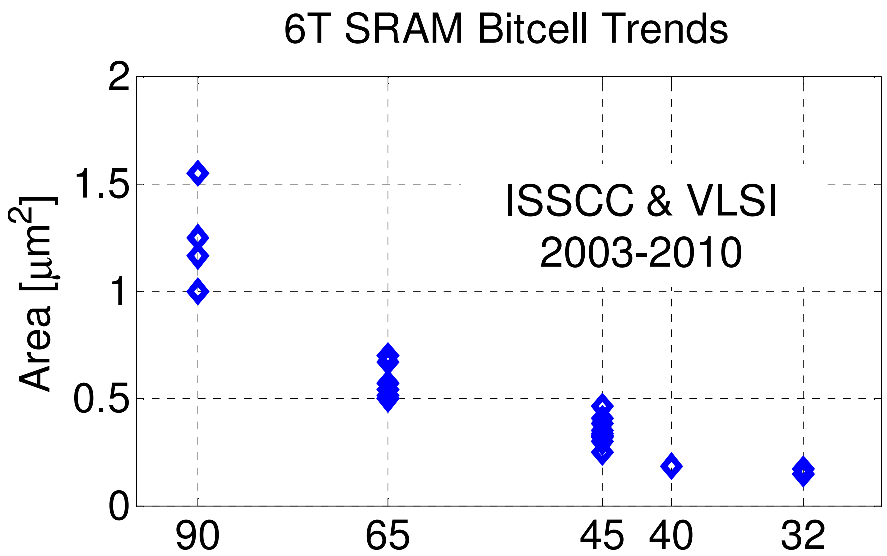
Loop Flattening & Spherical Sampling: Highly Efficient Model Reduction Techniques for SRAM Yield Analysis

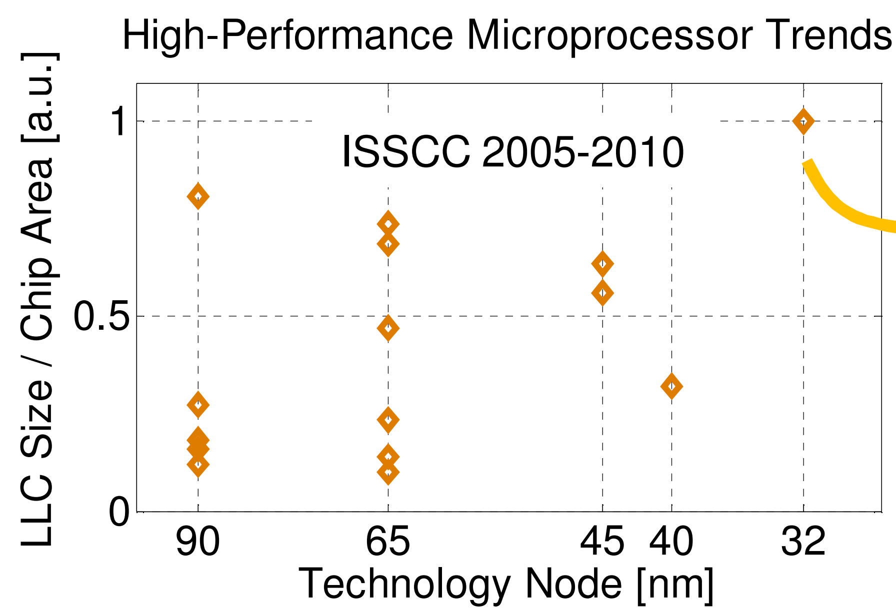
Masood Qazi[1], Mehul Tikekar[1], Lara Dolecek[1,2],
Devavrat Shah[1], Anantha Chandrakasan[1]

[1]Massachusetts Institute of Technology
[2]University of California Los Angeles

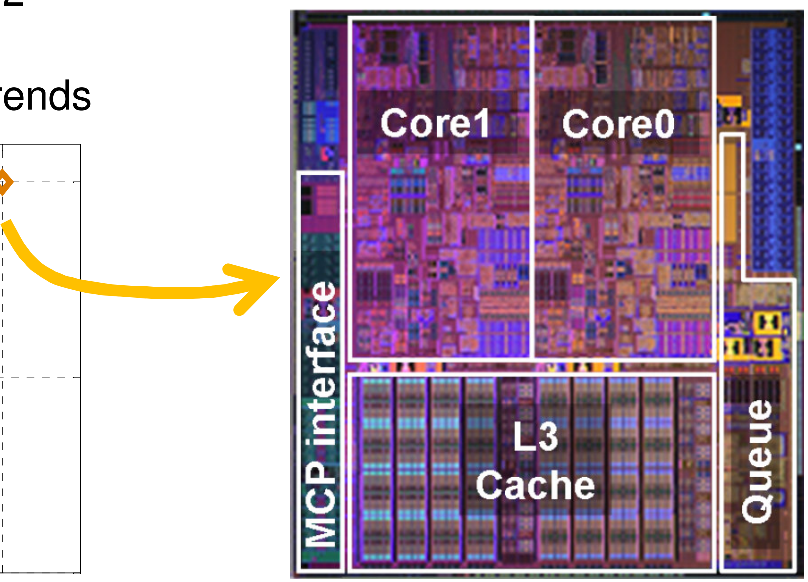
DATE 2010

SRAM Yield Challenge





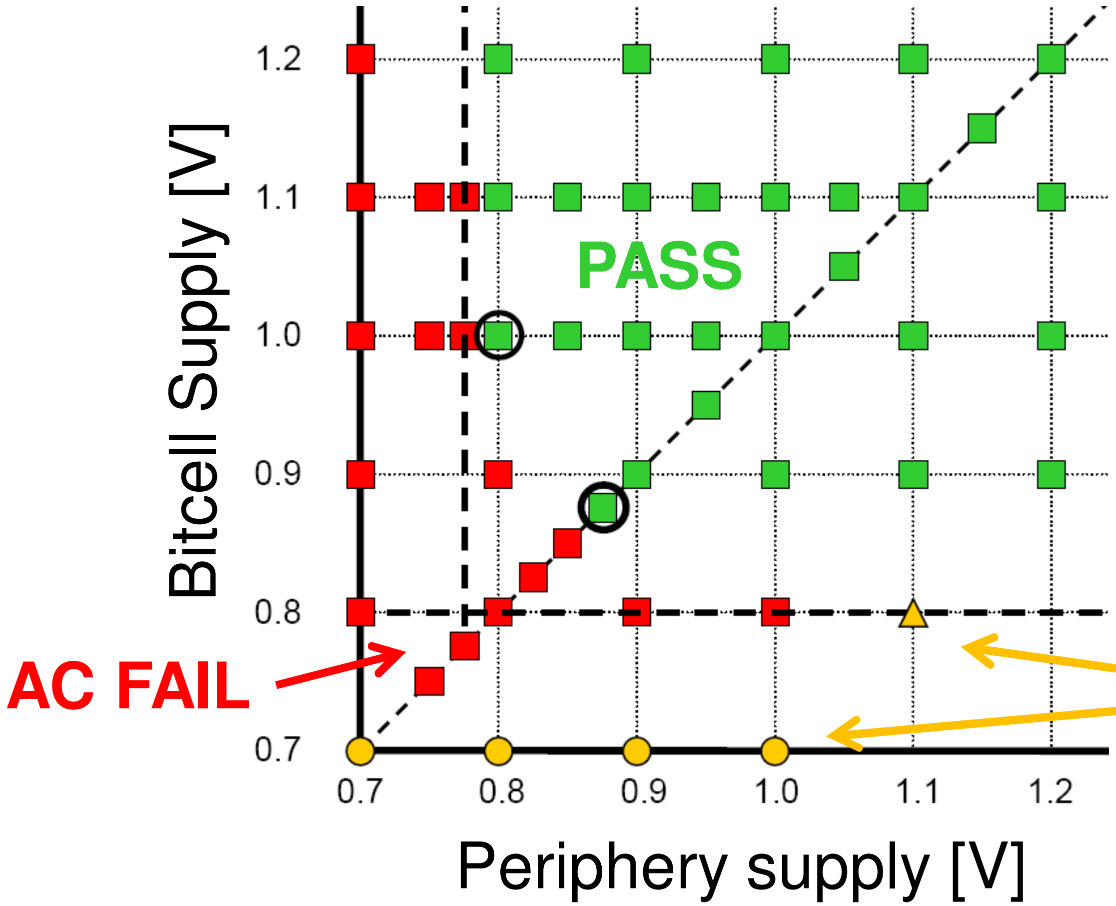
- Increased SRAM utilization must:
  - Address operating constraints (e.g. Vmin, performance) on the system level
  - Preserve chip yield



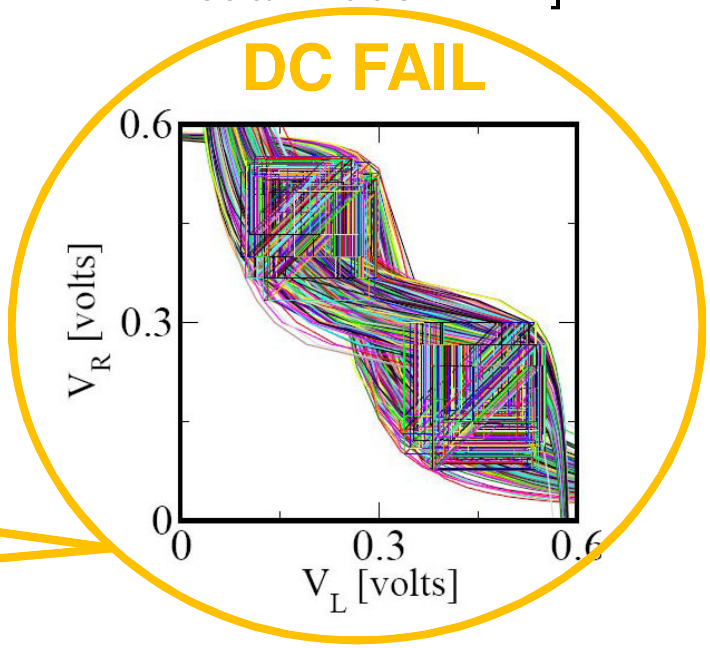
[Kurd et al. 2010 ISSCC]

Goal of this work: SRAM AC Yield

6T SRAM Array Operating Window in 65nm
[figure from Pille et al. 2007 ISSCC]



[figure from Bhavnagarwala et al. 2005 IEDM]



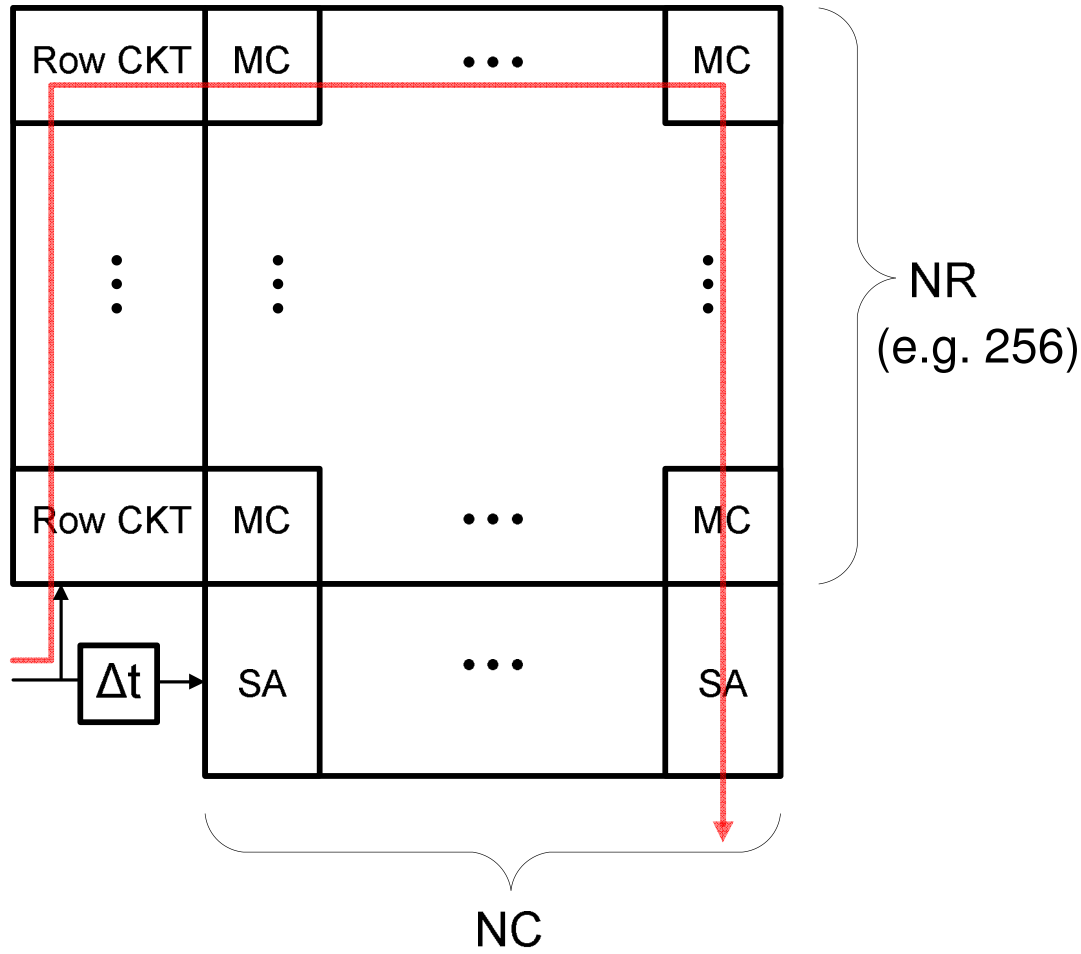
Read access time imposes stricter limit on yield

Outline

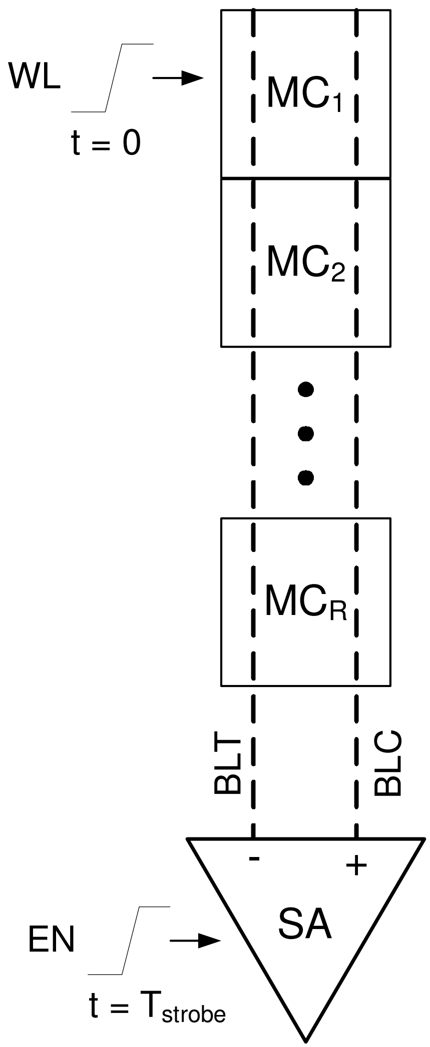
- Loop Flattening for the SRAM Critical Path
- Spherical Importance Sampling for High Dimensionality
- Experimental Results

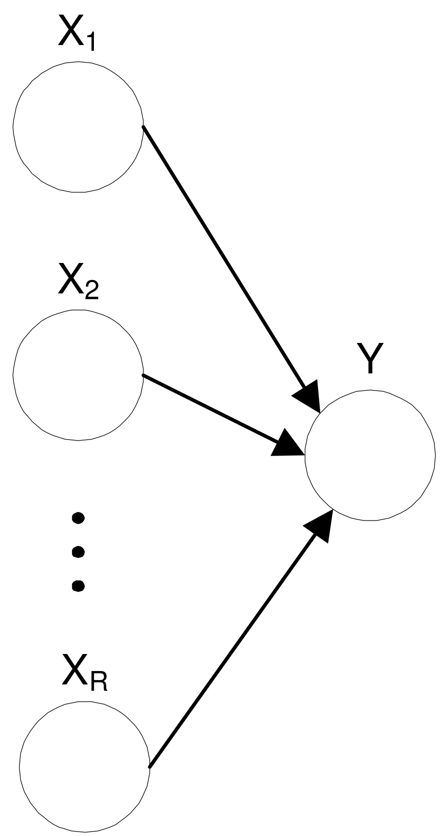
Challenges related to evaluating SRAM read access yield

- Repeated structures
- Cascade of correlated stages
- Simulating full circuit is prohibitive



The statistical analysis of memory critical paths requires insight into architecture





(e.g. R = 256)

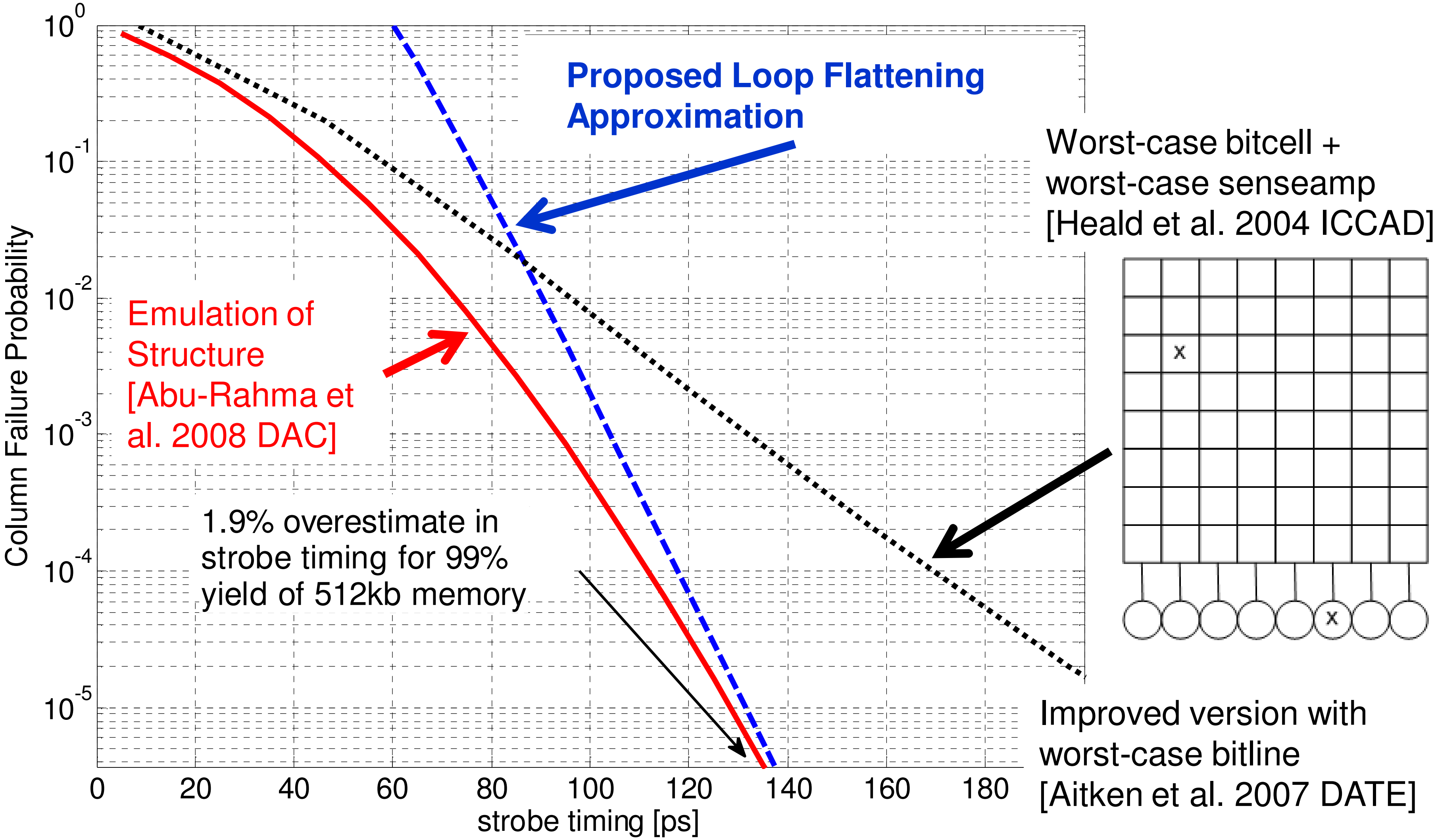
Conflicting intuitions:

1. Sense amplifier variation less severe because of less repetition
2. Sense amplifier variation just as critical because same number of paths

- Latter interpretation is more appropriate, enabling “loop flattening” estimate

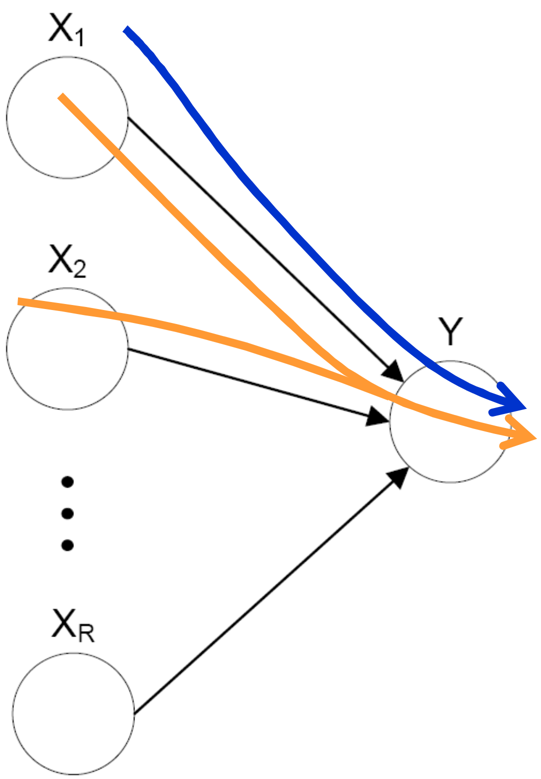
$$\hat{F} = R \cdot f, \quad \lim_{t \to \infty} \frac{F - \hat{F}}{F} = 0$$

Loop flattening approximation is very accurate at relevant levels of yield



Simulation in MATLAB with example parameters of 1mV/ps bitline discharge, 10% relative σ of read current, and 20mV SA offset.

Loop Flattening: Simplified Proof



Suppose X_i and Y are indep. additive zero-mean standard normal delays. Each path has a delay Z_i:

$$Z_i = X_i + Y$$

Delay of tree is slowest path: $T = \max_{1 \le i \le R} Z_i$

Associated failure is:

$$F = \mathrm{P}\left(\max_{1 \le i \le R} Z_i \ge t\right) = \mathrm{P}\left(\bigcup_{i=1}^{R} (Z_i \ge t)\right)$$

Loop Flattening is conservative union bound:

$$F \le \hat{F} = \sum_{i=1}^{R} \mathrm{P}(Z_i \ge t) = R \cdot \mathrm{P}(Z_1 \ge t)$$

Pair-wise intersection produces lower bound:

$$\begin{aligned} F \ge F^* &= \sum_{i=1}^{R} \mathrm{P}(Z_i \ge t) - \sum_{\substack{i=1 \\ i<j}}^{R} \mathrm{P}\left((Z_i \ge t) \bigcap (Z_j \ge t)\right) \\ &= \hat{F} - \frac{R(R-1)}{2} \mathrm{P}\left((Z_1 \ge t) \bigcap (Z_2 \ge t)\right) \end{aligned}$$

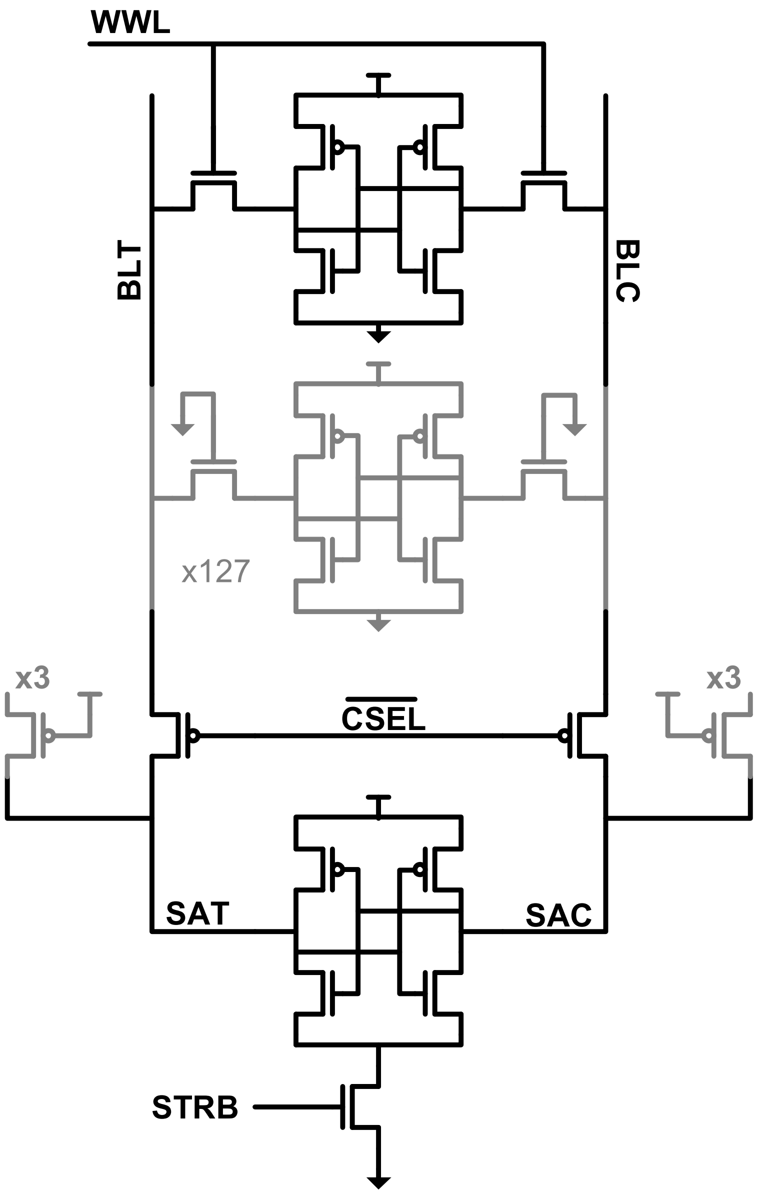
With these upper/lower bounds, it can be shown:

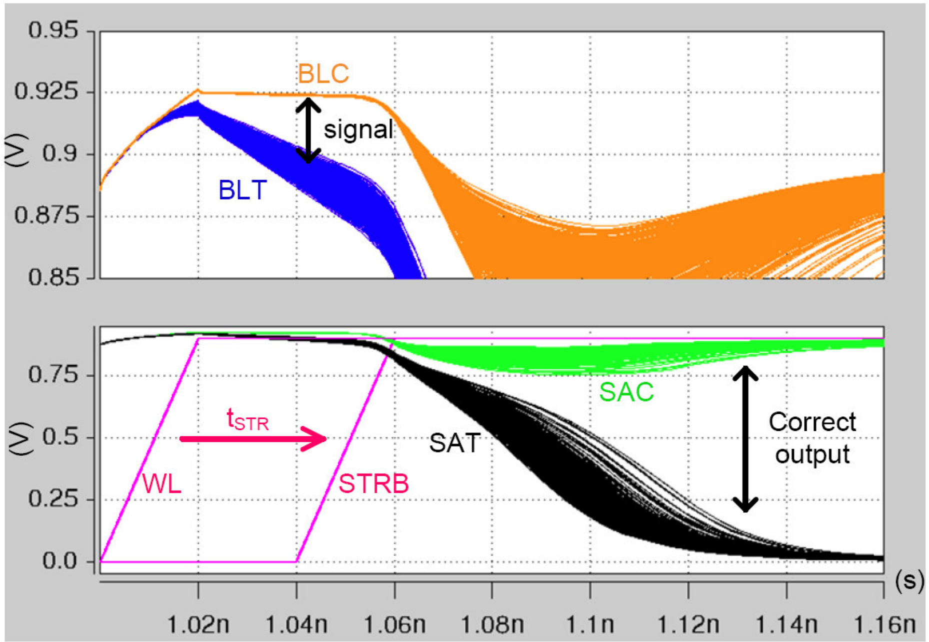
$$\lim_{t \to \infty} \frac{\hat{F} - F^*}{F^*} = 0 \Rightarrow \lim_{t \to \infty} \frac{\hat{F} - F}{F} = 0$$

Outline

- Loop Flattening for the SRAM Critical Path
- Spherical Importance Sampling for High Dimensionality
- Experimental Results

SPICE simulation of the small-signal read path is hard to simplify





Variable bitline signal interacts with **SA mismatch** to produce logic-level output after an **uncertain delay**

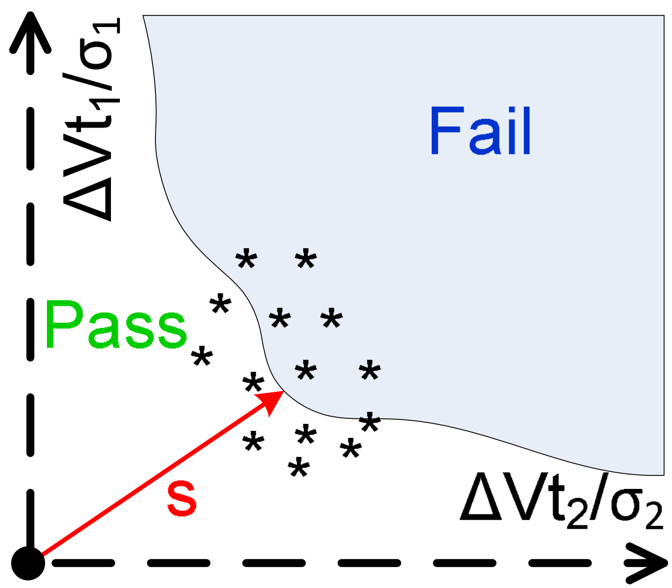
Techniques to Accelerate Monte Carlo Simulation

Techniques

- Extreme Value Theory
  - [Singhee et al. 2007 DATE], 11 dimensions after filtering
  - [Singhee et al. 2008 VLSID], 6 dimensions
- Improvements on Importance Sampling (all in 6 dimensions)
  - MixIS [Kanj et al. 2006 DAC]
  - Adaptive IS [Jaffari and Anis 2009 ICCAD]
  - Norm Minimization [Dolecek et al. 2008 ICCAD]
- Worst-case point
  - [Du and Chen 2000 ASME], [Antreich and Graeb 1991 ICCAD], [Bucklew 2004]

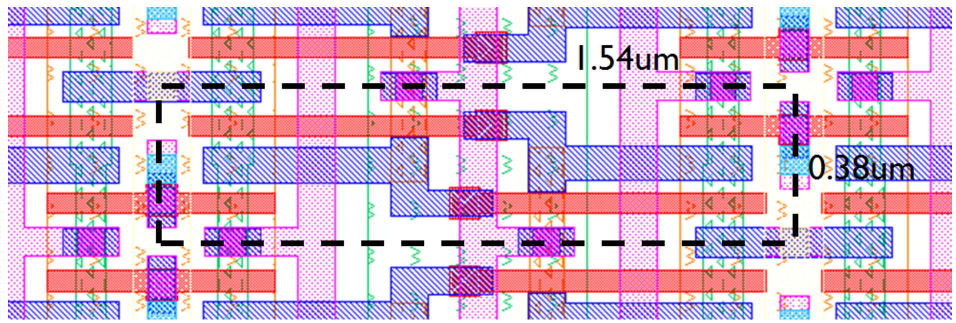
$$p_{MC} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad p_{IS} = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_i \cdot w_i$$

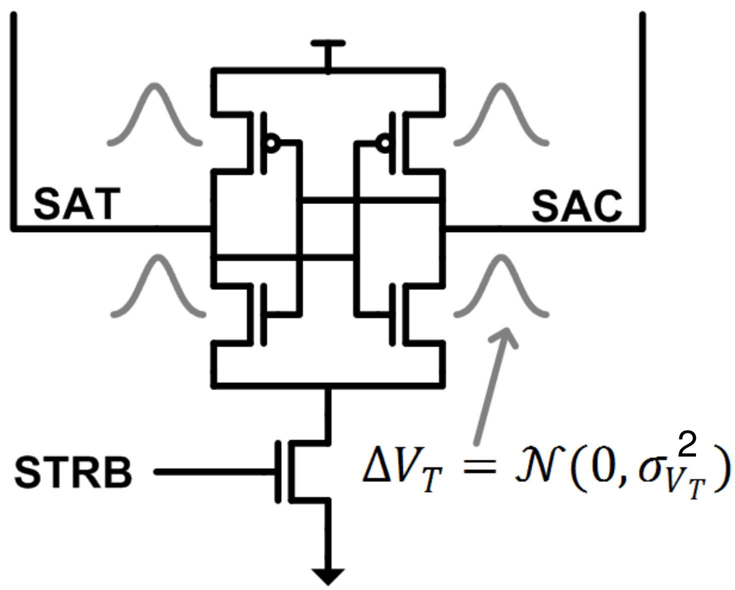
$$w_i = \exp\left(-\sum_{j=1}^{M} \frac{s_j(2y_{i,j} - 2\mu_j - s_j)}{2\sigma_j^2} \right)$$



Local Variation Simulation Setup

- SRAM sense-amplifier based read path with 12 normal random r.v.'s
- 45nm Free PDK [Stine et al. 2007 ICMSE]
  - Bitline extracted capacitance of 0.24fF/μm
  - 45nm HKMG PTM models [Zhao et al. 2006 TED]
- Device sizing:
  - Bitcell sized for better than 1×10^{-9} DC stability failure
  - Sense network area sized to occupy 8x bitcell area





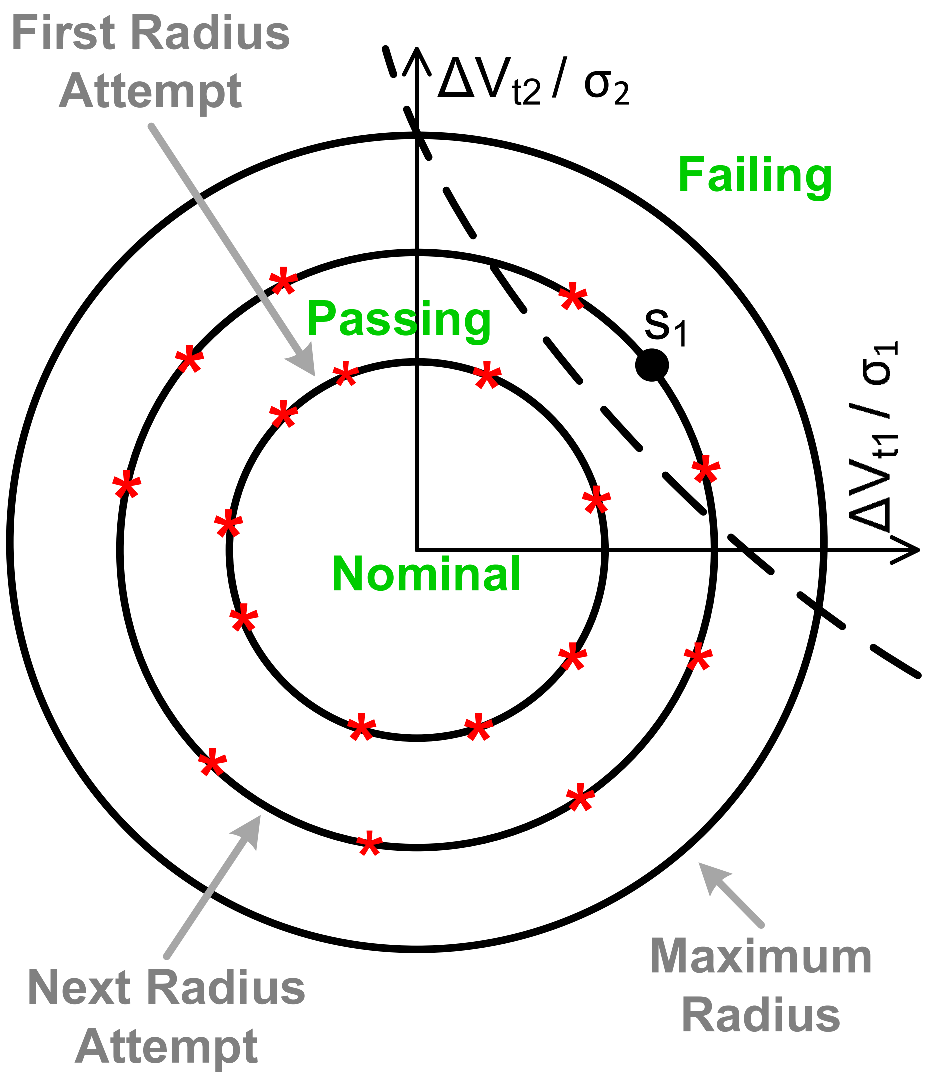
$$\sigma_{V_t} = \frac{1.8mV}{\sqrt{W_{\text{eff}} L_{\text{eff}}}}$$

[Kuhn 2007 IEDM]

Spherical IS, Step 1: Sampling on Shells

Global directional search:

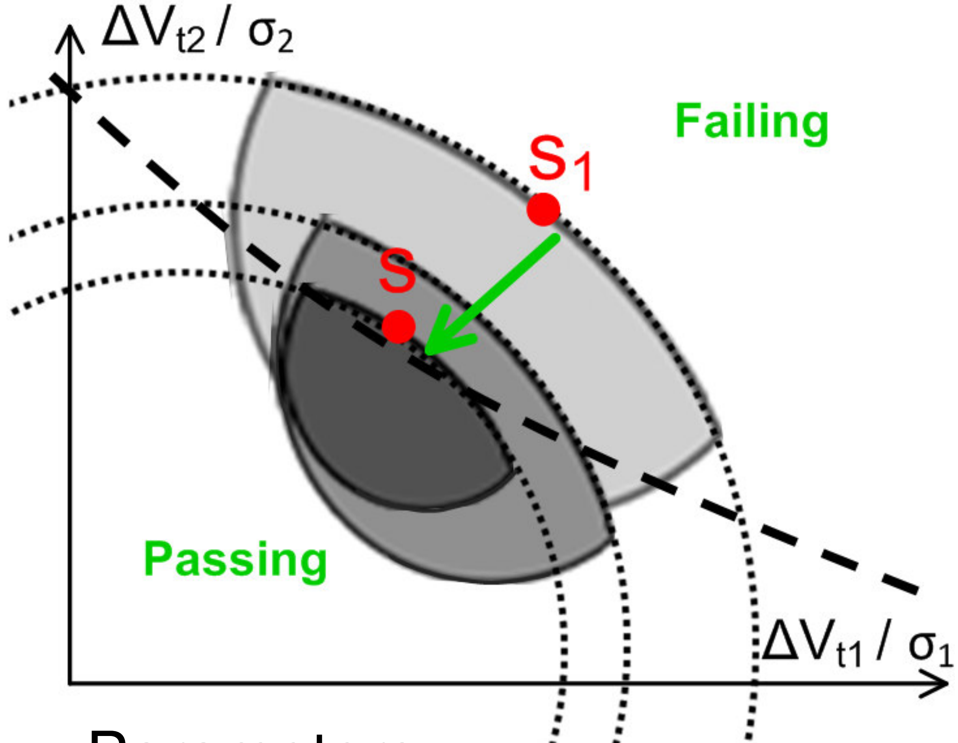
- Parameters:
  - N_1: number of samples per shell (500)
  - R_{high}: maximum distance related to failure floor (1×10^{-12})
- Perform bisection on radius until failures observed on shell (1 to $N_1/2$)
- Average over failing direction vectors to produce $\mathbf{s_1}$, an initial direction.



Spherical IS, Step 2: Local Exploration

Fine tune mean shift:

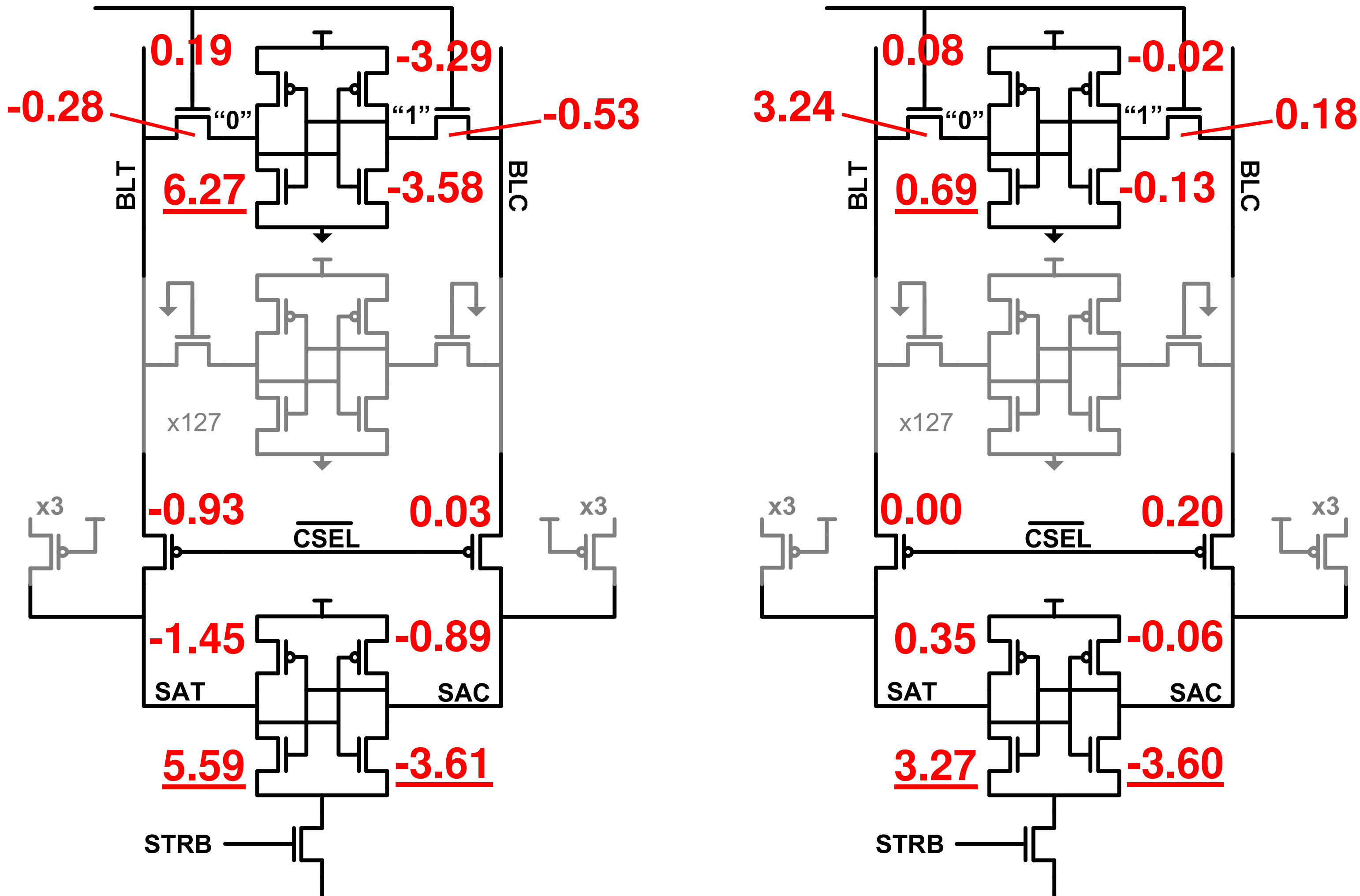
- Initialize shift estimate, **s**, with value from Step 1 (**s_1**)
- **A:** Sample within the volume of a local sphere centered at **s** that also has norm less than ||**s**||.
- **B:** Check if the point results in failure. If so, update value of **s** and shrink local radius (R_2) geometrically towards 0.05
- If less than N_2 runs have occurred, go back to **A** otherwise return **s** for standard I.S. simulation



Parameters:

- N_2: number of simulations runs (500)
- Initial local radius R_2 ($||s_1||/2$)
- Final local radius (0.05)

Mean Shift Evolution for $p_{path}=2\times10^{-9}$



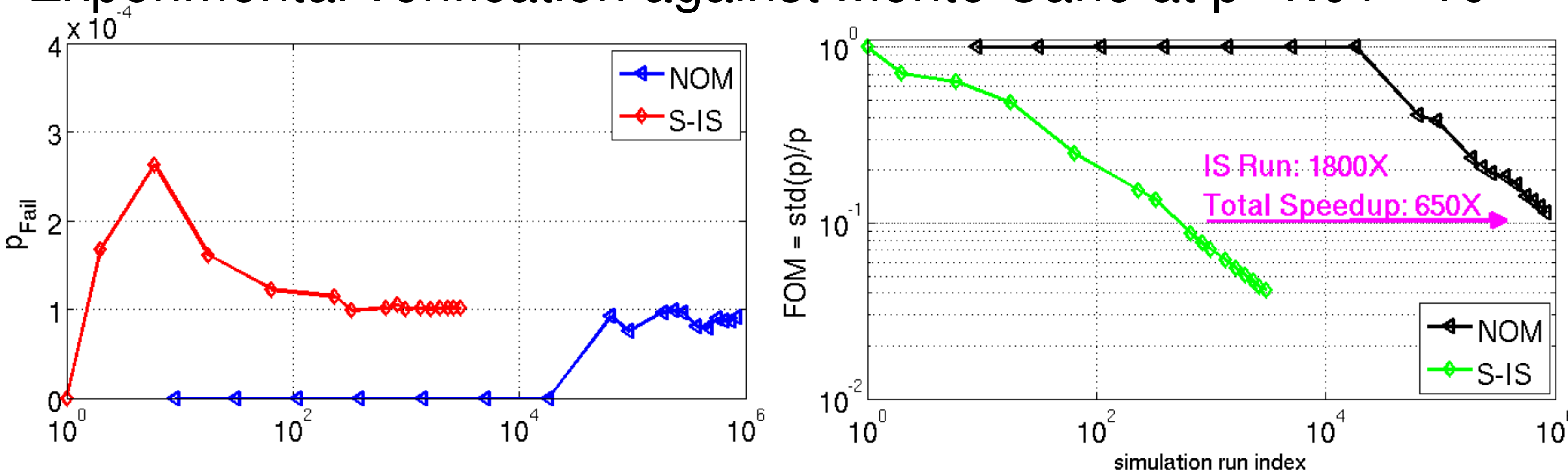
	Step 1	Step 2
# Runs:	1000	500
L_2 Norm:	10.55	5.90

Strobe Timing Simulation Results

The probability of column failure is determined for various timings.

t_{STR}	p_{path}	cost	speed-up	p_{col}
40ps	1.01×10^{-4}	1534	6.50×10^{2}	5.2×10^{-2}
50ps	9.08×10^{-6}	1660	6.63×10^{3}	4.6×10^{-3}
65ps	1.33×10^{-7}	2214	3.40×10^{5}	6.8×10^{-5}
80ps	1.91×10^{-9}	2423	2.16×10^{7}	9.8×10^{-7}

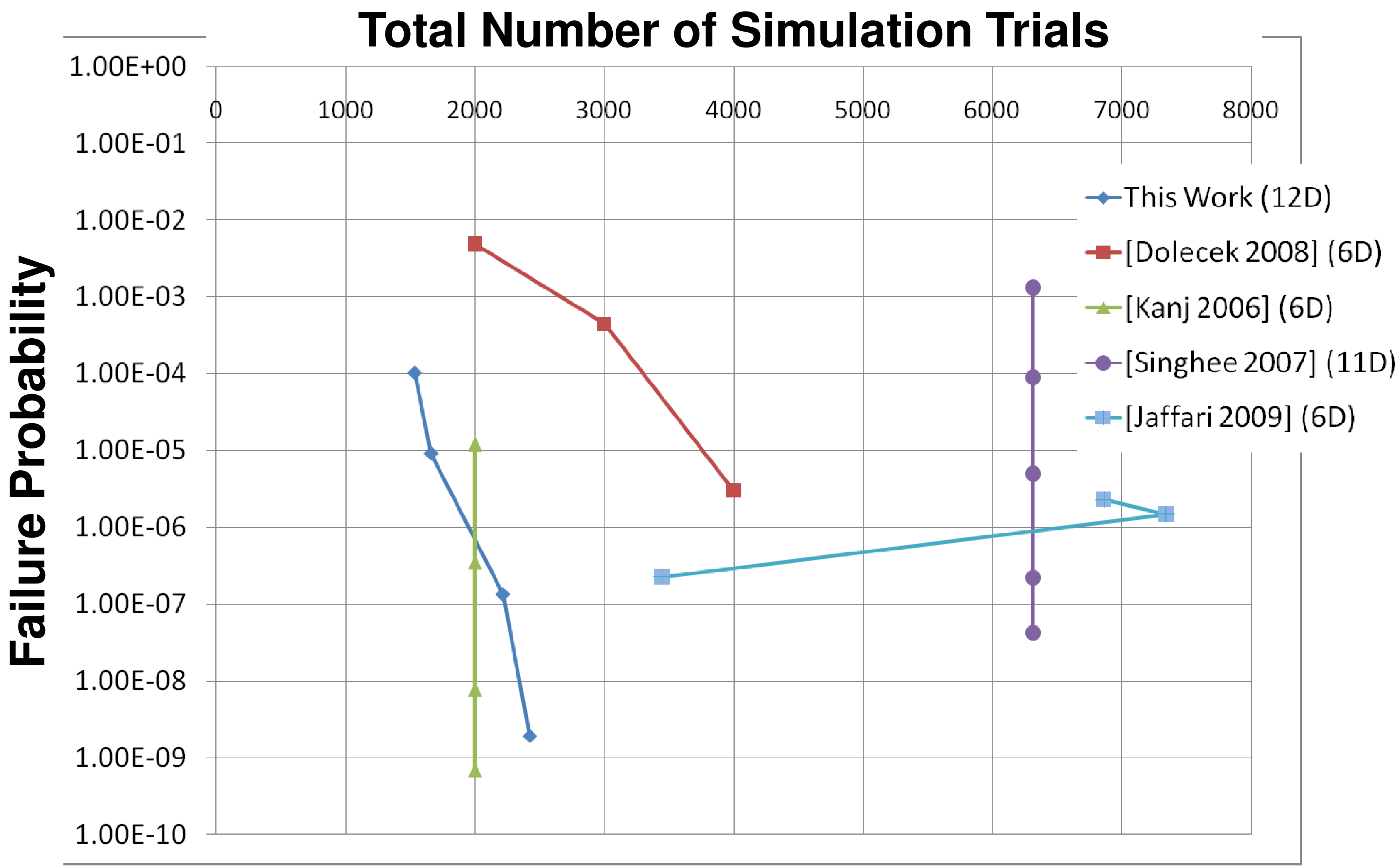
Experimental verification against Monte Carlo at $p=1.01\times10^{-4}$



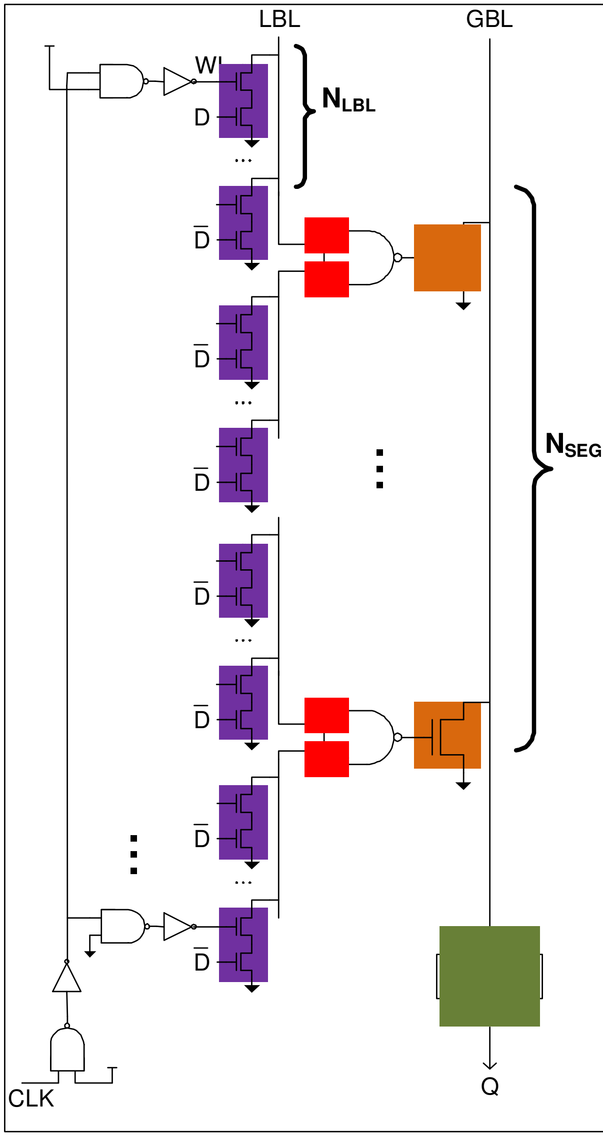
Simulation cost breakdown

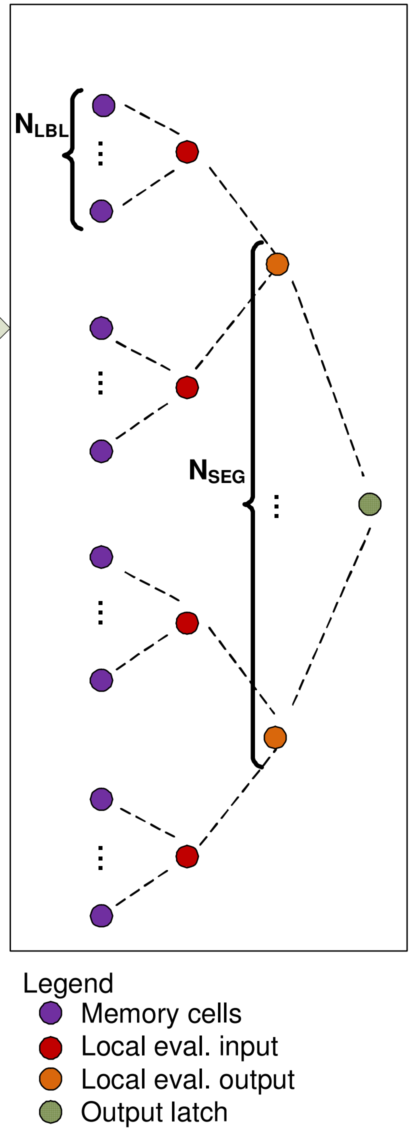
	This Work 12 Dimensions 2-step Spherical Samples			[Dolecek 2008 ICCAD] 6 Dimensions Uniform Exploration		
P	9.08×10^{-6}	1.33×10^{-7}	1.91×10^{-9}	4.9×10^{-3}	4.4×10^{-4}	3.0×10^{-6}
Step 1	500	1000	1000	-	-	-
Step 2	500	500	500	-	-	-
Total Exploration	1000	1500	1500	1000	1000	2000
IS run	660	714	923	1000	2000	2000
Total	1660	2214	2423	2000	3000	4000

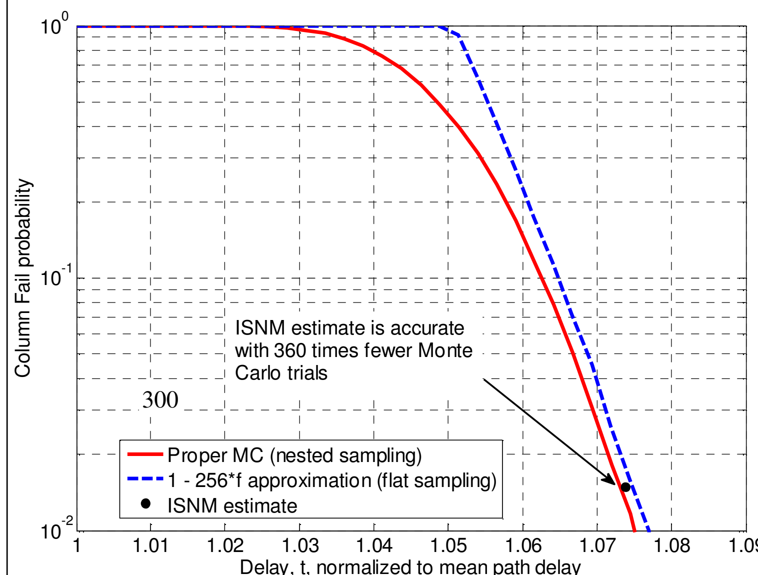
Cost versus Failure Probability



Future direction: loop flattening on the large-signal hierarchical read path







Observations:

- Additional depth in tree
- Correlation across stages
- Faster convergence of loop flattening

Conclusions

- **The apparent structure of integrated circuits differs under the statistical analysis of rare occurrences.**
- **The efficient application of Importance Sampling to speed-up Monte Carlo simulation requires exploration of the parameter space with:**
  - **Non-gaussian statistical sampling**
  - **Sequential focus on direction, then magnitude**

Acknowledgement: FCRP Focus Center for Circuit & System Solutions (C2S2) for funding.