

Exploiting Device Physics in Circuit Design for Efficient Computational Functions in Analog VLSI

Andreas G. Andreou
Electrical and Computer Engineering
Johns Hopkins University,
Baltimore, Maryland 21218 USA

Chapter 4, Low Voltage Low Power Integrated Circuits and Systems
IEEE Press, Edgar Sanchez Sinencio, ed., 1998

September 24, 1997

1 Introduction

Computation as performed by “real” systems is an irreversible physical process and as such it is associated with an inevitable amount of energy dissipation [1, 2]. This is true for both human engineered VLSI systems (Chapter 9 in [3]), and for Nature’s machinery, biological systems.

Biological organisms excel at solving hard problems in sensory communication and motor control –speech and vision– by sustaining high computational throughput while keeping energy dissipation at a minimal level. The total power consumed by the awaken human brain is about 10 Watts [4, 5]. The existence of such remarkably efficient computational structures is a result of an evolutionary necessity towards systems that are **truly autonomous**, and thus constrained to operate under strict bounds of size, weight and at temperatures where favorable conditions exist for the development of life as “we” know it of ($T = 300K$).

Neural systems operate from power supplies of approximately 100mV, only a few kT/q , while the bandwidth limitation at the **macroscopic** scale of constituent components is only a few hundred kHz, which is comparable and sometimes even lower than the bandwidth of the signals that have to be processed. Typical current values in biology are in the pico to nano ampere range [7]. Neural “wet-ware” are characterized by components that are inherently highly variable and noisy. Nevertheless, natural sensors and natural computing systems are capable of remarkable performance that our best technology is unable to match. Consider the mammalian retina, a thin structure (approximately $500\ \mu m$) at the back of the eyeball and an outpost of the brain. It serves as both sensor and preprocessor for the visual cortex. Nobel laureate David Hubel writes [6]:

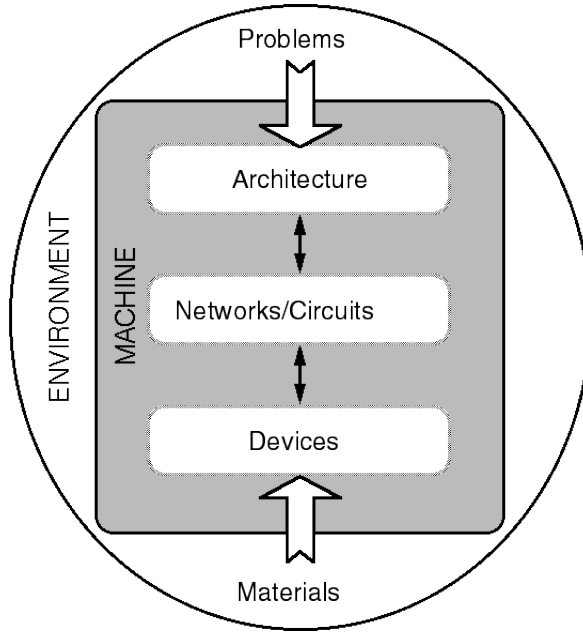


Figure 1: A physics of computation view for the synthesis of complex, highly integrated microsystems. This methodology optimizes the design at and between all levels of a system; from devices to circuits and network levels, and all the way to the architecture.

The eye has often been compared to a camera. It would be more appropriate to compare it to a TV camera attached to an automatically tracking tripod— a machine that is self-focusing, adjusts automatically for light intensity, has a self cleaning lense, and feeds into a computer with parallel processing capabilities so advanced that engineers are only just starting to consider similar strategies for the hardware they design No human inventions, included computer assisted cameras can begin to rival the eye. (*David Hubel: Eye, Brain, and Vision; page 33*).

The effectiveness and efficiency of biological systems stems partly from exploiting *prior* knowledge about the problems that they encounter [9]. Such information, in the form of *internal models*, reflects the statistical properties of the natural environments in which they function. The exploitation of such prior knowledge plays an important role in both the evolutionary development of neural structures as well as the adaptation mechanisms during system function. Exploitation of prior knowledge is also the key to successful algorithm design for speech and vision human engineered systems (see discussion in [8, 18]).

Perhaps equally important is the way *function* emerges from *form*. Systems of remarkable complexity not just in the number of elements but also in composition have evolved into a hierarchy of structures with remarkable computational capabilities that clearly the best digital technologies can not match.

1.1 From device physics to analog VLSI systems

At the most basic level, analog VLSI technology offers the possibility of exploring experimentally computation by truly complex, *real systems* which lie beyond digital computing and the symbolic

processing paradigm.

It is appropriate at this point to ask the question: what kind of computational primitives does one have? In CMOS silicon these are continuous functions (analog) of *time*, *space*, *voltage*, *current* and *charge*. To help manage the complexity in VLSI systems, these functions will be considered at three hierarchical levels: the *device level*, the *circuit level* and the *architectural level*.

Device level: At the lowest level, gain, is provided by MOS transistors operating in subthreshold region [11, 17, 13]. In this regime device physics yield the following functional form for the drain current in terms of the voltages at its four terminals.

$$I = I_0 S \mathcal{G}(\kappa V_{GB}) [\mathcal{H}(V_{SB}) - \mathcal{H}(V_{DB})] \quad (1)$$

where $\mathcal{G}$ and $\mathcal{H}$ are growing and decaying exponential functions respectively. The terminal voltages V_{GB}, V_{SB}, V_{DB} are referenced to the substrate and are normalized to the thermal voltage (kT/q). The constant I_0 depends on mobility (μ) and other silicon physical properties. S is a geometry factor, the width W to length L ratio the device. The Pauli exclusion principle dictates that the prefactors that normalize the voltages to thermal voltages in the exponentials be less than or at best equal to unity. The MOS transistor has excellent circuit properties as a voltage-input, current-output device (transconductance amplifier) with good fan-out capabilities (high transconductance $\mathcal{G}$) and good fan-in capability (almost zero conductance at the input).

The exponential functions of voltage in the square brackets of Equation 1, correspond to Boltzmann distributed charges at the source and drain.

$$I \propto [\mathcal{Q}_S - \mathcal{Q}_D] \quad (2)$$

The charge-based representation depicted in Equation 2, suggests that the MOS transistor in subthreshold is a highly linear device; a property that finds many uses in analog circuit design. This property was first observed by Kwabena Boahen and discussed in [24] where the concept of a *diffusor* was introduced. The view of an MOS transistor in subthreshold as a basic diffusive element allows for the effective implementation of systems that exploit properties of elliptic partial differential equations.

The transfer characteristics of MOS transistors are plotted in Figure 2 for both the above and subthreshold regime. The transconductance per unit current increases as the current decreases—throughout the above-threshold and transition regions—and reaches a maximum in the subthreshold region. In highly integrated VLSI systems, small geometry devices must be used to achieve high densities. Small device geometries and high transconductance per unit current makes the drain current strongly dependent on variations of the process-dependent parameters, in particular I_0 , which is the source for the variability observed in the drain currents of Figure 2. The apparent improvement in device matching for higher values of gate-source voltage, is simply a manifestation of reduced transconductance per unit current as the device enters the above threshold regime.

Our preference for subthreshold operation, (despite to what seems to be worse matching characteristics), is based on the observation that: **“Active devices should be used in the region where their transconductance per unit current is maximized”**. In this way

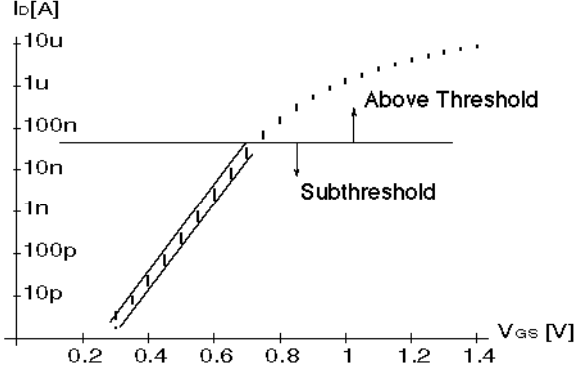


Figure 2: Measured drain current I_d versus gate-source voltage V_{GS} for 32 small geometry transistors ($4 \times 4 \mu m$) fabricated in a $2 \mu m$ n-well CMOS process; drain-source voltage of $V_{DS}=1.5$ Volts. The fuzziness in the current, (mismatch between devices), is constant in subthreshold (on a $\log(I)$ scale) and decreases as the device enters the transition and above threshold regime. (Data from [15]).

one can minimize the energy per operation and maximize the speed per unit power consumed, i.e. minimize the power-delay product:

$$\frac{\text{speed}}{\text{power}} = \frac{1/\tau}{I\Delta V} = \frac{g_m/C}{I^2/g_m} = \frac{1}{C} \left(\frac{g_m}{I} \right)^2 \quad (3)$$

A squared factor is obtained because both voltage swings (ΔV) and propagation delays (τ) are inversely proportional to the transconductance g_m for a given current level. However, only a linear factor is realized if the power supply voltage is not reduced to match the voltage swings $\Delta V \sim I/g_m$. When the device is operated in subthreshold, the drain-source conductance saturates at a few (kT/q), (see Equation 1). Power supplies of a few (kT/q) are also possible and thus power supplies can theoretically match the voltage swing levels. The capacitance C is analogous to an inevitable “mass” of the switching node. When physical structures are miniaturized, this capacitance is reduced and the power-delay product improves. This simple scaling “law” has been one of the driving forces towards high levels of system integration and miniaturization in the microelectronics industry.

The maximum useful frequency of operation possible with an MOS transistor, when operating in subthreshold is determined by its transition frequency f_T which has an upper limit f_{Tmax} of:

$$f_{Tmax} < \frac{\mu (kT/q)}{\pi L^2} \quad (4)$$

where μ is the effective carrier mobility and L is the device channel length. The transition frequency of a device is essentially the frequency where its gain-bandwidth product (as determined by the internal gain and parasitic capacitances of the transistor) is unity.

Circuit level: It is at this level where the synthesis of computational structures begins and manifests itself as the emergence of *networks*. Conservation laws, that is conservation of charge (Kirchoff’s Current Law), $\sum_i I_i = 0$, and conservation of energy (Kirchoff’s Voltage

Law), $\sum_i V_i = 0$, are used to realize simple constraint equations. The important concept of *negative feedback* is also exploited to trade the gain in the active elements for precision and speed in the circuits.

Aside from the benefits of a device with a large gain, the exponential relationships between the controlling voltages and the current depicted in Equation 1 endow the MOS transistor with some interesting circuit properties. There exists a powerful synthesis (and analysis) procedure which can be used to generate a wide variety of circuits that perform linear and non-linear operations in the current domain, and relies on the exponential form of current-voltage non-linearities. This procedure is based on what is known as the *Translinear Principle* [21] originally used in the context of bipolar transistors. The synthesized circuits are called *translinear* and may involve operations of one or more variables, such as products, quotients, power terms with fixed exponents, as well as scalar normalization of a vector quantity.

The application of the translinear principle to circuits implemented with MOS devices operating in subthreshold saturation, and an extension to the subthreshold ohmic regime is discussed in the next section. One fascinating aspect of translinear circuits is that while the currents in its constitutive elements (the transistors) are exponentially dependent on temperature, the overall input/output relationship is insensitive to isothermal temperature variations. The effect of small local variations in fabrication parameters can also be shown to be temperature independent.

To demonstrate how computational primitives emerge at the network level from device physics of the underlying technology, let us consider an example of a summing operation, *local aggregation*. Such linear addition of signals over a confined region of space occurs throughout the nervous system. Aggregation was discussed in Chapter 6 of [11], (also in [78]), and it is the basis for many neuromorphic silicon VLSI systems described therein. Here we take a close look at *diffusion*, the physical process that underlies local aggregation in the nervous system, contrast it with the process of diffusion in MOS transistors and come up with a novel network design technique.

The first network uses voltages and currents (Figure 3a). Its node equation is

$$\frac{dV_n}{dt} = \frac{4G}{C} \left(\frac{1}{4}(V_j + V_k + V_l + V_m) - V_n \right) \quad (5)$$

which is homologous to the diffusion equation since the term in large parenthesis is a first-order approximation to the Laplacian. However, this solution is not amenable to VLSI integration because transconductances (G) with a large linear range consume large amounts of area and power.

The second network uses charges (positive) and currents (Figure 3b). Its node equation is

$$\frac{dQ_n}{dt} = 4D \left(\frac{1}{4}(Q_j + Q_k + Q_l + Q_m) - Q_n \right) \quad (6)$$

Note that dQ_n/dt is the same as the current supplied to node n by the network. This solution is easily realized by exploiting diffusion in subthreshold MOS transistors. As shown in the device section, the current is linearly proportional to the charge difference across the channel (See Equation 2). Therefore, the diffusion process may be modeled using devices with identical

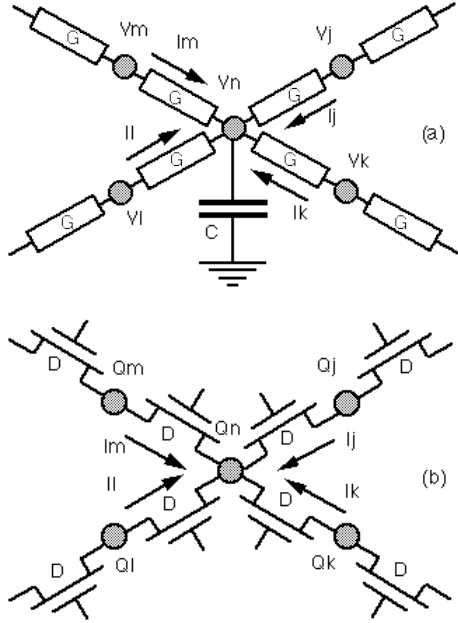


Figure 3: Simulating diffusion with (a) conductances and voltage/current variables or (b) diffusors and charge/current variables.

geometry S and identical gate voltages. The former guarantees they have the same diffusivity and the latter guarantees that the charge concentrations at all the source/drains connected to node n are the same and equal Q_n .

In both of these networks, the boundary conditions may be set up by injecting current into the appropriate nodes. In the voltage-mode network, the solution is the node voltages. They are easily read without disturbing the network. On the otherhand, the network depicted by Equation 6 represents the solution by charge concentrations Q_S and Q_D at source/drains—not the charge on the node capacitance. The source/drain charge cannot be measured directly without disturbing the network. It may be inferred from the node voltage.

Architectural level: At this level, differential equations from mathematical physics will be employed to implement useful signal processing functions, still in the form of constraint equations. For example, the *biharmonic* equation

$$\lambda \nabla^2 \nabla^2 \Phi + \Phi = \Phi_{in} \quad (7)$$

where $\nabla^2 \equiv \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplacian operator, constrains the sum of the fourth derivative of Φ and Φ itself to be equal to a fixed input Φ_{in} . From a *statistical* signal processing view-point, solutions to this equation could represent an *optimal estimation* Φ of the underlying smooth continuous function, given a set of noisy, spatially sampled observations Φ_{in} . The solution is optimal in the sense that it simultaneously minimizes the squared error and the energy in the second derivative—the parameter λ is the relative cost associated with the derivative term. A large value for λ favors smooth solutions while a small value favors a closer fit.

We have already seen how a diffusive grid can be used to compute a discrete-approximation

of the Laplacian. In Section 5 of this chapter we show how a model of early visual processing is related to the biharmonic equation and can be realized using diffusive networks.

In this chapter we begin at the basic technology device level and proceed all the way to the system level demonstrating how indeed a neuromorphic approach and a physics of computation view can yield fruitful results in analog VLSI.

The chapter is divided into six sections. Section 2 begins with an introduction to the physics of MOS transistors including floating gate MOS transistors (FGMOS). A basic review of subthreshold MOS and bipolar transistors device physics contrasting their operation, is provided since the large signal properties of the devices are *key* to the subject matter. In particular, we emphasize the translinear properties of bipolar transistors and those of MOS transistors in subthreshold. The current mode design methodology is introduced in Section 3, where the concepts of a current conveyor, a canonical block that is widely employed in minimal complexity circuits is presented. Circuit techniques based on the translinear principle that employ MOS transistors in subthreshold saturation and ohmic regimes are introduced in Section 4. In the same section, we discuss both translinear loops (TL) composed of generalized diodes and current sources, and translinear networks (TN) that include voltage sources as well. In Section 5 we discuss an analog VLSI system for early vision, a silicon retina that has evolved from the design methodology presented in the earlier sections. A general discussion in Section 6 concludes the chapter.

2 Technology and Devices

CMOS technology and, in particular, subthreshold MOS operation have long been recognized as the technology of choice for implementing digital VLSI and analog LSI circuits that are constrained by power dissipation requirements [13, 14]. CMOS has the highest integration density attainable today, making it especially attractive for systems that demand high overall computational throughput. Moreover, the physical properties of silicon and its native oxides, together with recent advances in micromachining of electromechanical elements [31], make silicon-based technologies the prime candidate for highly integrated, systems that are not only complex in the strict sense of a large number of gates but also in the design of the individual components.

In the last few years we have seen analog computation based on physical properties of silicon devices, emerge as an alternative to abstract mathematical algorithms implemented on digital hardware [11]. Much like biological systems, high processing throughput is attained through massive parallelism of slower but energy efficient analog circuits operating from low power supplies and operating in subthreshold CMOS [13, 14, 11] where the devices have the highest possible gain and the lowest noise. Our work [15, 16, 17, 18, 19] has followed a similar line of thinking.

At the lowest level, gain is provided by MOS transistors operating in subthreshold region [11, 17, 13]. Since the adopted design methodology depends critically on an in depth understanding of the underlying device physics, a charge-based transistor model [11, 17] that preserves the symmetry between the source/drain terminals of an MOS transistor is briefly discussed here. Which terminal of the device actually serves as the source or the drain is determined by the circuit, the bias conditions—and even the input signals. This symmetric view of an MOS transistor enabled us to extend the translinear principle to operation in the subthreshold ohmic regime [17].

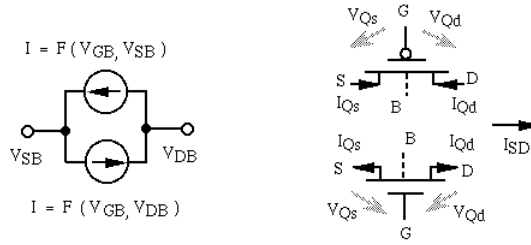


Figure 4: Large signal model for an NMOS transistor (left). Adopted conventions for current decomposition in PMOS and NMOS devices (right).

The MOS device has a very simple current-charge relationship because diffusion and drift are both proportional to the concentration gradient. As shown in Appendix A of [11] and in [17], as well as in Chapter XX of this book, this yields a quadratic expression for the current that consists of two independent opposing current components; the source driven current I_{Qs} and the drain driven current I_{Qd} . These currents are related to charge densities at the source Q_s° and Q_d° at the drain of the device.

The device drain–source current can thus be written as

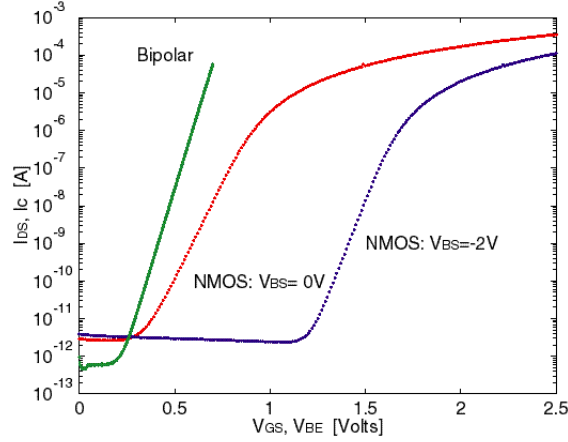


Figure 5: Measured current I_{DS} and I_C versus controlling voltage V_{GS} and V_{BE} respectively. The MOS transistor has dimensions of $(16 \times 16 \mu m^2)$ and is fabricated in a $1.2 \mu m$ n-well CMOS process and is biased at a drain-source voltage, $V_{DS}=1.5$ Volts. The current is measured at two different substrate voltage bias conditions. The bipolar transistor is a vertical device with an emitter area of $(16 \times 16 \mu m^2)$ fabricated in a $2 \mu m$ n-well CMOS process and biased with $V_{CE}=1.5$ Volts. $T = 301.5$ K.

$$I \equiv I_{Q_s} - I_{Q_d} = \mu \frac{W}{L + \left(\frac{\mu}{v_o} \frac{Q_s^{\circ 2}}{C_{ox}^{\circ} + C_{dep}^{\circ}} + \frac{kT}{q} Q_s^{\circ} \right)} \left[\left(\frac{1}{2} \frac{Q_s^{\circ 2}}{C_{ox}^{\circ} + C_{dep}^{\circ}} + \frac{kT}{q} Q_s^{\circ} \right) - \left(\frac{1}{2} \frac{Q_d^{\circ 2}}{C_{ox}^{\circ} + C_{dep}^{\circ}} + \frac{kT}{q} Q_d^{\circ} \right) \right] \quad (8)$$

W is the width, L is the length of the channel (μ) is the effective channel mobility and (v_o) is the saturation velocity of the carriers. The capacitances C_{ox}° and C_{dep}° are the gate oxide and depletion area capacitances of the channel. A key property of the MOS device that makes this possible is lossless channel conduction. Unlike a bipolar transistor, the controlling charge on the gate is isolated from the charge in transport by the almost infinite gate-oxide resistance. Therefore, there is no recombination between the current-carrying charge in the channel and the current-modulating charge on the gate.

The familiar ohmic/saturation dichotomy introduced in voltage-mode design can be reformulated in terms of the opposing drain and source driven current components. In the saturation region, $|I_{Q_d}| \ll |I_{Q_s}|$; $I \approx I_{Q_s}$; therefore the current is independent of the drain voltage. In ohmic, $I_{Q_d} \simeq I_{Q_s}$ and $I = I_{Q_s} - I_{Q_d}$ and therefore the current depends on the drain voltage as well as the source and gate voltages. The functional dependence of the current components on the terminal voltage is fixed and remains the same throughout the ohmic and saturation regions.

The charge densities at the source and drain terminals can be related to the terminal voltages. In general, the charge-voltage relationship is much more complicated than the current-charge one because both the mobile charge and the depletion charge are involved in the electrostatics. The device current in Equation 8 can thus be written as a function $\mathcal{F}$ of the terminal

voltages with a general functional form for the current-voltage relationship valid for all the regions of operation given by

$$I_{SD} \propto \mathcal{F}(V_{GB}, V_{SB}) - \mathcal{F}(V_{GB}, V_{DB}) \quad (9)$$

This functional form was first introduced by [34] for above threshold operation and is also discussed in [35]. For an n-type device, $\mathcal{F}$ is a nonpositive, monotonically decreasing function of V_{GB} and a monotonically increasing function of V_{SB} .

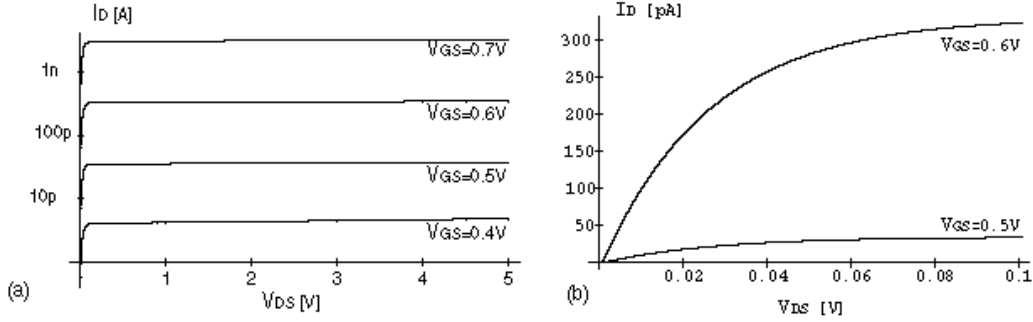


Figure 6: Measured output characteristics of NMOS transistors in subthreshold. Device dimensions are $W = L = 16\mu m$. The two sets of data emphasize the good saturation characteristics (Left) and the non-linear ohmic behavior at low drain source voltages (Right).

In the subthreshold region, the following factorization of $\mathcal{F}$ is also possible [14, 11].

$$I \propto \mathcal{G}(V_{GB}) [\mathcal{H}(V_{SB}) - \mathcal{H}(V_{DB})] \quad (10)$$

where $\mathcal{G}$ and $\mathcal{H}$ are exponential functions. This shows that the source-driven and drain-driven components are controlled independently by V_{SB} and V_{DB} . However, V_{GB} , acting through the surface potential, also controls both components in a symmetric and multiplicative fashion. In this mode of operation the MOS transistor has been called a *diffusor* [24] in analogy with the variable conductance electrical junctions in biological systems.

An expression for the current in an NMOS transistor operating in subthreshold can thus be written [14, 11] as

$$I_D \equiv I_{DS} = I_{n0} S \exp(\kappa_n V_{GB}/V_t) [\exp(-V_{SB}/V_t) - \exp(-V_{DB}/V_t)] \quad (11)$$

and for a PMOS transistor

$$I_D \equiv I_{SD} = I_{p0} S \exp(-\kappa_p V_{GB}/V_t) [\exp(V_{SB}/V_t) - \exp(V_{DB}/V_t)] \quad (12)$$

The terminal voltages V_{GB} , V_{SB} and V_{DB} are referenced to the substrate (“bulk”). The constant I_0 depends on mobility (μ) and other silicon physical properties. S is the width W to length L ratio (W/L) of the device. A plot for the transfer characteristics of MOS transistors in subthreshold and of a bipolar transistor is shown in Figure 5.

For devices that are biased with $V_{DS} \geq 4 V_t$, (saturation), the drain current for an NMOS transistor is reduced to

$$I_{DS} = S I_{no} \exp((1 - \kappa_n) V_{BS}/V_t) \exp(\kappa_n V_{GS}/V_t) + g_d V_{DS} \quad (13)$$

and for a PMOS transistor

$$I_D \equiv I_{SD} = S I_{p0} \exp(-(1 - \kappa_p) V_{BS}/V_t) \exp(-\kappa_p V_{GS}/V_t) + g_d V_{DS} \quad (14)$$

This shows explicitly the dependence on V_{BS} and the role of the bulk as a *back-gate* that underlies this. This equation, having only the dependence on V_{GS} and V_{BS} , is used for circuit designs where devices operate in saturation as transconductance amplifiers. When the device is operating as a controlled current source (saturation), its characteristics are not ideal. The variation of the current with drain voltage results in an output conductance. For a regular MOS transistor

$$g_d \equiv \frac{\partial I_{ds}}{\partial V_d} = \frac{\partial I_{ds}}{\partial L} \frac{\partial L}{\partial V_d} = \left(-\frac{I_{ds}}{L} \right) \left(-\sqrt{\frac{q\epsilon_{Si}}{2N_A}} \frac{1}{\sqrt{V_{DB}}} \right) \approx \frac{I_{ds}}{V_0} \quad (15)$$

The inverse square-root dependence of the depletion layer width is replaced by a nominal value V_0 which has units of volts and is proportional to L . This voltage is analogous to the Early voltage in bipolars and is determined by experiment.

The parameter κ is defined as

$$\kappa = \frac{C_{ox}^\diamond}{C_{ox}^\diamond + C_{dep}^\diamond} \quad (16)$$

The physical significance of κ is apparent if the observation is made that the oxide capacitance $C_{ox}^\diamond$ and depletion capacitance $C_{dep}^\diamond$ form a capacitive divider between the gate and bulk terminals that determines the surface potential [11]. Lighter doping reduces $C_{dep}^\diamond$, and pushes the divider ratio closer to unity. A larger surface potential also reduces $C_{dep}^\diamond$. The parameter κ takes values between 0.6 and 0.9.

2.1 Translinear Devices

A translinear element is a physical device whose transconductance and current through the device are linearly related, that is, the current is exponentially dependent on the controlling voltage. A two terminal p-n junction (diode), with its exponential I-V characteristics, is a translinear element and used often as an example in circuits [46].

Three-terminal devices are termed “translinear” if the relationship between the current and the controlling voltage is exponential *and* the two terminals across which the controlling voltage is applied exhibit true diode-like behavior, i.e., increasing the voltage on one terminal is exactly equivalent to decreasing the voltage on the other terminal by the same amount. In this case, a loop of such devices consists of voltage drops across pairs of control terminals and we exploit the linear transconductance-current relationship. Bipolar transistors have both properties whereas MOSFETs do not.

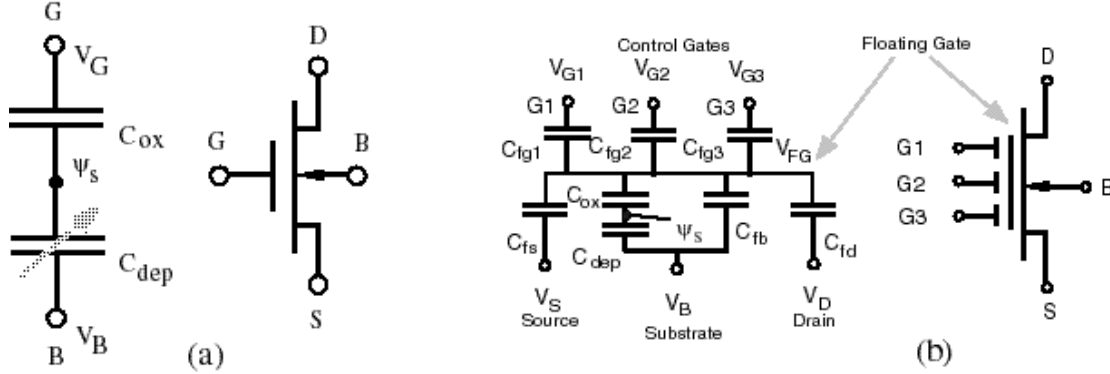


Figure 7: (a) Symbol and simple capacitance model for an NMOS transistor. The current between the drain and source is controlled by the difference between the surface potential ψ_s and the potential at the source terminal. The surface potential ψ_s is set by the potential at the gate and bulk terminals through the capacitive divider between C_{ox} and C_{dep} . (b) Symbol and capacitance model for a floating gate MOS (FGMOS) transistor. The device depicted in this figure has three control gates G_1 , G_2 and G_3 .

The large-signal device model equations for both the bipolar transistor and MOSFET in subthreshold are discussed in Appendix A where the approximations made during their derivations are clearly stated and the symbols are defined. In the active-forward region of operation, the function of a bipolar transistor as a *transconductance amplifier* is captured by the following equation

$$I_C = I_S e^{\frac{V_B - V_E}{V_t}} \quad (17)$$

where $V_t = (kT/q)$ and I_S is defined in Appendix A.

The magnitude of the transconductance from the base is identical to the magnitude of the transconductance from the emitter

$$g_m \equiv \frac{\partial I_C}{\partial V_{BE}} = - \frac{\partial I_C}{\partial V_E} \Big|_{V_B=c} = \frac{\partial I_C}{\partial V_B} \Big|_{V_E=c} = \frac{I_C}{V_t} \quad (18)$$

We now contrast the operation of a bipolar transistor as a translinear element with that of an MOS transistor operating in subthreshold. Much like a bipolar transistor, the MOSFET in subthreshold has exponential voltage current characteristics (see Figure 5). There are, however two fundamental differences between MOSFET and bipolar devices that have implications in the design of translinear circuits.

1. Unlike a bipolar transistor, the current in a MOSFET is controlled by the surface potential, which is capacitively-coupled to the gate (front-gate) and bulk (back-gate) terminals.
2. The MOSFET is symmetric with respect to the source and drain terminals while a bipolar is not.

In summary, the MOS transistor is a *four* terminal device with symmetric drain and source terminals, a result of lossless channel conduction, and an isolated control potential set by one or more control gates. As we will see in subsequent sections, the latter property of the MOS transistor is a mixed blessing when the design of translinear circuits is considered.

It should be pointed out that the voltage difference that controls the current in a MOSFET to yield the translinear behaviour, is the potential difference between the channel surface potential ψ_s and the potential at the source V_S and or drain V_D so that the current between the drain and source for an NMOS is given by

$$I_{DS} \sim \left[\exp \frac{(\psi_s - V_S)}{V_t} - \exp \frac{(\psi_s - V_D)}{V_t} \right] \quad (19)$$

Since the MOS transistor has two “gates” the relationship between $\psi_s(V_B, V_G)$ and the bulk or gate terminal voltages V_B and V_G can be obtained using the simple capacitive divider model depicted in Figure 7. The introduction of the parameter $\kappa \equiv C_{ox}^\diamond / (C_{ox}^\diamond + C_{dep}^\diamond)$ is convenient for modeling the effect of the two gates. Note that κ is a function of the surface potential ψ_s as $C_{dep}^\diamond$ is a function of the applied gate and substrate voltages.

In saturation (i.e. when $V_{DS} \geq 4V_t$), and when the current controlling voltages are referenced to source, Equation 19 can be rewritten as

$$I_{DS} = SI_{no} e^{\left(\frac{C_{dep}^\diamond}{C_{ox}^\diamond + C_{dep}^\diamond}\right) \frac{V_{BS}}{V_t}} e^{\left(\frac{C_{ox}^\diamond}{C_{ox}^\diamond + C_{dep}^\diamond}\right) \frac{V_{GS}}{V_t}} = SI_{no} e^{\frac{(1-\kappa)V_{BS}}{V_t}} e^{\frac{\kappa V_{GS}}{V_t}} \quad (20)$$

Equation 20 can be re-written as a function of dimensionless current quantities i_G and i_B . Each of these currents would correspond to the device current if the surface potential ψ_s could assume the voltage at the gate or bulk terminal. In essence these currents correspond to ideal diode junctions between the source and surface potential weighted by the appropriate capacitive divider ratio. Therefore, the equation for the drain current can be written as

$$I_{DS} \equiv SI_{no} i_G^\kappa i_B^{(1-\kappa)} \quad (21)$$

In subsequent sections, we will see how the latter formulation facilitates the analysis of MOS translinear circuits and an extension of it will be used to analyze FGMOS translinear loops. Since the dimensionless current quantities are related to the surface potential they will be called ψ -currents or *psi-currents*.

The transconductance from the gate is given by:

$$g_m \equiv \left. \frac{\partial I_{DS}}{\partial V_G} \right|_{V_B, V_S=c} = \frac{\kappa I_{DS}}{V_t} \quad (22)$$

and from the local substrate (backgate) terminal:

$$g_{mb} \equiv \left. \frac{\partial I_{DS}}{\partial V_B} \right|_{V_G, V_S=c} = \frac{(1-\kappa) I_{DS}}{V_t} \quad (23)$$

The conductance g_s at the source is given by:

$$g_s \equiv \left. \frac{\partial I_{DS}}{\partial V_S} \right|_{V_G, V_B=c} = \frac{I_{DS}}{V_t} \quad (24)$$

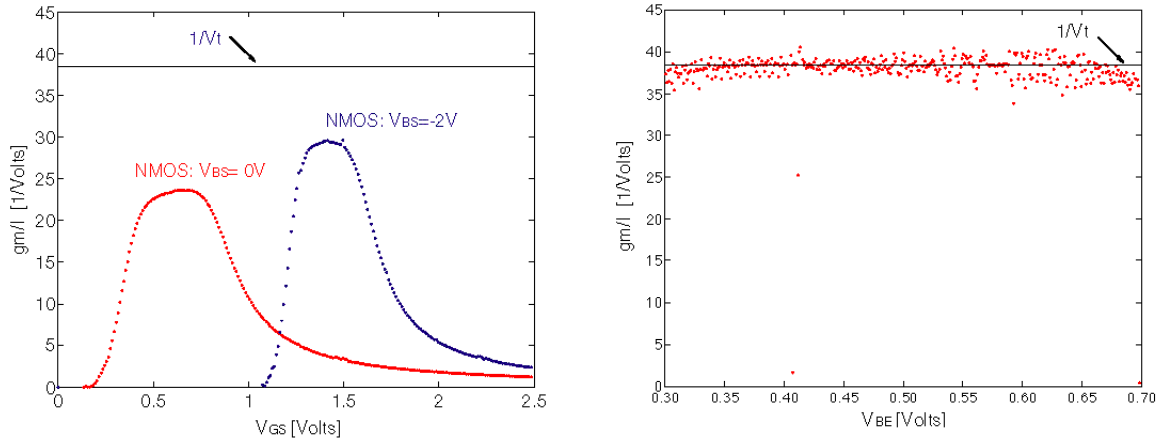


Figure 8: Normalized transconductance curves. The transconductance is computed through numerical differentiation of the data in Figure 2, and subsequent smoothing. (Left) For the MOS transistor; (Right) for the bipolar transistor. The dramatic decrease of the transconductance at low gate-source voltages is attributed to the leakage current.

The transconductances depicted in Equations 22, 23 and 24 are linear functions of the current –to a first order– and hence each MOS transistor in saturation has the equivalent of three different translinear elements. Note, that the source transconductance is equal to the gate transconductance by shorting the local substrate and the source of the transistor ($V_B = V_S$) in which case Equation 20 for the current becomes:

$$I_{DS} = SI_{no} e^{\frac{\kappa_n(V_G - V_S)}{V_t}} \quad (25)$$

In subsequent sections, we will see how shorting of the substrate to the bulk partially circumvents the non-idealities in the translinear properties of MOS transistors and enables the design of near ideal loops.

The translinear properties of the bipolar and MOS transistors in subthreshold are evident in Figure 5. In a logarithmic current scale, the transfer characteristics show linearity with respect to the controlling voltages. Plots of the normalized transconductance (g_m/I) are shown in Figure 8 and demonstrate how the bipolar device is an ideal translinear element while the MOS transistor in subthreshold only approximates it over a limited range.

2.2 The Floating-Gate MOS Transistor

In this subsection the MOS device model is extended to describe Floating-gate MOS transistors. The Floating-Gate MOS transistor (FGMOS) is an important element in the design of micropower neuromorphic systems. It can be used for long term non-volatile storage of neural network model parameters [36, 37, 38, 39] and for compensating parametric variations in device characteristics [40, 41, 42]. The floating gate can also be employed as a summing node to

perform mathematical operations in the charge domain with high linearity [43].

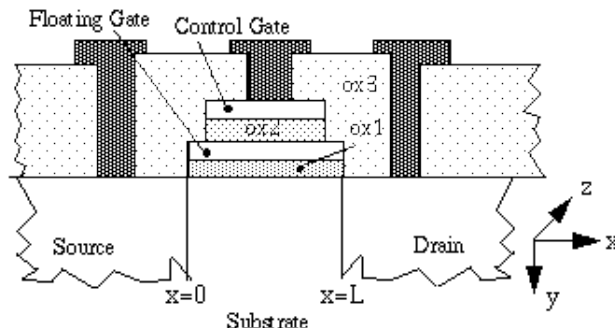


Figure 9: Simplified structure for an FGMOS device. The properties of the different dielectric layers (ox1, ox2, and ox3) that surround the two polysilicon gates depend on the particulars of the manufacturing process. Injecting charge on the floating gate involves current flow through these dielectric layers and thus programming the device is highly process dependent.

In an FGMOS transistor fabricated in a standard double-poly CMOS process, the floating gate is first polysilicon and the control gate is second polysilicon (see Figure 9). The floating gate controls the current in the channel beneath it and is capacitively-coupled to the control gate above it. The voltage on the floating gate is determined by the amount of charge deposited on it, as well as the voltages on the control gate, drain, source and substrate.

Several methods have been reported in the literature for metering charge onto or off from a floating gate. For a literature survey and reference to key papers in the field, please refer to [17].

The best way to incorporate a floating gate transistor as part of a system is in a *closed-loop* configuration. The relaxation phenomena in the different dielectrics [44] introduce hysteresis in the device characteristics. The time constants depend on the details of the manufacturing process. Thus open-loop control of the charge on the floating gate may be difficult.

2.3 Large Signal FGMOS Model

The large signal model of a floating gate MOS transistor can be derived by first obtaining the voltage on the floating gate as a function of the voltages on the nodes that are capacitively-coupled to it. Then, the large signal FGMOS equation is obtained by substituting this voltage for V_{GB} in the equation for the regular MOS transistor. The resulting model is not much different except for two extra terms arising from capacitive-coupling to the source and drain. This effect is significant and cannot be ignored in the design analog circuits using FGMOS transistors.

The voltage on the floating gate of the FGMOS can be calculated with the aid of the simple circuit model in Figure 7(b). If transient oxide interface states are ignored, the voltage on the floating gate is given by:

$$V_{FGB} = \frac{C_{fcg}V_{CGB} + C_{fs}V_{SB} + C_{fd}V_{DB} + C_{ox}\psi_s + Q_{fg}}{C_{sum}} \quad (26)$$

where

$$C_{sum} = C_{fcg} + C_{fs} + C_{fd} + C_{fb} + C_{ox},$$

Q_{fg} is the charge on the floating gate, C_{fcg} is the capacitance between the control gate and the floating gate, C_{fs} and C_{fd} are the capacitances between the floating gate and source and drain, respectively, C_{ox} is the oxide capacitance between the floating gate and the channel, and C_{fb} is the capacitance between the floating gate and the substrate along the edge of the channel. V_{CGB} is the control gate voltage, V_{FGB} is the floating gate voltage, V_{SB} is the source voltage, V_{DB} is the drain voltage, and ψ_s is the channel potential—all referenced to the bulk.

The surface potential ψ_s is eliminated using the kappa approximation, (see Chapter XX) and an expression for V_{FGB} is obtained in terms of the capacitances and the terminal voltages

$$V_{FGB} = \frac{C_{fcg}V_{CGB} + C_{fs}V_{SB} + C_{fd}V_{DB} + C_{ox}(1.5\phi_F + V_{SB} - \kappa V_{GB}^*) + Q_{fg}}{C_{sum}^*} \quad (27)$$

where $C_{sum}^* = C_{sum} - \kappa C_{ox}$.

Equation 27 yields the following equation for the current as a function of the terminal voltages

$$I = S I_{fno} e^{\frac{\zeta_C V_{CGB}}{kT/q}} e^{\frac{\zeta_S V_{SB} + \zeta_D V_{DB}}{kT/q}} \left(e^{\frac{-V_{SB}}{kT/q}} - e^{\frac{-V_{DB}}{kT/q}} \right) \quad (28)$$

where I_{fno} is given by

$$I_{fno} = I_{no} e^{\frac{\kappa Q_{fg} + \kappa C_{ox}(1.5\phi_F + V_{SB} - \kappa V_{GB}^*)}{C_{sum}^* kT/q}} \quad (29)$$

and $\zeta_C \equiv \kappa C_{fcg}/C_{sum}^*$, $\zeta_S \equiv \kappa C_{fs}/C_{sum}^*$, and $\zeta_D \equiv \kappa C_{fd}/C_{sum}^*$. The capacitive divider between the control gate and the floating gate reduces its log-lin slope coefficient ζ_C . Note that the current is also exponentially dependent on the drain and source voltages with log-lin slope coefficients of ζ_S and ζ_D due to capacitive coupling to the fringes of the floating gate.

Ideally, we would like to have $\zeta_C = 1, \zeta_S = \zeta_D = 0$. This may be achieved by adding some extra capacitance between the control and floating gates. The floating gate charge, Q_{fg} , is the only variable that can be changed after the chip is fabricated; it changes I_{fno} and shifts the I-V curve.

The output conductance of an FGMOS transistor has an additional term due to capacitive-coupling between the drain and the floating gate. This term is exponentially related to the drain voltage, with slope coefficient ζ_D . The output conductance is therefore given by

$$g_d^* = g_d + \zeta_D \frac{I_{ds}}{kT/q} \quad (30)$$

The extra term increases the output conductance significantly and therefore cascoding the FGMOS is strongly advised.

3 The Current-Mode Approach

VLSI system considerations and requirements for a high degree of integration make the subthreshold MOS transistor the device of choice for complex neuromorphic systems. Subthreshold operation offers the highest processing rates per unit power. Current-mode (CM) operation yields large dynamic range, simple and elegant implementations of both linear and nonlinear computations, and low power dissipation without sacrificing speed.

By taking advantage of the high subthreshold transconductance per unit current, voltage swings are kept to a few thermal voltages, and reasonable processing bandwidths are achieved. Dynamic power dissipation and supply noise are also reduced as a result of the smaller voltage swings. Smaller voltage swings eliminate the current that is wasted in charging and discharging parasitic capacitances, thereby allow us to use smaller current signals and cut quiescent power dissipation as well. Thus, this approach yields relatively fast analog circuits with power dissipation levels compatible with future trends in system integration. Fast digital circuits can also be designed using source-coupled logic gates and current steering (ECL-like circuits).

The essence of CM VLSI signal processing is signal representation and normalization in terms of a unit current. Since signals are represented by currents, the CM approach enables the design of systems that function over a wide dynamic range. The low end is limited by leakage currents in the junctions and by noise in the system; the high end is limited by degrading transconductance per unit current above threshold. Large dynamic range is essential in neuromorphic systems that receive inputs from real world environments (for example, silicon retinas). The value of the unit current ultimately determines the overall power dissipation and controls the temporal response of the system. Thus, power consumption can be managed and *adaptively* controlled to satisfy the temporal response requirements of the system.

We now introduce a the **current conveyor** concept to facilitate the analysis and synthesis of minimal complexity CM circuits.

3.1 The Current Conveyor

Sedra and Smith originated the notion of a current-conveyor [67]—a hybrid voltage/current three-port device (Figure 10). It is a versatile building block for analog signal processing applications. The current conveyor is characterized by small signal relationships between the voltages and currents at different nodes.

The design methodology presented here employs the concept of the current conveyor as a canonical element with minimal components that can be employed to aid the synthesis of analog VLSI systems. In analogy with the operational amplifier, there is a virtual short for voltage from node Y to node X and a virtual short for current from node X to node Z . Therefore, node X can transmit a current signal to node Y while simultaneously receiving a voltage signal from node Y . This original current-conveyor implementation (CC-I) has been exploited effectively in linear analog LSI circuit design [71]. The high functionality embodied in this very simple—yet elegant—circuit makes it a good candidate for analog VLSI systems where it satisfies the need for large fan-in and large fan-out. In the context of the CM systems described here, node X is used as a *communication* node, node Y as a *control* node, and node Z as an *output* node.

Although the original implementation of the current conveyor concept used five transistors,

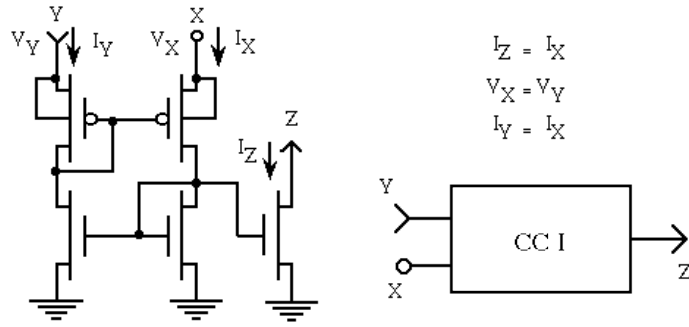


Figure 10: The original Sedra and Smith five-transistor (5T) current-conveyor circuit (CC-I) implemented using MOS transistors. In this conveyor, Y is a hybrid voltage-input/current-output node, X is a hybrid voltage-output/current-input node and Z is a current-output node.

it can in fact be realized with just one (the 1T form). A MOS transistor, in saturation, can transfer a current from its high-conductance source terminal to its low-conductance drain terminal or a voltage from its high-impedance gate terminal to its low-impedance source terminal. These two actions can be exploited to simultaneously achieve a voltage short from gate to source and a current short from source to drain (see Figure 11a). This dual role—obtained with a single device—captures the essence of the current-conveyor. As such, the gate is the control voltage (Y) node, the drain is the output current (Z) node, and the source is the hybrid input-current/output-voltage communication (X) node. In this minimalist implementation, we forsake the redundant current-output at node Y and introduce a nominally constant voltage offset between nodes X and Y .

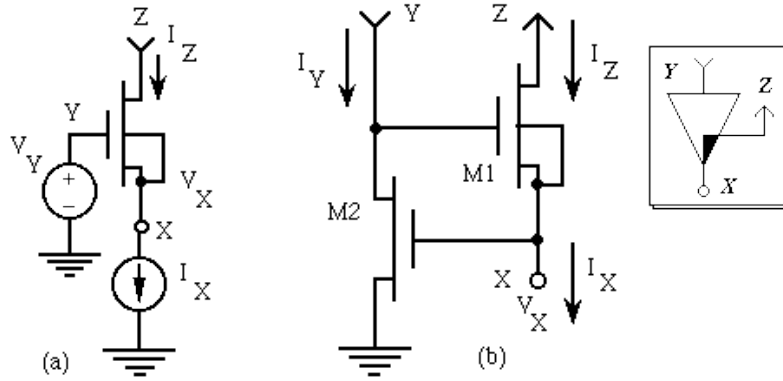


Figure 11: (a) A one-transistor (1T) current conveyor. (b) a two-transistor (2T) current-controlled current conveyor circuit and conveyor symbol. The white triangle represents the voltage buffering action from Y to Z while the in-laid black triangle represents the current buffering action from X to Z .

The large signal behavior for the 1T current-conveyor is described by the following set of equations

$$\begin{aligned} I_Z &= I_X \\ V_X &= \kappa V_Y - \frac{kT}{q} \ln \left(\frac{I_X}{S_1 I_o} \right) \end{aligned} \quad (31)$$

assuming the drain conductance is zero. Note that, unlike Sedra’s CC-I circuit, the 1T conveyor is highly nonlinear. However, the compressive log function does not seriously degrade the voltage following property. The small signal behavior of the 1T conveyor is due to the combined action of a single-transistor voltage follower and a single-transistor current buffer—realized by the same device. That is, if the current source at node X has an output conductance g_i , then the output conductance at node Z is approximately g_i/A_{v_0} and the input conductance at node X is g_m , where A_{v_0} is the intrinsic gain of the device and g_m is its transconductance.

A two-transistor (2T) current-controlled current conveyor is shown in Figure 11b. This circuit has a hybrid current-input/voltage-output communication node X , a current-input control node Y , and a current-output node Z . The voltage V_X at the communication node X is determined by the current supplied to node Y —instead of using a control voltage. The current-controlled conveyor’s large signal behavior is described by the following equations

$$\begin{aligned} I_Z &= I_X \\ V_X &= \frac{kT}{q} \frac{1}{\kappa_2} \ln \left(\frac{I_Y}{S_2 I_o} \right) \\ V_Y &= \frac{kT}{q} \frac{1}{\kappa_2 \kappa_1} \ln \left(\frac{I_Y}{S_2 I_o} \right) + \frac{kT}{q} \frac{1}{\kappa_1} \ln \left(\frac{I_X}{S_1 I_o} \right) \end{aligned} \quad (32)$$

The voltage buffering between nodes Y and X has been replaced with a generalized transimpedance (M_2) that generates a voltage V_X proportional to $\log(I_Y)$. As will be seen in a later section, in a current-mode circuit design style where all signals are implicitly represented by “log I” voltages, this nonlinearity is transparent. The log-ing action is achieved by diode-connecting M_2 —using the voltage-following action of M_1 from node Y to node X . M_1 sets V_X so as to make M_2 ’s current equal to I_Y . M_2 inverts and amplifies small changes in V_X to generate the requisite voltage at node Y .

Compared to the 1T conveyor, the small-signal conductance at the current-output node (Z) is lower by a factor of A_{v_0} and the small-signal conductance at the communication node (X) is increased by the same factor. More specifically, the conductance seen at node Z is approximately $g_i/(A_{v_1} A_{v_2})$, i.e. the buffering action of M_1 is improved by the gain A_{v_2} of M_2 . The conductance seen at node X is approximately $(A_{v_2} g_{m_1})$, i.e. the source conductance of M_1 times M_2 ’s gain.

The 2T current-controlled conveyor was used in a two-way communication scheme for a current-mode, clamped bit-line, associative memory design [15, 69]. In that application, the dual role of the 2T current-controlled current conveyor was fully exploited to simultaneously fan-in currents and fan-out voltage over the same wire. Säckinger independently proposed this arrangement to increase the output impedance of an MOS transistor current source, as we have shown here. He used this circuit in a high-performance op-amp design and dubbed it the

regulated cascode [70]. Indeed, the negative feedback arrangement used in the 2T conveyor is not new and is well-known to veteran analog circuit designers. One of its most elegant applications is the three-transistor Wilson current mirror [72].

In the next section, we introduce the **translinear principle** which provides a powerful design and analysis framework for linear and non-linear computational circuits in analog VLSI.

4 Translinear Principle

The *Translinear* principle [21] exploits the exponential current-voltage non-linearity in semiconductor devices and offers a powerful circuit analysis and synthesis [45] framework. Originally formulated for bipolar transistors [21], this principle enables the design of analog circuits that perform complex computations in the current-domain including products, quotients, and power terms with fixed exponents [21, 45]. Translinear circuits perform these computations without using differential voltage signals and are amenable to device-level circuit design methodology.

In this section, we provide an overview of nano-power translinear circuit design using MOS transistors operating in the subthreshold region. We contrast the bipolar and MOS subthreshold characteristics and extend the translinear principle to the subthreshold MOS ohmic region through a drain/source current decomposition. A front/back-gate current decomposition is adopted; this facilitates the analysis of translinear loops, including multiple input floating gate FGMOS transistors. Circuit examples drawn from working systems designed and fabricated in standard digital CMOS oriented process are used as vehicles to illustrate key design considerations, systematic analysis procedures, and limitations imposed by the structure and physics of MOS transistors. This performs phototransduction and amplification, edge enhancement and local gain control at the pixel level.

Most of the work on translinear circuits to date, use bipolar transistors and the emphasis is on high precision and high speed. One fascinating aspect of translinear circuits is their insensitivity to isothermal temperature variations, though the currents in its constitutive elements (the transistors) are exponentially dependent on temperature. The effect of small local variations in fabrication parameters can also be shown to be temperature independent. An excellent up-to-date overview of translinear current-mode analog circuits using bipolar transistors can be found in [46].

The increased commercial interest in analog CMOS LSI and VLSI has renewed interest in the translinear principle for MOS circuit design. A generalized form of the translinear principle was recently proposed for MOS operating above threshold [47]; this extension however does not follow the original definition of a translinear circuit [21]. This extension is simply a design principle that exploits conservation of energy (KVL) around circuit loops which have specific topological properties. A novel class of translinear circuits that employs multiple input gates, with floating gate MOS transistors in subthreshold, has been recently proposed and experimentally demonstrated [49]. Another exciting research area that emerged the last few years, is the synthesis of analog VLSI for sensory information processing which is the focus of discussion in this section.

In the next two subsections we discuss translinear circuits that employ translinear elements, both MOS operating in subthreshold and bipolar transistors. We follow the convention pro-

posed by Barrie Gilbert in [46] and make a distinction between a Translinear Loop (TL) and a Translinear Network (TN). We begin with Translinear loops.

4.1 Translinear Loops

In “strictly” TLs the translinear principle [21] can be stated as follows:

In a closed loop containing an equal number of oppositely connected translinear elements, the product of the current densities in the elements connected in the ClockWise (CW) direction is equal to the corresponding product for elements connected in the Counter ClockWise (CCW) direction.

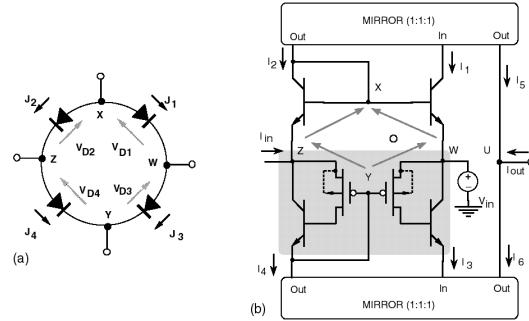


Figure 12: (a) A Translinear loop using ideal p-n junctions (diodes). (b) The translinear loop of the circuit (left), implemented using composite bipolar and subthreshold MOS transistors. The loop is employed in a current conveyor configuration where the bidirectional output current I_{out} equals to the bidirectional current I_{in} .

As an example let us consider the circuit of Figure 12(a) consisting of four *ideal* diodes in the loop X–Z–Y–W–X. Following the translinear principle, we can write:

$$\frac{\prod_{CCW} J}{\prod_{CW} J} = 1 \quad \text{or} \quad J_2 J_4 = J_1 J_3 \quad (33)$$

Note that the translinear principle is derived by beginning with Kirchoff’s voltage law or the principle of conservation of energy, so that:

$$\sum_{i=1}^{i=N/2} V_{D(2i-1)} - \sum_{i=1}^{i=N/2} V_{D(2i)} = 0 \quad (34)$$

Equation 33 follows from Equation 34 if the voltages are summed around loops of translinear devices.

In a circuit graph composed of two terminal elements such as ideal diodes (see Figure 12(a)), there is a direct relationship between the voltage difference among each pair of nodes transversed by the translinear device, and the current in the arc that joins the nodes. This is a consequence of having the voltage nodes that control the current be the same as the current–output nodes

of the device. In practical systems, the ideal diodes in Figure 12(a) would correspond to base-emitter junctions of bipolar transistors with shorted collector-base terminals.

Analogous behaviour can be obtained using translinear three terminal devices such as bipolar transistors, MOS transistors in subthreshold, or any other device that yields diode-like characteristics. However, in three terminal devices, the diode-control nodes in the circuit need not correspond to the current path. In bipolar transistors the diode control nodes are available and thus they can be used explicitly in constraint equations such as Equation 34. This is not true for MOS transistors! As we have seen already, one of the diode control nodes, (namely the node corresponding to the surface potential ψ_s) is not *directly* accessible. The situation becomes even more complex in MOS transistors with a floating gate (FGMOS) (see Figure 9) coupled to one or more controlling gates (see [48, 49] and references therein). At first sight, the floating gate appears to make the situation worse, but actually it opens the possibility for a new class of translinear circuits proposed and experimentally demonstrated recently by Minch et.al. [49].

Essentially the physical structure of FGMOS transistors offer an extra degree of freedom which can be exploited systematically through another set of constraint equations of the form:

$$V_{FGi} = \frac{Q_{FGi}}{C_{Ti}} + \sum_{j=1}^{j=N} \Lambda_{ij} V_{Gj} \quad (35)$$

where V_{FGi} , Q_{FGi} are the floating gate voltage and charge on the (*i*th) transistor and V_{Gij} is the voltage of the (*j*th) control gate. The total capacitance seen in the floating gate is C_{Ti} and Λ_{ij} is a design parameter that depends on the ratio of the control gate to floating gate capacitance, i.e. $\Lambda_{ij} \equiv C_{fgj}/C_{Ti}$. The details of a systematic analysis procedure for FGMOS translinear circuits can be found in [49].

4.1.1 Analysis of Translinear Circuits with MOS Transistors in Saturation

The current mirror is a trivial example of a translinear circuit. It has a single loop with two translinear elements, one CCW and the other CW.

Two current mirrors implemented with complementary devices and connected back-to-back yield the circuit shown in Figure 12(b). This loop includes four three-terminal devices and corresponds to the ideal diode example of Figure 12(a). The circuit can be readily recognized as a BiCMOS implementation of an AB stage in a digital oriented CMOS process where only one type (NPN) of bipolar transistors is available [52]. A composite structure made of an MOS in subthreshold and an NPN bipolar yields a pseudo-PNP device with good driving capabilities. Translinear loops using both PNP and NPN bipolar transistors were first studied by Fabre [50].

Applying the translinear principle to the loop X–Z–Y–W–X of Figure 12(b) yields the following constraint equation for the currents in the circuit:

$$I_2 I_4 = I_1 I_3 \quad (36)$$

This classical four junction loop can be combined with two current mirrors to implement a current conveyor [51] where $I_{out} = I_{in}$ and $V_Z = V_{in}$.

Our second example is the MOS transistor one-quadrant multiply-divide circuit shown in Figure 13(a). A large number of these CMOS multipliers have been employed in the implementation of a correlation-based motion-sensitive silicon retina [53].

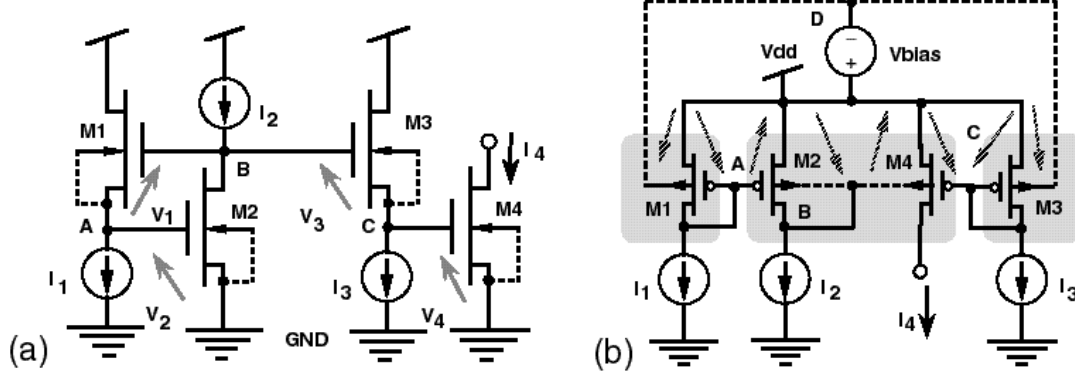


Figure 13: (a) A translinear circuit that performs one-quadrant normalized multiplication. I_1 , I_2 and I_3 are the inputs and I_4 is the output. (b) A four transistor translinear circuit that performs a one-quadrant normalized multiplication and exploits the back-gate in an MOS transistor. Device pairs $M1$, $M3$ and $M2, M4$ share local substrate terminals (in this case n-wells). I_1 , I_2 and I_3 are the inputs and I_4 is the output.

Applying the translinear principle to the loop GND–A–B–C–GND, we find a total of four equivalent diode junctions and obtain

$$I_1 I_2 = I_3 I_4 \quad \text{or} \quad I_4 = \frac{I_1 I_2}{I_3} \quad (37)$$

The above relationship can also be derived by summing the voltages around the loop (conservation of energy)

$$V_1 + V_2 - V_3 - V_4 = 0$$

Replacing the gate-source voltages for $M1, M2, M3, M4$ with their respective drain-source currents using Eq. (25) (assuming all devices are in saturation, have $V_{SB} = 0$, have negligible drain conductance, have identical κ , have identical I_0 and geometry S), we obtain

$$\frac{kT}{\kappa q} \left(\ln \left(\frac{I_1}{S I_0} \right) + \ln \left(\frac{I_2}{S I_0} \right) - \ln \left(\frac{I_3}{S I_0} \right) - \ln \left(\frac{I_4}{S I_0} \right) \right) = 0$$

or

$$\ln \left(\frac{I_1 I_2}{S I_0} \right) = \ln \left(\frac{I_4 I_3}{S I_0} \right)$$

from which Eq. (37) readily follows. Note that the assumption of identical κ holds true to a first order because $V_{SB} = 0$ and the gate of all transistors are within a few hundred millivolts from each other.

Yet another way of viewing the function of this circuit is that of a *log-antilog* block. Transistors $M1$ and $M2$ do the *log-ing*, $M4$ does the *antilog-ing* and $M3$ is a level shifter.

Another single quadrant multiplier is shown in Figure 13(b). This circuit was proposed and its function experimentally demonstrated in [58]. The operation of the circuit can be understood by noting that a single transistor (M_4) can perform a single quadrant multiplication because the voltages on the gate and bulk control the current in a multiplicative fashion (see Equation 20). Since in subthreshold the transistors saturate at only a few V_t of drain source voltage, the bulk terminal of the device can be connected to the drain without turning on the bulk-source junction.

An expression for the output current I_4 can be obtained by applying the translinear principle around the four loops (Vdd-A-Vdd), (Vdd-B-Vdd), (Vdd-C-Vdd), (Vdd-D-Vdd) to obtain the following equations for the *psi-currents* introduced in Equation 21.

$$\begin{aligned} i_{G1} &= i_{G2} \\ i_{B2} &= i_{B4} \\ i_{G4} &= i_{G3} \\ i_{B3} &= i_{B1} \end{aligned} \tag{38}$$

The actual currents in the four MOSFETs M_1, M_2, M_3, M_4 , can be written as a function of the *psi-currents*

$$\begin{aligned} I_1 &= I_{DS1} = SI_{no} i_{G1}^{\kappa_1} i_{B1}^{(1-\kappa_1)} \\ I_2 &= I_{DS2} = SI_{no} i_{G2}^{\kappa_2} i_{B2}^{(1-\kappa_2)} \\ I_3 &= I_{DS3} = SI_{no} i_{G3}^{\kappa_3} i_{B3}^{(1-\kappa_3)} \\ I_4 &= I_{DS4} = SI_{no} i_{G4}^{\kappa_4} i_{B4}^{(1-\kappa_4)} \end{aligned} \tag{39}$$

where the devices have been assumed to have the same S and I_{no} . If now the assumption is made that $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4$, Equations 38 and 39 yield the following expression for the output current I_4 in terms of the input currents I_1 , I_2 and I_3

$$I_1 I_4 = I_2 I_3 \quad \text{or} \quad I_4 = \frac{I_2 I_3}{I_1} \tag{40}$$

In the original implementation [58], it was suggested that, to improve accuracy, the voltage on the local substrate (n-well) of devices M_3 and M_1 should be set at a value close to that of node B , the local substrate of devices M_2 and M_4 . This is indeed necessary, as the bulk voltage determines κ of all transistors, which was assumed to be the same for all transistors. Another implicit assumption here is that the gate voltage is approximately the same for all transistors.

Our next example addresses the problem of converting a bidirectional current to two unidirectional currents which is the equivalent to a current-mode half-wave rectification. A translinear circuit that computes this nonlinear function is shown in Fig. 14(a). The bidirectional current I_{BD} is steered through transistor M_3 when $I_{BD} > 0$ and through transistors $M_{4,5}$ when $I_{BD} < 0$. Concentrating on transistors $M_{1,2,3,4}$, we identify a loop (VDD-A-B-C-VDD) and apply the translinear principle to yield

$$I_B I_B = I_1 (I_1 - I_{BD})$$

A second equation is obtained from the trivial loop VDD-C-VDD

$$I_2 = I_1 - I_{BD}$$

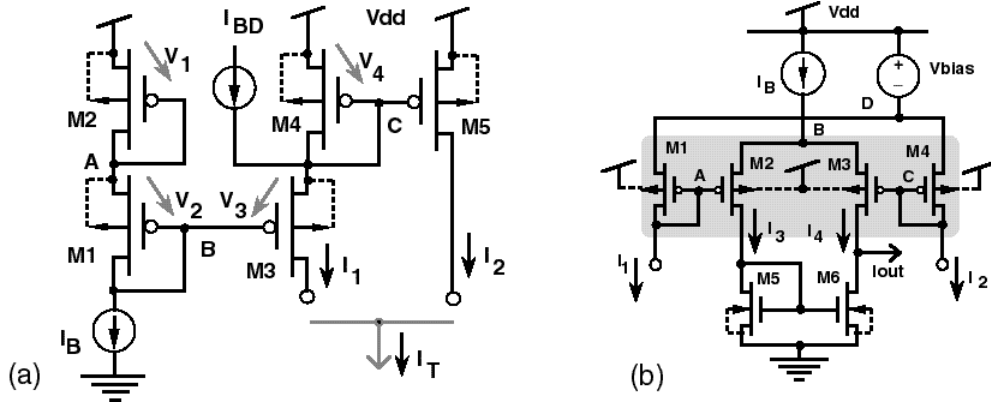


Figure 14: Circuit that converts a bidirectional current on a single wire into two unidirectional currents on separate wires. This is a current-mode absolute value circuit. The sign of the bidirectional input current is assumed to be positive when it adds positive charge to node C . The bidirectional current I_{BD} is the input, the unidirectional currents I_1 and I_2 are the outputs, and I_B sets the operating point of the circuit. (b) A translinear circuit that computes the normalized difference of two current signals. I_1 and I_2 are the inputs and the bidirectional current I_{out} is the output normalized to $I_1 + I_2$.

which can also be obtained by observing that transistors M_4 and M_5 form a simple current mirror.

These equations may be solved for I_1 and I_2 in terms of I_B and I_{BD}

$$\begin{aligned} I_1 &= \frac{1}{2} \left(I_{BD} + \sqrt{4I_B^2 + I_{BD}^2} \right) \\ I_2 &= \frac{1}{2} \left(-I_{BD} + \sqrt{4I_B^2 + I_{BD}^2} \right) \end{aligned} \quad (41)$$

Which shows that $I_1 \simeq |I_{BD}|$ and $I_2 \simeq 0$ when $I_{BD} \gg I_B$ and vice versa. The absolute value is obtained by connecting the two output wires together in which case

$$I_T \equiv I_1 + I_2 \simeq |I_{BD}| \quad \text{if } I_{BD} \gg I_B$$

This circuit has been employed in a CMOS integration of an autoadaptive linear recursive network for the separation of sources [23].

The next translinear circuit performs a current-ratio computation. This functional block is part of the readout amplifier in an analog VLSI system that integrates monolithically a one dimensional array of photodiodes and selective polarization film to form a *polarization contrast* retina [54].

The simple translinear circuit in Figure 14(b) is excellent for rescaling differential current signals and thus computing the contrast. I_1 and I_2 represent currents from two selected photodiodes. The heart of the computation circuit will be recognized as a Gilbert gain-cell [46] implemented in subthreshold MOS.

The analysis of this circuit is typical for translinear circuits that involve differential current signals. Application of the translinear principle around the loop A-B-C-D-A yields

$$I_1 I_4 = I_3 I_2 \Rightarrow \frac{I_3}{I_1} = \frac{I_4}{I_2}$$

Using basic algebra,

$$\frac{I_3 - I_4}{I_1 - I_2} = \frac{I_3 + I_4}{I_1 + I_2}$$

and

$$\Delta I^* = \frac{I_B}{I_{in}} \Delta I$$

where $\Delta I \equiv I_1 - I_2$, $I_{in} \equiv I_1 + I_2$, and similarly for $\Delta I^* \equiv I_3 - I_4$, $I_B \equiv I_3 + I_4$. The differential output current ΔI^* is a scaled version of the differential input current ΔI . The voltage between node *B* and *Vdd* should be such that the current source I_B stays in saturation.

The mirror composed of transistors *M5* and *M6* converts the unidirectional differential signal ΔI to the bidirectional signal I_{out} so that

$$I_{out} \equiv -I_B \frac{\Delta I}{I_{in}}$$

4.1.2 Analysis of Circuits with MOS Transistors in the Ohmic Regime

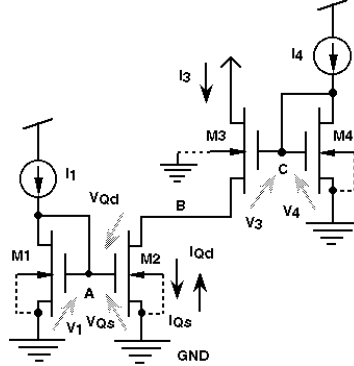


Figure 15: A translinear circuit that employs subthreshold MOS transistors in saturation and ohmic regime and computes the product of two input currents I_1 and I_2 normalized to $I_1 + I_2$.

In this subsection, we extend the translinear principle to subthreshold MOS transistors operating in the ohmic region. In Appendix A, (see Fig. 4), we show how the source-drain current of a MOS transistor can be decomposed into a source component I_{Q_s} and a drain component I_{Q_d} , and that these components superimpose linearly to yield the actual current $I_{SD} \equiv I_{Q_s} - I_{Q_d}$.

In the ohmic region, these components are comparable. Decomposition and linear superposition may be used to exploit the intrinsic translinearity of the gate-source and gate-drain “junctions.” This is the basis for extending the translinear principle to the ohmic region. On

the other hand, in the saturation region, we can exploit the translinearity of the gate-source “junction” directly because the drain component is essentially zero and decomposition is of no consequence.

The translinear circuits based on subthreshold ohmic operation are only possible because of the symmetry between drain and source operation of an MOS transistor. One could argue that decomposition is also possible with bipolar devices. However, while the difference of two exponentials is the *exact* form for MOS devices, it is only an *approximation* for bipolars, due to the fact that the forward and reverse current gains of the device never reach unity [79, 80] (see Equations 62 in Appendix A). This distinction is a fundamental and important difference between MOS and bipolar transistors arising from lossless transport in a MOS channel versus lossy transport in bipolars due to recombination in the base. It is possible to use CMOS compatible lateral bipolar transistors as symmetric devices [55] but at the expense of a large base current that increases the power dissipation in the system.

To demonstrate the application of the translinear principle to circuits that include MOS transistors in the ohmic regime, consider the one-quadrant current-correlator circuit in Fig. 15. Transistor M_2 operates in the ohmic region. Proper circuit operation requires that the output voltage is high enough to keep M_3 in saturation. This circuit was first introduced by Delbrück [56] and later incorporated in a larger circuit that implements the non-linear Hebbian learning rule in an auto-adaptive network [57, 59] and in a micropower autocorrelation system [60].

An expression relating the output current, I_3 , to the input currents, I_1 and I_4 , can be derived by treating the source-gate and the drain-gate “junctions” of the ohmic device as separate translinear elements and applying the translinear principle. For the two loops formed by nodes GND-A-GND and GND-A-B-C-GND in Fig. 15, we obtain

$$\begin{aligned} I_1 &= I_{Q_s} \\ I_3 I_{Q_s} &= I_4 I_{Q_d} \end{aligned} \tag{42}$$

In writing Eq. (42), we have tacitly assumed that the source-drain current of the MOS transistor can be decomposed into a source component I_{Q_s} and a drain component I_{Q_d} —controlled by their respective “junction” voltages V_{Q_s} and V_{Q_d} . These opposing components superimpose linearly to give the actual current passed by M_2 , i.e.,

$$I_{DS2} = I_{Q_s} - I_{Q_d}$$

Furthermore, both M_2 and M_3 pass the same current and therefore

$$I_3 = I_{Q_s} - I_{Q_d} \tag{43}$$

Substituting for I_{Q_s} and I_{Q_d} from Eqs. (42) and (43), the output current is given by

$$I_3 = \frac{I_1 I_4}{I_1 + I_4} \tag{44}$$

4.2 Translinear Networks

In the previous section we have discussed “strictly” translinear loops (TL). Translinear networks [46] differ from translinear loops in that they contain independent voltage sources and the following equation can be employed in their analysis

$$\frac{\prod_{CCW} J}{\prod_{CW} J} = \mathcal{G} e^{E/V_t} \quad \text{or} \quad J_2 J_4 \dots J_N / J_1 J_3 \dots J_{N-1} = \mathcal{G} e^{E/V_t} \quad (45)$$

where E is the independent voltage source and $\mathcal{G}$ is a constant coefficient that lumps device design and fabrication parameters. The above extension to the translinear principle was proposed by Hart [61].

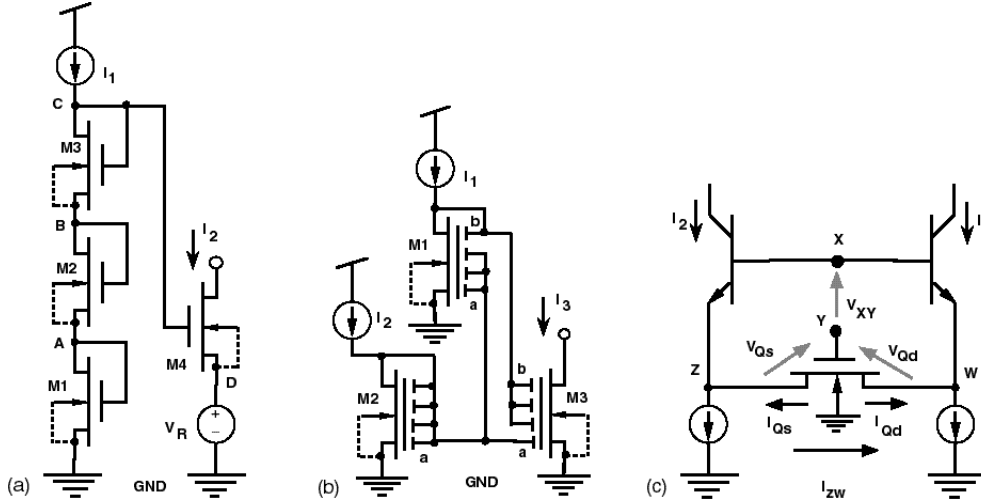


Figure 16: Translinear networks in subthreshold MOS. (a) Current-mode circuit that implements a normalized cubic non-linearity. I_1 is the input, I_2 is the output and the voltage source V_R normalizes the result. (b) Another translinear circuit using FGMOS transistors to compute the ratio of a cubic to square functions. I_1 and I_2 are the inputs and I_3 is the output. (c) BiCMOS translinear network that exploits the MOS subthreshold ohmic characteristics.

We begin the discussion of TNs using a simple circuit that has the topology of a current mirror and incorporates a voltage source in the loop (see Figure 16(a)). If the input current is I_1 , the output current is I_2 , and V_R is a constant voltage source, application of the translinear principle around the loop (GND-A-B-C-D-GND) yields

$$I_2 = \frac{I_1^3}{\mathcal{G} e^{V_R/V_t}} \quad (46)$$

The voltage source is necessary for circuit operation and it normalizes appropriate the output

current. This circuit has been employed in a small system that implements the Herault-Jutten independent component analyzer [23].

An FGMOS-based circuit that has the same functionality as the circuit in Figure 16(a) is shown in Figure 16(b). We begin the analysis of the circuit by noting that the current in the channel of an FGMOS is controlled by multiple gates that can be thought as extensions to the front gate of the transistor. As such, Equation 21 can be re-written for an N-input NMOS transistor as

$$I_{DS} \equiv S S_Q I_{no} i_B^{(1-\kappa)} \prod_{j=1}^{j=N} i_{Gj}^{\kappa/N} \quad (47)$$

where it has been assumed that all N gates of the i -th transistor have the same strength, i.e. have the same coupling capacitance to the floating gate. The charge Q_{FG} on the floating gate is incorporated through a geometry related multiplicative constant S_Q so that when the charge Q_{FG} is zero, $S_Q = 1$.

An expression for the output current I_4 can be obtained by applying the translinear principle around three loops that include the floating gates and source of transistor M_3 together with the floating gate nodes and sources of the other transistors. When the three loops are traversed, the following equations for *psi-currents* are obtained

$$\begin{aligned} i_{3a} &= i_{2a} \\ i_{3a} &= i_{1a} \\ i_{3b} &= i_{1b} \end{aligned} \quad (48)$$

We have adopted a notation where, for example, i_{3a} denotes the *psi-current* in device 3 controlled by the voltage on its gate a . Using Equation 47, the current at the source of M_1, M_2, M_3 , can be expressed as functions of *psi-currents* so that

$$\begin{aligned} I_1 &= I_{DS1} = S S_{Q1} I_{no} i_{1a}^{3\kappa_1/4} i_{1b}^{\kappa_1/4} \\ I_2 &= I_{DS2} = S S_{Q2} I_{no} i_{2a}^{\kappa_2} \\ I_3 &= I_{DS3} = S S_{Q3} I_{no} i_{3a}^{\kappa_3/4} i_{3b}^{3\kappa_3/4} \end{aligned} \quad (49)$$

where the devices are assumed to have the same S and I_{no} ; the back-gate contribution to the current in each device is eliminated as all transistors have the source shorted to the substrate. Now, by making the assumption that $\kappa_1 = \kappa_2 = \kappa_3 = \kappa_4$, and no charge on the floating gates, Equations 48 and 49 yield the following expression for the output current I_3 in terms of the input currents I_1 and I_2

$$I_3 = \frac{I_1^3}{I_2^2} \quad (50)$$

The assumption of equal κ is reasonable so long the voltage on the floating gate is such that all devices stay in subthreshold. An alternative way of obtaining a functional description of the circuit can be found in the paper by Minch et. al [49].

A TN that incorporates bipolar transistors, an MOS in subthreshold, and an independent voltage source V_{XY} is shown in Figure 16(c). We will use Equation 45 to derive the relationship

between I_1 , I_2 , I_{ZW} and V_{XY} . The MOS transistor is assumed to be ideal with $\kappa = 1$. A discussion of networks with non-ideal devices will be done in the next section.

A ratio relationship between I_{ZW} and $I_1 - I_2$ can be derived by employing Equation 45 applied to the two loops (X-Z-Y-X) and (X-Y-W-X) to yield the following equations

$$\begin{aligned} I_{Q_s} &= I_2 \mathcal{G} e^{-V_{XY}/V_t} \\ I_{Q_d} &= I_1 \mathcal{G} e^{-V_{XY}/V_t} \end{aligned} \quad (51)$$

Using the adopted conventions for current decomposition in NMOS transistors (see Figure 4), I_{Q_s} and I_{Q_d} ,

$$I_{ZW} = I_{Q_d} - I_{Q_s}$$

together with Equation 51 we obtain the following expression that relates the currents in the circuit.

$$I_{ZW} = (I_1 - I_2) \mathcal{G} e^{-V_{XY}/V_t} \quad (52)$$

It is immediately apparent that the current ratio $I_{ZW}/(I_1 - I_2)$ can be controlled both by a fixed parameter ($\mathcal{G}$) that is designed prior to the fabrication of the circuit and a variable quantity (e^{-V_{XY}/V_t}) that can be programmed (post-fabrication) during circuit/system operation. This property will be utilized in the design of **linear** MOS transistor-only spatial averaging networks.

4.2.1 Translinear Spatial Averaging Networks

Often, models of neural computation necessitate the realization of spatial averaging networks [11]. To demonstrate the analogies between linear and translinear networks as well as their subtle and important differences, we begin with networks that employ linear conductances, voltages and currents and contrast them with translinear current-mode [17] networks.



Figure 17: Building blocks for linear loaded networks. Using segments that employ ideal (left) linear and (right) non-linear elements.

A voltage-mode circuit model for a loaded network is shown in Figure 17(left) for which

$$I_{PQ} = (G_1/G_2)(I_Q - I_P)$$

This is a lumped parameter model where G_1 and G_2 correspond to resistances per unit length. The voltages on nodes P and Q , referenced to ground, represent the state of the network and can be read out using a differential amplifier with the negative input grounded.

The equivalent circuit using idealized non-linear conductances is shown in Figure 17(right). The difference in currents through the diodes D_1 and D_2 are linearly related to the current through the *diffusor* MOS transistor.¹ This relationship can be derived from Equation 11 describing subthreshold conduction, and the ideal diode characteristics where $I_D = I_S \exp[V_D/V_t]$. An expression can be derived for the current I_{PQ} in terms of the currents I_P and I_Q , the reference voltage V_r and the bias voltage V_C , when diodes are replaced by transistors

$$I_{PQ} = \left(\frac{SI_{n0}}{I_S} \right) \exp \left[\frac{\kappa_n V_C - V_r}{V_t} \right] (I_Q - I_P) \quad (53)$$

The current I_{n0} and S is the zero intercept current and geometry factor respectively for the diffusor transistor M_h . I_S is the reverse saturation current for the diode that is assumed to be ideal. The currents in these circuits are identical if

$$\frac{G_1}{G_2} = \left(\frac{SI_{n0}}{I_S} \right) \exp \left[\frac{\kappa_n V_C - V_r}{V_t} \right]$$

Increasing V_C or reducing V_r has the same effect as increasing G_1 or reducing G_2 . The state of this network is represented by the charge at the nodes P and Q . Since the anode of a diode is the reference level (zero negative charge), the currents I_P and I_Q represent the result. Unfortunately, the anode of a diode or a diode connected transistor is not a good current source.

When diodes are not explicitly available in the process, diode connected PMOS or NMOS transistors can be used as shown in Figure 18. When the loads are PMOS, the current I_{PQ} is given in terms of voltages normalized to (kT/q)

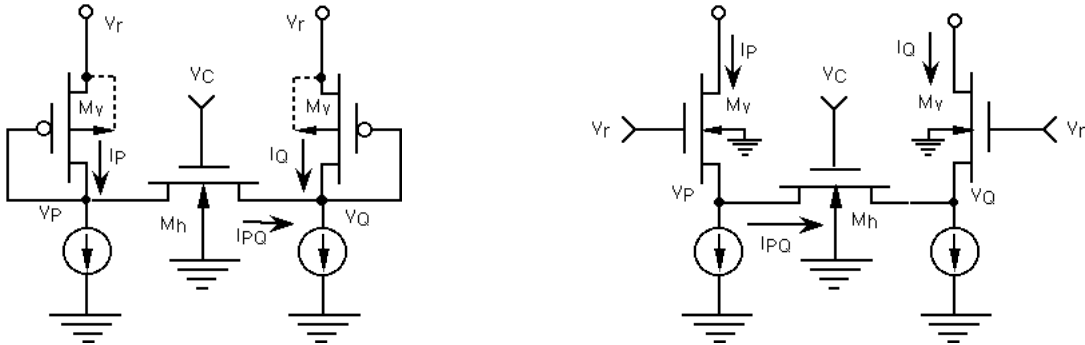


Figure 18: Current-mode building blocks for linear loaded networks using (top) PMOS transistor implementation, (bottom) NMOS single transistor current-conveyor implementation.

¹The diffusor is a term adopted in [24] to describe the exploitation of diffusion transport in MOS transistors to spread signals in a manner analogous to gap junctions between neural cells.

The parameter $\lambda \equiv \left(\frac{S_h}{S_v}\right) \exp(\kappa_n v_C - \kappa_n v_r)$ is the cost associated with the derivative energy—relative to the squared-error of the fit.

The diffusive network in Fig. 19 was recently described in terms of “pseudo-conductances” [63]. Here we use the charge/current-based formulation first proposed in [24] to explain its behaviour. This current-mode approach relies on an intuitive understanding of the device physics and yielded the insight which enabled us to extend the translinear principle to subthreshold MOS transistors in the ohmic region as well as the decomposition of the current into dimensionless components corresponding to ideal junction. We now have a comprehensive *current-mode* approach for analyzing subthreshold MOS circuits. The essence of this approach is the representation of variables and parameters by charge, current, and diffusivity. Voltages and conductances are not used explicitly.

Bult and Geelen proposed an identical network for linear current division above-threshold and used it in a digitally-controlled attenuator [64]; they also analyzed its subthreshold behavior. However, they stipulate that all gate voltages must be identical and control the division by manipulating the geometrical factor W/L of the devices. We have shown here, and previously in [24], that this constraint may be relaxed in subthreshold without disrupting linear operation. This is a real bonus because it allows us to modify the divider ratio or space constant of the network after the chip is fabricated by varying $(V_c - V_r)$. Tartagni et al. have demonstrated a current-mode centroid network [65] using subthreshold MOS devices whose operation is described by the current division principle.

4.3 A general result for MOS translinear loops

Three of the circuits discussed in the previous subsection, namely the translinear multiplier of Figure 13(b), the MOS implementation of the Gilbert gain stage in Figure 14(b), and the current correlator (Fig. 15) have been experimentally shown to exhibit near “exact” translinear behaviour even though they are built from MOS transistors and they do not have their source connected to the local substrate.

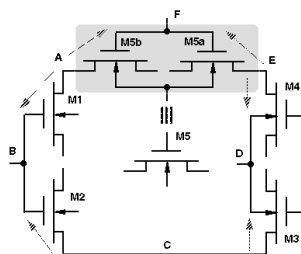


Figure 20: Translinear loop composed of five MOS transistors in subthreshold. All devices are in saturation except device M_5 which is in the ohmic regime and therefore can be decomposed as two devices in saturation, back to back sharing same gate and substrate.

A recent result by Eric Vittoz [30] can be employed to partially explain this rather surprising behaviour. He considers translinear loops constructed from MOS transistors in subthreshold saturation with common substrate connection (similar to the one shown in Figure 20). If

the pairing of transistors in the CW and CCW direction is such that they have their gates connected to gates and sources connected to sources and they are alternated (much like even-numbered and odd-numbered devices in Figure 20), Vittoz shows that the translinear loop does not suffer from the MOS transistor non-ideal translinear behaviour. He notes also that loops containing transistors in the ohmic regime can also be included in this formulation as they can be decomposed as two parallel connected saturated devices sharing common gate and common substrate (see Figure 4.)

However, to account for the near “exact” operation of the multiplier in Figure 13(b), Vittoz’s argument must be extended to include loops that go through the back gate of the MOS transistors as illustrated in Figure 20. The common substrate restriction can thus be removed and replaced by a local substrate connection, and the result still holds true. In a standard CMOS process, this will of course be possible only for one type of devices.

Now, we will re-examine the operation of the circuits in Figure 14(b), Figure 13(b), and Figure 15.

Consider the largest loop (A–B–C–D–A) in Figure 14(b). Devices M_1 and M_2 have common gate and common bulk and so do devices M_3 and M_4 . When adjacent devices are paired in different ways we observe that M_3 and M_2 share the same source and bulk which is the case also for M_1 and M_4 .

In the largest loop (A–B–C–D–A) of the translinear multiplier circuit of Figure 13(b), we can verify that transistors M_1 and M_2 as well as M_3 and M_4 share common gate and source. The alternative pairing finds M_4 and M_2 sharing same bulk and source, which is also the case for M_3 and M_1 .

When devices in the loop are operating in the ohmic regime, such as M_2 in the circuit of Figure 15, we can verify that the loop (GND–B–C–GND) incorporates two adjacent sets of devices. Note that M_3 and M_2 share same bulk and source/drain while M_3 and M_4 share bulk and gate; the bulk in this circuit is the same for all devices.

4.4 Translinear circuit dynamics

The dynamics of translinear circuits and systems have not been discussed in this paper. However, it was pointed out in [66] that, in networks with non-linear conductances without complementary non-linear reactances, the state equations that describe the dynamics of the system are non-linear. Given an architecture and a particular network, a method was outlined to test for stability [66].

5 A Contrast Sensitive Silicon Retina

The analog silicon system is modeled after neurocircuitry in the distal part of the vertebrate retina—called the outer-plexiform layer. Figure 21 illustrates interactions between cells in this layer [76]. The well-known center/surround receptive field emerges from this simple structure, consisting of just two types of neurons. Unlike the ganglion cells in the inner retina and the majority of neurons in the nervous system, the neurons that we model here have graded responses (they do not spike); thus this system is well-suited to analog VLSI.

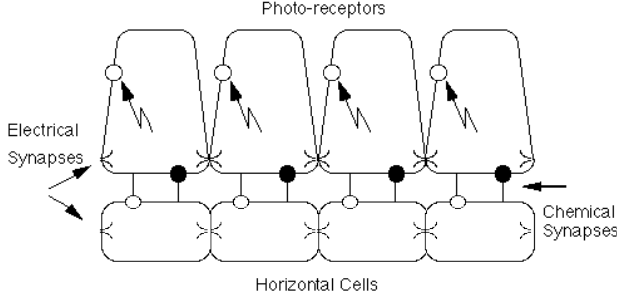


Figure 21: One-dimensional model of neurons and synapses in the outer-plexiform layer. Based on the red-cone system in the turtle retina.

The photoreceptors are activated by light; they produce activity in the horizontal cells through excitatory chemical synapses. The horizontal cells, in turn, suppress the activity of the receptors through inhibitory chemical synapses. The receptors and horizontal cells are electrically coupled to their neighbors by electrical synapses. These allow ionic currents to flow from one cell to another, and are characterized by a certain conductance per unit area.

In the biological system, contrast sensitivity—the normalized output that is proportional to a local measure of contrast—is obtained by shunting inhibition. The horizontal cells compute the local average intensity and modulate a conductance in the cone membrane proportionately. Since the current supplied by the cone outer-segment is divided by this conductance to produce the membrane voltage, the cone’s response will be proportional to the ratio between its photoinput and the local average, i. e. to contrast. This is a very simplified abstraction of the complex ion-channel dynamics involved. The advantage of performing this complex operation at the focal plane is that the dynamic range is extended (local automatic gain control).

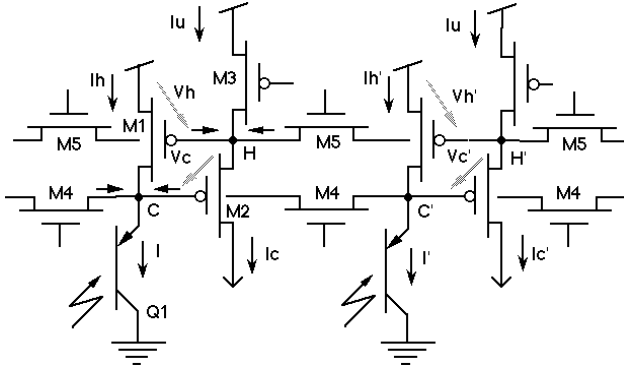


Figure 22: One-dimensional implementation of outer-plexiform retinal processing. There are two diffusive networks implemented by transistors M_4 and M_5 , which model electrical synapses. These are coupled together by controlled current-sources (devices M_1 and M_2) that model chemical synapses. Nodes H in the upper layer correspond to horizontal cells while those in the lower layer (C) correspond to cones. The bipolar phototransistor Q_1 models the outer segment of the cone and M_3 models a leak in the horizontal cell membrane. Note that the actual system has a six neighbor connectivity.

The basic analog MOS circuitry for a one dimensional pixel with two neighbor connectivity is shown in Figure 22. The analysis of the system can be found in [24, 17], here we present an outline and approximations to the main results.

We begin with the non-linear aspects of system operation, its *contrast sensitivity*. The non-linear operation that leads to a local gain-control mechanism in the silicon system is achieved through a mechanism that is qualitatively similar to the biological counterpart, but quantitatively different (see discussion in [24]). Referring to Figure 22, the output current $I_c(x_m, y_n)$ at each pixel, can be given (approximately) in terms of the input photocurrent $I(x_m, y_n)$ and a local average of this photocurrent in a pixel neighborhood (M, N) . This region may extend beyond the nearest neighbor. The fixed current I_u supplied by transistor M_3 normalizes the result.

$$I_c(x_m, y_n) = I_u \frac{I(x_m, y_n)}{\left(I(x_m, y_n) + \sum_{M, N} I(x_i, y_j)\right)} \quad (58)$$

At any particular intensity level, the outer-plexiform behaves like a linear system that realizes a powerful second-order regularization algorithm for edge detection. This can be seen by performing an analysis of the circuit about a fixed operating point. To simplify the equations we first assume that $\hat{g} = \langle I_h \rangle g$, where $\langle I_h \rangle$ is the local average. Now we treat the diffusers (devices M_4) between nodes C and C' as if they had a fixed diffusivity $\hat{g}$. The diffusivity of the devices M_5 between nodes H and H' in the horizontal network is denoted by h . Then the simplified equations describing the full two-dimensional circuit on a square grid are:

$$I_h(x_m, y_n) = I(x_m, y_n) + \hat{g} \sum_{\substack{i = m \pm 1 \\ j = n \pm 1}} \{I_c(x_i, y_j) - I_c(x_m, y_n)\}$$

$$I_c(x_m, y_n) = I_u + h \sum_{\substack{i = m \pm 1 \\ j = n \pm 1}} \{I_h(x_m, y_n) - I_h(x_i, y_j)\}$$

Using the second-difference approximation for the laplacian, we obtain the continuous versions of these equations

$$I_h(x, y) = I(x, y) + \hat{g} \nabla^2 I_c(x, y) \quad (59)$$

$$I_c(x, y) = I_u - h \nabla^2 I_h(x, y) \quad (60)$$

with the internode distance normalized to unity. Solving for $I_h(x, y)$, we find

$$\hat{g} h \nabla^2 \nabla^2 I_h(x, y) + I_h(x, y) = I(x_i, y_j) \quad (61)$$

This is the *biharmonic* equation used in computer vision to find an optimally smooth interpolating function $I_h(x, y)$ for the noisy, spatially sampled data $I(x_i, y_j)$; it yields the function with minimum energy in its second derivative [77]. The coefficient $\lambda = \hat{g} h$ is called the regularizing parameter; it determines the trade-off between smoothing and fitting the data.

A one dimensional solution to this equation can be obtained using Green's functions valid for vanishing boundary conditions at plus and minus infinity:

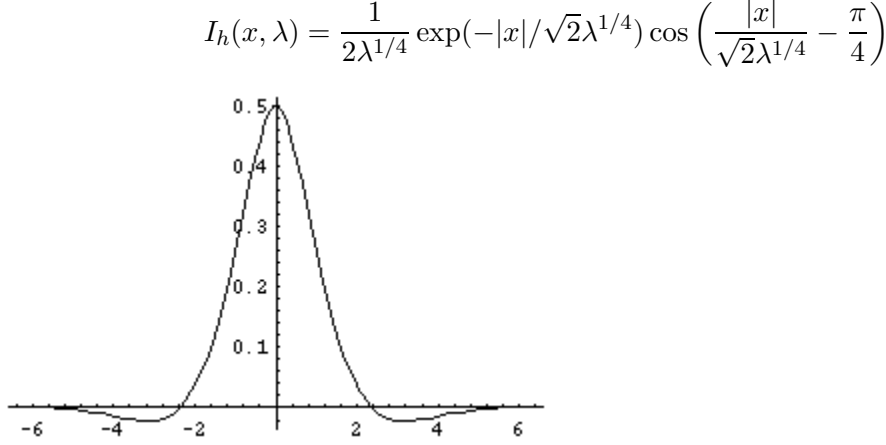


Figure 23: Plot for the one dimensional solution of the biharmonic equation; $\lambda = 1$

In the original work [24], the chip was fabricated with 90×92 pixels on a 6.8×6.9 mm die in a $2\mu\text{m}$ n-well double metal, double poly, garden variety digital oriented CMOS technology and was fully functional. More recently the same system has been fabricated with 230×210 pixels on a 1×1 cm die in a $1.2\mu\text{m}$ n-well double metal, double poly, digital oriented CMOS technology. The chip incorporates 590,000 transistors, 48,000 pixels, operating in subthreshold/transition region with power dissipation on the order of a few mW when powered from a 5V power supply. Temporal response is in the order of a few microseconds.

To find the energetic efficiency of this system we assume that a total of 18 low precision operations (OP) are performed per pixel. Six operations are necessary for the convolution with with bandpass kernel of Figure 23, six for the Laplacian operator (Equation 60) and six for the local gain control computation (Equation 58). If the system is biased so that at the pixel level the frequency response is 100Khz, approximately 1×10^{12} low precision calculations per second are performed in the (210×230) pixels. The power dissipation under the above biasing conditions is about 50mW when operating from 5 Volt power supplies. This is equivalent to 0.05 pW/OP. This performance is a result of an optimization done at the system level, by mapping the problem on an effective physical computational model, rather than trying to optimize the energetic efficiency of an individual gate.

An image captured through the silicon retina is shown in Figure 24. Note the edge enhancement properties of the system and the absence of a dynamic range (flat image).

6 Discussion

The exponential characteristics of a subthreshold MOS device offer the strongest nonlinearity relating a voltage and a current in solid state devices [74] (within the constant κ). When plotted on a logarithmic axis, it manifests itself as a linear function with a constant slope (see Figure 5). In this chapter we have seen how a bottom up design methodology can yield energy efficient



Figure 24: An image of the author as captured by the silicon system. There is a large gradient in illumination from left to right. A regular CCD camera does a poor job in imaging both the high and the low illumination regions.

circuits for computing and signal processing that exploit the translinear property and zero gate conductance of the MOS transistors.

The importance of this limiting steepness, and associated with it highest possible amplification, has long been recognized by engineers involved in the design of analog linear integrated circuits, and in their literature it is referred to as the “Boltzmann limited” slope. Carver Mead often points out the striking similarity between the electrical properties of excitable membranes and the MOS subthreshold characteristics, (see Figure 1 in [12]) as both exhibit the Boltzmann limited behaviour. Furthermore, he cites this similarity as one motivation for pursuing the synthetic approach in analog VLSI using subthreshold MOS devices. Having pursued such an approach, we are tempted to ask a question that has to do with differences rather than similarities. What is fundamentally different at this level of description, that could have implications at the **system** level?

A careful examination of the slopes in Figure 1 of [12] (also Figure 4.6 in [11]) reveals that in biological structures the prefactor in the exponent of the controlling node is larger than unity! That is, the transconductance per unit current is not limited to a value equal to or greater than (kT/q) mV per e-fold of current change. Ions in biological membranes are not limited by the Pauli exclusion principle, the conductance dependence is steeper in excitable membranes because of *correlated* charge control of the current (see discussion on page 55 of Hille [7]). In subthreshold MOS operation, the slope can only asymptotically achieve the minimum value of (kT/q) mV per e-fold of current change. The minimum value can however be seen in bipolar transistors and in junction field effect transistors when operating in subthreshold.

The ramifications of this fundamental difference can be appreciated if one attempts to realize physically an information processing system that operates in the neighbourhood of 300K from

power supplies that are only $4 \times (kT/q) \approx 100\text{mV}$ (biological hardware operate under these conditions). The advantage of reduced power supplies is reduced power dissipation and thus an improved figure for the power delay product (see Equation 3). Since the adopted figure of merit is quadratically related to the transconductance per unit current, a device with exponential voltage to current characteristics is always better. Bipolar transistors, field effect transistors operating in subthreshold, or any other barrier controlled device capable of power gain with the “Boltzmann limited” steepness, is “optimum” in this sense.

We now consider a very simple operation at this reduced power supplies, the quantization of a scalar signal for reliable communication. This could be an inverter circuit in VLSI or the generation of an action potential in biology. The effects of thermal agitation in the system make reliable operation of the quantizer possible only when the energy barriers that separate the two states are more than a few (kT) eV apart. This has been discussed extensively in the literature (see for example [1, 2]). The problem becomes more serious in large, complex information systems such as VLSI with millions of computational elements and where structural variability (i.e. “noise” in the individual components,) has to be taken into account (see transistor data in Figure 2. The problem of component variability in complex VLSI systems has been addressed by Mead and Conway in Chapter 9 of [3] and by Keyes in Chapter 4 of [75].

So, how is it possible for an information processing system that has the complexity of biological systems to operate reliably with power supplies of the order of a few (kT/q) Volts?

The issue of structural “noise” in biological systems can be addressed at the architectural level, through robust algorithms and representations much like it was done for our silicon retina, or through local *adaptation*–learning– mechanisms [41]. The problem of “noise” in a thermodynamic sense is a more difficult one. It can perhaps be addressed by the *fine* details of signal-amplification mechanisms that are found in biological systems. For example, biophysics of excitable membranes allow polyvalent charged entities of charge z to respond as a *unit* rather than independently to an applied potential energy differential. This is a cooperative phenomenon that produces Boltzmann limited, non-linear effects that are stronger than those possible in solid-state devices. This would correspond to an effective “cooling” of the system to a temperature (T/z) !

It is unlikely that the question posed in the previous paragraph has a simple answer and, therefore, our explanations must be inadequate. They do, however, point to some intriguing possibilities worth further consideration as better understanding of the chemistry and physics of computation in neural systems may contribute to long term fundamental improvements in solid state electronics and, in particular, in the fields of low energy, low voltage integrated microsystems.

7 Appendix A: Bipolar Transistor Model

The Ebers-Moll model [79, 80] for an npn bipolar transistor is

$$\begin{aligned} I_E &= -I_F + \alpha_R I_R \\ I_C &= \alpha_F I_F - I_R \end{aligned} \tag{62}$$

$$\begin{aligned} I_F &= I_{ES}(e^{\frac{qV_{BE}}{kT}} - 1) \\ I_R &= I_{CS}(e^{\frac{qV_{BC}}{kT}} - 1) \end{aligned} \quad (63)$$

$$\alpha_F I_{ES} = \alpha_R I_{CS} \quad (64)$$

where I_C and I_E are the collector and emitter currents respectively and V_{BE} is the base to emitter voltage,

V_{BC} is the base to collector voltage,

I_{ES} is the saturation current of emitter junction with zero collector current,

I_{CS} is the saturation current of collector junction with zero emitter current,

α_F common-base current gain.

α_R common-base current gain in inverted mode, i.e. with the collector functioning as an emitter and the emitter functioning as a collector.

By convention, the currents for bipolar transistors are positive when flowing into its terminals.

Combining Eqs. (62), (63), and (64), the collector current can be expressed as

$$I_C = \alpha_F I_{ES} \left((e^{\frac{qV_{BE}}{kT}} - 1) - \frac{1}{\alpha_R} (e^{\frac{qV_{BC}}{kT}} - 1) \right) \quad (65)$$

For an ideal device with common-base current gain, α_F , and common-base current gain α_R in inverted mode, very close to unity, the above equation becomes

$$I_C = I_{ES}(e^{\frac{qV_{BE}}{kT}} - e^{\frac{qV_{BC}}{kT}}) \quad (66)$$

However, regular bipolar transistors do not have both α_F and α_R near unity.

When the collector to base voltage equals zero or the collector is reverse biased with respect to the base, the above equation simplifies to the familiar

$$I_C = A_E J_{ES} (e^{\frac{qV_{BE}}{kT}} - 1) \approx A_E J_{ES} e^{\frac{qV_{BE}}{kT}} = I_S e^{\frac{V_{BE}}{V_t}} \quad (67)$$

where A_E is a design parameter, the area of the emitter junction. J_{ES} and I_S are the saturation current density and current for the emitter respectively. In this case, $I_R \ll I_F$ and the equations above give $I_C = -\alpha_F I_E$. Using the relation $I_E + I_C + I_B = 0$ (KCL) we get the familiar result

$$I_C = \frac{\alpha_F}{1 - \alpha_F} I_B \equiv \beta_F I_B \quad (68)$$

where β_F is the common-emitter current gain.

In a standard n-well CMOS process, a vertical pnp transistor is available for circuit design, but only in the common collector configuration since it has the p-substrate as the collector—the

n-well forms the base and a p-diffusion the emitter. This device is useful as a light sensor; the smallest possible phototransistor permitted by the 2-micron design rules has base dimensions of $16\mu m \times 16\mu m$ when the emitter has an area $A_E = 6\mu m \times 6\mu m$. The dark current in these minimum size phototransistors is approximately 100fA. These sensors show a linear response over at least eight orders of magnitude in light intensity. Experimentally determined responsivity of a device with $A_E = 100\mu m \times 100\mu m$ was $73.8A/W$ at $\lambda = 632.8nm$ and $118.5A/W$ at $\lambda = 834nm$. The β is approximately 200 with Early voltage $V_o = 48V$. The frequency response is limited to a few hundred Khz by the large base-collector capacitance.

In some processes, an npn transistor is offered through an extra implant in the n-well to form the base. Typical forward β for these devices are 60 for emitter area $A_E = 8\mu m \times 8\mu m$. The performance of such bipolar devices is limited by the collector resistance r_c which is in the $k\Omega$ range if there is no buried collector implant. The Early voltage of these devices is approximately 45V. At high collector currents (high injection conditions) the characteristics of bipolars deviate from exponential and their current gain β is also reduced. For the npn vertical bipolars, with a minimum emitter area, this high current effect becomes important at current levels above a few hundred microamps. At low collector currents the β is also limited by recombination in the base. Typical betas range from 20 at current levels of a few nanoamps, to their maximum at current levels of a few hundred microamps.

References

- [1] R. Landauer, “Irreversibility and heat generation in the computing process,” *IBM Journal of Research and Development*, pp. 183-191, July 1961.
- [2] R. Landauer, “Information is Physical,” *Proceedings of the 1992 Physics of Computation Workshop*, pp. 1-4, Dallas, Texas, 1992.
- [3] C.A. Mead and L. Conway, *Introduction to VLSI Systems*, Reading, MA, Addison-Wesley, 1980.
- [4] P. Wheeler L.C. Aiello, “The expensive-tissue hypothesis,” *Current Anthropology*, vol. 36, no. 2, pp. 199-201, April 1995.
- [5] M. Jones J. Allman A. Hakeem, G.R. Sandoval, *Brain and life span in primates*, chapter 5, Academic Press, fourth edition, 1996.
- [6] D. Hubel, *Eye, Brain and Vision*, Scientific American Library, 1988.
- [7] B. Hille, *Ionic Channels of Excitable Membranes*, Sunderland, MA, Sinauer Associates Inc., 1984.
- [8] A. G. Andreou, “On physical models of neural computation and their analog vlsi implementation,” in *Proceedings of the 1994 Workshop on Physics and Computation*, Los Alamitos CA, 1994, pp. 255-264, IEEE Computer Society Press.
- [9] H.B. Barlow, “Unsupervised Learning,” *Neural Computation*, Vol. 1, No. 3, pp. 295-311, Fall 1989.
- [10] S. Haykin, *Neural Networks; A comprehensive foundation*, McMillan College Publishing, New York, 1994.
- [11] C.A. Mead, *Analog VLSI and Neural Systems*, Reading, MA: Addison-Wesley, 1989.
- [12] C.A. Mead, “Neuromorphic electronic systems,” *Proceedings IEEE*, vol. 78, no. 10, pp. 1629-1636, Oct. 1990.
- [13] E.A. Vittoz and J. Fellrath, “CMOS analog integrated circuits based on weak inversion operation,” *IEEE Journ. of Solid-State Circuits*, SC-12, No. 3, pp. 224-231, June 1977.
- [14] E.A. Vittoz, “Micropower techniques,” in *VLSI Circuits for Telecommunications*, edited by Y. P. Tsividis and P. Antognetti, Prentice Hall, 1985.
- [15] A.G. Andreou and K.A. Boahen, “Synthetic neural circuits using current domain-signal representations,” *Neural Computation*, 1, pp. 489-501, 1989.
- [16] A. G. Andreou, K.A. Boahen, P.O. Pouliquen, A. Pavasović, R.E. Jenkins and K. Strohhahn, “Current-mode subthreshold MOS circuits for analog VLSI neural systems,” *IEEE Trans. on Neural Networks*, vol. 2, no. 2., pp. 205-213, March 1991.

- [17] A.G. Andreou and K.A. Boahen, "Neural Information Processing (II)," Chapter 8, in *Analog VLSI Signal and Information Processing*, M. Ismail and T. Fiez eds., McGraw-Hill, 1994.
- [18] A.G. Andreou, "Low Power Analog VLSI Systems for Sensory Information Processing," *Microsystems Technologies for Multimedia Applications: An Introduction*, edited by B. Sheu, Edgar Sanchez-Sinencio and M. Ismail, IEEE Press, Los Alamitos CA, 1995.
- [19] K.A. Boahen, "Retinomorphic Vision Systems," *Proceedings of MicroNeuro-96*, IEEE Computer Society Press, Los Alamitos CA, 1996.
- [20] R. Douglas, M. Mahowald and C.A. Mead, "Neuromorphic Analog VLSI," *Annu. Rev. Neurosci.*, vol. 18, pp. 255-281, 1995.
- [21] B. Gilbert, "Translinear circuits: A proposed classification," *Electronics Letters*, vol. 11, No. 1, pp. 14-16, 1975; errata vol. 11, No. 6, page 136.
- [22] A.G. Andreou, "Synthetic neural systems using current-mode circuits," *Proceedings of the IEEE 1990 International Symposium on Circuits and Systems*, New Orleans, pp. 2428-2432, May 1990.
- [23] M. Cohen and A.G. Andreou, "Current-Mode Subthreshold MOS Implementation of the Herault-Jutten Autoadaptive Network," *IEEE Journ. of Solid-State Circuits*, Vol. 27, No. 5, pp. 714-727, May 1992.
- [24] K.A. Boahen and A.G. Andreou, "A contrast sensitive silicon retina with reciprocal synapses," *Advances in Neural Information Processing Systems 4*, Moody, J.E., Hanson, S.J. and Lippmann, R.P. (eds.), Morgan Kaufmann Publishers, San Mateo, CA 1992.
- [25] A.G. Andreou and K.A. Boahen, "A 48,000 pixel, 590,000 transistor silicon retina in current-mode subthreshold CMOS," *Proceedings of the 37th Midwest Symposium on Circuits and Systems*, pp. 97-102, Lafayette, Louisiana, August 1994.
- [26] A.G. Andreou and K.A. Boahen, "A 590,000 transistor, 48,000 pixel contrast sensitive, edge enhancing CMOS imager-silicon retina," *Proceedings of the 16th Conference on Advanced Research in VLSI*, pp. 225-240, Chapel Hill, North Carolina, March 1995.
- [27] E. Vittoz, E. Dijkstra and D. Shiels, "Low-Power Design: A collection of CSEM Papers," edited volume, *Electronic Design Magazine*, Hasbrouck Heights, NJ, 1995.
- [28] Papers by Eric Vittoz at ISSCC96, the logitech mouse and the sun tracker.
- [29] E. Vittoz, "Analog VLSI signal processing: why, where and how?" *Journal of Analog Integrated Circuits and Signal Processing*, 6, pp. 27-44, June 1994.
- [30] E.A. Vittoz, "Analog VLSI Implementation of Neural Networks," in *Handbook of Neural Computation*, Institute of Physics Publishing and Oxford University Press, USA.

- [31] "Microsensors," edited by R.S. Muller, R.T. Howe, S.D. Senturia, R.L. Smith and R.M. White, IEEE Press, 1991.
- [32] E.A. Vittoz, "The design of high-performance analog circuits on digital CMOS chips," *IEEE Journ. of Solid-State Circuits*, SC-20, No. 3, pp. 657-665, June 1985.
- [33] Y.P. Tsividis, "Analog MOS integrated circuits—certain new ideas, trends and obstacles," *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 317-321, June 1987.
- [34] J.E. Meyer, "MOS Models and Circuit Simulation," RCA Review, Vol. 32, pp. 42-63, March 1971.
- [35] Y. Tsividis, "Operation and Modeling of the MOS Transistor," McGraw-Hill, New York, NY, 1987.
- [36] Chi-Kai Sin, A. Kramer, V. Hu, R. R. Chu, and P. K. Ko, "EEPROM as an analog storage device, with particular application in neural networks" *IEEE Trans. Electron Devices*, ED-39, No 6, pp. 1410-1419, June 1992.
- [37] M. Holler, S. Tam, H. Castro and R. Benson, "An electrically trainable artificial neural network ETANN with 10420 'floating gate' synapses," Proc. Int. Joint Conference on Neural Networks, Washington D.C., vol. II, pp. 191-196, June 1989.
- [38] T.H. Borgstrom, M. Ismail and S.B. Bibyk, "Programmable Current-Mode Neural Network for Implementation in Analog MOS VLSI," IEE Proceedings, Vol. 137, Pt. G, No. 2, pp. 175-184, April 1990.
- [39] E. Vittoz, H. Oguey, M.A. Maher, O. Nys, E. Dijkstra and M. Chevroulet, "Analog storage of adjustable synaptic weights," VLSI Design of Neural Networks, U. Ramacher and U. Ruckert eds., Kluwer Academic Publishers, Boston, MA, 1991.
- [40] E. Sackinger and W. Guggenbuhl, "An Analog Trimming Circuit based on a Floating Gate Device," *IEEE Journ. of Solid-State Circuits*, SC-23, No. 6, pp. 1437-1440, Dec. 1988.
- [41] C. A. Mead: "Adaptive Retina", in *Analog VLSI Implementation of Neural Systems*, C.A. Mead and M. Ismail eds., Kluwer Academic Publishers, Boston, MA, pp. 239-246, 1989.
- [42] L. R. Carley, "Trimming Analog Circuits Using Floating-Gate Analog Memory" *IEEE J. of Solid State Circuits*, vol. SC-24, pp. 1569-1575, Dec. 1989.
- [43] K. Yang and A.G. Andreou, "Multiple Input Floating-Gate MOS Differential Amplifier and Applications for Analog Computation," Proceedings of the 36th Midwest Symposium on Circuits and Systems, Detroit, MI, August 1993.
- [44] G. Cauwenberghs, C.F. Neugebauer and A. Yariv, "Analysis and verification of an analog VLSI incremental outer-product learning systems," IEEE Transactions on Neural Networks, Vol.3, No. 3, May 1992.

- [45] E. Seevinck, "Analysis and Synthesis of Translinear Integrated Circuits," Studies in Electrical and Electronic Engineering 31, Elsevier, Amsterdam, 1988.
- [46] B. Gilbert, "Current-mode circuits from a translinear viewpoint: a tutorial," in Chapter 2 of C. Toumazou, F.J. Lidgley and D.G. Haigh eds., "Analogue IC Design: The Current-Mode Approach," IEEE Circuits and Systems Series, vol. 2, Peter Peregrinus Ltd, London, 1990.
- [47] E. Seevinck and R.J. Wiergerink, "Generalized translinear circuit principle," *IEEE Journ. of Solid-State Circuits*, SC-26, No. 8, pp. 1198-1102, August 1991.
- [48] K. Yang and A.G. Andreou, "A multiple input differential amplifier based on charge sharing on a floating gate MOSFET," *Journal of Analog Integrated Circuits and Signal Processing*, 6, pp. 197-208, November 1994.
- [49] B.A. Minch, C. Diorio, P. Hasler and C.A. Mead, "Translinear circuits using subthreshold floating-gate MOS transistors," *Journal of Analog Integrated Circuits and Signal Processing*, this issue.
- [50] A. Fabre, "Dual translinear voltage/current convertor," *Electronics Letters*, vol. 19, No. 24, pp. 1030-1031, 1983.
- [51] A. Fabre, "Wideband translinear current convertor," *Electronics Letters*, vol. 20, pp. 241-242, 1984.
- [52] P.M. Furth and A.G. Andreou, "A high-drive low-power BiCMOS buffer using compound PMOS/NPN transistors," Proceedings of the 36th Midwest Symposium on Circuits and Systems, pp. 1369-1372, Detroit, Michigan, August 1993.
- [53] A.G. Andreou, K. Strohbehn and R. Jenkins, "A silicon retina for motion computation," Proceedings of the IEEE 1991 International Symposium on Circuits and Systems, Singapore, pp. 1373-1376, June 1991.
- [54] Z. Kalayjian, A.G. Andreou, L. Wolff, J. Williams "Integrated 1-D polarization imagers," Proceedings of the 29th Annual Conference on Information Sciences and Systems, Baltimore, MD, March 1995.
- [55] X. Arreguit, E.A. Vittoz, F.A. van Schaik and A. Mortara, "Analog implementation of low-level vision systems," Proceedings of European Conference on Circuit Theory and Design, Elsevier Publishing, pp. 275-280, September 1993.
- [56] T. Delbrück, "'Bump' circuits for computing similarity and dissimilarity of analog voltages," Proc. Int. Joint Conf. on Neural Networks, Seattle WA, pp. 475-479, 1991.
- [57] M. Cohen and A.G. Andreou, "MOS circuit for nonlinear Hebbian learning," *Electronics Letters*, Vol. 28, No. 9, pp. 809-810, April 1992.
- [58] M. van der Gevel and J.C. Kuenen, "Square-root X circuit based on a novel, back-gate-using multiplier," *Electronics Letters*, Vol. 30, No. 3, pp. 183-184, February 1994.

- [59] M.H. Cohen and A.G. Andreou, "Analog CMOS integration and experimentation with an autoadaptive independent component analyzer," *IEEE Transactions on Circuits and Systems: Part II: Analog and Digital Signal Processing*, Vol. 42, No 2, pp. 65-77, February 1995.
- [60] T.S. Lande, J.A. Nesheim and Y. Berg, "Auto Correlation in Micropower Analog CMOS," *Journal of Analog Integrated Circuits and Signal Processing*, 7, pp. 61-68, January 1995.
- [61] B.L. Hart, "Translinear circuit principle: a reformulation," *Electronics Letters*, vol. 15, no. 24, pp. 801-803, 22nd Nov., 1979.
- [62] Mead, C. A. and Mahowald, M. A., "A silicon model of early visual processing," *Neural Networks*, 1:91-97, 1988.
- [63] E. Vittoz and X. Arreguit, "Linear networks based on transistors," *Electronics Letters*, vol. 29, pp. 297-299, February 4, 1993.
- [64] K. Bult and G. J. G. M Geelen, "An inherently linear and compact MOST-only current division technique," *IEEE J. Solid-State Circuit*, vol. SC-27, no.12, pp. 1730-1735, December 1992.
- [65] M. Tartagni and P. Perona, "Computing centroids in current-mode technique," *Electronics Letters*, vol. 21, October 1993.
- [66] F. J. Pineda and A. G. Andreou, "An analog neural network inspired by fractal block coding," *Advances in Neural Information Processing Systems*, Volume 7, Morgan Kaufmann Publishers, San Mateo, 1995.
- [67] K. C. Smith and A. S. Sedra, "The Current Conveyor — a new circuit building block," *IEEE Proc.*, Vol. 56, pp. 1368-1369, Aug. 1968.
- [68] K.A. Boahen, P.O. Pouliquen, A.G. Andreou and R.E. Jenkins, "A heteroassociative memory using current-mode MOS analog VLSI circuits," *IEEE Trans. Circuits and Systems*, CAS-36, No. 5, pp. 747-755, May 1989.
- [69] K.A. Boahen, A.G. Andreou, P.O. Pouliquen and A. Pavasović, "Architectures for associative memories using Current-Mode analog MOS circuits," *Proceedings of the Decennial Caltech Conference on VLSI*, C. Seitz editor, MIT Press, 1989.
- [70] E. Säckinger and W. Guggenbühl, "A High-Swing, High Impedance MOS Cascode Circuit," *IEEE Jour. Solid-State Cir.*, Vol. 25, No. 1, pp. 289-298, Feb. 1990.
- [71] C. Toumazou, F.J. Lidgey and D.G. Haigh eds., "Analogue IC Design: The Current-Mode Approach," *IEEE Circuits and Systems Series*, vol. 2, Peter Peregrinus Ltd, London, 1990.
- [72] G. R. Wilson, "A Monolithic Junction FET-npn Operational Amplifier," *IEEE Jour. Solid-State Cir.*, Vol. SC-3, No. 4, pp. 341-348, Dec. 1968.

- [73] A. Pavasović, A. G. Andreou, and C. R. Westgate, "Characterization of subthreshold MOS mismatch in transistors for VLSI systems," *Journal of Analog Integrated Circuits and Signal Processing*, 6, pp. 75-84, June 1994.
- [74] W. Shockley, *Electrons and Holes in Semiconductors*, Princeton, NJ, D. van Nostrand Company, 1963 (page 90); J.B. Gunn, "Thermodynamics of Nonlinearity and Noise in Diodes," *Journ. Applied Physics*, Vol. 39, No. 12, pp. 5357-5361, 1968.
- [75] R.W. Keyes, *The Physics of VLSI Systems*, Addison-Wesley, Wokingham, England, 1987.
- [76] J. E. Dowling, "The retina: an approachable part of the brain," The Belknap Press of Harvard University, Cambridge, MA, 1987.
- [77] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory," *Nature*, 317, pp. 314-319, 1985.
- [78] C. Koch, "Seeing Chips: analog VLSI circuits for computer vision," *Neural Computation*, vol. 1, 2, pp. 184-200, 1989.
- [79] J. J. Ebers and J. L. Moll, "Large-signal behavior of junction transistors," *Proceedings IRE*, vol. 42, pp. 1761-1772, December 1954.
- [80] R. S. Muller and T. I. Kamins, "Device electronics for integrated circuits," New York, John Wiley and Sons, 1977.