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
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...
“SRC 1.0” = 2D Scaling

1984 SRC 10y goals
1990 Microtech 2000
1992 SIA/SRC NTRS
1998 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

2019 SIA Blueprint

2019 3x increase of federal investments in semiconductors

2030 Decadal Plan

“SRC 2.0”
Our 2030 Decadal Plan for Semiconductors


- 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

- The Analog Data Deluge
- The Growth of Memory and Storage Demands
- Communication Capacity vs. Data Generation
- ICT Security Challenges
- Compute Energy vs. Global Energy Production
Information along with Energy has been the Social-Economic Growth Engine of civilization since its very beginning.
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)

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…

\[ I = \log_2 N \]

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
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**

$$\beta = nf$$

**Power consumption**

$$P = E_{bit}nf$$
Lowest Barrier:

**The Boltzmann constraint**

*Distinguishability* $D$ implies low probability $\Pi$ of spontaneous transitions between two wells (error probability)

$D=\text{max}, \Pi=0 \quad D=0, \Pi=0.5 \ (50\%)$

**Classic distinguishability:**

$$\Pi_{\text{classic}} = \exp\left(-\frac{E_b}{k_BT}\right)$$

**Minimum distinguishable barrier:** $\Pi=0.5$

$$\frac{1}{2} = \exp\left(-\frac{E_b}{k_BT}\right) \quad E_b = kT\ln2$$

Shannon - von Neumann - Landauer limit
Scaling Limits: The Heisenberg Constraint

\[ \Delta x \Delta p \geq \frac{\hbar}{2} \]

\[ \Delta E \Delta t \geq \frac{\hbar}{2} \]

\[ \Delta p = \sqrt{2mE_b} \]

At this size, tunneling will destroy the state.

Minimal time of dynamical evolution of a physical system:

\[ a_{\text{crit}} \sim \frac{\hbar}{\sqrt{2mE_b}} \]

\[ t_{\text{min}} \sim \frac{\hbar}{2E_b} \]

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)

\[ a_{\text{min}} \sim 1.5 \text{ nm} \]
\[ t_{\text{sw}} \sim 40 \text{ fs} \]
\[ E_{\text{bit}} \sim 10^{-21} \text{ J} \]

Binary information throughput

\[ \beta = nf \]
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

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

$$k = 0.1, p = 0.64 \approx \frac{2}{3}$$

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

Instructions per second (IPS)

BITS (bit/s)

$$\beta = \alpha N_{tr} \cdot f$$

$$P = \beta E_{bit}$$
Computations per Year

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

1 ZIPS = $10^{15}$ MIPS

World’s installed computing capacity (MIPS) → Total bits → Total energy of computing

Energy per bit
Computing energy: Energy per bit in CPU

6 \times 10^{-18} \text{J/bit}
Total energy of computing a need to change ‘computational trajectory’


\[ MIPS = k \left( BITS \right)^p \]

Existing trajectory: \( p \approx \frac{1}{2} \)
- 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
Seismic shift #5: Computing energy is not sustainable

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.

Source: SRC Decadal Plan, 2020
A need to change ‘computational trajectory’

How can we get there?

bit utilization efficiency in computation!
Energy

Computing Instructions

Binary Transitions

10^{35} bit

BIT

10^{18} J

~3 \times 10^{11} \text{kW-h}

~10^{15} \text{BTU} = 1 \text{ Quad}

(2018)

MIPS

FLOPS

4 \times 10^{12} \text{ USD}

(2018)

World’s GDP (2018): 85,690 B$ = 8.57 \times 10^{13} \text{ USD}

World’s Energy use (2018): 6 \times 10^{20} J = 4 \times 10^{14} \text{ kW-h}

ICT Economy

Economic and Social Well-being
Global ICT ‘E-Economy’

- Total bits produced
  - $10^{35}$ bits
- Total energy
  - $10^{18} \text{ J} = 3 \times 10^{11} \text{ kW-h}$
- Average $$/kW-h$
  - 0.13 $\$
- Total compute energy cost
  - $3.6 \times 10^{10} \text{ USD}$

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

Global ‘E-Economy’

World’s GDP: $85,690 \text{ B} = 8.57 \times 10^{13} \text{ USD}$

World’s Energy use: $6 \times 10^{20} \text{ J} = 4 \times 10^{14} \text{ kW-h}$

World’s Energy cost: $21,700 \text{ B} = 2.17 \times 10^{13} \text{ USD}$

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 = $3 \times 10^{11}$ kW-h
- Average $$/kW\cdot h$
  - 0.13 $\$
- Total compute energy cost
  - $3.6 \times 10^{10}$ USD

- Discover new compute trajectories for $>1000$ ZIPS

*4.11E+17 J (chipmaking)*
Needed: Theory of Computation

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

[Diagram showing the interplay between data representation, high-dimensional computing, new physics, quantum computing, compute trajectory, topology, 2D→3D transformation, Rent’s rule hypothesis, fanout hypothesis, and new devices.]

For more information, visit: [https://www.semiconductors.org/events/webinardecadal-plan-for-semiconductors-new-compute-trajectories-for-energy-efficiency/](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
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.

Physics of Information

- Quantum mechanics
  \[ \Delta x \Delta p \geq \hbar \]

- Statistical physics
  \[ \Pi = \exp\left(-\frac{E_b}{k_B T}\right) \]

- 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.

Antoin Danchin, Bacteria as computers making computers, FEMS Microbiol. Rev. 33 (2009) 3
**Computations vs. binary transitions**

\[ \mu = 0.0005 \cdot \beta^{0.91} \]

**Estimates of computational power of human brain:**

**Binary information throughput:**

\[ \beta \approx 10^{19} \text{ bit/s} \]


(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**

\[ \mu \approx 10^8 \text{ MIPS} \]


(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.

- **SW**
  - Risk: low
  - Investments: $$$
  - Time horizon: short

- **HW**
  - Risk: medium
  - Investments: $$$$ (anticipating ~10-y needs of HW providers)
  - Time horizon: medium

- **ET**
  - Risk: high
  - Investments: $$$$$ (anticipating ~10-y needs of HW providers)
  - Time horizon: high

Enabling technologies:

SW development using off-the-shelf hardware offering.

HW developments anticipating ~10-y needs of HW providers.

ET developments anticipating ~10-y needs of HW providers.
Government investment in enabling technologies and codesign infrastructure as risk mitigation for long-term HW

Fundamental research on disruptive technologies, to guide strategic investment
Decadal Plan for Semiconductors: Setting the 2030 Goals

**Five Seismic Shifts**

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

Dec 2nd SIA-SRC Webinar
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 $p \approx 1$

• 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