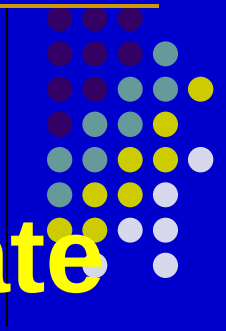


Modeling and Reduction of Gate Leakage during Behavioral Synthesis of NanoCMOS Circuits

Saraju P. Mohanty and Elias Kougianos

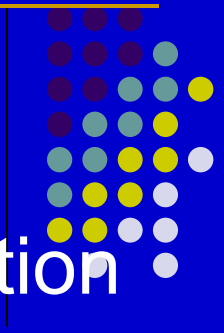
VLSI Design and CAD Laboratory

Homepage: http://www.vdcl.cse.unt.edu

University of North Texas, Denton, TX, USA.

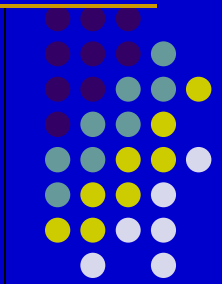
Email: smohanty@cs.unt.edu

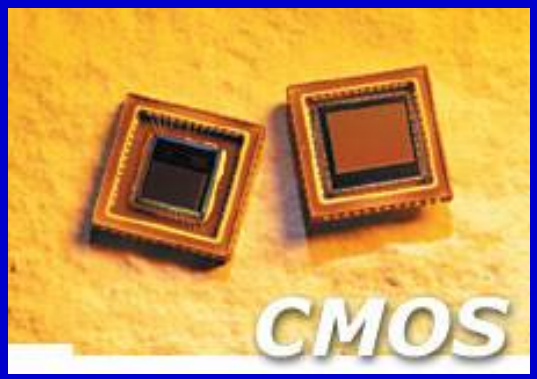
Outline of the Talk

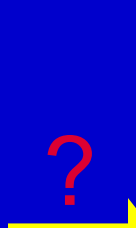


- CMOS scaling Trends and Power Dissipation redistribution
- Related Work
- First Principles Analytical Modelling of Gate Leakage
- Datapath Component Library
- Datapath Scheduling for Gate leakage reduction
- Experimental Results
- Conclusions and Future Directions

CMOS Driven Applications

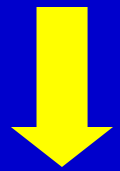






?

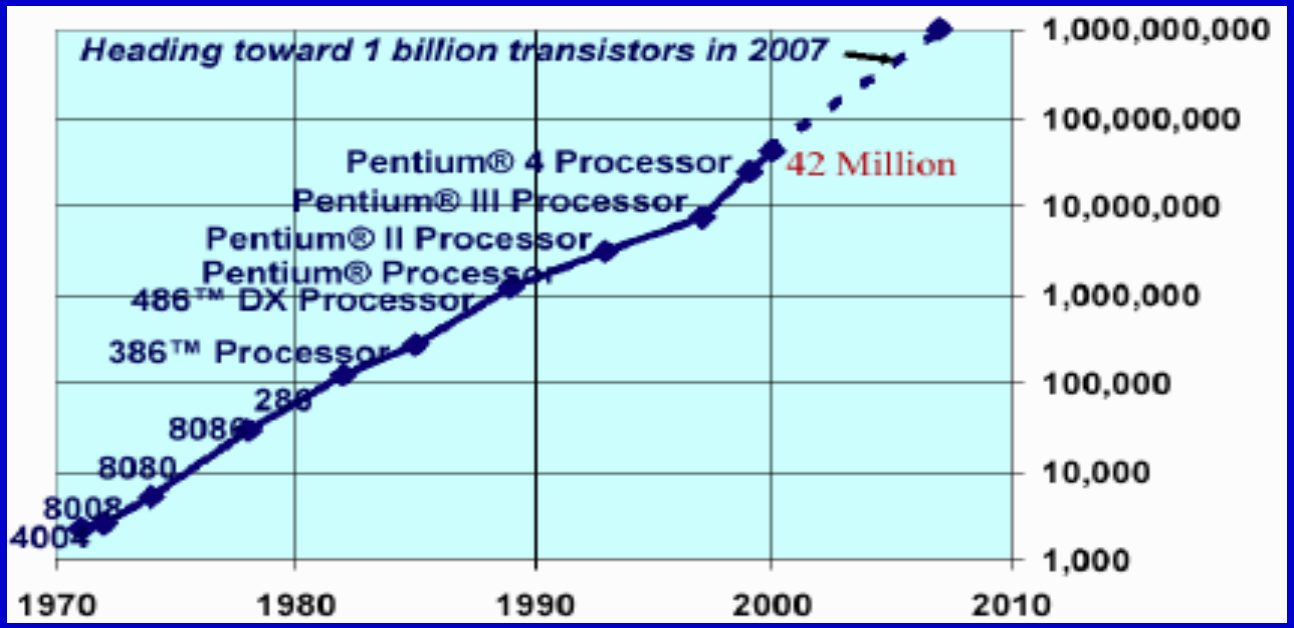
Energy costs,
Battery life,
Cooling costs



Low Power
Synthesis

Almost the entire industry today is driven by CMOS

Scaling Trend – Transistor Count



Increase in Transistor Count per chip

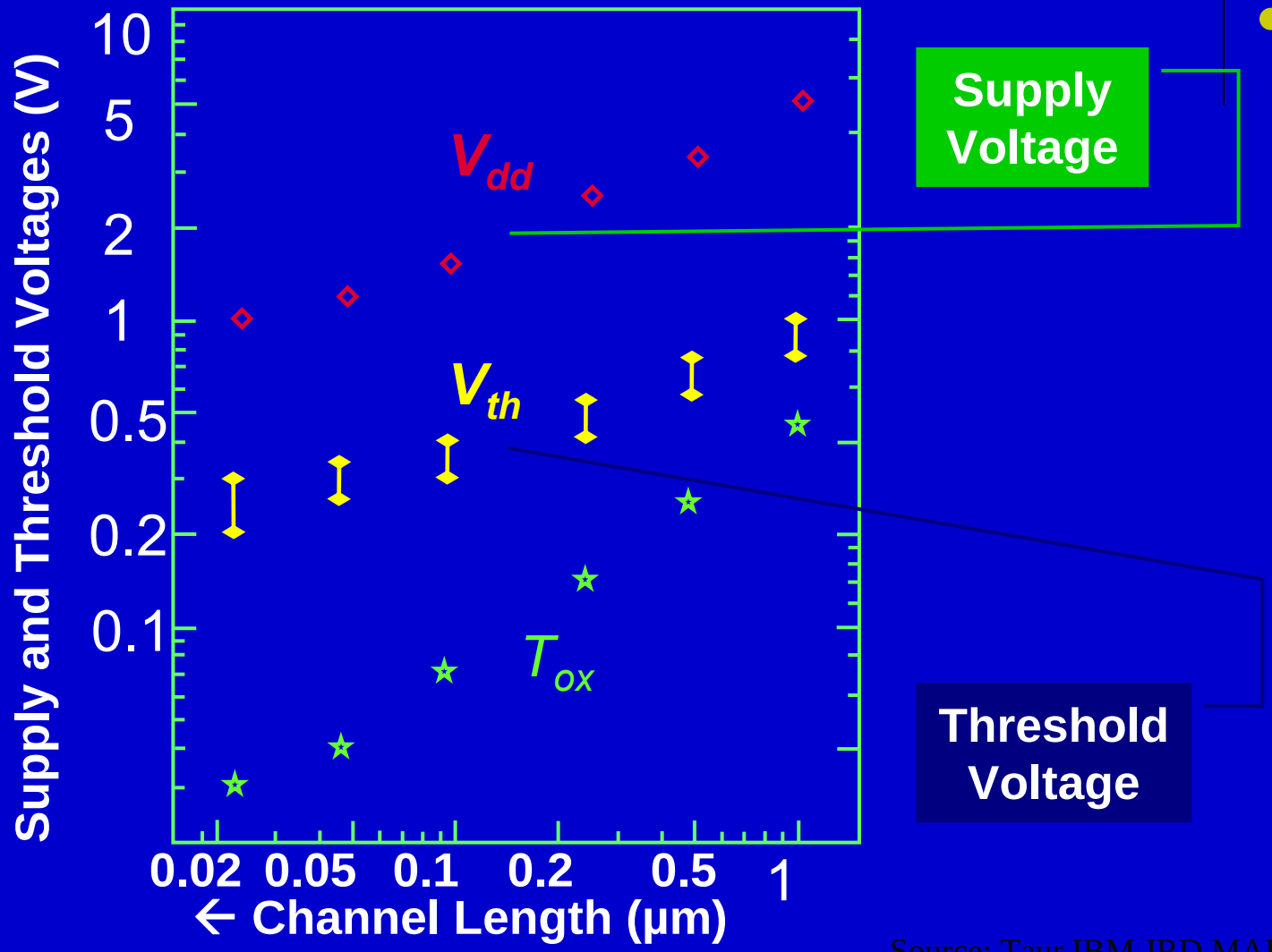
Operating frequency and throughput have increased.

1967

VLSI technology is the fastest growing technology in human history.

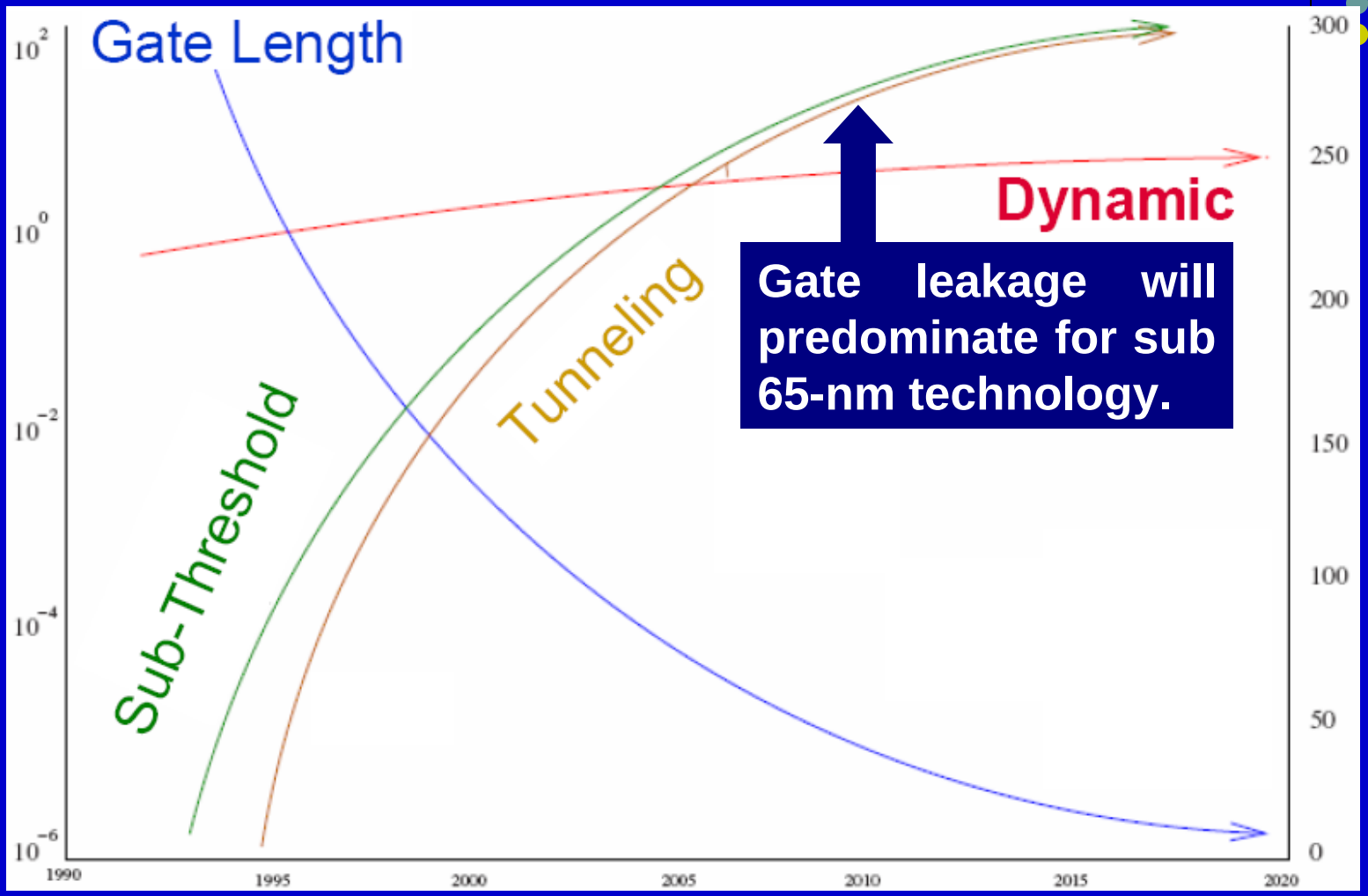
2007

What is Physically Scaled ?



Source: Taur IBM JRD MAR 2002

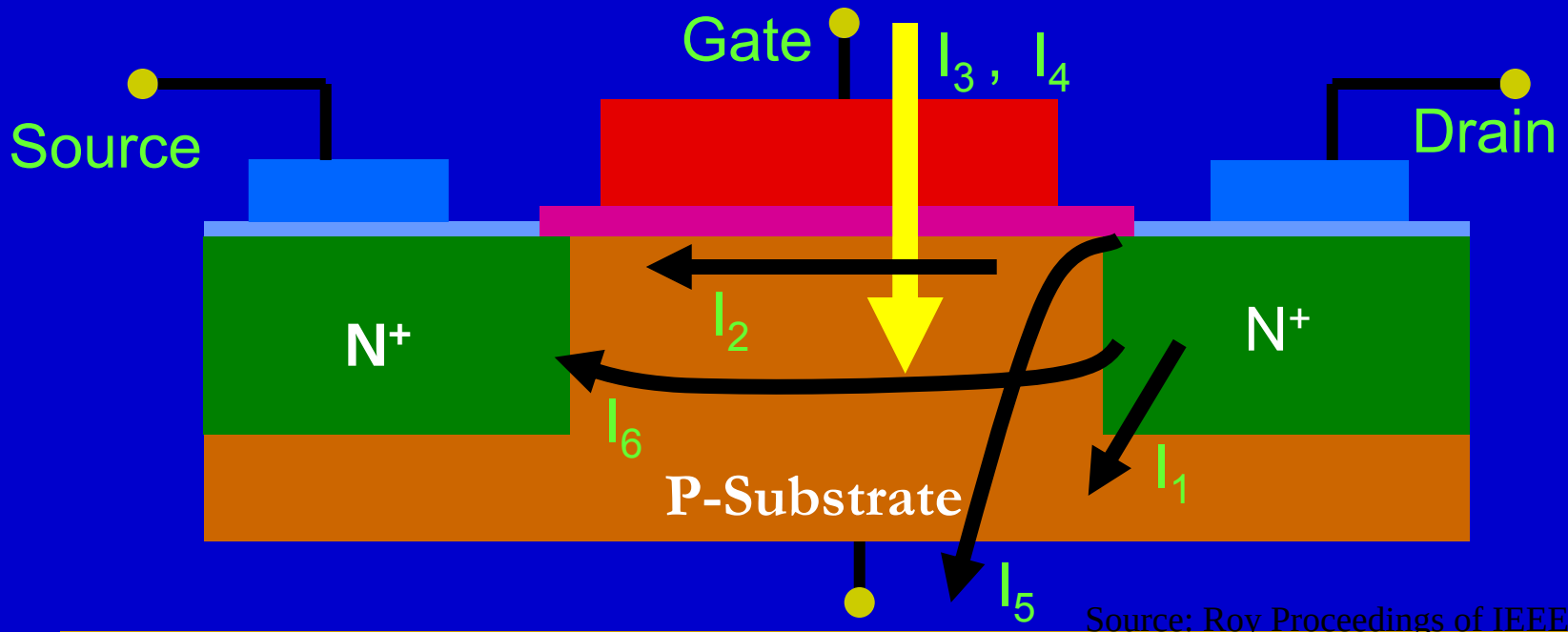
Power Dissipation : Redistribution



Source: Hansen Thesis 2004

Leakages in Nanometer CMOS

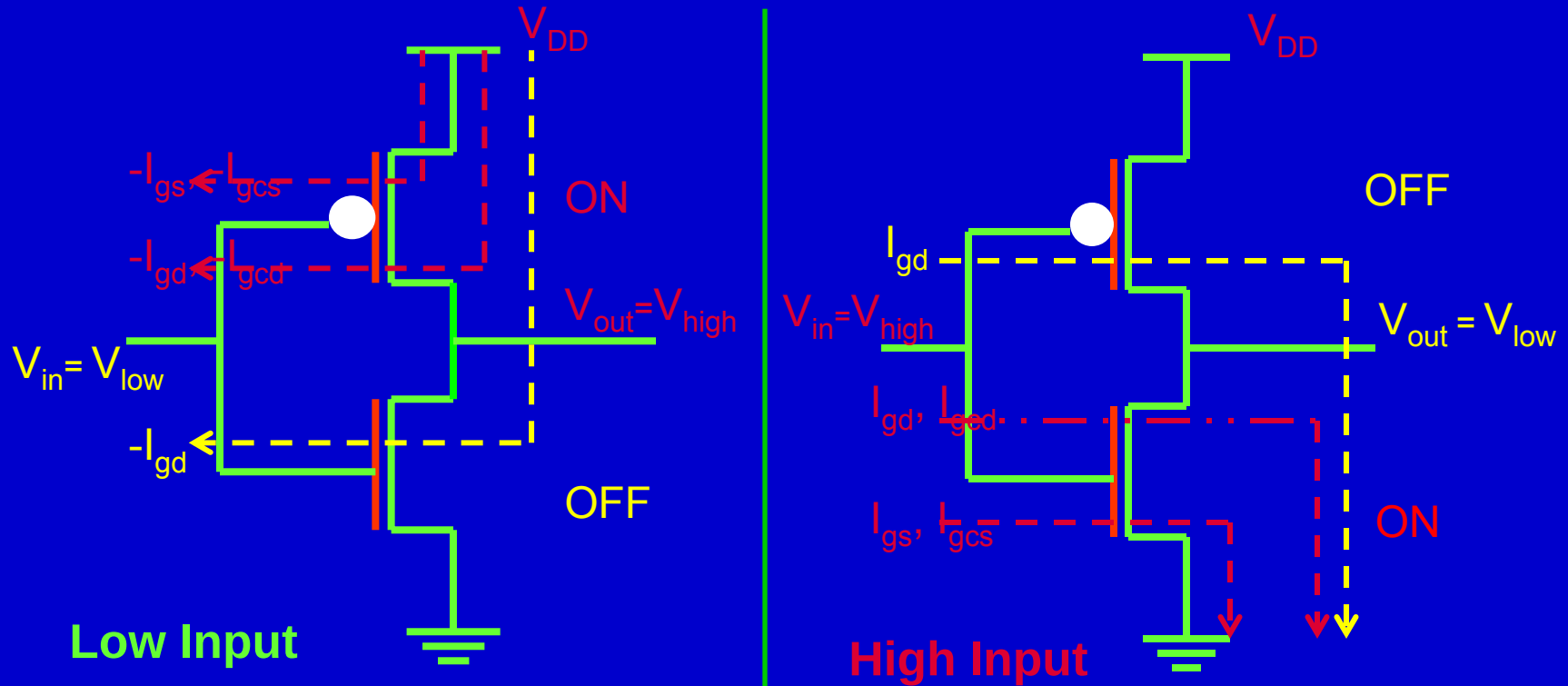
I_1 : reverse bias pn junction (both ON & OFF)
I_2 : subthreshold leakage (OFF)
I_3 : oxide tunneling current (both ON & OFF)
I_4 : gate current due to hot carrier injection (both ON & OFF)
I_5 : gate induced drain leakage (OFF)
I_6 : channel punch through current (OFF)



Source: Roy Proceedings of IEEE Feb2003

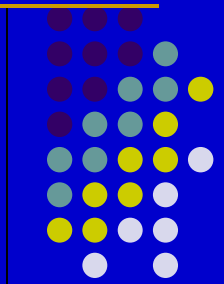
Tunneling Paths in an Inverter

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



NOTE: Gate to body component found to be negligible.

Related Works



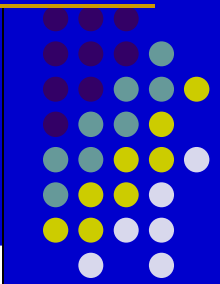
Behavioral Level Subthreshold:

- **Khouri, TVLSI 2002 :** Algorithms for subthreshold leakage power analysis and reduction using dual threshold voltage.
- **Gopalakrishnan, ICCD2003:** MTCMOS approach for reduction of subthreshold current

Logic / Transistor Level Tunneling:

- **Lee, TVLSI2004 :** Pin reordering to minimize gate leakage during standby positions of NOR and NAND gates.
- **Sultania, DAC2004 :** Heuristic for dual T_{ox} assignment for tunneling current and delay tradeoff.
- **Sirisantana, IEEE DTC Jan-Feb 2004:** Use multiple channel lengths and multiple gate oxide thickness for reduction of leakage.

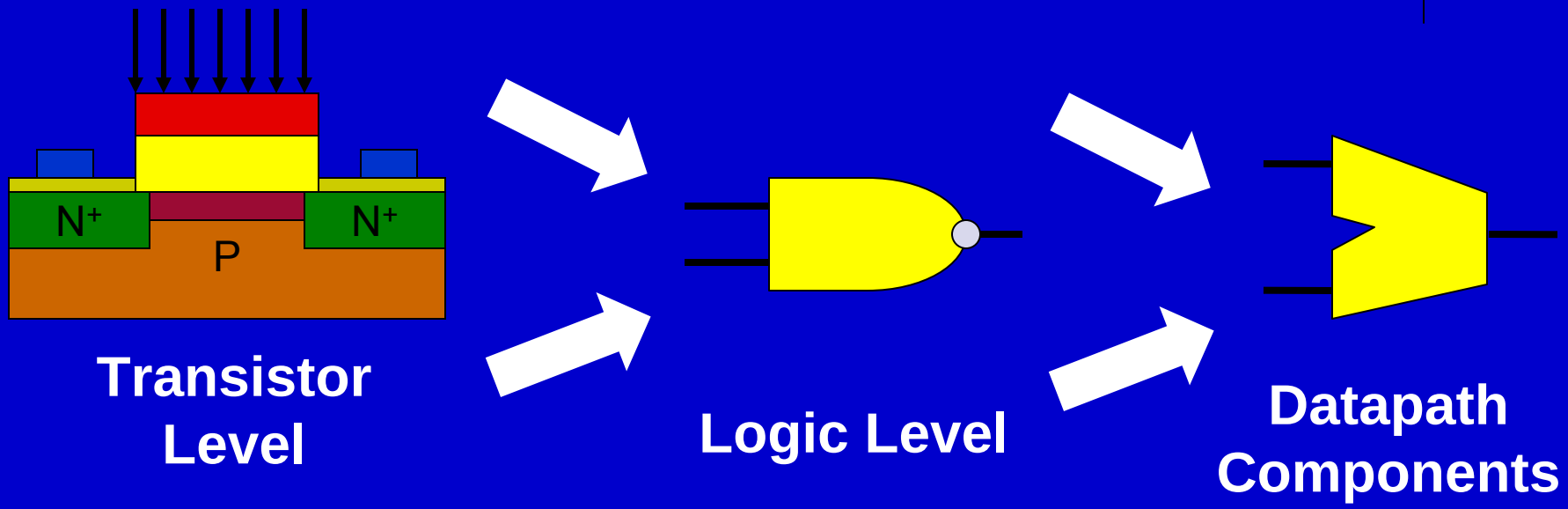
Contributions of Our Paper



Contributions of this paper are two fold.

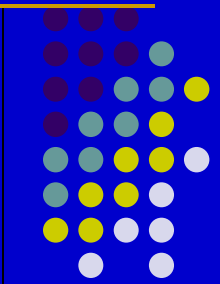
- First principle models for gate leakage (direct tunneling current) and propagation delay calculations of functional units.
- Algorithm for scheduling of the datapath operations such that overall gate leakage of a of a datapath circuit is reduced with minimal delay penalty.

Model for Tunneling Current (3 Level Hierarchy)



We observed that NAND gate has minimum leakage compared to all basic logic gates. Therefore we constructed datapath components using NAND.

Model for Tunneling Current (Assumption)



- We assumed that resources such as adders, subtractors, multipliers, dividers, are constructed using 2-input NAND.
- There are total n_{total} NAND gates in the network of NAND gates constituting a n-bit functional unit.
- n_{cp} number of NAND gates are in the critical path.

Model for Tunneling Current (Logic Level State Dependent)

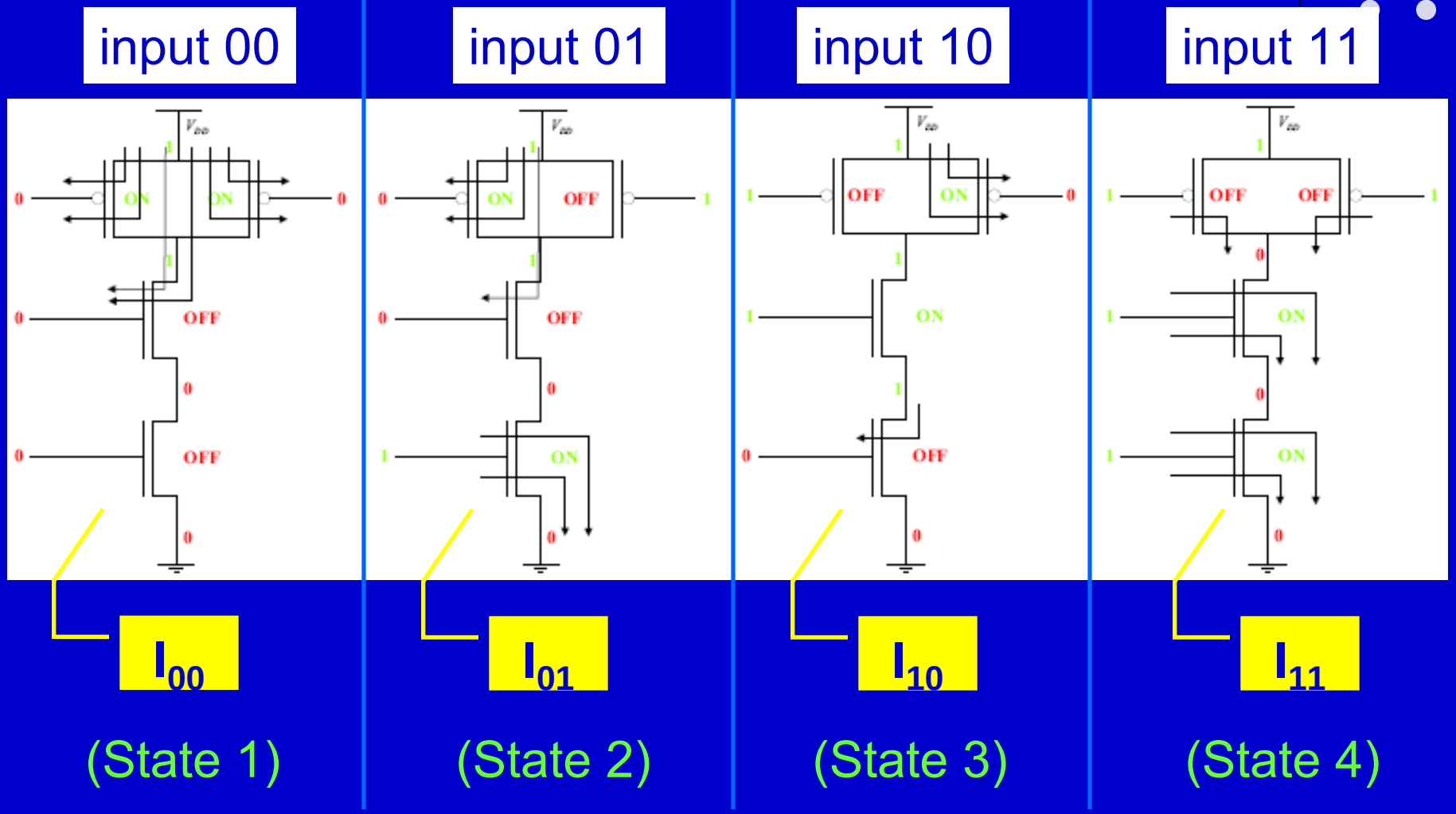
- Represent a datapath component circuit as a graph: vertex → NAND gate, edge → dependency.
- Average gate leakage of the circuit is the probabilistic mean of the gate leakage of the circuit:

$$I_{ox,FU} = E\left(\sum_{\forall\ NAND} I_{ox,NAND}(s(i))\right)$$

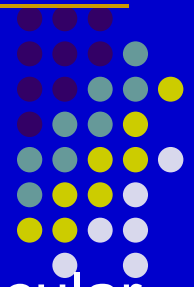
$$= \sum_{\forall NAND} \left(\sum_{j} I_{ox,\ NAND}(s(j)) * P(j)\right)$$

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

Model for Tunneling Current (Gate Leakage in 2-input NAND)



Model for Tunneling Current (Transistor Level)



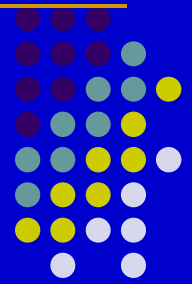
- The tunneling current for a NAND gate for a particular state is calculated as: $$I_{ox,NAND} = \sum_{MOSi \in NAND} I_{ox,i}$$
- The direct tunneling current of a MOS:

$$I_{DT} = \frac{WLq^3V_{ox}^2}{16\pi^2\phi_B T_{ox}^2} \exp\left[-\frac{4\sqrt{2m_{eff}}\,\phi_B^{1.5}T_{ox}}{3\hbar qV_{ox}}\left\{1-\left(1-\frac{V_{ox}}{\phi_B}\right)^{1.5}\right\}\right]$$

- The voltage across the MOS gate dielectric V_{ox} is expressed as follows: $V_{ox} = V_{gs} - V_{fb} - \Phi_S - V_{poly}$

NOTE: T_{ox} is calculated from physical oxide thickness T_{oxp} considering polysilicon depletion.

Model for Tunneling Current (Transistor Level)



- Voltage across polysilicon is: $V_{poly} = \frac{\varepsilon_{OX}^2 V_{OX}^2}{2q\varepsilon_{Si} N_{poly} T_{OX}^2}$
- We Arrange the above two equations to obtain a quadratic equation for Vox. By solving the quadratic equation:

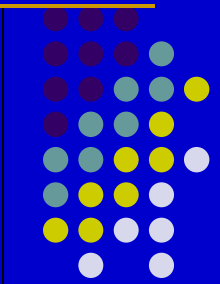
$$V_{OX} = \frac{\sqrt{1 - 2(V_{fb} + \psi_S - V_{gs})(\frac{\varepsilon_{OX}^2}{q\varepsilon_{Si} N_{poly} T_{OX}^2})} - 1}{(\frac{\varepsilon_{OX}^2}{q\varepsilon_{Si} N_{poly} T_{OX}^2})}$$

- The flat-band voltage V_{fb} can be obtained using the expression :

$$V_{fb} = \frac{q N_{channel} T_{OX}^2}{2 \in Si}$$

where, Ψ_S = 2 * Fermi-Level, for strong inversion.

Model for Propagation Delay



- The critical path delay of a *n*-bit functional unit using the NAND gates as building blocks :

$$T_{pd,FU} = \sum_{i=1 \rightarrow n_{cp}} 0.5(n_{fan\text{-}in} T_{pd,NMOS} + T_{pd,PMOS})$$

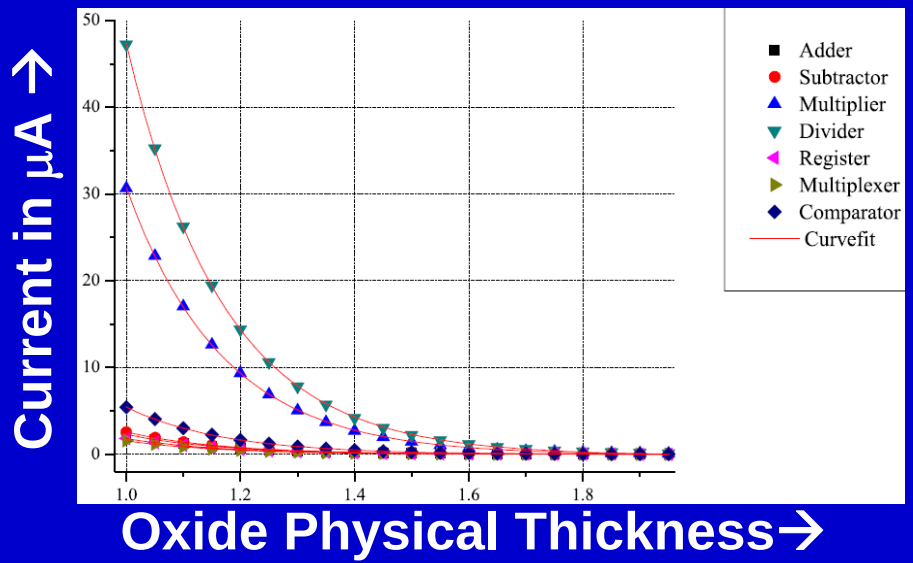
where, $n_{fan\text{-}in}$ is the effective fan-in factor.

- Using the physical-alpha-power model the delay of a MOS, where I_{DSat0} is the saturation drain current of the MOS for $V_{gs} = V_{dd}$.

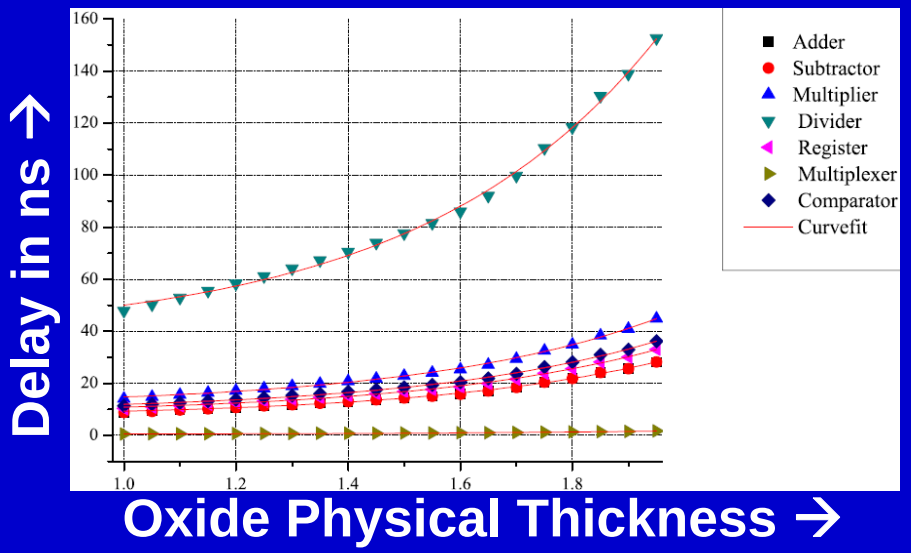
$$T_{pd,MOS} = \frac{0.5 C_L V_{dd}}{I_{DSat0}} + T_T \left\{ \frac{0.5 - \left(\frac{V_{dd} - V_{Th}}{V_{dd}} \right)}{\alpha + 1} \right\}$$

Cell Characterization : 65nm Tech (Components are of 16-bit size)

$$I_{ox,FU}(T_{ox}) = a\exp(-T_{ox}/\alpha) + b$$



$$T_{pd,FU}(T_{ox}) = c\exp(T_{ox}/\beta) + d$$

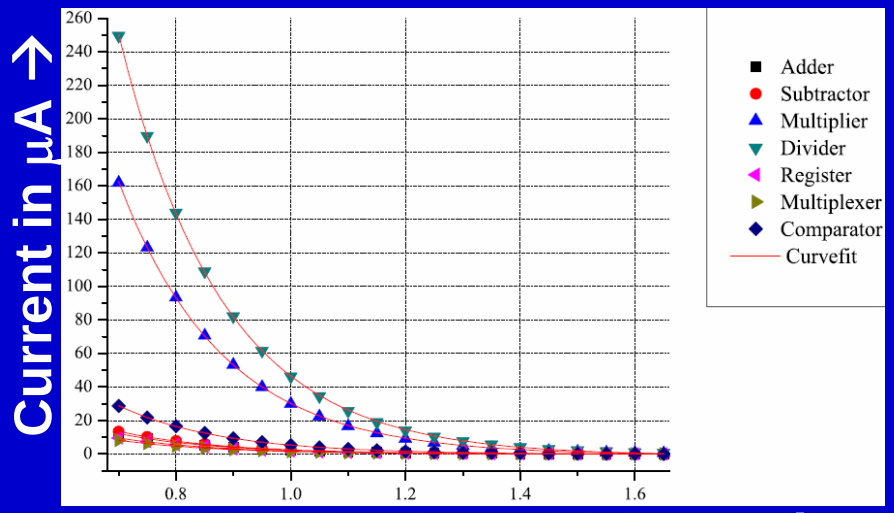


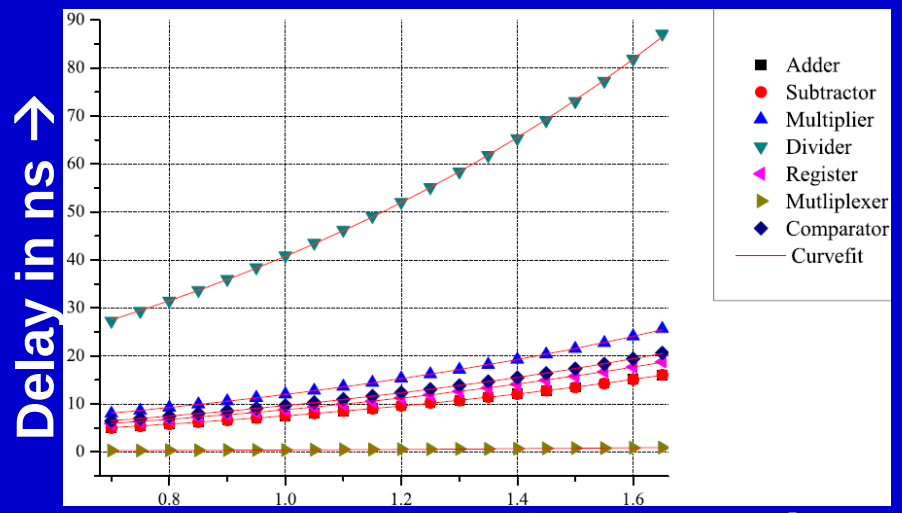
For a given length *L*, the width of the transistors is chosen as $W_{NMOS} = 4L$, $W_{PMOS} = 8L$. While changing the oxide thickness the channel length of the transistor is changed proportionately. Nominal L = 65nm, Nominal oxide physical thickness = 1.0nm

Cell Characterization : 45nm Tech (Components are of 16-bit size)

$$I_{ox,FU}(T_{ox}) = a\exp(-T_{ox}/\alpha) + b$$

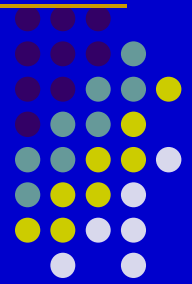
$$T_{pd,FU}(T_{ox}) = c\exp(T_{ox}/\beta) + d$$





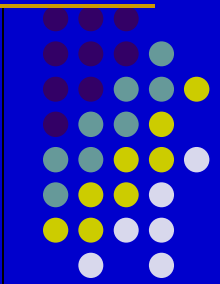
For a given length *L*, the width of the transistors is chosen as $W_{NMOS} = 4L$, $W_{PMOS} = 8L$. While changing the oxide thickness the channel length of the transistor is changed proportionately. Nominal L = 45nm, Nominal oxide physical thickness = 0.7nm

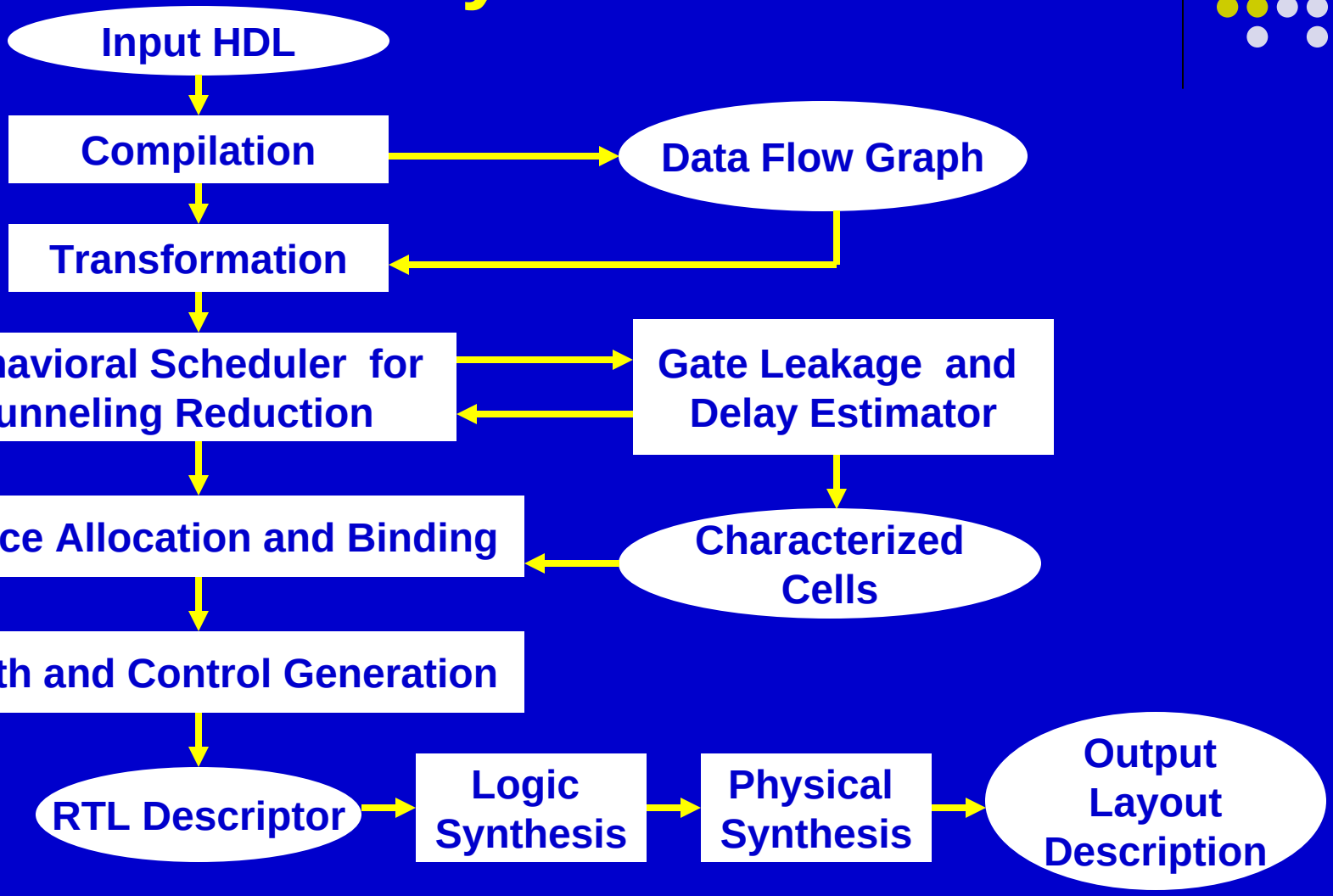
Dual-T_{ox} : Basis



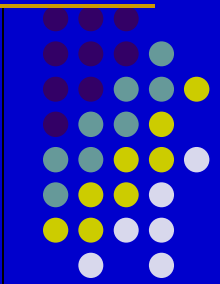
- Gate leakage : $I_{ox} \propto exp\ (-T_{ox} / \alpha)$
  - **I_{ox} decreases as T_{ox} increases**
- Propagation delay : $T_{pd} \propto exp\ (T_{ox} / \beta)$
  - **T_{pd} increases as T_{ox} increases**
- Thus combined use of high-T_{ox} resources and low-T_{ox} resources can reduce the gate leakage of a datapath with little compromise in circuit performance.

Dual-T_{ox} Based Behavioral Synthesis





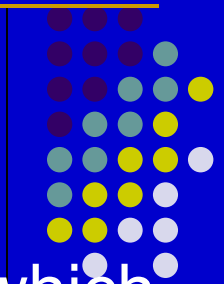
Dual-T_{OX} Assignment Algorithm (Basis)



- **Observation**: Gate leakage of Functional Units increases and propagation delay decreases as oxide thickness deceases.

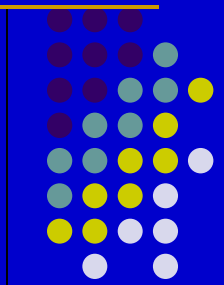
- **Strategy**: Maximize utilization of high-T_{OX} high leaky resources (e. g. multipliers) and low-T_{ox} low leaky resources (e.g. adder, subtractor) to improve chances of tunneling current reduction with minimal performance degradation.

Dual-T_{ox} Assignment Algorithm (Assumption)



- Datapath is represented as a sequencing DFG, which is a directed acyclic graph.
- Each vertex of the DFG represents an operation and each edge represents a dependency.
- Each vertex has attributes that specify the operation type.
- Each node connected to the primary input is assigned two registers and one multiplexer while the inner nodes of the DFG have one register and one multiplexer.
- The delay of a control step is dependent on the delays of the functional unit, the multiplexer, and register.

Dual-T_{ox} Assignment Algorithm (Objective Function)

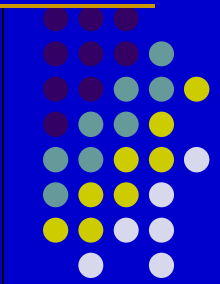


- The combined reduction of tunneling power dissipation and execution time translates to reduction of the current-delay-product (*CDP*).
- The objective of the scheduler is to minimize the *CDP* while assigning a schedule for the DFG. This implicitly facilitates minimization of tunneling current along with delay while considering resource constraints.
- Let N_c - number of control steps, and n_{FUc} - number of resources active in any step *c*. Then, *CDP* is calculated as,

$$CDP = \sum_{c=1}^{Nc} \sum_{r=1}^{nFUc} I_{ox,\,FU}(c,r) * T_{pd,\,FU}(c,r)$$

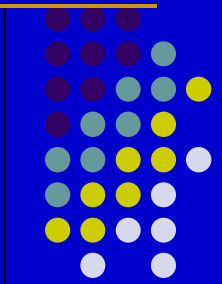
Here, $I_{ox,FU}(c,r)$ is tunneling current of *r*-th functional unit active in step *c* and $T_{pd,FU}(c,r)$ is its propagation delay.

Dual-T_{OX} Assignment Algorithm (Gaussian Variation)



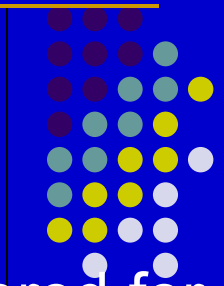
- We assume that the transistors inside the same resource have same oxide thickness, and transistor gate oxide thicknesses may differ for various functional units.
- To ensure that the results take process variations into account we assume that a given gate oxide thickness *Tox* can take any value in the range ($T_{ox} - \Delta T_{ox}$, $T_{ox} + \Delta T_{ox}$).
- We assume such variations to be Gaussian. We also maintain constant (*L / T*) and (*W / L*) ratios.
- Thus, all three parameters *Tox*, *L*, and *W* have Gaussian variation.

Dual-T_{OX} Assignment Algorithm (Steps)



1. Find total number of FUs of all available oxide thicknesses.
2. Get resource constrained *ASAP* and *ALAP* schedules.
3. Fix the number of cycles as the max. of *ASAP* and *ALAP* steps.
4. Find the vertices in critical path V_c and off-critical path V_{oc}.
5. Assume the above *ASAP* schedule as the current schedule.
6. For each $v \varepsilon V_c$ assign T_{oxH} to operations needing high-leaky resources and lowest thickness T_{oxL} to operations needing low-leaky resources.
7. While all $v \varepsilon V_{oc}$ of the current schedule are not considered for time stamping use heuristic to fix time stamp of the vertex with the *Tox* assignment for which *CDP* is minimum.
8. Find all vertices scheduled in every clock cycle. For operations using high leaky resources (e. g. multiplier or divider), if critical vertex has higher T_{ox} than off-critical then swap the T_{ox}.
9. Calculate Power and Delay for the scheduled DFG.

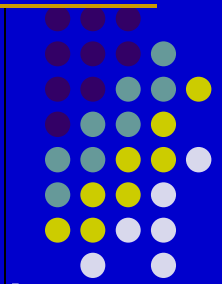
Dual-T_{ox} Assignment Algorithm (Heuristic)



While all $v \varepsilon V_{oc}$ of the current schedule are not considered for time stamping

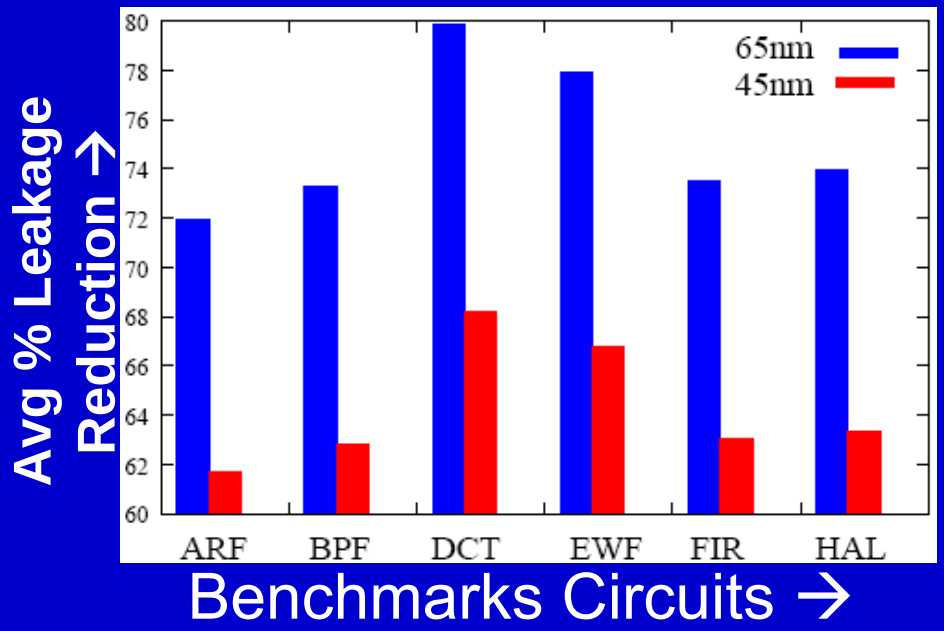
1. If vertex v is needs a high-leaky resource then
   1. assign the highest available thickness T_{oxH}.
2. Else assign the highest available thickness T_{oxL}.
3. Generate Gaussian a random numbers in the range $(T_{ox} - \Delta T_{ox}, T_{ox} + \Delta T_{ox})$.
4. Calculate the *CDP* of the current schedule for above T_{ox}.
5. For each off-critical vertex V_{oc} of the current schedule
   1. For every allowable control step c in the mobility range of v
      1. Assign next higher thickness if vertex needs high leaky resource and next lower thickness if vertex needs low leaky resource.
      2. Generate Gaussian random numbers in the range $(T_{ox} - \Delta T_{ox}, T_{ox} + \Delta T_{ox})$.
      3. Find the *CDP* of the current schedule for above T_{ox}.
6. Fix time stamp of the vertex with the current T_{ox} assignment for which *CDP* is minimum.
7. Remove the above time stamped vertex from V_{oc}.

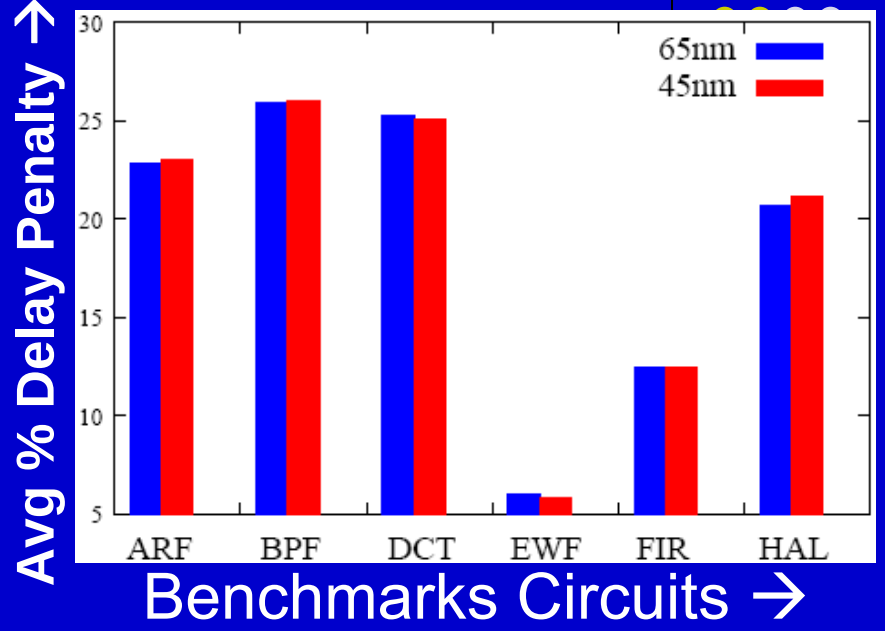
Experimental Results : Setup



- For both 65nm and 45nm technology we have chosen different oxide thickness in which higher thickness is 35% more than the lower thickness.
- First we carried out our experiments using resources of two different gate oxide thicknesses.
- The value of ΔT_{oxp} is assumed to be 10% of the original T_{oxp}.
- Nominal base T_{oxp} is chosen 0.7nm for 45nm tech and 1.0nm for 65nm tech.
- We also carried out experiments using functional units of three different gate oxide thicknesses.

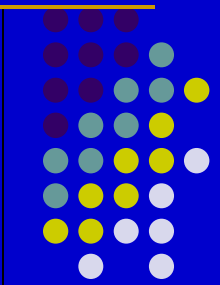
Experimental Results : % Average





- For three different gate oxide thicknesses: the maximum reduction improved in the range of 3 − 7% (the average reduction was improved by 2−5%) and there is increase in the average delay penalty in the range of 5 − 11%.
- This is observed for both 65nm and 45nm technology.

Conclusions and Future Works



- Gate leakage (tunneling current) is a major component of total power consumption of a low-end CMOS nanometer circuit.
- Dual-T_{ox} approach results significant reductions in tunneling current with minimal performance penalty.
- Development of optimal assignment algorithm is under progress.
- Tradeoff of leakage, area and performance needs to be explored.
- More sophisticated modeling to account process variation and mismatch is necessary.
- Dual-T_{ox} based design may need more masks for the lithographic process during fabrication.