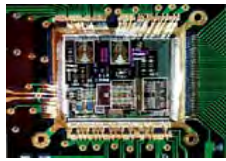
Lecture 5: Gate Leakage

CSCE 6730
Advanced VLSI Systems

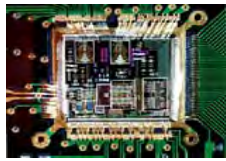
Instructor: Saraju P. Mohanty, Ph. D.

NOTE: The figures, text etc included in slides are borrowed from various books, websites, authors pages, and other sources for academic purpose only. The instructor does not claim any originality.

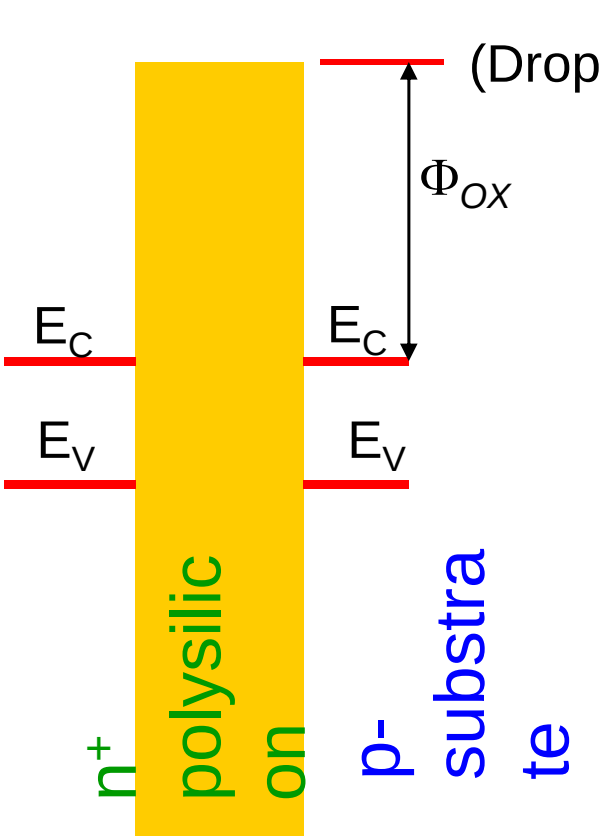


Scaling Trends and Effects: Summary

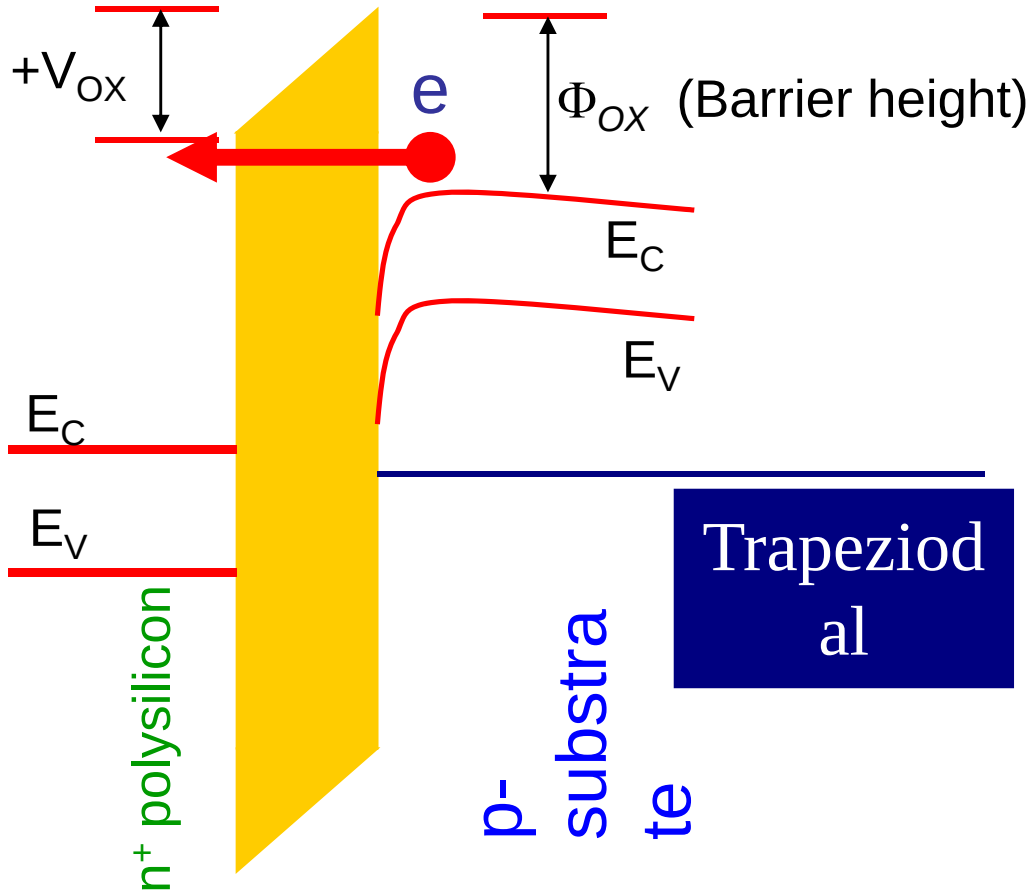
- Scaling improves
  - Transistor Density of chip
  - Functionality on a chip
  - Speed, Frequency, and Performance
- Scaling and power dissipation
  - Active power remains almost constant
  - Components of leakage power increase in number and in magnitude.
  - Gate leakage (tunneling) predominates for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



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



Flat-band Condition

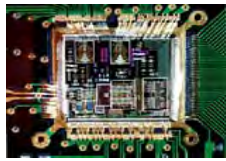


Direct Tunneling for positive bias

NOTE: For short channel MOS FN tunneling is negligible.

Source: AgarwalIEEPDTMay2005

E. Kougianos and **S. P. Mohanty**, "Metrics to Quantify Steady and Transient Gate Leakage in Nanoscale Transistors: NMOS Vs PMOS Perspective", in *Proceedings of the 20th IEEE International Conference on VLSI Design (VLSID)*, pp. 195-200, 2007.



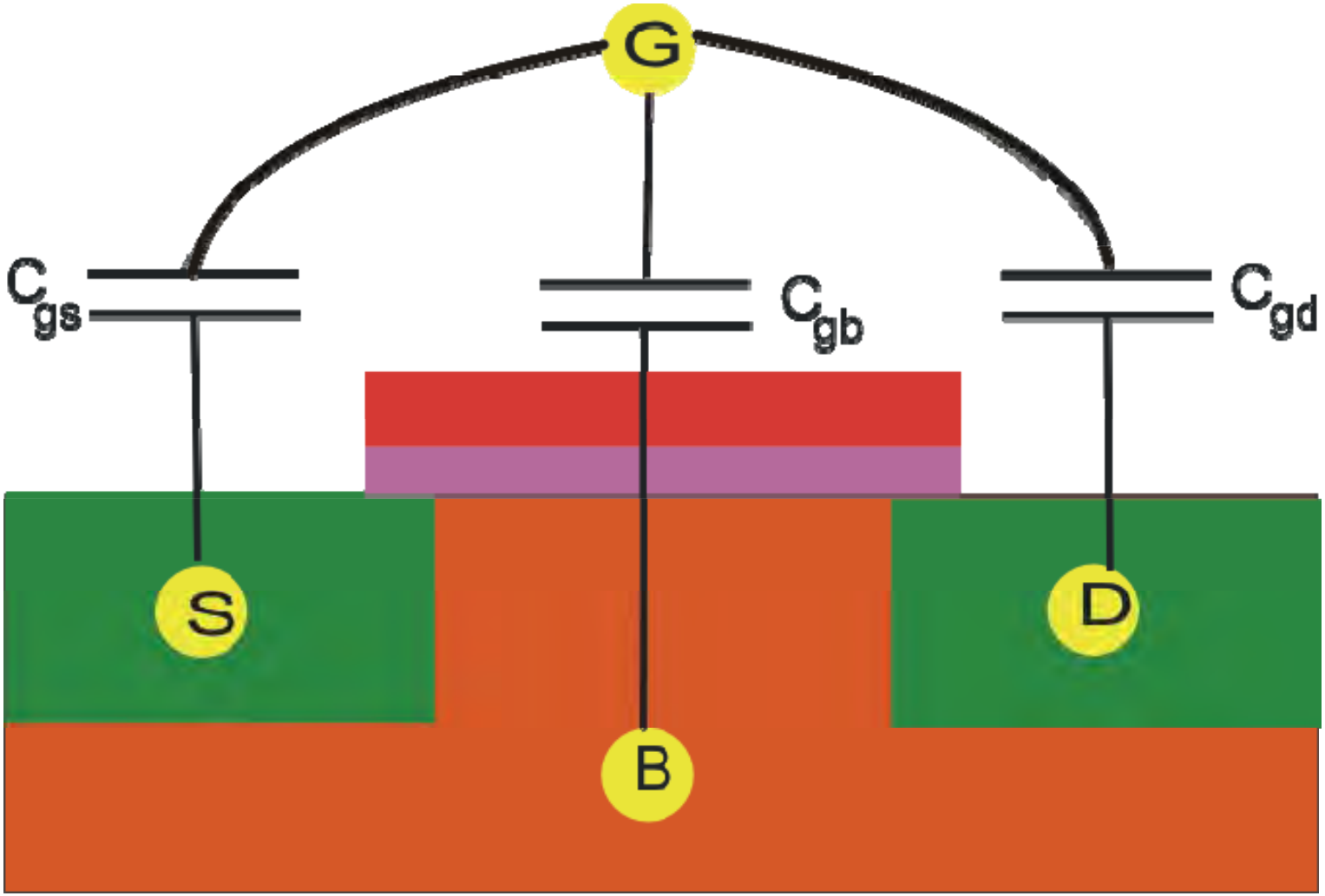
Outline

1. Both ON and OFF state gate leakage are significant.
2. During transition of states there is transient effect is gate tunneling current.
3. Three metrics: $I_{ON,}$ $I_{OFF,}$ and $C_{tunneling}$
4. $C_{tunneling}$: Manifests to intra-device loading effect of the tunneling current.
5. NMOS Vs PMOS in terms of three metrics.
6. Study process/supply variation on three metrics.

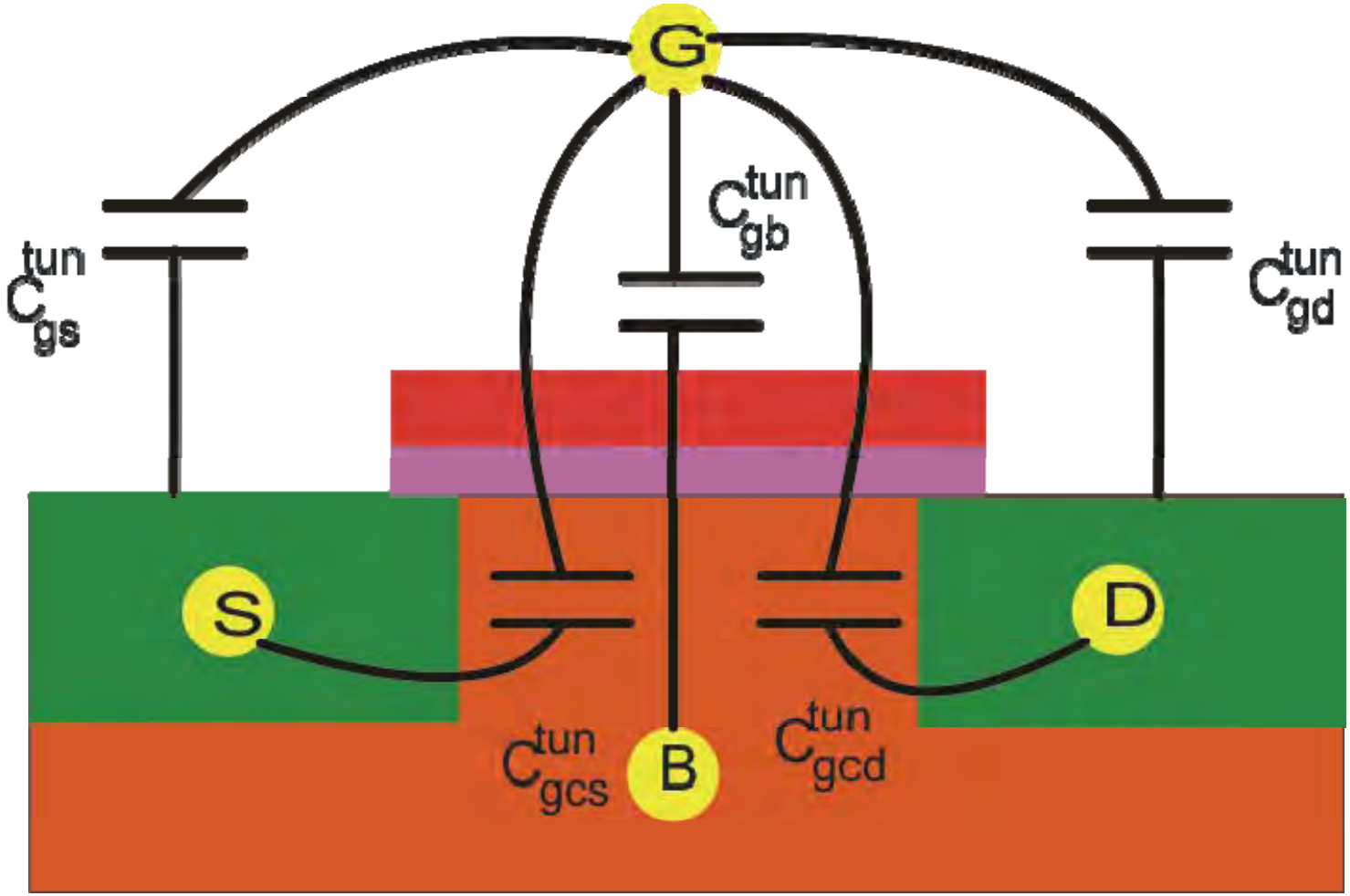
Salient Point

The metric, effective tunneling capacitance essentially quantifies the intra-device loading effect of the tunneling current and also gives a qualitative idea of the driving capacity of a Nano-CMOS transistor.

Gate Capacitance of a Transistor (Intrinsic)



Gate Capacitance of a Transistor (Tunneling: Proposed)



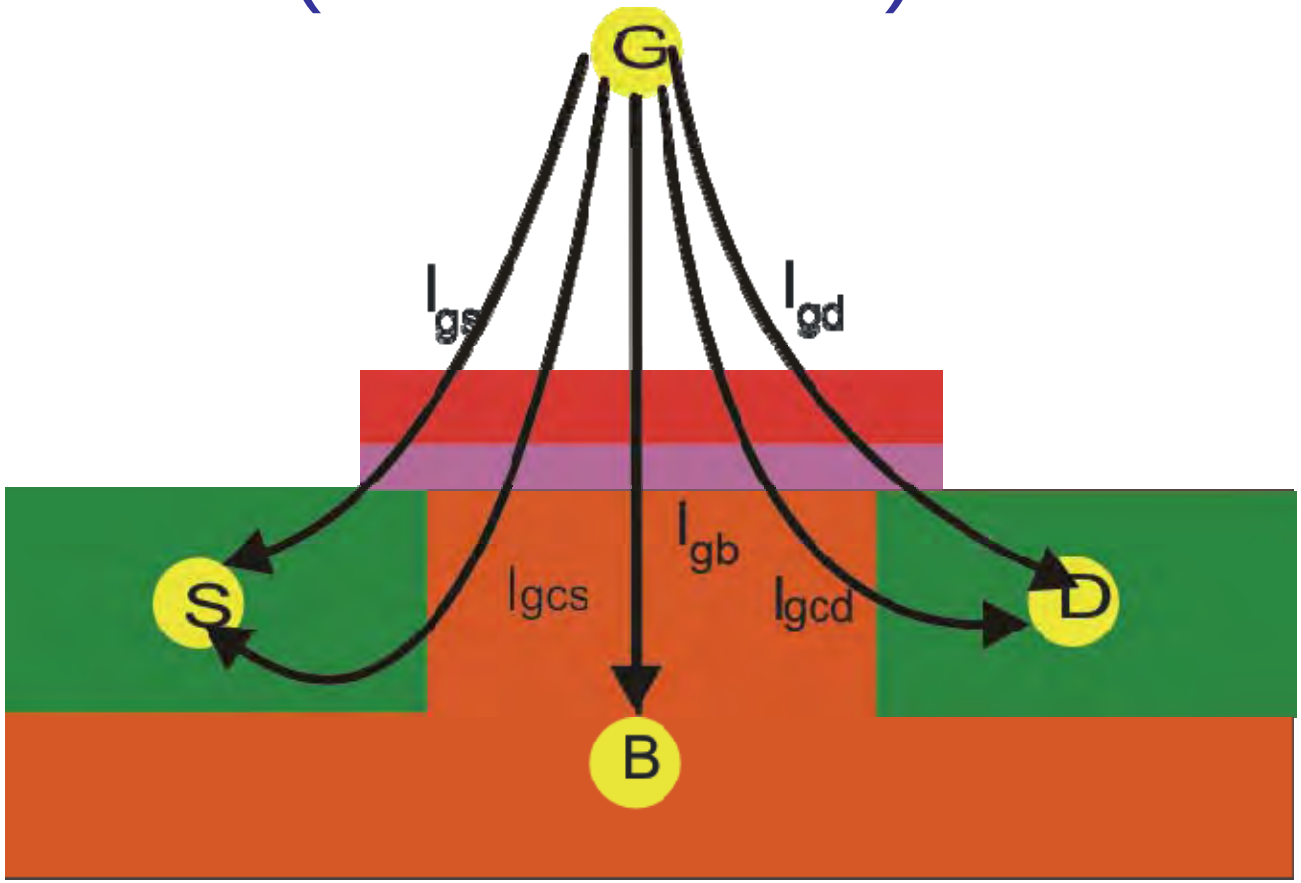
We propose that transient in gate tunneling current due to state transitions are manifested as capacitances.

Analysis
in a
Nano-CMOS Transistor

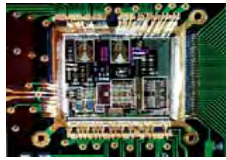
Outline: Nano-CMOS Transistor

- Dynamics of gate oxide tunneling in a transistor
- SPICE model for gate leakage
- ON, OFF, and transition states of a transistor
- Gate leakage in ON, OFF, and transition states of a transistor

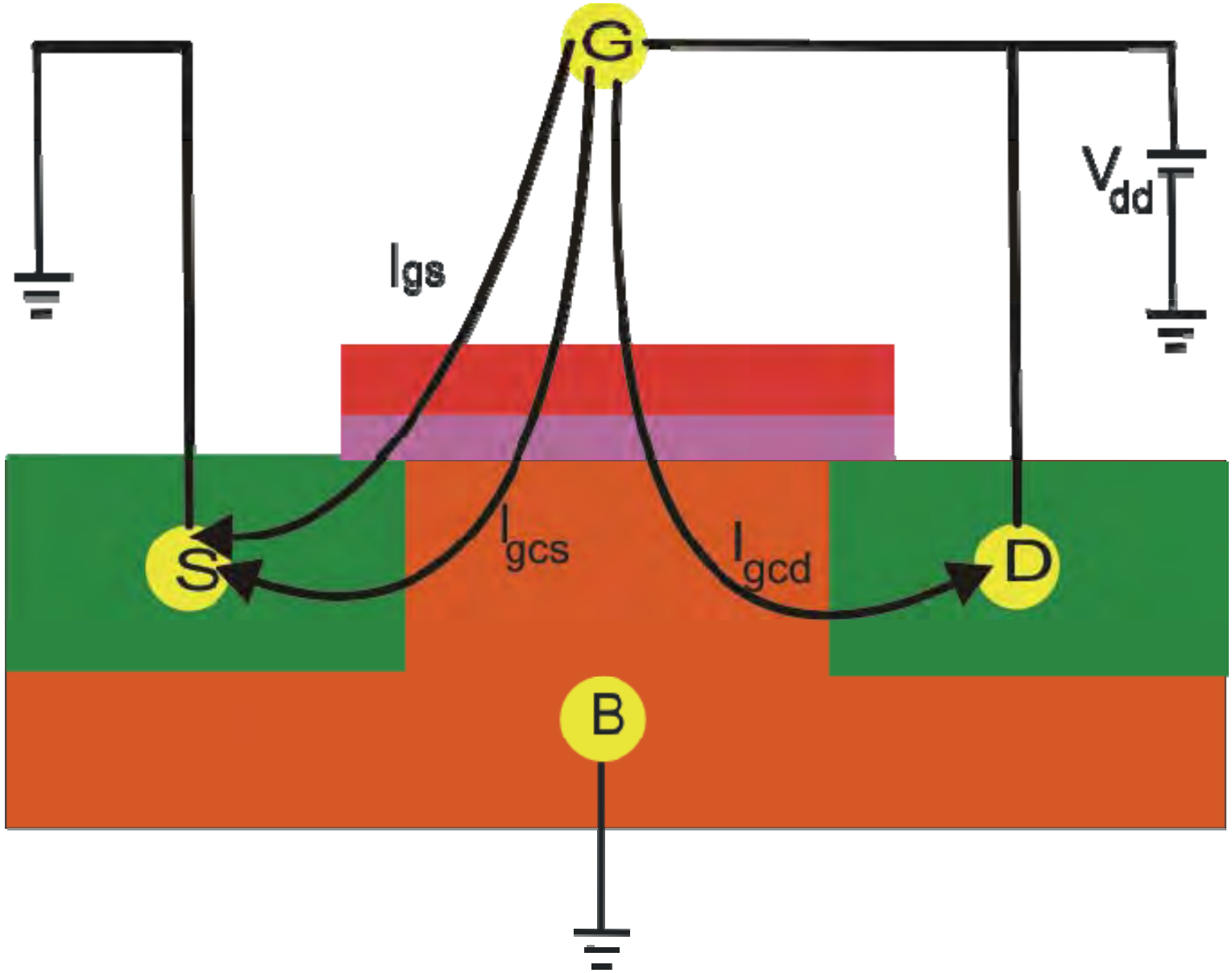
Gate Leakage Components (BSIM4 Model)

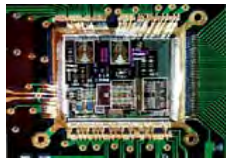


- I_{gs}, I_{gd}: tunneling through overlap of gate and diffusions
- I_{gcs}, I_{gcd}: tunneling from the gate to the diffusions via channel
- I_{gb}: tunneling from the gate to the bulk via the channel

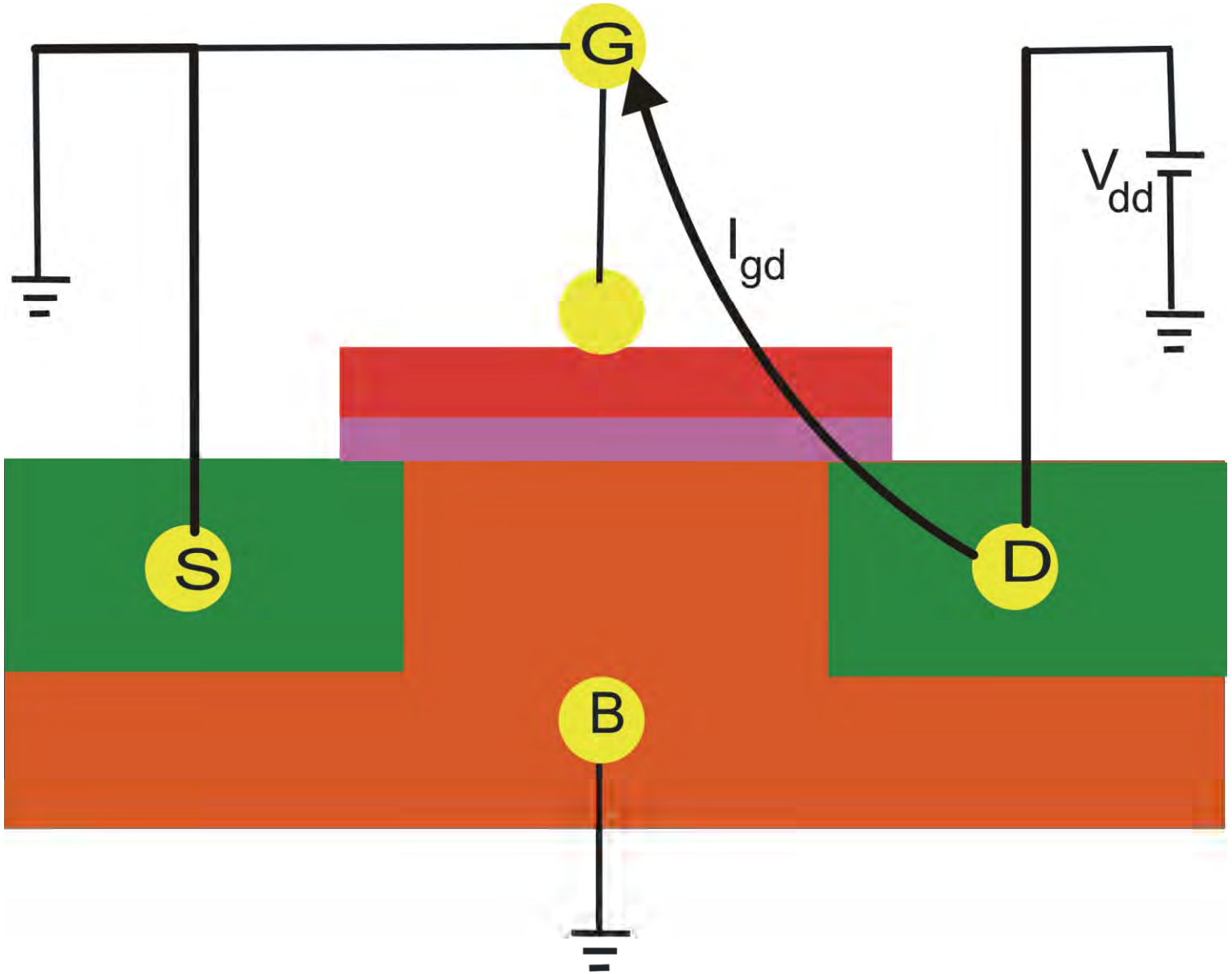


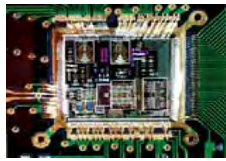
ON State: NMOS Transistor



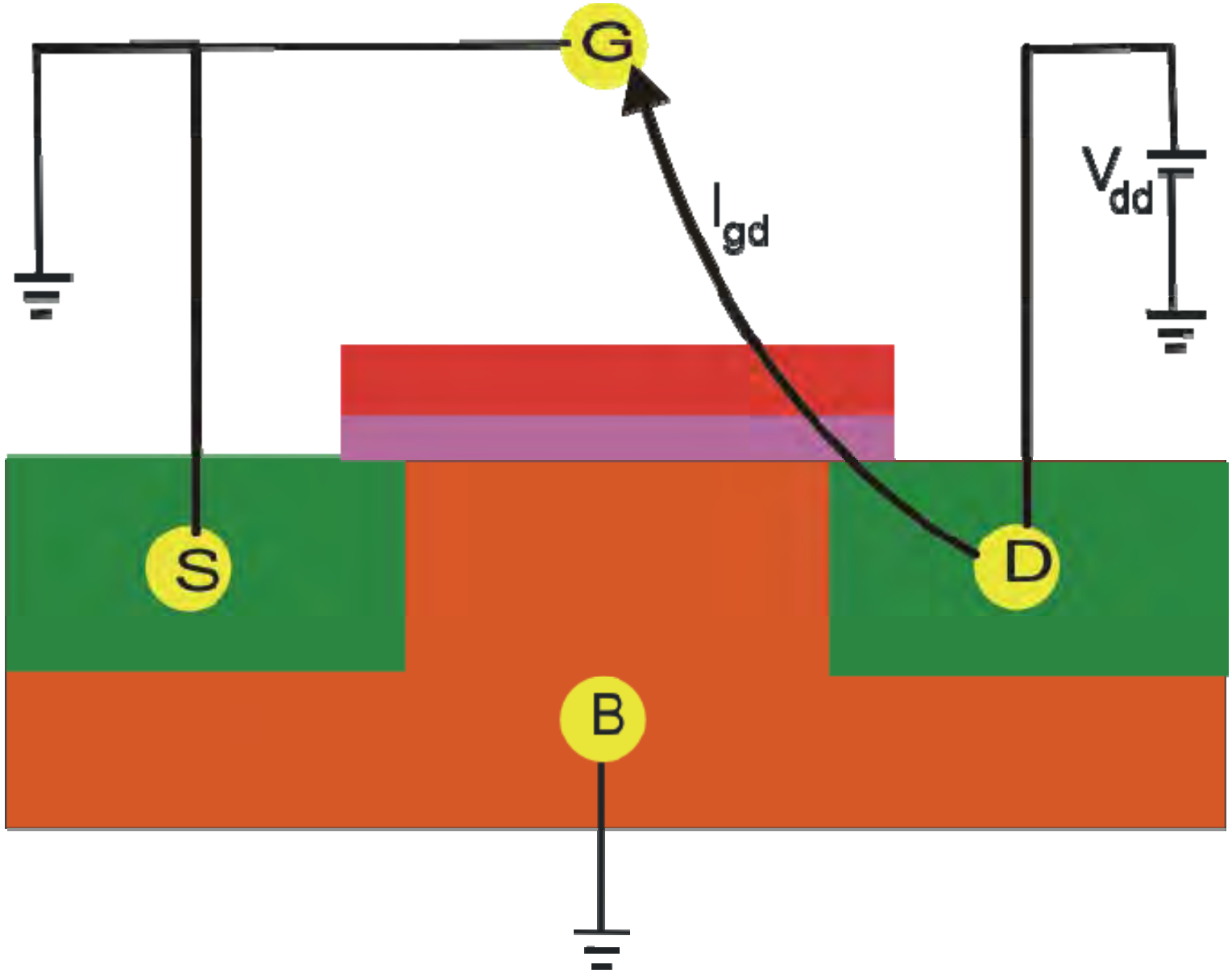


ON State: PMOS Transistor

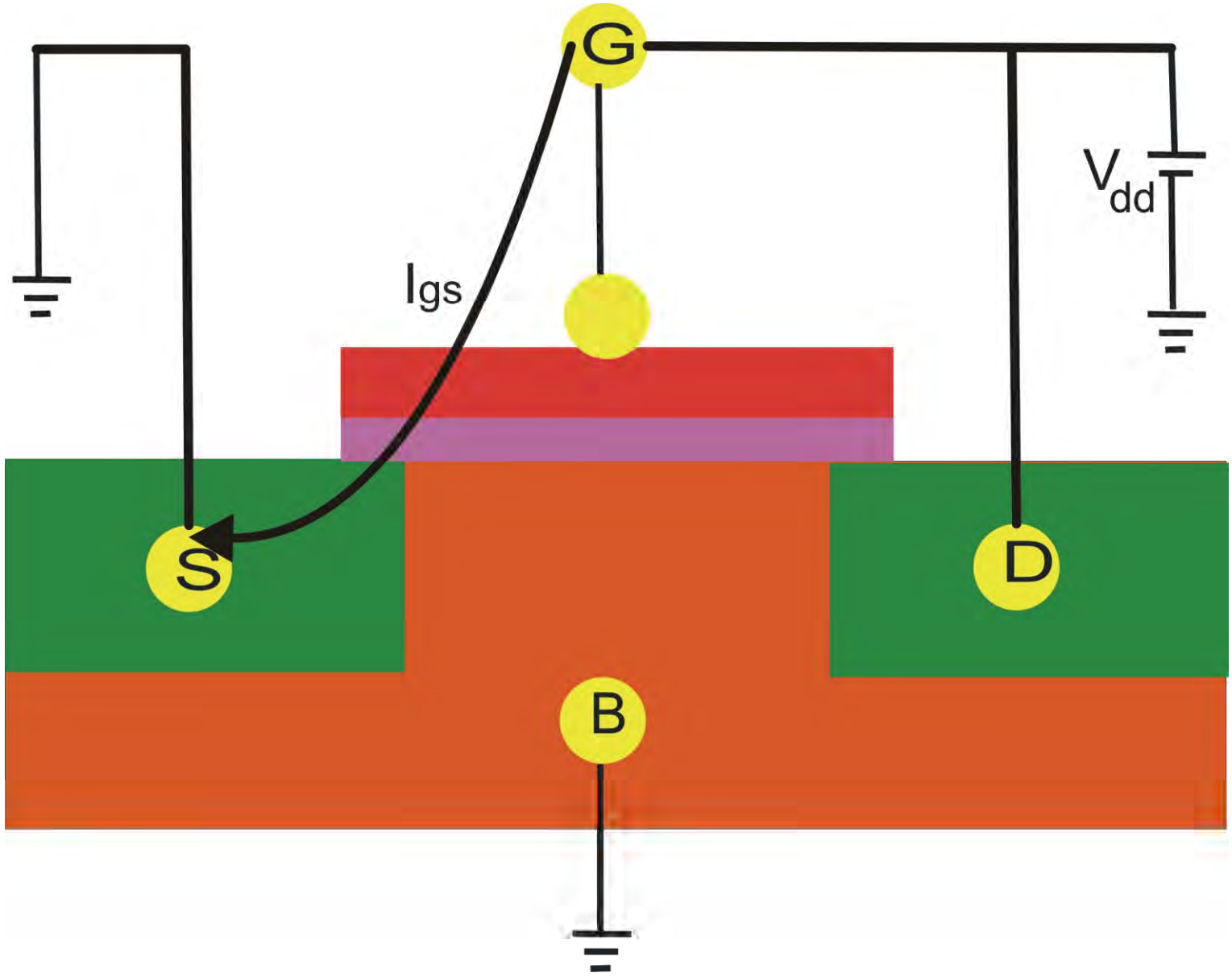




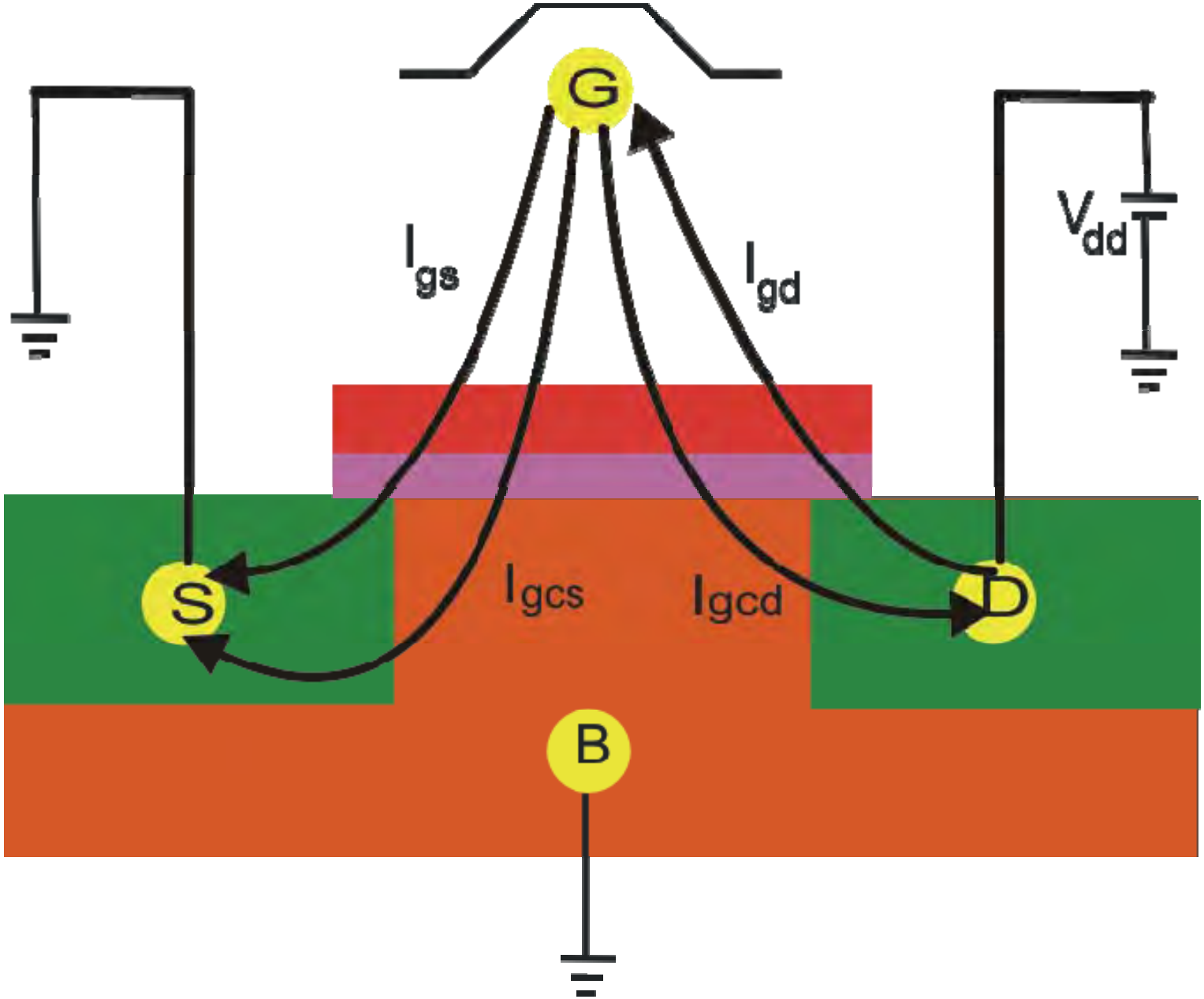
OFF State: NMOS Transistor

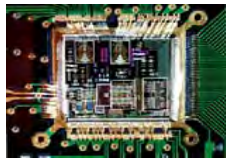


OFF State: PMOS Transistor

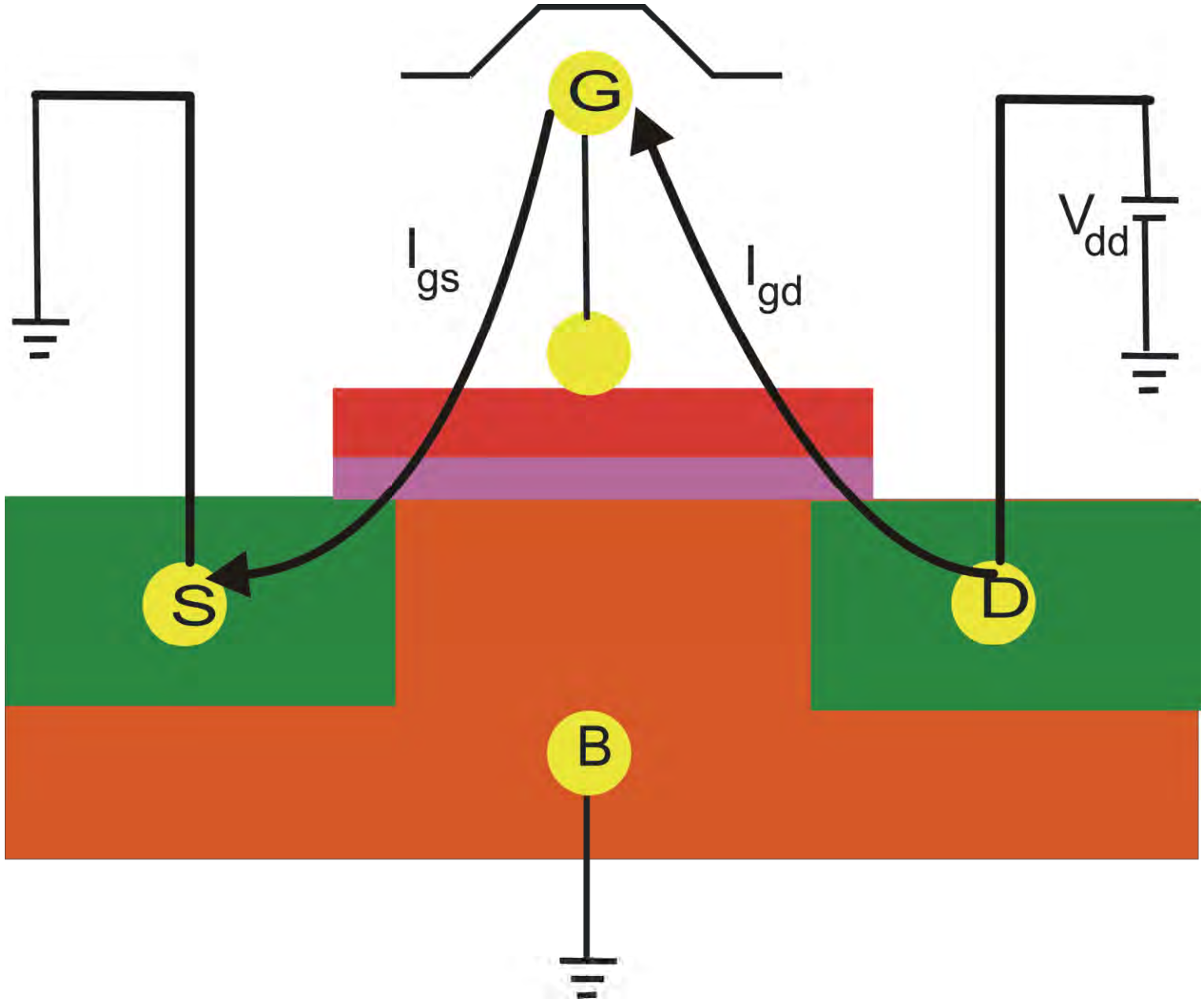


Transition State: NMOS Transistor





Transition State: PMOS Transistor



NMOS Gate Leakage Current (For a Switching Cycle)

Fig. 1

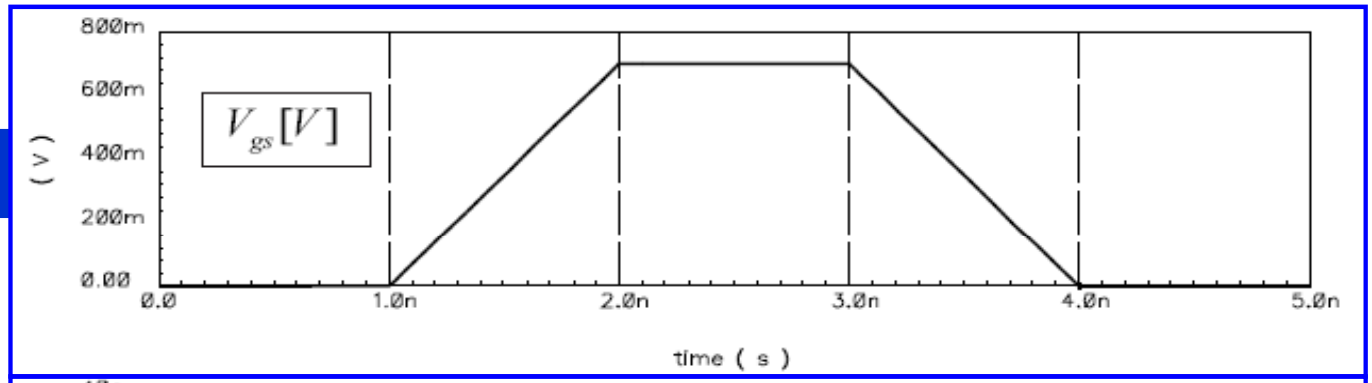
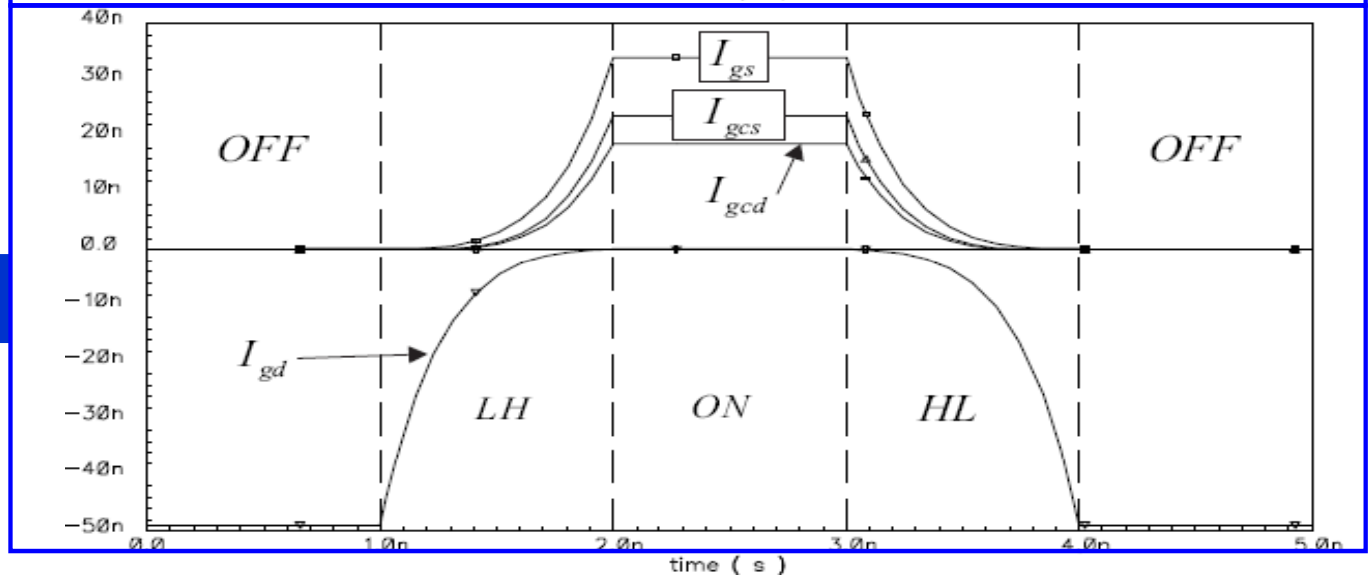
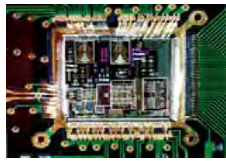


Fig. 2





PMOS Gate Leakage Current (For a Switching Cycle)

Fig. 1

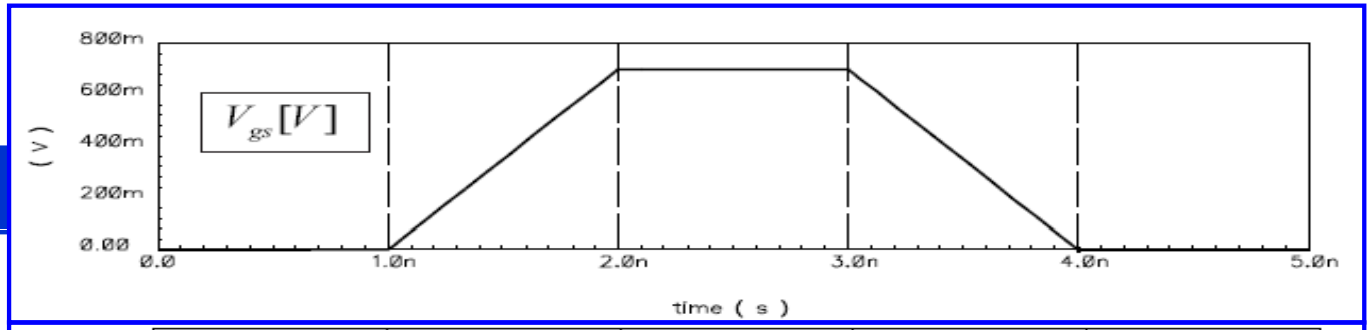
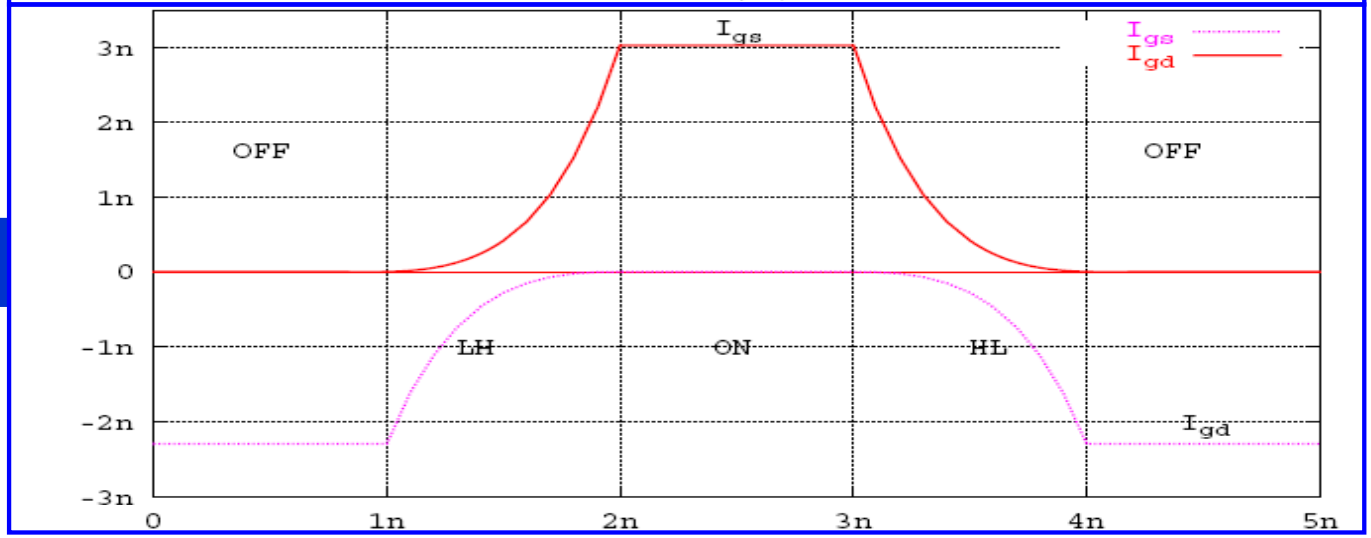
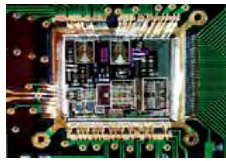


Fig. 2





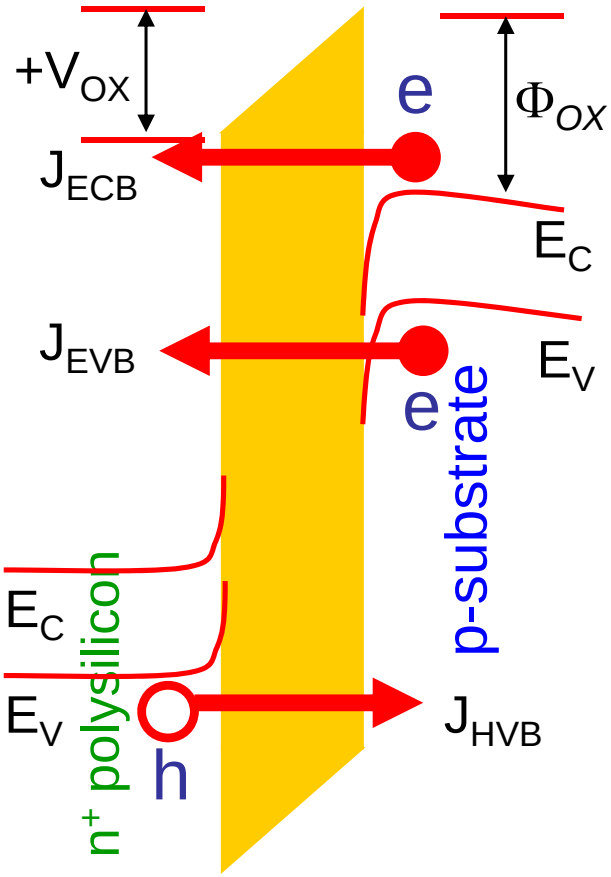
NMOS Vs PMOS: 3 Mechanisms of Tunneling

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

Three Metrics for Tunneling Current

Gate Leakage: Observation

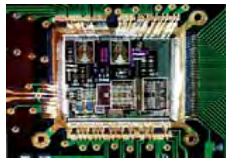
The behavior of the device in terms of gate tunneling leakage must be characterized not only during the steady states but also during transient states.

Gate Leakage: Metrics

- Gate leakage happens in ON state: I_{ON}
- Gate leakage happens in OFF state: I_{OFF}
- Gate leakage happens during transition: C_{eff}^{tun}

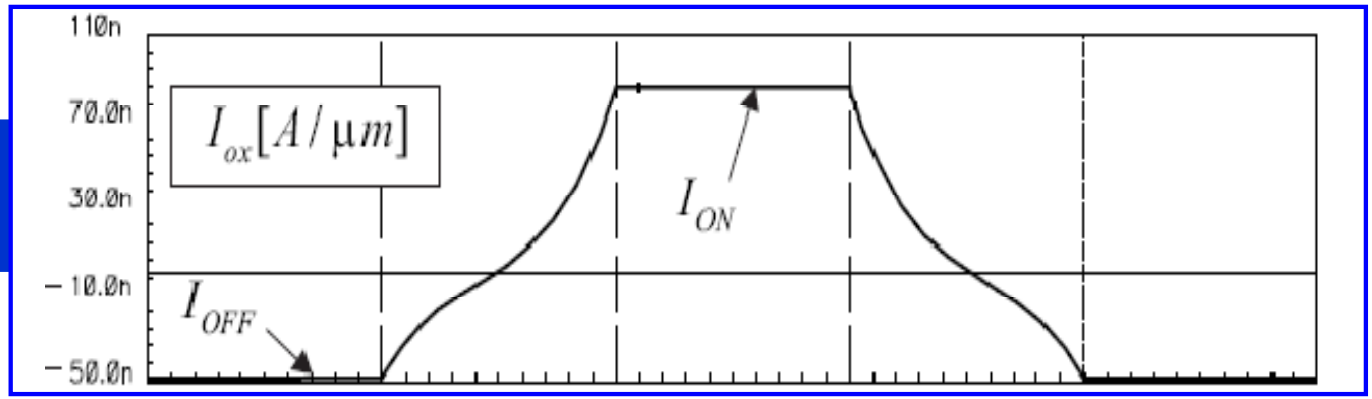
Gate Leakage for a Transistor

- 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, OFF, or transition

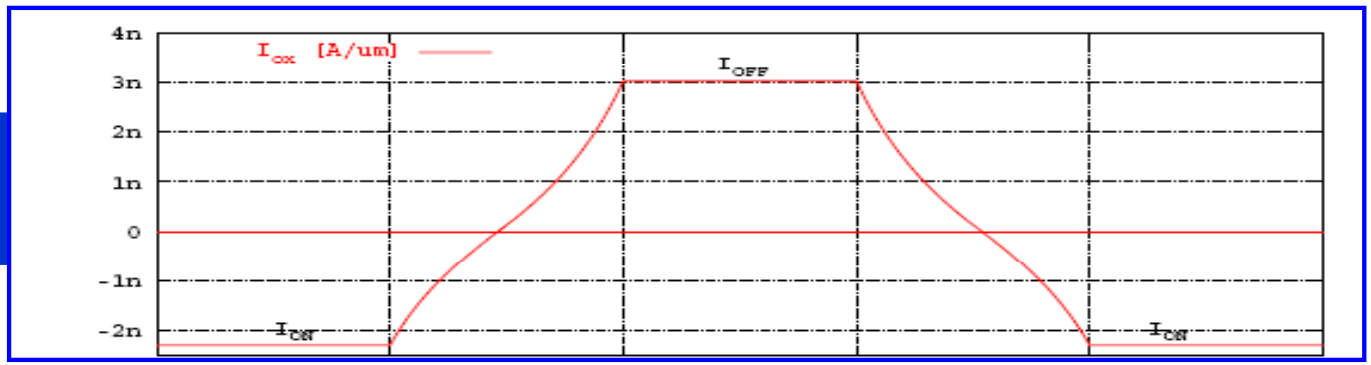


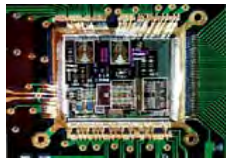
Gate Leakage Current (For a Switching Cycle)

For NMOS

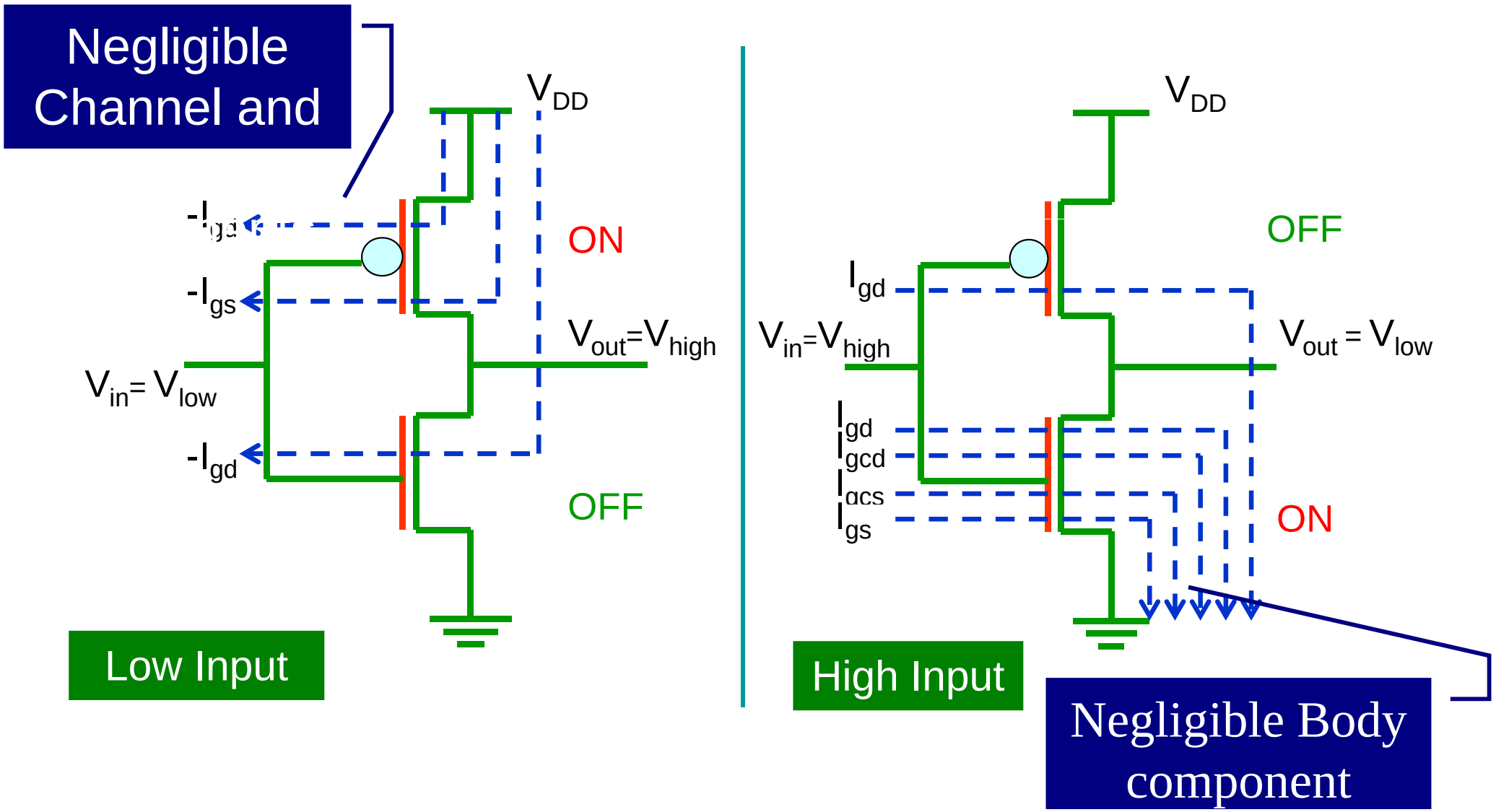


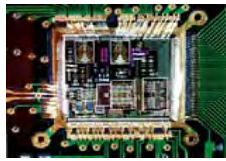
For PMOS



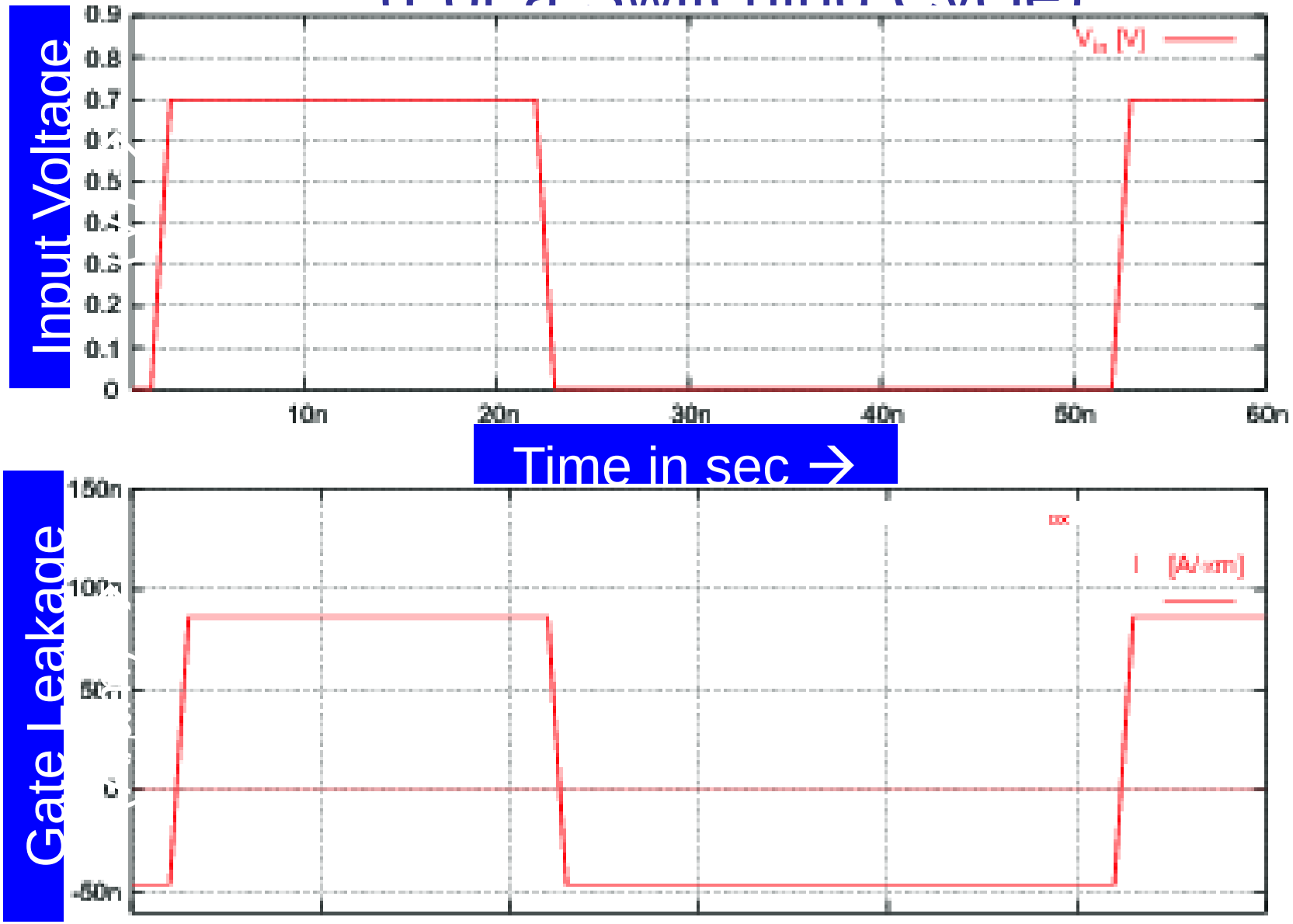


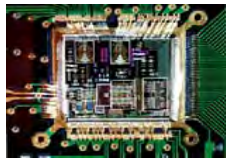
Inverter: Gate Leakage Paths (Putting NMOS and PMOS together)



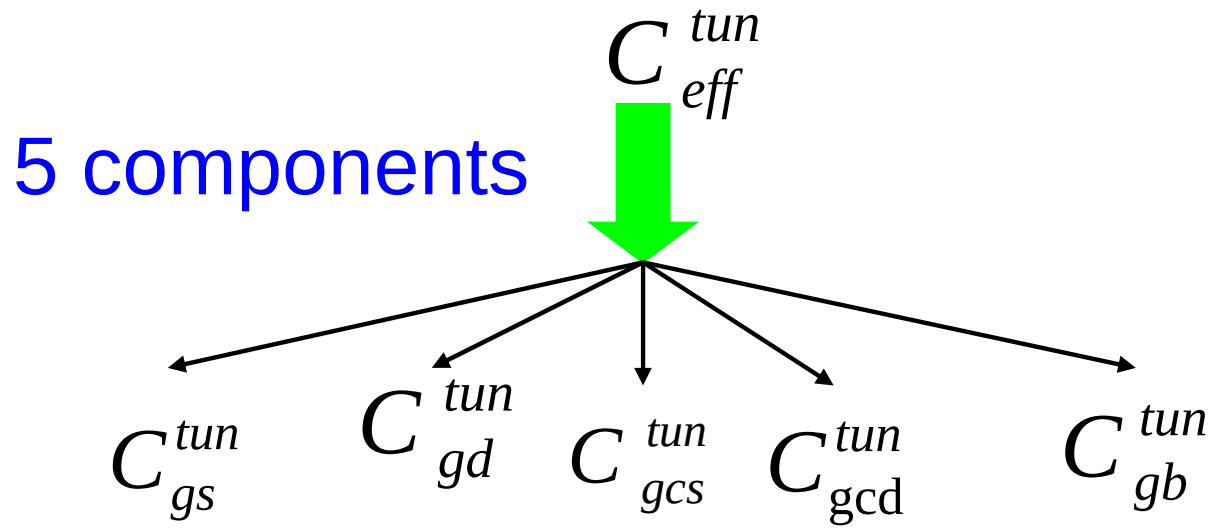


Inverter: Gate Leakage Current (For a Switching Cycle)





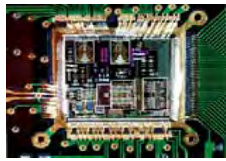
Transient Gate Leakage: C_{eff}^{tun}



We propose to quantify as:

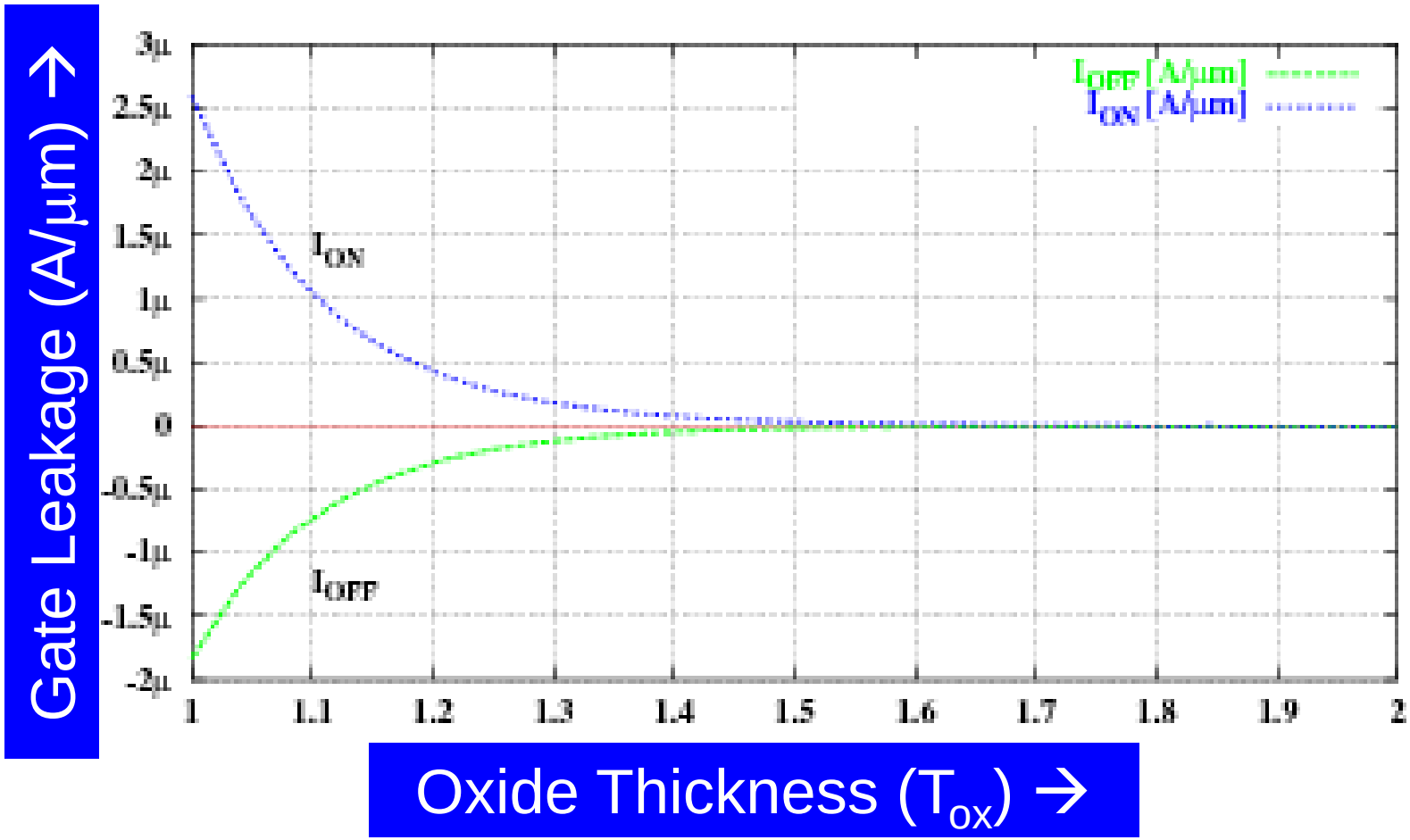
$$C_{eff}^{tun} = \frac{I_{ON} - I_{OFF}}{\left(\frac{dV_g}{dt}\right)}$$

$$= \frac{I_{ON} - I_{OFF}}{V_{DD}} t_r \text{ (for equal rise/fall time)}$$

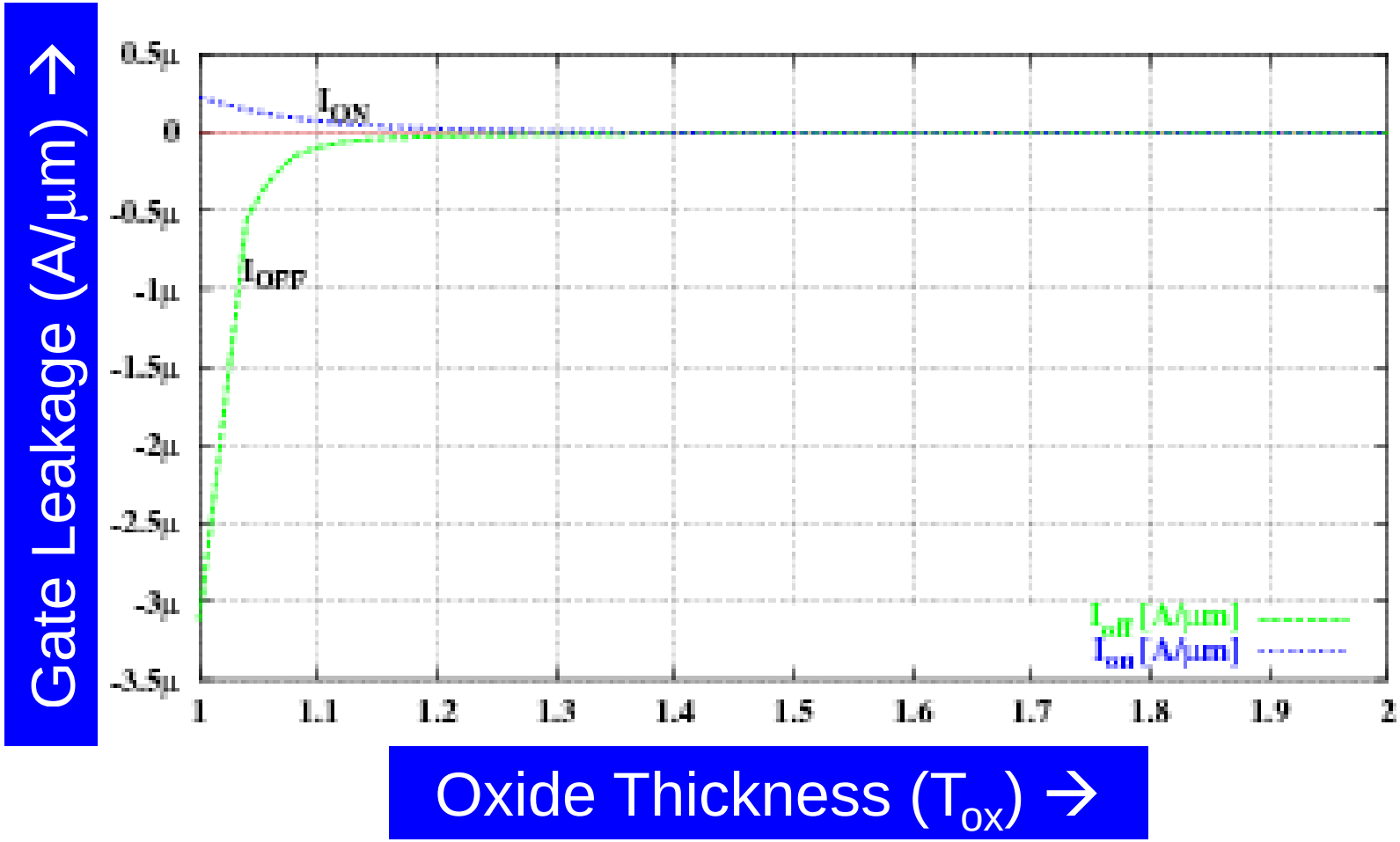


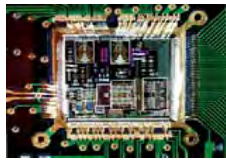
Effect of Process and Design Parameter Variation

I_{ON} / I_{OFF} Versus T_{ox}: NMOS

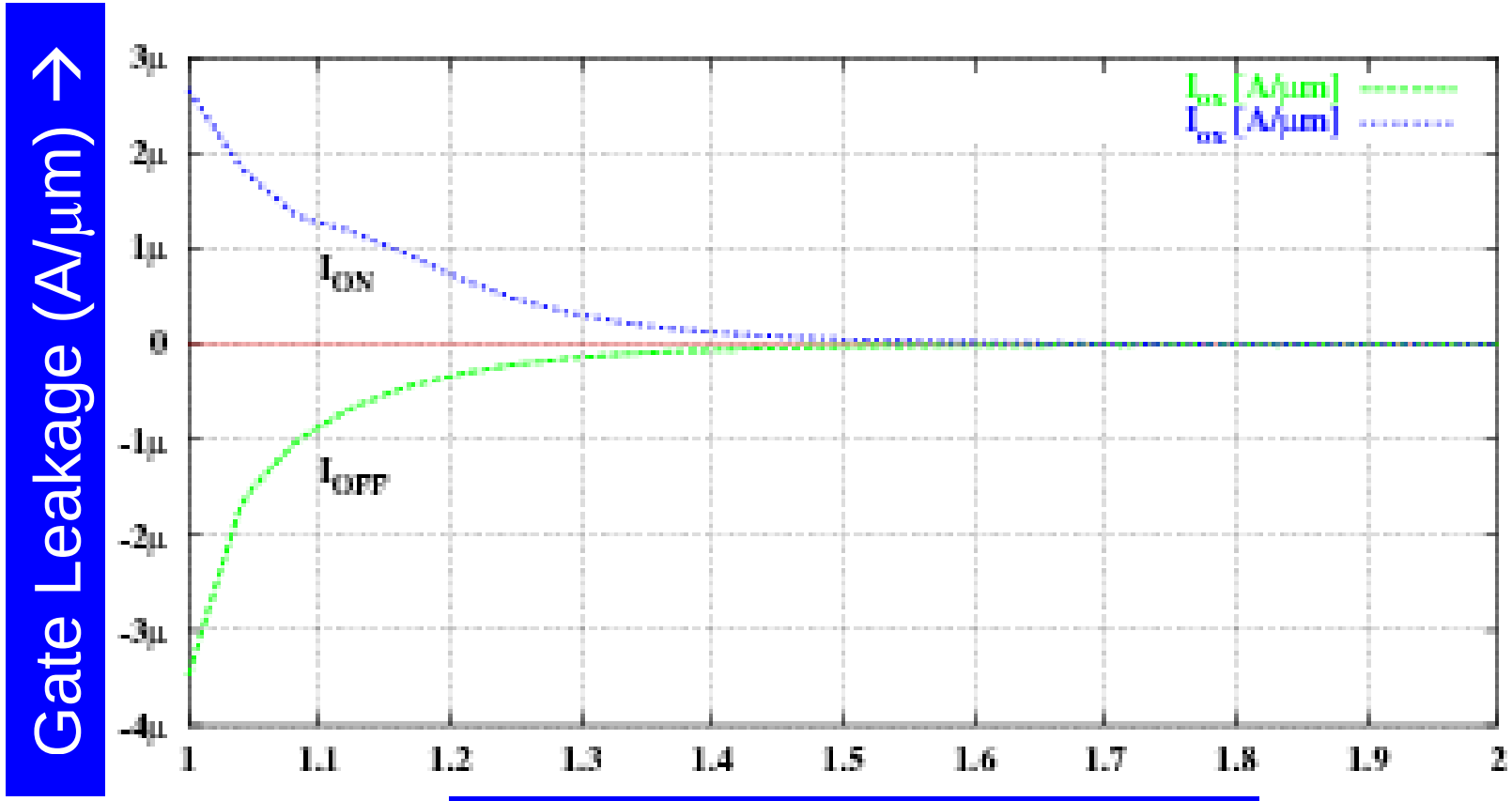


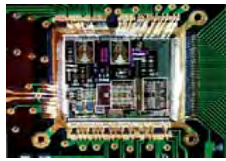
I_{ON} / I_{OFF} Versus T_{ox}: PMOS



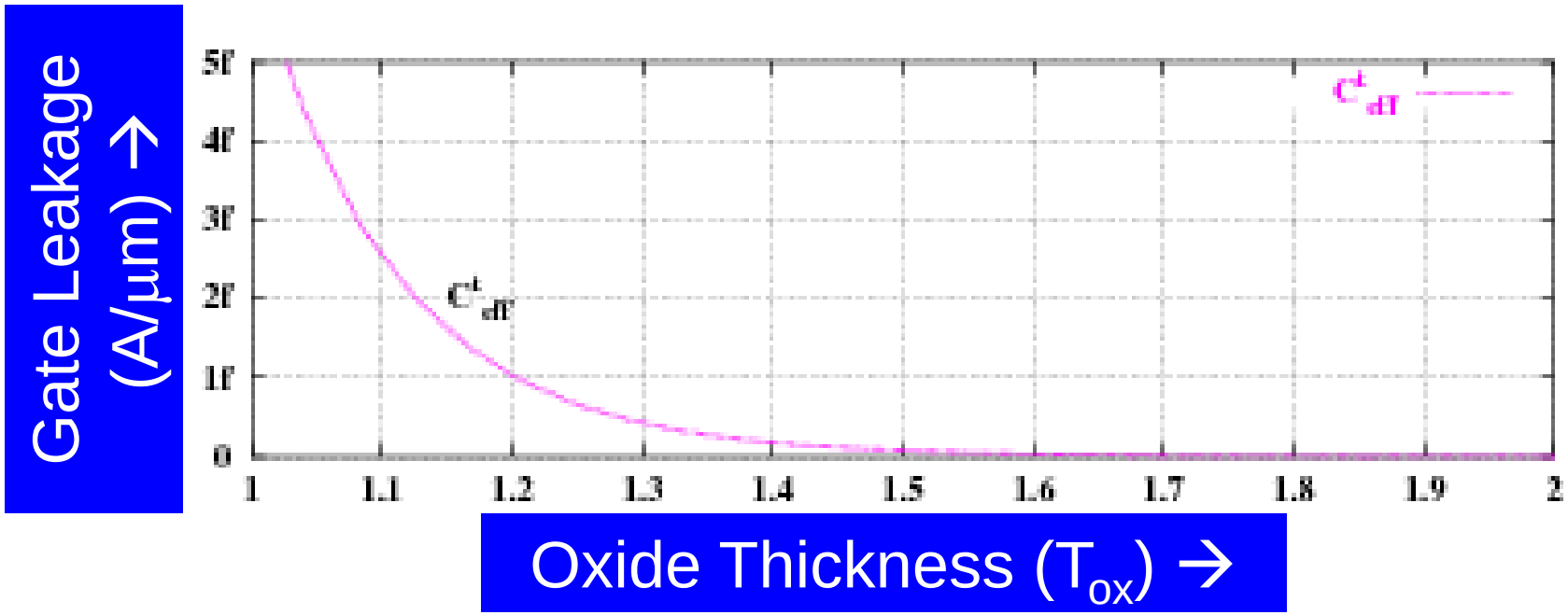


I_{ON} / I_{OFF} Versus T_{ox}: Inverter

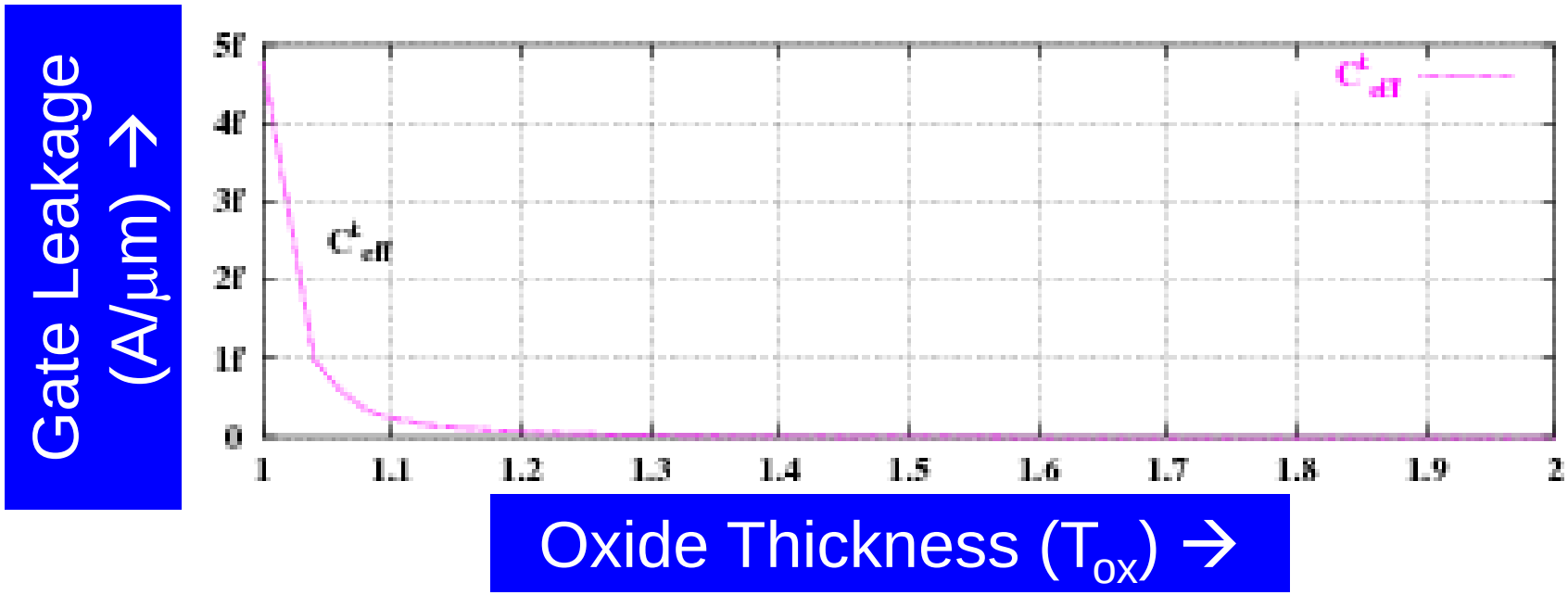


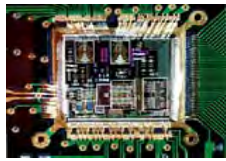


C_{eff}^{tun} Versus T_{ox}: NMOS

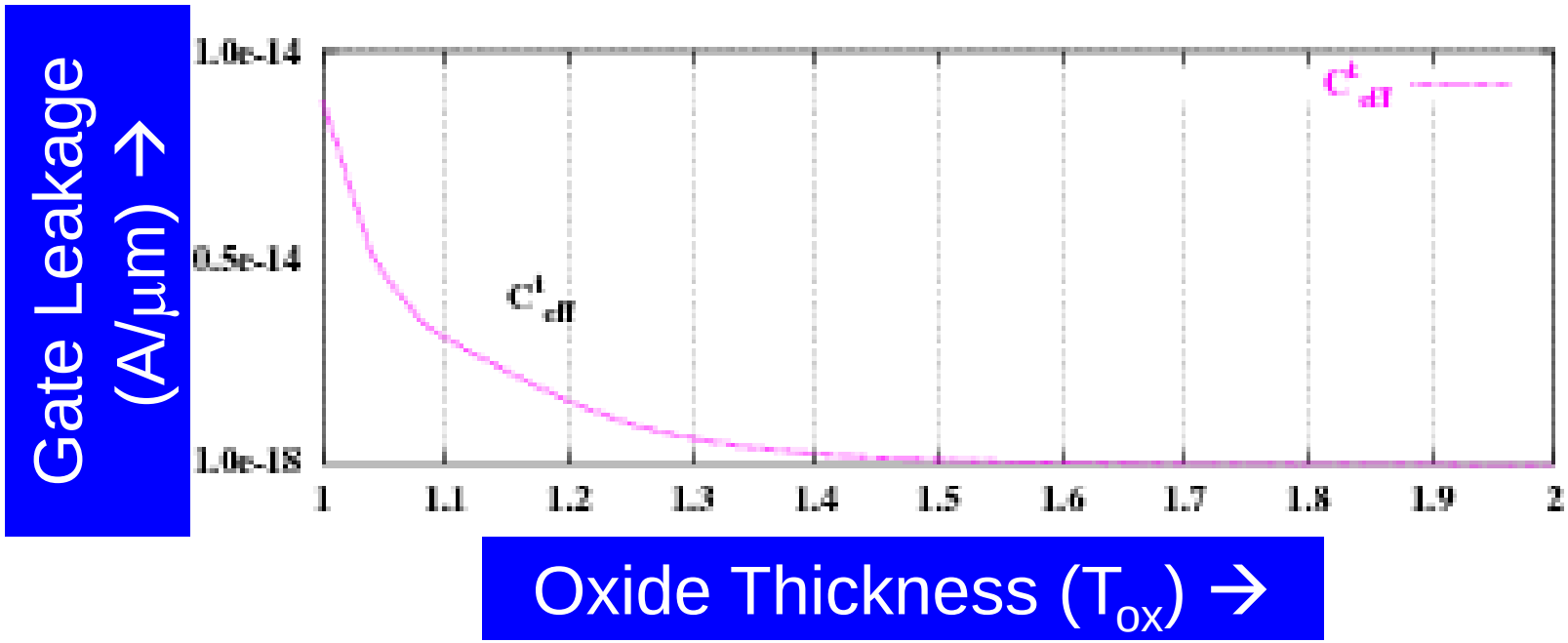


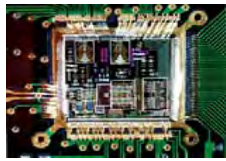
C_{eff}^{tun} Versus T_{ox}: PMOS





C_{eff}^{tun} Versus T_{ox}: Inverter

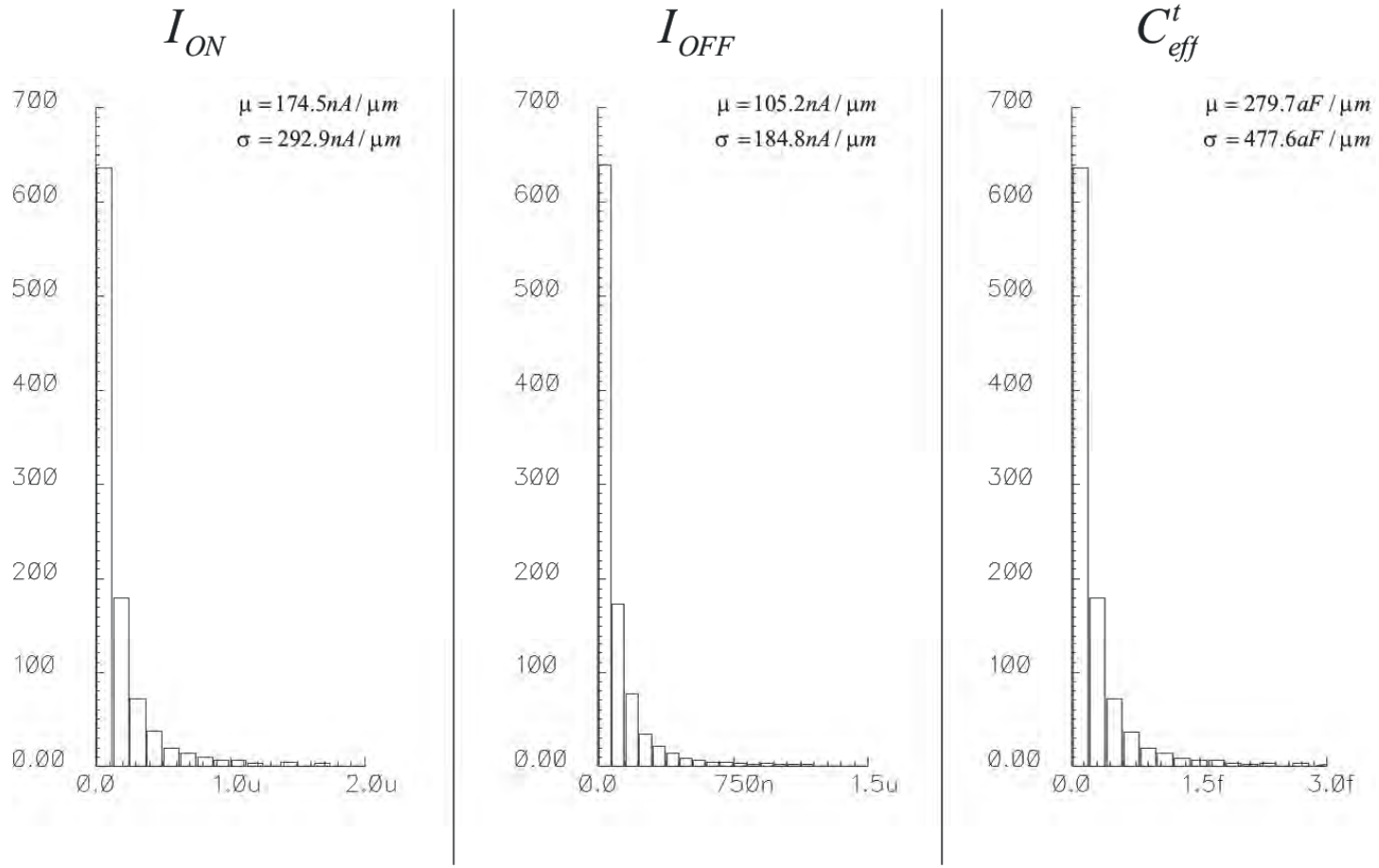


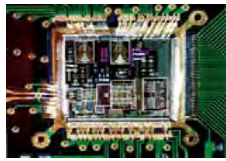


Statistical Analysis of the Metrics

Monte Carlo Simulations: (Modeling Variations)

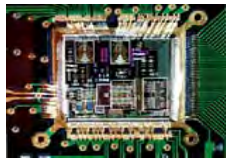
- Monte Carlo (N=1000) results.
  - 10% variation in gate oxide and supply assumed.





Monte Carlo Simulations: (Modeling Variations)

- All three metrics follow lognormal distribution.
- This is expected since gate T_{ox} and V_{dd} are assumed normally distributed and I_{ox} depends exponentially on both.
- Small parameter variation (10%) leads to large deviance in the metrics (2-3 sigma).



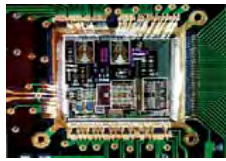
Gate Leakage in Nano-CMOS

- Both ON and OFF states contribute to gate oxide leakage.
- Transient effect is significant and can be captured via effective tunneling capacitance.
- I_{ON} and I_{OFF} metrics to quantify gate leakage current during steady state.
- C^{eff}_{tun} ≡ Effective tunneling capacitance at the input of a logic gate.

S. P. Mohanty and E. Kougianos, "Steady and Transient State Analysis of Gate Leakage Current in Nanoscale CMOS Logic Gates", in *Proceedings of the 24th IEEE International Conference on Computer Design (ICCD)*, pp. 210-215, 2006.

Outline

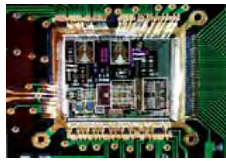
1. Both ON and OFF state gate leakage are significant.
2. During transition of states there is transient effect is gate tunneling current.
3. New metrics: I_{tun} and C_{tun}
4. C_{tun}: Manifests to intra-device loading effect of the tunneling current
5. NOR Vs NAND in terms of I_{tun} and C_{tun}
6. Study process/design variation on I_{tun} and C_{tun}



Salient Point

A new metric, the effective tunneling capacitance essentially quantifies the intra-device loading effect of the tunneling current and also gives a qualitative idea of the driving capacity of the logic gate.

How to quantify it at transistor and logic-gate level??

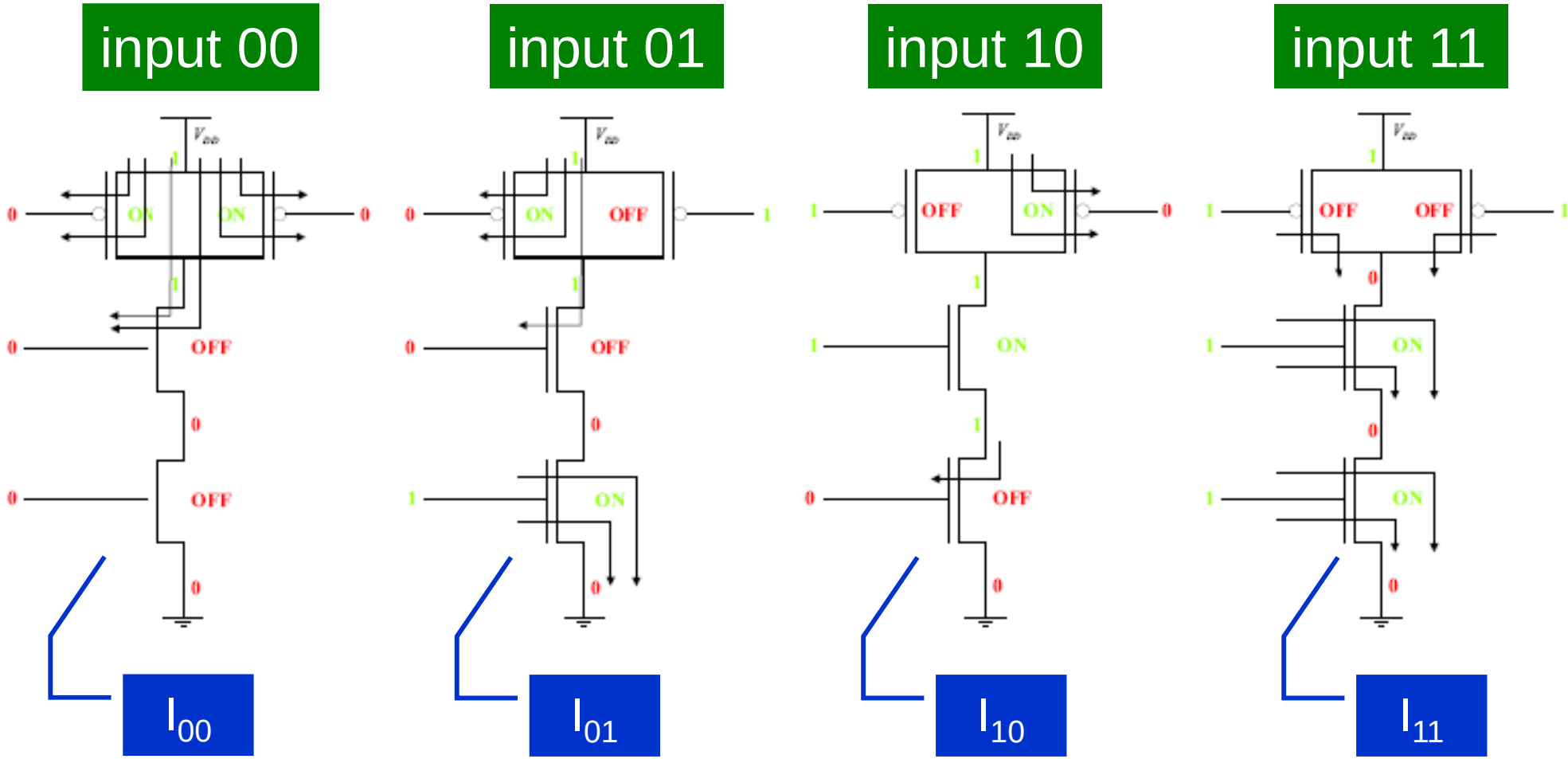


Transistor → Logic Gate

- How do we quantify the same metrics at logic level??
- State dependent or state independent??

Analysis
in
Logic Gates

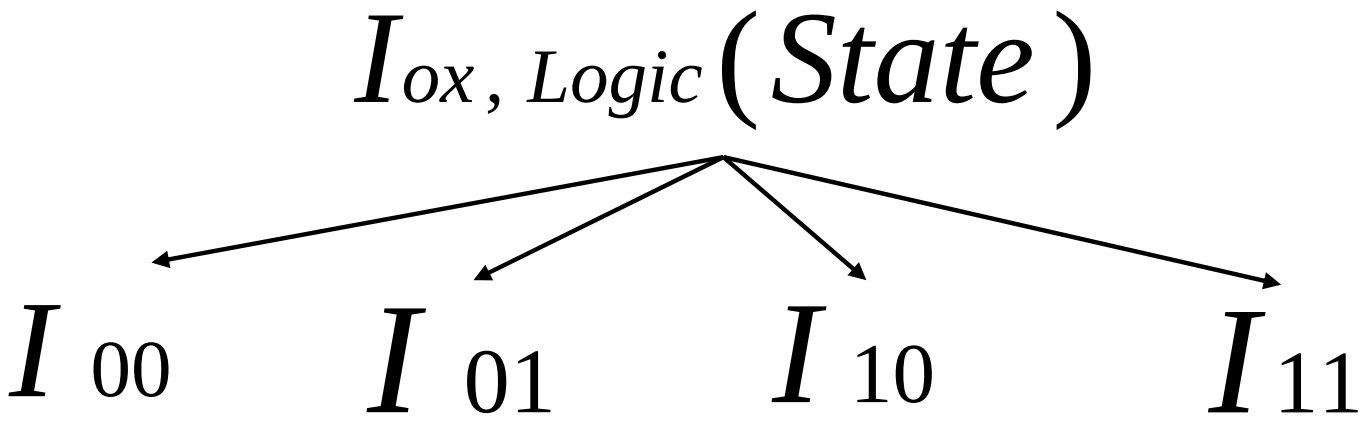
Gate Leakage in 2-input NAND (State Specific)

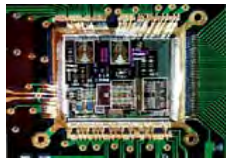


Gate Leakage in 2-input NAND (State Specific)

$$I_{ox,\,Logic}(State) = \sum_{MOS,i} I_{ox,\,i}$$

Four different states for 2-input NAND:





Gate Leakage in 2-input NAND (State Independent)

I_{tun} ≡ State Independent average gate leakage current of a logic gate

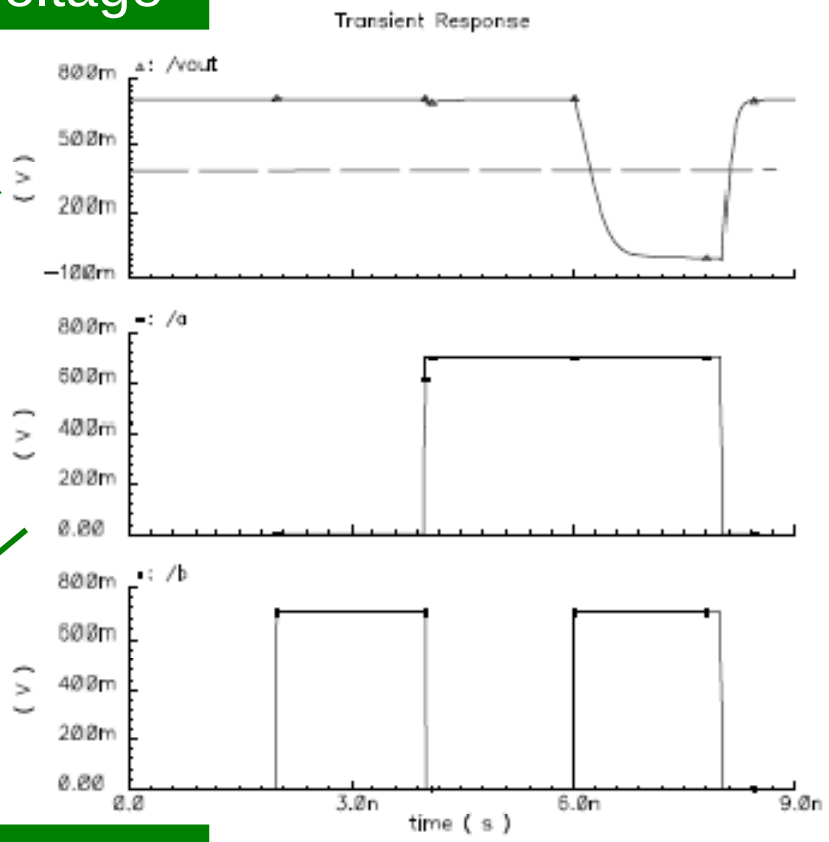
$$I_{tun} = \frac{1}{4}\left(I_{00} + I_{01} + I_{10} + I_{11}\right)$$

This is a measure of gate leakage of a logic gate during its steady state.

Gate Leakage in 2-input NAND (Transient Study)

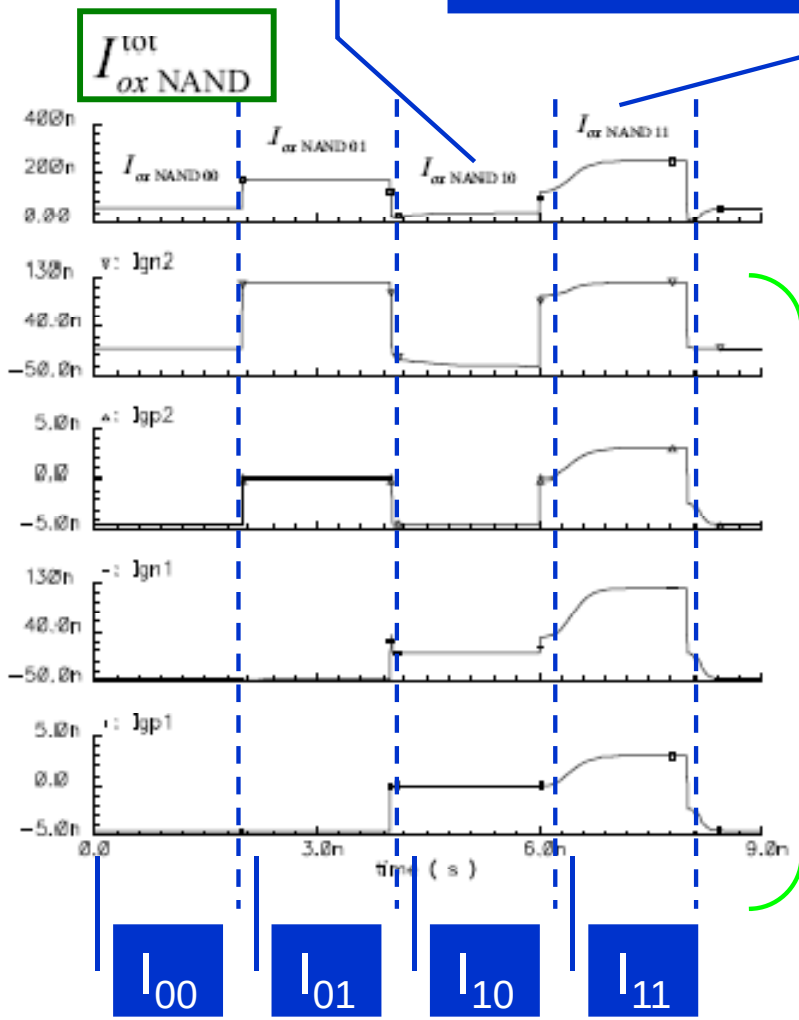
Output Voltage

Input Voltages



Best Case

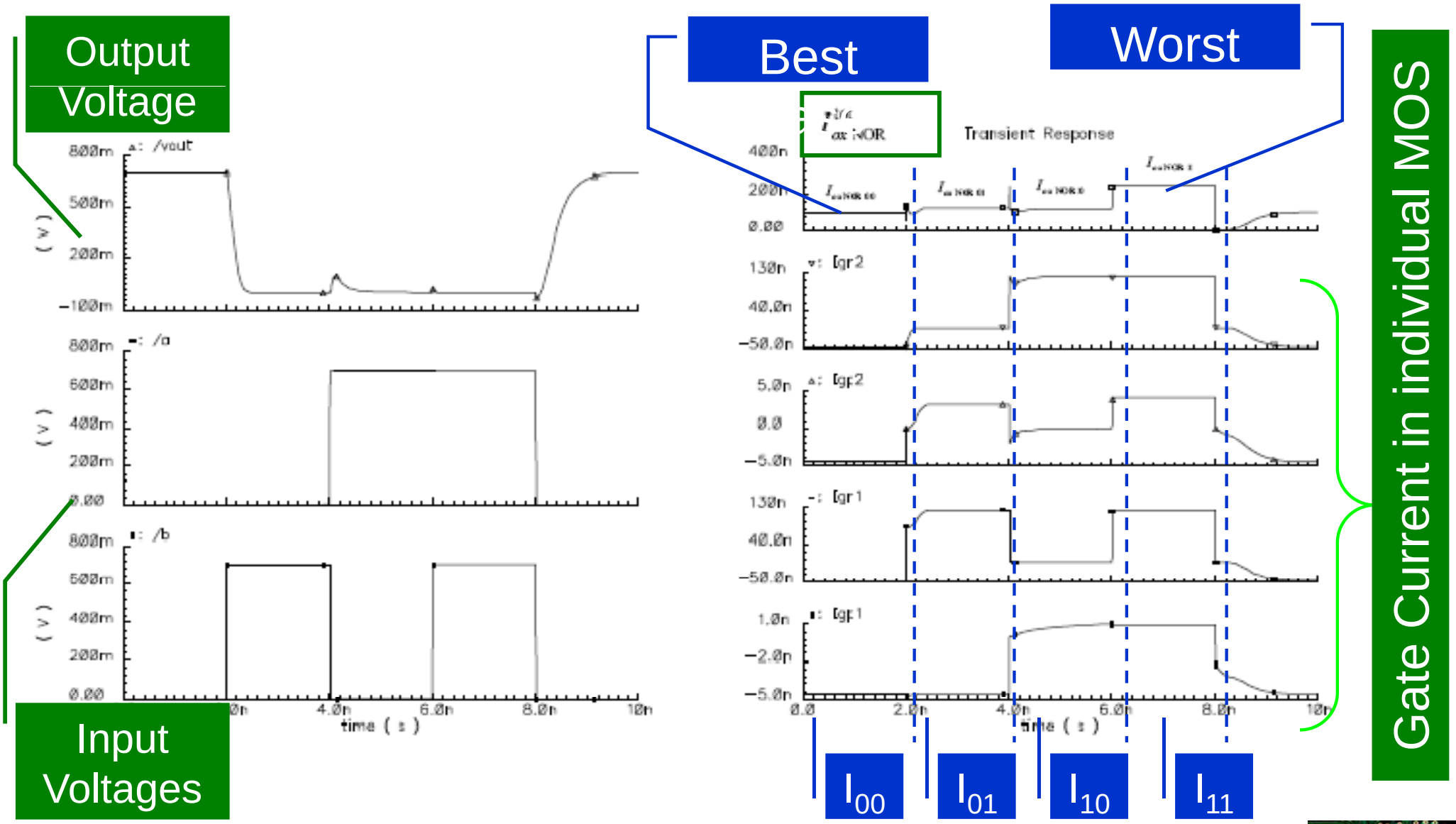
Worst



Gate Current in

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

Gate Leakage in 2-input NOR (Transient Study)



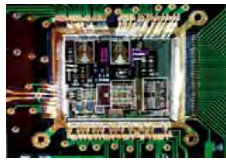
Gate Leakage in Logic Gate (Transient Study)

C_{tun} ≡ Effective tunneling capacitance at the input of a logic gate

We propose to quantify as:

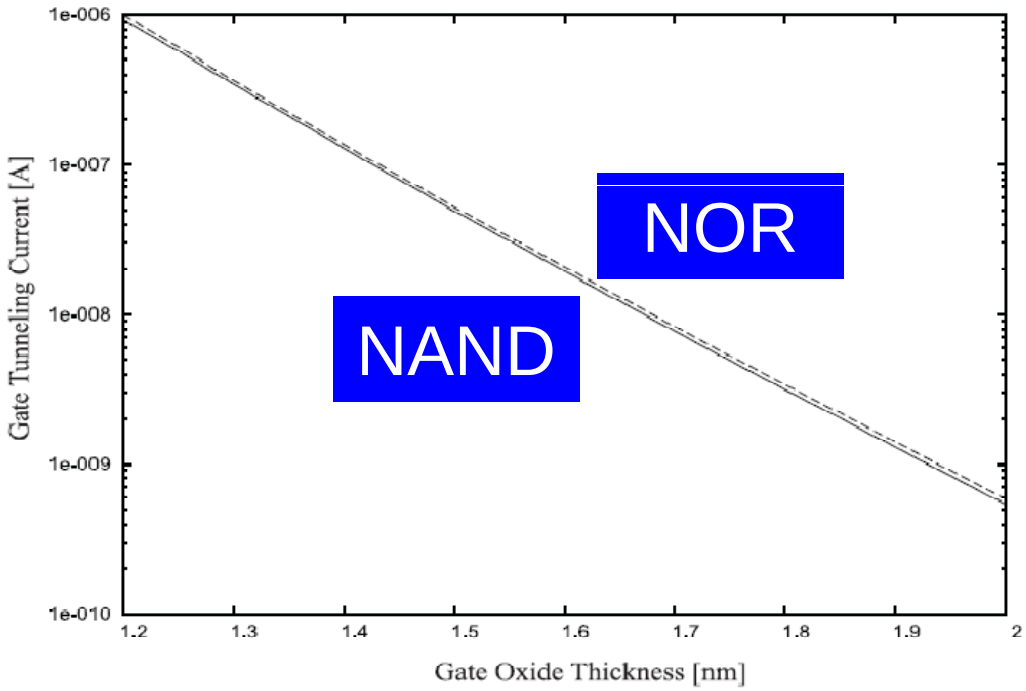
$$C^{tun} = \frac{I_{\max}^{\log ic} - I_{\min}^{\log ic}}{\left(\frac{dV_{in}}{dt}\right)}$$

$$= \frac{I_{\max}^{\log ic} - I_{\min}^{\log ic}}{V_{DD}} t_r \text{ (for equal rise/fall time)}$$

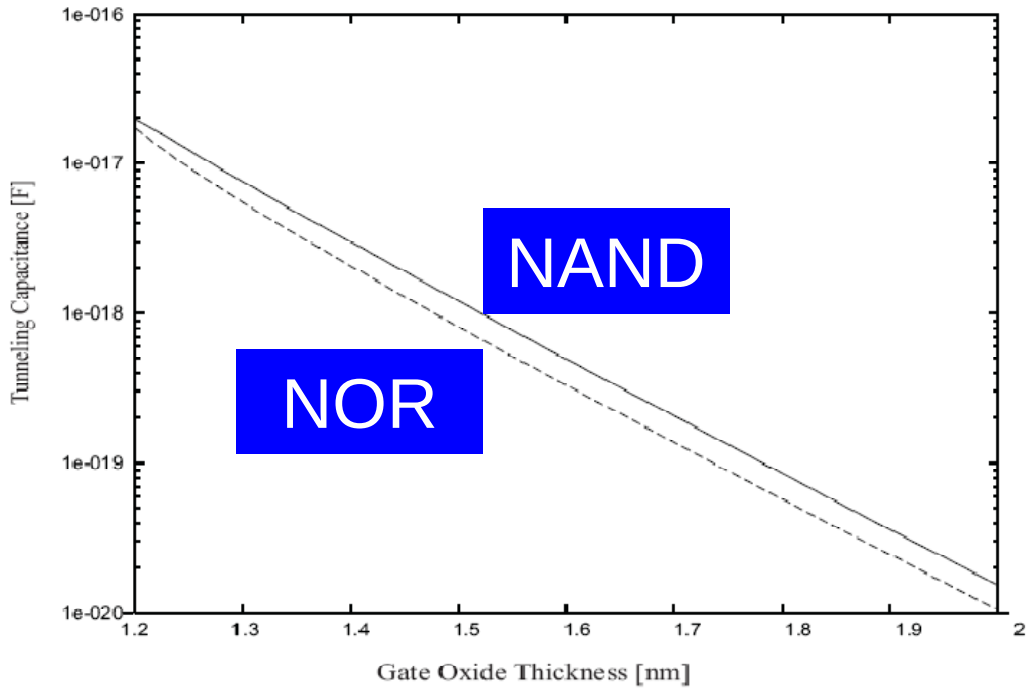


Effect of Process and Design Parameter Variation

Gate Leakage in 2-input Logic Gates (T_{ox} Variation)

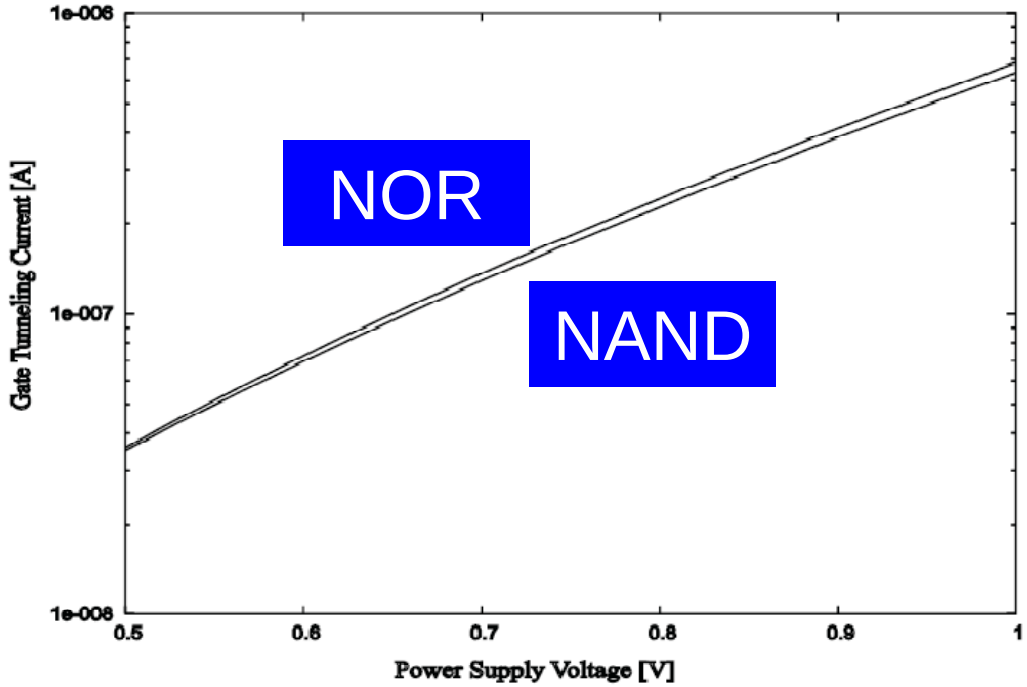


I_{tun} (logscale) versus T_{ox}

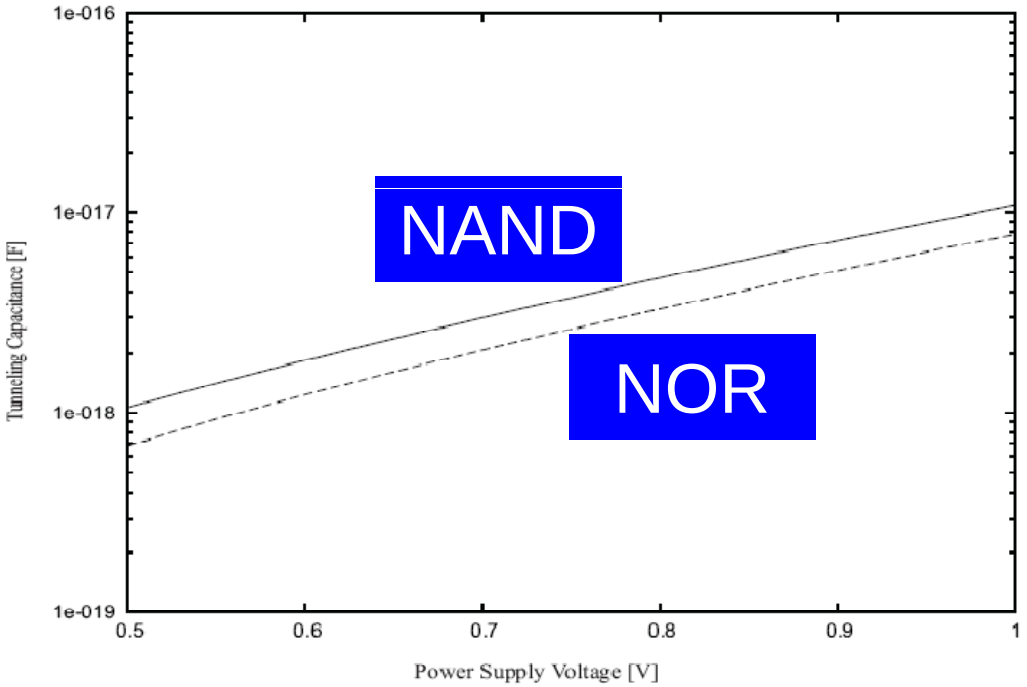


C_{tun} (logscale) versus T_{ox}

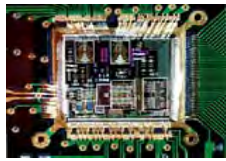
Gate Leakage in 2-input Logic Gates (V_{DD} Variation)



I_{tun} (logscale) versus V_{DD}



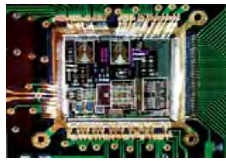
C_{tun} (logscale) versus V_{DD}



Observations

Gate Leakage in 2-input Logic Gates

- Both ON and OFF states contribute to gate leakage
- Transient effect is significant and can be captured via effective tunneling capacitance
- I_{tun} ≡ State Independent average gate leakage current of a logic gate
- C_{tun} ≡ Effective tunneling capacitance at the input of a logic gate
- I_{tun} is larger for NOR
- C_{tun} is larger for NAND



Usefulness of the Proposed Metrics

- The metrics allow designers to account for gate tunneling effect in nano-CMOS based circuit designs.
- I_{ON} and I_{OFF} - additive to static power consumption.
- C^{eff}_{tun} – additive to intrinsic gate capacitance

$$C_{logic} = C^{eff}_{tun} + C_{intrinsic}$$

- All three needs to be taken into account for effective total (switching, subthreshold, gate leakage) power optimization

Estimation of Gate Leakage

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

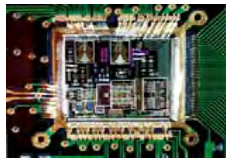
- Gate leakage estimation methods for logic level description of the circuit:
  - Pattern dependent estimation (R. M. Rao ISLPED 2003)
  - Pattern independent probabilistic estimation (R. M. Rao ISLPED 2003)

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

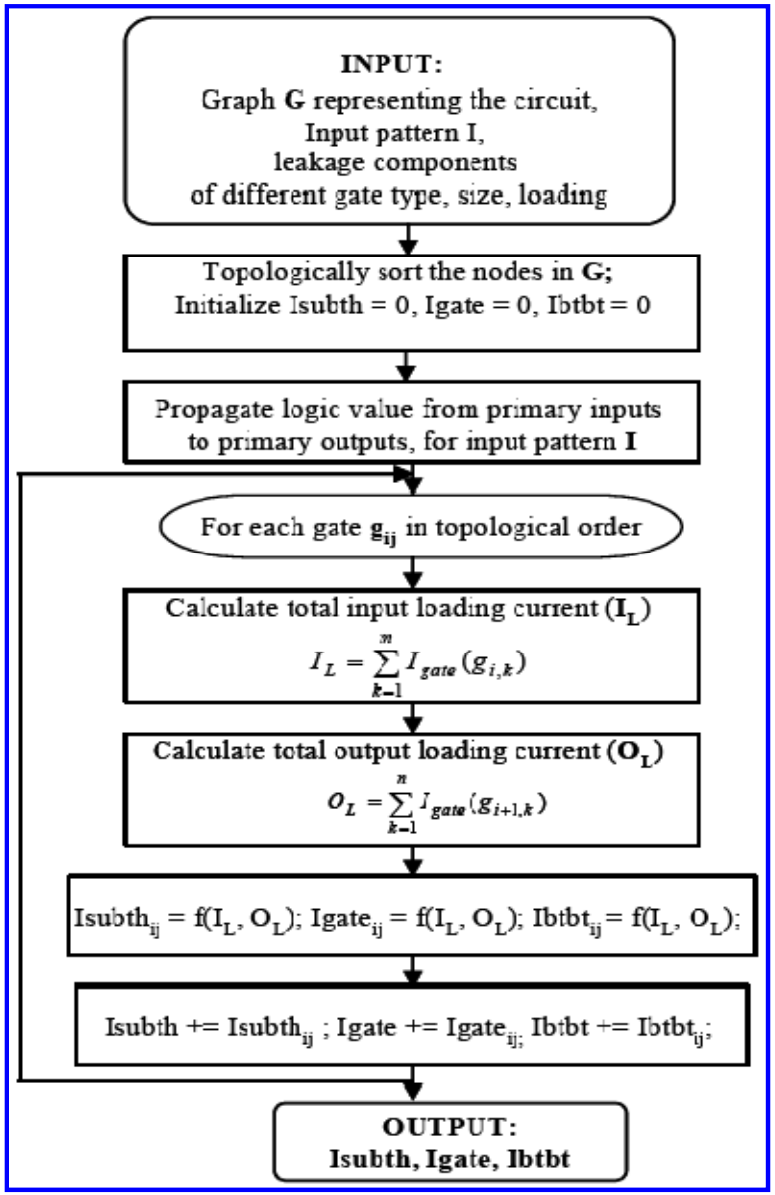
Estimation: Heuristic and Look-up Tables

- 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.

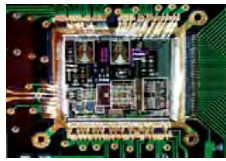
Source: Lee ISQED2003, TVLSI2003

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)

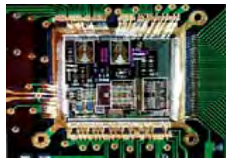


Optimization of Gate Leakage

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)

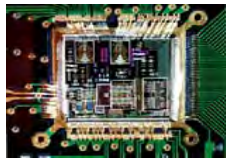


Dual T_{OX} Technique: Basis

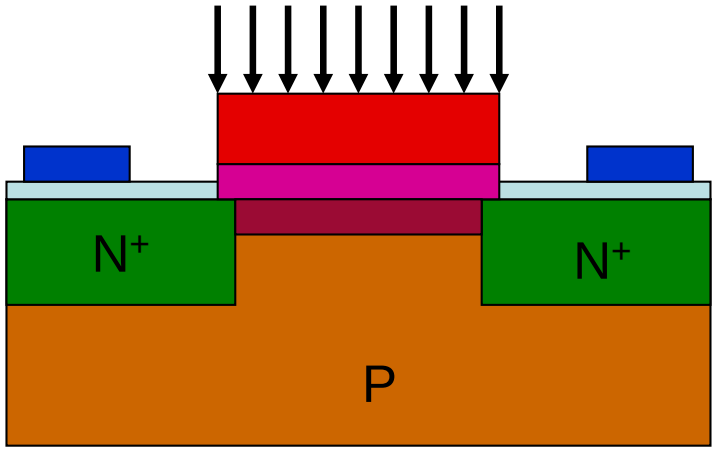
- Gate oxide tunneling current I_{oxide} (k is a experimentally derived factors):

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

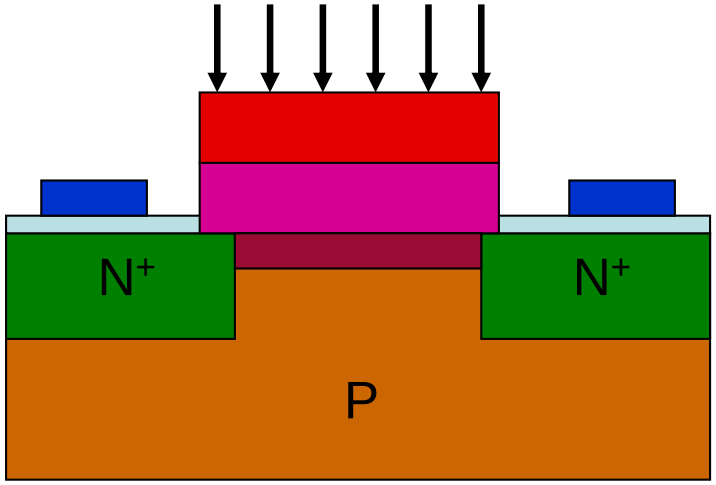
- Options for reduction of tunneling current:
  - Decreasing of supply voltage V_{dd} (*will play its role*)
  - Increasing gate SiO_2 thickness T_{oxide}



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

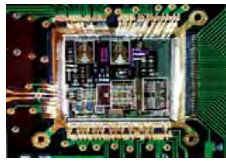
- Our approach – scale channel length (L) as well as T_{ox}; T_{ox} is almost linearly scaled with L_{eff}

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

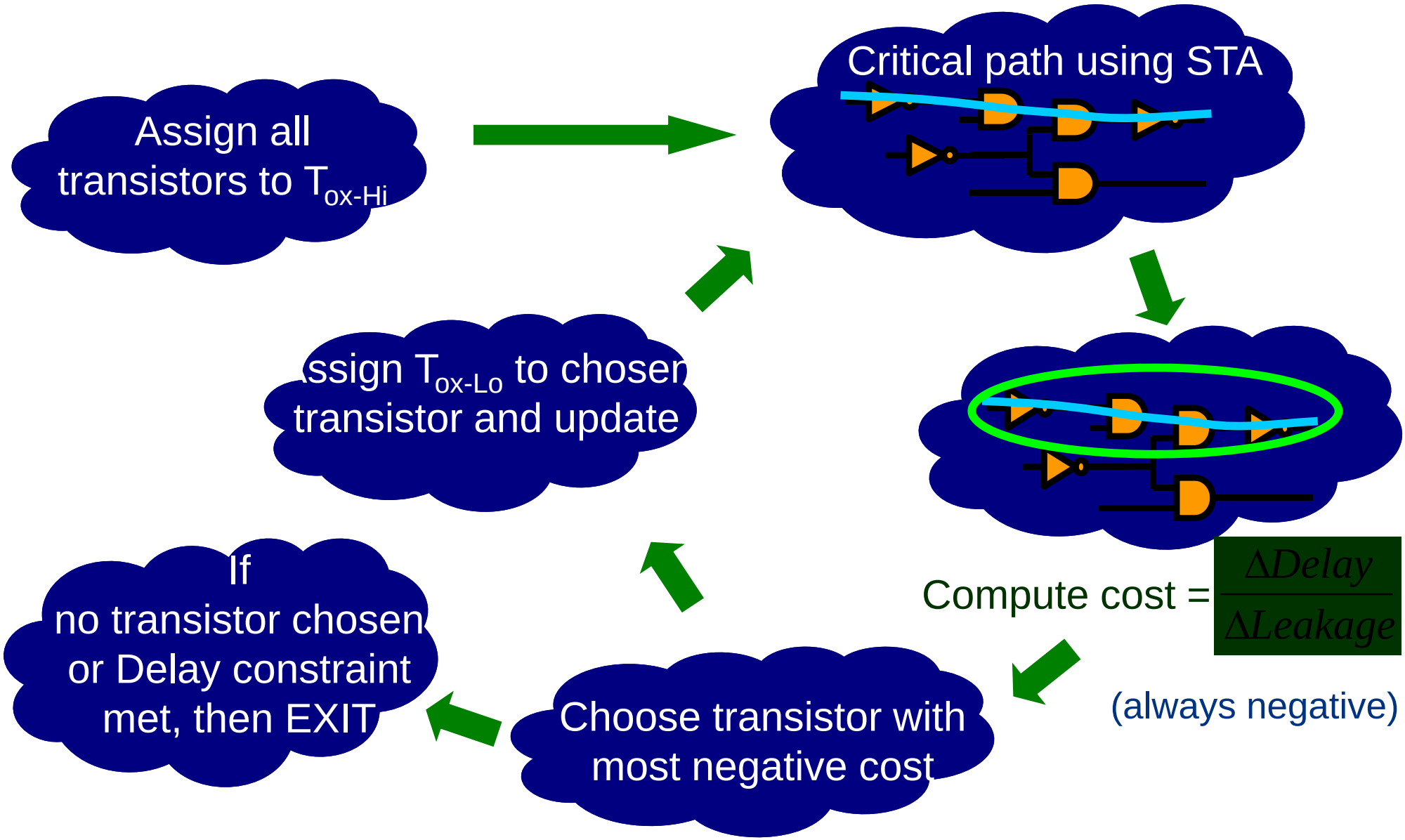
Advantages:

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

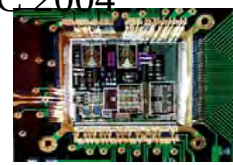
$$C_{micron} = \frac{\epsilon_{ox} L_{eff}}{T_{ox,eff}} = \text{constant}$$



Dual T_{OX} Technique: Algorithm



Source: Sultania DAC 2004

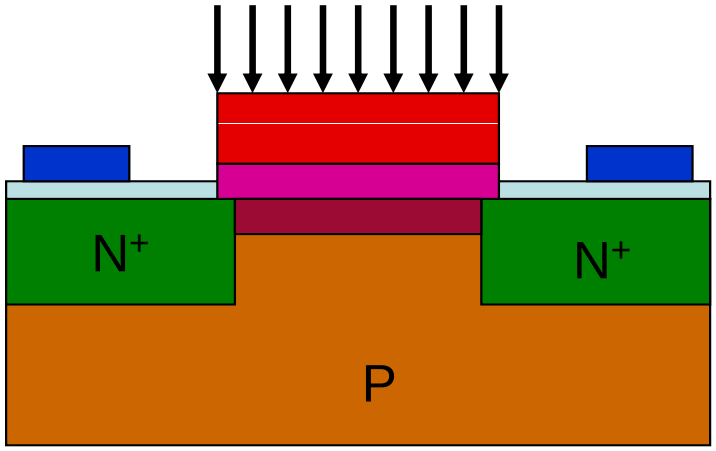


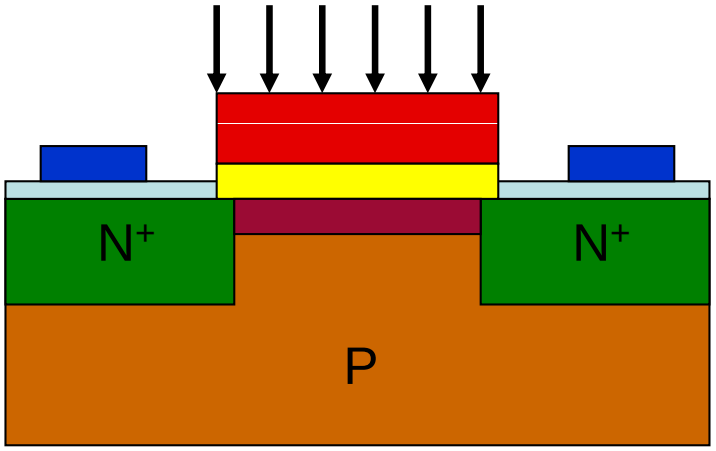
Dual T_{OX} Technique: Results

- 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

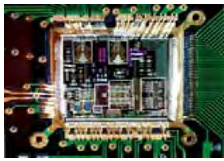
Dual K Technique: Basis



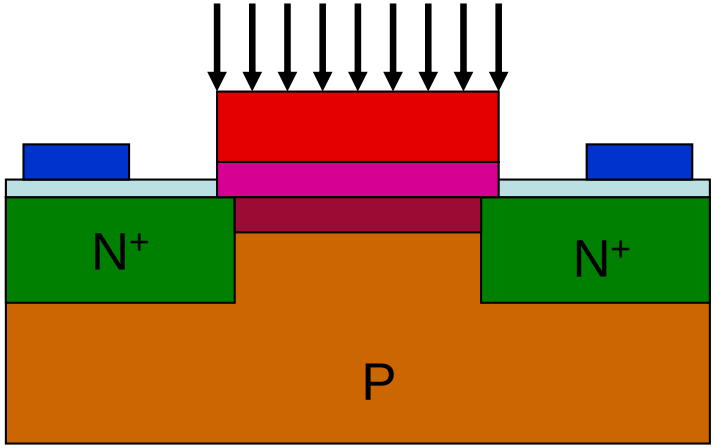


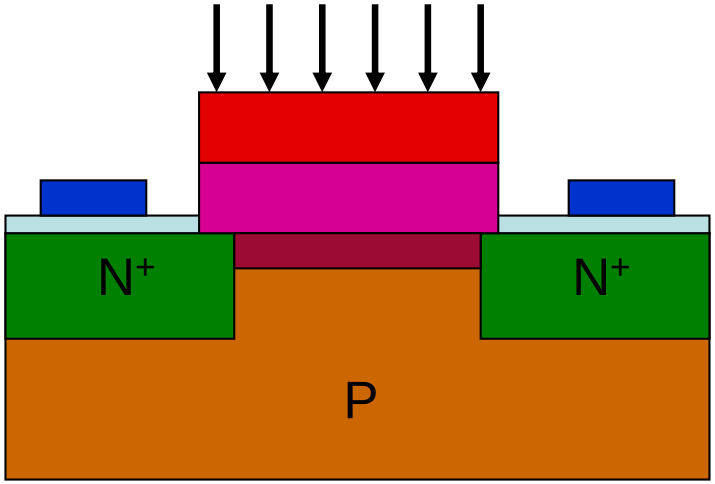
Low K_{gate} → Larger I_{gate} , Smaller delay

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



Dual K Technique: Basis

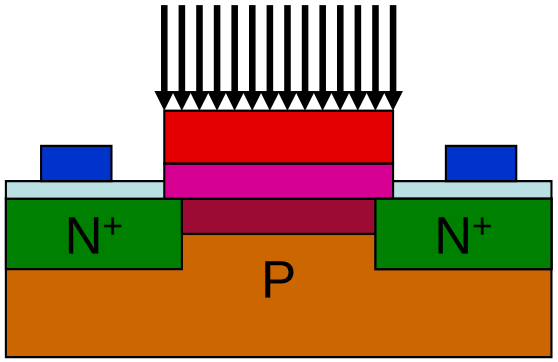




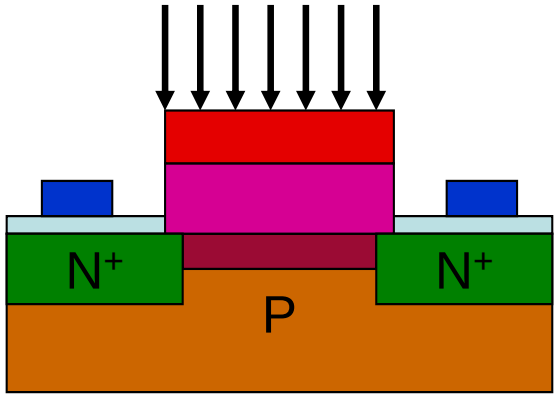
Low T_{gate} → Larger I_{gate}, Smaller delay

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

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

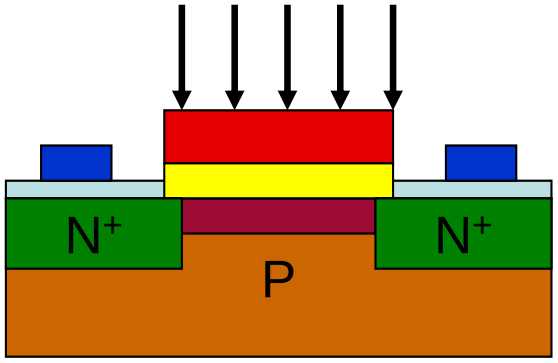


(1) K_1T_1

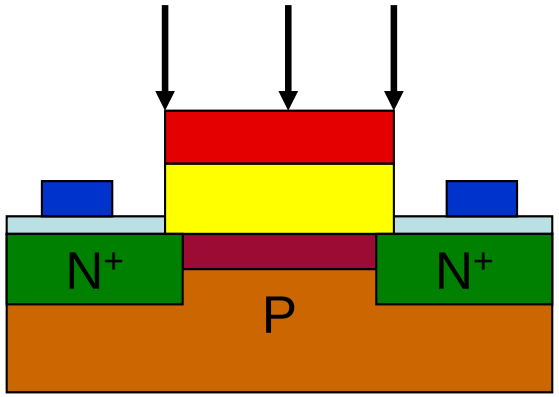


(2) K_1T_2

Tunneling Current ↓
Delay ↑



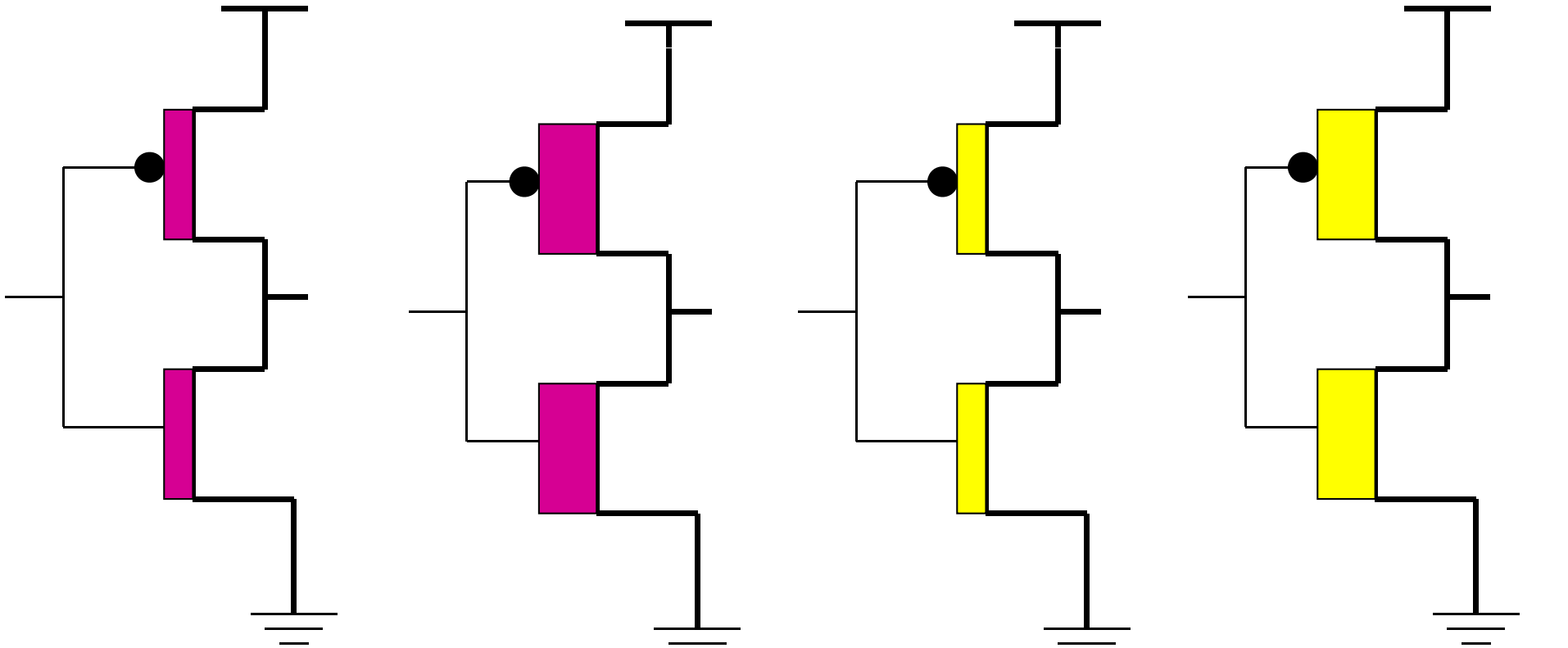
(3) K_2T_1



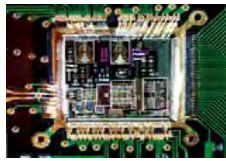
(4) K_2T_2

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}.



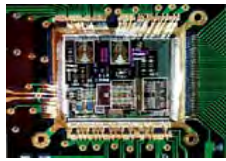
(1) K_1T_1 (2) $\overline{K_1}T_2$ (3) K_2T_1 (4) $\overline{K_2}T_2$



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.

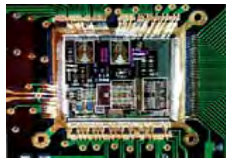
Source: Mukherjee ICCD 2005



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

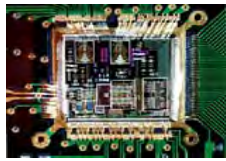
Step 5: Find the critical path delay D_{CP} for K_1T_1 assignment.

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

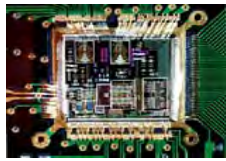


Dual K Technique: Characterization (How to Model High-K?)

- 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

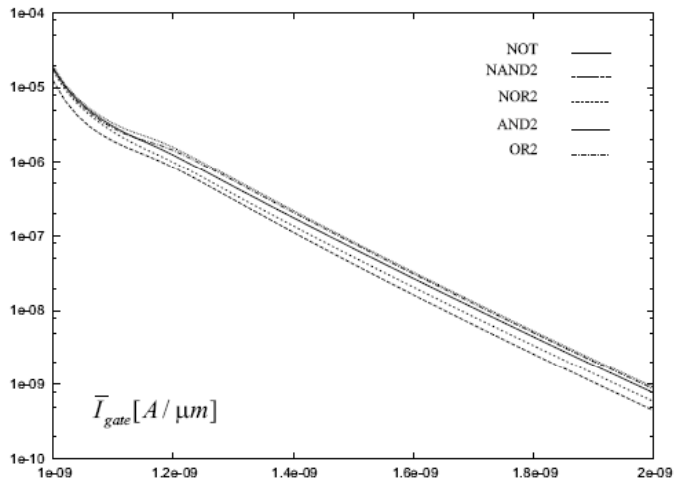
- The effect of varying oxide thickness T_{ox} was incorporated by varying TOXE in SPICE model.

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

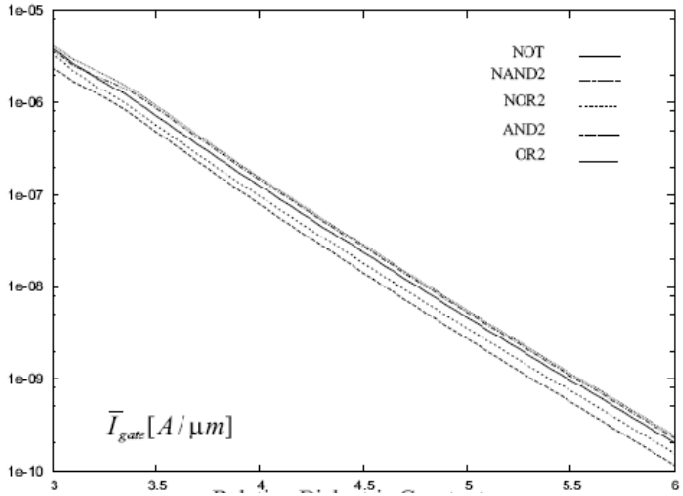
$$L^* = (T^*_{ox} / T_{ox})\, L$$

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

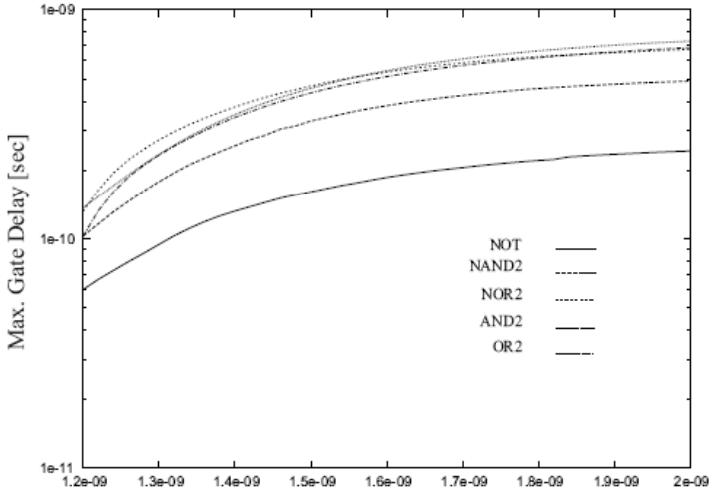
Dual K Technique: Characterization



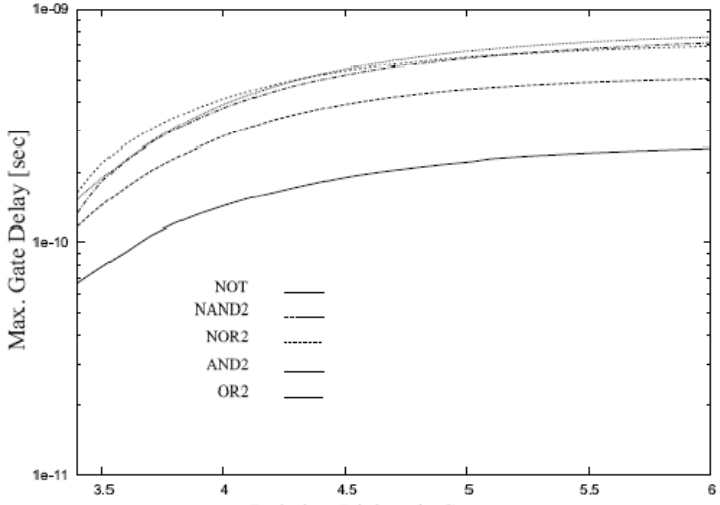
I_{gate} Vs Thickness



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.

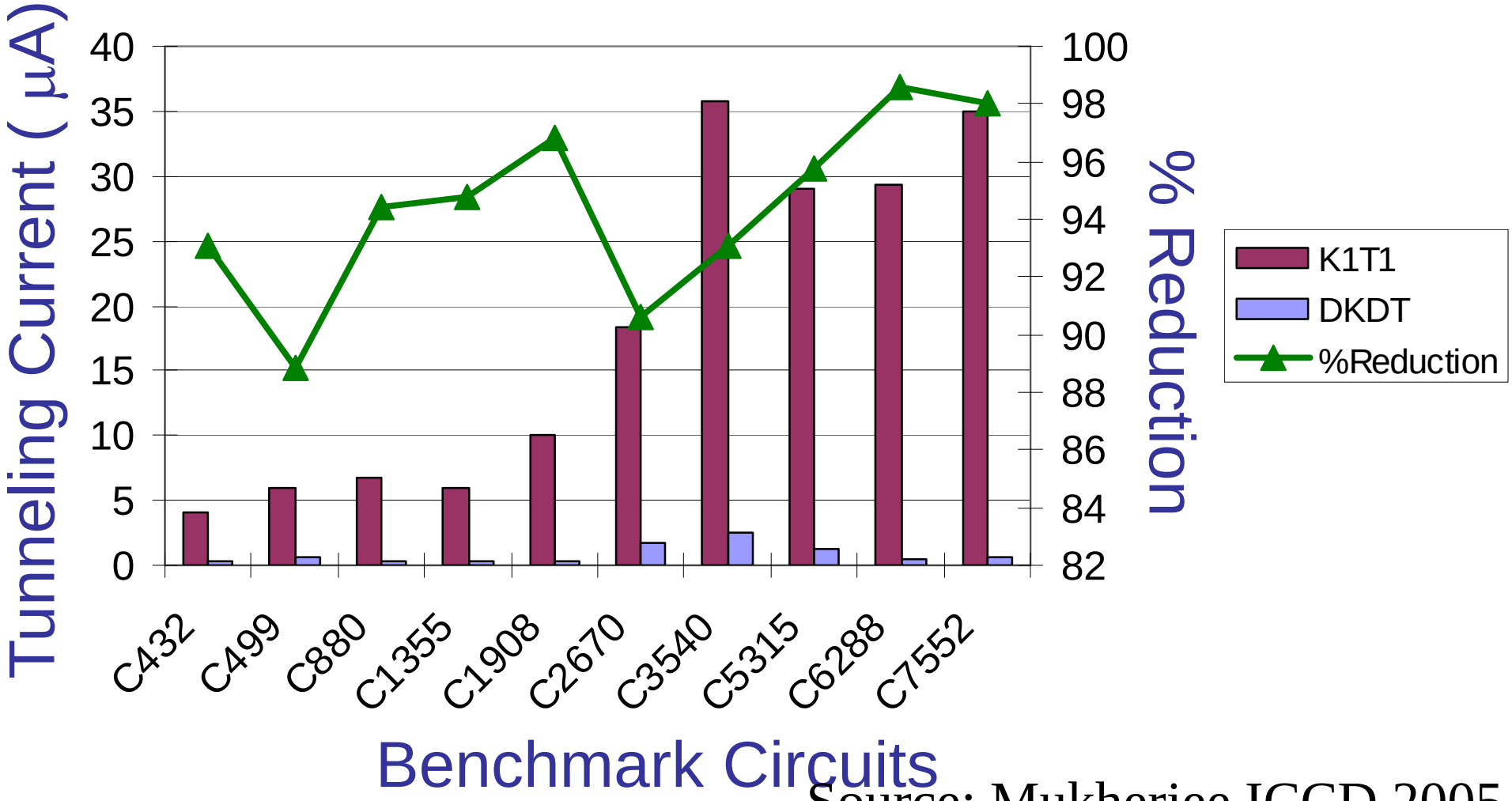
- Used K_1 = 3.9 (for SiO_2), K_2 = 5.7 (for SiON), T_1 = 1.4nm, and T_2 = 1.7nm for our experiments.

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

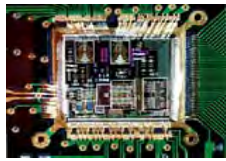
Source: Mukherjee ICCD 2005

Dual K Technique: Experimental Results

Tunneling Current and % Reduction



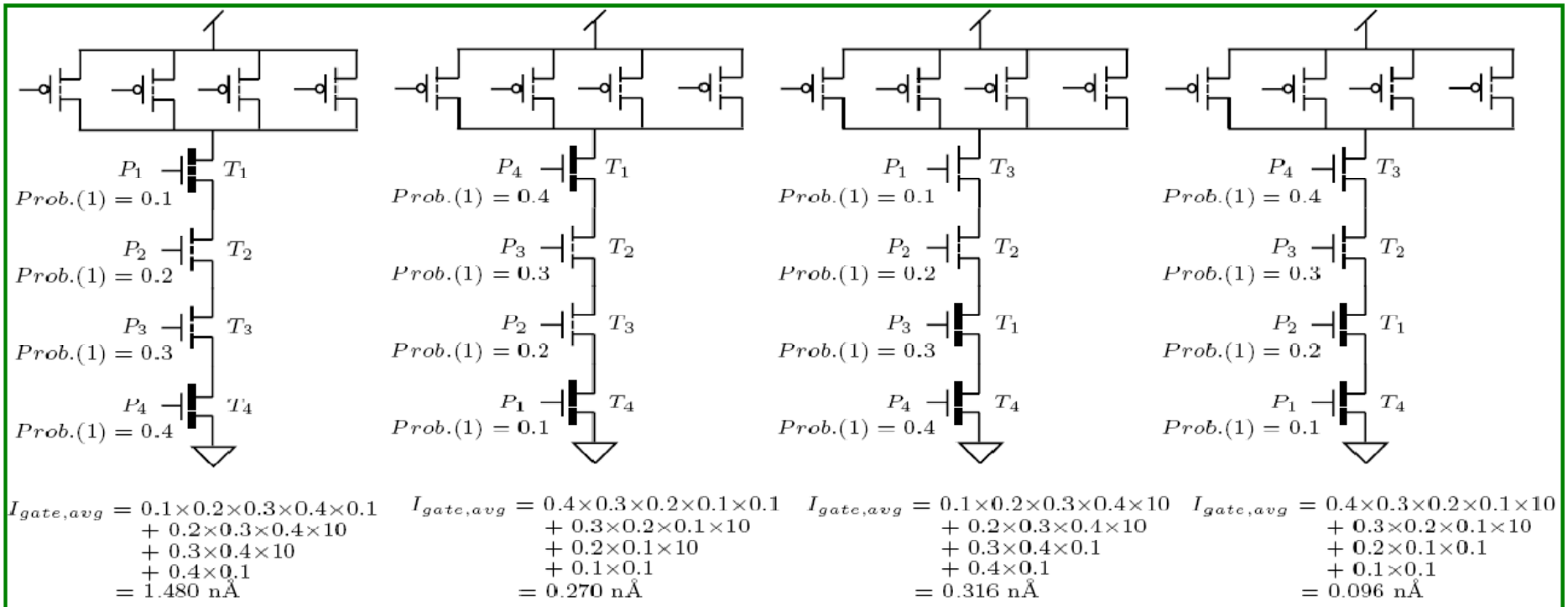
Source: Mukherjee ICCD 2005



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