Future Information Processing: From Materials to Systems

Victor Zhirnov
Semiconductor Research Corporation

Looking Beyond CMOS Technology for Future HPC

April 5-6, 2016, Hanover, MD 21076
Outline

- Data explosion facts
- Physics of information
- Energetics of computing
- New materials and processes for lower energy computing
- Nanoionic devices for new processor architecture
World’s technological installed capacity to store information

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

Analog World

Digital World

Informational Crust: A major tectonic plate shift
Storage Needs in 2040

A Thought System:
Ultimate Connectivity: Internet of Nanothings

IoT Grand Challenges

I. Giga-Nano-Tera (Billions of Nanosystems connected in a THz-network)

II. Exa-DataCenters: Semiconductor Technologies for Big Data
(Radically new energy-efficient technologies for storing and analyzing massive volumes of data)
What About Global Processing Capability Trends?


How can we increase MIPS?
Benchmark capability $\mu$ (IPS) as a function of $\beta$ (bit/s)

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

$R^2 = 0.9818$

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

$\beta$, bit/s

a measure of computational capability on device level
Turing-Heisenberg Rapprochement?

Instructions per second
a measure of computational capability on the processor level

$\mu = k \beta^p$

Switching time

Binary Information Throughput
a measure of computational capability on device level

Number of binary elements

How can we increase MIPS?

Alan Turing

Werner Heisenberg

Ludwig Boltzmann

Can computational theory suggest new devices?
Stan Williams @ Nanomorphic Forum
MPU: Benchmark capability $\mu$ (IPS) as a function of $\beta$ (bit/s)

$P = N \cdot f \cdot E_{bit} = \beta \cdot E_{bit}$

Power is the main issue for further scaling of high-performance computing

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

$R^2 = 0.9818$

Limits of Cooling

~100-200 W per chip

Power is the main issue for further scaling of high-performance computing

Limits of Cooling

$N_{tr} \cdot f$
What About Global Processing Capability Trends?

Projected Capability

Projected Binary Throughput

Energy


Instructions per second
a measure of computational capability on the processor level

\[ \mu = k \beta^p \]

“Turing-Heisenberg Rapprochement”

Binary Information Throughput
\[ \text{#}T/\text{Chip} \times \text{clock freq.} \]
a measure of computational capability on device level

Energy

\[ E_{\text{bit}} \sim 10^{-14} \]
Total energy of computing

World's energy production

Benchmark (system) - $10^{-14}$ J/bit

Joule/year

2010 2015 2020 2025 2030 2035 2040 2045
What is Information?

Information is measure of distinguishability

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

\[ I = K \ln N \]

\[ N_{\text{min}} = 2 \]

\[ I(N_{\text{min}}) = 1 \]

\[ 1 = K \ln 2 \]

A THEME: Minimal ICT Element

What is the smallest volume of matter needed for an ICT element? What is the smallest energy of operation?
Central Concept: **Energy Barrier**

How can a barrier be created and controlled in a physical system?
Lowest Barrier: 

**The Boltzmann constraint**

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

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

*Classic distinguishability:*

$$\Pi_{\text{classic}} = \exp(-\frac{E_b}{k_B T})$$

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

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

Shannon - von Neumann - Landauer limit

Thermal Noise
Scaling Limits: The Heisenberg Constraint

\[ E_b = k_B T \ln 2 \]

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

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

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

At this size, tunneling will destroy the state

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

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

\[ m^* = 0.15 - 0.20m_0 \]

\[ \sim 3-4 \text{ nm} \]

\[ \sim 1.5 \text{ nm} \quad (m = m_0) \]

\[ \sim 40 \text{ fs} \]

Minimal time of dynamical evolution of a physical system

CMOS scaling on track to obtain physical limits for electron devices

Prof. Mark Lundstrom/Purdue:

Why do we still operate so far above the fundamental limit: Why $10^5$-$10^7 \ k_B T \ln 2$ and not $k_B T \ln 2$?
Switching Energy: Energy of Full-cycle

\[ E_{\text{OFF-ON}} = E_b \]

\[ E_{\text{ON-OFF}} = E_b \]

\[ E_{\text{bit min}} = 2E_{b \text{ min}} + E_{\text{carrier}} \]

\[ kT \ln 2 \]

\[ E_{\text{SW min}} = 3k_B T \ln 2 \times N \]

\[ N - \text{the number of electrons} \]

\[ E_{\text{SW}} = 2E_b + NE_w = (N+2)k_B T \ln 2 \]
Binary switch and Interconnect abstraction: *Extended Well Model*

\[ L_{\text{min}} = 2a \cdot F \]
Connecting Binary Switches via Wires in 2D ($L > 2na$, $N$ electrons)

For logic operation, a binary switch needs to control at least two other binary switches.

\[
\Pi_{C&D} = \Pi_C \times \Pi_D = \left(1 - \left(1 - \frac{a}{L}\right)^N\right)^2
\]

$N_{\text{min}} = 5$

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\Pi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>0.19</td>
</tr>
<tr>
<td>3</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>0.47</td>
</tr>
<tr>
<td>5</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Minimum switching energy for connected binary switches

\[ E_{sw} = 2E_b + NE_b = (N+2)E_b \]

\( n = 2 \)
\( L = 4a \)

\( N_{min} = 5 \)

\( E_{sw} = 7k_B T \ln 2 \)

\( n = 4 \)
\( L = 8a \)

\( N_{min} = 14 \)

\( E_{sw} = 16k_B T \ln 2 \)

Communication between logic switches takes more energy than information processing (switch operations)
Operational reliability vs. Number of Electrons

- In interconnects, the number of electrons needs to be sufficient to guarantee successful communication between binary switches.

**Typical fan out (n=4) for logic**

$L = 8a$

**We need many electrons for reliable communication**

<table>
<thead>
<tr>
<th>N electrons</th>
<th>Operational reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>50%</td>
</tr>
<tr>
<td>20</td>
<td>75%</td>
</tr>
<tr>
<td>42</td>
<td>99%</td>
</tr>
</tbody>
</table>

$E \sim N \cdot E_b = N \cdot e \cdot V_{dd}$

$E \sim 42 \cdot 1.6 \cdot 10^{-19} \cdot 0.5 = 3 \cdot 10^{-18} J = 800k_BT$
Energy costs for fan-out: 2D vs. 3D

More Fan-Out (Branching) = More Computation

\[ E \approx 3 \times 10^{-19} J = 80k_B T \]

~ 10x energy reduction for FO4-6
Key Messages

- Topology optimization for energy reduction

- 1D structures could be enabling!
  - (Quasi) 1D (e.g. nanowires, CNT) components arranged in 3D structures
    - reduce fan-out costs
Energetics of computing: Logic

\[ y = 5 \times 10^{-5} x^{2.2629} \]
\[ R^2 = 0.9619 \]

Technology node, nm

1.00 - 1000.00

Pentium
Pentium Pro
PowerPC 750
Pentium III
Pentium IV
VIA C7
Athlon FX-57
PA6T-1682M
Athlon FX-69
Core i7920
PS3 Cell BE
Core i7 875K
Core i7 2600 K
I7-5960X
FX-8150
Ivy Bridge-EX-15

Top-down

\[ 10^{-16} \text{ J} \]
\[ 10^{-17} \text{ J} \]
A typical computer system

- CPU
  - 1st level cache
    - SRAM
    - SRAM
  - 2nd level cache
    - SRAM
    - DRAM
  - Main memory
    - Permanent Storage
  - SSD, HDD

**Density**
- Low
- Medium
- High

**Capacity**
- Low
- Medium
- High

**Speed**
- Low
- Medium
- High

**Volatile/Non-Volatile**
- Volatile
- Non-Volatile
Minimal Electronic Memory

\[ I_{o-b} = \frac{e}{\hbar} \cdot k_B T \cdot \exp \left( - \frac{E_b}{k_B T} \right) \]

\[ t_{o-b} = \frac{\hbar}{k_B T} \cdot \exp \left( \frac{E_b}{k_B T} \right) \]

\[ E_{b\text{min}} = k_B T \ln \left( \frac{k_B T}{\hbar} \cdot t_s \right) \]

\[ E_{b\text{min}} = 1.3 \text{ eV} \]

\[ t_s = \frac{e}{I_s} \sim 10 \text{ y} \]

\[ I_T = \frac{e}{\hbar} \cdot k_B T \cdot \exp \left( - \frac{2\sqrt{2m}}{\hbar} \cdot a \cdot \sqrt{E_b} \right) \]

\[ a_{\text{min}} = \frac{\hbar}{2\sqrt{2m}E_b} \ln \left( \frac{k_B T}{\hbar} \cdot t_r \right) \]

\[ a_{\text{min}} = 4.30 \text{ nm} \]

(Limited by the mass of electron)

Adjustments: effective mass, electrostatics etc.: \( a_{\text{min}} \sim 5 \text{ nm}, E_{\text{min}} \sim 2-3 \text{ eV} \)
Electron-based Nonvolatile Memory (Flash)

1. Basic Concept

- \( E_{bmin} > 1.7 \text{ eV} (>10 \text{ y retention}) \)
- \( E_b_{SiO_2} = 3.1 \text{ eV} \)
- \( a_{min} \approx 5 \text{ nm} \)

2. WRITE (F-N regime)

- \( eV_{write} > 2E_b \)
- \( >6 \text{ V} \)

3. READ

- \( eV_{read} < 2E_b \)
- \( <6 \text{ V} \)
- \( V_{read} \approx 5 \text{ V} \)
- \( T_{ox} > 10 \text{ nm} \)
- \( F_{min} > 10 \text{ nm} \)

4. Array

- \( C_{line} \approx 10^{-14} \text{ F} \)

- \( E \sim CV^2 \approx 10^{-12} - 10^{-13} J \)

Tunneling is a slow process at lower voltage!
I/O: Communication between an information processing system and the outside world

\[ \Pi_n = \left(1 - \left(1 - \frac{a}{L}\right)^N\right)^n \]

Communication cost per bit per unit length:

\[ C \sim \varepsilon_0 L = 3 \times 10^{-15} \text{ J/(bit\cdot m)} \]

<table>
<thead>
<tr>
<th>L</th>
<th>Joules</th>
<th>Joules</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 (\mu)m</td>
<td>2.86E-19</td>
<td>2.96E-19</td>
</tr>
<tr>
<td>1 mm</td>
<td>2.87E-18</td>
<td>2.96E-18</td>
</tr>
<tr>
<td>1 cm</td>
<td>2.87E-17</td>
<td>2.96E-17</td>
</tr>
<tr>
<td>10 cm</td>
<td>2.87E-16</td>
<td>2.96E-16</td>
</tr>
<tr>
<td>1 m</td>
<td>2.87E-15</td>
<td>2.96E-15</td>
</tr>
</tbody>
</table>

\[ E = \frac{CV^2}{2} \sim \frac{\varepsilon_0 L}{2} \left(\frac{k_B T}{e}\right)^2 \]
Energetics of computing: A Summary

V. Zhirnov, R. Cavin and L. Gammaitoni “Minimum Energy of Computing, Fundamental Considerations” (Chapter 7).

$E_{\text{bit}} \sim 10^{-14} \text{J}$
New Computing System?

\[ E_{\text{bit}} \sim 10^{-14} \text{J} \]

- **CPU**
  - 1st level cache: 10^{-17} J/bit
  - Transistor
  - 10^{-16} J/bit

- **Main memory**: 10^{-11} J/bit
- **2nd & 3rd level cache**: 10^{-11} J/bit

- **SSD/HDD**: 10^{-12} J/bit

- **Permanent storage**
  - Density: high
  - Energy: low
  - Speed: high
  - Non-Volatile
How can we go below 5 nm?

Electron-based Nonvolatile Memory (Flash): $F_{\text{min}} \sim 10\text{nm}$

$$L_{\text{min}} = \frac{\hbar}{\sqrt{2mE_b}}$$

Moving atoms instead of moving electrons?
Moving Atoms: Nanoionics
Resistive Switching by Redox-Based Mechanisms


HfO₂, TiO₂, Ag₂S...

valence change

O₂ ⇌ \frac{1}{2}O_2(g) + V^{**} + 2e'

electrochemical metallization

Atomic switch

ReRAM

<table>
<thead>
<tr>
<th>Feature size</th>
<th>Switching time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best projected</td>
<td>&lt;1 ns</td>
</tr>
<tr>
<td>Demonstrated</td>
<td>&lt;1 ns</td>
</tr>
<tr>
<td>Best demonstrated</td>
<td>&lt;5 nm</td>
</tr>
<tr>
<td>Demonstrated</td>
<td>0.3 ns</td>
</tr>
</tbody>
</table>

Memristor
Flash vs. ReRAM: Energy

Flash

- $e V_{\text{write}} > 2E_b$
- $V_{\text{write}} > 15-20 \ \text{Volt (ms-μs)}$

ReRAM

- $V_{\text{write}} \approx 1 \ \text{V (ns)}$
- $E \sim CV^2$
- $>100\times$ energy reduction
What is the smallest volume of matter needed for a memory cell?


In collaboration with RWTH Aachen Univ / Jülich Res. Ctr.


In collaboration with Micron Technology, Inc.
Discussion Questions

(1) What technology is "most likely to succeed" as the next big thing in HPC

(2) How long before it could be in a system doing something useful in HPC

(3) Key issues that need to be resolved for that technology
In sub-10 nm nanodevices, the ‘defects’ should not be treated as imperfections but instead as controllable entities. Due to these nonstoichiometric defects, the materials often behave as doped semiconductors.

HfO$_2$, TiO$_2$, …

**Chemiconductors** – semiconducting materials, whose stoichiometry can be varied by oxidative and/or reductive valency state changes. This is equivalent to variable doping with defects rather than with foreign species

- Typical example: metal oxides
Semiconductors vs. Chemiconductors

**Semiconductor**  
Si, Ge,…
- the ions forming donor and acceptor levels are due impurities (‘foreign bodies’) introduced into the host matrix
  - e.g. P, B atoms in Si
- The *dopants* (i.e. donors and acceptors) don’t change their positions
  - Increasingly difficult to put them in precise location
- Electrons are the only movable particles
  - Used to represent and sense state
  - Difficult at <10nm
- Rigid interfaces
  - EITHER Ohmic OR Blocking

**Chemiconductor**  
HfO₂, TiO₂,…
- the ions forming donor and acceptor levels are due to composition variation in the host matrix
  - lattice point defects (e.g. vacancies or interstitial atoms) can electrically act as donors or acceptors
  - e.g. ionized oxygen vacancy VO⁺² in TiO₂
  - the ions can move in electrical fields
    - e.g. under external bias
- Atoms and electrons are movable
  - e.g. Atoms – represent state; electrons – sense state
  - Operate at <10nm
- Adaptive interfaces
  - Can switch from Ohmic to Blocking
Metal Oxide ReRAM is Real!

• Panasonic MN101L ReRAM MCU
Energetics of computing: A Summary

Total energy of computing

- World's energy production
- Benchmark (system) - $10^{-14}$ J/bit
- Target (system) - $10^{-17}$ J/bit
- Landauer limit (device) - $3 \times 10^{-21}$ J/bit
Benchmark capability $\mu$ (IPS) as a function of $\beta$ (bit/s)

Estimates of computational power of human brain:

Binary information throughput:

$\beta \sim 10^{19}$ 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

$\mu \sim 10^8$ MIPS


(Estimate made from the analysis brain image processing)

What can we learn about information processing from Nature?

Basic algorithms need to work in very few steps!


1000x algorithmic efficiency

$10^{14}$ IPS $10^{19}$ bit/s $30$ W

$\sim 100-200$ W

$R^2 = 0.9806$

$\sim 500$ “raw” bit transitions per useful bit
Key Messages

- Future computing technologies will require new materials and processes for devices and interconnects e.g.,
  - 1-D structures
  - nanoionic chemiconductors,

- True 3-D integration
  - Topology optimization for energy reduction
    - Quasi 1D (e.g. nanowire) components arranged in 3D structures

- Advances in nanoionics memory could drive the emerging data-centric chip architectures
  - >100× energy reduction
Back-up Slides
Computing system Scaling
(Bell’s Law)

Volume, cm$^3$

1.00E+09 1.00E+07 1.00E+05 1.00E+03 1.00E+01 1.00E-01 1.00E-03 1.00E-05 1.00E-07 1.00E-09 1.00E-11 1.00E-13


IBM 709
IBM 5150
PDP-8
Compaq 1020
Intel 'Edison'

Smart Phone
Notebook
Wearable

Mainframe
PC

R$^2 = 0.9907$
In 1959, Richard Feynman suggested the possibility of building computers whose dimensions were ‘submicroscopic’. These submicroscopic computers remain outside of our grasp.
Minimal Turing Machine

- **2-4 DEC**: 144
- **2-bit Counter**: 12
- **Program Counter**: 24
- **Memory**: ~500 “raw” bit transitions per useful bit
- **Total Devices**: 320
- **Total “Raw” Transitions**: ~500 per useful bit
Miniaturizing Limits of a Computer?

A complete ‘computer’ must contain both logic unit and ‘tape’ (nonvolatile memory)

Si-μCell: A hypothetical 1μm³ Si computer

Binary Information Throughput

\[ N_{tr} = 320, \]
\[ t_{cycle} \approx 170\text{ns} \]
\[ \beta \approx 2 \times 10^9 \text{ bit/s} \]

Algorithmic performance

\[ \mu \approx 0.15 \text{ MIPS} \]

\[ \mu = k \beta^p \]

System’s “intelligence” limit for electron based systems?

~10^{11} \text{ MIPS/cm}^3
The formation of the electrical double-layer at metal-electrolyte interfaces is governed by similar physics as that for the depletion layer metal-semiconductor interfaces.

\[ W \approx 5-10 \text{nm} \]

\[ W \approx 0.5-1 \text{ nm} \]
Ions in liquid electrolytes play an important role in biological information processors such as the brain.

In the human brain, the distribution of Ca ions in dendrites may represent 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.

Based on the brain analogy, the binary state can be realized by a single ion that can be moved to one of two defined positions, separated by a membrane (the barrier) with voltage-controlled conductance.

Ions are heavy, but brain seems to use them efficiently!

Enabling Technology: Fluid nanoelectronics?
(Stuart Parkin, IBM)

• Fluid nanoelectronics utilizing liquid media may offer a new promising path to replace the foundation of today’s computing technologies

• Examples include nanoionic devices based on electrolyte-filled nanochannels
  – DNA memory
  – protonic transistors, etc.

• Such structures might be used to make an atom-based binary switch scalable to ~1 nm or below
  – Fluid nanoelectronic systems could be reconfigurable, with individual elements strung together to create wires and circuits that could be reprogrammed

• Although it is at a very early stage, fluid nanoelectronics could someday allow for very powerful and energy-efficient computing
Summarizing, what we have learned so far from fundamental physics

1) Minimum energy per binary transition

\[ E_{\text{bit}}^{\text{min}} = k_B T \ln 2 \]

2) Minimum distance between two distinguishable states

\[ \Delta x \Delta p \geq \hbar \]

\[ x_{\text{min}} = a = \frac{\hbar}{2\sqrt{2mkT} \ln 2} \approx 1.5\text{nm} \]

3) Minimum state switching time

\[ \Delta E \Delta t \geq \hbar \]

\[ t_{\text{sw}} = \frac{\hbar}{kT \ln 2} \approx 4 \times 10^{-14} \text{s} \]

4) Maximum 2D gate density:

\[ n = \frac{1}{x_{\text{min}}^2} \approx 5 \times 10^{13} \frac{\text{device}}{\text{cm}^2} \]

3 \times 10^{-21} \text{J}
Total Power Dissipation
(@\(E_{\text{bit}} = kT\ln(2)\))

\[
P_{\text{chip}} = \frac{n \cdot E_{\text{bit}}}{t} = 5 \cdot 10^{14} [cm^{-2}] \cdot \frac{3 \cdot 10^{-21} [J]}{4 \cdot 10^{-14} [s]}
\]

\[
E_{\text{bit}} = k_B T \ln 2 \approx 3 \cdot 10^{-21} J
\]

\[
P_{\text{chip}} \approx 4 \times 10^6 \frac{W}{cm^2}
\]

Limits of Cooling?

<table>
<thead>
<tr>
<th>Cooling method</th>
<th>W/cm(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free convection, air</td>
<td>0.25</td>
</tr>
<tr>
<td>Free convection, water</td>
<td>1</td>
</tr>
<tr>
<td>Forced convection, air</td>
<td>5</td>
</tr>
<tr>
<td>Forced convection, water</td>
<td>150</td>
</tr>
</tbody>
</table>

The circuit would vaporize when it is turned on!

6000 W/cm\(^2\)

Sun
Lattice point defects in nonstoichiometric metals oxides (e.g. vacancies or interstitial atoms) can electrically act as donors or acceptors.

Due to these nonstoichiometric defect levels, the materials often behave as doped semiconductors.

Interface controlled resistance (ICR)
Compaction effects due to system scaling can provide significant energy reduction.
MIPS as an indicator of hardware performance

P. Hilbert and M. Lopez, “The world’s technological capacity to store, communicate and compute information”, Science 332 (2011) 60

• Measuring the computational power of digital devices is not an easy task
  • Their characteristics are multidimensional

• The final choice of our unit of measurements was determined principally by the availability of relevant and consistent statistics, more than their robustness

• Dhrystone VAX MIPS as units of measurements
  • Extensive databases available
  • Conversion from SPEC to Dhrystone MIPS established
  • Conversion from FLOPS to Dhrystone MIPS established

A quote from anonymous scientist working at the frontiers of nanoelectronics:

“Heat removal is not an issue. Simply, engineers must invent better technologies for heat removal and cooling.”

“Carnot constraint” on cryogenic operation

\[ W_{\text{cool}} = \frac{T_a - T_c}{T_c} Q^- \]

Work to be done

Heat to be removed

Min. total power needed to run a 100 W chip:

at 77 K - 300 W
at 4.2 K - 7 kW
Due to tunneling, the power consumed by the device depends on both operating temperature and size that manifests itself with unexpectedly dramatic increases in total power consumption at cryogenic temperatures.

\[ E_{\text{bit}}^{\text{total}} = E_{\text{bit}} + \frac{T_a - T_{\text{dev}}}{T_{\text{dev}}} \]

\[ E_{\text{bit}} = \frac{T_a}{T_{\text{dev}}} \]

\[ E_{\text{bit}}^{\text{total}} = T_a \ln 2 + \frac{\hbar^2 (\ln 2)^2}{8ma^2} > k_B T_a \ln 2 \]

\[ T_{\text{dev}} \]