SRC® Semiconductor Research Corporation

PIONEERS IN COLLABORATIVE RESEARCH®

global GRC RESEARCH COLLABORATION

Fundamentals for Energy consumption in ICT devices

Victor Zhirnov

Semiconductor Research Corporation

NiPS Summer School 2014

Perugia, Italy, July 14-18, 2014

A Theme: Physics of Information

- 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

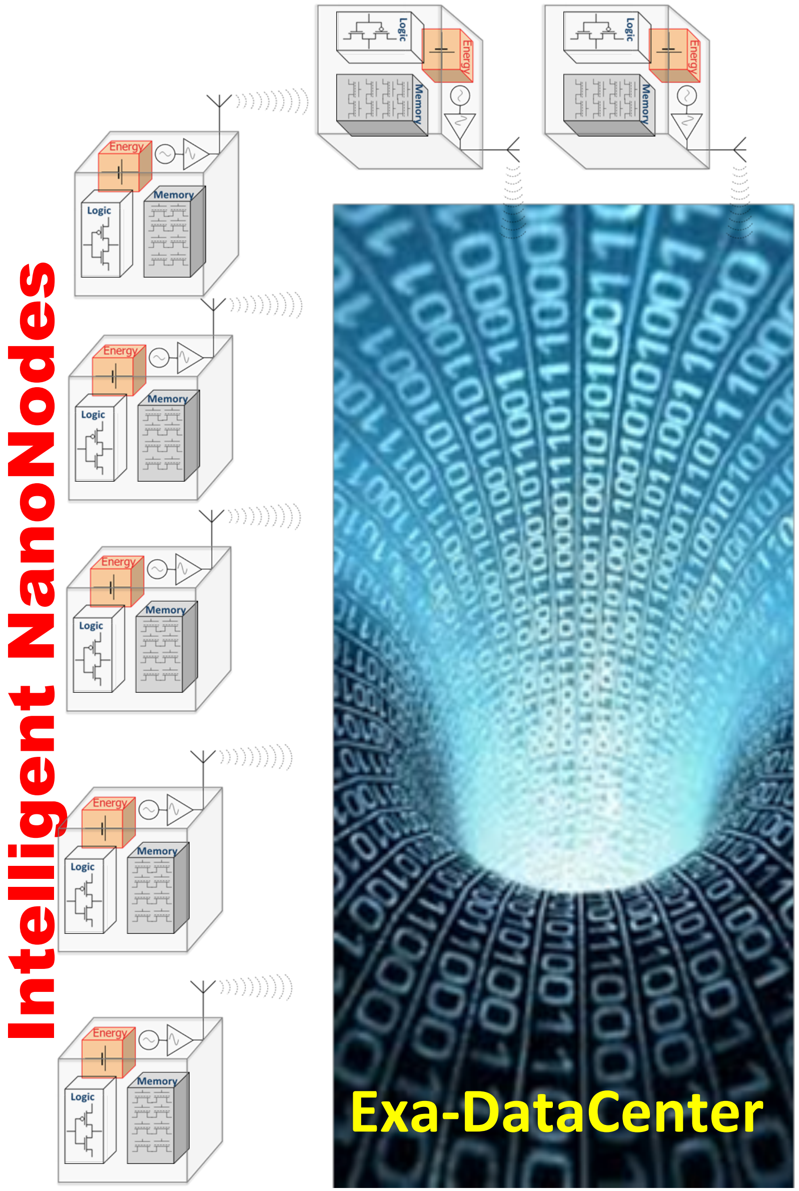
A Thought System:

Ultimate Connectivity: Internet of Nanothings

IoT Grand Challenges

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

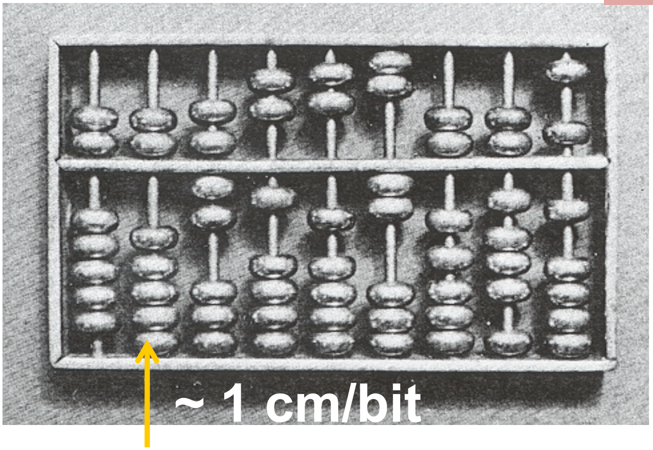
II. Exa-DataCenters: Semiconductor Technologies for Big Data
(Radically new energy-efficient technologies for storing and analyzing massive volumes of data)



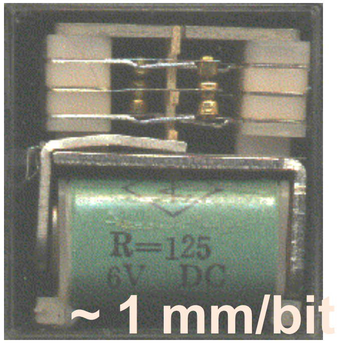
What is Information?

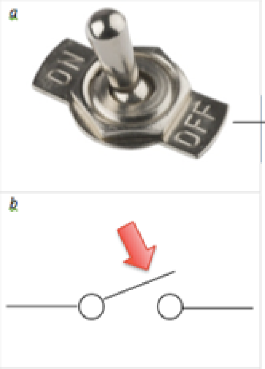
Information is measure of distinguishability

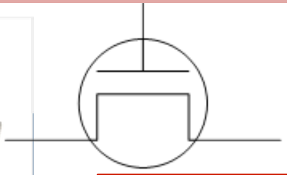
e.g. of a physical subsystem from its environment…



Information-bearing particles







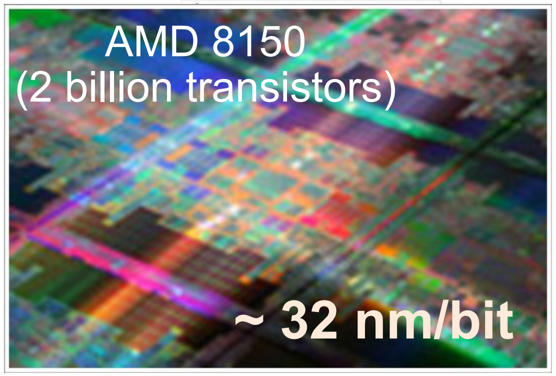
$$I = K \ln N$$

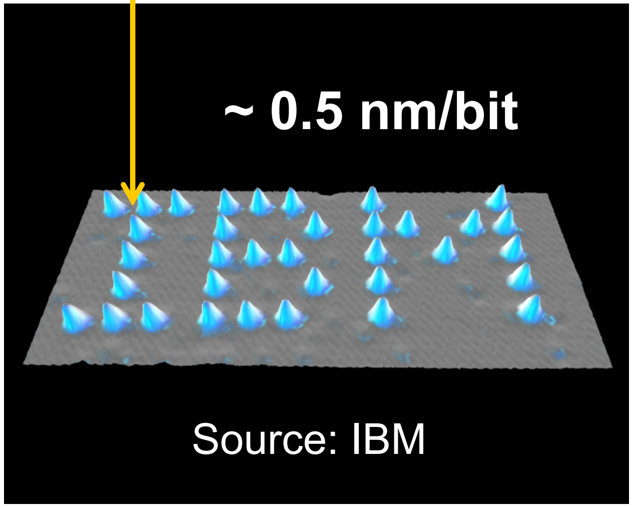
$$N_{\min} = 2$$

$$I(N_{\min}) = 1$$

$$1 = K \ln 2$$

$$K = \frac{1}{\ln 2}$$



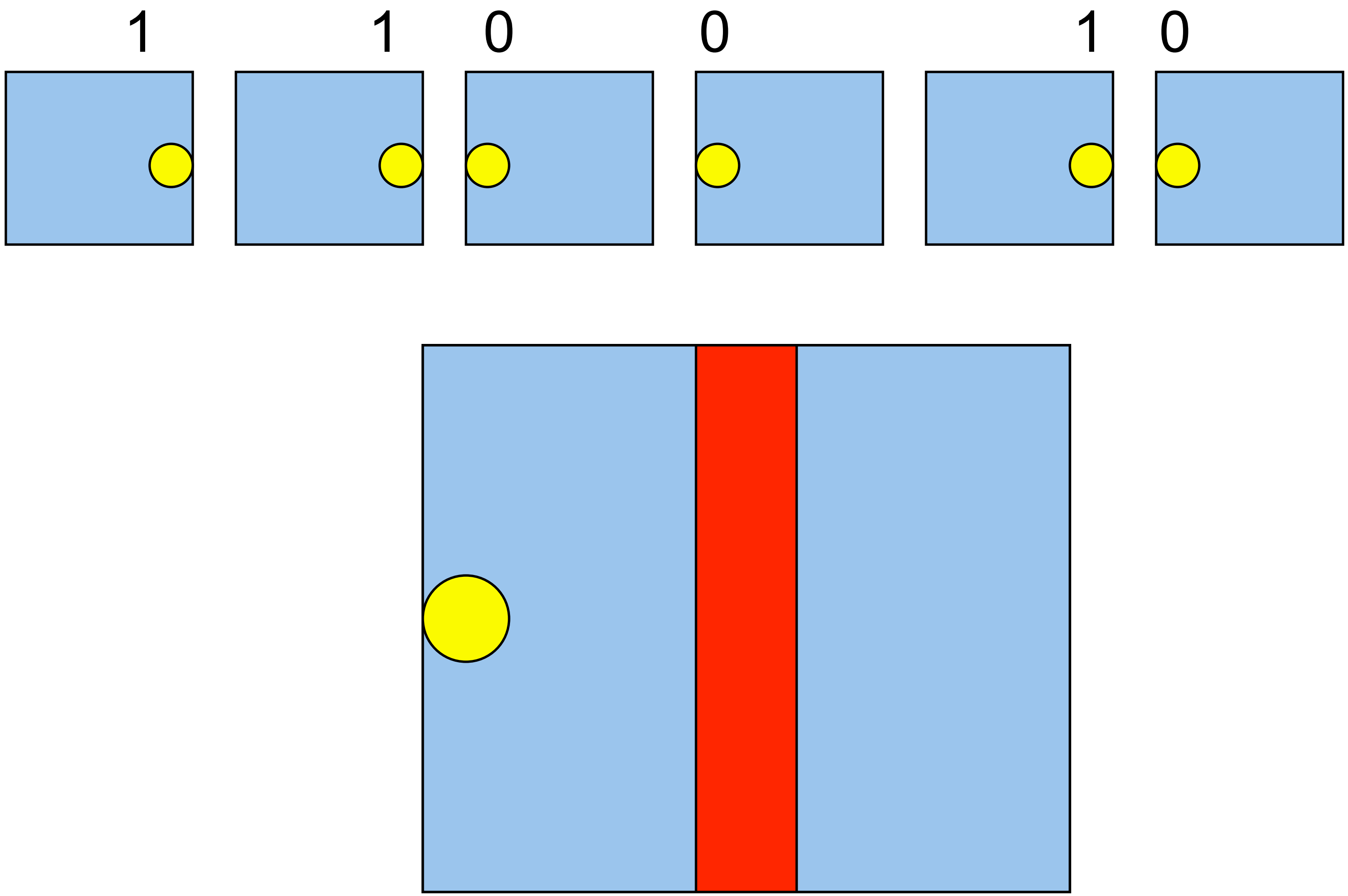


Source: IBM

A THEME: Minimal ICT Element

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

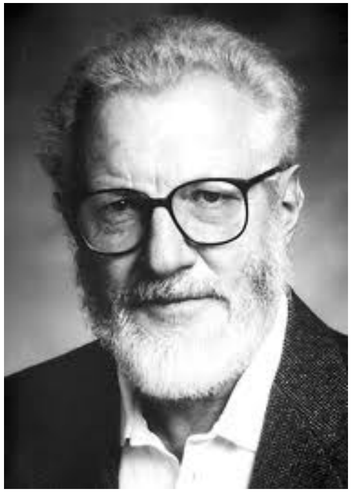
Particle Location is an Indicator of State



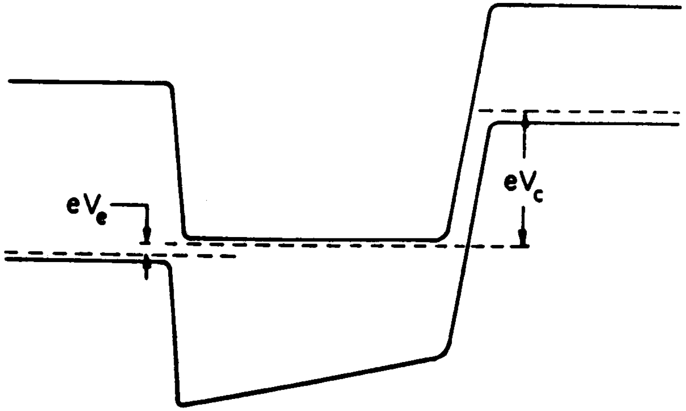
Kroemer's Lemma of Proven Ignorance

- 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*

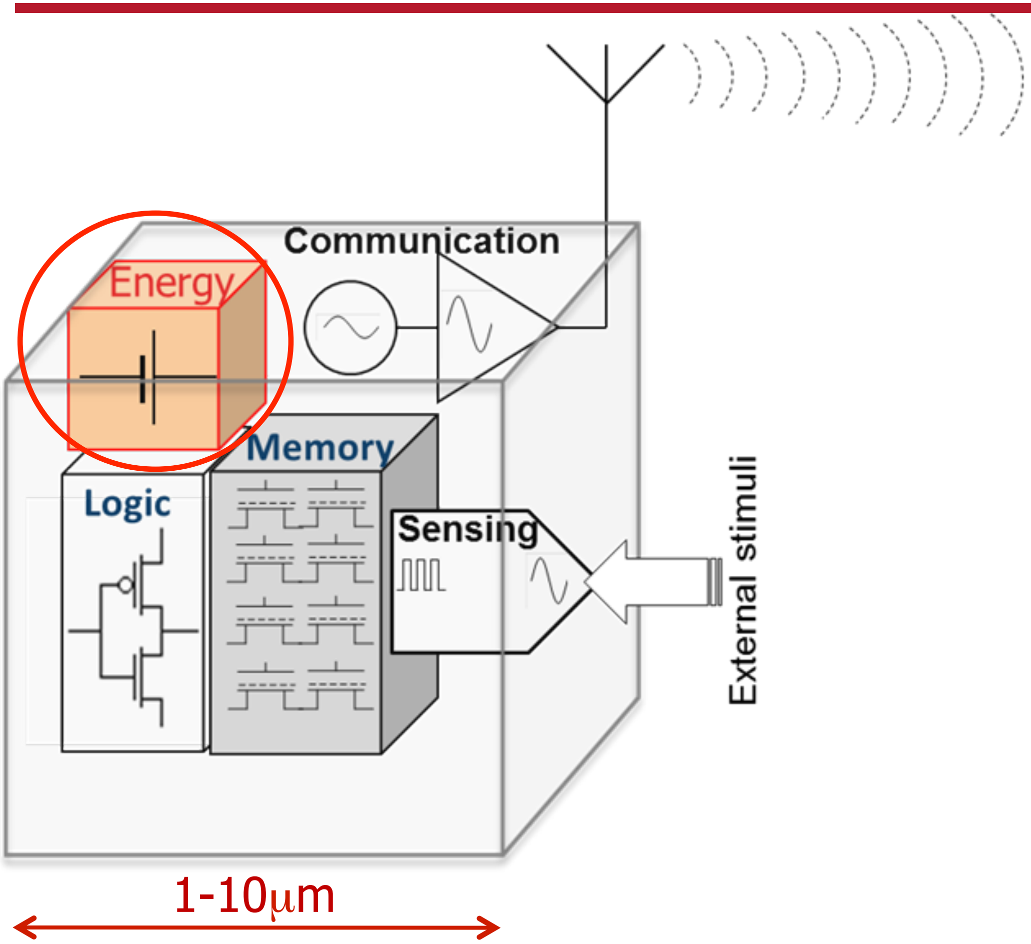
- If you can draw one, but don't, then *your audience won't know what are you talking about*

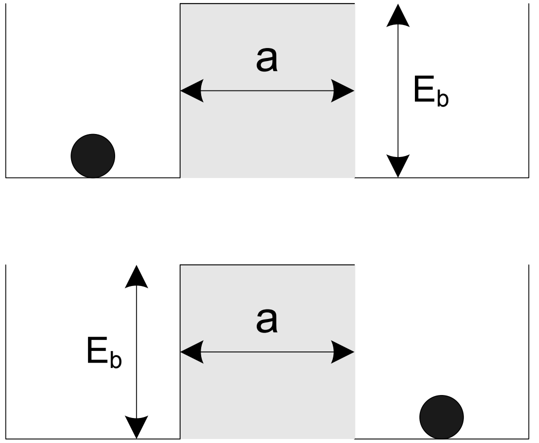


Herbert Kroemer
Nobel Lecture,
Dec. 8, 2000



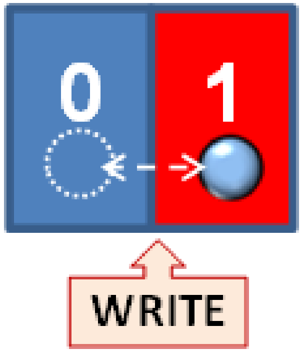
An abstract ICT-Energy element





Memory Logic Sensor

Energy

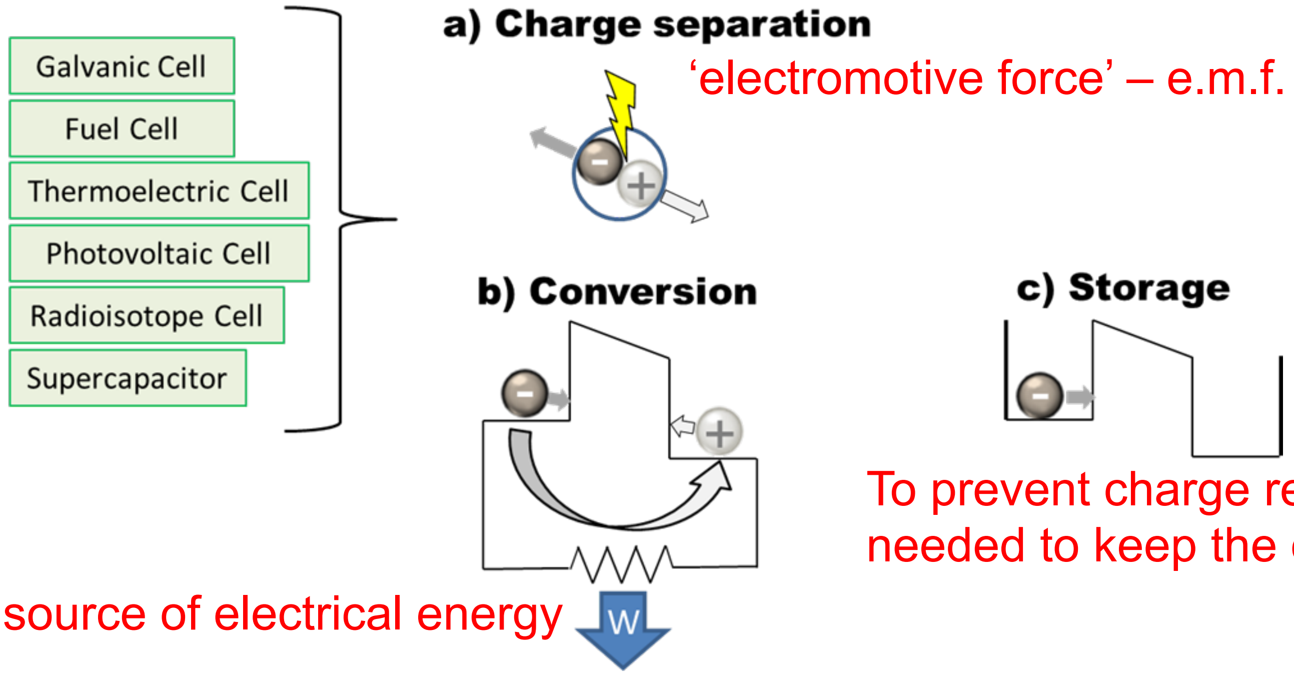


Central Concept: **Energy Barrier**

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



Galvanic Cell

Fuel Cell

Thermoelectric Cell

Photovoltaic Cell

Radioisotope Cell

Supercapacitor

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

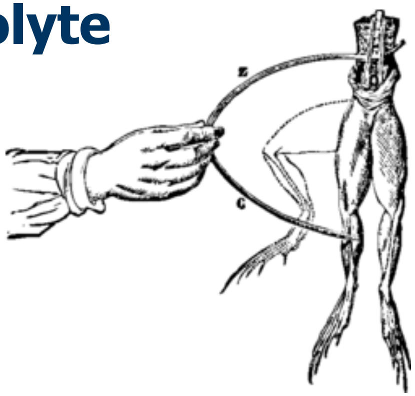
source of electrical energy

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



Alessandro Volta
(1745-1827)

University of Pavia

Charge Separation!

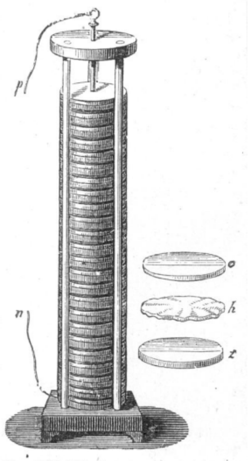
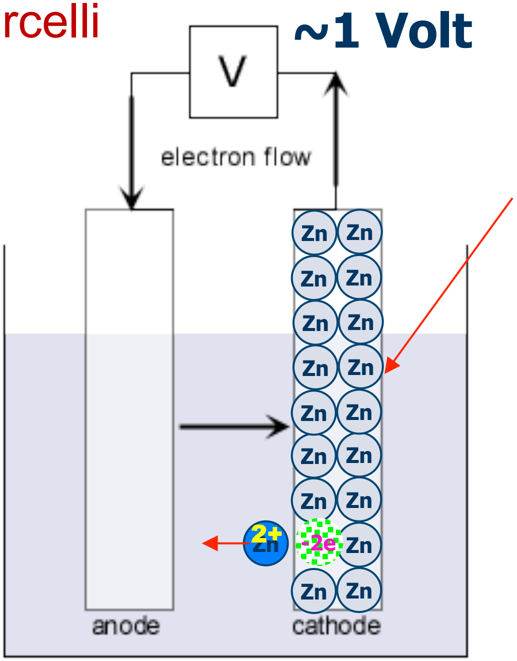


Fig. 255. Pile à colonne construite par Volta en 1800.

Inspired by Galvani, Alessandro Volta built the first battery

Scaling limits of micro-batteries

Reale Collegio di Vercelli



$Li \rightarrow Li^{+} + e^{-}$

$Zn \rightarrow Zn^{2+} + 2e^{-}$

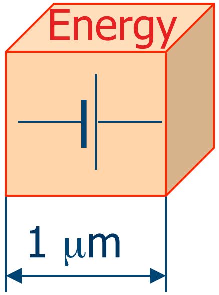
The galvanic cell consumes *atomic fuel* to produce electricity

ε~ 1eV/atom (~chemical bonding energy)

$$E = \varepsilon \cdot N$$

Number of atoms in cathode electrode

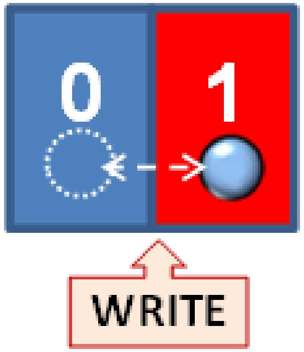
The energy output is limited by the *Avogadro's number, N_A*

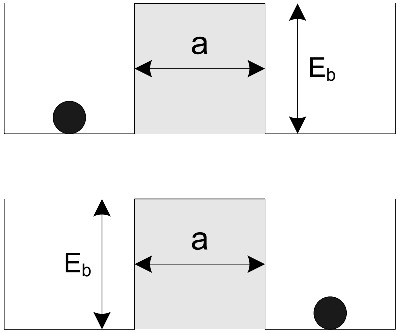


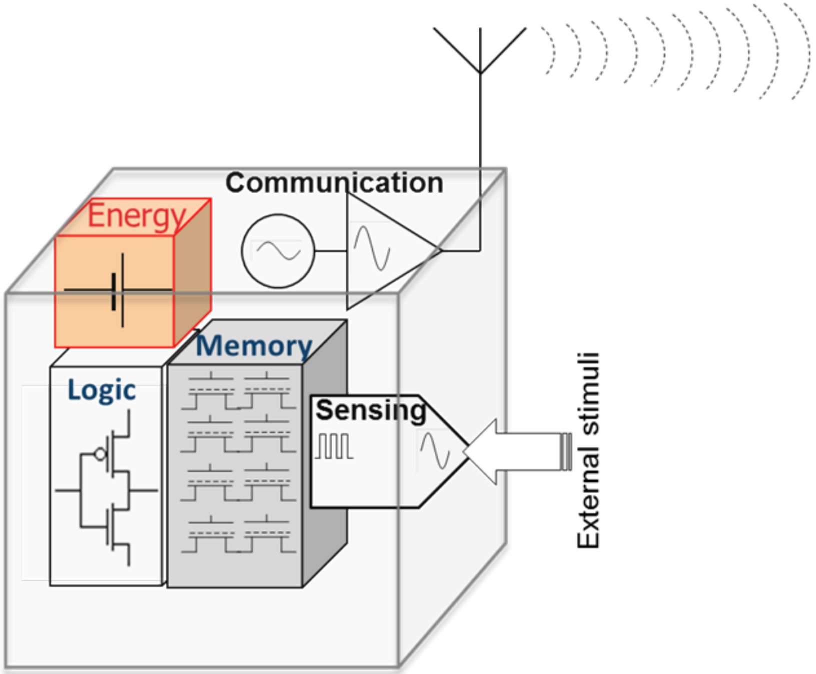
$$E_{\max} \sim eN_A \cdot 1V = 1.6\times 10^{-19} \cdot 6 \cdot 10^{23} \sim 10^{5}\,\frac{J}{mole} \sim 10^{4}\,\frac{J}{cm^{3}}$$

$$E \sim (10^{-4}\,cm)^{3} \cdot 10^{4} \sim 10^{-8}\,J$$

Information Processing Technology Desirata



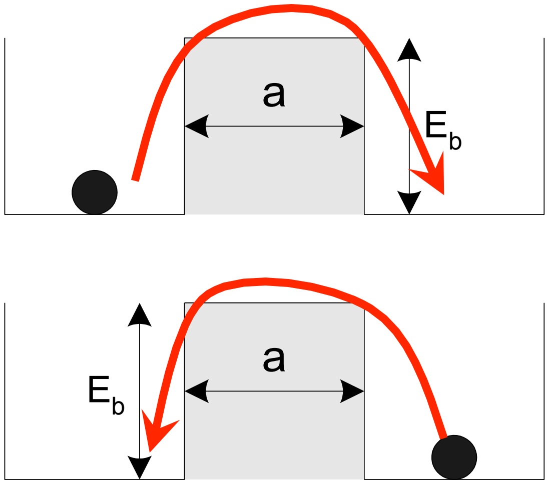




Designers and Users want:

- 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 Π of spontaneous transitions between two wells (error probability)

***D=max*, Π=0** ***D=0*, Π=0.5 (50%)**

Classic distinguishability:

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

Thermal Noise

Minimum distinguishable barrier: Π=0.5

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

Shannon - von Neumann - Landauer limit

Scaling Limits: The Heisenberg Constraint

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

$$\Delta x \Delta p \geq \frac{\mathrm{h}}{2}$$

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

$$a_{crit} \sim \frac{\mathrm{h}}{\sqrt{2mE_b}}$$

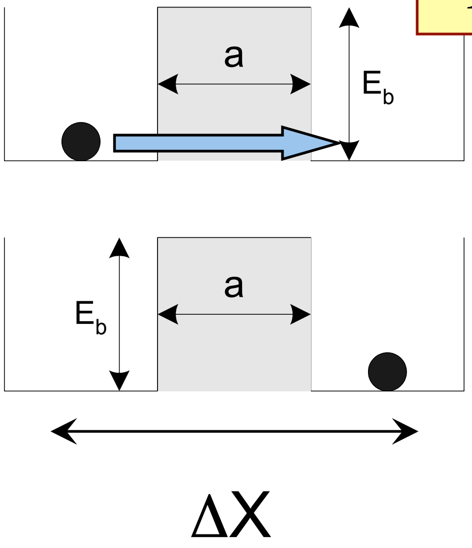
$$t_{\min} \sim \frac{h}{2E_b}$$

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

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

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

Plank's constant

h=6.62x10^{-34} Js

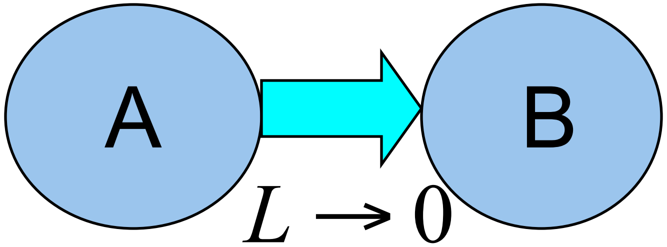
$$\boxed{\Delta t \geq \frac{h}{2\Delta E}}$$

Minimal time of dynamical evolution of a physical system

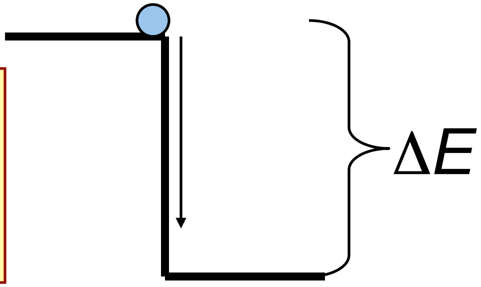
N. Margolus and L. B. Levitin, Physica D 120 (1998) 188

Quantum Resistance

Single –electron Conductance channel (mode)



Local Event –
no coordinate change



$$\Delta V = IR$$

$$\Delta V = \frac{\Delta E}{e}$$

$$I = \frac{e}{\Delta t}$$

R-?

$$\frac{\Delta E}{e} = \frac{e}{\Delta t} R$$

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

$$= \frac{6.64 \times 10^{-34} J \cdot s}{2 \times (1.6 \times 10^{-19} C)^2} = 12.9 k\Omega$$

Summary on Quantum resistance

Heisenberg's Energy-time relation

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

Ohm's Law:
V=IR

von Klitzing constant

$$R_0 = \frac{h}{2e^2} = 12.9k\Omega$$

$$G_0 = \frac{1}{R_0} = \frac{2e^2}{h}$$

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

Nobel Prize in 1985

$$I = n_{chan} \times G_0 \cdot V \cdot \Pi$$

Landauer formula

Summarizing, what we have learned so far from fundamental physics

1) Minimum energy per binary transition

Boltzmann →

$$E_{bit}^{\min} = k_B T \ln 2$$

3×10-21 J

2) Minimum distance between two distinguishable states

$$\Delta x \Delta p \geq \frac{\mathrm{h}}{2}$$

→ Heisenberg

$$x_{\min} = a = \frac{\hbar}{2\sqrt{2mkT\ln 2}} \sim 1nm$$

3) Minimum state switching time

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

→ Heisenberg

$$t_{sw} = \frac{\mathrm{h}}{2kT\ln 2} \sim 10^{-13} s$$

4) Maximum 2D gate density:

$$n = \frac{1}{x_{\min}^2} \approx 10^{14} \frac{device}{cm^2}$$

Total Power Dissipation (@E_{bit}= kTln(2))

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

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

$$\boxed{P_{chip} \sim 3 \times 10^{6} \frac{W}{cm^2}}$$

The circuit would vaporize when it is turned on!

Sun

Limits of Cooling?

Cooling method	W/cm²
Free convection, air	0.25
Free convection, water	1
Forced convection, air	5
Forced convection, water	150

Nanoscale Devices

$$\Pi_{error} = \exp\left(\frac{E_b}{k_B T}\right)$$

"Boltzman constraint" on minimum switching energy

$$E_b^{\min} = k_B T \ln 2$$

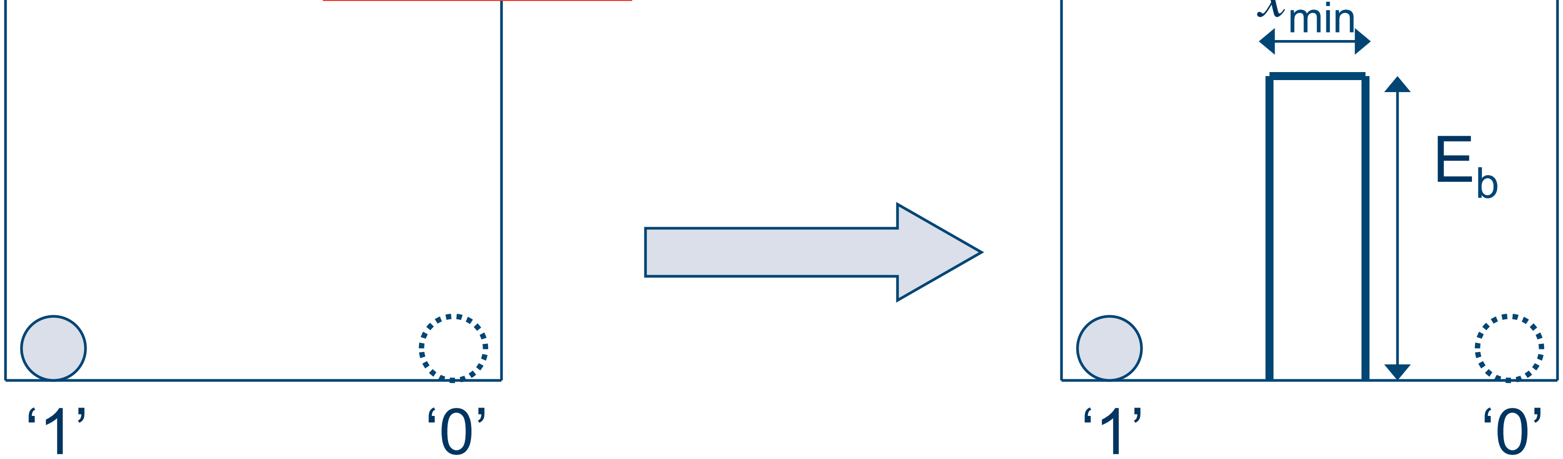
~10^-21 J

$$\Delta x \Delta p \geq \hbar$$

"Heisenberg constraint" on minimum device size

$$x_{\min} = \frac{h}{2\sqrt{2mkT \ln 2}}$$

~1nm

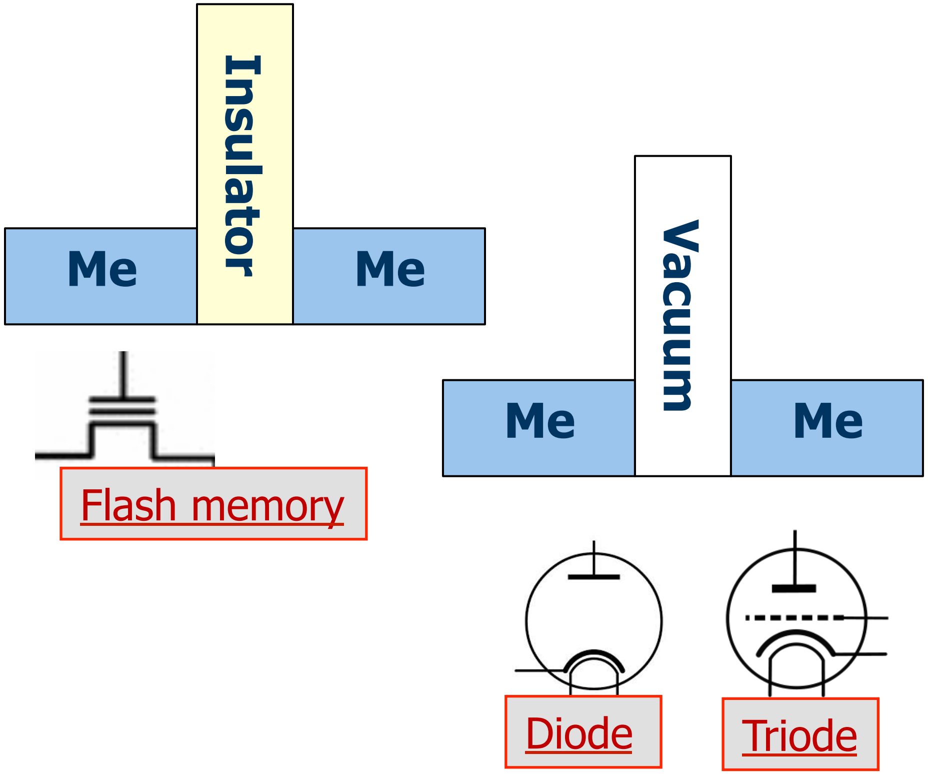


This structure cannot be used for representation/processing information

An energy barrier is needed to preserve a binary state

Barriers in electronic ICT: A Summary

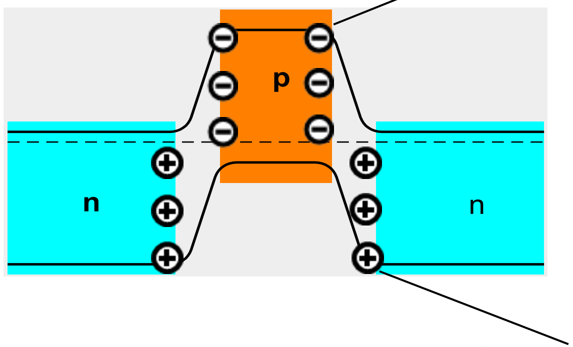
1. Metal-Insulator-Metal stack



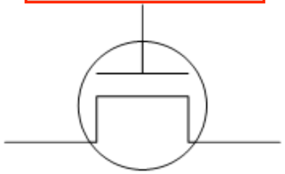
The height of these barriers cannot be changed

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

2. *pn*-junction



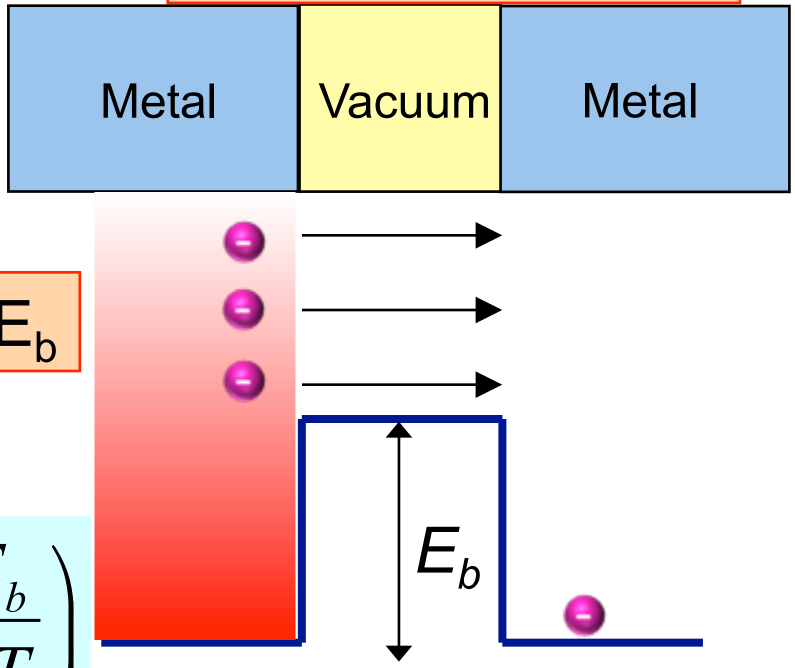
Transistor



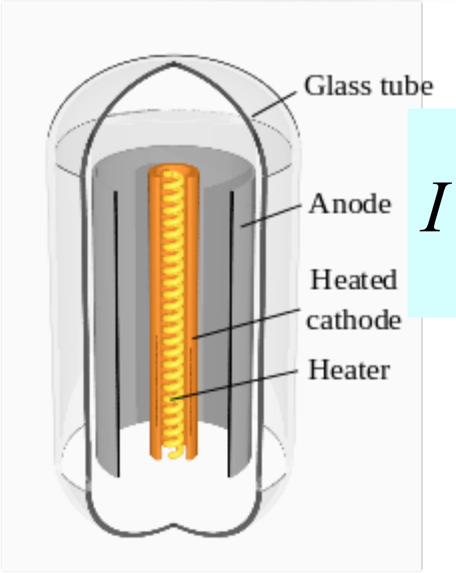
The height of these barriers can be changed

Energy dissipation in binary transitions: Example I Vacuum Tubes

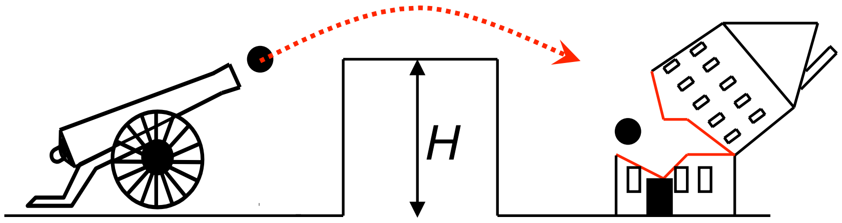
E_b~4-5eV~$200k_BT$



E>E_b



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

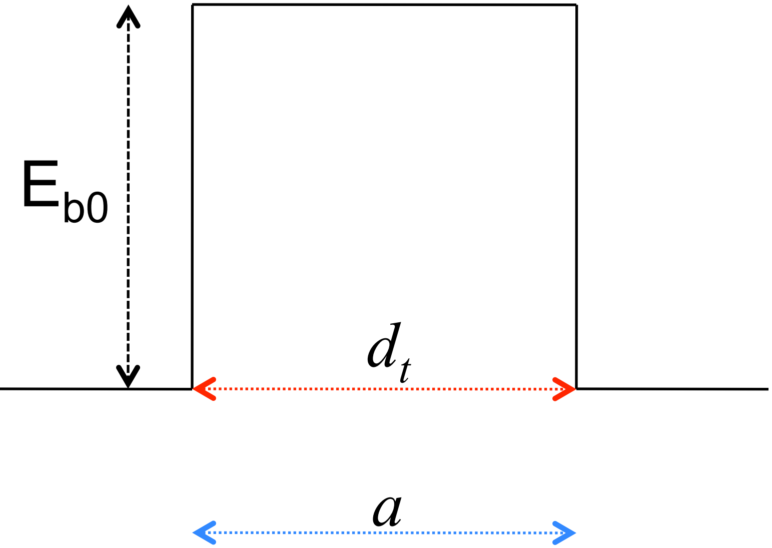


E_b=mgH

E_{kmin}>E_b

dissipation

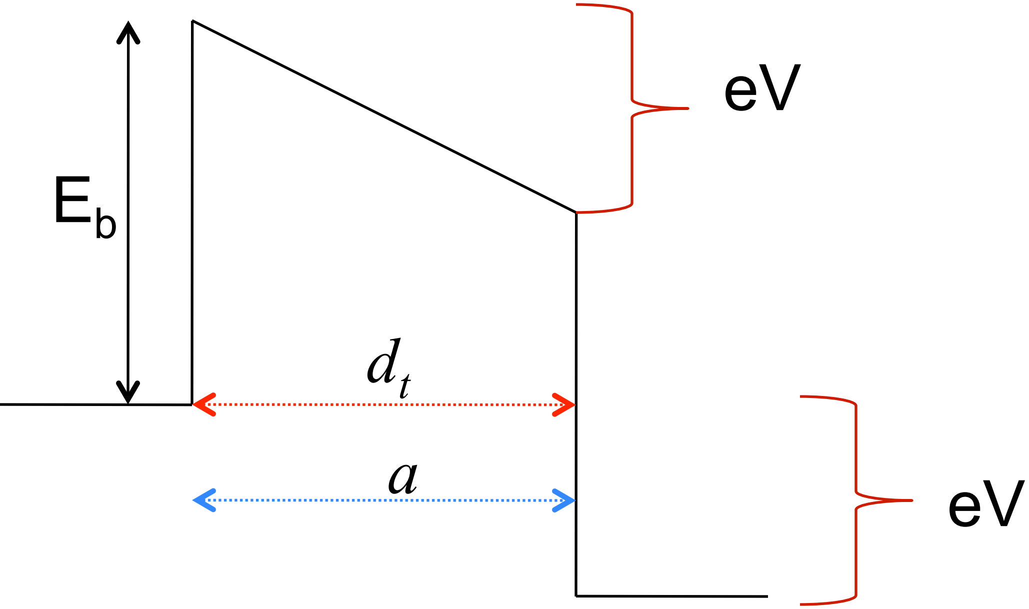
Ideal rectangular barrier (abrupt walls)



V=0

$d_t = a$

Ideal rectangular barrier (abrupt walls)

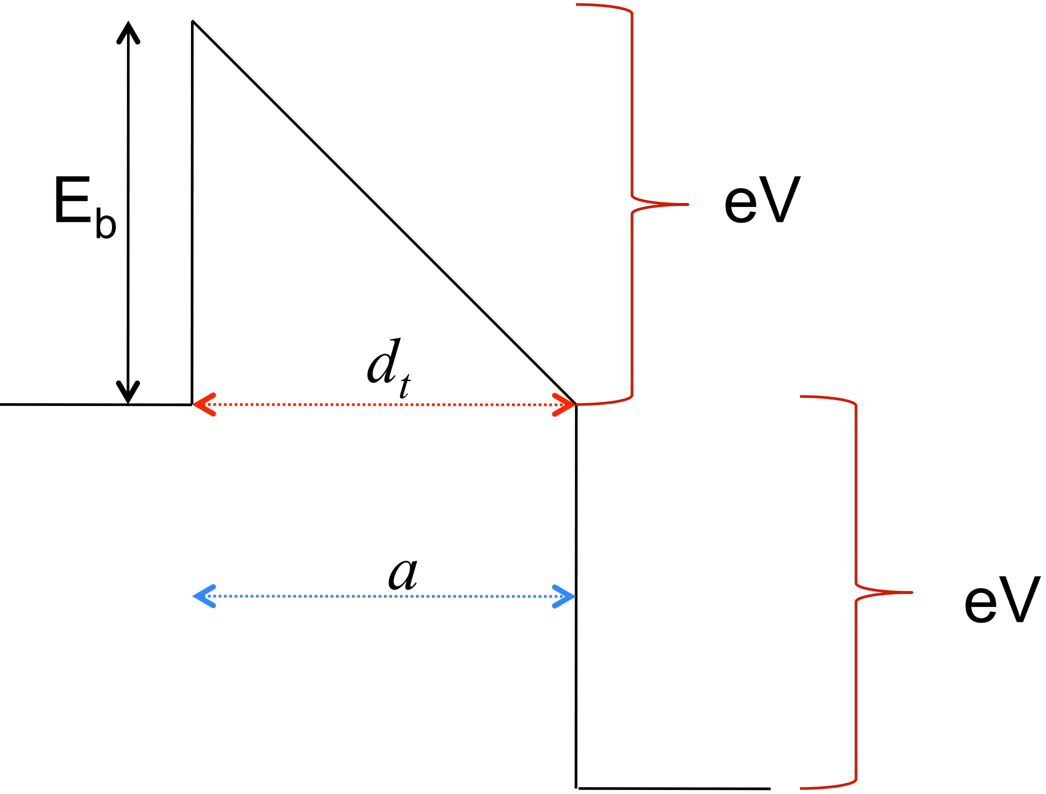


$eV<E_{b0}$

$E_b=E_{b0}$

$d_t=a$

Ideal rectangular barrier (abrupt walls)

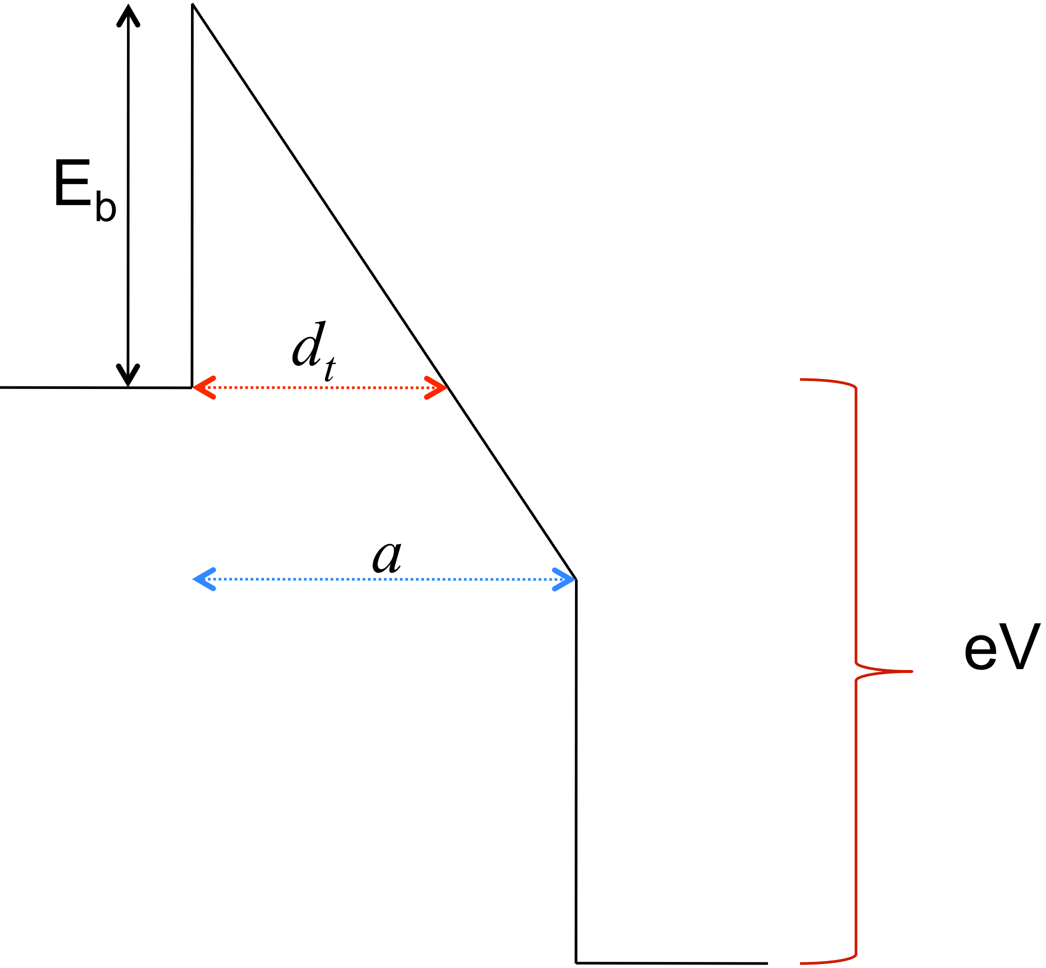


$eV=E_b$

$E_b=E_{b0}$

$d_t=a$

Ideal rectangular barrier (abrupt walls)



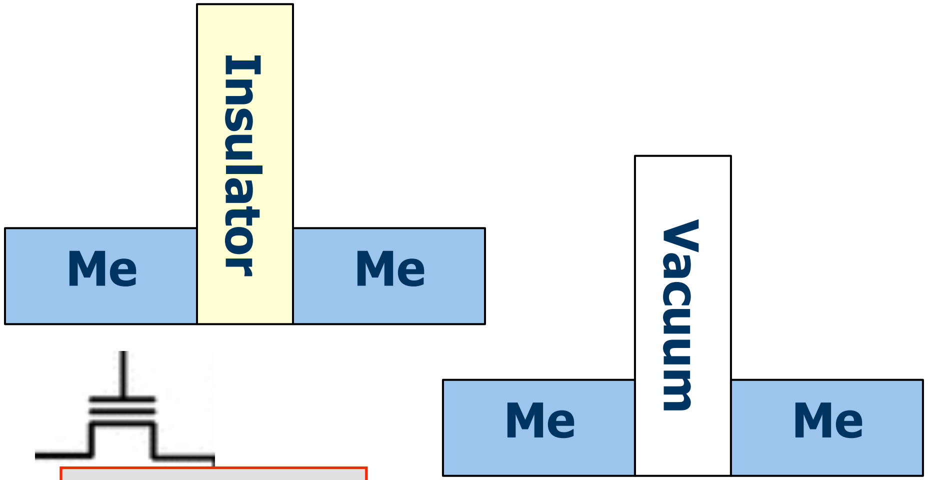
$eV>E_b$

$E_b=E_{b0}$

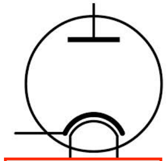
$d_t<a$

Barriers in electronic ICT: A Summary

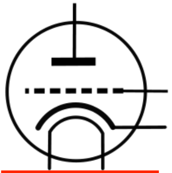
1. Metal-Insulator-Metal stack



Flash memory



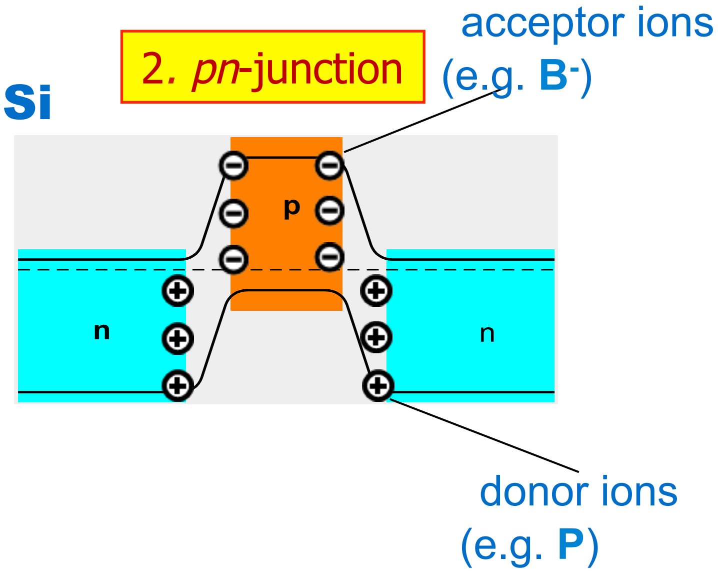
Diode



Triode

The height of these barriers cannot be changed

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

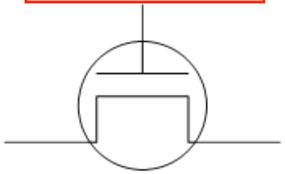


2. *pn*-junction

acceptor ions (e.g. **B**$^-$)

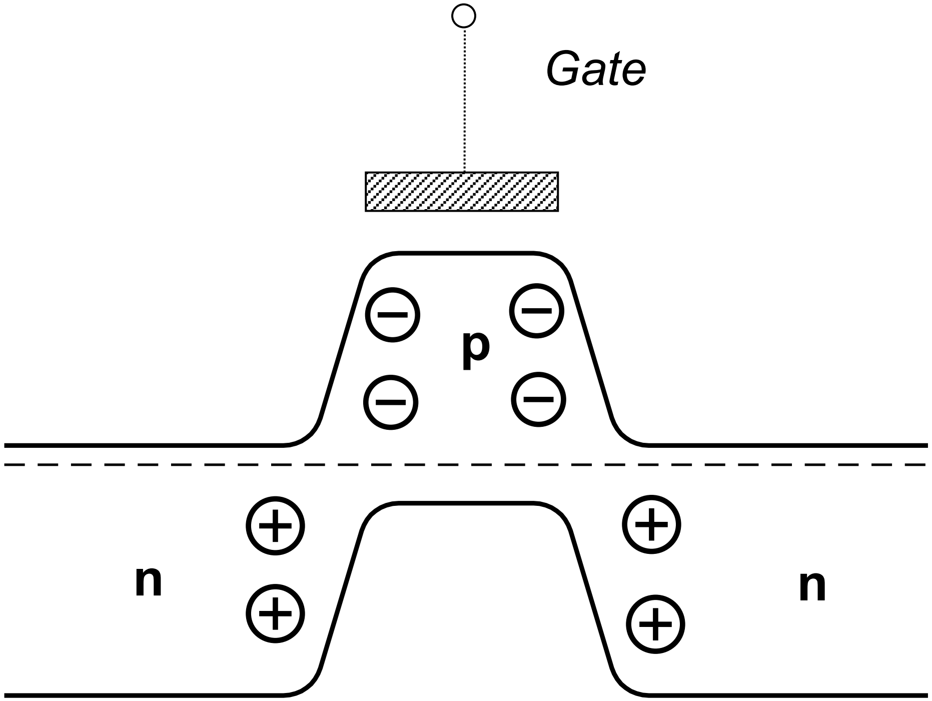
donor ions (e.g. **P**)

Transistor



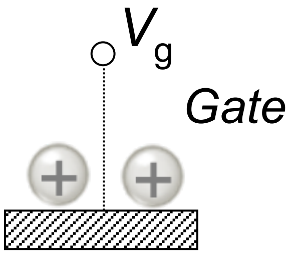
The height of these barriers can be changed

Barrier height control in a semiconductor system

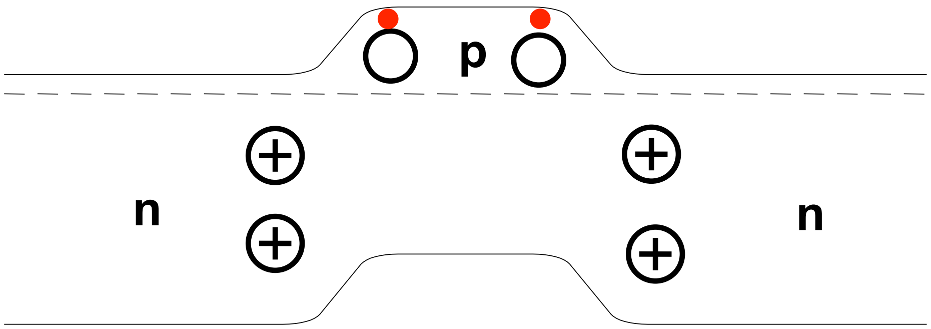


$$E_{b0} = E_g - k_B T\left(\ln\frac{N_V}{N_a^-} + \ln\frac{N_C}{N_d^+}\right)$$

Barrier height control in a semiconductor system



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



$$E_b \approx E_{b0} - eV_g$$

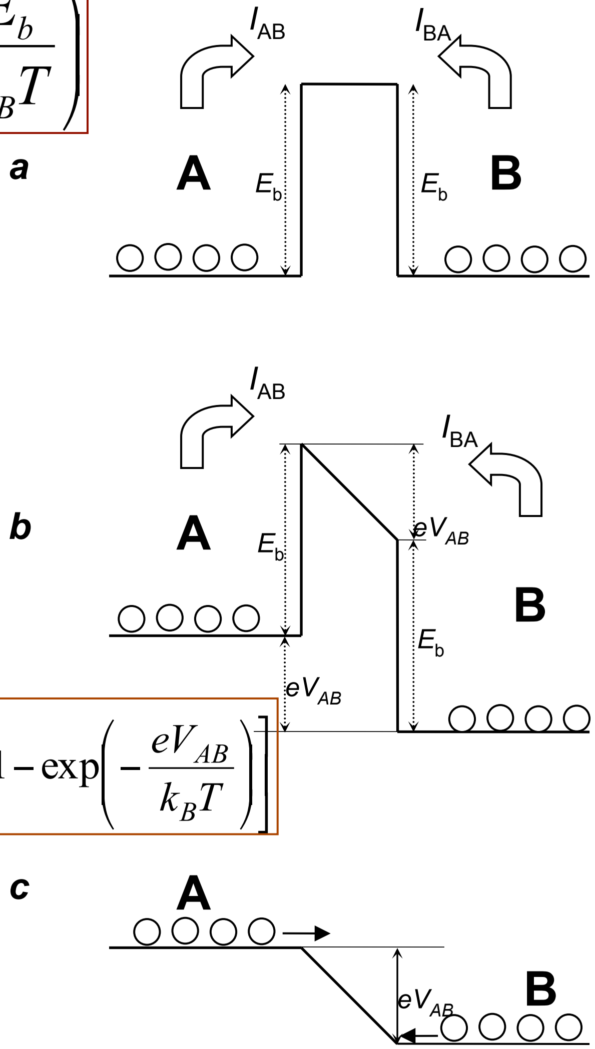
Fundamental operation of multi-electron binary switch:

$$N_A = N_0 \exp\left(-\frac{E_b}{k_B T}\right)$$

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

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

$$E_b = E_{b0} - eV_g$$



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

$$I = eN_0 \exp\left(-\frac{E_{b_0} - eV_g}{k_B T}\right)\left[1 - \exp\left(-\frac{eV_{AB}}{k_B T}\right)\right]$$

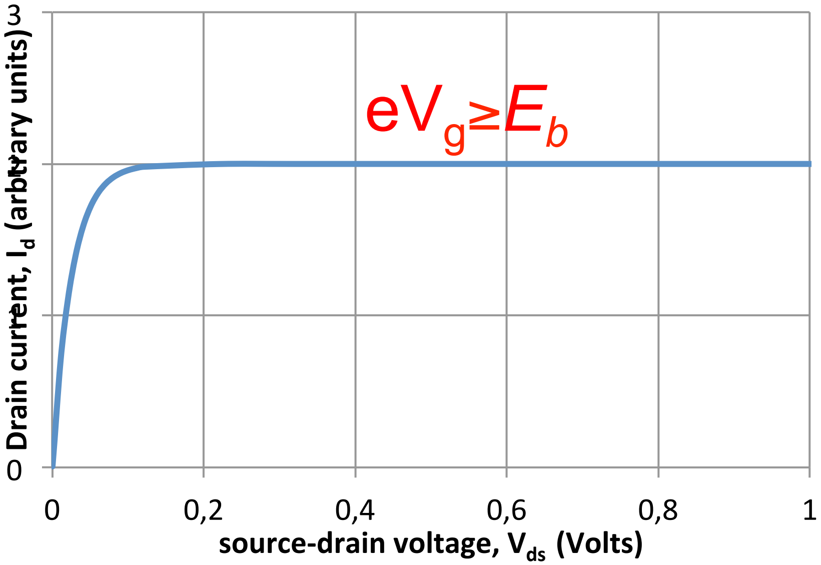
FET Equation

E_b

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

$$\boxed{S = \frac{\Delta V}{dec(\Delta I)}}$$

a 10 times increase



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

$$S = \frac{k_B T \ln 10}{e} = 60\frac{mV}{dec}$$

$$\frac{I_2}{I_1} = 2$$

$$S = \frac{k_B T \ln 2}{e}$$

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…**

$$E_{dis} = \frac{C_g V^2}{2}$$

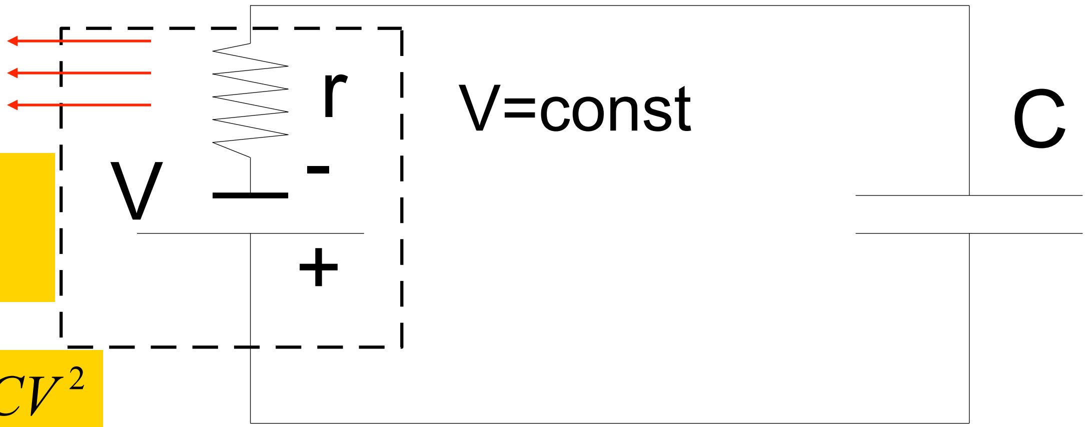
Energy to "deform" the barrier is equivalent to the energy of charging the control (gate) capacitor

Energy dissipated by charging of a capacitor

E_r

DISSIPATED ENERGY

$$E_{dis} = \frac{CV^2}{2}$$



STORED ENERGY

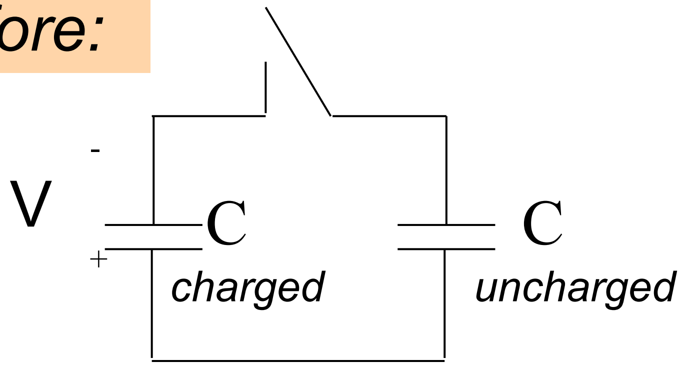
$$E_C = \frac{CV^2}{2}$$

$$E_r = \int_0^\infty \frac{V_r^2(t)}{r} = \frac{1}{r}\int_0^\infty \left(Ve^{-\frac{t}{rC}} \right)^2 dt = \frac{V^2}{r}\int_0^\infty e^{-\frac{2t}{rC}} dt = \frac{V^2}{r}\cdot\frac{rC}{2}\int_0^\infty e^{-\tau} d\tau = \frac{CV^2}{2}$$

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

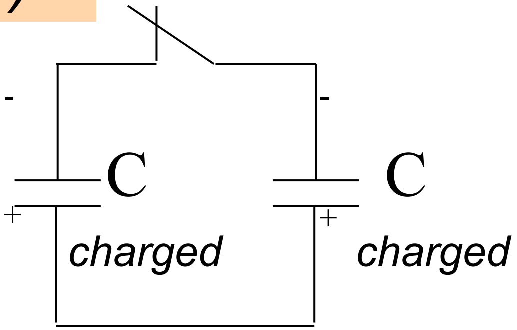
Before:



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

$$E_{before} = \frac{CV^2}{2}$$

After (*):



$$C^* = 2C$$

$$q^* = q = CV = C^*V^* = 2CV^*$$

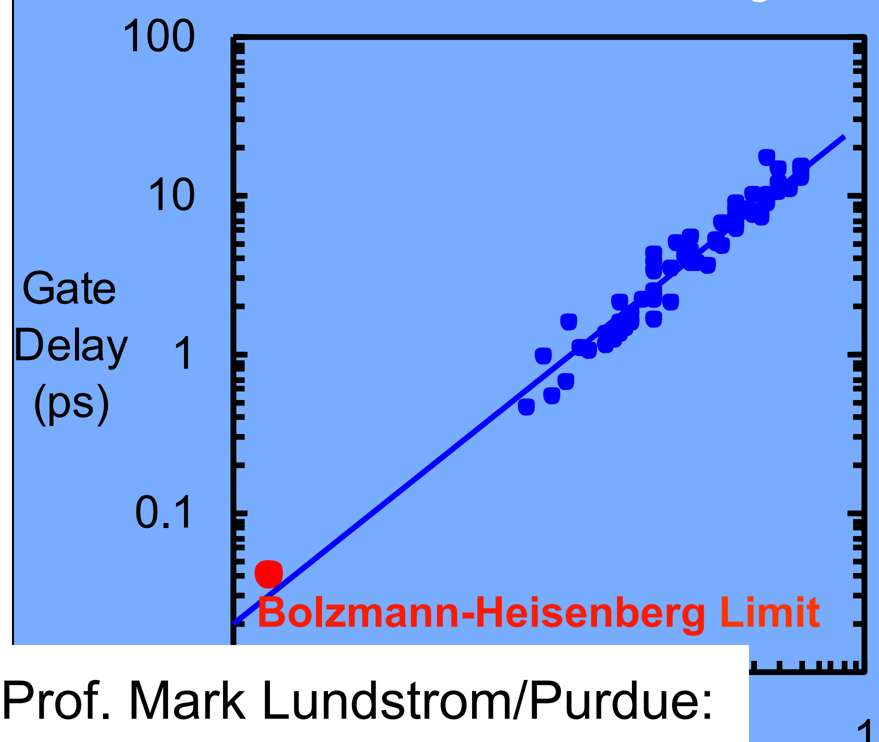
$$V^* = \frac{V}{2}$$

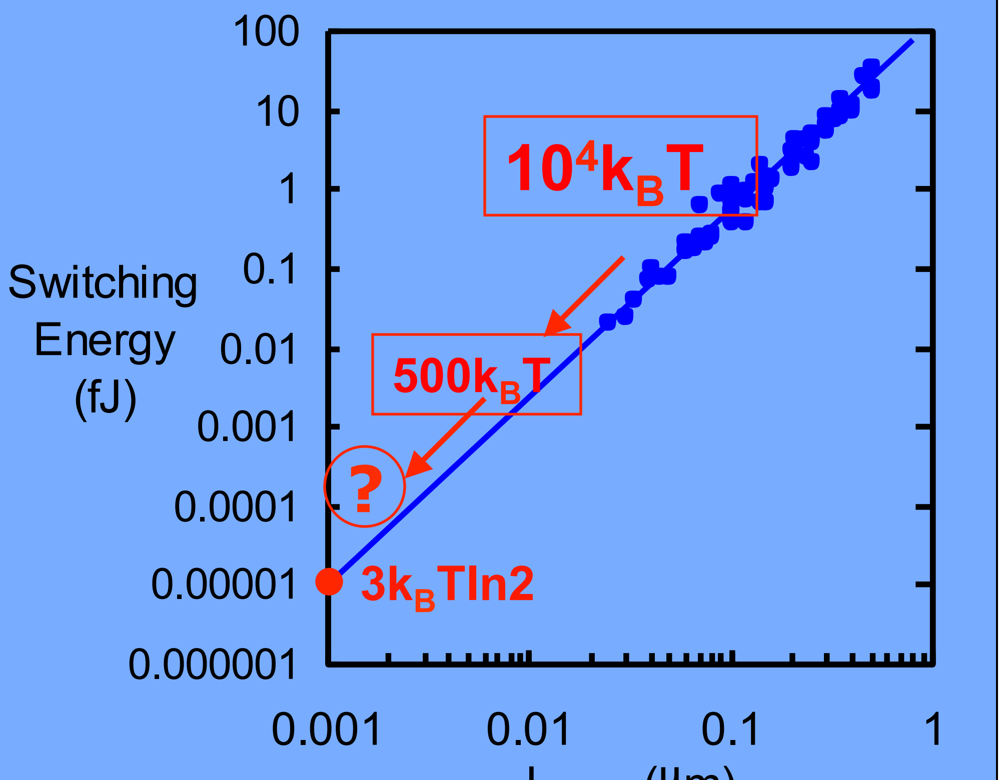
E_{after} is the sum of energy stored in two capacitors) after closing the switch

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

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 $k_BT\ln2$ and not $k_BT\ln2$?

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

Answer:

1) **System reliability costs**
2) **Communication costs**
3) **Fan-Out costs**

System Level Energetics I: *Reliable Switching*

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

In useful computation, Π_{err} <<0.5, hence barrier height larger than $k_B ln2$ is needed (larger <u>total</u> power consumption)

Question: How Much Larger?

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

The probability that all N switches in a circuit work correctly

$\mathbf{N\uparrow \rightarrow L\downarrow \rightarrow \Pi_{err}\uparrow}$ $\longrightarrow$ $\mathbf{\Pi_{err}\downarrow \rightarrow E\uparrow}$

(Heisenberg)

(Boltzmann&Heisenberg)

System Constraint on Minimum Energy per Bit

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

The probability that all N switches in a circuit work correctly

$$\Pi_{syst} > \Pi_{crit}$$

e.g., → 0.5 lower boundary

e.g., → 0.99 a “reasonable” boundary

$$\Pi_{err} = 1-\Pi_{crit}^{\frac{1}{N}}$$

$$\Pi_{err} = f(E_b)$$

$$E_{b_{\min}} = f(N)$$

$$N_{\max} \sim \frac{1}{a^2}$$

$$\Pi_{err} = \overset{\text{Boltzmann}}{\boxed{\exp\left(-\frac{E_b}{kT}\right)}} + \overset{\text{Heisenberg}}{\boxed{\exp\left(-\frac{2\sqrt{2m}}{\hbar}a\sqrt{E_b}\right) - \exp\left(-\frac{\hbar E_b + 2akT\sqrt{2mE_b}}{\hbar kT}\right)}}$$

Uniformly Scaled Information Processor

$$N_{\max} \sim \frac{1}{(20L_g)^2} \sim 10^{10}\,cm^{-2}$$

- The maximum possible number *N* of binary switches in a close-packed array is inversely proportional to the square of the barrier lengt *L* (e.g. the FET gate length L_g)
  - $N=f_1(L) \leftrightarrow L=f_2(N)$

$$L_g \sim \frac{1}{\sqrt{20N}}$$

- 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)

$$E_{b_{\min}} = f(N)$$

for $\Pi = \Pi_{crit}$

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

Π_{crit}=0.99

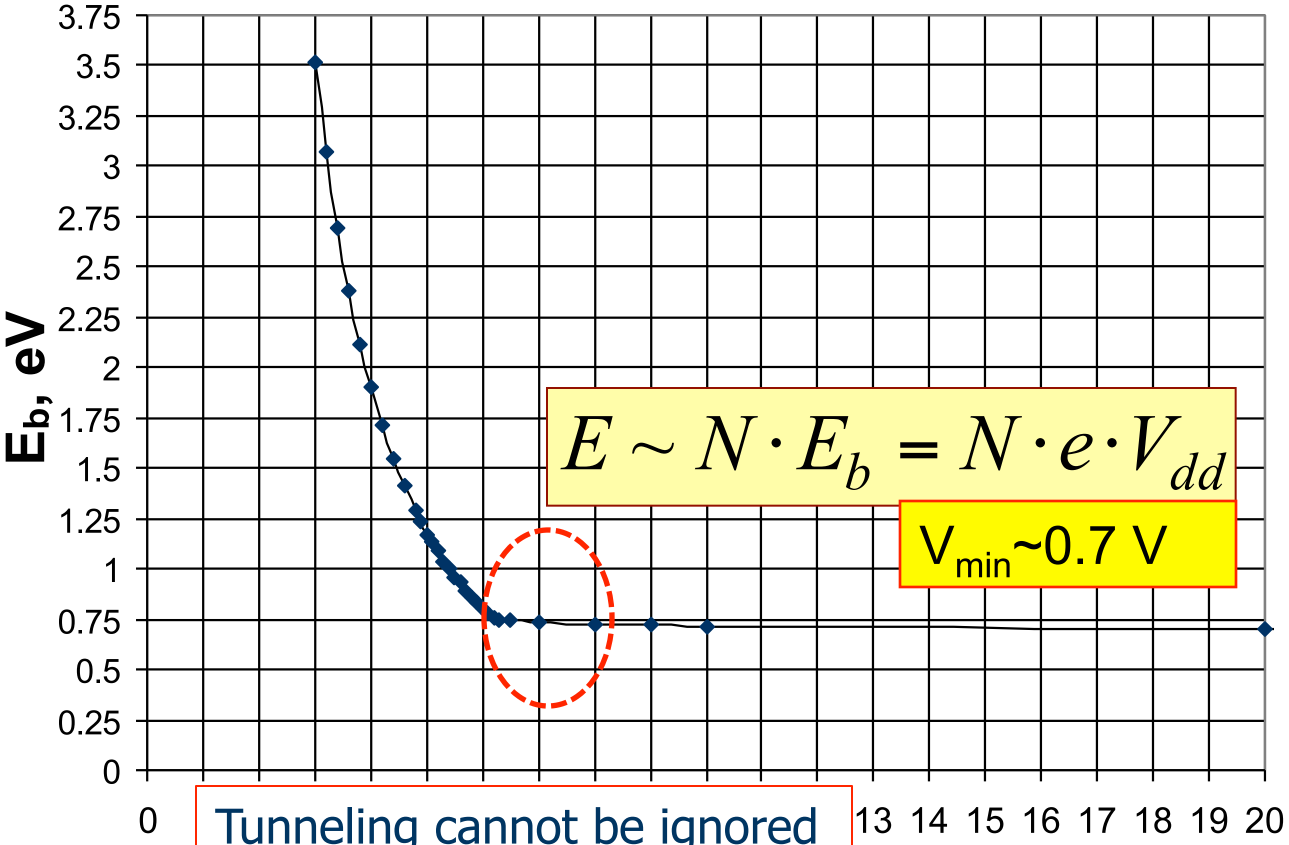
L_g, nm	*N*, cm^{-2}	$E_{b\min}$
100	2.50E+07	0.65
50	1.00E+08	0.67
30	2.78E+08	0.69
20	6.25E+08	0.70
10	2.50E+09	0.71
9	3.09E+09	0.72
8	3.91E+09	0.72
7	5.10E+09	0.73
6	6.94E+09	0.80
5	1.00E+10	1.17
4	1.56E+10	1.90
3	2.78E+10	3.52

Generic Challenges

- Energy – Errors dilemma

1cm×1cm

N~10^{10} cm^{-2}



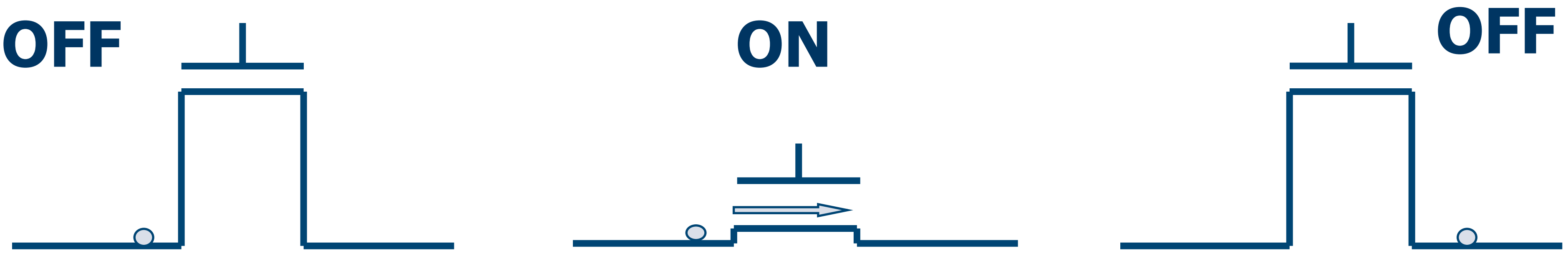
$$E \sim N \cdot E_b = N \cdot e \cdot V_{dd}$$

V_{min}~0.7 V

Tunneling cannot be ignored for ***a***<5nm, which sets a practical limit

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

Switching Energy: Energy of Full-cycle



$E_{OFF-ON} = E_b$

$E_{carrier}$

$E_{ON-OFF} = E_b$

$$E_{bit_{\min}} = 2E_{b_{\min}} + E_{carrier}$$

kTln2

We are fighting ambient thermal energy!

$$E_{SW\min} = 3k_B T \ln 2$$

x N

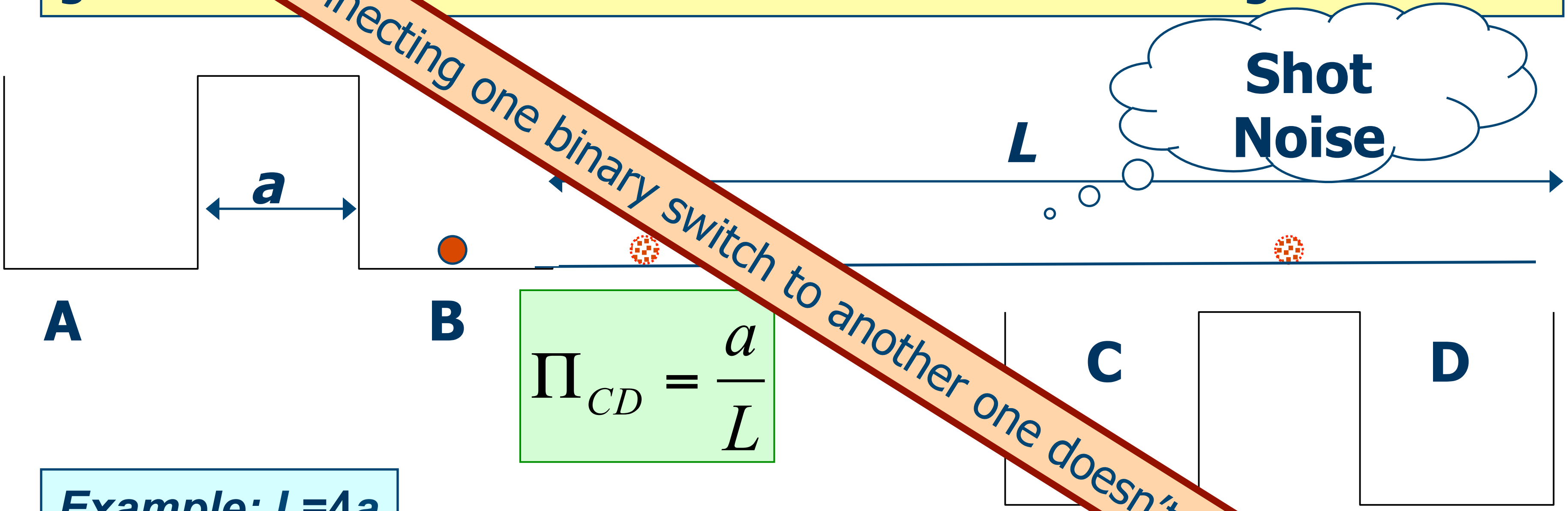
N – the number of electrons

$$E_{sw}=2E_b+NE_w=(N+2)k_BT\ln 2$$

Connecting Binary Switches via Wires: *Extended Well Model*

The problem is to 'place' the electron on the down stream gate – more than one electron is needed to 'charge' the line

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



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

Example: ***L=4a***

N=1→P<0.25

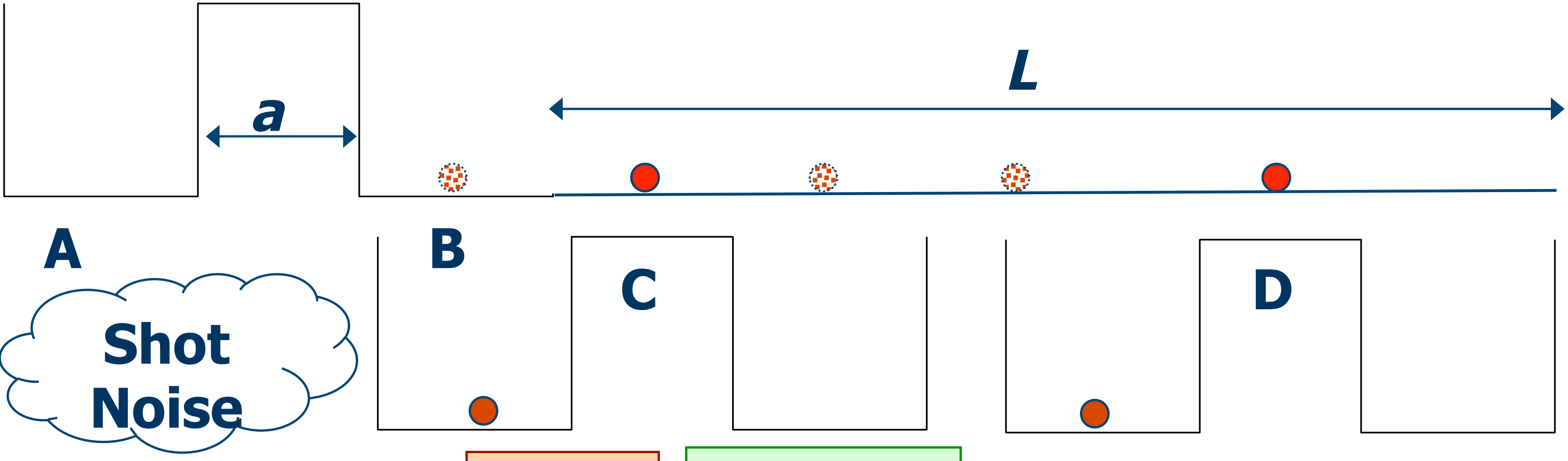
N – the number of electrons

In General:

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

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*>2*na

***n*- fan**

N – the number of electrons

***n*=2**

L*=4*a

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

N_{min}=5

N	Π
1	0.06
2	0.19
3	0.33
4	0.47
5	0.58
6	0.68

Minimum switching energy for connected binary switches

$$E_{sw}=2E_b+NE_b=(N+2)E_b$$

FO2

***n*=2**
L*=4*a

$$N_{min}=5$$

$$E_{sw}=7k_BT\ln 2$$

FO4

***n*=4**
L*=8*a

$$N_{min}=14$$

$$E_{sw}=16k_BT\ln 2$$

Communication between logic switches takes more energy than information processing (switch operations)

N	Π
1	0.00
2	0.00
3	0.01
4	0.03
5	0.06
6	0.09
7	0.14
8	0.19
9	0.24
10	0.29
11	0.35
12	0.41
13	0.46
14	0.51
15	0.56
16	0.60
17	0.65
18	0.68

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*=8*a

N electrons	Operational reliability
14	50%
20	75%
42	99%

We need many electrons for reliable communication

More electrons means more energy…

Mark Lundstrom/Purdue:

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

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

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

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

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

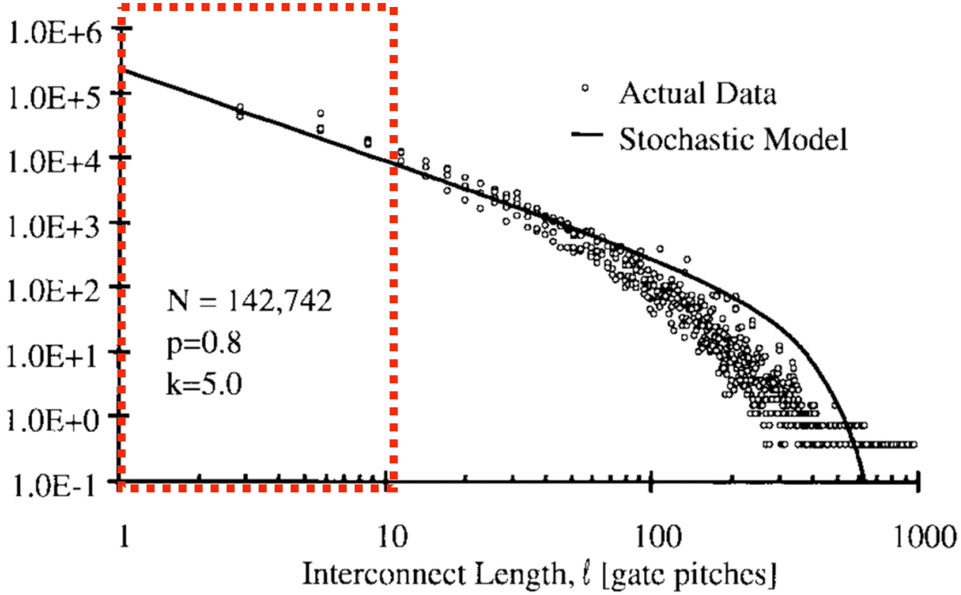
Long Interconnects

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

n=2 ***L=100a***

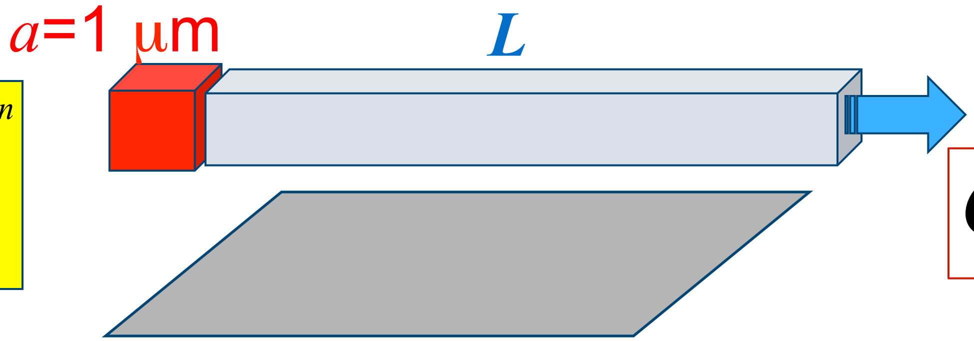
$$\Pi_n = \left(1 - \left(1 - \frac{a}{L}\right)^N\right)^n$$

N electrons	Operational reliability	
121	50%	E~120k_BT
198	75%	E~200k_BT
487	99%	E~500k_BT



Communication between an information processing system and the outside world

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



$$C \sim \varepsilon_0 L$$

$$E_{\min} \sim eNk_BT$$

$$E = \frac{CV^2}{2} \sim \frac{\varepsilon_0 L}{2}\left(\frac{k_BT}{e}\right)^2$$

L	Joules	Joules
100 µm	2.86E-19	2.96E-19
1 mm	2.87E-18	2.96E-18
1 cm	2.87E-17	2.96E-17
10 cm	2.87E-16	2.96E-16
1 m	2.87E-15	2.96E-15

Communication cost per bit per unit length: $\sim \frac{\varepsilon_0}{2}\left(\frac{k_BT}{e}\right)^2$ =3×10^{-15} J/(bit·m)

Wireless Communication for Micro-Scale Systems

Example: **Uniformly** radiated wireless communication

Bernardo Dessau,

La telegrafia senza filo

Augusto Righi, Bernardo D

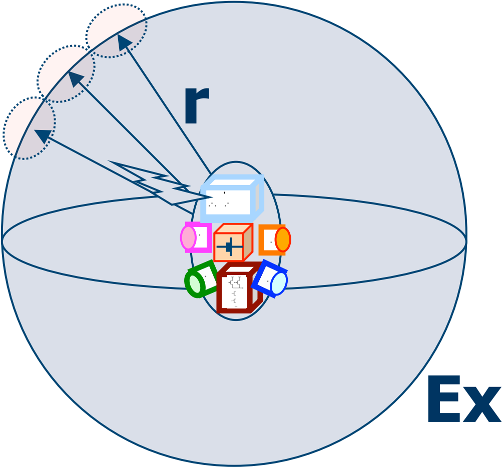
University of Perugia

$$E_{com} = N_{photons} \cdot E_{ph}$$

$$E_{ph} = h\nu = \frac{hc}{\lambda}$$

$$N_{photons} \sim \frac{4\pi r^2}{\lambda^2}$$

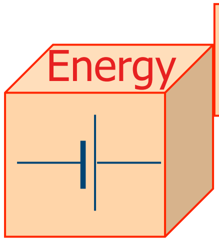
~Friis equation



$$E_{com} \sim \frac{4\pi r^2 hc}{\lambda^3}$$

Example: r=1m

λ~1 μm



$$E \sim 10^{-8} J$$

10 bit

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

Communication Scaling

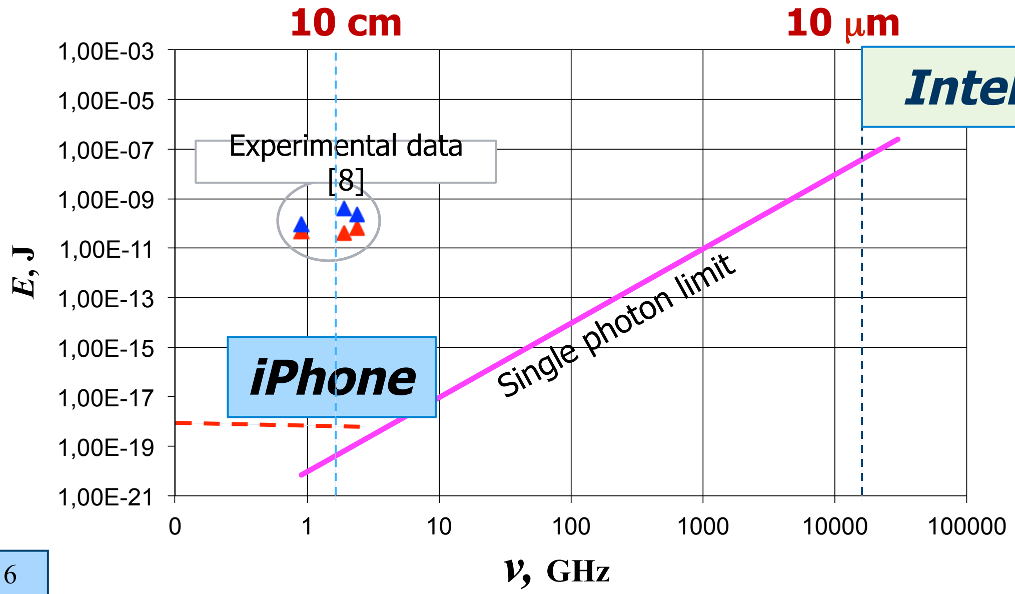
$$E_{com} \sim \frac{4\pi r^2}{\lambda^2}\cdot\frac{hc}{\lambda} = \frac{4\pi r^2 hc}{\lambda^3}$$

Scaling of omni-directional wireless is limited due to increased energy costs

iPhone~ 10 cm

$$N_{bit} = \frac{E_{total}}{E_{bit}} \sim \frac{10^6}{10^{-18}}$$

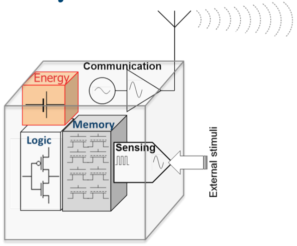
~10^{24} transmitted bits



Intelligent μ-cell

***r*=10 m**

~10 μm cell



$$N_{bit} = \frac{E_{total}}{E_{bit}} \sim \frac{10^{-5}}{10^{-7}}$$

~100 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*

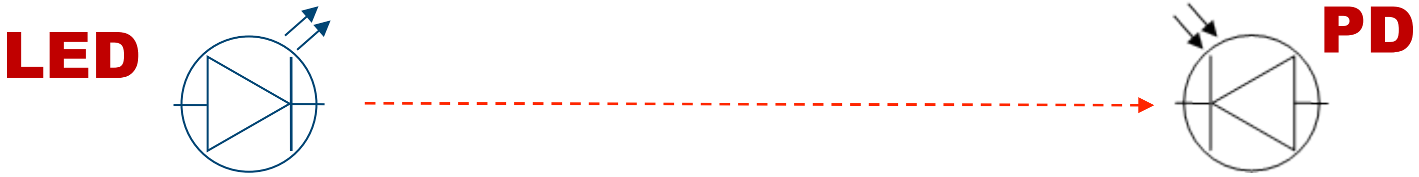
Thermal & Shot Noise – we need more electrons for reliable branched communication

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

Eli Yablonovitch/UC Berkeley

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

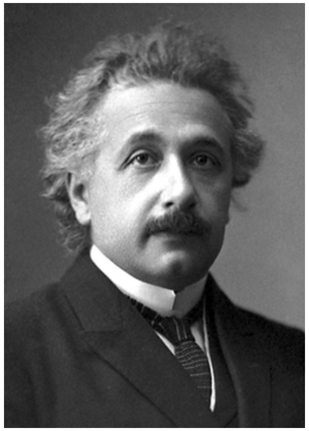
Directed transmission



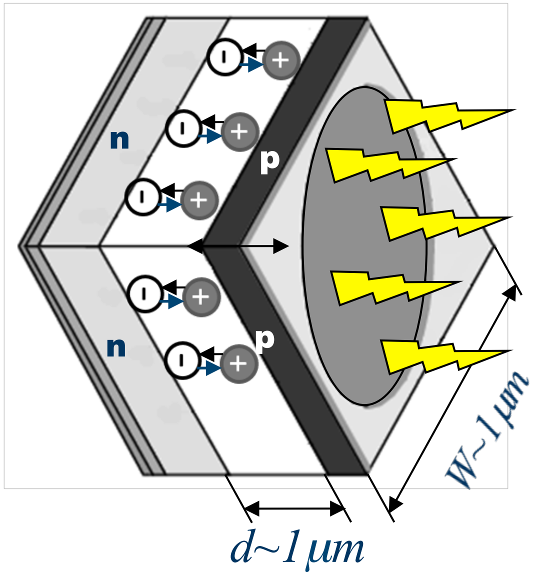
$$C_{pn} \sim \frac{\varepsilon_0 K d^2}{W}$$

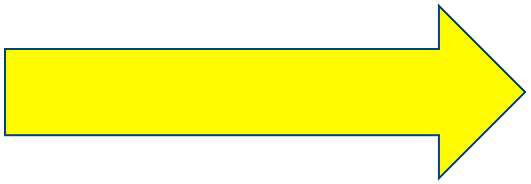
λ~1 μm

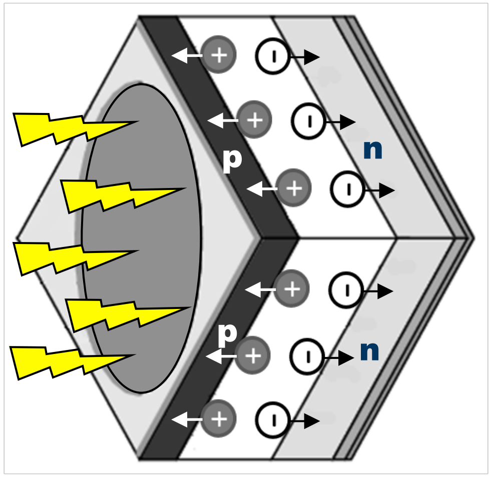
Nobel Prize in Physics



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





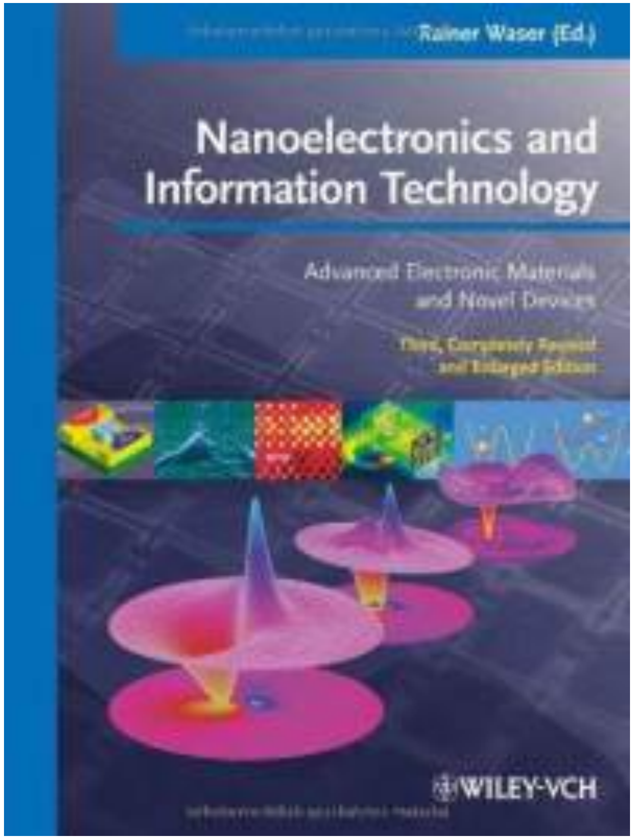


$$E_{LED} = \frac{C_{pn}^2 V_{th}^2}{2} \sim \frac{5\text{x}10^{-16} \cdot 1^2}{2} = 2.5 \cdot 10^{-16}\, J$$

Orientation problem

MINIMAL MEMORY ELEMENT

(Nonvolatile case)


Rainer Waser (Ed.)
Nanoelectronics and Information Technology
Advanced Electronic Materials and Novel Devices
Third, Completely Revised and Enlarged Edition
WILEY-VCH

V. V. Zhirnov and T. Mikolajick,
Chapter 26: *Flash Memories*,
in: **Nanoelectronics and Information Technology,**
by R. Waser (Ed.) Wiley-VCH 2012.

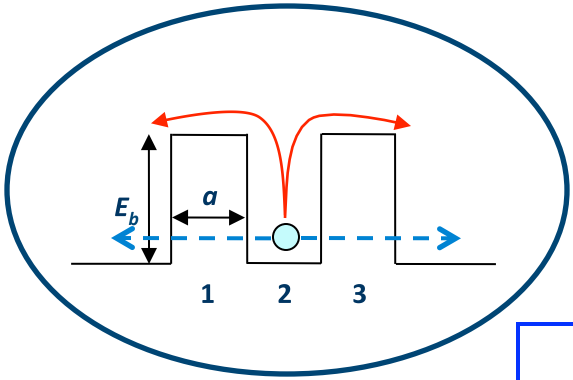
What is the smallest volume of matter needed for memory?

Minimal Electronic Memory

$$t_s = \frac{e}{I_s} \sim 10y$$

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

$$I_{o-b} = \frac{e}{h} \cdot k_B T \cdot \exp\left(-\frac{E_b}{k_B T}\right)$$



$$t_{o-b} = \frac{h}{k_B T} \exp\left(\frac{E_b}{k_B T}\right)$$

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

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

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

E_{bmin}=1.3 eV

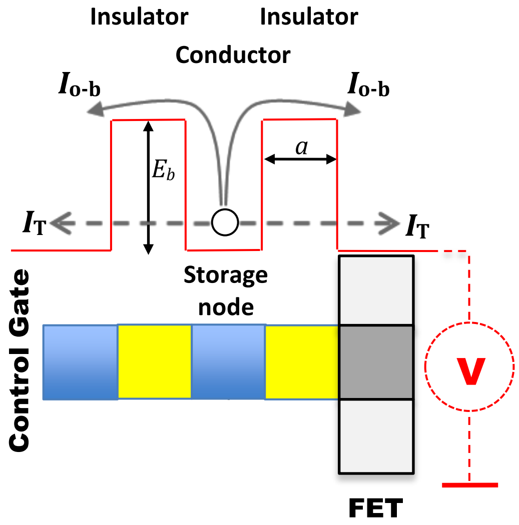
a_{min}=4.30 nm

(Limited by the mass of electron)

Adjustments: effective mass, electrostatics etc.: a_{min}~5 nm, E_{min}~2-3 eV

Electron-based Nonvolatile Memory (Flash)

1. Basic Concept

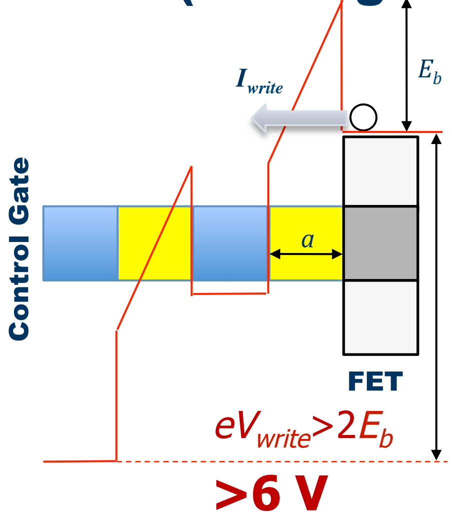


E_{bmin} >1.7 eV (>10 y retention)

$E_{b\,SiO2}$=3.1 eV

a_{min}~5 nm

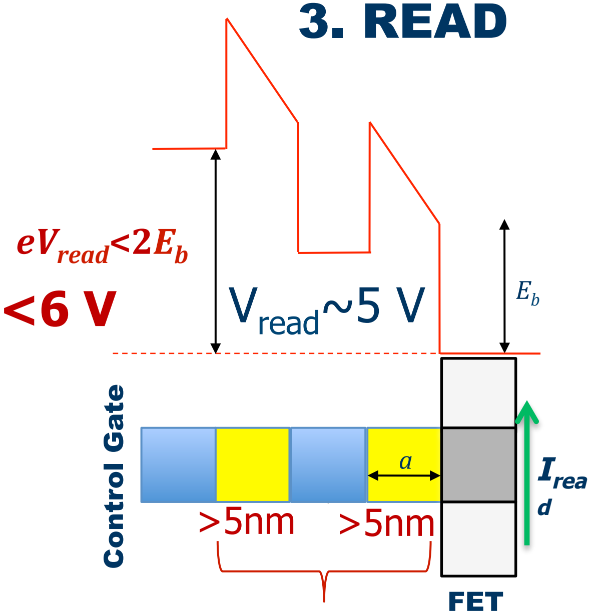
2. WRITE (F-N regime)



$V_{write\ min}$> 6-7 Volt (very slow)

V_{write} >**10-15 Volt** (ms-μs)

3. READ

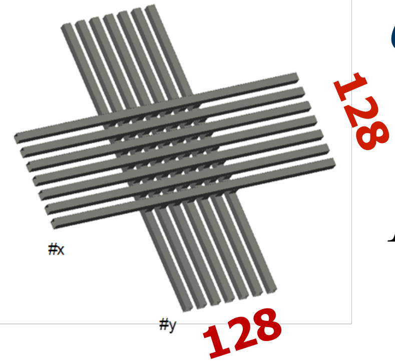


T_{ox}>10nm

$$\boxed{\frac{T_{ox}}{L_{ch}} \sim 1}$$

F_{min}>10nm

4. Array

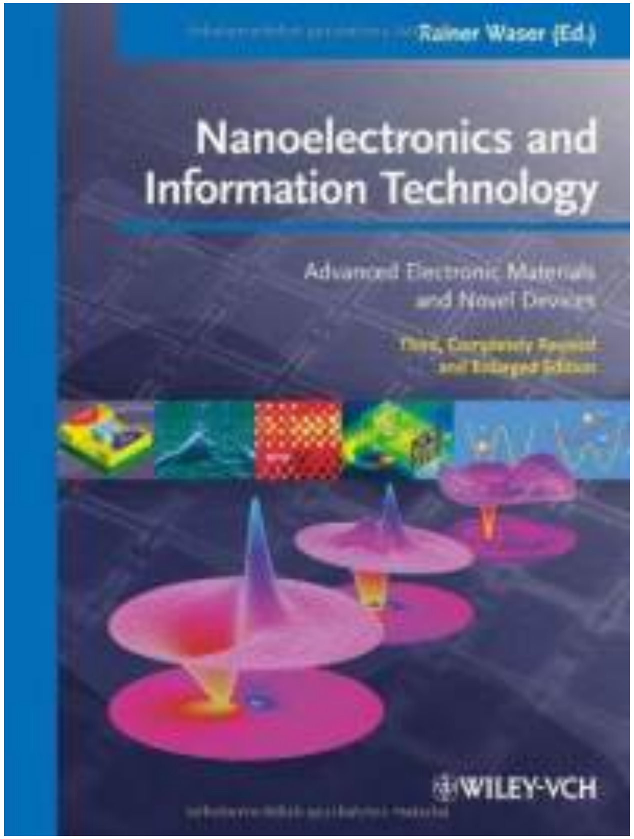


C_{line}~ 10^{-14} F

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

or 2×10^{-15} J/bit

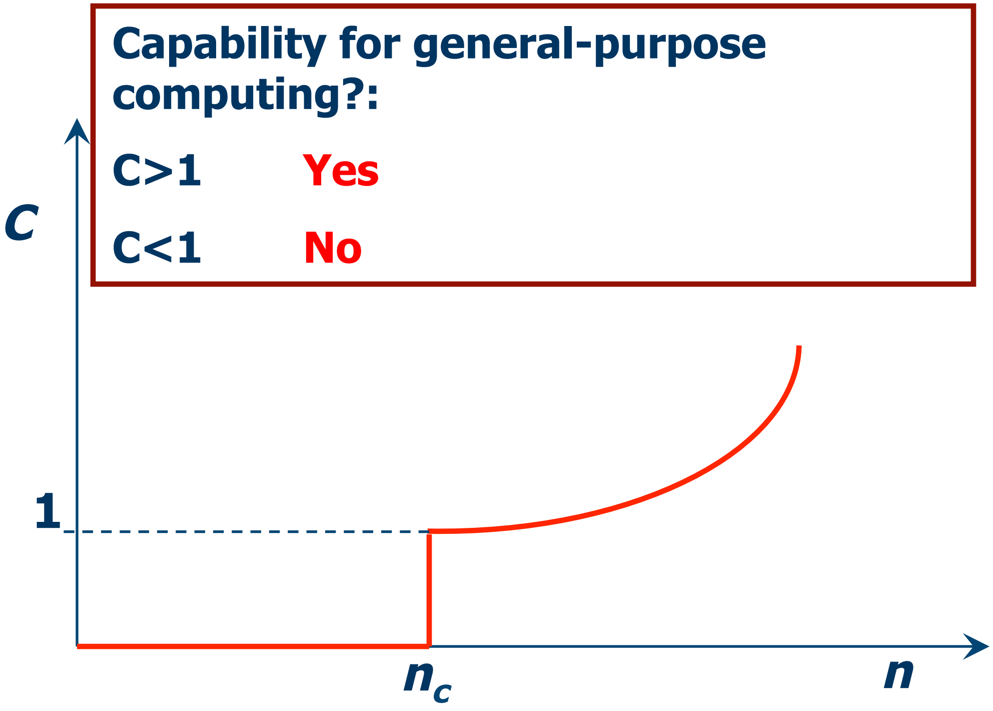
MINIMAL COMPUTER


Rainer Waser (Ed.)
Nanoelectronics and Information Technology
Advanced Electronic Materials and Novel Devices
Third, Completely Revised and Enlarged Edition
WILEY-VCH

R. Cavin, W. Joyner, and T. Noll, **Chapter 22**: *Performance Estimates for Microprocessors: at Technology Limits and in Practice*, in: **Nanoelectronics and Information Technology,** by R. Waser (Ed.) (Wiley 2012)

Von Neumann’s Threshold

“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)**

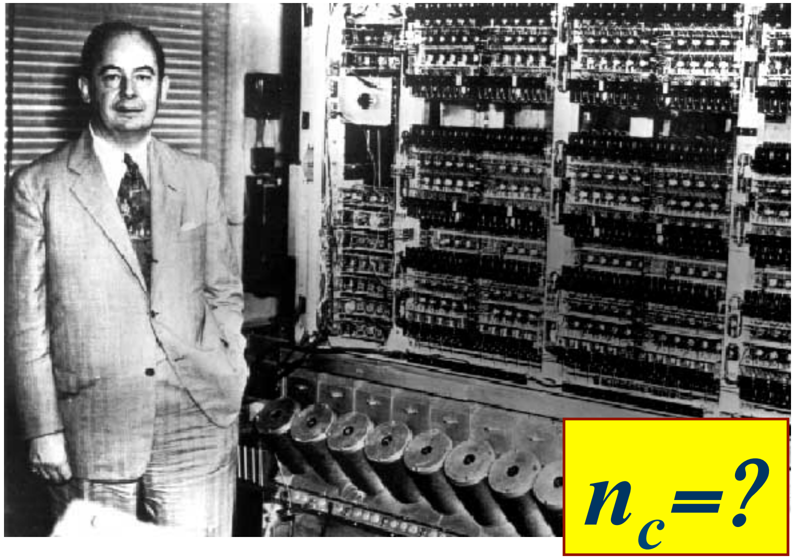


Capability for general-purpose computing?:

C>1 Yes

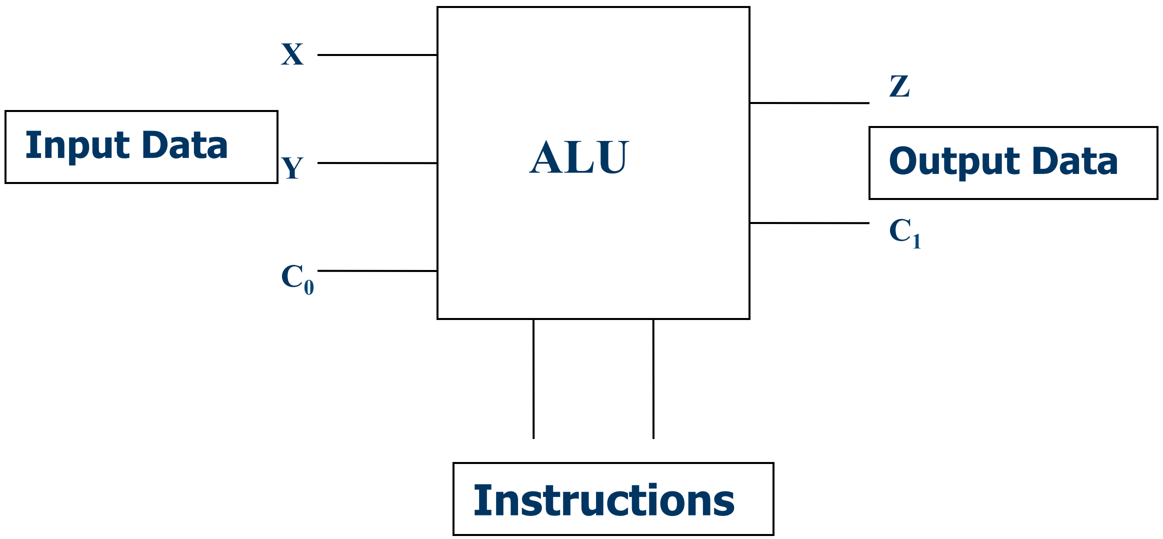
C<1 No

Von Neumann threshold



‘Minimal’ Turing Machine

1-bit ALU example – simple Turing Machine model



The minimal ALU does 2^2=4 operations on two 1-bit X and Y:

Operation 1: X AND Y

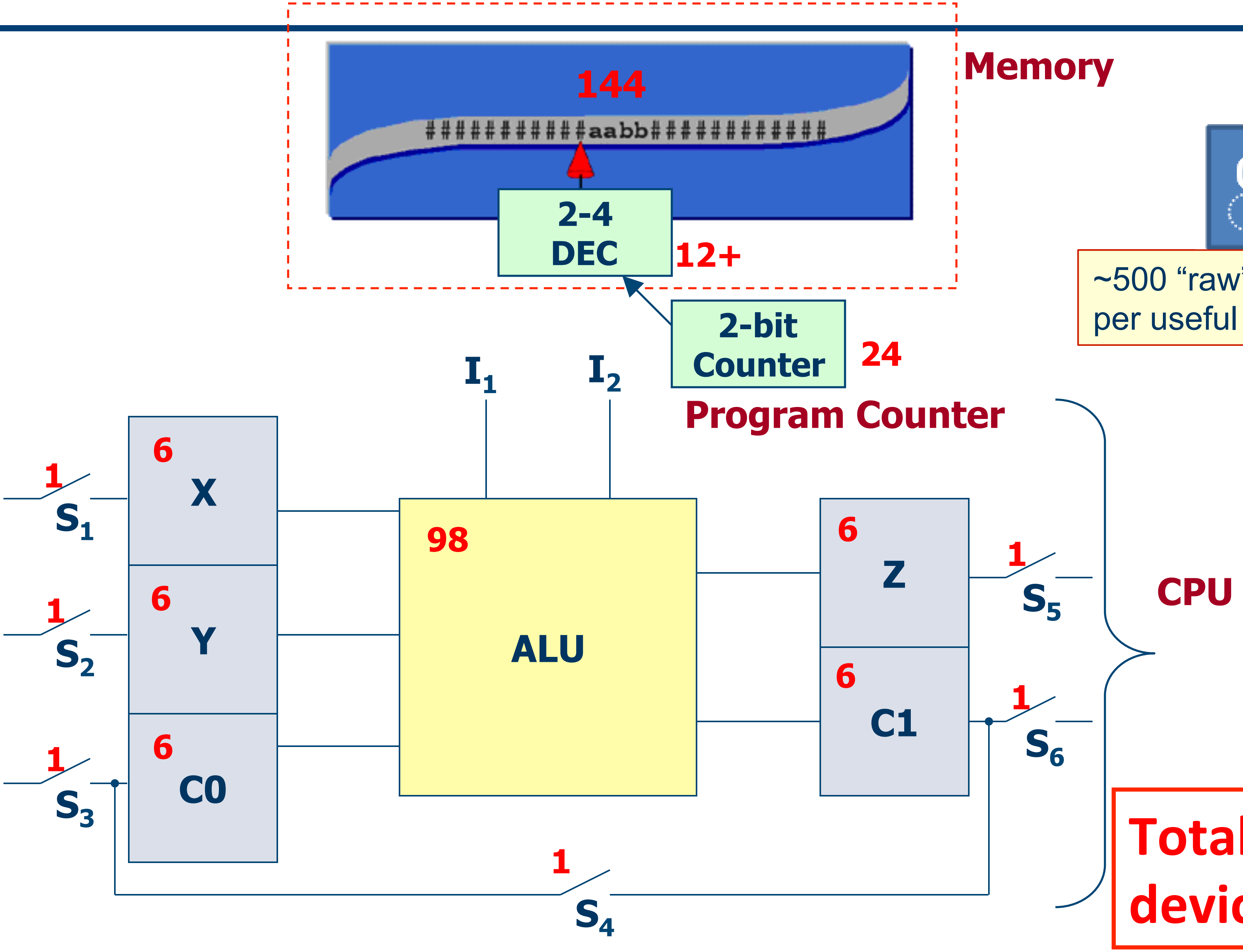
Operation 2: X OR Y

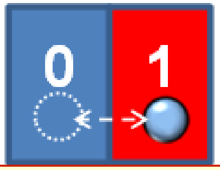
Operation 3: (X+Y)

Operation 4: (X+(NOT Y))

Supports functionally complete set of logic and arithmetic operations

Minimal Turing Machine





~500 "raw" bit transitions per useful bit

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

$$\Pi_{syst} > \Pi_{crit}$$ e.g., 0.5

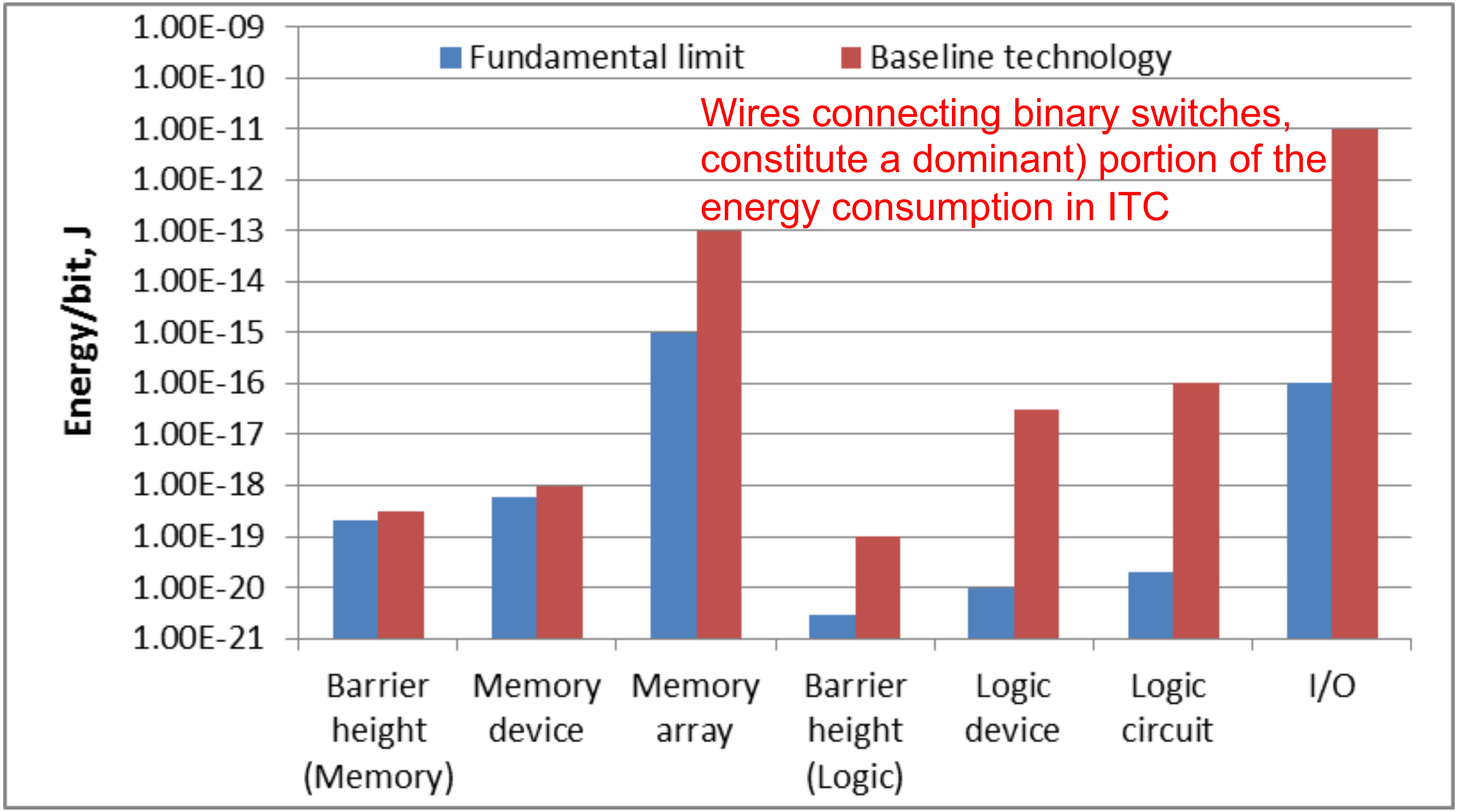
lower boundary

$$\Pi_{err} = 1 - \Pi_{crit}^{\frac{1}{N}} = 1 - 2^{-\frac{1}{320}} = 0.002$$

$$\Pi_{err} = \overset{\textbf{Boltzmann}}{\boxed{\exp(-\frac{E_b}{kT})}} + \overset{\textbf{Heisenberg}}{\boxed{\exp(-\frac{2\sqrt{2m}}{\hbar} a\sqrt{E_b}) - \exp(-\frac{\hbar E_b + 2akT\sqrt{2mE_b}}{\hbar kT})}}$$

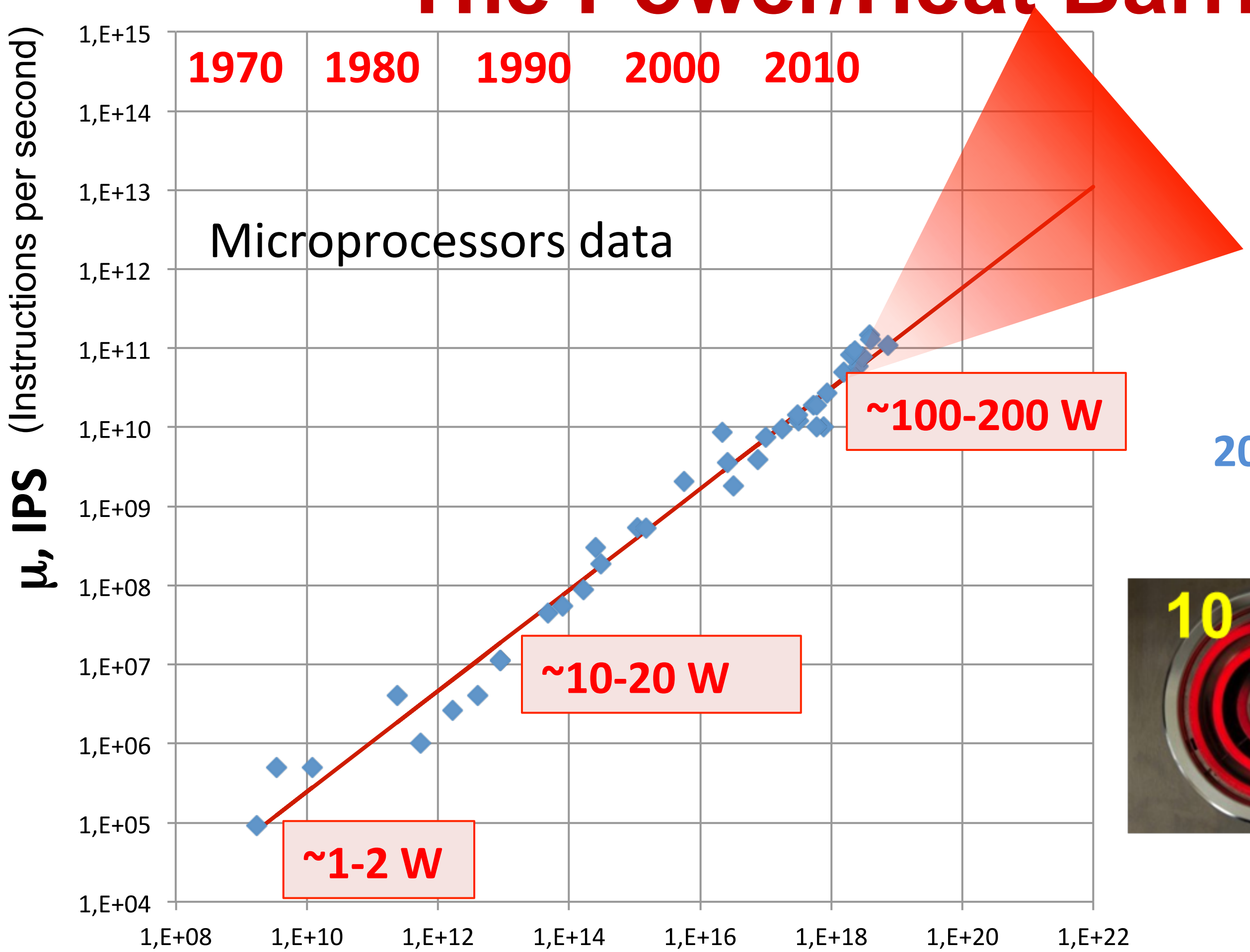
$$\boxed{E_{b_{\min}} \approx k_B T \ln \frac{1}{\Pi_{err}} \approx 6k_B T}$$

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

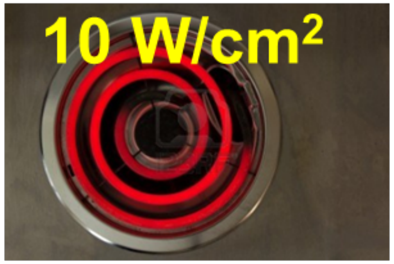


β, bit/s

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

Close to practical limits for cost-effective cooling

2013 microprocessor chip



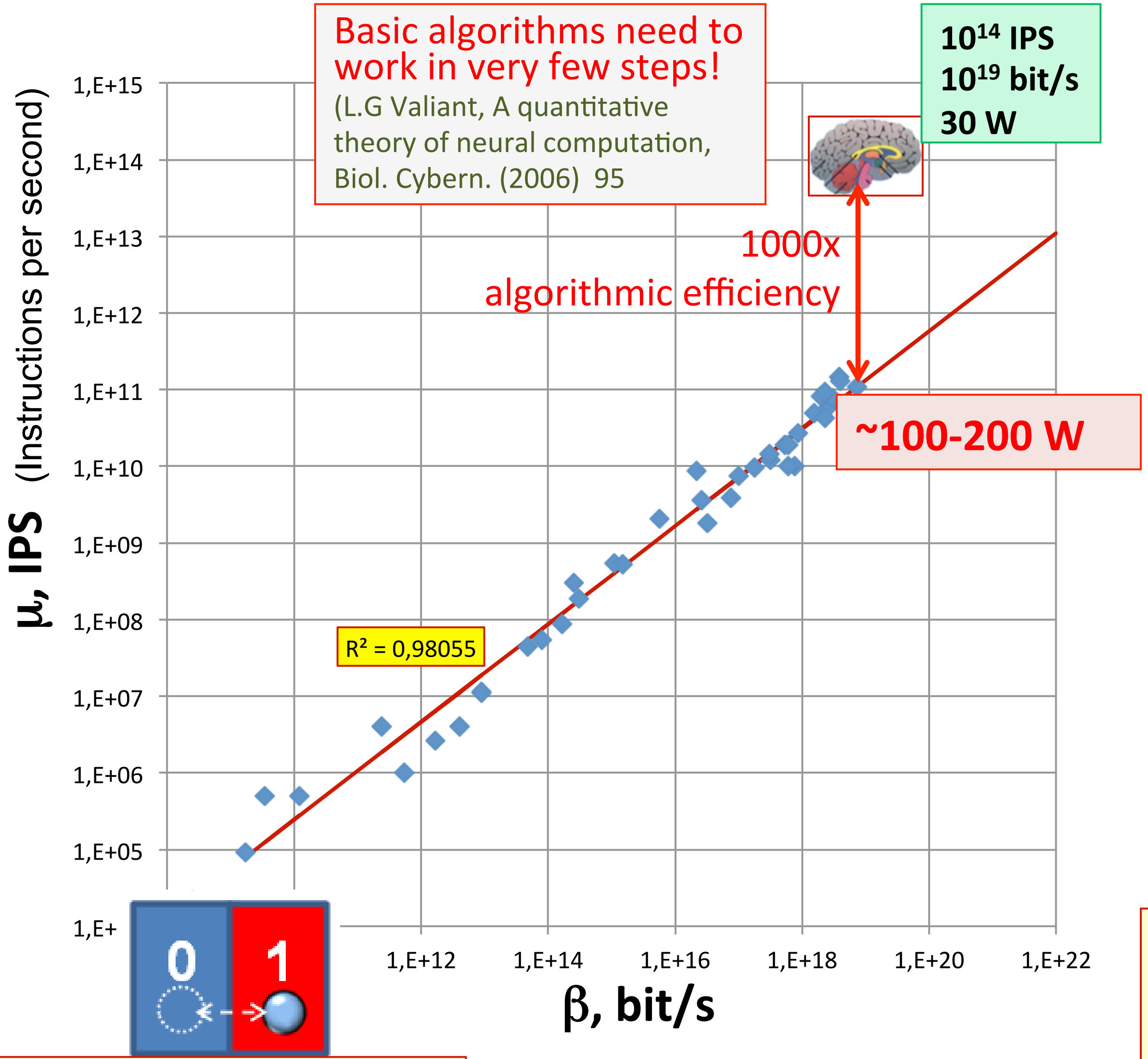
Hot Plate

Energy consumption by ICT is growing

Heat removal is a crucial issue for future computing

New ICT principles for greater energy efficiency needs to be discovered

Benchmark capability μ (IPS) as a function of β (bit/s)



Basic algorithms need to work in very few steps!

(L.G Valiant, A quantitative theory of neural computation, Biol. Cybern. (2006) 95

10^{14} IPS

10^{19} bit/s

30 W

1000x algorithmic efficiency

~100-200 W

~500 "raw" bit transitions per useful bit

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)

What can we learn about information processing from Nature?

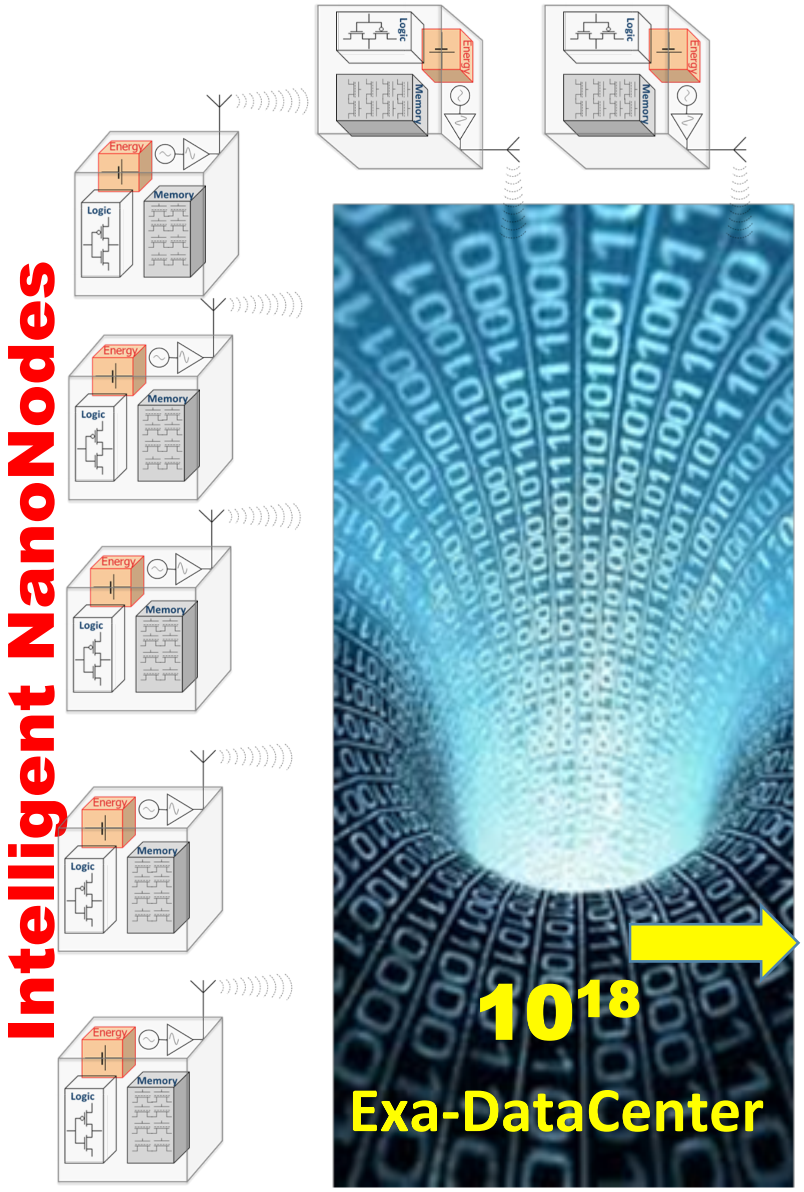
A Thought System:

Ultimate Connectivity: Internet of Nanothings

IoT Grand Challenges

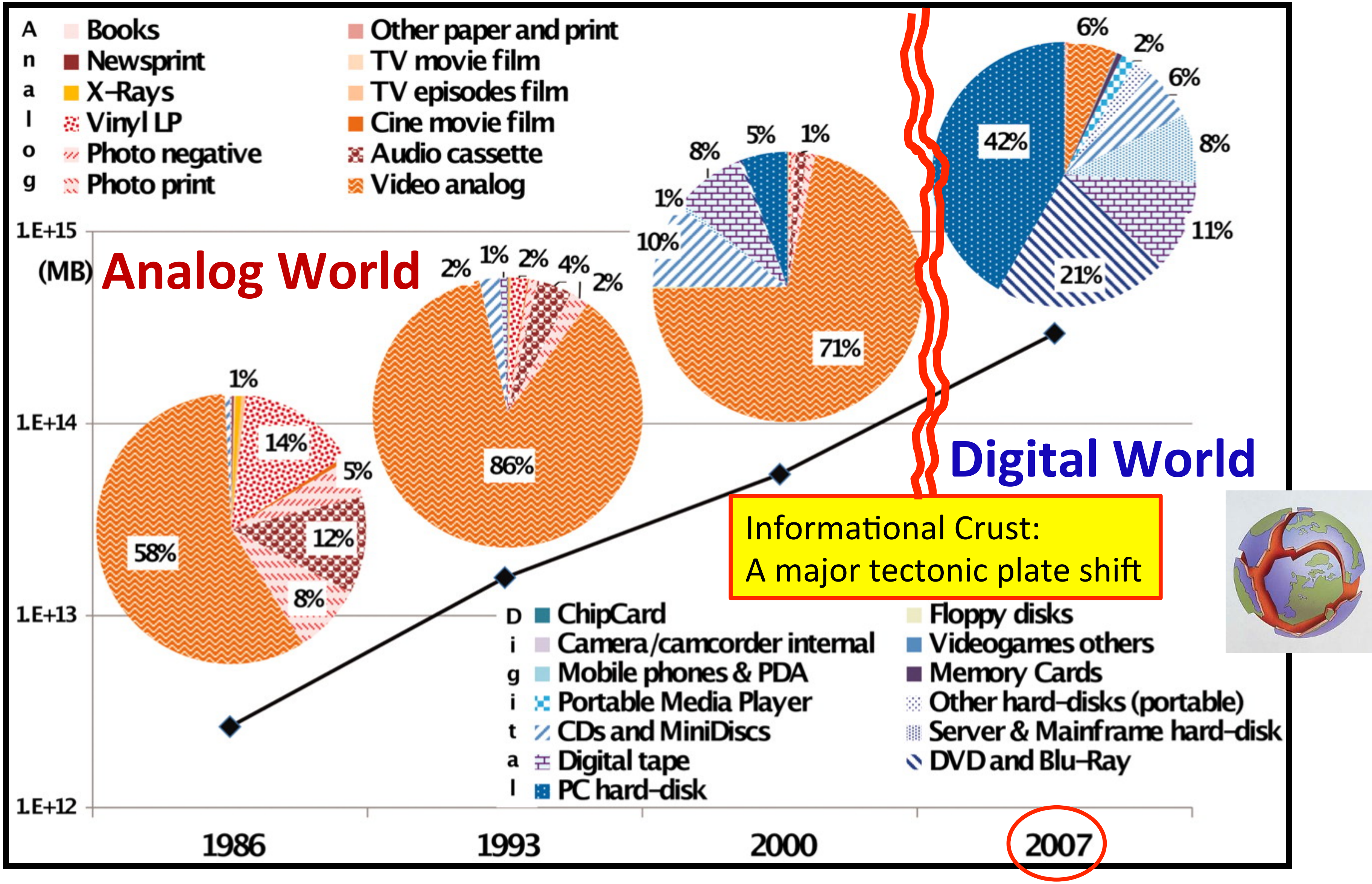
I. Giga-Nano-Tera (Billions of Nanosystems connected in a THz-network)

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

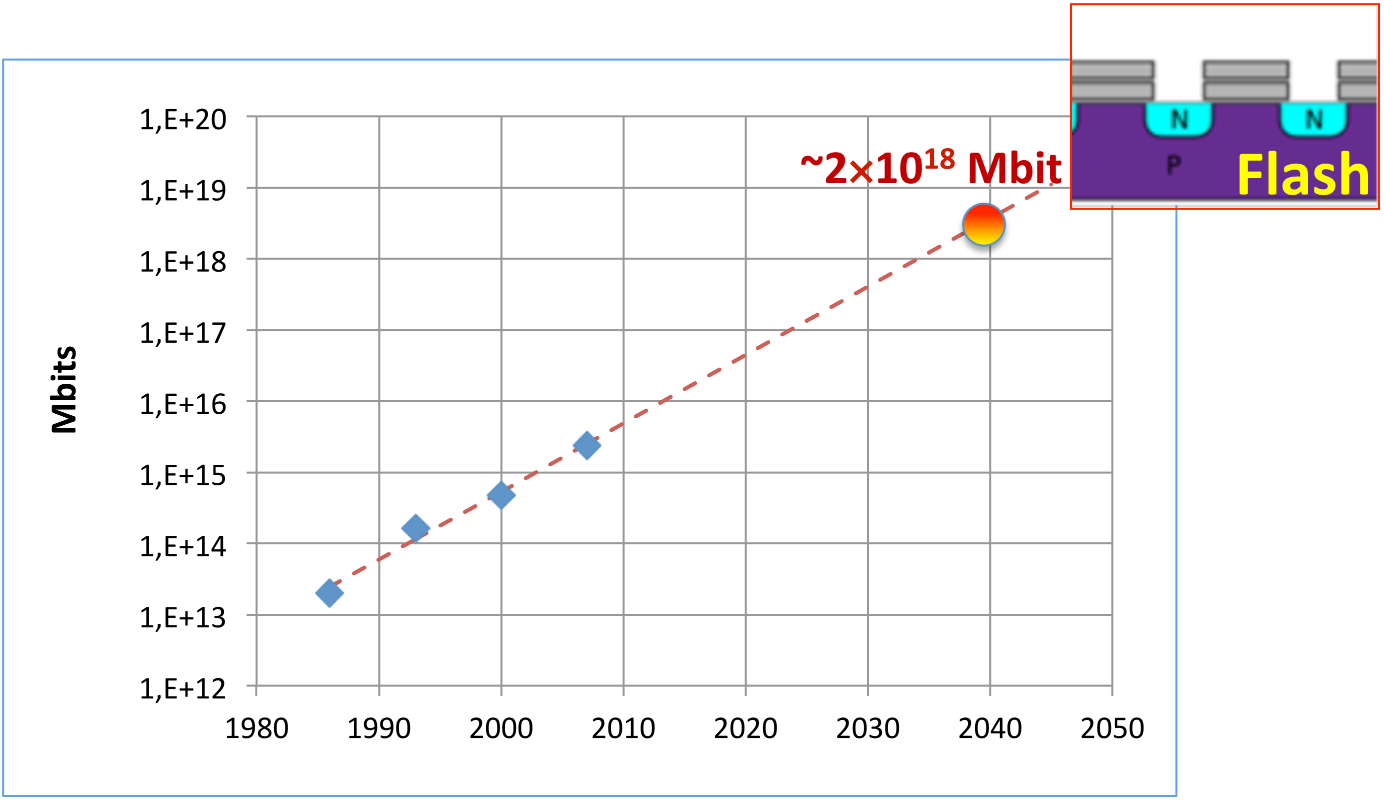


Analog World

Digital World

Informational Crust:
A major tectonic plate shift

Storage Needs in 2040



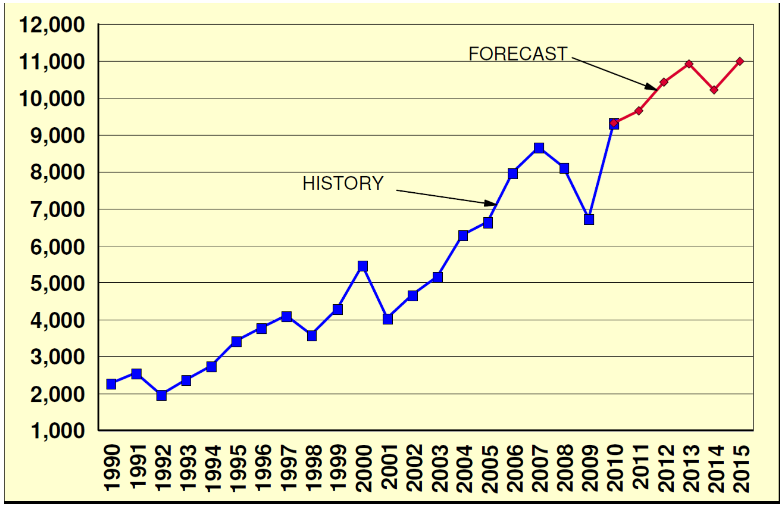
10^{10}kg of *wafer-grade Si*

Projected global supply:

~2×10^7 kg

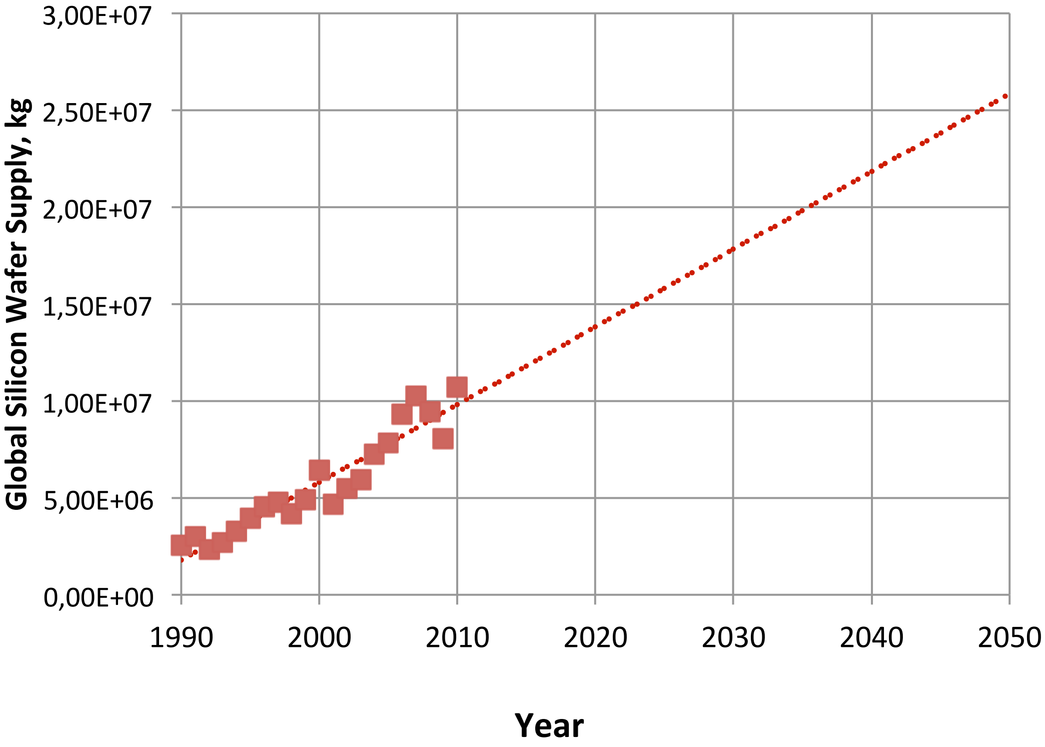
Radical Departures from current baseline technologies may be needed to address the exponential growth in the storage needs

Global Silicon Wafer Supply Trend



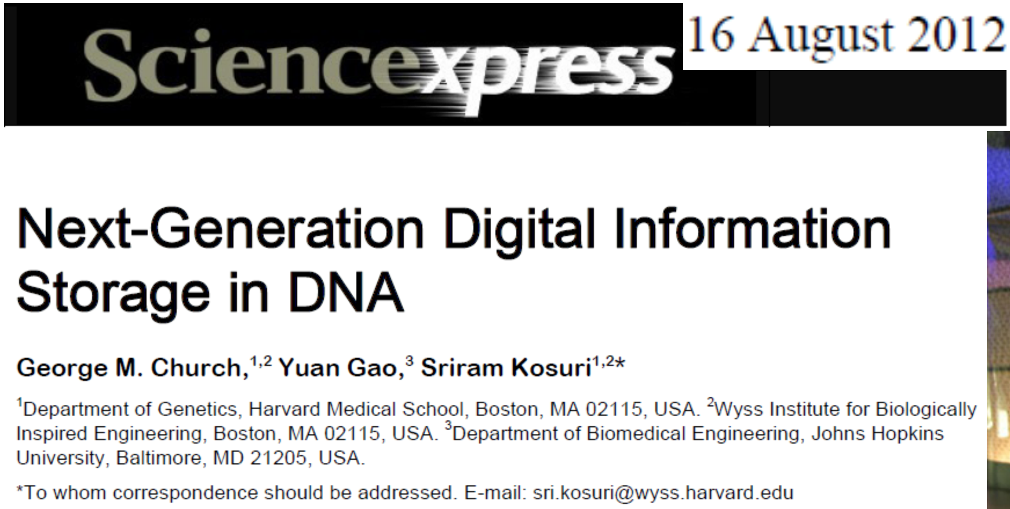
GLOBAL SILICON WAFER SALES FORECAST IN MILLIONS OF SQUARE INCHES

SILICON INDUSTRY 2011
By Richard M. Winegarner, Sage Concepts



Is there enough silicon to support the major tectonic plate shift in the Informational Crust?

Example I: DNA Memory



Sciencexpress 16 August 2012

Next-Generation Digital Information Storage in DNA

George M. Church,[1,2] Yuan Gao,[3] Sriram Kosuri[1,2]*

[1] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA. [2] Wyss Institute for Biologically Inspired Engineering, Boston, MA 02115, USA. [3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed. E-mail: sri.kosuri@wyss.harvard.edu

http://www.wired.com/wiredscience/2012/08/dna-data-storage/

DNA: The Ultimate Hard Drive?

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.

5.27×10^6 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

Recent Disclosures of DNA Memory Capability II

NEW!

7 FEBRUARY 2013 | VOL 494 | NATURE |

Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

Nick Goldman[1], Paul Bertone[1], Siyuan Chen[2], Christophe Dessimoz[1], Emily M. LeProust[2], Botond Sipos[1] & Ewan Birney[1]

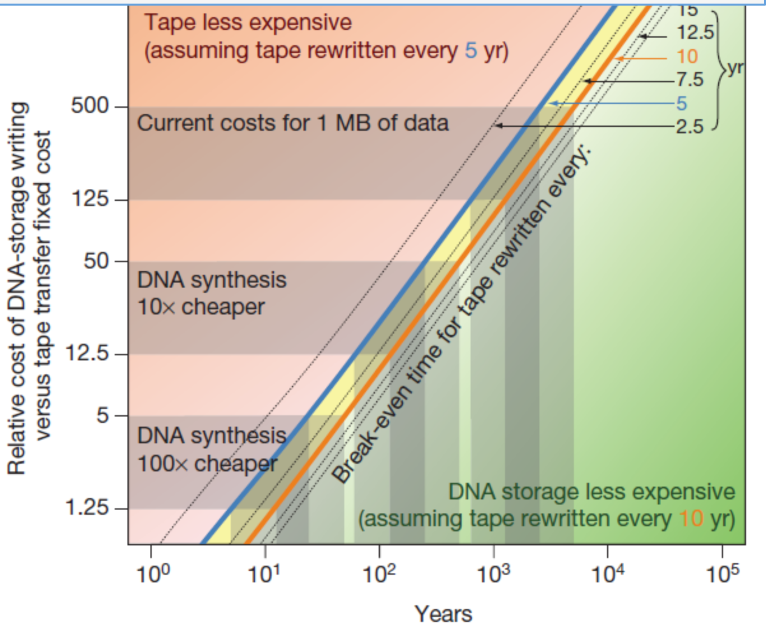
European Molecular Biology Laboratory
Heidelberg Grenoble
Hamburg Hinxton Monterotondo

Agilent Technologies

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.

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

So are you to my thoughts
as food to life,
Or as sweet-season'd
showers are to the ground;
And for the peace of you
I hold such strife
As 'twixt a miser and his
wealth is found;
Now proud as an enjoyer, and anon
Doubting the filching age will steal
his treasure;
Now counting best to be with you alone,
Then better'd that the world may see
my pleasure:
Sometime all full with feasting on
your sight,
And by and by clean starved for a look;
Possessing or pursuing no delight,
Save what is had or must from you
be took.
Thus do I pine and surfeit day by day,
Or gluttoning on all, or all away.



Memory Only

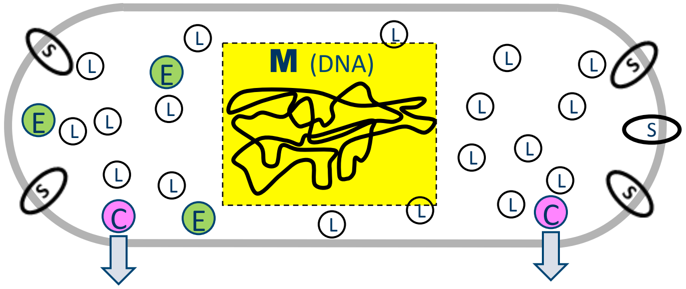
Given:

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

DNA memory µsystem



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):

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

Characteristic energy per bit (system-level):

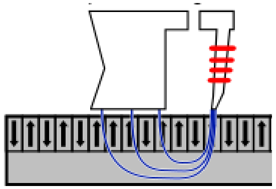
$$P_{DNA} < \frac{1.4 \cdot 10^{-13}}{2.4 \cdot 10^{-3}} = 5.8 \cdot 10^{-11} \frac{W}{GByte}$$

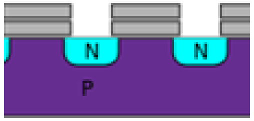
DNA-Inspired Memory

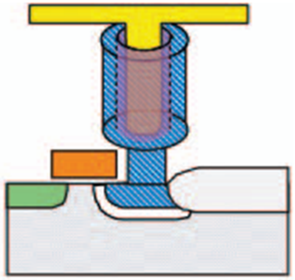
(On-Going Project with Micron Technology)

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



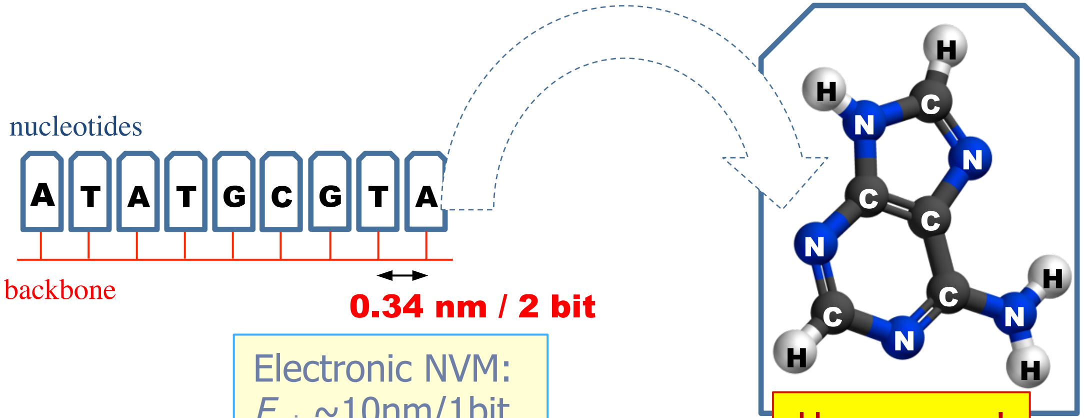




	HardDiskDrive	NAND flash	DRAM	DNA in cell
Read/Write latency	3-5 ms/bit	~100μs/bit	<10 ns/bit	<100μs/bit
Endurance (cycles)	unlimited	10^4-10^5	unlimited	unlimited
Retention	>10 years	~10 years	64 ms	>10 years
ON power (W/GB)	~0.04	~0.01-0.04	0.4	$<10^{-11}$
Aerial Density	~ 10^{11} bit/cm^2	~ 10^{10} bit/cm^2	~ 10^9 bit/cm^2	n/a
Volumetric Density	n/a	10^{16} bit/cm^3	~10^{13} bit/cm^3	10^{19} bit/cm^3

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), cytonine (C), guanine (G), and thymine (T).



Electronic NVM:
$F_{\min}$~10nm/1bit

Heavy mass!

$$a_{\min} = \frac{h}{2\sqrt{2mE_b}}$$

DNA is NOT a read-only memory

DNA memory operations

READ

- *Multi-access capability* by distinct computing units

WRITE

Vertical gene transfer - exact copying of the parental DNA

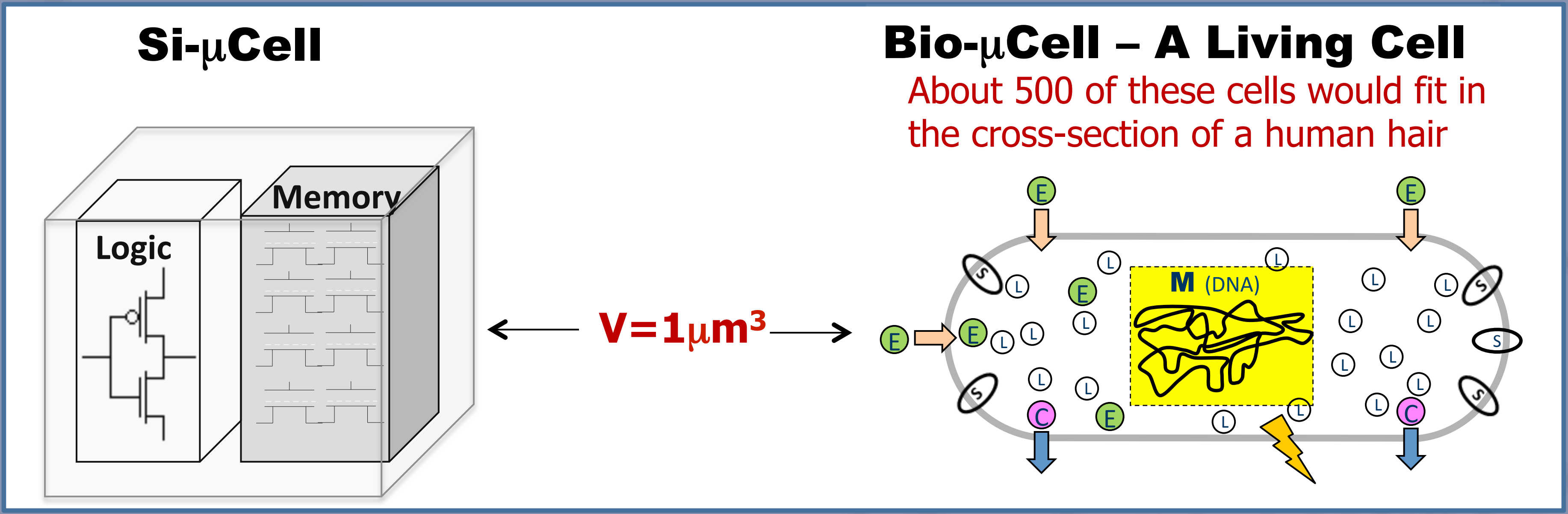
Lateral (horizontal) gene transfer :

(1) direct uptake (‘swallowing’) of a naked DNA by a cell,

(2) by a virus,

(3) by direct physical contact between two cells.

Nature Has Been Processing Information for a Billion Years



Si-µCell

Bio-µCell – A Living Cell

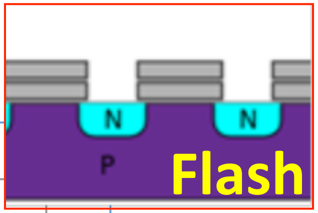
About 500 of these cells would fit in the cross-section of a human hair

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:

Memory:	1000x more
Logic:	>10x more
Power:	1000,000x **less**
Algorithmic efficiency:	1000x more

Lower-hanging fruit?

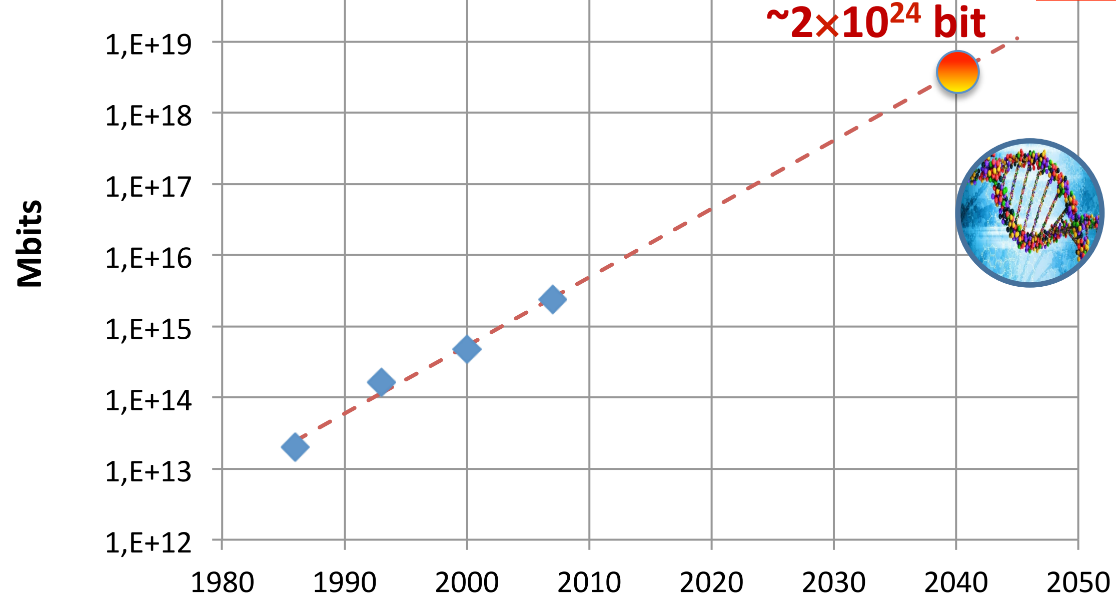
Storage Needs in 2040



10^{10}kg of *wafer-grade Si*

Projected global supply:

~2×10^{7} kg



~2×10^{24} bit

1kg of *DNA*

A bucket of DNA address the BigData storage capacity challenge?

Transmission Challenge

BigData Communication Challenge

Example: Fiber optic cables technology

~2×10^{24} bit

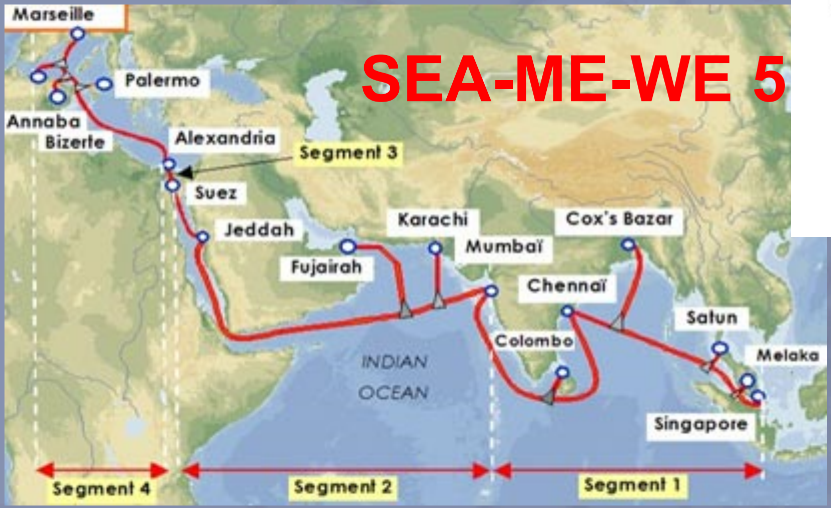
24 Tbit/s

3000 years

To be completed in 2016:

24 Tbit/s

Close to practical limits of current communication technologies

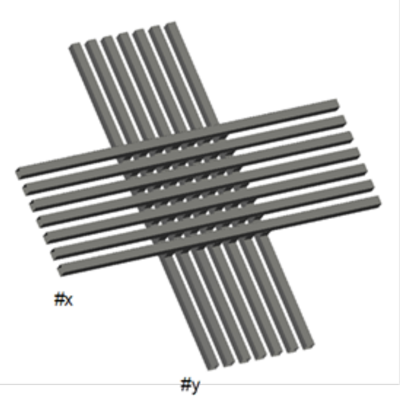


Conclusions

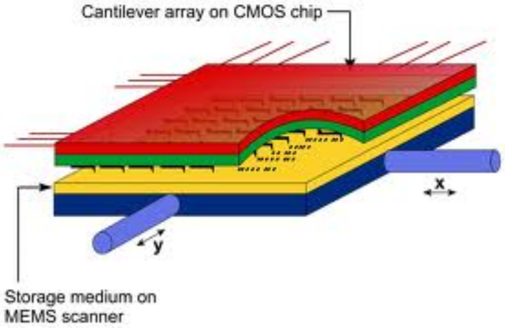
- ICT and Energy devices have a common soul
  - The universal principle of operation of all these elements is the creation and management of charge separation
  - Controllable energy barriers is a fundamental component in all ICT & Energy devices
- Memory and Communication are the main factors of energy 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-μCell 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 (~ 1nm memory elements)
  - Example: the IBM 'Millipede'



'Millipede'


Cantilever array on CMOS chip
x
y
Storage medium on MEMS scanner