 (2016). <https://doi.org/10.1109/TBCAS.2016.2592525>
31. S.-C.S. Lin, X. Mao, T. Jun Huang, Surface acoustic wave (SAW) acoustophoresis: Now and beyond. *Lab Chip* **12**, 2766–2770 (2012). <https://doi.org/10.1039/C2LC90076A>
32. G.R. Lockwood, J.W. Hunt, F.S. Foster, The design of protection circuitry for high-frequency ultrasound imaging systems. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **38**, 48–55 (1991). <https://doi.org/10.1109/58.67834>
33. G.R. Lockwood, D.H. Turnbull, D.A. Christopher, F.S. Foster, Beyond 30 MHz [applications of high-frequency ultrasound imaging]. *IEEE Eng. Med. Biol. Mag.* **15**, 60–71 (1996). <https://doi.org/10.1109/51.544513>
34. Y. Lu, H.Y. Tang, S. Fung, B.E. Boser, D.A. Horsley, Short-range and high-resolution ultrasound imaging using an 8 MHz aluminum nitride PMUT array, in *2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*. Presented at the 2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS), (2015), pp. 140–143. <https://doi.org/10.1109/MEMSYS.2015.7050905>
35. R.G. Maev, E.Y. Bakulin, E.Y. Maeva, F.M. Severin, High resolution ultrasonic method for 3D fingerprint representation in biometrics, in *Acoustical Imaging*, (Springer, Dordrecht, 2008), pp. 279–285
36. S. Michau, P. Mauchamp, R. Dufait, Piezocomposite 30MHz linear array for medical imaging: design challenges and performances evaluation of a 128 elements array, in *Ultrasonics Symposium, 2004 IEEE*. Presented at the Ultrasonics Symposium, 2004 IEEE, vol. 2 (2004), pp. 898–901. <https://doi.org/10.1109/ULTSYM.2004.1417880>
37. I.G. Mina, H. Kim, I. Kim, S.K. Park, K. Choi, T.N. Jackson, R.L. Tutwiler, S. Trolrier-McKinstry, High frequency piezoelectric MEMS ultrasound transducers. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **54**, 2422–2430 (2007). <https://doi.org/10.1109/TUFFC.2007.555>
38. T.A. Nappholz, H.L. Valenta Jr., S.M. Maas, K. Koestner, Method and apparatus for chronically monitoring the hemodynamic state of a patient using doppler ultrasound. US5188106 A. (1993)
39. C.G. Oakley, Calculation of ultrasonic transducer signal-to-noise ratios using the KLM model. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **44**, 1018–1026 (1997). <https://doi.org/10.1109/58.655627>
40. M. Pertijs, C. Chen, S. Raghunathan, Z. Yu, M. ShabaniMotlagh, Z. Chen, Z. y. Chang, E. Noothout, S. Blaak, J. Ponte, C. Prins, H. Bosch, M. Verweij, N. de Jong, Low-power receive electronics for a miniature real-time 3D ultrasound probe, in *2015 6th International Workshop on Advances in Sensors and Interfaces (IWASI)*. Presented at the 2015 6th International Workshop on Advances in Sensors and Interfaces (IWASI), (2015), pp. 235–238. <https://doi.org/10.1109/IWASI.2015.7184963>
41. V.C. Protopappas, D.A. Baga, D.I. Fotiadis, A.C. Likas, A.A. Papachristos, K.N. Malizos, An ultrasound wearable system for the monitoring and acceleration of fracture healing in long bones. *IEEE Trans. Biomed. Eng.* **52**, 1597–1608 (2005). <https://doi.org/10.1109/TBME.2005.851507>
42. Y. Qiu, J.V. Gigliotti, M. Wallace, F. Griggio, C.E.M. Demore, S. Cochran, S. Trolrier-McKinstry, Piezoelectric micromachined ultrasound transducer (PMUT) arrays for integrated sensing, actuation and imaging. *Sensors* **15**, 8020–8041 (2015). <https://doi.org/10.3390/s150408020>

43. S.B. Raghunathan, D. Bera, C. Chen, S. Blaak, C. Prins, M.A.P. Pertijs, J.G. Bosch, N. de Jong, M.D. Verweij, Design of a miniature ultrasound probe for 3D transesophageal echocardiography, in *2014 IEEE International Ultrasonics Symposium. Presented at the 2014 IEEE International Ultrasonics Symposium*, (2014), pp. 2091–2094. <https://doi.org/10.1109/ULTSYM.2014.0521>
44. K. Ranganathan, M.K. Santy, T.N. Blalock, J.A. Hossack, W.F. Walker, Direct sampled I/Q beamforming for compact and very low-cost ultrasound imaging. *Ultrason. Ferroelectr. Freq. Control IEEE Trans. On* **51**, 1082–1094 (2004). <https://doi.org/10.1109/TUFFC.2004.1334841>
45. N.K. Ratha, V. Govindaraju, *Advances in Biometrics: Sensors, Algorithms and Systems* (Springer Science & Business Media, London, 2008)
46. M. Sautto, D. Leone, A. Savoia, D. Ghisu, F. Quaglia, G. Caliano, A. Mazzanti, A CMUT transceiver front-end with 100-V TX driver and 1-mW low-noise capacitive feedback RX amplifier in BCD-SOI technology, in *ESSCIRC 2014 – 40th European Solid State Circuits Conference (ESSCIRC)*. Presented at the *ESSCIRC 2014 – 40th European Solid State Circuits Conference (ESSCIRC)*, (2014), pp. 407–410. <https://doi.org/10.1109/ESSCIRC.2014.6942108>
47. R.M. Schmitt, W.G. Scott, R.D. Irving, J. Arnold, C. Bardons, D. Halpert, L. Parker, Ultrasonic imaging of fingerprints using acoustical impediography, in *IEEE Ultrasonics Symposium, 2004. Presented at the IEEE Ultrasonics Symposium, 2004*, vol. 1 (2004), pp. 680–688. <https://doi.org/10.1109/ULTSYM.2004.1417814>
48. S. Sharma, T. Ytterdal, Low noise front-end amplifier design for medical ultrasound imaging applications, in *2012 IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC)*. Presented at the *2012 IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC)*, (2012), pp. 12–17. <https://doi.org/10.1109/VLSI-SoC.2012.7332069>
49. K.K. Shung, *Diagnostic Ultrasound: Imaging and Blood Flow Measurements* (Taylor and Francis, Boca Raton, 2006)
50. E.A. Simpson, H.S. Lay, G.R. Lockwood, Novel interconnection and fabrication method for high-frequency ultrasound arrays, (2010), pp. 76290P–76290P–13. <https://doi.org/10.1117/12.845383>
51. J. Song, S. Jung, Y. Kim, K. Cho, B. Kim, S. Lee, J. Na, I. Yang, O. Kwon, D. Kim, Reconfigurable 2D cMUT-ASIC arrays for 3D ultrasound image, (2012), pp. 83201A–83201A–6. <https://doi.org/10.1117/12.911263>
52. T.L. Szabo, *Diagnostic Ultrasound Imaging: Inside Out* (Academic Press, Burlington, 2004)
53. H. Tang, Y. Lu, S. Fung, J.M. Tsai, M. Daneman, D.A. Horsley, B.E. Boser, Pulse-echo ultrasonic fingerprint sensor on a chip, in *2015 Transducers – 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*. Presented at the *2015 Transducers – 2015 18th International Conference on Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS)*, (2015a), pp. 674–677. <https://doi.org/10.1109/TRANSDUCERS.2015.7181013>
54. H.Y. Tang, D. Seo, U. Singhal, X. Li, M.M. Maharbiz, E. Alon, B.E. Boser, Miniaturizing ultrasonic system for portable health care and fitness. *IEEE Trans. Biomed. Circuits Syst.* **9**, 767–776 (2015b). <https://doi.org/10.1109/TBCAS.2015.2508439>
55. K.E. Thomenius, Evolution of ultrasound beamformers, in *Ultrasonics Symposium, 1996. Proceedings.*, 1996 IEEE. Presented at the *Ultrasonics Symposium, 1996. Proceedings*, 1996 IEEE, vol. 2 (1996), pp. 1615–1622. <https://doi.org/10.1109/ULTSYM.1996.584398>
56. J. Tillak, J. Yoo, A 23 uW digitally controlled pMUT interface circuit for Doppler ultrasound Imaging, in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*. Presented at the *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, (2015), pp. 1618–1621. <https://doi.org/10.1109/ISCAS.2015.7168959>
57. Transesophageal Echocardiography (TEE) [WWW Document], (2017), URL http://www.heart.org/HEARTORG/Conditions/HeartAttack/SymptomsDiagnosisofHeartAttack/Transesophageal-Echocardiography-TEE_UCM_441655_Article.jsp#.WH9SCRuLSUK. Accessed 18 Jan 2017

58. Ultrasound | Analog Devices [WWW Document], (2016), URL <http://www.analog.com/en/products/application-specific/medical/ultrasound.html>. Accessed 19 Dec 2016
59. Ultrasound Imaging – Maxim [WWW Document], (2017), URL <https://www.maximintegrated.com/en/markets/healthcare/imaging.html>. Accessed 10 Jan 2017
60. D. Vilkomerson, T. Chilipka, Implantable Doppler system for self-monitoring vascular grafts, in *IEEE Ultrasonics Symposium, 2004. Presented at the IEEE Ultrasonics Symposium, 2004*, vol. 1 (2004), pp. 461–465. <https://doi.org/10.1109/ULTSYM.2004.1417762>
61. R.M. Vlad, S. Brand, A. Giles, M.C. Kolios, G.J. Czarnota, Quantitative ultrasound characterization of responses to radiotherapy in cancer mouse models. *Clin. Cancer Res.* **15**, 2067–2075 (2009). <https://doi.org/10.1158/1078-0432.CCR-08-1970>
62. P.N. Wells, Advances in ultrasound: from microscanning to telerobotics. *Br. J. Radiol.* **73**, 1138–1147 (2000)

Chapter 9

High-Density CMOS Neural Probes

Bogdan Raducanu, Carolina Mora Lopez, and Srinjoy Mitra

1 Introduction

Neural signals originate in neurons which are specialized cells that process and transmit information through the use of electrical pulses. Neurons are connected together via synapses that transmit information by chemical neurotransmitters in their ion channels. Together, neurons are forming complex neural networks, which build the central nervous system (including the brain and spinal cord) and the peripheral nervous system. As the structural and functional properties of the nervous system are rigorously investigated in different branches of neuroscience, various tools have been developed in studying both the chemical and the electrical properties of neurons and synapses.

On the electrical side, different signals can be captured at different levels of proximity from their source, depending on how the signals originating in the neurons are accessed. The electrical recordings may be intracellular or extracellular (depending on whether the signals are captured from inside the cell or from outside), minimally invasive (when recorded outside the cortex), or even completely noninvasive (when recorded outside the body).

B. Raducanu (✉) • C.M. Lopez
IMEC, Leuven, Belgium
e-mail: bogdan.raducanu@imec.be

S. Mitra
School of Engineering, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK

1.1 Applications

Recording the electrical activity neural signals in various forms can provide multiple types of insights. Starting from the outside of the brain, EEG recording can be used for the study and diagnosis of epilepsy, monitoring anesthesia levels and blood perfusion, or in intensive care units, for brain monitoring during an event. EEG has found applications in neuroscience, cognitive psychology, and brain machine interface. Furthermore, it can be used together with other noninvasive neuroimaging techniques such as fMRI, complimenting each other. Accessing the brain from the inside, using minimally invasive ECoG electrodes placed on the surface of the brain, can provide similar information as EEG, but with increased resolution and sensitivity. The applications include epilepsy diagnosis and, more recently, brain computer interfaces.

Electrical recording directly from within the cells and tissues is the domain of electrophysiology. Such invasive recording from the central or the peripheral nervous system is one of the primary sources of research in experimental neuroscience. In the human brain, implants can be also used for therapeutic applications in epilepsy, Parkinson's, etc. Furthermore, whether located in the brain or in the peripheral nervous system, neural implants can be used as brain machine interface to control prosthetic limbs. The primary focus of this chapter is on invasive neural recording and its requirements, problems, and some solutions.

1.2 Types of Recording

Intracellular recordings require delicate tools such as glass microelectrodes in order to access the inside of the cell. Given the size of the cell in the orders of micrometers, the operation of accessing the cell is difficult and time-consuming. The tools used in intracellular recording allow for studying of neurons themselves and are normally applied at small scales, as neurons need to be accessed individually. Furthermore, intracellular recording can be mostly done in anesthetized animals, leaving aside the rich information content in behaving animals.

However, as neurons form complex networks, it is required to study larger populations to understand the functionality of such complex systems. Fortunately, the neural signals can be accessed from the outside of the cell as well, by placing electrodes nearby. The distances and obstacles between the electrodes and the brain as well as the size of these electrodes determine the types of signals that may be captured.

At the closest level, at distances and electrode sizes comparable to the neurons themselves (tens of micrometer), the action potential (AP) signals can be captured (Fig. 9.1). An action potential signal is a short event consisting of a rapid rise and fall of the cell membrane potential of a neuron, which shows a consistent temporal trajectory for a given neuron. The recorded electrical signal consists of

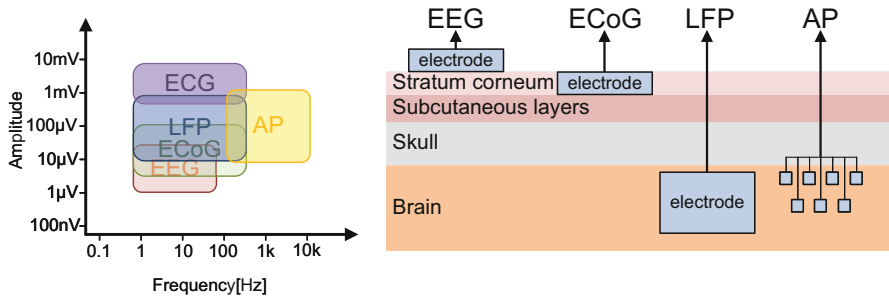


Fig. 9.1 Amplitude and frequency domain for different types of neural signals and capturing locations and methods for neural signals

the firing of individual neurons or a superposition of multiple neurons, usually with distinguishable temporal characteristics. An action potential (or spike) indicates that a specific neuron has fired. Various mathematical tools can be utilized to perform “spike sorting” that can detect individual neurons among the activities captured.

Increasing the size of the electrode produces an averaging of the nearby action potentials and instead allows for the recording of what is known as the local field potential (LFP). These signals carry components at lower frequencies than action potentials and provide an indication of the state of the nearby neural networks.

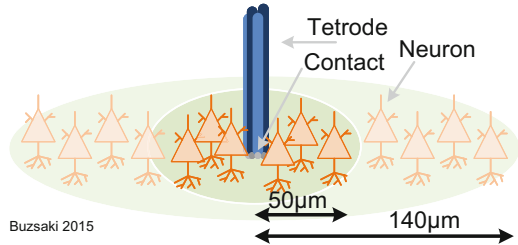
Given the signals desired to be recorded, multiple types of electrodes and systems can be envisioned and designed in order to record the specific information, depending on the required scenario. Methods that record large populations of neurons can provide information relevant for certain specific measurements; however, they provide limited details that can be used in understanding the workings of a brain.

For understanding the underlying function of the neural network, it is often relevant to access the AP of individual neurons, with the goal of recording from as many as possible. Although there is significant demand from neuroscientists for large-scale neural recording, the devices currently in widespread usage may still be rudimentary and not take advantage of the tremendous advantages that the semiconductor industry has made and could offer.

1.3 Types of Recording Devices

The simplest device used to record extracellular neural signals is a plain wire. A very thin insulated metal wire can be inserted into the brain, having only the tip exposed to the nearby neurons, forming an electrode. With diameters reaching close to the size of a neuron (μm up to tens of μm), the tip of such a wire within the cortical structures is bound to end up in proximity of neurons, capturing their action potentials. While this method is simple and low cost and can be scaled to multiple wires, the disadvantages come from the sheer complexity and size of a brain.

Fig. 9.2 A tetrode capturing signals from neurons. Neurons within a 50 μm distance from the tetrode tip where the contacts are located can be sorted, while neurons up to 140 μm can be captured



Practically, the number of wires placed in a brain is significantly low compared to the number of neurons, and a wire can only record signals at its tip.

To improve this method, multiple wires can be packed together; specifically, using four wires packed together will form a “tetrode,” a device terminating with four-contact electrodes at the tip (Fig. 9.2). Such a device improves signal quality and allows for the spatial localization of the nearby neurons, but will not improve the scale of the recording significantly [1]. The amplitude of neurons in close proximity is high enough to allow for spike sorting (i.e., assigning the spike to a specific neuron based on its individual shape). Neurons at larger distances may be recorded, but their amplitude is small enough to prevent a good-quality sorting. Furthermore, given their distance to the tetrode relative to its size, localization is not accurately possible anymore.

Extrapolating from these numbers and taking into account the brain of a small mammal such as a mouse or rat, the number of such tetrodes required to record every neuron becomes impractically large [2].

Given the limitations of wires, other methods of capturing the action potentials have been developed. As wires are insulated and can only capture the signals present at their tip, trying to pack more electrodes or contacts within the same volume as a wire can provide additional information.

Such devices, called neural probes, can be manufactured with various methods at dimensions comparable to the previously described tetrodes. Although high variability in shape, size, and materials exists, they are similar in their architecture and structure. Figure 9.3 shows the usual structure of such a probe, containing three key elements. Neural probes are usually built out of an insulating material, which contains exposed electrodes and should be partially inserted into the brain. To access the signals captured by the exposed electrodes, insulated conducting wires carry the signals outside the brain. Lastly, the non-implantable part of the probe contains electrical contacts to connect to an electronic readout system.

Depending on the spatial distribution of the electrodes, one form devices may take is a two-dimensional array, where electrodes are disposed in a planar manner. Another approach is to have a longitudinal device which is covered with electrodes along its length, similar to a wire. Furthermore, different combinations of the two arrangements can be envisioned, providing more flexibility in the spatial distribution of the recorded neurons [3–7].

The goal of a neural probe is to record as many neurons as possible using a much larger number of electrodes (compared to traditional methods), while also being as

small as possible to reduce the damage it causes to a brain when implanted. These two requirements are generally in contradiction, particularly when a certain signal quality is required. The materials used to build neural probes pose difficulties as well: since the device is implanted in a brain, it needs to have minimal reactivity with the surrounding tissue. This is not an easy task, as few materials can achieve this biocompatibility.

As the size of electrodes reduces, their number increases. Since the probes are generally very long compared to their cross section (high aspect ratio), the number of electrodes will be capped by an interconnection limitation – the number of wires that can be fitted without increasing the cross section.

One of the most modern and advanced ways to build neural probes is to use semiconductor technologies used in fabricating integrating circuits. This leverages the advanced processes’ lithographic methods which can produce shapes of μm or even nm in size.

Similar to the architecture of the probe shown in Fig. 9.3, an integrated circuit is built onto an insulating substrate – the bulk silicon, in which various materials are deposited. These materials can be used to make semiconductor devices such as transistors and interconnection lines, together forming an integrated circuit, with the most popular ones being the CMOS type.

With current technology, it is possible to fabricate tens or hundreds of wires within the same volume of a traditional wire that the neuroscientist could use 50 years ago. The modern CMOS process allows building circuits such as amplifiers within the same device which leads to a higher integration, making recording devices more compact as well as more cost-effective.

With such possibilities, a complete advanced silicon neural probe can be designed on a single chip, containing the electrodes, the interconnection wires, the required amplifiers and filters, and even analog to digital conversion, ideally allowing a single device to connect the brain to a computer.

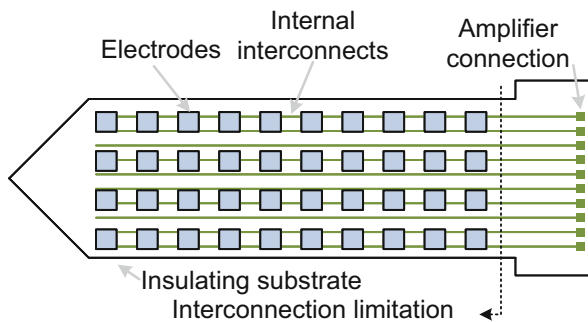


Fig. 9.3 Basic architecture of a neural probe: multiple electrodes are placed on an insulating substrate which internally contains interconnection wires to bring the signal outside of the brain for it to be amplified and analyzed. The limited number of wires compared to the number of electrodes can cause an interconnection limitation

2 CMOS Probes for In Vivo Neural Recording

Silicon neural probes, though not yet a staple in neurophysiology labs, have been extremely promising and already studied very widely. Along with its many advantages, they come with their own set of challenges, particularly in CMOS-based probes. The study of brain activity requires, besides implantable microelectrodes, electronic circuitry for accurately amplifying and conditioning the signals detected at the recording sites. Complex CMOS circuits are now used to execute these functions right on the probe.

While neural probes have become more compact and more dense in order to monitor large populations of neurons, the interfacing electronic circuits have also become smaller and more capable to handle large amounts of parallel recording channels. However, the interconnection of neural probes with application-specific integrated circuits (ASICs) to form fully implantable devices poses important power and area limitations to the circuit design and creates several trade-offs among different circuit blocks and specifications. For instance, implantable systems may dissipate only very low power in order to avoid heating of the surrounding tissue [8], but low-power telemetry usually achieves only limited bandwidths, making the transmission of many recording channels difficult. In the last years, researchers have proposed different kinds of neural interface architectures to deal with such trade-offs. These architectures are usually composed of low-noise neural amplifiers, filters, multiplexers, ADCs, and wired or wireless telemetry circuits. Although the design aspects of these circuit blocks and the important trade-offs related to optimal neural interfaces have been covered by several reviews and tutorials [8–13], a brief overview of this subject will follow.

2.1 *Generic Architecture*

The architecture of an implantable system for neural recording can be implemented in different ways, and this has a big impact on the design constraints of several component blocks. Jochum et al. described the three common architectures of an implantable system: distributed, mixed, and merged [12], as shown in Fig. 9.4. These architectures are composed of implantable neural probes, neural amplifiers, and other electronic circuits such as data converters, signal processors, telemetry transceivers, and power supplies [12]. In the distributed architecture shown in Fig. 9.4a, different circuit components are separated from the probe, making the wires more susceptible to noise coupling. The system reported by Sodagar et al. is an example of such an architecture [14]. Figure 9.4b shows the mixed architecture, in which the array of neural amplifiers and the analog multiplexer are placed very close to the probe (and even in the same substrate [4]), having just a few cables taking the nonsensitive signals to the external circuits. Such architecture was, for example, implemented by Song et al. [15]. The merged architecture in Fig. 9.4c

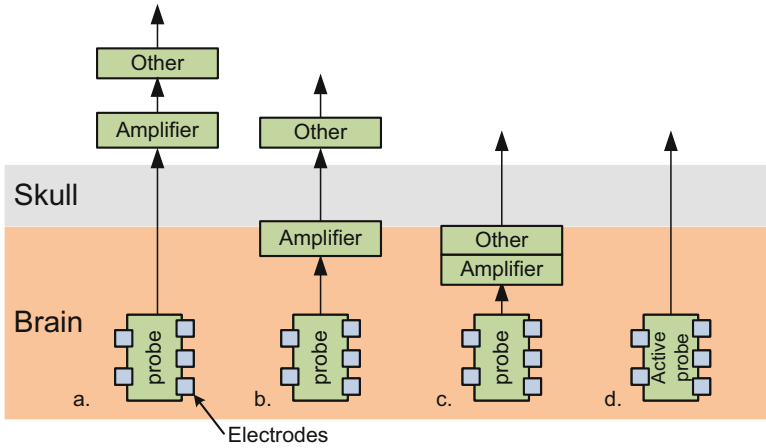
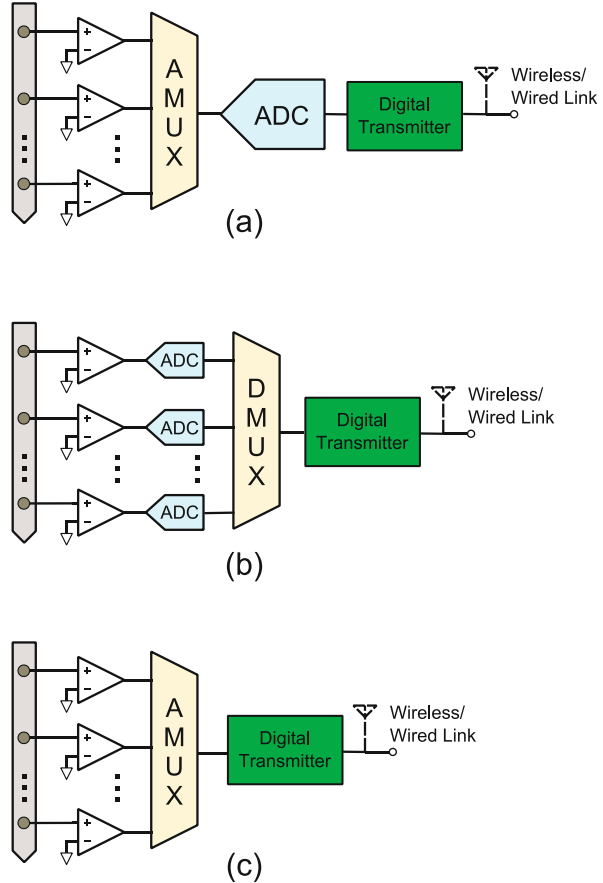


Fig. 9.4 Common architectures of an implantable neural interface (Modified from Jochum et al. [7]). (a) Distributed architecture separating the probe from all other circuits. (b) Mixed architecture minimizing the wires from the probe to the amplifier. (c) Merged architecture combining all assemblies into a compact implantable unit. (d) CMOS neural probe

integrates all different system components into a fully implantable package that provides robustness against noise and other external interferences. An example of such an architecture was described by Gosselin et al. [16]. A complete comparison of the advantages and disadvantages of these three architectures can be found in Jochum's work [12]. Finally, a further step in the system integration not yet discussed in Jochum's work is represented in Fig. 9.4d. In this case, a CMOS neural probe integrating the electrodes and signal processing into the same silicon substrate provides minimum cross talk and noise coupling and an optimal digital interface to minimize the package complexity. Examples of such systems have been recently reported by Lopez et al. [3] and Raducanu et al. [17].

The recording circuits inside an ASIC can also have different system-level architectures, which imposes various circuit constraints and trade-offs among the design specifications. Three traditional microsystem architectures have been described by Gosselin et al., and they are shown in Fig. 9.5 [9]. The architectures are built with an array of recording channels, an analog or digital multiplexer, one or several ADCs, and wireless or wired telemetry links. In the first approach (Fig. 9.5a), a fast ADC is shared among different recording channels using a time-division analog multiplexer (e.g., Sodagar et al. [14]). This architecture requires one output buffer per channel to drive the high-speed ADC, which can lead to excessive power consumption. The second approach (Fig. 9.5b) integrates a low-power, low-speed ADC into each recording channel and performs time-division multiplexing in the digital domain (e.g., Gosselin et al. [16]). In this case, the power-consuming buffers are not necessary and the channel-to-channel cross talk can be reduced, but it requires larger chip areas. The last approach (Fig. 9.5c) uses an analog multiplexer

Fig. 9.5 Neural recording microsystem architectures (Modified from Gosselin et al. [9]). **(a)** Analog time-division multiplexer (AMUX) and shared high-speed ADC. **(b)** Low-speed ADC in each channel and digital time-division multiplexer (DMUX). **(c)** Analog time-division multiplexer and analog transmission



to combine different channels into a single analog line that is sent to an external digitizer (e.g., Seung et al. [18]). The purpose of this architecture is to save power and area, but it may produce excessive cross talk if not designed properly. A design optimization technique was proposed by Chae et al. [11], in which the power and area of a multichannel architecture were quantified and analyzed in order to find the optimal number of ADCs for a given number of channels. This approach provides a useful tool to solve the power-area trade-off, considering other design parameters such as overall noise and gain. Apart from these traditional architectures, other innovative architectures have been reported in literature in the last year.

2.2 *System Requirements*

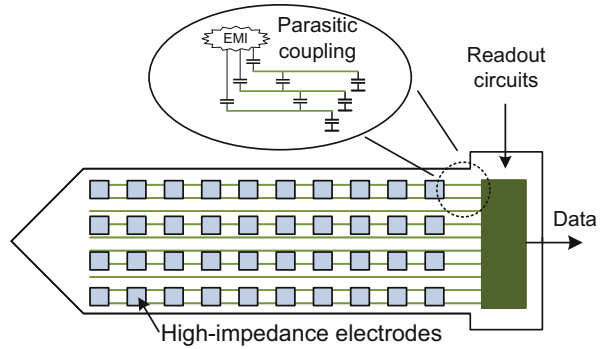
The biggest challenges of neural implants are related to the design of biocompatible, stable, and reliable microprobes, suitable for long-term chronic experiments or treatments. Despite many available designs and techniques, several engineering issues remain unsolved, preventing the widespread use of the current technologies into chronic and medical applications. One of these challenges is the development of reliable and biocompatible low-impedance electrode materials and probe substrate materials, capable of remaining stable for many years. Biocompatibility is also associated with probe dimensions, which can cause tissue displacement and damage during implantation. Thus, microprobes with reduced sizes are required to minimize the tissue response and to extend the lifespan of the implant under chronic conditions.

Regarding the functionality, it is clear from the state of the art that current trends in neural recording interfaces are toward the development of very high-density electrode arrays, capable of recording the activity of hundreds or thousands of neurons at the same time and providing information about different processes in the brain. This involves the design of neural probes with hundreds of recording sites arranged at small pitches, which is currently limited by the number of connecting wires that can be placed in the probe shank width. In order to record the neural signals coming from the electrodes with high SNRs, large arrays of low-noise neural amplifiers need to be integrated in smaller chip areas, including adequate filters for rejecting DC offsets and out-of-band signal components. Ideally, electrode arrays and readout circuitry should be integrated in the same implantable substrate to avoid reliability issues due to the assembly and packaging of hybrid microsystems. However, this imposes severe power constraints as heating of the surrounding tissue must be avoided. Thus, optimal circuit design techniques and architectures are required to further reduce the area and power consumptions of different circuit blocks in a neural interface. Furthermore, power-efficient data management and wireless transmission techniques that do not compromise the performance or functionality of the system are still to be developed.

2.3 *Signal Quality*

The main goal of the current neural probe technologies is to minimize the size of the implants while including as many recording site electrodes that can be placed and routed in the same substrate due to the width of the interconnections between the recording tips and the external connectors [2]. Achieving massively parallel neuronal recordings requires the development of high-aspect-ratio neural probes with a high density of recording sites, so that the recording yield is increased while the tissue displacement and damage are minimized. Additionally, smaller recording electrodes are required, which usually exhibit very high impedances.

Fig. 9.6 Capacitive and parasitic coupling of *parallel metal lines* in a high-aspect-ratio passive neural probe



Hence, as seen in Fig. 9.6, such high-aspect-ratio passive neural probes suffer from several drawbacks: (i) noise or electromagnetic interference (EMI) pickup due to the long metal wires connecting the recording sites to the electronic amplifiers, (ii) signal attenuation due to parasitic capacitances, (iii) cross talk between adjacent channels due to the proximity between metal wires, and (iv) a large number of cables connecting the probe to the electronic readout equipment, making certain chronic experiments difficult. Because electrical recording from single neurons is invasive, monitoring large numbers of neurons using large implanted devices inevitably increases the tissue damage; thus, there exists a trade-off between the probe size and the number of recording sites. Although existing neural probes can record from large numbers of neurons, they are still limited in the number of recording electrodes that can be placed and routed in the same substrate due to the width of the interconnections between the recording tips and the external connectors.

Several approaches have been proposed in literature to achieve high-density electrode arrays and to diminish the cross talk effects in dense interconnects. Du et al. proposed a neural probe with a dense electrode array at the tip (i.e., 64 electrodes) and readout circuitry assembled to the body of the probe [19]. To reduce cross talk in the 64 parallel lines running along the probe shank, electroplating was performed before the measurements. In this way, the impedance of the small electrodes was lowered until a tolerable level was achieved. Although electroplating is a common technique to reduce electrode impedances, this is a time-consuming (and, in some cases, nonreproducible) procedure usually applied on one electrode at a time. Moreover, depending on the used material and technique, the low durability of the plating can limit its use in chronic applications [20]. Bringing the readout circuits closer to the electrode sites (e.g., Olsson et al. [4], Du et al. [19]) also reduces the cross talk and other noise pickup as the length of the sensitive metal interconnections is reduced. However, in very long shanks, cross talk can still be a problem when the metal lines are placed very close to each other. Torfs et al. [21] implemented a switch matrix in the probe shank to effectively reduce the number of metal wires connecting the electrodes to the recording circuits. In that design, a large array of electrodes (i.e., 256) was included in the shank, and only a few of them (i.e., 8) could be digitally selected to be recorded. Connecting a switch matrix

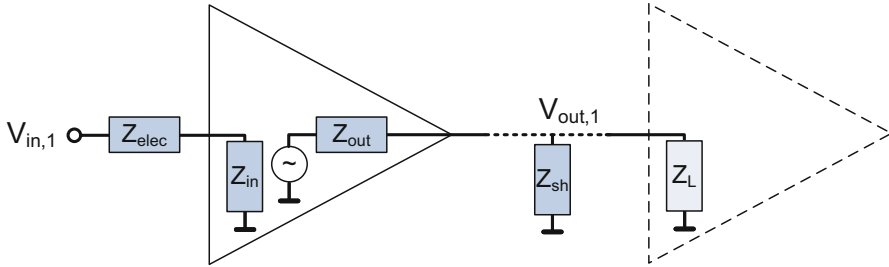


Fig. 9.7 Impedance conversion effect achieved by placing an amplifier very close to the electrode

to high-impedance nodes may, however, increase the parasitic capacitance of the lines and, therefore, increase the signal loss.

An effective way of reducing the electrode impedance is by placing an amplifier or buffer very close to the signal source. As a result, the high-impedance node at the electrode site is converted to a low-impedance node provided by the output impedance of the amplifier. This effect is depicted in Fig. 9.7. In order to guarantee cross talk and signal attenuation reduction, Z_{out} must be lower than Z_{elec} , and Z_{in} must be kept large enough in order not to produce additional signal attenuation. Thus, the proposed method to achieve a very high electrode density (i.e., spatial resolution) consists of the design of a full-CMOS active neural probe integrating in situ amplification under each electrode and a switch matrix in the shank to reduce the number of metal interconnections. As the metalline width and spacing decrease with the technology scaling, modern CMOS technologies allow for the placing of a large amount of parallel lines in single or multiple metal layers. An electrode array covering the full shank area enables the monitoring of neural activity along the entire shank length, thus maximizing the recording yield for the given probe dimensions. When the number of electrodes exceeds the number of recording channels, a switch matrix serves the purpose of selecting the desired brain targets to be recorded. This selectivity must provide enough flexibility to cover large brain areas with high spatial resolution or to distribute the recording sites along the implantation track in an arbitrary manner to be chosen by the user. This proposed solution, however, imposes severe power and area constraints for the design of the in situ amplifier due to the limited area under the electrodes (i.e., dictated by the electrode pitch and shank width) and the low heat dissipation requirements of a neural implant. Therefore, under these constraints, an optimal architecture is needed to still achieve the required low-noise performance for this application.

2.4 Power Budget

The power budget for an implanted ASIC is primarily dictated by the amount of heat that can be dissipated without increasing the temperature of the nearby tissue by 1 °C. Even though several IC design solutions depend on the total

power budget, this is not something that can be conclusively estimated without the complete information on actual neurophysiological setup. Unlike traditional ASICs, an implantable IC has a rather unusual shape and is encompassed in a heterogeneous environment of tissues, blood vessels, bone, fat, etc. Several approximations need to be done to derive a reasonable power budget for the IC. To begin with, for a central nervous system (CNS) implant, the brain is considered as a homogeneous volume of tissue. Suitable assumptions must be made on the part of the ASIC outside the brain and what it is connected to. Generally, this chunk of silicon is held on a PCB and cemented to the skull. It is often assumed that this part of the system is freely floating in air. Finally, one of the biggest challenges is to make judicious assumption on the self-heating of the chip itself. A piece of silicon with active circuits spread around a length of 20 mm will generate various local hotspots depending on the amount of power required.

A detailed thermal simulation could be done using some structural, circuit, and thermal conductivity approximation. An example of such simulation is shown in Fig. 9.8, which follows the approach suggested by Kim et al. [22]. In that work,

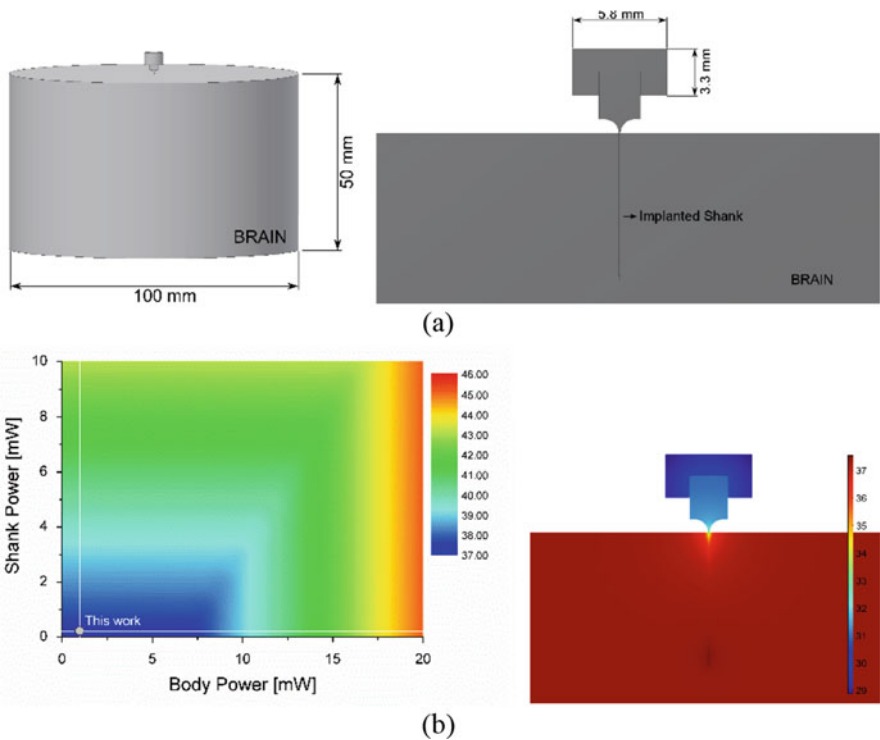


Fig. 9.8 (a) Modeling of the ASIC and the surrounding tissue for the numerical FEM thermal simulations. *Left*: three-dimensional view. *Right*: zoomed-in cross-sectional view of the neural probe and the brain. (b) Computed temperature distribution in the brain and the ASIC. *Left*: maximum temperature at any spot of the brain. *Right*: cross-sectional view for a specific shank and base power

the authors have studied the thermal influence of a powered Utah electrode array implanted in the brain by performing numerical simulations (i.e., finite element analysis) and in vitro and in vivo experimental measurements. They were able to predict that the temperature increased linearly with the power dissipation through the electrode array, with an amount of $0.051\text{ }^{\circ}\text{C}/\text{mW}$. This was an important result to set power dissipation limits for that specific implant. The results in Fig. 9.8b correspond to an active neural probe consisting of an implanted shank with dimensions of $10\text{ mm} \times 100\text{ }\mu\text{m} \times 50\text{ }\mu\text{m}$ and a probe base with dimensions of $3.3\text{ mm} \times 2.9\text{ mm} \times 50\text{ }\mu\text{m}$. The probe base was exposed to air (i.e., at $24\text{ }^{\circ}\text{C}$) and attached to a cylindrical polydimethylsiloxane (PDMS) structure (i.e., 3.3 mm height and 2.9 mm radius), simulating a custom package (see Fig. 9.8a). The volume of the brain tissue was chosen much larger than the implant (i.e., a cylinder of 50 mm radius and 50 mm height), and its boundaries were considered to be at body temperature. At the brain surface contacting with air, it was assumed that there was no heat transfer, thus forcing the heat conduction to occur only through the implant (i.e., worst-case analysis). The heat transfer from the implanted neural probe to its surrounding tissue was simulated using FEM modeling. The model included heat conduction, convection through blood flow, metabolic heat generation in the tissue, and heat generation by the ASIC [22]. Additionally, the tissues were assumed to be homogeneous and isotropic. The active neural probe, as a heat source, was modeled to have a uniform heat distribution throughout the volume of the body and a concentrated heat distribution throughout a specific portion of the shank (i.e., 1 mm long portion).

The material properties and physiological parameters of the tissues and other materials used in these simulations are summarized in Fig. 9.9 (values taken from [22]).

<i>Parameter</i>	<i>Brain</i>	<i>Skull</i>	<i>Scalp</i>	<i>Blood</i>	<i>Silicon</i>	<i>PDMS</i>
Density [kg/m^3]	1041	1990	1100	1060	2329	0.97
Specific Heat Capacity [$\text{J}/\text{kg}\cdot\text{K}$]	3640	1300	3150	3840	702	1460
Thermal Conductivity [$\text{W}/\text{m}\cdot\text{K}$]	0.528	0.65	0.342	0.53	12.4	0.15
Blood Perfusion Rate [$(\text{ml}/\text{s})/\text{ml}$]	$9.7\text{E-}3$	$9.9\text{E-}4$	$2.2\text{E-}3$	–	–	–
Metabolic Heat Generation [W/m^3]	10383	26	1100	–	–	–
Heat Transfer Coefficient [$\text{W}/\text{m}^2\cdot\text{K}$]	–	–	5	–	5	5

Fig. 9.9 Physical and physiological properties of the tissue and other materials used in the thermal simulations [22]

It is not always obvious that the most amount of power in such high-density electrodes is consumed while transmitting the data off-chip, both in wired or wireless fashion. This might be well beyond the average power consumed by the rest of the chip and can very likely be the primary cause of self-heating. While a tethered animal might require a long cable (5 m or more) to behave in unhindered manner, the additional power to push data on such a long cable could cause further problems. Thermal simulations show that if the site of communication power is not sufficiently far away from the brain, heat can easily flow toward the brain from outside.

One obvious method to reduce the communication power is to reduce the supply voltage, but that could come at a cost of reduced reliability over long distance. Wireless communication is not necessarily a boon for such high data rate. Calculating power budget for wireless neural probes can be much more involved and has to include both wireless channel efficiency and prescribed SAR (specific absorption rate) of the antenna [23].

2.5 *Programming the Probe*

Active neural probes are mixed-signal ASICs of unusual shape and requirement. Other than modern-day imagers, there are no custom ASIC that contains so many analog circuit blocks spread over such a wide piece of silicon. While imagers still retain the traditional rectangular shape of a generic custom ASIC, a neural probe not only has a nonuniform contour, but all its connection to the external world has to be obtained from only a small region on one end. This brings up additional problem in accessing the internal registers of the chip (for programming), providing supply pads and also distributing any sensitive analog signal.

A neural probe could have a narrow ($<100\ \mu\text{m}$) and long (10 mm) shaft connected to a rather broad base. As indicated in Fig. 9.10, in most cases, only one edge of the base could be used for pads. While the shaft and the base can have few thousand registers to program (for gain, bandwidth, and various other settings), unlike a digital ASIC, no clock or data tree can be created for guaranteed timing enclosure for all these registers. The only way to program such a chip is to have an extremely long chain of cascaded shift registers. Granted that the programming happens at a very low frequency and rather infrequently, this can still pose significant reliability issue. This unusual digital circuit block, physically distributed among the analog core, is not easy to simulate at all. Even though this can never be guaranteed to operate as traditional digital blocks, some precautionary methods can be built in. Firstly, the chain can be broken in as many small parts as possible. Secondly, the clock and data can be fed in from opposite direction for reduced possibility of setup-hold violation. Thirdly, it should be kept in mind that all the registers in one chain are tripping near simultaneously causing significant power/ground droop. And finally, Monte Carlo simulations can be done (considering the clock delays and supply drops) to ensure few thousand instances of shift registers

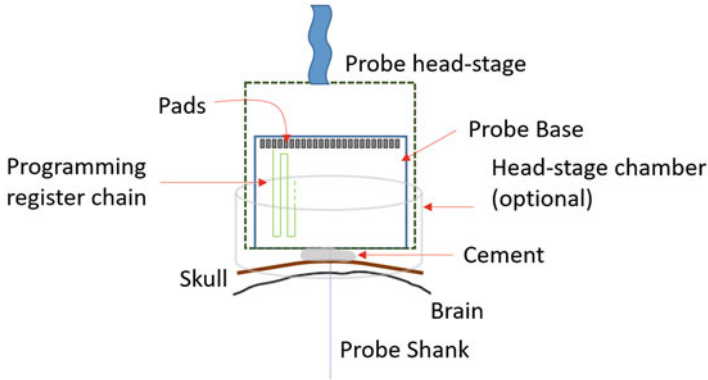


Fig. 9.10 A diagram of a neural probe fixed to the skull using a dental cement. The probe is generally held on a headstage and connected to a flexible cable. The probe base and the headstage often have size limitations for ease of use or due to a fixed chamber fitted to the animal brain (that connects to microdrive, etc.). The pads on the probe base could be most likely placed on one edge

can behave in an acceptable fashion. Another safety feature is to readout dummy data from the register chain before the actual programming is done.

Distribution of power supply both in the base and particularly in the shank is a major concern for such a probe shape. It is necessary to estimate the droop in power and ground lines as they can travel several millimeters from the edge of the chip where the pads are. This requires appropriate planning on supply distribution right from the beginning of the design and individual block simulation that considers the estimated drop. Even if the supply variations are not catastrophic for circuit blocks, they can create systematic mismatch from one corner of the chip to the other and should be considered while specifying the performance.

Similar concerns could arise while distributing any bias all along the chip. Even if a bias line connects only gates of transistors, it's possible to have bias drops due to small gate leakage in individual transistors. Hence, current mode biasing is preferred wherever possible.

3 State of the Art on CMOS Neural Probes

Although some limited attempts to implement CMOS neural probes have been reported in the last decade [4, 21, 24], it has been only in the last few years that such implementations succeeded to surpass the number of recording electrodes and channel and the overall functionality of the traditional passive and hybrid neural probes. In this section, we will discuss the most recently reported CMOS probes and their main design features.

3.1 A 966-Electrode Probe with 384 Channels

An active neural probe implemented in a 0.13 μm technology was reported by Lopez et al. [17], which is a scaled-up and modified version of the previous design of the same author [25]. The design consists of an implantable shank containing a large and dense electrode array with in situ buffering under each electrode. From this array, subgroups of 384 electrodes can be selected through a local switch matrix to be read out simultaneously. Due to the local buffering, an impedance conversion is performed very close to the electrodes, which leads to lower-signal cross talk and interferences. A high-level block diagram of this design is shown in Fig. 9.11.

The readout of the neural signals is achieved through 384 recording channels located in the non-implantable part of the probe (base). These channels consist of a low-noise instrumentation amplifier, band-splitting filters, programmable gain amplifiers, and driver buffers. The signals are finally digitized and sent to an external headstage through a serial interface. The channel-independent programmable settings (gains and bandwidths) in this probe enable the adaptation of the channel to the specific signal being recorded by it. Therefore, different gains could be applied, for example, to the signals coming from different brain regions. The band-splitting filter separates the AP and LFP signals to make maximum use of the ADC dynamic

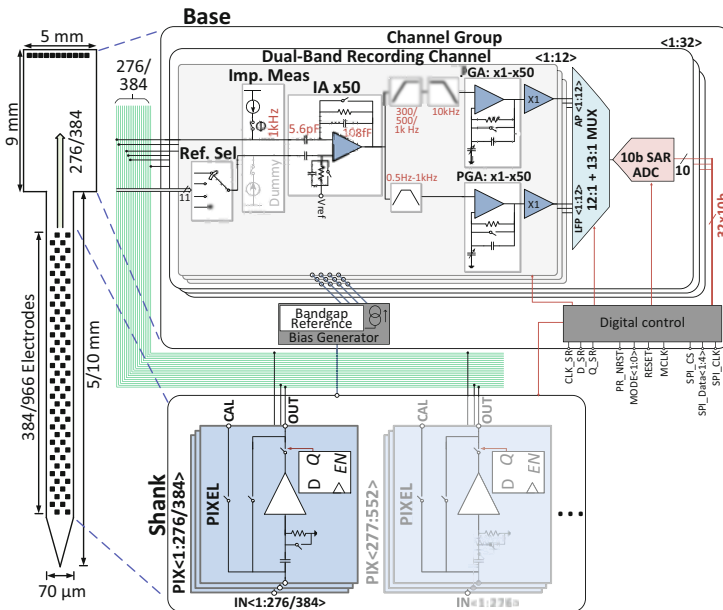


Fig. 9.11 High-level architecture of a CMOS neural probe with 966 electrodes and 384 parallel recording channels [17]

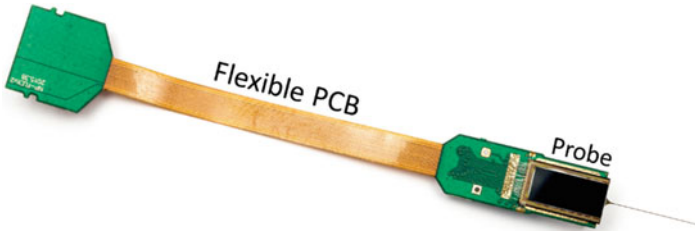


Fig. 9.12 CMOS neural probe die packaged on a flexible PCB with rigid ends

range. In this way, the ADC does not require a high resolution. The AP and LFP channel paths can also have independently programmable gains, and both signals are always recorded simultaneously.

As shown in Fig. 9.12, the probe is packaged in a flexible PCB with rigid ends. This PCB directly connects to the headstage. The small and light headstage includes a serializer to transmit the data in real time and with minimum latency to the PC through a very thin, long, and flexible cable. The same cable is used to send configuration commands to the probe. Such compact system enables the use of the probe in both acute and chronic experimental settings.

One of the important features of this design is the possibility to pseudo-randomly select combinations of electrodes that cover the full shank length, which translates to the possibility to record neural signals along the implantation track covering several brain regions. This has greatly broadened the type of experiments that can be currently realized with implantable neural probes, especially because different brain regions could be monitored simultaneously. Currently, this probe is being actively used by several European and American neuroscience groups and institutions, constituting one of the most advanced tools available.

3.2 A Time-Multiplexed Neural Probe with 1356 Parallel Recording Sites

Although the neural probe described above is scalable in the number of electrodes for a given shank length, the number of simultaneous readout channels is limited by the number of metal wires that can be accommodated in the shank width. In order to address this wiring bottleneck, the authors have reported in [26] a new architecture which relies on time-division multiplexing and techniques that reduce the associated drawbacks. This architecture maximizes the readout capability of a given inserted shank by simultaneously recording all the available electrodes.

By using time-division multiplexing directly inside the shank, this design achieves a 1:1 electrode-to-channel ratio and supports simultaneous recording of all the 1356 electrodes.

Such a feat is not an easy task to achieve: multiplexing requires high-speed switching circuits which are difficult to control in the difficult conditions of the shank. The limited cross section creates a bottleneck for the signal lines which further extends to the power lines. The consequence is that the power supply drop across the wires is unusually high, and coupling the switching circuits with a lack of decoupling capacitors creates a high ripple. Multiple circuit design techniques are used to mitigate these shortcomings.

At the signal side, the lack of a traditional anti-aliasing filter limiting the high bandwidth of the in situ amplifiers increases the in-band noise due to spectral folding, as the multiplexing operation effectively samples the input signal.

To overcome this, an alternative method of noise reduction is used in this architecture by integrating the signal over a period of time (Fig. 9.13). A simplified architecture of the device is shown in Fig. 9.14.

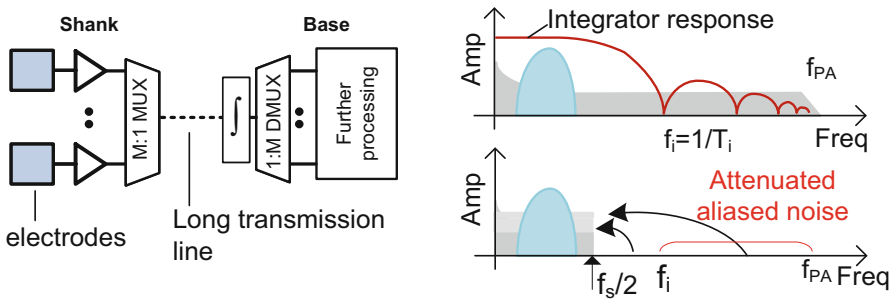


Fig. 9.13 *Left:* simplified multiplexing-demultiplexing architecture with integration. *Right:* the frequency response of the integrator attenuates the out-of-band noise that would normally fold in the band due to the multiplexing operation being equivalent to sampling

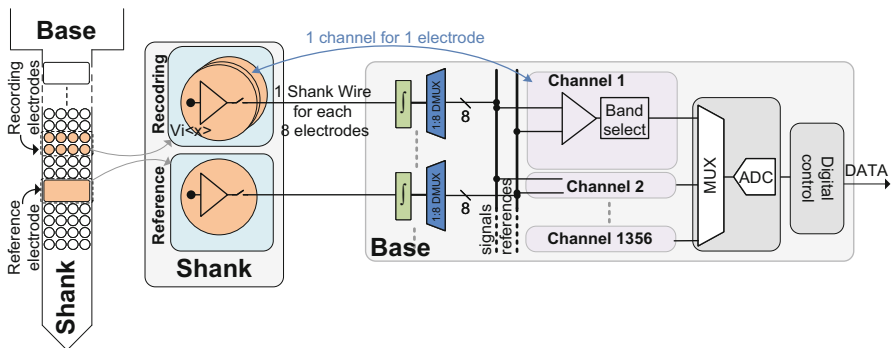


Fig. 9.14 Simplified architecture of a 1356 channel active neural probe. Time-division multiplexing and noise reduction techniques through integration are used to provide readout for the complete set of electrodes present in the implanted shank. Further techniques are used to solve the consequences of a limited power supply distribution across the shank

The main feature of this active neural probe is that it achieves 3.5 times increase in the number of simultaneous recording channels compared to the abovementioned design. Thus, the complete set of electrodes present on the implantable shank can be used for recording simultaneously, bringing in the readout of more neurons than previous implementations. This architecture advances the neural probes to a next generation, providing a new paradigm where thousands of electrodes can now be recorded from the volume where once was just a simple wire.

With such electrode density and layout, coupled with high channel count, neuroscientists are open to the possibility of new types of observations, leading to a further understanding of the brain.

4 Future Prospects

Neural probes are pushing the boundaries of what neuroscientists can achieve in terms of numbers of simultaneously recorded neurons. They are, so far, the only method to provide a high enough spatial and temporal resolution to image an individual spike from an individual neuron and apply this to a significantly large number of neurons.

Modern CMOS fabrication processes similar to integrated circuits create active neural probes, further embedding the electrodes as well as amplification and processing circuitry in a single device. Coupling such manufacturing with the help of advanced circuit techniques, the number of channels and electrodes can be pushed to even larger numbers, achieving unprecedented imaging possibilities. Future development is on track to increase the number of channels while decreasing the size and cost of CMOS neural probes. We might want to reach a point where all neurons that can be theoretically recorded by a long narrow shaft (within a cylindrical volume of 140 μ m radius [27]) are captured by three or four separate electrodes simultaneously. This is a situation that will let all the neurons to be triangulated and classified with very little ambiguity.

However, a higher number of electrodes come with its own problem. An obvious one is related to the shank bottleneck, as described in Fig. 9.3. This limits the maximum number of electrodes one can simultaneously record. Even though elegant time-multiplexing solutions can be implemented in the shank (as in Sect. 3.2), the base of the probe still needs to amplify, filter, and digitize all these electrodes using independent channels. However, more channels in the base can result in a larger dimension and hence explode the problems described in Sect. 2.5. Many of these problems could be eliminated if the size of the channels could be made smaller or even negligible.

More channels also mean more digital data and associated power to transmit them. Several options exist to reduce the power consumption of wired communication. LVDS and low-voltage LVDS and other custom current mode communication could be investigated. More channels also mean greater variations (in gain, BW, etc.) among channels. Ideally this could be calibrated out as a “factory setting”

before the probes are used by neurophysiologists. However, calibration of such a large analog chip with a wide number of performance variables could be quite time-consuming and expensive, considering the rather low volume of manufactured devices. Alternatively, it could be argued that a neural recording probe does not need factory calibration since there is no absolute AP shape or magnitude the user is interested in. As long as the captured signal is repeatable over time, the unique features of individual neurons can easily be distinguished and classified.

Other features such as on-board data analysis, signal compression, and wireless and battery-less operation have been designed, increasing the palette of neuroscience application. Several on-chip classification algorithms have been demonstrated, from simple threshold detectors to far complex ones [23], to process the data locally and transmit only the necessary signal. However, most of these lossy compression techniques have their limitations both on the engineering aspect (working within limited power budget) and also general acceptance among neuroscientists. One of the major problems in data compression that depends on a predefined set of parameter is the significant shift in the local environment of the probe over time. For a chronic recording, the probe can shift, new networks can form, and scar tissues can be generated. Among other techniques, compressive sampling has been shown as a very useful compression tool that requires very little on-chip computation capability (hence low power) and can automatically detect signal quality degradation to update its parameters [23].

Even though wireless recording probes are an interesting option from usability point of view, as explained before, wireless systems are not necessarily capable of handling such large data volume while operating for a long time from a light battery attached to the animal. However, there are alternative approaches that reduce the size of the entire probe to a few electrodes. Such a device, called neural dust, can be envisioned as a small integrated circuit comprising of a few electrodes, similar to the tip of a tetrode and comparable in size. Harvesting energy from the ambient brain or external magnetic fields and with the capability to transmit the recorded data using wireless technologies allows for the creation of an independent and small device that can be implanted anywhere in the brain [28]. Allowing multiple such devices, even millions to operate simultaneously inside the same brain, will provide the future of neuroscience imaging.

5 Conclusion

This chapter provides a general overview of various forms of neural recording and performs an in-depth analysis of in vivo CMOS neural probes. This comparatively new tool available to neurophysiologists, particularly the ultrahigh-density probes, has the potential to significantly alter the experimental methods and subsequent data processing. Here we describe various requirements, problems, and possible solutions in designing such CMOS circuits. Two state-of-the-art probes that use many of these novel techniques are described in detail. It is expected that high-

density neural probes will be commonly used in future neuroscience experiments, and neurophysiologists will require custom probes (in number of electrodes, length/width of shank, number of simultaneous recording channels, etc.) tailored to their experimental protocol. As the basic engineering problems get solved, it can be envisioned that such custom modifications can be easily automated using modern-day IC design tools.

References

1. G. Buzsáki, E. Stark, A. Berényi, et al., Tools for probing local circuits: high-density silicon probes combined with optogenetics. *Neuron* **86**, 92–105 (2015). <https://doi.org/10.1016/j.neuron.2015.01.028>
2. F. Yazicioglu, C.M. Lopez, S. Mitra, et al., Ultra – high – density in – vivo neural probes. in *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2014), pp. 2032–2035
3. C.M. Lopez, S. Mitra, J. Putzeys, et al., 22.7 A 966-electrode neural probe with 384 configurable channels in 0.13 μm SOI CMOS. in *Dig. Tech. Pap. IEEE International Solid-State Circuits Conference* (2016), pp. 392–393
4. R.H. Olsson, K.D. Wise, A three-dimensional neural recording microsystem with implantable data compression circuitry. *IEEE J. Solid State Circuits* **40**, 2796–2804 (2005). <https://doi.org/10.1109/JSSC.2005.858479>
5. J. Scholvin, J.P. Kinney, J.G. Bernstein, et al., Close-packed silicon microelectrodes for scalable spatially oversampled neural recording. *IEEE Trans. Biomed. Eng.* **63**, 120–130 (2016). <https://doi.org/10.1109/TBME.2015.2406113>
6. G. Santhanam, M.D. Linderman, V. Gilja, et al., HermesB: a continuous neural recording system for freely behaving primates. *IEEE Trans. Biomed. Eng.* **54**, 2037–2050 (2007). <https://doi.org/10.1109/TBME.2007.895753>
7. D. Han, Y. Zheng, R. Rajkumar, et al., A 0.45 V 100-channel neural-recording IC with sub- μW /channel consumption in 0.18 μm CMOS. *IEEE Trans. Biomed. Circuits Syst.* **7**, 735–746 (2013). <https://doi.org/10.1109/TBCAS.2014.2298860>
8. R.R. Harrison, The design of integrated circuits to observe brain activity. *Proc. IEEE* **96**, 1203–1216 (2008). <https://doi.org/10.1109/JPROC.2008.922581>
9. B. Gosselin, Recent advances in neural recording microsystems. *Sensors* **11**, 4572–4597 (2011). <https://doi.org/10.3390/s110504572>
10. K.D. Wise, A.M. Sodagar, Y. Yao, et al., Microelectrodes, microelectronics, and implantable neural microsystems. *Proc. IEEE* **96**, 1184–1202 (2008). <https://doi.org/10.1109/JPROC.2008.922564>
11. M.S. Chae, W. Liu, M. Sivaprakasam, Design optimization for integrated neural recording systems. *IEEE J. Solid State Circuits* **43**, 1931–1939 (2008). <https://doi.org/10.1109/JSSC.2008.2001877>
12. T. Jochum, T. Denison, P. Wolf, Integrated circuit amplifiers for multi-electrode intracortical recording. *J. Neural Eng.* **6**, 12001 (2009). <https://doi.org/10.1088/1741-2560/6/1/012001>
13. R. Rieger, J. Taylor, Design strategies for multi-channel low-noise recording systems. *Analog Integr. Circ. Sig. Process* **58**, 123–133 (2009)
14. A.M. Sodagar, G.E. Perlin, Y. Yao, et al., An implantable 64-channel wireless microsystem for single-unit neural recording. *IEEE J. Solid State Circuits* **44**, 2591–2604 (2009). <https://doi.org/10.1109/JSSC.2009.2023159>
15. Y.-K. Song, D.A. Borton, S. Park, et al., Active microelectronic neurosensor arrays for implantable brain communication interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* **17**, 339–345 (2009). <https://doi.org/10.1109/TNSRE.2009.2024310>

16. B. Gosselin, A.E. Ayoub, J.-F. Roy, et al., A mixed-signal multichip neural recording interface with bandwidth reduction. *IEEE Trans. Biomed. Circuits Syst.* **3**, 129–141 (2009). <https://doi.org/10.1109/TBCAS.2009.2013718>
17. C.M. Lopez, J. Putzeys, B.C. Raducanu, et al., A neural probe with up to 966 electrodes and up to 384 configurable channels in 0.13 μm SOI CMOS. *IEEE Trans. Biomed. Circuits Syst.*, 1–13 (2017). <https://doi.org/10.1109/TBCAS.2016.2646901>
18. S.B. Lee, H.-M. Lee, M. Kiani, et al., An inductively powered scalable 32-channel wireless neural recording system-on-a-chip for neuroscience applications. *IEEE Trans. Biomed. Circuits Syst.* **4**, 360–371 (2010). <https://doi.org/10.1109/TBCAS.2010.2078814>
19. J. Du, T.J. Blanche, R.R. Harrison, et al., Multiplexed, high density electrophysiology with nanofabricated neural probes. *PLoS One* **6**, e26204 (2011). <https://doi.org/10.1371/journal.pone.0026204>
20. S.A. Desai, J.D. Rolston, L. Guo, S.M. Potter, Improving impedance of implantable microwire multi-electrode arrays by ultrasonic electroplating of durable platinum black. *Front Neuroeng* **3**, 5 (2010). <https://doi.org/10.3389/fneng.2010.00005>
21. T. Torfs, A.A.A. Aarts, M.A. Erisimis, et al., Two-dimensional multi-channel neural probes with electronic depth control. *IEEE Trans. Biomed. Circuits Syst.* **5**, 403–412 (2011). <https://doi.org/10.1109/TBCAS.2011.2162840>
22. S. Kim, P. Tathireddy, R.A. Normann, F. Solzbacher, Thermal impact of an active 3-D microelectrode array implanted in the brain. *IEEE Trans. Neural Syst. Rehabil. Eng.* **15**, 493–501 (2007). <https://doi.org/10.1109/TNSRE.2007.908429>
23. J. Zhang, K. Duncan, Y. Suo, et al., Communication channel analysis and real time compressed sensing for high density neural recording devices. *IEEE Trans. Circuits Syst. I Regul. Pap.* **63**, 599–608 (2016). <https://doi.org/10.1109/TCSI.2016.2556123>
24. R.H. Olsson III, D.L. Buhl, A.M. Sirota, et al., Band-tunable and multiplexed integrated circuits for simultaneous recording and stimulation with microelectrode arrays. *IEEE Trans. Biomed. Eng.* **52**, 1303–1311 (2005). <https://doi.org/10.1109/TBME.2005.847540>
25. C.M. Lopez, A. Andrei, S. Mitra, et al., An implantable 455-active-electrode 52-channel CMOS neural probe. *IEEE J. Solid-State Circuits* **49**, 248–261 (2014). <https://doi.org/10.1109/JSSC.2013.2284347>
26. B.C. Raducanu, R.F. Yazicioglu, C.M. Lopez, et al., Time multiplexed active neural probe with 678 parallel recording sites. in *2016 46th European Solid-State Device Research Conference* (2016), pp. 385–388
27. G. Buzsaki, C.A. Anastassiou, C. Koch, The origin of extracellular fields and currents – EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012)
28. D. Seo, R.M. Neely, K. Shen, et al., Wireless recording in the peripheral nervous system with ultrasonic neural dust. *Neuron* **91**, 529–539 (2016). <https://doi.org/10.1016/j.neuron.2016.06.034>

Chapter 10

Photonic Interaction with the Nervous System

Patrick Degenaar

1 Introduction

There is great value and beauty in understanding the complexity and structure of our nervous systems for its own sake. But in addition, by learning its key principles, it is also possible to develop better electronic systems for the coming era of artificial intelligence and robotics. It is also of interest to those seeking clinical solutions to neurological conditions. The neuroscience field is therefore of particular interest to many talented scientists, engineers and clinicians.

But how does this relate to biophotonics?

To answer that question, the electronic operation of the nervous system first needs to be explained in a little more detail. A neuron – the unit cell of the nervous system – can be considered a sealed membrane, floating in an electrolyte solution. Ion pumps remove positively charged sodium ions, making the internal space negative with respect to the solution (i.e. polarised). Channels in the membrane can be opened to allow the cell to depolarise again. The state of the cell can thus be thought in pseudo-digital terms of being polarised/unpolarised or in analog terms as the transmembrane voltage potential.

A typical example is the communication cells in the eye, known as retinal ganglion cells. Ion channels on the membrane are normally in the closed state. Thus, the continuing pumping from membrane pumps polarises the cell. Then, when these cells receive chemical stimulus from neighbouring cells, chemical ion channels called receptors open allowing the cell to depolarise. If the cell depolarises past a

P. Degenaar (✉)
School of Engineering, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
e-mail: patrick.degenaar@newcastle.ac.uk

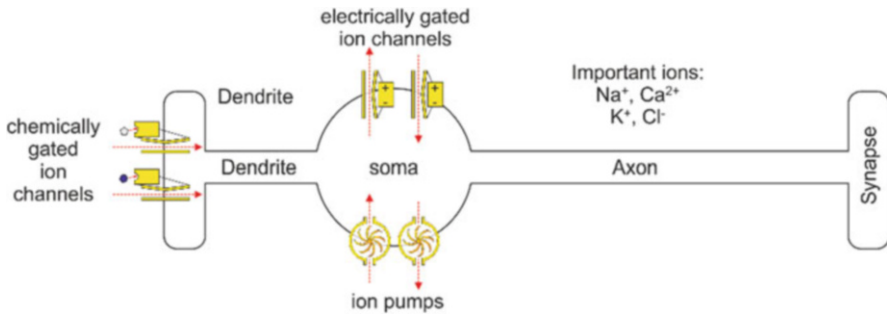


Fig. 10.1 Operation of the neuron. Continuously pumps in potassium and calcium while pumping out sodium and chlorine ions. This acts to polarise the neuron – i.e. put an electric field across the membrane. Electrically gated ion channels allow the reverse through diffusion, and chemically gated ion channels allow activation via neurotransmitters

threshold, electrically gated channels open forcing the cell to be fully depolarised for a short time. The cell then resets. Such and digital-like bursts of depolarization are known as action potentials. The phase and frequency of these provide information to downstream brain cells.

Figure 10.1 shows a typical nerve cell in terms of its main components. Traditionally, it was assumed that the neurons are essentially electrical and chemical in nature. Thus, these methods would be best for interaction. A biphasic injection of charge close to the cell membrane could alter the membrane potential sufficiently to impart an action potential. However, discoveries in 2003 showed that it is possible to control this cell optically. The reason for its importance is that traditional neuroelectronic interfaces suffer from many drawbacks, which can be summarised as follows:

<i>Inhibition</i>	Electrical stimuli can excite action potentials (depolarise neurons), but it is challenging to inhibit activity.
<i>Cellular targeting</i>	A pulse of stimulus will affect all neurons within its sphere of influence. It is challenging to target specific cells – e.g. inhibitory or excitatory.
<i>Scalability</i>	Scaling to larger densities of electrodes results in larger current densities through each electrode. However, there is a material limit to charge density beyond which there is rapid degradation.
<i>Closed loop</i>	Electrical stimulation pulses are much larger than neural electrical signalling. Thus, simultaneous stimulus will swamp recording electronics and the signals they are trying to detect, making closed-loop operation difficult.

There are many ongoing innovations in materials science and electronics which are trying to negate the above effects. But an alternative that became available in 2003 is the use of optogenetics – the genetic manipulation of the nervous system to become light sensitive.

2 Optogenetic Photosensitization of Cells

Between 2003 and 2005, four remarkably different approaches were published in the literature. This family of approaches became known as ‘optogenetics’ – the genetic manipulation of neural tissue to become light sensitive. This is described in Fig. 10.2 and can be classified into four domains:

- Optogenetic ion channel*: Genetic manipulation of cells to express optically activated ion channels on their cell membranes
- Optogenetic ion pumps*: Genetic manipulation of cells to express optically activated ion pumps on their cell membranes
- Optogenetic amplifiers*: Genetic manipulation of cells to express optically activated G-protein-coupled receptors (GPCR) (amplifiers)
- Optogenetic sensors*: Genetic manipulation of cells to express fluorescent proteins which respond to differences cell activity

The *optogenetics* name refers to the fact that, in order to get optically activatable ion channels, pumps and voltage sensors into the cell membrane, they need to be produced by the cell itself. That requires changing the genetic code of the cell. There are various methods to achieve this such as the use of Adeno Associated (AAV) and

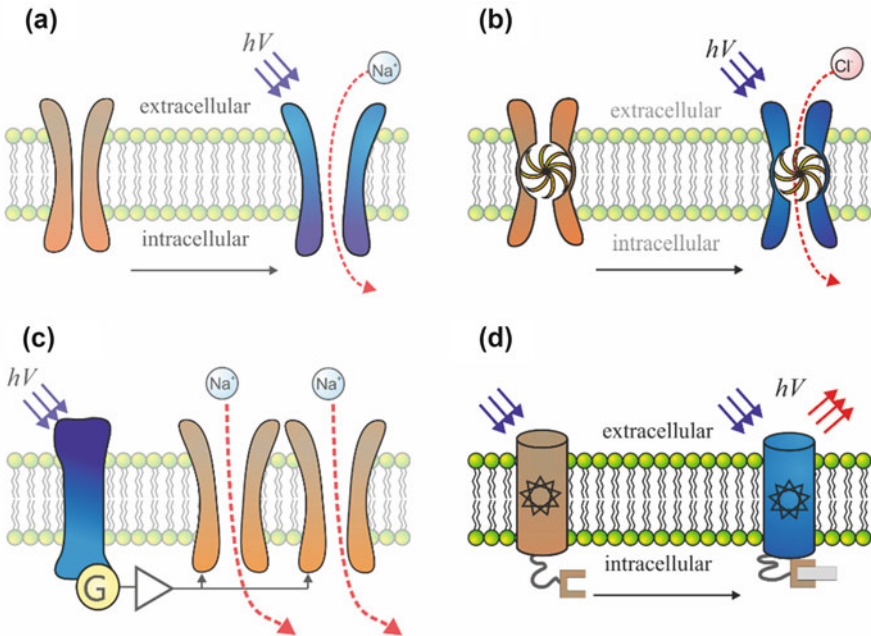


Fig. 10.2 Methods of optical-neural stimulus. (a) Via optically sensitive ion channels. (b) Via optically sensitive ion pumps. (c) Stimulation of optically sensitive signal transduction pathways (biological signal amplifiers). (d) Fluorescence sensors which respond to cellular activity

Lenti Virii. What is significant is that it allows for the genetic targeting of certain cell types. For example, in contrast to an electrical stimulus exciting excitatory and inhibitory cells equally, each pathway could be targeted with chromatically different proteins.

2.1 *Optogenetic Ion Channels*

At the end of 2003, Nagel et al. presented a light-activated ion channel derived from single-celled swamp algae called *Chlamydomonas reinhardtii* [1]. Previously, channelrhodopsin-1 had been published by the same team, but it primarily transported hydrogen ions. The great significance of the 2003 paper was that the presented channelrhodopsin-2 (ChR2) could significantly transport cation ions such as sodium and potassium. Furthermore, its photoswitching ability was demonstrated in cells which did not natively express them [1]. Then in 2005, a US team including Boyden et al. in collaboration with the German team published their efforts showing how channelrhodopsin-2 could be used in neurons [2]. One member of that team, Karl Deisseroth, later coined the term ‘optogenetics’ which is now commonly used across the field.

The specific required light intensity to activate a cell will vary according to a number of parameters: the number of channels expressed, the light intensity, the geometry of the cell and the activation and recovery kinetics of the opsin molecules [3, 4]. The common threshold of irradiance acknowledged by the community is $\sim 1 \text{ mW/mm}^2$ ($= 2 \times 10^{15}$ photons/ $\text{mm}^2 \text{ s}$). This is considerably less intense than the IR method by Fork in 1971 but is still high. Methods to provide sufficient stimulus must, therefore, be considered [5, 6].

The wild type of channelrhodopsin has an absorption peak at 470 nm (blue). Moving to the blue is undesirable, so there is scope for two more absorption peaks in the green and orange, which would allow for multiwavelength selectivity. This has been achieved with Zhang et al. [7] presenting a green variant in 2008 from the *Volvox carteri* organism and Lin et al. presenting an orange/red variant in 2011 called ReaChR [8]. There have also been various efforts to improve the light sensitivity [9], a review of which can be found in [10].

2.2 *Optogenetic Ion Pumps*

The second type of photosensitization agent is the ion pump. The concept of this is described in Fig. 10.2c. Incoming photons provide energy to pump ions to the other side of a membrane at a rate of one ion per photon. If this is compared to a 10 ms open period of a ChR2 channel, around 1000 \times more light is required for the same aggregate ion transfer. However, unlike channels, ion pumps can operate against concentration gradients. It is thought that the earliest life on the planet – bacterial-like organisms – known as *Archaea* used solar-powered ion pumps to survive in the

extreme conditions of early earth. Thus, such structures have existed for around 4 billion years. Matsuno et al. were the first to discover an alternative light-activated ion pump in 1977 [11], which they called halorhodopsin in a paper in 1981 [12]. It was later determined to be an efficient chloride transporter as well as a pH regulator. However, it remained an interest purely in the field of molecular biology until ChR2 was successfully demonstrated in neurons in 2005. Two competing teams (Zhang et al. [13] and Han et al. [14]) demonstrated halorhodopsin expression in 2007 derived from the *Natronomonas pharaonis* (NpHr) bacteria. Of particular significance was the fact that the absorption peak was 590 nm compared to 470 nm for ChR2. This allowed for blue light stimulus (ChR2) and yellow light inhibition (NpHr).

2.3 *Optogenetic G-Protein-Coupled Receptors*

The third form of photosensitization agent that needs to be considered is the G-protein-coupled receptor (GPCR). These are molecular amplifiers connected to ion channels which allow for great sensitivity while minimising noise [15]. In 2005 two independent groups demonstrated ectopic expression of melanopsin, a depolarising (stimulating) GPCR in neuron cells [16, 17]. The result showed that the photosensitivity was 1000× compared to channelrhodopsin-encoded cells. However, the kinetics was very slow. Response times were tens of seconds rather than tens of milliseconds for ChR2. In the early days, the fact that it was already expressed in the human genome was considered a particular regulatory advantage for moving towards clinical use. That is, as the GPCR components on the cell membrane already exist in the human retina, a long-term immune response would not be expected from their use. There was thus some excitement about its potential for clinical use. In 2008, Lin et al. demonstrated its potential for use in returning visual function in mouse retinæ with degeneration similar to the retinitis pigmentosa disease [18]. Since then, melanopsins have been overshadowed by the less complex and faster channelrhodopsins. Nevertheless, in the long term, they have very interesting possibilities. Even if fast kinetics cannot be achieved, there are many non-neuronal systems in the body that can benefit from photoactivation.

2.4 *Optogenetic Sensing*

Fluorescent imaging utilises proteins which absorb at one wavelength and emit in another. By separating emission and stimulation wavelengths, it is possible to quantify the presence of fluorescence proteins which in turn can quantify cell types, structures or activity. To record neural activity, it is possible to genetically combine green fluorescent protein (GFP) with a calcium-binding protein called calmodulin (CaM) to create GCaMP. Chen et al. in 2013 [19] reported an advanced version of

this which can explore basic neuroscience activity in the brain. When combined with two-photon imaging techniques, it has allowed for remarkable imaging of activity in cortical regions in live brain.

3 Optoelectronic Interrogation of the Nervous Tissue

Optogenetics typically has an irradiance requirement of ~ 1 mW/mm². If inefficiencies in the optical delivery systems are taken into account (e.g. 1–10% for virtual reality optics [20] to the eye), then a very intense source is required. Furthermore, scattering of light in neural tissue needs to be taken into account. In short, for thin tissue and dissociated culture, imaging of patterned light onto the target is appropriate. For deeper tissue, in addition to the loss of irradiance, light patterns become rapidly blurred with depth. Thus, either two-photon approaches or implantable light delivery systems need to be developed.

Laboratory experiments initially used high-brightness, e.g. Xe/Hg, lamps [21] progressing to laser sources [22, 23]. To achieve patterned stimulation, such systems would either raster scan a spot or use spatial light modulation schemes [6]. In recent years, for deeper tissue research, whether with organotypic slices, or in vivo brain, two-photon approaches have been used. Two-photon approaches utilise two photons of double the normal wavelength which can traverse the tissue to much greater depth. More typically this is for optogenetic monitoring of neural activity, rather than stimulus. A review of the technology has been provided by Schultz et al. [24].

Pohrer et al. were the first to demonstrate that high-brightness micro-light-emitting diodes (micro-LEDs) could be used for patterned neural illumination [25]. These have then been developed into high-power light-emissive arrays [26, 27]. For microscopy, standard optics can then be used to image the array to the target. For retinal prosthetics, virtual reality optics can be used. For deeper tissues, LEDs can be either used at a distance to drive light down an optic fibre or waveguide or generate light locally in tissue.

3.1 *Scattering of Light in the Brain Tissue*

A well-understood phenomenon in physics is the description of light as both waves and particles. The wave nature means that the irradiant density of light can be calculated as the intensity divided by the 3D surface of the wave at given distance from the source. For example, for a point source with a radial emitter, the irradiance at any point varies with $1/4\pi r^2$, where r is the radial distance from the source. The particle nature of governs absorption and scattering processes of light traversing through a medium. As such, a homogenous medium can be classified in terms of its coefficient of absorption and coefficient of scattering. The combination of the two provides its coefficient of extinction.

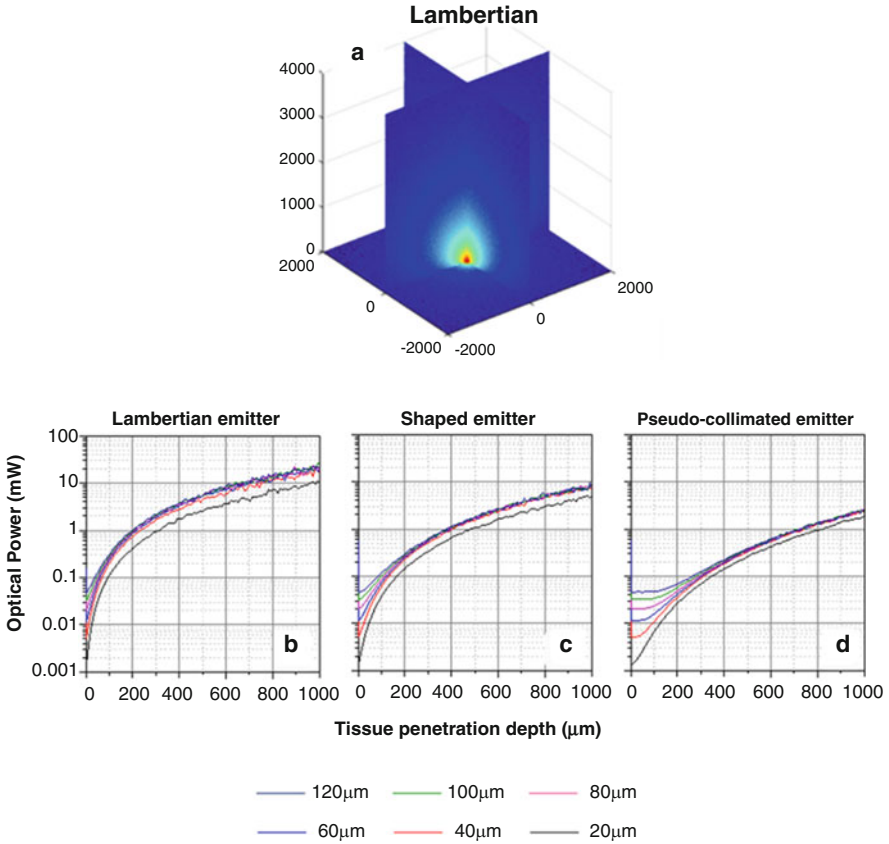


Fig. 10.3 The profile of related (normalised) (a) 3D views of the emission profiles of 80 μm emitters for Lambertian LED emission. (b–d) Optical power required to reach specific depths for emitters of different size and intensity

Common wavelengths are between 470 nm (ChR2, GFP) and 590 nm (HpHr). The interaction of these wavelengths with tissue can be considered in terms of particles. As a general rule, when light interacts with particles which are smaller than the wavelength, e.g. protein structures, Rayleigh scattering is dominant. For particles between 1λ and 10λ , e.g. subcellular structures such as nuclei and blood vessels, Mie scattering is dominant. Beyond that, absorption processes are dominant. Figure 10.3 shows the Monte Carlo modelling results of light traversal through tissue for different emitter types and emitter sizes.

Key design considerations:

Emitter type Even though tissue scatters light strongly, there is a benefit in terms of tissue penetration for shaped emitters such as collimated LEDs [27], lasers or optic fibre.

Emitter size (d–f) Demonstrates that the size of the emitter only matters in the initial penetration depth equal or smaller to the size of the emitter. If the emission is directly from an emitter, then it is best to use emitters as large as possible for the application, as emitter efficiencies reduce in proportion with current density (i.e. emitter radius).

3.2 Light Sources

Micro-light-emissive elements for optogenetics need to be small, efficient and bright. They also need to emit wavelengths that closely match the absorption peak of the target optogenetic photosensitizer (i.e. 470–590 nm). Organic light-emitting diodes (OLEDs) typically found on consumer devices are ideal for colour separation but are unfortunately too dim for useful stimulus. As such, high-radiance gallium nitride (GaN) light-emitting diodes can be utilised [20, 25, 28].

The main candidates for ultrabright fast structured light sources are micro-LEDs or vertical-cavity surface-emitting laser (VCSEL). However, at the time of writing, the latter is still at an early stage of development for the optical wavelengths [29]. Gallium nitride LEDs can operate with current densities up to 10KA/cm² and thus provide sufficient power. They have been shown to provide the required brightness and can be manufactured to high resolutions [20, 25]. The current density and emission density profiles can be seen in Fig. 10.4b. It should be noted that the efficiency drops with current density in a process known as droop. So it is typically best to use the larger LEDs where possible.

The gallium nitride technology primarily emits at 470 nm but can be pushed towards 500 nm. To get other colours, the lighting industry uses phosphor coatings, but this is less desirable for optogenetics. As such, 530 nm is a challenging wavelength for current semiconductor technologies. An alternative AlGaInP can provide wavelengths <590 nm. The emission profile of such emitters is shown in Fig. 10.4b and described below.

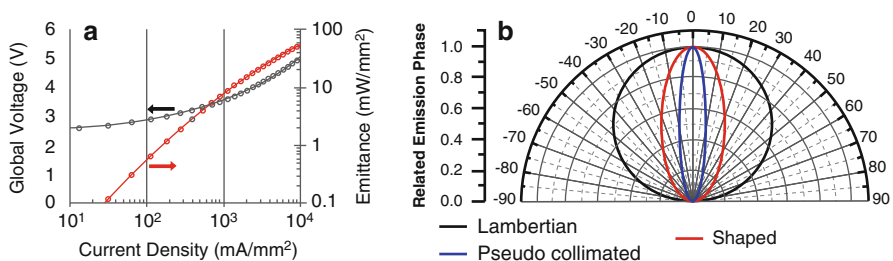


Fig. 10.4 (a) Micro-LED current-voltage profile. (b) Micro-LED emission profiles, (black) Lambertian emission which is common for most LEDs, including Cree (red)-shaped emission profile such as for Tyndall micro-LEDs (blue) pseudo-collimated for conceptual highly collimated emission

<i>Lambertian</i>	Is the standard emission profile for LEDs
<i>Shaped</i>	Includes rear beam-shaping optics which tighten the emission profile which has been previously developed at the Tyndall National Institute [27].
<i>Pseudo-collimated</i>	Conceptual LEDs which include both rear beam-shaping and frontal microlens optics [30].

3.3 Optoelectrodes (Optrodes) for Deep Tissue Penetration

Some of the earliest implantable optoelectrodes (optrodes) have either used light-guiding fibres with deposited electrode materials [31], integrated optic fibres with Utah recording electrodes [32] or made arrays of penetrating optic fibres [33]. The key issue with such approaches is that a number of emitted light channels can be limited and the optical emission is transverse through the cortical layer structures. Zorzos et al. [34] advanced on these designs by developing a probe with multiple light-guiding structures. Emission was improved to a 45° angle. However, optical multiplexing and connectivity are very challenging.

The alternative to light guiding is to generate the light directly at the target site deep in the neural tissue. McAlinden et al. demonstrated an optical probe fabricated directly from a gallium nitride LED substrate [35]. In tandem, Doroudchi et al. [36] and Cao et al. [37] have demonstrated silicon probes bonded on the mini (100–500 μm width/length dimensions)- or micro (sub 100 μm dimensions)-scale light-emitting diodes (LEDs). A key advantage of this approach is that the light emission can be in parallel with the cortical layering, which would improve the optical control.

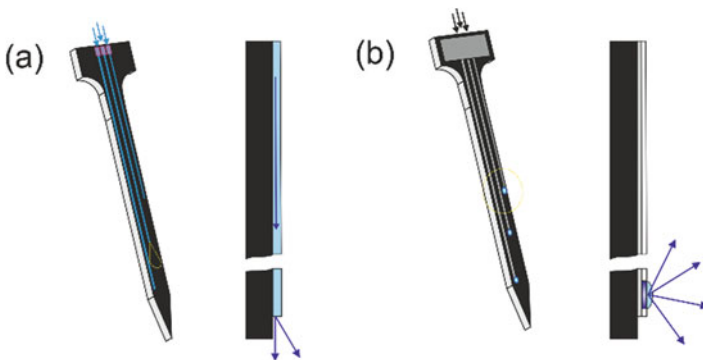


Fig. 10.5 Optrode types. (a) Waveguiding probe which receives external light and transmits down the shaft using light guides and (b) light-emissive probe which generates light locally using light-emissive elements

The two approaches are illustrated in Fig. 10.5. Each has their inherent advantages and disadvantages. Aside from emission direction, light-guiding probes are simpler, are less prone to biodegradation and do not generate heat. In contrast, optrodes with local emitters suffer from heat generation and potential biodegradation but are much easier to multiplex and determine emission profiles. It is this latter multiplexing capability that provides light-emissive probes with the potential to become high-density neural interfaces.

Light-emissive probes can be created in a number of forms. They may simply be passive structures with bonded LEDs and external driving. Alternatively, the probe can be shaped out of a CMOS chip [38], similar to high-density CMOS electronic recording probes. In this case, inbuilt electronics can be used to address both electronic recording sites and LED driving sites. Such probes can be cut out using standard laser cutting or deep reactive ion etching approaches. Section 4 will explore the circuit implications of such structures in more detail.

Penetration depth becomes important for long-term chronic studies. When probes are inserted into the brain tissue, there is a gliosis effect whereby glial scarring surrounds the immediate vicinity of the probe. Polikov et al. [39] demonstrate that this can be up to 100 μm from the probe surface. So, if, for example, the light penetration above the ChR2 threshold was only 100 μm , and the gliosis thickness was 100 μm , no net stimulus could be expected. Conversely, if the LED driving current is simply ramped up to compensate, there may be an unacceptable heating of tissue. As such, it is crucial for emissive probes to have both highly efficient emitters and drive circuits.

4 Electronics for Optical Neural Stimulation

Optogenetic cells will integrate photons over a period of milliseconds. Thus intensity can either be modulated directly over this period, or it can be an integral of the ON time within a defined period [40]. This latter form is known as pulse width modulation and can be performed with largely digital electronic methods but requires more precise timing. This section, therefore, explores how to scale such optical control over two-dimensional arrays and penetrating brain probes.

4.1 *Micro-LED Driving Circuits*

A light-emitting diode forms by combining electrons and hole injection layers and a series of quantum wells which act as a recombination layer. Although different to a standard p-n junction diode, it can be treated as having similar properties for the purposes of circuit development.

Figure 10.6a shows the equivalent circuit for a LED. The shunt resistance R_{SH} represents how much leakage current there will be in reverse. However, this

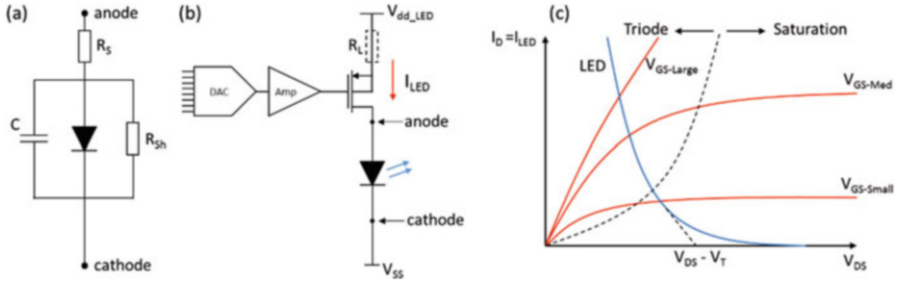


Fig. 10.6 LED driving circuits. (a) The equivalent circuit for a LED. (b) A simple pMOS driving circuit. R_L represents the line resistance to the circuit. (c) A load line plot – combining the current-voltage curves for the transistor and the LED. For low V_{GS} , the transistor operates in saturation, but more typically, it will operate in triode

approximation is poor as the reverse current in a LED is typically exponential. R_S (series resistance) represents a combination of the various resistances in series with the diode. This includes primarily the contact resistances of the anode and cathode and sheet resistances of the control lines. R_S will be the primary limit to the current at a given voltage. Thus,

$$\left. \begin{aligned} I_{LED} &= I_S e^{qV/nkT} \\ I_{LED} &= V/R_S \end{aligned} \right\} \text{whichever is the smallest}$$

That is, the LED current will either be determined by the Shockley equation (I_S = saturation current, q = charge on an electron, n = ideality factor, k = Boltzmann’s constant, T = temperature) or limited by the series resistance (R_S). The radiance from the LED is basically proportional to the current through the LED with an efficiency factor which decreases with radiance. This latter effect is known as droop and varies between LEDs.

To drive this structure, the simplest configuration is shown in Fig. 10.6b. In many cases, LEDs are common cathode with anode contacts for individual control. As such, for single driving transistor circuits, pMOS current drivers need to be used. To drive the pMOS, a digital to analog converter (DAC) can be connected to an inverting amplifier to provide the V_g driving voltage for M1. Figure 10.6c shows a load line plot scheme for driving the LED. The red lines represent the standard drain current curves for varying drain-source and gate-source voltages, i.e.

$$V_{DS} = V_{DD_LED} - V_{LED}$$

$$V_{GS} = V_{DD_LED} - V_g$$

The blue line represents the LED current variance with applied voltage across the transistor. The dotted line represents the boundary between the transistors operating in triode or saturation modes. It can be seen that, for low gate-source voltages (V_{GS}), the drive transistor is operating in triode mode. Thus,

$$I_{LED} = I_D = \frac{W}{L} K' \left(V_{GS} - V_T - \frac{V_{DS}}{2} \right) V_{DS}$$

where W and L are the widths and lengths of the MOSFTE and K' is the MOSFET parameter. As V_{DS} depends on the current through the LED, it means that the LED current and ultimately the light intensity can be susceptible to mismatch. That is, variation in the bonding arrangements can lead to variation in the series resistance of the LED. In turn, this will result in variances of the drain-source voltage (V_{DS}), thus changing the drain and LED currents. Ultimately this will lead to variance in radiance.

Commercial LED drivers thus often utilise large V_{DD_LED} voltages to ensure that the transistor driver operates in saturation. However, with most of the voltage drop over the transistor, the power efficiency decreases significantly, which is not desirable for implantable systems. A final reason for the mismatch is described in Fig. 10.6b. If there is line resistance (R_L), then there will be a slight voltage drop and the gate-source voltage becomes

$$V_{GS} = V_{DD_LED} - V_g - I_D R_L$$

For matrix arrays with long, thin drive lines, this can result in a slight pattern dimming across the matrix.

4.2 Light-Emissive Array Circuit Architecture

Optogenetic retinal prosthetics requires a high-radiance display, which can also be beneficial in microscopy. Standard organic LED displays are simply too dim so that they can be constructed from high-radiance gallium nitride LEDs. These, however, need to be driven electronically. Passive matrix systems arrange common LED anodes in rows and common LED cathodes in columns. As such, raster scanning rows and columns allow individual control, but not independent control of multiple LEDs. If the array as a whole is scanned at a sufficiently high frequency, it can be an effective display. However, this would reduce the effective intensity by a factor of N – where N is the number of rows. Therefore, active matrix systems with individual electronic control per pixel are required. To explore this further, we examine efforts to create a 90×90 high-radiance micro-LED matrix by our team (Soltan et al. [26]), as illustrated in Fig. 10.7.

The matrix was an evolution on previous efforts to develop a high-radiance display for optogenetic retinal prosthetics [41]. It consists mainly of three parts:

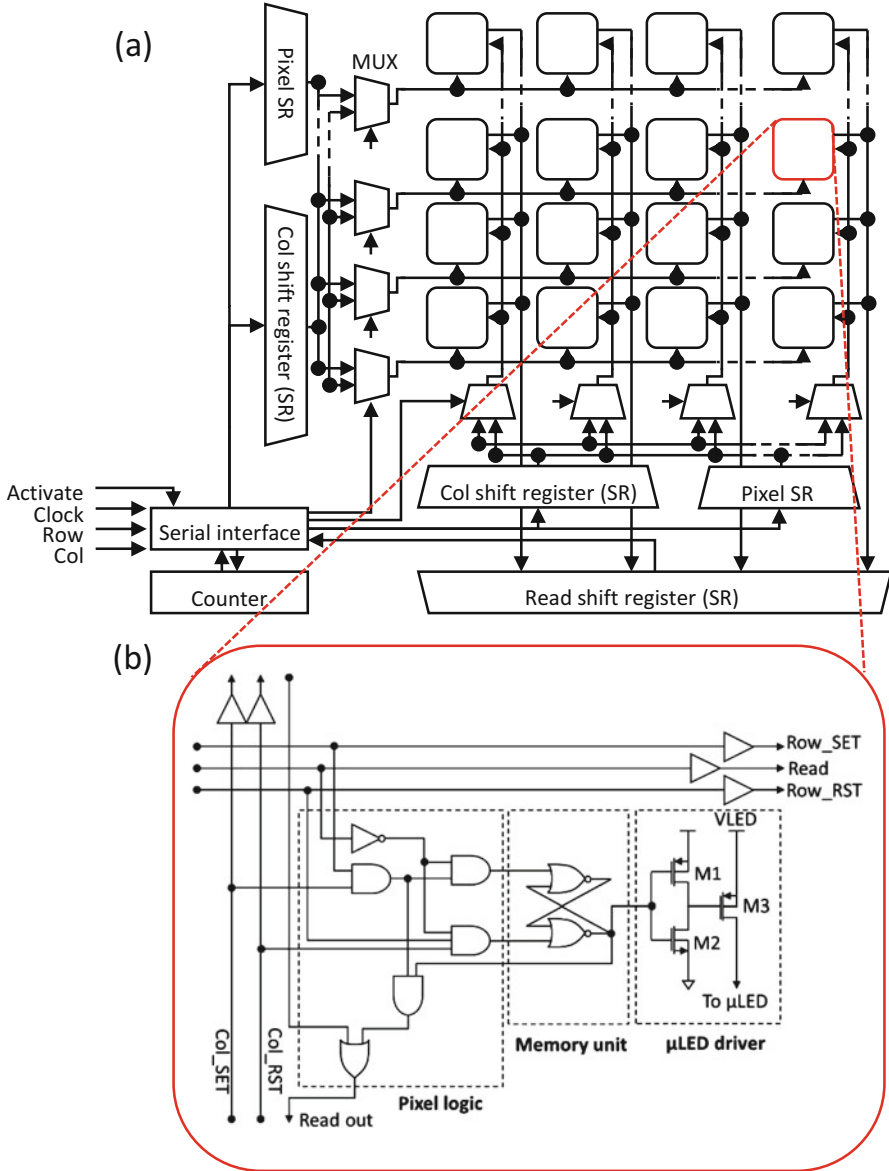


Fig. 10.7 Block diagram of the micro-LED CMOS chip. (a) The pixels are arranged in a matrix order. Shift registers read the information into both row and column, which then update pixel values. Updating can be performed in a single row, over multiple rows or at individual pixels. (b) The pixel circuit consists of three parts: the pixel logic, the memory unit and the micro-LED driver. Periodic buffers on the control lines regenerate control signals

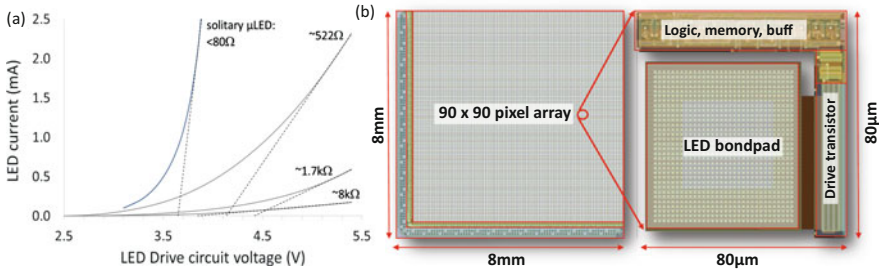


Fig. 10.8 (a) The LED current profile from the circuit shown in Fig. 10.7b for different series resistances of LED. (b) The layout of the 90×90 pixel array

the communication interface, the control logic and the micro-LED grid. To achieve active matrix functionality, each pixel has a memory unit determining the LED as ON or OFF. Figure 10.7b shows the pixel arrangement. It consists of logic control, a memory unit and a LED driver. It should be noted that, in this case, the digitally controlled amplitude modulation shown in Fig. 10.6b is absent. Incorporation of individual digital to analog converters would make pixel sizes large and thus reduce the resolution of a matrix. The DAC could be global, but it is more efficient to modulate an external V_{DD_LED} through a microcontrolled power supply.

The emitter radiance can be determined globally through voltage control (V_{DD_LED}) and locally through pulse width modulation. Global control of the V_{DD_LED} limits the overall voltage for the transistor driver V_{DS} and the LED V_{LED} , thus reducing current. In addition, it modifies the gate-source voltage ($V_{GS} = V_{DD_LED} - V_{G_LED}$). The resultant relationship between LED current and LED drive voltage (V_{DD_LED}) can be seen Fig. 10.8a. Individual control of LEDs can be achieved via pulse width modulation which is achieved by setting the pixel memory unit to ON or OFF at the appropriate rate.

In a normal imaging display, there is typically a normal distribution of intensities around the intensity midpoint. Thus, the most efficient method of updating the display is to raster the appropriate values in progressive or interlaced fashion. In a retinal display, it would be expected that the distribution is shifted bimodally between stimulus values and the minimum value (i.e. OFF). That is, while there is a distribution of intensities, the majority will be OFF for any given picture frame [42]. This is of advantage in terms of average power consumption, but it also means there are other options for updating the image.

Updating the pixel memory units can be performed by scanning the matrix row by row and updating accordingly. The 90-column shift register requires 90 clock cycles. Alternatively, there is also a pixel shift register which requires only seven clock cycles to update a pixel register. If the number of pixels to be updated is $<\log_2 N$ (in this case, $N = 90$, thus 6 pixels), then a pixel shift register can be used instead. The pixel shift register is also compatible with Address Event Representation information transfer, which is used in some neuromorphic systems.

Information transfer is via a four-wire system. In this case, a modified Serial Peripheral Interface (SPI) has been used. The four wires are as follows: activate

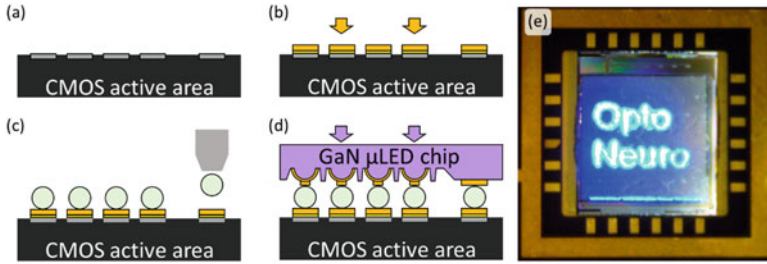


Fig. 10.9 Flip bonding process. (a) Preparation of the CMOS die. (b) Deposit a barrier layer, then gold onto the surface contact pads. (c) Deposit solder balls. (d) Flip-chip bonding using heat and pressure. (e) Final result

signal, row data line, column data line and clock signal. Fig. 10.8b shows the final layout of the chip. The ESD is in an L shape to allow for maximum use of the CMOS reticle.

The final consideration is the bonding between the CMOS and gallium nitride (GaN) micro-LED chips. Fig. 10.8b shows the pixel size as $80 \times 80 \mu\text{m}$, most of which is consumed by the bondpad. It would be possible to reduce the size further by placing the circuitry under the bondpad. The key limitation is bonding. Solder-ball flip-chip bonding processes are limited to around $80 \mu\text{m}$. The bonding arrangement and result is described in Fig. 10.9.

The CMOS pixel pads were post-processed using standard photolithographic techniques. A chrome-gold-nickel metal stack was deposited on each bondpad (Fig. 10.9b). Subsequently, $50 \mu\text{m}$ solder balls consist of Sn (96.5%), Ag (3%) and Cu (0.5%). Bonding is then achieved by pressing the two chips together at a defined pressure, while heating to melt the solder. This process mechanically and electrically combines the two chips together with results shown in Fig. 10.8e.

4.3 Penetrating Optrode Circuit Architecture

Two-dimensional emitter arrays are suitable for thin tissue, but not for penetrating brain structures deeper than a few hundred micrometres. As such, penetrating needs to be developed. Figure 10.10a shows an arrangement very similar to existing brain pacemakers used in deep brain stimulation. There is a control unit which has a battery, power management and microcontroller. A control line then extends to the brain-level implant and to the penetrating probes. However, for optical emission, if LEDs need to be driven, then there is a case for having local electronics to multiplex incoming signals and provide for control of multiple stimulator units. As such, the control line from the control unit can be considered as a network bus rather than as individual analog driving/sensing lines.

In some systems, communication between the control unit and the brain implant unit utilises a DC four-wire protocol such as SPI. However, if the wire were to break,

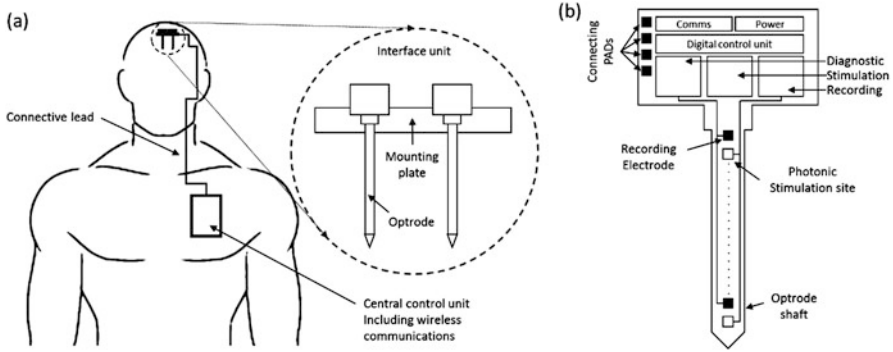


Fig. 10.10 Medical scheme for implanted optoelectrodes. (a) An array of implantable optrodes controlled by a central control unit in a similar fashion to existing brain pacemaker technology. (b) A conceptual intelligent CMOS-based optrode with communications, power management and analog circuits for diagnostics, optical stimulation and recording

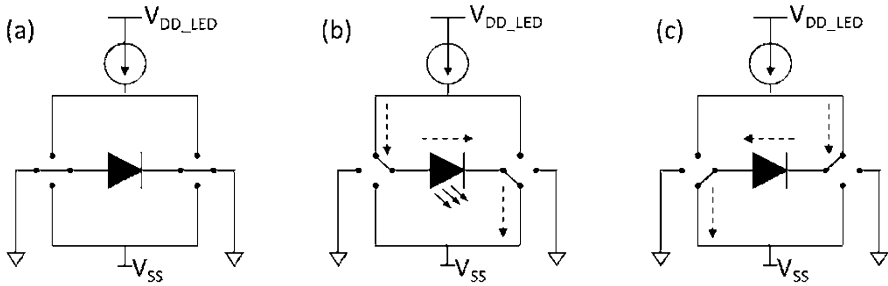


Fig. 10.11 Operational phases for an optrode LED. (a) OFF state with anode and cathode clamped to ground. (b) ON state with light emission. (c) REVERSE state with reverse current used for diagnostics

this would result in potentially dangerous DC leakage into the tissue. Thus, the best practice is to use an AC protocol which oscillates around the tissue ground. This does, however, mean a more complex power management unit on the implant.

Figure 10.10b describes how an intelligent optoelectrode can be developed. It can comprise individual electronics for communications and control, as well as power management, diagnostics, stimulation and recording. These can either be in the head or intriguingly placed along the shaft. At the time of writing, the long-term durability of doing so is not yet known and is thus the basis for further study.

The LEDs can be used to emit light as described above. One issue by doing so is that there will over the long term be a net electric field between the anode and the fluid. This may result in long-term electrolytic decay. As such, there is an advantage in specifically clamping the anode and cathode to the same potential as the tissue ground when not in use (Fig. 10.11a). Then during operation, two phases can be used: the first is shown in Fig. 10.11b where the LED is driven in forward direction. The second is shown in Fig. 10.11c with the LED driven in reverse direction. Very

little current will flow, but the objective is to balance the field on the anode and cathode contacts. A final point to note is that the figure describes a current driving process. However, as described previously, in most cases, the transistor will actually be in triode operation, and the current will be limited by the voltage drop across the series resistance.

An additional point to note is that the reverse phase described in Fig. 10.11c can be utilised as a sensor, as can the forward phase if it is below the LED threshold. Figure 10.6a shows the equivalent circuit for the LED. In reverse mode, current will pass through the series and shunt resistances. These will vary according to environmental conditions such as temperature and corrosion of the anode and cathode contacts. As such, Fig. 10.11 describes the concept of an H-bridge circuit which allows for driving the LEDs in both forward (LED ON) and reverse (LED REVERSE/DIAGNOSTIC) direction. Furthermore, as the anode and cathode are typically close to the optrode surface, it is best to ensure they are at the same potential as tissue ground when not in use. That is, this should reduce electrolytic degradation mechanisms.

Figure 10.12 shows the circuit diagram of the designed biphasic micro-LED driver which consists of digital to analog converter (DAC), an inverting amplifier and an H-bridge circuit. The inverting amplifier in combination with either M1 or M2 acts as a transconductance amplifier. As M2 provides a much smaller current in reverse mode, it is much smaller than M1. Diagnostic circuitry can be comprised of an amplifier connected to the anode (for sub-LED-threshold analysis) and cathode, then to an analog to digital converter (not shown). The three states can be described in the table below:

	M1	M2	M3	M4
OFF	OFF (V_{DD_LED})	OFF (V_{DD_LED})	OFF (V_{SS})	OFF (V_{SS})
ON	DAC→AMP	OFF (V_{DD_LED})	OFF (V_{SS})	ON (V_{DD_LED})
REVERSE	OFF (V_{DD_LED})	DAC→AMP	ON (V_{DD_LED})	OFF (V_{SS})

In addition, if the DAC→AMP cannot reach V_{SS} , then an additional switch can be used to clamp M1/M2 to V_{SS} to provide maximum current.

Such structures can be fabricated directly from CMOS as illustrated in Fig. 10.13. The circuits could be placed directly in a single system or as two separate pieces which are bonded together to ensure hermeticity against the tissue fluids. Electrode materials such as platinum or iridium oxide can be deposited onto a bondpad opening to form an electrode for recording. Bonding of micro-LEDs cannot take the same form as per Fig. 10.9, as lead is not biocompatible. As such, gold-tin eutectic methods can be used to lower the melting point of gold to allow bonding at a temperature that does not damage the other layers in the CMOS. An encapsulation layer is then required to protect the structure from fluids. Typically, this can be from silicone or parylene coatings.

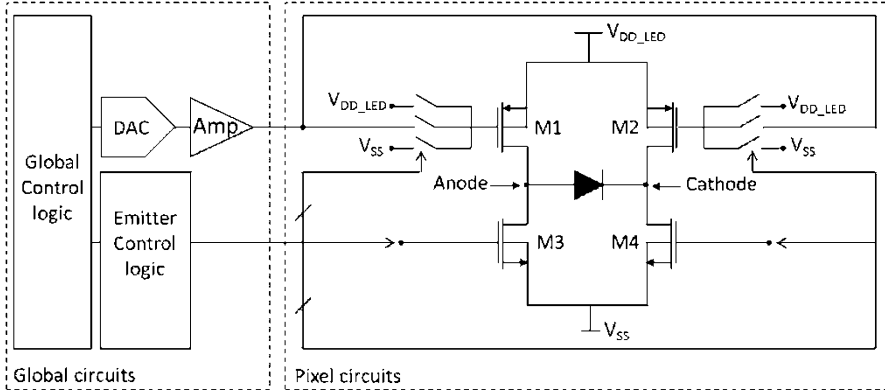


Fig. 10.12 A CMOS pixel driver for penetrating optoelectrodes. The left section is global to the optrode – consisting of logic control digital to analog conversion and an inverting amplifier. The pixel circuit consists of an H-bridge with analog input and switches to determine the direction of current flow according to Fig. 10.11

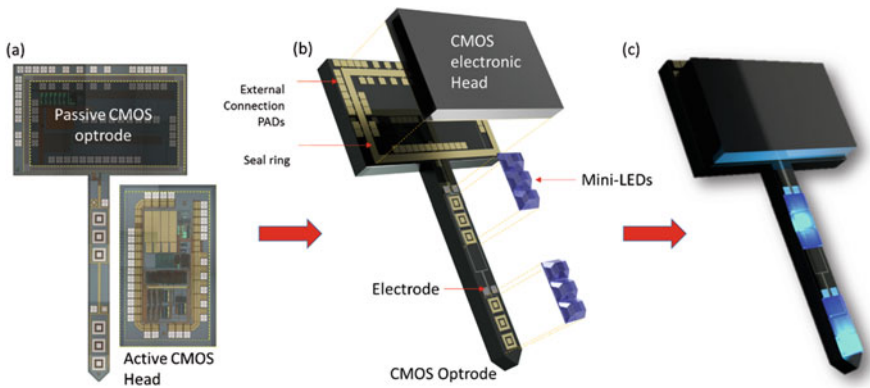


Fig. 10.13 CMOS optrode fabrication. (a) CMOS passive optrode and active head die. (b) Assembly of the CMOS head and mini LEDs. (c) An impression of how the optrode would look

5 Applications

The primary application of any neuroscience technique is to further our understanding of the basic science itself. The billion euro and billion dollar brain research programmes in the USA and EU at the time of writing are a testament to the importance given to this field. The advent of optogenetics has allowed considerable new discoveries in the field. Figure 10.13a shows a bar chart of the progression in terms of research papers. In the initial decade, the focus was on the technique, which has now matured. In vivo tools are still being developed, but the focus is now very much on the development of the science and how best to use the tools already developed.

Perhaps the most exciting implication of the optogenetics field is the impact to new forms of neuroprosthetic treatment. Optogenetic neuroprosthetics significantly increases complexity by requiring the use of gene therapy in addition to optoelectronics. However, genetic targeting provides a significant advantage. Furthermore, with simultaneous optical stimulation and electrical recording, closed-loop control can allow treatment for a number of conditions which are difficult at present.

Figure 10.13b shows the application space for neuroprosthetics. Each of these has potential to be benefited by optogenetic approaches. But in particular, visual prosthesis which has made limited progress over the years and closed-loop brain pacemakers are the primary targets for early adoption.

5.1 *Optogenetic Retinal Prosthesis*

Retinitis pigmentosa is a genetic condition that can be caused by a variety of aberrant genes and has a typical prevalence of 1:3000. It initially destroys the night vision rod cells in the retina which in turn leads to decay of day-vision cone cells. Without the capacity to transduce incoming light, individuals become blind. In 1992, Stone and co-workers noticed that the communication cells and a portion of the processing cells in the retina of those with retinitis pigmentosa were still intact [43]. The possibility of neuroprosthetic restoration was therefore born and has been the basis for significant effort in the last quarter century. At the time of writing, companies have managed to create and commercialise clinical grade retinal prosthesis: Second Sight in the USA and Retina Implant AG in Germany [44, 45]. These systems work and represent a tremendous advance in the field. However, the restored sight is very poor. In ‘pixel’ terms, the highest resolution is 1500 compared with 120,000,000 for a normal eye [46, 47]. Key bottleneck issues include:

ON/OFF pathway separation: The retina separates information into ON and OFF pathways. Information is contained in the difference between the two, but electronic stimulators tend to stimulate both similarly.

Electrode size: For any given electrode material, there will be a charge density limit beyond which degradation occurs, limiting electrode size and thus resolution.

Curvature: There is a limit to the field of view that can be restored when inserting flat chips into a curved retina. Surgically implanting a hemispherical sheet as thin as a human hair is equally difficult.

Optogenetic retinal prostheses are fundamentally different in that a gene therapy step is required to make cells light sensitive. Then, given that the eye is transparent, the light stimulus can be provided from externally without the requirement for invasive implantable electronics [28]. A detailed recent review of the state of the field was provided by Barrett et al. in 2014 [48].

The earliest attempt at an optogenetic retinal prosthesis was by Bi et al. [21], who demonstrated expression primarily in the retinal ganglion cells of the retina of a rd1 breed of blind mouse. The required thresholds were 1 mW/mm^2 which

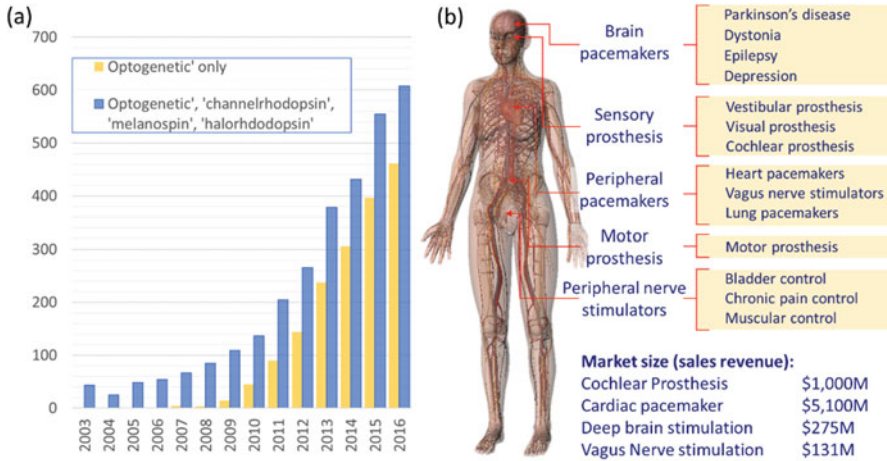


Fig. 10.14 (a) Journal publications in the optogenetics field. (Yellow) Publications with the search term 'optogenetics' – a term coined in 2005. (Blue) Publications with the search term 'optogenetics', 'channelrhodopsin', 'melanopsin' or 'halorhodopsin'. Note that there were ongoing studies into the basic molecular biology or cells from which they were derived prior to the field of optogenetics taking off. (b) Applications for neuroprosthetics

as per expected for channelrhodopsin. Later in 2008, Lagali et al. demonstrated targeted expression in ON type bipolar cells with an operational threshold of 10^{15} photons/mm²·s (100 μ W/mm²). In 2010, Busskamp demonstrated resensitization of cone cells via transfection with halorhodopsin [49]. In this case, photon fluxes of around 10^{14} photons/mm² s (10 μ W/mm²) for 5–10 ms were required to evoke strong responses. The demonstration of all layers being able to be stimulated gave impetus to clinical efforts. Furthermore, as per above, it turns out that the architecture of the retina allows lower thresholds of 0.1 mW/mm². This is still much higher than can be supplied with a typical virtual screen (~ 1 μ W/mm² on the retina), but advances in future opsin sensitivity may help further.

The concept of how an optogenetic retinal prosthesis could be implemented is illustrated in Fig. 10.14. It would comprise of the following primary components:

Gene therapy: Photosensitization of a chosen cell layer and type, e.g. ON bipolar cells using a gene therapy approach such as Adeno Associated (AAV) or *Lenti Virii*.

Stimulator: A high-radiance (>100 kcd/m² [20]) stimulator which could comprise of a laser scanner, spatial light modulator [6] or high-radiance micro-LED array [26, 41].

Headset: A wearable headset with a camera can be adapted from existing augmented reality systems to incorporate the stimulator and project light to the resensitized retina.

Embedded electronics: Much of the required processing can be performed on modern mobile central/graphics processor platforms with a microcontroller for the stimulator.

Software: Image processing software is required to maximise the useful information [50, 51] and present in the form that the retina is expecting [52, 53].

The key caveat is that optogenetics requires significant radiance for operation – equating to the midday equatorial solar irradiance. As such it needs to be enhanced via the use of a headset. Such headsets can be adapted from or benefit from the ongoing rapid developments in consumer-level virtual and augmented reality headsets.

At the time of writing, two companies, GenSight Biologics (France) and RetroSense (USA) are actively performing clinical trials. Both are utilising gene therapy to photosensitize new layers of the retina, but as yet, the outcome of these clinical efforts has not been published. However, earlier preclinical and early-stage clinical trials of gene therapies for various retinal disorders have largely not reported any adverse effects of retinal transfection [54]. Thus, there is strong hope that optogenetic approaches could significantly improve upon the visual return compared to existing devices.

5.2 *Visual Brain Prosthesis*

Compared to retinal prosthesis, visual brain prosthesis is much more invasive and potentially dangerous to the user. A key principle of medical care is to do no harm. In the case of retinal prosthesis, if something were to go wrong, the eye, which was in any case nonfunctional, can simply be removed. In the case of the brain, this is not straightforward. However, if the communication path between the eyes is destroyed, as is the case with optic neuropathies such as glaucoma, it is no longer possible to communicate from the eye, and a deeper visual brain prosthesis is required. The point of stimulation may then be in the lateral geniculate nucleus [55], which is the visual part of the thalamus, for the visual cortex itself [56] (Fig. 10.16).

Visual brain prosthesis is a much older technique with the first recorded stimulus of the visual cortex by Förster in 1929 [57]. A series of impressive experiments were performed in the 1960s and 1970s under Brindley and Lewin [58] and Dobbie and Mladejovsky [59], respectively. The results were impressive for their time, but the field stalled because of the significant technological challenges. Furthermore in 1992, Stone et al. [43] discovered that the many of the non-photosensory cells in the retinae of those blinded by the retinitis pigmentosa disease were still intact. Thus, the field shifted its attention to the less-invasive retinal prosthesis.

A clinical implementation of optogenetic visual cortical prosthesis is illustrated in Fig. 10.15. It has similarities to retinal prosthesis, but the stimulator is now a brain-penetrating optrode array, and part of the electronics would need to be implantable. The primary system components are thus:

Gene therapy: Photosensitization of a chosen cell layer in the visual cortex such as V1, using a gene therapy approach such as Adeno Associated (AAV) or *Lenti Virii*.

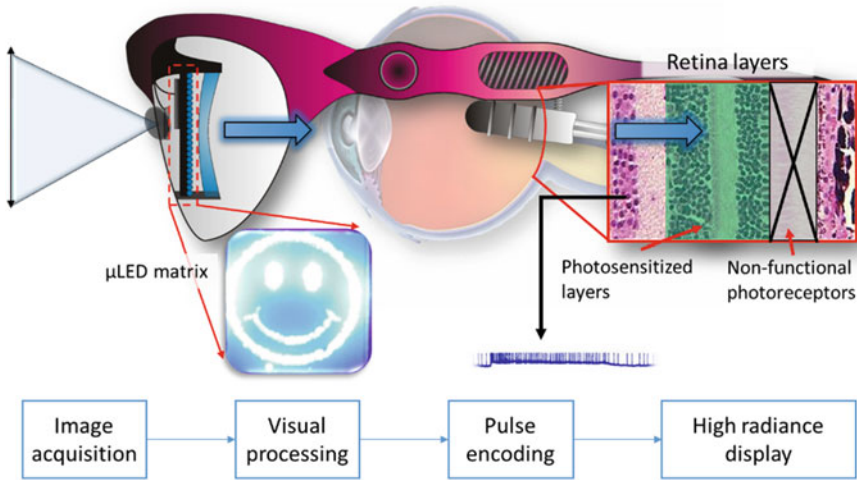


Fig. 10.15 Optogenetic retinal prosthesis. Optogenetics can be used to photosensitize bipolar cells or retinal ganglion cells in the retina. An external headset can then be used to capture the visual scene, process it and transmit it to the retina via ultrabright screens

Stimulator: An optrode array capable of stimulating brain tissue and surviving the corrosive conditions in brain tissue.

Headset: A wearable headset with a camera, embedded processing unit and wireless data/power transmission to the internal microcontrol unit.

Implantable microcontrol: An encased implantable unit which would receive video data which would need to be arranged in a form for driving the optrode array stimulator.

Software: Image processing software is required to maximise the useful information [50, 51] and present in the form that the visual cortex is expecting [52, 53].

The resolution that can be returned with visual prosthesis is debatable. On the one hand, it would be best to utilise dense arrays. But this may cause negative tissue responses. Furthermore, it may make probe insertion mechanically more difficult [60]. At the time of writing, functional optrode arrays are still in the process of development. Making such systems robust and functional in long-term chronic conditions will be challenging. But once solved will allow for rapid progress in the field of visual cortical prosthesis.

6 Final Notes and Future Perspectives

The discovery of Chr2 has been one of the most important findings in neuroscience for some time and parallels the discovery of GFP. Since its inception, the field of optogenetics has exploded (see Fig. 10.13). The tool has been adopted by

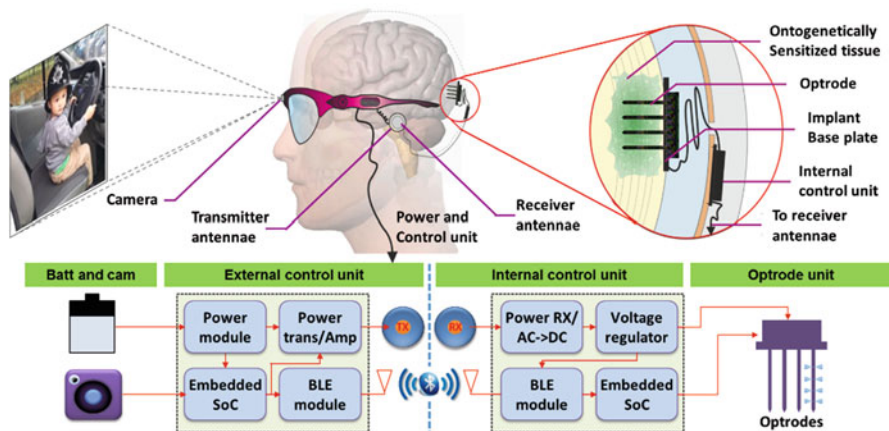


Fig. 10.16 Optogenetic visual cortical prosthesis. The target brain tissue – typically V1 of the visual cortex – needs to be sensitized. Then there will be external components managing video acquisition, processing and power-data transfer. Internally there needs to be a control unit and an optrode unit

neuroscience laboratories around the world to study a variety of basic science questions in numerous species. In 2010, optogenetics was chosen as the ‘Method of the Year’ across all fields of science and engineering by the interdisciplinary research journal *Nature Methods* [61]. At the time of writing, the first human trials to use optogenetic neuroprosthetics to restore sight to the blind are underway, with further trials planned for other disorders.

References

1. G. Nagel, T. Szellas, W. Huhn, S. Kateriya, N. Adeishvili, P. Berthold, et al., Channelrhodopsin-2, a directly light-gated cation-selective membrane channel. *PNAS* **100**, 13940–13945 (2003)
2. E.S. Boyden, F. Zhang, E. Bamberg, G. Nagel, K. Deisseroth, Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005)
3. K. Nikolic, N. Grossman, M.S. Grubb, J. Burrone, C. Toumazou, P. Degenaar, Photocycles of Channelrhodopsin-2. *Photochem. Photobiol.* **85**, 400–411 (2009)
4. N. Grossman, K. Nikolic, C. Toumazou, P. Degenaar, Modeling study of the light stimulation of a neuron cell with channelrhodopsin-2 mutants. *I.E.E.E. Trans. Biomed. Eng.* **58**, 1742–1751 (2011)
5. N. Grossman, V. Poher, M.S. Grubb, G.T. Kennedy, K. Nikolic, B. McGovern, et al., Multi-site optical excitation using ChR2 and micro-LED array. *J. Neural Eng.* **7**, 16004 (2010)
6. I. Reutsky-Gefen, L. Golan, N. Farah, A. Schejter, L. Tsur, I. Brosh, et al., Holographic optogenetic stimulation of patterned neuronal activity for vision restoration. *Nat. Commun.* **4**, 1509 (2013)

7. F. Zhang, M. Prigge, F. Beyriere, S. Tsunoda, J. Mattis, O. Yizhar, et al., Red-shifted optogenetic excitation: a tool for fast neural control derived from *Volvox carteri*. *Nat. Neurosci.* **11**, 631–633 (2008)
8. J.Y. Lin, P.M. Knutsen, A. Muller, D. Kleinfeld, R.Y. Tsien, ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013)
9. E.A. Ferenczi, J. Vierock, K. Atsuta-Tsunoda, S.P. Tsunoda, C. Ramakrishnan, C. Gorini, K. Thompson, S.Y. Lee, A. Berndt, S. Delp, K. Deisseroth, P. Hegemann, Optogenetic approaches addressing extracellular modulation of neural excitability. *Sci. Rep.* **6**, 23947 (2016). <https://doi.org/10.1038/srep23947>
10. F. Zhang, J. Vierock, O. Yizhar, L.E. Fenno, S. Tsunoda, A. Kianianmomeni, et al., The microbial opsin family of optogenetic tools. *Cell* **147**, 1446–1457 (2011)
11. A. Matsuno-Yagi, Y. Mukohata, Two possible roles of bacteriorhodopsin; a comparative study of strains of *Halobacterium halobium* differing in pigmentation, in *Biochem. Biophys. Res. Comm.*, vol. 78, pp. 237–243, 1977/09/09, (1977)
12. Y. Mukohata, Y. Kaji, Light-induced membrane-potential increase, ATP synthesis, and proton uptake in *Halobacterium-halobium* R1mR catalyzed by halorhodopsin – effects of N,N'-dicyclohexylcarbodiimide, triphenyltin chloride, and 3,5-di-tert-butyl-4-hydroxybenzylidenemalononitrile (SF6847). *Arch. Biochem. Biophys.* **206**, 72–76 (1981)
13. F. Zhang, L.P. Wang, M. Brauner, J.F. Liewald, K. Kay, N. Watzke, et al., Multimodal fast optical interrogation of neural circuitry. *Nature* **446**, 633–6U4 (2007)
14. X. Han, E.S. Boyden, Multiple-color optical activation, silencing, and desynchronization of neural activity, with single-spike temporal resolution. *PLoS One* **2**, e299 (2007)
15. K. Nikolic, J. Loizu, P. Degenaar, C. Toumazou, Noise reduction in analogue computation of *Drosophila* photoreceptors. *J. Comp. Electron.* **7**, 458–461 (2008)
16. Z. Melyan, E.E. Tarttelin, J. Bellingham, R.J. Lucas, M.W. Hankins, Addition of human melanopsin renders mammalian cells photoresponsive. *Nature* **433**, 741–745 (2005)
17. X. Qiu, T. Kumbalasiri, S. Carlson, K. Wong, V. Krishna, I. Provencio, et al., Induction of photosensitivity by heterologous expression of melanopsin. *Nature* **433**, 745–749 (2005)
18. B. Lin, A. Koizumi, N. Tanaka, S. Panda, R.H. Masland, Restoration of visual function in retinal degeneration mice by ectopic expression of melanopsin. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 16009–16014 (2008)
19. T.W. Chen, T.J. Wardill, Y. Sun, S.R. Pulver, S.L. Renninger, A. Baohan, et al., Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013)
20. P. Degenaar, N. Grossman, M.A. Memon, J. Burrone, M. Dawson, E. Drakakis, et al., Optobionic vision—a new genetically enhanced light on retinal prosthesis. *J. Neural Eng.* **6**, 035007 (2009)
21. A. Bi, J. Cui, Y.P. Ma, E. Olshevskaya, M. Pu, A.M. Dizhoor, et al., Ectopic Expression of a Microbial-Type Rhodopsin Restores Visual Responses in Mice with Photoreceptor Degeneration. *Neuron* **50**, 23–33 (2006)
22. R. Airan, K. Thompson, L. Fenno, H. Bernstein, K. Deisseroth, Temporally precise in vivo control of intracellular signalling. *Nature* **458**, 1025–1029 (2009)
23. X. Han, X. Qian, J.G. Bernstein, H.-h. Zhou, G.T. Franzesi, P. Stern, R.T. Bronson, A.M. Graybiel, R. Desimone, E.S. Boyden, Millisecond-timescale optical control of neural dynamics in the nonhuman primate brain. *Neuron* **62**, 191–198 (2009)
24. S.R. Schultz, C.S. Copeland, A.J. Foust, P. Quicke, R. Schuck, Advances in two-photon scanning and scanless microscopy technologies for functional neural circuit imaging. *Proc. IEEE* **105**, 139–157 (2017)
25. V. Poher, N. Grossman, G.T. Kennedy, K. Nikolic, H.X. Zhang, Z. Gong, et al., Micro-LED arrays: a tool for two-dimensional neuron stimulation. *J. Phys. D-Appl. Phys.* **41**, 094014 (2008)
26. A. Soltan, B. McGovern, E. Drakakis, M. Neil, P. Maaskant, M. Akhter, et al., High density, high radiance μ LED matrix for optogenetic retinal prostheses and planar neural stimulation. *IEEE Trans. BioCAS* **11**, 347–359 (2017)

27. P.P. Maaskant, H. Shams, M. Akhter, W. Henry, M.J. Kappers, D. Zhu, et al., High-speed substrate-emitting micro-light-emitting diodes for applications requiring high radiance. *Applied Physics Express* **6**, 22102–022102 (2013)
28. N. Grossman, K. Nikolic, et al., A non-invasive retinal prosthesis testing the concept. *Proc. EMBC Conf.* **2007**, 6364–6367 (2007)
29. D. Kasahara, D. Morita, T. Kosugi, K. Nakagawa, J. Kawamata, Y. Higuchi, et al., Demonstration of blue and green GaN-based vertical-cavity surface-emitting lasers by current injection at room temperature. *Appl. Phys. Express* **4**, 072103 (2011)
30. L. Chaudet, M. Neil, P. Degenaar, K. Mehran, R. Berlinguer-Palmini, B. Corbet, et al., Development of optics with micro-LED arrays for improved opto-electronic neural stimulation, presented at the *Proc. Photonics West*, (2013)
31. K. Tamura, Y. Ohashi, T. Tsubota, D. Takeuchi, T. Hirabayashi, M. Yaguchi, et al., A glass-coated tungsten microelectrode enclosing optical fibers for optogenetic exploration in primate deep brain structures. *J. Neurosci. Methods* **211**, 49–57 (2012)
32. J. Wang, F. Wagner, D.A. Borton, J. Zhang, I. Ozden, R.D. Burwell, et al., Integrated device for combined optical neuromodulation and electrical recording for chronic in vivo applications. *J. Neural Eng.* **9**, 016001 (2011)
33. T.V.F. Abaya, S. Blair, P. Tathireddy, L. Rieth, F. Solzbacher, A 3D glass optrode array for optical neural stimulation. *Biomed. Opt. Express* **3**, 3087–3104 (2012)
34. A.N. Zorzos, J. Scholvin, E.S. Boyden, C.G. Fonstad, Three-dimensional multiwaveguide probe array for light delivery to distributed brain circuits. *Opt. Lett.* **37**, 4841–4843 (2012)
35. N. McAlinden, D. Massoubre, E. Richardson, E. Gu, S. Sakata, M.D. Dawson, et al., Thermal and optical characterization of micro-LED probes for in vivo optogenetic neural stimulation. *Opt. Lett.* **38**, 992–994 (2013)
36. M. M. Doroudchi, K. P. Greenberg, A. N. Zorzos, W. W. Hauswirth, C. G. Fonstad, A. Horsager, et al., Towards optogenetic sensory replacement, presented at the 2011 IEEE EMBC conference, (2011)
37. H. Cao, L. Gu, S.K. Mohanty, J.C. Chiao, An integrated μ LED optrode for optogenetic stimulation and electrical recording. *I.E.E.E. Trans. Biomed. Eng.* **60**, 225–229 (2013)
38. H.B. Zhao, F. Dehkhoda, R. Ramezani, D. Sokolov, P. Degenaar, Y. Liu, et al., A CMOS-based Neural implantable optrode for optogenetic stimulation and electrical recording, in *2015 IEEE Biomedical Circuits and Systems Conference*, (2015), pp. 286–289
39. V.S. Polikov, P.A. Tresco, W.M. Reichert, Response of brain tissue to chronically implanted neural electrodes. *J. Neurosci. Methods* **148**, 1–18 (2005)
40. F.Y.B. Chen, D.M. Budgett, Y. Sun, S. Malpas, D. McCormick, P.S. Freestone, Pulse-width modulation of Optogenetic photo-stimulation intensity for application to full-implantable light sources. *IEEE Trans Biomed Circuits Syst* **11**(1), 28–34 (2017)
41. B. McGovern, R.B. Palmini, N. Grossman, E. Drakakis, V. Poher, M.A. Neil, et al., A New individually addressable micro-LED Array for photogenetic neural stimulation. *IEEE T. BioCAS* **4**, 469–476 (2010)
42. W. Al-Atabany, B. McGovern, K. Mehran, R. Berlinguer-Palmini, P. Degenaar, A processing platform for optoelectronic/Optogenetic retinal prosthesis. *IEEE Trans Biomed Eng* **60**(3), 781–791 (2013). <https://doi.org/10.1109/TBME.2011.2177498>
43. J.L. Stone, W.E. Barlow, M.S. Humayan, E. de Juan Jr, A.H. Milam, Morphometric analysis of macular photoreceptors and ganglion cells in retinas with retinitis pigmentosa. *Arch. Ophthalmol.* **110**, 1634–1639 (1992)
44. J.D. Dorn, A.K. Ahuja, A. Caspi, L.d. Cruz, G. Dagnelie, J.A. Sahel, et al., The detection of motion by blind subjects with the epiretinal 60-electrode (Argus II) retinal prosthesis. *JAMA Ophthalmol.* **131**, 183–189 (2013)
45. K. Stingl, K.U. Bartz-Schmidt, D. Besch, A. Braun, A. Bruckmann, F. Gekeler, et al., Artificial vision with wirelessly powered subretinal electronic implant alpha-IMS. *Proc. R. Soc. B Biol. Sci.* **280**, 20130077 (2013)
46. S. Picaud, J.-A. Sahel, Retinal prostheses: clinical results and future challenges. *C. R. Biol.* **337**, 214–222 (2014)

47. R.K. Shepherd, M.N. Shivdasani, D.A.X. Nayagam, C.E. Williams, P.J. Blamey, Visual prostheses for the blind. *Trends Biotechnol.* **31**, 562–571 (2013)
48. J.M. Barrett, R. Berlinguer-Palmini, P. Degenaar, Optogenetic approaches to retinal prosthesis. *Vis. Neurosci.* **31**, 345–354 (2014)
49. V. Busskamp, J. Duebel, D. Balya, M. Fradot, T.J. Viney, S. Siegert, et al., Genetic reactivation of cone photoreceptors restores visual responses in retinitis pigmentosa. *Science* **329**, 413–417 (2010)
50. W. Al-Atabany, B. McGovern, K. Mehran, R. Berlinguer-Palmini, P. Degenaar, A processing platform for optoelectronic/optogenetic retinal prosthesis. *IEEE T. BME* **PP**, 1–1 (2011)
51. W.I. Al-Atabany, M.A. Memon, S.M. Downes, P.A. Degenaar, Designing and testing scene enhancement algorithms for patients with retina degenerative disorders. *Biomed. Eng. Online* **9**, 27 (2010)
52. W.I. Al-Atabany, T. Tong, P.A. Degenaar, Improved content aware scene retargeting for retinitis pigmentosa patients. *Biomed. Eng. Online* **9**, 52 (Sep 2010)
53. S. Nirenberg, C. Pandarinath, Retinal prosthetic strategy with the capacity to restore normal vision. *PNAS* **109**, 15012–15017 (2012)
54. S.E. Boye, S.L. Boye, A.S. Lewin, W.W. Hauswirth, A comprehensive review of retinal gene therapy. *Mol. Ther.* **21**, 509–519 (2013)
55. J.S. Pezaris, R.C. Reid, Demonstration of artificial visual percepts generated through thalamic microstimulation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 7670–7675 (2007)
56. R.A. Normann, B.A. Greger, P. House, S.F. Romero, F. Pelayo, E. Fernandez, Toward the development of a cortically based visual neuroprosthesis. *J. Neural Eng.* **6**, 035001 (2009)
57. O. Förster, Beiträge zur Pathophysiologie der Sehbahn und der Sehsphäre. *J. für Psychologie und Neurologie* **39**, 463–485 (1929)
58. G.S. Brindley, W.S. Lewin, The sensations produced by electrical stimulation of the visual cortex. *J. Physiol.* **196**, 479–493 (1968)
59. W.H. Dobbelle, M.G. Mladejovsky, J.P. Girvin, Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. *Science* **183**, 440–444 (1974)
60. T. Parittotokkaporn, D.G.T. Thomas, A. Schneider, E. Huq, B.L. Davies, P. Degenaar, et al., Microtextured surfaces for deep-brain stimulation electrodes: a biologically inspired design to reduce lead migration. *World Neurosurg.* **77**, 569–576 (2012)
61. Nat.Meth.Editorial, Method of the Year 2010. *Nat. Methods* **8**, 1 (2011)

Chapter 11

Implantable Microsystems for Personalised Anticancer Therapy

Jamie R. K. Marland, Ewen O. Blair, Brian W. Flynn,
Eva González-Fernández, Liyu Huang, Ian H. Kunkler, Stewart Smith,
Matteo Staderini, Andreas Tsiamis, Carol Ward, and Alan F. Murray

1 Introduction

This work has its roots in an almost casual question asked by co-author Ian Kunkler in 2005. The subject of the conversation was the fact that radiotherapy can now offer targeted treatment in time and space with exquisite, pinpoint (submillimetre) accuracy. This allows radiotherapy-resistant cells, largely those in a hypoxic (oxygen-starved) state, to be given a greater dose than “normal” cancer cells. The problem is that:

- Hypoxic cells occupy small volumes.
- Hypoxic regions are not static.
- We do not know where they are.

Cancer scans, such as positron emission tomography (PET), can identify hypoxic regions with reasonable accuracy but are static measurements. There is therefore a strong and unmet clinical need for sensors that can provide a real-time, accurate,

J.R.K. Marland (✉) • E.O. Blair • B.W. Flynn • L. Huang • S. Smith • A. Tsiamis • A.F. Murray
School of Engineering, University of Edinburgh, King’s Buildings, Colin Maclaurin Road,
Edinburgh, EH9 3DW, UK
e-mail: jamie.marland@ed.ac.uk

E. González-Fernández • M. Staderini
School of Chemistry, University of Edinburgh, Joseph Black Building, David Brewster Road,
Edinburgh, EH9 3FJ, UK

I.H. Kunkler
Cancer Research UK Edinburgh Centre, Institute of Genetics and Molecular Medicine, University
of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh, EH4 2XR, UK

C. Ward
Division of Pathology Laboratories, University of Edinburgh, Western General Hospital, Crewe
Road South, Edinburgh, EH4 2XU, UK

three-dimensional map of the hypoxic status of the tumour's microenvironment. Ian Kunkler's question to the engineers was "can you do it?" The answer was "probably", and this chapter indicates how far we have come towards that goal. The first section reviews the important areas of cancer biology that impinge on the specifications of the sensors developed. Subsequent sections describe the sensor types that have been developed and the challenges addressed in integrating them with CMOS circuitry, packaging them and powering/communicating with them wirelessly. The work has been funded by a £5.2 M EPSRC Programme Grant, "Implantable Microsystems for Personalised Anti-Cancer Therapy (IMPACT)", and is not yet complete. The results reported in this chapter are, therefore, incomplete and preliminary. They do, however, highlight clearly the scientific and technical challenges posed by implanted medical sensing devices and some potential approaches to addressing them.

2 Cancer Biology and Therapy

2.1 Cancer and Radiotherapy

Cancer is a family of diseases characterised by uncontrolled cell growth. It often spreads through the lymphatic or vascular systems to distant organs such as the bone, liver, lung and brain (metastatic growth), leading to death. The International Agency for Research on Cancer (IARC) estimated that in 2012, there were 14.1 million new cancer cases, 8.2 million deaths and 32.6 million people living with cancer globally [1].

Surgery, radiotherapy and anticancer drug therapy form the cornerstone of multi-disciplinary anticancer treatment. Radiotherapy plays a key role in the management of a wide variety of common solid tumours (e.g. breast, lung, head and neck, prostate and cervix). Its aim is to eradicate cancer cells in the primary tumour and regional lymph nodes while minimising damage to normal tissues. Radiotherapy is given in approximately 50% of cancer patients and in 40% of patients who are cured. Normal tissue recovers better from radiotherapy damage than tumour tissue with curative radiation dose schedules. As a result, this differential effect (the therapeutic ratio) confers an important clinical advantage.

In clinical practice, ionising radiation is usually produced by high-energy radiotherapy treatment machines and delivered in small daily doses over several weeks.

Advances in radiation delivery allow dose to be applied with greater accuracy, sparing normal tissues. For example, stereotactic body radiotherapy (SBRT), which uses multiple radiation fields or radiation arcs, results in rapid fall-off in dose outside the target volume [2].

Sensitivity to radiation can be influenced by cancer stem cell number, oxygenation status, acidity, tumour repopulation between treatments and redistribution of surviving cycling cells following radiation-induced cell cycle blockade [2].

2.2 Tumour Hypoxia

Tumour hypoxia is important in tumour progression, metastatic growth and treatment resistance. Poorly oxygenated tumours are resistant to the effects of radiotherapy [3], with hypoxic cells three times more resistant to radiation than well-oxygenated cells [4]. Although atmospheric oxygen concentration is 21%, partial pressure is 5–9% in well-oxygenated tissues and approximately 3% in solid tissues. Hypoxia occurs at oxygen levels of 1% or less [5, 6].

Hypoxia follows when tumour growth outstrips the capacity of the vasculature to carry sufficient oxygen, allowing oxygen gradients to develop. The oxygen diffusion limit is approximately 100–200 μm from a blood vessel. Therefore, when a tumour reaches 1–2 mm^3 , diffusion limits cause hypoxic regions to form [7]. Tumour cells adjacent to blood vessels may have intermediate levels of hypoxia and impaired radiosensitivity, since the distance that oxygen can diffuse depends on the capacity of haemoglobin to release oxygen, oxygen uptake around the blood vessel and intravascular oxygen partial pressure [8]. Acute hypoxia occurs due to transient perfusion changes [9] and has been found in a range of solid tumours, including breast and head and neck cancer.

Hypoxia characterises many solid tumours [10], such as head and neck cancers, which, when treated with radiotherapy, are associated with poorer clinical outcomes [11]. These clinical observations were ascribed to the presence of hypoxic cells, which are linked with a more malignant phenotype [12]. Tumour cells adapt to hypoxia by activating the transcription factor hypoxia-inducible factor-1 (HIF-1), which induces expression of factors required for survival in the harsh conditions of the tumour microenvironment. HIF-1, composed of HIF-1 α and HIF-1 β subunits, modulates genes involved in angiogenesis, glycolysis, cell cycle, proliferation and metastasis [13]. Hypoxia and overexpression of HIF-1 α are associated with poor clinical outcome in many cancers [14].

Hypoxia (via HIF-1) increases levels of vascular endothelial growth factor (VEGF), which supports the growth of new blood vessels, which, in tumours, are frequently malformed, minimising oxygen perfusion and transportation of nutrients and wastes to and from the tumour site (Fig. 11.1). HIF-1 is also associated with acidosis in the tumour microenvironment. Activation increases glycolysis by enhancing expression of glucose transporters and glycolytic enzymes, leading to the production of lactate which is symported along with hydrogen ions from hypoxic cells via the HIF-1-regulated monocarboxylate transporter 4 [7, 15, 16].

Cell survival is dependent on alkaline intracellular pH. Acidic pH and hypoxia activate pH-regulating systems to maintain cellular pH within well-defined limits, but this causes extracellular pH to fall [17–19]. In solid tumours, pH falls with

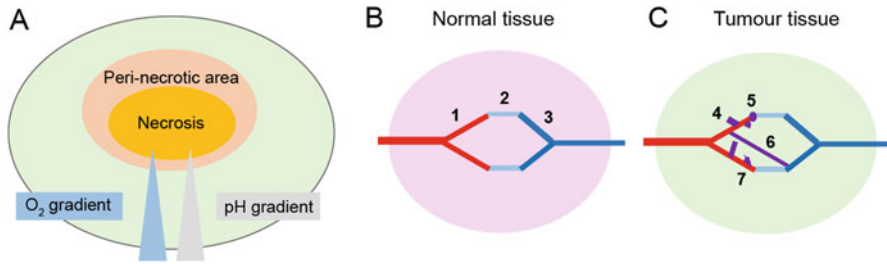


Fig. 11.1 Disrupted blood supply and hypoxia in tumours. (a) Illustrates the microenvironment in a solid tumour. Both pH and O₂ decrease with distance from the blood supply. The central core is often necrotic because of cell death. The peri-necrotic area contains dead cells and live cells which adapt to adverse conditions causing cancer progression. (b) Demonstrates the normal tissue where arterioles (1) transport oxygenated blood to the capillary bed (2) and where oxygen and nutrients are delivered and waste products removed via venules (3) to join the general circulation. (c) Shows some of the abnormalities (*in purple*) of the tumour vasculature which lead to hypoxic and acidic conditions in the tumour microenvironment: (4) blind end, (5) vessel occlusion, (6) arteriovenous shunt which bypasses the capillary bed and (7) breaks and bulges in the vessel walls

increased distance from blood vessels and can drop to pH 6.0 as lactic acid and hydrogen ions accumulate because of the malformed vasculature (Fig. 11.1) [7, 18].

Both hypoxia and acidosis cause resistance to radiotherapy. Oxygen is required to stabilise radiation-induced DNA damage [18], while lactate increases migration of cancer cells [20] and VEGF expression, thus inducing angiogenesis [21]. Lactate concentrations correlate with patient survival, metastatic growth and radioresistance in several tumours [22–24]; the antioxidant capacity of lactate may induce resistance to radiation [25].

2.3 Previously Developed Methods of Measuring Hypoxia

A number of techniques are used to measure tumour hypoxia (see review by Horsman et al. (2012), [26]). These include direct measurement of oxygen partial pressure distribution using polarographic electrodes or indirect measures such as injection of exogenous markers to label viable hypoxic cells or endogenous markers (i.e. genes/proteins regulated by hypoxia) [27]. The Eppendorf oxygen electrode can be directly implanted into tumours both preclinically or clinically, but this is an invasive procedure and has not been adopted in clinical practice.

While the Eppendorf electrode may be regarded as the gold standard in direct measurement of hypoxia, it cannot capture the three-dimensional distribution of hypoxia within a tumour which is required for hypoxic dose painting (“boosting” with extra radiation dose beyond that delivered to the rest of the tumour) nor the spatio-temporal changes in hypoxia in the dynamic microenvironment.

2.4 Real-Time Sensing of Hypoxia and Biologically Adapted Radiotherapy

Co-registered PET/computerised CT images can be utilised to identify hypoxic sub-volumes which may benefit from “dose painting” to improve eradication of local tumour, for example, in head and neck cancer (Fig. 11.2). However, these forms of imaging are all snapshots in time. The development of implanted real-time biosensors in a representative array which captured both the true level of hypoxia and its spatial and temporal changes would provide a powerful tool for realising biologically adapted radiotherapy.

This approach is being developed in the cross-disciplinary EPSRC-funded Implanted Microsystem Personalising Anti-Cancer Therapy (IMPACT) project (www.impact.eng.ed.ac.uk) and validated in veterinary ovine tumour models. This research approach presents many challenges including miniaturisation of all the components of the sensors, powering sufficient to last for 3–7 weeks of common radiotherapy scheduling, biofouling of the sensors, implantation of the sensors without significantly impairing the microenvironment or inducing an immune response. Fidelity of the signal correlating with the true biological change the sensors are seeking to measure is also key. In addition issues of toxicity, safety and patient acceptability are paramount. If proof of principle is shown in animal models, validation of the sensors in clinical trials will be needed to show safety and efficacy and improvement in tumour control and survival over and above standard care.

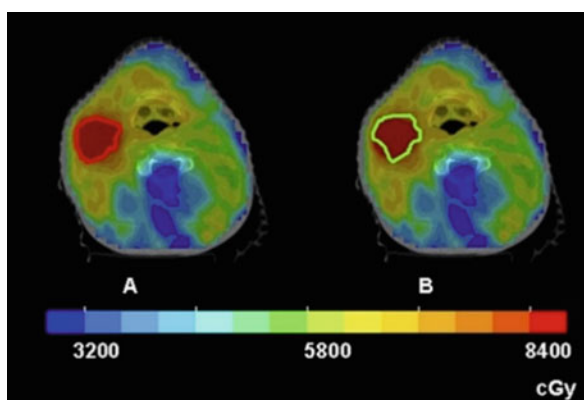


Fig. 11.2 Intensity-modulated radiotherapy dose distributions in colourwash display of patient 7, for whom the sequential hypoxia images were dissimilar. (a) Both sub-volumes of VH1 (the red contours) received 84 Gy. (b) When the same treatment plan was applied to VH2 (the green contour), part of the hypoxic volume did not receive the intended boost dose (Reprinted from, Z. Lin et al. The influence of changes in tumor hypoxia on dose-painting treatment plans based on 18F-FMISO positron emission tomography. *Int. J. Radiat. Oncol. Biol. Phys.* **70**(4), 1223, Copyright 2008, with permission from Elsevier)

3 Oxygen and pH Sensors

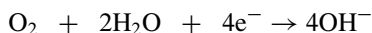
3.1 *Miniaturised Sensors for Hypoxia Monitoring*

Monitoring tumour hypoxia effectively requires miniaturised sensors for detecting dissolved oxygen concentration and pH. Microfabrication techniques developed for the microelectronics industry are ideal for this type of application. They have the advantage of being highly reproducible due to tight process control and are well suited to mass production since standard industrial processes and tools are used. If sensors are fabricated on silicon, then integration with CMOS instrumentation circuits to create a “system on chip” also becomes possible. Two sensor technologies amenable to CMOS integration are being developed for integration into the final IMPACT device: a Clark electrode oxygen sensor and an ion-sensitive field-effect transistor (ISFET) pH sensor.

3.2 *Implantable Clark Electrode Oxygen Sensors*

3.2.1 Principle of Operation

The Clark electrode oxygen sensor, first patented in 1959 [28], is an amperometric sensor. It uses the electrochemical reduction (defined as a gain of electrons or hydrogen atoms) of oxygen in an electrochemical cell to generate a measurable electric current proportional to the oxygen concentration:



Miniaturised Clark sensors have been produced by others using microfabrication methods [29–33]. They typically use a three-electrode configuration: a working electrode (WE) where the reaction of interest (oxygen reduction) occurs, a counter electrode (CE) that supplies the necessary current to balance the WE reaction and a reference electrode (RE) that provides a stable potential that the WE potential is set against [34]. In a Clark sensor, the WE is typically held at a potential of -0.7 V (against an Ag/AgCl RE), to cause the oxygen reduction reaction at its surface.

3.2.2 Electrode Materials

Platinum is commonly used as a WE and CE material in microfabricated Clark sensors because it is a noble metal and can therefore reliably support electrochemical reactions at its surface without being altered itself [29–31]. Microfabricating a reliable RE is more challenging, as it must be fabricated from a material capable of producing a fixed potential through reactions at its surface [35]. Platinum has

been used for the RE in some devices [29, 30]. This minimises the complexity of fabrication, but there are conflicting reports in the literature about whether bare platinum can provide a stable potential over time [36, 37]. Successful miniaturised liquid junction Ag/AgCl reference electrodes that can provide a more stable potential have been described [38], although their lifetime is limited to hours once the liquid electrolyte is added [39]. A feasible alternative for long-term measurements is to use an Ag/AgCl electrode covered with a solid polymer membrane to prevent rapid electrode deterioration [40], and this strategy is being pursued by IMPACT.

3.2.3 Electrolyte Designs

To allow reactions within the electrochemical cell, the circuit between electrodes must be completed by an electrolyte capable of conducting the ionic species taking part in the reaction. Early attempts at creating a microfabricated Clark sensor used a liquid electrolyte added to the sensor at the time of use, with a thin polydimethylsiloxane (PDMS) or silicone membrane to separate external media and the electrolyte [31, 38]. However, it is difficult to effectively miniaturise this design. An alternative fully solid-state approach developed more recently is to coat the electrodes in the solid polymer electrolyte Nafion [29, 30]. This is a perfluorinated ion exchange resin that is oxygen permeable and can conduct protons between electrodes [41]. We have developed a process for depositing Nafion layers from solution directly onto the sensor by spin coating, followed by a patterning step using reactive ion etching to remove excess material outside the sensor area (Fig. 11.3) [42].

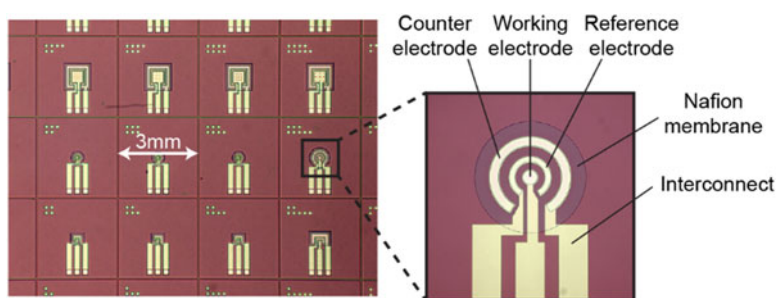


Fig. 11.3 Miniaturised Clark electrode sensors microfabricated on a 4" silicon wafer. Each 3×3 mm die contains a single sensor, interconnect and bonding pads. The sensor electrodes are covered in a patterned Nafion solid electrolyte membrane approximately 500 nm thick

3.2.4 Integration on CMOS

Within IMPACT, initial development of miniaturised Clark electrode sensors used custom microfabrication processes on blank silicon wafers. The next challenge will be integration of the sensors onto a foundry CMOS ASIC platform. Since the platinum electrode metal is not available in conventional CMOS processes, it must be deposited and patterned in post-processing steps on fabricated dies. This will be followed by fabrication of a suitable RE material and deposition of a Nafion membrane over the electrode area.

Fabrication of Clark sensors on CMOS allows integration with their instrumentation circuitry. This is typically a potentiostat circuit – a system designed to hold the WE at a constant potential relative to the RE by sinking or sourcing current at the CE as required [43]. A successful example was recently described by Chan et al., in which a Clark sensor with an array of WEs was integrated with a multiplexed input potentiostat in an implantable CMOS system on chip [29]. This potentiostat circuit design holds both the RE and CE at a fixed potential set by a band-gap reference. The WE potential is held by feedback at a potential set by a digital-to-analogue converter (DAC) (Fig. 11.4a). This arrangement allows the WE to be controllably biased relative to the RE. A current follower is used to sense the WE current, and its output voltage is amplified using a chopper-stabilised amplification stage (Fig. 11.4b). The amplified signal can then be digitised using standard ADC blocks.

3.3 Implantable ISFET pH Sensors

3.3.1 ISFET Structure and Function

Microfabricated pH sensors can be produced using ion-sensitive field-effect transistor (ISFET) semiconductor technology, first described by Piet Bergveld in 1970 [44]. ISFETs are analogous in construction and principle of operation to a conventional MOSFET. However, in an ISFET, the gate connection is left floating, and its insulator layer is exposed directly to the solution being sensed (Fig. 11.5a). The insulator material (typically silicon dioxide or nitride) is designed to have a surface charge that varies with pH. The charge influences conduction within the channel region, which can be detected using straightforward instrumentation circuitry.

To allow integration of the ISFET with on-chip instrumentation, a CMOS compatible fabrication process may be used that takes advantage of the metallisation layers available in the process to bring the gate connection up to the surface of the integrated circuit. There the top insulation acts as pH-sensitive layer, and depending on the application, it may be necessary to alter the fabrication process.

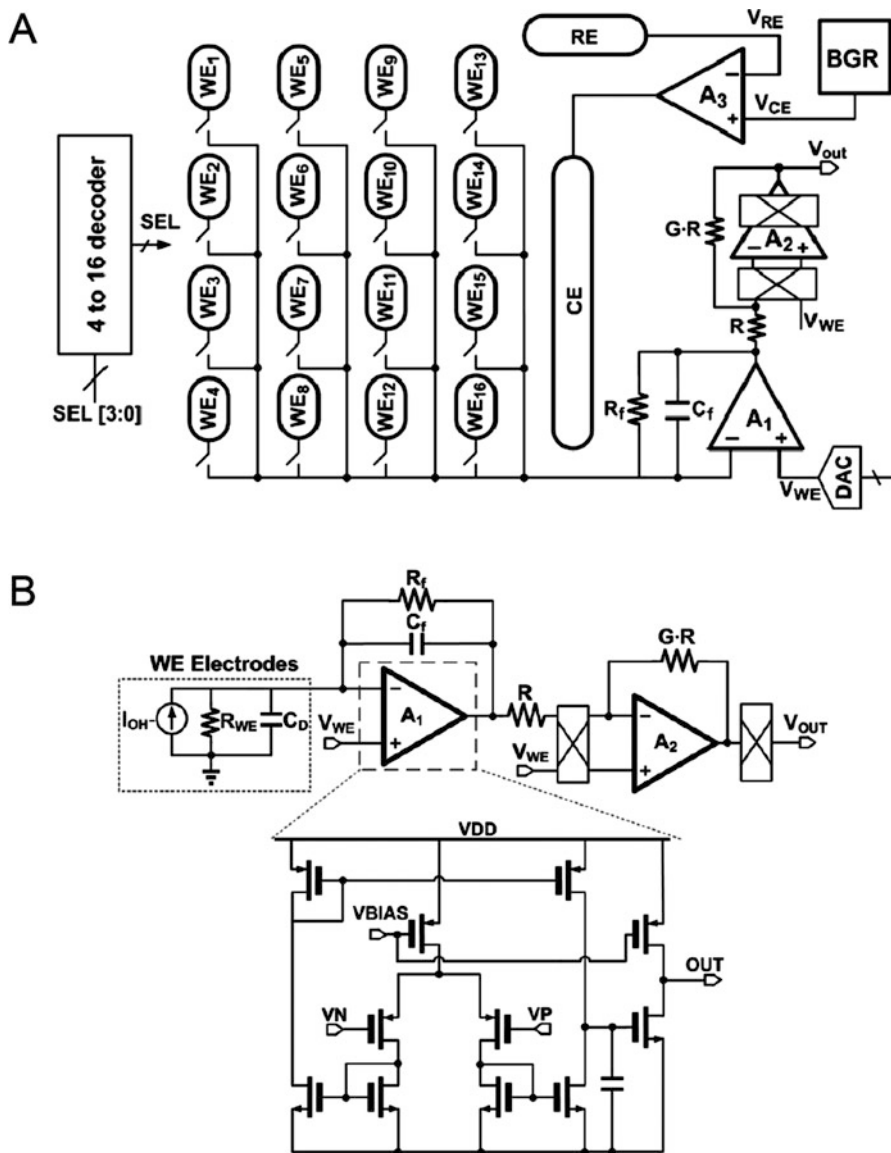


Fig. 11.4 CMOS instrumentation electronics for an integrated electrochemical Clark sensor. (a) Multiplexed potentiostat driving a WE array. (b) Current follower for sensing and amplifying WE currents (Adapted from [29], reproduced with permission of IEEE)

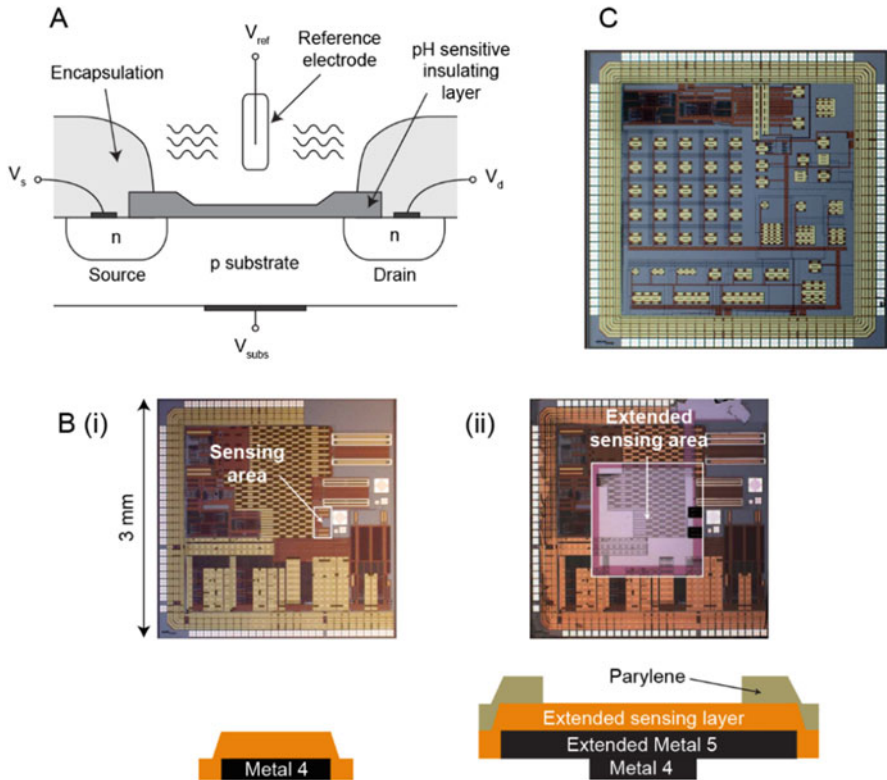


Fig. 11.5 IMPACT ISFET pH sensors. (a) Schematic showing the structure of a pH-sensitive ISFET. (b) Chip-level ISFET sensor post-processing, showing images of processed chip and schematic of extended layer structure (i) before, and (ii) after post-processing. (c) Optical microscopy image of a 3×3 mm IMPACT ISFET sensor chip

3.3.2 ISFET Fabrication for the IMPACT Implantable Device

IMPACT uses a commercial foundry to fabricate the ISFET-based application-specific integrated circuit (ASIC) chips designed for the project. The process uses $0.35 \mu\text{m}$ CMOS technology, with a Si_3N_4 sensing layer, offered by AustriaMicrosystems. Foundry-fabricated CMOS sensor chips were used to develop a set of multistep photolithographic on-chip post-processing techniques. These are not normally available in standard commercial processes and allow extended chip designs to be produced. Our work has focused on characterising the patterning, deposition and etching conditions of the devices. The clean room protocols we developed allow sensor integration or modification at chip level. Moving away from full wafer scale provides the opportunity to cost effectively investigate a large number of sensor designs. Figure 11.5b shows an example of a CMOS foundry chip, prior and after going through a number of post-processing steps, to modify an ISFET sensor.

3.3.3 Challenges and Solutions in ISFET Design

To date, three generations of chip designs have been taped-out, containing multiple ISFET devices that serve as a testing platform for characterising the sensors. Figure 11.5c presents a first-generation 3×3 mm fabricated ASIC die. However, in this generation, the sensors suffered from problems with trapped charge on the floating ISFET gates. Multiple solutions for addressing this problem were trialed in the second generation, such as sensors with a capacitively coupled control gate that tunes the operating point of the ISFETs or sensors with openings through all the metallisation layers, to allow discharging of the gates by UV light. Other examples of different ISFET sensor implementations, which have been designed and fabricated on the IMPACT chips, include:

- Differential sensor arrangement that can be integrated with a post-processed on-chip pseudo-reference electrode
- A 5×5 array of nominally identical sensors
- Sensors with proposed read-out instrumentation electronics
- Sensors with modified process designs and parameters

3.3.4 ISFET Instrumentation

Although an ISFET can be characterised as an isolated FET, integration with CMOS read-out instrumentation electronics is a common practice. One of the proposed circuit designs for IMPACT that can read out the pH-sensitive gate voltage of an ISFET is based on the work presented by Nakazato et al. [45, 46]. Figure 11.6a shows a schematic of the proposed source-drain follower. Transistors in groups (P_1, P_2), (N_1, N_2, N_3), (N_4, N_5, N_6) and (N_7, N_8, N_9) are nominally identical. N_7 to N_9

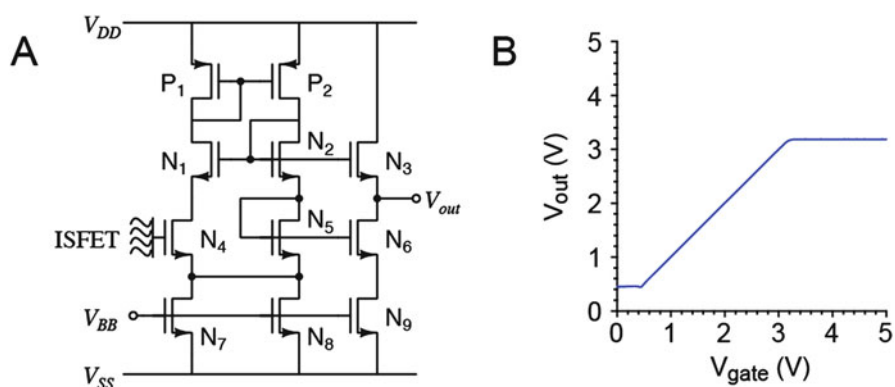


Fig. 11.6 ISFET instrumentation circuits. (a) Schematic of source-drain follower to read out pH-sensitive gate voltage on an ISFET. (b) Correlation between pH-sensitive ISFET gate voltage and output from proposed read-out electronics

are current sources controlled by a bias voltage (V_{BB}). The current and voltage in N_4 and N_5 are equal. The output (V_{out}) is equal to the pH-sensitive gate voltage of the ISFET N_4 . Figure 11.6b shows an example measurement from a fabricated device, correlating sensing device gate voltage and output from the proposed read-out electronics.

4 Biomarkers and Electrochemical Sensing

4.1 Basic Principles of Electrochemical Biosensors

4.1.1 Introduction to Electrochemical Biosensors

A biosensor can be defined as a self-contained device able to provide analytical information which integrates a biochemical recognition strategy with a physico-chemical transducer [47]. The bio-recognition element (e.g. antibody, enzyme) provides selectivity against the target analyte, while the transducer (e.g. optical, mass) allows conversion of the output from the molecular recognition event into a measurable electrical signal. If the transduction element is an electrode, the device is referred to as an electrochemical biosensor. Electrochemical biosensors have been attracting attention within the field of point-of-care (PoC) device development due to their potential to provide rapid, sensitive, low-cost and portable means of sensing for a huge range of biomarkers.

4.1.2 Classification According to Molecular Recognition Element

The molecular recognition element plays a key role in biosensor architecture, conferring selectivity against the target molecule. A huge range of recognition elements can be used as receptors to build up an electrochemical biosensor including enzymes (enzymatic sensors), antibodies and antibody fragments (immunosensors), nucleic acids (genosensors), aptamers (aptasensors), peptides and proteins, etc.

The use of enzymes as molecular recognition elements relies on the catalytic conversion of the substrate (analyte) to give a product (Fig. 11.7(1)). Enzymes have highly specific binding pockets allowing for high selectivity against the analyte. Within this group, glucose sensors should be highlighted, as they have been extensively studied in the electrochemical biosensor field [48]. Other enzymes that have been employed include lactate dehydrogenase [49], lactate oxidase [50] or urease [51].

Antibodies are proteins produced by the immune system that show high affinity towards their antigens (analytes). Immunosensor architectures typically require antibody (or the antigen-binding fragment) immobilisation onto a solid surface while maintaining its binding affinity [52]. One of the most common immunosensing designs is the sandwich assay [53, 54], based on the use of two antibodies for

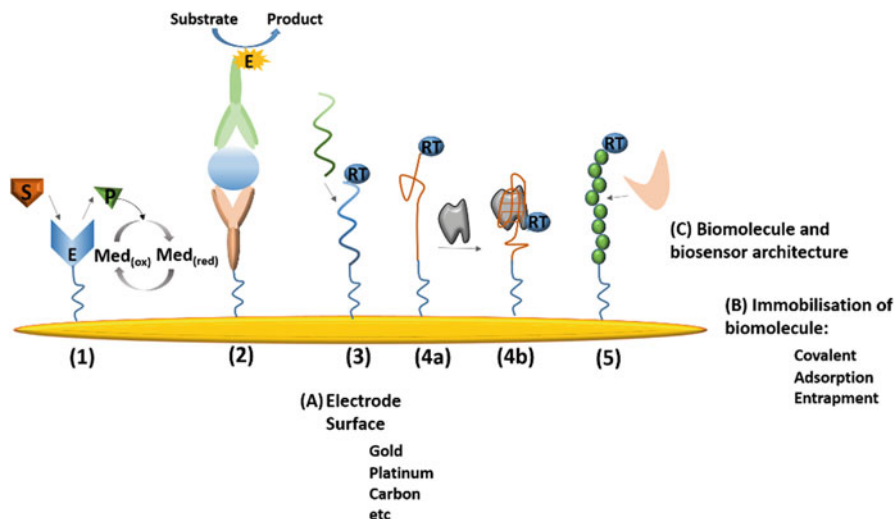


Fig. 11.7 Different components and architectures of electrochemical biosensors. (1) Enzymatic sensor using a redox mediator to shuttle the electrons between the enzyme and the electrode, where the enzyme is immobilised and converts a substrate into a product. (2) Sandwich immunosensor architecture that requires two antibodies: capture, immobilised onto the solid support and reporter antibody, which in this case is enzymatically labelled. (3) Nucleic acid-based hybridisation biosensor. The capture probe is labelled with a redox tag. (4) Aptamer-based biosensor. The redox-labelled aptamer will fold into a 3D conformation upon target binding (4a–4b). (5) Peptide-based affinity biosensor for protein detection. Binding of the target protein will lead to a change in conformation/dynamics of the redox-labelled peptide

the same analyte that gets “sandwiched”, providing high selectivity (Fig. 11.7(2)). Enzymatic labels are commonly used (e.g. ALP, HRP), and can be detected by addition of a substrate that will be transformed to a detectable electroactive product.

Nucleic acids are also well-studied bioreceptors for the construction of electrochemical biosensors able to detect specific DNA/RNA sequences. They require the immobilisation of a labelled capture probe (linear or stem-loop), single-stranded DNA or PNA [55] (peptide nucleic acids, which have a neutral backbone and show higher affinity for complementary sequences). Hybridisation with the complementary target sequence translates in a change of the electrochemical signal generated by the redox reporter [56] (Fig. 11.7(3)). Aptamers are also nucleic acid-based recognition elements; however, they are *in vitro* selected for binding non-nucleic acid molecules, from small molecules to cells, by a process called SELEX [57, 58]. They consist of a short (15–40 bp) single-stranded DNA or RNA oligonucleotides that folds in a 3D conformation that selectively binds the target molecule. Aptamers have been extensively used in the development of electrochemical biosensors in a variety of architectures including sandwich assays and binding-induced conformational change systems (Fig. 11.7(4)). The range of biomolecules that can be used as molecular recognition elements in electrochemical

biosensors is not constrained to those already mentioned and include others such as proteins and peptides [59] (Fig. 11.7(5)), which can act as a substrate for a certain enzyme, whole cells or synthetic molecularly imprinted polymers (MIPs).

4.1.3 Electrochemical Transduction Methods

Electrochemical transducers allow the use of multiple techniques in order to monitor a binding-induced change. Depending on which electrochemical characteristic is monitored, four main types of electrochemical biosensors can be considered: amperometry/voltammetry (measurement of current), impedance (resistance to electron transfer), conductometry (electrical conductivity) and potentiometry (potential or accumulated charge) [60].

Amperometry is generally used in enzymatic sensors or in assays that employ an enzymatic amplification. It is a simple technique, consisting in the measurement of a current due to the oxidation/reduction of an electroactive species at a fixed potential with respect to a reference electrode. In voltammetry, a current is also measured, but in this case, the potential applied to the working electrode is ramped with time over a potential window. This sweep of potential will result in the oxidation/reduction of an electroactive species, generating a current that is plotted against the applied potential and generally shows a peak or plateau proportional to the concentration of such species. Different potential profiles can be applied leading to different voltammetric techniques such as linear sweep voltammetry (LSV), cyclic voltammetry (CV), differential pulse voltammetry (DPV), square wave voltammetry (SWV) and alternating current voltammetry (ACV). The choice of technique is led by the specific requirements for each individual application. CV is widely used to understand electron transfer mechanisms, while pulse methods are typically more rapid and sensitive due to their ability to discriminate between faradaic and non-faradaic (charging) current.

Electrochemical impedance spectroscopy (EIS) [61, 62] allows investigation of the electrical resistance and capacitive properties of a system upon its perturbation with a small amplitude sinusoidal AC excitation signal (2–10 mV) at different frequencies in order to obtain an impedance spectrum. It has proven to be an adequate technique to monitor surface binding events without the need for redox-labelled components.

Conductometry measures changes in the conductivity of a solution connecting two electrodes as the result of a chemical reaction. It is basically used in enzymatic reactions that involve consumption/production of charged compounds, causing a change in the ionic composition of the solution that will be detected measuring the conductance between two electrodes [63, 64]. One of the main advantages of this technique is that no reference electrode is required.

Potentiometric sensors rely on measuring the potential of an electrochemical cell under conditions of negligible current flow. They are mainly based on the use of ion-selective electrodes (ISE) coated with a biological element, typically an enzyme that will either consume or produce the ion for which the selected ISE is specific for, such as H^+ (pH) or NH_4^+ .

4.2 A Practical Example: Design of Peptide-Based Electrochemical Biosensors for Protease Detection

Proteases are enzymes that play a well-established and important role in cancer progression and the apoptotic pathway, a tumour hallmark [65]. In this context, electrochemical biosensors containing peptides as recognition element have proven to be valuable tools for direct detection of enzymes [66]. Thus, detection of proteases was achieved by using peptide-based biosensors composed of a redox-labelled peptide immobilised onto an electrode surface that undergoes an enzymatic cleavage by the target enzyme. This triggers the release of the redox-labelled probe from the electrode surface leading to a measurable signal decrease (Fig. 11.8) [67]. Building on this approach, we developed an electrochemical peptide-based biosensor employing trypsin as a model protease. To this end, we modified a cleavable short peptide sequence (for trypsin) (Phe-Arg-Arg (FRR)) by adding a redox reporter at one end and an anchor containing a thiol moiety (cysteine) for attachment to the gold electrode at the other. Importantly, we also inserted a flexible spacer within the probe connecting the anchor moiety to the peptide sequence in order to promote the approach of the redox reporter to the electrode, thereby facilitating electron transfer and enhancing the measured electrical current (Fig. 11.8) [59].

We carried out a systematic study to explore the influence of the nature of the redox reporter and spacer (alkyl or PEG-based chain) on biosensor performance. Furthermore, in an effort to prevent non-specific binding to the probe film surface, our probes relied on a self-assembled monolayer (SAM) whose composition and structure were investigated to improve their antifouling capabilities. For the redox reporter, a comparison between ferrocene (Fc) and methylene blue (MB) was carried out. MB-tagged peptide probes combined with a polyethylene glycol (PEG)-based spacer showed enhanced electrochemical performance when compared to Fc-tagged probes, displaying a more stable background and reproducible response. We also

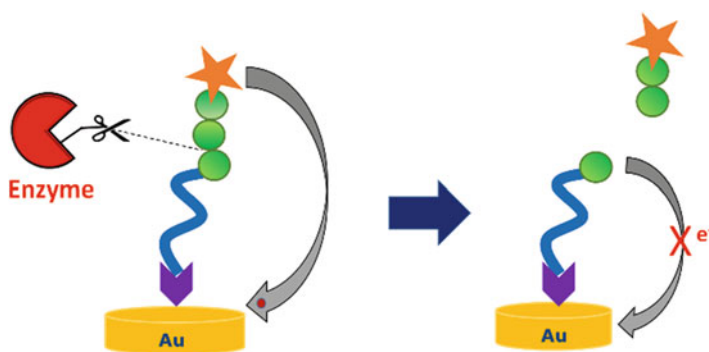


Fig. 11.8 Principle of detection of the peptide-based electrochemical sensor for protease detection

demonstrated that a sensing platform consisting of a ternary SAM configuration using a PEG-based dithiol as a co-adsorbent was able to minimise non-specific adsorption of proteins and was sensitive towards trypsin with a clinically relevant limit of detection (LoD) of 250 pM. In addition, further kinetic analysis, and the performance of a negative control with an uncleavable D-amino acid-based probe, confirmed the specificity of our probes. Altogether, these findings indicate that the use of methylene blue as a redox reporter, combined with a PEG spacer and immobilised onto an electrode surface as a ternary SAM configuration, represents a valuable platform for the generation of clinically relevant and quantitative electrochemical peptide-based protease biosensors.

4.3 Challenges and Opportunities of Integration on CMOS

In more recent times, the use of nanotools for electrochemical detection in a biological setting has come to the attention of the scientific community. Several efforts are being made to integrate miniaturised sensors with microelectronic technology, to create a chip-based sensing platform. In this context, CMOS process technology is widely used for the generation of modern integrated circuits, including microprocessors and microcontrollers offering great performance. Therefore, we are interested in the development of CMOS-integrated electrochemical biosensors for biological sensing [68].

Integration of electrochemical biosensors is a complex process due to multiple challenges concerning miniaturisation, the low currents involved, reference electrode stability, surface functionalisation and non-specific adsorption. A reference electrode, e.g. a silver-silver chloride (Ag/AgCl) electrode, has a stable and known potential and is used to control the potential applied to the working electrode [69]. However, its miniaturisation is not easy because of the difficulty to pack it within a chloride ion solution. Although some promising outcomes have been observed, further improvements are required in order to have a miniaturised stable reference electrode.

As mentioned above, improving the anti-fouling properties of the sensing platform is an important challenge in the development and integration of electrochemical biosensors. In an effort to avoid non-specific adsorptions from complex matrices, self-assembled monolayers containing anti-fouling features such as polyethylene glycol [70] or zwitterionic SAMs have been used [71]. In the particular case of electrochemical biosensors, the main challenge is producing a low-thickness coating to avoid hampering the electron transfer process while displaying low non-specific adsorption of proteins and good stability and reproducibility.

5 Packaging

5.1 Introduction

Packaging is a key component of any microelectronic system which needs to be protected from its environment [72–74]. This is particularly true of sensor systems, which typically require that parts of the electronics are in direct contact with its environment [75]. In order to achieve this, the insulation material which constitutes the packaging must be patterned to open windows directly to the chip surface [76, 77]. The packaging process is especially important in a liquid environment, where any exposed non-sensing electronics such as wire bonds will likely result in device failure. Other requirements for these materials include:

- Being an effective barrier to the surrounding environment
- Being physically durable
- Adhering well to the surface of the chip
- Being compatible with standard microfabrication or post-CMOS processing techniques

Adhesion of the insulation material to the chip surface is of particular importance. This is because the chip-package interface will be exposed to the solution when patterned. Ingress of the liquid along this interface will eventually cause the device to fail.

In the case of biosensors that are to be implanted in the body, this type of failure could also expose the organism to undesirable materials. It is therefore critical that not only the insulation material not fail while implanted but that the package itself is biocompatible. A biocompatible material is one which is not only non-toxic but does not cause any response from the body which would perturb the measurement site [78–81]. This can arise from inflammation or biofouling, where the body treats the sensor as a foreign object and encapsulates it in biomaterial [78–81]. Surface chemistry of the material is not the only factor in this response; some others include surface roughness, geometry, porosity and position in the body [79, 80]. It is therefore vital that the immune response to the chosen packaging materials is understood before any implantation is attempted. These tests can be exhaustive and typically consist of immersing samples of the material in both in vitro cell cultures and in vivo host bodies and studying the response from the body [79, 80]. The biological evaluation and testing of medical devices and their components are laid out in ISO 10993 [82].

Typical materials of choice are polymers, rubbers and resins [83, 84]. Commonly used examples of these include Parylene-C and PDMS [84, 85]. These all have strengths and weaknesses, for example, Parylene-C has good chemical resistance and can be deposited in a conformal, pinhole-free manner [86, 87]. However, it is susceptible to physical damage and can be easily scratched, especially since it is usually deposited as a thin film (0.5–5 μm). PDMS has better mechanical resistance, but ensuring its adhesion to many typical semiconductor materials requires attention [88, 89].

5.2 Packaging in the IMPACT Project

In the IMPACT project, several potential materials were selected and investigated. The materials are needed to be thoroughly characterised according to the parameters above, as well as compatibility with the selected implantation process and the wireless power and data transfer.

In order to assess the above parameters, test structures were designed to specifically assess physical durability, insulation, adhesion and the effectiveness and cleanliness of the patterning process. This methodology enables these areas to be benchmarked and optimised if required. The full assessment can be found here [90].

The process for patterning the epoxy consisted of selective exposure to UV light through a photomask. The areas exposed to UV were cured, which left the masked areas uncured and could be removed using acetone. A Clark oxygen sensor packaged in resin is shown in Fig. 11.9a. The central area consists of a 2 mm square window in the resin, exposing the sensing area.

The insulation of the resin was assessed by applying a voltage to an aluminium electrode, coated in the resin immersed in solution. Any current measured, not attributable to capacitive charging or electrical noise, was most likely due to ions reaching the electrode surface, indicating the resin had been compromised. An exemplar measurement is presented in Fig. 11.9b, and suggests there is no leakage of ions through the resin.

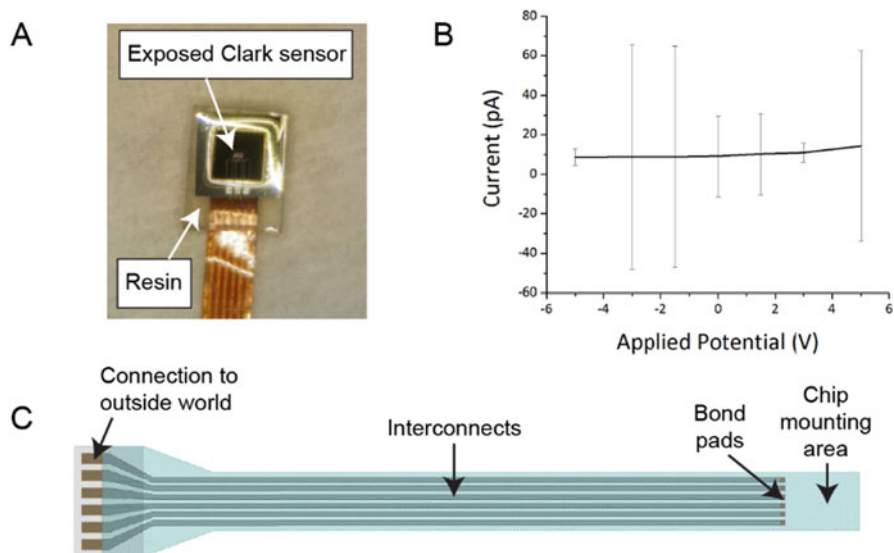


Fig. 11.9 Development of IMPACT sensor packaging. (a) Photograph of a 3 mm square Clark sensor packaged in resin with a 2 mm square window in the centre. (b) Leakage current measured through the resin, plotted against applied potential. (c) Schematic of the designed flexible PCB

5.3 *Chip Deployment*

For deploying the chip, it is important that the package is physically compatible with whatever means of implantation is employed. Usually this requires the package to be as small as possible while maintaining physical durability. Developing small packaging is also important for minimising invasiveness to the patient or animal. On the other hand, the small size of the chip may present a challenge for the personnel who are performing the procedure. This therefore requires close collaboration with the medical team, to ensure the factors are appropriately designed for.

In the initial animal model trials, a wired connection will be employed to the sensors. To achieve this, a flexible PCB strip was designed and is presented in Fig. 11.9c. This enables verification of sensor functionality in isolation from the wireless power and data transfer systems.

Once those systems are in place, the package material must also not interfere with or significantly attenuate the wireless signal. For this, further test structures will be developed in collaboration with those developing the wireless system, to ensure full compatibility with the packaging.

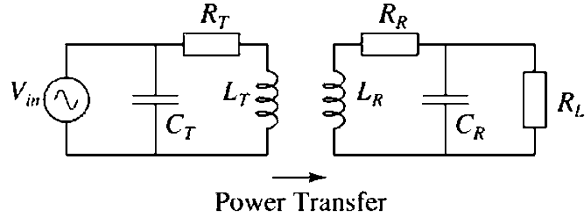
5.4 *Summary*

Packaging of biosensors requires significant attention to ensure the device does not fail. This takes the form of assessing a number of factors. In the IMPACT project, a methodical approach has been taken where each of these factors is characterised using specific test structures. Currently Epo-Tek ETOG116-31/1LB resin has demonstrated a good balance of properties; it can be patterned, is physically durable and appears to insulate effectively. However, adhesion to the chip surface needs to be improved [90].

6 **Wireless Operation: Challenges and Options**

In this section, we will discuss the application of wireless power transfer and communication system for implanted sensors. A magnetic resonance wireless power transfer (MRWPT) system is the primary choice for power and sensor activation. Bluetooth and frequency-shift keying (FSK) are options for transmitting data from the implanted sensors. Overall, the power and communication system must be robust and efficient and have minimal impact on the patient's daily life.

Fig. 11.10 Circuit schematic of a resonant power transfer system



6.1 Magnetic Resonance Wireless Power Transfer

6.1.1 Introduction to Magnetic Resonance Wireless Power Transfer

The concept of inductive power transfer is illustrated in Fig. 11.10. An AC signal applied to a transmitter coil (inductor L_T) generates a magnetic field. This can couple with another coil (inductor L_R) and transfer power to the receiver circuit. Power transfer is enhanced when the transmitter and receiver are both tuned LC circuits with the same resonant frequency. This is a non-radiative, near-field effect, and the maximum range is few times the diameter of the transmitter coil.

6.1.2 Wireless Power Transfer System

A wireless energy transmission system can be divided into three parts: a primary transmitter, an optional secondary power relay and a receiver (Fig. 11.11). Generally, the power transmitter's function is to create a magnetic field that intersects with the receiver. It can also provide automatic adjustment on resonance frequency. The primary function of power receiver block is to receive wireless power. It may also be able to send feedback information to the transmitter. The function of the optional power relay is to couple the energy from the transmitter to extend the transmission range of the system.

6.1.3 Challenges

(a) Exposure limit

The World Health Organization (WHO) allows two standards for electric field and magnetic field exposure limits. They are the International Commission on Non-Ionizing Radiation Protection (ICNIRP) guidelines [91, 92] (0.16 A/m H-field strength at 10 MHz) and IEEE Standard for Safety Levels with Respect to Human Exposure to RF-EM fields [93] (10 MHz is 1.63 A/m). These standards and human tissue characteristics restrict the maximum magnetic and electric field strengths which the implants can be exposed to.

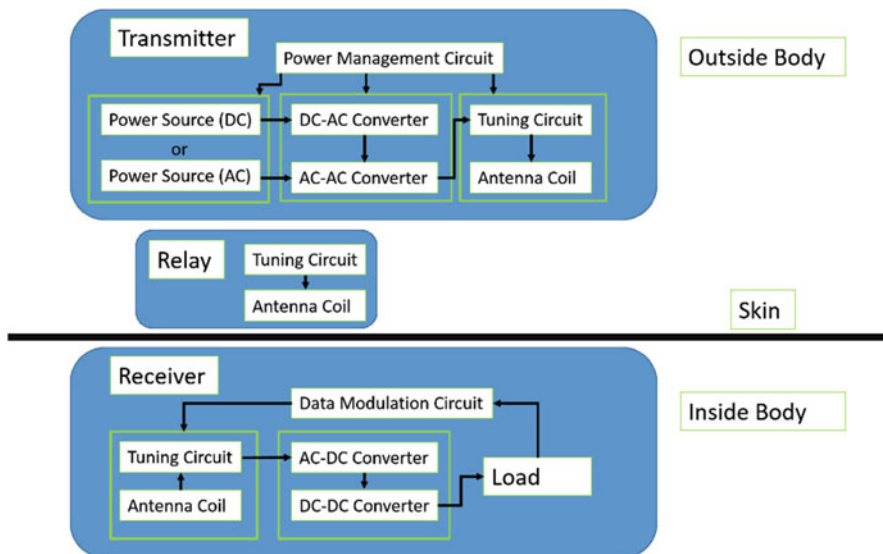


Fig. 11.11 Schematic of a magnetic resonance wireless power transfer system

(b) Implanted receiver size, coupling coefficient and operating frequency

The nature of implants means that the size of the receiver is limited. This limitation affects the magnetic coupling between transmitters and receivers. According to coupled mode theory, maximum coupling efficiency is achieved when the resonating transmitter and receiver are identical and the separation is constant [94]. This cannot be achieved for micro-implants, as the size of transmitter is, in most cases, much larger than the size of receivers. One possible solution is the application of relay coils, which can help to increase the coupling coefficient [95].

(c) Obtainable power, transfer efficiency and transfer distance

Because of the difference in sizes of transmitter and receiver coils and the low-quality factor of the receiver coil, the obtainable power for magnetic resonance transfer is generally in the order of a few mW. In a near-field system, the power density is inversely proportional to the square of the transmission distance, meaning that the transfer efficiency will fall quickly when the distance increases. Therefore, the transfer distance is typically less than 10 cm when the transmitter power is within exposure limits.

6.2 Options for Wireless Communication with Implants

6.2.1 General Considerations

Various options are available for wireless communication between implants and the outside world. The main differences between such devices are mostly in the transceivers on the implants. Factors to be considered include power consumption, size, operating frequency, maximum signal transmission range and data rate. Two options for communication with an implanted sensor system are discussed here: Bluetooth and frequency-shift keying (FSK).

6.2.2 Bluetooth

Bluetooth low energy (BLE) is a mature technology that provides an ideal solution for implant-reader communication. A BLE device can consume mW level power (Table 11.1) while maintaining a signal transmission range of 2–5 metres [96]. BLE devices can be as small as mm scale, which is small enough to be implanted inside the human body. A maximum data rate for BLE is around 10 kb/s, which is sufficient for most implants. The main problem of BLE is its operating frequency of 2.4 GHz, at which attenuation of EM radiation is relatively high [97].

6.2.3 Frequency-Shift Keying Modulation

Frequency-shift keying (FSK) is a digital modulation technique. In binary FSK, the frequency of a carrier signal is shifted between two discrete values representing digital “0” and digital “1” [98]. The scheme is simple, and only a few extra components are needed, which means only a limited space is needed for an FSK scheme in implants. Also, no extra power is needed for transmitting data as the information is carried by the implant resonant frequency, which can be detected by external readers.

Table 11.1 Bluetooth power class

Power class	Max output power (mW)	Range (m)
Class 1	100	> 100
Class 2	2.5	10
Class 3	1	1

6.3 Conclusion

Although magnetic resonance wireless power transfer remains the primary choice for WPT so far, multiple challenges such as exposure limits, transfer efficiency and maximum obtainable power set the upper limit for its application in powering implanted devices. For communication, both Bluetooth and FSK have high potential in this field because of their low power consumption and limited space requirements.

7 Summary

This chapter has described the challenges, successes and frustrations in developing a wireless sensor system to improve the treatment of solid cancers in humans. We have described the scientific and technical challenges posed by implanted medical applications and some potential approaches to addressing them. We have not, however, explored the other, often competing challenges posed by medical regulation frameworks that are often trying hard to keep pace with technology and often failing to do so. Part of the IMPACT project is a small, crucial work package in the area of “value systems” – meaning the series of implications that accrue during IMPACT’s scientific and engineering development that may affect its ability to be translated into clinical use. These may be negative (i.e. choosing materials that are not biocompatible) or positive (i.e. choosing power/communication frequencies that do not pose a health hazard). If we were to leave a final message for this chapter, alongside the many technical challenges and solutions that we have described, it would be that engineers should work more closely, at project level and from the outset, with experts in non-scientific disciplines such as medical regulation, ethics and relevant social sciences. If we do, we will (a) become better engineers and (b) get our ideas into the marketplace faster.

Acknowledgements This work was supported by the funding from the UK Engineering and Physical Sciences Research Council, through the Implantable Microsystems for Personalised Anti-Cancer Therapy (IMPACT) programme grant (EP/K034510/1).

References

1. IARC, *Globocan 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012* (WHO, Geneva, Switzerland, 2012)
2. M. Baumann, M. Krause, J. Overgaard, J. Debus, S.M. Bentzen, J. Daartz, et al., Radiation oncology in the era of precision medicine. *Nat. Rev. Cancer* **16**, 234–249 (2016)
3. J.C. Mottram, A factor of importance in the radio sensitivity of tumours. *Br. J. Radiol.* **9**, 606–614 (1936)
4. L.H. Gray, A.D. Conger, M. Ebert, S. Hornsey, O.C.A. Scott, The concentration of oxygen dissolved in tissues at the time of irradiation as a factor in radiotherapy. *Br. J. Radiol.* **26**, 638–648 (1953)

5. J.P. Ward, Oxygen sensors in context. *Biochim. Biophys. Acta* **1777**, 1–14 (2008)
6. J.A. Bertout, S.A. Patel, M.C. Simon, The impact of O₂ availability on human cancer. *Nat. Rev. Cancer* **8**, 967–975 (2008)
7. C. Ward, S.P. Langdon, P. Mullen, A.L. Harris, D.J. Harrison, C.T. Supuran, et al., New strategies for targeting the hypoxic tumour microenvironment in breast cancer. *Cancer Treat. Rev.* **39**, 171–179 (2013)
8. C. Bayer, K. Shi, S.T. Astner, C.A. Maftei, P. Vaupel, Acute versus chronic hypoxia: why a simplified classification is simply not enough. *Int. J. Radiat. Oncol. Biol. Phys.* **80**, 965–968 (2011)
9. D.J. Chaplin, P.L. Olive, R.E. Durand, Intermittent blood flow in a murine tumor: radiobiological effects. *Cancer Res.* **47**, 597–601 (1987)
10. P. Vaupel, F. Kallinowski, P. Okunieff, Blood flow, oxygen and nutrient supply, and metabolic microenvironment of human tumors: a review. *Cancer Res.* **49**, 6449–6465 (1989)
11. J. Overgaard, Hypoxic modification of radiotherapy in squamous cell carcinoma of the head and neck – a systematic review and meta-analysis. *Radiother. Oncol.* **100**, 22–32 (2011)
12. A.L. Harris, Hypoxia – a key regulatory factor in tumour growth. *Nat. Rev. Cancer* **2**, 38–47 (2002)
13. G.L. Semenza, Targeting HIF-1 for cancer therapy. *Nat. Rev. Cancer* **3**, 721–732 (2003)
14. J.P. Dales, S. Garcia, S. Meunier-Carpentier, L. Andrac-Meyer, O. Haddad, M.N. Lavaut, et al., Overexpression of hypoxia-inducible factor HIF-1 α predicts early relapse in breast cancer: retrospective study in a series of 745 patients. *Int. J. Cancer* **116**, 734–739 (2005)
15. G.L. Semenza, Regulation of cancer cell metabolism by hypoxia-inducible factor 1. *Semin. Cancer Biol.* **19**, 12–16 (2009)
16. M.S. Ullah, A.J. Davies, A.P. Halestrap, The plasma membrane lactate transporter MCT4, but not MCT1, is up-regulated by hypoxia through a HIF-1 α -dependent mechanism. *J. Biol. Chem.* **281**, 9030–9037 (2006)
17. B.A. Webb, M. Chimenti, M.P. Jacobson, D.L. Barber, Dysregulated pH: a perfect storm for cancer progression. *Nat. Rev. Cancer* **11**, 671–677 (2011)
18. R.A. Gatenby, K. Smallbone, P.K. Maini, F. Rose, J. Averill, R.B. Nagle, et al., Cellular adaptations to hypoxia and acidosis during somatic evolution of breast cancer. *Br. J. Cancer* **97**, 646–653 (2007)
19. R.J. Gillies, Z. Liu, Z. Bhujwala, 31P-MRS measurements of extracellular pH of tumors using 3-aminopropylphosphonate. *Am. J. Physiol.* **267**, C195–C203 (1994)
20. K. Goetze, S. Walenta, M. Ksiazkiewicz, L.A. Kunz-Schughart, W. Mueller-Klieser, Lactate enhances motility of tumor cells and inhibits monocyte migration and cytokine release. *Int. J. Oncol.* **39**, 453–463 (2011)
21. F. Vegran, R. Boidot, C. Michiels, P. Sonveaux, O. Feron, Lactate influx through the endothelial cell monocarboxylate transporter MCT1 supports an NF- κ B/IL-8 pathway that drives tumor angiogenesis. *Cancer Res.* **71**, 2550–2560 (2011)
22. U.G. Sattler, S.S. Meyer, V. Quennet, C. Hoerner, H. Knoerzer, C. Fabian, et al., Glycolytic metabolism and tumour response to fractionated irradiation. *Radiother. Oncol.* **94**, 102–109 (2010)
23. D.M. Brizel, T. Schroeder, R.L. Scher, S. Walenta, R.W. Clough, M.W. Dewhirst, et al., Elevated tumor lactate concentrations predict for an increased risk of metastases in head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **51**, 349–353 (2001)
24. S. Walenta, M. Wetterling, M. Lehrke, G. Schwickert, K. Sundfor, E.K. Rofstad, et al., High lactate levels predict likelihood of metastases, tumor recurrence, and restricted patient survival in human cervical cancers. *Cancer Res.* **60**, 916–921 (2000)
25. C. Groussard, I. Morel, M. Chevanne, M. Monnier, J. Cillard, A. Delamarque, Free radical scavenging and antioxidant effects of lactate ion: an in vitro study. *J. Appl. Physiol.* **89**(2000), 169–175 (1985)
26. M.R. Horsman, L.S. Mortensen, J.B. Petersen, M. Busk, J. Overgaard, Imaging hypoxia to improve radiotherapy outcome. *Nat. Rev. Clin. Oncol.* **9**, 674–687 (2012)

27. M. Nordsmark, J. Alsner, M. Busk, J. Overgaard, M.R. Horsman, Hypoxia and radiation therapy, in *Hypoxia and Cancer: Biological Implications and Therapeutic Opportunities*, ed. by G. Melillo (Springer, New York, 2014), pp. 265–281.
28. L.C. Clark, Electrochemical device for chemical analysis, U.S. Patent Office, 1959
29. W.P. Chan, M. Narducci, Y. Gao, M.-Y. Cheng, J.H. Cheong, A.K. George, et al., A monolithically integrated pressure/oxygen/temperature sensing SoC for multimodality intracranial neuromonitoring. *IEEE J. Solid-State Circuits* **49**, 2449–2461 (2014)
30. P. Wang, Y. Liu, H.D. Abruña, J.A. Spector, W.L. Olbricht, Micromachined dissolved oxygen sensor based on solid polymer electrolyte. *Sens. Actuators B* **153**, 145–151 (2011)
31. C.-C. Wu, T. Yasukawa, H. Shiku, T. Matsue, Fabrication of miniature Clark oxygen sensor integrated with microstructure. *Sens. Actuators B* **110**, 342–349 (2005)
32. A.M. Otto, M. Brischwein, E. Motrescu, B. Wolf, Analysis of drug action on tumor cell metabolism using electronic sensor chips. *Arch. Pharm.* **337**, 682–686 (2004)
33. M. Brischwein, E.R. Motrescu, E. Cabala, A.M. Otto, H. Grothe, B. Wolf, Functional cellular assays with multiparametric silicon sensor chips. *Lab Chip* **3**, 234–240 (2003)
34. R. Pethig, S. Smith, Electrochemical Principles and Electrode Reactions, in *Introductory Bioelectronics: For Engineers and Physical Scientists*, 1st edn., (Wiley, Chichester, 2012), pp. 177–213
35. M.W. Shinwari, D. Zhitomirsky, I.A. Deen, P.R. Selvaganapathy, M.J. Deen, D. Landheer, Microfabricated reference electrodes and their biosensing applications. *Sensors* **10**, 1679–1715 (2010)
36. C. Duarte-Guevara, V.V. Swaminathan, M. Burgess, B. Reddy, E. Salm, Y.-S. Liu, et al., On-chip metal/polypyrrole quasi-reference electrodes for robust ISFET operation. *Analyst* **140**, 3630–3641 (2015)
37. K.K. Kasem, S. Jones, Platinum as a reference electrode in electrochemical measurements. *Platinum Metals Review* **52**, 100–106 (2008)
38. H. Suzuki, T. Hirakawa, S. Sasaki, I. Karube, An integrated module for sensing pO₂, pCO₂, and pH. *Anal. Chim. Acta* **405**, 57–65 (2000)
39. H. Suzuki, A. Hiratsuka, S. Sasaki, I. Karube, Problems associated with the thin-film Ag/AgCl reference electrode and a novel structure with improved durability. *Sens. Actuators B* **46**, 104–113 (1998)
40. P. Hashemi, P.L. Walsh, T.S. Guillot, J. Gras-Najjar, P. Takmakov, F.T. Crews, et al., Chronically implanted, nafion-coated Ag/AgCl reference electrodes for neurochemical applications. *ACS Chem. Neurosci.* **2**, 658–666 (2011)
41. K.A. Mauritz, R.B. Moore, State of understanding of nafion. *Chem. Rev.* **104**, 4535–4586 (2004)
42. J.R.K. Marland, C. Dunare, A. Tsiamis, E. González-Fernández, E. Blair, S. Smith, et al., Test structures for optimizing polymer electrolyte performance in a microfabricated electrochemical oxygen sensor, in *International Conference on Microelectronic Test Structures*, (Grenoble, 2017), pp. 145–149
43. R. Pethig, S. Smith, Basic Sensor Instrumentation and Electrochemical Sensor Interfaces, in *Introductory Bioelectronics: For Engineers and Physical Scientists*, 1st edn., (Wiley, Chichester, 2012), pp. 259–296
44. P. Bergveld, Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Trans. Biomed. Eng.* **17**, 70–71 (1970)
45. K. Nakazato, M. Ohura, S. Uno, Source-drain follower for monolithically integrated sensor array. *Electronics Letters* **43**, 1255–1257 (2007)
46. K. Nakazato, M. Ohura, S. Uno, CMOS cascode source-drain follower for monolithically integrated biosensor array. *IEICE Trans. Electron.* **E91c**, 1505–1515 (2008)
47. D.R. Thévenot, K. Toth, R.A. Durst, G.S. Wilson, Electrochemical biosensors: recommended definitions and classification. *Biosens. Bioelectron.* **16**, 121–131 (2001)
48. J. Wang, Electrochemical glucose biosensors. *Chem. Rev.* **108**, 814–825 (2008)
49. S. Azzouzi, L. Rotariu, A.M. Benito, W.K. Maser, M. Ben Ali, C. Bala, A novel amperometric biosensor based on gold nanoparticles anchored on reduced graphene oxide for sensitive detection of l-lactate tumor biomarker. *Biosens. Bioelectron.* **69**, 280–286 (2015)

50. K. Rathee, V. Dhull, R. Dhull, S. Singh, Biosensors based on electrochemical lactate detection: a comprehensive review. *Biochemistry and Biophysics Reports* **5**, 35–54 (2016)
51. M.P. Massafra, S.I.C. de Torresi, Urea amperometric biosensors based on a multifunctional bipolymeric layer: comparing enzyme immobilization methods. *Sens. Actuators B Chem.* **137**, 476–482 (2009)
52. A. Makaraviciute, A. Ramanaviciene, Site-directed antibody immobilization techniques for immunosensors. *Biosens. Bioelectron.* **50**, 460–471 (2013)
53. V. Seraffín, L. Agüí, P. Yáñez-Sedeño, J.M. Pingarrón, Electrochemical immunosensor for the determination of insulin-like growth factor-1 using electrodes modified with carbon nanotubes-poly(pyrrole propionic acid) hybrids. *Biosens. Bioelectron.* **52**, 98–104 (2014)
54. G. Lai, H. Zhang, T. Tamanna, A. Yu, Ultrasensitive immunoassay based on electrochemical measurement of enzymatically produced polyaniline. *Anal. Chem.* **86**, 1789–1793 (2014)
55. P.E. Nielsen, Peptide nucleic acids (PNA) in chemical biology and drug discovery. *Chem. Biodivers.* **7**, 786–804 (2010)
56. A.A. Lubin, K.W. Plaxco, Folding-based electrochemical biosensors: the case for responsive nucleic acid architectures. *Acc. Chem. Res.* **43**, 496–505 (2010)
57. A.D. Ellington, J.W. Szostak, In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990)
58. C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990)
59. E. González-Fernández, N. Avlonitis, A.F. Murray, A.R. Mount, M. Bradley, Methylene blue not ferrocene: optimal reporters for electrochemical detection of protease activity. *Biosens. Bioelectron.* **84**, 82–88 (2016)
60. A. Bard, L. Faulkner, *Electrochemical Methods: Fundamentals and Applications* (Wiley, New York, 2001)
61. F. Lisdat, D. Schäfer, The use of electrochemical impedance spectroscopy for biosensing. *Anal. Bioanal. Chem.* **391**, 1555–1567 (2008)
62. E.P. Randviir, C.E. Banks, Electrochemical impedance spectroscopy: an overview of bioanalytical applications. *Anal. Methods* **5**, 1098–1115 (2013)
63. I.S. Kucherenko, D.Y. Kucherenko, O.O. Soldatkin, F. Lagarde, S.V. Dzyadevych, A.P. Soldatkin, A novel conductometric biosensor based on hexokinase for determination of adenosine triphosphate. *Talanta* **150**, 469–475 (2016)
64. T.T. Nguyen-Boisse, J. Saulnier, N. Jaffrezic-Renault, F. Lagarde, Highly sensitive conductometric biosensors for total lactate, D- and L-lactate determination in dairy products. *Sens. Actuators B Chem.* **179**, 232–239 (2013)
65. C. López-Otín, L.M. Matrisian, Emerging roles of proteases in tumour suppression. *Nat. Rev. Cancer* **7**, 800–808 (2007)
66. Q. Liu, J. Wang, B.J. Boyd, Peptide-based biosensors. *Talanta* **136C**, 114–127 (2015)
67. J. Adjémian, A. Anne, G. Cauet, C. Demaille, Cleavage-sensing redox peptide monolayers for the rapid measurement of the proteolytic activity of trypsin and α -thrombin enzymes. *Langmuir* **26**, 10347–10356 (2010)
68. P.S. Singh, From sensors to systems: CMOS-integrated electrochemical biosensors. *IEEE Access* **3**, 249–259 (2015)
69. E.D. Minot, A.M. Janssens, I. Heller, H.A. Heering, C. Dekker, S.G. Lemay, Carbon nanotube biosensors: the critical role of the reference electrode. *Appl. Phys. Lett.* **91**, 093507 (2007)
70. O.Y.F. Henry, J.L.A. Sanchez, C.K. O’Sullivan, Bipodal PEGylated alkanethiol for the enhanced electrochemical detection of genetic markers involved in breast cancer. *Biosens. Bioelectron.* **26**, 1500–1506 (2010)
71. T. Bertok, E. Dosekova, S. Belicky, A. Holazova, L. Lorencova, D. Mislovicova, et al., Mixed zwitterion-based self-assembled monolayer interface for impedimetric glycomic analyses of human IgG samples in an array format. *Langmuir* **32**, 7070–7078 (2016)
72. D. Lu, C.P. Wong, *Materials for Advanced Packaging* (Springer, Boston, 2009)
73. S.F. Al-Sarawi, D. Abbott, P.D. Franzone, A review of 3-D packaging technology. *IEEE Trans. Compon. Packag. Manuf. Technol. Part B* **21**, 2–14 (1998)

74. W.H. Ko, Packaging of microfabricated devices and systems. *Mater. Chem. Phys.* **42**, 169–175 (1995)
75. C.P. Wong, K.-S. Moon, Y. Li, *Nano-bio-electronic, Photonic and MEMS Packaging*, 1st edn. (Springer, New York, 2010)
76. N. Abramova, A. Bratov, Photocurable polymers for ion selective field effect transistors. 20 years of applications. *Sensors* **9**, 7097–7110 (2009)
77. C. Cotofana, A. Bossche, P. Kaldenberg, J. Mollinger, Low-cost plastic sensor packaging using the open-window package concept. *Sensors and Actuators A: Physical* **67**, 185–190 (1998)
78. D.J. Apple, N. Mamalis, S.E. Brady, K. Loftfield, D. Kavka-Van Norman, R.J. Olson, Biocompatibility of implant materials: a review and scanning electron microscopic study. *American Intra-Ocular Implant Soc. J.* **10**, 53–66 (1984)
79. M. Frost, M.E. Meyerhoff, In vivo chemical sensors: tackling biocompatibility. *Anal. Chem.* **78**, 7370–7377 (2006)
80. Y. Onuki, U. Bhardwaj, F. Papadimitrakopoulos, D.J. Burgess, A review of the biocompatibility of implantable devices: current challenges to overcome foreign body response. *J. Diabetes Sci. Technol.* **2**, 1003–1015 (2008)
81. S. Kirsten, M. Schubert, M. Braunschweig, G. Woldt, T. Voitsekhivska, K.-J. Wolter, Biocompatible packaging for implantable miniaturized pressure sensor device used for stent grafts: Concept and choice of materials, in *Electronics Packaging Technology Conference (EPTC), 2014 IEEE 16th*, (Singapore, 2014), pp. 719–724
82. ISO, ISO 10993-1:2009 Biological evaluation of medical devices – part 1: evaluation and testing within a risk management process, (2009)
83. Y. Qin, M.M.R. Howlader, M.J. Deen, Y.M. Haddara, P.R. Selvaganapathy, Polymer integration for packaging of implantable sensors. *Sens. Actuators B* **202**, 758–778 (2014)
84. M. Leineweber, G. Pelz, M. Schmidt, H. Kappert, G. Zimmer, New tactile sensor chip with silicone rubber cover. *Sensors and Actuators A: Physical* **84**, 236–245 (2000)
85. T. Datta-Chaudhuri, P. Abshire, E. Smela, Packaging commercial CMOS chips for lab on a chip integration. *Lab Chip* **14**, 1753 (2014)
86. J.B. Fortin, T.-M. Lu, *Chemical Vapor Deposition Polymerization: The Growth and Properties of Parylene Thin Films* (Springer, Boston, 2004)
87. E.M. Schmidt, J.S. McIntosh, M.J. Bak, Long-term implants of Parylene-C coated microelectrodes. *Med. Biol. Eng. Comput.* **26**, 96–101 (1988)
88. M.A. Eddings, M.A. Johnson, B.K. Gale, Determining the optimal PDMS–PDMS bonding technique for microfluidic devices. *J. Micromech. Microeng.* **18**, 067001 (2008)
89. J.C. Lötters, W. Olthuis, P.H. Veltink, P. Bergveld, The mechanical properties of the rubber elastic polymer polydimethylsiloxane for sensor applications. *J. Micromech. Microeng.* **7**, 145–147 (1997)
90. E.O. Blair, A. Buchoux, A. Tsiamis, C. Dunare, J.R.K. Marland, J.G. Terry, et al., Test structures for the characterisation of sensor packaging technology, in *2017 International Conference on Microelectronic Test Structures*, (Bordeaux, 2017)
91. International Commission on Non-Ionizing Radiation Protection, Guidelines for limiting exposure to time-varying electric and magnetic fields (1 Hz to 100 kHz). *Health Phys.* **99**, 818–836 (2010)
92. International Commission on Non-Ionizing Radiation Protection, Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz). International Commission on Non-Ionizing Radiation Protection. *Health Phys.* **74**, 494–522 (1998)
93. IEEE, C95.1-2005 – IEEE standard for safety levels with respect to human exposure to radio frequency electromagnetic fields, 3 kHz to 300 GHz, (New York, 2006)
94. P.T. Theilmann, *Wireless power transfer for scaled electronic biomedical implants: UC San Diego*, (2012)
95. F. Zhang, S.A. Hackworth, W. Fu, C. Li, Z. Mao, M. Sun, Relay effect of wireless power transfer using strongly coupled magnetic resonances. *IEEE Trans. Magn.* **47**, 1478–1481 (2011)

96. K. Townsend, R. Davidson, Akiba, C. Cufí, *Getting started with Bluetooth low energy: tools and techniques for low-power networking*, Revised 1st edn. (O'Reilly, Sebastopol, 2014)
97. I. Dove, *Analysis of Radio Propagation Inside the Human Body for in-Body Localization Purposes*, University of Twente, (2014)
98. G. Kennedy, *Electronic Communication Systems*, McGraw-Hill Inc., US 1992, p. 509

Chapter 12

Compressed Sensing for High Density Neural Recording

Jie Zhang, Tao Xiong, Srinjoy Mitra, and Ralph Etienne-Cummings

1 Introduction

1.1 *Practical Limitations of High Density Neural Recording Devices*

High density neural recording devices have become a very useful tool for neuro-physiologists to study electrical activities of the brain. These implanted devices have to be small in weight and volume, and yet contain many recording sites or electrodes. When deployed in the brain, they can record electrical signals generated by individual neurons (Action Potential or ‘spikes’) or a group of neurons (multi-unit activity and local field potentials) modern system neuroscience rely on the ability to record from a large number of spikes from single or multiple region of the brain, and to attribute them to individual neurons (spike sorting). Studying the neurons’ activities allows neuroscientists to discover the function and connectivity of brain and their role in cognition and behavior [1, 2]. Clinically, electrophysiology recordings can also be studied to diagnose neuropsychological illnesses such as depression, epilepsy, and traumatic brain injuries [3, 4]. Modern electrophysiology recording microsystems have evolved from single electrode in the 1950s to multi-

J. Zhang (✉)

Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA
e-mail: jzhang41@mit.edu

T. Xiong • R. Etienne-Cummings

Johns Hopkins University, 3400 Charles St., Baltimore, MD 21218, USA
e-mail: txiong1@jhu.edu; retinne@jhu.edu

S. Mitra

School of Engineering, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK
e-mail: srinjoy.mitra@glasgow.ac.uk

electrode array in the 1990s [5, 6]. Due to the advancement of silicon microchips and the corresponding lithographic technology, integration of very high density electrode on CMOS probes is now possible [7, 8]. The trend of high density electrode integration is expected to continue as researchers are currently designing silicon probes that contain > 1000 channels. Possibility of distributed network of extremely small (sub-100 μm) recording devices that can float in the brain is also being explored [9, 10].

Unfortunately, while the recording devices' size shrinks and electrode density increases, their energy capacity and requirement do not follow the same aggressive trend. As a result, power consumption of these devices becomes a bottleneck that prevents further integration of even higher density. In previous analysis, a system power breakdown has shown that data transmission accounts for the majority power consumption of the neural recording system [11]. To acquire the spikes, each electrode must sample at Nyquist rate of 30 kHz with at least 10 bits of resolution as required by post-processing. The data collected by the high density electrode array exceeds 300 Megabits-per-second (Mbps) as the number of integrated electrodes increases above 1000. The enormous amount of data poses significant challenges for the design of digital data readout interfaces. This is mostly due to the available power budget that can be dissipated close to the brain that does not result in a temperature rise that exceeds the safe limit of 1 °C. The challenge is even greater when neural probes with wireless transmission are considered. The limited weight of head mounted devices that can be carried by small laboratory animals like mice and rats (10% of body weight) puts a severe restriction on the battery life of these devices. A summary of the power consumption requirement, both for a wired and a wires system, is presented in the next two sections to demonstrate this issue.

In such a situation data compression is highly desirable. However, due to the same limitations mentioned above, the compression has to be power efficient and should retain the neurophysiology information. In this chapter we focus on a specific data compressing method, Compressed Sensing, that has shown a great promise in this respect.

1.1.1 Power Consumption in Wired Transmission

Figure 12.1a shows a typical setup for recording experiment using high density silicon probe [7]. The neural probe consists of recording contacts and Application Specific Integrated Circuit (ASIC) fabricated on the same silicon substrate. The recording contacts are inserted into the brain while the ASIC is placed outside of the skull. The recording contacts, equipped with active electrodes, amplifies the signals before sending them to the ASIC for additional filtering and digitization. The ASIC then transmits the digitized signal through a short PCB trace (5–10 cm) to the headstage, which then drives a long cable (> 10 m) to reach the computer.

In the wired recording setup, the power is typically delivered to the electronics through additional wires. Thus battery capacity and weight are not the major constraint in this setup. However, the power dissipation of the head stage and the ASIC still needs to be low to avoid heat generation which could increase the

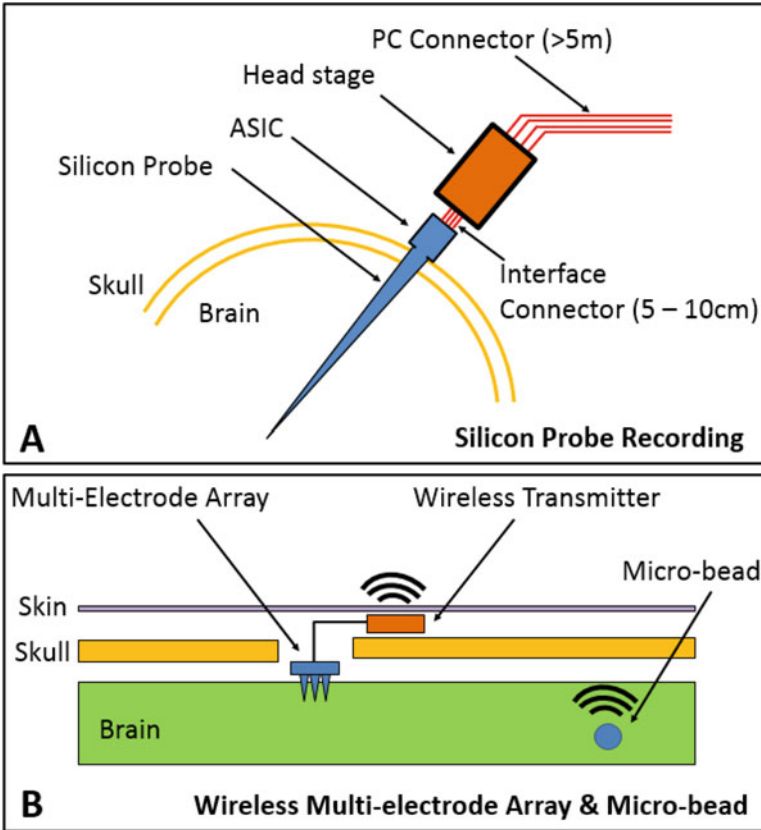


Fig. 12.1 (a) Wired silicon probe recording experiment setup. (b) Wireless multi-electrode array and miniature recording device setup

temperature of the brain to cause tissue damage. The heat dissipation at the ASIC also has to be kept low especially due to its proximity to the brain tissue.

The dynamic power consumption at the output of the ASIC can be calculated as:

$$P_{\text{dynamic}} = \rho_t (C_L \cdot V_{dd}^2 \cdot f) \quad (12.1)$$

where C_L is the load capacitance, f is the data rate or clock frequency, and ρ_t is the probability where a signal transition occurs at the output [12]. We take $\rho_t = 0.5$ and $C_L = 30$ pF. The load capacitance includes wire bonding capacitance, connector interface, and Printed Circuit Board (PCB) trace (1 pf/cm) to the head stage. Each electrode has to sample the signal at 30 KHz with 10 bits of resolution. Therefore a 1000 electrode silicon probe generates data at rate of 300 Mbps. Using standard digital output drivers with 3.3 V power rail, the dynamic power consumption is calculated to be around 50 mW. As a comparison we also estimated power consumption for the rest of the circuit. One electrode readout circuit typically

consumes around $20 \mu\text{W}$ [7]. A linear extrapolation indicates that 1000 recording channels would consume around 20 mW. Therefore, the power consumed in wired transmission at high data rate accounts for 62.5% of the entire system power consumption.

Power consumption on the order of 50mW may also induce significant amount of heat dissipation. Since silicon has high thermal conductivity, the heat may travel along the probe and rise the temperature of the brain tissue. The exact determination of the heat transfer is still an active area of study. Previous study has shown that 13 mW of heat dissipation by the ASIC in proximity of the electrode array induces around 1.3°C [13] temperature increase in the tissue, which is higher than the maximum permissible temperature increase in the cortex of 1°C .

As shown in Eq. (12.1), the dynamic power consumption is directly proportional to the data rate. The key to power reduction would be to efficiently compress the data prior to transmission. In this example, it can be shown that with 10x compression rate, the dynamic power consumption can decrease to 5mW from 50mW. As a result, the overall system power consumption can decrease from 80 to 25mW by reducing the transmission data rate.

1.1.2 Power Consumption in Wireless Transmission

Figure 12.1b shows a recording setup using wireless multi-electrode array (MEA) and distributed miniature brain implants. Wireless communication method is commonly used for neuroscience experiment where the animal must perform free movement tasks [2, 14, 15]. They are also commonly used for human experiments such as prosthetics and Brain Computer Interface (BCI). The miniature brain implants is a new generation of device under research [9, 10]. They are small enough to float within the deep brain to record neural activity and transmit the data wirelessly to the receiver outside of the brain. These devices are either powered through wireless power transfer or through a small battery carried by the animal. Hence, the wireless communication protocol must be power efficient to increase battery life, minimize RF radiation injected into the tissue, and avoid heating the brain tissue.

Unlike wired power consumption, the wireless transmission power estimation requires extra considerations: modulation method, channel SNR, transmission distance, data rate, and link budget. If the wireless radio is within the tissue or under the skull, we must also consider Specific Absorption Rate (SAR) and Equivalent Isotropic Radiated Power (EIRP). We provided a detailed analysis in [16]. To summarize our results, let's consider a $10\times$ data reduction that compresses 100 Mbps neural data to 10 Mbps. From the link budget perspective and considering same receiver sensitivity, with $10\times$ data compression at 10 Mbps, the same transmitter power for transmitting 100 Mbps at 2 m of distance can be used to transmit 10 Mbps at 5 m. This result assumes a non-line of sight communication channel in the UWB bandwidth of 7500 MHz and bit error rate of 10^{-6} . From the other perspective, with $10\times$ compression, a 10 dB reduction in transmit power can still achieve the desired transmission distance of <2 m. Furthermore, with reduced power, the SAR margin also improves significantly.

To summarize, we showed there is a significant amount of power saving from both wired and wireless communication perspective by efficiently reducing transmission data rate. In the next subsection, we elaborate on the compressed sensing technique and demonstrate an efficient compression scheme on the recording site as well as an analysis method on the receiver side.

1.2 Compressed Sensing

Compressive sensing originated as a theoretical framework regarding encoding and recovery of an S -sparse signal, \mathbf{x} , of length N [17, 18]. A signal is S -sparse if it can be well approximated by its largest S coefficients in a transform domain (a.k.a. a ‘dictionary’), where $S \ll N$. During encoding, the S -sparse signal, \mathbf{x} , is compressed linearly into a smaller measurement vector, \mathbf{y} , of length M , such that:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \quad (12.2)$$

where $S < M \ll N$, and \mathbf{A} is a sensing matrix of dimension $M \times N$. The compression rate (CR) achieved in this case is N/M . Normally, recovering \mathbf{x} , given \mathbf{y} and \mathbf{A} , is not trivial because this system of linear equations contains more unknown variables than the number of equations. Fortunately, considering matrix \mathbf{A} satisfies the Restricted Isometry Property (RIP) and \mathbf{x} is S -sparse, this underdetermined problem can be solved and \mathbf{x} can be recovered exactly with extremely high probability from \mathbf{y} by minimizing the S -sparse vector’s l_1 -norm through optimization [17]:

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{x} \quad (12.3)$$

RIP is the key factor to determine the optimal choices of sensing matrices. RIP describes how well a S -Sparse signal can be preserved after the projection using sensing matrix \mathbf{A} . From another perspective, RIP characterizes the nearly orthonormal characteristic of the matrix \mathbf{A} with respect to any S -sparse vectors. Mathematically, RIP is expressed as:

$$(1 - \delta_s) \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta_s) \|\mathbf{x}\|_2^2 \quad (12.4)$$

where \mathbf{A} is an $M \times N$ matrix and s is an integer, such that $1 \leq s \leq N$. Suppose there exist a constant $\delta_s \in (0, 1)$ that satisfies the above condition for every $M \times s$ submatrix \mathbf{A}_s of \mathbf{A} and for every S -sparse vector \mathbf{y} . Then, the matrix \mathbf{A} is said to satisfy the s -RIP with restricted isometry constant δ_s . This condition also implies that the S -sparse vectors do not belong to in the null-space of \mathbf{A} . Therefore, it is possible to recover the vector after a projection onto \mathbf{A} .

Many matrices such as the random Gaussian, random Bernoulli, and Partial Fourier matrices all satisfy the RIP universally with a small number of M . Choices of sensing matrix can be determined based on specific applications and desired performance trade-offs.

1.3 Sparse Representation

In practice, it is often difficult to observe a sparse signal \mathbf{x} in its natural form. But most of the time, this signal has a sparse representation \mathbf{v} with respect to a basis (or dictionary) \mathbf{D} such that:

$$\mathbf{x} = \mathbf{D}\mathbf{v} \tag{12.5}$$

where \mathbf{v} is an S -sparse vector representation of \mathbf{x} in the dictionary \mathbf{D} . Figure 12.2 shows some well-known pairs of sparse signals. In the frequency domain, a non-sparse constant function can be represented sparsely in the time domain by an impulse function. Vice versa, a non-sparse sinusoid in the time domain can be represented sparsely in the frequency domain with a pair of impulses at the frequency of oscillation. For a rectangular function, when the width d is large, the function becomes non-sparse in time domain. But as d increases, the function becomes more sparse in the frequency domain.

Some other well-known transforms include the Discrete Wavelet Transform (DWT), Gabor Transform and Discrete Cosine Transform (DCT). DWT and Gabor frames are popular sparsifying transforms for time-frequency sparse signal such as bio-potentials. DCT is often used to sparsify images in image acquisition applications.

The example in Fig. 12.2 also demonstrates another important concept in sparse representation: Mutual coherence. Let Φ be the orthonormal basis of \mathbb{R}^N where

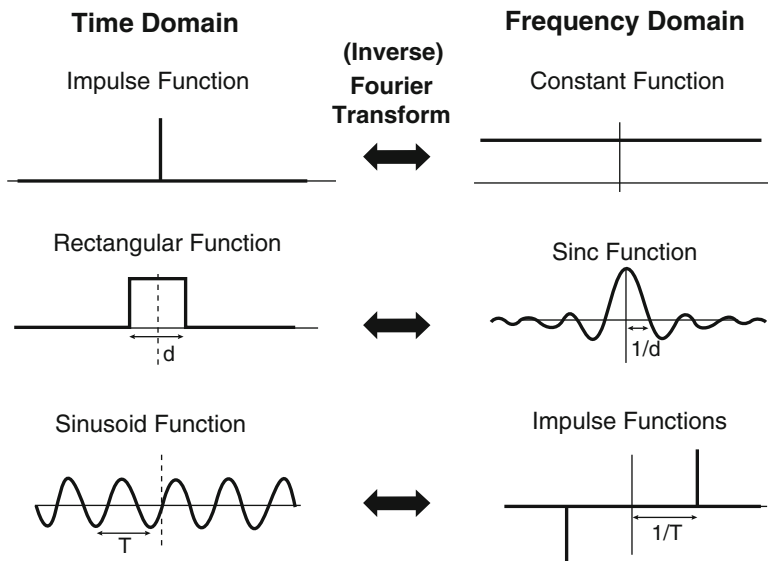


Fig. 12.2 Examples of sparse representations through Fourier and inverse Fourier transform

signal \mathbf{x} is measured, and \mathbf{D} be the orthobases of \mathbb{R}^N where \mathbf{x} has a sparse representation. The mutual coherence between the sensing basis Φ and representation basis \mathbf{D} is defined as:

$$\mu(\Phi, \mathbf{D}) = \sqrt{(N)} \max_{1 \leq k, j \leq N} |\langle \phi_k, d_j \rangle| \quad (12.6)$$

where ϕ_k and d_j are the k th and j th column of the corresponding matrices. The mutual coherence measures the largest correlation between any two columns of Φ and \mathbf{D} . If these two bases contain highly correlated items, the coherence measurement is large.

Mutual coherence is an important concept. It suggests that a signal, densely represented in one base, has a sparse representation in bases that has low mutual coherence with the sparse representation base. The example in Fig. 12.2 shows that time domain and frequency domain are a representation pair with low mutual coherence. Because of this, a signal sparse in time cannot be sparse in frequency domain and vice versa. From a sensing perspective, mutual incoherence means that the information contained in a sparse signal can be spread over a large number of bases when it is transformed into an incoherent transform. Because of this, it is possible to only use a few random samples to capture the information of the sparse signal in the incoherence domain.

1.4 Sampling and Sensing Matrix

Random matrices are common choice for sensing matrix. The two most relevant examples of random matrices are: Random Gaussian Matrices and Random Bernoulli Matrices. Every Gaussian matrix's entry is chosen as i.i.d Gaussian random variables with expectation 0 and variance $1/M$. Every Bernoulli matrix's entry takes value of either $1/\sqrt{M}$ or $-1/\sqrt{M}$ with equal probability. The random noise-like nature of these matrices gives them high probability of exhibiting low mutual coherence with common transform domains such as time, frequency, wavelet, etc. It has been shown that this probability is bounded by:

$$\mathbb{P}(|\|\mathbf{Ax}\|_2^2 - \|\mathbf{x}\|_2^2| \geq \delta_s \|\mathbf{x}\|_2^2) \leq 2e^{-c_0 \delta^2 M} \quad (12.7)$$

where $c_0 > 0$ is a constant and M is the number of rows in \mathbf{A} . It can be further demonstrated that when the number of samples, M , satisfies:

$$M \geq C \cdot S \cdot \log(N/M) \quad (12.8)$$

where C is a constant, we can recover an S -sparse signal from \mathbf{A} with high probability. This bound also shows that the number of samples needed for recovery is linearly dependent on signal sparsity.

1.4.1 Sparse Signal Recovery

Equation (12.3) can be modified to recover the S -sparse signal in any arbitrary bases:

$$\begin{aligned} \hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmin}} \quad & \|\mathbf{v}\|_1 \quad \text{s.t.} \quad \mathbf{y} = \mathbf{A}\mathbf{D}\mathbf{v} \\ \hat{\mathbf{x}} = & \mathbf{D}\mathbf{v} \end{aligned} \quad (12.9)$$

where \mathbf{v} is the sparse representation of signal \mathbf{x} in dictionary \mathbf{D} . Equation (12.9) can be solved via convex optimization using Basis Pursuit algorithm (BP) [19]. In a real system with noise, Eq. (12.9) can be modified to trade off sparsity with congruence of $\mathbf{A}\mathbf{x}$ and \mathbf{y} :

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{D}\mathbf{v}\|_2 + \lambda \|\mathbf{v}\|_1 \quad (12.10)$$

where λ is the coefficient that controls the trade-off between sparsity and reconstruction error, $\|\mathbf{y} - \mathbf{A}\mathbf{D}\mathbf{v}\|_2$. Basis Pursuit De-noising (BPDN) is used to solve Eq. (12.10).

One of the drawbacks of Basis Pursuit algorithms is the computational cost. As alternatives, a number of greedy matching pursuit methods can be used to improve computation speed [20, 21]. These matching pursuit algorithms approximate Eq. (12.9) by solving the following optimization problem:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{A}\mathbf{D}\mathbf{v}\|_2 \quad \text{s.t.} \quad \|\mathbf{v}\|_0 < s_0 \quad (12.11)$$

where s_0 is the constraints on the l_0 -norm of sparse vector \mathbf{v} . Previous work has shown that Orthogonal Matching Pursuit (OMP), a type of the matching pursuit algorithm, can reliably recover an S -sparse vector with length N given $\mathcal{O}(S \cdot \ln N)$ random linear measurements [21].

1.5 Dictionary Learning

Unfortunately, in many applications, most of the standard transforms cannot achieve high sparsity signal representations. To build a more compact representation dictionary, a number of dictionary learning algorithms have been developed.

Given a set of L training signal $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^L$, where $\mathbf{x}_i \in \mathbf{R}^N$, the process of building a signal dependent dictionary is to solve the following optimization problem:

$$\underset{D, \{\mathbf{v}_i\}_{i=1}^L}{\operatorname{argmin}} \sum_{i=1}^L \|\mathbf{x}_i - \mathbf{D}\mathbf{v}_i\|_2^2 \quad (12.12)$$

$$\text{s.t.} \|\mathbf{v}_i\|_0 \leq s_0, 1 \leq i \leq L$$

where $\mathbf{D} \in \mathbf{R}^{N \times P}$ is the signal dependent dictionary. P denotes the size of the dictionary and $\mathbf{v}_i \in \mathbf{R}^P$ is the sparse vector representing the training data \mathbf{x}_i in \mathbf{D} . s_0 is the bound on the ℓ_0 -norm of the sparse vector.

1.5.1 K-SVD Dictionary Learning Algorithm

Several dictionary learning algorithms were proposed to solve Eq. (12.17) [22–25]. Among these algorithms, the K-SVD algorithm has gained popularity due to its simple and efficient training computational steps. There are two iterative steps to the K-SVD algorithm: the sparse coding stage and dictionary update stage. At the beginning, a dictionary, $\mathbf{D} \in \mathbf{R}^{N \times P}$ is first initialized with random values. The columns are then normalized to have l_2 norm of 1. The sparse coding stage then finds a sparse representation, \mathbf{v}_i , for every training vector, \mathbf{x}_i . It finds \mathbf{v}_i from solving the following objective function using greedy matching pursuit methods [20, 21]:

$$\mathbf{v}_i = \min_{\mathbf{v}_i} \|\mathbf{x} - \mathbf{D}\mathbf{v}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{v}\|_0 \leq s_0 \quad (12.13)$$

where s_0 is the upper constraint on the number of non-zero items in the sparse vector \mathbf{v}_i . We combine the sparse vectors into a matrix $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_L\}$. With all the sparse vector computed, the dictionary update stage computes the new dictionary by solving Eq. (12.17) with fixed sparse vectors. However, finding the whole dictionary \mathbf{D} is not possible. So the dictionary update step searches for approximation of \mathbf{D} one column at a time. First, the penalty term in Eq. (12.17) can be rewritten as

$$\begin{aligned} \|\mathbf{X} - \mathbf{D}\mathbf{V}\|_F^2 &= \left\| \mathbf{Y} - \sum_{j=1}^K \mathbf{d}_j \mathbf{v}_T^j \right\|_F^2 \\ &= \left\| \left(\mathbf{Y} - \sum_{j \neq k} \mathbf{d}_j \mathbf{v}_T^j \right) - \mathbf{d}_k \mathbf{v}_T^k \right\|_F^2 \\ &= \|\mathbf{E}_k - \mathbf{d}_k \mathbf{v}_T^k\|_F^2 \end{aligned} \quad (12.14)$$

where \mathbf{v}_T^k is the k th row of matrix \mathbf{V} . Through this decomposition, $\mathbf{D}\mathbf{V}$ is expressed as the sum of K rank-1 matrices. Out of them, $K - 1$ terms are fixed except for k th term. Now, it is tempting to use find the rank-1 approximation to \mathbf{E}_k and update the column \mathbf{d}_k with Singular Value Decomposition (SVD). But this solution would likely yield a dense vector \mathbf{v}_T^k , which is not what we desire.

To avoid this problem, we concatenate the vector and matrices by only keeping the columns that correspond to non-zero entries of \mathbf{v}_T^k . To do this, we define \mathbf{v}_R^k as the concatenated version of \mathbf{v}_T^k , where \mathbf{v}_R^k only contains the non-zero entries of \mathbf{v}_T^k . We also concatenated \mathbf{E}_k in the same way by defining \mathbf{E}_k^R , which discard the column at the index where \mathbf{v}_T^k is zero. Therefore, Eq. (12.14) can be rewritten as:

$$\| \mathbf{E}_k - \mathbf{d}_k \mathbf{v}_T^k \|_F^2 = \| \mathbf{E}_k^R - \mathbf{d}_k \mathbf{v}_R^k \|_F^2 \quad (12.15)$$

we then seek rank-1 approximation of \mathbf{E}_k^R through SVD. $\mathbf{E}_k^R = \mathbf{U} \mathbf{\Delta} \mathbf{V}^T$. We can then update \mathbf{d}_k with the first column of \mathbf{U} and the coefficient vector \mathbf{v}_R^k as the first column of $\mathbf{V} \times \mathbf{\Delta}(\mathbf{1}, \mathbf{1})$.

Following the above procedure, the dictionary updating step finds the new values for all the columns in \mathbf{D} . The process then repeats the sparsity coding step to solve for new representation of \mathbf{V} , and iteratively solve for \mathbf{D} again.

2 Compressed Sensing for Neural Signal Compression

With a brief introduction on some basic concept of compressed sensing, in this section we outline the framework of using compressed sensing for neural signal compression.

2.1 Sampling Framework

To acquire the compressed samples of the neural signal, we must implement Eq. (12.2) at the front-end of sensor. Equation (12.2) can be rewritten as a system of linear equations:

$$\begin{aligned} y_1 &= A_{1,1}x_1 + A_{1,2}x_2 + \cdots + A_{1,N}x_N \\ y_2 &= A_{2,1}x_1 + A_{2,2}x_2 + \cdots + A_{2,N}x_N \\ &\quad \vdots \quad \vdots \quad \vdots \\ y_M &= A_{M,1}x_1 + A_{M,2}x_2 + \cdots + A_{M,N}x_N \end{aligned} \quad (12.16)$$

where $x_1 \cdots x_N$ are the neural signal's samples at discrete time 1 to N , $y_1 \cdots y_M$ are entries of compressed sample \mathbf{y} of length M ($M \leq N$), and \mathbf{A} is the sensing matrix of size $M \times N$.

Typically for high density neural signal acquisition, matrix \mathbf{A} is chosen as a random Bernoulli matrix due to its simplicity. A random Bernoulli matrix contains entries of either +1 or -1s. Hence, the system of Eq. (12.16) can be implemented using M digital accumulators or analog integrator. Depending on the corresponding value of \mathbf{A} , the accumulators either add or subtract digitized signal x_i from the value of the accumulator to generate y_i . Using the ideas of 'Rakeness' together with restricted isometry property, we can also construct optimized version random Bernoulli matrix that is ideal to capture signals that have uneven distribution of energy, such as spikes, where most of the signal energy concentrate under the localized spike waveform [26].

Other matrices such as the random Gaussian and optimized sensing matrices all contain fractional entries [27, 28]. Thus implementing Eq. (12.16) with these choices requires the use of digital or analog multipliers in addition to accumulators. These additional components consume a large amount of chip area. For example, an implementation of a Gaussian matrix using M-DAC occupies around 0.6 mm^2 [29]. This further makes these matrices not suitable for high density neural recording applications.

2.1.1 Analog Implementation

Using the popular choice of random Bernoulli sensing matrix, Eq. (12.16) can be implemented using a parallel bank of M analog integrators [30, 31]. The signal is split into different channels, multiplied by the corresponding sensing matrix value, and then accumulated onto the integrating capacitor. M ADCs then sample the integrator value at a reduced rate of fs/N , where fs is the signal's nyquist sampling rate.

The advantage of the analog implementation is the ability to allow sub-nyquist sampling by reducing ADC speed and energy. But the downside is the increased silicon chip area and complexity needed to implement the parallel integrators [11, 32].

2.1.2 Digital Implementation

Alternatively, we can move the compressed sensing stage after digitization [11, 33]. Here, Eq. (12.16) can be implemented using a parallel bank of M digital accumulators. The digitized signal is split into M channels. Depending on the sensing matrix value, the corresponding accumulator either adds the value or subtracts the value from the accumulator. At the rate of fs/N , the accumulator values are transmitted as the compressed measurement y .

Previous works have examined and cross-compared the analog method with the digital methods [11, 16, 34]. A thorough comparison and analysis involve many parameters, and the reconstruction quality heavily depends on signal SNR. Therefore, we think the best choice of implementation methods should be left to the designers, after considering the target signal of interests.

2.2 Reconstruction Framework

Here we describe different approaches to reconstruct the neural signal from the compressed samples. As outlined in [17, 18], reconstruction involves solving Eq. (12.9). But prior to solving this equation, we must choose an appropriate sparsifying dictionary, \mathbf{D} . The idea is to select \mathbf{D} such that the neural signal \mathbf{x} has

the most sparse representation. Once the dictionary \mathbf{D} is selected, Eq. (12.9) is a basis pursuit problem and can be solved via convex optimization algorithms or be approximated to a great accuracy using greedy matching pursuit algorithms. In this section we first address several considerations for the dictionary selection. We then describe the different algorithms used for solving Eq. (12.9).

2.2.1 Signal Agnostic Dictionary

It has been shown that neural action potential has a sparse representation in time-frequency representations such as Wavelet and Gabor transforms [11, 30]. Therefore it is natural to consider dictionary, \mathbf{D} , as a general time-frequency representation during reconstruction. The advantage of this choice is that no learning and adaptation step is needed to determine the optimal dictionary, \mathbf{D} . Hence, this choice offers an universal approach to neural signal reconstruction using compressed sensing.

But the drawback of this choice is the limited sparse representation capability of these general time-frequency dictionary. Low sparsity leads to reduced signal compression rate, because the number of compressed samples required for recovery, M , is proportional to the signal sparsity, S , shown in Eq. (12.8). Previous literature has reported that CS based on general time-frequency dictionary could achieve at most compression rate of 2.05, but cannot be recommended for higher signal compression rate [35].

2.2.2 Signal Dependent Dictionary

Alternatively, we could choose to construct a signal dependent dictionary to enhance signal sparsity and improve compression rate. Signal dependent dictionary has been widely used in a number of computer vision and media compression application [25, 36]. The idea is to learn a sparse representation dictionary given a rich training data set. A learned dictionary contains key distinct features of the training data set. For example, a dictionary trained using images could contains items representing key decomposable features of the images such as surfaces and edges oriented in different directions [25].

In electrophysiology recordings, it is widely accepted that each neuron's spike has a characteristic shape depending on its morphology and proximity to the recording electrode. This unique shape can be used to cluster and sort the spikes to their corresponding neuron [37–39]. This is more commonly known as spike sorting. Utilizing the underline property of spike sorting, we can also train a signal dependent dictionary based on prior information of the spike to sparsely represent future recorded spikes.

To train a dictionary, we must collect training data at full data rate. The training data should be rich enough ideally containing spike waveform from all the neurons in the neighborhood. With the training data, Eq. (12.17) can be solved to learn a signal dependent dictionary, \mathbf{D} . As mentioned in Sect. 1.5, a number of algorithms are available to solve Eq. (12.17).

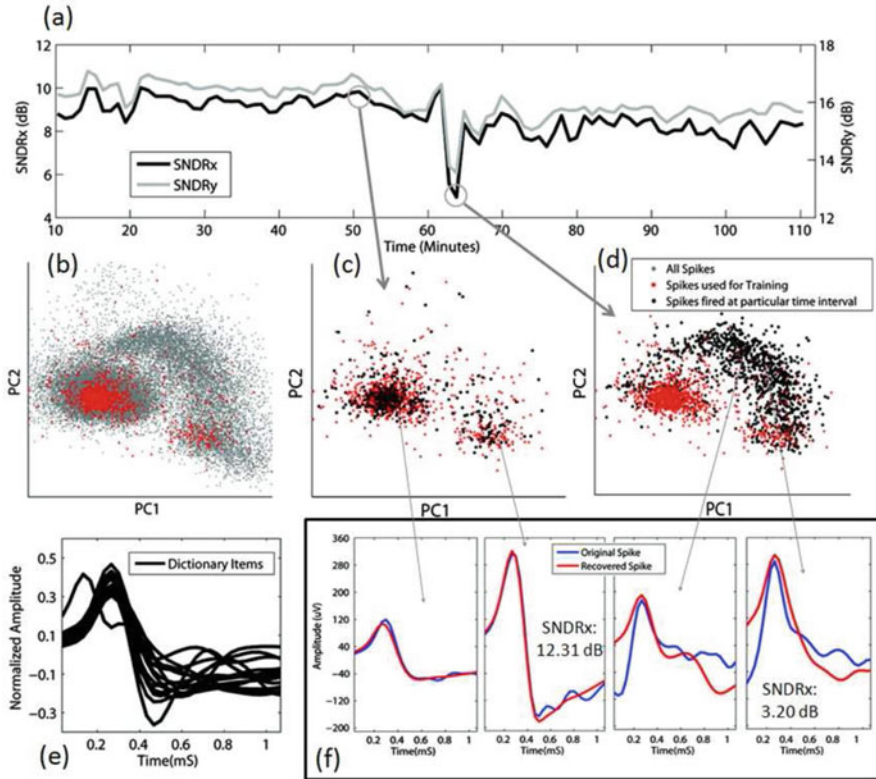


Fig. 12.3 (a) $SNDR_x$ and $SNDR_y$ of the recovery over a 2 h experiment. (b) *Gray dots*: PCA results of all the spikes collected in 2 h experiment. *Red dots*: spikes used for dictionary learning. (c) *Black dots*: spikes recorded around 50–51 min interval. They are well recovered since their shapes do not vary too much from the dictionary. (d) *Black dots*: spikes recorded around 62–63 min interval. A new type of spike starts to appear. As its shape varies significantly from the dictionary, the recovery results deteriorate, shown by a decrease of $SNDR_x$ and $SNDR_y$. (e) The trained dictionary (f) Temporal view of original spikes and recovered spikes at different time intervals

It has been shown that signal dependent dictionary can achieve compression rate on the order of 8–16 [32, 40]. This is much improved than the signal agnostic dictionary [35]. But despite the high compression rate, the biggest disadvantage of a learned dictionary is the lack of adaptation to signal variation. In an electrophysiology experiment, signal waveform could vary from day to day. In addition, the learned dictionary might not be able to well represent neuron’s waveforms that are not included in the training dataset.

For example, in this *in vivo* tetrode recording from the CA1 region of the rat hippocampus (Fig. 12.3), different neurons become active when the rat is sleeping (10–63 min) and running (> 63 min). Thus, dictionary learned using the training data collected at the beginning of the experiment when during sleeping is not able

to reconstruct the spikes that result from rat running. To quantify reconstruction quality, Signal to Noise and Distortion ratios, $SNDR_x$ and $SNDR_y$ are measured. They are defined as:

$$SNDR_x = \frac{1}{T} \sum_{i=1}^T 20 \log \frac{\|x_i\|_2}{\|x_i - \hat{x}_i\|_2}$$

$$SNDR_y = \frac{1}{T} \sum_{i=1}^T 20 \log \frac{\|y_i\|_2}{\|y_i - \hat{y}_i\|_2} \quad (12.17)$$

$$\hat{y}_i = A\hat{x}_i$$

$SNDR_x$ quantifies the reconstruction quality by directly comparing the reconstructed spike waveforms \hat{x} with the original signal x , averaged over a spike train containing T spikes. But most of the time, measuring $SNDR_x$ is infeasible due to lack of the original signal when the system transmits only the compressed measurements. Thus, we can measure $SNDR_y$ after mapping the reconstructed signal back into an estimate of the measurement \hat{y} using the same sensing matrix A . In Fig. 12.3, we can see that $SNDR_y$ closely follows $SNDR_x$, indicating it can be used as an indirect measurement for reconstruction quality.

In the experiment shown in Fig. 12.3, the recovery quality measured by $SNDR_x$ and $SNDR_y$ stays constant above 8 and 15 dB for the time intervals before 60 min. This is because the spikes detected during this time can be well represented by the learned dictionary. Examples of these spikes' PCA plot are shown in Fig. 12.3c, where the black dots represent the spikes detected between 50 and 51 min. Most of the black dots fall onto the PCA space covered by the training data, and therefore have similar shape than the dictionary. However, at around 61–62 min, when the rat is first placed on the treadmill, a lot more spikes are detected that do not fall into the PCA space covered by the training data (Fig. 12.3d), indicating firing of new units. These spikes have different shapes compared to the training data and the learned dictionary, as shown in Fig. 12.3f. Therefore, they cannot be recovered accurately using the learned dictionary. As a result, there exists a significant amount of mismatch between the original and the recovered signal. Both $SNDR_x$ and $SNDR_y$ experience a decrease of 4.5 and 3.5 dB, more than 10 and 8 standard deviation from their corresponding running averages.

2.2.3 Signal Dependent Dictionary with Closed-Loop Quality Feedback

To compensate for the degradation of reconstruction quality during signal variation, we can use $SNDR_y$ as an indicator to trigger additional dictionary learning (DL). Figure 12.4 illustrates the same experiment when an additional dictionary learning is triggered at 61–62 min interval when the measured $SNDR_y$ decreases by more than 4 standard deviation from its running average. Spikes from this time interval

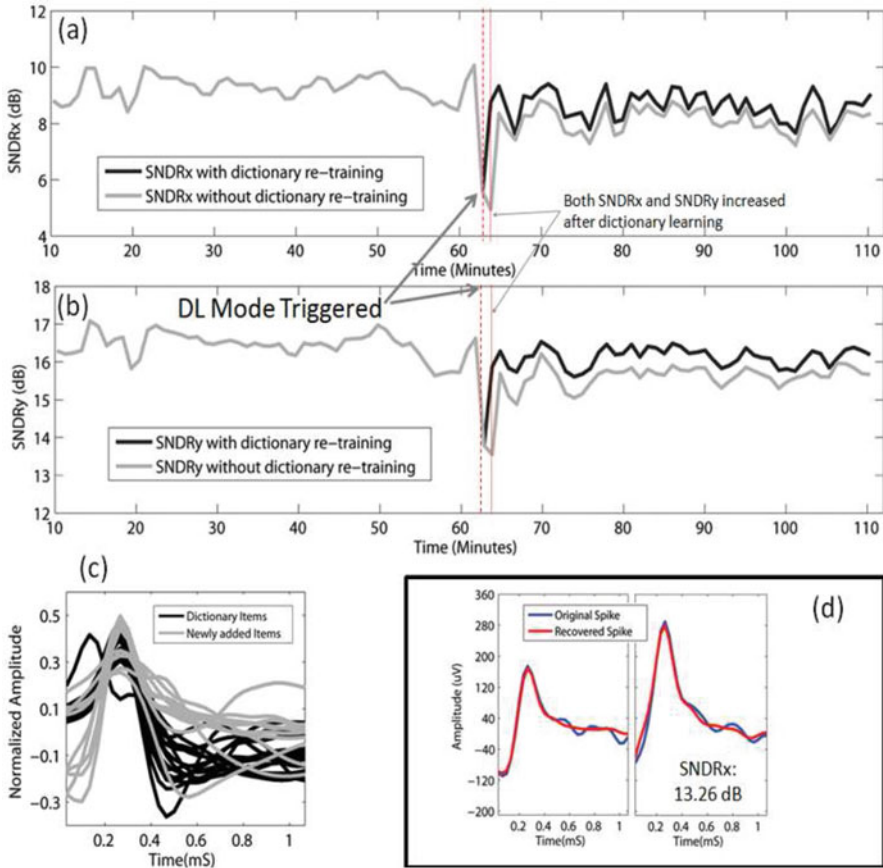


Fig. 12.4 (a) $SNDR_x$ of a 2 h experiment with and without dictionary retaining (b) $SNDR_y$ of a 2 h experiment with and without dictionary retaining (c) The dictionary after dictionary re-training. (d) Temporal view of original spikes and recovered spikes at 62–63 min interval

are used to learn a new dictionary. The new dictionary items are then added to form a dictionary that is used to recover signal after 62 min, shown in Fig. 12.4c. When the system switched from dictionary learning mode back to compression mode, we see an increase of both $SNDR_x$ and $SNDR_y$ at 62–63 min compared to $SNDR_x$ and $SNDR_y$ measurement in Fig. 12.3. Figure 12.4d shows the spikes appeared at 62–63 intervals can now be well recovered using the newly learned dictionary.

3 Compressed Sensing for Multi-Electrodes Compression and Analysis

In the previous section we have showed that compressed sensing can be efficiently used to compress data from high density electrode arrays. But the reconstruction and dictionary learning are based on single electrode assuming no signal correlation between electrodes in the proximity. But in many commonly used electrophysiology recording setup such as the tetrode systems and silicon probes, it is often the case that a neuron's activity is recorded by multiple electrodes in its proximity. Here in this section, we show that this redundancy of multiple features can greatly improve the reconstruction quality and spike sorting accuracy in neural recordings. We also demonstrate that compressed sensing and dictionary learning can also be used to implement unsupervised spike sorting analysis, which can greatly enhance scientists' ability to analyze a large amount of data from these high density electrode arrays.

3.1 Multi-Channel Dictionary Learning Using Joint-Group Sparsity

To enhance the detectability of units from single electrode, neuroscientists have employed multiple electrodes in close proximity to record a small region of interest. For example in the tetrode system [41, 42], 4 electrodes are placed in close proximity within a radius of $< 50 \mu\text{m}$. When implanted, it can simultaneously register neural activity in the nearby brain region. Depending on the location of neurons relative to each channel, the shapes and amplitudes of the spikes from same neuron could vary across four channels. These redundant features can be used to enhance spike sorting capabilities during multi-unit activities.

We can also take advantage of signal correlation across multiple electrode in our dictionary learning. We can bring the idea of joint-group sparsity into the dictionary learning and reconstruction steps. Joint-group sparsity is used to enforce that spikes seen at electrodes in close proximity are reconstructed by similar items from the dictionary. These constraints could further increase the compression rate while guaranteeing good reconstruction quality.

3.1.1 Signal Model

To illustrate this, we assume there are C multi-channels and G groups (clusters) of neural spikes in the data samples $\mathbf{X} \in \mathbb{R}^{N \times T}$. We incorporate a key feature in the signal model: joint-group sparsity. As shown in Fig. 12.5, dictionary \mathbf{D}_c is a concatenation of sub-dictionary $\mathbf{D}_{c,g}$, where c and g separately indicate the indices of channels and groups. Intuitively, if neural signal x_c associated with channel c

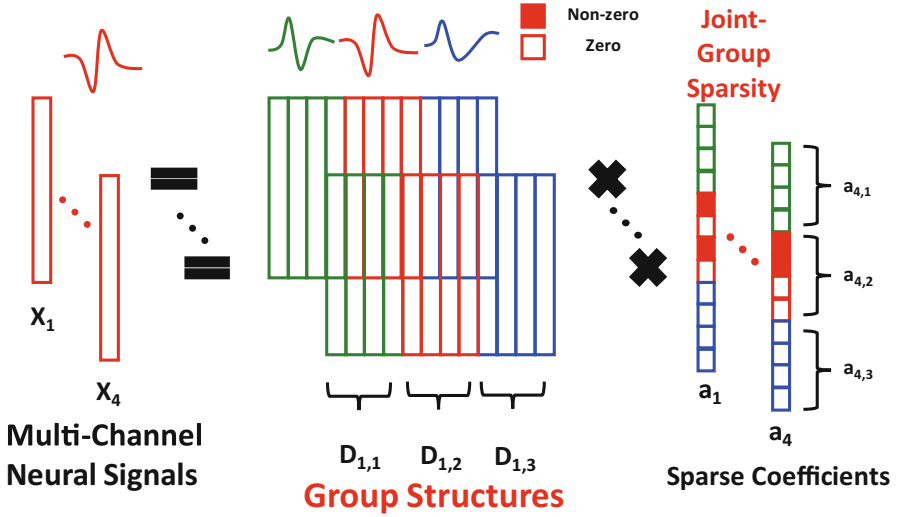


Fig. 12.5 Intuitive illustration of the proposed signal model with discriminative group structures (color-coded blocks) and joint-group sparsity (red filled) for multi-mode structured dictionary learning

belongs to group g , it should be ideally represented by the corresponding dictionary $\mathbf{D}_{c,g}$, in which only the elements of its sparse vector $\mathbf{a}_{c,g}$ are possibly non-zero. For example, $\mathbf{D}_{2,3}$ contains atoms which can sparsely represent the neural signals of group 3 collected from channel 2. $\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \dots, \mathbf{x}_{C,t}$, the windowed segments from raw data, indicate the neural signals collected by multi-channels at timestamp t . Since they are recorded by electrodes such as tetrodes simultaneously in the close proximity, $\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \dots, \mathbf{x}_{C,t}$ share a similar pattern in terms of particular shapes. Taking the correlation among channels into account, we implement a joint-group sparsity in our signal model to further improve the recovery quality, spike sorting accuracy, and compression ratio. \mathbf{a}_c indicates the sparse coefficient vector of channel c while $\mathbf{a}_{c,g}$ indicates the sub-vector of \mathbf{a}_c . Among C channels, \mathbf{A} is defined as $[\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_C] = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_G]^T$. \mathbf{A}_g ($g = 1, 2, \dots, G$) is the sub-matrix associated with group G . Based on these definitions, joint-group sparsity is defined as:

$$\|\mathbf{A}\|_{\text{group},0} = \sum_{l=1}^G I(\|\mathbf{A}_g\|_F > 0) = 1,$$

$$\|\mathbf{a}_c\|_0 \leq S, \forall c.$$

In our formulation, I is the indicator function and S denotes the sparsity. $\|\mathbf{A}\|_{\text{group},0}$ is constrained as 1 to determine the corresponding group g of the spike. $\|\mathbf{A}_g\|$ denotes the Frobenius norm. Therefore, the mathematical definition of the proposed signal model is:

$$\mathbf{x}_c = [\mathbf{D}_{c,1} \ \mathbf{D}_{c,2} \ \dots \ \mathbf{D}_{c,G}] [\mathbf{a}_{c,1}^\top \ \mathbf{a}_{c,2}^\top \ \dots \ \mathbf{a}_{c,G}^\top]^\top,$$

$$\|\mathbf{A}\|_{\text{group},0} = 1, \|\mathbf{a}_c\|_0 \leq S, \forall c.$$

Intuitively, a spike should be represented by atoms from a corresponding group \mathbf{D}_g , and also be constrained by the information given by neighboring electrodes. Taking neighboring spikes into account, the compression ratio $\frac{M}{N}$ can be further improved, which also promotes the performance of the neural recording systems in terms of the power efficiency. Therefore, the dictionary learning task can be formulated as:

$$\min_{\mathbf{D}_{c,g}, \mathbf{a}_{c,g}} \sum_{c=1}^C \|\mathbf{x}_{c,t} - \mathbf{D}_{c,g} \mathbf{a}_{c,g}\|_2 \text{ s.t.}$$

$$\|\mathbf{A}\|_{\text{group},0} = 1, \|\mathbf{a}_{c,g}\|_0 \leq S.$$

Here, we find out the best sparse representation $\mathbf{a}_{c,g}$ of each spike $\mathbf{x}_{c,t}$ in the training samples based on each sub-dictionary $\mathbf{D}_{c,g}$. Taking advantage of such a sparse domain with structures, we are able to significantly improve the reconstruction quality as well as enable online spike sorting.

3.1.2 Unsupervised Dictionary Learning

To make dictionary learning unsupervised, we can first adopt clustering techniques to initialize the dictionary with a discriminate group structure. For example, if we use the spectral clustering method [43], the dictionary initialization is divided into two stages: (1) initialization of the similarity matrix, and (2) spectral clustering. The similarity matrix \mathbf{E} represents the quantitative assessment of similarity between spikes. The similarity matrix is generated based on the nearest-neighbor method and the similarity of neural signals in the multi-channel is defined as:

$$e(t, t') = \sum_{c=1}^C \|\mathbf{x}_{c,t} - \mathbf{x}_{c,t'}\|_2,$$

$$t, t' \in \{1, 2, \dots, T\}, t \neq t'.$$

where $\mathbf{x}_{c,t}$ denotes the spike from c th channel and timestamp t and $e(t, t')$ denotes the summation of squared Euclidean distance between t th and t' th spikes from channels 1 to C (i.e., $C=1$ indicates single channel). The smaller the $e(t, t')$ is, the closer correlation the neural signals share with each other. The elements of t th row of similarity matrix \mathbf{E} are set to 1 if the corresponding indexes belong to the K smallest set while the others are set to 0. Then, we update the similarity matrix by $\mathbf{E} = \mathbf{E} + \mathbf{E}^T$.

After the similarity matrix \mathbf{E} is generated, we pre-define the group number G and then adopt the spectral clustering to group neural signals into G different clusters, providing prior information to help initialize the dictionary with group structures.

Given the clustering information \mathbf{g} , the dictionary \mathbf{D}_c of c th channel is defined as:

$$\mathbf{D}_c = [\mathbf{D}_{c,1} \ \mathbf{D}_{c,2} \ \dots \ \mathbf{D}_{c,G}].$$

$\mathbf{D}_{c,g}$ indicates the sub-dictionary of \mathbf{D}_c , in which its atoms are randomly picked up from the group of cluster g . We also obtain the mean shape, also defined as centroids $\mathbf{c}_{c,g}$ associated with a distinct cluster, which is used for template matching in the sparse coding stage. Centroids $\mathbf{c}_{c,g}$, representing the template and a particular pattern of groups, g are found by:

$$\mathbf{c}_{c,g} = \frac{1}{|\mathcal{S}_g|} \sum_{t \in \mathcal{S}_g} \mathbf{x}_{c,g}, \quad \mathcal{S}_g = \{t | g_t = g\}.$$

After initializing the dictionary \mathbf{D}_c ($c = 1$ indicates single channel, $c > 1$ indicates the multi-channel), the unsupervised multi-mode structured dictionary learning is basically divided into two stages in each iteration, similar to the K-SVD algorithm [25]: the sparse coding stage and the dictionary update stage.

In the sparse coding stage, we introduce the joint-group sparsity and then solve the sparse representation problem below using Orthogonal Matching Pursuit (OMP) [21],

$$\begin{aligned} \min_{\mathbf{a}_{c,g}} \sum_{c=1}^C \|\mathbf{x}_{c,t} - \mathbf{D}_{c,g} \mathbf{a}_{c,g}\|_2 \text{ s.t.} \\ \|\mathbf{A}\|_{\text{group},0} = 1, \|\mathbf{a}_{c,g}\|_0 \leq S. \end{aligned}$$

Here, we find out the best sparse representation $\mathbf{a}_{c,g}$ of each spike $\mathbf{x}_{c,t}$ in the training samples based on each sub-dictionary $\mathbf{D}_{c,g}$. Then, we use the linear combination coefficient $\lambda \in (0, 1)$ to balance the residual of the sparse representation and the squared Euclidean distance between the spike and its centroids. Thereby, the cluster g of the spike is determined by solving the problem below,

$$\min_g \sum_{c=1}^C \{\lambda \|\mathbf{x}_{c,t} - \mathbf{D}_{c,g} \mathbf{a}_{c,g}\|_2 + (1 - \lambda) \|\mathbf{D}_{c,g} \mathbf{a}_{c,g} - \mathbf{c}_{c,g}\|_2\}.$$

As shown in Fig. 12.6, the squared Euclidean distance for mean shape matching provides another evaluation of the similarity of spikes in the sparse representation stage. Given the group g of each spike, we define a trust region set \mathcal{S}_g associated

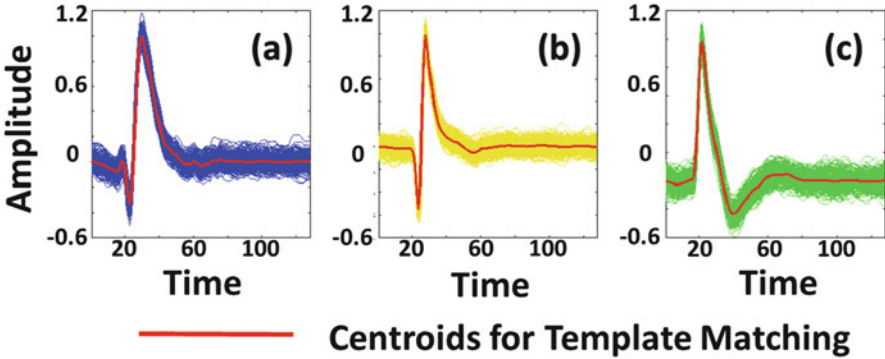


Fig. 12.6 The illustration of different groups of spikes with distinct shapes. The *red color-coded spikes* indicate the centroids (mean shape) associated with the corresponding groups. The mean shape matching provides another perspective of similarity in the sparse coding stage

with the group g . To construct the trust region set \mathcal{S}_g , we add the index of spike t into it if the spike is represented perfectly in the sparse coding stage, which indicates the reconstruction error is smaller than the pre-defined *error*. Intuitively, the trust region set \mathcal{S}_g only contains spikes with high reconstruction quality in each learning iteration.

In the dictionary update stage, we simply fix the sparse coefficients matrix \mathbf{A}_c and update each atom of the dictionary using the same approach in the K-SVD [25]. While the K-SVD updates the dictionary based on the whole training samples, our approach only updates it based on the current trust region set \mathcal{S} , which is the union of set \mathcal{S}_g . Iteratively, the trust region covers the entire training samples. Figures 12.7 and 12.8 illustrate that the trust region set \mathcal{S} approaches the entire training samples after several learning iterations. Furthermore, we dynamically update the centroid $\mathbf{c}_{c,g}$ depending on the clustering result obtained from the sparse representation stage. As shown in Fig. 12.8, the average recovery error converges as the trust region \mathcal{S} covers the entire training samples.

Taking advantage of the iterative refinement in the dictionary learning, we are able to correct the spike sorting error generated by the dictionary initialization, as shown in Fig. 12.9. Figure 12.9a shows the similarity matrix initialization step mistakenly clusters some spikes, which are denoted as blue dots and distributed in the cluster of green dots. But as shown in Fig. 12.9b, after the dictionary learning, the PCA result illustrates that the spike sorting performance is improved as mistaken spikes are sorted into the correct clusters.

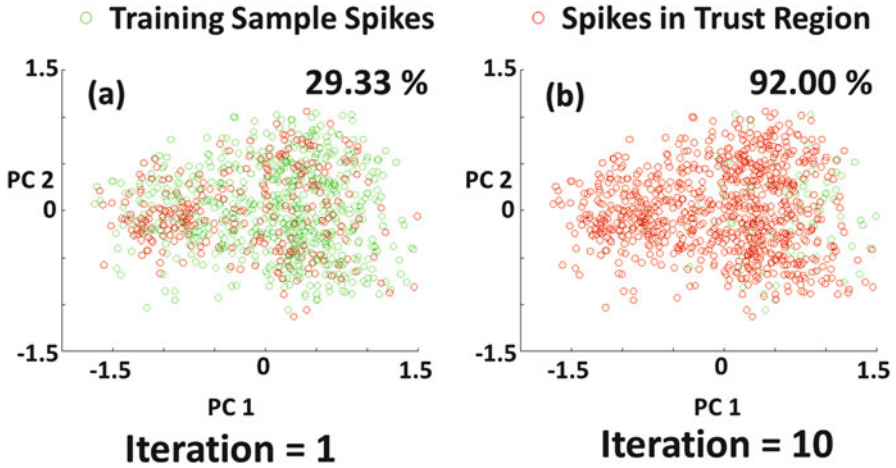


Fig. 12.7 An illustration of how the trust region \mathcal{S} performs in principal component analysis (PCA). As we iterate from 1 to 10, the percentage of spikes in the trust region increases from 29.33% to 92.00%, indicating that most spikes in the training samples satisfy the pre-defined reconstruction quality after 10 iterations

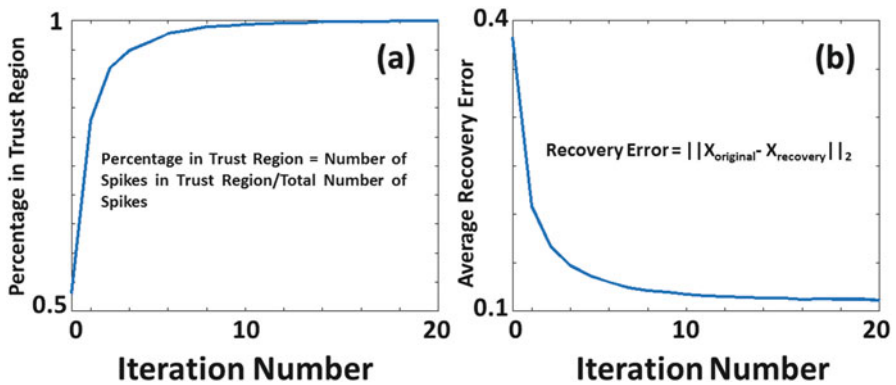


Fig. 12.8 (a) indicates the illustration of percentage change in trust region \mathcal{S} and (b) indicates average recovery error as dictionary learning iterates from 1 to 20

3.2 Reconstruction and Spike Sorting Using Group Sparsity

Once a structured dictionary is trained using join-group sparsity that enforces clustering information, reconstruction and spike sorting are relatively straightforward. Assuming using a random Bernoulli sensing matrix, \mathbf{A} , we can use Algorithm 1 to reconstruct the waveform $\hat{\mathbf{X}}$ and cluster information g from compressed measurement \mathbf{y}_c , where c denotes the channel number.

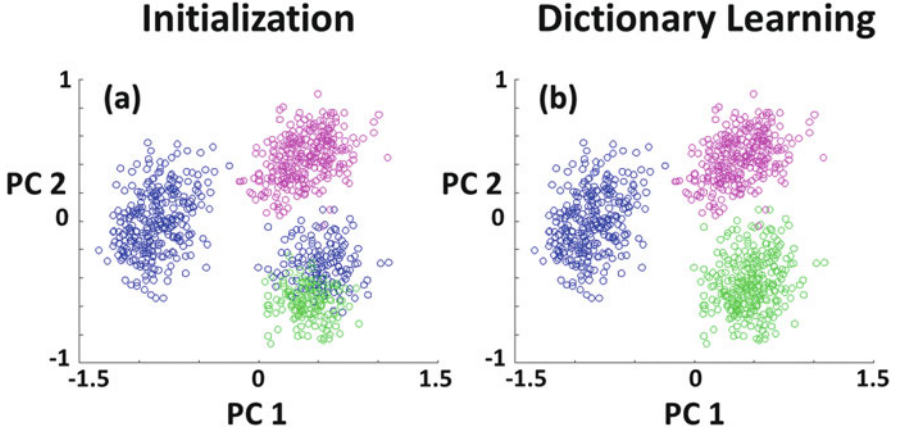


Fig. 12.9 An example of robustness of spike sorting from the perspective of sparse coding, visualized in the PCA domain. The iterative refinement helps correct the mistakenly sorted spikes generated from the initialization

Algorithm 1 Reconstruction and spike sorting approach

Require: The initialized dictionaries \mathbf{D}_c , the centroids $\mathbf{c}_{c,g}$, measurements \mathbf{y}_c , where $c = 1, 2, \dots, C$ ($C = 1$ indicates single channel) and random Bernoulli matrix \mathbf{A} . Number of clusters G , sparsity S and linear combination coefficient $\lambda \in (0, 1)$.

1: Solve the representation problem via Orthogonal Matching Pursuit [21],

$$\min_{\mathbf{a}_{c,g}} \sum_{c=1}^C \|\mathbf{y}_c - \mathbf{S}\mathbf{D}_{c,g}\mathbf{a}_{c,g}\|_2 \text{ s.t. } \|\mathbf{a}_{c,g}\|_0 \leq S, \forall g.$$

2: Determine the cluster g of spikes by solving following problem,

$$\min_g \sum_{c=1}^C \{\lambda \|\mathbf{y}_c - \mathbf{S}\mathbf{D}_{c,g}\mathbf{a}_{c,g}\|_2 + (1 - \lambda) \|\mathbf{y}_{c,g} - \mathbf{S}\mathbf{c}_{c,g}\|_2\}.$$

3: **Return** The recovered signal $\hat{\mathbf{x}}_c = \mathbf{D}_{c,g}\mathbf{a}_{c,g}$ and cluster g .

The unsupervised dictionary learning and reconstruction method using joint-group sparsity shows superior performance for both single channel recording and multi-channel recordings. The performance evaluation is done on pre-recorded dataset. Datasets used include Leicester dataset, a synthetic test dataset from a wavelet based cluster software, WavClus, from the University of Leicester [37]; MGH dataset, consists of primates recordings conducted at the Massachusetts General Hospital on monkeys ‘‘Pogo’’ and ‘‘Romeo’’ [44]; and HC-1 dataset, consists of tetrode recordings from mice hippocampus [45].

Table 12.1 Comparison of reconstruction performance (in SNDR) of different CS methods on “Leicester”

Database	CS approach	CR = 20:1	CR = 10:1
Easy	Proposed approach	10.44	11.60
	K-SVD + OMP	9.23	10.40
	Data dictionary + OMP	7.11	7.76
	Wavelet + OMP	-1.81	-0.82
Difficult	Proposed approach	8.64	10.21
	K-SVD + OMP	6.40	8.03
	Data dictionary + OMP	6.38	6.64
	Wavelet + OMP	-2.71	-1.78

Table 12.2 Comparison of classification performance (in CA) of different CS methods on “Leicester”

Database	CS approach	CR = 20:1	CR = 10:1
Easy	Proposed approach	97.62	98.08
	K-SVD + OMP	94.04	98.05
	Data Dictionary + OMP	93.88	95.77
Difficult	Proposed approach	90.24	95.87
	K-SVD + OMP	74.03	86.84
	Data Dictionary + OMP	73.24	78.22

3.2.1 Single Channel

For single channel, Tables 12.1 and 12.2, and Fig. 12.10 demonstrate the reconstruction and spike sorting performance on the Leicester database at a compression ratio of 20:1 and 10:1. The proposed approach outperforms the other CS-based approaches, and achieves an average gain of 2 dB and 4% in terms of SNDR and classification accuracy on the “Easy” database at the CR of 20:1. For the “Difficult” database, the approach attains more than 90% spike sorting success rate, while achieving a CR of 10:1–20:1. Tables 12.3 and 12.4 show the reconstruction and spike sorting performance of the MGH “Pogo” and “Romeo” databases, respectively. Here too, the proposed approach outperforms the other CS-based approaches. Especially, the proposed approach shows more than 90% spike sorting success rate at the CR of 10:1, and achieves an average gain of 30% over other methods.

3.2.2 Multi-Channel

For multi-channel dataset, hc-1, the joint-group sparsyt/Unsupervised dictionary learning shows average gain of 4–5 dB over other CS-based approaches in terms of SNDR in multi-channel reconstruction (Table 12.5). Figure 12.11 illustrates the

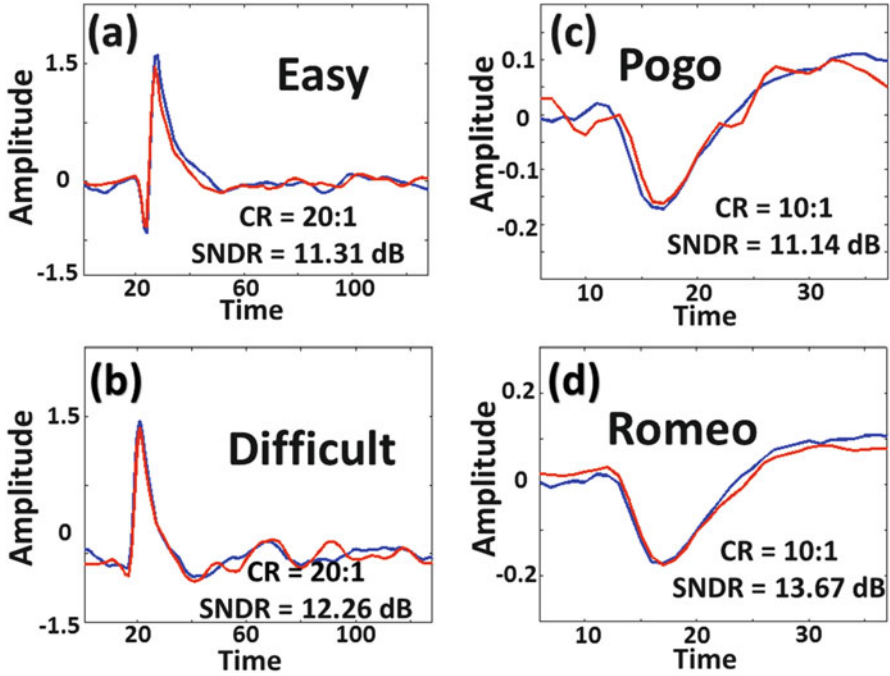


Fig. 12.10 Examples of reconstruction performance of single channel neural recordings. For (a)–(d), recovered signals (*red*) still preserve the major features of original signals (*blue*) at CR of 20:1 and 10:1, respectively. (a) and (b) demonstrate synthetic spikes from the Leicester database [45] while (c) and (d) demonstrate real spikes from the MGH database [44]

Table 12.3 Comparison of reconstruction and classification performance of different CS methods on “Pogo”

Database	CS approach	CR = 10:1	CR = 5:1
SNDR (dB)	Proposed approach	7.46	10.30
	K-SVD + OMP	2.73	7.34
	Data Dictionary + OMP	6.96	8.18
	Wavelet + OMP	-1.51	-1.17
CA (%)	Proposed approach	93.63	95.11
	K-SVD + OMP	51.07	64.07
	Data Dictionary + OMP	64.13	82.59

Table 12.4 Comparison of reconstruction and classification performance of different CS methods on “Romeo”

Database	CS approach	CR = 10:1	CR = 5:1
SNDR (dB)	Proposed approach	8.20	11.00
	K-SVD + OMP	3.14	7.27
	Data Dictionary + OMP	6.69	8.00
	Wavelet + OMP	-1.53	-1.27
CA (%)	Proposed approach	94.54	96.88
	K-SVD + OMP	44.32	57.98
	Data Dictionary + OMP	62.65	88.66

Table 12.5 Comparison of reconstruction performance (in SNDR) of different CS methods on “hc-1”

Database	Compressed sensing approach	CR = 4:1	CR = 8:1
hc-1-14521	Proposed approach	13.93	8.85
	K-SVD + OMP	6.03	3.18
	Wavelet + OMP	-0.05	-3.64
hc-1-14531	Proposed approach	12.18	9.55
	K-SVD + OMP	7.17	3.43
	Wavelet + OMP	-0.38	-1.97
hc-1-14921	Proposed approach	12.64	8.54
	K-SVD + OMP	7.73	4.12
	Wavelet + OMP	-0.24	-2.02

multi-channel reconstruction example on the hc-1 database at a CR of 8:1. The blue signals denote the original spikes recorded from the tetrodes setup, which shows similar pattern and correlation among the four channels as shown in Fig. 12.11. The red signals denote the spikes recovered by the proposed CS-based approach. As shown in Fig. 12.11, the recovered signals still preserve most of the features, even though only 12.5% of the information of the original signals is used for the reconstruction. The proposed approach is also able to sense and reconstruct neural signals in the continuous time domain, including the low activity region between spikes.

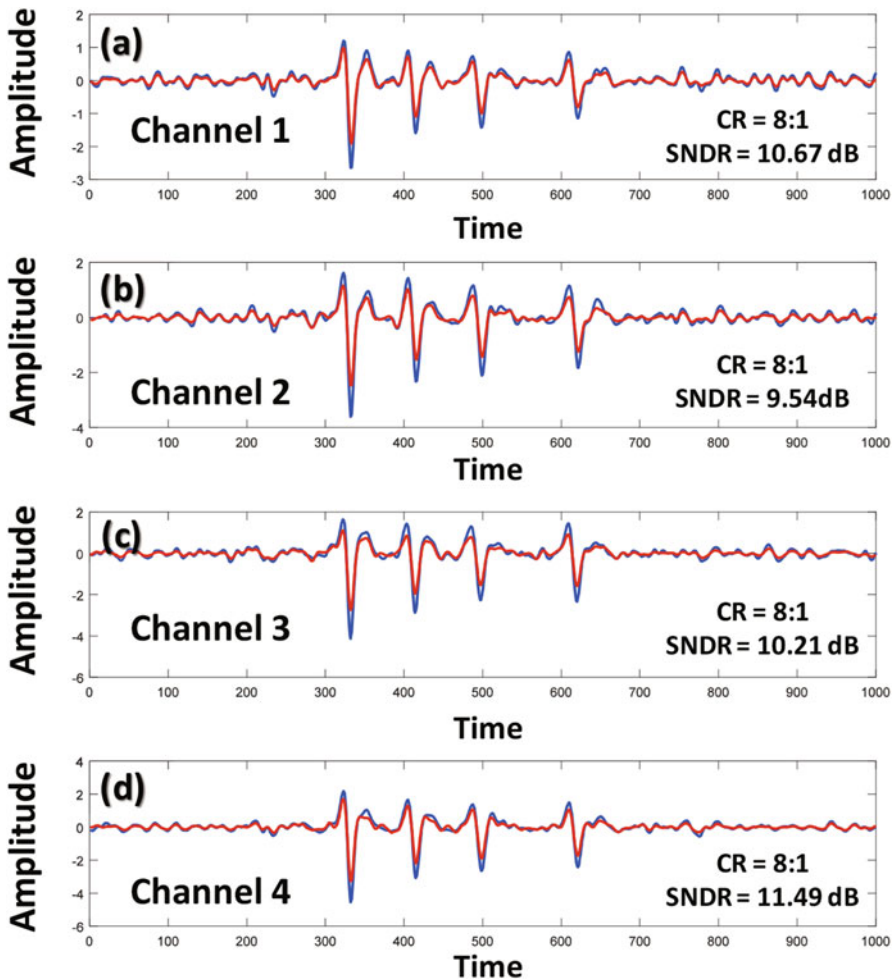


Fig. 12.11 An example of reconstruction performance of multi-channel neural recordings on the hc-1 database [45] at a CR of 8:1. *Blue and red spikes* indicate the original neural spikes and the recovered neural spikes, respectively

References

1. M.A. Wilson, B.L. McNaughton, Dynamics of the hippocampal ensemble code for space. *Science* **261**(5124), 1055–1059 (1993)
2. S. Mitra, J. Putzeys, F. Battaglia, C.M. Lopez, M. Welkenhuysen, C. Pennartz, C. Van Hoof, R.F. Yazicioglu, 24-channel dual-band wireless neural recorder with activity-dependent power consumption, in *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)* (IEEE, New York, 2013), pp. 292–293
3. R.J. Staba, C.L. Wilson, A. Bragin, I. Fried, J. Engel, Sleep states differentiate single neuron activity recorded from human epileptic hippocampus, entorhinal cortex, and subiculum. *J. Neurosci.* **22**(13), 5694–5704 (2002)

4. J.N. Aziz, K. Abdelhalim, R. Shulyzki, R. Genov, B.L. Bardakjian, M. Derchansky, D. Serletis, P.L. Carlen, 256-channel neural recording and delta compression microsystem with 3d electrodes. *IEEE J. Solid State Circuits* **44**(3), 995–1005 (2009)
5. D.H. Hubel, T.N. Wiesel, Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**(3), 574–591 (1959)
6. E.M. Maynard, C.T. Nordhausen, R.A. Normann, The Utah intracortical electrode array: a recording structure for potential brain-computer interfaces. *Electroencephalogr. Clin. Neurophysiol.* **102**(3), 228–239 (1997)
7. C.M. Lopez, A. Andrei, S. Mitra, M. Welkenhuysen, W. Eberle, C. Bartic, R. Puers, R.F. Yazicioglu, G.G. Gielen, An implantable 455-active-electrode 52-channel CMOS neural probe. *IEEE J. Solid State Circuits* **49**(1), 248–261 (2014)
8. R. Shulyzki, K. Abdelhalim, A. Bagheri, M.T. Salam, C.M. Florez, J.L. Perez Velazquez, P.L. Carlen, R. Genov, 320-channel active probe for high-resolution neuromonitoring and responsive neurostimulation. *IEEE Trans. Biomed. Circuits Syst.* **9**(1), 34–49 (2015)
9. D. Seo, J.M. Carmena, J.M. Rabaey, M.M. Maharbiz, E. Alon, Model validation of untethered, ultrasonic neural dust motes for cortical recording. *J. Neurosci. Methods* **244**, 114–122 (2015)
10. A. Khalifa, J. Zhang, M. Leistner, R. Etienne-Cummings, A compact, low-power, fully analog implantable microstimulator, in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, New York, 2016)
11. F. Chen, A.P. Chandrakasan, V.M. Stojanović, Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors. *IEEE J. Solid State Circuits* **47**(3), 744–756 (2012)
12. A.P. Chandrakasan, S. Sheng, R.W. Brodersen, Low-power CMOS digital design. *IEICE Trans. Electron.* **75**(4), 371–382 (1992)
13. S. Kim, R. Normann, R. Harrison, F. Solzbacher et al., Preliminary study of the thermal impact of a microelectrode array implanted in the brain, in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2006. EMBS'06* (IEEE, New York, 2006), pp. 2986–2989
14. D.A. Borton, M. Yin, J. Aceros, A. Nurmikko, An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. *J. Neural Eng.* **10**(2), 026010 (2013)
15. H. Gao, R.M. Walker, P. Nuyujukian, K.A. Makinwa, K.V. Shenoy, B. Murmann, T.H. Meng, Hermese: a 96-channel full data rate direct neural interface in 0.13 μm CMOS. *IEEE J. Solid State Circuits* **47**(4), 1043–1055 (2012)
16. J. Zhang, K. Duncan, Y. Suo, T. Xiong, S. Mitra, T.D. Tran, R. Etienne-Cummings, Communication channel analysis and real time compressed sensing for high density neural recording devices. *IEEE Trans. Circuits Syst. Regul. Pap.* **63**(5), 599–608 (2016)
17. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
18. D.L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
19. S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit. *SIAM Rev.* **43**(1), 129–159 (2001)
20. D. Needell, J.A. Tropp, Cosamp: iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26**(3), 301–321 (2009)
21. J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* **53**(12), 4655–4666 (2007)
22. M.S. Lewicki, B.A. Olshausen, Probabilistic framework for the adaptation and comparison of image codes. *JOSA A* **16**(7), 1587–1601 (1999)
23. M.S. Lewicki, T.J. Sejnowski, Learning overcomplete representations. *Neural Comput.* **12**(2), 337–365 (2000)
24. K. Engan, S.O. Aase, J.H. Husøy, Multi-frame compression: theory and design. *Signal Process.* **80**(10), 2121–2140 (2000)
25. M. Aharon, M. Elad, A. Bruckstein, K-svd: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **54**(11), 4311–4322 (2006)

26. M. Mangia, R. Rovatti, G. Setti, Rakeness in the design of analog-to-information conversion of sparse and localized signals. *IEEE Trans. Circuits Syst. Regul. Pap.* **59**(5), 1001–1014 (2012)
27. J.M. Duarte-Carvajalino, G. Sapiro, Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization. DTIC Document, Technical Report (2008)
28. M. Elad, Optimized projections for compressed sensing. *IEEE Trans. Signal Process.* **55**(12), 5695–5702 (2007)
29. D. Gangopadhyay, E.G. Allstot, A.M. Dixon, K. Natarajan, S. Gupta, D.J. Allstot, Compressed sensing analog front-end for bio-sensor applications. *IEEE J. Solid State Circuits* **49**(2), 426–438 (2014)
30. Z. Charbiwala, V. Karkare, S. Gibson, D. Marković, M.B. Srivastava, Compressive sensing of neural action potentials using a learned union of supports, in *2011 International Conference on Body Sensor Networks (BSN)* (IEEE, New York, 2011), pp. 53–58
31. M. Shoran, M.H. Kamal, C. Pollo, P. Vanderghenst, A. Schmid, Compact low-power cortical recording architecture for compressive multichannel data acquisition. *IEEE Trans. Biomed. Circuits Syst.* **8**(6), 857–870 (2014)
32. J. Zhang, Y. Suo, S. Mitra, S.P. Chin, S. Hsiao, R.F. Yazicioglu, T.D. Tran, R. Etienne-Cummings, An efficient and compact compressed sensing microsystem for implantable neural recordings. *IEEE Trans. Biomed. Circuits Syst.* **8**(4), 485–496 (2014)
33. M. Zhang, A. Bermak, Compressive acquisition CMOS image sensor: from the algorithm to hardware implementation. *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **18**(3), 490–500 (2010)
34. D.E. Bellasi, L. Benini, Energy-efficiency analysis of analog and digital compressive sensing in wireless sensors. *IEEE Trans. Circuits Syst. Regul. Pap.* **62**(11), 2718–2729 (2015)
35. C. Bulach, U. Bihl, M. Ortmanns, Evaluation study of compressed sensing for neural spike recordings, in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, New York, 2012), pp. 3507–3510
36. Q. Zhang, B. Li, Discriminative k-svd for dictionary learning in face recognition, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, New York, 2010), pp. 2691–2698
37. R.Q. Quiroga, Z. Nadasdy, Y. Ben-Shaul, Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Comput.* **16**(8), 1661–1687 (2004)
38. T. Sasaki, N. Matsuki, Y. Ikegaya, Action-potential modulation during axonal conduction. *Science* **331**(6017), 599–601 (2011)
39. P.H. Thakur, H. Lu, S.S. Hsiao, K.O. Johnson, Automated optimal detection and classification of neural action potentials in extra-cellular recordings. *J. Neurosci. Methods* **162**(1), 364–376 (2007)
40. B. Sun, W. Zhao, X. Zhu, Training-free compressed sensing for wireless neural recording using analysis model and group weighted-minimization. *J. Neural Eng.* **14**(3), 036018 (2017)
41. C.M. Gray, P.E. Maldonado, M. Wilson, B. McNaughton, Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex. *J. Neurosci. Methods* **63**(1), 43–54 (1995)
42. K.D. Harris, D.A. Henze, J. Csicsvari, H. Hirase, G. Buzsáki, Accuracy of tetrode spike separation as determined by simultaneous intracellular and extracellular measurements. *J. Neurophysiol.* **84**(1), 401–414 (2000)
43. A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, in *Advances in Neural Information Processing Systems*, vol. 14 (2001), pp. 849–856
44. W.F. Asaad, E.N. Eskandar, Encoding of both positive and negative reward prediction errors by neurons of the primate lateral prefrontal cortex and caudate nucleus. *J. Neurosci.* **31**(49), 17772–17787 (2011)
45. D.A. Henze, Z. Borhegyi, J. Csicsvari, A. Mamiya, K.D. Harris, G. Buzsáki, Intracellular features predicted by extracellular recordings in the hippocampus in vivo. *J. Neurophysiol.* **84**(1), 390–400 (2000)

Chapter 13

Very Large-Scale Neuromorphic Systems for Biological Signal Processing

Francky Catthoor, Srinjoy Mitra, Anup Das, and Siebren Schaafsma

1 Introduction

Various technological advancements have made it possible to record and accumulate extremely large volume of biological data in a very short amount of time. However, interpreting the data is an altogether different problem. These recording techniques might result in high-resolution images (of the brain), high-density neural signals, genetic or metabolic data from blood sample, continuous visual/auditory signal, etc. Processing of large-scale biological data is still in its infancy and limited to algorithmic complexities which are within the reach of customized ICs. But the trend is rapidly driving beyond this possibility. Deep neural networks are often called into action for processing and interpreting the resulting ‘big data’. One school of thought aims at a minimal amount of local processing, followed by a transfer of the resulting data to the cloud where all the massive global centralized analysis will take place. This is the basis of the cloud computing paradigm which has gained enormous momentum. It has the big advantage of being able to manage and maintain all the data centrally where a high efficiency can be supported and where massive signal fusion can be obtained. However, as with all centralized approaches, it has the disadvantage of having to transfer all the data to the more central cloud. This entails inevitably a longer latency and increased local energy consumption in the transfer of the data. For some important applications where feedback loops are important, e.g. to enable autonomous computing with continuous learning, this

F. Catthoor (✉) • A. Das • S. Schaafsma
IMEC, Leuven, Belgium
e-mail: catthoor@imec.be

S. Mitra
School of Engineering, University of Glasgow, University Avenue, Glasgow G12 8QQ, UK

latency is detrimental. In addition, when the available energy is very scarce and the total signal bandwidth is high, sending all the data to the outside of the sensor node or the local hub can rapidly become prohibitive.

For such applications, we want a more dedicated local computing platform solution to support different algorithms. It is well known that the human brain is much more suited to interpret various biological signals, and that too at a low power and latency. We would hence want to have a dedicated so-called neuromorphic computing platform for distributed local processing. Neuromorphic computing, from its inception has been driven by a vision of emulating the computing power of the brain, both in architecture and performance [1, 2]. Over the last few years, there has been a regained interest by several research groups including corporate R&D. One major reason for this is the advent of new non-CMOS devices that could integrate many more synapses (also probably neurons) within a small silicon area. Unfortunately, considering the relentless progress in standard von-Neumann computing architecture, the slow turnaround time of individual neuromorphic chips have often resulted in a suboptimal solution compared to the state-of-the-art digital processors. Hence it has been very difficult to show much commercial relevance to such chips. One of the important reasons behind this is the interconnect scaling problem. Neuromorphic engineers have mostly focused on creating and solving relatively small problems that would demonstrate the superiority of a brain-inspired hardware over a more traditional approach [3–5]. These ICs were built around the assumption that they could be connected together to obtain the number of neurons necessary for solving large-scale, practical and relevant problems.

1.1 Human Brain-Scale Architecture

In order to justify the advantage of an ultra-low-power full-custom mixed-signal neuromorphic ASIC, it is now essential to demonstrate the feasibility of an architecture that is much larger or even of similar size to that of the human brain. Neuromorphic engineers are aware that it is not a simple problem and there has always been an effort to progress in this direction [6, 7].

However, none of the previous attempts have reached the neuron count or the synaptic density that can be considered anything remotely close to the human brain; 10^{10} neurons and 10^{15} synapses [8]. This is true even if we ignore the unrealistic volume and high power consumption of a benchtop neuromorphic chip compared to a mammalian cortex. In recent reviews [7, 9] of high-density neuromorphic architecture, the authors compared the biggest systems till date [10–13]. While none of them crosses the one million neuron mark/chip, the largest conceived multichip, multi-PCB rack consists of 460 million neurons. On the other hand, the largest functional brain model built to date [14] has 2.3 million neurons and runs $9000\times$ slower than real-time on a 16-core PC. Much larger simulations (10^{10} neurons using a 96-rack blue gene supercomputer [15]) have been reported, but without any bio-realistic model or function. As apparent, neither the hardware nor the software has reached the scale necessary for real-time implementation of existing

algorithms or development of novel models for truly human brain-scale function. Even though the number of neurons and synapse targeted in this work doesn't consider a specific application and is rather arbitrary, we expect that a realistic architecture and technology proposal for truly human brain-scale neuromorphic hardware is overdue.

The applications can include classifying visual images of organs/cells (for neuroscientific or point-of-care medical applications), natural images or sounds (for autonomous robots), high-density neural recording (for brain-computer interface), etc. Many of these applications require local processing power (with limited energy and area budget) and cannot only rely on the latency of cloud computing on large servers. This chapter is a white paper that provides an overview of a proposed architecture and the necessary technology that would be highly suitable for these applications. We provide a tutorial approach and do not go deep into implementation details or experimental work.

2 System-on-Chip and Neural Networks

The problem faced by a scalable and ultra-dense neuromorphic system is not unlike the interconnection problem in ultra-scaled modern system-on-chips (SoCs). These SoCs contain many different processing cores communicating with each other and with many distributed memories in the layered background memory organization through an intra- and inter-tile communication network. Tiles are formed by a group of tightly connected cores (processors), between which the activity exceeds a certain threshold level. One important design feature of the SoCs relate to the length of the interconnection between the clusters. State-of-the-art solutions have relatively long time-multiplexed connections that need to be near continuously powered up and down, reaching from the ports of the data producers/consumers (inside the tiles or between different tiles) up to the ports of the communication switches. SoC inter-tile communication network solutions are based on different types of buses (shared or not) and networks-on-chip (NoCs) that are much more power efficient. One approach to build a large-scale neuromorphic system is to borrow the tile-based communication framework of a modern system-on-chips [16, 17]. However, today's neuromorphic systems with tile-based SoC style communication do not yet offer the required flexibility and energy efficiency of a human brain-scale system, which we address in this work.

Most standardized communication solutions today target chip-to-chip protocols. A well-established example is PCI Express standard. PCIe is more specialized to efficiently transport large blocks of memory at a high speed. These existing standards share with our proposal the interesting properties of enabling parallel concurrent bus transfers and hierarchical composition. However, these solutions are typically not flexible in letting a huge amount of sources exchange data with an equally huge amount of destinations. In a neuromorphic architecture, both the source and destinations can be in the millions! Due to the communication

protocol overhead (required to enable the standardization across several internal data send/receive options), they will then consume much more than the minimal energy spent to transmit the data words when only a dedicated and fully matched protocol is used. Hence, the flexibility comes at a high price of energy loss. In this architecture we consider implementing a very large number of long-range synaptic (global synapse) connections for physically separated neuron clusters. Hence we propose a strongly integrated intra-chip/3D stack solution, utilizing circuit switching and not packet switching, for a very simple synchronization scheme. Our architecture is composed of a heavily pruned and partitioned global synaptic network with a mesh or a crossbar-type organization used in a local synapse array. Later in this chapter, we will justify this choice and also discuss the difference with a few popular on-chip communication structures, like the AMBA-style time-multiplexed segmented buses, the fully connected crossbar meshes and the multistage NoCs.

Typically, neural network algorithms use a multilayer network with one or more hidden layers and sometimes also with recurrent feedback layer. These layers, in theory, should be fully connected with one another to allow all possible flexibility, which leads to a densely connected neural array. In practice, only a limited amount of these connections will be active, but as it is not known upfront (at design time) where the active connections are situated, many hardware implementations also foresee a crossbar of some type. For an array with N neurons in the input layer and M neurons in the output layer, $N \times M$ synapses are then necessary. However, typically at least one so-called hidden neuron layer is also present with K neurons, as shown

in Fig. 13.1a. One example of this is the restricted Boltzmann machine or its successor deep belief network [18]. The figure on the left shows the logical connection, while the right-hand figure is an abstract physical implementation. Here a crossbar connection between N - K and K - M indicates that all possible connections between these layers are realizable. The crossbar can be considered as passive

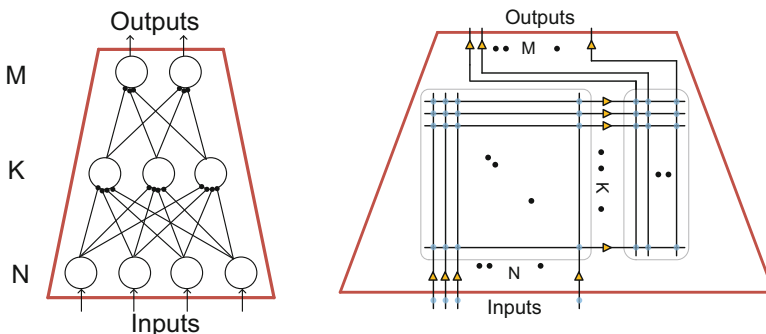


Fig. 13.1 Conventional neural network arrays organized in layers: illustration with three neuron arrays (input, output and hidden layers) and two dense local synapse arrays. Logical diagram on the *left* and possible implementation (using crossbar) on the *right*. Neurons and synapses are marked in *yellow* and *blue*, respectively

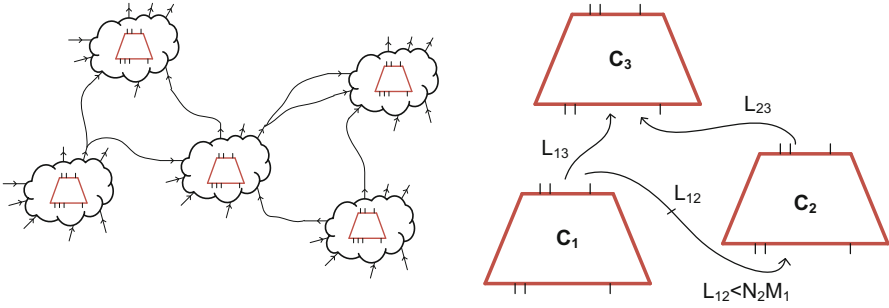


Fig. 13.2 (left) Connection between neural clusters. (right) Two neural clusters are connected with a subset of all possible connections between the neuron layers of the clusters

synaptic connections that create the link between pre- and postsynaptic neurons. However, these local synapses do not need to be passive but can store a weight (even show synaptic plasticity), depending on the implementation method [19, 20].

Neurons will typically need to communicate across layers as well. But not all connections need to be present, so no full crossbar is needed here. That is partly the case already between the input and output layer when a hidden layer is present (illustrated in Fig. 13.1 where not all possible connections are realizable between the N input neurons and M output neurons, except when $K = N \times M$, which would lead to too much overhead). This is especially so when several clusters of densely connected neural arrays are communicating with each other (see Fig. 13.2, right where two clusters (C_1 and C_2) are shown with their mutual L_{12} ($< N_2 \times M_1$) connections that form a subset of all possible connections). However, upfront at design time of the SoC, it is not known which specific neuron connections (e.g. which of the L_{12} connections in Fig. 13.2) are needed later.

In this work, we do not make any specific assumption on the communication protocol between neurons or neuron clusters. The communication can be completely spike driven (e.g. address-event representation [21]) where the only information transferred between two neurons is the spike timing. The information could also consist of a certain number of bits (say 8 bits or 16 bits) representing the synaptic weights or a combination of both. In any of these cases, one primary requirement is the existence of a very large number of synaptic connections, both short and long range. Hence, a connecting line between neuron clusters (or neurons) can be considered either a single wire or a bus of W wires (where W represents the resolution in synaptic weight). The basic description of the envisioned system does not depend on the size of W , unless it is too large. Since an 8-bit (or at most 16-bit) resolution is most likely more than enough to represent the synaptic strength, the wiring complexity due to this issue ($W = 1$, or $W > 1$) is insignificant compared to communicating between 10^{10} neurons.

3 State-of-the-Art Overview for Scalable Synapse Networks

Researchers focused on large-scale neuromorphic computing mostly address the dense local synapse array using traditional SRAMs or emerging non-volatile memories like phase-change memory (PCM), resistive RAM (ReRAM) or STT-MRAM. Several examples of this can be found in recent literature [22–27]. These are very well suited for supporting local 2D arrays of synaptic connections between neuron layers. The problems of scaling this up to human brain size with billions of neurons are many [28], with a main bottleneck on the large fan-in/fan-out requirements, power consumption and noise margin challenge due to the accumulated sneak currents. Looking more into detail at the global synapse communication problem, as formulated above, there is however also a clear need for further scalable solutions. A similar observation can be made for inter-core communication networks in SoCs. Some prior art approaches to solve the global intercluster communication bottleneck with low energy, while still covering a wide application range, are now discussed with more details. Existing solutions can roughly be divided into two main categories.

3.1 *Restricted Connectivity*

A first set of solutions is characterized by a restricted connectivity. Somewhat regular locally connected architectures are usually used in this approach, similar to systolic arrays.¹ Two main options are available for time multiplexing: local sequential global parallel (LSGP) or the opposite, local parallel global sequential (LPGS). The concepts are illustrated in Fig. 13.3. They are quite effective for local arrays that have a very limited global communication which, namely, has to pass through the time-multiplexed global connection scheme. The Spinnaker project of the University of Manchester [13] is mostly based on this. With heavy time multiplexing for the longer connections, this restricts the maximal bandwidth available for the global data connections. This is motivated for neural algorithms which are quite local in nature and which communicate very infrequently to neurons that are spatially distant. However, if one would size up the number of time-multiplexed buses to have trillions of global synapses connected billions of neurons in a sufficiently flexible way, a huge amount of long wiring (and associated switching energy) resources would be necessary.

¹A homogeneous network of tightly coupled data processing units (DPUs) called cells or nodes.

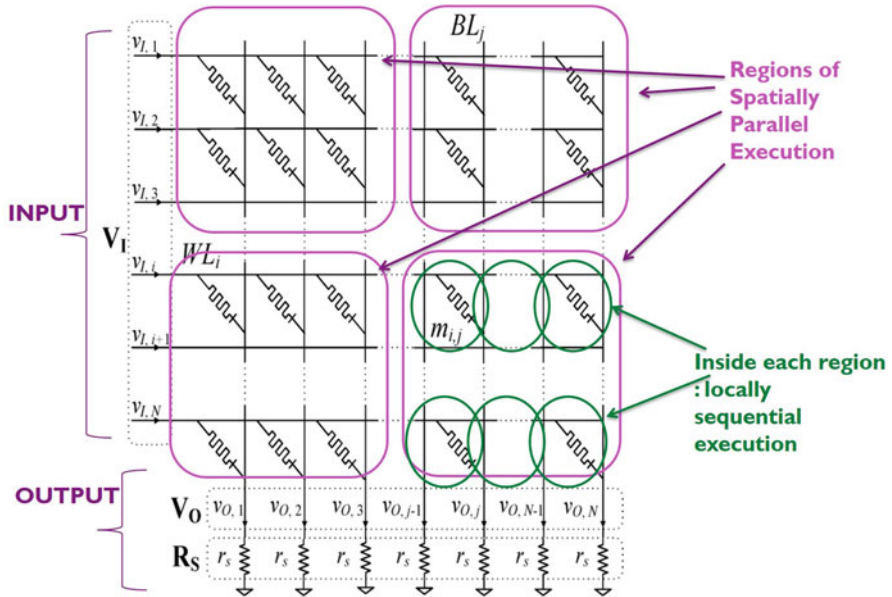


Fig. 13.3 LSGP solution for local synapse array time multiplexing

3.2 Full Connectivity

In a second category of prior art solutions, full connectivity is maintained. This is clearly not scalable to dimensions of 10^{10} neurons as present in the human brain. Of course, the brain does not have all-to-all connectivity, and a reduction from the theoretical N^2 ($=10^{20}$) connections to about 10^{15} synapses is observed in the human brain [8, 29]. Furthermore, most of these are inactive for a large part of the instantiated processing. Even then, 10^{15} (or a few orders lower) is still a gigantic number, and to our knowledge, there has been no effort to reach anywhere close to this level of configurable connectivity with sufficient bandwidth for nonlocal communication. Though nowhere close to this number, strong time-multiplexing architectures are sometime used to scale up the synaptic count within a neural cluster. They also pay a price in the storage of the time-multiplexed synapse weights and their access because the solution is typically a partly distributed on-chip and partly centralized off-chip main memory storage. The TrueNorth chip [16], which is mostly based on on-chip SRAMs combined with off-chip denser memories, follows this principle.

To implement hidden layers more effectively, we believe it is best to use LPGS where the highly dynamic global connectivity can be exploited in a flexible time-multiplexed software-enabled way. Since intraneural cluster connections are generally ‘static’, it is most suitable to link that to the spatially parallel hardware domain (i.e. without time multiplexing). One then still has to make sure that

interconnections are not too long (e.g. by limiting the intra-cluster size). This creates a first new sub-branch of solutions. An alternative new sub-branch is obtained if one opts for a more dynamic architectural solution. These two additional sub-branches are further discussed below.

3.2.1 Full Connectivity: Predefined

The first sub-branch comprises of solutions with pseudo-static predefined full connectivity. Multistage networks have some form of crossbar implementation requiring N^2 transfers. This necessitates an enormous area and power budget for very large values of N . A partial solution exists in power-gating all connections not required during the actual running of an application instance, restricting the overall energy consumption. However, the area requirement still remains and, consequently, a strong energy overhead for scaled technology nodes due to the needlessly long lines in the oversized layout. The TrueNorth project uses this approach because it has a very dense crossbar communication on-chip but reduces the maximal bandwidth drastically for the inter-chip board-level communication where heavy time multiplexing is exploited [16]. This approach has allowed to support reasonable synapse array sizes with a high amount of intra-array flexibility. It is sufficient for multilayer neural algorithms that are built out of stages of neural kernels with local feedforward communication. Even though TrueNorth has reached one million neurons on a single chip, the inter-chip solution does not provide such a massive bandwidth. For neural algorithms that have more complex combinations of longer distance feedforward and feedback connections, this methodology cannot promise a sufficiently flexible scaling up beyond what can fit on a single chip.

This solution is still not attractive in our specific target context due to the lack of full scalability and of sufficient configurable bandwidth in the semi-global connections. While relying on SRAM or large non-volatile memories with a high energy cost per access, it requires a potentially much larger energy budget compared to fully customized (nonflexible) communication solutions. So, it is not well suited for embedded or potentially even portable usage and instead can target mainly ‘shared servers in the cloud’.

3.2.2 Full Connectivity: Dynamic

Solutions in the second sub-branch have dynamic full connectivity. They exploit the property that longer intercluster connections are needed more rarely. It is not known upfront where these connections are situated though, so a run-time layer is needed to accommodate the actual transfers at instantiation time. Until now this branch is very rarely explored. To the best of our knowledge, only a few recent instances are published in this direction. One way to achieve dynamic full connectivity is exploiting hardware-based control protocols using some type of statically allocated network-on-chip (NoC) or multistage network approach. This approach is adopted

in particular in the ETH Zurich approach [17, 30]. A LPGS solution is used there to obtain a parallel implementation of a quite strongly connected ‘static’ intra-cluster organization and a largely sequential (time-multiplexed) implementation of more sparsely connected time-varying intercluster communication. In [31] another version of a NoC is presented but with fixed in hardware at design time. Even though both these NoC approaches are stated as scalable, in practice the size and energy still go up clearly super-linearly, and they will also become prohibitive when billions of neurons have to be connected in a configurable way and only the target of the data messages routed through this fixed network can be programmed. Patent application US2015/058268 (IBM) presents a hierarchical, scalable neuromorphic system for synaptic and structural plasticity. However, the obtained scalability is again still limited: local connections are performed with ‘sparse crossbar tables’. Clearly that does not allow realizing global connections in a sufficiently flexible way. Since the system is dimensioned at design time, this solution does not achieve high scalability, flexibility and low power simultaneously.

Hence, to reach our specific target goal, we have to further alleviate the intermediate length interconnection problems encountered in global data communication networks connecting many computation clusters. We believe the latter is possible by exploring the dynamic sub-branch with fully virtualized middleware-based run-time approaches steering a special type of programmable switch network. The envisioned scaled-up neuromorphic architecture outlined in the rest of this chapter will have local clusters of closely packed CMOS neurons connected via highly dense local synapse matrix (most likely in some form of memresistive non-volatile technology like an RRAM, PCM or STT-MRAM array). These clusters will be connected to one another using configurable global synapses. We expect a very advanced technology node (<25 nm) will be necessary to reach the required neuronal density and a vertical metal stack of 30 layers. The metal layers are required to accommodate the local and global synaptic connectivity all on single chip. In fact there are no clear mechanical limit exists on the number of metal layers which can be stacked (including the TFT transistors sandwiched in between them). Conventional belief is that the current way with up to 12 layers is only there for cost and necessity, not for any fabrication restrictions. Hence, with the advent of scaled CMOS technology, the possibility of such a technology is a highly realistic scenario.

In the following description, we do not make any assumption on the method through which the local synapse matrix is designed and whether any synaptic plasticity is inherent to the architecture.

4 Summary of Main Features of Proposed Approach

In this section, we describe a method for designing data communication network wherein the intermediate length interconnection problems are solved so that full dynamic connectivity and scalability are achieved while drastically reducing the power consumption. This approach describes a data communication network con-

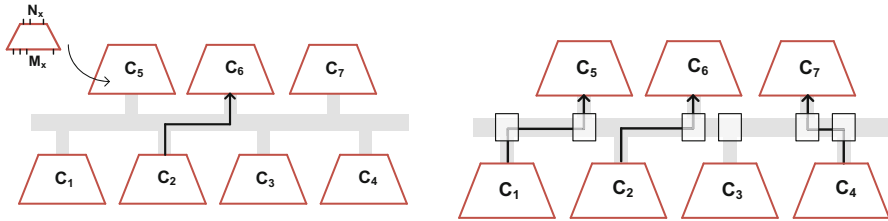


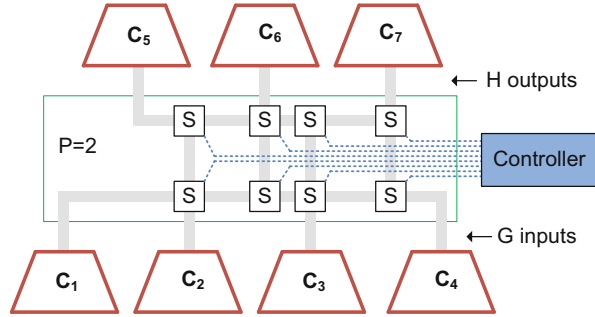
Fig. 13.4 Illustration of a simple bus connecting seven neuron clusters (C_x) sharing a W -bit-wide bus. *Left*: In a shared bus, once connection between C_2 and C_6 is established, it cannot be used for any other connections. *Right*: A segmented bus where many simultaneous connections are possible

necting multiple computation clusters. The communication network is designed for receiving data via G input ports, $G > 1$ (input signals from one or more clusters), and for sending signals out to one or more clusters via H output ports ($H > 1$). This network is used for interconnecting clusters, and the corresponding control signals are arranged for concurrently activating up to P parallel data buses. This will result in a bidirectional parallel interconnection between P of the G input ports ($P < G$) and P of the H output ports ($P < H$), via paths of activated segments of the segmented bus network, the segments being linked by means of segmentation switches.

Let's consider a simplified example where this intermediate network is supposed to connect seven neural clusters (C_1 – C_7). For simplicity, we assume outputs of C_1 – C_4 can only connect to inputs of C_5 – C_7 without any feedback or recurrent connections. Figure 13.4 (left) shows a single shared bus connecting G input ports ($G = M_1 + M_2 + M_3 + M_4$) to H output ports ($H = N_5 + N_6 + N_7$), where M_x is the number of output ports of cluster C_x and N_x is its input ports. While using shared bus, only one connection between any pair of input-output cluster is possible at a given time (shown by the arrow in Fig. 13.4 left), resulting in underutilization of the bus. A simple segmented bus allows to overcome this problem by breaking the bus into multiple segments. As seen from Fig. 13.4 (right), $P = 1$ can allow few more possible connections (even though it guarantees only one in worst-case scenario). A higher number for P would allow for many more possible simultaneous connections. It is to be noted that, to guarantee all possible connections at all time, we would require $G \cdot H$ separate buses. However, such a full connection is rarely exercised in a typical neural processing. A more realistic scenario is the one where at different time intervals a different connection topology is active. Figure 13.5 shows a segmented bus where $P = 2$, along with the control plane.

By providing a run-time-controlled parallel segmented bus network for interconnecting clusters and using up to P parallel data buses, we propose a bidirectional parallel interconnection in a fully dynamical fashion. This is analogous in principle to virtualization approach which is employed in modern high-performance computing platforms, however, with difference in its implementation. The applications are not written with direct reference to the processor and storage resources but in terms of virtual operators and virtual data. The latter are then run-time managed to be

Fig. 13.5 The parallel segmented bus with vertical stubs ($P = 2$) is shown with its control signals. The bus connects $G (= M_1 + M_2 + M_3 + M_4)$ input ports to $H (= M_5 + M_6 + M_7)$ output ports



allocated/assigned in space and time to the physical resources. This is particularly useful when a large amount of dynamism is present and when the actual worst-case concurrency is nearly never occurring at run-time. In this method, one can fully virtualize the global synaptic connections between neuron clusters. Hence, the bandwidth allocation needed at design time can be reduced from the physical maximum to what is maximally happening concurrently, namely, P . Consequently, the active wire length is reduced, and hence the energy overhead can also be drastically lowered. The proposed solution allows for scaling by adapting the number of parallel buses, P .

4.1 Determining the Number of Parallel Buses (P)

One method to estimate P is to use from a profiled histogram of concurrently occurring intercluster connection patterns in at least one given application. P will then typically be the maximum number of concurrently required data interconnections of the required connection patterns in a profiled histogram (see Sect. 5 for more information). Some safety margin for accommodating unforeseen extra communication capacity can be implemented by selecting P parallel buses among S available buses of the segmented bus network during pruning based on a particular application.

This method can also include a step to perform time-division multiplexing of the concurrently required data interconnections of the profiled histogram. Here the time-multiplexing factor should not exceed the ratio of the realizable clock frequency of the implementation fabric and the minimal frequency at which the application needs to be repeatedly executed on this fabric. The latter corresponds to any required data rate for the synaptic communication that needs to be achieved between inputs and outputs of the clusters. The time-division multiplexing of the synapse hardware should be preferably organized according in a LPGS scheme as already mentioned in Sect. 2.

4.2 Active BEOL Network

In order to save space and to reduce the routing length and capacitance of the most active communication paths, we propose to implement this data communication network, at least partly, in the back-end-of-line (BEOL) fabric. By doing so, the scalability and in particular the energy efficiency of the proposed solution are even more improved.

A cartoon of a BEOL device is shown in Fig. 13.6 (red). The entire metal stack above the active silicon layer (front end of line: FEOL) for the BEOL is completely passive in all traditional CMOS processes. However, thin-film technology (TFT) devices are now possible to be integrated between the BEOL layers [32]. Materials like gallium-indium-zinc-oxide (GIZO) are used for such devices, which exhibit extremely low leakage.

The realization of some of the segmentation switches in BEOL allows directly reducing the vertical wire length in a substantial way. Since metal wires do not have to go back and forth to the CMOS (FEOL) layer for every active device in the segmented bus, a huge number of vias and corresponding parasitic capacitance can be eliminated. As a significant amount of switches can be removed from the FEOL layer, the horizontal wire length is also reduced and subsequently reducing the overall area. It is expected that the global intercluster connections is stable for long periods of time, so the BEOL devices do not have to switch at themost

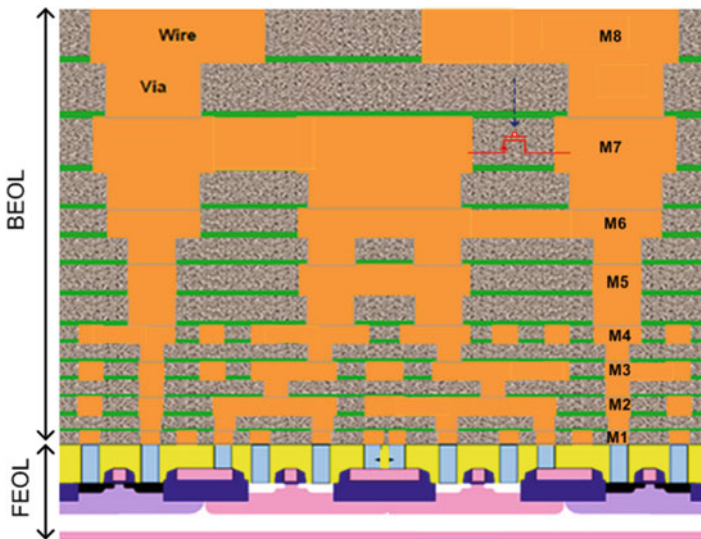


Fig. 13.6 Cross section of a typical CMOS wafer with 8 metal BEOL connecting FEOL devices. BEOL devices acting as a switch (shown in red) can considerably reduce the vertical wiring. The control line is indicated in blue

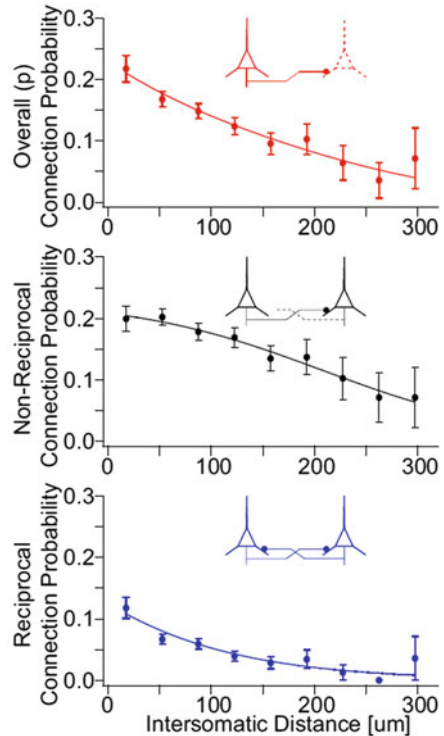
advanced clock rates. This is an added benefit since standard FEOL devices would have excessive leakage in advanced technology nodes and static switches would need strong speed optimization.

5 Detailed Description of the Network Data Plane

5.1 Design-Time Connection Profiling

One of the primary assumptions in this architecture is related to the number of global connections required between neural clusters that are far apart. It is obvious that only a few of such long-range global connections are simultaneously required at the highest level, but not statically the same over time. In reality, connection lengths between clusters are distributed with decreasing upper bound as a function of intercluster distance [33, 34]. Figure 13.7 shows the connection probability of the neurons depending on the distance between the neural clusters in the human brain. This substantiates our claim that these are monotonously decreasing, but they do not become negligible so the current solutions mentioned in the state-of-the-art overview of Sect. 2 do not scale up sufficiently well.

Fig. 13.7 Illustration of the probability synaptic connection as function of the distance between neurons [33]



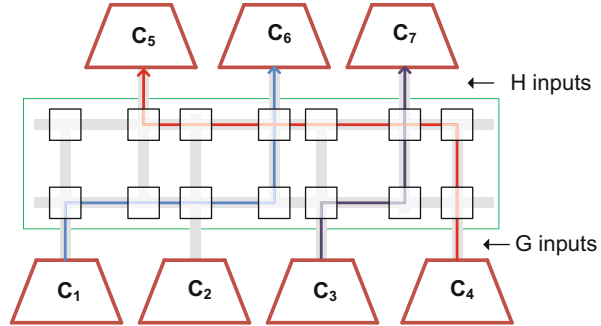
For the actual implementation, these histograms will be obtained upfront from executing a representative set of target applications with their corresponding long test stimuli. The histogram data could be obtained from profiling the intercluster data communication network and the processing clusters that are communicating. This can, for instance, happen with a neural application simulation framework like CARLsim of UC Irvine [35]. Such a simulator may run for days or even weeks to allow us to trace the spike trains across different neurons and synapse groups. Contemporary CAD tools for complex SoCs exhibit similar execution times and are well accepted in ASIC design community. We have extended this framework to produce the required activity information for both the neurons and the synapses, and this can be fed into an extended network architecture simulator to analyse the activity on the global synapse network, i.e. on the dynamic segmented bus in our case. Hence, the profiling can already be completed at design time, but the stored histogram information can also be adapted with new information derived at run-time. For instance, assume that at design time a broad targeted set of applications has a particular histogram shape (which can in principle be any ‘monotonically decreasing’ curve), but at run-time, only one particular application (e.g. image recognition) is running for a very long time. Then the specific shape of the histogram will differ during that period, which can be exploited in dimensioning the actively powered-up communication architecture organization and in this way further reducing the dynamic and leakage energy.

5.2 Run-Time Configuration

We envision to build a system that has a rather broad range of target applications. This broad target market is also needed to amortize the huge nonrecurring engineering (NRE) cost of advanced CMOS technologies, so as to make these custom ICs economically viable. Hence, programmable and heavily reusable platforms are needed. It is expected that most neuromorphic applications will need a very large number of neurons to be active in parallel. This results in a large interconnection bottleneck if a rather broad target application domain is envisioned.

The solution is a middleware based run-time approach where the required data connections are dynamically allocated to reduce the number of parallel global bus. Due to the middleware control, true dynamic full connectivity can be achieved. This solves the global intercluster communication bottleneck with low energy while still covering a wide range of applications. By exploiting the upper-bound distance graph combined with the profiling information of the histogram, less parallel bandwidth allocation is needed at design time. The use of a segmented bus network optimized for the required distance graph is proposed here. It connects available clusters (C_x) to each other via switches. As already motivated, this can be achieved by allocating only P parallel buses, where P is the maximal amount of simultaneously active global transfers. Figure 13.8 shows a use-case of the segmented bus where three simultaneous connections are possible even when $P = 2$.

Fig. 13.8 Representative block scheme for a dynamically controlled segmented bus network (with $P = 2$)



The actual value of P can be obtained from one of the profiled histograms discussed in the previous subsection. The histograms have the number of simultaneous active connections on the x -axis and an indication of how often this occurs (e.g. as a percentage between 0% and 100%) in a representative application benchmark set on the y -axis. If a threshold is then imposed on the minimum percentage of occurrence (y -axis in the histogram), one can typically discard the right-hand side of the histogram with the larger amounts of simultaneous active connections. At run-time, it could still be possible that the limit imposed at design time is exceeded. In that case, the communication then have to be delayed to the next cycle. This is especially easy to achieve in an asynchronous intercluster communication protocol. This restriction of the maximal P should be combined with activity-based floor planning to further reduce the energy (see Sect. 4). In that case, a block ordering is performed based on the profiling information. By utilizing the segmented bus network topology, also existing physical bus libraries like AMBA-lite [36] can be reused wherever possible (if the control protocol allows this).

This approach substantially deviates from the prior art solutions and enables truly scalable ultra-low-energy global connections. Here, the dynamic run-time flexibility of biochemical connections in the brain is mimicked by similar flexibility and energy efficiency in a middleware-controlled time-shared segmented bus network. It should be noted that apart from the feedforward network examples shown above, feedback and recurrent connections can be easily implemented in such neuron clusters.

6 Detailed Description of the Hierarchical Floor Plan Organization Aspects

Until now we have concentrated on the data plane architecture of the communication network between neural clusters without going into details of their topological arrangement and the overall floor plan (Fig. 13.9).

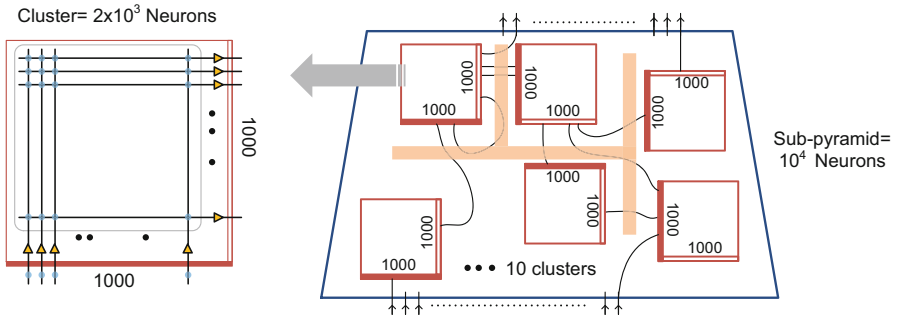


Fig. 13.9 Architecture of a sub-pyramid consisting of 1000 clusters and a total of 2×10^4 neurons. For simplicity, in each cluster $M = 1000, N = 1000, K = 0$

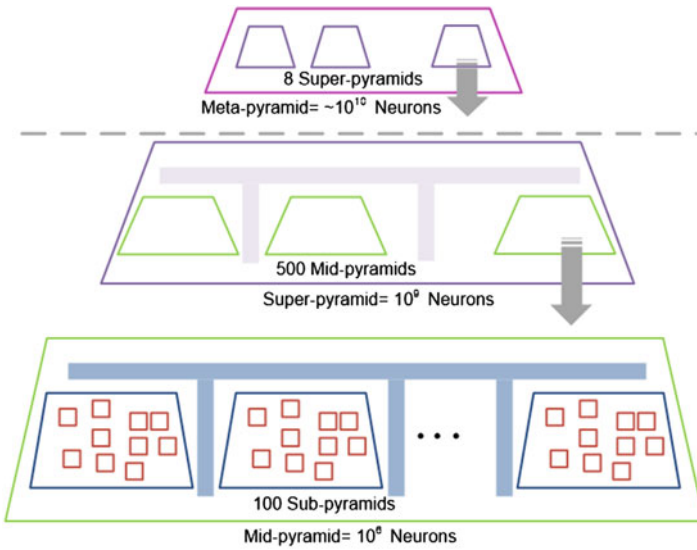


Fig. 13.10 A complete hierarchy of pyramids with expected number of neurons

The core element in this network is the clusters shown in Fig. 13.2. We assume a group of 2000 neurons are arranged within a three-layered (N, K and M neurons in input, hidden and output layers) topology with all-to-all connection between each layer. This connection could be created by an NVM (e.g. RRAM or PCM or STT-MRAM) matrix and physically placed right above the silicon neurons. The value of K (number of hidden layer neurons) that has to be chosen upfront could be anything between 0 and $\max\{M, N\}$. Since 2000 neurons in each cluster are divided between N, M and K , the worst-case situation in terms of global connectivity is when $N = M = 1000$ and $K = 0$. In this case, all the cluster will have 1000 input and 1000 outputs that can be connected to other clusters (Fig. 13.10).

A group of ten clusters would form a sub-pyramid with 2×10^4 neurons. This organization will have a pyramidal shape since it assumes that there will be fewer output neurons than input neurons. The connections (shown in orange in Fig. 13.10) between the clusters will consist of segmented buses. More clusters will be hierarchically organized in successive pyramidal structure where a mid-pyramid would contain 100 sub-pyramids and a super-pyramid with 500 sub-pyramids. Finally, a meta-pyramid may contain 8 or 16 super-pyramids taking the total number of neurons close to the desired level of 10^{10} .

Figure 13.10 shows the hierarchical pyramid structure in its most simplistic form. The number of output neurons in each of these pyramids is considered to be much smaller than the number of input neurons. Each pyramid will have its own segmented bus architecture with some limitation on the total number of simultaneous connections (i.e. P). It can be noticed that the number of internal blocks in a sub-pyramid and mid-pyramid and a super-pyramid has been progressively increased (from 10 to 100 to 500). These numbers are arbitrary, but the assumption is that the segmented bus inside a sub-pyramid will be more active (hence will require more switches) than that of the upper hierarchies. The need for the meta-pyramid will only arise when the bus width is more than one ($W > 1$; i.e. synaptic weights are non-binary). In this illustration, we assumed $W = 8$.

This virtual topological arrangement of the pyramidal structure doesn't fully show the physical structure behind. A possible physical organization is shown in Fig. 13.11. Here the silicon neurons are situated within the FEOL with local routing

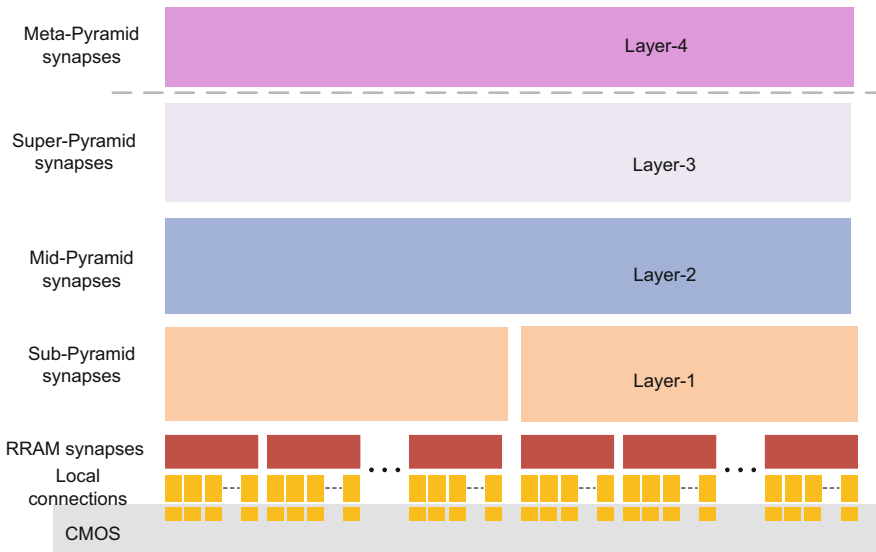


Fig. 13.11 Physical structure of the envisioned architecture. The dashed horizontal line suggests a potential chip boundary. In such a case, the meta-pyramid will be built among few ICs consisting of super-pyramids

restricted to a few bottommost metal layers (say $M1$ and $M2$). The RRAM synapse crossbar is placed right on top of the neurons, and the remaining metal layers are used for constructing the segmented bus in various pyramidal networks. It can be assumed that six to eight metal layers will be sufficient for each hierarchy of pyramid. As mentioned before, the switches of the segmented bus will be mostly in the BEOL fabric, and it can be envisioned that part of the controller architecture can also be accommodated in the BEOL.

Such an architecture in a scaled CMOS node (<25 nm) is quite feasible, and it is not even necessary to have the entire network on a single chip. One can envision a small PCB with many of these chips on it. Each chip will consist a few mid-pyramid or even a complete super-pyramid. Due to much lower number of connections in the segmented bus at these levels, going beyond the boundary of a single IC will not necessarily have a high impact on the power budget.

7 Ways to Reduce the Number of Buses and Switches

Here we formulate a method to reduce the size of the segmented bus whose switching activity would consume the bulk of the energy consumption. By exploiting an upper bound to the profile histogram of intercluster distance, less bandwidth allocation is required. However, in this way, it remains difficult or even impossible to exploit the detailed profiling info of the histogram. Hence, a worst-case upper-bound distance-based network would have to be allocated. So, still a needlessly high-energy overhead is expected in practical realizations due to long sequential ‘data pass’ sequences across the network-on-chip (NoC) links. That is also true for NoCs that exploit energy-optimized spatial time multiplexing as in [37] or in the NoCs which are used in other neuromorphic network proposals [17].

Instead we propose to further reduce the power and energy overhead in the following way. The maximum amount of simultaneously active connections can be obtained from the maximum in the profiled histogram. Typically, this upper-bounding can simply happen on the individual intercluster connections. However, this leads to a pessimistic bound because it ignores the cross-correlations that are present among these individual connections. So, we can obtain a tighter bound by taking into account the cross-correlation of the connection patterns (and an updated profiled histogram). That upper bound/maximum determines the parameter P in the proposed data communication network. However, given that this is still profiling-based and not fully deterministic, in order to provide some slack, it is possible to over-dimension this with a designer-based margin. Additional buses can be added in the segmented bus network to arrive at a total of R . In that case normally only P of them are needed, and a run-time decision is therefore to select P out of R that needs to be activated.

Let us now assume that a segmented bus is used in a sub-pyramid with less output ports than input (hence $G > H$, as shown in Fig. 13.5). Simply put, $P*(G + H)$ switches are necessary to make use of all possible connections. However, significant area saving can be obtained by not using a ‘full’ switch matrix topology on the

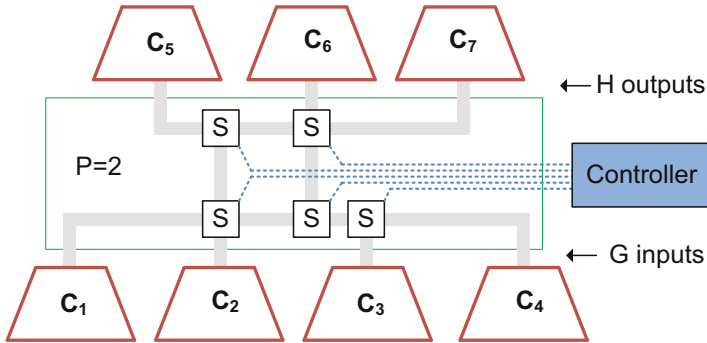


Fig. 13.12 A pruned embodiment of the dynamically controlled segmented bus network in Fig. 13.5

segmented bus. This can best be decided based on the simultaneously required intercluster connections which are needed to execute the typical applications. This can be derived again from the histogram of correlated connection patterns. When only the top $x\%$ (designer-defined) most occurring connection patterns are taken, not all of the $P \cdot G$ potential switch positions will have to be present. This is now illustrated with a simple example: the default switch topology of Fig. 13.5 is compared with the pruned topology of Fig. 13.12. Here three of the eight initial switches have been removed on the right-hand side and have been decided based on the histogram of correlated connection patterns. In this case it means that the direct connection from cluster C₇ to C₃ or to C₄ is not so active. Moreover, it also implies that the simultaneous connection from cluster C₆ to the C₄ is not sufficiently often required together with the cluster C₃ to C₅ (or C₃ to C₇, etc.) connection. When these more rarely occurring connection patterns would be present at run-time, it means that they have to wait for the next available time slot, and hence a latency is induced on these connections. Hence, the application running on the platform has to be able to tolerate this. If that is not the case for a subset of the connection patterns, then these latency-critical patterns have to be included by constraint on top of the $x\%$ of patterns to be kept.

More drastic saving on the energy consumption is however potentially reachable by activating less than P buses and much less segments and switches than the maximal amount $P \cdot G$ at power-up. Only a limited set of switches can be activated, and the other are left in full power-down mode which means that they cannot be quickly (i.e. in a few clock cycles) activated any more. This is especially important for the devices which are still required to be implemented in the FEOL layer, where leakage is expected to increase significantly for further scaled nodes. In addition, this also allows reducing dynamic energy significantly because of the earlier-discussed activity-based floor planning.

Figure 13.13 shows a 3D view of the pyramidal structure where the neural clusters are located in the FEOL. Apart from the crossbar synapses, all intra-cluster and feedback connections require a virtualization step (through the segmented

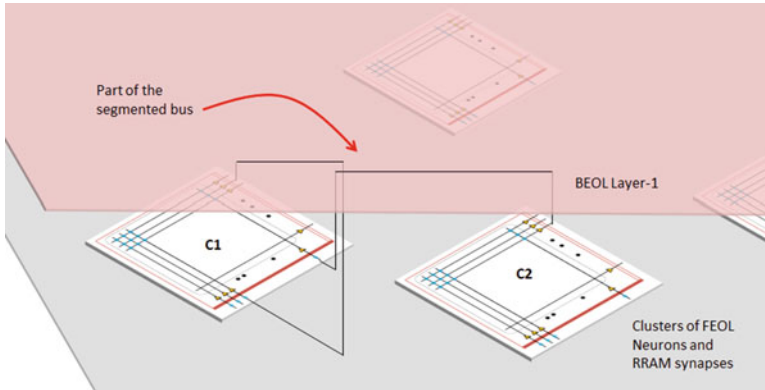


Fig. 13.13 A 3D view of the neural clusters and possible connections between them. All intra-cluster and feedback connections will be via BEOL switches in the segmented bus. In reality, all the neural clusters will be densely packed in Layer-0

BEOL bus) to form a real synaptic connection between two neurons. This figure demonstrates just two of the many possible connections (an output neuron in $C2$ connected to an input neuron in $C1$ and a feedback connection within $C1$).

8 Summary of the Network Control Plane

As shown in [38], the control plane of such a segmented bus has to be designed with care. One important aspect in the control plane is the cost of the potential storage for the switch positions. We believe an efficient implementation for this can be realized based on the distributed loop buffer concept which is described in [39] [patent EP1958059 B1] and initially intended for conventional instruction-set processor programming. This is a very energy-efficient solution to realize the look-up-table storing the (instruction) control bits for potentially huge amount of three- and four-way BEOL switches. In a multicore SoC, the platform already contains the hierarchy which is required to efficiently use the distributed loop buffer control. For neuromorphic synapse control, however, these concepts should be used in a re-instantiated form. It should also be combined with the suitable instruction bit compression/encoding schemes that are used in conventional microprocessors.

A typical three-way switch for a segmented bus is shown in Fig. 13.14. The control of the switches does not need to be updated (and hence rerouted) often, because it is expected that for long periods of time, the training/learning will not require to change the global synapse weights. That makes these switches ideally suited for a realization with the BEOL TFT devices. These switches are inherently ultra-low energy due to their negligible leakage (orders of magnitude lower than

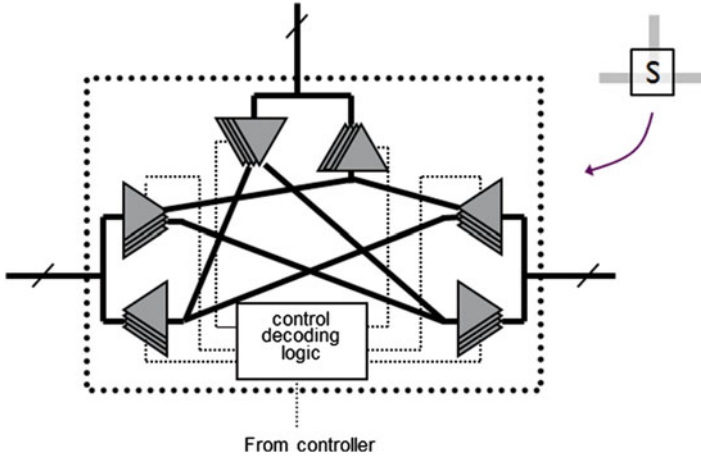


Fig. 13.14 Details of three-way TFT switches that can be used in the segmented bus. Similar four-way switches are also possible

CMOS devices) and their low dynamic power (due to BEOL fabrication). Their main limitation is the restricted clock frequency they can reach (10–100 lower than CMOS).

9 Illustration with Local and Global Synapse Networks

In order to visualize a much larger array of neural cluster and corresponding connections between them, let's start with a scaled-down illustration. Assume a group of neural clusters arranged in a way such that the segmented bus connecting them has ten input ports ($G = 10$) and eight output ports ($H = 8$), and only three of these needs to be simultaneously connected (i.e., $P = 3$). Even with such a restriction, the total number of possible switches in the segment bus will be $(10 + 8) * 3 = 54$. However, armed with the knowledge of the profiled histogram, we can assume that not all input cluster will need to be connected to any random output cluster. In fact, the probability of connection will decrease as the clusters are far apart, possibly in an exponential fashion [34]. Hence, some of these switches can be pruned out without affecting the final implementation. The diagram below shows an innermost region with three possible switches on a few vertical stubs (four stubs within dark blue region), with adjacent regions having two out of three switches on each stub (six stubs in light blue region) and an outer region with just one of three switches on each stub. This can be called as an onion-like layered architecture where the switches become less and less dense the farther away they are from the core. The total number of switches can now be counted as $4 * 3 + 6 * 2 + 8 * 1 = 32$. Though this doesn't seem to be a big difference in this simple case, a tenfold reduction (or

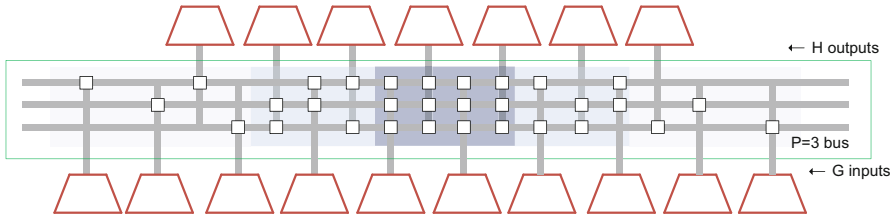


Fig. 13.15 A simplified illustration of a P -3 bus with switch pruning. The number of switches on the vertical stubs coming from the input and output ports follows an ‘onion’ structure, with fewer switches on the edge (controller not shown)

even more) in number of switches is possible in a large array of clusters. Since each switch requires a connection from the controller (not shown here) and corresponding memory bits, pruning will be extremely important as array size grows (Fig. 13.15).

Though the choice of the core and the size of the ‘onion’ requires specific application examples, we can expect that the most active switches will be present at the centre. It is possible to enable the centre to move dynamically by a virtualization layer which has to be supported in the control plane of this segmented bus. Further explanation of this scheme is beyond the scope of this chapter.

Now assume the neurons are organized in groups of 10 clusters, each with 1000 in/out neurons, further assuming that all the 10×1000 ($=10$ k) input neurons can potentially receive connections from all the 10 k output neurons. If we have a $P = 100$ segmented bus (that will allow 100 simultaneous connections) for this massive communication scheme, then there are $(10 \text{ k} + 10 \text{ k}) \times 100 = 2 \text{ M}$ possible switches. As before, $P = 100$ only signifies the lower limit of possible simultaneous connections in this bus. Like the above example, in practice one can prune the number of switches due to the correlated connection pattern information in the profile histogram. Here we can consider an innermost region with 1 k stubs having all possible 100 switches, then 2 k outer stubs with 30 switches, next 3 k stubs with 20 switches and outermost 4 k stubs with just 10 out of 100 switches. The results in $(100 \text{ k} + 60 \text{ k} + 60 \text{ k} + 40 \text{ k}) = 260 \text{ k}$ switches, an order of magnitude smaller than the original 2 M switches. The 260 k switches required for one group at layer 0 can be realized in TFT BEOL technology with a smaller area footprint, roughly the size of 64 k RRAM bit cells.

As shown in Fig. 13.10, the total amount of physically implemented neurons in this instantiation is eight billion, a number close to our target value of 10^{10} . This consists of 1000 input and output layer neurons per local array cluster and four million clusters. However, the number of physically implemented synapses is much lower than 10^{15} ! Each cluster has about one million synapses (1000×100 crossbar plus 1000 input synapses) leading to about 3×10^{12} in total, for the overall architecture. However, this many physical synapse can still support the equivalent of 10^{15} logical ones. As explained in Sect. 3, this highly desirable property is due to our virtualization approach where we exploit the characteristic that not all potentially usable parallel synapses in the neural connection graph are active simultaneously.

Because of the $P = 100$ assumptions at the different layers of the pyramid system architecture in our instantiation of Fig. 13.10, we have a significant reduction of the required physical synapses. Still, we must store a very large amount of potential control words to steer the virtualized control plane of the segmented bus networks. Fortunately, as hinted in Sect. 7, that can happen at a much-reduced update rate and these control words are stored in the most dense mass storage device technology can offer. Moreover, we envision several architecture innovations to compress these control words (with dynamic scenario clustering [40] and distributed loop buffer concepts [41]) so that the area and energy overhead can remain reasonable.

To that end, time multiplexing of hardware (synapse and neuron) resources could be of great benefit. This has not been exploited in the architecture shown in Fig. 13.10, with eight billion physically distinct neurons. If this could time multiplexed with a factor M , then both the physical neuron nodes and the segmented bus network hardware could be scaled with that factor. This will also reduce the physical control plane realization in complexity. The storage of the amount of synapse-related information remains similar, but it can be more heavily compressed now, and also the wire routing overhead is reduced.

Without any time multiplexing even with much scaled memristive arrays and BEOL TFT switch technology, it would also be extremely challenging to squeeze this entire pyramidal stack in a single chip. However, given the quite reasonable amount of $P = 100$ parallel buses connecting the super-pyramids, it should be relatively easy to realize the layer 2 segmented bus network across multiple chip stacks residing on a shared package. This is true even with 8–12 bits data, multiplexed over 100 pins (bit-serial communication), in and out for each bus connection from the previous or to the next super-pyramid. A possible chip boundary is shown with horizontal dashed line in Fig. 13.11.

Even before a complete single chip neuromorphic solution for large-scale classification is possible, the above solution can be realized in a combined board-level implementation with the local synapse solutions as shown in Fig. 13.16. Here the sensor is a separate IC that can be neuromorphic in nature (like silicon retina, cochlea, etc.), but doesn't necessarily need to be so, as long as the neural clusters receive the information in the right format. The neural clusters themselves can be divided in between a few ICs each having their own pyramids. Only the controller for the switches will be realized in the FEOL or can come from an off-chip microprocessor to enable flexible testing.

10 Conclusions

In summary, we propose an idea for a highly scaled neuromorphic platform to incorporate human brain size network. Energy and area optimization is targeted in all phases of the proposed design incorporating the impact of scaled process technology. This significantly improves intercluster communication energy consumption and area overhead, by extending already known principles to a (much) larger scale.

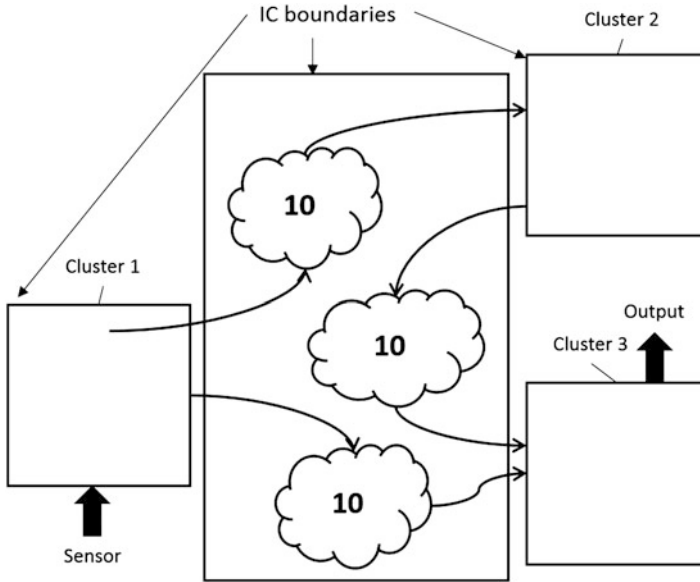


Fig. 13.16 Example of a board-level implementation illustrating the interaction between multiple local synapse arrays and global synapse network

These results should be reusable for different realizations of the global intercluster communication organization. The approach can most probably be used also for the training phase of the neural network, when the initial segmented bus template is somewhat over-dimensioned for the training, and then ‘restricted’ (power-down mode) in the energy-optimized trained execution phase.

To our knowledge, this is the first attempt to consider the theoretical possibility of a single system-in-package (SiP) battery-driven neuromorphic system of this scale (10^{10} neurons) that can be meaningfully connected with a very large number of synapses (10^{12} physical and 10^{15} logical). This will be needed for the embedded wearable compact and ultra-energy-efficient application targets that we envision here. In doing so, we took the liberty of various assumptions but always considering technological and biological realism. We envision that such hardware would be extremely useful in today’s renewed interest in neuromorphic application. Apart from the scale, the flexibility of the device would be ideal to bring down NRE cost for algorithm developers who would be interested in a generic neuromorphic platform.

Acknowledgement The authors would like to acknowledge the interesting discussions with their colleagues Rudy Lauwereins, Diederik Verkest, Soeren Steudel, Marc Van Bladel and Aneta Markova during the preparation of the material in this paper and also the results produced by the MSc students Francesco Dell’Anna, Ahmed Ammar, Ahmed Abdelmoneem, Thibaut Marty and Gagandeep Singh. Many of them have also contributed to some quantitative data in this material. We also acknowledge the support of the Horizon 2020 NeuRAM3 EC project.

References

1. C. Mead, *Analog VLSI and Neural Systems* (Addison-Wesley, Reading, 1989)
2. M. Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function* (California Institute of Technology, Pasadena, 1992)
3. R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gomez-Rodriguez, L. Camunas-Mesa, R. Berner, M. Rivas-Perez, T. Delbruck, S.C. Liu, R. Douglas, P. Haffiger, G. Jimenez-Moreno, A. Civit Ballcells, T. Serrano-Gotarredona, A.J. Acosta-Jimenez, B. Linares-Barranco, CAVIAR: A 45k neuron, 5M synapse, 12G connects/s AER hardware sensory-processing-learning-actuating system for high-speed visual object recognition and tracking. *IEEE Trans. Neural Netw.* **20**(9), 1417–1438 (2009)
4. S. Mitra, S. Fusi, G. Indiveri, Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI. *IEEE Trans. Biomed. Circuits Syst.* **3**(1), 32–42 (2009)
5. S.C. Liu, A. Van Schaik, B.A. Minch, T. Delbruck, Asynchronous binaural spatial audition sensor with 2²² channel output. *IEEE Trans. Biomed. Circuits Syst.* **8**(4), 453–464 (2014)
6. J. Hasler, B. Marr, Finding a roadmap to achieve large neuromorphic hardware systems. *Front. Neurosci.* **7**(7), 1–29 (2013)
7. S. Furber, Large-scale neuromorphic computing systems. *J. Neural Eng.* **13**(5), 51001 (2016)
8. B. Pakkenberg, D. Pelvig, L. Marnar, M.J. Bundgaard, H.J.G. Gundersen, J.R. Nyengaard, L. Regeur, Aging and the human neocortex. *Exp. Gerontol.* **38**(1–2), 95–99 (2003)
9. J. Hsu, IBM's new brain. *IEEE Spectr.* **51**(10), 17–19 (2014)
10. IBM, Lawrence Livermore National Laboratory and IBM Collaborate to Build Brain-Inspired Supercomputer, (2016), Available <http://www-03.ibm.com/press/us/en/pressrelease/49424.wss>. Accessed 25 Aug 2016
11. S. Scholze, H. Eisenreich, S. Hoppner, G. Ellguth, S. Henker, M. Ander, S. Hanzsche, J. Partzsch, C. Mayr, R. Schuffny, A 32 GBit/s communication SoC for a waferscale neuromorphic system. *Integr. VLSI J.* **45**(1), 61–75 (2012)
12. B.V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A.R. Chandrasekaran, J.M. Bussat, R. Alvarez-Icaza, J.V. Arthur, P.A. Merolla, K. Boahen, Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations. *Proc. IEEE* **102**(5), 699–716 (2014)
13. S.B. Furber, F. Galluppi, S. Temple, L.A. Plana, The SpiNNaker project. *Proc. IEEE* **102**(5), 652–665 (2014)
14. C. Eliasmith, T.C. Stewart, X. Choo, T. Bekolay, T. Dewolf, Y. Tang, D. Rasmussen, A large-scale model of the functioning brain. *Science* (80-.) **338**, 1202–1205 (2012)
15. T.M. Wong, R. Preissl, P. Datta, M.D. Flickner, R. Singh, S.K. Esser, E. McQuinn, R. Appuswamy, W.P. Risk, H.D. Simon, D.S. Modha, IBM internal Research Report 10 14. **10502**, 1–3 (2012)
16. P.A. Merolla, J.V. Arthur, R. Alvarez-Icaza, A.S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, D.S. Modha, A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* (80-.). **345**(6197), 668–673 (2014)
17. S. Moradi, N. Imam, R. Manohar, G. Indiveri, A memory-efficient routing method for large-scale spiking neural networks. *Eur. Conf. Circuit Theory Des.* **2013**, 1–4 (2013)
18. G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
19. T. Serrano-gotarredona, T. Prodromakis, B. Linares-Barranco, A proposal for hybrid memristor-CMOS spiking neuromorphic learning systems. *IEEE Circ. Syst. Magaz.* 74–88, 2nd quarter (2013)
20. S.H. Jo, T. Chang, I. Ebong, B.B. Bhadviya, P. Mazumder, W. Lu, Nanoscale memristor device as synapse in neuromorphic systems. *Nano Lett.* **10**(4), 1297–1301 (2010)
21. K.A. Boahen, Point-to-point connectivity between neuromorphic chips using address events. *IEEE Trans. Circuits Syst. II Analog Digit. Signal Process.* **47**(5), 416–434 (2000)

22. C. Gamrat, O. Bichler, D. Roclin, Memristive based device arrays combined with spike based coding can enable efficient implementations of embedded neuromorphic circuits. Tech. Dig. Int. Electron Devices Meet. IEDM **2016**, 4.5.1–4.5.7 (2016)
23. G. Piccolboni, G. Molas, J.M. Portal, R. Coquand, M. Bocquet, D. Garbin, E. Vianello, C. Carabasse, V. Delaye, C. Pellissier, T. Magis, C. Cagli, M. Gely, O. Cueto, D. Deleruyelle, G. Ghibaudo, B. De Salvo, L. Perniola, Investigation of the potentialities of vertical resistive RAM (VRRAM) for neuromorphic applications. Tech. Dig. Int. Electron Devices Meet. IEDM **2016**, 17.2.1–17.2.4 (2016)
24. G.W. Burr, P. Narayanan, R.M. Shelby, S. Sidler, I. Boybat, C. Di Nolfo, Y. Leblebici, Large-scale neural networks implemented with non-volatile memory as the synaptic weight element: Comparative performance analysis (accuracy, speed, and power). Tech. Dig. Int. Electron Devices Meet. IEDM **2016**(408), 4.4.1–4.4.4 (2016)
25. S. Kim, M. Ishii, S. Lewis, T. Perri, M. Brightsky, W. Kim, R. Jordan, G.W. Burr, N. Sosa, A. Ray, J. Han, C. Miller, K. Hosokawa, C. Lam, NVM Neuromorphic Core with 64k-cell (256-by-256) Phase Change Memory Synaptic Array with On-Chip Neuron Circuits for Continuous In-Situ Learning. (2015), pp. 443–446
26. D. Lee, J. Park, K. Moon, J. Jang, S. Park, M. Chu, J. Kim, J. Noh, M. Jeon, B.H. Lee, B. Lee, B.G. Lee, H. Hwang, Oxide based nanoscale analog synapse device for neural signal recognition system. Tech. Dig. - Int. Electron Devices Meet. IEDM **2016**, 4.7.1–4.7.4 (2016)
27. S.B. Eryilmaz, D. Kuzum, S. Yu, H.S.P. Wong, Device and system level design considerations for analog-non-volatile-memory based neuromorphic architectures. Tech. Dig. Int. Electron Devices Meet. IEDM **2016**, 4.1.1–4.1.4 (2016)
28. S. Yu, P.Y. Chen, Y. Cao, L. Xia, Y. Wang, H. Wu, Scaling-up resistive synaptic arrays for neuro-inspired architecture: Challenges and prospect. Tech. Dig. Int. Electron Devices Meet. IEDM **2016**, 17.3.1–17.3.4 (2016)
29. Y. Tang, J.R. Nyengaard, D.M.G. De Groot, H.J.G. Gundersen, Total regional and global number of synapses in the human brain neocortex. *Synapse* **41**(3), 258–273 (2001)
30. G. Indiveri, F. Corradi, N. Qiao, *Neuromorphic Architectures for Spiking Deep Neural Networks* (IEEE IEDM intl. conf., Washington DC, 2015), pp. 68–71
31. D. Vainbrand, R. Ginosar, Scalable network-on-chip architecture for configurable neural networks. *Microprocess. Microsyst.* **35**(2), 152–166 (2011)
32. K.K. Hidetomo Kobayashi, T. Ohmaru, S. Yoneda, Processor with 4.9-us Break-even Time in Power Gating Using Crystalline In-Ga-Zn-Oxide Transistor, in *Cool Chips Conference* (2013)
33. R. Perin, T.K. Berger, H. Markram, A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci. U. S. A.* **108**(13), 5419–5424 (2011)
34. R.B. Levy, A.D. Reyes, Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *J. Neurosci.* **32**(16), 5609–5619 (2012)
35. M. Beyeler, K.D. Carlson, T.S. Chou, N. Dutt, J.L. Krichmar, CARLsim 3: A user-friendly and highly optimized library for the creation of neurobiologically detailed spiking neural networks. *Proc. Int. Jt. Conf. Neural Netw.* **2015** (2015)
36. ARM, AMBA-lite. Available <https://www.arm.com/products/amba-open-specifications.php>. Accessed 13 Sept 2016
37. A. Leroy, D. Milojevic, D. Verkest, F. Robert, F. Catthoor, Concepts and implementation of spatial division multiplexing for guaranteed throughput in networks-on-chip. *IEEE Trans. Comput.* **57**(9), 1182–1195 (2008)
38. K. Heyrman, A. Papanikolaou, F. Gathoor, P. Veelaert, W. Philips, Control for power gating of wires. *IEEE Trans. Very Large Scale Integr. Syst.* **18**(9), 1287–1300 (2010)
39. F. Catthoor, P. Raghavan, A. Lambrechts, M. Jayapala, A. Kritikakou, J. Absar, *Ultra-low Energy Domain-Specific Instruction-set Processors* (Springer, Dordrecht, 2010)
40. S.V. Gheorghita, F. Vandeputte, K. De Bosschere, M. Palkovic, J. Hamers, A. Vandecappelle, S. Mamagkakis, T. Basten, L. Eeckhout, H. Corporaal, F. Catthoor, System-scenario-based design of dynamic embedded systems. *ACM Trans. Des. Autom. Electron. Syst.* **14**(1), 1–45 (2009)
41. M. Jayapala, F. Barat, T.V. der Aa, F. Catthoor, G. Deconinck, H. Corporaal, Clustered L0 buffer organisation for low energy embedded processors. *IEEE Trans. Comput.* **54**(6) (2005)

Erratum to: CMOS Nano-Pore Technology



Sina Parsnejad and Andrew J. Mason

Erratum to:
Chapter 1 in: S. Mitra, D.R.S. Cumming (eds.), *CMOS Circuits for Biological Sensing and Processing*,
https://doi.org/10.1007/978-3-319-67723-1_1

There was a mistake in the 20th reference in chapter 1. The reference with the correct journal name should read as below:

20. H. Li et al., Ultracompact microwatt CMOS current readout with picoampere noise and kilohertz bandwidth for biosensor arrays. *IEEE Trans Biomed Circuits Syst.* **12**(1), 35–46 (2018)

The updated online version of the chapter can be found under
https://doi.org/10.1007/978-3-319-67723-1_1

© Springer International Publishing AG 2018
S. Mitra, D.R.S. Cumming (eds.), *CMOS Circuits for Biological Sensing and Processing*, DOI 10.1007/978-3-319-67723-1_14

E1

Index

A

- Acoustic impedance, 188
- Active back-end-of-line network, 324–325
- Active electrode (AE) architectures, EEG recording
 - capacitively coupled chopper architecture, 174–175
 - capacitively coupled inverting architecture, 178
 - capacitive non-inverting architecture, 178–179
 - digitally assisted architecture, 181–183
 - digitally interfaced architecture, 179–180
 - noncontact sensing, 183
- Adenosine triphosphate (ATP), 136
- Amperometric electrochemical biosensors
 - nano-pores
 - array implementations, 10
 - Faraday's law, 6
 - input offset currents, 10
 - noise model, 7–9
 - readout circuit, 7
 - TIA, 7–10
 - role, 1
 - sensor array
 - bacteria/viruses counting, 93–96
 - circuit schematics, 85
 - high-density sensor array, 87, 88
 - transducers, 270
- Application-specific integrated circuits (ASICs)
 - CMUTs, 203
 - discrete components, 197
 - frequency requirements, 198–199
 - fully integrated system, 198

IP blocks, 200

MFCV

- human trials, 150, 152–156
 - MATLAB[®] analysis, 152–156
 - micro-photograph, 149
 - sEMG signal processing, 141, 142, 148
 - system block architecture, 147–148
 - system validation, 150
 - multiple gain stages, 199
 - neural probes
 - power budget, 221–223
 - program, 224
 - noise performance, 199
 - physical dimensions, 198
 - piezoelectric transducers, configurations of, 200–201
 - PMUTs, 203
 - power consumption, 199
 - PVDF transducers, 202
 - PZT transducer, 202
 - wired silicon probe recording, 286–288
- Application-specific standard product (ASSP)
 - integrating multimodal sensors, 83
- Array-enabled molecular mass sensing. *See* Multichannel nano-pore arrays
- ASICs. *See* Application-specific integrated circuits (ASICs)

B

- Back-end-of-line thin-film technology (BEOL TFT), 334, 335
- Basis pursuit (BP) algorithm, 292
- Basis pursuit de-noising (BPDN), 292
- Bead arrays, 32

- Bernoulli matrix, 291, 294
- Bilayer lipid membrane (BLM), 2–3, 16, 17
- Biocompatible packaging
chip deployment, 275
IMPACT project, 274, 275
materials, 273
- Biological signal processing. *See*
Neuromorphic computing
- Biomolecular sensing
enzymes, 25, 28
NMR system (*see* Miniature nuclear
magnetic resonance system)
- Biophotonics, 233
- Biosensors. *See also* Electrochemical
biosensors
microultrasound imaging
acoustic propagation, 188
attenuation parameters, 191–192
biomarkers, 190–191
chemical samples, 190
circuit and system approaches, 190
endoscopic imaging, 189
frequency, 191
lateral resolution, 188–189
multimodal sensor array
bacteria/viruses counting, 93–96
DNA single-base polymerization,
92–93
enzyme sensor with redox mediator,
89–93
glucose sensor, 91
- Bit-stream cross correlator
correlation stage, 145–146
MATLAB[®] simulations, 144–145
maximum detector stage, 146–147
proposed architecture, 143–144
- BLE. *See* Bluetooth low energy (BLE)
- BLM. *See* Bilayer lipid membrane (BLM)
- Bluetooth low energy (BLE), 278
- C**
- Capacitive micromachined ultrasonic
transducers (CMUTs), 201, 203
- Charge detection method, 89
- Chlamydomonas reinhardtii*, 236
- Clark electrode oxygen sensors
CMOS integration, 264
electrode materials, 262–263
electrolyte designs, 263
operation principle, 262
- CMFB. *See* Common-mode feedback circuit
(CMFB)
- CMOS. *See* Complementary metal-oxide
semiconductor (CMOS) integrated
circuits
- CMRR. *See* Common-mode rejection ratio
(CMRR)
- CMUTs. *See* Capacitive micromachined
ultrasonic transducers (CMUTs)
- Common-mode feedback (CMFB) technique,
110–111, 178
- Common-mode feedforward (CMFF)
technique, 178, 179
- Common-mode rejection ratio (CMRR), 173
- Complementary metal-oxide semiconductor
(CMOS) integrated circuits
active back-end-of-line network, 324–325
electrical detection (*see* Portable diagnostic
inspection system)
- metabolomics
capsule-based pH sensor, 35
foundry-processed pH sensing, 37–39
gas sensor, 35
micromachining method, 34
Moore's law, 33–34
photodiode setup, for cholesterol
quantification, 40, 41
photodiode setup, for protein detection,
40, 42
smart self-powering system, 35–36
- multimodal sensor array
amperometric sensor array, 85, 87,
93–96
bacteria/viruses counting, 93–96
biomolecule detections, 97
design principles, 78–79
DNA single-base polymerization,
92–93
enzyme sensor with redox mediator,
89–93
glucose sensor, 91
high-density sensor array, 87–88
impedimetric sensor array, 85, 87
photometric sensor, 80–81, 83
potentiometric sensor, 80, 83, 84, 87–89
- muscle fatigue (*see* Electromyography
(EMG) muscle fatigue analysis)
- nano-pores
amperometric sensing, 6–10
electrical model, 5–6
interface circuits, 10–15
multichannel arrays, 16–18
- neural probes
basic architecture, 215
goal, 214–215

- implantable system, architecture of, 216–217
 - microsystem architectures, 217–218
 - power budget, 221–224
 - programming, 224–225
 - signal quality, 219–221
 - silicon neural probes, 216
 - 66 electrodes and 384 parallel recording channels, 226–227
 - system requirements, 219
 - tetrode capturing signals, 214
 - time-division multiplexing architecture, 227–229
 - NMR system (*see* Miniature nuclear magnetic resonance system)
 - Compressive sensing, neural recording devices
 - multi-electrodes compression
 - join-group sparsity, 300, 305–307
 - multi-channel reconstruction, 307, 309, 310
 - signal model, 300–302
 - single-channel reconstruction, 307, 308
 - spike sorting, 306
 - unsupervised dictionary learning, 302–305
 - neural signal compression
 - analog implementation, 295
 - digital implementation, 295
 - reconstruction framework, 295–299
 - sampling framework, 294–295
 - signal agnostic dictionary, 296
 - signal dependent dictionary, 296–299
 - RIP, 289
 - Conductometry, 270
 - Contact printing, 29–30
 - Current-mode circuits, 79
 - Cyclic voltammetry (CV), 270
- D**
- Deep belief network, 316
 - Deep neural networks, 313
 - Deep reactive ion etching (DRIE), 53
 - Design-time connection profiling, 325–326
 - Digital microfluidics (DMF) device
 - applications, 116
 - droplet actuation, 117–119
 - fabrication, 117–119
 - magnet and coil codesign, 116–117
 - NMR experiments, 119–121
 - DNA single-base polymerization, 92–93
 - Double-stranded DNA (dsDNA) microarrays, 29
 - DRIE. *See* Deep reactive ion etching (DRIE)
 - Dual-side single-photon avalanche photodiode sensor
 - array characterization, 65–67
 - cross-inter-arrival time histogram, 67, 70
 - DCR distribution histograms, 66, 68
 - dual-side imaging, 67–69, 71
 - electrical and optical performance, 72, 74
 - fabrication flowchart, 58, 59, 64–65
 - functional architecture, 62–64
 - PDP, 60, 62, 66, 69
 - pixel farm sample, 60, 61
 - pixel structure, 57, 58
 - resolution enhancement, 69, 72, 73
 - second metallization, 64, 66
 - SOI body doping profile, 60, 61
 - substrate etching, 65, 66
 - timing jitter measurements, 60, 62
 - wire bonding, 65, 67
 - Dynamically controlled segmented bus network, 326–327
- E**
- Electrochemical biosensors, 77–78
 - foundry-processed CMOS, 37
 - implantable microsystems
 - CMOS-integration, 272
 - molecular recognition element, 268–270
 - protease detection, 271–272
 - transducers, 270
 - nanopores, 1–2
 - Electrochemical impedance spectroscopy (EIS), 270
 - Electrochemical transducers, 270
 - Electroencephalography (EEG) recording, 212
 - AE architectures
 - capacitively coupled chopper architecture, 174–175
 - capacitively coupled inverting architecture, 178
 - capacitive non-inverting architecture, 178–179
 - digitally assisted architecture, 181–183
 - digitally interfaced architecture, 179–180
 - noncontact sensing, 183
 - applications, 163
 - brain-computer interface study, 164
 - clinical use cases, 163
 - components, 166
 - electrical specifications, 170–171
 - CMRR, 173
 - electrode offset tolerance, 173

- Electroencephalography (EEG) recording
(*cont.*)
input impedance, 173
noise contribution, 172–173
number of wires, 174
power dissipation and supply voltage,
172
electrodes, 167–168
evolution, 164, 165
IA architectures
capacitively coupled chopper
architecture, 174–175
current-balancing architecture, 176–177
digitally assisted architecture, 177
neurofeedback records, 164
readout circuits
active electrode acquisition, 170
illustration, 169
passive electrode acquisition, 169
requirements, 168–169
simplified diagram, 166
state-of-the-art devices, 165
Electroless plating, 87–88
Electromyography (EMG) muscle fatigue
analysis
biofeedback cycle, 134
ergonomics, 134
reliable metrics, 133
Electrophysiology recordings, 285
Electrowetting-on-dielectrics (EWOD), 116
Enzyme printed microarray, 30–31
Enzyme reactions
free energy profile, 26
lock and key model, 25
Michaelis-Menten equation, 27–28
Enzyme sensor with redox mediator, 89–91
- F**
Faraday's law, 6
Field-programmable gate arrays (FPGAs), 118,
195
Flexible CMOS single-photon avalanche diode
image sensor
dual-side single-photon image sensor
array characterization, 65–67
cross-inter-arrival time histogram, 67,
70
DCR distribution histograms, 66, 68
dual-side imaging, 67–69, 71
electrical and optical performance, 72,
74
fabrication flowchart, 58, 59, 64–65
functional architecture, 62–64
PDP, 60, 62, 66, 69
pixel farm sample, 60, 61
pixel structure, 57, 58
resolution enhancement, 69, 72, 73
second metallization, 64, 66
SOI body doping profile, 60, 61
substrate etching, 65, 66
timing jitter measurements, 60, 62
wire bonding, 65, 67
ultrathin-body single-photon image sensor
DCR density vs. PDP, 56–57
dual-side illumination, 55, 56
flexible electronics solutions, 49, 50
Geiger-mode performances, 56
Frequency-shift keying (FSK), 278
- G**
Gaussian matrix, 291
GeneChips, 31, 38
GenSight Biologics (France), 253
Gilbert circuit, 85, 86
Global synapse networks, 333–336
Glucose sensor, 91
G-protein-coupled receptor (GPCR), 237
Green fluorescent protein (GFP), 237
- H**
High-density bead arrays, 32
High-density CMOS neural probes
basic architecture, 215
goal, 214–215
implantable system, architecture of,
216–217
microsystem architectures, 217–218
power budget, 221–224
programming, 224–225
signal quality, 219–221
silicon neural probes, 216
66 electrodes and 384 parallel recording
channels, 226–227
system requirements, 219
tetrode capturing signals, 214
time-division multiplexing architecture,
227–229
High density neural recording devices
compressive sensing
multi-electrodes compression, 300–310
neural signal compression, 294–299
RIP, 289
dictionary learning
K-SVD algorithm, 293–294
optimization problem, 292–293

limitations, 285–286
 power consumption
 wired transmission, 286–288
 wireless transmission, 288–289
 sampling and sensing matrix
 Bernoulli matrix, 291
 Gaussian matrix, 291
 sparse signal recovery, 292
 sparse representation, 290–291
 Human brain-scale architecture, 314–315
 Human Genome Project, 32

I

IEEE Standard for Safety Levels with Respect to Human Exposure to RF-EM fields, 276
 Illumina bead array, 32, 33
 Impedimetric sensor array
 circuit schematics, 85
 high-density sensor array, 87, 88
 Implantable microsystems
 electrochemical biosensors
 CMOS-integration, 272
 molecular recognition element, 268–270
 protease detection, 271–272
 transducers, 270
 miniaturised sensors
 Clark electrode oxygen sensors, 262–264
 ISFET pH sensors, 264–268
 packaging
 chip deployment, 275
 IMPACT project, 274, 275
 materials, 273
 wireless power transfer
 BLE device, 278
 FSK, 278
 magnetic resonance, 276–277
 Implantable Microsystems for Personalised Anti-Cancer Therapy (IMPACT) project, 258, 261, 274, 275
 Input referred noise (IRN), 110–111
 In situ synthesised microarrays, 31–32
 Instrumentation amplifier (IA) architectures, EEG recording
 capacitively coupled chopper architecture, 174–175
 current-balancing architecture, 176–177
 digitally assisted architecture, 177
 Integrated arrays, 16
 International Agency for Research on Cancer (IARC), 258

International Commission on Non-Ionizing Radiation Protection (ICNIRP)
 guidelines, 276
 Intracellular recordings, 212
 Ion channels, 2
 Ion-sensitive field-effect transistor (ISFET)
 foundry-processed CMOS, 37–40
 pH sensors
 challenges and solutions, 267
 fabrication, 266
 instrumentation, 267–268
 structure and function, 264–266
 system-on-chip process, 35
 Ion Torrent genome sequencer, 32, 38, 39, 42
 IRN. *See* Input referred noise (IRN)
 ISFET. *See* Ion-sensitive field-effect transistor (ISFET)

K

KLM model, 192–193
 K-SVD dictionary learning algorithm, 293–294

L

Lead zirconate titanate (PZT) transducer, 202
 Light-emissive array circuit architecture, 244–247
 Light-emitting diodes (LEDs), 238, 240–241
 LNA. *See* Low-noise amplifier (LNA)
 Localised muscle fatigue analysis, 133–134
 Local parallel global sequential (LPGS) solution, 318, 319, 321
 Local sequential global parallel (LSGP) solution, 318, 319
 Local synapse networks, 333–336
 Lock and key model, 25
 Low-noise amplifier (LNA), 109–111

M

Magnetic nanoparticles (MNPs), 105
 Magnetic resonance wireless power transfer (MRWPT) system
 challenges, 276–277
 inductive power transfer, 276
 schematics, 276, 277
 Mason's model, 192
 Massively parallel DNA sequencing, 32, 40
 Maximum voluntary contractions (MVCs), 150
 Mean absolute relative difference (MARD) metrics, 142
 Medical US. *See* Microultrasound (μ US) imaging

- Metabolomics
 - analytical platforms, 24–25
 - CMOS technology
 - capsule-based pH sensor, 35
 - foundry-processed pH sensing, 37–39
 - gas sensor, 35
 - micromachining method, 34
 - Moore's law, 33–34
 - photodiode setup, for cholesterol quantification, 40, 41
 - photodiode setup, for protein detection, 40, 42
 - smart self-powering system, 35–36
 - enzymes (*see* Enzyme reactions)
 - microarray technologies
 - bead arrays, 32
 - DNA, 28–29
 - enzymes, 29
 - high-throughput sequencing, 28
 - printing, 29–31
 - in situ synthesised microarrays, 31–32
- MFCV. *See* Muscle fibre conduction velocity (MFCV)
- Michaelis-Menten model, 27–28
- Microarray technologies
 - bead arrays, 32
 - DNA, 28–29
 - enzymes, 29
 - high-throughput sequencing, 28
 - printing, 29–31
 - in situ synthesised microarrays, 31–32
- Micro-LED driving circuits, 242–244
- Micromachining method, 34
- Microultrasound (μ US) imaging
 - biosensing
 - acoustic propagation, 188
 - attenuation parameters, 191–192
 - biomarkers, 190–191
 - chemical samples, 190
 - circuit and system approaches, 190
 - endoscopic imaging, 189
 - frequency, 191
 - lateral resolution, 188–189
 - continuous wave and contact measurement modes, 204–205
 - electronic systems
 - chipset implementation, 195–196
 - clinical US system, block diagram of, 193
 - frequencies, 196–197
 - KLM model, 192–193
 - Mason's model, 192
 - overall system architecture, 192–195
 - receiving (RX) circuits, 194–195
 - transmitting (TX) circuits, 193–194
- integrated circuits
 - CMUTs, 203
 - discrete components, 197
 - frequency requirements, 198–199
 - fully integrated system, 198
 - IP blocks, 200
 - multiple gain stages, 199
 - noise performance, 199
 - physical dimensions, 198
 - piezoelectric transducers, configurations of, 200–201
 - PMUTs, 203
 - power consumption, 199
 - PVDF transducers, 202
 - PZT transducer, 202
- miniaturised electronics, 204–205
- minimally invasive medical devices, 204
- Miniature nuclear magnetic resonance system
 - B_0 -field calibration
 - chip prototype platform, 125–126
 - Hall sensors, 122–125
 - off-chip current driver, 122
 - readout circuit, 122, 124, 125
 - Wheatstone bridge, 124
 - coil miniaturization, 107
 - digital pulse generator, 111–112
 - DMF device
 - applications, 116
 - droplet actuation, 117–119
 - fabrication, 117–119
 - magnet and coil codesign, 116–117
 - NMR experiments, 119–121
 - electronic-automated sample management, 115–121
 - ^1H proton magnetic moments, 102–104
 - human bladder cancer cell detection, 114, 115
 - key components, 102
 - magnet miniaturization, 106
 - MNPs, 105
 - physics, 103
 - portable magnet, 121–126
 - prototype, 108, 109
 - Rabi oscillation, 104
 - vs. traditional NMR instruments, 101
 - transceiver (TRX)
 - architecture, 109–110
 - LNA, 110–111
 - miniaturization, 107–108
 - noise matching, 110–111
 - "1-chip" NMR system, 112–115
 - water detection, 114, 115

- Miniaturised sensors
 - Clark electrode oxygen sensors, 262–264
 - ISFET pH sensors, 264–268
 - Minimally invasive US imaging, 189, 204
 - MNPs. *See* Magnetic nanoparticles (MNPs)
 - Moore’s law, 33–34
 - More-than-Moore technologies, 33
 - MOSFET sensors, 37, 84, 85
 - Motor unit action potentials (MUAPs), 136, 137, 139
 - MRWPT system. *See* Magnetic resonance wireless power transfer (MRWPT) system
 - Multichannel nano-pore arrays, 16–18
 - Multi-electrodes compressive sensing
 - join-group sparsity, 300, 305–307
 - multi-channel reconstruction, 307, 309, 310
 - signal model, 300–302
 - single-channel reconstruction, 307, 308
 - spike sorting, 306
 - unsupervised dictionary learning, 302–305
 - Multimodal sensor array
 - amperometric sensor array
 - bacteria/viruses counting, 93–96
 - circuit schematics, 85
 - high-density sensor array, 87, 88
 - biomolecule detections, 97
 - design principles, 78–79
 - DNA single-base polymerization, 92–93
 - enzyme sensor with redox mediator, 89–93
 - glucose sensor, 91
 - impedimetric sensor array
 - circuit schematics, 85
 - high-density sensor array, 87, 88
 - photometric sensor
 - circuit schematics, 81
 - 64 current-mode ADCs, 83
 - wide-swing cascode current mirror, 80–81
 - potentiometric sensor
 - cascode source-drain follower, 79, 80
 - 64 current-mode ADCs, 83
 - high-density sensor array, 87–88
 - principles, 89
 - voltage to current converter, 84
 - Muscle fatigue
 - analysis types, 133–134
 - EMG (*see* Electromyography (EMG) muscle fatigue analysis)
 - Muscle fibre conduction velocity (MFCV)
 - advantage, 139
 - ASIC
 - human trials, 150, 152–156
 - micro-photograph, 149
 - sEMG signal processing, 137, 138, 148
 - system block architecture, 147–148
 - system validation, 150
 - cross-correlation
 - bit-stream cross correlator, 143–147
 - CMOS cross correlator, 141–143
 - sEMG detection points, 139–140
 - mathematical model, 139
 - wearable device, 156–158
 - Mutual coherence, 291
- N**
- Nano-pores
 - amperometric sensing
 - array implementations, 10
 - Faraday’s law, 6
 - input offset currents, 10
 - noise model, 7–9
 - readout circuit, 7
 - TIA, 7–10
 - challenges, 4
 - electrical model, 5–6
 - implementations, 2–4
 - interface circuits
 - capacitive feedback approach, 11–12
 - enhanced capacitive feedback, 12–13
 - fully active feedback control, 13–14
 - offset cancelation, 12–15
 - op-amp performance, 15
 - TIA noise performance, 10–11
 - ion channels (*see* Ion channels)
 - operation principles, 2
 - solid-state nano-pores, 3–4, 6
 - types, 2
 - Neural network algorithms, 316
 - Neural recording probes. *See* High-density CMOS neural probes
 - Neural signal compressive sensing
 - analog implementation, 295
 - digital implementation, 295
 - reconstruction framework, 295–299
 - sampling framework, 294–295
 - signal agnostic dictionary, 296
 - signal dependent dictionary, 296–299
 - Neuroelectronic interfaces, drawbacks of, 234
 - Neuromorphic computing
 - global synapse networks, 333–336
 - human brain-scale architecture, 314–315
 - local synapse networks, 333–336
 - non-CMOS devices, 314
 - scalable synapse networks
 - data communication network design, 321–323

- Neuromorphic computing (*cont.*)
 dense local synapse array, 318
 dynamic full connectivity, 320–321
 full connectivity, 319–321
 network data plane, 325–333
 parallel buses determination, 323
 predefined full connectivity, 320
 restricted connectivity, 318, 319
 SoCs, 315–317
- Neuron, operation of, 234
- Neuroprosthetics
 retinitis pigmentosa, 251–254
 visual brain prosthesis, 253–255
- NimbleGen, 32
- Non-contact printing, 29–30
- Nuclear magnetic resonance (NMR). *See*
 Miniature nuclear magnetic
 resonance system
- O**
- Oligonucleotide DNA microarrays, 30
- On-chip NMR sensing system, 112–115
- “One drug fits all,” 23
- Operation in subthreshold region, 79
- Optogenetics
 cell photosensitization, 235–236
 domains, 235
 electronics
 light-emissive array circuit architecture,
 244–247
 micro-LED driving circuits, 242–244
 penetrating optrode circuit architecture,
 247–250
 fluorescent imaging, 237–238
 GPCR, 237
 ion channels, 236
 ion pumps, 236–237
 journal publications, 252
 neural tissue
 LEDs, 238, 240–241
 light scattering, 238–240
 optrodes, 241–242
 two-photon approaches, 238
 neuroprosthetics
 retinitis pigmentosa, 251–254
 visual brain prosthesis, 253–255
- Organic light-emitting diodes (OLEDs), 240
- Orthogonal matching pursuit (OMP), 292
- Penetrating optrode circuit architecture,
 247–250
- Peptide-based electrochemical sensor, 271–272
- Personalised anticancer therapy. *See*
 Implantable microsystems
- Personalised medicine
 vs. conventional medicine, 23–24
 individual patient’s genome, 23–24
 metabolome (*See* Metabolomics)
- Photometric sensor
 circuit schematics, 81
 64 current-mode ADCs, 83
 wide-swing cascode current mirror, 80–81
- Photon-counting biomedical imaging
 applications, 47–49
 flexible CMOS
 dual-side single-photon image sensor,
 57–74
 ultrathin-body single-photon image
 sensor, 49–57
 SOI fabrication, 50–53
 substrate transfer process, 53–55
- Piezoelectric micromachined ultrasonic
 transducers (PMUTs), 201, 203
- PMUTs. *See* Piezoelectric micromachined
 ultrasonic transducers (PMUTs)
- Polyvinylidene fluoride (PVDF) transducers,
 202
- Portable diagnostic inspection system
 advantages, 77
 applications, 78
 stand-alone portable system, 96–97
- Positron emission tomography (PET), 257
- Post-CMOS process, 79, 86, 87
- Potentiometric sensor
 cascode source-drain follower, 79, 80
 64 current-mode ADCs, 83
 high-density sensor array, 87–88
 principles, 89
 voltage to current converter, 84
- Precision healthcare. *See* Personalised
 medicine
- Printed microarrays
 alginate gel enzyme entrapment, 31
 contact printing vs. non-contact printing,
 29–30
 disadvantage, 29
 DNA microarrays, 29–30
- Protein in situ array (PISA), 32
- P**
- Packaging. *See* Biocompatible packaging
- PCI Express standard (PCIe), 315
- R**
- Rabi oscillation, 104
- Radiotherapy, 257, 258

Random matrices, 291
 Redox potential detection method, 89–93
 Restricted Boltzmann machine, 316
 Restricted isometry property (RIP), 289
 Retina Implant AG (Germany), 251
 Retinal ganglion cells, 233–234
 Retinal prosthesis, 251–254
 RetroSense (USA), 253

S

Scalable synapse neuromorphic computing
 data communication network design,
 321–323
 dense local synapse array, 318
 dynamic full connectivity, 320–321
 full connectivity, 319–321
 network data plane
 buses and switches reduction, 330–332
 design-time connection profiling,
 325–326
 hierarchical floor plan, 327–330
 run-time configuration, 326–327
 three-way TFT switches, 332–333
 parallel buses determination, 323
 predefined full connectivity, 320
 restricted connectivity, 318, 319
 Second Sight (USA), 251
 Sensor system-on-chip (SSOC), 35, 37, 40
 Shared op-amp structure circuitry, 17–18
 Silicon neural probes, 216
 Single-photon avalanche photodiode (SPAD)
 dual-side single-photon image sensor,
 57–74
 ultrathin-body single-photon image sensor,
 49–57
 SOC. *See* System-on-chip (SOC)
 Solid-state nano-pores, 3–4, 6
 SSOC. *See* Sensor system-on-chip (SSOC)
 Stand-alone portable diagnostic inspection
 system, 96–97
 Stereotactic body radiotherapy (SBRT),
 258
 Surface Electromyography for Noninvasive
 Assessment of Muscles (SENIAM),
 137
 System-in-package (SiP) battery-driven
 neuromorphic system, 336
 System-on-chip (SOC), 34, 315–317

T

TIA. *See* Transimpedance amplifier (TIA);
 Trans-impedance amplifier (TIA)
 Time-correlated single-photon counting
 (TCSPC), 48
 Time-domain signal representation, 79
 Time-multiplexed neural probe, 227–229
 Transimpedance amplifier (TIA), 7, 124, 125
 TrueNorth chip project, 319, 320
 Tumour hypoxia
 disrupted blood supply, 259–260
 Eppendorf oxygen electrode, 260
 intensity-modulated radiotherapy, 261

U

Ultrasound (US) imaging. *See* Microultrasound
 (μ US) imaging
 Ultrathin-body single-photon avalanche
 photodiode sensor
 DCR density vs. PDP, 56–57
 dual-side illumination, 55, 56
 flexible electronics solutions, 49, 50
 Geiger-mode performances, 56
 Unsupervised dictionary learning, 302–305

V

Vermon, 204
 Vertical Hall sensor (VHS), 123–124
 Very large-scale neuromorphic systems. *See*
 Neuromorphic computing
 VHS. *See* Vertical Hall sensor (VHS)
 Visual brain prosthesis, 253–255
 VisualSonics, 204
Volvox carteri organism, 236
 von-Neumann computing architecture, 314

W

Wearable muscle fatigue tracking system,
 156–158
 Wired silicon probe recording, 286–288
 Wireless multi-electrode array transmission,
 288–289
 Wireless power transfer (WPT), microsystems
 BLE device, 278
 FSK, 278
 magnetic resonance, 276–277