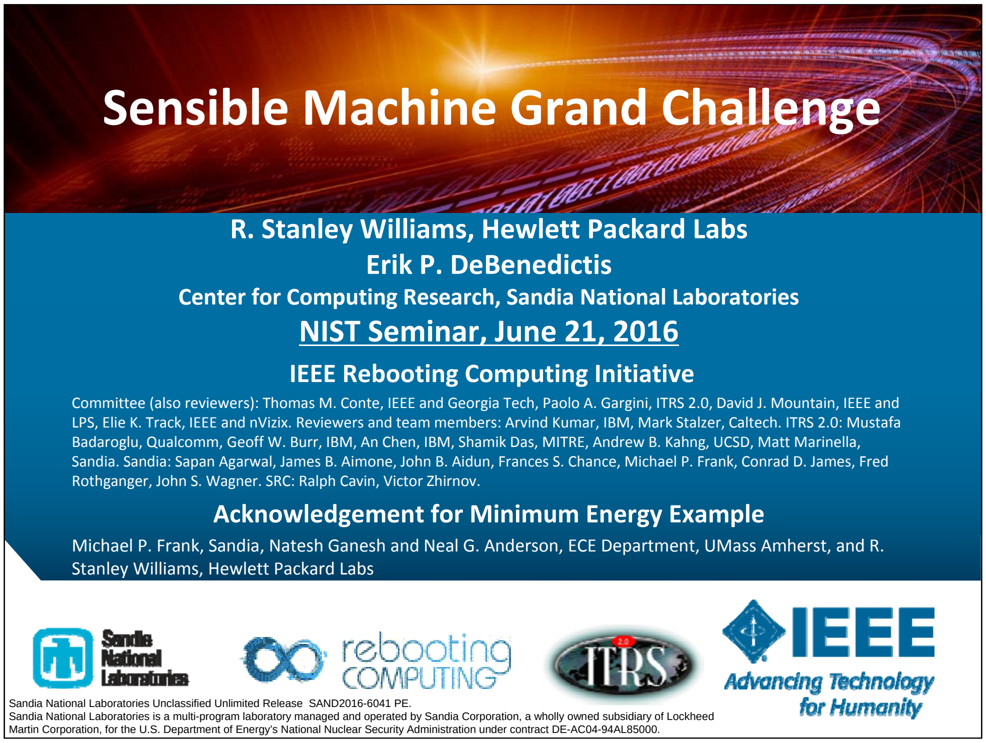


Sensible Machine Grand Challenge

R. Stanley Williams, Hewlett Packard Labs
Erik P. DeBenedictis
Center for Computing Research, Sandia National Laboratories

NIST Seminar, June 21, 2016

IEEE Rebooting Computing Initiative

Committee (also reviewers): Thomas M. Conte, IEEE and Georgia Tech, Paolo A. Gargini, ITRS 2.0, David J. Mountain, IEEE and LPS, Elie K. Track, IEEE and nVizix. Reviewers and team members: Arvind Kumar, IBM, Mark Stalzer, Caltech. ITRS 2.0: Mustafa Badaroglu, Qualcomm, Geoff W. Burr, IBM, An Chen, IBM, Shamik Das, MITRE, Andrew B. Kahng, UCSD, Matt Marinella, Sandia. Sandia: Sapan Agarwal, James B. Aimone, John B. Aidun, Frances S. Chance, Michael P. Frank, Conrad D. James, Fred Rothganger, John S. Wagner. SRC: Ralph Cavin, Victor Zhirnov.

Acknowledgement for Minimum Energy Example

Michael P. Frank, Sandia, Natesh Ganesh and Neal G. Anderson, ECE Department, UMass Amherst, and R. Stanley Williams, Hewlett Packard Labs

Sandia National Laboratories Unclassified Unlimited Release SAND2016-6041 PE.
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories
rebooting COMPUTING
ITRS 2.0
IEEE
Advancing Technology for Humanity

The Sensible Machine project has been adopted by IEEE Rebooting Computing and the ITRS 2.0 (now IRDS) activity.

The speaker is with Sandia National Laboratories.

Note

- Stan Williams will give a talk in September
- Stan and I coordinated for this talk
- This talk leaves out much of the physical science and neuromorphic content that Stan is likely to stress

Scope of Talk – Outline

High level, non-technical:

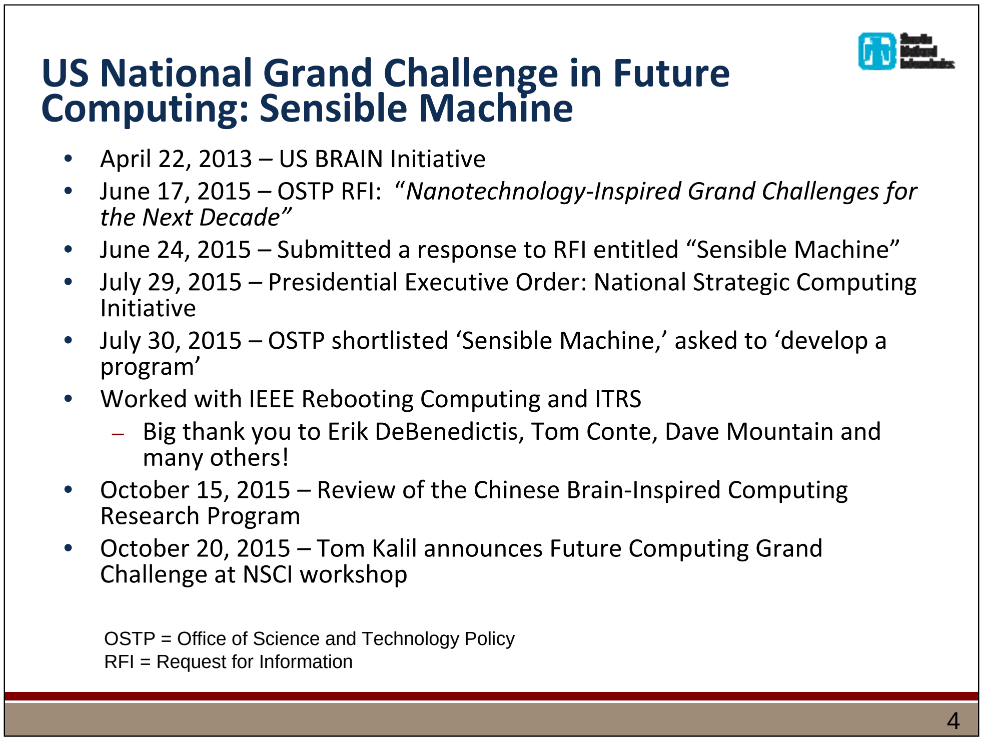
Proposal of a nanotechnology Grand Challenge and its acceptance by the US Government. It is a challenge, not the solution, so we avoid favoring any technical approach

High level, technical:

New theory on the limit of computation showing that many orders of magnitude increase in energy efficiency is possible, but more than the von Neumann architecture will be necessary, such as learning machines

Possible technical validation:

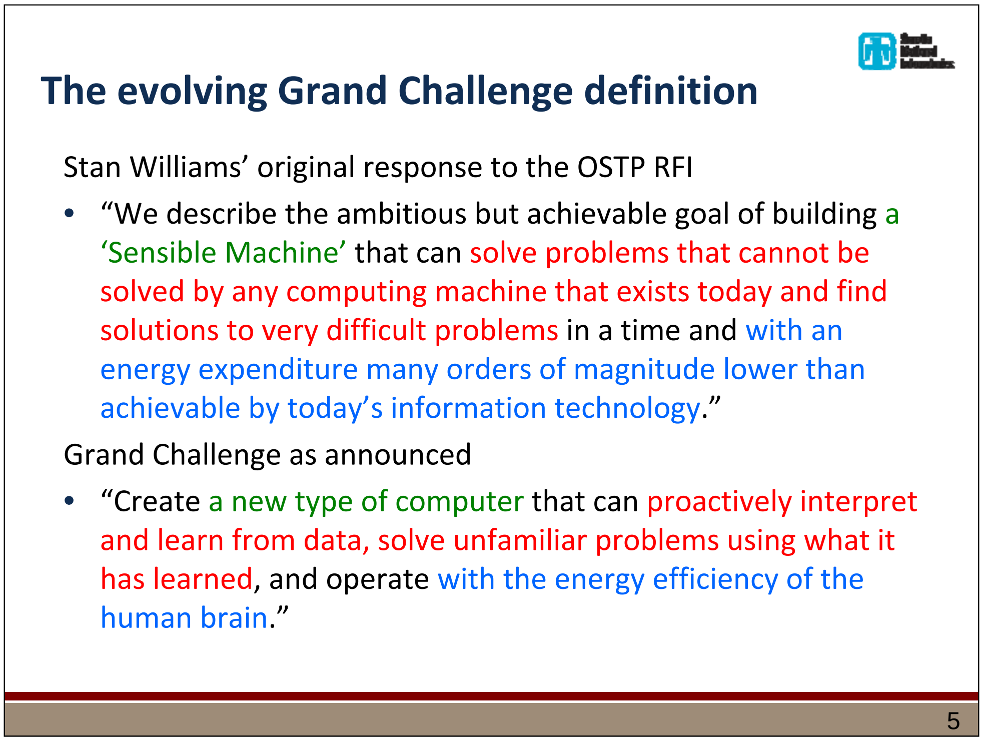
Can the theoretical limits be approached with a real system, even if impractical? As a starting point, the talk will review a superconducting nSQUID circuit that has been built and tested in a similar configuration



US National Grand Challenge in Future Computing: Sensible Machine

- April 22, 2013 – US BRAIN Initiative
- June 17, 2015 – OSTP RFI: *“Nanotechnology-Inspired Grand Challenges for the Next Decade”*
- June 24, 2015 – Submitted a response to RFI entitled “Sensible Machine”
- July 29, 2015 – Presidential Executive Order: National Strategic Computing Initiative
- July 30, 2015 – OSTP shortlisted ‘Sensible Machine,’ asked to ‘develop a program’
- Worked with IEEE Rebooting Computing and ITRS
  - Big thank you to Erik DeBenedictis, Tom Conte, Dave Mountain and many others!
- October 15, 2015 – Review of the Chinese Brain-Inspired Computing Research Program
- October 20, 2015 – Tom Kalil announces Future Computing Grand Challenge at NSCI workshop

OSTP = Office of Science and Technology Policy
RFI = Request for Information



The evolving Grand Challenge definition

Stan Williams’ original response to the OSTP RFI

- “We describe the ambitious but achievable goal of building a ‘Sensible Machine’ that can solve problems that cannot be solved by any computing machine that exists today and find solutions to very difficult problems in a time and with an energy expenditure many orders of magnitude lower than achievable by today’s information technology.”

Grand Challenge as announced

- “Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.”

Additional NSCI detail motivating this talk

“While it continues to be a national priority to advance conventional digital computing—which has been the engine of the information technology revolution—current technology falls far short of the human brain in terms of both the brain’s sensing and problem-solving abilities and its low power consumption. Many experts predict that fundamental physical limitations will prevent transistor technology from ever matching these twin characteristics. We are therefore challenging the nanotechnology and computer science communities to look beyond the decades-old approach to computing based on the Von Neumann architecture as implemented with transistor-based processors, and chart a new path that will continue the rapid pace of innovation beyond the next decade.”

- Target problem solving and digital computers using devices other than transistors in non von Neumann architectures

Structure of a US Nanotechnology-Inspired Future Computing Program

1. Devices and Materials – *in situ* and *in operando* test and measurement
   - Most likely materials will be adopted from Non-Volatile Memory
   - Already more than a decade of experience in commercial grade foundries
   - One promising path forward utilizes electronic synapses and axons
2. Chip Processing and Integration – Full Service Back End of Line on CMOS
   - What facilities are available for general use in the US?
   - DoE Nanoscale Science Research Centers (NSRCs) – e.g. CINT
   - Fabbing CMOS in Asia and sending wafers to Europe for BEOL?
3. Chip Design – System-on-Chip: Accelerators, Learning and Controllers
   - Compatible with standard processors, memory and data bus

Structure of a US Nanotechnology-Inspired Future Computing Program

4. System Software, Algorithms & Apps – Make it Programmable/Adaptable
   - At least two thirds of the effort will be in firmware and software
   - Will this require an open source model?
5. Simulation of Computational Models and Systems
   - Develop a suite of tools of compact models and detailed analyses
6. Architecture of the Brain and Relation to Computing and Learning
   - Theories of Mind: Albus, Eliasmith, Grossberg, Mead, many others
7. Connect Theory of Computation with Neuroscience and Nonlinear Dynamics
   - What is the computational paradigm? What do spikes really do?
   - Boolean, CNN, Bayesian Inference, Energy-Based Models, Markov Chains

Scope of Talk – Outline

High level, non-technical:

Proposal of a nanotechnology Grand Challenge and its acceptance by the US Government. It is a challenge, not the solution, so we avoid favoring any technical approach

High level, technical:

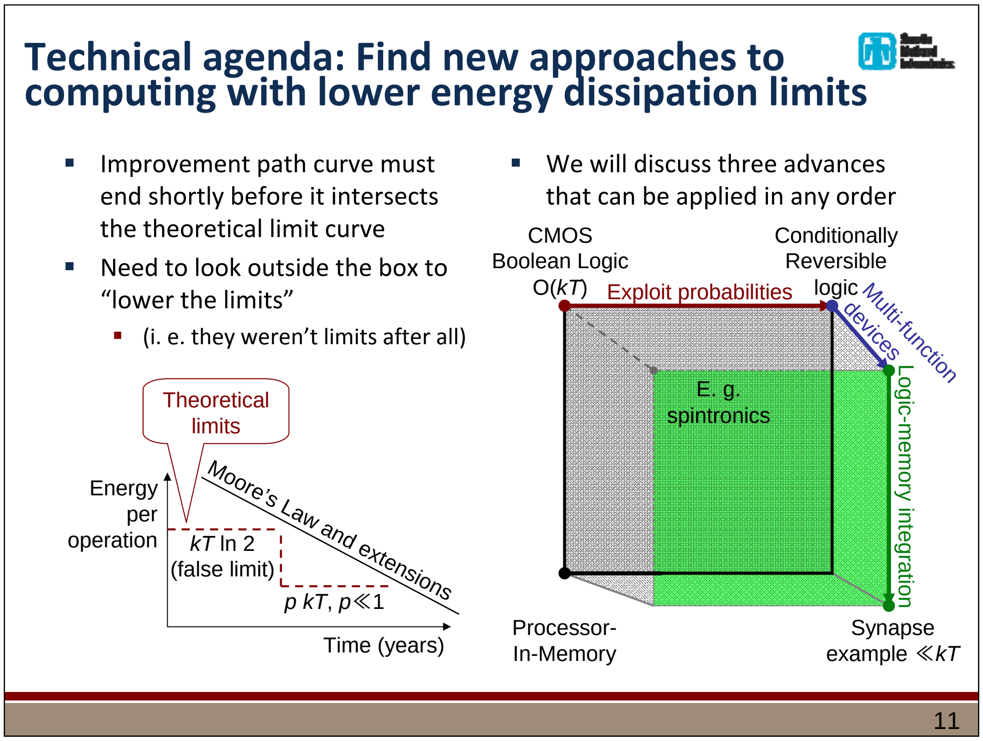
New theory on the limit of computation showing that many orders of magnitude increase in energy efficiency is possible, but more than the von Neumann architecture will be necessary, such as learning machines

Possible technical validation:

Can the theoretical limits be approached with a real system, even if impractical? As a starting point, the talk will review a superconducting nSQUID circuit that has been built and tested in a similar configuration

OSTP's advice leads us to *kT* limits

- Clock rate not scaling anymore
  - Clock rate scaling in products slowed due to excessive energy consumption
  - Also, OSTP said to worry about energy efficiency not speed
- Density continues to scale (so what is the problem?)
  - Memory density scales just fine
  - Logic density could scale except for excess heat dissipation due to leakage current. So microprocessors are built with a lot of memory
- Leakage current due to kT/q subthreshold slope limiting reduction in power supply voltage
- Beyond CMOS transistors are research topics
    - TFETs, piezotronic transistors, etc.
  - Positive: Lower power supply voltage without leakage
  - Negative: $p_{\text{error}} = \exp(-e_{\text{signal}} / kT)$ errors
- *kT* is the root of many of today's limits



Technical agenda: Find new approaches to computing with lower energy dissipation limits

- Improvement path curve must end shortly before it intersects the theoretical limit curve
- Need to look outside the box to “lower the limits”
  - (i. e. they weren’t limits after all)

- We will discuss three advances that can be applied in any order

Narrative: When Moore’s Law was proposed in the 1960s, the ultimate limits of computing were understood to be way way out in the future. Von Neumann understood some limit around kT, but Landauer worked out a strategy for computing minimum energy. While Landauer never called this a “limit,” society has interpreted it as such. Today’s logic roadmaps predict industrial progress to about 10,000 kT per CMOS-like gate. The fact that the theoretical limit is so close now creates an argument that “Moore’s Law is ending.”

To continue Moore’s Law or some similar rate of improvement requires effectively lowering the limits, as illustrated on the left. Limits cannot be reduced if they are real limits rooted in physical law. However, it may be possible to find new approaches to computing that have a lower energy limit than CMOS-type circuits when solving the same problem.

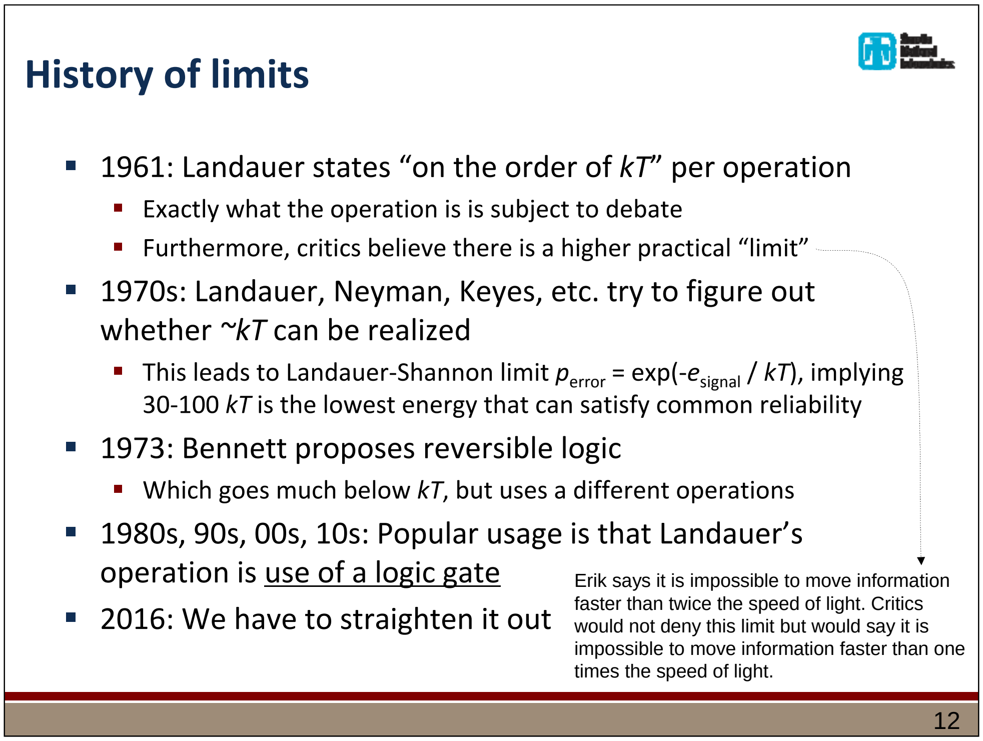
The right shows a “cubic” roadmap of sorts. Industry uses CMOS Boolean logic (upper left) which has an O(kT) limit. This talk will describe three approaches for reducing minimum energy, each corresponding to a dimension of the cube:

Horizontal (red): Probabilities may be exploited. There are two parts to this activity. First, some logical operations fundamentally dissipate heat (e. g. NAND). However, dissipation can be reduced if the system can be designed so the fraction of the time it does these heat-dissipating operations is lower and the fraction of time it does nothing is higher. Second, some logical operations do not fundamentally dissipate heat (e. g. the reversible Toffoli gate)..

Into page (blue): Without reducing energy per device, the energy of a system can be reduced by making the devices do more. For example, replacing four Boolean NAND gates of four transistors each with a single device would reduce energy consumption 16 fold even if all the devices dissipated the same energy.

Down (green): The previous ideas may not help the processor component of a von Neumann processor very much. If the gates in a CPU or ALU are only active a small percentage of the time, the logic could probably should be redesigned because it is inefficient. However, a memory cell that is idle most of the time is not considered wasteful because it is storing data. In addition, logic-memory integration reduces the need to move data over long distances. This data movement consumes energy.

The ideas on this slide establish a connection to brain-inspired computing: This slide deck will use a type of synapse as an example, yet a synapse uses all three of the features above.

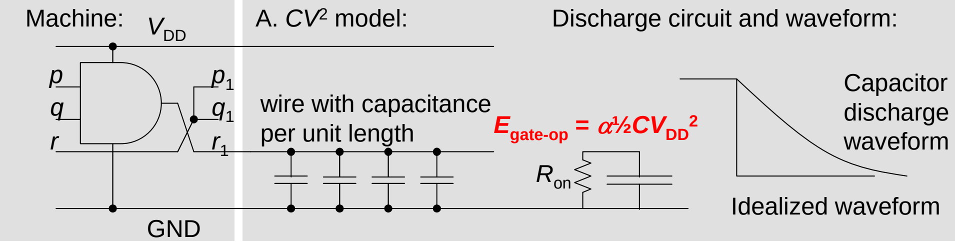


History of limits

- 1961: Landauer states “on the order of *kT*” per operation
  - Exactly what the operation is is subject to debate
  - Furthermore, critics believe there is a higher practical “limit”
- 1970s: Landauer, Neyman, Keyes, etc. try to figure out whether ~*kT* can be realized
  - This leads to Landauer-Shannon limit p_{error} = exp(-e_{signal} / *kT*), implying 30-100 *kT* is the lowest energy that can satisfy common reliability
- 1973: Bennett proposes reversible logic
  - Which goes much below *kT*, but uses a different operations
- 1980s, 90s, 00s, 10s: Popular usage is that Landauer’s operation is <u>use of a logic gate</u>
- 2016: We have to straighten it out

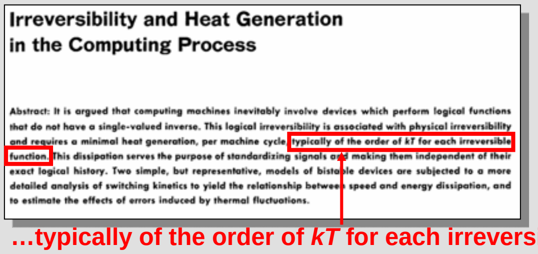
Erik says it is impossible to move information faster than twice the speed of light. Critics would not deny this limit but would say it is impossible to move information faster than one times the speed of light.

Models of computer energy dissipation


Machine:
V_{DD}
p
q
r
p_1
q_1
r_1
GND
A. CV^2 model:
wire with capacitance per unit length
Discharge circuit and waveform:
R_{on}
Capacitor discharge waveform
Idealized waveform

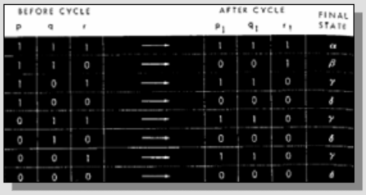
$E_{gate\text{-}op} = \alpha ½CV_{DD}^2$

B. Information erasure model [Landauer 61]:



Irreversibility and Heat Generation in the Computing Process

Abstract: It is argued that computing machines inevitably involve devices which perform logical functions that do not have a single-valued inverse. This logical irreversibility is associated with physical irreversibility and requires a minimal heat generation, per machine cycle, typically of the order of *kT* for each irreversible function. This dissipation serves the purpose of standardizing signals and making them independent of their exact logical history. Two simple, but representative, models of bistable devices are subjected to a more detailed analysis of switching kinetics to yield the relationship between speed and energy dissipation, and to estimate the effects of errors induced by thermal fluctuations.



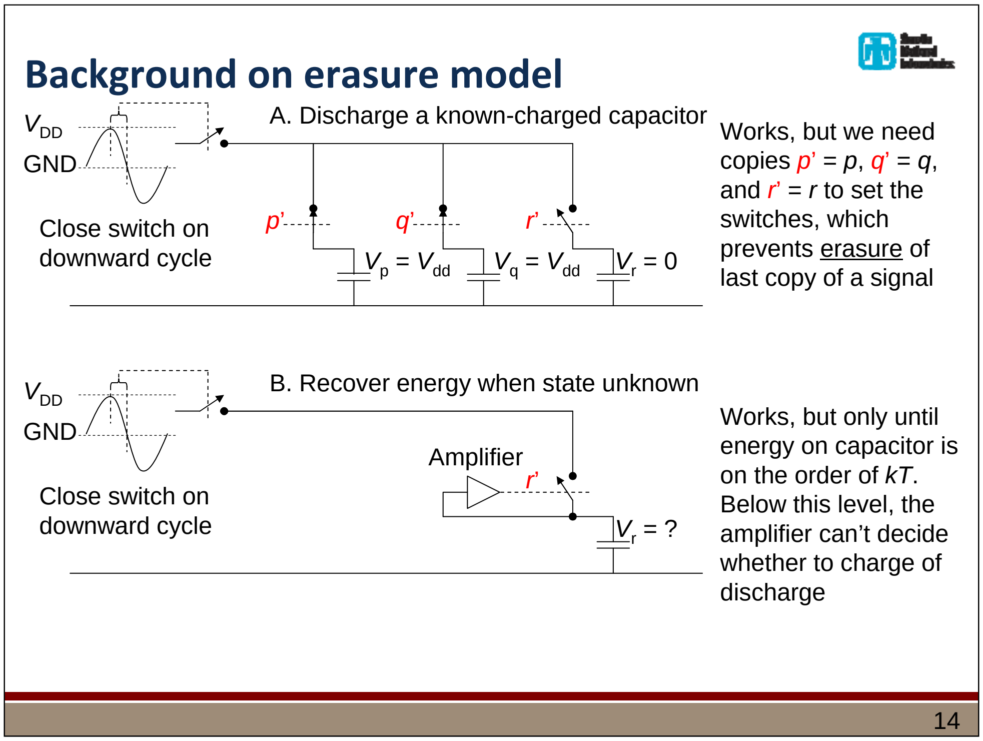
BEFORE CYCLE				AFTER CYCLE			FINAL STATE
p	q	r		p_1	q_1	r_1	
1	1	1	→	1	1	1	α
1	1	0	→	0	0	1	β
1	0	1	→	1	1	0	γ
1	0	0	→	0	0	0	δ
0	1	1	→	1	1	0	γ
0	1	0	→	0	0	0	δ
0	0	1	→	1	1	0	γ
0	0	0	→	0	0	0	δ

…typically of the order of *kT* for each irreversible function

[Landauer 61] Landauer, Rolf. "Irreversibility and heat generation in the computing process." *IBM journal of research and development* 5.3 (1961): 183-191.

See also http://rebootingcomputing.ieee.org/images/files/pdf/RCS4DeBenedictisposter.pdf

This is a tutorial-type viewgraph for explaining limits.



Background on erasure model

V_{DD}
GND
Close switch on downward cycle
A. Discharge a known-charged capacitor
p'
q'
r'
$V_p = V_{dd}$
$V_q = V_{dd}$
$V_r = 0$

Works, but we need copies *p'* = *p*, *q'* = *q*, and *r'* = *r* to set the switches, which prevents erasure of last copy of a signal

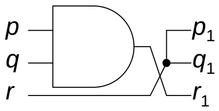
V_{DD}
GND
Close switch on downward cycle
B. Recover energy when state unknown
Amplifier
r'
V_r = ?

Works, but only until energy on capacitor is on the order of *kT*. Below this level, the amplifier can't decide whether to charge of discharge

This is a tutorial-type viewgraph for explaining limits.

Landauer’s method from the paper’s example

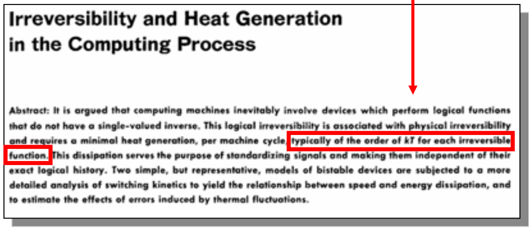
System:



prob	p	q	r		p1	q1	r1	Si (k's)	State	Sf (k's)
0.125	1	1	1	→	1	1	1	0.25993	α	0.25993
0.125	1	1	0	→	0	0	1	0.25993	β	0.25993
0.125	1	0	1	→	1	1	0	0.25993	γ	0.367811
0.125	1	0	0	→	0	0	0	0.25993	δ	0.367811
0.125	0	1	1	→	1	1	0	0.25993	γ	0
0.125	0	1	0	→	0	0	0	0.25993	δ	0
0.125	0	0	1	→	1	1	0	0.25993	γ	0
0.125	0	0	0	→	0	0	0	0.25993	δ	0
								2.079442	Sf (k's)	1.255482
									Si-Sf (k's)	0.823959

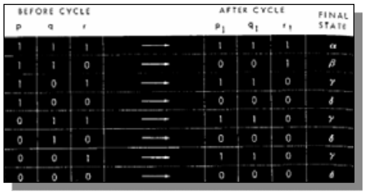
Typically of the order of *kT* for each irreversible function

From source:



Irreversibility and Heat Generation in the Computing Process

Abstract: It is argued that computing machines inevitably involve devices which perform logical functions that do not have a single-valued inverse. This logical irreversibility is associated with physical irreversibility and requires a minimal heat generation, per machine cycle, typically of the order of kT for each irreversible function. This dissipation serves the purpose of standardizing signals and making them independent of their exact logical history. Two simple, but representative, models of bistable devices are subjected to a more detailed analysis of switching kinetics to yield the relationship between speed and energy dissipation, and to estimate the effects of errors induced by thermal fluctuations.



BEFORE CYCLE				AFTER CYCLE			FINAL STATE
p	q	r		p_1	q_1	r_1	
1	1	1	→	1	1	1	α
1	1	0	→	0	0	1	β
1	0	1	→	1	1	0	γ
1	0	0	→	0	0	0	δ
0	1	1	→	1	1	0	γ
0	1	0	→	0	0	0	δ
0	0	1	→	1	1	0	γ
0	0	0	→	0	0	0	δ

[Landauer 61] Landauer, Rolf. "Irreversibility and heat generation in the computing process." *IBM journal of research and development* 5.3 (1961): 183-191.

Narrative:

The AND gate with p, q, r is the “machine” analyzed in Landauer’s paper: An AND gate with a wire that goes along for the ride.

The white on black spreadsheets are shown in the visual notation of Landauer, except we added additional columns to make points about the arithmetic.

Landauer used equal probabilities of 1/8 for each input (orange) – which was reasonable given that IBM produced gates for general use. The 1/8 probability leaves the most flexibility to the engineer/buyer/user.

We augmented the original table with additional columns for Si and Sf and the arithmetic. The lower four states (rows) are merged into the top for rows; we put a zero in the Sf column to let the spreadsheet give a consistent tally of the column.

The example yields .83 kT, which is approximate kT (note: there is an arithmetic error in the paper). Thus justifies the statement of “the order of kT” in the paper’s abstract.

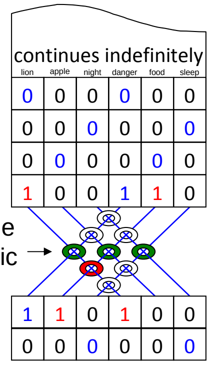
Backup: Details

- Each input combination gets a row
  - Each input combination *k* has probability p_k, p_k's summing to 1
  - S_i (i for input) is the sum of all $p_k \log p_k$'s
- Each unique output combination is analyzed
  - Rows merge if the machine produces the same output
  - Each output combination *k* has probability p_k, p_k's summing to 1
  - S_f (f for final) is the sum of all $p_k \log p_k$'s
- Minimum energy is $S_\mathrm{i} – S_\mathrm{f}$
- Notes
  - Inputs that don't merge do not raise minimum energy
  - Inputs that merge raise minimum energy based on their probability

Example: a learning machine

This “learning machine” example exceeds energy efficiency limits of Boolean logic. The learning machine monitors the environment for knowledge, yet usually just verifies that it has learned what it needs to know. Say “causes” (lion, apple, and night) and “effects” (danger, food, and sleep) have value 1.

Example input:
{lion, danger } {apple, food } {night, sleep } {lion, danger } {apple, food } {night, sleep } {lion, danger } {apple, food } {night, sleep } {lion, danger, food } {apple, food } {night, sleep } { lion, danger } {lion, danger }



continues indefinitely

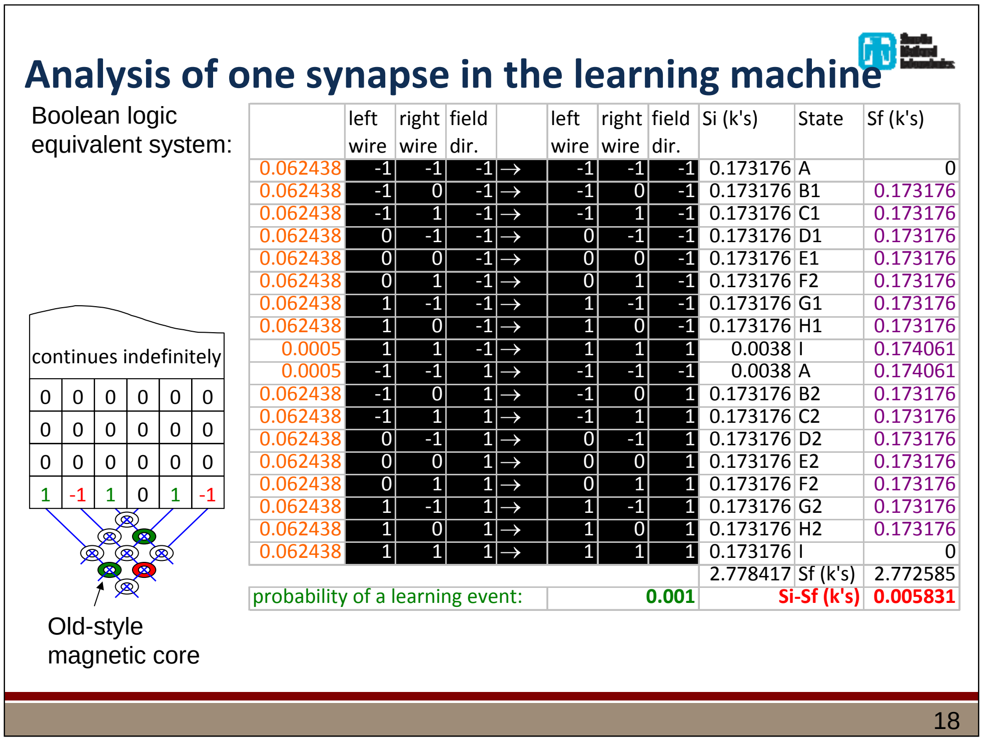
lion	apple	night	danger	food	sleep
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	0	1	1	0

Old-style magnetic cores →

1	1	0	1	0	0
0	0	0	0	0	0

Functional example:
Machine continuously monitors environment for {1, 1} or {-1, -1} pairs and remembers them in state of a magnetic core. Theoretically, there is no need for energy consumption unless state changes.

Signals create currents;
core flips a ±1.5



Analysis of one synapse in the learning machine

Boolean logic equivalent system:

	left wire	right wire	field dir.		left wire	right wire	field dir.	Si (k's)	State	Sf (k's)
0.062438	-1	-1	-1	→	-1	-1	-1	0.173176	A	0
0.062438	-1	0	-1	→	-1	0	-1	0.173176	B1	0.173176
0.062438	-1	1	-1	→	-1	1	-1	0.173176	C1	0.173176
0.062438	0	-1	-1	→	0	-1	-1	0.173176	D1	0.173176
0.062438	0	0	-1	→	0	0	-1	0.173176	E1	0.173176
0.062438	0	1	-1	→	0	1	-1	0.173176	F2	0.173176
0.062438	1	-1	-1	→	1	-1	-1	0.173176	G1	0.173176
0.062438	1	0	-1	→	1	0	-1	0.173176	H1	0.173176
0.0005	1	1	-1	→	1	1	1	0.0038	I	0.174061
0.0005	-1	-1	1	→	-1	-1	-1	0.0038	A	0.174061
0.062438	-1	0	1	→	-1	0	1	0.173176	B2	0.173176
0.062438	-1	1	1	→	-1	1	1	0.173176	C2	0.173176
0.062438	0	-1	1	→	0	-1	1	0.173176	D2	0.173176
0.062438	0	0	1	→	0	0	1	0.173176	E2	0.173176
0.062438	0	1	1	→	0	1	1	0.173176	F2	0.173176
0.062438	1	-1	1	→	1	-1	1	0.173176	G2	0.173176
0.062438	1	0	1	→	1	0	1	0.173176	H2	0.173176
0.062438	1	1	1	→	1	1	1	0.173176	I	0
								2.778417	Sf (k's)	2.772585
probability of a learning event:							0.001		Si-Sf (k's)	0.005831

continues indefinitely

0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	-1	1	0	1	-1

Old-style magnetic core

The learning machine circuit has two trit or 3-state inputs, making the table larger.

The underlying Excel spreadsheet allows the user to type in the probability of a learning event and uses that value to fill in the orange probabilities on the left. This is critical to the result.

Note that only states A and I merge; the other states do not merge and contribute the same amount to Si and Sf (i. e. they contribute nothing to the dissipation).

The result is .005 kT, much lower than Landauer’s example.

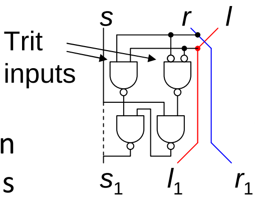
Note: the result is not fundamental (like kT ln 2) but is roughly proportional to the “probability of a learning event.”

Why is the “limit” so low? (I) Probabilities

- The “limit” depends where you look in Landauer’s article
  - Word limit does not appear in the article
  - “on the order of *kT*” (abstract) *kT* ln 2 per bit erased (body)
  - 0.82 *kT* or 1.18 *kT* (he made a math error) in the example
- Actually, the “limit” assumes
  - The system is in thermodynamic equilibrium (p = .125)
  - Input bits have a full bit of information, $p_0 = p_1 = 0.5$
- However, the body of the paper very clearly talks about the probabilities of input states (or combinations)
- The example exploits the fact that synapses usually <u>verify that they have learned what the need to know</u> and actually change state with low probability

Why is the “limit” so low? (II)
Aggregation principle

- The Landauer’s minimum energy stays the same or rises when a function is broken up into pieces – it cannot decrease
  - If splitting into pieces produces intermediate variables that have to be erased, minimum energy will increase
  - If the pieces digitally restore signals, they can’t be aggregated
- A single magnetic core implements the 4-gate sub circuit →



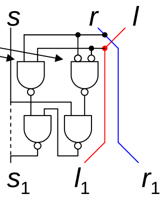
Notes: like Landauer’s “machine,” but *r* and *l* are trits & *s*, s_1 are state

- The magnetic core application was engineered to exploit this aggregation
  - Ask a question if you want details

Comparison to CMOS and a modern nanotechnology implementation

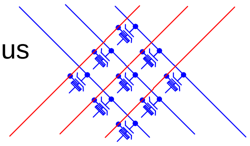
CMOS implementation:

Trit inputs



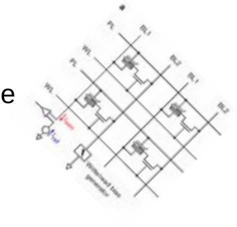
Notes: like Landauer's "machine," but r and l are trits & s, s_1 are state

Array analogous to cores above



Possible MeRAM implementation:

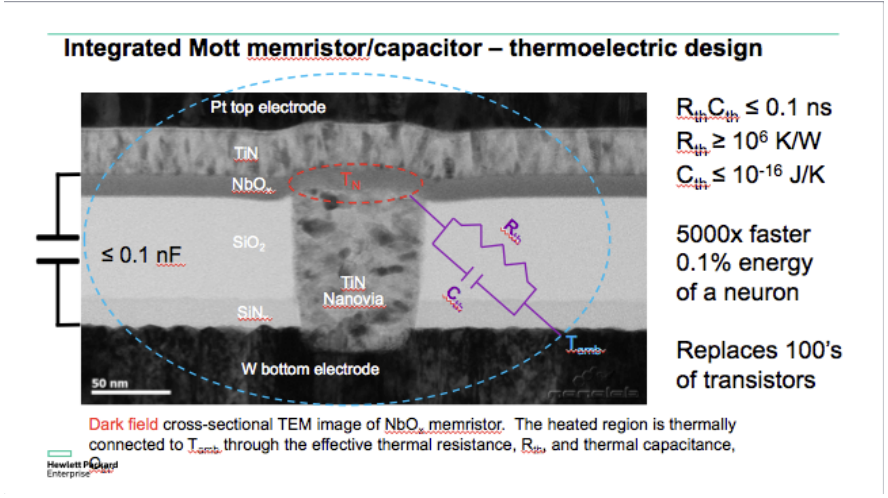
Magnetoelectric RAM is based on a device where voltage exceeding a threshold causes a nanomagnet to flip. Losses are negligible in absence of state change.



Jia-mian Hu, et al. "High-density magnetoresistive random access memory operating at ultralow voltage at room temperature." *Nature communications* 2 (2011): 553

Memristor-class device

- Late-breaking public info (you'll hear about this from Stan)



Integrated Mott memristor/capacitor – thermoelectric design

$R_{th}C_{th} \leq 0.1$ ns
$R_{th} \geq 10^6$ K/W
$C_{th} \leq 10^{-16}$ J/K

5000x faster
0.1% energy
of a neuron

Replaces 100's
of transistors

Dark field cross-sectional TEM image of NbO_x memristor. The heated region is thermally connected to T_{amb} through the effective thermal resistance, R_{th}, and thermal capacitance, C_{th}

Why is the “limit” so low? (III)
Logic-memory integration

- The preceding methods won’t help very much for the processor component of the von Neumann architecture
- A logic design is considered inefficient if the inputs to a large number of gates are nearly always 0 or 1. The design can be improved irrespective of anything in this slide deck.
- However, it is not poor design for a state-containing device (memory cell) to be idle most of the time – because it is serving the useful purpose of storing information
- While the preceding methods are independent of architecture, they give the biggest energy efficiency boost for processor-in-memory and neuromorphic

Scope of Talk – Outline

High level, non-technical:

Proposal of a nanotechnology Grand Challenge and its acceptance by the US Government. It is a challenge, not the solution, so we avoid favoring any technical approach

High level, technical:

New theory on the limit of computation showing that the Grand Challenge vision of many orders of magnitude increase in energy efficiency is possible, but the theory is general to many technical approaches

Possible technical validation:

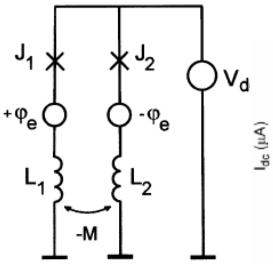
Can the theoretical limits be approached with a real system, even if impractical? As a starting point, the talk will review a superconducting nSQUID circuit that has been built and tested in a similar configuration

Can we find a device or circuit that might be able to reach the limit described?

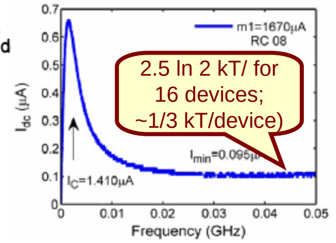
- Requirements
  - Row, column addressable (i. e. the array)
  - Addressed cell can be set to 1 or -1; all other cells unchanged
  - Zero dissipation if cell unaddressed or value already correct
  - Minimum energy ($T\Delta S$) if cell changes state
- Literature
  - P. Zulkowski and M. DeWeese, “Optimal finite-time erasure of a classical bit,” *Physical Review E* 89.5 (2014): 052140.
  - Uses a protocol for raising/lowering barriers and tilt
  - Dissipation $-T\Delta S + O(1/t_f)$, Landauer’s minimum as time limit $t_f \rightarrow \infty$
    - we can have a lot of discussion on this if you like
- Is there a circuit that does this?

Semenov's nSQUID circuit

A. Circuit

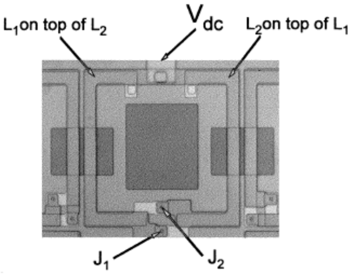


B. Measurements

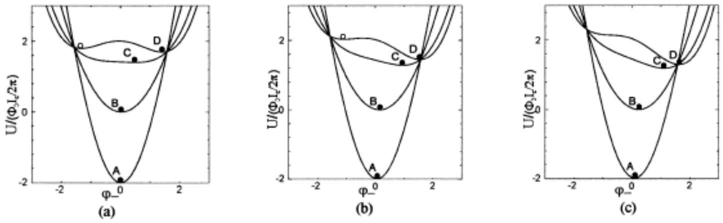


2.5 ln 2 kT/ for 16 devices; ~1/3 kT/device)

C. Micrograph



D. Behavior

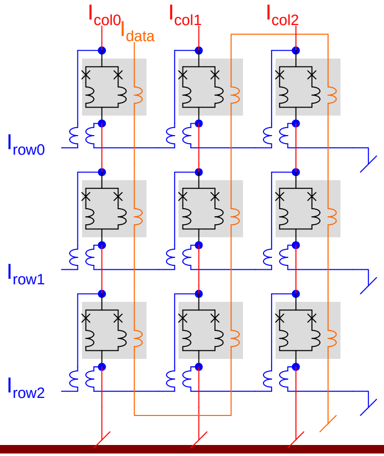


(a) (b) (c)

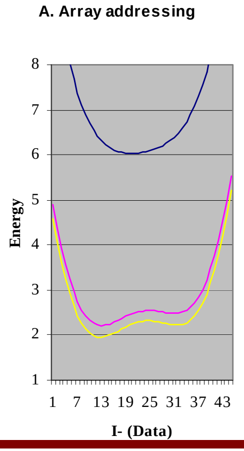
V. K. Semenov, G. V. Danilov, and D. V. Averin, "Negative-inductance SQUID as the basic element of reversible Josephson-junction circuits," *Applied Superconductivity, IEEE Transactions on* 13.2 (2003): 938-943.

Addition of addressing

- Author proposes addressing, which was not present in Semenov's work



- Excel spreadsheet of wells
  - Top: addressed
  - Lower: Un- and half-addressed



Conclusions

- Public believes "Moore's law is ending" due to imminent approach to (superficial generalizations of) "*kT*" limits; we show the limits are further out that commonly believed
- However, pushing out "limits" requires new approaches to computing as well as new devices. Approaches:
  - Optimize for probabilities in input data and intermediate variables
  - Find devices with higher level functions but the same dissipation
  - Use memory–intensive architectures (e. g. neural networks)
- This is a bridge between the brain and computing
- We don't have a complete working example, but Semenov may have constructed and tested a suitable circuit in a different context and measured 1/3 *kT*

Clarification: The limits we know of are leakage current, *kT*, O(*kT*), *kT* ln 2, p_{error} = exp(-e_{signal} / *kT*), 100 *kT*. We'll call these *kT* limits that differ by constant factors.

Further work

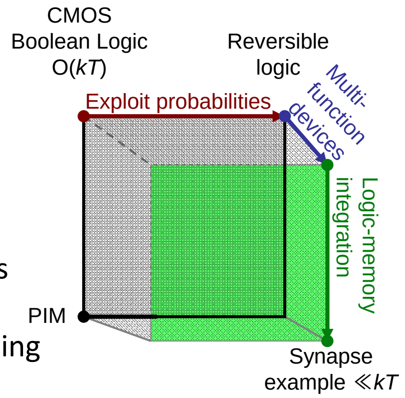
- Stan Williams calls the device research area “Nonlinear dynamics and Chaos”
  - Stan will talk at the NIST seminar series in September
  - Examples: some memristors, spintronic devices, artificial synapses
- There is a wide open area to develop integrated logic-memory architectures that can reach extreme sub-*kT* limits
- In fact, it might be possible to build an ultra-energy efficient brain. Of course, we’d have to figure out how the brain works.

NONLINEAR DYNAMICS AND CHAOS
Second Edition
NONLINEAR DYNAMICS AND CHAOS
Steven H. Strogatz
NONLINEAR DYNAMICS AND CHAOS
Klaus Mainzer • Leon Chua
LOCAL ACTIVITY PRINCIPLE

By the way, there is no device called a “nonlinear dynamics and chaos,” but it is instead a method of characterizing behavior

Roadmap and agenda (tentative)

- The cube forms a research agenda
- Each dimension can be explored separately
- Most of the vertices form recognized computer classes
- The lower-right corner represents a way to integrate neuromorphic computing into a general computing agenda
- Author have an ICRC paper with a guide to a roadmap based on E_r, the parasitic overhead energy of a gate


CMOS
Boolean Logic
O(kT)
Reversible
logic
Exploit probabilities
Multi-
function
devices
Logic-memory
integration
PIM
Synapse
example $\ll kT$