ILP Based Leakage Optimization During Nano-CMOS RTL Synthesis: A DOXCMOS Vs DTCMOS Perspective

Saraju P. Mohanty
Dept. of Comp. Science & Engineering
University of North Texas, USA.
Email: saraju.mohanty@unt.edu

Bijaya K. Panigrahi
Dept. of Electrical Engineering
Indian Institute of Technology, Delhi.
Email: bkpanigrahi@ee.iitd.ac.in

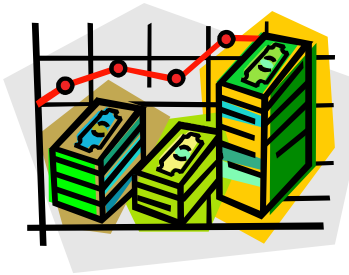
Acknowledgment: This research is supported in part by NSF award numbers CCF-0702361 and CNS-0854182.

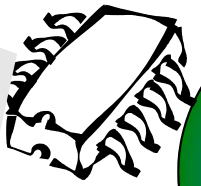
Outline of the Talk

- Introduction
- Prior Related Research
- DOXCMOS and DTCMOS Technology
- ILP Based Gate Leakage Optimization
- Architecture Component Library
- Experimental Results
- Conclusions

Introduction and Motivation

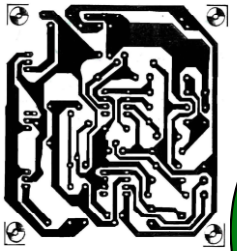
Why Low Power?



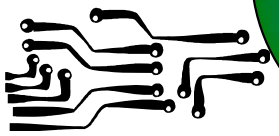


Packaging costs

Chip and system cooling costs



Power supply rail

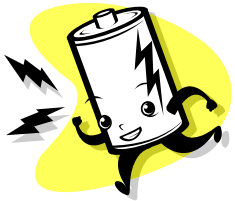


Power affects

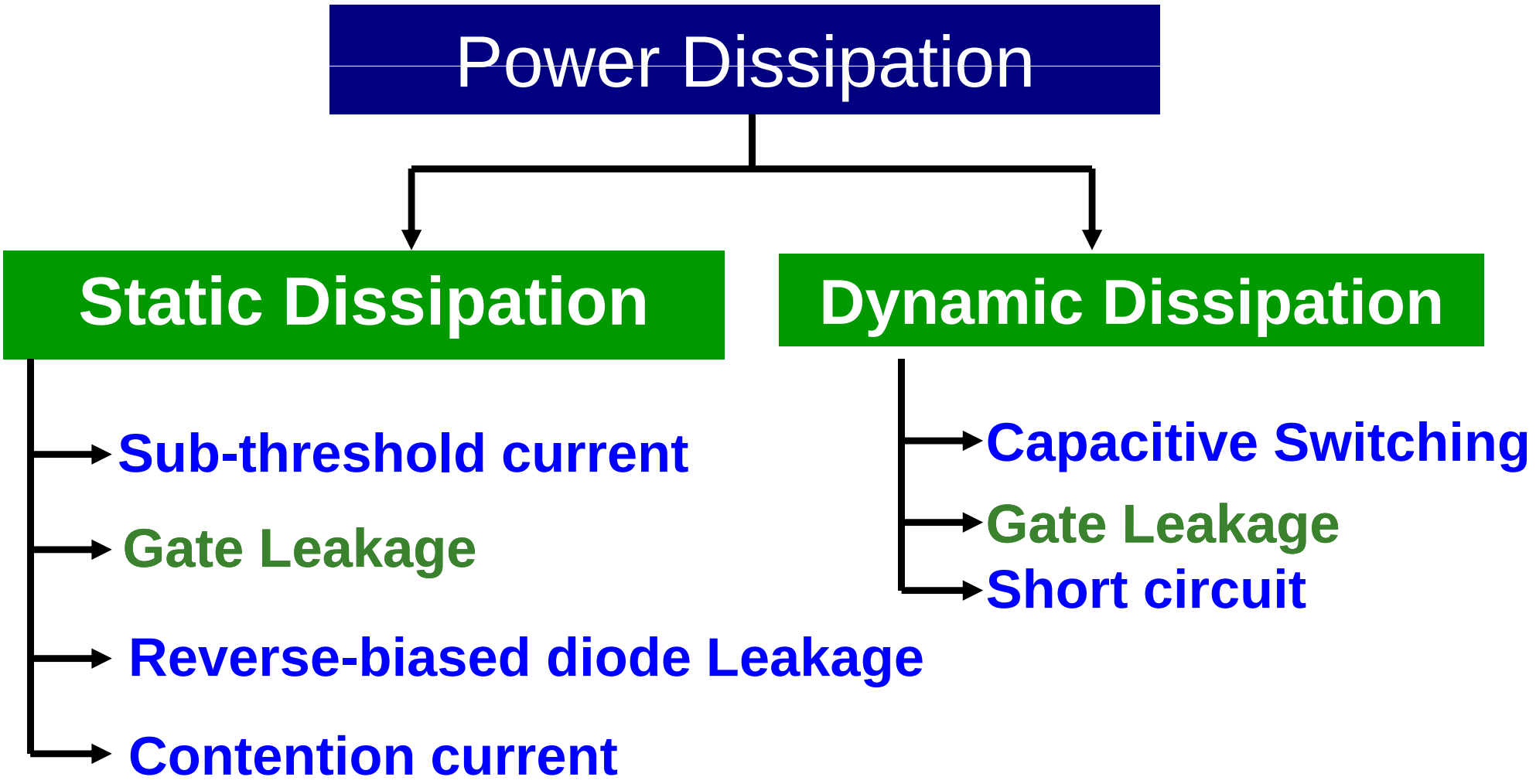
Noise and reliability

Environmental

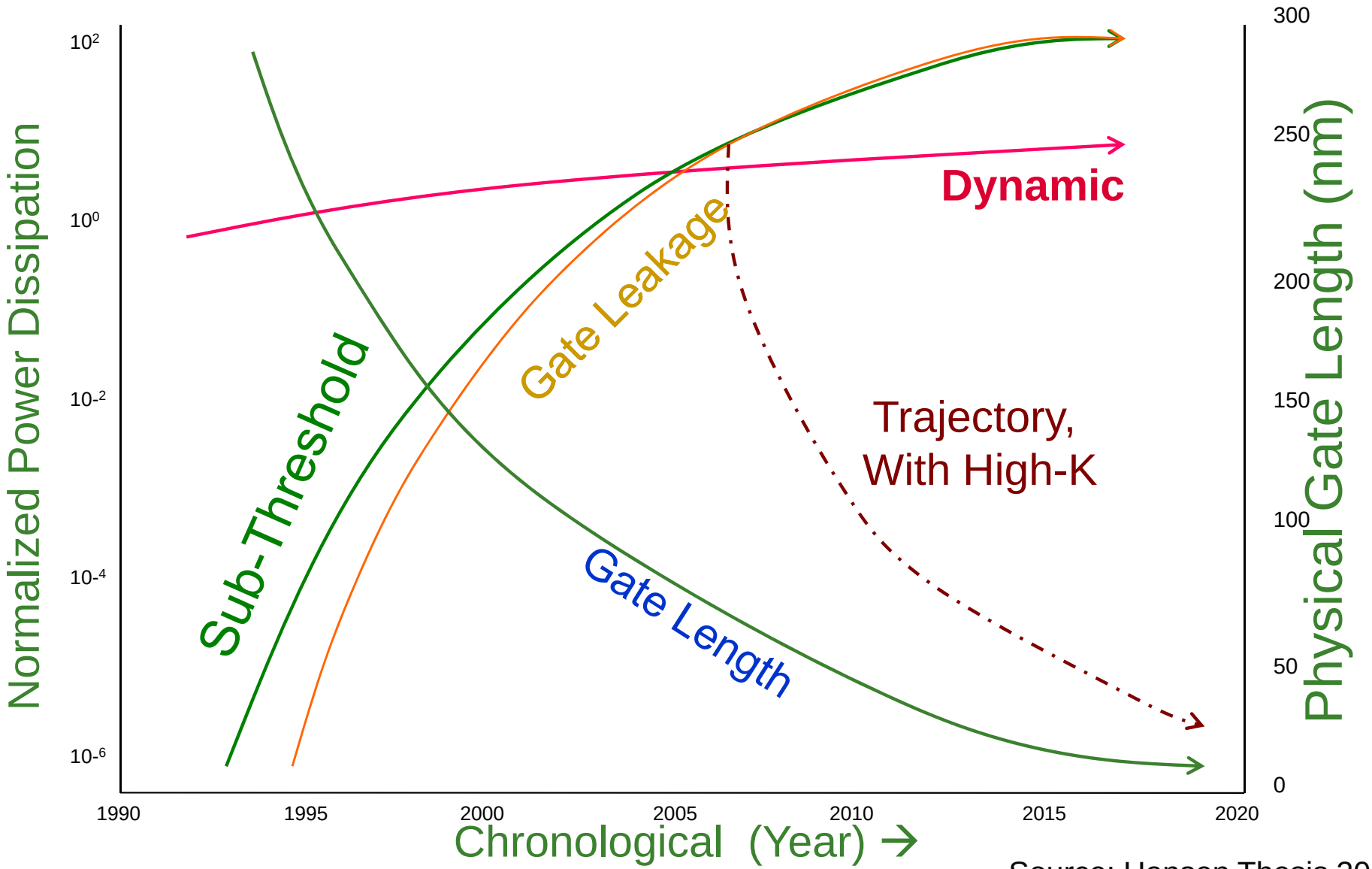
Battery life



Power Dissipation in CMOS

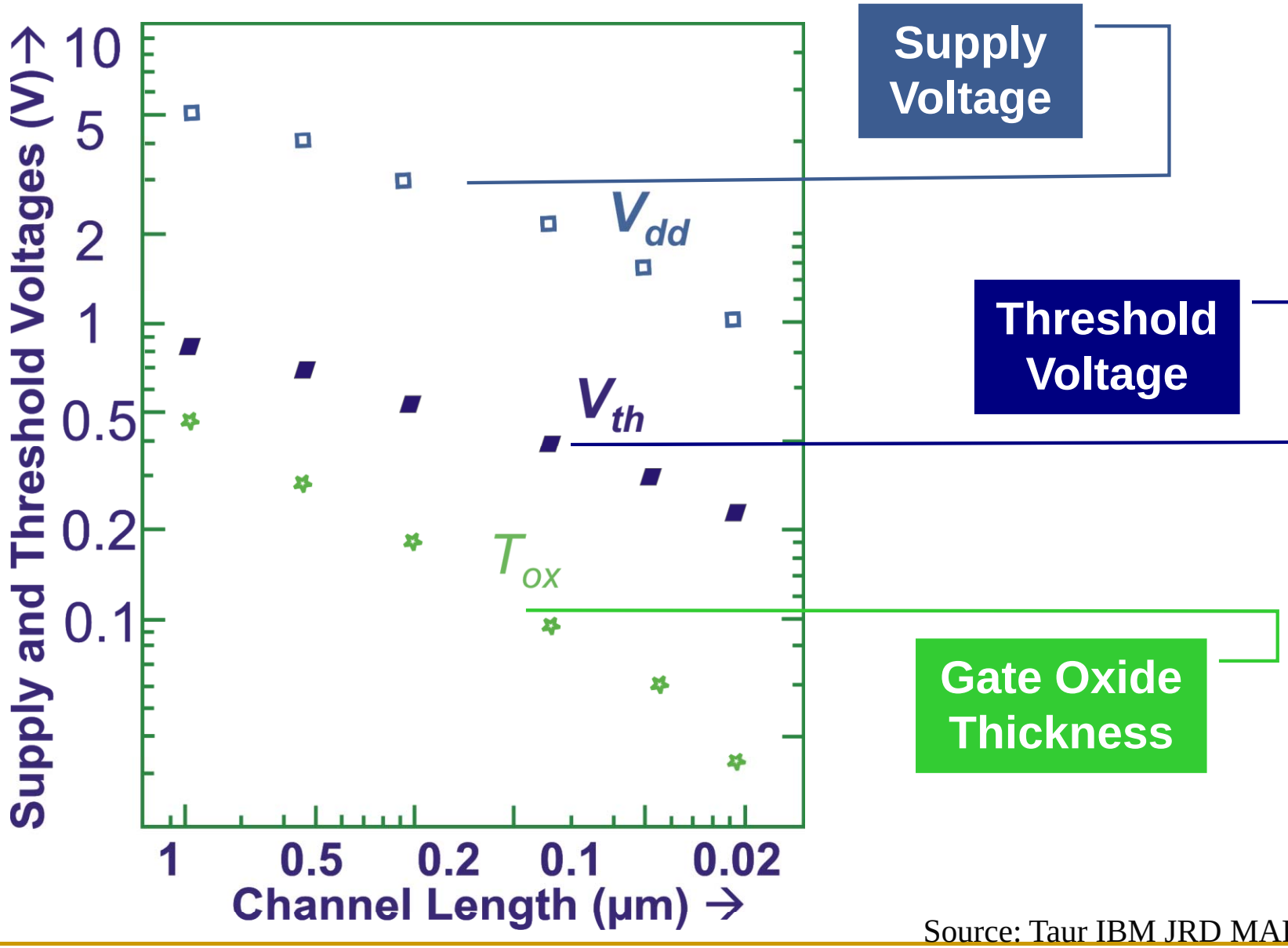


Power Dissipation Redistribution



Source: Hansen Thesis 2004

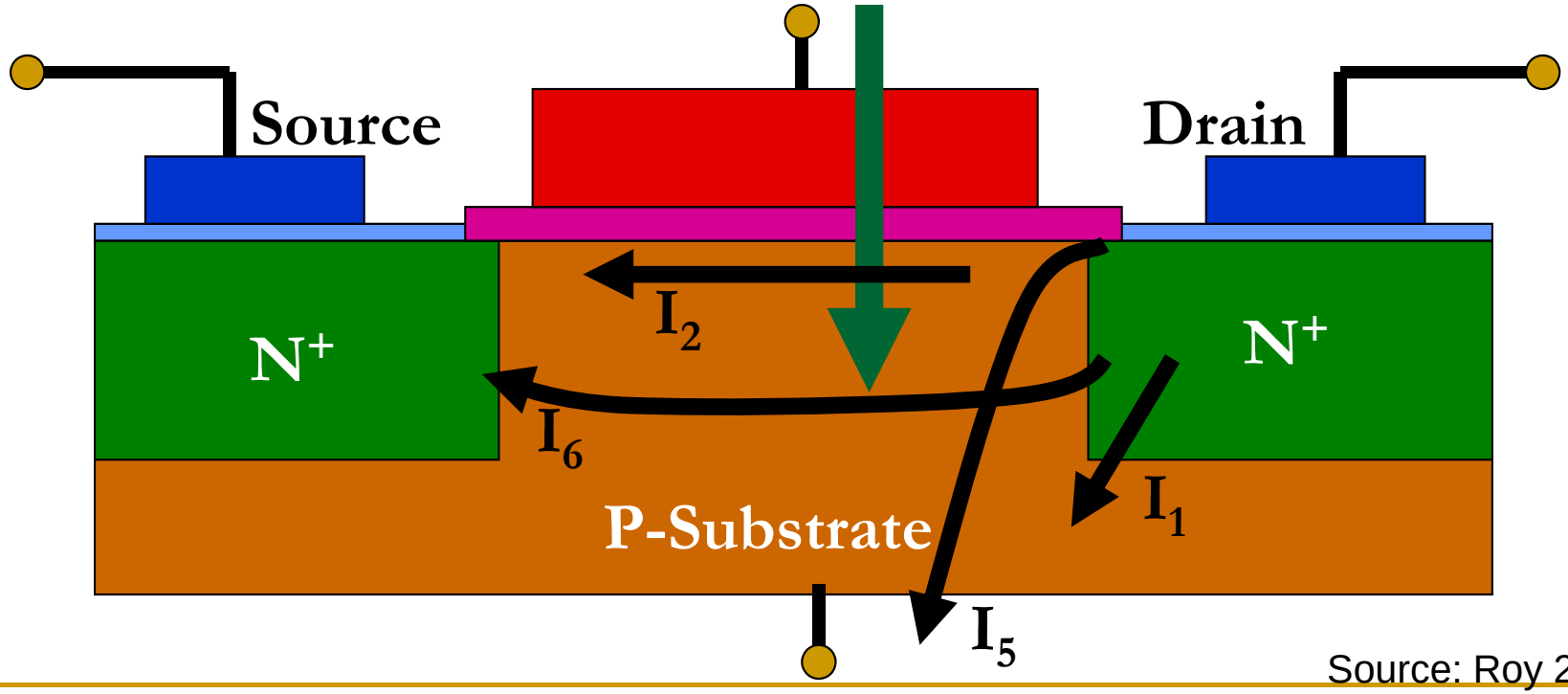
What are Scaled ?



Source: Taur IBM JRD MAR 2002

Leakages in CMOS

I_1 : reverse bias pn junction (both ON & OFF)
I_2 : subthreshold leakage (OFF)
I_3 : Gate Leakage 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 2003

Contributions of this Paper

- Explores DOXCMOS/DTCMOS technology for architectural level leakage and delay tradeoffs.
- Presents an approach that schedules operations of a sequencing data flow graph (DFG) and maps the operations to RTL library for optimization.
- The algorithm minimizes leakage delay product (LDP) for given resource constraints.
- Leakage accounts both gate and subthreshold.
- The RTL library is constructed for DOXCMOS and DTCMOS technologies.
- Comparative study of DOXCMOS versus DTCMOS based optimization is presented.

Related Prior Research

Related Research: RTL

Subthreshold Leakage:

- Khouri - TVLSI 2002 : Algorithms for subthreshold leakage power analysis and reduction using dual-V_{Th}.
- Gopalakrishnan - ICCD2003: Dual-V_{Th} approach for reduction of subthreshold current through binding.
- Tang - DAC 2005: A heuristic approach using dual-V_{Th}.
- Dal – ISQED 2006: Power island partitioning for reduction.

Gate Leakage:

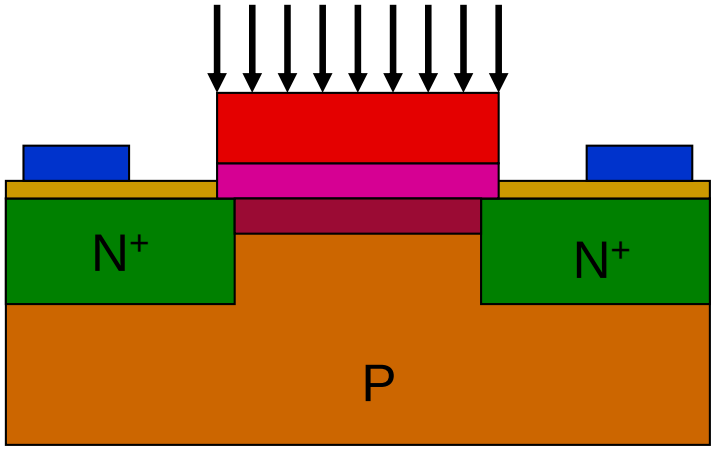
- Mohanty - VLSI Design 2006: Dual-T_{ox} approach for reduction of gate leakage current.
- Mohanty - ISQED 2006: Simulated annealing algorithms using dual-K or dual-T_{ox}.

Related Research: Logic / Transistor Level Gate Leakage Reduction

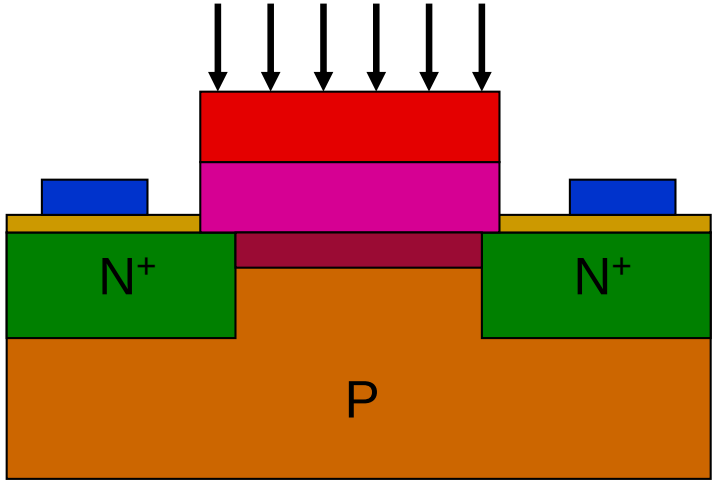
- Lee - TVLSI2004 : Pin reordering to minimize gate leakage during standby positions of logic gates.
- Sirisantana - IEEE DTC Jan-Feb 2004: Use multiple channel lengths and multiple gate oxide thickness for reduction of leakage.
- Sultania – TVLSI Dec 2005 and Sultania - DAC2004 : Heuristic for dual-T_{ox} assignment for gate leakage and delay tradeoff.
- Mukherjee - ICCD 2005: Introduced dual-K approach for reduction of gate leakage.

DOXCMOS / DTCMOS Technology: The Key Idea

Low T_{ox} Vs High T_{ox} Device

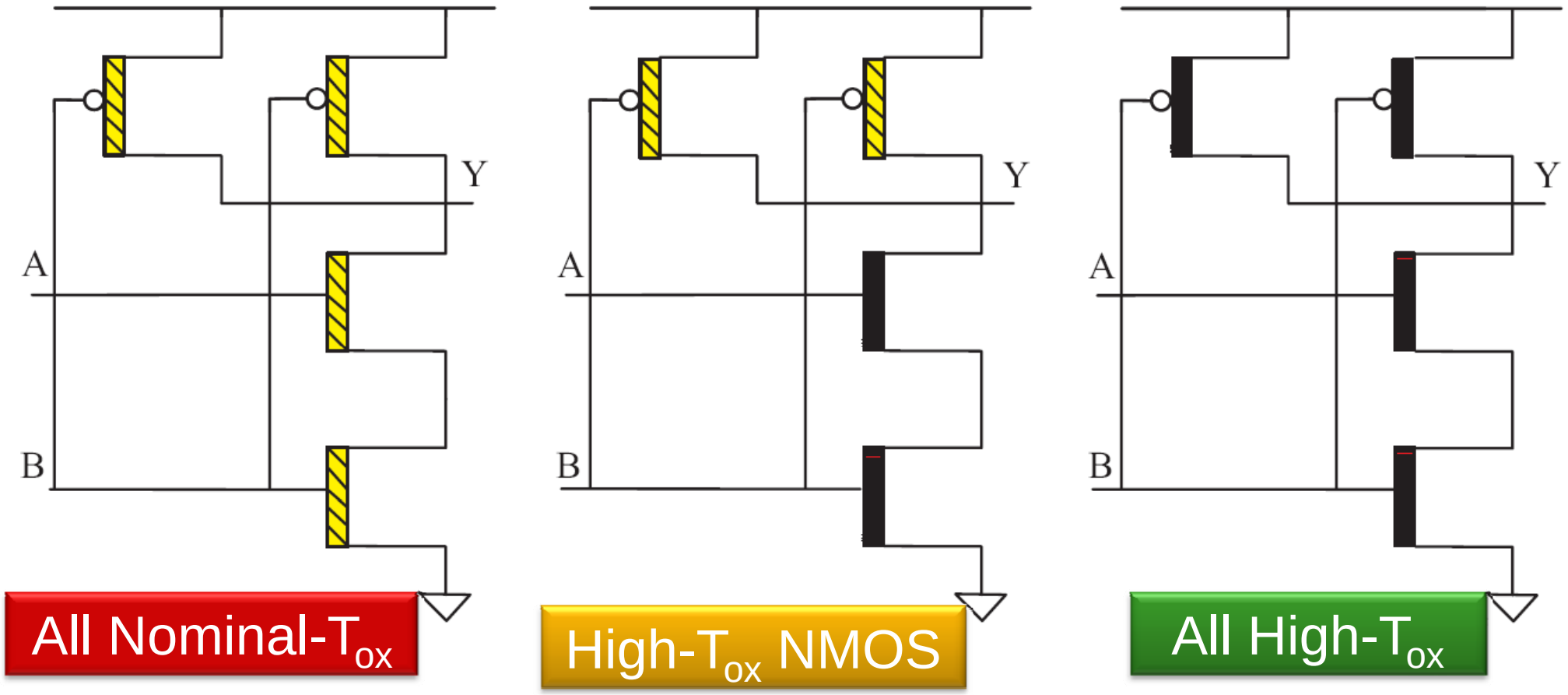


Low T_{ox} → Larger I_{gate} , Smaller delay



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

DOXCMOS Technology

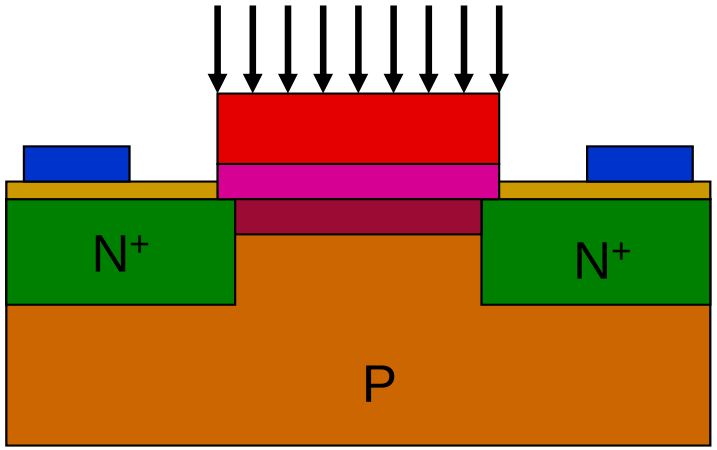


All Nominal-T_{ox}

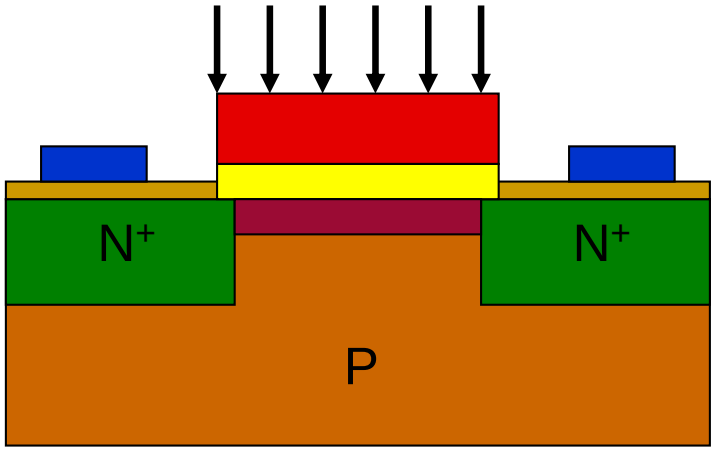
High-T_{ox} NMOS

All High-T_{ox}

Low V_{th} Vs High V_{th} Device

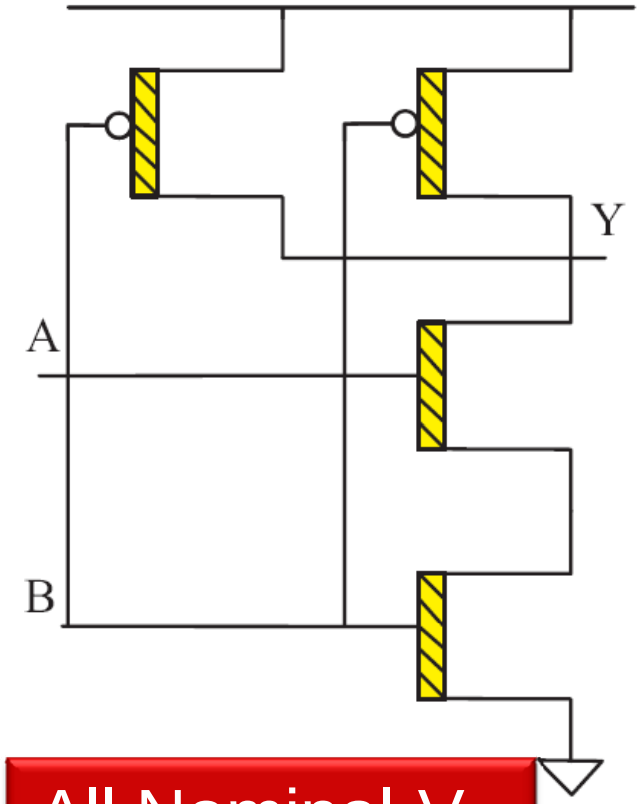


Low V_{th} → Larger I_{sub}, Smaller delay

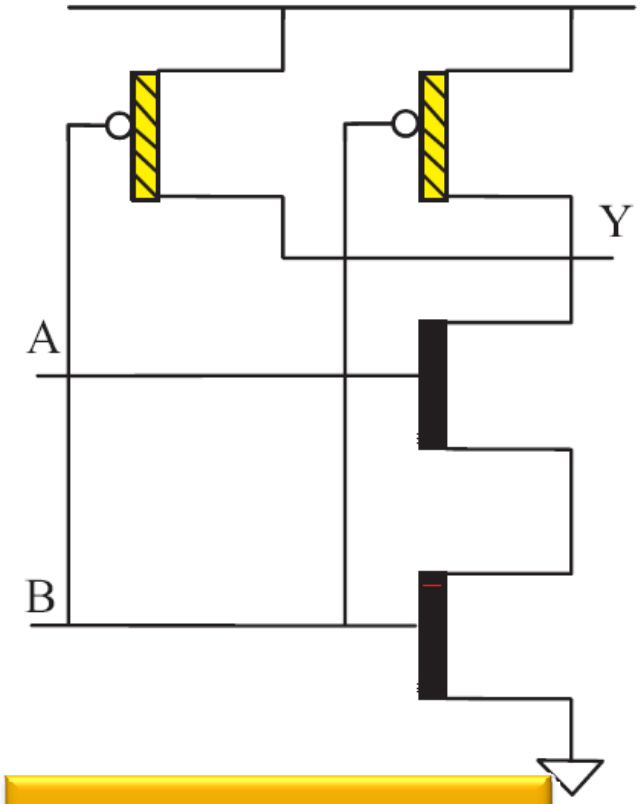


High V_{th} → Smaller I_{sub}, Larger delay

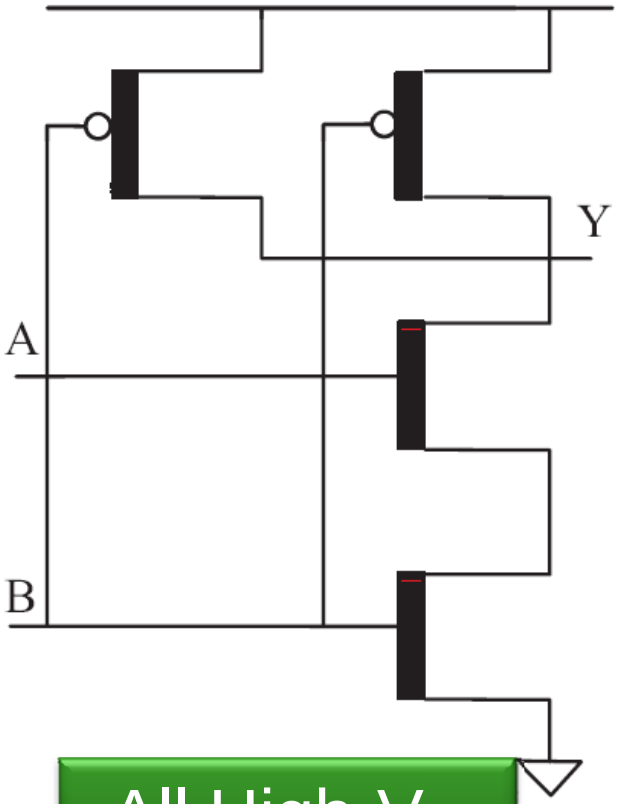
DTCMOS Technology



All Nominal-V_{th}



High-V_{th} NMOS

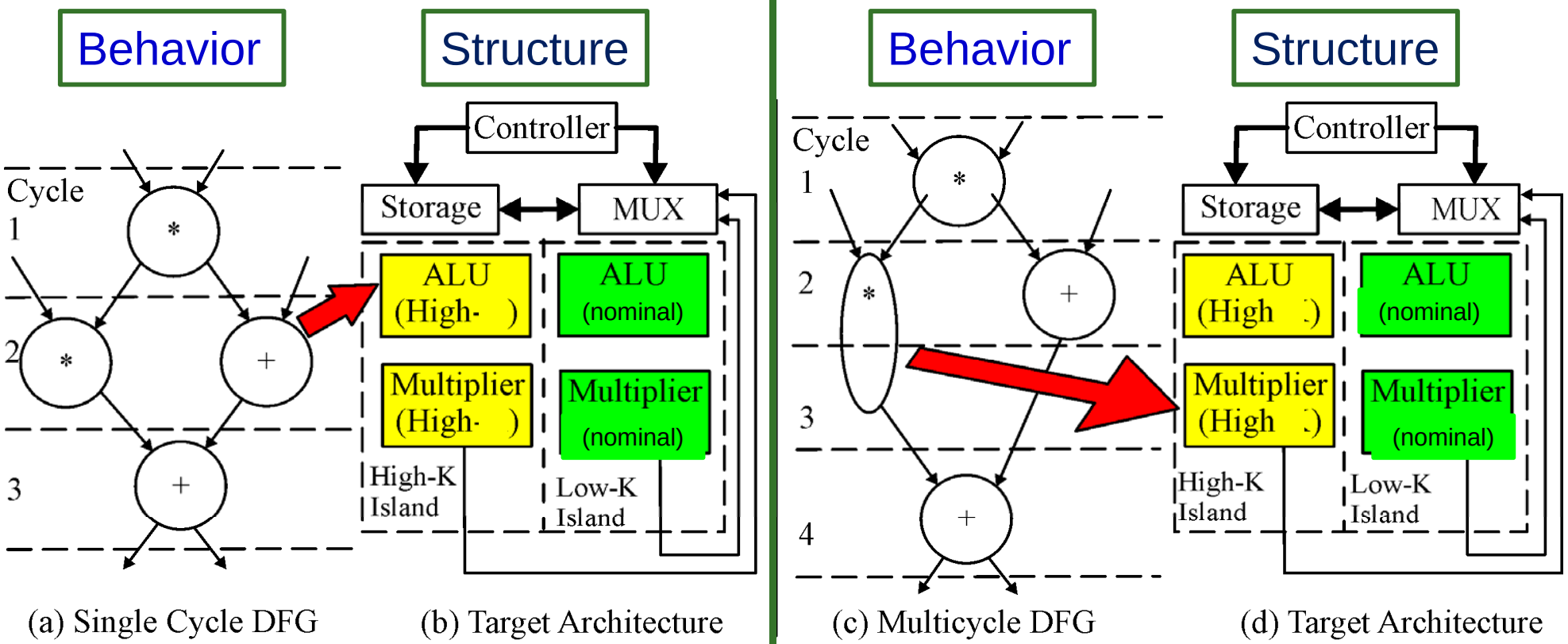


All High-V_{th}

The Key Idea : At RTL

- Each functional unit (or RTL component, e.g. adder) have is made of same T_{ox} or V_{th}.
- The DOXCMOS library has units made of high-T_{ox} and nominal-T_{ox} transistors.
- The DTCMOS library has units made of high-V_{th} and nominal-V_{th} transistors.
- A mix of RTL units of all-nominal case and all-high case will serve leakage and delay trade-offs and will go well with industry trend.

The Key Idea : Target Architecture



ILP-Based Leakage Optimization

Problem Formulation

- Given an unscheduled data flow graph $G_U(V,E)$, it is required to find the scheduled data flow graph $G_S(V,E)$ with appropriate resource binding such that the total-leakage and delay product (LDP) is minimized under given resource constraints.

Objective Function: Minimize: $LDP(DFG)$

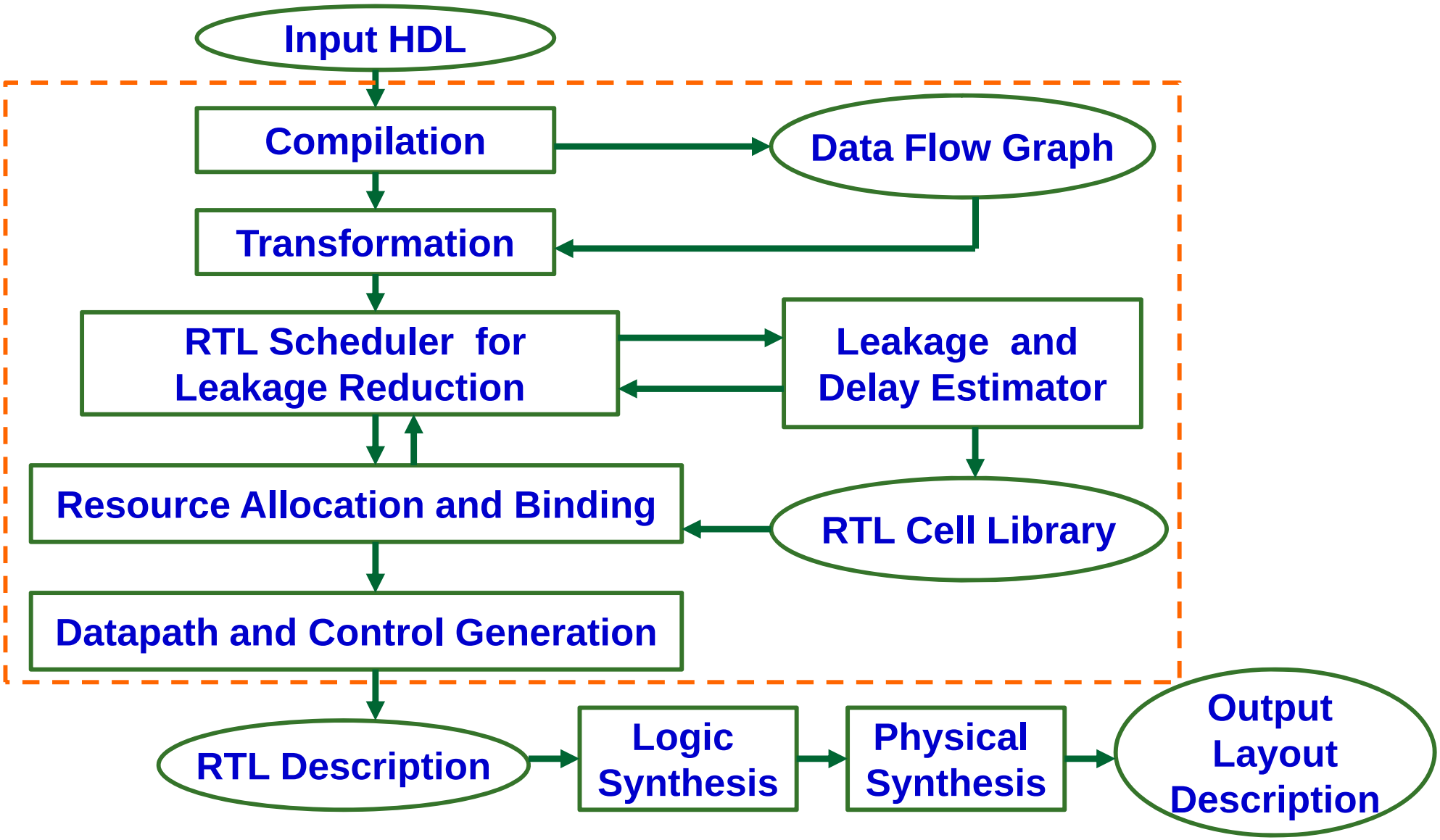
Constraints: $\text{Allocated}(R_{t,K}) \leq \text{Available}(R_{t,K}), \forall c \in N$

Leakage and Delay Product Calculation:

$$LDP(DFG) = \sum_{c=1}^{N} LDP_c$$

$$= \sum_{c=1}^{N} \sum_{\forall v_{i,c}} P_{leakage}(v_{i,c}) \times d_c$$

RTL Optimization for Leakage



ILP-Based Optimization

- ILP-based algorithm minimizes leakage while performing simultaneous scheduling, allocation, and binding.
- ILP-based algorithm provides optimal solution.
- ILP-based provides useful solution to the problems in a reasonable time at the architecture level unlike logic or circuit level where chip-complexity is too high.
- ILP has exponential time-complexity, however is not an issue for RTL abstraction of the circuit, where chip-complexity is low.
- ILP can be modeled using AMPL and easily solved and the design flow can be fully automated.

LDP Optimization Flow …

1. Preprocess given behavioral description to construct a sequencing DFG.
2. Perform simulations to estimate gate leakage and delay of RTL units.
3. Construct resource allocation table and available resource table based on input resource constraints.
4. Obtain ASAP and ALAP schedules of the input DFG.
5. Determine the number of different resources for each T_{ox} or V_{th} using the resource allocation table.
6. Modify both ASAP and ALAP schedules obtained above using the number of resources found in previous step.

LDP Optimization Flow

7. Construct the mobility graph based on above schedules.
8. Fix the total number of clock cycles as the maximum of modified ASAP and ALAP schedules' control step.
9. Model the ILP formulations of the DFG using AMPL.
10. Obtain the final solution by solving the ILP formulations.
11. Estimate the gate and subthreshold leakage and delay.
12. Postprocess scheduled sequencing DFG to generate gate or subthreshold leakage optimal RTL description.

ILP Formulations …

- Objective Function: The objective is to minimize the LDP of the whole DFG over all control steps. This can be expressed using decision variable as:

$$\text{Minimize: } LDP(DFG)$$

$$\text{Minimize: } \sum_{c}\sum_{i}\sum_{t} X_{i,t,c} \times LDP(i,t)$$

ILP Formulations

Uniqueness Constraints: These constraints are represented as, $\forall i,\ 1 \le i \le V$,

$$\sum_{c}\sum_{t} X_{i,t,c} = 1$$

Precedence Constraints: These constraints should also ensure the multicycling and are modeled as, $\forall i, j, v_i \in Pred_{vj}$,

$$\sum_{t}\sum_{d=S_i}^{E_i} d \times X_{i,t,d} - \sum_{t}\sum_{e=Sj}^{E_j} e \times X_{j,t,e} \le -1$$

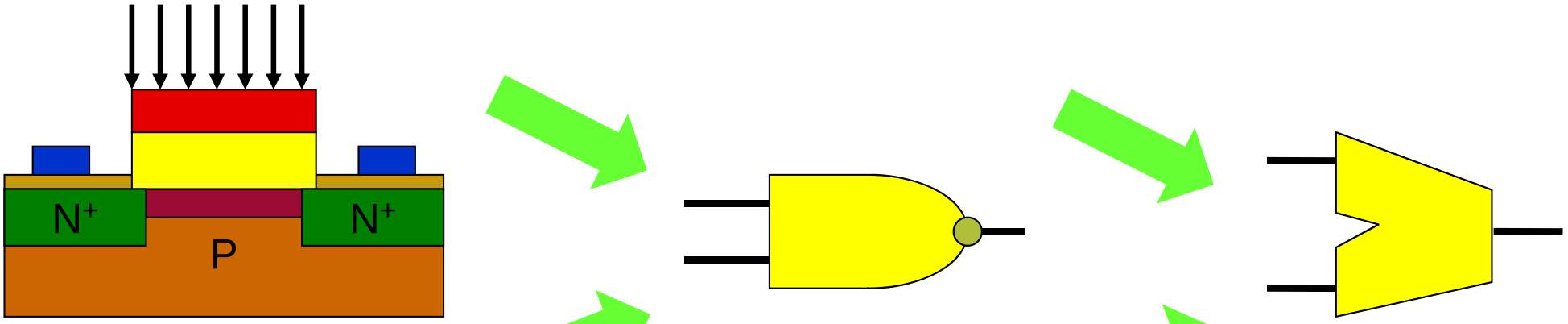
Resource Constraints: These constraints ensure that each cycle uses resources not exceeding available number of resources and are enforced as, $\forall t$ *and* $\forall c,\ 1 \le c \le N$,

$$\sum_{i \in R_{k,t}} X_{i,t,c} \le M_{k,t}$$

RTL Components Library

Datapath Component Library …

3 Level Bottom-up Hierarchical Approach



Transistor Level

Logic Level

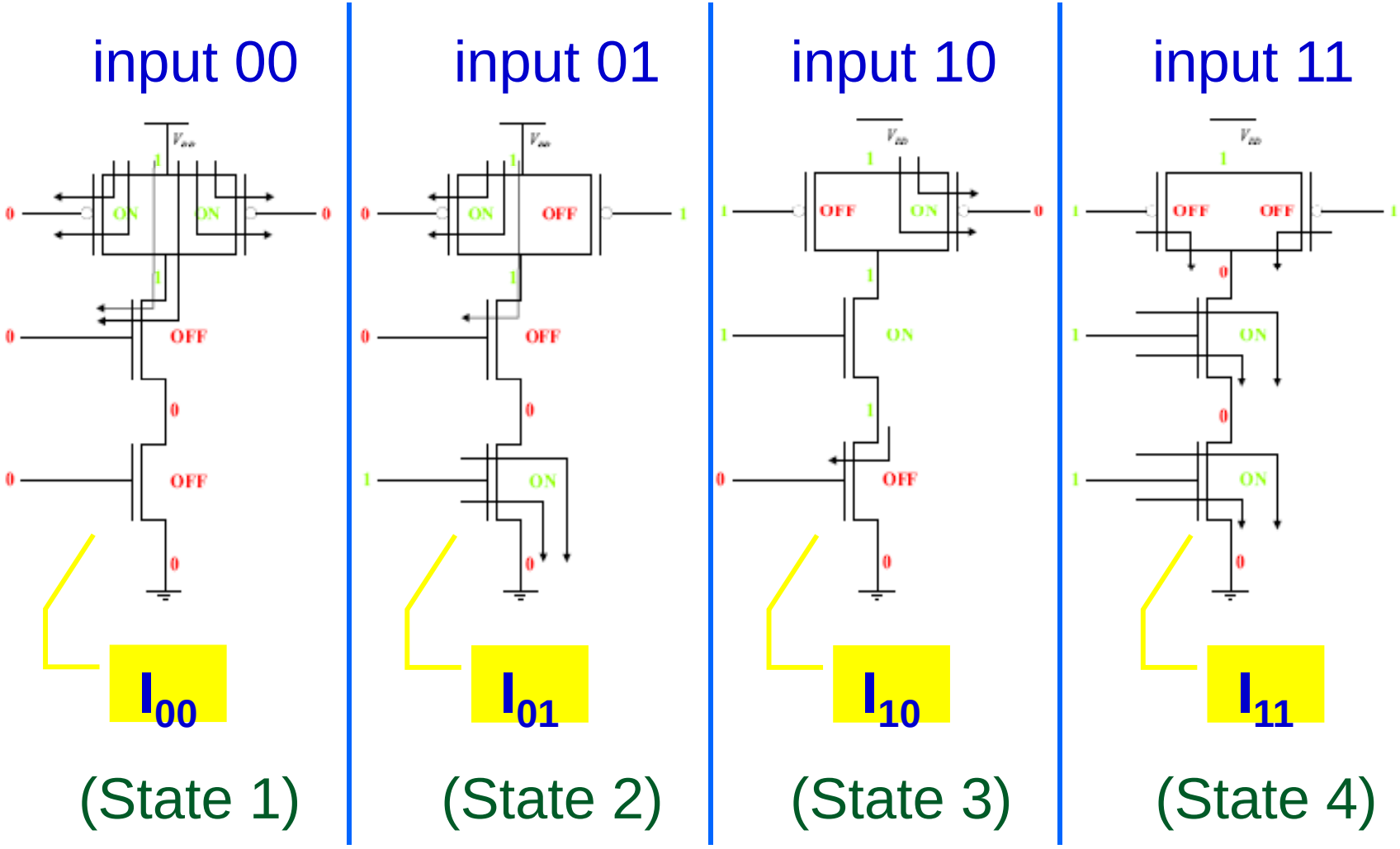
Datapath Components

It is observed that a NAND gate has least gate leakage compared to all other basic logic gates. Therefore the datapath components are constructed using NAND logic gates.

Datapath Component Library …

- First the NAND gate is characterized using analog simulations and then the functional units.
- It is assumed that there are total n_{total} NAND gates in the network of NAND gates constituting an n-bit functional unit out of which n_{cp} are in the critical path of the logic netlist.
- The effect of interconnect wires are not considered and the focus is on the gate leakage dissipation and propagation delay of the active units only.

Datapath Component Library … (NAND Gate)



Datapath Component Library …

- Gate leakage current for a specific state of a logic gate is then calculated by:

$$I_{gate\,\text{Logic}_{\text{state}}} = \sum_{\forall MOS_i} \left| I_{gate\,\text{MOS}}[i] \right|$$

- Gate leakage of a n-bit RTL unit is calculated as:

$$I_{gate,\,\text{R}} = \sum_{j=1}^{n_{total}} \Pr ob(state)\, I_{gate,\,\text{NAND}\,j_{\text{state}}}$$

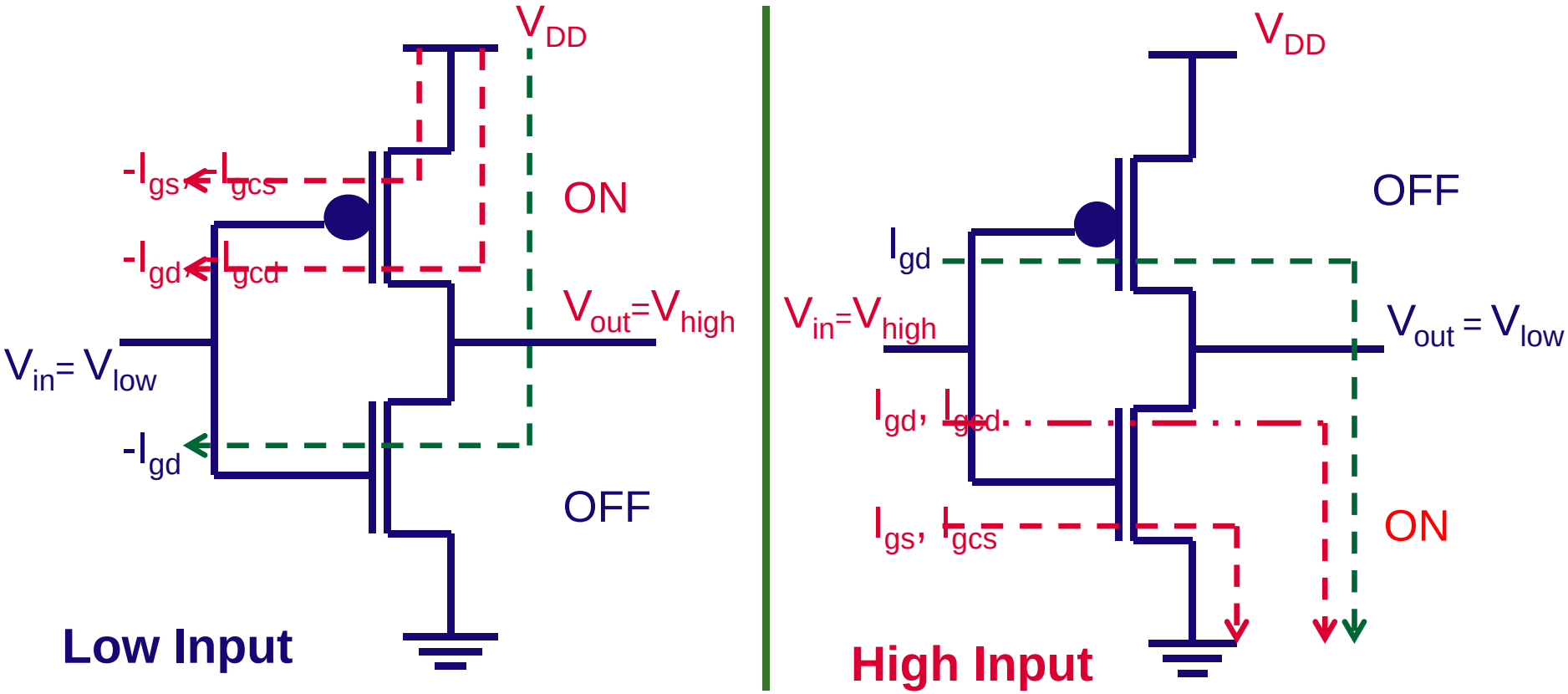
- Subtreshold leakage of a n-bit RTL unit is:

$$I_{sub,\,\text{R}} = \sum_{j=1}^{n_{total}} \Pr ob(state)\, I_{sub,\,\text{NAND}\,j_{\text{state}}}$$

- Delay of an *n*-bit functional unit is: $T_{pd\,\text{R}} = \sum_{i=1}^{n_{cp}} T_{pd\,NANDi}$

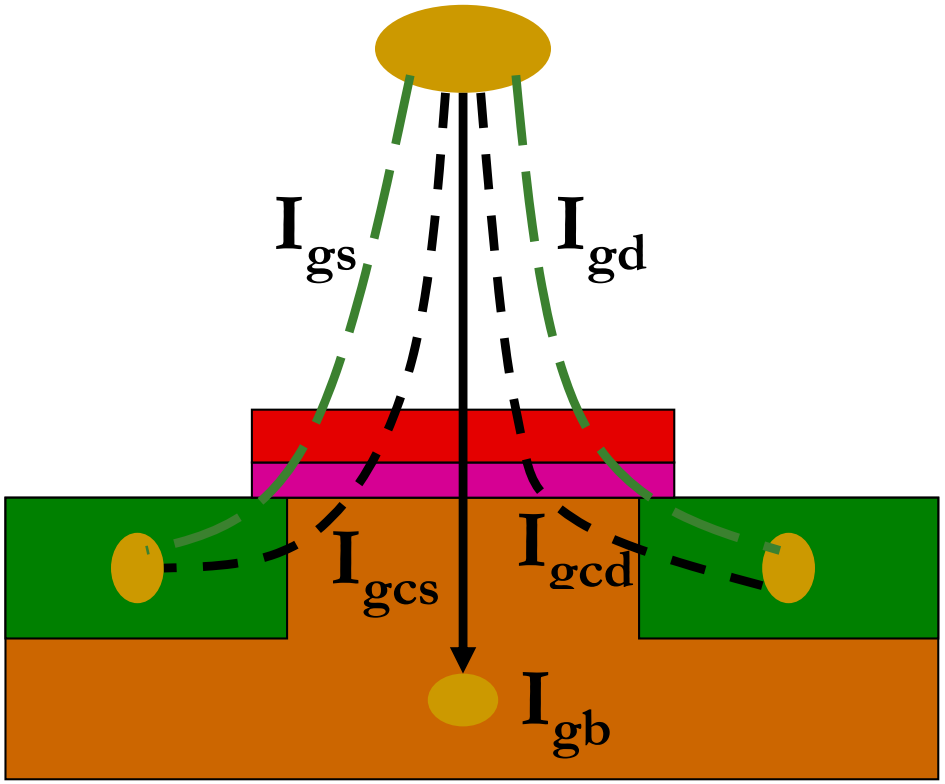
Datapath Component Library …
(Inverter Showing Components)

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



NOTE: Gate to body component found to be negligible.

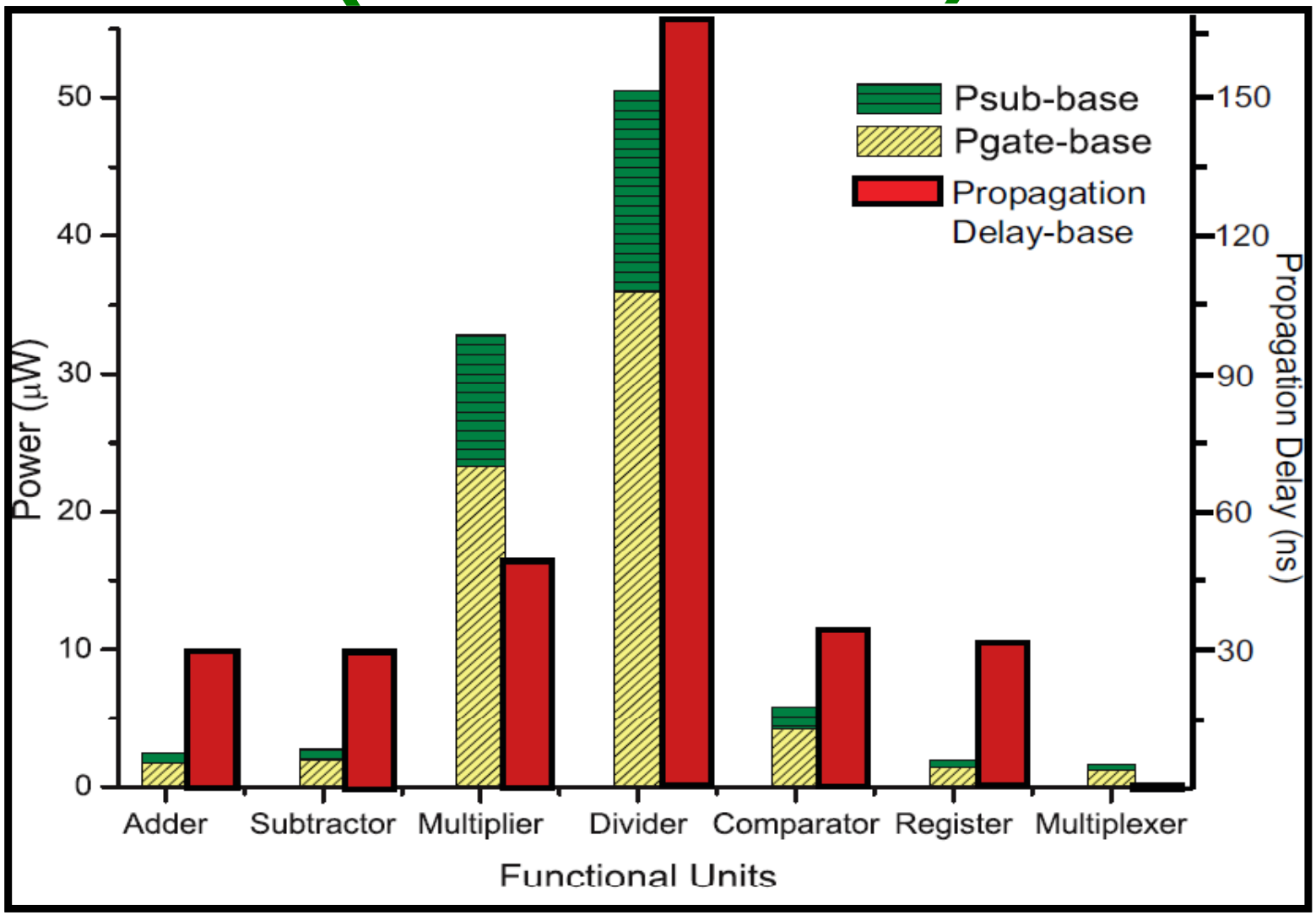
Datapath Component Library …
(A CMOS Transistor)



BSIM4 Model

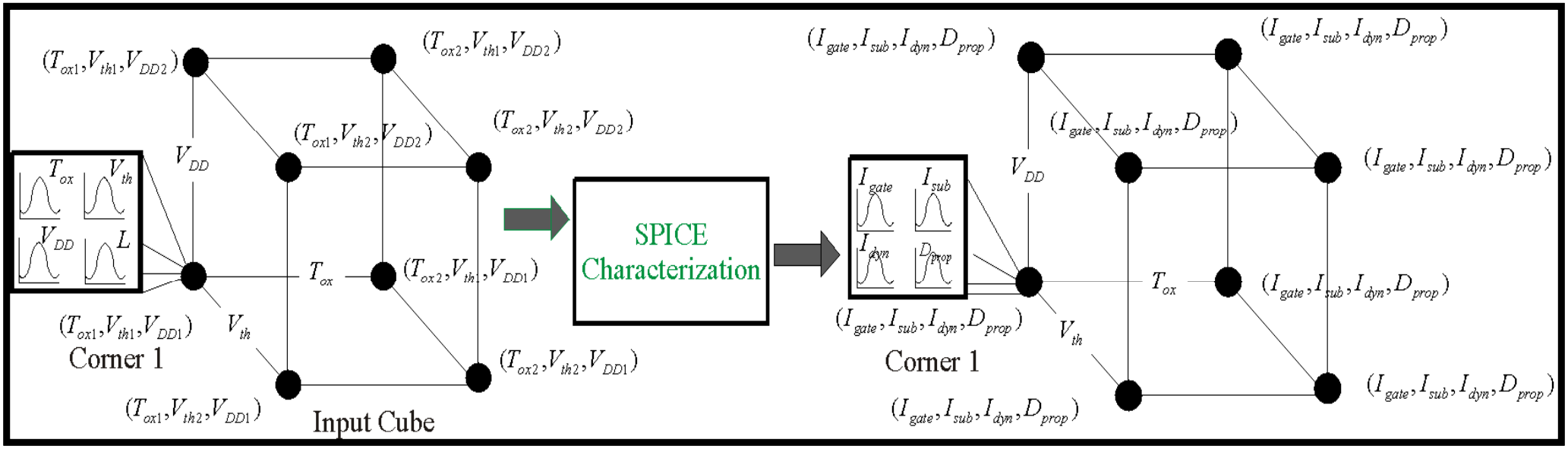
- Calculated by evaluating both the source and drain components
- For a MOS, $I_{gate} = (|I_{gs} + I_{gd} + I_{gcs} + I_{gcd} + I_{gb}|)$
- Values of individual components depends on states, ON or OFF

Datapath Component Library: (Baseline Data)



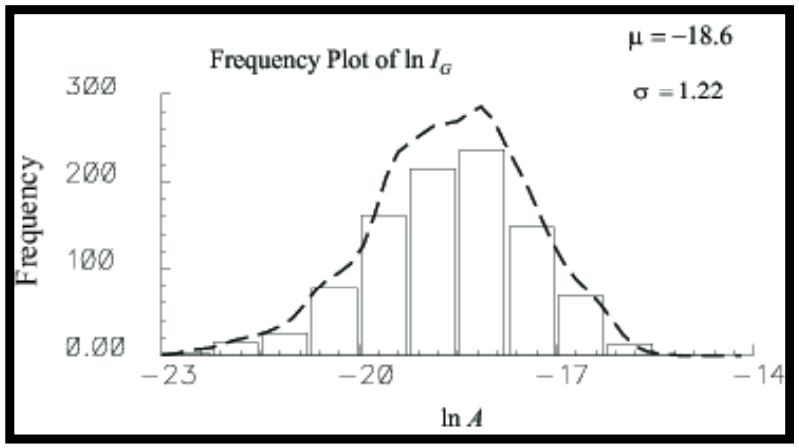
Datapath Component Library: Statistical Data …

- Through Monte Carlo simulations the input process and design variations are modeled.

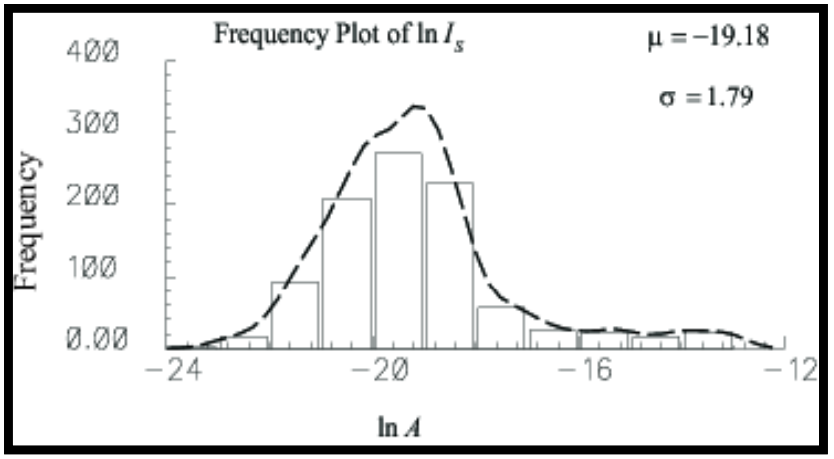


- Correlation are considered at transistor level.

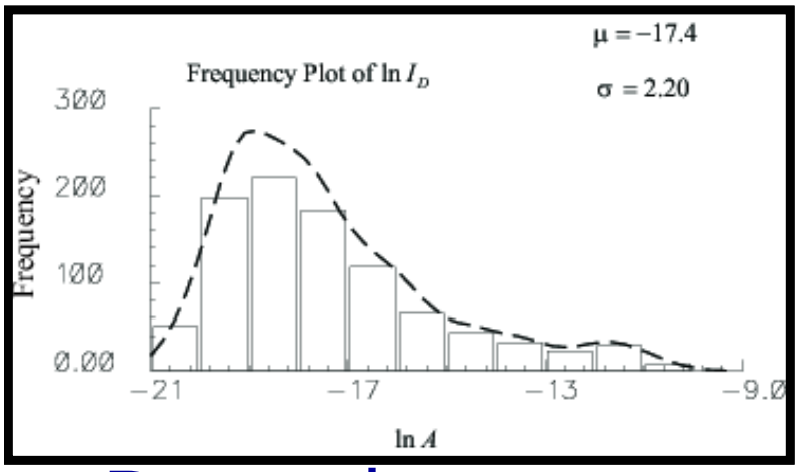
Datapath Component Library: Statistical Data …



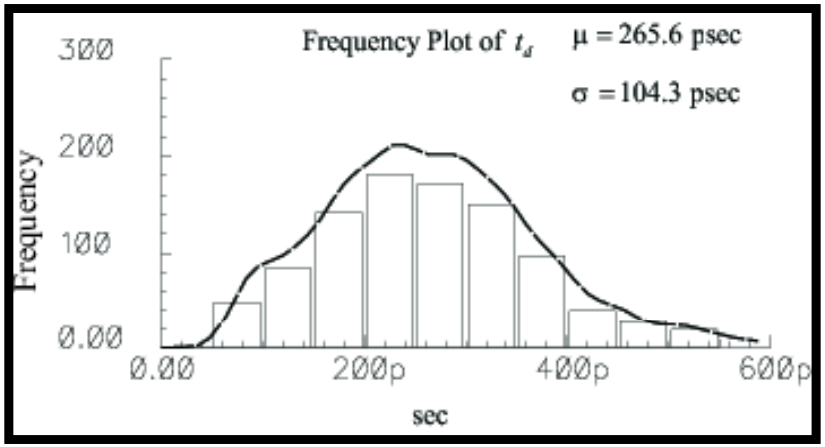
Gate leakage current



Subthreshold leakage current



Dynamic current



Propagation delay

Datapath Component Library: Statistical Data …

- The PDF of a current component of a functional unit is calculated as:

$$I_{dyn}^{FU} = \text{Statistical Summation over N} \left(I_{dyn}^{NAND} \right)$$

$$I_{sub}^{FU} = \text{Statistical Summation over N} \left(I_{sub}^{NAND} \right)$$

$$I_{gate}^{FU} = \text{Statistical Summation over N} \left(I_{gate}^{NAND} \right)$$

- The PDF of delay can be calculated as:

$$D_{prop}^{FU} = \text{Statistical Summation over } N_{CP} \left(D_{prop}^{NAND} \right)$$

Datapath Component Library: Statistical Data …

- Using Central Limit Theorem (CLT) the logic level distributions are translated to RTL.
- Assuming that the distributions for each gate are statistically independent of each other, the mean and variance of the leakages are calculated as:

$$\mu_R = \sum_{i=1}^{n_{total}} \mu_{\mathrm{NAND}_i}$$

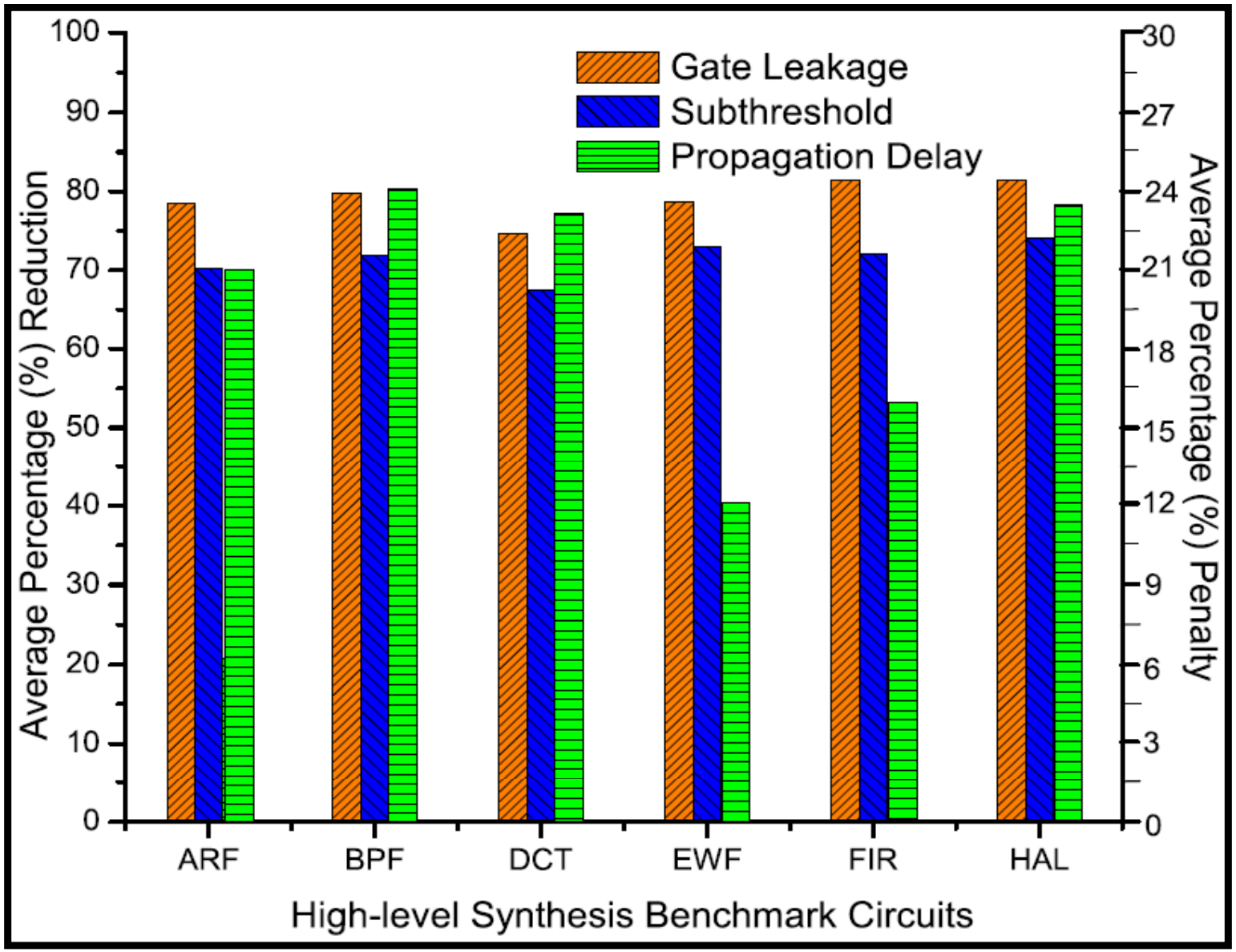
$$\sigma_R = \sqrt{\sum_{i=1}^{n_{total}} \sigma^2_{\mathrm{NAND}_i}}$$

Experimental Results

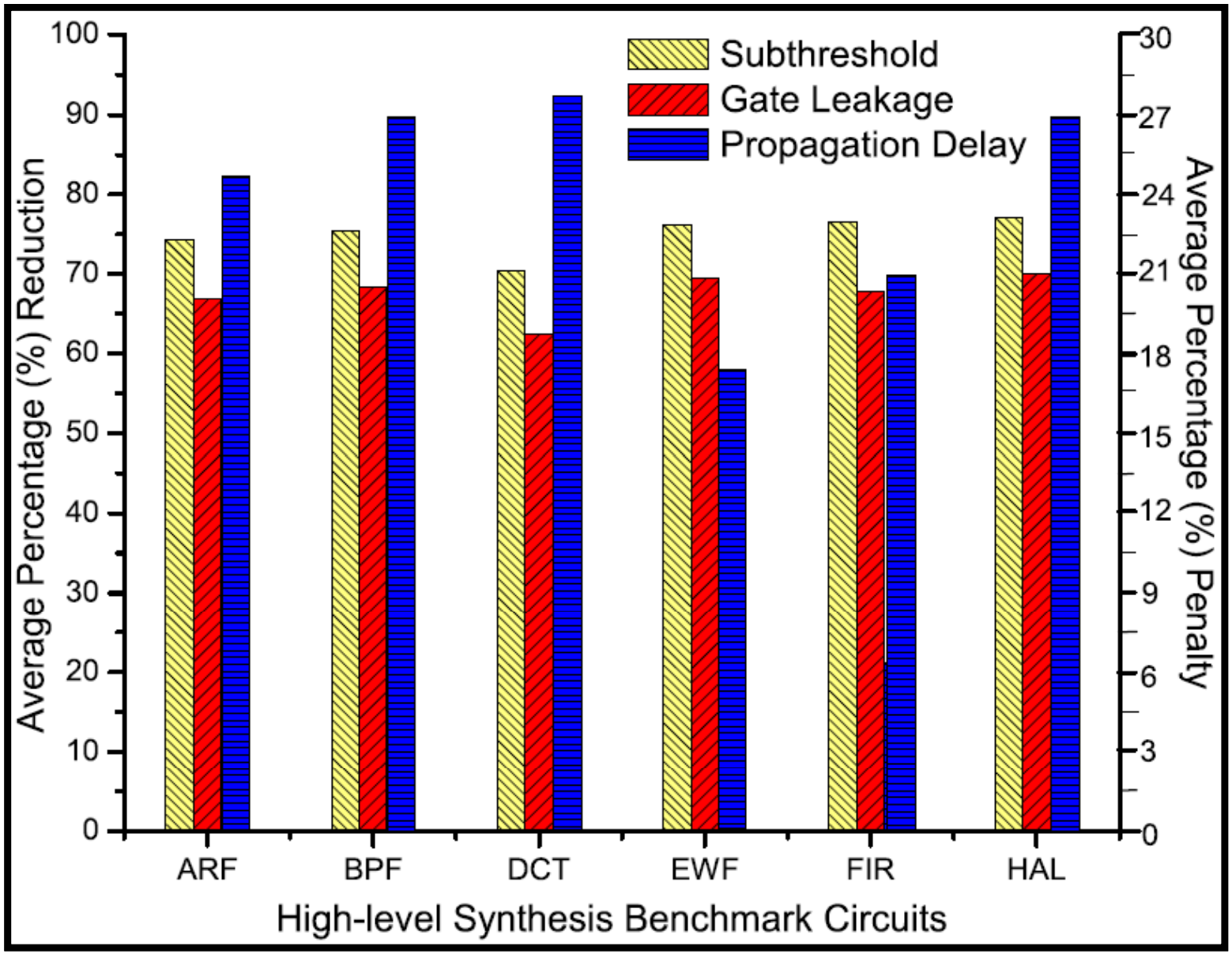
Experimental Results : Setup

- 45nm CMOS technology with baseline T_{ox} = 1.4nm and V_{th} = 0.22V.
- High-T_{ox} = 1.7nm and High-V_{th} = 0.25V.
- Algorithm implemented in C and integrated in in-house high-level synthesis tool.
- Experiments are performed on data intensive signal processing benchmark circuits whose applications are immense in day-to-day life.

Experimental Results : DOXCMOS



Experimental Results : DTCMOS



Experimental Results: Analysis ...

- For the DOXCMOS technology:
  - Gate leakage reduction: 74% to 81%.
  - Subthreshold leakage reduction: 67% to 74%.
  - Delay penalty of 12% to 23%.
- For the DTCMOS technology:
  - Gate leakage reduction: 70% to 77%.
  - Subthreshold leakage reduction: 62% to 70%.
  - Delay penalty of 17% to 27%.

Experimental Results: Analysis

- For DOXCMOS technology, the reduction of subthreshold leakage is due to change of V_{th} with T_{ox}, related by the following expression:

$$V_{Th} = V_{fb} + 2\phi_F + \left(\frac{T_{ox}}{\epsilon_{ox}}\right)\sqrt{2q\epsilon_{Si}N_{sub}\left(2\phi_F + V_{bs}\right)}$$

- For DTCMOS technology, the reduction in gate leakage due to increase in V_{th} is not straight forward. It is due to the reduction in gate tunneling current density with increase of V_{th}, as voltage across oxide drops by V_{th}.

Experimental Results: Comparison

DOXCMOS Technology	DTCMOS Technology
One parameter varied.	Several parameters varied.
Less area overhead: T_{ox} and L.	More area overhead: additional transistors.
Gate/subthreshold directly affected.	Subthreshold directly and gate indirectly.
Higher reduction.	Lower reduction.

Conclusions and Future Research

Conclusions

- An ILP based algorithm is presented for leakage optimization during architectural synthesis.
- The algorithm uses DOXCMOS and DTCMOS for leakage optimization under resource constraints.
- Experiments proved that both the techniques are quite effective. The percentage reductions in leakage is higher compared to existing literature.
- It is observed that DOXCMOS technology outperforms the DTCMOS technology and may be cheaper from fabrication point of view as well as only parameter need to be controlled.

Future Research

- Evaluation of the impact of these techniques on the area, capacitance and dynamic power dissipation is ongoing research.
- While the optimization is performed based on the mean value of power and delay distributions, the future optimization will account the variances as well to more accurately account process variations which are important for nano-CMOS.


Thank You !!!
The presentation is available at:
http://www.cse.unt.edu/~smohanty