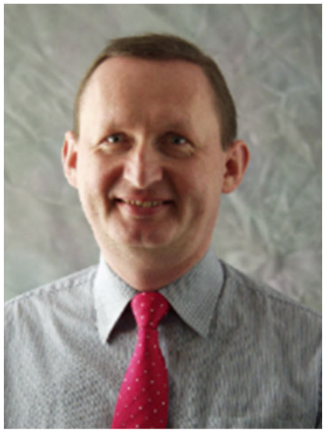
Semiconductor Research Corporation

New Compute Trajectories for Energy-Efficient Computing



Victor Zhirnov, Chief Scientist,
Semiconductor Research Corporation

Princeton Plasma Physics Laboratory

June 15, 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
  - Co-design Challenges and Decadal Plan
- Summary

Suhp lhu#P lfurhdnfwurqlfv#Frqvruwxp #Vlqfh#1<;5

Sulydwh#Vhfwru#bqg#Jqwhudjhqf|/#Sduwflsdwlrq#bqg#Jryhuqdqfh

DARPA NSF NIST

AMD ANALOG DEVICES arm EMD PERFORMANCE MATERIALS

GLOBALFOUNDRIES IBM intel LOCKHEED MARTIN

Mentor A Siemens Business Micron MUBADALA NORTHROP GRUMMAN

NXP QUALCOMM Raytheon Technologies SAMSUNG

SK hynix TEL Texas Instruments tsmc Veeco

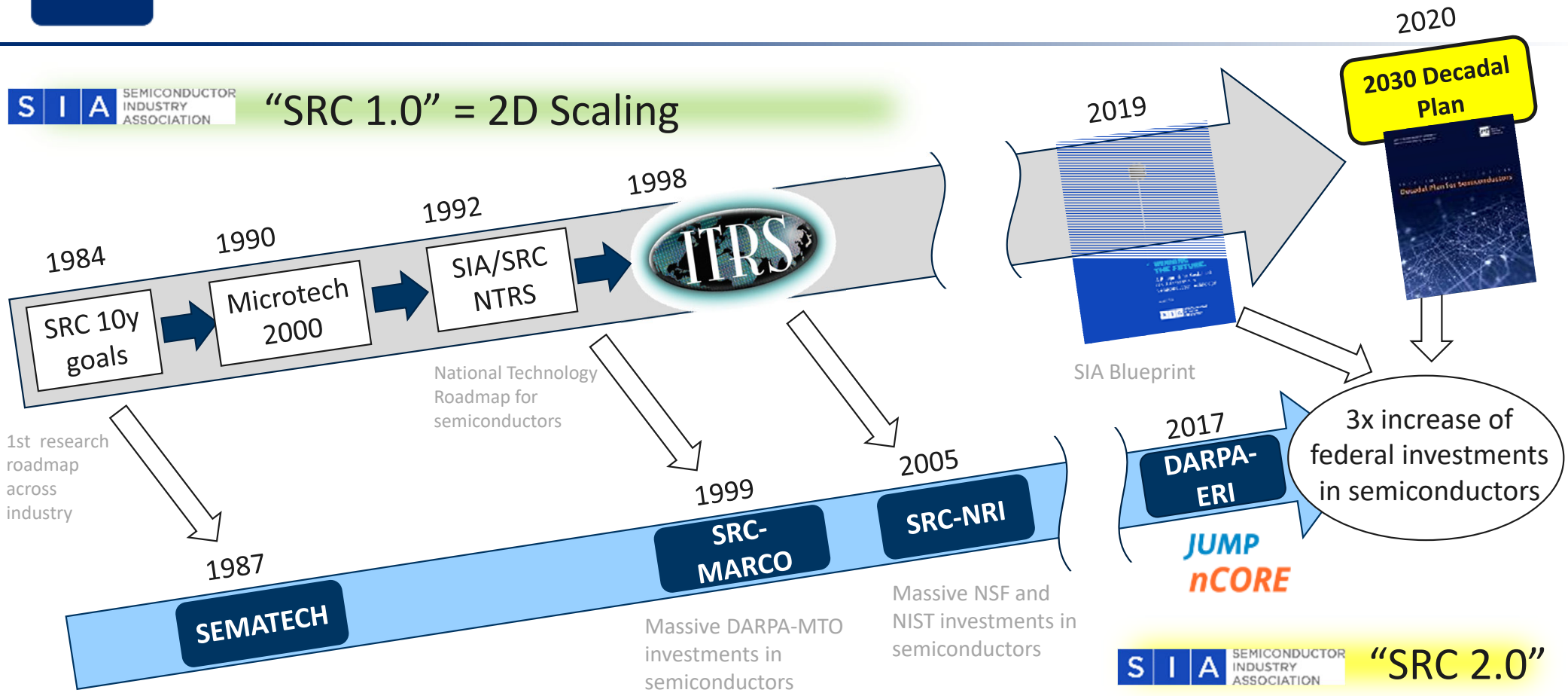
SRC is a trusted advisor with a vast network, community, and shared dedication to research, prototyping, and workforce training in advanced semiconductor technologies

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…


SEMICONDUCTOR INDUSTRY ASSOCIATION
"SRC 1.0" = 2D Scaling
1984
SRC 10y goals
1990
Microtech 2000
1992
SIA/SRC NTRS
1998
ITRS
2019
2020
2030 Decadal Plan
SIA Blueprint
1st research roadmap across industry
National Technology Roadmap for semiconductors
1987
SEMATECH
1999
SRC-MARCO
Massive DARPA-MTO investments in semiconductors
2005
SRC-NRI
Massive NSF and NIST investments in semiconductors
2017
DARPA-ERI
JUMP
nCORE
3x increase of federal investments in semiconductors
SEMICONDUCTOR INDUSTRY ASSOCIATION
"SRC 2.0"

SRC®

Decadal Plan Participants

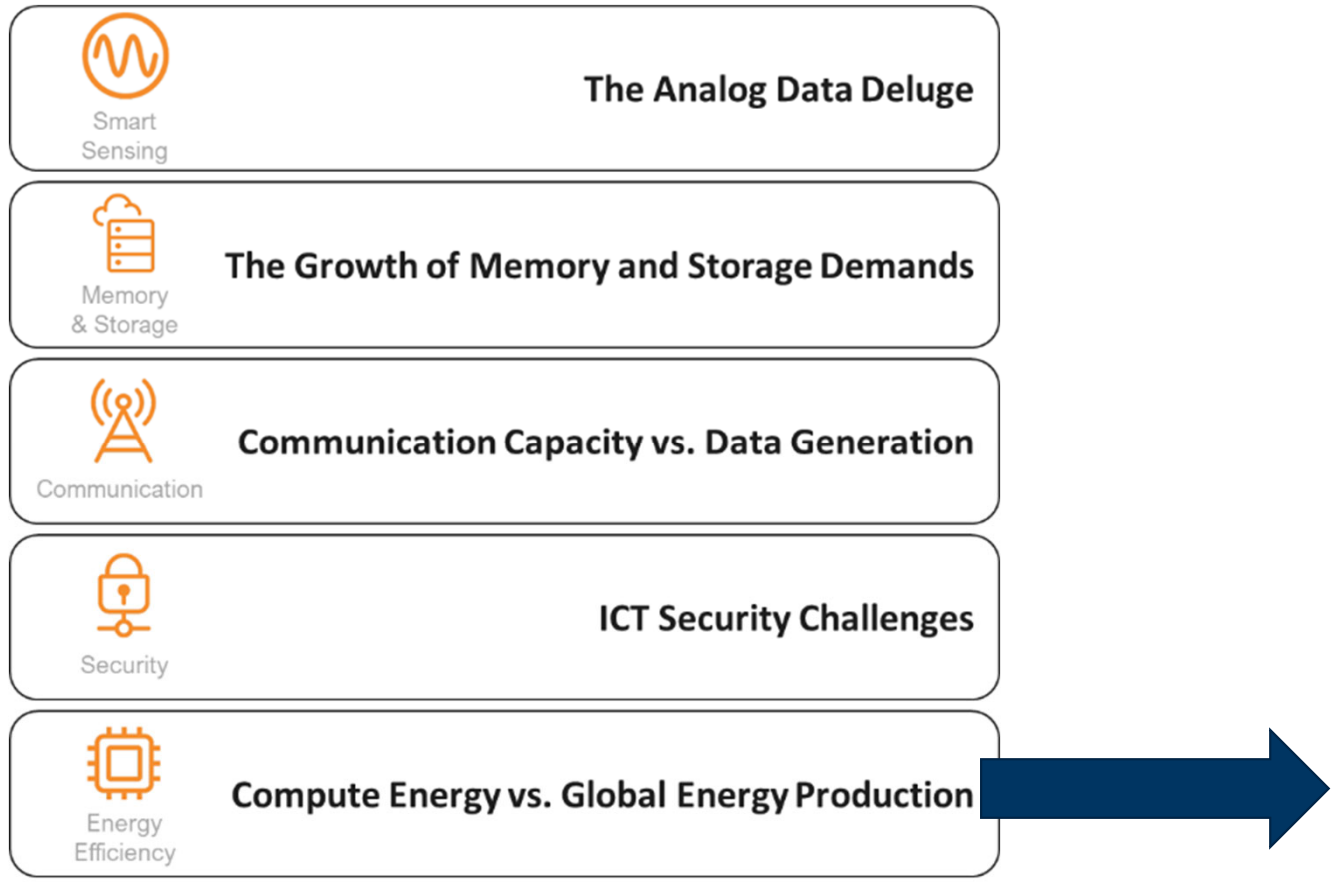
IBM
intel
SIEMENS
QUALCOMM
Texas Instruments
TEL TOKYO ELECTRON
MUBADALA
NXP
AMD
SAMSUNG
SK hynix
arm
ANALOG DEVICES
tsmc
Western Digital
Micron
MITRE
OXFORD INSTRUMENTS
MEDIATEK
amazon
SIA SEMICONDUCTOR INDUSTRY ASSOCIATION
U.S. DEPARTMENT OF ENERGY
Sandia National Laboratories
NIST
Pacific Northwest NATIONAL LABORATORY

Our 2030 Decadal Plan for Semiconductors

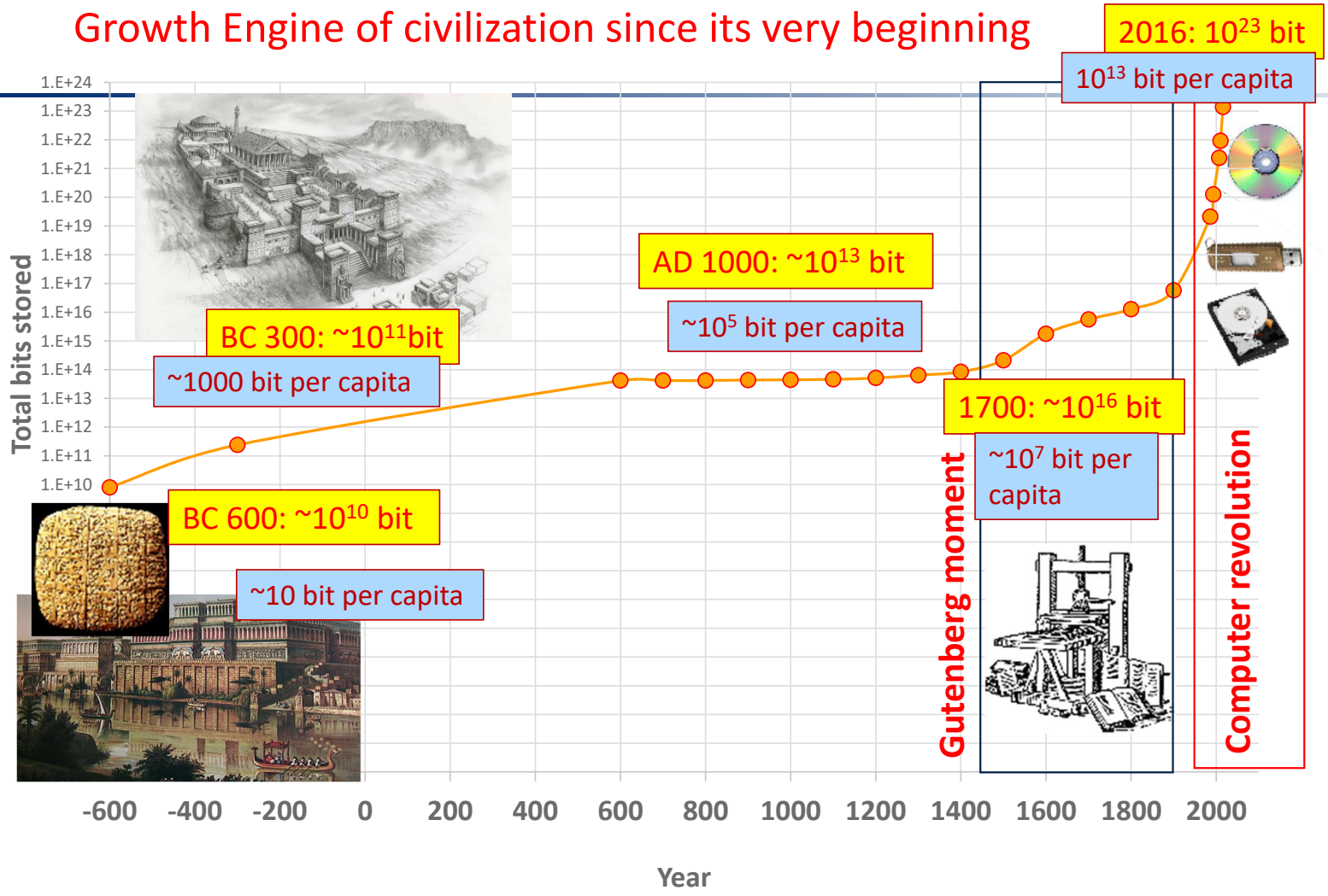
https://www.src.org/about/decadal-plan/ (released on January 25, 2021)

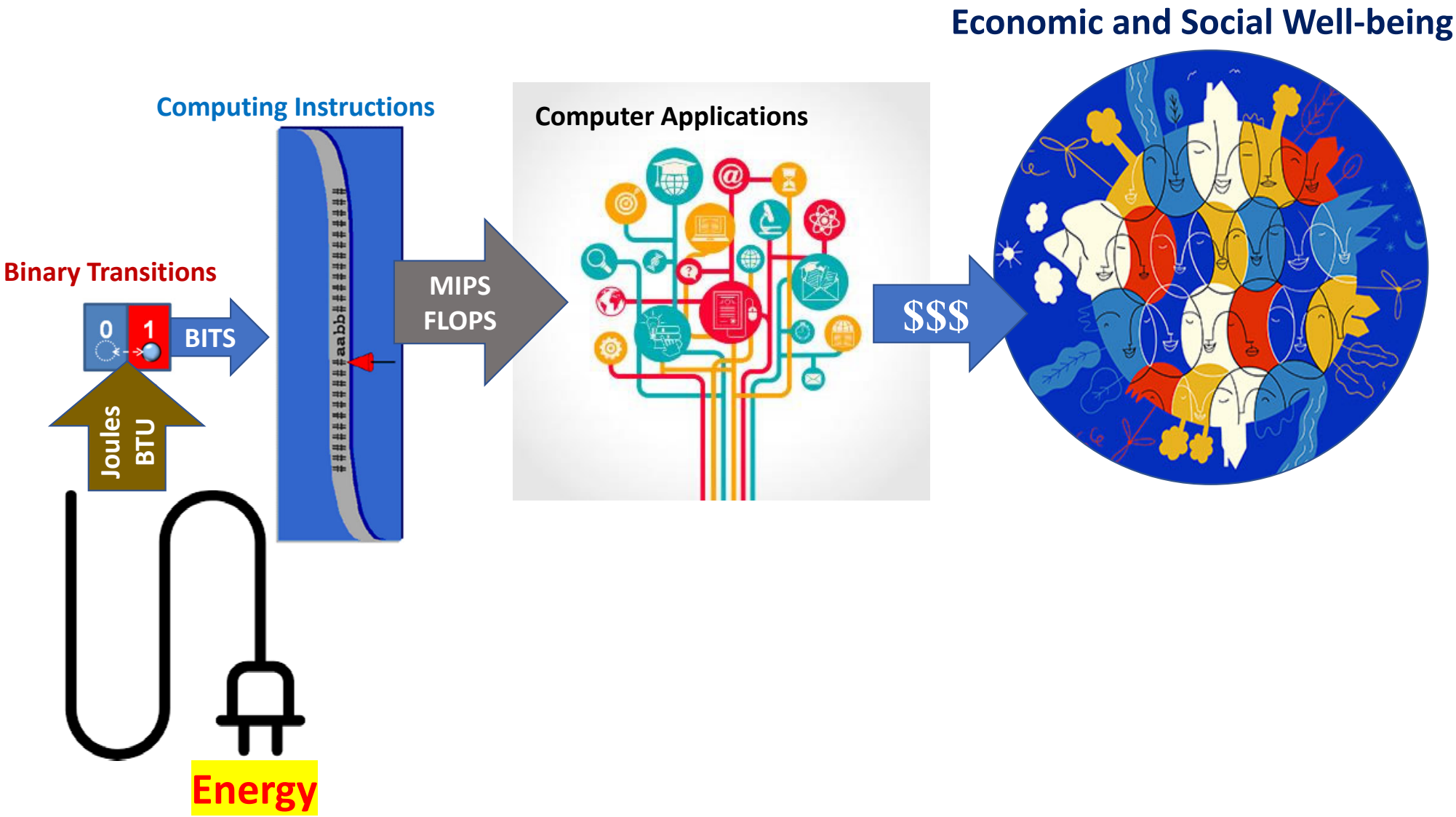
SIA and SRC call for +34B in semiconductor R&D throughout the 20s

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



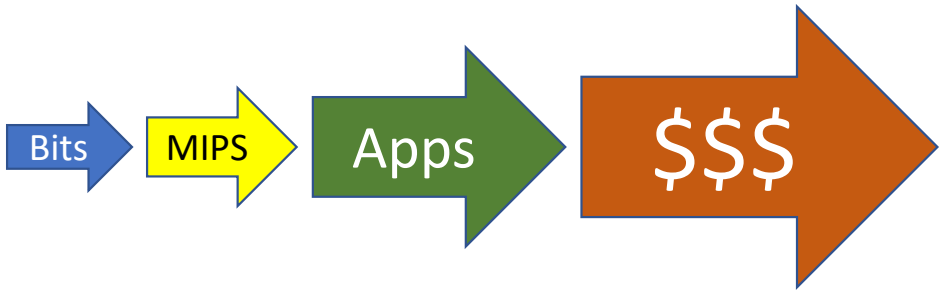
Information along with Energy has been the Social-Economic Growth Engine of civilization since its very beginning




Economic and Social Well-being
Computing Instructions
Computer Applications
Binary Transitions
0
1
BITS
Joules
BTU
MIPS
FLOPS
$$$
Energy

The global ICT Ecosystem

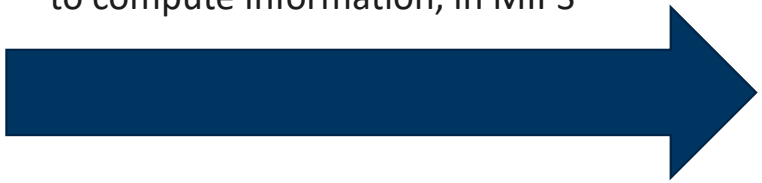
Figure courtesy of Analog Devices, Inc.



G1: Computation

- *Personal computers (PC)*
- *Professional computers*
- *Supercomputers*
- *Game consoles*
- *Electronic calculators*
- *Mobile phones and PDAs*
- *Digital Signal Processors (DSP)*
- *Microcontroller (MCU)*
- *Graphic Processing* Units (GPU)

World's technological installed capacity to compute information, in MIPS

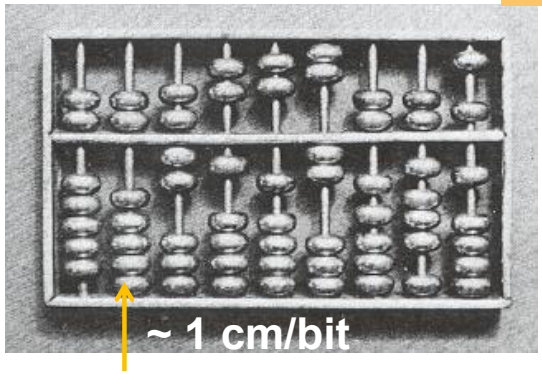


Hilbert and Lopez, Science (2011) 332 pp. 60-65

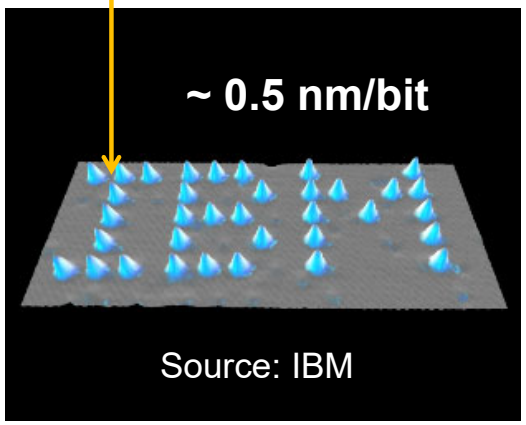
What is Information?

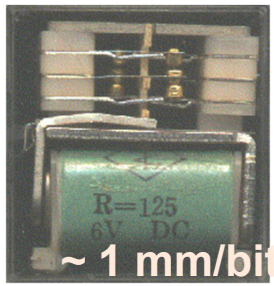
Information is measure of distinguishability

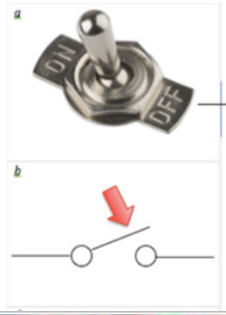
e.g. of a physical subsystem from its environment…

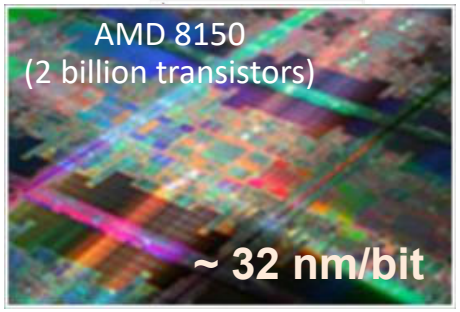


Information-bearing particles







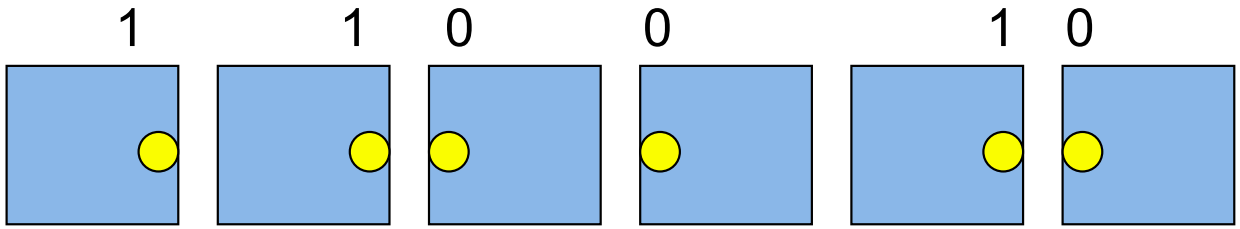


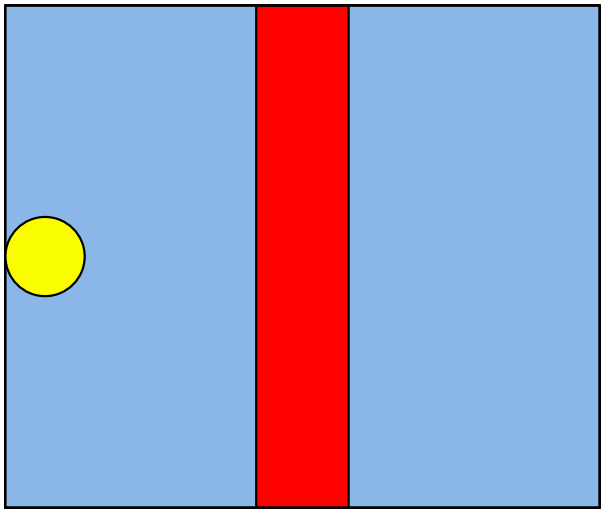
$$I = \log_2 N$$

A THEME: Minimal ICT Element

What is the smallest volume of matter needed for an ICT element? What is the smallest energy of operation?

Particle Location is an Indicator of State

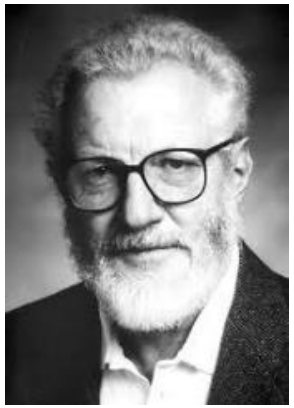




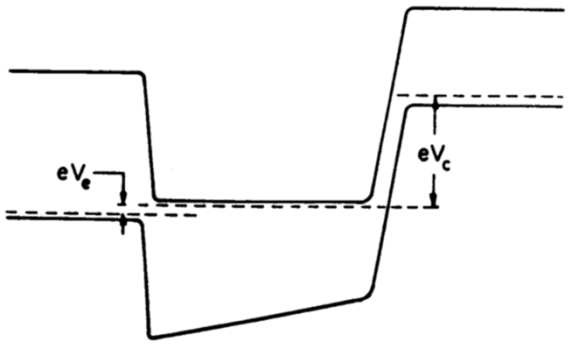
Kroemer's Lemma of Proven Ignorance

- If in discussing a semiconductor problem, you cannot draw an Energy-Band-Diagram, this shows that *you don't know what are you talking about*

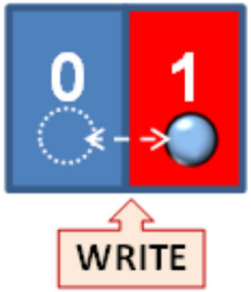
- If you can draw one, but don't, then *your audience won't know what are you talking about*

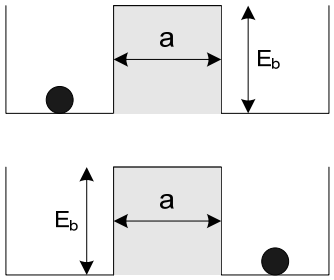


Herbert Kroemer
Nobel Lecture, Dec. 8, 2000



Information Processing Technology Desirata 1.0





Designers and Users want:

- Highest possible integration density (***n***)
  - *To keep size small*
  - *To increase functionality*
- Highest possible speed (***f=1/t***)
  - *Speed sells!*
- Lowest possible power consumption (*P*)
  - *Decrease demands for energy*
  - *The generation of too much heat means costly cooling systems*

Binary information throughput

$$\beta = nf$$

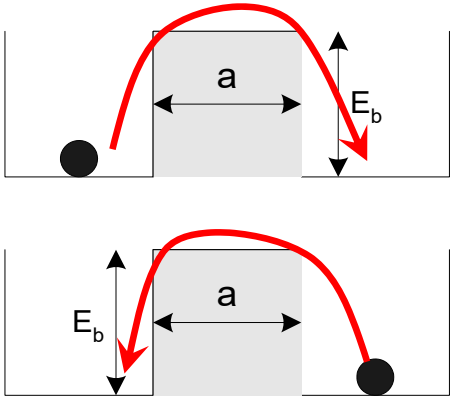
$$P = E_{bit} nf$$

Power consumption

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

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

Thermal Noise

Shannon - von Neumann - Landauer limit

Scaling Limits: The Heisenberg Constraint

$$\Delta p = \sqrt{2mE_b}$$

$$\Delta x \Delta p \geq \frac{\hbar}{2}$$

$$\Delta E \Delta t \geq \frac{h}{2}$$

$$a_{crit} \sim \frac{\hbar}{\sqrt{2mE_b}}$$

$$t_{\min} \sim \frac{h}{2E_b}$$

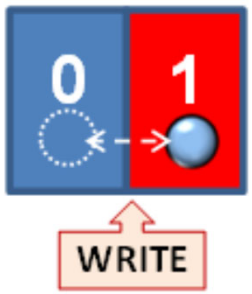
At this size, tunneling will destroy the state

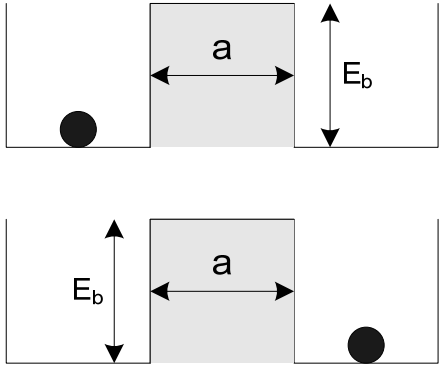
Minimal time of dynamical evolution of a physical system

N. Margolus and L. B. Levitin, Physica D 120 (1998) 188

Physics of Information: Central Question

- What are the smallest volume of matter and energy needed to create a bit of information?





Example: binary switch
(Theoretical limits)
a_{min} ~ 1.5 nm

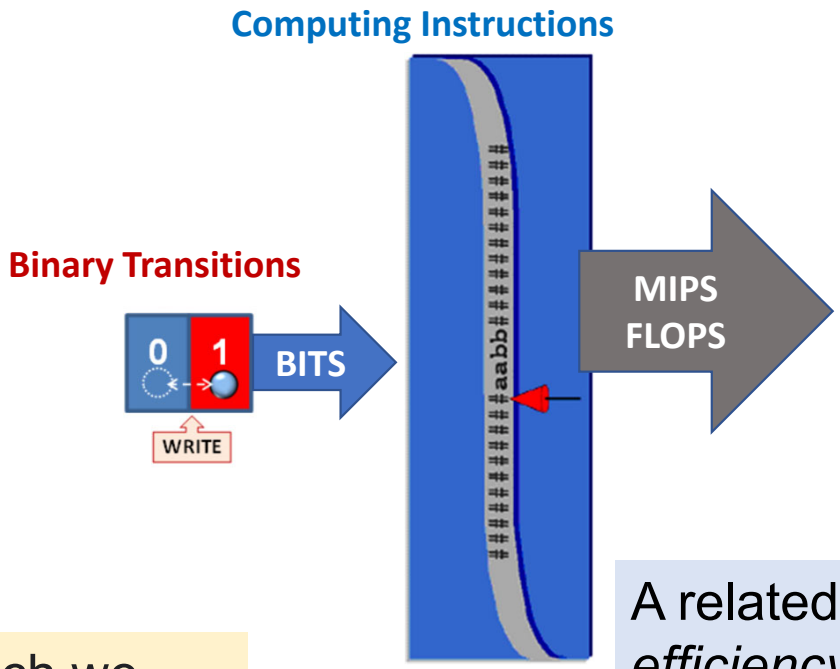
t_{sw} ~ 40fs

E_{bit}~10^{-21} J

Binary information throughput

$$\beta = nf$$

What is Compute Trajectory?



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

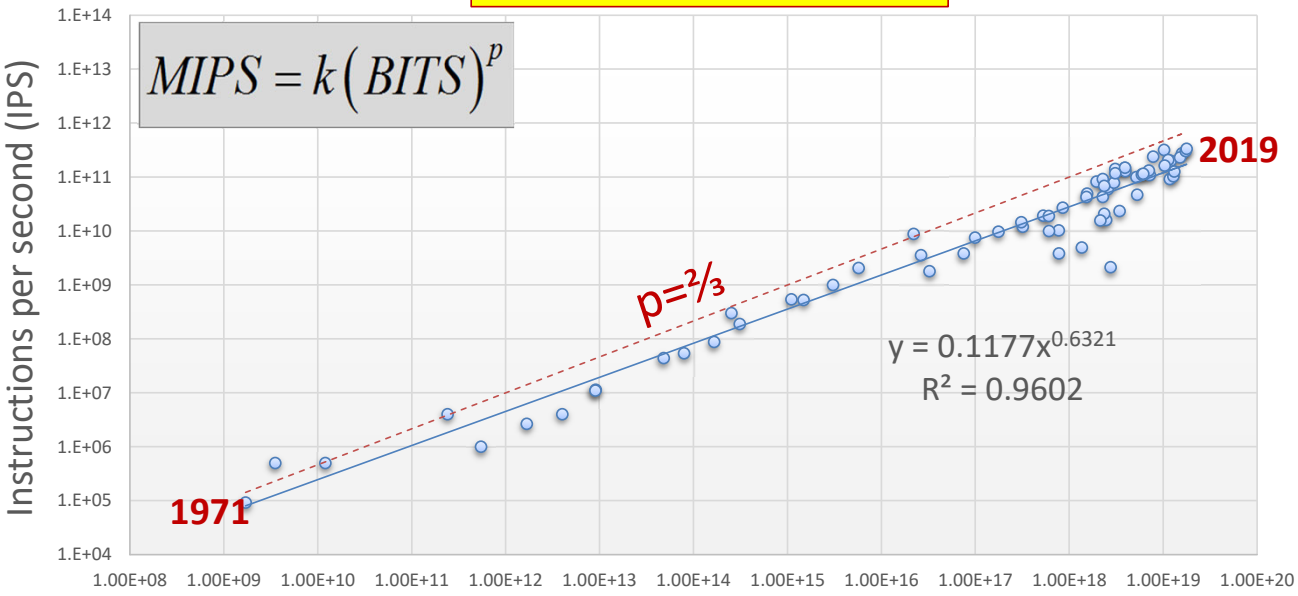
Bits → MIPS

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

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

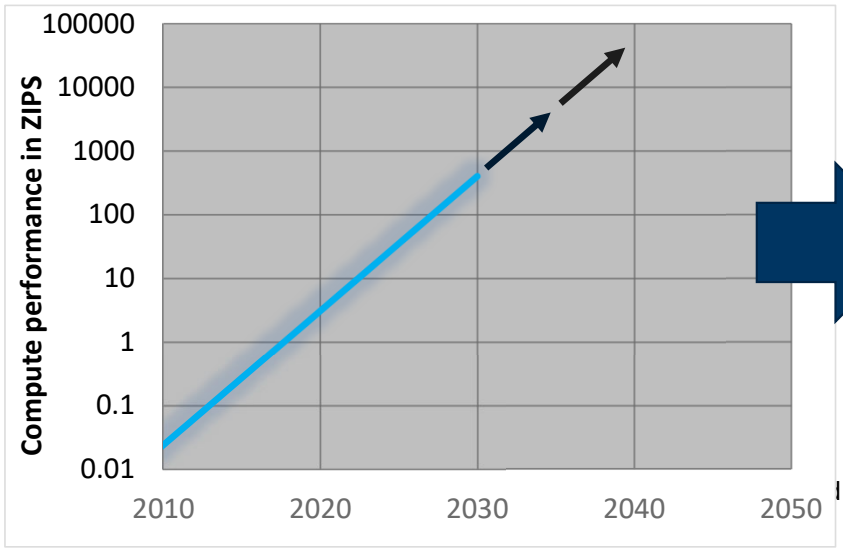
k=0.1, p= 0.64 ≈ ⅔

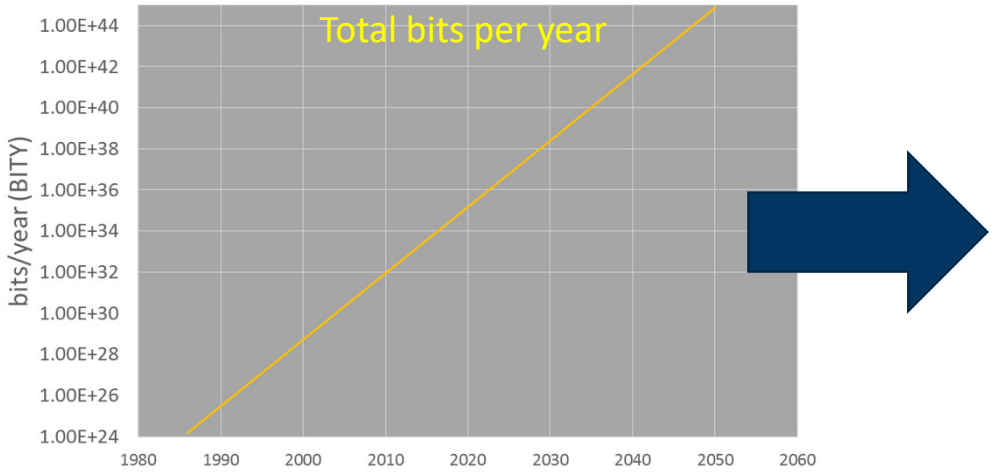


$$\beta = \alpha N_{tr} \cdot f \quad \Rightarrow \quad P = \beta E_{bit}$$

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 21064 EV4	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 E3-1290 v2	2012
Intel	Ivy Bridge-EX-15	2013
Intel	i7-5960X	2014

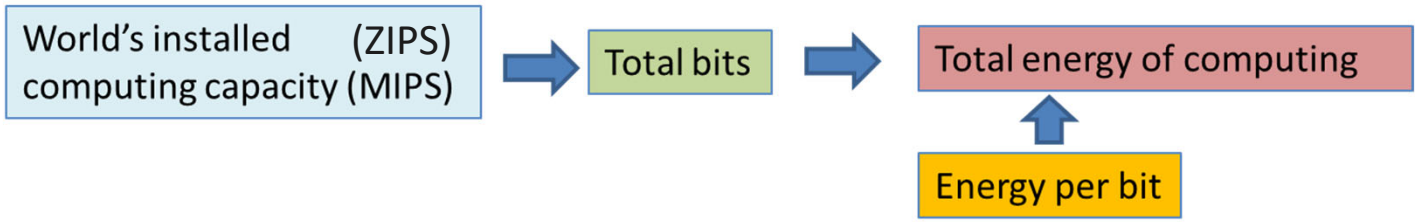
Computations per Year



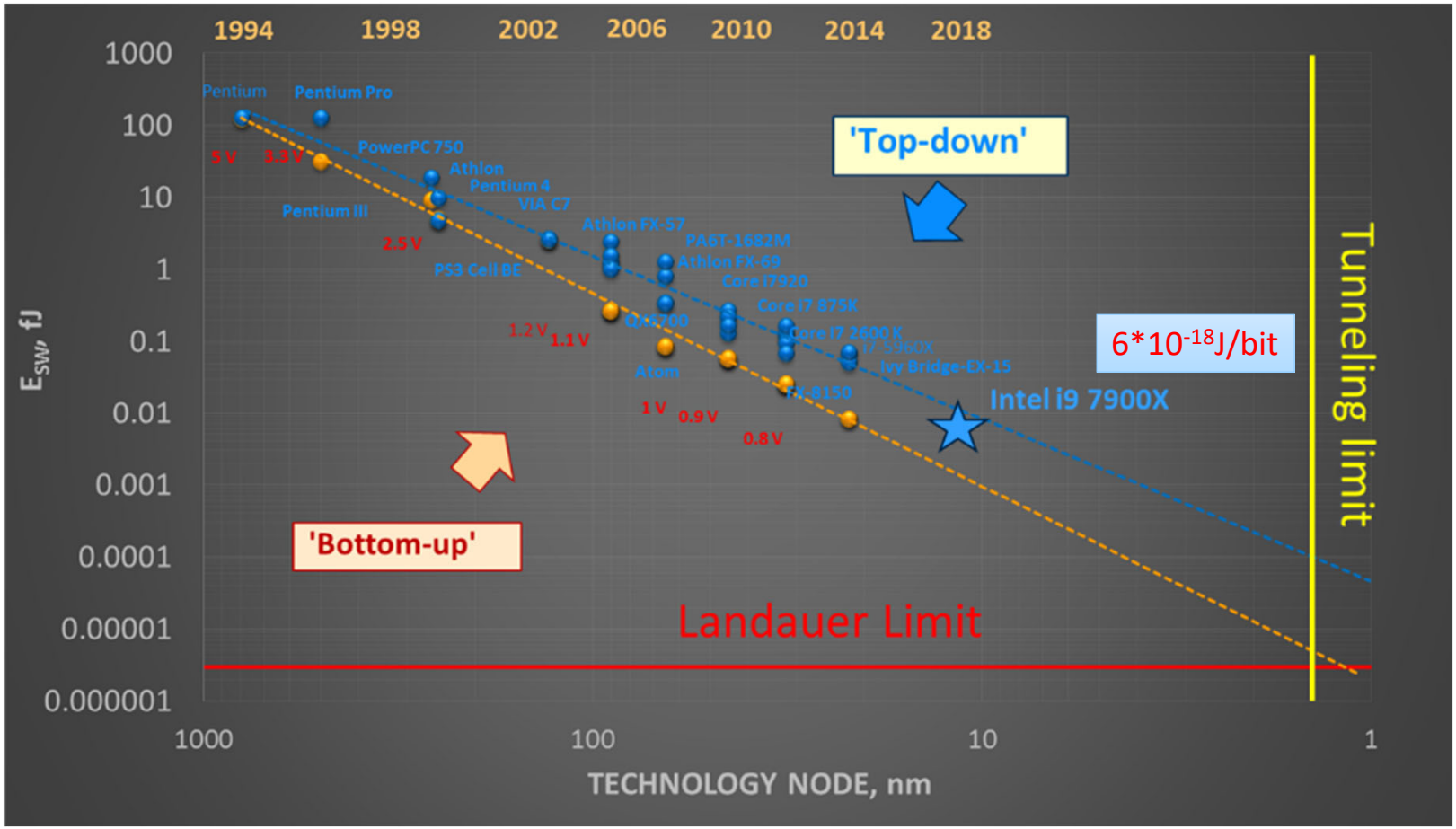


l P. Lopez, “The World's Technological Capacity to Store, Communicate,

1 ZIPS= 10^{15} MIPS



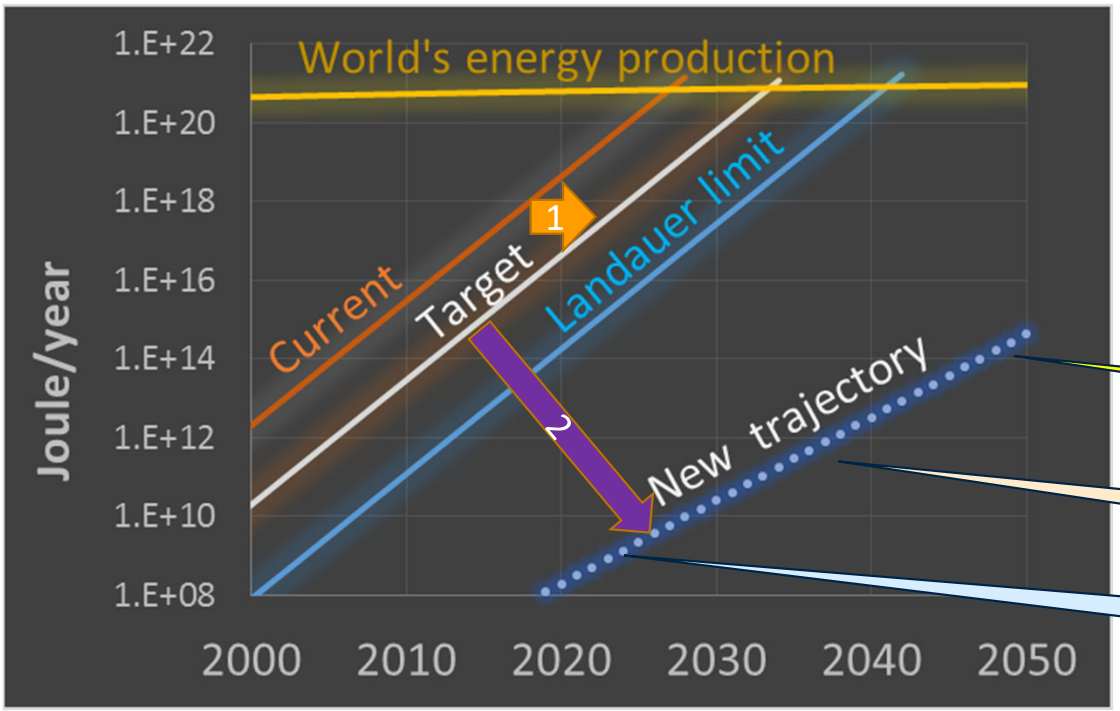
Computing energy: Energy per bit in CPU



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

Current: 10^{-17} J/bit

Target: 10^{-18} J/bit

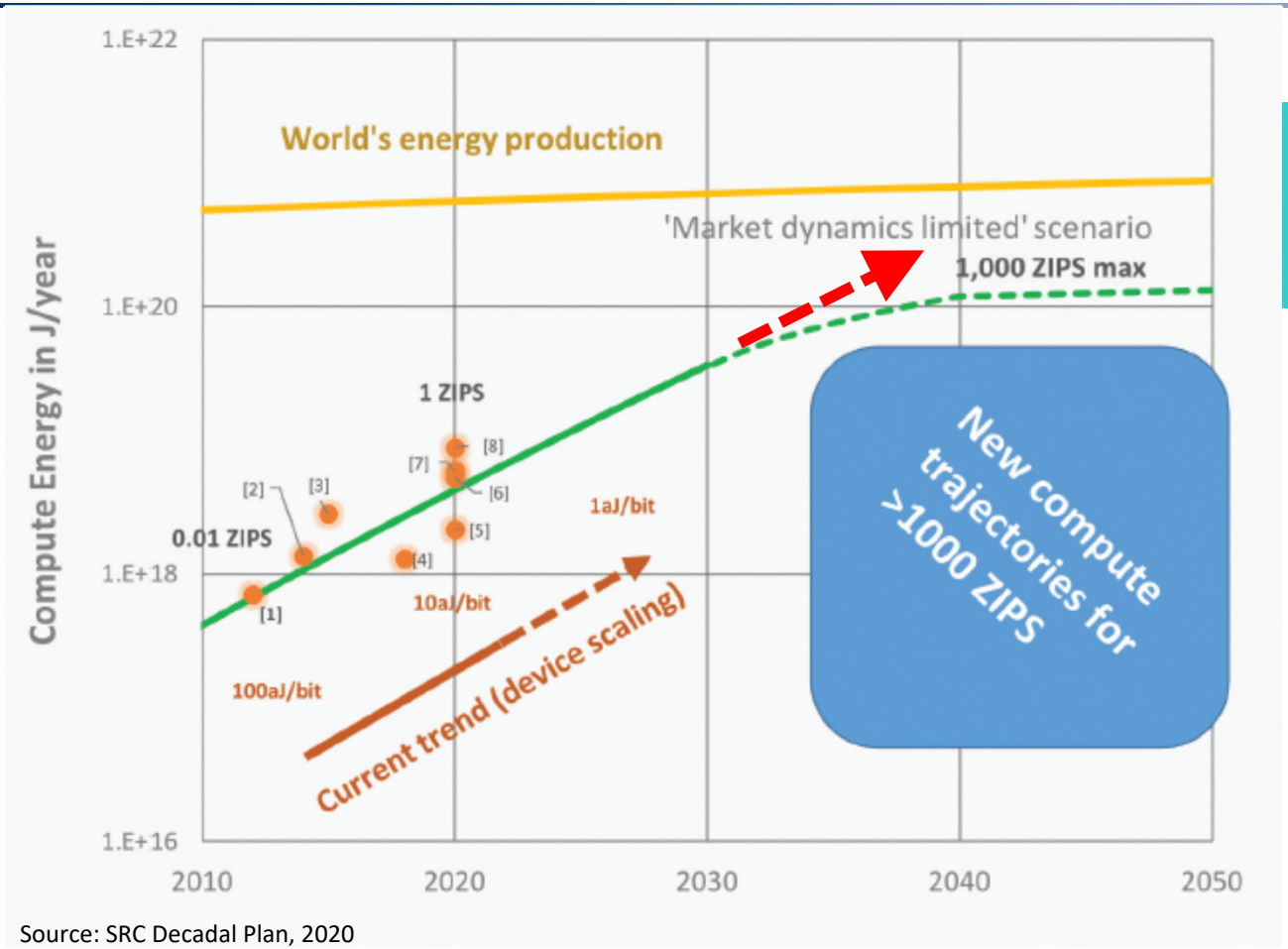
Landauer limit: 10^{-21} J/bit

New trajectory: $p \approx 1$

- Quantum computing
- Neuromorphic
- AI engines

SRC Select Disclosure

Seismic shift #5: Computing energy is not sustainable



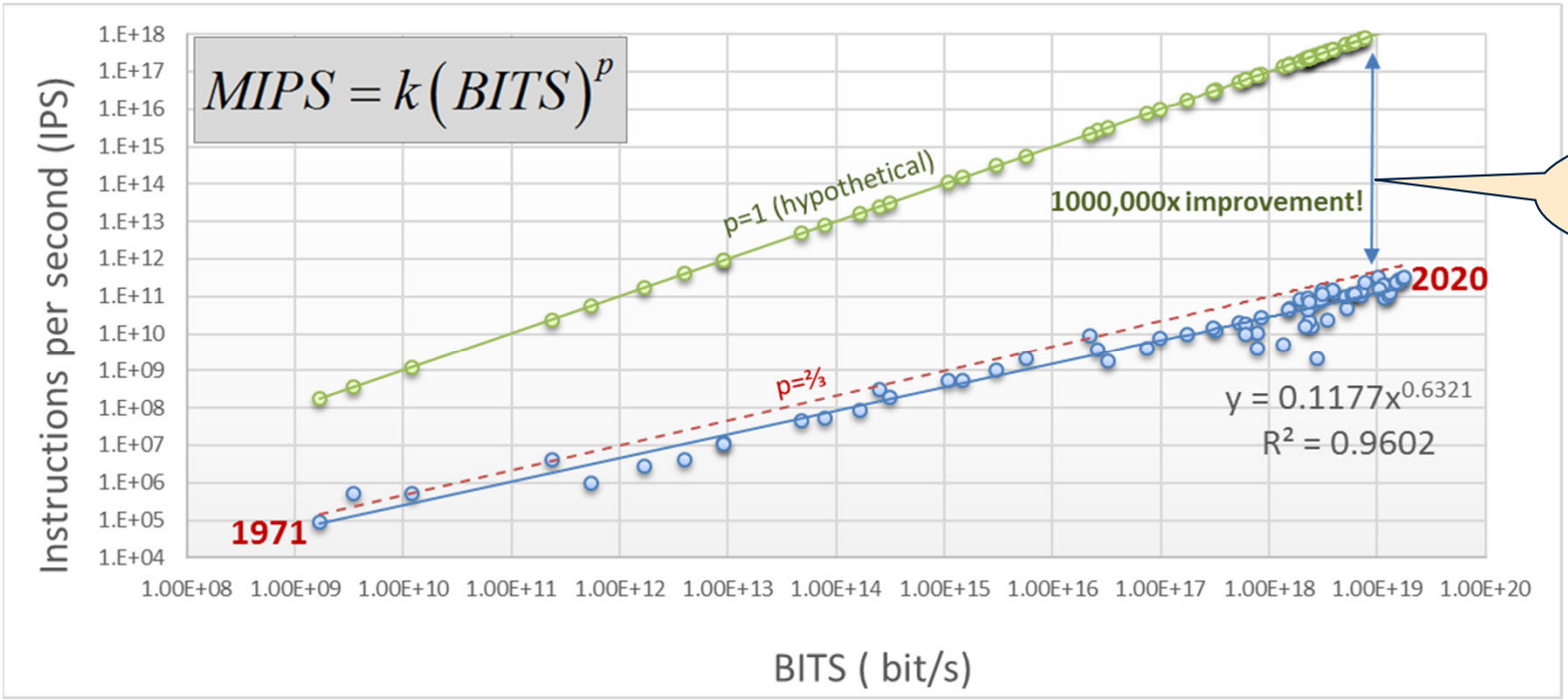
Source: SRC Decadal Plan, 2020

Why Seismic Shift?

Computing will not be 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.

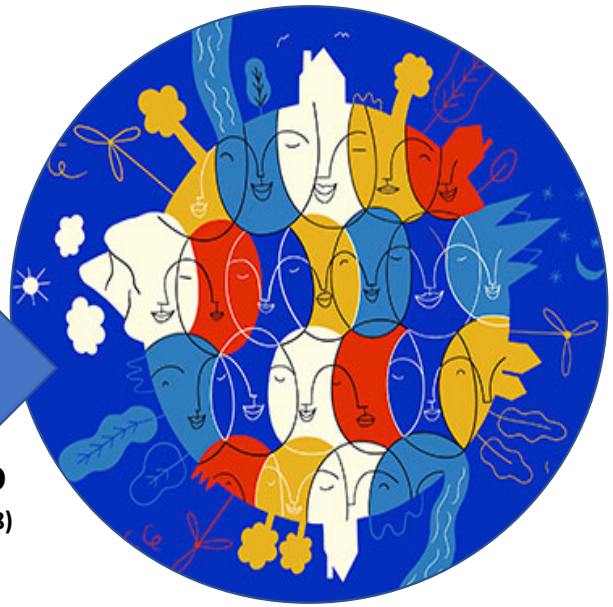
A need to change ‘computational trajectory’



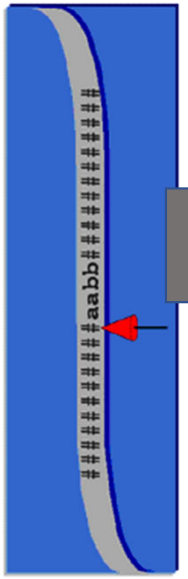
How can we get there?

bit utilization efficiency in computation!

Economic and Social Well-being



Computing Instructions



Computer Applications

Binary Transitions

BITS

10^{35} bit

(2018)

Joules

BTU

MIPS

FLOPS

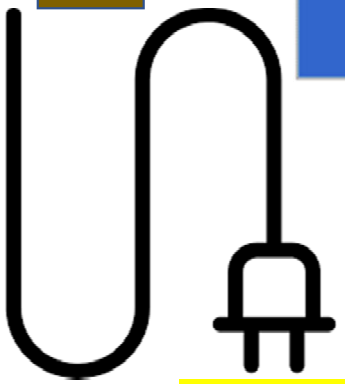
$\$\$$

4×10^{12} USD

(2018)

World's GDP (2018): 85,690 B$=$8.57\times10^{13}$ USD

World's Energy use (2018): 6×10^{20} J $=4\times10^{14}$ kW-h



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

(2018)

ICT Economy

Global ICT 'E-Economy'

- Total bits produced
  - 10^{35} bits
- Total energy
  - 10^{18} J=3× 10^{11} kW-h
- Average $/kW-h
  - 0.13 $
- <u>Total compute energy cost</u>
  - **3.6× 10^{10} USD**

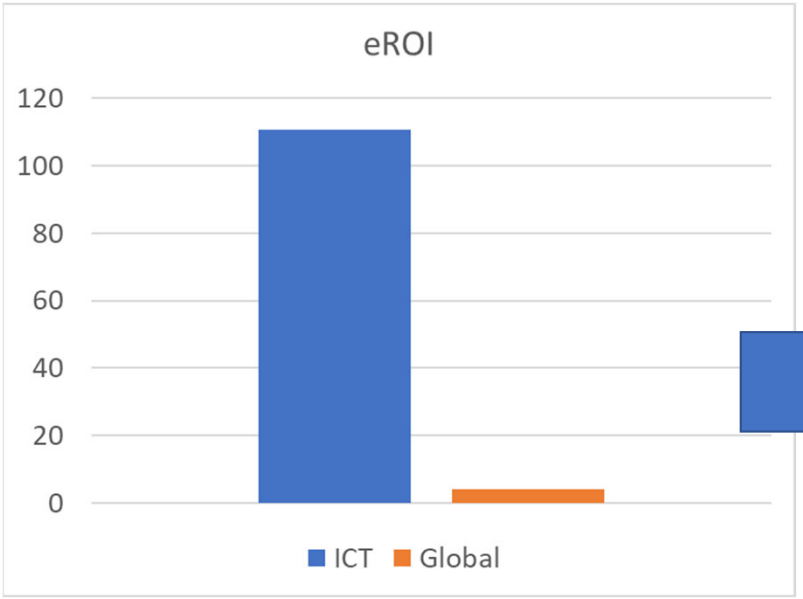
Global ICT "E-ROI": $\frac{4 \cdot 10^{12}}{3.6 \cdot 10^{10}} \approx 100$

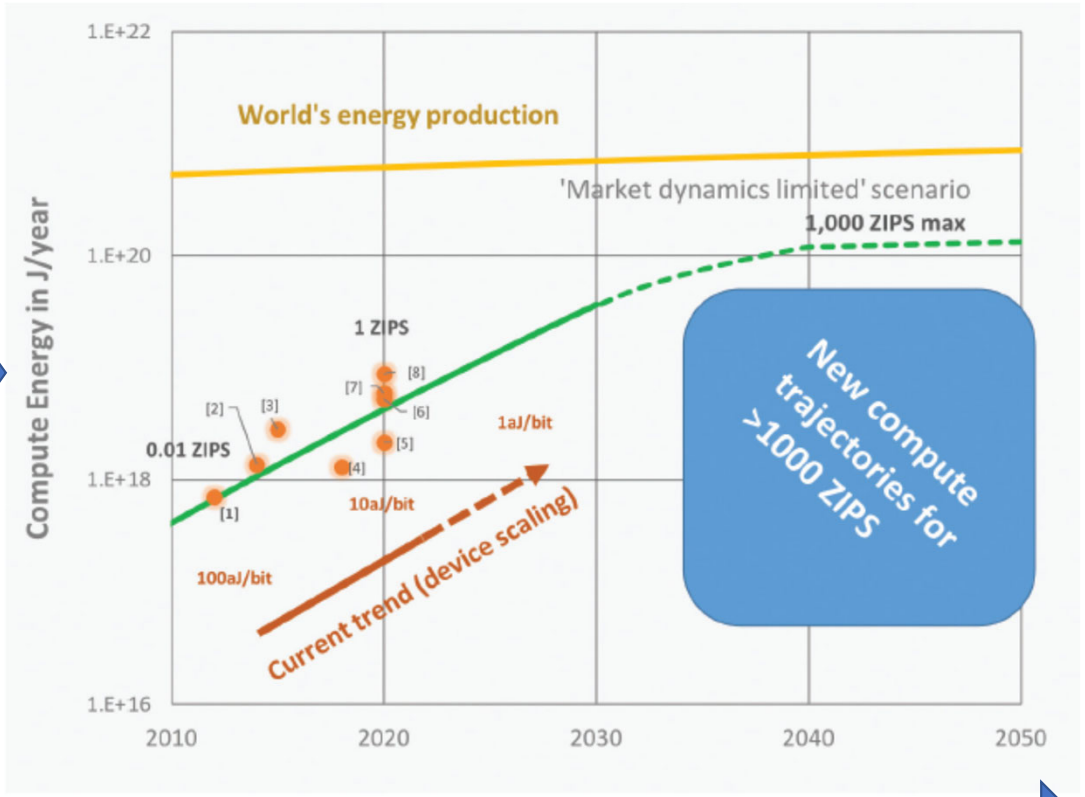
Global 'E-Economy'

World's GDP:	\$85,690 B=8.57×10^{13} USD
World's Energy use:	6×10^{20} J =4×10^{14} kW-h
World's Energy cost:	\$21,700 B=2.17×10^{13} USD

Global "E-ROI": $\frac{8.57 \cdot 10^{13}}{2.17 \cdot 10^{13}} \approx 4$

High ROI drives accelerated growth



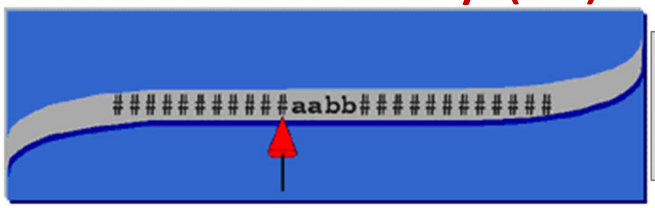


- Total bits produced
  - 10^{35} bits
- Total energy
  - **10^{18} J=3× 10^{11} kW-h** **4.11E+17 J (chipmaking)**
- Average $/kW-h
  - 0.13 $
- Total compute energy cost
  - **3.6× 10^{10} USD**

Discover new compute trajectories

Needed: Theory of Computation

Memory (M)



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

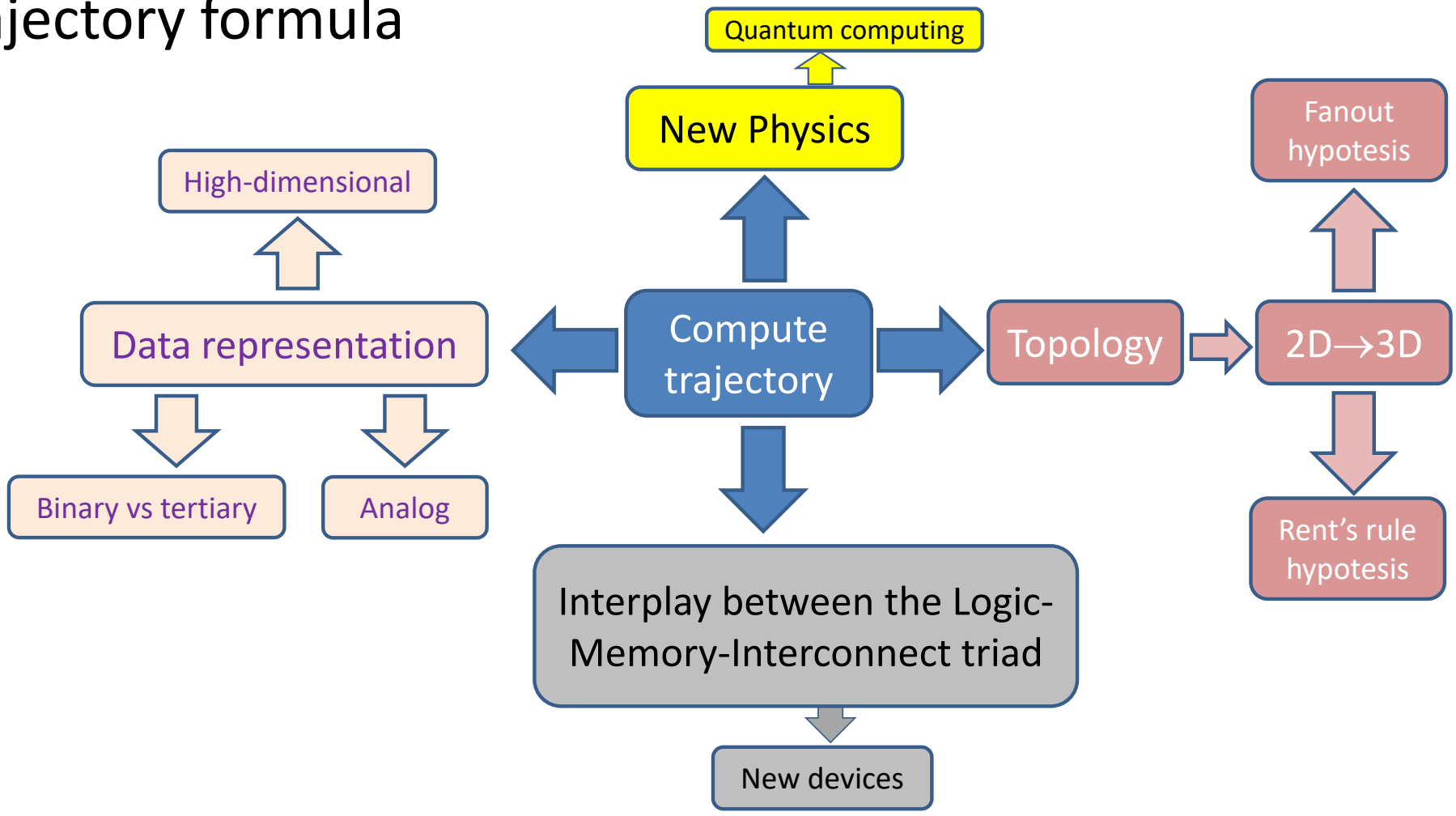
Existing trajectory: $p \approx \frac{2}{3}$

?

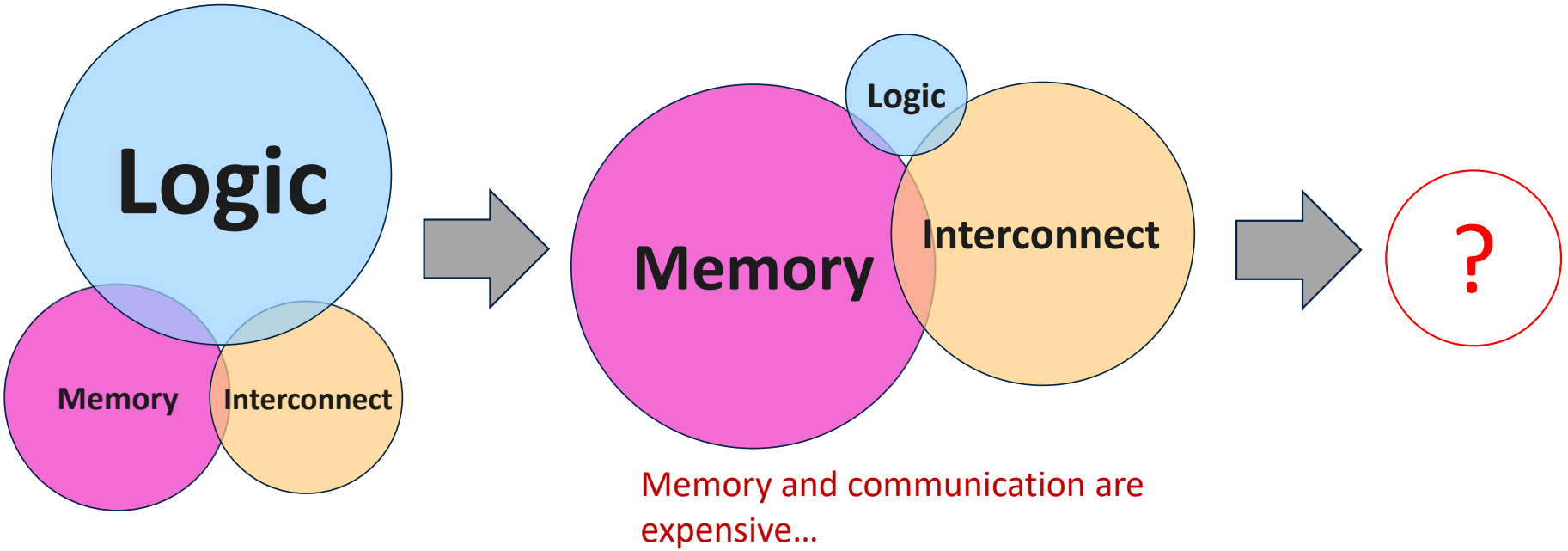
The theoretical basis for performance measurements for computers is much less solid than the theoretical basis for information storage and communication (e.g. Shannon limit etc.)

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

https://www.semiconductors.org/events/webinardecadal-plan-for-semiconductors-new-compute-trajectories-for-energy-efficiency/



Three Cornerstones of Computing



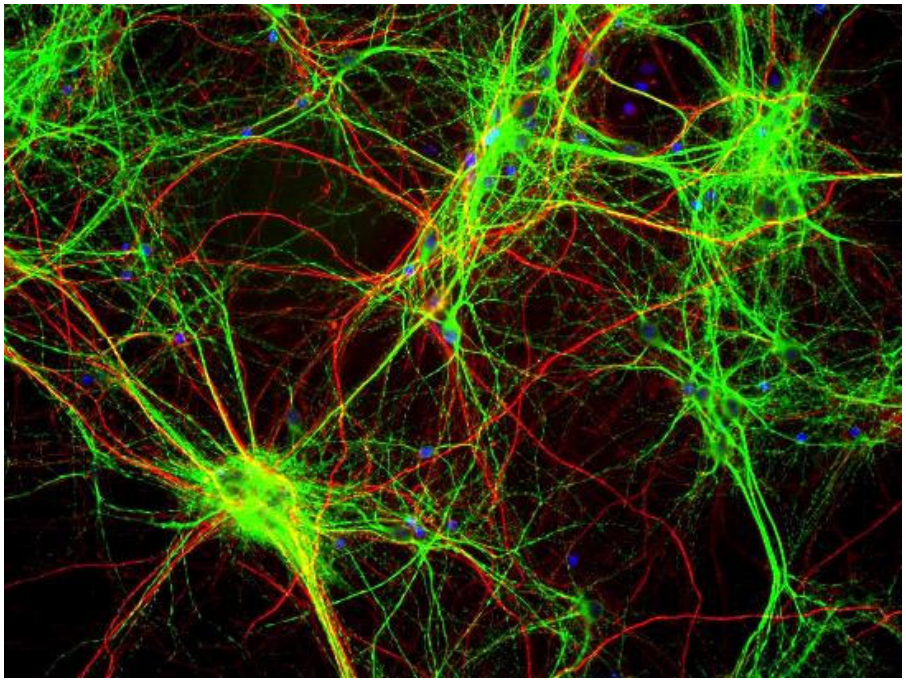
Memory and communication are expensive…

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

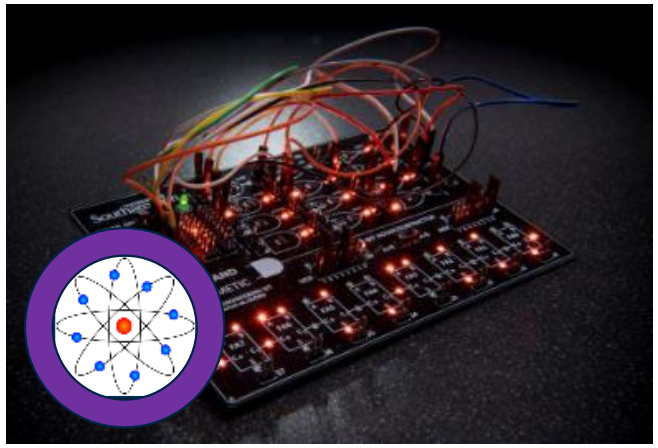
S. L. Smith et al, "Dendritic spikes enhance stimulus selectivity in cortical neurons in vivo", Nature 503 (2013) 115

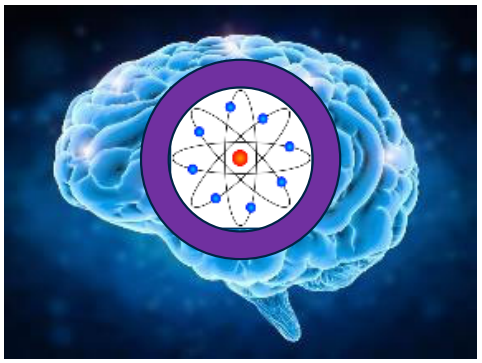
C. Koch, "Computation and single neuron", Nature 385 (1997) 207

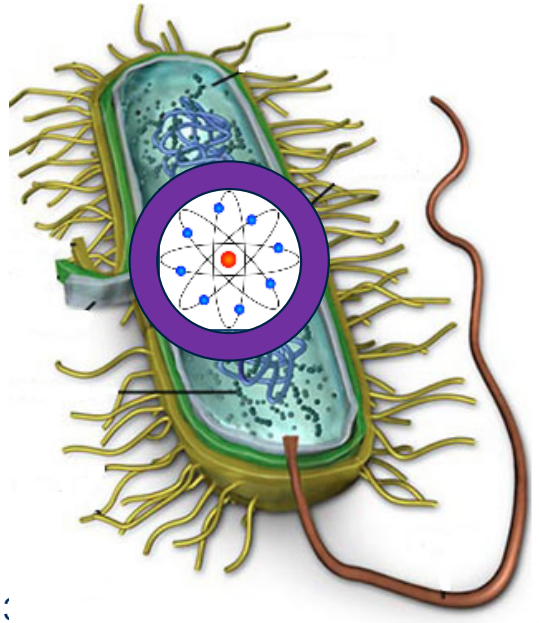
Physics of Information

- Quantum mechanics $\Delta x \Delta p \geq \hbar$
- Statistical physics $\Pi = \exp(-\frac{E_b}{k_B T})$
- Thermodynamics $\Delta E = T\Delta S + P\Delta V + \sum_i \mu_i \Delta N_i$

"Living Materials"

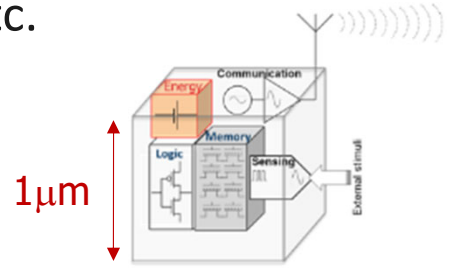






Living Cell as a General-Purpose Processor

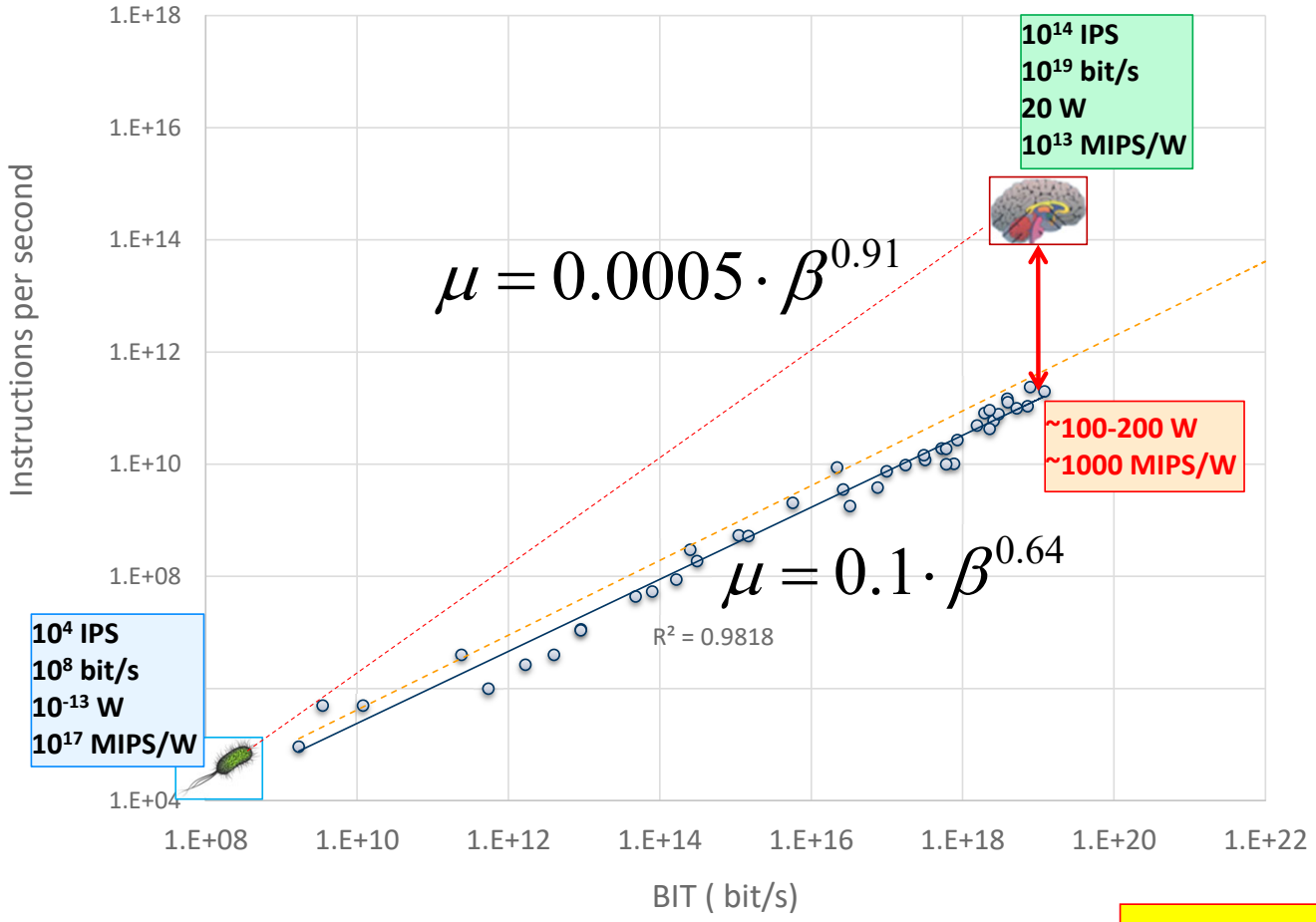
- Single-cell living organisms, such as bacteria, have the formal attributes of a Turing Machine, i.e. a machine expressing a program.
  - Cell can be re-programmed!
- In fact, the cell can be thought of as von Neumann's Universal Constructor, as the cell *expresses the output of its information processing on the matter* constituting the building blocks of the cell itself
  - *computer making computers*.
- In addition, single-cell organisms exhibit the ability to learn, to communicate with each other, various complex social behavior, etc.



"Living Materials"

Antoin Danchin, Bacteria as computers making computers, FEMS Microbiol. Rev. 33 (2009) 3

Computations vs. binary transitions



Estimates of computational power of human brain:

Binary information throughput:

β ~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

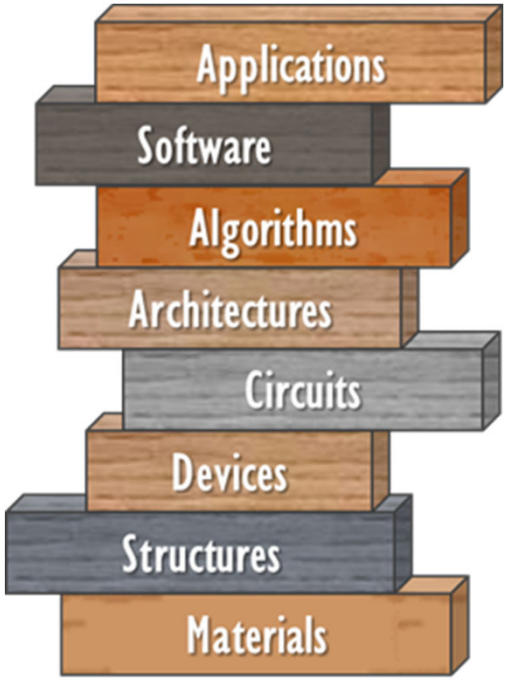
μ ~ 10^8 MIPS

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!

Co-design Challenges and the Decadal Plan for Semiconductors



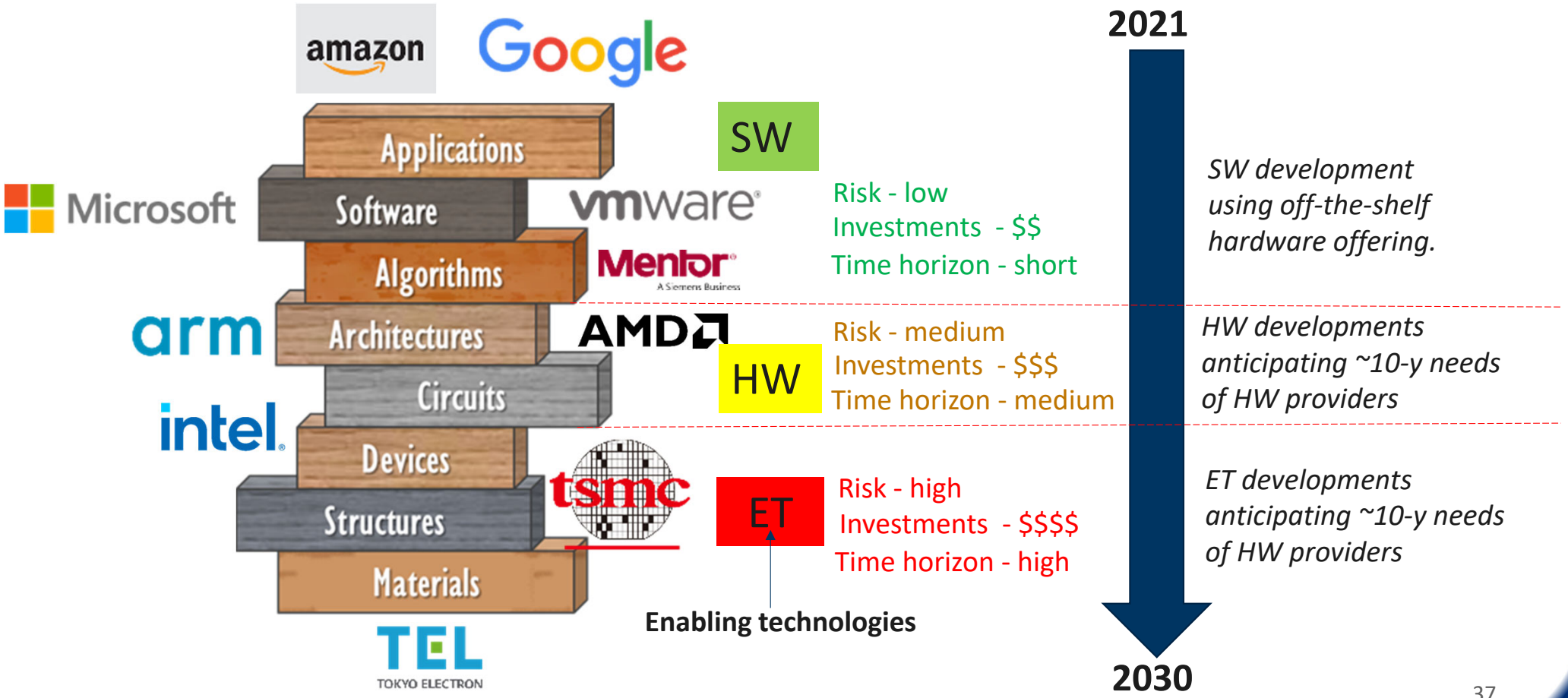
Needed: *True codesign optimization* across all layers from materials to applications. A number of emerging *codesign challenges* anticipated over the next decade are outlined in the 2030 Decadal Plan for Semiconductors

https://www.src.org/about/decadal-plan/

The biggest challenge for the future ICT systems is the **absence of unified codesign framework**, with **most current codesign effort being ad hoc**, and **task-specific**. Different organizations have different definitions of what 'codesign' is

- *for semiconductor companies the codesign occurs mainly on device-to-circuits level for emerging products,*
- *IT companies usually consider HW/SW codesign using off-the-shelf hardware offerings.*

CoDesign 2030 Challenge #2: Different risk models and time horizons between SW, HW and ET providers



Critical role of Government

Government investment in enabling technologies and codesign infrastructure as risk mitigation for long-term HW

Critical role of Universities

Fundamental research on disruptive technologies, to guide strategic investment

5363#Ghfdgdo#Sodq#0Wlphdqh#dqg#Qh{w#Vwhsv

Mdq0534<

Odxqfk##
Ixqgdphqwdo#
Vwxglhv#Ehjlq

*Ghfdgdo#Sodq#
h{hfxwlyh#frpplwwhh#
irupdwlrq#dqg#
zhhnd|#
whohfrqihuhqfhv#
+qrz#rq#&;3,*

Rfw0534<

8#Wrslfdo
Zrunvkrsv#rq#
Vhlvplf#Vkliw

*Rfw#4<#0#Rfw#53
Wkdqnv#GRH$*

Rfw05353

Iqwhulp#Uhsruw##
Zhelqdu

*Fdoo0wr0Dfwlrq#
Rxwdgh/#
H{hfxwlyh#Vxppdu|/#
dqg#Jrdov*

*Ghf#5qg#VID0VUF#
Zhelqdu*

Mdq05354

Ixoo#Uhsruw

*£458#sdjhv/#
ixoo#uhihuhqfhv*

*Loozvwudwlrq#ri#
Wuhqgv/#Nh|#
Fkdoohqjhv/#
Surplvlqj#
Whfkqrorjlhv*

5354

Rxwuhdfk
Hgxfdwlrq
Sxeolf#Zhelqdu

Phpehu#Hyhqw
Vrolflwdwlrqv

Gulyh#Qhz#
Uhvhdufk

53556.

Vxppdul}h
Frpplwwhh#Uhylhz
Uhiuhvk#wkh#Uhsruw

Rxwuhdfk
Sxeolf#Zhelqdu
Phpehu#Hyhqw

Gulyh#Qhz#
Uhvhdufk

2030

Decadal Plan

Five Seismic Shifts

- Smart Sensing
- Memory & Storage
- Communication
- Security
- Energy Efficiency

Dec 2nd SIA-SRC Webinar

Key Research Areas

Qryho#Pdwhuldov

6G#Khwhurjhqhrxv#Lqwhjudwlrq

Dgydqfhg#Sdfndjlqj/#Lqfo1#Glvdjjuhjdwlrq

Lqwhjudwhg#Skrwrqlfv

Frpsxwh#Hiilflhqf|/#Lqfo1#DL#)#Txdqwxp

Wkh#Phpru|#)#Vwrudjh#Sdudgljp

Frppxqlfdwlrqv

Hgjh#Lqwhooljhqfh

Djloh#dqg#Grpdlq#Vshflilf#Ghvljq

Qryho#Dufklwhfwxuhv#)#Dojrulwkpv

Vhfxulw|/#Sulydf|/#dqg#Wuxvw

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.

Thank You!

Todd Younkin, President and CEO: todd.younkin@src.org

Victor Zhirnov, Chief Scientist: Victor.Zhirnov@src.org