Techniques for Dynamic Power Reduction

Dynamic Power Reduction

Power Dissipation in CMOS

- Dynamic P.D. → $P_{sw} = C_L V^2_{DD} f_{CLK} \alpha$
- Static P.D. → $P_{stat} = V_{DD} I_{leak}$

- Reducing the device *switching capacitance*, C_L.
- Choosing a lower *power supply*, V_{DD}.
- Reducing the *frequency of operation*, f_{CLK}.
- Reducing the *activity factor*, α.

Technology change

It is brought about by scaling both the device and the process

- Scaling results in reduced device dimensions and in addition,
  - Reduced supply voltage
  - Reduced capacitances
  - Reduced delay
  - Increased leakage due to reduced V_{th}

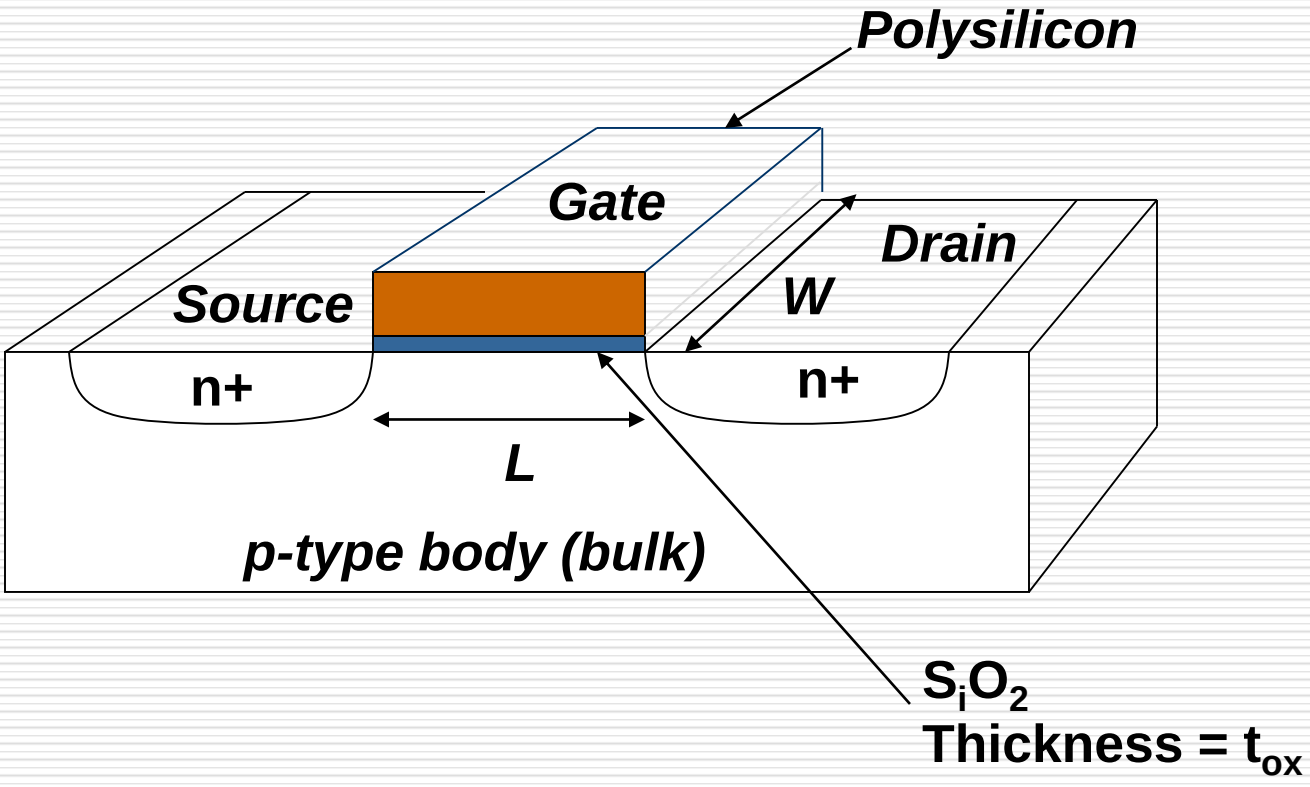
A Simplistic View

- Assume: Dynamic power dominates
  - Power reduces as square of supply voltage; should reduce with device scaling
  - Power reduces linearly with capacitance; should reduce with device scaling
  - Delay is proportional to *RC* time constant; *R* is constant with scaling, *RC* should reduce
- Power reduces with scaling

Device Scaling

1. Constant Field Scaling
2. Constant Voltage Scaling
3. Lateral Scaling

Bulk nMOSFET



Technology Scaling

- A scaling factor (*S*) reduces device dimensions as 1/*S*
- Successive generations of technology have used a scaling ***S* = √2**
- This doubles the number of transistors per unit area.
- This approach produced 0.25μ, 0.18μ, 0.13μ, 90nm and 65nm technologies, continuing on to lower technology nodes
- Constant Voltage scaling : Vdd is kept constant
- Lateral scaling : A 5% gate shrink (*S* = 1.05) is commonly applied to boost speed as the process matures.

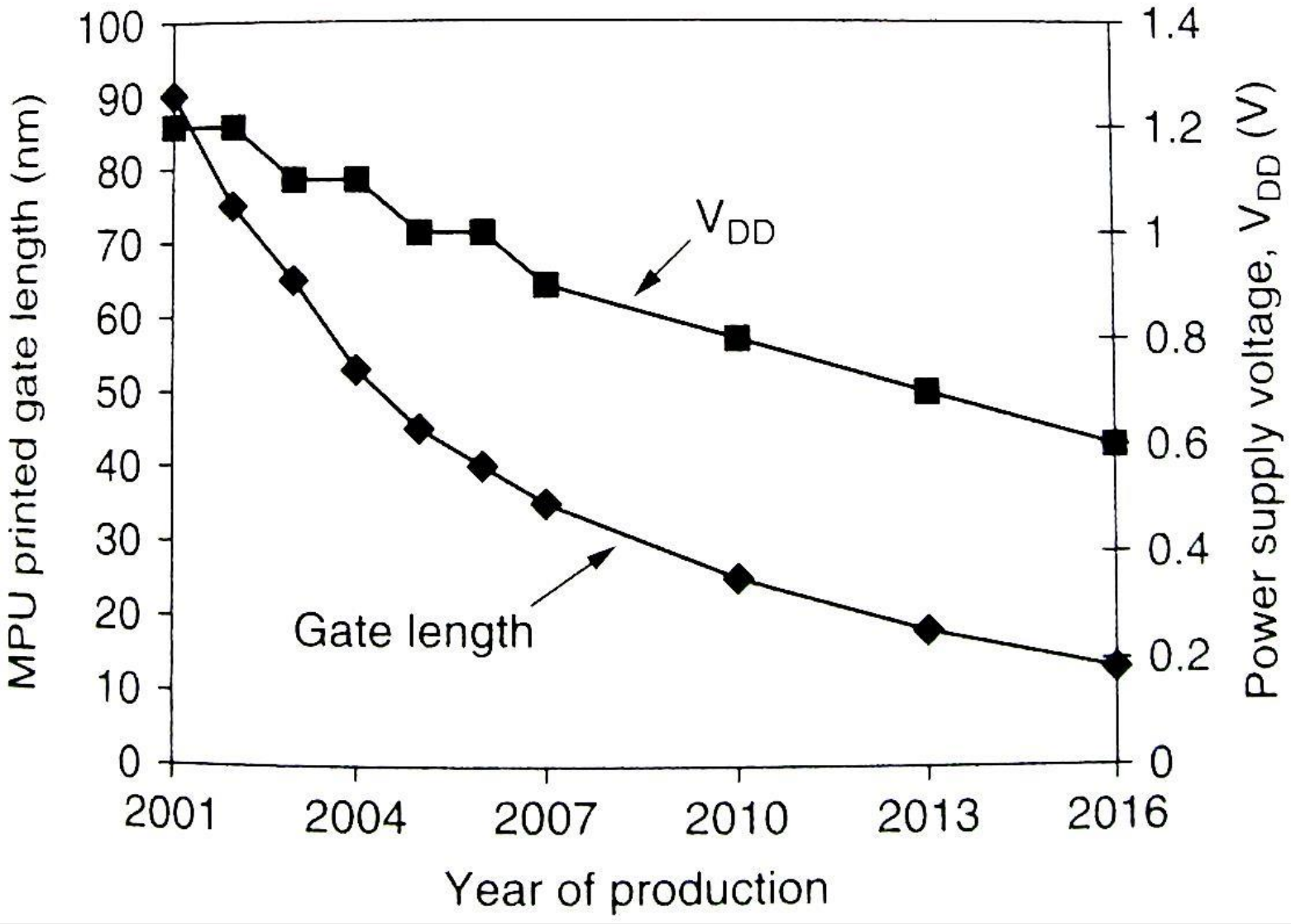
Table 4.12 Influence of scaling on MOS device characteristics

Parameter	Sensitivity	Constant Field	Lateral	Constant Voltage
Scaling Parameters				
Length: L		$1/S$	$1/S$	$1/S$
Width: W		$1/S$	1	$1/S$
Gate oxide thickness: t_{ox}		$1/S$	1	$1/S$
Supply voltage: V_{DD}		$1/S$	1	1
Threshold voltage: V_{tn}, V_{tp}		$1/S$	1	$1/S$
Substrate doping: N_A		S	1	S
Device Characteristics				
β	$\frac{W}{L}\frac{1}{t_{ox}}$	S	S	S
Current: I_{ds}	$\beta(V_{DD} - V_t)^2$	$1/S$	S	$1/S$
Resistance: R	$\frac{V_{DD}}{I_{ds}}$	1	$1/S$	1
Gate capacitance: C	$\frac{WL}{t_{ox}}$	$1/S$	$1/S$	$1/S$
Gate delay: τ	RC	$1/S$	$1/S^2$	$1/S$
Clock frequency: f	$1/\tau$	S	S^2	S
Dynamic power dissipation (per gate): P	CV^2f	$1/S^2$	S	$1/S^2$
Chip area: A		$1/S^2$	1	$1/S^2$
Power density	P/A	1	S	1
Current density	I_{ds}/A	S	S	S

S is generally $\sqrt{2}$

PARAMETER	SCALING MODEL	
	Constant field	Constant voltage
Length (L)	$1/\alpha$	$1/\alpha$
Width (W)	$1/\alpha$	$1/\alpha$
Supply voltage (V)	$1/\alpha$	1
Gate-oxide thickness (t_{ox})	$1/\alpha$	$1/\alpha$
Current ($I = (W/L)(1/t_{ox})V^2$)	$1/\alpha$	α
Transconductance (g_m)	1	α
Junction depth (X_j)	$1/\alpha$	$1/\alpha$
Substrate doping (N_A)	α	α
Electric Field across gate oxide (E)	1	α
Depletion layer thickness (d)	$1/\alpha$	$1/\alpha$
Load Capacitance ($C = WL/t_{ox}$)	$1/\alpha$	$1/\alpha$
Gate Delay (VC/I)	$1/\alpha$	$1/\alpha^2$
	RESULTANT INFLUENCE	
DC power dissipation (P_s)	$1/\alpha^2$	α
Dynamic power dissipation (P_d)	$1/\alpha^2$	α
Power-delay product	$1/\alpha^3$	$1/\alpha$
Gate Area ($A = WL$)	$1/\alpha^2$	$1/\alpha^2$
Power Density (VI/A)	1	α^3
Current Density	α	α^3

Trends of technology and power supply voltage



@ 2002

Voltage Scaling

- Dominant component of power consumption for a properly designed CMOS circuit is proportional to the square of the supply voltage.
- Operating circuits at the lowest possible voltage is the key to minimizing the energy consumed per operation.
- Computational throughput is maintained through appropriate architectural design.

Voltage Scaling Approaches

In deep-submicron technologies, there are many approaches to the selection of optimal power supply voltage:

1. Reliability-Driven Voltage Scaling
2. Technology-Driven Voltage Scaling
3. Energy x Delay Minimum Based Voltage Scaling
4. Voltage Scaling Through Optimal Transistor Sizing
5. Voltage Scaling Using Threshold Reduction
6. Architecture-Driven Voltage Scaling

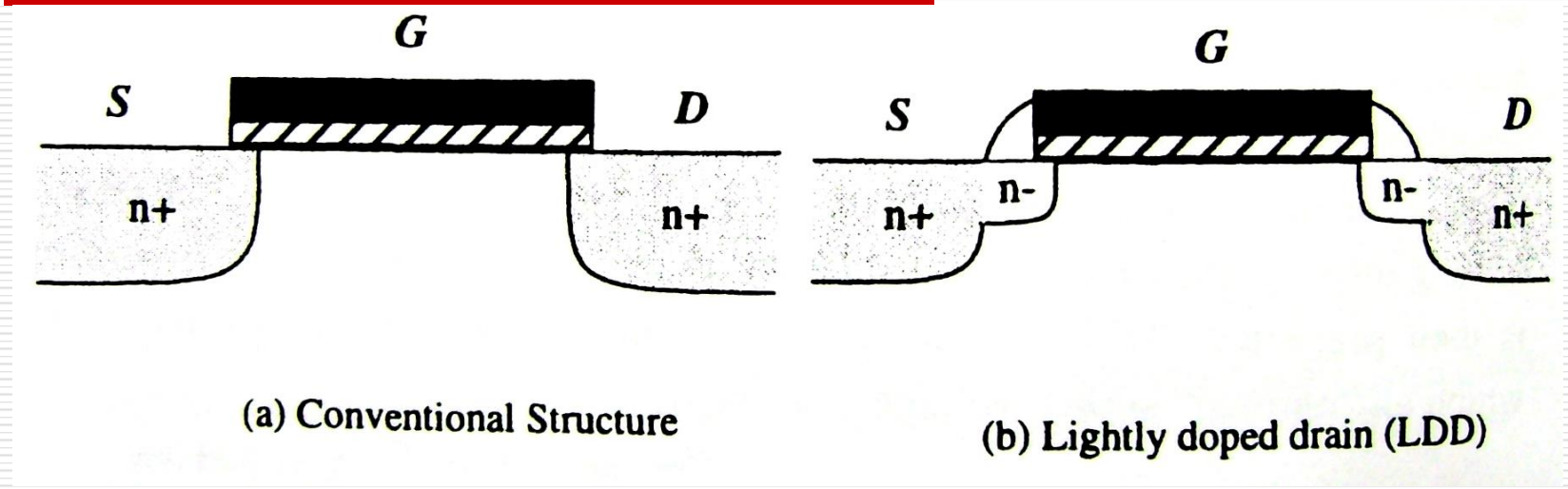
1.Reliability-Driven Voltage Scaling

- Based on optimization of the trade-off between speed, long-term reliability, and power dissipation.
- Constant-voltage scaling results in higher electric fields that create hot carriers (short-channel effect).
- Result – the devices degrade with time (V_{TH} change, g_m degradation, increase in subthreshold current) leading to eventual breakdown.

Reliability-Driven Voltage Scaling

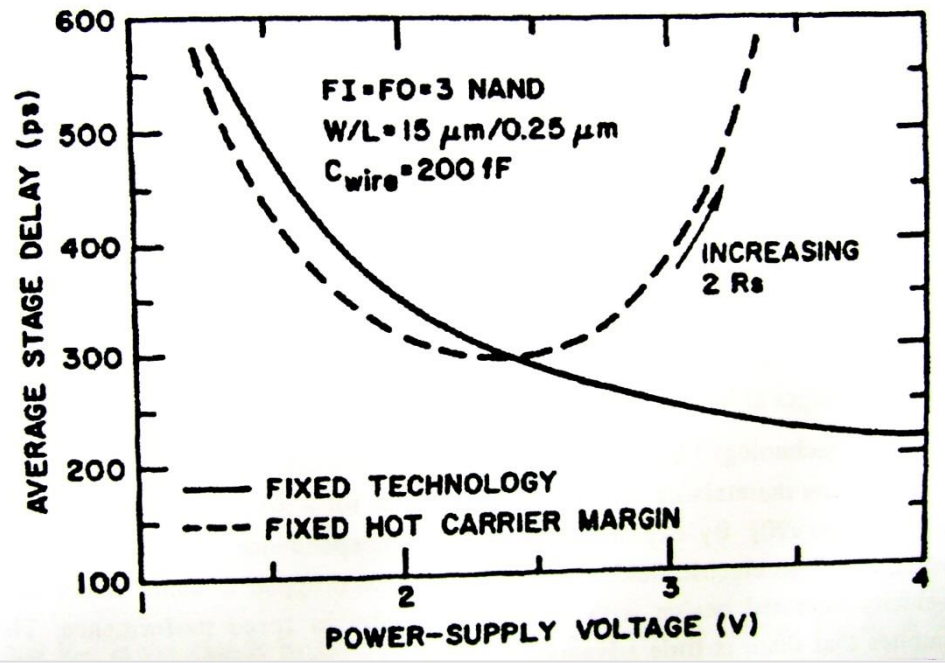
- One approach to reduce the number of hot carriers is to change the device structure to reduce the fields inside or to limit the high field regions so that carriers do not gain enough energy from the field which may cause problems.
- Since high fields occur only near the drain, the potential gradient is made more gradual at the drain side by introducing Lightly Doped Drain area (LDD) between the heavily doped n+ region and the channel.

Drain doping profile for a conventional device and an LDD device:



- LDD improves the reliability but also increases the series resistance of the drain and source which results in reduced performance.
- So, in this approach, power supply voltage selection is based on a trade-off between performance and reliability.

Reliability-driven supply voltage selection:



- The delay starts to increase with increasing supply voltage since the parasitic resistance of the LDD structure must be increased to maintain the CHC margin and this limits the circuit performance.

2. Technology-Driven Voltage Scaling

For long channel devices: $I = \frac{1}{2} \mu_n C_{OX} W/L(V_{dd} - V_t)^2$;

$$\text{Delay} \propto C_L.R_C = C_L.(V_{dd}/I)$$

Technology-Driven Voltage Scaling

- But when the device shrinks below 1.0μ, because of velocity saturation, the current is no longer a quadratic function of the voltage but linear; hence, the current drive is significantly reduced and is given by,

$$I \propto C_{OX}\, W\, (V_{dd} - V_t)\, v_{max};$$

$$\text{Delay} \propto C_L.R_C = C_L.(V_{dd}/I)$$

i.e, delay for submicron circuits is relatively independent of supply voltages at high electric fields.

Technology-Driven Voltage Scaling

- In this scaling approach, the power supply voltage is chosen based on maintaining the speed-performance for a given submicron technology.
- By exploiting the relative independence of delay on supply voltage at high electric fields, the voltage can be dropped to some extent for a velocity-saturated device with little penalty on speed performance.
- In other words, if performance advantage we get by increasing supply voltage is not significant, why not save power by reducing Vdd that compromises the performance a little bit?

Technology-Driven Voltage Scaling

$$I_D = \frac{W_{eff}\mu_{eff}C_{ox}\left(V'_G - \frac{V_{DS}}{2}\right)V_{DS}}{L_{eff}\left(1 + \frac{V_{DS}}{E_C L_{eff}}\right)} \qquad (V_{DS} \le V_{DSAT})$$

$$I_D = \frac{W_{eff}\mu_{eff}C_{ox}\left(V'_G - V_{DSAT}\right)E_C}{2} \qquad (V_{DS} > V_{DSAT})$$

Where,

$$V_{DSAT} = \frac{E_C L_{eff} V'_G}{E_C L_{eff} + V'_G};$$

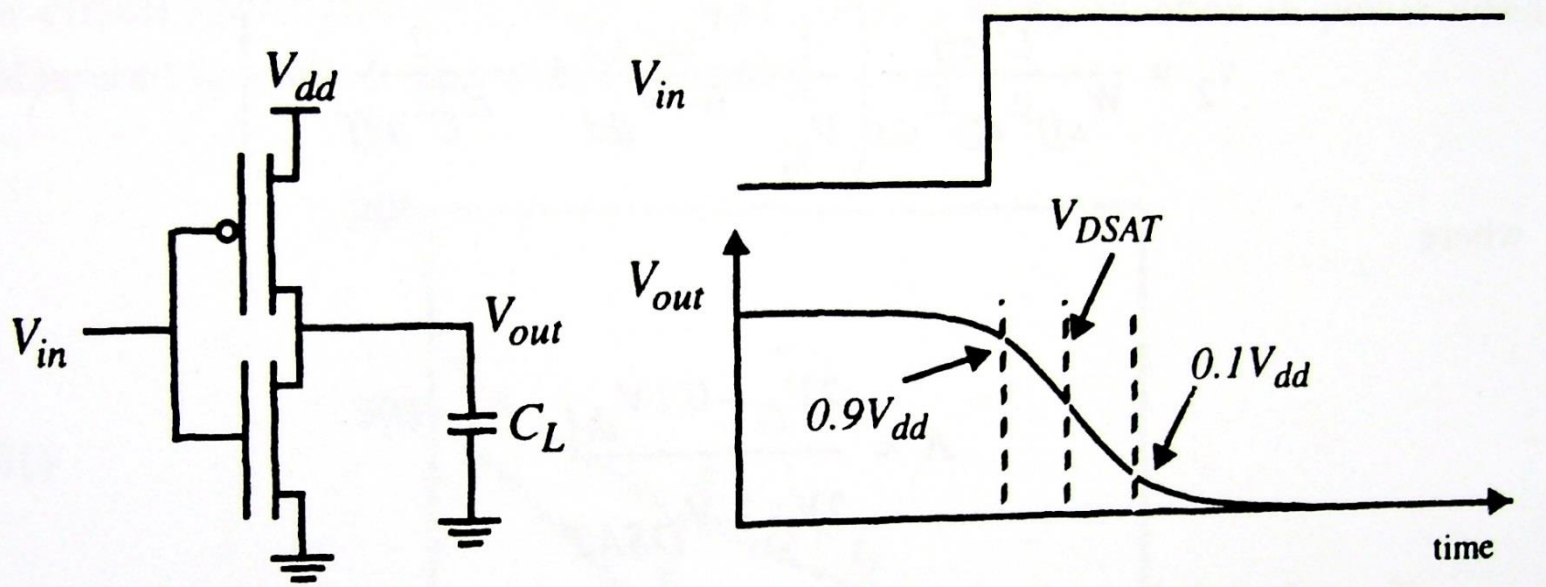
$V'_G = (V_{GS} - V_t)$;

$E_C = 2v_{sat}/\mu_{eff}$

= critical electrical field at which the carrier velocity is saturated.

Technology-Driven Voltage Scaling

Model for computing the delay and "critical voltage" [Kakumu90], which provides a lower limit on the supply voltage.



Model for computing delay and critical voltage.

Technology-Driven Voltage Scaling

- Delay of the circuit is the time taken for the output to transition from $0.9V_{dd}$ to $0.1V_{dd}$.
- For this range, the NMOS transistor goes through two regions: from $0.9V_{dd}$ to V_{DSAT} (saturation) and from V_{DSAT} to $0.1V_{dd}$ (linear). Let these times be defined as τ_1 and τ_2.

The delays are then determined as,

$$\tau_1 = \frac{2C_L\left(0.9V_{dd} - V_{DSAT}\right)}{W_{eff}\mu_{eff}C_{ox}\left(V_G' - V_{DSAT}\right)E_C}$$

Technology-Driven Voltage Scaling

$$\tau_2 = \frac{C_L L_{eff}}{W_{eff}\mu_{eff}C_{ox}}\left(\frac{1}{V_G'}\ln\frac{V_{DSAT}}{0.1V_{dd}}A + \frac{2}{E_C L_{eff}}\ln A\right)$$

Where,

$$A = \frac{2V_G' - 0.1V_{dd}}{2V_G' - V_{DSAT}}$$

Technology-Driven Voltage Scaling

- To understand the two delay terms some simplifications can be made. For τ_1, it can be assumed that,

$$(0.9V_{dd} - V_{DSAT})/(V'_G - V_{DSAT}) \approx 1.$$

- This implies, τ_1 is to first order independent of the power supply voltage.
- For τ_2, the second term in the bracket is negligible and therefore τ_2 can be approximated as:

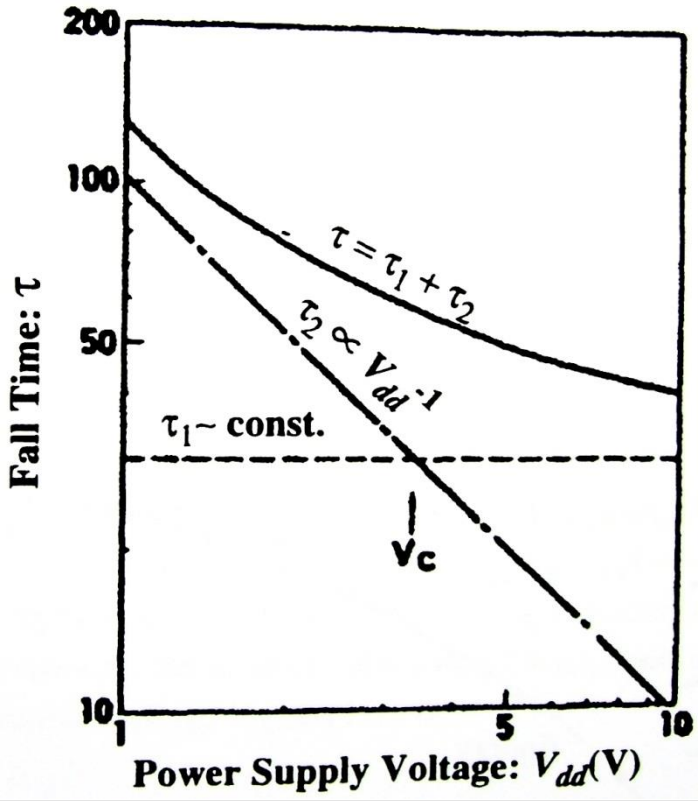
$$\tau_2 \approx \frac{C_L L_{eff}}{W_{eff}\mu_{eff}C_{ox}V'_G} \ln \frac{V_{DSAT}}{0.1V_{dd}} A$$

Technology-Driven Voltage Scaling

- Since the ln term is nearly constant, it can be seen that τ_2 is inversely proportional to V_{dd}.

- Therefore, the total delay = $\tau_1 + \tau_2$ contains a term which is independent of supply voltage and one term which is dependent on supply voltage.

The plot of the fall time as a function of the power supply voltage on a log-log plot is shown below.



When V_{dd} becomes larger than the critical voltage, V_C (the voltage at which $\tau_1 = \tau_2$), the delay is dominated by τ_1 and does not improve significantly with an increase in voltage and hence regard V_C as the lower limit on supply voltage.

3. Energy x Delay Minimum Based Voltage Scaling: [Burr91]

- This approach involves minimizing the *energy x delay* product.
- For a fixed technology (both feature size and threshold voltage), there is a supply voltage that optimizes the quadratically reduced energy per computation (due to a lower supply voltage) and the increased circuit delays.
- This "optimum" supply voltage can be computed analytically assuming that the transistors operate mostly in the saturation region.

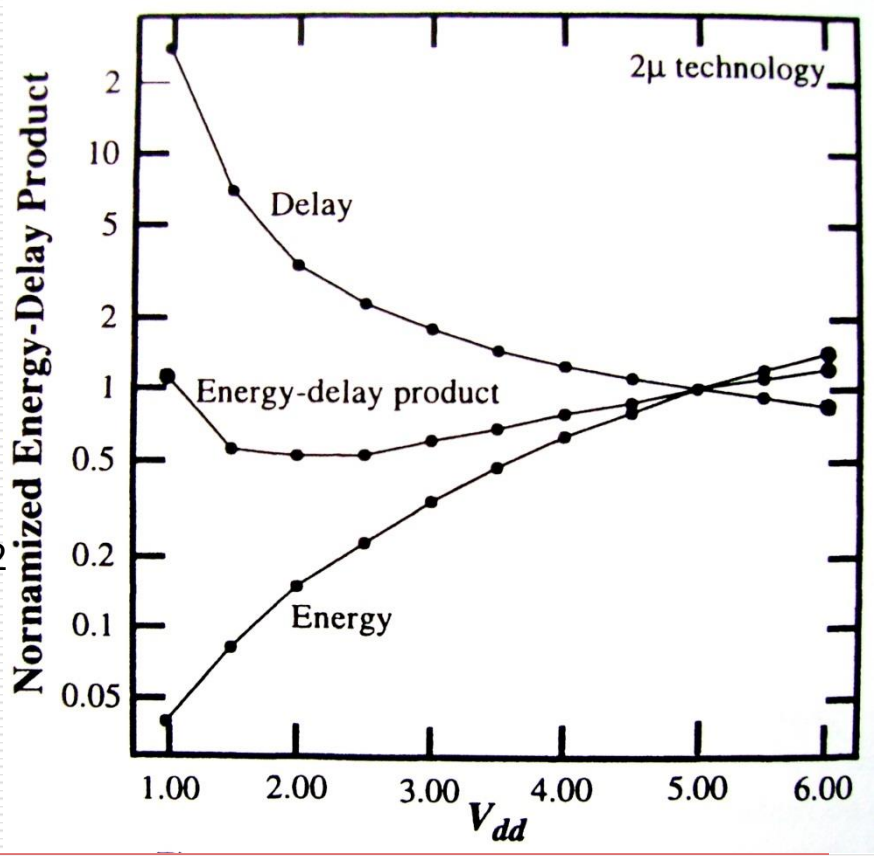
Energy x Delay Minimum Based Voltage Scaling

Ex: Ring Oscillator Circuit

Energy = $k1.C_L.V^2_{dd}$

$$T_d = \frac{C_L \times V_{dd}}{I} = \frac{C_L \times V_{dd}}{\frac{\mu C_{ox}}{2}(W/L)\left(V_{dd} - V_t\right)^2}$$

Energy.T_d = K $C_L V^2_{dd} C_L V_{dd}/(V_{dd}-V_t)^2$



Energy x Delay Minimum Based Voltage Scaling

- Taking the derivative of the *energy x* T_d and solving for the supply voltage for minimum of this product gives optimal value of Vdd in this strategy

i.e, $d/dV_{dd}\,[Energy.T_d] = d/dV_{dd}\,[K\,C_L V^2_{dd} V_{dd}/(V_{dd}-V_t)^2] = 0$

Solving this we get, $V_{dd} = 3V_t$

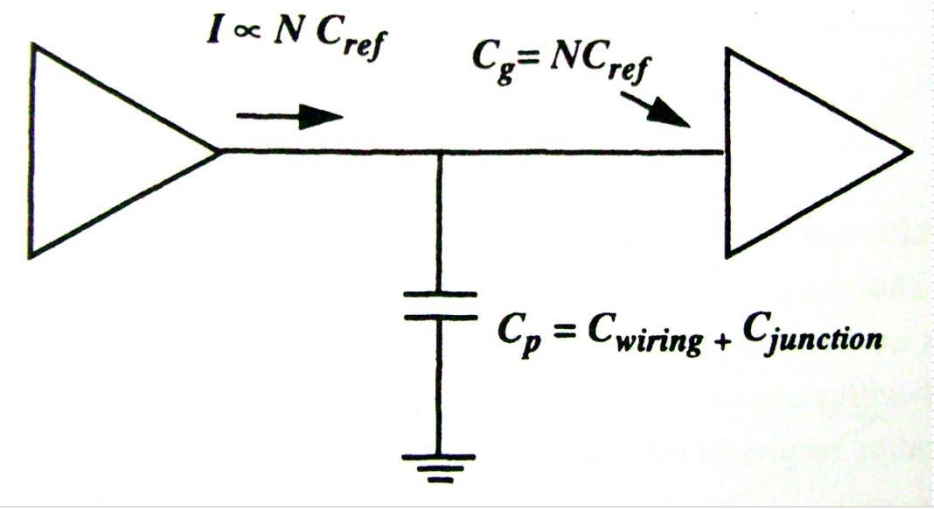
- In this case, power consumption is lowered at the expense of a reduced computational throughput.
- Not desirable for throughput constrained functions.

4. Voltage Scaling Through Optimal Transistor Sizing:

- Optimized transistor sizing will play an important role in reducing power consumption.
- Equalize all delay paths so that a single critical path does not unnecessarily limit the performance of the entire circuit.
- The W/L ratios are uniformly raised for all the devices, yielding a uniform decrease in the gate delay and hence allowing for a corresponding reduction in voltage and power.
- If voltage is allowed to vary, the optimal sizing for low power is quite different from that required for high speed operation.

Voltage Scaling Through Optimal Transistor Sizing:

- First stage is driving $\mathbf{C_g}$ (gate capacitance of the second) **+** $\mathbf{C_p}$ (substrate coupling and interconnect)
- $\boldsymbol{C_g = NC_{ref}}$, for both stages, where C_{ref} is the gate capacitance of a unit transistor.



Circuit model for analyzing the effect of transistor sizing

Voltage Scaling Through Optimal Transistor Sizing:

- Delay through the first gate at a supply voltage $\mathbf{V_{ref}}$ is given by:

$$T_d = \frac{C_L \times V_{dd}}{I} = \frac{C_L \times V_{dd}}{\frac{\mu C_{ox}}{2}(W/L)\left(V_{dd} - V_t\right)^2}$$

$$T_N = K\frac{\left(C_p + NC_{ref}\right)}{\left(NC_{ref}\right)}\frac{V_{ref}}{\left(V_{ref} - V_t\right)^2} = K(1 + \beta/N)\frac{V_{ref}}{\left(V_{ref} - V_t\right)^2}$$

$\beta = C_p/C_{ref}$; Input gate capacitance of both stages = $N.C_{ref}$

K = terms independent of device width and voltage.

Voltage Scaling Through Optimal Transistor Sizing:

- For a given supply voltage V_{ref}, the speed up of a circuit whose W/L ratios are sized up by a factor of N over a reference circuit using unit transistors (N=1) is given by:

$$T_N/T_1 = \mathbf{(1+\beta/N)/(1+\beta)}$$

- In order to evaluate the energy performance of the two designs at the same speed, the voltage of the scaled design is allowed to vary so as to keep the delay constant.

- Assuming that the delay scales as $1/V_{dd}$ (ignoring V_t) the supply voltage, V_N, where the delay of the scaled design and the reference design are equal is given by:

$$V_N = \frac{(1+\beta/N)}{(1+\beta)} V_{ref}$$

Unit Transistor

$N = 1$

$Vdd = V_{ref}$

$T_1 = (1+ß)/V_{ref}$

Sized up transistor

$N = N$

$Vdd = V_N$

$T_N = (1+ß/N)/V_N$

Equalizing the delays we have,

$T_1 = (1+ß)/V_{ref} = T_N = (1+ß/N)/V_N$

Hence $V_N = \frac{(1+ß/N)}{(1+ß)} V_{ref}$

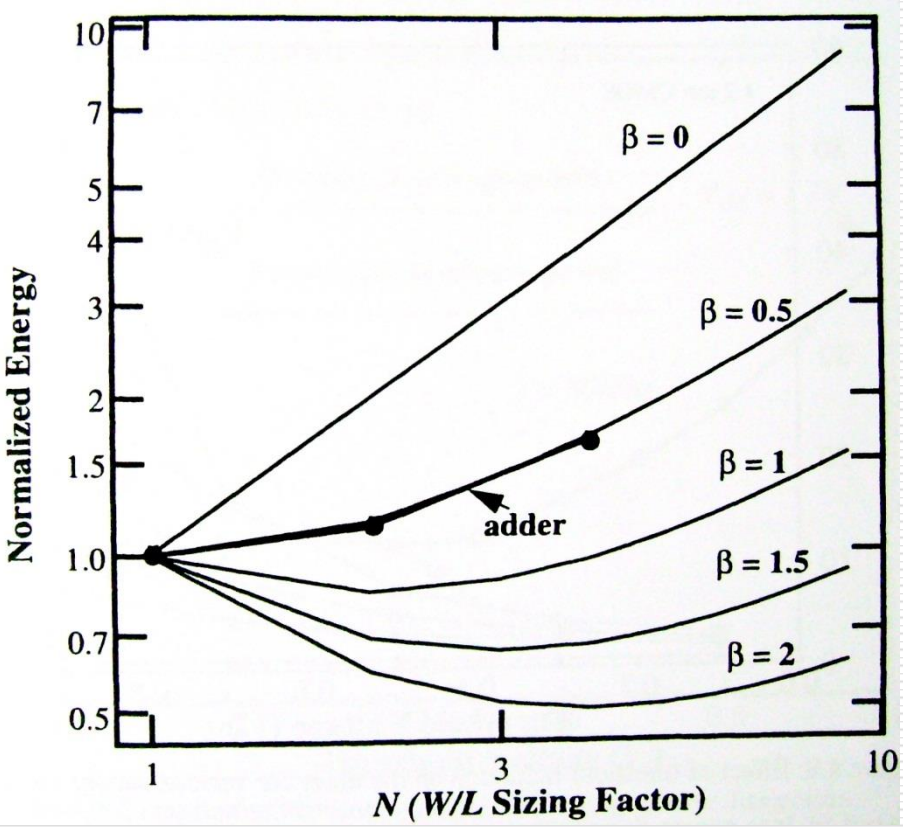
Voltage Scaling Through Optimal Transistor Sizing:

- Under these conditions, the energy consumed by the sized up design as a function of N is given by:

$$Energy(N) = \left(C_p + NC_{ref}\right)V_N^2 = \frac{NC_{ref}(1+\beta/N)^3 V_{ref}^2}{(1+\beta)^2}$$

Voltage Scaling Through Optimal Transistor Sizing:

After normalization (Energy(N)/Energy(1)), the plot is given below:



Voltage Scaling Through Optimal Transistor Sizing:

- In this analysis, it is assumed that the parasitic capacitance is independent of device sizing.
- When there is no parasitic capacitance contribution (i.e, $\beta=0$), the energy increases with N, and the solution utilizing devices with the smallest W/L ratios results in the lowest energy.
- At high values of β, when parasitic capacitances begin to dominate over the gate capacitances, the energy decreases temporarily with increasing device sizes and then starts to increase, resulting in an optimal value for N.

Voltage Scaling Through Optimal Transistor Sizing: Summary

- Td α 1/ ((W/L).Vdd)
- If W/L is sized up by N, Vdd can be scaled down to save power for the same delay Td
- For a sized up gate, the new

$$V_N = \frac{(1+\beta/N)}{(1+\beta)} V_{ref}$$

$$Energy(N) = \left(C_p + NC_{ref}\right)V_N^2 = \frac{NC_{ref}(1+\beta/N)^3 V_{ref}^2}{(1+\beta)^2}$$

- Find Energy (1) for a reference(unsized) gate
- Plot Energy(N)/Energy(1) wrt sizing factror N for various β (i.e. Cp/Cref)

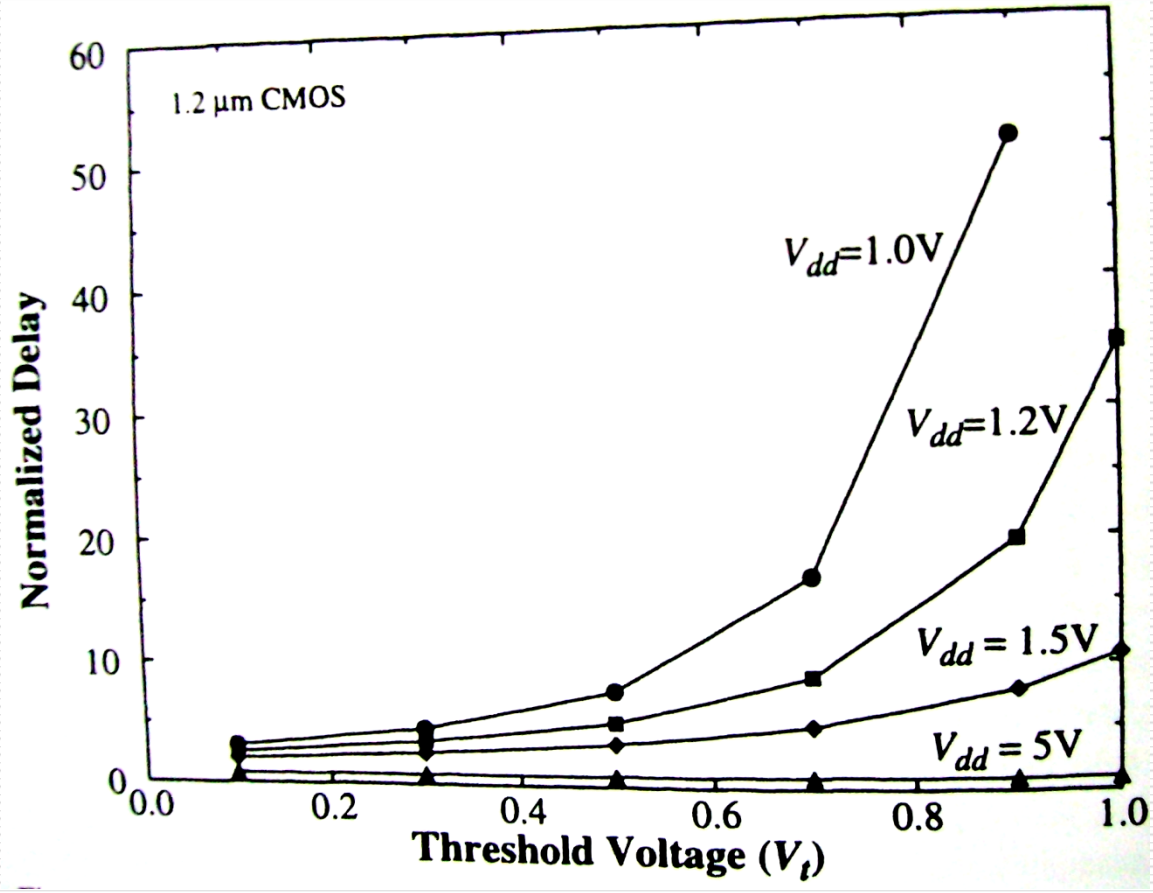
5. Voltage Scaling using Threshold Reduction

- Reducing the threshold voltage of the device allows the supply voltage to be scaled down (and hence lower switching power) without loss in speed.

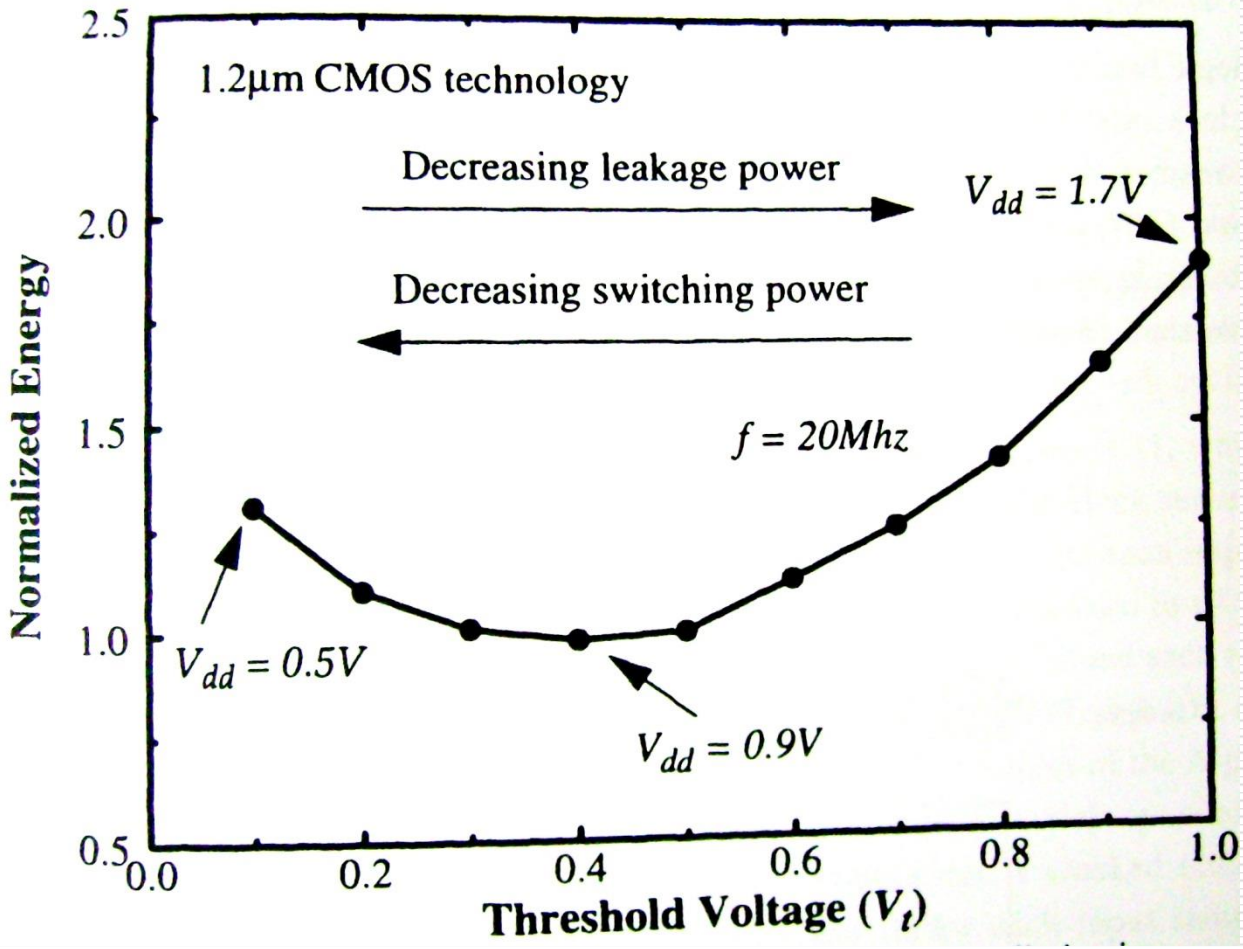
Ex: A circuit with V_{dd}=1.5V, V_t=1V will have approx. the same performance as the circuit with V_{dd}=1V, V_t=0.5V

- The limit is set by the requirement to retain adequate noise margins and the increase in the subthreshold currents.

Voltage Scaling using Threshold Reduction



Plot of energy vs. threshold voltage for a fixed throughput for a 16-bit datapath ripple carry adder:



6. Architecture-Driven Voltage Scaling

- The “reliability” and “technology based” approaches discussed earlier are focused on reducing the voltage while maintaining the device speed, and do not attempt to achieve the minimum possible power.
- In this approach, the architecture is modified to compensate for the reduced circuit speed that comes with operating even below the “critical voltage”.
- Two types:
  - Trading Area for Lower Power Through Hardware Duplication.
  - Trading Area for Lower Power Through Hardware Pipelining.

1. Trading Area for Lower Power Through Hardware Duplication

Ex: A conventional uni-processor implementation of some logic function F(IN):

Let f_{sample} (= $1/T_{sample}$) be the throughput clock rate

$$C_{ref} = C_{in\text{-}reg} + C_F + C_{out\text{-}reg}$$

$$P_{ref} = C_{ref} V^2_{ref} f_{sample}$$

Timing Diagram

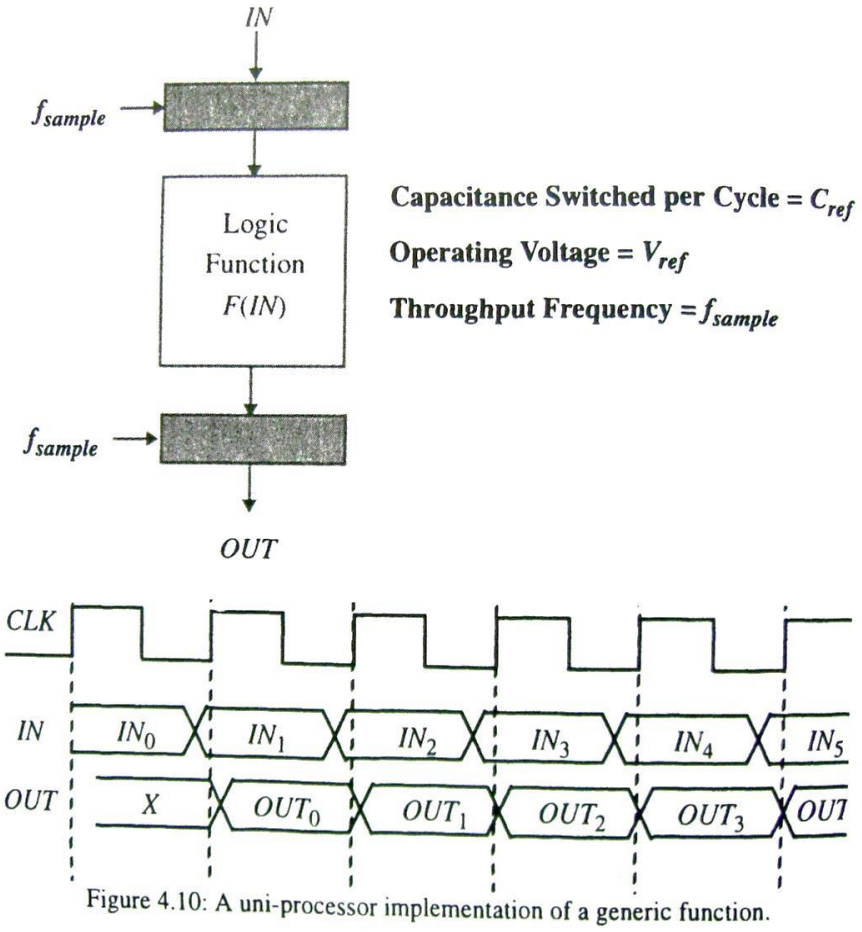
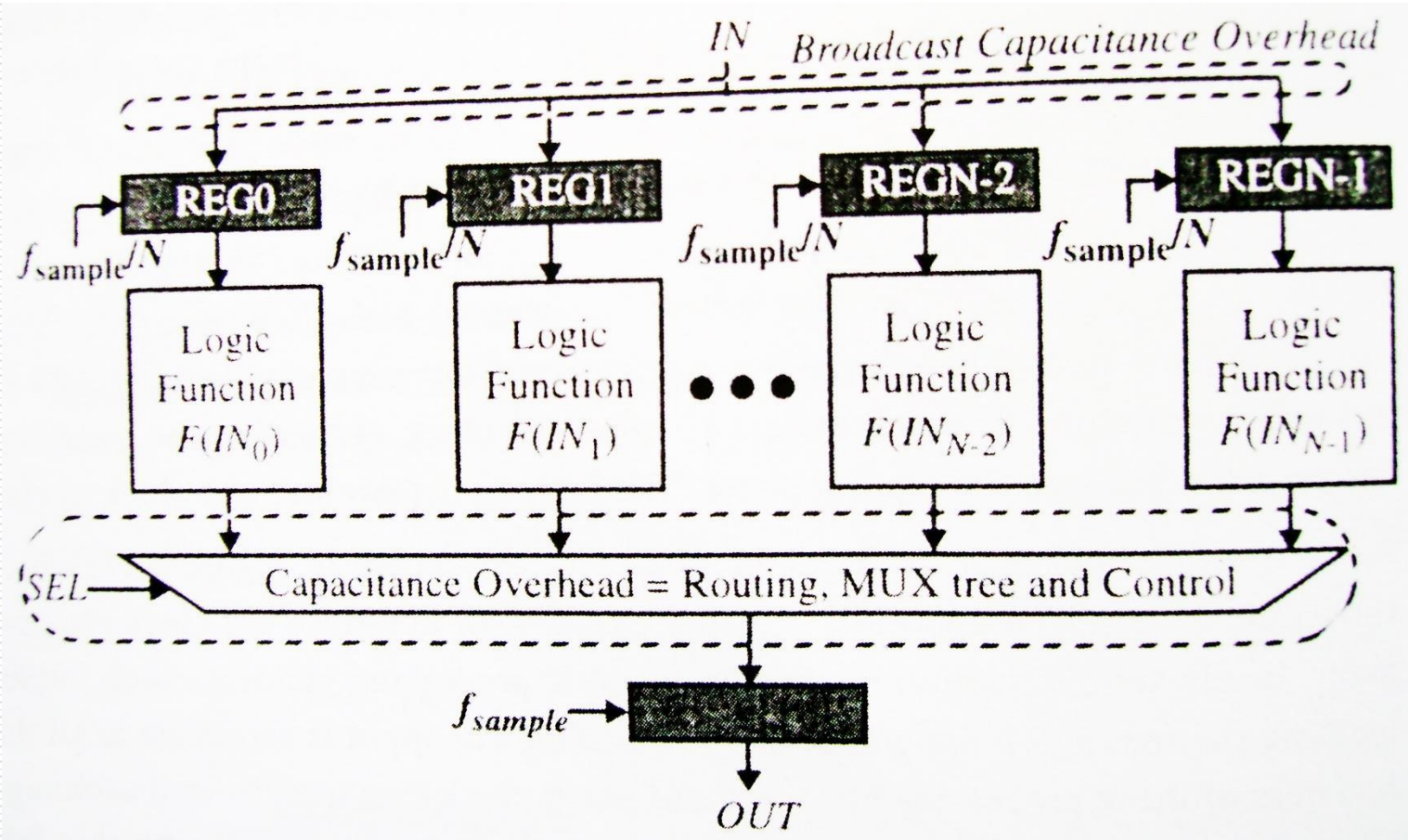


Figure 4.10: A uni-processor implementation of a generic function.

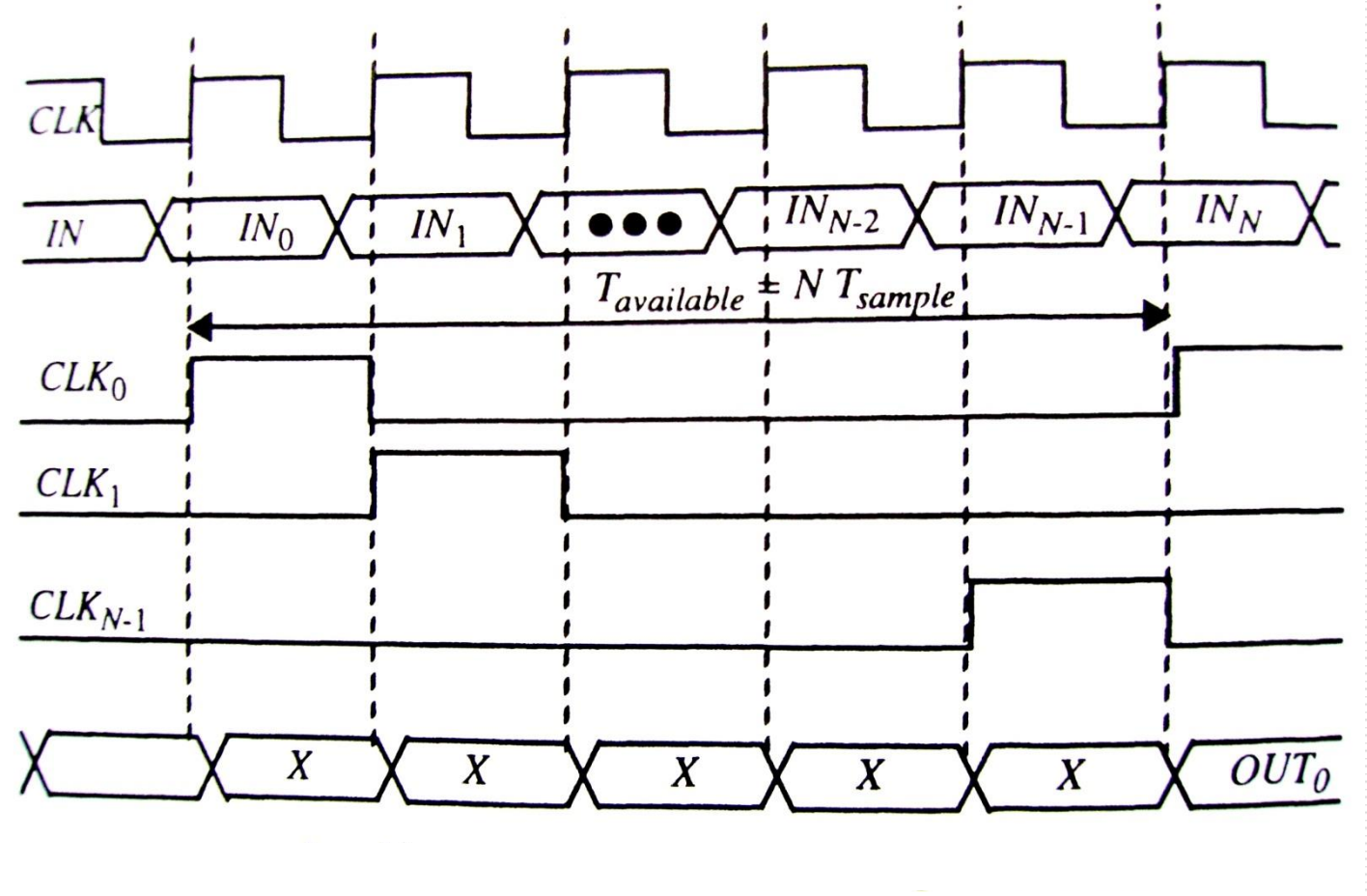
A multi-processor implementation of a generic function F(IN)

- Parallelism to reduce the clock requirement
- Hardware duplication – N processing elements
- Input is broadcast to all the registers
- The outputs of the N elements are multiplexed and sent to an output register which operates at f_{sample}
- Time available to compute the function F for each input sample, $T_{available} = N/f_{sample} = N.T_{sample}$
- Supply voltage V_{ref} can be lowered to V_N , to increase the circuit delays until the critical path is extended to become equal to the new, slower clock cycle

A multi-processor implementation of a generic function F(IN)



Timing Diagram



Determination of V_N

- The power supply voltage corresponding to N-level parallelism, V_N, is determined below:

$$T_N = K\frac{V_N}{(V_N - V_t)^2} = N \bullet T_1 = N \bullet K\frac{V_{ref}}{(V_{ref} - V_t)^2}$$

$$T_d = \frac{C_L \times V_{dd}}{I} = \frac{C_L \times V_{dd}}{\frac{\mu C_{ox}(W/L)}{2}(V_{dd} - V_t)^2}$$

  - T_N -> the critical path delay for each computation module of the N processor implementation
  - T_1 -> critical path delay for the uni-processor.

Determination of V_N

- The ratio of V_N to V_{ref} is given by:

$$\frac{V_N}{V_{ref}} = N \bullet \frac{\left(V_N - V_t\right)^2}{\left(V_{ref} - V_t\right)^2} = N \bullet \frac{\left(\frac{V_N}{V_{ref}} - \frac{V_t}{V_{ref}}\right)^2}{\left(1 - \frac{V_t}{V_{ref}}\right)^2} = N \bullet \frac{\left(\frac{V_N}{V_{ref}} - \beta\right)^2}{(1-\beta)^2}$$

where,

$$\beta = \frac{V_t}{V_{ref}}$$

Determination of V_N

The quadratic relationship of above equation can be solved to obtain V_N as a function of V_{ref} and V_t.

$$\frac{V_N}{V_{ref}} = \frac{2\beta + \gamma + \sqrt{4\beta\gamma + \gamma^2}}{2}$$

$$ax^2 + bx + c = 0,$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

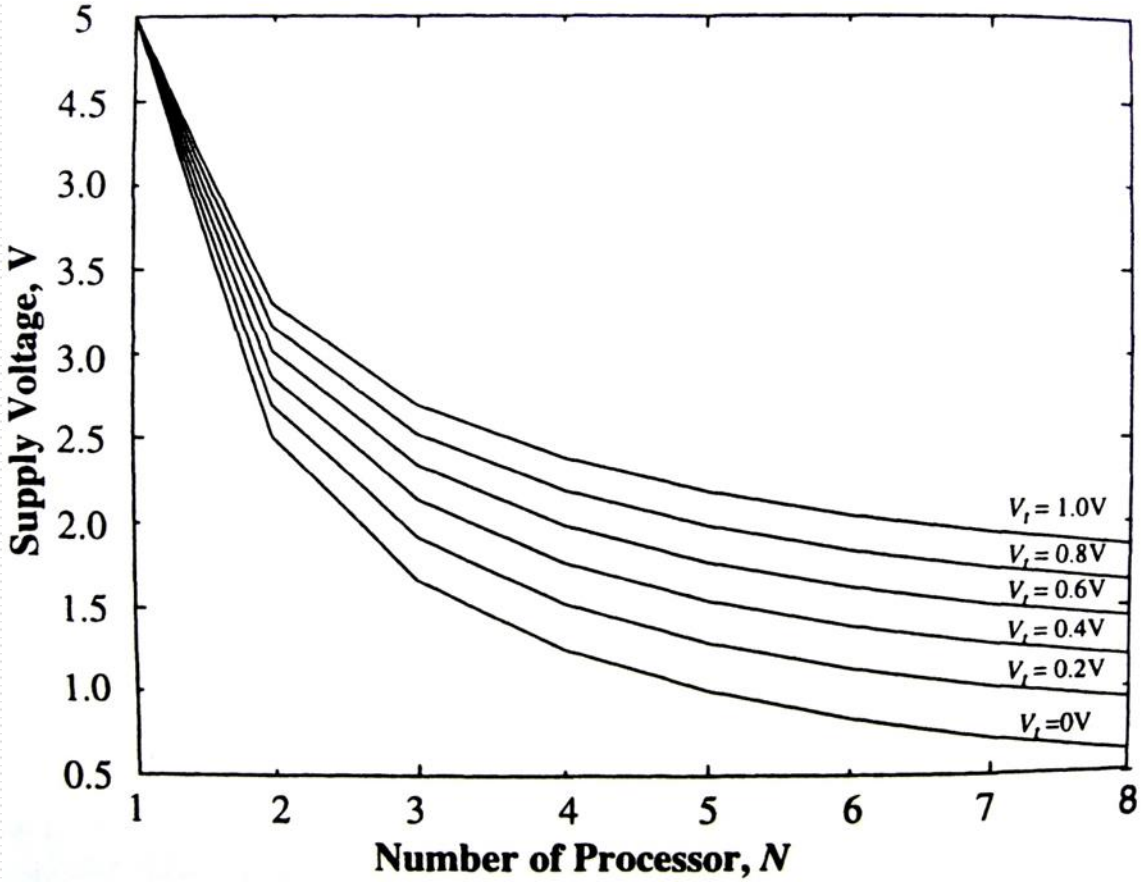
where,

$$\gamma = \frac{\left(1 - \beta^2\right)}{N}$$

If $V_t = 0$, then above equation degenerates to:

$$V_N = \frac{V_{ref}}{N}$$

Supply voltage reduction vs. number of processors as a function of V_t



Total Power

- The parallel architecture shown above does have overhead circuitry which requires extra power.
- Therefore the total power consumption of the parallel implementation is:

$$P_N = P_{processing} + P_{overhead}$$

- $P_{processing}$ is the power consumed by the input registers, the logic blocks which are clocked at f_{sample}/N, and the output register which is clocked at f_{sample}.

Processing Power

- The processing power is then given by,

$$P_{processing} = N\left(C_{inreg} + C_F\right)V_N^2 \frac{f_{sample}}{N} + C_{outreg} V_N^2 f_{sample}$$

$$= \left(C_{inreg} + C_F + C_{outreg}\right)V_N^2 f_{sample} = C_{ref} V_N^2 f_{sample}$$

Overhead Power

- The overhead power, $P_{overhead}$, consists of 3 components:
  - routing capacitance due to broadcast of the input which is switched at f_{sample}.
  - the output routing from the processors
  - multiplexer overhead circuitry operating at the sample rate.

- The overhead power is given by:

$$P_{overhead} = C_{overhead}(N)\, V^2_N\, f_{sample}$$

Power Improvement

- If the overhead components of power are assumed to be zero, then the power consumption can be arbitrarily reduced by making the computation parallel.
- The power in this case is given by:

$$P_N = P_{processing} = C_{ref} V^2_N f_{sample}$$

- The power improvement over uni-processor is,

$$\frac{P_N}{P_1} = \frac{V_N^2}{V_{ref}^2}$$

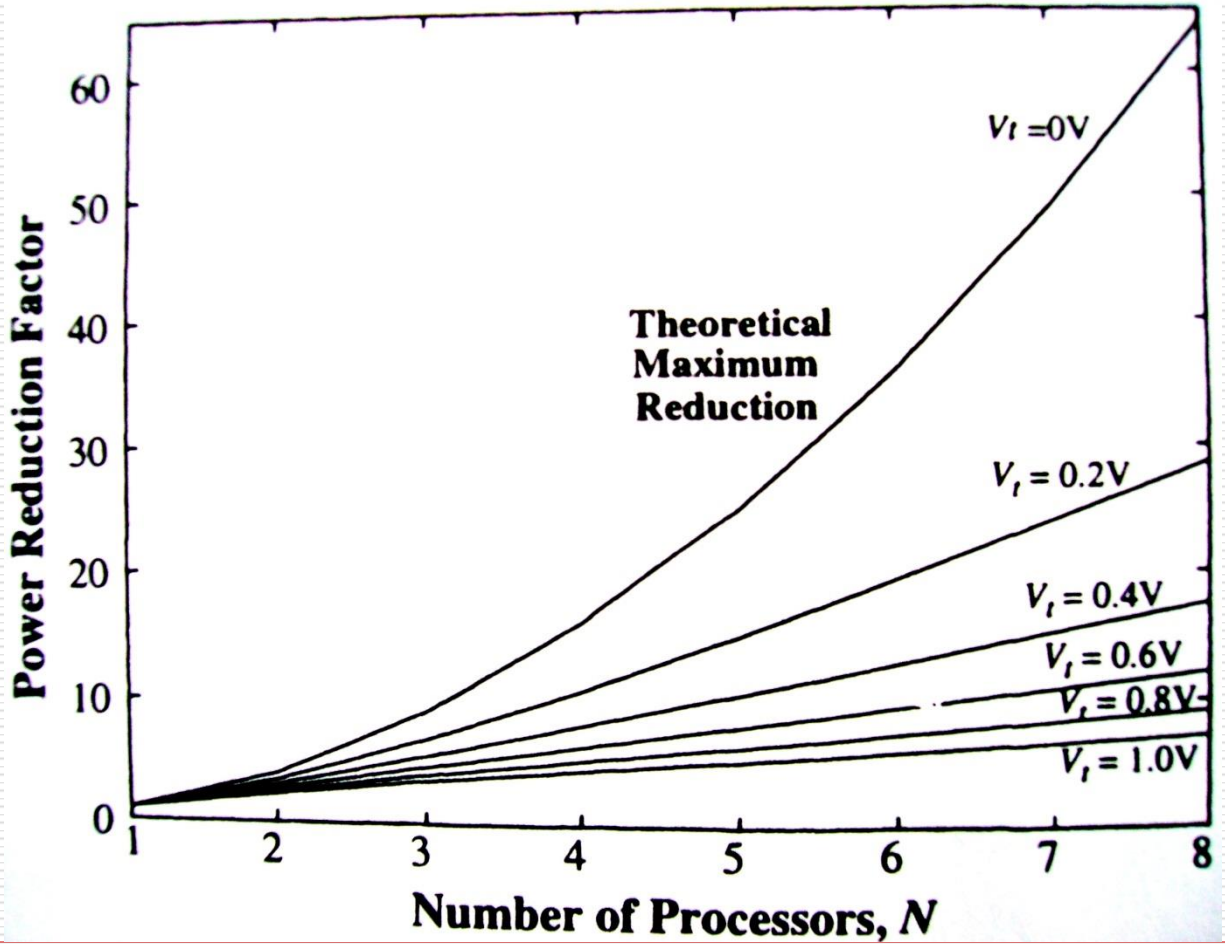
Power Improvement

- By substituting the values of V_N and V_{ref}, assuming $V_t = 0$, we get,

$$\frac{P_N}{P_1} = \frac{V_N^2}{V_{ref}^2} = \frac{1}{N^2}$$

- This is the limit in power reduction achievable with architecture driven voltage scaling alone. This ignores leakage currents and the overhead circuitry required to parallelize a circuit.

Power reduction (P_1/P_N) vs. number of processors as a function of V_t



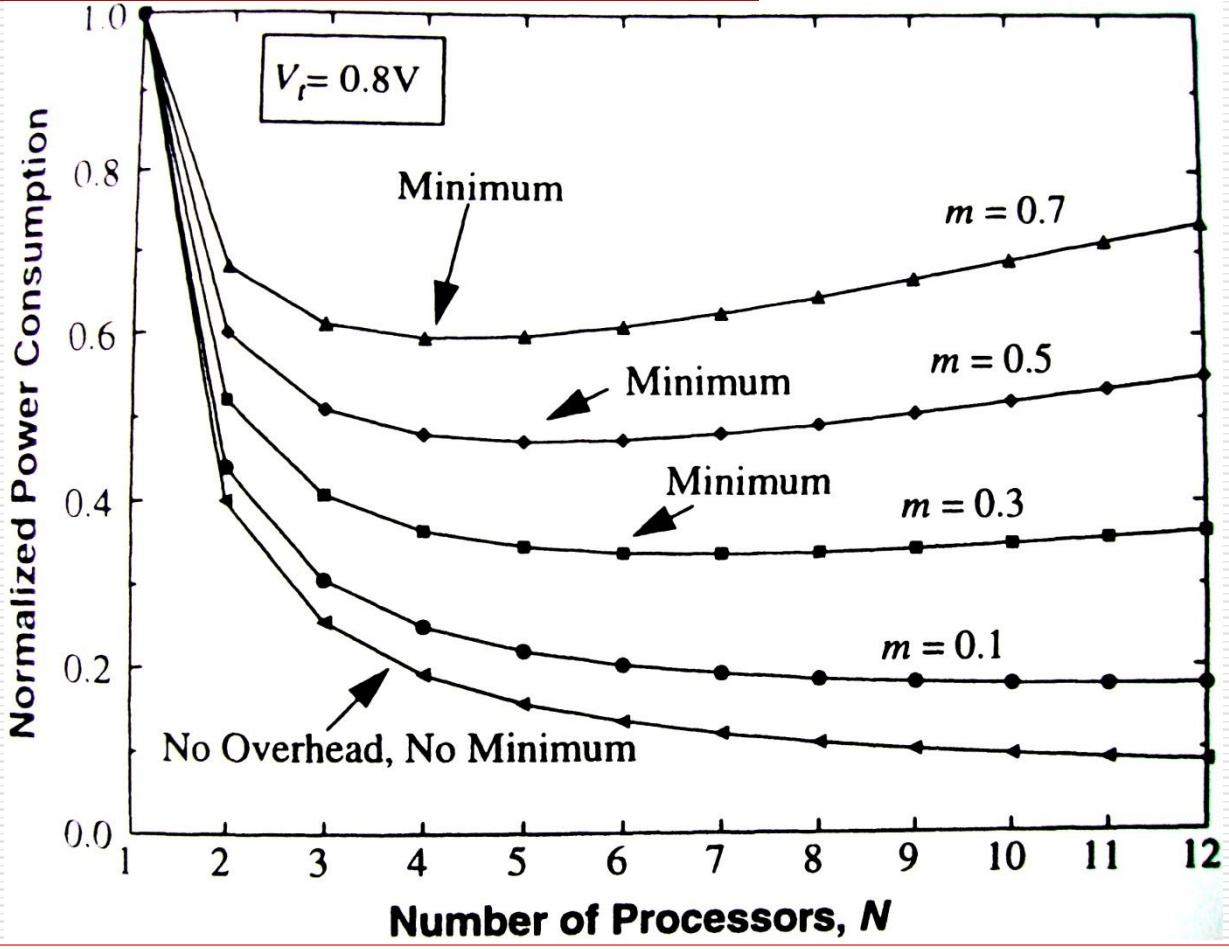
Optimal Supply Voltage for Architecture Driven Voltage Scaling

$$P_N = C_{ref}\, V^2_N\, f_{sample} + C_{overhead}(N)\, V^2_N\, f_{sample}$$

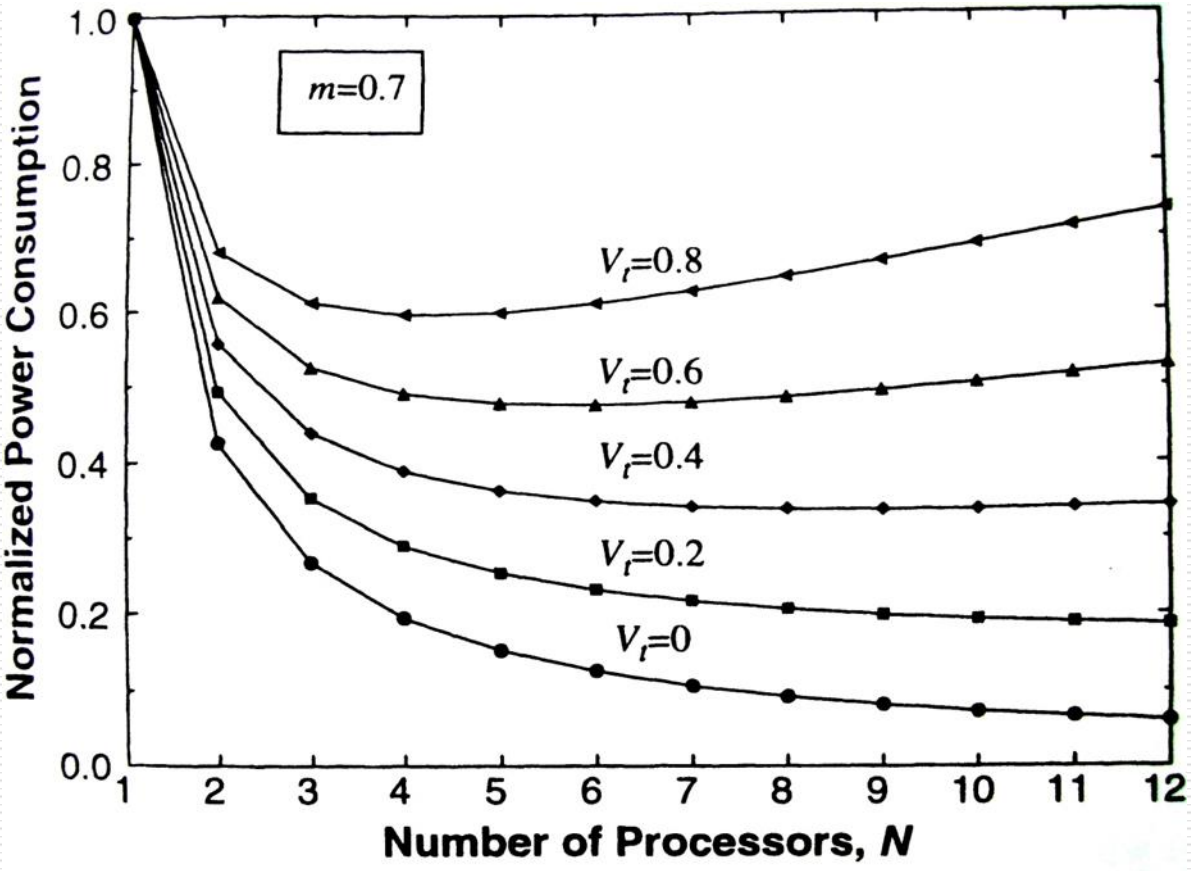
$$P_{normalized} = \frac{P_N}{P_1}$$

$$P_{normalized} = \left(1 + \frac{C_{overhead}(N)}{C_{ref}}\right)\left(\frac{V_N}{V_{ref}}\right)^2$$

Power vs. N for as a function of overhead capacitance (m=$C_{overhead}/C_{ref}$)



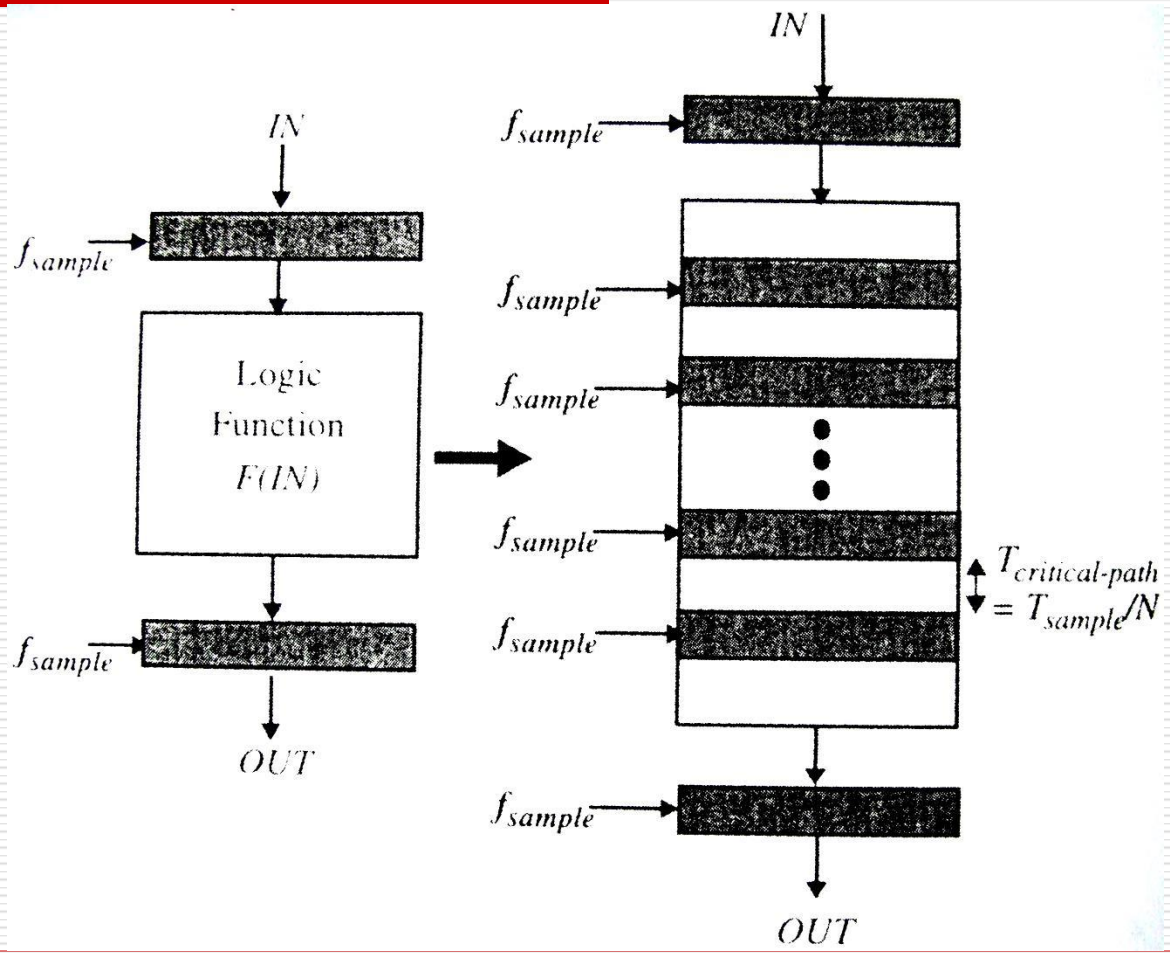
Power vs. N for various threshold voltages



Drawbacks:

- Increase in the area.
- Increase in latency → is equal to N cycles.
- But for most signal processing applications like video compression, throughput is the critical metric and latency can be tolerated.

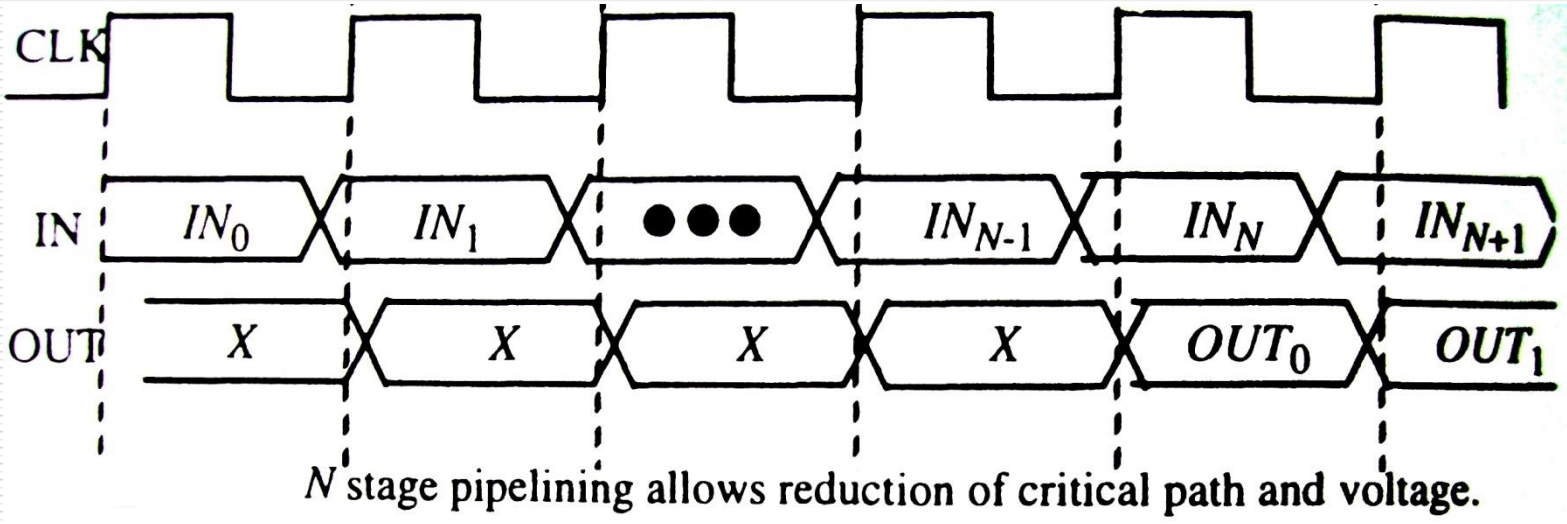
2. Trading area for Lower Power Through Hardware Pipelining



Trading area for Lower Power Through Hardware Pipelining

- The task is divided equally between N stages
- The critical path for a stage is 1/N th of the uniprocessor system
- The throughput clock rate is f_{sample}
- Thus the N-1 registers are clocked at f_{sample}
- This has the effect of increased time available for the critical path by a factor N for each of these N stages
- The new critical path delay is N times the original
- Supply voltage can be reduced to save power

Trading area for Lower Power Through Hardware Pipelining



N stage pipelining allows reduction of critical path and voltage.

Trading area for Lower Power Through Hardware Pipelining

- The supply voltage at which N stage pipelined implementation will have the same throughput is

$$\frac{V_N}{V_{ref}} = \frac{2\beta + \gamma + \sqrt{4\beta\gamma + \gamma^2}}{2}$$

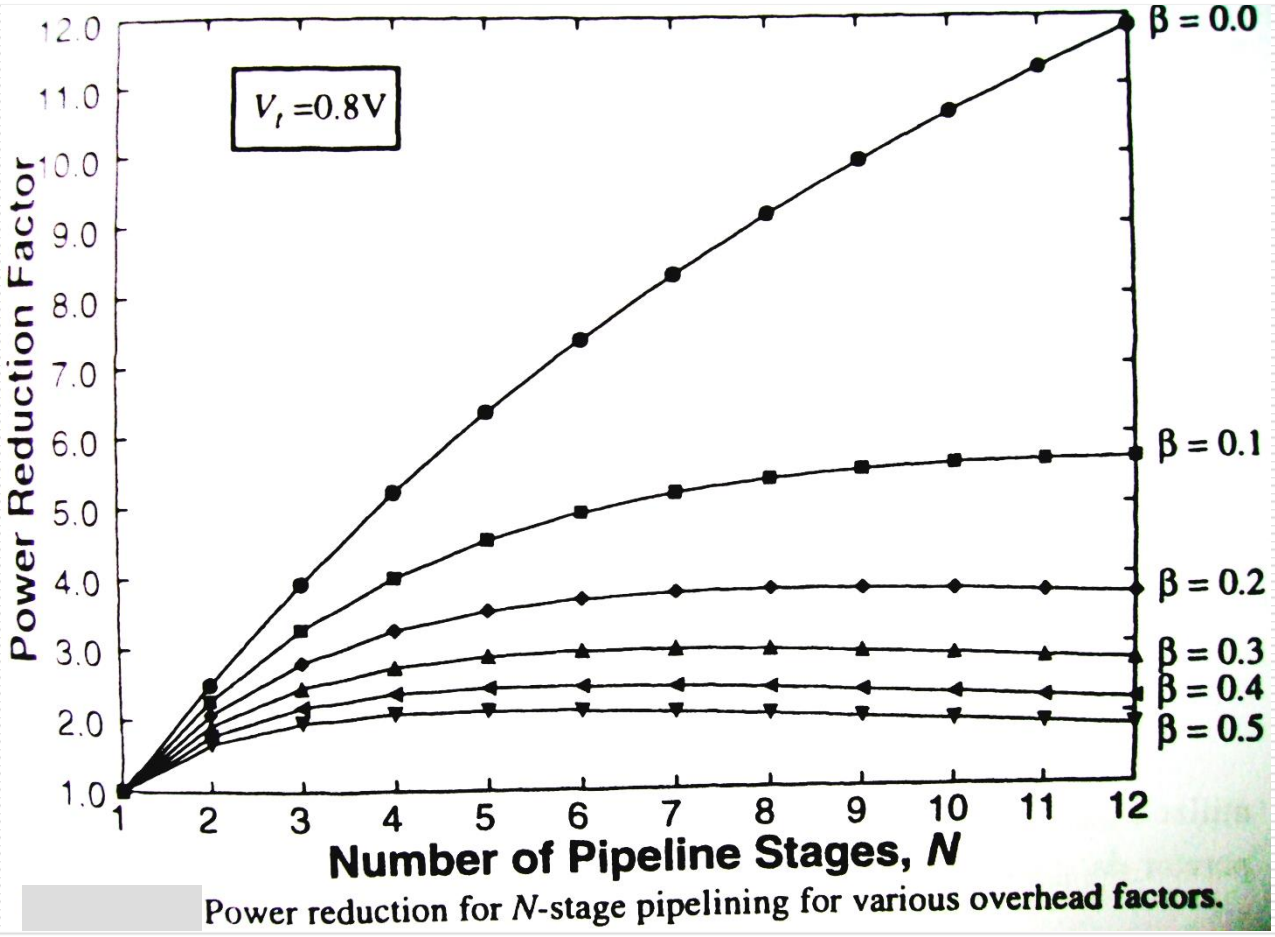
Trading area for Lower Power Through Hardware Pipelining

$$\frac{P_N}{P_1} = \frac{\left(C_{ref} + (N-1)\,C_{reg}\right)V_N^2 f_{sample}}{C_{ref} V_{ref}^2 f_{sample}} = (1 + \beta(N-1))\frac{V_N^2}{V_{ref}^2}$$

$C_{ref} = 2\ C_{reg} + C_F$; Non–pipelined capacitance

where $\beta = C_{reg}/C_{ref} < 0.5$

Trading area for Lower Power Through Hardware Pipelining



Power reduction for N-stage pipelining for various overhead factors.