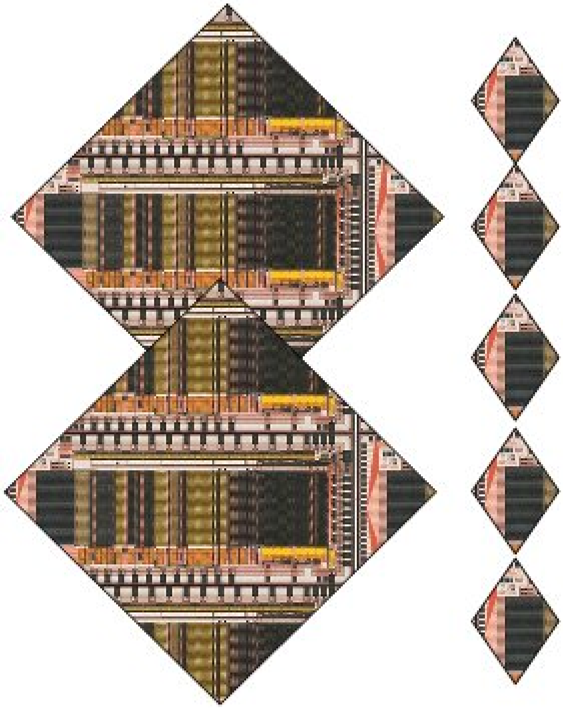
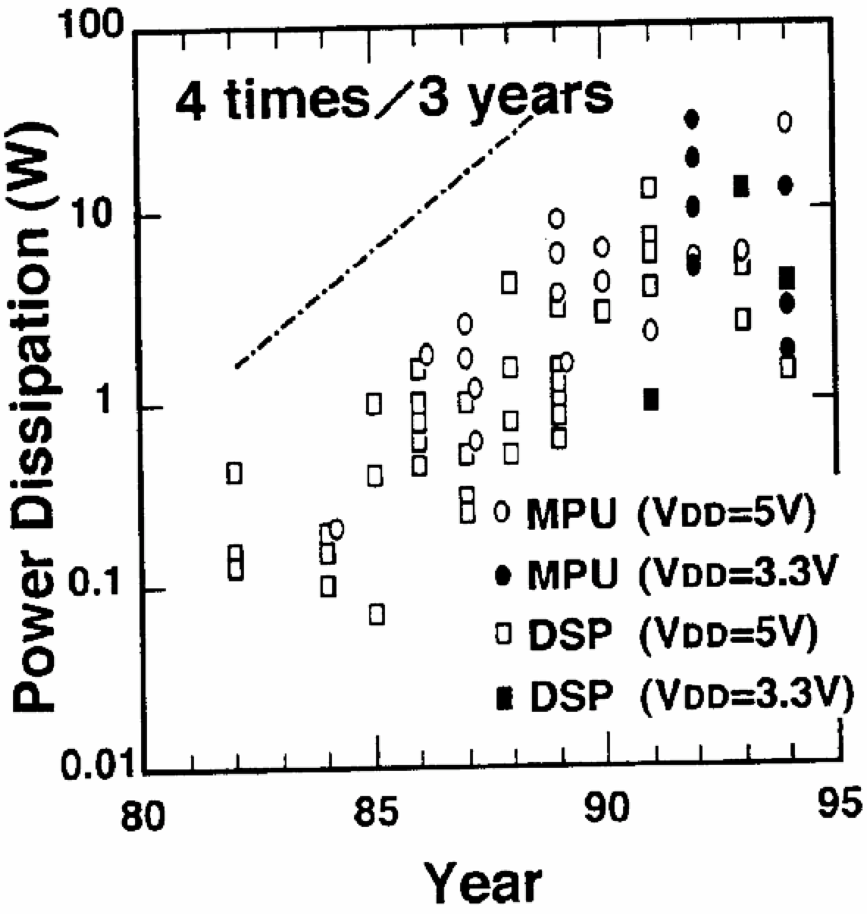
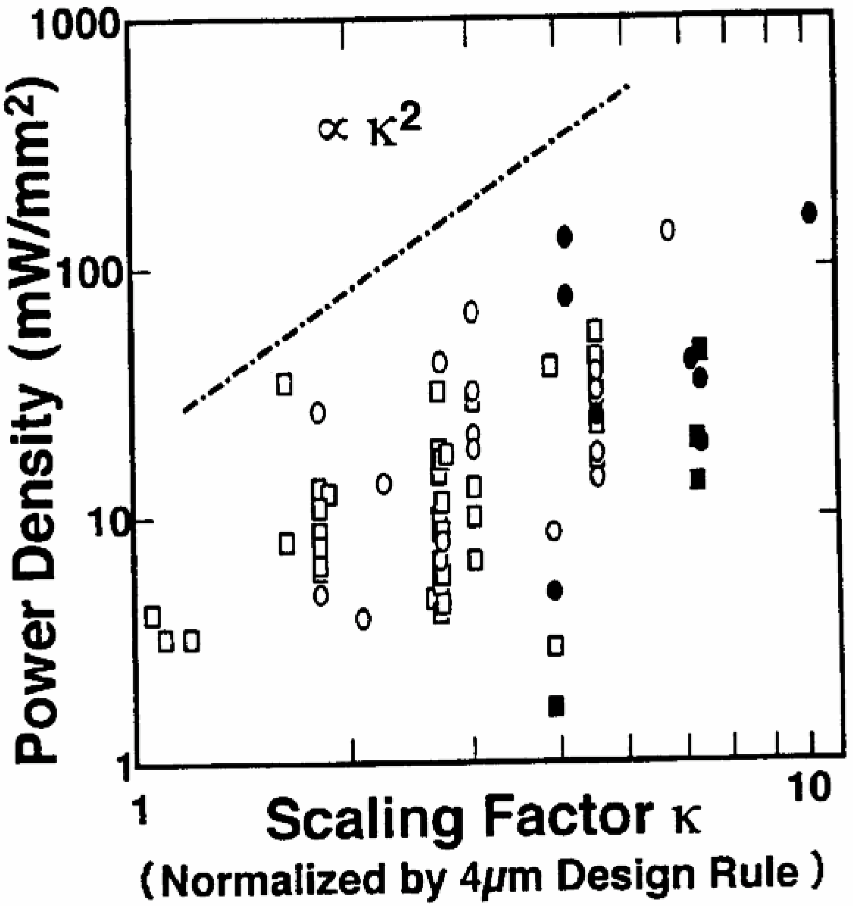


Low Power Design in VLSI

Evolution in Power Dissipation:

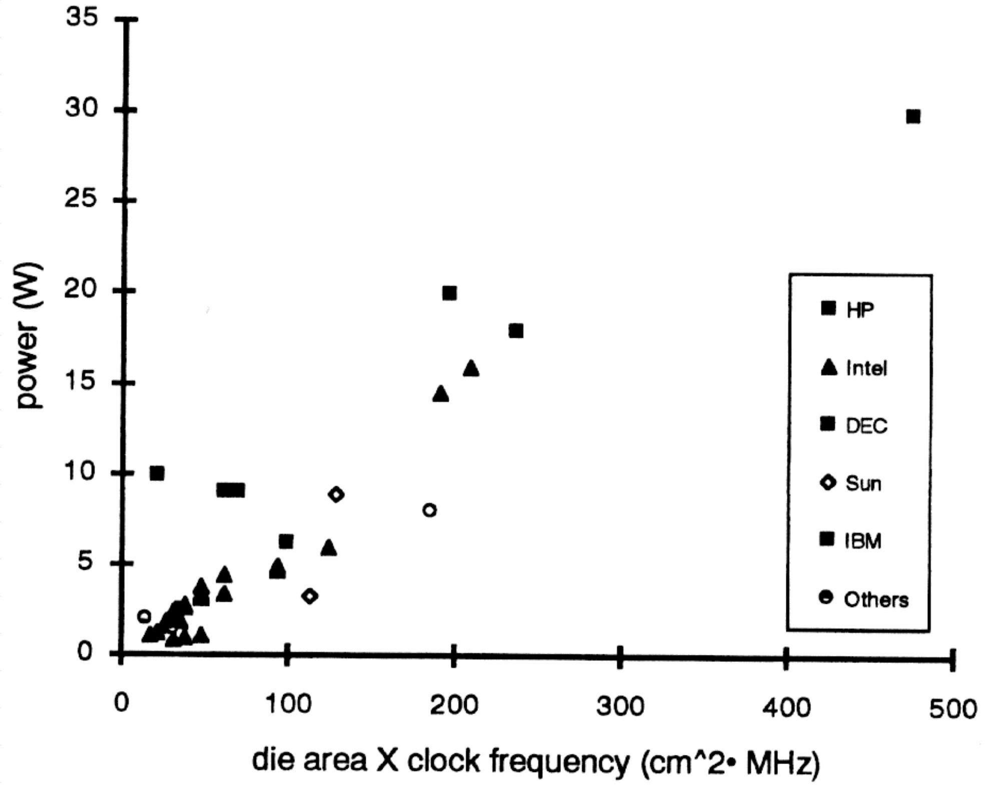




Why worry about power?
Heat Dissipation

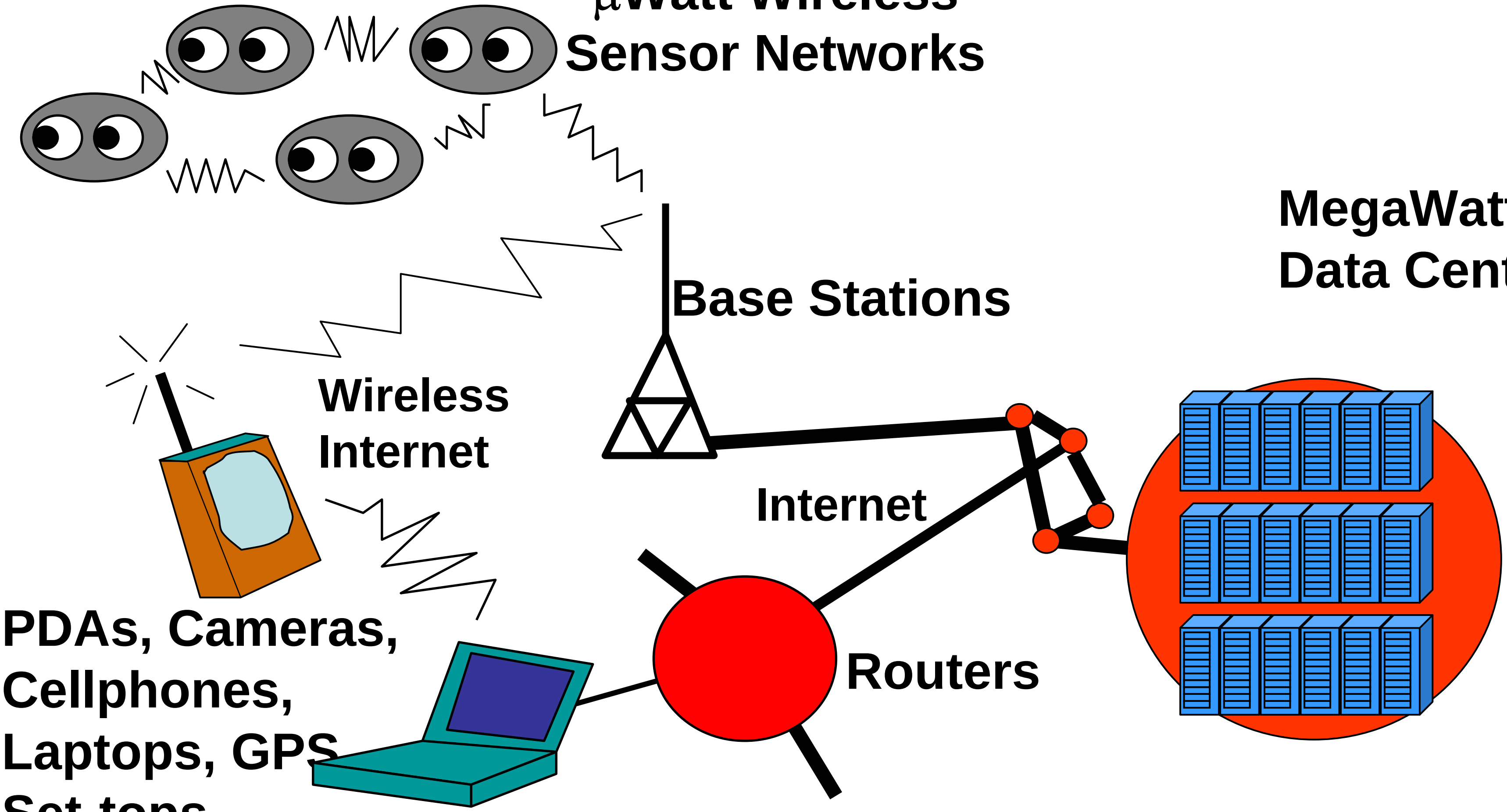
microprocessor power dissipation

source : arpa-esto



DEC 21164

Computers Defined by Watts not MIPS:



μWatt Wireless
Sensor Networks

MegaWatt
Data Centers

Base Stations

Wireless
Internet

Internet

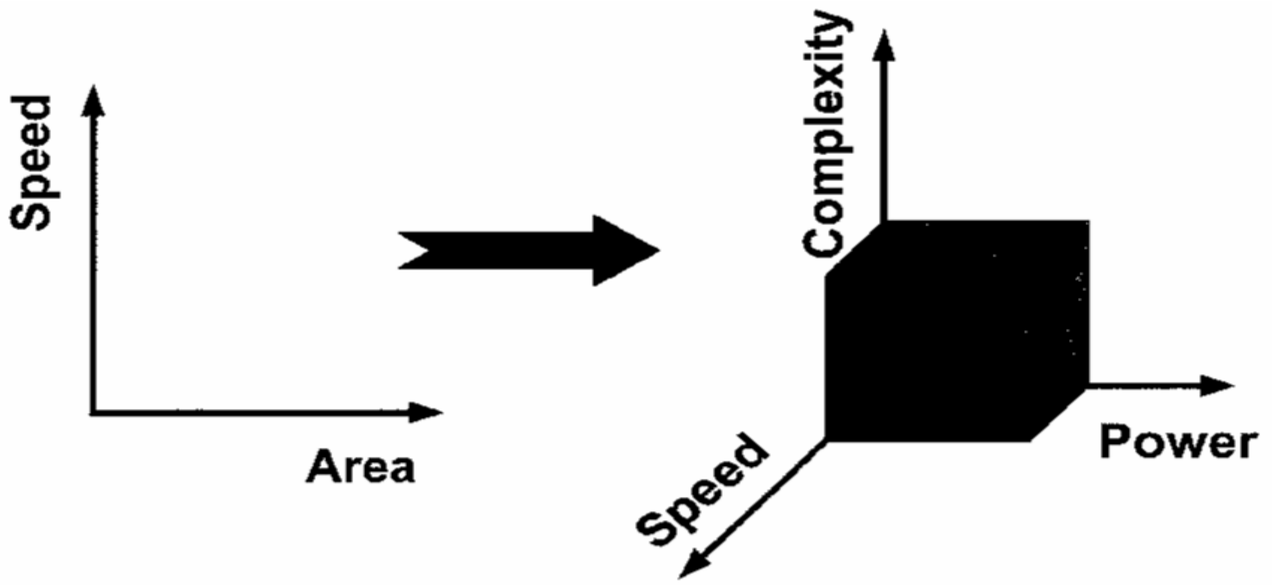
Routers

PDAs, Cameras,
Cellphones,
Laptops, GPS,
Set-tops,
0.1-10 Watt Clients

Why Low Power ?

- **Growth of battery-powered systems**
- **Users need for:**
  - **Mobility**
  - **Portability**
  - **Reliability**
- **Cost**
- **Environmental effects**

IC Design Space:



Power Impacts on System Design:

- **Energy consumed per task determines battery life**
  - **Second order effect is that higher current draws decrease effective battery energy capacity**
- **Current draw causes IR drops in power supply voltage**
  - **Requires more power/ground pins to reduce resistance R**
  - **Requires thick&wide on-chip metal wires or dedicated metal layers**
- **Switching current (dI/dT) causes inductive power supply voltage bounce $\propto$ LdI/dT**
  - **Requires more pins/shorter pins to reduce inductance L**
  - **Requires on-chip/on-package decoupling capacitance to help bypass pins during switching transients**
- **Power dissipated as heat, higher temps reduce speed and reliability**
  - **Requires more expensive packaging and cooling systems**

Facts ...

Moore´s Law - *doubling transistors every 18 months*

- Power is proportional to **die area** and **frequency**!
- In the same technology a new architecture has 2-3X in Die Area
- Changing technology implies 2X frequency

SCALING TECHNOLOGY ...

• Decreasing voltage	(0.7 scaling factor)
• Decreasing of die area	(0.5 scaling factor)
• Increasing C per unit area	43% !!!

This implies that the power density increase of 40% every generation !!!

Temperature is a function of power density and determinates the type of cooling system needed.

<u>VARIABLES</u>

- PEAK POWER (worst case)

  Today´s packages can sustain a power dissipation over 100W for up to 100msec >>> *cheaper package if peaks are reduced*

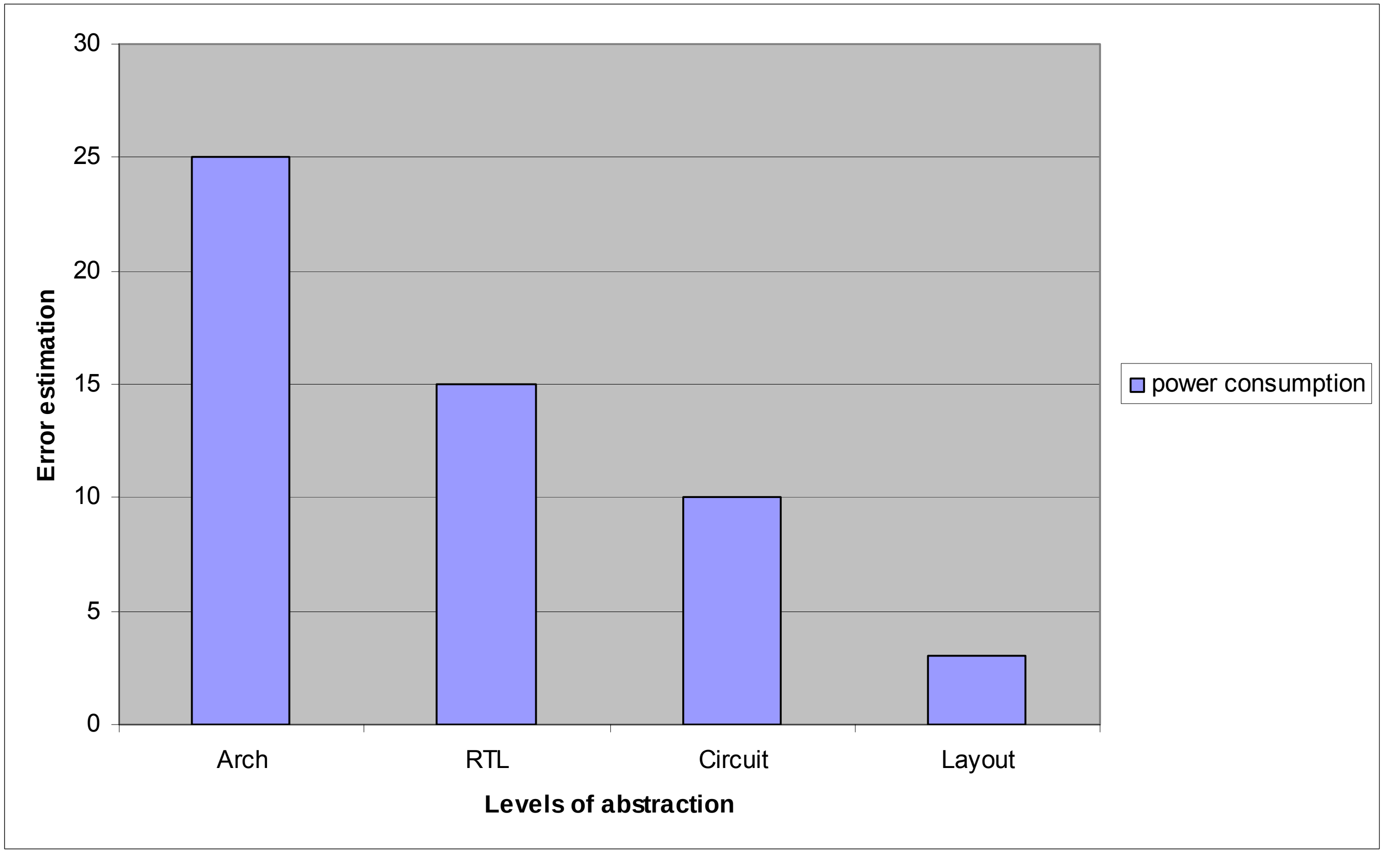
- ENERGY SPENT (for a workload)

  More correlated to battery life

Low Power Strategies:

- OS level : PARTITIONING, POWER DOWN
- Software level : REGULARITY, LOCALITY, CONCURRENCY
  (Compiler technology for low power, instruction scheduling)
- Architecture level : PIPELINING, REDUNDANCY, DATA ENCODING
  (ISA, architectural design, memory hierarchy, HW extensions, etc)
- Circuit/logic level : LOGIC STYLES, TRANSISTOR SIZING, ENERGY RECOVERY
  (Logic families, conditional clocking, adiabatic circuits, asynchronous design)
- Technology level : *Threshold reduction, multi-threshold devices, etc*

Power Consumption Estimation:



Due to the relative high error rate in the architectural estimation *(no vision of the total area, circuit types, technology, block activity, etc)*

IMPORTANT DESIGN DECISIONS MUST BE DONE AT ARCHITECTURAL LEVEL!

- Accurate power evaluation is done at late design phases
- Needs of good feedback between all the design phases

  - *Correlation between power estimation from low level to high level*

TRY TO IMPROVE ACCURACY AT HIGH LEVEL

- *Critical path based power consumption analysis*

  (CIRCUIT TYPES, TECHNOLOGY, ACTIVITY FACTOR)

- *Thermal images based correlation analysis*

  (HOTTEST SPOTS LOCATION, COOLEST SPOTS LOCATION, TEMPERATURE DIFFERENCES, TEMPERATURE DISTRIBUTION)

Architectural Power Evaluation:

Architectural design partition

- Power consumption evaluation at block level

  - *Power density of blocks (SPICE simulation, statistical input set, technology and circuit types definition)*

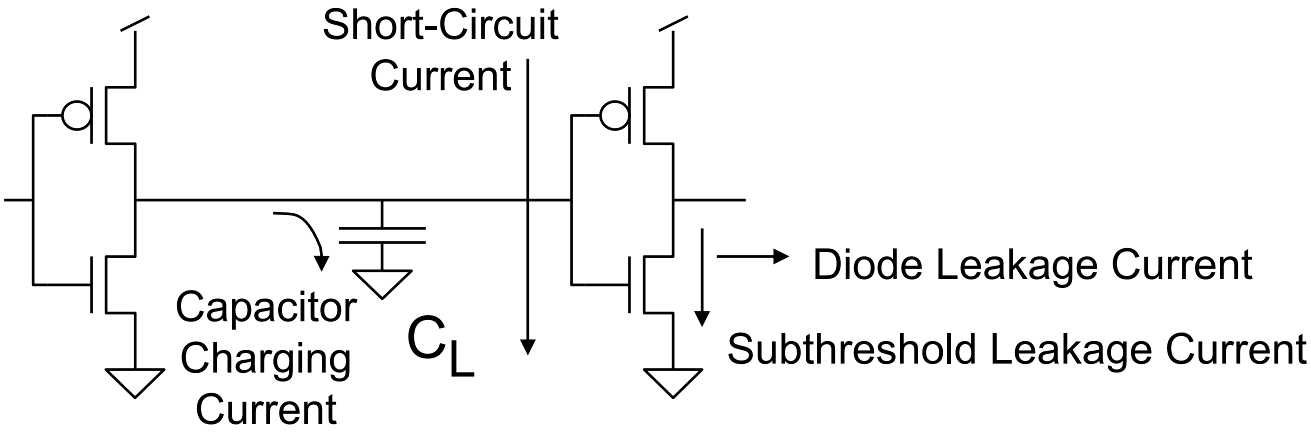
  - *Activity of blocks and sub-blocks (running benchmarks)*

  - *Area (feedback from VLSI design, circuits and technology defined)*

- Try do define scaling factors that allow to remap the architectural power simulator when technology, area and circuit types change

- Try to reduce the error estimation at high level

Power Dissipation in CMOS:



Primary Components:

- **Capacitor Charging (85-90% of active power)**
  - **Energy is ½ CV^2 per transition**
- **Short-Circuit Current (10-15% of active power)**
  - **When both p and n transistors turn on during signal transition**
- **Subthreshold Leakage (dominates when inactive)**
  - **Transistors don’t turn off completely**
- **Diode Leakage (negligible)**
  - **Parasitic source and drain diodes leak to substrate**

Sources of Power Dissipation:

- Dynamic power dissipations: whenever the logic level changes at different points in the circuit because of the change in the input signals the dynamic power dissipation occurs.
  - Switching power dissipation.
  - Short-circuit power dissipation.
- Static power dissipations: this is a type of dissipation, which does not have any effect of level change in the input and output.
  - Leakage power.

Switching Power Dissipation:

- Caused by the charging and discharging of the node capacitance.

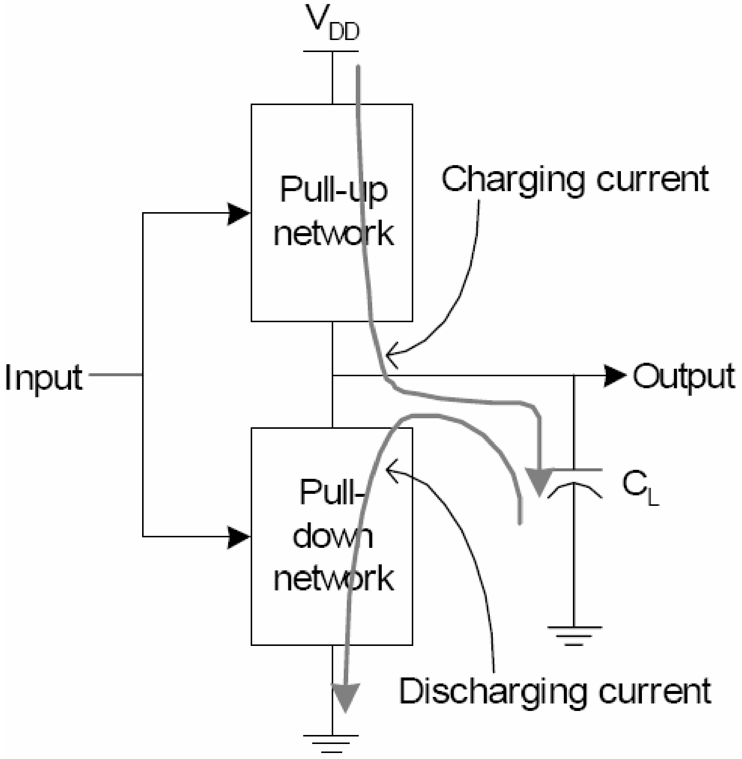


Figure 1: Switching power dissipation [1].

Switching Power Dissipation (Contd.):

- $P_{s/w} = 0.5 * \alpha * C_L * V_{dd}^2 * f_{clk}$
    - C_L physical capacitance, V_{dd} supply voltage, α switching activity, f_{clk} clock frequency.
  - $C_L(i) = \Sigma_j C_{IN}^j + C_{wire} + C_{par(i)}$
    - C_{IN} the gate input capacitance, C_{wire} the parasitic interconnect and C_{par} diffusion capacitances of each gate[I].
- Depends on:
  - Supply voltage
  - Physical Capacitance
  - Switching activity

Short circuit power dissipation:

- Caused by simultaneous conduction of n and p blocks.

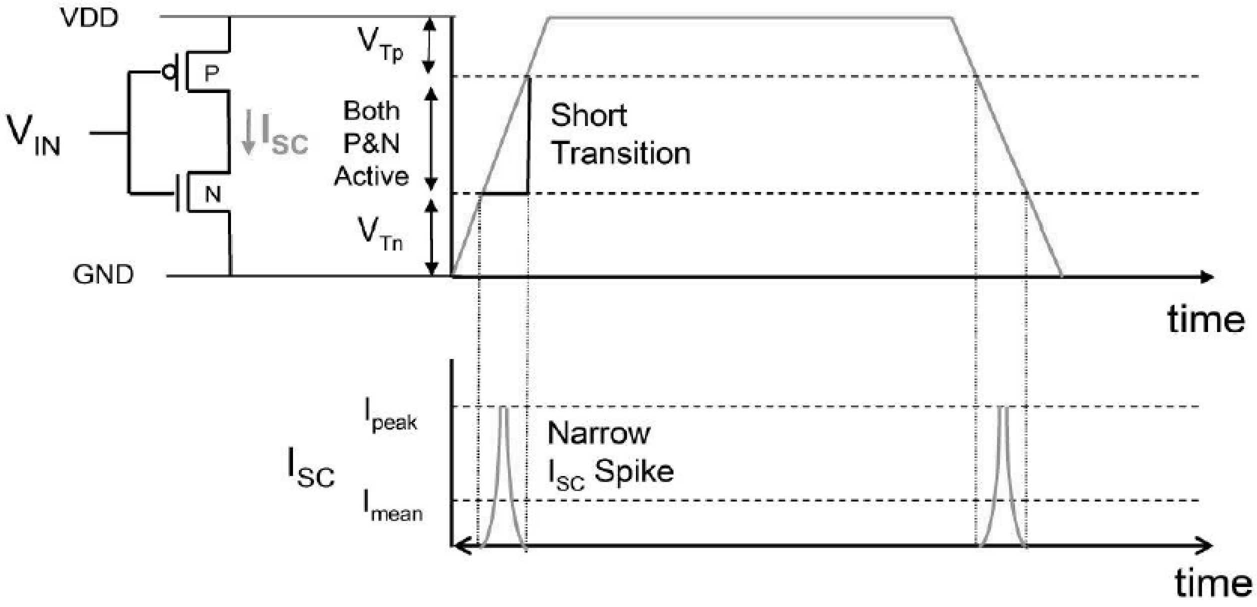


Figure 2: Short circuit current

Short circuit power dissipation (contd.):

$$P_{sc} = I_{sc} \cdot V_{DD} = \frac{1}{12} \cdot k \cdot \tau \cdot (V_{DD} - 2V_T)^3 \cdot f$$

where k = (k_n = k_p), the trans conductance of the transistor,

τ = (t_{rise} = t_{fall}), the input/output transition time, V_{DD} = supply voltage,

f = clock frequency, and V_T = (V_{Tn} = $|V_{Tp}|$), the threshold voltage of MOSFET.

- Depends on :
  - The input ramp
  - Load
  - The transistor size of the gate
  - Supply voltage
  - Frequency
  - Threshold voltage.

Leakage power dissipation:

- Six short-channel leakage mechanisms are there:
  - I_1 Reverse-bias p-n junction leakage
  - I_2 Sub threshold leakage
  - I_3 Oxide tunneling current
  - I_4 Gate current due to hot-carrier injection
  - I_5 GIDL (Gate Induced Drain Leakage)
  - I_6 Channel punch through current

- I_1 and I_2 are the dominant leakage mechanisms

Leakage power dissipation (contd.)

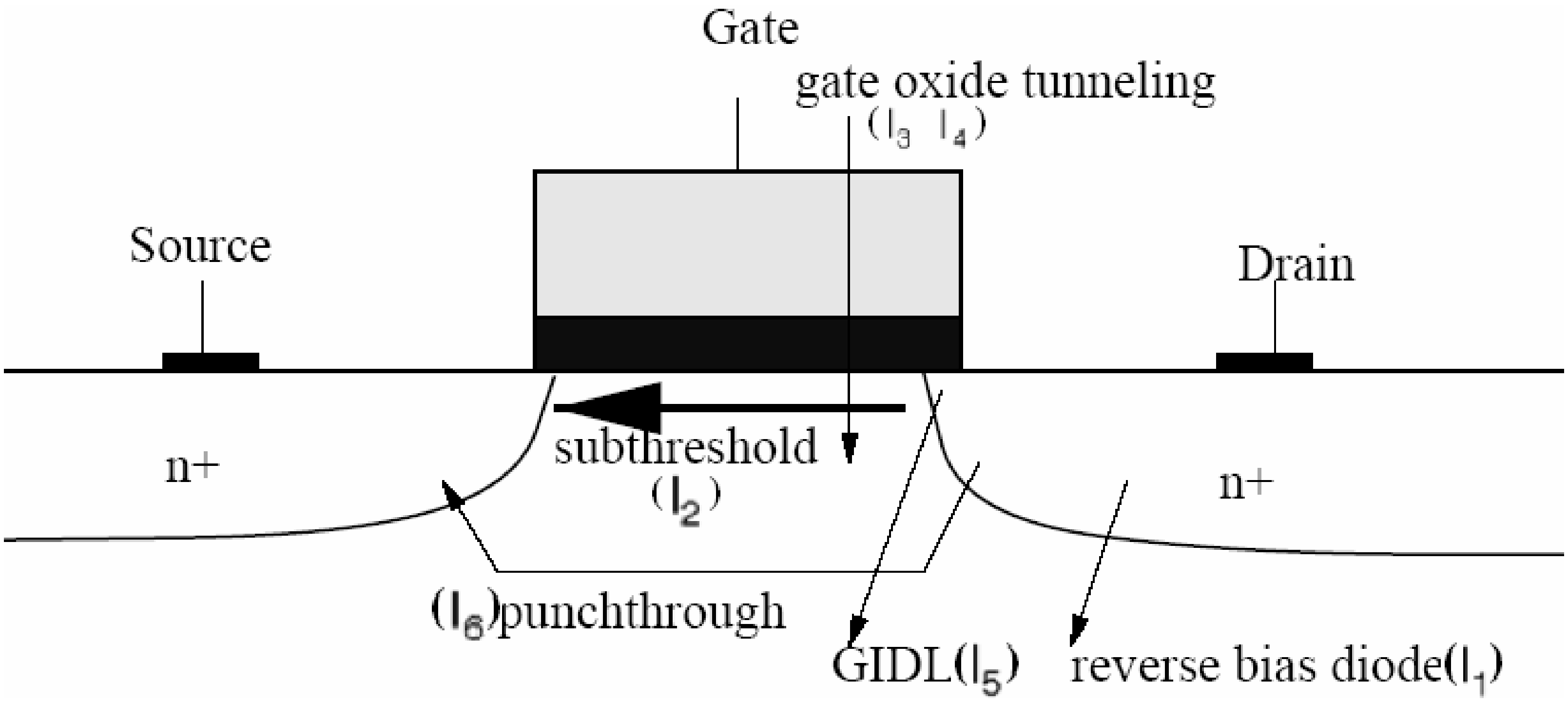


Figure 3: Summary of leakage current mechanism [2]

PN Junction reverse bias current:

- The reverse biasing of p-n junction cause reverse bias current
  - Caused by diffusion/drift of minority carrier near the edge of the depletion region.

$$I_{\text{reverse}} = A \cdot J_s \cdot \left[e^{\frac{q \cdot V_{\text{bias}}}{k \cdot T}} - 1 \right]$$

where V_{bias} = the reverse bias voltage across the p-n junction, J_s = the reverse saturation current density and A = the junction area.

Sub Threshold Leakage Current:

- Caused when the gate voltage is below $V_{th.}$

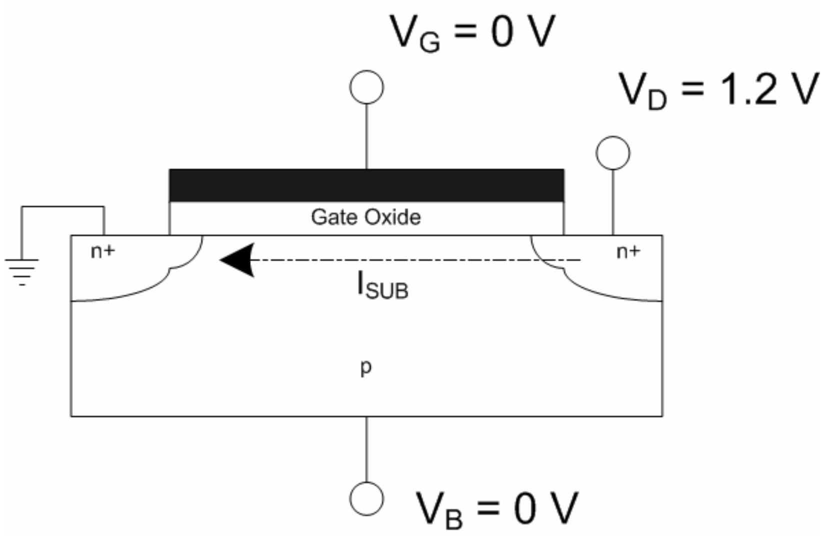


Fig 4: Sub threshold current[2]

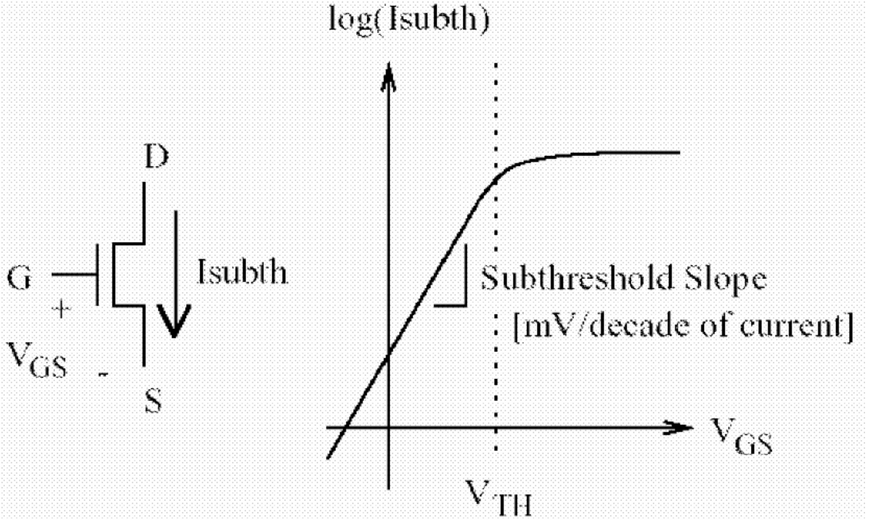


Fig 5: Subthreshold leakage in a negative-channel metal–oxide–semiconductor (NMOS) transistor.[2]

Contribution of Different Power Dissipation:

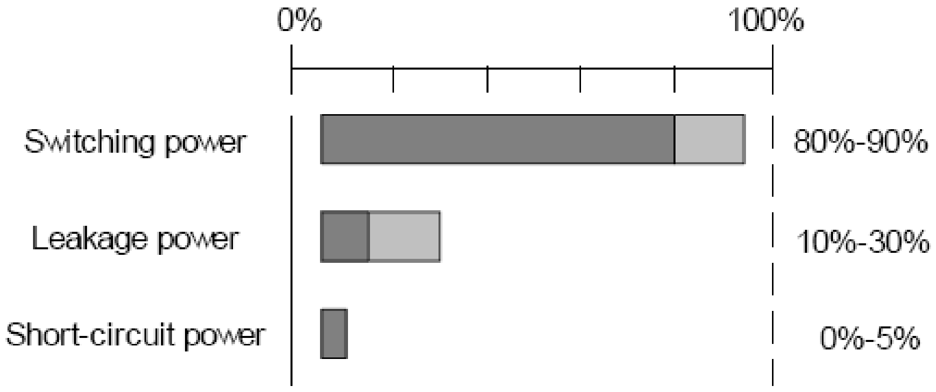


Fig 6: Contribution of different powers[1]

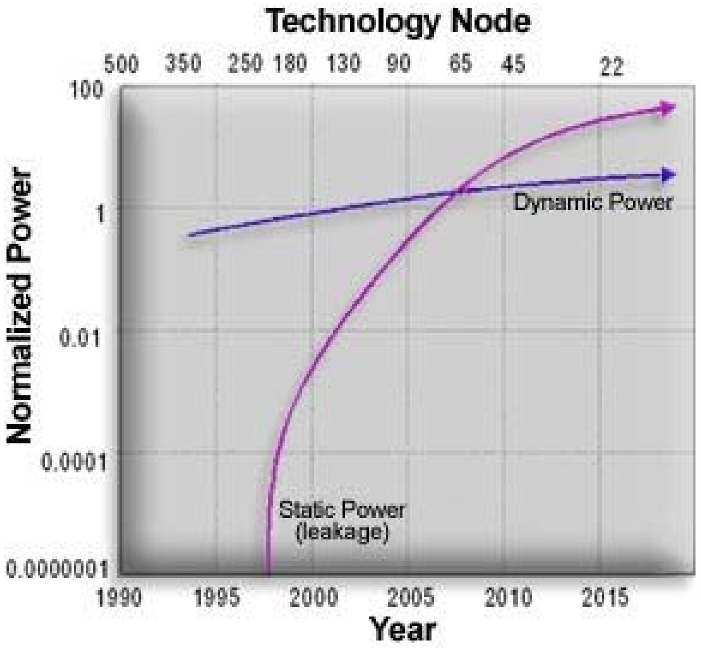


Fig 7:Static power increases with shrinking device geometries [7].

Degrees of Freedom

- The three degrees of freedom are:
  - Supply Voltage
  - Switching Activity
  - Physical capacitance

Reducing Power:

- **Switching power $\propto$ activity*½ CV^2*frequency**
  - **(Ignoring short-circuit and leakage currents)**
- **Reduce activity**
  - **Clock and function gating**
  - **Reduce spurious logic glitches**
- **Reduce switched capacitance C**
  - **Different logic styles (logic, pass transistor, dynamic)**
  - **Careful transistor sizing**
  - **Tighter layout**
  - **Segmented structures**
- **Reduce supply voltage V**
  - **Quadratic savings in energy per transition – BIG effect**
  - **But circuit delay is reduced**
- **Reduce frequency**
  - **Doesn’t save energy just reduces rate at which it is consumed**
  - **Some saving in battery life from reduction in current draw**

Supply Voltage Scaling

- Switching and short circuit power are proportional to the square of the supply voltage.
- But the delay is proportional to the supply voltage. So, the decrease in supply voltage will results in slower system.

$$dealy = \text{constant}\frac{C_L}{V_{DD}\left(1-\frac{V_T}{V_{DD}}\right)^2}$$

- Threshold voltage can be scaled down to get the same performance, but it may increase the concern about the leakage current and noise margin.

Supply Voltage Scaling (contd.)

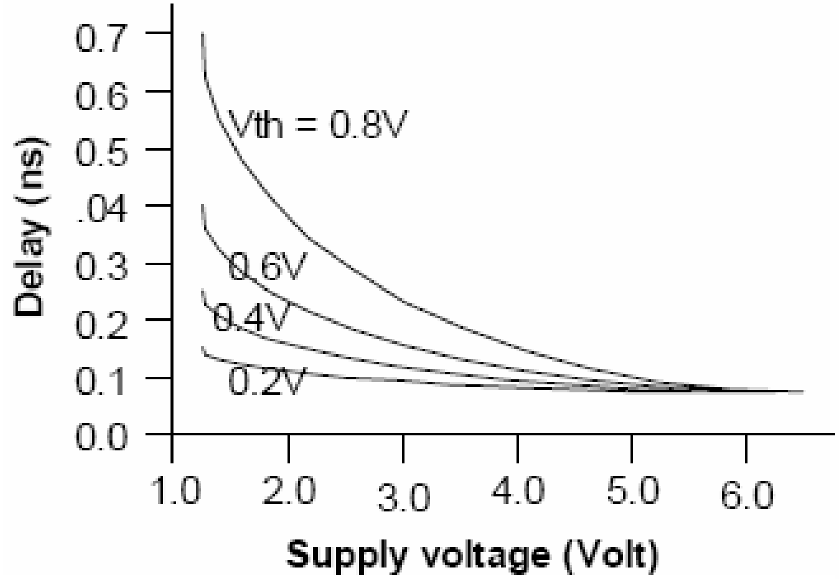


Fig 8: Scaling supply and threshold voltages [4]

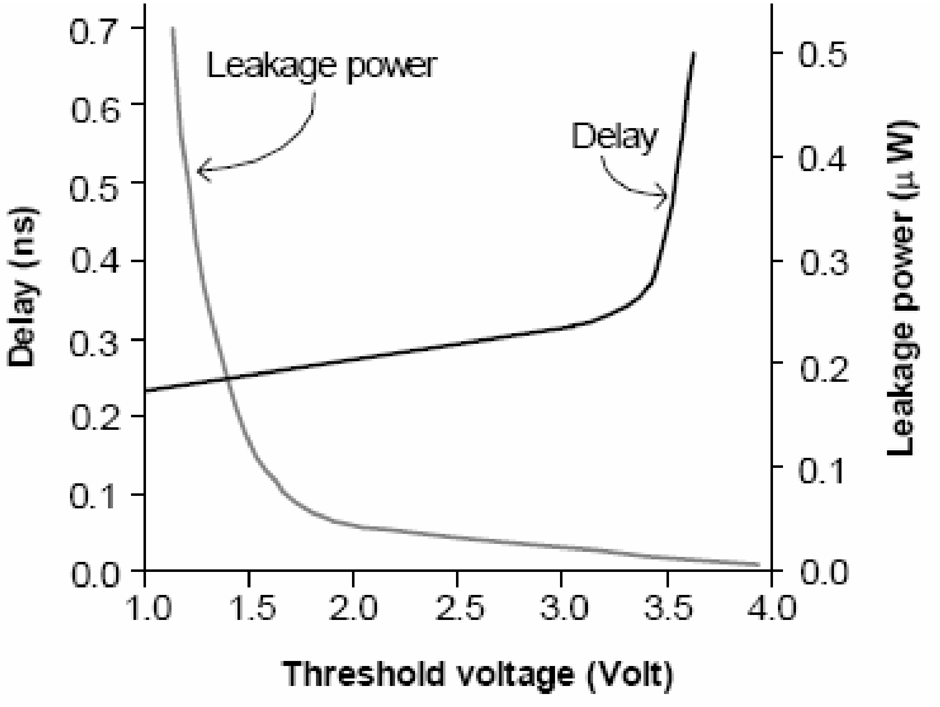


Fig 9: Scaling of threshold voltage on leakage power and delay[4]

Switching Activity Reduction

- Two components:
  - f: The average periodicity of data arrivals
  - α: how many transitions each arrival will generate.
- There will be no net benefits by Reducing f.
- α can be reduced by algorithmic optimization, by architecture optimization, by proper choice of logic topology and by logic-level optimization.

Physical capacitance reduction

- Physical capacitance in a circuit consists of three components:
  - The output node capacitance (C_L).
  - The input capacitance (C_{in}) of the driven gates.
  - The total interconnect capacitance (C_{int}).
- Smaller the size of a device, smaller is $C_{L.}$
- The gate area of each transistor determines $C_{in.}$
- C_{int} is determine by width and thickness of the metal/oxide layers with which the interconnect line is made of, and capacitances between layers around the interconnect lines.

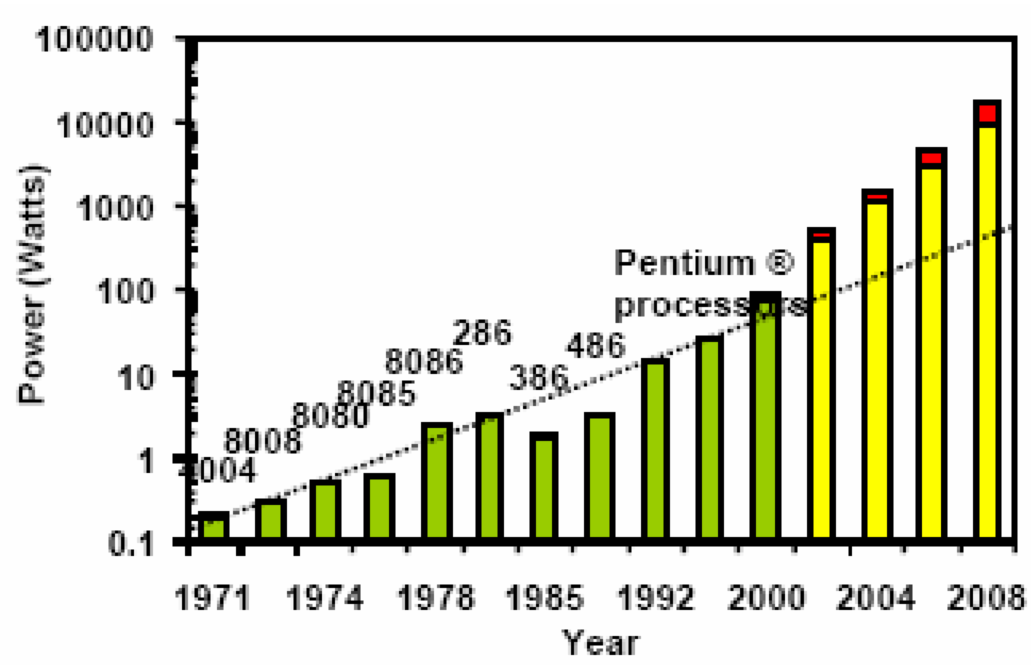
Issues

- Technology Scaling
  - – Capacitance per node reduces by 30%
  - – Electrical nodes increase by 2X
  - – Die size grows by 14% (Moore's Law)
  - – Supply voltage reduces by 15%
  - – And frequency increases by 2X

  This will increase the active power by 2.7X

Issues (contd.)

- To meet frequency demand V_t will be scaled, resulting high leakage power.

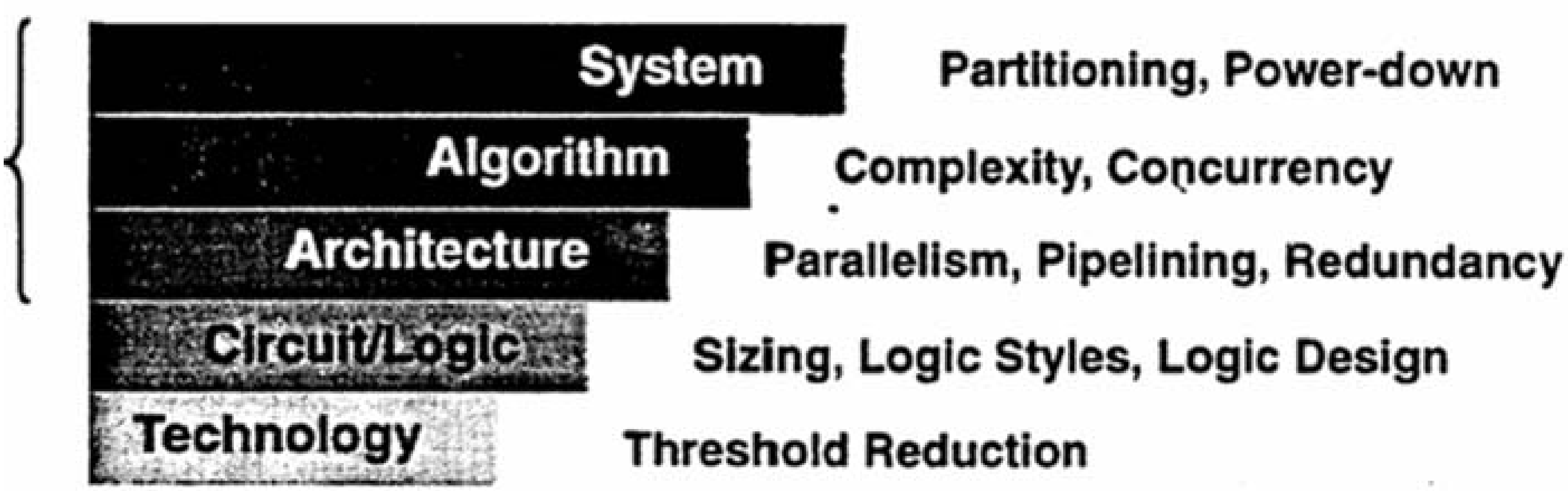


*Source: Intel

Fig 10:Total power consumption of a microprocessor following Moore’s Law

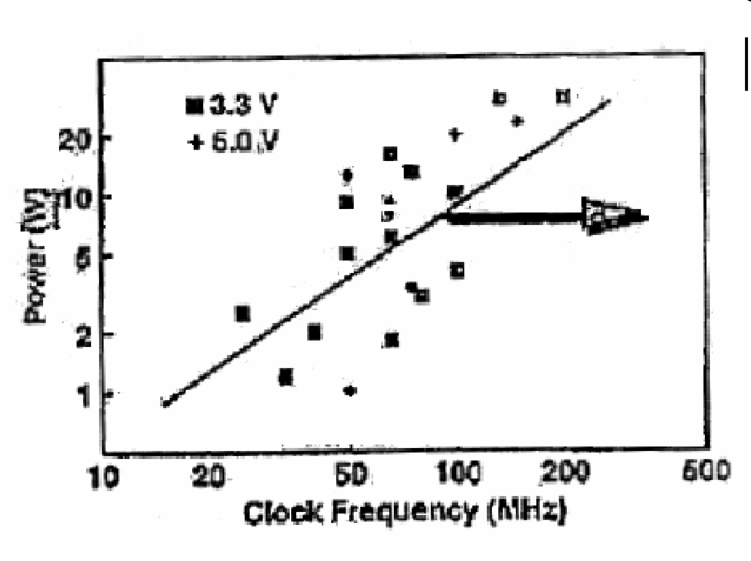
Ultra Low Power System Design:

- Power minimization approaches:
  - Run at minimum allowable voltage
  - Minimize effective switching capacitance

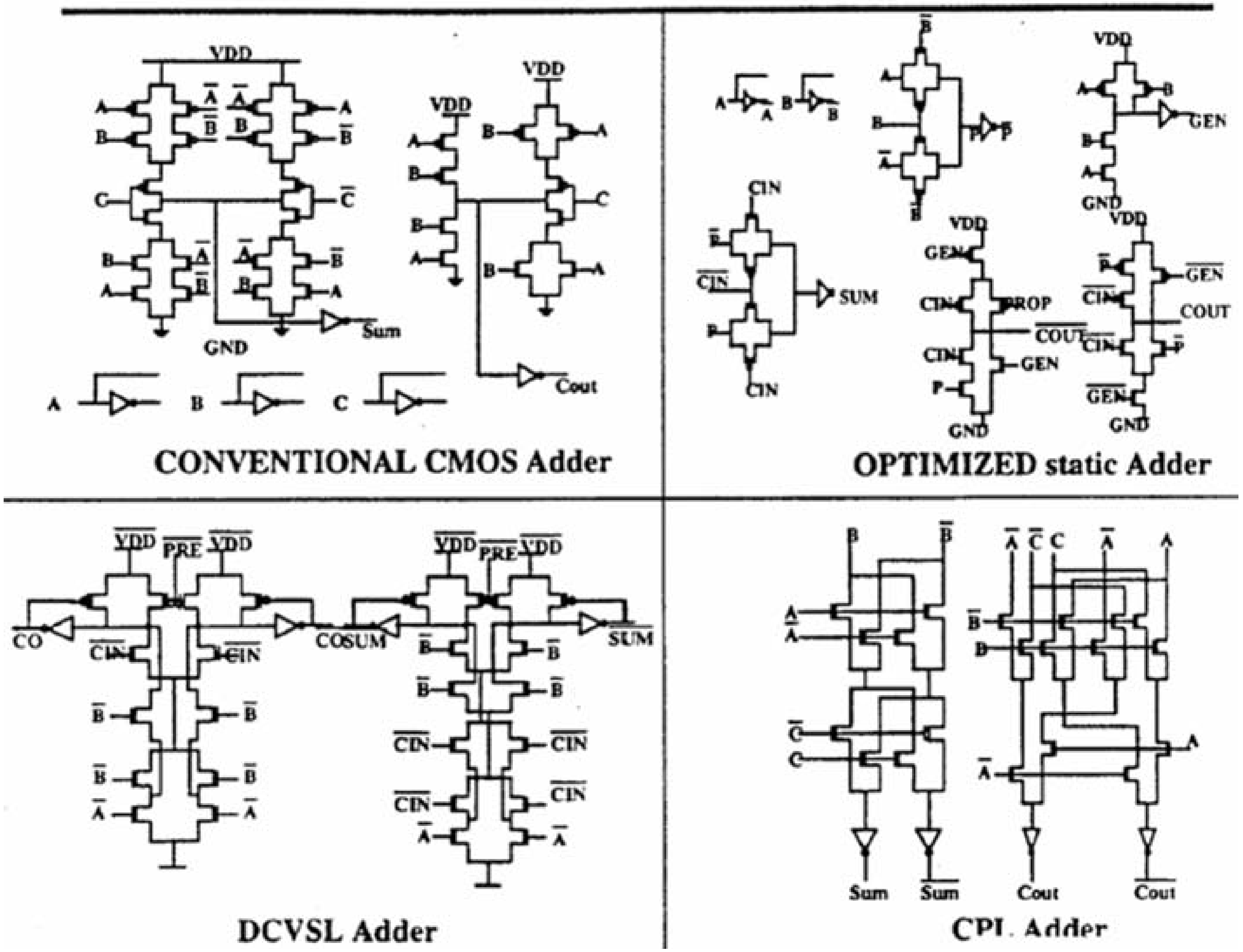


Process

- Progress in SOI and bulk silicon
  - (a) 0.5V operation of ICs using SOI technology
  - (b) 0.9V operation of bulk silicon memory, logic, and processors
- Increasing densities and clock frequencies have pushed the p luce power supply



Choice of Logic Style



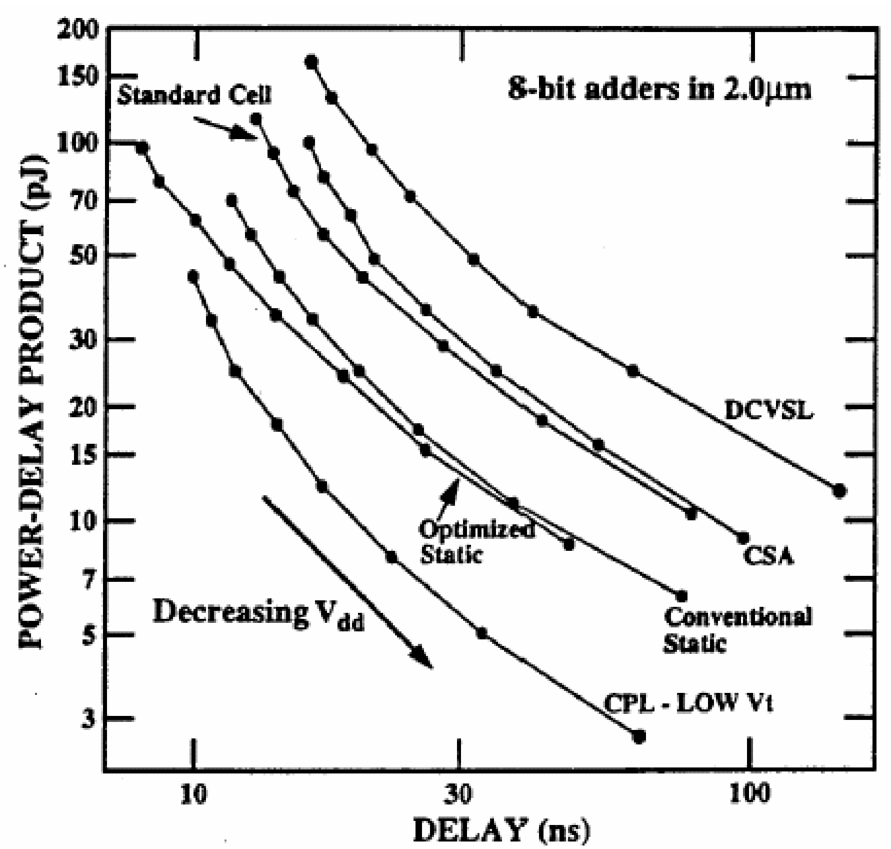
CONVENTIONAL CMOS Adder

OPTIMIZED static Adder

DCVSL Adder

CPL Adder

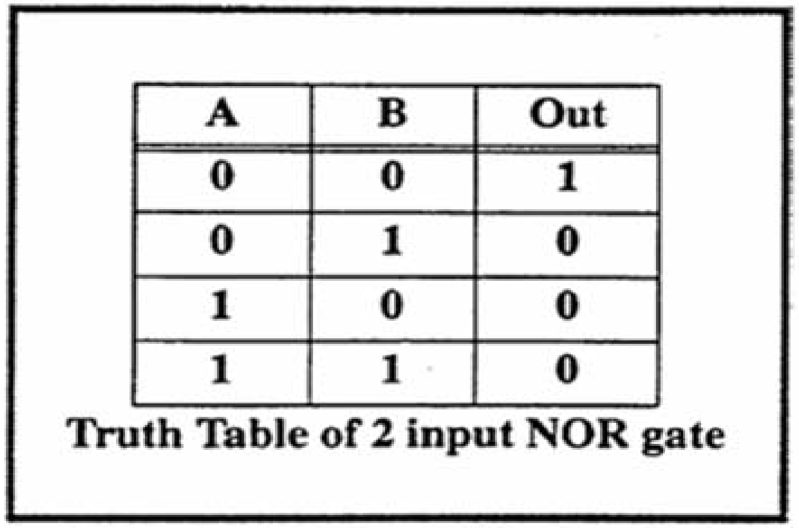
Choice of Logic Style



- Power-delay product improves as voltage decreases
- The “best” logic style minimizes power-delay for a

Power Consumption is Data Dependent

- Example : Static 2 Input NOR Gate



A	B	Out
0	0	1
0	1	0
1	0	0
1	1	0

Truth Table of 2 input NOR gate

Assume :

$P(A=1) = \frac{1}{2}$

$P(B=1) = \frac{1}{2}$

Then :

$P(Out=1) = \frac{1}{4}$

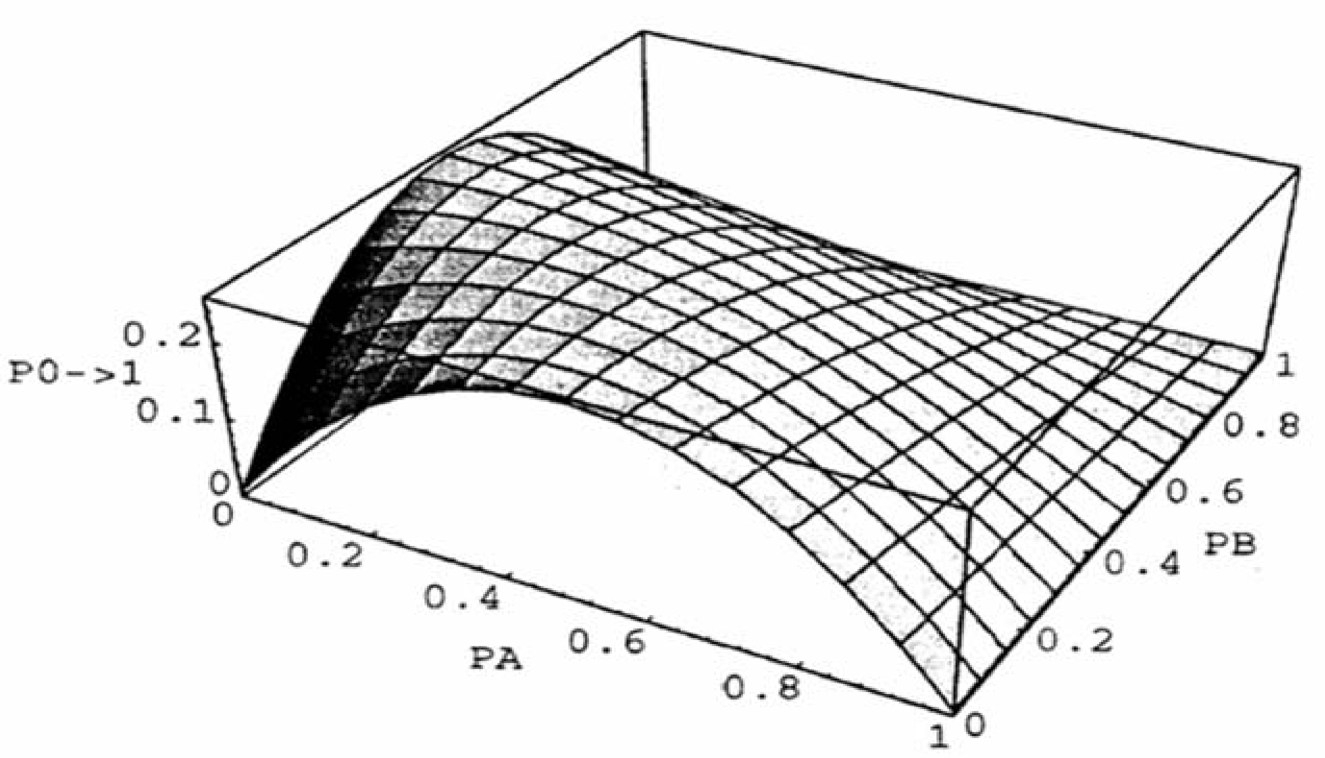
$P(0 \rightarrow 1)$

$= P(Out=0).P(Out=1)$

$=3/4 * 1/4 = 3/16$

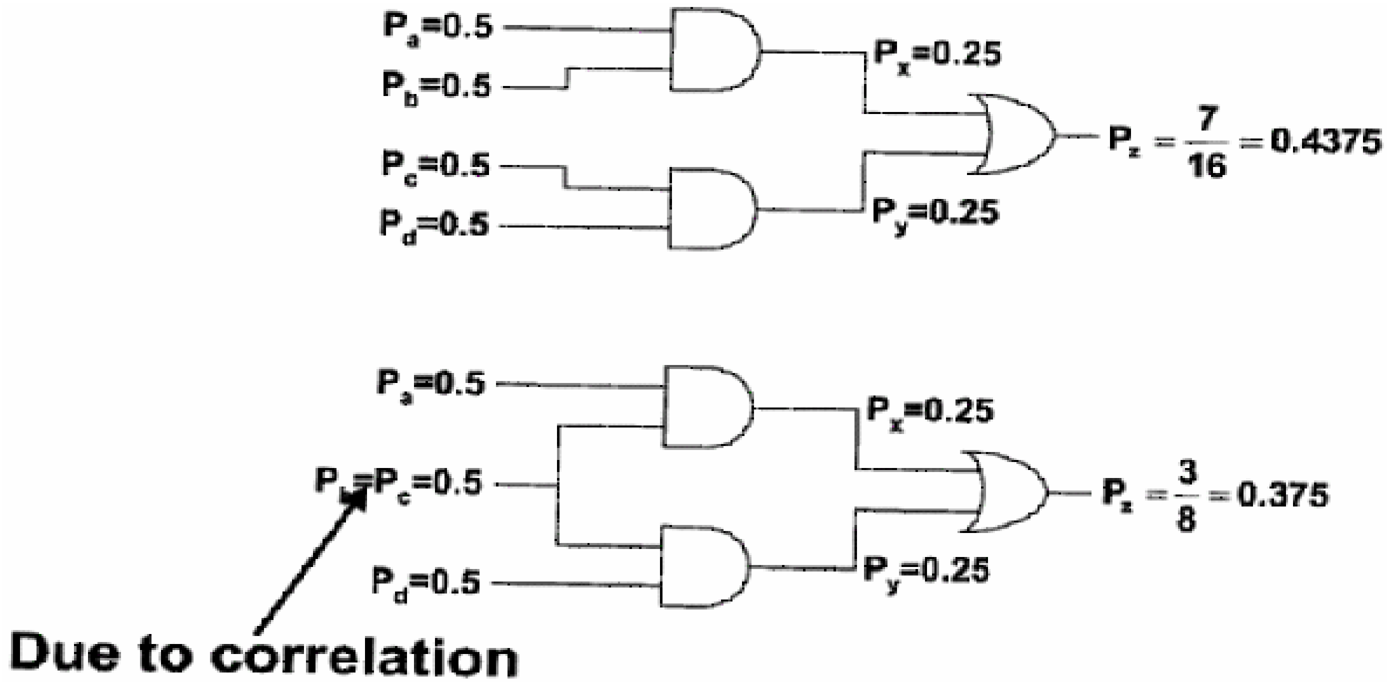
$C_{EFF} = 3/16 * C_L$

Transition Probability of 2-input NOR Gate



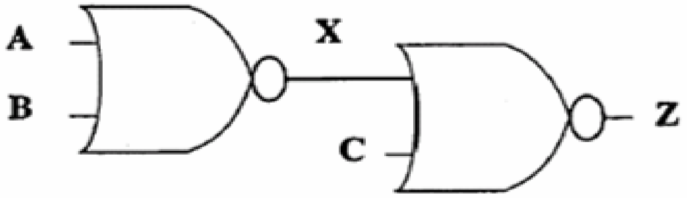
as a function of input probabilities

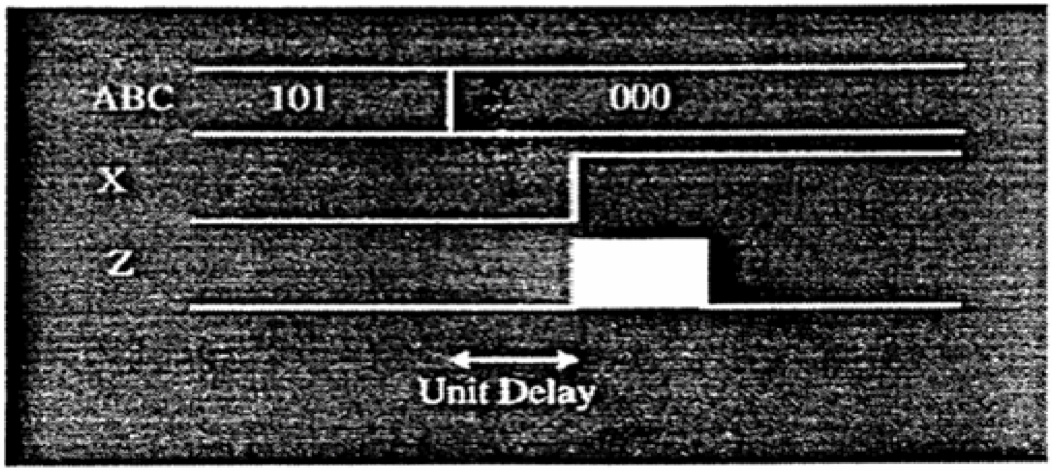
Switching Activity (α) : Example



Glitching in Static CMOS

also called: dynamic hazards

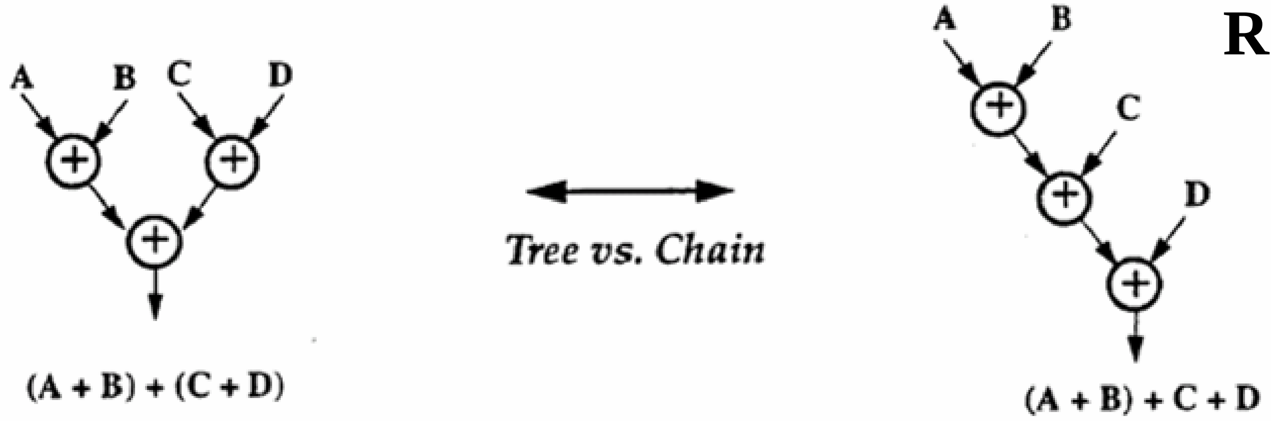




At the Datapath Level…

Irregular

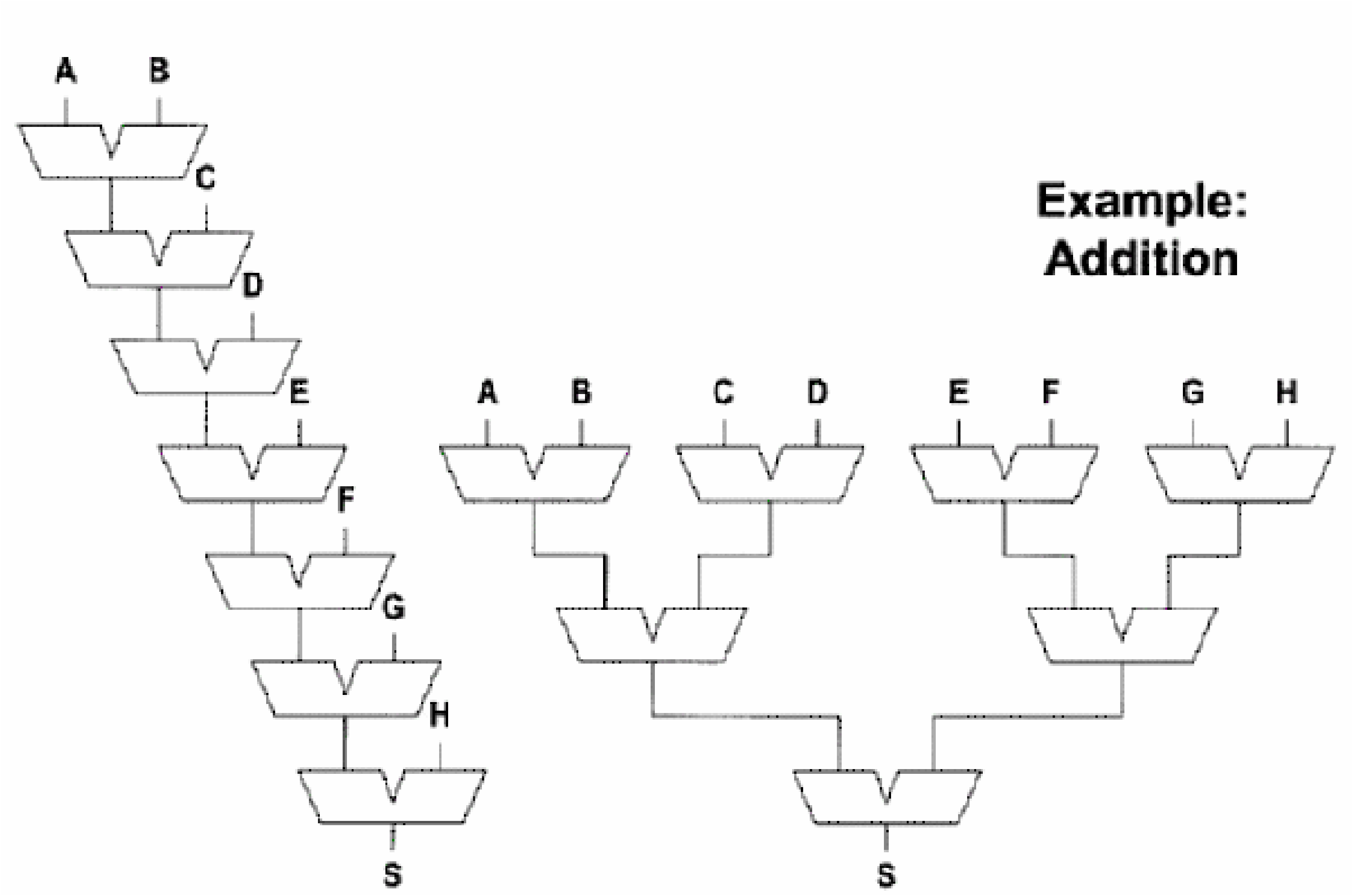
Reusable



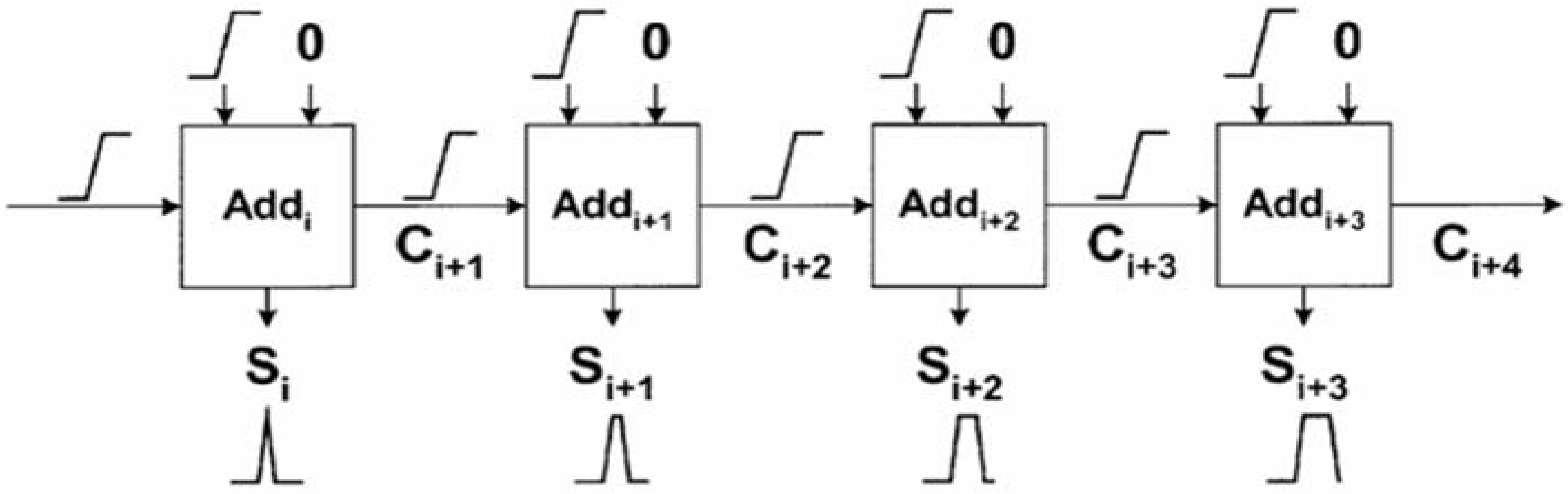
Inputs	Normalized # of Transitions	
	Tree	Chain
4	1	1.45
8	1	2.5

- Can be reduced by reducing the logic depth and balancing signal paths.

Balancing Operations



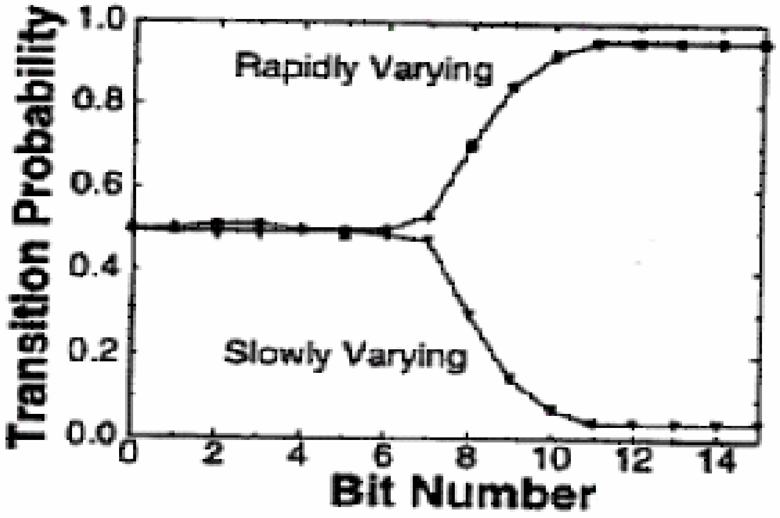
Carry Ripple



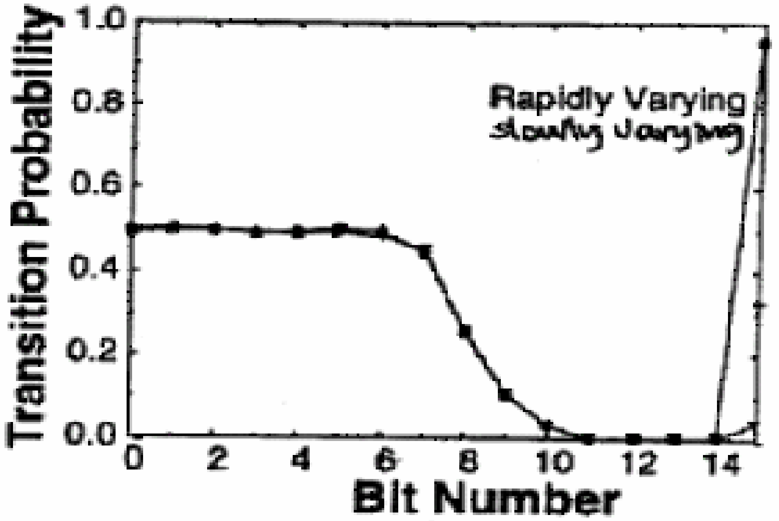
Transitions due to carry propagation

Data Representation

Two's Complement

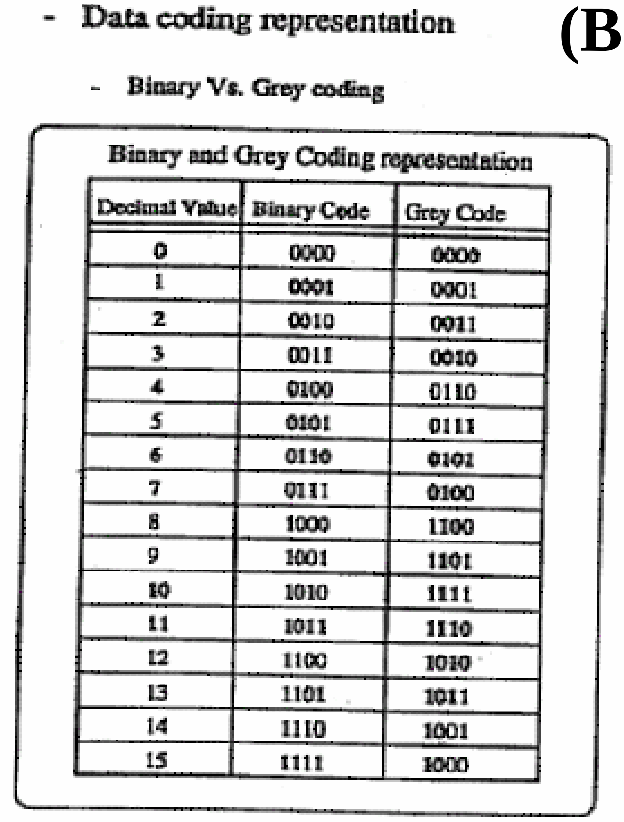


Sign Magnitude



- **Sign-extension activity significantly reduced using sign-magnitude representation**

Low Power Design Consideration (cont’)

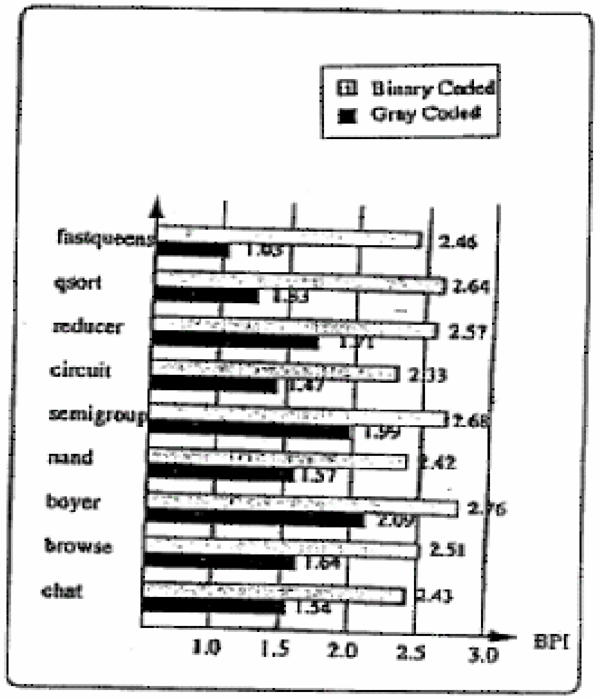


- Data coding representation
  - Binary Vs. Grey coding

Binary and Grey Coding representation

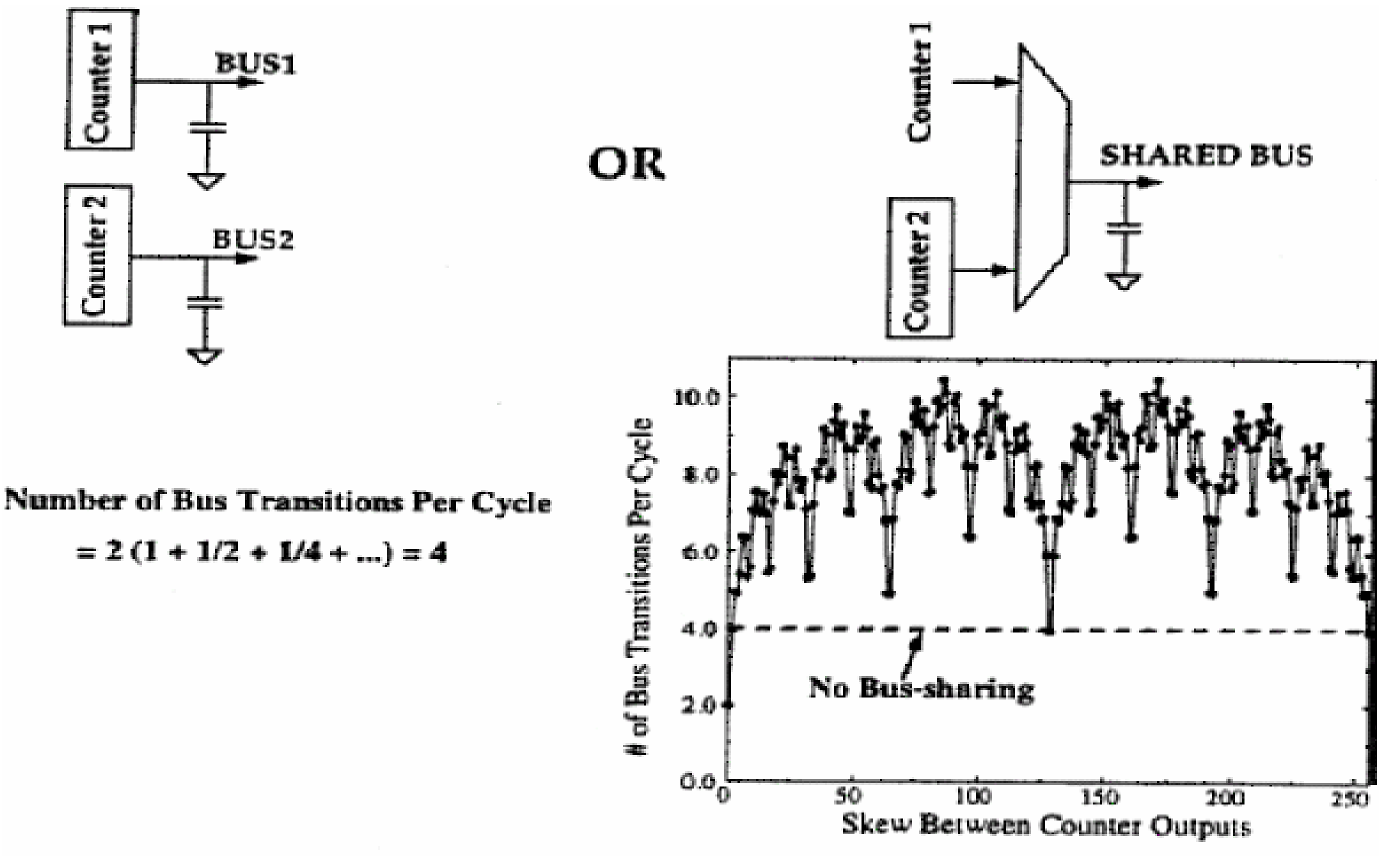
Decimal Value	Binary Code	Grey Code
0	0000	0000
1	0001	0001
2	0010	0011
3	0011	0010
4	0100	0110
5	0101	0111
6	0110	0101
7	0111	0100
8	1000	1100
9	1001	1101
10	1010	1111
11	1011	1110
12	1100	1010
13	1101	1011
14	1110	1001
15	1111	1000

(Binary v.s. Gray Encoding)



Resource Sharing Can Increase Activity

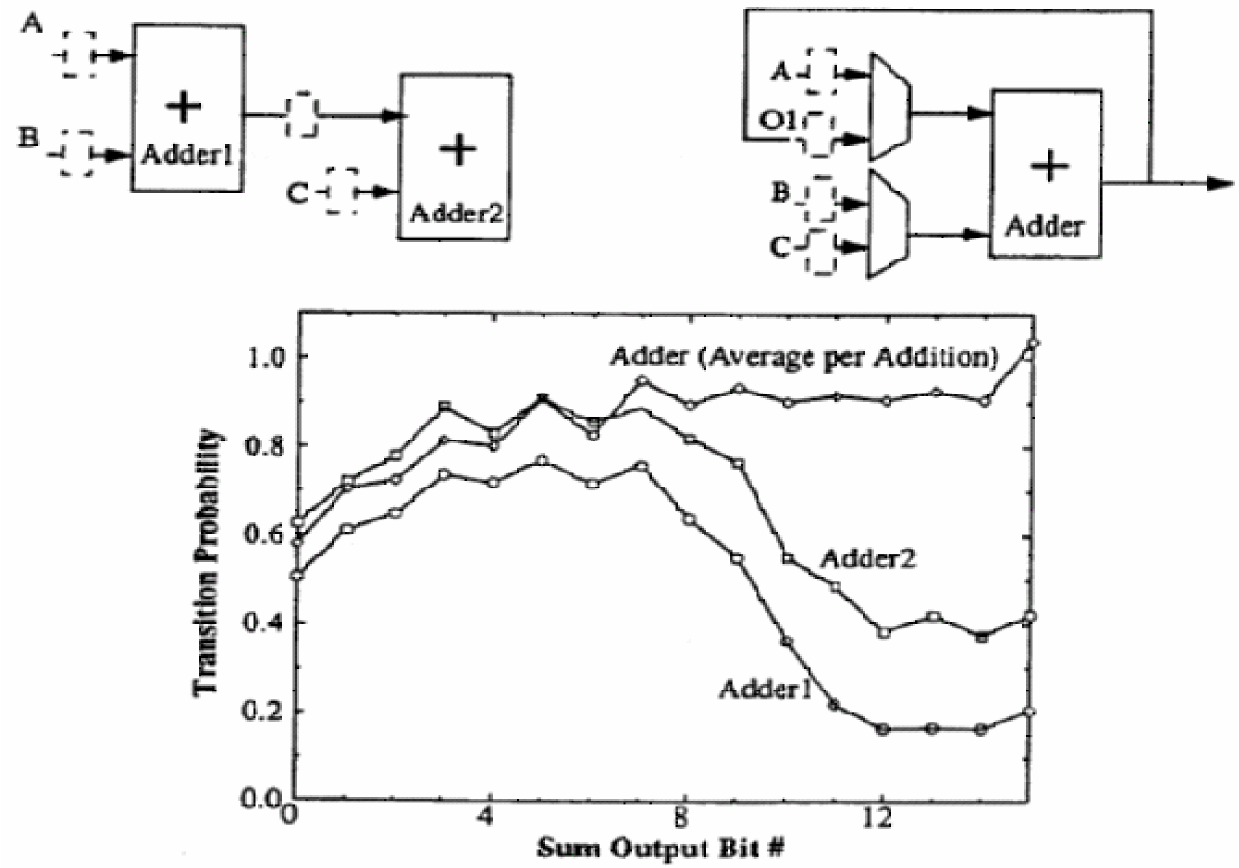
(Separate Bus Structure)



Number of Bus Transitions Per Cycle

$= 2\,(1 + 1/2 + 1/4 + \ldots) = 4$

Resource Sharing Can Increase Activity (cont'd)

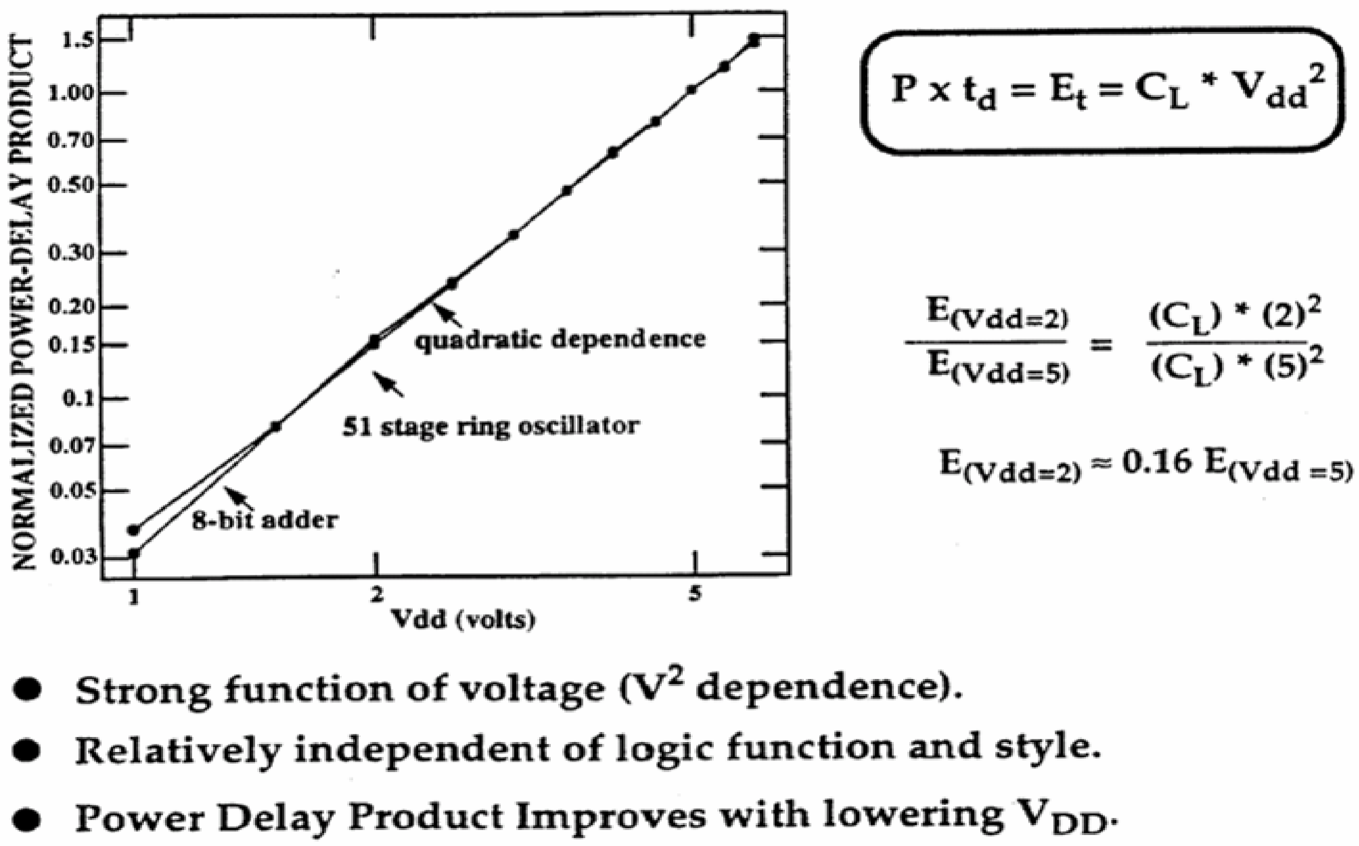


Operating at the Lowest Possible Voltage

- Desire to operate at lowest possible speeds (using low supply voltages)
- Use Architecture optimization to compensate for slower operation

Approach : Trade-off **AREA** for lower **POWER**

Reducing V_{dd}



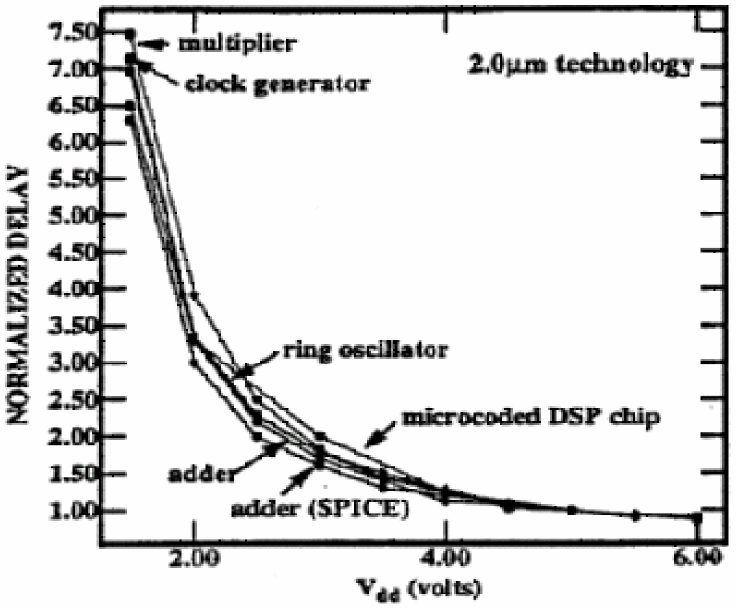
$$\boxed{P \times t_d = E_t = C_L * V_{dd}^2}$$

$$\frac{E_{(Vdd=2)}}{E_{(Vdd=5)}} = \frac{(C_L) * (2)^2}{(C_L) * (5)^2}$$

$$E_{(Vdd=2)} \approx 0.16\ E_{(Vdd=5)}$$

- Strong function of voltage (V^2 dependence).
- Relatively independent of logic function and style.
- Power Delay Product Improves with lowering V_{DD}.

Lowering V_{dd} Increases Delay



$$\boxed{T_d = \frac{C_L * V_{dd}}{I}}$$

$$I \sim (V_{dd} - V_t)^2$$

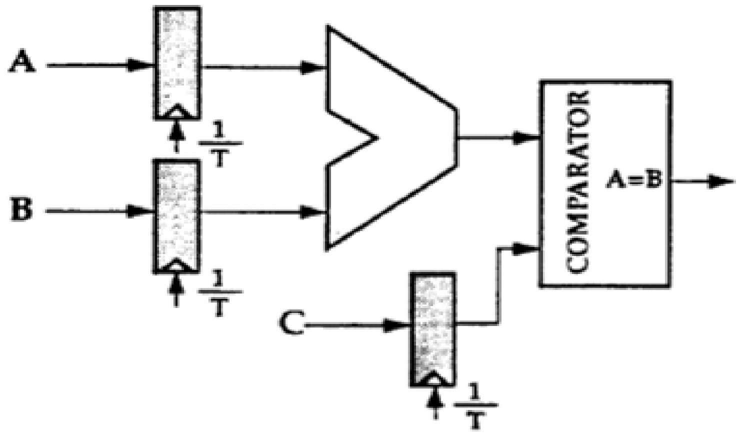
$$\frac{T_{d(Vdd=2)}}{T_{d(Vdd=5)}} = \frac{(2) * (5 - 0.7)^2}{(5) * (2 - 0.7)^2}$$

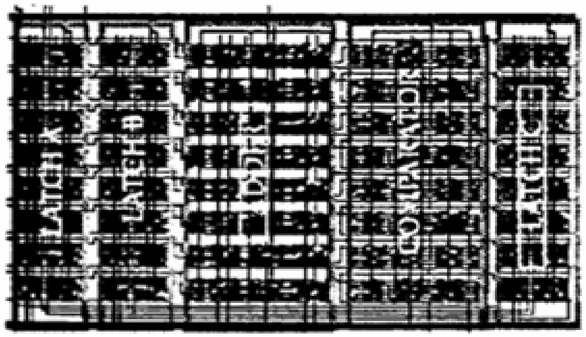
$$\approx 4$$

- Relatively independent of logic function and style.

- **Concept of Dynamic Voltage Scaling (DVS)**

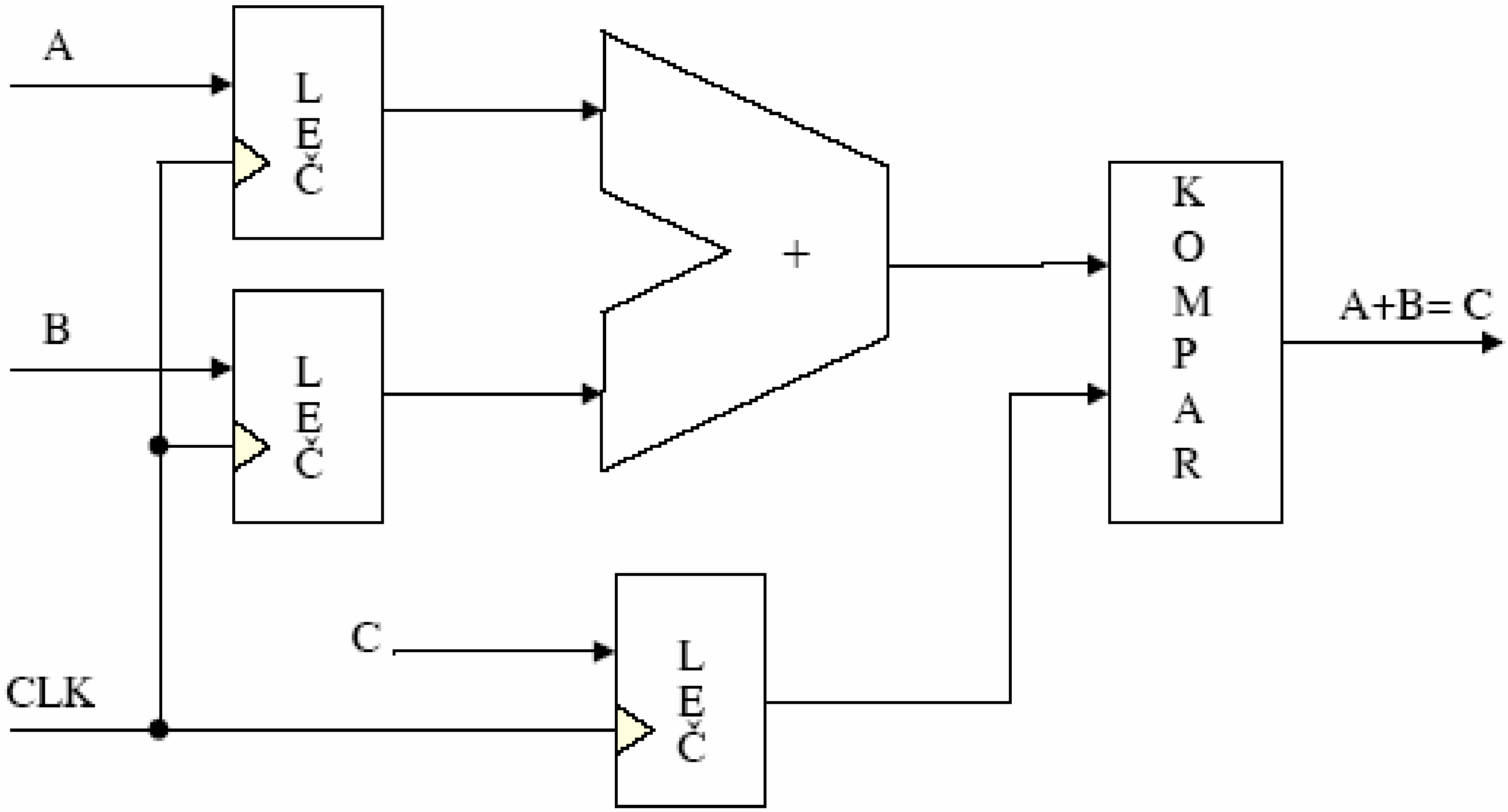
Architecture Trade-offs : Reference Data Path



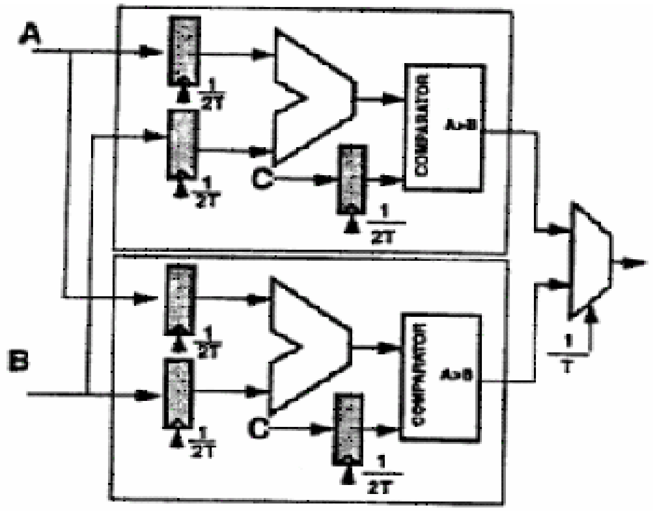


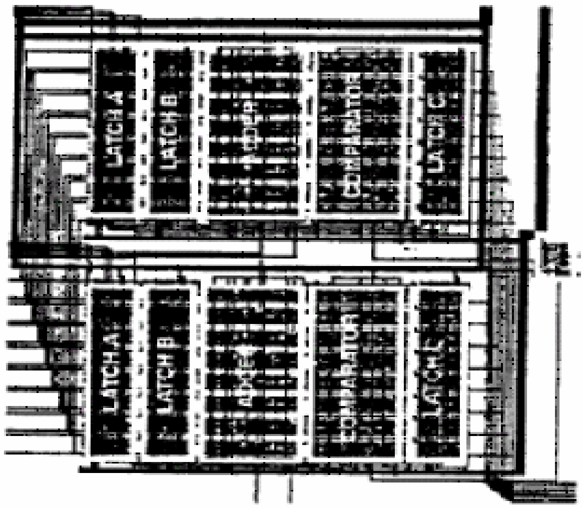
Area = 636 x 833 u^2

- Critical path delay ⇒ $T_{adder} + T_{comparator}$ (= 25ns)
  ⇒ f_{ref} = 40Mhz
- Total capacitance being switched = C_{ref}
- $V_{dd} = V_{ref} = 5V$
- Power for reference datapath = $P_{ref} = C_{ref} V_{ref}^2 f_{ref}$


A
B
C
CLK
L E Č
L E Č
L E Č
+
K O M P A R
A+B= C

Parallel Data Path

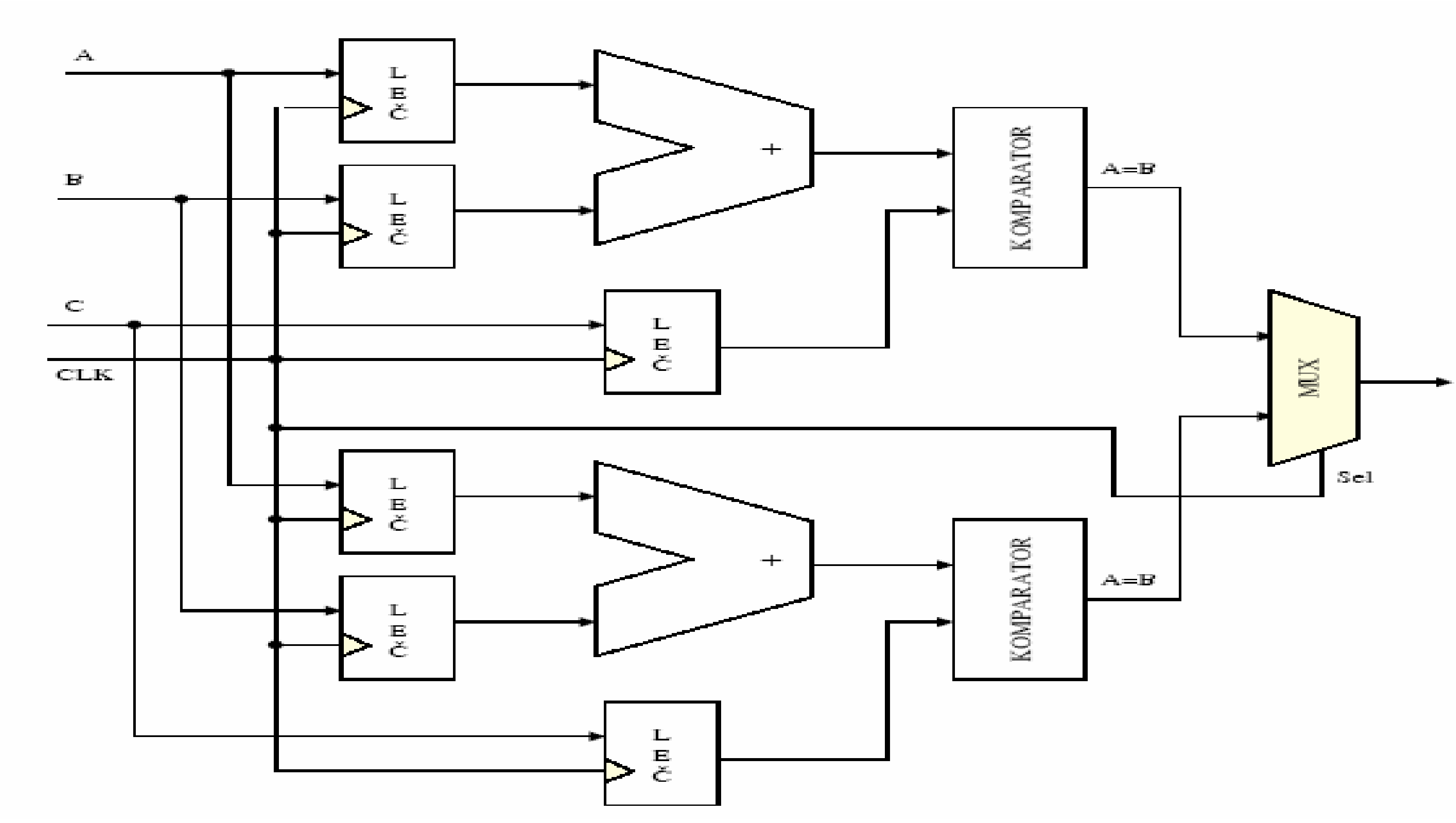




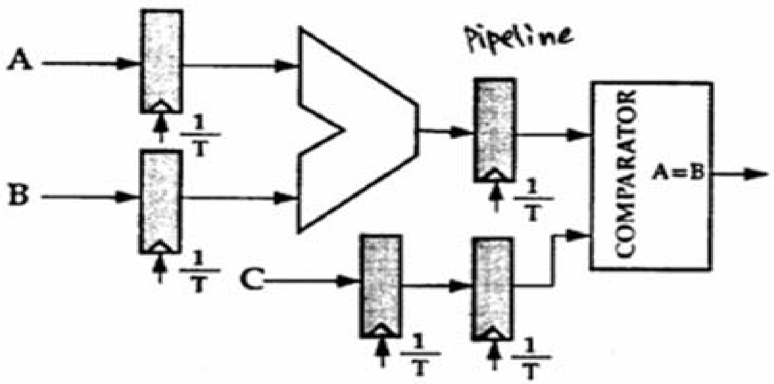
Area = 1476 x 1219 μ^2

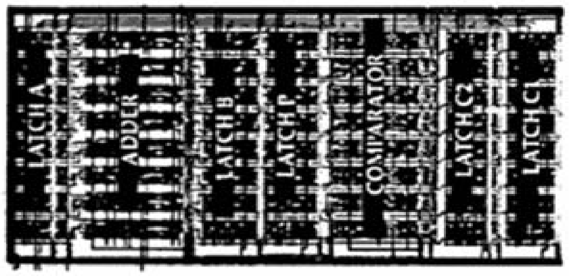
- The clock rate can be reduced by half with the same throughput ⇒ $f_{par} = f_{ref} / 2$
- $V_{par} = V_{ref} / 1.7$, $C_{par} = 2.15C_{ref}$
- $P_{par} = (2.15C_{ref})\ (V_{ref}/1.7)^2\ (f_{ref}/2) \approx 0.36\ P_{ref}$

Paralelna implementacija dela datapath :



Pipelined Data Path

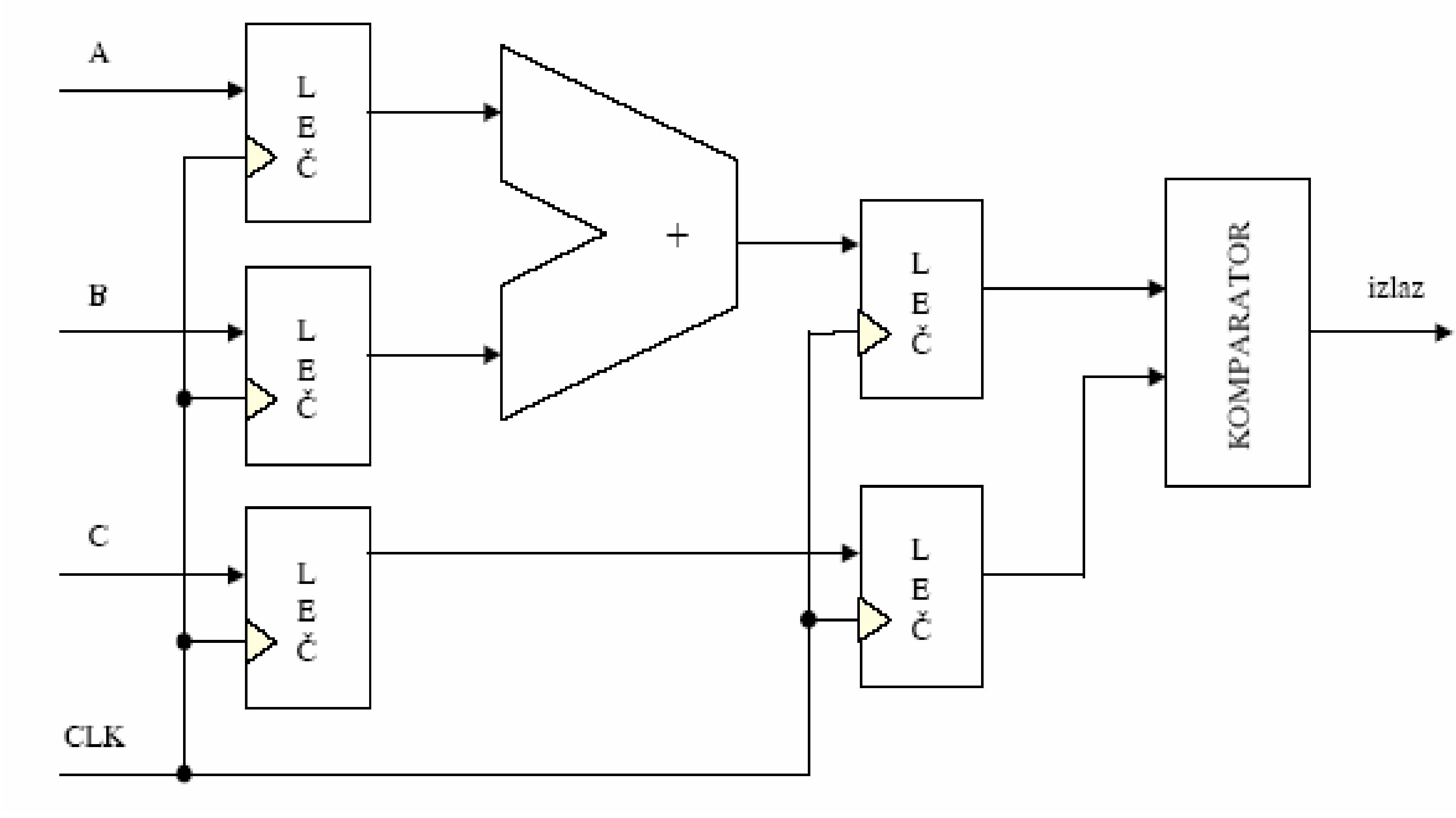




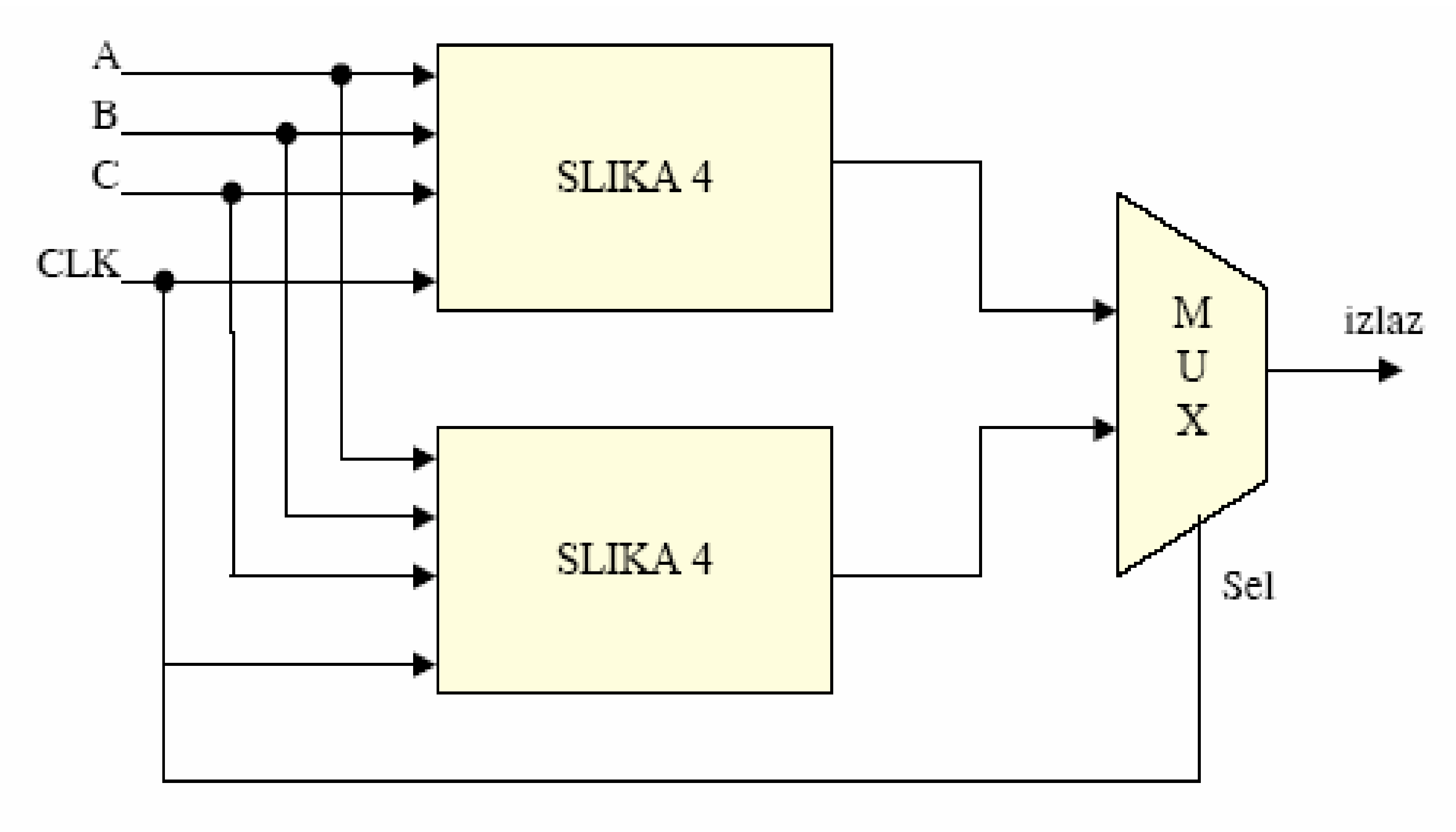
Area = 640 x 1081 u^2

- $f_{pipe} = f_{ref}$
  $C_{pipe} = 1.1C_{ref}$
  $V_{pipe} = V_{ref}/1.7$
- Voltage can be dropped while maintaining the original throughput.
- $P_{pipe} = C_{pipe}\,V_{pipe}^2\,f_{pipe} = (1.1C_{ref})\,(V_{ref}/1.7)^2\,f_{ref} = 0.37\,P_{ref}$

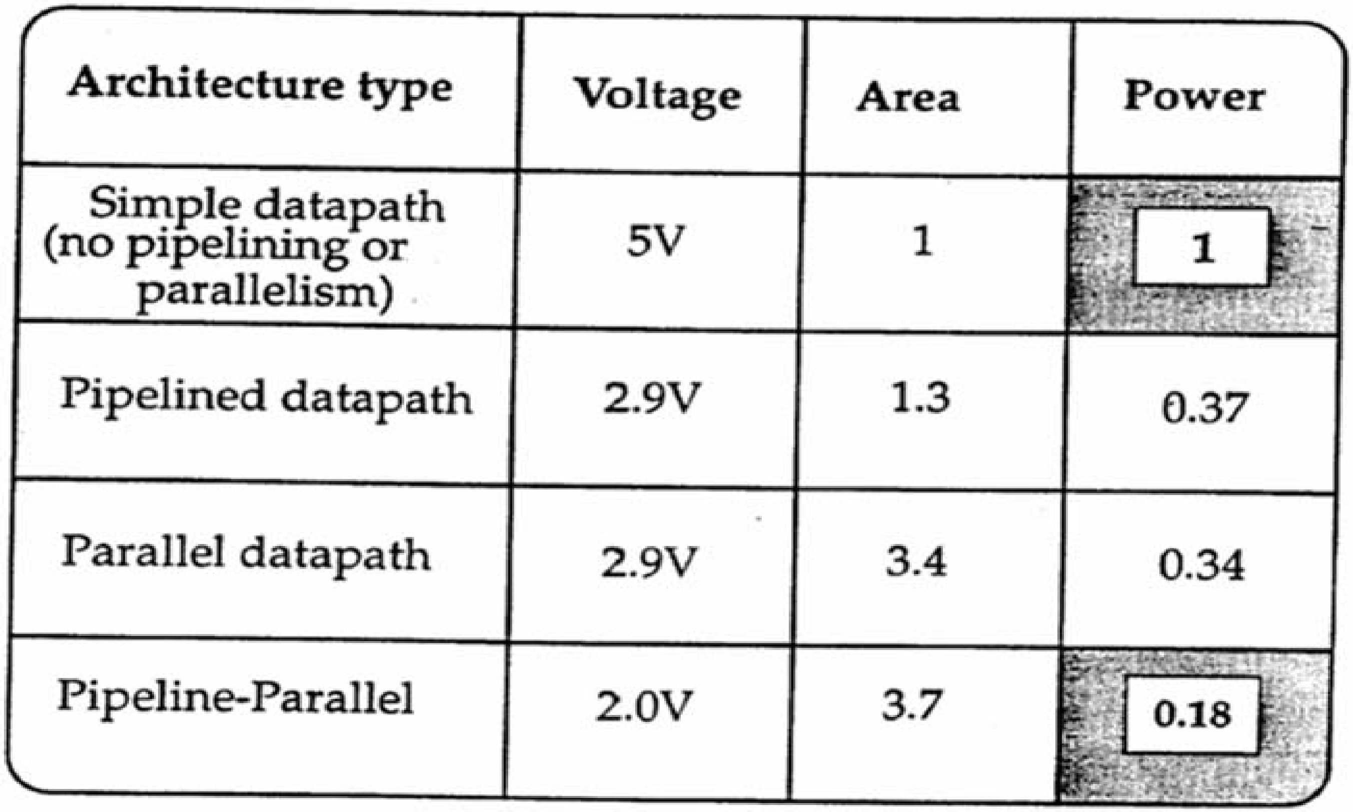
Protočna implementacija:



Paralelno-protočna implementacija:

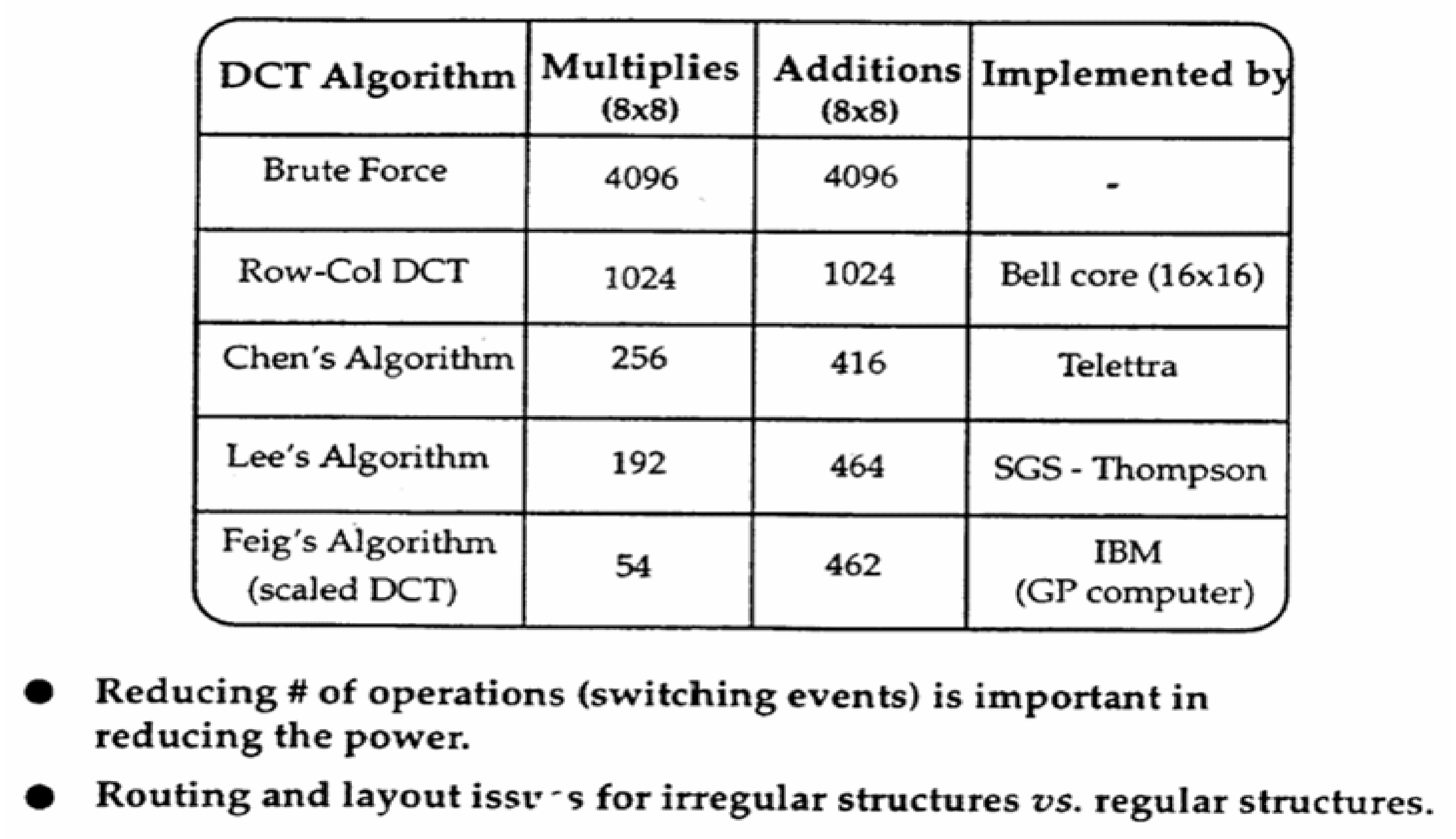


A Simple Data Path : Summary



Architecture type	Voltage	Area	Power
Simple datapath (no pipelining or parallelism)	5V	1	1
Pipelined datapath	2.9V	1.3	0.37
Parallel datapath	2.9V	3.4	0.34
Pipeline-Parallel	2.0V	3.7	0.18

Computational Complexity of DCT Algorithms

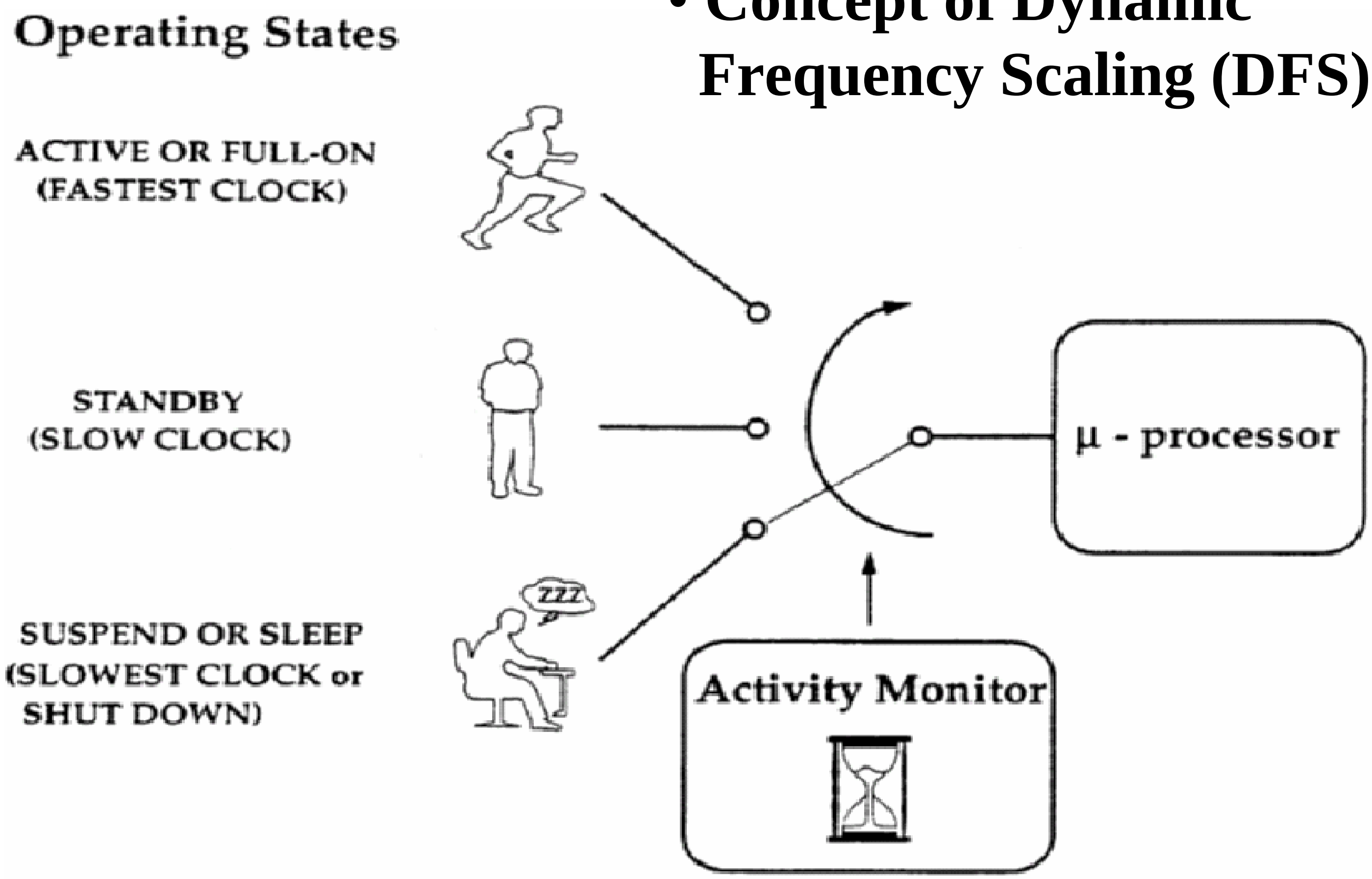


DCT Algorithm	Multiplies (8x8)	Additions (8x8)	Implemented by
Brute Force	4096	4096	-
Row-Col DCT	1024	1024	Bell core (16x16)
Chen's Algorithm	256	416	Telettra
Lee's Algorithm	192	464	SGS - Thompson
Feig's Algorithm (scaled DCT)	54	462	IBM (GP computer)

- **Reducing # of operations (switching events) is important in reducing the power.**
- **Routing and layout issv-s for irregular structures *vs.* regular structures.**

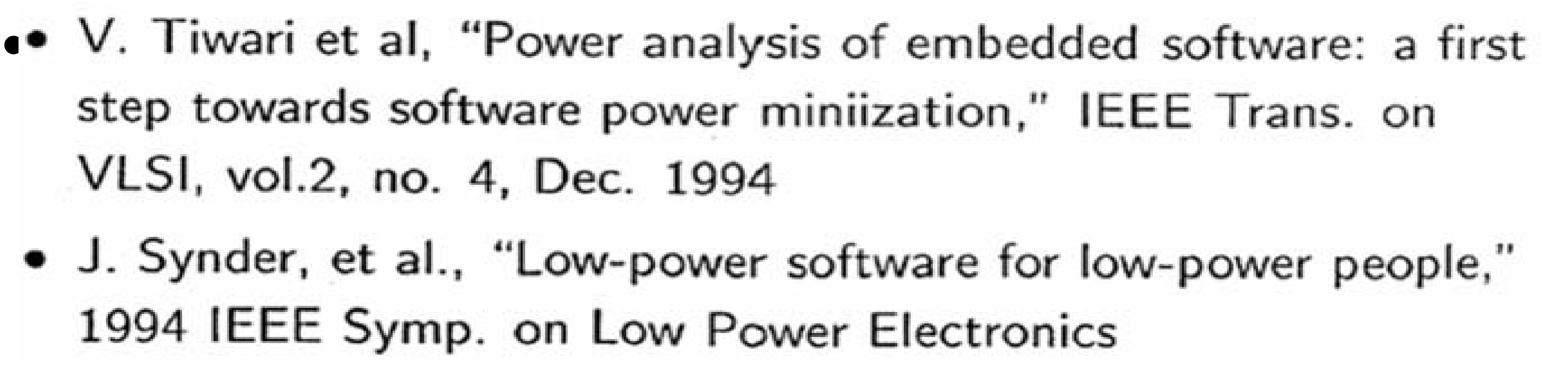
Power Down Techniques

- **Concept of Dynamic Frequency Scaling (DFS)**



Energy-efficient Software Coding

- Potential for power reduction via software modification is relatively unexploited.
- Code size and algorithmic efficiency can significantly affect energy dissipation
- Pipelining at software level- VLIW coding style

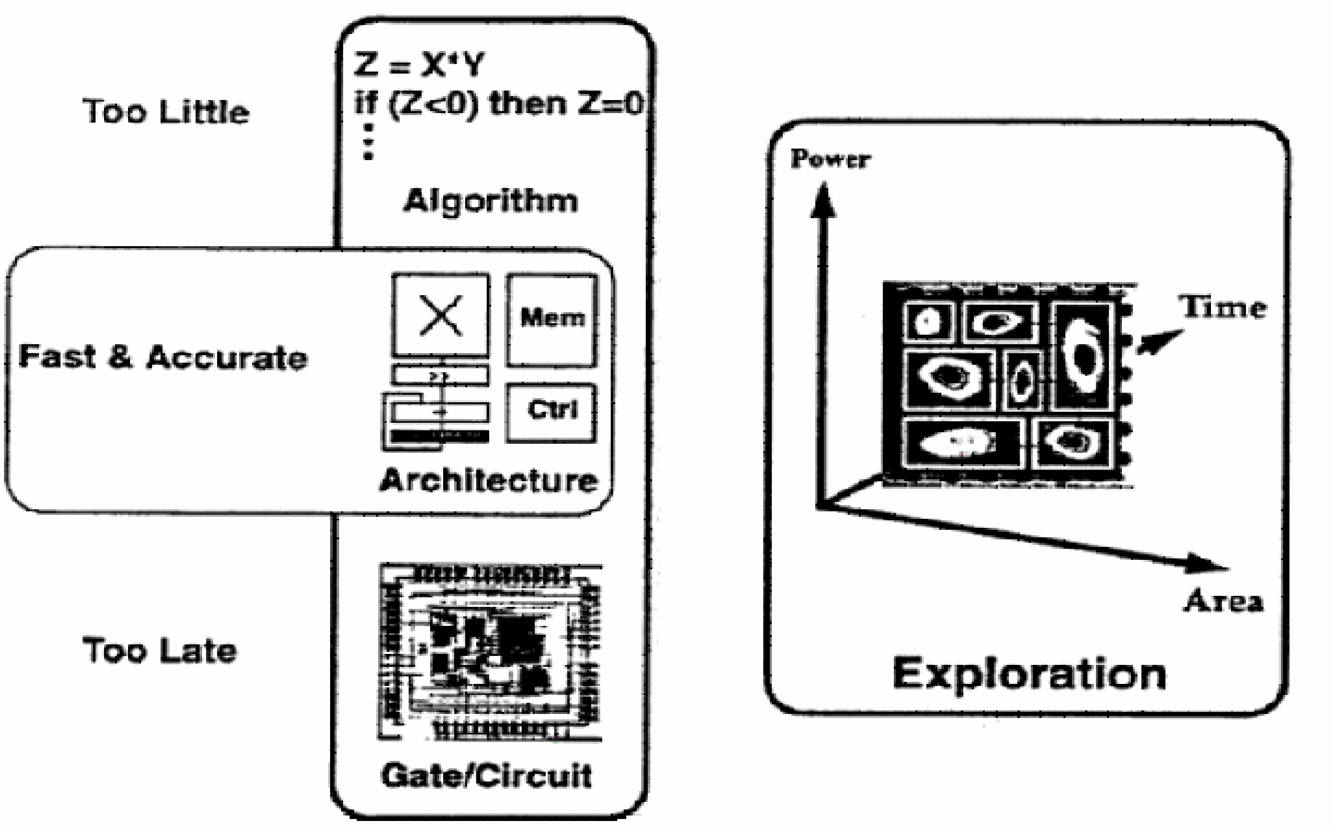


- V. Tiwari et al, “Power analysis of embedded software: a first step towards software power miniization,” IEEE Trans. on VLSI, vol.2, no. 4, Dec. 1994
- J. Synder, et al., “Low-power software for low-power people,” 1994 IEEE Symp. on Low Power Electronics

Power Hunger – Clock Network (Always Ticking)

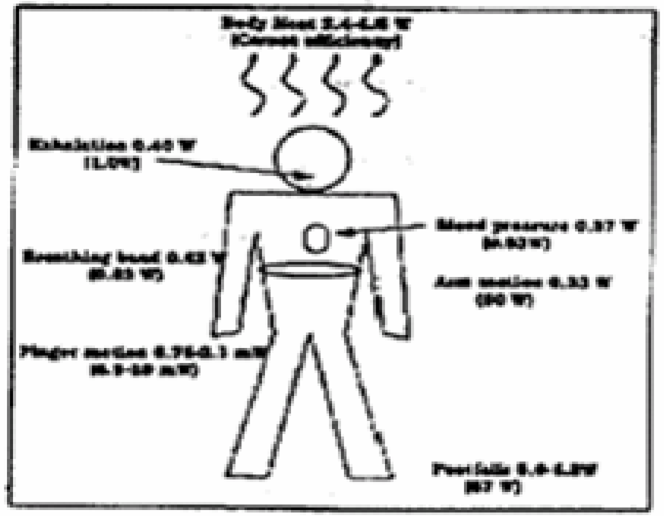
- H-Tree – design deficiencies based on Elmore delay model
- PLL – every designer (digital or analog) should have the knowledge of PLL
    - Multiple frequencies in chips/systems – by PLL
    - Low main frequency, But
    - Jitter and Noise, Gain and Bandwidth, Pull-in and Lock Time, Stability …
- Local time zone
- Self-Timed
- Asynchronous => Use Gated Clocks, Sleep Mode

Power Analysis in the Design Flow



Human Wearable Computing - Power

- Wearable computing – embedding computer into clothing or creating a form that can be used like clothing
- Current computing is limited by battery capacity, output curre t for recharging



Conclusions

- **High-speed design is a requirement for many applications**
- **Low-power design is also a requirement for IC designers.**
- **A new way of THINKING to simultaneously achieve both!!!**
- **Low power impacts in the cost, size, weight, performance, and reliability.**
- **Variable V_{dd} and Vt is a trend**
- **CAD tools high level power estimation and management**
- **Don’t just work on VLSI, pay attention to MEMS – lot of problems and potential is great.**