

Neuromorphic architectures: challenges and opportunities in the years to come

Andreas G. Andreou

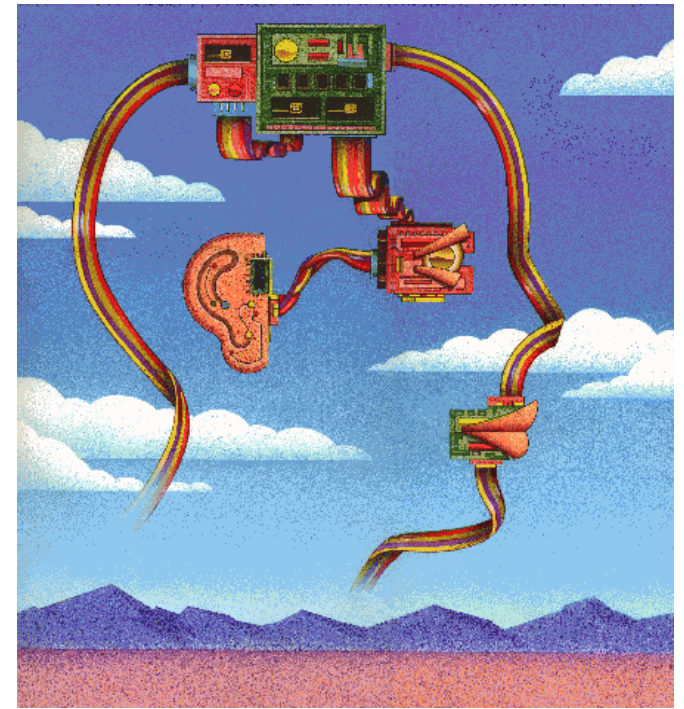
andreou@jhu.edu

Electrical and Computer Engineering

Center for Language and Speech Processing

Johns Hopkins University

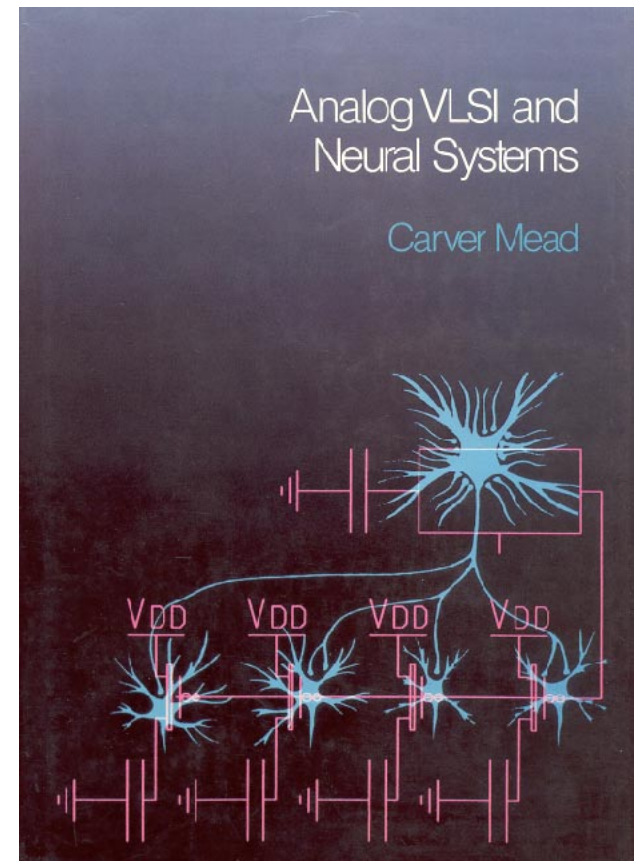
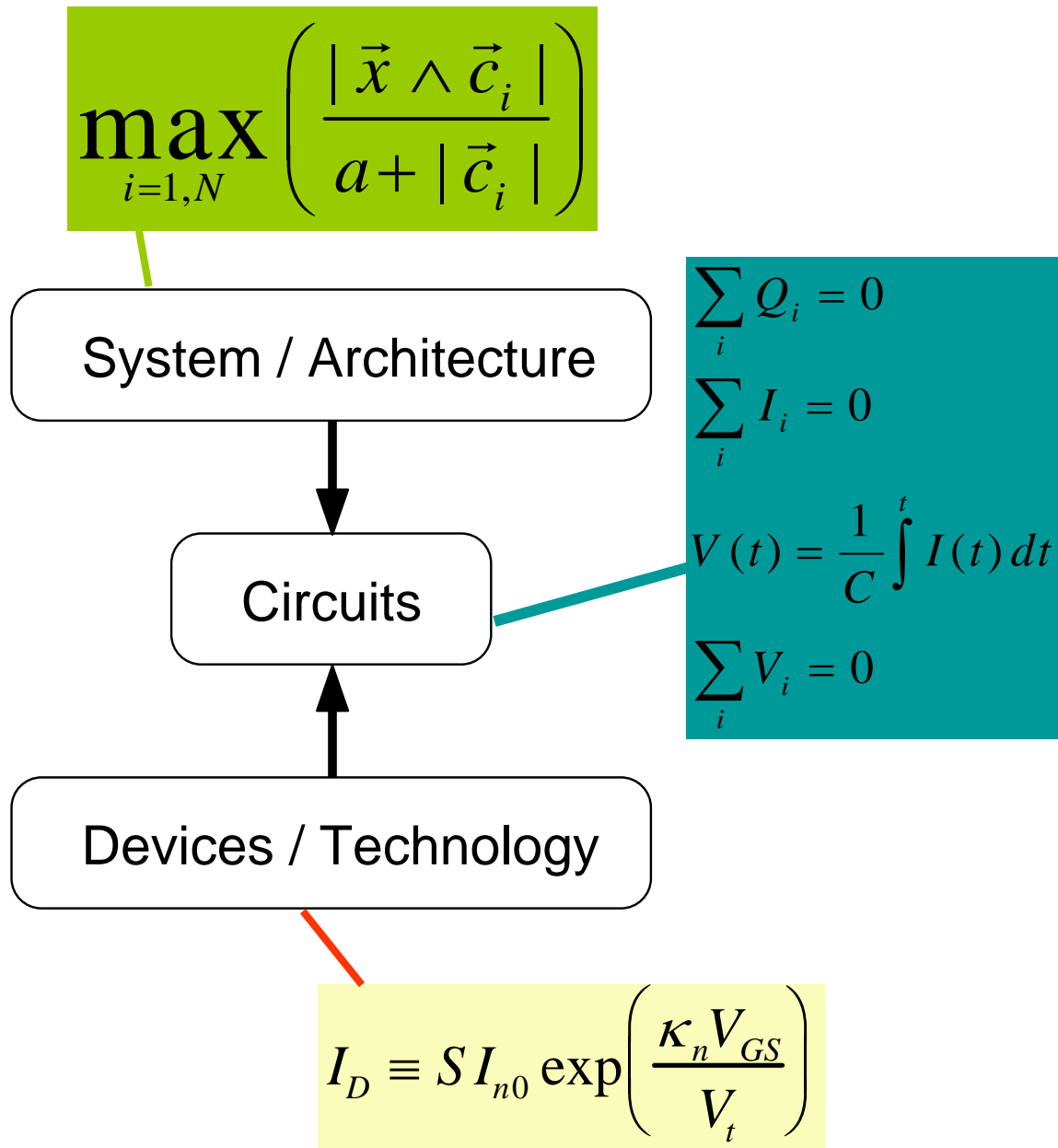
<http://www.ece.jhu.edu/faculty/andreou/AGA/index.htm>



Part I: Neuromorphic architectures

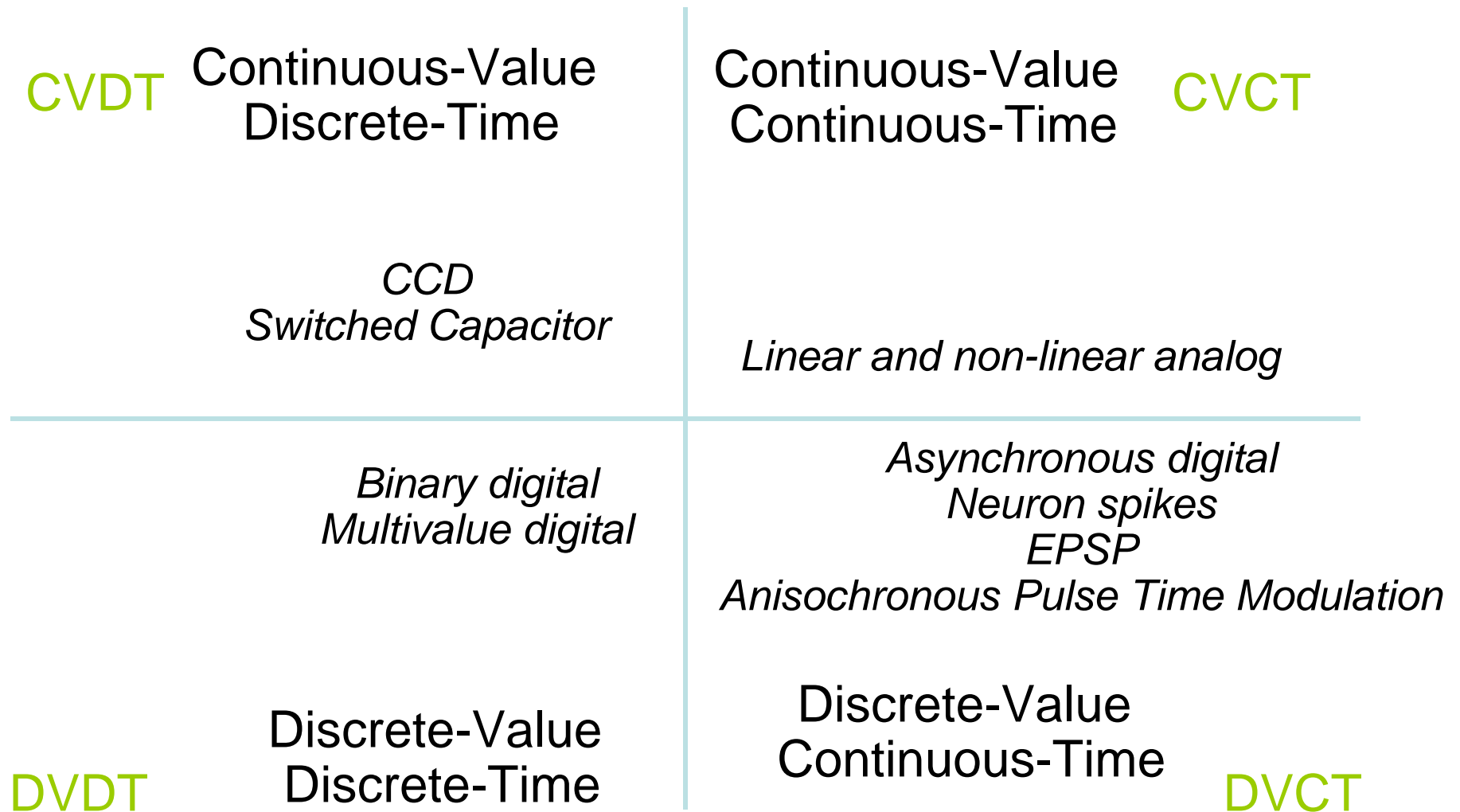
What did we learn the last 25 years?

1986: Let the physics do the work!



October 1986 (1st Draft)

circuits: analog, digital and beyond ...



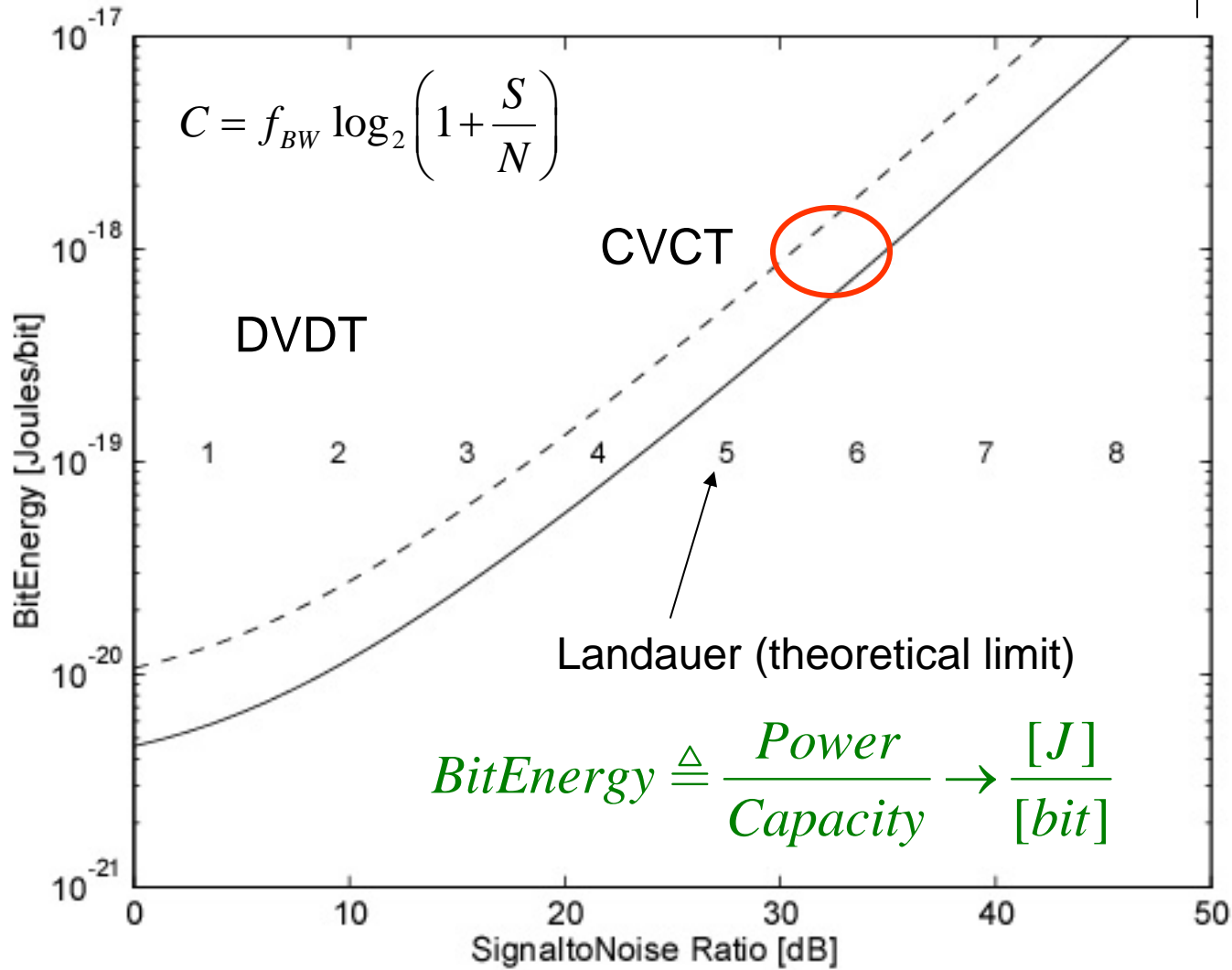
P.M. Furth and A.G. Andreou, "Comparing the bit-energy of continuous and discrete signal representations," *Proceedings of the Fourth Workshop on Physics and Computation (PhysComp96)*, T.Toffoli, M. Biafore and J. Leao eds., New England Complex Systems Institute, pp. 127-133, Boston, MA, November 1996.

the energy costs of computing

$\sim 10^{-16}$

8-9 bits

DVDT
practical limit at
10nm CMOS



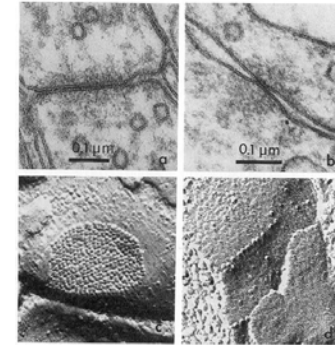
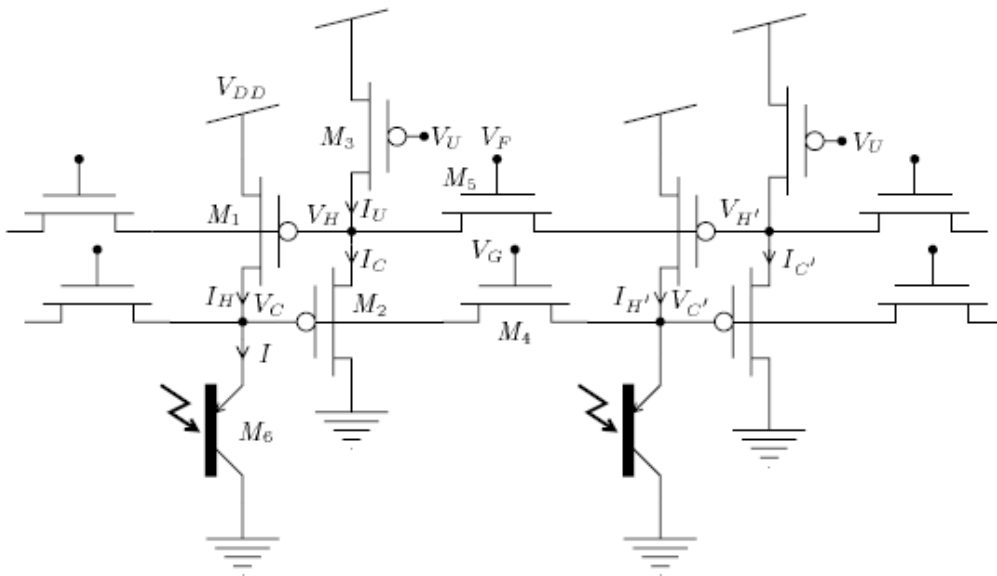
A Contrast Sensitive Silicon Retina with Reciprocal Synapses

Kwabena A. Boahen

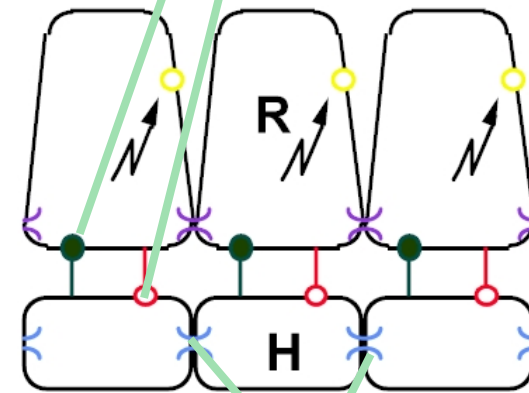
Computation and Neural Systems
California Institute of Technology
Pasadena, CA 91125

Andreas G. Andreou

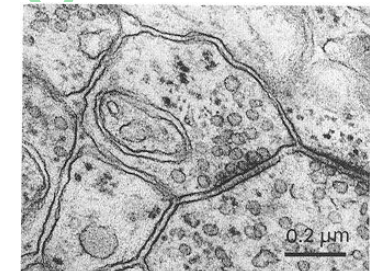
Electrical and Computer Engineering
Johns Hopkins University
Baltimore, MD 21218



Chemical synapses



Electrical synapses



the mathematical abstraction of biology

1. Photons to electrons: transduction and amplification

$$I_{in}(x_m, y_n) = \beta I_{ph}(x_m, y_n) \leftarrow \Phi(x_m, y_n)$$

2. Local gain control: source coding

$$I_{out}(x_m, y_n) = I_u \frac{I_{in}(x_m, y_n)}{I_{in}(x_m, y_n) + \psi \sum_{M,N} I_{in}(x_i, y_j)}$$

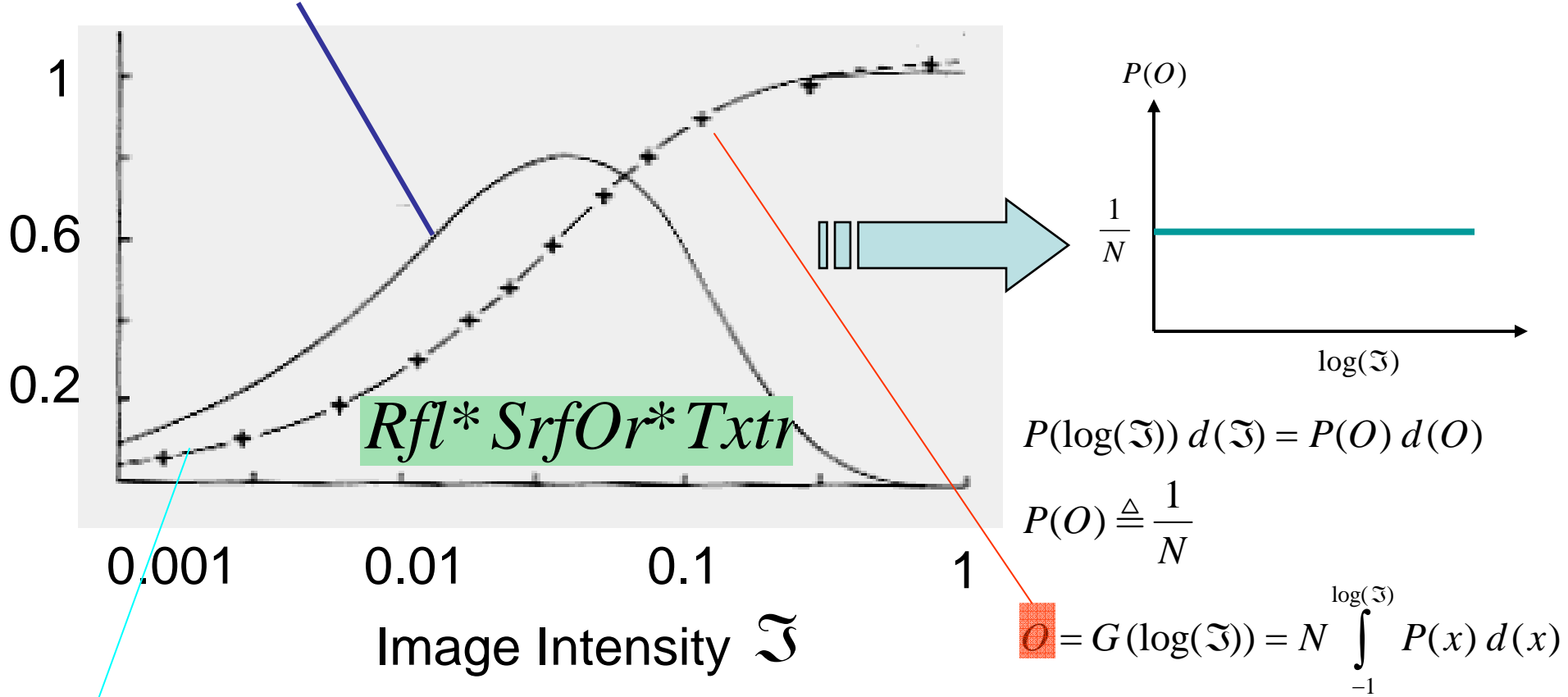
3. Spatial filtering: optimal smoothing

$$i_h(x_m, y_n) + \lambda \nabla^2 \nabla^2 i_h(x_m, y_n) = i_{in}(x_m, y_n)$$

$$i_{out}(x_m, y_n) = -\nabla^2 i_h(x_m, y_n)$$

the statistics of natural scenes

$pdf(\log(\mathfrak{I}))$ W. Richards, "Lightness scale from image intensity distributions", *Applied Optics*, vol. 21, no. 14, 2569-2582, 1982



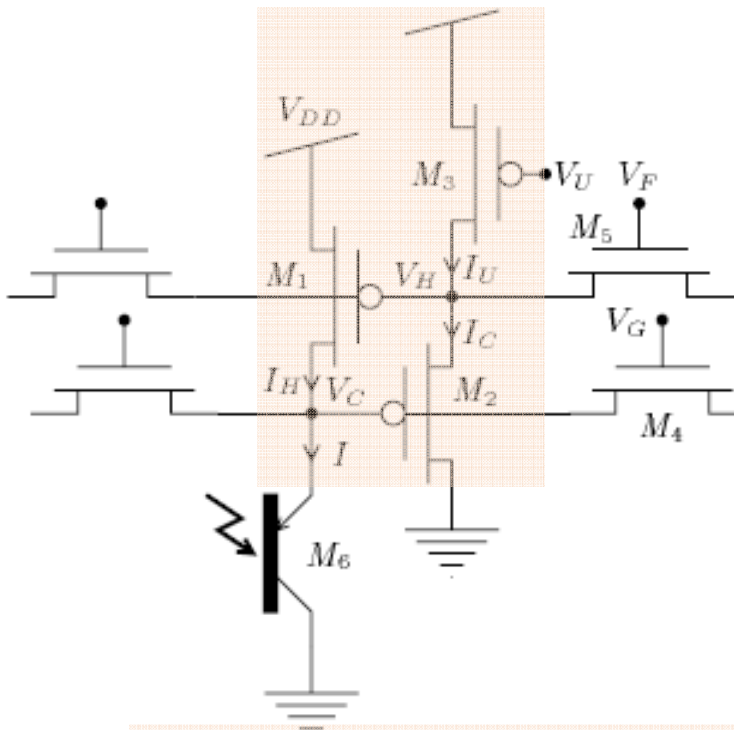
$$\frac{O}{m} = \frac{(\log(\mathfrak{I}))^n}{(\log(\mathfrak{I}))^n + c^n}$$

$m = 12.5$
 $n = 1$
 $I_s = 3$

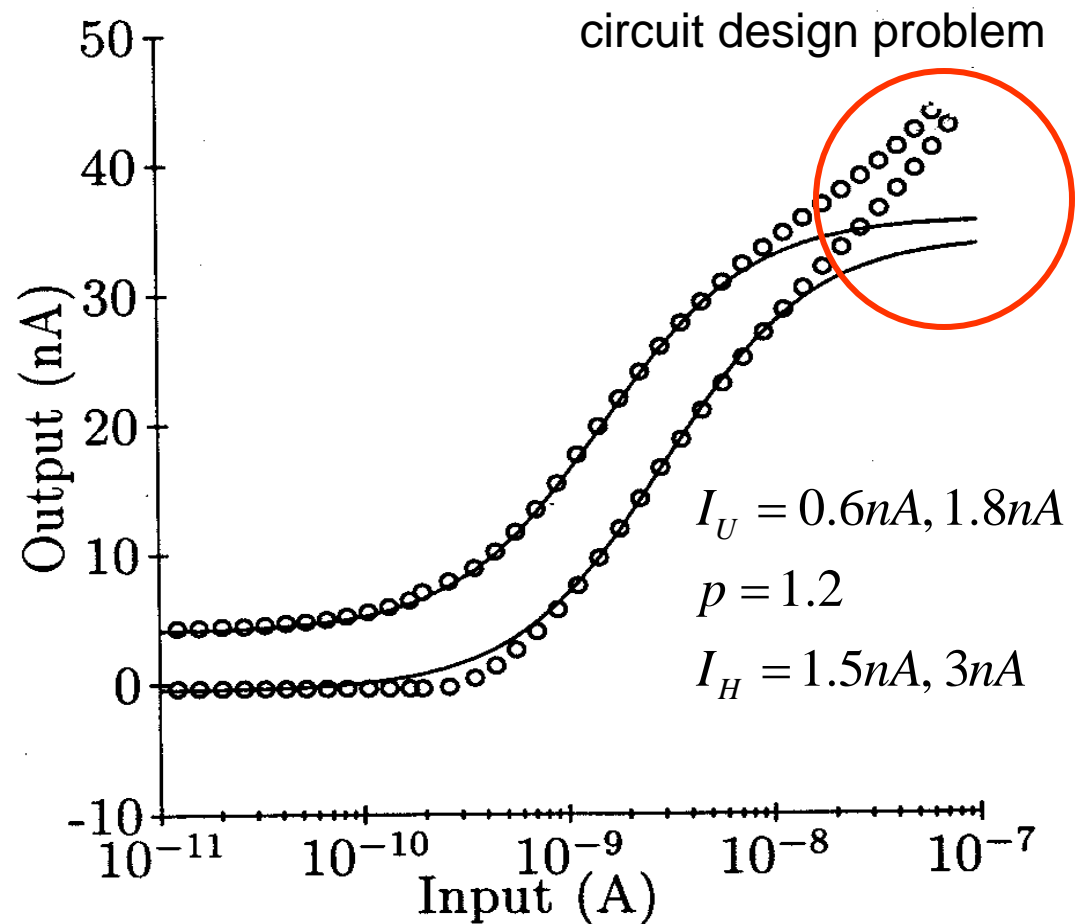
Naka-Rushton Equation

matching signals to circuits!

challenge: matching the wide dynamic range of signals to limited dynamic range of analog computing hardware

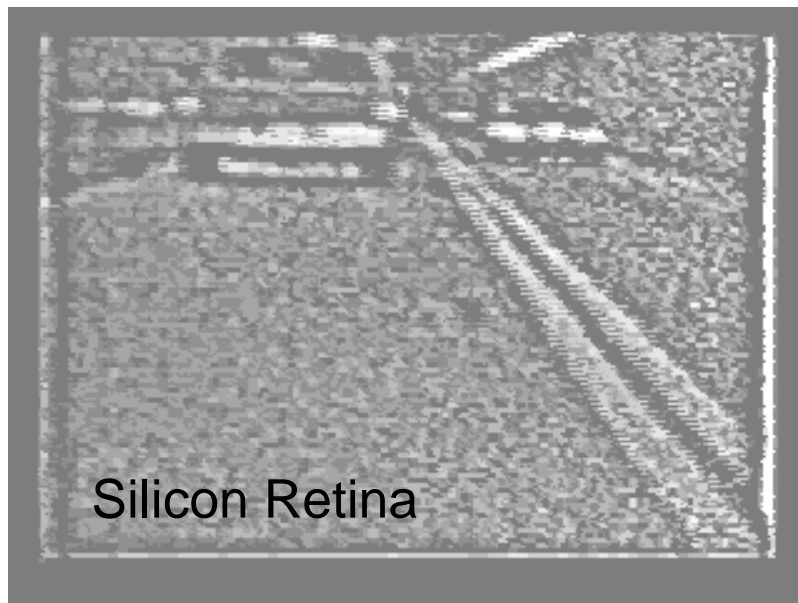


$$\frac{I_{out}}{I_U} = \frac{I_C}{I_U} = \frac{I^P}{I^P + I_H^P}$$



non-linear analog processing to do "source" coding

dealing with the dynamic range problem



(210 x 230 pixels) X

(6 OPS per pixel for second order smoothing) X

(6 OPS per pixel for Laplacian) X

(6 OPS per pixel for gain control) X

(10^5 OPS per second --100 kHz temporal response--)

= $5 \times 10^4 \times 2 \times 10^2 \times 10^5$

10^{12} OPS with 50mW total power dissipation at

5 Volts power supply

Subthreshold CMOS

560,000 transistors, era 1995

embedded analog computing in digital memories

Analog Integrated Circuits and Signal Processing, 13, 211-222 (1997)
 © 1997 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Winner-Takes-All Associative Memory: A Hamming Distance Vector Quantizer

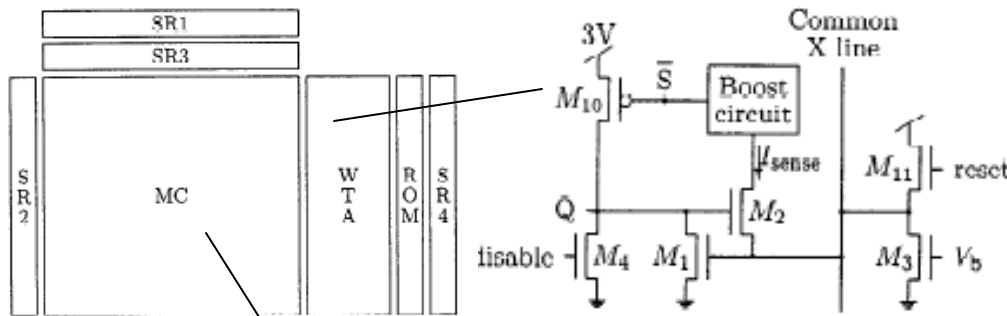
PHILIPPE O. POULIQUEN,¹ ANDREAS G. ANDREOU¹, AND KIM STROHBEHN²

¹[philippe.andreou]@olympus.ece.jhu.edu, ²aleph0@apicomm.jhuapl.edu

¹Electrical and Computer Engineering, Center for Language and Speech Processing, Johns Hopkins University, 3400 N. Charles Street, Baltimore MD

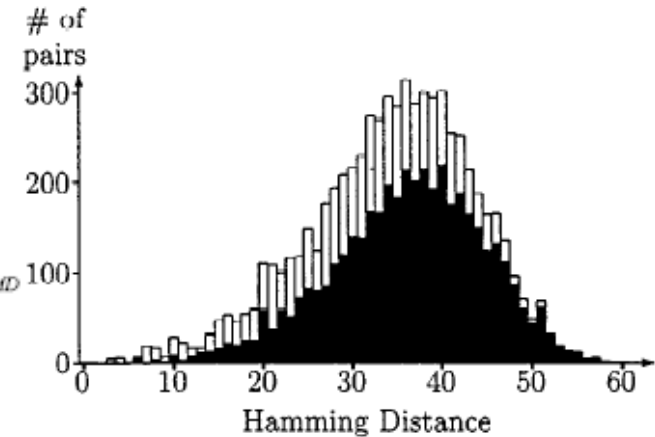
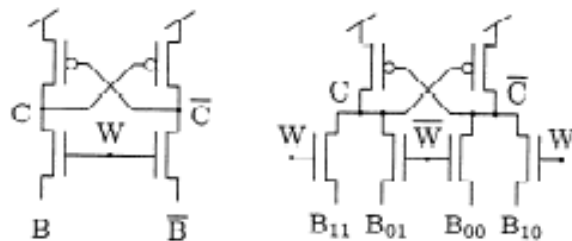
²Applied Physics Laboratory, Johns Hopkins University, Laurel MD 20723 USA

Received June 28, 1996; Accepted



Given the input pattern w , we need to find the template c_j that maximizes

$$\frac{|w \wedge c_j|}{\alpha + |c_j|}$$

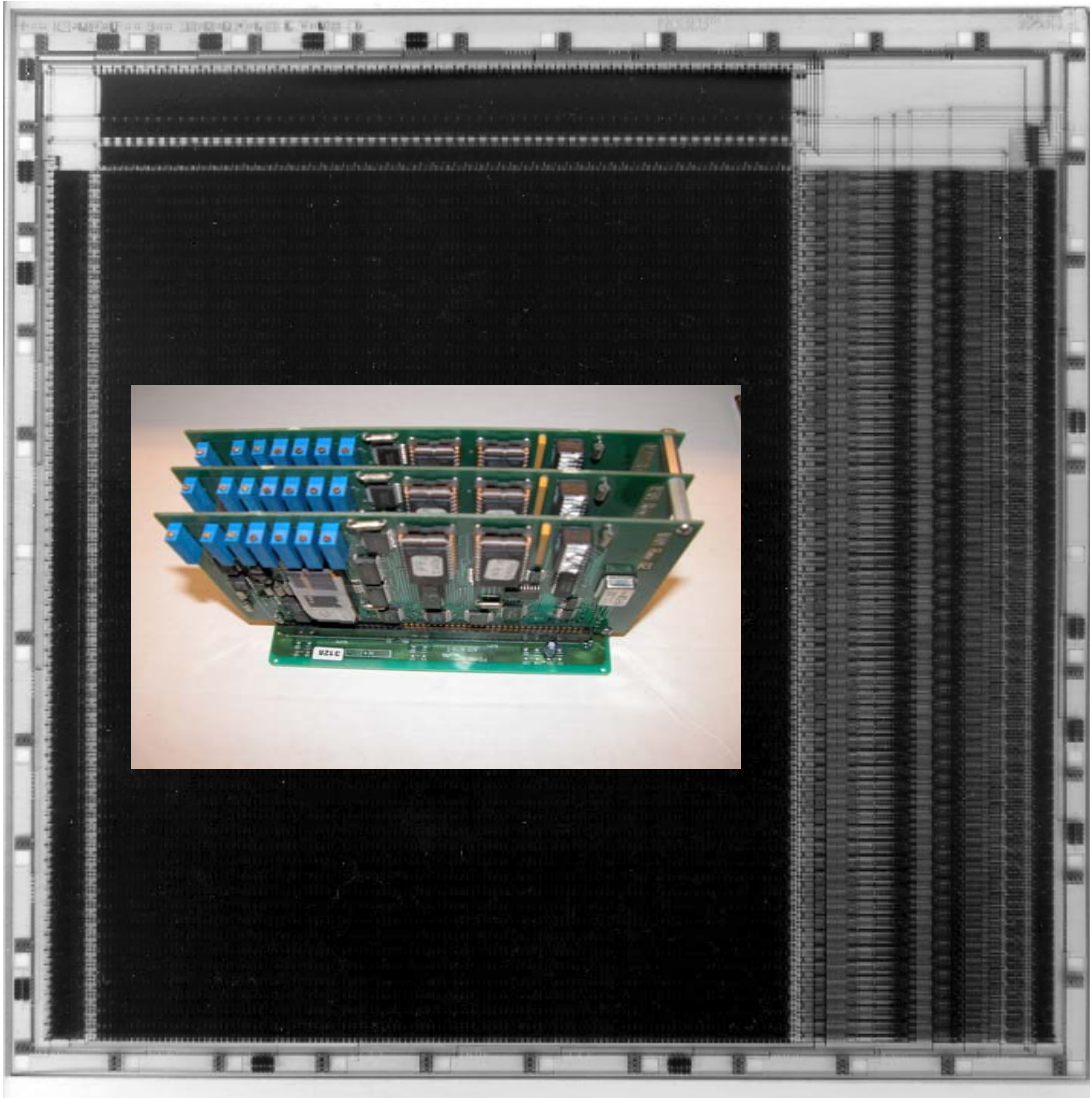


exploiting problem statistics!

500 MHz and therefore the energy per classification is **600 μJ**. The Pentium-Pro is worse, because it requires 50W at 150 MHz and more than 10000 cycles for a single pattern matching. In contrast, the total current in the WAM is: $(124 \times 116 \times 10)$ nA continuous bias current for the memory cells at 5V. Computation time is approximately $70 \mu s$ for a total energy per classification of approximately **100 nJ**. The power dissipation in

minimal complexity CMOS circuits

precision on demand architecture



1. Memory and processing are integrated in a single structure; this is analogous to the synapse in biology.
2. The system has an internal model that is related to the problem to be solved (*prior* knowledge). This is the template set of patterns to be classified.
3. The system is capable of learning i.e. templates can be changed to adapt to a different character set (different problem). This is done at the expense of storage capacity—we use a RAM based cell instead of a more compact ROM cell—.
4. The system processes information in a parallel and hierarchical fashion in a variable precision architecture. I.e. given the statistics of the problem, most of the computation is carried out with low precision (three or four bit) analog hardware.
5. The system is fault tolerant and gracefully degrades. The same structures that is used in the *precision-on-demand* architecture can also be used to reconfigure the system for defects in the fabrication process. The components of the chip that are worse matched can be disabled during operation.

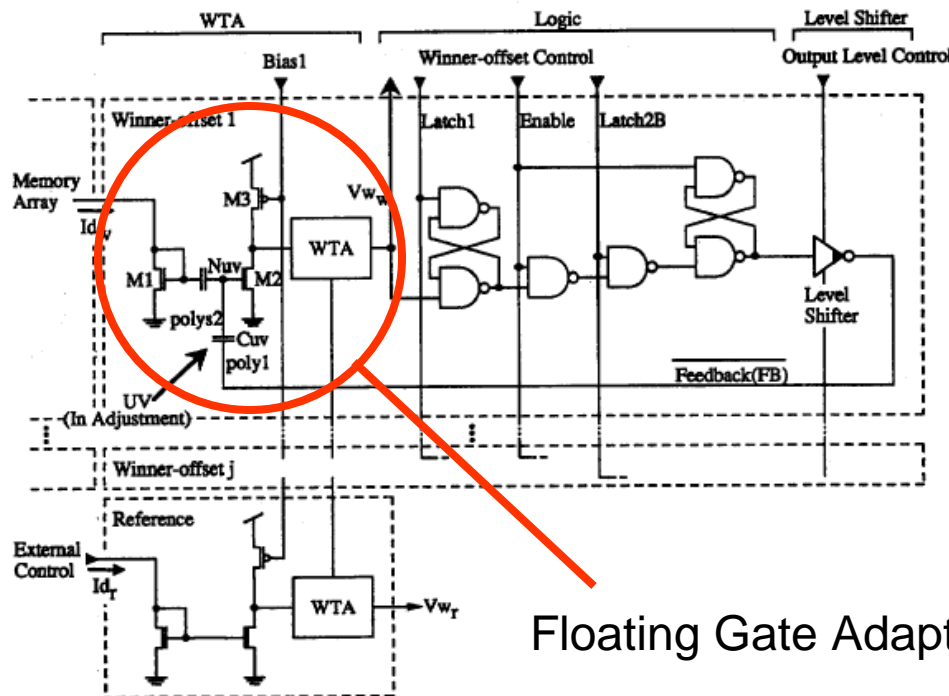
dealing with device mismatch ... again

ISCAS 94

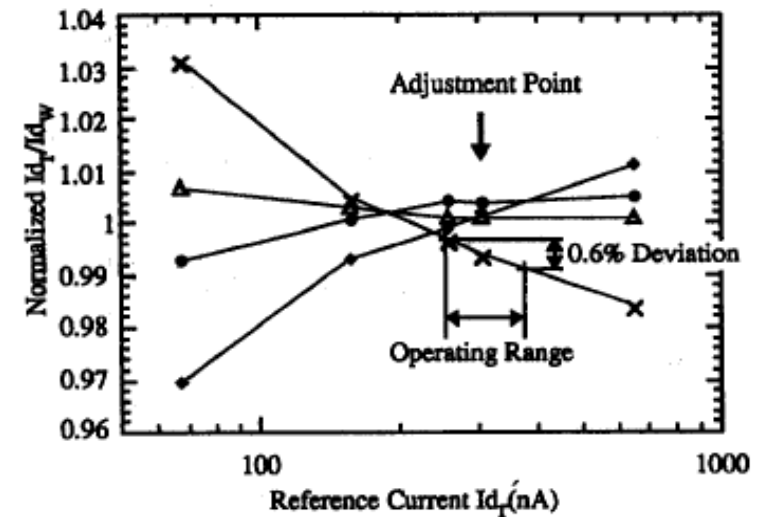
Storage Enhancement Techniques for Digital Memory Based, Analog Computational Engines

Hitoshi Miwa
 Device Development Center, Hitachi Ltd.
 2326 Imai, Ome-shi, Tokyo 198, Japan
 (+81) 428-32-1111
 miwa@ddc.hitachi.co.jp

Kewei Yang, Philippe O. Pouliquen,
 Nagendra Kumar, Andreas G. Andreou
 The Johns Hopkins University
 Department of Electrical and Computer Engineering
 Baltimore, MD 21218 USA
 (+1) 410-516-8361
 kewei@olumpus.ecc.jhu.edu



Floating Gate Adaptation

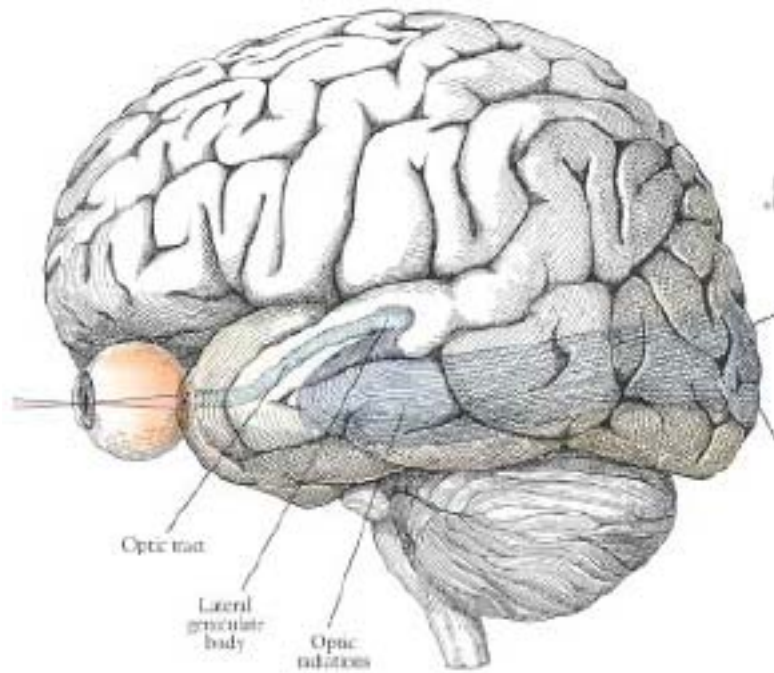


Part II: What is the real problem?

Physical world 3-dimensional
world's problems N-dimensional

natural and synthetic computing structures

The Brain



IBM Blue Gene/L

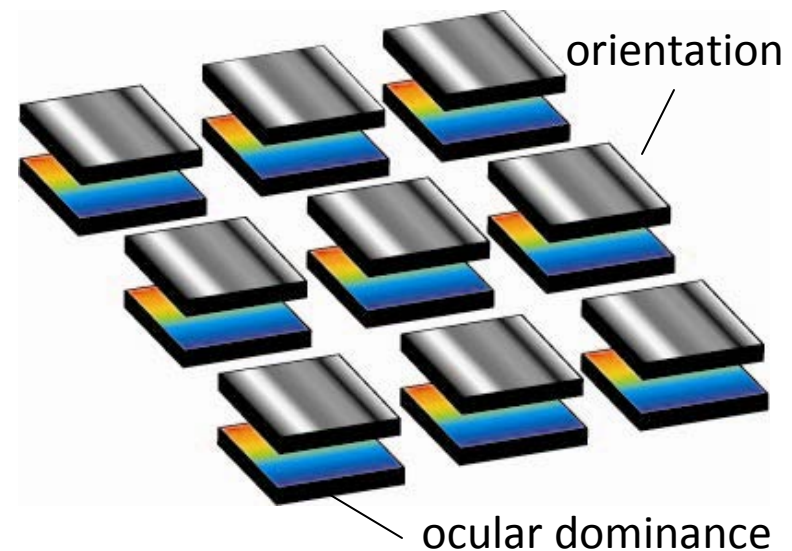


exist in three dimensional physical space but can deal with problems in hyper-dimensional spaces

visual representation of the world through cortical maps

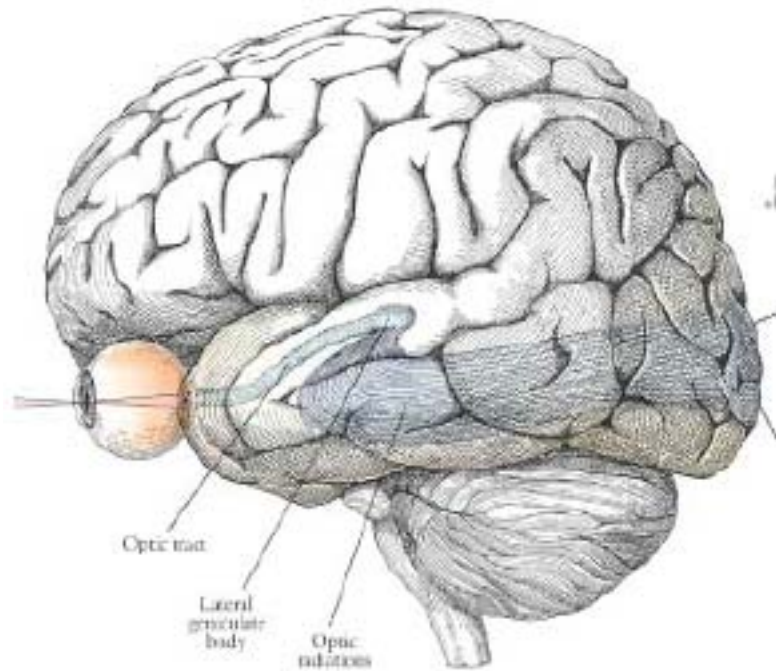
- Multiple stimulus modalities such as orientation, spatial frequency, ocular dominance are mapped into $2D+\delta$ patches on the surface of the cortex (V1)
- Conflicting constraints
 - **Maximize coverage** –every location in the physical space is mapped to all possible combinations of stimulus modalities-
 - **Minimize wiring length** and metabolic costs–neurons with similar stimulus response should exist in closed proximity on the cortical surface (smoothness of mapping).

The “ice-cube” model for stimulus representation in V1 (Hubel and Wiesel 1977) suggests stimulus modalities in orthogonal dimensions



natural and synthetic computing structures: another view

The Brain



15W

IBM Blue Gene/L

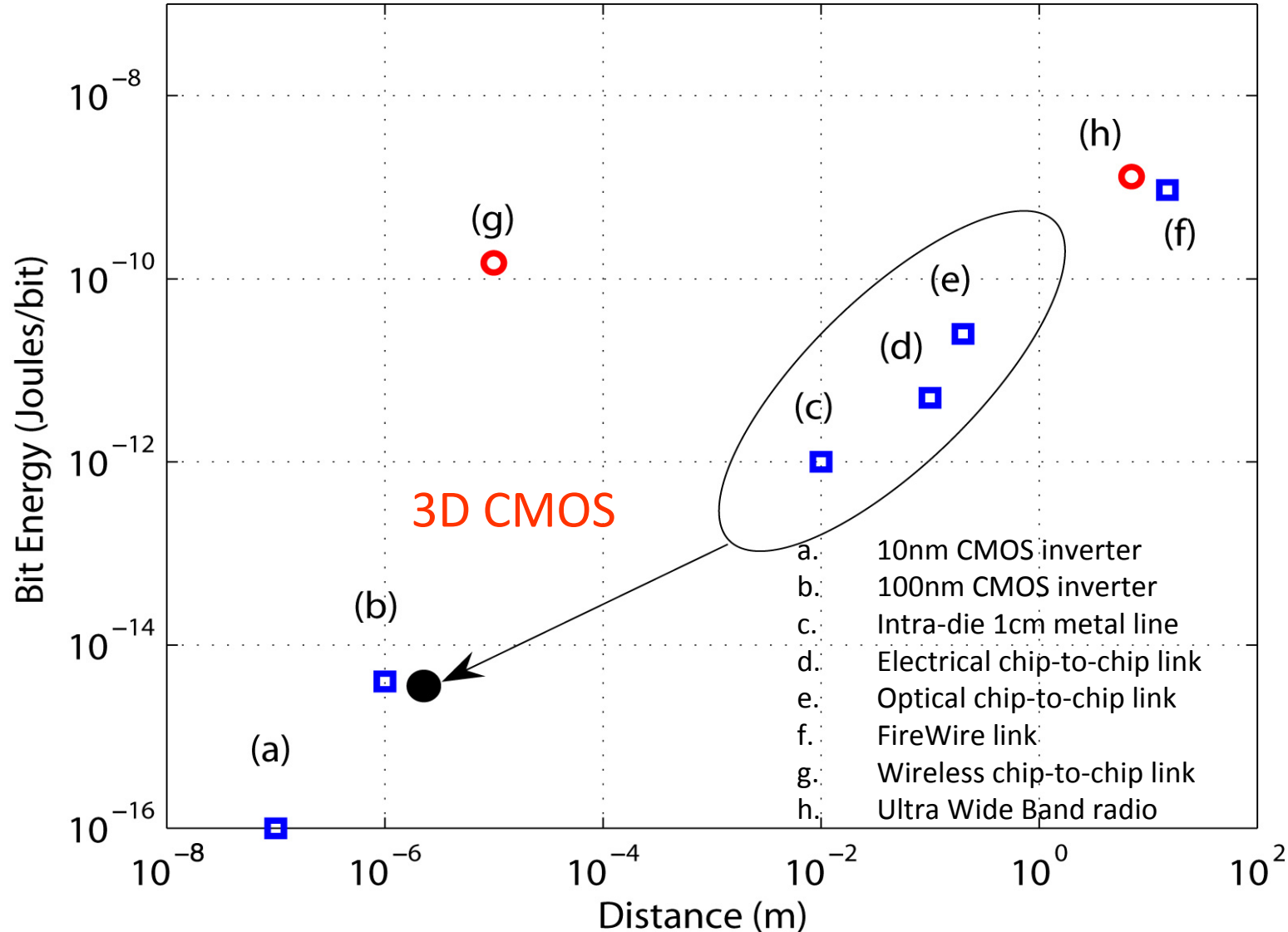


125 KW

5 racks



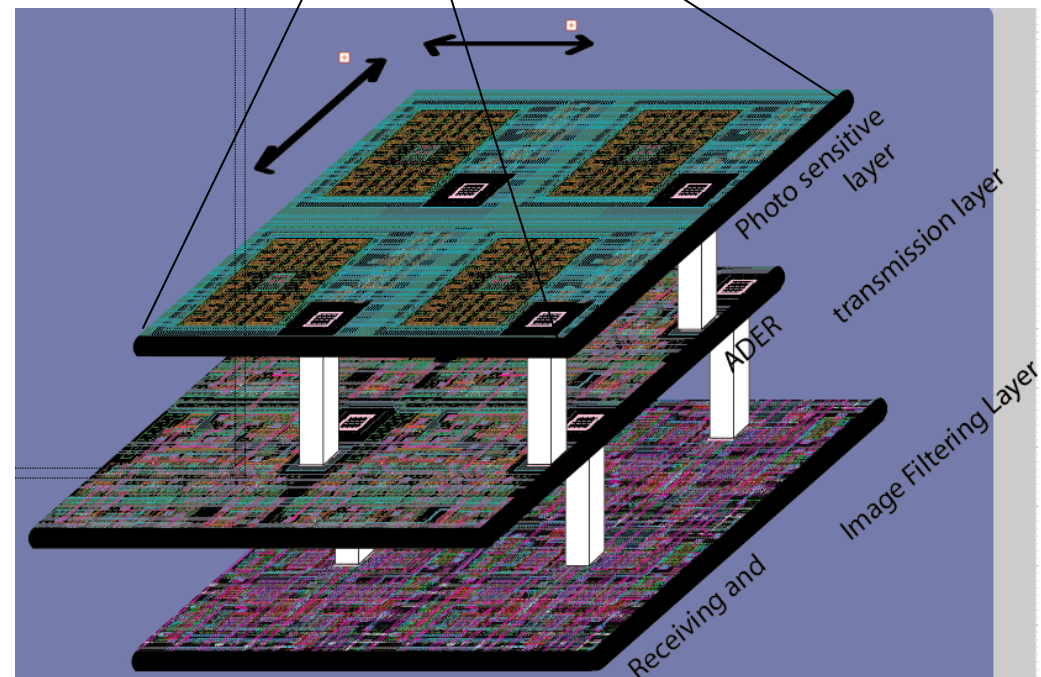
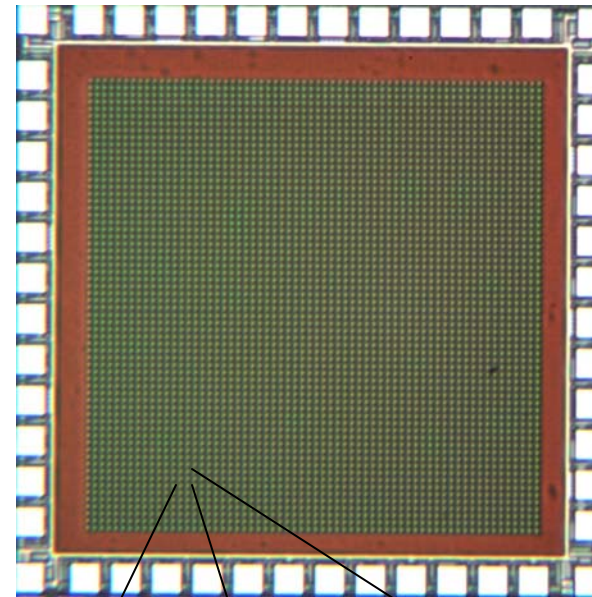
the energy costs of communication



M.A Marwick and A.G. Andreou, "Retinomorphic system design in three dimensional SOI-CMOS," *Proceedings of the 2006 IEEE International Symposium on Circuits and Systems*.

3D silicon cortex (a dynamical system approach)

Supply Voltage	1.5V
Technology	MITLL 0.18 μ m
3DL1	
Array size	64 \times 64
Spatial Processing	8 orient
Filter time	3-4ns
External Communications protocol (2 phase async.)	ADER
Internal Communications	4 phase asynchronous
Dynamic Range	6 bit
Minimum Frame Rate	300Hz
Maximum Frame Rate	20kHz



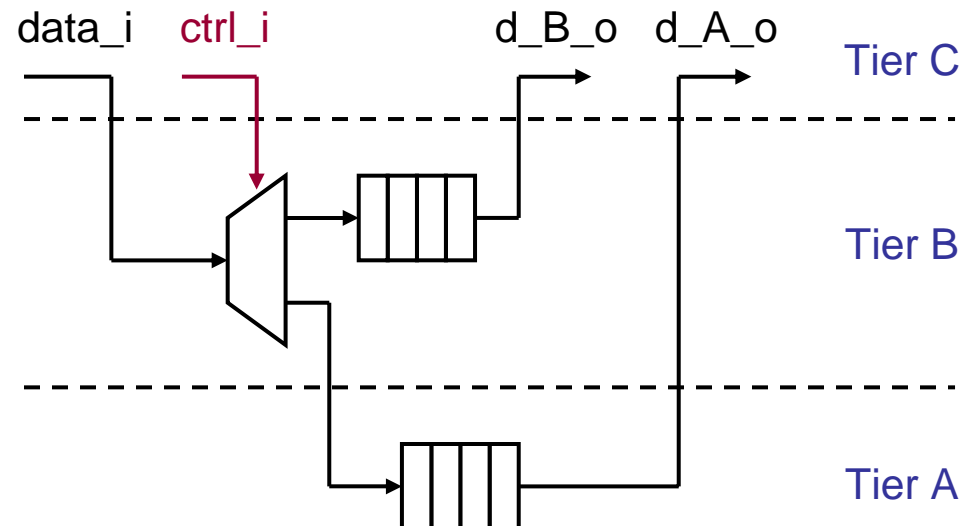
Towards Field Programmable Spiking Array in 3D CMOS

Results



Block Diagram

- 5 asynchronous handshake buffers in each path
 - (4 deep FIFO + 1 MUX)
- Utilizes all three tiers
 - Handshake between tiers



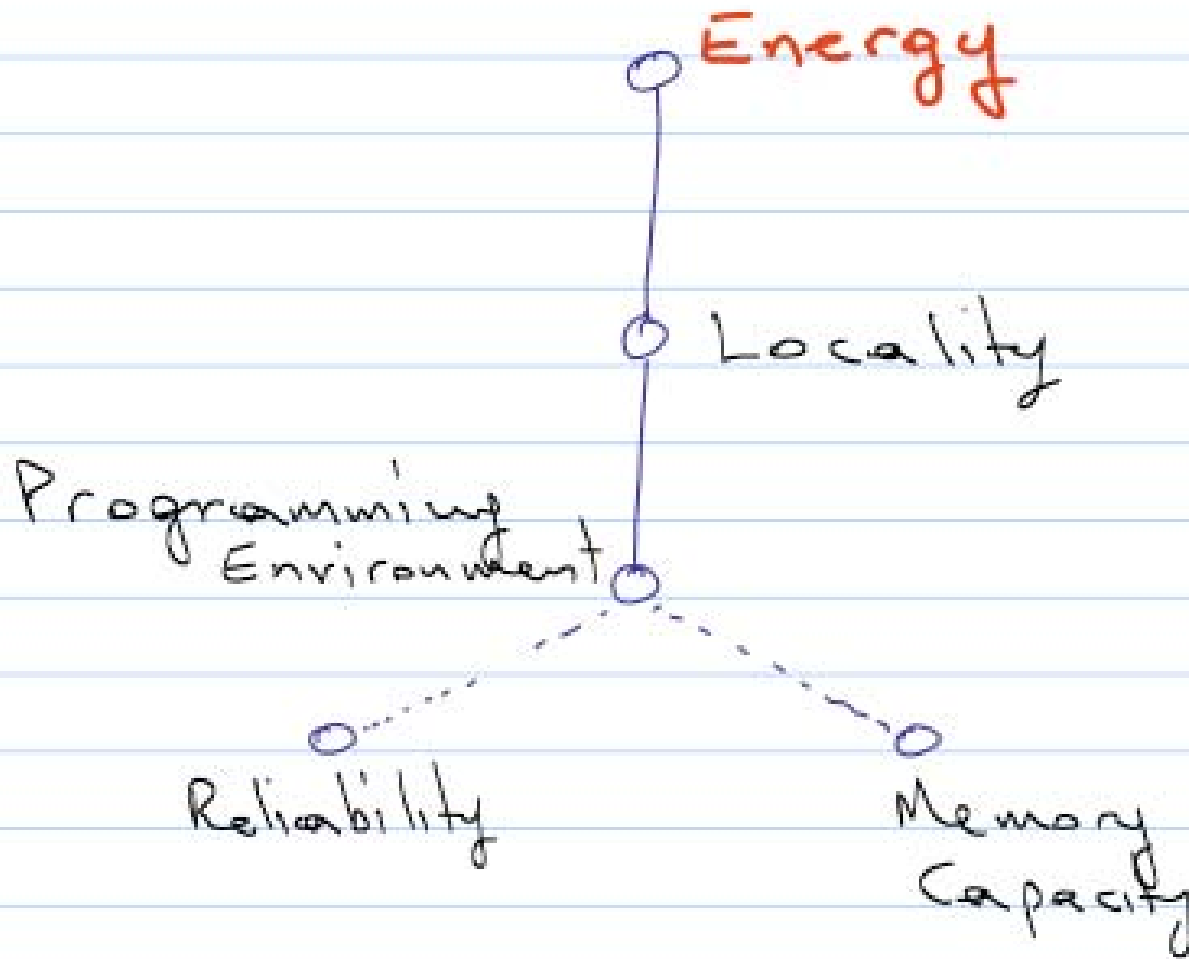
Part III: Final remarks

Report on the Second Kavli Futures Symposium

“Real Problems for Imagined Computers”

Jan 12-14 2009, Muelle, San Carlos, Costa Rica

Group B: Andreou, Dally, Roukes, Cornwell, Nair, Simon



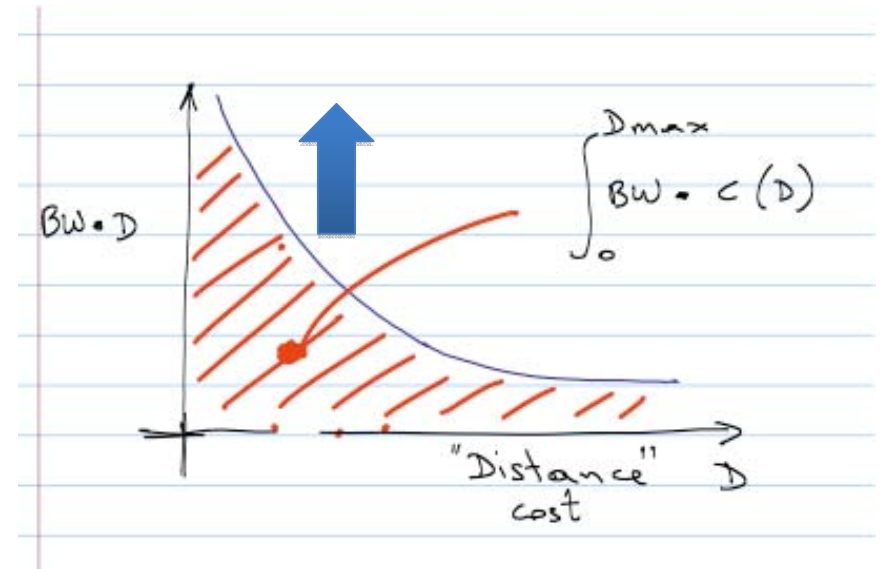
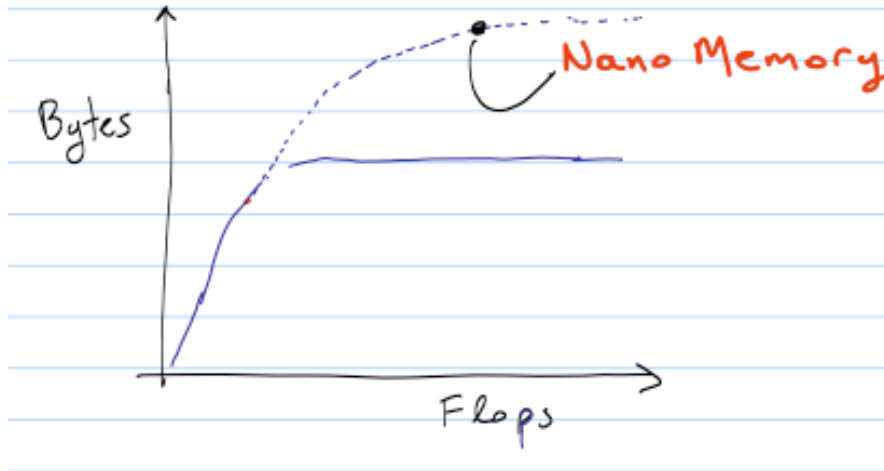
The Tyrannies

Commodity

Compatibility

Nano to the rescue

- Improving memory density
 - Makes higher Bytes per FLOP economically feasible
 - Improves the capacity at each level



On hardware, algorithms and architectures

Computational and energy efficiency can only be achieved through co-development of algorithms to application specific, reconfigurable architectures.

summary

- Early 80s: Carver Mead's neuromorphic manifesto towards new ways of computing inspired by biology

device physics based approach
exploit statistics of the problem
parallel distributed processing
processors in memory
adaptation
learning
analog circuits

- Late 90s: Some neuromorphic apostles took the wrong turn and got lost on the way.

- Today: Good news! The problems of the world have not yet been solved, and the apostles are 15 years older and hopefully wiser!