Review of Register-transfer Level Design Flow and a Look at Industrial Practice

Prof. Kurt Keutzer
EECS
keutzer@eecs.berkeley.edu

Feedback

- Good
  - Entire flow of CAD
  - Real world perspective
  - Examples of algorithms
  - Class structure → problem → formulation → algorithm
  - No homework
  - Dialog with class/ “Socratic approach”
- Bad
  - No homework
  - Questions to class sometimes unclear
  - Too fast, need more details
  - Energy of lecture up and down some days

Actually, this was last year’s feedback!!!
Feedback – this year

- Good
  - Entire flow of CAD
  - Real world perspective
  - Examples of algorithms
  - Good slides
  - No homework
  - Questions invited in class
- Bad
  - No homework
  - Too fast, need more details – more examples of algorithms
  - Slides not updated on webpage before class – slides don’t match the lecture
  - Can’t see fonts on slides
  - Don’t know what to read before class

Responding to feedback

- Graduate course without an undergraduate “Intro to CAD” course – it’s a challenge for all of us
  - No homework -
    - we voted – it was your call
  - Too fast, need more details –
    - A lot of material but a modest amount of class work
    - more examples of algorithms? Welcome to graduate school!
- Slides not updated on webpage before class – slides don’t match the lecture
  - I add new questions every year – don’t want to update before class
  - Should be updated immediately after class – my bad
- Can’t see fonts on slides – get glasses, really!
- Don’t know what to read before class – check the web page – reading assignments have been there
- Speak up!! Raise your questions/concerns during class
Design Process

- **Design**: specify and enter the design intent

**Verify:**
verify the correctness of design and implementation

**Implement:**
refine the design through all phases

How are designers describing ICs?
Current Practice: HDL at RTL Level

```vhdl
module foobar (q, clk, s, a, b);
  input clk, s, a, b;
  output q; reg q; reg d;
always @(a or b or s) // mux
begin
  if( !s )
    d = a;
  else if( s )
    d = b;
  else
    d = 'bx;
end // always @(a or b or s)

always @(clk) // latch
begin
  if( clk == 1 )
    q = d;
  else if( clk !== 0 )
    q = 'bx;
end // always @(clk)
endmodule
```

Verification

- **Design**: specify and enter the design intent
- **Verify**: verify the correctness of design and implementation
- **Implement**: refine the design through all phases
What are the three phases of verification?

Design Verification

Is the design consistent with the original specification?

Is what I think I want what I really want?
Implementation Verification

Is the implementation consistent with the original design intent?
Is what I implemented what I wanted?

Manufacture Verification (Test)

Is the manufactured circuit consistent with the implemented design?
Did they build what I wanted?
What is the work horse of design verification?

Is the design consistent with the original specification?
Is what I think I want what I really want?
Types of software simulators

- Circuit simulation
  - Spice, Advice, Hspice
  - Timemill + Ace, ADM
- Event-driven gate/RTL/Behavioral simulation
  - Verilog - VCS, NC-Verilog, Turbo-Verilog, Verilog-XL
  - VHDL - VSS, MTI, Leapfrog
- Cycle-based gate/RTL/Behavioral simulation
  - Verilog - Frontline, Speedsim
  - VHDL - Cyclone
- Domain-specific simulation
  - SPW, COSSAP
- Architecture-specific simulation

What are the work horses of implementation verification?

- What are the key techniques employed?
Use static analysis techniques to verify:

- use static timing analysis
- longest path delay calculation
- false path elimination

Use static analysis techniques to verify:

- formal equivalence-checking techniques
- testing based
- BDD based
- SAT based
- structural
What are the work horses of manufacturer test?

- What are the key techniques employed?

Manufacture Verification (Test)

- Scan-based methodology for high stuck-at fault coverage
- Test vector generation using:
  - Podem-like methods
  - SAT
- Coverage grading using fault simulation
What kind of flow is used to design most ASICs?

RTL Synthesis Flow

![RTL Synthesis Flow Diagram]
**RTL Synthesis**

```verilog
module foobar (q, clk, s, a, b);
    input clk, s, a, b;
    output q; reg q; reg d;
    always @(a or b or s) // mux
        begin
            if(~s )
                d = a;
            else if( s )
                d = b;
            else
                d = 'bx;
        end // always @(a or b or s)
endmodule
```

---

**Logic Optimization**

- Perform a variety of transformations and optimizations
  - Structural graph transformations
  - Boolean transformations
  - Mapping into a physical library
- What are the key algorithms?
Physical Design

- Transform sequential circuit netlist into a physical circuit
  - *place* circuit components
  - *route* wires
  - transform into a mask
- Or for FPGA’s
  - *place* look-up tables
  - *route* wires
- What are the key algorithms?

Ok – now what do companies really do?

- What do they build?
- How do real designers put things together?
- Where does the time go?
- And what challenges do they face?
Nvidia GeForce4 Ti Architecture

- 0.15u, 8LM, 63M txtr, 505 signals/801 balls, 18W

A. Khan – Cadence - 2002

- 0.5um ➔ 0.18um
- 5x5 mm² ➔ 21.7x21.3 mm²
- ~0.8M ➔ 287.5M transistors
- 3LM ➔ 6LM (8LM)
- ~100 ➔ 150 MHz (622 MHz)

# Cirrus Logic, Inc. IC, 3Ci™
* Sony Computer Entertainment, Inc. & Sony Corporation
Graphics Synthesizer®-32. Copyright 2000
Sony Computer Entertainment, Inc.
**The Design and Implementation of a First-Generation CELL Processor (IBM, Sony, Toshiba)**

- 164b PPE + 8 SPEs; typ. clock 4GHz+
- 234M transistors
- 3 clock networks
- 12ps skew (top-2 metal grids)
  - 850 individually-tuned elements
- Design verified for valid operation with thermal transients

**The Implementation of a 2-core Multi-Threaded Itanium®-Family Processor (Intel)**

- Power reduction primary design priority
  - 130W single uP
  - 2 uPs, 90nm, 2+GHz, 26.5MB cache (L1-L3); 1.72B transistors, 21.5x27.7 sq. mm
    - 300W $\Rightarrow$ 100W
- Manage power to a dynamically-adjustable limit (Max. performance/watt)
  - Multi-L, V$_c$; real-time clock, VDD control (ammeter, thermal monitors)

---

**Table:**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Frequency (GHz)</th>
<th>Power (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current (25°C)</td>
<td>2.6</td>
<td>90</td>
</tr>
<tr>
<td>Current (45°C)</td>
<td>2.6</td>
<td>95</td>
</tr>
<tr>
<td>Current (67°C)</td>
<td>2.6</td>
<td>100</td>
</tr>
<tr>
<td>Current (80°C)</td>
<td>2.6</td>
<td>115</td>
</tr>
<tr>
<td>Current (90°C)</td>
<td>2.6</td>
<td>115</td>
</tr>
<tr>
<td>Current (95°C)</td>
<td>2.6</td>
<td>115</td>
</tr>
<tr>
<td>Current (99°C)</td>
<td>2.6</td>
<td>115</td>
</tr>
<tr>
<td>Current (100°C)</td>
<td>2.6</td>
<td>120</td>
</tr>
</tbody>
</table>

**Diagram:**

Ok – now what do people really do?

- What do they build?
- How do real designers put things together?
- Where does the time go?
- And what challenges do they face?

EDA Design Flow for Implementation

[Diagram showing the EDA Design Flow for Implementation]
### Tools in Stargen Flow

<table>
<thead>
<tr>
<th>Design Tools</th>
<th>Design Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTL creation, checking, debugging</td>
<td>Emacs, vi, Everest HDL-Lint, Novas Debussy</td>
</tr>
<tr>
<td>Verilog simulation/testbench</td>
<td>Synopsys VCS, VERA</td>
</tr>
<tr>
<td>RTL floorplanning</td>
<td>Icenergy SOCPrototype</td>
</tr>
<tr>
<td>Logic synthesis</td>
<td>Synopsys Design Compiler</td>
</tr>
<tr>
<td>Physical synthesis</td>
<td>Synopsys Physical Compiler</td>
</tr>
<tr>
<td>Static timing analysis</td>
<td>Primetime</td>
</tr>
<tr>
<td>Design for test analysis and scan chain insertion</td>
<td>Synopsys DFT Compiler or Mentor DFT Advisor</td>
</tr>
<tr>
<td>Gate netlist floorplanning</td>
<td>Synopsys (Avanti) Jupiter</td>
</tr>
<tr>
<td>Clock tree synthesis, routing</td>
<td>Synopsys (Avanti) Astro</td>
</tr>
<tr>
<td>Extraction</td>
<td>Synopsys (Avanti) Star-RCXT</td>
</tr>
<tr>
<td>Signal integrity</td>
<td>PrimetimeSi, AstroXTalk, AstroRail</td>
</tr>
<tr>
<td>DRC/LVS</td>
<td>Mentor Calibre</td>
</tr>
<tr>
<td>Equivalency checking</td>
<td>Synopsys Formality</td>
</tr>
<tr>
<td>Memory BIST</td>
<td>TBD Tools</td>
</tr>
<tr>
<td>ATPG, IEEE1149</td>
<td>TBD Tools</td>
</tr>
</tbody>
</table>

### Ok – now what do people really do?

- What do they build?
- How do real designers put things together?
- Where does the time go?
- And what challenges do they face?
TOSHIBA CHIP DESIGN
FLOW AND CYCLE TIME

Current Status of RTL Design Flow

- Current RTL design flow is able to produce
  - 10 - 100M logical gate ASIC platforms:
  - Significant portions of high speed microprocessors - e.g. Alpha, Pentium Pro > 1M gate-equivalents > 1GHz.
  - Rapid turnaround ASIC → structured ASIC
  - 6000 (down from 8500) IC designs/year go through this flow
  - Not providing productivity improvement needed to keep up with Moore’s Law
- Successful flow, but stalled out – why?
Ok – now what do people really do?

- What do they build?
- How do real designers put things together?
- Where does the time go?
- And what challenges do they face?

What happens in a process generation?

- In the transition from one process generation (e.g. .18u) to another (e.g. .13u)
- Critical dimensions shrink by a scaling factor $S$, typically $S = 2^{1/2}$
  - Has a “squaring” effect on density $S^2 = 2$ – i.e. same number of transistors in half the area
  - Hence Moore’s Law – 2X density increase every 18 months
  - Ideally has a similar effect on performance $1.3 \times 1.3 = 1.7$
- Vdd reduces by a scaling factor $U$ – more on this later
- As Vdd reduces Vth reduces in order to get performance gains

- And what’s the point of all this?
Defining a technology node – “half pitch”

Figure 4  Definition of Metal Half Pitch

Definition of “half-pitch” – ITRS Roadmap

Drawn and Effective

Drawn – e.g. 90nm

Effective –
ASIC – 70nm
MPU e.g. 45nm

Physical Structure
Technology Scaling Models

• Full Scaling (Constant Electrical Field)
  ideal model — dimensions and voltage scale  
  together by the same factor $S$

• Fixed Voltage Scaling
  most common model until 1990’s
  only dimensions scale, voltages remain constant

• General Scaling
  most realistic for today’s situation —
  Voltages and dimensions scale with different factors –
  $U$ and $S$ respectively

Scaling Relationships for Long Channel Devices

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$, $L$, $t_{ox}$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td></td>
</tr>
<tr>
<td>$V_{DD}$, $V_T$</td>
<td>$1/S$</td>
<td>$1/U$</td>
<td>$1$</td>
<td></td>
</tr>
<tr>
<td>$N_{SUB}$</td>
<td>$V/W_{depl}^2$</td>
<td>$S$</td>
<td>$S^2/U$</td>
<td>$S^2$</td>
</tr>
<tr>
<td>Area/Device</td>
<td>$WL$</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>$C_{ex}$</td>
<td>$1/t_{ox}$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$C_L$</td>
<td>$C_{ex}WL$</td>
<td>$1/S$</td>
<td>$1/S$</td>
<td>$1/S$</td>
</tr>
<tr>
<td>$k_n$, $k_p$</td>
<td>$C_{ox}W/L$</td>
<td>$S$</td>
<td>$S$</td>
<td>$S$</td>
</tr>
<tr>
<td>$I_{av}$</td>
<td>$k_{n,p} V^2$</td>
<td>$1/S$</td>
<td>$S/U^2$</td>
<td>$S$</td>
</tr>
<tr>
<td>$I_P$ (intrinsic)</td>
<td>$C_I V / I_{av}$</td>
<td>$1/S$</td>
<td>$U/S^2$</td>
<td>$1/S^2$</td>
</tr>
<tr>
<td>$P_{av}$</td>
<td>$C_I V^2 / I_{P}$</td>
<td>$1/S^2$</td>
<td>$S/U^3$</td>
<td>$S$</td>
</tr>
<tr>
<td>$PDP$</td>
<td>$C_I V^2$</td>
<td>$1/S^3$</td>
<td>$1/SU^2$</td>
<td>$1/S$</td>
</tr>
</tbody>
</table>

Table 3.1: Scaling Relationships for Long Channel Devices

Weste & Harris – CMOS VLSI Design
Result: A Quadruple-Whammy

Complexity

Heterogeneity

Time-to-Money

DSM Effects

Increasing Device and Context Complexity

- Exponential increase in device complexity—increasing with Moore’s law (or faster)!
- System context in which devices are deployed (e.g. cellular radio) are increasing in complexity as well exponential increases in design productivity

*We have exponentially more transistors!*
Scope of design growing exponentially

Lines of Verilog grows with Moore’s Law

<table>
<thead>
<tr>
<th>GPU</th>
<th>Technology</th>
<th>Transistors</th>
<th>Frequency</th>
<th>Placeable Instances</th>
<th>Flops</th>
<th>Models C vs V</th>
<th>Directed Arch Tests</th>
<th>GDS2 File size</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gen1</td>
<td>0.25u</td>
<td>9M</td>
<td>125MHz</td>
<td>1M</td>
<td>-50K</td>
<td>90K/300K</td>
<td>300</td>
<td>300K</td>
<td>6 W</td>
</tr>
<tr>
<td>Gen2</td>
<td>0.18u</td>
<td>25M</td>
<td>250MHz²</td>
<td>1.5M</td>
<td>-200K</td>
<td>600K/300K</td>
<td>6000</td>
<td>2GB</td>
<td>10 W</td>
</tr>
<tr>
<td>Gen3</td>
<td>0.15u</td>
<td>57M</td>
<td>350MHz²</td>
<td>3M</td>
<td>-500K</td>
<td>600K/800K</td>
<td>25000</td>
<td>4.5 GB</td>
<td>12 W</td>
</tr>
<tr>
<td>Gen4</td>
<td>0.13u</td>
<td>~120M</td>
<td>450MHz²</td>
<td>5.5M</td>
<td>-750K</td>
<td>700K/1.3M</td>
<td>50000</td>
<td>8 GB</td>
<td>15 W</td>
</tr>
</tbody>
</table>

1 - dual edge clocking

Deep Submicron Effects

The design of each transistor is getting more difficult due to physical effects associated with very small (deep submicron) geometries.

Chris Malachowsky - NVidia
Important Deep Submicron Effects

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   4a. Dynamic power consumption
   4b. Static power and leakage
5. Electromigration
6. Variability
7. Reliability
Signal integrity

- Crosstalk/Crosscoupled capacitances can not only cause delay degradation in static CMOS digital circuits – it can cause errors in:
  - Analog circuits
  - Dynamic CMOS circuits
- Solutions as before:
  - Shielding
  - Better analysis
  - Signal encoding

<table>
<thead>
<tr>
<th>Wire Length (mm)</th>
<th>Incremental Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0%</td>
</tr>
<tr>
<td>0.5</td>
<td>20%</td>
</tr>
<tr>
<td>1.0</td>
<td>40%</td>
</tr>
<tr>
<td>1.5</td>
<td>60%</td>
</tr>
<tr>
<td>2.0</td>
<td>80%</td>
</tr>
<tr>
<td>2.5</td>
<td>100%</td>
</tr>
<tr>
<td>3.0</td>
<td>120%</td>
</tr>
</tbody>
</table>

- Same Direction
- Opposite Direction

Important Deep Submicron Effects

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   4a. Dynamic power consumption
   4b. Static power and leakage
5. Electromigration
6. Variability/Reliability
**Power, timing inter-related:**

*Instance IR drop impact on delay*

A. Khan - Cadence

- Buffers get different VDD voltage
- Causes timing closure problems if not accounted for
  - Additional failed paths
  - Race conditions
- Working solution – better analysis tools

---

**Important Deep Submicron Effects**

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   - 4a. Dynamic power consumption
   - 4b. Static power and leakage
5. Electromigration
6. Variability/Reliability
**Power dissipation going forward**

- **Sun’s Surface**
- **Rocket Nozzle**
- **Nuclear Reactor**
- **Hot Plate**

**Power Ingredients**

- **Dynamic Dissipation**
  \[ P_{dyn} = C_L V_{DD} V_{sw} f \]

- **Short-Circuit Currents**
  \[ P_{sc} = V_{DD} I_{sc} \]

- **Static Dissipation**
  \[ P_{stat} = V_{DD} I_{leak} \]
**Dynamic Power Dissipation**

\[ P_{\text{dyn}} = C_L V_{DD} V_{sw} f \]

How do these scale?

- \( C_L \)
- \( V_{DD} \)
- \( V_{sw} \)
- \( f \)

---

**Scaling Relationships for Long Channel Devices**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Relation</th>
<th>Full Scaling</th>
<th>General Scaling</th>
<th>Fixed Voltage Scaling</th>
</tr>
</thead>
<tbody>
<tr>
<td>( W, L, t_{ox} )</td>
<td>( 1/S )</td>
<td>( 1/S )</td>
<td>( 1/S )</td>
<td></td>
</tr>
<tr>
<td>( V_{DD}, V_{T} )</td>
<td>( 1/S )</td>
<td>( 1/U )</td>
<td>( 1 )</td>
<td></td>
</tr>
<tr>
<td>( N_{\text{SUB}} )</td>
<td>( V/W_{\text{dep}} )</td>
<td>( S )</td>
<td>( S^2/U )</td>
<td>( S^2 )</td>
</tr>
<tr>
<td>Area/Device</td>
<td>( WL )</td>
<td>( 1/S^2 )</td>
<td>( 1/S^2 )</td>
<td>( 1/S^2 )</td>
</tr>
<tr>
<td>( C_{\text{ox}} )</td>
<td>( 1/t_{ox} )</td>
<td>( S )</td>
<td>( S )</td>
<td>( S )</td>
</tr>
<tr>
<td>( C_L )</td>
<td>( C_{\text{ox}} W/L )</td>
<td>( 1/S )</td>
<td>( 1/S )</td>
<td>( 1/S )</td>
</tr>
<tr>
<td>( k_w, k_p )</td>
<td>( C_{\text{ox}} W/L )</td>
<td>( S )</td>
<td>( S )</td>
<td>( S )</td>
</tr>
<tr>
<td>( t_{sw} )</td>
<td>( k_{n,p} V^2 )</td>
<td>( 1/S )</td>
<td>( S/U^2 )</td>
<td>( S )</td>
</tr>
<tr>
<td>( t_p ) (intrinsic)</td>
<td>( C_L V / t_{sw} )</td>
<td>( 1/S )</td>
<td>( U/S^2 )</td>
<td>( 1/S^2 )</td>
</tr>
<tr>
<td>( P_{sw} )</td>
<td>( C_L V^2 / t_p )</td>
<td>( 1/S^2 )</td>
<td>( S/U^3 )</td>
<td>( S )</td>
</tr>
<tr>
<td>PDP</td>
<td>( C_L V^2 )</td>
<td>( 1/S^3 )</td>
<td>( 1/SU^2 )</td>
<td>( 1/S )</td>
</tr>
</tbody>
</table>

Table 3.1: Scaling Relationships for Long Channel Devices

\( f \) – scaling faster than \( 1/S \) - why?
**Dynamic Power Dissipation**

- Dynamic Dissipation
  \[ P_{\text{dyn}} = C_L V_{\text{DD}} V_{\text{sw}} f \]

  How do these scale?

  - \( C_L \) - 1/S
  - \( V_{\text{DD}} \) – Full 1/S, Fixed 1, General \( U \)
  - \( f \) - S ASIC, \( >> S \) for uP

  All adds up to higher power density for high-performance designs

---

**Important Deep Submicron Effects**

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
   - 2a. Delay degradation in static CMOS
   - 2b. Signal integrity in Analog and Dynamic CMOS
3. IR drop
4. Power
   - 4a. Dynamic power consumption
   - 4b. Static power and leakage
5. Electromigration
6. Variability/Reliability
Leakage current components

1: pn junc reverse bias
2: weak inversion
3: DIBL – drain induced barrier lowering
4: GIDL – gate induced drain leakage
5: Punchthrough
6: Narrow width effect
7: gate oxide tunneling
8: hot carrier injection

These Slides are derived from Design of High-Performance Microprocessor Circuits, A. Chandrakasan, W. Bowhill, F. Fox, IEEE, 2001

Leakage Current Trends

\[ I_{\text{off}} = I_0 \exp\left(-\frac{qV_t}{mkT}\right) \]

- Leakage current exponentially increases with reduction in threshold voltage \( V_t \)
- Leakage projected to grow to 40% of total power on future microprocessors
**Subthreshold Leakage – Borkar Intel**

![Graph showing subthreshold leakage vs. temperature](image)

**Important Deep Submicron Effects**

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   4a. Dynamic power consumption
   4b. Static power and leakage
5. Electromigration
6. Variability
7. Reliability
**Electromigration (1)**

Limits dc-current to 1 mA/µm

Rabaey, Nikolic

---

**Solution – Materials: Copper**

- With cladding and other effects, Cu ~ 2.2 μΩ·cm vs. 3.5 for Al(Cu) ⇒ 40% reduction in resistance
- Electromigration improvement; 100X longer lifetime (IBM, IEDM97)
  - Electromigration is a limiting factor beyond 0.18 µm if Al is used (HP, IEDM95)

Rabaey, Nikolic
**Important Deep Submicron Effects**

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   4a. Dynamic power consumption
   4b. Static power and leakage
5. Electromigration
6. Variability – covered in earlier lecture
7. Reliability – covered in earlier lecture

**DSM Effects**

- During the golden era of ASIC (3.5u to .18u) there was little concern for physical effects during design.
  - Area
  - Delay
  - Power
- Were all relatively predictable and insurable
- With smaller geometries (.18u and below) physical effects began to demand greater and greater attention in design. Basic design parameters were neither predictable or insurable
- But moving higher in the design hierarchy (RTL → Behavioral → “System-level”) requires building on top of predictable layers
- As a result the design process has been stalled at the RTL level as we try to bring netlist implementation back to the predictability and reliability that we had 20 years ago!
- Focus of research in the RTL design flow has been on managing these DSM effects
Process Variability

- Increased device variation
- Increased interconnect variation
- Result
  - Slower/unpredictable speed of digital circuits
  - Greater uncertainty about parameter values
  - Over-conservatism
  - Reduced signaling robustness
  - Increased clock skew

The march of technology

Is this worth a huge (multi $B$) investment?
Where does it Come from?

- Inter-Die Variation
- Intra-Die Variation

Device Parameter Variations

- W, L variations
  - Due to photolithography proximity effect or etching
  - Layout density dependent
  - Location dependent
- Tox variation
  - Well controlled by a product spec.
- Vth variation
  - Due to doping
### Device Worst Case Model

**Comparison Corner vs Real Data**

- FF: Fast NMOS & Fast PMOS
- SF: Slow NMOS & Fast PMOS
- FS: Slow NMOS & Slow PMOS

3 corner Model: TT, SS, FF
5 corner model: all

- Corners from SPICE document

---

### Within Wafer Channel Length Variation

- The most important source of variation is $L_{gate}$ variation
  - Determines speed of the transistor through effective channel length
Sources of CD Variation

- Among lithography people, Lgate is known as CD – the critical dimension (because it’s so important)

SIA thinks that CD should be allowed a 10% error budget – why?

<table>
<thead>
<tr>
<th>Source</th>
<th>CD Error</th>
<th>Defects</th>
<th>Aberrations</th>
<th>Lens Heating</th>
<th>Focus</th>
<th>Leveling</th>
<th>Dose</th>
<th>Power</th>
<th>Pressure</th>
<th>Flow Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wafer Flatness Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reticule Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stepper Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Etch Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resist Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PEB Reflectivity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 39a Lithography Technology Requirements—Near Term**

<table>
<thead>
<tr>
<th>Year</th>
<th>Technology Node</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRIE</td>
<td>Half pitch (nm)</td>
<td>160</td>
<td>165</td>
<td>170</td>
<td>175</td>
<td>180</td>
<td>185</td>
<td>190</td>
</tr>
<tr>
<td></td>
<td>Contacts (nm)</td>
<td>200</td>
<td>195</td>
<td>190</td>
<td>195</td>
<td>200</td>
<td>205</td>
<td>210</td>
</tr>
<tr>
<td></td>
<td>Overlay (nm mean + 3 sigma)</td>
<td>65</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>CD control (nm, 3 sigma, post-etch)</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>MPU</td>
<td>Half pitch</td>
<td>230</td>
<td>219</td>
<td>218</td>
<td>216</td>
<td>214</td>
<td>210</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>Gate length (nm, in resist)</td>
<td>140</td>
<td>120</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Contact length (nm, post-etch)</td>
<td>140</td>
<td>120</td>
<td>100</td>
<td>90</td>
<td>80</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>Gate CD control (nm, 3 sigma, post-etch)</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

76
Every nm of CD spread reduction means about $10 of more revenue per CPU chip.

Source: ITRS 1999

Decomposition of Sources of Variability

- Reticule error, lens aberration → Intra-field
- Exposure/focus → Die-to-die
- CVD, coating, develop, etching → Across wafer
- Batch to batch variability in materials, long term variability in equipment → Lot-to-lot
- Optical diffraction, proximity effect, micro-loading in etching and development → Pattern dependent
**Why is Intra-field Variation Critical?**

Field size increases by 12%/year to accommodate 59% more components per year by Moore’s Law.

- Spatial variability is mainly systematic instead of random.
- It could be compensated by process optimization or circuit design, or special mask design.

**Lgate Varies Depending on Local Layout Patterns**

- For full characterization, gates classified by their local layout patterns

- Gate are classified by:
  - **A) Distance** to neighboring gate (proximity effect)
  - **B) Left vs right** neighbor position (coma effect)
  - **C) Vertical vs horizontal orientation**
**Spatial Maps for Different Gate Categories**

- Two spatial profiles are statistically different
- Separate models need to be used at the CAD level

![Spatial Maps for Different Gate Categories](image)

**Interconnect Parameter Variations**

- Line width(w), spacing(s)
  - Due to photolithography proximity effect or etching
  - Layout dependent
  - Location dependent
- Metal thickness(T)
  - Due to erosion, dishing
  - Layout density dependent
- Inter-layer Dielectric (ILD) thickness(H)
  - Due to CMP
  - Dielectric Constant(ro)
**All layer Cu/Low-K Interconnect**

Current process technology for interconnect with multiple layers of metal/dielectric

From TSMC 0.13um technology

---

**Planarity of Al Metal CMP Processes**

- Chemical-mechanical polish (CMP) rate is different for sparse and dense areas
- Tiling: adds new features in sparse areas to ensure better planarity
- Design problem: determine location and amount of dummy features needed to achieve a planarity

(Grobman, DAC2001)
Planarity In Copper CMP Processes

- For Cu processes have two problems
  - Oxide erosion
  - Copper ‘dishing’

(Grobman, DAC2001)

Intra-Die ILD Thickness Variation

- Within die variation
**ILD Thickness Variation**

- TSMC specs on ILD Variation
  - Variation is up to 20%
  - Modest (3% ?) impact on timing

<table>
<thead>
<tr>
<th>Dielectric layers</th>
<th>Thickness</th>
<th>% Var</th>
<th>Dielectric constant</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOX</td>
<td>3500</td>
<td>± 17.1%</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>ILD</td>
<td>7000</td>
<td>± 21.4%</td>
<td>4.0</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD1a</td>
<td>11300</td>
<td>± 20%</td>
<td>3.7</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD1b</td>
<td>2000</td>
<td>± 3%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>IMD2a</td>
<td>11300</td>
<td>± 20%</td>
<td>3.7</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD2b</td>
<td>2000</td>
<td>± 3%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>IMD3a</td>
<td>11300</td>
<td>± 20%</td>
<td>3.7</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD3b</td>
<td>2000</td>
<td>± 3%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>IMD4a</td>
<td>11300</td>
<td>± 20%</td>
<td>3.7</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD4b</td>
<td>2000</td>
<td>± 3%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>IMD5a</td>
<td>11300</td>
<td>± 20%</td>
<td>3.7</td>
<td>See NOTE 1.</td>
</tr>
<tr>
<td>IMD5b</td>
<td>2000</td>
<td>± 3%</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>PASS1</td>
<td>10000</td>
<td>± 10%</td>
<td>4.2</td>
<td>See NOTE 2.</td>
</tr>
<tr>
<td>PASS2</td>
<td>7000</td>
<td>± 10%</td>
<td>7.9</td>
<td>Conformal material.</td>
</tr>
</tbody>
</table>

NOTE 1: The dielectric layers of ILD, IMD1a, IMD2a, IMD3a, IMD4a and IMD5a outside the metal are overetched 1000.

---

**Tiling for Better Planarity**

- **Untiled reticle (768A)**
  - (unmanufacturable)

- **Conventional Rule-Based Tiling (702A)**
  - (9% uniformity improvement)

- **Model-Based Tiling (152A)**
  - (80% uniformity improvement)
    - Motorola DSP
      - (Grobman, DAC2001)
**Important Deep Submicron Effects**

1. Rising relative delay of interconnect
2. Cross-coupled capacitance
3. IR drop
4. Power
   - 4a. Dynamic power consumption
   - 4b. Static power and leakage
5. Electromigration
6. Variability
7. Reliability

---

**Reliability**

- **Soft Error FIT/Chip (Logic & Mem)**
- **Extreme device variations**
- **Time dependent device degradation**
- **Burn-in may phase out…? Chip infant mortality?**
SER Mitigation in Microprocessors

Manufacturing Techniques
- SOI technology (e.g., AMD Opteron™ processor) reduces SER compared to bulk
  - Charge generated below BOX cannot be collected
- Eliminate alpha-producing materials from chip environment
  - BPSG
- Low-alpha package materials and underfill
- Low-alpha lead for C4 bumps reduces emissivity by >1000X

Design Techniques
- Parity: Detects SEU upsets
- ECC: Corrects single bit fails; detects dual-bit fails.
  - Reduces SER by several orders of magnitude in the protected arrays.
  - Typically, over 90% of SER-susceptible bit are either parity or ECC-protected in Server processors
- Cache Line Interleaving: Physical separation of logically adjacent bits to greatly reduce multibit fails
- Scrubbing: Cache scrubbing in background to do background correction of SEU
- Hardening: selective nodes made SEU-resistance by device size tweaks

Operational Techniques
- For mission-critical applications, operate at highest voltage possible
  - SER is strongly voltage dependent
- Try to avoid high-altitude operation! In fact, try to operate in deep mines!
  - At airline altitudes, Neutron flux is ~300X higher than sea level