Dependable Design in Nanoscale CMOS Technologies: Challenges and Solutions

Vikas Chandra
ARM R&D
Reliability challenges

- Reasons of unreliable transistors
  - Random manufacturing defects
  - Significant increase in variability
  - Increasing electric field
  - Thin gate oxides
  - Voltage, Temperature variations
  - …
Atomistic scale devices

The simulation Paradigm now

A 22 nm MOSFET
In production 2010

A 4.2 nm MOSFET
In production 2023

Source: A. Asenov
Types of variability

- **Spatial**
  - Variations due to the manufacturing process
  - Systematic, process and apparatus induced variations
  - Random variations

- **Temporal**
  - Mainly due to aging and wearout
  - NBTI
  - Gate oxide degradation

- **Dynamic**
  - Workload dependent
  - Voltage fluctuation
  - Temperature variation
Spatial variations

- Simplified Manufacturing Process Steps

1. Single crystal Si wafer
2. Resist coat
3. Expose
4. Post-exposure bake (PEB)
5. Develop
6. Photolithography
7. Reactive Ion Etch
8. Implant / doping
9. Cu deposit
10. Chemical mechanical polishing (CMP)
The Lithography Challenge: Reducing Feature Size

- Wavelength scaling has stopped!
  - Glass does not transmit
  - Source not bright enough
  - Reticle/mask too expensive to manufacture

- Deep sub-wavelength lithography
  - Finer lines than the point of a pen!

---

Data: Tim Brunner, IBM

Source: Stephen Renwick, Nikon
Lithography Variability

- Several sources of variation in lithography
  - Defocus variation
  - Exposure dose (intensity) variation
  - Mask errors
  - Overlay/mask alignment variation
Etch Variability

- Etching process has randomness
  - Poisson process for ions hitting the resist
  - Plasma gas flow can have turbulence

- Etch chuck temperature profile is radial – etch rate profile is radial

- Typically CD (linewidth) droops near wafer edge

Source, A. Singhee, IBM
Material removal depends on wire density and width
Surface topography changes across the die with Copper density
Wire resistance and capacitance variation
Focus error for upper metal layers – wire width errors
Random Dopant Fluctuation

- Doping/implant is a random process
- Number of dopants in channel ~100
- Dopant count is not repeatable
- Dopant position is not repeatable
- Large variations in threshold voltage
  \[ \sigma_{V_t} = \left( \frac{4\sqrt{4q^3 \varepsilon_{Si} \phi_B}}{2} \right) \cdot \frac{T_{ox}}{\varepsilon_{ox}} \cdot \frac{4\sqrt{N}}{\sqrt{W_{eff}L_{eff}}} \propto \frac{1}{\sqrt{W_{eff}L_{eff}}} \]
  - ~10-15% \( \sigma(V_t) \) at 45 nm and increasing
    - Typical \( \pm 3\sigma \) tolerance range \( >= \pm 30\% \)!

M. Hane, et. al., SISPAD 2003
Variability Challenges For Design: ITRS 2007

- Lots of RED ahead

- Economics of purely process solution are infeasible
  - Mask cost today up to $100,000
  - Litho tool cost today ~$50,000,000

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized mask cost from public and IDM data</td>
<td>1.0</td>
<td>1.3</td>
<td>1.7</td>
<td>2.3</td>
<td>3.0</td>
<td>3.9</td>
<td>5.1</td>
<td>6.6</td>
<td>8.7</td>
</tr>
<tr>
<td>% $V_{dd}$ variability: % variability seen in on-chip circuits</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>% $V_{dd}$ variability: doping variability impact on $V_{dd}$ (minimum size devices, memory)</td>
<td>31%</td>
<td>35%</td>
<td>40%</td>
<td>40%</td>
<td>40%</td>
<td>58%</td>
<td>58%</td>
<td>81%</td>
<td>81%</td>
</tr>
<tr>
<td>% $V_{dd}$ variability: includes all sources</td>
<td>33%</td>
<td>37%</td>
<td>42%</td>
<td>42%</td>
<td>58%</td>
<td>58%</td>
<td>81%</td>
<td>81%</td>
<td>81%</td>
</tr>
<tr>
<td>% $V_{dd}$ variability: typical size logic devices, all sources</td>
<td>16%</td>
<td>18%</td>
<td>20%</td>
<td>20%</td>
<td>26%</td>
<td>26%</td>
<td>36%</td>
<td>36%</td>
<td>36%</td>
</tr>
<tr>
<td>% CD variability</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>% circuit performance variability circuit comprising gates and wires</td>
<td>46%</td>
<td>48%</td>
<td>49%</td>
<td>51%</td>
<td>60%</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
<td>63%</td>
</tr>
<tr>
<td>% circuit total power variability circuit comprising gates and wires</td>
<td>56%</td>
<td>57%</td>
<td>63%</td>
<td>68%</td>
<td>72%</td>
<td>76%</td>
<td>80%</td>
<td>84%</td>
<td>88%</td>
</tr>
<tr>
<td>% circuit leakage power variability circuit comprising gates and wires</td>
<td>124%</td>
<td>143%</td>
<td>186%</td>
<td>229%</td>
<td>255%</td>
<td>281%</td>
<td>287%</td>
<td>294%</td>
<td>331%</td>
</tr>
</tbody>
</table>

- Need more process and variability-aware design
Temporal variations

- Infant mortality: *Increasing manufacturing defects*
- Normal lifetime: *Increasing transient errors*
- Wearout: *Acceleration of aging phenomena*
Temporal unreliability

- Infant mortality
  - Marginal parts due to random manufacturing defects
  - Gate-to-source shorts
  - Small opens, poor vias & contacts
  - Mitigated by Burn-in

- Normal Lifetime
  - Soft errors in memory and logic
  - Mitigated by design, architecture and ECC

- Wearout
  - Transistor degradation (NBTI)
  - Gate oxide breakdown (GBD)
  - Mitigated by circuit, architecture techniques and overdesign
Infant mortality

- Also known as Early Life Failures (ELF)
  - Do not affect the circuit initially, but they get worse over time

- Due to manufacturing defects that are random in nature
  - Particles in interlevel oxide creating shorts between metal layers
  - Insulator cracks
  - Thin oxide defects
  - Metallization problems
  - Via defects
  - …

- ELFs follow log-normal failure distribution
  - Short mean lifetime and high sigma
  - Failure rate decreases over time
Burn-in testing

- Burn-in is stress testing for weeding out ELF defects
  - “Age” the circuits just beyond the infant mortality period
  - Weak (defective) parts break due to accelerated aging
  - Employs voltage and temperature to accelerate device aging

- Stress conditions
  - **Voltage stress:** Typically 30-40% over nominal \( V_{dd} \)
  - **Temperature stress:** Typically \( >120^\circ \text{C} \)
  - **Stress time:** Typically 10’s of hours
    - Decreases as failure rate decreases
Temperature and Voltage stress

- Temperature acceleration factor
  \[ TAF = e^{\frac{E_a}{K} \left( \frac{1}{T_{stress}} - \frac{1}{T_{use}} \right)} \]

- Voltage acceleration factor
  \[ VAF = H e^{\gamma \left(V_{stress} - V_{use} \right)} \]

- TAF targets: electromigration, metallization problems, contact/via defects etc

- VAF targets: gate oxide defects
VAF and TAF trends

- Supply voltage is saturated

  - $\Delta V = V_{\text{stress}} - V_{\text{use}}$
    - 40% of 3.3V $\rightarrow$ 1.32V
    - 40% of 1V $\rightarrow$ 0.4V

- VAF goes down exponentially

- On chip temperature is going up

- TAF goes down exponentially

- Burn-in testing running out of steam?
Normal lifetime unreliability (Soft errors)

- Mechanism of soft errors due to high energy particles

Particle strike creates hole electron pairs

Source: Ziegler, et al., IBM J. of R&D, 1996
Source: R. Baumann, IEEE TDMR, 2001

Source: P. Roche, ST, IRPS 2006
Impact on storage logic

- Particle strike flips the stored value
- The flipped value stays due to regenerative feedback
- Corrupts the state of the system

6T bit cell

Latch
Impact on combinational logic

- Causes glitch at gate outputs
- Can be latched if transition happens during latching window
  - Can result in timing failure
  - Errors can be masked by electrical and logical masking
- Decreasing cycle time exacerbates this problem

![Diagram showing latching window and timing failure](image-url)
Soft error trends

- Substantial increase in soft error susceptibility with technology scaling!

Source: R. Baumann, TI, SemaTech 2004
Wearout - NBTI basics

- NBTI stands for Negative Bias Temperature Instability
  - Degradation in PMOS performance over device lifetime
  - Due to traps at Si-SiO₂ interface
  - Instability refers to gradual shift in transistor parameters with time

- Impact on transistor performance
  - $V_t$
  - $I_{ds}$, $g_m$, $I_{off}$

- Temporal behavior of NBTI induced aging

NBTI stands for Negative Bias Temperature Instability. Degradation in PMOS performance over device lifetime due to traps at Si-SiO₂ interface. Instability refers to gradual shift in transistor parameters with time.

- Impact on transistor performance
  - $V_t$ (up)
  - $I_{ds}$, $g_m$, $I_{off}$ (down)

- Temporal behavior of NBTI induced aging

Rapid increase or $|ΔI_{ds}|$ (slower rate)
NBTI: Degradation – Recovery

<table>
<thead>
<tr>
<th>Silicon</th>
<th>Gate oxide</th>
<th>Poly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si</td>
<td>*</td>
<td>H</td>
</tr>
<tr>
<td>Si</td>
<td>H</td>
<td>H₂</td>
</tr>
<tr>
<td>Si</td>
<td>*</td>
<td>H</td>
</tr>
<tr>
<td>Si</td>
<td>H</td>
<td></td>
</tr>
</tbody>
</table>

Negative Bias: Si-H bond disassociation
Zero Bias: Si-H bond recovery

- $V_{DD}$
- $V_{DD}$ or 0

Stress stage
Recovery stage

$\Delta V_t$

Si-H bond disassociation
Si-H bond recovery
Impact on logic circuits

- Temporal $V_t$ shift in PMOS affects critical performance metrics

- Combinational circuits
  - $F_{\text{max}}$ decreases $\downarrow$
  - Timing failure as circuits age

- Storage cells (SRAM, latch)
  - Static Noise Margin $\downarrow$
  - Read and write stability $\downarrow$
  - Parametric yield loss
Circuit degradation

- Average degradation of ~8% in 3 years
- Degradation more dominant for PMOS dominated designs
- Complex circuits seem to degrade less

Source: K. Kang, IRPS, 2007
Gate oxide scaling trend

To reduce power, Vdd is scaled
- $t_{ox}$ is reduced to reduce $V_t$
- Performance increases, as well as leakage

$t_{ox}$ scaling has hit a plateau
- Leakage, reliability…

Source: Nature, June 1999

Source: Intel, 2005
Gate oxide degradation

- Traps start to form in the Gate Oxide
  - Non overlapping
  - Do not conduct

- Hard Breakdown
  - Silicon in the breakdown spots melts
  - Oxygen is released
  - Silicon Filament is formed from Gate to Substrate (Hard Breakdown)

- As more and more traps are created
  - Traps start to overlap
  - Conduction Path is created
  - Soft breakdown (SBD)

- Thermal Damage
  - Conduction leads to heat
  - Heat leads to thermal damage
  - Thermal Damage leads to Traps

- Silicon in the breakdown spots melts
- Oxygen is released
- Silicon Filament is formed from Gate to Substrate (Hard Breakdown)
Temporal oxide degradation

- Gate leakage fluctuates as the gate oxide degrades

Source: H. Wang et al, IEEE TDMR, 2007
Design Characteristic – Digital logic

- CMOS logic inherently acts as noise rejecter
Design Characteristic – Digital logic

- Ring oscillators

41 stage ring oscillator
  - Leakage current goes up after successive breakdowns
  - Still functional after multiple breakdowns
  - Oscillation frequency slows down

Source: B. Kaczer, Trans on Electron Devices, Mar 2002
Dynamic variations: Temperature

- Thermal map – 1.5 GHz Itanium map

[Source: Intel Corporation and Prof. V. Oklobdzija]
Dynamic variations: Voltage, Power

Voltage variations

Power variations

Source: D. Hathaway, SLIP 2005

Source: Naftziger et al, JSSC 2006
Design with margins

- Variability leads to margins

Uncertainly leads to overheads in performance and power
- Increasing intra- and inter-chip variation with process scaling
- Sources: lithography, manufacturing (dopant fluctuation, pattern density effects), crosstalk noise, temperature variation, aging…

Worst-case scenarios are highly improbable
- Significant gain for circuits optimized for the common case
Adaptive designs

- Reduce guardbands due to variations
  - Spatial, temporal and dynamic

- Respond to variations by dynamic adaptation

- Three components required for adaptability
  - Failure prediction
  - Failure detection
  - Failure recovery
Failure prediction

- Predict the errors before they affect design functionality
  - More applicable to slow changing variations

- Adapt by changing frequency and/or voltage

- Possible ways to detects errors
  - Canary circuits: These circuits fail before the actual design fails
  - Pre-sampling: Sample the same data at different points in time
  - Aging monitor: Detect a transition in a guardband period
Failure prediction: Canary circuits

- SRAM example for choosing minimum Data Retention Voltage (DRV)
  - Use replica bitcells (canary bitcells) inspired by canary birds
  - Use Canary bitcells in closed-loop VDD scaling

Source: J. Wang et al, CICC 2007
Failure prediction: Pre-sampling

- Key features of AVERA cell
  - Scan circuit re-used for error checking and analysis
  - Circuit timing degradation detected by pre-sampling LA-LB
  - C-element for error correction

Source: M. Zhang, IOLTS '07
Failure prediction: Aging detector

Detect transitions during $T_g$

Flip-Flop with Aging-resistant Built-in Aging Sensor

Flip-Flop with Aging-resistant Built-in Aging Sensor

Source: Agarwal, Mitra et al, VTS '07
Failure detection

- Detect errors which affect functionality
  - Fast changing errors
    - Soft errors, transient errors due to voltage glitch etc.
  - Slow changing errors
    - Aging induce timing errors
    - Temperature induce timing errors

- Failure detection methods
  - Software
  - Redundancy
  - Coding
  - Path-level delay fault detection
  - …
Failure detection

- Error detection by double sampling

Source: D. Ernst et al, Micro, 2003
Transient faults such as SEU manifest themselves as voltage pulses.

Temporal redundancy (sampling at 2 points in time) detects such an event.
- Error is flagged when the delayed sample does not agree with the first sample.

The error signal can be used for recovery.

Source: Anghel & Nicolaidis '01
Transient error mitigation

- Add redundancy to detect and correct transient errors (e.g. BISER FF)

<table>
<thead>
<tr>
<th>A B</th>
<th>00</th>
<th>11</th>
<th>01</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-element (A, B)</td>
<td>1</td>
<td>0</td>
<td>Previous value retained</td>
<td>Previous value retained</td>
</tr>
</tbody>
</table>

Source: S. Mitra, Stanford
Failure recovery

- Local recovery
  - Inject correct value into pipeline
  - Stall for one cycle and continue

- Instruction replay
  - Invalidate instructions in pipeline
  - Re-execute from failing instruction

- Checkpointing with roll-back
  - Periodically, save system state in memory
  - On error, roll back to last saved state

Source: Jim Tschanz, Intel
Failure recovery

- **Razor: Local error detection and correction on the fly**
  - Upon failure: Overwrite main flip-flop with correct data from the shadow latch
  - Ensure that the shadow latch is always correct by conventional design

Source: S. Das et al, JSSC 2006
Failure recovery

- Error correction by instruction replay

Source: K. Bowman, ISSCC 2008
Energy-error tradeoff

- Adaptive designs have much lower $V_{opt}$ than worse case designs
- Or alternatively, adaptive designs can run much faster at the same voltage

Source: K. Bowman, ISSCC 2008

D. Ernst et al, IEEE Computers 2004

Nominal: $V_{CC} = 1.2V$ & Temp = 60°C
Worst-Case: 10% $V_{CC}$ Droop & Temp = 110°C

Source: K. Bowman, ISSCC 2008
Conclusions

- Variations are becoming dominant with technology scaling
  - Spatial variations
  - Temporal variations
  - Dynamic variations

- Designing with margins is not a sustainable proposition
  - Too much power, performance overhead

- Resilient designs are needed which can adapt to variations
  - Three components required for adaptability
    - Failure prediction
    - Failure detection
    - Failure recovery
Fin