A Carnot Bound for General Purpose Information Processors?

Sadasivan Shankar (Intel), Ralph K. Cavin III (SRC), Victor V. Zhirnov (SRC)

1st Berkeley Symposium on Energy Efficient Electronic Systems

June 11 & 12, 2009, Berkeley, CA

Main Points

- Energy/Power minimization is a universal macro-constraint for on-chip architectures
  - Performance should not (& *does not have to*) be sacrificed
- Many new directions to leverage scaling
  - New materials, devices, topologies
  - Functional diversification
    - Power sources, capacitors
    - Application-specific processors
- How is maximum computational performance related to device physics?
  - Architecture and software need consideration to enable scaling => *Thermodynamics of Computation at System Level is a more systematic way to leverage scaling*
  - *A new methodology based on statistical physics and quantum mechanics have been developed for addressing thermodynamics of switching-based systems*

Outline

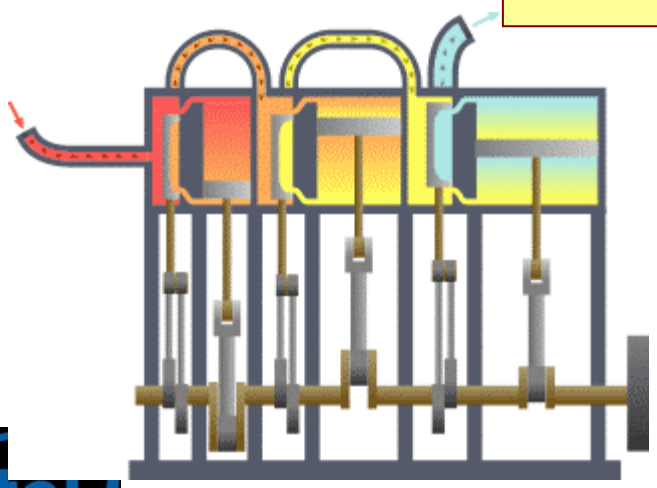
- What ?
  - Basic goals
- Why do it ?
  - Context of power versus MIPs
- How ?
  - Methodology used in the analysis
- Where ?
  - Utility

Fundamental limits for *information engines?*

The discipline of Thermodynamics resulted from the practical need to increase the *EFFICIENCY* of heat engines

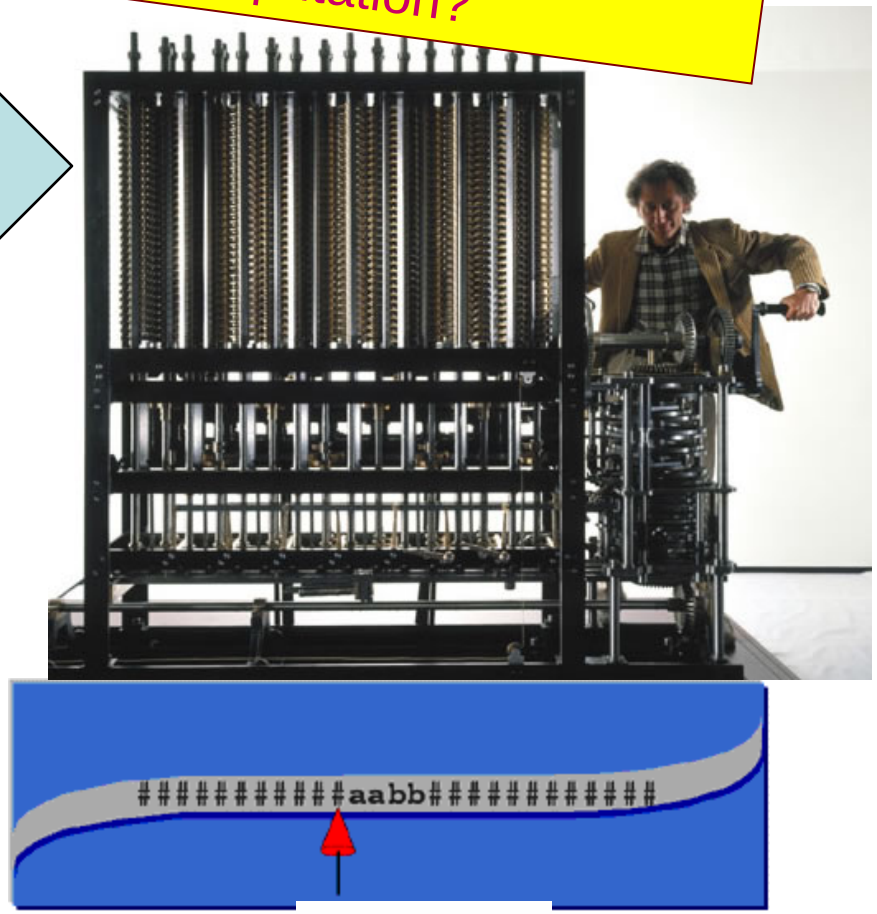
Carnot's formula - the fundamental limits on engine's efficiency

$$\eta = 1 - \frac{T_1}{T_2}$$



Source: Wikipedia

Efficiency boundaries for computation?



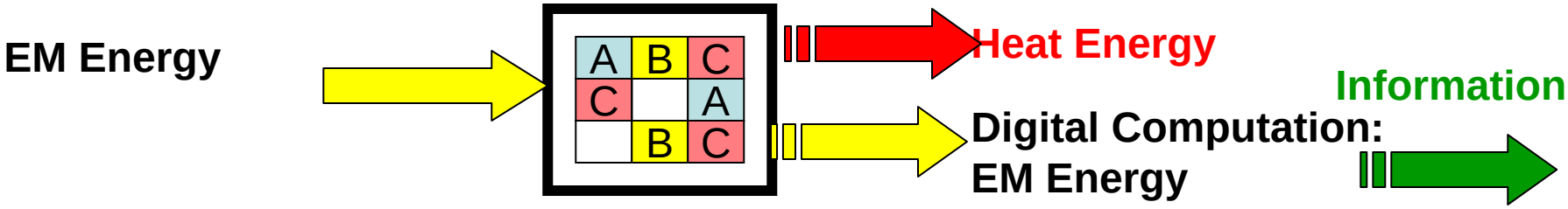
System Reliability Perspectives

- Current approach: System reliability through device reliability
  - All N devices in the logic system operate correctly $E_b \uparrow$
- Requiring all ideal devices may not end with 'ideal' system
  - Locally optimized components may not result in globally optimized system
  - **Global system optimization**: $E_b \downarrow$??

Computing Engine Premise

Heat Engine

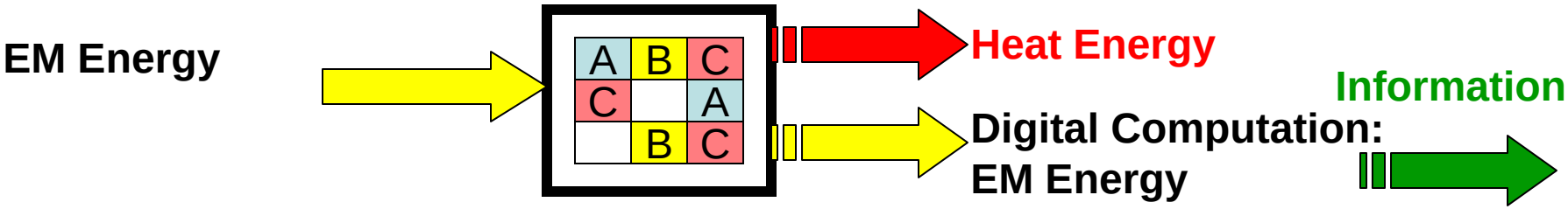
Computing Engine



- Similar to a heat engine, a computing engine can be visualized

Thermodynamics of Computation *@System Level*

- Thermodynamics is the study of energy transformation properties common to all systems
- Goal is to use thermodynamics, which incorporates relations between system's components and determines the most energy efficient systems

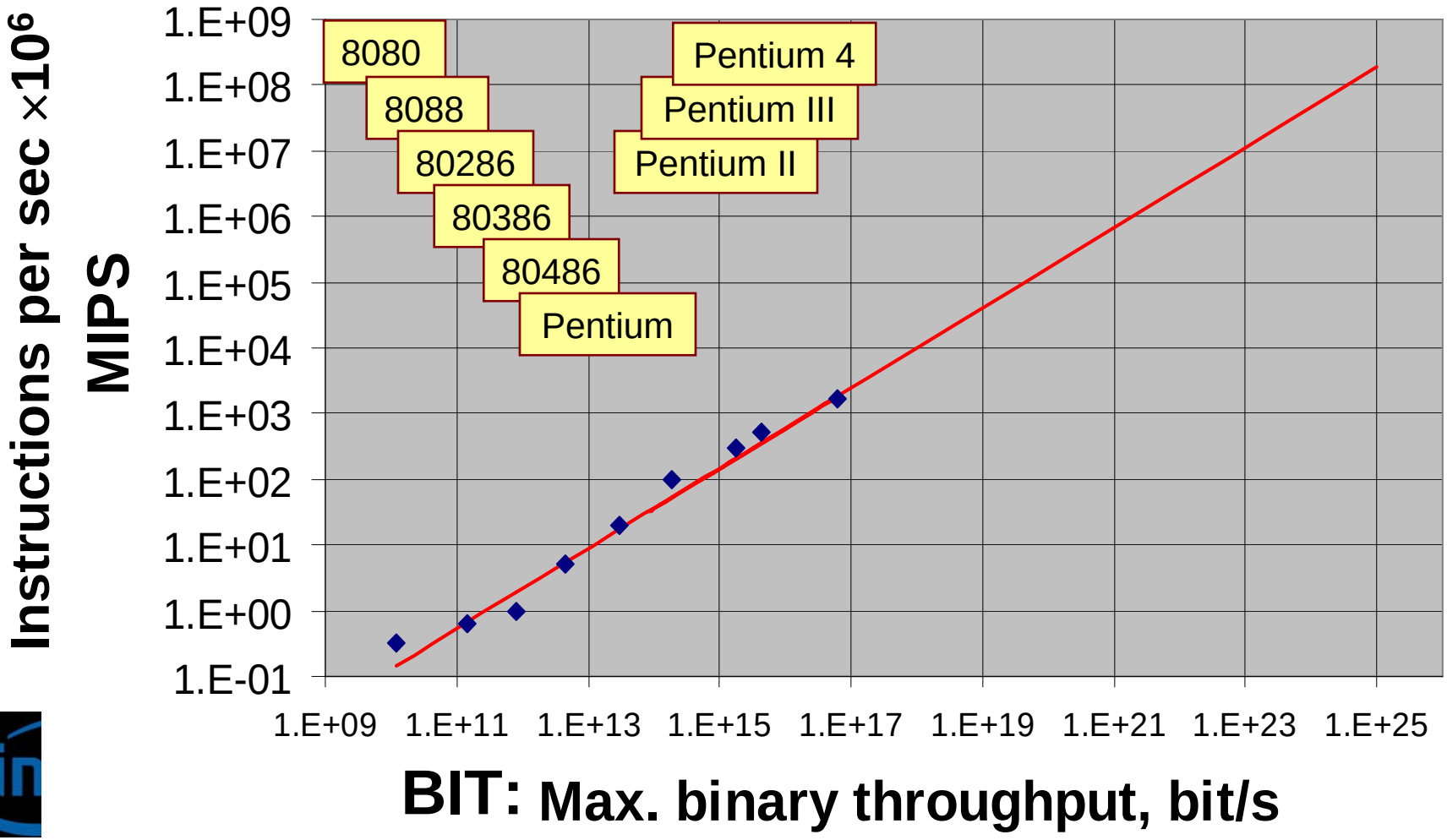


Previous Work

Acknowledgement: V. Zhirnov, R. Cavin

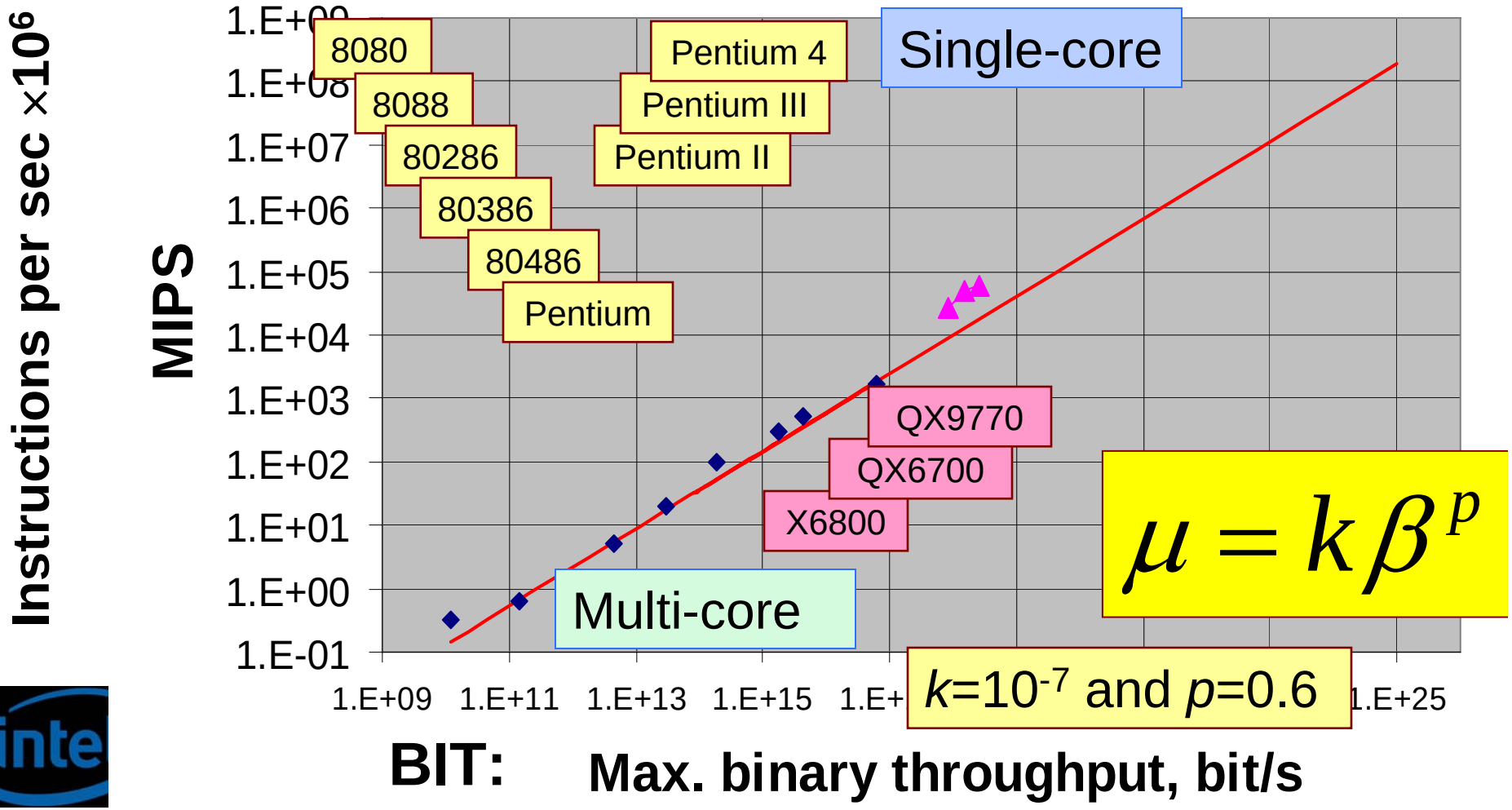
Computing Power: MIPS (μ) vs. BIT (β)

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

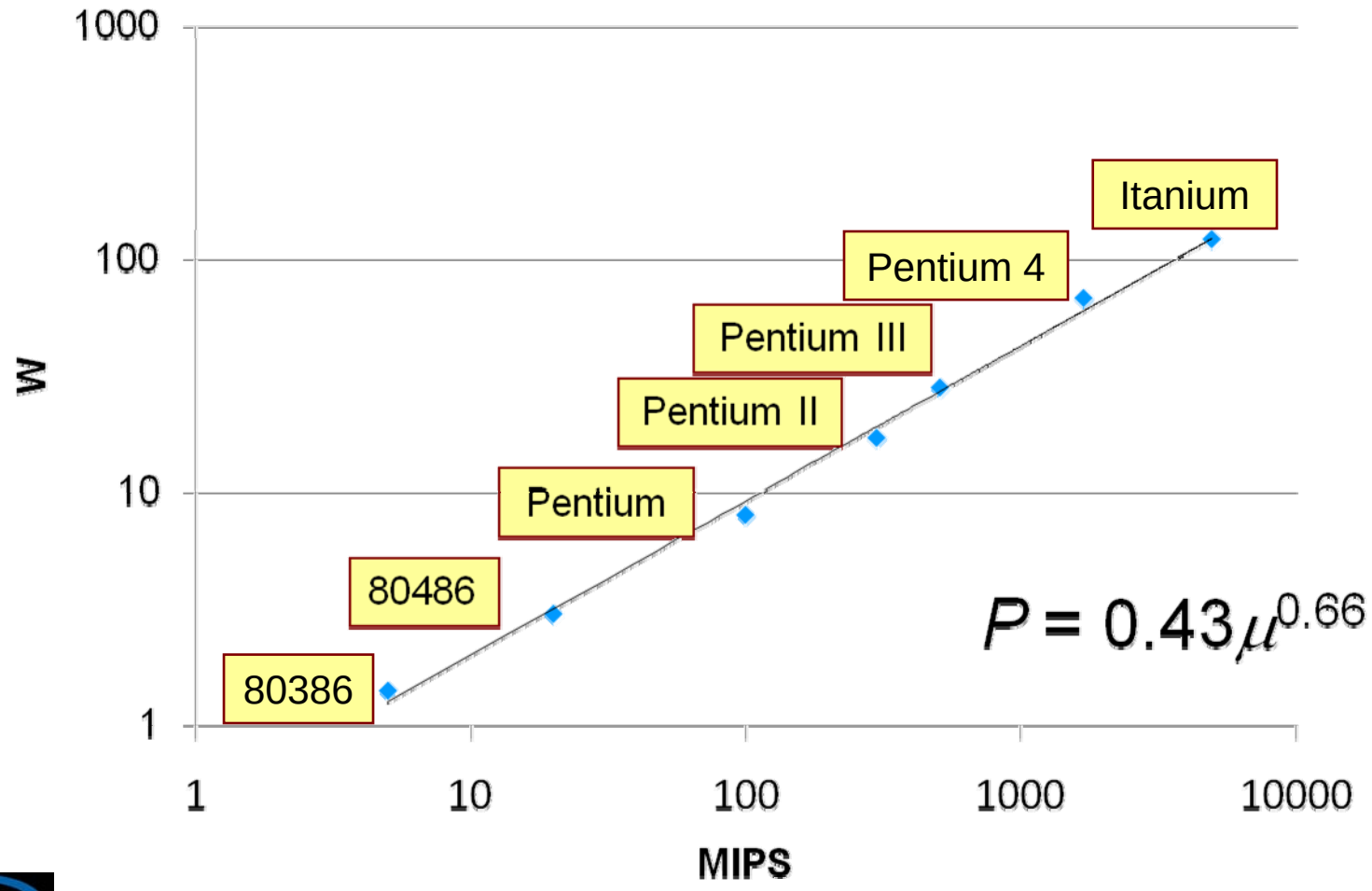


Computing Power: MIPS (μ) vs. BIT (β)

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



MPU Watt- MIPS relations



Observations

1) 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 processor:

How can we increase MIPS?

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

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

2) There also appears to be a functional relationship between electrical power consumption, and the millions of instructions executed per section, μ, executed by a processor:

How can we decrease WATTS?

$$P = k \cdot \mu^{p}$$

k=0.43 and p=0.66

Turing-Heisenberg Rapprochement?

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

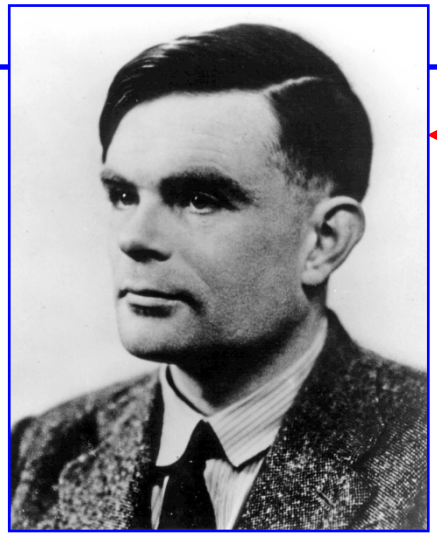
$$\mu = k\beta^{p}$$

$$\beta = \frac{n_{bit}}{t_{sw}}$$

Number of binary elements

Switching time

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!

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

"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

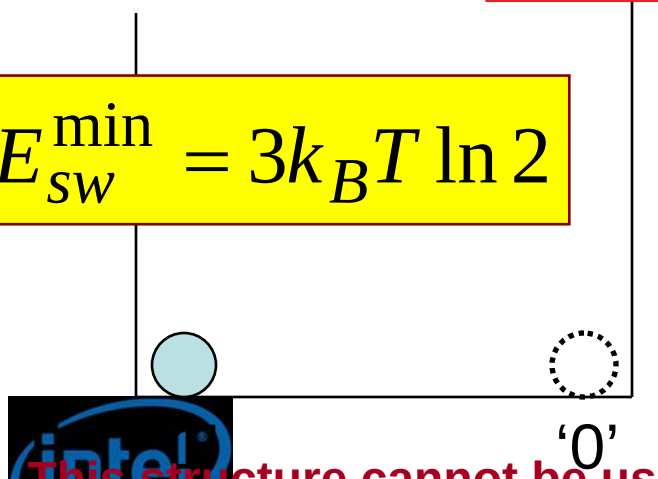
Nanoscale Devices

Π_{error}=0.5

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

~10^{-21} J

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



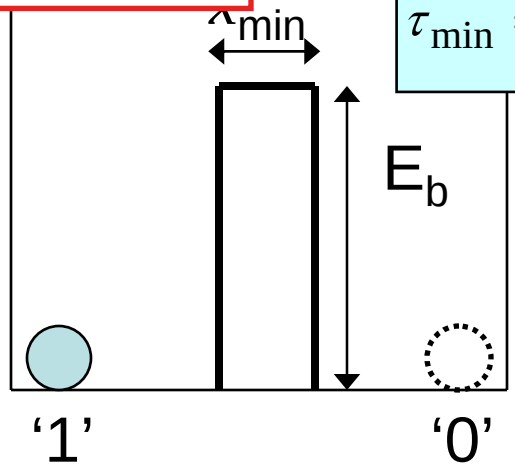
This structure cannot be used for representation/processing information

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

~1.5 nm

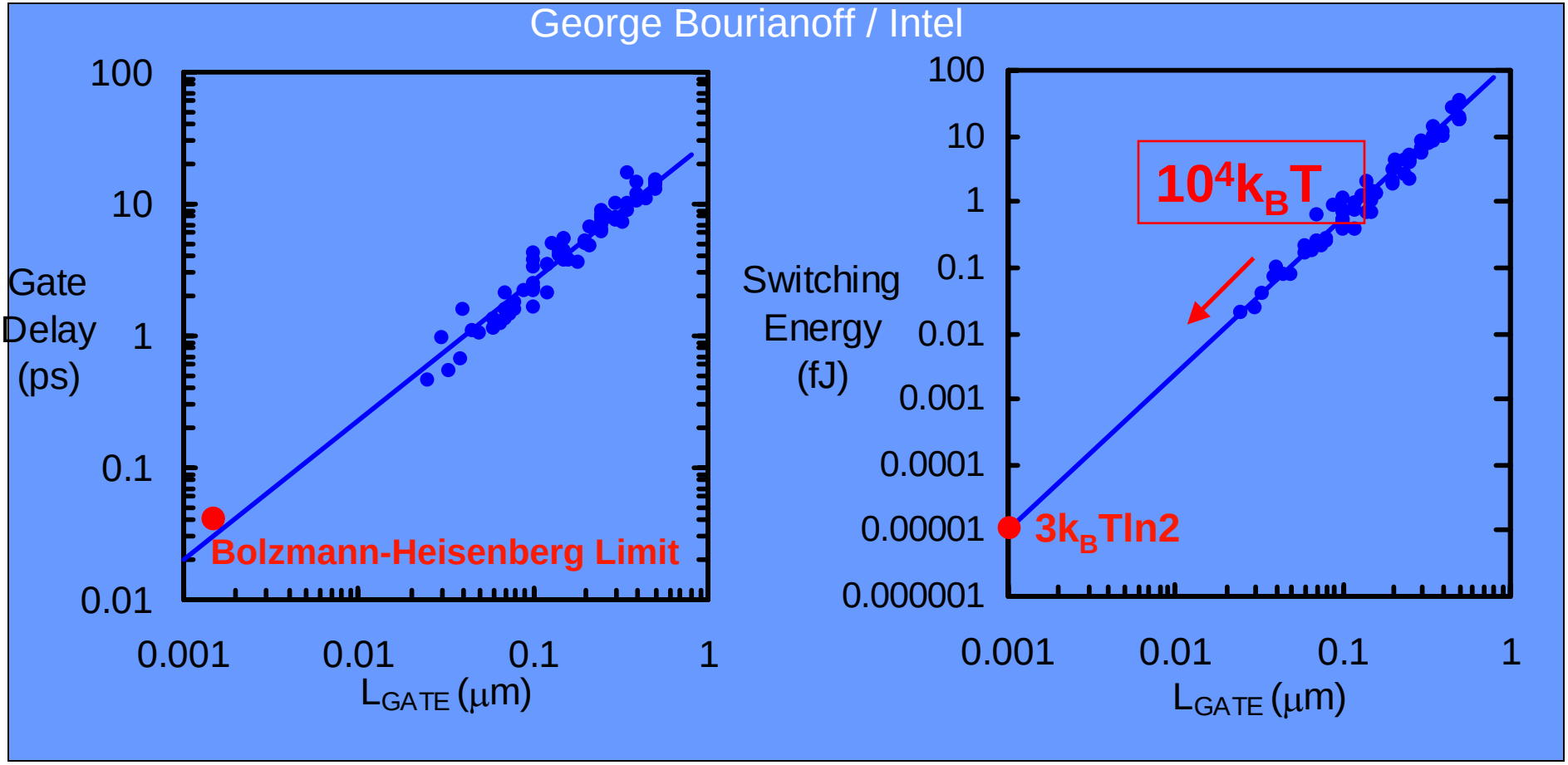
$$\tau_{\min} = \frac{\hbar}{kT \ln 2}$$

~40 fs



An energy barrier is needed to preserve a binary state

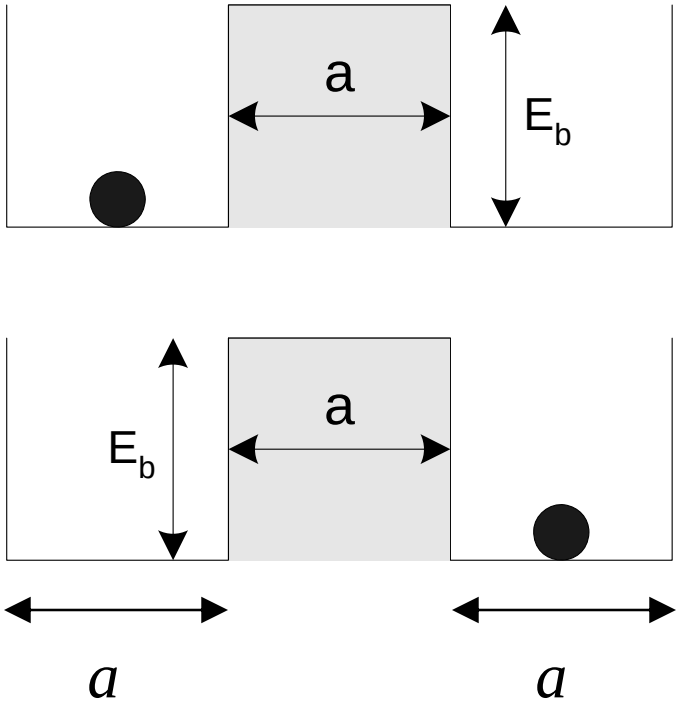
CMOS scaling on track to obtain physical limits for electron devices



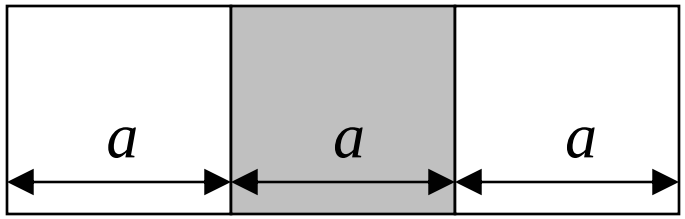
- **Long way to go => challenges ahead; opportunities abound**
- **Question: Why are we at 10,000 $k_B T$?**

Binary switch abstraction:

Generic floorplan and energetics



Generic Floorplan of a binary switch



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

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

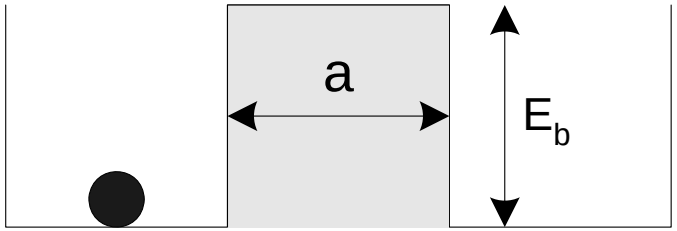
$$E_{sw_{\min}} = 3k_BT$$

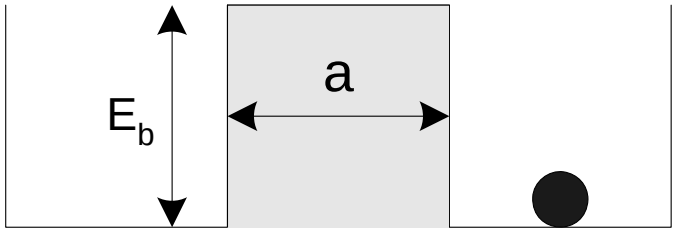
$$\varepsilon = k_BT\left(\frac{J}{tile}\right)$$

Π_{error}=0.5

Two-well bit – Universal Device Model

White spaces are required to provide for isolation and interconnect

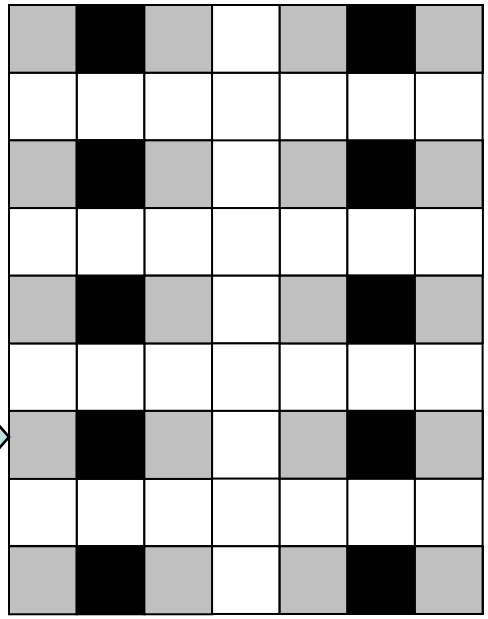




eneric Floorplan
a binary switch

Array

Optimum tiling



Device density

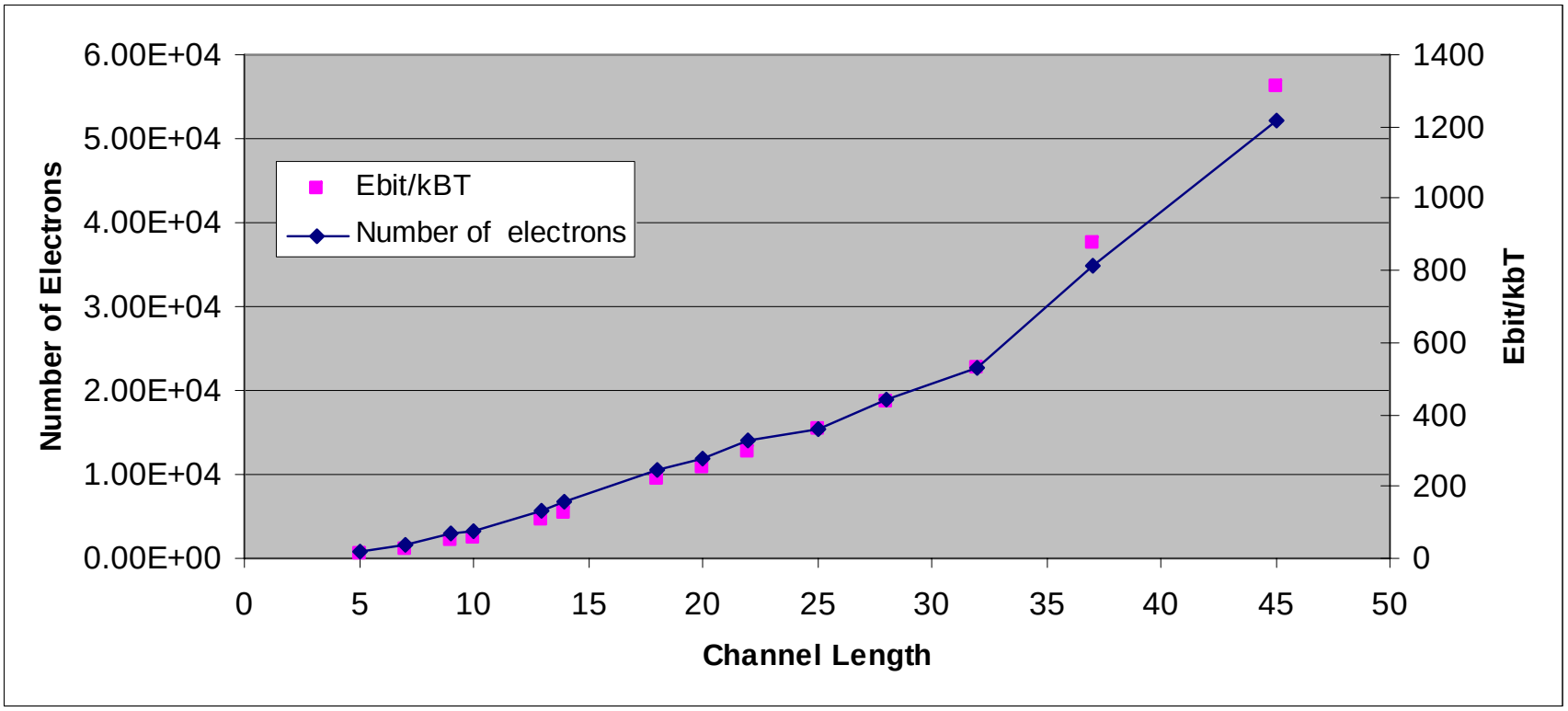
1) Upper Bound

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

2) IC (ITRS)

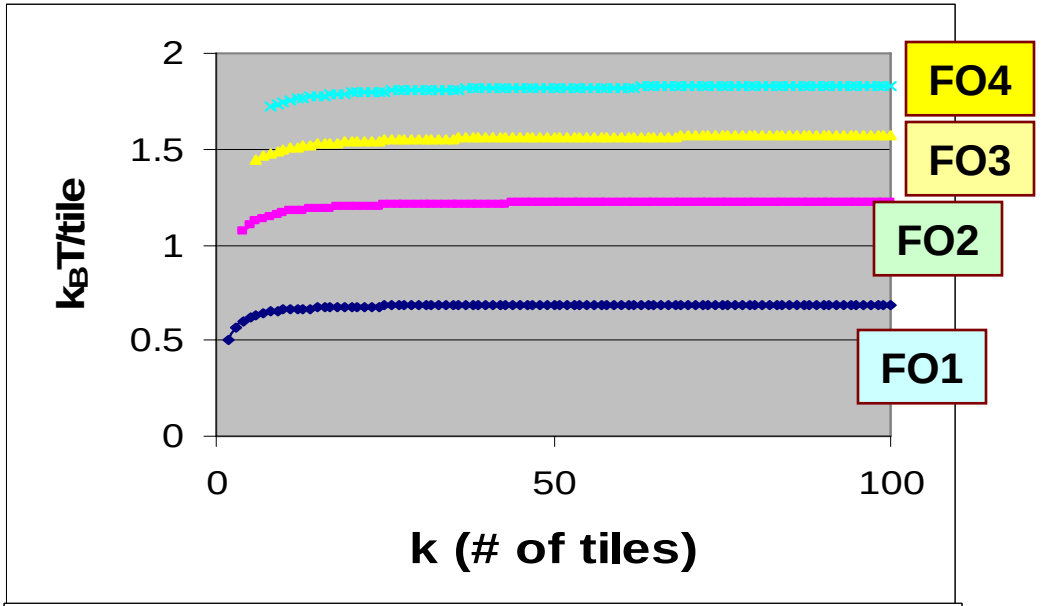
$$n_{MPU} = \frac{1}{(20a)^2}$$

More electrons means more energy



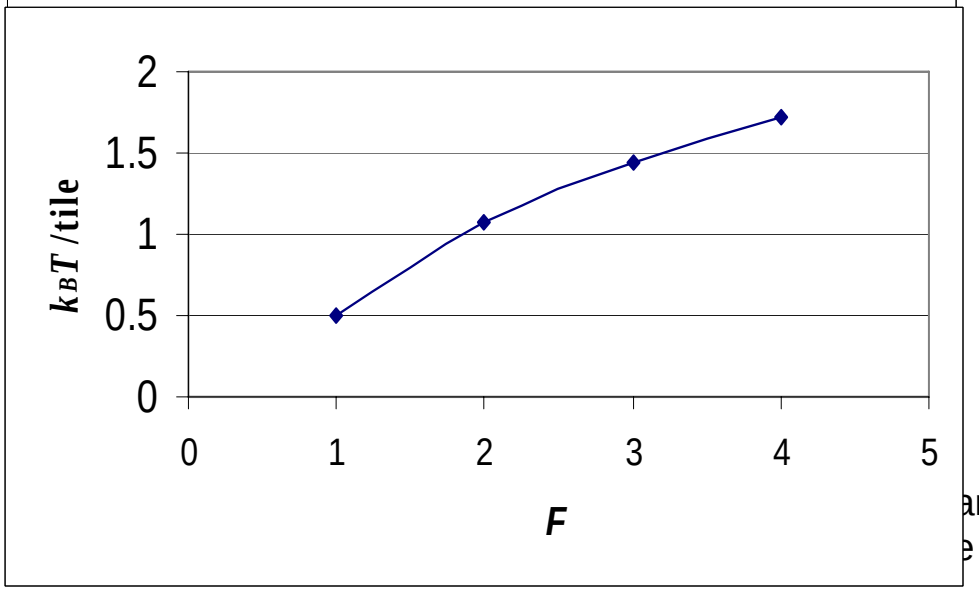
- We need a significant number of electrons for branched communication between binary switches

In the limits: Energy per interconnect tile



Long interconnect limit

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



Minimum interconnect limit

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

Π=0.5

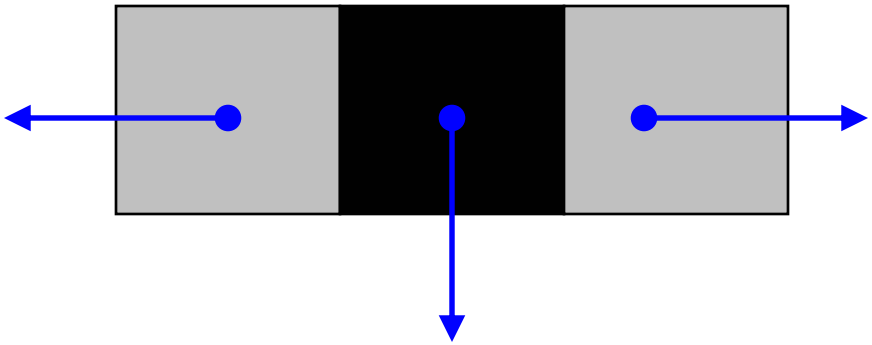
$$\varepsilon \sim k_B T/\text{tile}$$

Floor space Expenses of Communication between Binary Switches

Assumption: For each of 3 tiles of Binary Switch we need at least:

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

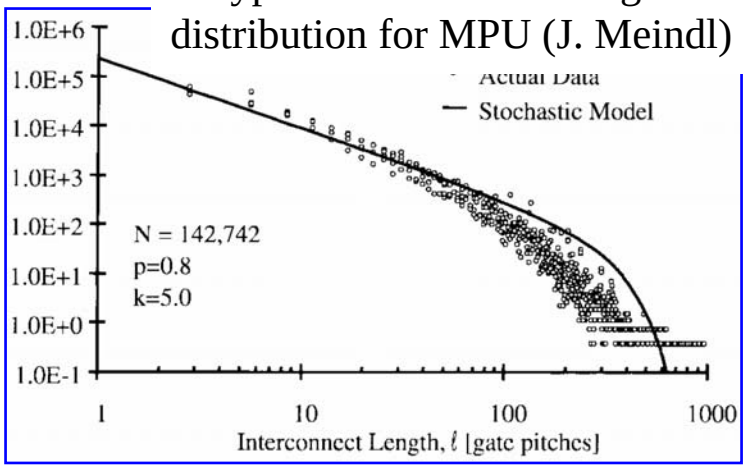
Total **6 interconnect** tiles per binary switch



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

Reality check:

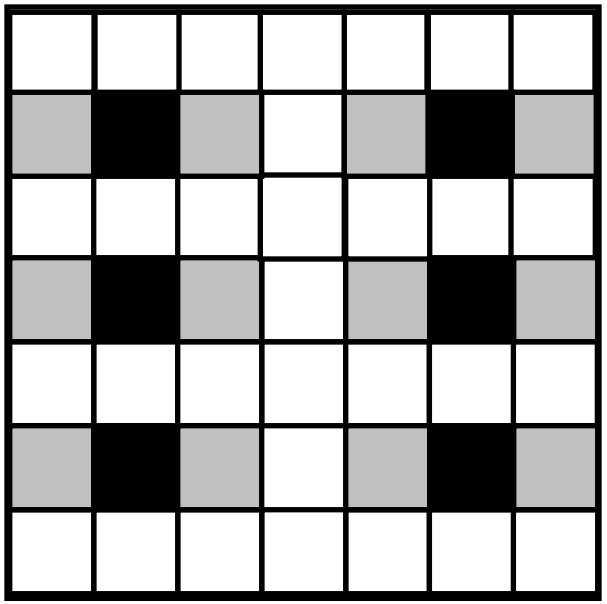
A typical interconnect length distribution for MPU (J. Meindl)



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 floor plan, energetics and speed

Switching energy of one binary switch in a circuit



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: τ_{min}/tile

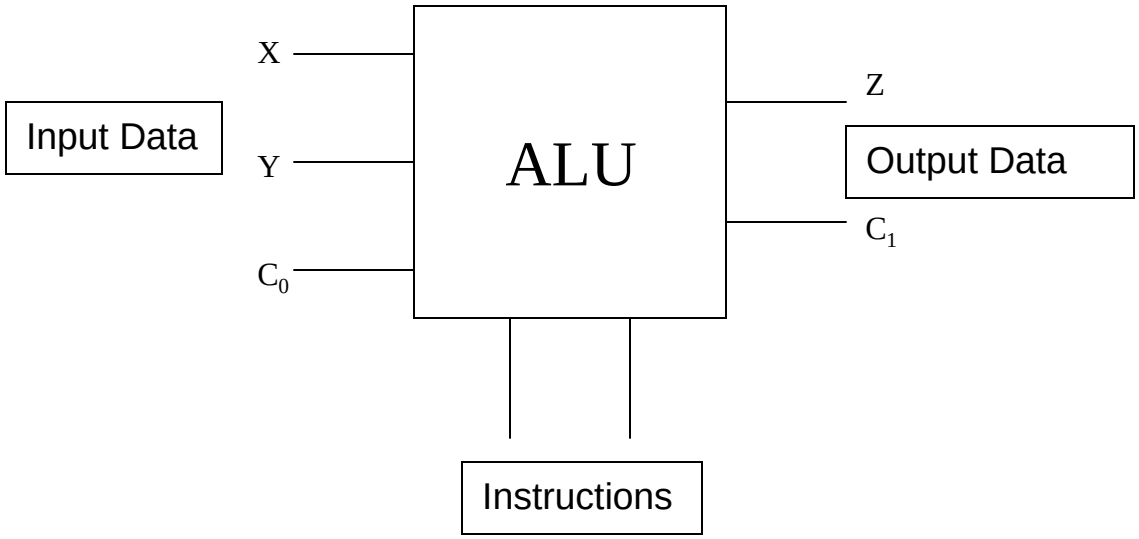
$$\tau_{min}=\frac{\hbar}{kT\ln 2}$$

~40 fs

S. Shankar
une 2009

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

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**))

Jan Rabaey,
Digital Integrated Circuits
2

Minimal ALU abstraction: *Energetics*

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

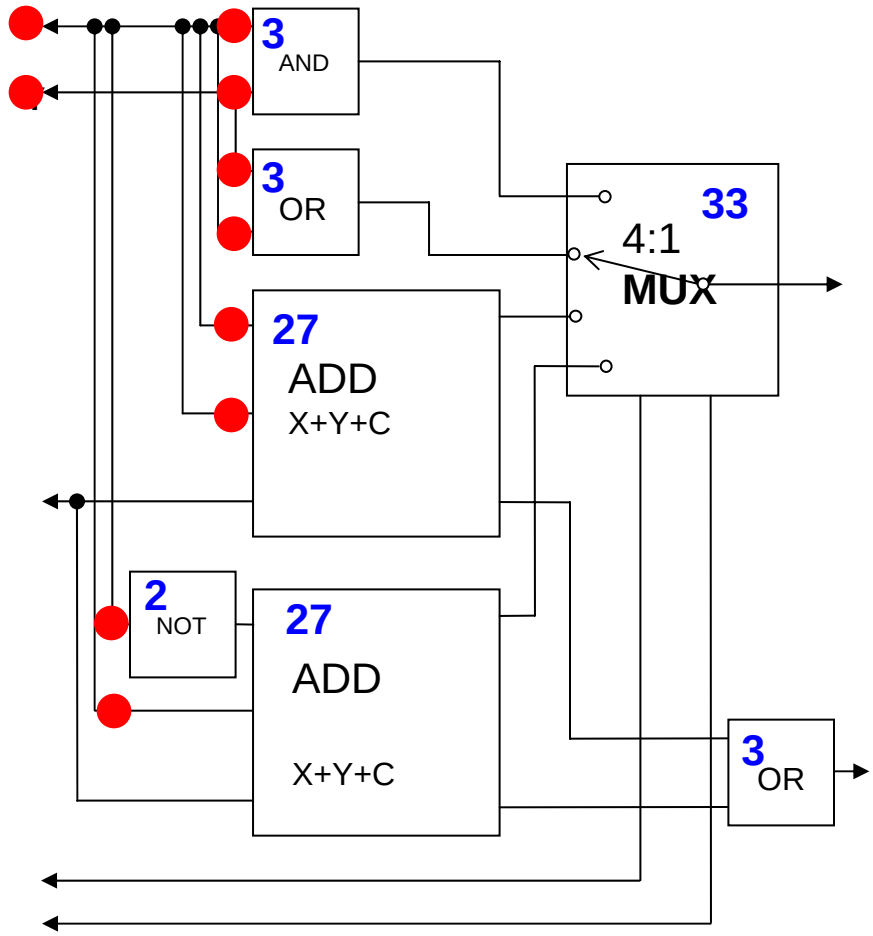
Energy efficiency: $\eta = \dfrac{E_{op}}{E_{ALU}}$

$$E_{AND} = \frac{9}{2} \cdot 3 \cdot k_B T \ln 2 \sim 10 k_B T$$

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

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

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



All 4 units execute even though only one output is used

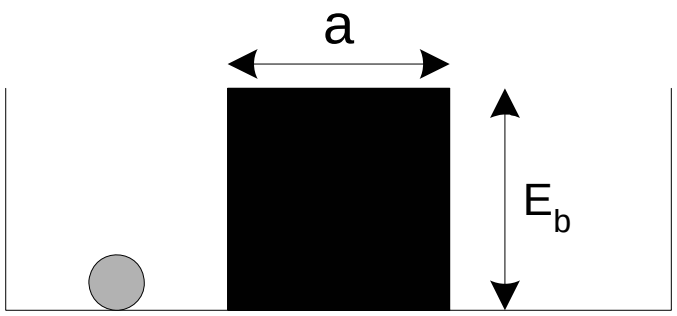
Total: 98 devices

23

Current Work: System Layout

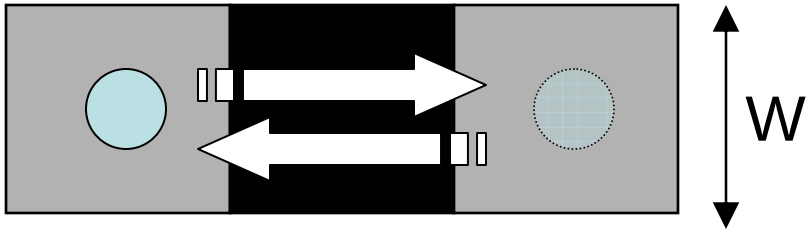
Expression of *Computing System*
in a *Geometrical Representation*

Binary Switch - Basics



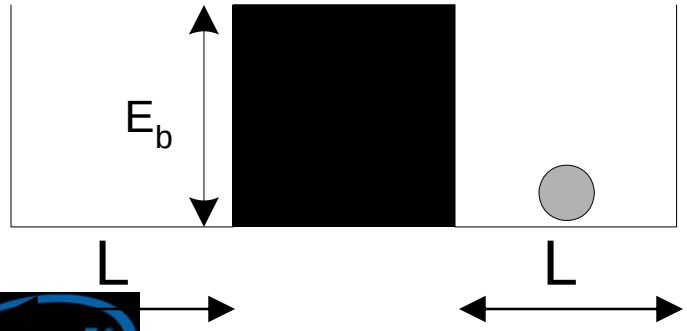
Key Characteristics:

1. ***Confinement (Energy)***
2. ***Barrier (Energy)***
3. ***Information carrier (Charge)***



Geometrical Parameters:

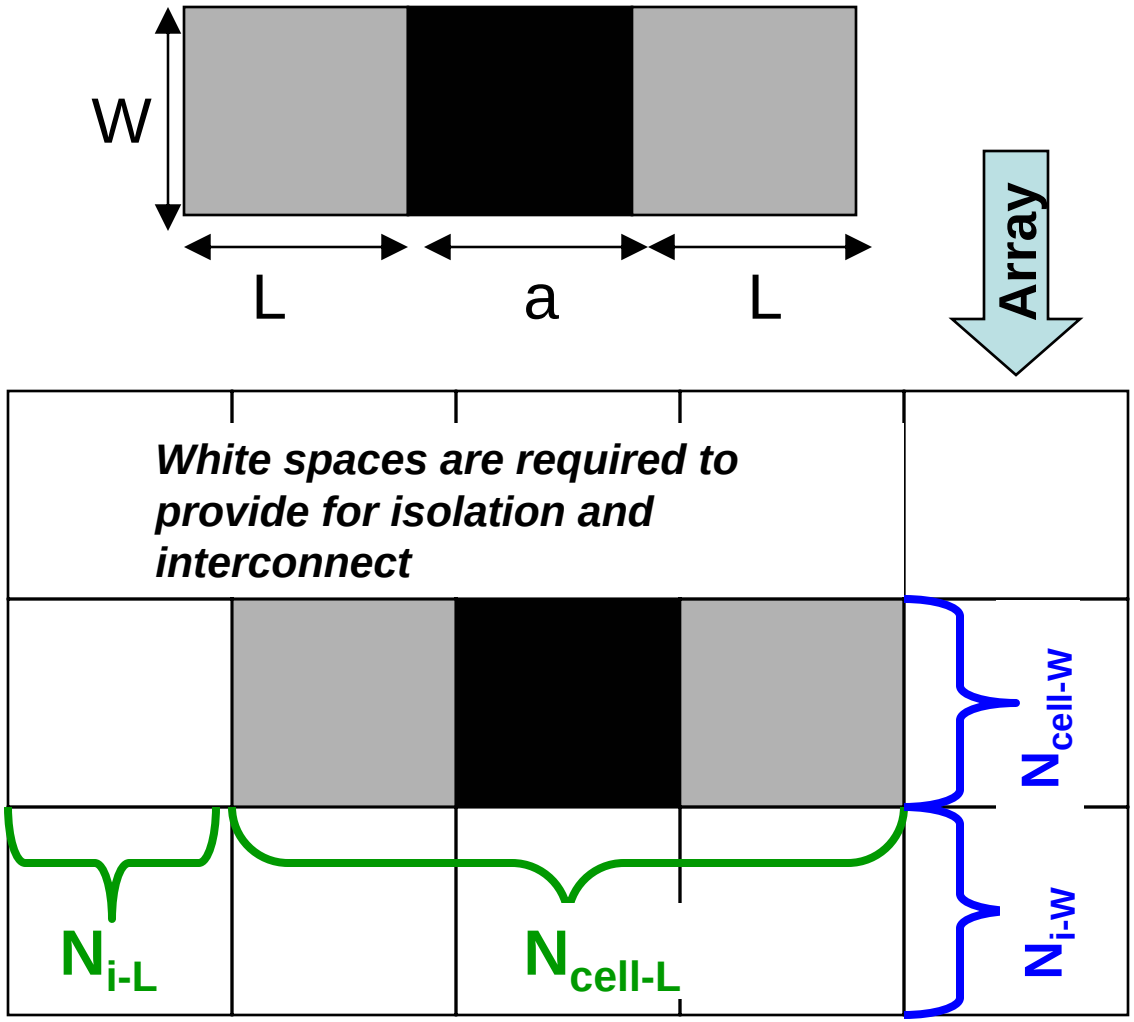
1. ***Confinement Width (W) & Length (L)***
2. ***Barrier Length (a)***
3. ***Information carrier (Charge)***



System Parameters:

1. ***Barrier Energy (Eb)***
2. ***Temperature T***
3. ***Charge (e)***

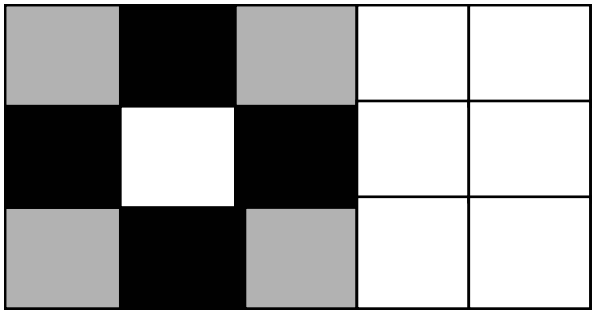
Binary Switch – Floor Plan



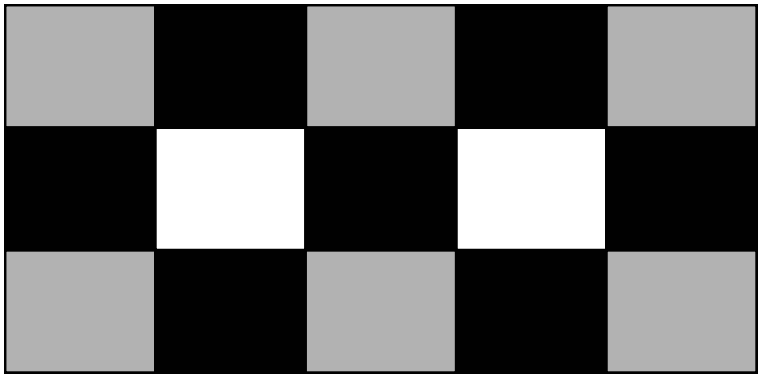
Floor Planning Examples

1 Switch

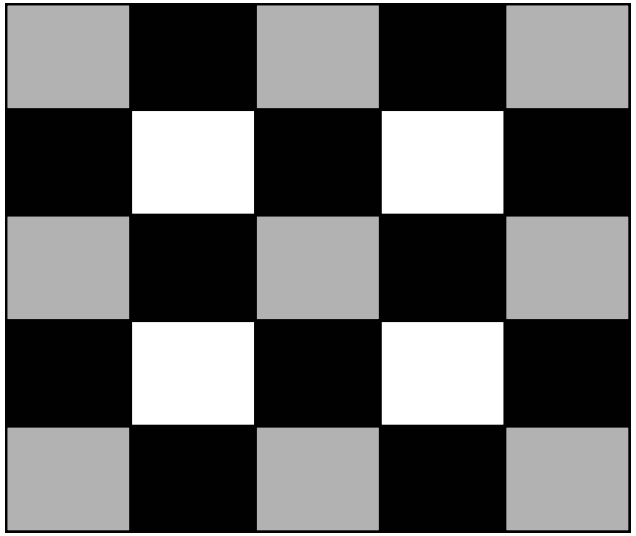
2 Switches



4 Switches



7 Switches



12 Switches

Floor Planning Examples

Density

Examples	N_{cell-L}	N_{i-L}	N_{cell-W}	N_{i-W}	α
1-S	3	1	1	1	8
2-S	5	1	1	1	12
4-S	3	2	3	1	20
7-S	5	1	3	1	24
12-S	5	1	5	1	36

$$n = \frac{1}{\alpha a^2}$$

Upper Bound

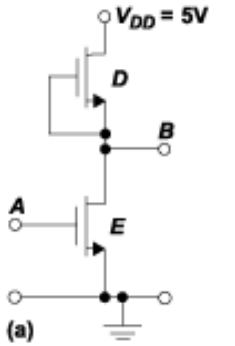
$$n_{\max} = \frac{1}{8a^2}$$

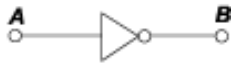
$$n = \frac{1}{(N_{cell-w} + N_{i-w})(N_{cell-L} + N_{i-L})a^2}$$

IC (ITRS)

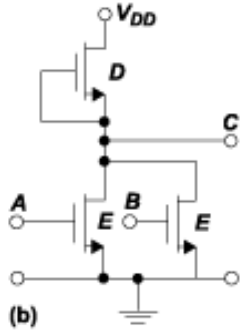
$$n_{MPU} = \frac{1}{(20a)^2}$$

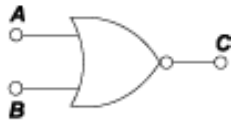
Logical Switches - Illustration



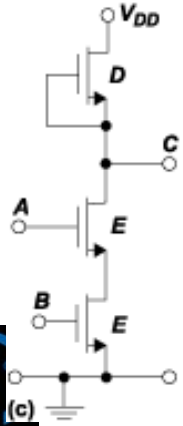


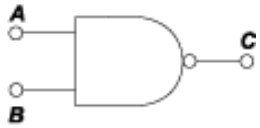
Inverter - 2 switches



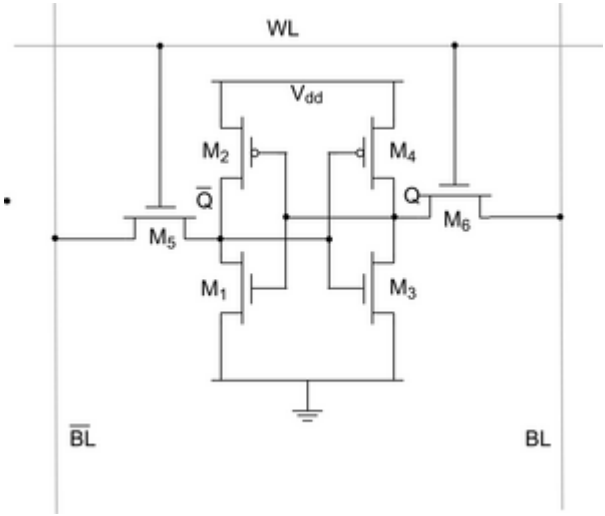


NOR - 3 switches





NAND - 3 switches



SRAM - 6 switches

Floor Planning Examples - Density

Examples	N_{cell-L}	N_{i-L}	N_{cell-W}	N_{i-W}	α
Inverter	5	1	1	1	12
NOR	5	1	3	1	24
NAND	7	1	1	1	16
6-T SRAM	7	1	5	1	48

- Reflects the principle of floor planning
- Actual Layout may have other considerations

$$n = \frac{1}{\alpha a^2}$$

Upper Bound:
Inverter

$$n_{max} = \frac{1}{12a^2}$$

Lower Bound:
6-T SRAM

$$n_{MPU} = \frac{1}{48a^2}$$

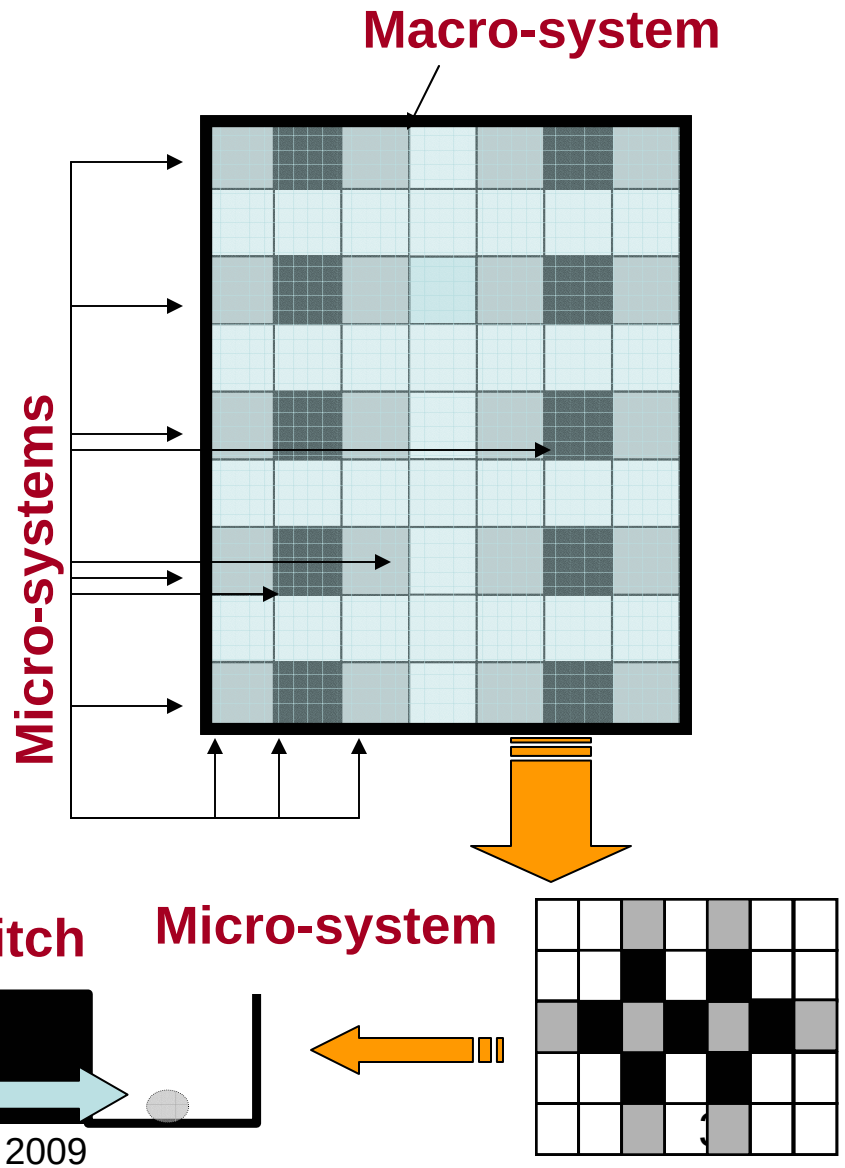
Current Work: System Thermodynamics

Mapping the *Geometrical Representation* to a *Thermodynamic System*

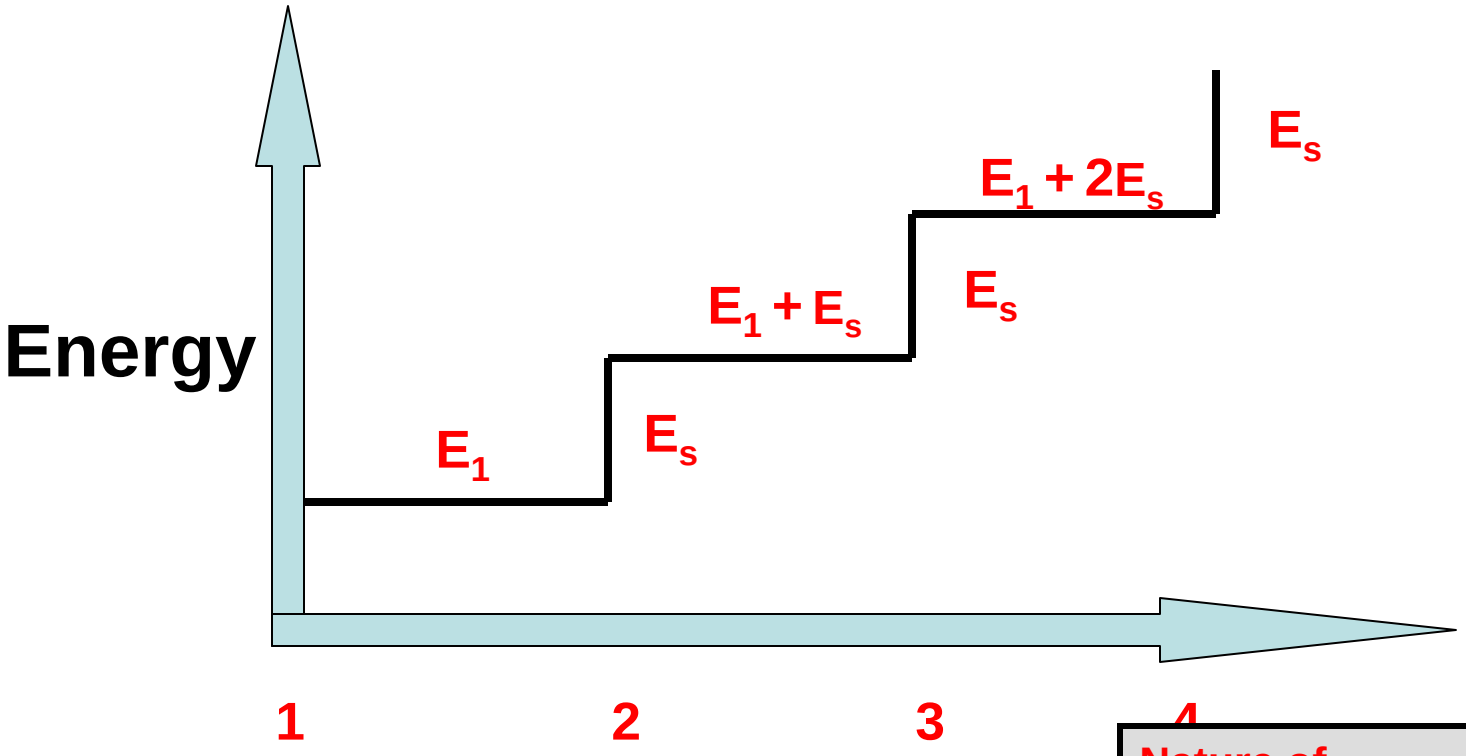
Thermodynamics of Computing

Basic Premise

- Switches are operating at quantum limits, while the macro-system is thermodynamic
- Macro-system is in thermal equilibrium with surroundings (*canonical ensemble*)
- Micro-systems (multi-switch systems; NAND etc..) are sub-domains which can be thermodynamically represented by average energy
- Probability determined by operating “NITS” (NIT =2 is bit)

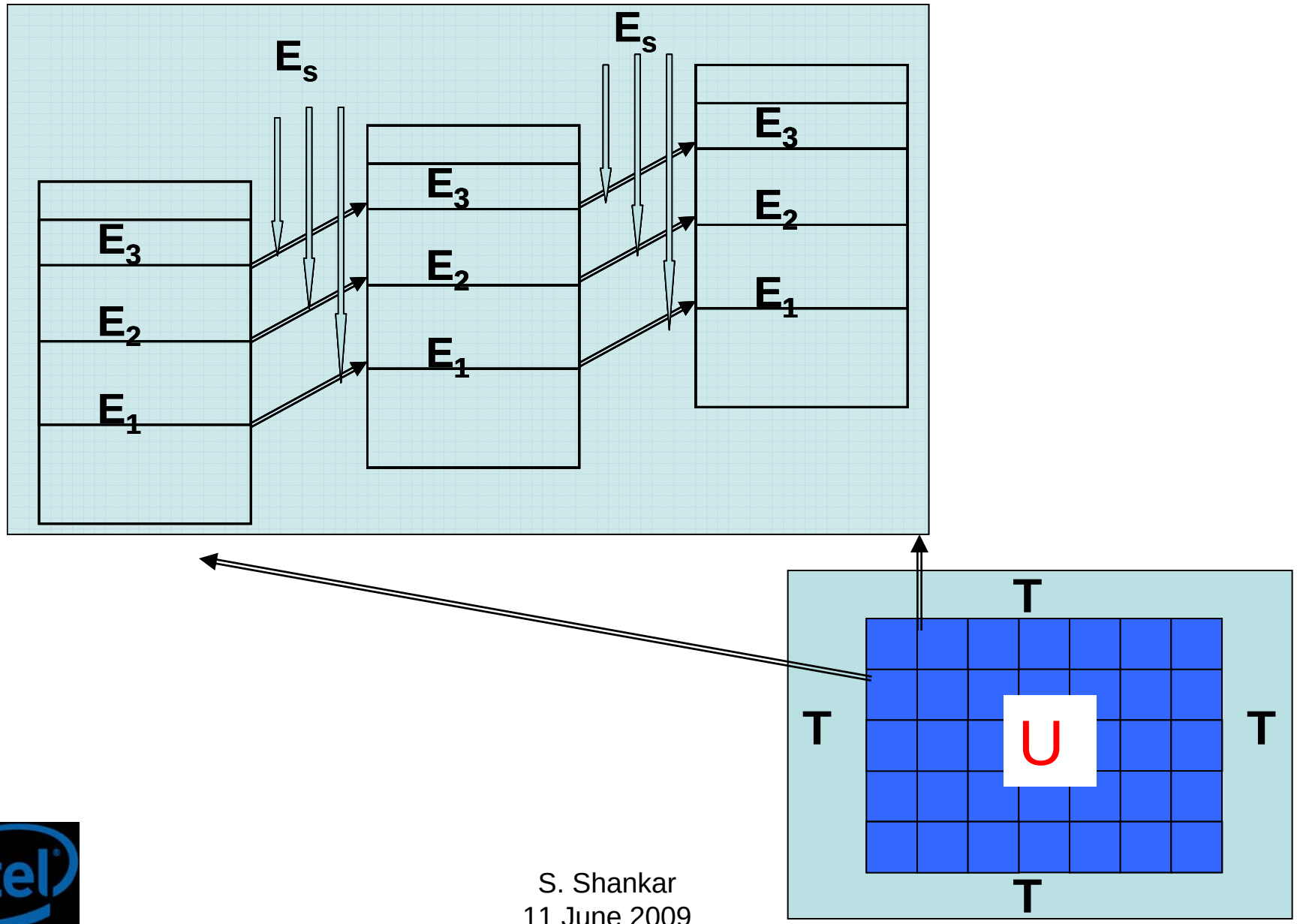


NIT Switching Energy



Nature of Switching	NIT
Binary	2
Ternary	3
Quarternary	4

Ensemble – Micro System

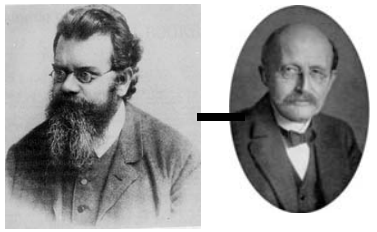


S. Shankar
11 June 2009

Basic Equations

System Entropy

$$S = -k_B \sum_i^N P_i Log P_i$$



Equation

Probability

$$P_i = \frac{e^{-E_i/k_B T}}{Z}$$

$$Z = \sum^N g_i \, e^{-E_i/k_B T}$$

Energy of a quantum device

$$E_i = \frac{2(n_x^2 + n_y^2 + n_z^2)\left(h^2/4\right)}{2m_{electron} a^2}$$

D - ***Linear Dimension of the System***

g_i - ***Statistical weight of each of the micro-system***

E_i – Energy is estimated from quantum mechanics

N – energy states

Z – Partition function

System Free energy

$$A = E - TS$$

Entropy of a Single Switch System

- Entropy is determined by statistical mechanics

$$S = \frac{E_i}{T} - \frac{N}{T} E_s (Nit - 1) + k_B N \ln Z + k_B \ln Nit$$

Total energy

Switching Energy

Quantum States

Switching state

Nit – ***Nit = 2 is bit, Nit = 4 is qit etc….***

k_B - Boltzmann constant

N – number of states in the system (1 for single state switching)

Et – Total Energy is estimated from statistical mechanics

Z – Partition function

Simple Illustration

- For a binary switch, the minimum energy is determined by the need to maintain binary transition (bit) and energy of the particle in an isolated level
- For a classical switch, the following are the limits

Example Micro-Systems	Bits	N	P_i	E_{Min}
Binary	1	$D^2/8a^2$	$1/2^N$	kT log 2
Inverter	1	$D^2/12a^2$	$1/2^N$	kT log 2
NOR	2	$D^2/24a^2$	$1/4^N$	2kT log 2
NAND	2	$D^2/16a^2$	$1/4^N$	2kT log 2
6-T SRAM	16	$D^2/48a^2$	$1/2^{16N}$	16kT log 2

- E_{Min} is idealistic and is determined by the bits processed in the micro-systems

Summary

- We have developed a general methodology for applying thermodynamic principles for information engines like the Carnot principle to heat engines
  - More fundamental than simplistic capacitance based formalism currently being used
- Two applications
  - Similar to heat engines, will identify the ideal Compute Engine – Carnot's Compute Engine for ideal computing. This would serve as a limiting case for realistic architectures
  - Evaluate theoretical efficiencies for different architectures based on physics
- Lessons from Biological Computation
  - the brain appears to operate at a device switching energy of a few hundred kT
- Continue work on estimating minimum energy needed of various simple systems
  - Include realistic heat terms (dissipation)
- Estimate available energy