On physical models of neural computation and their analog VLSI implementation

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

Abstract

We examine computation in a framework where the problem is essentially that of extracting a signal from noise, a filtering *(selective amplification) or an* estimation *problem. Our discussion is relevant to computational tasks in sensory communication such as vision, speech, and natural language processing. We consider "real" systems, both natural (neural systems) and human engineered (silicon integrated circuits), where information processing takes the form of an irreversible physical process. We argue, and demonstrate experimentally, that it is possible to see the emergence of truly complex processing structures that are commensurate with the physical properties of the computational substrate and therefore are energetically efficient.*

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 problems in sensory communication and motor control, by sustaining high computational throughput with minimal energy dissipation.[1] Their effectiveness stems partly from exploiting *prior* knowledge about the problems that they encounter [4]. Such information in the form of *internal models,* reflects the statistical properties of the natural environments in which the systems function. Since the environment is rarely fixed, model *adaptation* and *self-organization* is necessary.

[1] Jim Bower, a neuroscientist at CalTech, argues that the energetic efficiency of the brain is so high that it does not produce enough heat to keep itself warm and thus we see the evolutionary development of a hairy skull.

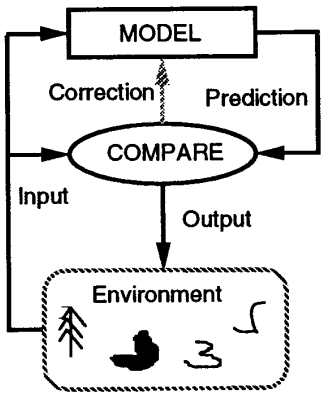

Figure 1: An adaptive system and its interaction with the environment. External inputs to the system are compared with its internal state (model), to produce an output. Information from the input and from the comparison process may be used to adjust the parameters of the internal model. (Adapted from Carver Mead in [6]).

Sensory communication problems can therefore be cast into the canonical form an adaptive system [5, 6]. At each level of processing there is an internal model that is refined by experience as shown by the functional–block-level description in Figure 1. Adaptation is a pervasive property of neural systems and is found at many different levels of a hierarchical neural organization. For example, adaptation can be found in the electromechanical properties of sensory transducers, in the network properties of neurons, and all the way to the abstraction of high level cognitive processes. The classical information theory formalism and the canonical model of a communication channel can be used to develop theories of how such statistical internal models help to "optimaly" encode signals in neural pathways (see for example [7]).

However, the question of *how* such sophisticated processing is actually carried out by the "real" sys-

0-8186-6715-X/94 $4.00

tem, still remains open. "Real" computing structures must satisfy strict constraints of size, weight, utilization of energetic resources, and the ability to operate at temperatures where favorable conditions exist for the development of life as "we" know it –vicinity of 300K–. Algorithms, based on statistical methods and self-organizing techniques for data processing, are notorious for their enormous computational requirements when implemented on digital computers. How is such sophisticated processing done in neural structures?

Carver Mead [8, 6] has eloquently argued that an answer to this question could perhaps be found if one looks at algorithms and information processing structures that emerge from the physical properties of the computational substrate. Furthermore, Mead and coworkers propose an *analysis by synthesis* approach, where analog methods and VLSI technology can be used to prototype such "not-so-conventional" information processing *systems*. From this perspective, analog VLSI technology can be viewed as a modeling tool [8, 9] aimed at capturing the behavior of neurons, networks of neurons, or the complex mechanical-electrical-chemical information processing in biological systems. Computationally, analog VLSI models can be more effective compared to software simulations. More important, they are "real" models, constrained by fundamental physical limitations and scaling laws. Constraints such as: power dissipation, physical extent of computing hardware, density of interconnects, gain-bandwidth product limitations in the gain elements, precision and noise in the characteristics of the basic elements, signal dynamic range, and robust behavior and stability, may force the development of more realistic models. Our work [10] follows a similar line of thought.

From a more practical viewpoint, it is believed that such ideas could also lead to the development of VLSI systems that are more effective in solving sensory communication problems.

In this paper, we discuss how such a methodology has led to the development of an *analog* VLSI silicon system for early vision processing [11]. The architecture is inspired by the processing performed at the outer plexiform layer of the vertebrate retina. It is mapped onto silicon using circuits of minimal complexity that exploit native properties of subthreshold MOS transistors. High computational throughput at low levels of energy dissipation is achieved by employing analog processing in a massively parallel architecture; a paradigm that minimizes the "mismatch" between the physics of the problem and the physics of the computational substrate. We begin with a discussion of analog VLSI.

2 Analog VLSI

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*. The understanding of complex information processing in neural systems through a discussion at different levels, is an approach that was first introduced by Marr and Poggio [12]; also discussed extensively in [13].

Device level: At the lowest level, gain, is provided by MOS transistors operating in subthreshold region [8, 10, 14]. In this regime device physics yield the following functional form for the drain current in terms of the voltages at its four terminals.

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

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 constant κ be less than or 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.

$$I \propto [Q_S - Q_D] \qquad (2)$$

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 [11] 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.

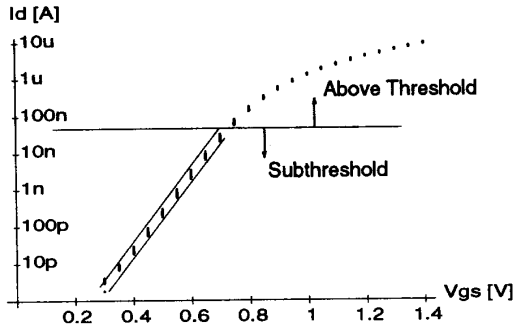

Figure 2: Measured drain current I_d versus gate-source voltage V_{GS} for 32 small geometry transistors ($4 \times 4\mu m$) fabricated in a $2\mu m$ n-well CMOS process; drain-source voltage of V_{DS}=1.5 Volts. The fuzziness in the current, (mismatch between devices), is constant in subthreshold (on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

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

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

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:

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

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 nonlinearities. This procedure is based on what is known as the *Translinear Principle* [16] 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, can be found in [10]. 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 [8], (also in [17]), 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 diffusion process is described by the following equation:

$$\frac{dN}{dt} = D\nabla^2 N(x,y) \tag{5}$$

N is the concentration of the diffusing species and D is their diffusivity. Equation 5 applies to the 2-D case where the concentration is assumed uniform in the third dimension and N is the number of particles per unit area. Two alternative analog simulations of this process on a discrete grid are shown in Figure 3.

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

$$\frac{dV_n}{dt} = \frac{4G}{C}\left(\frac{1}{4}(V_j + V_k + V_l + V_m) - V_n\right) \tag{6}$$

which is homologous with Equation 5 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

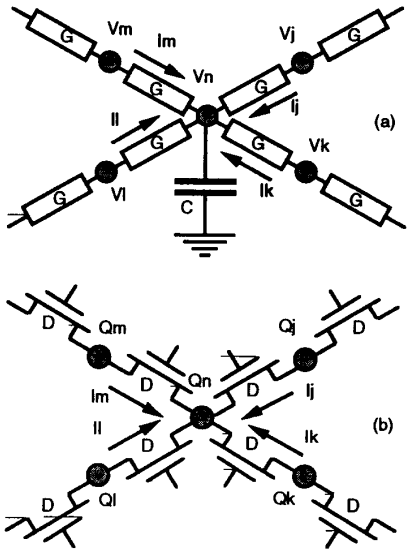

Figure 3: Simulating diffusion with (a) conductances and voltage/current variables or (b) diffusors and charge/current variables.

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

$$\frac{dQ_n}{dt} = 4D\left(\frac{1}{4}(Q_j + Q_k + Q_l + Q_m) - Q_n\right) \tag{7}$$

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 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 in Figure 7 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

$$\lambda \nabla^2 \nabla^2 \Phi + \Phi = \Phi_{in} \tag{8}$$

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 the next section we show how a model of early visual processing is related to the biharmonic equation and can be realized using diffusive networks.

3 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 4 illustrates interactions between cells in this layer [18]. 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.

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.

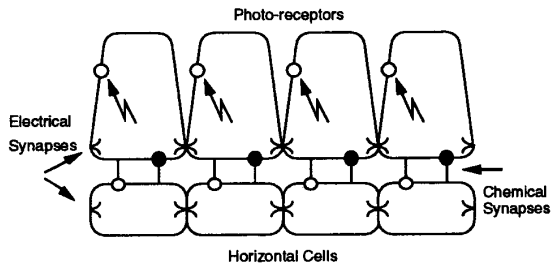

Figure 4: One-dimensional model of neurons and synapses in the outer-plexiform layer. Based on the red-cone system in the turtle retina.

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).

The basic analog MOS circuitry for a one dimensional pixel with two neighbor connectivity is shown in Figure 5. The analysis of the system can be found in [11, 10], 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 acheived through a mechanism that is qualitatively similar to the biological counterpart, but quantitatively different (see discussion in [11]). Refering to Figure 5, 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.

$$I_c(x_m, y_n) = I_u \frac{I(x_m, y_n)}{\left(I(x_m, y_n) + \sum_{\mathrm{M,N}} I(x_i, y_j)\right)} \tag{9}$$

At any particular intensity level, the outer-

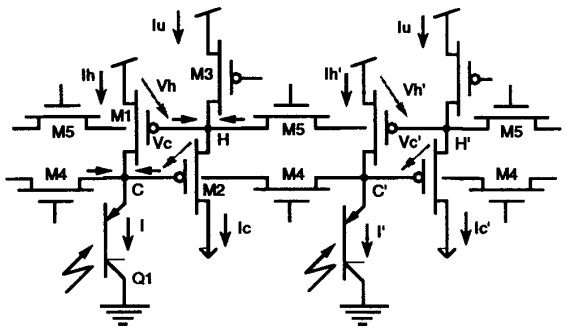

Figure 5: One-dimensional implementation of outer-plexiform retinal processing. There are two diffusive networks implemented by transistors M_4 and M_5, which model electrical synapses. These are coupled together by controlled current-sources (devices M_1 and M_2) that model chemical synapses. Nodes 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.

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 diffusors (devices M_4) between nodes C and C' as if they had a fixed diffusitivity $\hat{g}$. The diffusitivity 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:

$$I_h(x_m, y_n) = I(x_m, y_n) + \hat{g} \sum_{\substack{i=m\pm1 \\ j=n\pm1}} \{I_c(x_i, y_j) - I_c(x_m, y_n)\}$$

$$I_c(x_m, y_n) = I_u + h \sum_{\substack{i=m\pm1 \\ j=n\pm1}} \{I_h(x_m, y_n) - I_h(x_i, y_j)\}$$

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

$$I_h(x,y) = I(x,y) + \hat{g}\nabla^2 I_c(x,y) \quad (10)$$

$$I_c(x,y) = I_u - h\nabla^2 I_h(x,y) \quad (11)$$

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

$$\hat{g}h\nabla^2\nabla^2 I_h(x,y) + I_h(x,y) = I(x_i, y_j) \quad (12)$$

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 [22]. 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:

$$I_h(x,\lambda) = \frac{1}{2\lambda^{1/4}} \exp(-|x|/\sqrt{2}\lambda^{1/4}) \cos\left(\frac{|x|}{\sqrt{2}\lambda^{1/4}} - \frac{\pi}{4}\right)$$

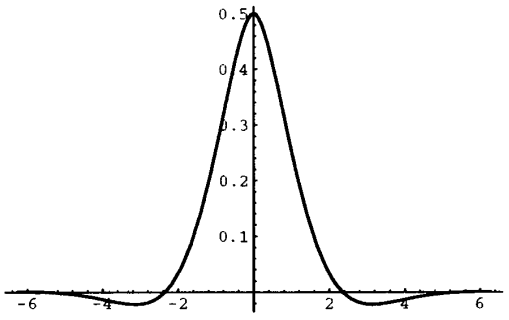

Figure 6: Plot for the one dimensional solution of the biharmonic equation; $\lambda = 1$

In the original work [11], the chip was fabricated with 90 x 92 pixels on a 6.8 x 6.9 mm die in a 2μm 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 x 210 pixels on a 1 x 1 cm die in a 1.2μm 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 6, six for the Laplacian operator (Equation 11) and six for the local gain control computation (Equation 9). 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 x 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 7. Note the edge enhancement properties of the system and the absence of a dynamic range (flat image).

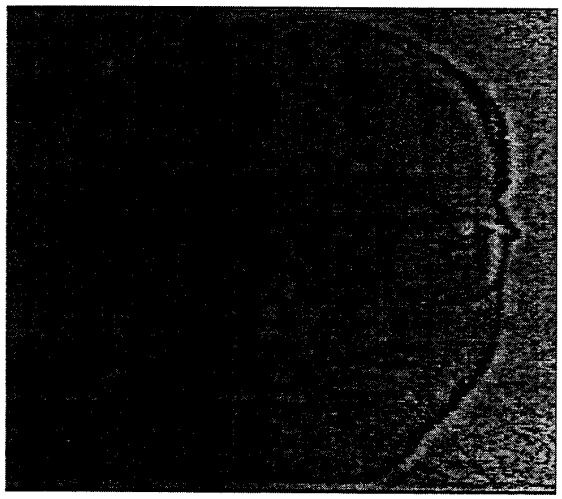

Figure 7: An image of the author as captured by the silicon system.

4 Discussion

On the approach: In the previous sections, we have seen how an analysis by synthesis methodology [8] using analog computation and VLSI technology has led to the development of an energetically efficient analog VLSI system for early vision. Crucial to the success of our endeavor is a hierarchical view of information processing as discussed in Chapters 1 and 7 of [13]. Marr strongly believed however, that computational theory should be on top of the hierarchy and plays the most important role, while the particulars of the implementation have only a peripheral role (see Figure 8a). Our work suggests that it may be beneficial to view the different levels from a slightly different perspective, one that is depicted in

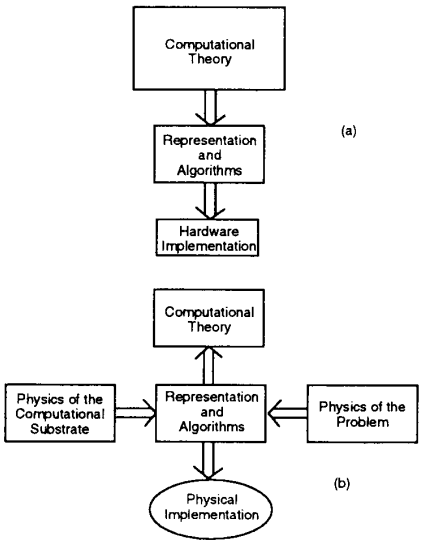

Figure 8: (a) Marr's three levels of looking at complex information processing systems. (b) what could be called the "physics of computation" view.

Figure 8b. We begin with the physics of the problem and the physics of the computational substrate. Good algorithms and representations emerge as a result of constraints imposed at this level.

On algorithms and architectures: We have experimentally demonstrated that in considering possible algorithms and architectures for solving sensory communication problems one need not be restricted to a particular model of computation. The a-priori assumption should be made that a structure exists (within a well defined set of "real" constraints consistent with the computational substrate). This incorporates the best possible model of computation. The particular mapping of a model to a computational substrate is thus guided by fundamental limitations of the basic elements, the properties that make the solution scalable, and the existence of a synthesis procedure that enables the emergence of a complex structure.

For example, it is easier today to write software that implement a filter function on a digital computer, than to implement the filter function using ASIC digital circuits, than to design and implement analog filters as analog integrated circuits, than to design and manufacture a filter based on the physical properties of some mechanical silicon micro-structure. Given the subject matter of this paper, there is no reason to believe that the last solution is not the preferred solution given adequate research resources to solve "algorithmic" and technological problems.

On physical models: The Ising spin model and

the dynamical systems formalism employed by Hopfield [19] is one example of a physical model that has become popular in the field of neural networks. The Hopfield model is of great intellectual value because it demonstrates how a physical dynamical system can be employed for information processing tasks. However, its practical value is limited. The charge-based formulation and analog VLSI implementation of the silicon retina presented here is another example of a physical model that could also be cast in the dynamical systems framework (a relaxation network). It is mathematically interesting, and at the same time perhaps more practical. Indeed, by judiciously employing "physical models" of computation such as the Hopfield network [19] or a detailed biophysical model of a retina [11] the inherent parallelism, nature of physical laws [21] is exploited in the computational process.

The biologically motivated solution to early vision processing is attractive from a computational perspective because *contrast*, an invariant representation of the visual world, has been obtained with a front-end that is robust, small, and extremely low power (a few mW). There is also an engineering benefit because subsequent processing stages are not burdened with handling and processing signals of wide dynamic range.

It can be argued that the analog VLSI retina model has an *a-priori* internal model of the world; one that assumes that the intensity is either uniform or, in the case of non-uniform illumination, is a linear function of space. The output of the system is the difference between the input intensity field and the model. As such, the output is a measure of the second spatial derivatives (or the Laplacian) of the intensity field. In the field of computer vision, linear methods based on regularization theory are used to impose smoothness constraints [22] on the discretely sampled and noisy real world data. These computational demanding algorithms are run on general purpose digital hardware.

In the physical realization of a computational systems, the same "regularization" benefits could be beneficial in dealing with the "noise" introduced by the variability in gain of MOS transistors (see Figure 2).[2]. Thus we see how in the organization of the system one could account for the properties of the computational substrate at the architectural level, that which is irrelevant when implementing algorithms on general purpose, digital computers. In digital computers and symbolic processing machines, structural variability and noise in the basic elements is handled at a much lower level, at the gate level. Switching levels are choosen so that adequate noise margin is introduced for large scale reliable computation.

In the context of the biological model, the function of the horizontal cells (corresponding to nodes H) is to compute "optimaly" a smoothed version of the image (through a convolution with the kernel shown in Figure 6) while the cones (corresponding to nodes C) perform edge detection by taking the Laplacian of the smoothed image as given by Equation 11. The space constant of the solutions is $\lambda^{1/4}$ or $(\hat{g}h)^{1/4}$. The model suggests that specialized structures in biological systems could mitigate some type of "wet-ware regularization" to compensate for the inherent random variations in the neuronal characteristics, which in turn could lead to robust performance in the presence of "noise". The latter statement is just a hypothesis subjected to experimental verification.

The notion of an "optimal" computation step has been introduced by Bialek and Owens [23]. They have considered the signal and noise characteristics of the photoreceptors in the outer retina, and they have derived "optimal" temporal filters to further process the receptor signals. Our work [11] addresses a similar problem in the space domain where "noise" is introduced by the structural variability in the gain of the individual elements, and spatial smoothing is needed to increase the information capacity of the system.

The contrast sensitive silicon retina, is an architecture that yields the ON-center/OFF-surround response at the level of the cone (photoreceptor) network. Even though from an engineering perspective one can employ this function for edge enhancement, and we have done so, the question of why such structure exists in the neural system is still open. To put it more succinctly; is edge enhancement the goal or is it simply an emerging property from a computational function that is aimed at dealing with signals of large dynamic range using imprecise components?

Analog VLSI and neural systems: a discussion in contrasts. The exponential characteristics of a subthreshold MOS device offer the strongest nonlinearity relating a voltage and a current in solid state devices [24] (within the constant κ). When plotted on a logarithmic axis, it manifests itself as a linear function with a constant slope (see Figure 2).

The importance of this limiting steepness 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 to the striking similarity between the electrical properties of excitable mem-

[2]Noise here denotes structural variability, as opposed to noise in a thermodynamic sense

branes and the MOS subthreshold characteristics, (see Figure 1 in [6]) 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 two 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 [6] (also Figure 4.6 in [8]) reveals that in biological structures the constant κ in the exponent (see Equation 1) is larger than unity! That is, the slope is not limited to a value equal to or greater than (kT/q) mV per e-fold of current change. Not being 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 [25]). 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 $4 \times (kT/q) \approx 100$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). [3]

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 [26].

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 arhitectural level, through robust algorithms and representations much like it was done for our silicon retina, or through local *adaptation* –learning– mechanisms. 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. This would correspond to an effective "cooling" of the system to a temperature (T/z)! At lower temperatures, undesirable, thermally activated events would become less frequent, resulting in a more reliable system operation. 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.

Acknowledgments: The research was supported by NSF grant ECS-9313934; Paul Werbos is the program monitor. The author would like to thank Professor Carver Mead of Caltech for encouraging this work. Stimulating discussions with K. Strohbehn, F. Pineda, G. Cauwenberghs and R. Jenkins are acknowledged. The substantial contributions of Kwabena Boahen (Buster) who is a co-author in the original papers are especially acknowledged. The author is thankful to Paul Furth for proofreading the document. Chip fabrication was provided by MOSIS.

[3] 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.

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 Work-

shop, pp. 1-4, Dallas, Texas, 1992.

[3] C.A. Mead and L. Conway, *Introduction to VLSI Systems,* Reading, MA, Addison-Wesley, 1980.

[4] H.B. Barlow, "Unsupervised Learning," Neural Computation, Vol. 1, No. 3, pp. 295-311, Fall 1989.

[5] T. Kohonen, *Self-Organization and Associative Memory,* Springer Verlag, (2nd edition), Berlin, Heidelberg New York, 1988; A.L. Gorin, S. Levinson, A. Gertner and E. Goldman, " Adaptive Aqcuisition of Language," Computer Speech and Language, Vol. 5, No. 2, pp. 101-132, April 1991; S. Haykin, *Neural Networks; A comprehensive foundation,* McMillan College Publishing, New York, 1994.

[6] C.A. Mead, "Neuromorphic electronic systems," *Proceedings IEEE,* vol. 78, no. 10, pp. 1629-1636, Oct. 1990.

[7] R. Linsker, "Self-organization in a perceptual network," IEEE Computer, pp. 105-117, March 1986; J.J. Atick and N.A. Redlich, "Towards a Theory of Early Visual Processing," Neural Computation, Vol. 2, No. 3, pp. 308-320, Fall 1990; Z. Li and J.J. Atick, "Toward a Theory of the Striate Cortex," Neural Computation, Vol. 6, No. 1, pp. 127-146, January 1994.

[8] C.A. Mead, *Analog VLSI and Neural Systems,* Reading, MA: Addison-Wesley, 1989.

[9] M. Mahowald and R. Douglas, "A silicon neuron," Nature, vol. 354, 19/26, December 1991.

[10] 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.

[11] 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.

[12] D. Marr and T. Poggio, "From understanding computation to understanding neural circuitry," Neurosciences Res. Program Bulletin 15, pp. 470-488, 1977.

[13] D. Marr, *Vision,* W.H. Freeman and Company, New York, 1992.

[14] E. A. Vittoz and J. Fellrath, "CMOS analog integrated circuits based on weak inversion operation," *IEEE J. of Solid-State Circuits,* vol. SC-12, no. 3, pp. 224–231, June 1977.

[15] 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.

[16] B. Gilbert, "Translinear Circuits: A Proposed Classification," *Electronics Letters,* vol. 11, No. 1, pp. 14-16, 1975; B. Gilbert, "A Monolithic 16-Channel Analog Array Normalizer," *IEEE Journ. of Solid-State Circuits,* vol. SC-19, No. 6, 1984.

[17] C. Koch, "Seeing Chips: analog VLSI circuits for computer vision," *Neural Computation,* vol. 1, no. 2, pp. 184–200, 1989.

[18] J. E. Dowling, "The retina: an approachable part of the brain," The Belknap Press of Harvard University, Cambridge, MA, 1987.

[19] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Nat. Acad. Sci., USA, 79, pp. 2554-2558, 1982.

[20] B.R. Gossick, *Hamilton's Principle and Physical Systems,* Academic Press, New York, 1967.

[21] D.W. Hillis and B.M. Boghosian, "Parallel scientific computation," Science, vol. 261, p856, 1993.

[22] T. Poggio, V. Torre and C. Koch, "Computational vision and regularization theory," *Nature,* 317, pp. 314-319, 1985.

[23] W. Bialek and W. Geoffrey Owen, "Temporal filtering in retinal bipolar cells: Elements of an optimal computation?", Biophysical Journal, vol. 58, pp. 1227-1233, Nov. 1990.

[24] 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.

[25] B. Hille, *Ionic Channels of Excitable Membranes,* Sunderland, MA, Sinauer Associates Inc., 1984.

[26] R.W. Keyes, *The Physics of VLSI Systems,* Addison-Wesley, Wokingham, England, 1987.