Gate Leakage Analysis and Reduction in Nanoscale CMOS circuits

Saraju P. Mohanty
University of North Texas, Denton, TX 76203.
Email: smohanty@cs.unt.edu
Homepage: http://www.cs.unt.edu/~smohanty/

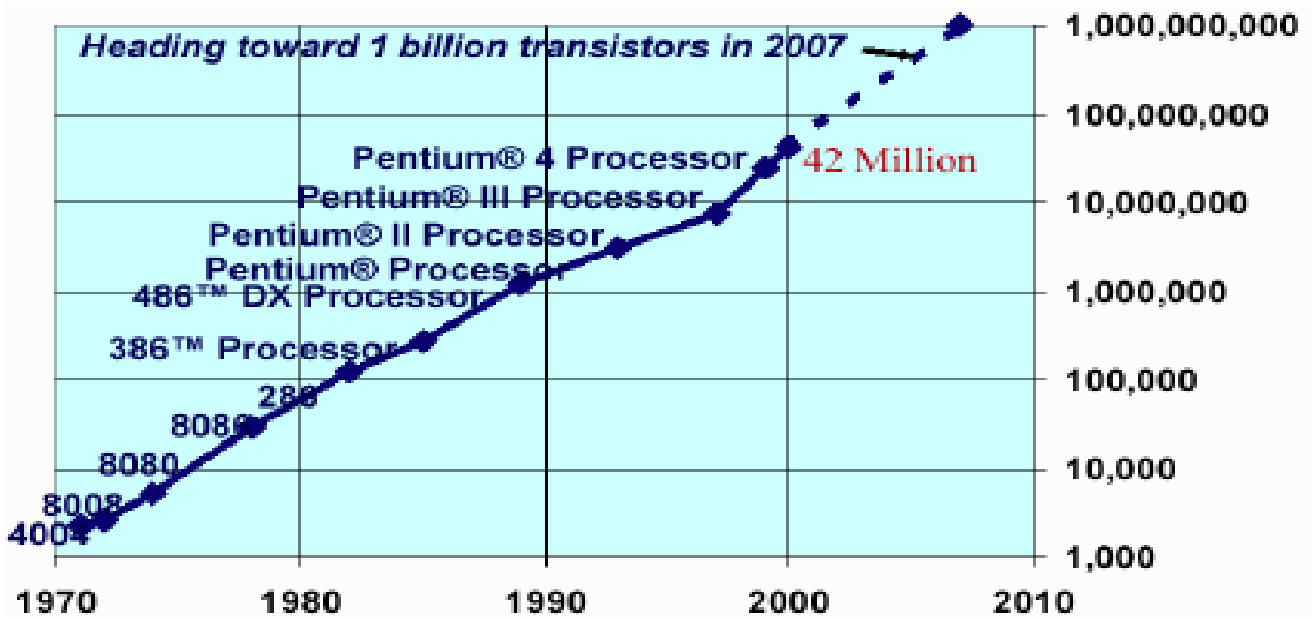
Outline of the Talk

- CMOS scaling –Trends and Effects
- Power consumption redistribution due to scaling
  - Components of Power Dissipation
  - Components of Leakage
- Gate leakage analysis
- Gate leakage variation with process and design parameters
- Gate leakage reduction

CMOS Driven Applications

Almost the entire industry today is driven by CMOS

Scaling Trend – Transistor Count



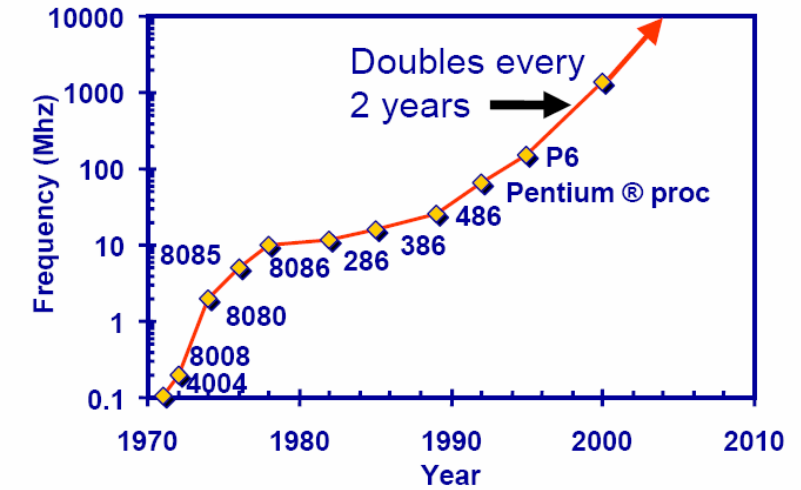
Increase in Transistor Count per chip

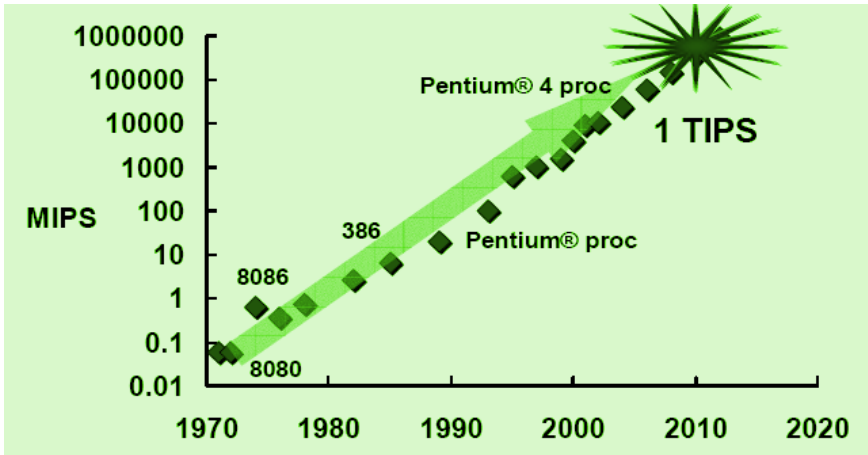
1967

VLSI technology is the fastest growing technology in the human history.

2007

Scaling Trend – Frequency and Performance

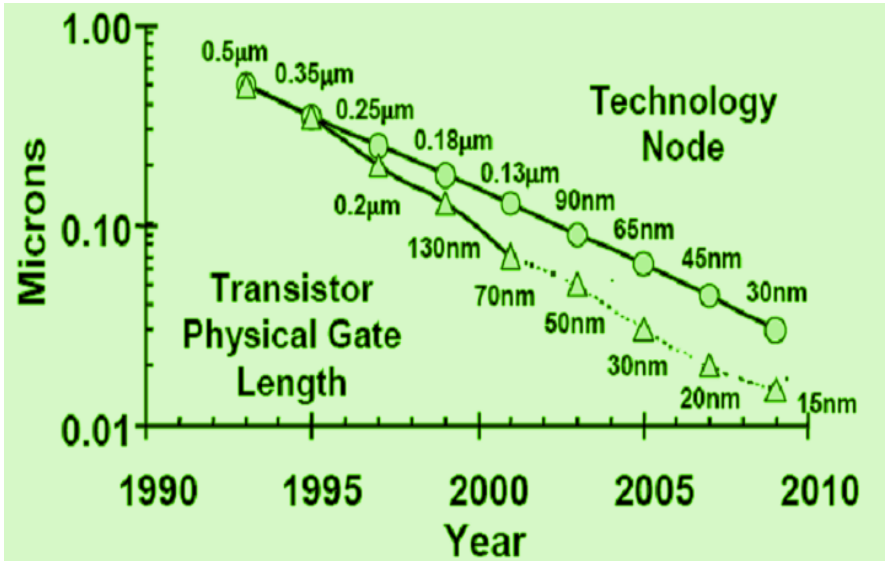


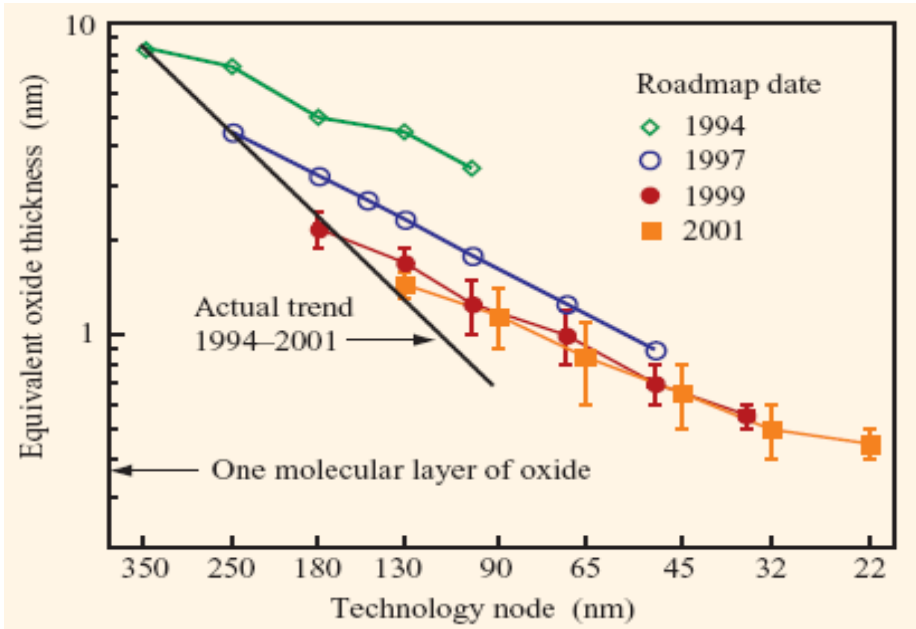


- With scaling the transistors are becoming twice as fast as the previous generation.
- Applications are also being targeted for TIPS level performance.

Source: Pedram ASPDAC 2004

What is Physically Scaled ? (Gate Length and Gate thickness)

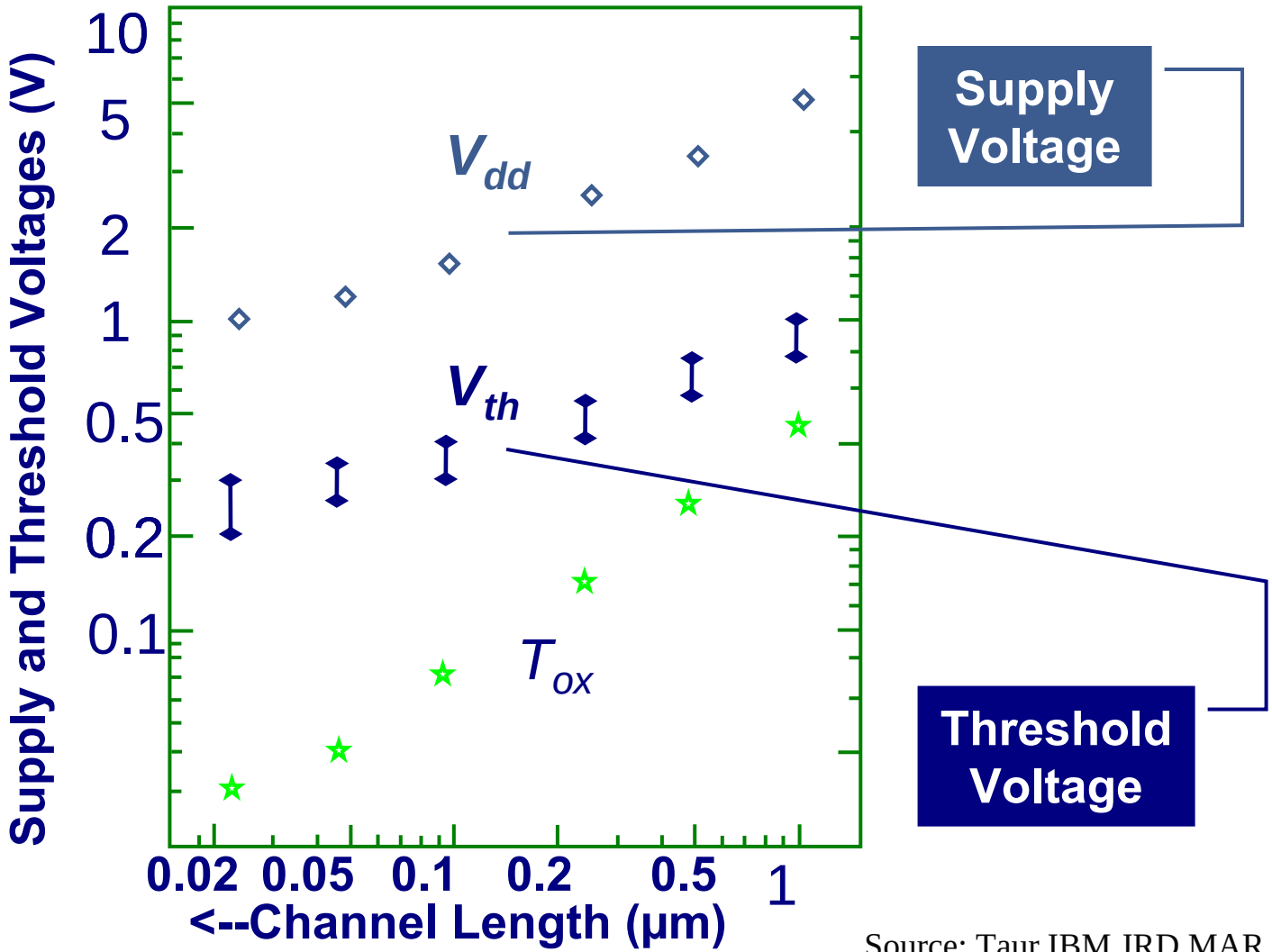




- Gate length of the transistor has been decreasing with technology scaling.
- All the other dimensions including gate oxide thickness have been scaled down to support this trend

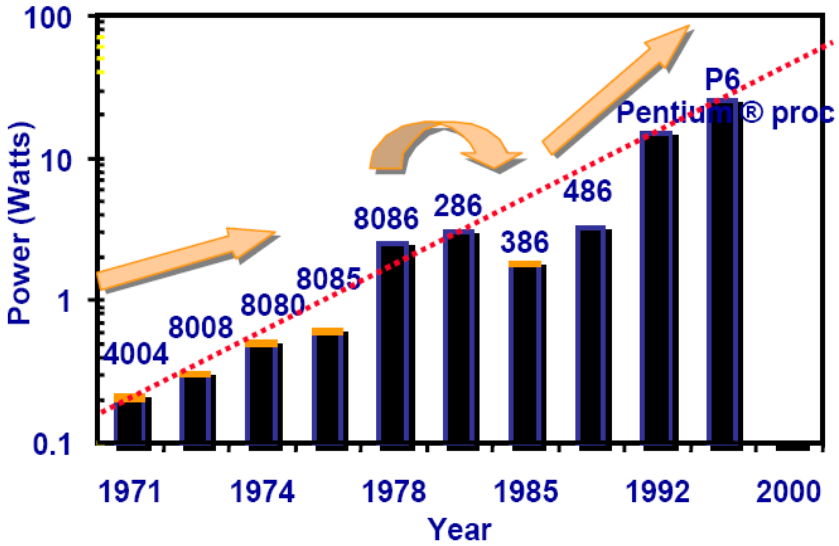
Source: Pedram ASPDAC 2004, Osburn IBM JRD Mar2002

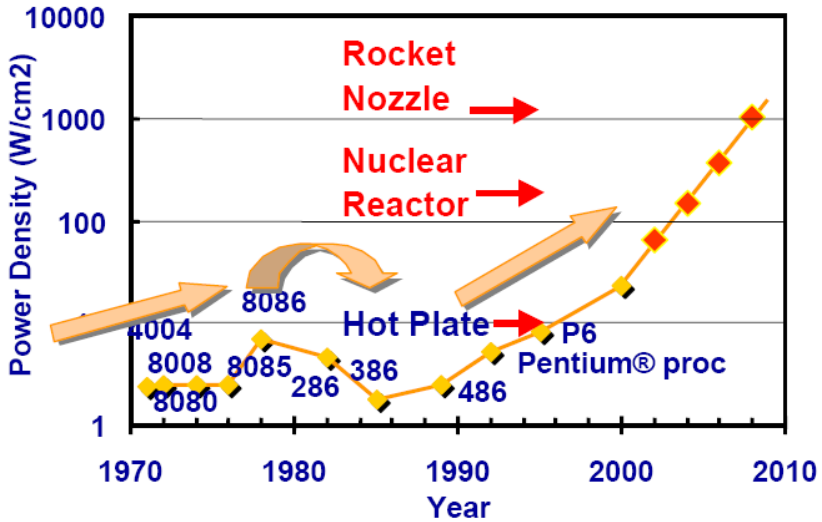
Other Parameters Scaled?



Source: Taur IBM JRD MAR 2002

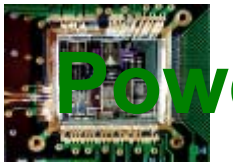
Scaling Trend – Power Dissipation





- Power dissipated by the transistor has manifested itself most emphatically along with scaling.
- The power density is increasing exponentially

Source: Intel



Power Dissipation Components in CMOS

Total Power Dissipation

Static Dissipation

- Sub-threshold current
- Tunneling current
- Reverse-biased diode Leakage
- Contention current

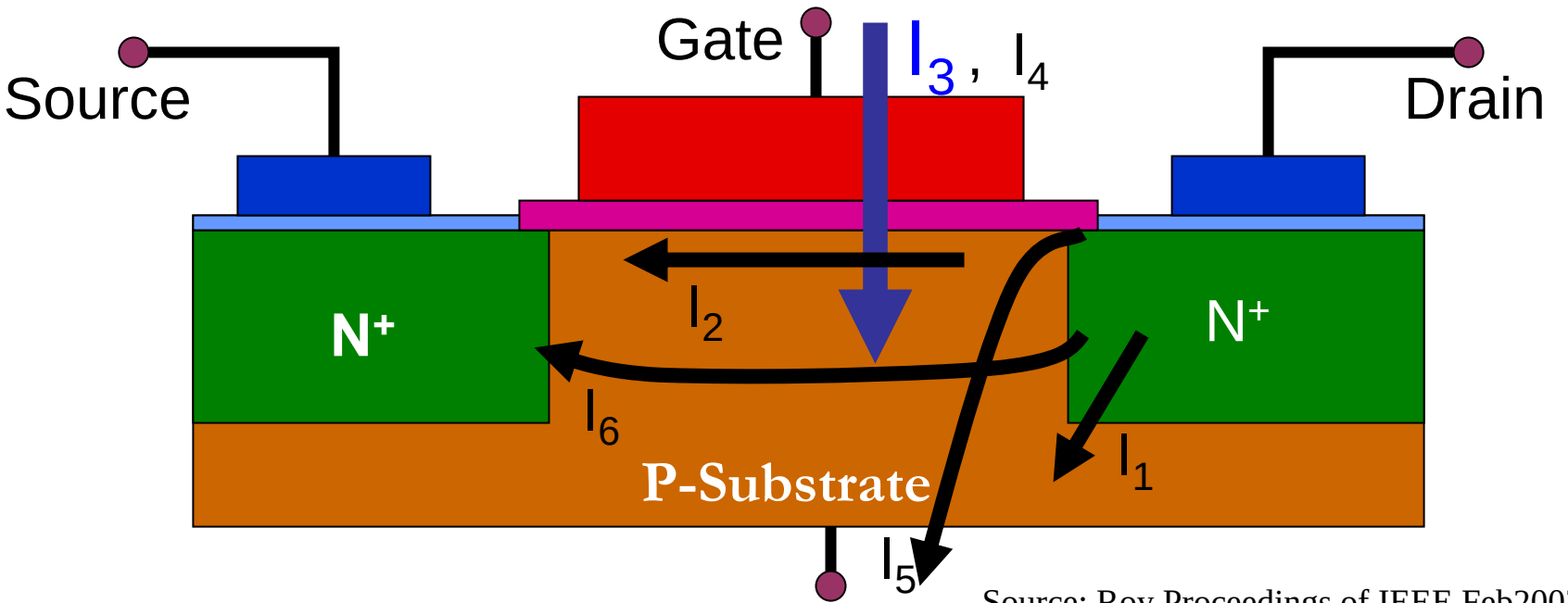
Dynamic Dissipation

- Capacitive Switching
- Tunneling current
- Short circuit

Source: Weste and Harris 2005

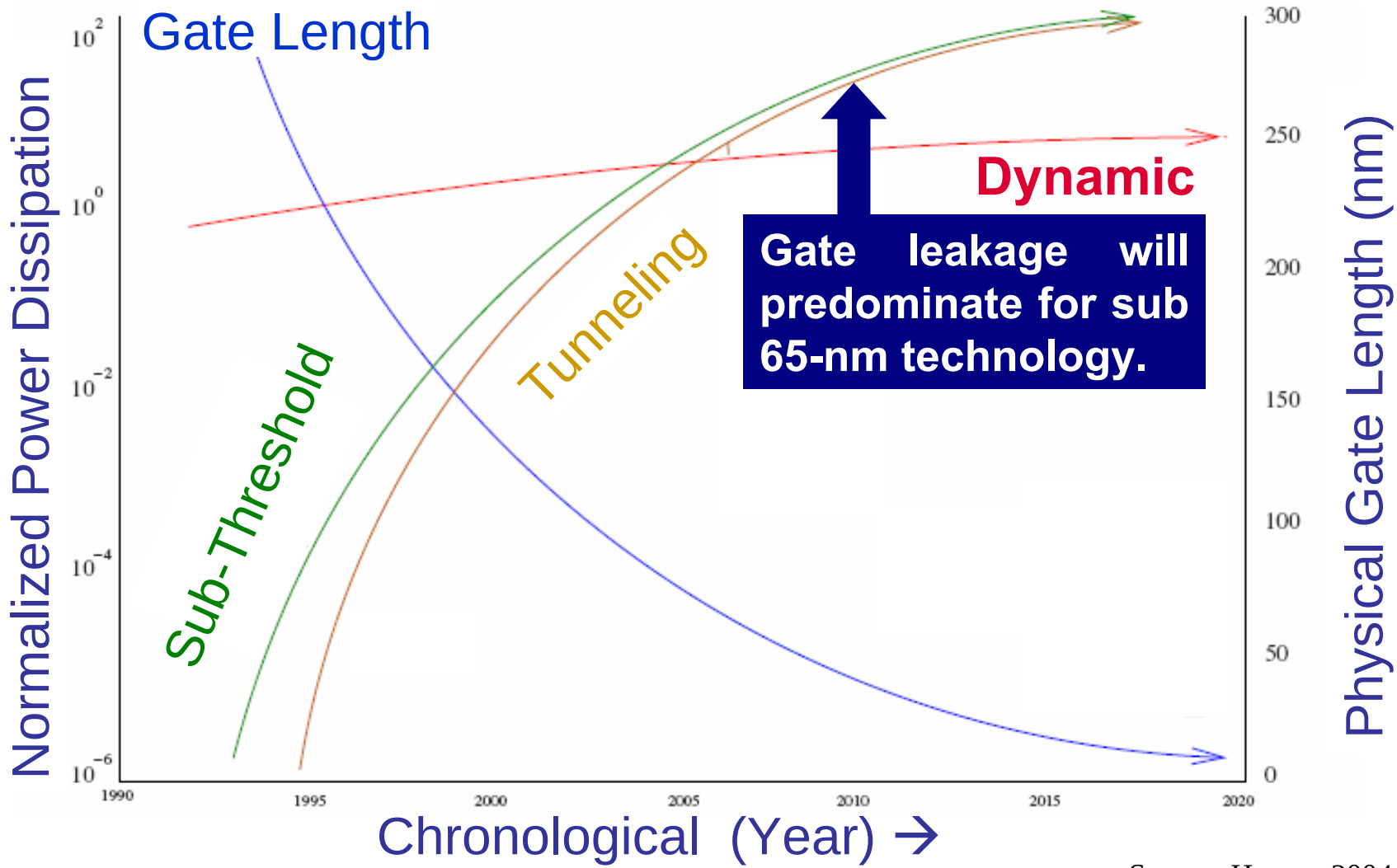
Leakages in Nanoscale CMOS

I_1 : reverse bias pn junction (both ON & OFF)
I_2 : subthreshold leakage (OFF)
I_3 :oxide tunneling current (both ON & OFF)
I_4 : gate current due to hot carrier injection (both ON & OFF)
I_5 : gate induced drain leakage (OFF)
I_6 : channel punch through current (OFF)



Source: Roy Proceedings of IEEE Feb2003

Power Dissipation : Redistribution

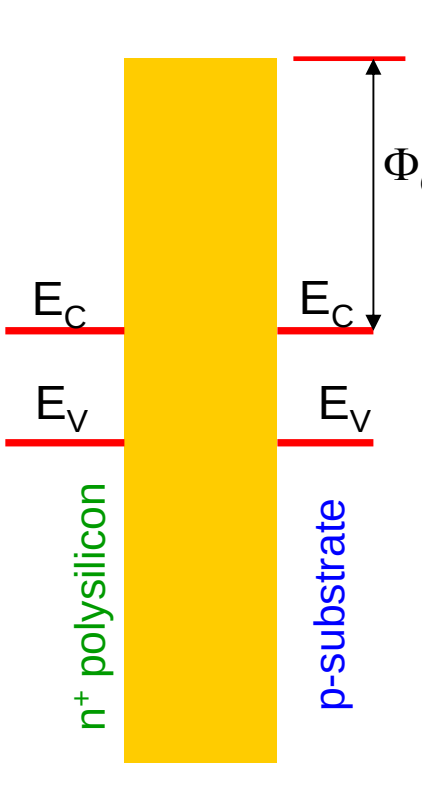


Source: Hansen 2004

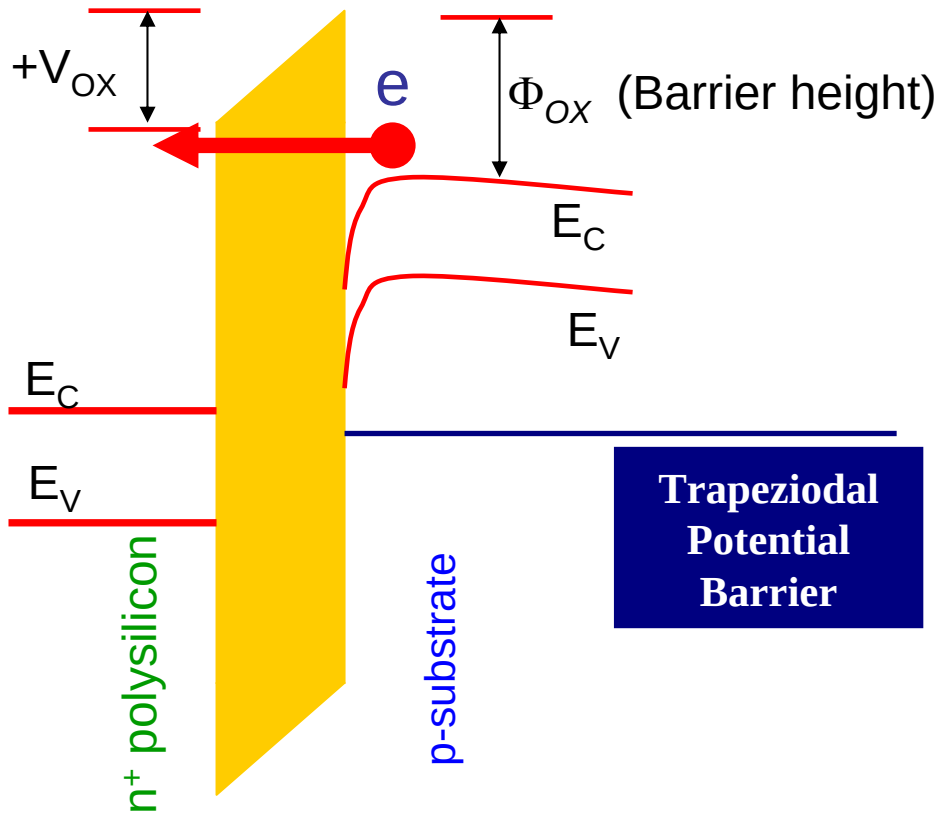
Scaling Trends and Effects : Summary

- Scaling improves
  - Transistor Density of chip
  - Functionality on a chip
  - Speed and frequency of operation
  - Higher performance
- Scaling and power dissipation
  - Active power remains almost constant
  - Components of leakage power increase in number and in magnitude.
  - Gate leakage (tunneling) predominates for sub 65-nm technology.

Energy-Band Diagram Showing Tunneling (Direct Tunneling Occurs when: $V_{OX} < \Phi_{OX}$)



Flat-band Condition

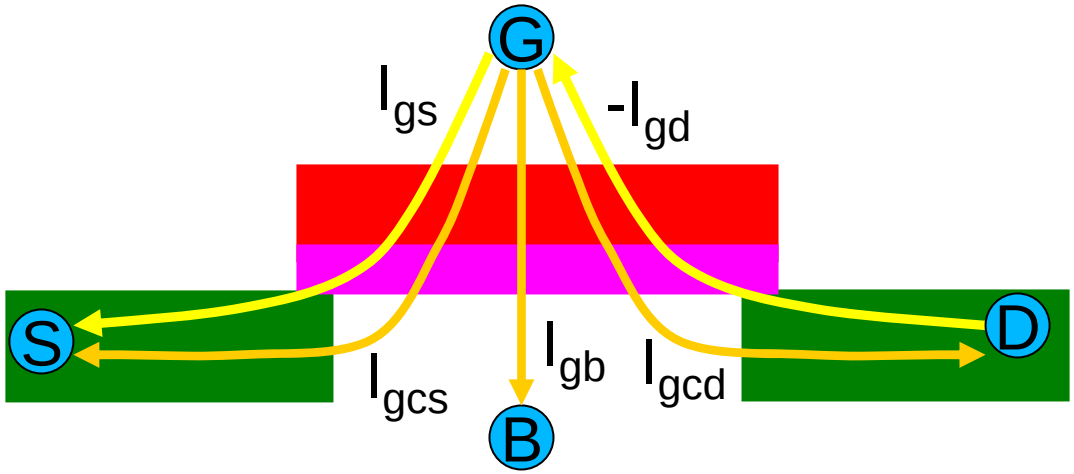


Direct Tunneling for positive bias

NOTE: For short channel MOS FN tunneling is negligible.

Source: AgarwalIEEPDTMay2005

Gate Leakage Components

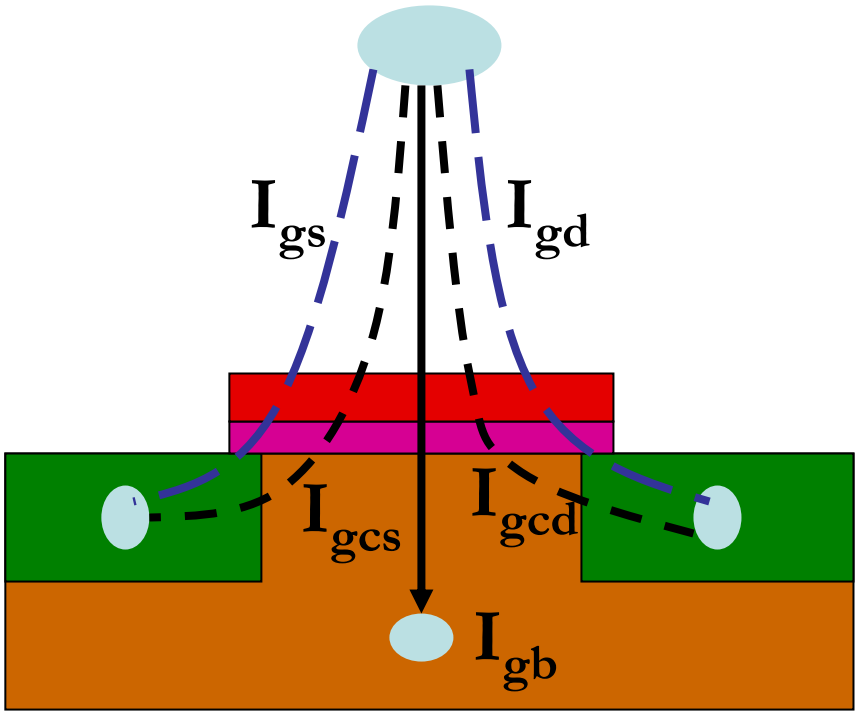


Gate oxide tunneling current components in BSIM4.4.0 model.

- I_{gs}, I_{gd}: Components due to the overlap of gate and diffusions
- I_{gcs}, I_{gcd}: Components due to tunneling from the gate to the diffusions via the channel and
- I_{gb}: Component due to tunneling from the gate to the bulk via the channel.

Note: all the currents are with respect to gate.

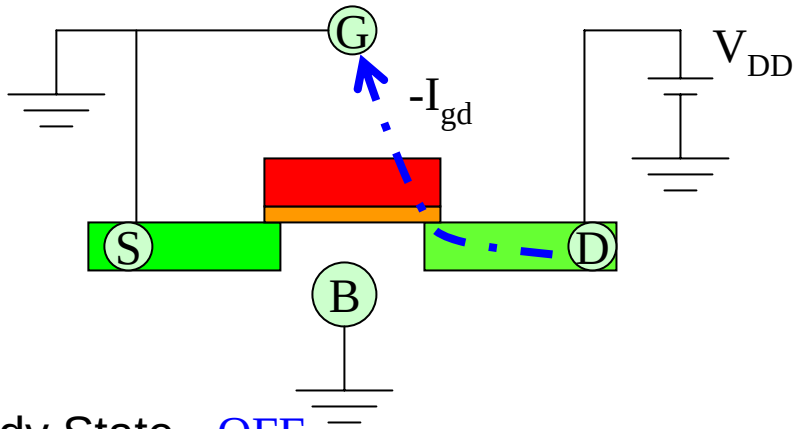
Gate Leakage for a MOS: I_{ox}



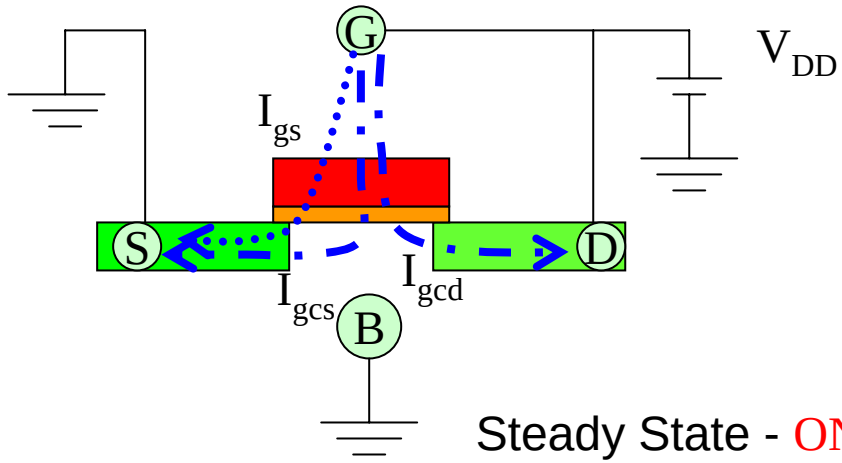
BSIM4 Model

- Calculated by evaluating both the source and drain components
- For a MOS, $I_{ox} = (|I_{gs}| + |I_{gd}| + |I_{gcs}| + |I_{gcd}| + |I_{gb}|)$
- Values of individual components depends on states, ON or OFF

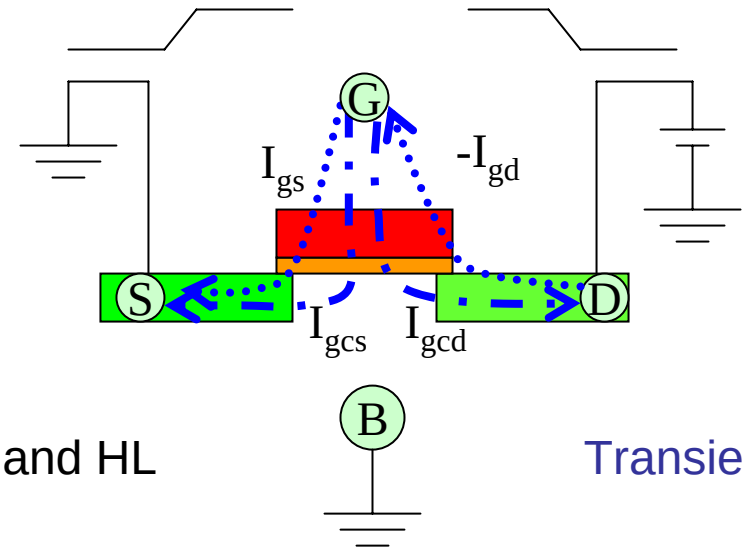
NMOS: Gate Leakage Paths



Steady State - OFF



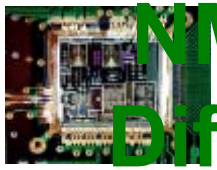
Steady State - ON



LH and HL

Transient State

Negligible Body component



NMOS Gate Leakage Components in Different Phases of a Switching Cycle

Fig. 1

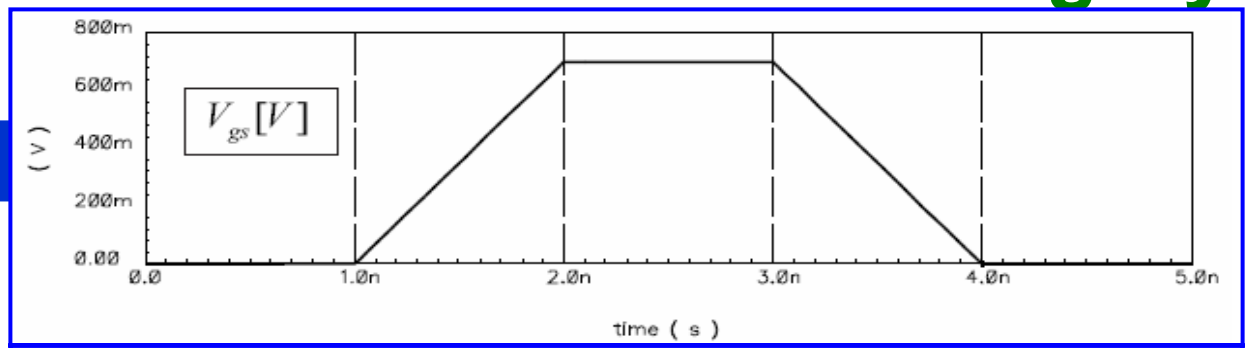


Fig. 2

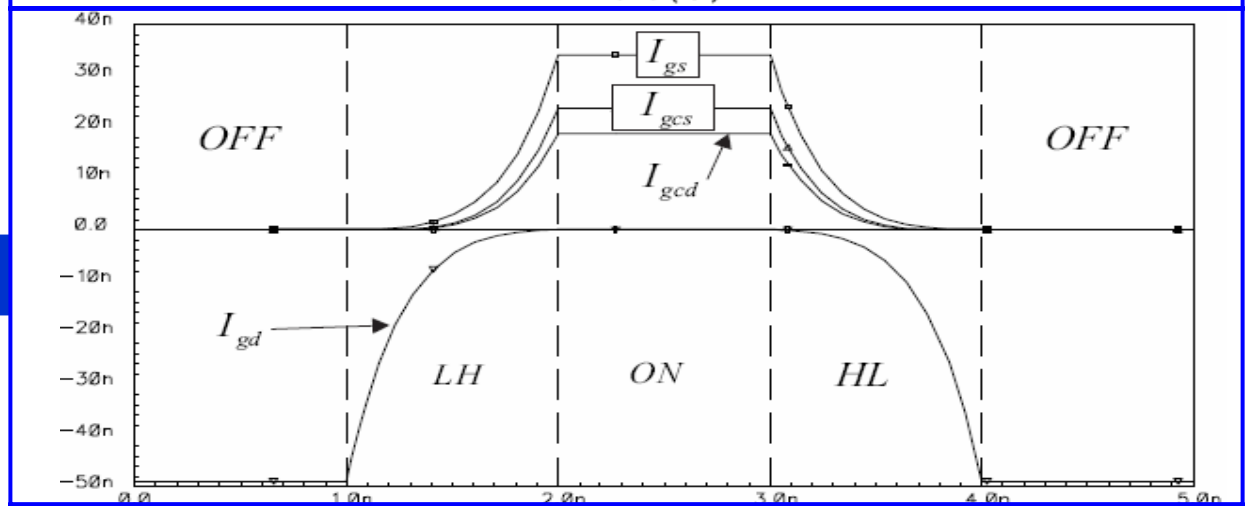
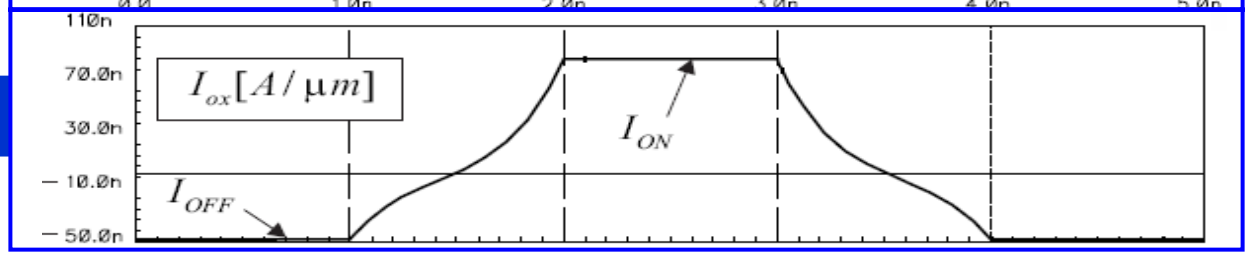
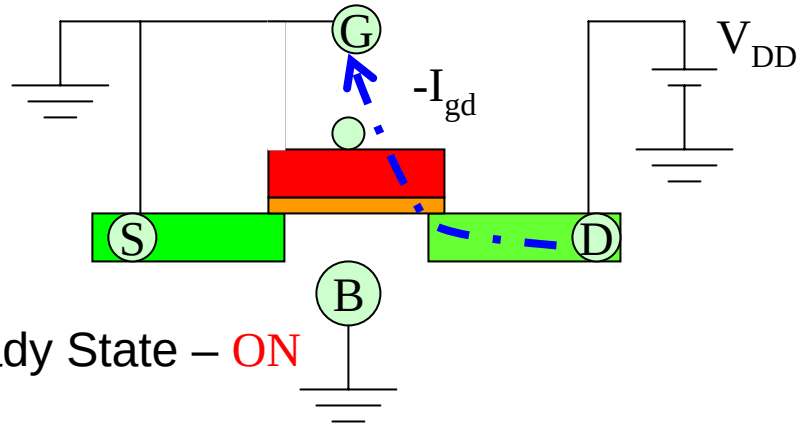


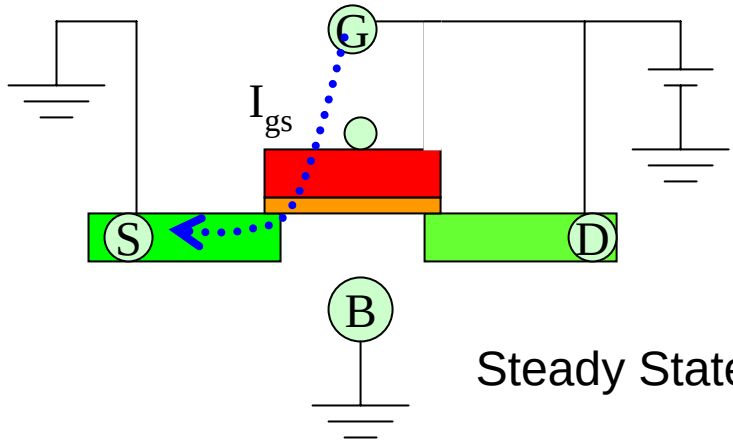
Fig. 3



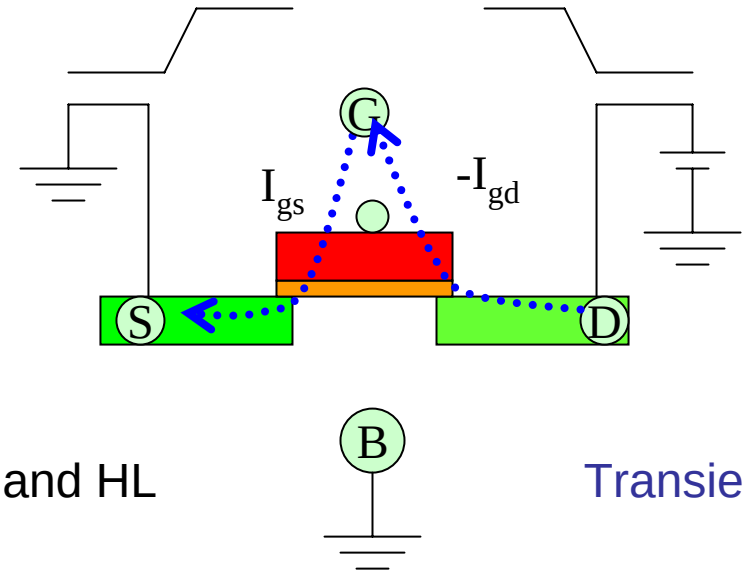
PMOS: Gate Leakage Paths



Steady State – ON



Steady State – OFF



LH and HL

Transient State

Negligible Channel and Body components

PMOS Gate Leakage Components in Different Phases of a Switching Cycle

Fig. 1

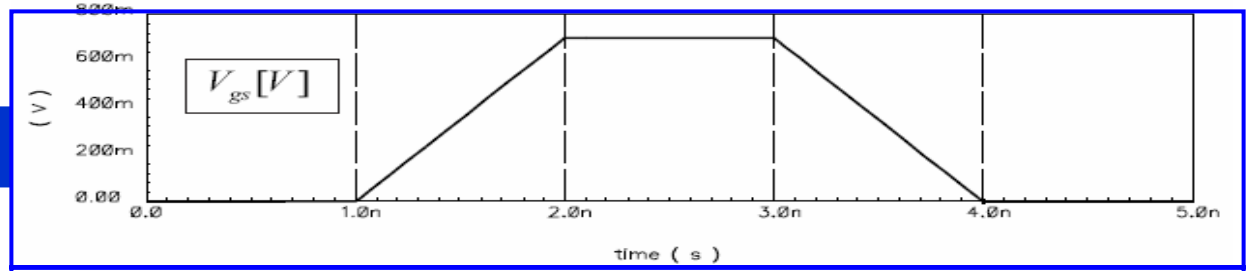


Fig. 2

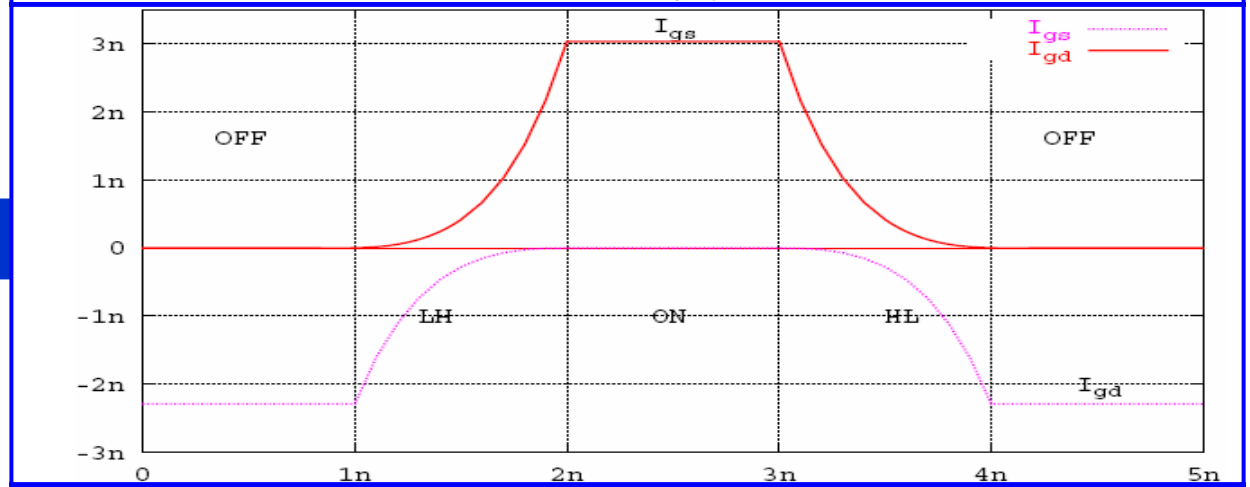
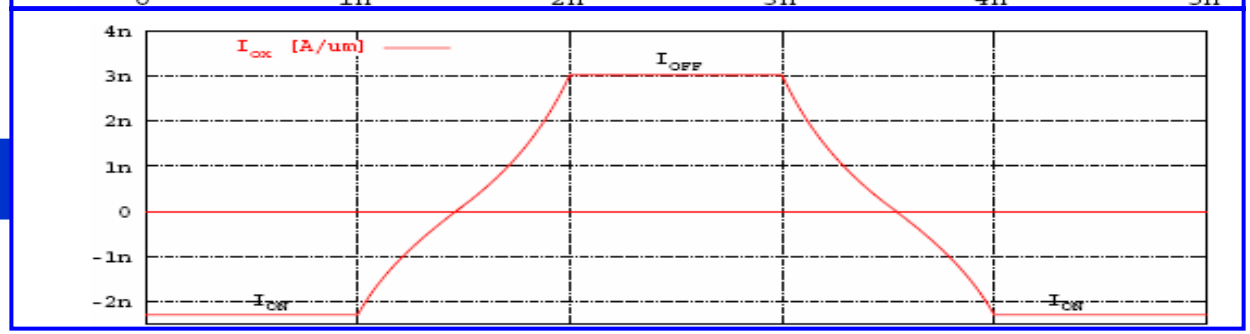


Fig. 3



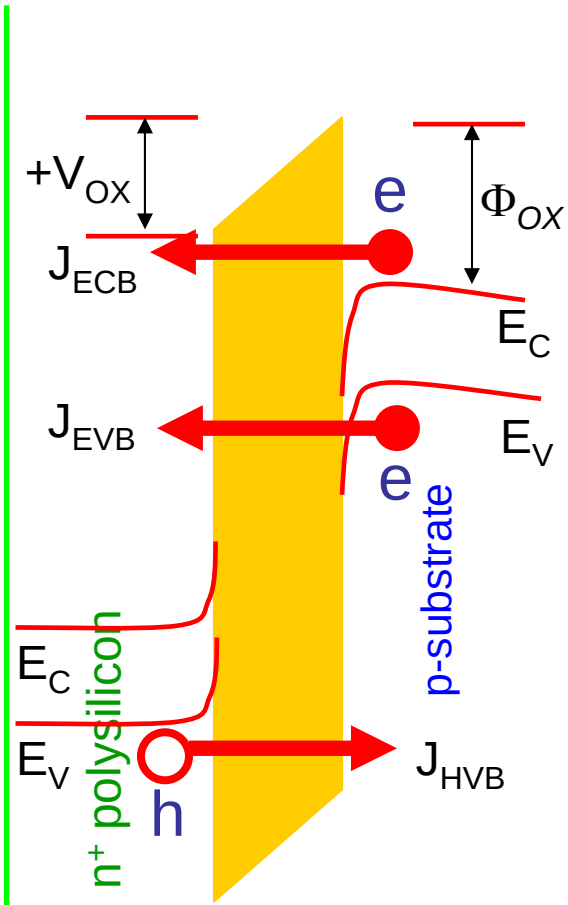
NMOS Vs PMOS: 3 Mechanisms of Tunneling

Three major mechanisms for direct tunneling:
1.electron tunneling from conduction band (ECB)
2.electron tunneling from valence band (EVB)
3.hole tunneling from valance band (HVB)

For NMOS:
- ECB controls gate-to-channel tunneling in inversion
- EVB controls gate-to-body tunneling in depletion-inversion
- ECB controls gate-to-body tunneling in accumulation

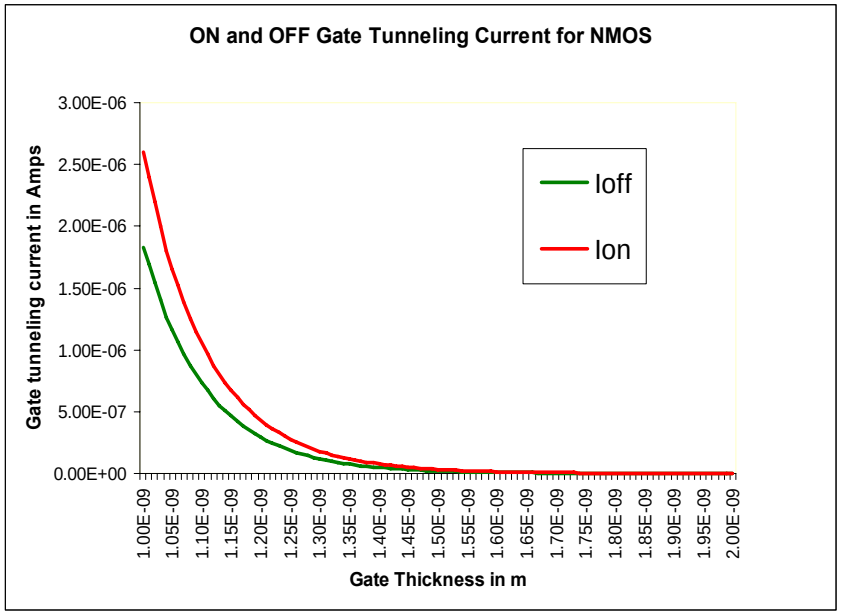
For PMOS:
- HVB controls the gate-to-channel tunneling in inversion
- EVB controls gate-to-body tunneling in depletion-inversion
- ECB controls gate-to-body tunneling in accumulation

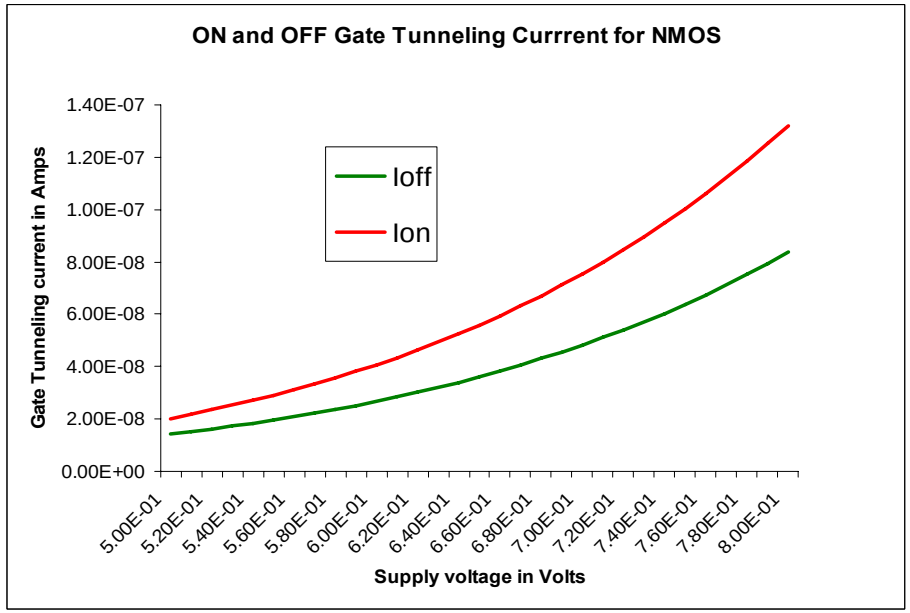
PMOS < NMOS: Φ_{OX} for HVB (4.5 eV) is higher than Φ_{OX} for ECB (3.1 eV), the tunneling current associated with HVB is less than that with ECB.



Source: Roy Proceedings of IEEE Feb2003

Gate Leakage: Effect of Parameter Variation (NMOS)

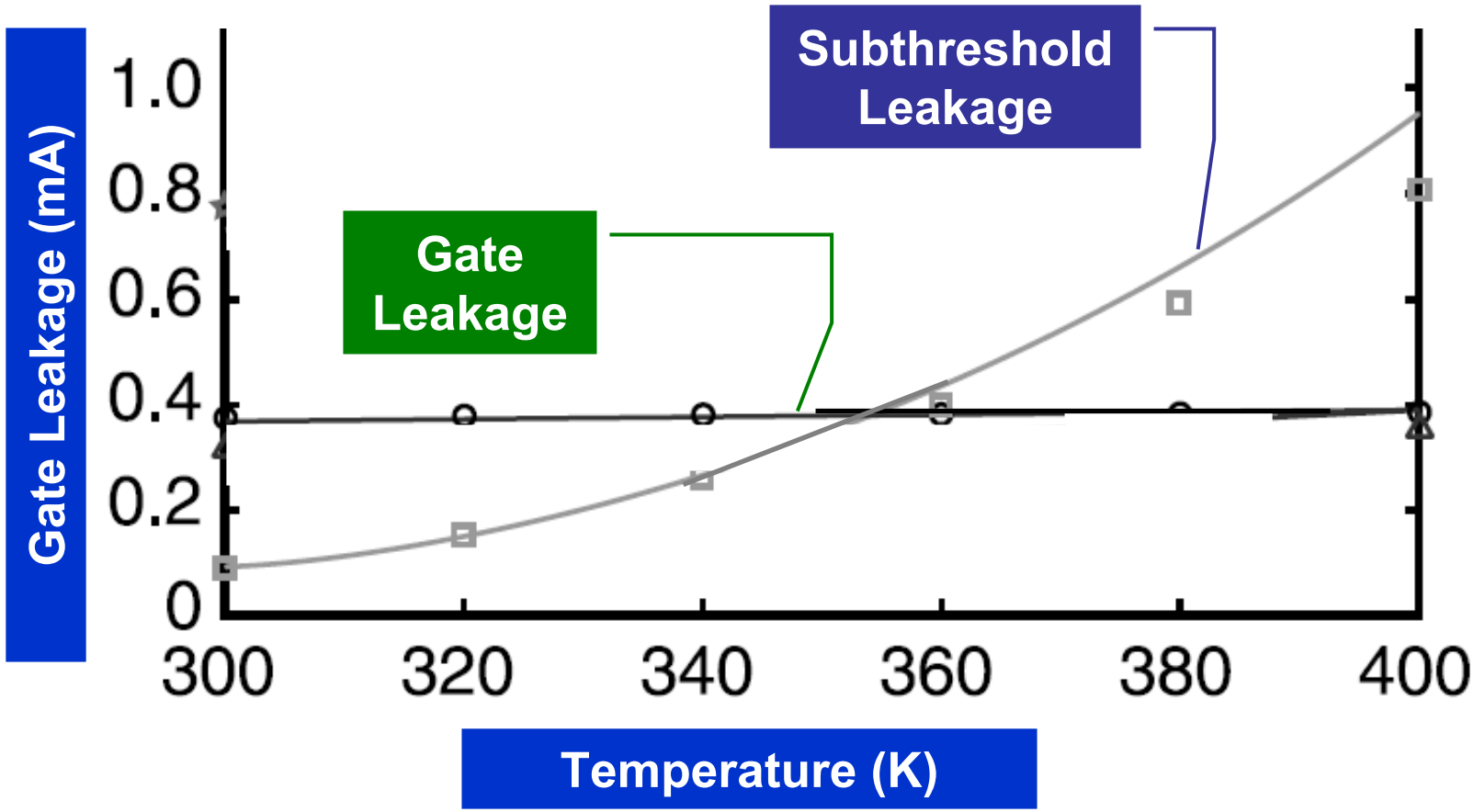




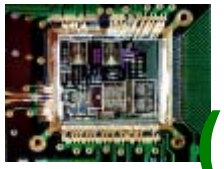
Gate Leakage Vs T_{ox}

Gate Leakage Vs V_{DD}

Gate Leakage: Effect of Parameter Variation (NMOS)

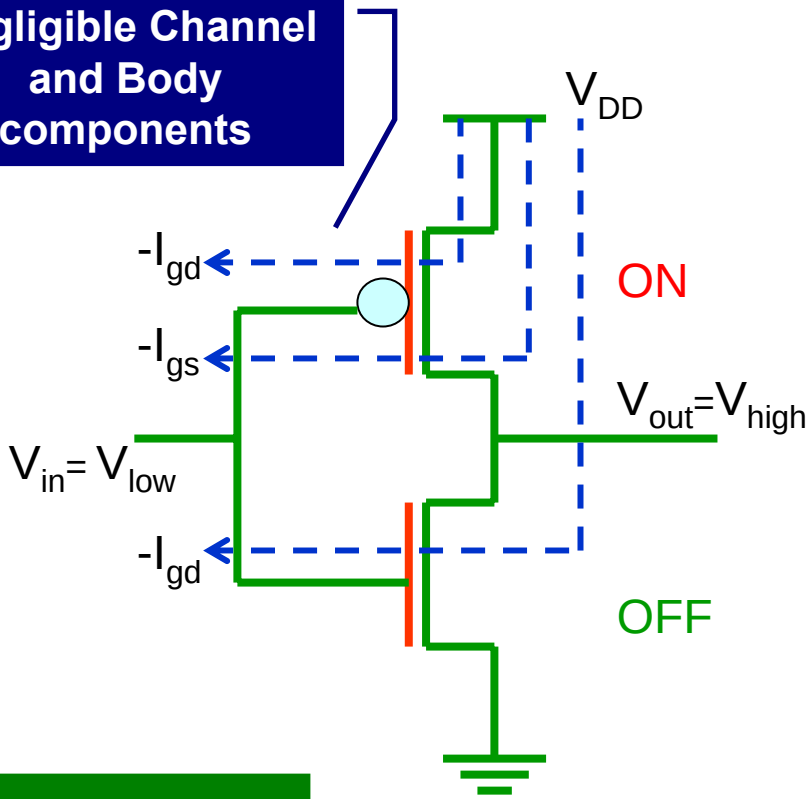


Source: Agarwal IEE Proc. CDT May 2005

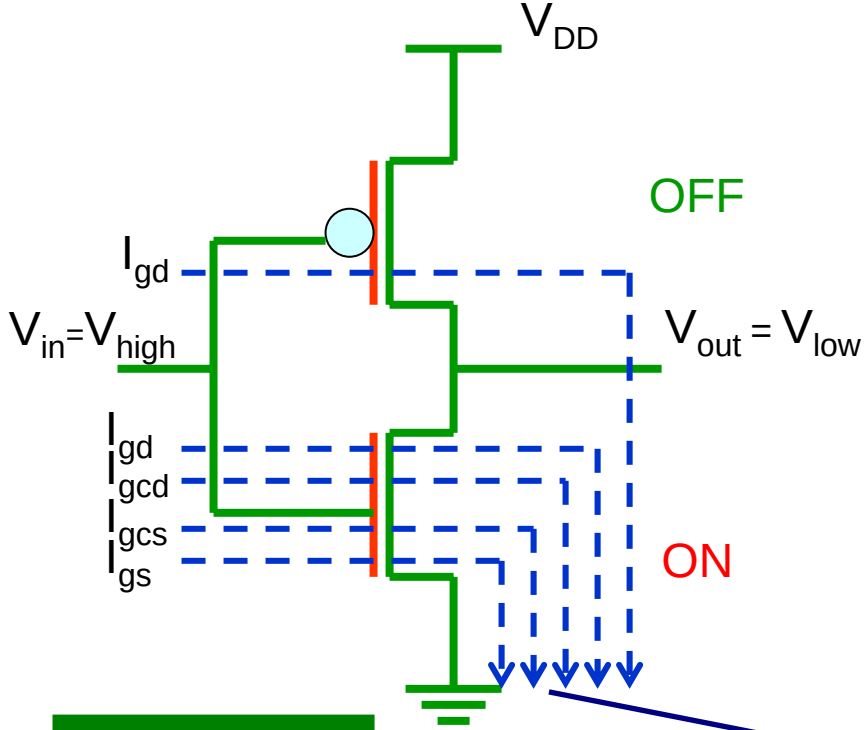


Inverter: Gate Leakage Paths (Putting NMOS and PMOS together)

Negligible Channel and Body components



Low Input

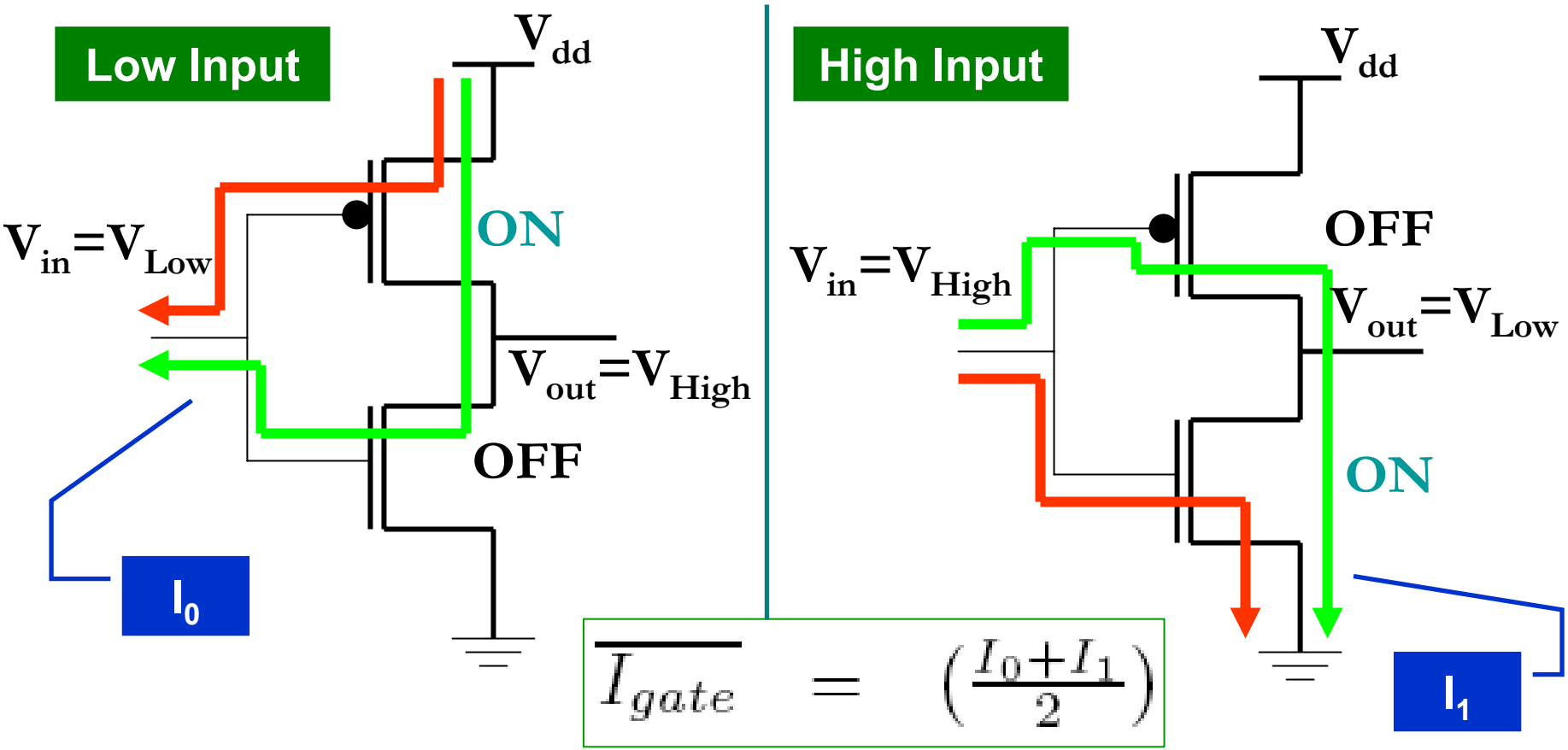


High Input

Negligible Body component

Inverter: Average Gate Leakage

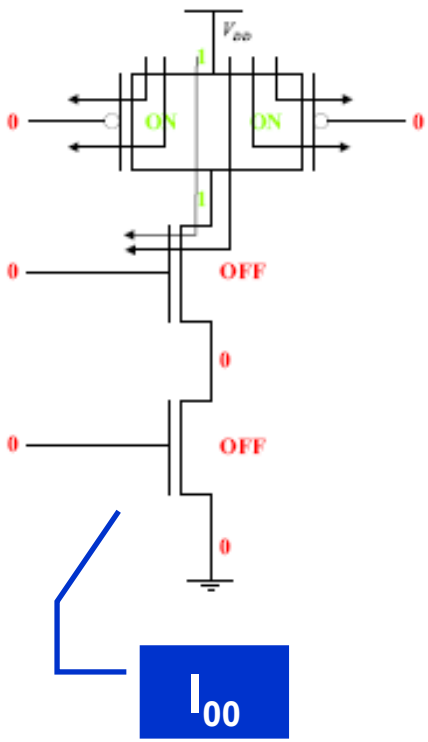
- **Low Input :** Input supply feeds the tunneling current.
- **High Input :** Gate supply feeds the tunneling current.



$$\overline{I_{gate}} = \left(\frac{I_0+I_1}{2}\right)$$

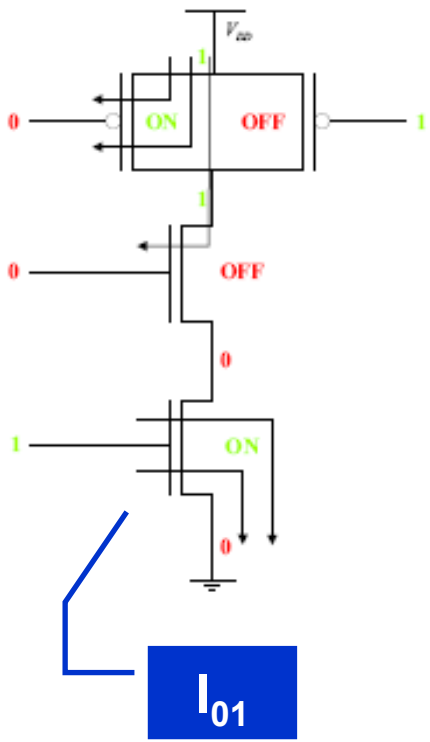
Gate Leakage in 2-input NAND

input 00



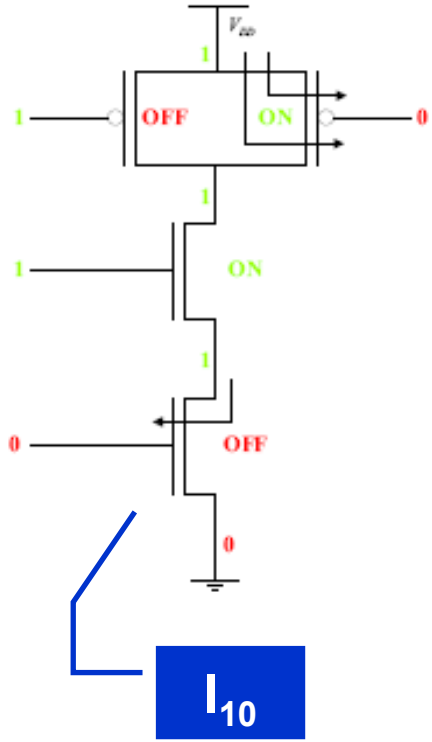
I_{00}

input 01



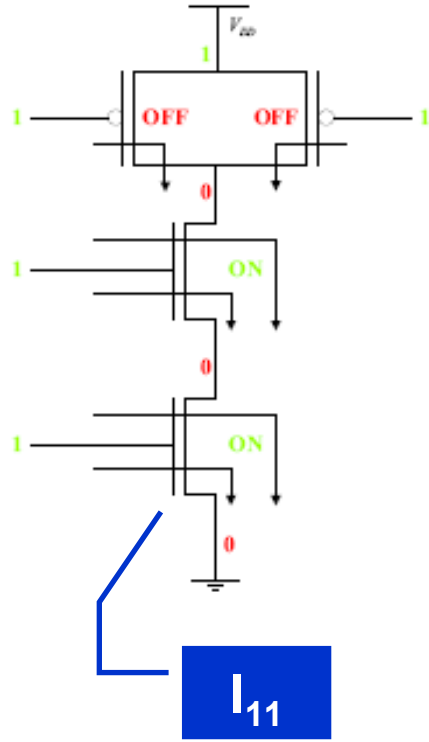
I_{01}

input 10



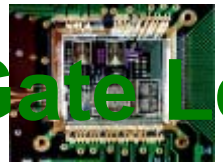
I_{10}

input 11



I_{11}

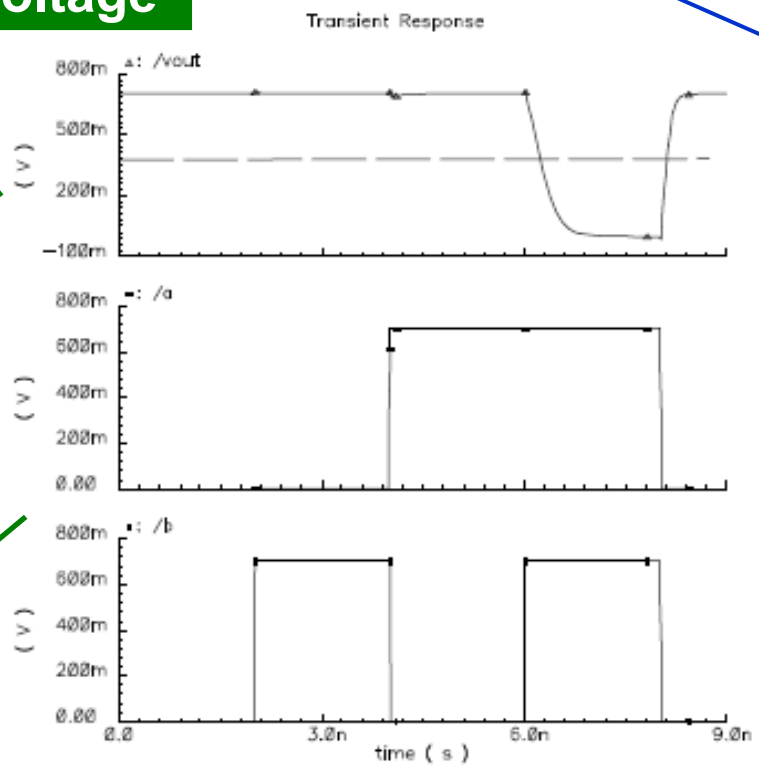
$$\overline{I_{gate}} = \left(\frac{I_{00}+I_{01}+I_{10}+I_{11}}{4}\right)$$



Gate Leakage in 2-input NAND: Transient Study

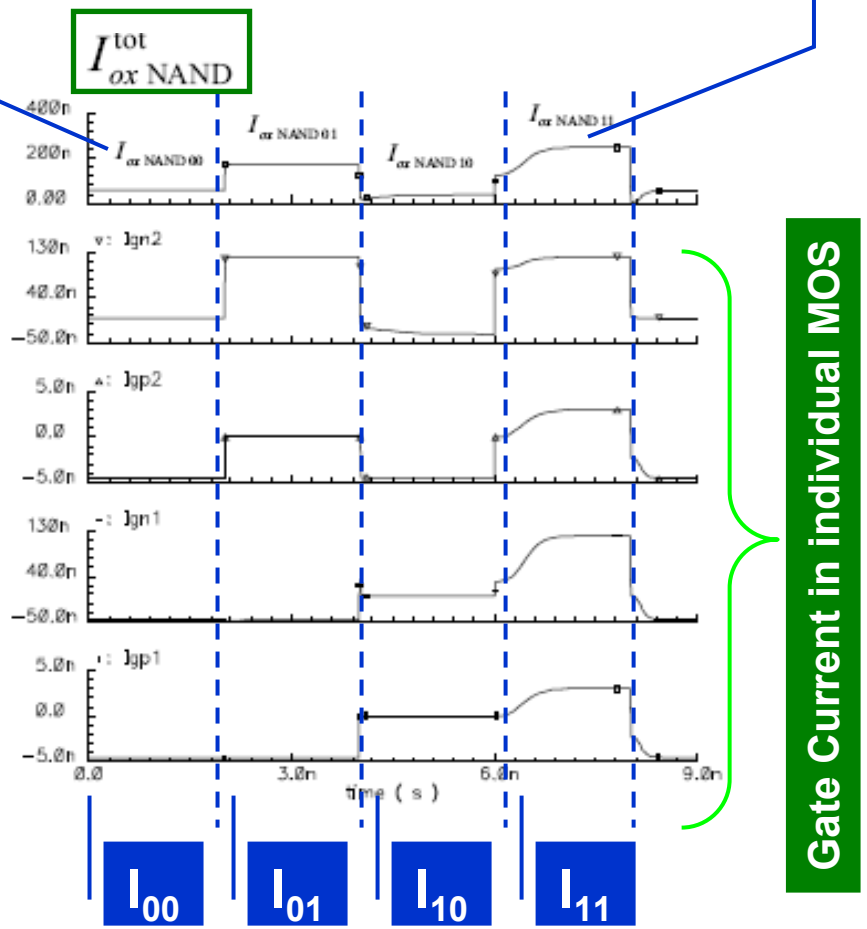
Output Voltage

Input Voltages

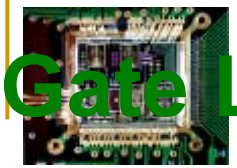


Best Case

Worst Case



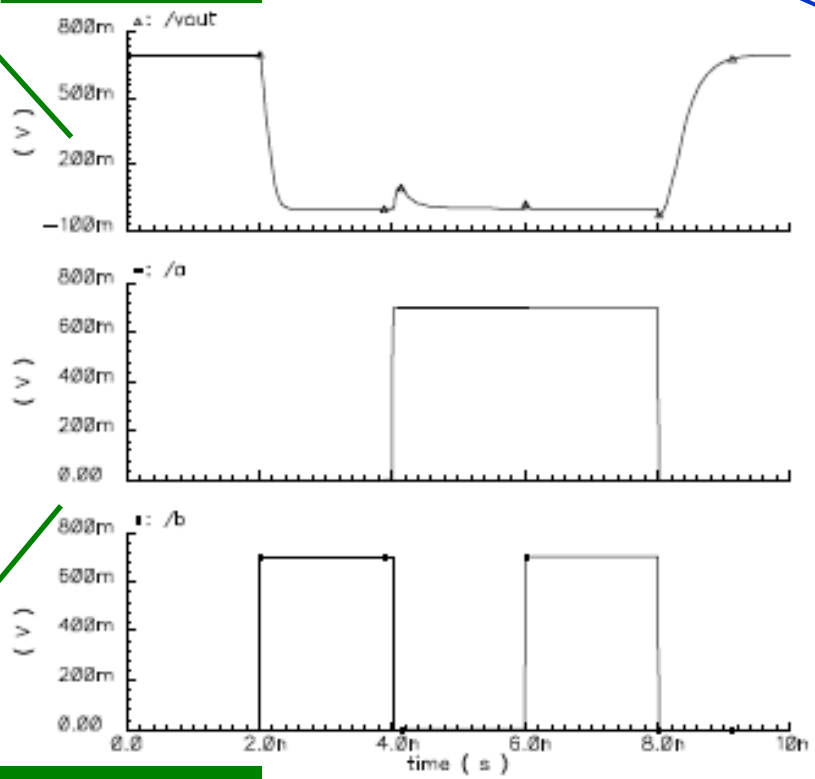
Gate Current in individual MOS



Gate Leakage in 2-input NOR: Transient Study

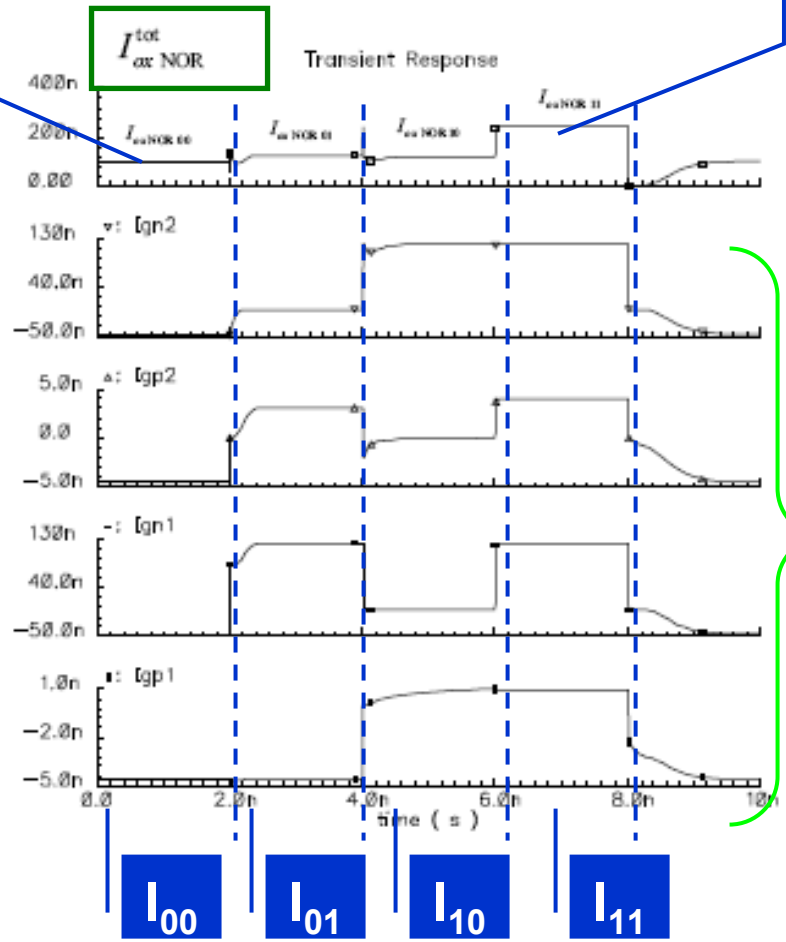
Output Voltage

Input Voltages



Best Case

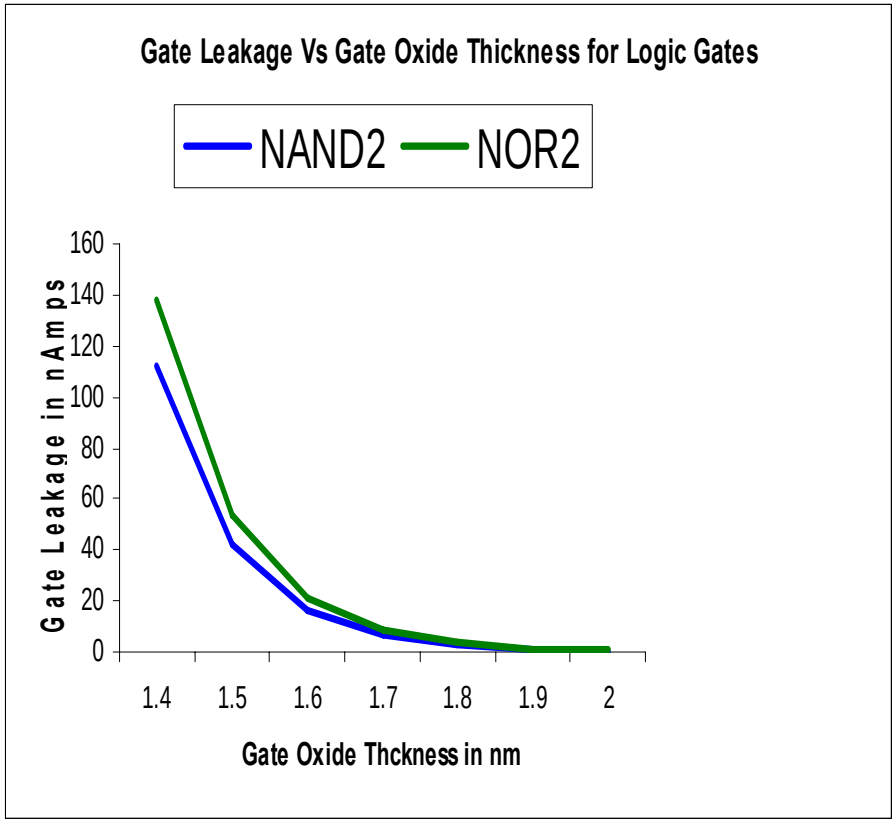
Worst Case



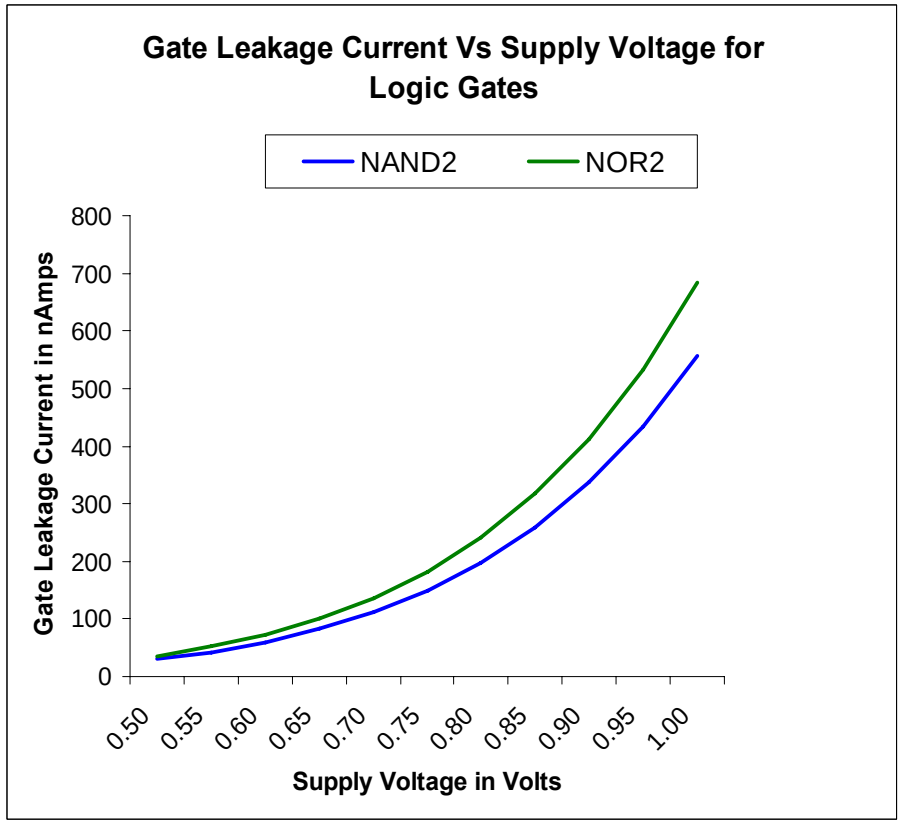
Gate Current in individual MOS

I_{00} I_{01} I_{10} I_{11}

Gate Leakage in 2-input Logic Gates (Average Current's Dependence on Parameters)



Gate Leakage Vs T_{ox}



Gate Leakage Vs V_{DD}

Gate Leakage Estimation

- What we have observed?
  - Gate leakage is input state dependent
  - Gate leakage is dependent on position of ON/OFF transistors
  - Gate leakage is sensitive to process variation

- Gate leakage estimation methods for logic level description of the circuit:
  - Pattern dependent estimation (R. M. Rao ISLPED 2003)
  - Pattern independent probabilistic estimation (R. M. Rao ISLPED 2003)

Estimation: Pattern Dependent

- For an given input vector switch-level simulation is performed
- State of internal nodes is determined for the input vector
- Unit width gate leakage of a device is determined for different states
- The total gate leakage is computed by scaling the width of each device by unit-width leakage in that state and adding the individual leakages:

$$I_{ox} = \sum_{MOS} I_{ox,MOS}(s(i)) * W_{MOS}$$

Source: R. M. Rao ISLPED2003

Estimation: Pattern Independent

- Probability analysis in conjunction with state-dependent gate leakage estimation is used.
- The average gate leakage of the circuit is the probabilistic mean of the gate leakage of the circuit:

$$I_{ox,avg} = E(\Sigma_{MOS}\, I_{ox,MOS}(s(i)) * W_{MOS})$$

$$= \Sigma_{MOS}\, W_{MOS} * (\Sigma_j\, I_{ox,MOS}(s(j)) * P(j))$$

where P(j) is the probability of occurrence of state j.

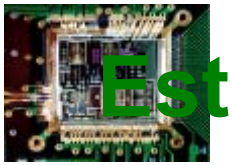
Source: R. M. Rao ISLPED2003

Estimation: Heuristic and Look-up Tables

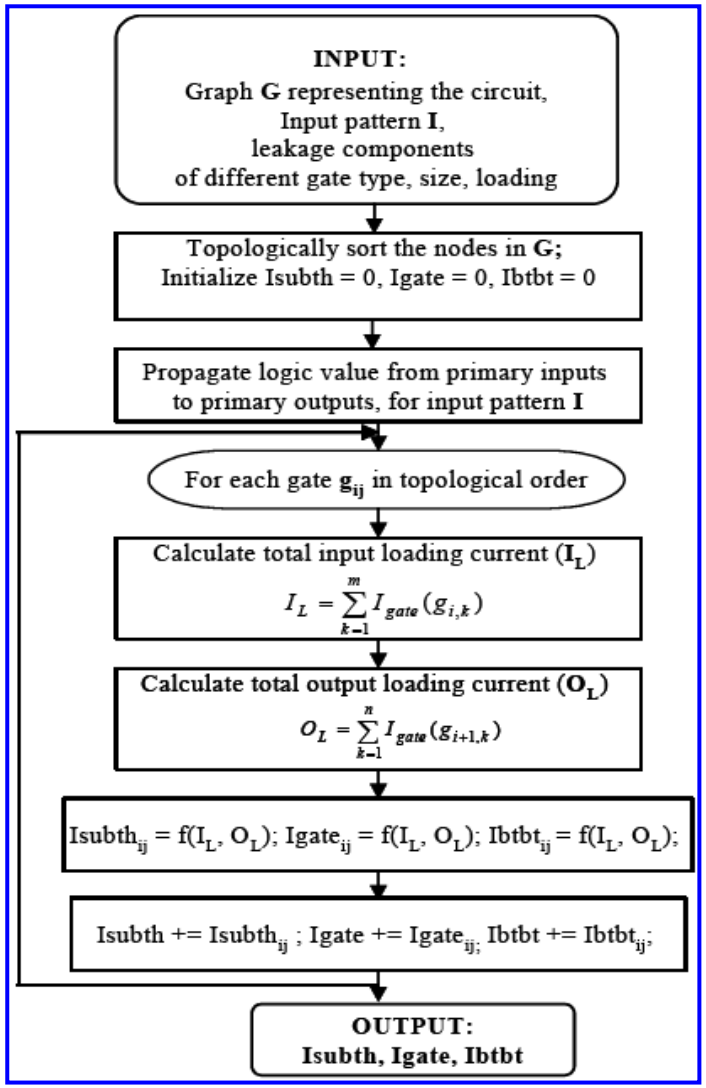
- Interaction between gate leakage and subthreshold leakage are used to develop heuristic based estimation techniques for state-dependent total leakage current.

- Heuristics based on lookup tables are available to quickly estimate the state-dependent total leakage current for arbitrary circuit topologies.

Source: Lee ISQED2003, TVLSI2003



Estimation: Loading Effect on leakage



1. Represent circuit as graph: vertex → logic gate and edge → net
2. Sort vertices in topological order and initialize leakage values to zero
3. Propagate input vector and assign a logic state to each gate
4. Calculate total input and output loading current due to gate leakage
5. Calculate the leakage of the individual logic gates
6. Compute the leakage of the total circuit by adding leakage of individual gates.

Source: Mukhopadhyay DATE2005 and TCAD 2005 (to appear)

Techniques for Gate Leakage Reduction

Research in Gate leakage is catching up and have not matured like that of dynamic or subthreshold power. Few methods:

- Dual T_{OX} (Sultania DAC 2004, Sirisantana IEEE DTC Jan-Feb 2004)
- Dual K (Mukherjee ICCD 2005)
- Pin and Transistor Reordering (Sultania ICCD 2004, Lee DAC 2003)

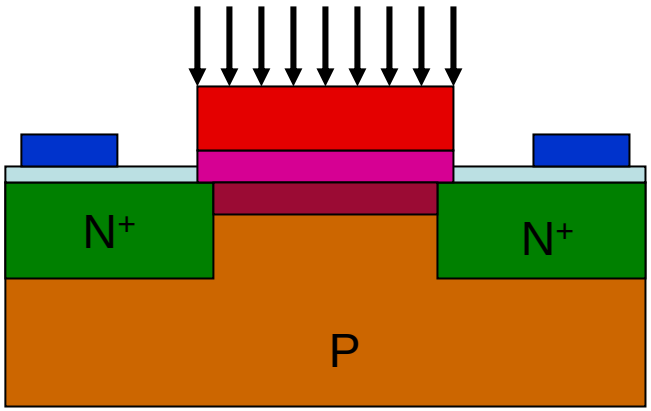
Dual T_{OX} Technique: Basis

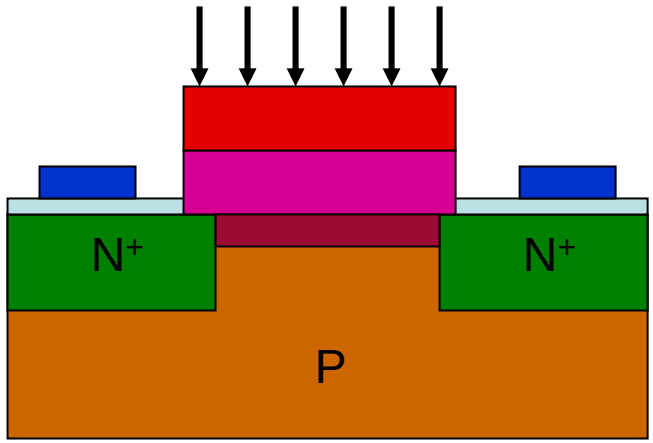
- Gate oxide tunneling current I_{oxide} (k is a experimentally derived factors):

$$I_{oxide} \alpha (V_{dd} / T_{gate})^2 \exp (- k T_{gate}/V_{dd})$$

- Options for reduction of tunneling current:
  - Decreasing of supply voltage V_{dd} (*will play its role*)
  - Increasing gate SiO_2 thickness T_{oxide}

Dual T_{OX} Technique: Basis





Low T_{gate} → Larger I_{gate} , Smaller delay

High T_{gate} → Smaller I_{gate} , Larger delay

Dual T_{OX} Technique: Approach

- Our approach – scale channel length (L) as well as T_{ox}; T_{ox} is almost linearly scaled with L_{eff}

$$\text{Aspect Ratio} = \frac{L_{eff}}{T_{ox,eff}} = \text{constant}$$

Advantages:

- Reduces DIBL effect
- Constant Input Gate Capacitance for a given W_{eff},

$$C_{micron} = \frac{\epsilon_{ox} L_{eff}}{T_{ox,eff}} = \text{constant}$$

Dual T_{OX} Technique: Algorithm

Assign all transistors to $T_{ox\text{-}Hi}$

Critical path using STA

Compute cost = $\frac{\Delta Delay}{\Delta Leakage}$ (always negative)

Choose transistor with most negative cost

If no transistor chosen or Delay constraint met, then EXIT

Assign $T_{ox\text{-}Lo}$ to chosen transistor and update

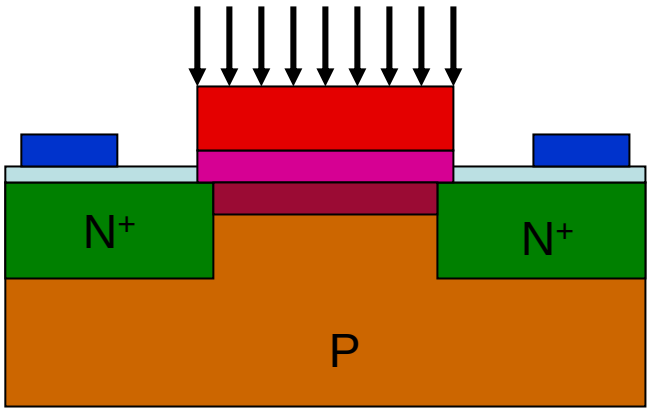
Source: Sultania DAC 2004

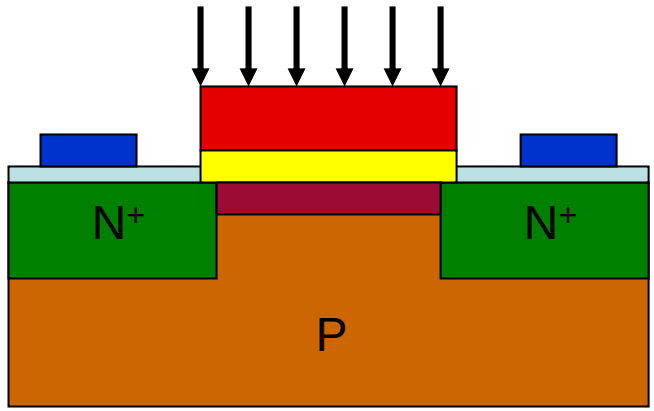
Dual T_{OX} Technique: Results

- Iterative algorithm that
  - Generates delay/leakage tradeoffs
  - Meets delay constraint
- For same delay an average leakage reduction of 83% compared to the case where all transistors are set to $T_{ox\text{-}Lo}$.
- Minor changes in design rules and an extra fabrication step is required, extra mask required.

Source: Sultania DAC 2004

Dual K Technique: Basis

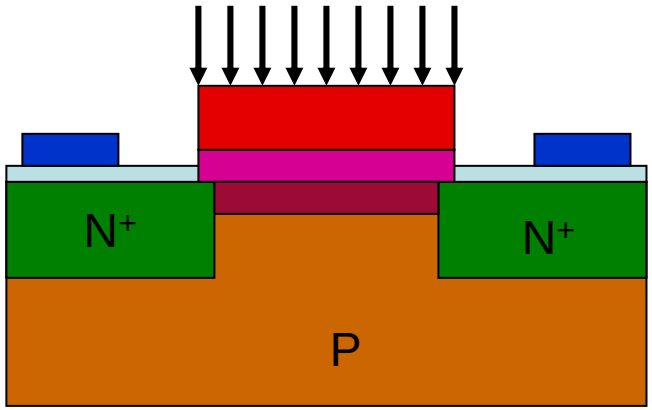


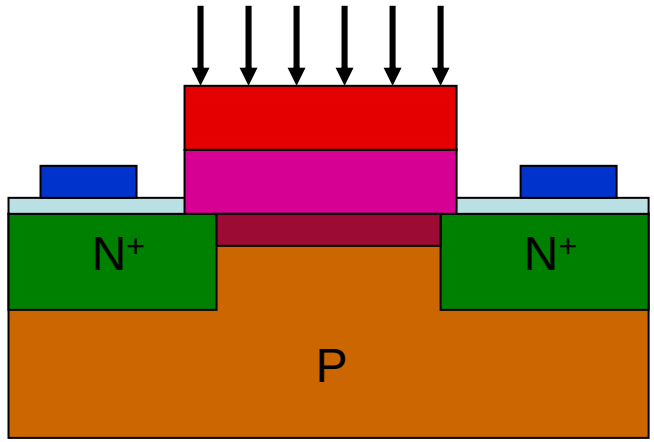


Low K_{gate} → Larger I_{gate} , Smaller delay

High K_{gate} → Smaller I_{gate} , Larger delay

Dual K Technique: Basis

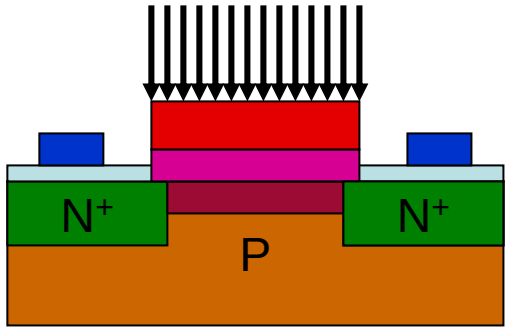




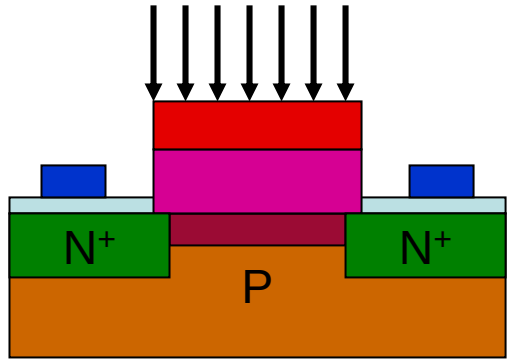
Low T_{gate} → Larger I_{gate}, Smaller delay

High T_{gate} → Smaller I_{gate}, Larger delay

Dual K Technique: Basis (Four Combinations of K_{gate} & T_{gate})

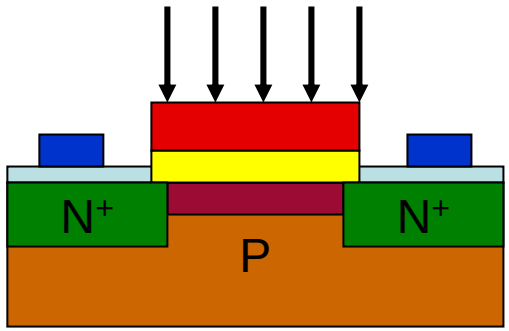


(1) K_1T_1

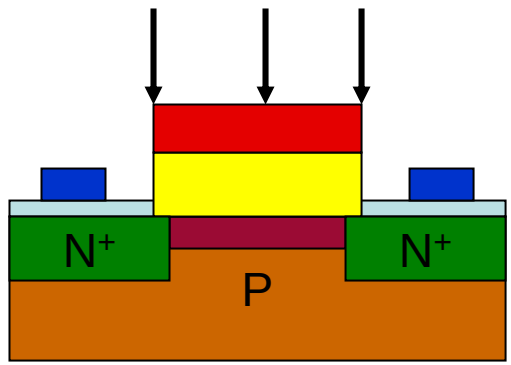


(2) K_1T_2

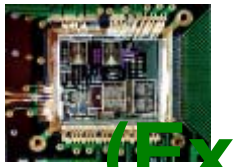
Tunneling Current ↓
Delay ↑



(3) K_2T_1

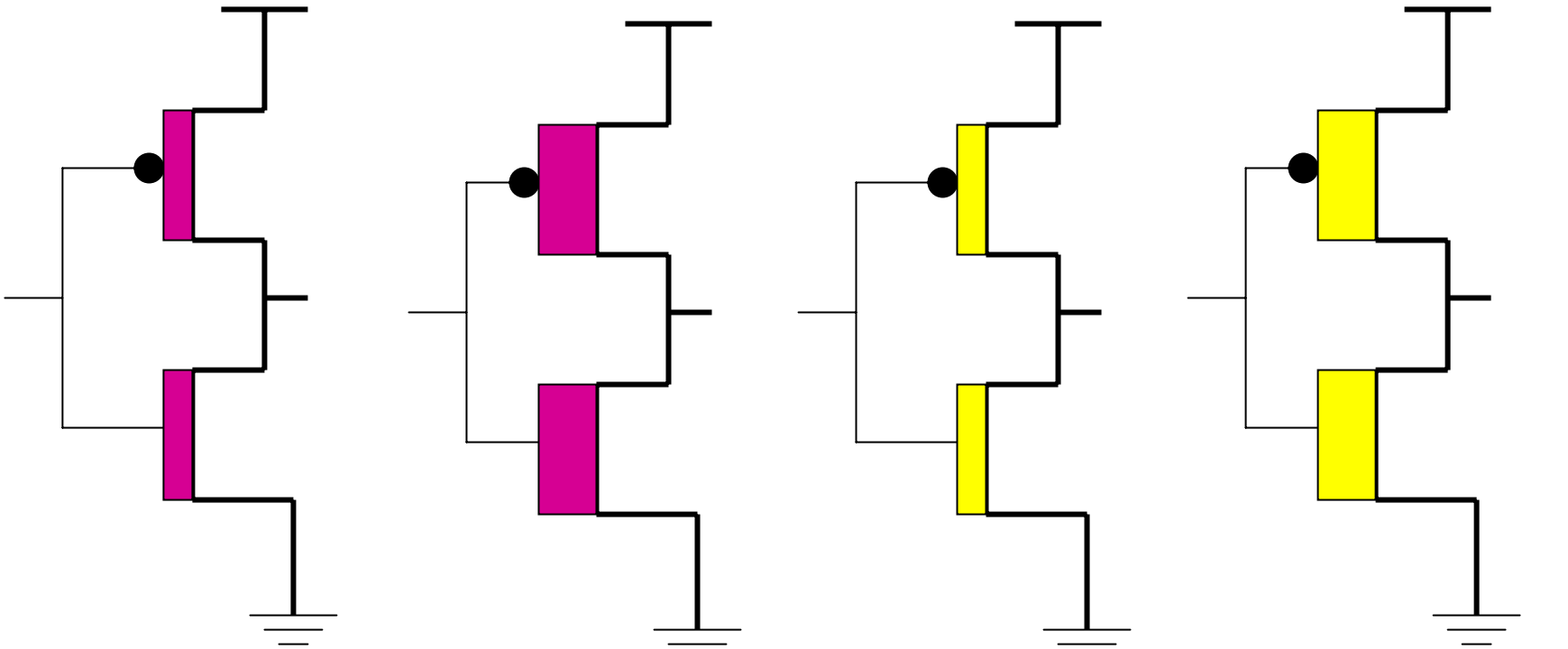


(4) K_2T_2



Dual K Technique: Basis (Example: Four Types of Logic Gates)

Assumption: all transistors of a logic gate are of same K_{gate} and equal T_{gate}.



(1) K_1T_1 (2) K_1T_2 (3) K_2T_1 (4) K_2T_2

Dual K Technique: Basis

Use of multiple dielectrics (denoted as K_{gate}) of multiple thickness (denoted as T_{gate}) will reduce the gate tunneling current significantly while maintaining the performance.

Source: Mukherjee ICCD 2005

Dual K Technique: New Dielectrics

- Silicon Oxynitride (SiO_xN_y) (K=5.7 for SiON)
- Silicon Nitride (Si_3N_4) (K=7)
- Oxides of :
  - Aluminum (Al), Titanium (Ti), Zirconium (Zr), Hafnium (Hf), Lanthanum (La), Yttrium (Y), Praseodymium (Pr),
  - their mixed oxides with SiO_2 and Al_2O_3

- **NOTE**: I_{gate} is still dependent on T_{gate} irrespective of dielectric material.

Dual K Technique: Strategy

- **Observation**: Tunneling current of logic gates increases and propagation delay decreases in the order K_2T_2, K_2T_1, K_1T_2, and K_1T_1 (where, $K_1 < K_2$ and $T_1 < T_2$).

- **Strategy**: Assign a higher order K and T to a logic gate under consideration
  - To reduce tunneling current
  - Provided increase in path-delay does not violate the target delay

Source: Mukherjee ICCD 2005

Dual K Technique: Algorithm

Step 1: Represent the network as a directed acyclic graph G(V, E).

Step 2: Initialize each vertex $v \in$ G(V, E) with the values of tunneling current and delay for K_1T_1 assignment.

Step 3: Find the set of all paths $P\{\Pi_{in}\}$ for all vertex in the set of primary inputs (Π_{in}), leading to the primary outputs Π_{out} .

Step 4: Compute the delay D_P for each path $p \in P\{\Pi_{in}\}$.

Source: Mukherjee ICCD 2005

Dual K Technique: Algorithm

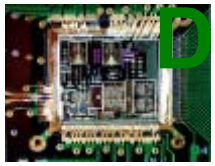
Step 5: Find the critical path delay D_{CP} for K_1T_1 assignment.

Step 6: Mark the critical path(s) P_{CP}, where P_{CP} is subset $P\{\Pi_{in}\}$.

Step 7: Assign target delay $D_T = D_{CP}$.

Step 8: Traverse each node in the network and attempt to assign K-T in the order K_2T_2, K_2T_1, K_1T_2, and K_1T_1 to reduce tunneling while maintaining performance.

Source: Mukherjee ICCD 2005

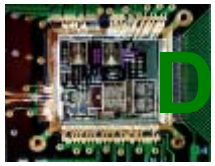


Dual K Technique: Characterization (How to Model High-K?)

- The effect of varying dielectric material was modeled by calculating an equivalent oxide thickness (T^*_{ox}) according to the formula:

$$\mathbf{T^*_{ox} = (K_{gate} / K_{ox})\, T_{gate}}$$

- Here, K_{gate} is the dielectric constant of the gate dielectric material other than SiO_2, (of thickness T_{gate}), while K_{ox} is the dielectric constant of SiO_2.



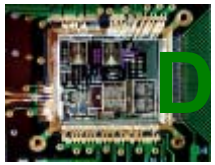
Dual K Technique: Characterization

- The effect of varying oxide thickness T_{ox} was incorporated by varying TOXE in SPICE model.

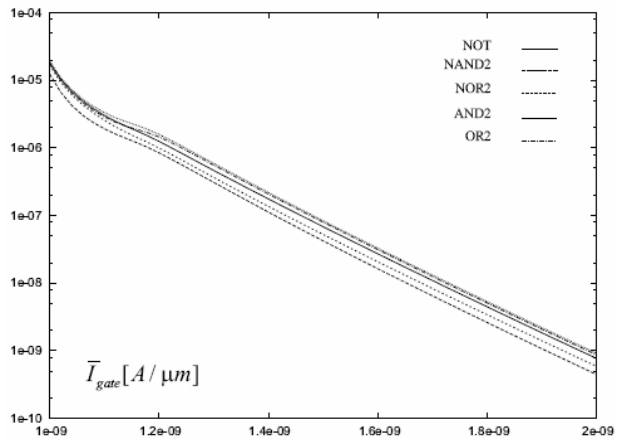
- Length of the device is proportionately changed to minimize the impact of higher dielectric thickness on the device performance :

$$L^* = (T^*_{ox} / T_{ox})\, L$$

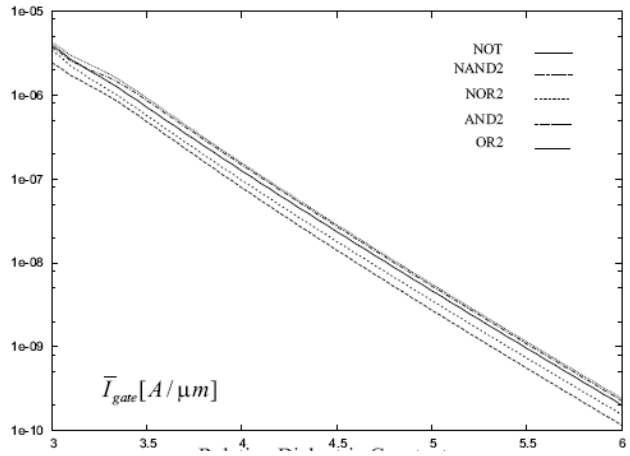
- Length and width of the transistors are chosen to maintain (W:L) ratio of (4:1) for NMOS and (8:1) for PMOS.



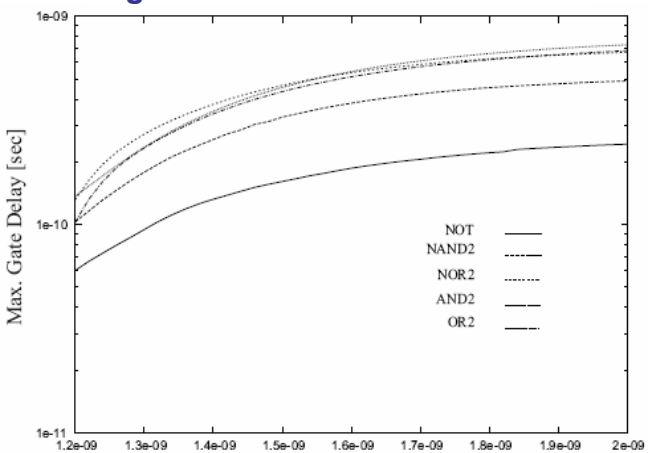
Dual K Technique: Characterization



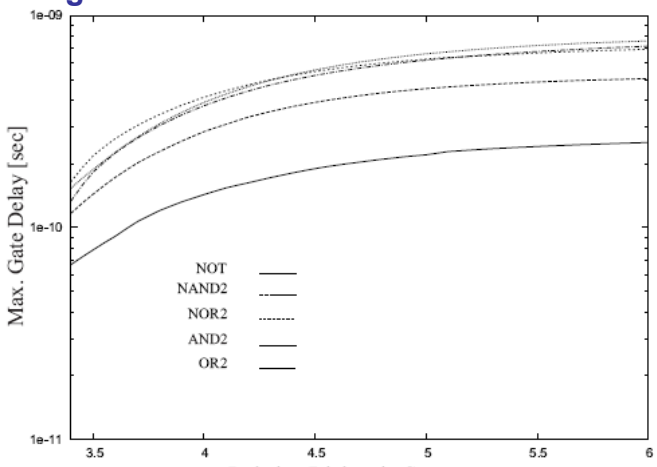
I_{gate} Vs Thickness



I_{gate} Vs Dielectric Constant

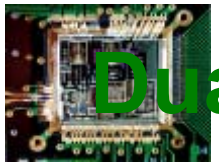


T_{pd} Vs Thickness



T_{pd} Vs Dielectric Constant

Source: Mukherjee ICCD 2005



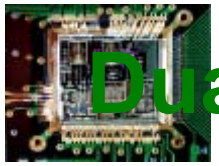
Dual K Technique: Experimental Setup

- DKDT algorithm integrated with SIS, and tested on the ISCAS'85 benchmarks.

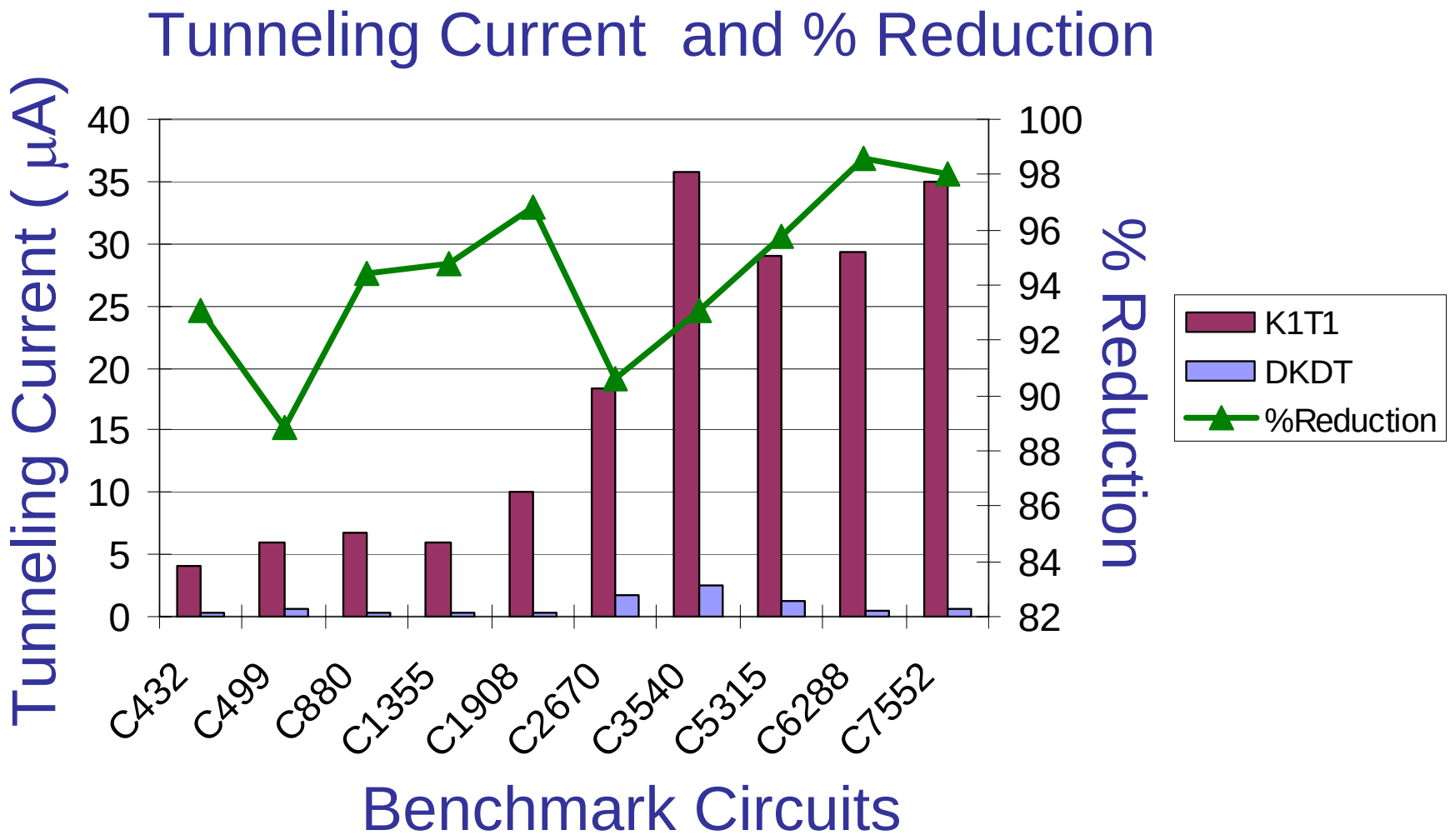
- Used $K_1 = 3.9$ (for SiO_2), $K_2 = 5.7$ (for SiON), $T_1 = 1.4nm$, and $T_2 = 1.7nm$ for our experiments.

- T_1 is chosen as the default value from the BSIM4.4.0 model card and value of T_2 is intuitively chosen

Source: Mukherjee ICCD 2005



Dual K Technique: Experimental Results

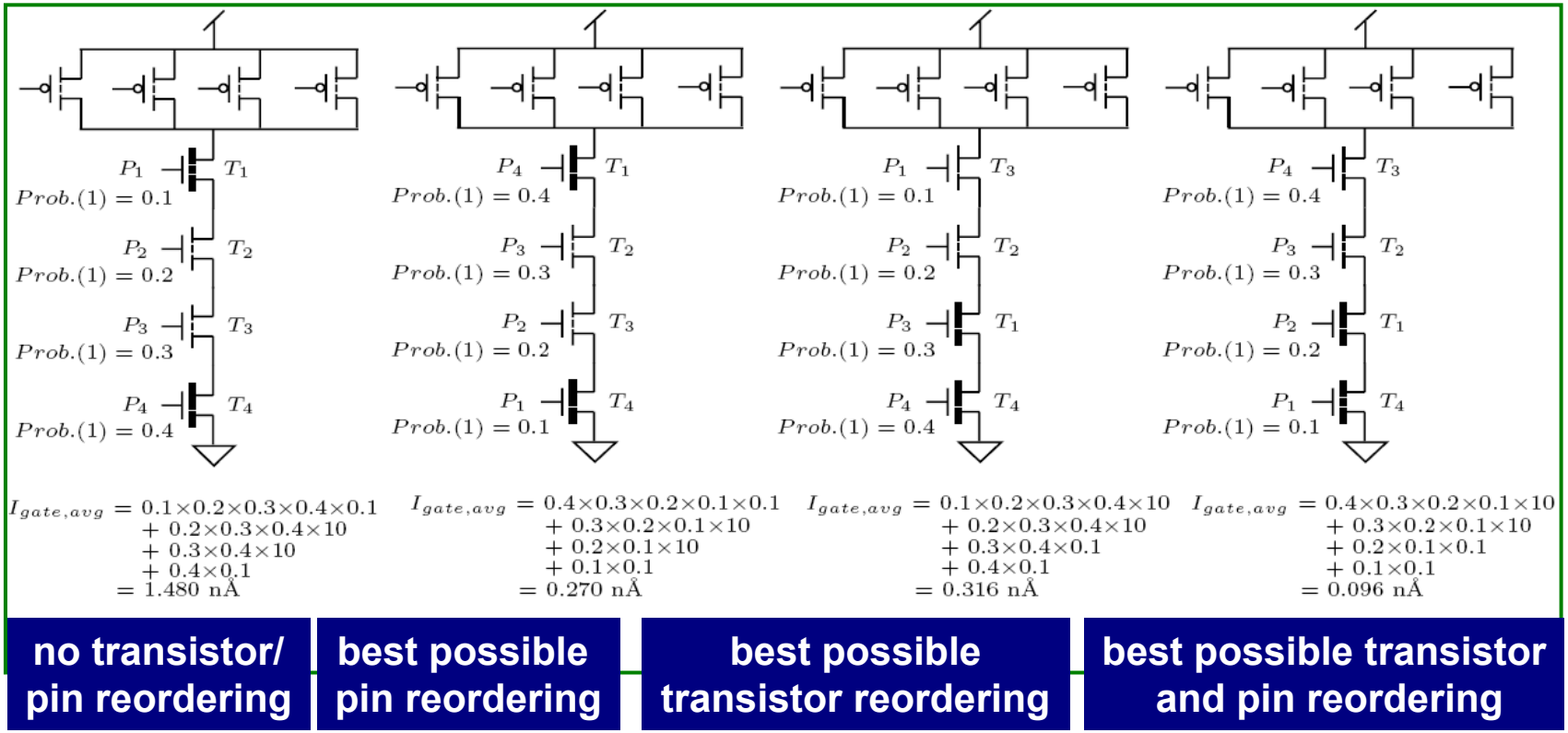


Source: Mukherjee ICCD 2005

Pin Reordering with Dual-Tox

A key difference between the state dependence of I_{sub} and I_{gate}

- I_{sub} primarily depends on the number of OFF in stack
- I_{gate} depends strongly on the position of ON/OFF transistors



no transistor/ pin reordering	best possible pin reordering	best possible transistor reordering	best possible transistor and pin reordering

- Results improve by 5-10% compared to dual-Tox approach.

Source: Sultania ICCD 2004

Conclusions and Future Research

- Gate leakage is an major component of power consumption in nano-scale CMOS circuits.
- Gate leakage is present in both ON and OFF state of a MOS device.
- Few research works so far have addressed its estimation in CMOS circuits.
- Few research works address its reduction in CMOS circuit.
- Use of high-K is expected to be a stable solution for the gate leakage problem, which is largely unaddressed from modeling and synthesis flow point of view.

Thank You