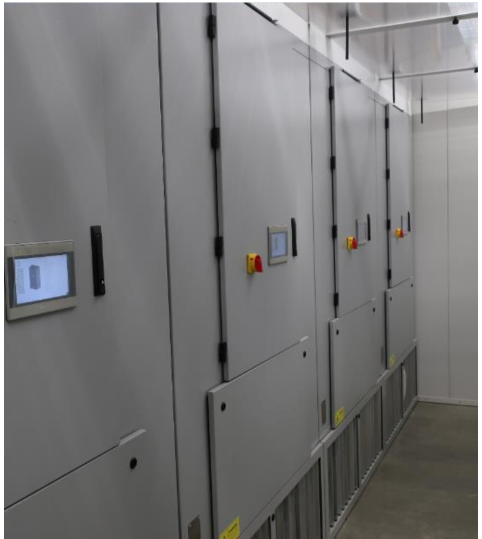
REVIEWING THE RELATIONSHIP BETWEEN INFORMATION AND ENERGY, AND THE PHYSICAL LIMITS OF COMPUTATION

Dr Jon Summers
Scientific Leader in Data Centers, RISE
Adjunct Professor of Fluid Mechanics, Lulea University of Technology, SWEDEN
Senior Lecturer in Thermofluids, University of Leeds, UK

ICT DIVISION, Research Institutes of Sweden

PPAM 2019 Bialystok, Poland, September 8-11, 2019






RISE ICE

- 15 projects, from the ground to the cloud
- 20 employees
- 3 MEUR turnover
- Established 2016

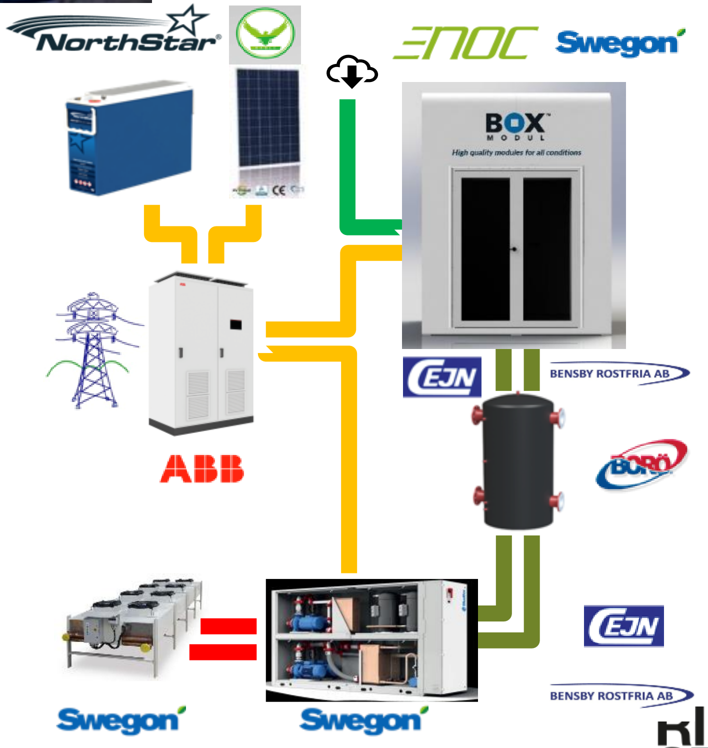
RISE

Best Data Center Initiative of the Year

with RISE SICE North Research Data Center

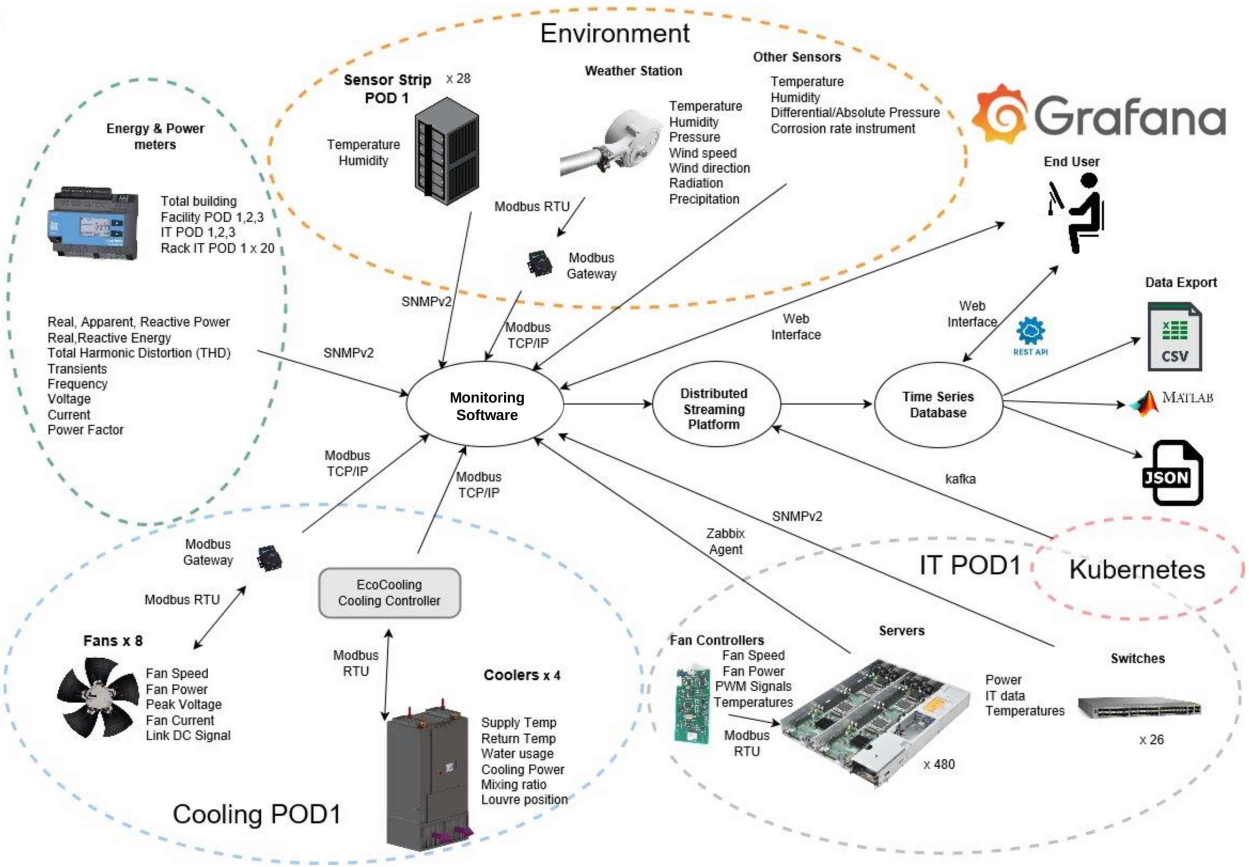
Sponsored by DCPRO

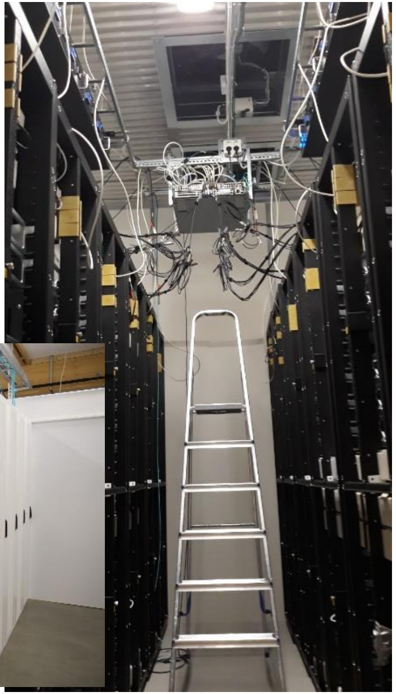
RISE ICE facilities



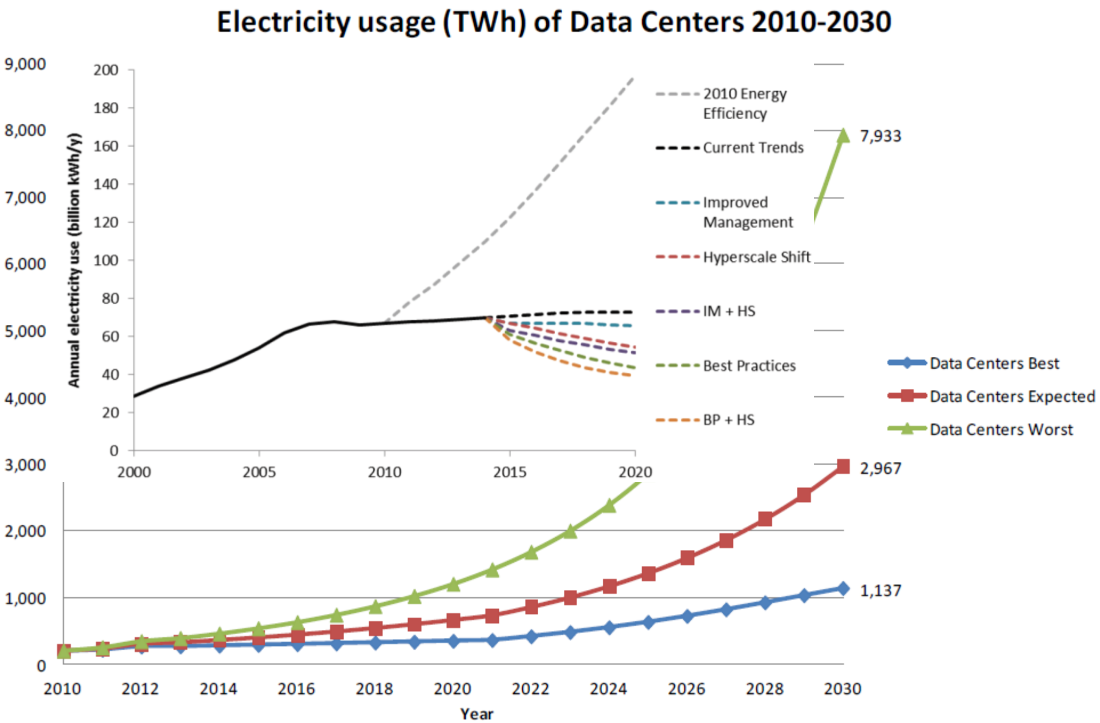
RI. SE

Open source monitoring system



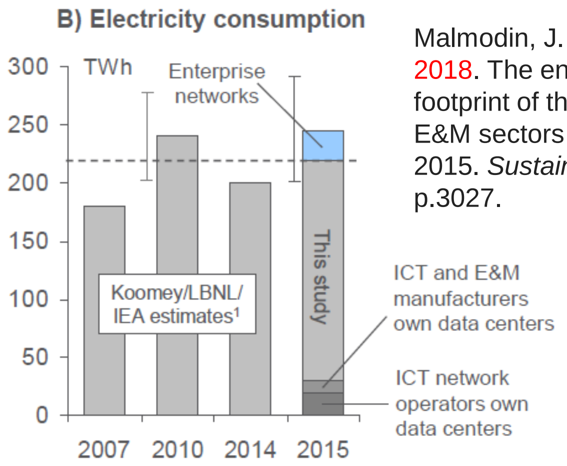


Multiples studies on consumption

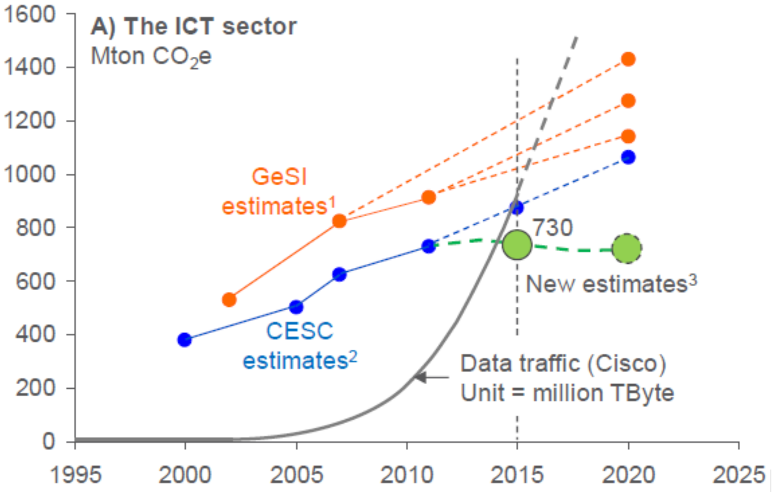


Andrae, A. and Edler, T., 2015. On global electricity usage of communication technology: trends to 2030. *Challenges*, *6*(1), pp.117-157.

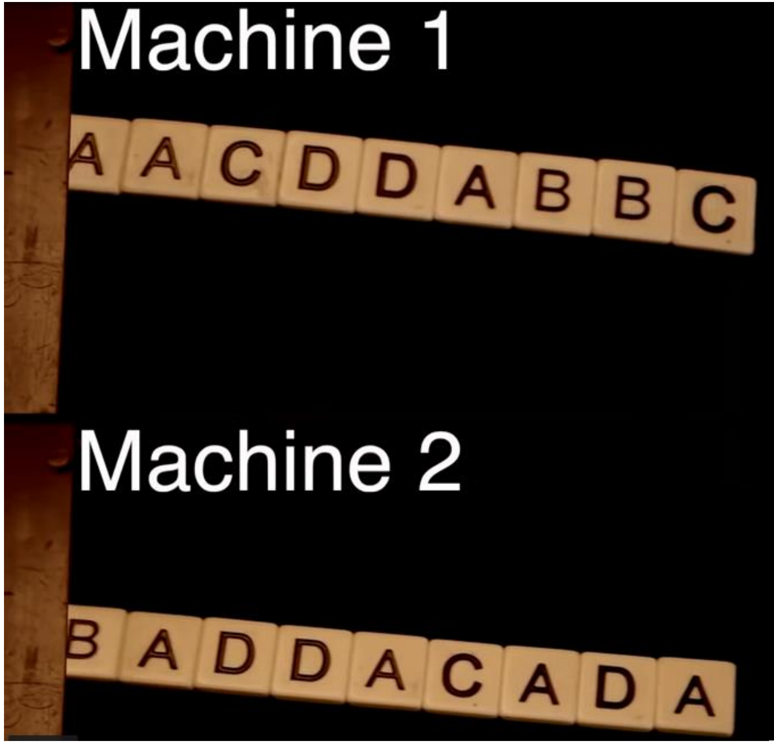
Shehabi, A., Smith, S., Sartor, D., Brown, R., Herrlin, M., Koomey, J., Masanet, E., Horner, N., Azevedo, I. and Lintner, W., 2016. United states data center energy usage report.



Malmodin, J. and Lundén, D., 2018. The energy and carbon footprint of the global ICT and E&M sectors 2010–2015. *Sustainability*, *10*(9), p.3027.



What is information? A quick look back at Shannon



Machine 1 produce random information
P(A)=P(B)=P(C)=P(D)=0.250

Machine 2 produces information according
P(A)=0.500
P(B)=0.125
P(C)=0.125
P(D)=0.250

Shannon asked how to predict the next symbol

- **Machine 1**. Ask binary questions. Is it (A or B) or (C or D)? If A or B is YES. Then ask a second binary question. Is it A or B? If B is YES, then the uncertainty of Machine 1 is 2 questions per symbol.
- **Machine 2**. Ask binary questions. Is it A? If YES then only 1 question. If NO then ask question: Is it D? If YES then we asked 2 questions. If NO then ask question: Is it B or C?

RI. SE

P(A)=0.500
P(B)=0.125
P(C)=0.125
P(D)=0.250

How many questions on average?

- Machine 2: Ask 1 question 50% of the time to guess A, 2 questions 25% of the time to guess D, 3 questions 25% of the time to guess B or C. So the average number of questions to ask is:
- #questions = 1 x P(A) + 2 x P(D) + 3 x P(B) + 3 x P(C), which is 1.75 questions per symbol on average.
- So if we need to guess 100 symbols from both machines we would need to ask 200 questions for Machine 1, but 175 questions for Machine 2. Machine 2 is producing less information! Less uncertainty or surprise in its output.

Shannon's information entropy

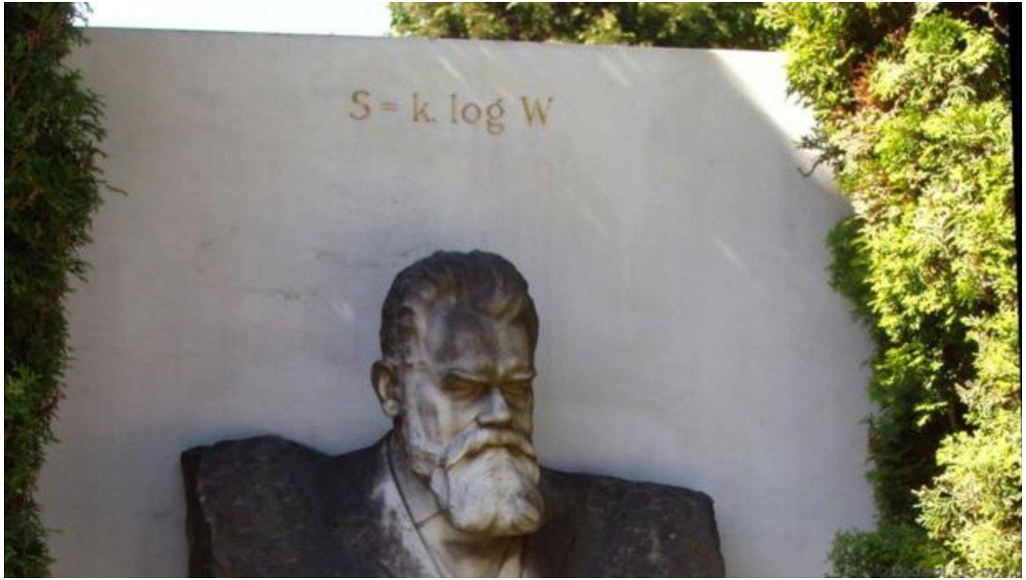
- Claude Shannon called this measure of uncertainty, information entropy, using the symbol H.
- The units of H are based on the uncertainty of a fair coin flip. Shannon used "bit" based on a fair coin flip.
- $H(p_1, p_2 \dots\dots p_n) = \sum_i p_i \times \#questions_i$
- $\#questions_i = log_2(\#outcomes) = log_2\left(\frac{1}{p_i}\right)$
- $H(p_1, p_2 \dots\dots p_n) = -\sum_i p_i log_2(p_i)$

Boltzmann (thermodynamic) Entropy

- In terms of a dilute gas the "thermodynamic" entropy, S, is written as

  $$S = k_B ln\mathrm{W}$$

  where W is the number of real microstates of the gas



- In statistical mechanics, a microstate is a specific microscopic configuration of a thermodynamic system that the system may occupy with a certain probability in the course of its thermal fluctuations.

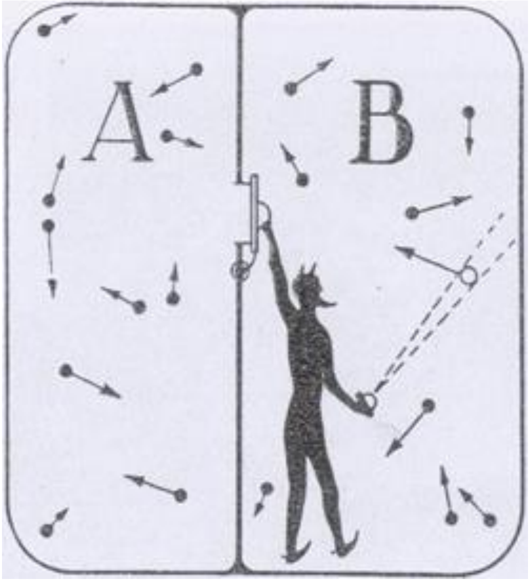
Boltzmann's constant

- A universal constant that relates a gas molecules kinetic energy to its temperature, so it is measure in J/K – Joules per Kelvin
- The value is 1.380649×10^{-23} J/K and is equal to the Universal Gas Constant, R, divided by Avogadro's constant, N_A. R is the energy required to raise 1 mole of a substance by 1 Kelvin and N_A is the number of molecules in 1 mole of a substance.
- Energy required to raise 1 molecule by 1 Kelvin.
- So $S = -k_B \sum_i p_i ln(p_i)$ which reduces to $S = k_B ln\mathrm{W}$ if all probabilities, p_i, are equal.

RI. SE

Maxwell’s Demon

- Thought experiment proposed by James Clerk Maxwell in 1867.
- The Demon sorts hot and cold (respectively fast and slow moving) particles
- End up with oven (A) and fridge (B) and no energy **consumed**!
- Contradicts second law of thermodynamics (creating order from disorder – reducing Entropy).
- Does it imply a relationship between **information** and **energy**?



Szilard's answer

- An attempt to explain the paradox of Maxwell's demon was put forward by Szilard in 1929.
- Szilard argued that there must be an entropic cost associated with the Demon's acquisition of ***information***.
- Boltzmann's statistical mechanics definition of entropy involves probability of microstates.
- Therefore an increase in ***information*** corresponds to a decrease in ***entropy***. (Brillouin later used ***negentropy***)
- There is still much debate, see the 2015 article – "A few exciting words": Information and Entropy revisited.

Szilard, Leo. über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen Journal Zeitschrift für Physik Volume 53, Issue 11-12 (1929) , pp 840-856.

Brillouin, Leon. "Maxwell's demon cannot operate: Information and entropy. I."*Journal of Applied Physics* 22.3 (1951): 334-337.

Bawden, David, and Lyn Robinson. ""A few exciting words": Information and entropy revisited." *Journal of the Association for Information Science and Technology* (2015).

Information is a physical entity

- Information is physical: writing on stone, printing text in a book – difficult to reverse so *thermodynamic entropy* always increases.
- In 1961, Rolf Landauer, while working at IBM proposed that the acquisition of one bit of information through erasure of 0 or 1 required dissipation of at least $k_B \sum_i p_i ln(p_i) T$ J of energy at a temperature T - probabilities of 0 or 1 are p_1 and p_2 and are equal at 0.5.

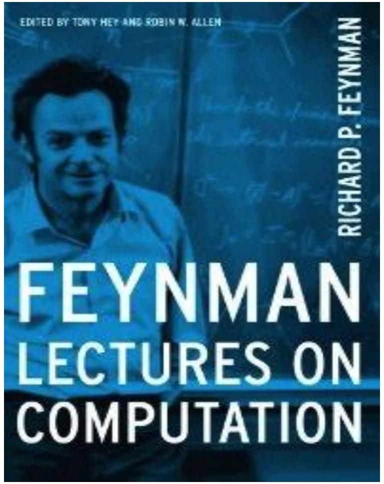
- In principle this assumed no thermodynamic objection to a logically reversible operation.
- In 2012 a team of French researchers published in Nature experimental verification to support the Landauer principle.

Rolf Landauer (1961), "Irreversibility and heat generation in the computing process" , *IBM Journal of Research and Development* **5** (3): 183–191

Antoine Bérut; Artak Arakelyan; Artyom Petrosyan; Sergio Ciliberto; Raoul Dillenschneider; Eric Lutz (8 March 2012), "Experimental verification of Landauer's principle linking information and thermodynamics", *Nature* **483** (7388): 187–190

Information and Energy

- Of course, there are critics such as Norton who identifies thermal fluctuations as a missing component of Landauer's argument
- Rolf Landauer demonstrated that the minimum dissipation of energy in the erasure of 1 bit of information is **2.9zJ** (z = Zepto = = 10^{-21}) at 300K (27^{O}C).
- Bennett's digital tape machine as discussed in Feynman's Lectures on Computation shows that at room temperature a tape carrying a full fuel load, **2.9zJ per bit**, carries zero information.
- Could this value of 2.9zJ per bit be a physical limit of computation?



Norton, John D. "Waiting for landauer." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 42.3 (2011): 184-198.

Bennett, C.H., 1982. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, *21*(12), pp.905-940.

RI.
SE

Data centre power consumption

Consider the contributing factors to Data Centre Power Consumption:

$$\text{Power} = (N_{compute} \times \textbf{Power/compute}) \times \text{PUE}$$

Competition between:
Demand UP and
Consolidate DOWN

Efficiency of ICT equipment is a function of Moore's Law

10 years of PUE have helped to **reduce overhead** of a data centre end use energy consumption

RI.
SE

Power consumption of IT hardware

Power/compute = N_{tr} x Freq x E_{tr} x CompUE

Number of transistors per datacom has increased for 50 years doubled ever 2 years by Moore's law and indicates performance.

Clock speeds have not really increased since 2005 as it has a significant effect, but is now variable.

Energy consumption per transistor is key to total power consumption.

Compute Usage Effectiveness Overhead from power supply, xDD, RAM, etc.

Note also that Power/compute = α **x** C x V^2 x Freq + leakage

RI.SE

Energy consumption of a transistor

$$E_{tr} = E_{FACTOR} \times (k_B \times T)$$

Energy/Entropy Factor related to the approach of state changes in Field Effect Transistors (FETs): Depends on Voltage and materials.

Physical constant used statistical mechanics, called the Boltzmann constant with a value of 1.38 x 10^{-23} J/K

Temperature at which the transistor is operating.

Summary of relationship of E_{FACTOR} with Data Center Power

$$\underline{\textbf{Power}} = (N_{compute} \times \textbf{Power/compute}) \times PUE$$

$$\textbf{Power/compute} = N_{tr} \times Freq \times \textbf{E}_{\textbf{tr}} \times CompUE$$

$$\textbf{E}_{\textbf{tr}} = \underline{E_{FACTOR}} \times (k_B \times T)$$

Effect of multicore and scaling of gate lengths

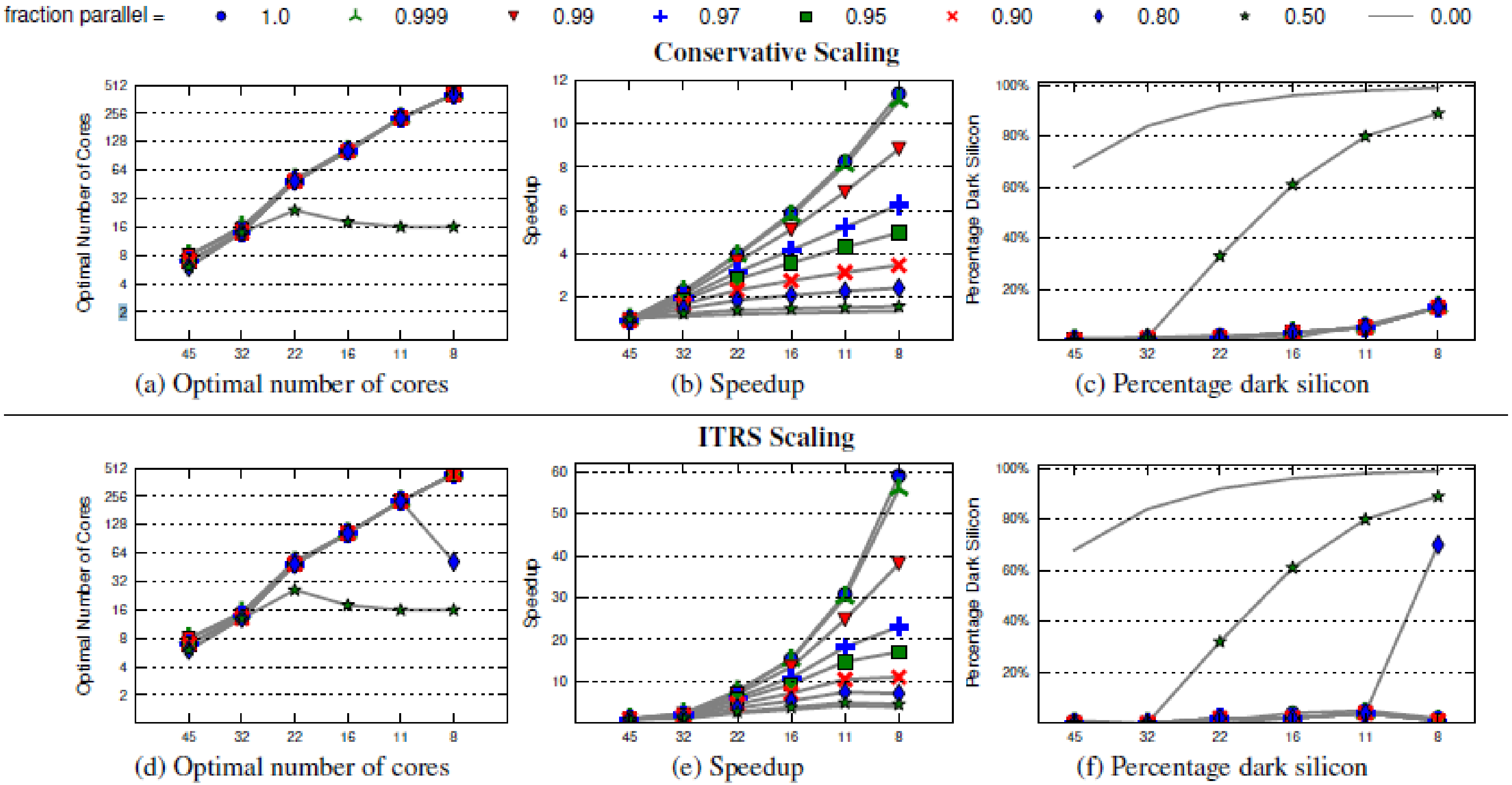


Figure 4: Amdahl's law projections for the dynamic topology. Upperbound of all four topologies (x-axis: technology node).

Esmaeilzadeh, H., Blem, E., Amant, R.S., Sankaralingam, K. and Burger, D., 2011, June. Dark silicon and the end of multicore scaling. In *Computer Architecture (ISCA), 2011 38th Annual International Symposium on* (pp. 365-376). IEEE.

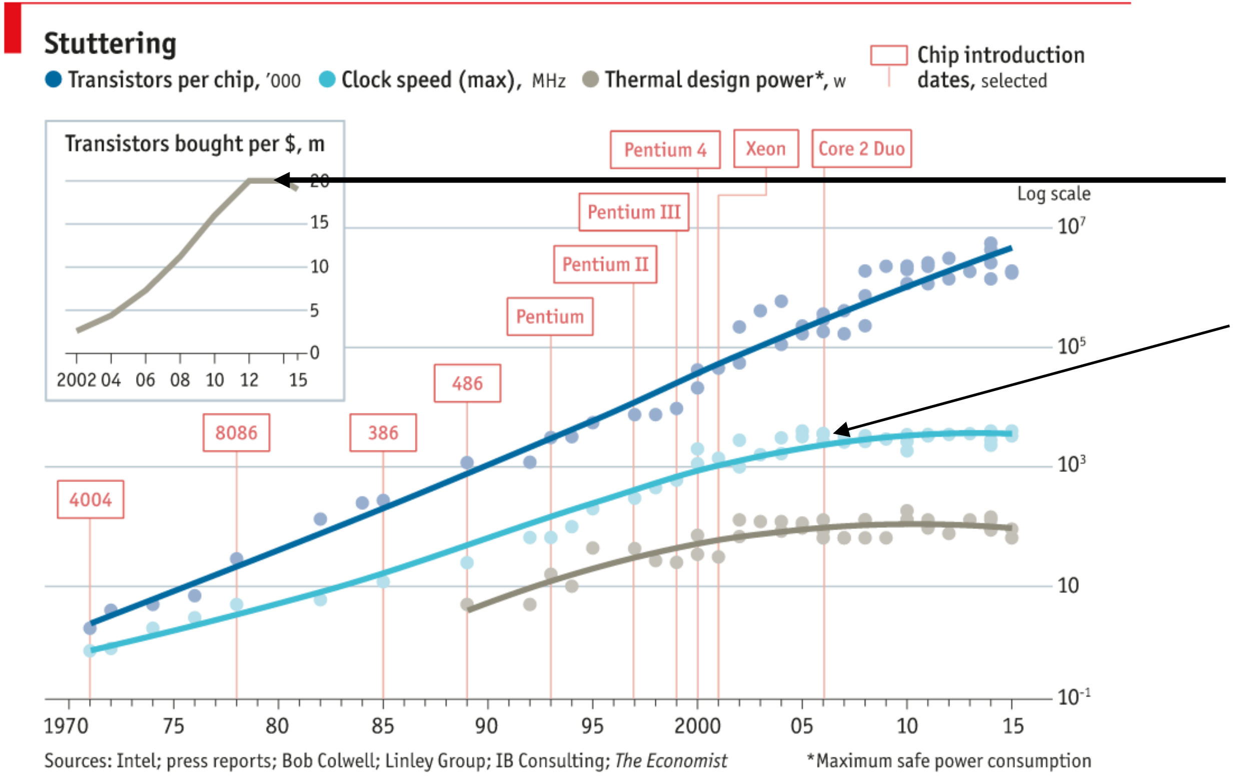
Switching Energy up to today.

Processor Architecture	Year	Feature Size	Switching Energy (zJ)
Pentium 486	1989	600nm	41137803
Pentium M	2003	130nm	325143
Core	2006	65nm	560819
Nehalem	2008	45nm	156566
Sandy Bridge	2012	32nm	74555
Ivy Bridge	2014	22nm	28994
Broadwell	2015	14nm	24852
Intel Xeon Plat 8180	2017	14nm	18743
AMD Epyc	2017	14nm	13889
Qualcomm Centriq 2400	2018	10nm	10256

Dark Silcon
Not all transistors are operational.

New TriGate FinFETS ~ 3D!

E_{FACTOR} is linked to Moore's Law



Cost of transistors is going up. Peaked at 20 million per $ in 2015

End of Dennard scaling.

Moore's Law: Self-fulfilling prophecy to provide double the number of transistors in the same area every two years.

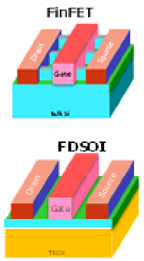
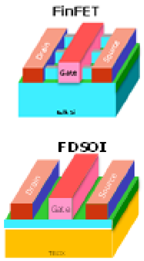
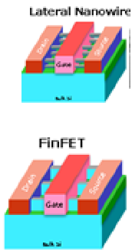
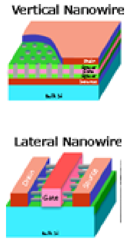
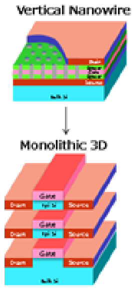
Cross, T. "After Moore's Law: Double, double, toil and trouble." *The Economist, Technology Quarterly, Quarter* 1 (2016).

Getting E_{FACTOR} down.

- millivolt, transistor size and materials may reduce the E_{FACTOR} , or going 3D, like gate-all-around.
- **Waldrop** quotes "*My bet is that we run out of money before we run out of physics*" [Rock's Law]

Waldrop, M. Mitchell. "The chips are down for Moore's law." *Nature News* 530.7589 (2016): 144.

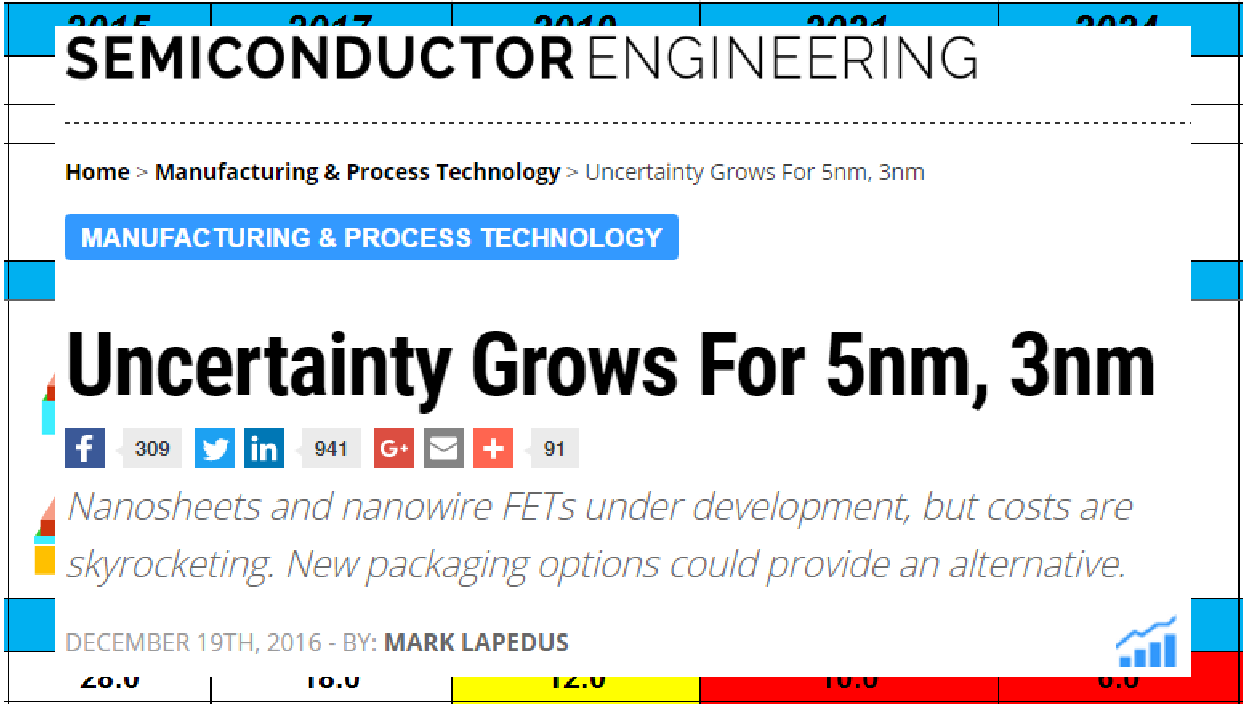
Carballo, Juan-Antonio, Wei-Ting Jonas Chan, Paolo A. Gargini, Andrew B. Kahng, and Siddhartha Nath. "ITRS 2.0: Toward a re-framing of the Semiconductor Technology Roadmap." In Computer Design (ICCD), 2014 32nd IEEE International Conference on, pp. 139-146. IEEE, 2014.

2015	*2017*	*2019*	*2021*	*2024*
P70M56	P48M36	P42M24	P32M20	P24M12G1
"16/14"	"11/10"	"8/7"	"6/5"	"4/3"
finFET FDSOI	finFET FDSOI	finFET LGAA	finFET LGAA VGAA	VGAA, M3D
				
28.0	18.0	12.0	10.0	6.0

Getting E_{FACTOR} down.

- millivolt, transistor size and materials may reduce the E_{FACTOR} , or going 3D, like gate-all-around.

- **Waldrop** quotes "*My bet is that we run out of money before we run out of physics*" [Rock's Law]



Waldrop, M. Mitchell. "The chips are down for Moore's law." *Nature News* 530.7589 (2016): 144.

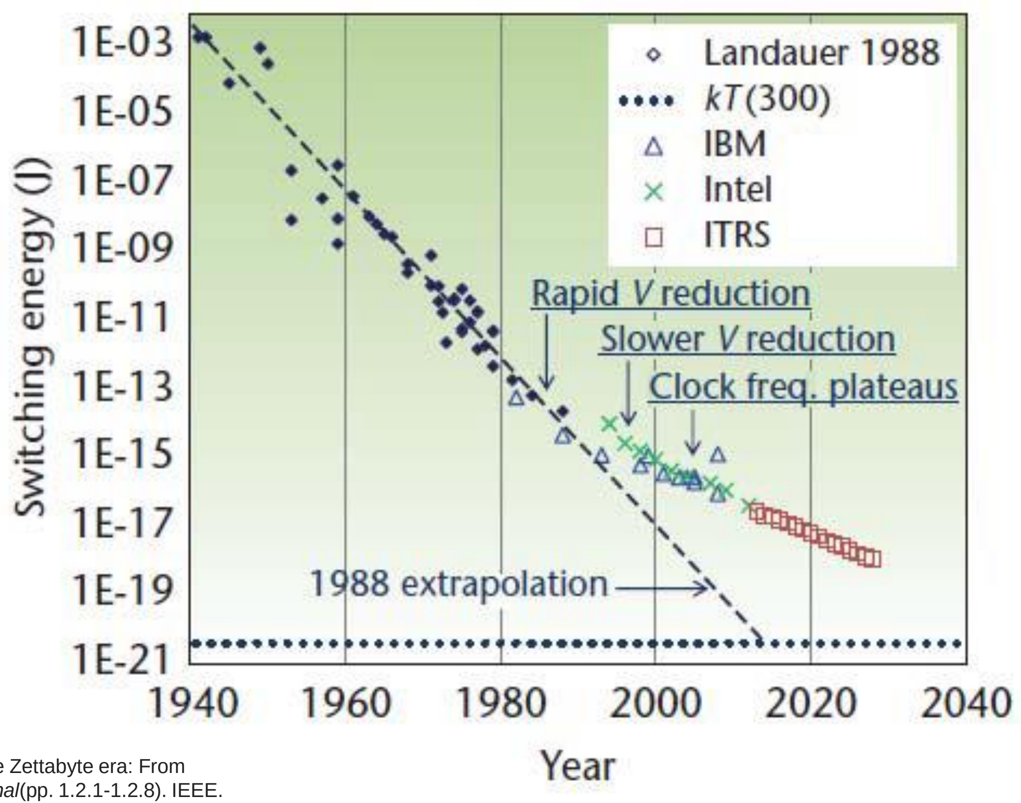
Carballo, Juan-Antonio, Wei-Ting Jonas Chan, Paolo A. Gargini, Andrew B. Kahng, and Siddhartha Nath. "ITRS 2.0: Toward a re-framing of the Semiconductor Technology Roadmap." In Computer Design (ICCD), 2014 32nd IEEE International Conference on, pp. 139-146. IEEE, 2014.

Processing information costs energy.

1E-17 J is 10,000 zJ and the E_{FACTOR} is 3,500 today.

By 2030 the curve shows 1,000 zJ with an E_{FACTOR} of 350.

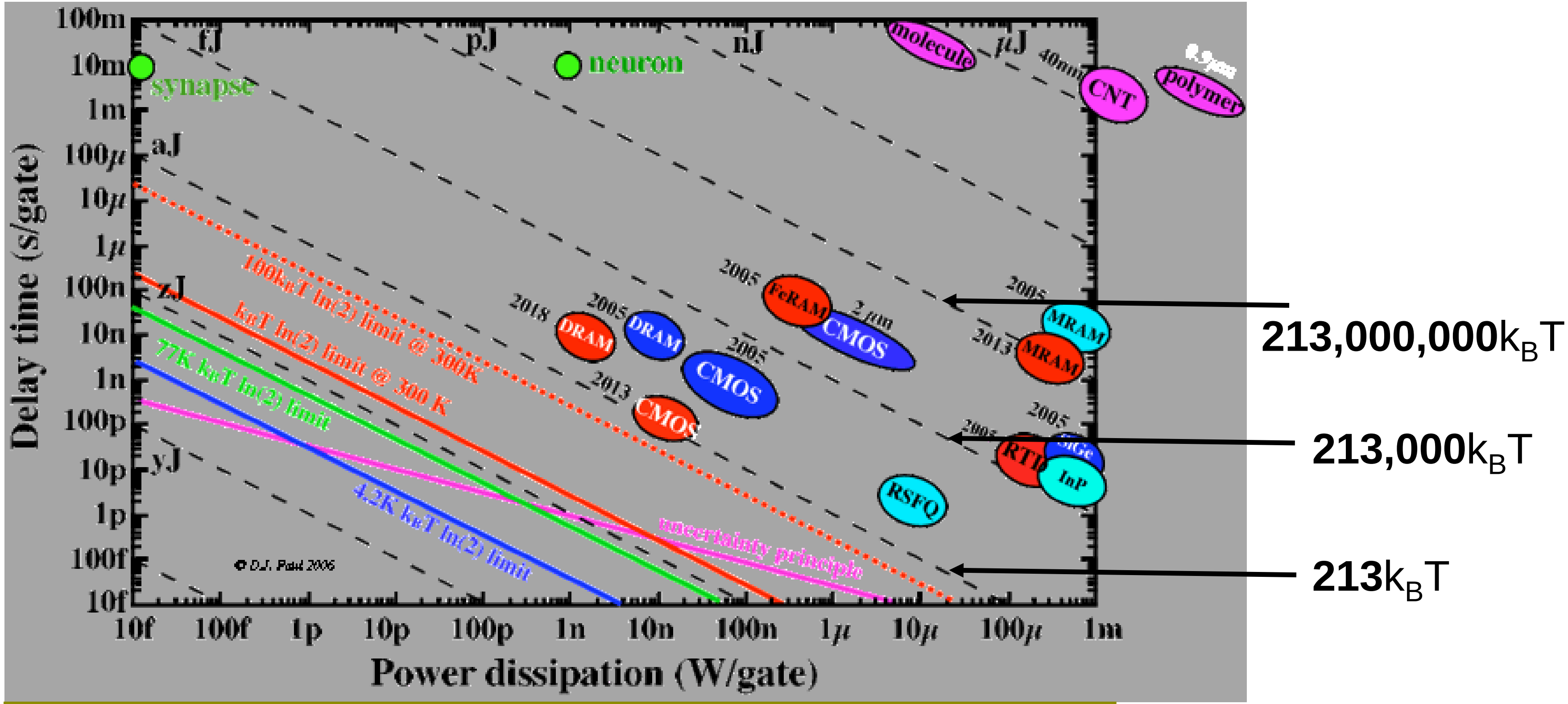
Landauer puts the physical limit of E_{FACTOR} at 0.69.



Ionescu, A.M., 2017, December. Energy efficient computing and sensing in the Zettabyte era: From silicon to the cloud. In *Electron Devices Meeting (IEDM), 2017 IEEE International*(pp. 1.2.1-1.2.8). IEEE.

The switching rate is also important

From Professor Douglas Paul at the University of Glasgow
http://userweb.eng.gla.ac.uk/douglas.paul/SiGe/limits.html



Power (W)= Switch Energy (J) x Switching Rate (s^{-1})

What are the practical limits of E_{FACTOR}?

$$E_{tr} = E_{FACTOR} \times k_B \times T$$

- Frank argues that to measure a signal in the correct state with an error of p_e ($<10^{-40}$) requires the signal energy to be greater than $ln(1/p_e)k_BT$, that is around ***$100k_BT$***.
- Bennet gave an interesting example of DNA polymerization that occurs in cell division to use ~***$40k_BT$*** of energy per step.
- If we cannot get **E_{FACTOR}** down, then we reduce temperature, **T**!

Frank, Michael P. "Approaching the physical limits of computing." *Multiple-Valued Logic, 2005. Proceedings. 35th International Symposium on*. IEEE, 2005.

Bennett, Charles H. "The thermodynamics of computation—a review. "*International Journal of Theoretical Physics* 21.12 (1982): 905-940.

Superconducting Computing!

- IBM ran a project from 1973-1983 on this – terminated due to the success of Si.

- At 4K and an **E_{FACTOR}** of 1,500, a cryotron (Buck's superconducting switch) would use **83 zJ** and switching frequency of less than **125 THz limited by Planck Constant.**

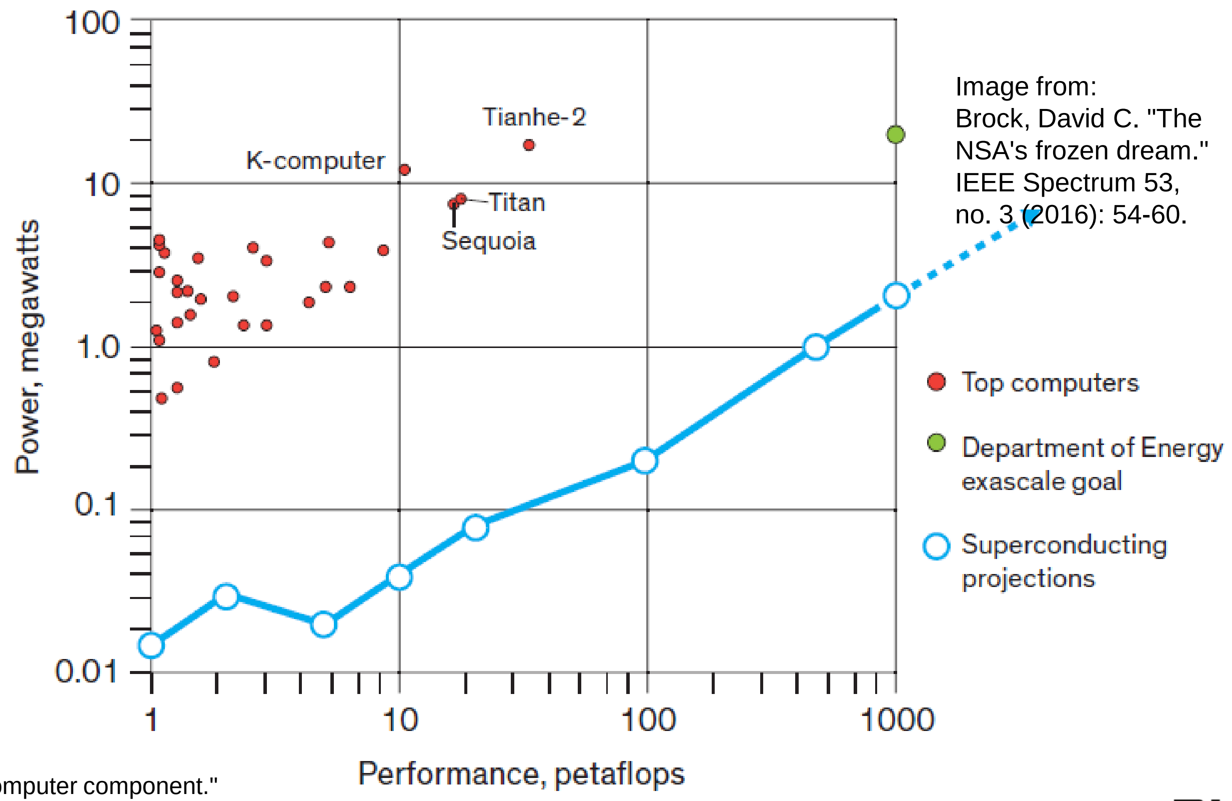


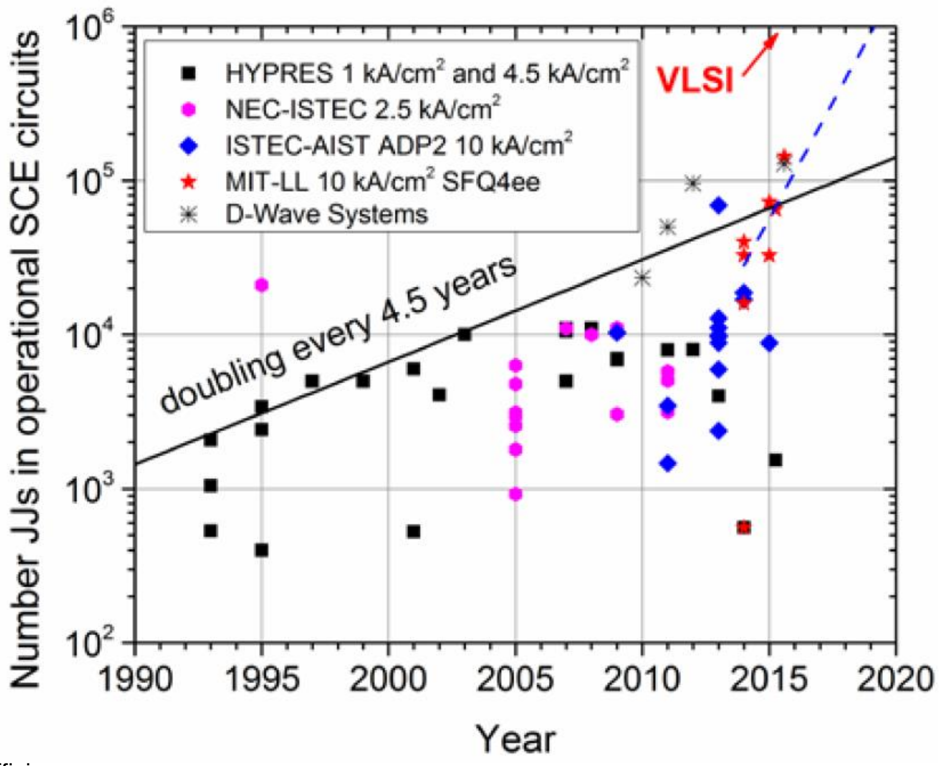
Image from:
Brock, David C. "The NSA's frozen dream." IEEE Spectrum 53, no. 3 (2016): 54-60.

Buck, Dudley A. "The cryotron-a superconductive computer component." *Proceedings of the IRE* 44, no. 4 (1956): 482-493.

So what about the cryogenics cost?

- A CPU with 5 billion cryotron gates and operating at 3GHz would consume 1.25W of power compared to 160W plus the cryogenics overhead (commercially available use 20kW).
- Candidate for 3D.

BUT



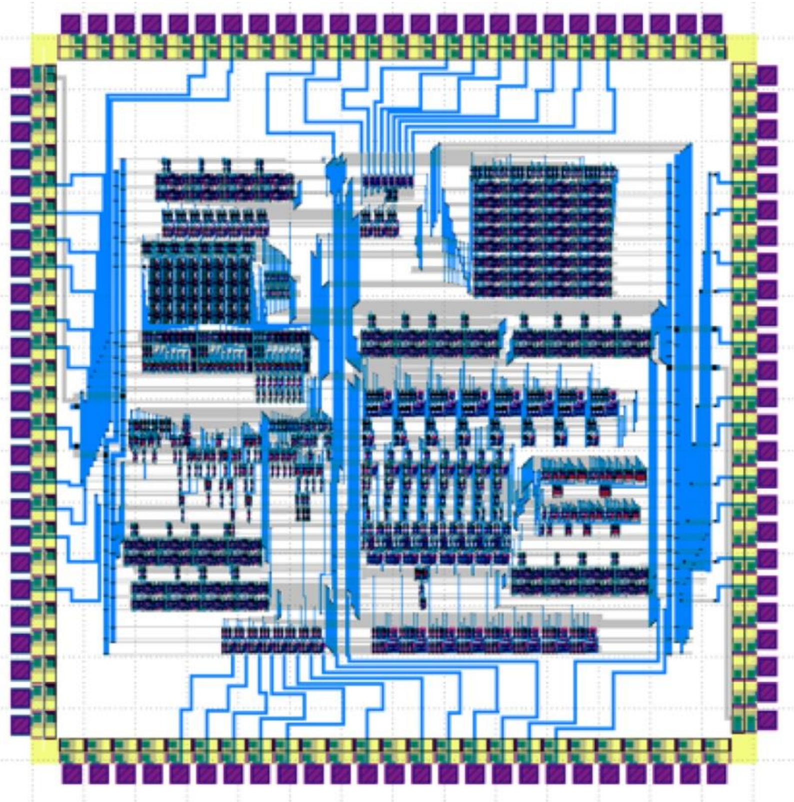
Tolpygo, Sergey K. "Superconductor digital electronics: Scalability and energy efficiency issues (Review Article)." *Low Temperature Physics* 42, no. 5 (2016): 361-379.

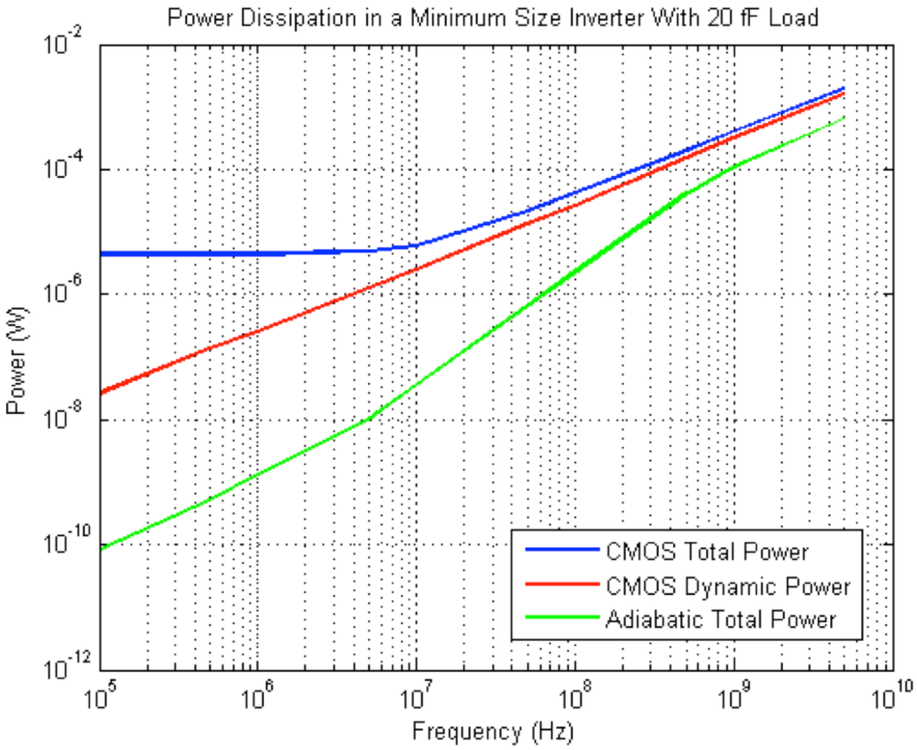
Reversible processes/logic.

- Computing is logically irreversible, it is not possible to travel backwards through the logic gates and get back to the initial state.
- Bennett, Fredkin, Toffoli, Feynman, Frank and others preset theories behind reversible computing. It is a subject area in its own right.
- Feynman talked about computing through reversible logic gates and then ***de-computing*** (i.e. reversing the computation), thus not having to dissipate heat on the basis that this would also be physically reversible.

Bennett, C. H. Logical reversibility of computation. IBM J. Res. Dev., 17, 6 (1973), 525-532.
Fredkin, E. F., Toffoli, T. Conservative logic. Int. J. Theo. Phys., 21, 3/4 (1982), 219-253.
Feynman, R. P. Quantum mechanical computers. Found. Phys., 16, 6 (1986), 507-531.
Frank, M. P. (2005, May). Introduction to reversible computing: motivation, progress, and challenges. In *Proceedings of the 2nd Conference on Computing Frontiers* (pp. 385-390). ACM.

Reversible logic





Campos-Aguillón, César O., Rene Celis-Cordova, Ismo K. Hänninen, Craig S. Lent, Alexei O. Orlov, and Gregory L. Snider. "A Mini-MIPS microprocessor for adiabatic computing." In Rebooting Computing (ICRC), IEEE International Conference on, pp. 1-7. IEEE, 2016.

Reversible computing

- **Snider** and workers presented a paper in 2012, which demonstrated experimentally that there is no Landauer limit in computation – they observed a dissipation of $0.04kT < 0.69kT$. This was achieved using *adiabatically clocked reversible circuits*.

Snider, G.L., Blair, E.P., Thorpe, C.C., Appleton, B.T., Boechler, G.P., Orlov, A.O. and Lent, C.S., 2012, August. There is no Landauer Limit: Experimental tests of the Landauer principle. In *Nanotechnology (IEEE-NANO), 2012 12th IEEE Conference on* (pp. 1-6). IEEE.

Neuromorphic computing

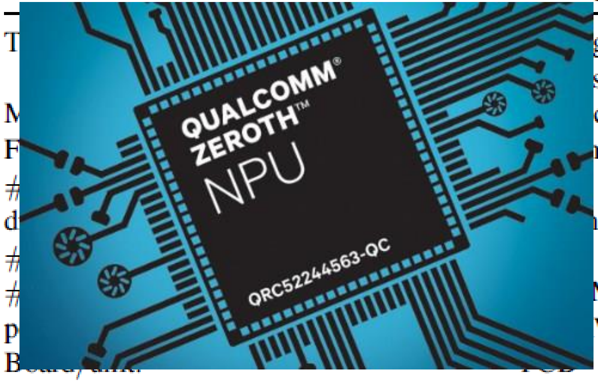
Platform:	Human brain	Neurogrid	BrainScaleS	TrueNorth	SpiNNaker
Technology:	Biology	Analogue, sub-threshold	Analogue, over threshold	Digital, fixed	Digital, programmable
Microchip:		Neurocore	HiCANN		18 ARM cores
Feature size:	10 μm[a]	180 nm	180 nm	28 nm	130 nm
# transistors:		23 M	15 M	5.4 B	100 M
die size:		1.7 cm^2	0.5 cm^2	4.3 cm^2	1 cm^2
# neurons:		65 k	512	1 M	16 k
# synapses:		~100 M	100 k	256 M	16 M
power:		150 mW	1.3 W	72 mW	1 W
Board/unit:		PCB	20 cm wafer	PCB	PCB
# chips:		16	352	16	48
# neurons:		1 M	200 k	16 M	768 k
# synapses:		4 B	40 M	4B	768 M
power:		3 W	500 W	1 W	80 W
Reference system:	1.4 kg		20 wafers in 7 × 19″ racks		600 PCBs in 6 × 19″ racks
# neurons:	100 B		4 M		460 M
# synapses:	10^{15}		1 B		460 B
power:	20 W		10 kW		50 kW
Energy/connection:	10 fJ	100 pJ	100 pJ	25 pJ	10 nJ
Speed versus biology:	1×	1×	10 000×	1×	1×

2,337,540 k_B T

Furber, Steve. "Large-scale neuromorphic computing systems." *Journal of Neural Engineering* 13, no. 5 (2016): 051001.

Neuromorphic computing

Platform:	Human brain	Neurogrid	BrainScaleS	TrueNorth	SpiNNaker
T		gue, sub-threshold	Analogue, over threshold	Digital, fixed	Digital, programmable
M		core	HiCANN		
F		n	180 nm		
#			15 M		
d		^{2}	0.5 cm^2		
#			512		
#		M	100 k		
p		W	1.3 W		
B			20 cm wafer		
			52		
			00 k		
			0 M		
			00 W		
			0 wafers in 7 × racks		
			M		460 M
			B		460 B
			0 kW		50 kW
			00 pJ	25 pJ	10 nJ
			0 000×	1×	1×



biology:

Furber, Steve. "Large-scale neuromorphic computing systems." *Journal of Neural Engineering* 13, no. 5 (2016): 051001.

RI. SE

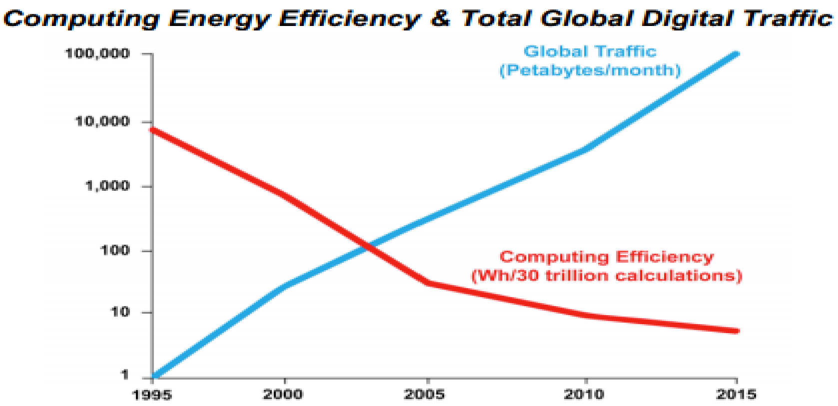
What technology will do the computing of the future?

- CMOS with millivolt, reduce feature size, going 3D and using new materials (error rate, gate-all-around, leakage/quantum effects, heat issue, EUV, new materials still in the lab) [$E_{tr} = 100k_BT$]
- Superconducting (switch count per unit volume too low) [$E_{tr} > h/(t_{delay})$]
- Quantum (problem specific, still the challenge of error correction) [$E_{tr} > h/(t_{delay})$]
- Reversible (complex logic) [$E_{tr} = 0.04k_BT$]
- Dark silicon/multicore (software development needed) [$N_{tr} < N_{tr}$]
- Approximate computing (specialised application) [low bit operations]
- Neuromorphic (potential energy efficiency issues, application specific and massively parallel)

RI.SE

Store, transmit and compute digital information bottom up approach for 2030

- Battle between digital growth and energy efficiency of compute, storage and transmission of digital information.
- **Xu** based on **Hilbert** and Lopez:



Data Source: IEEE Transactions, Into the Exacloud, Entropy Economics

Year	Storage	Communi-cation	General-Purpose Computing	Special-Purpose Computing
1986	21 PB	59 PB	0.3 PIPS	0.44 PIPS
2007	277 EB	537 EB	6.39 EIPS	189 EIPS
2030	140 ZB	272 ZB	18 ZIPS	2 570 ZIPS

Xu ZW. Cloud-sea computing systems: Towards thousand-fold improvement in performance per watt for the coming Zettabyte era. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(2): 177–181 Mar. 2014.

Hilbert, M. and López, P., 2011. The world's technological capacity to store, communicate, and compute information. science, 332(6025), pp.60-65.

SUMMARY

- We do need better predictions of the future of energy consumption by the microelectronics infrastructure, but this depends on what is next for the core technology.
- Clear that the way that we do computation has reached an exciting point. One such individual from the Rebooting Computing project did say “the semiconductor industry has been boring for the last 40 years, it is not getting interesting!”
- Whilst a focus on energy consumption of the devices and infrastructure is important, perhaps even more important for the future is the consumption of resource and material, such as rare earth metals.
- As well as the core data centres and supercomputing centres, more energy is consumed by the digital networks and with 5G poised to grow, future network energy consumption needs to be checked.
- *Reduce digital profligacy and increase digital sobriety.*

Visit https://bodentypeDC.eu

PODS available for test partners.

BODEN TYPE DC ONE

GA 768875

RI. SE