ESE535:
Electronic Design Automation

Day 24: April 22, 2015
Statistical Static Timing Analysis
Delay PDFs? (2a)
Today

• Sources of Variation
• Limits of Worst Case
• Optimization for Parametric Yield
• Statistical Analysis

Behavioral (C, MATLAB, …)
  ↓ Arch. Select Schedule
  ↓ RTL
  ↓ FSM assign
  ↓ Two-level Multilevel opt.
  ↓ Covering Retiming
  ↓ Gate Netlist
  ↓ Placement Routing
  ↓ Layout
  ↓ Masks
Central Problem

- As our devices approach the atomic scale, we must deal with statistical effects governing the placement and behavior of individual atoms and electrons.

- Transistor critical dimensions
  - Atomic discreteness
  - Subwavelength litho
  - Etch/polish rates
  - Focus
- Number of dopants
- Dopant Placement
Oxide Thickness

[Asenov et al. TRED 2002]

Fig. 1. (a) Typical profile of the random Si/SiO$_2$ interface in a $30 \times 30$ nm$^2$ MOSFET, followed by (b) an equiconcentration contour obtained from DG simulations, and (c) the potential distribution.
Line Edge Roughness

- 1.2μm and 2.4μm lines

From:
http://www.microtechweb.com/2d/lw_pict.htm
Light

- What is wavelength of visible light?
Phase Shift Masking

Source
http://www.synopsys.com/Tools/Manufacturing/MaskSynthesis/PSMCreate/Pages/default.aspx
Line Edges (PSM)

Source:
Intel 65nm SRAM (PSM)

Source:
http://www.intel.com/technology/itj/2008/v12i2/5-design/figures/Figure_5_lg.gif
Statistical Dopant Placement

[Bernstein et al, IBM JRD 2006]
$V_{th}$ Variability @ 65nm

No. of devices = 3,481

$\sigma_{V_T} = 25.58$ mV

$L = 46$ nm

$W = 122$ nm

[Bernstein et al, IBM JRD 2006]
Gaussian Distribution

# ITRS 2011 Variation (3\(\sigma\))

| Table DESN10  Design for Manufacturability Technology Requirements |
|---------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Normalized mask cost from public and IDM data | 1.0             | 1.3             | 1.7             | 2.2             | 2.8             | 3.7             | 4.9             | 6.4             | 8.4             | 11.0            | 14.4            | 18.8            | 24.6            | 32.1            | ?               | ?               |
| \% \(V_{dd}\) variability: % variability seen in on-chip circuits | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             | 10%             |
| \% \(V_{th}\) variability: doping variability impact on \(V_{th}\) (minimum size devices, memory) | 40%             | 40%             | 40%             | 45%             | 45%             | 50%             | 50%             | 55%             | 55%             | 60%             | 60%             | 65%             | 65%             | 70%             | 70%             | 75%             | 75%             |
| \% \(V_{th}\) variability: includes all sources | 42%             | 42%             | 42%             | 47%             | 47%             | 53%             | 53%             | 58%             | 58%             | 63%             | 63%             | 68%             | 68%             | 74%             | 74%             | 79%             | 79%             |
| \% \(V_{th}\) variability: typical size logic devices, all sources | 20%             | 20%             | 20%             | 23%             | 23%             | 25%             | 25%             | 28%             | 28%             | 30%             | 30%             | 33%             | 33%             | 35%             | 35%             | 38%             | 38%             |
| \% CD variability | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             | 12%             |
| \% circuit performance variability circuit comprising gates and wires | 42%             | 42%             | 42%             | 45%             | 45%             | 47%             | 47%             | 50%             | 50%             | 52%             | 52%             | 55%             | 55%             | 57%             | 57%             | 60%             | 60%             |
| \% circuit total power variability circuit comprising gates and wires | 51%             | 51%             | 51%             | 55%             | 55%             | 59%             | 59%             | 63%             | 63%             | 68%             | 68%             | 72%             | 72%             | 77%             | 77%             | 81%             | 81%             |
| \% circuit leakage power variability circuit comprising gates and wires | 126%            | 126%            | 126%            | 129%            | 129%            | 132%            | 132%            | 135%            | 135%            | 138%            | 138%            | 141%            | 141%            | 145%            | 145%            | 148%            | 148%            |
Example: $V_{th}$

- Many physical effects impact $V_{th}$
  - Doping, dimensions, roughness
- Behavior highly dependent on $V_{th}$

\[
I_{DS} = \frac{\mu_n C_{OX}}{2} \left( \frac{W}{L} \right) \left[ (V_{GS} - V_{th})^2 \right]
\]

\[
I_{DS} = I_s \left( \frac{W}{L} \right) e^{\left( \frac{V_{GS} - V_{th}}{nkT/q} \right)} \left( 1 - e^{-\left( \frac{V_{DS}}{kT/q} \right)} \right) (1 + \lambda V_{DS})
\]
Impact Performance

- $V_{th} \rightarrow I_{ds} \rightarrow \text{Delay (} C_{load}/I_{ds} \text{)}$
Impact of $V_{th}$ Variation

![Graph showing the impact of $V_{th}$ variation on delay. The x-axis represents $V_{dd}$ (V) ranging from 0.0 to 1.4, and the y-axis represents delay in seconds ranging from $10^{-13}$ to $10^{-6}$. Different curves represent various $V_{th}$ values, such as $V_{th} = 500\, mV$, $V_{th} = 450\, mV$, $V_{th} = 400\, mV$, $V_{th} = 350\, mV$, $V_{th} = 300\, mV$, $V_{th} = 250\, mV$, and $V_{th} = 200\, mV$.](image-url)
Variation in Current FPGAs

(a) Measured Delay

(b) CAD Delay

(c) Correlation

[Gojman et al., FPGA2013]
Reduce Vdd
(Cyclone IV 60nm LP)

[Gojman et al., FPGA2013]
Scale of Variations

Die-to-Die (D2D) Variations

Within-Die (WID) Variations
  - Systematic
  - (Uncorrelated) Random

Wafer Scale
Die Scale
Feature Scale

Source: Noel Menezes, Intel ISPD2007
Nature of correlated variation

- **CDs of transistors that are close track**
  - Tracking diminishes with distance

Source: Noel Menezes, Intel ISPD2007
Old Way

• Characterize gates by corner cases
  – Fast, nominal, slow
• Add up corners to estimate range

• Preclass:
  – Slow corner: 1.1
  – Nominal: 1.0
  – Fast corner: 0.9
Corners Misleading

- Probabilistic Timing
- Traditional Timing

Non-probabilistic Timing - Inaccurate
Non-probabilistic Timing - Conservative

$T_{\text{clock}}$ at 98th percentile
Worst-case $T_{\text{clock}}$

probability density, $f(T_{\text{clock}})$

Clock Period ($T_{\text{clock}}$)

$D_0$, $E_{\max\{D\}}$
Parameteric Yield

Probability Distribution

Delay

Sell Premium

Sell nominal

Sell cheap

Discard
Phenomena 1: Path Averaging

- $T_{path} = t_0 + t_1 + t_2 + t_3 + \ldots + t_{(d-1)}$
- $T_i$ – iid random variables
  - Mean $\tau$
  - Variance $\sigma$
- $T_{path}$
  - Mean $d \times \tau$
  - Variance $= \sqrt{d} \times \sigma$
Sequential Paths

- $T_{\text{path}} = t_0 + t_1 + t_2 + t_3 + \ldots + t_{(d-1)}$
- $T_{\text{path}}$
  - Mean $d \times \tau$
  - Variance $= \sqrt{d} \times \sigma$

- 3 sigma delay on path: $d \times \tau + 3\sqrt{d} \times \sigma$
  - Worst case per component would be: $d \times (\tau + 3 \sigma)$
  - Overestimate $d$ vs. $\sqrt{d}$
SSTA vs. Corner Models

- STA with corners predicts 225ps
- SSTA predicts 162ps at 3σ
- SSTA reduces pessimism by 28%

Source: IBM, TRCAD 2006

[Slide composed by Nikil Mehta]

Penn ESE535 Spring 2013 -- DeHon

Source: IBM, TRCAD 2006
Phenomena 2: Parallel Paths

- Cycle time limited by slowest path
- $T_{cycle} = \max(T_{p0}, T_{p1}, T_{p2}, \ldots T_{p(n-1)})$
- $P(T_{cycle} < T_0) = P(T_{p0} < T_0) \times P(T_{p1} < T_0) \ldots$
  
  $= [P(T_p < T_0)]^n$

- $0.5 = [P(T_p < T_{50})]^n$
- $P(T_p < T_{50}) = (0.5)^{(1/n)}$
System Delay

- $P(T_p<T_{50}) = (0.5)^{1/n}$
  - $N=10^8 \Rightarrow 0.999999993$
    - $1-7\times10^{-9}$
  - $N=10^{10} \Rightarrow 0.99999999993$
    - $1-7\times10^{-11}$
Gaussian Distribution

System Delay

- \( P(T_p < T_{50}) = (0.5)^{(1/n)} \)
  - \( N=10^8 \rightarrow 0.999999993 \)
    - \( 1 - 7 \times 10^{-9} \)
  - \( N=10^{10} \rightarrow 0.99999999993 \)
    - \( 1 - 7 \times 10^{-11} \)

- For 50% yield want
  - 6 to 7 \( \sigma \)
  - \( T_{50} = T_{\text{mean}} + 7\sigma_{\text{path}} \)
System Delay

μ = 1, σ = 0.03

Bowman, et al.  
ISSCC 2001
System Delay

T50 Delay vs Paths

(MaxDelay-MeanDelay)/Sigma vs Paths

Paths: 1, 100, 10000, 1e+06, 1e+08, 1e+10, 1e+12

T50
Corners Misleading

Phenomena 1

Phenomena 2

[Orshansky+Keutzer DAC 2002]
SSTA gain relative to $3\sigma$ corner analysis: Random

$$\text{gain} = \frac{1 + 3\sigma_{\text{rand}}}{\Phi^{-1}(Y^{1/N_p}) \sigma_{\text{rand}}}$$

$\Phi^{-1}$ is the inverse cumulative distribution function of the standard normal distribution.

99.9% yield corner
3σ random corner
Overdesign

SSTA provides a potential gain relative to a 3σ corner-based STA for purely random variation
Source: Noel Menezes, Intel ISPD2007
But does worst-case mislead?

• STA with worst-case says these are equivalent:
But does worst-case mislead?

- STA Worst case delay for this?
But does worst-case mislead?

- STA Worst case delay for this?
Does Worst-Case Mislead?

- Delay of off-critical path may matter
- Best case delay of critical path?
- Worst-case delay of non-critical path?
Probability of Path Being Critical

Figure 1: Probability that a path shows up in top 50 paths
(Data from Monte Carlo simulation of a 90nm microprocessor block)

[Source: Intel DAC 2005]
What do we need to do?

• Ideal:
  – Compute PDF for delay at each gate
  – Compute delay of a gate as a PDF from:
    • PDF of inputs
    • PDF of gate delay
Delay Calculation

AND rules

<table>
<thead>
<tr>
<th>$i_1 \rightarrow$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_2 \downarrow$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>$MIN(l_1, l_2) + d$</td>
<td>$l_2 + d$</td>
<td>$MIN(l_1, l_2) + d$</td>
</tr>
<tr>
<td></td>
<td>$MIN(u_1, u_2) + d$</td>
<td>$u_2 + d$</td>
<td>$u_2 + d$</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$l_1 + d$</td>
<td>$MAX(l_1, l_2) + d$</td>
<td>$l_1 + d$</td>
</tr>
<tr>
<td></td>
<td>$u_1 + d$</td>
<td>$MAX(u_1, u_2) + d$</td>
<td>$MAX(u_1, u_2) + d$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$MIN(l_1, l_2) + d$</td>
<td>$l_2 + d$</td>
<td>$MIN(l_1, l_2) + d$</td>
</tr>
<tr>
<td></td>
<td>$u_1 + d$</td>
<td>$MAX(u_1, u_2) + d$</td>
<td>$MAX(u_1, u_2) + d$</td>
</tr>
</tbody>
</table>
What do we need to do?

• Ideal:
  – compute PDF for delay at each gate
  – Compute delay of a gate as a PDF from:
    • PDF of inputs
    • PDF of gate delay
  – Need to compute for distributions
    • SUM
    • MAX (maybe MIN)
For Example

• Consider entry:
  – MAX(u₁,u₂)+d
MAX

• Compute MAX of two input distributions
  – Preclass 2(b)
• Add that distribution to gate distribution.
Continuous

• Can roughly carry through PDF calculations with Gaussian distributions rather than discrete PDFs
Dealing with PDFs

• Simple model assume all PDFs are Gaussian
  – Model with mean, $\sigma$
  – Imperfect
    • Not all phenomena are Gaussian
    • Sum of Gaussians is Gaussian
    • Max of Gaussians is not a Gaussian
MAX of Two Identical Gaussians

- Max of Gaussians is not a Gaussian
- Can try to approximate as Gaussian
- Given two identical Gaussians A and B with $\mu$ and $\sigma$
- Plug into equations
- $E[\text{MAX}(A,B)] = \mu + \sigma/\pi^{1/2}$
- $\text{VAR}[\text{MAX}(A,B)] = \sigma^2 - \sigma/\pi$
  [Source: Nikil Mehta]
More Technicalities

• Correlation
  – Physical on die
  – In path (reconvergent fanout)
    • Makes result conservative
      – Gives upper bound
      – Can compute lower

Graphics from: Noel Menezes (top) and Nikil Mehta (bottom)
Max of Gaussians with Correlation

- Max of identical Gaussians

[Blaauw et al. TRCAD v27n4p589]
MAX of Two Identical Gaussians

- Given two identical Gaussians A and B with \( \mu \) and \( \sigma \)
- Plug into equations
  - \( \text{E}[\text{MAX}(A,B)] = \mu + \frac{\sigma}{\pi}^{1/2} \)
  - \( \text{VAR}[\text{MAX}(A,B)] = \sigma^2 - \frac{\sigma}{\pi} \)

[Source: Nikil Mehta]

Extreme of correlated: is just the input Gaussian
# SSTA vs. Monte Carlo Verification Time

## TABLE II

**MONTE CARLO VERSUS EinsStat COMPARISON**

<table>
<thead>
<tr>
<th>Test case</th>
<th>Gates</th>
<th>EinsStat CPU</th>
<th>Monte Carlo</th>
<th>Monte Carlo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Samples</td>
<td>Sequential CPU</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dd:hh:mm:ss</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
<td>1 sec.</td>
<td>100000</td>
<td>5:57</td>
</tr>
<tr>
<td>2</td>
<td>3042</td>
<td>2 sec.</td>
<td>100000</td>
<td>2:01:15:10</td>
</tr>
<tr>
<td>3</td>
<td>11937</td>
<td>7 sec.</td>
<td>10000</td>
<td>0:20:33:40</td>
</tr>
<tr>
<td>4</td>
<td>70216</td>
<td>59 sec.</td>
<td>10000</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Source: IBM, TRCAD 2006
Using SSTA in FPGA CAD

- Le Hei
  - FPGA2007
  - SSTA Synthesis, Place, Route
- Kia
  - FPGA2007
  - Route with SSTA

<table>
<thead>
<tr>
<th>Circuit</th>
<th>Delay Impr. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ex5p</td>
<td>6.58</td>
</tr>
<tr>
<td>alu4</td>
<td>1.50</td>
</tr>
<tr>
<td>misex3</td>
<td>5.76</td>
</tr>
<tr>
<td>apex2</td>
<td>3.24</td>
</tr>
<tr>
<td>apex4</td>
<td>2.57</td>
</tr>
<tr>
<td>pdc</td>
<td>4.74</td>
</tr>
<tr>
<td>seq</td>
<td>4.37</td>
</tr>
<tr>
<td>des</td>
<td>3.73</td>
</tr>
<tr>
<td>spla</td>
<td>4.82</td>
</tr>
<tr>
<td>ex1010</td>
<td>1.83</td>
</tr>
<tr>
<td>frisc</td>
<td>2.84</td>
</tr>
<tr>
<td>elliptic</td>
<td>0.17</td>
</tr>
<tr>
<td>bigkey</td>
<td>0.35</td>
</tr>
<tr>
<td>s298</td>
<td>7.10</td>
</tr>
<tr>
<td>tseng</td>
<td>5.93</td>
</tr>
<tr>
<td>diffeq</td>
<td>4.16</td>
</tr>
<tr>
<td>dsip</td>
<td>7.37</td>
</tr>
<tr>
<td>s38417</td>
<td>7.56</td>
</tr>
<tr>
<td>s38584.1</td>
<td>5.43</td>
</tr>
<tr>
<td>clma</td>
<td>-1.17</td>
</tr>
<tr>
<td>Mean</td>
<td>3.95</td>
</tr>
</tbody>
</table>

Table 6: Comparison of mean delay and standard deviation between deterministic and stochastic flows under various process variation assumptions (based on the geometric mean of 20 MCNC designs).
Impact of SSTA in High-Level Synthesis

<table>
<thead>
<tr>
<th>Design (#ops)</th>
<th>#ALU, #MUL</th>
<th>$p_{clk}$ (ns)</th>
<th>Latency (cycles)</th>
<th>Reduction</th>
<th>Run time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS[15]</td>
<td>HLS-tv(Y)</td>
<td></td>
</tr>
<tr>
<td><strong>DIFF (18)</strong></td>
<td>3, 3</td>
<td>3.5</td>
<td>32</td>
<td>28 (94.5%)</td>
<td>12.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0</td>
<td>29</td>
<td>24 (90.7%)</td>
<td>17.2%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5</td>
<td>26</td>
<td>22 (93.2%)</td>
<td>15.4%</td>
</tr>
<tr>
<td><strong>LATT (22)</strong></td>
<td>3, 2</td>
<td>3.5</td>
<td>47</td>
<td>36 (94.3%)</td>
<td>23.4%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0</td>
<td>42</td>
<td>32 (94.3%)</td>
<td>23.8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5</td>
<td>37</td>
<td>30 (90.2%)</td>
<td>18.9%</td>
</tr>
<tr>
<td><strong>AR (28)</strong></td>
<td>2, 3</td>
<td>3.5</td>
<td>57</td>
<td>45 (93.9%)</td>
<td>21.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0</td>
<td>51</td>
<td>40 (93.9%)</td>
<td>21.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5</td>
<td>45</td>
<td>36 (90.8%)</td>
<td>20.0%</td>
</tr>
<tr>
<td><strong>EWF (34)</strong></td>
<td>2, 3</td>
<td>3.5</td>
<td>46</td>
<td>37 (93.6%)</td>
<td>19.6%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.0</td>
<td>42</td>
<td>34 (93.6%)</td>
<td>19.0%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.5</td>
<td>38</td>
<td>33 (91.5%)</td>
<td>13.2%</td>
</tr>
<tr>
<td><strong>avg.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>(92.9%)</strong></td>
</tr>
</tbody>
</table>

- Scheduling and provisioning
  - ALU/MUL $\sigma=5\% \ t_{\text{nominal}}$

[Jung&Kim ICCAD2007]
Summary

- Nanoscale fabrication is a statistical process
- Delays are PDFs
- Assuming each device is worst-case delay is too pessimistic
  - Wrong prediction about timing
  - Leads optimization in wrong direction
- Reformulate timing analysis as statistical calculation
- Estimate the PDF of circuit delays
- Use this to drive optimizations
Big Ideas:

• Coping with uncertainty
• Statistical Reasoning and Calculation
Admin

- Reading for Monday on canvas
- Milestone Thursday