



Low-Power VLSI

Seong-Ook Jung

2011. 5. 6.

sjung@yonsei.ac.kr

*VLSI SYSTEM LAB, YONSEI University
School of Electrical & Electronic Engineering*



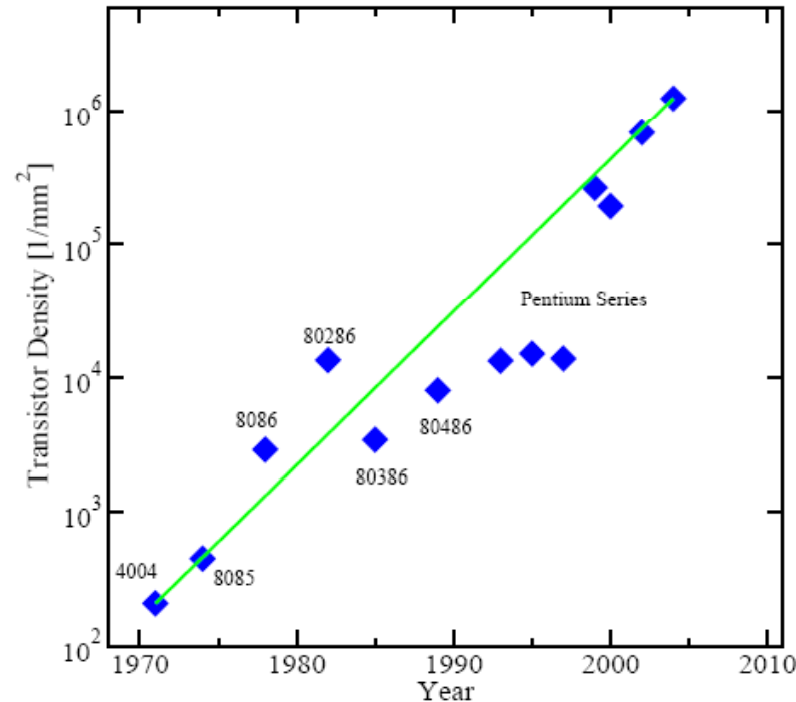
Contents

- 1. Introduction**
- 2. Power classification**
- 3. Power – performance relationship**
- 4. Low power design**
 - 1. Architecture and algorithm level**
 - 2. Block and logic level**
 - 3. Circuit level**
 - 4. Device level**
- 5. OMAP processor**
- 6. Summary**

Introduction

VLSI
SYSTEM LAB.

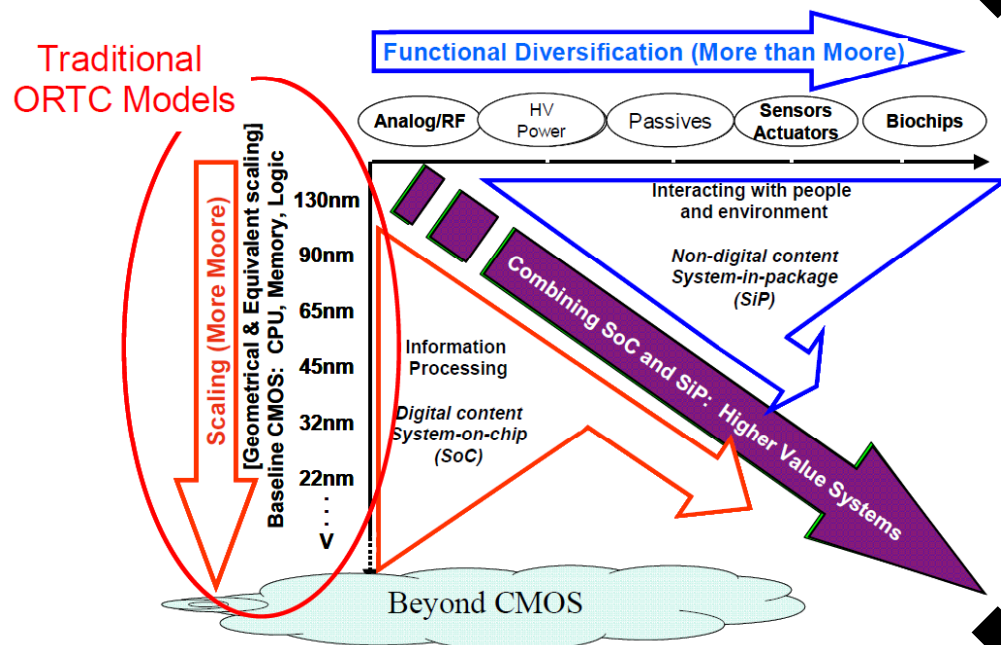
Technology Scaling



◆ Technology scaling : Moore's law

- The number of transistors that can be placed on an integrated circuit has doubled approximately every 18 months

Development Trend



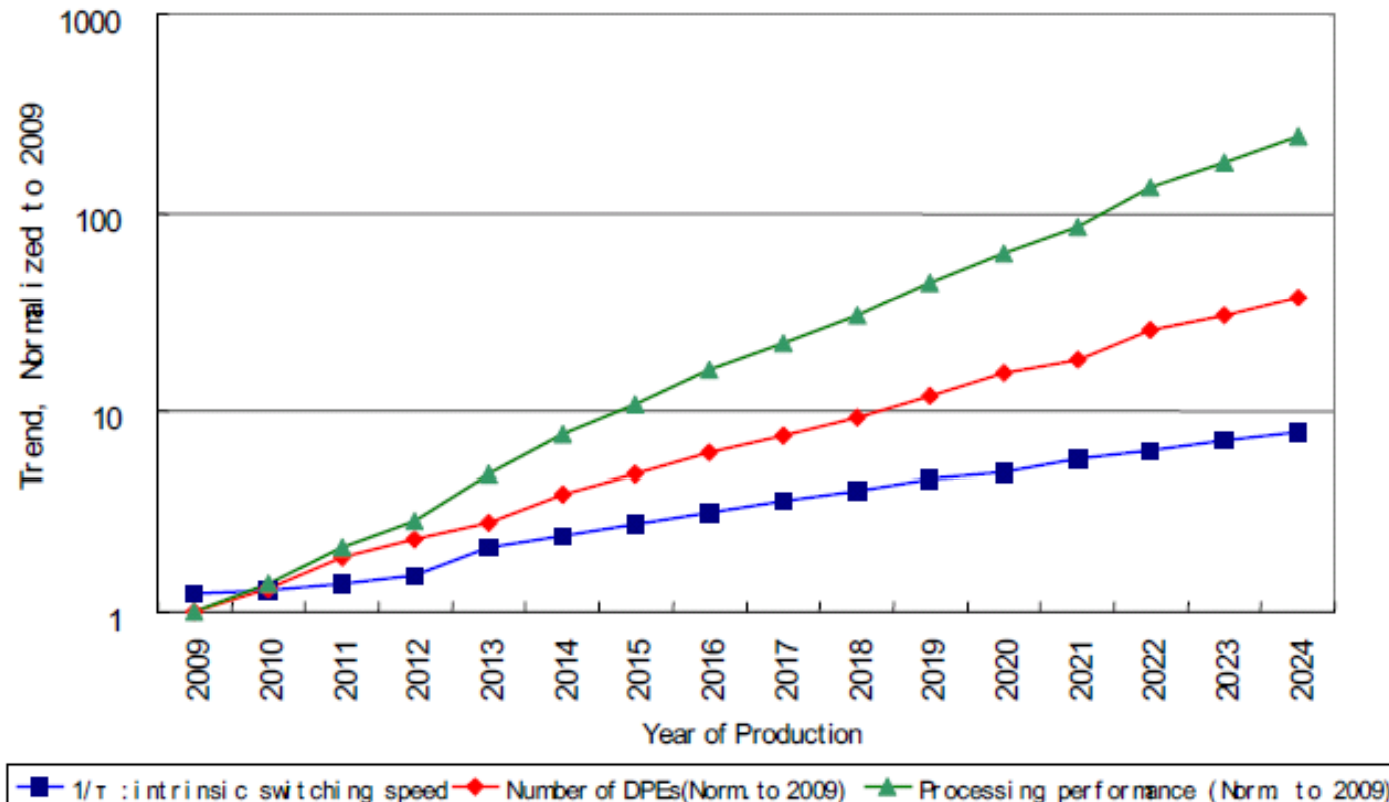
◆ Scaling (More Moore)

- More devices are integrated in a chip
- New scaling road map
 - ❖ Not only 'geometrical scaling' for 2D device, but also 'equivalent scaling' for 3D device
- Beyond bulk CMOS
 - ❖ FinFET, SOI...

◆ Functional diversification (More than Moore)

- Several functions are merged in a chip

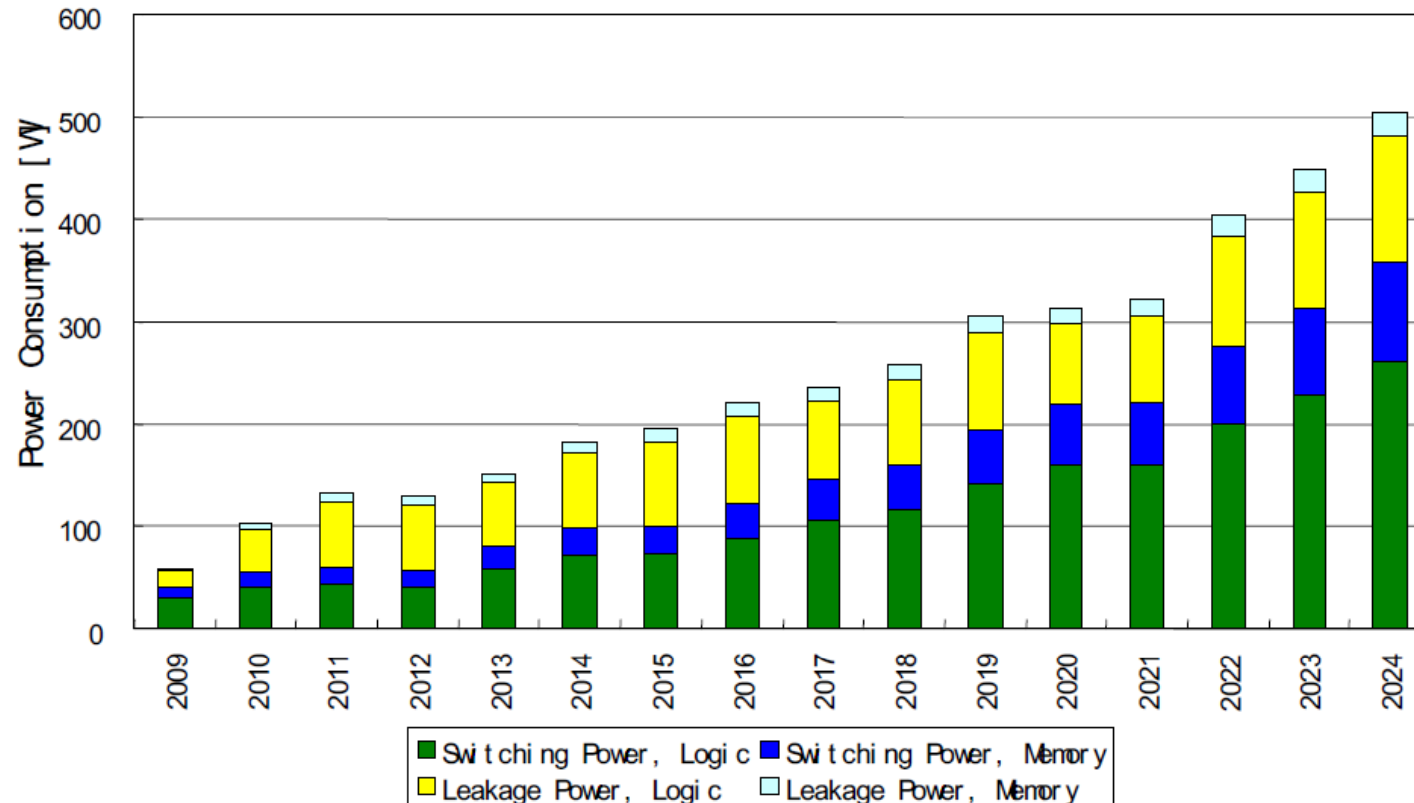
SoC Performance



◆ SoC performance : exponentially increase!!

- Thanks to both device technology and design methodology

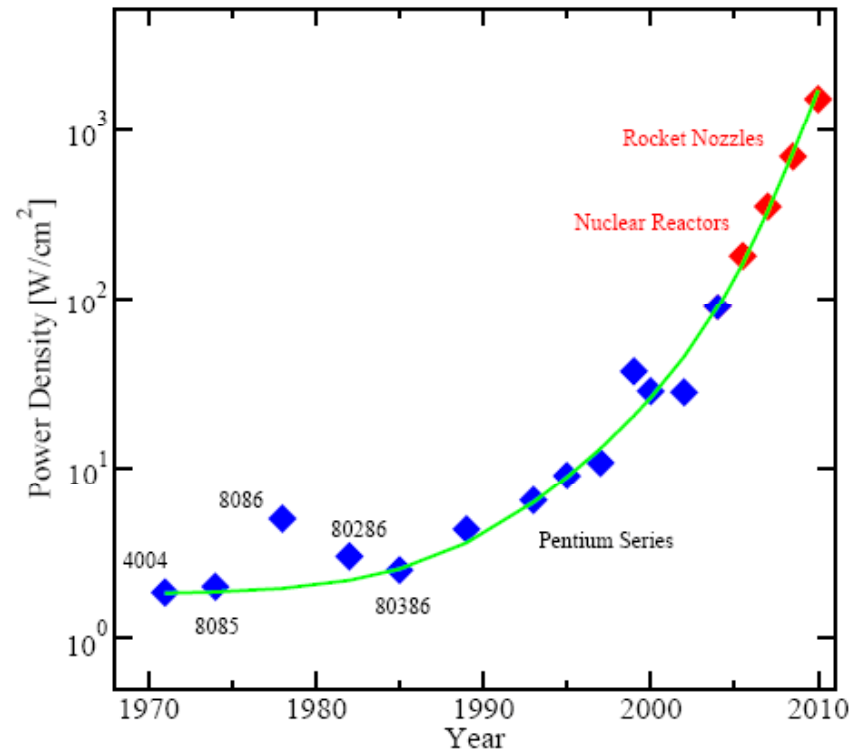
SoC Power Consumption Problem



◆ SoC power consumption : 'also' severely increase

- After 15 years, x10 power is required...

SoC Power Density Problem



◆ Power density : exponentially increase!!

- Power consumption per die area (W/cm^2)
- We would soon reach power densities of nuclear power plants or rocket nozzles in a few years!!

Process Variation Problem

◆ Process variation : Result of scaling

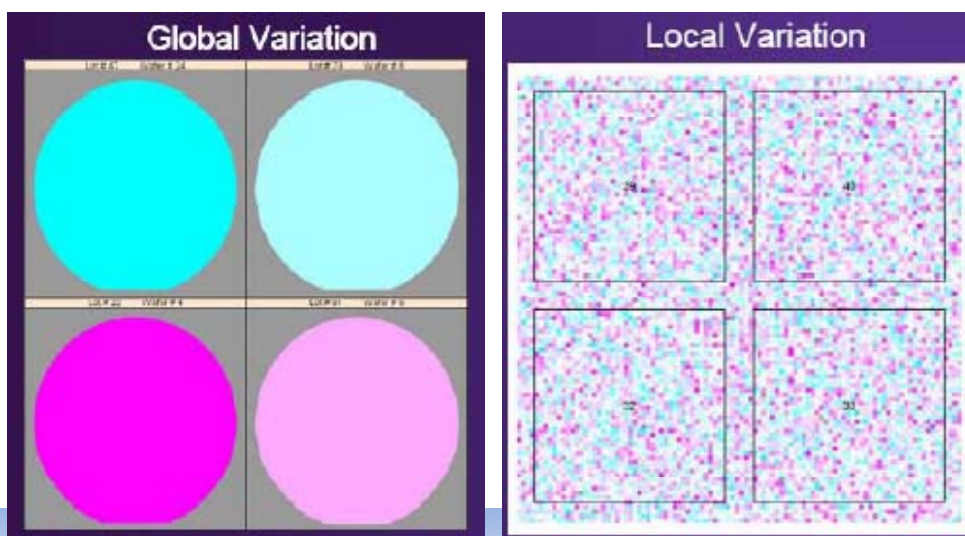
● Global variation and local variation

❖ Global variation

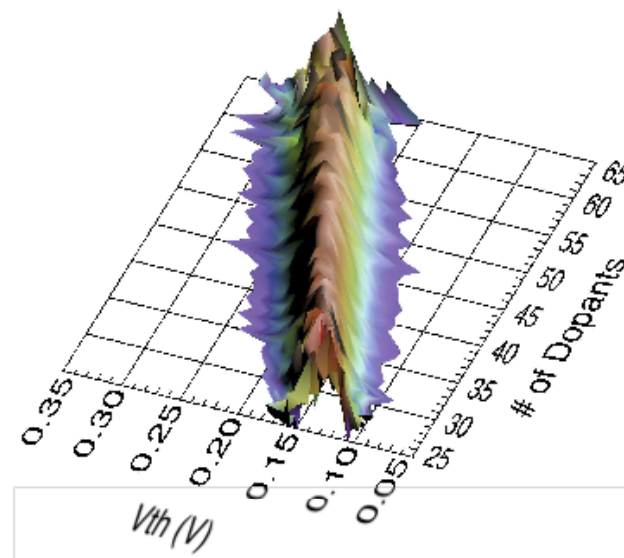
- Comes from fabrication, lot, wafer processes
- Different process corner (NMOS-PMOS : SS/SF/TT/FS/FF)

❖ Local variation

- Truly random variation between device with identical layout

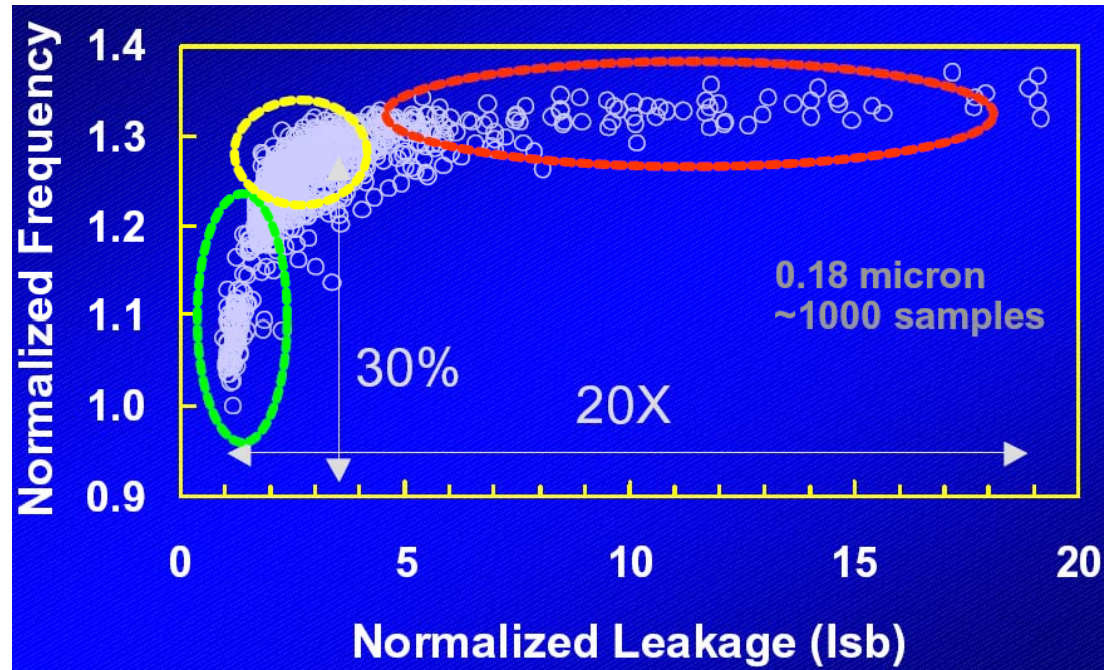


[3] Synopsis, 2005



[4] <http://cnx.org>

Process Variation Problem



◆ Performance variation due to process variation

- Frequency difference $\approx 30\%$
 - Leakage current difference $\approx \times 20$
- ⇒ Process variation should be considered in SoC design

Effect of the Process Variation

◆ Low Voltage / Low Power limitation

- $I_D \propto W/L \cdot (V_{DD} - V_{TH})^\alpha$
- V_{TH} variation $\Rightarrow I_D$ variation \Rightarrow Performance Variation !!
- Need more design margin due to process variation $\Rightarrow V_{DD} \uparrow$

◆ Yield limitation

- Because of process variation, failure probability $\uparrow \Rightarrow$ Yield \downarrow

저전력 SoC

2009 ITRS SPECIAL TOPICS

ENERGY

Energy consumption has become an increasingly important topic of public discussion in recent years because of global CO₂ emission. Since semiconductor electronics are broadly applicable to energy collection, conversion, storage, transmission, and consumption/usage, it is not surprising that the ITRS addresses many factors of significance to energy issues. In general, the ITRS documents the impressive trends and, more importantly, sets aggressive targets for future electronics energy efficiency, for example, computational energy/operation (per logic and per memory-bit state changes). The most detailed targets relate directly to semiconductor materials, process, and device technologies, which form the bases of integrated-circuit manufacturing and components, respectively.

- Low power VLSI design !!!
- Low process variation (high yield) design

Power Classification

VLSI
SYSTEM LAB.

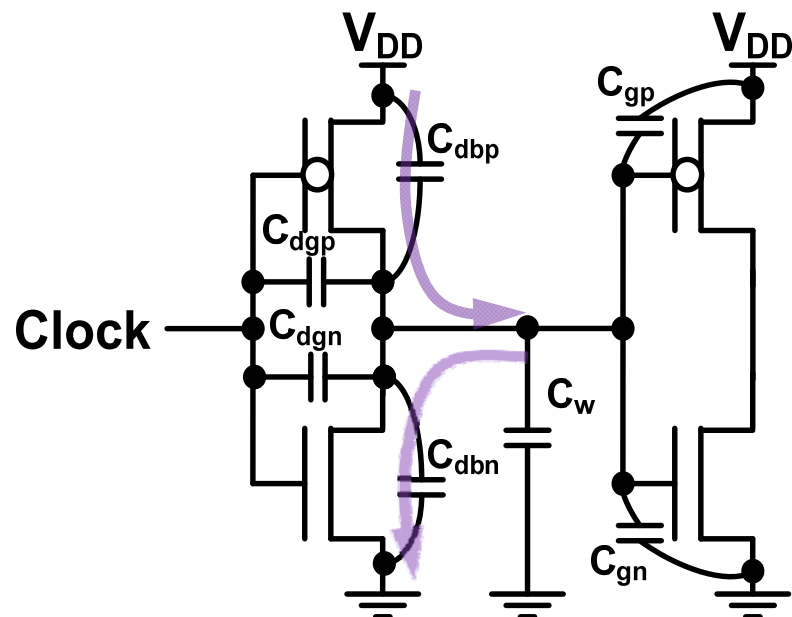
Power Classification

- ◆ Power consumption of CMOS circuits

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}}$$

$$P_{\text{dynamic}} = P_{\text{sw}} + P_{\text{sc}}$$

Switching Power



$$I = C_L dV/dt = C_L \Delta V f$$

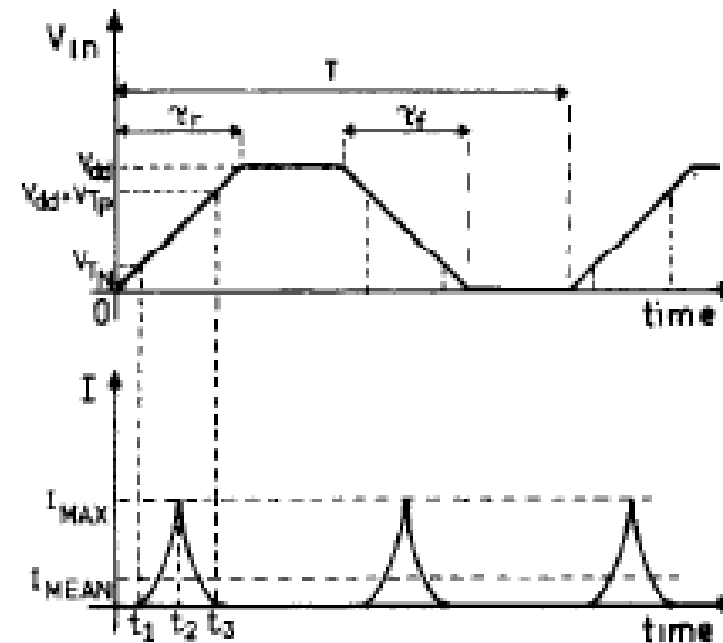
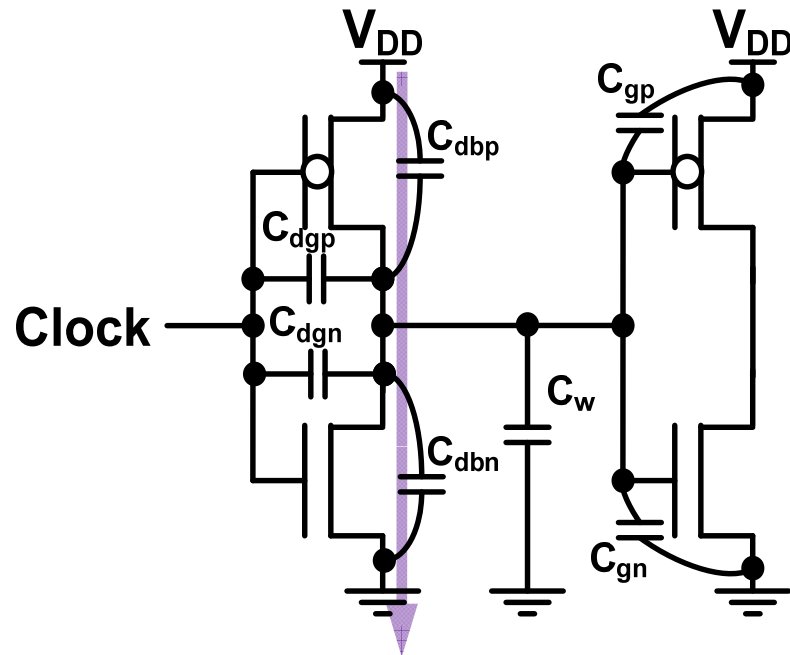
$$P_{sw} = I V_{DD} = C_L \Delta V V_{DD} f$$

In digital circuit, $\Delta V = V_{DD}$

$$P_{sw} = I V_{DD} = C_L V_{DD}^2 f$$

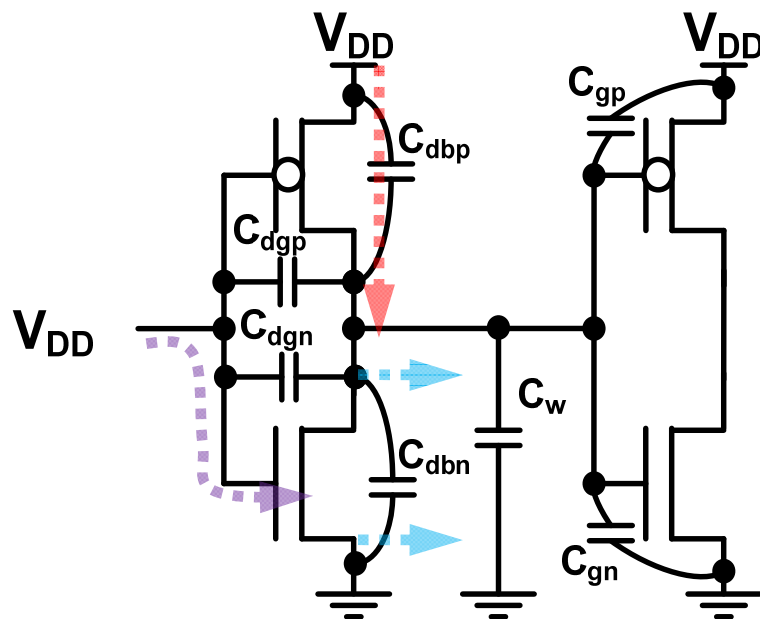
- ◆ P_{sw} is due to the charge and discharge (output transition) of the capacitors driven by the circuit according to input transition.
- ◆ $P_{sw} = C_L V_{DD}^2 f$

Short Circuit Power



- ◆ P_{sc} is caused by the simultaneous conductance of PMOS and NMOS during input and output transitions.
- ◆ $P_{sc} = (\beta/12)(V_{DD}-2V_{TH})^3 (t_3-t_1)$

Static Power : P_{sub} , P_{gate} & P_{junc}



◆ P_{sub}

- Sub- V_{TH} leakage : $|V_{GS}| < |V_{TH}|$

- $P_{sub} \propto \text{Exp}[(V_{GS} - V_{TH})/m v_T] V_{DD}$

◆ P_{gate}

- Ideal MOSFET : $I_{gate} = 0$
- In short channel MOSFET, I_{gate} exists because of thin T_{OX}

- $P_{gate} \propto WL (V_{GS}/T_{OX})^2 V_{DD}$

◆ P_{junc}

- Reverse PN junction leakage

- $P_{junc} \propto \text{Exp}[V_D/v_T - 1] V_{DD}$

Power – Performance Relationship

VLSI
SYSTEM LAB.

V_{DD} Reduction

◆ Power consumption equation

- $P_{sw} = C_L V_{DD}^2 f$
- $P_{sc} = (\beta/12) (V_{DD} - 2V_{TH})^3 (t_3 - t_1)$
- $P_{sub} \propto \text{Exp}[(V_{GS} - V_{TH})/m v_T] V_{DD}$
- $P_{gate} \propto WL (V_{GS}/T_{OX})^2 V_{DD}$
- $P_{junc} \propto \text{Exp}[V_D/v_T - 1] V_{DD}$

◆ Case.1 : $V_{DD} \downarrow$

- All power consumption \downarrow
 - However...
 - ❖ Delay $\propto C_L V_{DD}/I_D \propto C_L V_{DD}/(V_{DD} - V_{TH})^\alpha$
 - ❖ If $V_{DD} \downarrow$, Delay \uparrow
- ⇒ **Performance loss**

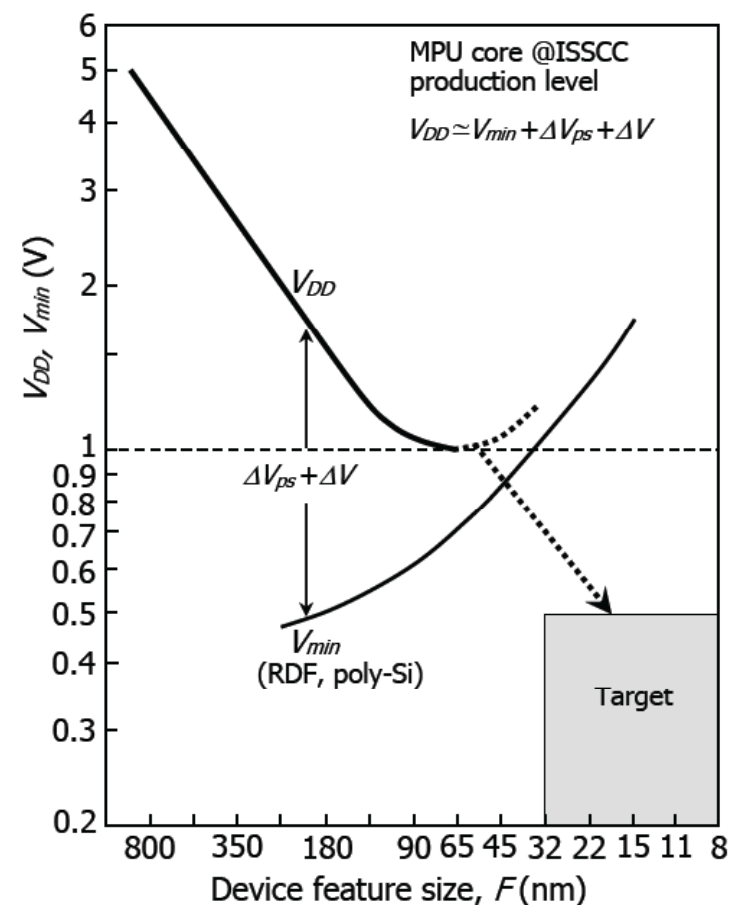
V_{DD} Scaling Limitation

◆ Low V_{DD} limitation with process variation

- $V_{DD.min} = V_{T0} + K\sigma(V_T)$
 - ❖ $\sigma(V_T)$: 1-sigma of V_T variation
 - $\propto T_{ox} N_A^{0.25} (LW)^{-0.5}$
- Significant increment of $\sigma(V_T)$ with technology scaling ($LW \downarrow \downarrow$)

⇒ V_{DD} scaling meets the limitation!!

⇒ Process variation tolerant circuit design technique is required!!



High V_{TH}

◆ Power consumption equation

- $P_{sw} = C_L V_{DD}^2 f$
- $P_{sc} = (\beta/12) (V_{DD} - 2V_{TH})^3 (t_3 - t_1)$
- $P_{sub} \propto \text{Exp}[(V_{GS} - V_{TH})/m v_T] V_{DD}$
- $P_{gate} \propto WL (V_{GS}/T_{OX})^2 V_{DD}$
- $P_{junc} \propto \text{Exp}[V_D/v_T - 1] V_{DD}$

◆ Case.2 : $V_{TH} \uparrow$

- $P_{sc} \downarrow$ and especially, $P_{sub} \downarrow$
 - However...
 - ❖ Delay $\propto C_L V_{DD}/I_D \propto C_L V_{DD}/(V_{DD} - V_{TH})^\alpha$
 - ❖ If $V_{TH} \uparrow$, Delay \uparrow
- ⇒ **Performance loss**

Low Frequency

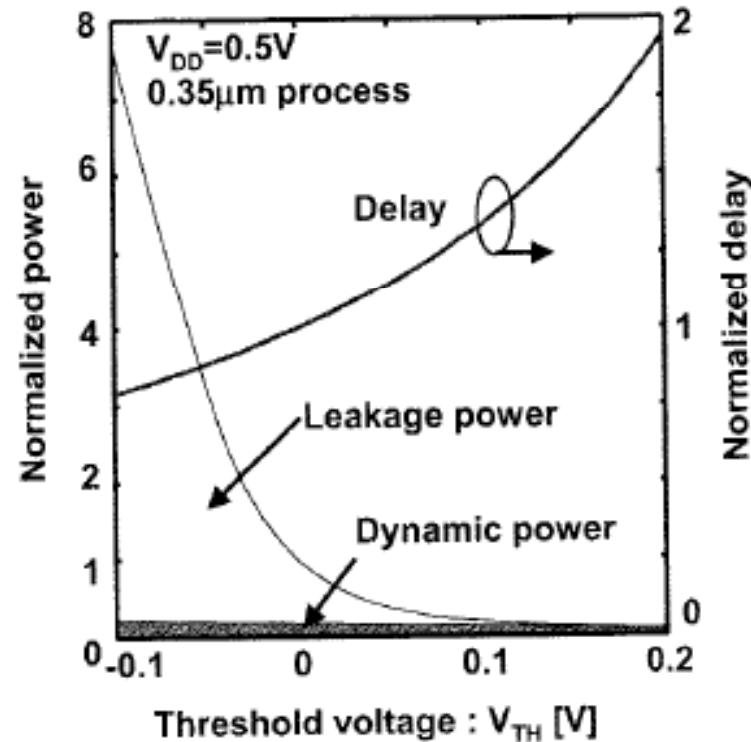
◆ Power consumption equation

- $P_{sw} = C_L V_{DD}^2 f$
- $P_{sc} = (\beta/12) (V_{DD} - 2V_{TH})^3 (t_3 - t_1)$
- $P_{sub} \propto \text{Exp}[(V_{GS} - V_{TH})/m v_T] V_{DD}$
- $P_{gate} \propto WL (V_{GS}/T_{OX})^2 V_{DD}$
- $P_{junc} \propto \text{Exp}[V_D/v_T - 1] V_{DD}$

◆ Case.3 : $f \downarrow$

- $P_{sw} \downarrow$
- However...
 - ❖ Throughput $\propto f$
 - ⇒ **Performance loss**

Tradeoff



- ⇒ Tradeoff between low power and high performance
- ⇒ Low power design :
 - power reduction without performance degradation

Low power design

VLSI
SYSTEM LAB.

Low Power Design Methodology

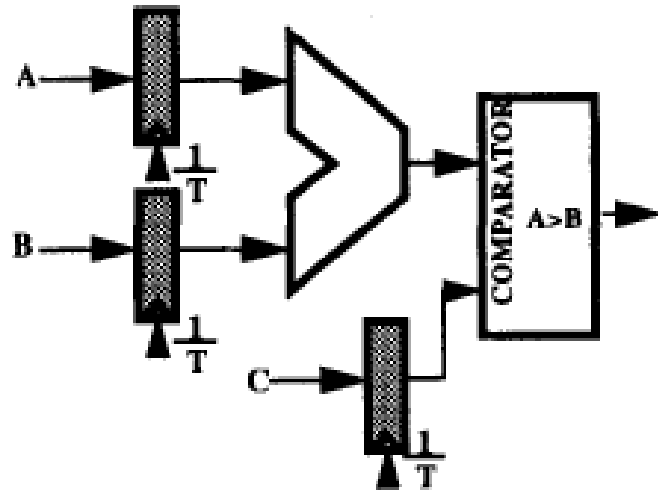
◆ To make low power SoC...

- Architecture and algorithm levels
 - ❖ Parallelism, Pipeline ...
- Block and logic levels
 - ❖ V_{DD} / Frequency scheduling by monitoring workload (AVFS)
 - ❖ Temperature management to reduce leakage current
- Circuit level
 - ❖ Circuit type (Dynamic, static, ...)
 - ❖ Circuit technique (Dual V_{DD} , Dual V_{TH} , MTCMOS, ...)
- Device level
 - ❖ Control the process parameter
 - Halo doping, retrograde well...
 - ❖ Low leakage new device
 - SOI, FinFET ...

Architecture and Algorithm Levels

VLSI
SYSTEM LAB.

Parallelism



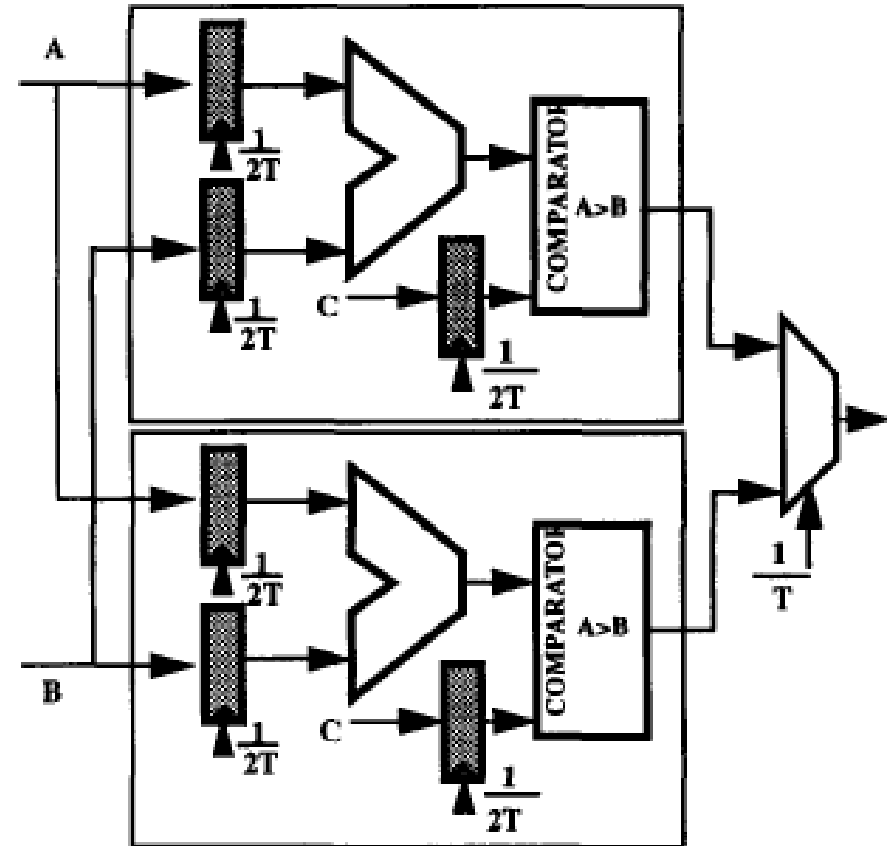
< A simple adder comparator DP >

$$P_{ref} = C_{ref} V_{ref}^2 f_{ref} \quad P_{par} = C_{par} V_{par}^2 f_{par}$$

$$\frac{P_{par}}{P_{ref}} = \frac{C_{par}}{C_{ref}} \frac{f_{par}}{f_{ref}} \frac{V_{par}^2}{V_{ref}^2} = (N + \delta) \frac{1}{N} \frac{V_{par}^2}{V_{ref}^2}$$

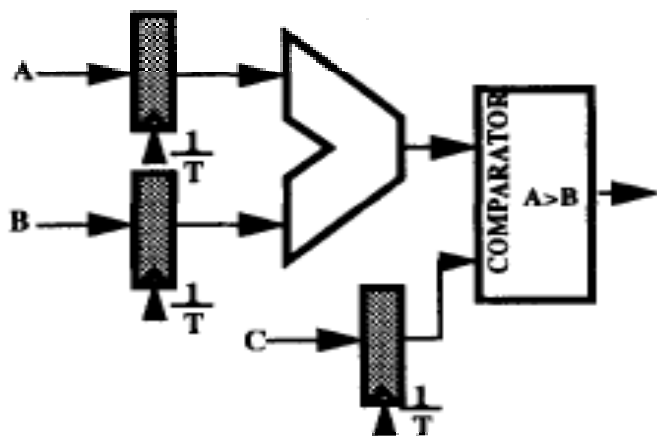
N: # of parallelism

δ : a slight increase in capacitance due to the extra routing

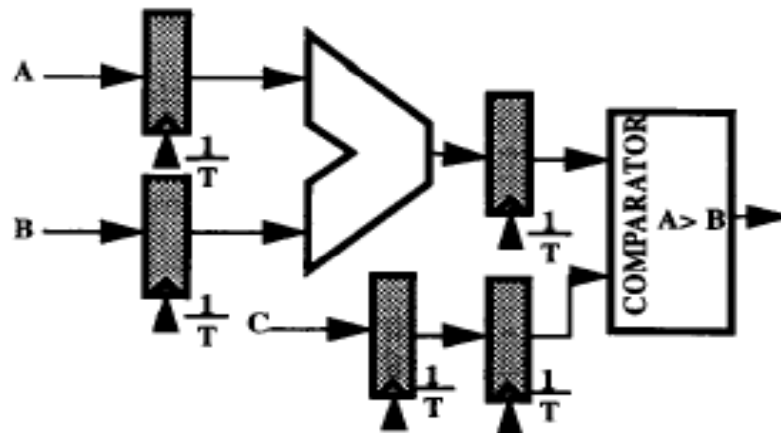


< Parallel implementation >

Pipeline



< A simple adder comparator DP >



< Pipeline implementation >

$$P_{ref} = C_{ref} V_{ref}^2 f_{ref}$$

$$P_{pipe} = C_{pipe} V_{pipe}^2 f_{pipe}$$

$$\frac{P_{pipe}}{P_{ref}} = \frac{C_{pipe}}{C_{ref}} \frac{f_{pipe}}{f_{ref}} \frac{V_{pipe}^2}{V_{ref}^2} = (1 + \delta) \frac{V_{pipe}^2}{V_{ref}^2}$$

N: # of pipeline stage

δ : a slight increase in capacitance due to the extra latch

Circuit Level

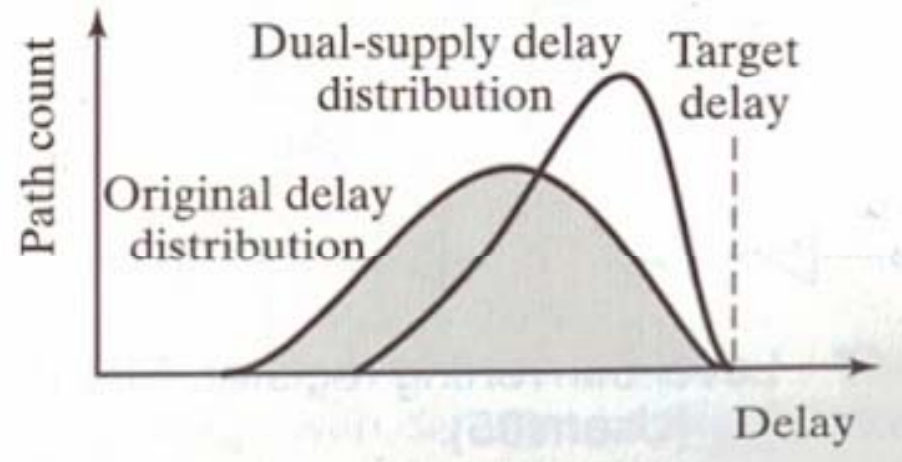
VLSI
SYSTEM LAB.

Circuit Level Low Power Techniques

◆ Low power techniques

- Multiple channel length
- Stacked transistor
- Dual V_{DD}
- Dual V_{TH}
- MTCMOS (Multi Threshold voltage CMOS)
- DVS (Dynamic Voltage Scaling) : open-loop / closed loop

Critical Path



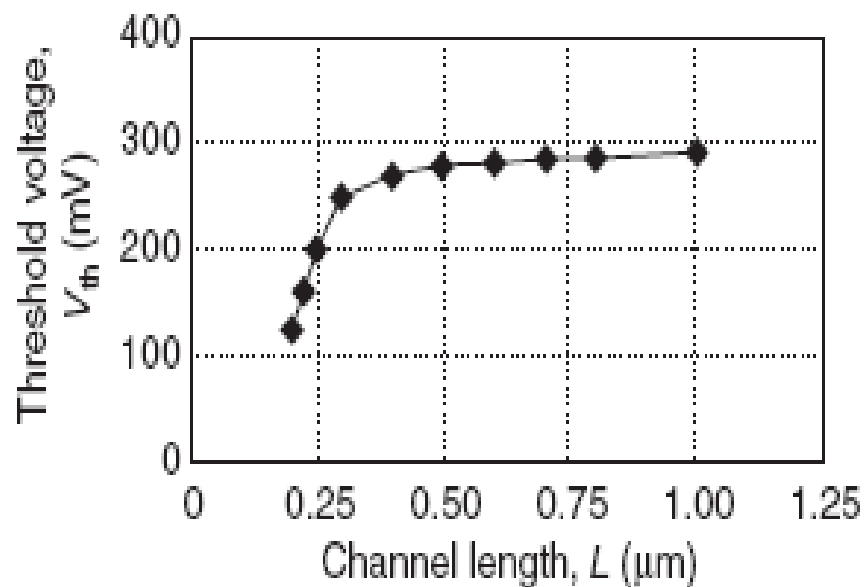
◆ Critical Path : The worst case delay path

- Determines SoC's maximum performance
- # of critical path \ll # of non-critical path
- Fast non-critical path is just wasteful...
 - ⇒ **By increasing non-critical path's delay, we may achieve power reduction because of tradeoff relation between power & performance**

Multiple Channel Length

◆ Threshold voltage roll-off

- Longer L
 - ❖ Higher V_t
 - ❖ Low leakage with low performance
 - ❖ Used in non-critical path



Stacked Transistor

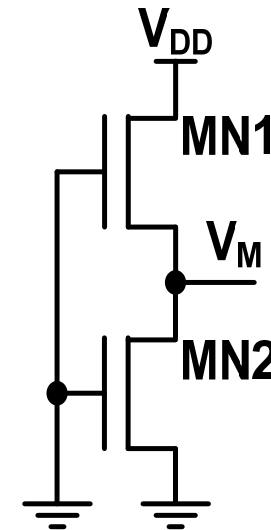
◆ V_M level

- $V_M > 0$ due to leakage current.
 - ❖ Negative V_{GS_MN1}
 - ❖ Positive V_{SB_MN1}
 - Increase in V_{TH} by body effect

$$P_{sub} \approx \left(e^{\frac{-(V_{gs} - V_{th})}{mv_T}} \right) V_{dd}$$

→ Large reduction in I_{sub}

◆ Primary input vector control to utilize the stack effect in the standby mode



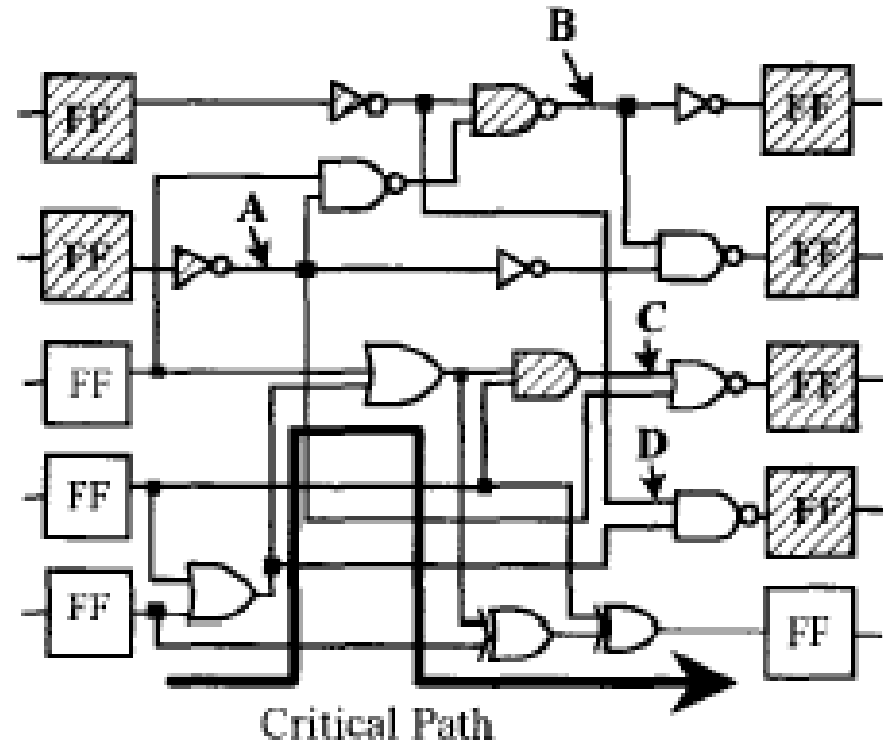
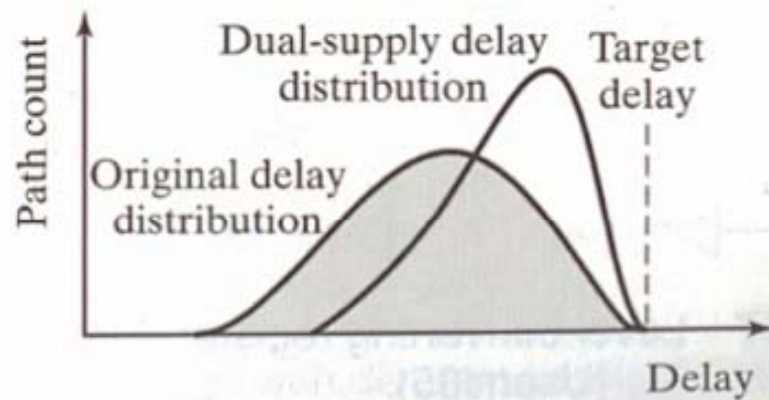
High V_t Low V_t

	High V_t	Low V_t
2 NMOS	10.7X	9.96X
3 NMOS	21.1X	18.8X
4 NMOS	31.5X	26.7X
2 PMOS	8.6X	7.9X
3 PMOS	16.1X	13.7X
4 PMOS	23.1X	18.7X

Dual V_{DD}

◆ Basic idea

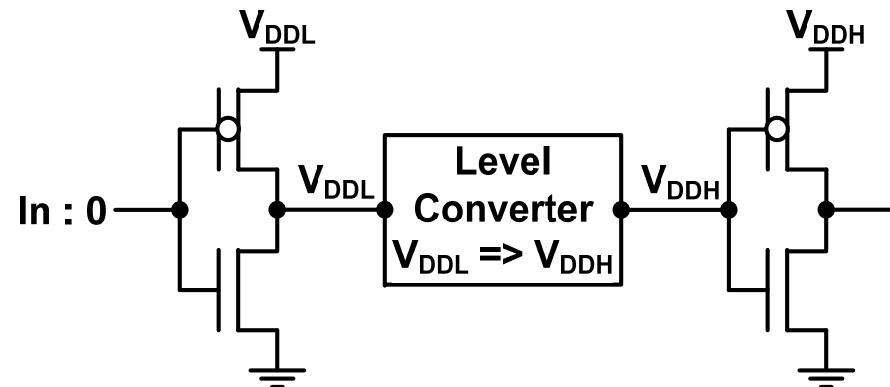
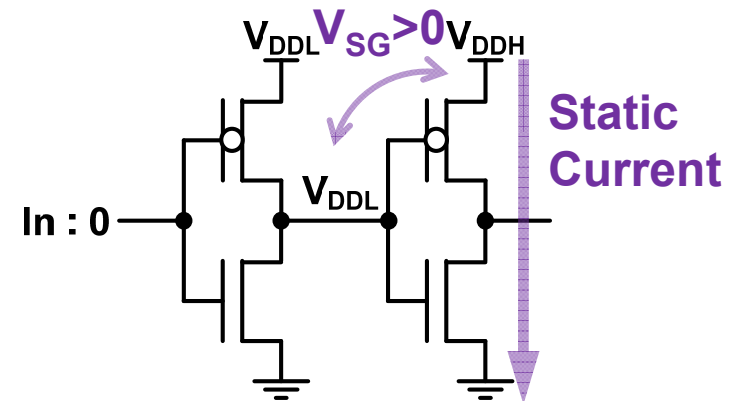
- V_{DDL}
 - ❖ Logic gates off the critical path
- V_{DDH}
 - ❖ Logic gate on the critical path
- Reduce power without degrading the performance



Dual V_{DD} : Design Issue & Target

◆ Issue

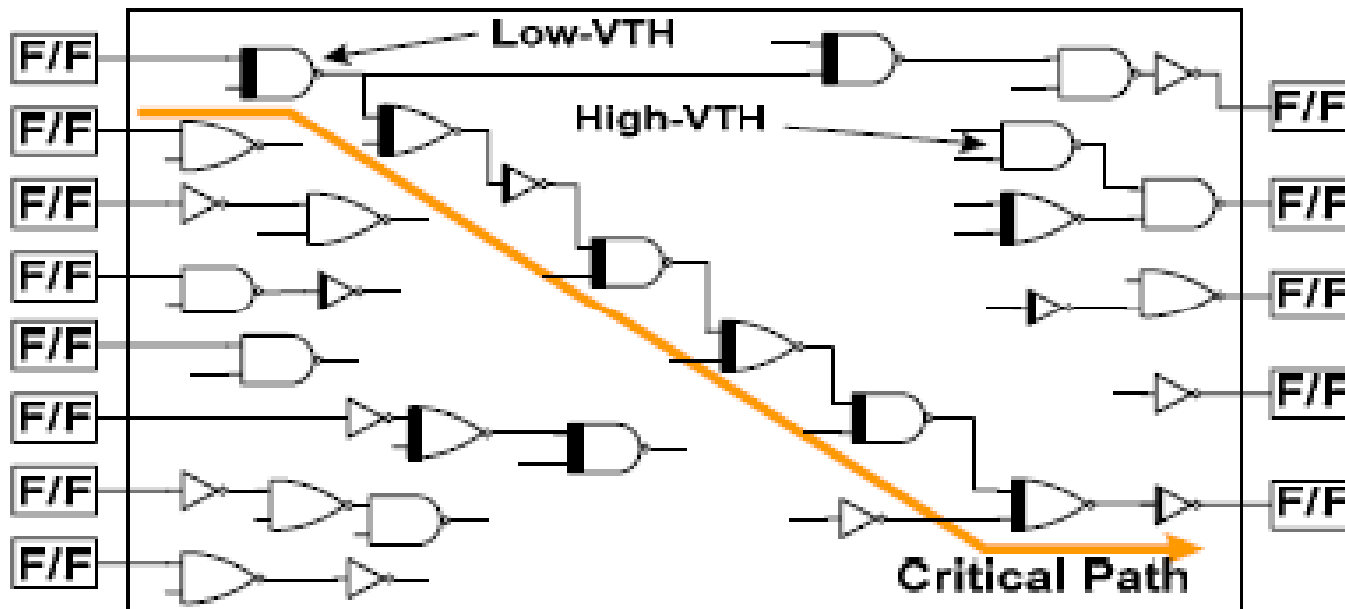
- Static current flow at a V_{DDH} gate if it is directly drive by a V_{DDL} gate
 - Level converter is needed
- ⇒ Overhead of area and power



◆ Design target

- For a give circuit, choose gates for V_{DDL} application to minimize power consumption while maintaining performance with consider level converter.

Dual V_{TH} Voltages



◆ HVt

- Assigned to transistors in noncritical path.
- Leakage saving in both standby and active modes

◆ LVt

- Assigned to transistors in critical path
- Maintained performance

MTCMOS : Basic

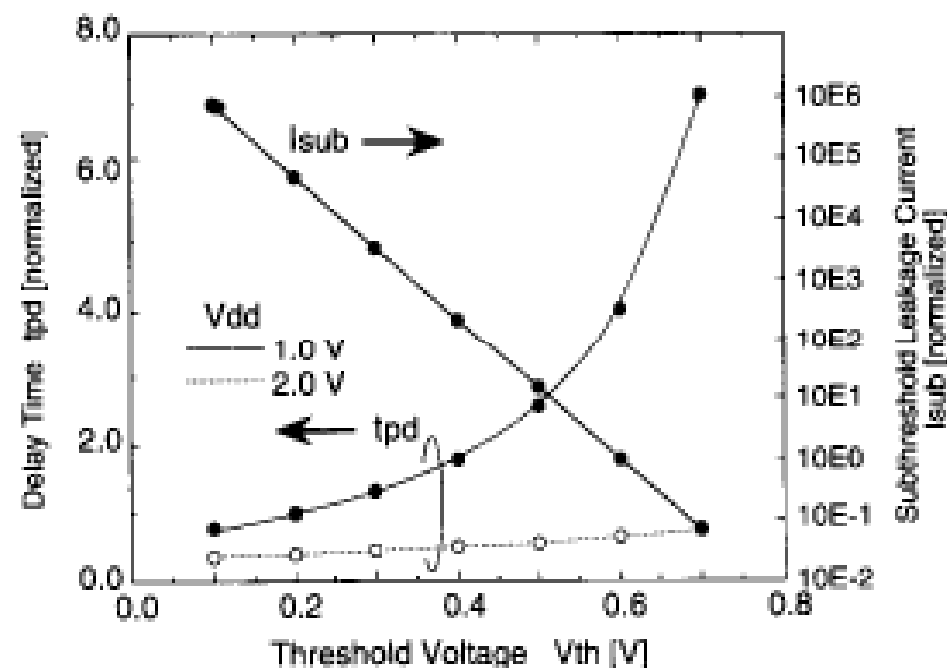
◆ MTCMOS : Multiple Threshold voltage CMOS

◆ Low power & low Energy

- $E_{TOT} = E_{STD} + E_{ACT} = P_{static} * t_{STD} + P_{dynamic} * t_{ACT}$
- Portable device : $t_{STD} \gg t_{ACT}$

◆ Basic circuit scheme

- Two different V_t
 - ❖ HVt (0.5~0.6V)
 - ❖ LVt (0.2~0.3V)
- Two operating mode
 - ❖ Active
 - ❖ Standby



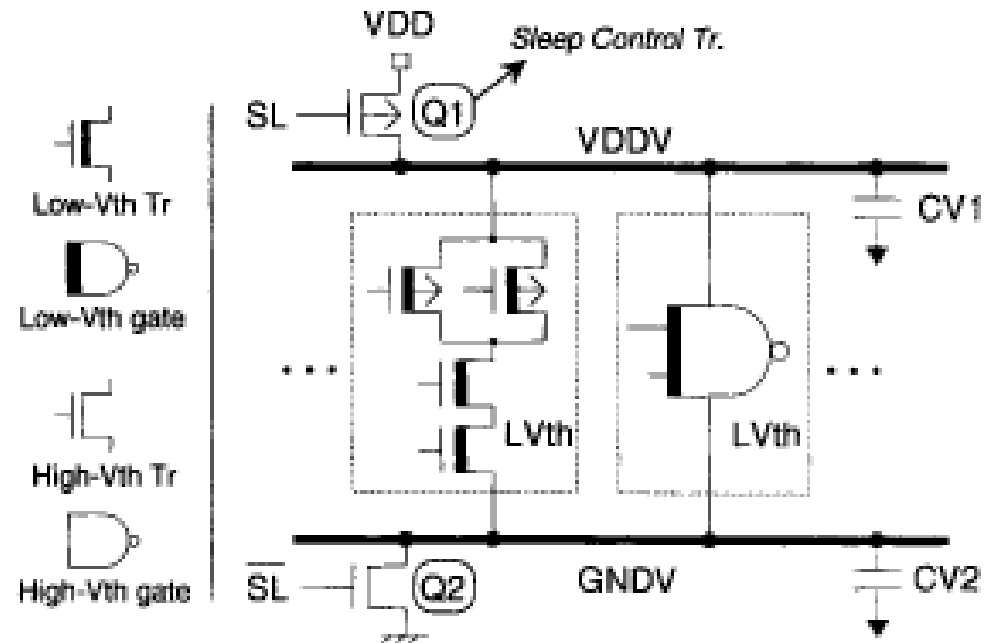
MTCMOS : Scheme

◆ Active mode

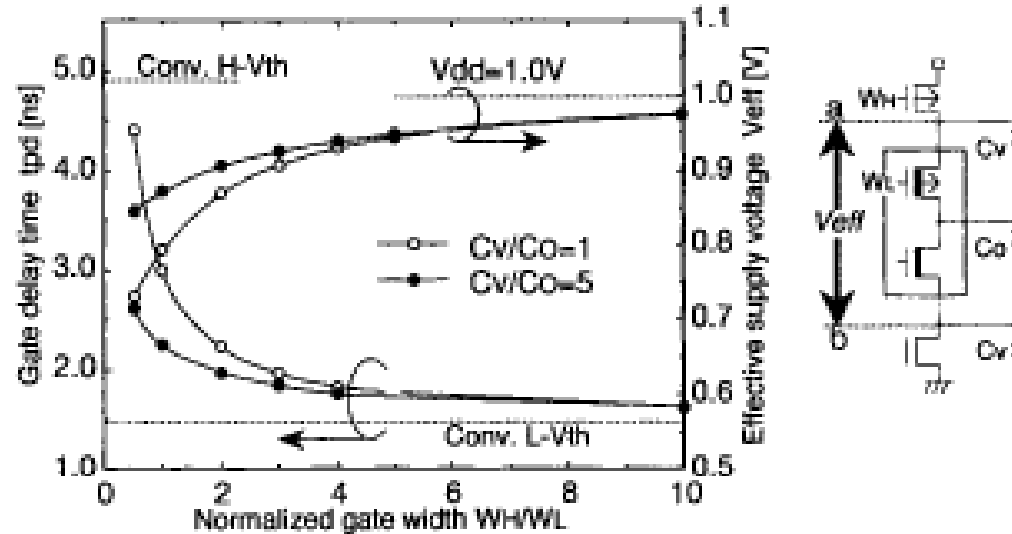
- $SL=1 / \underline{SL}=0$
- $V_{DDV} \approx V_{DD} / V_{GNDV} \approx V_{GND}$
- LVt operating frequency

◆ Standby mode

- $SL=0 / \underline{SL}=1$
- V_{DDV} & V_{GNDV} = floating
- HVt leakage



MTCMOS : Constraint



◆ Performance constraint according to

- Normalized foot/head switch size : W_H/W_L
- Normalized cap on VDDV/VGNDV : C_V/C_O

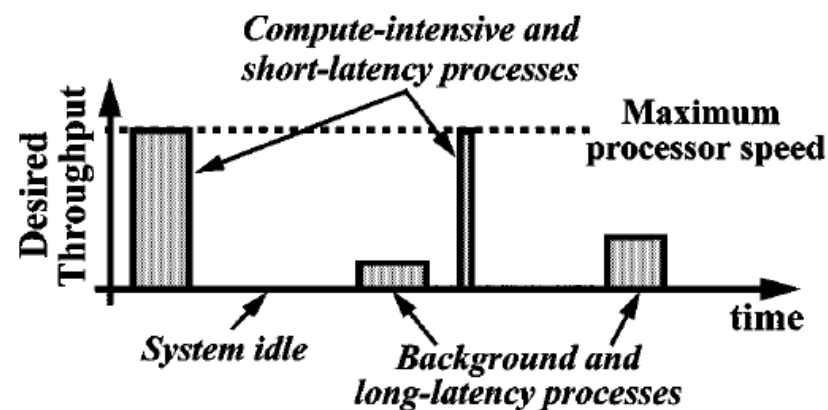
◆ Area penalty

- Relatively small because Head/Footswitches are shared by all logic gates on a chip (global foot switch)

DVFS : Basic Concept

◆ Basic concept

- $P_{\text{dynamic}} = CV_{\text{DD}}^2f$
- V_{DD} and frequency scaling simultaneously
- V_{DD} scaling
 - ❖ A best way to get low P_{dynamic} because $P_{\text{dynamic}} \propto V_{\text{DD}}^2$
- Frequency scaling
 - ❖ Operating frequency = throughput
 - ❖ Not all task requires maximum throughput
 - ❖ By controlling the frequency, SoC improves energy efficiency



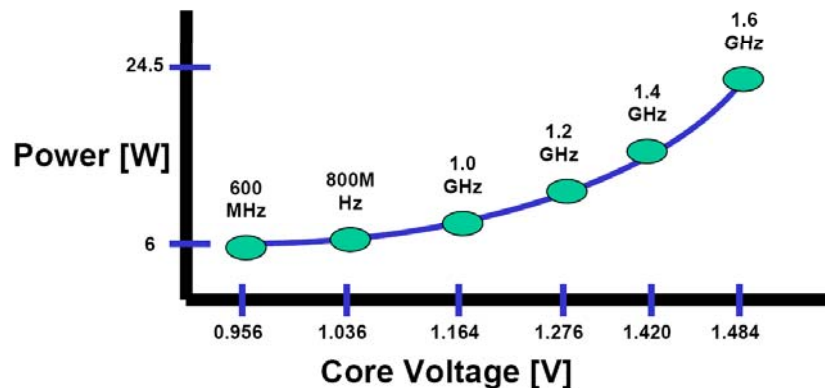
DVFS : Open loop VS. Closed Loop

◆ Open loop system

- Can not adapt to PVT variations
- Need more design margin
- Example
 - ❖ Enhanced SpeedStep technology of Intel

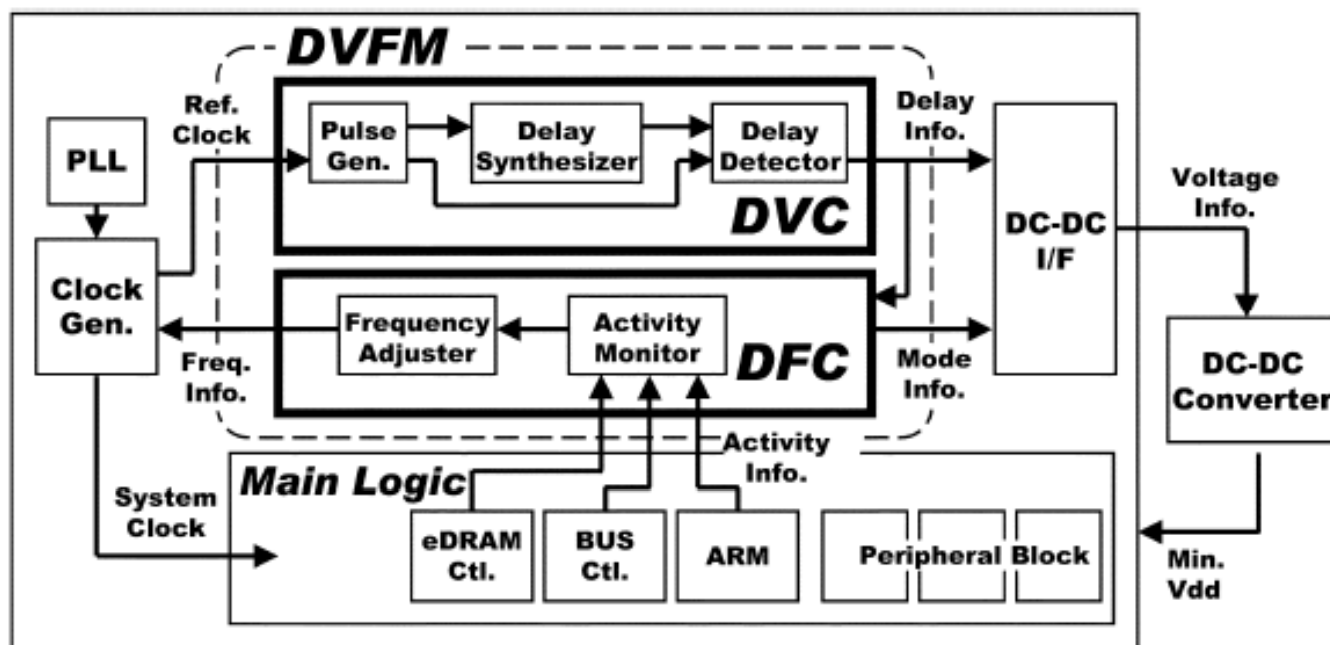
◆ Closed loop system

- Can adapt to PVT variations
- Need less design margin
- Example
 - ❖ Intelligent Energy Management technology of ARM
 - ❖ SmartReflex2 of TI OMAP processor



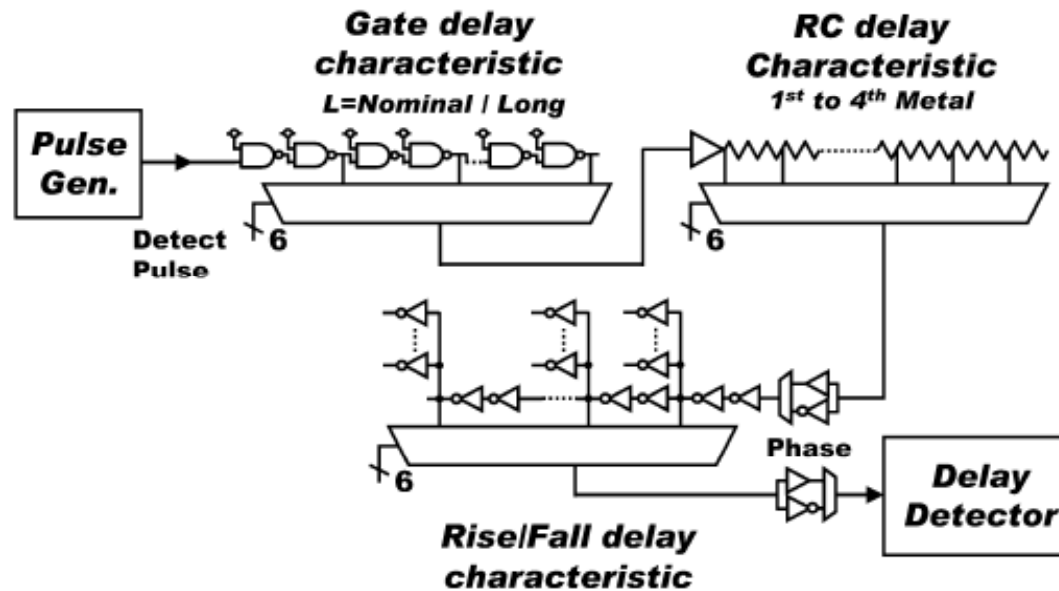
DVFS (SONY, PDA)

◆ Block Diagram



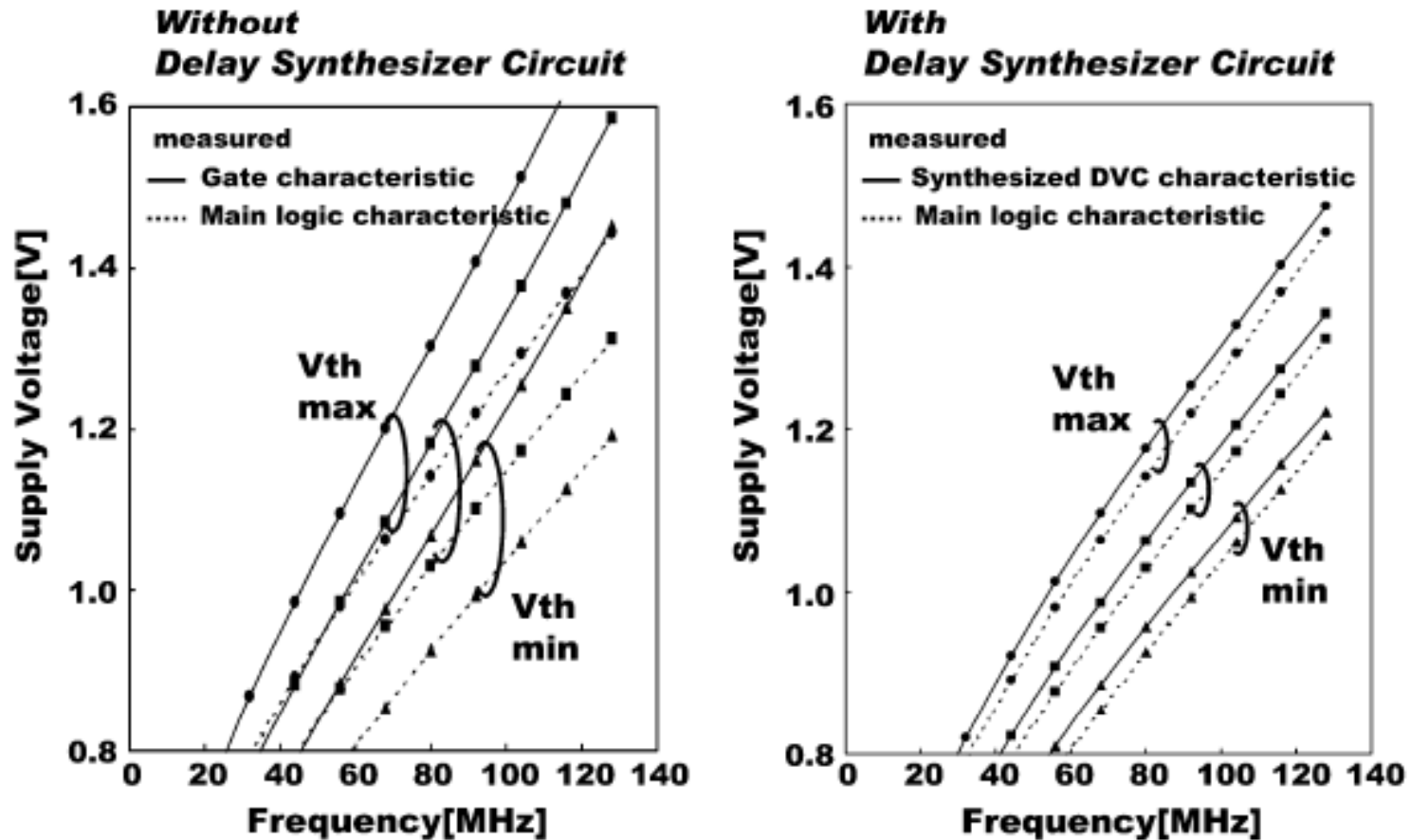
- Closed loop system

Delay Synthesizer Structure



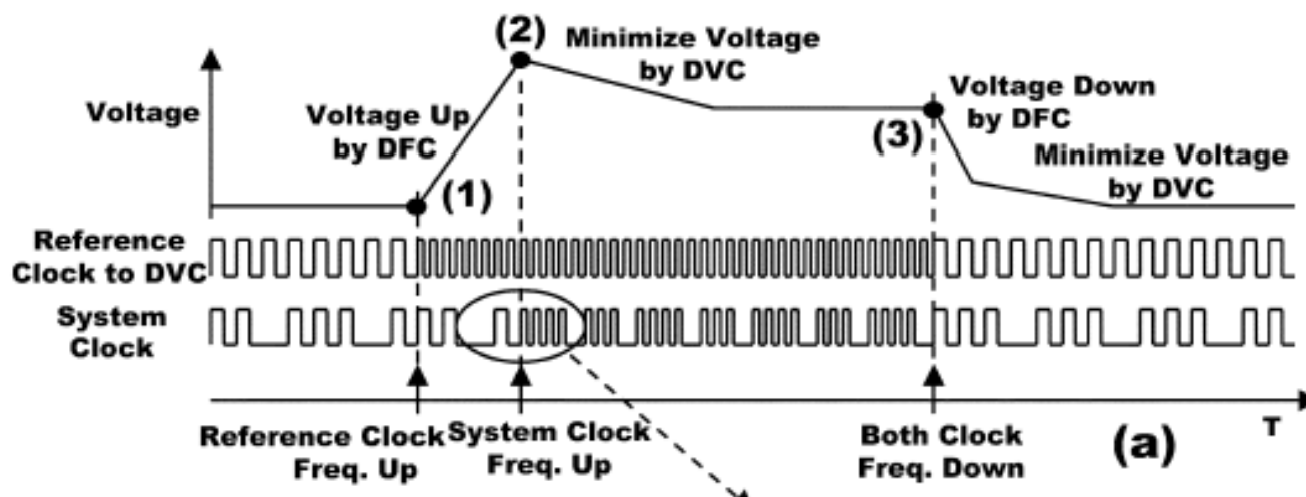
- Composed not only a simple transistor delay factor, but also wire delay and rise/fall delay
 - ❖ Gate delay component : one of nominal gate length and another of long gate length
 - ❖ RC delay component : wires from each of the four metal layers and its total length is 14mm

Delay Synthesizer Effect



Operation (DVC+DFC)

◆ Operation procedure



- Low → High : The main logic clock frequency is changed after the DVC confirms the voltage has increased enough
- High → Low : Both the DVC reference clock and the system clock are changed simultaneously

Device Level

VLSI
SYSTEM LAB.

Device Level Low Power Technique

◆ FinFET

- FinFET : Vertical structure
 - ❖ Planar MOSFET width = FinFET height

- $\sigma(V_T) \propto T_{ox} N_A^{0.25} (LW)^{-0.5}$
 - ❖ As scaling goes on, variation of planar MOSFET get worse

➢ V_{DD} scaling is impossible

❖ However, FinFET's $\sigma(V_T)$ doesn't degraded

- FinFET width doesn't occupy the active area
- As scaling goes on, $L \cdot W$ of FinFET can be maintained
- V_{DD} scaling is possible \Rightarrow **low power !!**

	Planar MOSFET	FinFET	
L	$1/\alpha$	$1/\alpha$	$1/\sqrt{\alpha}$
W	$1/\alpha$	α	$\sqrt{\alpha}$
W/L	1	α^2	α
LW	$1/\alpha^2$	1	1
A_{vt}	$1/\sqrt{\alpha} (1/\alpha)$	$1/\sqrt{\alpha} (1/\alpha)$	$1/\sqrt{\alpha} (1/\alpha)$
$\sigma(V_T)$	$\sqrt{\alpha} (1)$	$1/\sqrt{\alpha} (1/\alpha)$	$1/\sqrt{\alpha} (1/\alpha)$
V_{DD}	$\sqrt{\alpha} (1)$	$1/\sqrt{\alpha} (1/\alpha)$	$1/\sqrt{\alpha} (1/\alpha)$
I_{DS}	$\sim \alpha^{1.1}$	$\sim \alpha^{1.9}$	$\sim \alpha^{0.9}$
$\tau (MOS)$	$\sim \alpha^{-2.1}$	$\sim \alpha^{-1.9}$	$\sim \alpha^{-0.9}$
$P (= V_{DD} I_{DS})$	$\sim \alpha^{1.6}$	$\sim \alpha^{1.4}$	$\sim \alpha^{0.4}$
$P_r (MOS)$	$\sim \alpha^{-0.5}$	$\sim \alpha^{-0.5}$	$\sim \alpha^{-0.5}$
W_{min}/L_{min}	$F = 45 \text{ nm}$ ($\alpha = 1$)	45/45 nm	45/45 nm
	$F = 11 \text{ nm}$ ($\alpha = 4$)	11/11 nm (aspect ratio = 1)	180/11 nm (16)

$A_{vt} \propto t_{ox} N_{sub}^{0.25}$, $\sigma(V_T) = A_{vt}/\sqrt{LW}$, $I_{DS} = \beta (V_{DD} - V_T)^{1.2}$ for constant N_{sub} , $\tau (MOS) = V_{DD} C_G / I_{DS}$

OMAP Processor

VLSI
SYSTEM LAB.

OMAP Processor

◆ OMAP Processor

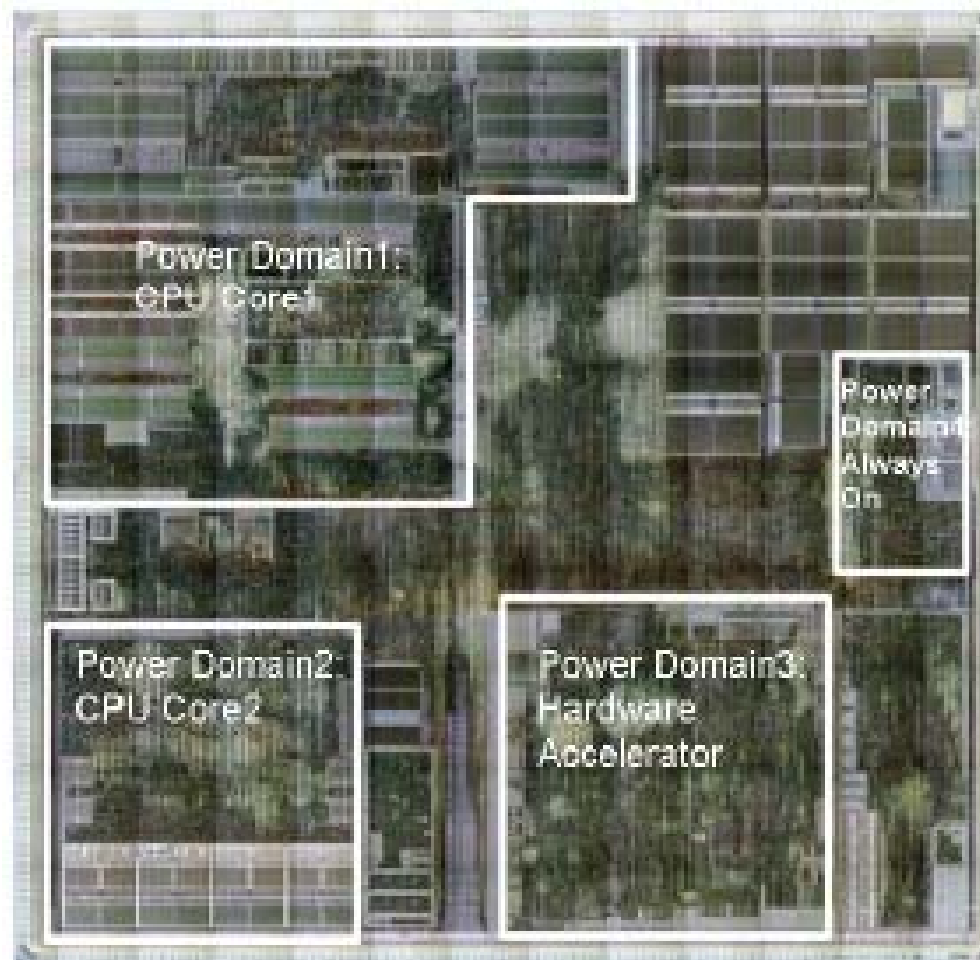
- Dual core platform
- Multimedia hardware accelerators for video and graphics
- Frame buffers
- Various dedicated and general purpose interfaces

◆ Power saving mode

- Idle (Clock stopped)
- Retention for low leakage
- Fast re-start and power-off mode
- Power gating technique

Power Domains

- ◆ 5 power domains
 - Processor core 1
 - Processor core 2
 - Hardware accelerator (Graphic)
 - Always on
 - Rest of the chip (including the interconnects and various peripherals)



Power Gating

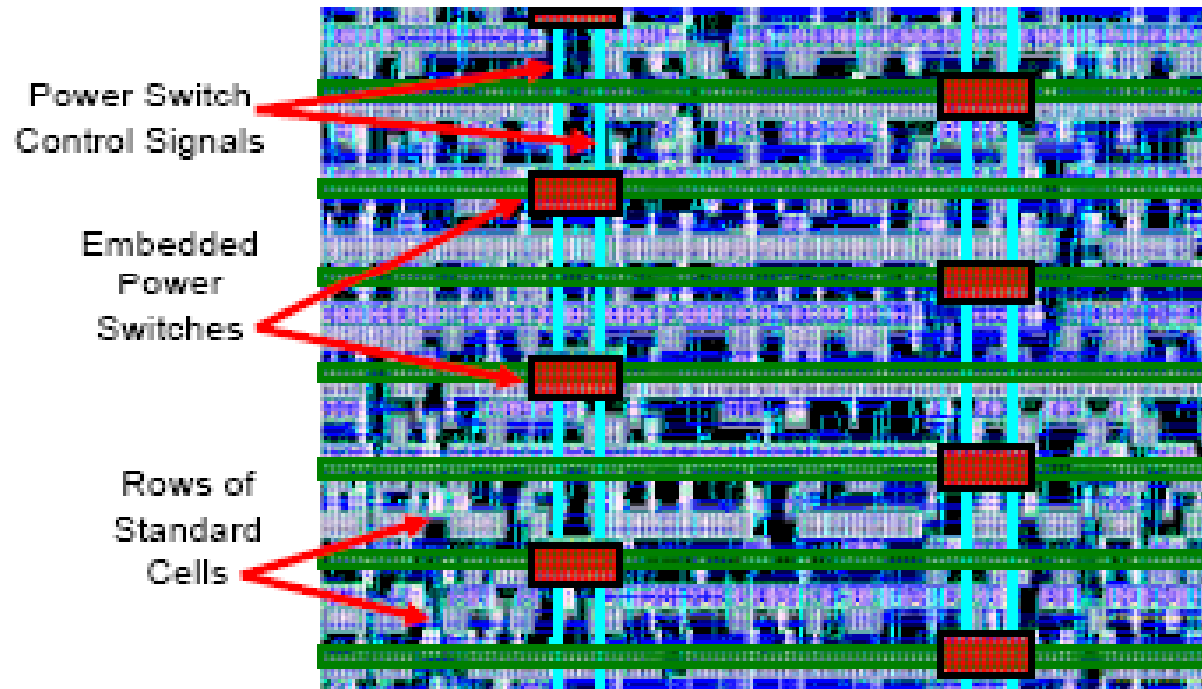
◆ Power gating

- Global mesh built with the highest metal layer distributes power and ground across the chip
- Local mesh is broken to reflect the power domain partitioning
- Power switch makes connection between global mesh and local mesh according to operating modes and switch control
 - ❖ If power domain is on, its power switches connect its local plane to the global plane., i.e., the constant power supply
 - ❖ Otherwise that plane drifts to a potential near ground

◆ Power switch

- Embedded in power domains
 - ❖ by placing power switches at a regular pitch in a staggered manner
 - ❖ by placing power switches around hard Ips
- Header switch
 - ❖ 90um PMOS with 200uA current driving capability at worst case
 - ❖ Multiple fingers and redundant vias

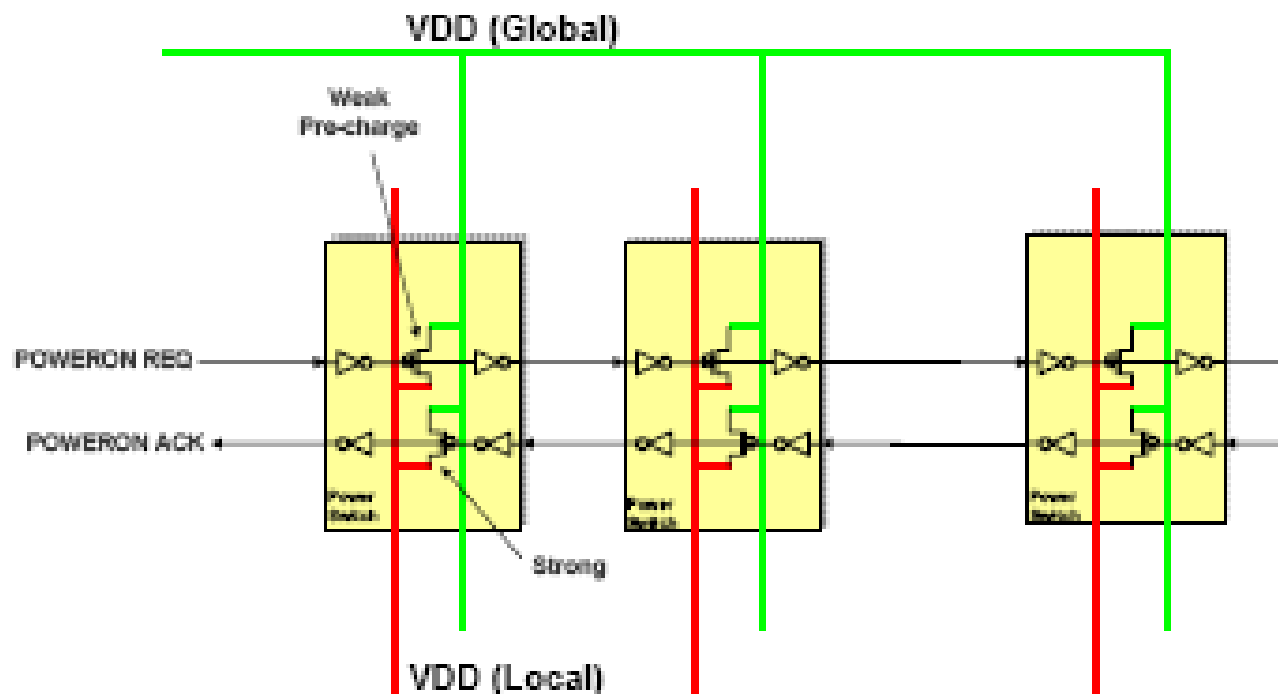
Embedded Power Domains



◆ Other power management cells

- Retention flip-flops
- Constantly powered buffers to transport critical signals through a power domain potentially off
- Isolation cells to prevent the propagation of a non-state

Power Switching Control



◆ Current surges and dynamic IR drop

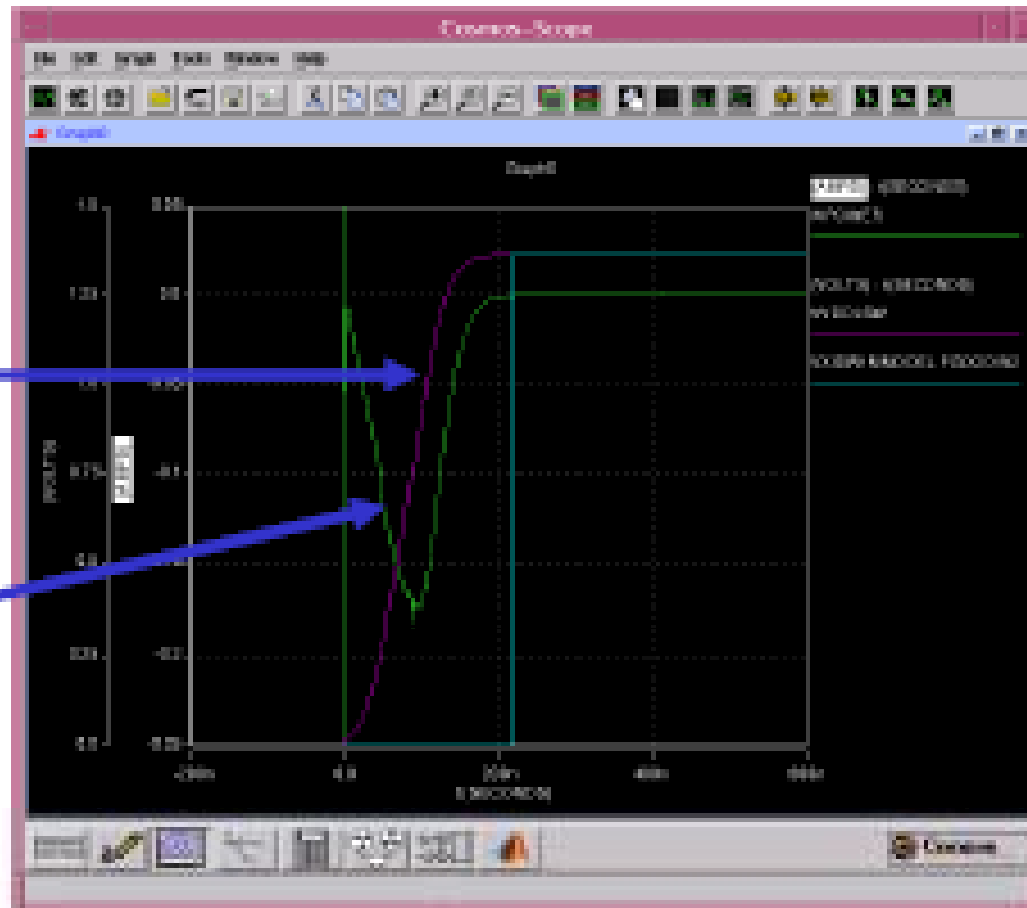
● Two-pass turn-on mechanism

- ❖ Weak PMOS to sinks low current for power restore: Turn-on first
- ❖ Strong PMOS to deliver current for normal operation: Turn-on next

Current Surge and Power Restore

Local VDD restored
within 200ns

170mA current peak

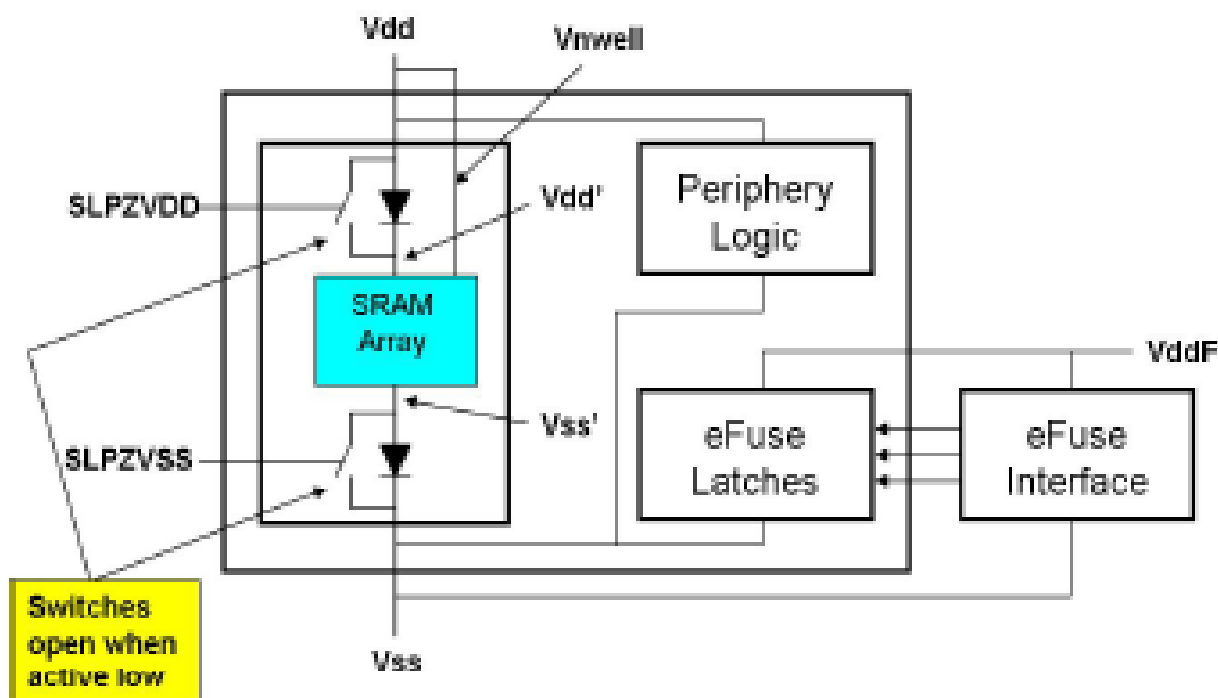


Leakage Current Reduction

◆ In off mode

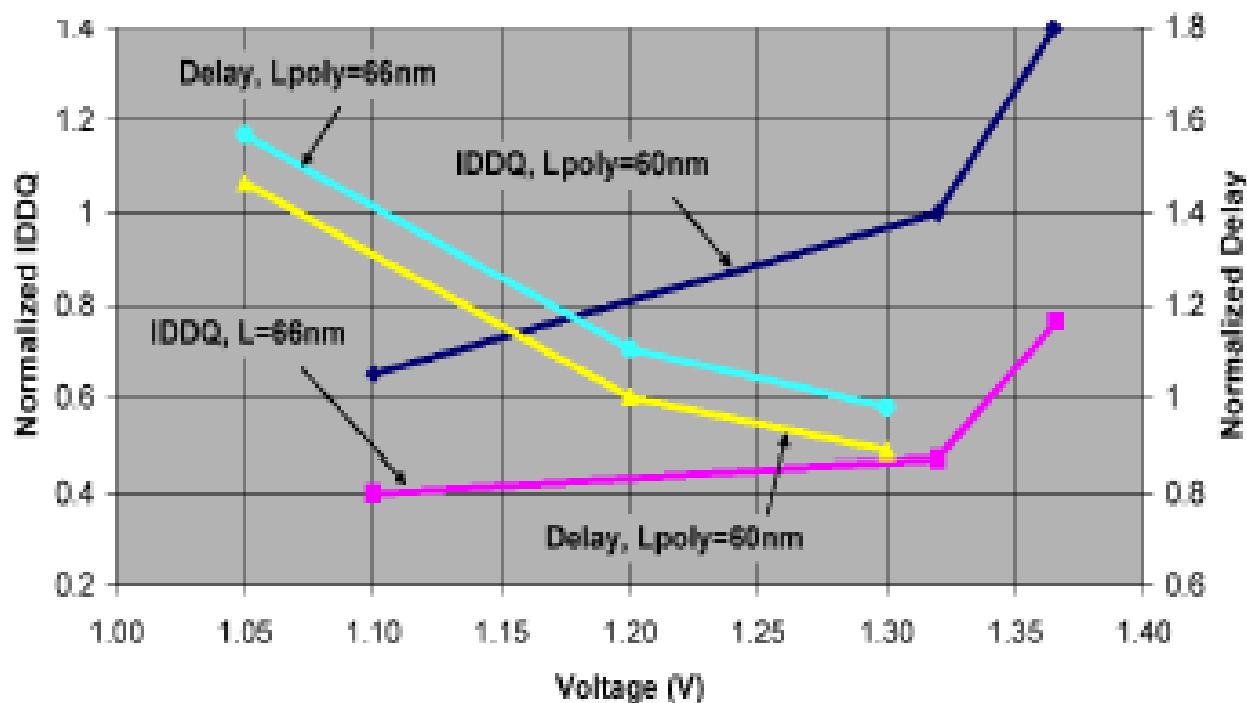
- Leakage current comes from power switches and power management cells
- 4 power switches per Kgate
 - ❖ ~40X leakage reduction

SRAM Retention



- ◆ Footer and header diodes
 - In active mode, the diodes are bypassed
 - During retention mode, one diode is enabled and
 - ❖ Field across the array is reduced
 - ❖ Reverse body bias
 - Leakage saving (x2)

Dual Gate Length



- ◆ Dual gate length
 - Standby mode: 30% leakage reduction
 - Active mode: active leakage current saving: very useful if many blocks are idle in active mode
- ◆ Vdd scaling during the slow active mode
 - 300mV scaling: 2X leakage reduction



Summary

VLSI
SYSTEM LAB.

Summary

◆ Green SoC design

⇒ Low power & process variation tolerant SoC design

$$\text{◆ } P = \underbrace{P_{sw} + P_{sc}}_{P_{dynamic}} + \underbrace{P_{sub} + P_{gate} + P_{junc}}_{P_{static}}$$

◆ Power and performance : Trade-off

◆ Low power design

- **Architecture and algorithm level** : parallelism, pipe line
- **Block and logic level** : workload monitoring, V_{DD} /frequency scheduling
- **Circuit level**
 - ❖ Long channel : Reduce I_{leak} by using V_{TH} roll off ($V_{TH} \uparrow$)
 - ❖ Stacked MOSFET : Reduce I_{leak} by using body effect ($V_{TH} \uparrow$) & negative V_{GS}
 - ❖ Dual V_{DD} : Use low V_{DD} at non-critical path
 - ❖ Dual V_{TH} : Use low V_{TH} at non-critical path
 - ❖ MTCMOS: Use high V_{TH} sleep TR (low leakage in stand-by mode) & low V_{TH} logic (high performance in active mode)
 - ❖ DVFS : Reduce dynamic power by controlling both V_{DD} & frequency
- **Device level** : FinFET