Sensible Machine Grand Challenge

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

Notre Dame Solid State Physics Seminar Series, April 1, 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

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

Survey of technical approaches:

However, there is existing work in approaches that could further developed into a solution of the Grand Challenge. Some of these are mentioned without showing favoritism

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

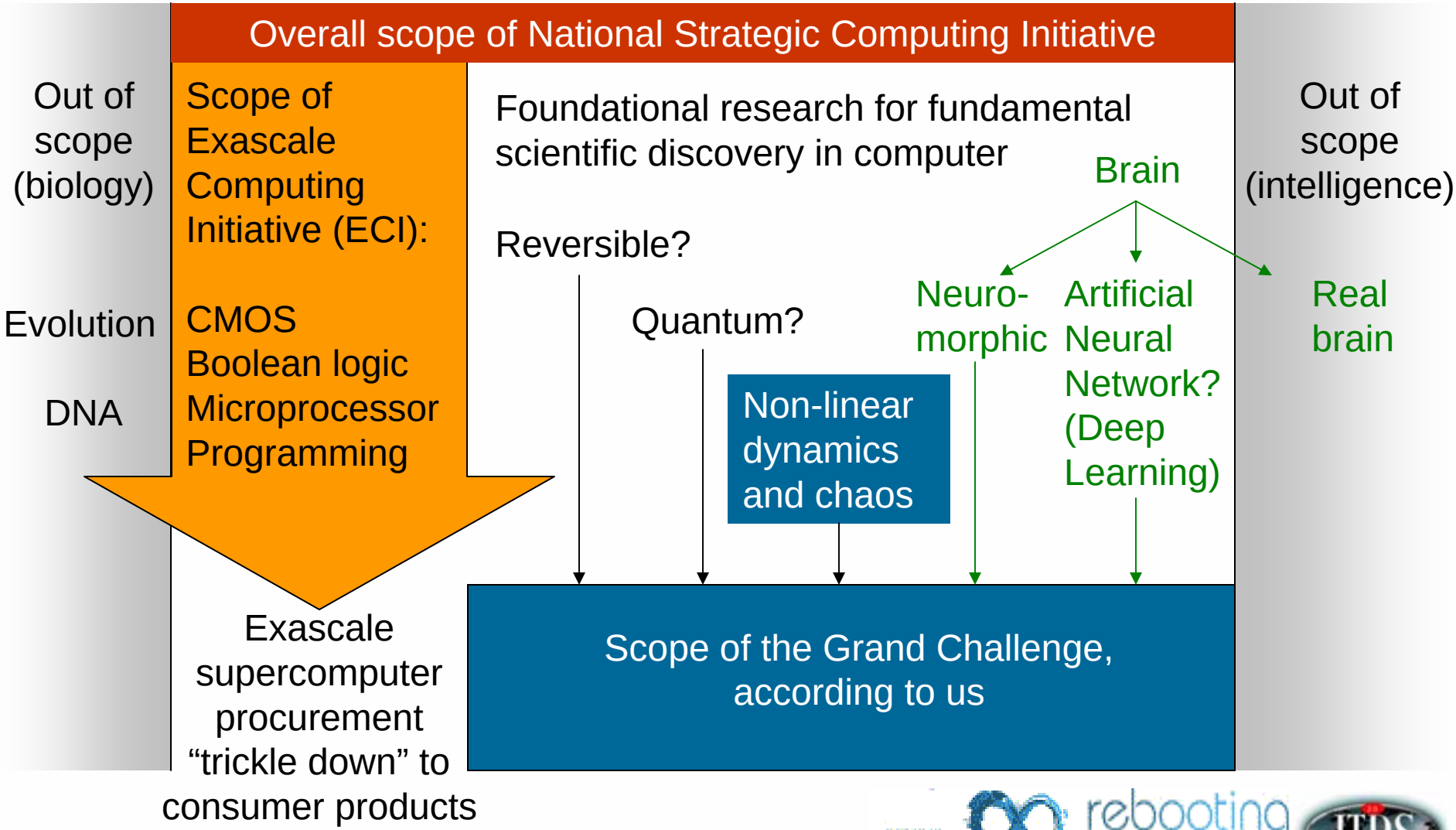
Stan Williams' original response to the OSTP RFI

- "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."

Scope of the Grand Challenge, according to us



IEEE rebooting COMPUTING ITRS

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:

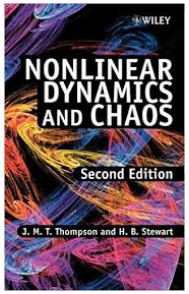
New theory on the limit of computation showing that the Grand Challenge vision of many orders of magnitude increase in energy efficiency is possible, but the theory is general to many technical approaches

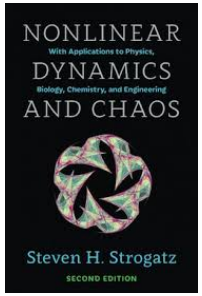
Survey of technical approaches:

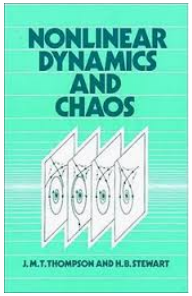
However, there is existing work in approaches that could further developed into a solution of the Grand Challenge. Some of these are mentioned without showing favoritism

Outlining the direction of a hardware solution

- Physical limits of current technology are due to two effects
  - CV^2 energy on wires
  - Landauer's "Limit" of O(kT) per gate-op (big debate here)
- Question: Are these fundamental limits or just attributes of the current approach?
- Stan's suggestion in RFI response
  - View devices at the level of nonlinear dynamics and chaos,
  - …instead of Boolean logic





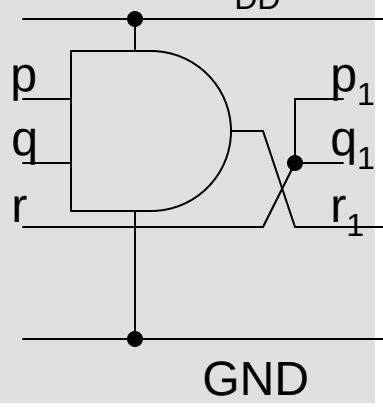


By the way, there is no device called a "nonlinear dynamics and chaos," but it is instead a method of characterizing behavior

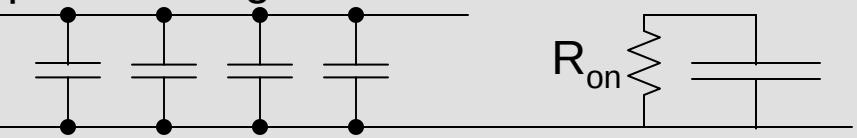
IEEE rebooting COMPUTING ITRS

Models of computer energy dissipation

System:



A. CV^2 model:



Discharge circuit and 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

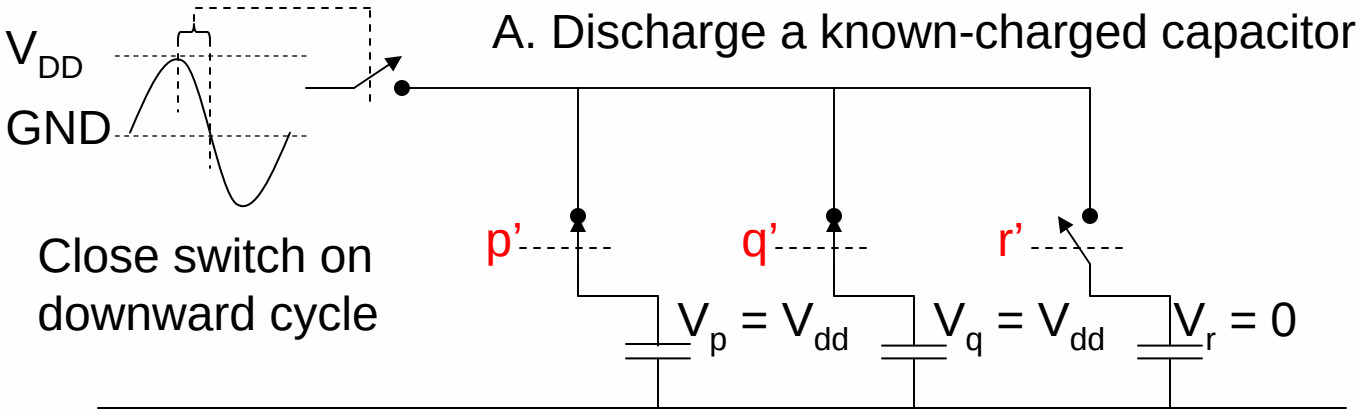
…typically of the order of kT for each irreversible function

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

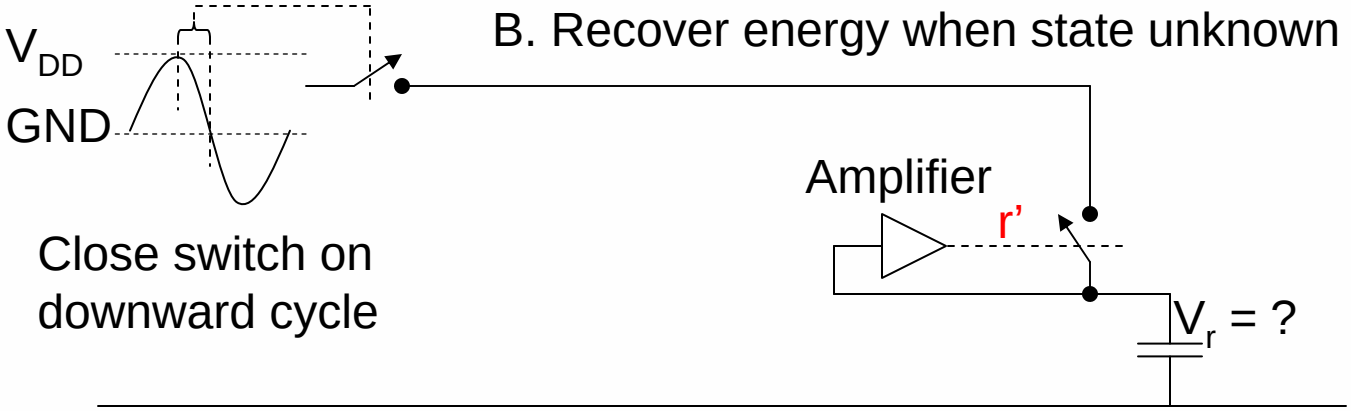
See also http://rebootingcomputing.ieee.org/images/files/pdf/RCS4DeBenedictisposter.pdf

IEEE rebooting COMPUTING ITRS

Background on erasure model



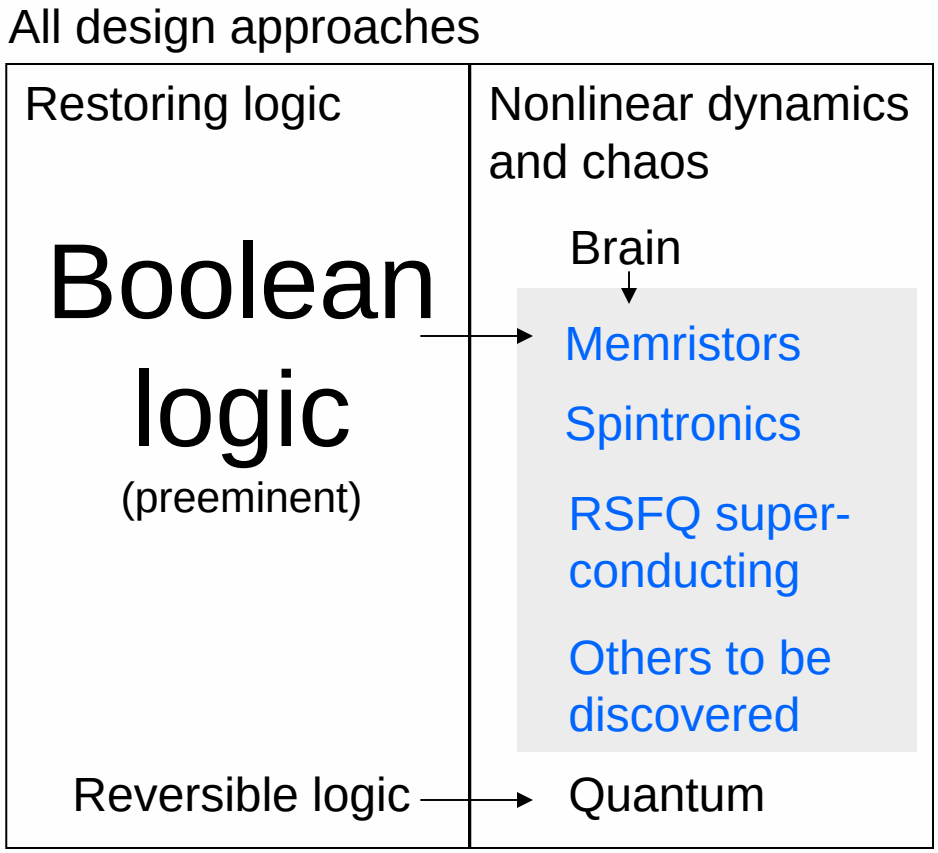
Works, but we need copies p' = p, q' = q, and r' = r to set the switches, which prevents erasure of last copy of a signal



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

Outlining the direction of a hardware solution

Use ideas from the brain and current approaches (Boolean logic) to be more energy efficient (without changing erasure debate)

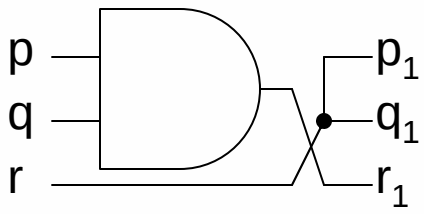


Technical agenda

- Let's consider the brain's function when learning
  - With intuition of what is to come, engineer a machine for it
  - Run our machine through Landauer's minimum energy assessment
- We should not expect Landauer's typical results
  - In the 1960s, typical circuits were Boolean gates (AND, in the example)
  - Landauer's paper was correct about O(kT) minimum energy for typical circuits of that time, but this would have been an "atypical" circuit
  - (We'll get $\ll$kT minimum energy)
- Then, generalize the intuition
  - Can we reverse-engineer Landauer's energy assessment to find a systematic way to generate systems where minimum energy $\ll$kT
- Hmm, the brain inspired us to create better computers

Landauer's method from the paper's example

System:



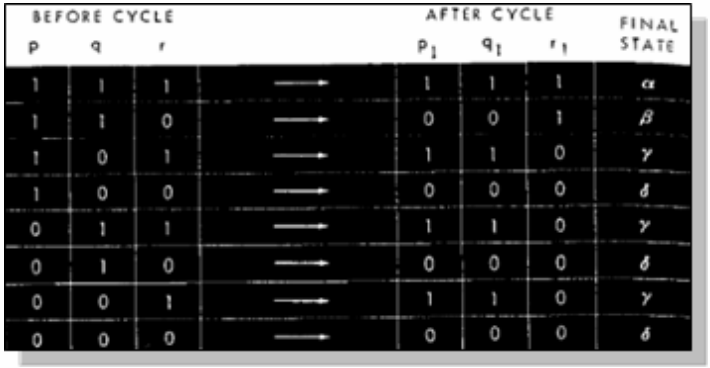
prob	p	q	r		p1	q1	r1	Si (k's)	State	Sf (k's)
0.125	1	1	1	→	1	1	1	0.25993	α	0.25993
0.125	1	1	0	→	0	0	1	0.25993	β	0.25993
0.125	1	0	1	→	1	1	0	0.25993	γ	0.367811
0.125	1	0	0	→	0	0	0	0.25993	δ	0.367811
0.125	0	1	1	→	1	1	0	0.25993	γ	0
0.125	0	1	0	→	0	0	0	0.25993	δ	0
0.125	0	0	1	→	1	1	0	0.25993	γ	0
0.125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

IEEE rebooting COMPUTING ITRS

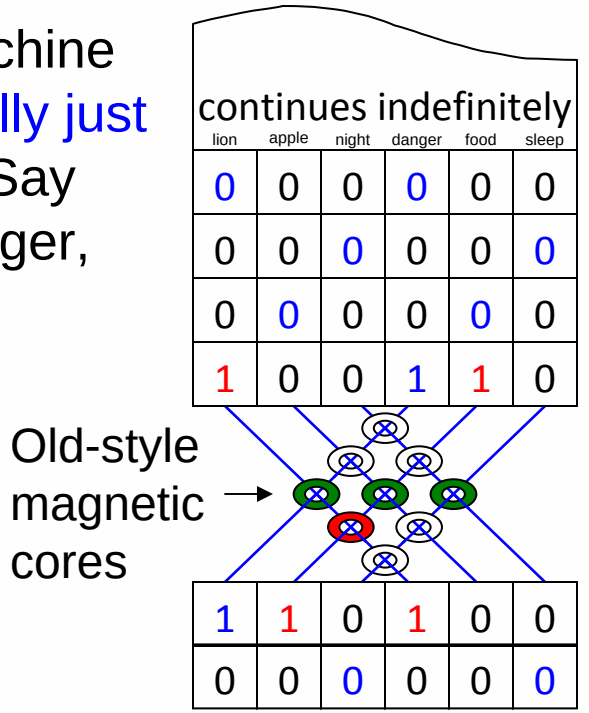
Backup: Details

- 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 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 }



Signals create currents;
core flips a ±1.5

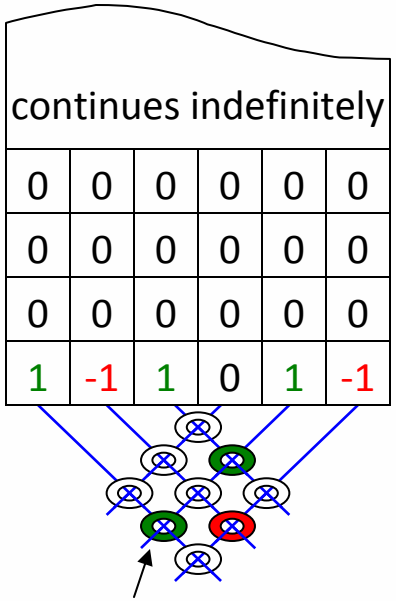
Functional example:
Machine continuously monitors environment for {1, 1} or {-1, -1} pairs and remembers them in state of a magnetic core. Theoretically, there is no need for energy consumption unless state changes.

IEEE rebooting COMPUTING IRS

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



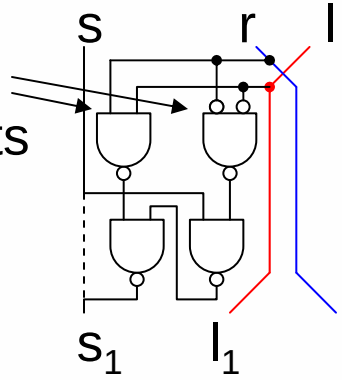
Old-style magnetic core

IEEE rebooting COMPUTING ITRS

Comparison to CMOS and a modern nanotechnology implementation

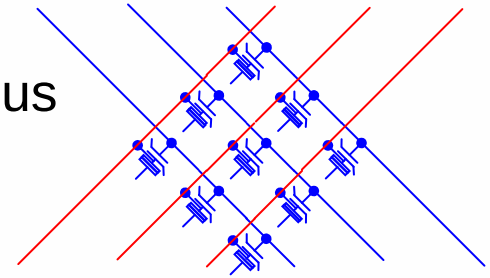
CMOS implementation:

Trit inputs



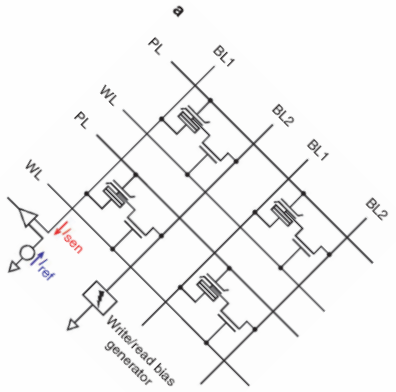
Notes: like Landauer's "machine," but r and l are trits & s, s_1 are state

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

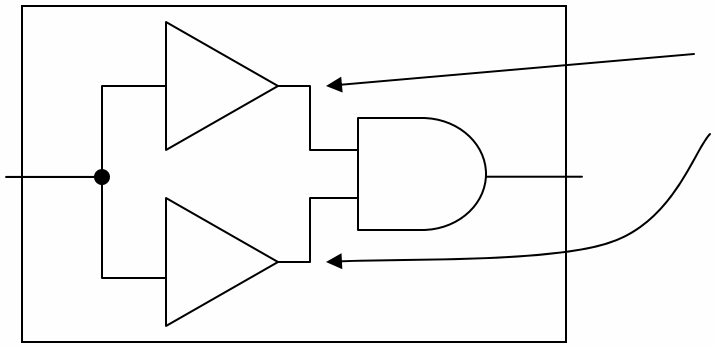
IEEE rebooting COMPUTING ITRS

Generalizing the intuition I

- Create and optimize devices for the needed function, instead of creating switches for Boolean logic
  - Different information representations OK (voltage, magnetic field)
  - Devices need not restore logic each time (just sometimes)
- Try to simplify in comparison to Boolean logic
  - Any function can be realized by Boolean logic because it is universal, most of such circuits constructed from minimal energy Boolean logic gates will not have minimal energy
- Exploit probabilities – design so the most probable inputs are the ones with small (or zero) minimal energy
- Use logic-in-memory to avoid unnecessary data movement
- More on subsequent slides (aggregation and free stuff)

Generalizing the intuition II (aggregation)

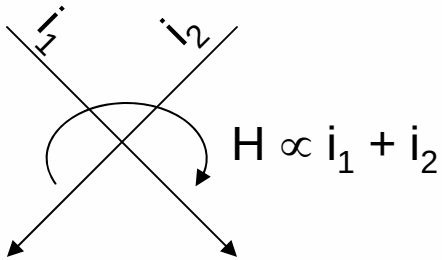
- Circuit to below is a non-inverting buffer when viewed as a whole
  - Minimum energy is zero
- If viewed as three separate gates
  - AND gate has minimum energy of .823 kT
  - CV^2 energy will involve three signals
- In general, energy $E_{min}(f(g(x))) \leq E_{min}(f(x)) + E_{min}(g(x))$
- So make integrated structures that do a lot, if you can



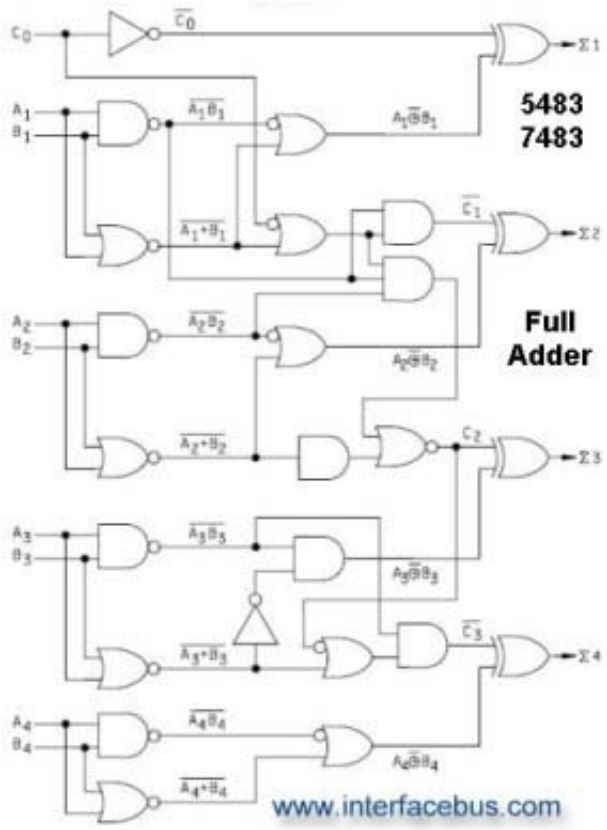
One of these copies will need to be erased, generating heat

Generalizing the intuition II (free computations)

- If you want to add or multiply with gates, you get a collection of gates, like the 7483 4-bit adder →
  - Each gate has a minimum energy of kT ln 2
- However, you can add two currents and express the result as a magnetic field by simply having two wires near each other ↘
  - We believe even the most ardent skeptics will admit these wires dissipate ≪ kT energy



- We did addition, comparison with memory, and conditional branch



IEEE rebooting COMPUTING ITRS

Conclusions to section

- We analyzed the brain's function when learning
  - With foresight of what was to come, we engineered a machine for it
  - We ran our machine through Landauer's minimum energy assessment
- This is unlikely to have been forefront on Landauer's mind
  - In the 1960s, typical circuits were Boolean gates (AND, in the example)
  - Landauer's paper was correct about O(kT) minimum energy for typical circuits of that time, but this would have been an "atypical" circuit
  - Actually, we got $\ll kT$
- We then generalized our foresight
  - We reversed-engineered Landauer's energy assessment to find a way to systematically generate systems with $\ll kT$ minimum energy
- Gee, wasn't the brain a nice inspiration?

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

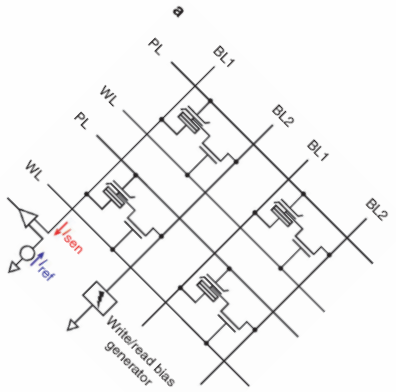
Survey of technical approaches:

However, there is existing work in approaches that could further developed into a solution of the Grand Challenge. Some of these are mentioned without showing favoritism

Spintronics (repeat slide)

Many spintronics devices seem to include state and offer the possibility of unusually low minimum energy.

Discussed previously in this slide deck.



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

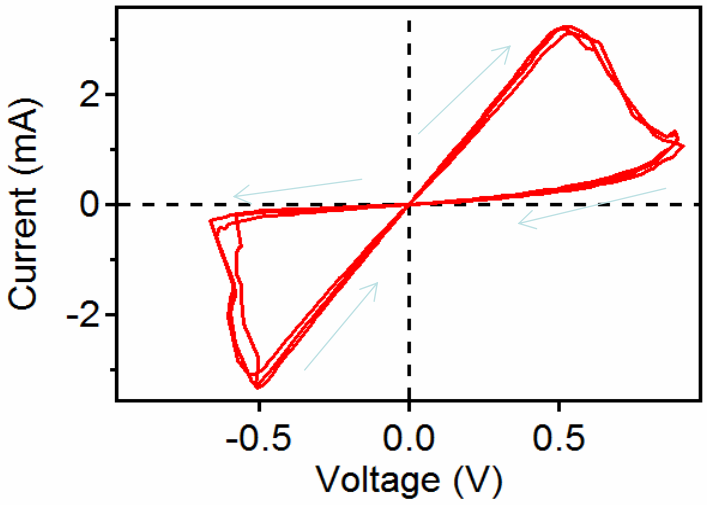
Superconducting electronics (MAGIC cells)

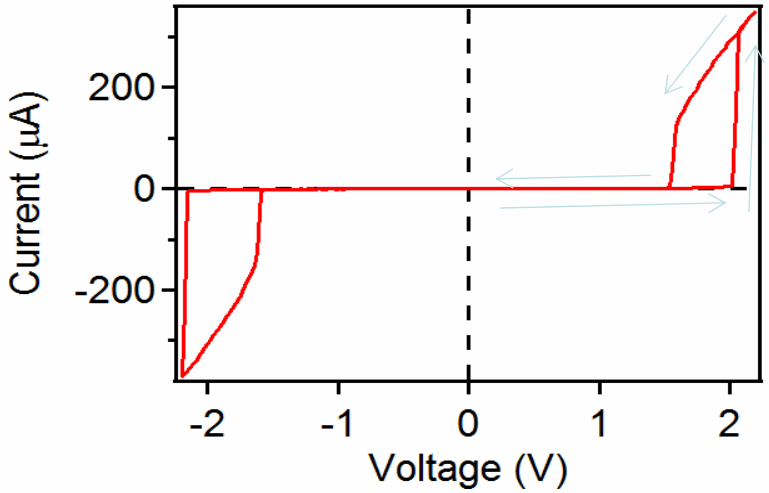
- Ref. Memory And loGIC (MAGIC) cells
- Superconducting Josephson junction circuits tend to hold state for no expenditure of energy
- Diagram to right
  - "Same-scale images of one-bit slices of (a) the carry-look-ahead accumulator and (b) Kogge-Stone adder. The images are extracted from microphotographs of first-order decimation filter shown in […]. Informally, we could say that the accumulator "exploded" into the much larger adder after inherent memory of the RSFQ accumulator has been eliminated."

Semenov, Vasili K. "Magic cells and circuits: New convergence of memory and logic functions in superconductor devices." Applied Superconductivity, IEEE Transactions on 23.3 (2013): 1700908-1700908.

Neural processes can be emulated with memristors

- Nonvolatile Memristor
  - Emerging digital memory/storage
  - Synapse in neuromorphic circuit
- Locally Active (e. g. “Mott”) memristor
  - Emerging neuronal compute device
  - Passive selector in crossbar memories





Leon Chua, *IEEE Trans. Circuit Theory* 18, 507 (1971).

Conclusions

- One of us proposed a nanotechnology Grand Challenge
  - The US Government adjusted the definition to address political sensitivities
  - In the process of adoption, it turned into an arrangement between 24 people and two professional organizations
- The Grand Challenge vision has a technical basis
  - Energy efficiency limits of O(kT) per function were identified the 1960s
    - Now called the “Landauer Limit” (even though “limit” is not in the paper)
  - By reverse engineering the devised in the 1960s, we can find a general strategies to lower energy, beating so-called limit by orders of magnitude
  - This talk uses a learning example inspired by the brain
- Intent is to encourage/fund research from materials to new software

Backup

- The following were in the original slide deck, but were not presented at Notre Dame by choice of presenter
- However, they may be useful for reference after the fact

Technical paths to a solution

- Hardware: Thermodynamic limits are within the planning horizon
  - Need to reframe problem to encompass a larger solution space, and then find a solution in the larger space
- Software: Speed of light limits speed of von Neumann computers
  - OSTP statements hint that "learning" might be a post-software model

OSTP: 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.

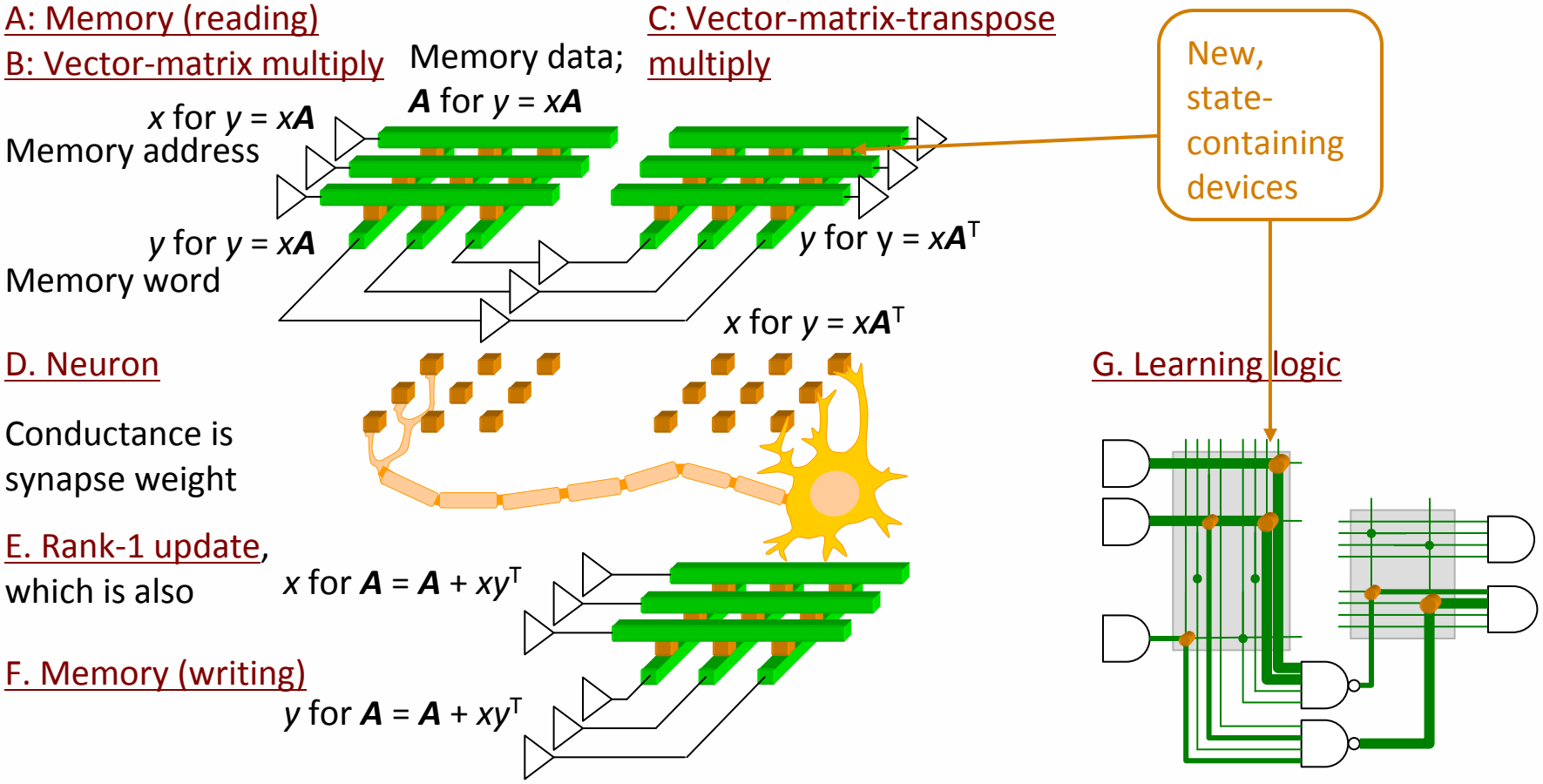
IEEE rebooting COMPUTING ITRS

Backup: Theoretical Analysis of Improved Energy Efficiency

Diagram is the same calculation as in Landauer's paper. In lieu of Boolean logic with O(kT) energy/gate, diagram is for a learning machine directly, with 1% probability of seeing input data to be learned and 0.01% probability of seeing contradictory data.

Probability of data to be learned:								**0.01**		
Probability of conflicting data:								**0.0001**		
Probability	left wire	right wire	field dir.		left wire	right wire	field dir.	Si (k's)	State	Sf (k's)
0.000001	-1	-1	-1	→	-1	-1	-1	0.000014	A	0.000921
0.001400	-1	0	-1	→	-1	0	-1	0.009201	B1	0.009201
Seven copies of row above for sequential input combinations (states C1-H1)										
0.000099	1	1	-1	→	1	1	1	0.000913	I	
0.000099	-1	-1	1	→	-1	-1	-1	0.000913	A	
0.140014	-1	0	1	→	-1	0	1	0.275269	B2	0.275269
Seven copies of row above for sequential input combinations (states C2-H2)										
0.009901	1	1	1	→	1	1	1	0.045694	I	0.046052
						Si (k's)		2.038824	Sf (k's)	2.038263
								Si-Sf (k's)		**0.000561**

Generalization of memristor applications



Can A-F form a general-purpose computer? Can process be repeated?
Narrative online (see notes)