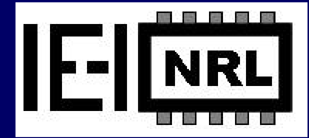


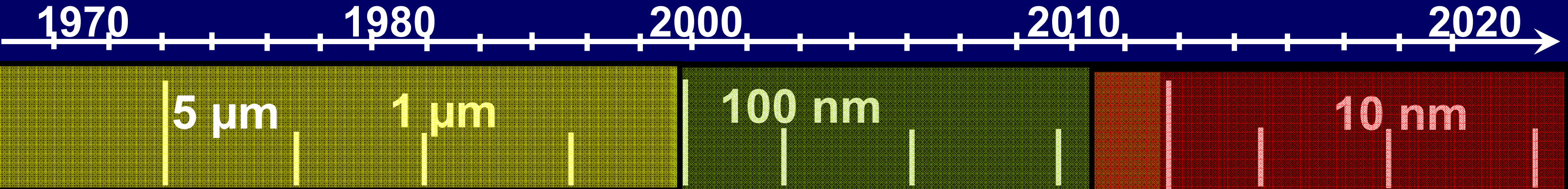
Process Variations & Process-Adaptive Design for the Nano-meter Regime

Kaushik Roy

Electrical & Computer Engineering

Purdue University

Exponential Increase in Leakage



Silicon Micro- electronics

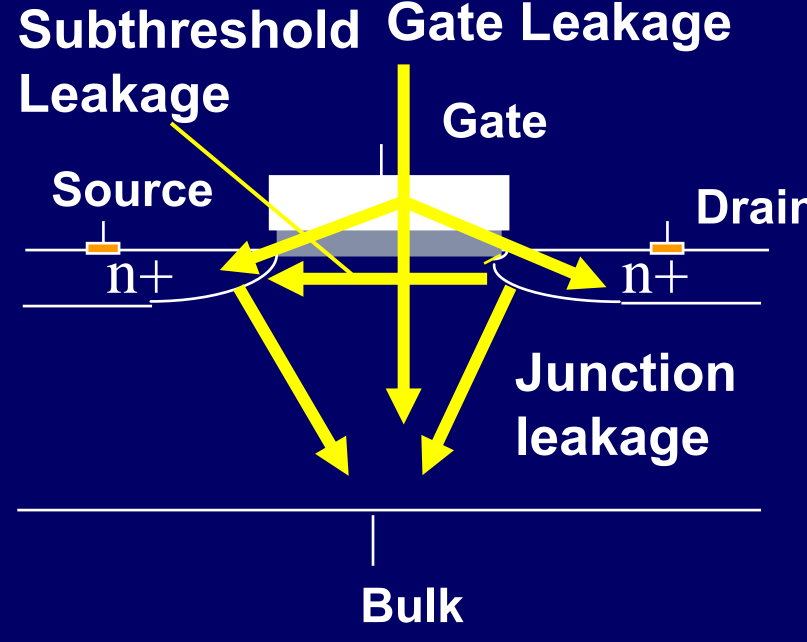
$$\frac{I_{ON}}{I_{OFF}} = 10^6$$

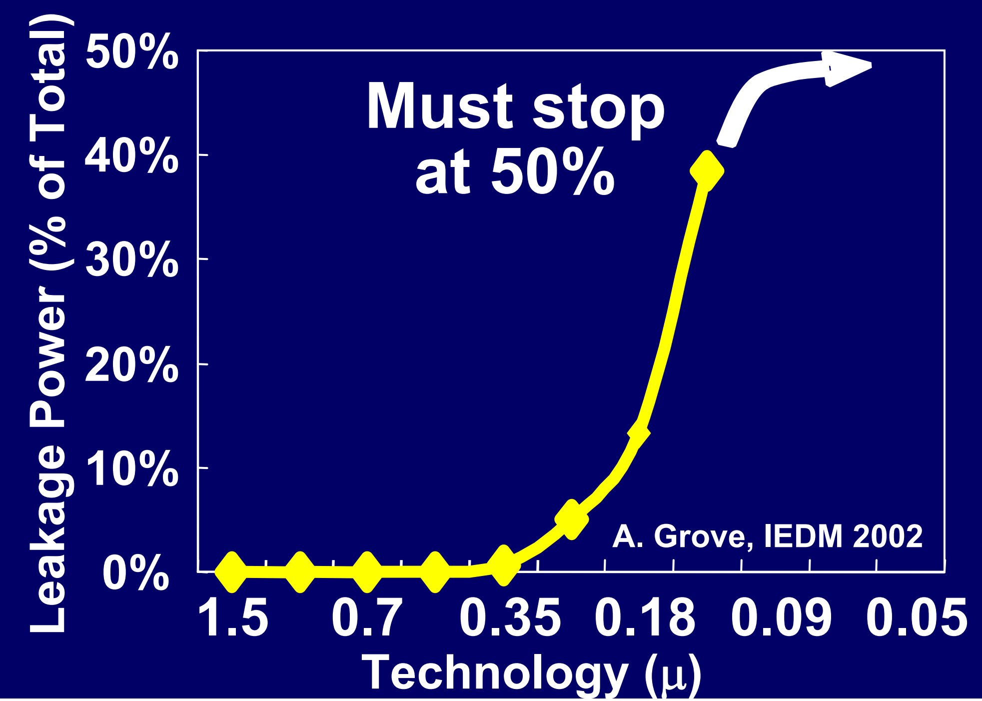
Silicon Nano- electronics

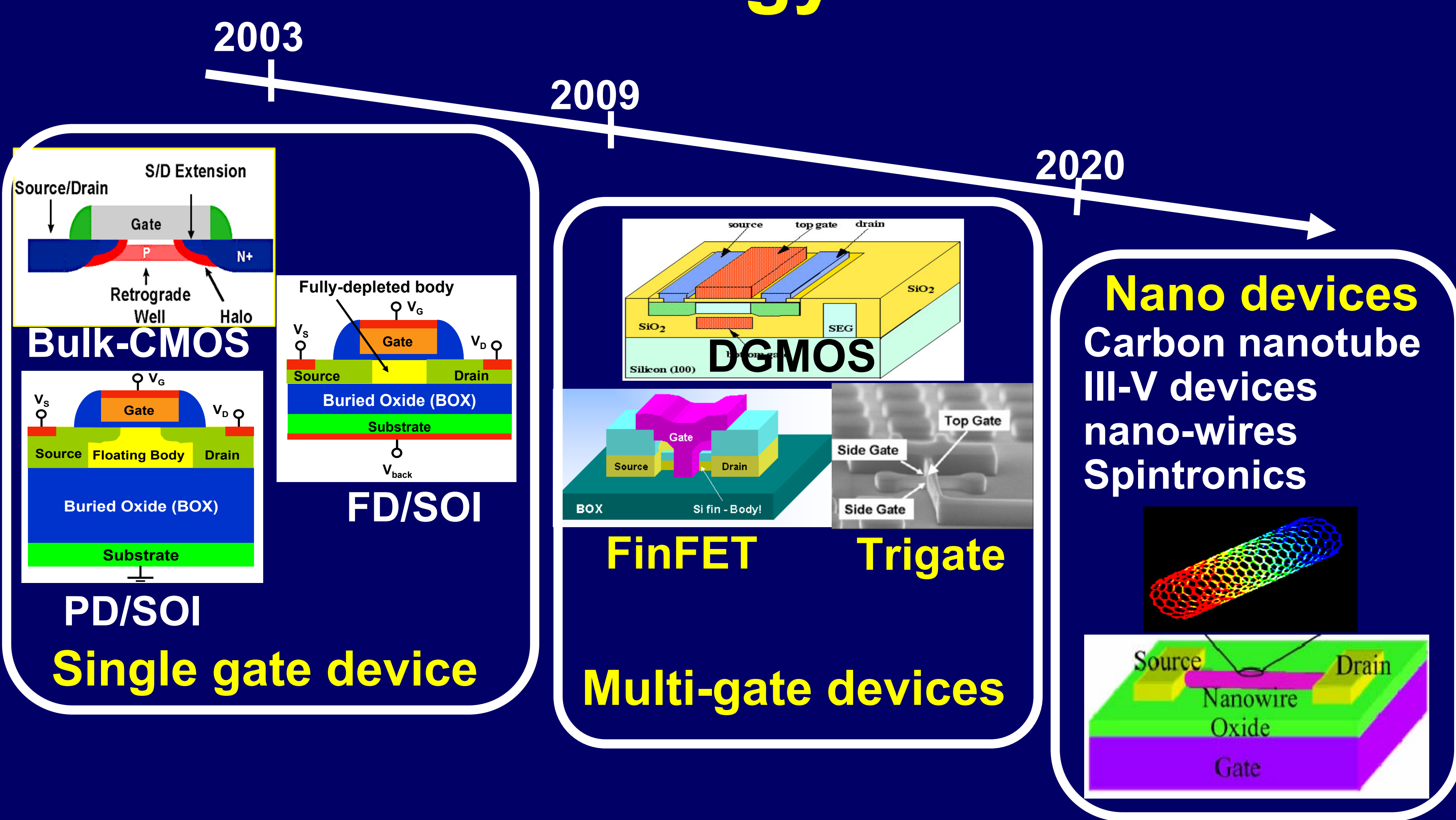
$$\frac{I_{ON}}{I_{OFF}} = 10^3$$

Non-Silicon Technology

$$\frac{I_{ON}}{I_{OFF}} \sim 10^{2\sim6}$$

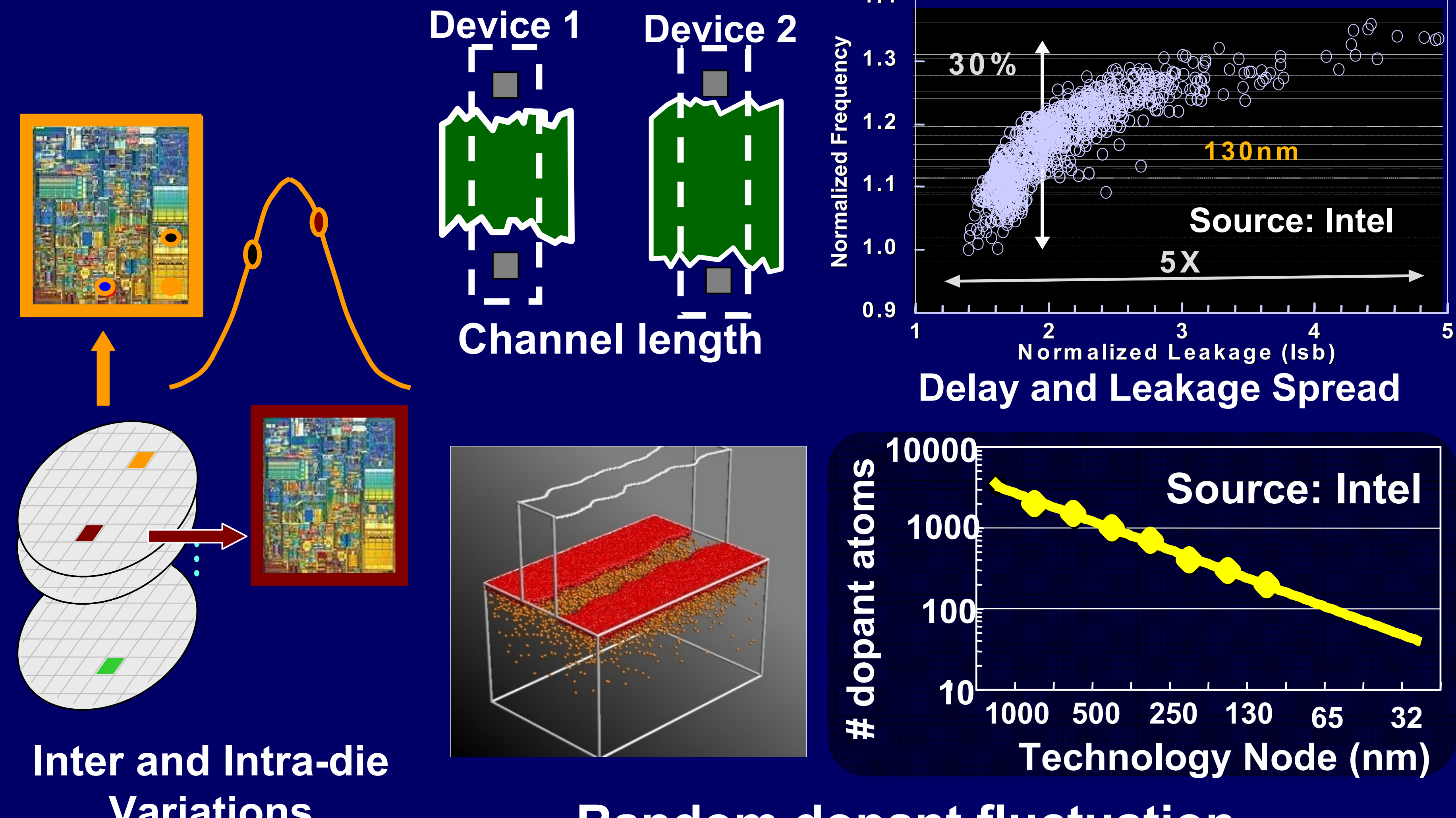






Design methods to exploit the advantages of technology innovations

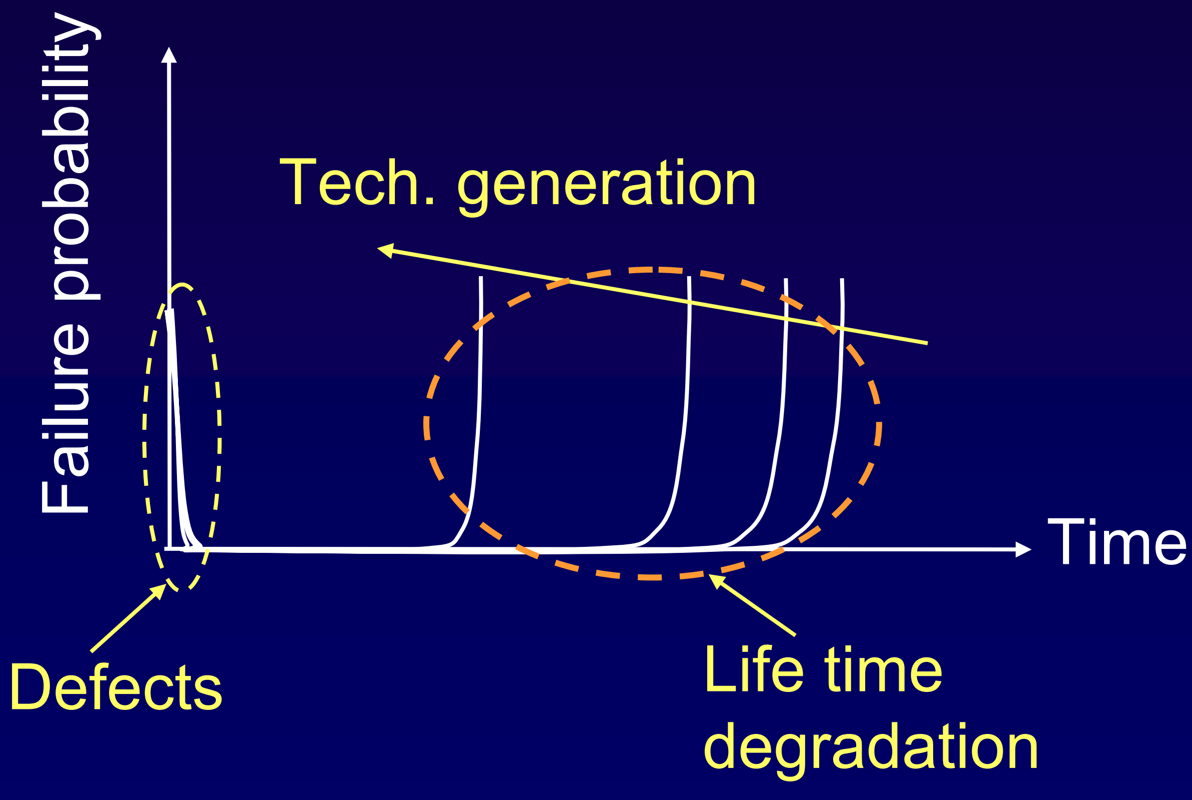
Variation in Process Parameters



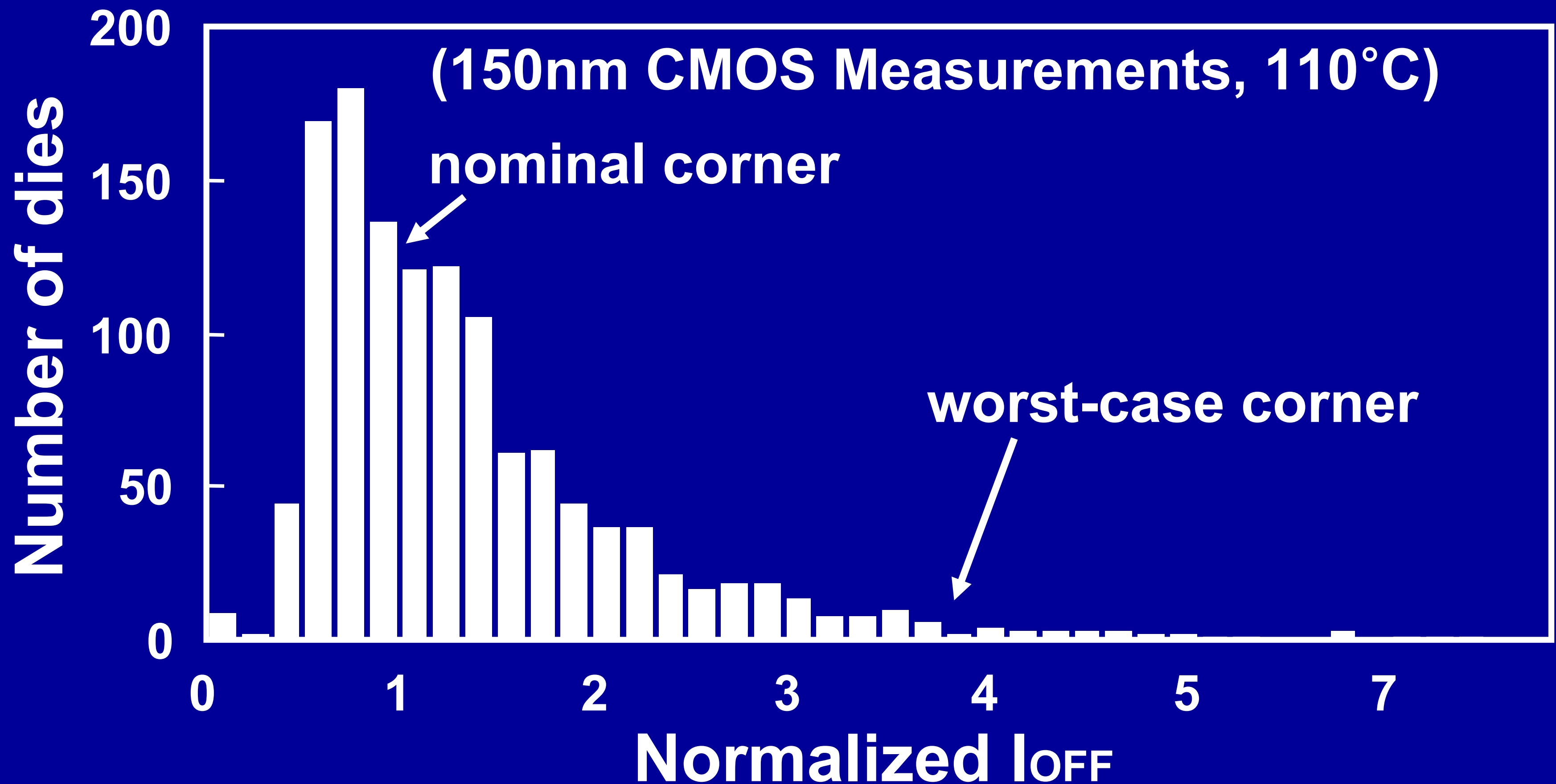
Device parameters are no longer deterministic

Reliability

Temporal degradation of performance -- NBTI

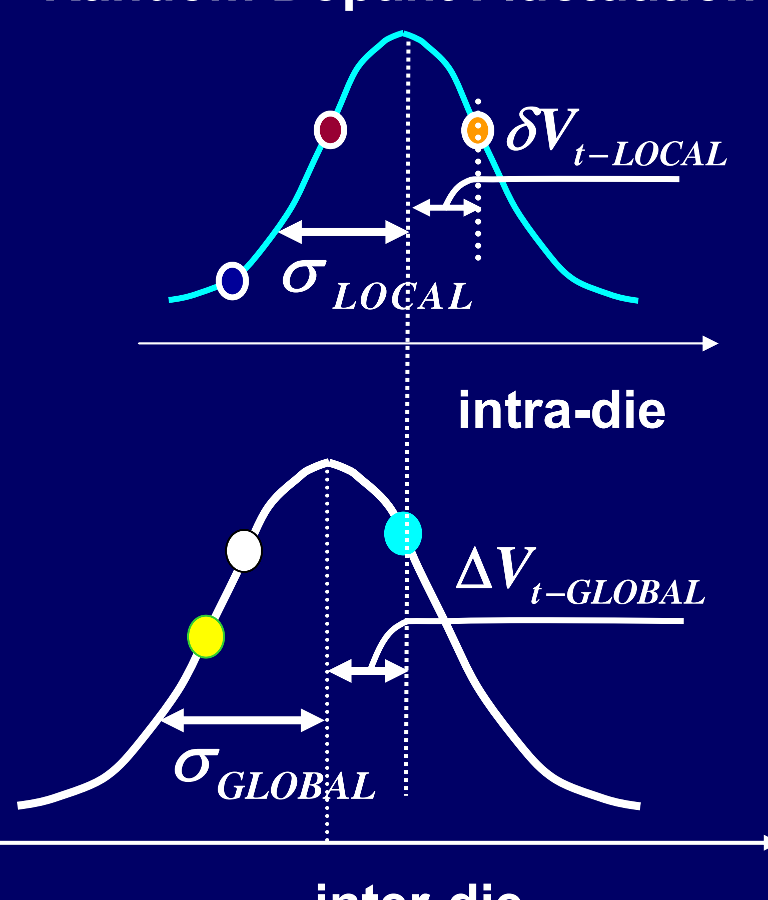


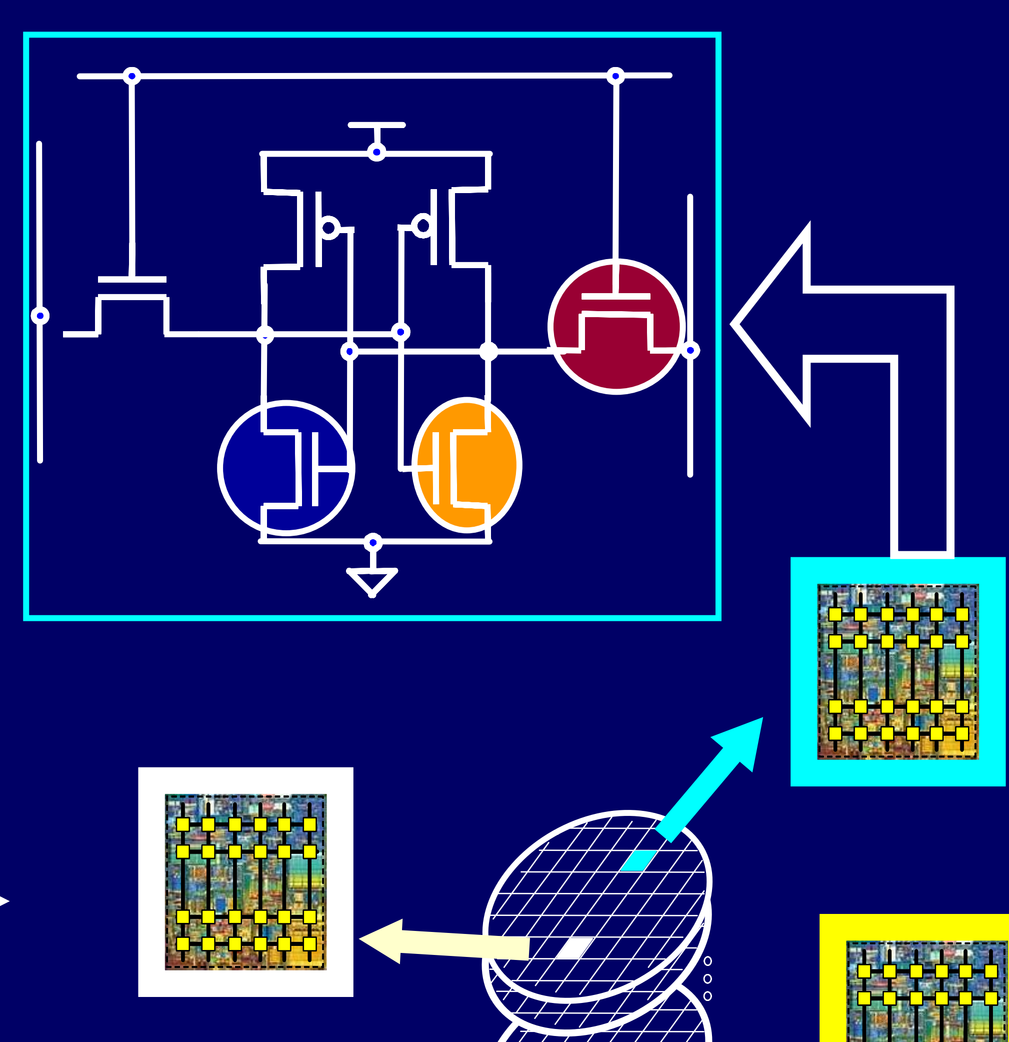
Pessimistic Design Hurts Performance



- Substantial variation in leakage across dies
- 4X variation between nominal and worst-case leakage
- Performance determined at nominal leakage
- Robustness determined at worst-case leakage

Global and Local Variations

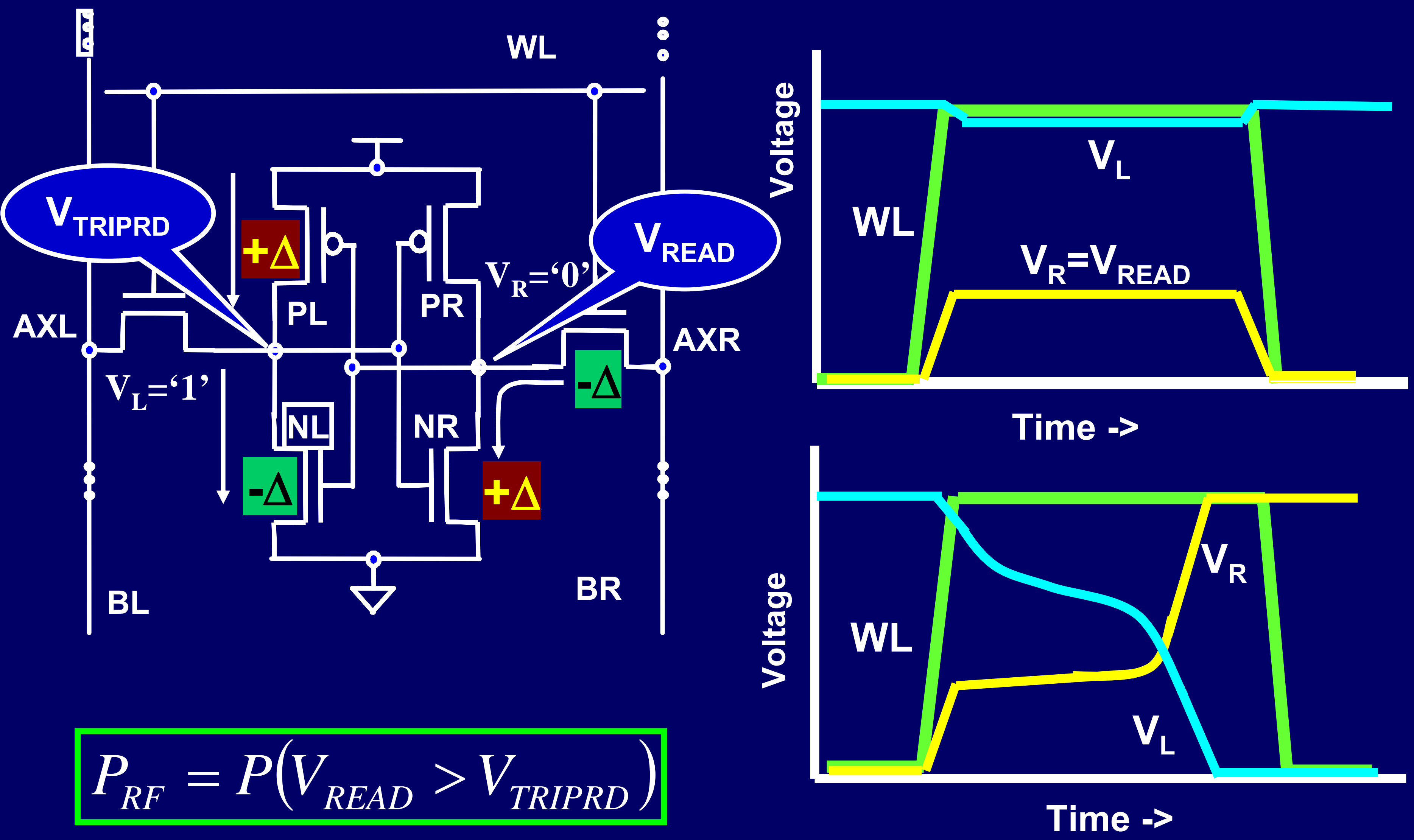




Process Tolerance: Memories

S. Mukhopadhaya, Mahmoodi, Roy
VLSI Circuit Symposium 2006, JSSC 2006, TCAD

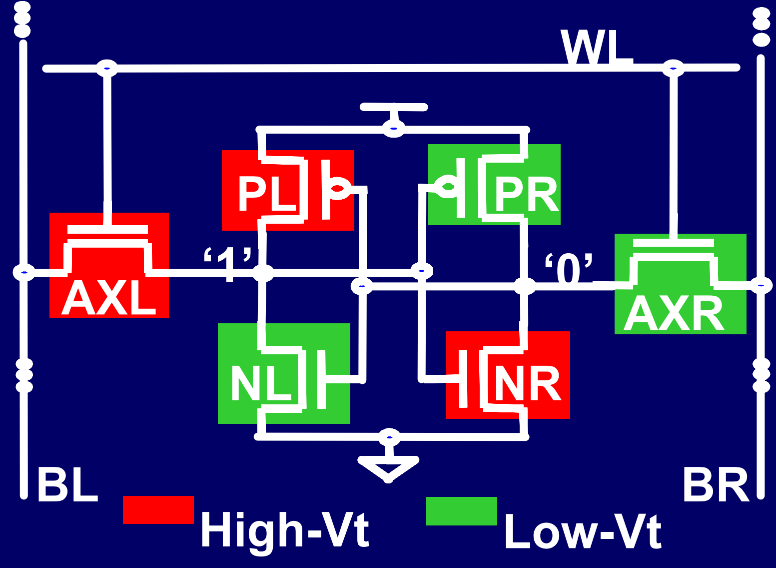
Parametric Failures: Read



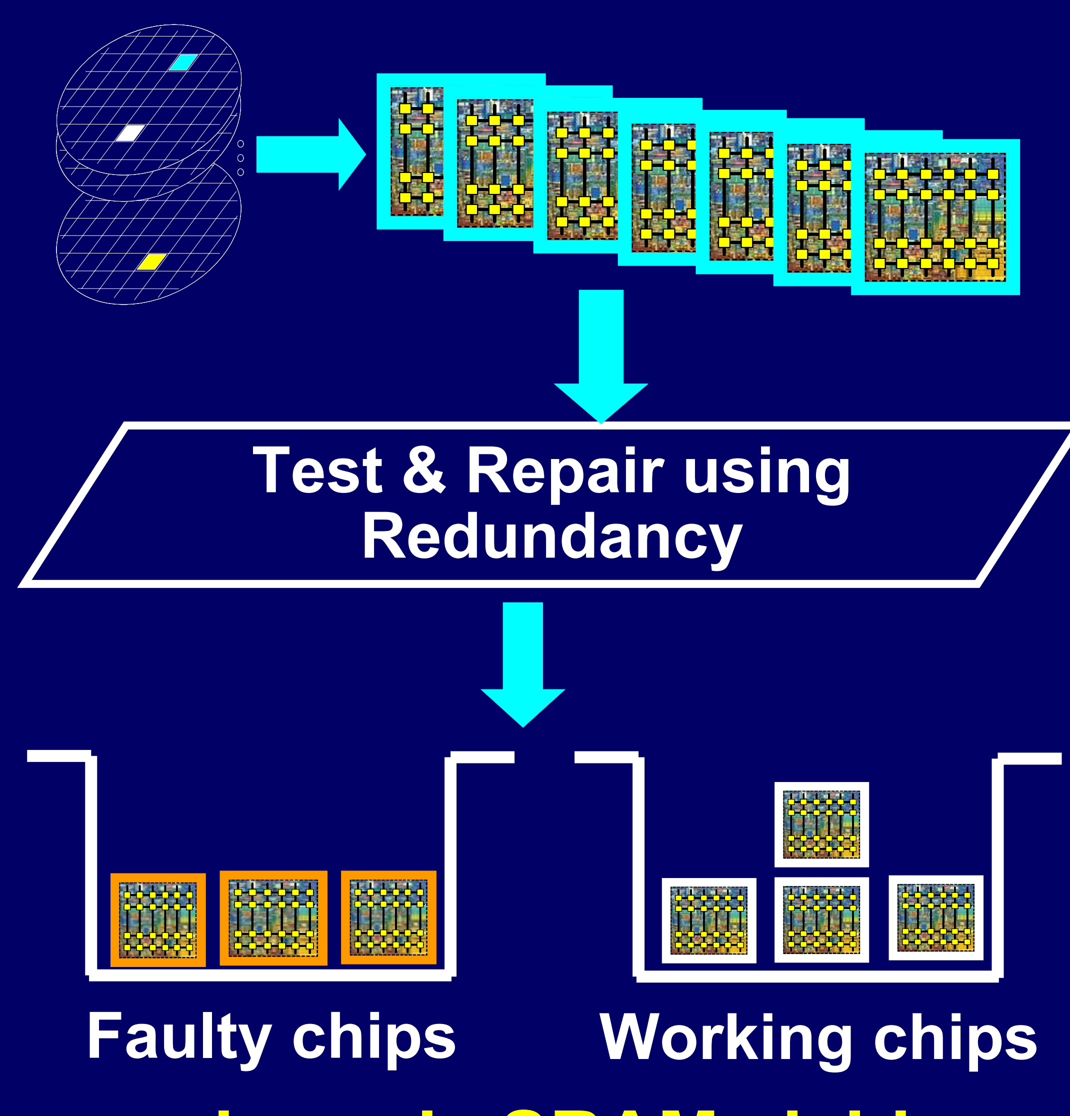
$$P_{RF} = P\left(V_{READ} > V_{TRIPRD}\right)$$

Read failure => Flipping of Cell Data while Reading

Parametric Failures in SRAM

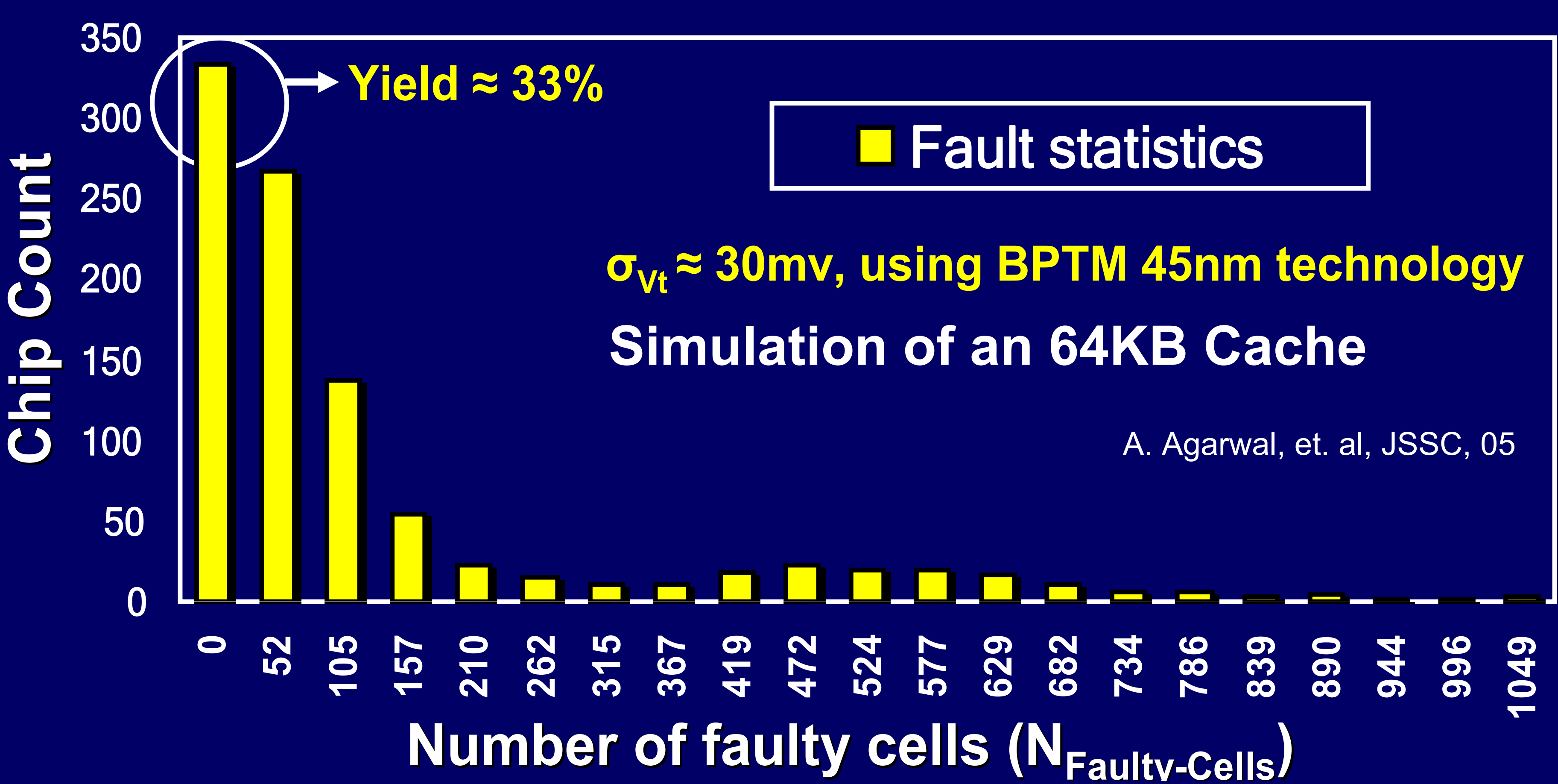


Parametric failures
- Read Failures
- Write Failures
- Access Failures
- Hold Failures



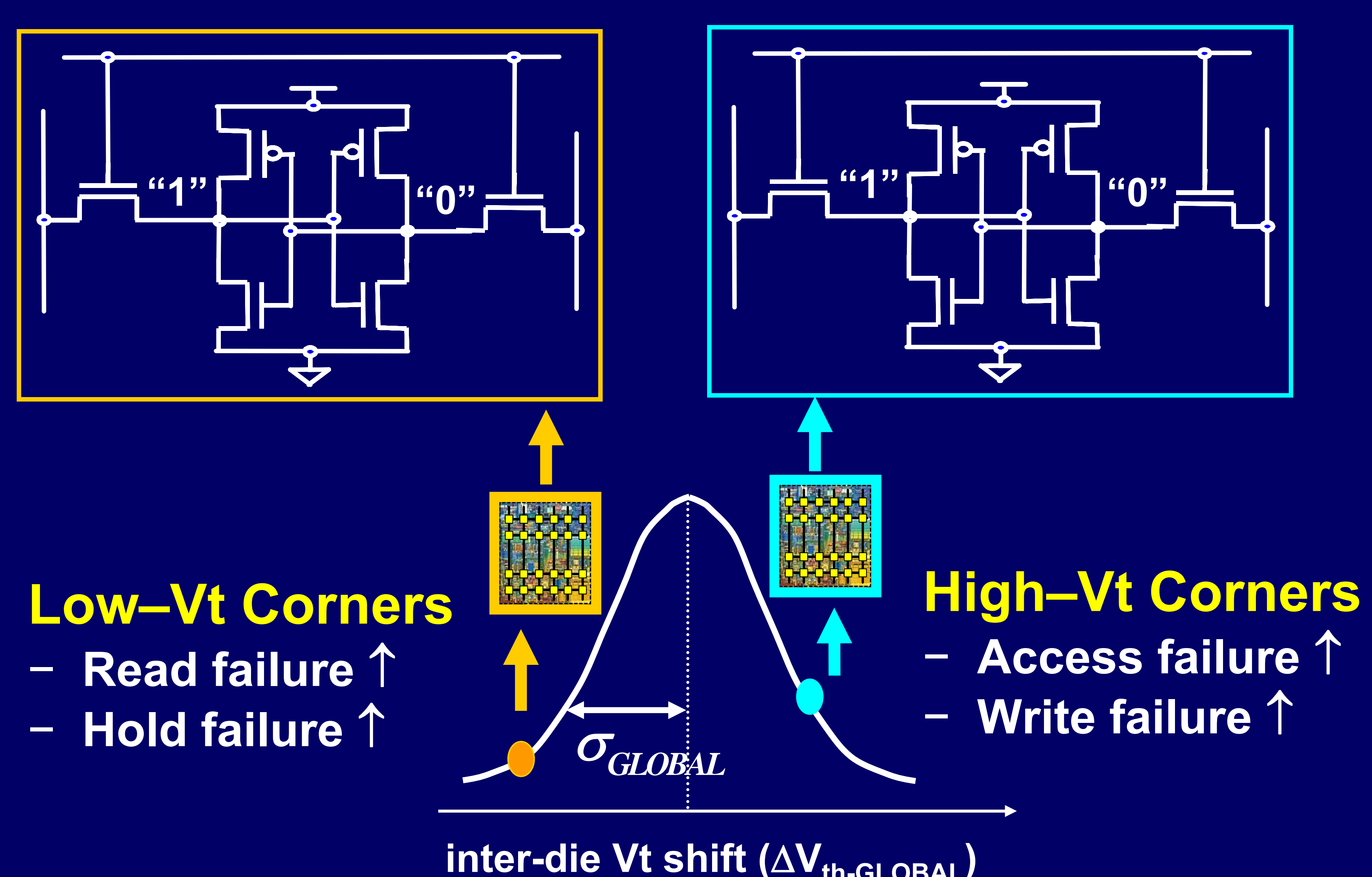
Parametric failures can degrade SRAM yield

Process Variations in On-chip SRAM

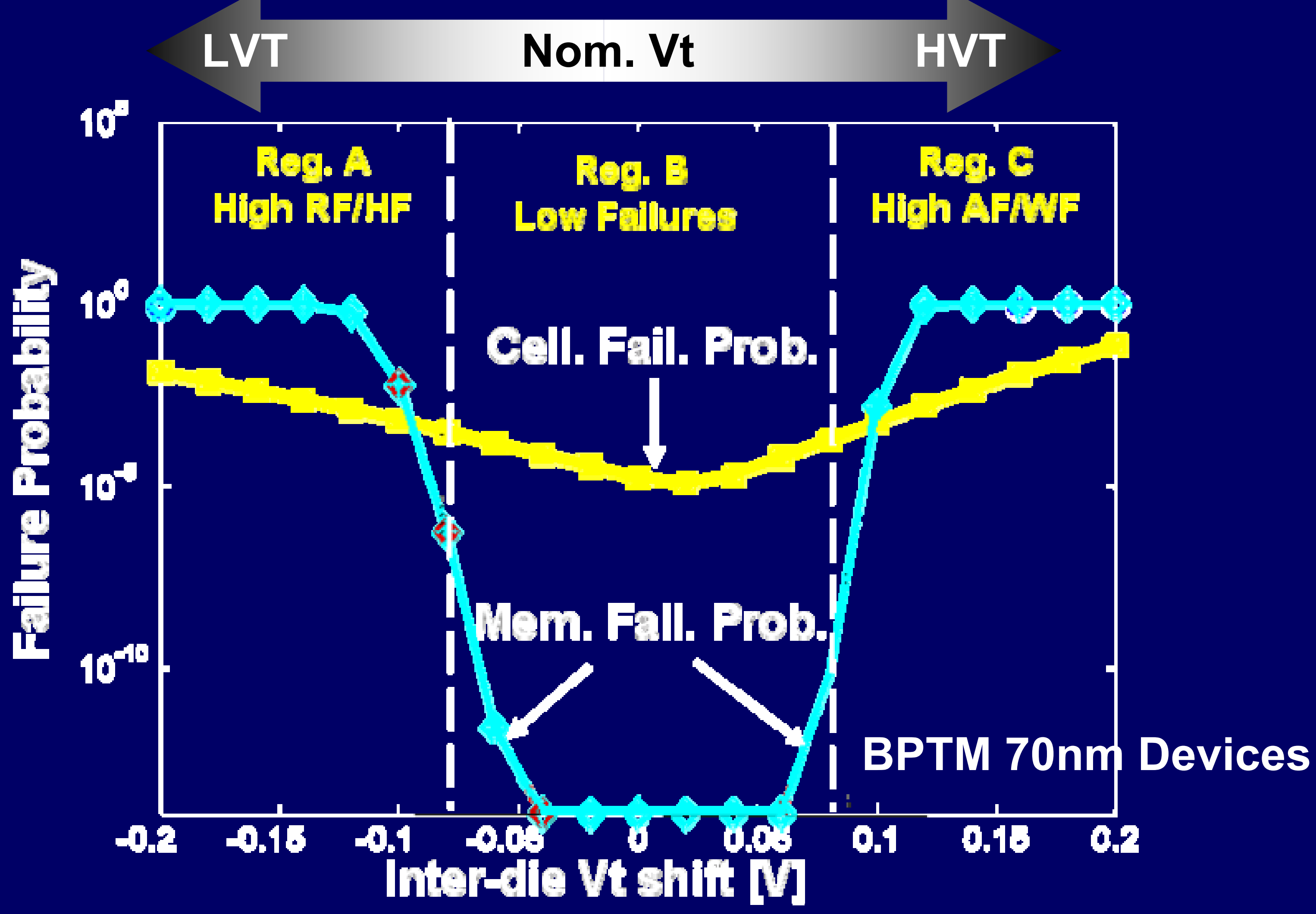


Parametric failures →Yield degradation

Inter-die Variation & Cell Failures

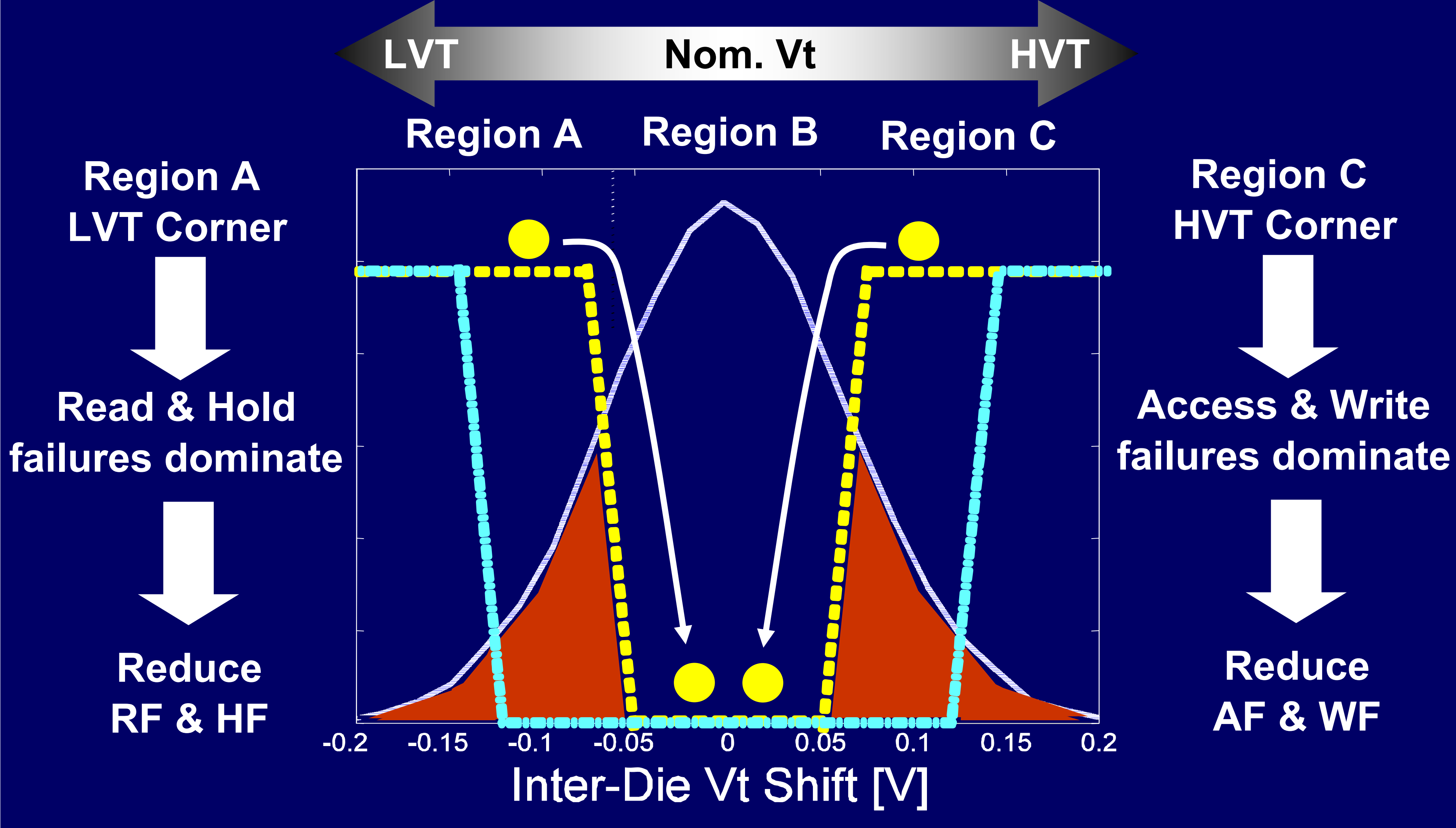


Inter-die Variation & Memory Failure



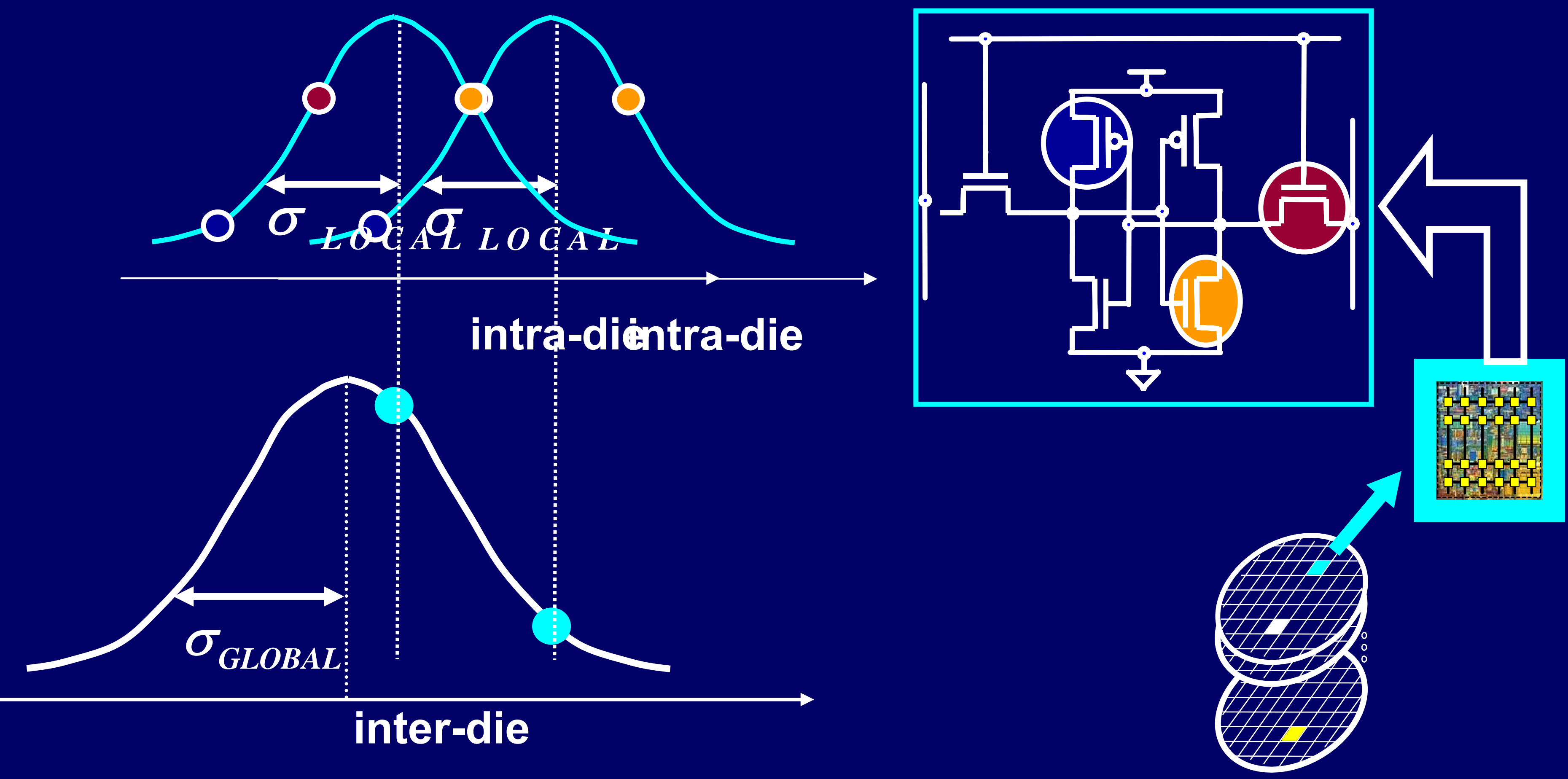
Memory failure probabilities are high when inter-die shift in process is high

Self-Repairing SRAM Array



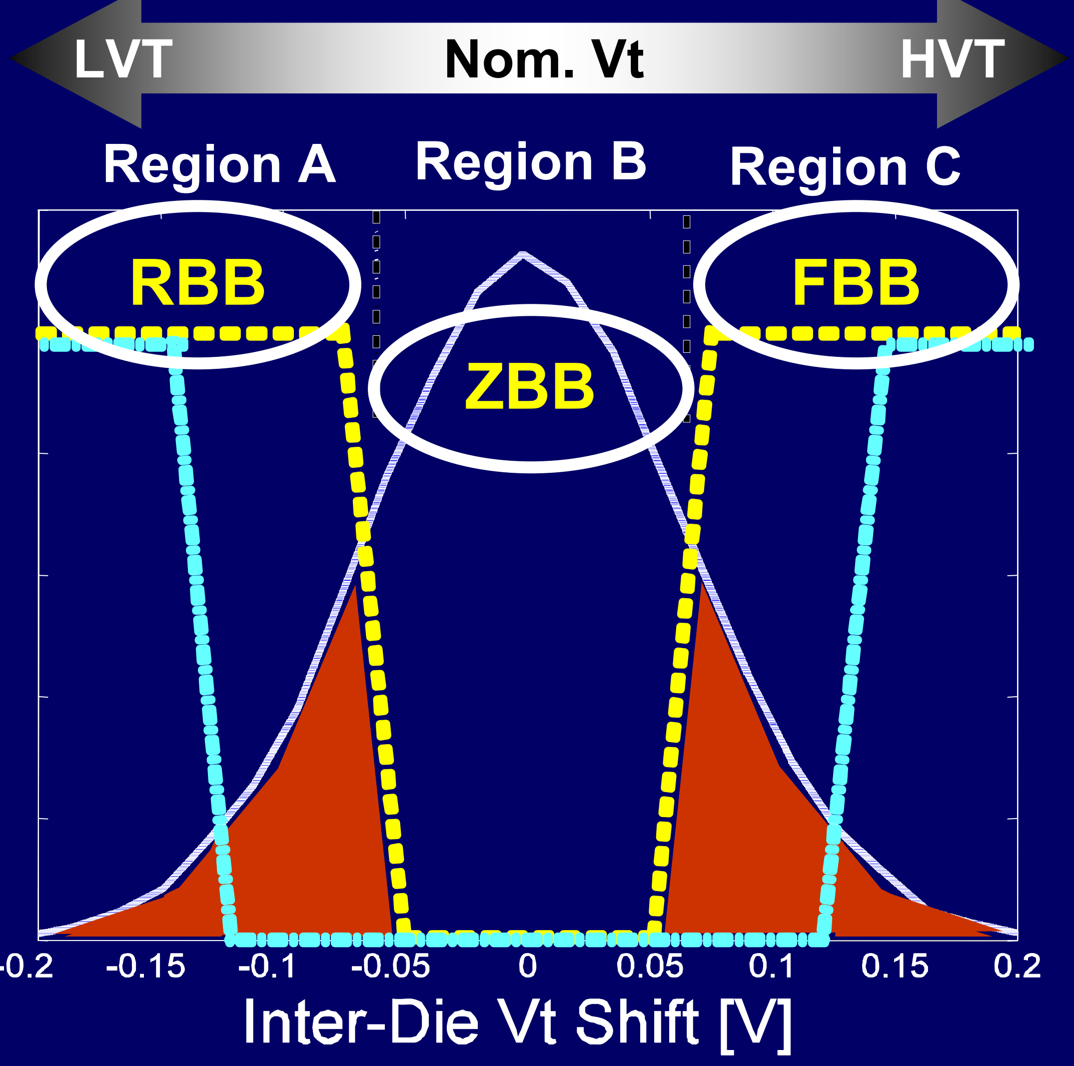
Reduce the dominant failures at different inter-die corners to increase width of low failure region

Post-Silicon Repair: Proposed Approach



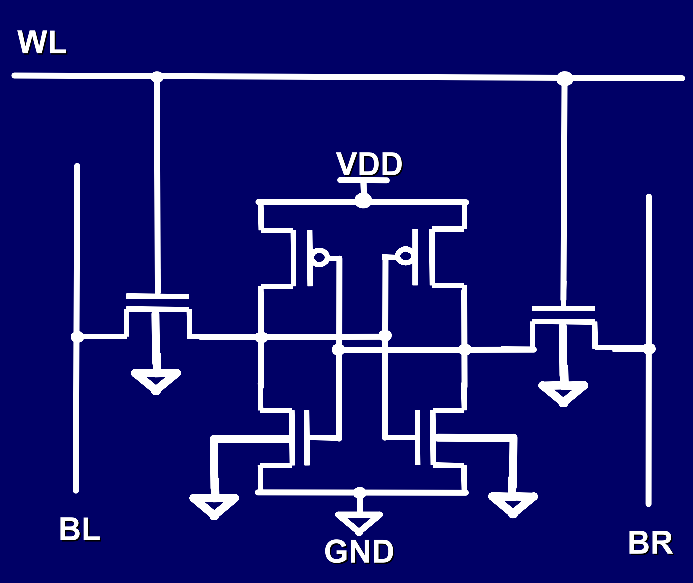
Apply correction to the global variation to reduce number of failures due to local variations

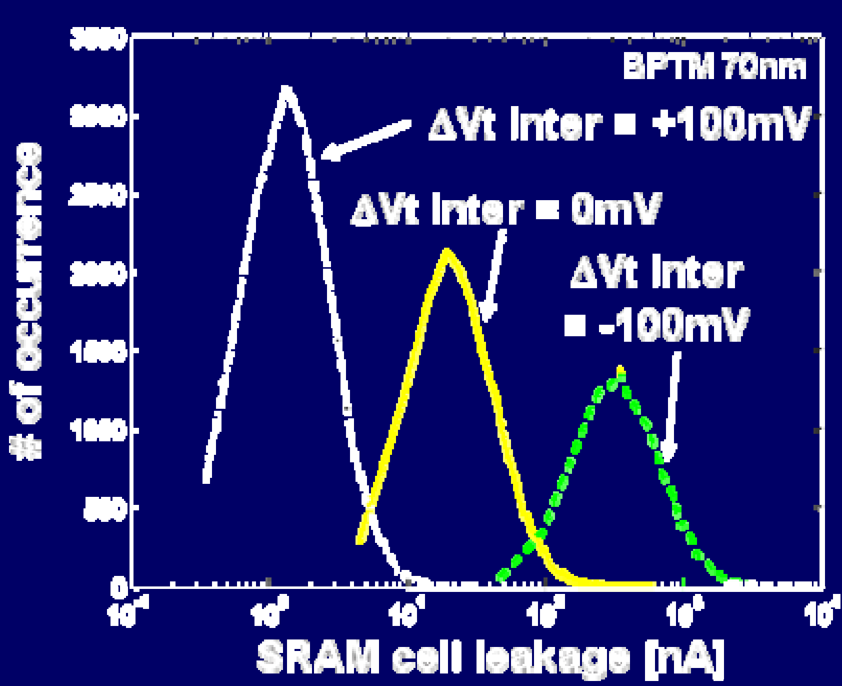
Self-Repairing SRAM Array



Reduce the dominant failures at different inter-die corners to increase width of low failure region

How to identify the inter-die Vt corner under a large intra-die variation ?

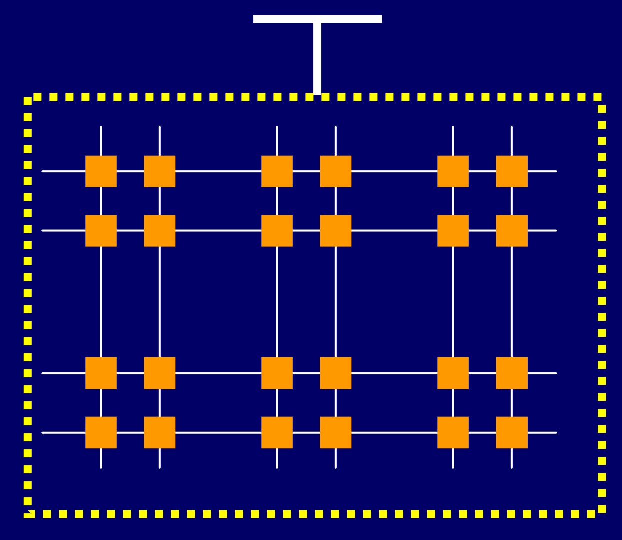




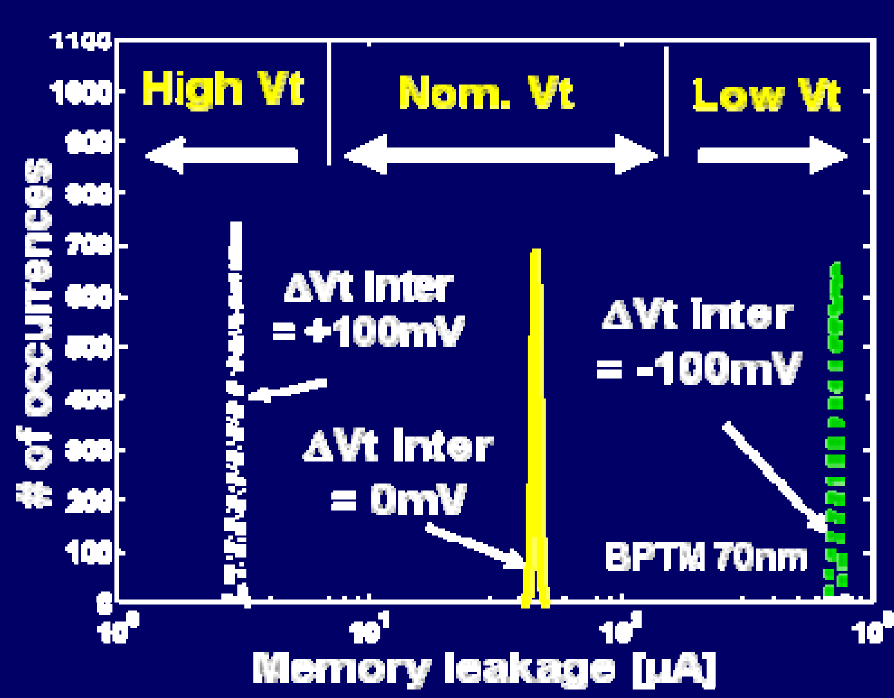
Monitor circuit parameters, e.g. leakage current

Effect of inter-die variation can be masked by intra-die variation

Array Leakage Monitoring



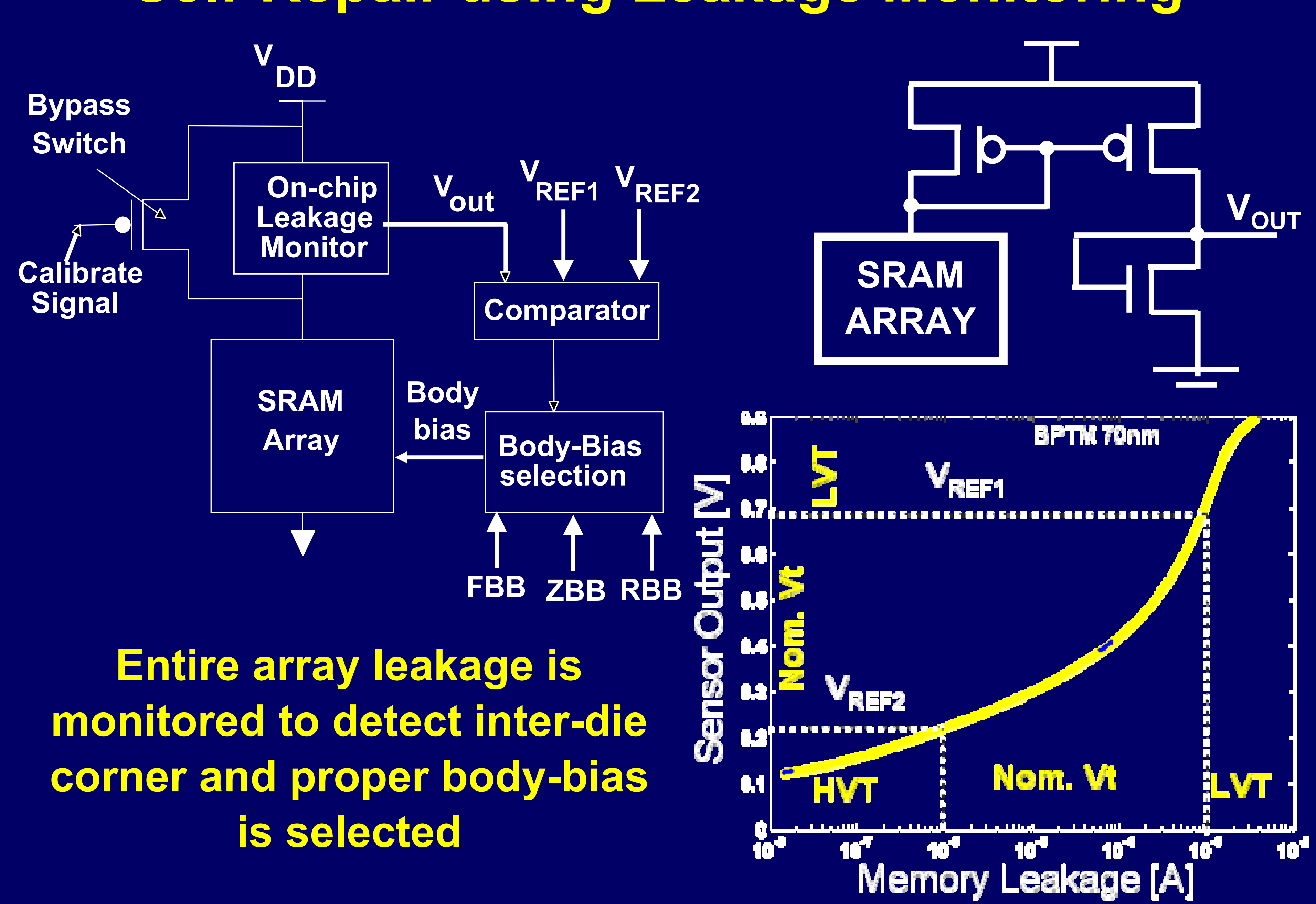
$$Y = \sum_{i=1}^{N} X_i \Rightarrow \frac{\sigma_Y}{\mu_Y} = \frac{1}{\sqrt{N}} \frac{\sigma_X}{\mu_X}$$



- Adding a large number of random variables reduces the effect of intra-die variation

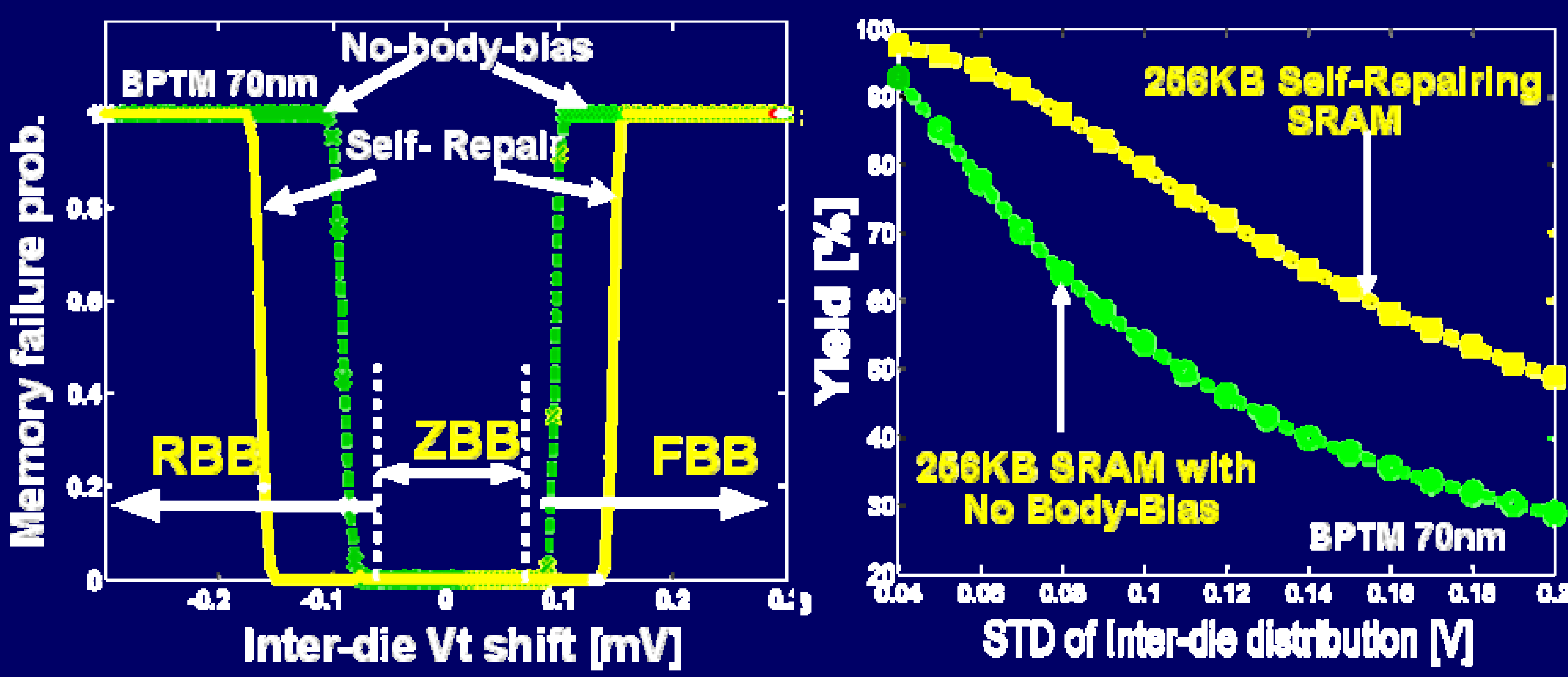
Leakage of entire SRAM array is a reliable indicator of the inter-die Vt corner

Self-Repair using Leakage Monitoring



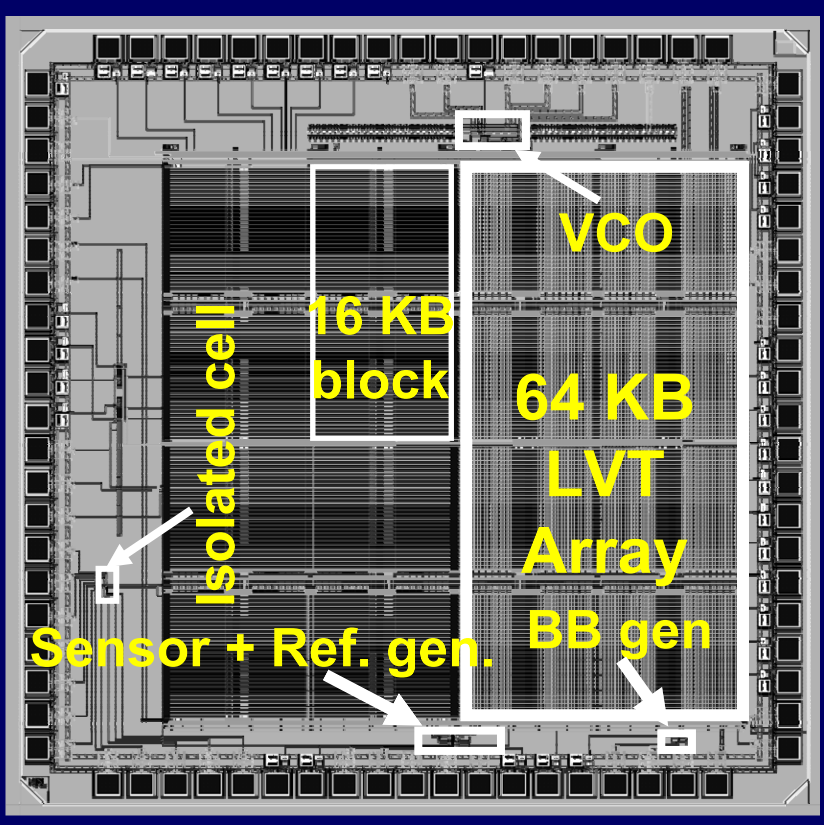
Entire array leakage is monitored to detect inter-die corner and proper body-bias is selected

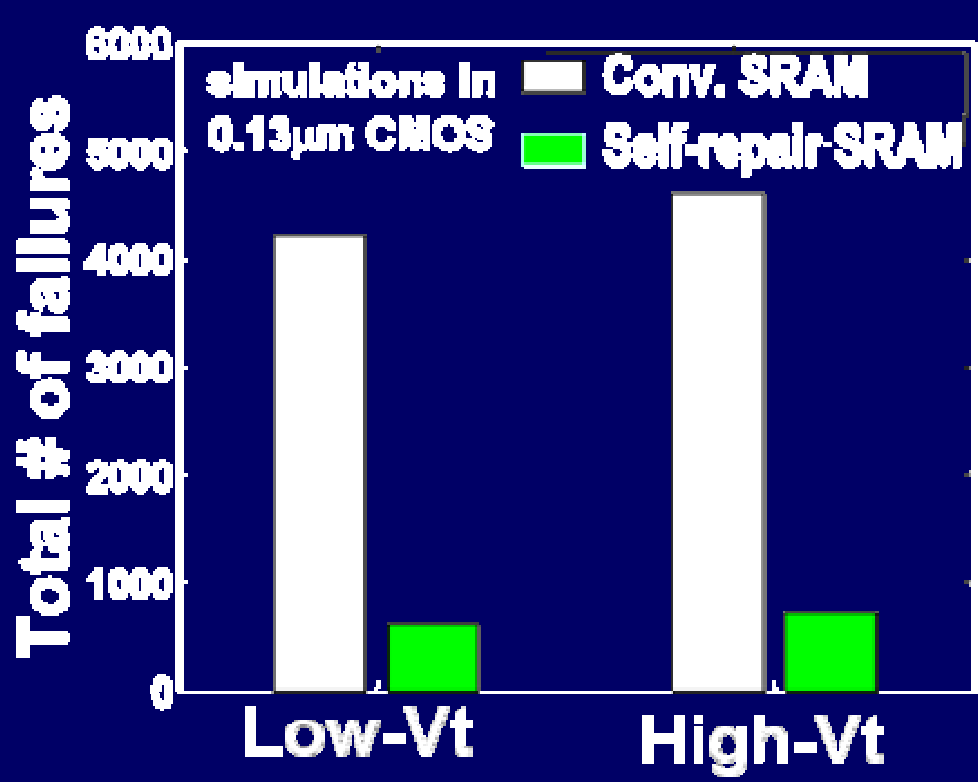
Yield Enhancement using Self-Repair



Self-Repairing SRAM using body-bias can significantly improve design yield

Test-Chip of Self-Repairing SRAM





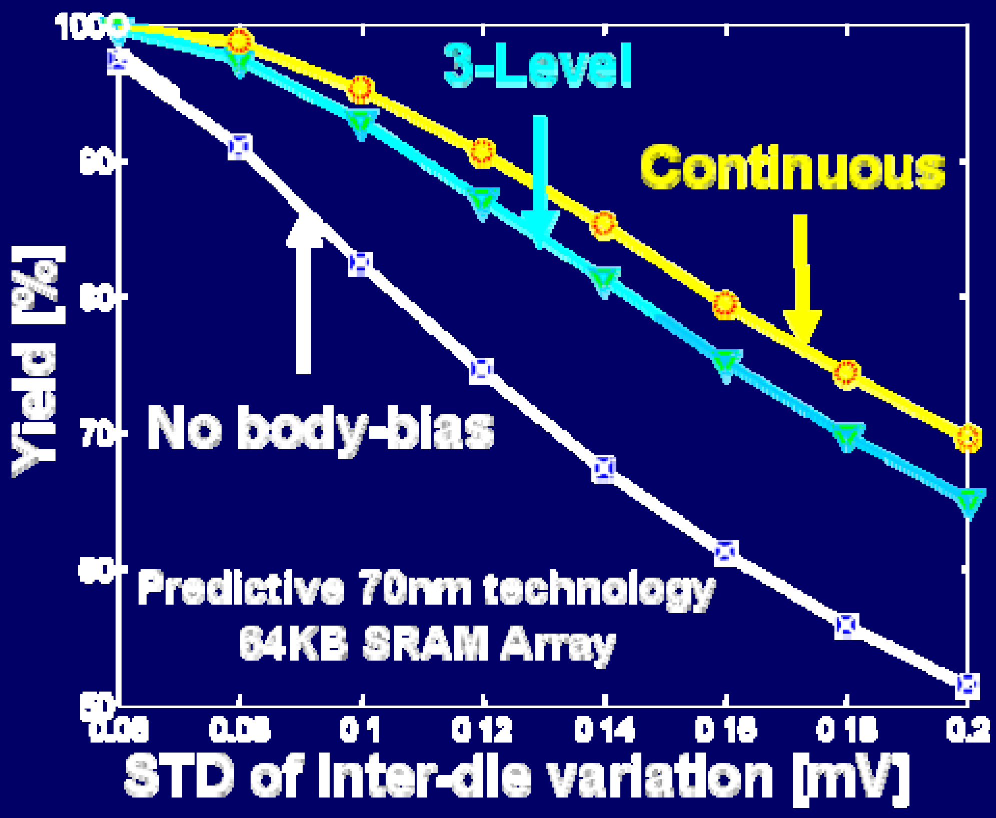
Technology : IBM 0.13 μm

128KB SRAM

Dual-Vt Triple-well tech.
Number of Trans: ~ 7 million
Die size: 16mm2
VLSI CKT Symp. 2006, ITC 2005

Simulation results for 1MB array designed in IBM 0.13μm

Continuous vs Quantized Body Bias

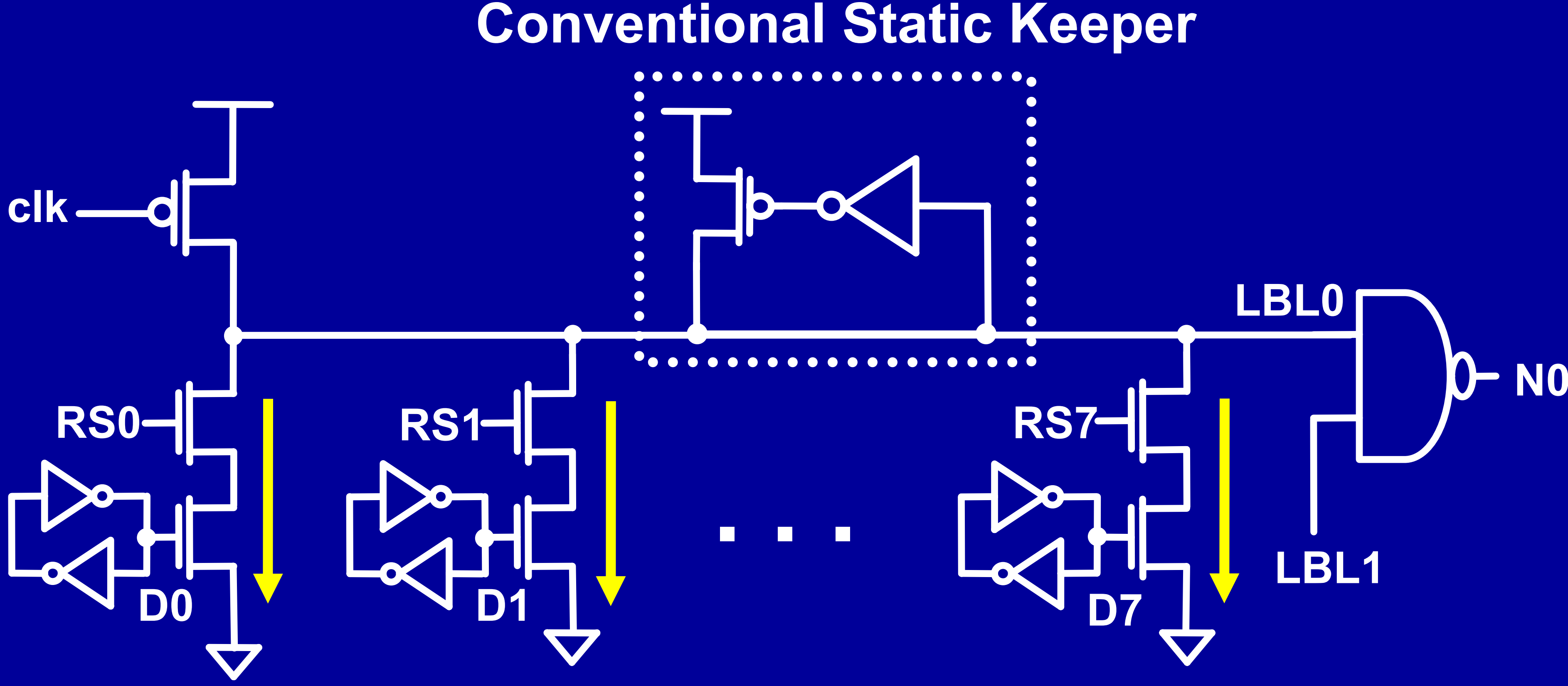


Quantized (3 Level: FBB, ZBB, RBB) body bias scheme is a cost effective solution with good yield enhancement possibility

Process Tolerance: Register Files

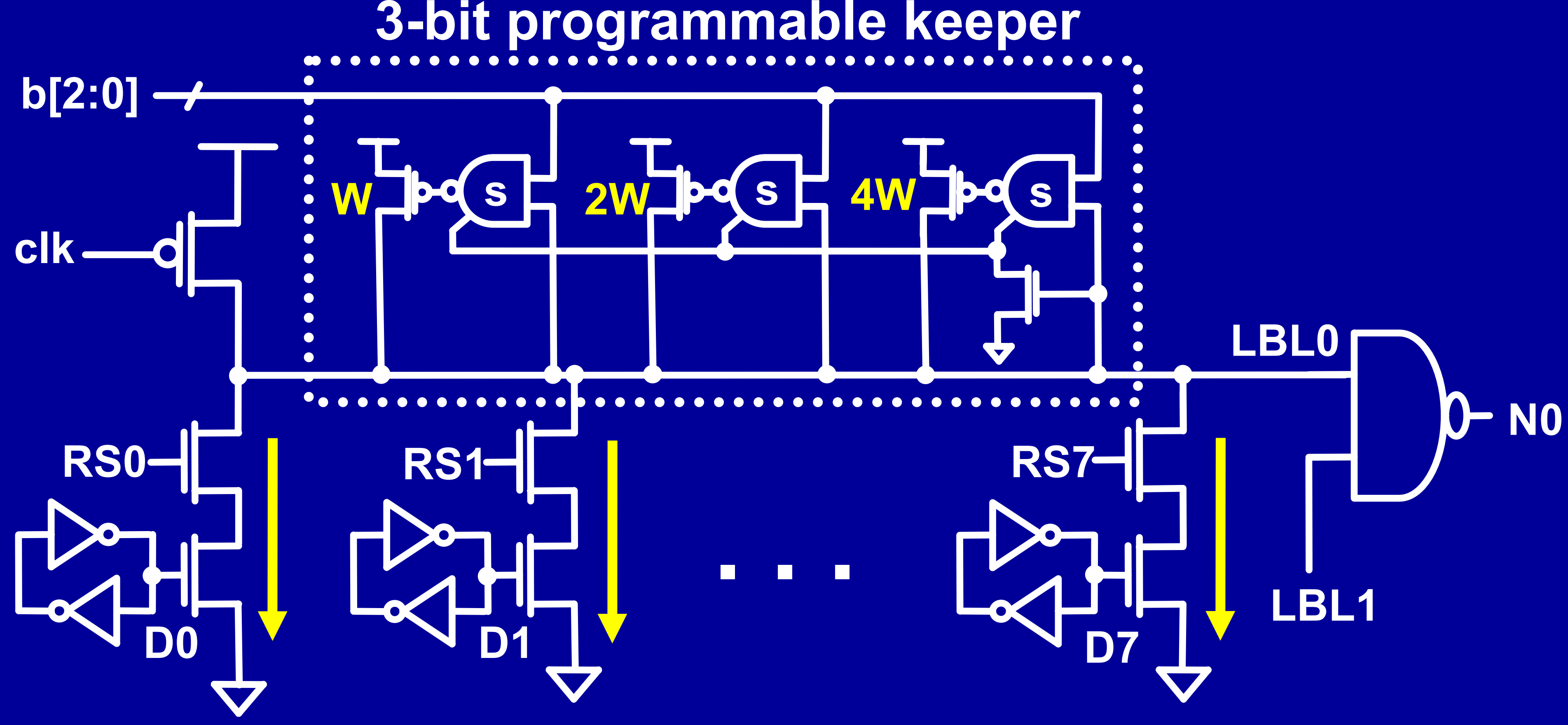
Kim et. al. VLSI Circuit Symposium 2004

Process Compensating Dynamic Circuit Technology



- Keeper upsizing degrades average performance

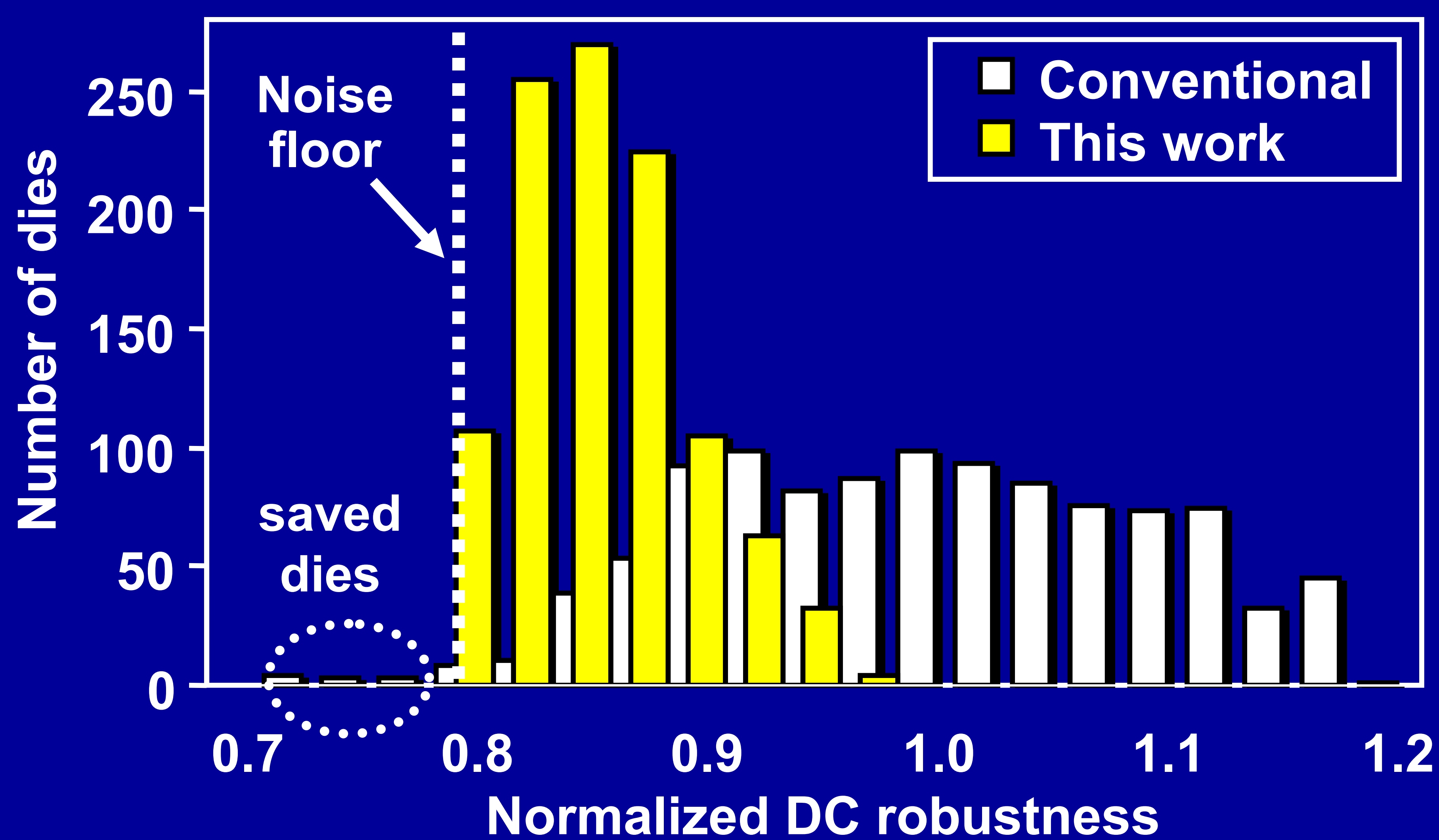
Process Compensating Dynamic Circuit Technology



C. Kim et al. , VLSI Circuits Symp. '03

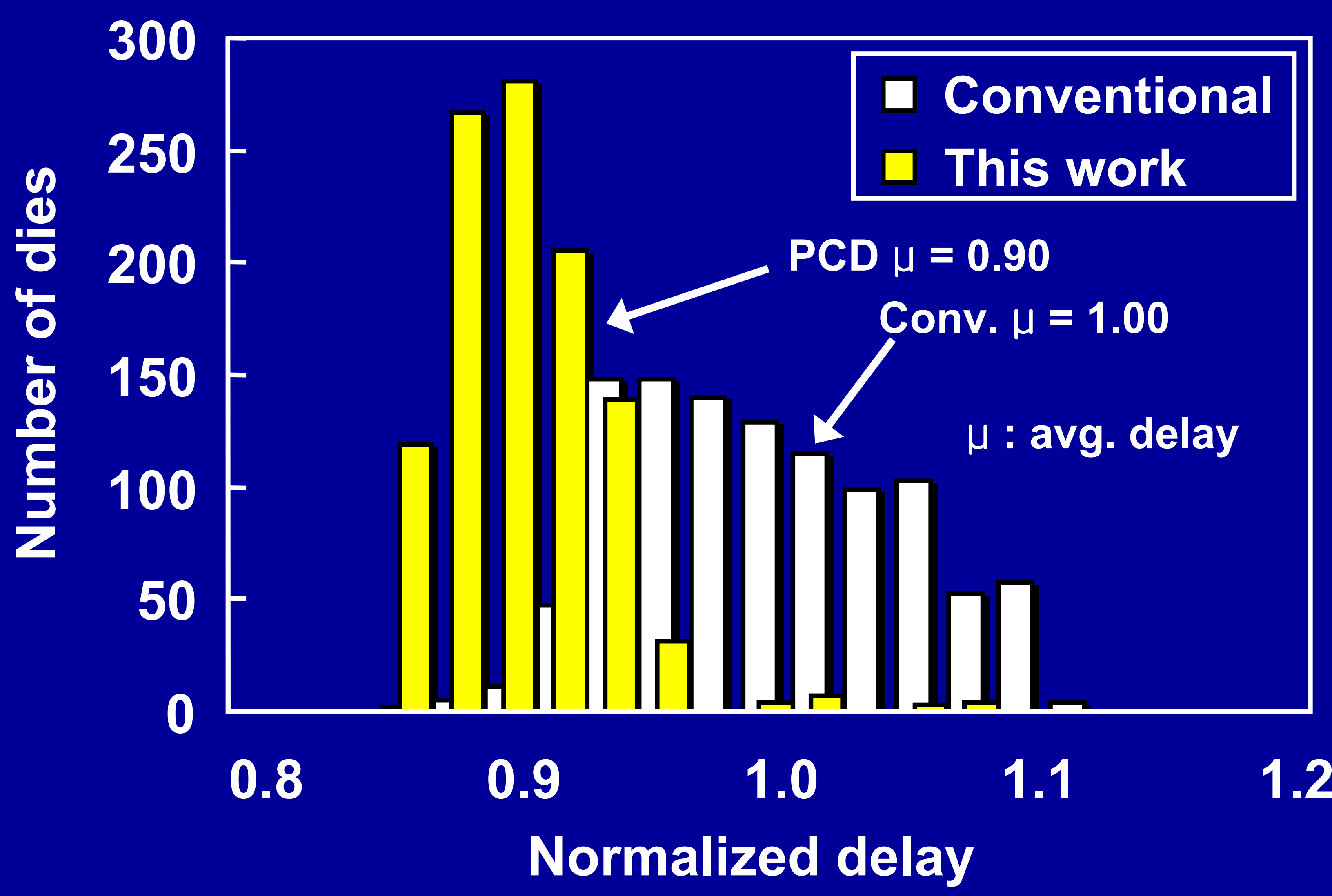
- Opportunistic speedup via keeper downsizing

Robustness Squeeze



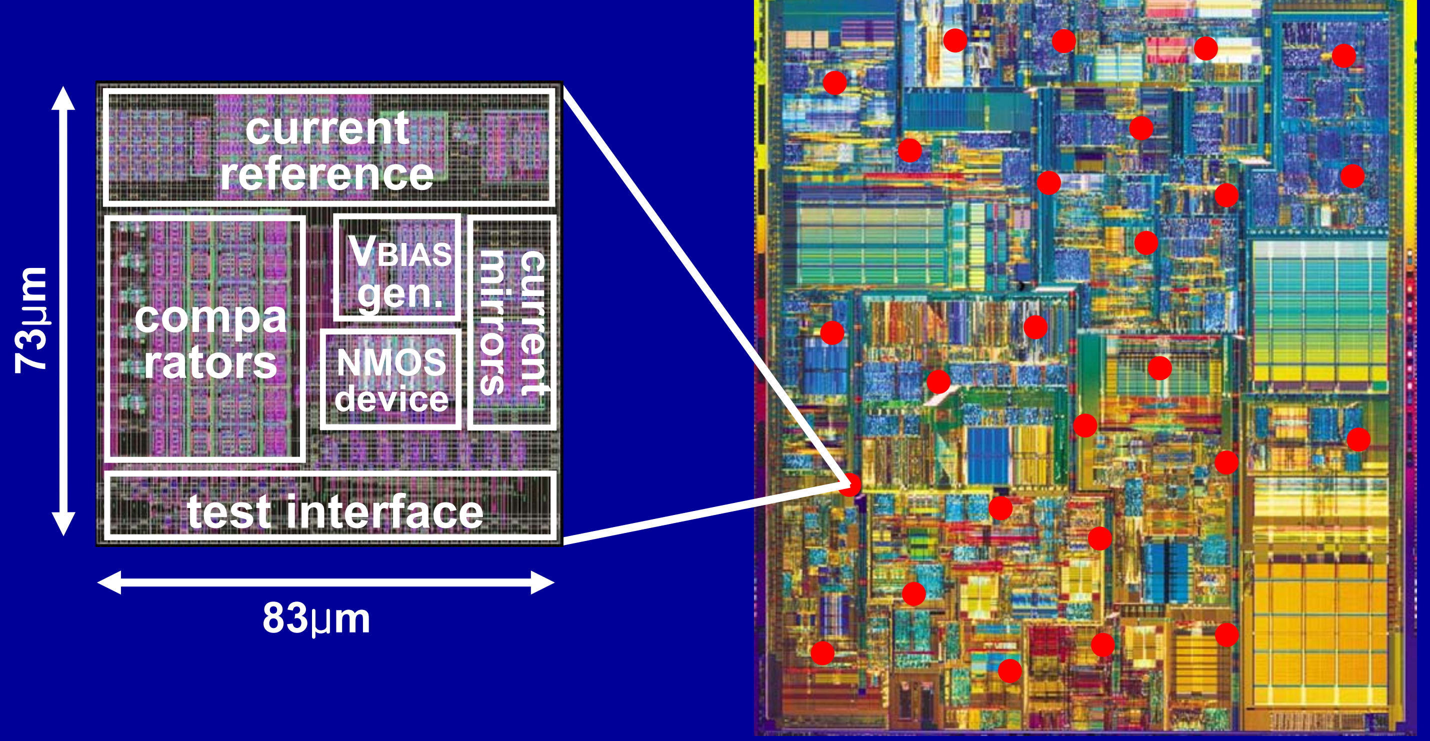
- 5X reduction in robustness failing dies

Delay Squeeze



- 10% opportunistic speedup

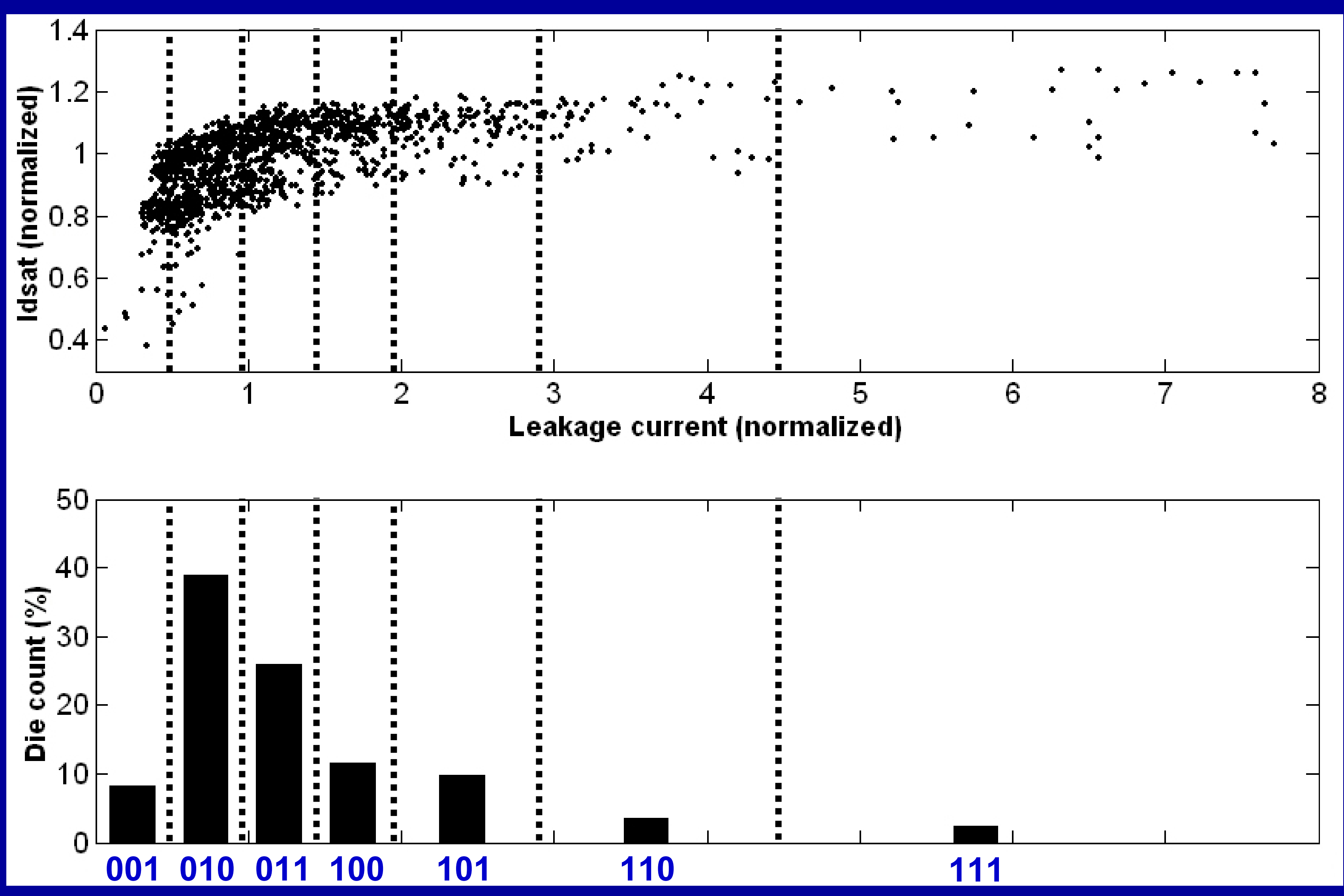
On-Die Leakage Sensor For Measuring Process Variation



C. Kim et al. , VLSI Circuits Symp. '04

- High leakage sensing gain – 90nm dual-Vt, Vdd=1.2V, 7 level resolution, 0.66 mW @80Cº

Leakage Binning Results

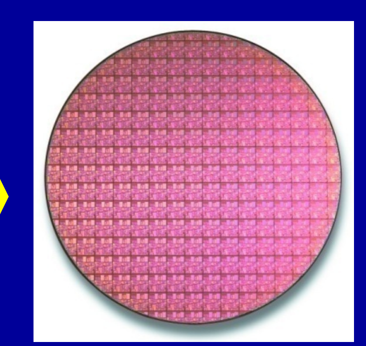


Output codes from leakage sensor

Self-Contained Process Compensation

Fab

Wafer test

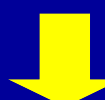


Process detection

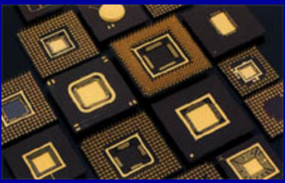
Leakage measurement

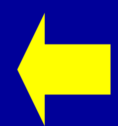
On-die leakage sensor

Program PCD using fuses

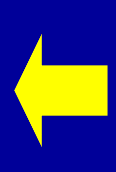


Assembly

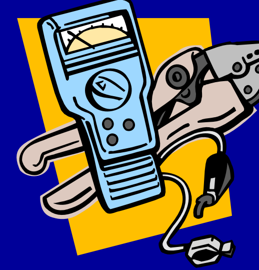


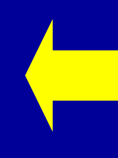


Burn in

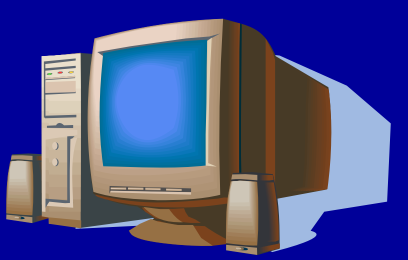


Package test





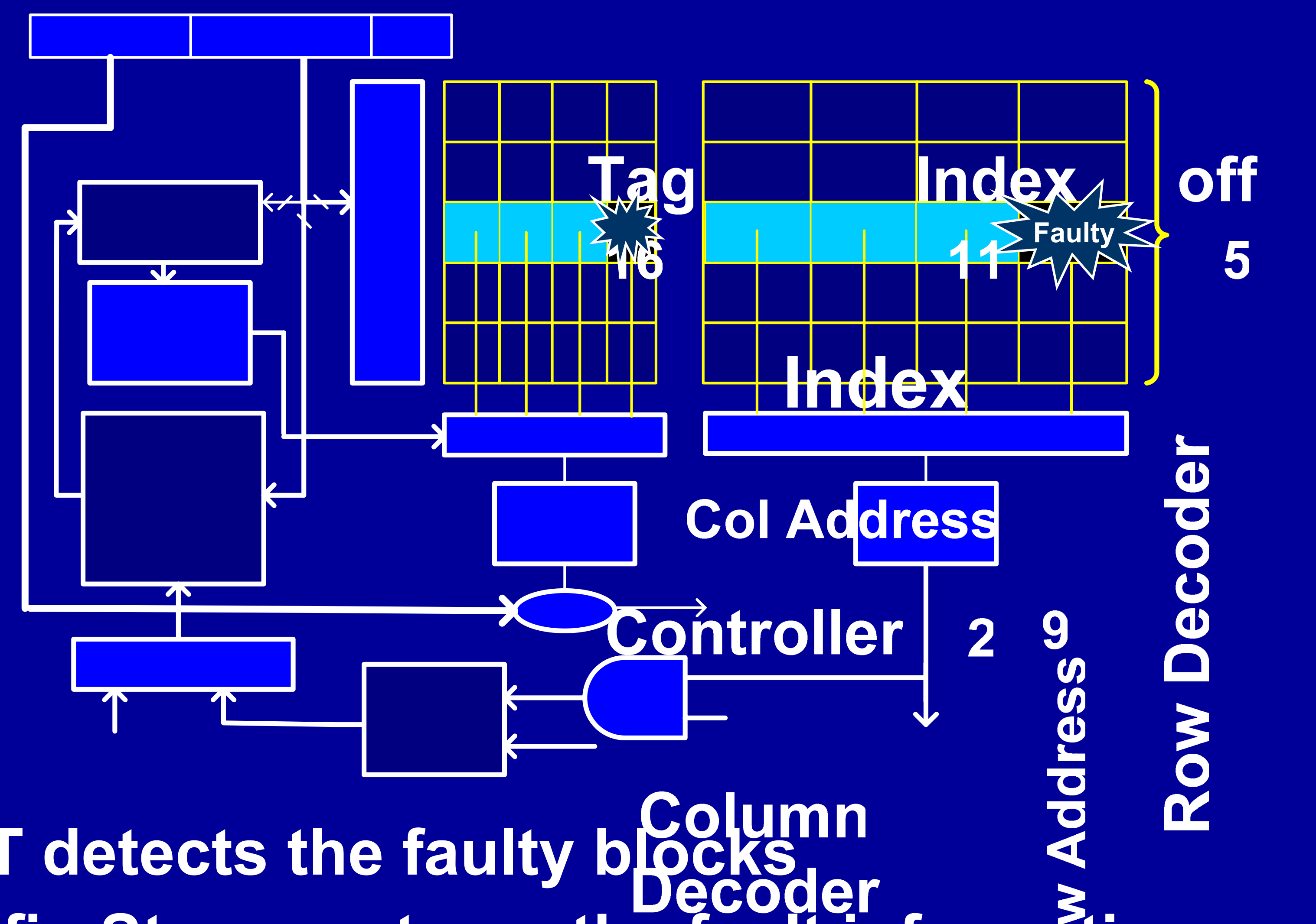
Customer



Self-Repair: Architecture Level

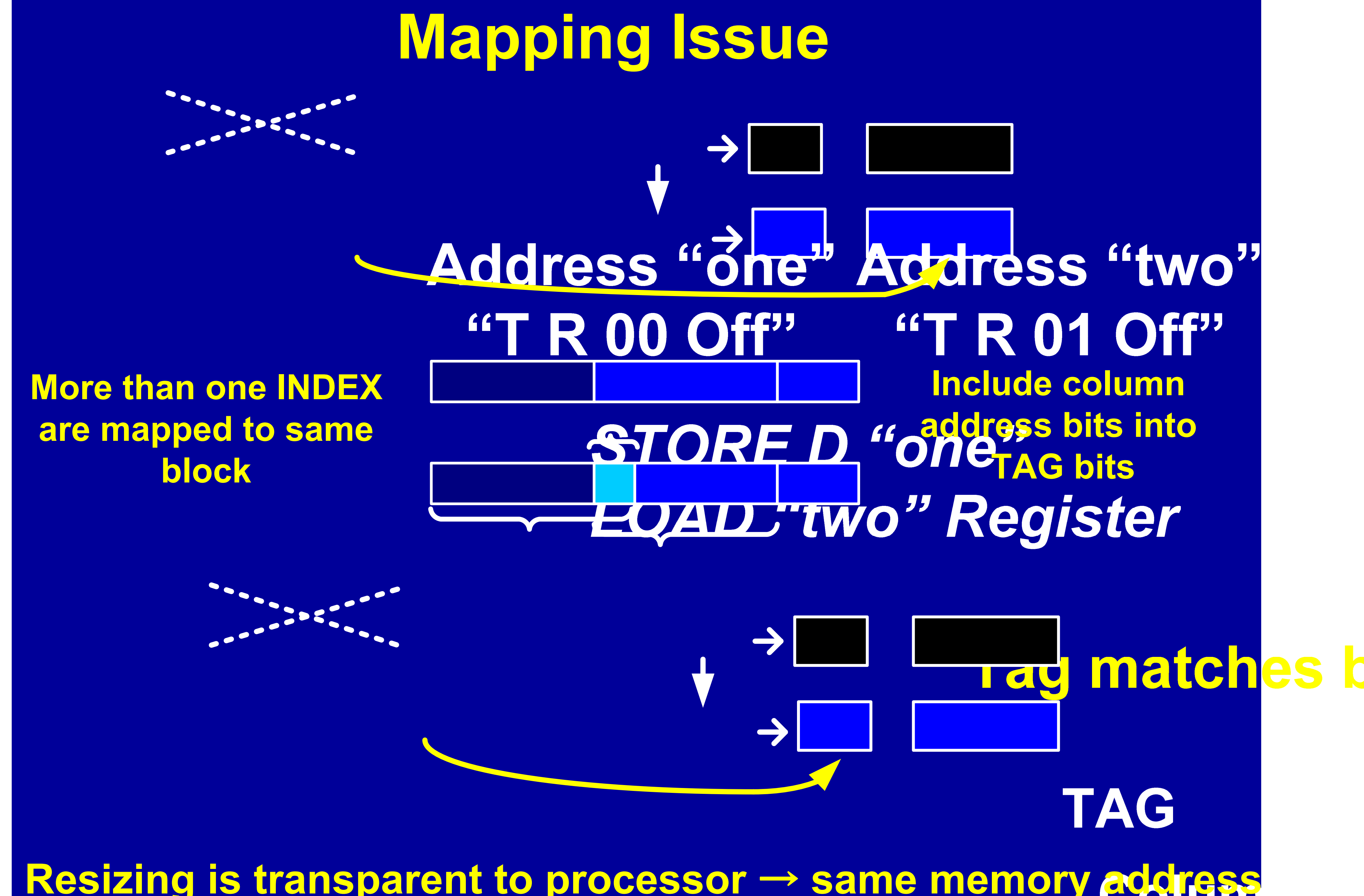
Agarawal, Roy TVLSI 2005

Fault-Tolerant Cache Architecture

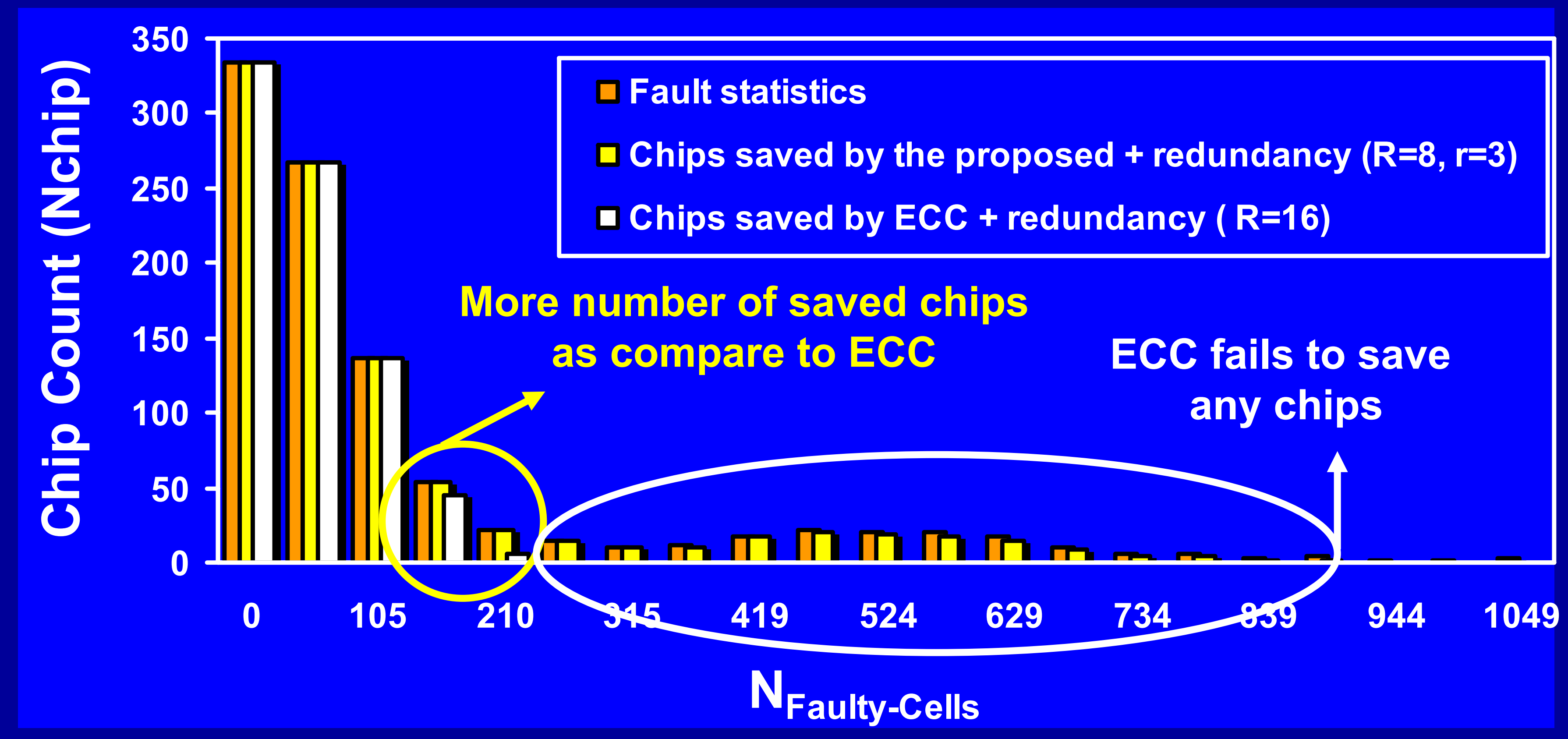


- BIST detects the faulty blocks
- Config Storage stores the fault information

Idea is to resize the cache to avoid faulty blocks during regular operation

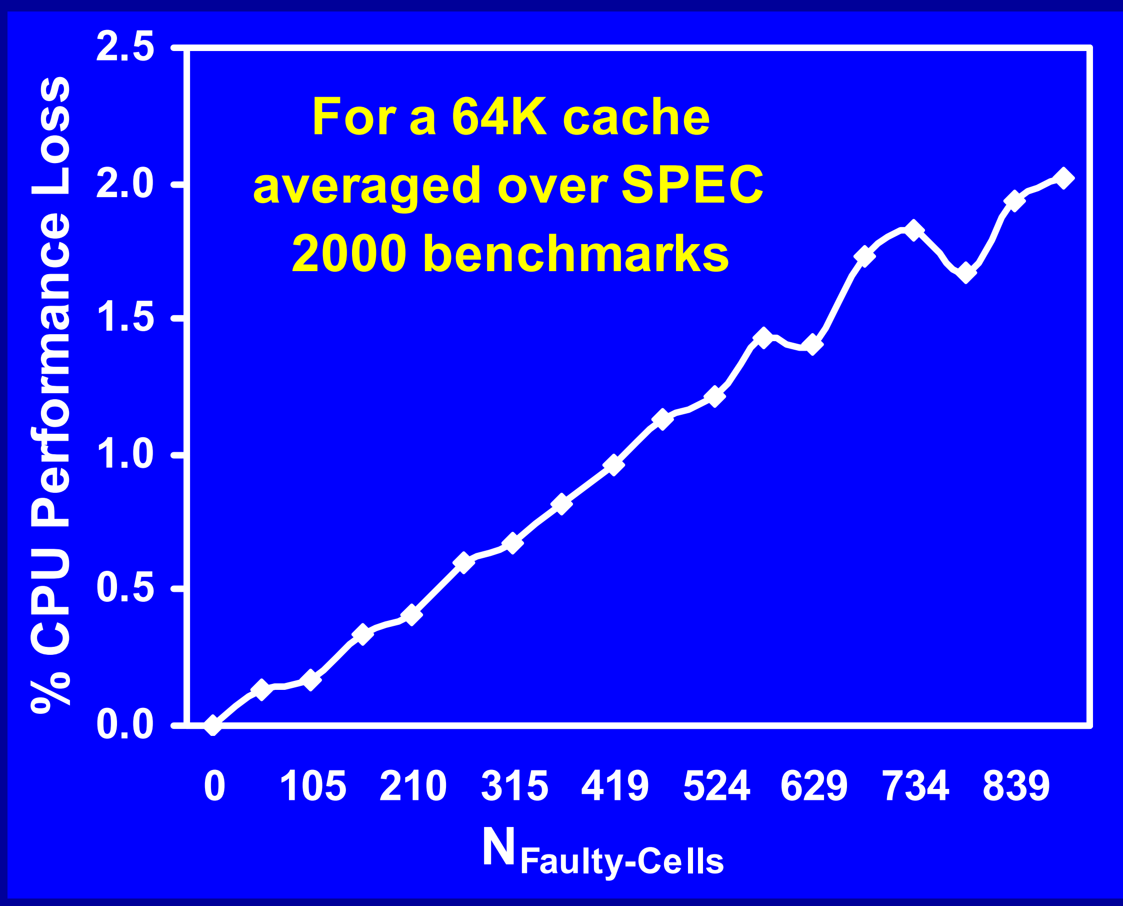

Mapping Issue
Address "one"
Address "two"
"T R 00 Off"
"T R 01 Off"
More than one INDEX are mapped to same block
Include column address bits into TAG bits
STORE D "one"
LOAD "two" Register
Tag matches b
TAG
Resizing is transparent to processor → same memory address
Colum

Fault Tolerant Capability



- Proposed architecture can handle more number of faulty cells than ECC, as high as 890 faulty cells
- Saves more number of chips than ECC for a given $N_{Faulty\text{-}Cells}$

CPU Performance Loss



- Increase in miss rate due to downsizing of cache
- Average CPU performance loss over all SPEC 2000 benchmarks for a cache with 890 faulty cells is ~ 2%

Logic: Process Tolerance

Sizing, Dual-Vt, Synthesis, etc.

- Mostly worst-case design methdology
- Guard-banding with larger transistors/larger library elements
- Selective guard-banding to reduce delay in longer paths
  - Transistor sizing
  - Dual-Vt
- Logic optimization and simultaneous sizing/dual-Vt

Logic: A New Paradigm for Low-Voltage, Variation Tolerant Circuit Synthesis Using Critical Path Isolation (CRISTA)
-- Better than worst-case design

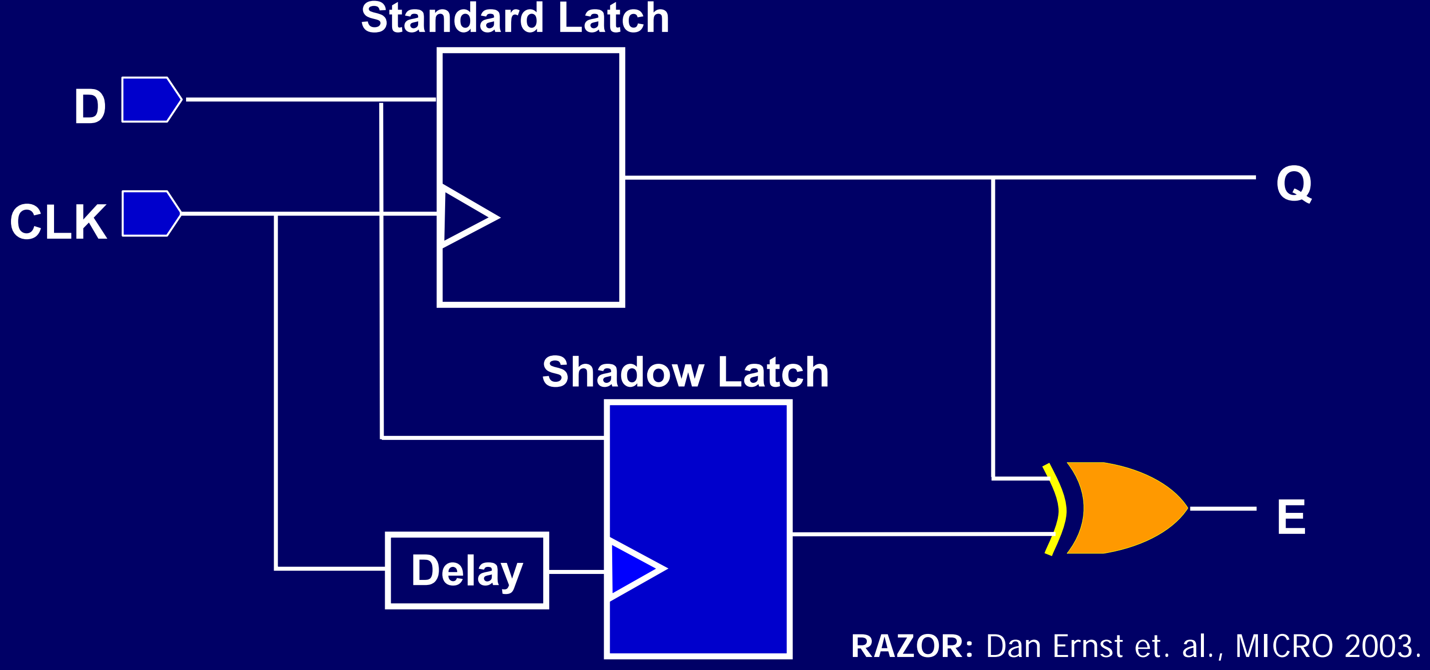
Ghosh, Bhunia, Roy -- ICCAD 2006

Vdd Scaling and Process Tolerance: Conventional Solutions

- Low power:
  - Reduce the supply voltage
    - Error rate increases
  - Dual-Vt/dual-VDD assignment
    - Number of critical paths increases
- Robustness:
  - Increase supply voltage
    - Power dissipation increases
  - Upsize the gates
    - Switching capacitance increases

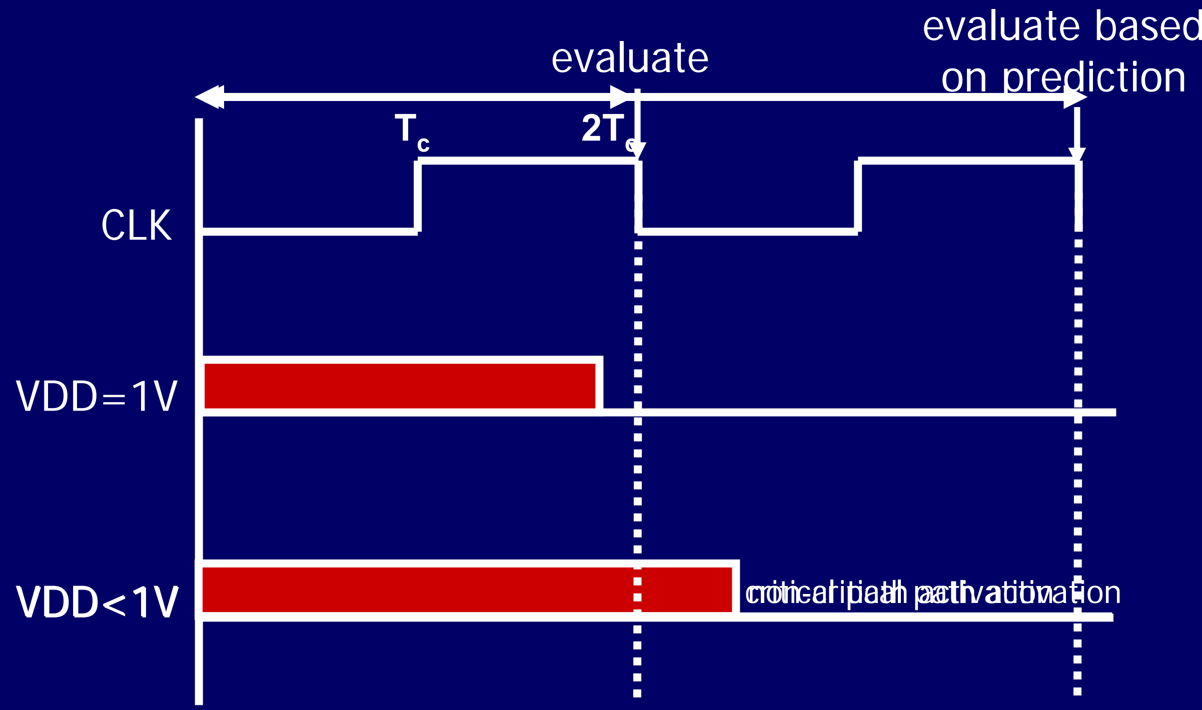
Low power and robustness: conflicting requirements

Razor Approach



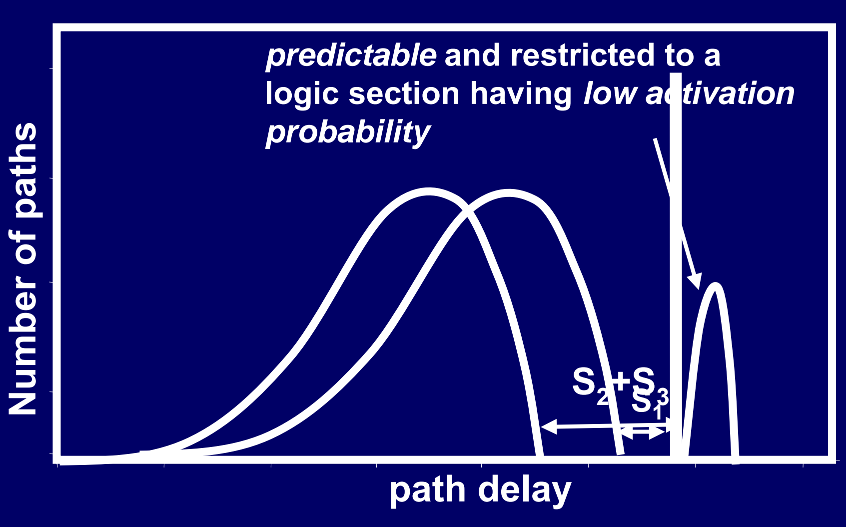
- *Post-Silicon* technique for *dynamic* supply scaling and *timing error* detection/correction
- Error correction overhead is 1% for a 10% error rate

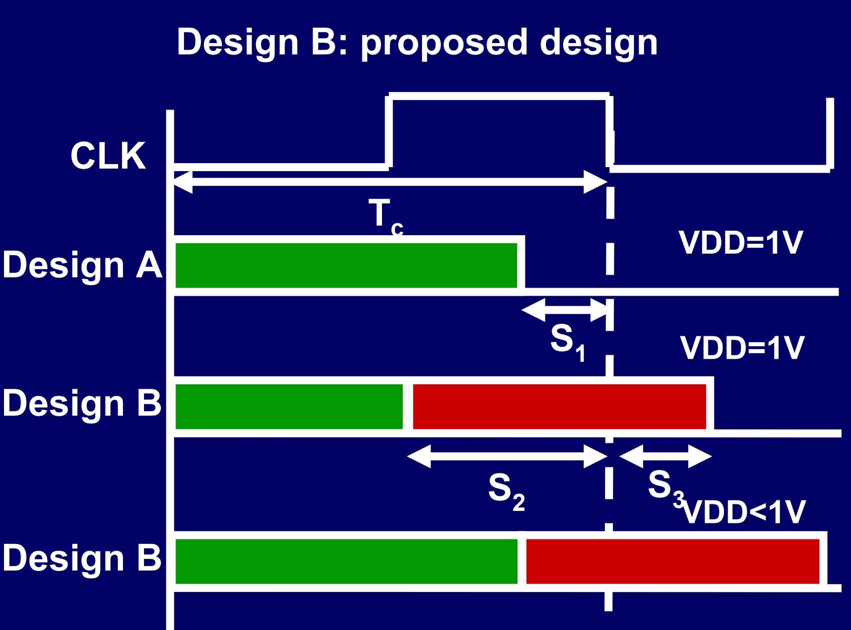
CRISTA: Basic Idea



- Important points:
  - Design time technique (as opposed to post-Si)
  - Scale down the supply while making ***delay failures predictable***
  - ***Avoid*** the failures by ***adaptive clock stretching***
  - Ensure that critical paths are activated ***rarely***

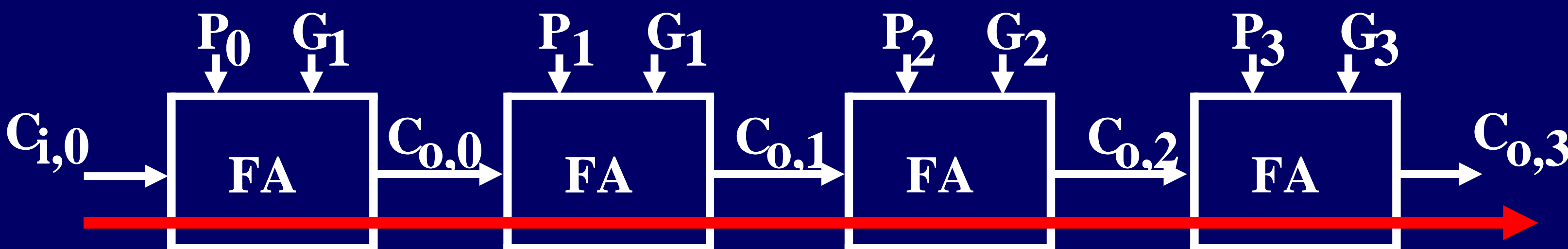
Design Considerations for CRISTA





- Few predictable critical paths
- Low activation probability of critical paths
- Slack between critical and non-critical paths under variations

Case Study: Adder



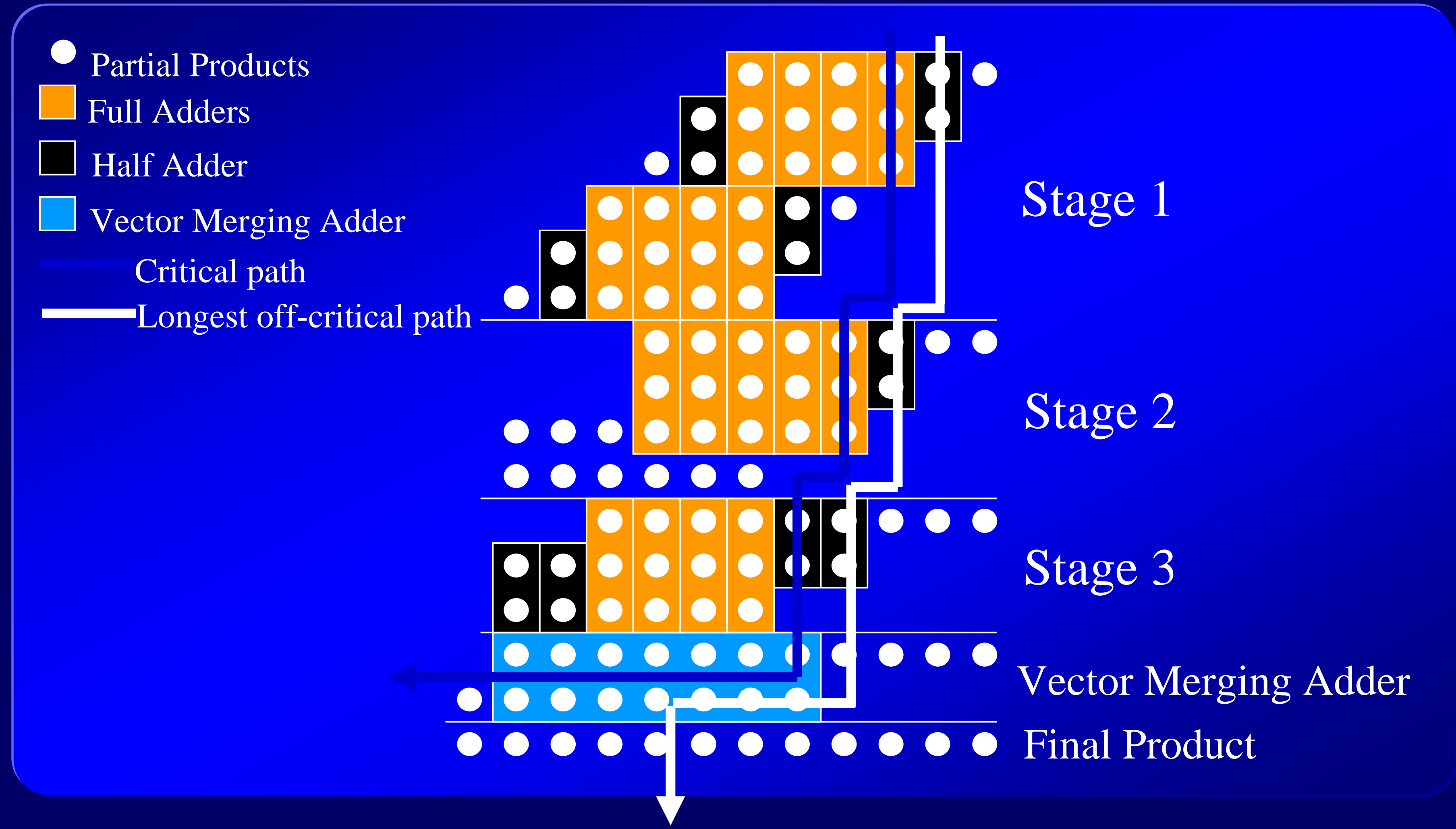
- Interesting features:
  - Single critical path (activated by $P_0P_1P_2P_3=1$ & $C_{i,0}=1$)
  - Low activation probability of critical path

VDD = 1V, TCLK = 260ps	VDD = 0.8V, TCLK = 260ps
Crit. path delay=260ps longest non-crit. path delay=165ps *P = 13uW (1-cycle)*	Crit. path delay=330ps longest non-crit. path delay=260ps *P = 7.4uW (rare 2-cycles, decoder)*

44% power saving by reducing voltage and, operating critical path at 2-cycle and other paths at 1-cycle

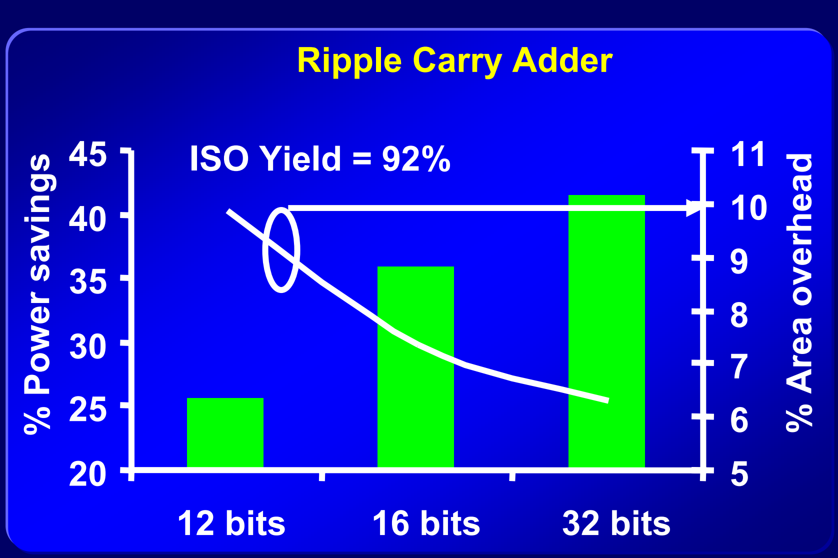
Can we apply same technique to any random logic?

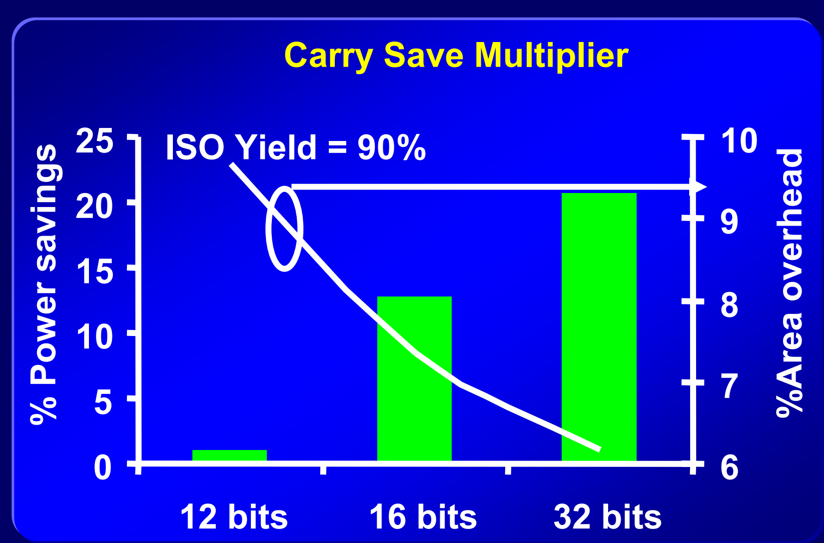
Wallace Tree Multiplier

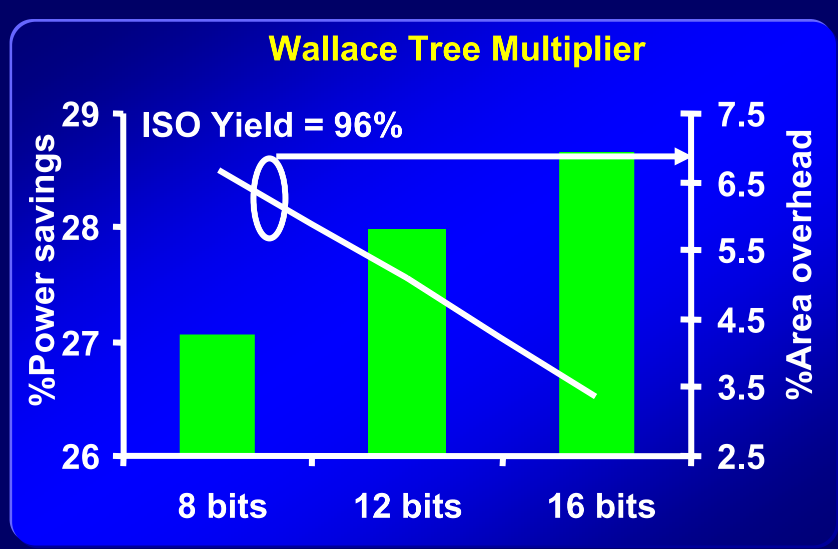


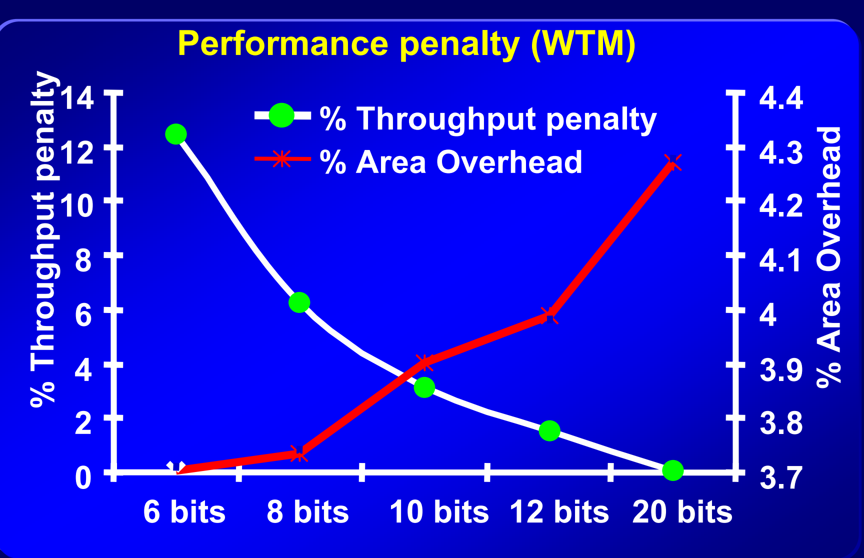
- 29% power saving with ~4% area overhead

Simulation Results



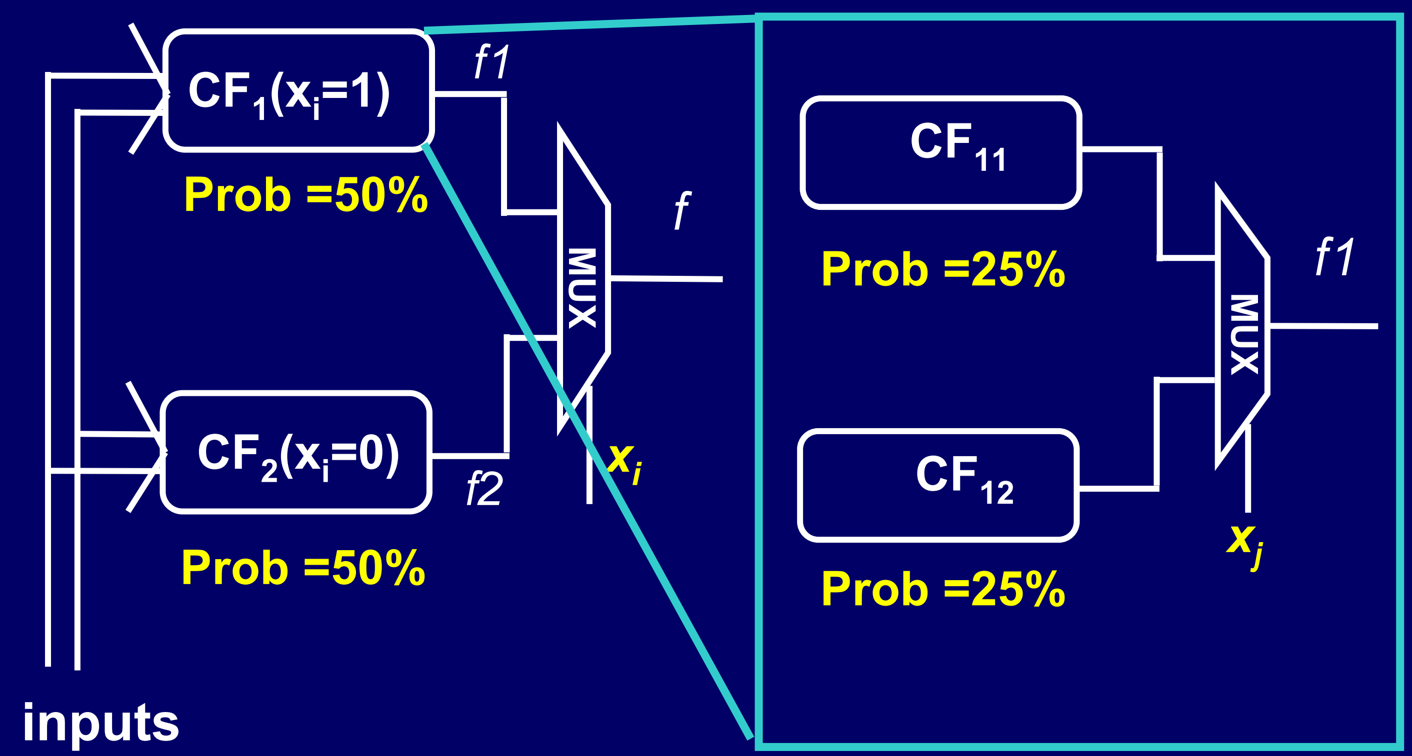






Random Logic: Shannon's Expansion

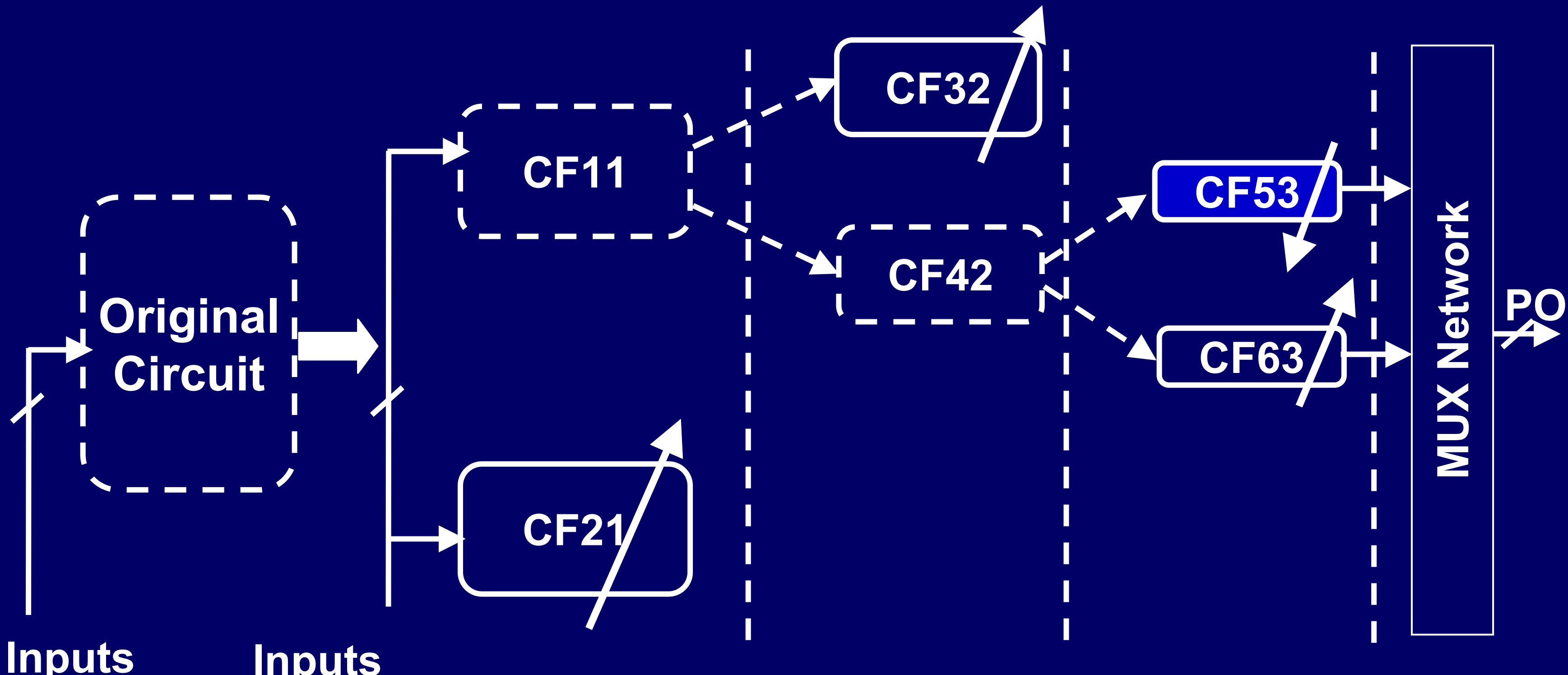
$$f(x_1,\ldots,x_i,\ldots,x_n) = x_i \bullet f(x_1,\ldots,x_i=1,\ldots,x_n) + x_i' \bullet f(x_1,\ldots,x_i=0,\ldots,x_n)$$
$$= x_i \bullet CF_1 + x_i' \bullet CF_2$$
$$CF_1 = f(x_1,\ldots,x_i=1,\ldots,x_n); \quad CF_2 = f(x_1,\ldots,x_i=0,\ldots,x_n)$$



Activation probability of cofactors can be reduced

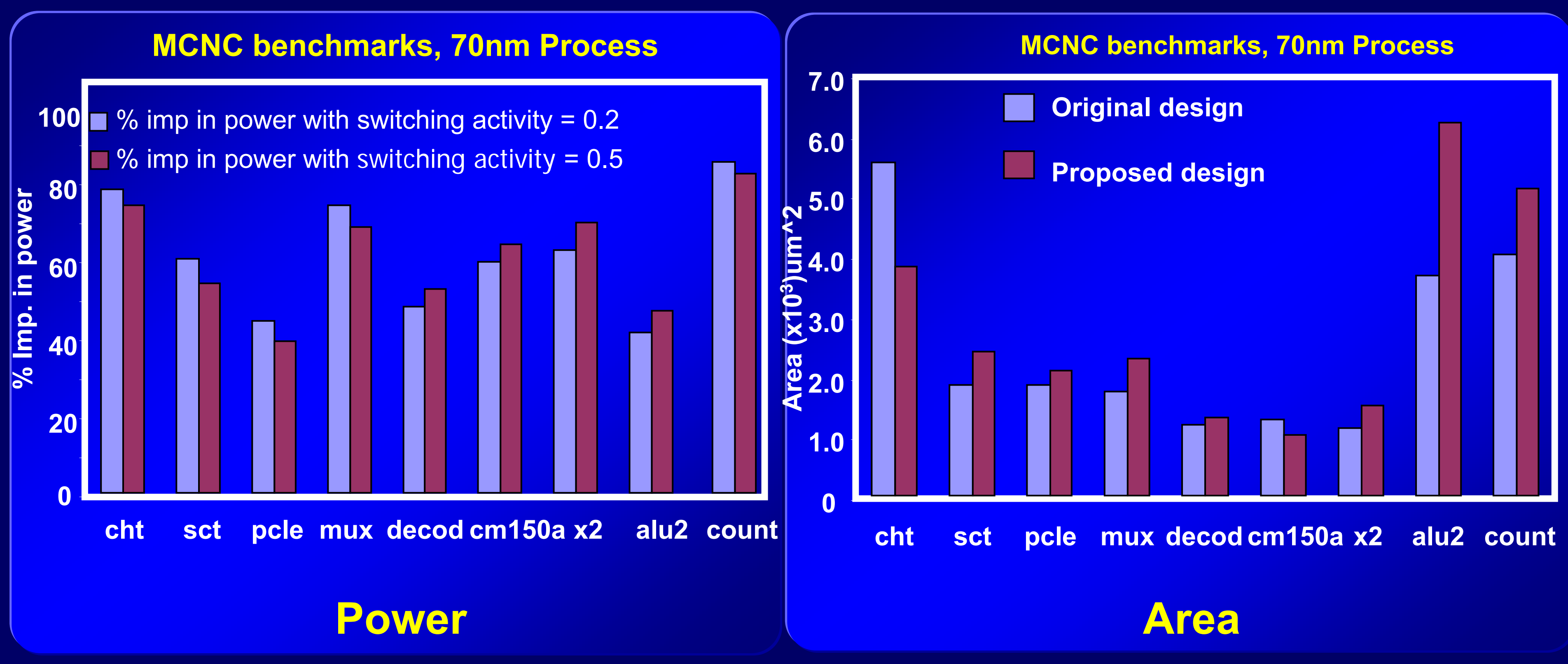
How to choose *Control Variable* ?

Further Isolation and Slack Creation by Sizing



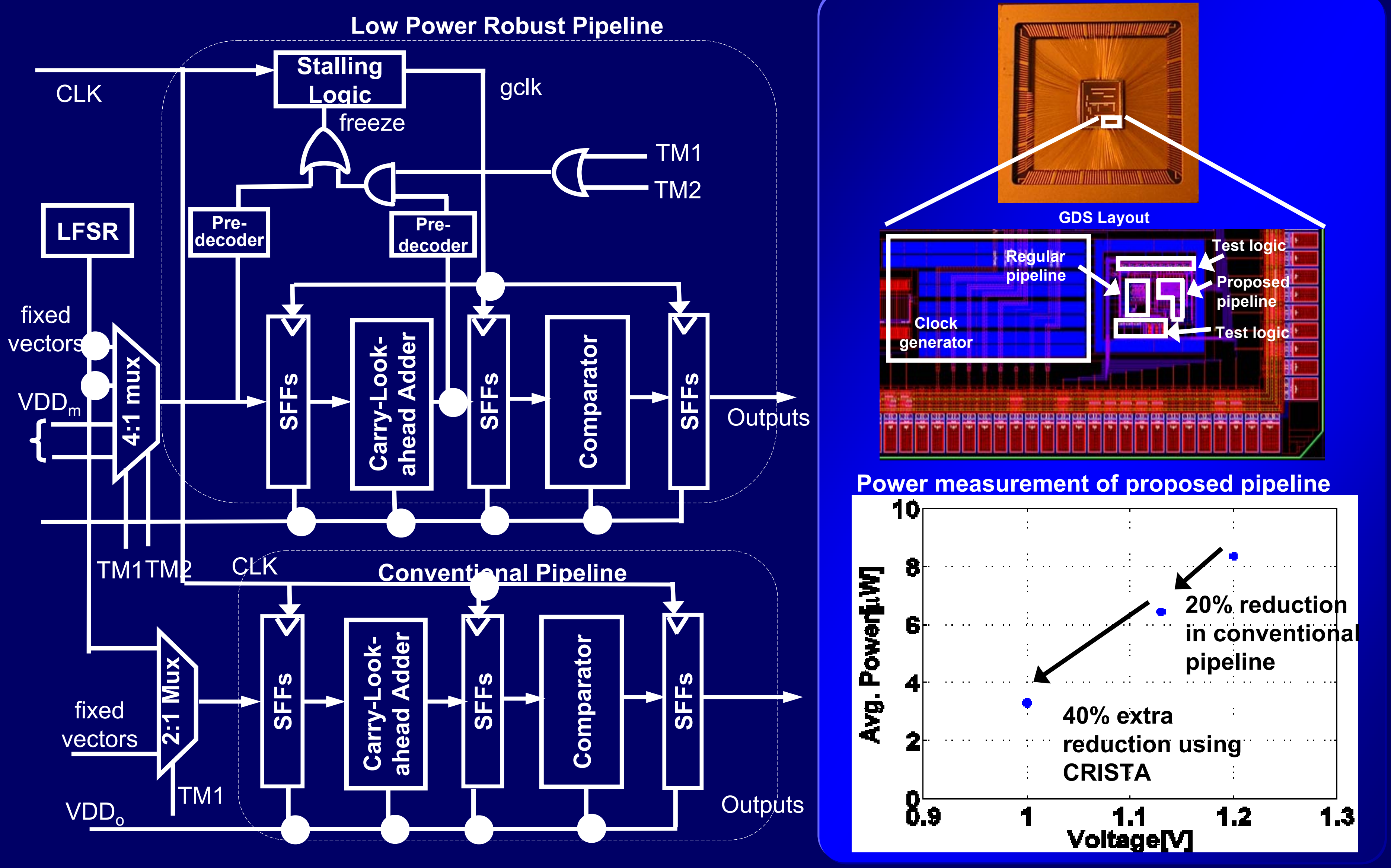
- Slack creation strategy
  - Lagrangian Relaxation based sizing (*B.C. Paul et. al., DAC 2004*) is used
  - Non-critical paths are selectively made faster
  - Critical paths are slightly slowed down

Simulation Results



- Average power saving = ~45%
- Average area overhead = 18%
- Avg performance penalty=5.9% (with 4 control variables) for signal prob=0.5

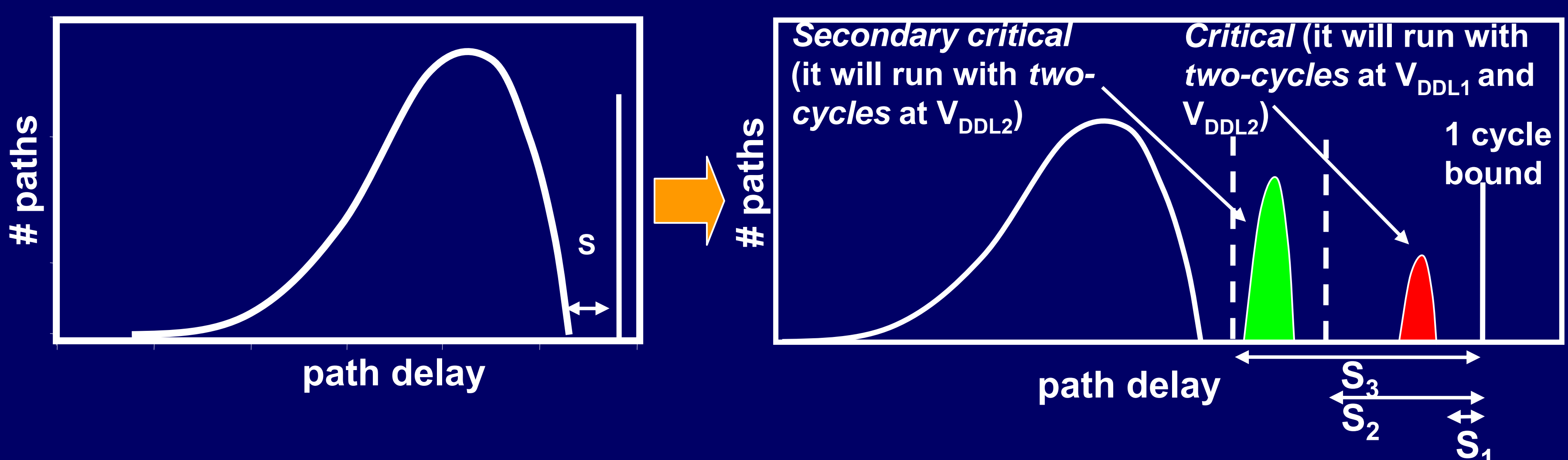
Two-Stage Pipeline with Test Logic



~40% power saving with ~13% performance penalty

CRISTA in Thermal Design

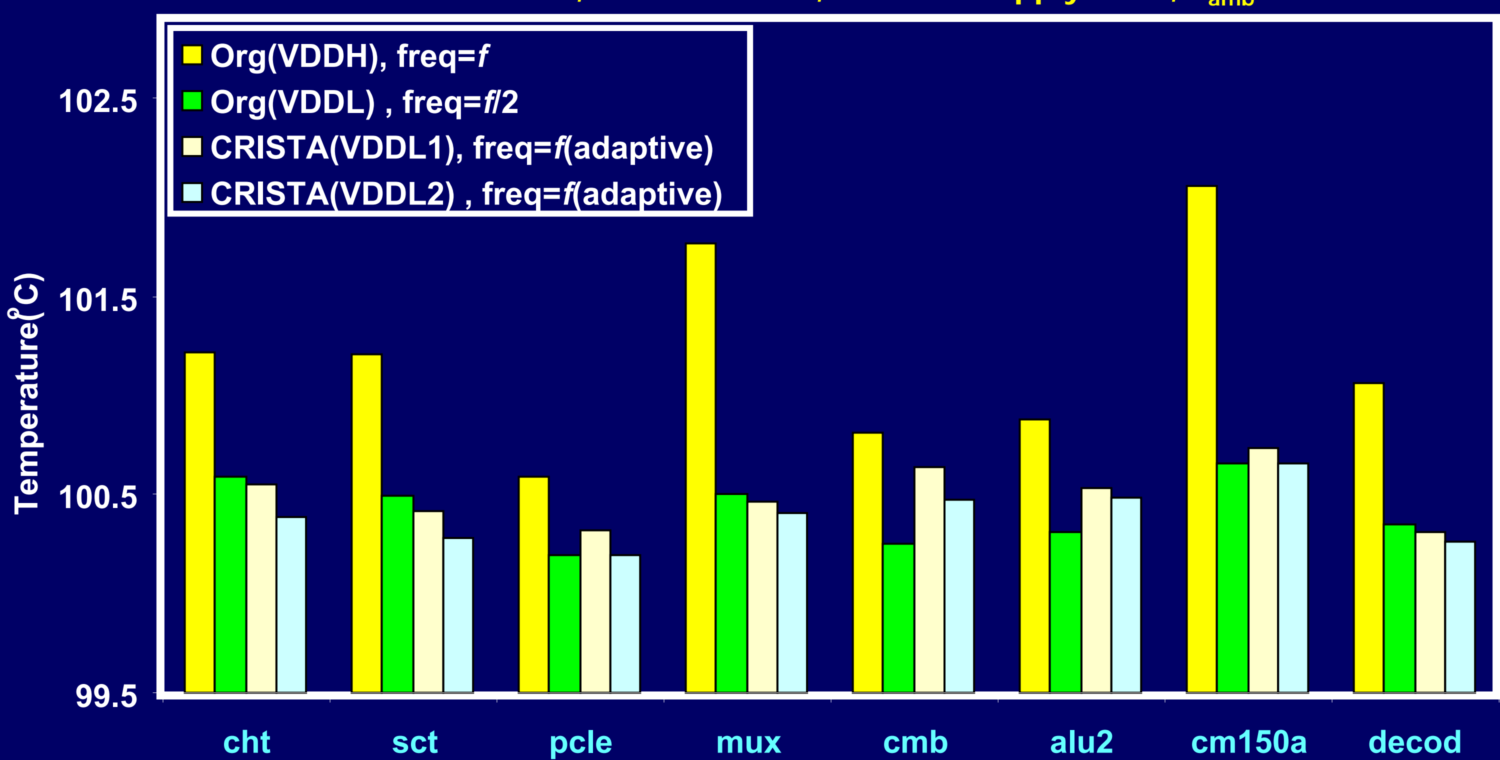
Temperature Aware Design using CRISTA*



- Scale down the supply at elevated temperatures while making *delay failures predictable*
- Maintain *rated* frequency
- *Avoid* the failures by *adaptive clock stretching*
- Ensure that critical paths are activated *rarely*
- Tradeoff between temperature and performance loss

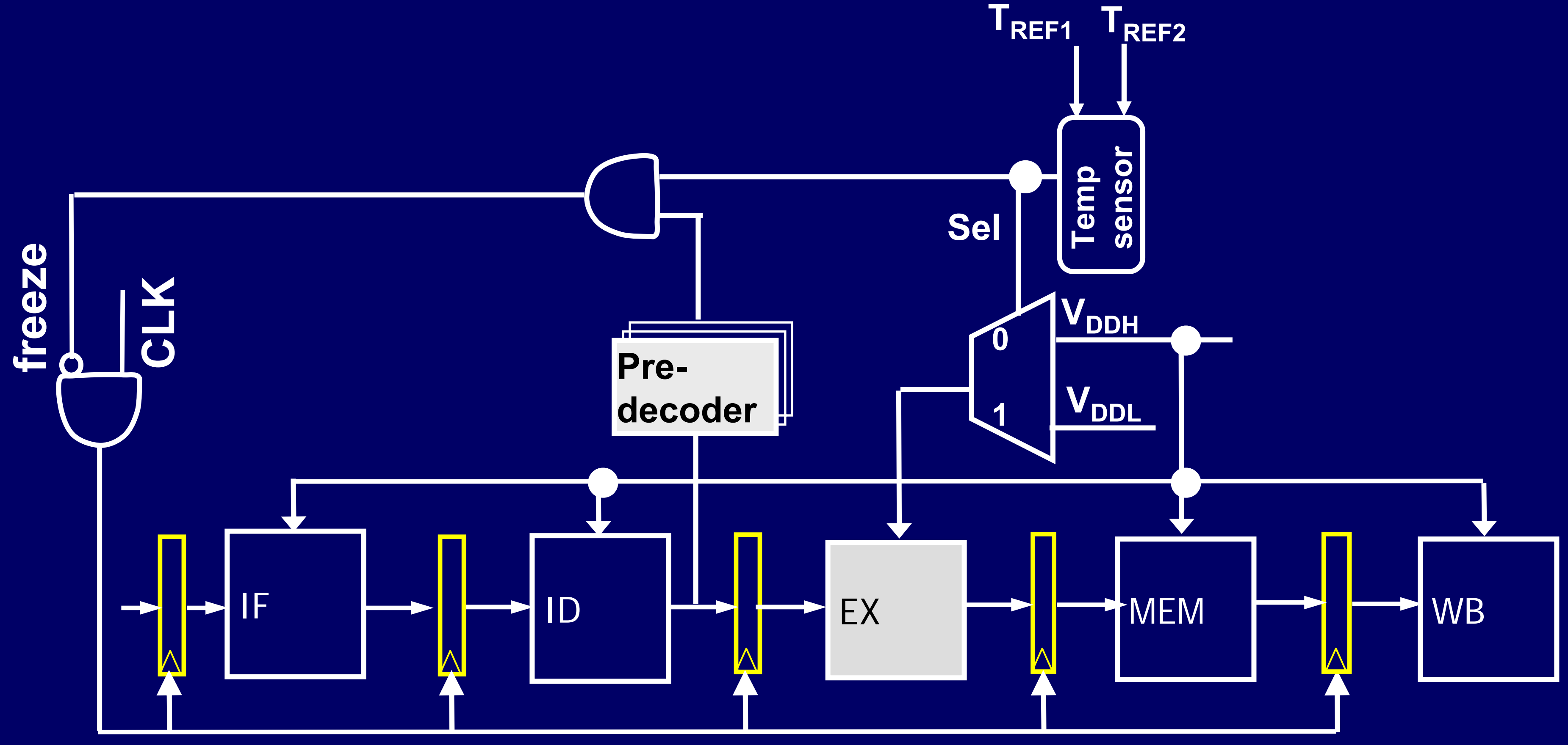
S Ghosh, S Bhunia and K Roy, Low overhead circuit synthesis for temperature adaptation using dynamic voltage scheduling, DATE 2007.

Simulation Results



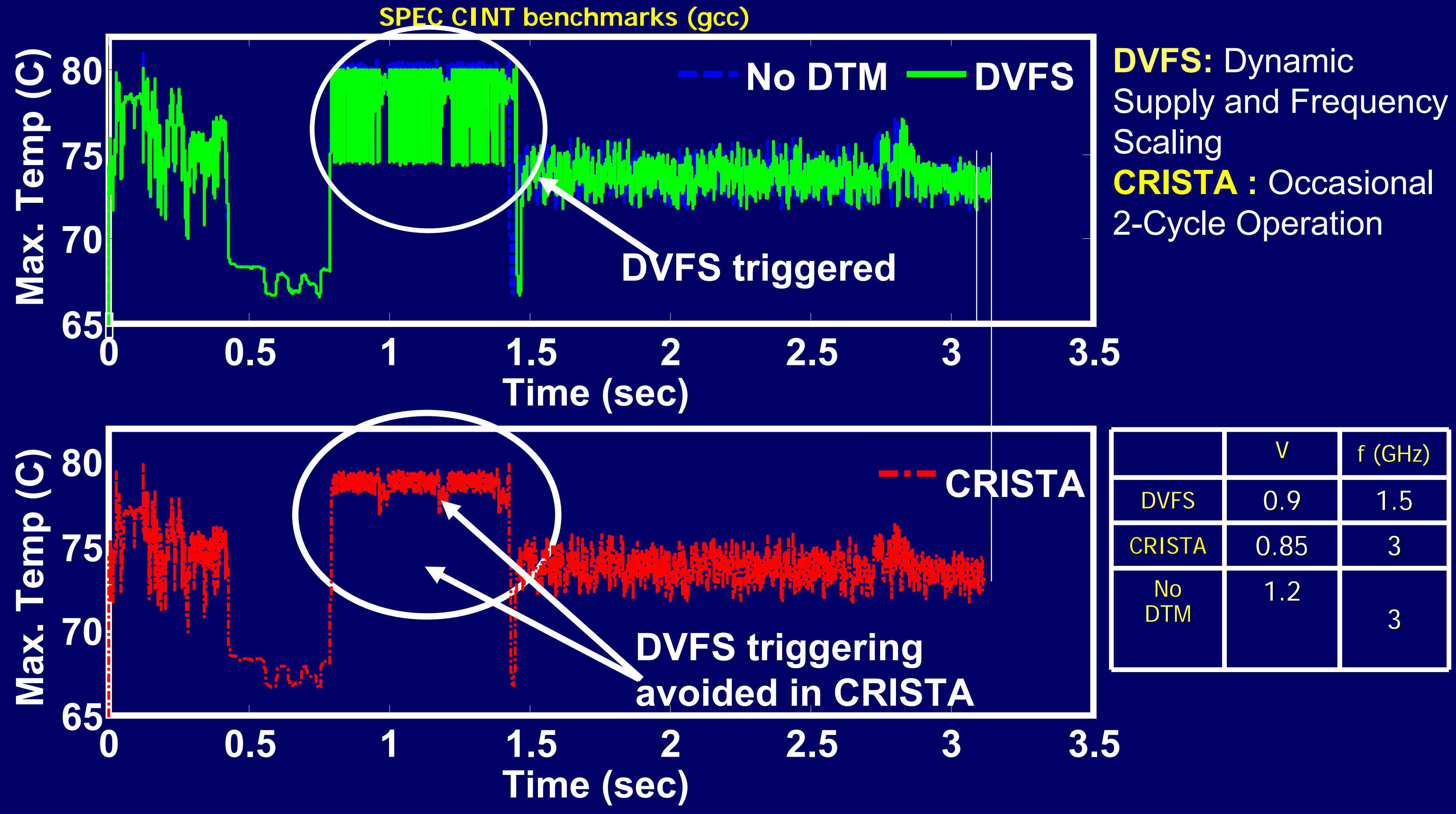
- Similar temperature reduction as conventional technique (i.e. supply scaling with half frequency)
- Avg performance penalty=5.9% (at V_{DDL1}) and 15.2% (at V_{DDL2}) compared to 50% in conventional technique
- Avg area overhead ~14%

Micro-Architectural Implementation



- Applied to execution stage (adders and multipliers) of in-order pipeline
- Can be an *add-on* to standard DVFS

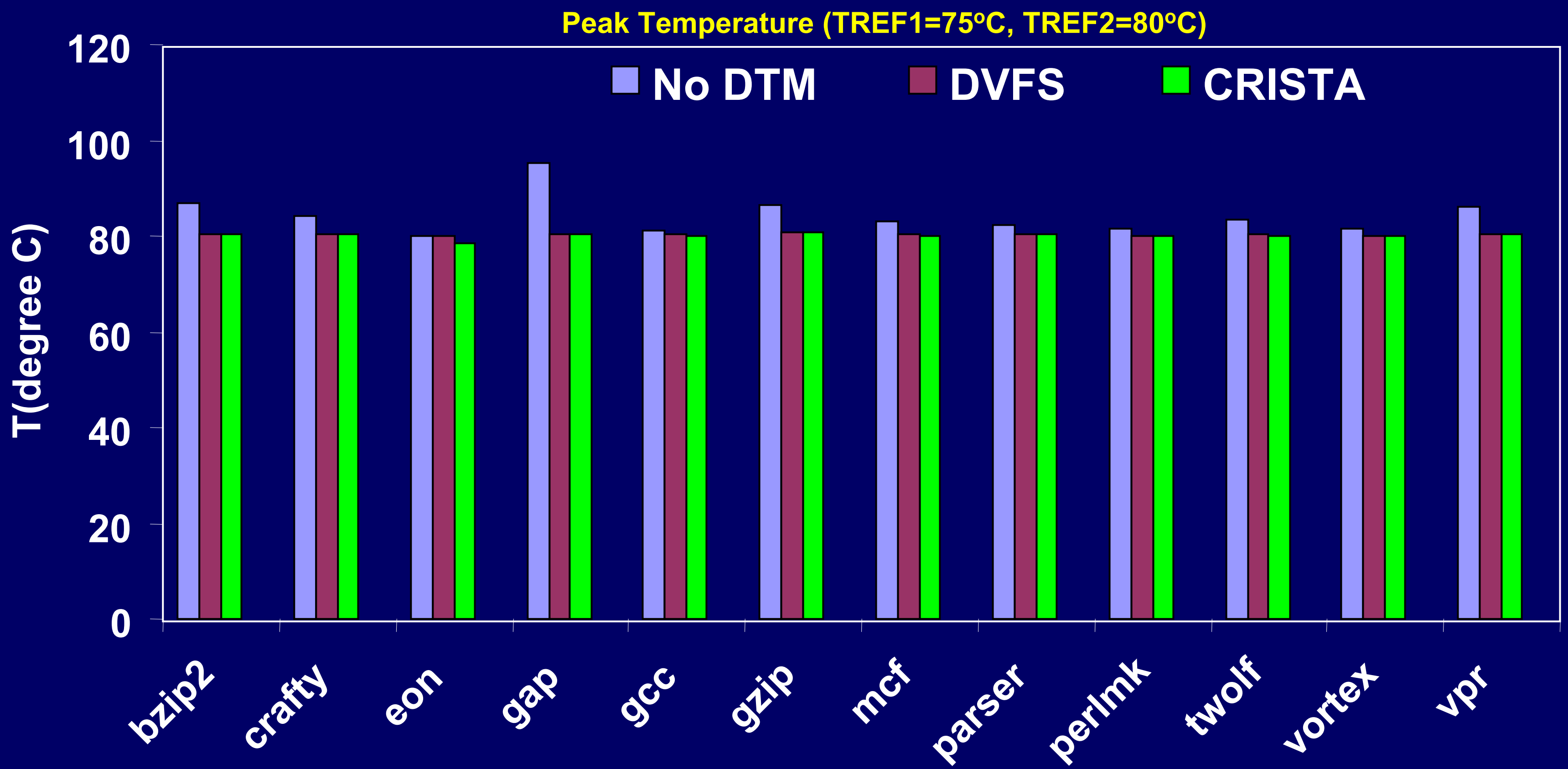
Simulation Results



	V	f (GHz)
DVFS	0.9	1.5
CRISTA	0.85	3
No DTM	1.2	3

- Avoids requirement of full-chip DVFS

Simulation Results



- Average reduction in peak temperature ~6.6% with ~3.4% throughput loss and ~4.2% area overhead

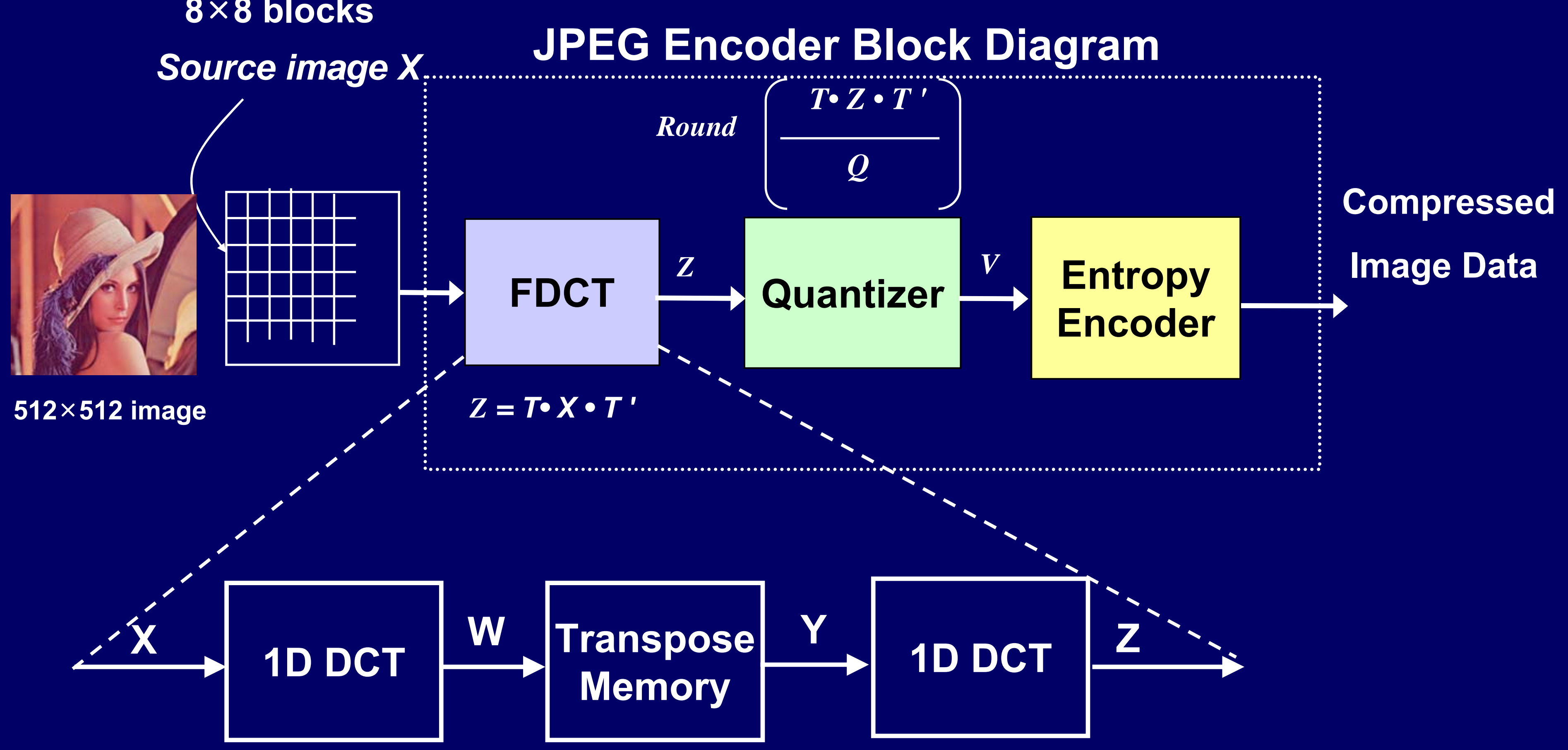
VDD Scaling, Process Variation, and Quality Trade-off: DCT
-- Quality knob for tuning

Banerjee, Karakonstantis, Roy
Design Automation and Test in Europe (DATE) 2007

Basic Idea

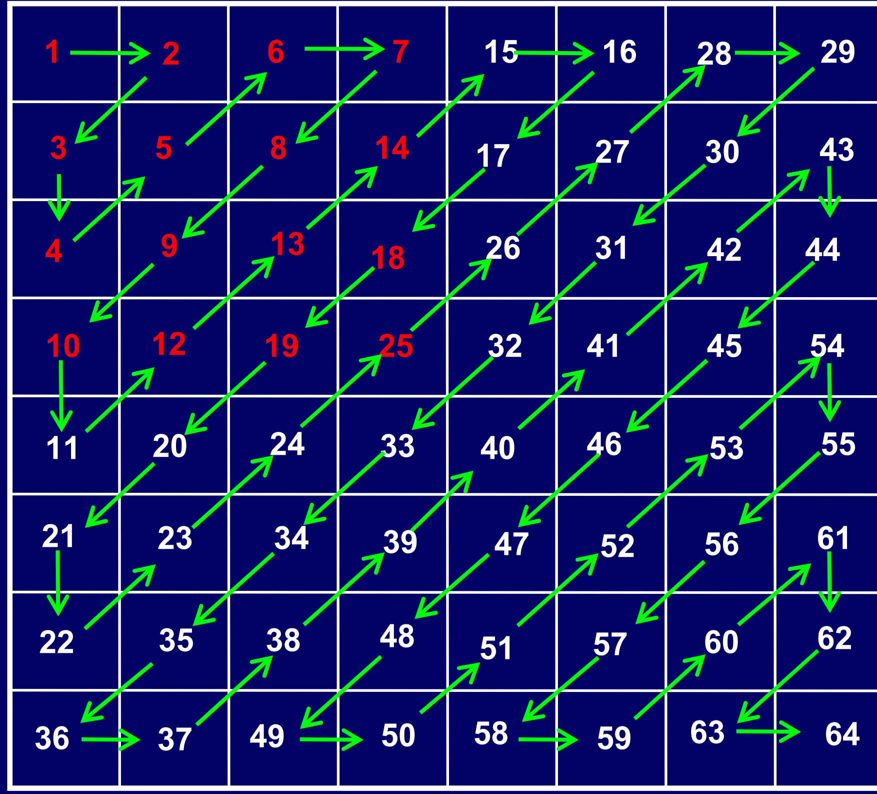
- All computations are “not equally important” for determining outputs
- Identify important and unimportant computations based on output “sensitivity”
- Compute important computations with “higher priority”
- Delay errors due to variations/ Vdd scaling “affect only” non-important computations
- “Gradual degradation” in output with voltage scaling and process variations

DCT Based Image Compression Process



- DCT is used in current international image/video coding standards
  - JPEG, MPEG, H.261, H.263

Energy Distribution of a 2D-DCT Output

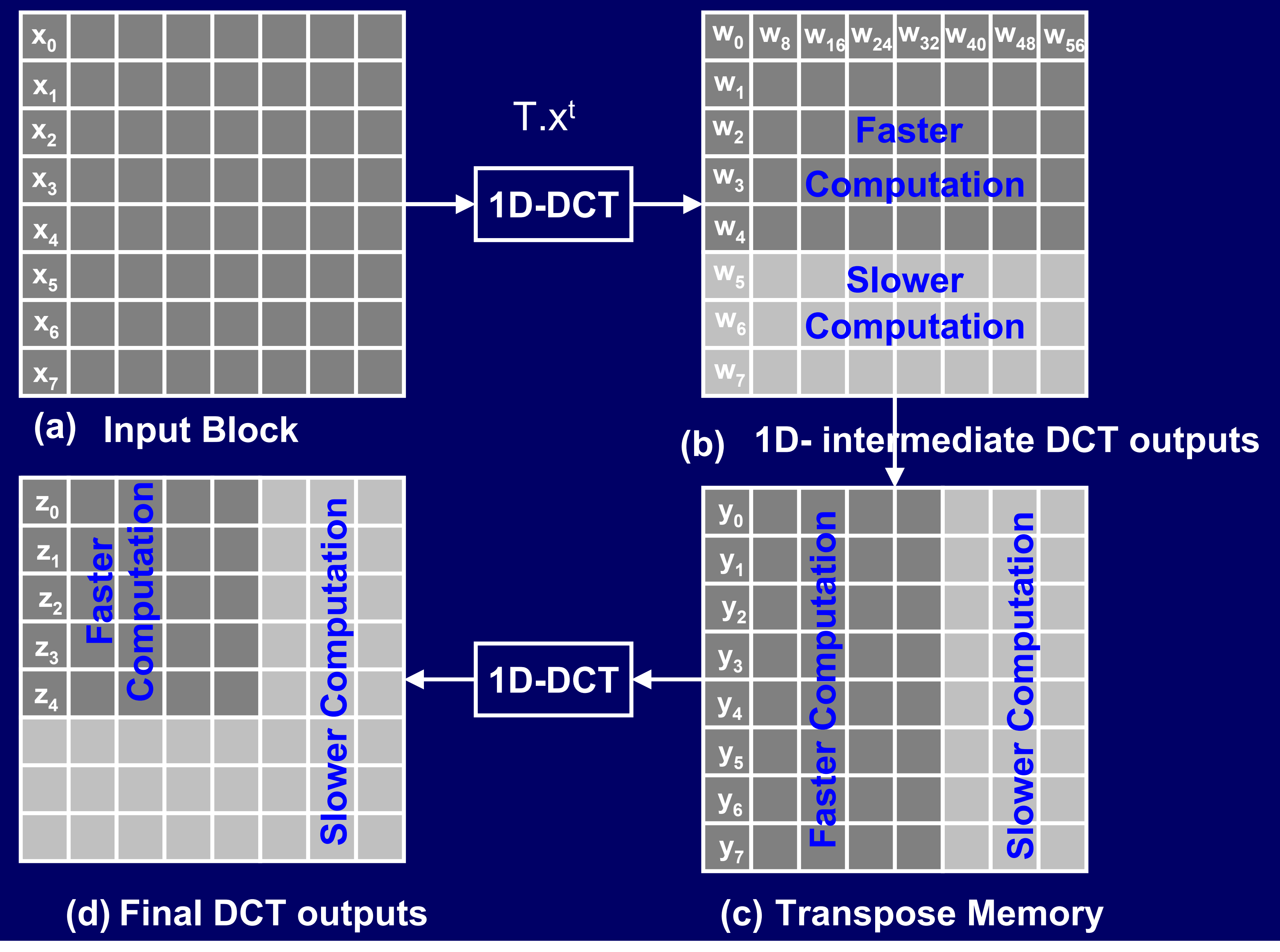


High energy components (important outputs 75% energy)

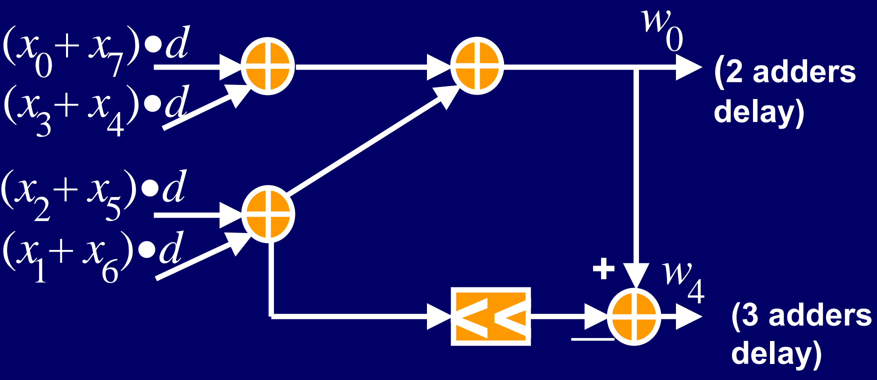
Low energy components (less important outputs)

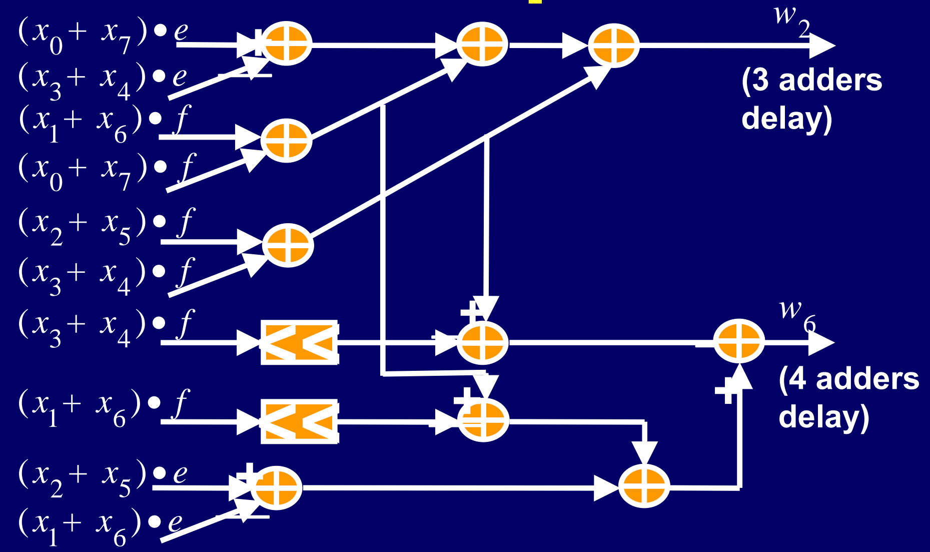
Can important components be computed with higher priority ?

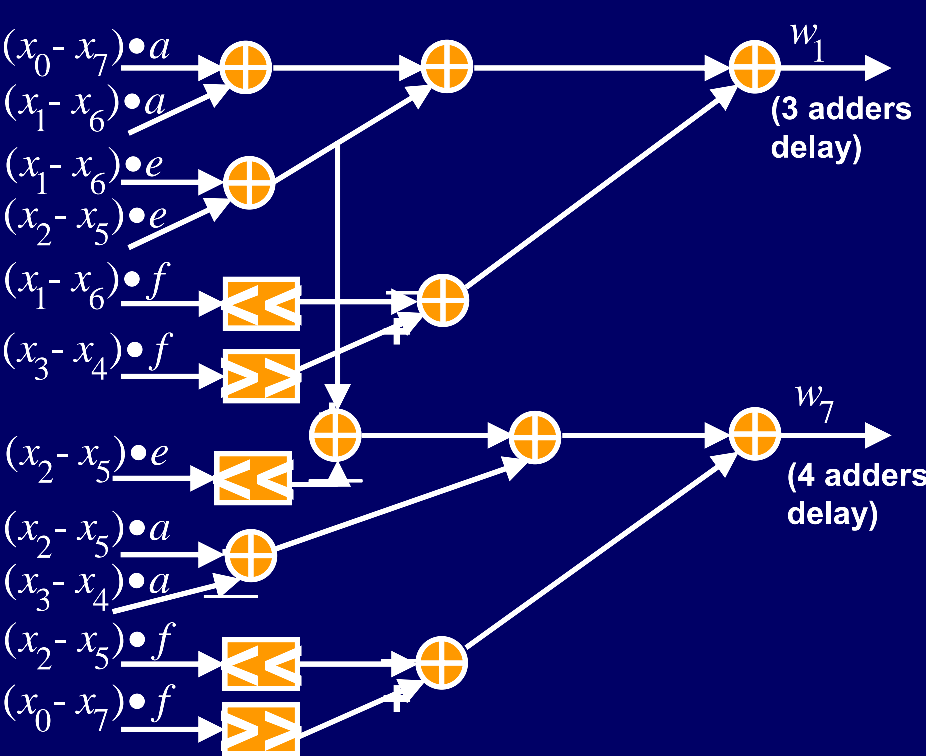
Design Methodology

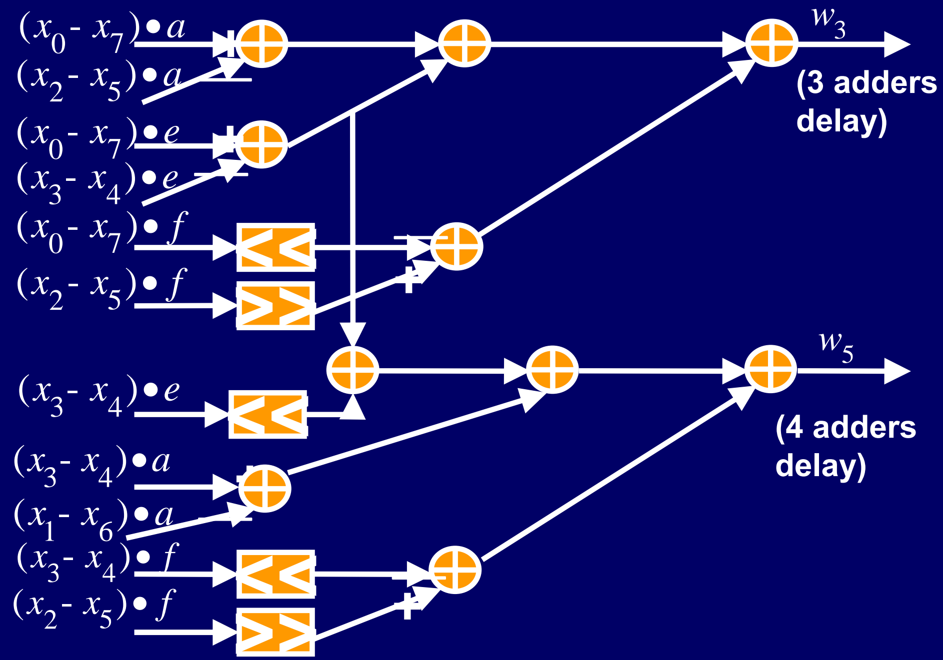


Path Delays for 1D-DCT outputs





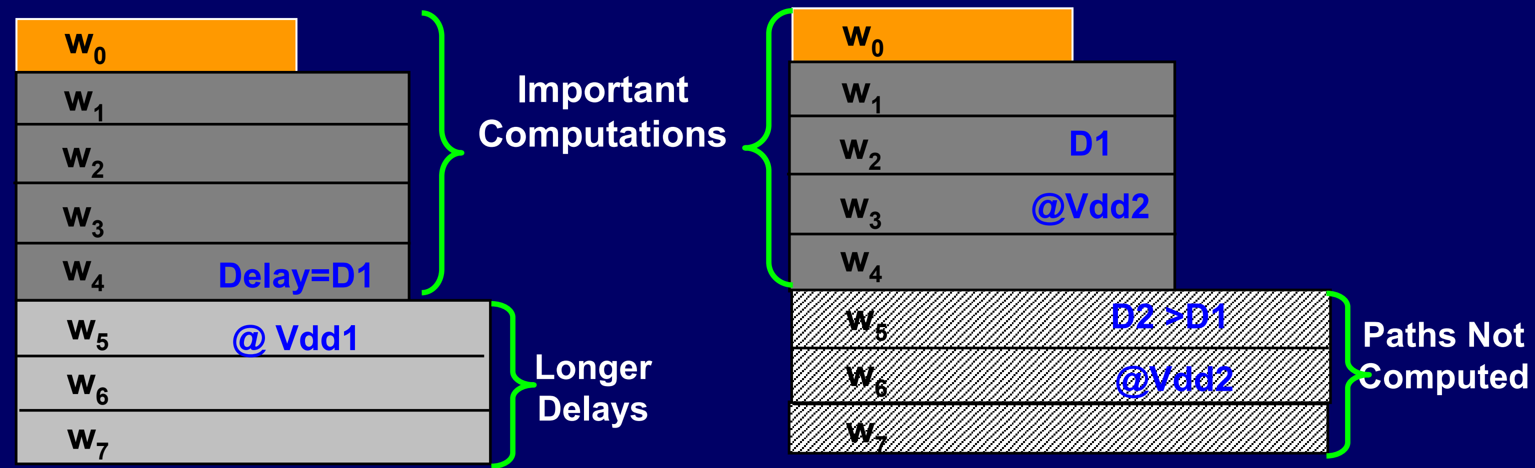


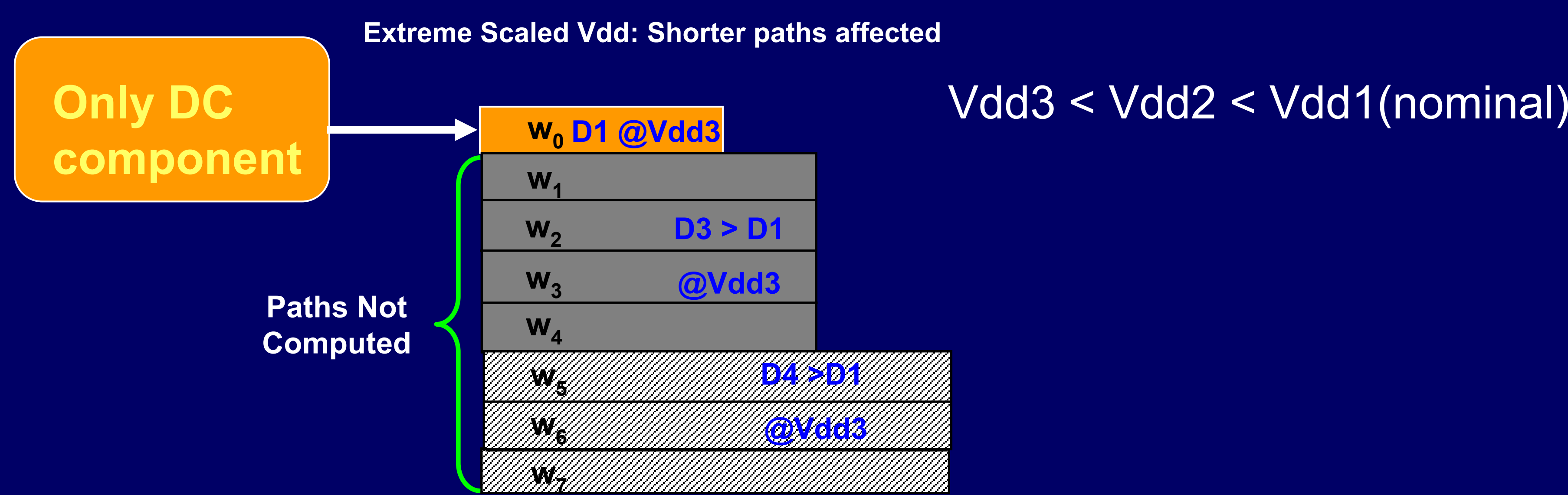


Proposed DCT under Vdd scaling

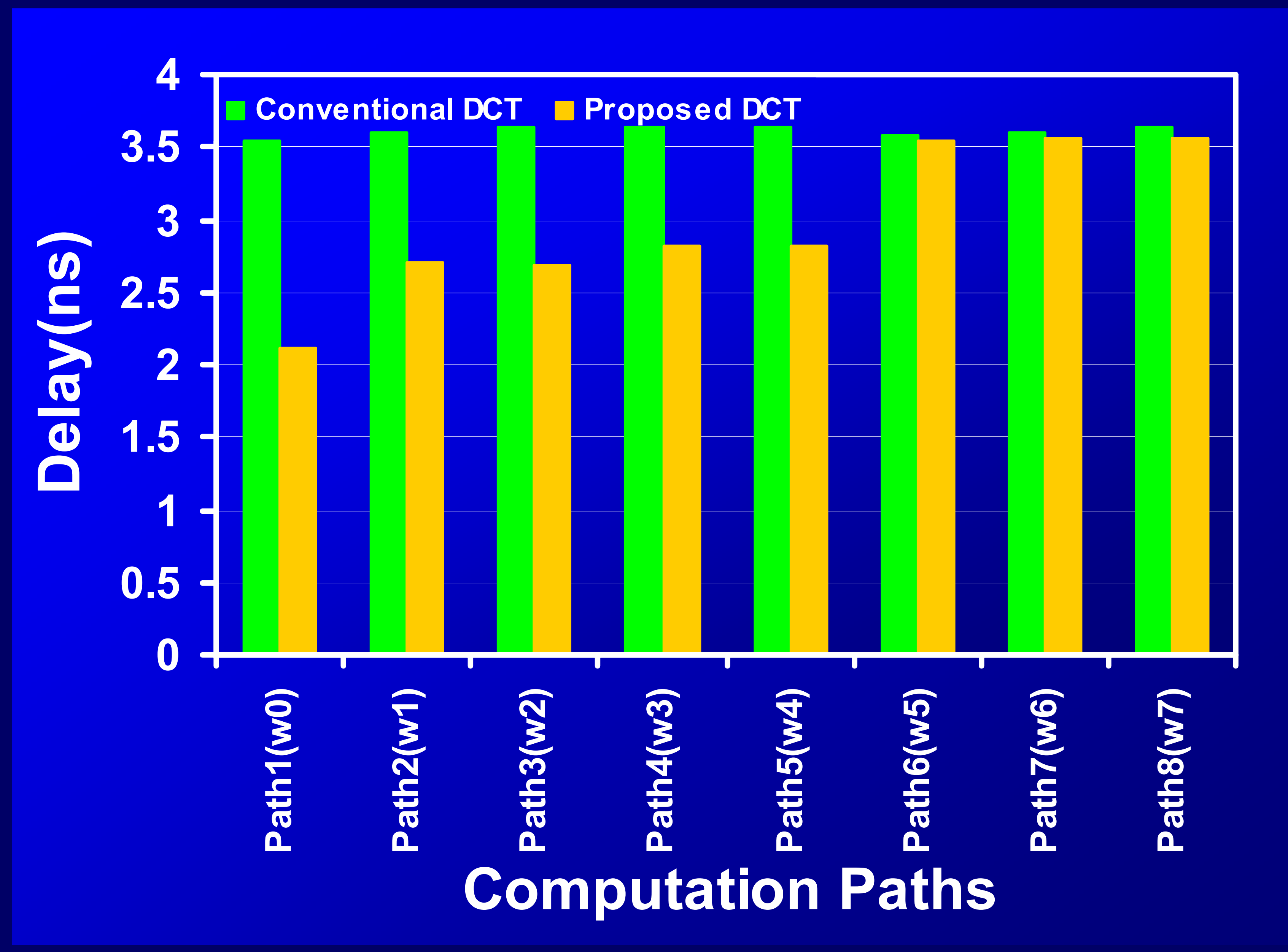
Proposed Design with high/low delay paths

Scaled Vdd: Longer paths under Vdd scaling





1D-DCT Path Delay Comparisons



Effect of Vdd Scaling

	Conventional WTM DCT	CSHM DCT (2 alphabet)	Proposed DCT
1.0 V			
0.9 V	FAILS	FAILS	
0.8 V	FAILS	FAILS	

Different Architectures at Nominal Voltage

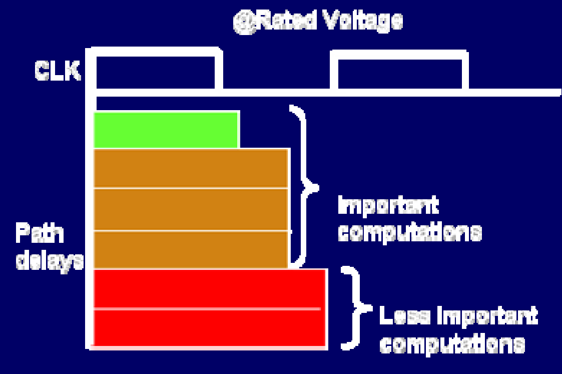
1.0V	CSHM DCT (2 alphabets)	DCT with WTM	Proposed DCT
Power (mW)	25.1	29.8	26
Delay (ns)	3.2	3.64	3.57
Area (um^2)	80490	108738	90337
PSNR (dB)	21.97	33.23	33.22

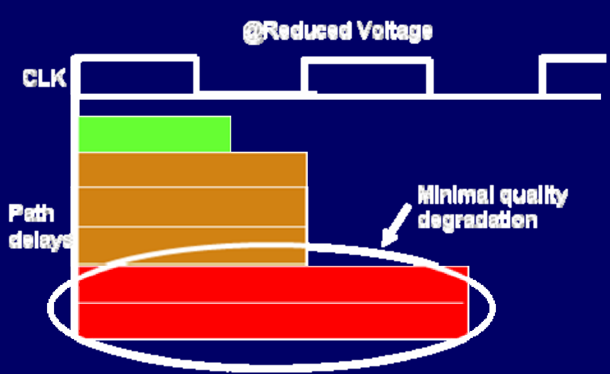
Proposed Architecture at Reduced Voltage

	Proposed DCT Vdd=0.9V	Proposed DCT Vdd=0.8V
Power (mW)	17.53(41.2%)	11.09(62.8%)
PSNR (dB)	29	23.41

- Graceful degradation of proposed DCT architecture under Vdd scaling (Vdd can be scaled to 0.75V)
- Conventional architectures fails

CRISTA for DSP Systems





- Scale down supply voltage
- Design/algorithm such that failures can only occur in less critical section -- minimal quality degradation

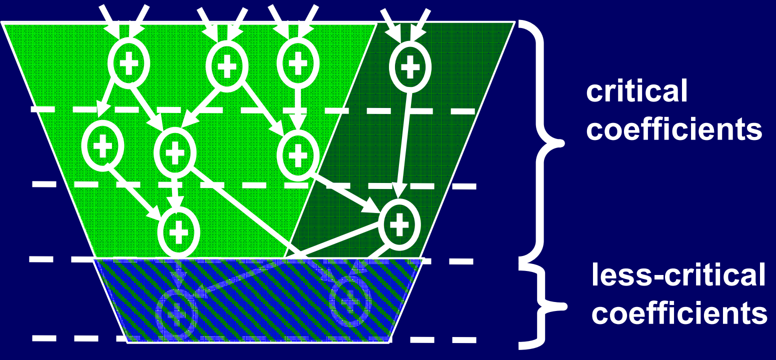
1. Discrete Cosine Transform (DCT)

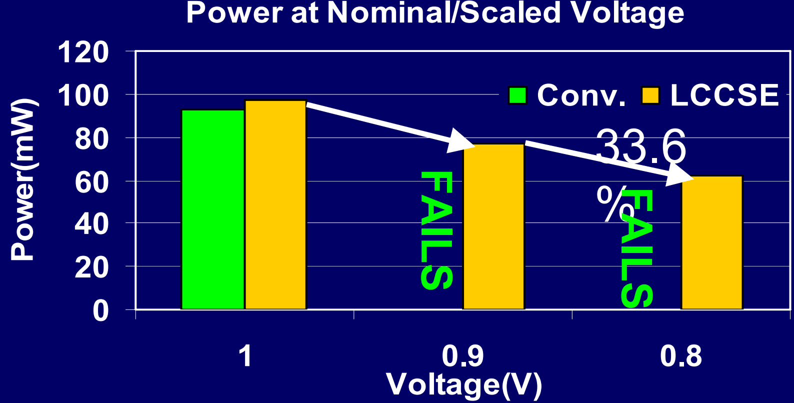
	Conventiona l WTM DCT	CSHM DCT (2 alphabet)	Proposed DCT
1.0 V			
0.9 V	FAILS	FAILS	
0.8 V	FAILS	FAILS	

- Graceful degradation in quality under Vdd scaling (Vdd can be scaled to from 1V to 0.8V) with ~60% improvement in power
- Conventional architectures *fail*

For Other DSP Systems

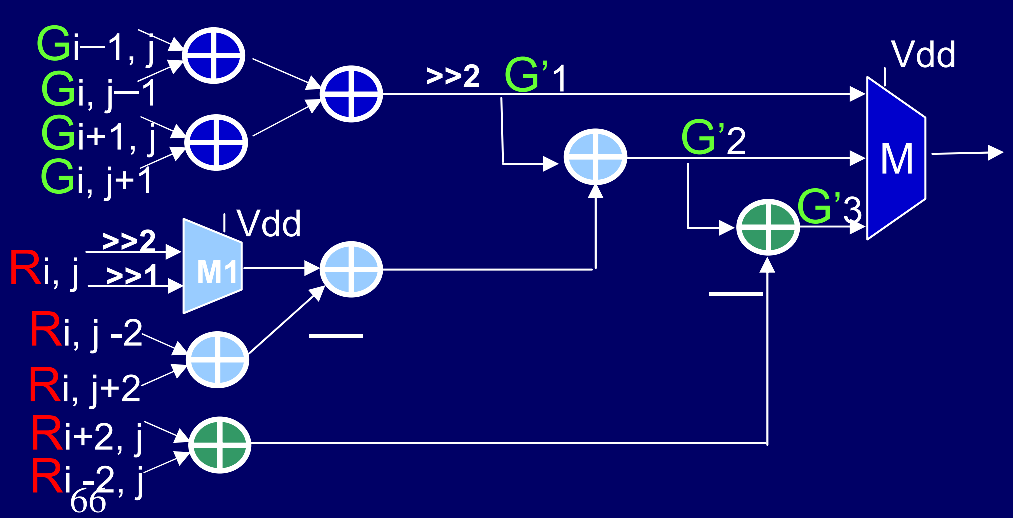
2. Finite Impulse Response (FIR)





3. Color Interpolation

- Bilinear component is critical and gradient component is less-critical
- Design architecture such that failures can only occur in gradient term

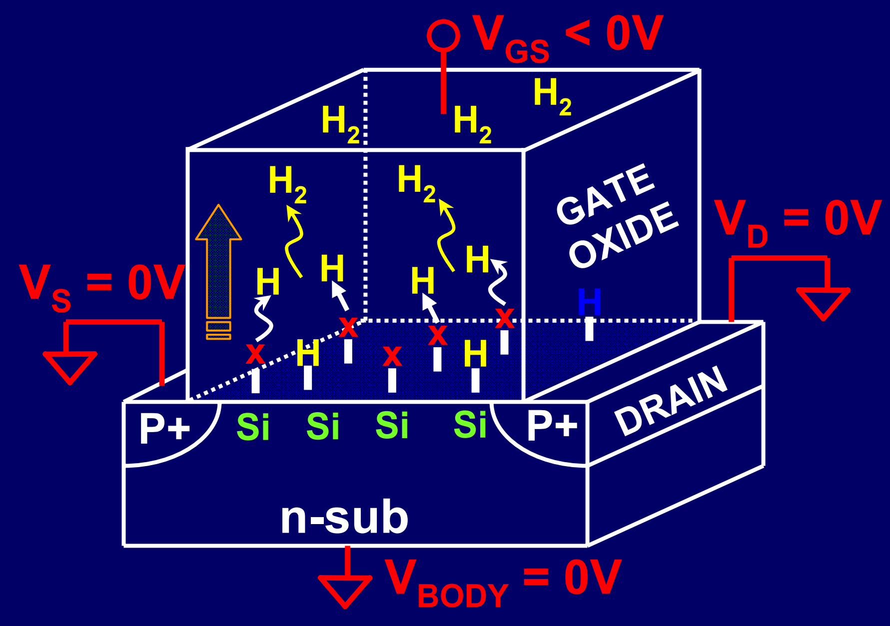


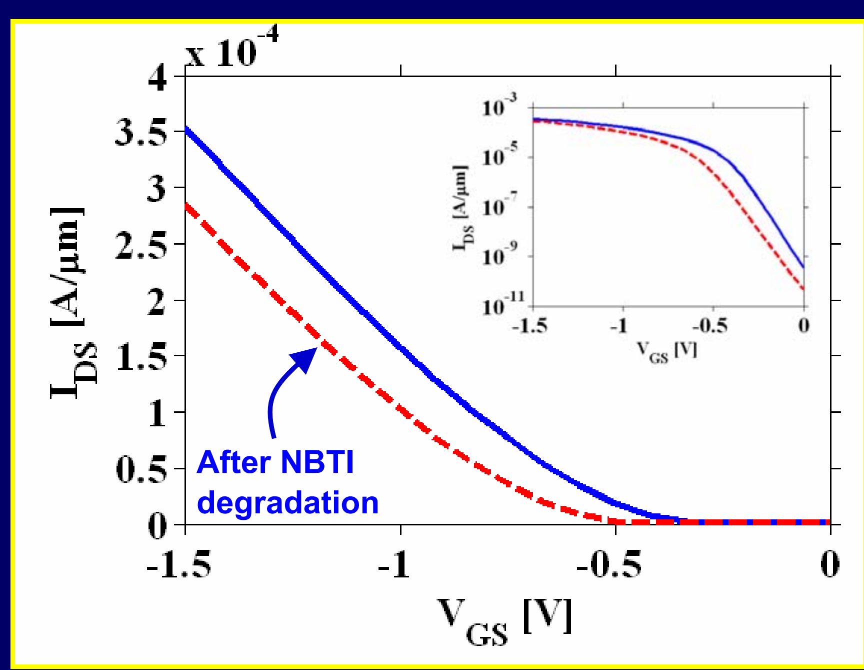
	Conv	Proposed
1.0 V		
0.9 V	FAILS	
0.8 V	FAILS	

Temporal Degradation: BTI

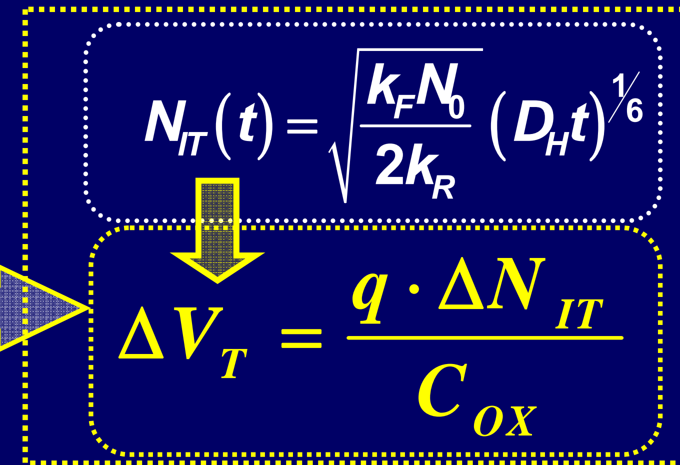
Kang, Roy, et. al. – TCAD, DAC-07

Bias Temperature Instability





- PMOS specific *Aging Effect*
- Generation of (+) traps
- Reaction-Diffusion (RD) model*
- Time exponent ~ 1/6

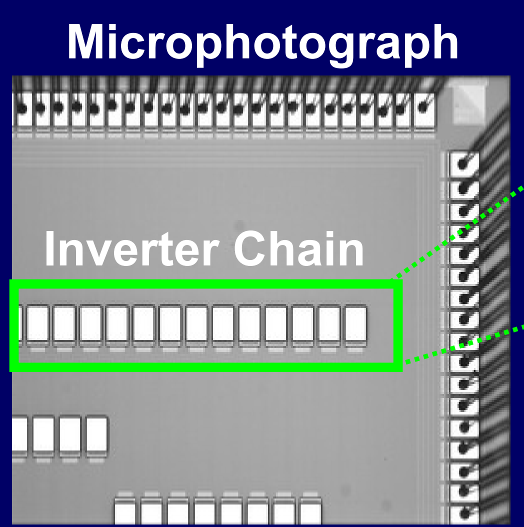


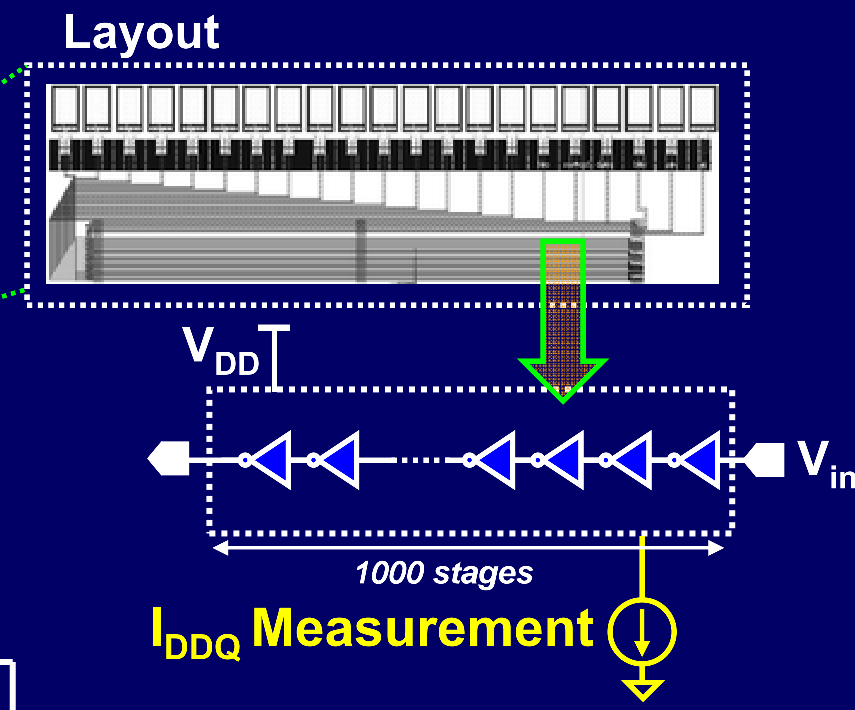
$$N_{IT}(t) = \sqrt{\frac{k_F N_0}{2k_R}} (D_H t)^{1/6}$$

$$\Delta V_T = \frac{q \cdot \Delta N_{IT}}{C_{OX}}$$

**M. A. Alam, IEDM'03*

I_{DDQ} based NBTI Characterization

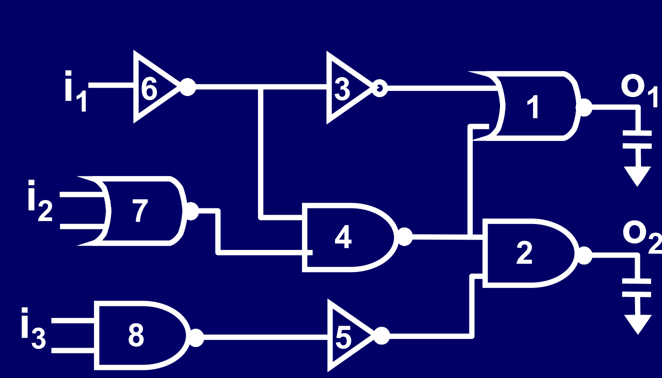




Technology	CMOS 130nm
Die Size	20 (mm²)
I/O Pin	209
T_{ox}	1.6 (nm)
V_{DD}	1.2 (V)

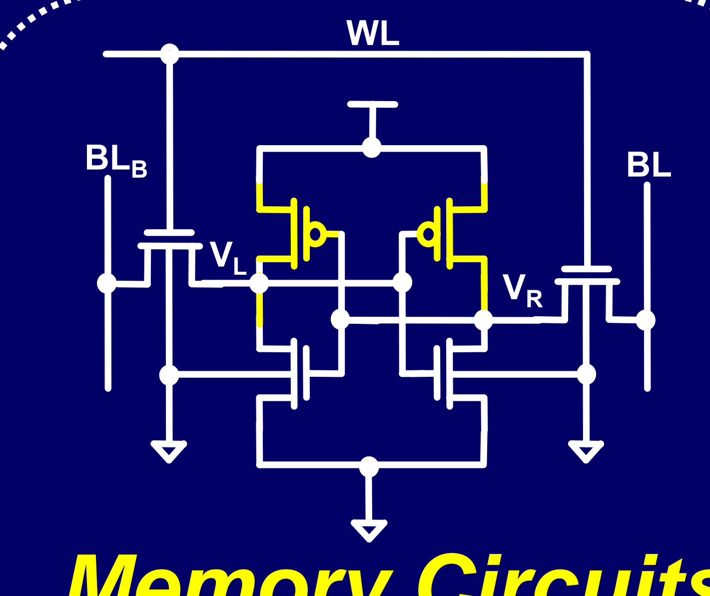
- Test Circuit Fabricated
- 1000 stage INV chain
- DC Stress signal @V_{in}
- I_{DDQ} measurement *@GND*

NBTI in Digital Circuits



Logic Circuits

- f_{MAX} decreases ↓
- Timing failure with time

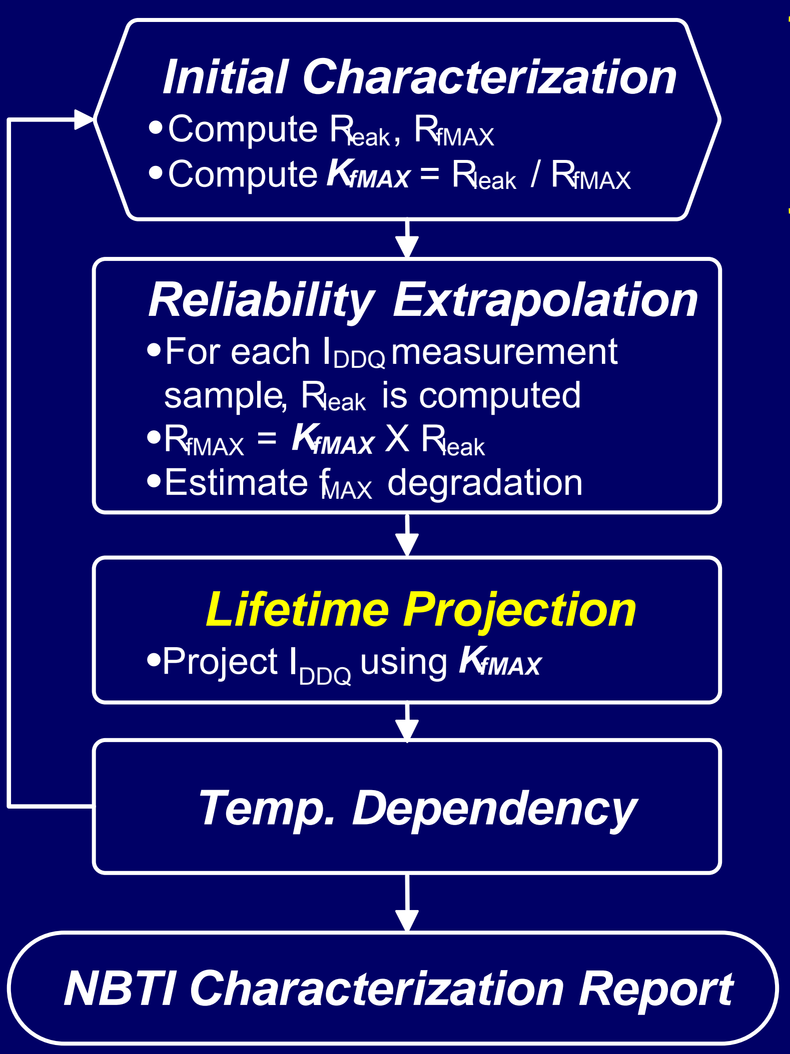


Memory Circuits

- Static Noise Margin (SNM) ↓
- Read & Write Stability
- Parametric Yield ↓

Temporal V_{Th} increase in PMOS affects critical performance factors of digital VLSI circuits

I_{DDQ} based Characterization Technique

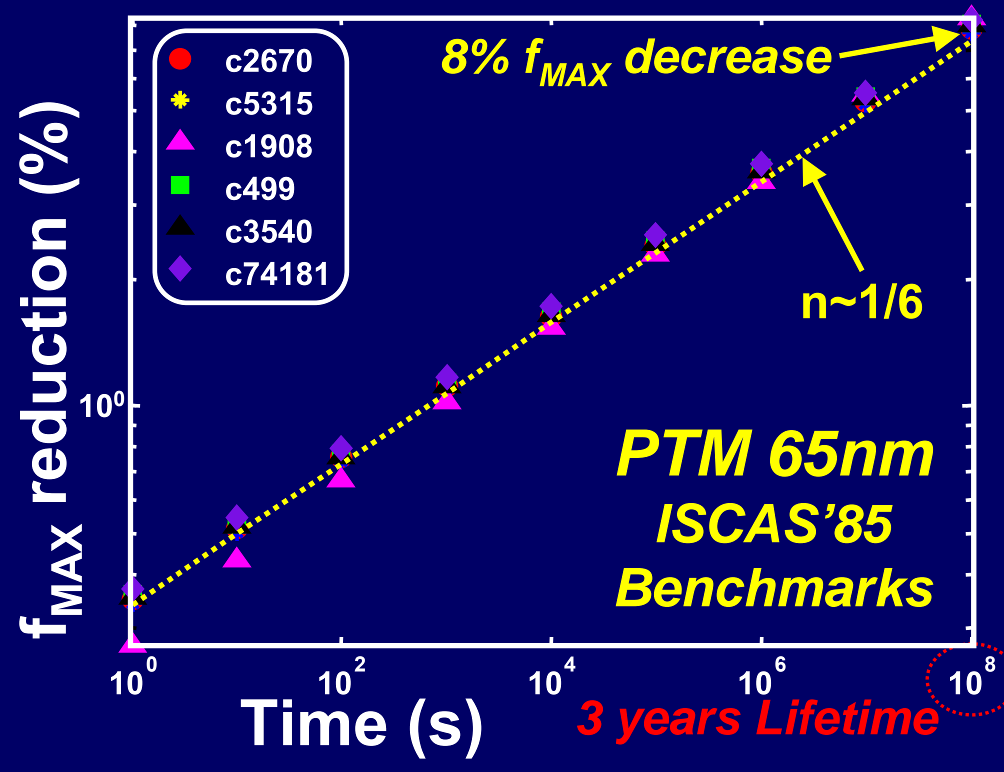


- Circuit-level NBTI Reliability Characterization
- I_{DDQ} test is used
- Expensive f_{MAX} testing is avoided (or minimized)
- Accurate circuit level performance degradation can be predicted
- IC specific burn-in to qualify the target produce
- Efficient way of field monitoring: dynamic local signature of produce usage
- Possible usage in other reliability sources; HCI

NBTI: Random Logic Circuits

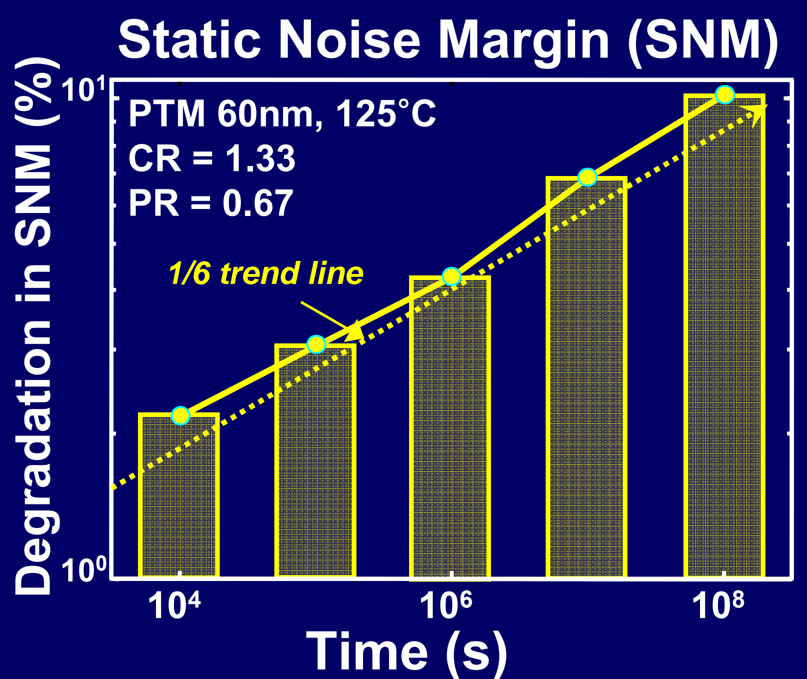
Delay Degrad. STD cells

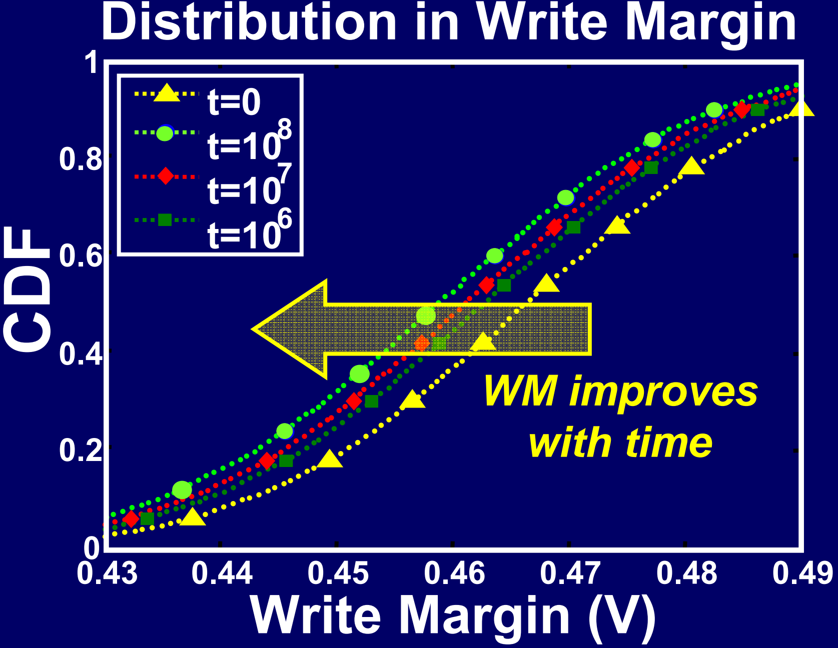
Logic Cell	fanin	Delay (ps)		Δ (%)
		t=0	3 years	
INV	1	13.77	16.77	21.8
NAND	2	16.86	19.88	17.9
NAND	3	19.57	22.45	14.8
NOR	2	17.26	21.89	26.8
NOR	3	23.80	30.19	26.9



- ISCAS'85 Benchmark Circuits, PTM 65nm
- Gate delay: analytical delay model considering NBTI
- Circuit delay: NBTI-aware Static Timing Analysis (STA)
- Circuit f_{MAX} → time exponent n ~ 1/6

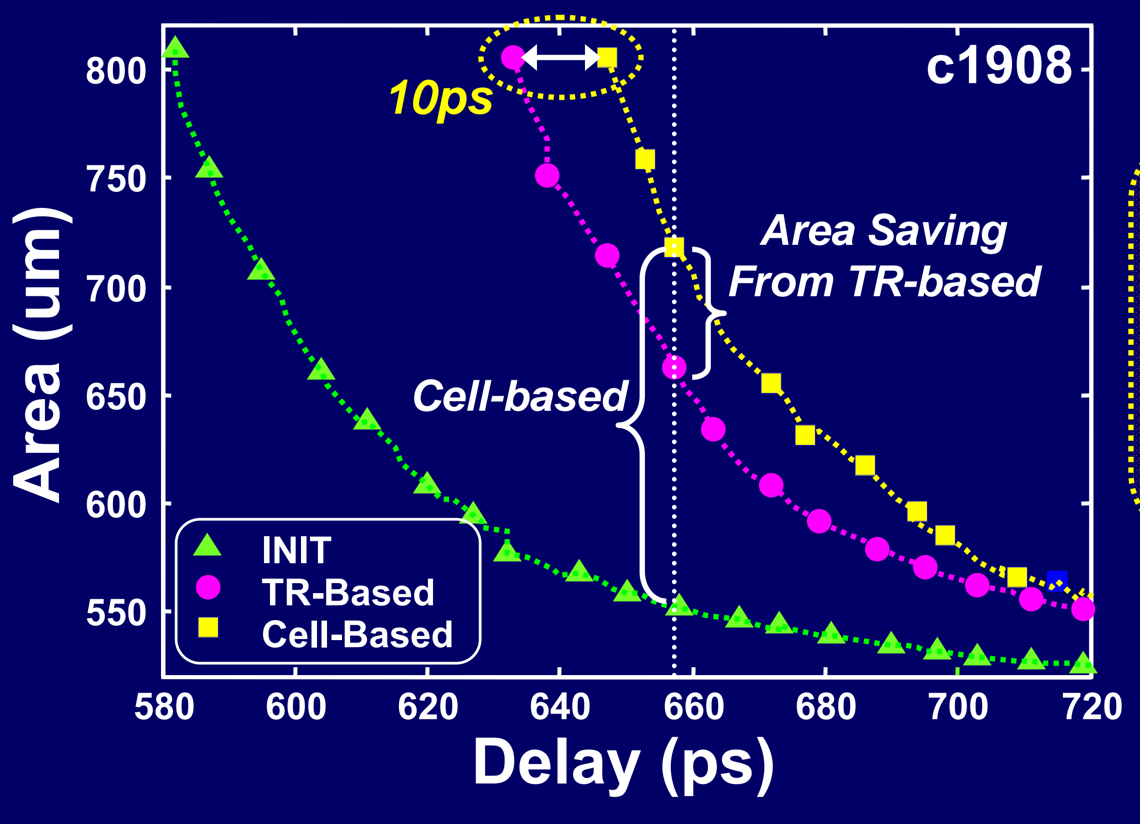
NBTI: 6T SRAM Cell





- SNM degrades by more than 10% in 3 years
- **% SNM Degradation → time exponent n ~ 1/6**
- **WM improves** with time under NBTI

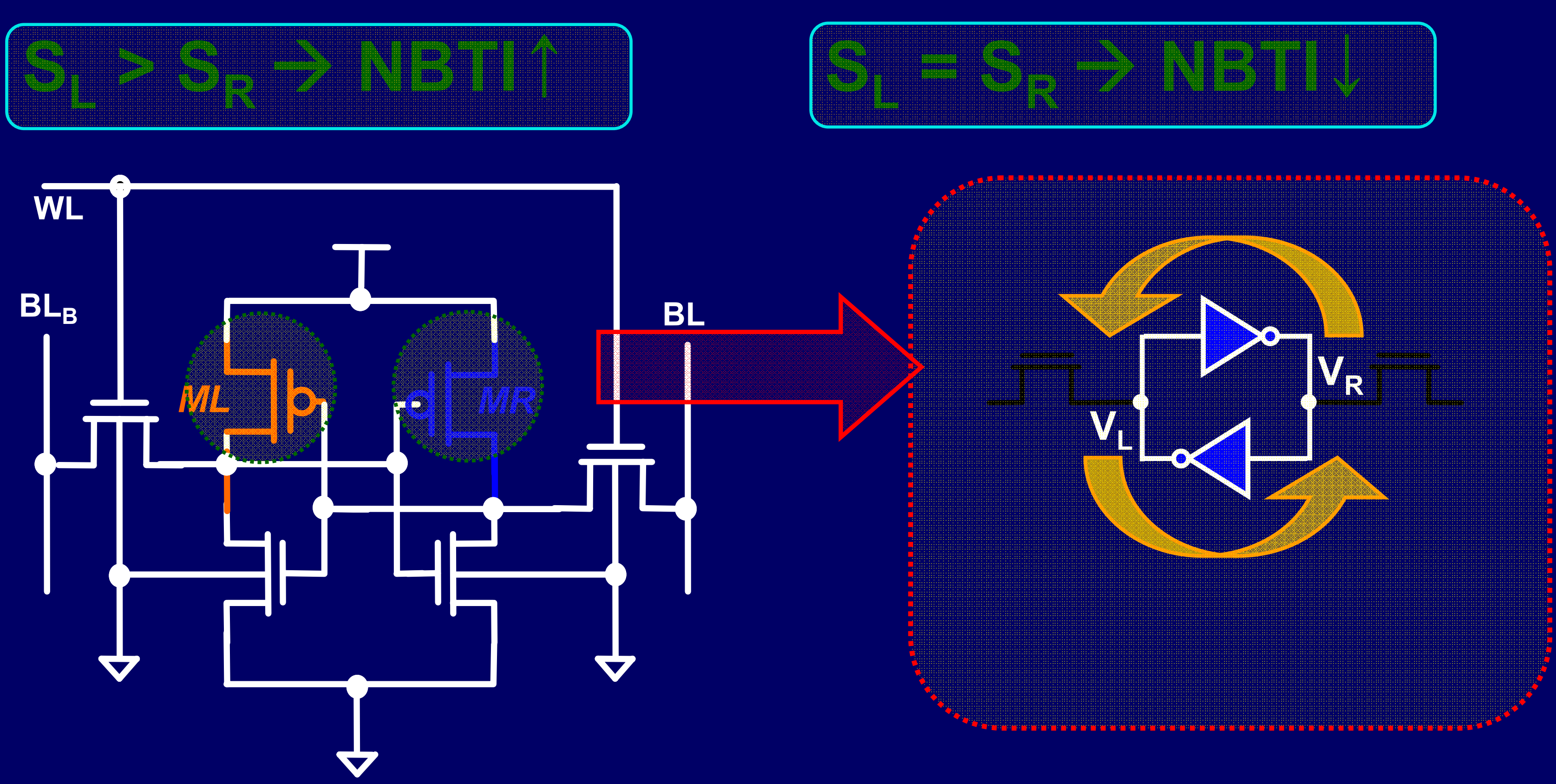
Design for Reliability under NBTI



- **Simulation Setup**
  - Synthesized in PTM 65nm
  - 1/6 V_{Th} degradation model
  - 125°C Stress temperature
  - 50% Signal Probability at PI's

- Gate Sizing applied to guarantee lifetime functionality of design
- **11.7% overhead for Cell-based sizing**
- **6.13% overhead for TR-based sizing**
  - 45% improvement in area overhead
  - Runtime complexity for TR-based sizing is identical to that of Cell-based sizing

Memory Solution #1: Periodic Cell-Flipping*



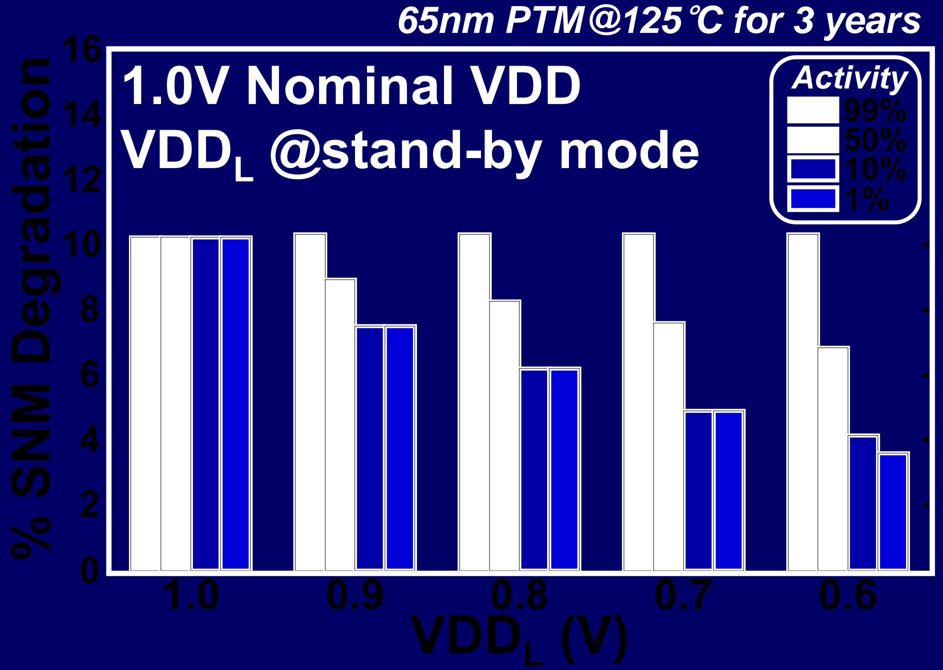
- **Balance the signal probability on both side of the cell → $S_L = S_R$**
- **Still incurs close to 10% degradation in SNM**

*S. Kumar et al., ISQED'06

Solution #2: Standby V_{DD} Scaling

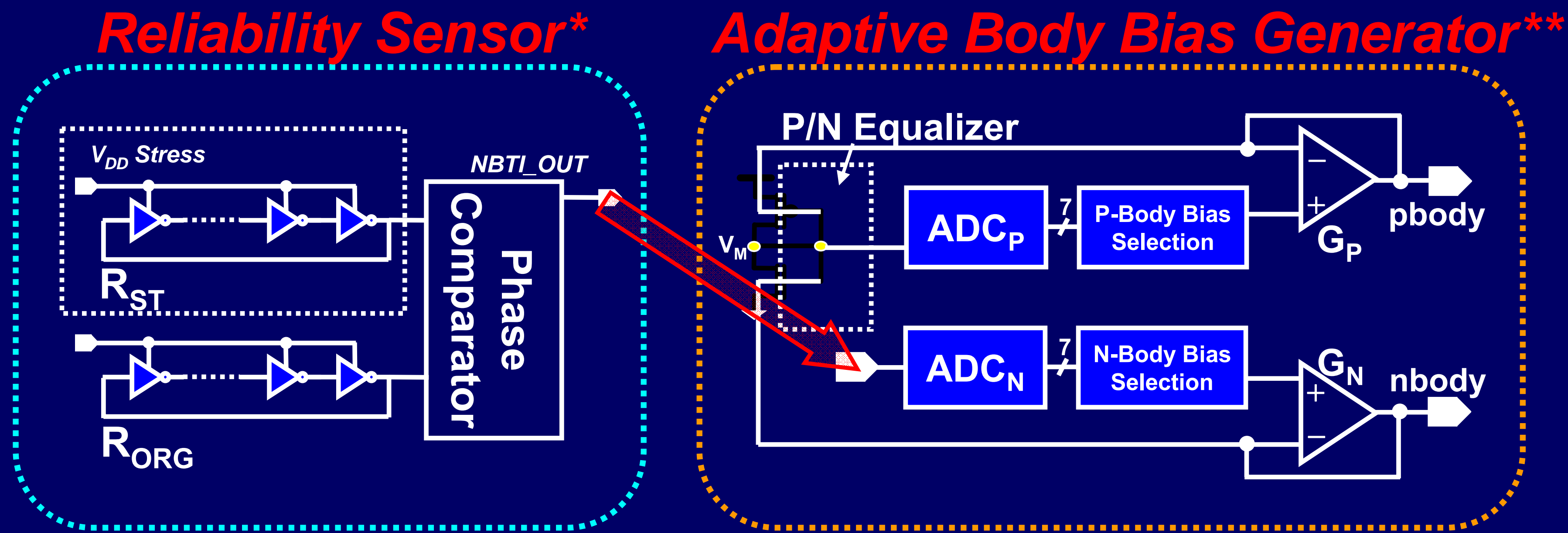
- Normal Mode → VDD
- Standby Mode → VDD_L

- Normal Mode → NBTI
- Standby Mode → Min. NBTI



- **NBTI is a strong function of V_{DD}**
- **Lower V_{DD} during standby mode → MIN V_{DD}**
- ***Effective solution with low design effort***

Solution #3: Sensing & Correction



- **Sense reliability degradation → adaptive correction using circuit techniques**
- ***T. Kim et al., VLSI Circuit Symposium 2007**
- **** K. Kang et al., Design Automation Conf. 2007**
- ***Need to properly consider design overhead***

Conclusions

- Process Variation and Process Tolerance is becoming important
- There is a need to optimize designs considering power/performance/yield