Semiconductor Research Corporation

PIONEERS IN COLLABORATIVE RESEARCH®

global RESEARCH COLLABORATION

"Minimizing" the Effects of Physical Limits

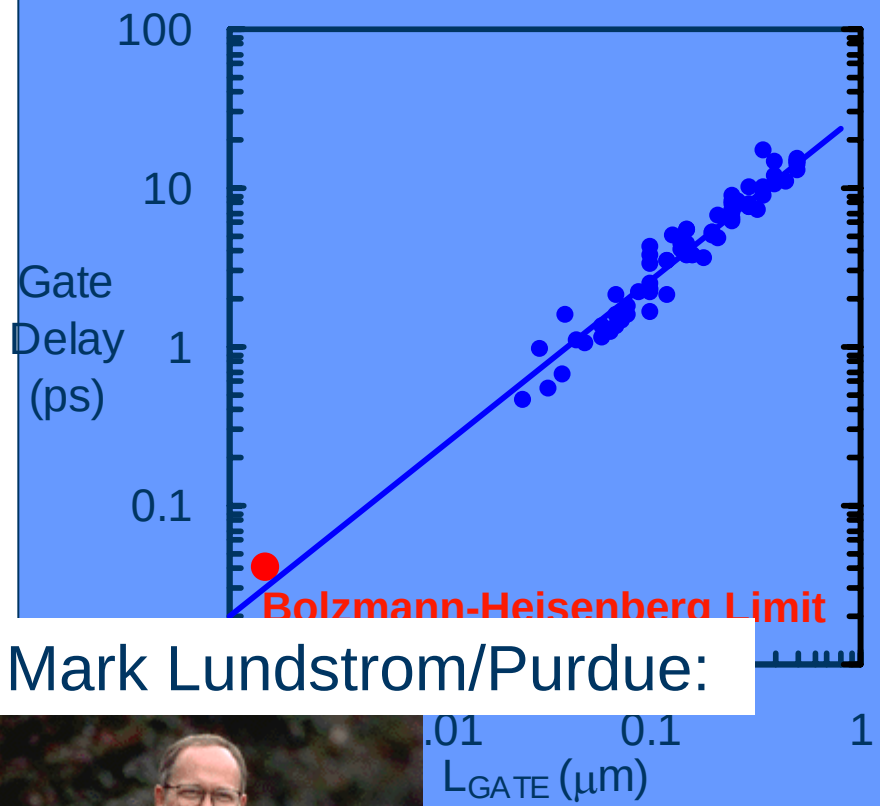
Victor Zhirnov

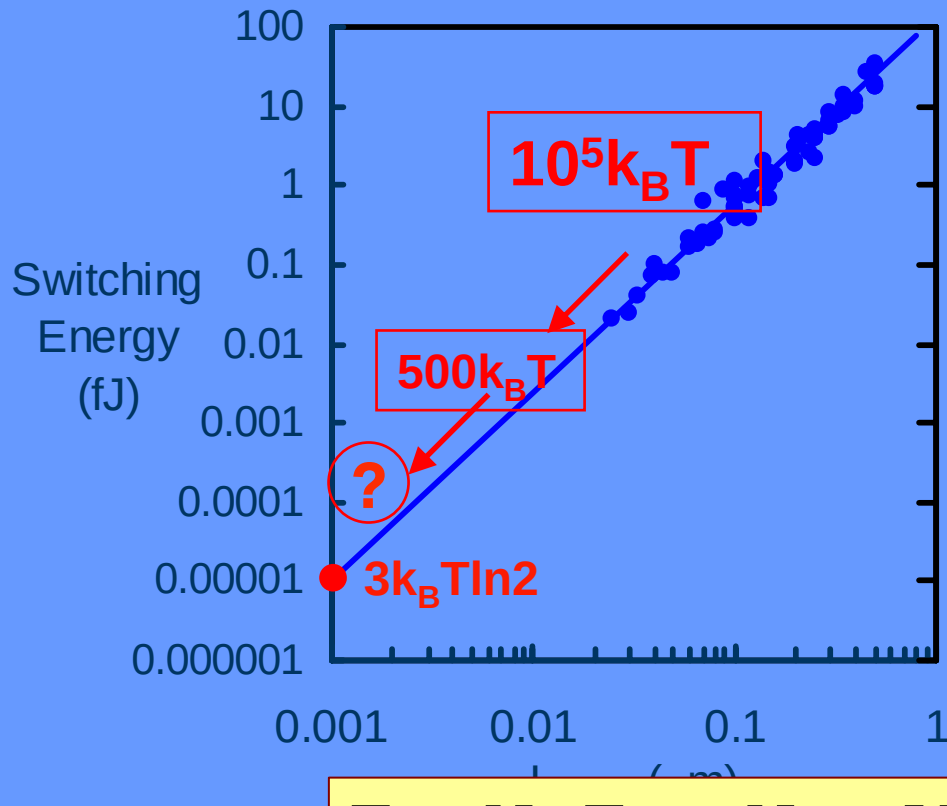
Semiconductor Research Corporation

Stanford University, November 8 & 9, 2007

CMOS scaling on track to obtain physical limits for electron devices

George Bourianoff / Intel





Mark Lundstrom/Purdue:

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

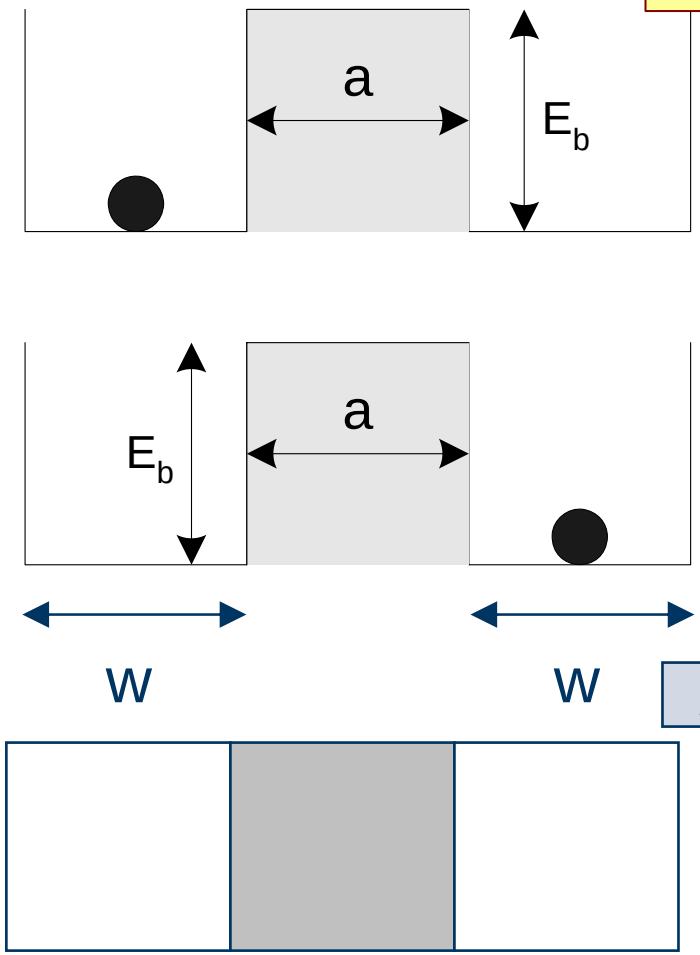
$$E \sim N \cdot E_b = N \cdot e \cdot V_{dd}$$

Answer:

1) **System reliability costs**
2) **Long communication costs**
3) **Fan-Out costs**

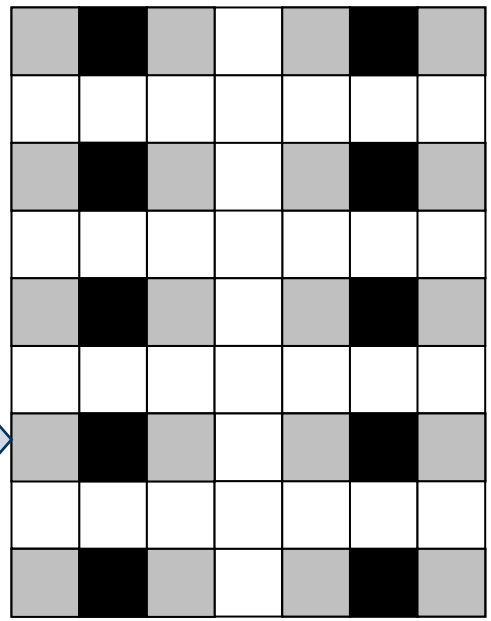
Two-well bit – **Universal Device Model**

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



Generic Floorplan
of a binary switch

Array



Optimum tiling)

White spaces are required to provide for isolation and interconnect

Device density

1) Upper Bound

$$n_{\max} = \frac{1}{8a^2}$$

2) IC (ITRS)

$$n_{MPU} = \frac{1}{20a^2}$$

System Level Energetics I: *Reliable Switching*

Computation at Π_{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

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

$\Pi_{err}\downarrow \rightarrow E\uparrow$ (Boltzmann&Heisenberg)

System Constraint on Minimum Energy per Bit

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

The probability that all N switches in a circuit work correctly

$$\Pi_{syst} > \Pi_{crit}$$ e.g., 0.5 lower boundary; 0.99 a "reasonable" boundary

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

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

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

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

Boltzmann Heisenberg

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

Uniformly Scaled Information Processor

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

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

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

- The minimum barrier height in binary switches and therefore minimum operating voltage is a function of
  - L_g (device level)
  - N (system level)

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

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

for $\Pi=\Pi_{crit}$

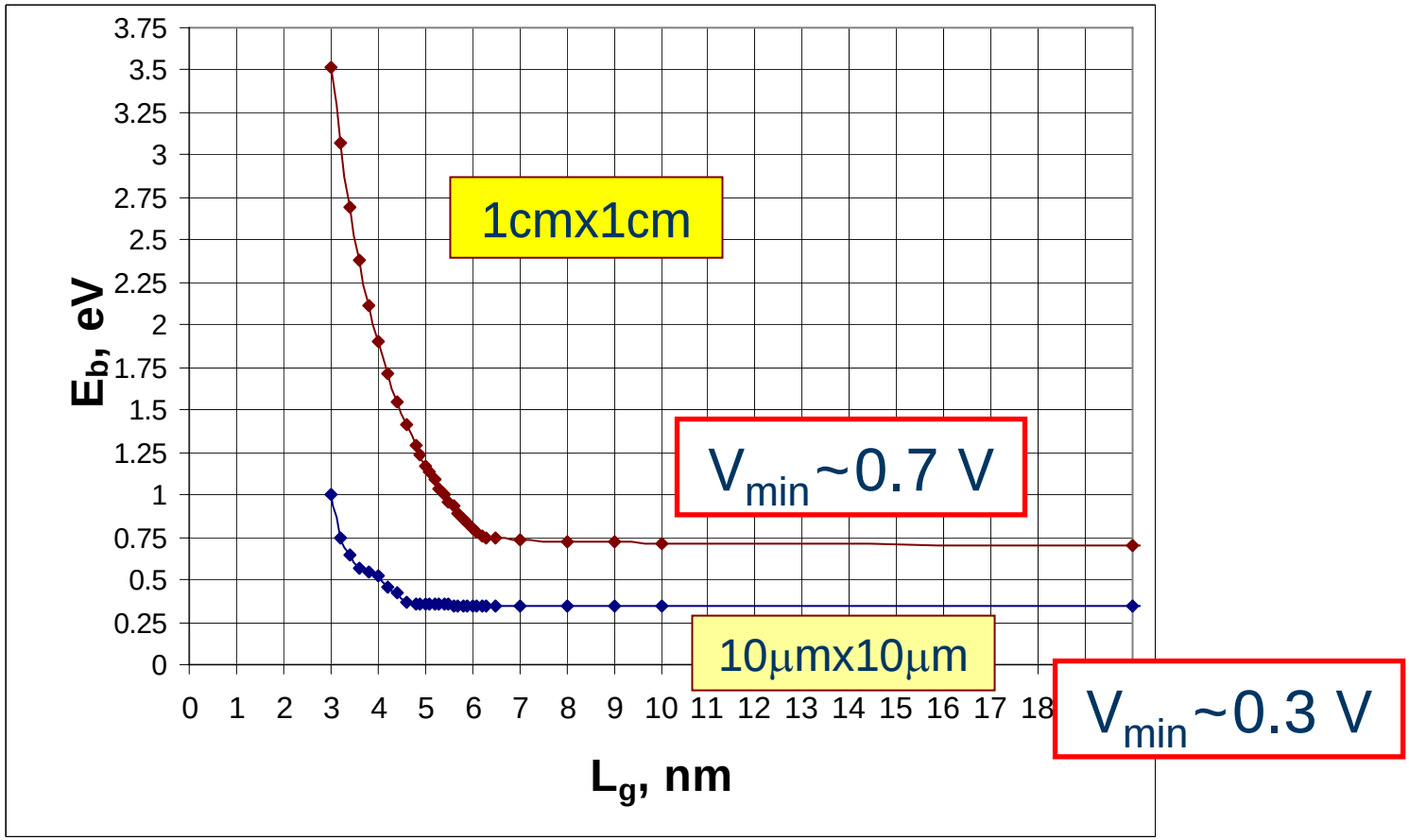
$$\Pi_{syst} = (1-\Pi_{err})^N$$

Π_{crit}=0.99

L_g, nm	N, cm^{-2}	E_{bmin}
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

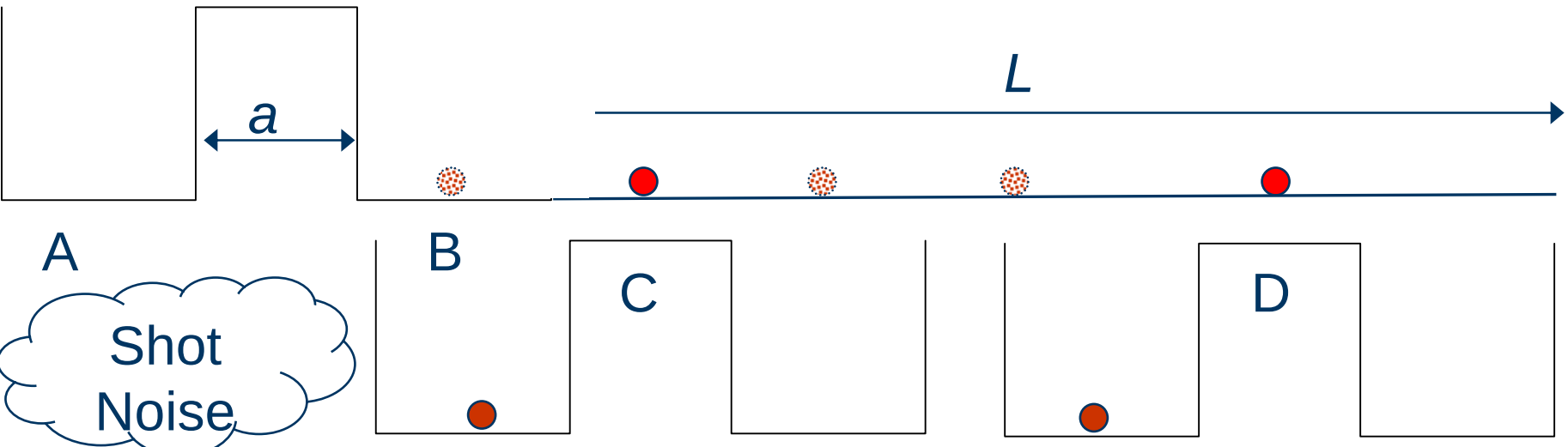
Compaction Effects on Minimum Operation Voltage

- Energy – Errors dilemma



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 out

N – the number of electrons

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

n=2

L=4*a*

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

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

FO2

$n=2 \quad L=4a$

$N_{min}=5$

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

FO4

$n=4 \quad L=8a$

$N_{min}=14$

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

Communication between binary switches takes more energy than information processing

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

$L=8a$

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
			40	1.07E+03
			22	6.05E+02

Messages: Long Interconnects

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

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

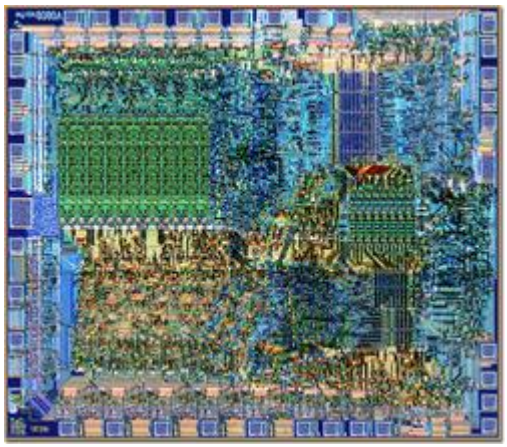
$$\Pi_n = \left(1 - \left(1 - \frac{a}{L}\right)^N\right)^n$$

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

We need to minimize the number of long lines!
Architecture Implications?

Scaling of 8080 MPU

1974



Technology:	NMOS
Feature size:	6 μm
# of transistors:	4500
Die size:	5 mm x 4 mm
Voltage:	5V, 12 V
Frequency:	2 MHz
Power:	1.5 W

2020

Technology:	CMOS
Feature size:	6 nm
# of transistors:	4500
Die size:	5 μm x 4 μm
Voltage:	~0.5 V
Frequency:	~2 MHz-1GHz
Power:	~10nW-10μW

Different Facets of Scaling

Device Scaling

Decrease the physical size

Functional Scaling

Increase system capability and/or application space

System Scaling

Decrease physical size of the system and increase both system capability and application space

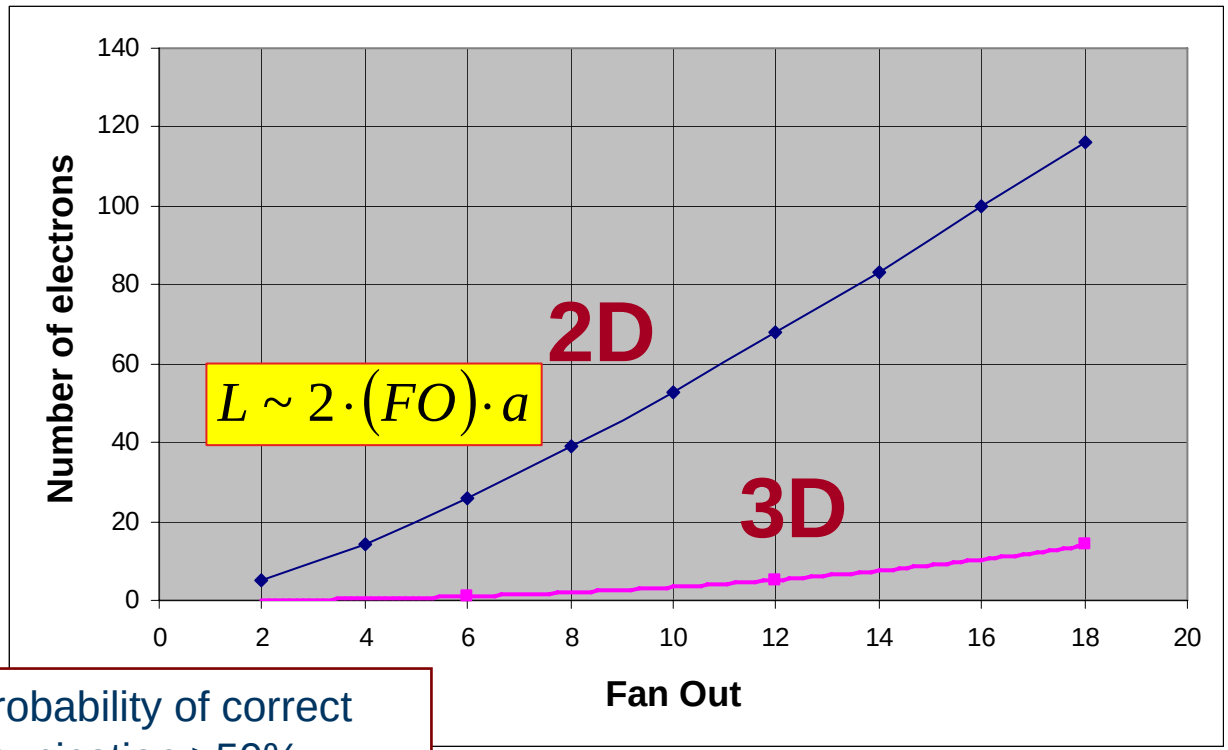
Example:

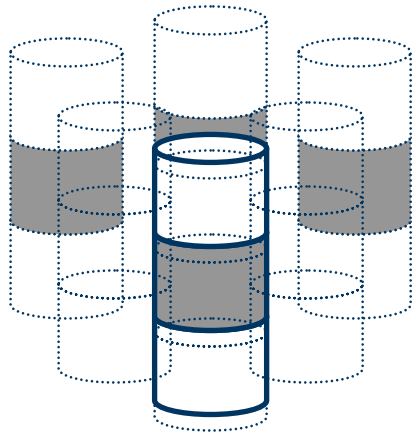
Ultra Mobile Platform

Extreme Microsystems

Electronic cell

Energy costs for fan-out: 2D vs.3D





Generic topology of a 3D binary switch

For probability of correct communication >50%

More Fan-Out (Branching) = More Computation

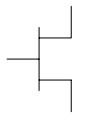
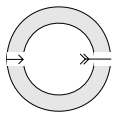
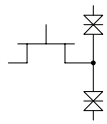
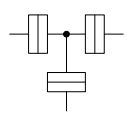
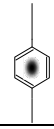
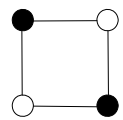
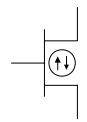
$$E \sim N \cdot E_b = N \cdot e \cdot V_{dd}$$

$$E \sim 3 \cdot 1.6 \cdot 10^{-19} \cdot 0.3 = 1.4 \cdot 10^{-19} J = 35 k_B T$$

Emerging Research Logic Devices

2003 ITRS PIDS/ERD Chapter

3D

Device	 FET	 RSFQ	1D structures	 Resonant Tunneling Devices	 SET	 Molecular	 QCA	 Spin transistor
Cell Size	100 nm	0.3 µm	100 nm	100 nm	40 nm	Not known	60 nm	100 nm
Density (cm^{-2})	3E9	1E6	3E9	3E9	6E10	1E12	3E10	3E9
Switch Speed	700 GHz	1.2 THz	Not known	1 THz	1 GHz	Not known	30 MHz	700 GHz
Circuit Speed	30 GHz	250–800 GHz	30 GHz	30 GHz	1 GHz	<1 MHz	1 MHz	30 GHz
Switching Energy, J	2×10^{-18}	$>1.4\times10^{-17}$	2×10^{-18}	$>2\times10^{-18}$	$>1.5\times10^{-17}$	1.3×10^{-16}	$>1\times10^{-18}$	2×10^{-18}
Binary Throughput, GBit/ns/cm^2	86	0.4	86	86	10	N/A	0.06	86

We HAVE IDENTIFIED NO VIABLE EMERGING LOGIC TECHNOLOGIES for Information Processing beyond CMOS

Key Messages

- Compaction effects due to scaling can provide significant energy reduction
  - Fixed Architecture
- Alternative devices must offer size or energy benefits
  - 1D structures
  - Molecular
- Topology optimization for energy reduction
    - Quasi 1D (e.g. nanowire) components arranged in 3D structures

Additional Remarks

- Memory optimization
  - Low-Voltage Non-volatile memory to replace SRAM
    - Quasi 1D (e.g. nanowire) components
- Communication
  - CNT antennas

Semiconductor Research Corporation

PIONEERS IN COLLABORATIVE RESEARCH®

global GRC RESEARCH COLLABORATION

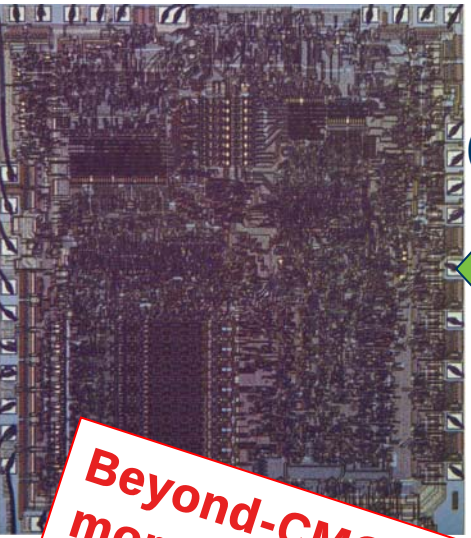
Back-Up

Control Logic Unit for Autonomous Micro-Scale Systems

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

L_g~5nm!

$$N_{tr} = 10^{10} \cdot \left(10^{-3}\right)^2 = 10{,}000$$

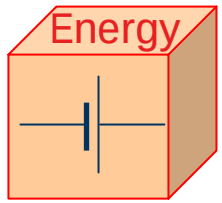


Intel 8080 (5000 trans)

Aggressive scaling is mandatory for μ-scale systems

In principle, the full MPU capabilities can be realized within 10um square if we can sustain scaling to ultimate CMOS (~5 nm gate lengths)

Beyond-CMOS devices for more functionality at less device count?



E_{bit}~*2x10^{-18} J/bit*

$$E \sim 10^{-5}\ J$$

Max number of binary transitions

~5x10^{12}

Complexity of Logic Unit

MAXIMUM

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

$L_g \sim 5nm!$

$$N_{tr} = 10^{10} \cdot \left(10^{-3}\right)^2 = 10{,}000$$

MINIMUM

Logic Unit must contain a minimum number of switches(e.g. transistors) if it is to do useful computation

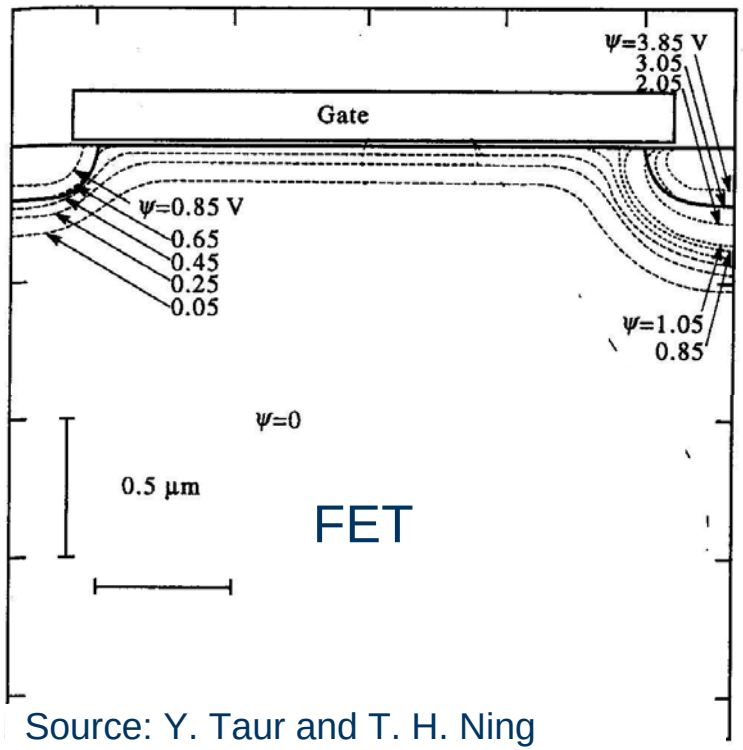
"*if one constructs the automaton (A) correctly, then any additional requirements about the automaton can be handled by sufficiently elaborated instructions. This is only true if A is sufficiently complicated, if it has reached a certain minimum of complexity" (*J. von Neumann)

If we consider a one-bit MPU as the minimum useful element, then the von Neumann threshold is ~150-200 switches

~100 ALU ~100 memory

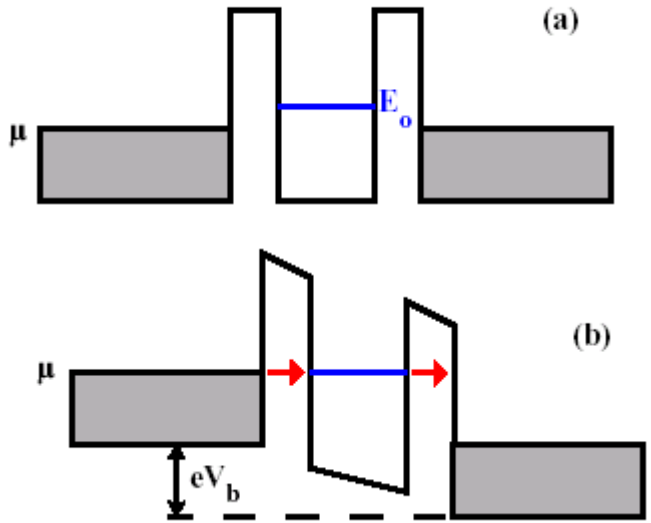
Energy Barriers in Materials

- Any electronic device contains at least one energy barrier, which controls electron flow. The barrier properties, such as height, length, and shape determine the characteristics of electronic devices.



Source: Y. Taur and T. H. Ning
"Fundamentals of modern VLSI devices"
Cambridge university Press 1998

Resonant Tunnel Diode



R. Compano (Ed.)
Technology Roadmap for Nanoelectronics
(European Communities, 2001)

Guglielmo Marconi