Delay Uncertainty Reduction by Interconnect and Gate Splitting

Vineet Agarwal Jin Sun
Alexander Mitev Janet Wang
ECE Department
University of Arizona
{vagarwal,sunj,mitev,wml}
@ece.arizona.edu

Performance Variations

- The critical dimensions in today's digital circuits are decreasing continuously.
- The impact of manufacturing process e.g. Chemical Mechanical Polishing (CMP) leads to variations in circuit parameters.
  - Result in 'wide-spread' appearance of timing profile
  - Cause yield deterioration due to time violations
- Various techniques have been proposed to deal with these CMP related performance perturbations.

Sizing Technique

- Timing variation reduction using sizing technique
  - Lancelot [Jacobs00][Raj04]
  - Lagrangian Relaxation[Choi04]
  - Geometric programming[Singh05]
- Disadvantages:
  - Incur an increase in circuit area
  - Increases the mean delay value at the outputs
- A gate splitting mechanism with trade-off between variation reduction and area increase[Neiroukh05][Wu05]

Our Main Contributions

- Include interconnect splitting to reduce variability
- Provide proofs to reveal the relationship between delay variation reduction and splitting
- Implement proposed methodology in TURGIS
  - Timing Uncertainty Reduction by Gate-Interconnect Splitting
  - Integrated with existing SSTA tool
  - Apply placement algorithms

Splitting Technique

- Definition:
  - Splitting is defined as the technique of substituting a parent (larger) entity by two children (smaller) entity of half the parent size and connecting them in parallel in place of the parent entity.
- Note:
  - Not exactly the same as ‘hardware redundancy’
  - Split original component into two equal size components
  - Keep the same total component size

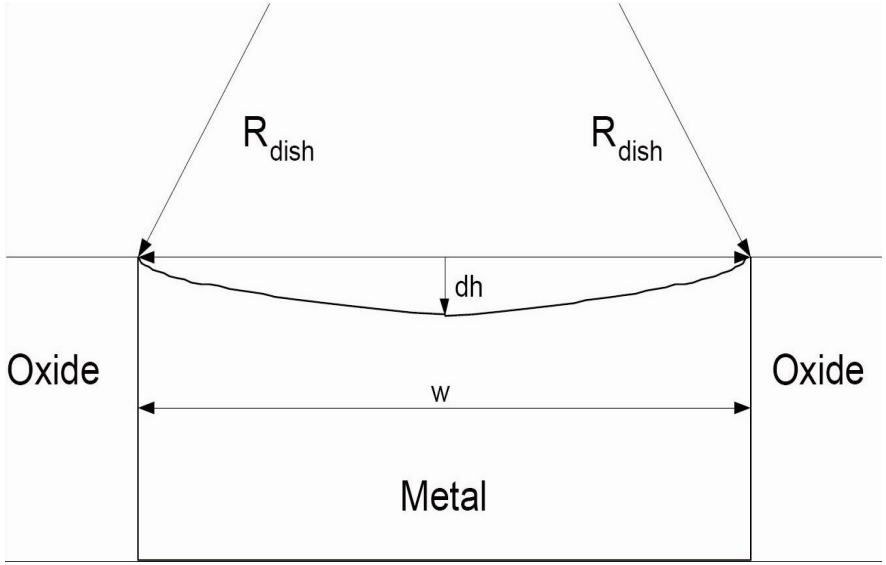
Dishing Effect

- Metal Dishing Model

  [Chang04]

  - dh can be written as

$$dh(w, R_{dish}) = R_{dish} - \sqrt{R_{dish}^2 - \frac{w^2}{4}}$$

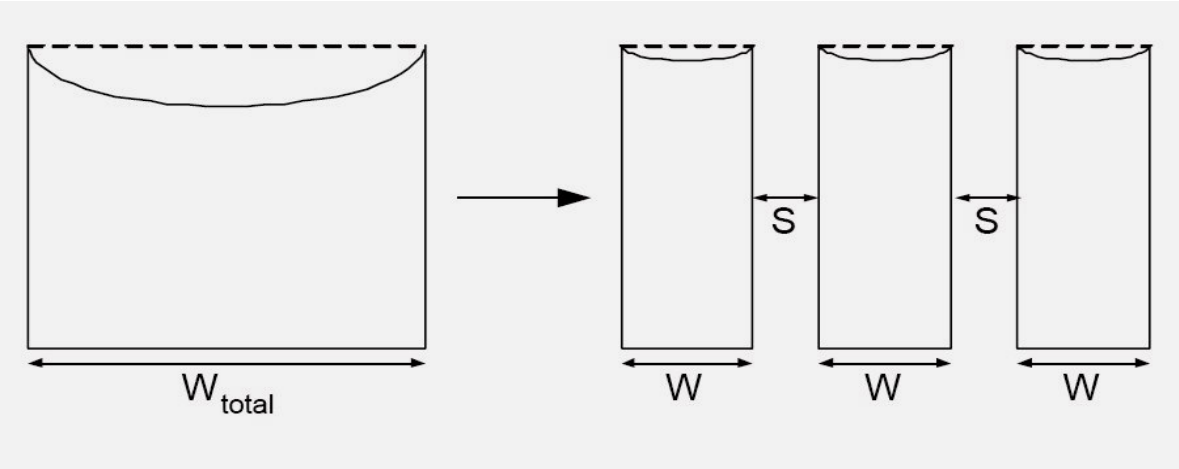


  - R_{dish} is assumed to be 4x-6x of metal width (w).
  - Interconnect resistance per unit length

$$r_d = \frac{\rho}{wh + 0.5w\sqrt{R_{dish}^2 - \frac{w^2}{4}} - R_{dish}^2 \sin^{-1}(\frac{w}{2R_{dish}})}$$

Interconnect Splitting

- Interconnect splitting to subside dishing effect



  - A wide metal line is replaced by N lines in parallel
  - Dishing effect is less prominent
  - The parasitic capacitance is nearly the same

Interconnect Splitting

- The only difference is in line resistance.
  - The delay of original wire

$$d_{org} = r_d(w, R_{dish})C_{total}$$

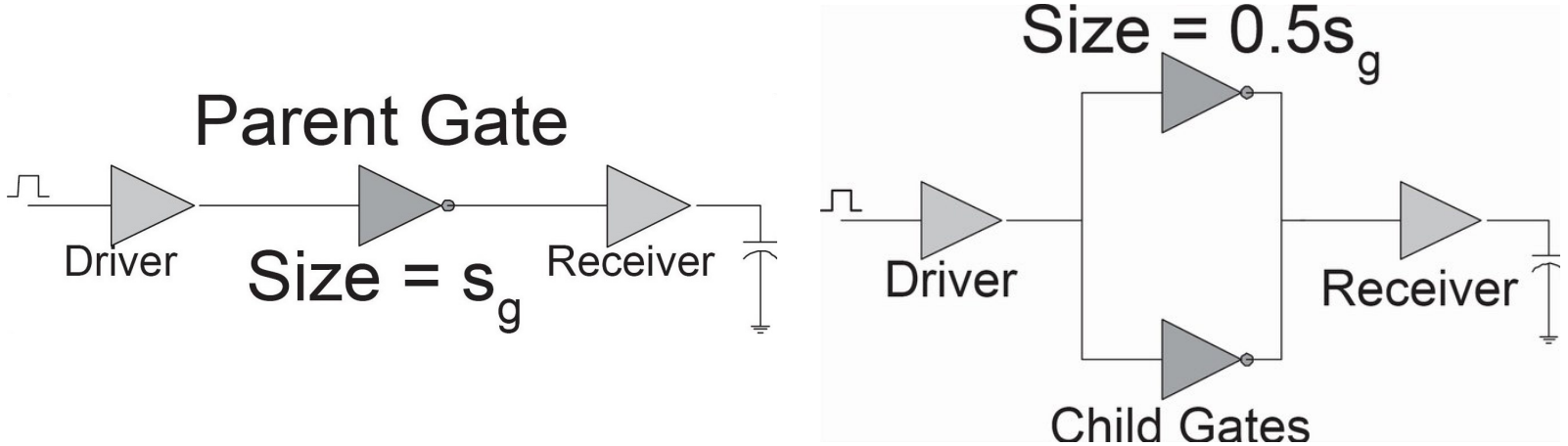
  - The delay of the split configuration

$$d_{split} = (r_d^1 \parallel r_d^2 \parallel \cdots \parallel r_d^N) \quad \text{where} \quad r_d^i = r_d(w/N_s, R_{dish})$$

  - N_S is the number of split wires.
  - N_S = 3 or 4 is suggested.

Gate Splitting

- An example of equivalent inverter splitting



- The two children gates
  - do not present extra load on driver gate
  - do not provide extra driving power to receiver
  - Functionally equivalent to the parent gate

Gate Splitting

- Gate splitting should not be viewed as multi-fingered gates.
  - Multi-fingered gates have fixed distances between two adjacent gates.
  - Gate splitting allows to adjust the distance according to trade-offs.
- The intuition behind gate splitting mechanism:
  - The variance of the sum of two less-than-perfectly correlated random variables is always less than that of two perfectly correlated ones.

Delay Metric with Splitting Technique

- The Elmore delay model for former example
  - delay at output of parent gate

$$d_s^p = 0.69[R_D C_D + (R_D + R_w / N_s)(C_L + C_w) s_g + \frac{r}{s_g} C_R]$$

  - delay at output of split children gates

$$d_s^c = 0.69[R_D C_D + (R_D + R_w / N_s)(C_L + C_w)(s_1 + s_2) + \frac{r}{(s_1 + s_2)} C_R]$$

  - s_g is a function of L_{eff} , and thus a random variable.

Delay Metric with Splitting Technique

- With following assumption
  - s_g is normally distributed, $s_g \sim N(\mu_s, \sigma_s)$
  - $s_1, s_2 \sim N(\mu_s^{'}, \sigma_s^{'})$ with $\mu_s^{'} = 0.5\mu_s$
  - The correlation coefficient between them is ρ

  the two delay values can be rewritten as

$$d_s^p = \left(s_g + \frac{1}{s_g} \right) \qquad d_s^c = \left(s_c + \frac{1}{s_c} \right)$$

- Note:
  - $s_c = s_1 + s_2$, $s_c \sim N(\mu_s, \sigma_{s_c})$ where $\sigma_{s_c} = \sigma_s^{'}\sqrt{2(1+\rho)}$
  - Proofs can be extended to non-Gaussian cases.

Delay Metric with Splitting Technique

- We can derive the ratio
  - for $\rho = 0$

$$\frac{\sigma(d_s^c)}{\sigma(d_s^p)} = 0.7282 - 0.0076\mu_s + 0.0077\mu_s^2 - 0.0015\mu_s^3$$

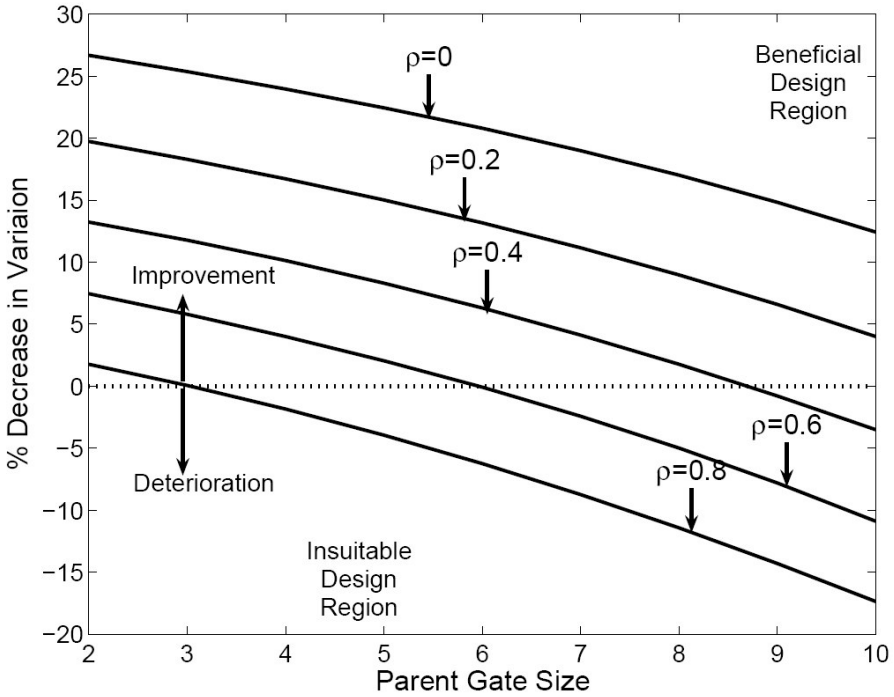
  - for $\rho = 0.5$

$$\frac{\sigma(d_s^c)}{\sigma(d_s^p)} = 0.8485 - 0.1825\mu_s + 0.0007\mu_s^2 - 0.0001\mu_s^3$$

- Note:
  - There are similar expressions for non-Gaussian cases.
  - The ratio is less than 1 for different μ_s and ρ .

Delay Metric with Splitting Technique

- Relative decrease in delay variance



- The reduction exhibits inherent advantage of splitting any gate.
- Not all the gates exhibit a decrease in variation.
- The beneficial design region (above the dotted line)

TURGIS Algorithm

- *Timing Uncertainty Reduction by Gate and Interconnect Splitting*
  - To optimally choose the splitting candidates
  - SSTA: to provide a list of split candidates
  - $d_f = d(G_f, T_i)$: to offer delay estimation adjustment which is the distance between a spin-off gate G_f and steiner tree T_i .
  - Apply splitting to those elements contributing most to output delay variance

TURGIS Algorithm Flow

$\mathcal{A} \leftarrow$ STA$(circuit)$
$\mathcal{PO} \leftarrow$ Extract primary outputs$(circuit)$
$\mathcal{PI} \leftarrow$ Extract primary inputs$(circuit)$
for each $\lambda \in \mathcal{PO}$
  do
    $slp_\lambda = \{\emptyset\}$
    PUSH(slp_λ, λ)
    parent node $\leftarrow \lambda$
    while parent node $\notin \mathcal{PI}$
      do
        fan-ins $\leftarrow$ Find Fan-In(parent node)
        $\nu \leftarrow$ FIND LARGEST(fan-ins)
        PUSH(slp_λ, ν)
        parent node $\leftarrow \nu$
for each $g \in slp$
  do
    $s_g \leftarrow$ Compute Sensitivity(g)
    if $s_g > threshold$
      then
        $\mathcal{R} \leftarrow$ Locate(g)
        if $\mathcal{R} \subset$ BeneficialDesignRegion
          then circuit $\leftarrow$ Split(g)
return (circuit)

- 2 operations in TURGIS
  - Random variable comparison
    (FIND LARGEST( ))
  - Sensitivity computation
    (Compute Sensitivity( ))

Random Variable Comparison

- For any 2 random variables x and y
  - x is termed as statistically larger than y if

$$\left(\mu_x - \mu_y\right) \Big/ \sqrt{\sigma_x^2 + \sigma_y^2} \geq 3$$

  - otherwise y is statistically larger than x if

$$\left(\mu_y - \mu_x\right) \Big/ \sqrt{\sigma_x^2 + \sigma_y^2} \geq 3$$

  - If none of the above is true

$$\frac{\partial \Gamma(x, y)}{\partial x} \geq \frac{\partial \Gamma(x, y)}{\partial y} \Rightarrow x \geq y$$

Output Delay

- For any circuit with
  - N split elements
  - size profile $S = \{s_1, s_2, \cdots s_N\}$

  The output delay can be modeled in form of

$$d_0(S) = \alpha + \sum_{i=1}^{N} \beta_i s_i + \sum_{i=1}^{N} \gamma_i s_i^2 + \sum_{i=1}^{N-1} \sum_{i=1}^{N} \delta_{ij} s_s s_j$$

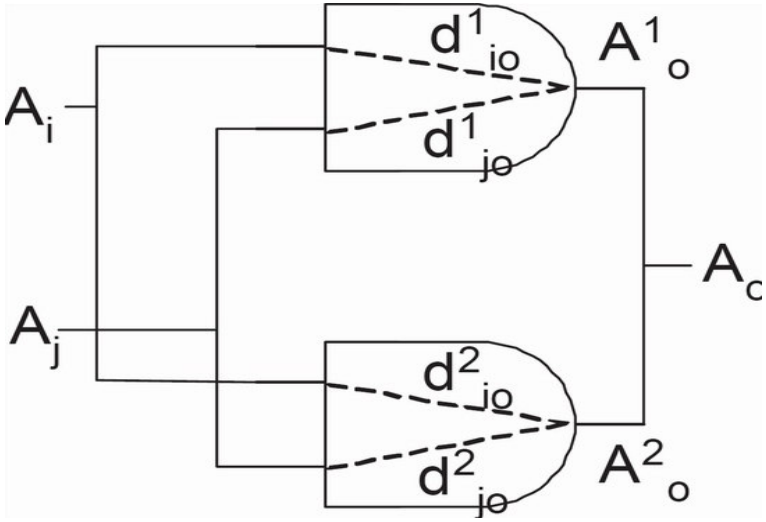
- To establish the coefficients
  - Response Surface Modeling
  - Lest Square Fitting

Sensitivity Computation

- Steps to compute element's sensitivity
  - $F(S) = f_{s_1 s_2 \cdots s_N}(s_1, s_2, \cdots s_N)$
  - $\mu(d_0) = \int\int \cdots \int d_0 F(S) ds_1 ds_2 \cdots ds_N$
  - $\sigma^2(d_0) = \int\int \cdots \int [d_0 - \mu(d_0)]^2 F(S) ds_1 ds_2 \cdots ds_N$
  - $\mu(d_0 \mid s_i) = \int\int \cdots \int d_0 F(S \mid s_i) ds_1 \cdots ds_{i-1} ds_{i+1} \cdots ds_N$
  - $sen(g_i) = \frac{1}{\sigma^2(d_0)} \int [\mu(d_0 \mid s_i)]^2 f(s_i) ds_i$

MAX and MIN Operator

- Applied to include multi-driver effects
- Definition:



- $A_0^1 = \max(A_1 + d_{io}^1, A_j + d_{jo}^1)$
- $A_0^2 = \max(A_i + d_{io}^2, A_j + d_{jo}^2)$
- $A_0 = \min(A_0^1, A_0^2)$

- Can be confirmed using SPICE simulations

Operator Implementation

- Using Cumulative Distribution Function (CDF)
  - MAX operator

$$A_0 = \max(A_a, A_b) \Rightarrow C_0(t) = C_a(t)C_b(t)$$

  - MIN operator

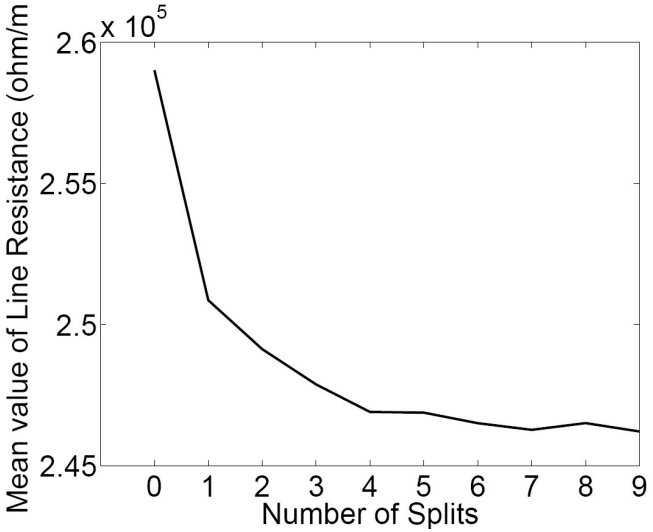
$$A_0 = \min(A_a, A_b) \Rightarrow C_0(t) = C_a(t) + C_b(t) - C_{ab}(t,t)$$

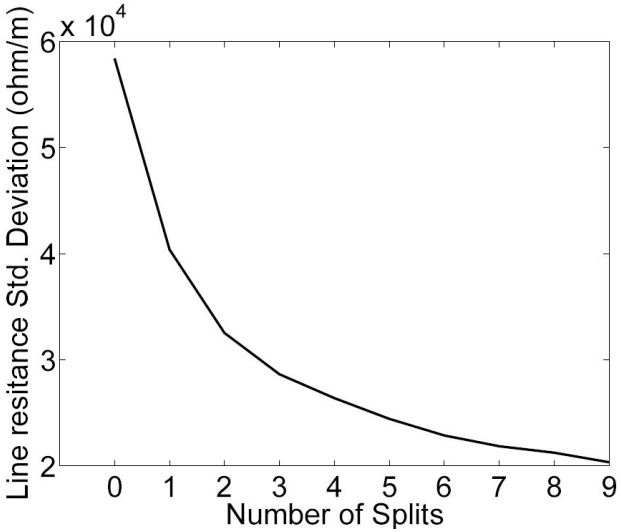
Experimental Results

- Experiment Environment
  - Done in MATLAB and HSPICE
  - Berkeley PTM interconnect model
  - A computer running Windows OS with 1.5GHZ clock frequency and 512MB RAM
- Experiment Condition
  - A interconnect of length 1 centimeter
  - Width and height are assumed normal distribution
  - R_{dish} is assumed 4x the wire width

Experimental Results

- Mean and Deviation of Interconnect resistance

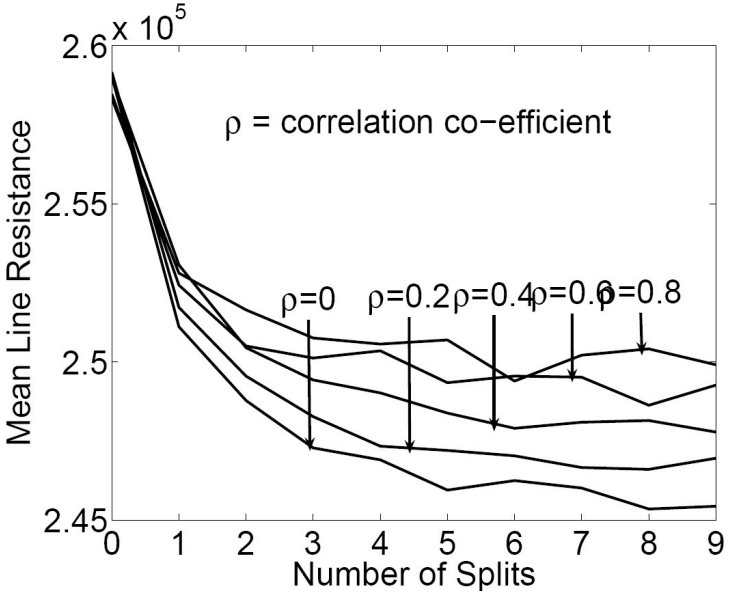


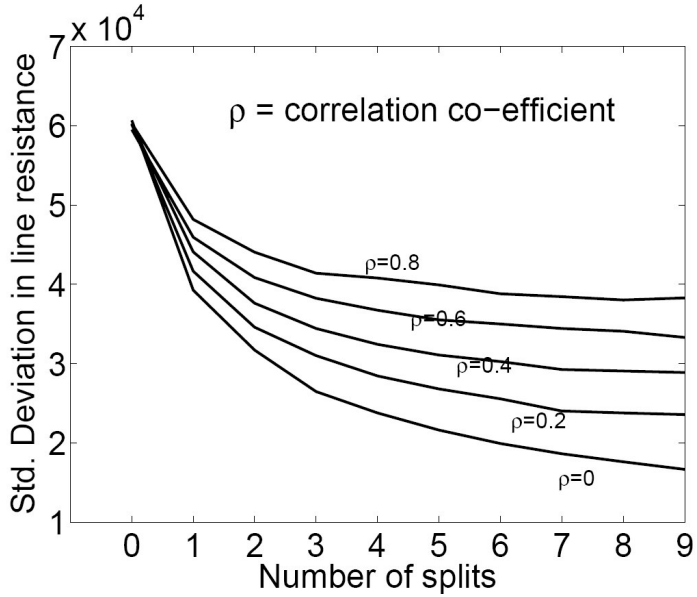


  - Both decrease with increasing splits number.
  - Variation reduction of up to 55% is achieved.
  - 3 is the optimal number of splits.

Experimental Results

- The spatial distance between split lines





  - Reduction reduces with increasing correlation coefficient.
  - Trade-off between variation reduction and cell area

Experimental Results

- Results on Various Length NAND Gate Chains

Number of gates	Sizing Profile	Original Chain Statistics			Split Gate Chain Statistics			Improvement (% Decrease)		
		μ	σ	σ/μ	μ	σ	σ/μ	μ	σ	σ/μ
5	$\Re$	87.83	1.404	0.015	87.83	1.024	0.011	0.00	27.06	26.67
5	$\aleph$	82.03	0.949	0.011	81.96	0.805	0.009	0.08	15.17	18.18
10	$\Re$	198.65	3.395	0.016	198.34	2.514	0.012	0.15	25.95	25.00
10	$\aleph$	175.54	0.995	0.005	175.42	0.847	0.004	0.06	14.87	20.00
20	$\Re$	398.09	3.322	0.008	397.61	2.719	0.006	0.12	18.15	25.00
20	$\aleph$	362.93	1.121	0.003	362.44	0.756	0.002	0.13	32.56	33.33

$\Re \Rightarrow$ Random Sizing $\aleph \Rightarrow$ All gates size = 6

Experimental Results

- Original Output Delay statistics for ISCAS Benchmark circuits

Circuit name	Original statistics (ps)			
	N_T	μ	σ	(σ/μ)
c17	6	42.97	2.98	0.069
c432	160	588.20	33.22	0.053
c499	202	962.49	133.86	0.140
c880	383	200.36	6.26	0.049
c1355	546	1212.65	95.54	0.079
c1908	880	1444.95	159.95	0.100
c2670	1193	410.43	33.38	0.110
c3540	1669	2308.53	52.19	0.032

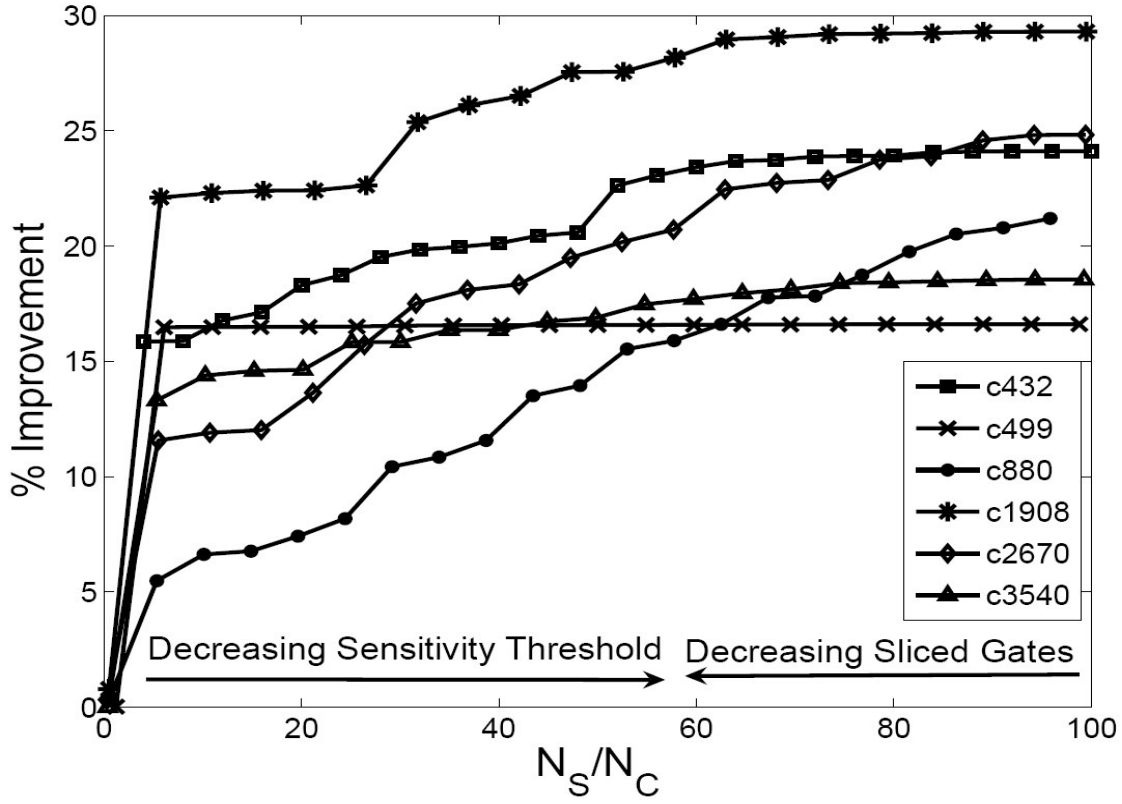
Experimental Results

- Output Delay statistics for ISCAS Benchmark circuits split on all the gates

Circuit name	Split on All gates Statistics (ps)				Improvements (% Decrease)				CPU time (sec)
	N	μ	σ	(σ/μ)	$\frac{N}{N_T}$	$\Delta\mu$	$\Delta\sigma$	(σ/μ)	
c17	6	42.32	2.31	0.054	1	1.51	22.48	21.73	0.02
c432	160	558.26	24.01	0.041	1	5.09	27.72	22.64	0.22
c499	202	915.06	116.42	0.129	1	4.92	13.02	7.85	0.35
c880	383	193.45	5.06	0.038	1	3.44	19.16	22.44	0.51
c1355	546	1156.67	82.55	0.072	1	4.61	13.59	8.86	0.89
c1908	880	1332.45	111.48	0.075	1	7.78	30.30	25.00	2.52
c2670	1193	385.22	23.85	0.085	1	6.14	28.55	22.73	4.21
c3540	1669	2089.71	42.26	0.028	1	9.47	19.02	12.50	19.95
	Average Improvements				0	5.37	21.73	17.97	

Experimental Results

- Average Improvement in Output Delay variance on ISCAS Benchmark by TURGIS



Conclusion

- Using Interconnect splitting to reduce variation
- Overcoming disadvantages of previous techniques
- Noticeable improvements in experiments
  - Improvement of 5.37% in mean value
  - Improvement of 21.73% in variance