Circuit and Design Automation Techniques for Leakage Minimization of CMOS VLSI Circuits

Massoud Pedram

University of Southern California
EE Dept.
pedram@ceng.usc.edu

Samsung Microelectronics
Seoul, South Korea
Oct 27, 2006

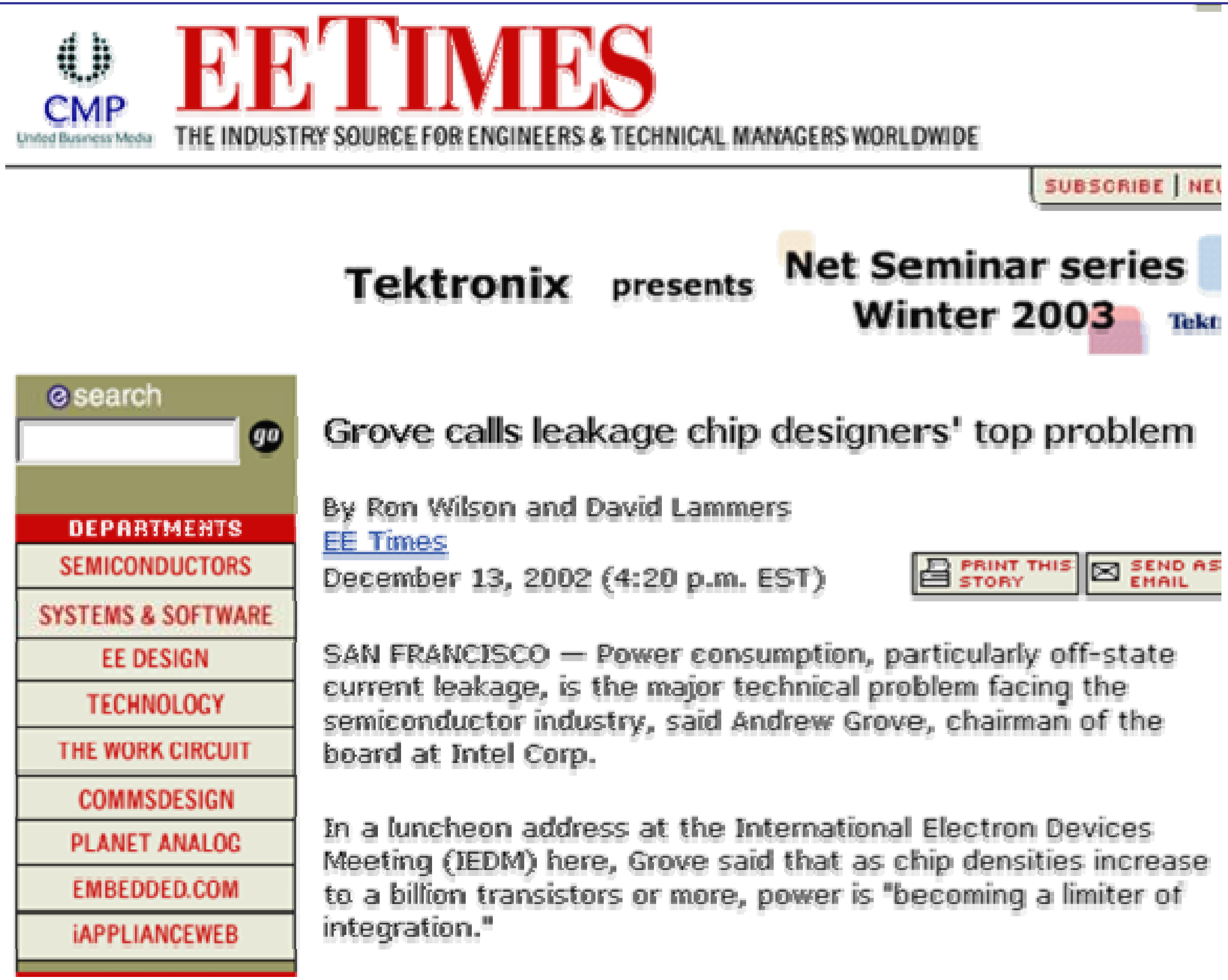
Realities

- Power has emerged as the #1 limiter of design performance beyond the 65nm generation.
- Dynamic and static power dissipation limit achievable performance due to fixed caps on chip or system cooling capacity.
- Power related signal integrity issues (IR drop, L di/dt noise) have become major sources of design re-spins.

Transistors (and silicon) are free.
Power is the only real limiter.
Optimizing for frequency and/or area may achieve neither.

Pat Gelsinger, Senior Vice President & CTO, Intel

Industry Views (Intel)



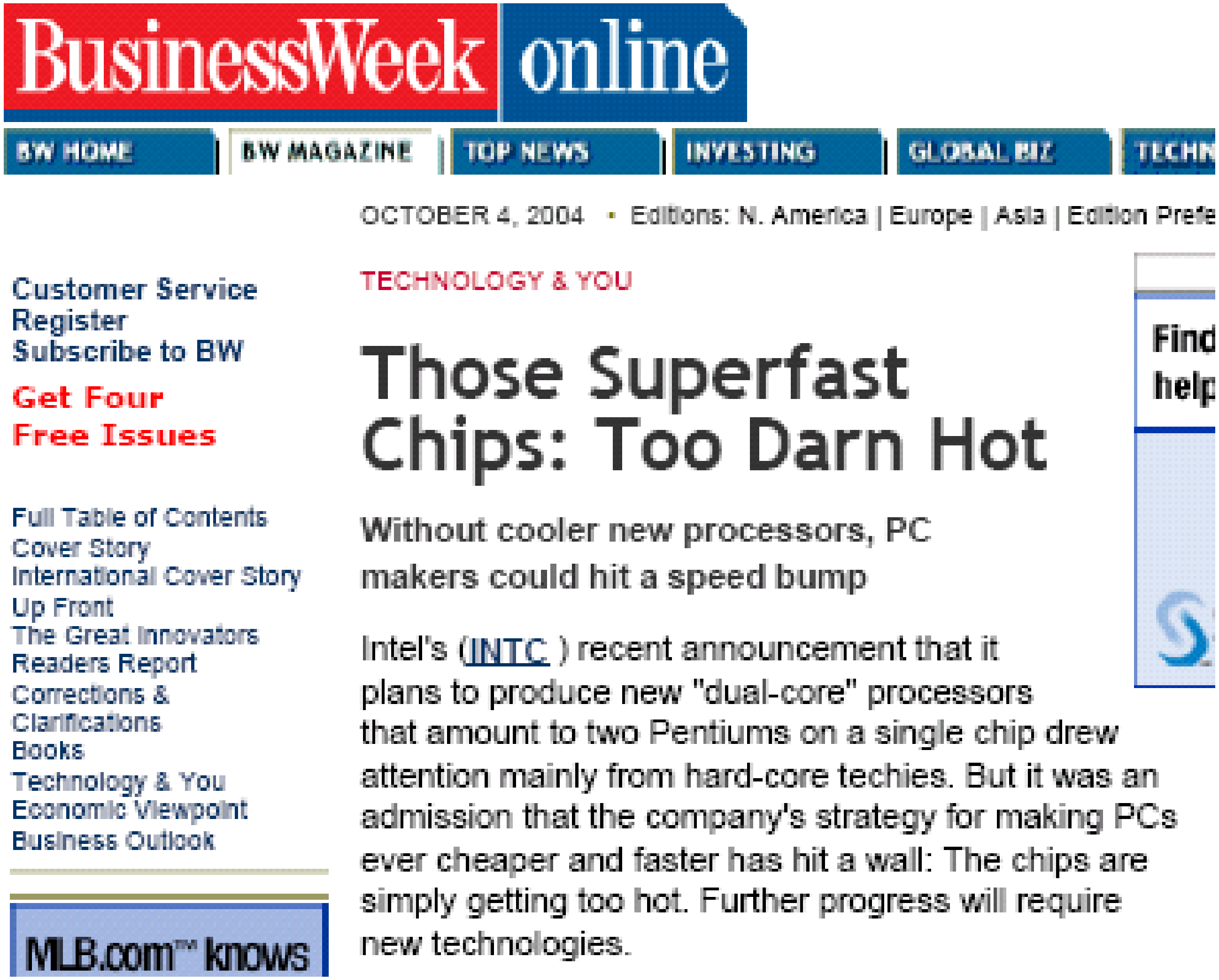
Grove calls leakage chip designers' top problem

By Ron Wilson and David Lammers
EE Times
December 13, 2002 (4:20 p.m. EST)

SAN FRANCISCO — Power consumption, particularly off-state current leakage, is the major technical problem facing the semiconductor industry, said Andrew Grove, chairman of the board at Intel Corp.

In a luncheon address at the International Electron Devices Meeting (IEDM) here, Grove said that as chip densities increase to a billion transistors or more, power is "becoming a limiter of integration."

Industry Views



OCTOBER 4, 2004 • Editions: N. America | Europe | Asia | Edition Prefe

TECHNOLOGY & YOU

Those Superfast Chips: Too Darn Hot

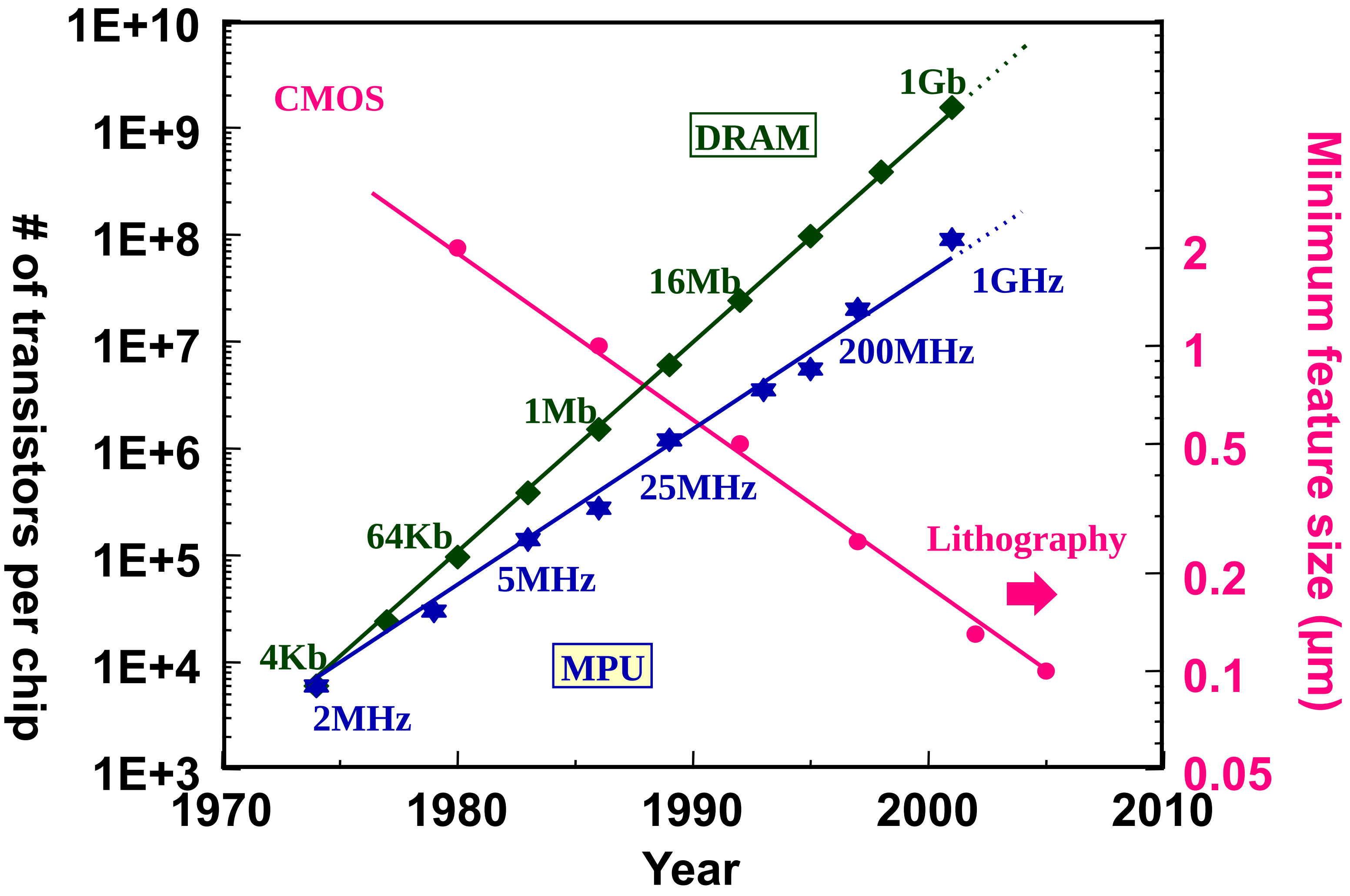
Without cooler new processors, PC makers could hit a speed bump

Intel's (INTC) recent announcement that it plans to produce new "dual-core" processors that amount to two Pentiums on a single chip drew attention mainly from hard-core techies. But it was an admission that the company's strategy for making PCs ever cheaper and faster has hit a wall: The chips are simply getting too hot. Further progress will require new technologies.

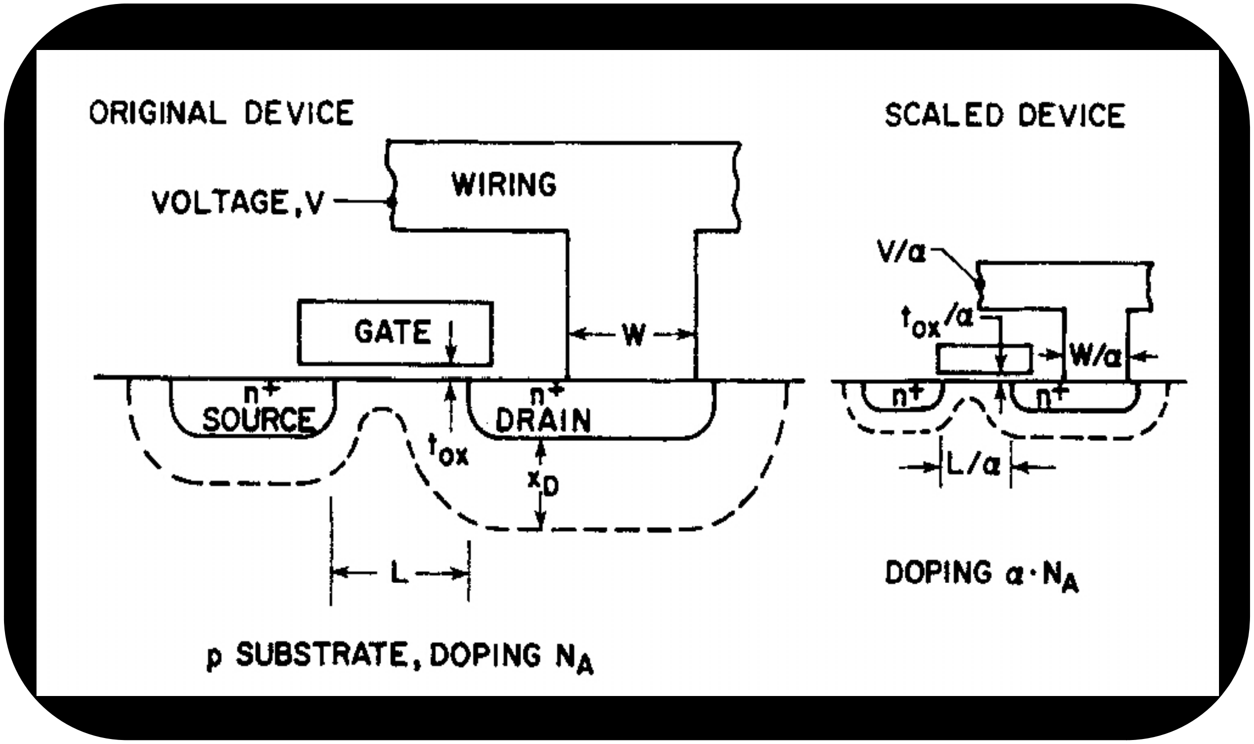
Outline

- Technology Trends
- Power Dissipation 101
- Sources of Leakage in CMOS VLSI Circuits
- Circuit and Device Techniques for Leakage Minimization
- Conclusion

Circuit Density and Performance Trends



Constant-Field MOSFET Scaling



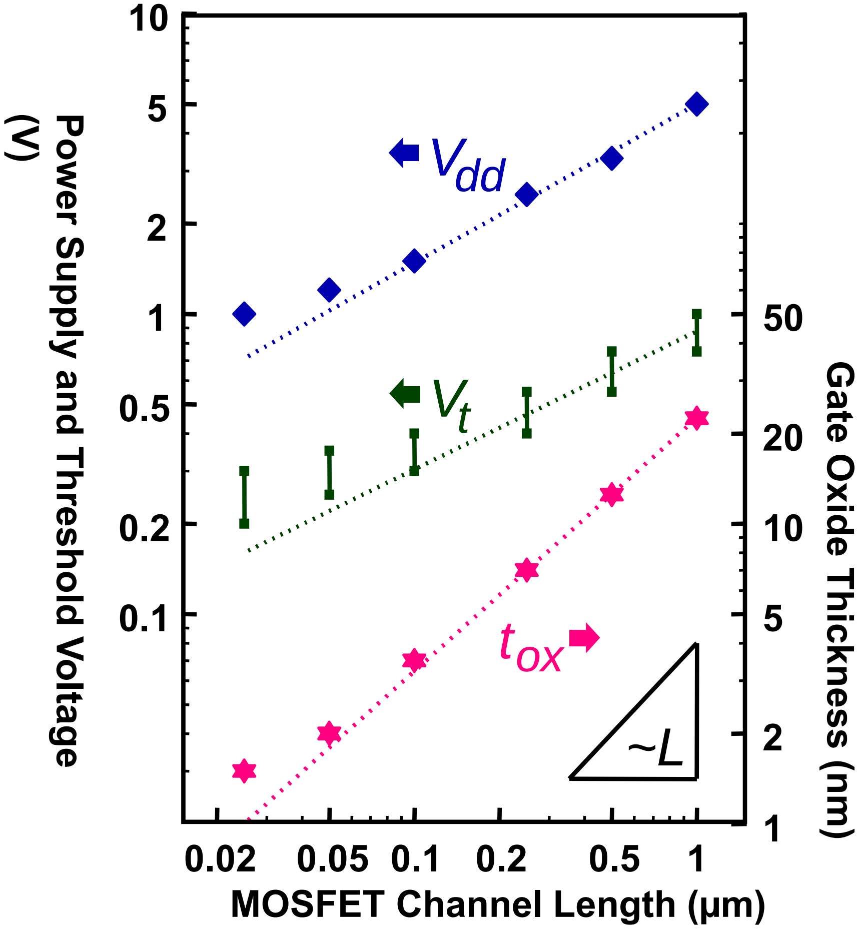
Source: B. Davari, IBM, 1999

- L, W, t_{ox}, x_D, V_{DD}, V_T, C, I, and τ scale by $1/\alpha$.
- Area, power dissipation, and charges scale by $1/\alpha$.
- Power dissipation and charges per unit area do not scale.

V_{dd}, V_{th} and t_{ox} Scaling

- V_{dd} scaling needed to reduce power and maintain device reliability
  - V_{th} scaling needed to maintain switching speeds
  - t_{ox} scaling needed to maintain the current drive and keep V_{th} variations under control when dealing with short-channel effects.
- V_{th} does not scale much since the inverse subthreshold slope, which represents transistor turn-off rate, is dominated by temperature, not V_{th} or V_{dd}.

Source: Y Taur, 2002

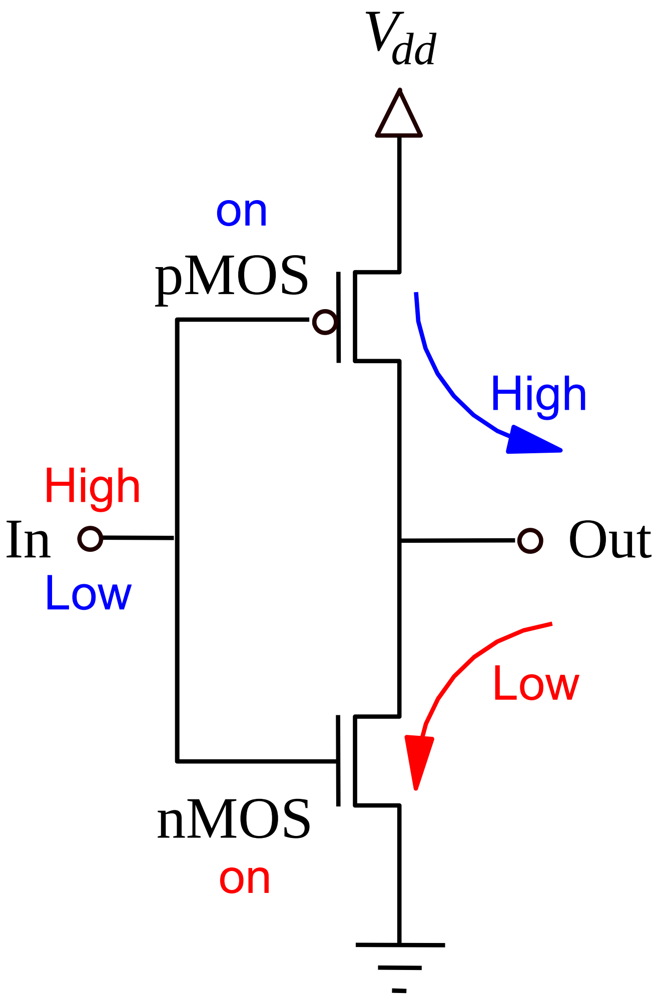


Non-CMOS, Non-Si Replacements?

- Single Electron Transistor
- Resonant Tunneling Diode
- Josephson Junctions
- Carbon Nano-Tubes

- No credible candidate on the horizon that shows the promise to replace CMOS ULSI yet.

The Beauty of CMOS Logic



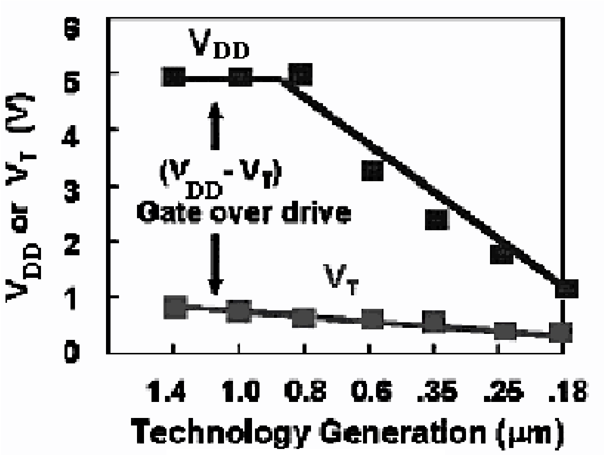
- Negligible standby power dissipation.
- Uses most of the power budget to switch output state
  - Low leakage.
- Highly scalable.
- Low energy dissipation per output value change
  - Low activity factor.
- The only known logic circuit style with such properties.

Limiting factors to CMOS scaling

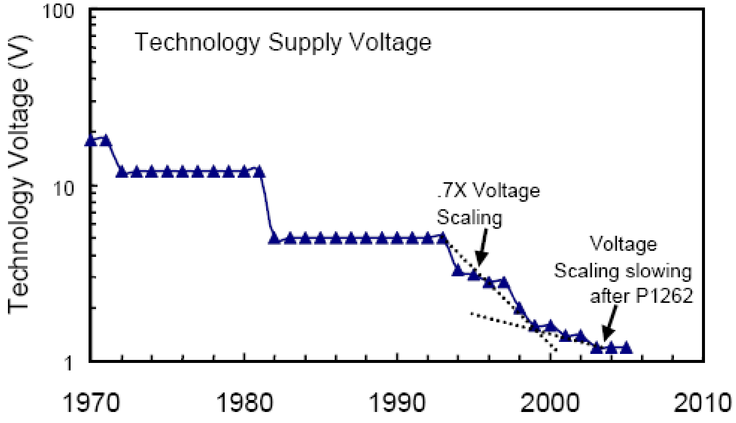
- Fundamental non-scaling effects are caused by the fact that neither the thermal voltage kT/q nor the silicon bandgap changes with scaling.
  - The first results in non-scaling of the inverse substhreshold voltage slope.
  - The latter results to non-scalability of built-in junction potential, depletion layer width, and short channel effects.
- Because of the field dependence of the carrier mobility, the gate speed will not improve linearly with dimensional scaling.
- There is adverse impact on device reliability due to high electric field stress.

CMOS Scaling Summary

- Scaling increases:
  - Transistor density and functionality
  - Speed of operations
  - Power density and parametric variability.
- Maximum integration density is limited by the power density while maximum circuit speed is limited by the parametric variability.
- Situation is expected to become worse since voltage scaling is slowing/stopping.



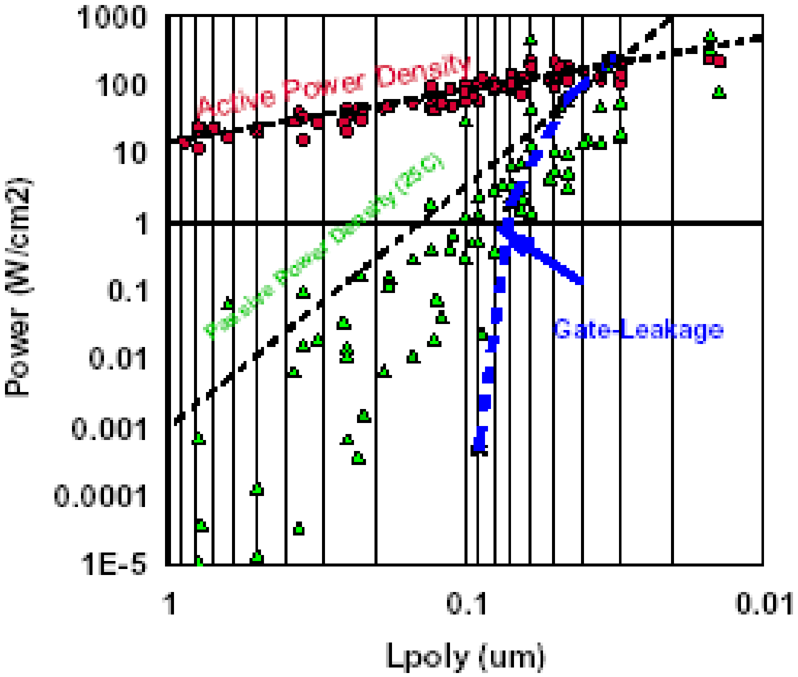
Source: Intel



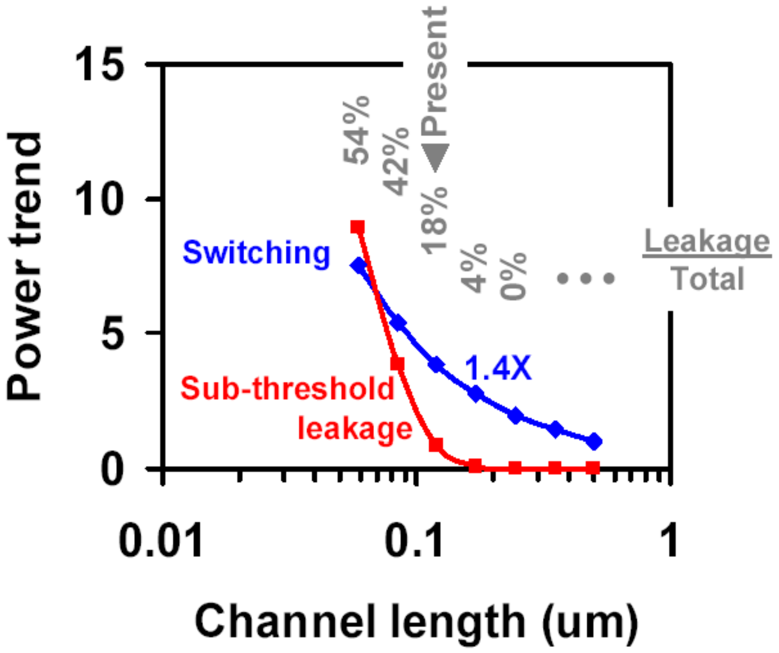
Source: Scott Thompson, TI

Leakage vs. Total Power

- A significant part of total power at 90nm and below
  - Sub-threshold leakage is increasing due to V_{th} scaling.
  - Gate leakage is increasing due to gate oxide scaling.
- Leakage in active mode is a major issue.

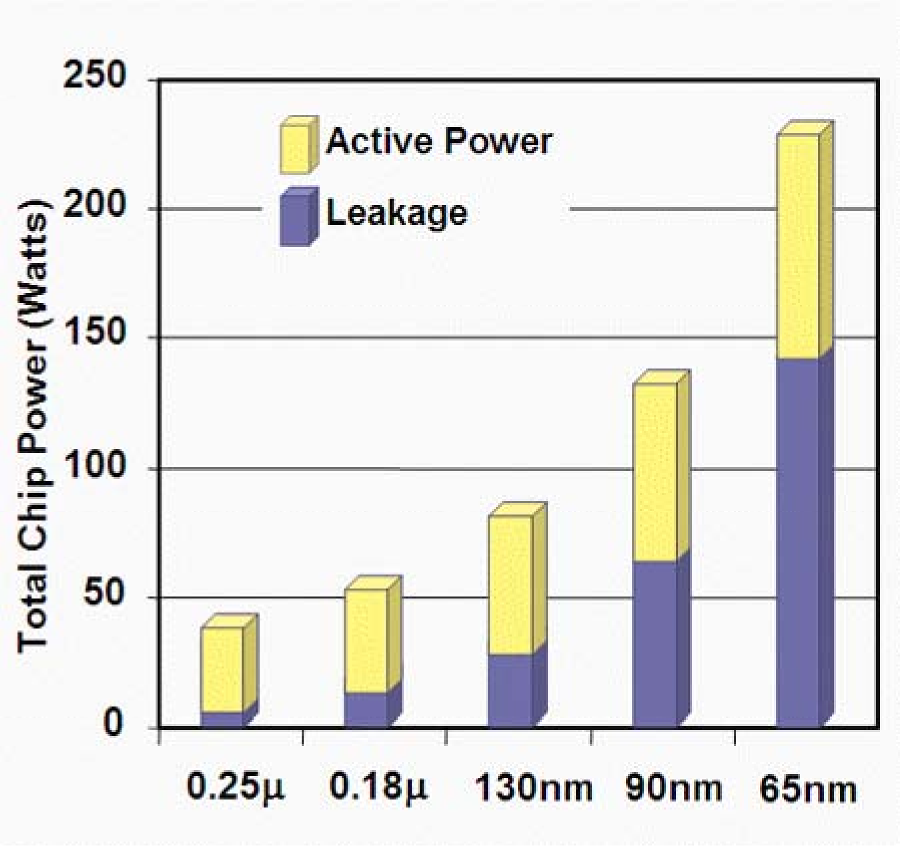


Source: Puri, et al 2003



Source: Chandrakasan, et al 2002

Trends in Power Across Process Technologies



Source EE-Times, August 06, 2004 Issue

Outline

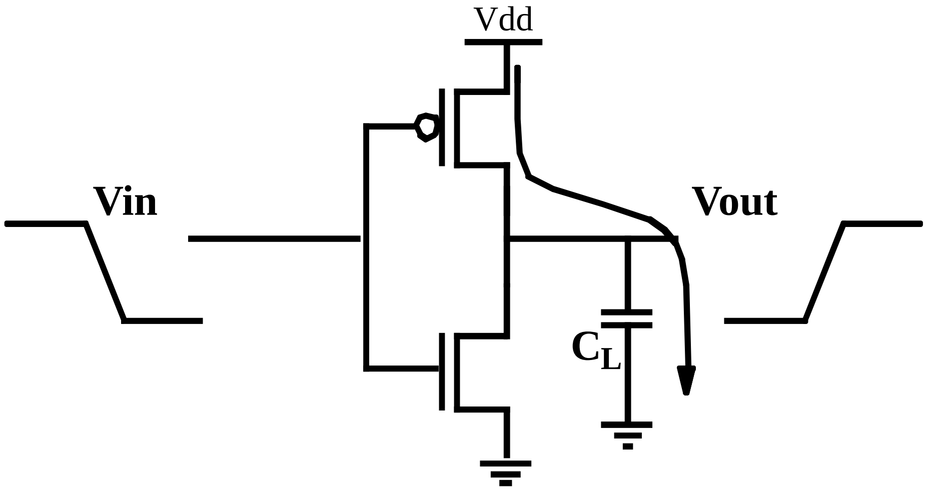
- Technology Trends
- Power Dissipation 101
- Sources of Leakage in CMOS VLSI Circuits
- Circuit and Device Techniques for Leakage Minimization
- Conclusion

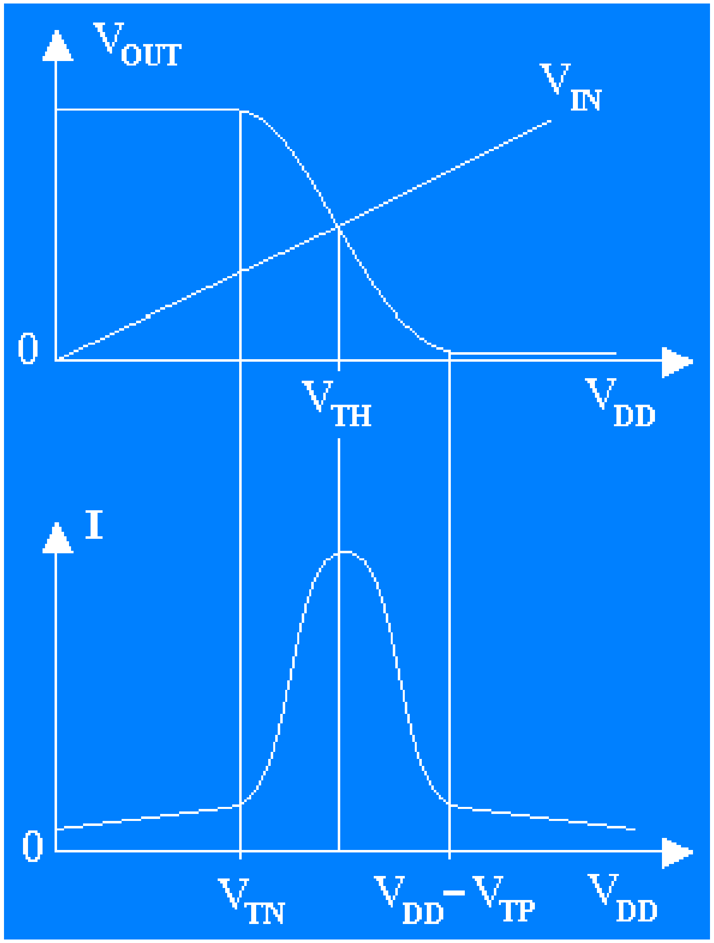
Dynamic Power Dissipation

- The dynamic power dissipation is a function of:
  - Frequency
  - Capacitive loading
  - Voltage swing
  - Activity factor.

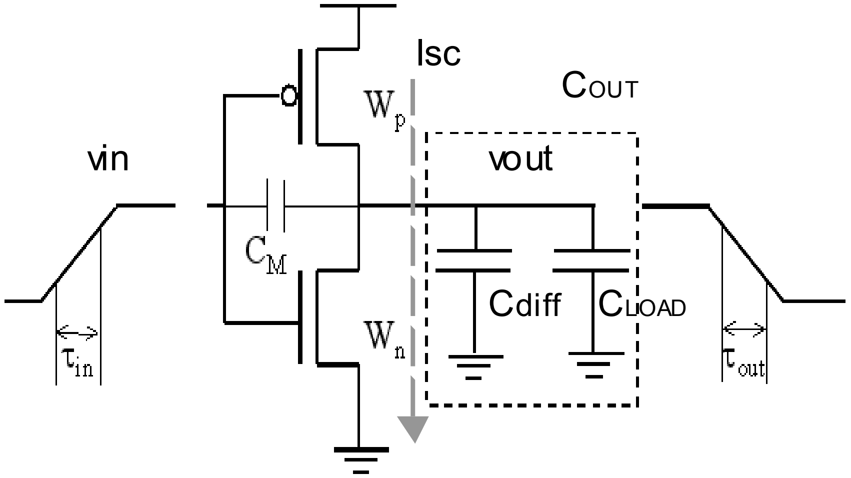
$$\text{E} = \text{Energy/transition} = \frac{1}{2} \cdot C_L \cdot V_{dd}^2$$

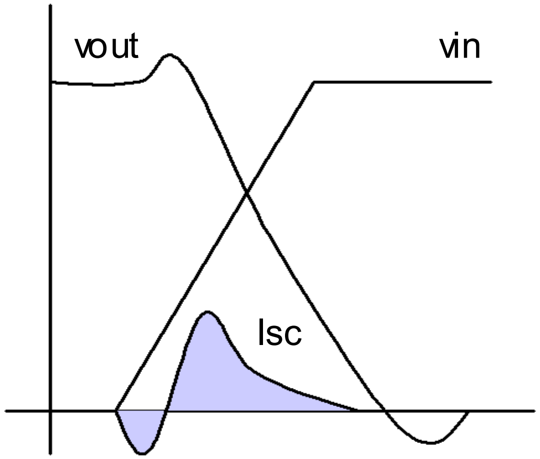
$$\text{P} = \text{Power} = E \cdot f \cdot a = \frac{1}{2} \cdot C_L \cdot V_{dd}^2 \cdot f \cdot a$$

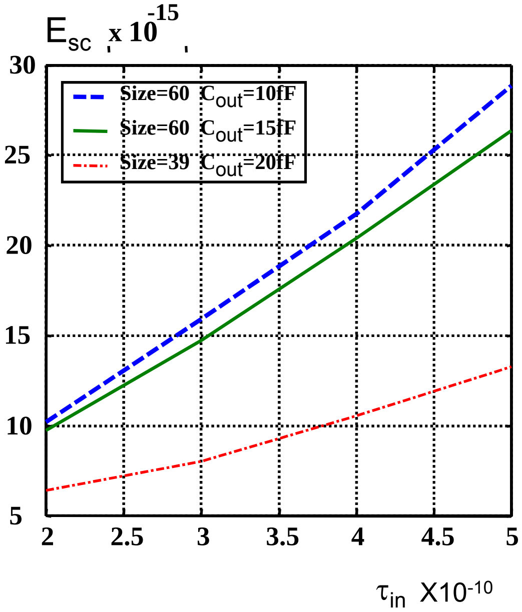




Short Circuit Power Dissipation







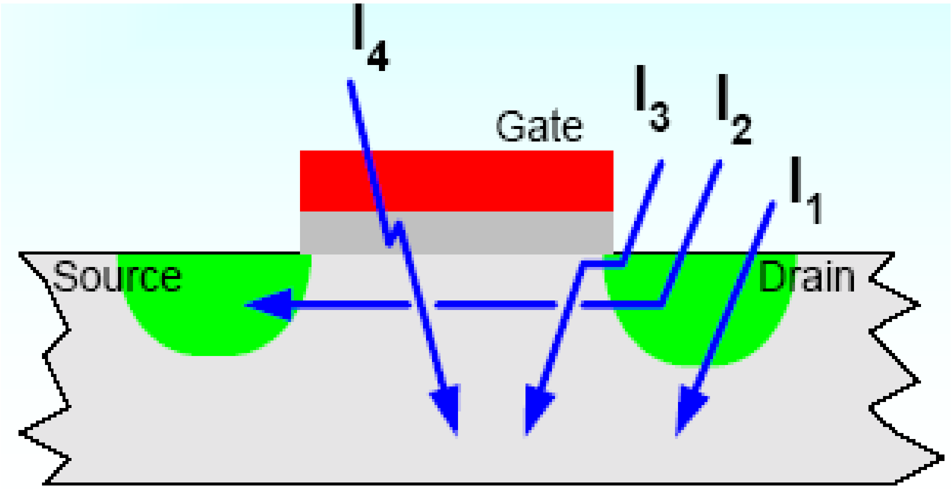
$$E_{sc}\left(\tau_{in}, W, C_{out}\right) = \sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} m_{ijk}\frac{W^{i}\tau_{in}^{\ j}}{C_{out}^{\ k}}V_{DD}$$

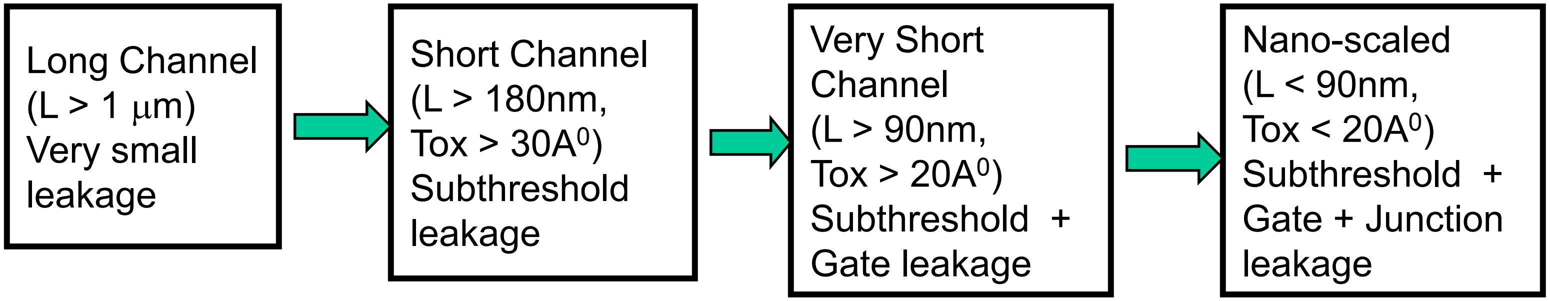
Outline

- Technology Trends
- Power Dissipation 101
- Sources of Leakage in CMOS VLSI Circuits
- Circuit and Device Techniques for Leakage Minimization
- Conclusion

Leakage Components in Bulk CMOS

- I_1 Diode reverse bias current
- I_2 Subthreshold current
- I_3 Gate induced drain leakage
- I_4 Gate oxide tunneling





Long Channel (L > 1 μm) Very small leakage

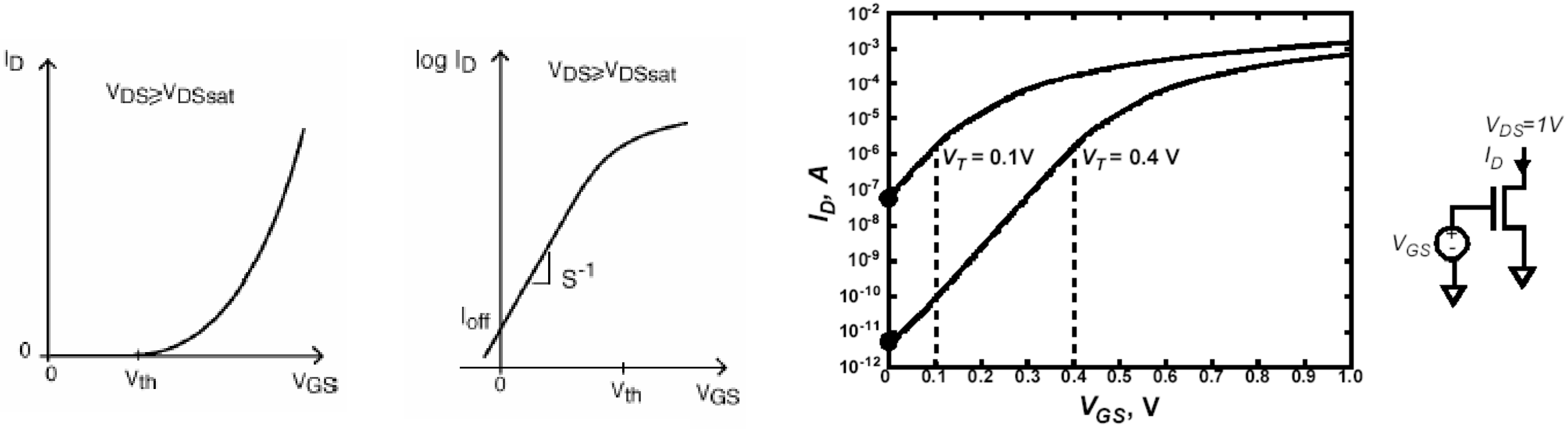
Short Channel (L > 180nm, Tox > 30A^0) Subthreshold leakage

Very Short Channel (L > 90nm, Tox > 20A^0) Subthreshold + Gate leakage

Nano-scaled (L < 90nm, Tox < 20A^0) Subthreshold + Gate + Junction leakage

Subthreshold Leakage Current

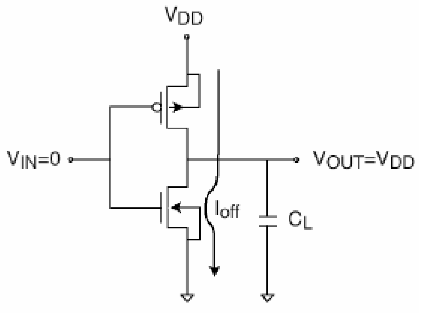
I_2: Transfer characteristics of MOSFET for V_{GS} near V_{th}:



$$I_{sub} = \frac{W}{L}\mu_e v_T^2 C_{sth}\, \mathrm{e}^{\frac{V_{GS}-V_{th}+\eta V_{DS}}{n v_T}}\left(1-\mathrm{e}^{\frac{-V_{DS}}{v_T}}\right) \propto e^{\frac{V_{GS}-V_{th}+\eta V_{DS}}{n v_T}} = 10^{\frac{V_{GS}-V_{th}+\eta V_{DS}}{S}}$$

- The inverse subthreshold slope, S, is equal to the voltage required to increase I_D by 10X, i.e., $S = \frac{nkT}{q}\ln 10$
  - If n = 1, S = 60 mV/dec at 300 K
  - We want S to be small to shut off the MOSFET quickly
  - In well designed devices, S is 70 - 90 mV/dec at 300 K.

Modeling Subthreshold (I_{sub}) and off (I_{off}) Currents



- Increases exponentially with reduction in $V_{th.}$
- Modulation of V_{th} in a short channel transistor.
  - L ↓ ⇒ V_{th} ↓: "V_{th} Rolloff"
  - V_{DS} ↑ ⇒ V_{th} ↓:"Drain Induced Barrier Lowering"
  - V_{SB} ↑ ⇒ V_{th} ↑: "Body Effect".
- If V_{DS} = 0 ⇒ I_{sub} = 0
- If long-channel device w/ $V_{DS} > 3nv_T \Rightarrow I_{sub} = \frac{W}{L}\mu_e v_T^2 C_{sth} e^{\frac{V_{GS}-V_{th}}{nv_T}}$
- With $n = 1+\frac{\gamma}{2\sqrt{2\Phi_f}} = 1+\frac{C_{sth}}{C_{ox}} = 1+\frac{C_{dep}+C_{it}}{C_{ox}}$ $\quad I_{off} = I_{sub}(V_{GS}=0) = \frac{W}{L}\mu_e v_T^2 C_{sth} e^{-\frac{V_{th}}{nv_T}}$
- Key dependencies of the subthreshold slope:
  - T_{ox} ↓ ⇒ C_{ox} ↑ ⇒ n ↓ ⇒ sharper subthreshold
  - N_A ↑ ⇒ C_{sth} ↑ ⇒ n ↑ ⇒ softer subthreshold
  - V_{SB} ↑ ⇒ C_{sth} ↓ ⇒ n ↓ ⇒ sharper subthreshold
  - T ↑ ⇒ softer subthreshold.

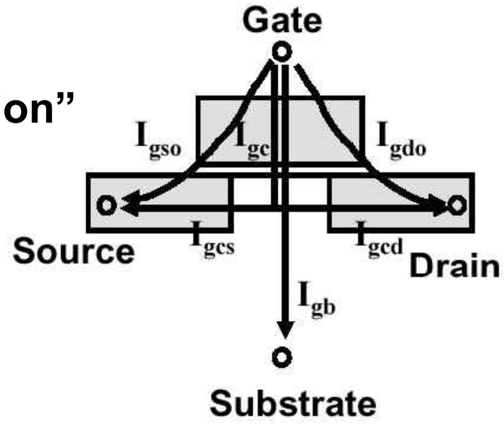
- **Occurs when transistor is "off"**

Gate Oxide Tunneling

- I_4: *Gate oxide tunneling* of electrons that can result in leakage when there is a high electric field across a thin gate oxide layer. Electrons may tunnel into the conduction band of the oxide layer; this is called Fowler-Nordheim tunneling.
- In oxide layers less than 3–4 nm thick, there can also be direct tunneling through the silicon oxide layer. Mechanisms for direct tunneling include electron tunneling in the conduction band (ECB), electron tunneling in the valence band (EVB), and hole tunneling in the valence band (HVB).
- Direct tunneling of electrons through gate oxide is the dominant source. This current depends exponentially on the oxide thickness and the V_{DD} [BSIM 4].

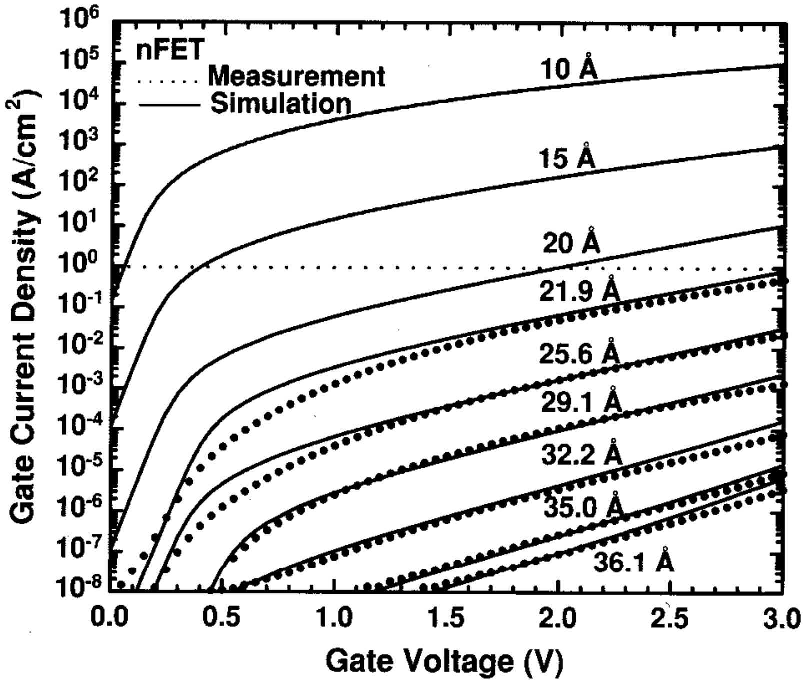
$$J_{DT} = A_g \left(\frac{V_{gs}}{T_{ox}}\right)^2 e^{\frac{-B_g\left(1-\left(1-\frac{V_{gs}}{\Phi_{ox}}\right)^{1.5}\right)}{\frac{V_{gs}}{T_{ox}}}}$$

- **Occurs when transistor is “on”**

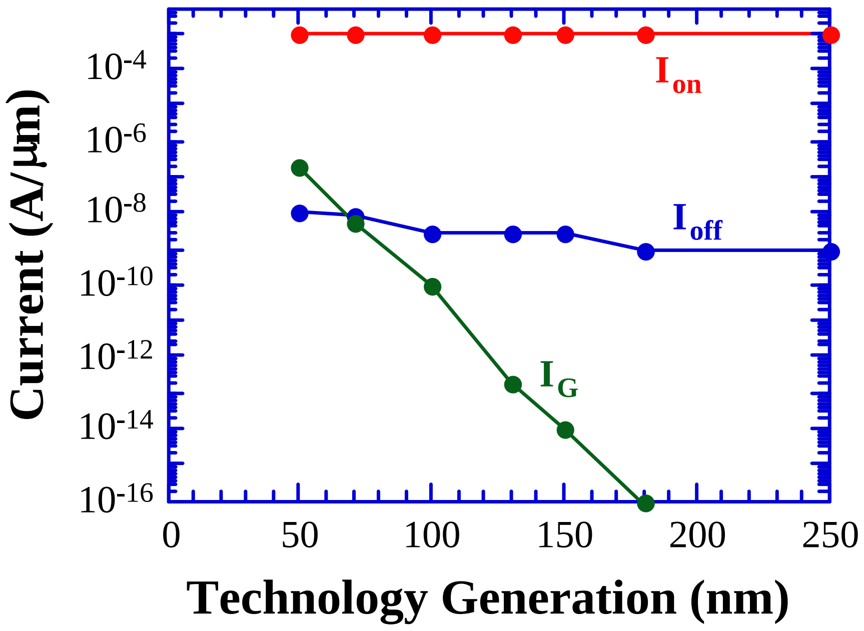


GATE ON/OFF CURRENTS

- As oxide thickness decreases, gate current becomes more important; eventually it dominates the *off* current (I_D at V_G = 0).

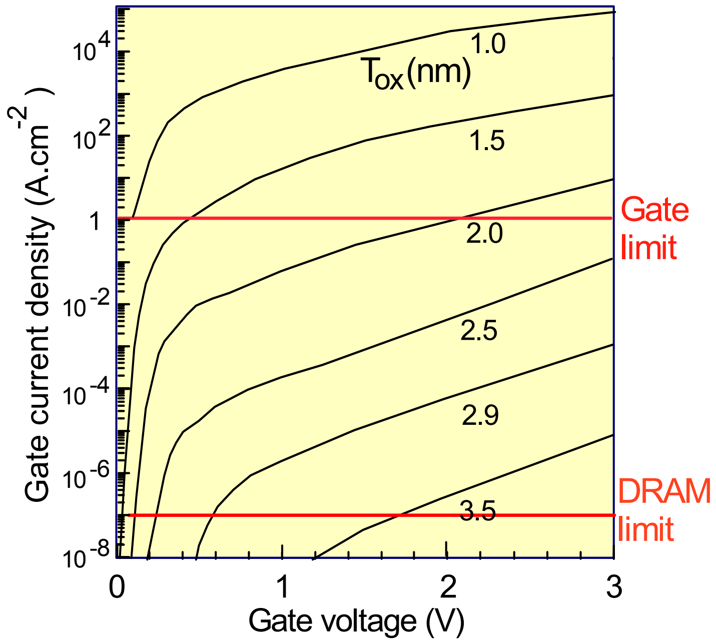


Source: D. Frank, IBM, 2002

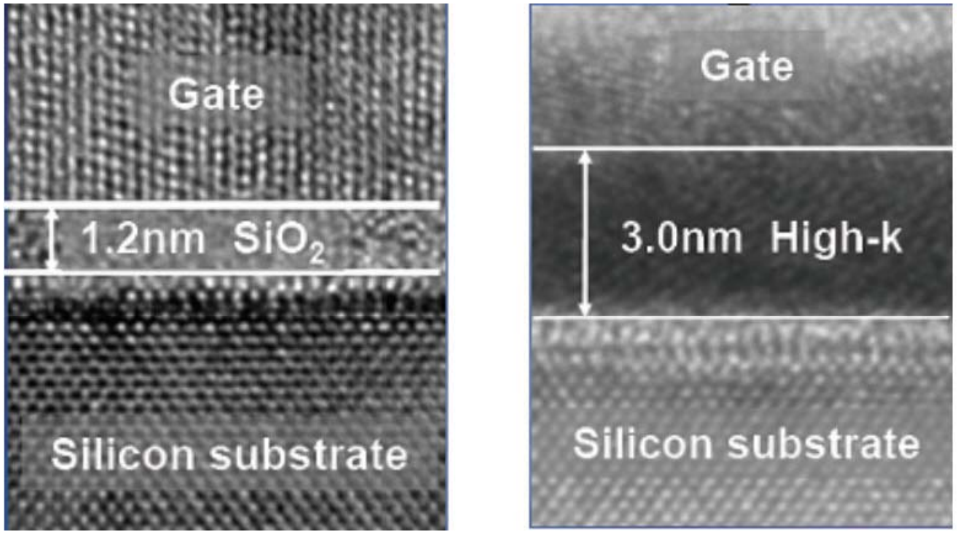


Source: Dieter K. Schroder, Arizona State

Why High K Dielectric



Source: John Roberston, Cambridge Univ.



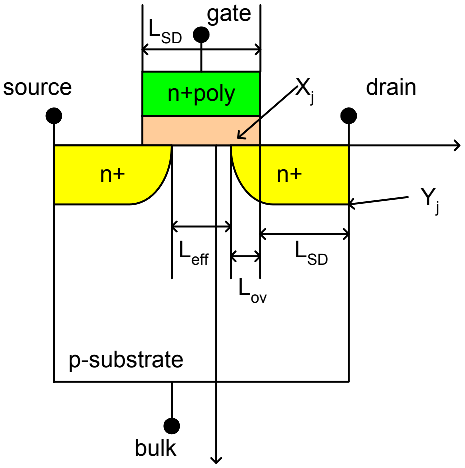
- SiO2 layers <1.6 nm have high leakage current due to direct tunnelling. Not insulating.
- Maintain C/area for S-D current: $C = \dfrac{K\varepsilon_0}{t}$
- Replace SiO2 with thicker layer of new oxide with higher K.
- Equivalent oxide thickness, 'EOT'.

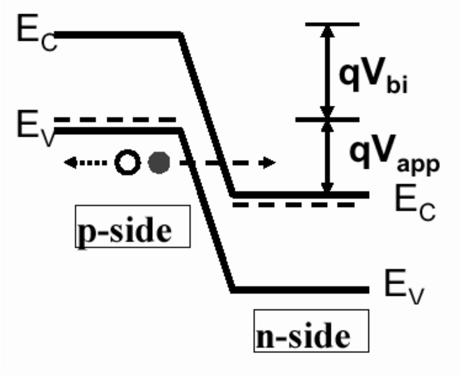
Junction Leakage

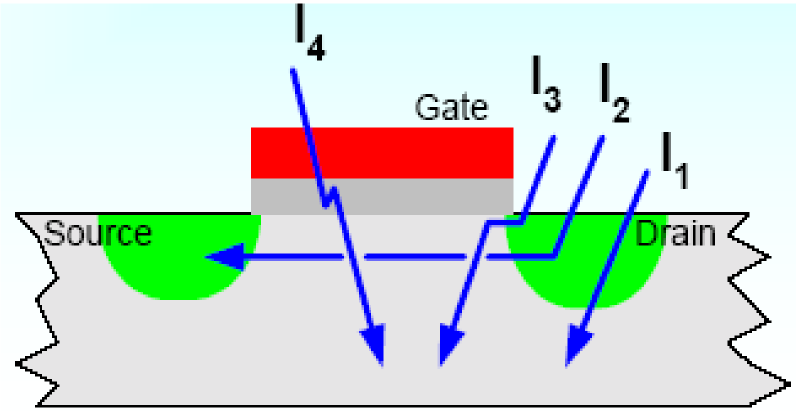
- I_1: *Junction leakage* that results from minority carrier diffusion & drift near the edges of depletion regions, and also from generation of electron-hole pairs in the depletion regions of reverse-bias junctions.
- Diode reverse bias current

$$I_{junc} = I_s\left(1 - e^{-\frac{V_{DB}}{V_{th}}}\right)$$

where V_{DB} is drain to bulk (substrate) voltage.







Outline

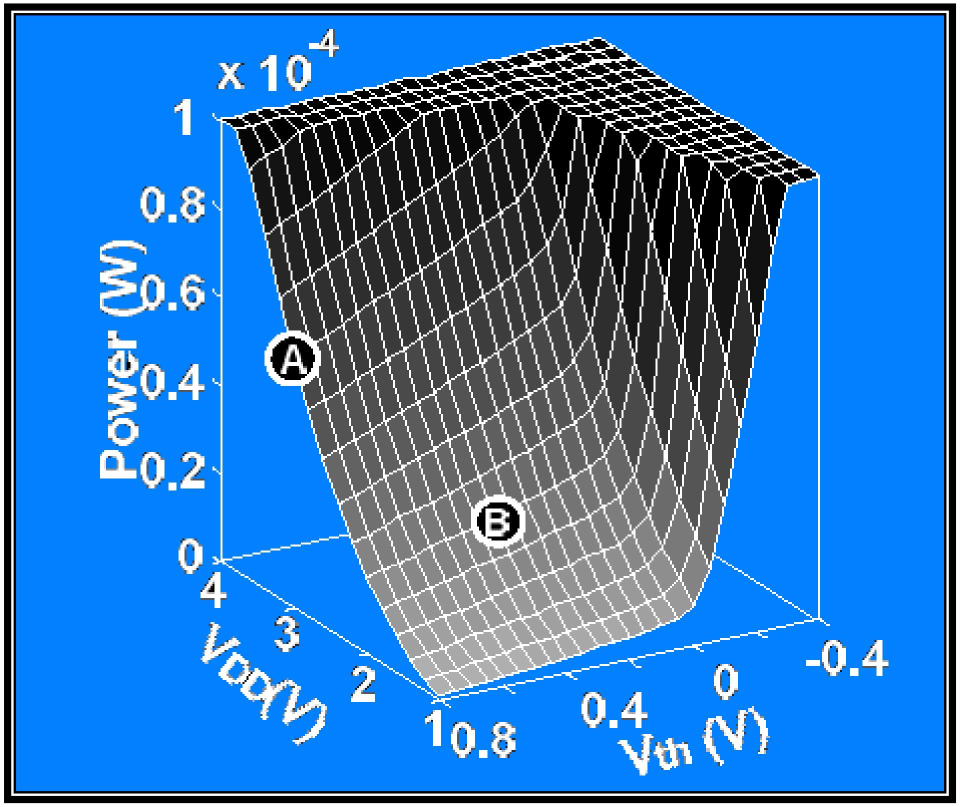
- Technology Trends
- Power Dissipation 101
- Sources of Leakage in CMOS VLSI Circuits
- Circuit and Device Techniques for Leakage Minimization
- Conclusion

Leakage Reduction Techniques

- Lowering and/or turning off V_{dd} (voltage islands and power domains)
- Non-minimum channel length transistors
- Dual-V_{th} design
- Transistor stacking
- Body bias control (static and/or adaptive)
- MTCMOS (sleep transistors, power gating)
- SOI technology
- Cooling and/or refrigeration

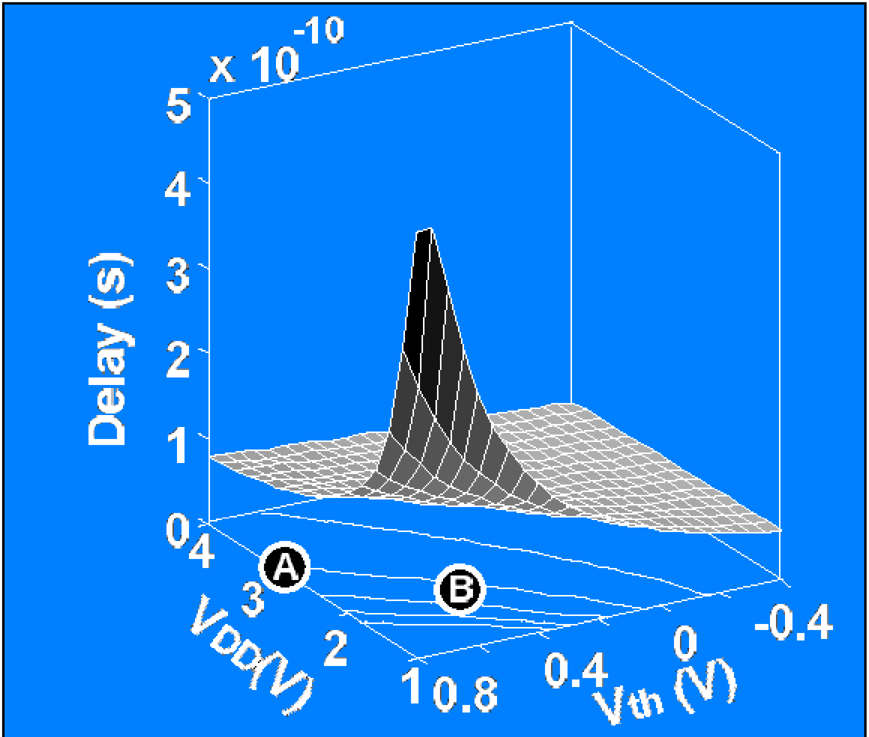
Power dependence on V_{TH} and V_{DD}

Power : $P = p_t \cdot f_{CLK} \cdot C_L \cdot V_{DD}^2 + I_0 \cdot 10^{-\frac{V_{th}}{S}} \cdot V_{DD}$



Delay dependence on V_{TH} and V_{DD}

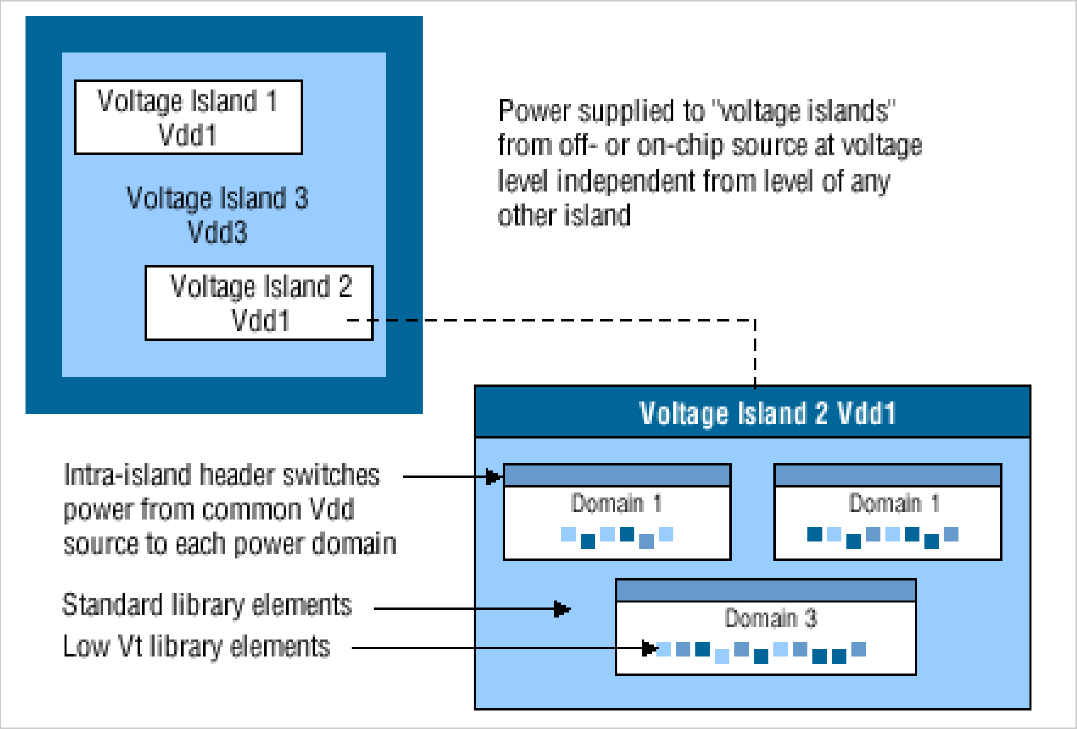
$$\text{Delay} = \frac{k \cdot Q}{I} = \frac{k \cdot C_L \cdot V_{DD}}{(V_{DD} - V_{th})^{\alpha}} \quad (\alpha = 1.3)$$



Delay dependence on V_{DD} & V_{TH}

Voltage Islands and Power Domains

- Voltage islands - Areas (logic and/or memory) on chip supplied through separate, dedicated power feed.
- Power domains - Areas within an Island fed by same Vdd source but independently controlled via on intra-island header switches.

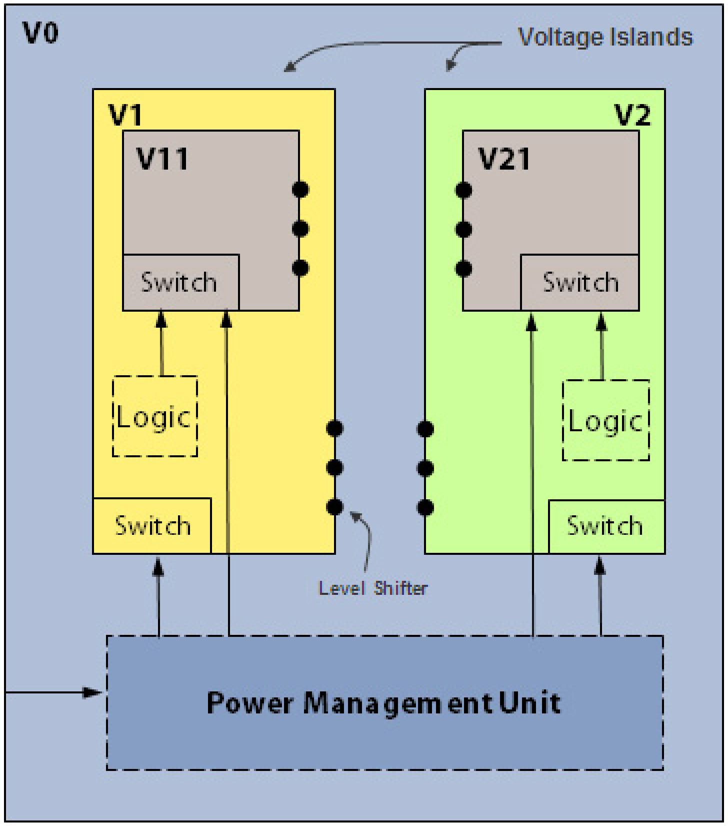

Voltage Island 1
Vdd1
Voltage Island 3
Vdd3
Voltage Island 2
Vdd1
Power supplied to "voltage islands" from off- or on-chip source at voltage level independent from level of any other island
Voltage Island 2 Vdd1
Intra-island header switches power from common Vdd source to each power domain
Domain 1
Domain 1
Standard library elements
Low Vt library elements
Domain 3

Source: IBM Blue Logic® Cu-08 voltage islands

Ldrawn = 70 nm
Up to 72-million wireable gates
Power supply: 1.0 V with 1.2-V option
Power dissipation: 0.006 µW/MHz/ gate
Gate delays: 21 picoseconds (2-input NAND gate)
Eight levels of copper for global routing.

Voltage Island Powering and Switching Control

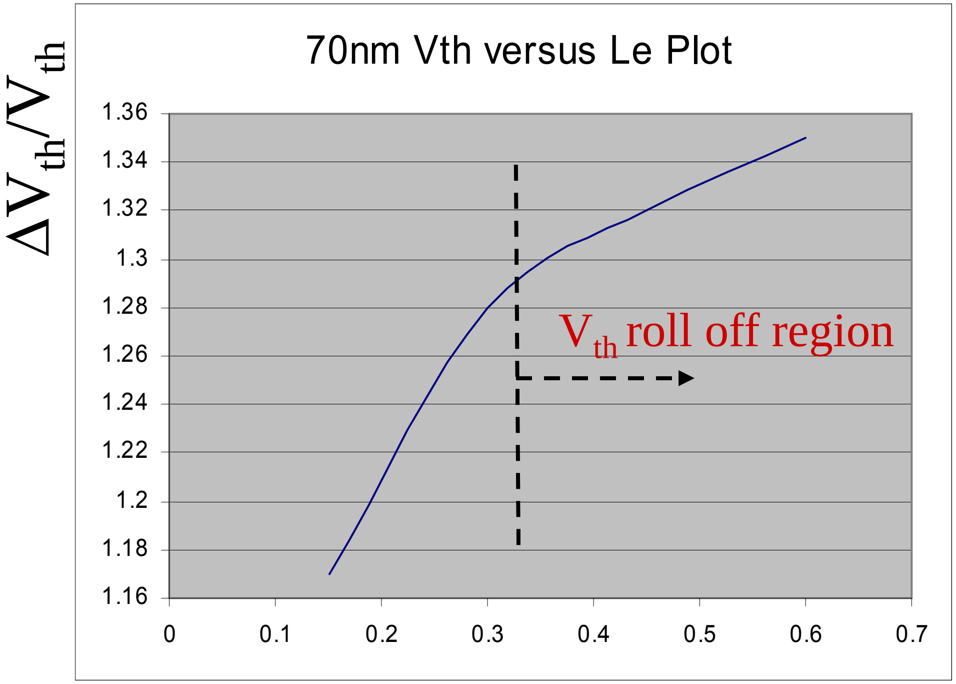
- Any additional voltage levels must be available from off-chip sources or generated with on-chip regulators. Signal interfaces between regions running on different supplies must have appropriate level converters.
- If blocks are powered down, electrical "fencing" is required to prevent indeterminate state from corrupting still active blocks. In addition, powered down blocks probably need to save and restore state.



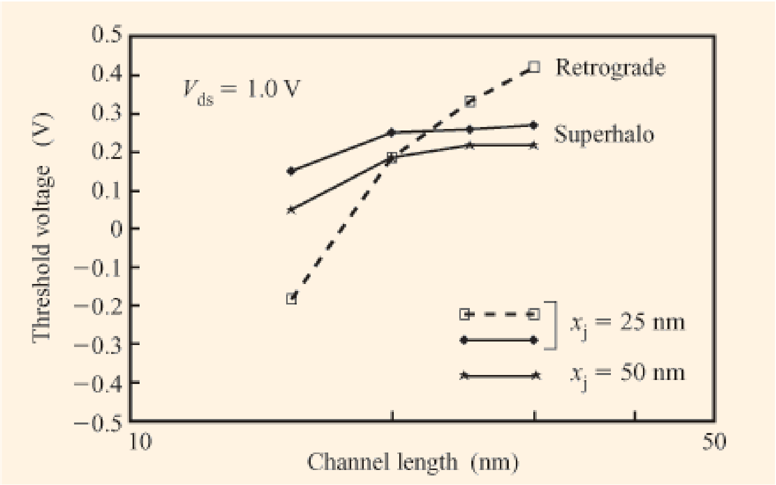
Source EE-Times, August 06, 2004 Issue

Using Non-minimum Channel Lengths

- As L_e increases, V_{th} goes up and leakage is reduced exponentially.
- Extent of leakages reduction is a function of the V_{th} roll off effect.



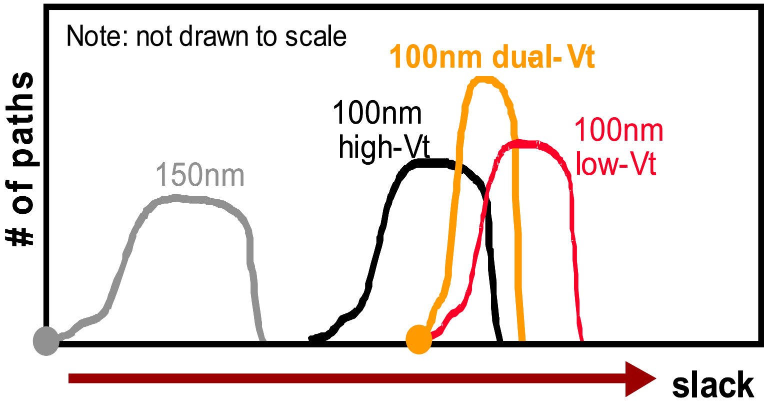
Source: Clark, SLPED 2004



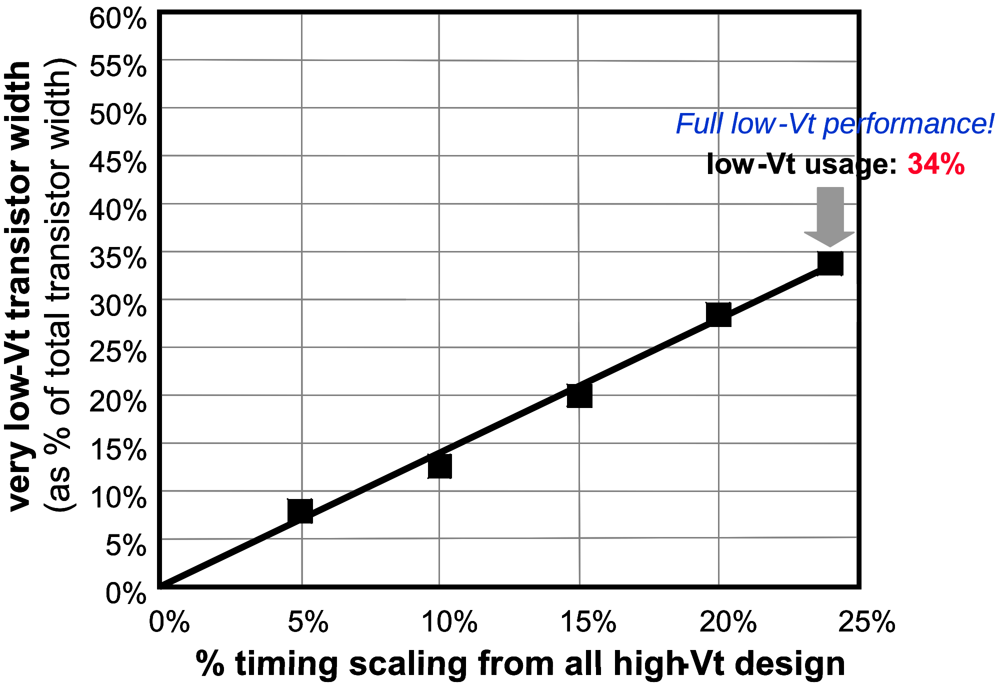
Source: Taur, IDEM 1998

Dual V_{th} Design

- Use two V_T's (e.g., 0.6V and 0.3V for V_{DD} = 2.5V)
  - Use the lower threshold for gates on critical path
  - Use the higher threshold for gates off the critical path.
- Improves performance without an increase in power.



Source: Vivek De, Intel



Full low-Vt performance!
low-Vt usage: 34%

Dual V_{th}: Observations and Questions

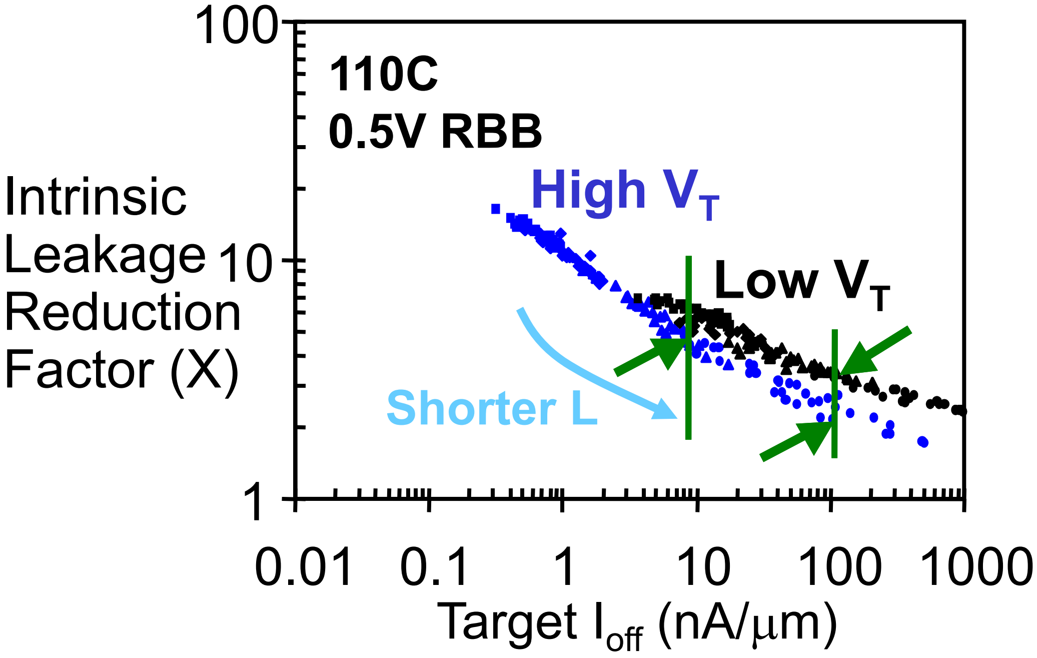
- In the dual-V_{th} approach, leakage current is controlled by managing threshold voltages using manufacturing process techniques so that the threshold of a gate is fixed.
- The high V_{th} cells are identical to the low V_{th} cells except for processing changes.
- No change to the library characterization methodology is required, but the high V_{th} library does require a separate characterization. Modeling of the library can continue to be done with NLDM format look-up tables, and the cells can be stored in separate Liberty files.
- How can we assign threshold voltages to transistors?
  - Not all non-critical gates can be made high V_{th}.
- What are their optimal values?
  - Delay and leakage sensitivities to the V_{th} values.

Variable-Threshold CMOS

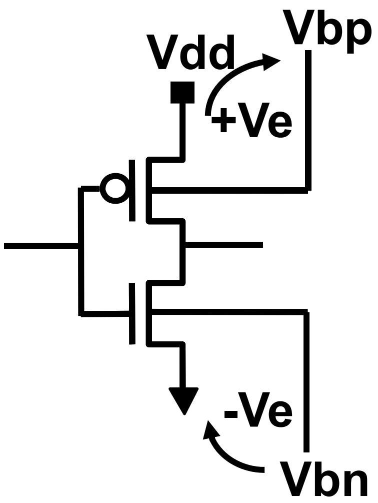
- Another approach to managing threshold voltage is the use of substrate biasing.
  - Normally, the substrate in a digital gate is tied to GND for NMOS transistors and to V_{dd} for PMOS transistors. Substrate biasing is used in conjunction with process-based V_{th} management.
- As the bias voltage for any of these devices is reduced below 0V under reverse bulk bias (RBB), V_{th} increases and leakage current decreases. Under forward bulk bias (FBB), V_{th} decreases and performance increases at the expense of leakage current.
- The variable threshold or VTCMOS approach uses RBB in standby modes to manage leakage power without degrading performance.
  - This approach assumes a low V_{dd} with low V_{th} devices are used in active mode to meet overall performance requirements, although it can be combined with dual- and multi- V_{dd} approaches. A standby mode using RBB raises the effective V_{th} to block leakage current.

Reverse and Forward Body Biasing

- Reverse Body Biasing (RBB)
  - No Bias = Low V_{th}
  - Apply Reverse Bias = High $V_{th.}$
- Forward Body Biasing (FBB)
  - No bias = High V_{th}
  - Apply Forward Bias = Low $V_{th.}$



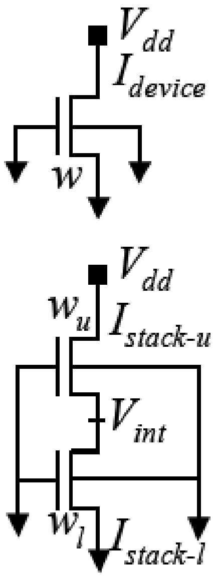
Body Biasing

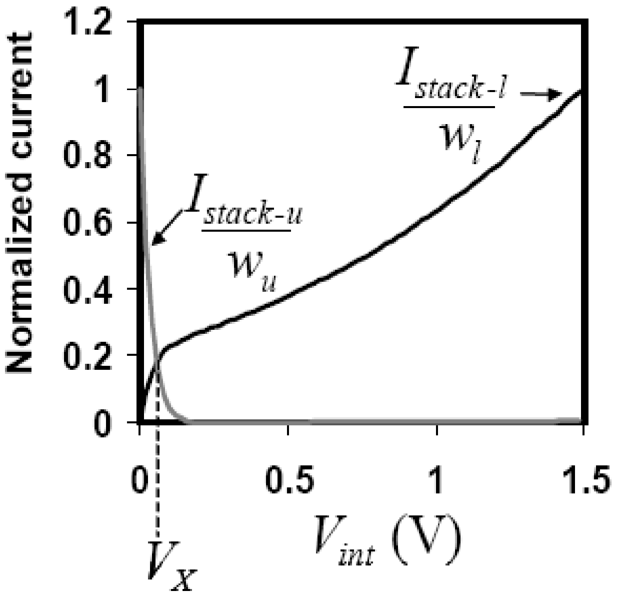


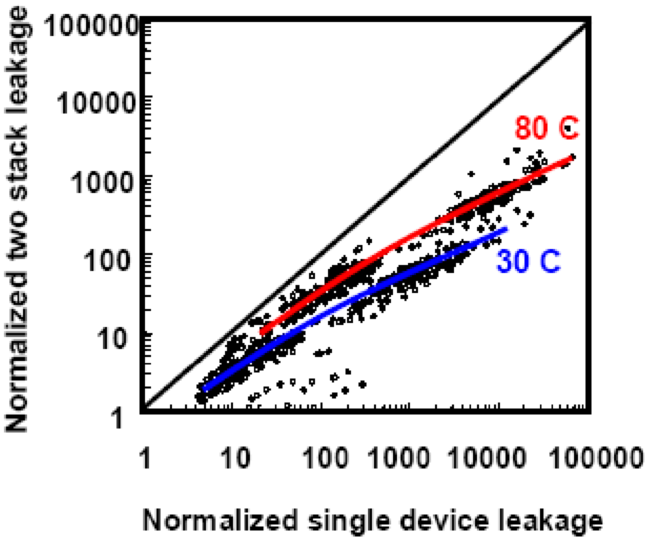
RBB becomes less effective at shorter L and lower V_{th}

Source: A. Keshavarzi, 1999 & 2001 , SLPED

Stack Forcing



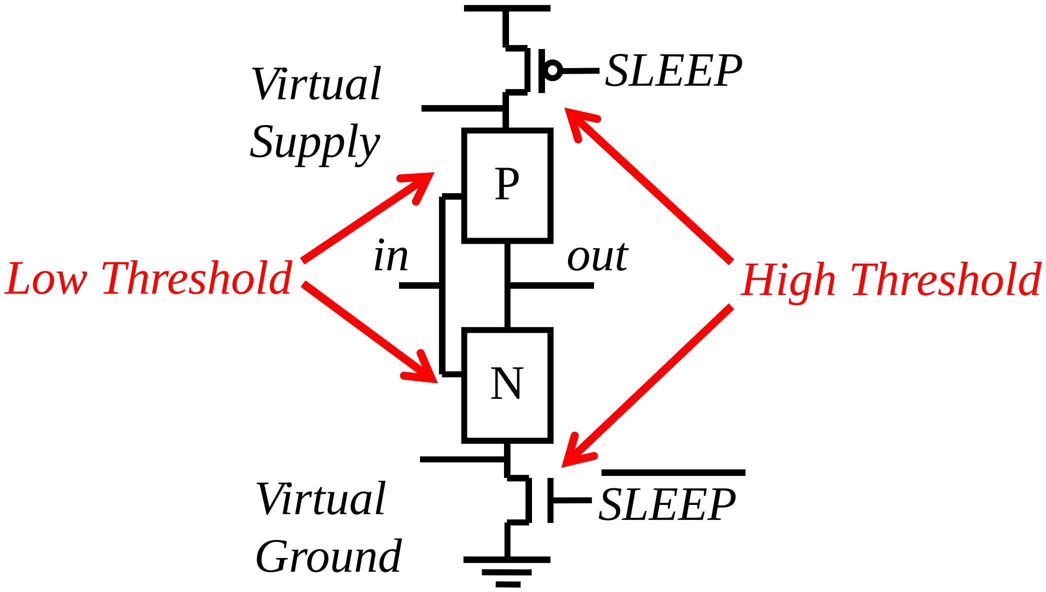




- Another low-leakage power design solution is “stack forcing,” in which a single transistor is divided into two or more transistors to increase resistance for leakage currents.

Multi-Threshold CMOS (MTCMOS)

- It is also called guarding, power gating, ground gating, using sleep transistor, etc.
- A high-V_{th} is used to disconnect low-V_{th} transistors from the ground (V_{dd}).

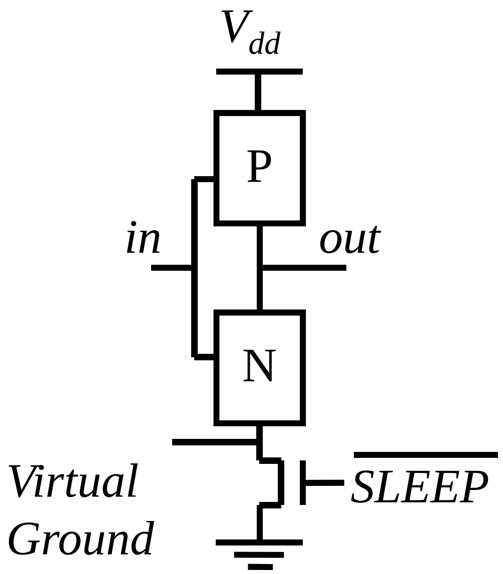


Sleep Mode Approach

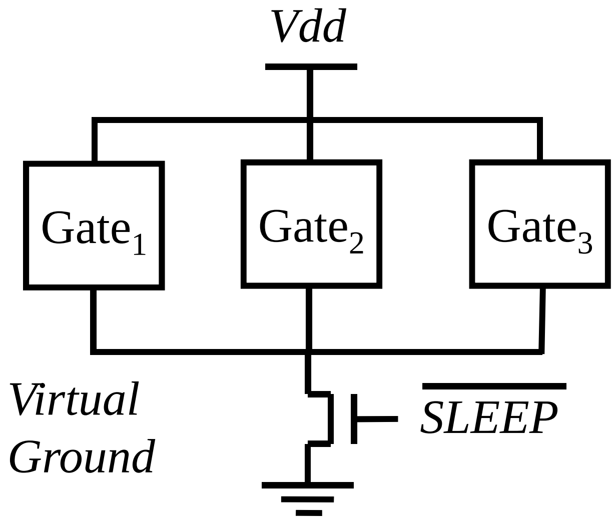
- The high Vt sleep mode approach is a global leakage power reduction technique that requires fewer changes to the design. The design is implemented using low Vt transistors to meet high-speed performance goals. High Vt NMOS and PMOS sleep control transistors are added to form virtual supply rails ($V_{DD,V}$, GND_V.)
- A sleep mode signal, SL, controls the operation of these high Vt transistors. In active mode, these sleep transistors function as real power and grounds with small on-resistances. In sleep mode, these transistors block the leakage currents otherwise present in the low Vt circuitry.

Simplifications

- Instead of two sleep transistors, one can be used.
- Usually NMOS:
  - $\mu_n > \mu_p$ → smaller size.
  - However, PMOS usually has a lower leakage.

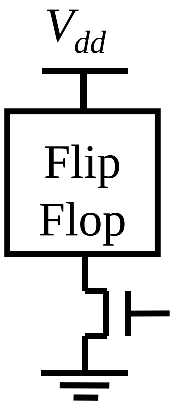


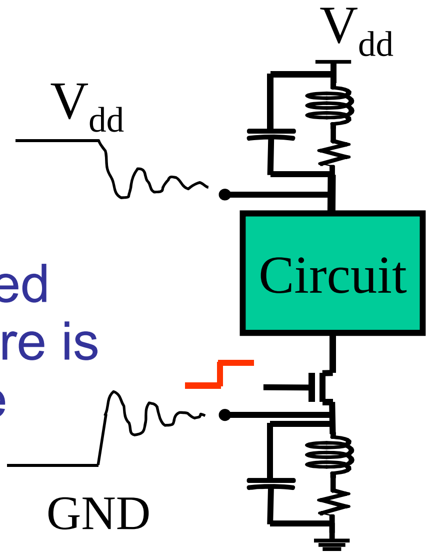
- One sleep transistor can be shared between several gates
  - Reduction in the number of sleep transistors, area overhead, dynamic and leakage power dissipations
  - Increase in the complexity of design optimization process.



Problems with MTCMOS

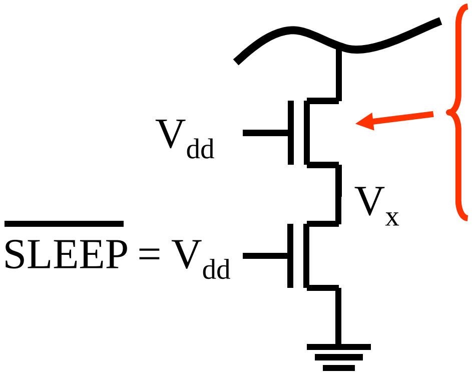
- The electro-migration effect on vias and wires
  - The number of vias is determined based on the average and the maximum allowable currents for each via.
- MTCMOS cannot be easily applied to Flip Flops
  - May use balloon latches, etc.
- In an SoC, not all IP blocks are guarded
  - Short circuit current will result if a guarded output drives a regular input.
- Ground bounce problem
  - During the sleep period, internal nodes are charged to V_{dd}; When the sleep transistor is turned on, there is a current spike flowing to the ground (due to large V_{DS}); This creates large V_{dd} and ground noise.





Sleep Transistor Sizing

- Reduction in the high to low transition due to,
  - Reduction in the gate drive from V_{dd} to $V_{dd} – V_x$.
  - Increase in the threshold voltage of NMOS due to the body effect.
- Increase the sleep transistor width to solve the problem
  - Increase in the area overhead, dynamic power and leakage.
- Technology scale down →have to enlarge the sleep transistor



$$V_{gs} < V_{dd}$$

$$V_{th} = V_{th_0} + \gamma' V_{app} - \eta V_{ds}$$

[Kao-DAC97]

Important Questions

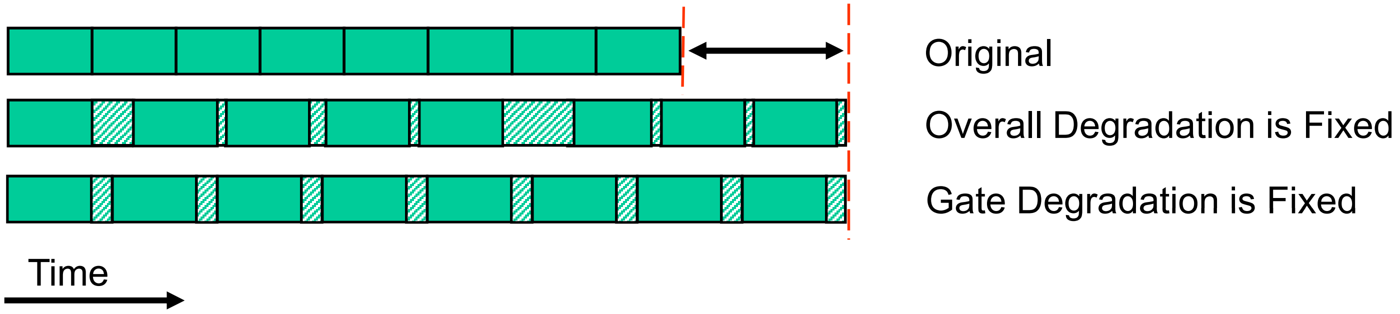
- How many sleep transistors?
  - Affects the area overhead, the dynamic power overhead, and the leakage power saving.
- How to cluster gates?
  - Affects the routability and the size of the sleep transistors.
- What size to choose for the sleep transistors?
  - Affects the delay and area overhead, the dynamic power overhead and the leakage power saving.

Sizing: Exhaustive Approach

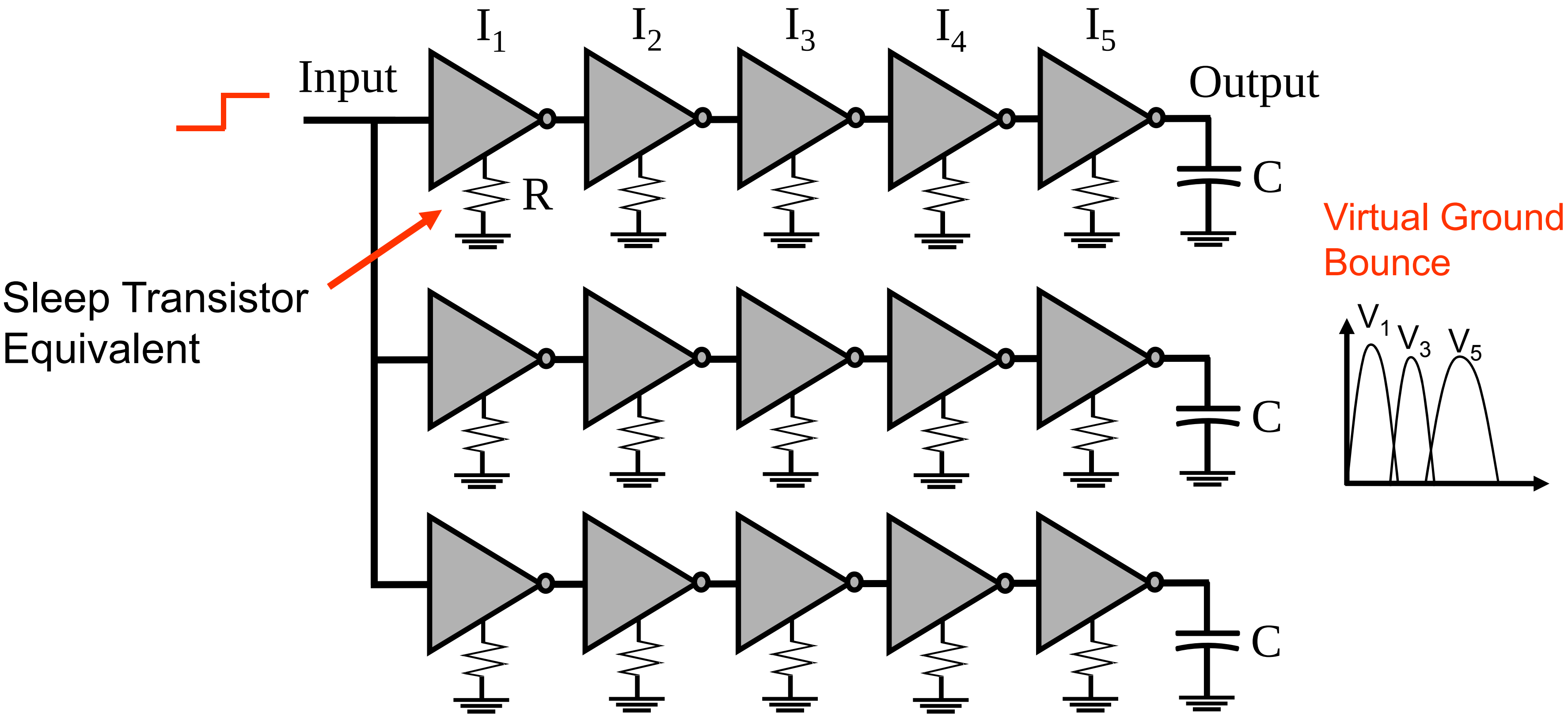
- Objective: Find size of the sleep transistor to meet a user-specified *overall* delay degradation ($\Delta d/d$).
- Exhaustively simulate the entire circuit with the sleep transistor in place
  - Finds the optimum size,
  - Works well for library cells,
  - But is impractical for real-sized circuits.

Gate-based Sizing: A Conservative Approach

- Limit the delay degradation of each gate to *Δd/d.*
- Find the optimum size of the sleep transistor needed for each gate,
  - A pessimistic approach, but easier to do.
  - Assumes both low-to-high and high-to-low transitions degrade.
- Combine the sleep transistors for different gates.



Mutual Exclusion-Circuit A

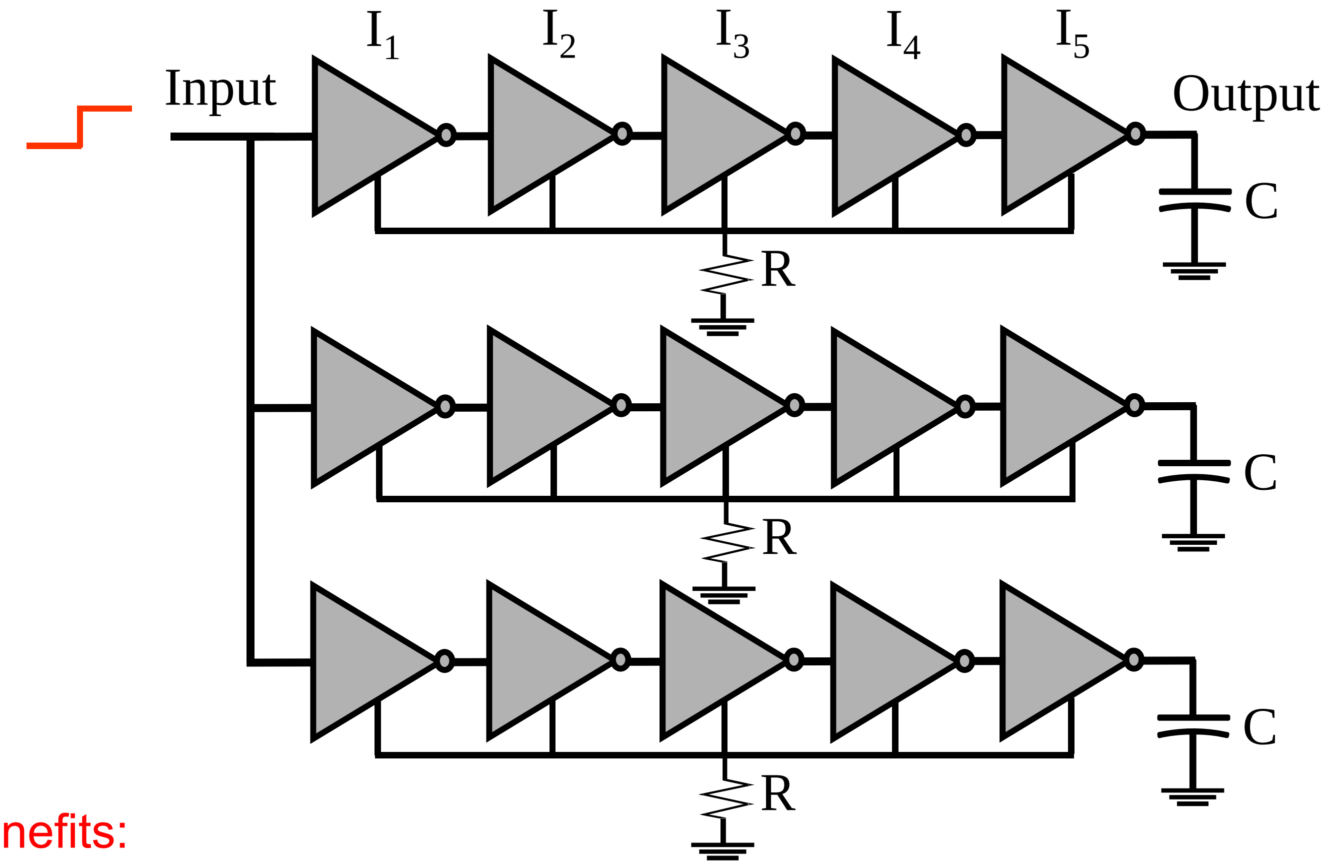


Sleep Transistor Equivalent

Virtual Ground Bounce

[Kao-DAC98]

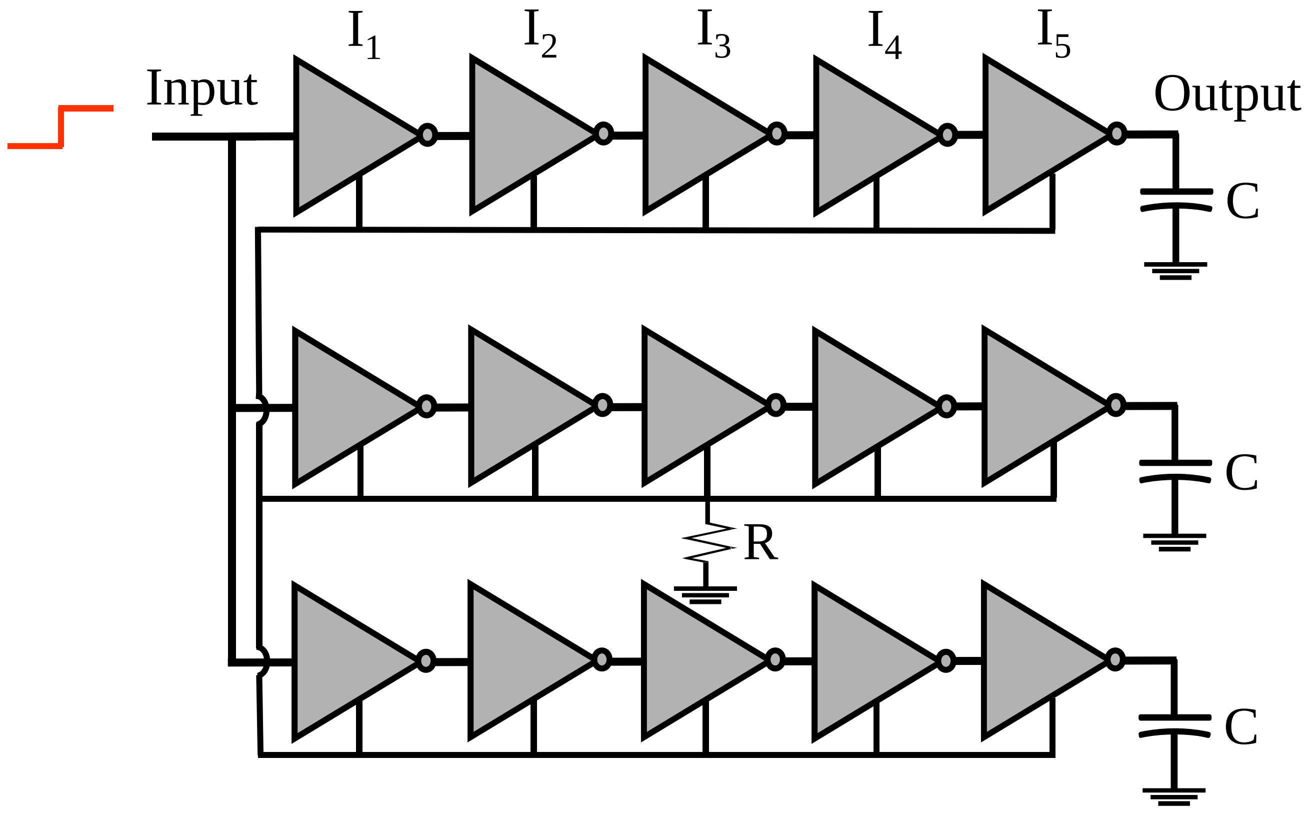
Mutual Exclusion-Circuit B



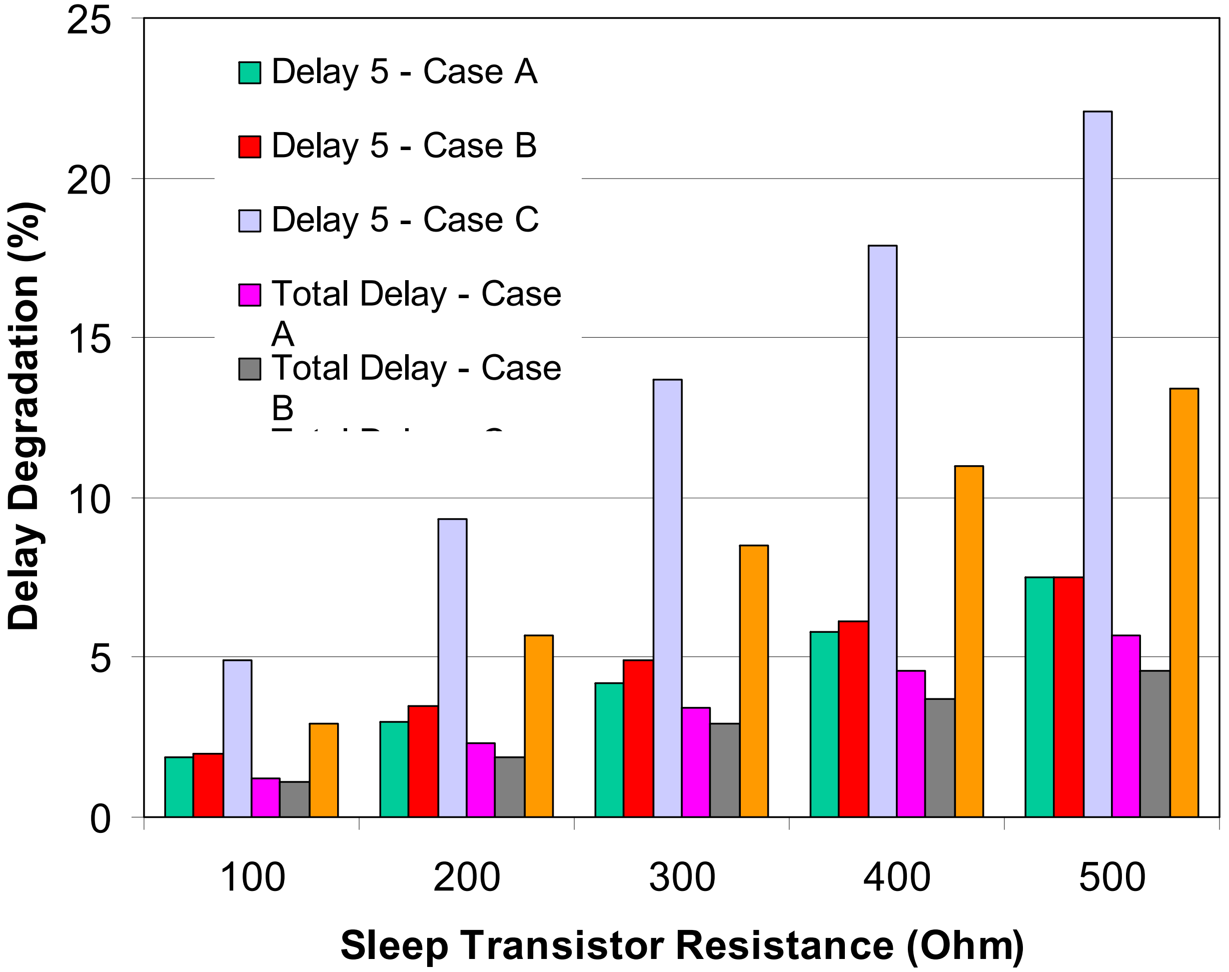
Benefits:

- Reduction in the area overhead, leakage and dynamic power.
- Decrease in the virtual ground bounce due to increase in the parasitic capacitance.

Mutual Exclusion-Circuit C

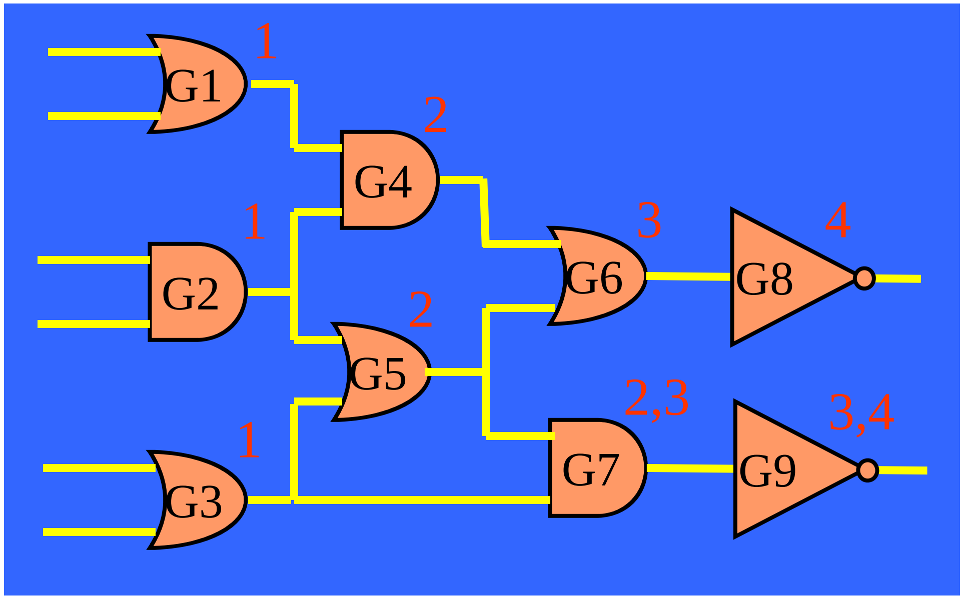


Percentage of Delay Degradation



Mutual Exclusion-based Sizing

- Find all possible transition times of gates by assuming a unit delay model for each gate.
- Form groups of gates that have mutually exclusive transition times.
- Connect each group to a properly-sized sleep transistor.
- Merge the parallel sleep transistors.



$$Group1 = \{ G1, G4, G6, G8 \}$$

$$Group2 = \{ G2, G5, G9 \}$$

$$Group3 = \{ G3, G7 \}$$

Merging Parallel Sleep Transistors

$$V(t) = \min(V_1(t), V_2(t))$$

$$\downarrow$$

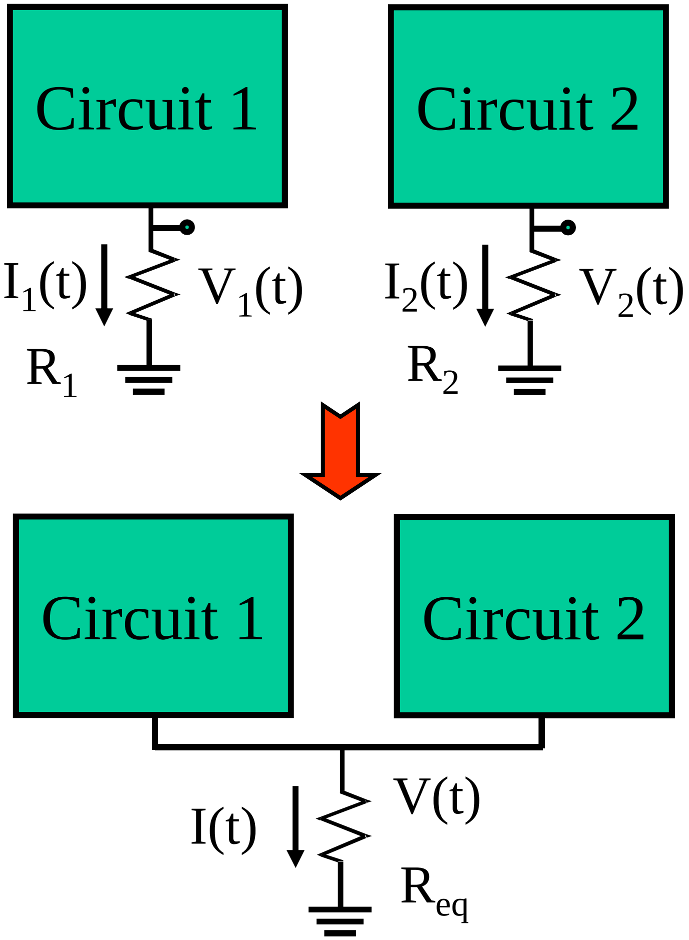
$$R_{eq} = \frac{\min(V_1(t), V_2(t))}{I(t)}$$

$$\downarrow$$

$$R_{eq} = \frac{\min(V_1(t), V_2(t))}{I_1(t) + I_2(t)}$$

$$\downarrow$$

$$R_{eq} = \frac{\min(V_1(t), V_2(t))}{\dfrac{V_1(t)}{R_1} + \dfrac{V_2(t)}{R_2}} \overset{V_1(t) = V_2(t)}{=} \frac{R_1 R_2}{R_1 + R_2}$$



Comparison

Circuit	Sizing Method	Sleep Transistor Resistance
Inverter	Optimal	340Ω
3 inverter chains	Mutual Exclusion (ME)	113Ω
3 inverter chains	Optimal	180Ω

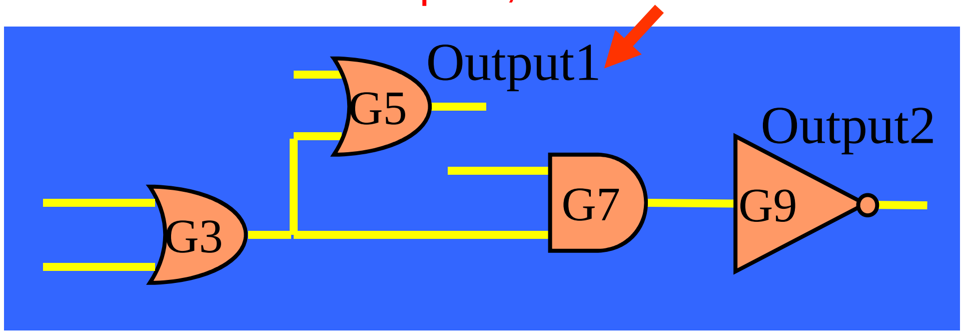
Also, the sleep transistor resistance for one inverter chain when utilizing the mutual exclusion principle

- The ME-based sizing results in ~60% overestimation for the size of the sleep transistor!!!
  - Because in practice only half of gates switch from high to low.

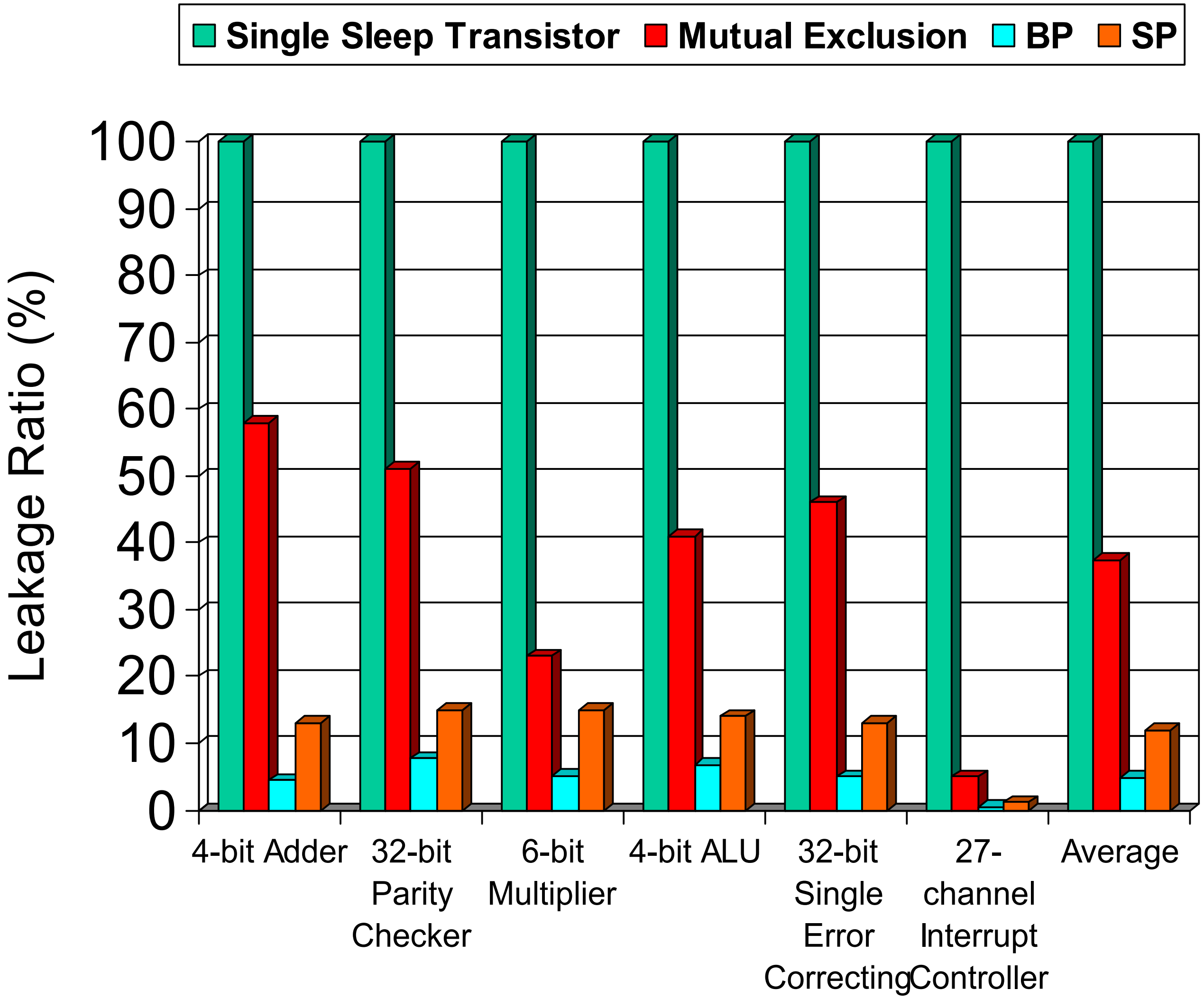
How to Improve the Results

- The ME-based method provides an upper bound on the size of a sleep transistor.
- To improve the result,
  - Use logical information in addition to structural information to determine mutual exclusion conditions.
  - Limit the delay degradation of the entire circuit, not every output.
- Two heuristics for grouping cells to connect to a single sleep transistor
  - Bin packing (BP)
  - Set partitioning (SP)

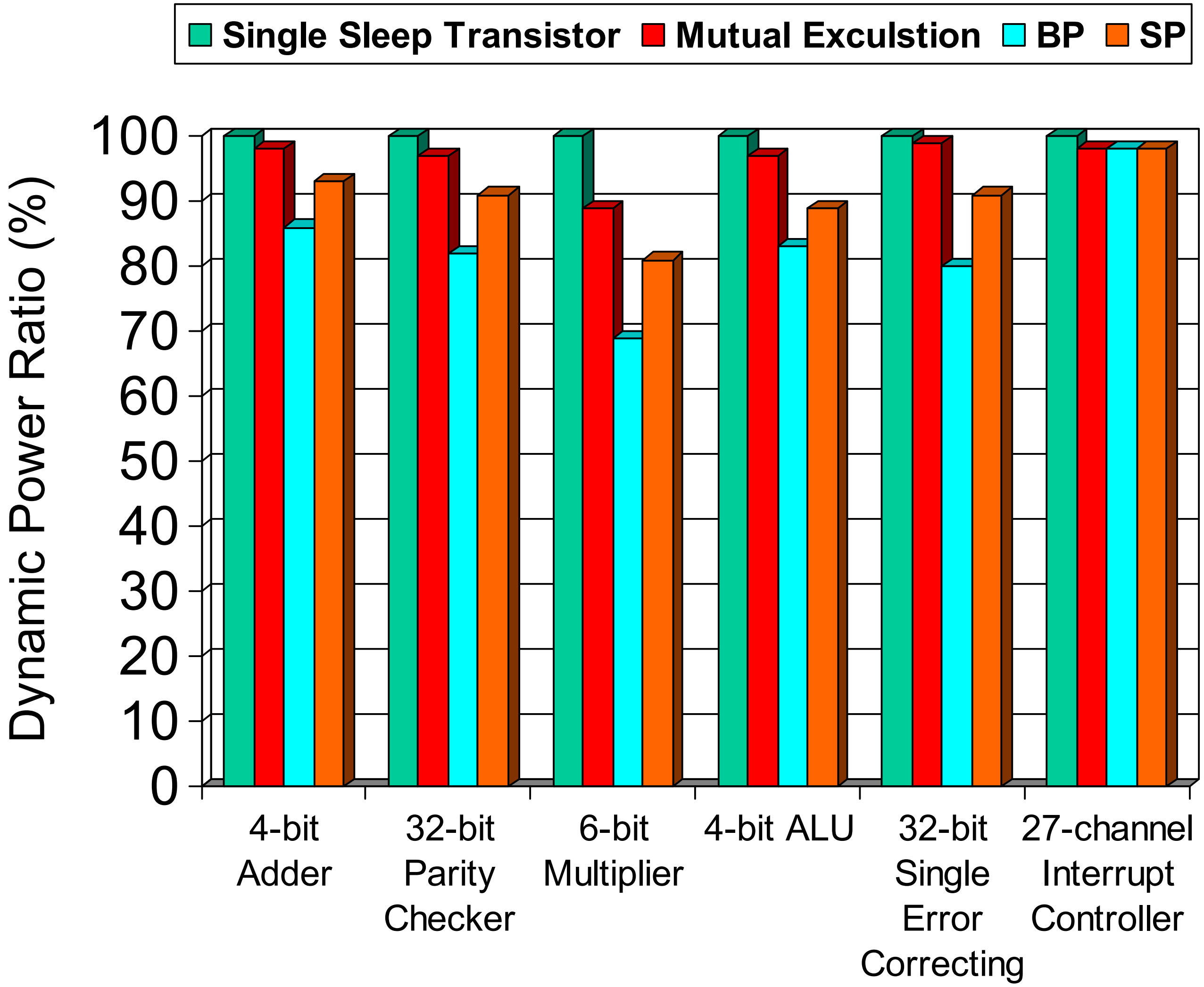
This output is not on the critical path; no need to limit its delay degradation.



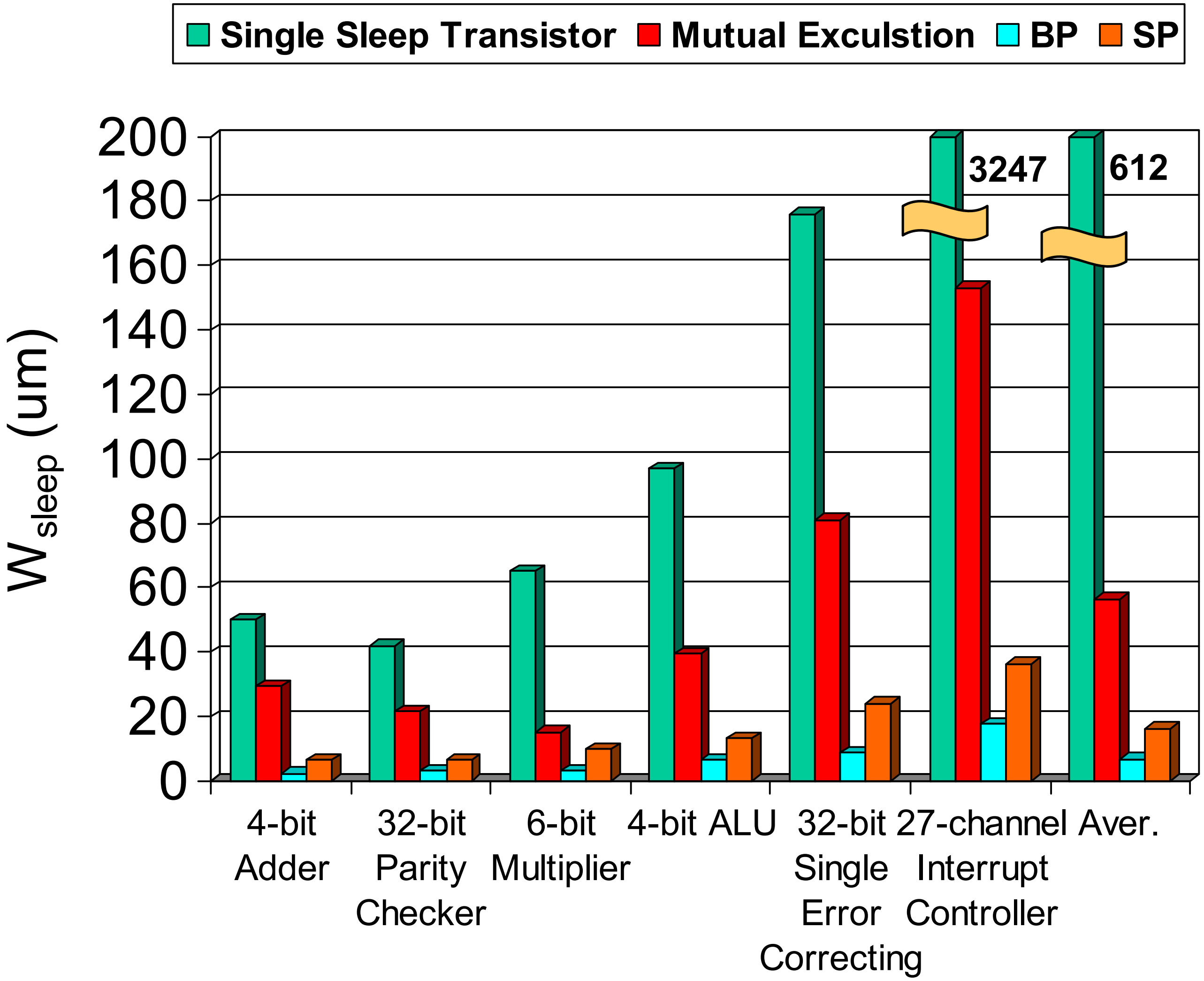
Results: Leakage Reduction



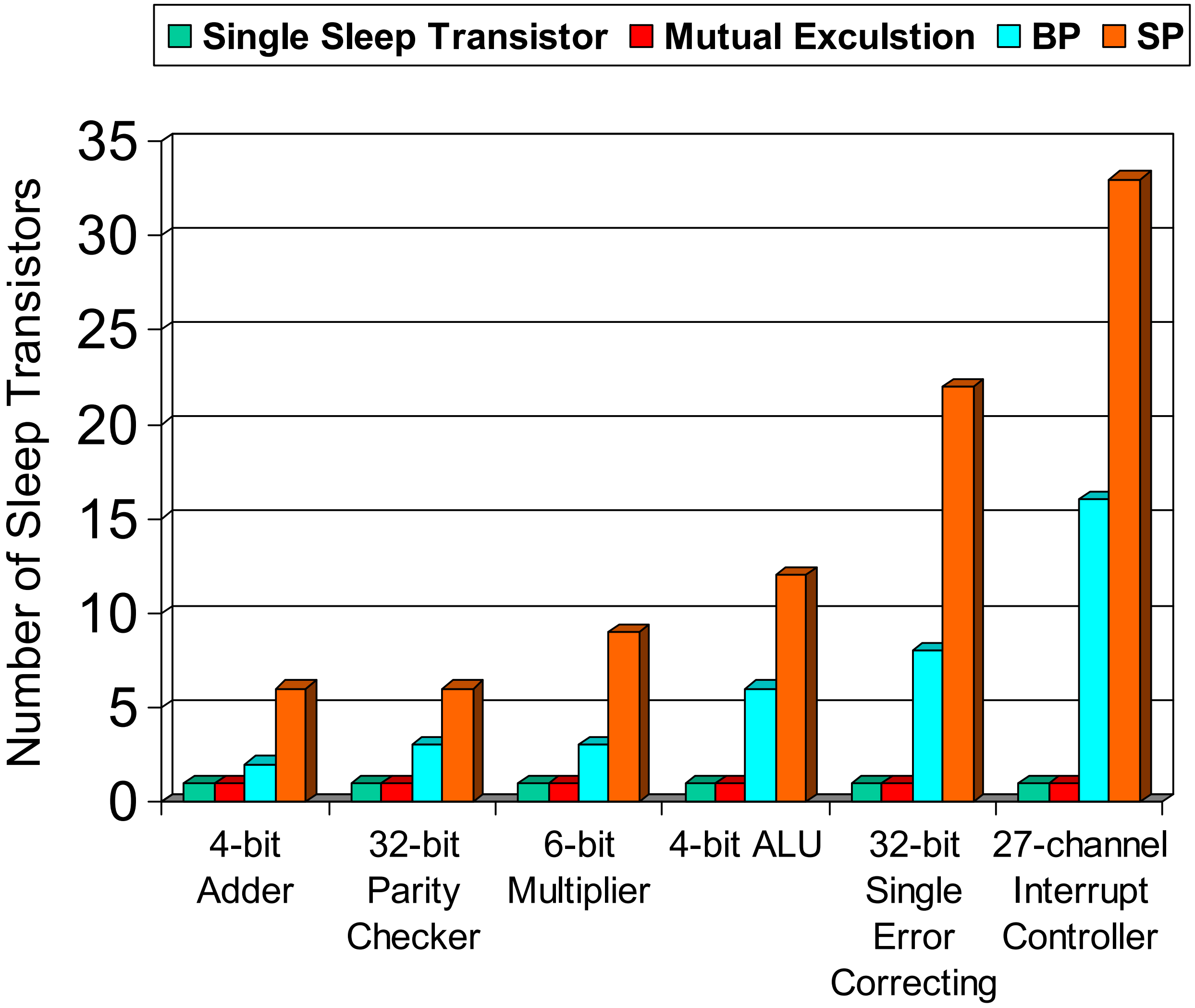
Results: Dynamic Power Reduction in Active Mode



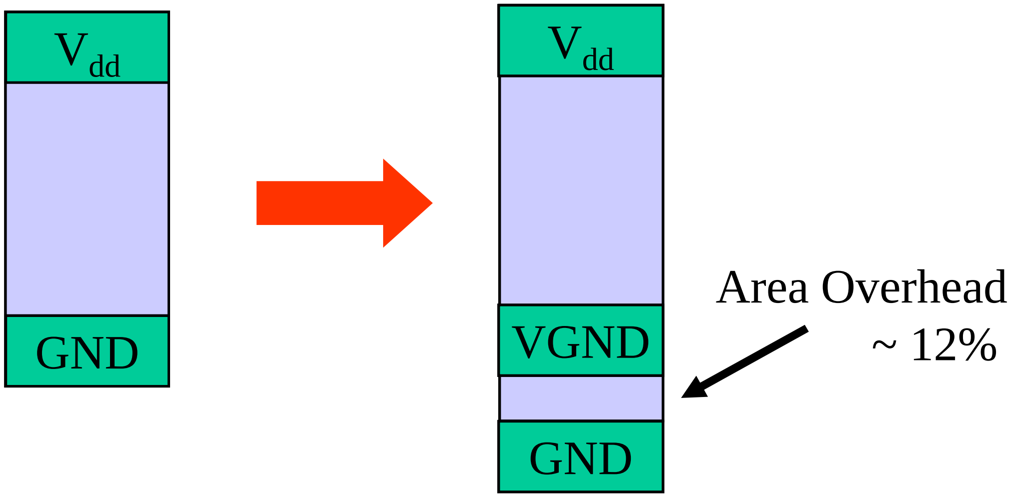
Results: Total Width of the Sleep Transistors



Results: Number of Sleep Transistors



Samsung’s MTCMOS Design Methodology

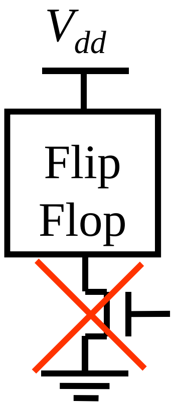


- Design new cells that have sleep transistors.
- Use conventional P&R methodology.

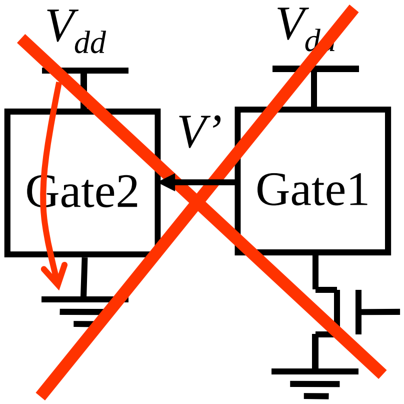
[Won-ISLPED03]

Problems

- MTCMOS cannot be applied to Flip Flops
  - Data loss
  - Can copy the data to an external memory
    - Delay and dynamic power overhead
    - Energy overhead of external memory

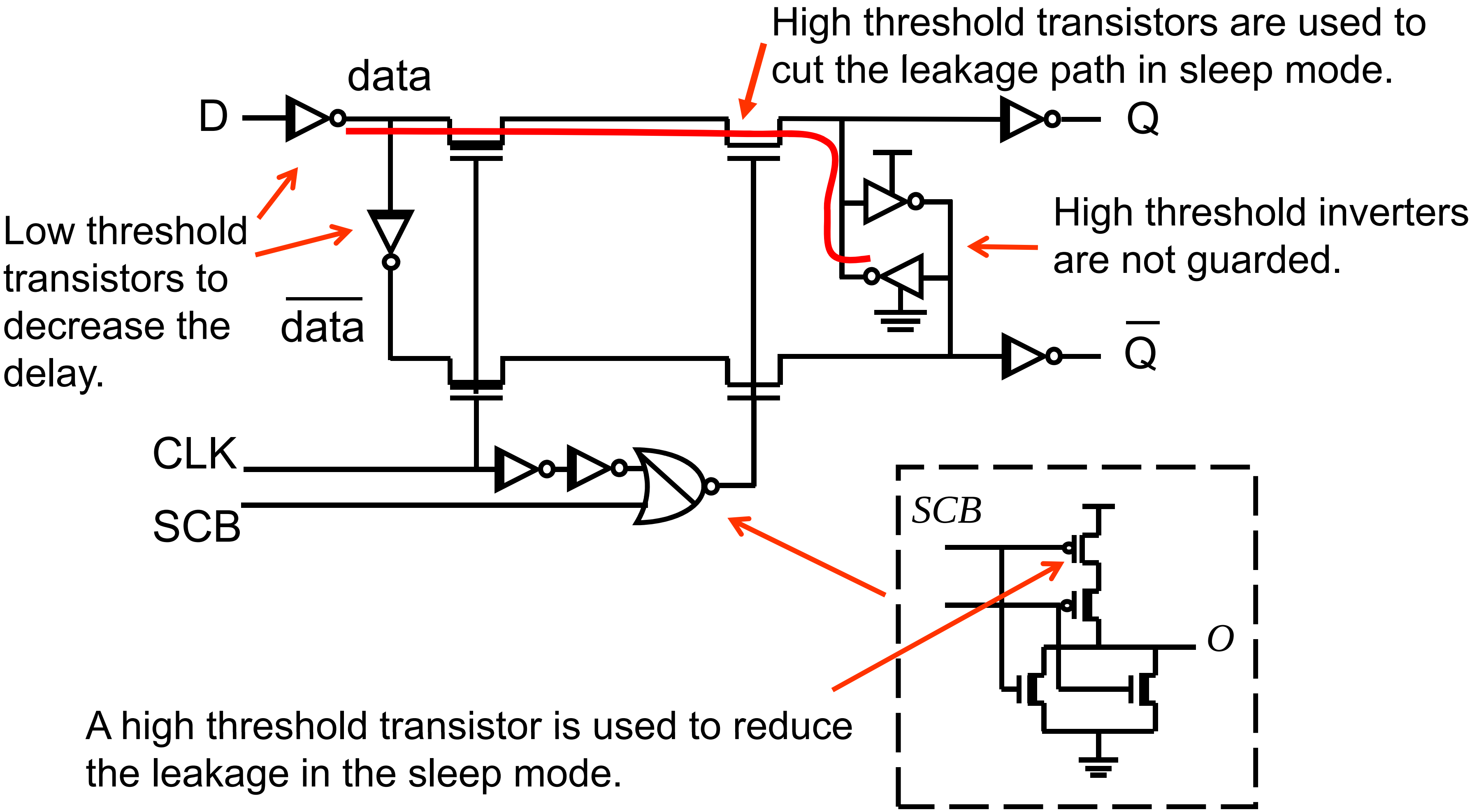


- In an SoC, not all IP blocks are guarded
  - Short circuit current if a guarded output drives a regular input.



$0 < V' < V_{dd}$

Complementary Pass-Transistor Flip Flop (CPFF)



High threshold transistors are used to cut the leakage path in sleep mode.

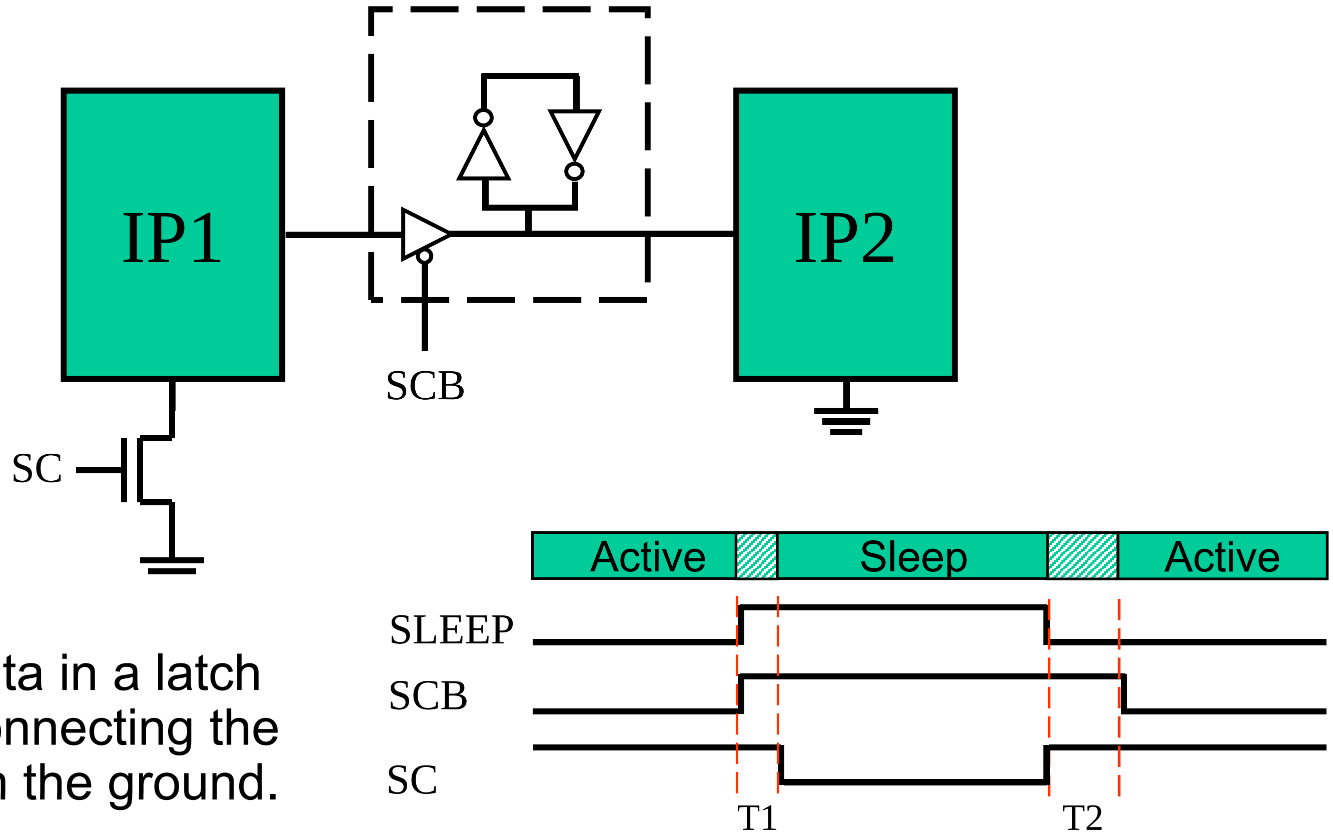
Low threshold transistors to decrease the delay.

High threshold inverters are not guarded.

A high threshold transistor is used to reduce the leakage in the sleep mode.

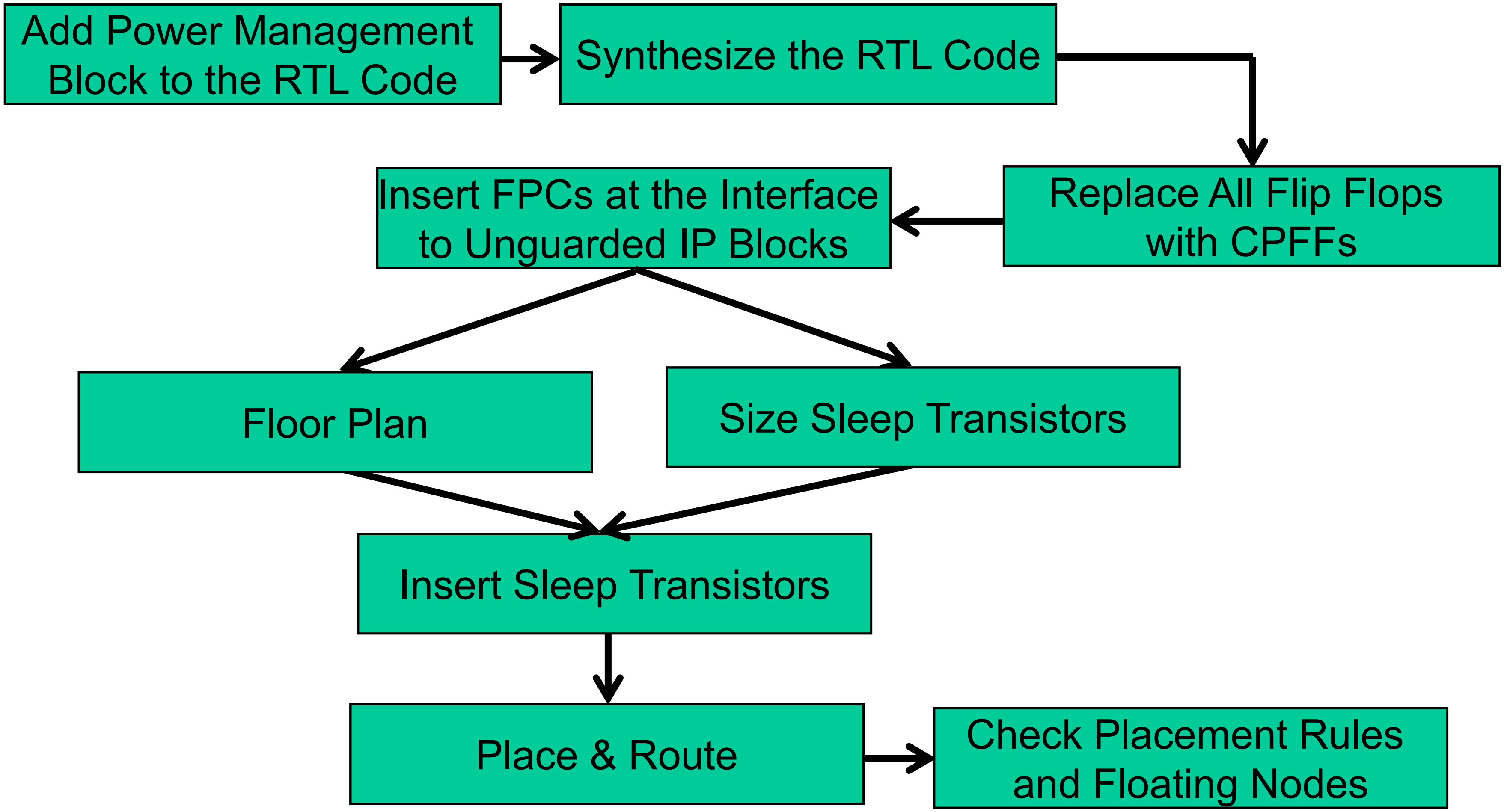
Preventing Short Circuit Current

Floating Prevention Circuit (FPC)



- Store the data in a latch before disconnecting the module from the ground.

Design Flow



Add Power Management Block to the RTL Code

Synthesize the RTL Code

Replace All Flip Flops with CPFFs

Insert FPCs at the Interface to Unguarded IP Blocks

Floor Plan

Size Sleep Transistors

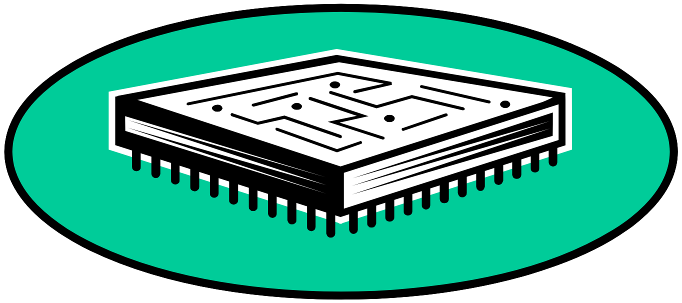
Insert Sleep Transistors

Place & Route

Check Placement Rules and Floating Nodes

DSP Core

- The method was applied to a 16-bit DSP chip
- 0.18μm, V_{dd} = 1.8V
- Inserted 324 sleep transistors with the size of 5μm
- Ground bounce: average=9mV, max=49mV
- Performance degradation = 2%

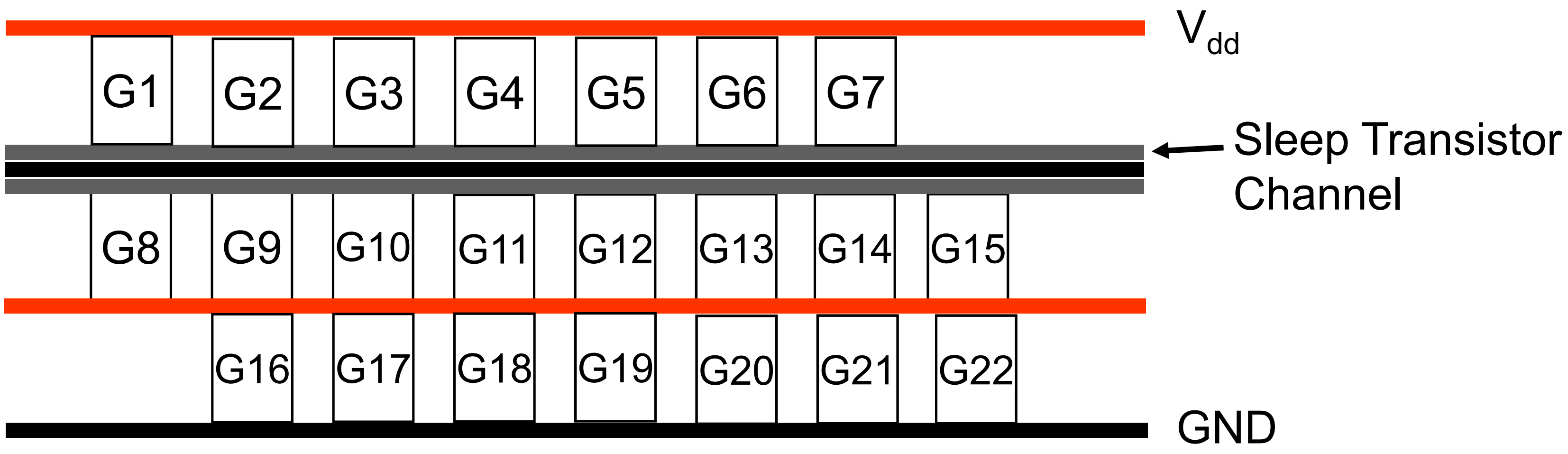


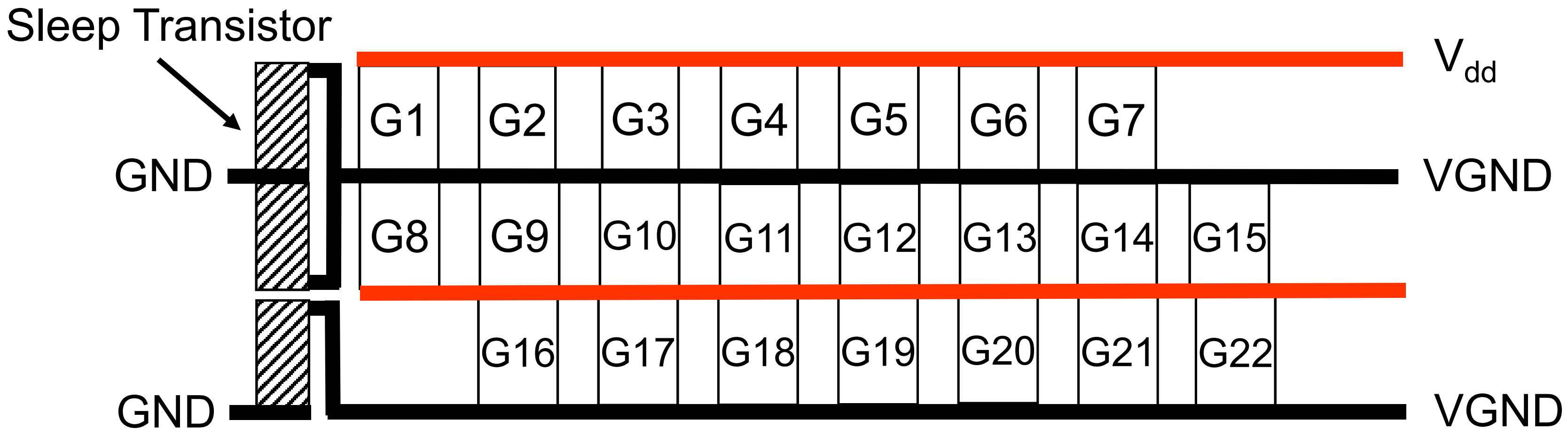
A 32-bit RISC Processor used in a PDA

Chip Size	Process	# Gates	Clock	Total Sleep Transistors Width	Power Dissipation
5.7mm × 5.7mm	0.18µm 5-metal	1,914K	333MHz	18mm	270mW

Leakage Power	Reduction
2µW	6000x

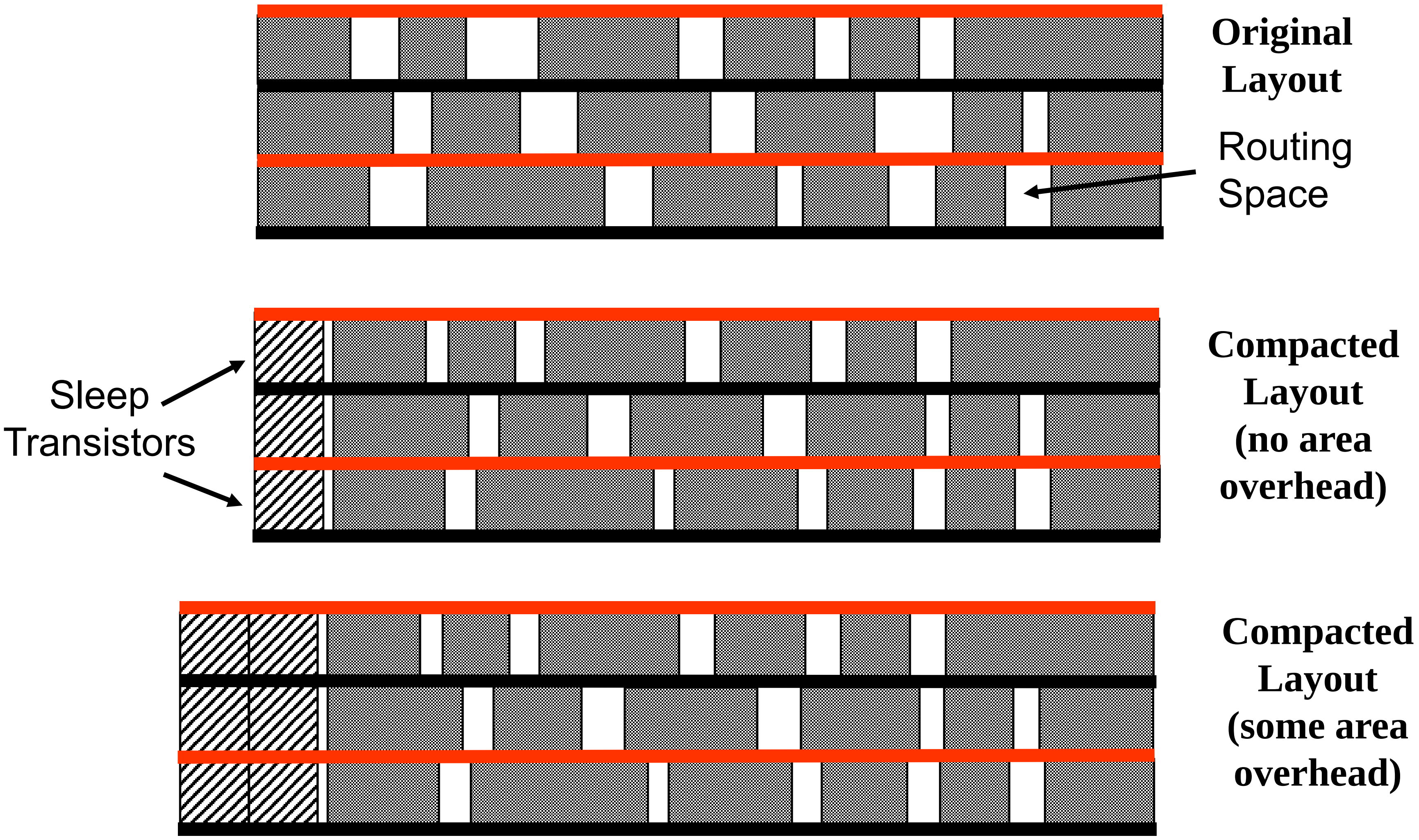
Post-Layout Leakage Power Minimization Based on Distributed Sleep Transistor Insertion





[Babighian-ISLPED04]

Post-Layout Leakage Power Minimization Based on Distributed Sleep Transistor Insertion



Original Layout

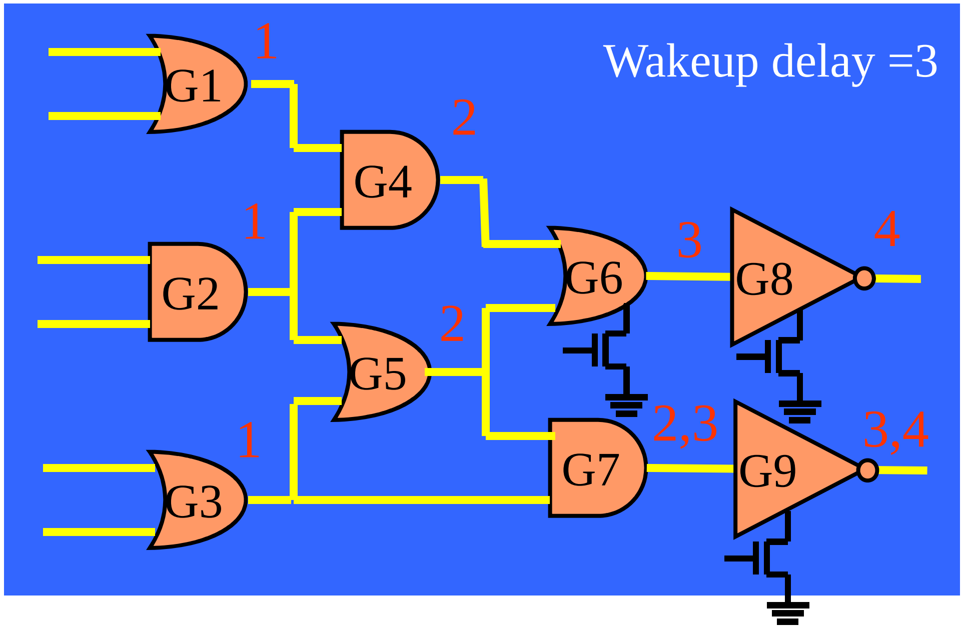
Routing Space

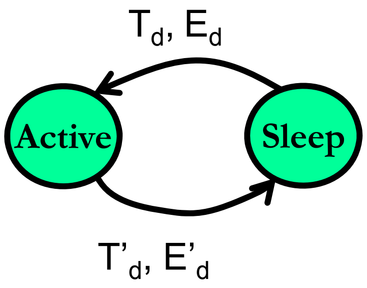
Sleep Transistors

Compacted Layout (no area overhead)

Compacted Layout (some area overhead)

Reducing the Re-activation (Wakeup) Delay



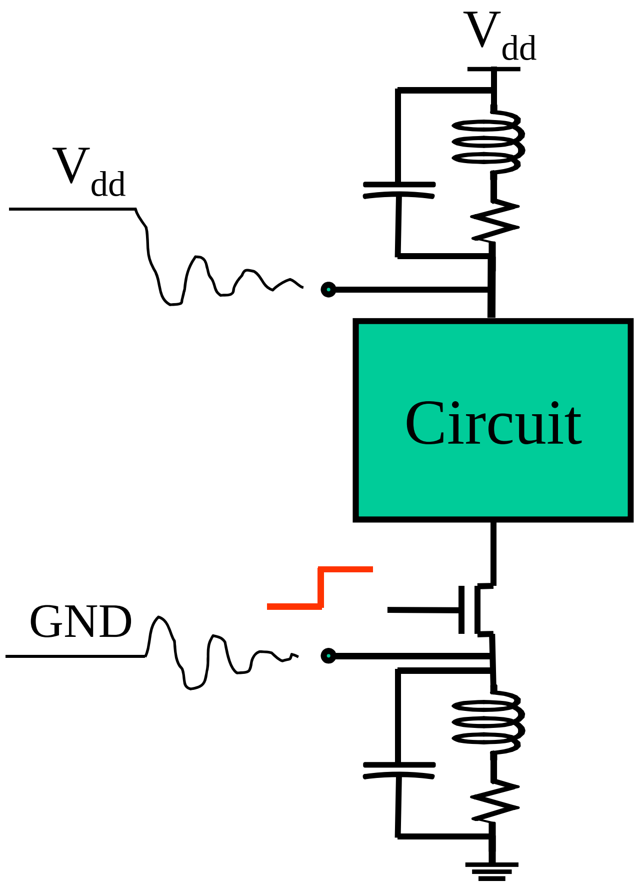


- Find minimum arrival times of inputs of all gates.
- Use sleep transistors for gates whose input arrival times are larger than the re-activation delay.
- Results (average): 80% leakage reduction, 19% overall power reduction, 2.5% area overhead, 5% delay overhead, and zero re-activation delay.

[Babighian-ISLPED04]

Minimizing Ground Bounce

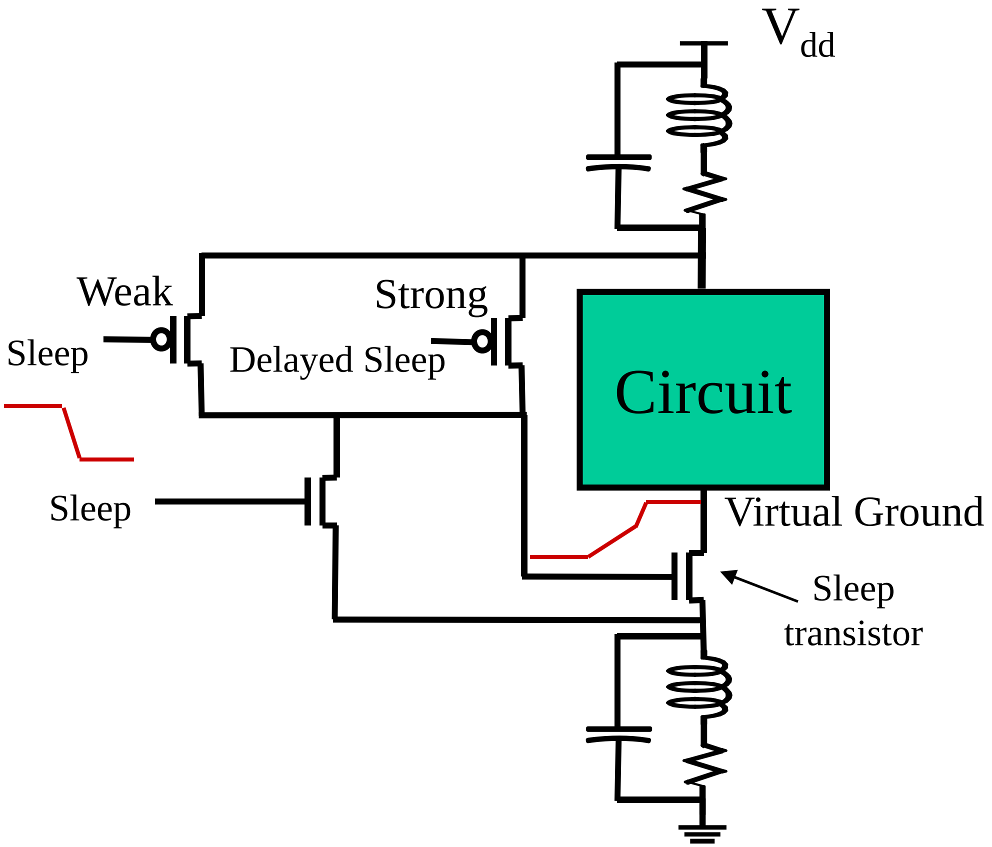
- During the sleep period, internal nodes are charged up.
- When the sleep transistor is turned on hard, there is a current spike flowing to the ground (due to the large V_{ds} of the sleep transistor).
- This sinks a lot of current into the GND terminal, potentially creating a large ground bounce noise.



[Kim-ISLPED03]

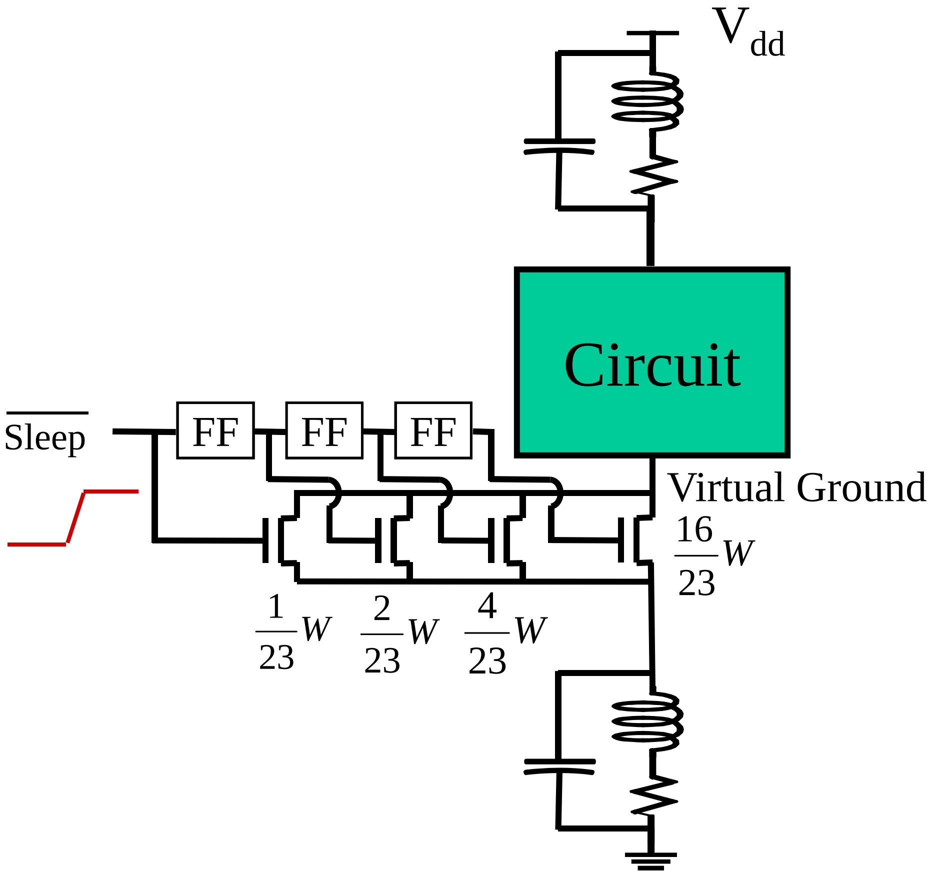
IBM's First Solution

- Turn-on the sleep transistor in two steps:
  1. Using a weak PMOS: $V_{gs} < V_{dd}$ for the sleep transistor. Originally, V_{ds} is high. So, the peak current is controlled.
  2. Using a strong PMOS: $V_{gs} = V_{dd}$ for the sleep transistor. V_{ds} is however low. Therefore, the peak current reduces.



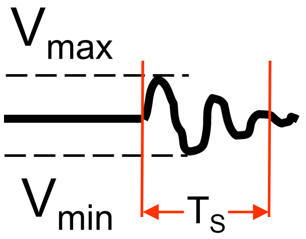
IBM's Second Solution

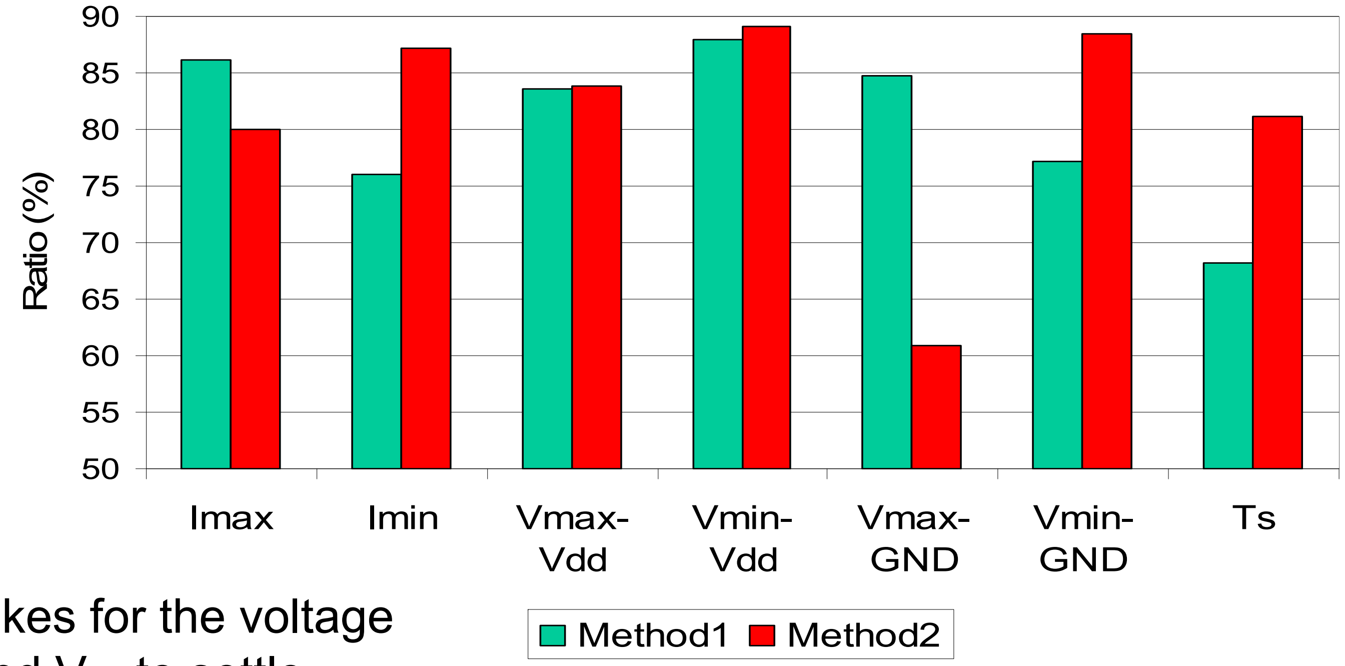
- Use several sleep transistors.
- Turn them on with some delay.
- The resistance between the virtual ground and the ground is reduced as the V_{ds} of the sleep transistor is lowered. This reduces the peak current.



Results

- Applied to a 16-bit ALU (with a multiplier)
- Designed at 0.13μm, 1.5V, operating at 500MHz.

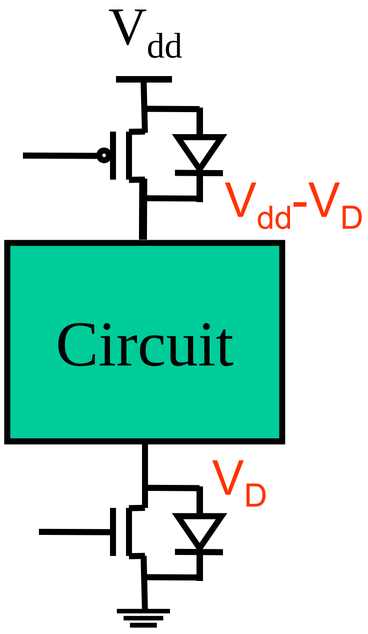




T_s is the time it takes for the voltage of both ground and V_{dd} to settle within ±5% of their final values.

Virtual Power/Ground Rail Clamp (VRC)

- Reduce the virtual supply and ground voltages using two diodes
  - This allows state retention.
  - It reduces noise during transition to active mode.
  - However, the leakage reduction is small.



[Kumagai-SVLSIC98]

Park Mode

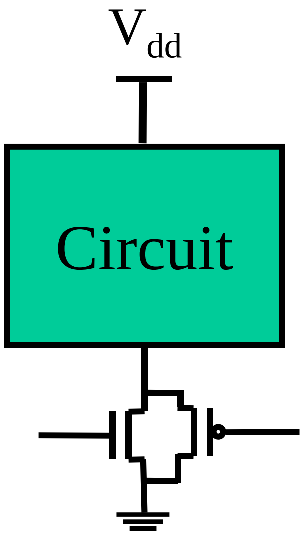
- Use a normally-on PMOS transistor to clamp the virtual ground (Park Mode)
  - Reduces leakage and bounce noise during wakeup.
  - Keeps the internal state.
- Turn off the PMOS transistor when in the Sleep Mode to achieve higher leakage saving. However, internal state will be lost.
- Empirical results for a 32-bit Carry Look Ahead adder designed in 0.13µm technology for various supply voltages.

	0.9V	1.5V
Park	2.3x	2.68x
Sleep	43x	22.8x

Leakage reduction compared to active mode

	1.1V	1.3V	1.5V
Park	2.44%	1.07%	2.83%
Sleep	7.42%	3.49%	6.26%

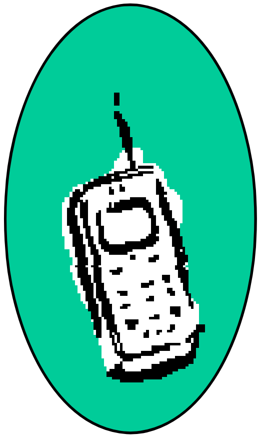
Clock Frequency Reduction



[Kim-ISLPED04]

Toshiba's Mixed MTCMOS and Dual V_{th} Method

- Used to reduce the leakage power in a DSP core for W-CDMA cell phones.
- Cell phones spend a significant amount of time in the standby mode.
  - High leakage power dissipation.
- Note however that the phones have to exchange some information with base stations every 100ms.



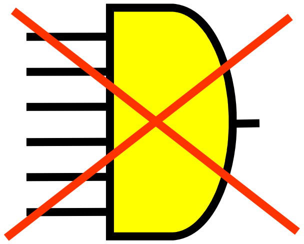
[Usami-ISLPED02]

Combining MTCMOS and Dual V_{th}

- Using MTCMOS for the entire circuit means the flip-flop values have to be saved and restored every 100ms.
    - Significant delay and power overhead.
- Solution: use MTCMOS for selected gates only (i.e., those with low V_{th} transistors which are on the timing-critical paths of the circuit). All other gates in the circuit use high V_{th} or dual V_{th} transistors.
    - Use Dual V_{th} for the flip-flops.
- Assume one sleep transistor per cell.
    - Simplifies the analysis.

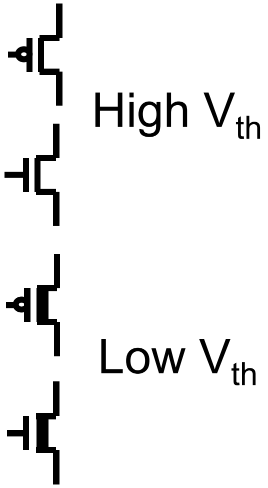
MTCMOS Cell Generation

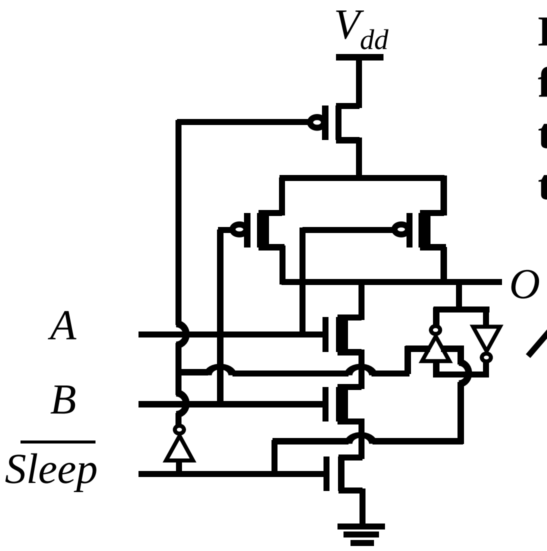
- Exclude Flip Flops and Latches
- Exclude cells with small drive
  - Unlikely to be on the critical path
- Exclude high fanin gates
  - Can be implemented by using 2-input gates.
- Develop complex library cells to speed up the timing-critical paths of the circuit, thereby, reducing the number of gates that must be implemented in MTCMOS.
- Overall 56 MTCMOS cells were developed using low-V_{th} transistors for logic.



The Floating Node Problem and Solutions

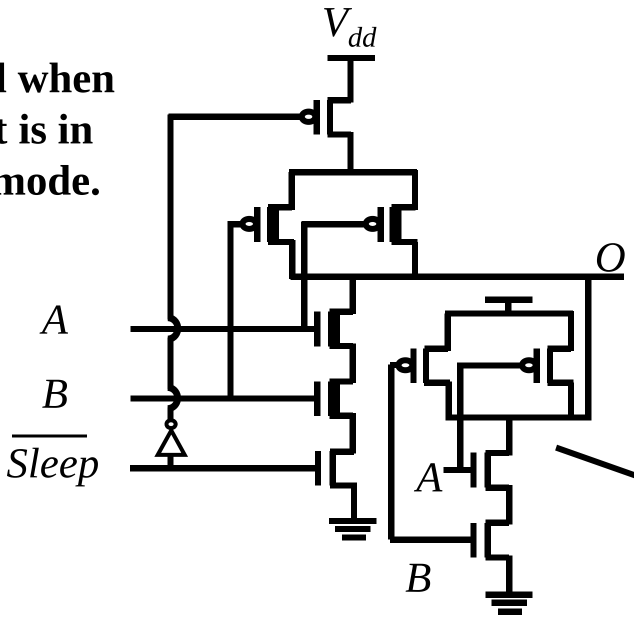
- An MTCMOS gate should not drive a regular gate (due to possibility of static current flow).
- Use a latch-type or bypass-type cell.
  - Note that two sleep transistors are used.





Latch Type

Latch is functional when the circuit is in the sleep mode.



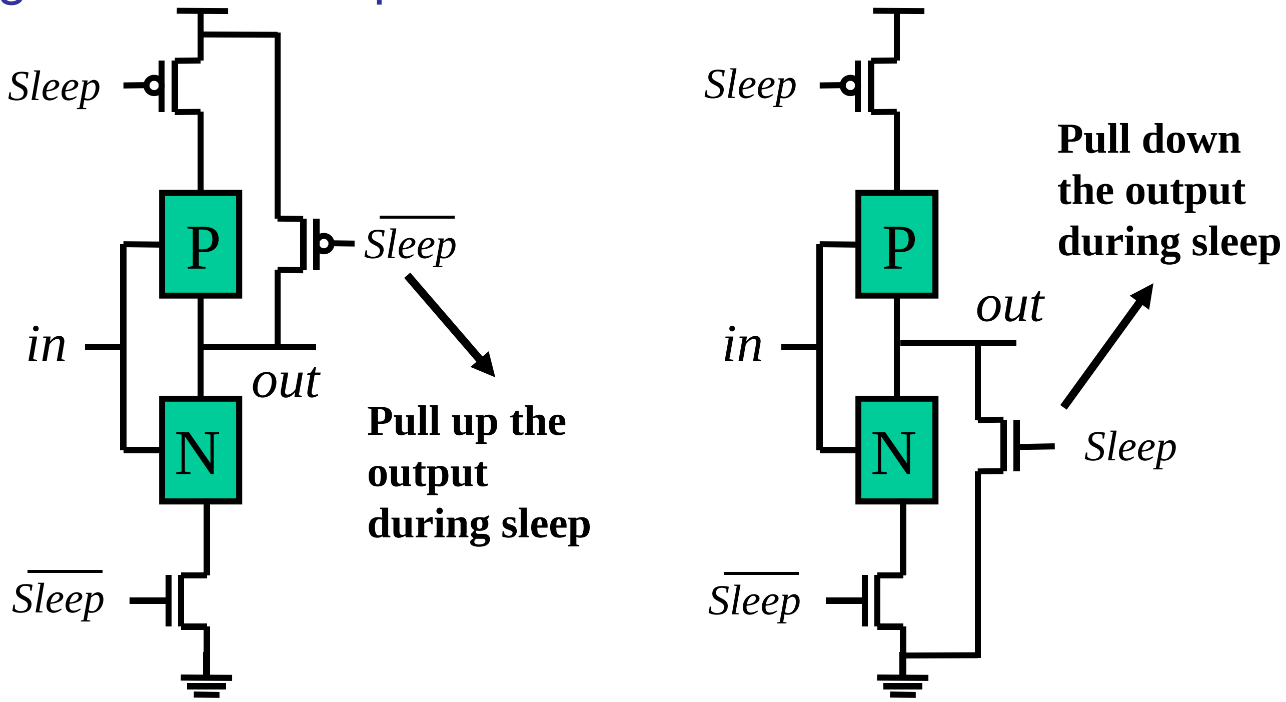
Bypass Type

The high V_{th} copy of the logic cell holds the output value during the sleep mode.

Another Solution

- Use transistors to pull-up or pull-down outputs of MTCMOS gates during the sleep mode
  - Smaller number of transistors, but almost the same area overhead.
    - The area overhead is dominated by the sleep transistor size.
  - Results in extra switching activity in the circuit every time the circuit goes to the sleep mode.

However, be careful when only one sleep transistor (say NMOS type) is used, in which case the output can only be set to one value (i.e., 1 in this case).



Pull up the output during sleep

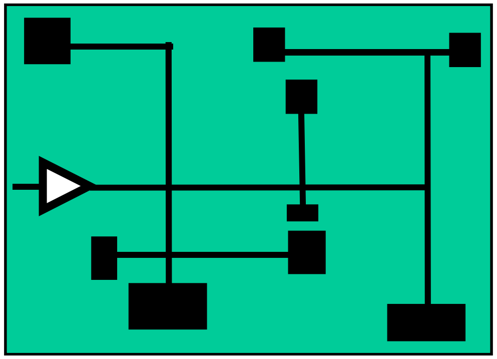
Pull down the output during sleep

Applying the Technique

- It is not possible to direct the conventional tools to use MTCMOS cells for critical paths and high-V_{th} cells for non-critical ones.
- A high-V_{th} circuit was developed first.
- Critical paths were identified.
- Cells on the critical paths were replaced by MTCMOS cells,
  - Started from output and continued backward until the timing constraint was met.

Driving Sleep Transistors

- Many sleep transistors and long wires.
  - Electro-migration problem, etc.
- A clock-tree-synthesis tool was used to generate a buffer-tree to drive the sleep transistors.
  - Only tree-construction and buffer placement, no skew of course.

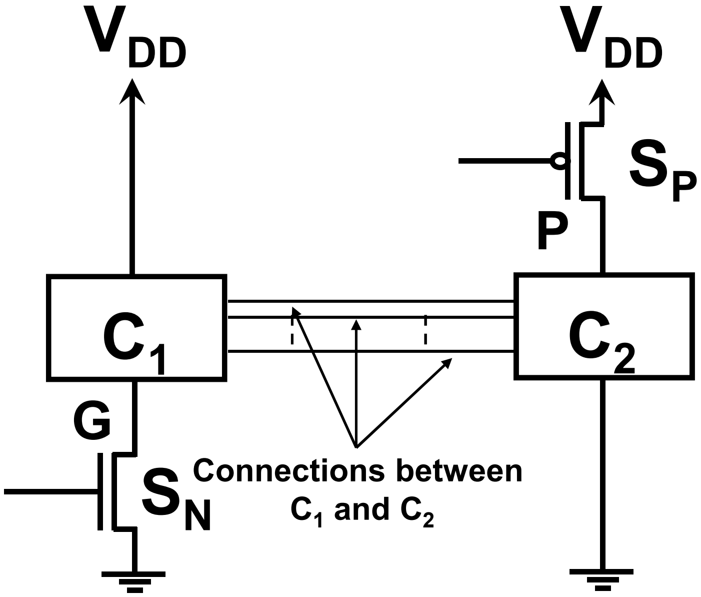


Experimental Setup and Result

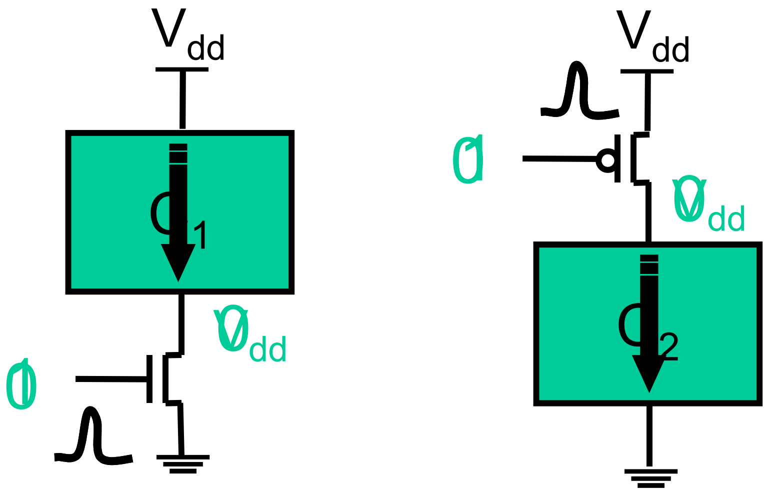
- Applied to a 34K-cell module.
- High-V_{th}=0.55V, Low-V_{th}=0.35V, 0.18μm, V_{dd}=1.5V @ 100MHz.
- 30 out of 53 levels of gates on the critical path were replaced by MTCMOS cells to meet the timing constraint.
  - Reduction from 10.27ns to 8.85ns (a 14% improvement).
- 12% of total cells were replaced with MTCMOS cells.
  - Area overhead=10%.
- Leakage at 85°C,
  - Active mode: 86μA
  - Standby mode: 28μA ≅ leakage of a high-V_{th} design

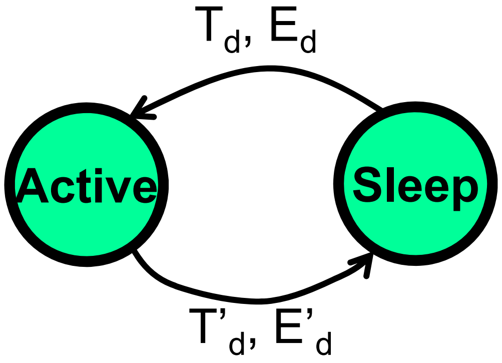
Charge Recycling MTCMOS

- Charge recycling technique uses both NMOS and PMOS sleep transistors.
- Circuit C is divided into 2 sub-circuits:
  - Sub-circuit C_1 is connected to S_N
  - Sub-circuit C_2 is connected to S_P

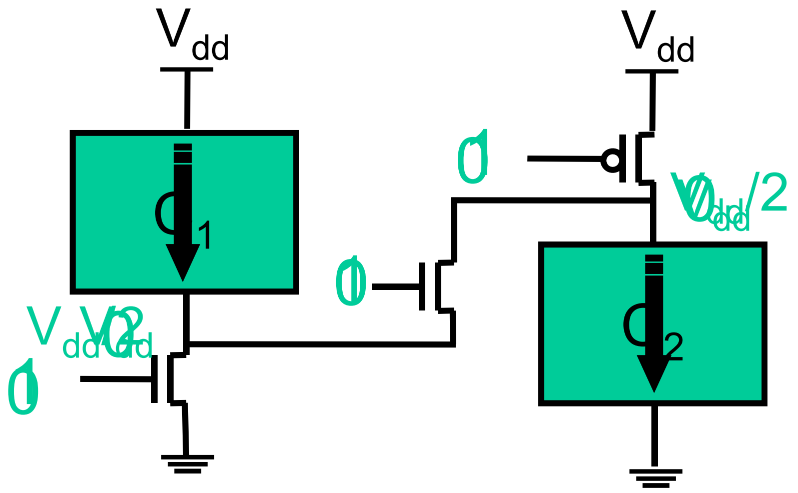


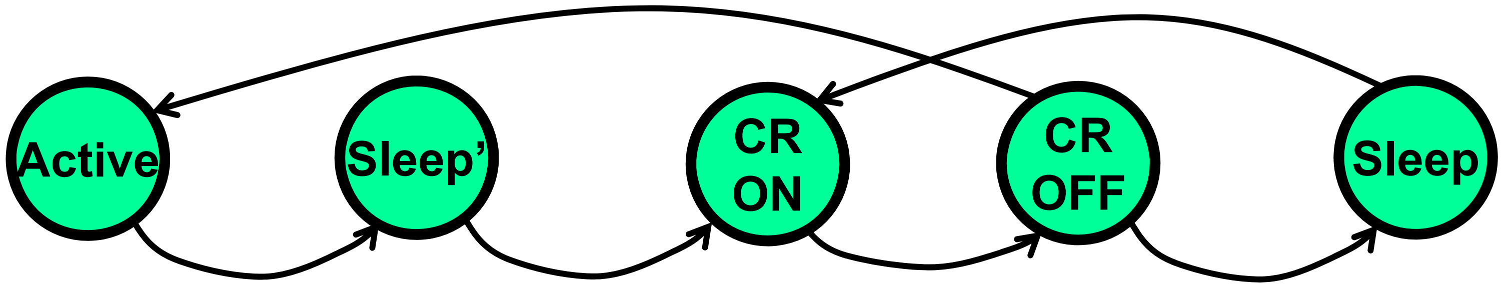
Mode Transitions in This Configuration





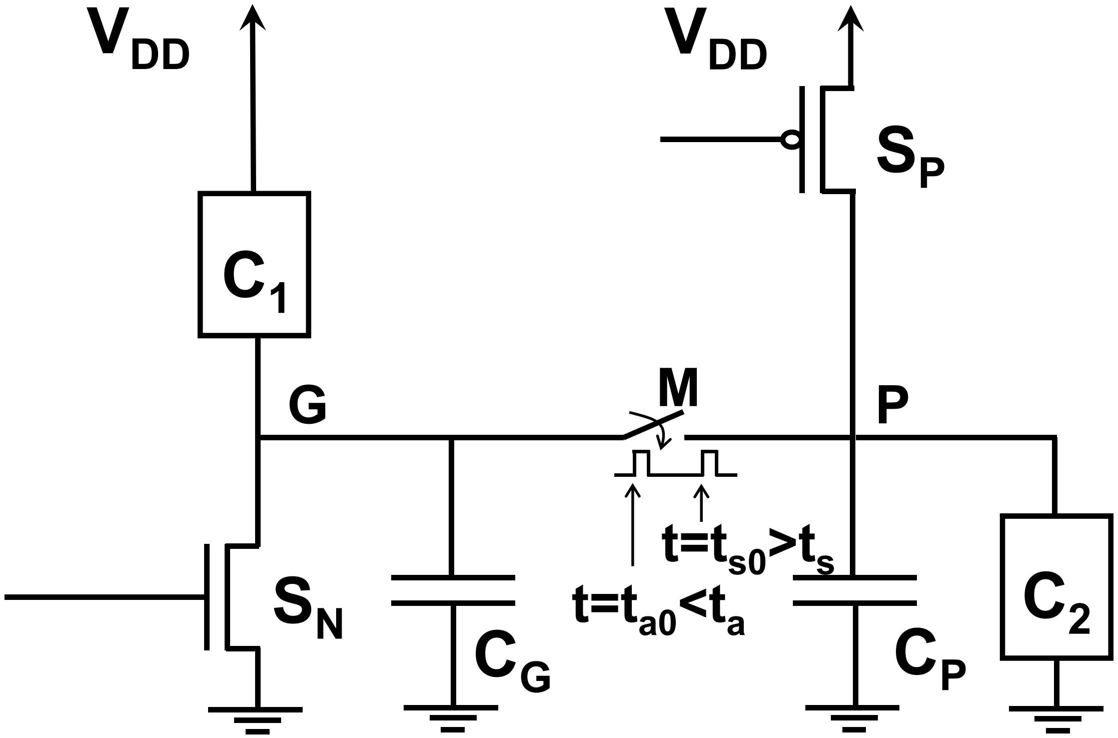
Our Solution: Charge Recycling (CR)





Energy Consumption in CR

- Replacing CR element with an ideal switch, M:



Energy Consumption in CR (cont.)

- Energy consumption during mode transition:

$$E_{sleep-active} = (1-\alpha)C_P V_{DD}^2$$

$$E_{active-sleep} = (1-\beta)C_G V_{DD}^2$$

where we have:

$$\alpha = \frac{C_G}{C_G + C_P} \quad \text{and} \quad \beta = \frac{C_P}{C_G + C_P}$$

$C_P =$ Total Virtual Power Capacitance

$C_G =$ Total Virtual Ground Capacitance

Energy Saving Ratio (ESR) in CR

- Energy consumption in one cycle for the conventional MTCMOS and CR-MTCMOS:

$$E_{conv.} = C_G V_{DD}^2 + C_P V_{DD}^2$$

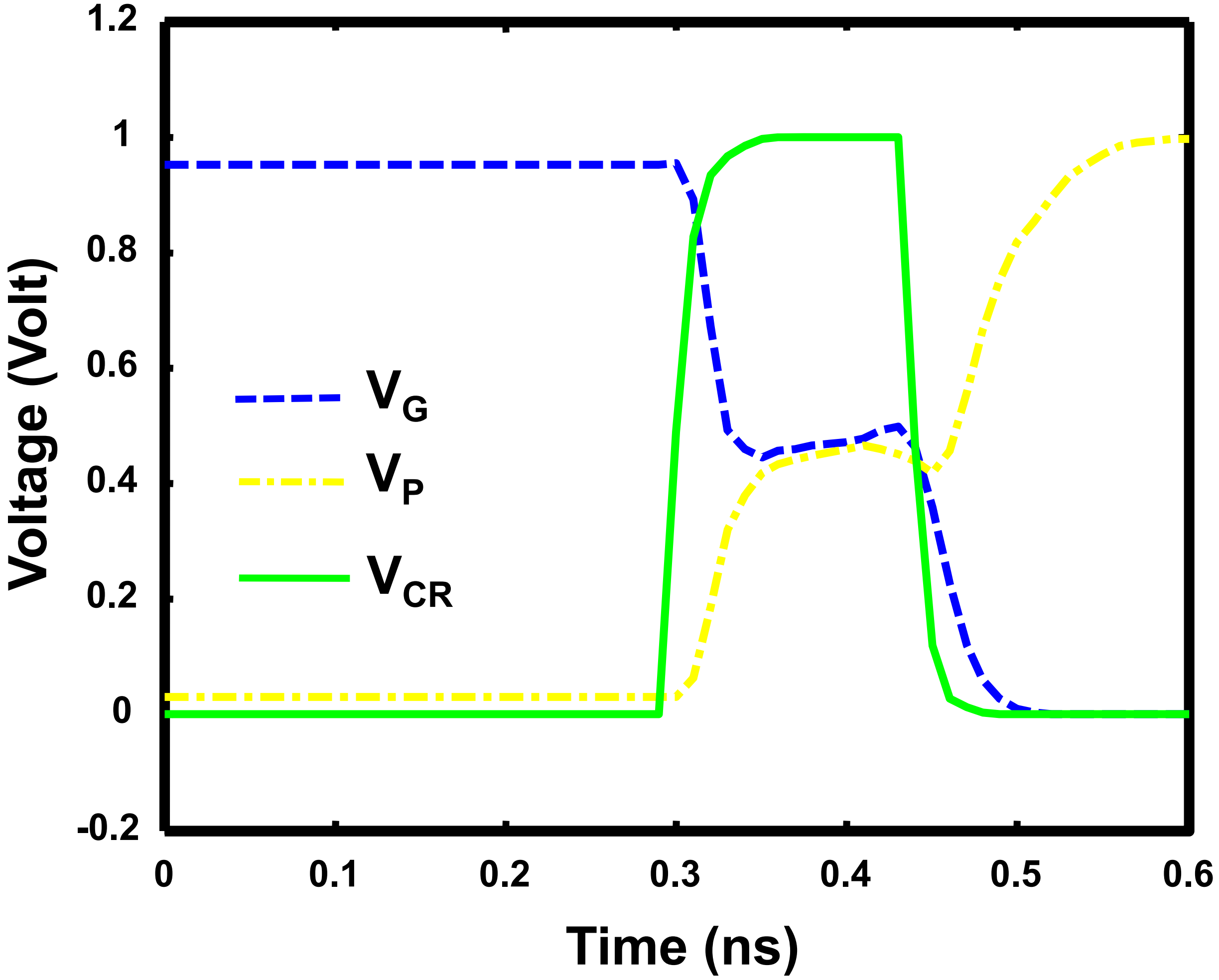
$$E_{CR} = \alpha C_G V_{DD}^2 + \beta C_P V_{DD}^2$$

- The energy saving ratio is:

$$ESR(X) = \frac{E_{total} - E_{cr\_total}}{E_{total}} = \frac{2X}{(1+X)^2}$$

- ESR is maximum when $X=1$, i.e., when $C_G = C_P$.

Charge Recycling Operation



Effect of the Threshold Voltage

- Condition for a complete CR (assuming C_G=C_P):

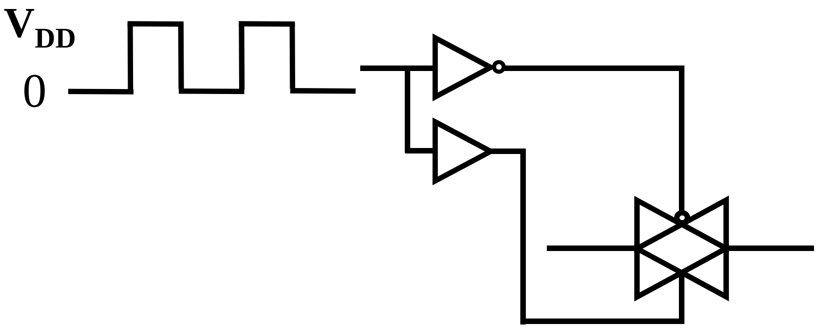
$$Min\left\{V_{tn},\ \left|V_{tp}\right|\right\}\leq\frac{V_{DD}}{2}$$

  - If $V_{tn}=|V_{tp}|$, then a simple NMOS will be adequate.

- Solution: decrease $\mathrm{V_t}$
  - Trade off: leakage current increases.

Effect of Transistor Sizing

- The larger the transmission gate (TG), the faster the charge recycling operation

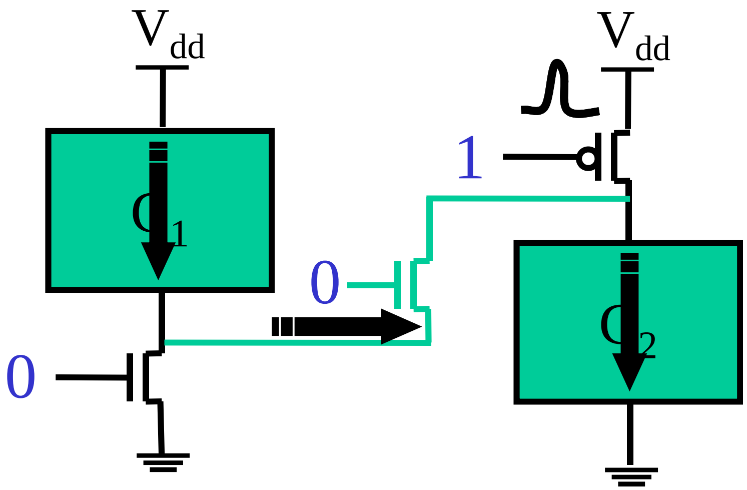


$$\frac{E_{tg-total}}{E_{total}} = \frac{4C_{tg}V_{DD}^2}{(C_G + C_P)V_{DD}^2} = \frac{4C_{tg}}{C_G + C_P}$$

- Trade off: larger TG switching power penalty
  - C_{tg} denotes the input cap of NMOS and PMOS in TG.

Leakage Analysis

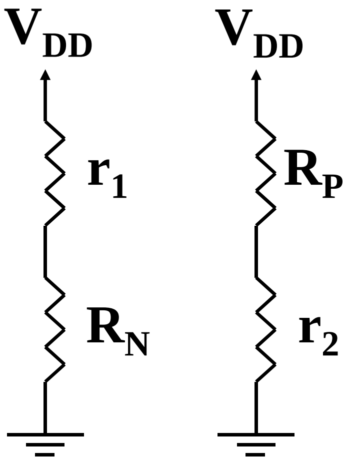
- TG adds a new leakage path:



- Transistors in the TG must be high V_t transistors.

Leakage Paths in Conventional Technique

- The equivalent leakage model for the sleep mode:

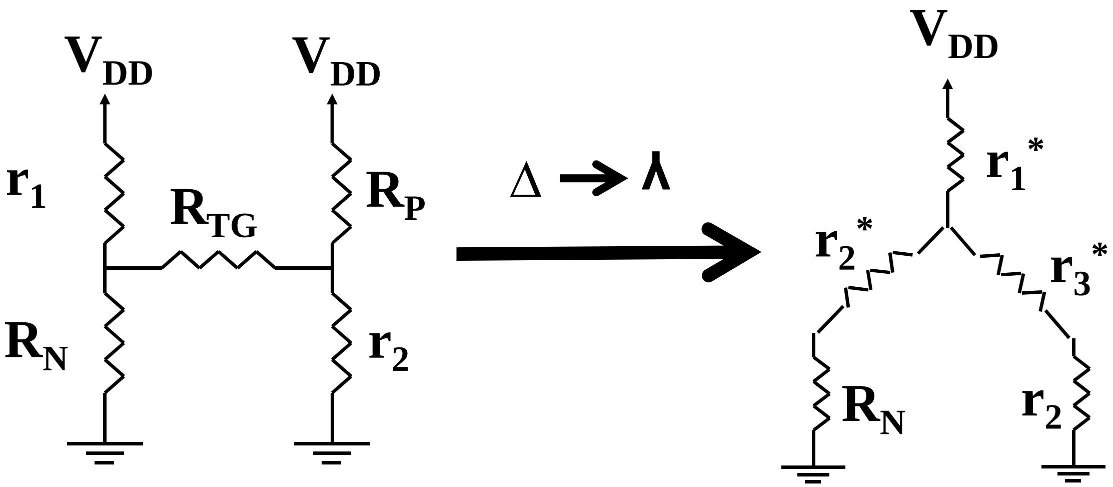


- Leakage is calculated by writing KVL equations (R_N=R_P=R):

$$P_{leakage-conv.} = \frac{2V_{DD}^2}{R}$$

Leakage Paths in CR Technique

- The equivalent leakage model in the sleep mode:



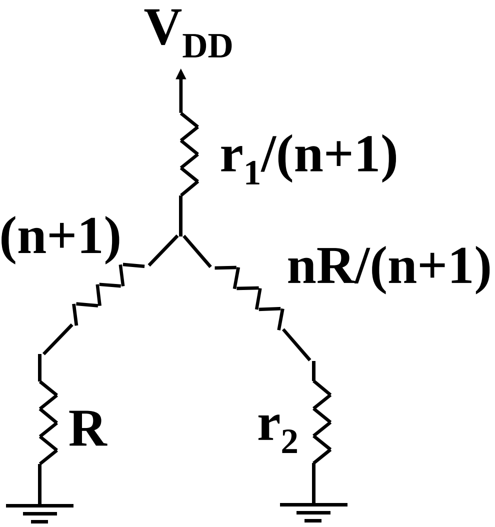
where (R_N=R_P=R and R_{TG}=nR):

$$r_1^* = \frac{r_1 R_P}{r_1 + R_{TG} + R_P} = \frac{1}{n+1} r_1$$

$$r_2^* = \frac{r_1 R_{TG}}{r_1 + R_{TG} + R_P} = \frac{n}{n+1} r_1$$

$$r_3^* = \frac{R_P R_{TG}}{r_1 + R_{TG} + R_P} = \frac{n}{n+1} R$$

Leakage in CR Technique



- Leakage is calculated by writing KVL equations:

$$P_{leakage-CR} = \left(2 + \frac{1}{n}\right)\frac{V_{DD}^2}{R}$$

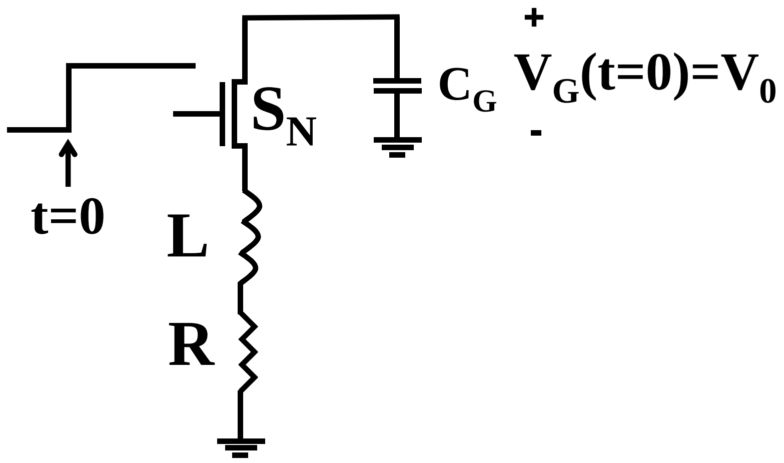
- Leakage has increased by a factor of 1/2n .

Leakage in CR Technique (cont.)

- If n=2, there is 25% increase in the leakage:
  - For short and medium sleep periods, this increase is negligible compared to the saving that we get from the CR technique
  - For long sleep periods, we must use larger *n* by choosing transistors with smaller W/L ratios in the TG
  - This is also beneficial from the layout area point of view
  - Potential disadvantage: CR takes longer to complete.

Ground Bounce (GB) Analysis

- Simple wake-up circuit model for GB analysis:



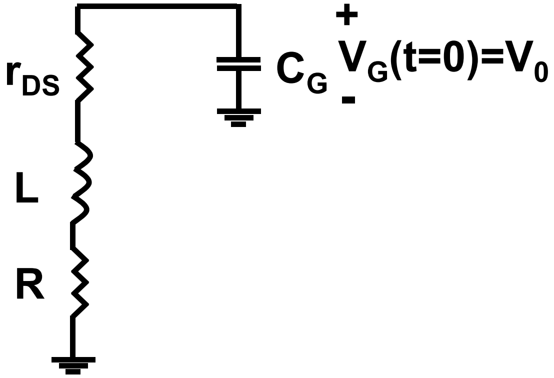
- For conventional MTCMOS: $V_0=V_{DD}$
- For CR MTCMOS: $V_0=V_{DD}/2$

Ground Bounce Analysis (cont.)

- It is well known that:
  - The positive GB peak occurs when S_N operates in saturation region
  - When operating in the saturation region, drain-source current of S_N, and thus the GB value, dose not depend on the V_0 value
  - In CR-MTCMOS, the positive GB peak value remains unchanged
  - The negative GB peak occurs when S_N operates in linear region. This changes in CR-MTCMOS.

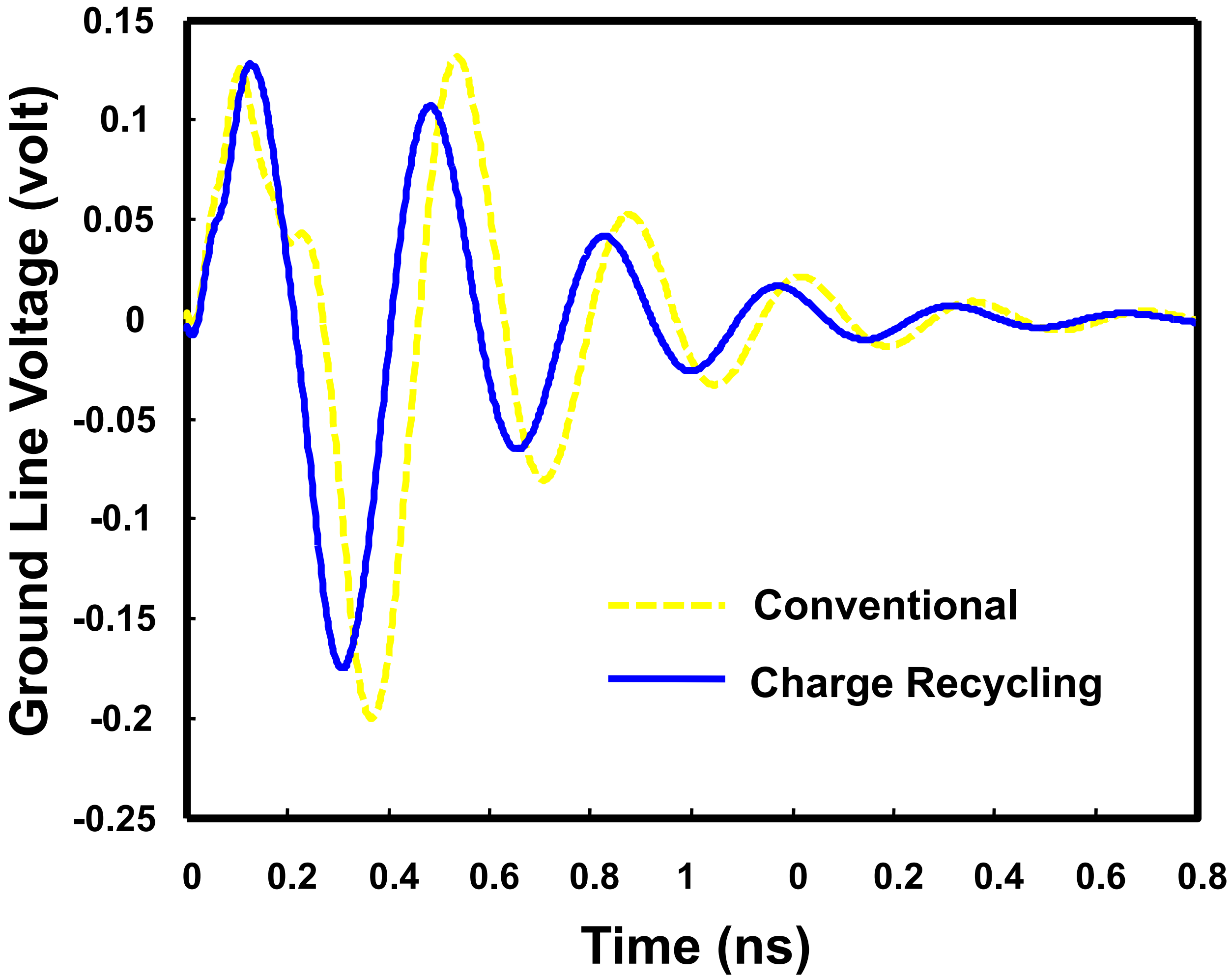
Ground Bounce Analysis (cont.)

- Equivalent circuit model when the negative GB peak happens (r_{DS} is the ON resistance of S_N):



- RLC circuit with initial voltage value V_0 on C_G.
- In CR-MTCMOS:
  - V_0 is reduced by 50%
  - Negative GB peak value is reduced
  - Settling time is lowered.

Ground Bounce



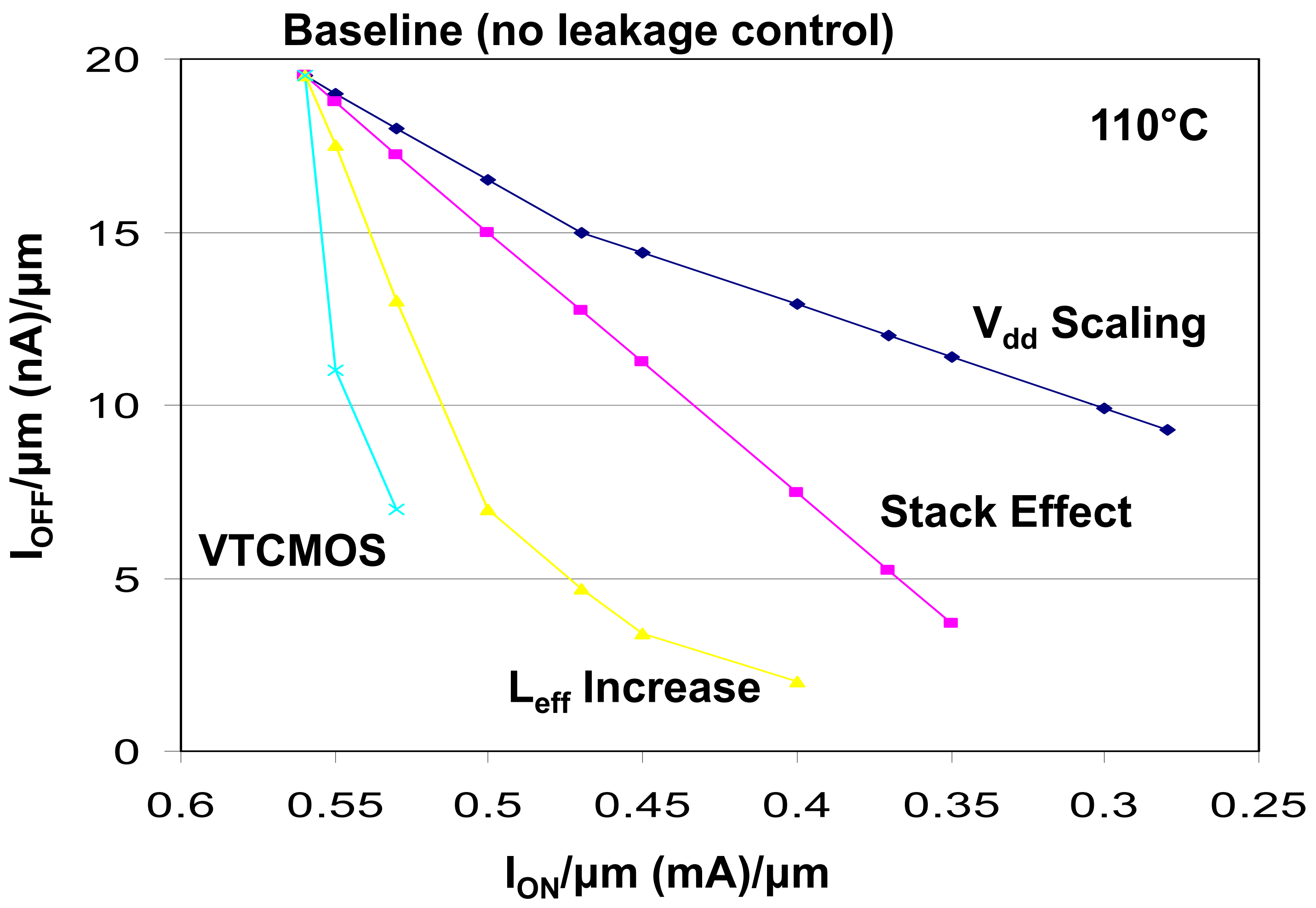
Experimental Results for the 90nm CMOS Node

Circuit	Wake up Time (ps)		Mode Transition Energy Cons. (pJ)		Energy Saving (%)	Wake Up Time Reduction (%)
	Conv.	CR	Conv.	CR		
9Sym	494	489.61	29	16	45%	0.9%
C432	240	232.73	10	5.7	43%	3%
C1355	132	125.42	12	7.2	40%	5%
C1908	267	275.63	38	20.5	46%	-3%
C2670	578	573	123	72.6	41%	0.9%
C3540	1500	1545	490	276.9	43%	-3%
C5315	1320	1307	638	357.3	44%	0.1%
C6288	2100	2047	1047	628.2	40%	2.5%
C7552	2310	2402	1532	842.6	45%	-4%

Leakage Reduction for 130nm Technology

Technique	Simulation Results	Theoretical Model
Reduction in V_{dd} by 30%	2.2X	1.9X
Increase in L_{eff} by 30%	9.3X	8.7X
Stack Effect	12.0X	11.5X
Reverse Bias by 30% of V_{dd}	2.3X	2.1X

I_{off} – I_{on} Curve



Normalized I_{off}/I_{on} Degradation: Scaling Trends

$\zeta = \frac{\partial I_{OFF}}{\partial I_{ON}} \Big/ \frac{I_{OFF}}{I_{ON}}$	Changing V_{dd}	Changing L_e	Stack Effect	VTCMOS
130nm	1.1	3.1	2.2	20
100nm	1	3.1	2.1	9
70nm	0.8	2.8	1.9	7.5

Higher values are better.

Conclusions

- Leakage currents are rising fast and must be controlled by circuit design and optimization tools
- Gate leakage is rising at the fastest rate, but is expected to be controlled by the introduction of high-K dielectric material; thus, subthreshold leakage remains the most worrisome component of standby power dissipation
- Voltage islands, Dual-V_{th} designs, and MTCMOS technique appear to be the most effective solutions for minimizing the subthreshold leakage current.
- Most of the power reduction techniques described depend on trading off timing slack for power at the cell level. Even the block-level approaches are often combined with the cell level approaches.
- Thus most designs driving for low-power, high performance applications will require analysis across multiple voltages, whether supply, bias or both.

References (I)

- Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998.
- D. K. Schroder, *Semiconductor Material and Device Characterization*, 2nd ed. J. Wiley, New York, NY, 1998.
- F. Fallah and M. Pedram, "Standby and active leakage current control and minimization in CMOS VLSI circuits." *IEICE Transactions Fundamentals of Electronics, Communications and Computer Sciences*, 2005.
- K. Roy, S. Mukhopadhyay, H. Mahmoodi, “Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits,” *Proc. of the IEEE*, Vol. 91, No. 2, Feb. 2003.
- Y. Taur, “CMOS scaling and issues in sub-0.25 um systems,” in *Design of High-Performance Microprocessor Circuits*, A. Chandrakasan, W. J. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, ch. 2, pp. 27–45.
- L.T. Clark, R. Patel, T. S. Beatty, “Managing standby and active mode leakage power in deep sub-micron design,” Int’l Symposium on Low Power Electronics and Design, Aug. 2004, pp. 274-279
- V. De, Y. Ye, A. Keshavarzi, S. Narendra, J. Kao, D. Somasekhar, R. Nair, and S. Borkar, “Techniques for leakage power reduction,” in Design of High-Performance Microprocessor Circuits, A. Chandrakasan, W. Bowhill, and F. Fox, Eds. Piscataway, NJ: IEEE, 2001, ch. 3, pp. 52–55.

References (II)

- S. Borkar, “Design Challenges of Technology Scaling”, *IEEE MICRO*, July-August 1999.
- B. Davari et al., “CMOS Scaling for High Performance and Low Power - The Next Ten Years,” *Proc. of the IEEE*, vol. 87, no. 4, Apr. 1999, pp. 659-667.
- Y Taur, “CMOS design near the limit of scaling,” *IBM Journal of Research and Development*, Volume 46, Numbers 2/3, 2002.
- D. Frank, “Power-constrained CMOS scaling limits,” *IBM Journal of Research and Development*, Volume 46, Numbers 2/3, 2002.
- R. Puri, L. Stok, J. Cohn, D. Kung, D. Pan, D. Sylvester, A. Srivastava, “Pushing ASIC Performance in a Power Envelope,” *Proc. of Design Automation Conference*, Jun. 2003.
- J.M.C. Stork, “Technology Leverage for Ultra-Low Power Information Systems,” *HP Labs Technical Reports*, Dec. 2005.
- A. Chandrakasan, S. Narendra, J. Kao, “Subthreshold leakage modeling and reduction techniques,” *Proc. of Int’l Conf. on Computer Aided Design*, Nov. 2002, pp. 141-148.
- B. Chatterjee, M. Sachdev, S. Hsu, R. Krishnamurthy, S. Borkar, “Effectiveness and scaling trends of leakage control techniques for sub-130nm CMOS technologies,” Proc. of ISLPED, Aug. 2003, pp. 122-127.

References (III)

- T. Kuroda, et al., “A 0.9V, 150MHz, 10mW, 4mm2, 2-D discrete cosine transform core processor with variable-threshold-voltage scheme,” *ISSCC Digest of Technical Papers*, Feb. 1997, pp.166-167.
- S. Mutoh, et al., “1-V Power Supply High-Speed Digital Circuit Technology with Multi-threshold Voltage CMOS,” *IEEE Journal of Solid-state Circuits*, Aug. 1995, pp. 847-854.
- Z. Cheng, M. Johnson, L. Wei, and K. Roy, “Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks,” *Int’l Symp. On Low Power Electronics and Design*, Aug, 1998, pp. 239-244.
- E. Acar, A. Devgan, R. Rao, Y. Liu, H. Su, S. Nassif, J. Burns, “Leakage and leakage sensitivity computation for combinational circuits, “ *Proc. of the international symposium on Low power electronics and design*, Seoul, Korea, Aug. 2003, pp. 96-99.
- L.Wei, Z. Chen, M. Johnson, K. Roy, Y. Ye, and V. De, “Design and optimization of dual threshold circuits for low voltage low power applications,” *IEEE Trans. VLSI Systems*, pp. 16–24, Mar. 1999.