Maps for Energy Efficiency in Computing

Part 2: *Limits and Energy Estimates*

Sadasivan (Sadas) Shankar
SLAC National Laboratory,
Materials Science and Engineering, Stanford University

Energy Efficiency Scaling (EES2): Roadmap Meeting #9
DOE/EERE Advanced Manufacturing, Materials, and Technologies Office (AMMTO)

Online Meeting
August 16-17, 2023

Questions to be Addressed (*Recap*)

- Top-down estimates have shown > 24 orders of magnitude in energy as a bit is translated to an *instruction* for *simulation of an Application*

- What are can be learned from biological and quantum systems?
- How can we use this to help design energy efficient computing?

- Two parts:
  - Part 1: Estimate Energy for information processing on the other different dimensions
  - Part 2: Quantify and Identify Pathways for using the lessons from Nature and Quantum Information Processing
  - Part 3: New formalisms for Energy Efficient Computing

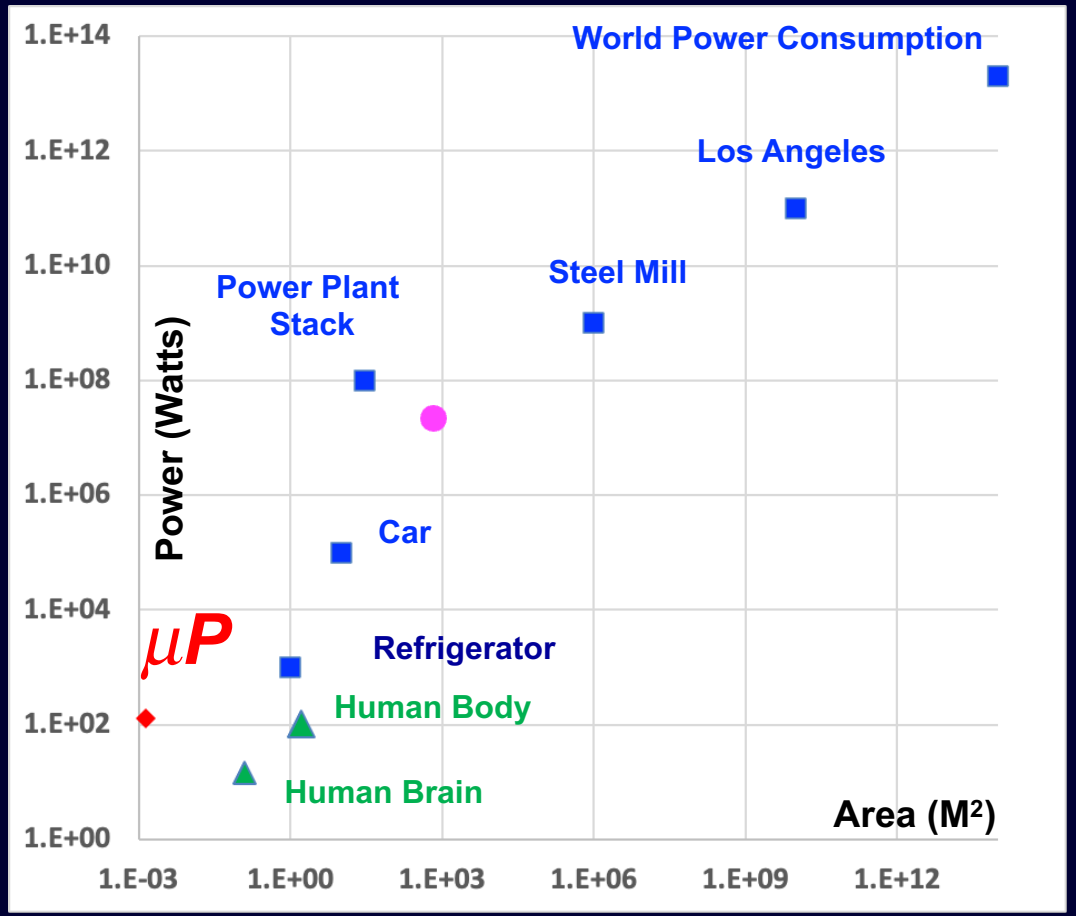
Outline

- Energy comparisons with other systems
- Energy limits in computing
- ~~Three~~ Four Different Domains in Computing
- Summary
- Energy Estimates for different Neural Networks & Neuromorphic Architectures (students presentation)

Energy comparisons

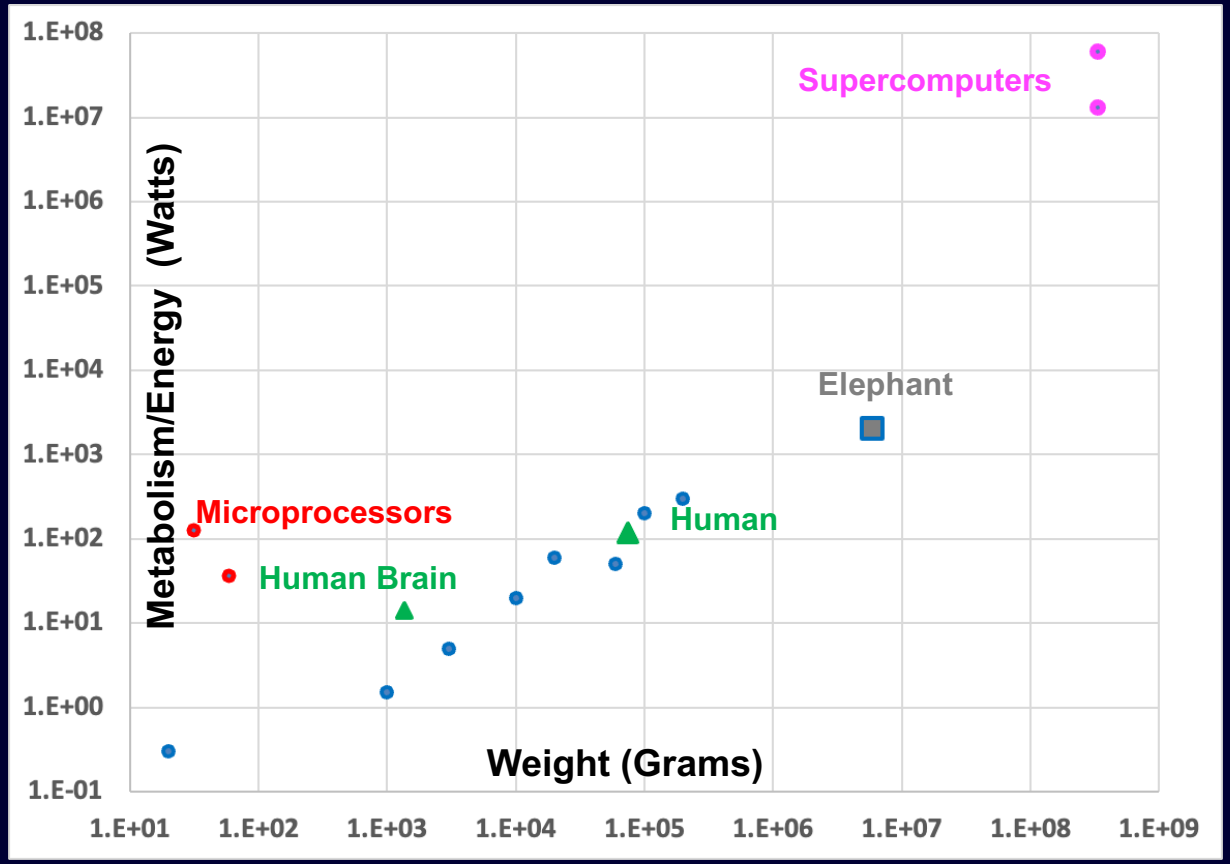
Natural and Synthetic Systems

Human-made Synthetic Systems



- Energy Consumption Scaling with Area
- Supercomputer is consistent with other human-made designs
- Higher Power is associated with larger areas, except Processor

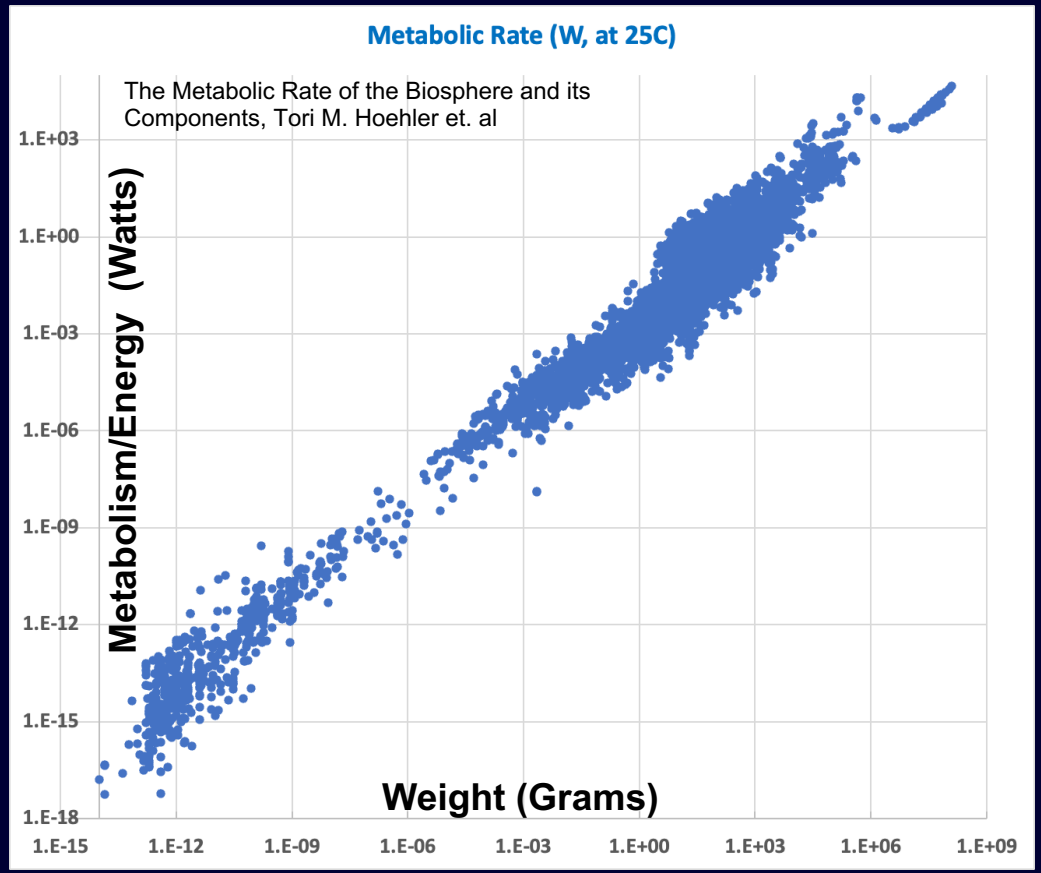
Natural Systems (Specific)



- Computing systems do not lie on the natural system allometric curves
- Microprocessor energy intensity (power/mass) is higher than of larger computers
- Evolution will **not** let microprocessor survive naturally

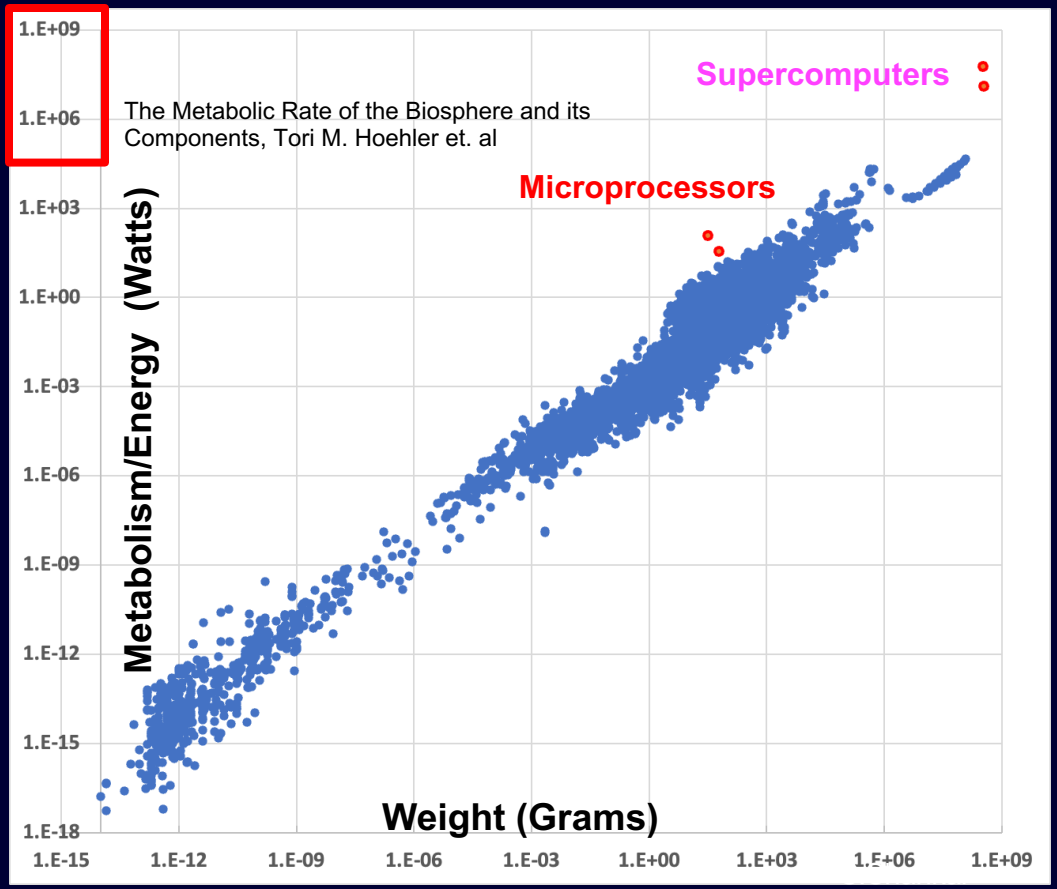
S.Shankar

Natural Systems (Ecosystem)



- Biosphere of comprising >10,000 metabolic rate measurements made on >2,900 individual species
  - From insects, trees to both mammals in land and in water
- Availability of energy constrains the potential abundance, distribution, and productivity of life
- The organism-level data, which are dominated by animal species, have a geometric mean among basal metabolic rates/gm of 5.16E-04

Natural Systems (Ecosystem)



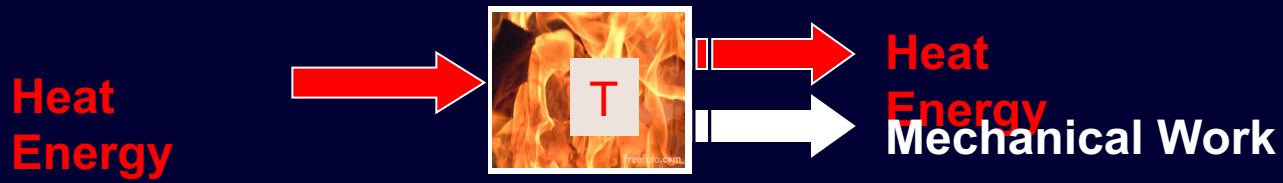
- Computing systems do not lie on the natural system allometric curves
- Power (or Energy) intensity of computing is evident
  - Off the charts
- Microprocessor energy intensity (power/mass) is higher than of larger computers
- Evolution will **not** let microprocessor or computer survive naturally
  - AGI will be power hungry

Energy Limits of Computing

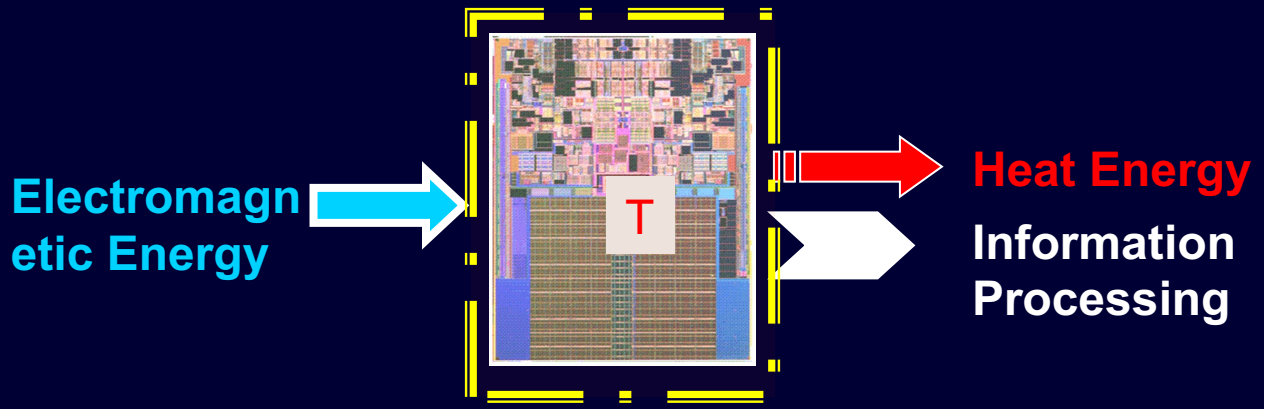
Reversible Computing

Computer as a Thermodynamic System

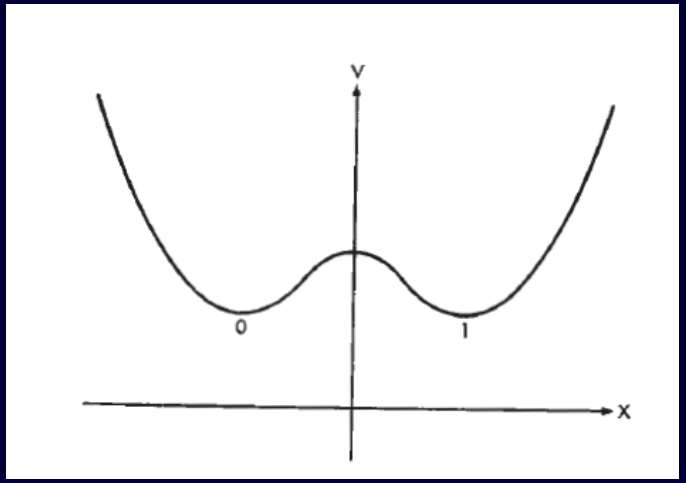
(a). Heat Engine



(b). Information Processing Engine



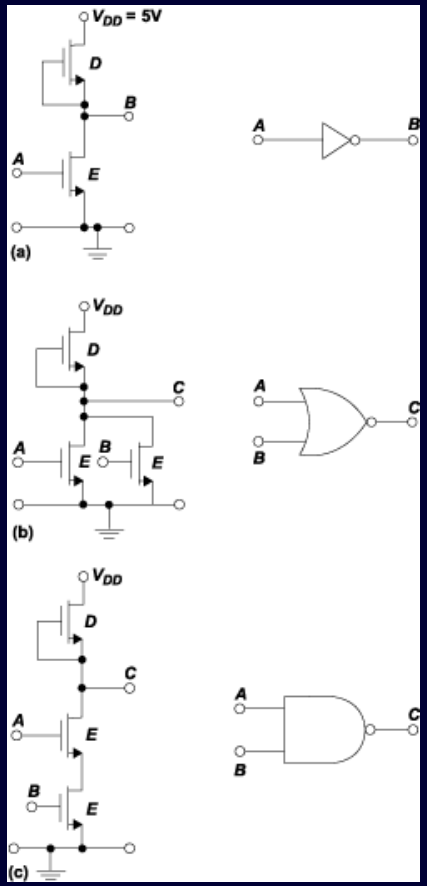
Logical Switches: Logical Irreversibility

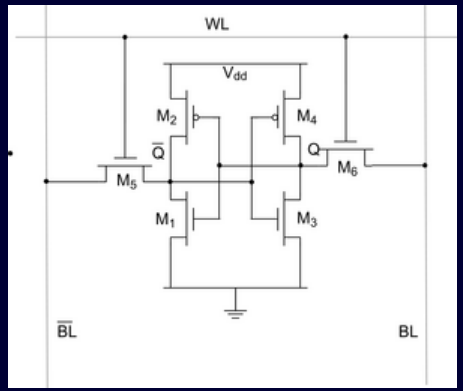


irreversible. We shall call a device *logically irreversible* if the output of a device does not uniquely define the inputs. We believe that devices exhibiting logical irreversibility are essential to computing. Logical irreversibility, we believe, in turn implies physical irreversibility, and the latter is accompanied by dissipative effects.

Landauer, 1961

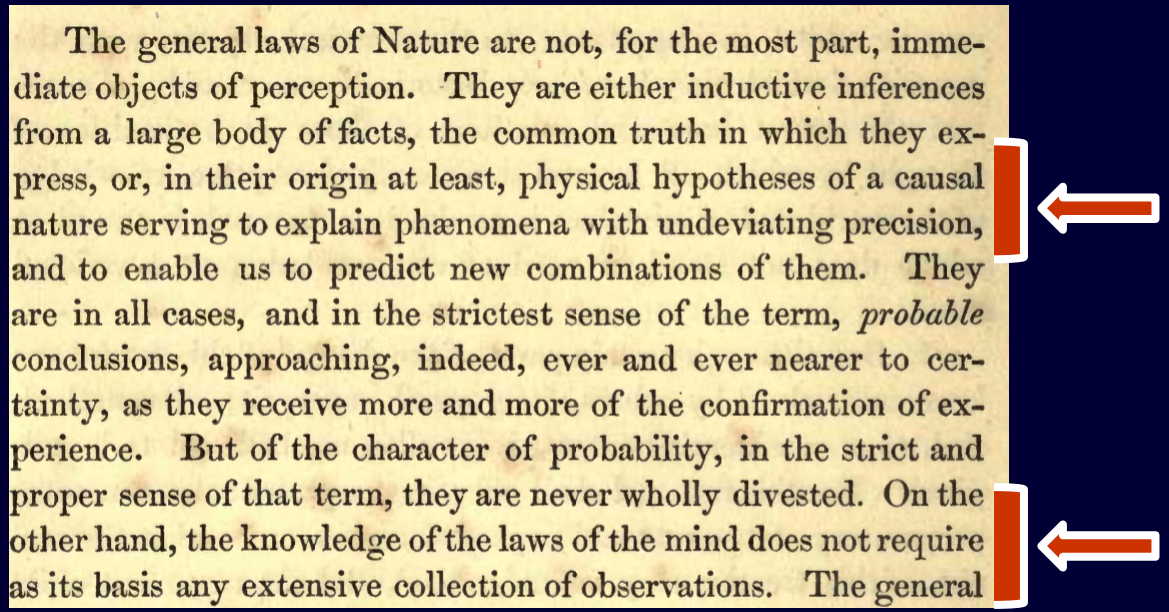
Logical Switches: Illustration (1)





SRAM - 6 switches

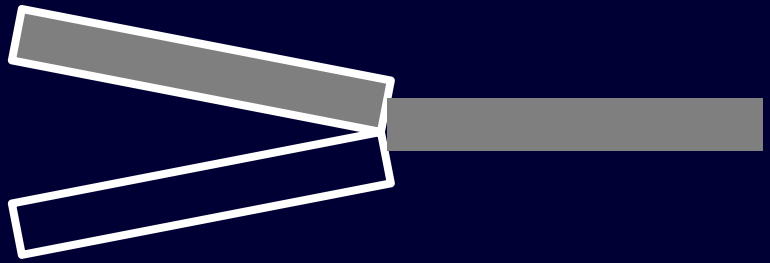
Boolean Logic vs Physical System

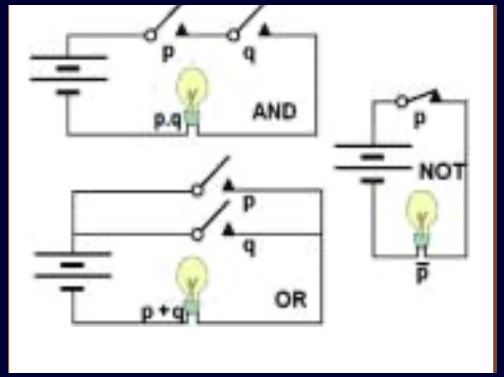


The general laws of Nature are not, for the most part, immediate objects of perception. They are either inductive inferences from a large body of facts, the common truth in which they express, or, in their origin at least, physical hypotheses of a causal nature serving to explain phænomena with undeviating precision, and to enable us to predict new combinations of them. They are in all cases, and in the strictest sense of the term, *probable* conclusions, approaching, indeed, ever and ever nearer to certainty, as they receive more and more of the confirmation of experience. But of the character of probability, in the strict and proper sense of that term, they are never wholly divested. On the other hand, the knowledge of the laws of the mind does not require as its basis any extensive collection of observations. The general

George Boole, 1854: *The Laws of Thought*

Boolean Logic vs Physical System





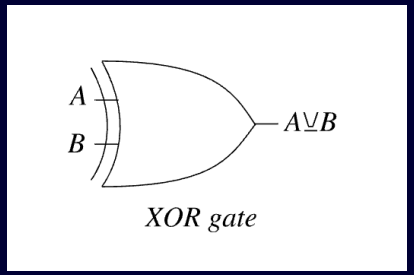
- OR gate implemented in a two connector pipe or in circuits
  - The Boolean logic operation does not follow conservation laws
  - However the conservation laws are still valid in the physical system
- No scientific reason for all internally consistent logical frameworks to follow conservation laws (note: the difference between *hardware vs architecture*)
  - But the converse is not true, as all conservation laws have their own consistent set of logical rules

Logical Switches: Illustration (2)

	Input	Input	Input	Output
OR	0	0	-	0
	0	1	-	1
	1	0	-	1
	1	1	-	1
NOR	0	0	-	1
	0	1	-	0
	1	0	-	0
	1	1	-	0
XOR	0	0	-	0
	0	1	-	0
	1	0	-	1
	1	1	-	1

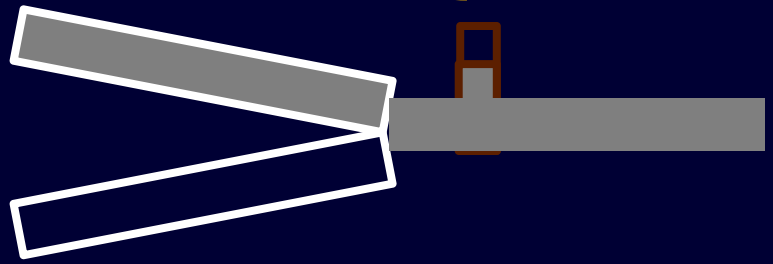
Logical Switches: Illustration (3)

(A OR B) AND (NOT (A) OR NOT (B))



Physics-based Gates

	Input	Input	Input	Output	Output
XOR	0	0	-	0	-
	0	1	-	0	-
	1	0	-	1	-
	1	1	-	1	-
Conserv	0	0	-	0	0
	0	1	-	0	1
	1	0	-	1	0
	1	1	-	1	1



Summary (1)

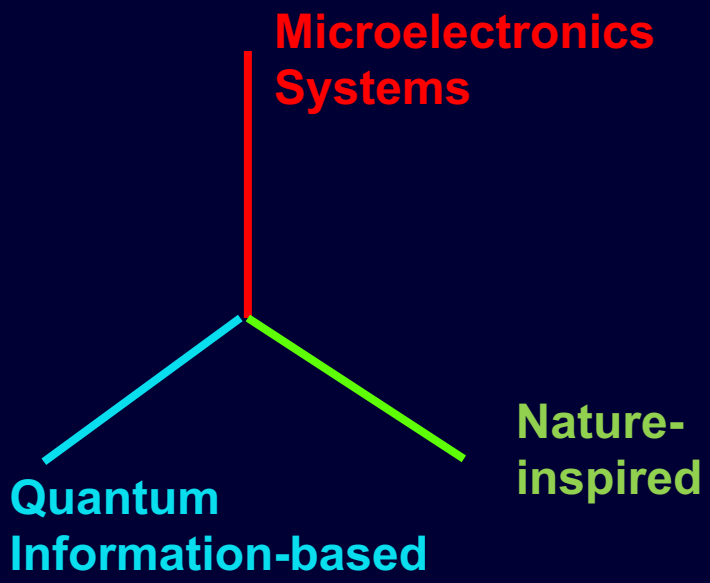
- Reversible gates increase transistor count (and the corresponding communication and memory overhead)
  - Still driven by Boolean logic
- Logical reversibility does not have to be tied to physical reversibility
  - Boolean logic is different from physical laws (e.g. unitary formalism)

Characteristics	**Reversible Gates**	**Current Formalism**
Intent	As logic is synonymous with physics, use reversible logic to minimize energy loss	Logic and energy conservation can be independent, use existing logic and conservative flow to minimize energy loss
Inputs	New gates: In addition to conventional 2 inputs, there is a control input	Conventional gates
Minimum Energy	Arbitrarily small as long as there is no memory erasure	Arbitrarily small
Dissipation	Assumes no dissipative process	Assumes no dissipative process

Energy Estimates in Computing:
Lessons from other domains

Three Different Dimensions for Computing
(Sep 14, 2022, EES2 Presentation)

- Microelectronics Systems: The current trajectories, using scaling and specialized architectures are on the Microelectronics Basis
- Nature-inspired represents all the information processing in nature from neuron synapses to photosynthesis
- Quantum Information-based is based on using quantum representations as units of computation and finding algorithms that can simulate quantum and classical processes

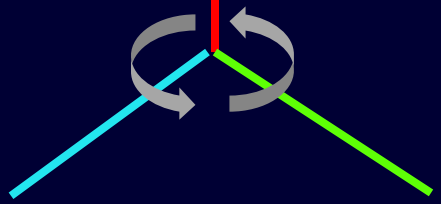


Map for Energy Efficiency

(Sep 14, 2022, EES2 Presentation)

Headroom for energy efficiency

Systems in Microelectronics
(Up to 10^8 or 100 Million)



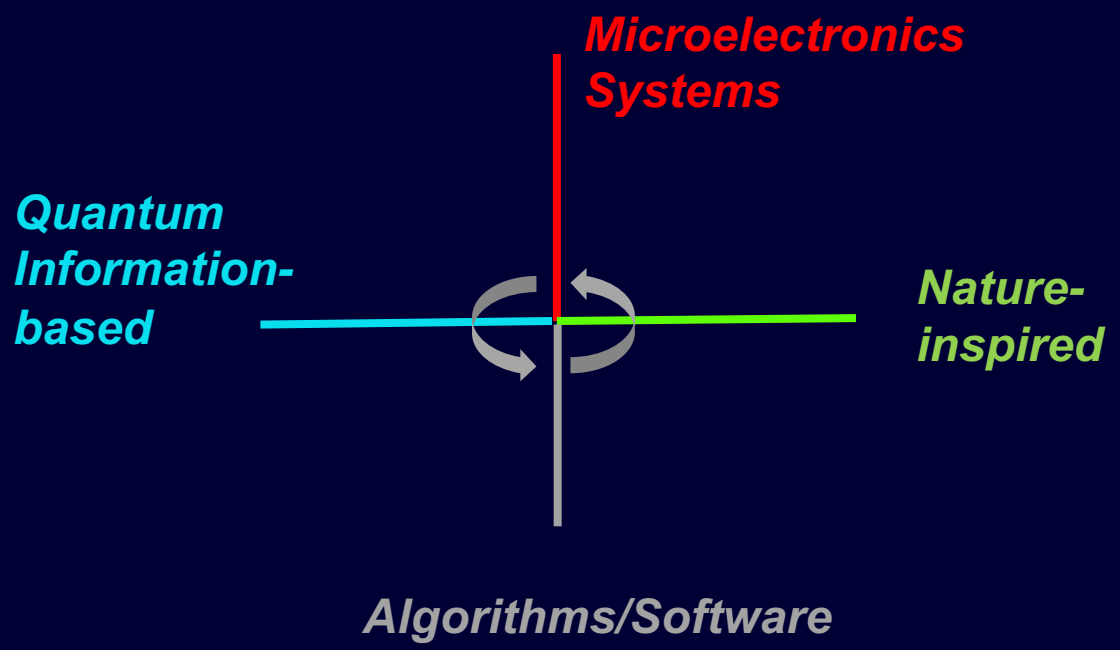
Quantum
(Up to 10^4 to 10^7 or 10,000 to 10 Million)

Nature-inspired
(Up to 10^6 to 10^8 or 1 to 100 Million)

~~Three~~ Four Different Domains in Computing

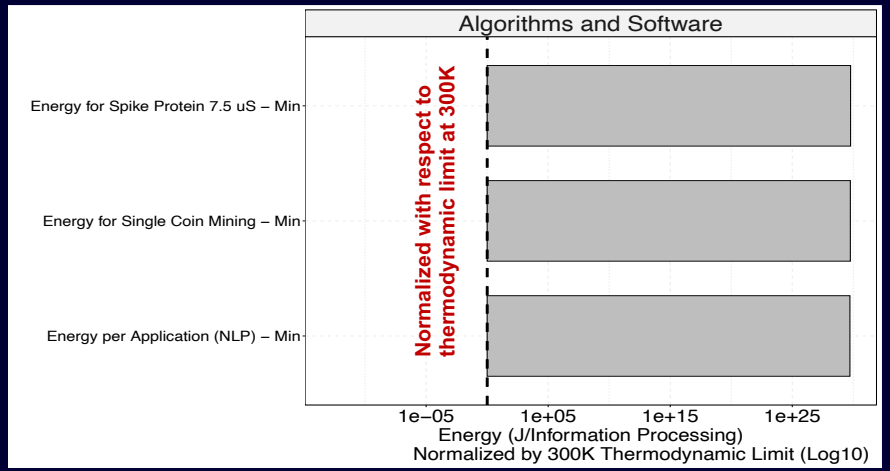
(Sep 14, 2022, EES2 Presentation)

- Microelectronics Systems:
- Nature-inspired Systems:
- Quantum Information-based Systems:
- Systems of Algorithms/Software:



Four Different Domains in Computing (1)

(normalized to the thermodynamic limit at 300K)

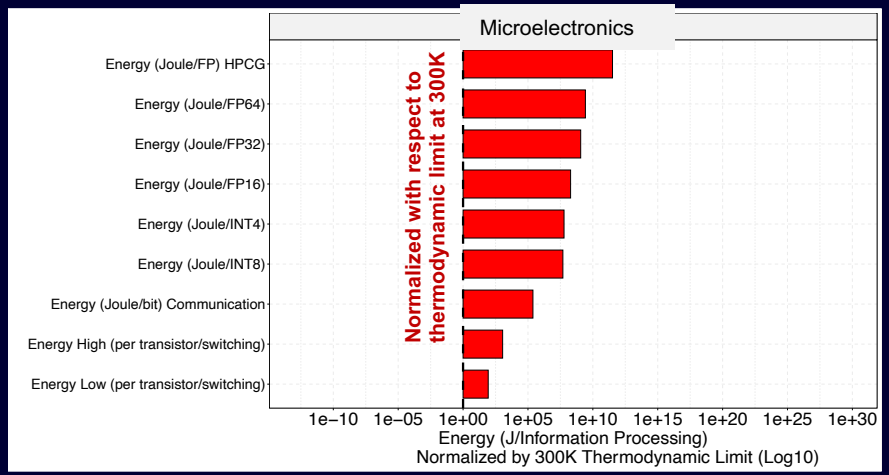


V. Shankar

- Algorithms and Software indicate significant demands on compute cycles
- Independent of scientific or ML algorithms, intrinsically energy intensive, given the relatively higher power requirements
- No signs of reduction in compute needs

Four Different Domains in Computing (2)

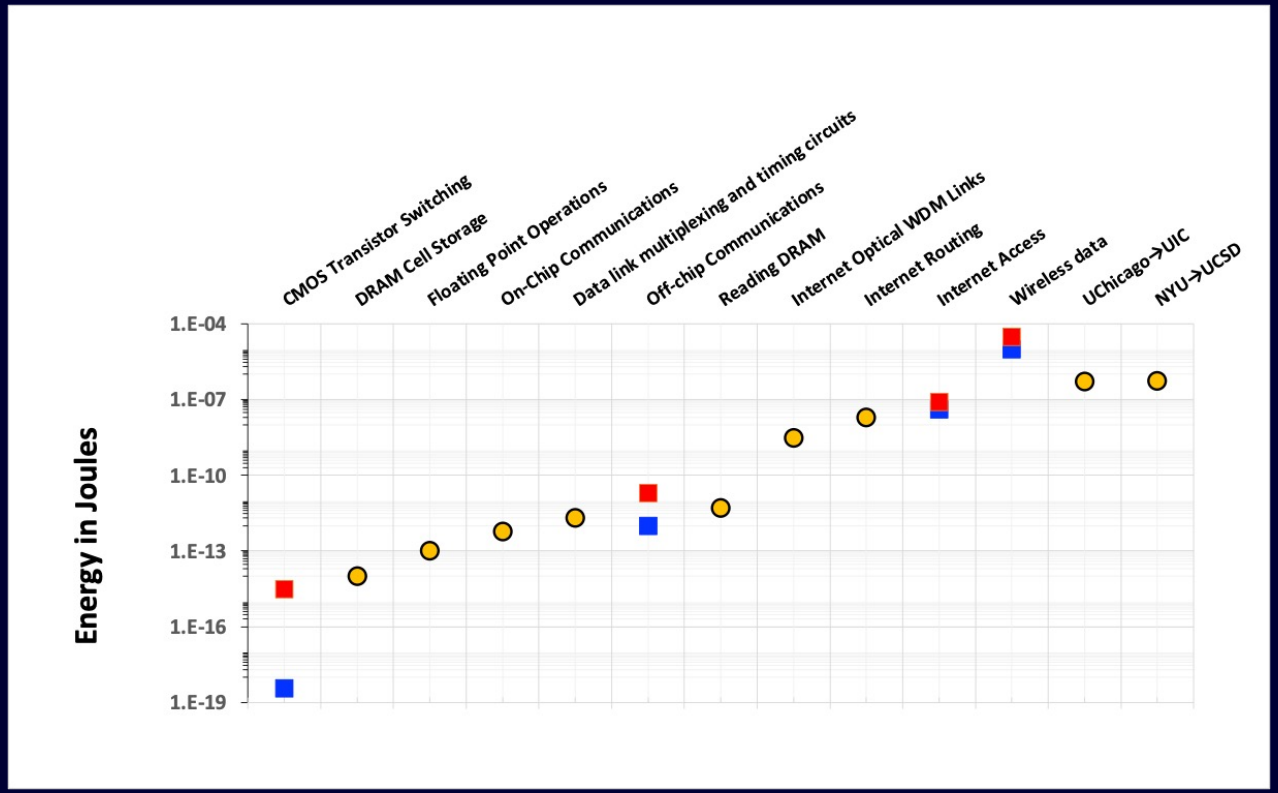
(normalized to the thermodynamic limit at 300K)



V. Shankar

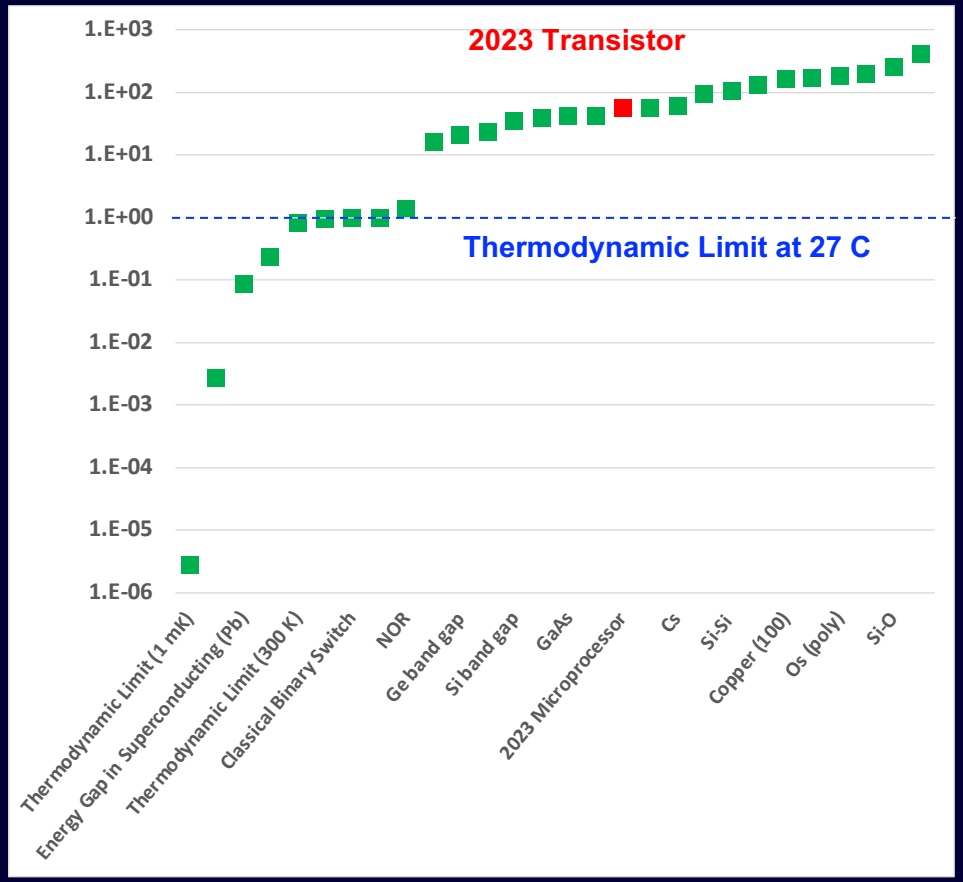
- Microelectronic Systems driven by Architecture, Hardware, and Information Processing themselves provide additional margins for energy reduction
  - Hint: 6-T SRAM
- Applications drive instructional requirements which in turn, drive energy needs

Computing: *From Transistor to Systems*



- Although transistors are energy efficient, other aspects add energy requirements
  - Memory
  - Floating point operations,
  - Communications on-chip
  - Communications off-chip
  - Routing
- All these aspects provide headroom for system level energy efficiency

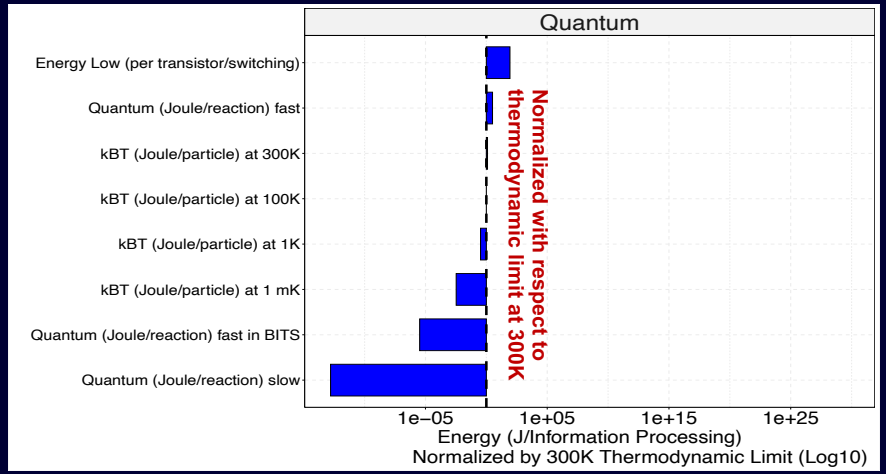
Computing: *From Transistor to Atoms*



- Normalized Energy Estimates with respect to thermodynamic limit at 27 C
- To reduce energy from the current state, material innovations at the atomic level may be needed
- Cryo engineering and Quantum effects are also relevant

Four Different Domains in Computing (3a)

(normalized to the thermodynamic limit at 300K)

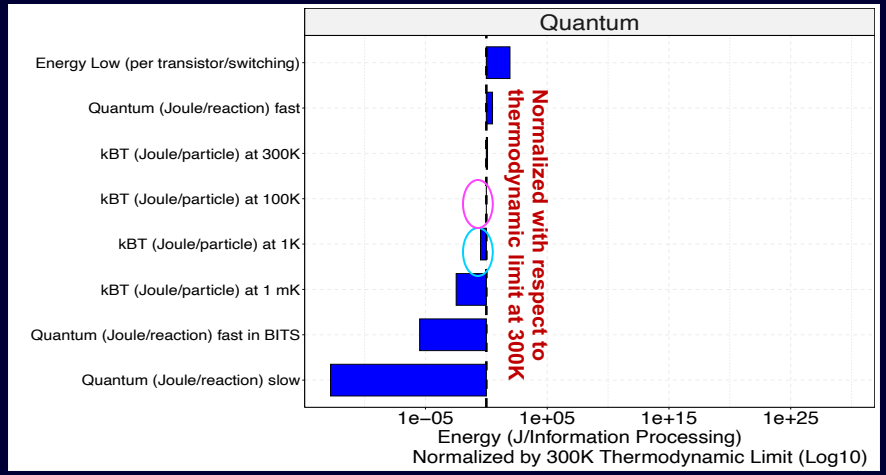


V. Shankar

- Quantum information processing indicates lower energy due to:
  - Lower temperatures
  - Computing representations, information bases etc..
- Reducing rate of processing seem to indicate lower energy
  - Consistent with classical computing

Four Different Domains in Computing (3b)

(normalized to the thermodynamic limit at 300K)



V. Shankar

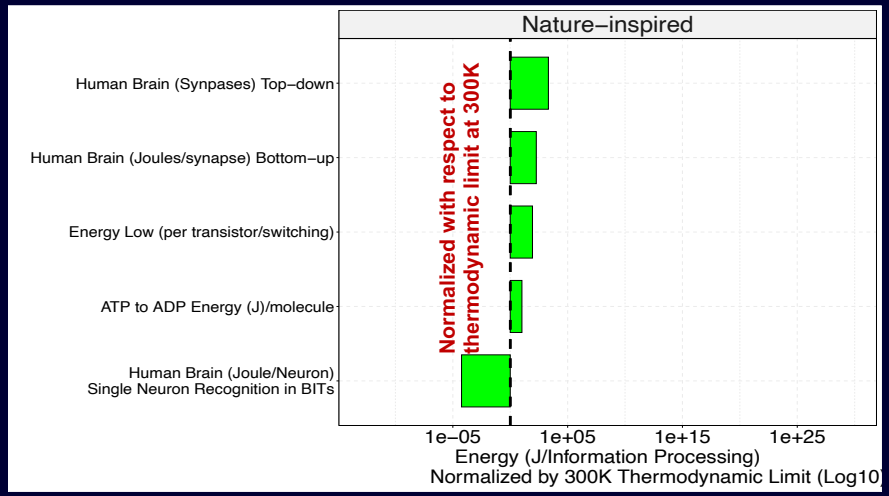
Energy gaps for elemental superconductors

Energy gaps for High Temperature compound superconductors

- Quantum information processing indicates lower energy due to:
  - Lower temperatures
  - Computing representations, information bases etc..
- Reducing rate of processing seem to indicate lower energy
  - Consistent with classical computing

Four Different Domains in Computing (4)

(normalized to the thermodynamic limit at 300K)



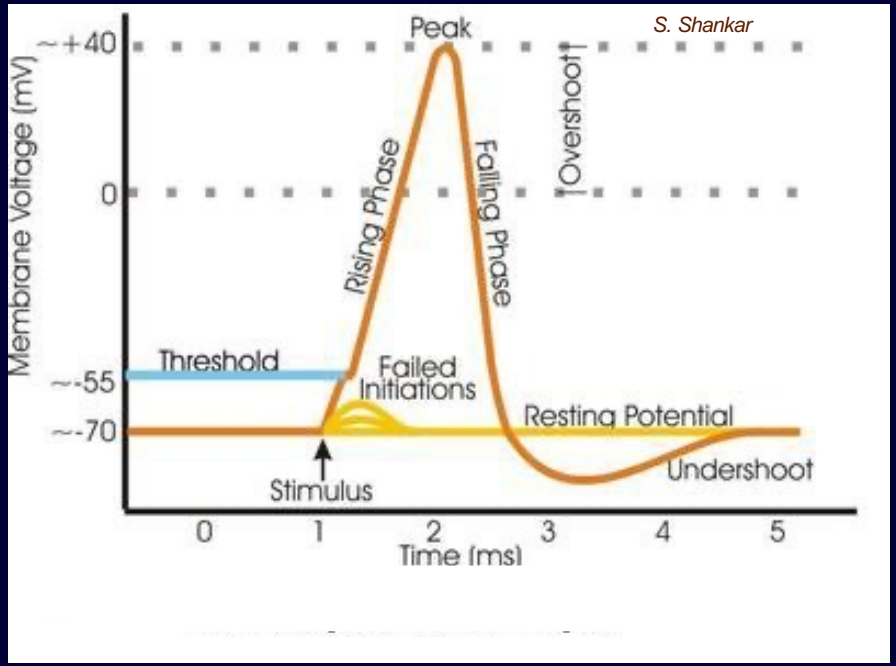
V. Shankar

- The quantal of energy is close to the thermodynamic limit
- Biological systems including human synapse operates close to the thermodynamic limit
- Information processing is more efficient

Single Neuron Switching as a Spike

Membrane Voltage
+40 to -70 mV

Thermodynamic Limits
25- 40 meV

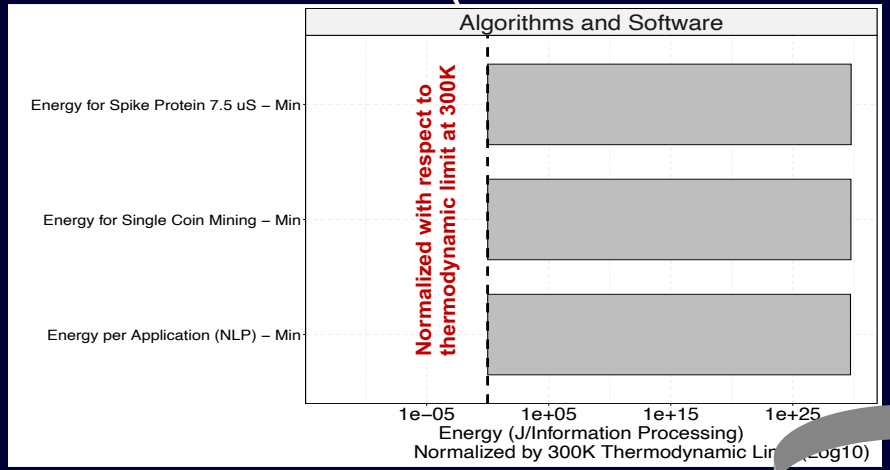


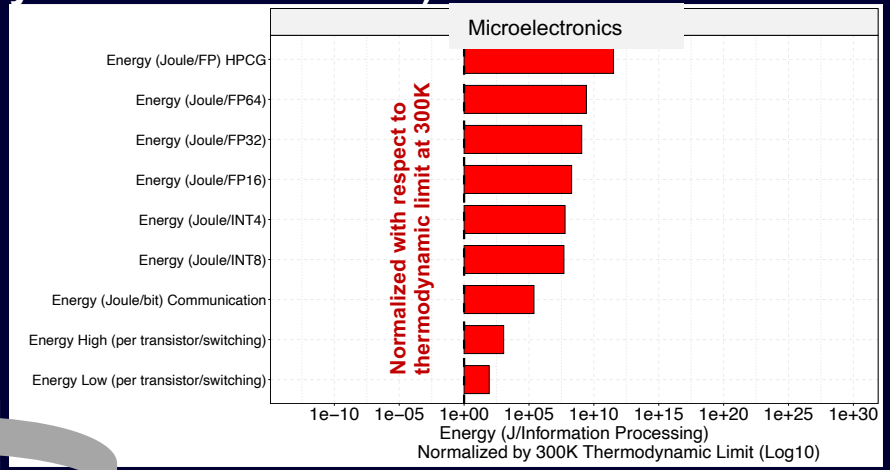
- **Conservation of Energy/Neuron across all species**

S.Shankar

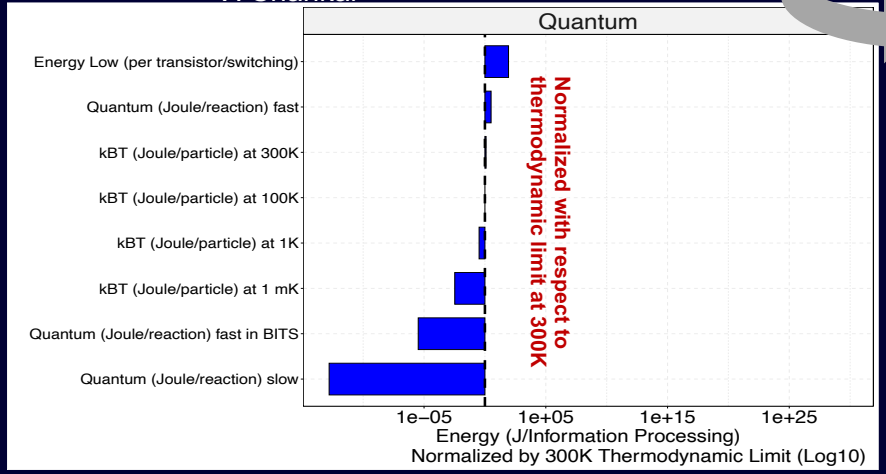
Map for Energy Efficiency: Combining Solutions and Concepts

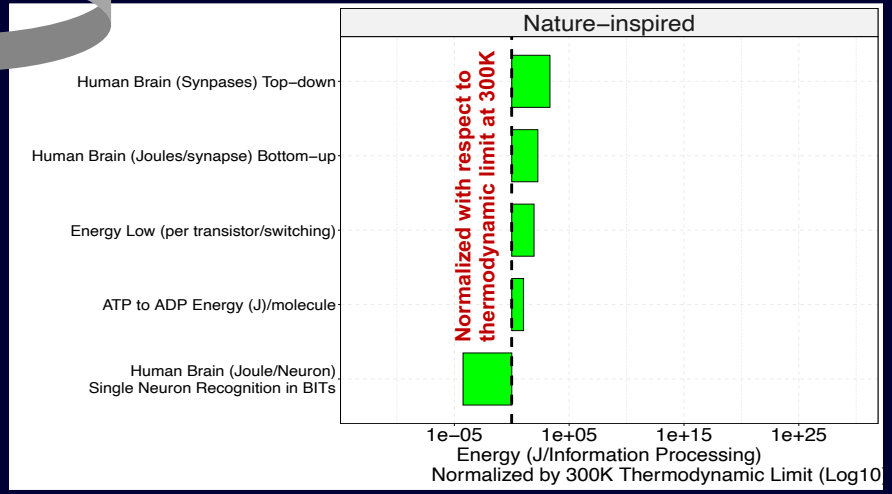
(normalized to the thermodynamic limit at 300K)





V. Shankar





Summary

Summary (2a)

- Energy density-wise, compared to natural and synthetic systems, computing appears to be more energy intensive
  - Incompatible with the principles of evolution
  - Microprocessors and Large-scale computing platforms are outliers compared to Natural

- Lower energy limits of computing is determined only by thermodynamics and cryogenic constraints
  - Large computers still follow human design of other energy intensive systems
  - Information abstraction, representations, and translation
  - Signal/Noise

Summary (2b)

- Energy in Computing is a “3E”+ problem
  - Most efficient energy (electronic) is converted to least efficient energy (heat) at the speed of computing
    - *First*, to compute
    - *Second,* to cool the heat generated (electrical energy to thermal energy at the speed of computing)
    - *Third,* Cooling and Refrigeration
    - *Fourth*, Digitalization and Energy/Information distribution network

- Energy Efficiency in Computing can be achieved by combining concepts:
  - Combing Architecture/Hardware solutions in conjunction with Algorithms/Software
  - From nature-inspired and quantum information

- Measurement of energy across the layers of computing is a key element……

Summary (2c)

- Energy density-wise, compared to natural and synthetic system, computing appears higher
  - Incompatible with the principles of evolution
  - Microprocessors and Large-scale computing platforms are outliers
- Lower energy limits of computing is determined only by thermodynamics and cryogenic constraints
  - Information abstraction, representations, and translation
  - Signal/Noise
- Map for Energy Efficiency in Computing

Acknowledgements and References

Thanks!

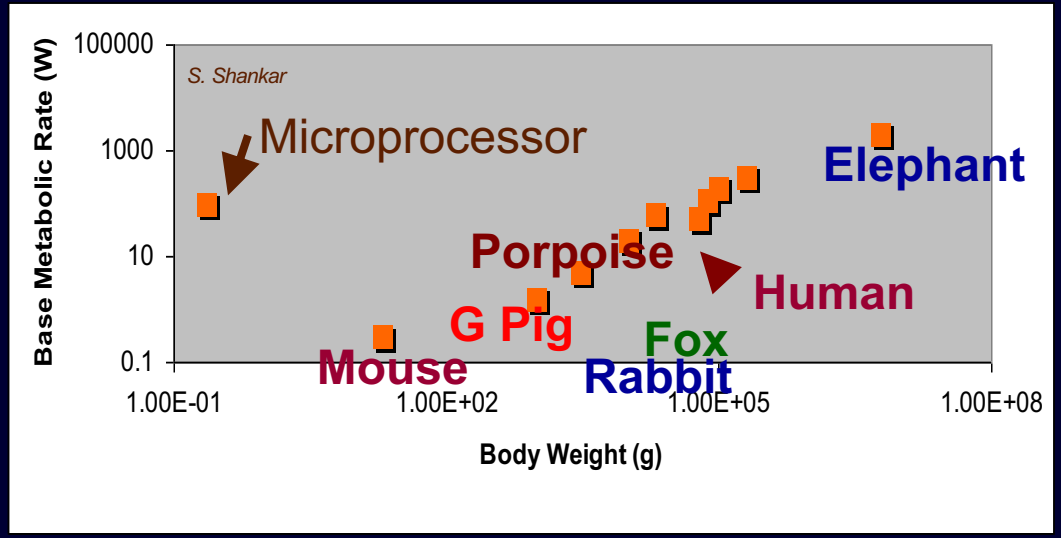
- Partially funded by the U.S. Department of Energy's Office of Science contract DE-AC02-76SF00515 with SLAC (T. Kaarsberg/EES2 Team)
- A. Reuther (MIT-LL)
- SLAC National Lab support (P. McIntyre)

Papers

1. Shankar, S. 2021 "*Lessons from Nature for Computing: Looking beyond Moore's Law with Special Purpose Computing and Co-design",* 2021 IEEE High Performance Extreme Computing Conference (HPEC) (pp. 1-8).
2. Shankar, S, Reuther, A, 2022, "*Trends in Energy Estimates for Computing in AI/Machine Learning Accelerators, Supercomputers, and Compute-Intensive Applications*", 2022 IEEE High Performance Extreme Computing Conference (HPEC)
3. Shankar, S., *Energy Estimates Across Layers of Computing: From Devices to Large-Scale Applications in AI/Machine Learning in Natural Language Processing, Scientific Computing, and Crypto coin Mining* (submitted to IEEE HPEC, 2023)
4. *A Logical Framework for Information Processing* (in preparation)
5. *Energy as a Design Variable for Computing* (in preparation)

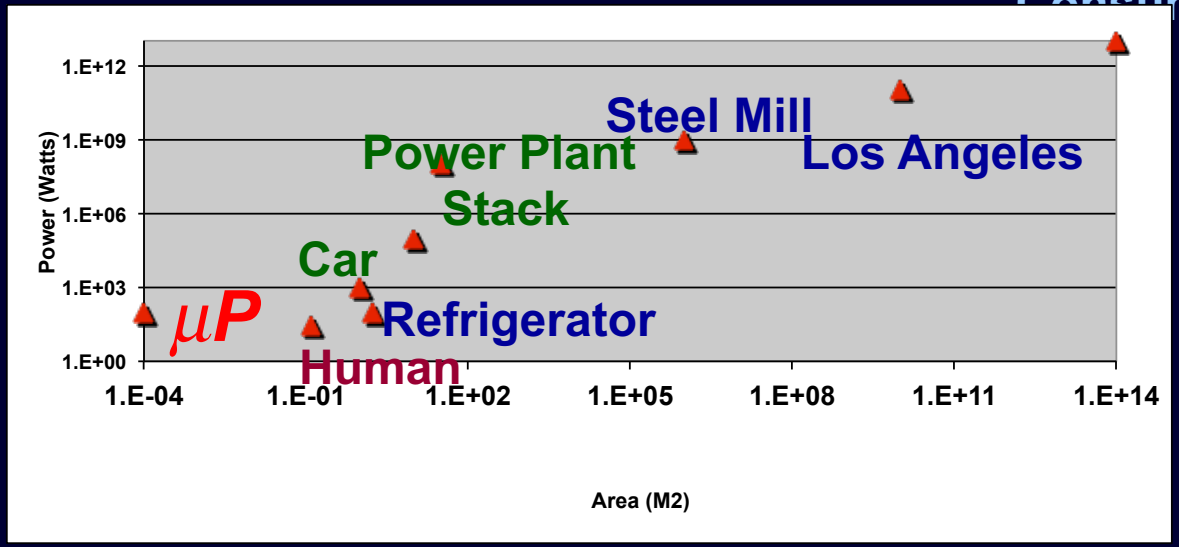
Backup Slides

Natural Systems



- Evolution will **not** let microprocessor survive naturally
- Limiting merging with biological *Singular Point* due to energy mis-match

Man-made Systems

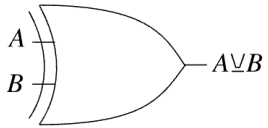


- Energy Consumption Scaling with Area
- Higher Power is associated with larger areas, except μ Processor

Logical Switches: Illustration (3)

XOR

Download Wolfram Notebook



XOR gate

A connective in logic known as the "exclusive or," or exclusive disjunction. It yields true if exactly one (but not both) of two conditions is true. The XOR operation does not have a standard symbol, but is sometimes denoted $A \veebar B$ (this work) or $A \oplus B$ (Simpson 1987, pp. 539 and 550-554). $A \veebar B$ is read "A aut B," where "aut" is Latin for "or, but not both." The circuit diagram symbol for an XOR gate is illustrated above. In set theory, $A \veebar B$ is typically called the symmetric difference. The XOR function is implemented as Xor[*predicate1*, *predicate2*, ...].

The binary XOR operation $A \veebar B$ is identical to nonequivalence $A \not\equiv B$. $A \veebar B$ can be implemented using AND and OR gates as

$$A \veebar B = (A \wedge !B) \vee (!A \wedge B) \tag{1}$$

$$= (A \vee B) \wedge (!A \vee !B), \tag{2}$$

where $\wedge$ denotes AND and $\vee$ denotes OR, and can be implemented using only NOT and NAND gates as

$$A \veebar B = (A \barwedge !B) \barwedge (!A \barwedge B) \tag{3}$$

Logical Switches: Illustration (4)

	Input	Input	Input	Output	Output	Output
TOF	0	0	0	0	0	0
	0	0	1	0	0	1
	0	1	0	0	1	0
	0	1	1	0	1	1
	1	0	0	1	0	0
	1	0	1	1	0	1
	1	1	0	1	1	1
	1	1	1	1	1	0
Conserv	0	0	0	0	0	0
	0	0	1	0	0	1
	0	1	0	0	1	0
	0	1	1	0	1	1
	1	0	0	1	0	0
	1	0	1	1	0	1
	1	1	0	1	1	0
	1	1	1	1	1	1

Physics-based Gates