Fundamentals for Energy consumption in ICT devices

Victor Zhirnov
Semiconductor Research Corporation

NiPS Summer School 2014

Perugia, Italy, July 14-18, 2014
The question of the limiting energetics of ICT systems is open

Limiting energy projections for many electronic components are needed to comprehend system scaling limits, e.g.
- Logic, Memory/Storage, Communication, Energy Sources etc.

Physics for ICT energetics
- Energy Source - Avogadro’s Law
- Logic and Memory - Boltzmann-Heisenberg relations
- Communication - Einstein relation
- Computation – Turing Machine
A Thought System: Ultimate Connectivity: Internet of Nanothings

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 is Information?

Information is measure of distinguishability

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

Information-bearing particles

\[ I = K \ln N \]

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

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

\[ 1 = K \ln 2 \]

\[ K = \frac{1}{\ln 2} \]

A THEME: Minimal ICT Element

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

How can a barrier be created and controlled in a physical system?
‘Energy cell’ – a generic abstraction

For electronic ICT-energy technology, the universal principle of operation is the creation and management of charge separation.

**a) Charge separation**

‘electromotive force’ – e.m.f.

**b) Conversion**

**c) Storage**

To prevent charge recombination a barrier is needed to keep the opposite charges apart.

**Source of electrical energy**
The Origins of Charge-Based ICT - Energy elements

Luigi Galvani (1737-1798)
University of Bologna
Discovered the electrical effect of two dissimilar metals in contact with electrolyte

University of Pavia

Charge Separation!

Inspired by Galvani, Alessandro Volta built the first battery
The galvanic cell consumes \textit{atomic fuel} to produce electricity

\[ \varepsilon \approx 1 \text{eV/atom} \sim \text{(chemical bonding energy)} \]

\[ E = \varepsilon \cdot N \]

Number of atoms in cathode electrode

The energy output is limited by the Avogadro’s number, \( N_A \)

\[ E_{\text{max}} \sim eN_A \cdot 1V = 1.6 \times 10^{-19} \cdot 6 \cdot 10^{23} \sim 10^5 \frac{J}{\text{mole}} \sim 10^4 \frac{J}{\text{cm}^3} \]

\[ E \sim (10^{-4} \text{ cm})^3 \cdot 10^4 \sim 10^{-8} J \]
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
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:**

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

Minimum distinguishable barrier: $\Pi=0.5$

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

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

Quantum Mechanical Tunneling

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

$\Delta x \geq a$

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

$a \sqrt{2mE_b} \leq \frac{\hbar}{2}$

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

At this size, tunneling will destroy the state

$1 - \frac{2\sqrt{2m}}{\hbar} (a \sqrt{E_b}) \geq 0$

$1 - x \approx e^{-x}$

$\exp \left( - \frac{2\sqrt{2m}}{\hbar} (a \sqrt{E_b}) \right) \geq 0$

$\Pi_{quant} = \exp \left( - \frac{2\sqrt{2m}}{\hbar} (a \sqrt{E_b}) \right)$

Wentzel-Kramers-Brillouin (WKB) approximation
Example: Quantum Resistance

Heisenberg’s Energy-time relation

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

Plank’s constant

\[ h = 6.62 \times 10^{-34} \text{ Js} \]

Minimal time of dynamical evolution of a physical system

Quantum Resistance

Single -electron Conductance channel (mode)

\[ \Delta V = IR \]

\[ \Delta V = \frac{\Delta E}{e} \]

\[ I = \frac{e}{\Delta t} \]

Local Event – no coordinate change

\[ \frac{\Delta E}{e} = \frac{e}{\Delta t} R \]

\[ R = \frac{\Delta E \Delta t}{e^2} = \frac{h}{2e^2} = \]

\[ = \frac{6.64 \times 10^{-34} \text{ J} \cdot \text{s}}{2 \times (1.6 \times 10^{-19} \text{ C})^2} = 12.9 \text{k}\Omega \]
Summary on Quantum resistance

Heisenberg’s Energy-time relation:
\[ \Delta E \Delta t \geq \frac{\hbar}{2} \]

Ohm’s Law:
\[ V = IR \]

von Klitzing constant:
\[ R_0 = \frac{h}{2e^2} = 12.9 \text{k}\Omega \]
\[ G_0 = \frac{1}{R_0} = \frac{2e^2}{h} \]

It was experimentally discovered in the 1980s in Quantum Hall Experiments.

Nobel Prize in 1985

Landauer formula:
\[ I = n_{chan} \times G_0 \cdot V \cdot \Pi \]
Summarizing, what we have learned so far from fundamental physics

1) Minimum energy per binary transition
   \[ E_{bit}^{\text{min}} = k_B T \ln 2 \]
   \[ = 3 \times 10^{-21} \text{ J} \]

2) Minimum distance between two distinguishable states
   \[ \Delta x \Delta p \geq \frac{\hbar}{2} \]
   \[ x_{\text{min}} = a = \frac{\hbar}{2\sqrt{2mkT \ln 2}} \sim 1 \text{ nm} \]

3) Minimum state switching time
   \[ \Delta E \Delta t \geq \frac{\hbar}{2} \]
   \[ t_{sw} = \frac{\hbar}{2kT \ln 2} \sim 10^{-13} \text{ s} \]

4) Maximum 2D gate density:
   \[ n = \frac{1}{x_{\text{min}}^2} \approx 10^{14} \frac{\text{device}}{\text{cm}^2} \]
Total Power Dissipation
(@$E_{\text{bit}} = kT\ln(2)$)

\[ P_{\text{chip}} = \frac{n \cdot E_{\text{bit}}}{t} = 10^{14} \text{[cm}^{-2}] \cdot \frac{3 \cdot 10^{-21} \text{[J]}}{10^{-13} \text{[s]}} \]

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

\[ P_{\text{chip}} \sim 3 \times 10^6 \frac{W}{cm^2} \]

The circuit would vaporize when it is turned on!

### Limits of Cooling?

<table>
<thead>
<tr>
<th>Cooling method</th>
<th>W/cm²</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>

6000 W/cm² Sun
Nanoscale Devices

"Boltzmann constraint" on minimum switching energy

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

\[ \sim 10^{-21} \text{ J} \]

This structure cannot be used for representation/processing information

"Heisenberg constraint" on minimum device size

\[ x_{\text{min}} = \frac{\hbar}{2\sqrt{2mkT \ln 2}} \]

\[ \sim 1 \text{ nm} \]

\[ \Pi_{\text{error}} = \exp\left(\frac{E_b}{k_B T}\right) \]

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

An energy barrier is needed to preserve a binary state
By doping, it is possible to create a built-in field and energy barriers within semiconductor.

1. Metal-Insulator-Metal stack

The height of these barriers cannot be changed.

2. \textit{pn}-junction

The height of these barriers can be changed.

- Acceptor ions (e.g. B\textsuperscript{+})
- Donor ions (e.g. P)

- Flash memory
- Diode
- Triode
Energy dissipation in binary transitions: Example I Vacuum Tubes

\[ I = e \cdot f_0 \exp \left( \frac{E_b}{kT} \right) \]

\[ E_b \approx 4-5\text{eV} \approx 200k_B T \]

\[ E_{k\text{min}} > E_b \]

\[ E_b = mgH \]
Ideal rectangular barrier (abrupt walls)

\[ E_{b0} \]

V=0

\[ d_t = a \]
Ideal rectangular barrier (abrupt walls)

\[ E_b = E_{b0} \]

\[ d_t = a \]
Ideal rectangular barrier (abrupt walls)

\[ eV = E_b \]

\[ E_b = E_{b0} \]

\[ d_t = a \]
Ideal rectangular barrier (abrupt walls)

$E_b > E_b$

$E_b = E_{b0}$

$d_t < a$

$d_t$
Barriers in electronic ICT: A Summary

1. Metal-Insulator-Metal stack

- Flash memory

The height of these barriers cannot be changed

2. \(pn\)-junction

- Acceptor ions (e.g. \(B^+\))
- Donor ions (e.g. \(P\))

By doping, it is possible to create a built-in field and energy barriers within semiconductor

- Transistor

The height of these barriers can be changed
Barrier height control in a semiconductor system

\[ E_{b0} = E_g - k_B T \left( \ln \frac{N_V}{N_a} + \ln \frac{N_C}{N_d} \right) \]
Barrier height control in a semiconductor system

\[ E_{b0} = E_g - k_B T \left( \ln \frac{N_V}{N_a} + \ln \frac{N_C}{N_d} \right) \]

\[ E_b \approx E_{b0} - eV_g \]
Fundamental operation of multi-electron binary switch:

\[ N_A = N_0 \exp \left( -\frac{E_b}{k_B T} \right) \]
\[ I_{AB} = e \cdot N_A = eN_0 \exp \left( -\frac{E_b}{k_B T} \right) \]

\[ N_B = N_0 \exp \left( -\frac{E_b + e\Delta V_{AB}}{k_B T} \right) \]

\[ E_b = E_{b0} - eV_g \]

\[ I = eN_0 \exp \left( -\frac{E_{b0}}{k_B T} \right) - eN_0 \exp \left( -\frac{E_b + eV_{AB}}{k_B T} \right) = eN_0 \exp \left( -\frac{E_b}{k_B T} \right) \left[ 1 - \exp \left( -\frac{eV_{AB}}{k_B T} \right) \right] \]

\[ I = eN_0 \exp \left( -\frac{E_{b0} - eV_g}{k_B T} \right) \left[ 1 - \exp \left( -\frac{eV_{AB}}{k_B T} \right) \right] \]
FET Equation

\[ I_{ds} = I_0 \exp \left( \frac{eV_g - eV_t}{k_B T} \right) \left[ 1 - \exp \left( - \frac{eV_{ds}}{k_B T} \right) \right] \]

\[ S = \frac{\Delta V}{\text{dec}(\Delta I)} \]

\[ \frac{I_2}{I_1} = \frac{I_0 \exp \left( - \frac{E_{b_2}}{k_B T} \right)}{I_0 \exp \left( - \frac{E_{b_1}}{k_B T} \right)} = \exp \left( - \frac{E_{b_2} - E_{b_1}}{k_B T} \right) = \exp \left( - \frac{\Delta E_b}{k_B T} \right) = 10 \]

\[ S = \frac{k_B T \ln 10}{e} = 60 \frac{mV}{\text{dec}} \]

\[ eV_g \geq E_b \]

\[ \frac{I_2}{I_1} = 2 \]

\[ S = \frac{k_B T \ln 2}{e} \]
Barrier Height Control in Charge Transport Devices

Conditional Change of State

Poisson’s equation:

\[ \nabla^2 \varphi = -\frac{\rho}{\varepsilon_0} \]

Changes in the barrier height require changes in charge density/distribution

\[ C = \frac{\Delta q}{\Delta \varphi} \]

Operation of ALL charge transport devices includes charging/discharging capacitances to change barrier height controlling charge transport

- FET
- SRAM, DRAM, flash
- RTD, SET…

Energy to “deform” the barrier is equivalent to the energy of charging the control (gate) capacitor

\[ E_{dis} = \frac{C_g V^2}{2} \]
Energy dissipated by charging of a capacitor

By charging a capacitor to the energy $E = CV^2/2$ from a constant voltage power supply, an equal amount of energy $(CV^2/2)$ is also dissipated.
Energy dissipated by discharging of a capacitor

Before:

\[
E_{\text{before}} = \frac{CV^2}{2}
\]

\(E_{\text{before}}\) is the sum of energy stored in two capacitors) before closing the switch

After (*):

\(C^* = 2C\)

\(q^* = q = CV = C^*V^* = 2CV^*\)

\(V^* = \frac{V}{2}\)

\[
E_{\text{after}} = \frac{C^*V^*^2}{2} = \frac{2C \cdot \left( \frac{V}{2} \right)^2}{2} = \frac{CV^2}{4} = \frac{E_{\text{before}}}{2}
\]

\(E_{\text{after}}\) is the sum of energy stored in two capacitors) after closing the switch
CMOS scaling on track to obtain physical limits for electron devices

George Bourianoff / Intel

Prof. Mark Lundstrom/Purdue:

Why do we still operate so far above the fundamental limit: Why $10^4 k_B T \ln 2$ and not $k_B T \ln 2$?

Answer:
1) System reliability costs
2) Communication costs
3) Fan-Out costs

$E \sim N \cdot E_b = N \cdot e \cdot V_{dd}$
System Level Energetics I: Reliable Switching

Computation at $\Pi_{err}=0.5$, and hence at $E_b=k_B T \ln 2$ is impossible.

In useful computation, $\Pi_{err} \ll 0.5$, hence barrier height larger than $k_B \ln 2$ is needed (larger total power consumption).

**Question: How Much Larger?**

$$\Pi_{syst} = (1 - \Pi_{err})^N$$

The probability that all $N$ switches in a circuit work correctly.

($N \uparrow \rightarrow L \downarrow \rightarrow \Pi_{err} \uparrow$) (Heisenberg)

($\Pi_{err} \downarrow \rightarrow E \uparrow$) (Boltzmann&Heisenberg)
System Constraint on Minimum Energy per Bit

\[ \Pi_{syst} = (1 - \Pi_{err})^N \]

The probability that all N switches in a circuit work correctly

\[ \Pi_{syst} > \Pi_{crit} \quad \text{e.g.,} \quad 0.5 \quad 0.99 \]

a "reasonable" boundary

\[ \Pi_{err} = 1 - \Pi_{crit} \]

\[ \Pi_{err} = f(E_b) \]

\[ E_{b_{min}} = f(N) \]

\[ N_{max} \sim \frac{1}{a^2} \]

Boltzmann

\[ \Pi_{err} = \exp(-\frac{E_b}{kT}) + \exp(-\frac{2\sqrt{2m}}{\hbar} a\sqrt{E_b}) - \exp(-\frac{\hbar E_b + 2akT\sqrt{2mE_b}}{\hbar kT}) \]

Heisenberg
The maximum possible number $N$ of binary switches in a close-packed array is inversely proportional to the square of the barrier length $L$ (e.g. the FET gate length $L_g$)

- $N = f_1(L) \leftrightarrow L = f_2(N)$

- The switching time $t_{sw}$ of an individual switch is directly proportional to $L$

- The minimum barrier height in binary switches and therefore minimum operating voltage is a function of

\[ L_g \sim \frac{1}{\sqrt{20N}} \]

- $L_g$ (device level)
- $N$ (system level)

\[ E_{b_{\text{min}}} = f(N) \]

for $\Pi = \Pi_{\text{crit}}$

<table>
<thead>
<tr>
<th>$L_g$, nm</th>
<th>$N$, cm$^2$</th>
<th>$E_{b_{\text{min}}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>2.50E+07</td>
<td>0.65</td>
</tr>
<tr>
<td>50</td>
<td>1.00E+08</td>
<td>0.67</td>
</tr>
<tr>
<td>30</td>
<td>2.78E+08</td>
<td>0.69</td>
</tr>
<tr>
<td>20</td>
<td>6.25E+08</td>
<td>0.70</td>
</tr>
<tr>
<td>10</td>
<td>2.50E+09</td>
<td>0.71</td>
</tr>
<tr>
<td>9</td>
<td>3.09E+09</td>
<td>0.72</td>
</tr>
<tr>
<td>8</td>
<td>3.91E+09</td>
<td>0.72</td>
</tr>
<tr>
<td>7</td>
<td>5.10E+09</td>
<td>0.73</td>
</tr>
<tr>
<td>6</td>
<td>6.94E+09</td>
<td>0.80</td>
</tr>
<tr>
<td>5</td>
<td>1.00E+10</td>
<td>1.17</td>
</tr>
<tr>
<td>4</td>
<td>1.56E+10</td>
<td>1.90</td>
</tr>
<tr>
<td>3</td>
<td>2.78E+10</td>
<td>3.52</td>
</tr>
</tbody>
</table>
Generic Challenges

- **Energy – Errors dilemma**

Tunneling cannot be ignored for $a<5\text{nm}$, which sets a practical limit.

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

$$ V_{\text{min}} \sim 0.7 \text{ V} $$

$$ \sim \exp\left(-\frac{2\sqrt{2m}}{\hbar}(a\sqrt{E_b})\right) $$

$1\text{cm} \times 1\text{cm}$

$N \sim 10^{10} \text{ cm}^{-2}$
Switching Energy: Energy of Full-cycle

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

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

\[ E_{bit_{min}} = 2E_{b_{min}} + E_{carrier} \]

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

\[ E_{sw} = 2E_b + NE_w = (N+2)k_B T \ln 2 \]

We are fighting ambient thermal energy!
Connecting Binary Switches via Wires: Extended Well Model

The problem is to ‘place’ the electron on the downstream gate – more than one electron is needed to ‘charge’ the line.

Example: $L = 4a$
$N = 1 \rightarrow P < 0.25$

In General:

$$\Pi_{CD} = \frac{a}{L}$$

$$\Pi = 1 - \left(1 - \frac{e}{L}\right)^N$$

$N$ – the number of electrons.

Note: Connecting one binary switch to another one doesn’t yet do computation!
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.

$L > 2na$, $n$- fan

$N$ – the number of electrons

$n=2$

$L=4a$

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

$N_{min} = 5$

<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

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

**FO2**
- \( n=2 \)
- \( L=4a \)
- \( N_{min}=5 \)
- \( E_{sw} = 7k_B T \ln 2 \)

**FO4**
- \( n=4 \)
- \( L=8a \)
- \( N_{min}=14 \)
- \( E_{sw} = 16k_B T \ln 2 \)

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

\[ L = 8a \]

<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>

We need many electrons for reliable communication.
More electrons means more energy...

<table>
<thead>
<tr>
<th>Year</th>
<th>Node</th>
<th>MPU gate</th>
<th>N electron</th>
<th>$E_{\text{bit}}/k_B T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003</td>
<td>100</td>
<td>45</td>
<td>1215</td>
<td>5.63E+04</td>
</tr>
<tr>
<td>2004</td>
<td>90</td>
<td>37</td>
<td>812</td>
<td>3.76E+04</td>
</tr>
<tr>
<td>2005</td>
<td>80</td>
<td>32</td>
<td>532</td>
<td>2.26E+04</td>
</tr>
<tr>
<td>2006</td>
<td>70</td>
<td>28</td>
<td>439</td>
<td>1.87E+04</td>
</tr>
<tr>
<td>2007</td>
<td>65</td>
<td>25</td>
<td>360</td>
<td>1.53E+04</td>
</tr>
<tr>
<td>2008</td>
<td>57</td>
<td>22</td>
<td>331</td>
<td>1.28E+04</td>
</tr>
<tr>
<td>2009</td>
<td>50</td>
<td>20</td>
<td>280</td>
<td>1.08E+04</td>
</tr>
<tr>
<td>2010</td>
<td>45</td>
<td>18</td>
<td>245</td>
<td>9.47E+03</td>
</tr>
<tr>
<td>2012</td>
<td>35</td>
<td>14</td>
<td>155</td>
<td>5.39E+03</td>
</tr>
<tr>
<td>2013</td>
<td>32</td>
<td>13</td>
<td>134</td>
<td>4.66E+03</td>
</tr>
<tr>
<td>2015</td>
<td>25</td>
<td>10</td>
<td>77</td>
<td>2.37E+03</td>
</tr>
<tr>
<td>2016</td>
<td>22</td>
<td>9</td>
<td>69</td>
<td>2.12E+03</td>
</tr>
<tr>
<td>2018</td>
<td>18</td>
<td>7</td>
<td>40</td>
<td>1.07E+03</td>
</tr>
<tr>
<td>2020</td>
<td>14</td>
<td></td>
<td>22</td>
<td>6.05E+02</td>
</tr>
</tbody>
</table>

We need a significant number of electrons for branched communication between binary switches

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

$$E \sim 22 \cdot 1.6 \cdot 10^{-19} \cdot 0.7 = 2.5 \cdot 10^{-18} J = 600 k_B T$$

Mark Lundstrom/Purdue:

Why do we still operate so far above the fundamental limit: Why $10^5 k_B T \ln 2$ and not $k_B T \ln 2$?
Long Interconnects

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

\[ n=2 \quad L=100a \]

<table>
<thead>
<tr>
<th>N electrons</th>
<th>Operational reliability</th>
<th>Energy ( E )</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>50%</td>
<td>(<del>120k_BT</del>)</td>
</tr>
<tr>
<td>198</td>
<td>75%</td>
<td>(<del>200k_BT</del>)</td>
</tr>
<tr>
<td>487</td>
<td>99%</td>
<td>(<del>500k_BT</del>)</td>
</tr>
</tbody>
</table>

\[ \Pi_n = \left(1 - \left(1 - \frac{a}{L}\right)^N\right)^n \]
Communication between an information processing system and the outside world

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

\[ a = 1 \ \mu m \]

\[ C \sim \varepsilon_0 L \]

\[ E_{\text{min}} \sim eNk_B T \]

\[ E = \frac{CV^2}{2} \sim \frac{\varepsilon_0 L}{2} \left(\frac{k_B T}{e}\right)^2 \]

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

Communication cost per bit per unit length:

\[ \sim \frac{\varepsilon_0}{2} \left(\frac{k_B T}{e}\right)^2 \]

\[ = 3 \times 10^{-15} \text{ J/(bit}\cdot\text{m}) \]
Wireless Communication for Micro-Scale Systems

Example: **Uniformly** radiated wireless communication

\[ E_{\text{com}} = N_{\text{photons}} \cdot E_{\text{ph}} \]

\[ E_{\text{ph}} = h \nu = \frac{hc}{\lambda} \]

\[ N_{\text{photons}} \sim \frac{4 \pi r^2}{\lambda^2} \]

\[ E_{\text{com}} \sim \frac{4 \pi r^2 hc}{\lambda^3} \]

\[ E \sim 10^{-8} J \]

Example: \( r = 1 \text{m} \)

\[ \lambda \sim 1 \mu \text{m} \]

\[ E_{\text{com}} = 4 \pi \cdot 1^2 \cdot \frac{6.62 \cdot 10^{-34} \cdot 3 \cdot 10^8}{(4 \cdot 10^{-6})^3} \sim 10^{-9} \frac{J}{\text{bit}} \]

\[ 10 \text{ bit} \]
Scaling of omni-directional wireless is limited due to increased energy costs.

\[ E_{\text{com}} \sim \frac{4\pi r^2}{\lambda^2} \cdot \frac{hc}{\lambda} = \frac{4\pi r^2 hc}{\lambda^3} \]

Experimental data

\[ N_{\text{bit}} = \frac{E_{\text{total}}}{E_{\text{bit}}} \sim 10^6 \frac{10^{-18}}{10^{-18}} \]

\[ r = 10 \text{ m} \]

\[ N_{\text{bit}} = \frac{E_{\text{total}}}{E_{\text{bit}}} \sim 10^{-5} \]

\(~10^2 4\text{ transmitted bits}\)

\(~10\text{ mm cell}\)

\(~100\text{ transmitted bits}\)
New Interconnect paradigm?

The main problem of interconnects is the statistical behavior of discrete charges – *Electrons are free to move along the line*

Thermal & Shot Noise – we need more electrons for reliable branched communication

Eli Yablonovitch/UC Berkeley

Are “deterministic interconnects” possible? e.g. Photons transition point-to-point? Others?

Can we decrease the number of information-bearing particles in communication between binary switches?
Directed transmission

\[ C_{pn} \sim \frac{\epsilon_0 K d^2}{W} \]

\[ \lambda \sim 1 \, \mu m \]

\[ eV_{th} \sim E_{ph} = \frac{hc}{\lambda} \]

\[ E_{LED} = \frac{C_{pn}^2 V_{th}^2}{2} \sim \frac{5 \times 10^{-16} \cdot 1^2}{2} = 2.5 \times 10^{-16} \, J \]
MINIMAL MEMORY ELEMENT
(Nonvolatile case)


What is the smallest volume of matter needed for memory?
Minimal Electronic Memory

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

\[ I = G_0 \cdot V \cdot \Pi = \frac{2e^2}{h} \cdot \frac{k_B T}{2e} \cdot \exp \left( -\frac{2\sqrt{2m}}{\hbar} (a\sqrt{E_b}) \right) \]

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

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

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

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

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

\[ a_{\text{min}} = \frac{\hbar}{2\sqrt{2mE_b}} \ln \frac{k_B T}{h} t_r \]

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

(Limited by the mass of electron)

Adjustments: effective mass, electrostatics etc.: \( a_{\text{min}} \sim 5 \text{ nm}, E_{\text{min}} \sim 2-3 \text{ eV} \)
1. **Basic Concept**

- $E_{b_{min}} > 1.7$ eV (>10 y retention)
- $E_b_{SiO_2} = 3.1$ eV
- $a_{min} \sim 5$ nm

2. **WRITE (F-N regime)**

- $V_{write_{min}} > 6-7$ Volt (very slow)
- $V_{write} > 10-15$ Volt (ms-$\mu$s)

3. **READ**

- $eV_{read} < 2E_b$
- $< 6$ V

- $V_{read} \sim 5$ V
- $T_{ox} > 10$ nm
- $F_{min} > 10$ nm

4. **Array**

- $C_{line} \sim 10^{-14} F$

\[
E \sim C_{line} V^2 \sim 2.5 \cdot 10^{-13} J / line
\]

or $2 \times 10^{-15} J / bit$
MINIMAL COMPUTER

“If one constructs the automaton (A) correctly, then any additional requirements about the automaton can be handled by sufficiently elaborated instructions. This is only true if A is sufficiently complicated, if it has reached a certain minimum of complexity” (J. von Neumann)

**Capability for general-purpose computing?**

| C > 1 | Yes |
| C < 1 | No  |

\[ n_c = ? \]

**Von Neumann threshold**
The minimal ALU does $2^2 = 4$ operations on two 1-bit X and Y:

- Operation 1: X AND Y
- Operation 2: X OR Y
- Operation 3: (X+Y)
- Operation 4: (X+(NOT Y))

Supports functionally complete set of logic and arithmetic operations.
Minimal Turing Machine

- Memory
  - 2-4 DEC
  - 2-bit Counter
  - ~500 “raw” bit transitions per useful bit

Program Counter
- I<sub>1</sub>, I<sub>2</sub>, 24

ALU
- X, Y, C0
- 98

CPU
- Z, C1
- S<sub>1</sub>, S<sub>2</sub>, S<sub>3</sub>, S<sub>4</sub>, S<sub>5</sub>, S<sub>6</sub>

Total: 320 devices
System Constraint on Minimum Energy per Bit

\[ \Pi_{syst} = (1 - \Pi_{err})^N \]

The probability that all \( N \) switches in a circuit work correctly

\[ \Pi_{syst} > \Pi_{crit} \quad \text{e.g.,} \quad 0.5 \quad \text{lower boundary} \]

\[ \Pi_{err} = 1 - \Pi_{crit} = 1 - 2^{-\frac{1}{320}} = 0.002 \]

Boltzmann

\[ \Pi_{err} = \exp\left(-\frac{E_b}{kT}\right) + \exp\left(-\frac{2\sqrt{2m}}{\hbar} a\sqrt{E_b}\right) - \exp\left(-\frac{\hbar E_b + 2akT\sqrt{2mE_b}}{\hbar kT}\right) \]

Heisenberg

\[ E_{b_{\text{min}}} \approx k_B T \ln \frac{1}{\Pi_{err}} \approx 6k_B T \]
Charge based computing: A Summary

Wires connecting binary switches, constitute a dominant portion of the energy consumption in ITC
A difficult problem for continuing scaling: The Power/Heat Barrier

Energy consumption by ICT is growing

New ICT principles for greater energy efficiency need to be discovered

Energy consumption by ICT is growing

Heat removal is a crucial issue for future computing
Benchmark capability $\mu$ (IPS) as a function of $\beta$ (bit/s)

Basic algorithms need to work in very few steps!

$R^2 = 0.98055$

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

1000x algorithmic efficiency

$\sim 100-200$ W

~500 “raw” bit transitions per useful bit

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?
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)
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

Radical Departures from current baseline technologies may be needed to address the exponential growth in the storage needs.
Is there enough silicon to support the major tectonic plate shift in the Informational Crust?
Next-Generation Digital Information Storage in DNA

George M. Church,1,2 Yuan Gao,3 Sriram Kosuri1,2*
1Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. 2Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. 3Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA.
*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu

Researchers stored an entire genetics textbook in less than a picogram of DNA — one trillionth of a gram — an advance that could revolutionize our ability to save data.

5.27×10⁶ bit

DNA memory can be stable ~ 100y+

**HARDWARE**: Agilent Oligo Library Synthesis microarray platform

- Agilent Technologies, a spin-off of Hewlett-Packard (1999), originally a semiconductor company, which became now a global company offering products & services in communications, electronics, semiconductor, test and measurement, life sciences and chemical analysis industries.
- Example of a successful convergence of semiconductor and bio industries
Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

All 154 of Shakespeare’s sonnets and audio clip from Martin Luther King’s famous “I have a dream” speech, were encoded in DNA by a EMBL & Agilent team.

The team projects that, based on the current progress in DNA read and write technologies, this technique could be scaled up to store all of the data in the world.
Given:
| Memory:  | 9.6 Mbit |
| Power:   | $10^{-13}$ W |
| Task time*: | 2400s=40min |

**Simplifying Assumption:**
The entire DNA information content is read and written at least once during one cell division cycle.

**Characteristic access time per bit:**

$$t_{bit} \sim \frac{2400}{2 \cdot 9.6 \cdot 10^6} \sim 100 \mu s$$

**Characteristic energy per bit (system-level):**

$$E < \frac{10^{-13} W \cdot 2400s}{2 \cdot 9.6 \cdot 10^6} = 2.5 \cdot 10^{-17} \frac{J}{bit}$$

**Characteristic energy per bit (system-level):**

$$P_{DNA} < \frac{1.4 \cdot 10^{-13}}{2.4 \cdot 10^{-3}} = 5.8 \cdot 10^{-11} \frac{W}{GByte}$$
DNA-Inspired Memory  
(On-Going Project with Micron Technology)

DNA-inspired memory

- DNA volumetric memory density far exceeds (1000x) projected ultimate electronic memory densities
- Potential for very low-energy memory access
- **Goal:** Demonstrate a miniaturized, on-chip integrated DNA storage

<table>
<thead>
<tr>
<th></th>
<th>HardDiskDrive</th>
<th>NAND flash</th>
<th>DRAM</th>
<th>DNA in cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read/Write latency</td>
<td>3-5 ms/bit</td>
<td>~100μs/bit</td>
<td>&lt;10 ns/bit</td>
<td>&lt;100μs/bit</td>
</tr>
<tr>
<td>Endurance (cycles)</td>
<td>unlimited</td>
<td>10⁴-10⁵</td>
<td>unlimited</td>
<td>unlimited</td>
</tr>
<tr>
<td>Retention</td>
<td>&gt;10 years</td>
<td>~10 years</td>
<td>64 ms</td>
<td>&gt;10 years</td>
</tr>
<tr>
<td>ON power (W/GB)</td>
<td>~0.04</td>
<td>~0.01-0.04</td>
<td>0.4</td>
<td>&lt;10⁻¹¹</td>
</tr>
<tr>
<td>Aerial Density</td>
<td>~10¹¹ bit/cm²</td>
<td>~10¹⁰ bit/cm²</td>
<td>~10⁹ bit/cm²</td>
<td>n/a</td>
</tr>
<tr>
<td>Volumetric Density</td>
<td>n/a</td>
<td>10¹⁶ bit/cm³</td>
<td>~10¹³ bit/cm³</td>
<td>10¹⁹ bit/cm³</td>
</tr>
</tbody>
</table>
Memory Hardware

- All data about structure and operation of a living cell are stored in the long DNA molecule
  - Nonvolatile memory
- DNA coding uses a base-4 (quaternary) system
  - The information is encoded digitally by using four different molecular fragments, to represent a state: adenine (A), cytosine (C), guanine (G), and thymine (T).

DNA memory operations

READ
- Multi-access capability by distinct computing units

WRITE
Vertical gene transfer - exact copying of the parental DNA
Lateral (horizontal) gene transfer:
1. direct uptake (‘swallowing’) of a naked DNA by a cell,
2. by a virus,
3. by direct physical contact between two cells.

Electronic NVM:
- \( F_{\text{min}} \approx 10\text{nm}/1\text{bit} \)

Heavy mass!

\[
\alpha_{\text{min}} = \frac{h}{2\sqrt{2mE_b}}
\]

DNA is NOT a read-only memory
Nature Has Been Processing Information for a Billion Years

Si-μCell

Bio-μCell – A Living Cell
About 500 of these cells would fit in the cross-section of a human hair

V=1μm³

Our studies show that the Si-μCell cannot match the Bio-μCell in the density of memory and logic elements, nor operational speed, nor operational energy:

<table>
<thead>
<tr>
<th></th>
<th>Memory:</th>
<th>Logic:</th>
<th>Power:</th>
<th>Algorithmic efficiency:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si-μCell</td>
<td>1000x more</td>
<td>&gt;10x more</td>
<td>1000,000x less</td>
<td>1000x more</td>
</tr>
<tr>
<td>Bio-μCell</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lower-hanging fruit?
Storage Needs in 2040

10^{10} kg of wafer-grade Si
Projected global supply: ~2 \times 10^7 kg

A bucket of DNA address the BigData storage capacity challenge?
Close to practical limits of current communication technologies.

To be completed in 2016:

Example: Fiber optic cables technology

SEA-ME-WE 5

24 Tbit/s

～2×10^24 bit

3000 years

Big Data Communication Challenge
Conclusions

- ICT and Energy devices have a common soul
  - The universal principle of operation of all these elements is the creation and management of charge separation
  - Controllable energy barriers is a fundamental component in all ICT & Energy devices

- Memory and Communication are the main factors of energy consumption by ICT rather than Logic

- We suggest that inspiration for future ultra-low energy ICT can be derived from organic systems, i.e., at the intersection of chemistry, biology, and information processing
Short-term lessons

- Memory access is the most severe limiting factor of Si-\( \mu \)Cell Computer.
  - not enough nonvolatile memory bits
  - Memory access to support computations takes too much energy
- Organizing solid-state memory in cross-bar arrays, while an elegant solution at larger scale, but it contributes to excessive energy dissipation due to line charging during R/W access.
  - Access to the DNA memory is array-less and can be viewed as similar to access to tape or hard disc drive.
  - Multiple W/R heads for independent access
- Desirable attributes for future memory technology
  - Array-less organization for energy minimization
  - Multiple R/W heads for independent access
  - Moving atoms for ultimate density (\(~ 1\)nm memory elements)
  - Example: the IBM ‘Millipede’