



Nanoelectronic Scaling Tradeoffs: What does Physics have to say?

Victor V. Zhirnov
Semiconductor Research Corporation

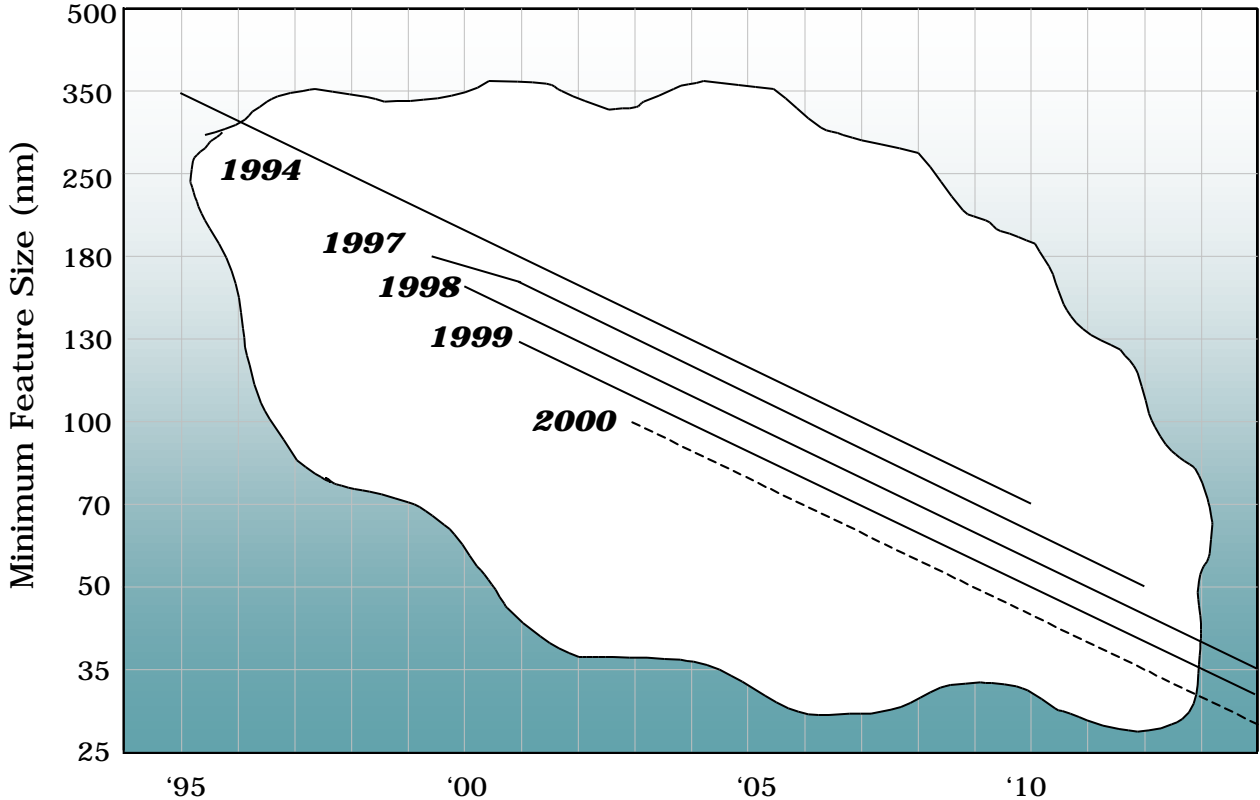
Outline

- ◆ Technology Roadmap on Semiconductors
- ◆ Fundamentals of information processing,
- ◆ Fundamental limits to scaling
- ◆ Thermal Limits
- ◆ Message: *We suggest that the benefits from nanoelectronics research may, in the --*
- ◆ *Short term* lie with the invention of new structures, materials and processes that extend the CMOS technology platform
 - ❖ Radical thermal solutions are needed
- ◆ *Long term* enable invention of entirely new information processing technologies

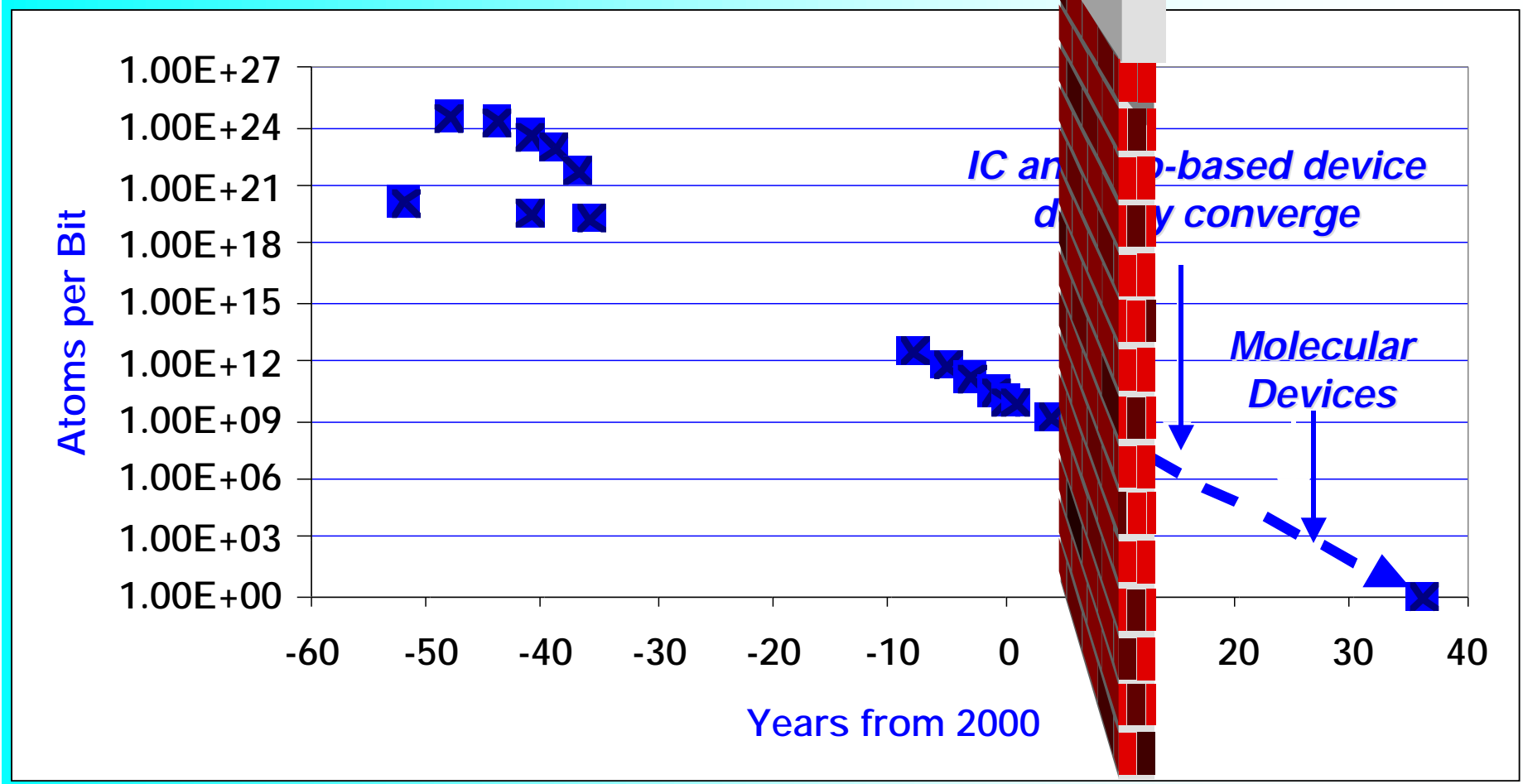
International Technology Roadmap on Semiconductors

- ◆ A very detailed industrial perspective on the future requirements for micro/nano electronic technologies
 - ❖ Goal is to continue exponential gains in performance/price for the next fifteen years
- ◆ Built on worldwide consensus of leading industrial, government, and academic technologists
- ◆ Provides guidance for the semiconductor industry and for academic research worldwide
- ◆ Content: critical requirements and judgment of status
- ◆ Projects that by 2016, half-pitch spacing of metal lines will be 22 nanometers and device gate lengths will be 9 nanometers

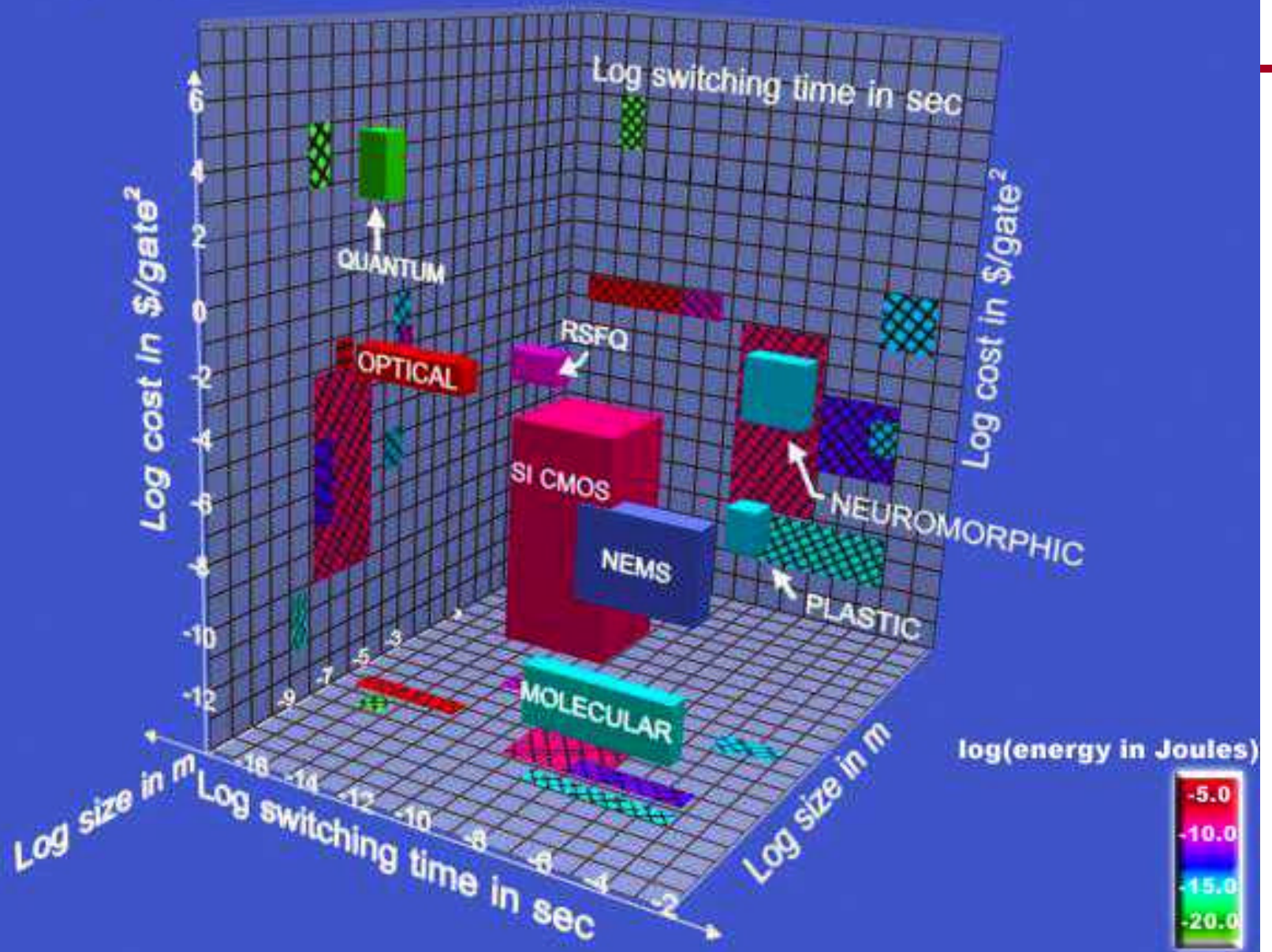
Moore's Law: Minimum Feature Size



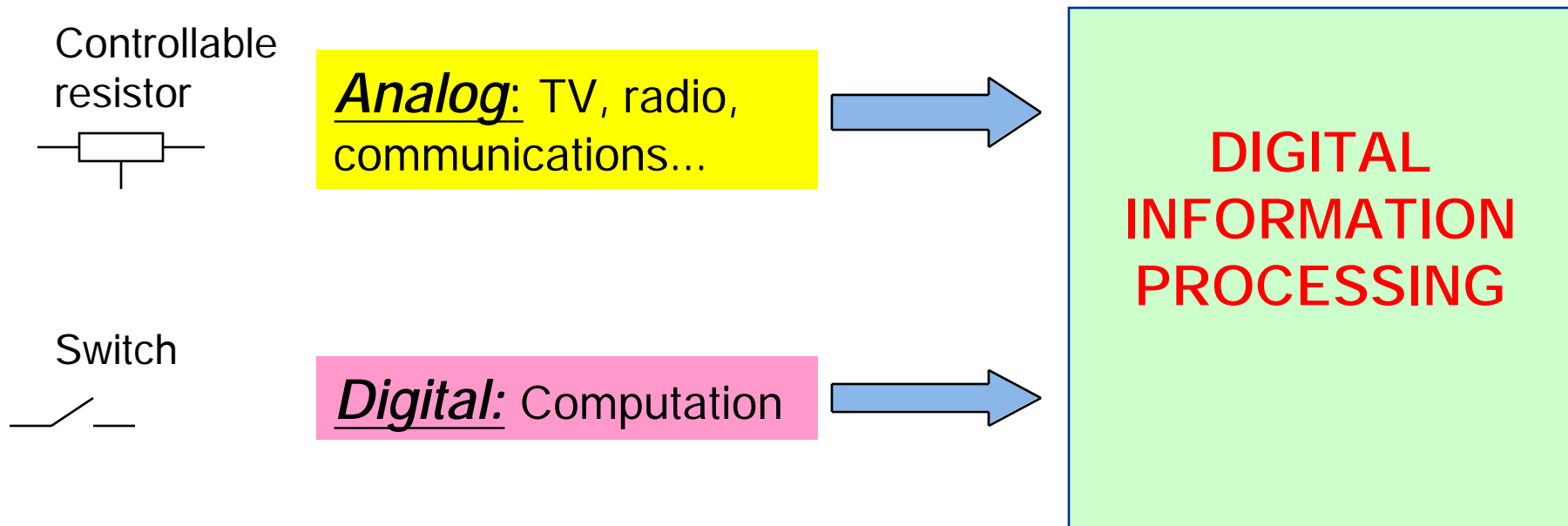
*Devices will soon be on the
"molecular" or "atomistic" scale*



Emerging Technology Parametrization



Evolution of Electronics



General Purpose Computer (GPC) accepts arbitrary types of data and sets of instructions to perform arbitrary tasks of transmission, processing, and storing the information

Parameters of GPC:

- ◆ Number of components (integration density/functional complexity)
- ◆ Speed
- ◆ Energy consumption

What is Information?

Information is...

- ◆ ...Measure of distinguishability
- ◆ ...A function of a priory probability of a given state or outcome among the universe of possible states.

$$I = K \ln N$$

$$N_{\min} = 2$$

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

$$1 = K \ln 2$$

The binary choice: YES/NO, 1/0 etc

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

$$N = 2^n$$

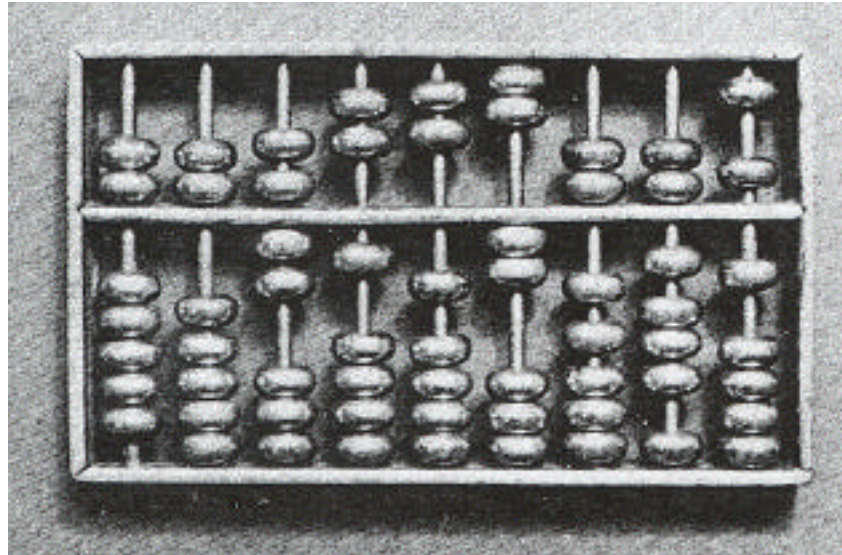
$$I = K \ln 2^n = \frac{1}{\ln 2} n \ln 2 = n = \log_2 N$$

Constituents of the Information Theory

- ◆ Constituents of the Information Theory
 - Sender and recipient
 - Symbols (microstates) as elementary units of information
- ◆ Information carriers

Information is physical!

The Abacus, an ancient digital calculating device

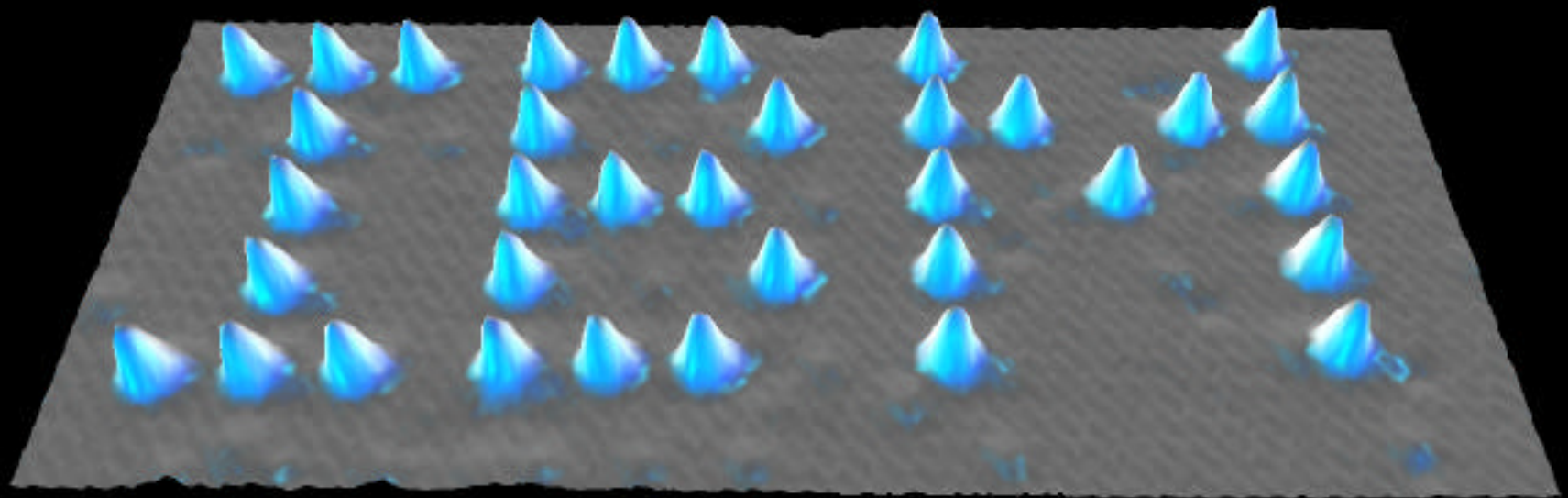


Information is represented in digital form

Each column denotes a decimal digit

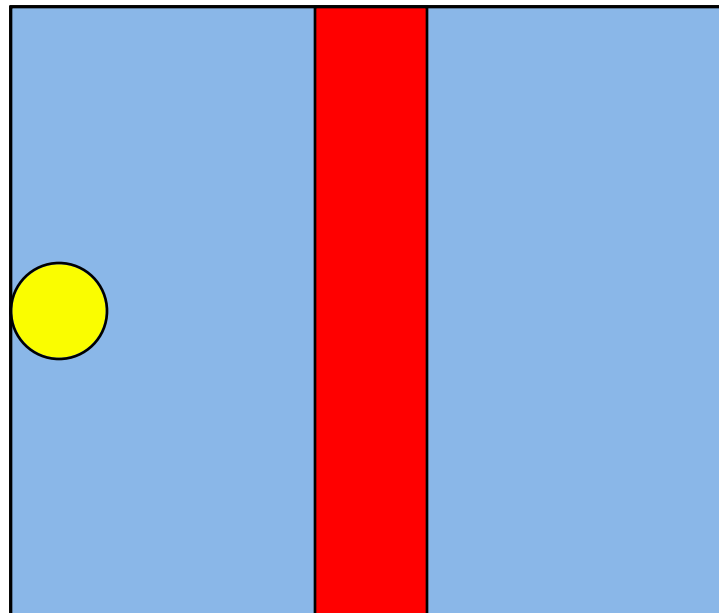
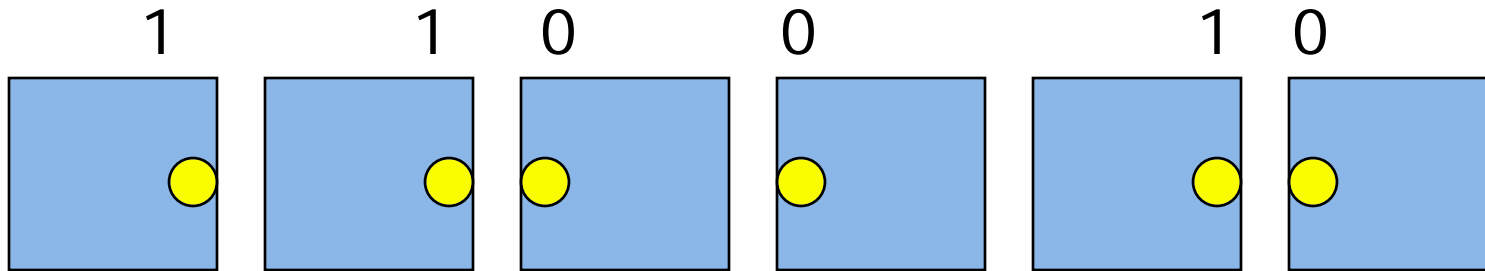
Binary representation: two possible positions for each bead

A bead in the abacus is a memory device, not a logic gate

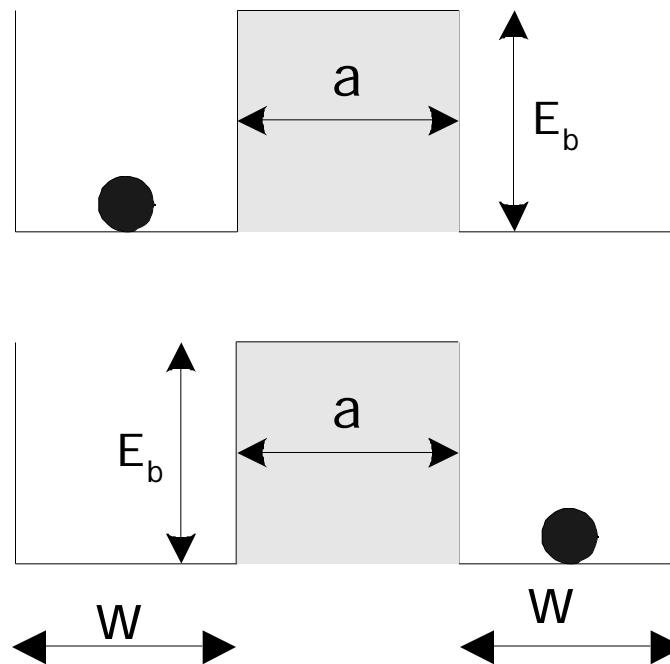


Source: IBM

Particle Location is an Indicator of State



Two-well bit



A physical system as a computing medium

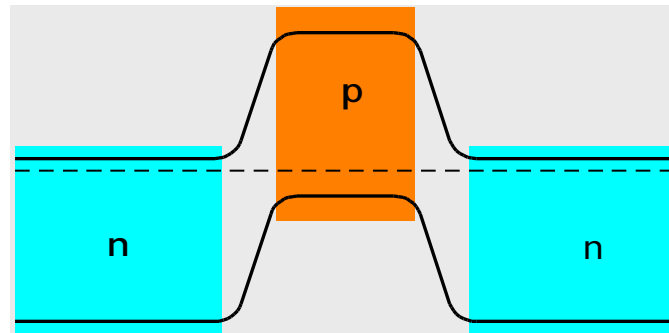
- ◆ We need to create a bit first. Information processing always requires physical carrier, which are material particles.
- ◆ First requirement to physical realization of a bit implies creating *distinguishable* states within a system of such material particles.
- ◆ The second requirement is *conditional* change of state.
- ◆ The properties of *distinguishability* and *conditional change of state* are two fundamental properties of a material subsystem to represent information. **These properties can be obtained by creating *energy barriers* in a material system.**

Kroemer's Lemma of Proven Ignorance

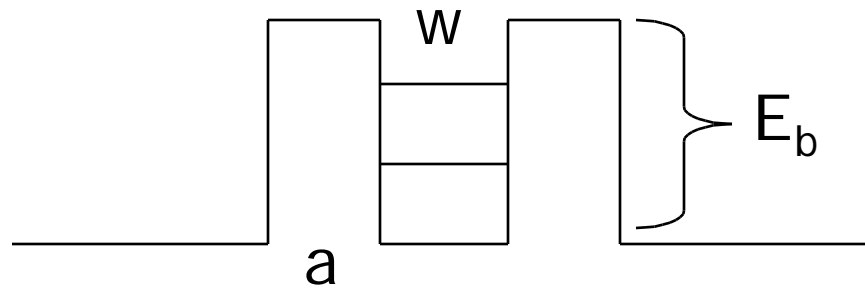
- ◆ If in discussing a semiconductor problem, you cannot draw an Energy-Band-Diagram, this shows that *you don't know what are you talking about*
- ◆ If you can draw one, but don't, then *your audience won't know what are you talking about*

Barrier engineering in semiconductors

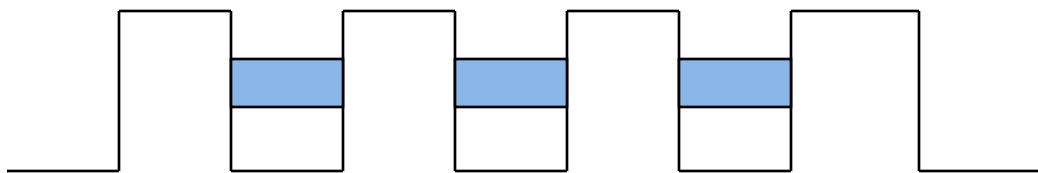
By doping, it is possible to create a built-in field and energy barriers of controllable height and length within semiconductor. It allows one to achieve conditional complex electron transport between different energy states inside semiconductors that is needed in the physical realization of devices for information processing.



Heterojunction barriers



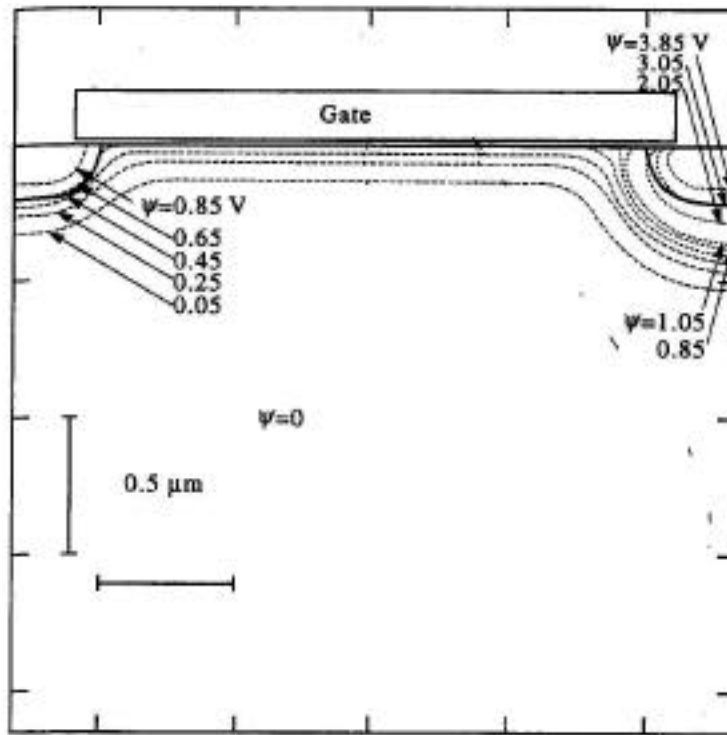
Double barrier



Superlattice

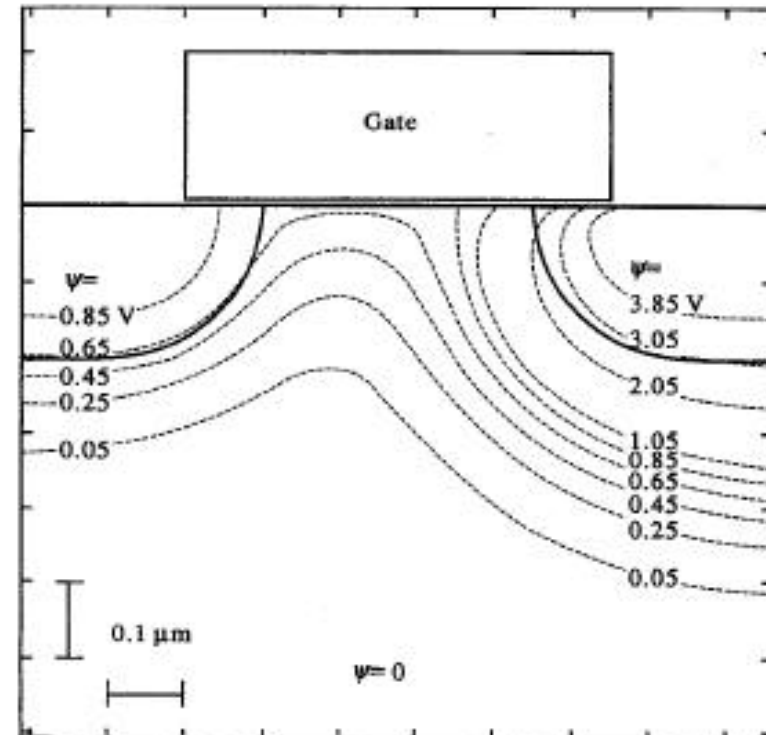
Example: Field Effect Transistor

Long Channel



(a)

Short Channel



(b)

It is possible to derive MOSFET I-V equation from the two-well one-barrier model

Ideal von Neumann's Computer

Designers and Users want:

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

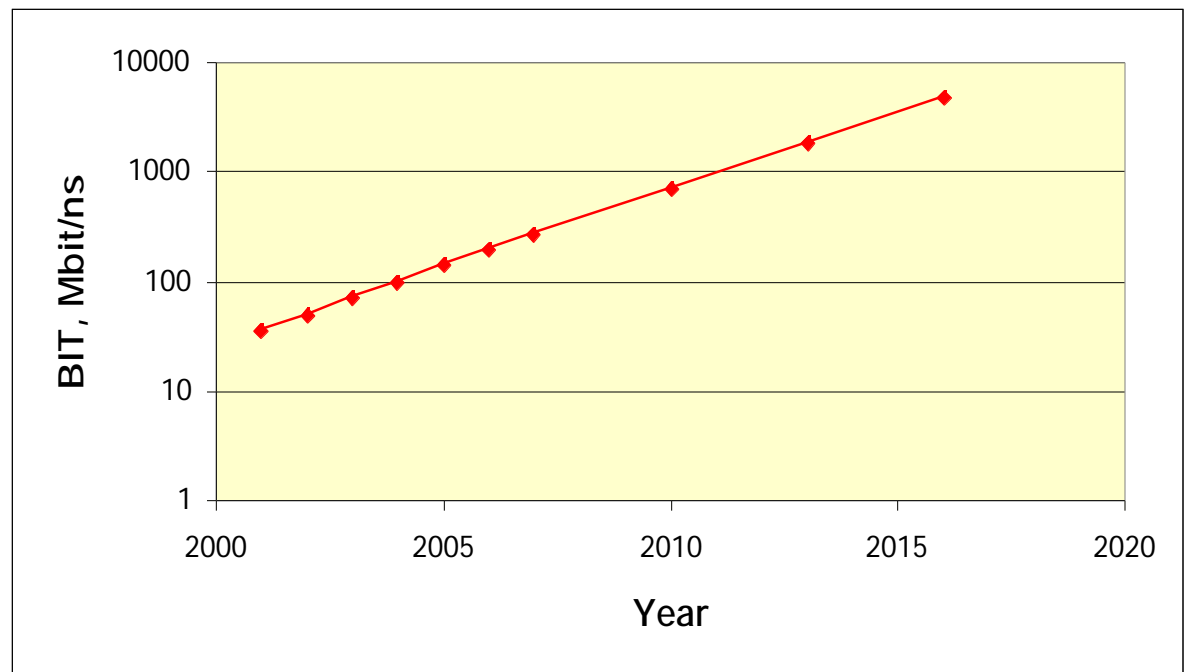
Binary Information Throughput (BIT)

BIT is the maximum number of binary transition per unit time

$$BIT = n_{bit} f$$

- one measure of computational capability

n_{bit} – the number of binary states (e.g. transistors)
 f – switching frequency



Energetics of Computation

$$P = E_{bit}nf$$

Requirements for an ideal computer:

(integration density)

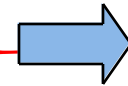
$n = \max$

(switching frequency)

$f = \max$

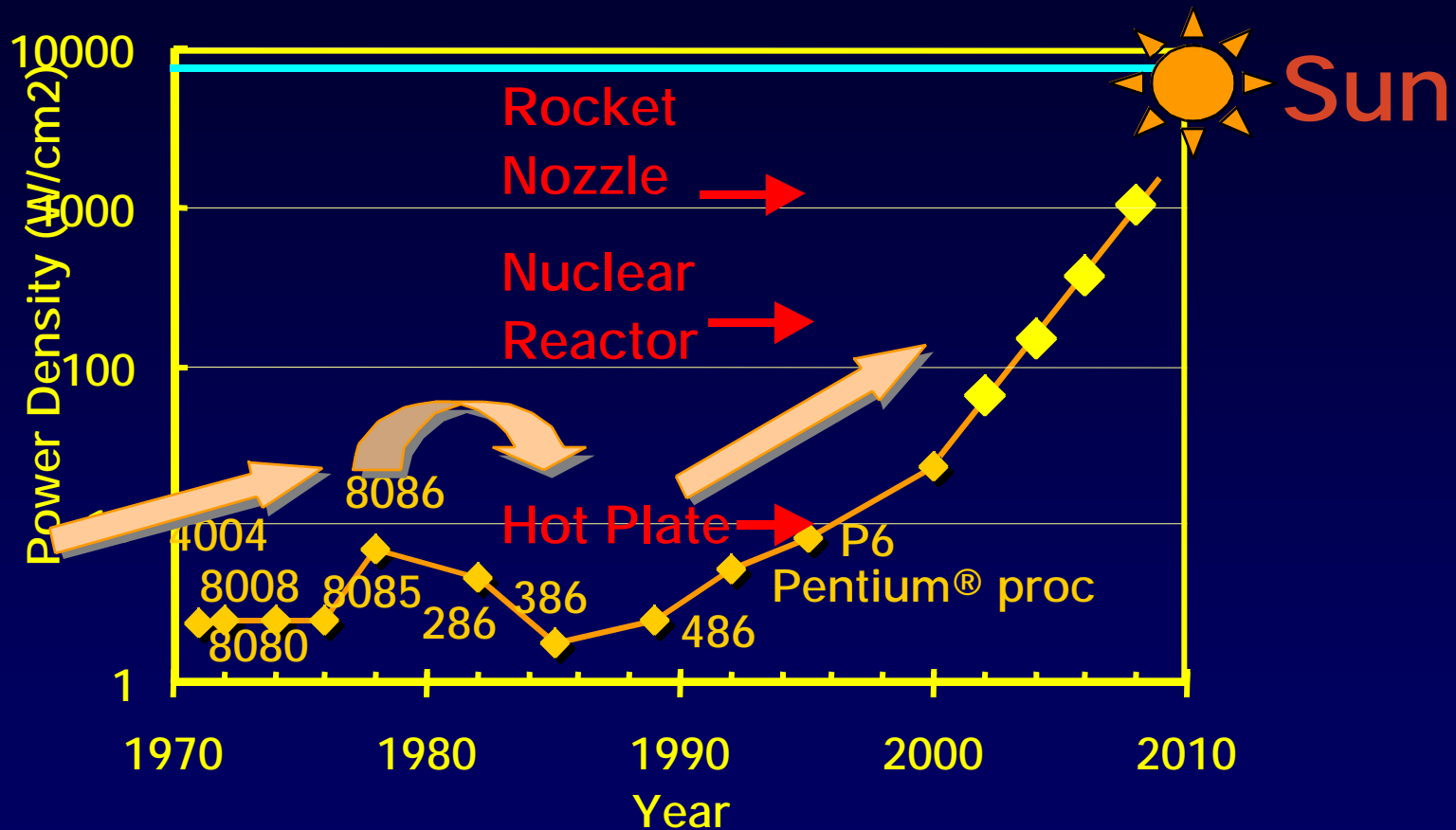
(power)

$P = \min$



BIT = max

Power density will increase



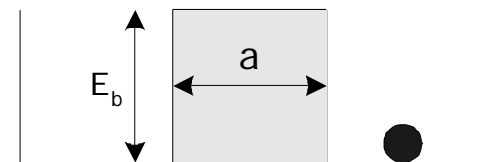
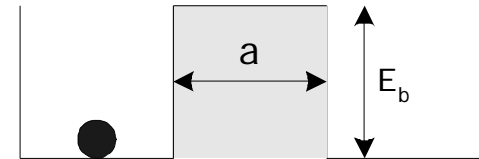
Power density too high to keep junctions at low temp

Lowest Barrier: Distinguishability Barrier

Distinguishability D implies low probability Π of spontaneous transitions between two wells (error probability)

$$D = \max, \Pi = 0$$

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



Classic distinguishability:

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

Minimum distinguishable barrier: $\Pi = 0.5$

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



$$E_b = kT \ln 2$$

Shannon - von Neumann - Landauer limit

Smallest Size: The Heisenberg Barrier

$$\begin{array}{ccc} x & p & \hbar \\ E & t & \hbar \end{array} \quad \longrightarrow \quad \begin{array}{l} a_{crit} = \frac{\hbar}{\sqrt{2mE_b}} \\ t_{min} = \frac{\hbar}{E_b} \end{array}$$

$$E_b = kT \ln 2$$

Classic and Quantum Distinguishability @ $\Pi=0.5$

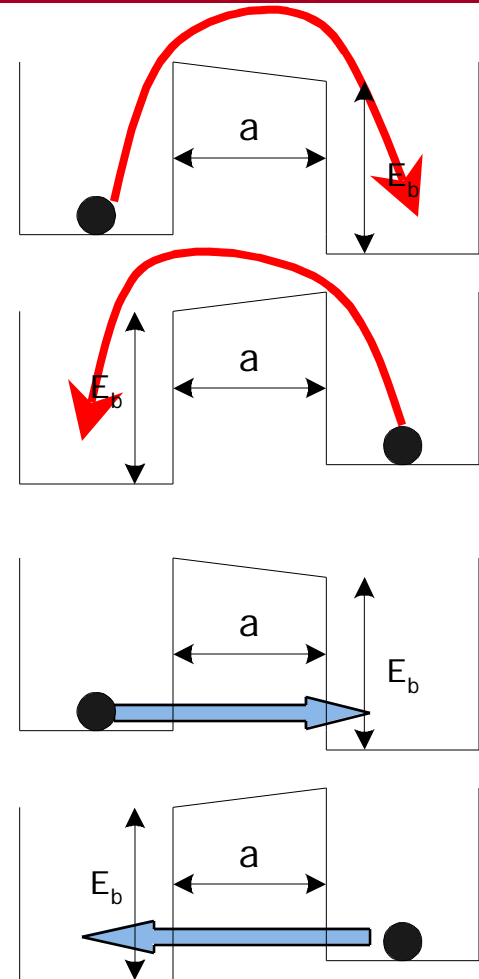
$$P_{\text{classic}} = \exp\left(-\frac{E_b}{k_B T}\right)$$

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

WKB: (Tunneling)

$$P_{\text{quantum}} = \exp\left(-\frac{2\sqrt{2m}}{\hbar} a \sqrt{E_b}\right)$$

$$E_b^{\text{min}} = \frac{\hbar^2 \ln^2 2}{8ma^2}$$

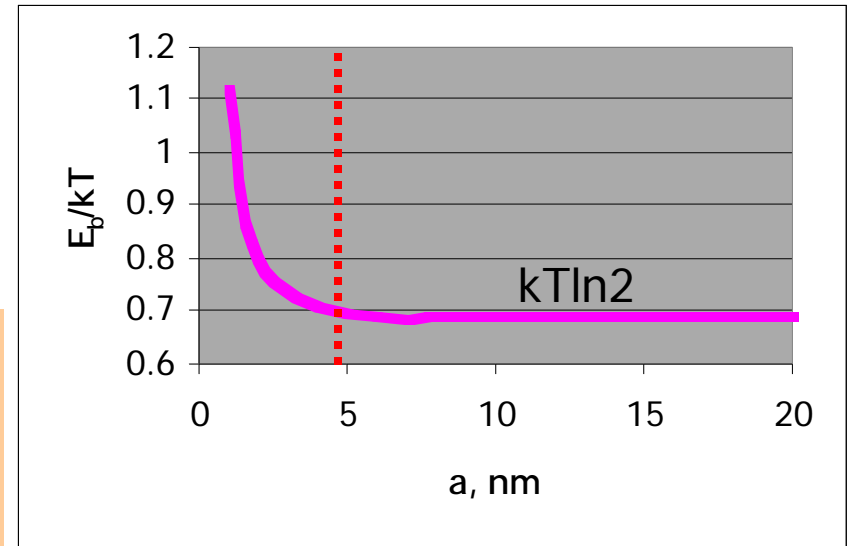


Total Distinguishability @ $\Pi=0.5$

$$\begin{aligned}
 \text{error} &= \text{classic}^+ \text{ quantum}^- \text{ classic} \text{ quantum}^= \\
 &= \exp\left(-\frac{E_b}{kT}\right) + \exp\left(-\frac{2\sqrt{2m}}{\hbar} a\sqrt{E_b}\right) - \exp\left(-\frac{\hbar E_b + 2akT\sqrt{2mE_b}}{\hbar kT}\right)
 \end{aligned}$$

Generalized expression for the minimum energy barrier to create a bit

$$E_b^{\min} = kT \ln 2 + \frac{\hbar^2 (\ln 2)^2}{8ma^2}$$



Least Energy Computer

1) Minimum distance between two distinguishable states (Heisenberg)

$$x_{\min} = a = \frac{\hbar}{\sqrt{2mkT \ln 2}} = 1.5 \text{ nm}(300 \text{ K})$$

2) Minimum state switching time (Heisenberg)

$$t_{st} = \frac{h}{2kT \ln 2} = 1.2 \times 10^{-13} \text{ s}(300 \text{ K})$$

3) Maximum gate density

$$n = \frac{1}{x_{\min}^2} = 4.6 \times 10^{13} \frac{\text{gate}}{\text{cm}^2}$$

4) Maximum binary throughput

$$BIT_{\max} = 2m \frac{(kT \ln 2)^2}{\hbar^3} = 10^7 \frac{\text{Tbit}}{\text{ps}}$$

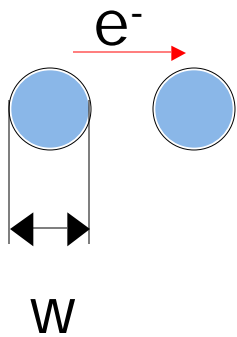
Total Power Consumption at Minimal Energy per bit - $\{kT\ln(2)\}$

$$P_{bit} = \frac{E_b}{t_{sc}} = \frac{kT\ln 2}{t_{sc}} = \frac{2}{h} (kT\ln 2)^2 \quad P_{chip} = nP_{bit}$$

$$P_{chip} = 4.74 \times 10^6 \frac{W}{cm^2} \quad T=300 \text{ K}$$

The circuit would vaporize when it is turned on!

Single Electron Devices Don't Avoid the Power Problem



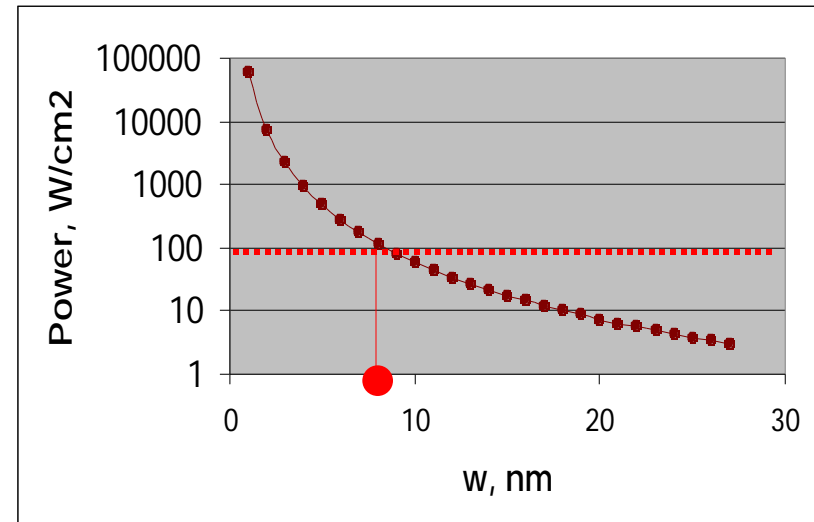
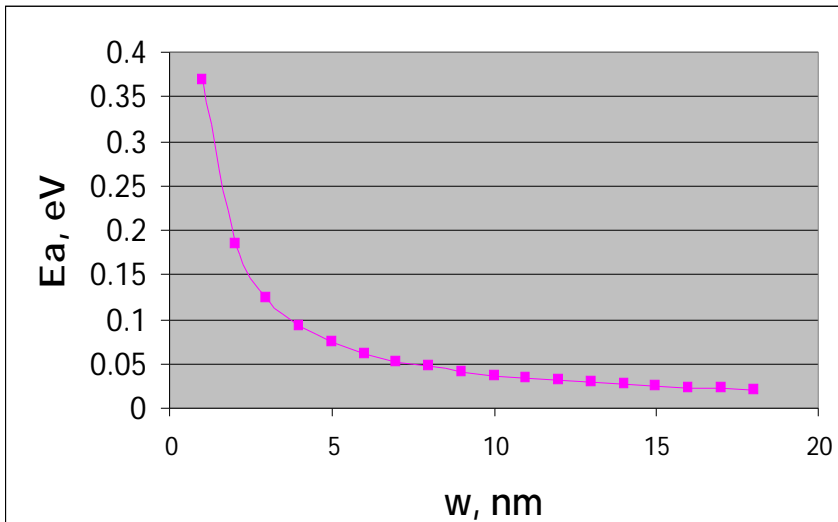
Electrostatic energy to add an electron to a particle with size w

$$E_a = \frac{e^2}{4\pi\epsilon\epsilon_0 w}$$

$$E_{\text{bit}} > E_a$$

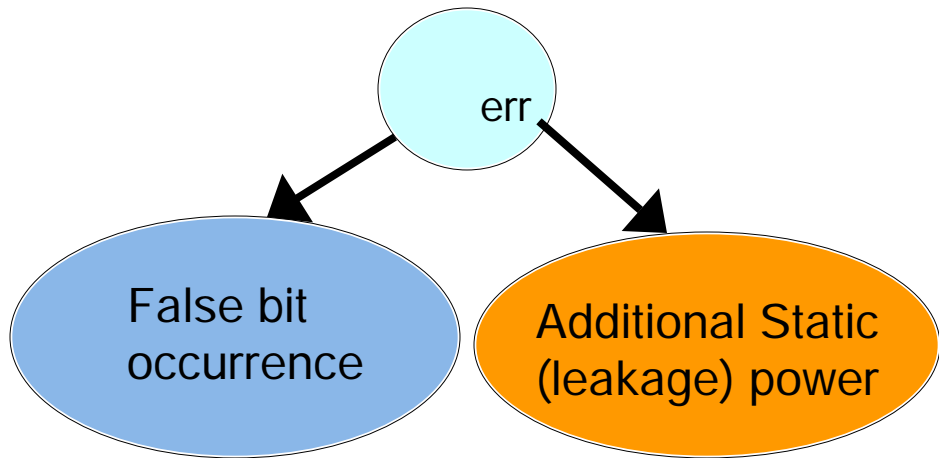
$$P = E_a f n = \frac{f E_a}{W^2}$$

$$f = 10 \text{ GHz}$$

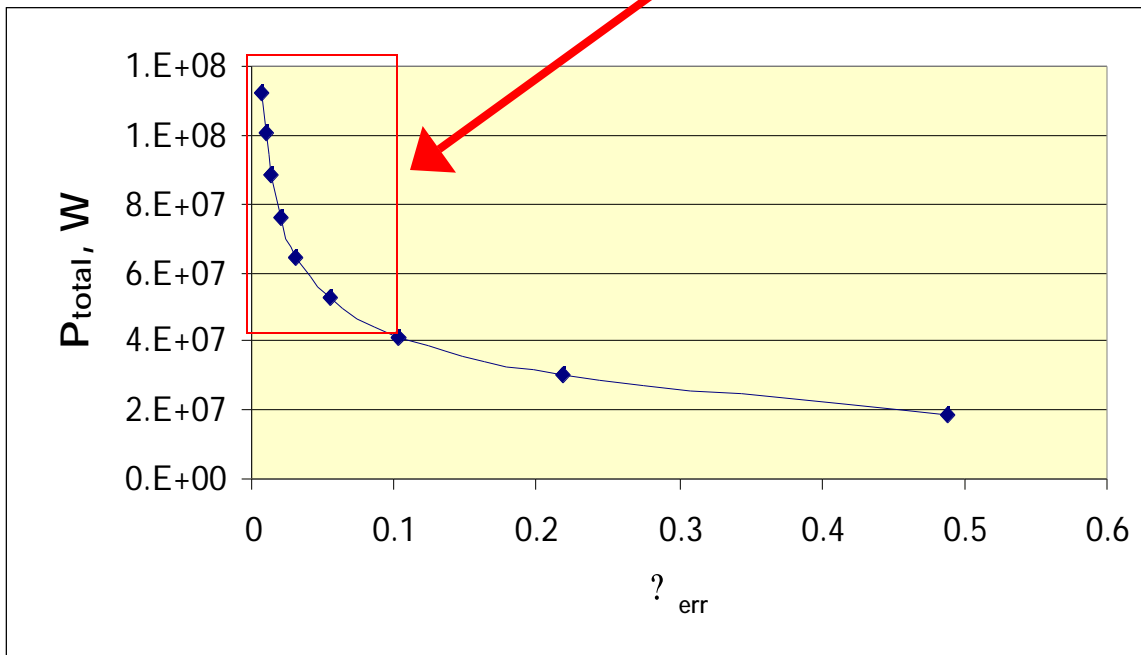


In sub-10 nm single electron devices, the minimum energy to add/remove electron is much larger than kT , and it increases as size decreases.

Power vs. Error trade-off



Computation at $\epsilon_{err}=0.5$ is impossible
In useful computation, $\epsilon_{err} \ll 0.5$,
hence much larger total power is
needed



$a=1\text{nm};$
 $n=10^{14}\text{ bit/cm}^2;$
 $f=3 \times 10^{13}\text{ Hz}$

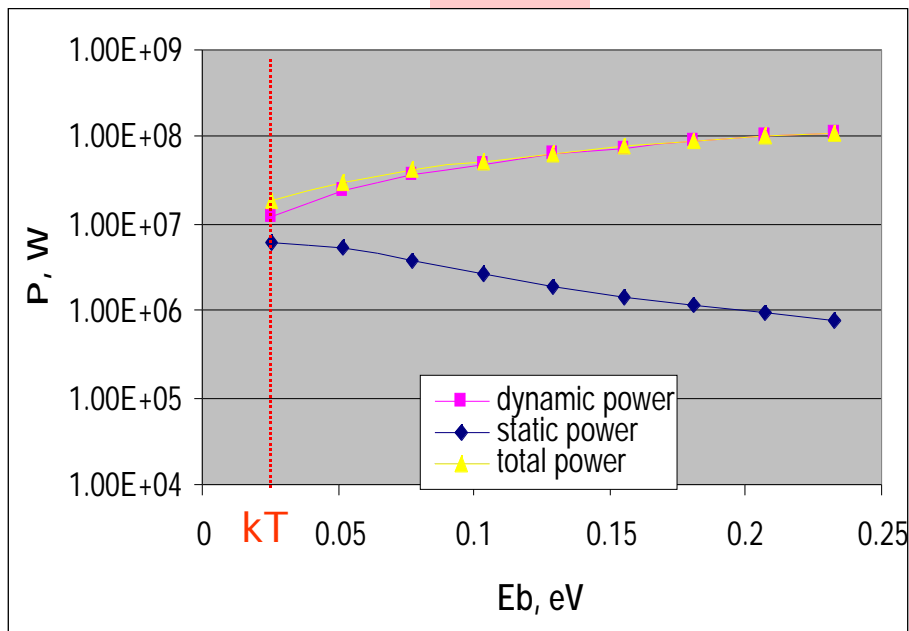
Dynamic and Static Power

$$P_d = I_{on} V = \alpha n e f \frac{E_b}{e} = \alpha E_b f n$$

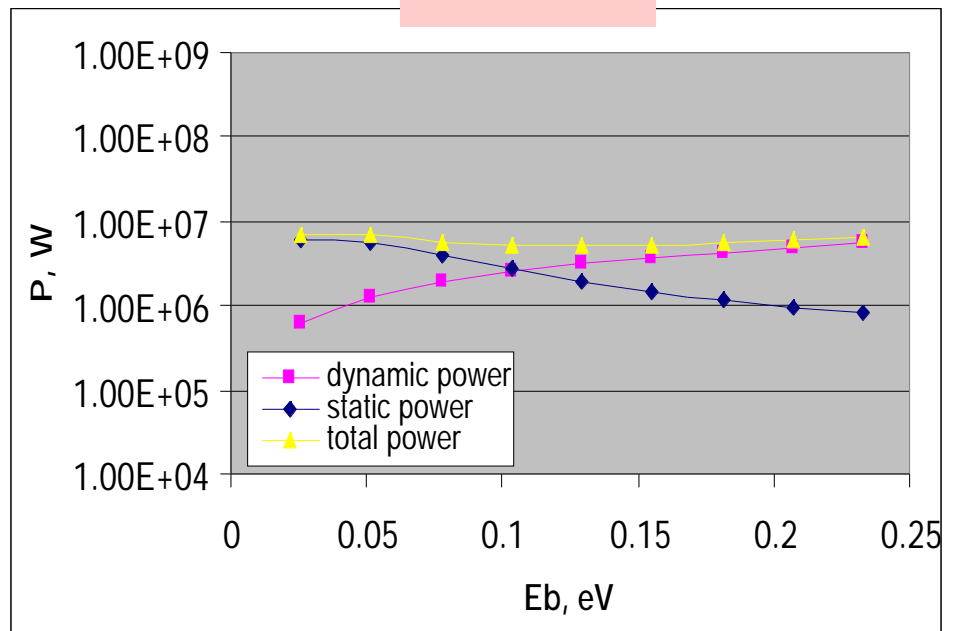
$$P_s = i_{off} V = n e f_{err}(E_b) \frac{E_b}{e} = \frac{P_d}{\alpha} f_{err}(E_b)$$

$a=1\text{nm}; n=10^{14}\text{ bit/cm}^2; f=3 \times 10^{13}\text{ Hz}$

=1



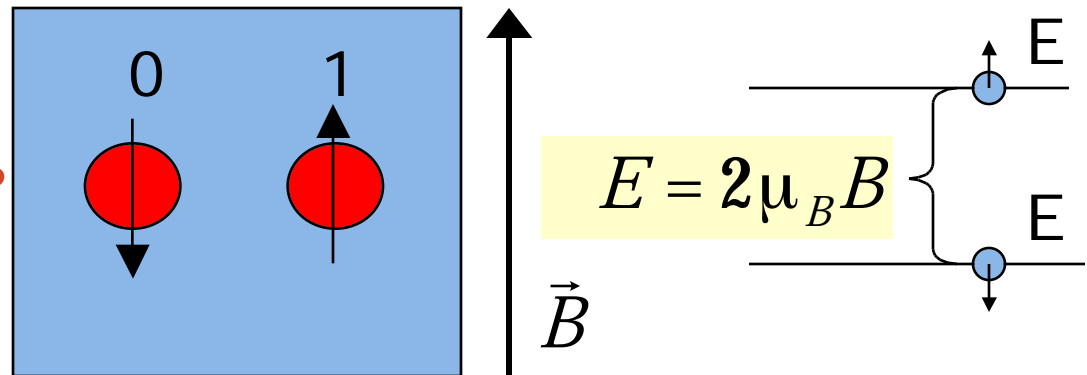
=0.05



Will Spintronics Alleviate the Power Problem?

◆ Expectations:

- ❖ ultra low power ???
- ❖ ultra high density ???



- ◆ A quote: "Spintronics would use much less power than conventional electronics, because the energy needed to change a spin is a minute fraction of what is needed to push charge around"

Example:

$T=300\text{ K}$

$B=1.5\text{ T}$

(practically viable)



$$E=2 \mu_B B = 2 \cdot 9.27 \cdot 10^{-24} \text{ J / T} \cdot 1.5 \text{ T} = 2.78 \cdot 10^{-23} \text{ J} = \mathbf{1.74 \cdot 10^{-4} \text{ eV}}$$

$$err = \exp \left(-\frac{2\mu_B B}{kT} \right) = \mathbf{0.99}$$

$$(\mu_B = 9.27 \cdot 10^{-24} \text{ J / T})$$

- ◆ Is the very low energy to change state an advantage (e.g. low dynamic power) or a disadvantage (e.g. high error probability) for applications of spin devices in information processing?

How much heat a solid system can tolerate?...

ITRS 2001 projects 93 W/cm² for MPU in 2016

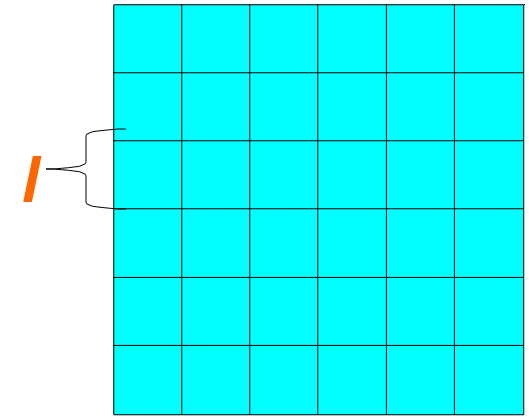
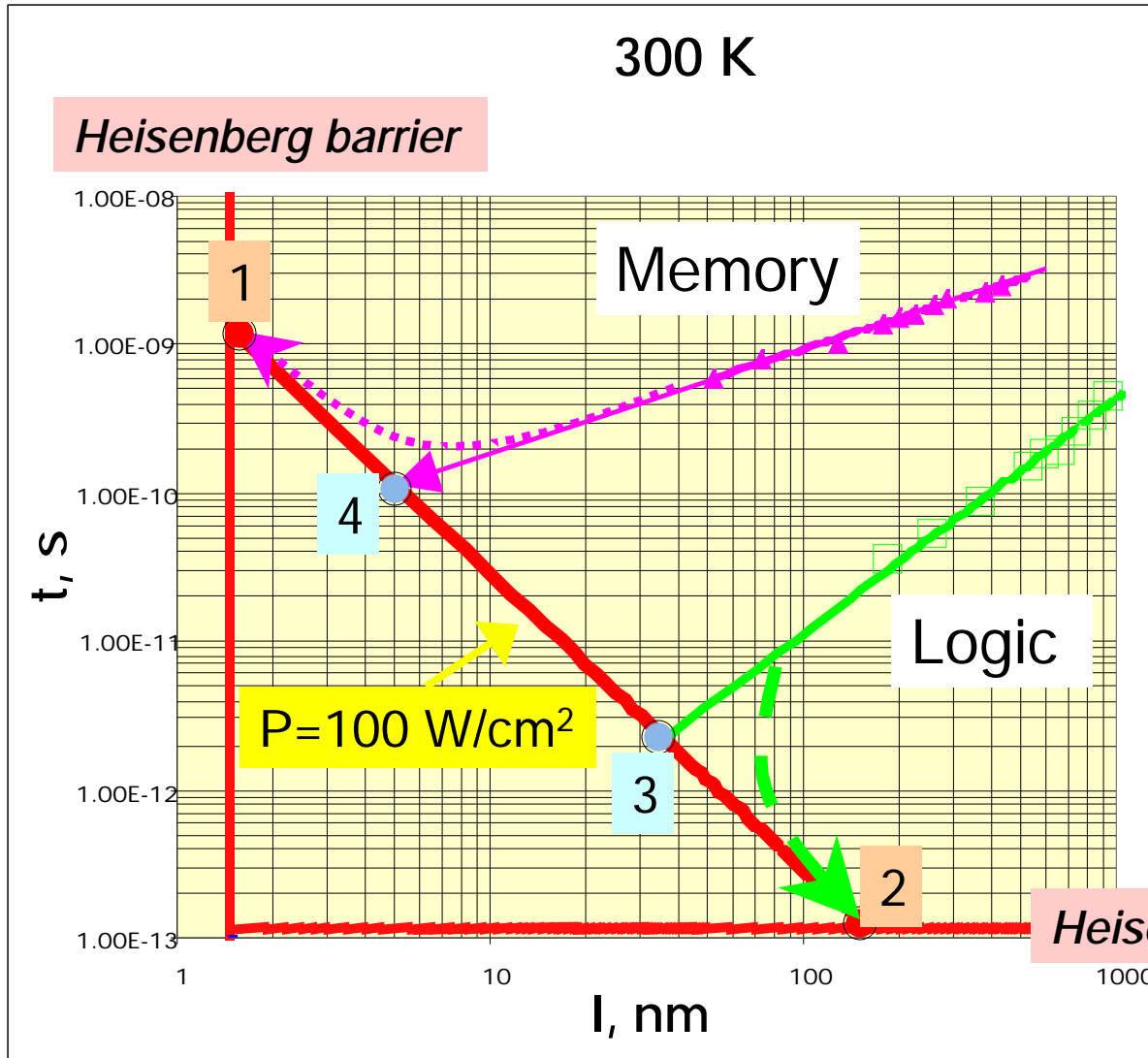
Several hundred W/cm² is close to known limits of heat removal from a 2-dimensional solid material structure with $T_{\max} = 125^{\circ}\text{C}$

Experimental demonstrations of on-Si cooling systems (without active devices):

680 W/cm² thermoelectric (Zheng et al.)

790 W/cm² microchannel (Tuckerman and Pease)

...and implications to the Roadmap: Inflexion of ITRS vectors?



l is the "cell size":

$$l = n^{-1/2}$$

$$l_{min} = a$$

$$l_{MPU} = (11-15)a$$

Implications for Nanoelectronics Utilizing ElectronTransport

- ◆ Scaling to molecular dimensions may not yield performance increases
 - ❖ We will be forced to trade-off between speed and density
- ◆ Optimal dimensions (depending on speed/density trade-offs) for electronic switches should range between 5 and 50 nm, and this may be achievable with silicon technology
 - ❖ Within the scope of ITRS projections

Fundamentals of Heat Removal

Quotes from anonymous scientists working at the frontiers of nanoelectronics:

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

“Heat can be dissipated somewhere else”

Three fundamentals of heat removal:

1) The Newton's Law of Cooling: $q = h(T_h - T_a)$

(h-heat transfer coefficient)

2) The Ambient: $T_a = 300 \text{ K}$!!! $W_{cool} = \frac{T_a - T_c}{T_c} Q$ Heat to be removed

3) The Carnot's theorem: Work to be done

Newton's law of cooling

$$Q = h^* A^{**} (T_h - T_a)$$

$$\max (T_h - T_a) = 100K$$

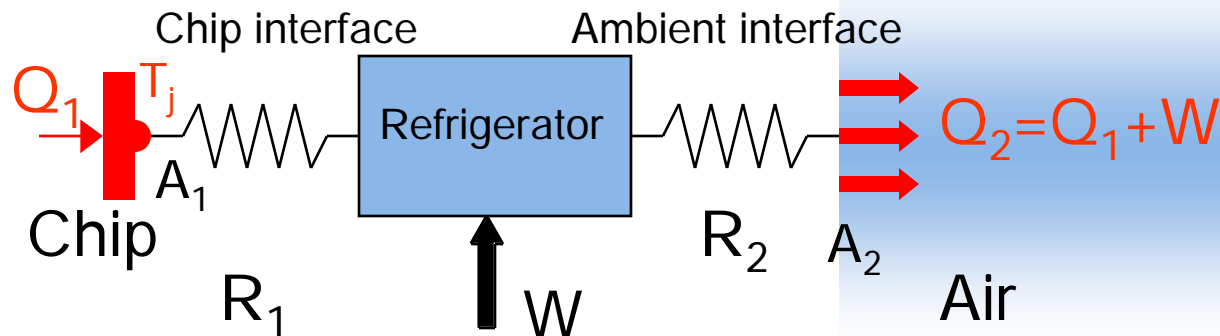
Cooling method	h , W/cm ² ·K
Air, natural convection	0.001
Air, forced convection	0.01
Water, natural convection	0.1
Water, forced convection	1
Boiling	10

Max $P = 1000 \text{ W/cm}^2$? ($A = \text{const}$)

*h – the heat transfer coefficient

**A - area

The ambient interface



$$R = \frac{1}{Ah} \quad \text{- thermal resistance}$$

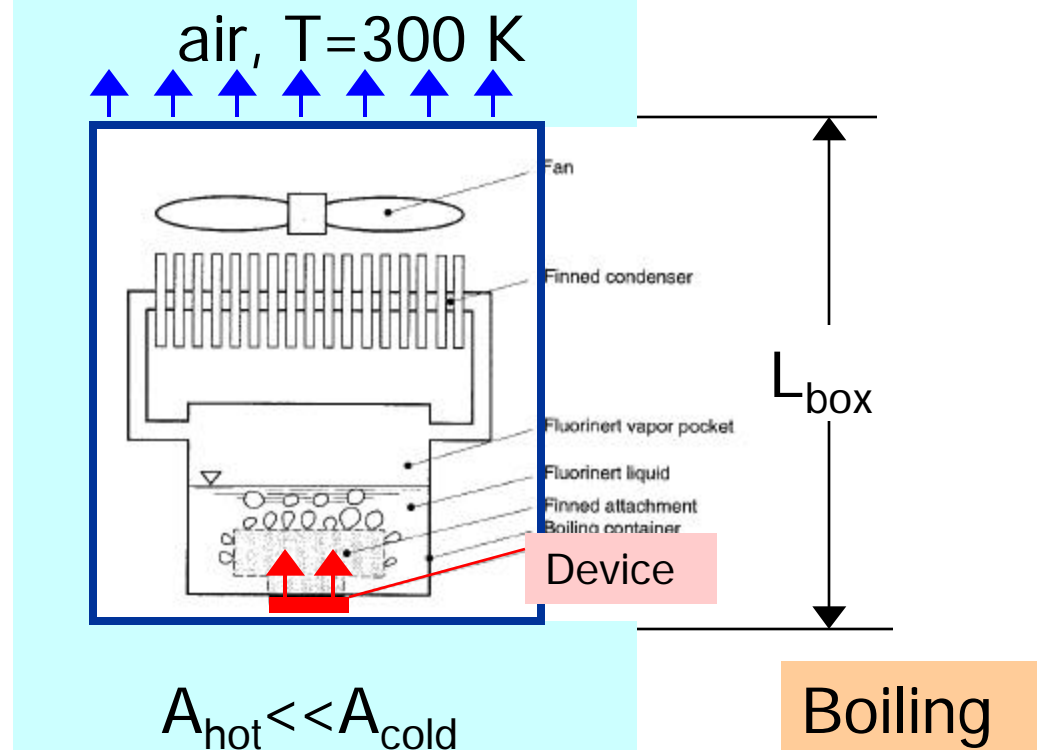
$$Q_2 > Q_1$$

$$h_2 < h_1$$

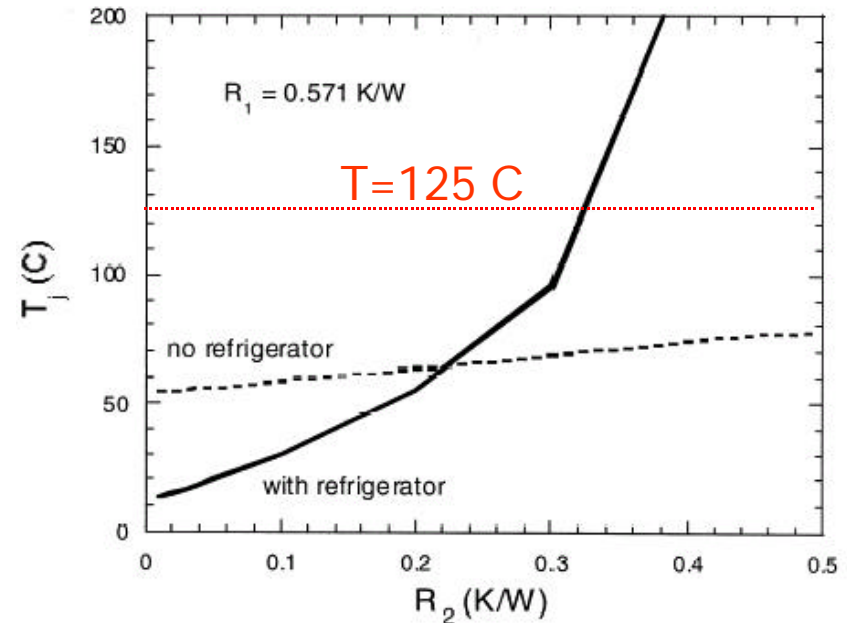
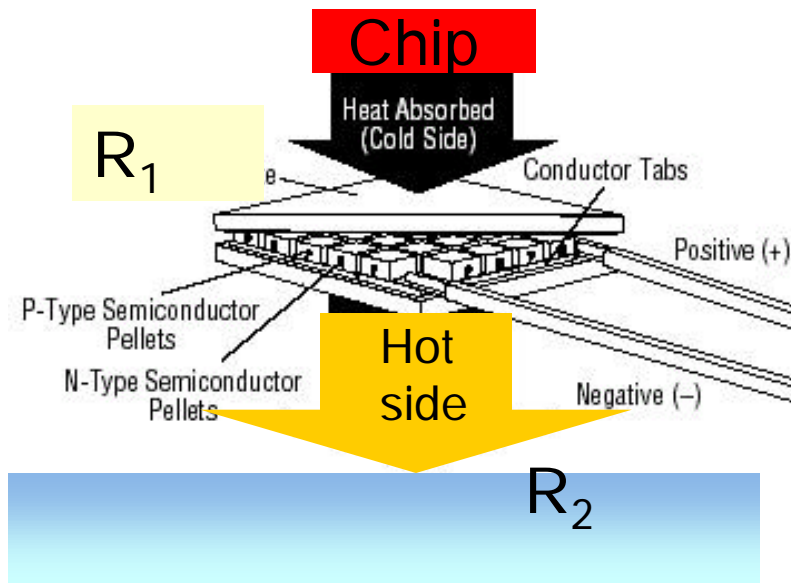


$$A_2 > A_1$$

Surface extension is essential for removal of high heat fluxes to the environment



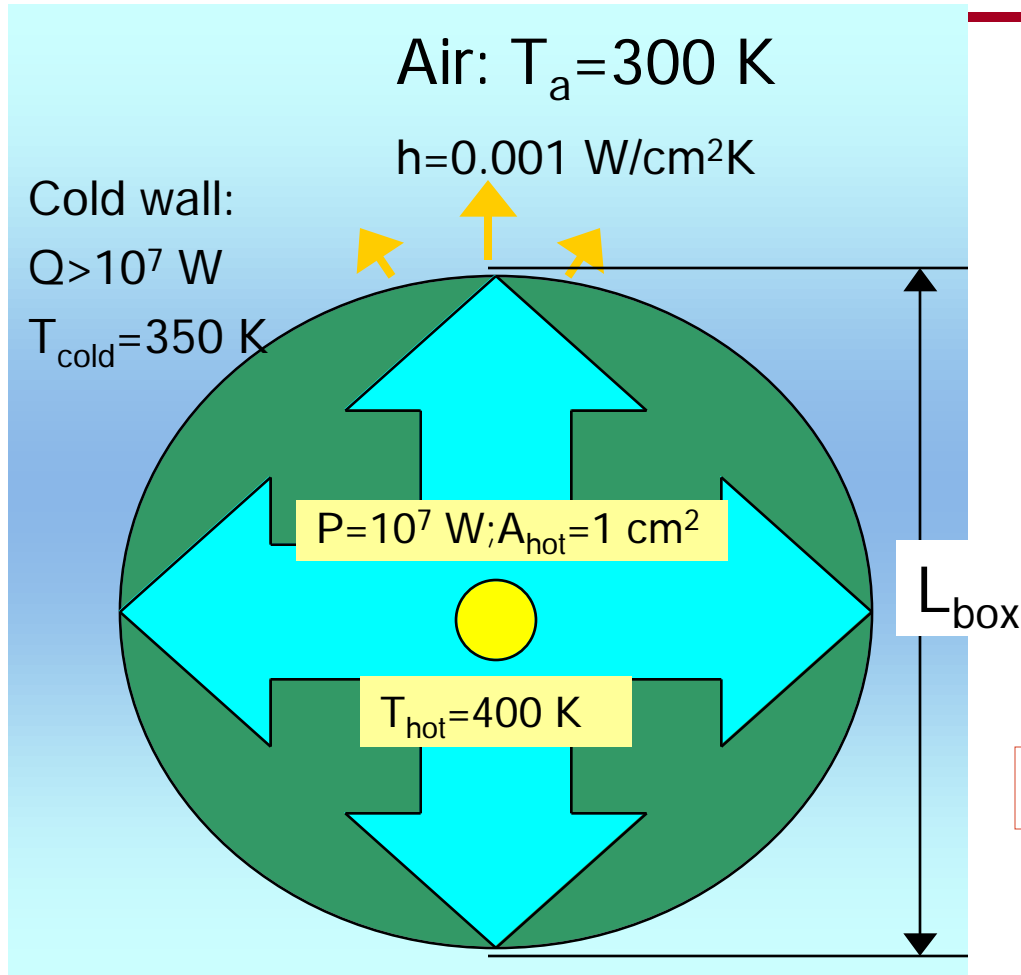
Thermoelectric cooling



Heat rejection to the ambient is a universal element of all cooling methods and requires surface extension

Effect of varying R_2 (e.g. changing air's humidity, temperature etc.) on the heat removal rate and junction temperature, for a thermoelectric cooler, compared with air cooling by a fan.

How Much Volume is Needed to Transfer Chip Heat to the Ambient?



Example:

$$n = 10^{14} \text{ bit/cm}^2$$

$$t = 0.01 \text{ ps}$$

$$\text{BIT} = 10^4 \text{ Tbit/ps}$$

$$a = 1 \text{ nm}$$

$$E_{\text{bit}} = 0.08 \text{ eV } (\sim 4kT)$$

$$\text{error} = 10\% / \text{device}$$

$$P = 10^7 \text{ W/cm}^2$$

Box size:

$$A_{\text{cold}} > 2 \times 10^8 \text{ cm}^2$$

$$L_{\text{box}} > 141 \text{ m}$$

Very Big Box!

Example: computer min. size vs power

P, W	Approx. box dimensions, cm
1	3x3x0.5
10	10x8x1
100	30x20x8
1000	100x50x40
10000	182x182x182

$$Q > P$$



$$A_{\text{cold}} > \frac{P}{h_{nc} T}$$

Carnot's Refrigerator and Cryogenic Computation

The efficiency of heat engines dramatically drops at $T \ll T_a$

$$W_{cool} = \frac{T_a - T_c}{T_c} Q^-$$

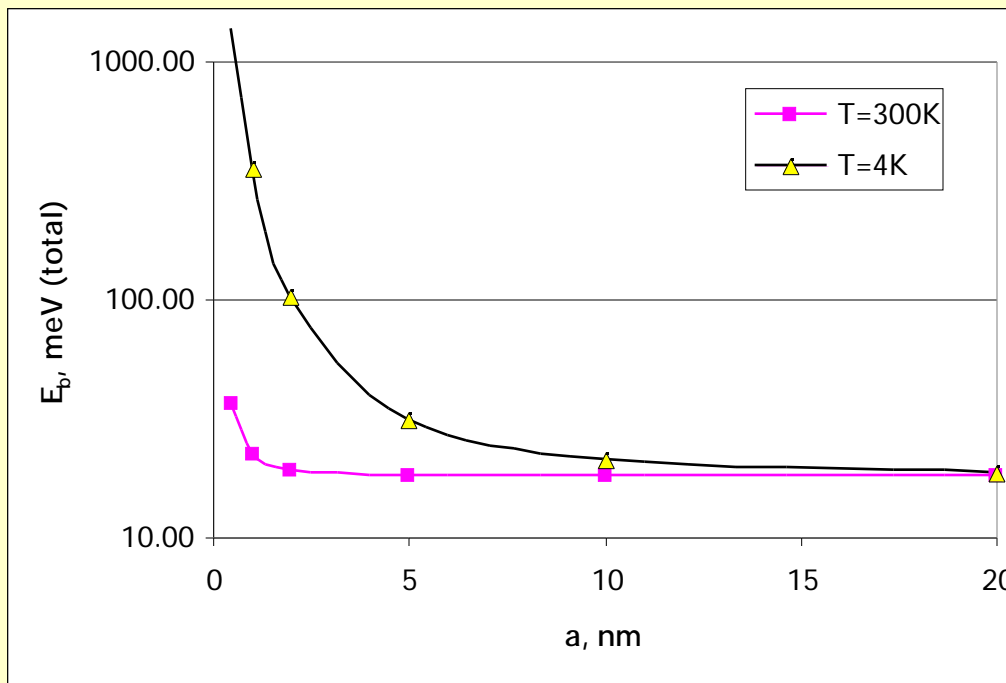
Min. total power needed to run a 100 W chip:

at 77 K - 300 W

at 4.2 K - 7 kW

Cryogenic Computation with nanodevices

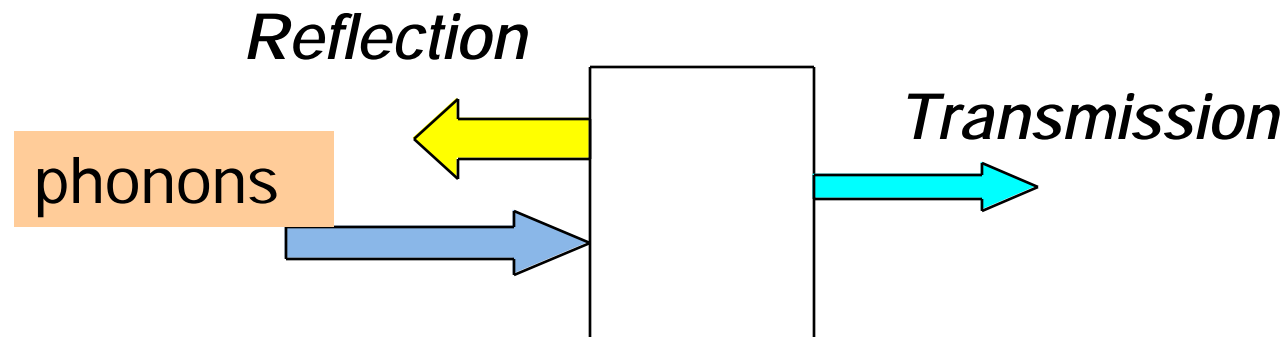
$$E_{bit}^{total} = E_{bit} + \frac{T_a - T_{dev}}{T_{dev}} E_{bit} = \frac{T_a}{T_{dev}} E_{bit} = \frac{T_a}{T_{dev}} k_B T_{dev} \ln 2 + \frac{\hbar^2 (\ln 2)^2}{8ma^2} =$$
$$= k_B T_a \ln 2 + \frac{T_a}{T_{dev}} \frac{\hbar^2 (\ln 2)^2}{8ma^2} > k_B T_a \ln 2$$



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.

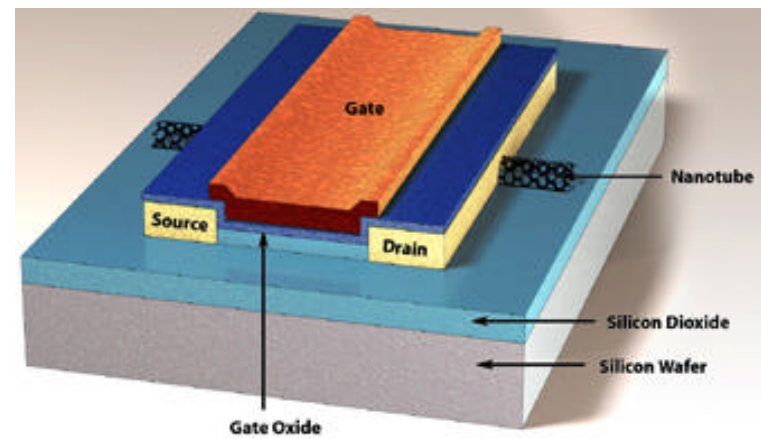
The barriers dilemma

- ◆ Energy barriers are key components to provide *Information Flow*
- ◆ Energy barriers are negative factor *for Heat Flow*
- ◆ Can we think of radically new ways of heat removal based on coherent *heat flows*, e.g. *heat lasers* or *solitons*?



A driver for physical layout?

A question – What to do?



Energy efficiency of CMOS

Does practical CMOS operate far from fundamental limits?

◆ 2016 ITRS 22-nm Node:

- ❖ x_{\min} : Channel length **9 nm**
- ❖ E_{sw} : Switching energy **2×10^{-18} J**
- ❖ E_b : S-Ch barrier height **~ 0.4 eV**
- ❖ Electrons/switching event **~ 50**
- ❖ Energy/electron **4×10^{-20} J ~ 12 kT**

Fundamental limits

$$x_{\min} = a = \frac{\hbar}{\sqrt{2mkT \ln 2}} = 1.5 \text{ nm}(300 \text{ K})$$

$$E_{\text{sw}} > kT \ln 2 = 3 \times 10^{-21} \text{ J}$$

$$E_b > kT \ln 2 = 0.02 \text{ eV}$$

1

$$3 \times 10^{-21} \text{ J} \sim kT$$

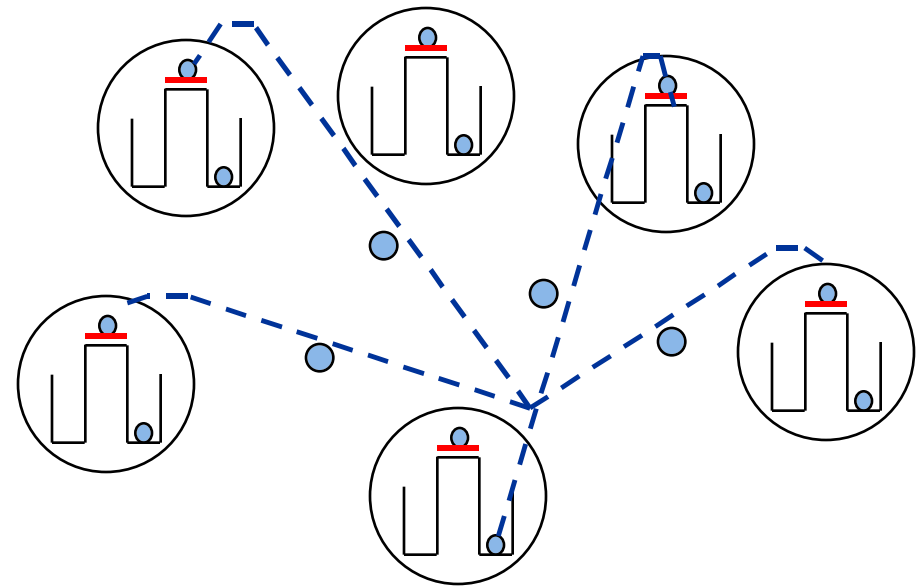
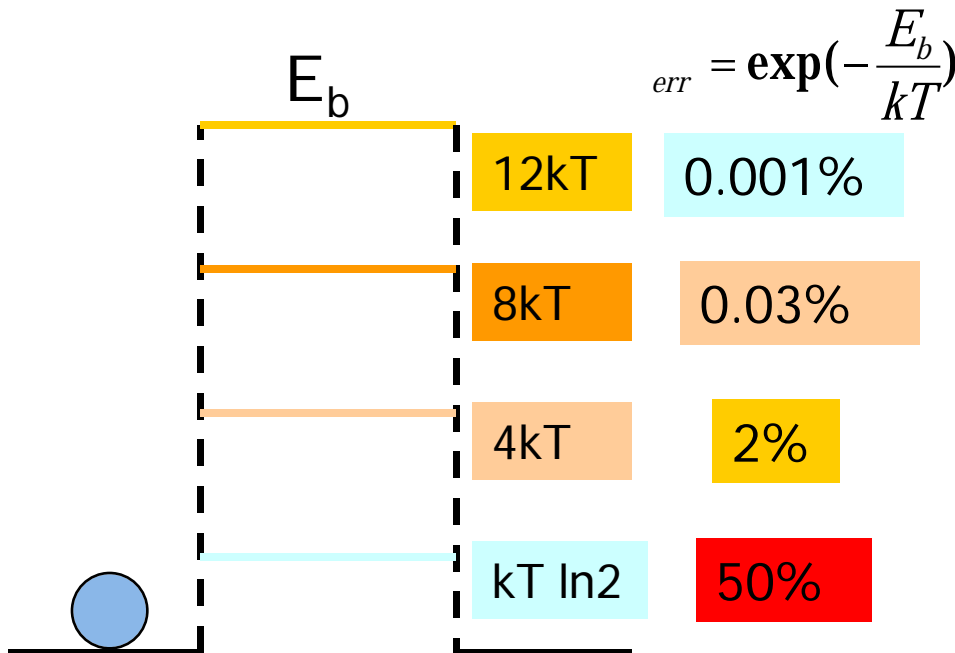
Can we decrease the energy of CMOS?

$2 \times 10^{-18} \text{ J}$ \longrightarrow $3 \times 10^{-21} \text{ J}$

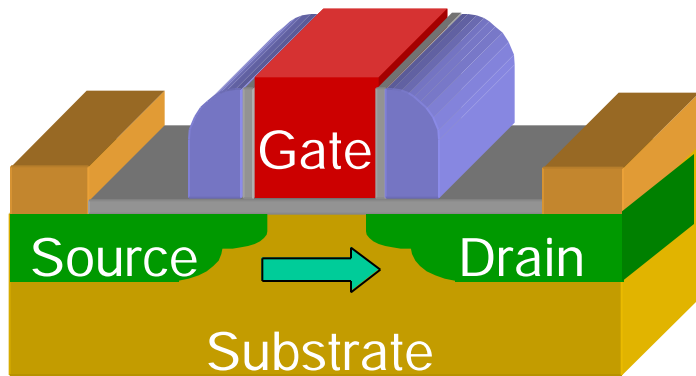
Yes (in principle):

Decrease the barrier height:
 $0.4 \text{ eV} \rightarrow 0.02 \text{ eV}$

Decrease the number of electron per switching event:
 $50 \rightarrow 1$

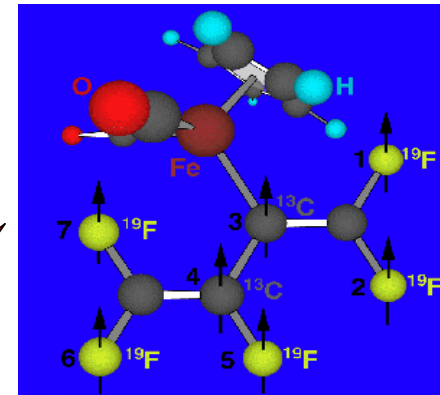


What to do? (Cont'd)

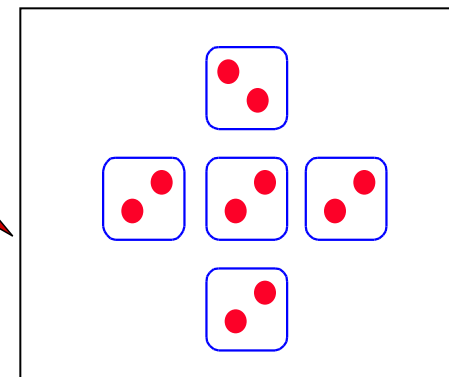


Silicon MOSFET

*Conventional von Neumann
Architecture*



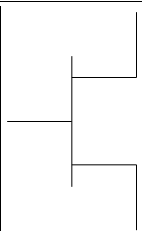
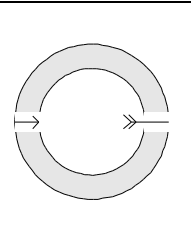
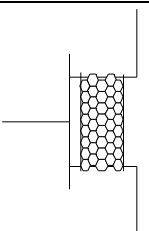
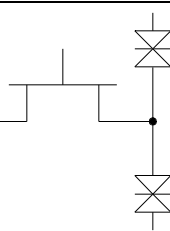
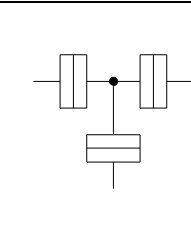
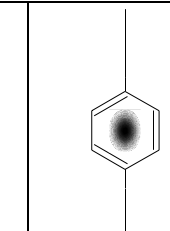
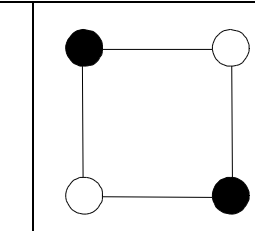
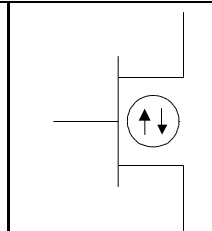
Quantum Computer



Cellular Non Linear Network

*New Information
Processing Architectures*

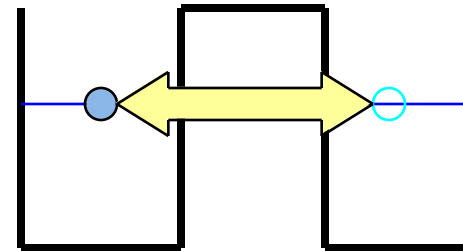
2003 ITRS: Emerging Research Devices

Device								
	FET	RSFQ	1D struct	RTD	SET	Molecular	QCA	Spin transistor
Density (dev/cm ²)	3E9	10 ⁶	3E9	3E9	6E10	10 ¹²	3E10	3E9
Switch Speed	700 GHz	1.2 THz	?	1 THz	1GHz	?	30 MHz	700 GHz
Circuit Speed	30 GHz	400 GHz	30 GHz	30 GHz	100 MHz	1 MHz	1 MHz	30 GHz
Switching energy, J	2E-18	2x 10 ⁻¹⁹ [>1.4E-17]	2E-18	>2E-18	10 ⁻¹⁸ [>1.5E-17]	1.3x10 ⁻¹⁶	E: [> 10 ⁻¹⁸] M:>4x10 ⁻¹⁷	2 x 10 ⁻¹⁸
Binary throughput, GBit/ns/cm ²	86	0.4	86	86	10	?	0.06	86

Classic to Quantum transition

- ◆ Classic memory bits become indistinguishable, which limits our ability to use them for computation

BUT



- ◆ The superposition of indistinguishable states is a key concept of *Quantum Computation*
- ◆ *A quantum bit or qbit* is a physical system with two quantum states

Power of quantum computing

- ◆ Quantum information storage
 - ❖ N quantum bits stores 2^N complex numbers
 - Consider information in 300 entangled qubits
 $2^{300} = 10^{90}$
 - Compare to the total number of atoms in the Universe:
 $N_{\text{atoms}} = 10^{80}$

If dramatic improvement of the information throughput can be achieved, the cryogenic operation might be affordable

Neuromorphic Computing

- ◆ Implies computational schemes and systems resembling operation of human brains.
 - ❖ The potential capabilities of neuromorphic computers could be close to those of the brain, thus enabling e.g. artificial intelligence
- ◆ Properties of brain:
 - ❖ Mass – 1.5 kg
 - ❖ Volume – 1.5 l
 - ❖ Energy consumption – ~10 W
 - ❖ Information stored – $1e14$ bits
 - ❖ $1e13$ bits/s

Conclusions

- ❖ Fundamental considerations suggest that the potential benefits from replacing CMOS devices with new types of electron transport devices may be limited
- ❖ Search for radically new methods of heat removal is one of the most critical research directions
- ❖ The exploration of alternative approaches to von Neumann type computing, such as brain or Reversible/Quantum Computation, is becoming a strategic imperative.
 - We need a concerted effort in these areas because of the long lead times for the introduction of radically new technologies