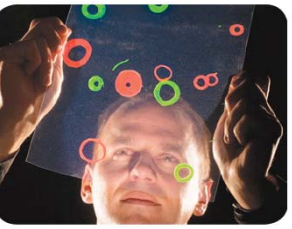
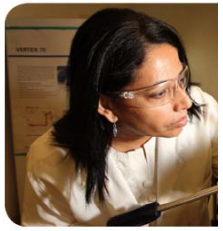
SAND2022-4954C





Using Neuromorphic Computing Methods for General Computer Performance Growth

IEEE Nano 2016

August 23, 2016

Erik P. DeBenedictis

Sandia National Laboratories

Exceptional service in the national interest

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 Unclassified Unlimited Release SAND 2016-6041 PE

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Executive summary

- Moore's Law
  - To Moore, scaling looked like it could proceed for a long time (1965)
  - Landauer's *kT* reasoning was unimaginably far away (1961)
- Today
  - Leakage current troubles due to subthreshold slope ln 10 *kT*/*q*
  - Landauer's reasoning now a limit at *kT* ln 2/binary op/clock (wrong)
  - Neuromorphic computing is cited as rejuvenating Moore's Law
- Key question
  - What about "neuromorphic" excepts it from thermodynamic limits?
- Answer
  - Nothing; theory has been misinterpreted
  - If we can understand the theory, maybe we can understand the scaling for neuromorphic systems.

kT is at the root of multiple problems

- Clock rate is not scaling anymore
  - Reason: excessive energy consumption
- Density continues to scale
  - Memory density scales just fine
  - Logic density could scale except for excess heat dissipation due to leakage current
- Leakage current is due to kT/q subthreshold slope limiting reduction in power supply voltage
- Beyond CMOS transistors are research topics
    - TFETs, piezotronic transistors, etc.
  - Benefit: Lower power supply voltage without leakage
  - Limit: thermal errors with probability $p_{error} = \exp(-e_{signal} / kT)$
- Hence *kT* limits speed, density, supply voltage, and reliability

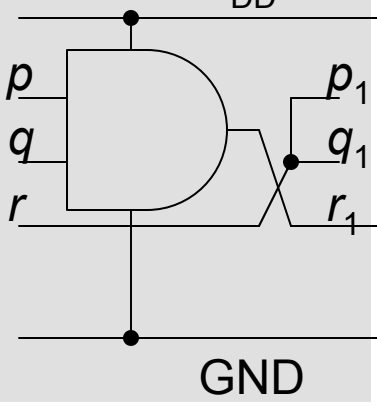
History of *kT* limits

- 1961: Landauer states "on the order of *kT*" per operation
  - Exactly what the operation is is subject to debate
  - Furthermore, critics believe it is not a tight "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 use of a logic gate
- 2016: We have to straighten it out

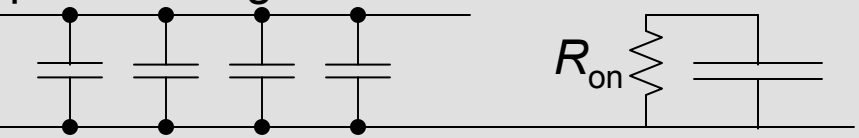
Erik says it is impossible to move information faster than twice the speed of light. Critics would not deny this limit but would say it is impossible to move information faster than one times the speed of light.

Models of computer energy dissipation

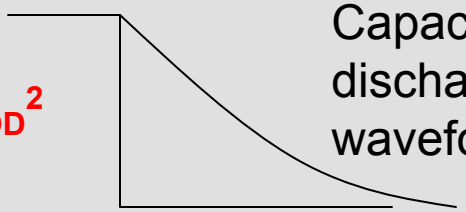
Machine:



A. CV^2 model:



Discharge circuit and waveform:



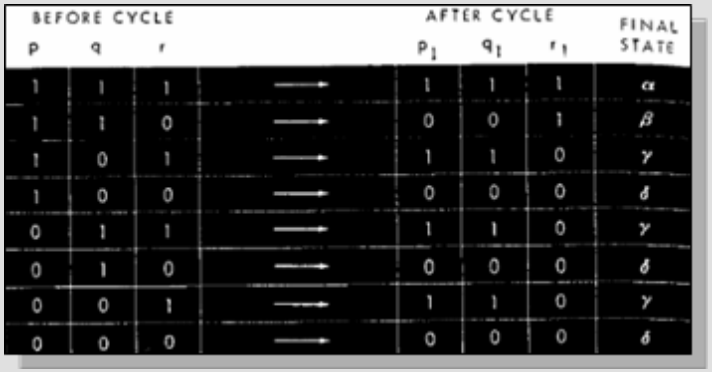
Capacitor discharge waveform

Idealized waveform

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.



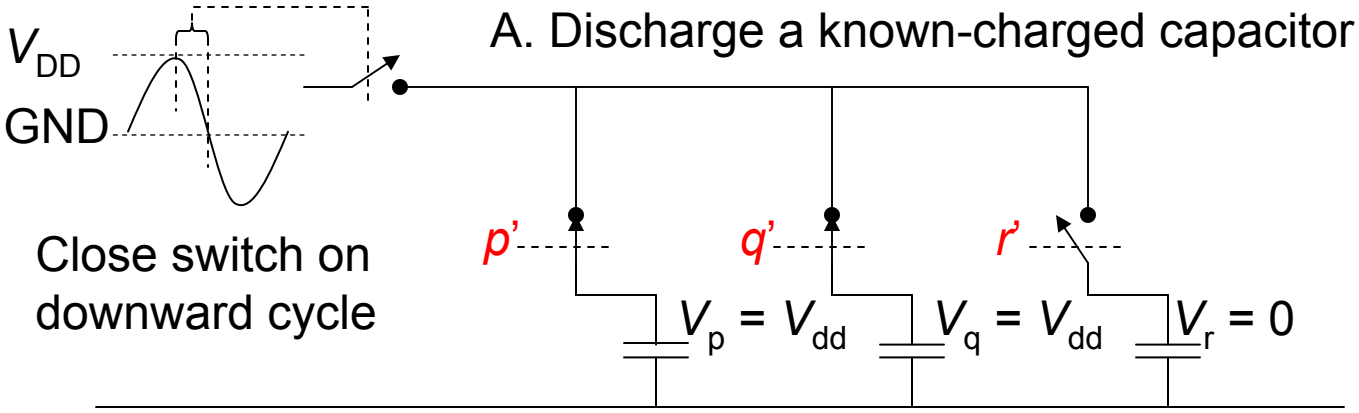
BEFORE CYCLE p	q	r	AFTER CYCLE p_1	q_1	r_1	FINAL STATE
1	1	1	1	1	1	α
1	1	0	0	0	1	β
1	0	1	1	1	0	γ
1	0	0	0	0	0	δ
0	1	1	1	1	0	γ
0	1	0	0	0	0	δ
0	0	1	1	1	0	γ
0	0	0	0	0	0	δ

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

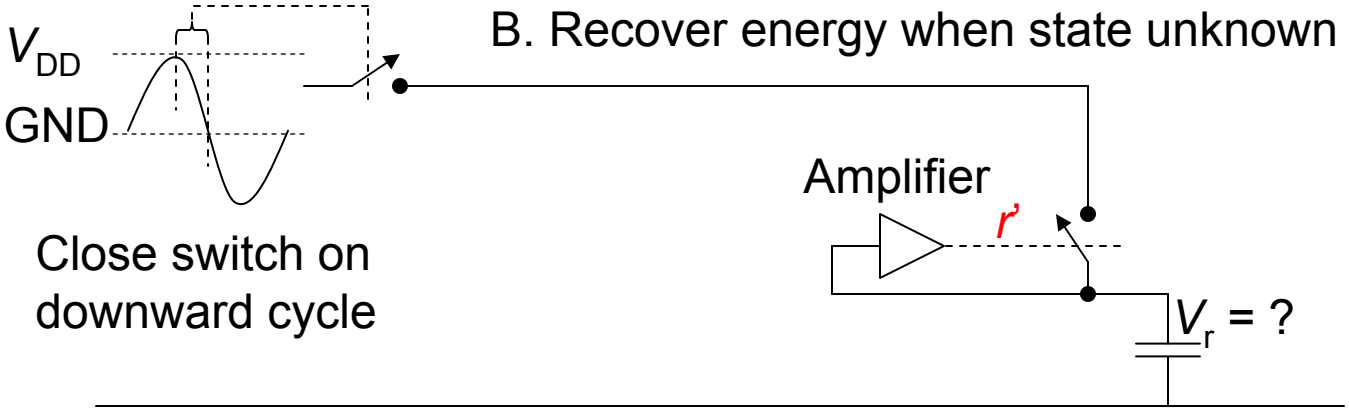
[Landauer 61] Landauer, Rolf. "Irreversibility and heat generation in the computing process." *IBM journal of research and development* 5.3 (1961): 183-191.

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

Background on erasure model



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



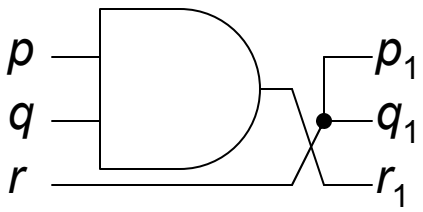
Works, but only until energy on capacitor is on the order of *kT*. Below this level, the amplifier can't decide whether to charge of discharge

Agenda for next several slides (speaker flip back and forth if you can)

- Look at Landauer's paper, paying attention to the message in
  - Title + abstract + diagrams
  - Body of article
- Go through example in the 1961 paper
  - An AND gate yields limit of 0.82 kT, which is O(kT)
  - This is the "Landauer Limit" per popular usage
- Go through a modern example of neuromorphic synapse
  - Synapse yields limit of 0.0058 kT, which is not O(kT) so to speak
  - This is consistent with body of paper but not popular usage
- Why?

Landauer's method from the paper's example

System:



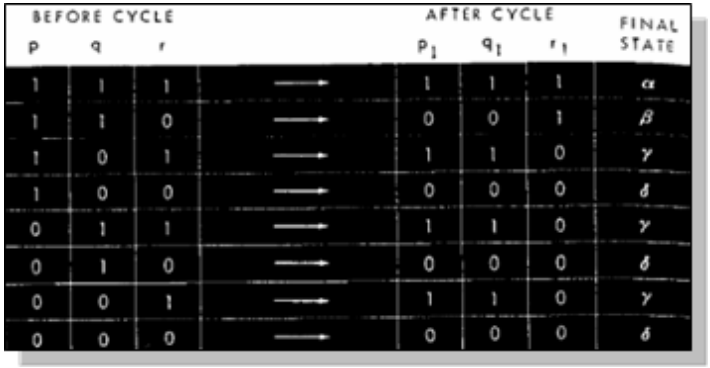
prob	p	q	r		p1	q1	r1	Si (k's)	State	Sf (k's)
0.125	1	1	1	→	1	1	1	0.25993	α	0.25993
0.125	1	1	0	→	0	0	1	0.25993	β	0.25993
0.125	1	0	1	→	1	1	0	0.25993	γ	0.367811
0.125	1	0	0	→	0	0	0	0.25993	δ	0.367811
0.125	0	1	1	→	1	1	0	0.25993	γ	0
0.125	0	1	0	→	0	0	0	0.25993	δ	0
0.125	0	0	1	→	1	1	0	0.25993	γ	0
0.125	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 p	q	r		AFTER CYCLE p_1	q_1	r_1	FINAL STATE
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.

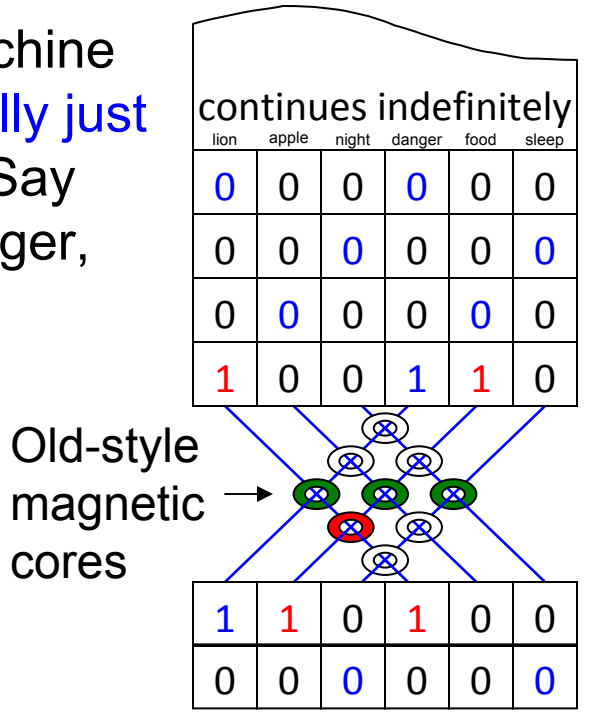
Backup: Details

- 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 $S_i - S_f$
- Notes
  - Inputs states that don't merge do not raise minimum energy
  - Inputs that merge raise minimum energy based on their probability
  - Assumption: All input combinations equally probable

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 }

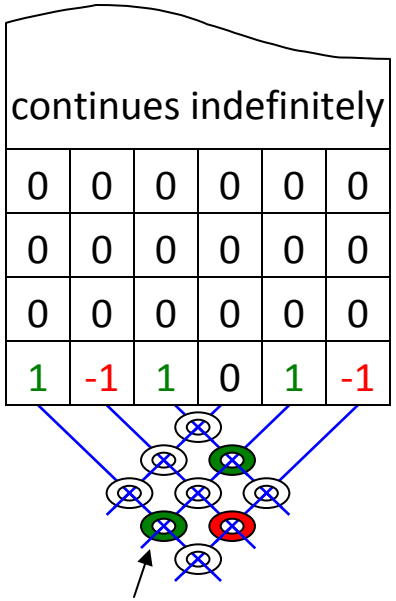


Functional example:
Machine continuously monitors environment for {1, 1} or {-1, -1} pairs and remembers them in state of a magnetic core. Theoretically, there is no need for energy consumption unless state changes.

Signals create currents;
core flips a ±1.5

Analysis of one synapse in the learning machine

Boolean logic equivalent system:



Old-style magnetic core

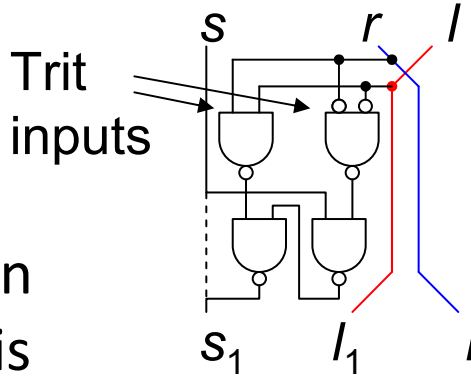
	left wire	right wire	field dir.		left wire	right wire	field dir.	Si (k's)	State	Sf (k's)
0.062438	-1	-1	-1	→	-1	-1	-1	0.173176	A	0
0.062438	-1	0	-1	→	-1	0	-1	0.173176	B1	0.173176
0.062438	-1	1	-1	→	-1	1	-1	0.173176	C1	0.173176
0.062438	0	-1	-1	→	0	-1	-1	0.173176	D1	0.173176
0.062438	0	0	-1	→	0	0	-1	0.173176	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**

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 (but "limitation")
  - "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 verify that they have learned what the need to know 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 →
- The magnetic core application was engineered to exploit this aggregation
  - Ask a question if you want details

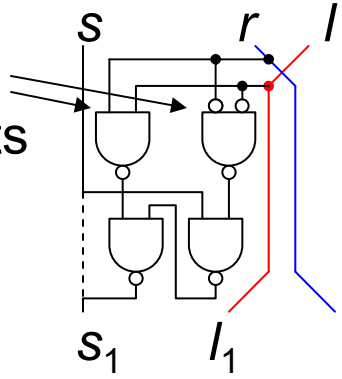


Notes: like Landauer's "machine," but *r* and *l* are trits & *s*, s_1 are

Comparison to CMOS and a modern nanotechnology implementation

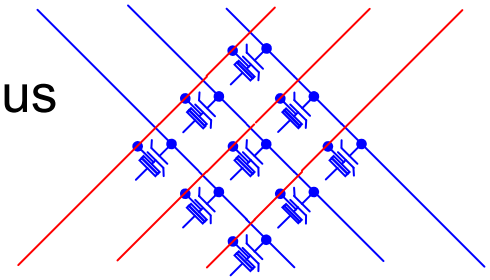
Sandia National Laboratories

CMOS implementation:



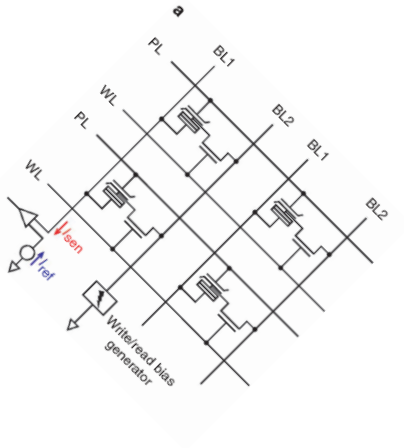
Notes: like Landauer's "machine," but *r* and *l* are trits & *s*, s_1 are

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.

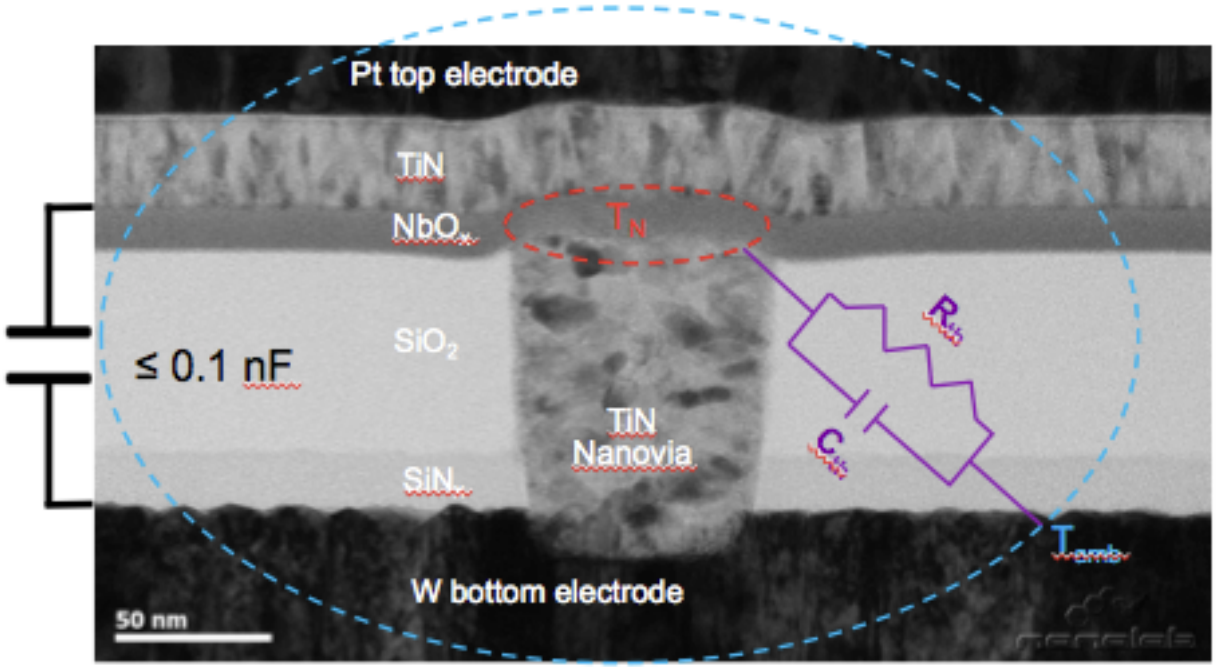


Jia-mian Hu, et al. "High-density magnetoresistive random access memory operating at ultralow voltage at room temperature." *Nature communications* 2 (2011): 553

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}.

Hewlett Packard Enterprise

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

Sandia National Laboratories

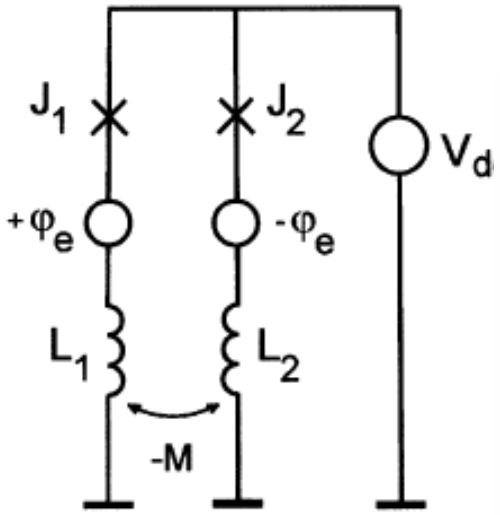
- The preceding methods won't help very much for the processor component of the von Neumann architecture
- A logic design is considered inefficient if the inputs to a large number of gates are nearly always 0 or 1. The design can be improved irrespective of anything in this slide deck.
- However, it 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

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

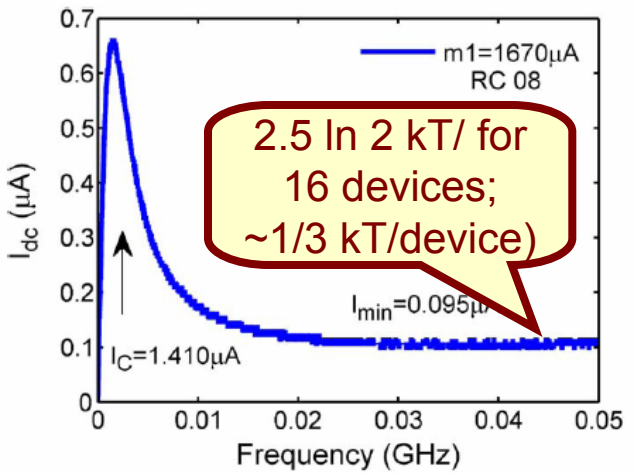
- Requirements
  - Row, column addressable (i. e. the array)
  - Addressed cell can be set to 1 or -1; all other cells unchanged
  - Zero dissipation if cell unaddressed or value already correct
  - Minimum energy ($T\Delta S$) if cell changes state
- Literature
  - P. Zulkowski and M. DeWeese, "Optimal 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

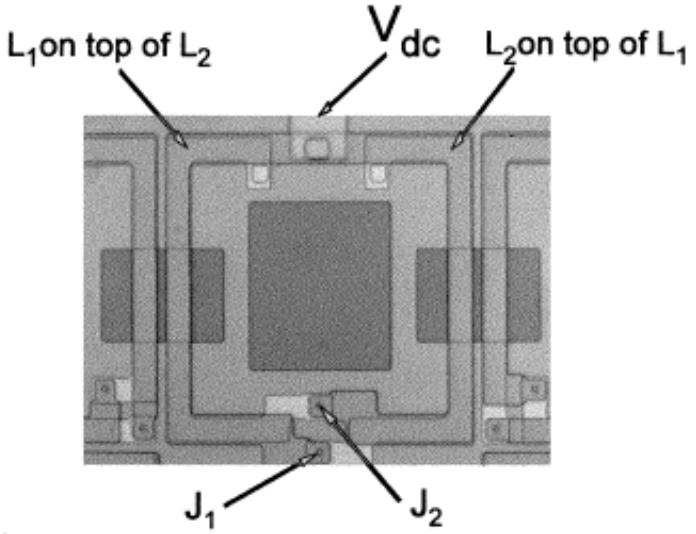
A. Circuit



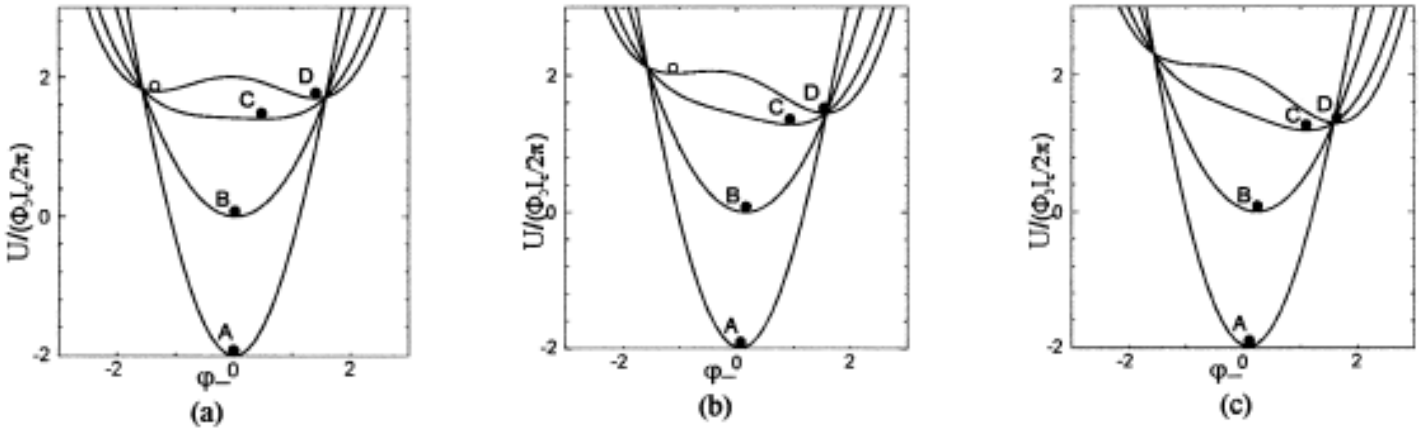
B. Measurements



C. Micrograph



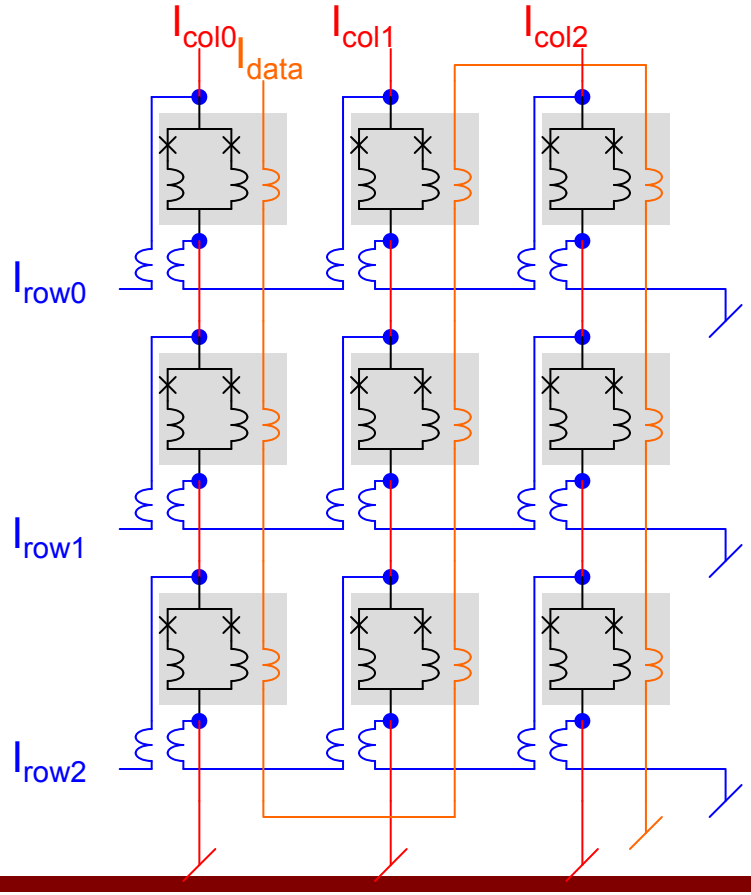
D. Behavior



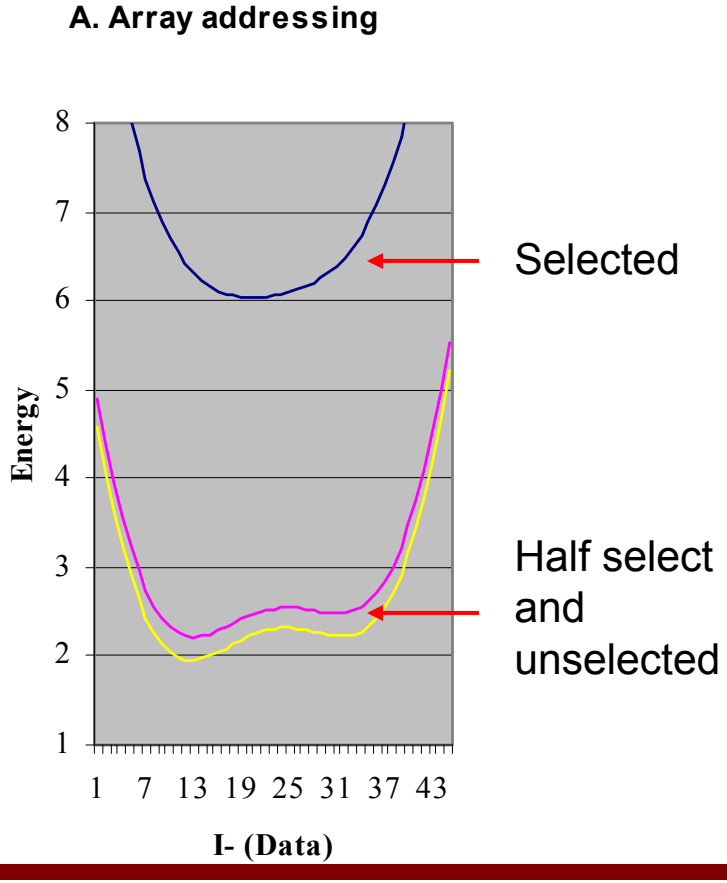
V. K. Semenov, G. V. Danilov, and D. V. Averin, "Negative-inductance SQUID as the basic element of reversible Josephson-junction circuits," *Applied Superconductivity, IEEE Transactions on* 13.2 (2003): 938-943.

Addition of addressing

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

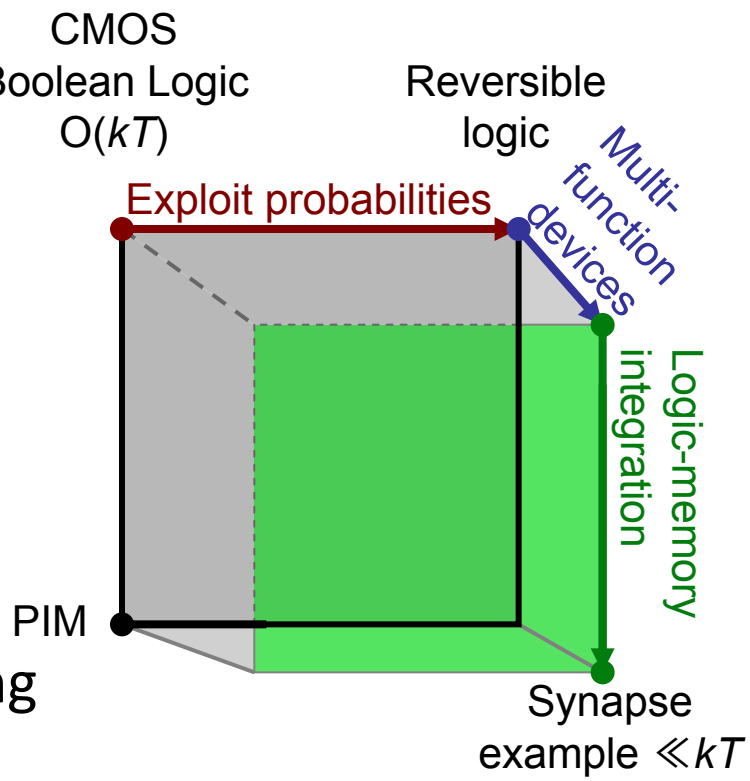


- Excel spreadsheet of wells
  - Top: addressed
  - Lower: Un- and half-addressed



Roadmap and agenda/Conclusions

- 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

Conclusions

- Public believes "Moore's law is ending" due to imminent approach to (improperly interpreted) "*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.