SEMICONDUCTOR INDUSTRY ASSOCIATION

SIA/SRC Webinar

Decadal Plan for Semiconductors: New Compute Trajectories for Energy Efficiency

Wednesday, April 14 at 11:00 am EDT

Victor V. Zhirnov, Ph.D.
Chief Scientist
Semiconductor Research Corporation

Maryam Cope
Director, Government Affairs
SIA

Rob Clark
Senior Member of the Technical Staff
TEL

Carlos Diaz
Senior Director, Research and Development
TSMC

Stephen Kosonocky
Senior Fellow
AMD

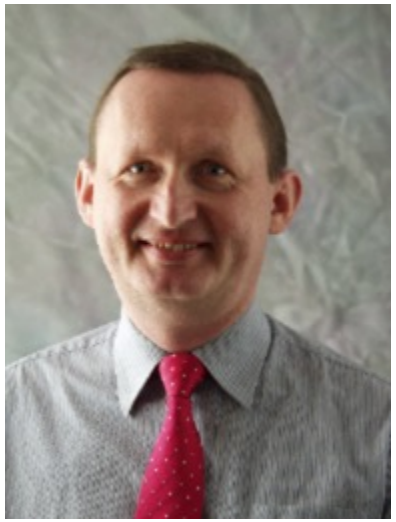
Heike Riel
IBM Fellow, Head Science & Technology
IBM Research

Gilroy Vandentop
Director of Corporate University Research
Intel

Jim Ang
Chief Scientist for Computing in the Physical and Computational Sciences Directorate
PNNL

Semiconductor Research Corporation

New compute trajectories for energy efficiency



Victor Zhirnov, Chief Scientist,
Semiconductor Research Corporation

April 14, 2021

Outline

- SRC and Decadal Plan for Semiconductors
  - Five seismic shifts in information and communication technologies
- Compute Needs after 2030
  - Energy challenge
  - New compute trajectories
- Summary

The Case for a Decadal Plan for Semiconductors (2030 ICT research goals)

The current hardware-software (HW-SW) paradigm in information and communication technologies (ICT) has reached its limits and must change. It is important to identify significant trends that are driving information technology and what roadblocks/challenges the industry is facing. A Decadal Plan for Semiconductors is needed that will transform the semiconductor industry by:

- supporting the strategic visions of semiconductor companies
- placing 'a stake in the ground' to motivate and challenge the best and brightest university faculty and students to be a key part of the solution
- guiding a (r)evolution of research programs
  - **3x increase of federal research spending relevant to the semiconductor industry**

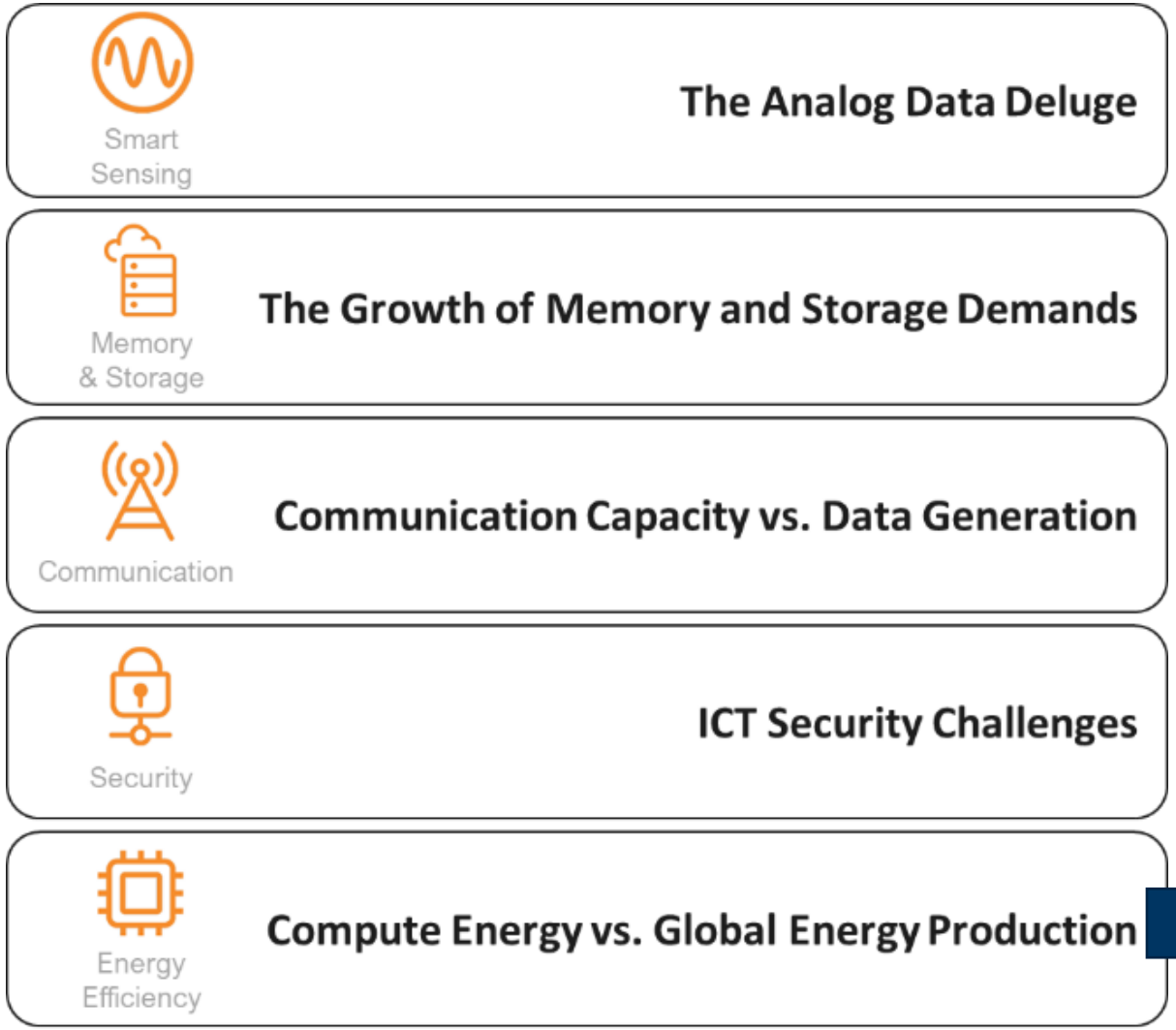
Because the future can't wait, we bring the best minds together to achieve the unimaginable…

Our 2030 Decadal Plan for Semiconductors

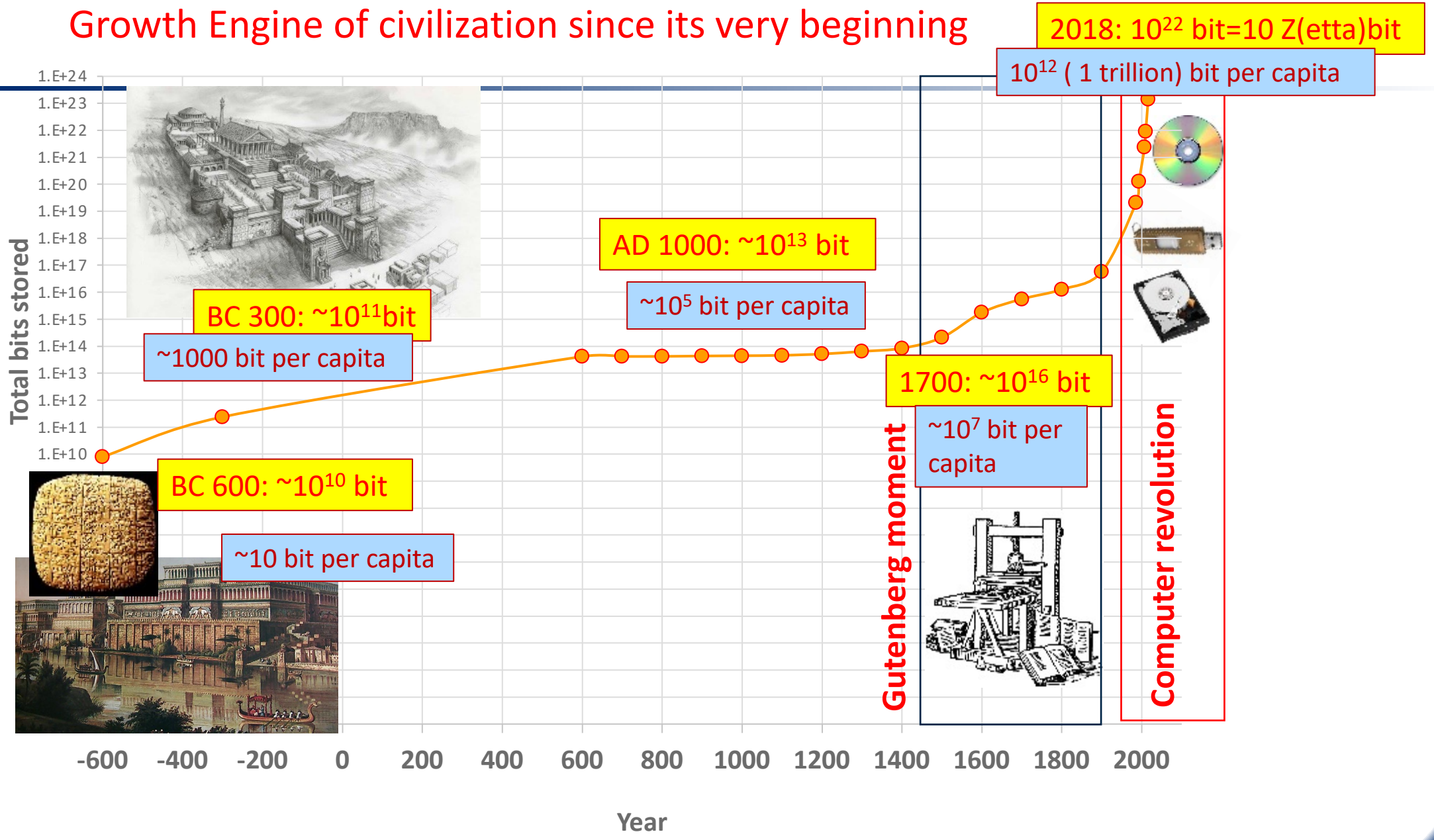
https://www.src.org/about/decadal-plan/ (released on January 25, 2021)

SIA and SRC call for +34B in semiconductor R&D throughout the 20s

2021 NDAA Passes on 1-1-2021 Indicating Increased Appetite for hardware R&D



Information along with Energy has been the Social-Economic Growth Engine of civilization since its very beginning



Economic and Social Well-being

Computing Instructions

Computer Applications

Higher Standard of Living

4×10^{12} USD

(2018)

Binary Transitions

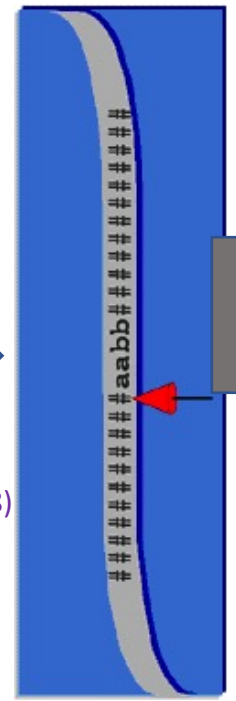
BITS

10^{35} bit

(2018)

Joules

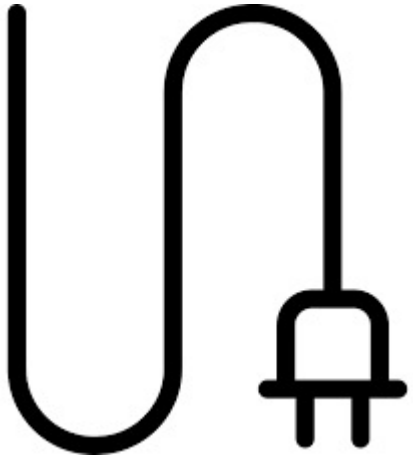
BTU



MIPS

ZIPS

FLOPS

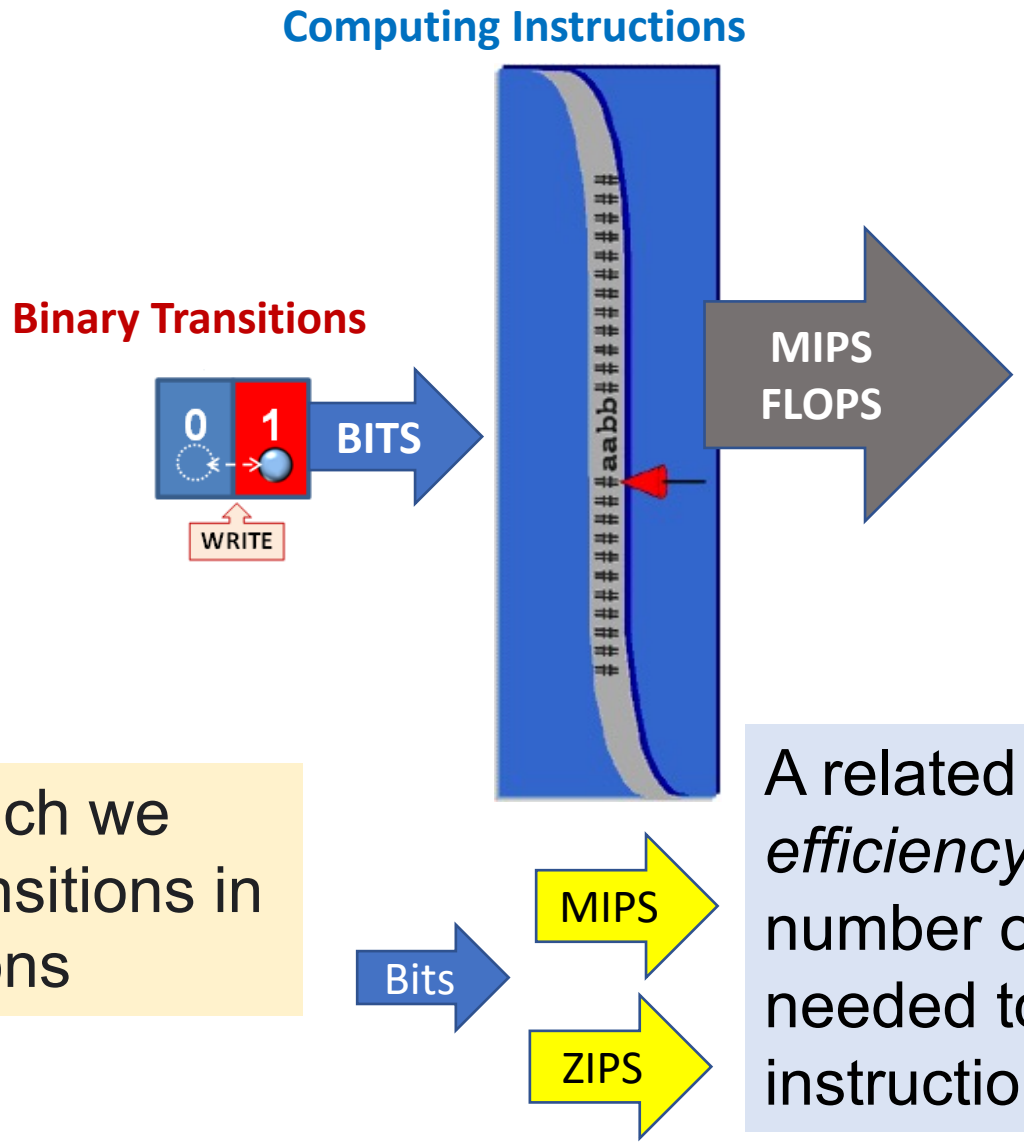


~10^{18} J ~3× 10^{11} kW-h ~10^{15} BTU=1 Quad

(2018)

Energy

What is Compute Trajectory?



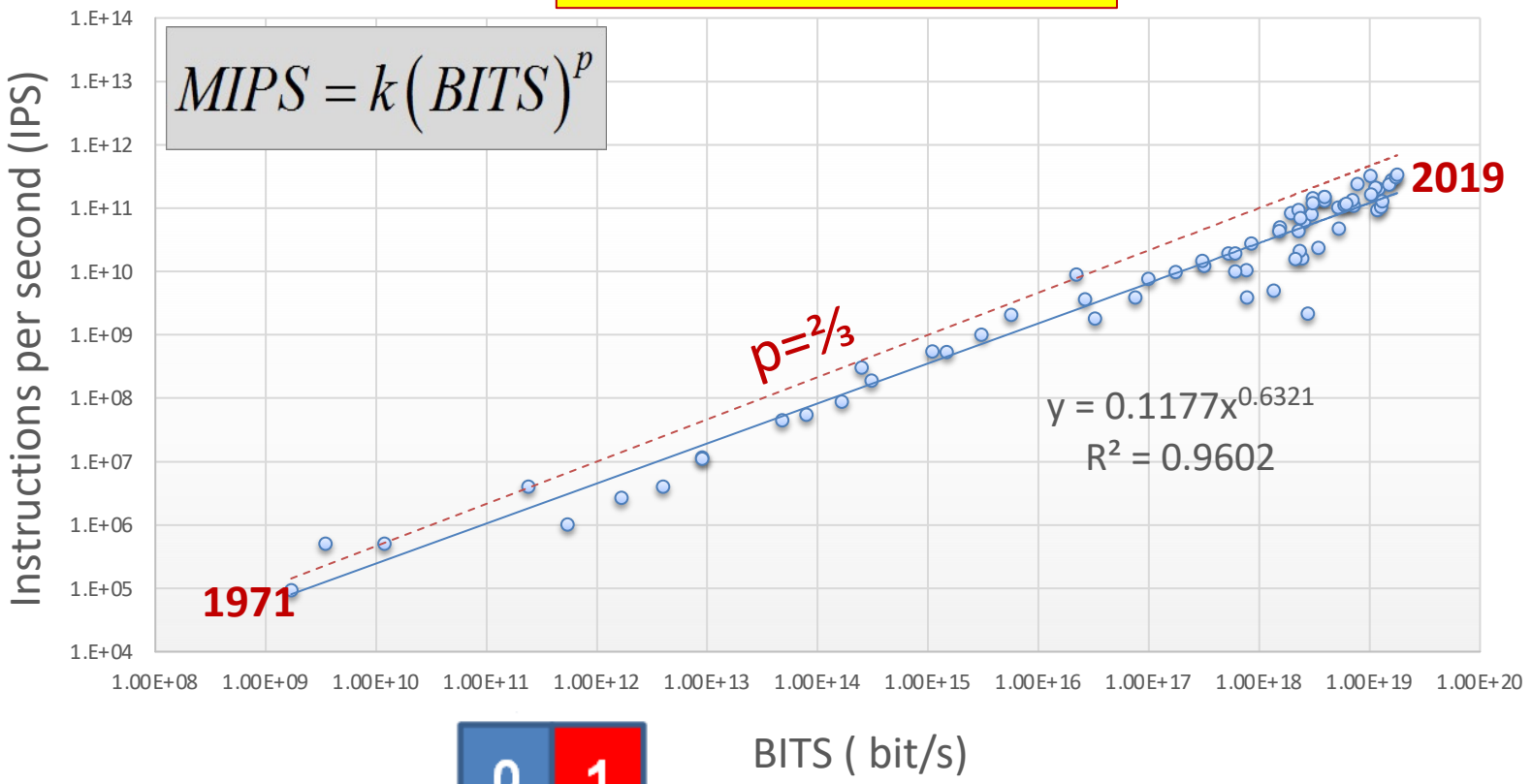
It is the way in which we convert binary transitions in compute instructions

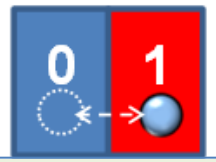
A related question is *bit-utilization efficiency* in computation, i.e., the number of single bit transitions needed to implement a compute instruction.

CPU operations vs. binary transitions

$$\mu = f(\beta) = k\beta^p$$

k=0.1, p= 0.64 ≈ ⅔





$$\beta = \alpha N_{tr} \cdot f \quad \Rightarrow \quad P = \beta E_{bit}$$

Company	Model	Year
Intel	4004	1971
Intel	8080	1974
MOS Technology	6502	1975
Motorola 68000	68000	1979
Intel	286	1982
Motorola	68020	1984
Intel	386DX	1985
ARM	ARM2	1986
Motorola	68030	1987
Motorola	68040	1990
DEC	Alpha 21064EV4	1992
Intel	486DX	1992
Motorola	68060	1994
Intel	Pentium	1994
Intel	Pentium Pro	1996
IBM - Motorola	PowerPC 750	1997
Intel	Pentium III	1999
AMD	Athlon	2000
AMD	Athlon XP 2500+	2003
Intel	Pentium 4 Ext. Edition	2003
Centaur - VIA	VIA C7	2005
AMD	Athlon FX-57	2005
AMD	Athlon 64 3800+ X2	2005
IBM	Xbox360 "Xenon"	2005
Sony-Toshiba-IBM	PS3 Cell BE	2006
AMD	Athlon FX-60	2006
Intel	Core 2 Extreme X6800	2006
Intel	Core 2 Extreme QX6700	2006
P.A. Semi	PA6T-1682M	2007
Intel	Core 2 Extreme QX9770	2008
Intel	Core i7 920	2008
Intel	Atom N270	2008
AMD	E-350	2011
AMD	Phenom II X4 940	2009
AMD	Phenom II X6 1100T	2010
Intel	Core i7 980X	2010
Intel	Core i7 2600K	2011
Intel	Core i7 875K	2011
AMD	8150	2011
Intel	Xeon E3-1290v2	2012
Intel	Ivy Bridge-EX-15	2013
Intel	i7-5960X	2014

Limits to Binary Logic Switch Scaling—A Gedanken Model

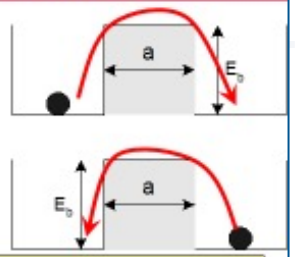
VICTOR V. ZHIRNOV, RALPH K. CAVIN, III, FELLOW, IEEE,
JAMES A. HUTCHBY, SENIOR MEMBER, IEEE, AND GEORGE I. BOURIANOFF, MEMBER, IEEE

Invited Paper

Lowest Barrier: The Boltzmann constraint

Distinguishability ***D*** implies low probability Π of spontaneous transitions between two wells (error probability)

***D=max*, Π=0** **D=0, Π=0.5 (50%)**



Classic distinguishability: $\Pi_{classic} = \exp(-\frac{E_b}{k_B T})$

Minimum distinguishable barrier: Π=0.5

$\frac{1}{2} = \exp(-\frac{E_b}{k_B T}) \Rightarrow E_b = kT\ln 2$

Shannon - von Neumann - Landauer limit

14

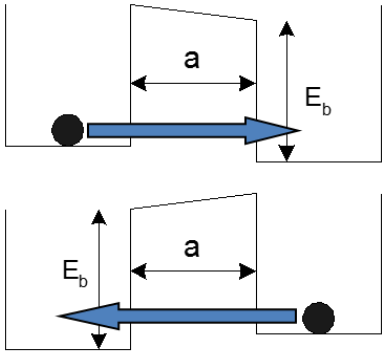
- The limit for E_{bit} is given by the Shannon–von Neumann–Landauer (SNL) expression for smallest energy to process a bit

$0.7k_BT = 0.02eV = 3\times10^{-21}J$

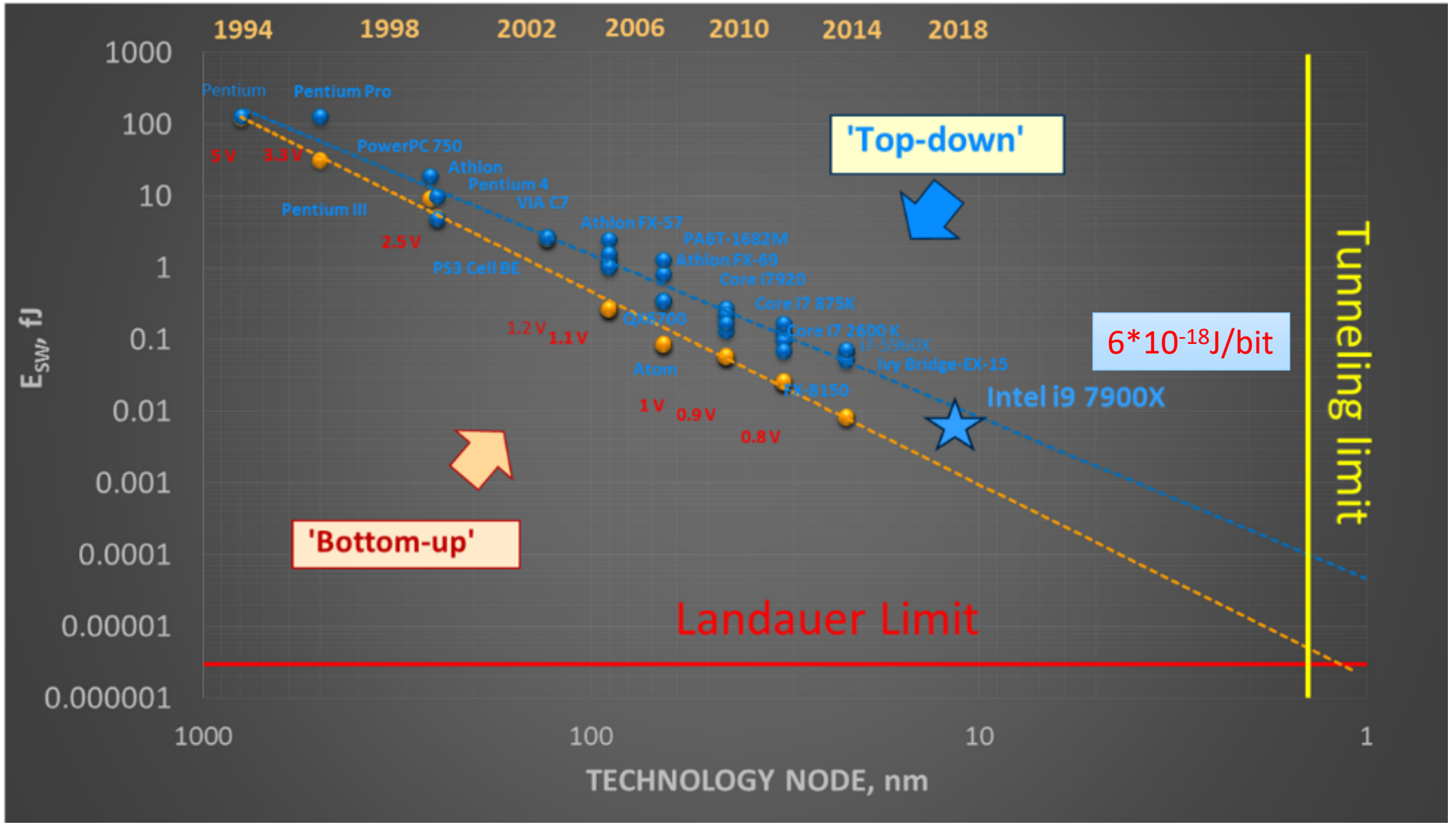
- If we endeavor to construct a model for a computer operating at SNL limit at 300 K, the minimum size and switching time of binary switches can be estimated based on the Heisenberg Uncertainty relations

$$a_{\min} = \frac{\hbar}{\sqrt{2mkT\ln 2}} \sim 1.5nm \qquad t_{sw} = \frac{\hbar}{kT\ln 2} \sim 0.04ps$$

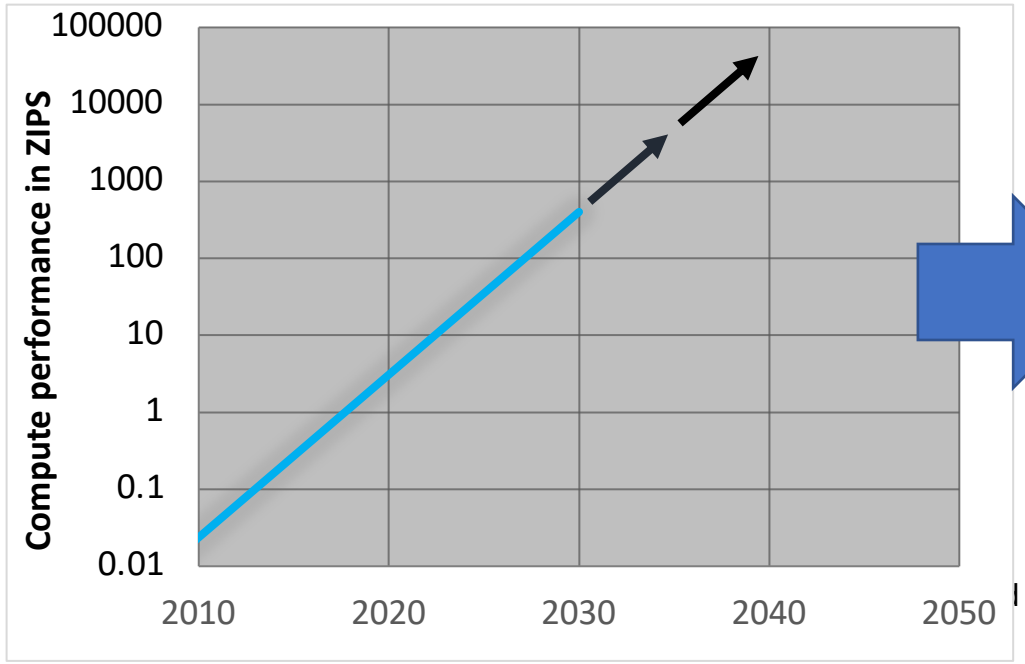
$$\Delta x \Delta p \geq \hbar \qquad \Delta E \Delta t \geq \hbar$$

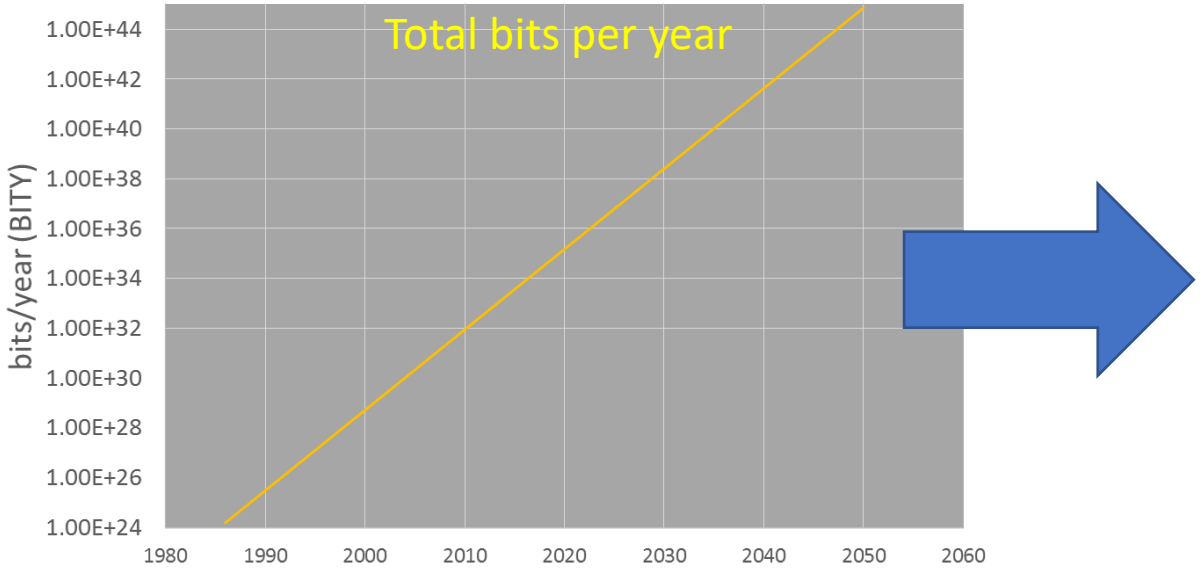


Computing energy: Energy per bit in CPU



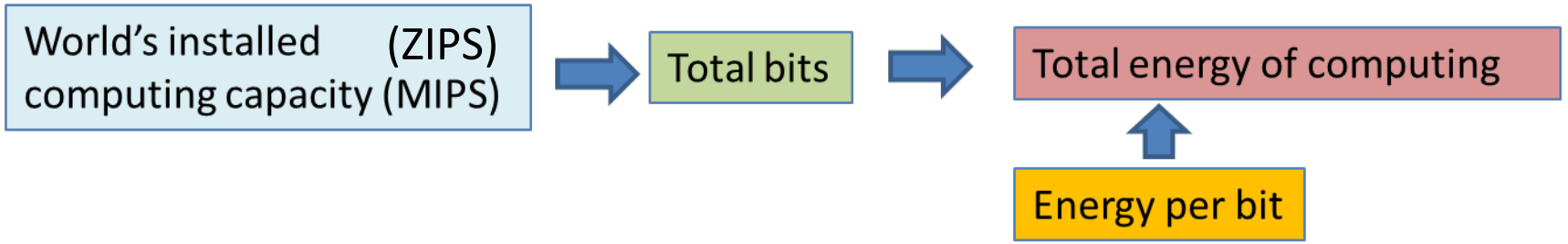
Computations per Year





I P. Lopez, “The World's Technological Capacity to Store, Communicate,

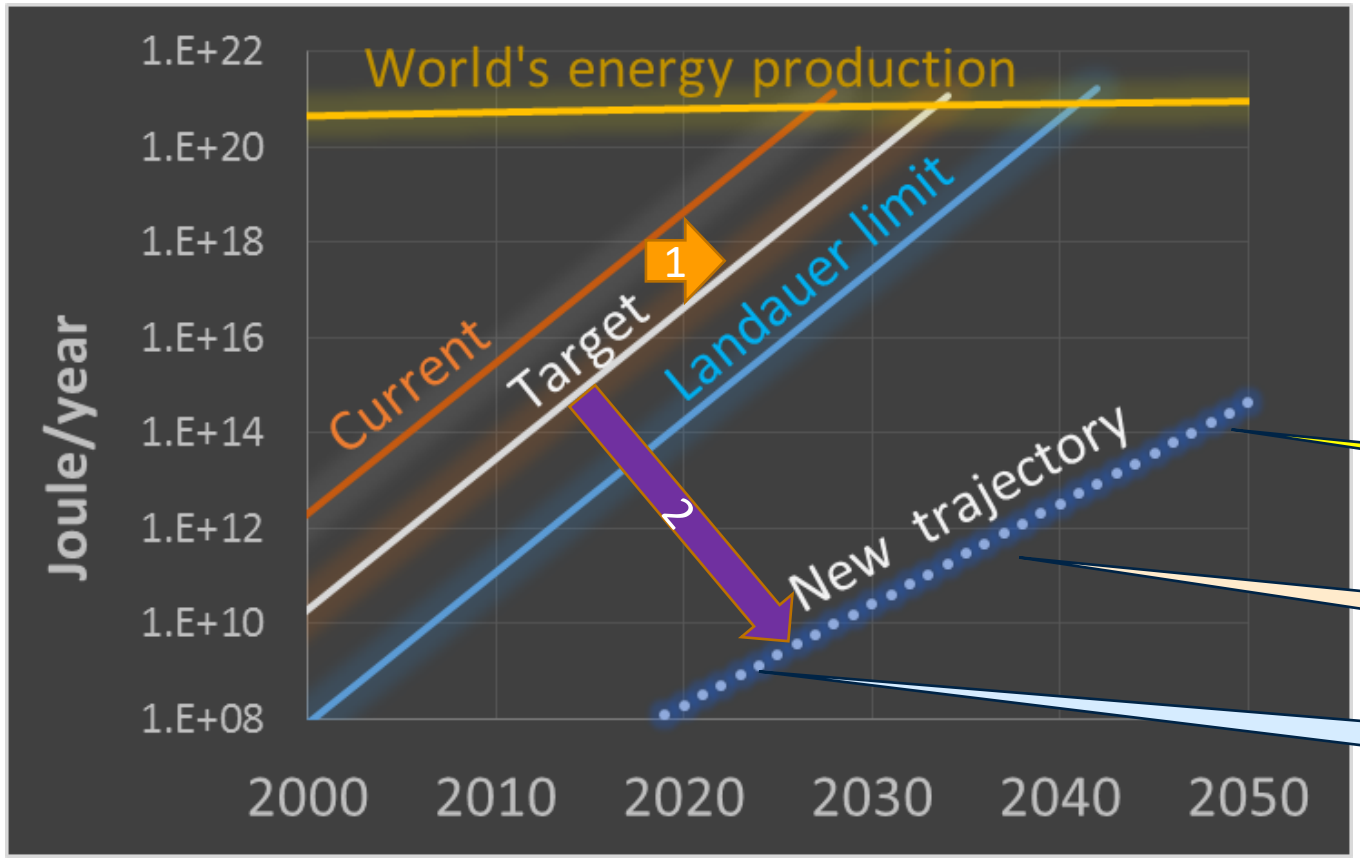
1 ZIPS= 10^{15} MIPS



Total energy of computing: A need to change 'computational trajectory'

(based on research by Hilbert and Lopez: M. Hilbert and P. Lopez, "The World's Technological Capacity to Store, Communicate, and Compute Information", Science 332 (2011) 60-65

$$MIPS = k\left(BITS\right)^{p}$$



Existing trajectory: $p \approx 2/3$

Current: 10^{-16} J/bit

Target: 10^{-18} J/bit

Landauer limit: 10^{-21} J/bit

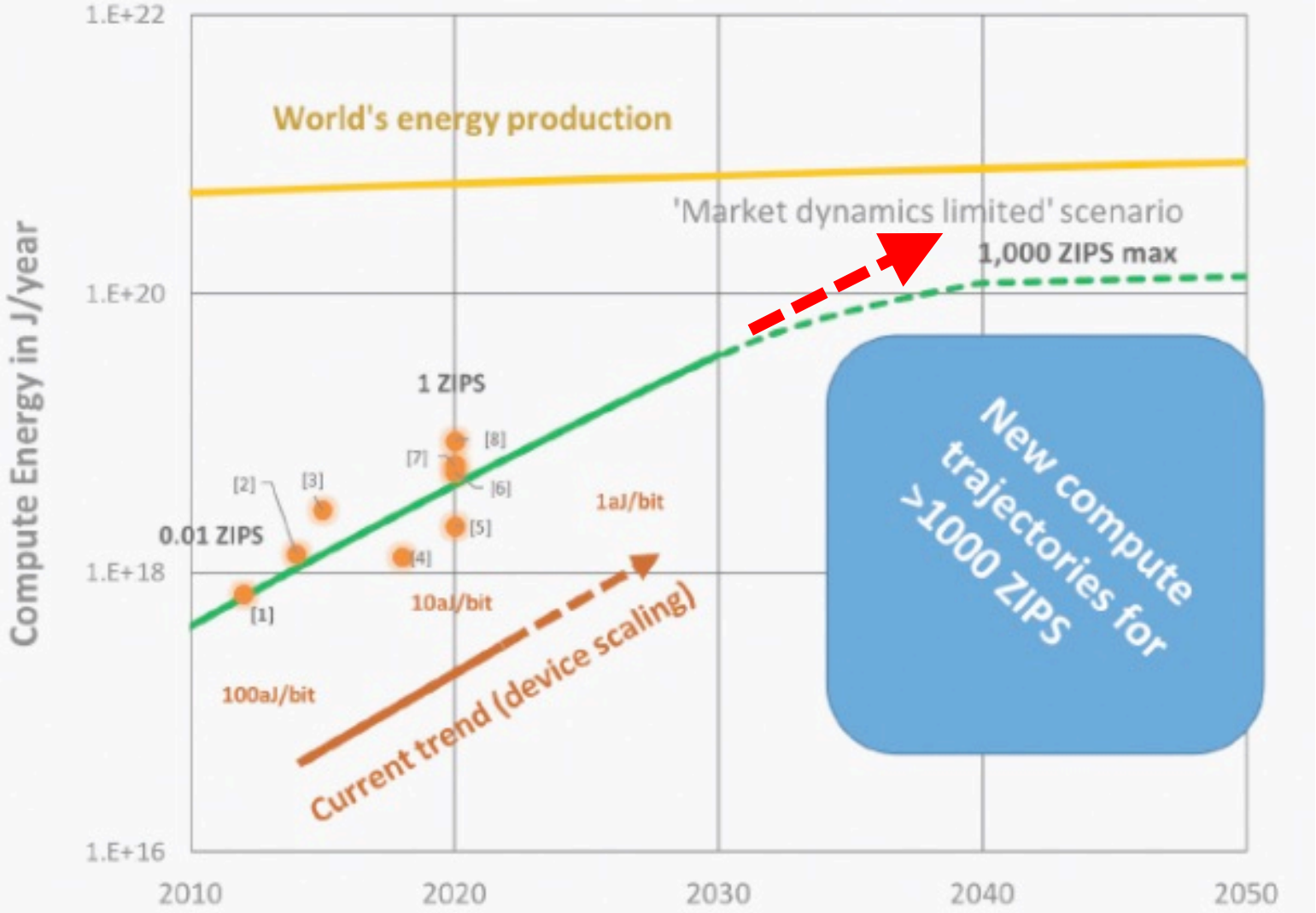
New trajectory: $p \approx 1$

Quantum computing

Neuromorphic

AI engines

Seismic shift #5: Computing growth may not be sustainable



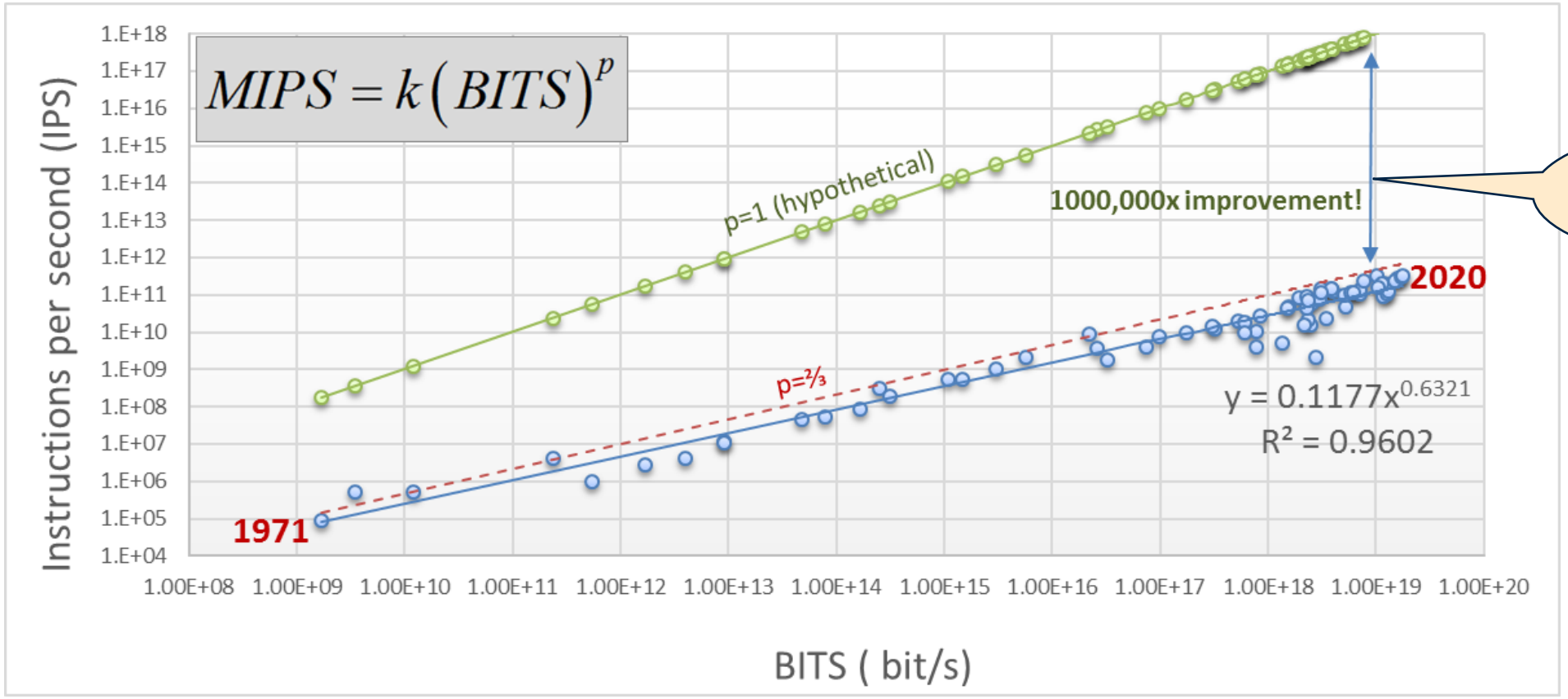
Source: SRC Decadal Plan, 2020

Why Seismic Shift?

Computing growth may be not sustainable by 2040, as its energy requirements would exceed the estimated world's energy production

Need: Discover computing paradigms/architectures with a radically new 'computing trajectory' demonstrating >1,000,000x improvement in energy efficiency. Changing the trajectory not only provides immediate improvements but also provides many decades of buffer and is much more cost effective than attempting to increase the world's energy supply dramatically.

A need to change ‘computational trajectory’

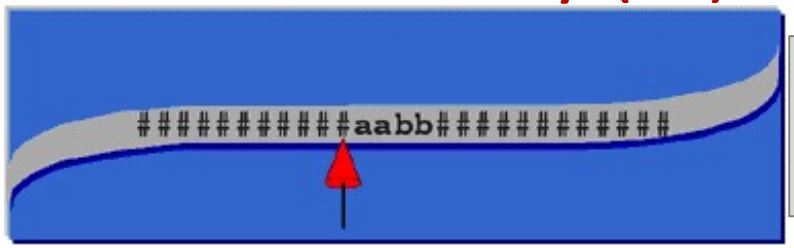


How can we get there?

bit utilization efficiency in computation!

Needed: Theory of Computation

Memory (M)



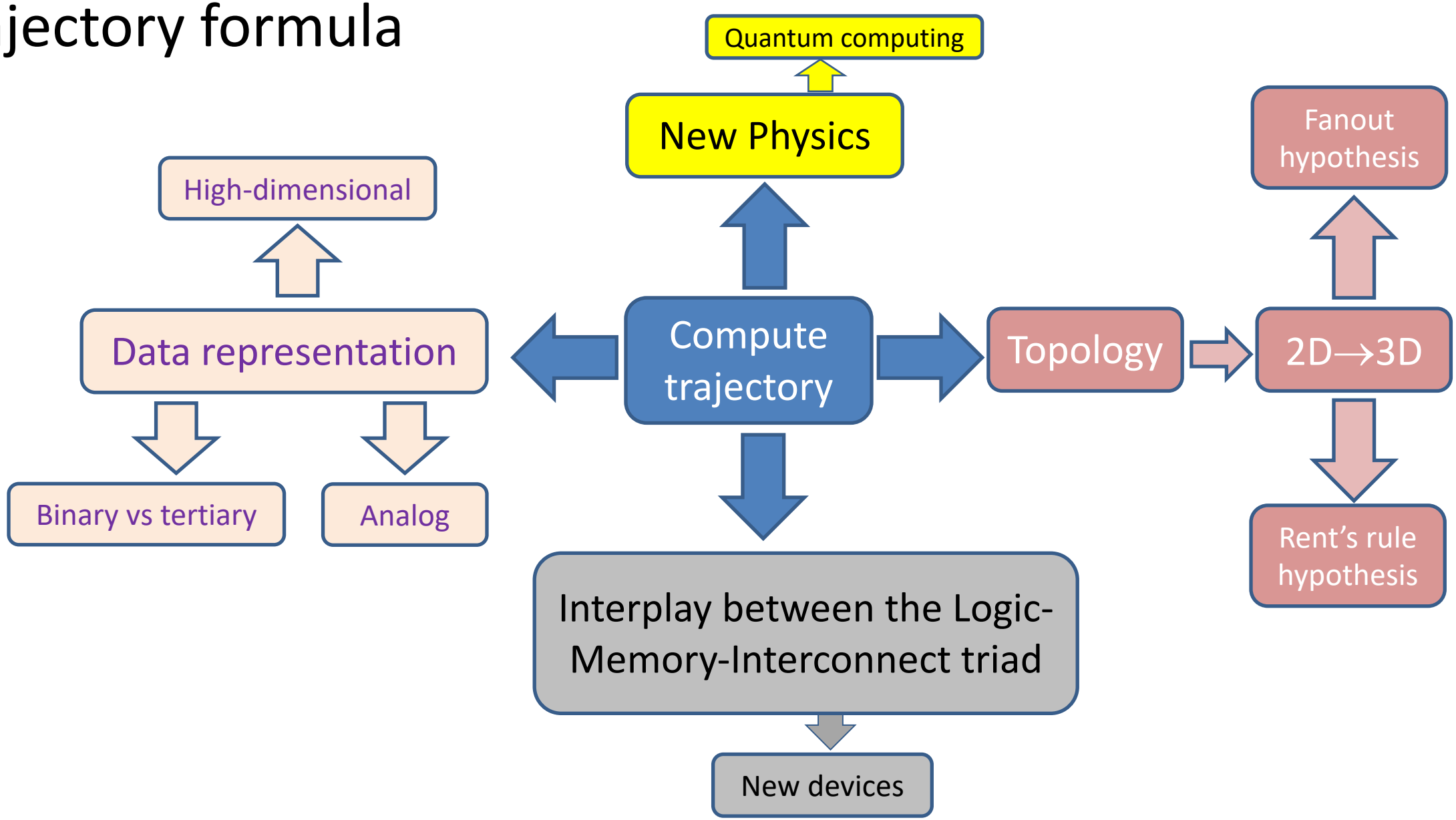
$$MIPS = k\left(BITS\right)^{p}$$

Existing trajectory: $p \approx \frac{2}{3}$

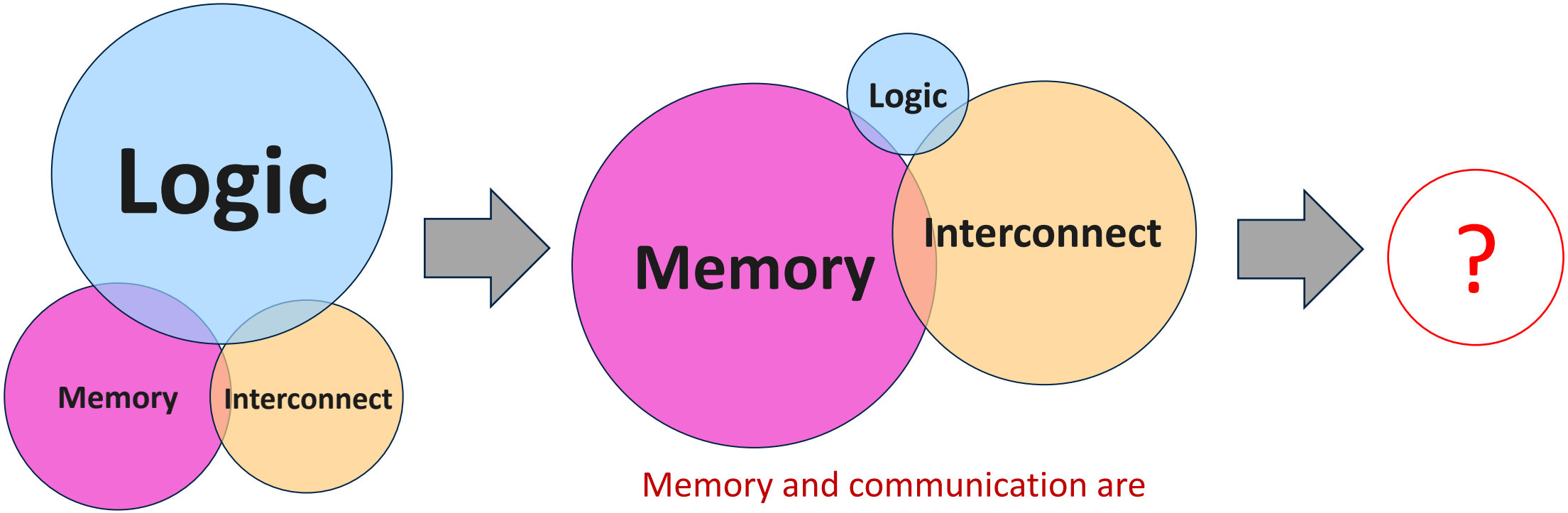
?

The theoretical basis for performance measurements for computers is much less solid than the theoretical basis for information storage and communication (e.g. Shannon limit etc.)

Hypotheses of the origin of the exponent p in the compute trajectory formula



Three Cornerstones of Computing



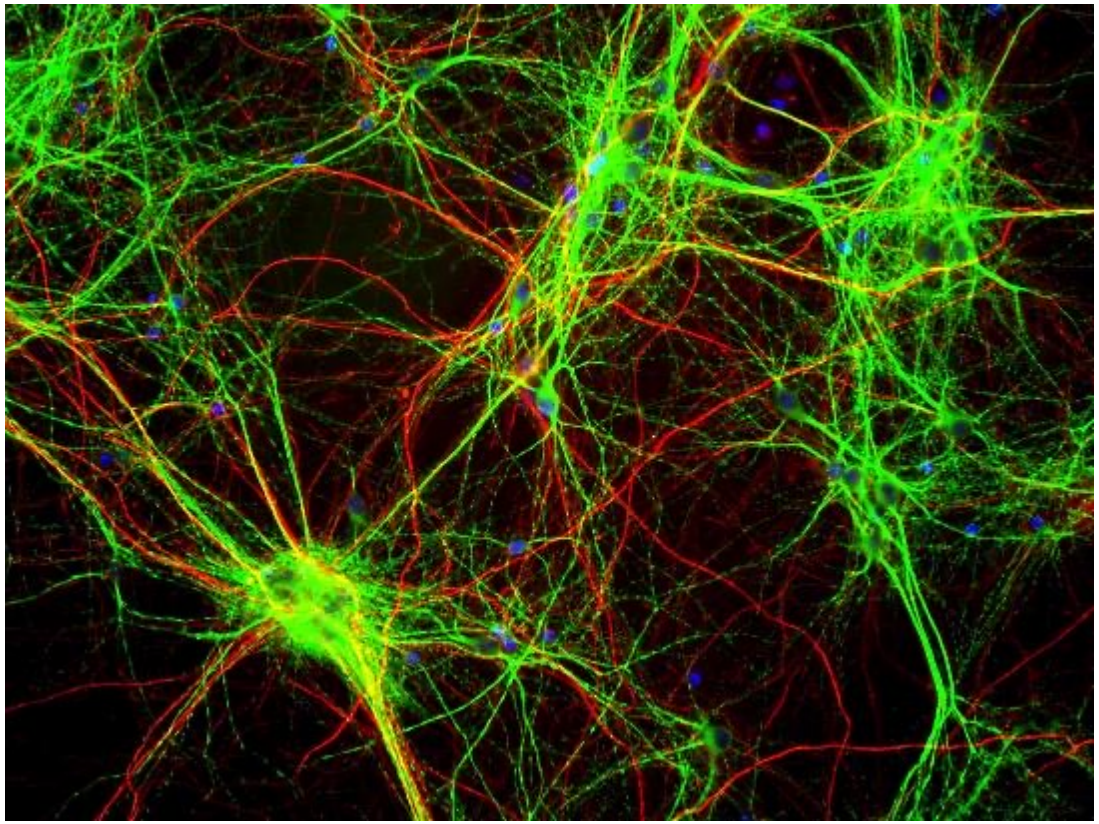
Memory and communication are expensive…

1990 2021 2030

Brain computes BOTH with interconnects and with memory

In the human brain, the distribution of **Ca** ions in dendrites represents a crucial variable for processing and storing information.

Ca ions enter the dendrites through voltage-gated channels in a membrane, and this leads to rapid local modulations of calcium concentration within dendritic tree



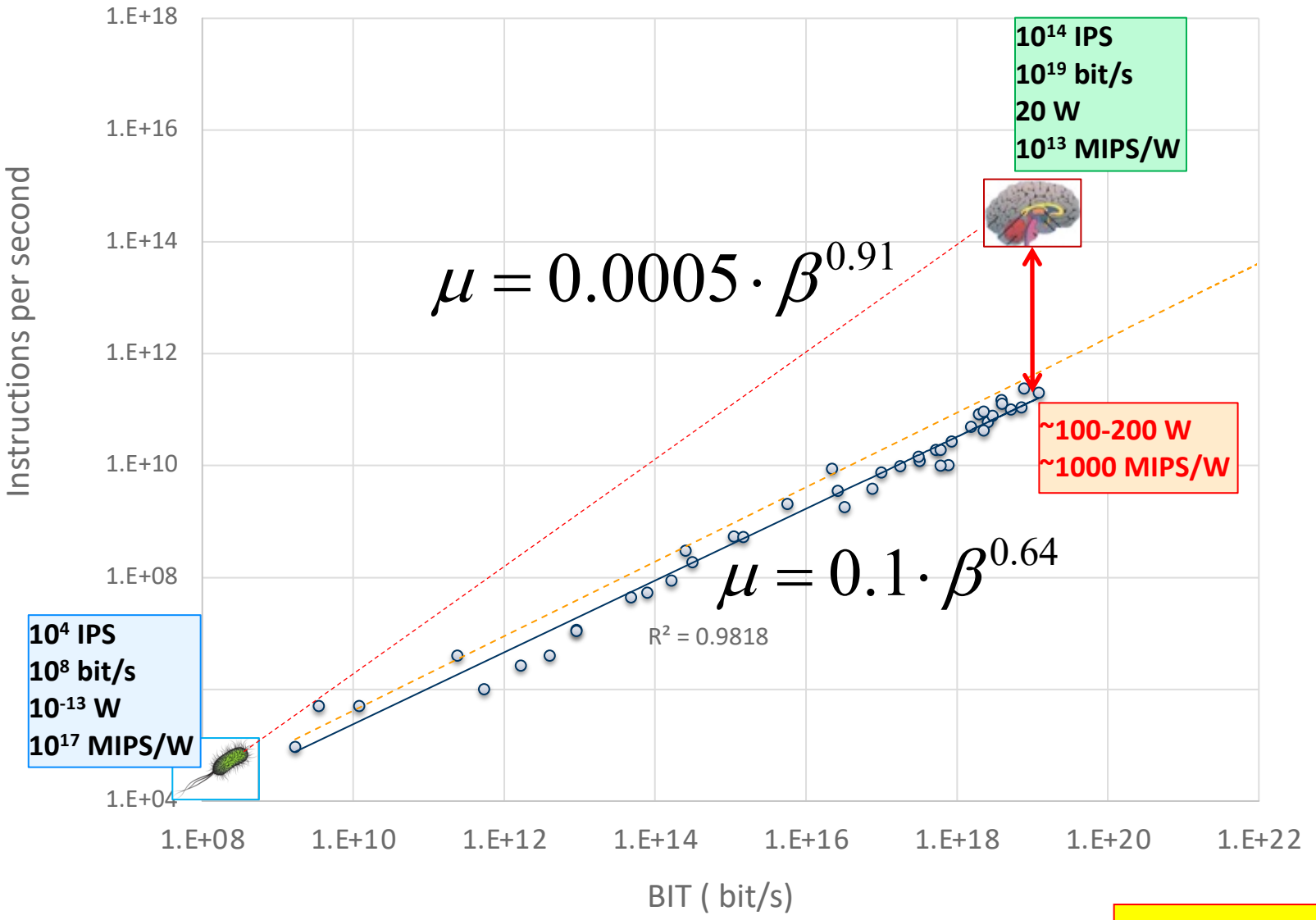
DENDRITES ARE LIKE MINI-COMPUTERS IN YOUR BRAIN

Source: FUTURITY

S. L. Smith et al, “Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo”, Nature 503 (2013) 115

C. Koch, “Computation and single neuron”, Nature 385 (1997) 207

Computations vs. binary transitions



Estimates of computational power of human brain:

Binary information throughput:

$\beta \sim 10^{19}$ bit/s

Gitt W, "Information - the 3rd fundamental quantity", Siemens Review 56 (6): 36-41 1989

(Estimate made from the analysis of the control function of brain: language, deliberate movements, information-controlled functions of the organs, hormone system etc.

Number of instruction per second

$\mu \sim 10^{8}$ MIPS

H. Moravec, "When will computer hardware match the human brain?" J. Evolution and Technol. 1998. Vol. 1

(Estimate made from the analysis brain image processing)

Alternative trajectory may exist!

Summary

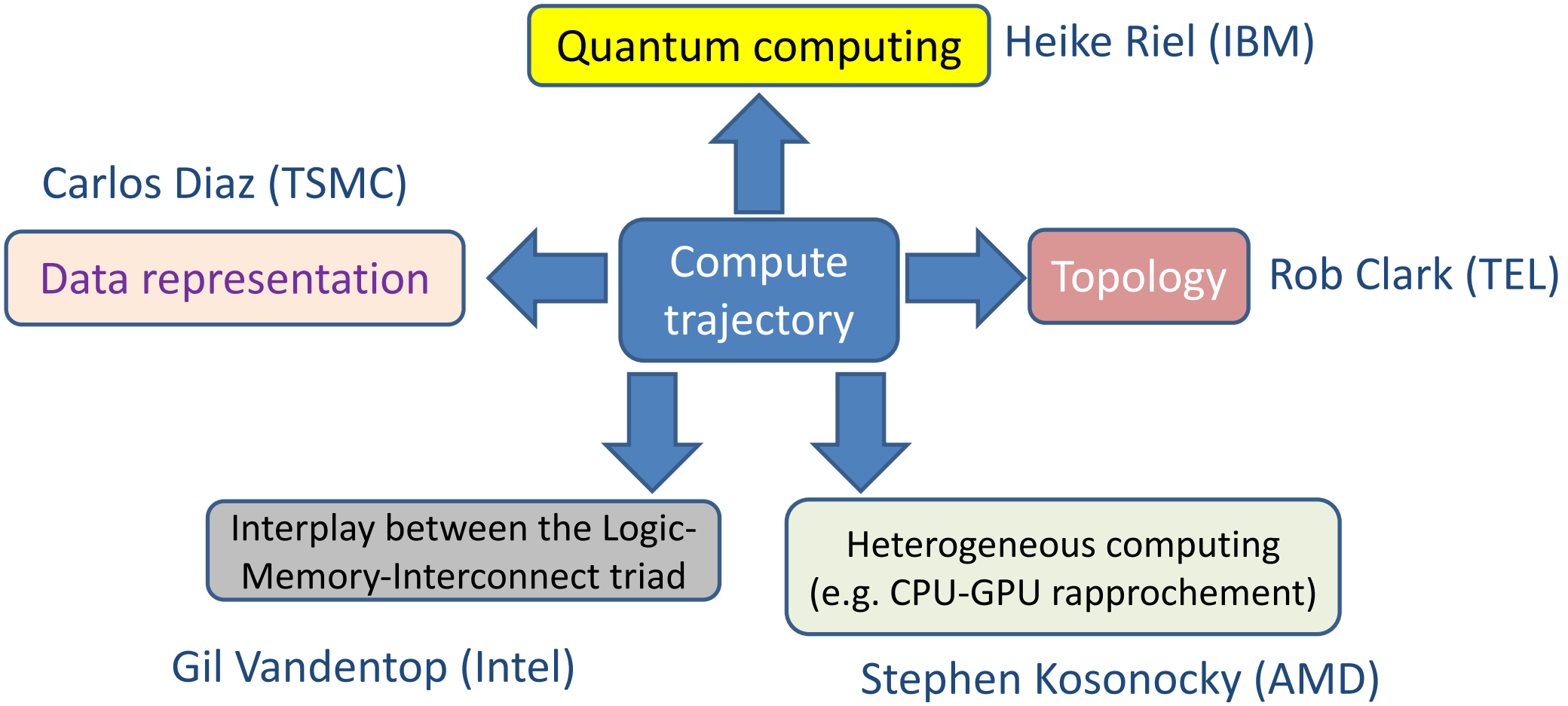
- It is paramount to restore U.S. leadership in microelectronic technologies and innovation
- The Decadal Plan for semiconductor research is instrumental to address on-going seismic shifts in information & communication technology (ICT)
  - The Decadal Plan provides an executive overview of the global drivers and constraints for the future ICT industry, rather than to offer specific solutions
    - The document identifies the 'what', not the 'how'
    - e.g. Discover compute trajectories with $p \sim 1$

$$MIPS = k\left(BITS\right)^{p}$$

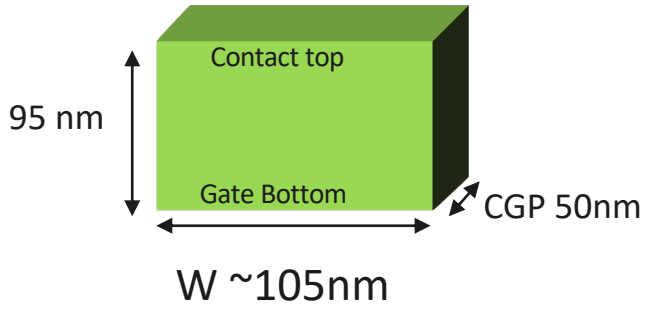
HOW

- With the 2030 Decadal Plan for Semiconductors released in January 2021, now is the crucial time to drive the conversion of the high-level Grand Goals of the Decadal Plan into a detailed Semiconductor Agenda toward 2030.

Research directions towards new compute trajectories



Why do we need 3D Architectures? – 5 nm example



How big is the transistor?

How big is the Chip?

-Processed Volume: STI Bottom to CMP Top

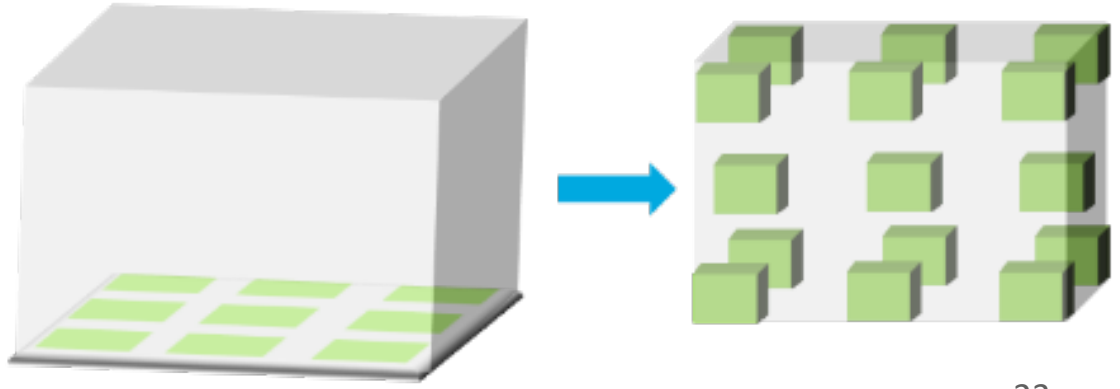
89,000,000 μm^2 die x 4.77 µm "height"

$$\frac{.0053\ \mu m^2/\text{transistor} \times 11.8E9\ \text{transistors/die}}{89{,}000{,}000\ \mu m^2/\text{die}} \times 100\% \sim 70\%$$

$$\frac{5E\text{-}4\ \mu m^3/\text{transistor} \times 11.8E9\ \text{transistors/die}}{426{,}000{,}000\ \mu m^3/\text{die}} \times 100\% \sim 1.4\%$$

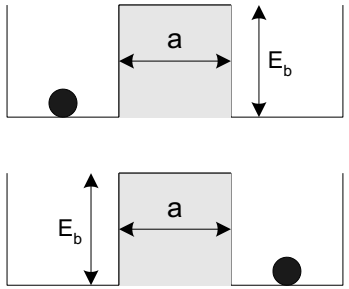
Most of the available Si Surface area is occupied by devices already, and devices are becoming (much) harder to scale down.

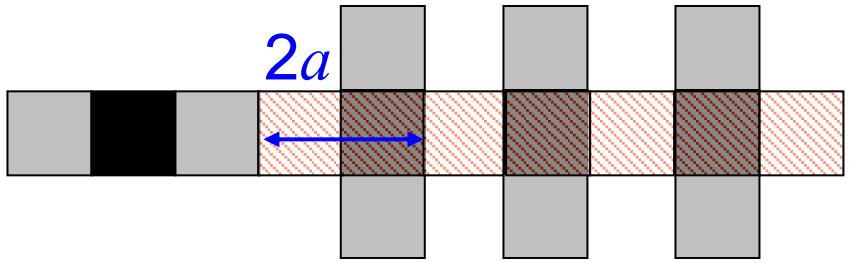
BUT, only a fraction of the processed volume of the chip is devices.

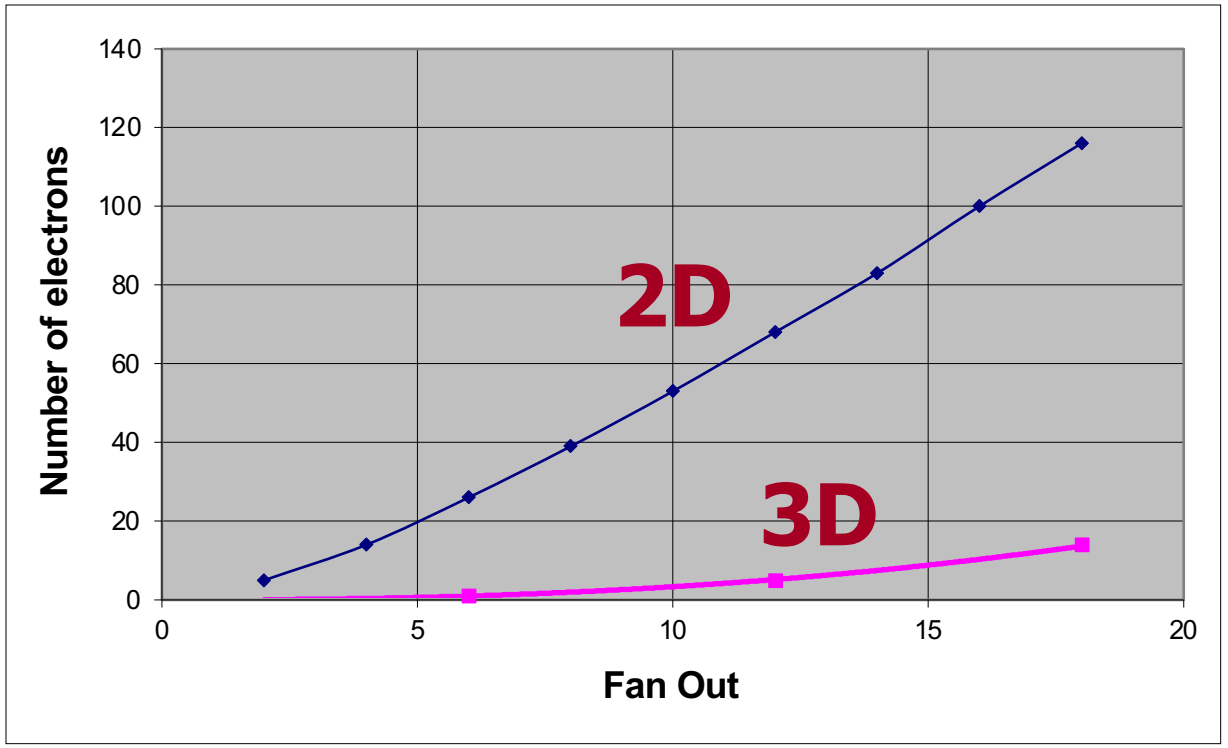


Energy costs for fan-out: 2D vs.3D

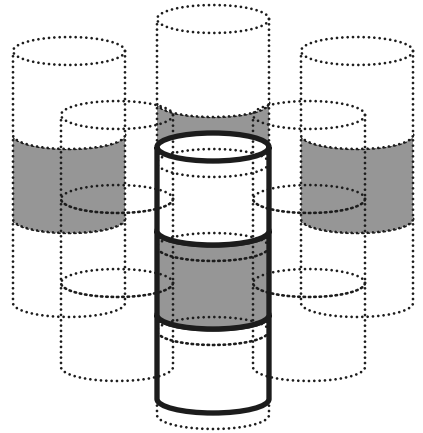
$$L_{\min} = 2a \cdot F$$







~ 10x energy reduction for FO4-6



Generic topology of a 3D binary switch

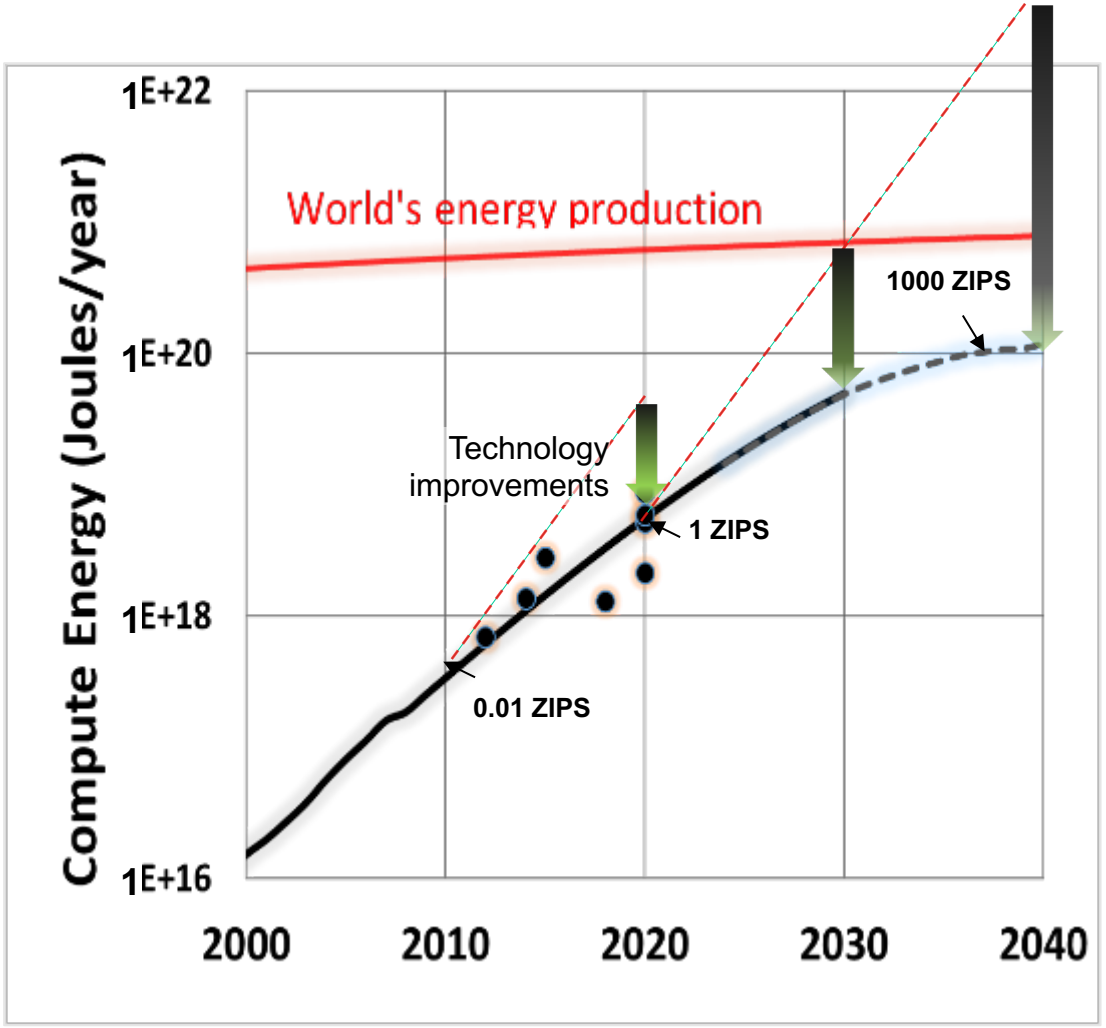
More Fan-Out (Branching) = More Computation

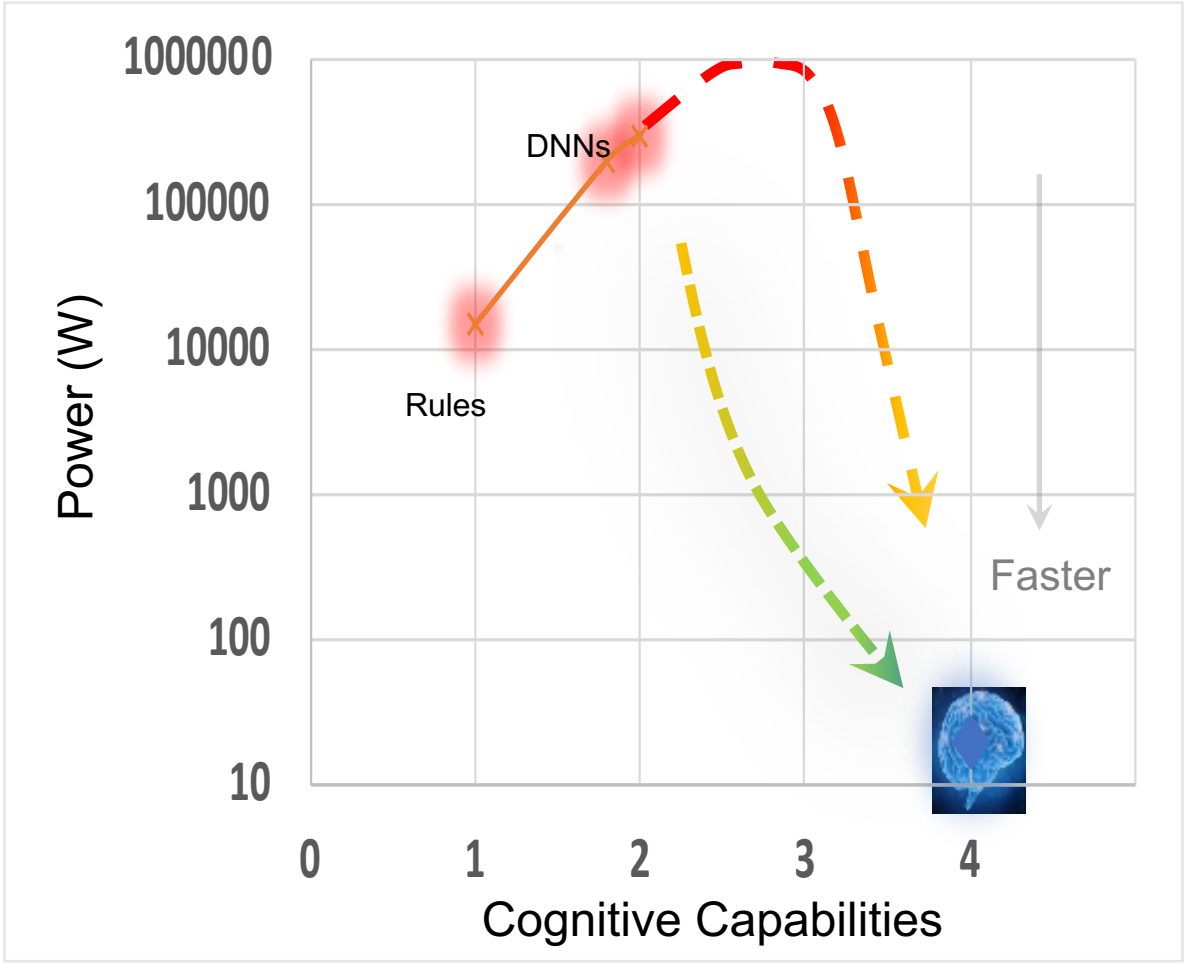
Convergence and Branching are necessary attributes of logical inference

Opportunities in 3D systems

- 2D layout results in long interconnects,
  - increased energy of operation, since more electrons are required for reliable switching.
  - Separates memory from logic by design – Van Neumann
- In a hypothetical 3D topology for binary switches, the generic shape of the binary switch corresponding to the basic energy diagram would be a vertical cylinder,
  - 3D organization of switches would allow for ‘stacked’ configurations, without as many additional wires as in 2D layout
  - It could enable ‘wireless’ communication between the sending devices A and several receiving devices by electrostatic coupling, which could dramatically reduce the number of electrons needed for branched communication (fan-out)
- The advantages of vertical 3D topology increase with fan-out, thus suggesting a larger-fan-out design approach might be desirable for 3D
  - Even for low fan-out, at least one order of magnitude in energy reduction could be expected
- More Fan-Out (Branching) = More Computation per step

Challenge: Enhanced energy efficiency while boosting computational performance and keeping cost in check



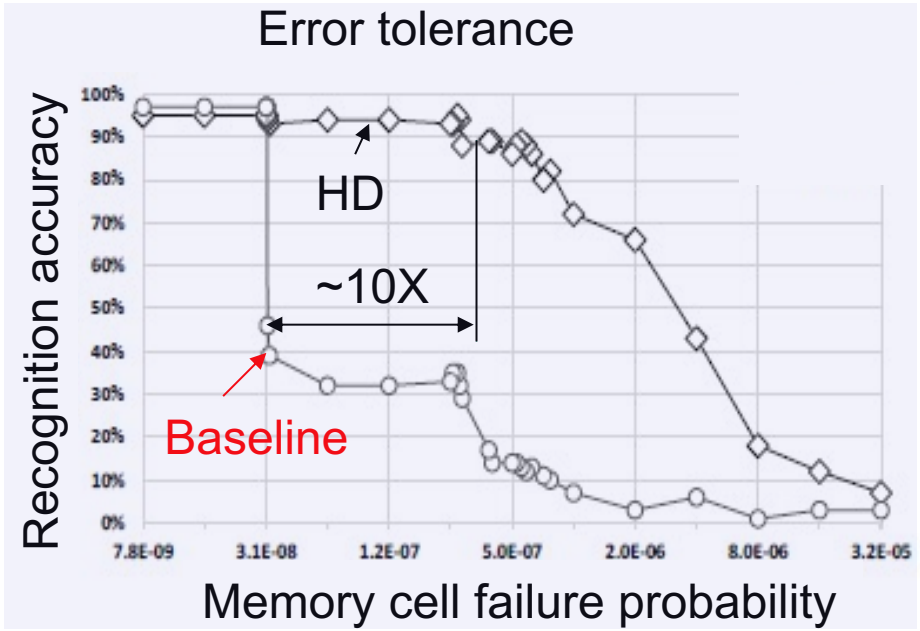


Ref. SRC-SIA Decadal Plan for Semiconductors

New Information representation and processing – a key knob to energy-efficient next generation AI capabilities

- High-dimensional computing can significantly enhance AI's computing energy-efficiency given its intrinsic one-pass & continuous learning capabilities, error robustness, and better scalability features compared to DNN

Representation Space	Conventional Computing	High Dimensional Computing
Elements	Low dimensional binary vectors	Vectors
Dimension → Cardinality	64 → 2^{64} = 1.8E19	1000 → 2^{1000} = 1.1E301
Type	Local	**Distributed**
Value / Meaning	Pattern itself	**Pattern relations**
Error tolerance	Low	**High**



Scalability

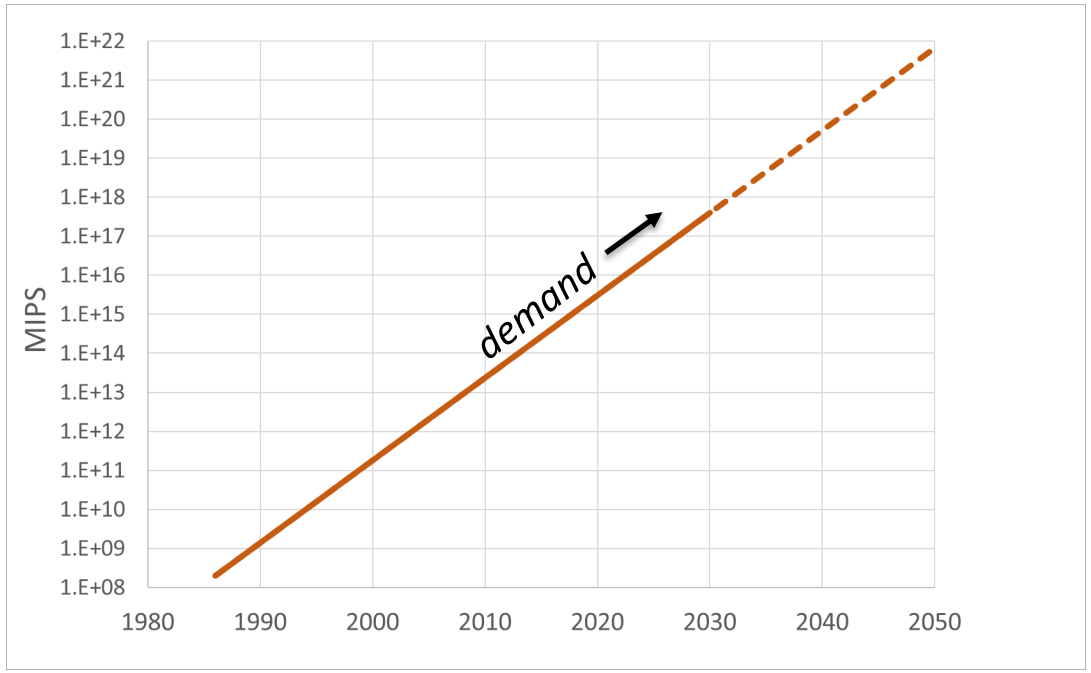
Problem Size Index	Memory (Kb) HD	Memory (Kb) Baseline
2	670	39
3	680	532
4	690	13837
5	700	373092
	$\propto n$	$\propto 27^n$

References, c.f.: A. Rahimi et al., 2016

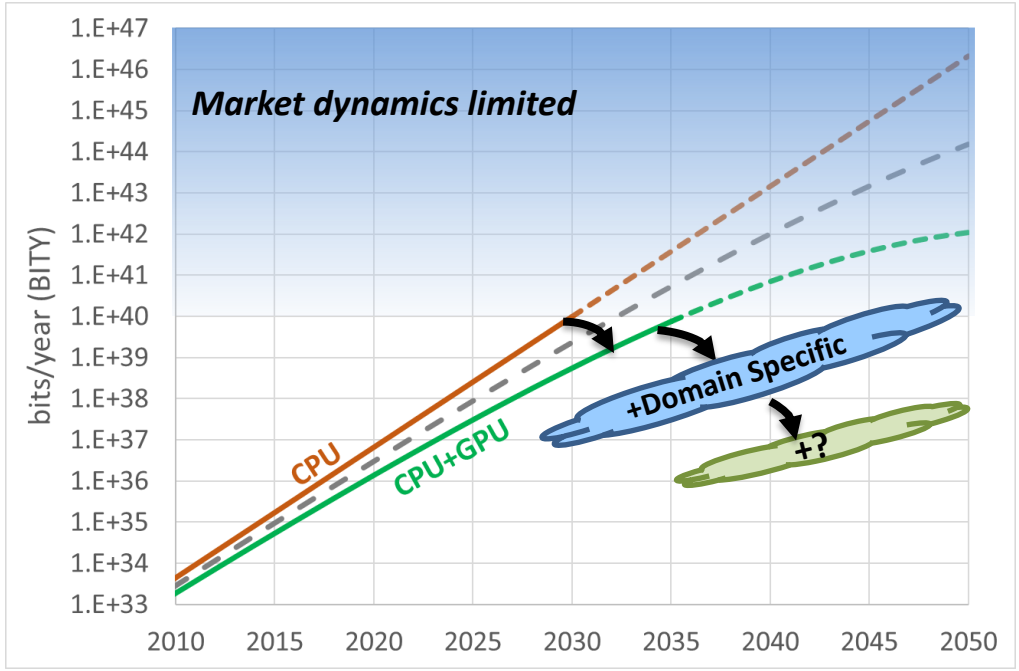
Changing the Compute Energy Trajectory: Heterogeneous processing

The current trend is toward the increasing use of GPU based architectures for different computational tasks, including general-purpose computations and machine learning/artificial intelligence.

Measure of aggregate worldwide computational performance (MIPS)



Total number of "raw" binary transactions required to service worldwide computational demand

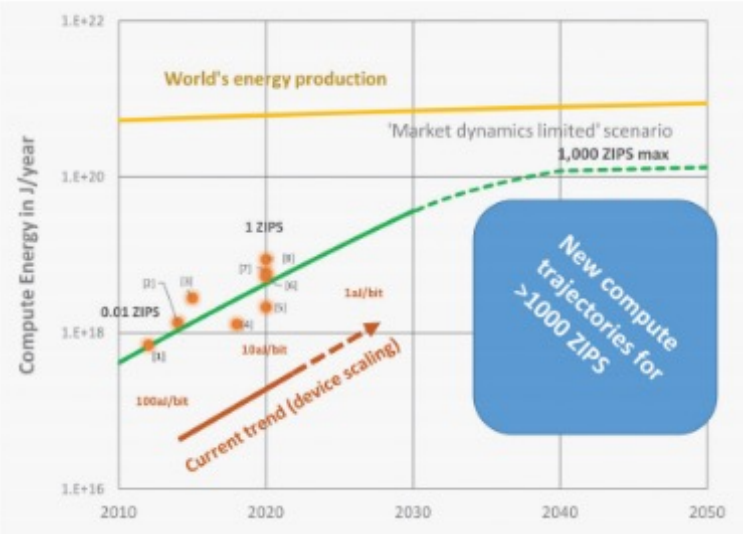


Source: Decadal Plan for Semiconductors

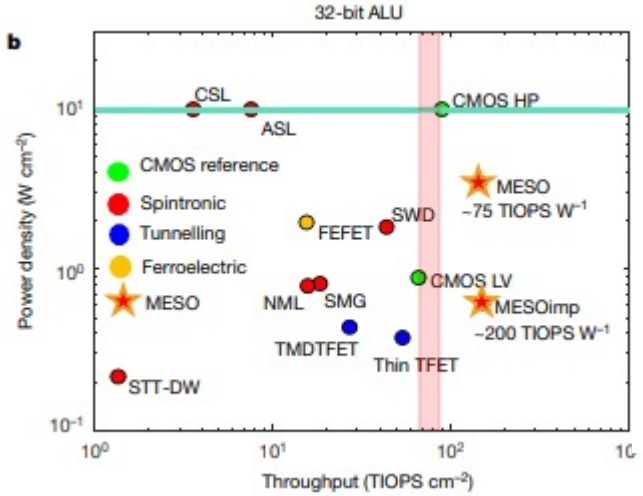
The increased use of GPUs and domain specific architectures helps to decrease the total number of "raw" binary transitions required for computing without sacrificing the computation capacity (MIPS), shifting energy limitations further off in time.

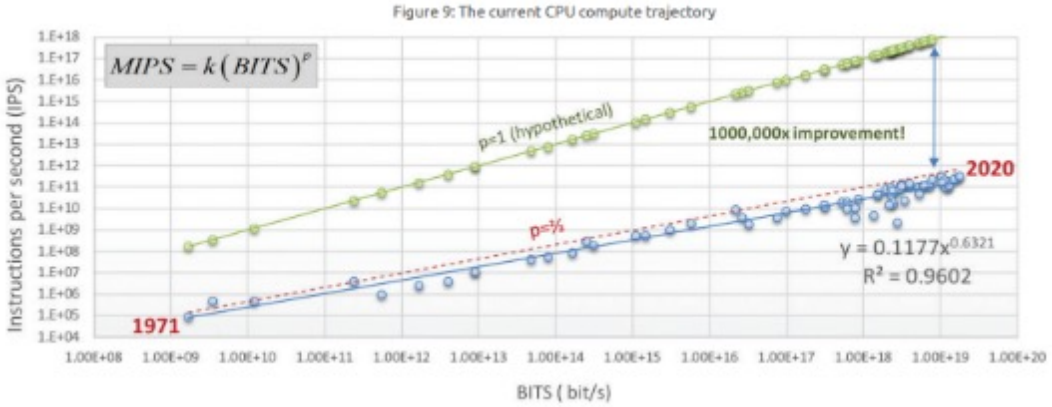
Changing the Compute Energy Trajectory

"The Problem and Opportunity"



Many device options exist and new approaches are needed



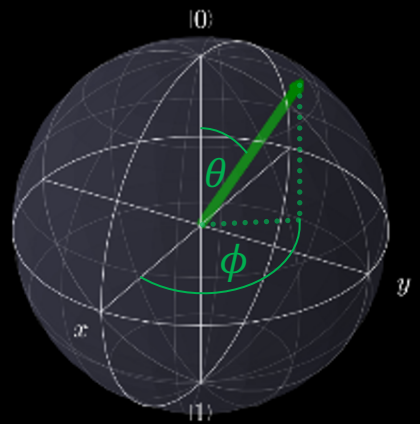


1) Fundamental studies around the Logic-Memory-Interconnect Triad and a better understanding of details within the historical p=2/3 trend may highlight opportunities.
2) A new architecture for Compute in Memory or Compute in interconnect is one of our best hopes since the brain functions so efficiently but has not yet shown us the way there for silicon.
3) There has been a lot of work to date around compute in memory, without an energy efficiency breakthrough, because moving the data remains energy intensive.
4) MESO devices are the most promising beyond CMOS option. There are many materials and device challenges remaining. We need to accelerate the progress towards these new device solutions.

Quantum Computing – Changing the Game

Quantum works different !

Quantum Bit - Qubit



Superposition: can be an arbitrary point on the Bloch sphere (α |0⟩ + β |1⟩)

0 *and* 1

Entanglement & Interference

n qubits – 2^n basis states

➔ Quantum laws can enable exponential increase of computing power

Quantum is built for difficult problems !

Problems we can't address adequately today

Problems we can address today

937 x 947 = ?

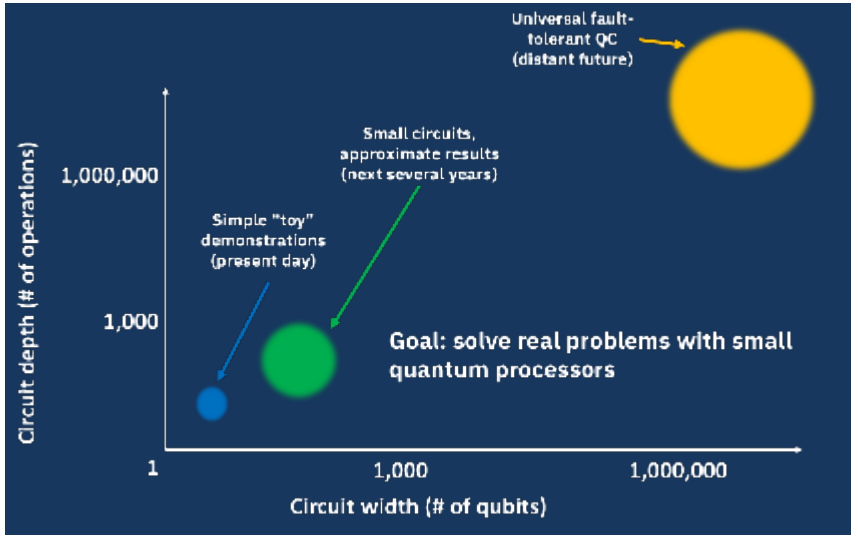
Problems we can address with quantum

e.g. Simulating Quantum Mechanics
Factoring: 887339 = ? x ?

Quantum Applications:

- Simulating Quantum Systems
- Algebraic Problems: ML, Differential Equations, Factoring
- Database Search: Quantum Monte Carlo, Optimization, Graph Problems

Development of Quantum Computing



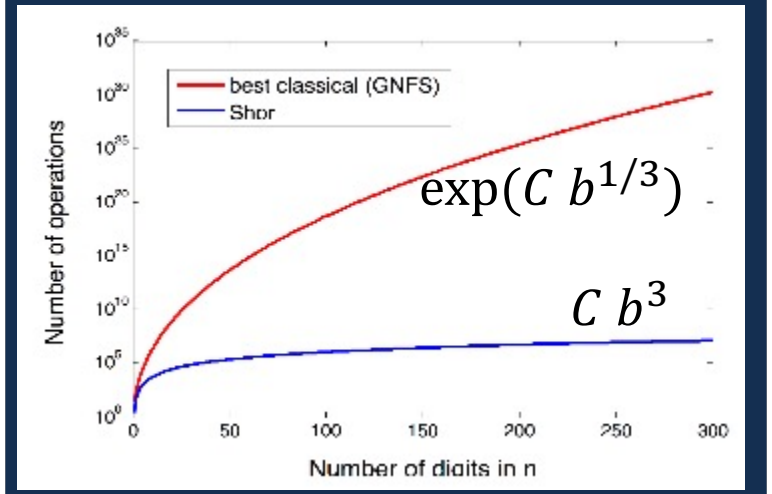
➔ Goal: Demonstrate Quantum Advantage

Commercial advantage to solving real world problems with quantum computers

➔ A new path to solve some of the hardest problems

Quantum Computing – Benefit

Shor Algorithm
Universal Quantum Computer



Exponential speed-up for factoring:
A task taking **300 years** (2^{33} seconds) on a classical computer might take a minute (~ **30 seconds**) on a quantum computer

➔ This can lead to speed up and reduced energy consumption

Today's examples – speed and energy gain

Example 1:
Measure probability distribution by sampling computational state-space of dimension 2^{53} (ca 10^{16}).

- 200s on QC and $2 \cdot 10^5$ days on HPC
- Roughly: $5 \cdot 10^5$ J vs $3 \cdot 10^{12}$ J

Example 2:
Boson sampling on 76 photons implemented in photonic QC.

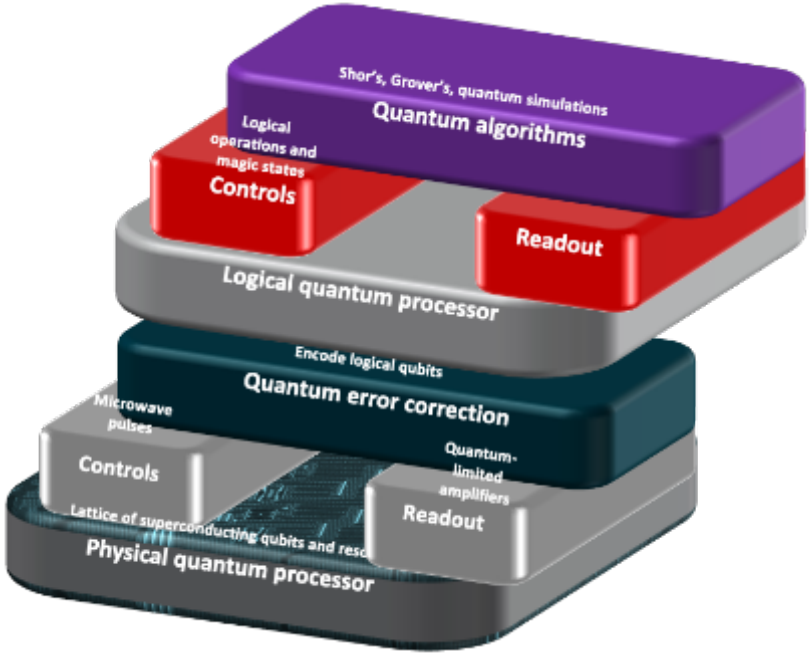
- 20s on QC and $6 \cdot 10^8$ years on HPC

Example 3:
Simulations of hard random quantum circuits:

Energy cost (MWh)		
Electra	Summit	QPU
96.8	21.1	4.2×10^{-4}

Villalonga et al. Quantum Sci. Technol. 5 (2020) 034003

But much more to be considered



- Energy overhead of error correction?
- Scaling to millions of qubits
- Integration of electronics
- …

➔ Further understanding and work needed to benchmark

➔ Potential to save energy exists

Gambetta, Chow, Steffen, npj Quantum Information 3, 2 (2017)]

Panel Discussion Questions

- Application Drivers and Challenge Problems
  - How will the R&D called for in the SRC Decadal Plan impact future energy efficient computing solutions across many applications?
  - Do you foresee specific resource needs to drive the seismic shift in computing hardware and software to support these future applications?

Panel Bonus Question

- Energy efficiency is driving computing to be more:
  - Heterogeneous – leveraging architectural specialization
  - Distributed – integrating large scale HPC/Data Centers networked to Edge Servers, to Edge Computing and IOT devices
  - What can we do to ensure software investments stay aligned with this evolving hardware infrastructure?