Ultra Low Power Bioelectronics

This book provides, for the first time, a broad and deep treatment of the fields of both ultra low power electronics and bioelectronics. It discusses fundamental principles and circuits for ultra low power electronic design and their applications in biomedical systems. It also discusses how ultra energy-efficient cellular and neural systems in biology can inspire revolutionary low power architectures in mixed-signal and RF electronics.

The book presents a unique, unifying view of ultra low power analog and digital electronics and emphasizes the use of the ultra energy-efficient subthreshold regime of transistor operation in both. Chapters on batteries, energy harvesting, and the future of energy provide an understanding of fundamental relationships between energy use and energy generation at small scales and at large scales. A wealth of insights and examples from brain implants, cochlear implants, biomolecular sensing, cardiac devices, and bio-inspired systems make the book useful and engaging for students and practicing engineers.

Rahul Sarpeshkar leads a research group on Bioelectronics at the Massachusetts Institute of Technology (MIT), where he has been a professor since 1999. This book is based on material from a course that Professor Sarpeshkar has taught at MIT for 10 years, where he has received both the Junior Bose Award and the Ruth and Joel Spira Award for excellence in teaching. He has won several awards for his interdisciplinary bioengineering research including the Packard Fellow Award given to outstanding faculty.

Ultra Low Power Bioelectronics

Fundamentals, Biomedical Applications, and Bio-inspired Systems

RAHUL SARPESHKAR
Massachusetts Institute of Technology

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521857277

© Cambridge University Press 2010

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2010

Printed in the United States of America

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication Data

Sarpeshkar, Rahul
Ultra low power bioelectronics : fundamentals, biomedical applications, and bio-inspired systems / Rahul Sarpeshkar.
p. cm.
ISBN 978-0-521-85727-7 (hardback)
1. Low power systems. 2. Biomimicry. 3. Bioelectronics. 4. Bionics. 5. Medical electronics. I. Title.
TK7881.15.S27 2010
621.381–dc22
2009047509

ISBN 978-0-521-85727-7 Hardback

Additional resources for this publication: www.cambridge.org/9780521857277

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Contents

Section I

Foundations

1 The big picture

It is the harmony of the diverse parts, their symmetry, their happy balance; in a word it is all that introduces order, all that gives unity, that permits us to see clearly and to comprehend at once both the ensemble and the details.

It is through science that we prove, but through intuition that we discover.

Henri Poincaré

This book, *Ultra Low Power Bioelectronics*, is about ultra-low-power electronics, bioelectronics, and the synergy between these two fields. On the one hand it discusses how to architect robust ultra-low-power electronics with applications in implantable, noninvasive, wireless, sensing, and stimulating biomedical systems. On the other hand, it discusses how bio-inspired architectures from neurobiology and cell biology can revolutionize low-power, mixed-signal, and radio-frequency (RF) electronics design. The first ten chapters span feedback systems, transistor device physics, noise, and circuit-analysis techniques to provide a foundation upon which the book builds. Chapters that describe ultra-low-power building-block circuits that are useful in biomedical electronics expand on this foundational material, followed by chapters that describe the utilization of these circuits in implantable (invasive) and noninvasive medical systems. Some of these systems include cochlear implants for the deaf, brain implants for the blind and paralyzed, cardiac devices for non-invasive medical monitoring, and biomolecular sensing systems. Chapters that discuss fundamental principles for ultra-low-power digital, analog, and mixed-signal design unify and integrate common themes woven throughout the book. These principles for ultra-low-power design naturally progress to a discussion of systems that exemplify these principles most strongly, namely biological systems. Biological architectures contain many noisy, imprecise, and unreliable analog devices that collectively interact through analog and digital signals to solve complex tasks in real time, with precision, and with astoundingly low power. We provide examples of how bio-inspired systems, which mimic architectures in neurobiology and cell biology, lead to novel systems that operate at high speed and with low power. Finally, chapters on batteries, energy harvesting, and the future of energy discuss tradeoffs between energy density and power density, which are essential in architecting an overall low-power system, both at small scales and at large scales.

The book can serve as a text for senior or graduate students or as a reference for practicing engineers in the fields of

- Ultra-low-power Electronics: Chapters 1 through 22, 25, and 26.
- Biomedical Electronics: Chapters 1 through 22, 25, and 26.
- Bio-inspired Electronics: Chapters 1 through 18, 21 through 26.
- Analog and Mixed-Signal Electronics: Chapters 1 through 24.

In this busy day and age, many people with an interest in these fields may not have the time to read a whole book, especially one of this size. Therefore, the book has been written so that a reader interested in only a chapter or two can read the chapter and delve deeper if he/she would like. There is a slight amount of redundancy in each chapter to enable such sampling, with interconnections among the various chapters outlined throughout every chapter. The index should also be useful in this regard. Every reader should read Chapter 1 (this chapter). Chapter 2 on the fundamentals of feedback is also essential for a deeper understanding of many chapters. Chapters 1 through 10 provide a firm foundation, necessary for a deep understanding of the whole book.

Throughout this book, intuitive, geometric, and physical thinking are emphasized over formal, algebraic, and symbolic thinking. Physical intuition is extremely important in getting systems to work in the world since they do not always behave like they do in simulations or as the mathematical idealizations suggest they do. When the mathematics becomes intractable, usually the case in all but the simplest linear and idealized systems, intuitive and physical thinking can still yield powerful insights about a problem, insights that allow one to build high-performance circuits. Practice in physical thinking can lead to a lightning-fast understanding of a new circuit that lots of tedious algebra simply can never provide. Nevertheless, one must attempt to be as quantitative as possible for a deep understanding of any system and for theory and experiment to agree well. Thus, the book does not aim to substitute qualitative understanding for quantitative understanding; rather it attempts to maximize insight and minimize algebraic manipulations. We will always aim to look at problems in a physically insightful and original way such that the answer is intuitive and can be obtained exactly and quickly because the picture in our heads is clear.

Feedback is so fundamental to a deep understanding of how circuits work and how biology works that we shall begin this book with a review of feedback systems in Chapter 2. We shall see in this chapter that feedback is ubiquitous in physical, chemical, biological, and engineering systems even though the importance of feedback has been largely unappreciated. Throughout the book, we shall draw on our knowledge of feedback systems to derive or interpret results in a simple way that would not be possible without the use of this knowledge. For example, our discussion of physics in an MOS transistor will often use feedback analogies to understand the physics of their operation intuitively in Chapters 3 and 4. The equations of electron velocity saturation in an MOS transistor will be represented as a feedback loop in Chapter 6. We shall often avoid tedious Kirchoff's current law algebraic equations by simply drawing a feedback loop to provide all the answers for any transfer function, noise or offset analysis, robustness analysis, or dynamic analysis that we may need. We shall use feedback interpretations to understand how the noise in a transistor is affected by internal feedback within it. A deep understanding of feedback and circuits can enable a unified understanding of several systems in the world.

In both biomedical and bio-inspired electronics, it is important to deeply understand the biology. To understand and mimic biological systems in this book, we shall use circuits as a primary language rather than mathematics. Several nonlinear partial differential equations and structures in biology then translate into simple intuitive, lumped or distributed circuits. For example, we use such circuits to mimic the inner ear or cochlea, to understand the retina in the eye, to understand and mimic the heart, to mimic the vocal tract, to mimic spiking (pulsatile) neurons in the brain, and to understand and mimic biochemical gene–protein and protein–protein molecular networks within cells. Such circuits can help make engineers and physicists more comfortable with biology because it is described in a familiar language. Distributed circuits will help us understand Maxwell's equations and antennas intuitively. Circuits will help us quickly understand chemical reactions. Circuits will even help us understand the energy efficiency of cars.

In the rest of this chapter, we shall summarize some themes, ideas, principles, and biomedical and bio-inspired system examples that are discussed in depth elsewhere in the book. In this introductory chapter, the aim is to provide an intuitive 'big picture' without getting caught up in details, citations, proofs, mathematical equations and definitions, subtleties, and exceptions, which are addressed in the remaining chapters of the book. We shall start by discussing the importance of ultra-low-power electronics. We shall describe a power-efficient regime of transistor operation known as the subthreshold regime, which is enabling in low-power design. We shall then discuss important connections between information, energy, and power. We shall highlight some key themes for designing ultra-low-power mixed-signal systems that have analog and digital parts. We shall discuss examples of biomedical application contexts for low-power design, and fundamental principles of low-power design that are applicable to all systems, analog or digital, electronic or biological. After providing some numbers for the amazing energy efficiency of biological systems, we shall briefly discuss examples of systems inspired by neurobiology and by cell biology. Then, we provide a discussion of batteries and other energy sources, highly important components of low-power systems at small scales and at large scales. Finally, we shall conclude with a summary of the book's sections and some notes on conventions followed in the book.

1.1 Importance of ultra-low-power electronics

Ultra-low-power electronics in this book usually refers to systems that operate anywhere from a pico- to a milliwatt. However, the principles of ultra-low-power design are useful in all kinds of systems, even in low-power microprocessors, that

Since ultra-energy-efficient biological systems also operate with Boltzmann exponential devices, subthreshold operation is highly useful in mimicking their operation. Thus, subthreshold operation is enabling in bio-inspired systems as well.

1.3 Information, energy, and power

Information is always represented by the states of variables in a physical system, whether that system is a sensing, actuating, communicating, controlling, or computing system or a combination of all types. It costs energy to change or to maintain the states of physical variables. These states can be in the voltage of a piezoelectric sensor, in the mechanical displacement of a robot arm, in the current of an antenna, in the chemical concentration of a regulating enzyme in a cell, or in the voltage on a capacitor in a digital processor. Hence, it costs energy to process information, whether that energy is used by enzymes in biology to copy a strand of DNA or in electronics to filter an input.[2] To save energy, one must then reduce the amount of information that one wants to process. The higher the output precision and the higher the temporal bandwidth or speed at which the information needs to be processed, the higher is the rate of energy consumption, i.e., power. To save power, one must then reduce the rate of information processing. The information may be represented by analog state variables, digital state variables, or by both. The information processing can use analog processing, digital processing, or both.

The art of low-power design consists of decomposing the task to be solved in an intelligent fashion such that the rate of information processing is reduced as far as is possible without compromising the performance of the system. Intelligent decomposition of the task involves good architectural system decomposition, a good choice of topological circuits needed to implement various functions in the architecture, and a good choice of technological devices for implementing the circuits. Thus, low-power design requires a deep knowledge of devices, circuits, and systems. This book shall discuss principles and examples of low-power design at all of these levels. Figure 1.1 shows the "low-power hand". The low-power hand reminds us that the power consumption of a system is always defined by five considerations, which are represented by the five fingers of the hand: 1) the task that it performs; 2) the technology (or technologies) that it is implemented in; 3) the topology or architecture used to solve the task; 4) the speed or temporal bandwidth of the task; and, 5) the output precision of the task. As the complexity, speed, and output precision of a task increase, the rate of information processing is increased, and the power consumption of the devices implementing that task increases.

[2] In Chapter 22, we shall see that, technically, if one operates infinitely slowly and in a manner that allows the states of physical variables to be recovered even after they have been transformed, energy need not be dissipated. In practice, in both natural and artificial systems, which cannot compute infinitely slowly, and which always have finite losses, there is always an energy cost to changing or maintaining the states of physical variables.

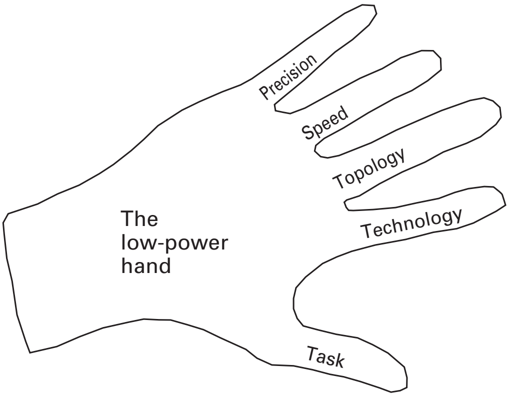


Figure 1.1. The low-power hand.

1.4 The optimum point for digitization in a mixed-signal system

The problem of low-power design may be formulated as follows: Suppose we are given an input **X**, an output function $\mathbf{Y}(t) = f(\mathbf{X}, t)$, basis functions $\{i_{out1} = f_1(\mathbf{v}_{in}), i_{out2} = f_2(d\mathbf{v}_{in}/dt), i_{out3} = f_3(\int \mathbf{v}_{in}), ..\}$ formed by the current-voltage curves of a set of technological devices, and noise-resource equations for devices in a technology that describe how their noise or error is reduced by an increase in their power dissipation for a given bandwidth; such noise-resource equations are described in Equation (22.4) in Chapter 22. Then, find a topological implementation of the desired function in terms of these devices that maximizes the mutual information between the actual output $\mathbf{Y}(t)$ and the desired output $f(\mathbf{X},t)$ for a fixed power-consumption constraint or per unit system power consumption. Area may or may not be a simultaneous constraint in this optimization problem. A high value of mutual information, measured in units of bits per second, implies that the output encodes a significant amount of desired information about the input, with higher mutual information values typically requiring higher amounts of power consumption.

Hence, low-power design is in essence an information-encoding problem. How do you encode the function you want to compute, whether it is just a simple linear amplification of a sensed signal or a complex function of its input, into transistors and other devices that have particular basis functions given by their current-voltage curves? Note that this formulation is also true if one is trying the minimize the power of an actuator or sensor, since information is represented by physical state variables in both, and we would like to sense or transform these state variables in a fashion that extracts or conveys information at a given speed and precision. In non-electrical systems, through (current) and across (voltage) variables play the roles of current and voltage, respectively. For example, in a fluid-mechanical system, pressure is analogous to voltage while volume velocity of fluid flow is analogous to current.

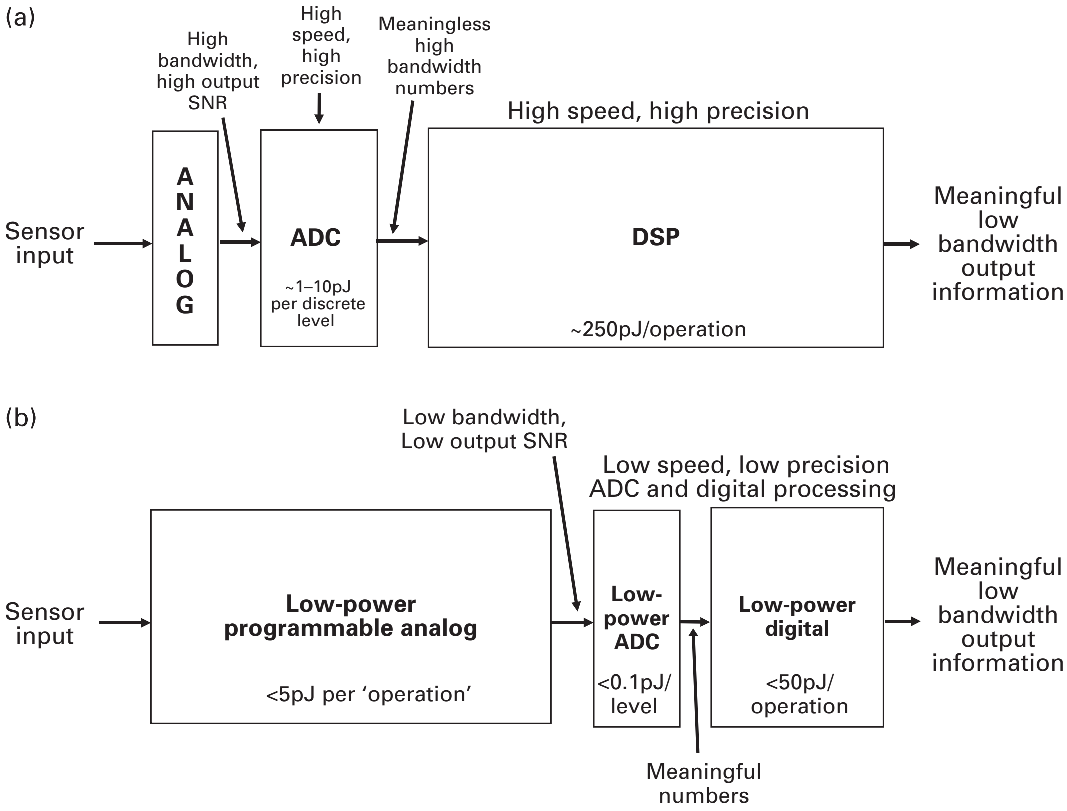

(a)
High bandwidth, high output SNR
High speed, high precision
Meaningless high bandwidth numbers
High speed, high precision
Sensor input
ANALOG
ADC
~1–10pJ per discrete level
DSP
~250pJ/operation
Meaningful low bandwidth output information
(b)
Low bandwidth, Low output SNR
Low speed, low precision ADC and digital processing
Sensor input
Low-power programmable analog
<5pJ per 'operation'
Low-power ADC
<0.1pJ/ level
Low-power digital
<50pJ/ operation
Meaningful low bandwidth output information
Meaningful numbers

Figure 1.2a, b. (a) A traditional ADC-then-DSP architecture; (b) A more energy-efficient mixed-signal architecture. The numbers shown represent bounds that are constantly improving for both analog and digital technologies and that vary with implementation details.

power of an electronic system is lowered by delaying digitization. Figure 1.2 (a) shows a method of processing information where digitization is immediate, while Figure 1.2 (b) shows a method for processing information where digitization is delayed. The delaying of digitization leads to a significant lowering of power in the ADC and in the digital portions of the system but an increase in the analog power consumption. There is an optimal point for digitization where the overall power is minimized that varies depending on what is being computed. The lowering of power arises not only because speed and/or precision are lowered in the ADC or in the digital processing but also because it is easier to design energy-efficient ADCs and low-power digital systems at lower speeds and/or precisions. We shall discuss ultra-low-power ADCs in Chapter 15 and principles for ultra-low-power digital design in Chapter 21.

Several biomedical applications can exploit the architecture of Figure 1.2 (b) to lower power consumption. In these applications, the meaningful information

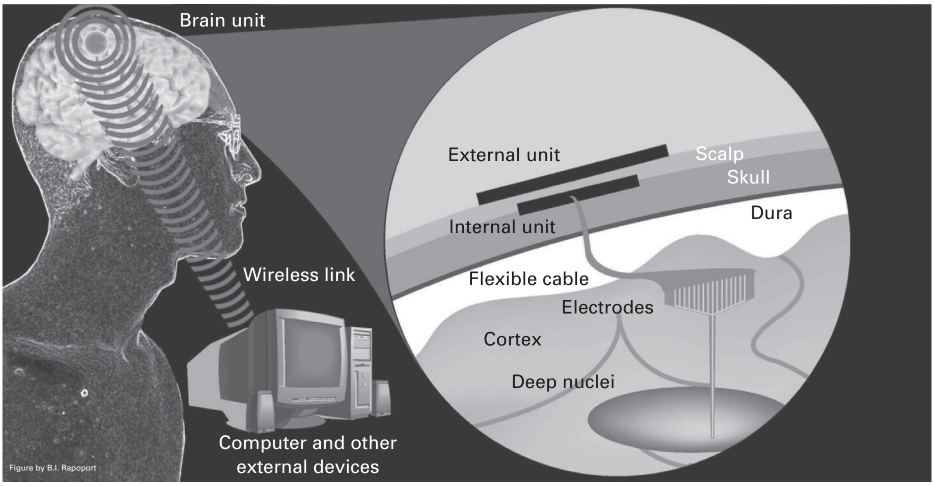


Figure 1.4. Configuration of a brain implant or brain-machine interface (BMI).

receive power and data wirelessly from an external unit as in cochlear implants. The received power can periodically recharge an implanted battery or constantly power the internal unit wirelessly. The external unit connects in a wired (or wireless) fashion with an imager and a processor, which can be worn on a set of 'glasses'. The external unit is capable of wireless communication to a computer via a standard Ultra Wide Band (UWB), Zigbee, or Bluetooth interface for programming, monitoring, and debugging functions. The external unit and internal unit are mechanically aligned via electronic magnets. The overall configuration is one of several possibilities but presents several advantages including stable coil coupling, a relatively benign bio-environment between skin and scalp for the implanted unit, minimization of tissue heating within the skull, the ability to be relatively stable during brain and head motion, and good RF link efficiency.

Figure 1.4 also illustrates how the same BMI is useful for the treatment of paralysis. In this case, neural electrical-recording (sensing) electrodes from the surface of a motor region of the brain can be used to decode the intention of a paralyzed patient to move their limb. The decoded information can then be used to stimulate a muscle or robot arm in the patient. The decoding can be fully or partially done in the external unit based on data conveyed by the internal unit. The external unit then wirelessly relays motor commands to the arm. Brain implants for epilepsy require simultaneous recording and stimulation circuitry since the detection of a seizure must trigger electrical stimulation to suppress it.

The examples of Figure 1.3 and Figure 1.4 represent two biomedical application contexts where several ultra-low-power RF, sensor, analog processing, electrode-recording, and electrode-stimulation building-block circuits, which are described in the book, are useful. Implants for several other biomedical applications, e.g., cardiac pacemakers, spinal cord stimulators, deep-brain stimulators for the treatment of Parkinson's disease, vagal-nerve stimulators, etc., utilize several of these same building-block circuits, to architect slightly different systems that all operate by and large with the same technology base. In fact, biocompatibility design, hermetic design, mechanical design, and electrode design share several similarities in all of these applications as well. We shall also discuss how to architect far-field energy-harvesting RF circuits for non-invasive cardiac medical applications, which will make the reader familiar with the basic principles of antenna design. Antenna-based RF communication systems transmit information through a relatively thick portion of the body, e.g., for deep implants such as electronic pills, used for diagnosis of gastrointestinal disorders. Chapters 10 through 18 discuss several circuits useful in low-power biomedical applications.

Chapter 19 focuses on implantable electronics, with an emphasis on cochlear implants and brain implants. It provides a concrete example of the power savings possible when analog preprocessing is used to delay digitization. Chapter 20 focuses on noninvasive medical electronics with an emphasis on cardiac devices and biomolecular sensing.

1.6 Principles for ultra-low-power design

The principles for low-power design discussed in the book apply to all systems that aim to use power efficiently, independent of their absolute power consumption. These principles are true for both biological systems and electronic systems. They include:

1. Encoding the task in the computational basis functions of technology in an efficient fashion to save power, e.g., the use of exponentials to efficiently compute logarithms in electronics or the use of chemical binding in biology to multiply.
2. Using energy-efficient regions of operation of technological devices, e.g., the use of the exponential subthreshold regime of transistor operation in analog and digital electronics and the use of exponential Boltzmann relations in biology.
3. Delaying digitization via analog preprocessing to reduce the information bandwidth in the computation, e.g., as in the filtering and mixing operations in a radio today or in the image preprocessing done by the retina in the eye.
4. Decomposing tasks into more energy-efficient slow-and-parallel architectures rather than fast-and-serial ones, e.g., as in several low-power digital architectures today or in slow-and-parallel computation in the brain.
5. Balancing the costs of computation and communication, e.g., as in a wireless biomedical system transmitting high-level versus low-level information or in the division of energy for computation versus communication in the brain's neuronal computing cells.
6. Reducing the amount of information that needs to be processed, e.g., as in automatic-gain-control circuits in analog electronics, in gated-clock circuits in digital electronics, in event-driven asynchronous analog and digital systems,

in calcium gain-control circuits in the photoreceptor of the eye, and in gated bio synthesis cell-regulation circuits.

7. Using feedback, knowledge, and learning to improve the efficiency of the computation via error-correction, compression, and optimization mechanisms, e.g., as in digitally calibrated analog-to-digital converters, auto-zeroing analog systems, negative auto-regulation feedback loops within cells, or in adaptive power-supply circuits in digital design.
8. Architecting the task such that its circuits do not need to be simultaneously fast and precise, e.g., as in comparators in electronics or in neuronal comparators in biology.
9. Operating slowly or 'adiabatically' with passive and active components that consume little power, e.g., as in adiabatically clocked digital circuits in electronics, in high-quality-factor circuits in electronics, or in high-quality-factor mechanical transmission lines in the biological inner ear or cochlea. The quality factor is a measure of the ratio of the reactive energy to dissipative energy in a system.

These low-power principles are discussed in detail in Chapter 22. Many of the low-power principles that we have listed apply to both low-power analog and low-power digital design although they are often manifested in seemingly different ways. Low-power digital design is discussed in Chapter 21. Chapter 22 discusses several similarities between low-power analog and low-power digital design. In both analog and digital systems, natural or artificial systems, robustness and flexibility trade off against the efficiency of the architecture. Extra degrees of freedom are always needed to attain robustness and flexibility, which compromise its efficiency. A good architecture must be designed to be efficient and robust without being needlessly flexible. We show in Chapter 22 that biological systems obey *all* of the energy-saving principles listed above in an exemplary fashion. They also obey another important principle that we term *collective analog* or *hybrid computation* and that we explain in Chapter 22 [3], [4]. In addition, they provide us with clues for building ultra-low-power systems that force us to think outside the box of traditional engineering designs.

1.7 Ultra-low-power information processing in biology

Biology has designed architectures where many noisy, imprecise, unreliable analog devices collectively interact through analog and digital signals to solve a task in real time, precisely, and with astoundingly low power. For example, a single neuron in the brain performs highly complex real-time pattern recognition, spatio-temporal filtering, and learning tasks with $\sim$0.66 nW of power. The $\sim$22 billion neurons of the brain consume only $\sim$14.6 W of power in an average 65 kg male. Neurons collectively interact to perform sensorimotor tasks in noisy environments in real time that no computer can yet match. A single cell in the body performs $\sim$10 million energy-consuming biochemical operations per second on its noisy molecular inputs with ~1 pW of average power. Every cell implements a ~30,000 node gene-protein molecular interaction network within its confines. All the ~100 trillion cells of the human body consume ~80 W of power at rest. The average energy for an elementary energy-consuming operation in a cell is about 20 kT, where kT is a unit of thermal energy. In deep submicron processes today, switching energies are nearly $10^4 - 10^5$ kT for just an elementary 0→1 digital switching operation. Even at 10 nm, the likely end of business-as-usual transistor scaling in the future, it is unlikely that we will be able to match such energy efficiency. Unlike traditional digital computation, biological computation is tolerant to error in elementary devices and signals. Nature illustrates that it is significantly more energy efficient to compute with error-prone devices and signals and then correct for these errors through feedback-and-learning architectures than to make every device and every signal in a system robust, as in traditional digital paradigms thus far.

We can learn from Nature, for she has had 1 billion years of experimentation to evolve magnificent nanotechnological, low-power devices, circuits, topologies, and architectures in environments where food, and consequently energy, was scarce. What we learn can help inspire the design of novel engineering systems termed *bio-inspired systems*. Such inspiration from nature must always be combined with perspiration and rational design from engineering if the final result is to be useful. In this book, we shall always take an engineering approach and show that bio-inspired designs can and do result in impressive engineering architectures. Birds are not airplanes and airplanes are not birds, but one can shed insight into the operation of the other. A humble, open, and curious mindset toward ideas in nature is all that is needed for appreciating this field.

We shall find on several occasions that bio-inspired algorithms, architectures, and circuits frequently have biomedical applications. Not surprisingly, it helps to mimic how the biology works if one is attempting to fix it. For example, we shall discuss an asynchronous stochastic sampling strategy inspired by how the auditory-nerve works in Chapter 19 that enables an approximately 6× reduction in electrode stimulation power while maintaining a high effective rate of sampling. Bio-inspired systems, which have just scratched the surface of what will likely be possible in the future, already show that there are several clever ideas in biology that are useful in engineering, and not just for low-power design. Now, we shall highlight one example of a neuromorphic electronic system, i.e., an electronic system inspired by neurobiology. The term *neuromorphic electronics* was coined by the founder of the field, Carver Mead. We shall describe an *RF cochlea*, a fast, power-efficient radio-frequency spectrum analyzer useful in broadband wireless communication systems.

1.8 Neuromorphic system example: the RF cochlea

The biological inner ear, or cochlea, performs spectrum analysis on its incoming sound input through the use of a mechanical transmission-line architecture.

The transmission-line architecture of the cochlea has exponentially tapered time constants from its beginning or high-frequency base location to its ending or low-frequency apex location. The net result is that the cochlea separates sounds based on their frequencies, with high-frequency sounds stimulating the base or beginning of the cochlea and low-frequency sounds stimulating the apex or end of the cochlea. Thus, the cochlea functions as a spectrum analyzer that performs a frequency-to-space or frequency-to-location transformation on its sound input. It operates over an ultra-broadband 100:1 sound-carrier-frequency range of ~100 Hz to 10 kHz with good sensitivity. The cochlea also performs gain control and compression on its sound input via a distributed amplification system within the transmission line that is highly energy efficient. Therefore, we can hear over a 120 dB input dynamic range with a minimum detectable signal of 0.05 Å at our ear drums at our most-sensitive frequency of ~3 kHz. The power consumption of the cochlea is ~14 μW even though it implements at least ~1 billion floating-point operations per second. Nonlinearity and gain-control in the cochlea are important for enhancing signals in noisy environments and for our ability to hear speech in noise. An appendix in Chapter 23 discusses how we can estimate power and/or information processing rates in the ear, the eye, the brain, and the body.

The ear operates remarkably like an ultra-broadband super-radio for sound waves with 3,500 output spectral channels operating in parallel. The outer ear or pinna is a directional antenna. The middle ear is an impedance-matching transformer that matches the impedance of the antenna to the impedance of the inner ear or cochlea. The piezoelectric outer hair cells in the cochlea function like amplifiers that enhance the passive resonant gain of its membrane-and-fluid transmission-line structure. The inner hair cells in the cochlea function like rectifying demodulators that detect modulations of the carrier sound waves propagating through the cochlea. Finally, the auditory-nerve output spikes or pulses sample and partially quantize the inner-hair-cell cochlear output for eventual communication to the brain. Figure 19.1 in Chapter 19 reveals some of the anatomy of the ear.

We show in Chapter 23 that the exponentially tapered time-constant architecture of the cochlear transmission-line allows the cochlea to implement the fastest and most-efficient spectrum-analysis architecture that is currently known. For a spectrum analyzer with N output bins operating over a given input-frequency range, the cochlear architecture only consumes $O(N)$ resources in time and $O(N)$ resources in hardware to perform spectrum analysis. In contrast, a constant fractional-bandwidth analog filter bank spectrum analyzer consumes $O(N)$ resources in time and $O(N^2)$ resources in hardware. A fast-fourier-transform (FFT) fixed-bandwidth spectrum-analyzer consumes $O(N\log_2 N)$ resources in time and $O(N\log_2 N)$ resources in hardware. A traditional swept-sine spectrum analyzer consumes $O(N^2)$ resources in time and $O(1)$ resources in hardware. Thus, the cochlear architecture has a good tradeoff between the use of temporal resources and hardware resources, and for a given amount of hardware resources its architecture for spectrum analysis leads to the quickest performance.

The RF cochlea is an extremely recent neuromorphic architecture, which will undoubtedly evolve over time [5]. However, it has already excited interest and enthusiasm in the radio-engineering community. In general, inspiration from the ear and auditory system may lead to revolutionary RF architectures in the future.

1.9 Cytomorphic electronics

Circuits in cell biology and circuits in electronics may be viewed as being highly similar with biology using molecules, ions, proteins, and DNA rather than electrons and transistors. Just as neural circuits have led to biologically inspired neuromorphic electronics, cellular circuits can lead to a novel biologically inspired field that we introduce in this book and term *cytomorphic electronics*. We will show that there are many similarities between spiking-neuron computation and cellular computation in Chapter 24.

Figure 1.6 illustrates that there are striking similarities between chemical reaction dynamics (Figure 1.6 (a)) and electronic current flow in the subthreshold regime of transistor operation (Figure 1.6 (b)). Electron concentration at the source is analogous to reactant concentration; electron concentration at the drain is analogous to product concentration; forward and reverse current flows in the transistor are analogous to forward and reverse reaction rates in a chemical reaction; the forward and reverse currents in a transistor being exponential in voltage differences at its terminals are analogous to reaction rates being exponential in the free energy differences in a chemical reaction; increases in gate voltage lower energy barriers in a transistor increasing current flow are analogous to the effects of enzymes or catalysts in chemical reactions that increase reaction rates; and, the stochastics of Poisson shot noise in subthreshold transistors are

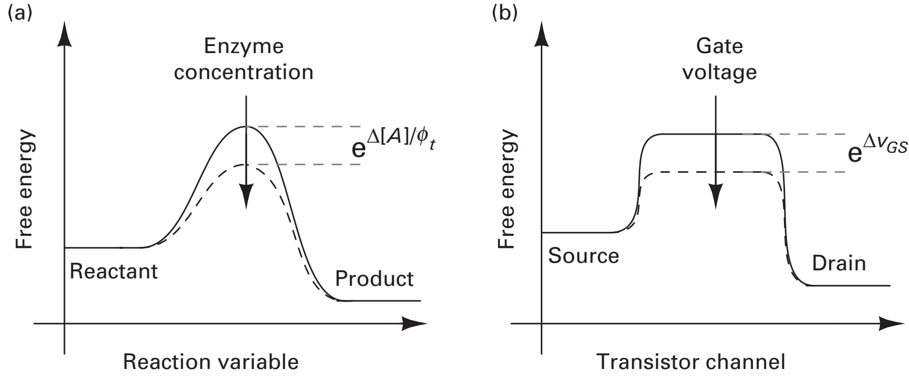


Figure 1.6a, b. Similarities between chemical reaction dynamics and subthreshold transistor electronic flow. Reprinted with permission from [6] (© 2009 IEEE).

analogous to the stochastics of molecular shot noise in reactions. These analogies suggest that one can mimic and model large-scale chemical-processing systems in biological and artificial networks very efficiently on an electronic chip at time scales that could potentially be a million times faster. No one thus far appears to have exploited the detailed similarity behind the equations of chemistry and the equations of electronics to build such networks. The single-transistor analogy of Figure 1.6 is already an exact representation of the chemical reaction $A \rightleftarrows B$ including stochastics, with forward electron flow from source to drain corresponding to the $A \rightarrow B$ molecular flow and backward electron flow from drain to source corresponding to the $B \rightarrow A$ molecular flow. In Chapter 24, we shall build on the key idea of Figure 1.6 to show how to create current-mode subthreshold transistor circuits for modeling arbitrary chemical reactions. We can then create large-scale biochemical reaction networks from such circuits for modeling computation within and amongst cells.

Since extracellular cell-cell networks also rely on molecular binding and chemical reactions, networks such as hormonal networks or neuronal networks can be efficiently modeled using such circuits. Thus, in the future, we can potentially attempt to simulate cells, organs, and tissues with ultra-fast highly parallel analog and hybrid analog-digital circuits including molecular stochastics and cell-to-cell variability on large-scale electronic chips. Such molecular-dynamics simulations are extremely computationally intensive, especially when the effects of noise, nonlinearity, network-feedback effects, and cell-to-cell variability are included. Stochastics and cell-to-cell variability are highly important factors for predicting a cell's response to drug treatment, e.g., the response of tumor cells to chemotherapy treatments. We will show in Chapter 24 that circuit, feedback, and noise-analysis techniques described in the rest of this book can shed insight into the systems biology of the cell. For example, flux balance analysis is frequently used to reduce the search space of parameters in a cell. It is automatically implemented as Kirchhoff's current law in circuits since molecular fluxes map to circuit currents. Similarly, Kirchhoff's voltage law automatically implements the laws of thermodynamic energy balance in chemical-reaction loops. Robustness analysis of the circuit using feedback techniques can shed insight in the future into which genes, when mutated, will lead to disease in a network, and which will not. Circuit-design techniques can also be mapped to create synthetic-biology circuits that will perform useful functions in the future.

1.10 Energy sources

A highly important component of any low-power electronic system is the battery or energy source. Therefore, the last two chapters of the book are devoted to this topic. Chapter 25 discusses batteries and electrochemistry in some depth. It shows how a simple physical interpretation of the equations of electrochemistry lead to electrical large-signal and small-signal equivalent circuits that make the operation of batteries intuitive. These equivalent circuits are also useful for modeling recording and stimulation electrodes in biomedical systems. A new and simple formula for battery operation that characterizes the loss in battery capacity with increasing usage and increasing current draw is described. A discussion of how battery–ultra-capacitor and fuel-cell–battery hybrids can enable advantageous operation in many systems is presented. We shall see that, in low-power systems, a battery does not have a longer lifetime only because its power draw is low. Its lifetime is longer also because it becomes capable of more charge-recharge cycles and because its geometry within a fixed volume can be architected such that it has more charge capacity. Furthermore, when the battery supplies current, its output voltage is higher, increasing its energy efficiency. Thus, the rate of use of energy is intimately linked to its storage and generation.

Chapter 26 reviews prior work on energy harvesting, including the use of body motion, body heat, and solar energy in self-powered battery-free biomedical and portable systems. The fundamental principles of piezoelectric motion-energy harvesting, thermoelectric body-heat harvesting, and photovoltaic electricity generation are discussed. Circuit models for energy harvesting are found to be similar to circuit models for RF energy harvesting discussed in depth in Chapters 16 and 17. We shall find that the principles of low-power electronic design described in this book apply not just to electronic systems and at small scales, but are also useful for non-electrical systems and at large scales, e.g., in low-power cars. An equivalent circuit model of a car can help us understand issues related to the energy efficiency of transportation. Thus, the transport energy efficiency of, say, a cheetah versus an electric car or a bicycle, or a bird versus an airplane can be compared. The current power consumption of the world is largely dominated by transport, heating, and electricity costs and is a staggering 15 TW today. We shall summarize ideas actively being researched for architecting low-power systems that function with renewable sources of energy like solar power or biofuels, a highly likely necessity in our planet's future.

We shall now conclude our top-down view of some of the book's contents. In the next section, we will present a brief bottom-up view of the book's organization.

1.11 An overview of the book's chapters and organization

The book is organized into seven sections, each with several chapters. We shall discuss each section of the book briefly.

1. The **Foundations** section of the book contains ten chapters including this overview chapter, Chapter 1. This section contains a review of feedback

systems in Chapter 2, an in-depth discussion of the device physics of the MOS transistor in Chapters 3 through 6, and a discussion of noise in devices and circuits in Chapters 7 and 8, respectively. The section concludes with more material on feedback systems and feedback-circuit-analysis techniques in Chapters 9 and 10, respectively.

2. The **Low-Power Analog and Biomedical Circuits** section of the book is formed by Chapters 11 through 15. This section contains various circuits that are useful for low-power biomedical electronics and analog electronic systems in general. The foundational material from the first section enables design and analysis of these circuits.
3. The **Low-Power RF and Energy-Harvesting Circuits for Biomedical Systems** section of the book is formed by Chapters 16 through 18. It contains a description of energy-efficient power and data radio-frequency (RF) links that are uniquely suited to biomedical systems.
4. The **Biomedical Electronic Systems** section of the book contains a chapter on ultra-low-power implantable electronics, Chapter 19, and a chapter on ultra-low-power noninvasive medical electronics, Chapter 20. In Chapter 19, exemplary systems for cochlear implants for the deaf, brain implants for the blind and paralyzed, and other implantable systems are discussed. Building-block low-power circuits from the previous chapters and new circuits unique to implantable electronics show how large systems can be architected. In Chapter 20, cardiac devices for noninvasive medical monitoring are discussed. Principles for biomolecular sensing, such as in bioMEMS and microfluidic systems, are also discussed.
5. The **Principles for Ultra-low-power Analog and Digital Design** section of the book contains one chapter on principles for ultra-low-power digital design, Chapter 21, and one chapter on principles for ultra-low-power analog and mixed-signal design, Chapter 22. Similarities in principles for low-power analog and digital design are discussed. Ten principles for ultra-low-power design are discussed, all of which are obeyed by ultra-energy-efficient biological systems.
6. The **Bio-inspired Systems** section of the book comprises two chapters. The first chapter, Chapter 23, on *neuromorphic electronics* discusses electronics inspired by neurobiology. The second chapter, Chapter 24, discusses the novel form of electronics that we have termed *cytomorphic electronics*, i.e., electronics inspired by cell biology. Applications of neuromorphic and cytomorphic electronics to engineering and medicine are discussed.
7. The **Energy Sources** section of the book comprises Chapters 25 and 26. Chapter 25 on batteries and electrochemistry discusses how batteries work from a unique circuit viewpoint and presents important tradeoffs between

energy density and power density. Chapter 26 discusses energy harvesting in portable and biomedical systems at small scales and at larger scales. We show how some of the principles of low-power design that we have studied apply not only at small scales and in electronics but also at large scales and in non-electrical systems.

1.12 Some final notes

Problem sets, errata, and instructor material for the book will be available at the Cambridge University Press website (http://www.cambridge.org/). The author's full name will serve as an index term for navigating this site.

The book attempts as far as possible to follow IEEE convention for algebraic symbols: A dc variable is denoted by I_A, a small-signal variable is denoted by i_a, a total-signal variable is denoted by i_A, and a frequency-response or Laplace variable is denoted by $I_a(f)$ or $I_a(s)$, respectively. Thus, $i_A = I_A + i_a$. However, there are situations in the book where this convention has not been followed to improve clarity, or where doing so does not matter much in the discussion.

I have tried to write this book such that it is clear, accessible, and rewards the reader with a broad and deep knowledge in the fields of both ultra-low-power electronics and bioelectronics. As I have discussed, physically intuitive ways of thinking are emphasized throughout this book. It is worth noting that it was Einstein's impressive physical intuition that enabled him to see that the equations of special relativity, which had already been stated by Lorenz, had a transparently simple derivation if one accepted that the speed of light is a constant in nature. It is no surprise, then, that Einstein said, "The only real valuable thing is intuition." While most of us, including the author, do not have Einstein's physical intuition, we can take inspiration from Einstein to always strive for his simple intuitive way of understanding phenomena. It is by developing our scientific intuition that we can begin to see connections that illuminate the unity of all nature.

References

[1] R. Sarpeshkar, C. D. Salthouse, J. J. Sit, M. W. Baker, S. M. Zhak, T. K. T. Lu, L. Turicchia and S. Balster. An ultra-low-power programmable analog bionic ear processor. *IEEE Transactions on Biomedical Engineering*, **52** (2005), 711–727.

[2] A. T. Avestruz, W. Santa, D. Carlson, R. Jensen, S. Stanslaski, A. Helfenstine and T. Denison. A 5 μW/channel Spectral Analysis IC for Chronic Bidirectional Brain-Machine Interfaces. *IEEE Journal of Solid-State Circuits*, **43** (2008), 3006–3024.

[3] R. Sarpeshkar. Analog versus digital: extrapolating from electronics to neurobiology. *Neural Computation*, **10** (1998), 1601–1638.

feedback loops in circuits are often intentionally exploited to create circuits that perform faster, that respond only to changing inputs, or that reduce leakage rates on a capacitor, etc.

Feedback is at the heart of how *all* circuits work, whether electrical or non-electrical in nature. All nontrivial circuits arise from feedback interactions between devices in physics, chemistry, or biology that create a circuit, sometimes called a network or system. Thus, a deep understanding of feedback is important to an understanding of all systems, since feedback principles are at the heart of how simple devices, when hooked together, exhibit complex behavior in a circuit or system. We will begin with a few examples that illustrate how amazingly broad the reach of feedback circuits is in all of science and engineering.

2.1 Feedback is universal

Feedback circuits ensure that we are all alive and well! Several homeostatic systems use molecular negative-feedback circuits to ensure that our temperature, the pH of our blood and intracellular fluids, our weight, and the concentrations of various nutrients and molecules in our cells and bodies are regulated to be within an acceptable range. Gene-protein feedback networks within cells ensure normal functioning of the cell under various environmental conditions and developmental cycles. Cell growth is regulated to ensure normal growth and functioning of the body's organs and to prevent cancer. Positive and negative feedback circuits in the immune system help combat infection and implement self healing and repair. A malfunction in any of these feedback systems in our cells or bodies can lead to debilitating diseases like diabetes, hyperthyroidism, cancer, or autoimmune disease.

The sensory and motor systems of our bodies make extensive use of feedback circuits to improve their performance and dynamic range of operation. For example, pupil accommodation in the eye and calcium-based feedback circuits in our photoreceptors allow us to see over a wide range of light levels. Auditory-gain-control systems in our ears exploit nonlinearity and feedback in our outer hair cells to enable us to hear over a wide range of sound intensity levels. Our ability to continually gaze at moving objects is made precise via eye-movement pursuit and tracking feedback circuits. Our motor systems use sensory feedback to improve their performance, e.g., visual feedback for locomotion and auditory feedback for speech production.

The brain is a massive feedback circuit with extensive excitatory and inhibitory connections amongst its cells, termed *neurons*. There are significantly more feedback connections formed by the cells of the brain among themselves than feedforward connections received from the inputs to the brain. The pulsatile signal variables of neurons, known as *spikes*, are created by transient positive-feedback action in sodium conductances in neurons much as in a relaxation oscillator in electrical circuits. Learning may be viewed as feedback information from errors,

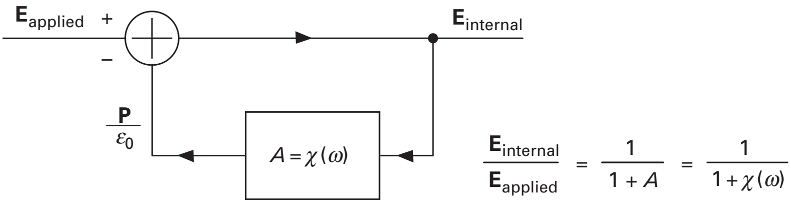

$\mathbf{E}_{\text{applied}}$
+
−
$\mathbf{E}_{\text{internal}}$
$\frac{\mathbf{P}}{\varepsilon_0}$
$A = \chi(\omega)$
$\frac{\mathbf{E}_{\text{internal}}}{\mathbf{E}_{\text{applied}}} = \frac{1}{1+A} = \frac{1}{1+\chi(\omega)}$

Figure 2.1. Block diagram of the negative-feedback loop that creates the dielectric constant.

failure, or from the structure of the input space that is ported into the adaptive architecture of the brain to improve its performance over time. Evolution comprises slower feedback from the environment to the genes.

All chemical reactions are bidirectional and have two embedded feedback loops within them that help establish their equilibrium: an increase in the rate of the backward reaction if the concentration of products becomes excessive and a decrease in the rate of the forward reaction as the concentration of reactants is used up. Several biochemical reaction pathways in the body have enzyme-mediated feedback loops that function by products enhancing or decreasing the rates of the reactions that created them. Bonds between atoms in a molecule or between molecules lead to an equilibrium separation established by forces that effectively implement a feedback mechanism that pushes the constituents apart if they get too close and that pulls them back together if they get too far apart.

In physics, feedback mechanisms are ubiquitously present, although they are not often described as such. For example, the electric field emanating from a charge in water is reduced by a factor of 81 from the electric field emanating from the same charge in air or in vacuum. Therefore, we say that water has a dielectric constant that is 81 times larger than that of air. Such an effect arises because polar water molecules strongly oppose any electric field within them by creating their own anti-aligned polarization electric field that subtracts from the input field and thus attempts to neutralize it via negative-feedback action. The small residual output electric field that remains after near-perfect neutralization is 81 times weaker because the negative-feedback loop gain A of this feedback loop is 80, and, as we shall discuss later in the chapter, the effect of such negative-feedback action is to weaken the response to the input by a factor of $1/(1+A)$. Figure 2.1 shows the operation of the feedback loop and the basic negative-feedback equations that explain this dielectric effect. As a consequence, electrostatic forces between charges in water are highly attenuated compared with forces in air. Therefore, ionic substances like salt easily dissolve in water due to the weakening of the forces which hold them together. The variable $\chi(\omega)$ in Figure 2.1 is called the *susceptibility*, which in our viewpoint is simply the frequency-dependent loop gain $A(\omega)$.

Phase transitions that cause a substance to change from one form, e.g., solid ice, to another form, e.g., liquid water, at a particular magical temperature known as the *melting point* are due to a positive-feedback loop. At the magical temperature, the thermal energy is sufficiently large such that the breaking of a few bonds between ice molecules causes bonds between other molecules to more easily break since they are now collectively more loosely held. Thus, the breakage of a few bonds causes easier breakage of more bonds through positive-feedback action until the entire substance melts into a new stable form, water. The whole process is similar to the positive-feedback action seen when a slight tear in a garment can lead to a larger tear and finally tearing of the entire garment. Presumably, at the magical melting point, the collective positive-feedback loop gain of the system is greater than 1, the critical value at which a positive-feedback loop just goes unstable.

Maxwell's equations inherently have feedback embedded into them because the distributions of charges and motions of charges that create electric and magnetic fields are themselves affected by these same fields. Thus, fields and charged matter form a reciprocal feedback network with each affecting the other.

A stable downward-hanging pendulum has forces at equilibrium that are an example of negative-feedback action that restore it to its lowest point. An unstable upward-hanging inverted pendulum has forces at equilibrium that are an example of positive-feedback action that move it away from its highest point.

Feedback circuits are so ubiquitous in control systems in engineering that the terms *feedback* and *control* are almost used synonymously even though control systems could be purely feed-forward in principle. Industrial-process control, aircraft control, mechanical system control, robotic systems, space-flight systems, guidance and navigation systems all exploit feedback circuits to function.

Distributed computing systems such as those present in the internet are composed of multiple interacting feedback loops within the communication network. All communication systems that have the possibility of sending messages from a transmitter to a receiver and then receiving one in turn via one or more hops in the network may be viewed as feedback networks. Social communication networks are an example of such networks as well.

Feedback is the essential ingredient that makes circuits work. Without feedback, analog circuits would be simple, imprecise, feed-forward, impractical curiosities. With feedback, analog systems become rich, complex, precise, practical systems that can make our hearts pound with excitement. Without a deep understanding of feedback, every analog circuit appears like a special case. With a deep understanding of feedback, almost all analog circuits begin to look like examples of a general pattern that has been studied before. That is why we tackle feedback systems before we tackle circuits in this book. Digital circuits incorporate feedback into a discrete dynamical system termed a *finite state machine*. Here, the output of the system feeds back and affects the processing of its input after a delay of one clock cycle. The discrete state of the system is concomitantly updated after this clock cycle as well. Digital systems also use positive feedback to create static-memory circuits, which store digital bits of information.

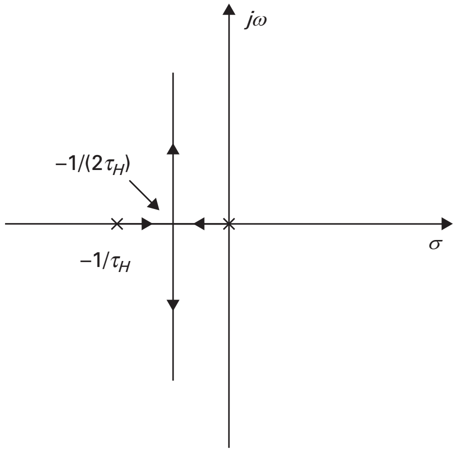


Figure 2.22. An example of the value rule for root loci.

Figure 2.22 illustrates a simple application of Equation (2.28) for $a(s)f(s) = 1/(\tau_L s(\tau_H s + 1))$. We would like to compute the value of K at which the pole at the origin and the pole at the $-1/\tau_H$ just meet before they depart from the real axis in the root-locus plot. This point is at $-1/(2\tau_H)$ by Grant's rule. Thus, the contribution of the normalized distance of the integrator pole at the origin is $\tau_L/(2\tau_H)$, the contribution of the normalized distance of the $-1/\tau_H$ pole is $\tau_H/(2\tau_H) = 1/2$. Thus, the value of K at which the two poles just meet before they depart from the real axis is given from Equation (2.28) to be the product of these normalized distances, i.e., $\tau_L/(4\tau_H)$. This leads to the well-known result that, if an integrator and a lowpass filter are placed together in a negative-feedback loop with $K = 1$, then the closed-loop poles are at 'critical damping' when the integrator's time constant is four times that of the lowpass filter's time constant.

The root-locus rules can be remembered by the mnemonic **B****eaches** **R****eally** **D****o** **M****ake** **A** **C****alm** **R****elaxing** **V****acation** where the underlined terms signify the Beginning, Real-axis, Departure, Mean, Asymptote, Complex-singularity, Remote, and Value rules respectively.

2.7 Example of a root-locus plot

Figure 2.23 (a) shows a feedback system composed of a motor with a lowpass transfer function $1/(\tau_m s + 1)$ between its input voltage and output angular velocity $\Omega_{out}(s)$ and an electrical controller transfer function $K(\tau_z s + 1)/(\tau_e s)$ composed of an integrator pole and a high-frequency zero. Figure 2.23 (b) plots the root-locus plot for this system as K varies from 0 to ∞. The root-locus plot of Figure 2.23 (b) is important in several feedback systems, including position and velocity control of motors, in transimpedance-amplifier feedback loops,

3 MOS device physics: general treatment

Intuition will tell the thinking mind where to look next.

Jonas Salk

To deeply understand any electronic circuit, whether it is low power or not, it is essential to have a good mastery of the devices from which that circuit is made. In this chapter, we will begin our study of device physics with the metal oxide semiconductor (MOS) transistor, the most important active device in electronics today. The MOS transistor is a field effect transistor (FET) and MOSFETs are abbreviated as nFETs if their current is due to electron flow and as pFETs if their current is due to hole flow. In this chapter, we shall focus on fundamental principles and on exact mathematical descriptions that are applicable to transistors built in technologies with relatively long dimensions. In later chapters, we shall study practical approximations needed to simplify these exact mathematical descriptions (Chapter 4), study small-signal dynamic models of the MOS transistor (Chapter 5), and discuss effects observed in deep submicron transistors with relatively short dimensions (Chapter 6).

Figure 3.1 shows a zoomed-in view of an n-channel FET or nFET built in a standard bulk complementary metal oxide semiconductor (CMOS) process [1]. There are four terminals referred to as the gate (G), source (S), drain (D), and bulk (B), respectively. The control terminal, the metal-like polysilicon gate, is insulated from the silicon bulk via a silicon dioxide insulator; the source and drain terminals are created with n+ regions in the p-type silicon bulk. The voltage v_{GS} refers to the voltage between the gate and source terminals and v_{DS} refers to the voltage between the drain and source terminals. It is conventional to either reference all voltages to the source, i.e., make the source voltage ground, to create a source-referenced description or to reference all voltages to the bulk, i.e., make the bulk voltage ground, to create a bulk-referenced description. By convention, the drain terminal always has a higher voltage than the source terminal in nFETs. Conventional positively charged current, i_{DS} or i_D, flows from drain to source in the nFET while negatively charged electron current flows from source to drain in an nFET. In the particular example of Figure 3.1, the bulk is tied to the source such that v_{BS} is zero and all voltages are referenced to the source. Due to the symmetry of the source and drain terminals, bulk-referenced descriptions are more symmetric than source-referenced models but source-referenced descriptions are more widespread in use.

bottom-plate constant-voltage terminal of an oxide capacitance with the gate being the other terminal. The bulk is then screened from the effects of the gate near the source end of the channel such that the electron charge at this end changes with changes in gate voltage but the bulk charge and surface potential there remain nearly constant. The change in electron charge concentration at the source end of the channel is then well modeled by the change in charge across a linear C_{ox} capacitor with a constant-voltage bottom plate and a changing top-plate gate voltage, and therefore is linear with changes in gate voltage.

The above-threshold square-law dependence of the transistor saturation current on the gate voltage arises because increasing gate voltage increases the amount of mobile charge at the source linearly (as would be expected of a capacitor with a fixed voltage on its bottom plate), and the maximum lateral electric field in the channel also increases linearly with this increased charge yielding net square-law behavior for the drift current (drift current increases proportionately with more charge and with more electric field).

3.2 Intuitive model of MOS transistor operation

Figure 3.4 shows an intuitive model of the MOS transistor that is extremely useful in understanding its operation in both the subthreshold and above-threshold regimes. The transistor has a distributed oxide capacitance, modeled by the C_{ox} capacitors in the figure, and a distributed depletion capacitance due to bulk dopant charge, modeled by the C_{dep} capacitors in the figure. The surface potential varies along the length of the transistor and is comprised of the set of voltages at the junctions where capacitors meet in a distributed capacitive-divider-like configuration. At the source and drain ends of the channel, two diode-like elements represent pn junctions

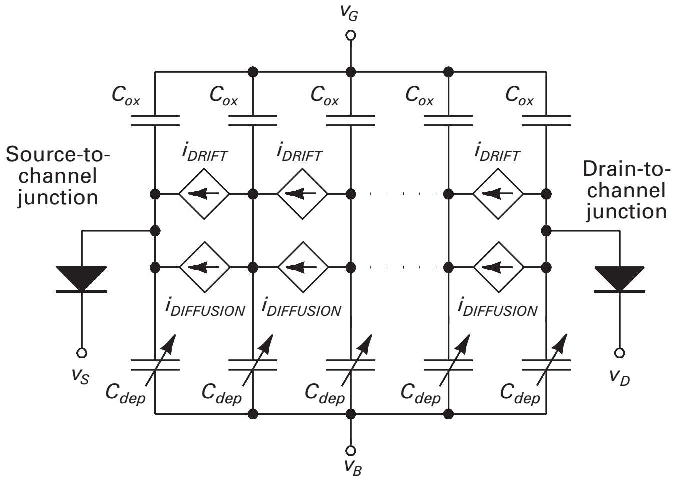


Figure 3.4. Intuitive model of an MOS transistor.

4 MOS device physics: practical treatment

Although this may seem a paradox, all exact science is dominated by the idea of approximation.

Bertrand Russell

In Chapter 3, we discussed an intuitive model to describe transistor operation shown in Figure 3.4. We will now use this intuitive model to simplify Equation (3.23) into more practical and insightful forms in subthreshold and above-threshold operation. To do so, we will use three approximations listed below.

1. **The $\boldsymbol{\kappa}$ approximation.** We approximate $\gamma\sqrt{\psi_S}$ by a constant term $\gamma\sqrt{\psi_{se}}$ around an operating point ψ_{se} plus a Taylor-series linear term $\left(\gamma/\sqrt{(2\psi_{se})}\right)(\psi_S - \psi_{se})$ to describe deviations from the operating point. This approximation is equivalent to modeling the distributed nonlinear depletion capacitance C_{dep}, shown in Figure 3.4 as a linear capacitance with some fixed value around a given ψ_{se}. The surface potential ψ_s varies with x in above-threshold operation, such that C_{dep} varies along the channel in Figure 3.4. We ignore this variation and assume one value for C_{dep} throughout the channel given by $\gamma C_{ox}/(2\sqrt{\psi_{se}})$ from Equation (3.11). The chosen operating point ψ_{se} is usually at the subthreshold to above-threshold transition where we have good knowledge of the surface potential and, in addition, is at the source channel end in source-referenced models. The value of $\kappa = C_{ox}/\left(C_{ox} + C_{dep}\right)$ is always between 0 and 1 and given by a capacitive-divider ratio. The κ approximation is useful in weak inversion and strong inversion. **The parameter $\boldsymbol{\kappa = 1/n}$ by definition, where $\boldsymbol{n}$ is termed the subthreshold slope coefficient in some texts**.
2. **The Taylor-series square-root approximation.** In weak inversion, we assume that the $\phi_t\exp(\ldots)$ term in $\gamma C_{ox}\sqrt{(\psi_S + \phi_t\exp((\psi_S - 2\phi_F - v_{CB})/\phi_t)))}$, an expression that describes the total charge at the source channel end or drain channel end, is much smaller than the ψ_S term in the expression. Therefore, we use a Taylor-series approximation for the square root again to evaluate the total charge in the expression. This approximation is useful in weak inversion only. Note that $v_{CB} = v_{SB}$ at the source end and $v_{CB} = v_{DB}$ at the drain end.
3. **The diode-clamp approximation.** In strong inversion, we will assume that the diode-like element of Figure 3.4 clamps the surface potential at the source end of the channel to $(2\phi_F + 6\phi_t) + v_{SB} = \phi_0^s + v_{SB}$ due to the operation of a strong negative-feedback loop. The surface potential at the drain end of the channel

is also clamped by a diode-like element to $\phi_0^s + v_{DB}$ if the transistor is not in saturation such that the diode-like element at the drain end is on. If the transistor is saturated, the diode-like element at the drain end is off and the drain surface potential is at a value determined by the capacitive divider formed by C_{ox} and C_{dep} in weak inversion. The mobile charge in strong inversion decreases along the channel as the bottom-plate surface potential on the distributed C_{ox} capacitors changes from a low value at the source end to a high value at the drain end, thus increasing the dopant charge and decreasing the total gate charge as we move from the source to the drain. In saturation, the mobile charge is assumed to abruptly go to zero at a pinchoff point along the channel very near the drain end where the mobile charge concentration is so low that the strong-inversion approximation no longer holds.

The diode-clamp approximation is only useful in strong inversion and is invalid in weak inversion. Weak-inversion operation occurs when the surface potential is less than or equal to $\phi_0^w + v_{SB} = 2\phi_F + v_{SB}$ at the source end of the channel. Note that we always use the w superscript for weak-inversion operation and the s superscript for strong-inversion operation. Thus, ϕ_0^w is the voltage across the diode-like elements above which weak-inversion operation is no longer valid while ϕ_0^s is the diode-clamp voltage across the diode-like elements in strong inversion, with $\phi_0^s - \phi_0^w = 6\phi_t$. In between the two extremes, where $\phi_0^w < \psi_S < \phi_0^s$ or equivalently, $2\phi_F < \psi_S < 2\phi_F + 6\phi_t$, we have moderate-inversion operation and neither the subthreshold exponential-junction nor the above-threshold diode-clamp approximation are valid.

We shall begin by discussing how the κ approximation is useful in multiple ways in subthreshold and above-threshold operation, especially when combined with the diode-clamp approximation. **The κ approximation could also be termed the n approximation since κ is defined to be $1/n$.**

4.1 The κ approximation

The κ approximation is useful in six different aspects of transistor operation, all of which may be derived from the intuitive model of Figure 3.4.

4.1.1 The capacitive divider of weak inversion

In weak inversion, κ is defined to be

$$\kappa = \frac{\partial \psi_S}{\partial v_G} \tag{4.1}$$

In weak inversion, the source and drain terminal voltages have little effect on the surface potential, which is constant throughout the channel and has a value

Table 4.1 Table summarizing source- and body-referenced equations

	Weak inversion	Strong inversion
$\kappa_x \equiv 1/n_x$	$i_{DS} = \mu\phi_t \frac{W}{L}(Q_{IL} - Q_{I0})$	$i_{DS} = \frac{\mu\kappa_s}{2C_{ox}}\frac{W}{L}\left(Q_{I0}^2 - Q_{IL}^2\right)$
Body-referenced	$i_{DS} = I_0 e^{\kappa_0 v_{GB}/\phi_t}\left(e^{-v_{SB}/\phi_t} - e^{-v_{DB}/\phi_t}\right)$ $I_0 = \mu C_{ox}\phi_t^2 \frac{W}{L}\left(\frac{1-\kappa_{sa}}{\kappa_{sa}}\right)e^{-\kappa_0 V_{T0}/\phi_t}$ $V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0}, \quad \boxed{\phi_0 = 2\phi_F}$ $\kappa_{sa} = \frac{1}{1+\frac{\gamma}{2\sqrt{\psi_{sa}}}}, \quad \kappa_0 = \frac{1}{1+\frac{\gamma}{2\sqrt{\phi_0}}} = \frac{1}{n_0}$ $\psi_{sa} = \left(-\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} + (v_G - V_{FB})}\right)^2$	$i_{DS} = \frac{\kappa_0 \mu C_{ox}}{2}\frac{W}{L}\left[\left(v_G - V_{T0} - \frac{v_{SB}}{\kappa_0}\right)^2 - \left(v_G - V_{T0} - \frac{v_{DB}}{\kappa_0}\right)^2\right]$ $V_{T0} = V_{FB} + \phi_0 + \gamma\sqrt{\phi_0}, \quad \boxed{\phi_0 = 2\phi_F + 6\phi_t}$ $v_{DSAT} = \kappa_0(v_{GB} - V_{T0}), \quad v_{SB} = 0$ $i_{DSAT} = \frac{\kappa_0 \mu C_{ox}}{2}\frac{W}{L}(v_{GB} - V_{T0})^2$
Source-referenced	$i_{DS} = I_{0S} e^{\kappa_s v_{GS}/\phi_t}(1 - e^{-v_{DS}/\phi_t})$ $I_{0S} = \mu C_{ox}\phi_t^2 \frac{W}{L}\left(\frac{1-\kappa_{sa}}{\kappa_{sa}}\right)e^{-\kappa_s V_{TS}/\phi_t}$ $i_{DS} = i_{DSAT}(1 - e^{-v_{DS}/\phi_t}), V_{DSAT} \approx 4\phi_t$ $V_{TS} = V_{T0} + \gamma\left(\sqrt{(\phi_0 + V_{SB})} - \sqrt{\phi_0}\right)$	$i_{DS} = \mu C_{ox}\frac{W}{L}\left[(v_{GS} - V_{TS})v_{DS} - \frac{v_{DS}^2}{2\kappa_s}\right]$ $V_{TS} = V_{T0} + \gamma\left(\sqrt{(\phi_0 + V_{SB})} - \sqrt{\phi_0}\right)$ $v_{DSAT} = \kappa_s(v_{GS} - V_{TS})$ $i_{DSAT} = \frac{\kappa_s \mu C_{ox}}{2}\frac{W}{L}(v_{GS} - V_{TS})^2$

curvature only if v_{GS} is below V_{T0}^w. The threshold voltage V_{T0}^s is greater than V_{T0}^w by approximately $6\phi_t/\kappa_0$.

The saturation current, I_T, at the edge of subthreshold operation at $v_{GB} = V_{T0}^w$, $v_{SB} = 0$, may be approximated by substituting an overdrive voltage of $(v_{GS} - V_{TS}) \approx \phi_t/\kappa_0$ in the above-threshold square-law relation given by Equation (4.46) with κ_0 assumed near 0.5. The form of Equation (4.34) then suggests that $I_{0,}$ the leakage current in digital circuits when $v_{GS} = 0$, may be estimated by attenuating I_T exponentially with one decade of attenuation every $(\phi_t/\kappa_0)\ln(10)$ V from $v_{GS} = V_{T0}^w$ down to $v_{GS} = 0$. For example, if κ_0 is near 0.6, we have nearly 100 mV per decade of attenuation; if $V_{T0}^s = 0.5$, $V_{T0}^w \approx 0.25$ V, and I_T is computed to be 1 μA, then the leakage current when $v_{GS} = 0$ is estimated by attenuating 1 μA over two-and-a-half 100 mV-decades to approximately 3.2 nA; if V_{T0}^s were 0.4 V, I_0 would be 32 nA. This example illustrates why leakage currents in modern semiconductor processes have been rising dramatically with reductions in threshold voltages. The speeds available in subthreshold operation have been rising dramatically as well.

4.6 Moderate inversion

An empirical equation for describing the current that is valid in weak inversion, moderate inversion, and strong inversion has been proposed and is often called

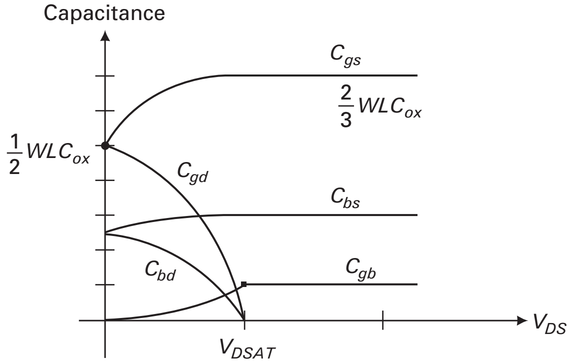


Figure 5.7. Dependence of various intrinsic capacitances on V_{DS} in an above-threshold MOSFET.

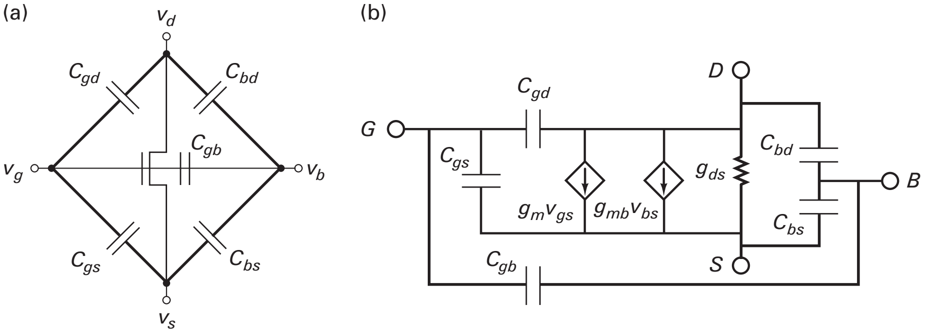


Figure 5.8a, b. The MOSFET "capacitance diamond" in (a) and complete small-signal model in (b).

$Q_{IL} = 0$, and the shape of the inversion charge distribution along the channel is a square-root function as revealed by Equation (5.33) and Figure 5.6. If we increase the source terminal voltage by an infinitesimal amount Δv_S, the diode-clamp mechanism raises the surface potential at the source end of the channel by Δv_S, which lowers the gate charge by $C_{ox}\Delta v_S$, and lowers the magnitude of the inversion charge by $\left(C_{ox} + C_{dep}\right)\Delta v_S$ at this end. This lowering of charge at the source end lowers the total charge in the channel. The *total* charge change along the channel is then given by

$$\begin{aligned} -\frac{\Delta Q_G^{TOT}}{\Delta v_s} &= \kappa_S \frac{\Delta Q_I^{TOT}}{\Delta v_s} \\ &= \kappa_S \left(\frac{C_{ox}}{\kappa_S} W \int_0^L \sqrt{1 - \frac{x}{L}} dx \right) \\ &= C_{ox} WL \int_0^1 \sqrt{z} dz \\ C_{gs} &= \frac{2}{3} WLC_{ox} \end{aligned} \tag{5.42}$$

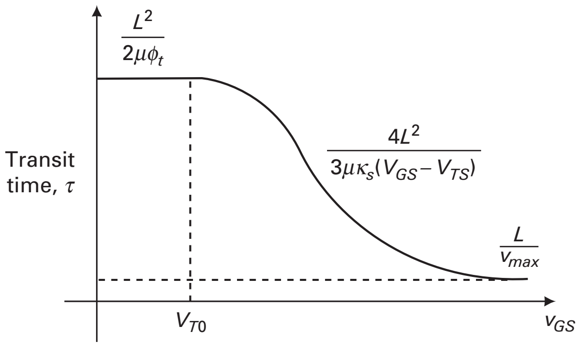


Figure 5.11. Transit time for a MOSFET in strong and weak inversion.

For v_{max} given by 10^5 m/s and $\mu = 500\,\text{cm}^2/\text{Vs}$ we find that L_c is 25 nm. The thermal velocity of electrons is given approximately by$\sqrt{kT/m_e}$, where m_e is the mass of the electron, and is about 0.67×10^5 m/s at 300 K. This discussion is highly oversimplified but provides an intuitive way of understanding why diffusive transport and subthreshold operation are getting increasingly faster. Chapter 6 will discuss, in a more rigorous fashion, interesting regimes of transistor operation in very-short-channel transistors where the charge density is relatively large but the mechanisms of transport are well described by thermal and scattering parameters in the transistor rather than by drift.

5.7 The 'beta' of an MOS transistor

Figure 5.12 (a) shows the small-signal equivalent of a source-and-bulk-grounded MOS transistor. The base (b) is analogous to the gate, the emitter (e) is analogous to the source, and the collector (c) is analogous to the drain. The input current $i_{gs}(s)$ through C_{gs} in Figure 5.12 (a) generates a current through the MOS g_m generator of $i_{gmm} = \left[g_m/\left(C_{gs}s\right)\right]i_{gs}(s)$ if we assume that C_{gb} may be neglected. The capacitances C_{gb} and C_{bd} are tied to the bulk terminal and effectively appear as grounded capacitances to the gate and drain respectively. Figure 5.12 (b) shows the small-signal equivalent of a bipolar transistor (BJT). The input current $i_b(s)$ through the r_π and $1/(C_\pi\ s)$ impedances generates a current through the bipolar g_m generator of $i_{gmb} = [g_m r_\pi/(r_\pi Cs + 1)]i_b(s)$. If we define $\beta_m(s) = i_{gmm}(s)/i_{gs}(s)$ to be the beta of an MOS transistor, analogous to $\beta_b(s) = i_{gmb}(s)/i_b(s)$, then we get

$$\begin{aligned}\beta_m(s) &= \frac{g_m}{C_{gs}s}\\ \beta_b(s) &= \frac{g_m r_\pi}{r_\pi C_\pi s + 1} = \frac{\beta_0}{r_\pi C_\pi s + 1}\end{aligned} \tag{5.58}$$

Figure 5.13 plots the 'beta' $\beta_m(s)$ of an MOS transistor and contrasts it with the beta $\beta_b(s)$ of a bipolar transistor with $C_\pi = C_{gs}$ and the g_m at the same value such that an intuitive understanding of the physics is more valuable to a circuit designer than an exact mathematical expression with empirically fit parameters.

When transistors are operated at very high frequencies, extrinsic parasitics such as the substrate and terminal resistances and diode-junction terminal capacitances can affect operation. At very high frequencies, finite transit-time effects caused by distributed charge within the transistor can cause its effective small-signal transconductance and capacitance values to become a function of frequency rather than being constant. We shall summarize such effects within the context of an EKV model, which is described in more detail in [3]. Since many RF systems operate well below the f_T of their devices, such distributed-charge effects are usually unimportant and may be neglected to a first approximation.

When transistor channel-length dimensions are comparable to a mean free length (10–15 nm typically), an electron shooting out of the source may suffer no collision with any impurity or lattice wave such that it is not scattered. The electron simply makes its way to the drain in a ballistic or scatter-free fashion. The limits of transistor transport when there is no scattering in the channel are referred to as ballistic transport [4], [5]. We shall discuss models for transistor operation that are ballistic and nearly ballistic, i.e., there are very few scattering events in the channel [6].

At low nanometer channel lengths, tunneling from the source to the drain can dominate the current flow. In fact, many advanced carbon nano tube (CNT) and single-electron transistors at nanometer lengths function via bidirectional quantum mechanical tunneling from source to drain and from drain to source [7], [8]. In conventional nanoscale transistors of the future, quantum-mechanical tunneling from the source to the drain will increase subthreshold current but reduce above-threshold current. Source-to-drain tunneling represents an ultimate limit on transistor performance [9].

Quantum-mechanical tunneling effects from the channel to the gate become significant when the gate-oxide thickness is lower than 3 nm. Gate-oxide leakage via quantum-mechanical tunneling is already significant in the transistors of today that do not have high-dielectric-constant insulators and therefore require a thin insulator to function. In fact, limiting gate-oxide tunneling is an important constraint for ensuring transistor scaling into the future [10].

We conclude by discussing the scaling of transistors in the future as their channel length decreases. We discuss four scaling laws that have been found to empirically represent the scaling of transistor bulk doping, supply voltage, oxide thickness, and threshold voltage with channel length [11]. It is expected that semi-classical transistor operation as described in this chapter will represent transistor characteristics well till channel lengths reach 10 nm [6].

6.1 The dimensionless EKV model

As in several fields of physics, the EKV model represents all transistor variables such as charge Q, current I, voltage V, and channel distance from source x in dimensionless

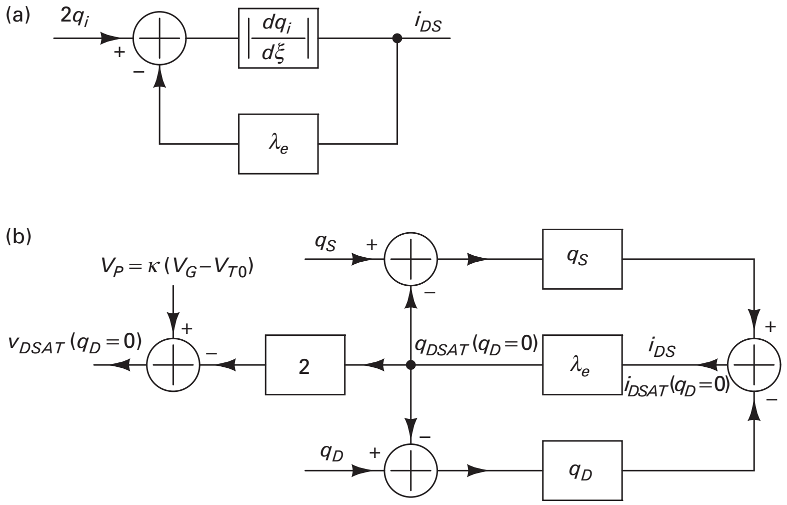


Figure 6.2a, b. (a) The feedback loop of velocity saturation present at every point in the channel. (b) The feedback loop of velocity saturation with q_S and q_D boundary charges.

$$\begin{aligned}
i_{DS} &= q_i \frac{-2\dfrac{dq_i}{d\xi}}{1-\dfrac{2}{2E_c}\dfrac{\phi_t}{L_{eff}}\dfrac{dq_i}{d\xi}} \\
i_{DS} &= 2q_i \frac{\left\|\dfrac{dq_i}{d\xi}\right\|}{1+\lambda_e\left\|\dfrac{dq_i}{d\xi}\right\|} \qquad (6.12)\\
\lambda_e &= \frac{\phi_t}{E_c L_{eff}}
\end{aligned}$$

Note that $dq_i/d\xi$ is negative since the unsigned dimensionless charge q_i decreases as ξ changes from 0 to 1, which is why we have used the absolute value in the second line of Equation (6.12). The parameter λ_e is a dimensionless electric field and is a measure of the strength of the short-channel effect. When L_{eff} is large and λ_e approaches 0, the current i_{ds} approaches its above-threshold value. Figure 6.2 (a) shows a feedback-loop interpretation of Equation (6.12). As λ_e increases, the feedback gain in Figure 6.2 (a) increases, attenuating the current from the above-threshold value given by just the feed-forward path to that given by Equation (6.12). Since i_{DS} is constant throughout the channel, the top line of Equation (6.12) is easily integrated as ξ ranges from 0 to 1 as we move from source to drain. We then get

$$i_{DS} = \frac{q_S^2 - q_D^2}{1+\lambda_e(q_S - q_D)} \qquad (6.13)$$

We notice that the above-threshold quadratic terms of Equation (6.5) have been modified by short-channel effects with a negative-feedback loop gain of voltages to get currents, which is not possible in equations with dimensions. The re-incorporation of the constants of Equation (6.1) together with expressions for Q_S and Q_D derived in the previous chapters yields the following set of body-referenced and source-referenced equations. The body-referenced equations correspond to our discussion thus far and the source-referenced equations are derived by straightforward extensions of them as in Chapters 3 and 4.

Body-referenced equations:

$$I_{DS} = \frac{\kappa\mu_0 C_{ox}}{2}\frac{W}{L_{eff}}\frac{(V_G - V_{T0} - \frac{V_S}{\kappa})^2 - (V_G - V_{T0} - \frac{V_D}{\kappa})^2}{1 + \dfrac{V_{DS}}{2E_c L_{eff}}}$$

$$I_{DSAT} = \frac{\kappa\mu_0 C_{ox}}{2}\frac{W}{L_{eff}}\frac{(V_G - V_{T0} - \frac{V_S}{\kappa})^2}{1 + \dfrac{\kappa(V_G - V_{T0}) - V_S}{2E_c L_{eff}}}$$

$$\boxed{\begin{array}{ll} I_{DSAT} = WC_{ox}\left(V_G - V_{T0} - \dfrac{V_S}{\kappa}\right)\nu_{sat}; E_c L_{eff} \to 0 \\ g_{mSAT} = WC_{ox}\nu_{sat}; \qquad E_c L_{eff} \to 0 \end{array}} \tag{6.25}$$

$$Q_{DSAT} = -C_{ox}\left(V_G - V_{T0} - \frac{V_S}{\kappa}\right)\frac{\dfrac{\kappa(V_G - V_{T0}) - V_S}{2E_c L_{eff}}}{1 + \dfrac{\kappa(V_G - V_{T0}) - V_S}{2E_c L_{eff}}}$$

$$V_{DSAT} - V_S = \frac{\kappa(V_G - V_{T0}) - V_S}{1 + \dfrac{\kappa(V_G - V_{T0}) - V_S}{2E_c L_{eff}}}$$

Source-referenced equations:

$$I_{DS} = \mu_0 C_{ox}\frac{W}{L_{eff}}\frac{(V_{GS} - V_{TS} - \frac{V_{DS}}{2\kappa_S})V_{DS}}{1 + \dfrac{V_{DS}}{2E_c L_{eff}}}$$

$$I_{DSAT} = \frac{\kappa_S\mu_0 C_{ox}}{2}\frac{W}{L_{eff}}\frac{(V_{GS} - V_{TS})^2}{1 + \dfrac{\kappa_S(V_{GS} - V_{TS})}{2E_c L_{eff}}}$$

$$\boxed{\begin{array}{ll} I_{DSAT} = WC_{ox}(V_{GS} - V_{TS})\nu_{sat}; E_c L_{eff} \to 0 \\ g_{mSAT} = WC_{ox}\nu_{sat}; \qquad E_c L_{eff} \to 0 \end{array}} \tag{6.26}$$

$$Q_{DSAT} = -C_{ox}(V_{GS} - V_{TS})\frac{\dfrac{\kappa_S(V_{GS} - V_{TS})}{2E_c L_{eff}}}{1 + \dfrac{\kappa_S(V_{GS} - V_{TS})}{2E_c L_{eff}}}$$

$$V_{DSAT} = \frac{\kappa_S(V_{GS} - V_{TS})}{1 + \dfrac{\kappa_S(V_{GS} - V_{TS})}{2E_c L_{eff}}}$$

6.10 Scaling of transistors in the future

Transistor channel length, L, is decreasing by ~1.4× in every 18-month technology revision today. On a semilog plot, transistor dimensions appear to obey an approximately constant geometric or exponential scaling law with time, as first observed by Moore [17] and now termed Moore's law. Moore's law is slowing as we reach fundamental physical limits that make well-controlled fabrication at small dimensions increasingly more expensive and difficult. The doping of the bulk N_A, supply voltage V_{DD}, oxide thickness t_{ox}, and threshold voltage V_{T0} of transistors appear to obey some simple empirical trends with L. These trends as postulated by Mead [11] are summarized below:

$$\begin{aligned} N_A &= (4 \times 10^{16})L^{-1.6} \\ V_{DD} &= 5L^{0.75} \\ t_{ox} &= \max(21L^{0.77}, 14L^{0.55}) \\ V_{T0} &= 0.55L^{0.23}, \end{aligned} \tag{6.52}$$

where L is in μm units, N_A is in cm^{-3} units, V_{DD} is in V units, t_{ox} is in nm units, and V_{T0} is in V units. The scaling of N_A and V_{DD} with L are necessary to limit punchthrough, DIBL, drain-junction breakdown, drain tunneling, and drain-conductance degradation; to ensure that κ is not too low and that the gate continues to have an effect on the surface potential, t_{ox} must then be lowered as well; the scaling of V_{T0} is based on the practicalities of doping implants and materials issues and to limit 'subthreshold leakage' in the off-state of digital transistors. Empirically, it is often found that

$$\sigma_{vth} = \frac{A_{vth}(L)}{\sqrt{WL}}, \tag{6.53}$$

with σ_{vth} being the standard deviation of random threshold-voltage mismatch across transistors in a process with width and length dimensions W and L respectively. A regression fit by the author of A_{vth} versus L data compiled in [18] reveals that

$$A_{vth}(L) = 17L^{0.71}, \tag{6.54}$$

where A_{vth} is in mV. The scaling of the threshold-voltage mismatch in Equation (6.53) with transistor dimensions is similar to the scaling of $1/f$ noise in transistors. As we discuss in Chapter 7, $1/f$ noise may be viewed as a dynamically varying threshold voltage caused predominantly by the occurrence of random trap-filling processes in the oxide of the transistor. Equation (6.54) can be viewed as an addition to the four scaling laws of Equation (6.52).

Some ultimate limits of scaling based on the Shannon limit of energy dissipation per bit ($kT \ln 2$ as explained in Chapter 22), tunneling from source to drain (~1 nm), and quantum channel transit times ($L/v_{quantum}$) are discussed in [9] and [1]. They predict a power dissipation of 3.7 MW/cm^2 if we continue scaling as usual. However, as is correctly pointed out, it is unlikely that more than 100 W/cm^2

where ${Q_I}^{TOT}$ is the total electron charge, and $\overline{Q_I}$ is the mean charge concentration per unit area across the whole channel. That is, the noise within a subthreshold MOS transistor may be viewed as being the same as that of an equivalent sheet resistor with charge concentration per unit area of $\overline{Q_I}$.

7.4 Unity between thermal noise and shot noise

Thermal noise and shot noise are often mistakenly viewed to be different forms of white noise, i.e., noise with a flat power spectrum. However, our discussion and results suggest that

1. If there is no increase in variance due to large-electric-field effects, a good approximation is to assume that only diffusion currents cause noise in all electronic devices, and that drift currents are noiseless.
2. Shot noise is fundamental and due to the mathematics of Poisson processes.
3. *Thermal noise is shot noise due to internal thermally generated diffusion currents in physical devices*. The latter statement is true, to good approximation even if the dominant external current in the device is due to drift as long as the electric fields in the devices are such that the drift velocity is well below the thermal velocity.

This method of viewing white noise in electronic devices helps one avoid double counting shot noise and thermal noise sources and views them as different ways of expressing the same white-noise source. Consider the example shown in Figure 7.10 where we have a modest dc voltage V_{DC} across a resistor and we would like to know the noise current variance in this situation. Is the current noise given by

$$\left[2q\left(\frac{V_{DC}}{R}\right) + 4kTG\right]\Delta f? \tag{7.28}$$

Such a view might arise because we get shot noise from the dc current flow and thermal noise from the fact that it is a resistor and exhibits Nyquist-Johnson noise. In fact, measurements show that, if V_{DC} is a reasonable commonly used voltage of a few volts such that the electric fields in the resistor are not so large as to cause hot electron effects or drift velocities comparable to thermal velocities, then the noise in the resistor is simply $4kTG\Delta f$ and is independent of V_{DC}. The internal diffusion currents which cause noise in the resistor are unchanged by V_{DC} and

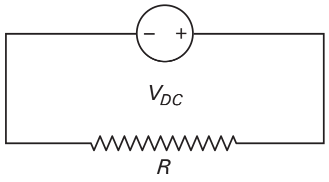


Figure 7.10. Finding the noise of a resistor R with a dc voltage source V_{DC} connected across its terminals.

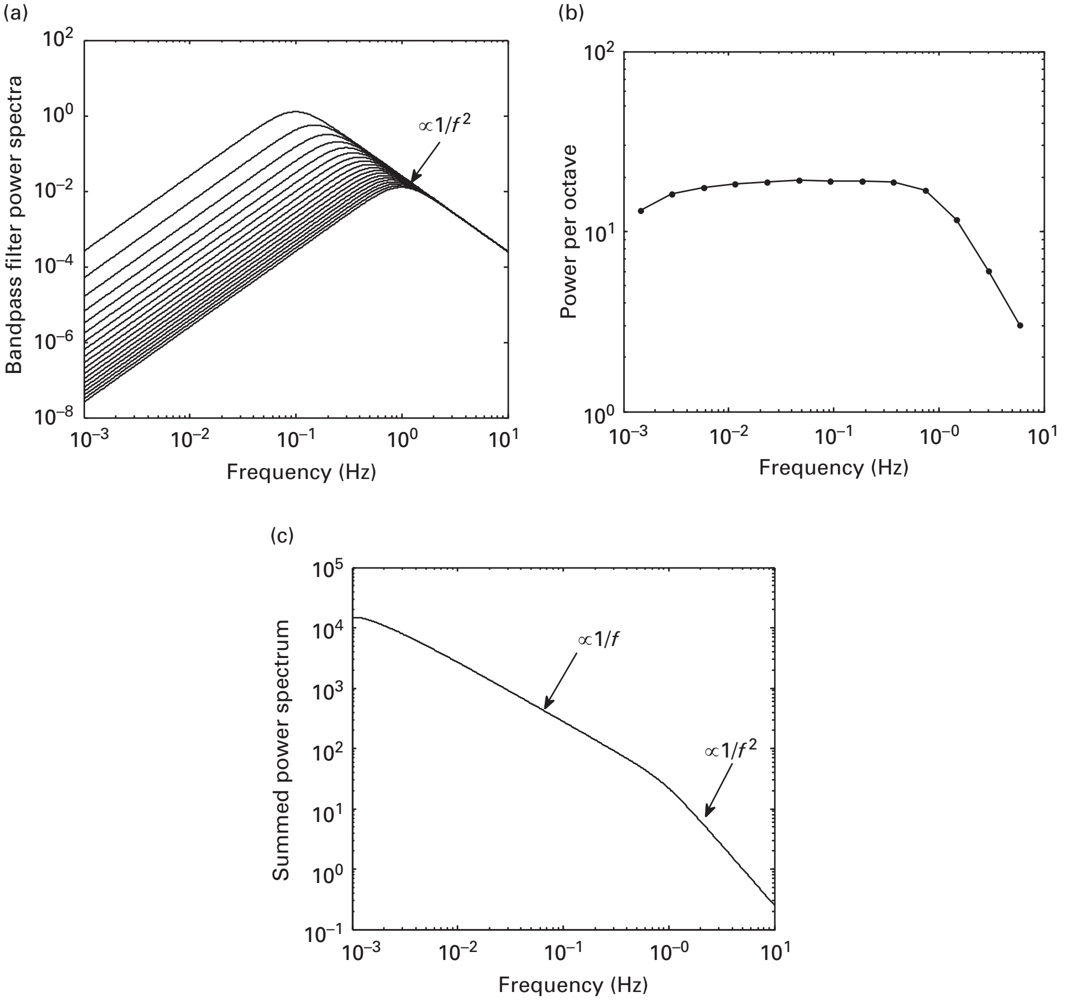


Figure 7.15a, b, c. Summing up the response of many constant-Q bandpass filters (with uniformly distributed center frequencies) produces a $1/f$ power spectrum. Individual bandpass filter responses are shown in part (a). Part (b) shows that the sum of these responses has constant power per octave (or any other constant frequency ratio). Part (c) plots the summed response and shows that it has a $1/f$ power spectrum.

7.8 Some notes on 1/*f* noise

We shall now itemize some points of note regarding $1/f$ noise:

1. Transistors that are pFETs usually have much less $1/f$ noise than nFETs presumably because electrons need much more energy to enter the oxide in pFETs than in nFETs: the mean energy of electrons in the channel of a pFET is near the bottom of the valence band rather than near the top of the conduction band in nFETs. Thus, many low-noise circuits are built with pFET transistors in their very first input-sensing gain stage.

Since the f_T of transistors is lower in the subthreshold regime rather than in the above-threshold regime, the flicker-noise corner frequency is lower in the subthreshold regime than in the above-threshold regime. Therefore, it is common to see increasingly more $1/f$ noise in a circuit as its mode of operation is changed from the subthreshold regime to the above-threshold regime. Figure 12.15 in Chapter 12 reveals experimental measurements of this phenomenon.

9. The complete small-signal noise model that incorporates white noise and $1/f$ noise into the noise generator of Figure 7.9 and that is valid in both the linear and saturation regions of operation is listed below.

$$
\begin{aligned}
\overline{i_n^2} &= \left[2qI_{sat}\left(1+e^{-V_{DS}/\phi_t}\right)+\frac{qB_{imp}}{C_{ox}^2WL}\frac{I_{sat}^2\kappa_s^2\left(1-e^{-V_{DS}/\phi_t}\right)}{\phi_t^2}\frac{1}{f}\right]\Delta f \quad \leftarrow \quad \text{Subthreshold} \\
\overline{i_n^2} &= \left[4kT\mu\frac{W}{L}C_{ox}(V_{GS}-V_{TS})\frac{2}{3}\left(\frac{1+\eta+\eta^2}{1+\eta}\right)\right. \\
&\quad \left.+\frac{qB_{imp}}{C_{ox}^2}\frac{2\kappa\mu C_{ox}I_{sat}}{L^2}\left(1-\eta^2\right)^2\frac{1}{f}\right]\Delta f \quad \leftarrow \quad \text{Above threshold}
\end{aligned}
\tag{7.40}
$$

10. In deep submicron devices where oxide capacitances are small, it is possible to see individual filling and emptying of traps such that $1/f$ noise manifests as discrete jumps in current noise or input-referred voltage noise in a random telegraph fashion.
11. It is possible to reduce $1/f$ noise by cycling the gate of a MOSFET between accumulation, which empties the trap since it is energetically unfavorable for the electron to remain in it, and regular-inversion operation, which is not exercised long enough for the trap to refill [9]. Other techniques for reducing $1/f$ noise and transistor threshold-voltage mismatch effects, which are the dc manifestation of $1/f$ noise from extremely slow-time-constant traps, have been reviewed in [10]. These include autozeroing, correlated double sampling, and chopper stabilization. In Chapter 8, we shall discuss the technique of lock-in amplification, upon which chopper stabilization is based, which also mitigates the effects of $1/f$ noise.

7.9 Thermal noise in short-channel devices

In short-channel devices that have high lateral and vertical electric fields, the phenomena of velocity saturation, vertical mobility reduction, carrier heating, and channel-length modulation all conspire to affect thermal noise in above-threshold operation. These phenomena are described in Chapter 6 and a review of this chapter is helpful for understanding this section. The thermal noise in subthreshold operation is largely unaffected so we shall focus only on above-threshold operation. In the velocity-saturated above-threshold regime, electrons near the drain end of the channel are at a drift velocity v_{sat} that is near their

γ can almost double and in some reported measurements can be even as high as 10. High values of V_{GS} also help attenuate noise by making Q_S and Q_D more equal, thus reducing lateral electric fields and all the associated effects of velocity saturation.

7.10 Thermal noise in moderate inversion

Equation (7.27) for the thermal noise is valid in all regions of operation as long as electric fields are not high enough to create hot electrons. Thus, it applies in weak inversion, moderate inversion, and strong inversion. In weak inversion and saturated operation, it can be evaluated as

$$\overline{\Delta I_{ds}^2} = 2qI_{DSAT}\Delta f \tag{7.41}$$

In strong inversion and saturated operation, it can be evaluated as

$$\overline{\Delta I_{ds}^2} = 4kT\left(\frac{2}{3}\frac{g_{msat}}{\kappa_S}\right)\Delta f \tag{7.42}$$

In moderate inversion, there is no exact closed-form expression for Equation (7.27) in terms of the terminal voltages of the transistor. Nevertheless, in the spirit of the empirical EKV equation of Chapter 4 (Equation 4.48), we can invent a formula that interpolates between the two regimes:

$$\boxed{\overline{\Delta I_{ds}^2} = \frac{2qI_{DSAT}}{1+\dfrac{(3/4)\kappa_s(V_{GS}-V_{TS})}{2\phi_t}}\Delta f} \tag{7.43}$$

Figure 7.17 reveals that simulations of the thermal noise in a transistor of 200 μm width and 2 μm length in a UMC 0.13 μm process are consistent with

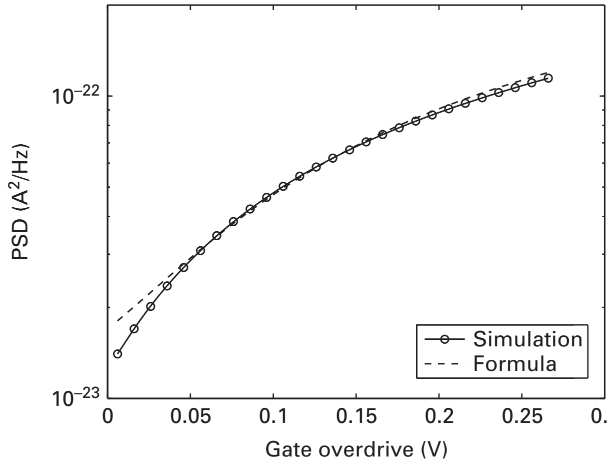


Figure 7.17. Thermal noise in moderate inversion.

8 Noise in electrical and non-electrical circuits

When we tug on a single thing in nature, we find it attached to everything else.

John Muir

Devices that dissipate energy, such as resistors and transistors, always generate noise. This noise can be modeled by the inclusion of current-noise generators in the small-signal models of these devices. When several such devices interact together in a circuit, the noise from each of these generators contributes to the total current or total voltage noise of a particular signal in a circuit. In this chapter, we will understand how to compute the total noise in a circuit signal due to noise contributions from several devices in it. We shall begin by discussing simple examples of an RC circuit and of a subthreshold photoreceptor circuit. We shall see that the noise of both of these circuits behaves in a similar way. We shall discuss the equipartition theorem, an important theorem from statistical mechanics, which sheds insight into the similar noise behavior of circuits in all physical systems. We shall then outline a general procedure for computing noise in circuits and apply it to the example of a simple transconductance amplifier and its use in a lowpass filter circuit. We will then be armed with the tools needed to understand and predict the noise of complicated circuits, and to design ultra-low-noise circuits.

We shall conclude by presenting an example of an ultra-low-noise micro-electro-mechanical system (MEMS), a capacitance-sensing system capable of sensing a 0.125 parts-per-million (ppm) change in a small MEMS capacitance (23-bit precision in sensing) [1]. This system will help us understand how mechanical noise and electrical circuit noise both determine the minimum detectable signal of the sensor. In fact, we will see that noise in non-electrical systems can be treated and understood with the same fundamental concepts that we have used in electrical systems. In Chapter 24, we will see that such noise concepts can be extended to understand noise in chemical and molecular systems in biology as well.

The example will also help us understand the important technique of lock-in detection, an ingenious technique for removing non-fundamental $1/f$ noise in circuits such that detection limits are only set by fundamental thermal noise. Lock-in detection techniques require the use of multiplier or mixer circuits. Lock-in amplifiers that implement these multipliers with passive switch-based mixers are often referred to as chopper-modulated amplifiers. The lock-in techniques will be presented in the context of a vibration sensor. However, such techniques can be adapted for several applications including DNA and bio-molecular detection in cantilever-based MEMS systems or in low-power Electroencephalogram (EEG) amplifiers [2], [3].

8.1 Noise in an RC lowpass-filter circuit

Figure 8.1 (a) shows a simple RC circuit. Figure 8.1 (b) shows the same circuit with the resistor replaced with a noise generator of $4kTRf$ in series with it. Figure 8.1 (c) is the circuit that we use for computing the noise, obtained by grounding v_{IN}. When computing noise, we ground all input voltage sources and open all input current sources such that we can focus on the noise of the circuit even when no input is present. We find that:

$$\begin{aligned}
v_{out}(s) &= \frac{v_{in}(s)}{\tau s+1}, \qquad \tau = RC = \frac{1}{2\pi f_c}, \qquad s = \jmath\omega \\
\overline{v_{out}^2}(\omega) &= \frac{\overline{v_{in}^2(\omega)}}{|1+\jmath\omega\tau|^2} = \frac{\overline{v_{in}^2(\omega)}}{1+\omega^2\tau^2} \\
\overline{v_{out}^2} &= \int_0^\infty \frac{\overline{v_{in}^2(f)}df}{1+(2\pi f)^2/(2\pi f_c)^2} \\
&= \int_0^\infty \frac{4kTRdf}{1+(f/f_c)^2} \\
&= 4kTR\int_0^\infty \frac{f_c du}{1+u^2} \\
&= 4kTRf_c \tan^{-1}(u)\Big|_0^\infty \\
&= 4kTR \times \frac{1}{2\pi RC} \times \frac{\pi}{2} \\
\overline{v_{out}^2} &= \frac{kT}{C}
\end{aligned} \tag{8.1}$$

The calculations of Equation (8.1) illustrate how we calculate the output voltage noise measured at the capacitor of the RC circuit. We first find the transfer function from the noise generator to the output as though it were a regular input. Then, we evaluate the mean-square output noise per unit bandwidth between frequencies f and $f + df$ by multiplying the mean-square input noise per unit bandwidth due to the noise generator between these frequencies and the square of the magnitude of the transfer function that we have just computed at frequency f. Finally, we integrate the mean-square output noise over all frequencies from 0 to ∞ to compute the total noise to be kT/C.

The surprisingly simple kT/C result arises from the fact that the noise per unit frequency bandwidth is $4kTR$, the bandwidth is $1/RC$ in ω units or $1/(2\pi RC)$ in

We have glibly assumed that the noise generators from different devices in a circuit are all independent and uncorrelated when making our calculations. This assumption may seem strange since each generator affects the circuit's voltages and currents. Therefore each noise generator alters the operating-point parameters of the whole circuit, thereby affecting the noise produced by all the other generators. Thus, the noise from various generators ought to be correlated. The key to this paradox is *small signal*. We assume that the net effect of all the noise generators is still small enough such that the operating-point parameters of the circuit are barely changed. To first order then, all the noise generators remain uncorrelated because the electrons in one device flow without regard to electrons in other devices and the noise from one device does not alter the operating-point parameters of another device.

In certain RF circuits it may be necessary to include induced gate noise in an MOS device with inherent correlation between a drain-to-source current noise generator and a gate-to-source current noise generator in the elementary noise model of the device itself (see Equations (7.45) and (7.46) in Chapter 7). The induced gate-noise generator is only of importance when the operating RF frequencies are very near the f_T of the device, a bad strategy for ultra-low-power design, as we discuss in Chapter 22. Thus, we shall not focus much on the latter noise generator in this book. It is discussed extensively in other texts, e.g. [7]. In such cases, one can use Equation (7.19) in Chapter 7 to account for correlations, and refer back to the actual input of the system at the very end.

8.6 An ultra-low-noise MEMS capacitance sensor

Figures 8.14 (a) and 8.14 (b) illustrate the principle of lock-in signal detection in the context of a MEMS capacitance sensor, similar to the one used in several commercial accelerometers and vibration sensors. The capacitances C_{s1} and C_{s2} are differential capacitances that differ from one another ever so slightly, with one increasing by ΔC when the other decreases by ΔC and vice versa. The differential nature of the capacitance arises because the V_x node is formed by a mechanically mobile plate common to both capacitances: in response to a mechanical linear acceleration in one direction, the mobile plate moves towards the static plate of C_{s1} and away from the static plate of C_{s2}; in response to a mechanical linear acceleration in the other direction, the mobile plate moves towards the static plate of C_{s2} and away from the static plate of C_{s1}. The mobile plate of mass m is tethered via a spring of stiffness k to a substrate, such that the mechanical equilibrium displacement of the mobile plate in response to an acceleration a is $(ma)/k$. For the relatively small accelerations that we shall be considering, the sensor may be assumed linear with the fractional capacitance change of either capacitance proportional to its fractional displacement. For a typical k/m of $2\pi \times 20$ kHz, a few mg of acceleration will cause displacements on the order of 0.01 Å, significantly less than the nominal $2\,\mu$m plate separation. Thus, we need to detect

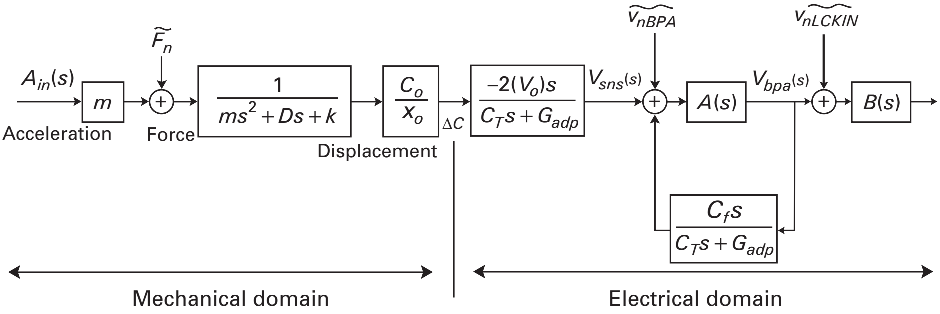


Figure 8.18. Illustration of a block diagram for noise analysis in both the electrical and mechanical domains.

8.6.2 Noise analysis

Figure 8.18 reveals the transfer function for the complete mechano-electrical system with mechanical and electrical noise sources. The noise input $\widetilde{F_n}$ represents the power spectral density of the mechanical force noise, $\widetilde{v_{nBPA}}$ represents the input-referred power spectral density of the bandpass preamplifier with transfer function $A(s)$, and $\widetilde{v_{nLCKIN}}$ represents the input-referred power spectral density of the lock-in amplifier with transfer function $B(s)$.

The mechanical force noise PSD $\widetilde{F_n}$ is related to the mechanical damping D just as the electrical current noise PSD is related to the electrical conductance, i.e.,

$$\widetilde{F_n^2} = 4kTD, \tag{8.29}$$

analogous to

$$\widetilde{i_n^2} = 4kTG \tag{8.30}$$

In fact, the transfer function from force to displacement shown in Figure 8.18 is identical to that of a parallel LCR circuit with L analogous to $1/k$, C analogous to m, D analogous to G, the force analogous to an input current, and the velocity analogous to a voltage. Thus, the mechanical quality factor or Q of the system is given by

$$Q = \frac{R}{\sqrt{L/C}} = \frac{RC}{\sqrt{LC}} = \frac{\omega_{res}m}{D} \tag{8.31}$$

Since ω_{res}, m, and Q are easily measured for a spring-mass system, we can measure D in terms of Q. From Figure 8.18, the input-referred mechanical acceleration noise PSD in units of g, the gravitational acceleration, are then given by dividing Equation (8.29) by m^2g^2. Thus, the input-referred acceleration noise PSD due to mechanical noise is given from Equations (8.29) and (8.31) to be

$$\widetilde{a_{mech}}^2 = \frac{4kT\omega_{res}}{Qm(9.8)^2}\ \mathrm{g^2/Hz} \tag{8.32}$$

feedback path gain and major loop gain are at their highest values such that a good phase margin for this case guarantees a good phase margin for other cases where the feedback path gain has an attenuating value, and the major loop gain and crossover frequency are consequently lower. The shape of the net transmission is controlled by a well-defined passive component C_c and the minor loop has about 90° of phase margin as well.

9.4 The closed-loop two-pole τ-and-Q rules for feedback systems

Frequently, well-behaved loop transmissions in circuit design can be approximated with a feedback loop that has two time constants and a dc loop gain A_{lp} especially near their crossover frequency:

$$L(s) = A_{lp}\left(\frac{1}{\tau_{big}s + 1}\right)\left(\frac{1}{\tau_{sml}s + 1}\right) \tag{9.12}$$

The closed-loop transfer function of a system with this loop transmission entirely in its feedforward path and with a unity-gain feedback path is then given by

$$\begin{aligned} H_{cl}(s) &= \frac{L(s)}{1 + L(s)} \\ &= \frac{\left(\dfrac{A_{lp}}{A_{lp} + 1}\right)}{\dfrac{\tau_{big}\tau_{sml}}{A_{lp} + 1}s^2 + (\tau_{big} + \tau_{sml})s + 1} \end{aligned} \tag{9.13}$$

Equation (9.13) can be rewritten in a form that describes the closed-loop system as a second-order system with its transfer function in a canonical form

$$H_{cl}(s) = \frac{A_{cl}}{\tau_{cl}^2 s^2 + \dfrac{\tau_{cl}s}{Q_{cl}} + 1} \tag{9.14}$$

By performing some simple algebra on Equation (9.13), the values of A_{cl}, τ_{cl}, and Q_{cl} in Equation (9.14) can be found to be

$$\begin{aligned} A_{cl} &= \frac{A_{lp}}{A_{lp} + 1} \\ \frac{1}{\tau_{cl}} &= \omega_n = \sqrt{\frac{1 + A_{lp}}{\tau_{sml}\tau_{big}}} \\ Q_{cl} &= \frac{\sqrt{(1 + A_{lp})}}{\sqrt{\dfrac{\tau_{big}}{\tau_{sml}}} + \sqrt{\dfrac{\tau_{sml}}{\tau_{big}}}} \end{aligned} \tag{9.15}$$

9.9 The ‘fake label’ circuit-analysis trick

1. We first re-label all dependent sources with new and unique names, e.g., i_{x1}, v_{yz}, etc. We erase their dependencies on their control variables.
2. The circuit cannot tell the difference between a dependent source with a ‘fake label’ and an independent source with a ‘real label’. A current source is a current source, independent of what it is called.
3. Superposition now applies to all sources! Find via superposition the values of all control variables as a function of the independent sources and the newly independent sources. Find via superposition the values of any desired output variables as a function of the independent sources and the newly independent sources. It’s acceptable to short or open the newly independent sources as is customary during superposition. Form a block diagram.
4. Now reinsert the dependencies of the newly independent sources on their control variables to make them dependent again and thus complete a feedback block diagram.
5. Analyze the feedback block diagram via block-diagram simplifications and Black’s formula and obtain insight into the circuit.

Figures 9.22 (a) and (b) illustrate why this trick works. We are essentially breaking feedback loops involving the control variables (Figure 9.22 (a)) and then reforming them (Figure 9.22 (b)) by removing dependencies and then reinserting them. Superposition applies at the adder block in Figure 9.22 (a). The figures show how control variables of dependent generators can lead to feedback loops. Once the values of the dependent generators have been computed by this trick, the independent and dependent generators values can be used to determine any output variables of interest, once again by superposition at the adder as in Figure 9.22 (c).

9.10 A circuit example

Figure 9.23 shows the small-signal circuit for a source-degenerated single-transistor amplifier. We’d like to find the input-output transfer function $v_{out}(s)/v_{in}(s)$ using the strategy described above. Since v_{in} sets the value of v_g, v_s is the only unknown control variable that determines the value of the g_m and g_{mb} generators. Thus, we label the g_m and g_{mb} generators with fake labels i_{x1} and i_{x2} respectively to make them independent sources as shown in Figure 9.24 (a) and compute via superposition the value of v_s as a function of v_{in}, i_{x1}, and i_{x2}. We also compute v_{out} as a function of the same three sources by superposition. Since i_{x1} and i_{x2} are in parallel with each other, their transfer functions to both v_s and v_{out} are identical.

The superposition subcircuit obtained by opening i_{x1} and i_{x2} is shown in Figure 9.24 (b) while that obtained by shorting v_{in} is shown in Figure 9.24 (c). From these subcircuits, we can see by inspection that

We shall introduce return-ratio analysis by exploiting the fake-label concept described in Chapter 9. Bode himself did not describe return-ratio analysis with fake labels per se, although they are implicit in his description. We shall then compute the return ratio, R, for dependent generators and passive impedances and apply them to create formulas for computing transfer functions in circuits in terms of return ratios. We shall show that the robustness of a circuit to parameter variations in one of its elements is proportional to the reciprocal of the return difference, $1-R$, of that element; thus, the return ratio of an element is a measure of the robustness of the circuit to variations in this element's parameters. After providing three examples of application of return-ratio analysis in an inverting operational-amplifier circuit, a resistive-bridge circuit, and a bridged-T network, we present Blackman's impedance theorem, a special case of the return-ratio formulas, useful for computing impedances in circuits. We show examples of application of Blackman's impedance formula to a cascode impedance, a cascode impedance with a resistive load, and driving-point impedances of single-transistor circuits. Then, we discuss Middlebrook's extra-element theorem and illustrate its application to an example. We then show that Thevenin's theorem is also a special case of the return-ratio formulas. We conclude the chapter with two examples that illustrate the application of return-ratio analysis. The first example shows how return-ratio techniques may be used in a hierarchical fashion to analyze ever-more complex circuits built up on simpler circuits that have themselves been analyzed by return-ratio techniques. The second example analyzes a super-buffer circuit via return-ratio techniques and shows the equivalence of return-ratio techniques to more conventional feedback-block-diagram techniques such that the two may be compared.

10.1 Return ratio for a dependent generator

Figure 10.1 illustrates a dependent generator in a big linear circuit that has been replaced with a fake label that allows us to pretend temporarily that it is an independent generator of value $i_{fakelabel}$. The transfer function from the independent generator to the dependent generator's control variable v_ε is given by $Z_{fakelabel}$, which is a transfer impedance due to the other elements of the circuit. Thus, we may write

$$v_\varepsilon = i_{fakelabel} Z_{fakelabel} \tag{10.1}$$

However, in reality, if we put the dependency back in

$$i_{fakelabel} = -g_m v_\varepsilon, \tag{10.2}$$

we can define a return ratio R_{dep} for the dependent generator to be

$$R_{dep} = -g_m Z_{fakelabel}, \tag{10.3}$$

the g_{m2} generator and rapidly compute the output impedance with Blackman's formula. As one gets more practiced, return-ratio techniques can be done mentally, without the need for an explicit feedback block diagram.

10.11 Summary of key results

Since several inter-related ideas were presented in this chapter, it is worth summarizing the five key ideas derived from them:

1. The return ratio of a dependent generator is $-g_m Z_{fakelabel}$. The return ratio of a parallel passive impedance Z is $-Z_{fakelabel}/Z$ and the return ratio of a series passive impedance Z is $-Z/Z_{fakelabel}$. Return ratios are assumed negative, hereafter.
2. The transfer function of an element with a varying parameter g_m such as the g_m of a dependent generator or the Z of a passive impedance is given by

$$\begin{aligned} TF_{gm} &= TF_0\left(\frac{1+R_{outputnulled}}{1+R_{inputnulled}}\right) \\ &= \frac{TF_0}{1+R_{inputnulled}} + TF_\infty\left(\frac{R_{inputnulled}}{1+R_{inputnulled}}\right) \end{aligned} \tag{10.67}$$

$$TF_0 R_{outputnulled} = TF_\infty R_{inputnulled}$$

3. Middlebrook's extra-element theorem for a parallel passive impedance Z is given by

$$TF_Z = TF_{open}\left(\frac{1}{1+\dfrac{Z_d}{Z}}\right) + TF_{short}\left(\frac{\dfrac{Z_d}{Z}}{\dfrac{Z_d}{Z}+1}\right) \tag{10.68}$$

where Z_d is the impedance across the element. The extra-element theorem for a series passive impedance Z is given by

$$TF_Z = TF_{short}\left(\frac{1}{1+\dfrac{Z}{Z_d}}\right) + TF_{open}\left(\frac{\dfrac{Z}{Z_d}}{\dfrac{Z}{Z_d}+1}\right) \tag{10.69}$$

4. Blackman's formula for a driving-point impedance in a circuit is given by

$$Z_{gm} = Z_{gm=0}\left(\frac{1+R_{nodeshorted}}{1+R_{nodeopen}}\right), \tag{10.70}$$

where $R_{nodeshorted}$ and $R_{nodeopen}$ are return ratios for an element with varying parameter g_m computed when the impedance node of interest is shorted or opened respectively.
5. Thevenin's theorem is a special case of return-ratio analysis.

Section II

Low-power analog and biomedical circuits

11 Low-power transimpedance amplifiers and photoreceptors

In the study of this membrane [the retina] . . . I felt more profoundly than in any other subject of study the shuddering sensation of the unfathomable mystery of life.

Santiago Ramón y Cajal

With this chapter, we begin our study of circuits by understanding a classic feedback topology used to sense currents and convert them into voltage, i.e., transimpedance amplifiers. Transimpedance amplifiers are widely used in several sensor and communication applications including microphone preamplifiers, amperometric molecular and chemical sensors, patch-clamp amplifiers in biological experiments, photoreceptors for optical communication links, barcode scanners, medical pulse oximeters for oxygen-saturation measurements, and current-to-voltage conversion within and between chips. After a brief introduction to transimpedance amplifiers, we shall focus on a specific application, i.e., the creation of a photoreceptor, a transimpedance amplifier for sensing photocurrents on a chip. The photoreceptor example will serve as a good vehicle for concretely illuminating several issues that are typical in sensors and transimpedance-amplifier design. It will also serve as a good application for illustrating how the fundamentals of device physics, feedback systems, and noise affect the operation of real circuits and systems.

The photoreceptor discussed here was inspired by the operation of photoreceptors in turtle cones and was first described in [1]. A photoreceptor similar to the one described in this chapter is used to create the low-power pulse oximeter described in Chapter 20 and in [2]. The photodiode basics described here are also useful in understanding how low-power imagers (see Chapter 19) and solar cells (see Chapter 26) work. Transimpedance amplifiers are useful in understanding how biomolecular amperometric sensors and electrochemical sensors, which are briefly discussed in Chapter 20, work. Many of the feedback principles described in this chapter, such as the two-pole Q approximation and the two-pole time-constant rule, are useful in the design of other feedback systems as well.

11.1 Transimpedance amplifiers

Figure 11.1 (a) illustrates a classic transimpedance amplifier topology. A current i_{IN} at a node with capacitance C_{in} is sensed and converted to a voltage

To understand the tutorial example of this chapter, i.e., a photoreceptor that senses photocurrents via a transimpedance amplifier topology, it is useful but not strictly necessary to understand how photocurrents are generated in silicon. Thus, we shall begin by reviewing some background material on the transduction of light to electrons. Readers unfamiliar with basic semiconductor device physics or readers only interested in transimpedance amplifiers may skip the following section. They will then need to abstract a photodiode as a device that generates a photocurrent proportional to the photon current or equivalently to the light intensity striking it.

11.2 Phototransduction in silicon

Figure 11.2 (a) shows a pn junction and Figure 11.2 (b) reveals an energy band diagram of the pn junction with zero voltage across it. The Fermi level, marked E_F, represents the average energy of an electron, which at zero voltage is the same all along the junction. The lowermost energy level of the conduction band (the conduction band edge) and the uppermost energy level of the valence band (the valence band edge) are drawn in bold wavy lines in the figure. They maintain a constant energy difference between each other, the bandgap energy, while the absolute values of these energies change due to the built-in depletion region or equivalently the built-in potential of the junction. In the n-type region, the Fermi level is near the bottom of the conduction band edge reflecting the fact that lots of donor dopant atoms have contributed to creating a large population of electrons in the conduction band. In the p-type region, the Fermi level is near the top of the valence band edge reflecting the fact that lots of acceptor dopant atoms have contributed to creating a large population of holes in the valence band.

An optical photon with an energy higher than the bandgap energy (1.12 eV in silicon) can create an electron-hole pair when it strikes a silicon atom: the photon promotes an electron from a valence-band energy level to a conduction-band

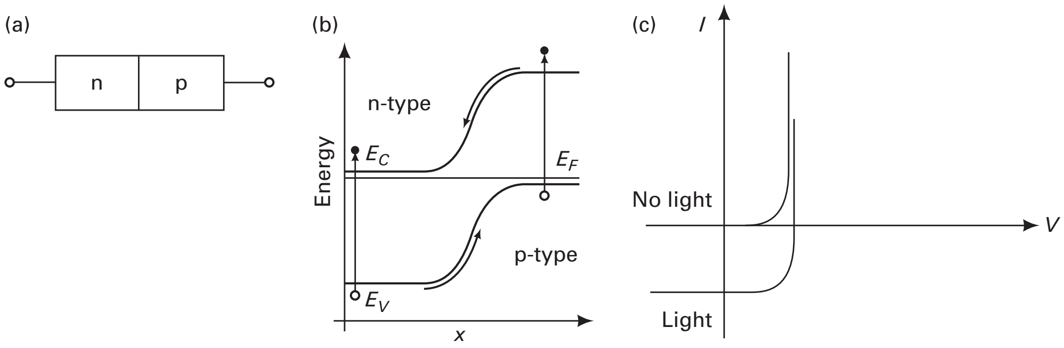


Figure 11.2a, b, c. An pn junction (a), its energy band diagram (b), and its $I-V$ curve with and without the presence of light (c).

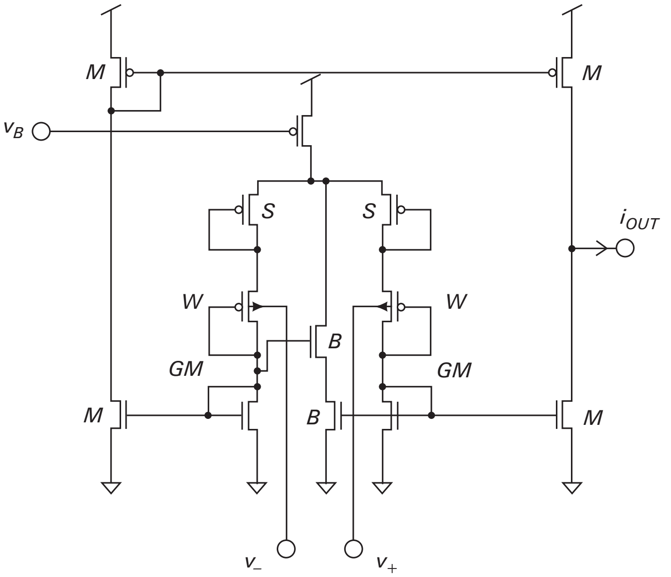


Figure 12.2. The wide-linear-range (WLR) transconductance amplifier. Reproduced with kind permission of Springer Science and Business Media from [2].

its input differential voltage are gentle and gradual. The gentler slope of the *I–V* curve then ensures that the saturation current I_B is reached at a larger V_L.

Figure 12.2 reveals a well-input wide-linear-range transconductance amplifier that exploits four techniques for reducing the G_m/I_B ratio, namely, the use of well inputs, source degeneration, gate degeneration, and bump linearization. These techniques are implemented in each of the two arms of the differential-pair transistors of Figure 12.2, i.e., in the *W*, *S*, *GM*, and *B* transistors respectively. The *M* transistors serve to mirror the currents in each differential arm to the output such that the output current is the difference in current between one differential arm and the other.

The first technique exploits the use of well inputs rather than gate inputs for reducing G_m: the *W* transistors in Figure 12.2 have their wells tied to the inputs of the amplifier. Since well inputs have a lower transconductance of g_{mb} for a given current than gate inputs, which have a transconductance of g_m, there is a gentler increase in current with differential voltage that reduces G_m. A second technique for reducing G_m exploits the well-known negative-feedback technique of source degeneration: any increases in current in a differential arm of the amplifier cause increases in voltage across the *S* transistors, which then reduce the source voltages of the pFET *W* transistors, thus attenuating the original current increase via negative feedback. The third technique for reducing G_m exploits a novel scheme that we term gate degeneration in analogy with source degeneration: any increases in current in a differential arm of the amplifier cause increases in voltage across the *GM* transistors, which then increase the gate voltages of the pFET *W* transistors, thus attenuating the original current increase via negative feedback.

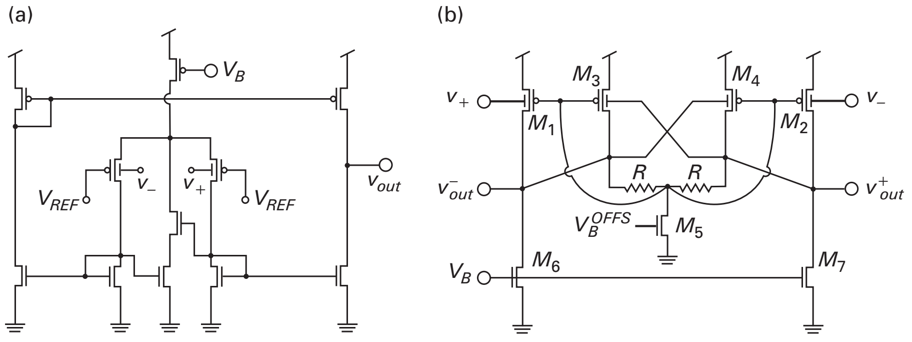


Figure 12.19a, b. A low-voltage subthreshold transconductance amplifier is shown in (a) and a low-voltage above-threshold transconductance amplifier is shown in (b). Reproduced with kind permission from [4] (©2005 IEEE).

processes, a linear range that exceeds V_{DD} is not needed: a linear range of $V_{DD}/2$ provides rail-to-rail operation with a common mode voltage of $V_{DD}/2$. Such a linear range can be obtained by using a subset of the linearization techniques described in this chapter rather than by using all of them. Figure 12.19 (a) reveals a subthreshold design suited for low-voltage operation and Figure 12.19 (b) reveals an above-threshold design suited for low-voltage operation [4]. Before we describe either design, we shall briefly digress to discuss seven general principles for low-voltage analog design. We shall then show how these principles manifest themselves in the particular circuits of Figure 12.19 (a) and 12.19 (b).

1. Since each transistor in a series stack causes a loss in saturated operating range of at least V_{DSAT} (V_{GS} if the transistor is diode connected), *series stacking of transistors should be minimized* in low-voltage designs.
2. *Weak-inversion operation* minimizes V_{DSAT}; therefore, unless high-speed requirements do not prohibit its use, it is preferable in low-voltage operation.
3. It is advantageous to not set the well-to-source voltage $v_{WS} = 0$ but to *use* v_{WS} *as a degree of freedom in low-voltage design*: the well can serve to modulate the dc biasing current while the gate serves as the primary ac input, or vice versa. The use of two control inputs rather than one is always advantageous as we have seen in the example of gate degeneration but it is particularly so in low-voltage design. If only one control input is used, say v_{GS}, the dc voltage V_{GS} must equal $V_{DD}/2$ to maximize voltage headroom; deviations in V_{GS} from $V_{DD}/2$, which are needed to alter the dc biasing current, will then compromise voltage headroom on the v_{gs} ac input at either the top or the bottom rail; however, if V_{WS} is also available as a control input, V_{GS} can be fixed at $V_{DD}/2$ maximizing input ac voltage headroom while V_{WS} can be varied to alter the dc biasing current.
4. If V_{DD} is small and less than the junction turn-on voltage, V_{WS} *can be biased* to be $-V_{DD}$ without danger of turning on the parasitic bipolar transistor in Figure 12.8. Thus, V_{T0} can be reduced to improve overdrive gate voltage in strong inversion, and rail-to-rail operation on the well voltage may be possible.

Table 13.1 Comparison between the two-transconductor (Figure 13.8 (c)) and three-transconductor (Figure 13.12 (b)) second-order filter topologies

Topology	$V_{\max}$	Total output noise power	Total bias current (I_{TOT})	$SNR_{\max}$	Power
Figure 13.8 (c)	$\frac{V_L}{Q}$	$\frac{NqV_L}{2C}$	$I_B\left(Q+\frac{1}{Q}\right)$	$\left(\frac{CV_L}{Nq}\right)\frac{1}{Q^2}$	$2\pi(qV_{DD})N(Q^3+Q)f_n(SNR_{\max})$
Figure 13.12 (b)	V_L	$\approx\left(\frac{NqV_L}{2C}\right)Q$	$I_B\left(2+\frac{1}{Q}\right)$	$\left(\frac{CV_L}{Nq}\right)\frac{1}{Q}$	$2\pi(qV_{DD})N(2Q+1)f_n(SNR_{\max})$

which is Q times higher than the filter of Figure 13.8 (c) whose maximal *SNR* or input dynamic range is given by Equation (13.24). If the bias current needed in a transistor to create $G_m/C=\omega_n$ is I_B, i.e.,

$$I_B = 2\pi f_n C V_L \tag{13.31}$$

then Table 13.1 reveals that the power consumption of the filter of the two-transconductor filter of Figure 13.8 (c) versus the alternate three-transconductor filter of Figure 13.12 (b) is given by

$$\boxed{\begin{array}{l} P_{2Gm} = 2\pi(qV_{DD})N(Q^3+Q)f_n(SNR_{\max}) \\ P_{2Gm} = 2\pi(qV_{DD})N(2Q+1)f_n(SNR_{\max}) \end{array}} \tag{13.32}$$

Thus, the alternate three-transconductor filter has a significantly better speed-precision scaling law for power than the two-transconductor filter. Equation (13.32) is a dramatic example of how topology plays a key role in determining power. Here, the other four determinants of power, namely, the task complexity (high-Q active filter), bandwidth (f_n), precision ($SNR_{\max}$), and technology (qV_{DD}) are identical in both topologies, but one topology does significantly better because of a better mapping of the task to the circuit implementing it.

To create bandpass resonant transfer functions, in the two-transconductor topology of Figure 13.8 (c), the ground terminal of C_1 is used as the input terminal and the input terminal is grounded; to create highpass resonant transfer functions, the ground terminal of C_2 is used as the input terminal and the input terminal is grounded. To create bandpass resonant transfer functions in the three-transconductor topology of Figure 13.12 (b), the ground terminal of C_1 is used as the input terminal and the input terminal is grounded; to create highpass resonant transfer functions, the ground terminal of C_2 is used as the input terminal and input terminal is grounded.

In deep submicron processes, the dc gain of $G_m - C$ filters can be fairly low due a small output resistance R_o and a large V_L, which yields a small transconductance G_m. The small value of R_o not only degrades the dc gain of the filter but its Q as well. Thus, for example, the two-transconductor topology of Figure 13.8 (c) yields

$$Q_{eff} \approx \frac{Q}{1+\left(\frac{1+Q^2}{G_mR_o}\right)} \qquad A_{dc} \approx \frac{1}{1+\left(\frac{1+Q^2}{G_mR_o}\right)} \tag{13.33}$$

The three-transconductor topology of Figure 13.12 (b) scales better in this regard as well and yields

$$Q_{eff} \approx \frac{Q}{1+\left(\dfrac{1+2Q}{G_m R_o}\right)};\; A_{dc} \approx \frac{1}{1+\left(\dfrac{1+2Q}{G_m R_o}\right)} \tag{13.34}$$

The solution in both cases is to use cascoding to improve R_o.

13.7 Higher-order G_m–C filter design

Higher-order $G_m - C$ filter design may proceed via two approaches: The first approach is to start with a high-order filter prototype such as a ladder filter, look up tabulated values for elements of the filter, and then use element-replacement techniques to synthesize a $G_m - C$ filter. The second approach is to look up tabulated pole and zero locations, decompose the poles and zeros into complex conjugate pairs, use element-replacement or state-space synthesis techniques to make first-order or second-order filters from these decompositions, and then to cascade the first-order or second-order filters. To obtain a good scaling law for power in terms of the bandwidth, precision, or Q of the filter, it is important to choose a $G_m - C$ topology with balanced differential-voltage transfer functions as we have illustrated for second-order filter design.

13.8 A $-s^2$-plane geometry for analyzing the frequency response of linear systems

Figures 13.14 (a), 13.14 (b), and 13.14 (c) illustrate how a simple $-90°$ rotation followed by a squaring of the frequency, i.e., a $(-js)^2 = -s^2$ mapping of the s-plane, is

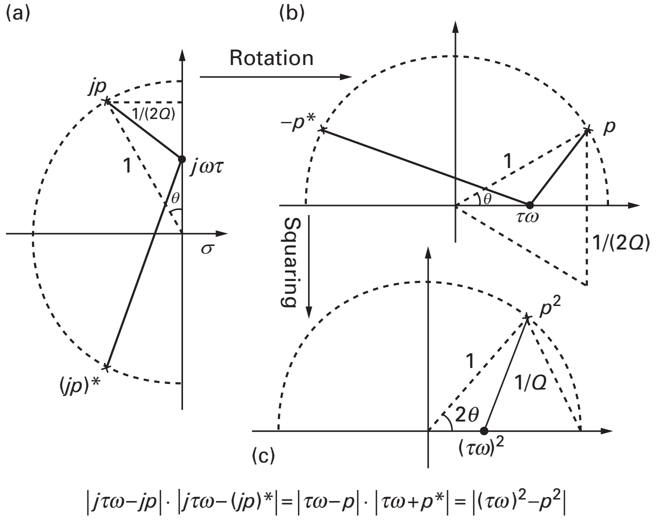


Figure 13.14a, b, c. Two-pole to one-pole geometry transformation.

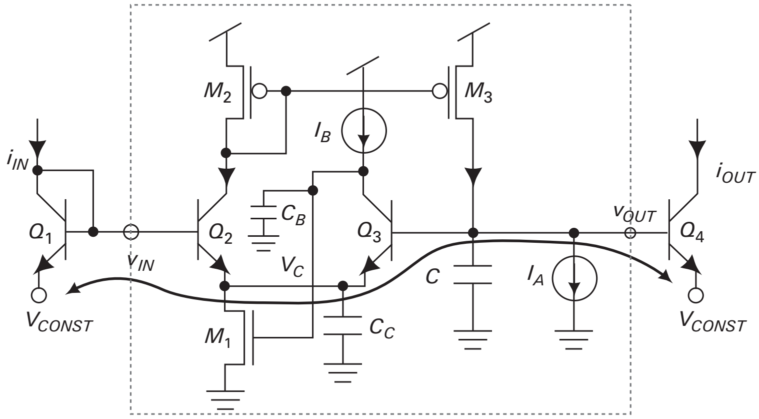


Figure 14.9. Dynamic translinear lowpass filter, circuit #3.

bias current of I_B; thus, the common-source node of the differential pair, v_C, quickly follows the voltage v_{OUT} with an offset voltage of $(kT/q)\ln(I_B/I_s)$. A second major negative-feedback loop changes v_{OUT} via capacitive charging in a slow fashion such that v_{OUT} and consequently v_C is driven to a value such that the current through Q_2 equilibrates at I_A. It is the slow charging of capacitor C that causes the circuit to function like a current-mode lowpass filter when the input current i_{IN} determines v_{IN} and the voltage v_{OUT} is used to create an output current i_{OUT}. The overall circuit from v_{IN} to v_{OUT} may be viewed as a dynamic floating voltage source with an equilibrium dc offset voltage of $v_{OUT} - v_{IN} = (kT/q)\ln(I_B/I_A)$ and a small-signal time constant of $C(kT/q)/I_A$. The additive dc offset between v_{IN} and v_{OUT} causes a multiplicative dc gain between i_{IN} and i_{OUT} of I_B/I_A and the dynamics of the floating voltage source manifest as a first-order lowpass filter in the i_{IN}-i_{OUT} current-mode input-output system.

The marked translinear loop in Figure 14.9 yields

$$\begin{aligned} i_{IN} \cdot I_B &= (i_C + I_A) \cdot i_{OUT} \\ i_C &= \frac{C\phi_t}{\eta}\frac{1}{i_{OUT}}\frac{di_{OUT}}{dt} \end{aligned} \tag{14.11}$$

such that we may once again conclude that

$$\begin{aligned} i_{IN}I_B &= I_A i_{OUT} + \frac{C\phi_t}{\eta}\frac{di_{OUT}}{dt} \\ \frac{I_{out}(s)}{I_{in}(s)} &= \frac{I_B/I_A}{1 + s\left(\frac{C\phi_t}{I_A\eta}\right)} \end{aligned} \tag{14.12}$$

It is interesting to note the similarities between the lowpass filter circuits of Figure 14.5 (a), Figure 14.7, and Figure 14.9: In all cases, a log-encoded input voltage created by the input current drives a source-follower-with-a-capacitor

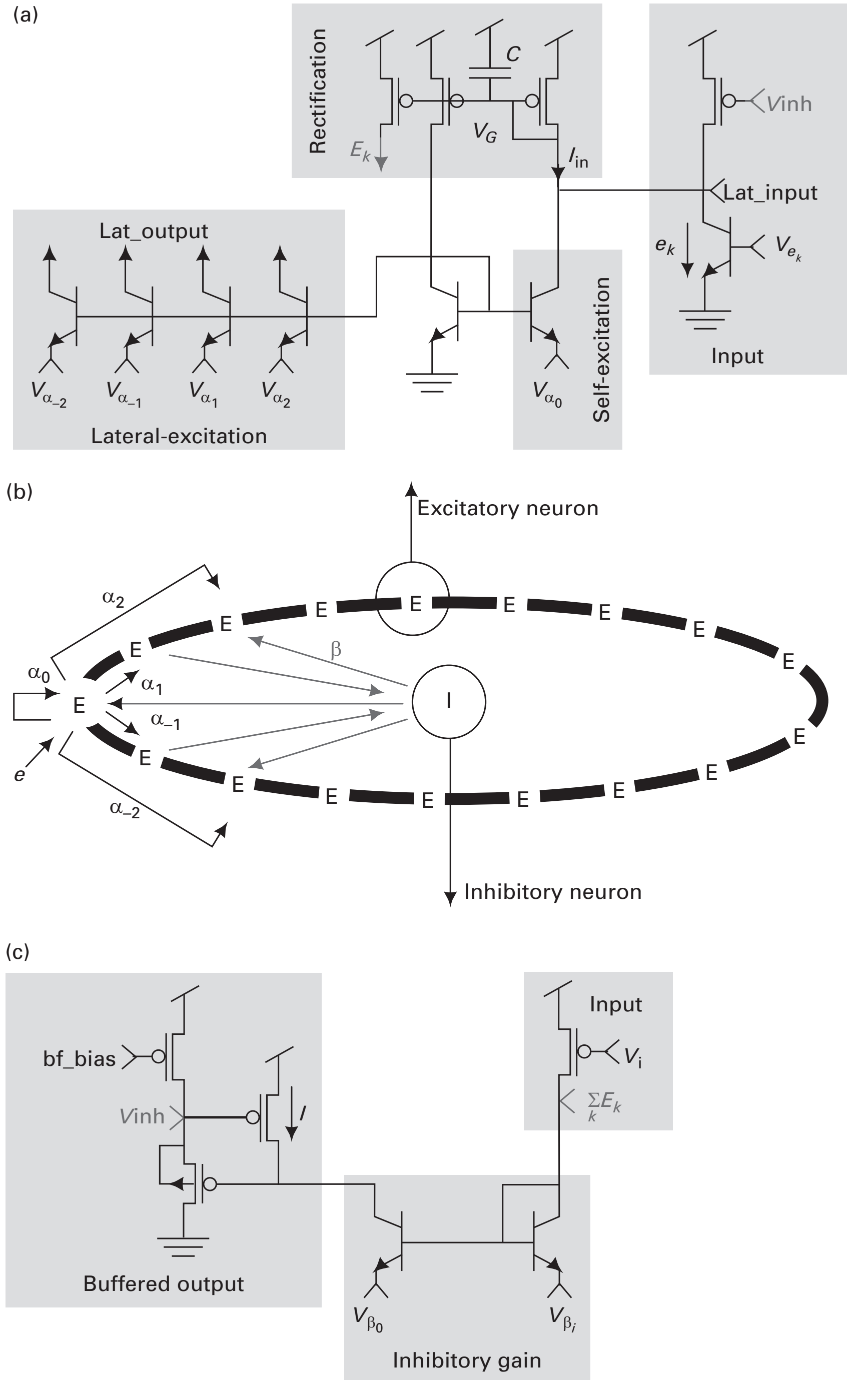


Figure 14.22a, b, c. A circuit model of excitation and inhibition in the cortex (featured on the cover of *Nature* in July 2000). Reprinted by permission from Macmillan Publishers Ltd: *Nature* [11] ©2000.

15 Ultra-low-power and neuron-inspired analog-to-digital conversion for biomedical systems

Although nature commences with reason and ends in experience, it is necessary for us to do the opposite; that is, to commence with experience and from this to proceed to investigate the reason.
Leonardo da Vinci

An analog-to-digital converter converts real-world continuous analog signals into symbolic discrete digital numbers. It is often abbreviated as an ADC, A-to-D, or A/D. ADCs are ubiquitous in all electronic systems. A digital-to-analog converter performs the inverse function and is correspondingly abbreviated as a DAC, D-to-A, or D/A. Figure 15.1 shows the input-output curve of an ADC [1]. The input and output are equal to each other within a quantization error of $\pm\,\Delta/2$, a consequence of the fact that we need to round up or round down real numbers to the nearest integer to represent them digitally. The digital numbers are usually represented with binary digits or bits. If, because of the input statistics, any error between $[-\Delta/2, +\Delta/2]$ is equally likely, then, from evaluation of the second moment of a flat probability distribution, the rms error of the quantized representation of a real number can be shown to be $\Delta^2/12$. If the ADC samples its input at a sampling frequency f_S, the power spectrum of the quantization noise is well approximated as being white from 0 to $f_S/2$ and therefore having a noise per unit bandwidth of $\Delta^2/(12(f_S/2))$. If the precision of the converter is N bits, digital numbers between 0 and $(2^N - 1)$ represent analog signals between 0 and a full-scale voltage V_{FS} with V_{FS} corresponding to 2^N. Thus $V_{FS}/2^N = \Delta$. Hence a sine wave with amplitude $V_{FS}/2$ that spans the full $[0, V_{FS}]$ input range of the converter has a signal-to-noise ratio (SNR) given by

$$SNR = \frac{\frac{(V_{FS}/2)^2}{2}}{\Delta^2/12} = \frac{3}{2}2^{2N} \tag{15.1}$$

The latter relationship is frequently used to evaluate N, the equivalent number of bits or ENOB of the converter.

Several ADCs usually have their power consumption well described by

$$\begin{aligned} P_{ADC} &\propto f_S 2^N \\ P_{ADC} &= E_q.f_S.2^N \end{aligned} \tag{15.2}$$

where E_q is relatively constant with N and represents the energy per quantization level. Such power scaling arises because ADCs that are more precise need to have on the ith clock cycle: each of the M converters begins its integration-of-the-input phase on a clock cycle that is sequentially staggered amongst the converters. Thus, the Mth converter will just be sampling its input on the Mth clock cycle when the first converter is in its terminating cycle of conversion.

15.3 Computational ADCs and time-to-digital ADCs

An interesting feature of the neuron-inspired ADC is that polynomial functions of the analog input current can be computed in a sequential Taylor-series-like fashion. We exploit successive *Charge* = *Current* × *Time* relationships to perform multiplicative operations. Figure 15.12 illustrates the process: ADC1 quantizes the first timing variable, T_{x1}, in the same manner as the neuron-inspired ADC. Therefore, we know that

$$T_{x1} = T_{clk} \frac{I_{in}}{I_{ref}} \tag{15.24}$$

Now ADC2 begins digitizing I_{in} after ADC1 has finished a sequence comprised of one Time-to-Voltage and one Voltage-to-Time operation. The effective integration time for ADC2's sample-and-hold phase is then T_{x1} as shown in Figure 15.12. The time interval T_{x1} that ADC2 operates on is free of comparator-delay and charge-injection errors; furthermore, T_{x1} is referenced to a following clock edge instead of a preceding clock edge for ADC2. Thus, the quantized output of ADC2 will reflect

$$T_{x2} = T_{x1} \frac{I_{in}}{I_{ref}} \tag{15.25}$$

From Equations (15.24) and (15.25), we get

$$T_{x2} = T_{clk} \left(\frac{I_{in}}{I_{ref}} \right)^2 \tag{15.26}$$

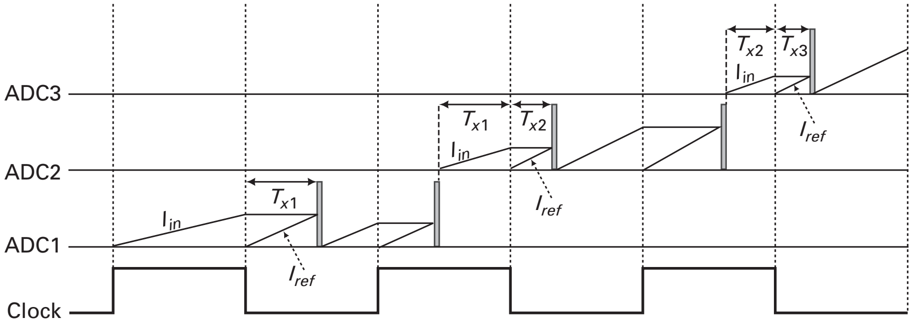


Figure 15.12. Computational ADCs. Reproduced with kind permission from [14] (©2006 IEEE).

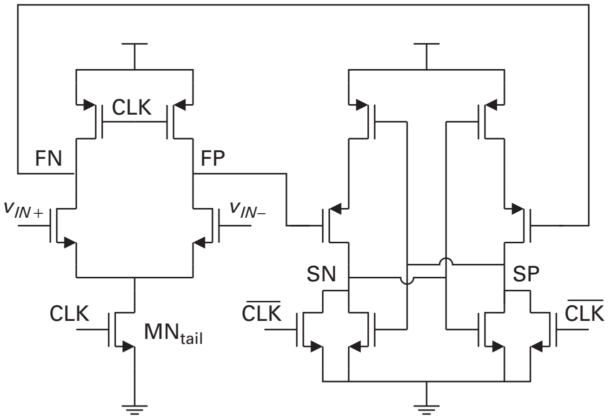


Figure 15.17. A highly energy-efficient comparator circuit. Reproduced with kind permission from [6] (©2008 IEEE).

15.7 Digital correction of analog errors

If an ADC is not thermal-noise limited, nonlinearities, gain, and offset errors in its components may limit its precision. Unlike thermal noise, these errors are predictable. Therefore, it is possible to correct for these errors on the digital output of the ADC if one can learn what these errors are [22]. To do so, we first form a parametric digital model of how various errors in its analog components cause error. For example, gain errors in a pipelined ADC from its amplifying stages have predictable effects on its digital output with errors in early amplification stages mattering exponentially more than errors in later amplification stages. Then, we experimentally determine the error of the ADC by sweeping its input and comparing its digital output with that from a slow highly precise ADC. We then learn the parameters of the model of the ADC that best describes the observed errors in a least-squares sense using standard learning techniques such as gradient descent on a multi-parameter space. Finally, we store the learned parameters digitally and use them and the error model to perform digital error correction on the outputs of the ADC while it is running. The learning can typically be done in a continuous online fashion with the calibrating ADC constantly running at a relatively low bandwidth in the background and constantly providing information to update the parameters of the error model. The ADC described in [23] uses such a strategy with a fast pipelined ADC being constantly calibrated in the background by a slow algorithmic ADC.

The overall scheme is energy efficient because one can use open-loop highly inaccurate, low-gain analog components that perform at high speed but with low precision in an energy-efficient fashion. If the known errors are well modeled and do not change often, the calibrating ADC can be slow and highly precise and perform calibrations relatively infrequently. Thus, the main ADC and the calibrating ADC may both be operated in an energy-efficient fashion because neither is simultaneously fast *and* precise. The digital error-correction power does now add to the overall power but the costs of such error correction are usually modest.

15.8 Neurons and ADCs

Intriguingly, highly energy-efficient ADCs appear to be converging to the digitization strategies used by 0.5 nW neuronal cells in the brain that also use comparators and time-based strategies for performing vector pattern recognition. Vector pattern recognition is a generalization of A-to-D conversion from scalar inputs to vector inputs. The neuron-inspired algorithmic ADC that we described in this chapter was inspired by the operation of integrate-and-fire pulsatile or spiking neurons in the brain and naturally led to energy-efficient conversion, computational ADCs and time-to-digital architectures.

The highly energy-efficient ADC of Figure 15.13 also operates much like an integrate-and-fire neuron even though the authors do not explicitly claim to have arrived at their architecture through neuronal inspiration. The neuron is often well modeled as a linear current-controlled oscillator (CCO) since the frequency of pulsatile spike firing of the neuron is proportional to its input current. Thus, if we count the number of pulses fired by a neuron in a fixed time period or fixed number of clock cycles, we have a quantized representation of the current input. The topology of Figure 15.13 replaces the neuronal CCO with a VCO, and the neuronal spike generator with an edge-sensitive XOR. Both topologies quantize their input by simply counting the number of edges in a frequency input.

Closed-loop ΣΔ architectures have strong analogies to adaptive neurons. In certain neurons, the firing of spikes leads to the accumulation of calcium in the cell, which, in turn, then turns on a calcium-dependent potassium conductance that inhibits or subtracts current from the charging input current of the neuron [24]. The subtraction is analogous to the subtractive input of a ΣΔ and the feedback current is analogous to the feedback variable of a ΣΔ architecture. However, in most neurons, the time constant of the feedback is slow such that the presence of a slow integration in the feedback path leads to a differentiating ΣΔ A/D that only outputs events when the input changes.

In Chapter 22, we shall see that successive approximation ADCs are a special example of a more general analysis-by-synthesis architecture, where the synthesized DAC model of the input provides iterative 1 or 0 event feedback that helps us analyze the input. Such analysis-by-synthesis architectures can be efficiently generalized to create hybrid state machines with spiking neurons as we discuss in Chapters 22 and 23.

The numerous analogies between neurons and ADC architectures suggest that we may find further inspiration from neurons on how to build interesting computational and energy-efficient ADCs and in signal-to-symbol conversion in general. We shall return to such bio-inspired and other mixed-signal architectures in Chapters 22, 23, and 24.

Section III

Low-power RF and energy-harvesting circuits for biomedical systems

16 Wireless inductive power links for medical implants

I do not think there is any thrill that can go through the human heart like that felt by the inventor as he sees some creation of the brain unfolding to success . . . Such emotions make a man forget food, sleep, friends, love, everything.

Nikola Tesla

Implanted medical devices are rapidly becoming ubiquitous. They are used in a wide variety of medical conditions such as pacemakers for cardiac arrhythmia, cochlear implants for deafness, deep-brain stimulators for Parkinson's disease, spinal-cord stimulators for the control of pain, and preliminary retinal implants for blindness. They are being actively researched in brain-machine interfaces for paralysis, epilepsy, stroke, and blindness. In the future, there will undoubtedly be electronically controlled drug-releasing implants for a wide variety of hormonal, autoimmune, and carcinogenic disorders. All such devices need to be small and operate with low power to make chronic and portable medical implants possible. They are most often powered by inductive radio-frequency (RF) links to avoid the need for implanted batteries, which can potentially lose all their charge or necessitate re-surgery if they need to be replaced. Even when such devices have implanted batteries or ultra-capacitors, an increasing trend in upcoming fully implanted systems, wireless recharging of the battery or ultra-capacitor through RF links is periodically necessary.

Figure 16.1 shows the basic structure of an inductive power link system for an example implant. An RF power amplifier drives a primary RF coil which sends power inductively across the skin of the patient to a secondary RF coil. The RF signal on the secondary coil is rectified and used to create a power supply that powers internal signal-processing circuits, electrodes and electrode-control circuits, signal-sensing circuits, or telemetry circuits depending on the application. The power consumption of the implanted circuitry is eventually borne by external batteries that power the primary RF coil; if an RF link is energy efficient, most of the energy in the primary RF coil will be transported across the skin and dissipated in circuits in the secondary. It is also important for an RF link to be designed such that the power-supply voltage created in the secondary is relatively invariant to varying link distances between the primary and secondary, due to patient skin-flap-thickness variability, device placement, and device variability.

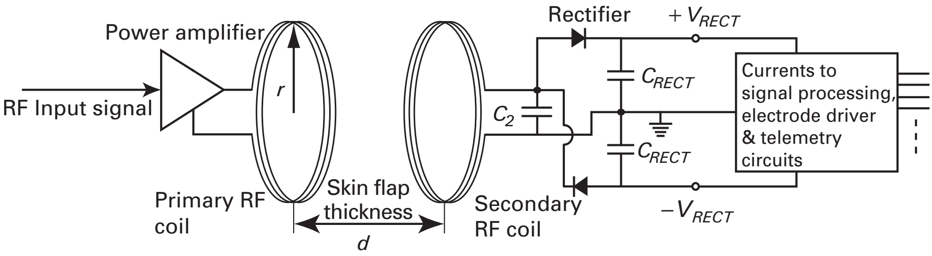


Figure 16.1. An example of a low-power bionic implant system. Reproduced with kind permission from [10] (©2007 IEEE).

RF power links for biomedical systems need to achieve good energy efficiency such that needless amounts of external power are not used to power an internal system. Small losses that are important in low-power systems may be insignificant in higher-power systems. For example, in milliwatt-level implanted systems such as in pacemakers or energy-efficient cochlear-implant processors [1], switching losses in the power amplifier and rectifier can be a significant portion of the overall power and hurt efficiency. In this chapter, we discuss how to design RF power links for maximum link energy efficiency and also to obtain acceptable robustness to inter-coil separation.

It is generally advantageous to separate the power and data transfer functions of a wireless link as some authors have done in the past [2], [3], [4]. Power signals carry no information, and power transfer efficiency is maximized for narrowband (high-Q) links that operate at low frequencies to minimize losses in body tissue. On the other hand, data signals carry information and therefore require larger link bandwidths, which are more easily obtained at higher operating frequencies. Separating the two functions therefore allows them to be independently optimized, improving overall performance.

In this chapter, we shall focus only on inductive power links and describe the design of inductive data links in a separate chapter, Chapter 18. We first discuss the theory of linear coupled resonators, a feedback analysis for understanding the power link, and derive expressions for its efficiency. Then, we discuss the design of a bionic implant power system, with attention to efficiency at low power levels. Finally, we present experimental results from a working system and discuss higher-order effects.

16.1 Theory of linear inductive links

A pair of magnetically coupled resonators is shown in Figure 16.2 and represents a model of our RF link with the primary external resonator on the left and the secondary implanted resonator on the right. The mutual inductance between the primary and secondary is represented by M. The resistances R_1 and R_2 are implicit resistances due to coil losses in the inductances L_1 and L_2 while C_1 and C_2 are

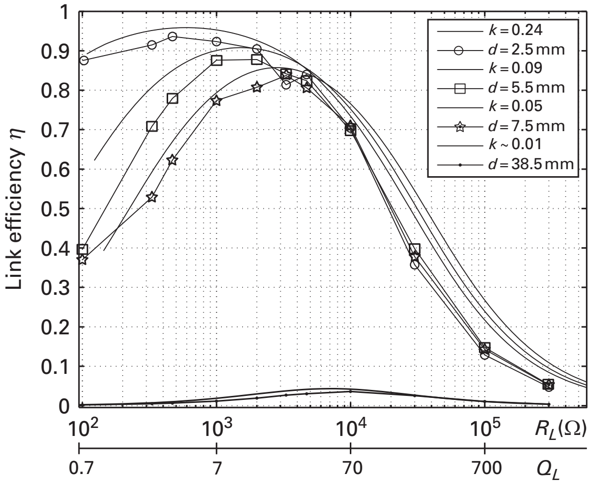


Figure 16.9. Efficiency of power transfer between coupled resonators for varying loads. This particular data were produced by ignoring switching losses in a discrete Class-E amplifier implementation. Reproduced with kind permission from [10] (©2007 IEEE).

The optimal Q_L that optimizes the overall efficiency

$$\begin{aligned}\eta &= \eta_{prm}\eta_{scnd}\\ &= \left(\frac{k^2Q_1\dfrac{Q_2Q_L}{Q_2+Q_L}}{1+k^2Q_1\dfrac{Q_2Q_L}{Q_2+Q_L}}\right)\left(\frac{Q_2}{Q_2+Q_L}\right)\end{aligned} \tag{16.34}$$

is found by differentiating Equation (16.34) with respect to Q_L. We find that the optimum

$$\boxed{Q_{L,opt} = \frac{1}{k}\sqrt{\frac{Q_2}{Q_1}}} \tag{16.35}$$

and that the maximally achievable efficiency η_{MAX} attainable at this optimum is given by substituting $Q_{L,opt}$ for Q_L in Equation (16.34). We find that

$$\boxed{\eta_{MAX} = \frac{k^2Q_1Q_2}{(kQ_1+1)(kQ_2+1)}} \tag{16.36}$$

Figure 16.9 shows experimental results of the overall power transfer efficiency. These results were obtained from a coupled-resonator system with no power amplifier or rectifiers in Figure 16.1 such that we could focus on only the essential resonator circuit of Figure 16.2 for now. The data are shown for four separation distances of the RF link. We see that the peak efficiency shifts to lower load resistance as the coils are moved closer together in accord with the theoretical

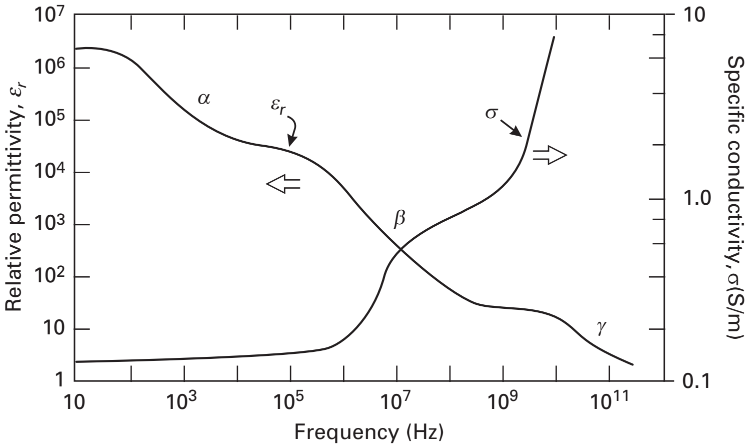


Figure 16.18. Relative permittivity and specific conductivity of skin as a function of frequency. Reproduced with kind permission from [11].

absorption at these frequencies. The three sigmoid-like changes in permittivity ε_r with frequency, marked α, β, and γ, are believed to be due to the motion of extracellular and cytoplasmic ions at membrane and tissue interfaces, the motion of dipoles fixed in cell membranes, and the motions of water dipoles respectively [11], [12]. Models for the electric properties of biological tissue may be found on the web at [13].

The overall characteristics of skin can be approximated by introducing a series-connected R-and-C bridging impedance element that couples the primary and secondary loops. The capacitive coupling due to this bridging element can sometimes help efficiency by introducing an additional coupling path although it often causes unpredictable changes in resonant frequency.

References

[1] R. Sarpeshkar, C. D. Salthouse, J. J. Sit, M. W. Baker, S. M. Zhak, T. K. T. Lu, L. Turicchia and S. Balster. An ultra-low-power programmable analog bionic ear processor. *IEEE Transactions on Biomedical Engineering*, **52** (2005), 711–727.

[2] L. Theogarajan, J. Wyatt, J. Rizzo, B. Drohan, M. Markova, S. Kelly, G. Swider, M. Raj, D. Shire and M. Gingerich, Minimally invasive retinal prosthesis. *Proceedings of the IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 99–108, 2006.

[3] M. Ghovanloo and S. Atluri. A Wide-Band Power-Efficient Inductive Wireless Link for Implantable Microelectronic Devices Using Multiple Carriers. *IEEE Transactions on Circuits and Systems I: Regular Papers*, **54** (2007), 2211–2221.

[4] R. Sarpeshkar, W. Wattanapanitch, S. K. Arfin, B. I. Rapoport, S. Mandal, M. W. Baker, M. S. Fee, S. Musallam and R. A. Andersen. Low-Power Circuits for Brain-Machine Interfaces. *IEEE Transactions on Biomedical Circuits and Systems*, **2** (2008), 173–183.

17 Energy-harvesting RF antenna power links

I do not think that the wireless waves I have discovered will have any practical application.

Heinrich Rudolph Hertz

Ultra-low-power systems have the potential to operate in a battery-free fashion by harvesting energy from their environment. Such energy may take the form of solar energy in a solar-powered system, chemical energy from carbohydrates in an enzyme-based system, mechanical energy from vibrations in the system's platform, thermal energy in systems that exploit temperature differences between themselves and their environment, or radio-frequency electromagnetic energy in the environment. In Chapter 26, we shall discuss several forms of energy harvesting. In this chapter, we will focus on energy harvesting with radio-frequency antennas or *rectennas* as they are sometimes called.

Radio-frequency electromagnetic energy is increasingly becoming ubiquitous due to the growing presence of cellular phones, local area networks, and other wireless devices. Systems that operate by harvesting electromagnetic energy need an antenna for sensing electromagnetic waves and a rectifier for converting the sensed ac energy to a dc power supply. The created power supply can then be used to power an ultra-low-power system such as a radio-frequency identification (RF-ID) tag in a grocery store or a medical monitoring tag on the body of a person (see Chapter 20 on medical monitoring). In this chapter, we shall discuss important principles and building blocks for creating such RF-energy-harvesting systems including antennas and rectifier circuits. We shall discuss an example of a complete functioning experimental system to illustrate system-level tradeoffs. We begin by reviewing the fundamentals of antenna operation.

Antennas are complicated distributed circuits that serve to transmit or receive electromagnetic energy. When functioning as receivers, they sense distributed electromagnetic wave input signals in free space to create a local electrical signal between a pair of lumped output terminals. When functioning as transmitters, they convert a local electrical signal between a pair of lumped input terminals to an electromagnetic wave that is radiated in a distributed fashion into free space.

A quantitative and detailed understanding of antennas could and does form the subject of several books that analyze antennas using Maxwell's equations of electromagnetism. For example, see [1]. Rather than repeat such analysis here, we shall focus on providing an intuitive understanding that is useful for our purposes and that is practical for low-power circuit design.

17.1 Intuitive understanding of Maxwell's equations

To begin, we note that a spatially discrete approximation to Maxwell's equations can be simulated in an 'analog computer' made up of an infinite LCR network of infinitesimally small and invisible inductors, capacitors, and resistors in free space that are connected to each other in the mesh of Figure 17.1. This ingenious analog-circuit simulation of Maxwell's equations was formulated by Gabriel Kron as described in [2]. The formulation cleverly configures loop currents and node voltages in the mesh such that distributed parameters such as the electric field ***E***, magnetic field ***H***, or current density ***J*** are proportional to the voltage across the capacitors, the current in the inductors, or the current in the resistors, respectively, as shown; the constants of proportionality are given by ε, μ, and σ, which represent the capacitive, magnetic, and conductive properties of the medium. Kirchhoff's voltage law (KVL), Kirchhoff's current law (KCL), and

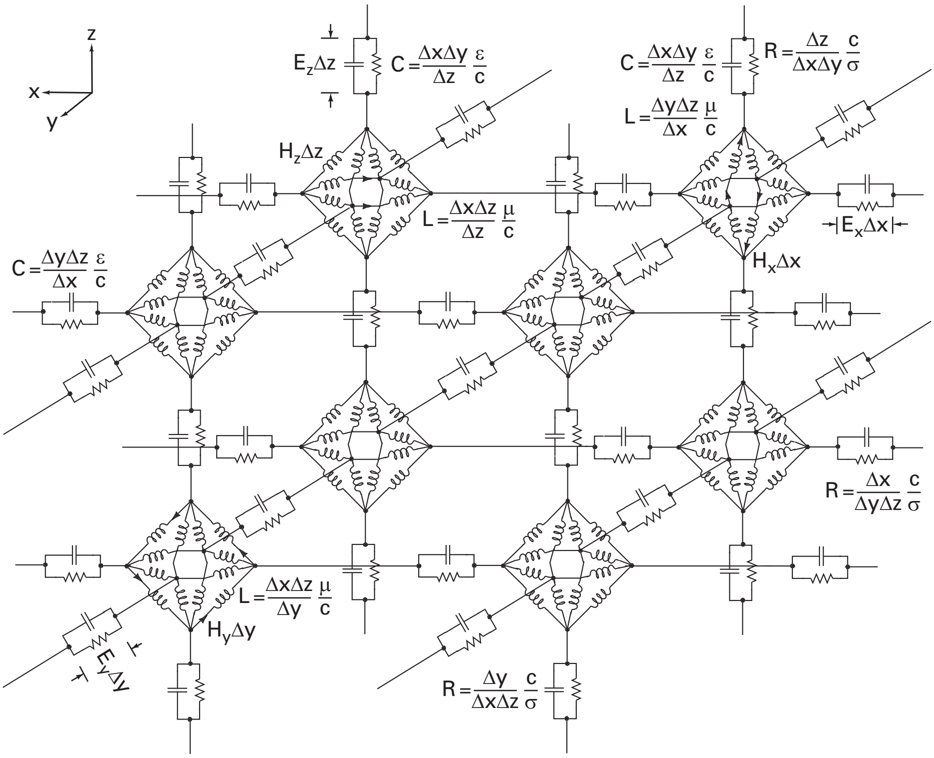

z
x
y
$E_z\Delta z$
$C=\frac{\Delta x\Delta y}{\Delta z}\frac{\varepsilon}{c}$
$C=\frac{\Delta x\Delta y}{\Delta z}\frac{\varepsilon}{c}$
$R=\frac{\Delta z}{\Delta x\Delta y}\frac{c}{\sigma}$
$L=\frac{\Delta y\Delta z}{\Delta x}\frac{\mu}{c}$
$H_z\Delta z$
$L=\frac{\Delta x\Delta z}{\Delta z}\frac{\mu}{c}$
$E_x\Delta x$
$H_x\Delta x$
$C=\frac{\Delta y\Delta z}{\Delta x}\frac{\varepsilon}{c}$
$R=\frac{\Delta x}{\Delta y\Delta z}\frac{c}{\sigma}$
$L=\frac{\Delta x\Delta z}{\Delta y}\frac{\mu}{c}$
$H_y\Delta y$
$E_y\Delta y$
$R=\frac{\Delta y}{\Delta x\Delta z}\frac{c}{\sigma}$

Figure 17.1. Circuit model of Maxwell's equations. Reproduced with kind permission from [2] (©1944 IEEE).

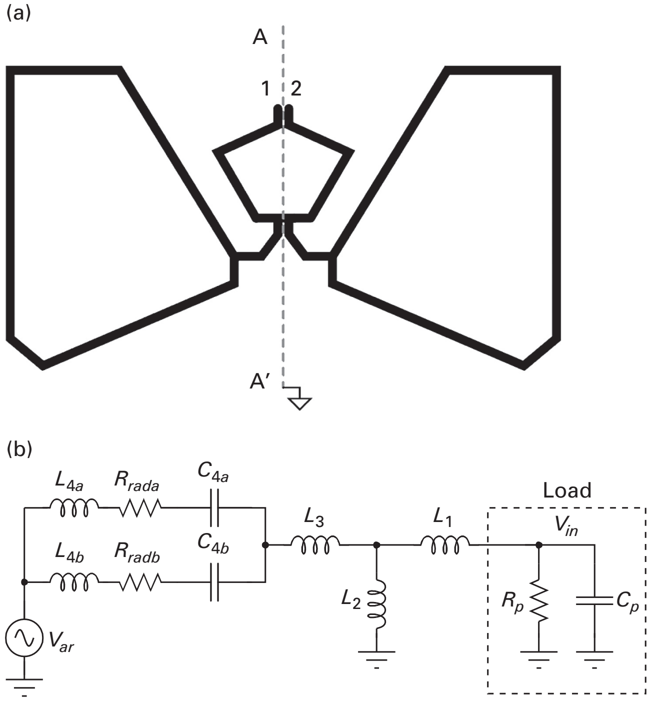


Figure 17.10a, b. Bugs Bunny antenna (a) and coupled-resonator analog (b). Figure (a) reproduced with kind permission from [10] (©2007 IEEE).

have been opened out from being near each other to being spread out and pointing away from each other. Note that the radiation resistance for a short dipole is proportional to the square of the frequency due to the inverse-squared dependence on λ in Equation (17.27). Thus, the resistance is not constant with frequency as in a traditional series LCR circuit.

A natural way to create a relatively broadband antenna is to let the distributed reactance of the antenna and the lumped reactance of the load form an implicit matching network between the antenna radiation resistance and the load resistance. The antenna, matching network, and load are then all part of one structure and appear less separable. While a dipole is well approximated by a simple series LCR circuit, more complex antennas that incorporate loops, dipoles, and distributed conductive structures within them can serve to create a more complex implicit matching network. We now provide an example of how a second-order matching network is implicitly created by an antenna and a load such that the radiation resistance of the antenna is matched to the resistive impedance of a load.

Figure 17.10 (a) shows the layout of a ‘Bugs-Bunny-like’ antenna composed of two ‘ears’, a small loop, and two curved chair-like wire segments that couple each of the ears to the small loop. The terminals of the antenna are marked as 1 and 2 in Figure 17.10 (a) and form the input to the small loop. The terminals of the antenna are connected to two pins of a surface-mounted chip on a printed circuit board

17.18 Summary

Since we have covered a lot of ground in this chapter, it is worth summarizing its main points. We have discussed how to harvest energy from an RF antenna and rectify it to create a power supply suitable for battery-free ultra-low-power electronics. Such systems are important for medical monitoring in body sensor networks and in telemedicine (see Chapter 20). We have shown experimental measurements from a ~900 MHz RF-ID tag that powers up at 6 μW of received RF power in a 0.18 μm traditional CMOS process with no special devices. These measurements agree well with theories of antenna operation and Bode-Fano broadband impedance-matching networks, and with a general model of rectifier operation that was specifically applied to our CMOS H-bridge rectifier. We will return to themes of energy harvesting in Chapter 26.

References

[1] J. D. Kraus and R. J. Marhefka. *Antennas for All Applications* (New York: McGraw-Hill, 2002).

[2] G. Kron. Equivalent circuit of the field equations of Maxwell-I. *Proceedings of the IRE*, **32** (1944), 289–299.

[3] Thomas H. Lee. *The Design of CMOS Radio-Frequency Integrated Circuits*. 2nd ed. (Cambridge, UK; New York: Cambridge University Press, 2004).

[4] W. Gosling. *Radio Antennas and Propagation* (Oxford, UK: Newnes, 1998).

[5] Constantine A. Balanis. *Antenna Theory: Analysis and Design*. 3rd ed. (Hoboken, NJ: John Wiley & Sons, Inc., 2005).

[6] Carver Mead. *Collective Electrodynamics: Quantum Foundations of Electromagnetism* (Cambridge, MA: MIT Press, 2000).

[7] R. M. Fano. Theoretical limitations on the broadband matching of arbitrary impedances. *Technical Report (Research Laboratory of Electronics, Massachusetts Institute of Technology)*, **41** (1948).

[8] L. J. Chu. Physical limitations of omnidirectional antennas. *Technical Report (Research Laboratory of Electronics, Massachusetts Institute of Technology)*, **64** (1948).

[9] T. G. Tang, Q. M. Tieng and M. W. Gunn. Equivalent circuit of a dipole antenna using frequency-independent lumped elements. *IEEE Transactions on Antennas and Propagation*, **41** (1993), 100–103.

[10] S. Mandal and R. Sarpeshkar. Low-power CMOS rectifier design for RFID applications. *IEEE Transactions on Circuits and Systems I: Regular Papers*, **54** (2007), 1177–1188.

and circuits of the internal-unit pulse-width-demodulation receiver. We present experimental results for a complete transceiver system that achieves an energy efficiency of 1 nJ/bit for data rates that are a few Mbps. An important effect regarding the asymmetry of rising and falling edges that is inherent to impedance modulation is predicted by theory. Such theory leads to predictions of the bit-error rate in such links, which are confirmed by experiment [3].

We then analyze the fundamental limits of energy efficiency of an impedance-modulation RF communication link. Such limits determine the minimal energy needed to transmit a bit of information. For a given distance of communication, they can be used to evaluate a figure of merit for RF communication links. A low value of transmitter power increases the power cost needed to detect faint signals at the receiver while reducing transmitter power. In contrast, a high value of transmitter power reduces the power cost needed to detect strong signals at the receiver while increasing transmitter power. Not surprisingly, transceiver power is minimized when the power used for transmission and reception is balanced at an optimum value such that one does not spend too much power in either transmission or reception. We shall compute this optimum value. We show how this optimum depends on the strength of the coupling in the link, i.e., whether the resonators are near each other or far apart. Links composed of resonators that are relatively near each other have well balanced transmitter and receiver power dissipation at their optimum and a relatively low energy per bit of information communicated. Links composed of resonators that are relatively far apart require most of the power to be spent in the transmitter at their optimum and a relatively high energy per bit of information communicated at this optimum.

We also outline the pros and cons of using incoherent (no fine carrier phase information is present) envelope-detector-based receivers versus coherent (fine carrier phase information is present) mixer-based receivers. Systems with either kind of receiver minimize total power when transmission and reception power costs are balanced. Coherent receivers are relatively more efficient in strongly coupled near-field links that are capable of a relatively high bandwidth of information flow while incoherent receivers are relatively more efficient in weakly coupled near-field links that are capable of a low bandwidth of information flow.

We use our RF link as an example that illustrates how to evaluate whether a designed system obeys Federal Communications Commission (FCC) regulations. In very-low-power RF links, there is considerably more freedom in choosing an optimal carrier frequency if one can operate at very low power levels that do not violate such regulations.

We then describe seven considerations that determine the choice of RF carrier frequency in a wireless telemetry system. We review RF antenna-based links for implants that operate at relatively high RF carrier frequencies. Such links utilize one far-field RF link to establish communication between an implant within the body and an RF receiver a short distance away from it. They do so primarily because they can exploit small antenna sizes for the implant, are convenient, and have the potential to operate at higher bandwidths. Their primary disadvantage is that transmission of RF energy through the body at these high frequencies is subject to significant transmission loss such that the data bandwidths and energy efficiency of such links have thus far been modest. We conclude by estimating the skin depth for RF wave propagation through the body at various carrier frequencies.

18.1 Impedance modulation in coupled parallel resonators

The coupled parallel resonator topology is shown in Figure 18.1. The inductance L_1 and capacitance C_1 form a parallel resonator in the external primary while L_2 and C_2 form a parallel resonator in the implanted secondary. The resistances R_1 and R_2 are parasitic series coil resistances that determine the quality factor $Q_1 = (\omega_{res}L_1)/R_1$ and $Q_2 = (\omega_{res}L_2)/R_2$ of the primary and secondary resonators respectively at the resonant RF operating frequency ω_{res}. The transistor with a dc bias of V_{BIAS} and a small-signal RF input of v_{in} has a small-signal transconductance of g_m and serves along with the RF load to create a common-source amplifier that amplifies v_{in} to a larger voltage $v_1 = i_{in}R_{eff}$ where R_{eff} is the effective resistance seen at the primary. The mutual inductance $M = k\sqrt{(L_1L_2)}$ serves to provide bidirectional coupling between the primary and secondary such that a voltage v_2 is created in the secondary.

Impedance modulation is accomplished by shorting or opening C_2 in the secondary with '0' or '1' data bits respectively such that the impedance in the

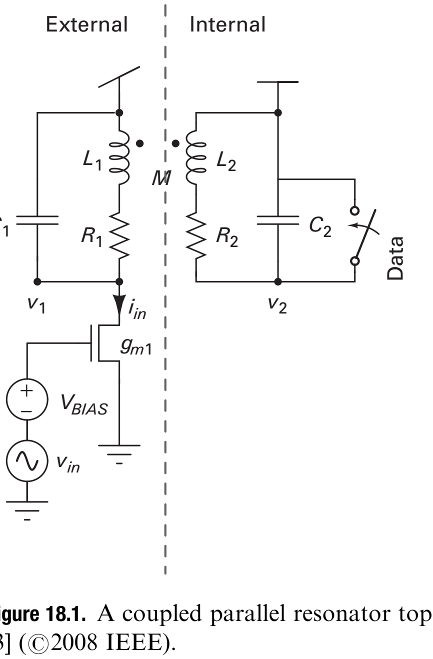


Figure 18.1. A coupled parallel resonator topology. Reproduced with kind permission from [3] (©2008 IEEE).

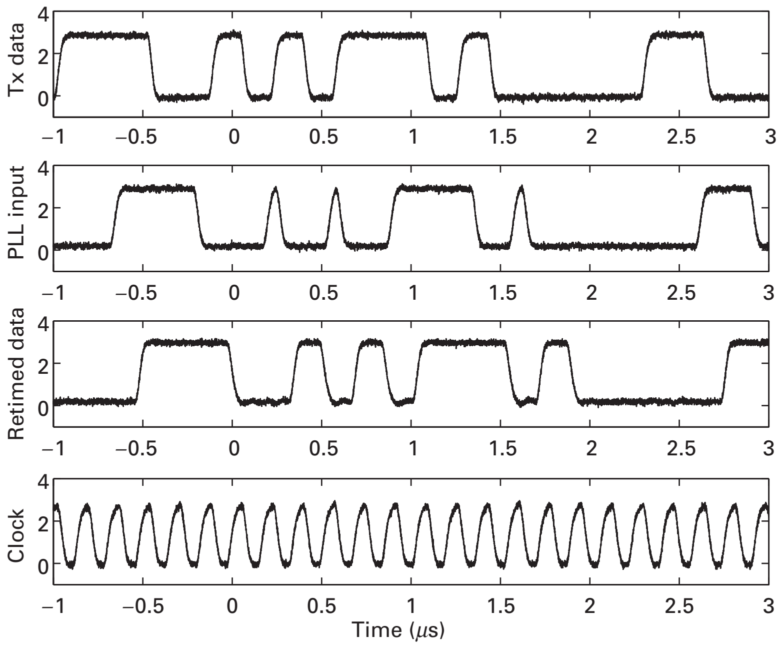


Figure 18.12. Transmitted and received data and recovered clock waveforms measured for the uplink at 5.8 Mb/s with coils 2 cm apart. Reproduced with kind permission from [3] (©2008 IEEE).

surface mounted on the printed circuit boards and aligned parallel to each other at various separations for testing. Implanted coils are typically less than 2 cm on a side and operate at link distances between 0.3 cm and 1 cm (the typical skin flap thickness). Therefore, to reduce the coupling constant to more typical values the link was tested over larger distances between 1.5 cm and 5 cm. A square coil of these dimensions is well fit by formulas for circular coils with radius 2 cm and the same square area as the 3.5 cm coils. The finite output impedance of the M_1 transistor in Figure 18.2 reduces Q_1 to 10 while Q_2 reduces from its simulated value of 30 to an actual experimental value of 25.

Figure 18.12 shows the transmitted and received data and recovered clock waveforms measured for the uplink at 5.8 Mb/s with coils 2 cm apart. The PLL synchronizes data edges to the rising edges of the clock. The isolated ‘1’ bits at the output of the comparator are significantly narrower than the ‘0’ bits due to the asymmetry of the 0-to-1 and 1-to-0 transitions. At this separation, the PLL locks between 1 Mb/s and 5.8 Mb/s. The upper end of the lock range is set by the loop filter, which saturates at V_{DDH} in Figure 18.6. The lower end of the lock range is set by the loop-filter output voltage driving the well-input voltage of the G_m WLR transconductor in Figure 18.4 to a sufficiently low value such that its shunting bipolar input transistors are activated (see Chapter 12). The rms clock jitter was measured to be 7.2 nS, t_{dr} varied between 65 ns and 140 ns with link distance

Similarly, we may write

$$E_{bit}^{strong} = \left(3\left(\frac{\left(1+k^2Q_1Q_2\right)^{2/3}}{k^{4/3}Q_1^2Q_2^{2/3}}\right)\left(\frac{V_{DDp}^2V_{Lm}V_{DDm}}{R_1^2}\right)^{1/3}(8kT\gamma)^{1/3}\right)\left(\frac{\left(\text{erfc}^{-1}(P_e)\right)^{2/3}}{f(R/B)^{1/3}}\right)\left(\frac{1}{R^{2/3}}\right) \tag{18.38}$$

$$\boxed{E_{bit}^{strong} = E_{0s}\left(\frac{\left(\text{erfc}^{-1}(P_e)\right)^{2/3}}{f(R/B)^{1/3}}\right)\left(\frac{1}{R^{2/3}}\right)}$$

Although we have approximated the weakly coupled and strongly coupled regimes through approximations of the more exact Equations (18.20) and (18.22), it is possible to formulate our analysis more exactly without approximations for the purposes of numerical computation: *Minimize $P_{TOT} = P_{PA} + P_{mxr}$ subject to the constraint that*

$$\frac{\frac{P_{PA}^2}{P_{th}}}{P_{PA}+\frac{P_{char}^2}{P_{mxr}}} = \left(\left(\text{erfc}^{-1}(P_e)\right)^2\right)\left(\frac{R/B}{f(R/B)}\right) \tag{18.39}$$

At the minimum, compute P_{TOT}/R to find the minimum energy per bit. Figure 18.16 (a) plots the minimum energy per bit derived from such numerical computation as a function of k and for various values of data rate R. The curves saturate at high values of k since the modulation depth $m_{eff} = m/(1+m)$ cannot exceed 1. The values used for numerical computation were $Q_1 = Q_2 = 25$, $f_{res} = 25\,\text{MHz}$, $B = f_{res}/(2Q_1)$, $V_{DDp} = 2\,V_{Lm} = 0.4\,\text{V}$, $V_{Lp} = 35\,\text{mV}$, $V_{DDm} = 2.0\,\text{V}$, $\gamma = 1$, $P_e = 10^{-4}$, $L_1 = 0.5\,\mu\text{H}$. The choice of $V_{DDp} = 2\,V_{Lm}$ ensures that saturation

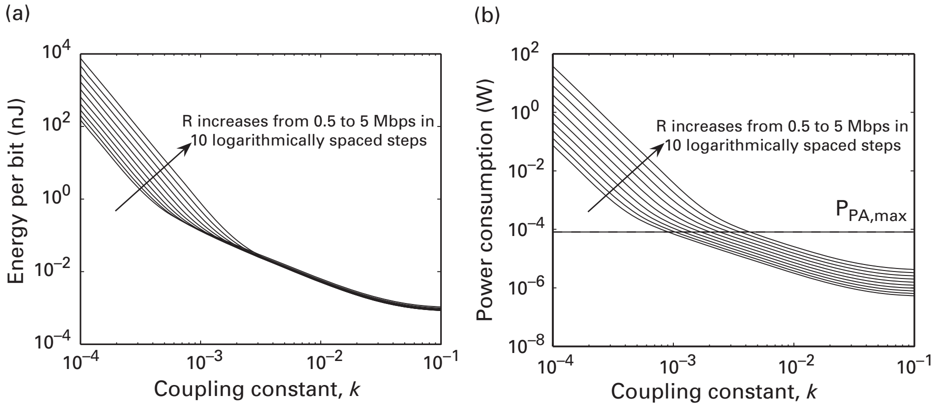


Figure 18.16a, b. (a) Plots of the minimum energy per bit derived as a function of the coupling constant k and various values of data rate R. (b) Total link power, P_{TOT}, as a function of k for various R.

value for our design, we get $P_{rad} = 9.5$ pW. The maximum radiated power density (W/m^2) at a distance R from the coil is given by

$$P_{dens} = \frac{D_0 P_{rad}}{4\pi R^2} = \frac{E_{rad}^2}{Z_0} \tag{18.42}$$

where $D_0 = 1.5$ is the maximum gain produced by a small loop antenna, E_{rad} is the intensity of the radiated electric field, and $Z_0 = 120\pi$ is the impedance of free space. Thus, from Equation (18.42), we find for $R = 30$ m that $E_{rad} = 0.69\ \mu\text{V/m}$, which is well below the FCC specification. Higher-frequency operation increases R_{rad} as does the use of more turns or a bigger coil.

18.11 Seven considerations in choosing a carrier frequency

When choosing a carrier frequency for an RF link, seven considerations are often balanced against each other:

1. **Bandwidth** – Higher frequency provides more bandwidth at constant Q.
2. **Power** – Higher baseband bandwidth, usually obtained at a higher carrier frequency, consumes more power. In a given technology, the power consumption of certain components of RF blocks, e.g., active transistors that must have adequate power gain at the RF frequency, also increases with carrier frequency.
3. **Antenna size** – Higher carrier frequencies require smaller antennas since the effectiveness of an antenna in transducing electromagnetic energy is small if its total length is significantly less than half the carrier wavelength.
4. **Communication range** – For a given power consumption, higher carrier frequency leads to a smaller communication range in accord with the Friis formula that we discussed in Chapter 17.
5. **Device Q** – Inductors generally have better Q's with increasing carrier frequency, due to steeper increase of inductive reactance with frequency than that of skin-effect or other resistive losses. Capacitors generally have better Q's with decreasing carrier frequency due to the steep increase of capacitive reactance with decreasing frequency and a decrease of skin-effect and other resistive losses with decreasing carrier frequency.
6. **The channel** – The transmission characteristics of the channel can result in higher or lower loss. For example, human tissue is extremely lossy near 433 MHz and significantly less so near 25 MHz. Diffraction, scattering, and directionality effects are wavelength dependent with higher carrier frequencies being more prone to reflection, scattering, or absorption by obstacles of a fixed size in their path than lower carrier frequencies that diffract more easily around obstacles.
7. **Federal Communications Commission (FCC) regulations** – FCC regulations on RF spectral bandwidths and intensities at various carrier frequencies are complex and need to be followed in any design.

18.12 RF antenna links for implants

Applications such as electronic pills are used for diagnosis of gastrointestinal (GI) problems. In such applications, a patient swallows a pill, and for a few hours imaging (via implanted LEDs and imagers within the pill), temperature, and pH information about the patient's GI tract can be wirelessly monitored. Such monitoring is useful in diagnoses of Crohn's disease and Celiac disease. In applications such as these, the size budget for the receiving coil/antenna in the pill is very small, and the distance between the transmitter and receiver r is relatively large. Thus, k, evaluated from Equation (16.7) in Chapter 16, is quite small, and inductive RF links become relatively inefficient. Systems that operate with far-field links at higher carrier frequencies and smaller antennas have the potential to be advantageous. Therefore, even though transmission of RF energy through the body beyond a few tens of MHz is quite lossy because of the highly conductive nature of the body at these frequencies, antenna-based systems have been and are being explored at such frequencies.

There are currently four frequency bands that have been or are being investigated. The **MICS, 402–405 MHz band**, has been the most successful thus far with low-power 0.8 Mbps commercial systems and chips for electronic pills already having been developed [7]. The latter chips also work in the **433 MHz ISM band**. Antenna, matching, fading, and body losses for 2 m communication ranges with the swallowed pill can lead to losses of 40 dB. Thus, ~99% of the power is dissipated within the body. For frequencies in the 300 MHz to 6 GHz range, the US specification is that the specific absorption rate (SAR) of this power in the body cannot exceed 1.6 W/kg in spatial peaks, e.g., in the head, say, and the whole-body average cannot exceed 0.08 W/kg. A system reported in [8] has achieved 2 Mbps at a carrier frequency of 144 MHz. Because of their promise of small antenna sizes, high bandwidth, good power efficiency, and low-cost transceivers, **3.1 GHz–10.6 GHz UWB** impulse systems have also been researched for such applications. The system reported in [9] has a 25 dB loss for every 2 cm of meat-tissue thickness in the 3 GHz–5 GHz band. UWB systems for cochlear-implant applications, which require shorter link distances, have been tested by bringing the knuckled fists of each hand together [10]. A rectenna system for transmitting RF power at 915 MHz in the **902–928 MHz band** (see Chapter 17) has achieved an RF link gain of −33 dB across 1.5 cm of bovine tissue [11]. A chip for reporting neural data through 2 cm of skin and skull in the brain and ~2 cm in air has achieved 0.33 Mbps in the 433 MHz band [12]. An integrated inductive powering system at 2.64 MHz on this chip has thus far achieved 1% power-link efficiency.

18.13 The skin depth of biological tissue

A useful parameter that characterizes RF plane-wave propagation with frequency f through a medium with conductivity σ, dielectric permittivity ε, and magnetic

Section IV

Biomedical electronic systems

19 Ultra-low-power implantable medical electronics

When one door of happiness closes another opens; but we often look so long at the closed one that we do not see the one which has opened for us.

Helen Keller

Implantable electronics refers to electronics that may be partially or fully implanted inside the body. Several implanted electronic systems today have revolutionized patients' lives. For example, more than 130,000 profoundly deaf people in the world today have a cochlear implant in their inner ear or cochlea that allows them to hear almost normally [1]. Some cochlear-implant subjects have word-error recognition rates in clean speech that are better than those of normal hearing subjects, and they understand telephone speech easily. Cochlear implants electrically stimulate the auditory nerve, the nerve that conducts electrical impulses from the ear to the brain, with an ac current. Cochlear implant subjects are so profoundly deaf that even the feedback-limited $\sim 1000\times$ gain of a hearing aid is not large enough to help them hear. Therefore, an implant that directly stimulates their auditory nerve is necessary. Cochlear implants today are partially implanted systems: the electrodes and a wireless receiver are implanted inside the body while a microphone, processor, and wireless transmitter are placed outside the body.

Patients with Parkinson's disease have had their quality of life significantly improve because of a deep brain stimulator (DBS) that has been fully implanted inside their bodies [2]. This stimulator provides ac electrical current stimulation to a highly specific region in their brain, which is most commonly the subthalamic nucleus (STN). The stimulation is an effective treatment for their uncontrollable shaking and inability to initiate movement. Some Parkinson's patients recover so well that they are even capable of dancing. Pacemakers, which were introduced several decades ago, have helped patients with disorders in their heart's rhythms lead normal and safe lives. Implanted cardiac defibrillators have been similarly beneficial.

While these and other FDA-approved clinical treatments are already in the market place, many believe that we have only scratched the surface of what will be possible in the future. For example, treatments to help cure paralysis by recording from electrical signals in the motor regions of the brain, and decoding these electrical signals to stimulate a muscle or robot arm, have already been demonstrated in patients [3]. Retinal implants for the blind function by electrically

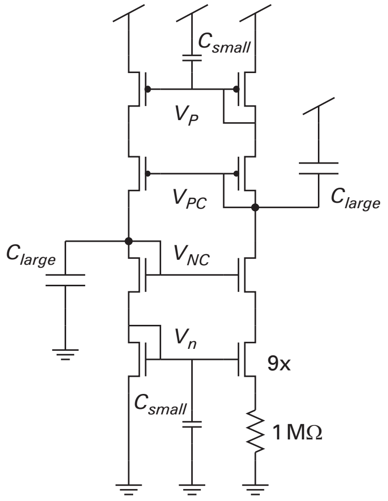


Figure 19.12. A power-supply-noise-immune and temperature-invariant constant-G_m reference for subthreshold operation. The reference also generates voltage biases that are useful for setting cascode bias voltages on a chip. Reproduced with kind permission from [17] (©2005 IEEE).

19.2.8 Robust current-and-voltage biasing

Temperature-independent biasing is *crucial* for robust operation in subthreshold circuits where currents have an exponential sensitivity to temperature. This is true even within the body where the temperature can fluctuate by 1.5 degrees centigrade in a healthy subject between day and night, and by up to 5 degrees centigrade in a subject with a fever. Figure 19.12 illustrates the PTAT biasing circuit, which leads to constant G_m-biasing in subthreshold, and therefore leads to temperature-independent biasing of all bandwidth and time-constant parameters throughout the chip. Any residual linear PTAT variation in the absolute value of a bias current (rather than exponential) has little effect on the operation of other parameters on the chip because it is automatically cancelled out, e.g., in the log ADC, or creates small variations in the noise and power that do not affect operation significantly, e.g., in the envelope detector. The relative constancy of body temperature makes such higher-order effects inconsequential.

The nominal reference current generated to bias DACs on the chip at equilibrium is

$$I_{DACbias} = \frac{25\ \text{mV} \times \ln(9)}{10^6\ \Omega} = 57\ \text{nA} \tag{19.16}$$

In practice, the current is rarely exactly that predicted by Equation (19.16), due to parameter variations and mismatch. For example, we measured 45 nA. Such variability in the absolute value of the reference does not matter as long as there are a sufficient number of DAC bits to alter currents on the chip to what is needed. The reference current is scaled up and down with mirroring in a

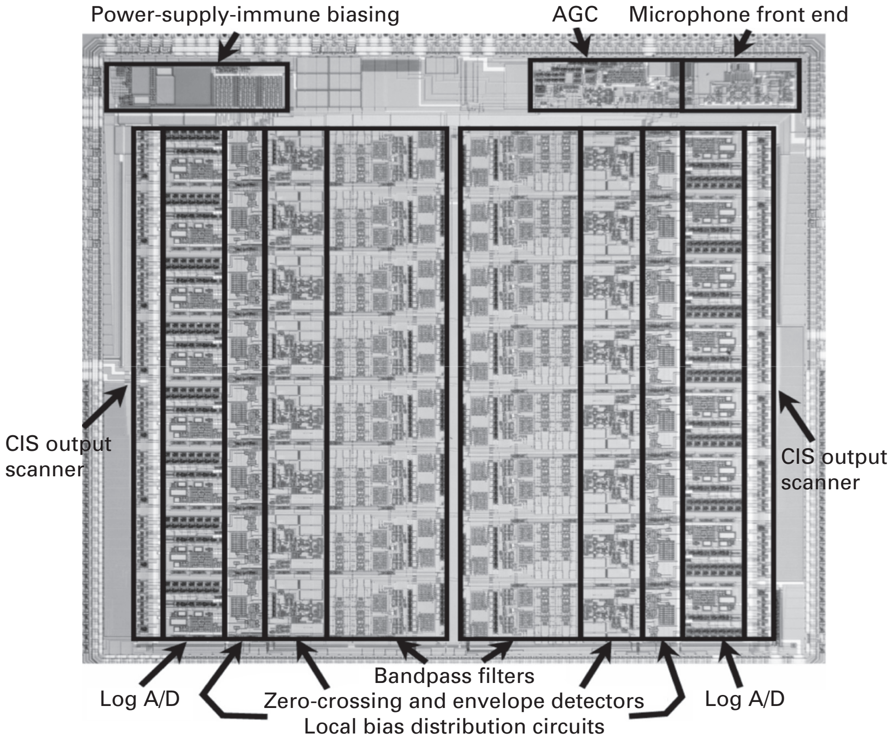

Power-supply-immune biasing
AGC
Microphone front end
CIS output scanner
CIS output scanner
Bandpass filters
Log A/D
Zero-crossing and envelope detectors
Log A/D
Local bias distribution circuits

Figure 19.13. The ultra-low-power programmable analog bionic ear processor. Reproduced with kind permission from [18] (©2005 IEEE).

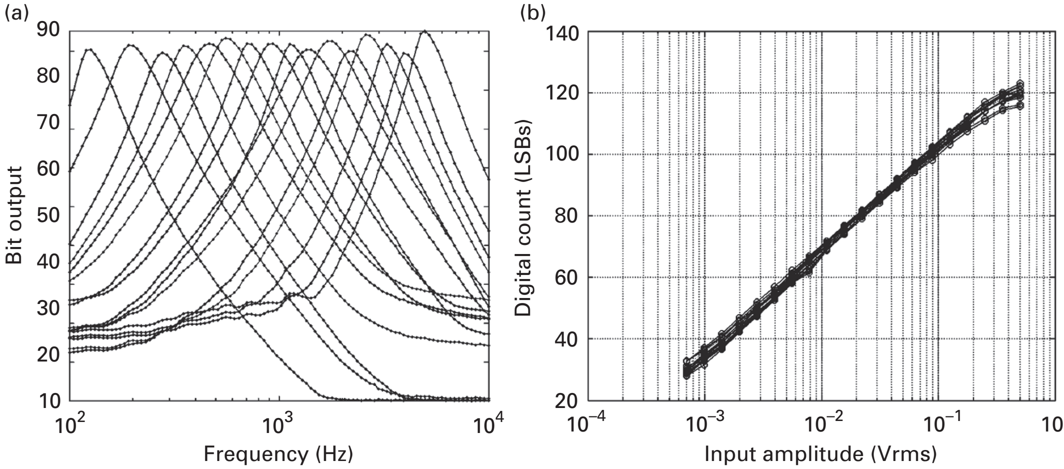


Figure 19.14a, b. Output bits from the bionic ear processor chip for different input frequencies in (a) and versus input amplitude in (b). (a) Reproduced with kind permission from [18] (©2005 IEEE). (b) Reproduced with kind permission from [17] (©2005 IEEE).

is encoded into a set of 60 electrodes as shown in Figure 19.20. Inspired by CIS, electrodes 1, 7, 13, 20,. . .,54, which are sufficiently far apart from each other such that electrode interactions are small, are simultaneously stimulated with charge-balanced information corresponding to the first DCT coefficient. Then, electrodes 2, 9, 16,. . .,55 are simultaneously stimulated with information corresponding to the second DCT coefficient. After all 6 DCT coefficients for the first column have been presented in this fashion, we turn our attention to the second column and repeat the same procedure. Now, we use information from the second column to stimulate the 60 electrodes rather than information from the first column. After all 10 columns of the block image have been represented via stimulation onto the electrodes, we pause briefly between frames, and then repeat the procedure for the next frame. If we devote 10 ms to each column, i.e., we have 100 pulses per second on any electrode with 600 DCT coefficients per second on 6 simultaneous electrodes, and create a frame pause of 20 ms, the overall frame period is $10 \times 10\,\text{ms} + 20\text{ ms} = 120\text{ ms}$ and corresponds to a frame rate of 8 Hz, which is quite adequate for vision. Jumping spiders and humans can reconstruct images by column scanning as is evident by the fact that a person walking past a door that is slightly ajar can form a complete image of what is on the other side of the door [33]. Humans constantly scan images via directed eye movements known as saccades. Non-columnar scanning patterns can also of course be implemented. Figure 19.21 shows that image reconstruction with this simple scheme for letters and faces is not perfect but reasonable. More electrodes (600 electrodes are now possible [34]), more DCT coefficients, and better basis functions could certainly improve the image representation. Time and experiments on real subjects will eventually tell if such an approach will actually succeed on blind patients. The success of similar strategies in cochlear implants does give us hope. Low-power image-processing circuits inspired by biology are discussed in Chapter 23.

19.6 Brain-machine interfaces for paralysis, speech, and other disorders

Experiments using brain-machine interfaces (BMIs) have shown that it is possible to predict intended limb movements by decoding simultaneous recordings from 10–100 neurons. See [3] for a report of the first human trials of such devices, and the references therein for recent reviews of this field. These findings have suggested a potential approach for treating paralyzed patients by recording and decoding neural data from the motor regions of their brain and using the decoded results to stimulate a prosthetic arm or muscle, or to control a computer. Such approaches could generalize to treating patients with speech disorders by recording and decoding neural data from the speech regions of their brain and using the decoded results to control their vocal tract or an artificial vocal tract, e.g., like the one described in [35]. For all such recording prosthetics, low-power neural decoding is very important. Low-power analog decoding architectures that are potentially 50× more power efficient than digital decoders have been described in [36]. Since the

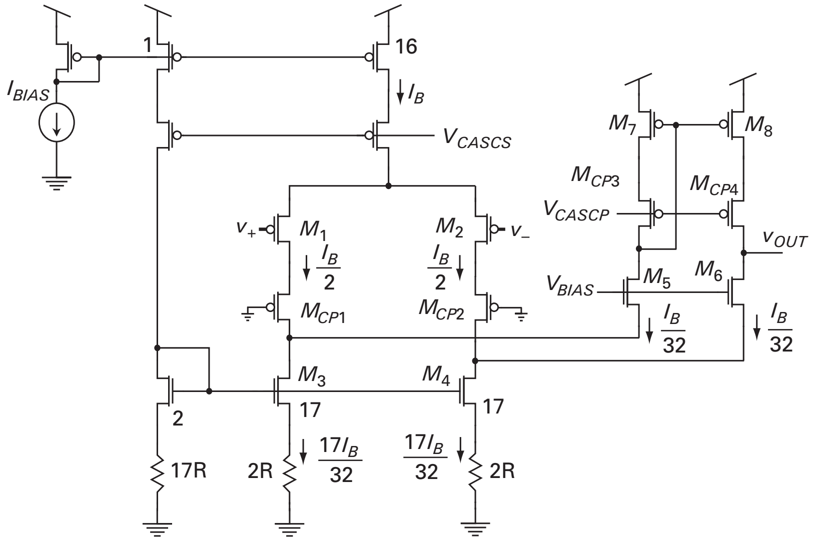


Figure 19.22. A low-noise micropower differential transconductance amplifier designed for neural-recording applications. Reproduced with kind permission from [38] (©2007 IEEE).

combined with digital processing in the external unit such that flexible configuration of the implanted analog system through external digital analysis becomes possible [29]; the flexibility of a higher-power digital system, is then only periodically necessary, saving power. However, the full flexibility of a high-power external digital system is always present if needed. As another example of the use of analog preprocessing with delayed digitization, a novel low-power spectral-analysis IC for deep-brain stimulation with local-field-potential-based recordings has been recently reported [37].

Ultra-low-power neural amplifiers are described in [38]. They are built with low-power OTAs, capacitive gain-setting elements, and adaptive elements (Chapter 11) to set floating-node voltages in a closed-loop inverting-amplifier topology. The neural amplifier described in [38] has 40 dB of gain, 3.05 μV rms input-referred noise in a 45 Hz to 5.3 kHz bandwidth, with a power consumption of 7.56 μW in a 2.8 V process. It is currently one of the most energy efficient and lowest power differential neural amplifiers. The key circuit of this neural amplifier is shown in Figure 19.22. Due to the distribution of currents in this folded-cascode topology, the negligible contribution of noise of the 2R resistors because ${g_m}^{M1,2}R_{1,2}$ is $\gg$1, and the self shunting of noise in cascode devices, this amplifier's noise is almost entirely due to the input devices M_1 and M_2, the lowest it can be in a differential amplifier. The transistors M_3 and M_4 ensure that the impedance looking into their drains is high such that most of the differential pair's output current is shunted to the folded output cascode.

Figure 19.23 shows that, in typical multi-electrode neural systems, the input-referred noise of recording electrodes has a probability distribution with some electrodes having low noise floors and others having high noise floors. The power needed to build an amplifier with a fixed bandwidth and input-referred noise

20 Ultra-low-power noninvasive medical electronics

It has long been an axiom of mine that the little things are infinitely the most important.

Sir Arthur Conan Doyle

Noninvasive medical electronics refers to electronics for medical instruments that do not invade or penetrate the body. The sensors in these instruments can and often do contact the body. Examples of such sensing include:

- Electrocardiogram (EKG or ECG) measurements of heart function.
- Photoplethysmographic (PPG) measurements of blood-oxygen saturation, or pulse oximetry.
- Phonocardiogram (PCG) measurements of heart sounds.
- Electroencephalogram (EEG) measurements of brain function.
- Magnetoencephalogram (MEG) measurements of brain function.
- Electromyogram (EMG) measurements of muscle function.
- Electrooculogram (EOG) measurements of eye motion.
- Electrical impedance tomography (EIT): measurements to infer composition of the body's tissues. Impedance cardiography (ICG) is a further specialization within the field.
- Temperature measurements.
- Blood-pressure (BP) measurements.
- Pulmonary auscultation (lung-sound) measurements.
- Biomolecular detection of small molecules, DNA, proteins, cells, viruses, or microorganisms for point-of-care or lab-on-a-chip applications, which often exploit BioMEMS (Bio Micro Electro Mechanical Systems) and microfluidic technologies, mostly in a noninvasive fashion thus far.

When such sensing is done chronically, for example as heart tags on patients with a high risk for myocardial infarction (MI), i.e., a heart attack, it is often called *wearable electronics*. The various sensors on the body may form a body sensor network (BSN) or body area network (BAN) that communicate with each other and/or the patient's cell phone, or with an RF-ID, Bluetooth, Zigbee, MICS, UWB, or other wireless receiver in the home, hospital, or battlefield [1]. Such sensors have applications in emergency, surgical, intensive-care, bedside, ambulatory, athlete, farm-animal, soldier, infant, home-care, or elderly monitoring. The need for chronic, wireless, and portable monitoring makes low-power operation important. Chronic monitoring is helpful for alarm situations, e.g., for automatically calling 911 after sudden cardiac arrest (SCA): defibrillation within the first six minutes after a heart attack is critical to saving life. Chronic monitoring is also important for providing long-term diagnostic information at a relatively cheap cost. It is widely recognized that medical costs could be greatly impacted through intelligent, networked, wireless medical-monitoring systems.

In this chapter, we will study some examples of ultra-low-power electronics for noninvasive medical monitoring applications, especially those involving cardiac function for body sensor networks or wearable systems. We shall then review developments in biomolecular sensing. For such applications, high-precision ultra-low-noise electronics for sensing, e.g., the 23-bit-precise MEMS capacitance sensor of Chapter 8, rather than ultra-low-power electronics is currently more important. Chapter 8 discusses how noise from the sensor affects the minimum detectable signal and sensitivity of an overall system including sensors and electronics. The reader may find it helpful to review the chapter since the noise-analysis and block-diagram methods of the chapter apply to all sensing systems. As we have discussed throughout the book, ultra-low-power analog design is all about minimizing noise within a tight power constraint. Thus, a good ultra-low-power analog designer is automatically a good ultra-low-noise high-precision designer, just without the power constraint. Many of the ultra-low-power circuits that we describe in this chapter such as an EKG amplifier rely on principles for good low-noise circuit design.

We shall begin by describing an analog electronic chip that models the heart and circulatory system with electrical circuit analogs: pressure corresponds to voltage and the volume velocity of blood flow corresponds to current. Such analogs can help electrical engineers rapidly and intuitively understand mechanical systems like the heart, as we will show. Understanding the heart is useful for understanding several noninvasive measurements including the EKG, PPG, PCG, and BP measurements. We then describe how the EKG arises in the body and how it relates to normal cardiac function. Our understanding of the heart and EKG will then tie together EKG, PCG, and BP measurements and reveal the relationships between them. Then, we shall describe how to build a micropower EKG amplifier and make it insensitive to 60 Hz noise. This noise typically overwhelms EKG signals that are extremely tiny compared with it. We shall then describe how pulse oximeters work (based on the PPG) and relate them to other measures of cardiac function like the EKG. We briefly describe how low-power pulse oximeters may be built and present PPG measurements from such an oximeter. Then, we shall study an example of an ultra-low-power battery-free medical tag that harvests RF energy to function. It exploits PCG-based cardiac sensing for ultra-low-power operation, is useful for measurements of heart-rate and blood-pressure variations, and can enable audio localization and audio alarms in case the patient needs attention. We shall then describe work on a low-power communication system that uses the conductive body itself as a ‘wire’ to transmit information between sensors implanted within the body or on it.

We shall conclude by reviewing work on sensors for biomolecular detection in the optical, mechanical, and electrical domains. In our brief review, rather than focus on circuit techniques, many of which are further instantiations of low-noise analog techniques that we have already described, and which can be found in the cited papers, we shall focus more on the key principles behind these schemes. We shall discuss ‘label free’ and ‘labeled’ detection of molecules, including examples for DNA detection.

20.1 Analog integrated-circuit switched-capacitor model of the heart

Figure 20.1 shows an artist’s depiction of the heart, which is about the size of one’s fist. The heart is a four-chambered organ with upper right and left atria and lower right and left ventricles. By convention, the right and left refer to the heart of the person that the reader is viewing, not the reader’s right and left, so they appear flipped in Figure 20.1. The right atrium receives blood from the upper regions of the body via the superior vena cava, and from the lower regions of the body via the inferior vena cava. It pumps this blood to the right ventricle. The right ventricle in turn pumps this blood to the left and right lungs via the pulmonary artery, where the blood’s hemoglobin molecules bind oxygen. The left atrium receives blood from the lungs via the pulmonary veins and pumps the blood to the left ventricle. The left ventricle in turn pumps this blood to the rest of the body via the aorta. The left ventricle is the biggest, most important, most contractile, and most muscular chamber. It pumps 75 ml to 90 ml of blood with every heart beat, wringing the blood out in a twisting helical action due to muscles that surround its walls. In typical, healthy individuals who are not super-athletic, the healthy heart beats about once every second at rest. The heart is itself sustained by blood from coronary arteries that branch off the aorta.

From the description above, it may appear that the heart is a serial four-phase pump with the right atrium, right ventricle, left atrium, and left ventricle being activated in turn. A healthy heart actually operates on two phases known as diastole and systole. During diastole, the right and left atria contract and pump in synchrony while the ventricles relax and fill with blood. During systole, the right and left ventricles contract and pump in synchrony while the atria relax and fill with blood. During diastole, blood from the atria is conveyed to the ventricles via the opening of valves between the atrial and ventricular chambers; these valves are otherwise closed. During systole, blood from the ventricular chambers is conveyed to body organs via the opening of valves in the output arteries of the heart; these valves are otherwise closed. The valves ensure that the flow of blood is always in the right direction much like diodes regulate the direction of current flow in an electrical circuit.

Figure 20.2 illustrates an analog electrical circuit model of the heart. Fluid pressure of the blood is mapped to voltage while volume velocity of blood is mapped to current. With this mapping, compliance, the inverse of stiffness, is

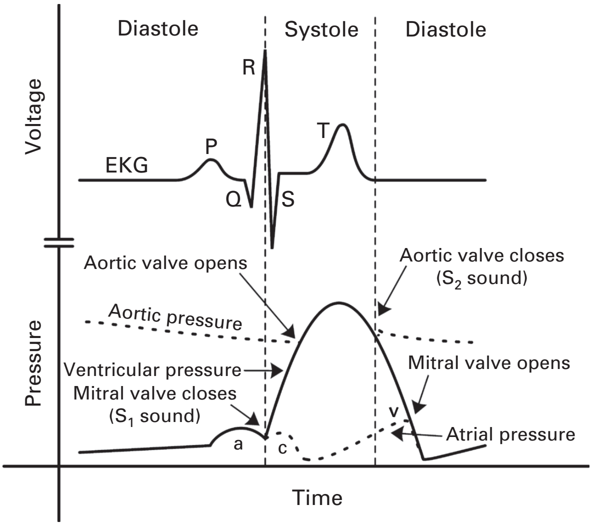


Figure 20.6. The timing relationships between electrical (EKG), blood pressure (BP), and heart-sound (S_1 and S_2) waveforms.

an oscilloscope. The recording was obtained by using skin electrodes called 'dermatrodes'. These electrodes are typically made with a gel electrolyte that permits a 'salt-bridge' connection between electronic conduction in artificial instruments and ionic conduction in the body's tissues. This particular trace corresponds to a difference in voltages between the left arm and right arm like in the example above. Referring to Figures 20.5 and 20.6, the P wave is the initial small bump in the EKG trace and corresponds to atrial depolarization. The P wave is very tiny in Figure 20.5 and more exaggerated in Figure 20.6. The large QRS complex refers to the undershoot-overshoot-undershoot (Q, R, and S, respectively) that follow the small P wave and correspond to ventricular depolarization; in general, the structure of the complex depends on the electrode configuration and the subject. The T wave is the broad bump after the QRS complex and corresponds to ventricular repolarization. Figure 20.6 shows the temporal relationships between the EKG, PPG (S_1 and S_2 sounds), and BP variations. The mechanical BP responses initiate near the peaks of the EKG waveforms such that the EKG is predictive of them: the ventricular pressure begins to rise near the QRS peak and the ventricular pressure begins to fall near the T peak; the atrial contraction begins near the P peak. The S_1 sound usually occurs soon after the QRS peak and the S_2 sound usually occurs after the T-wave.

In healthy subjects, when the Ca^{2+} ions depolarize the atrial walls, $\boldsymbol{H}(t)$ is small and points downward and to the left leading to the P wave component of the electrocardiogram (keep in mind, all coordinates are in body coordinates of the subject facing the reader). The vector $\boldsymbol{H}(t)$ then swings back to the right as Ca^{2+} ions begin to charge up the left-ventricular septal wall; it then grows large in signals [8]. This advantage arises because the common-mode signal is now attenuated to a value given by

$$\begin{aligned} V_{CM}^{body} &= I_{60}\left(\frac{Z_{gnd}}{1+A_{lp}} \,||\, \frac{1}{C_{body}s}\right) \\ &\approx I_{60}\frac{Z_{gnd}}{1+A_{lp}} \end{aligned} \tag{20.4}$$

where A_{lp} is the loop gain of the common-mode feedback loop at 60 Hz. The impedance of the grounding electrode is effectively reduced by feedback to a low value such that the interfering 60 Hz signal does not affect the common-mode voltage on the body much.

The benefit of such active grounding is that the power-supply voltage and CMRR requirements of the differential-input amplifier needed to obtain a given SNR can then be greatly reduced. The reduction in power-supply voltage obviously reduces power. Since large-size components need more power to be consumed to maintain bandwidth, the decrease in the size of components needed for good matching also lowers power consumption. The symmetry of a truly differential topology is also beneficial for attaining good matching and CMRR. For all of these reasons, we shall use the technique of active grounding in the micropower EKG amplifier that we now describe.

Figure 20.7 shows the schematic of the micropower EKG amplifier, which is based on a classic two-gain-stage instrumentation-amplifier topology with

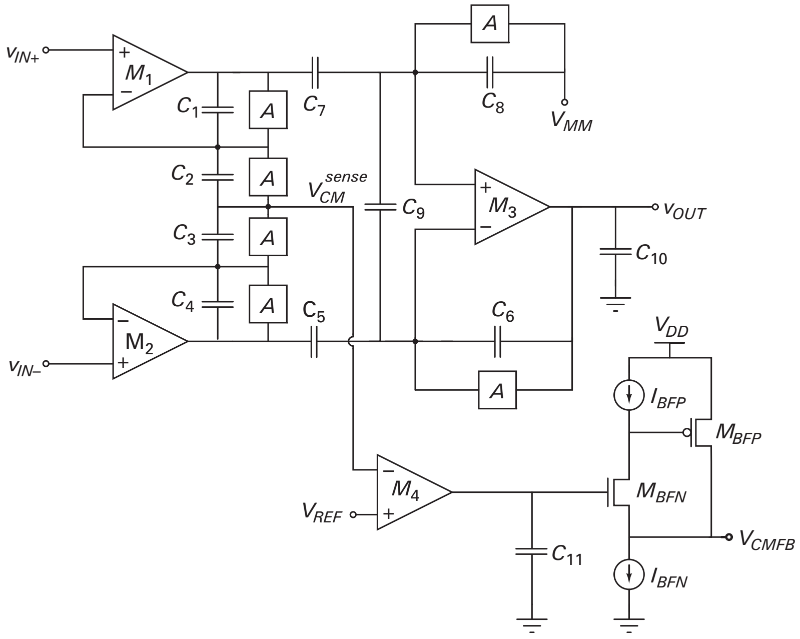


Figure 20.7. Micropower electrocardiogram amplifier schematic. Reproduced with kind permission from [10] (©2009 IEEE).

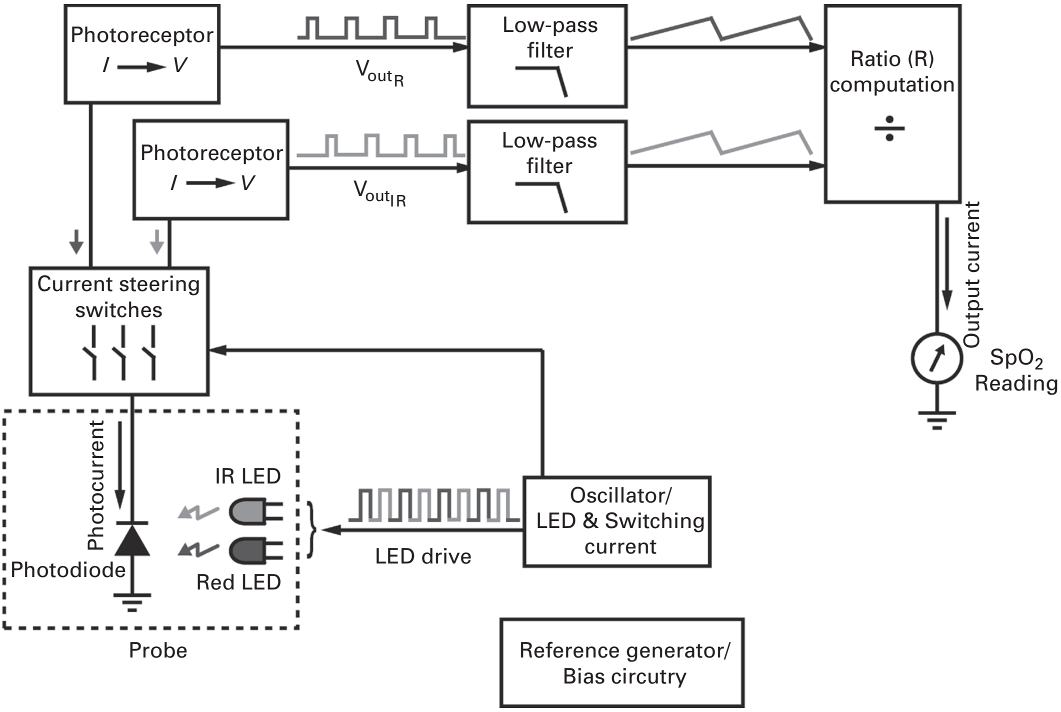


Figure 20.9. Architecture of a low-power pulse oximeter. Reproduced with kind permission from [11] (©2009 IEEE).

variable to regulate the level of an anesthetic. It is also an important vital sign. If oxygen saturation is not in the 80%–100% range, it indicates that the heart and/or the lung is not functioning as one would normally expect. Increasingly, pulse oximeters are ubiquitous in several clinical settings as *the* vital-sign monitor, for example, in intensive care, operating rooms, emergency care, birth and delivery, neonatal and pediatric care, sleep studies, and veterinary care. They are likely to be an important part of wireless medicine in the future. Their popularity stems from the fact that they are easy to use and provide valuable information: at one glance, they can provide vital-sign information about heart rate, oxygen saturation, and indirectly about blood-pressure variations. They are also noninvasive. It is remarkable that one can extract oxygen-saturation information noninvasively without having to prick a finger and do a blood test. The basic idea behind the operation of the pulse oximeter is that well-oxygenated blood is composed mostly of oxyhemoglobin, which is ‘blood red’ and does not absorb red light much, while de-oxygenated blood is composed mostly of deoxyhemoglobin, which is ‘dark red’ and absorbs red light more. Therefore, by shining red LED light through the index finger, where there is a good artery that is not buried deeply beneath the skin, and monitoring absorption of the light on the other side of the finger with a photoreceptor, we can obtain information about the oxygen content

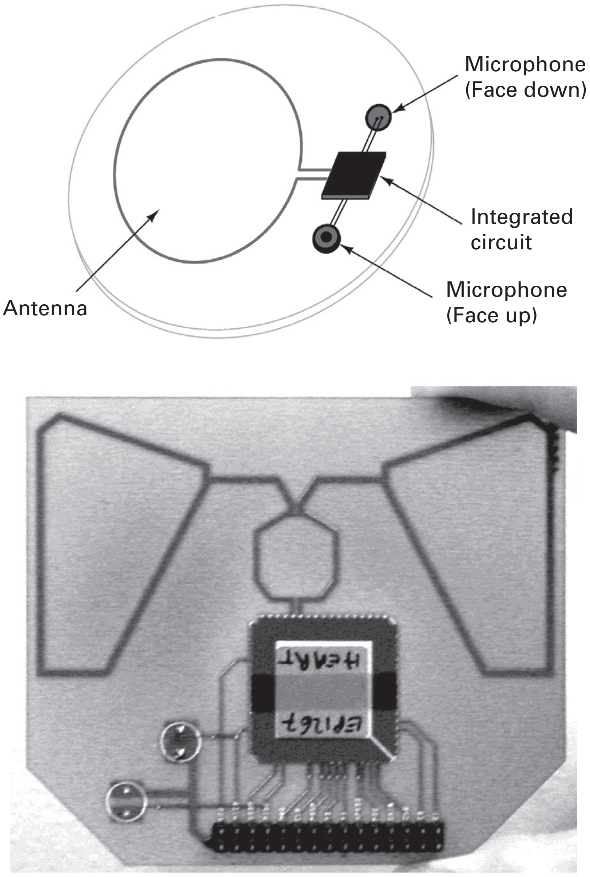


Figure 20.12. A low-power battery-free cardiac medical tag that functions through RF energy harvesting. A cartoon version with two microphones is shown at the top while the actual experimental tag is shown in the bottom figure. Reproduced with kind permission from [12] (©2009 IEEE).

arrival times between sounds from two speakers separated by 12 ft at a microphone can help localize where the microphone is located. Even when the loudspeakers output relatively low-frequency 230 Hz tone bursts, a localization accuracy of 0.6 m with a standard deviation of 0.4 m is achievable [12]. The loudspeakers can turn on only when alarm signals are received from a tag. The loudspeaker tones themselves then provide an audible alarm and can trigger other power-hungry sensors such as video cameras to turn on.

Commercial microphones contain built-in JFET preamplifiers as we have described in Chapter 19. Since heart sounds are relatively loud and low in bandwidth (typically 10–250 Hz), the microphone can be biased with currents far below those specified by the manufacturer to save power. The JFET then operates in its insensitive linear regime. For example, a Panasonic omni-directional condenser electret microphone (WM-63PR) is a small cheap thin device (thickness = 1.3 mm,

20.7 Biomolecular sensing

Biomolecular sensing has grown explosively in recent years due to rapid advances in miniaturization and fabrication in the fields of BioMEMS and microfluidics. We shall discuss some principles behind biomolecular sensing and some recent developments. The brief account below should not be seen as a review of the field but rather as a collection of examples that paint a picture that is relevant to the themes of this book. This picture is painted from the perspective of a bioelectronic engineer. It is too early to tell where this young and still maturing field will have its highest impact.

The word BioMEMS is used as a catch-all phrase to describe just about any system for biological applications that exploits MEMS or nanotechnology to function. Therefore, it is sometimes used to describe microfluidic systems as well. A good overview of BioMEMS for sensing and medical diagnostic applications may be found in [20]. The review in [21] discusses possibilities for therapeutics and drug delivery. The description in [22] provides examples of the potential of BioMEMS for tissue engineering. We shall focus almost exclusively on sensing.

Six common biological sensors include sensors for whole cells, viruses, microorganisms like bacteria, DNA, proteins, and small molecules. Examples of small molecules include glucose, lactate (a common by-product of anaerobic metabolism), H^+ ions (pH sensing), urea, or neurotransmitters such as dopamine. The sensing is invariably accomplished in the following three steps, which are summarized below:

1. The binding of one unit of a specific target species to one unit of a complementary capture species changes the analog state of a mechanical, optical, or electrical variable. The target species can be labeled with a molecular tag or fluorescence marker to facilitate optical detection when the binding event occurs; or it can be labeled with a magnetic bead to facilitate electrical detection when the binding event occurs. Certain forms of sensing that we shall discuss do not require the use of such labels and are therefore termed *label free*.
2. The change in the mechanical, optical, or electrical state is passively transduced to yet another domain, e.g., a mechanical state change is passively transduced to an electrical state change, or a mechanical state change is passively transduced to an optical state change. In most sensors, 0 or 1 such passive conversions between domains is typical.
3. The final transduced signal is then actively amplified, usually in the electrical or optical domain.

The detection of a species with high specificity (only the target species triggers detection), high sensitivity (a few molecules or units of the target species are detectable), in a short time (a second), over a wide dynamic range in concentration (e.g., fM to 0.1M), and in a low-cost fashion is challenging. Label-free sensing is more convenient than labeled sensing but performance and cost are key determinants in measuring the pros and cons of one sensing scheme versus another.

Section V

Principles for ultra-low-power analog and digital design

21 Principles for ultra-low-power digital design

A small leak will sink a great ship.

Benjamin Franklin

In this chapter, we shall review important principles for ultra-low-power digital circuit and system design. We shall focus on operation with extremely low power-supply voltages and on subthreshold operation, although we shall provide some analysis of moderate-inversion and strong-inversion operation with the EKV model as well. As Chapter 6 on deep submicron effects in transistors discussed, because threshold voltages scale significantly less strongly than power-supply voltages, subthreshold operation is an increasingly dominant fraction of the voltage operating range. Subthreshold operation has become and will continue to get increasingly fast such that ultra-low-power operation in this regime does not sacrifice bandwidth in many applications. In biomedical and bioelectronic applications, subthreshold operation is ideal since bandwidth requirements are typically modest while energy efficiency is of paramount importance. An insightful paper by Meindl, that was way ahead of its time, pioneered subthreshold digital design [1]. An analysis by Burr and Peterson analyzed the optimal energy efficiency of ultra-low-power subthreshold circuits [2]. A more recent publication by Vittoz [3], the pioneer of subthreshold analog design, has analyzed issues in subthreshold digital design using his EKV model. Through such pioneering and other work, subthreshold digital design has been revived and is an active field of research in several academic and industrial institutions.

We shall begin by discussing the operation of a subthreshold CMOS inverter. Operation in the subthreshold regime is highly subject to transistor mismatch. We present equations that help quantify transistor sizing and a lower limit to the power-supply voltage needed for robust subthreshold operation. The CMOS inverter serves as a good vehicle for understanding the three kinds of power dissipation in digital CMOS circuits, namely, dynamic power, static power, and short-circuit power. Dynamic power is dominant during switching and is due to the dissipation caused by on current flowing in a PMOS transistor charging a node capacitance or by on current flowing in an NMOS transistor discharging a node capacitance. Static power or 'leakage power' is caused by off current or subthreshold leakage current or background current present in an NMOS or PMOS transistor that has been turned off. Static power dominates during static operation when there is no switching and node voltages are static. Short-circuit power is caused by charging and discharging pathways being simultaneously active in a circuit such that on currents flow in a short-circuit fashion from V_{DD} to ground rather than to or from node capacitances. Short-circuit power contributes during switching when the NMOS and PMOS devices have comparable currents. Short-circuit power can typically be neglected unless the rise or fall times of the input are significantly larger than the rise or fall times of the output.

At an optimal value of the power-supply voltage, we shall show that the balance between dynamic energy and static energy leads to the lowest total energy dissipated per cycle of operation. For many computations, this optimal point causes the power-supply voltage to be in the subthreshold regime where the on-current to off-current ratio is maximum. At this optimum power-supply voltage, the threshold voltage may be set by body biasing to fulfill frequency-of-operation requirements set by the bandwidth of the task. Alternatively, if thresholds are fixed, the optimal power-supply voltage and threshold voltage determine an optimal frequency of operation. Increased switching activity causes more dynamic power dissipation and shifts the optimal point to occur at lower power-supply voltages. Reduced duty cycles in the computation cause more static power dissipation and shift the optimal point to occur at higher power-supply voltages.

The dynamic power dissipation of nodes with high activity factors (frequent switching activity) such as clock nodes can be considerable. We shall discuss how gated clocks and adiabatic clocks can greatly reduce such power dissipation. Adiabatic clocks are based on a more general computational paradigm termed adiabatic computing. While the paradigm of adiabatic computing is completely general and applies to all logic nodes, not just to clocked nodes, its high overhead has thus far made it of practical importance only in circuits with high node capacitances.

We then briefly review the key ideas behind architectural and algorithmic techniques for power reduction including parallelism, pipelining, ordering, symmetry, and algorithmic-efficiency improvements. Since many of these techniques have been reviewed previously [4], and in texts on low-power digital design [5], our treatment will focus only on presenting the key insights. Recent examples of systems that embody the principles described in this chapter include a sensor processor in [6], a microcontroller in [7], and a JPEG co-processor in [8].

21.1 Subthreshold CMOS-inverter basics

Figure 21.1 (a) shows a classic CMOS inverter circuit and Figure 21.1 (b) shows its input-output characteristic. The inverter inverts its digital inputs from 1 (V_{DD}) to 0 (GND or ground) and vice versa. In our case, the power-supply voltage V_{DD} is sufficiently small such that both the NMOS and PMOS transistors operate with subthreshold currents for all inputs. That is, V_{DD} is less than either $|V_{TP}|$ or V_{TN}. If we desire the switching transition point of the inverter to be at a symmetrical $V_{DD}/2$, as shown in Figure 21.1 (b), then we must require that

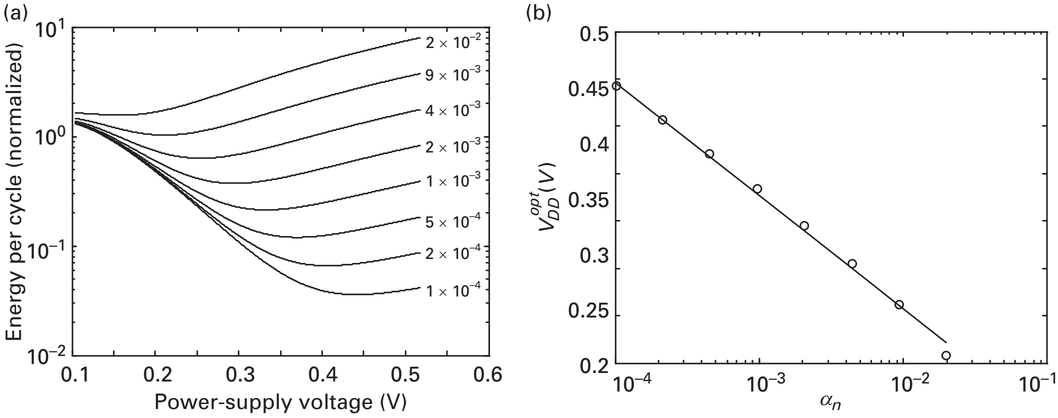


Figure 21.3a, b. (a) Variation of energy efficiency with power-supply voltage in weak inversion for various α_n, a dimensionless measure of the activity factor. (b) Variation of the optimal power-supply voltage in weak inversion with α_n.

$V_{DD}^{opt} = 3\phi_t/\kappa$, its minimum possible value. The inclusion of short-circuit power dissipation of Equation (21.20) has no discernible effect on the curves of Figure 21.3 (a) or those of Figure 21.3 (b).

Substitution of the optimal V_{DD}^{opt} extracted from Equation (21.31) in Equation (21.30) along with simple algebra reveal that the optimal value of E_{TOT}^{sub}, E_{opt}^{sub}, is given by

$$E_{opt}^{sub} \approx (\beta N_{TOT} N_{LD})\left(C_L V_{DD}^2\right)\left(\alpha_n + 2\alpha_n^{1.25}\right) \qquad (21.32)$$

We notice from Equation (21.32) that the dynamic energy and static energy terms are balanced at the optimum in that both decrease with decreases in α_n, and both terms are of comparable magnitude. We see in Figure 21.3 (a) that the optimum energy is lower at lower values of α_n in accord with Equation (21.32). At low values of α_n, corresponding to small activity factors in Equation (21.29) or Equation (21.30), the relative importance of static energy increases, and V_{DD}^{opt} increases in Equation (21.31) to reduce static energy, consistent with Figure 21.3 (b). Hence, lowering the activity factor of a computation always saves energy and always increases the optimum power-supply voltage.

Duty cycling the input lowers the effective value of f_{clk} in Equation (21.28) and increases the contribution of static energy. Thus, duty cycling the input effectively acts to increase N_{LD} in Equation (21.29), which can also be directly seen from Equation (21.27). The effective increase in N_{LD}, which is a measure of the logic depth of the computation, with the use of duty cycling may appear physically strange at first. The increase in N_{LD} with duty cycling is a mathematical energy equivalence: a large logic depth corresponds to a slow f_{clk} with lots of static devices dissipating energy while a few dynamic devices in the logic path actually switch and compute as the signal propagates through the depth of the logic; duty cycling

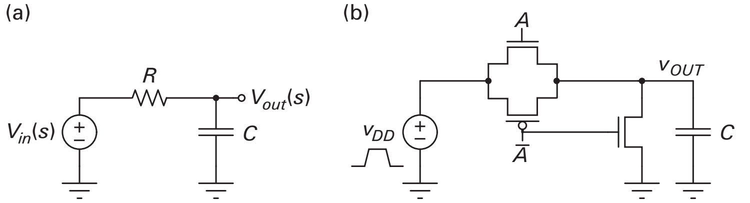


Figure 21.7a, b. (a) An RC circuit. (b) The fundamental canonical circuit of adiabatic computing.

The average power dissipated through the resistor R is given by

$$\begin{aligned} P_R(j\omega) &= \frac{|I_R(j\omega)|^2 R}{2} \\ P_R(j\omega) &= \left(\frac{|V_{in}(j\omega)|^2}{2R}\right)\left(\frac{\omega^2(RC)^2}{1+\omega^2(RC)^2}\right) \end{aligned} \tag{21.48}$$

If $\boldsymbol{\omega} \ll \boldsymbol{RC}$, we can ignore the ω^2 term in the denominator and find that the power dissipated is given by

$$P_R(j\omega) = \left(\frac{|V_{in}(j\omega)|^2}{2R}\right)\omega^2(RC)^2 \tag{21.49}$$

The energy dissipated per charge-discharge cycle is the average power integrated over a period:

$$\begin{aligned} E_R(j\omega) &= \left(\frac{|V_{in}(j\omega)|^2}{2R}\right)\omega^2(RC.RC) \times \frac{2\pi}{\omega} \\ E_R(j\omega) &= \left((2\pi)C\frac{|V_{in}(j\omega)|^2}{2}\right)(\omega RC) \\ E_C(j\omega) &\equiv C\frac{|V_{in}(j\omega)|^2}{2} \\ Q &\equiv \frac{1}{\omega RC} \\ &\boxed{E_R(j\omega) = E_C(j\omega)\left(\frac{2\pi}{Q}\right)} \end{aligned} \tag{21.50}$$

Thus, if the input angular frequency $\omega \ll 1/(RC)$, the $E_R(j\omega)$ energy dissipated per charge-discharge cycle is significantly less than the average reactive energy per cycle $E_C(j\omega)$. The parameter $1/(\omega RC) \equiv Q$ is the quality factor of this system, which is proportional to the ratio of the reactive energy to dissipated energy $E_C(j\omega)/E_R(j\omega)$; a high-$Q$ system dissipates little energy per cycle compared with the stored reactive energy. If $\omega \to 0$, $Q \to \infty$, the energy dissipation goes to zero and we have charged the capacitor adiabatically with no heat dissipated in the resistor.

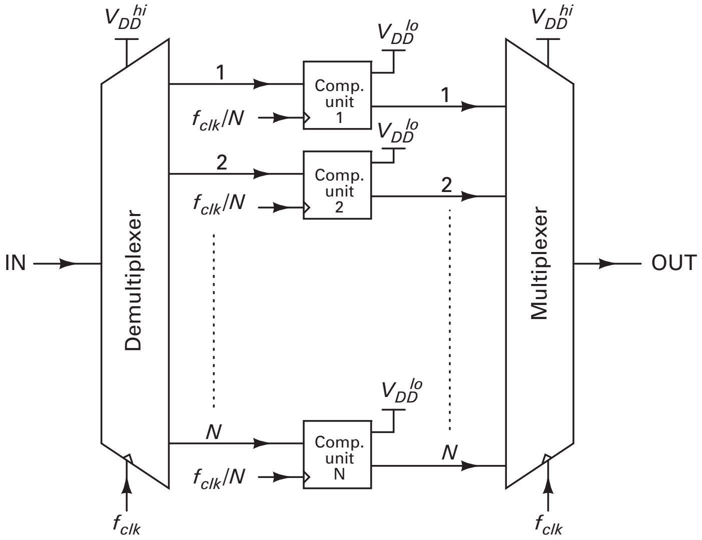


Figure 21.9. An architecture that illustrates the basic ideas behind low-power parallel computing.

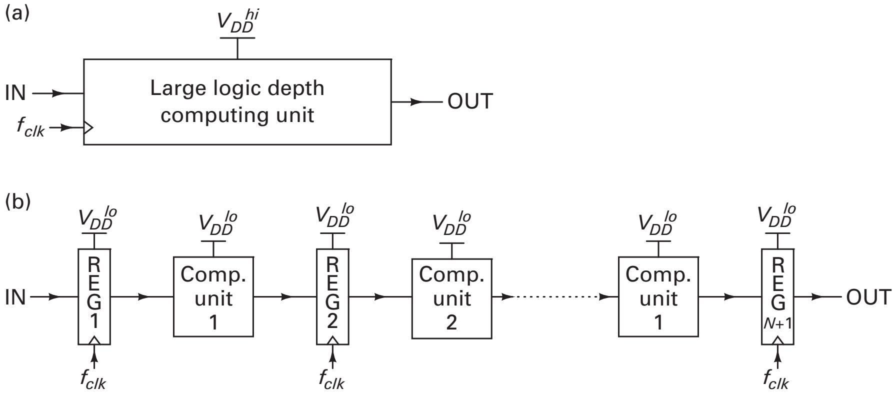


Figure 21.10a, b. (a) A large logic depth computing unit. (b) A pipelined version of the same computing unit that is more energy efficient.

do it faster. Rather than increasing speed, however, if the decrease in logic depth between pipelined stages is used to lower power-supply voltage and maintain throughput, there is more energy-efficient operation. Figures 21.10 (a) and (b) illustrate the concept. There is, however, in a fashion analogous to parallelism, an optimal amount of logic depth between registers. If there is too much pipelining, i.e., there are frequent registers between computational stages of small logic depth, the register power dissipation increases annulling the power savings in the computational stages. Pipelining is more amenable as an architectural low-power

22 Principles for ultra-low-power analog and mixed-signal design

If you shut the door to all errors, truth will be shut out.

Rabindranath Tagore

In this chapter, we present ten general principles for ultra-low-power analog and mixed-signal design. We shall begin by comparing the paradigms of analog computation and digital computation intuitively and then quantitatively. The quantitative analysis will be based on fundamental relationships that dictate the reduction of noise and offset with the use of power, area, and time resources in any computation. It shall reveal important tradeoffs in how the power and area resources needed for a computation scale with the precision of computation in analog versus digital systems. From these results, we shall discuss why, from power considerations, there is an optimum amount of analog preprocessing that must be performed before a signal is digitized. If digitization is performed early and at high speed and high precision, as is often the case, the power costs of analog-to-digital conversion and digital processing become large; if digitization is performed too late, the costs of maintaining precision in the analog preprocessing become large; at the optimum, there is a balance between the two forms of processing that minimizes power.

There are detailed similarities between power-saving principles in analog and digital paradigms because they are both concerned with how to represent, process, and transform information with low levels of energy. We shall itemize and discuss several of these similarities. We shall derive from Shannon's theorem on information theory that there is a lower bound on the energy needed to process a bit of information, kT ln(2), irrespective of whether the information is represented and processed in an analog fashion or in a digital fashion. We shall present five considerations that determine power in all systems, namely task, technology, topology, speed, and precision.

Our analysis suggests that the optimal method for designing the ultra-low-power systems of the future involves collective analog and hybrid computation. Such computation is a hybrid mix of the paradigms of analog and digital computation and is one of the secrets behind the awe-inspiring energy efficiency of biological systems. Chapter 23 dwells on systems involving neuronal computation and Chapter 24 dwells on systems involving cellular computation in some depth.

These systems embody some of the general principles of collective analog and hybrid computation outlined in this chapter. Here, we shall just outline the key ideas and an example that illustrates how such systems can be architected in principle.

A general trend in such bio-inspired systems and in highly energy-efficient systems built by engineers with no knowledge of biology is the increasing presence of feedback interactions between analog and digital portions of a mixed-signal system to improve energy efficiency. Such feedback is already implicitly present in well-known ADC architectures such as sigma-delta and successive-approximation converters, and in ADCs with digital error correction and calibration, all of which represent highly energy-efficient ADC topologies (see Chapter 15). We shall show that such architectures represent a special case of a more general architecture termed a *hybrid state machine* (HSM), a generalization of finite state machines from the digital domain to the hybrid analog-digital domain. Indeed, Chapter 15 on low-power analog-to-digital conversion has already provided an example of such an HSM to create an ADC. Here, we shall show why such mixed-signal feedback architectures are likely to be increasingly important in the future of low-power design and how they may be used to architect general digitally programmable analog systems and hybrid cellular automata of high energy efficiency in the future.

We discuss how to maintain robustness and flexibility in low-power systems and how robustness and flexibility trade with efficiency. We shall find that feedback and learning provide functions in analog systems analogous to error correction and data compression in digital systems: thus feedback is extremely important in providing a better robustness-efficiency tradeoff in analog systems just as error correction, redundancy, and compression do in digital systems.

We shall outline ten principles for architecting low-power systems, whether analog, digital, or mixed-signal. The digital manifestation of some of these principles has been discussed in Chapter 21 on ultra-low-power digital design. The analog application of these principles occurs throughout Chapters 11 to 20. We show how general-purpose energy-inefficient systems need to evolve to more energy-efficient special-purpose systems as time and learning allow them to acquire knowledge about their environments.

Applications of many of these principles in concrete contexts can be seen in the biomedical systems described in Chapters 16, 17, 18, 19, and 20. In many highly miniature implantable and noninvasive biomedical systems, the sensing, wireless, power-management, and actuation/stimulation power consumption dominate the power budget of the system. Thus, the costs of analog-to-digital conversion and digital processing can be largely irrelevant. In such cases, the principles for low-power design articulated in this chapter are still applicable to just the analog, RF, and power-management portions of such systems, and the latter chapters have actually applied them in a concrete way. In such systems, ultra-low-power digital processing as discussed in Chapter 21 may still be helpful in improving the energy efficiency of the dominant analog systems via feedback, calibration, learning,

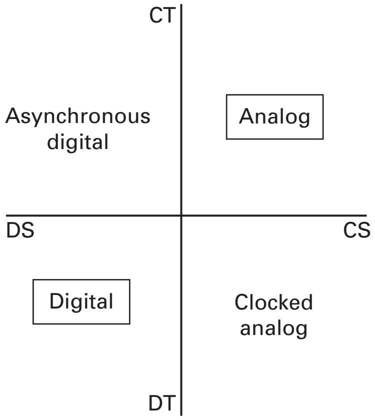


Figure 22.1. The four types of mixed-signal systems. Reprinted with kind permission from [11].

data compression, and mismatch-compensation functions. We discuss how sensors and actuators are also analog processing systems that automatically fit into the framework for low-power mixed-signal design discussed in this chapter. However, their physical state variables are not necessarily electrical.

22.1 Power consumption in analog and digital systems

Figure 22.1 shows that systems can be continuous or discrete in their signal levels (CS or DS), and operate in continuous time or discrete time (CT or DT). Thus, there are four kinds of systems corresponding to each of the four quadrants of Figure 22.1. Examples of systems in the first analog quadrant include continuous-time filters or the retina in the eye. Examples of systems in the second asynchronous digital quadrant include communication systems and pulsatile neuronal systems. Examples of systems in the third digital quadrant include microprocessors and digital signal processors. Examples of systems in the clocked analog quadrant include switched-capacitor filters and synchronous analog-to-digital converters. For simplicity, we will begin by studying the boxed analog and digital systems in the first and third quadrants of Figure 22.1, respectively. Such systems are purely analog, being continuous in both signal levels and time, or purely digital, being discrete in both signal levels and time. Systems in the first and third quadrants represent extremes of the two forms of computation and thus represent a good place to begin; systems in the second and fourth quadrants can then be understood by extrapolation from these extremes. Later, we shall discuss examples of hybrid systems that operate with continuous and discrete signals, and in continuous and discrete time.

We begin with an intuitive comparison of the pros and cons of analog versus digital computation.

Table 22.1 Comparison of analog and digital computation

Analog	Digital
1. Compute on a continuous set, e.g., [0,1], [0, V_{DD}].	1. Compute on a discrete set, e.g., {0,1}, {0, V_{DD}}.
2. The basis functions for computation arise from the *physics* of the computing devices: current-voltage curves of NFETs, PFETs, capacitors, resistors, floating-gate devices, KVL, KCL, etc. The amount of computation squeezed out of a single transistor or device is high.	2. The basis functions for computation arise from the *mathematics* of Boolean logic: logical relations like AND, OR, NOT, NAND, XOR, etc. The transistor is used as a switch, and the amount of computation squeezed out of a single transistor is low.
3. One wire represents many bits of information.	3. One wire represents one bit of information.
4. Computation is offset-prone since it is sensitive to the parameters of the physical devices.	4. Computation is not offset-prone since it is relatively insensitive to the parameters of the physical devices.
5. Noise due to thermal fluctuations in physical devices.	5. Noise due to round-off error and temporal aliasing.
6. Signal not restored at each stage of the computation.	6. Signal restored at each stage of the computation.
7. In a cascade of analog stages, noise starts to accumulate and build up.	7. Round-off error does not accumulate significantly for many computations.
8. Not easily programmable.	8. Easily programmable.
EFFICIENT	**ROBUST**

The intuitive comparison in Table 22.1 suggests that the large number of degrees of freedom exploited by analog computation in each device make it efficient. However, they also make it sensitive to errors in these degrees of freedom. In contrast, the few degrees of freedom exploited by digital computation in each device cause it to be less efficient but the loss of many degrees of freedom is traded for high levels of robustness. Furthermore, there is no general-purpose scalable signal-restoration paradigm in analog computation since we do not know where to restore the signal to in general. Therefore, unlike an arbitrarily complex digital system, an arbitrarily complex purely analog system will always accumulate enough noise such that its function is eventually compromised. To quantify the intuition of the table, we must begin by understanding how noise in devices and noise accumulation affects the precision of analog computation.

22.1.1 Noise in MOS transistors

The input-referred noise voltage $\overline{v_n^2}$ in a transistor with transconductance g_m, width W, length L, oxide capacitance C_{ox} over an operating frequency range of (f_l, f_h) at room temperature T due to thermal noise and $1/f$ noise (or offset) is given by

$$\overline{v_n^2} = \int_{f_l}^{f_h} \left(\frac{4\gamma kT}{g_m}\right) df + \left(\frac{B_{imp}}{C_{ox}^2 WL}\right) \frac{df}{f} \tag{22.1}$$

where γ is a noise factor, and B_{imp} is the trap-impurity density, a determinant of the transistor's threshold-voltage offset and its $1/f$ noise. Chapters 7 and 8 provide an in-depth discussion of noise in a transistor. The g_m versus I relationship in a saturated transistor is given by

Table 22.2 Scaling relationships between power, speed, and precision

Power $= f$ (Task, Technology, Topology, Speed, Precision)		
Task	Speed-precision law	Comment
1. First-order active filtering.	$P_{GmC} \sim (f_{crnr})(SNR)$	1. SNR^2 in above-threshold operation. See Chapter 12.
2. Second-order active filtering.	$P_{2Gm} \sim (Q^3 + Q) f_{crnr}(SNR)$ $P_{3Gm} \sim (2Q + 1) f_{crnr}(SNR)$	2. P_{2Gm} and P_{3Gm} represent two topologies. See Chapter 13.
3. Passive filtering or adiabatic operation.	$P_{adbtc} \sim \frac{f_{in} SNR}{Q}$	3. The Q can be for a tuned or untuned system. See Chapters 21 and 15.
4. Adaptive Class-A current-mode filtering.	$P_{adpt\text{-}clssA} \sim (f_{crnr})(SNR_{max})$ $SNR_{max} \neq$ dynamic range	4. SNR_{max} not being equal to the dynamic range implies that power can be saved. See Chapter 14.
5. Analog-to-digital conversion.	$P_{ADC}^{nthrm} \sim f_s 2^{Nbits}$ $P_{ADC}^{thrm} \sim f_s 2^{2Nbits}$	5. The superscripts *nthrm* and *thrm* correspond to non-thermal-noise-limited and thermal-noise-limited topologies. See Chapter 15.
6. Operational amplification.	$P_{amp} \sim (GBW)\left(\frac{1}{\overline{v_n^2}}\right)$ $P_{amp} \sim (GBW) SNR$	6. $GBW =$ Gain-bandwidth product; $\overline{v_n^2} =$ input-referred thermal noise.
7. RF Impedance-modulation communication.	$P_{trnscvr}^{wk} \sim SNR(\Delta f)$ $P_{trnscvr}^{strng} \sim SNR^{1/3}(\Delta f)^{1/3}$	7. The *wk* and *strng* superscripts are for weak and strong coupling. The power includes the transmitting power-amplifier and receiving mixer power. See Chapter 18 for an in-depth discussion of bit-error rate and further details.
8. Sensing.	$P_{snsr} \sim f_{snsr} SNR_{max}$	8. Noise of sensor must be added to electronic noise to determine SNR. SNR_{max} is not necessarily equal to dynamic range in an adaptive sensing system with gain control. See Chapter 19 on imagers and other sensing, Chapter 8, and Chapter 20.
9. Active non-resonant oscillators, e.g., ring, relaxation.	$P_{osc} \sim f_{osc} SNR$	9. SNR corresponds to period jitter.
10. LCR tuned oscillators.	$P_{LCR} \sim \frac{f_{osc} SNR}{Q_{osc}}$	10. Since only the dissipated energy is provided by the active system rather than all the energy, watches and high-Q oscillators are energy efficient. See Chapter 13 for a discussion of LCR systems.
11. Narrowband LCR tuned amplifiers and sensors.	$P_{tuned-amp} \sim \frac{f_{crr}}{Q} SNR$	11. The f_T of the transistors used must be larger than f_{crr} to get significant gain near f_{crr}. Note that signal bandwidth f_{crr}/Q determines power not carrier bandwidth f_{crr}. The SNR degradation of amplifiers is proportional to $(1 + f_{crr}/f_T)$ such that f_{crr} must be significantly less than f_T.
12. Digital systems.	$P_{dgtl} \sim f_{clk} E_{TOT}$ $P_{dgtl} \sim f_{clk}$ Poly$(\log_2(1 + SNR))$	12. Poly( ) is for polynomial function. See Chapter 21 for details on E_{TOT}.

allows flexibility of use with several stimulation circuits and paradigms. Thus, the optimal point for digitization was not picked for pure power-efficiency reasons but also picked to preserve robustness, flexibility, and modularity. The fact that 86 patient parameters could be altered with 373 programmable bits on the chip, and that back-compatibility and modularity were preserved as well imply that a good tradeoff between flexibility and efficiency was made: a 35x power reduction (see Chapter 19) with most of the flexibility intact. Indeed, it was the flexibility and programmability in the analog system that allowed testing of the chip on a deaf subject: She replaced her external digital processor with this chip processor, and was able to understand speech with it on her first try.

22.4 Common themes in low-power analog and digital design

Table 22.3 below summarizes 11 common themes in low-power analog and digital design, many of which have manifested in other chapters of the book.

Table 22.3 Common themes in low-power analog and digital design

Low-power analog	Low-power digital
1. Power $\sim$ Speed $\times$ Poly($2^{\text{precision}}$). So*, slow-and-parallel operation* is a big win. For example, in the bionic-ear processor, many slow-and-parallel moderate-precision log ADCs digitize information more efficiently at its output than one high-speed high-precision linear ADC at its input.	1. Power $\sim$ Speed $\times$ Poly(precision). So, *slow-and-parallel operation* is a big win: $N(f/N)C\left(V_{DD}^{lo}\right)^2$ is lower than $fC\left(V_{DD}^{hi}\right)^2$ in above-threshold operation. Parallelism has an overhead cost in the input front-end latches and output multiplexing stages, which need to operate fast to maintain the original speed, so there is an optimum amount of parallelism. See Chapter 21.
2. *Noise and offset management* to maintain speed and precision while lowering power is crucial. Lowering V_{DD} by too much can hurt dynamic range due to the need to operate transistors in saturation. A deep understanding of low-noise design and feedback design is essential.	2. V_{DD} *and threshold-voltage management* to maintain speed and limit static subthreshold leakage power while lowering dynamic and overall power is crucial. Feedback techniques that optimize V_{DD} for energy efficiency or for "just in time" computing are very useful.
3. *Compressive functions* like AGCs that reduce dynamic range at their output save power for later stages, which can operate at lower levels of *SNR* at any instantaneous setting of the AGC, e.g., 80 dB to 60 dB dynamic range reduction by the AGC in the bionic-ear processor.	3. *Compressive functions* like AND that reduce switching activity save power for later stages.
4. *"Pipelined" designs* with many stages of computation that are each doing less are often better than designs where one stage is doing too much. However, there is usually an optimum, e.g., a cascaded-gain amplifier has a better gain-bandwidth product for the same power than a single-stage amplifier but too many stages of gain will hurt its noise performance.	4. *Pipelining* a computation allows one to lower V_{DD} for each simpler stage of the computation. The reduced logic depth reduces the impact of leakage power and improves energy efficiency. However, there is usually an optimum amount of pipelining because latches between pipelined stages consume overhead power.

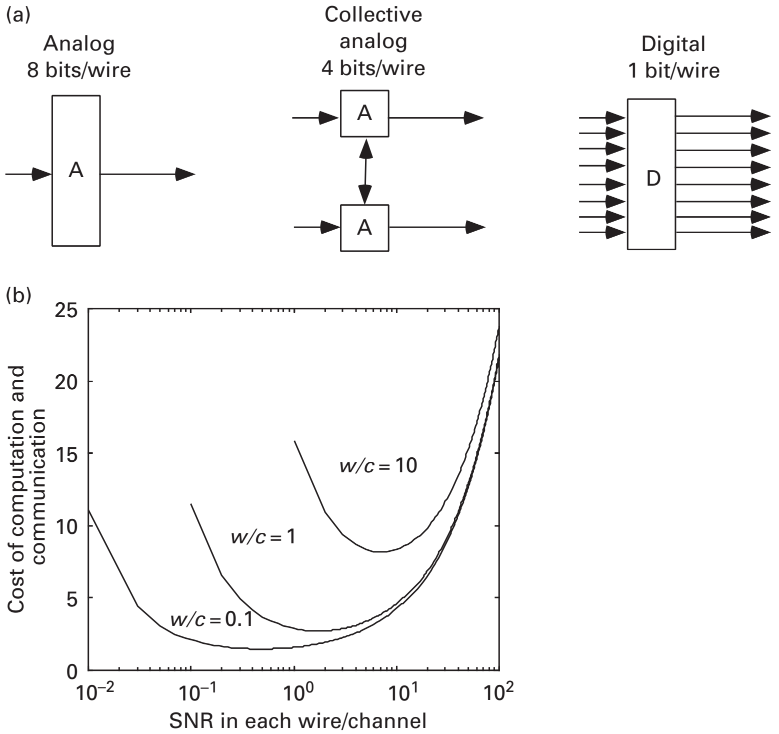


Figure 22.7a, b. (a) The idea behind collective analog computation. (b) The optimal *SNR* per computing channel in collective analog computing. Reprinted with kind permission from [11].

with one computing unit as in traditional analog computation. Thus, the paradigm incorporates the best ideas from the analog and digital worlds, and is likely a very important reason for the energy efficiency of biological computation [11].

One key point to note is that the interactions between analog units in the collective analog topology of Figure 22.7 (a) can be digital. In fact, this is often advantageous because the digital signals then serve the dual role of performing signal restoration within each unit and of being the interaction signals amongst analog units. Periodic signal restoration of analog variables via digitization is critical for enabling scaling of collective analog systems to systems of arbitrary complexity. The digitization must not be done too early such that the costs of signal restoration rise and the power of analog computation is not exploited; the digitization must not be done too late such that the costs of maintaining precision on a single analog channel rise. Thus, just as in the architecture of Figure 22.5 (a), there is an optimal point for digitization [11].

A second key point to note is that collective analog computation is implemented in various embodiments in nature, e.g., computation with cell-cell interactions amongst neurons in the brain, computation with gene-protein and protein-protein interactions amongst molecular state variables within the cell, computation with

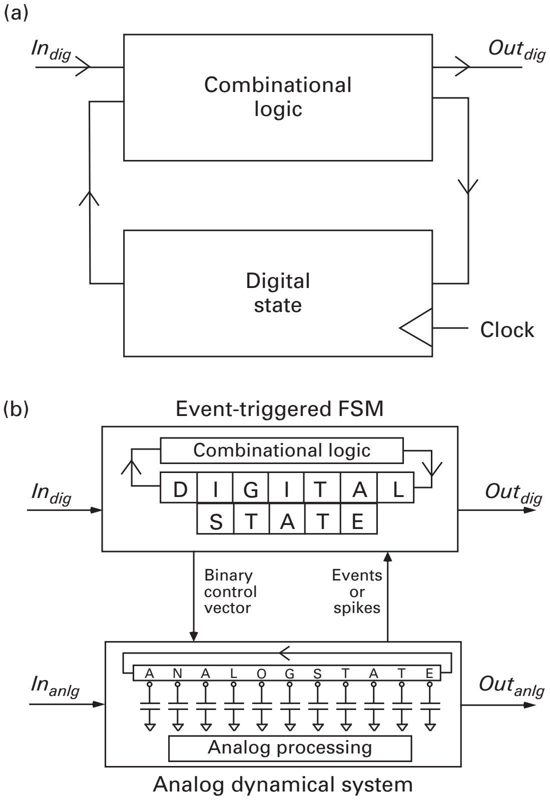


Figure 22.9a, b. (a) A finite state machine. (b) A hybrid state machine (HSM). Reprinted with kind permission from [13].

The events may be periodic synchronous edges as in traditional digital systems but both synchronous and asynchronous edges are allowed. Thus, HSMs operate over all four domains in Figure 22.1. The digital state in the FSM alters parameters and topology in the ADS via the binary control vector in Figure 22.9 (b) that controls the configuration of switches in the ADS. Hence the analog and digital portions of an HSM each affect and control each other, via the feedback loop shown in Figure 22.9 (b). Local analog feedback within the ADS and local digital feedback, present in every FSM, implement feedback loops within the analog and digital portions as well.

The HSM architecture of Figure 22.9 (b) is a generalization of several special-purpose mixed-signal dynamical systems that already exist in the field of analog-to-digital converters (ADCs). We shall provide six examples.

1. The ΣΔ ADC shown in Figure 15.5 has an ADS that is composed of a differential integrator and comparator. The comparator outputs events that are fed back to the integration input of the ADS via a binary control vector. The binary control vector is a simple 1-bit signal that alters DAC parameters that are input to the ADS. The FSM is represented by the digital intelligence that feeds back a 1

9. Just as clock-gating reduces energy in digital systems, HSMs can enable event-driven analog systems that minimize energy by only turning on high-precision and/or high-speed analog operation in certain states, and by turning off analog devices that are not needed.
10. Both neuronal computation amongst cells in the brain and within molecules in the cell are well described by an HSM framework (see Chapters 23 and 24). Therefore, we have an existence proof from biology of the power of mixed-signal computing with feedback. In summary, the robustness-efficiency and flexibility-efficiency tradeoffs of a system are greatly improved by exploiting the best of the analog and digital worlds.

One challenge in implementing mixed-signal HSM architectures on a large scale is the cross-coupling of undesirable digital signals into analog units, which compromises analog efficiency. A second challenge is that of designing feedback and calibration loops that automatically compensate for device variability in deep submicron processes in a manner that trades robustness for efficiency in an intelligent fashion. Fortunately, such problems are somewhat alleviated in collective analog systems that do not need high levels of local precision, in deep submicron processes where area is a relatively abundant resource, and in fully differential implementations. Furthermore, in energy-efficient subthreshold systems, digital voltage and current levels are modest such that power-supply crosstalk is inherently smaller. Another possible solution is to separate analog and digital systems on different dies and/or technologies that interact via parallel 3D interconnect.

Biology solves such problems very cleverly by using highly localized electrochemical coupling of signals across synapses in neurons or highly specific chemical binding of molecules in cells. Time will tell whether we will be able to architect such systems on a larger scale successfully as nature has done. Efforts to create algorithmically programmable analog systems in engineering have been explored for image-processing applications [16], [17].

22.8 General principles for low-power mixed-signal system design

We shall now present ten principles for low-power design that apply to digital, analog, and mixed-signal systems. These principles summarize and distill the essence of several concepts and examples discussed throughout the book. Just as in the seven benefits of feedback of Chapter 2, there is redundancy and overlap in these principles: each principle emphasizes a somewhat different aspect of the same underlying truth.

22.8.1 Encode the computation in the technology efficiently

As we have discussed extensively throughout the book, low-power design is inseparable from information-processing design. Given a function $\boldsymbol{Y}(t) = f(\boldsymbol{X}, t)$, basis functions $\{i_{out1} = f_1(\mathbf{v}_{in}), i_{out2} = f_2(d\mathbf{v}_{in}/dt), i_{out3} = f_3(\int \mathbf{v}_{in}), \ldots\}$ formed by the

Section VI

Bio-inspired systems

23 Neuromorphic electronics

It is the unification and simplification of knowledge that gives us hope for the future of our culture. To the extent that we encourage future generations to understand deeply, to see previously unseen connections, and to follow their conviction that such endeavors are noble undertakings of the human spirit, we will have contributed to a brighter future.

Carver Mead

Biological systems are *the* most energy-efficient systems in the world. For example, the ~ 22 billion neuronal cells of the brain dominate the $\sim 14.6\,\mathrm{W}$ brain power consumption of an average 65 kg male [1], [2]. These numbers imply a power consumption of $\sim 0.66\,\mathrm{nW}$ per neuron. The hybrid analog-digital brain performs *at least* 6×10^{16} FLOPS (floating-point operations per second) such that its energy efficiency is a staggeringly low 0.24 fJ/FLOP. This energy efficiency is about 5–6 orders of magnitude more efficient than that of even the most energy-efficient digital microprocessor or digital signal processor ever built. The human eye's retina consumes nearly $\sim 3.4\,\mathrm{mW}$, making the 135 million photoreceptor array in the eye the lowest power wide-dynamic-range imager and image compressor ever built. The retina in the eye performs sophisticated analog gain control, analog spatial filtering, and analog temporal filtering such that nearly $\sim 36\,\mathrm{Gb/s}$ of wide-dynamic-range raw image data from its photoreceptor array is compressed to $\sim 20\,\mathrm{Mb/s}$ of useful optic-nerve spiking output information. The human ear's nano-fluidic, electro mechanical, and electro chemical technology implements more than 1 GFLOPS of analog filtering, analog gain control, and analog spectral-analysis computations in $\sim 14\,\mu\mathrm{W}$ of power and is a marvel of nanotechnology. The human ear could run on a AAA battery for 15 years. The entire human body only has a basal metabolic power consumption of ~ 81 W [2]. The references and computations that lead to energy estimates for the brain, the body, the eye, and the ear are provided in the Appendix section of this chapter. The Appendix also provides numbers for the power consumption of the other organs of the body. In Chapter 24, we shall discuss the energy consumption of the body's cells.

The overall engineering specifications of the brain, the eye, and the ear, which are listed in Tables 23.1, 23.2, and 23.3 respectively, are awe-inspiring. The brain has a fantastic 3D interconnect technology that enables a fan-in and fan-out of nearly 6,000 connections per neuron compared with ~ 6 in the logic gates of today's microprocessors. Its architecture enables it to vastly outperform even the

Table 23.1 Specifications of the cochlea (and auditory system)

The cochlea and auditory system	
Input dynamic range	~ 120 dB
Output dynamic range of nerve firing	~ 40 dB
Power dissipation	$\sim 14\,\mu$W
Power supply voltage	~ 150 mV
Detection threshold at 3 kHz	0.05 Angstroms at eardrum
Frequency range	20 Hz – 20 kHz
Number of auditory nerve fibers	$\sim 35{,}000$
Filter bandwidths	$\sim 1/3$ octave
Phase locking threshold	~ 5 kHz
Inter-aural time discrimination	$\sim 10\,\mu$s
Loudness discrimination	1 dB
Frequency discrimination at 1 kHz	2 Hz
Computational rate	$> \sim 1$ GFLOPS
Length of uncoiled pea-sized cochlea	35 mm

Table 23.2 Specifications of the retina (and visual system)

The retina and visual system	
Dynamic range	10^8:1
Power dissipation of retina	~ 3.4 mW
Power supply voltage	~ 70 mV
Detection threshold of rod photoreceptor	1 photon
Bandwidth of response	12 Hz (rod), 55 Hz (cone)
Number of optic nerve fibers	~ 1.2 million
Spatial acuity at 20/20 vision	~ 2 arc minutes (60 minutes/degree)
Output bit rate of optic nerve	~ 20 Mb/s
Vernier hyperacuity	0.2 arc minutes
Number of photoreceptor cells	~ 5.5 million (cones); $\sim$130 million (rods)
Number of retinal synapses	~ 1 billion
Computational rate	> 10 GFLOPS
Area of $\sim$160 μm thick retina	2500 mm^2

most advanced supercomputers and robots at the tasks it has evolved to be good at, for example, sensorimotor processing, control, pattern recognition, and learning. The eye can detect a single photon and has a spatial acuity of 2 arc minutes that is as good as that determined by diffraction limits of optical physics. The visual system exhibits vernier hyper-acuity that allows it to detect line misalignments that are an order-of-magnitude below that limited by the spatial sampling of the photoreceptors from the eye. At its thermal-noise-limited threshold, the ear can detect a vibration of one-twentieth the diameter of a hydrogen atom at the ear drum. The auditory system can tell the difference in the time of arrival between the left ear and the right ear to within 10 μs even though the time constants of its cells are only 1 ms. The piezoelectric amplifying outer hair cells of the ear have a

Table 23.3 Specifications of the brain

The brain	
Kinds of two-terminal synapses	50–100
Kinds of one-terminal ion channels	50
Power supply voltage	$\sim 125\,\text{mV}$
Number of synapses	240×10^{12}
Number of neurons	~ 22 billion
Number of glial cells	~ 200 billion
Brain power dissipation	$14.6\,\text{W}$
Average firing rate	$\sim 5\,\text{Hz}$
Computational rate	$> 6 \times 10^{16}$ FLOPS
Energy efficiency	< 0.24 fJ/FLOP
Average connections per neuron	~ 6000
Surface area of cortex	$\sim 2500\,\text{cm}^2$
3D synapse grid size of $\sim 2\,\text{mm}$ thick cortex	$1.1\,\mu\text{m}$

piezoelectric coefficient that is a few orders of magnitude larger than that of any artificial piezoelectric known to man. Taken together, these specifications manifest the fantastic technology integration, sophisticated signal processing, and clever architectural design that nature has accomplished over more than a billion years of evolution. Since food, which ultimately provides energy, was a very scarce resource in the history of animal evolution, biological systems have been pushed to get more and more out of less and less by amazingly energy-efficient designs [3], [2]. As Thoreau said, "Nature is full of genius," and we can learn from her. She has figured out how to take many imprecise, noisy parts and put them together to create precise, complex, robust systems that operate in real time with phenomenally high energy efficiency.

Electronics inspired by neurobiology was termed *neuromorphic electronics* by the field's founder, Professor Carver Mead [4], [5]. Mead pointed out that the use of the physical basis functions of analog computation, the intimate integration of logic and memory through mostly local wiring, and the learning capabilities of neurobiological systems were key ingredients to their energy efficiency [4]. It is worth pointing out that *all* biological systems are energy efficient, not just neurobiological ones. For example, in Chapter 24, we shall discuss the amazing energy efficiency of a single cell that performs sophisticated biochemical computations in its gene-protein network. In Chapter 24, we will also discuss how to build *cytomorphic electronics*, a novel form of electronics introduced in this book to describe electronics inspired by cellular and molecular biology. In general biologically inspired engineering systems can be termed *biomorphic systems*. Examples of fluid-mechanical biomorphic systems include airplanes, which take inspiration from the winged flight of a bird. Examples of chemical or material biomorphic systems include the design of coatings for "self-cleaning" windows, which take inspiration from the design of a lotus leaf. Through its clever use of hydrophobicity and hydrophilicity at the leaf surface, the lotus plant's leaves remain

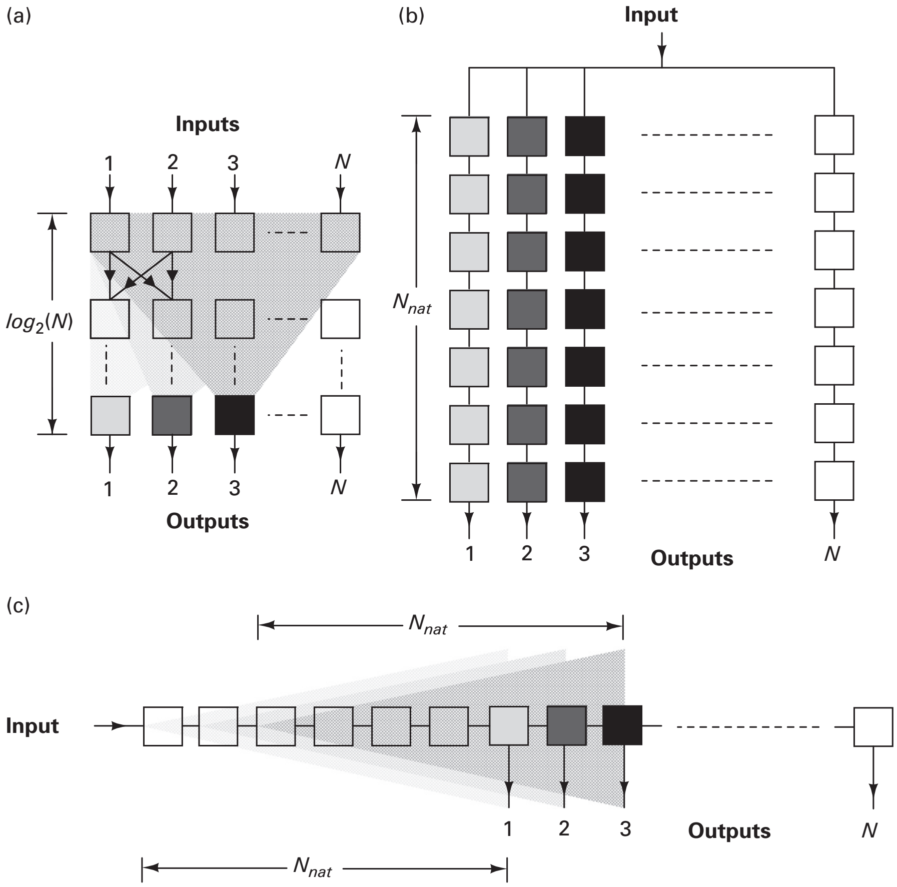


Figure 23.4a, b, c. Comparison of the FFT (a), analog filter bank (b), and cochlear architecture (c) for spectrum analysis. Reprinted with kind permission from [12] (©2009 IEEE).

need N_{nat} filtering blocks per stage. Since there are N such stages, which, unlike in the cochlea, are not shared in an exponentially tapered transmission line, the hardware costs, as measured by the total number of blocks, scales like $N \times N_{nat} \propto N^2$ if β is fixed. Figure 23.4 (b) illustrates the architectural scaling in this case. The time taken for spectrum analysis in the filter bank is given by the sum of the settling times of the N_{nat} filtering blocks in each stage, which still scales like N as in the cochlea. The fast Fourier transform (FFT), which, unlike the cochlea and constant-Q filter banks has fixed-bandwidth frequency bins, takes $O(N\ln(N))$ time (measured by the number of multiply-and-add operations) and $O(N\ln(N))$ hardware (measured by the number of multipliers and adders) to perform spectrum analysis as shown in Figure 23.4 (a). The output bins of the cochlea and filter banks are available and updated in parallel, which allows them to continuously monitor the whole spectrum. This behavior is in contrast to that of most commercial RF spectrum analyzers, which are of the swept-sine or signal, subtract these out from the signal, and amplify the subtracted residue to then detect weaker signals. The latter strategy may be useful in RF, where a dominant interferer can make it hard to detect desired weaker signals.

We shall end our tour of the cochlea by noting that, while biology has inspired several interesting architectures like the RF cochlea and companding architecture in engineering, engineering has in turn been useful to biology. A simple feedback and circuit model of the cochlea showed how fast 100 kHz amplification with slow 1 ms outer hair cells is possible in the cochlea, a two-decade-long mystery in hearing [23]. The essence is to realize that negative feedback in the cochlea serves to speed up time constants and increase Q via a simple root-locus plot as it does in several feedback systems (see discussions of feedback in Chapters 2, 9, and 10). Thus, the open-loop time constant of the outer hair cell is not predictive of its closed-loop time constant in the cochlea in much the same manner that the open-loop 10 Hz time constant of an operational amplifier is not predictive of its 1 MHz closed-loop performance in a negative-feedback buffer. In fact, the work in [23] showed that the gain-bandwidth product of the outer hair cells is large enough such that the use of negative feedback and sufficient gain bandwidth easily enables fast amplification and provides a good fit to experimentally observed cochlear data. Circuit models of biology can undoubtedly exploit the intuitive power of circuits and feedback to shed insight into the partial differential equations of biology in several other systems in the future, insight that is not easily attained by highly computationally intensive brute-force computer simulations. We shall return to this theme in the next chapter, where we discuss circuit models of gene-protein and protein-protein networks within the cell.

23.5 A bio-inspired analog vocal tract

Figure 23.9 shows a conceptual diagram of how speech synthesis and hearing analysis can be put together in a feedback loop analogous to the phase locked loop (PLL) discussed in Chapter 18. The feature comparator compares a spectral representation of the output synthesized by a vocal tract with a spectral representation of what has just been heard. The spectral representations are created by auditory processors that may be sophisticated and mimic the cochlea, e.g., a companding spectral representation; or they may be more straightforward spectral representations, e.g, that output by the bionic-ear processor described in Chapter 19 or an FFT. The results of the comparisons are used to drive a vocal-tract controller that outputs articulatory motor-control information that drives the vocal tract to produce sound. Such motor information could include configurations for articulators like the tongue, lips, jaw, velum, and larynx position. The feedback loop is architected such that the speech synthesized is both a good match to what has just been heard and can be produced by relatively smooth articulator dynamics that minimize articulator state changes and consequently muscle effort in the vocal tract. The auditory processors and feature comparator

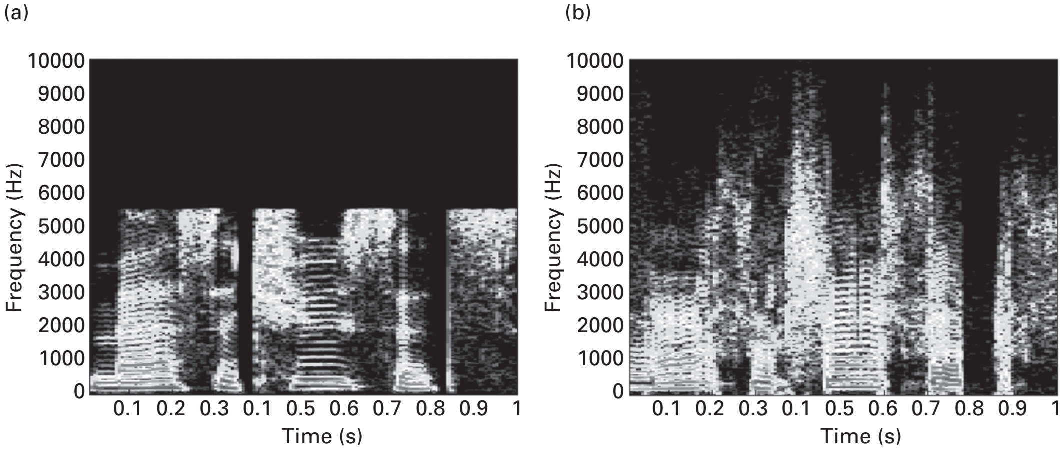


Figure 23.11a, b. (a) Spectrogram with intentional lowpass filtering above 5 kHz. (b) Spectrogram output by the analog vocal tract that reintroduces missing components in (a). Reprinted with kind permission from [18] (©2008 IEEE).

use of transconductor-capacitor circuits [18]. Feedback circuits ensure that dc offset does not greatly accumulate and degrade performance in the transmission line. More details of the analog vocal tract, including a description of how cross-sectional areas for various sounds are determined and related to articulator movements to create a babble code book are described in [18].

Figure 23.11 (a) shows a spectrogram of the word "Massachusetts" with intentional lowpass filtering at 5.5 kHz. Figure 23.11 (b) shows a re-synthesized version of this spectrogram using the vocal tract in a SLL configuration such that we try to match the sound that we have just heard. The spectrogram of the original recording (Figure 23.11 (a)) and the synthesized sound (Figure 23.11 (b)) show good matching between trajectories and formants indicating good performance by the analysis-by-synthesis technique. Furthermore, high-frequency speech components that were missing in Figure 23.11 (a) have been re-introduced by the AVT in Figure 23.11 (b). The AVT synthesizes all and only speech signals and thus provides a measure of signal restoration. Such signal-restorative properties are important in robust speech recognition in noise. Indeed, tests of the recognition of vowel-consonant syllables with the AVT in an SLL show improved performance in noise.

23.6 Bio-inspired vision architectures

Two examples of bio-inspired vision chips include the silicon retina and fly-vision-inspired motion-detection chips. The silicon retina, an electronic model of the retina, a thin sheet of tissue in the eye, performs spatial bandpass filtering and compression and was first built by Carver Mead [44]. Examples of silicon retinas may be found in [45], [46]. Chips for detection of image motion, inspired by Reichardt motion-detector circuits in the fly [7], may be found in [48], [49].

amacrine cells, and ganglion cells are hooked to each other in a highly parallel 2.5D architecture to do parallel processing of the retinal image. In the cochlea, 3,500 inner hair cells and 10,000 outer hair cells work in parallel with each other, coupled to one another via the motions of the basilar membrane, the tectorial membrane, and the cochlear fluid.

23.10.5 Balance computation and communication costs

Estimates of energy consumption by neurons in the brain show that they spend about as much energy in communicating with action potentials as in computing with them [55].

23.10.6 Exploit collective analog or hybrid computation

We have already discussed this point in Section 23.9 and partly with reference to Figure 22.7 in Chapter 22.

23.10.7 Reduce the amount of information that needs to be processed

If the retina did not preprocess and compress the information that it sends to the brain, we would be processing about ~36 Gb/s of wide-dynamic-range raw image data from its photoreceptor array rather than ~20 Mb/s of useful optic-nerve spiking output information. If we assume that the energy consumption per bit does not change, the power consumption of the brain would then need to be ~22 kW, more than that of 20 furnaces! The auditory nerve is architected to fire spikes more often in response to sound envelope onsets than to steady-state sound envelopes [24]. Throughout the nervous system and in the brain, there are integrators or lowpass filters that form parts of an adaptive feedback loop such that neurons only fire in response to changes in their inputs rather than to constant inputs [10]. They are constantly learning to adapt away useless information that they can predict. This principle is so ubiquitous in the organization of the nervous system, that we refer readers to [10] for at least 100 such examples.

23.10.8 Use feedback and feedforward architectures for improving robustness and energy efficiency

A good example of feedforward processing in the nervous system is the *vestibular ocular reflex* (VOR) [10]. The motion of the head to the right triggers a compensatory reflexive motion of the eyes to the left such that the image that the eye is watching does not blur when the head is moved. The use of feedforward processing allows for quick fast responses but requires constant learning and calibration to ensure that the eye-motion to head-motion gain is just right. It avoids the need for complex feedback loops that would need to model the dynamics of the head and the eye in detail to prevent overshoots and instability while maintaining fast response times.

The use of integrators and lowpass filters in feedback loops serve to improve energy efficiency by reducing the amount of information that needs to be processed, while simultaneously removing dc errors and device mismatch that prevent robust operation. Feedback loops for gain control in the cochlea and photoreceptors ensure relatively constant SNR operation over a very wide dynamic range of input signals, which would otherwise compromise information fidelity and energy efficiency. Feedback and feedforward gain control is ubiquitous amongst neurons in the brain at multiple temporal and spatial scales [1].

Chapter 2 provides several examples of feedback loops in biology.

23.10.9 Separate speed and precision in the architecture if possible

The somatic regions of neurons shown in Figure 23.15 function as a simple thresholding comparator that generates spikes. Comparators, as we explain in Chapter 22, use preamplification with modest gain to obtain precision and positive feedback to obtain speed.

The retina in the eye has a central foveal region made up of cone cells that are relatively slow but that have precise spatial and contrast resolution. The peripheral region of the eye is made up of rod cells that are relatively fast but that have imprecise spatial and contrast resolution.

The right brain is, in general, better at getting the 'big picture' quickly but is relatively imprecise. It is important for recognizing novel threats and responding to predators quickly without getting bogged down in irrelevant detail that could lead to death. In contrast, the left brain is, in general, better at getting the 'details' slowly and is relatively precise [56]. Interestingly, some remarkable experiments [56] have shown that patients with left-brain damage can reproduce the overall big-picture of a visual scene from memory but are unable to reproduce its details. In contrast, patients with right-brain damage can reproduce minute details in a visual scene from memory but are unable to tie them together in a big picture.

23.10.10 Operate slowly and adiabatically if possible

Since the biological cochlea can be described as a high-Q coupled-resonator transmission-line structure, it is inherently adiabatic as per the discussion in Chapter 21. That is one of the key reasons for its incredibly low power dissipation of $\sim 14\,\mu$W.

Dendrites in neurons in the brain are architected with distributed RC transmission lines, which because of their distributed nature have an effectively smaller time constant for their inputs. For example, a well-known result in RF electronics is that the effective RC time constant associated with a distributed polysilicon gate input to a transistor with gate length L is $(RL)(CL)/3$ rather than $(RL)(CL)$, where R and C are the resistance and capacitance per unit length of the polysilicon wire respectively. The 1/3 factor arises because the entire R and the entire C do not filter the incoming gate input at any point in the distributed RC line such that different points in the channel see different amounts of incremental filtering. The effective speedup in time constant due to a distributed structure rather than a lumped structure implies that, for a given input frequency of operation, the operation is more adiabatic: the cutoff frequencies of the structure are now larger relative to the input frequency. The brain's average rate of spike production is near 5 Hz and the time constants of its neuronal structures vary from 100 ms to 0.1 ms. Thus, the brain operates in a relatively slow fashion in some of its structures.

23.11 Other work

Neuromorphic electronics is still in its infancy because understanding nature with insight has taken time and will continue to take time. Nevertheless, we can pick increasingly high-hanging fruit as engineering helps understand biology better and biology in turn inspires better engineering, leading to a powerful positive-feedback loop. Besides some of the RF, auditory, speech, vision, and hybrid analog-digital examples presented in this chapter, we mention a few examples for the reader interested in exploring further. Spike-based communication circuits for interfacing chips termed *address event representation* have been proposed to connect chips, e.g., [57], [58]. Biosonar systems inspired by the bat are being actively researched and applied to artificial sonar systems [59]. Bio-inspired robotic systems have been important in creating ingeniously simple-but-clever robots [60], [50]. Learning and pattern-recognition circuits have been built [61], [62], [63]. A learning architecture inspired by neural encoding-decoding architectures has been proposed for decoding of neural signals in paralysis prosthetics [51], [64]. A table of some of the labs working in the general area of neuromorphic electronics may be found in [13]. Finally, the bio-inspired revolution is expanding beyond neurons and beyond electrical engineering to other biomorphic domains: as just one example in a large space, extremely large single-crystal structures inspired by biological material formation have led to very innovative materials-science designs [65].

In the next chapter, we shall study computation and circuits within cells, which can inspire cell-inspired or cytomorphic architectures in the future.

23.12 Appendix: Power and computation in the brain, eye, ear, and body

The brain's neuronal cells output $\sim$1 ms pulses (spikes) at an average rate of 5 Hz [55]. The 240 trillion synaptic connections [1] amongst the brain's neurons thus lead to a computational rate of *at least* 10^{15} synaptic operations per second. A synapse implements multiplication and filtering operations on every spike and sophisticated learning operations over multiple spikes. If we assume that synaptic multiplication is at least one floating-point operation (FLOP), the 20 ms second-order filter impulse response due to each synapse is 40 FLOPS, and that synaptic learning requires at least 10 FLOPS per spike, a synapse implements at least 50 FLOPS of computation per spike. The nonlinear adaptation-and-thresholding computations in the somatic regions of a neuron implement almost 1200 floating-point operations (FLOPS) per spike [66]. Thus, the brain is performing

24 Cytomorphic electronics: cell-inspired electronics for systems and synthetic biology

Any living cell carries with it the experience of a billion years of experimentation by its ancestors.

Max Delbrück

The cells in the human body provide examples of phenomenally energy-efficient sensing, actuation, and processing. An average $\sim 10\ \mu$m-size human cell hydrolyzes several energy-carrying adenosine-tri-phosphate (ATP) molecules within it to perform nearly $\sim 10^7$ ATP-dependent biochemical operations per second [1]. Since, under the conditions in the body, the hydrolysis of each ATP molecule provides about 20 kT (8×10^{-20} J) of metabolic energy, the net power consumption of a single human cell is an astoundingly low 0.8 pW! The $\sim$100 trillion cells of the human body thus have an average resting power consumption of $\sim$80 W, consistent with numbers derived from the Appendix of Chapter 23.

The cell processes its mechanical and chemical input signals with highly noisy and imprecise parts. Nevertheless, it performs complex, highly sensitive, and collectively precise hybrid analog-digital signal processing on its inputs such that reliable outputs are produced. Such signal processing enables the cell to sense and amplify minute changes in the concentrations of specific molecules amidst a background of confoundingly similar molecules, to harvest and metabolize energy contained in molecules in its environment, to detoxify and/or transport poisonous molecules out of it, to sense if it has been infected by a virus, to communicate with other cells in its neighborhood, to move, to maintain its structure, to regulate its growth in response to signals in its surround, to speed up chemical reactions via sophisticated enzymes, and to replicate itself when it is appropriate to do so. The $\sim$30,000-node gene-protein and protein-protein molecular interaction networks within a cell that implement and regulate these functions are a true marvel of nanotechnology. The nanotechnology of man appears crude and primitive when contrasted with that in nature's cells.

In this chapter, we show that the equations that describe subthreshold transistor operation and the equations that describe chemical reactions have strikingly detailed similarity, including stochastic properties. Therefore, any chemical reaction can be efficiently and programmably represented with a handful of subthreshold (or bipolar) transistors that comprise an analog circuit. Intracellular protein-protein biochemical reaction networks can then potentially be modeled by hooking such circuits to each other. DNA-protein networks can also be efficiently modeled with such analog circuits. DNA-protein networks can be modeled even more efficiently with hybrid analog-digital circuits that approximate nonlinear analog characteristics with more approximate digital ones. Since extracellular cell-cell networks also rely on molecular binding and chemical reactions, networks such as hormonal networks or neuronal networks can also be efficiently modeled. Thus, in the future, we can potentially attempt to simulate cells, organs, and tissues with ultra-fast highly parallel analog and hybrid analog-digital circuits including molecular stochastics and cell-to-cell variability on large-scale electronic chips. Such molecular-dynamics simulations are extremely computationally intensive especially when the effects of noise, nonlinearity, network-feedback effects, and cell-to-cell variability are included. Stochastics and cell-to-cell variabililty are highly important factors for predicting a cell's response to drug treatment, e.g., the response of tumor cells to chemotherapy treatments [2]. We will show in this chapter that circuit, feedback, and noise-analysis techniques described in the rest of this book can shed insight into the systems biology of the cell [3]. For example, flux balance analysis is frequently used to reduce the search space of parameters in a cell [4]. It is automatically implemented as Kirchhoff's current law in circuits since molecular fluxes map to circuit currents. Similarly, Kirchhoff's voltage law automatically implements the laws of thermodynamic energy balance in chemical-reaction loops. We shall provide other examples throughout the chapter. An excellent introduction to systems biology may be found in [5]. Robustness analysis of the circuit using return-ratio techniques can shed insight in the future into which genes, when mutated, will lead to disease in a network, and which will not. Circuit-design techniques can also be mapped to create synthetic-biology circuits that perform useful functions in the future [6].

Circuits in biology and circuits in electronics may be viewed as being highly similar with biology using molecules, ions, proteins, and DNA rather than electrons and transistors. Just as neural circuits have led to biologically inspired neuromorphic electronics, cellular circuits can lead to a novel biologically inspired field that we introduce in this chapter and term *cytomorphic electronics*. In fact, we will show that there are many similarities between spiking-neuron computation and cellular computation. The hybrid state machines (HSMs) described in Chapters 22 and 23 (see Figure 22.9 (b)) provide a useful framework for thinking about cellular circuits with active DNA genes in the cell being represented by digital variables and protein concentrations represented by analog variables. DNA-protein interactions are represented by interactions between the analog and digital parts of an HSM. Networks of hybrid state machines can model networks of cells. The ZY transmission-line architectures that we discussed in Chapter 23 for describing computation in the retina, the cochlea, the vocal tract, and neuronal dendrites are also useful for understanding nonlinear reaction-diffusion partial differential equations in cells. Ingenious nonlinear transmission lines are exploited by the cell to create spatially decaying molecular-concentration profiles that are robustly invariant to changes in protein-production rates during development.

Leonardo da Vinci, perhaps the first bio-inspired engineer in mankind, said: ‘Human subtlety will never devise an invention more beautiful, more simple or more direct than does nature because in her inventions nothing is lacking, and nothing is superfluous.’ One is indeed awed and humbled as one learns more and more about the ingenious operation of even a single cell. We have much to learn from nature in architecting clever electronics, algorithms, and nanostructures. This chapter will only attempt to scratch the surface.

24.1 Electronic analogies of chemical reactions

Figure 24.1 illustrates that there are striking similarities between chemical reaction dynamics (Figure 24.1 (a)) and electronic current flow in the subthreshold regime of transistor operation (Figure 24.1 (b)): electron concentration at the source is analogous to reactant concentration; electron concentration at the drain is analogous to product concentration; forward and reverse current flows in the transistor are analogous to forward and reverse reaction rates in a chemical reaction; the forward and reverse currents in a transistor are exponential in voltage differences at its terminals analogous to reaction rates being exponential in the free energy differences in a chemical reaction; increases in gate voltage lower energy barriers in a transistor increasing current flow analogous to the effects of enzymes or catalysts in chemical reactions that increase reaction rates; and the stochastics of Poisson shot noise in subthreshold transistors are analogous to the stochastics of molecular shot noise in reactions. These analogies suggest that one can mimic and model large-scale chemical-processing systems in biological and artificial networks very efficiently on an electronic chip at time scales that could be a million to billion times faster. No one, thus far, appears to have exploited the detailed similarity behind the equations of chemistry and the equations of electronics to build such networks. The single-transistor analogy of Figure 24.1 is already an exact representation of the chemical reaction $A \rightleftarrows B$ including

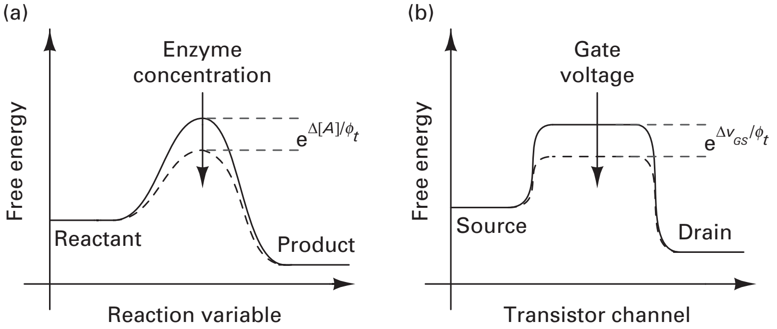


Figure 24.1a, b. Similarities between chemical reaction dynamics (a) and subthreshold transistor electronic flow (b). Reprinted with permission from [9] (©2009 IEEE).

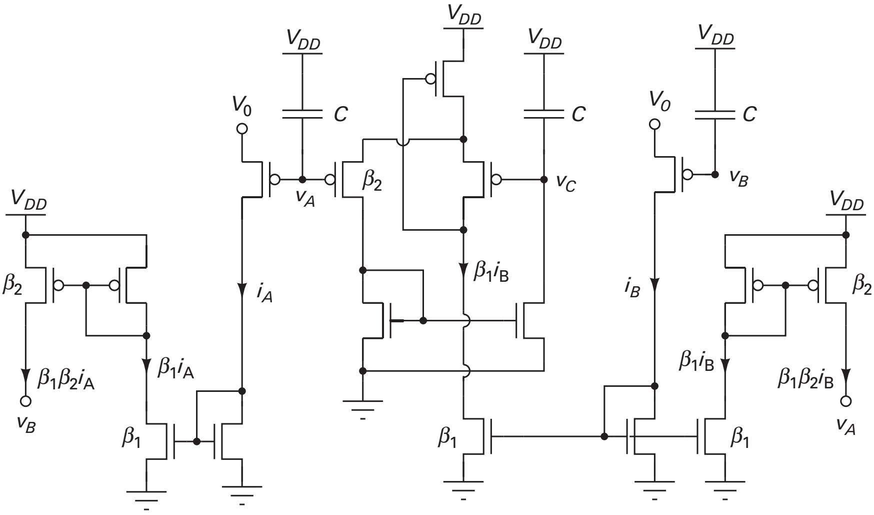


Figure 24.4. Second-order log-domain chemical reaction circuit. Reprinted with permission from [9] (©2009 IEEE).

forward-reaction current source (that exhibits Poisson noise statistics) hooked to a conductance proportional to K_d, which is in parallel with a capacitance (and the current source). The voltage on the capacitance represents the concentration of C. Due to the reverse-reaction current flowing through it, the conductance proportional to K_d exhibits Poisson statistics as well. In effect, we have a parallel RC circuit fed by a current source with shot noise from the current source and shot noise from the current flowing through the R. Note that the power spectral density of the current through R is NOT $4kTG$ but $2qI$ in such chemical resistances as we discussed when we derived Equation (24.7). If the forward-reaction current source has a value I, the voltage v_C will equilibrate at a value such that the backward-reaction current through R balances it. Thus, we may expect the noise voltage on the capacitor C to be given by

$$\begin{aligned}\overline{v_n^2} &= (2qI + 2qI)\left(R^2\right)\left(\frac{\pi}{2}\right)\left(\frac{1}{(2\pi)RC}\right)\\ &= \frac{q(IR)}{C}\end{aligned} \tag{24.18}$$

The signal-voltage power on the capacitor is given by

$$v_C^2 = (IR)^2 \tag{24.19}$$

Thus, the net SNR of the circuit in this scenario is given by

$$\begin{aligned}SNR &= \frac{v_C^2}{\overline{v_n^2}}\\ &= \frac{C(IR)}{q}\end{aligned} \tag{24.20}$$

maximize objective functions known to be of biological importance like cell growth. In essence, analog chips that model chemical reaction networks can serve as 'special-purpose' ALUs that implement nonlinear dynamical systems optimized for simulating biochemical reaction networks.

24.3 Analog circuit models of gene-protein dynamics

Figure 24.7 illustrates the basics of gene-protein interactions in cells. An *inducer* molecule, e.g., glucose or S_X, may enter a cell and cause biochemical reaction events that eventually, e.g., via a protein-protein network, lead to the activation of a particular protein called a *transcription factor* X. When activated, $X \rightarrow X^*$, with X^* being the active form of the transcription factor. The activation of the transcription factor most often occurs because a molecule, typically the inducer, binds to the transcription factor and changes its shape. The activated transcription factor X^* can then bind to DNA near or within specific *promoter* binding sites on the DNA that have a particular sequence of A, T, C, or G nucleotides within them. These sites are effective in binding the particular transcription factor. *Transcription* is the process whereby the enzyme RNA polymerase (RNAp in Figure 24.7) converts the DNA sequence of a gene into a corresponding messenger RNA (mRNA) transcript sequence that is eventually translated into a protein. The binding of the transcription factor causes the transcription rate of a gene near the promoter to be increased if the transcription factor is an *activator* or decreased if the transcription factor is a *repressor*. In Figure 24.7, a repressor transcription factor Y that is activated to Y^* by an inducer S_Y is shown. Ribosome molecules in

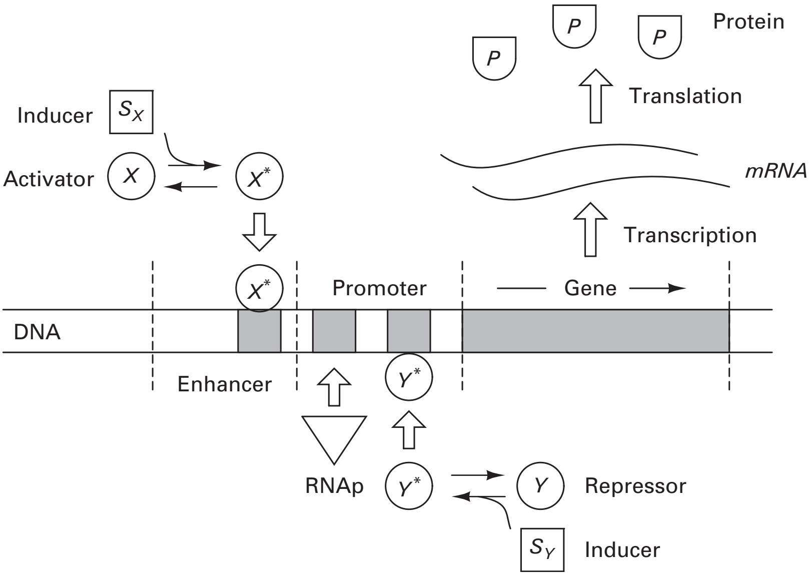


Figure 24.7. Basics of induction, transcription, and translation in cells.

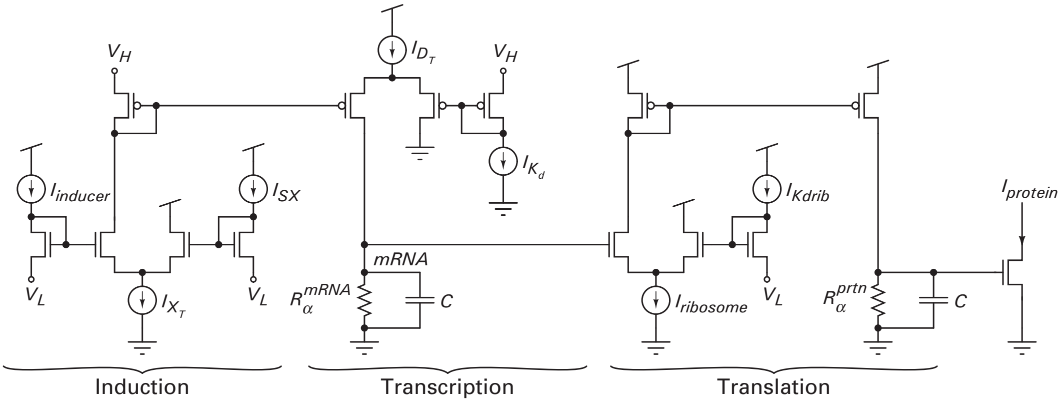


Figure 24.8. Basic induction, transcription, and translation circuit.

the cell *translate* the mRNA transcript into a sequence of corresponding amino acids to form a protein. The final translated protein can act as a transcription factor for other genes in a gene-protein network or affect other proteins in a protein-protein network or both. The translated protein can also serve as an activator or repressor for its own gene. Readers interested in further details of molecular biology should consult [1].

Figure 24.8 reveals a circuit model of induction, transcription, and translation. If the inducer concentration, $I_{inducer}$, is significantly greater than I_{SX}, the K_d for inducer-transcription-factor binding, most of the transcription-factor molecules X will be transformed into an active state X^*. The current output of the subthreshold differential pair that represents induction quantitatively models this process:

$$I_{X^*} = I_{X_T}\left(\frac{\frac{I_{inducer}}{I_{SX}}}{\frac{I_{inducer}}{I_{SX}}+1}\right) \tag{24.21}$$

Equation (24.21) is an exact model of Michaelis-Menten binding as per Equation (24.3) with I_{X_T} representing the total concentration of transcription factor, X_T, whether activated or not. Similarly, if the activated transcription-factor concentration I_{X^*} is significantly greater than K_d, the dissociation constant for DNA-transcription-factor binding, the rate of production of mRNA transcripts by the enzyme RNA polymerase will be near its maximal value, assuming, without loss of generality, that the transcription factor in question is an activator. The current output, I_m, of the subthreshold differential pair that represents transcription in Figure 24.8 quantitatively models this process:

$$I_m = I_{D_T}\left(\frac{\frac{I_{X*}}{I_{Kd}}}{\frac{I_{X*}}{I_{Kd}}+1}\right) \tag{24.22}$$

24.5 Stochastics in DNA-protein circuits

Poisson noise in mRNA-production flux can be mimicked by Poisson electronic current noise in a manner analogous to that discussed previously for protein-production flux. However, the noise levels in mRNA production for some genes can be high enough such that extremely low currents and extremely small capacitances become necessary in electronics to mimic the same low signal-to-noise ratio (SNR) in biology. The resultant noise in electronics is then not well controlled or predictable. Therefore, it is sometimes advantageous to artificially introduce a controlled level of noise in a relatively quiet electronic circuit to mimic high-noise signals in biology. Figure 24.12 illustrates a circuit for doing so.

In Figure 24.12, the current-mode integrator with output capacitor C and $I_{R\alpha}$ implement a current-mode version of the $R^{mRNA}C$ lowpass filter in Figure 24.11. That is, $2I_{R\alpha}$ and C correspond to I_A and C in Figure 14.9 in Chapter 14 on current-mode circuits. Similarly, v_{mRNA} and i_{mRNA} correspond to v_{OUT} and i_{OUT} in Figure 14.9 respectively. Instead of a constant $2I_{R\alpha}$ in a traditional current-mode lowpass filter, however, the leak current $2I_{R\alpha}$ is pseudo-randomly switched on and off with a duty cycle of 0.5. Thus, the average value of the leak current that sets the lowpass filter time constant is $I_{R\alpha}$ as in a traditional circuit but the random switching introduces a stochasticity in this leak current. The log voltage on the current-mode capacitor is exponentiated and converted to a current i_{mRNA} that encodes the level of mRNA as in any current-mode circuit. The current i_{mRNA} is gained up by β_{SNR} and used to control the frequency, f_{CCO}, of a current-controlled oscillator (CCO). The output switching frequency of the oscillator is proportional to i_{mRNA} according to $f_{CCO} = \beta_{SNR} i_{mRNA}/q_{CCO}$, where q_{CCO} depends on design parameters internal to the CCO. Thus, as mRNA levels rise, the control current and switching frequency of the CCO rise in proportion. The linear feedback shift register (LFSR) converts the digital output of the CCO to a random switching signal via a classic pseudo-random-number generator technique [13]. Thus, the output of the LFSR randomly switches the $I_{R\alpha}$ current on and off with a switching frequency f_{CCO} that is proportional to the mRNA level encoded by i_{mRNA}. Consequently, as mRNA levels rise, a consequence of a higher mRNA production rate, the arrival

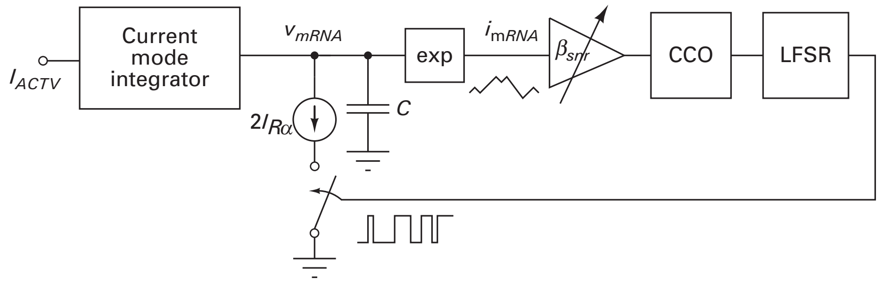


Figure 24.12. Artificial noise generation circuit for low SNR. Reprinted with permission from [14] (©2009 IEEE).

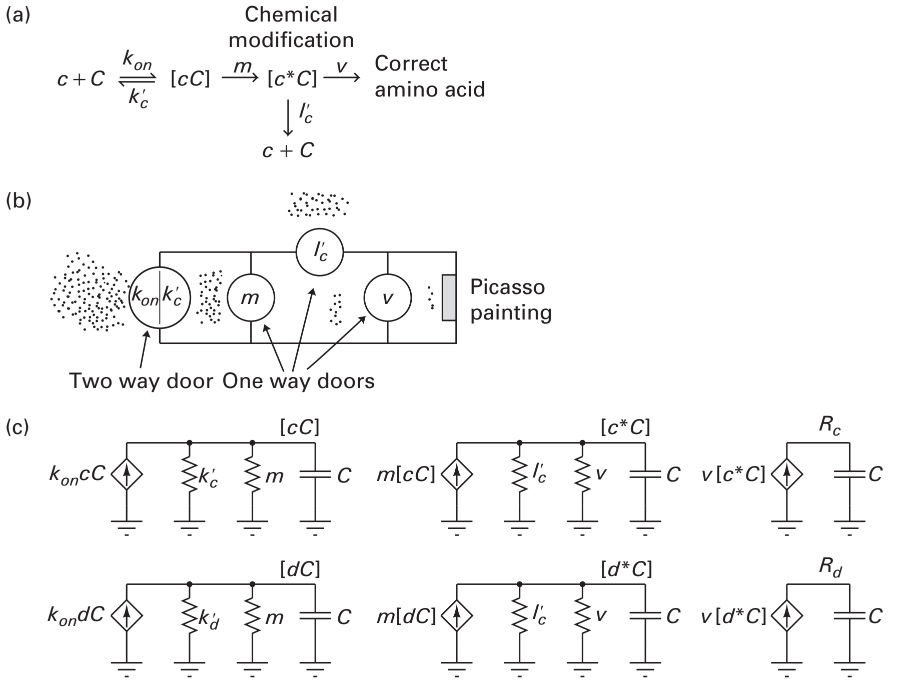


Figure 24.16a, b, c. (a) Kinetic proofreading. (b) Picasso-lover analogy. (c) Circuit analogy.

into meaningful network motifs such as the FFL sub-block that we have discussed [5]. In contrast, representing complex systems via a set of random connections amongst network nodes does not yield much insight into the function of the complex system although it may be mathematically equivalent to a circuit.

As an example of how circuits can shed insight into systems biology, we illustrate how the language of circuits can provide a useful interpretation of kinetic proofreading circuits in cells, which are ubiquitously present in many of its subsystems. Kinetic proofreading allows the cell to reduce the discrimination error rate between highly similar molecules without spending too much time or designing highly specific recognition molecules to do so [16]. Figure 24.16 (a) illustrates the biochemical process; Figure 24.16 (b) provides a Picasso-lover analogy of the process that is adapted from a textual description of this analogy in [5]; Figure 24.16 (c) provides a circuit model of the process.

Figure 24.16 (a) illustrates the chemical-reaction cascade involved in the binding of a specific tRNA codon molecule, which contains a three-letter snippet of RNA, to a complementary three-letter snippet in an mRNA transcript to create a bound species. The tRNA codon is denoted by *c*, the corresponding mRNA snippet is denoted by *C*, and the bound species is denoted by *cC*. Such binding triggers events within the ribosome protein translating machinery in the cell. These events translate a three-letter codon word formed from an alphabet of 4 molecular letters to a particular amino acid. Each of the 64 possible codons is translated by the paradigm with g_m changed from a finite value to 0. Such analyses can be considerably more subtle and robust to model error than digital-circuit techniques that have attempted to do the same by treating all variations as 'on' or 'off' and modeled all DNA circuits as being purely digital.

24.8 Hybrid analog-digital computation in cells and neurons

Computation in a neuronal network and computation in a gene-protein network share intriguing similarities at the signal, circuit, and system level. Table 24.1 below highlights some of these similarities.

These analogies suggest that neuronal networks in the brain have similarities to ultra-fast, highly plastic, large-scale gene-protein networks with lots of connections per node. The speed of neuronal responses is matched to the time scales

Table 24.1 Similarities between neuronal and cellular computation

Property	Neuronal equivalent	Cellular equivalent
1. Basic computational unit device	A neuron	A gene
2. Discrete symbolic digital output of device	A spike	An mRNA transcript
3. Translation of symbol to signal	Post-synaptic potential (PSP)	Translation of mRNA transcript to protein
4. Processing of signals to create symbols	Analog and digital dendritic processing of multiple inputs to neuron	Analog logic circuits on DNA from multiple inputs
5. Connection weighting	Synaptic weight	K_d and transcription-factor binding
6. Kinds of connections	Excitatory and inhibitory	Activatory and repressory
7. Dale's rule	Output connection signs of a single neuron are correlated	Output connection signs of a single gene are correlated
8. Topological similarities	Shunting inhibition near soma with many excitatory inputs at other dendritic locations	Common repressor on DNA promoter with many activator transcription factor inputs.
9. Adaptation	Learning	Evolution
10. Connections per node	~6000	~12
11. Number of nodes	~22 billion	~30,000
12. Distributed ZY transmission-line processing	In dendrites	Reaction-diffusion networks within and outside cells.
13. Power consumption	~0.66 nW per neuron 22 billion neurons in the brain	~1 pW per cell ~100 trillion cells in the body

Section VII

Energy sources

25 Batteries and electrochemistry

I have been so electrically occupied of late that I feel as if hungry for a little chemistry: but then the conviction crosses my mind that these things hang together under one law.

Michael Faraday

Ultra-low-power electronic systems often utilize batteries. The battery is frequently the heaviest and most critical portion of a portable system. The battery's size, weight, energy density, power density, form factor, cycling properties and lifetime can dictate constraints under which a portable electronic system must operate. Such constraints include the overall power dissipation of the system, the number of recharges that are possible in the system, the expected lifetime of the system, the overall size of the system, and the overall cost of the system. In this chapter, we shall provide a brief introduction to batteries with an emphasis on intuition and knowledge that is useful for providing insight into battery operation and ultra-low-power system design.

First, we shall begin with an exposition of the basic principles by which chemical energy is converted to electrical energy in a battery. Then, we shall discuss an equivalent circuit for a battery that intuitively describes several effects that are observed in practical batteries. We shall discuss the basics of lithium-ion and zinc-air batteries, two of the most common and most efficient batteries in existence in ultra-low-power electronic systems, particularly biomedical ones. We shall also briefly discuss the benefits of fuel cells and ultra capacitors, devices that are capable of being used in conjunction with or instead of batteries, and that are likely to be important in the future. Throughout our discussion, we shall focus on fundamental tradeoffs and practical characteristics that are seen in all batteries, irrespective of differences in their detailed chemistry.

25.1 Basic operation of a battery

A battery is a device that converts chemical energy to electrical energy by transforming the energy of a chemical reaction to electrical energy in a controlled fashion. The energy conversion and chemical reaction ideally only occur when the battery is used, i.e., when it discharges. The conversion is performed directly without the burning of chemical fuel in a battery. Thus, batteries have a

25.5 Large-signal equivalent circuit of a battery

Figure 25.3 shows a large-signal equivalent circuit of a battery that represents Equations (25.12) and (25.13) in circuit form with exponential elements (κ or $(1-\kappa)$ elements) shown as diodes. The pre-exponential I_S or equilibrium current of the diode is marked on the Figure 25.3 for each diode in terms of the respective i_0 and the mass-transport I_L parameters. The open-circuit voltages for the cathode and anode half cells are represented with voltage sources with the total open-circuit battery voltage being given by the difference between the cathodic voltage source (positive) and the anodic voltage source (negative). The dark-face diodes correspond to the exponential terms that are dominant during battery discharge while the light-face diodes correspond to the exponential terms that are dominant during battery charge. The conventional current I_{batt} flows through an external circuit impedance that the battery performs electrical work on from the cathode to the anode while electron current flows from the anode to the cathode. The circuit also models the electrolyte ohmic resistance due to its finite ionic conductivity via the R_{elec} resistances. The electrolyte resistance can be improved by increasing concentrations of its ionic species and by geometric design: increasing cross-sectional area between the anode and cathode and decreasing the distance between anode and cathode. The ionic conductivity of a battery usually increases with temperature such that R_{elec} falls at high temperatures. The significance of Q_A and Q_C in Figure 25.3 with respect to the state of charge (SOC) and usage of the battery are explained later. For now, we assume that all the parameters in Figure 25.3 such as i_{0A} and I_L remain constant in the battery independent of how long the battery has been discharged. Unless otherwise noted, we shall focus primarily on the discharging properties of the battery when it is used in a circuit.

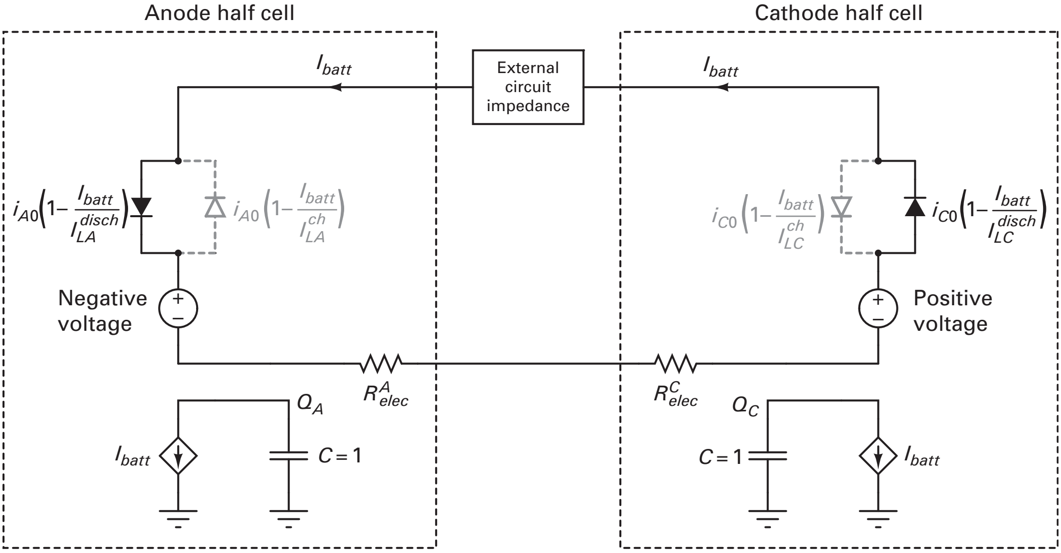

Anode half cell
Cathode half cell
I_{batt}
External circuit impedance
I_{batt}
$i_{A0}\left(1-\frac{I_{batt}}{I_{LA}^{disch}}\right)$
$i_{A0}\left(1-\frac{I_{batt}}{I_{LA}^{ch}}\right)$
$i_{C0}\left(1-\frac{I_{batt}}{I_{LC}^{ch}}\right)$
$i_{C0}\left(1-\frac{I_{batt}}{I_{LC}^{disch}}\right)$
Negative voltage
Positive voltage
R_{elec}^{A}
R_{elec}^{C}
Q_A
Q_C
I_{batt}
$C=1$
$C=1$
I_{batt}

Figure 25.3. Large-signal dc equivalent circuit of a battery.

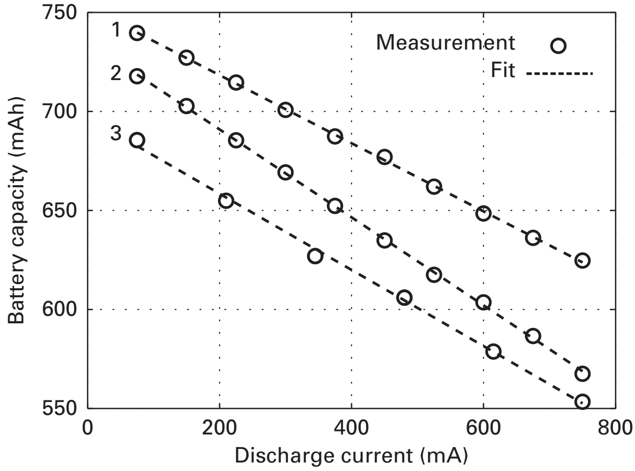


Figure 25.6. Loss in battery capacity with increasing discharge current.

Figure 25.6 plots Q_{ah} versus I_{batt} for three different lithium-ion batteries. We see that Equation (25.19) is a good fit to all three of them. For battery 1, we found $Q_{ah}^{max} = 753\,\text{mA h}$ and $\tau = 0.17\,\text{h}$; for battery 2, we found $Q_{ah}^{max} = 735\,\text{mA h}$ and $\tau = 0.22\,\text{h}$; for battery 3, we found $Q_{ah}^{max} = 697\,\text{mA h}$ and $\tau = 0.19\,\text{h}$. If Q_{ah} were not a function of I_{batt}, the curves in Figure 25.6 would all be flat horizontal lines with $\tau = 0$. Instead, they are lines with a finite slope of $-\tau$ as predicted by Equation (25.19).

From the discussions leading to Equations (25.17), (25.18), and (25.19), we can define an effective normalized SOC variable for the battery $Q_n(t)$ that is given by

$$\begin{aligned} Q_n(t) &= \left(1 - \frac{\int_0^t I_{batt}(\eta)d\eta}{Q_{ah}^{max}} \right) \\ Q_n(t) &= \frac{\int_0^t I_{batt}(\eta)d\eta}{Q_{ah}^{max}} \\ Q_n(t) &= 1 - Q_u(t) \end{aligned} \tag{25.20}$$

From Equation (25.17), as the battery is discharged, $Q_n(t)$ falls such that internal variables within it like I_L change according to

$$I_L(t) = I_L^{max} Q_n(t) \tag{25.21}$$

Note that $I_L(t)$ and I_L^{max} are based on combining parameters at the anode and the cathode, which may actually be different, into one effective parameter in Equation (25.21). Similarly, as the battery discharges, internal concentrations of reactants such as, for example, the number of lithium ions in the anode falls, or concentrations of reactants and products get more equalized. Thus as the battery reaction

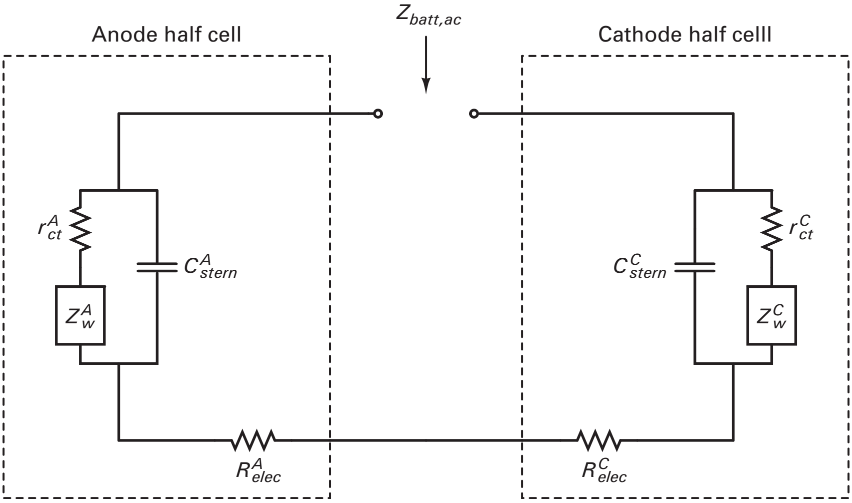


Figure 25.8. Small-signal ac equivalent circuit of a battery.

anodic and cathodic half circuits to create a complete small-signal equivalent for the battery, Z^{ac}_{batt} [2]. That is

$$\begin{aligned} Z^{ac}_{batt}(\omega) &= \left(r^C_{ct} + Z^C_W(\omega)\right) \| \left(\frac{1}{j\omega C^C_{stern}}\right) \\ &+ \left(r^A_{ct} + Z^A_W(\omega)\right) \| \left(\frac{1}{j\omega C^A_{stern}}\right) + \left(R^C_{elec} + R^A_{elec}\right) \end{aligned} \tag{25.34}$$

The anodic side has impedance formulas analogous to ones that we have explicitly discussed for the cathodic side.

The electrolyte resistance is determined by evaluating the resistance of all three-dimensional electrolytic conduction paths from anode to cathode. In the case of two nearby rectangular electrodes of equal surface area $A = A_C = A_A$, separated by a length L that is much smaller than any dimension of the electrode, to a good approximation,

$$\begin{aligned} R_{elec} &= R^C_{elec} + R^A_{elec} \\ &= \frac{\rho_{elec}(L/2)}{A} + \frac{\rho_{elec}(L/2)}{A} \end{aligned} \tag{25.35}$$

In Equation (25.35), ρ_{elec} is the specific resistivity of the electrolyte, which is a function of the ionic concentrations and ion mobilities. The specific resistivity of an electrolyte varies from 2–10 Ω cm for aqueous electrolytes, 10–50 Ω cm for inorganic electrolytes, 100–1000 Ω cm for organic electrolytes, 1000 to 10^7 Ω cm for polymer electrolytes, and 10^5–10^8 Ω cm for inorganic solid electrolytes [1].

For electrodes that are widely separated, or in more complex three-dimensional geometries, the exact evaluation of R_{elec} may involve a three-dimensional integral. As a simple example, consider a spherical anodic electrode of radius R sourcing

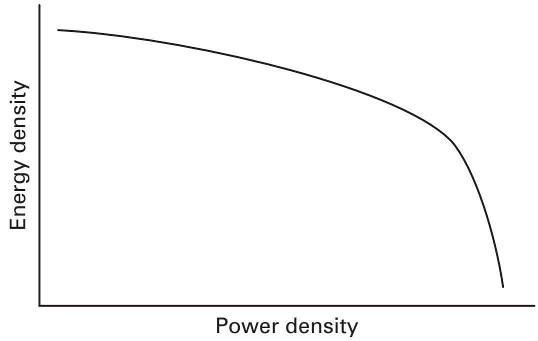


Figure 25.9. Example Ragone curve.

Often, the terms energy density and power density are used to describe volumetric numbers while the adjective 'specific' refers to gravimetric numbers. For simplicity, we shall use energy density and power density to describe both volumetric and gravimetric numbers.

When the power drawn from a battery is near its absolute maximum, $I_{batt}\tau \approx Q_{ah}^{max}$, such that Q_{ah} in Equation (25.19) is 0, and consequently the energy capacity or energy density of the battery is near 0; the open-circuit voltage of the battery is also severely degraded under such extreme conditions. When I_{batt} is almost 0, Q_{ah} in Equation (25.19) is near Q_{ah}^{max}, the battery's open-circuit value is near its maximum with almost no activation, ohmic, or concentration polarization losses, and consequently the energy capacity or energy density of the battery is at its maximum. These extremes of operation determine the limits of maximum energy density and maximum power density on the Ragone curve. Increasing the surface area/volume of the electrodes by having thin electrodes increases the maximum value of I_L, thus reducing τ and improving the power density but also compromises Q_{ah}^{max}, thus reducing the energy density.

Primary non-rechargeable batteries are built to have good energy densities such that they last long, and therefore tend to compromise power density. Secondary rechargeable batteries do not need to run as long as primary batteries since they can be recharged. Therefore, they usually have higher power densities and lower energy densities.

An example that illustrates the degradation in energy density with small size due to surface-area-versus-volume considerations is provided by the common cylindrical alkaline 1.5 V cells that dominate the primary battery market: Typical average ampere-hour capacities for AAAA, AAA, and AA batteries are 0.56 A h, 1.15 A h, and 2.67 A h, respectively as cylindrical radii and cell diameters increase from the smallest-size AAAA to the largest-size AA. These data were averaged from various manufacturers and are taken from [1]. The corresponding energy densities adapted from [1] are then given by 465 J/g, 564 J/g, and 627 J/g, respectively and improve as a larger fraction of the volume is devoted to energy storage in the battery.

Fuel cells can optimize power-generation by scaling up to high fuel-supply rates and large-area electrodes if high power densities are required and independently

3. Batteries that do not need to be discharged as deeply in an ultra-low-power electronic system fade more slowly and can be recharged several more times.
4. Low-power operation minimizes all polarization losses in the battery (activation, ohmic, and concentration) and thus maximizes the battery's voltage. The lack of losses improves the efficiency of energy extraction and thus prolongs the time between recharges.

An ultra-low-power electronic system reduces battery weight and size, consequently system weight and size, and consequently system cost. The materials cost of packaging is reduced because less material is needed in a smaller system, batteries with lower performance specifications are significantly cheaper to make, costs due to heating concerns are reduced, increased system lifetimes amortize costs over a longer time, a small lightweight product significantly increases the value of the product to the consumer, and enables scaling to larger more-complex systems to become possible. A small, portable low-power system can make solutions in tightly constrained spaces of the body with low tissue heating possible in implantable medical applications that could not have been possible otherwise. Indeed, implantable batteries in cardiac pacemakers have already revolutionized patients' lives. Ultra-low-power electronics will undoubtedly enable several more improvements in medical implants for patient treatment as we have discussed in Chapter 19.

We have discussed the harvesting of energy from RF sources in Chapters 16 and 17 for implantable (see Chapter 19) and for noninvasive medical applications (see Chapter 20), respectively. In the succeeding Chapter 26, we shall discuss other sources of energy that can be harvested such as solar energy, vibratory energy from the body, thermal energy from body heat, and chemical energy from organic or bio-molecules. We shall see that insights from ultra-low-power system design, which we have focused on mostly for small-scale medical applications, and mostly in electronics, are also important at larger scales and in non-electronic applications like transportation. To meet the 15 TW power budget of the world with renewable energy sources alone, a necessity after oil is highly scarce in 3–4 decades, is challenging. Just as low-power electronic design and battery design can undergo joint optimization to create a win-win scenario for both, low-power transportation design and renewable energy generation can undergo joint optimization to create a win-win solution for both.

References

[1] David Linden and Thomas B. Reddy. *Handbook of Batteries*, 3rd ed. (New York: McGraw-Hill, 2002).

[2] Allen J. Bard and Larry R. Faulkner. *Electrochemical Methods: Fundamentals and Applications*, 2nd ed. (New York: Wiley, 2001).

[3] S. Dearborn. Charging Li-ion batteries for maximum run times. *Power Electronics Technology Magazine*, (2005), 40–49.

26 Energy harvesting and the future of energy

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

Richard P. Feynman

Energy surrounds us, is within us, and is created by us. In this chapter, we shall discuss how systems can harvest energy in their environments and thus function without needing to constantly carry their own energy source. The potential benefits of an energy-harvesting strategy are that the lifetime of the low-power system is then not limited by the finite lifetime of its energy source, and that the weight and volume of the system can be reduced if the size of the energy-harvester is itself small. The challenges of an energy-harvesting strategy are that many energy sources are intermittent, can be hard to efficiently harvest, and provide relatively low power per unit area. Thus, energy-harvesting systems are usually practical only if the system that they power operates with relatively low power consumption.

We shall begin by discussing energy-harvesting strategies that have been explored for low-power biomedical and portable applications. First, we discuss the use of strategies that function by converting mechanical body motions into electricity. A circuit model developed for describing energy transfer in inductive links in Chapter 16 is extremely similar to a circuit model that accurately characterizes how such mechanical energy harvesters function. Thus, tradeoffs on maximizing energy efficiency or energy transfer are also similar. Energy harvesting with RF energy is discussed extensively in Chapters 16 and 17, so we shall not discuss it in this chapter. Then, we discuss the use of thermoelectric strategies that function by converting body heat into electricity. A fundamental thermodynamic principle limits the energy efficiency of a 'heat engine', whether in an internal combustion engine in a car, in a refrigerator, or in a thermoelectric device powered by body heat. The limiting efficiency is called the *Carnot efficiency*. The Carnot efficiency and models of heat flow from the body will help us understand the limits of operation of thermoelectric energy harvesting.

This book has largely discussed ultra-low-power systems at relatively small spatial scales in biomedical and in bio-inspired systems, mostly in the 10^{-12} W to 10^{-2} W range. In this final chapter, we shall see that principles of low-power design are also relevant to systems at large spatial scales with gigantic power consumption, e.g., a 40 kW gasoline-powered car moving at 30 mph, which if operated for 1 hour each day leads to an average power consumption of 1.67 kW.

The average human being on Earth consumes 2.5 kW of power such that our planet's current aggregate power consumption is roughly 15 TW. The average power consumption of people in richer countries is higher than that in poorer countries. For example, the average person in the United States consumes 10.4 kW [1].[1] We have been able to sustain such power consumption thus far largely because the 46,400 J/g energy density of gasoline, the 53,600 J/g energy density of natural gas, and the 32,500 J/g energy density of coal, and their relative abundance, have enabled us to burn energy at a profligate rate. In comparison, a well-optimized lithium-ion battery for portable applications operates at 650 J/g. Gasoline is currently cheaper per liter than bottled water in the United States.

For every kW h of oil, natural gas, or coal that is consumed, 250 g, 190 g, and 300 g, respectively, of CO_2 is dumped into our atmosphere [2]. This means that 5.5 tons of CO_2 is generated on average per person per year, increasing CO_2 levels by ~2.5 ppm (parts per million) per year today [3]. The accumulation of CO_2 has increased the atmospheric concentration from 280 ppm in pre-industrial times to ~390 ppm today [4]. The pace of CO_2 emissions is expected to increase significantly as India, China, and other developing nations output more CO_2. For every ppm increase in CO_2, the average Earth temperature appears to rise due to a greenhouse effect [3]. Many climatologists believe that there will be serious and irreversible consequences to world climate, partly due to positive-feedback loops, if the CO_2 concentrations increase significantly beyond 550 ppm.

The profligate burning of fossil fuels will lead to their inevitable extinction, which is not only catastrophic for energy and climate reasons, but also because they are quite useful for making several materials like plastics cheaply. Due to the need for minimizing fossil-fuel CO_2 emissions that impact climate change and due to the exhaustion of these fossil-fuel energy sources, our planet will need to function increasingly on renewable energy sources. These sources include solar power, wind power, hydroelectric power, wave power, tidal power, geothermal power, and biofuels. Since the areal power densities of these sources are relatively small, it is imperative that our power consumption be reduced. Most of our power consumption arises from transportation, heating, electricity usage, and material-synthesis costs.

We discuss how electric cars, powered by batteries driving motors, enable improvements in transport energy efficiency, i.e., energy consumed per person-km, over those of gasoline-powered cars. We shall discuss an equivalent circuit for a car, which will allow us to draw on principles of low-power design in electronics to understand how power consumption in cars can and is being reduced. We shall compare the energy efficiency of advanced electric cars versus cheetahs, the fastest land animals on earth. Even though legged locomotion is significantly less

[1] Interestingly, the average national per-capita income of a person in K$ divided by 4 is a good predictor of that nation's average per-person power consumption in kW.

efficient than wheels on flat terrains, we shall see that animals have impressively good transport energy efficiency when compared with even highly energy-efficient electric cars.

We will focus on two renewable sources that are likely to be very important in our future, namely, solar photovoltaics and biofuels. The basic principles of phototransduction described in Chapter 11 will be useful for understanding how solar photovoltaic cells function. We shall delve deeper into phototransduction in this chapter to understand the limits of solar-cell efficiency. Solar photovoltaic sources are important at small scales, e.g., for solar photovoltaic cells that power portable and biomedical applications, and also at large scales, e.g., for 300 MW electric generators. Solar energy is widely viewed as the most important renewable energy source because of its relatively high power density and ubiquitous presence [5]. We shall discuss some challenges in making solar electricity generation cost effective. We conclude by discussing biofuels, which are created by plants storing the energy of sunlight in chemical bonds through the process of photosynthesis. Biofuels represent an energy-dense method for the storage and distribution of solar energy. Such biofuels could be useful in cars and in implantable biomedical systems in the future.

26.1 Sources of energy

Figure 26.1 shows six common sources of energy that we can harvest. We have discussed RF energy harvesting in near-field systems in Chapter 16 for biomedical implants and in far-field systems for cardiac monitoring in Chapters 17 and 20. In general, ambient RF energy from cell phones and wireless devices in the environment may be harvested. Implantable biomedical systems can potentially harness the energy of blood flow or the energy of airflow during respiration to function; work in this area is just beginning. Ultra-low-power outdoor monitoring applications can exploit potential differences between two points on a tree trunk, which can vary by a few hundreds of mV, to operate [6]. In this chapter, we shall primarily focus on inertial-motion, heat, and solar energy harvesting.

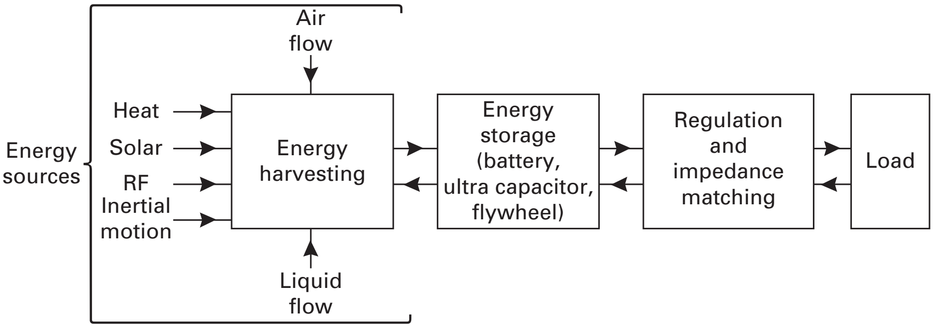

Air flow
Energy sources
Heat
Solar
RF
Inertial motion
Liquid flow
Energy harvesting
Energy storage (battery, ultra capacitor, flywheel)
Regulation and impedance matching
Load

Figure 26.1. A typical energy-harvesting architecture.

Electrical equivalent:
Every capacitor holds its charge unless it is charged or discharged by an electrical current.

2. Newton's second law:

$$F = mdv/dt \tag{26.1}$$

F is the force, m is the mass, and v is the velocity of the moving or stationary mass.
Electrical equivalent:

$$I = CdV/dt \tag{26.2}$$

I is the current, C is the capacitance, and V is the voltage on the capacitor.

3. Newton's third law:
For every action, there is an equal and opposite reaction.
Electrical equivalent:
In any two-terminal electrical element, whether active or passive, dependent or independent, linear or nonlinear, the current flowing into one terminal on the element is equal to the current flowing out of the other terminal of the element.

In the formulation above, current is analogous to a force, capacitance is analogous to a mass, and voltage is analogous to a velocity. The electrical equivalent of Newton's third law is such that it is automatically satisfied and represented in any circuit. Mutual interactions between two bodies are represented as a floating current between two nodes such that one of the currents through the two-terminal element creates a sink current on the node that it is attached to while its paired current creates a source current on the node that it is attached to. Thus, Newton's third law is nothing more or less than stating that a floating current source between two nodes may always be represented as a grounded sink current at one node and a grounded source current at the other node. The automatic and natural representation of Newton's third law by a circuit makes electrical representations of mechanical systems powerful because one is relieved from the burden of having to constantly keep track of symmetric pushing and pulling between bodies. Furthermore, force balancing is also automatic. Since the voltage on a capacitor stops changing when all the currents flowing towards (or away from) it sum to zero, Kirchhoff's current law is the law of force balance. Vector forces require 3D electrical circuits because the electrical analogies of mechanical systems hold separately for each of the x, y, and z components of force and velocity. For example, Figure 17.1 shows how circuit descriptions of Maxwell's equations conceptually represent vectors.

In the formulation above, capacitance is a mass. If

$$\begin{aligned} F &= k\int vdt \\ I &= \frac{1}{L}\int Vdt, \end{aligned} \tag{26.3}$$

far. Its delivered power density of $\sim 0.2\,\mathrm{W/m^2}$ is in excess of what an average 10%-efficient solar cell might deliver in indoor environments. An efficient charge pump for such thermal harvesters is described in [18].

26.5 Power consumption of the world

Mackay estimates the average power consumption of an affluent British citizen today in his book [1]. If we adapt his units of 1 kW h/day to simple kW units with the conversion factor 1 kW h/day = 41.67 W, we find that this consumption may be broken down as shown in Table 26.1.

The costs of Table 26.1 are estimated for an affluent British citizen. An average British citizen actually consumes 5.2 kW, an average European citizen consumes 5.46 kW, while an average American citizen consumes nearly 10.4 kW. The world average is 2.5 kW with great variance across nations. Since there are

Table 26.1 Power consumption of the world

Item	Power consumption	Comment
1. Car usage	1.67 kW	30 mph at 30 mpg for ~1 hour at 10 kW h/liter for gas with 3.8 liters = 1 gallon. Or equivalently, the cost of an average 42 kW car driven for ~1 hour each day.
2. One transatlantic flight per year on a Boeing 747	1.25 kW	Such planes operate at 0.14 mpg but amortize this cost over ~ 400 passengers such that they effectively operate at ~ 60 mpg per person. The power consumption of a Boeing 747 is ~150 MW.
3. Heating	1.540 kW	Not important in some geographical areas.
4. Material synthesis energy costs	2.08 kW	It costs energy to manufacture appliances.
5. Electric lighting	0.167 kW	Estimated for an average home.
6. Electric gadgets	0.208 kW	Washers, dryers, cell phones, etc.
7. Material transport	0.500 kW	Trucking and transportation costs to move materials.
8. Food	0.625 kW	This energy cost in food only tracks industrial energy flows associated with food, not the natural embedded energy in food. For example, it costs energy to transport food, and to maintain animals to be used later as food.
9. Defense	0.167 kW	These national costs are amortized per person.
Total	**8.207 kW**	Does not include the cost of imported goods, which bear their own energy costs, at 1.667 kW.

approximately 6 billion people on our planet today, the power consumption of the world is 15 TW. The electricity consumption of the world is 2 TW. However, since typical generating stations burn fossil fuels like coal to generate electricity and are only 40% efficient, the actual power consumption due to electricity use is 5 TW. We notice that a large fraction of the power consumption of the world revolves around transportation, heating, and electricity costs. This book has already discussed principles for lowering power in electrical systems. Now, we shall discuss some principles for the design of low-power transportation systems of the future.

26.6 A circuit model for car power consumption

Figure 26.5 shows a circuit model of a car that is useful for understanding factors that affect its power consumption. We shall use current to represent force and voltage to represent velocity in accord with Equations (26.1), (26.2), (26.3), and (26.4). Thus, mass is represented by a capacitance, mechanical damping by a conductance, and mechanical compliance by an inductance. A chemomechanical dependent force i_{ENG} due to the burning of fuel along with a Norton-equivalent mechanical admittance G_{ENG} represents the characteristics of the engine power source. The fuel-to-mechanical work efficiency is typically 25% such that the $i_{ENG}v_{ENG}$ power output by the engine requires $4i_{ENG}v_{ENG}$ power to be extracted from the chemical fuel. The motions of the engine are periodic. The engine force is conveyed via gears to provide force to the car wheels. The transformer in Figure 26.5 represents a lossless gearbox (and transmission) that performs an impedance transformation. The reflected admittance of the secondary wheel-and-road side to the primary engine side must be such that most of the power output by the engine is dissipated in the reflected admittance, not in G_{ENG}, which is usually the case. As the characteristics of the impedance in the secondary change with flat, uphill, or downhill road conditions, the gear ratios are changed such that this efficiency is preserved.

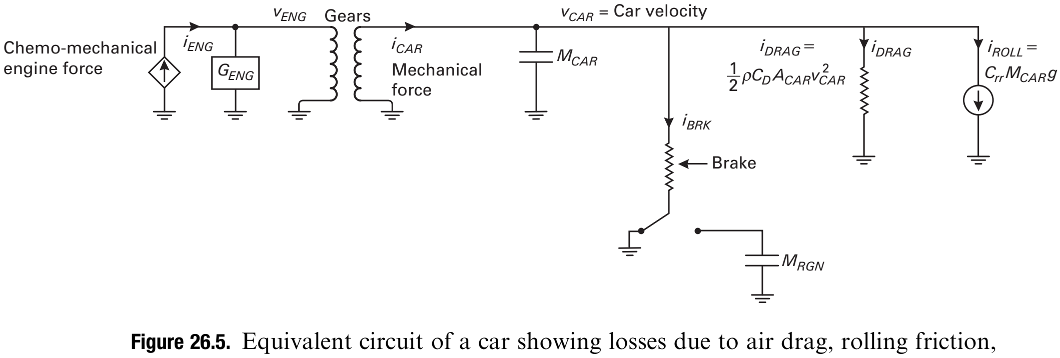


Figure 26.5. Equivalent circuit of a car showing losses due to air drag, rolling friction, braking, and chemical-to-mechanical energy conversion.

cycles ($\sim$ 100,000 miles), and it takes 3.5 hours for a full battery recharge although a full recharge may rarely be necessary. It implements regenerative braking. The lithium-ion battery has several short-circuit protection features including in-built fuses that disconnect it in situations of high temperature and pressure. The battery is architected to be safe even during collisions.

Most importantly, the Tesla Roadster's transport energy efficiency is $\sim$500 N, about 6 times better than that of an average car. The transport efficiencies of several lighter, lower-range, and low-speed electric cars are not significantly different from that of the Roadster and some are much worse [1]. To be fair to the average gasoline car, though, the Tesla Roadster uses high-grade electric energy, while the average car needs to extract its energy from fossil fuel. Most electricity generating fossil-fuel plants are 30%–40% efficient such that one could argue that the real improvement of a Roadster is a factor of 2× to 2.4×. Nevertheless, the Tesla Roadster does illustrate that direct conversion between high-grade forms of energy, e.g., electrical to mechanical rather than from chemical-to-heat-to-mechanical as in a gasoline car, is efficient. In the future, if electricity is generated in solar plants, such a car could indeed have a zero-emission footprint, especially if it is manufactured in plants using solar electricity as well. Even though the energy density of the lithium-ion battery that was used is 11 times less than that of gasoline, the weight of the car is manageable because the heavy gasoline engine is replaced with an electric motor.

26.8 Cars versus animals

Another impressive example in transportation engineering is the cheetah (*Acinonyx jubatus*), the fastest land animal. Its top sprint speed has been measured to be 30 m/s, i.e., 68 mph [22]. It can accelerate to 68 mph in 3 seconds, faster than a Tesla Roadster, which gets to 60 mph in 3.9 seconds, and faster than most high-performance cars [23], [24]. Since the average cheetah weighs nearly 50 kg, we can estimate that its mechanical power output during this acceleration is $(1/2)50(30^2)/3 = 7.5\,\text{kW}$. Its rudder-like tail enables it to make incredibly quick turns during its chase of a prey animal. The cheetah uses its spring-like backbone to partly store and regenerate energy in each stride and is airborne for more than half its stride. Its transportation efficiency for aerobic speeds, which can typically be maintained for long distances only if they are less than half the top speed [25], has been measured to be equivalent to 0.14 ml of oxygen consumption per g.km [26]. From the energetics of a glucose or carbohydrate reaction, and from the weight of an average cheetah, these numbers work out to an energy efficiency of 132 N. The cheetah's transport energy efficiency is 4 times better than that of a highly energy-efficient electric car. What is more impressive is that this transport energy efficiency is achieved even though the cheetah has to make do with a 25% efficient engine (fuel-to-mechanical-work) and that it runs with legs, not as optimal as wheels on flat terrain. For example, the energy efficiency of humans walking at

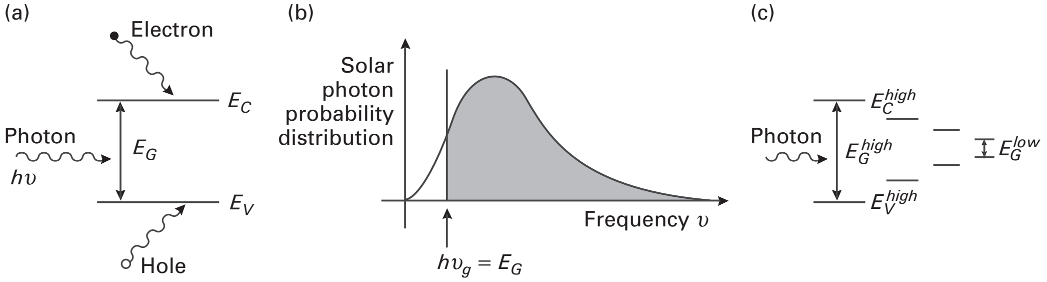


Figure 26.8. Generation of an electron-hole pair by a photon in a solar cell in (a). The only photons that can generate such electron-hole pairs must have energy greater than the bandgap energy such that only a fraction of the incoming solar black-body spectrum at 6000 K can be converted to electricity as shown in (b). Solar photovoltaics that attempt to improve energy efficiency use materials with multiple bandgaps as shown in (c) to increase their photon collection efficiency.

From the shaded area in Figure 26.8 (b), we can compute the electric energy generated by the high-energy photons. The ratio of the electric energy to the total incoming radiation energy then yields an ultimate limit for the solar-cell efficiency.

Shockley and Quiesser showed that their ultimate limit could be attained if the only method for electron-hole pair destruction at 300 K is radiative, i.e., incoming thermal energy at 300 K from the environment creates electron-hole pairs, which then recombine to generate outgoing 300 K blackbody radiation that balances such generation. In this limit, the value of I_S is as low as it can possibly be, and the open-circuit voltage of the junction asymptotes to the bandgap voltage, $V_G = E_G/q$, of the semiconductor. Any other forms of recombination, e.g., due to impurities in the semiconductor, decrease the minority-carrier lifetime τ, and consequently decrease the minority-carrier diffusion length and increase I_S. Hence, the open-circuit voltage given in Equation (26.29) is reduced. The use of pure semiconductors is thus important for achieving high efficiency, but making pure materials is expensive. The solar cell must be thick enough such that the probability of absorbing a photon is high. Figure 11.6 (a) and Figure 11.6 (b) in Chapter 11 show that bluer photons are absorbed at shallower depths while redder photons are absorbed at deeper depths. The solar cell should not be too thick since the chance for electron-hole recombination is then increased. Hence, there is an optimal thickness in solar cells that maximizes efficiency. However, maximum power transfer is not attained at this optimum. Topologies are now being explored in which electron-hole pairs are created by photons along one spatial dimension while electrons and holes travel a short distance in an orthogonal direction to create an electrical voltage; thus, the electron-hole pairs are given little opportunity to recombine [41].

Figure 26.8 (c) illustrates an idea for increasing the fraction of photons that contribute to electrical energy in a solar cell. If we have multiple pn junctions made of materials with progressively smaller bandgaps, we can first extract the

the enzymes that are used to oxidize the fuel lose their efficacy after some time, making them unattractive in long-term implants or in car applications that may need years of battery operation. Cells solve these problems by constantly degrading and regenerating enzymes needed for various biological processes such that they always maintain their efficacy.

26.12 Energy use and energy generation

Low-power systems can enable sources of energy that would normally be impractical for powering an application to become practical. In inductive links (Chapter 16), in piezoelectric harvesters, in electric motors, and in electric cars, we have seen that a low-power system can improve the energy efficiency of an overall system by altering the effective reflected load seen by the energy source. A system is most energy efficient when there is minimal power transfer but only achieves 50% efficiency at maximal power transfer. In the chapter on batteries (Chapter 25), we saw that a low-power system does not increase battery lifetime merely because of a low-power draw but also because it enables higher energy density, higher efficiency, and lower fade capacity in the battery. In electric cars, the use of a relatively light and efficient electric motor rather than a heavy engine enabled an energy source with significantly lower energy density, i.e., a battery, to become practical for powering a car. The cost effectiveness of solar electricity is improved if electricity consumption can be reduced, thus enabling green electricity rather than 'red' electricity. The principles of adiabatic design in Chapter 21, the Shannon limit on the minimum energy needed to compute in Chapter 22, and the Ragone-curve tradeoff between energy density and power density in Chapter 25 all reveal that if you can pull energy out of a source slowly, you can waste less of it and create a higher capacity to store it. The central take-home lesson from these numerous examples is that energy use and energy generation are deeply linked. We must try to optimize them jointly rather than treat them as two separate problems.

References

[1] David J. C. MacKay. *Sustainable Energy – Without the Hot Air* (Cambridge, UK: UIT Cambridge, Ltd., 2009).

[2] Department for Environment Food and Rural Affairs (DEFRA), *Guidelines for Company Reporting on Greenhouse Gas Emissions*, London (2005); available from: http://www.defra.gov.uk/environment/business/reporting/pdf/envrpgas-annexes.pdf.

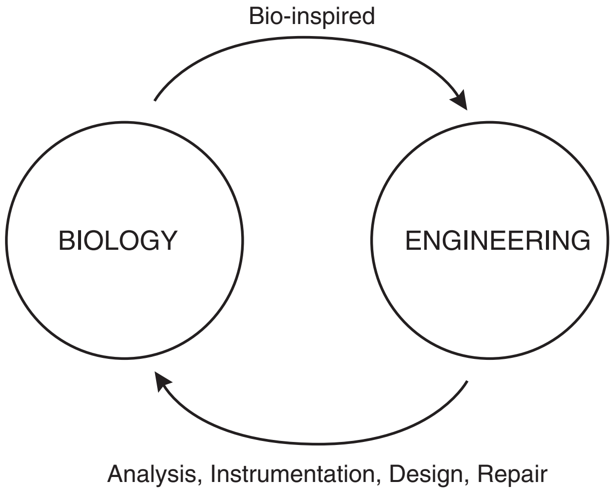
[3] D. L. Green and Scientific American. *Oil and the Future of Energy: Climate Repair, Hydrogen, Nuclear Fuel, Renewable and Green Sources, Energy Efficiency* (Guilford, Conn.: The Lyons Press, 2007).

[4] T. R. Karl and K. E. Trenberth. Modern global climate change. *Science*, **302** (2003), 1719–1723.

Epilogue

Information is represented by the states of physical devices. It costs energy to transform or maintain the states of these physical devices. Thus, energy and information are deeply linked. It is this deep link that allows us to articulate information-based principles for ultra-low-power design that apply to biology or to electronics, to analog or to digital systems, to electrical or to non-electrical systems, at small scales or at large scales. The graphical languages of circuits and feedback serve as powerful unifying tools to understand or to design low-power systems that range from molecular networks in cells to biomedical implants in the brain to energy-efficient cars.

A vision that this book has attempted to paint in the context of the fields of ultra-low-power electronics and bioelectronics is shown in the figure below. Engineering can aid biology through analysis, instrumentation, design, and repair (medicine). Biology can aid engineering through bio-inspired design. The positive-feedback loop created by this two-way interaction can amplify and speed progress in both disciplines and shed insight into both. It is my hope that this book will bring appreciation to the beauty, art, and practicality of such synergy and that it will inspire the building of more connections in one or both directions in the future.



Epilogue The two-way flow between biology and engineering.