Semiconductor Research Corporation

PIONEERS IN COLLABORATIVE RESEARCH®

Computation vis-à-vis Physics: A Framework

VIA 2020 Forum, July 10 & 11, 2008
Ralph K. Cavin, III, Victor V. Zhirnov & Sadasivan Shankar (Intel)

Outline

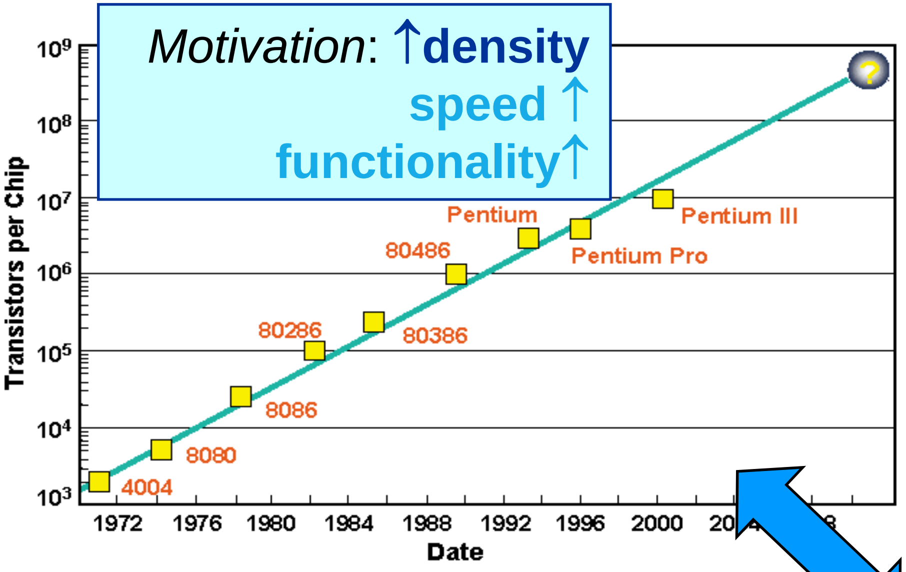
- Relations between maximum computational performance and device physics
  - Correlations between computer performance and technology capability
  - Von Neumann threshold
- Binary switch abstraction
  - Generic floorplan and energetics
  - Connected Binary Switches
  - Floorspace, Timing and Energy for Communication between Binary Switches
- ‘Minimal’ Turing Machine
  - System scaling limits
  - Energetics and efficiency

Moore’s Law: *Binary Information Throughput*

Motivation: ↑**density**

speed ↑

functionality↑



Source: Stan Williams, Hewlett Packard

What is the ultimate number of binary transitions per second in a 1cm^2 chip area?

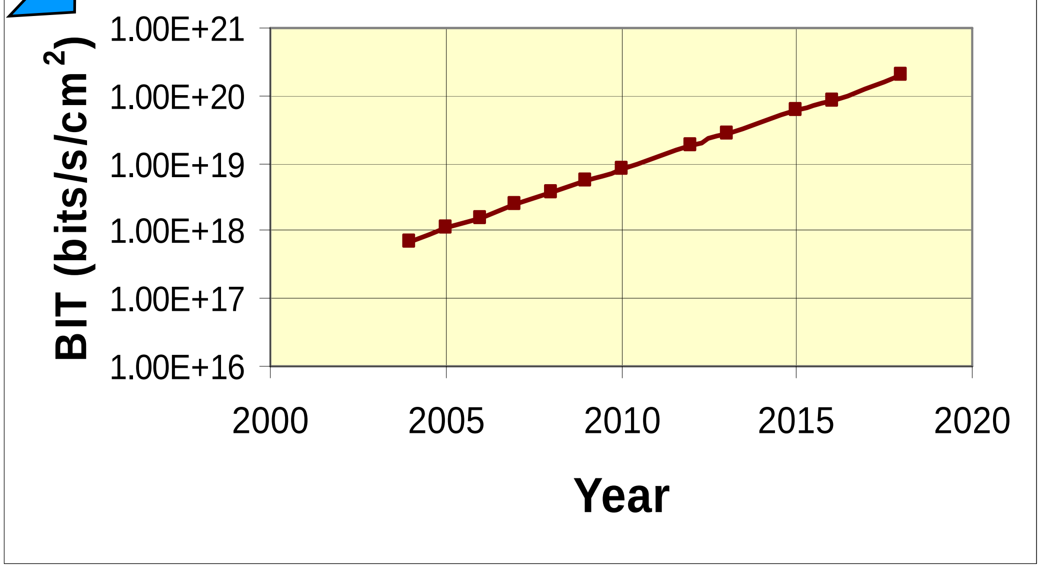
$$\beta = n_{bit}\, f$$

BIT

- a measure of computational capability on device level

n_{bit} – the number of binary states

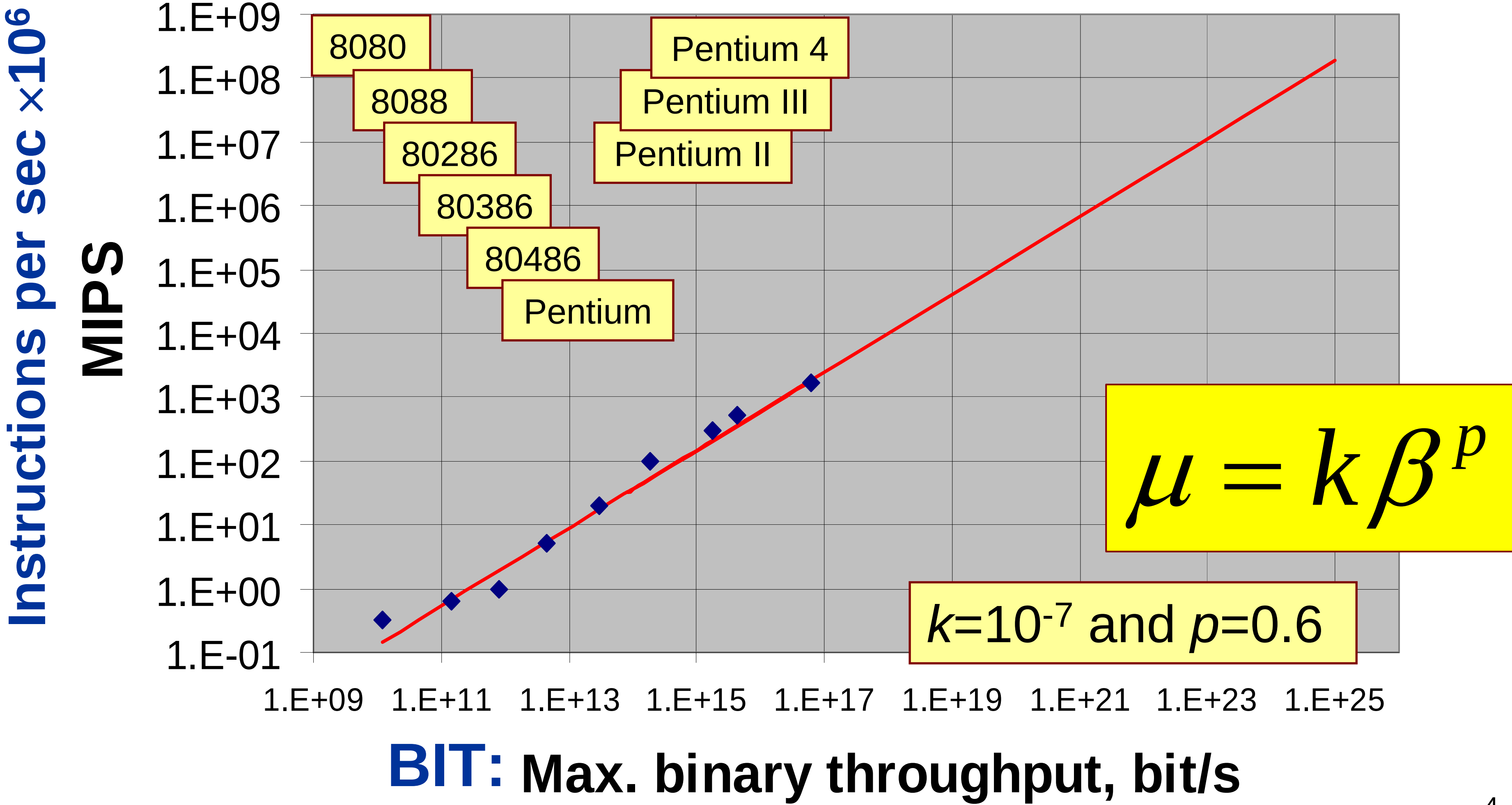
f-switching frequency



Why scaling? – To increase the *Binary Information Throughput* (BIT)

Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and
TSCP Benchmark Scores



$$\mu = k\beta^{p}$$

k=10^{-7} and p=0.6

THESIS

There appears to be a functional relationship between ultimate technology capability defined as the maximum number of binary transitions per unit time, *β*, and the millions of instructions executed per section, *μ*, executed by a single processor:

Number of binary elements

$$\mu = k\beta^{p} = \frac{n_{bit}}{t_{sw}}$$

Switching time

How can we increase MIPS?

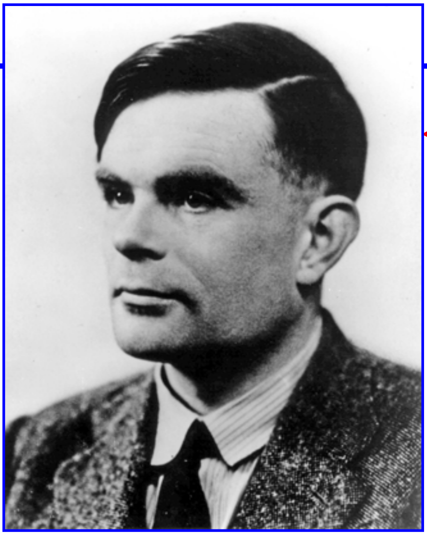
$k=10^{-7}$ and $p=0.6$

Turing-Heisenberg Rapprochement?

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

$$\mu = k\beta^p$$

Binary Information Throughput
a measure of computational capability on device level



Alan Turing

Werner Heisenberg *Ludwig Boltzmann*

Can computational theory suggest new devices?
Stan Williams @ Nanomorphic Forum

We think that all devices operating in an equilibrium with thermal environment are governed by these relations, no matter what state variables are chosen!

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

“Boltzman constraint” on minimum switching energy

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

$$\Delta E \Delta t \geq \hbar$$

“Heisenberg constraints” on device size and speed

Nanoscale Devices

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

~10^-21 J

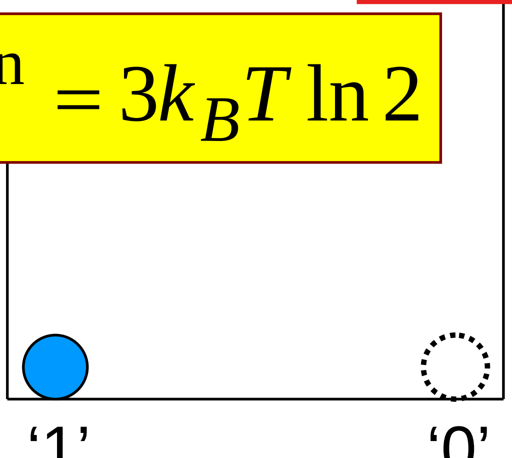
$$E_{sw}^{\min} = 3 k_B T \ln 2$$

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

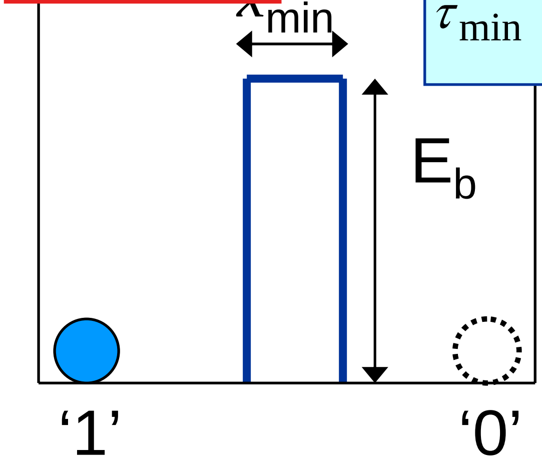
~1.5 nm

$$\tau_{\min} = \frac{\hbar}{kT \ln 2}$$

~40 fs

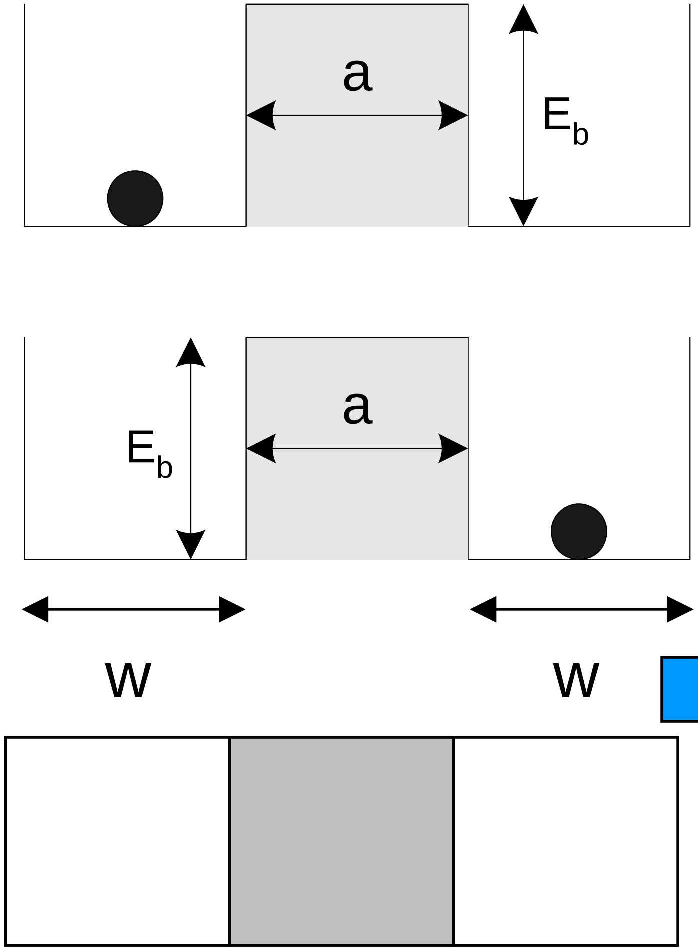


This structure cannot be used for representation/processing information



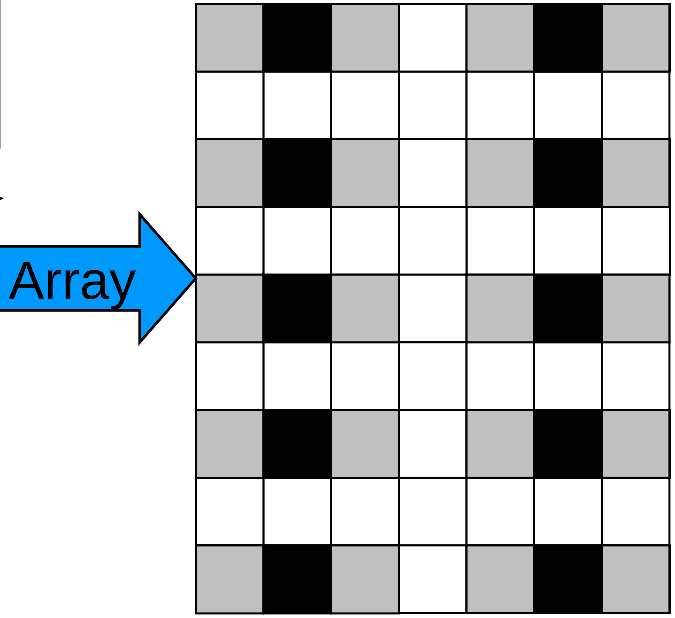
An energy barrier is needed to preserve a binary state

Two-well bit – Universal Device Model



Generic Floorplan
of a binary switch

Joyner tiling



Device density

1) Upper Bound

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

2) IC (ITRS)

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

How can we increase MIPS?

$$n = \frac{1}{8x_{\min}^2} = 6\times10^{12}\,\frac{gate}{cm^2}$$

Number of binary elements

Boltzmann-Heisenberg limit

$$\beta_{\max} \approx 10^{26}\,\frac{bits}{s\cdot cm^2}$$

$$\mu = k\beta^{p} \longrightarrow = \frac{n_{bit}}{t_{sw}}$$

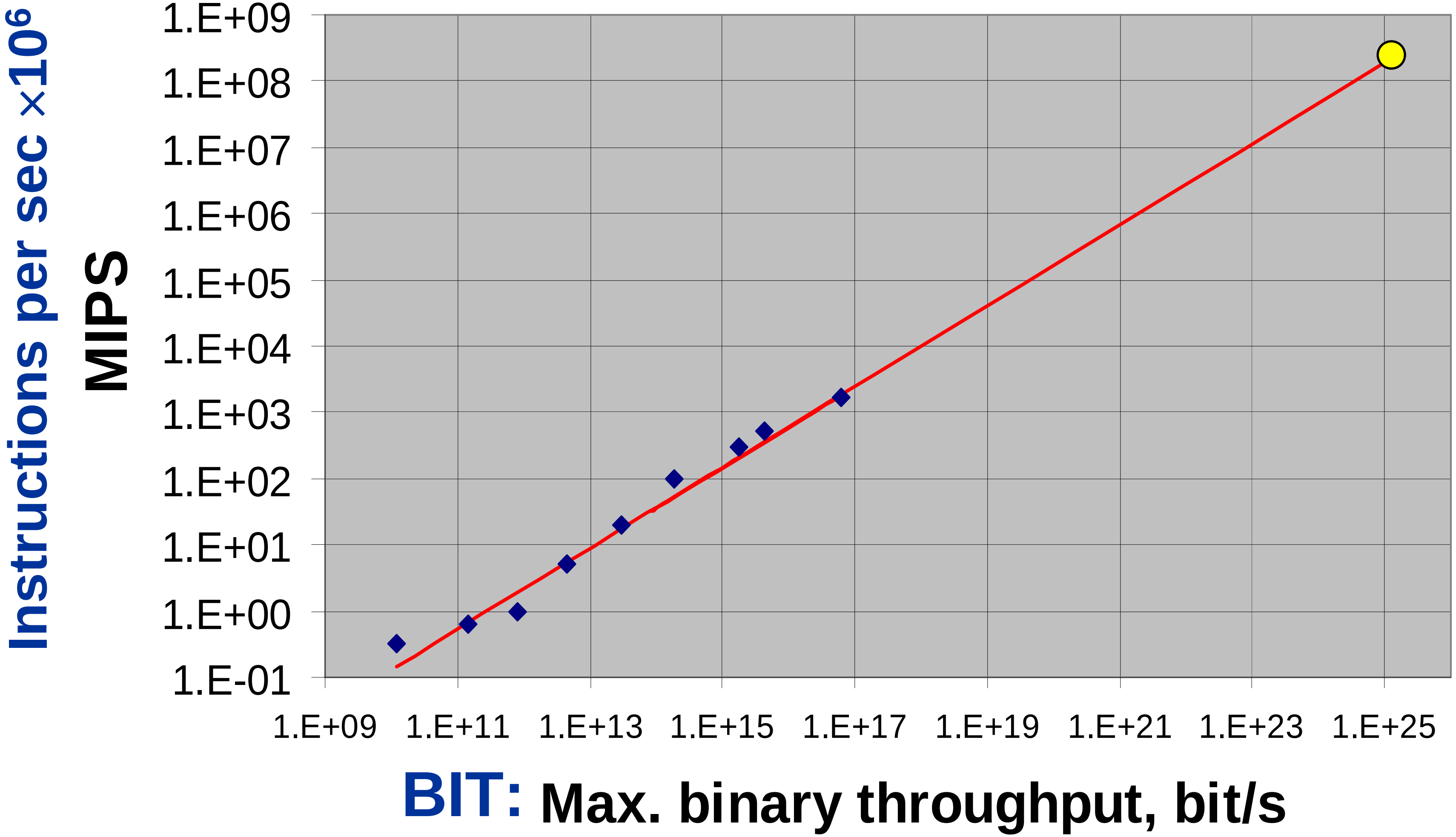
$$t_{sw} = \frac{\hbar}{kT\ln 2} = 4\times10^{-14}\,s(300K)$$

$$\mu_{\max} \approx 4\cdot10^{8}\,MIPS$$

k=10^{-7} and *p*=0.6

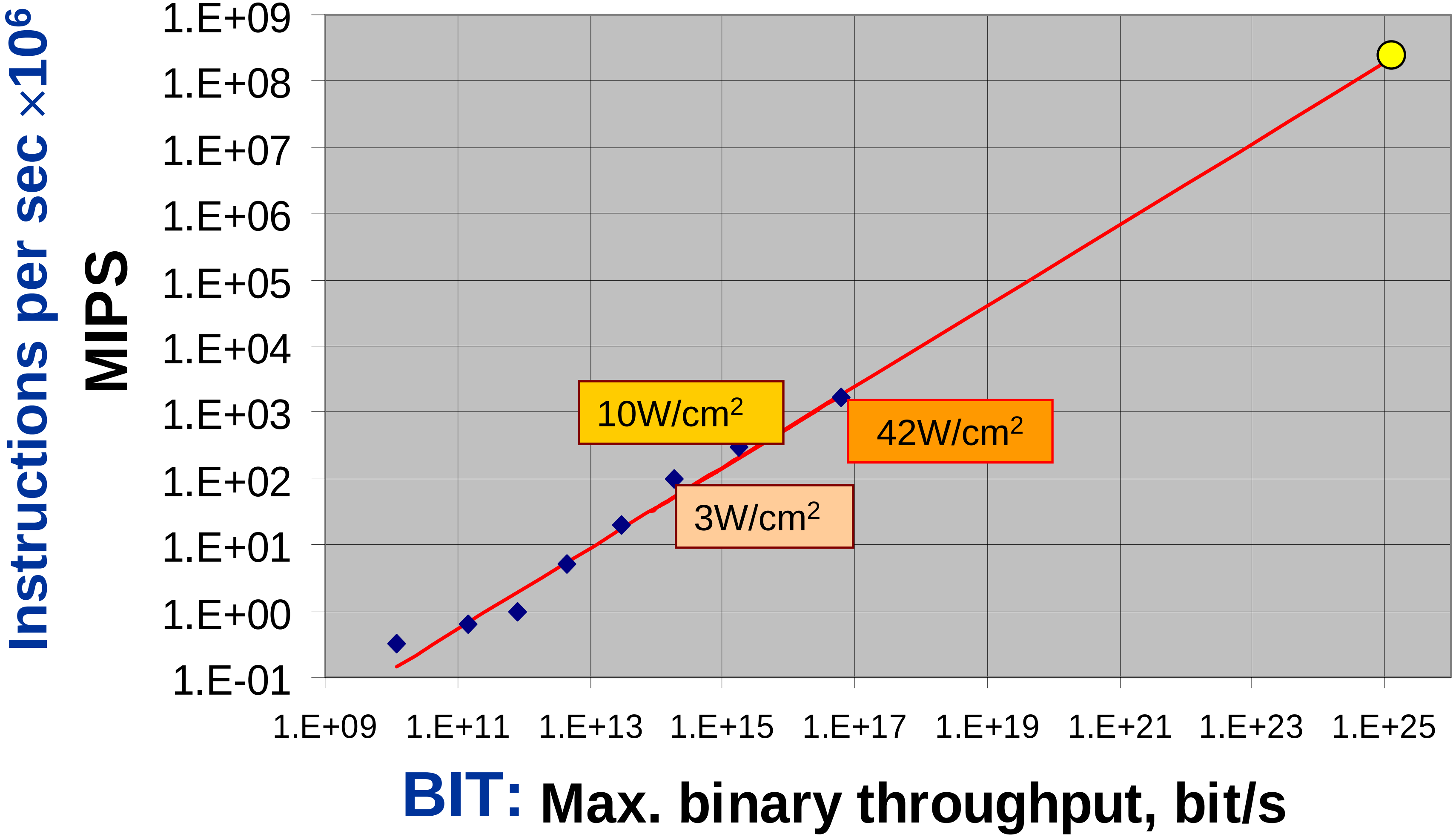
Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and
TSCP Benchmark Scores



Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



Total Power Dissipation (@E_{bit}= kTln(2)) - A Catastrophe!

$$P_{chip} = \frac{n \cdot E_{bit}}{t} = 6 \cdot 10^{12} [cm^{-2}] \cdot \frac{10^{-20}[J]}{4 \cdot 10^{-14}[s]}$$

$$E_{bit} = 3 k_B T \ln 2 \approx 10^{-20}\, J$$

$$\boxed{P_{chip} = 1.5 \times 10^{6} \frac{W}{cm^2}}$$

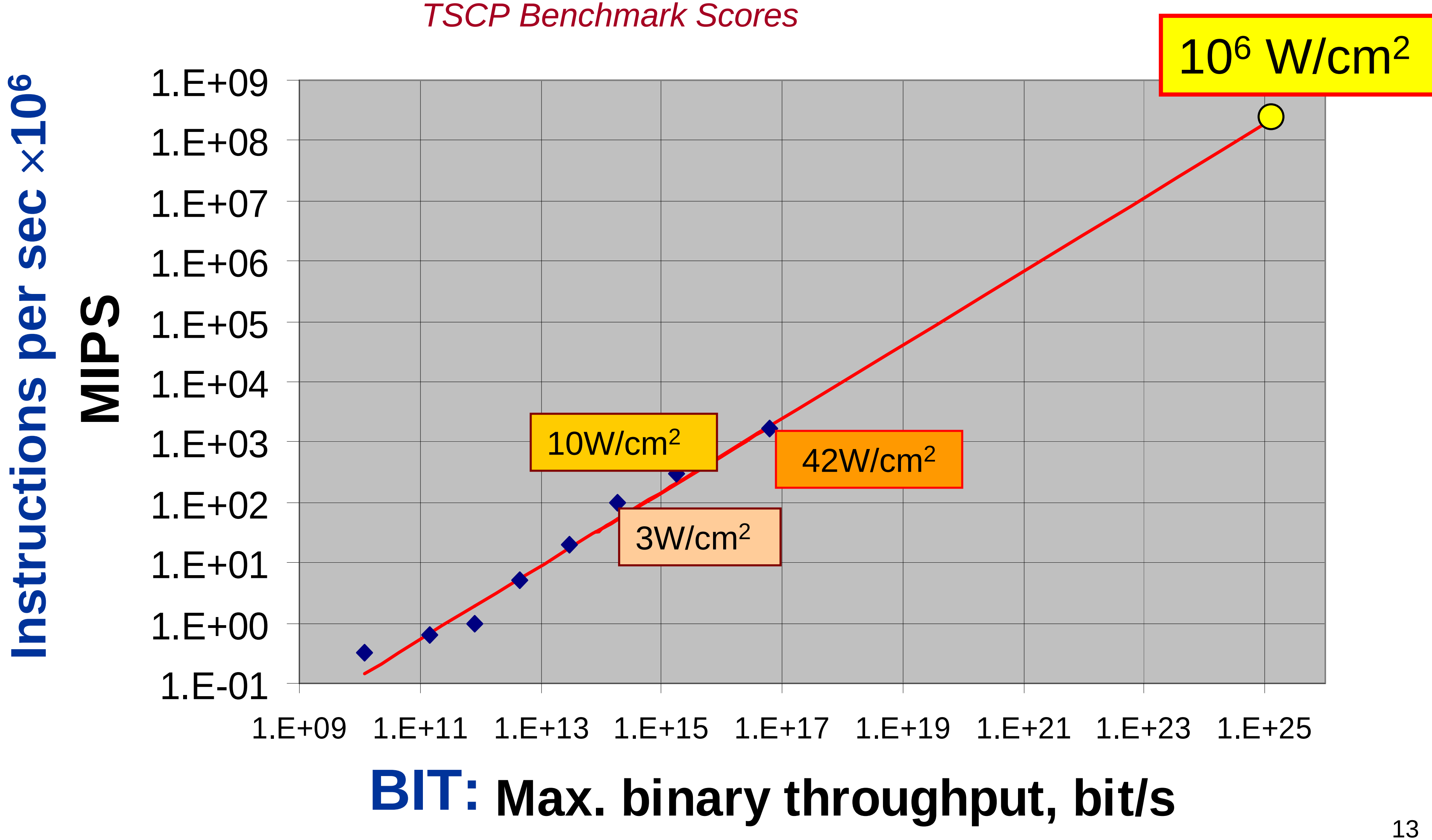
T [illegible]0 K

The circuit would vaporize when it is turned on!

Limits of Cooling?

Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



Energy Costs of Computation:
Energy Consumed and Heat generated

Since each binary transition requires energy E_{bit}, the total power dissipation growth is in proportional to the information throughput:

BIT

$$P = \frac{n_{bit}}{t_{sw}} \cdot E_{bit} = \beta \cdot E_{bit}$$

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

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

MIPS

$$\mu = f(\beta)$$

Can we change *f* ?

We don’t know how to remove that much heat!!

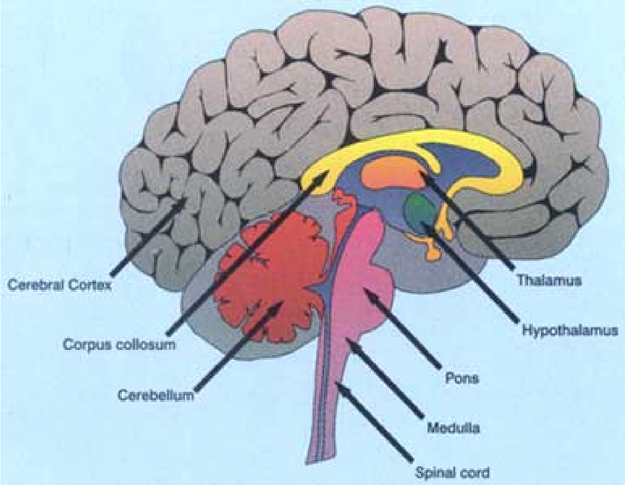
A universal relation for information processing devices

Biological Computation?

‘Computers Are Like Brains? Don’t They Wish’

The Wall Street Journal, July 9 2008

Most complex information-management system in the universe…



	Dell 8250 (Pentium® 4)	Brain
Mass	~25 kg	1.4 kg
Volume	34200 cm^3	1350 cm^3
MIPS	~10^3 MIPS	10^8 MIPS
BIT	<10^{16} bit/s	10^{19}bit/s
Power	**200 W**	**30 W (max)**
	~ 5 MIPS / W	**3x10^6 MIPS / W**
	5x10^6 k$_B$T / bit	**700 k$_B$T/bit**

When will computer hardware match the human brain?

A CMOS machine at the limits of scaling would use prodigious amounts of power

Dell 8250 (Pentium®4)

10^3 MIPS

200 W

x 10^5

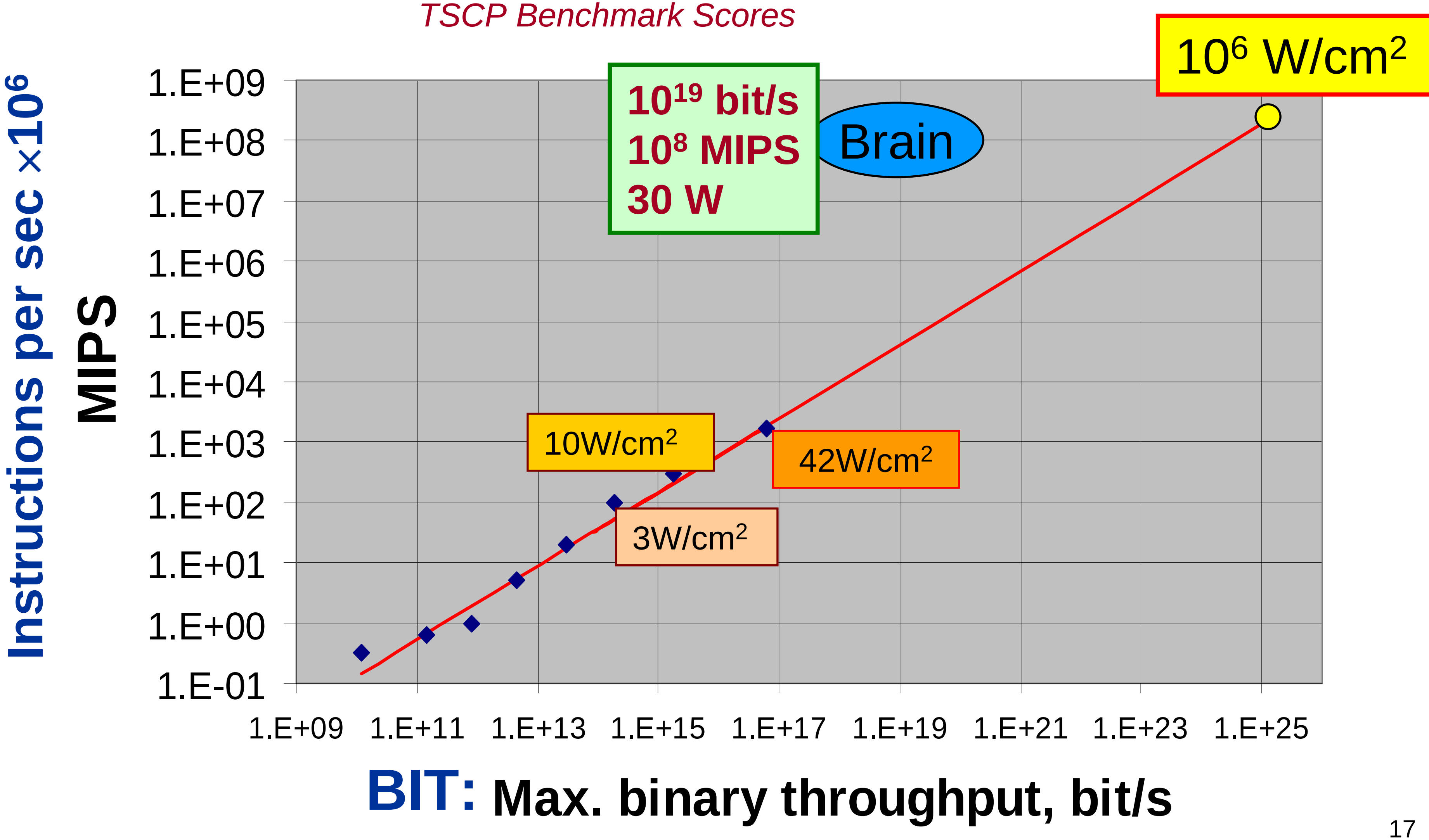
x 10^5

10^8 MIPS

=20,000 000 W

Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



PIONEERS IN COLLABORATIVE RESEARCH®

Chip Multiprocessors

Ralph K. Cavin III and Victor V. Zhirnov

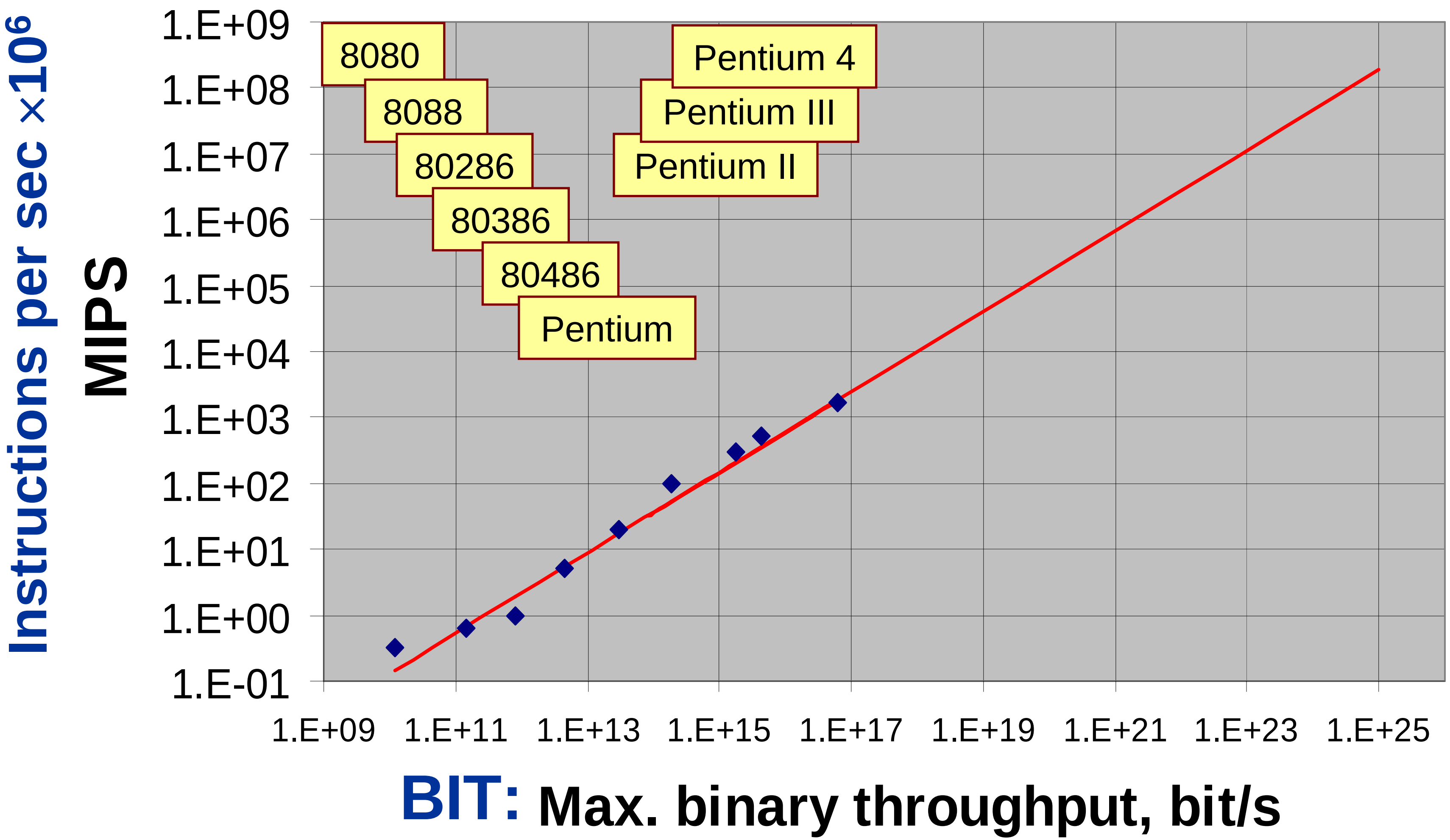
Semiconductor Research Corporation

IEEE/ACM International Symposium on Nanoscale Architectures

San Jose, CA, October 21-22, 2007

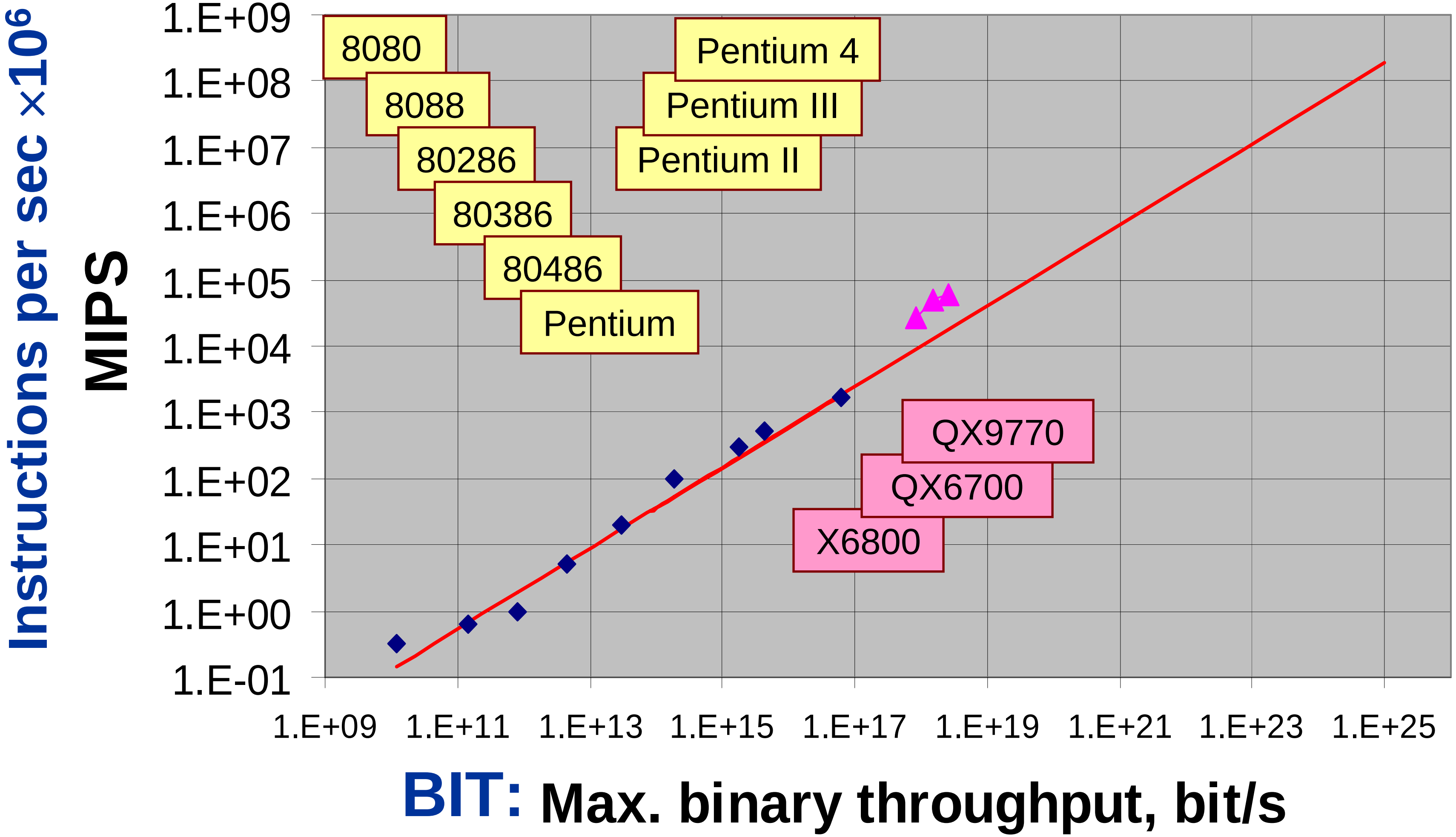
Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



Multi-Core Architectures

Multi-Core Architectures: A number *K* of light-weight processors instead of one heavy-weight processor

A Multi-Core processor consists of a total of *N* binary switches organized in *K* supercells or cores. Each core in this organization is a lighter-weight general-purpose information processor, containing *M* binary switches: *M*=*N*/*K*

$$M < N \Rightarrow E_{b_{\min}}\left(M\right) < E_{b_{\min}}\left(N\right)$$

for the same error probability

Favorable Multi-Core Postulates

1) The collective action of all *K* cores is equivalent to the action of the single-core

2) All processors are engaged in useful work

3) Each core contains an error-detecting mechanism

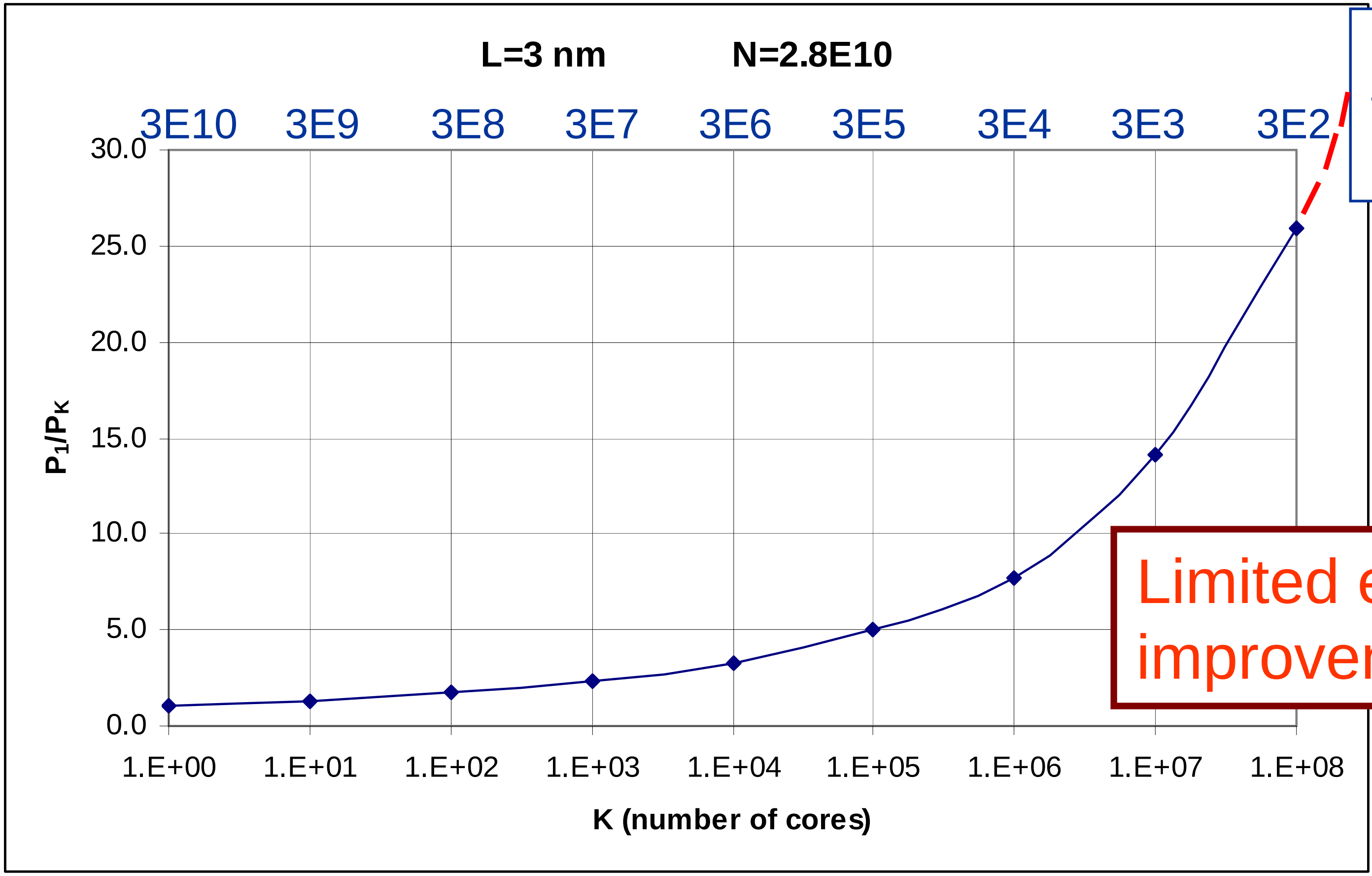
4) The other cores are able to wait until the failed microtask computation on a core repeats the microtask to generate correct answer

Extreme Multi-Core Analysis

Power consumption by K cores:

$$P_K = K \cdot \frac{M}{t_{sw}(M)} \cdot E_{b\min}(M) = \frac{N}{t_{sw}(M)} \cdot E_{b_{\min}}(N)$$

Energy improvement factor



Number of transistors in each core

Limited energy improvement in 2D

“Coreness” / “Weight”- Dilemma

- The is a limit for a maximum number of transistors in 1cm^2 of chip area

$$N_{\max} \sim 10^{10}\, cm^{-2}$$ (L$_g$=5 nm)

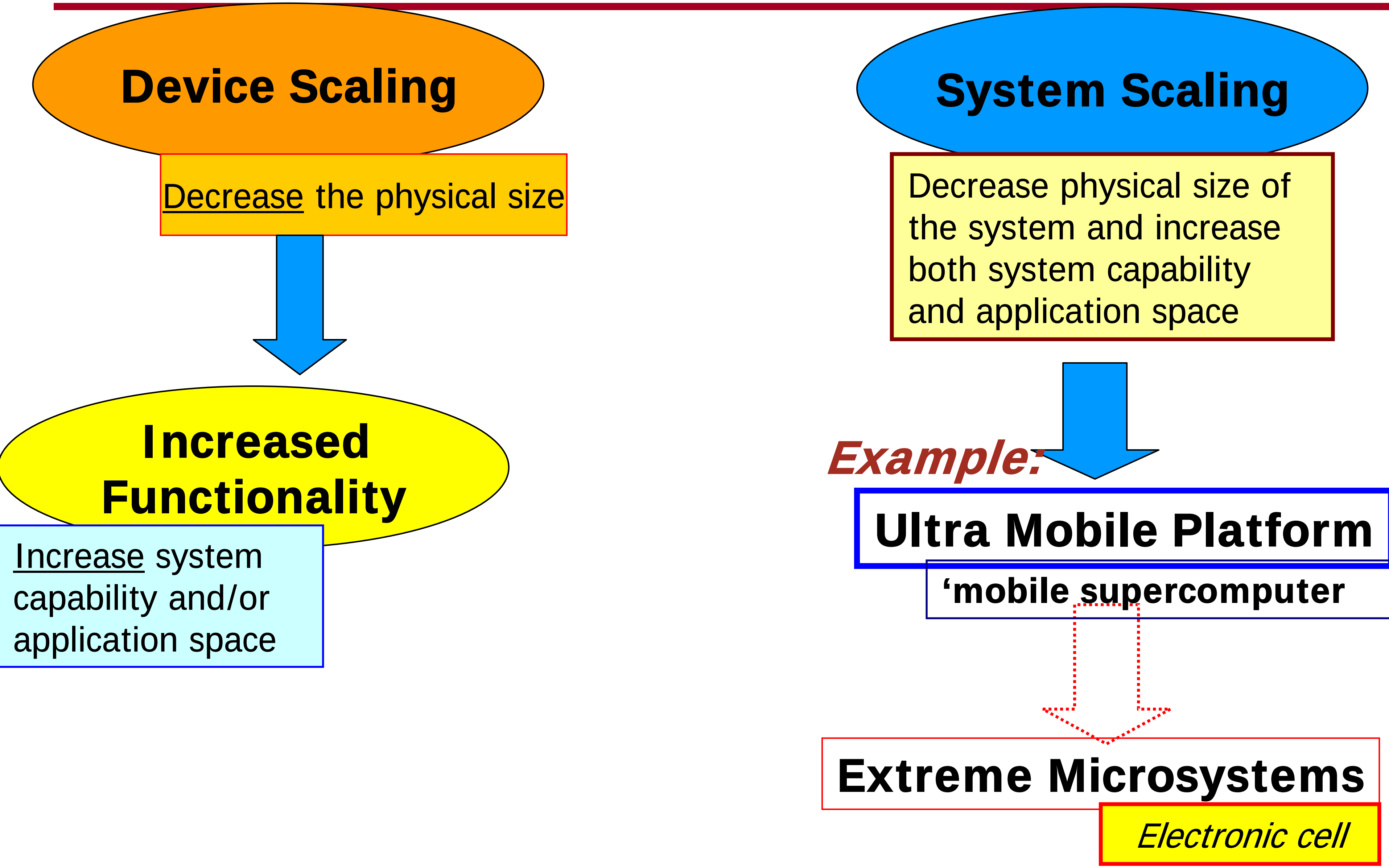
- A Multi-Core Information processor consists of a total of *N* binary switches organized in *K* supercells or cores.
  - Each core in this organization is a lighter-weight general-purpose information processor, containing *M* binary switches: *M=N/K*
  - In the limit:

$$M = \frac{10^{10}}{K}$$

What is smallest M?

System scaling limits need to be understood

Different Facets of Scaling



Device Scaling

Decrease the physical size

Increased Functionality

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

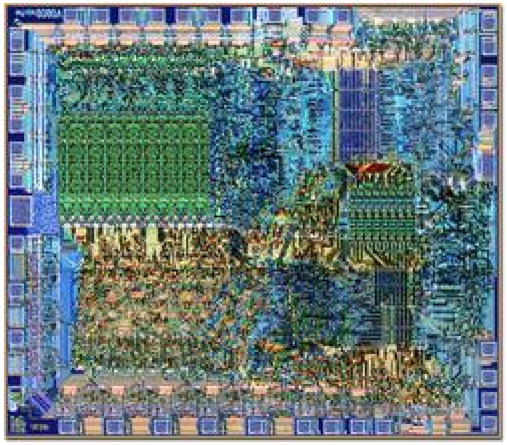
'mobile supercomputer

Extreme Microsystems

Electronic cell

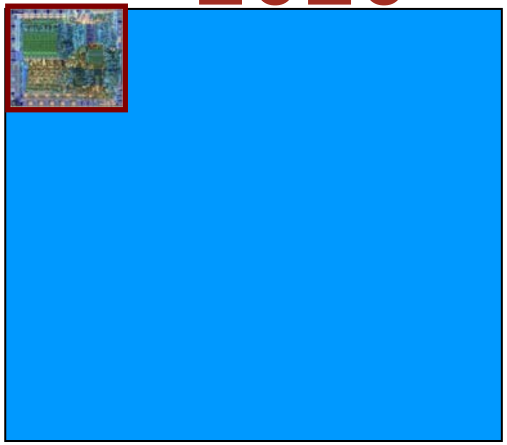
Scaling of 8080 MPU

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

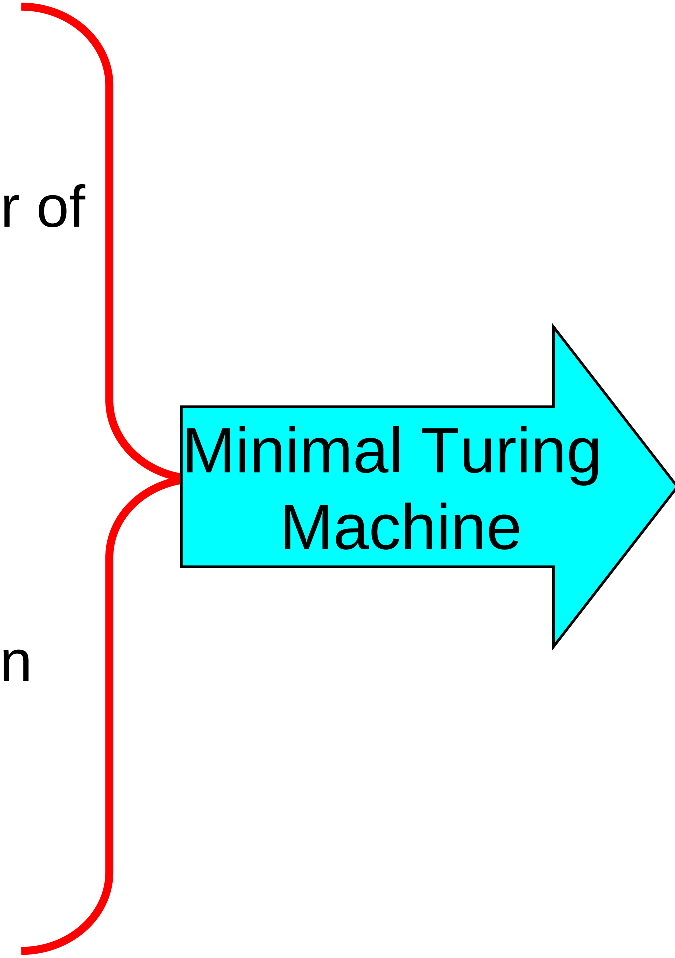


Technology:	CMOS
Feature size:	6 nm
# of transistors:	4500
Die size:	5 μm x 4 μm
Voltage:	~0.5 V
Frequency:	~2 MHz-1GHz
Power:	~10nW-10μW

System scaling limits

- **Multi-core CPU**
  - What is the maximum possible number of cores in multi-core processors

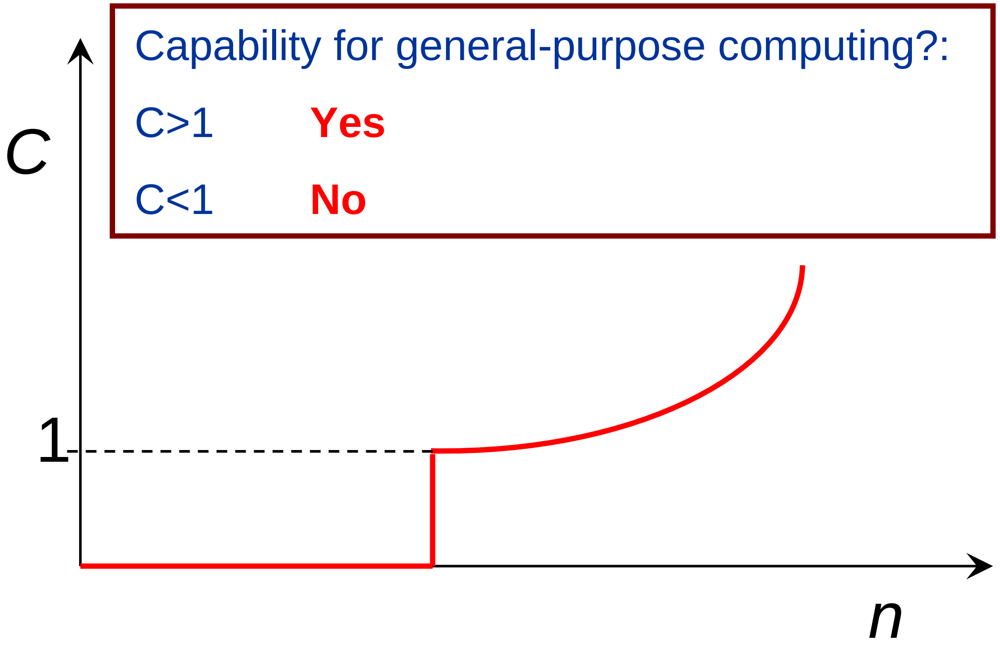
- **‘Mobile supercomputers’**
  - What is the smallest possible size of an intelligent ‘piconode’?



Minimal Turing Machine

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)

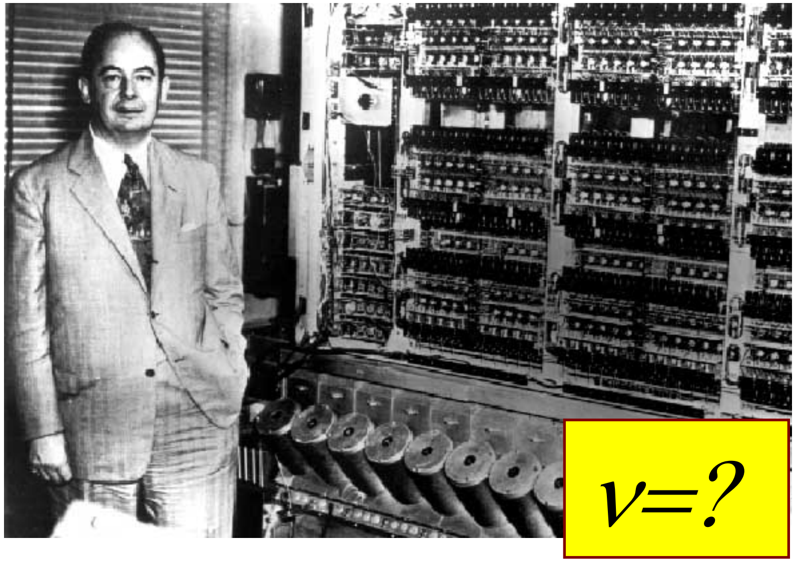


Capability for general-purpose computing?:

C>1 **Yes**

C<1 **No**

Von Neumann threshold

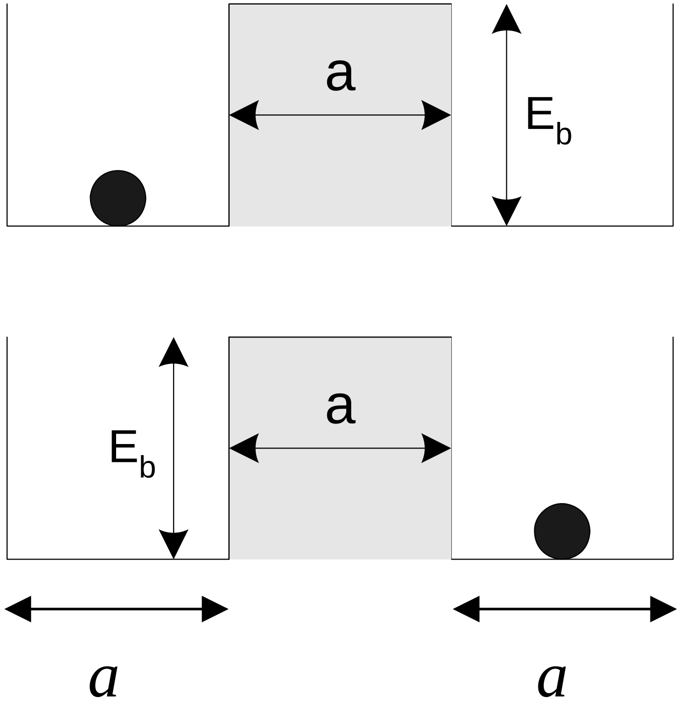


ν=?

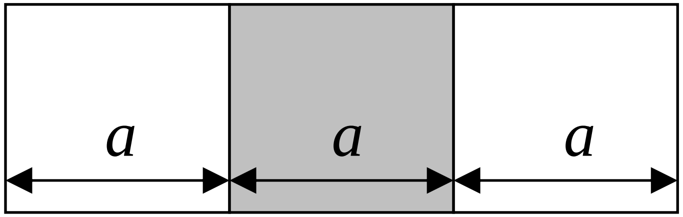
‘Minimal’ Turing Machine

Binary switch abstraction:

Generic floorplan and energetics



Generic Floorplan of a binary switch



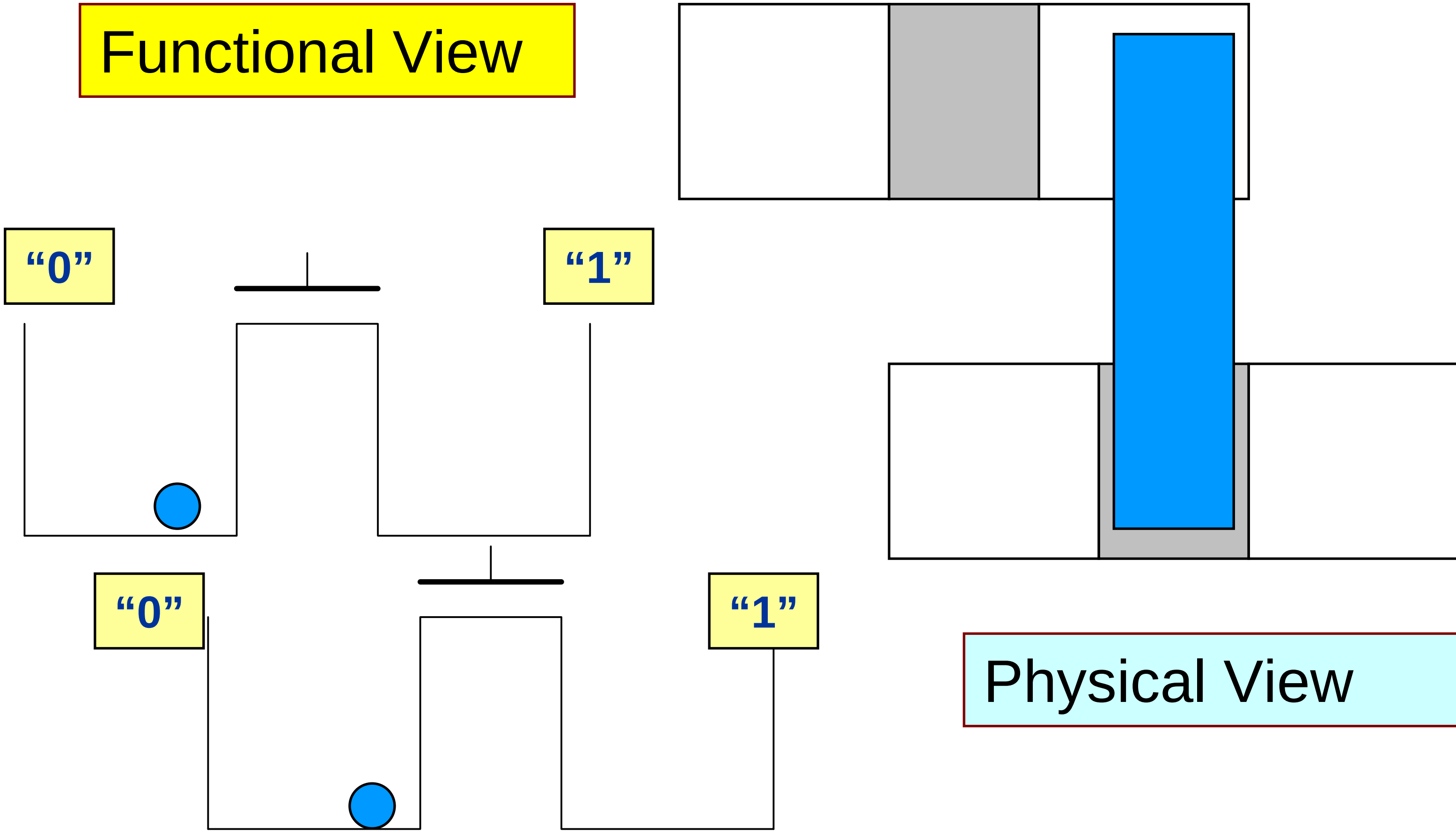
$$a = \frac{\hbar}{\sqrt{2mkT\ln 2}} = 1.5nm$$

$$Area_{\min} = 3a^2$$

$$E_{sw_{\min}} = 3k_BT$$

$$\varepsilon = k_BT \left(\frac{J}{tile}\right)$$

Connected Binary Switches



Functional View

“0”

“1”

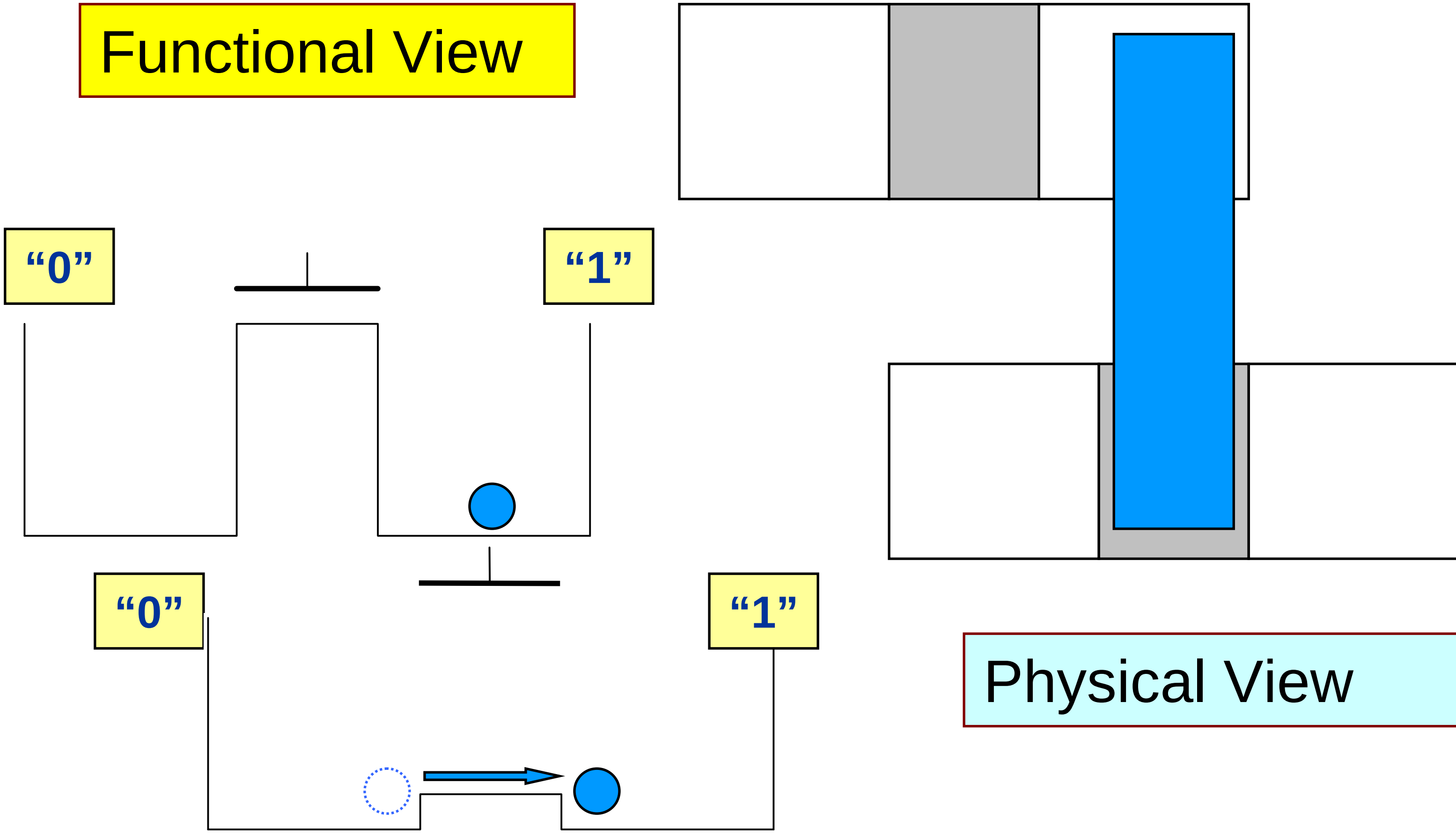
“0”

“1”

Physical View

Information-bearing charge

Connected Binary Switches

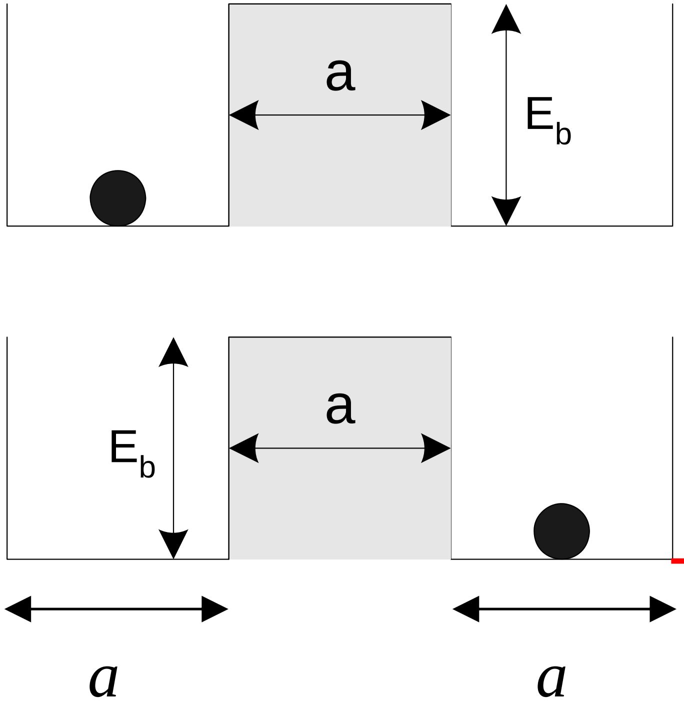


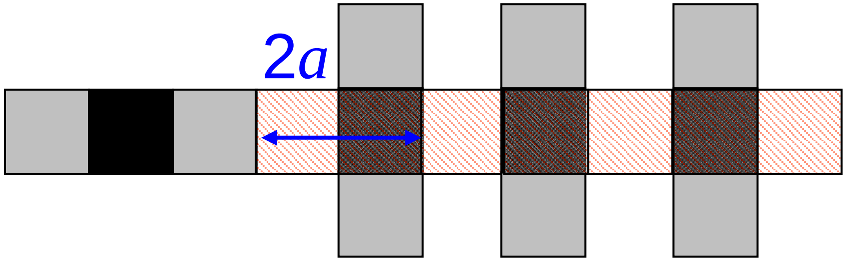
Functional View

Physical View

Information-bearing charge = Barrier-forming charge

Interconnect abstraction: *Extended Well Model*

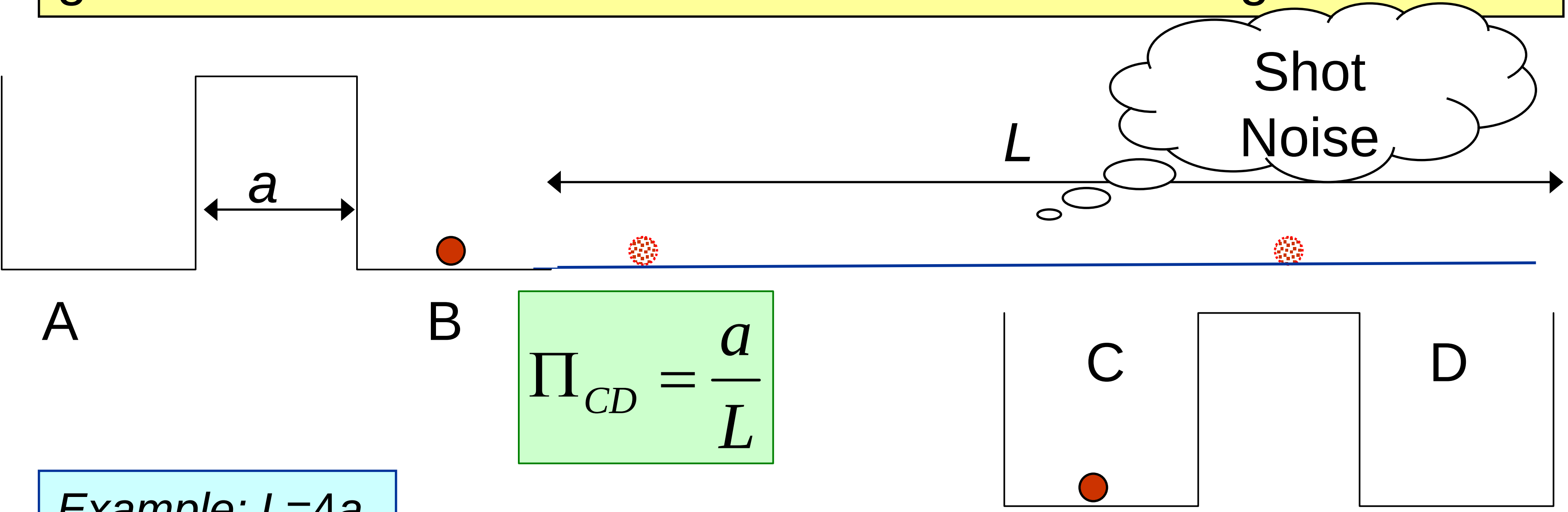




$$\boxed{L_{\min} = 2a \cdot F}$$

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



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

Example: *L*=4*a*

N=1→Π<0.25

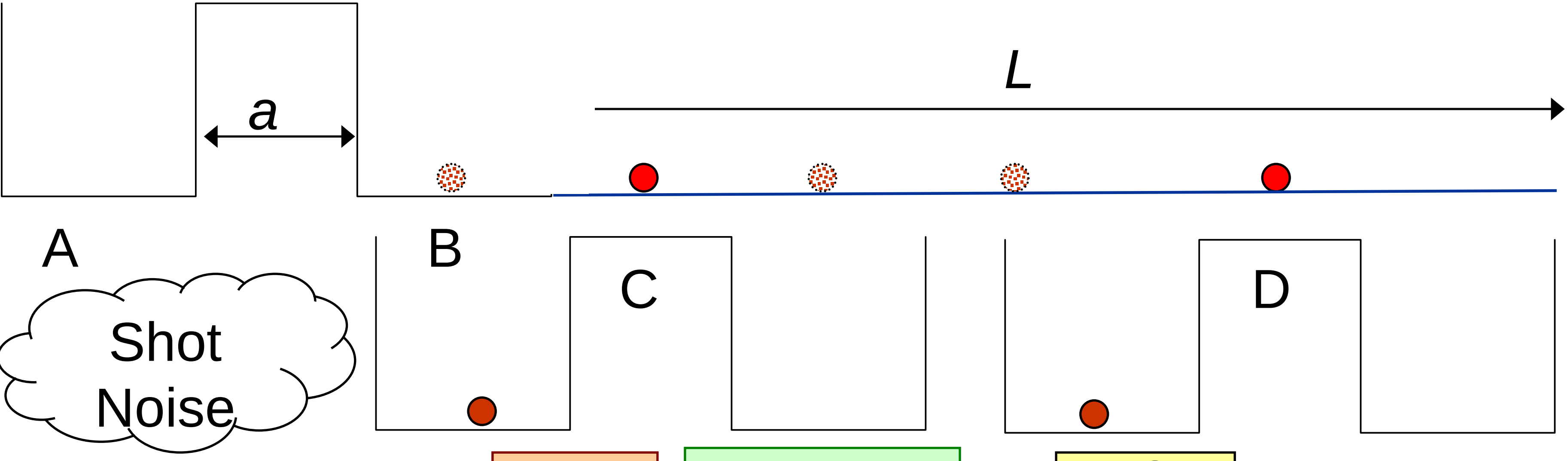
In General:

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

N – the number of electrons

Connecting Binary Switches via Wires (*L>4a, N electrons*)

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



$L>2Fa$

F- fan out

N – the number of electrons

$F=2$

$L=4a$

$N_{min}=5$

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

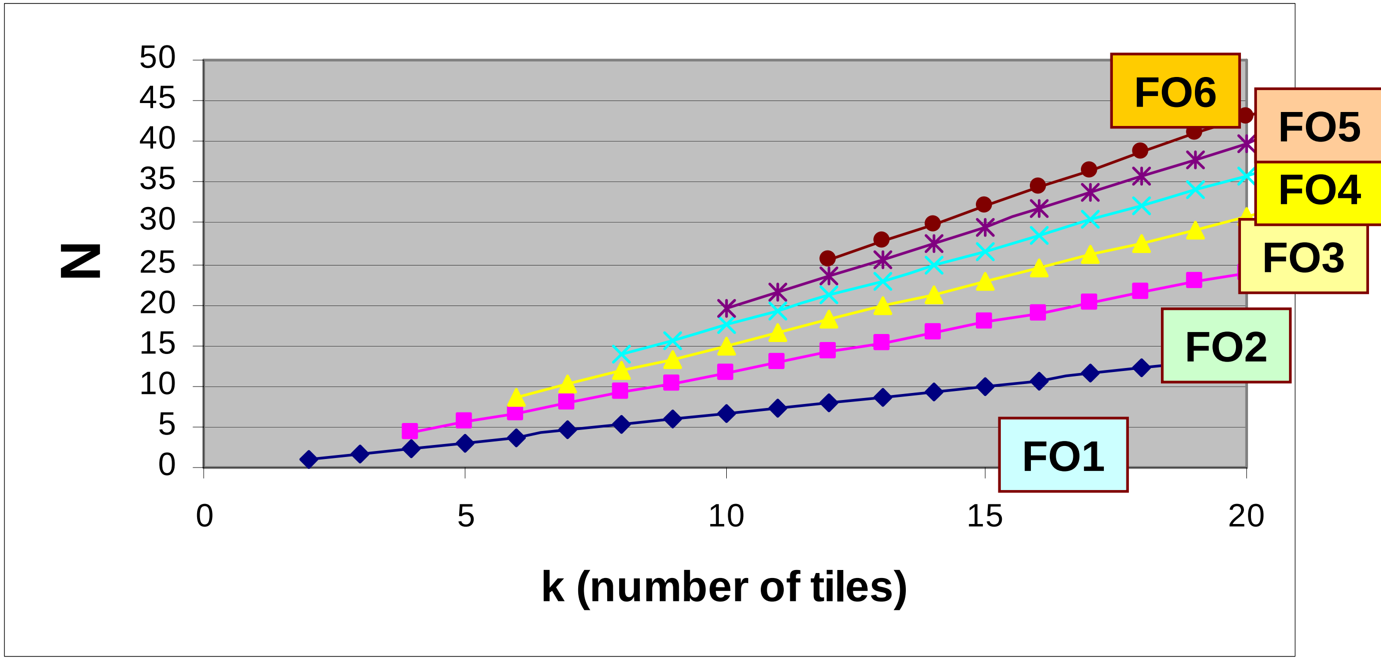
N	Π
1	0.06
2	0.19
3	0.33
4	0.47
5	0.58
6	0.68

Minimum number of electrons in interconnect line for communication and fan-out

N - number of electrons *F – fan-out* *k – number of tiles*

$$\Pi = \left(1-\left(1-\frac{a}{L}\right)^{N}\right)^{F} \xrightarrow[\boxed{\Pi=\frac{1}{2}}]{\boxed{\frac{L}{a}=k}} \frac{1}{2} = \left(1-\left(1-\frac{1}{k}\right)^{N}\right)^{F}$$

$$N = \frac{\ln\left(1-\frac{1}{2^{F}}\right)}{\ln\left(1-\frac{1}{k}\right)}$$



Minimum switching energy for connected binary switches

$$E_{sw}=3E_b+NE_w=(N+3)k_BT\ln 2$$

Minimum fan out

$F=2 \quad L=4a$

$N_{min}=5$

$E_{sw}=8k_BT\ln 2$

Typical fan out

$F=4 \quad L=8a$

$N_{min}=14$

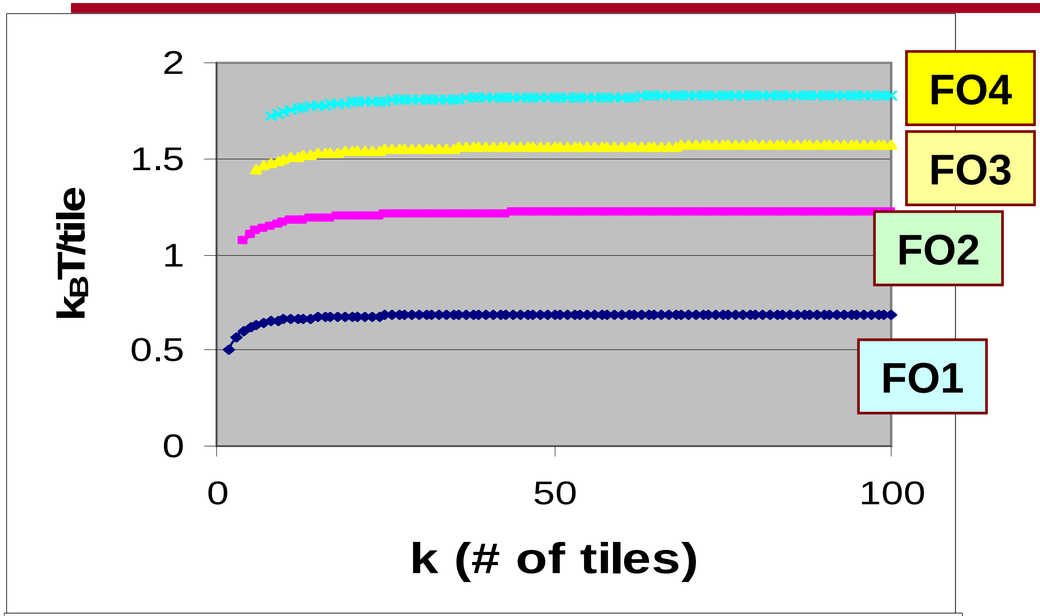
$E_{sw}=17k_BT\ln 2$

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

Communication between binary switches takes more energy than does changing switch state

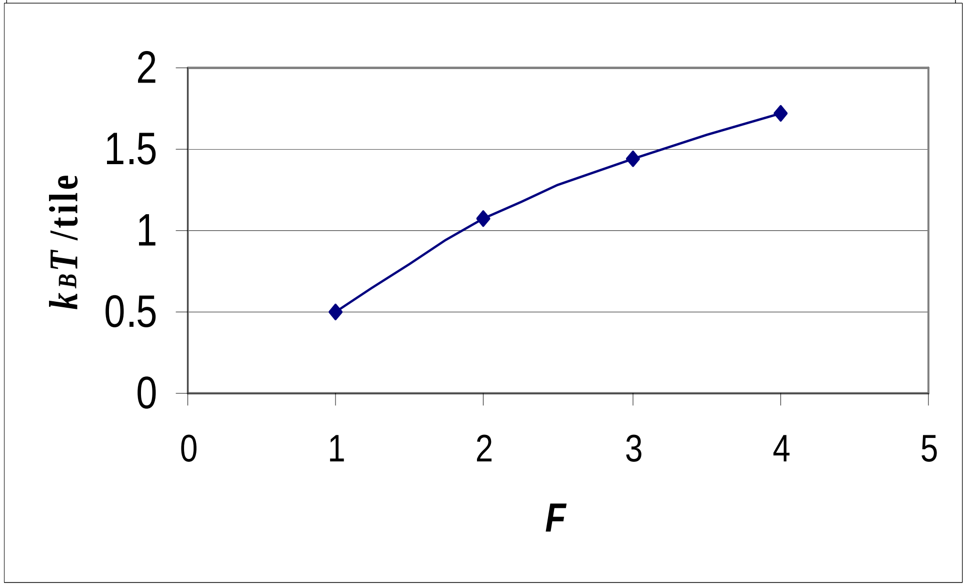
Can we make communication more energy efficient?

Energy per interconnect tile



Long interconnect limit

$$\langle \varepsilon \rangle = 1.33 \ \frac{k_B T}{tile}$$



Minimum interconnect limit

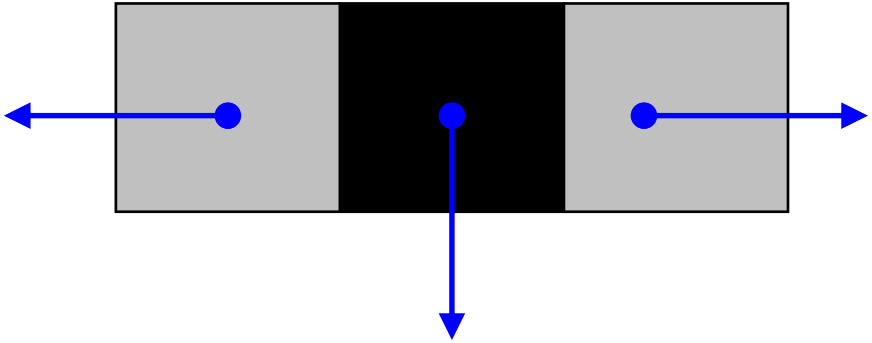
$$\langle \varepsilon \rangle = 1.18 \ \frac{k_B T}{tile}$$

$$\varepsilon \sim k_B T/\text{tile}$$

Floorspace Expenses of Communication between Binary Switches

Assumption: For each of 3 tiles of Binary Switch and for a fan-out **of three**, we need at least:

One contacting interconnect tile (3 total) and one connecting interconnect tile (3 total)

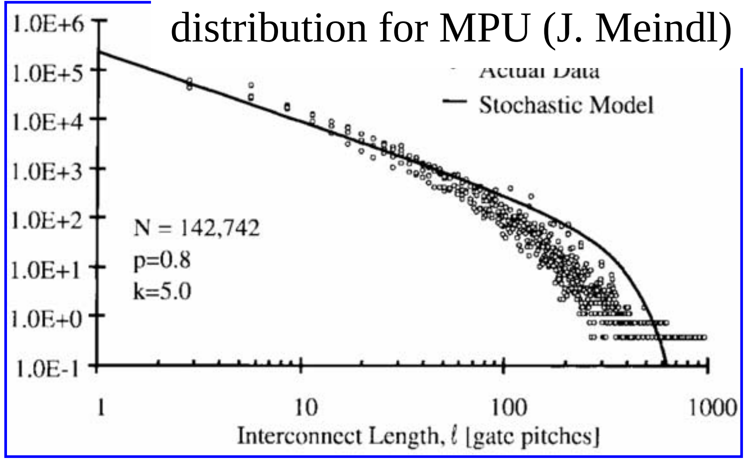


Total **6 inteconnect** tiles per binary switch

$$\boxed{L_{\text{int}} \sim 6a}$$

Reality check:

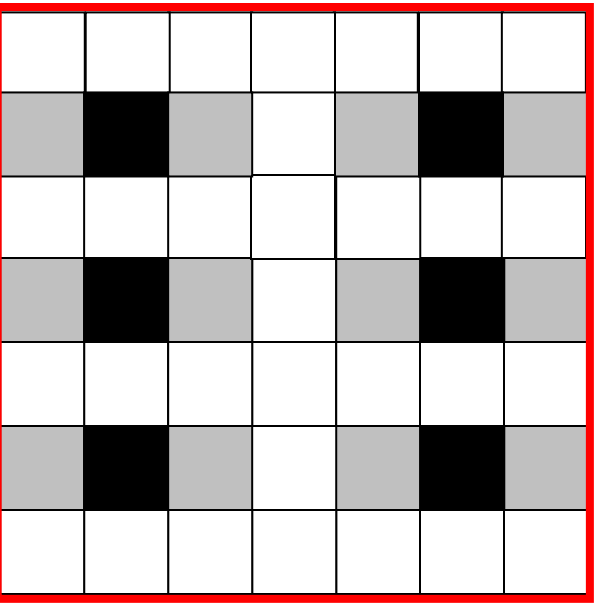
A typical interconnect length distribution for MPU (J. Meindl)



n, cm^{-2}	$\bar{L}(n)/L_g$
1.E+02	4.1
1.E+04	6.4
1.E+06	8.3
1.E+08	9.7
1.E+10	10.5

Digital circuit abstraction:
Generic floorplan and energetics and speed

Switching energy of one binary switch in a circuit (FO3)



3 switch tiles

$$E_{sw}=3E_b+6E_b=9k_BT\ln 2$$

6 wire tiles

Operational energy of a circuit of *n* binary switches:

(50% activity)

$$E_{op}=\frac{9}{2}nk_BT\ln 2$$

$$Area_{\min}=n\cdot 8a^2$$

Joyner tiling

Switching delay of one binary switch in a circuit:

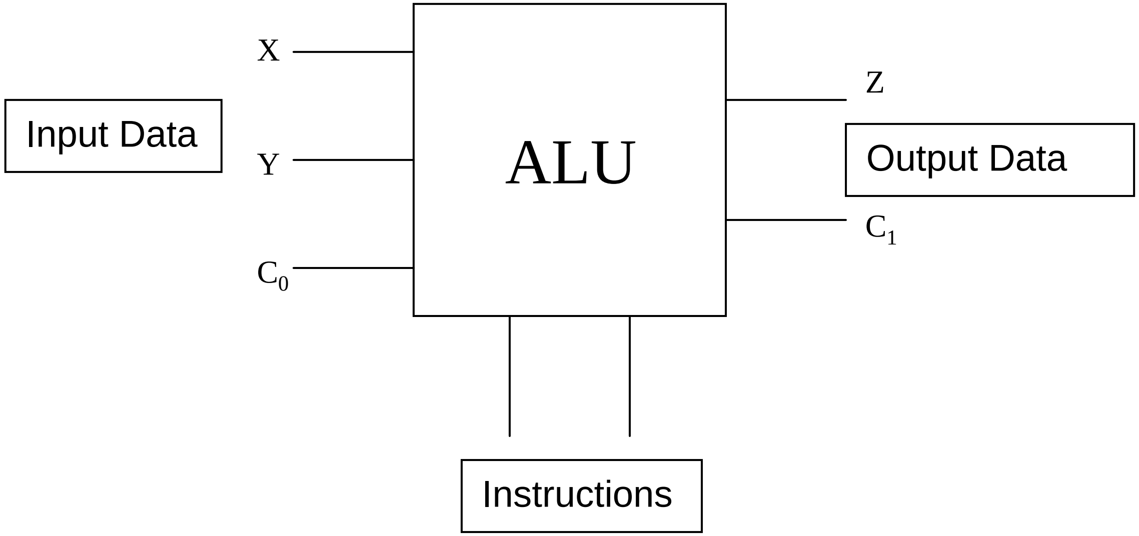
Speed: $\tau_{\min}$/tile

$$\tau_{\min}=\frac{\hbar}{kT\ln 2}$$

~40 fs

$$t_{sw}=9\,\tau_{min}$$

1-bit ALU



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**))

Jan Rabaey,
Digital Integrated Circuits

Minimal ALU abstraction: *Energetics*

$$E_{ALU} = \frac{9}{2} \cdot 98 \cdot k_B T \ln 2 \sim 300 k_B T$$

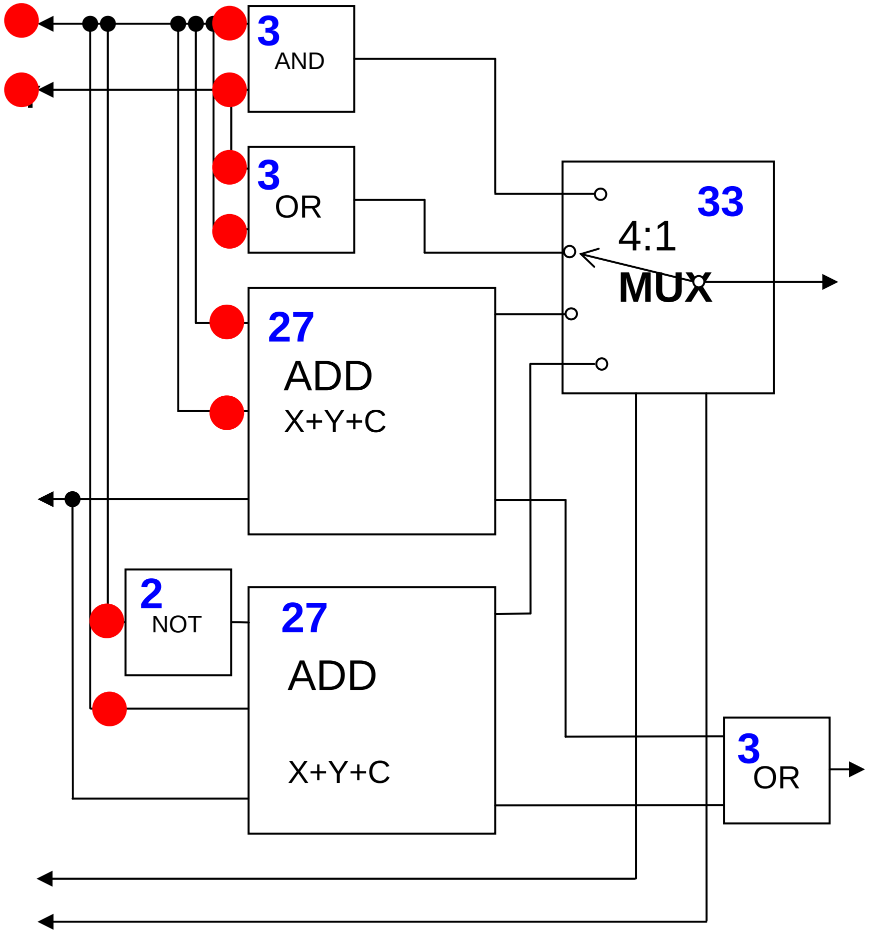
Energy efficiency: $$\eta = \frac{E_{op}}{E_{ALU}}$$

$$E_{AND} = \frac{9}{2} \cdot 3 \cdot k_B T \ln 2 \sim 9 k_B T$$

$$\eta_{AND} \sim 3\%$$

$$E_{ADD} = \frac{9}{2} \cdot 27 \cdot k_B T \ln 2 \sim 84 k_B T$$

$$\eta_{ADD} \sim 28\%$$

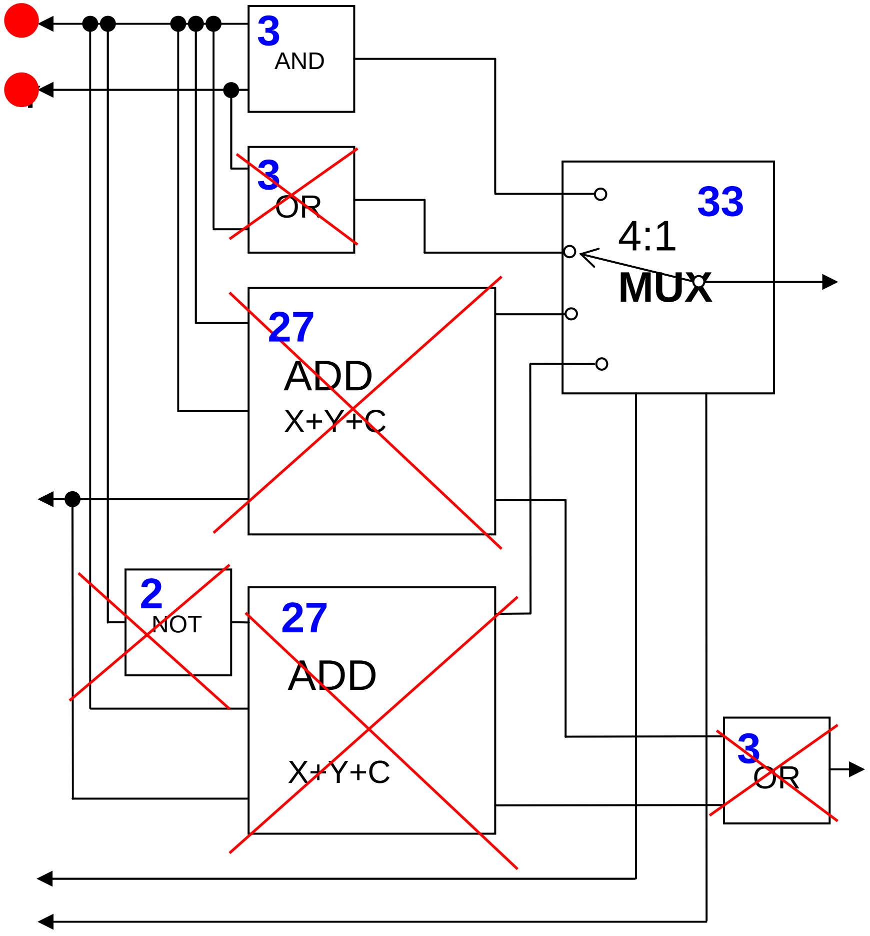


All 4 units execute even though only one output is used

Total: 98 devices

Can we increase ALU efficiency?

- De-parallelize inputs ?
  - Two input selectors are needed
    - Two 1:4 DMUX
      - 33 devices each



Example: AND operation

Active device count: 101

$$\eta_{AND} \sim 3\%$$

Carnot's equivalent for Computational Engine?

Thermodanomic entropy analysis may provide new insight on chip design

Minimal ALU abstraction: *Timing*

$t_1 = 9\,\tau_{min}$

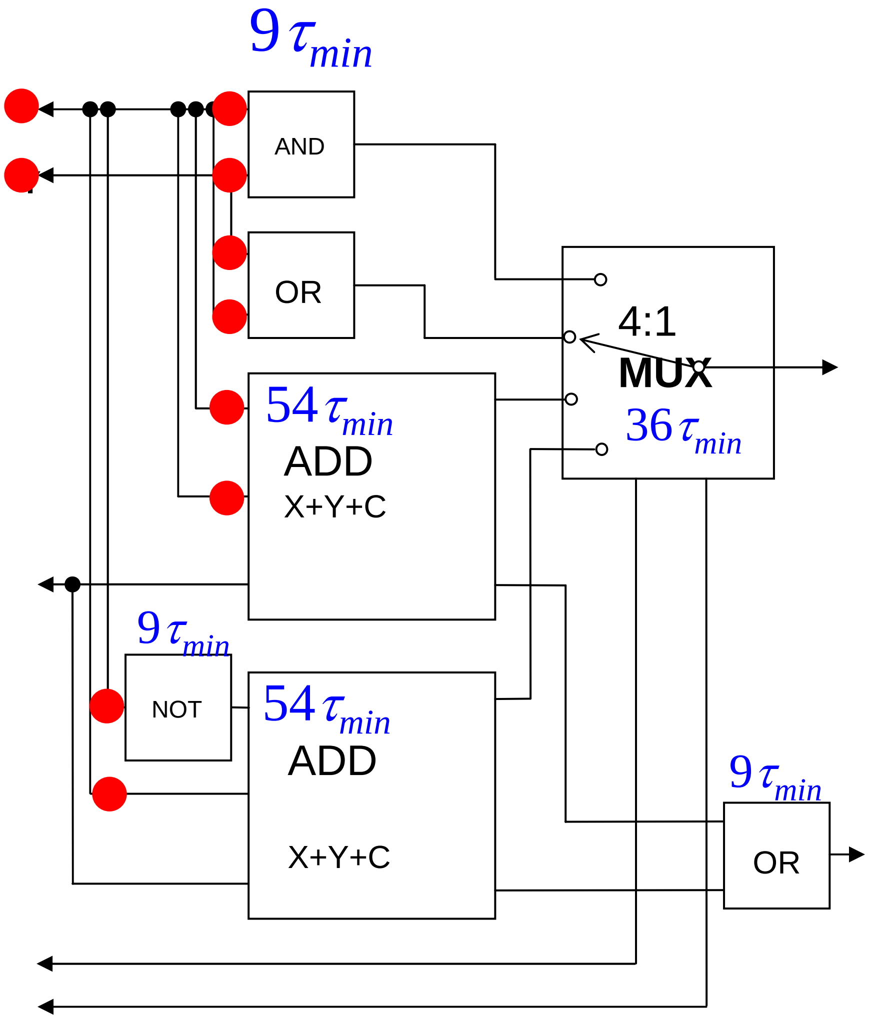
~360 fs

$t_n = n \times 9\,\tau_{min}$

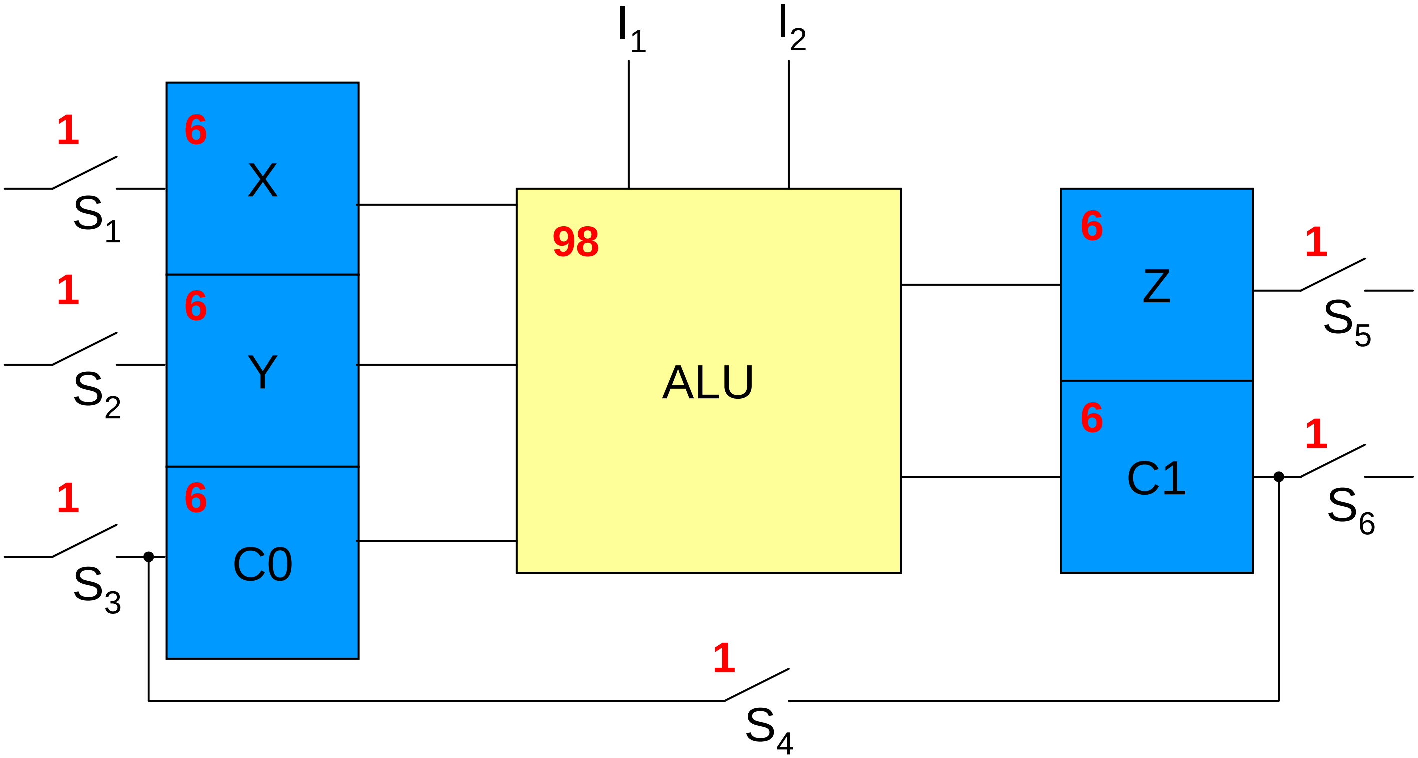
n= # cascades

$t_{ALU} \sim 50\tau_{\mathrm{min}}$

~2 ps

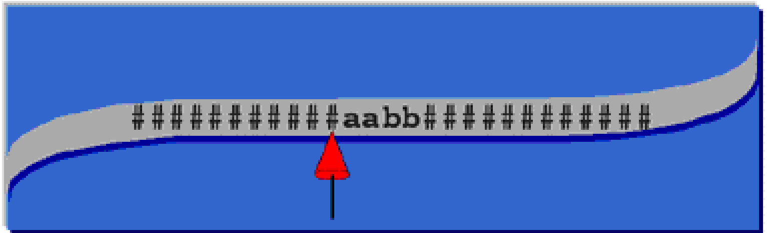


Minimal CPU



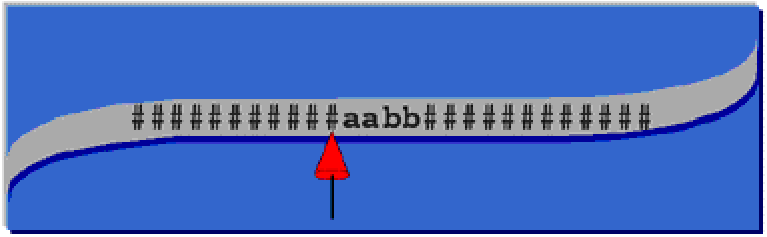
Total: 134 devices

$$E_{CPU} = \frac{9}{2} \cdot 135 \cdot k_B T \ln 2 \sim 420 k_B T / cycle$$



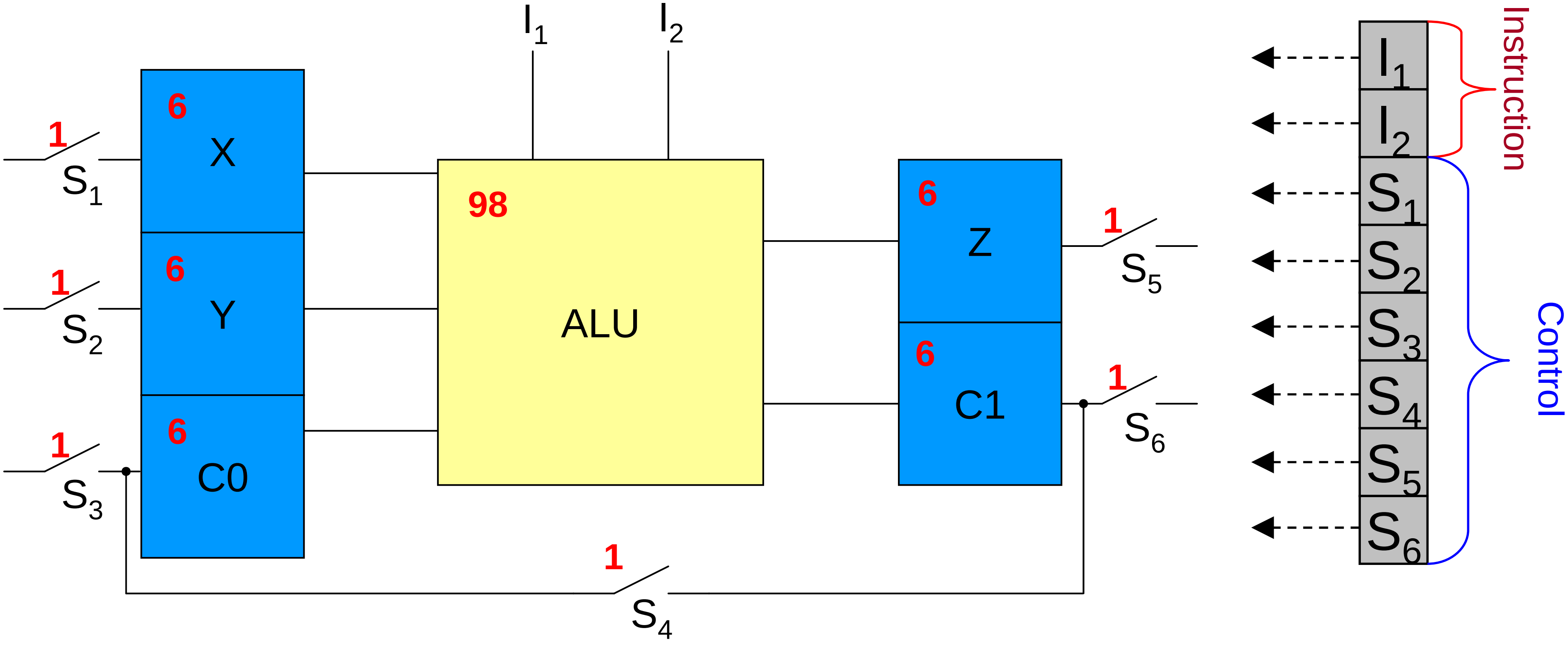
Not included

Minimal Turing Machine



Memory

Program Counter



8 bit per cycle

Program Memory per operation

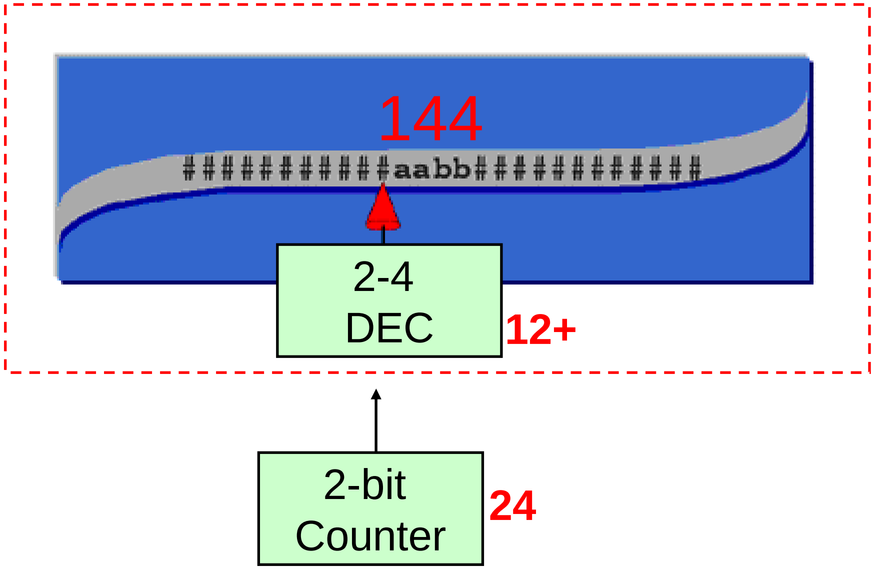
Operation 1: **X** AND **Y**
Operation 2: **X** OR **Y**
Operation 3: (**X**+**Y**)
Operation 4: (**X**+(NOT **Y**))

3 cycles per ALU operation

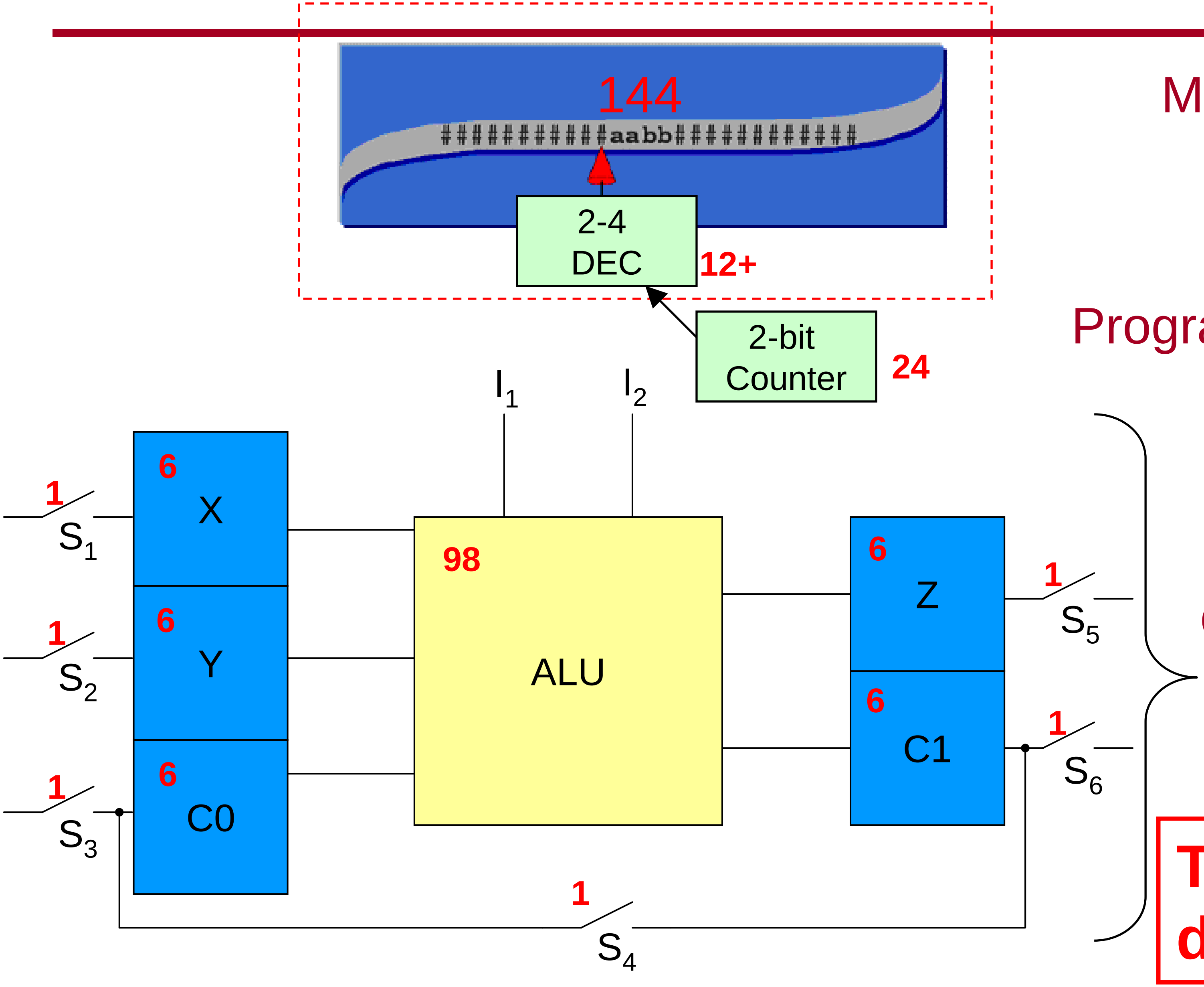
IN
Op
OUT

8 bit per cycle

24 bit Memory per operation



Minimal Turing Machine



Memory

Program Counter

CPU

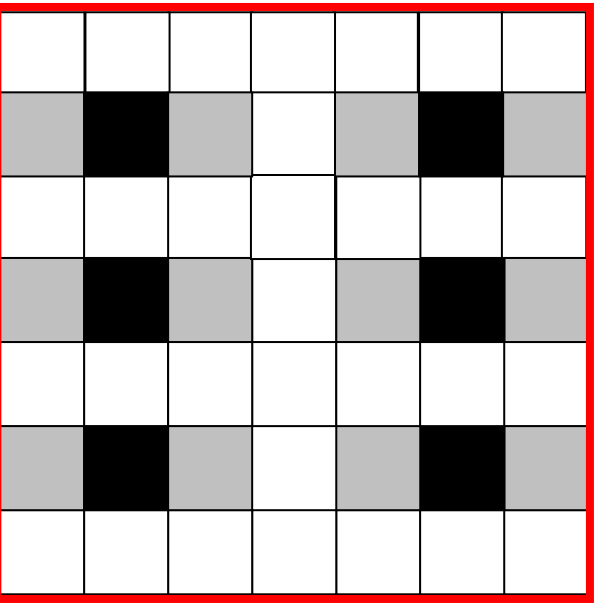
Total: 314 devices

Turing Machine Implementation:
Generic floorplan and energetics

Von Neumann threshold: $n=314$

Joyner tiling:

$$Area_{\min} = n \cdot 8a^2 = 314 \cdot 8a^2 \approx 2500a^2 = 50a \times 50a$$



$$a_{min} = 1.5\text{ nm}$$

$$Area_{\min} = 75nm \times 75nm$$

Operational energy of the *Minimal Turing Machine*

$$E_{op} = \frac{9}{2} n k_B T \ln 2 \approx 980 k_B T / cycle = 4 \cdot 10^{-18} \frac{J}{cycle}$$

Per full CPU operation:

$$E_{op} = 3 \cdot 4 \cdot 10^{-18} \frac{J}{cycle} \approx 10^{-17} \frac{J}{operation}$$

Minimal Turing Machine:
A summary

Devices: 314

$Area = 75nm \times 75nm$

Device density: 5.6×10^{12} cm^{-2}

Energy per cycle $= 4\cdot10^{-18}\,\frac{J}{cycle}$

Time per cycle

~2 ps

Power: 2μW

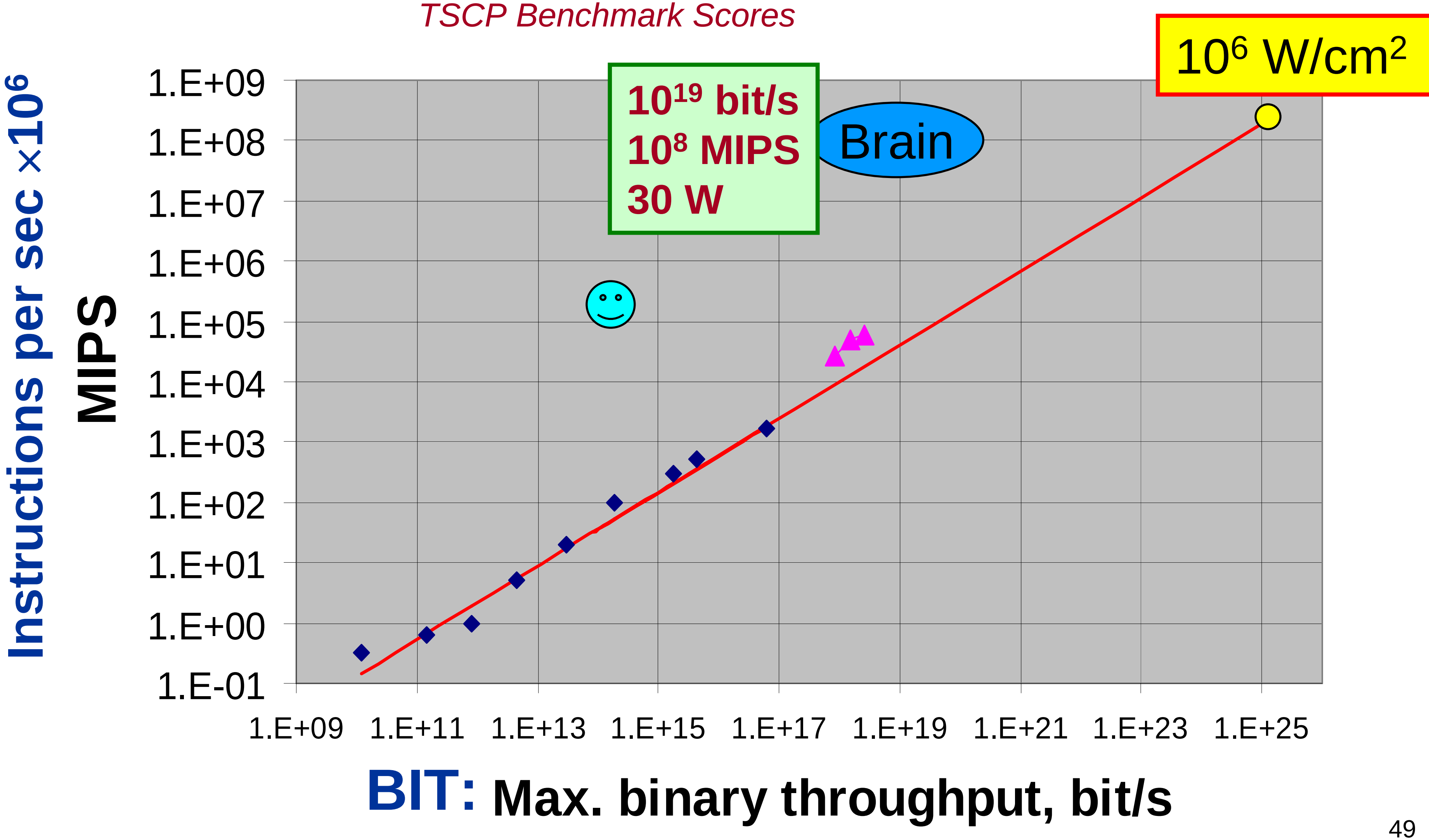
Power density : ~30 kW/cm^2

BITS: 10^{14} bit/s

MIPS: 2×10^5

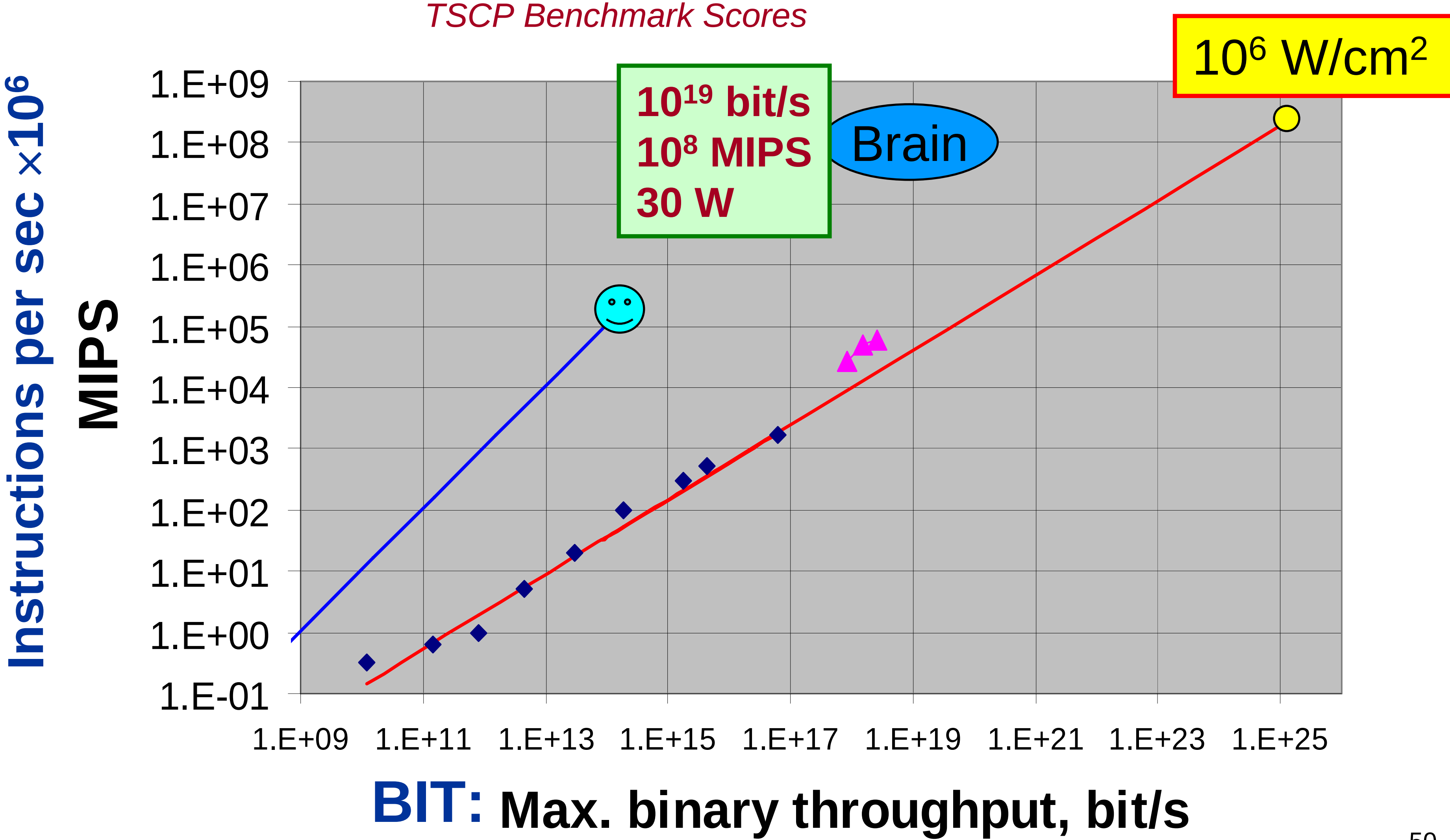
Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



Computing Power: MIPS (μ) vs. BIT (β)

<u>Sources:</u> *The Intel Microprocessor Quick Reference Guide* and
TSCP Benchmark Scores

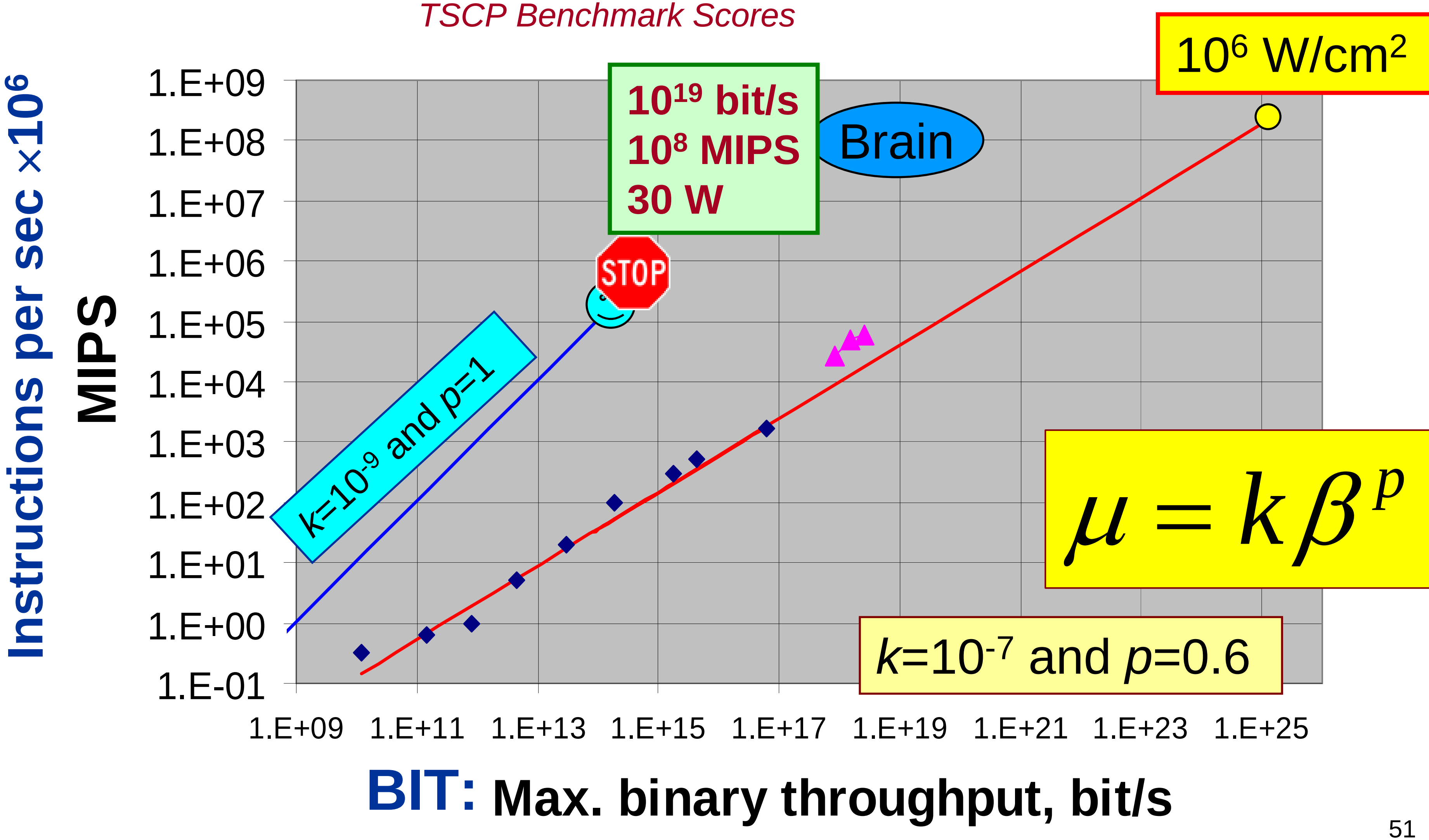


10^{19} bit/s
10^{8} MIPS
30 W

10^{6} W/cm^{2}

Computing Power: MIPS (μ) vs. BIT (β)

Sources: *The Intel Microprocessor Quick Reference Guide* and *TSCP Benchmark Scores*



10^{19} bit/s
10^{8} MIPS
30 W

10^{6} W/cm^{2}

$k=10^{-9}$ and $p=1$

$$\mu = k\beta^{p}$$

$k=10^{-7}$ and $p=0.6$

Summary

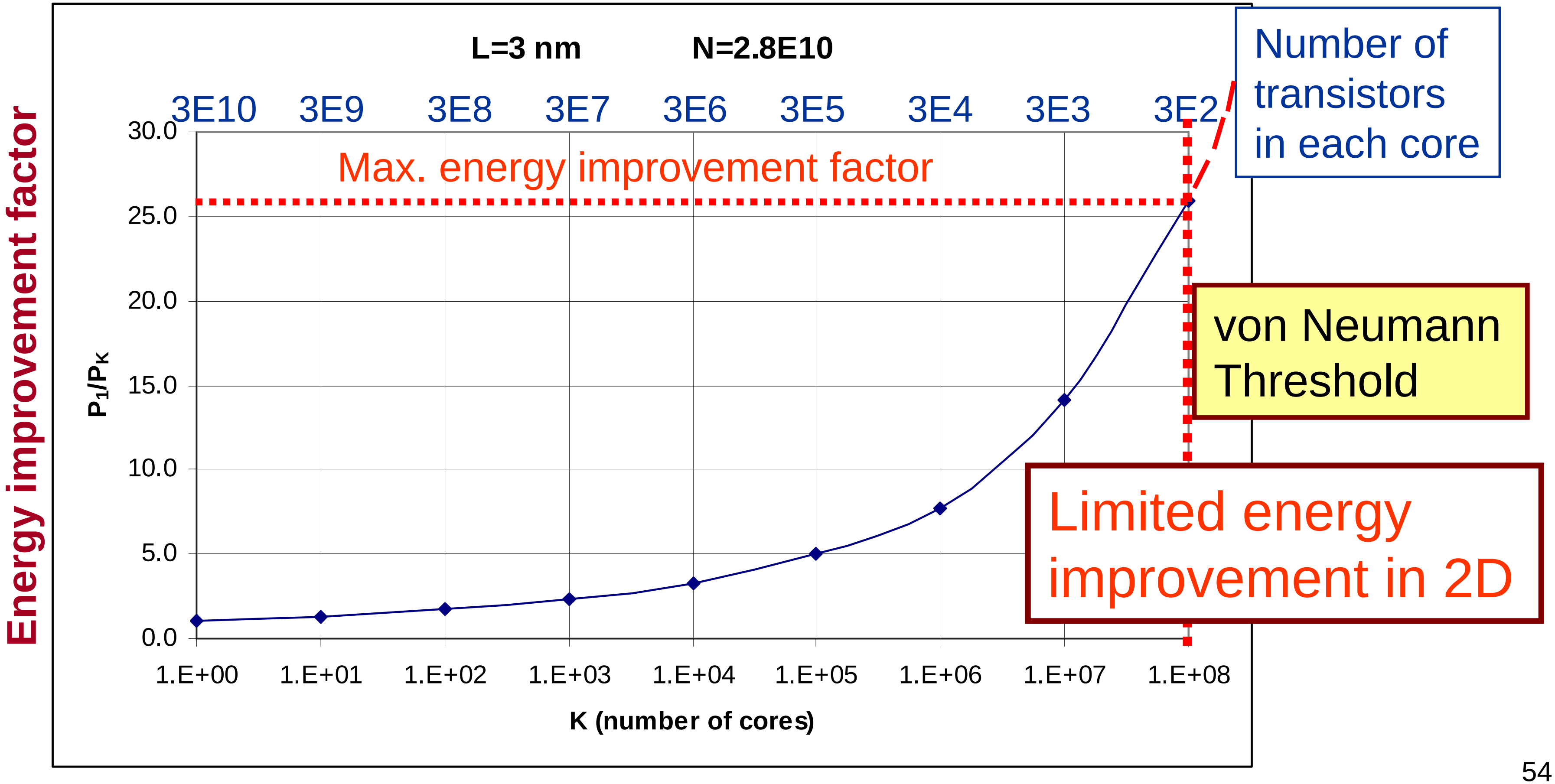
- The Minimal Turing Machine lies on the different performance trajectories from conventional computers
  - It has slope to meet brain performance
- More detailed physics based analysis is needed
  - System thermodynamics of computation
    - **Carnot's equivalent for Computational Engine?**
- Lessons from Biological Computation?
- Candidates for beyond-CMOS nano-electronics should be evaluated in the context of system scaling
  - e.g. spintronic minimal Turing Machine?

Back-up

Extreme Multi-Core Analysis

Power consumption by K cores:

$$P_K = K \cdot \frac{M}{t_{sw}(M)} \cdot E_{b\min}(M) = \frac{N}{t_{sw}(M)} \cdot E_{b_{\min}}(N)$$



Max. energy improvement factor

Number of transistors in each core

von Neumann Threshold

Limited energy improvement in 2D