Lecture 5: Nano-CMOS High-Level Synthesis

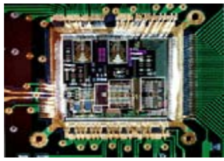
CSCE 6730
Advanced VLSI Systems

Instructor: Saraju P. Mohanty, Ph. D.

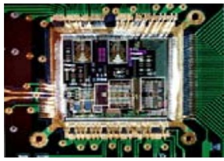
NOTE: The figures, text etc included in slides are borrowed from various books, websites, authors pages, and other sources for academic purpose only. The instructor does not claim any originality.

Outline of the Talk

- Issues in Nano-CMOS
- Challenges in The Context of HLS
- Proposed Techniques in Current Literature
- Conclusions

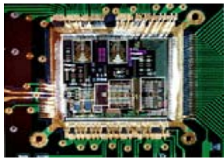


Issues in Nano-CMOS



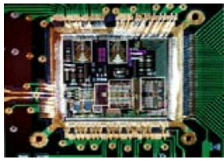
Issues in Nano-CMOS Circuits …

- **Variability**: Variability in process and design parameters has increased. They affect design decisions, yield, and circuit performance.
- **Leakage**: Leakage is increasing. Affects average as well as peak power metrics. Most significant for applications where system goes to standby mode very often, e.g. PDAs.
- **Power**: Overall chip power dissipation increasing. Affect energy consumption, cooling costs, packaging costs.



Issues in Nano-CMOS Circuits

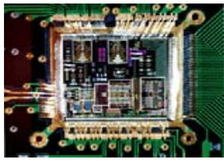
- **Thermals or Temperature**: Maximum temperature that can be reached by a chip during its operation is increasing. Affects reliability and cooling costs.
- **Reliability**: Circuit reliability is decreasing due to compound effects from variations, power, and thermals.
- **Yield**: Circuit yield is decreasing due to increased variability.



Variability: Origin and Sources

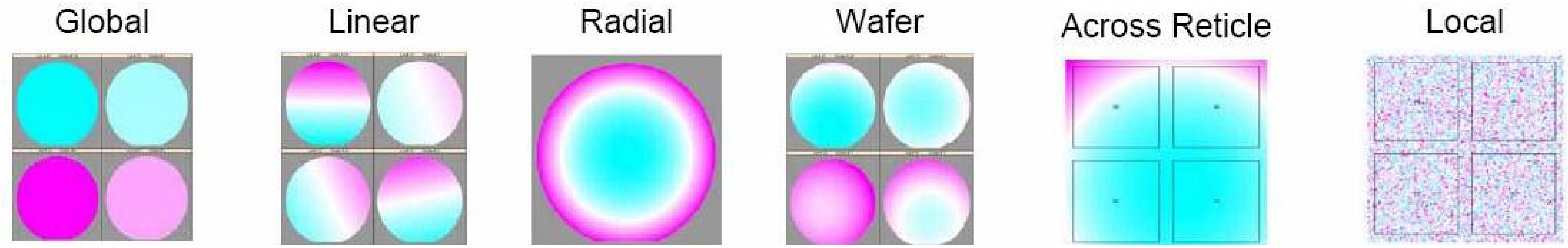
- Ion implantation
- Chemical mechanical polishing (CMP)
- Chemical vapor deposition (CVD)
- Sub-wavelength lithography
- Lens aberration
- Materials flow
- Gas flow
- Thermal processes
- Spin processes
- Microscopic processes
- Photo processes

Source: Singhal, DAC Booth 2007

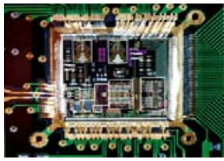


Variability: Types …

- **Parametric Variations**
  - **Wafer**
    - **Global**
    - **Linear**
      - **Caused by Materials/Gas Flow**
    - **Radial**
      - **Caused by Thermal/Spin Processes**
  - **Reticle**
    - **Caused by Photo Processes**
  - **Local**
    - **Caused by Random Microscopic Processes**



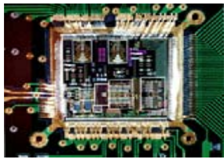
Source: Singhal, DAC Booth 2007



Variability: Types …

Global Variations

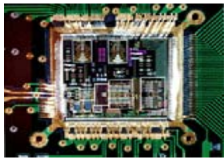
Fab Process	Lot Process	Wafer Process
From Plant to Plant	From Lot to Lot in a Plant	From Wafer to Wafer in a Lot



Variability: Types …

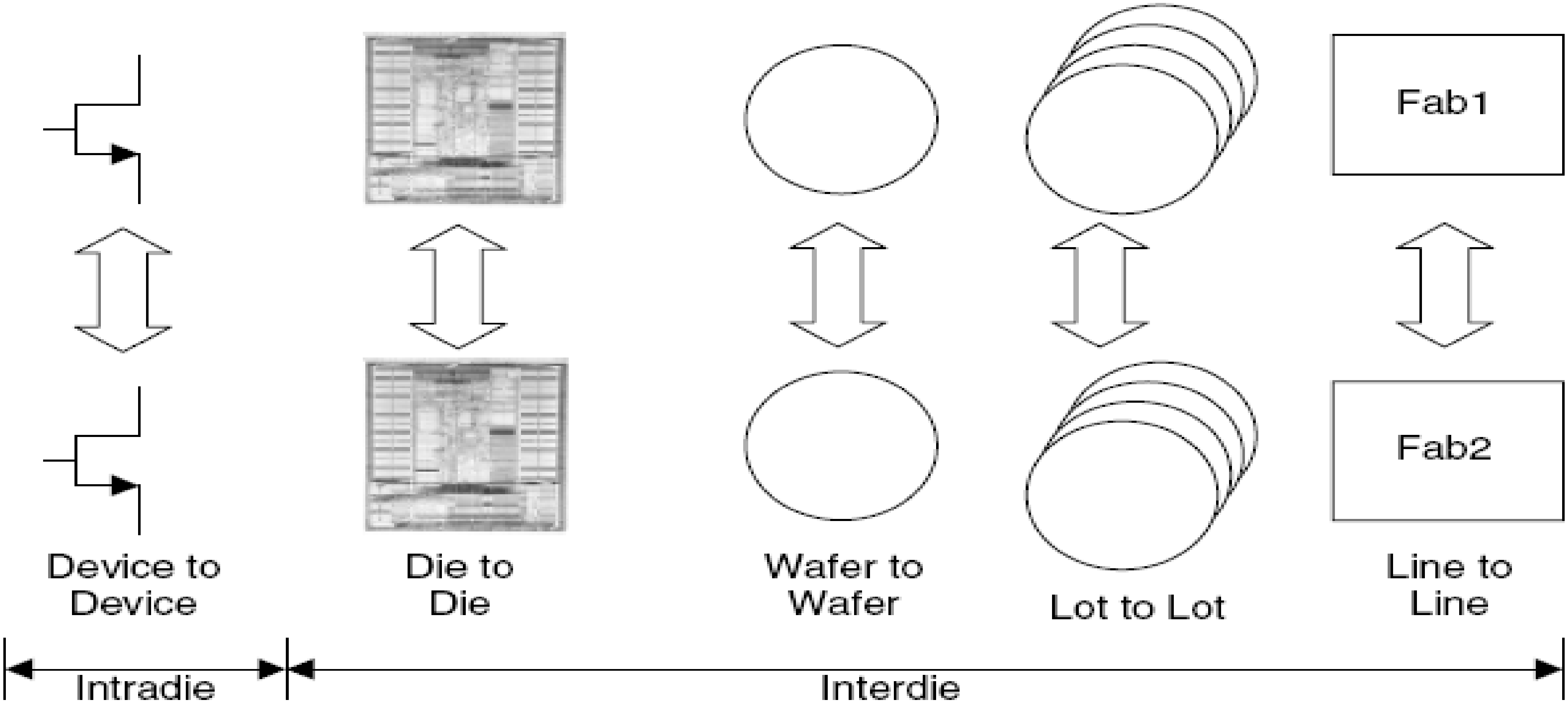
Variability Classifications

- Inter-Die or Intra-Die
- Random or Systematic
- Correlated or Uncorrelated
- Spatial or Temporal

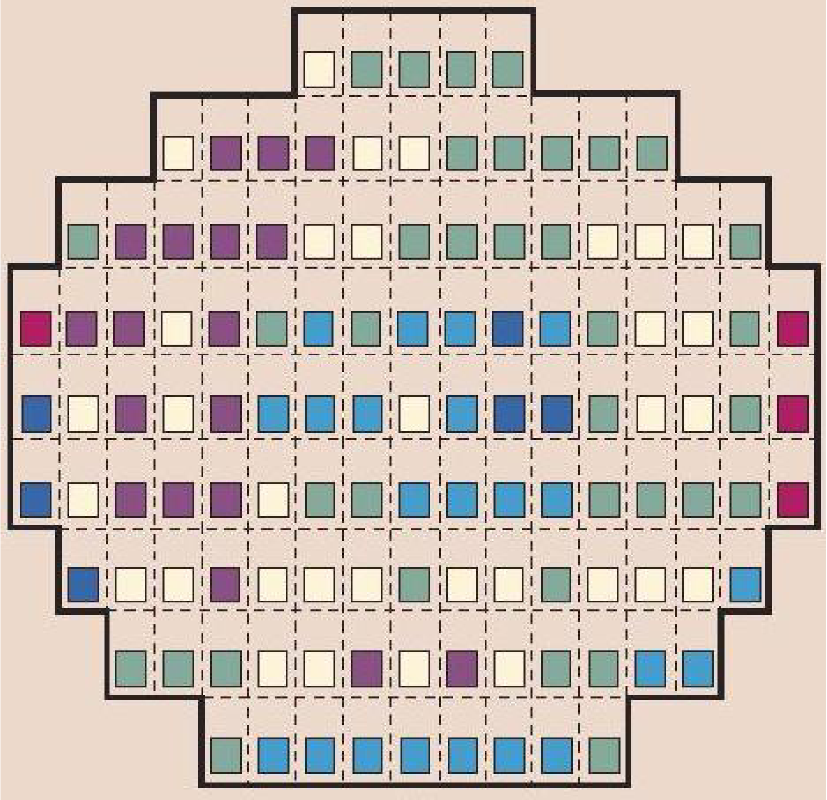


Variability: Types

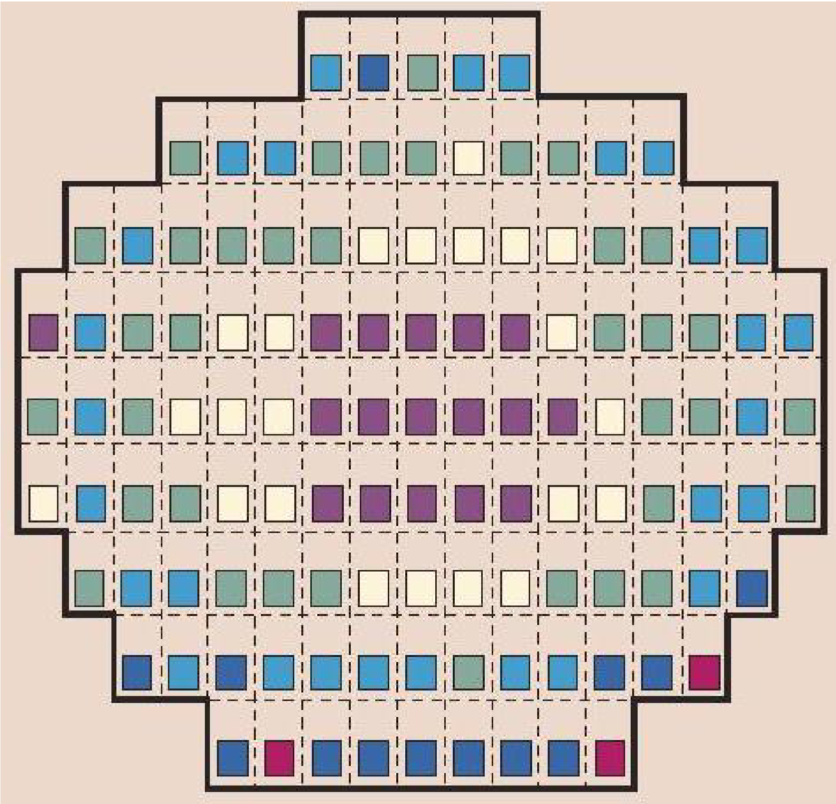
- Process variations are classified as:
  - Inter-die and Intra-die.



Variability: The Impact in a Wafer …



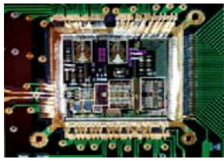
Source–drain resistance is different for different chips in a same die.



Gate-to-source and gate-to-drain overlap capacitance is different for different chips in a same die.

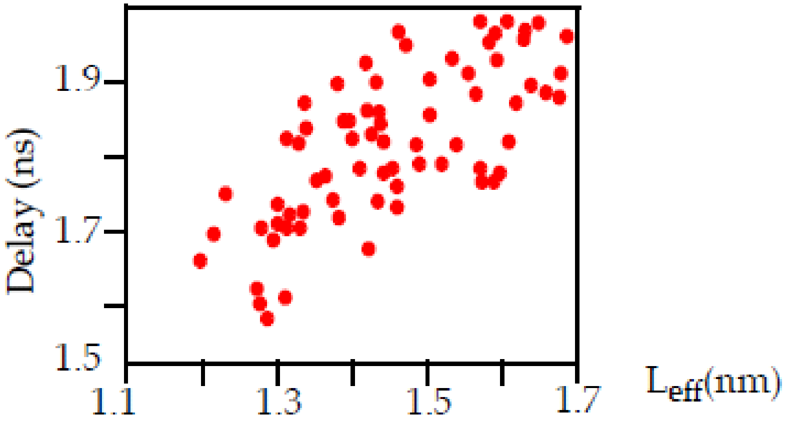
Fast 🟥 🟦 🟦 🟩 ⬜ 🟪 Slow

Source: Bernstein et al., IBM J. Res. & Dev., July/Sep 2006.



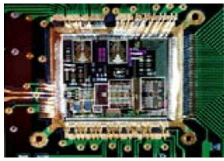
Variability: The Impact in a Wafer

- The impact of process variations is seen as design yield loss.
- Digital circuits are typically optimized for speed and power.
- Analog circuits are designed to meet as many as five to ten performance metrics.
- Variations in process parameters have a resounding effect on the performance metrics of analog/mixed-signal and RF circuits.
- Figure showing impact of effective transistor channel length on the speed of an adder cell.

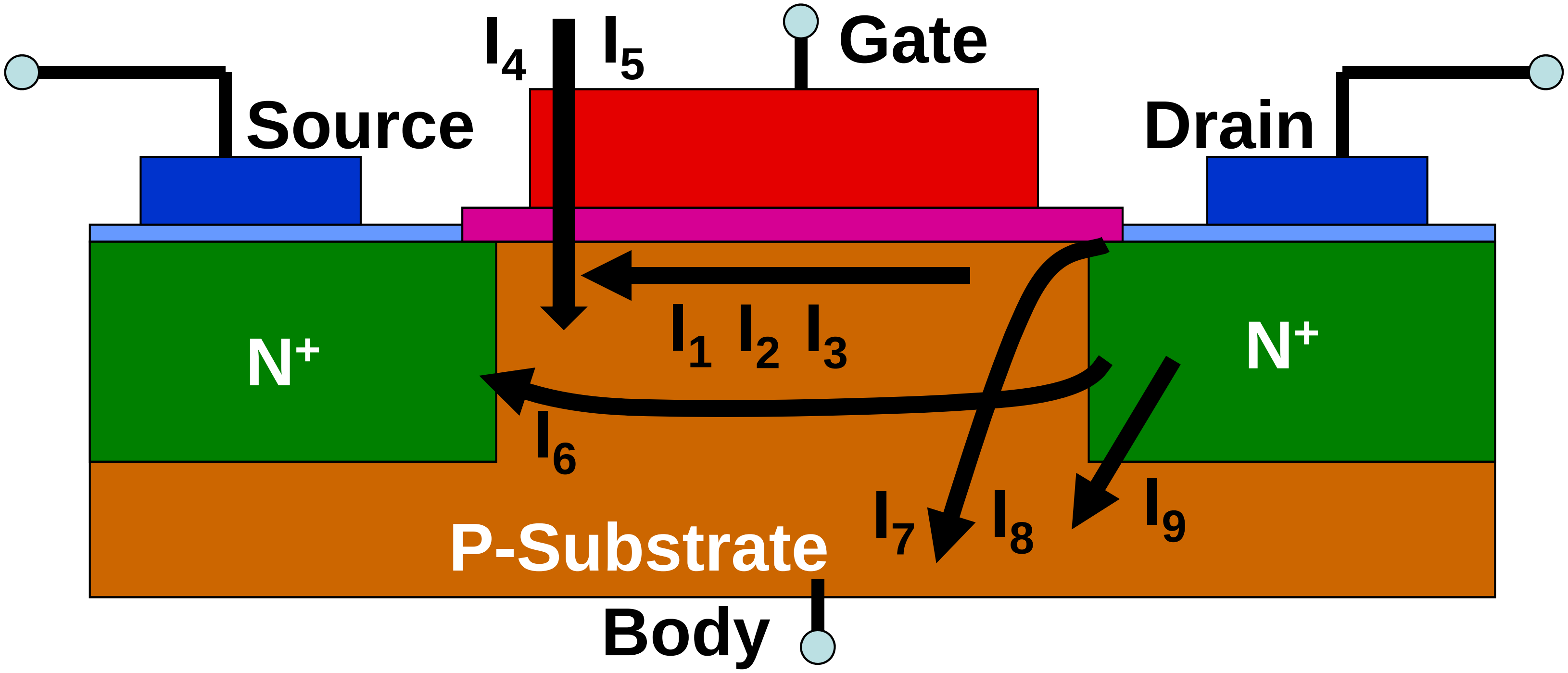


Variability: The 15 Device Parameters

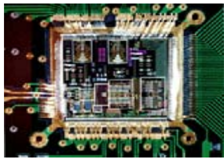
1) V_{DD}: supply voltage
2) V_{Thn}: NMOS threshold voltage
3) V_{Thp}: PMOS threshold voltage
4) t_{gaten}: NMOS gate dielectric thickness
5) t_{gatep}: PMOS gate dielectric thickness
6) L_{effn}: NMOS channel length
7) L_{effp}: PMOS channel length
8) W_{effn}: NMOS channel width
9) W_{effp}: PMOS channel width
10) N_{gaten}: NMOS gate doping concentration
11) N_{gatep}: PMOS gate doping concentration
12) N_{chn}: NMOS channel doping concentration
13) N_{chp}: PMOS channel doping concentration
14) N_{sdn}: NMOS source/ drain doping concentration
15) N_{sdp}: PMOS source/ drain doping concentration.



Power and Leakage …



I_1 : drain-to-source active current (ON state)
I_2 : drain-to-source short circuit current (ON state)
I_3 : subthreshold leakage (OFF state)
I_4 : gate Leakage current (both ON & OFF states)
I_5 : gate current due to hot carrier injection (both ON & OFF states)
I_6 : channel punch through current (OFF state)
I_7 : gate induced drain leakage (OFF state)
I_8 : band-to-band tunneling current (OFF state)
I_9 : reverse bias PN junction leakage (both ON & OFF states)

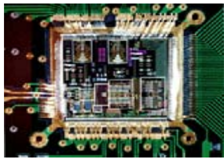


Power and Leakage

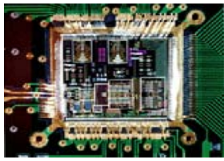
- The relative prominence of these components depend on:
  - Technology Node: 65nm, 45nm, or 32nm
  - Process : SiO_2/Poly or High-κ/Metal-Gate

- SiO_2/Poly
  - Dynamic
  - Subthreshold
  - Gate
- **High-κ/Metal-Gate**
  - **Dynamic**
  - **Subthreshold**
  - **Gate-Induced Drain Leakage (GIDL)**

- BTBT tunneling is important for sub-45nm.

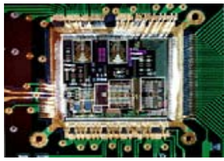


Challenges in The Context of HLS



High-Level Synthesis : An Effective Approach

- High-level synthesis (HLS) is defined as the translation from behavioral hardware description of chip to its register-transfer level (RTL) structural description.
- Allows exploration of design alternatives, including low power, prior to layout of the circuit in actual silicon.
- An efficient way to cope with system design complexity.
- Can facilitate early design verification.
- Can increase design reuse.



Nano-CMOS HLS: Goal

- Variability-driven statistical HLS is stated as: Given an unscheduled data flow graph (DFG), it is required to find a scheduled data flow graph with appropriate resource binding such that specified costs for the circuit are minimized statistically while accounting for variability and satisfying constraints.
- The resource, latency, and/or yield constrained optimization problem can be formulated as follows:

$$\text{Minimize: } PDF_{Cost,\ DFG}\ (Mean,\ Variance) \quad \ldots \quad (1)$$

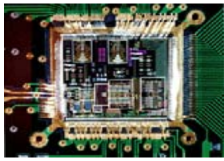
such that following resource, latency, and yield constraints, are satisfied:

$$Allocated\ (FU_{k,i}) \ \leq\ Available\ (FU_{k,i}),\ \ for\ each\ cycle\ c \quad \ldots \quad (2)$$

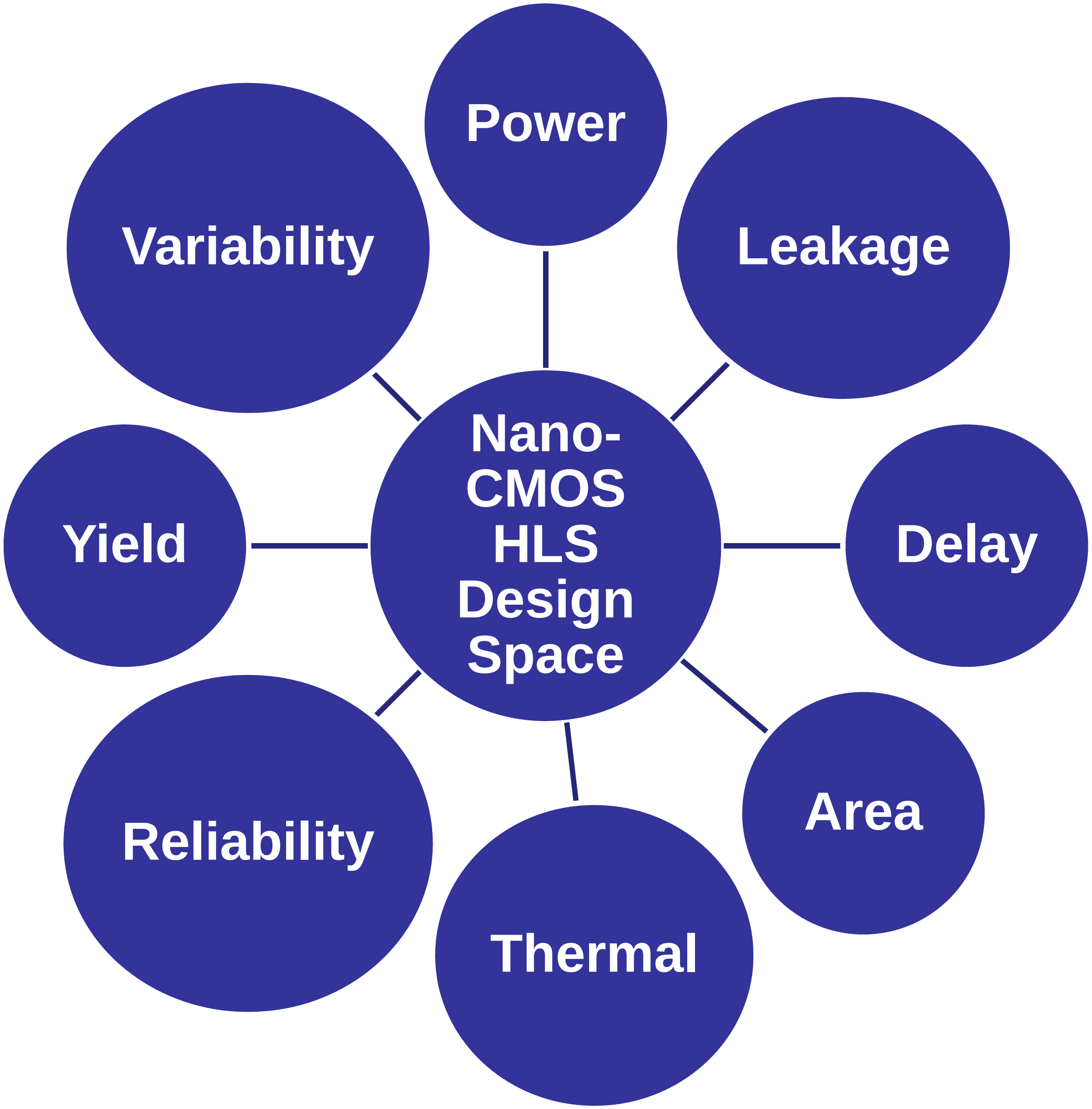
$$Expected\ [PDF_{DFG,\ Delay,\ Critical}\ (Mean,\ Variance)] \leq Delay_{DFG,\ Target} \quad (3)$$

$$Yield_{Circuit} \ \geq Yield_{Target} \quad \ldots \quad (4)$$

NOTE: PDF is probability density function.



Nano-CMOS HLS: Design Space



Nano-CMOS HLS Design Space

Power

Leakage

Delay

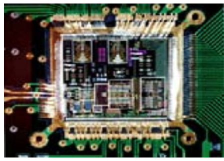
Area

Thermal

Reliability

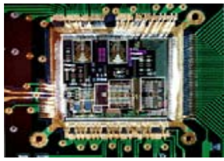
Yield

Variability

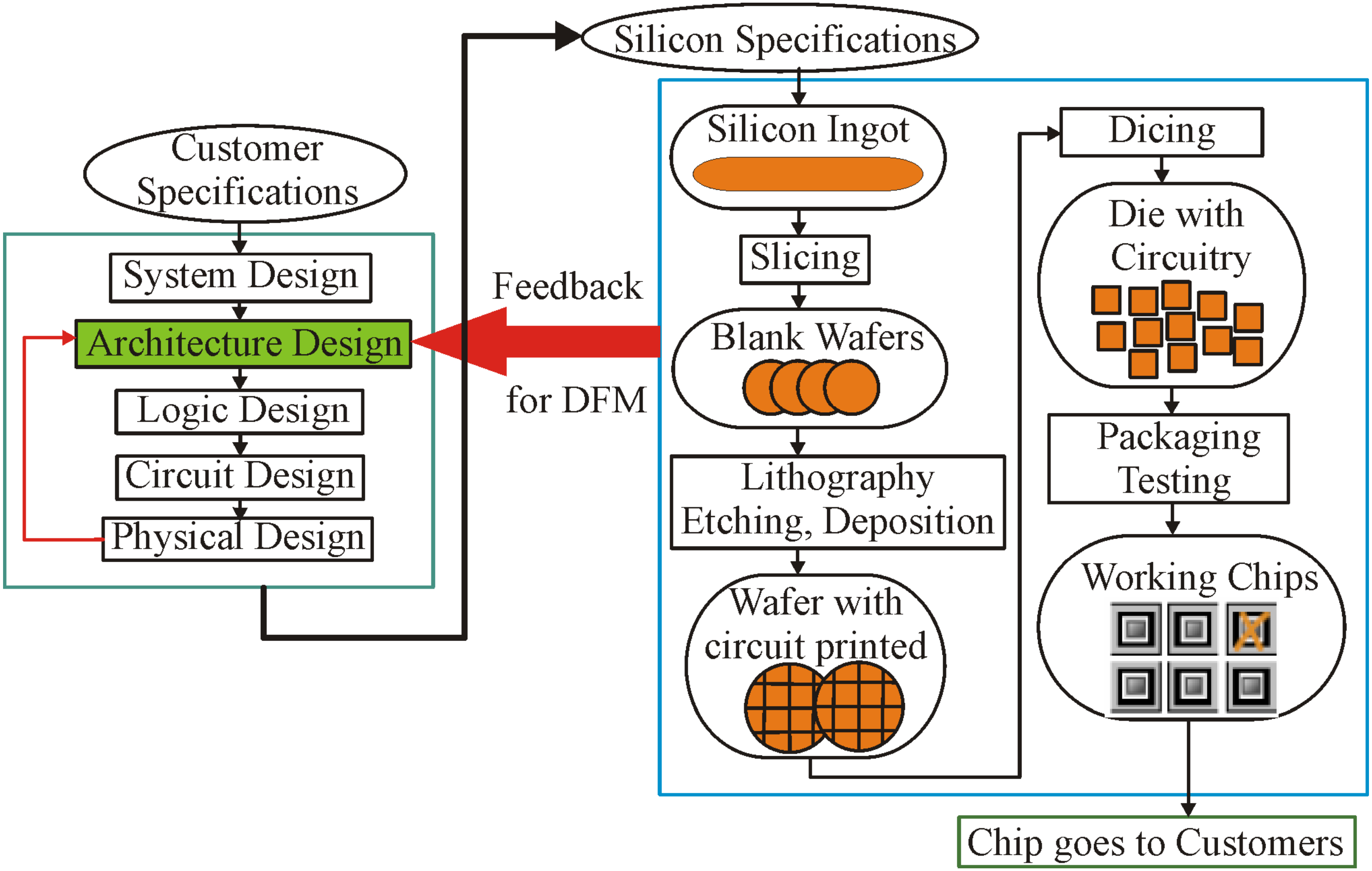


Nano-CMOS HLS: Challenges

- Unified consideration of axes of design space exploration for trade-offs.
- Determination of statistical models for variability of different nano-CMOS technologies.
- Propagation of the statistics to different levels of circuit abstraction.
- Performing statistical modeling of power, leakage, and delay for different RTL components.
- Estimating power, leakage, delay, area, and yield be estimated during HLS in the presence of variations.



Nano-CMOS HLS: Feedback Needed



Customer Specifications

System Design

Architecture Design

Logic Design

Circuit Design

Physical Design

Feedback for DFM

Silicon Specifications

Silicon Ingot

Slicing

Blank Wafers

Lithography Etching, Deposition

Wafer with circuit printed

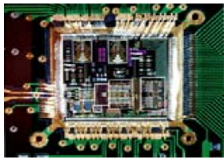
Dicing

Die with Circuitry

Packaging Testing

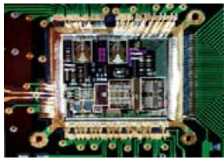
Working Chips

Chip goes to Customers

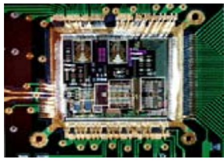


Nano-CMOS HLS: Questions

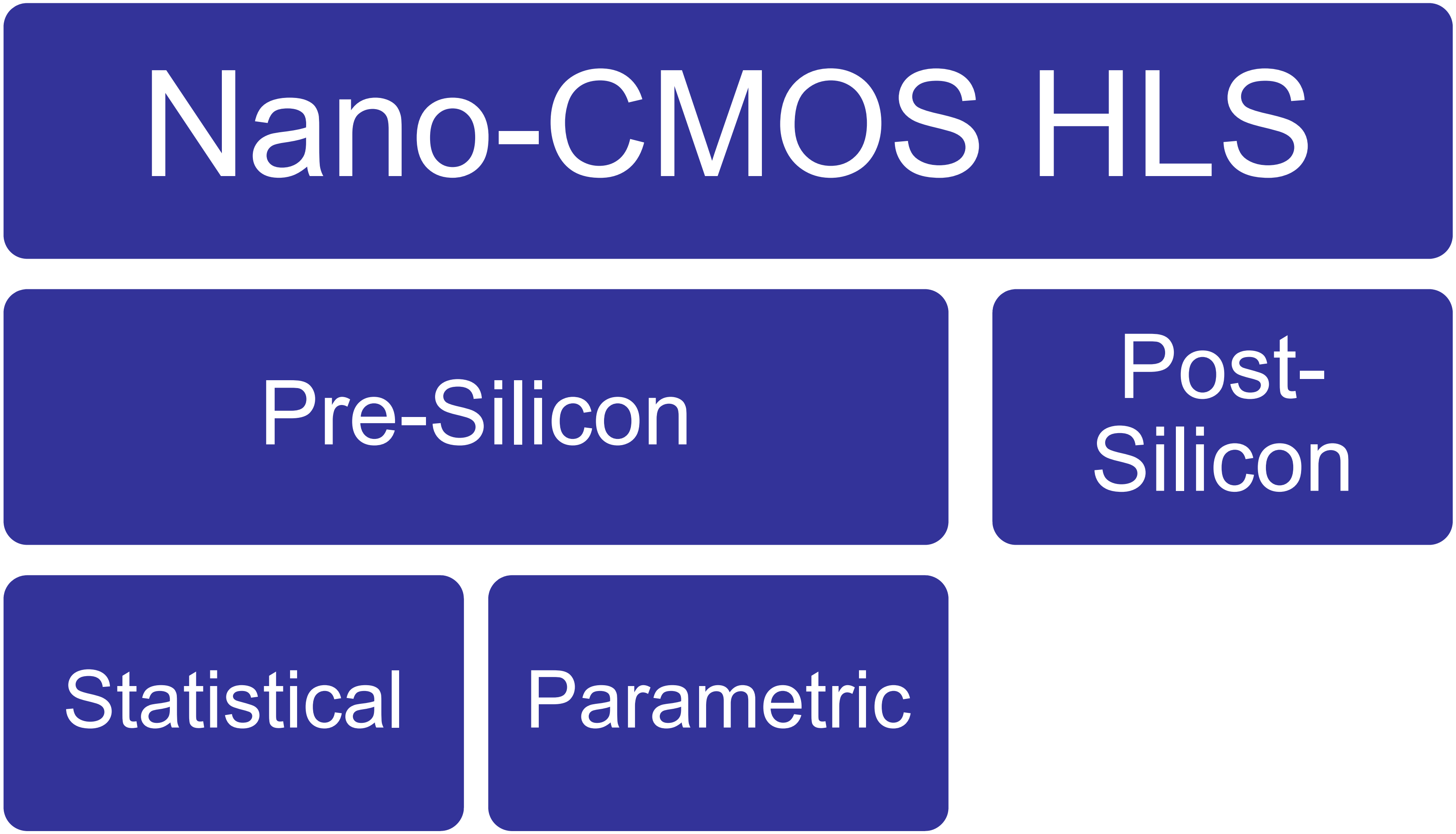
- How do the HLS phases (e.g. scheduling, binding) affect power, leakage, area, and yield in presence of variations?
- How do we judiciously consider design corners (e.g. V_{DD}, V_{Th}) to obtain a global power, leakage, and performance optimal circuit for given circuit constraints (from specifications)?



Proposed Approaches



Nano-CMOS HLS : Approaches



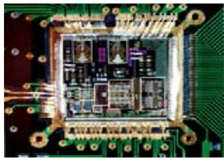
Nano-CMOS HLS

Pre-Silicon

Post-Silicon

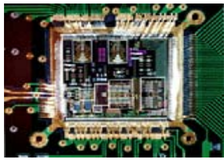
Statistical

Parametric

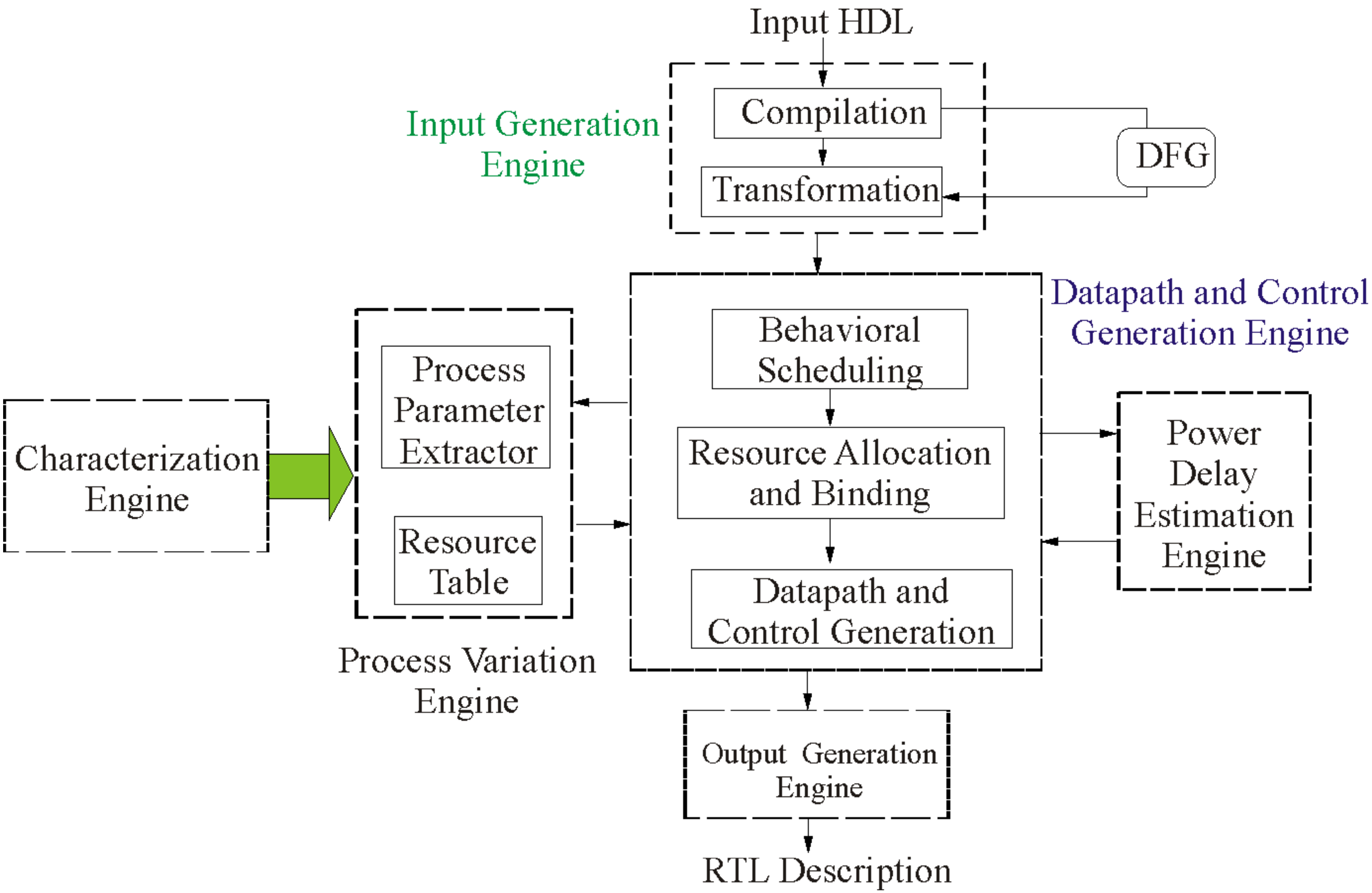


Statistical Nano-CMOS HLS for Power and Leakage

Source: S. P. Mohanty and E. Kougianos, "Simultaneous Power Fluctuation and Average Power Minimization during Nano-CMOS Behavioral Synthesis", in *Proceedings of the 20th IEEE International Conference on VLSI Design (VLSID)*, pp. 577-582, 2007.



Proposed Statistical Nano-CMOS HLS Framework



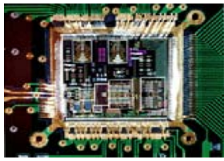
Statistical HLS : Formulation

Minimize: $I_{Total}^{DFG}\left(\mu_I^{DFG},\sigma_I^{DFG}\right)$

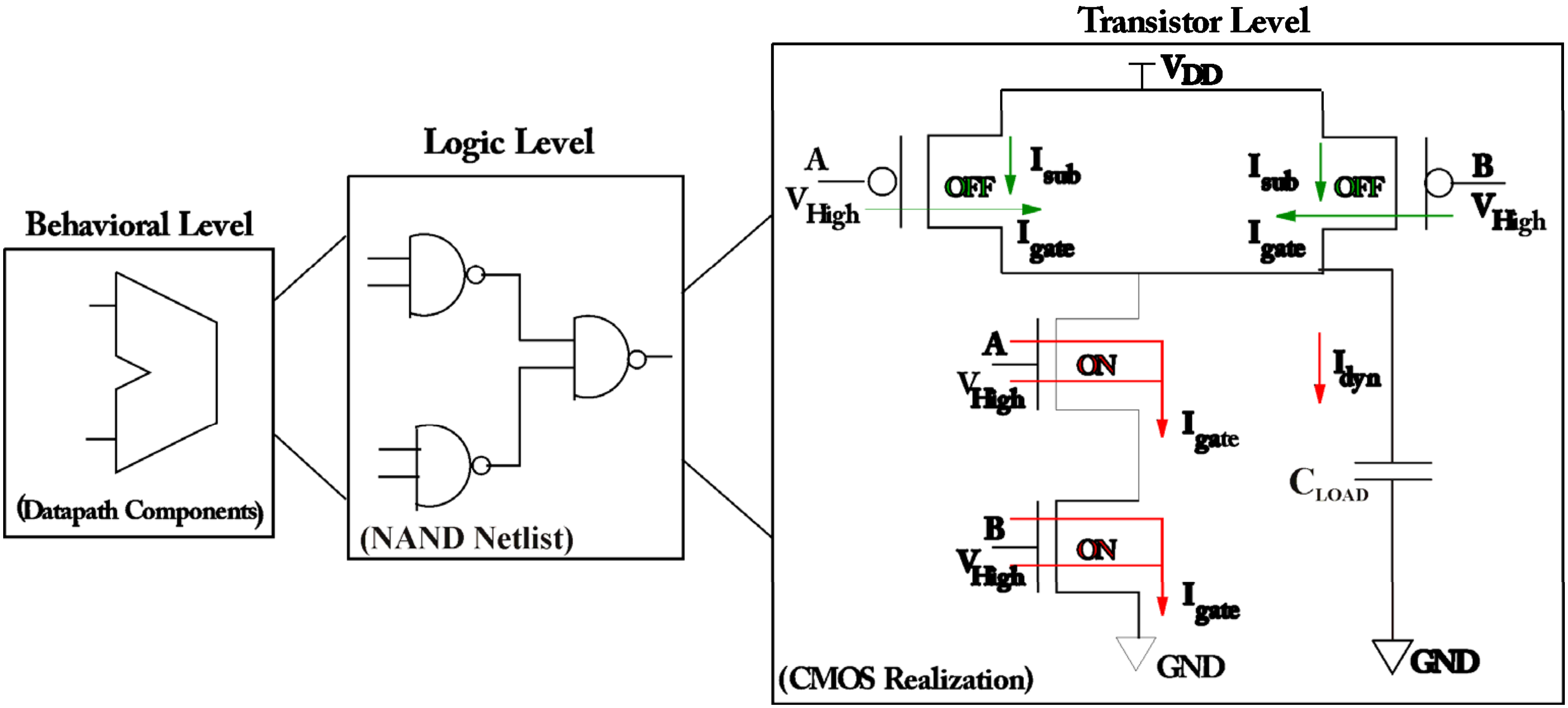
Subjected to (Resource/Time Constraints):

$$\text{Allocated}\left(FU_{k,i}\right) \le \text{Available}\left(FU_{k,i}\right), \forall \text{cycle } c$$

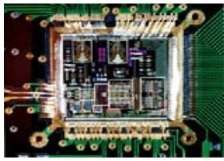
$$D_{CP}^{DFG}\left(\mu_D^{DFG},\sigma_D^{DFG}\right) \le D_{Con}\left(\mu_D^{Con},\sigma_D^{Con}\right)$$



Statistical HLS : Library …

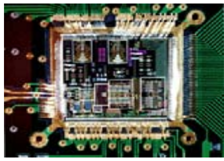


- 3 level hierarchical approach.



Statistical HLS : Library …

- It is assumed that resources such as adders, subtractors, multipliers, dividers, are constructed using 2-input NAND.
- There are total N 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.



Statistical HLS : Library …

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

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

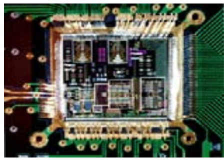
$$I_{sub}^{FU} = \text{Statistical Summation over N} \left(I_{sub}^{NAND}\right)$$

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

- The PDF of delay can be calculated as:

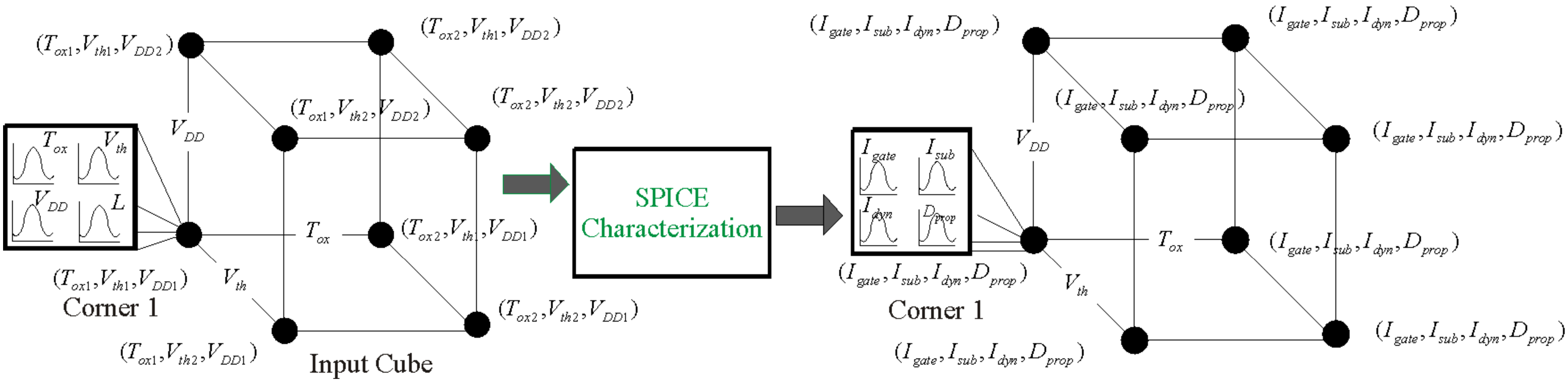
$$D_{prop}^{FU} = \text{Statistical Summation over } N_{CP} \left(D_{prop}^{NAND}\right)$$

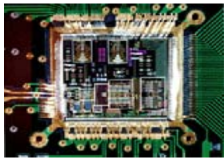
- Correlation needs to be considered.



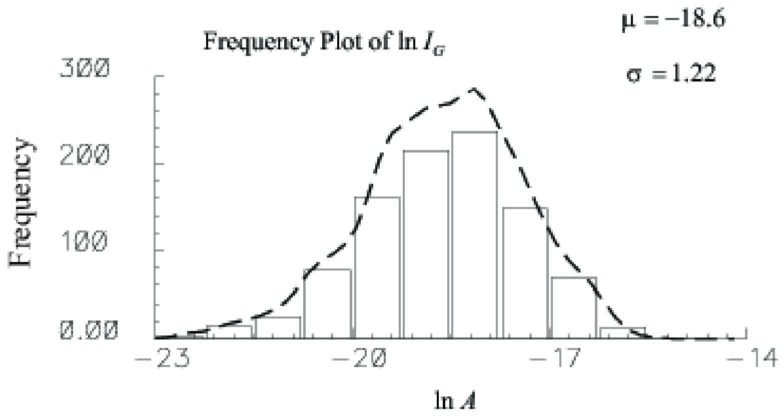
Statistical HLS : Library …

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

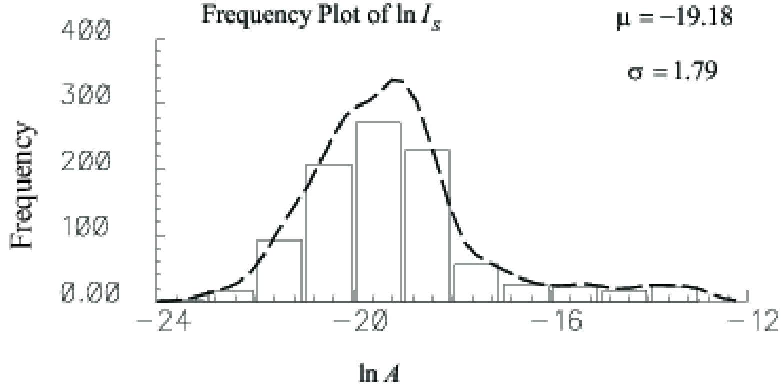




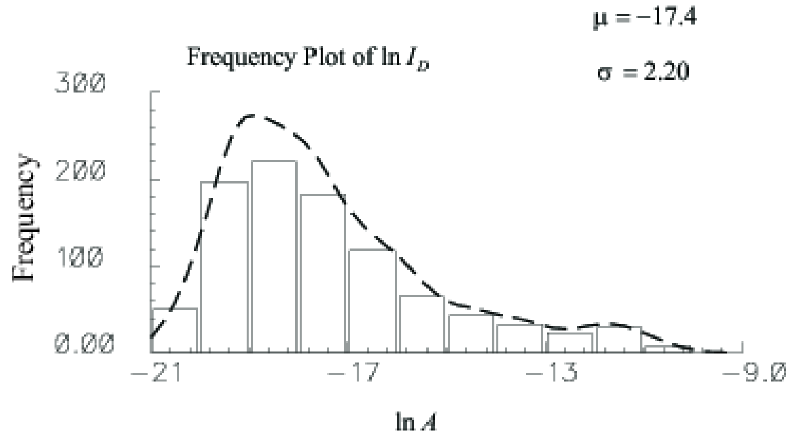
Statistical HLS : Library …
(PDFs of Currents and Delay)



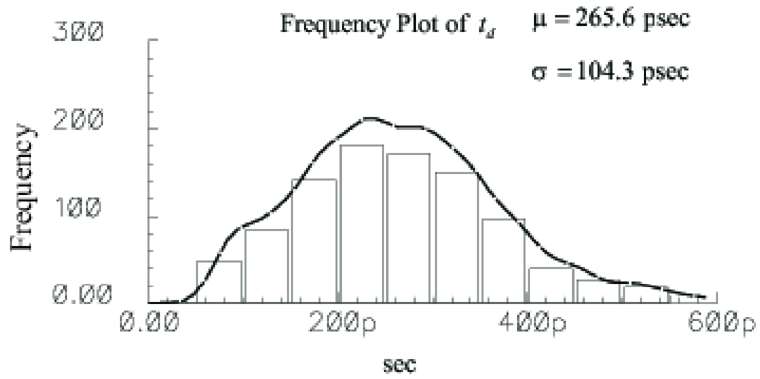
Gate leakage current



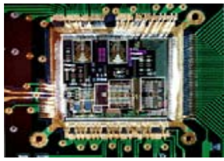
Subthreshold leakage current



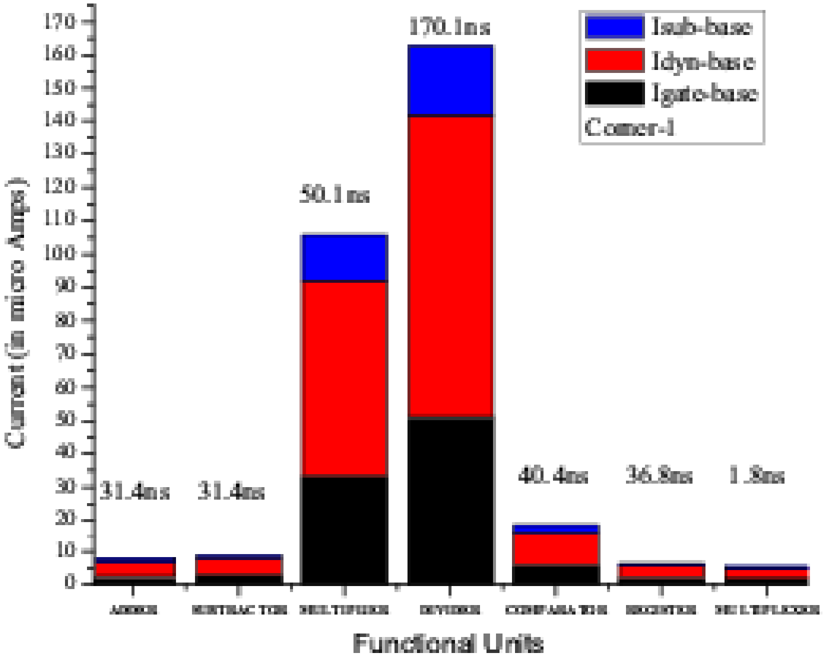
Dynamic current



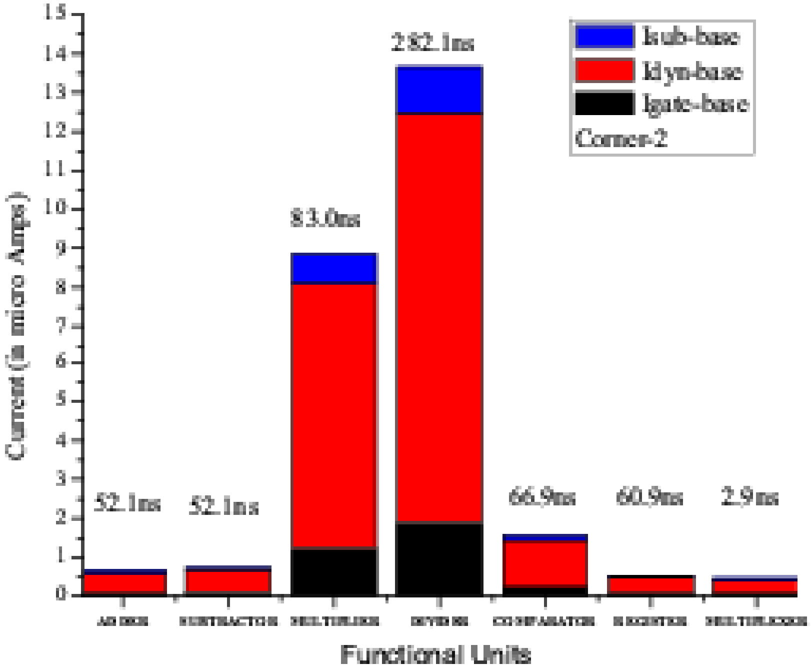
Propagation delay



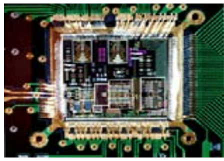
Statistical HLS : Library (Relative Contributions)



(Corner – 1)

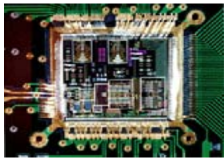


(Corner – 2)



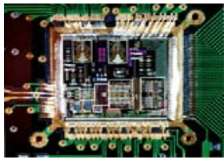
Statistical HLS : Optimization …

```
Simulated Annealing Algorithm (UDFG, Constraints, Library)
{
    Perform ASAP and ALAP scheduling.
    Temp = Initial Temperature.
    While there exists a schedule with available resources.
        i = Number of iterations.
        Perform resource constrained ASAP and  ALAP.
        Initial Solution ← ASAP Schedule.
        S ← Allocate-Bind().
        Initial Cost ← Statistical-Cost(S).
        While (i > 0)
            Generate random transition from S to S*.
            Δ-Cost ← Statistical-Cost(S*) − Statistical-Cost(S).
            if{ (Δ-Cost > 0) or ( $e^{\Delta\text{-Cost}/Temp}$ > random[0,1) ) } then S ← S*.
            i ← i − 1.
        end While
        Decrement available resources.
        Temp ← Cooling Rate x Temp.
    end While
    return S.
}
```

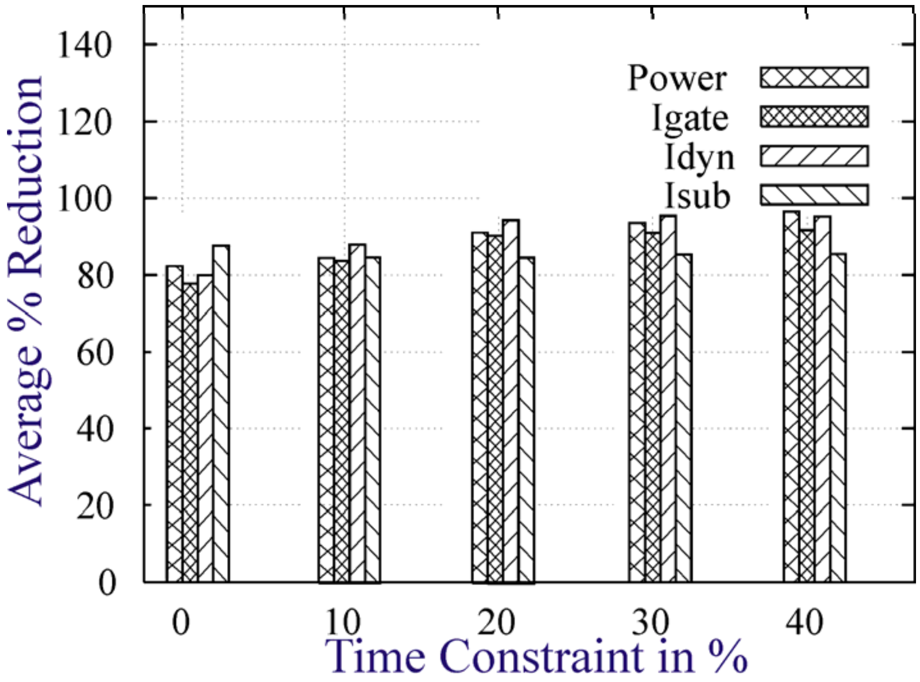


Statistical HLS : Optimization

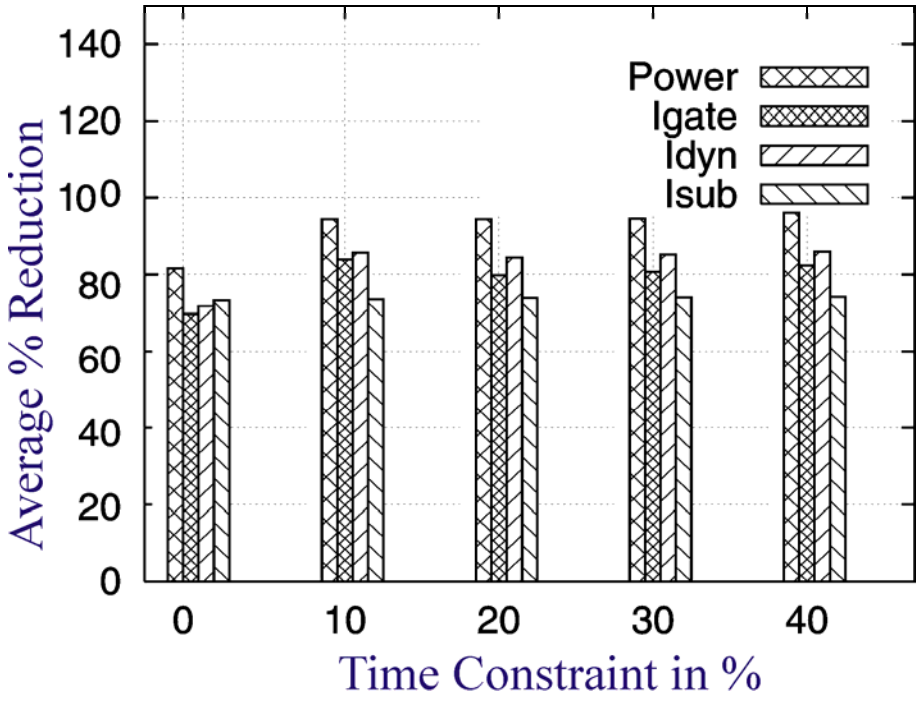
```
Statistical-Cost (S, Library)
{
    $I_{dyn}^{c}$ = Statistical Summation over all FU in c $\left(I_{dyn}^{FU}\right)$
    $I_{sub}^{c}$ = Statistical Summation over all FU in c $\left(I_{sub}^{FU}\right)$
    $I_{gate}^{c}$ = Statistical Summation over all FU in c $\left(I_{gate}^{FU}\right)$
    $I_{total}^{c}$ = Statistical Summation $\left(I_{dyn}^{c}, I_{sub}^{c}, I_{gate}^{c}\right)$
    $I_{total}^{DFG}$ = Statistical Summation over all cycles $\left(I_{total}^{c}\right)$
    $Cost_{I}^{DFG} = \mu_{I}^{DFG} + 3 \times \sigma_{I}^{DFG}$
    Similarly calculate delay cost $Cost_{D}^{DFG}$ of the DFG.
    $Cost = Cost_{I}^{DFG} \times Cost_{D}^{DFG}$
Return Cost.
}
```



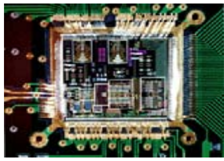
Statistical HLS : Results



(For ARF Benchmark)



(For BPF Benchmark)



Parametric Nano-CMOS HLS for Leakage

Source: S. P. Mohanty, R. Velagapudi, and E. Kougianos, "Physical-Aware Simulated Annealing Optimization of Gate Leakage in Nanoscale Datapath Circuits", in *Proc. 9th IEEE International Conference on Design Automation and Test in Europe (DATE)*, pp. 1191-1196, 2006.

Parametric HLS : Formulation

Minimize: $I_{Total}^{DFG}\left(\text{Parameters} : \kappa, T_{gate}, V_{Th}, V_{DD}, L_{eff}, W\right)$

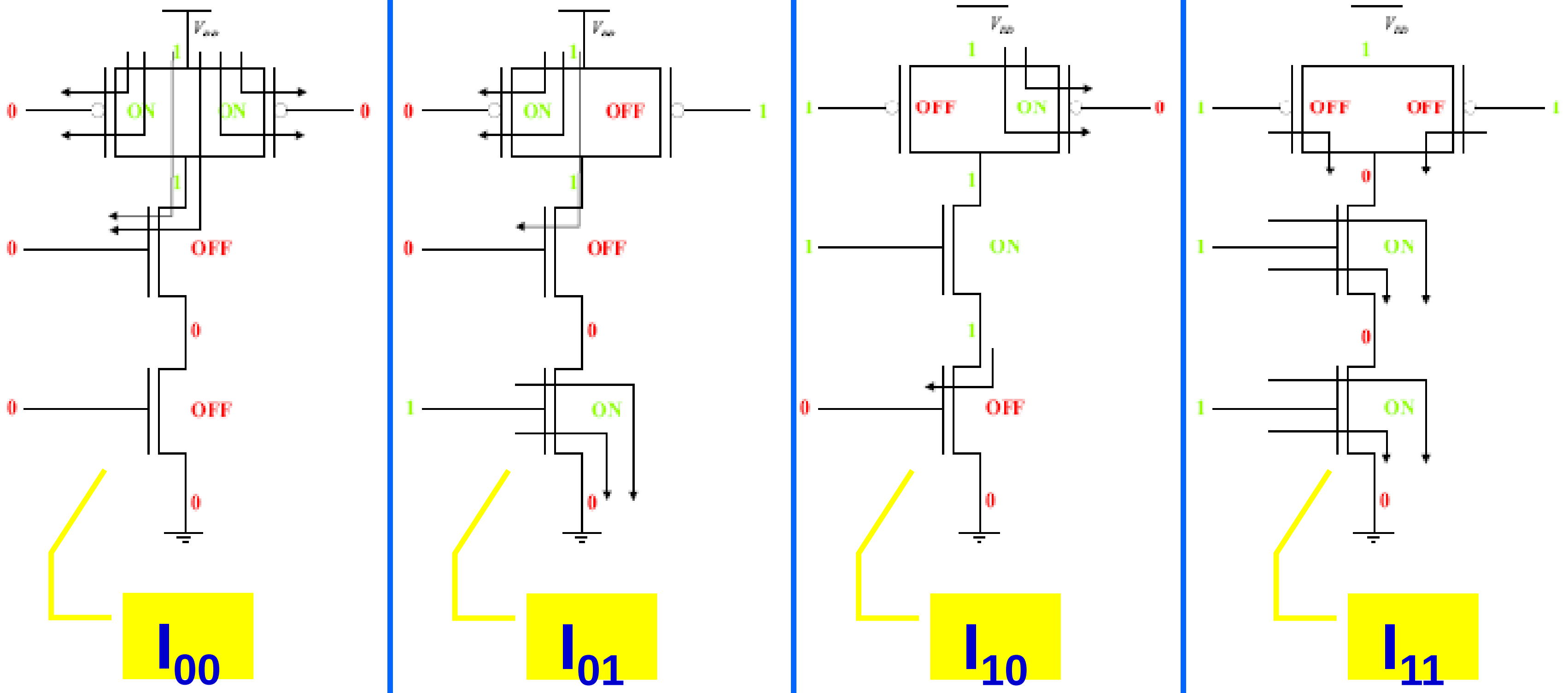
Subjected to (Resource/Time Constraints):

$$\text{Allocated}\left(FU_{k,i}\right) \leq \text{Available}\left(FU_{k,i}\right), \forall \text{cycle } c$$

$$D_{CP}^{DFG}\left(\text{Parameters} : \kappa, T_{gate}, V_{Th}, V_{DD}, L_{eff}, W\right) \leq D_{Con}$$

Parametric HLS : Library …

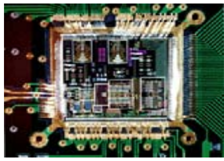
input 00	input 01	input 10	input 11



I_{00}	I_{01}	I_{10}	I_{11}
(State 1)	(State 2)	(State 3)	(State 4)

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

(Assuming all states to be equiprobable.)



Parametric HLS : Library …

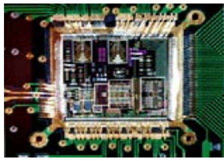
- We calculate the direct tunneling current (I_{oxFU}) of an *n*-bit functional unit as:

$$I_{ox\ FU} = \sum_{i=1}^{N} I_{ox\ NANDi}$$

where $I_{oxNANDi}$ is the *average gate oxide tunneling current* dissipation of the i^{th} 2-input NAND gate in the functional unit, assuming all states to be equiprobable.

- Similarly, the propagation delay and silicon area of an *n*-bit functional unit are

$$T_{pd\ FU} = \sum_{i=1}^{N_{CP}} T_{pd\ NANDi} \qquad A_{FU} = \sum_{i=1}^{N} A_{NANDi}$$



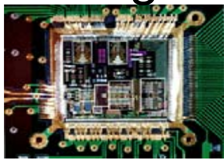
Parametric HLS : Library …

- At logic level we used BPTM BSIM4 models for analog simulation to find I_{ox} and T_{pd}.
- Due to unavailability of silicon data we used an analytical estimate for area calculations.

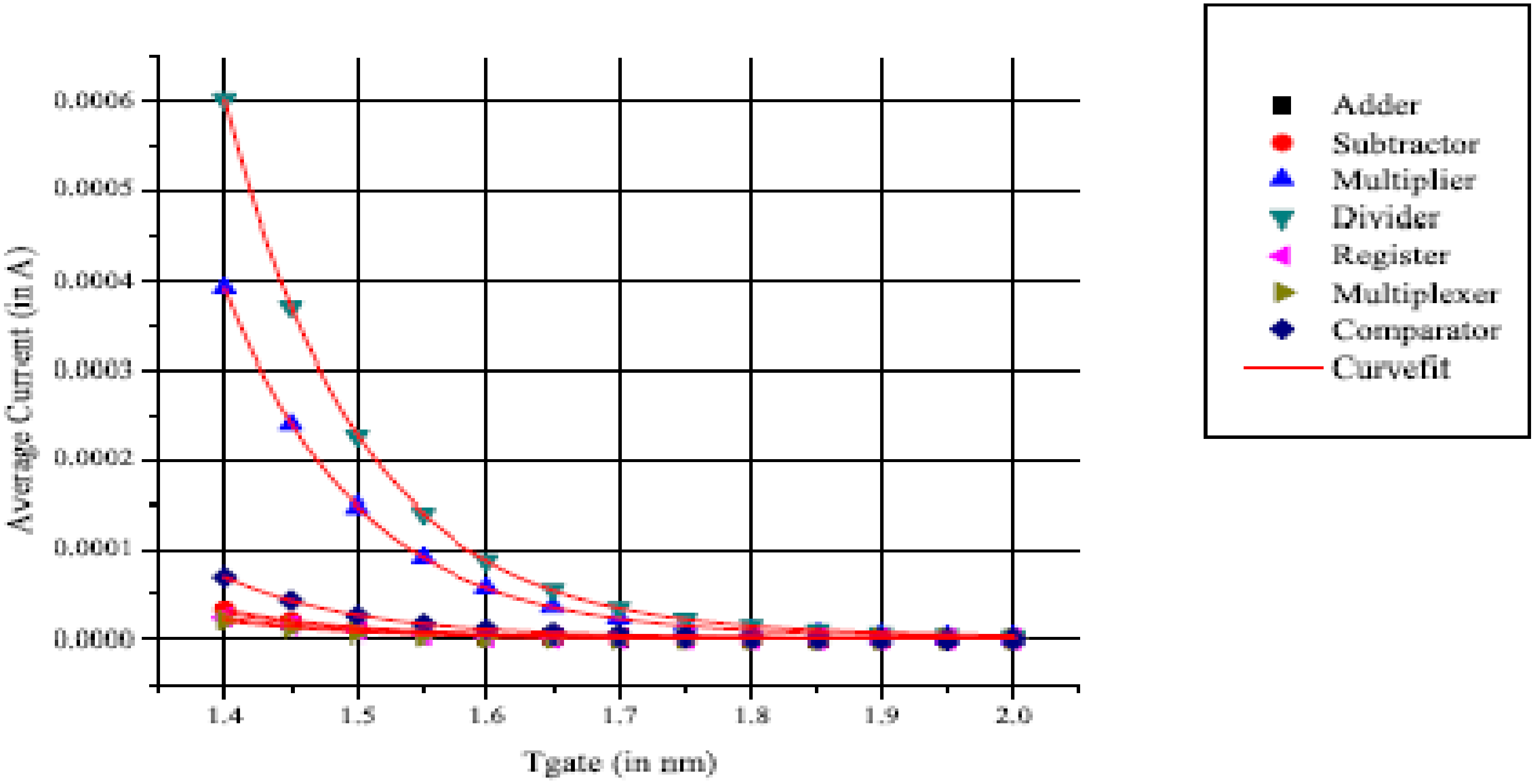
$$A_{NAND} = K_{inv}\left(1 + 4(n_{in} - 1)\sqrt{\frac{AR_{NAND}}{K_{inv}}}\right) * \left(1 + \frac{\left(\frac{W_{NMOS}}{f} - 1\right)(1 + \beta_{NAND})}{\sqrt{K_{inv}AR_{NAND}}}\right)$$

where,
W_{NMOS} = NMOS width,
f = Minimum feature size for a technology,
k_{inv} = Area of minimum size inverter with respect to f^2,
AR_{NAND} = aspect ratio of NAND gate,
n_{in} = number of inputs, and
β_{NAND} = ratio of PMOS width to NMOS width.

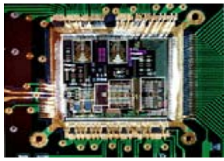
Source: Bowman TED 2001 Aug



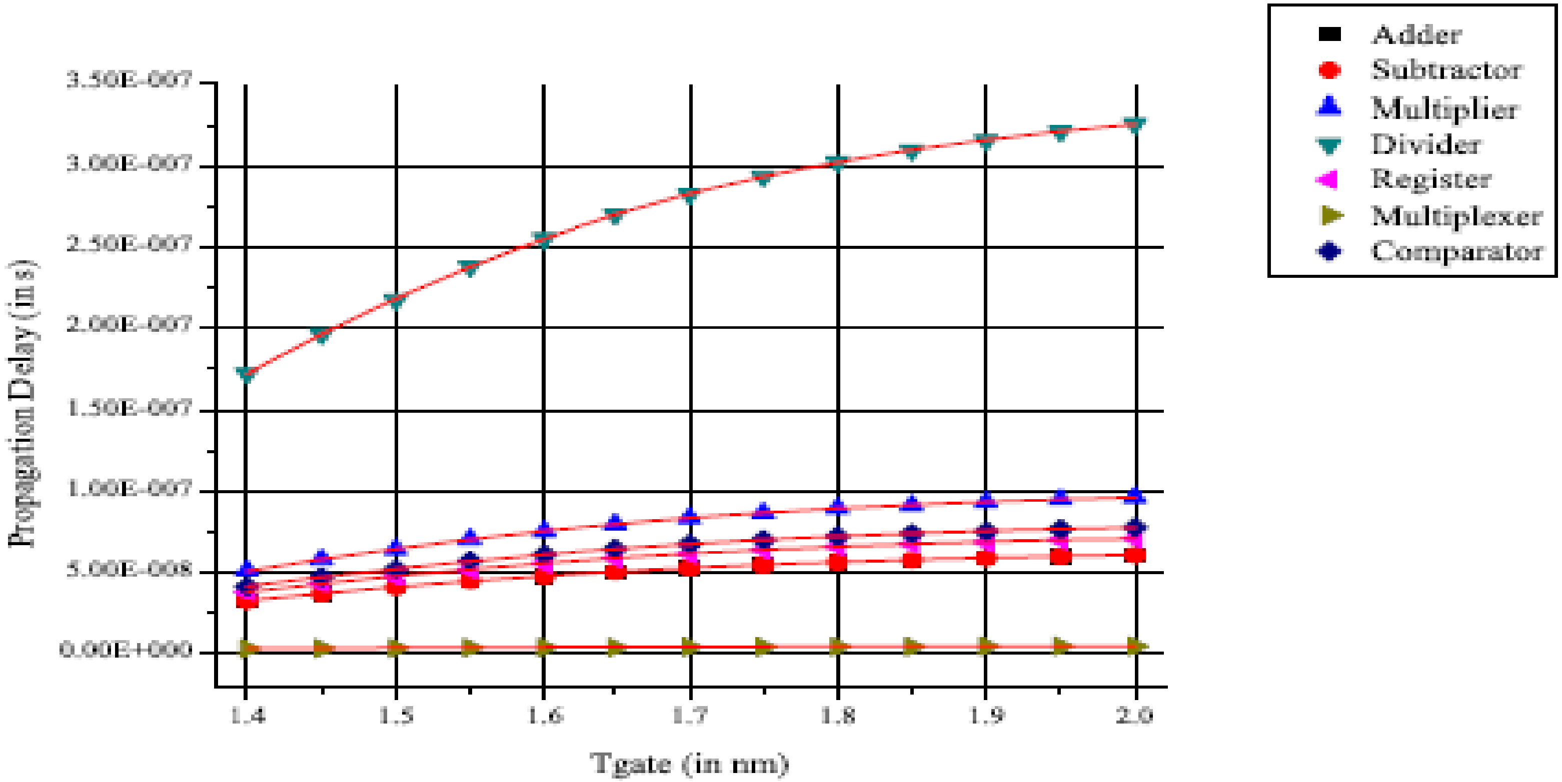
Parametric HLS : Library …



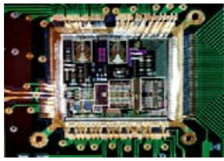
$$I_{ox}(\mu A) = A\exp\left(-T_{ox}/\alpha\right) + \beta$$



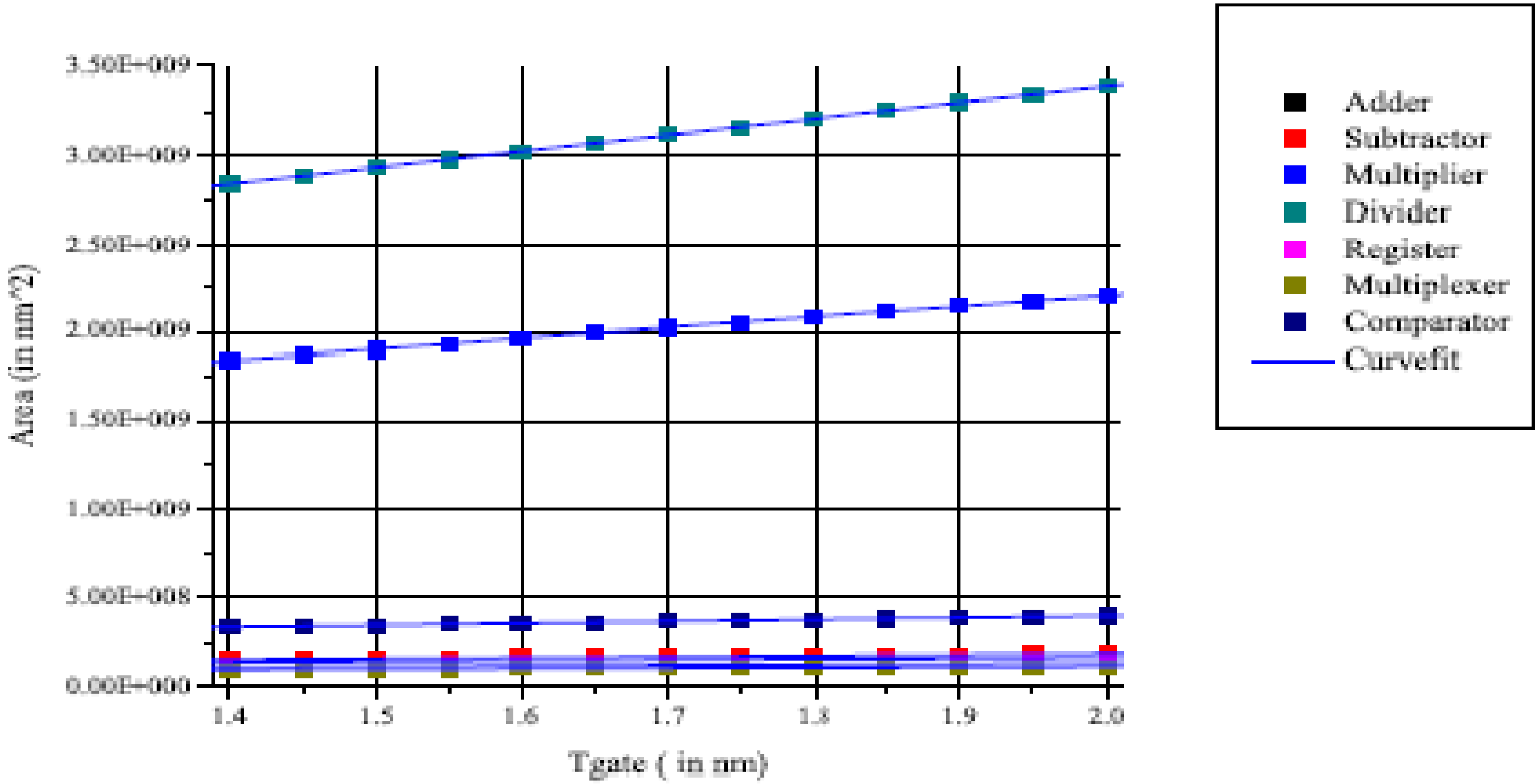
Parametric HLS : Library …



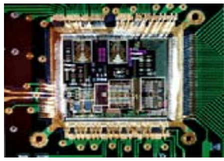
$$T_{pd}(ns) = \left({(A_1 - A_2)} \Big/ {\left(1 + \left({T_{ox}}/{\beta}\right)^{\nu}\right)} \right) + A_2$$



Parametric HLS : Library



$$A\left(nm^{2}\right) = \alpha T_{ox} + \beta$$

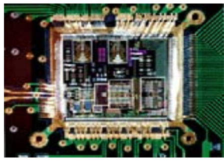


Parametric HLS : Optimization …

- The objective is to reduce both the gate leakage and area of the circuit for given time constraints.
- The objective function used by the optimization algorithm is:

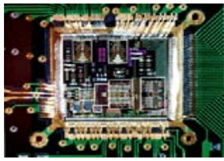
$$\text{Cost} = a*I_{ox} + b*A$$

- I_{ox} of the circuit is calculated as the sum of tunneling current of all the nodes in the circuit. A is the sum of areas of all the allocated resources. 'a' and 'b' are the weights of current and area respectively. 'a' and 'b' are chosen in such a way the effect of current and delay are normalized.

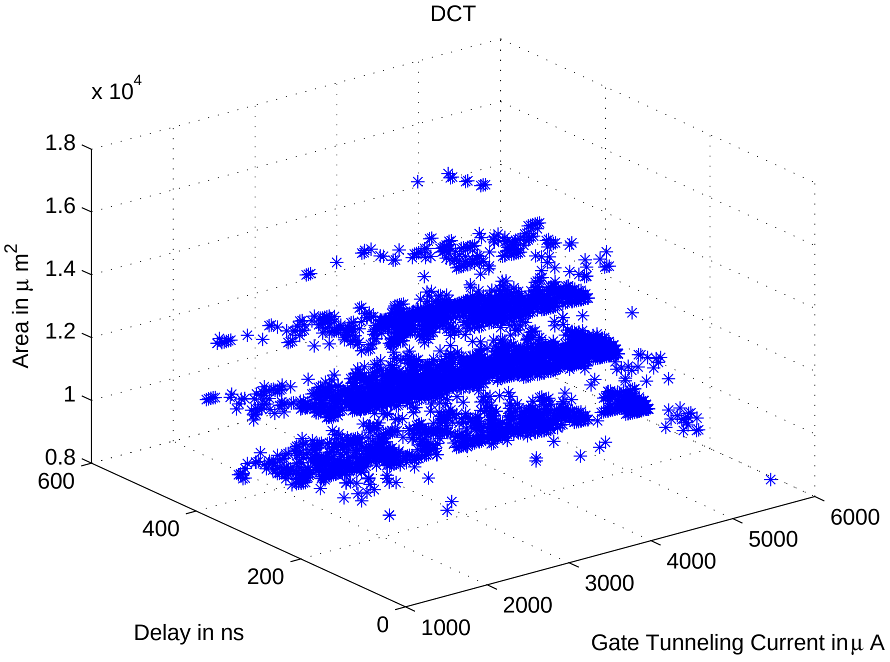


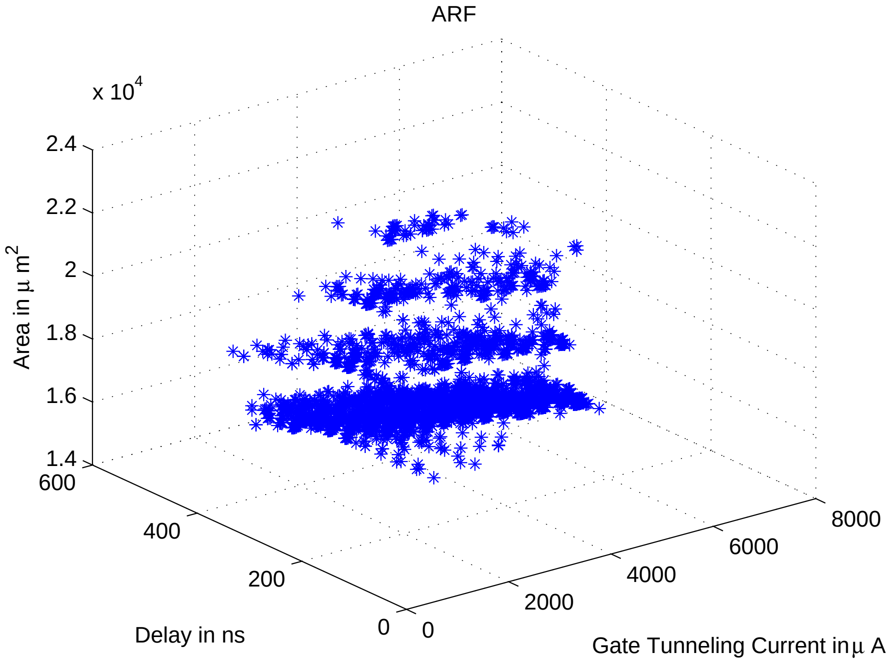
Parametric HLS : Optimization …

```
Initial Temperature $t \leftarrow t_o$ and available Resources ← Resource constraints.
While there exists a schedule with available resources.
    $i$ = Number of iterations.
    Perform resource constrained ASAP and resource constrained ALAP.
    Make initial Solution as ASAP Schedule.
    $S \leftarrow$ Allocate Bind() and Initial Cost ← Cost($S$).
    While ($i > 0$)
        Generate a random thicknesses in range of ($T_{ox}$ - $T_{oxL}$ $T_{ox}$ + $T_{ox}$)
        Generate random transition from $S$ to $S^*$.
        $\Delta C \leftarrow$ Cost($S$) - Cost($S^*$)
        if( $\Delta C > 0$ ) then $S \leftarrow S^*$.
        else if( $e^{\Delta C/t}$ > random[0,1) ) then $S \leftarrow S^*$.
        $i \leftarrow i - 1$.
    end While.
    Decrement available resources.
    $t \leftarrow$ Cooling Rate × $t$.
end While.
return $S$.
```

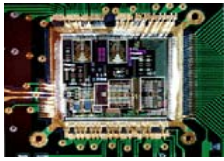


Parametric HLS : Optimization

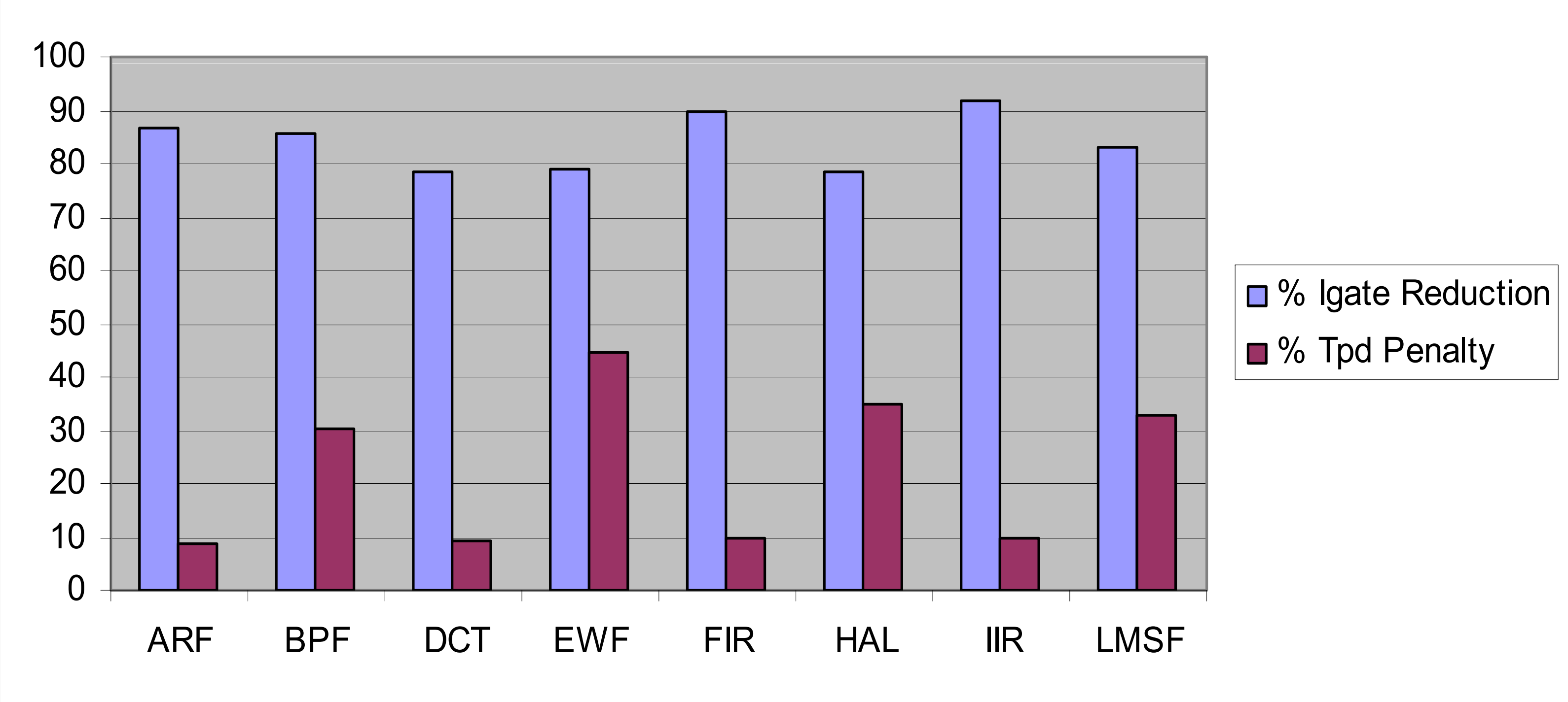




Each layer corresponds to a different resource constraint, each time the number of T_{oxH} multipliers are decreased a new layer is formed. We observed that the number of design corners reduces when we use more multipliers of T_{oxH} thickness, since delay increases and mobility of the nodes is restricted in order to satisfy the time constraint.



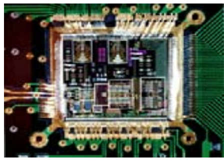
Parametric HLS : Results



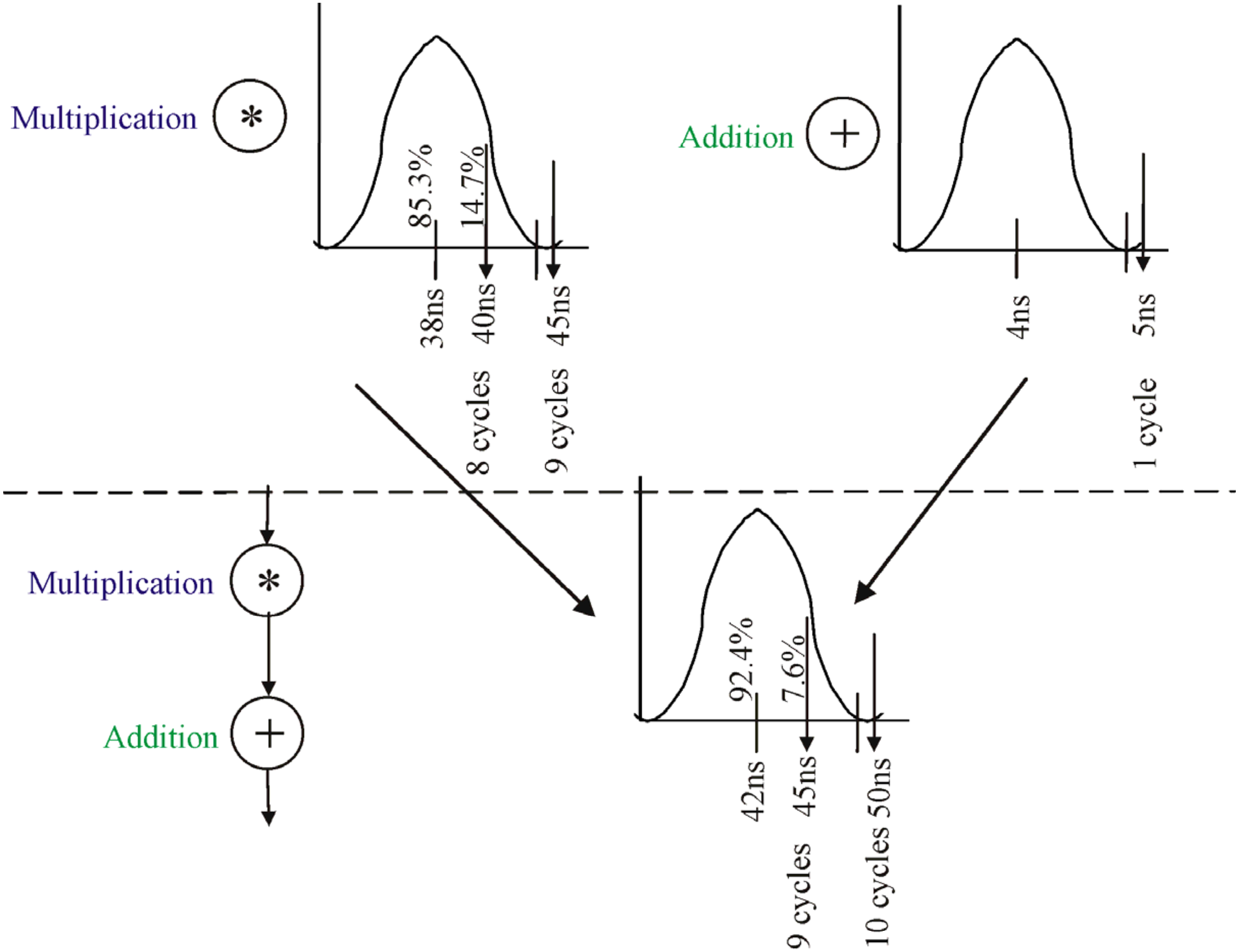
Results presented for different benchmarks for a delay trade-off factor of 1.4, T_{oxL} is 1.4nm and T_{oxH} is 1.7nm.

Statistical Nano-CMOS HLS for Timing

Source: Jongyoon Jung, Taewhan Kim, "Timing Variation-Aware High-Level Synthesis", in *Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2007, pp. 424-428.

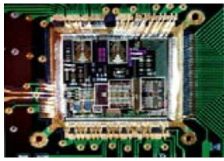


Statistical Timing HLS : Tradeoff



Statistical Timing HLS : Algorithm

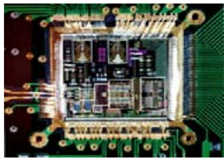
- Branch-and-bound algorithm for scheduling and binding.
- The search process is speeded up using window-based search.
- Window is maximum number of consecutive clock cycles satisfying resource constraints.



Statistical Timing HLS : Results

Results Compared Over Traditional List Scheduling

Benchmarks	Yield Constraint	Yield Obtained	Yield Penalty	Latency Reduction
Avg. of 4	90%	92.9%	7.1%	18.8%
Avg. of 4	80%	88.1%	11.9%	20.2%

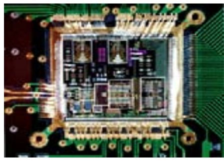


Statistical Nano-CMOS HLS for Post-Silicon Tuning

Source: Feng Wang, Xiaoxia Wu, and Yuan Xie, "Variability-Driven Module Selection With Joint Design Time Optimization and Post-Silicon Tuning", in *Proceedings of the Asia and South Pacific Design Automation Conference (ASPDAC)*, 2008, pp. 2-9.

Silicon Tuning HLS : Approach

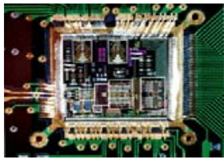
- Two stage module selection:
  - Stage 1: An iterative algorithm for power and timing variability aware module selection.
  - Stage 2: A sequential conic program (SCP) to determine the optimal body bias for post-silicon tuning which influences design-time module selection.



Silicon Tuning HLS : Results

Power Yield For 99% Performance Yield Constraint

Benchmarks	Power Constraint	Yield for Design Time Variation Aware Selection	Yield for Post Silicon Tuning + Design Time Variation Aware Selection	Improvements
Avg. of 6	No	66%	88%	38%
Avg. of 6	Yes	83%	92%	11%



Summary and Conclusions

- Most of the variability aware analysis and optimization works are at circuit or logic level.
- Work at architecture level and during HLS is slowly making progress.
- Pre-silicon and post-silicon approaches are introduced to improve power and timing yield.
- The main challenge in this unified consideration of variability, power, and timing.
- Another challenge is translation of process and physical level information to architecture level to close design-to-silicon loop.