Variability Sources

- **Physical**
  - Changes in characteristics of devices and wires.
  - Caused by IC manufacturing process & wear-out (electro-migration).
  - Time scale: $10^9$ sec (years).

- **Environmental**
  - Changes in VDD, Temperature, local coupling.
  - Caused by the specifics of the design implementation.
  - Time scale: $10^{-6}$ to $10^{-9}$ sec (clock tick).
Process Variations

- Control of minimum features does not track feature scaling
  - Relative device/interconnect variations increase
- Sources:
  - Random dopant fluctuations
  - Feature size, oxide thickness variations
- Effects:
  - Speed
  - Power, primary leakage
  - Yield
Increasing Process Variations

- Increase in variation of process parameters with scaling
- Worst-case design getting more expensive
- “Better than worst-case” design must be error tolerant

Percentage of total variation accounted for by within-die variation (device and interconnect)

Variability in sub 100nm Technologies

- Higher fractional (%) variability with finer design rules and larger wafers (Table source ITRS)

<table>
<thead>
<tr>
<th>L (nm)</th>
<th>250</th>
<th>180</th>
<th>130</th>
<th>90</th>
<th>65</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vt (mV)</td>
<td>450</td>
<td>400</td>
<td>330</td>
<td>300</td>
<td>280</td>
<td>200</td>
</tr>
<tr>
<td>σ-Vt (mV)</td>
<td>21</td>
<td>23</td>
<td>27</td>
<td>28</td>
<td>30</td>
<td>32</td>
</tr>
<tr>
<td>σ-Vt/Vt</td>
<td>4.7%</td>
<td>5.8%</td>
<td>8.2%</td>
<td>9.3%</td>
<td>10.7%</td>
<td>16%</td>
</tr>
</tbody>
</table>

- Lower voltages – less head room
- High speed and RF functions in CMOS
- Design time and risk for analog sections starting to exceed digital sections
- Pressure to ramp products to volume and early profitability
- Mask costs $1M at 90nm, $2M at 65nm
Vt Distribution

- High Freq
- High Isb

- Low Freq
- Low Isb

Sources of Variations

- Random Dopant Fluctuations

- Sub-wavelength Lithography
Is Variability for Real?
Variability $\neq$ Statistics
Most of the variability is Systematic
The key question is how much is known at design time
Most of variability is NOT getting worse each generation

Achieving Sub-wavelength Resolution

Design | Mask | Wafer
---|---|---
250nm | OPC | PSM
180nm | 0° | 190°
90nm and Below | 0° | OPC
**Variation Components (Wafer)**

- Global variation comes from Fab, Lot, Wafer processes
- Linear variation is due to materials and gas flow
- Radial variation is due to thermal and spin processes
- Wafer level variation is the sum of global, linear, and radial variation
- Affects mainly single-ended circuit performance measures like switching speed, gain, dynamic power etc.

**Variation Components (Reticle, Local)**

- Reticle variation is due to optical processes
- Local variation comes from totally random microscopic processes. It affects mainly differential circuit performance measures like differential amplifier offset voltage, current mirrors, DACs, etc. Becoming important for digital design (30% of global variation at 130nm).
Parametric Variation Components - Summary

- **Global** variation between chips: Includes Fab, Lot, Wafer, and Reticle level variations. Captured by case files for digital design.
- **Local** variation: Truly random variation between common centroid devices with identical layout. The magnitude is a function of geometry. Critical for analog matching.
- **Position** dependent variation: Due to across chip gradients. Important for large analog blocks, global clocks, and long critical paths.
- **Proximity** and orientation dependent variation: Systematic and deterministic rather than random. Can be modeled, extracted, simulated, reduced by process changes, design techniques, and mask compensation.
- Position and proximity dependent variation needs to be folded into global and local variation for pre-layout simulation.
- **Age** dependent degradation.
- Static and time dependent thermal, $V_{dd}$, and Xtalk variation.

Environemntal Variations: Power Supply

VDD Noise ≈ $\mu I_t R_g + \mu L + R_g^2 C_d (1 - e^{-V_t})$ ~ Same
Environmental variations: Thermal

Temperature varies within the chip

- Power 4 Server Chip: 2 CPU on a chip
  - The CPUs can be much hotter than the caches

Causes Larger Frequency Distribution

Courtesy Intel
**Frequency & SD Leakage**

![Graph showing normalized frequency and leakage]

- Low Freq: Low Isb
- High Freq: Medium Isb
- High Freq: High Isb

**Variation-tolerant Design**

- Balance power & frequency with variation tolerance
- Transistor size: small, large
- Logic depth: large, small
- Low-Vt usage: low, high

![Charts showing balance between power, frequency, and critical paths]

- # uArch critical paths: less, more
**Approaches**

- **Worst-case design**
  - Leaves too many crumbs on the table. Huge concurrency overhead for performance.
- **Regular design strategies to reduce variation**
- **Careful choice of logic styles**
- **Self-adapting design.**
  - Turns on-line knobs (Vdd, Vt) to guarantee operation of the design. Uses one-time correction for systematic errors
- **Alternative Timing Approaches**
  - Self-timed or clockless design
    - Defers the decisions to the system level. Comes with large overhead
  - Pseudo-synchronous design
    (e.g. Razor)
    - Allows for occasional timing errors. Limited operation range.

**Problem:**  
**Predictability \( \approx (\text{Chip Variability})^{-1} \)**

- Std library abstractions break: don’t “hide” the details anymore, as we scale down
- Correlated random variations hit ckt level
- Demise of context-free layout design rules
- Local printability problems
- Global effects
- Cu thickness distribution

30 March 2005 Slide 20
Yesterday’s Freelance Layout

No layout restrictions

Transistor Orientation Restrictions

Transistor orientation restricted to improve manufacturing control
Transistor Width Quantization
“Fabrics” Idea: Atomic Regularity
(Make the Variability Small... Everywhere)

- Starting from basic manufacturing shapes \(\rightarrow\) circuits \(\rightarrow\) logic \(\rightarrow\) routing
  everything is extremely regular
- Means radical re-architecting of flows
- How much predictability? At what cost?
- Initial motivation was “what’s after ASICs”, now more generally aimed at “predictability”

Today’s designs

Tomorrow’s designs

Regular/Structured Integrated System

- Regular Circuits
- Regular Geometry Fabric

Regular Fabrics – A Plethora of Choices

Trade-off between area, performance, power and time-to-market (factors 5 to 10)

VPGA CMU

River PLA Berkeley

Structured ASIC (e.g. LSI RapidChip)
Fabric Architectures: Via Patterned Gate Array

- Configurable with 4 masks for top vias
  - Base architecture can be like an FPGA – but replace expensive switches with mask-config vias

- Many possible interconnect options: std cell routing, or fully regular top-level patterned routing

<table>
<thead>
<tr>
<th>Network switch (80k Gates)</th>
<th>Area (um²)</th>
<th>Slack (ns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard ASIC flow</td>
<td>1752048</td>
<td>-2.521</td>
</tr>
<tr>
<td>Regular Logic VGA flow</td>
<td>1960000</td>
<td>-2.982</td>
</tr>
</tbody>
</table>

- Array offers fully predictable geom. patterning

Fabric Analysis: Enhanced Manufacturability for Regular Ckt Fabrics

- Reduced CMP effects
  - Copper dishing < 40Å
  - Post-CMP Copper thickness variation is less than 2.3%
  - Highly promising as a manufacturable 'logic' replacement structure
Fabric-level Custom Circuit Design: Limited-Switch Dynamic Logic (LDSL)

- Merges latch with every output stage
- Speed of domino – with less power
  
  ![Diagram of domino and LSDL comparison]
  
  Domino
  LSDL [Montoye 03]

- Experiment: 16-bit Kogge-Stone adders, full domino vs LDSL
  - 58 extracted 0.18um fab run models
  - Monte-Carlo for chip-to-chip & mismatch

- LDSL: good for size & speed
  - ~20% less area; ~2X faster

- LDSL: good for regularity, var tolerance
  - Cells more regular in content and size.
  - Less variation in pattern density.
  - More tolerant of manufacturing variation

Fabric-Level Flow Design: “Regularizing” Cell-Based Flows

Complementary approach:

- Regularize a library-based flow
  - Every cell is identical—except for vias
  - Regularity issues handled by cell layout generators, exclusively
  - Extends lifetime of existing flows

- Looking at impacts on performance
  - Ex: granularity of available cell sizes (i.e. library size) is reduced

[Sechen, Washington]
30 March 2005 Slide 33
**RapidChip® Platform ASIC**

- **Configurable Platform**
  - Families of pre-manufactured slices
  - Sea of transistors for high density, high performance user-configurable logic
  - Up to 5 layers of metal personalization
  - Flexible approach to IP:
    - Diffused only when performance dictates, eg high speed SerDes
    - On-demand for most other IPs, eg processors
  - Rich portfolio of soft IPs available

---

**Today’s Reconfigurable FPGA Platform**

- High-speed 3.125 Gbps Serial Transceivers
- >500 DSP datapaths
- 10 Million gates
- PowerPC™ Processor 400+ MHz
- Programmable IO
- 10Mbit Dual-Port™ RAM
Delay and Power Variability in CMOS

Goal: Investigate the effects of variations in $V_{th}$, $L_{poly}$, $W$, $t_{ox}$ and $V_{dd}$ on the performance of a family of representative circuits.

- Quantify the statistical variability of circuit delay and power (active).
- Identify single parameter contributions to overall variability levels.

Circuits under study:
- NAND chain (six stages)
- Adders (16-bits, various architectures)
- Logic styles: Static, Dynamic Domino, Passgate
- All transistor sizes optimized for minimum delay under an area constraint

Experimental Setup:
- 90nm, pd-SOI technology
- Industrial research site
- All parameter distributions set by predictive BSIMSOI models, ITRS (2003)

Monte Carlo Simulation I

Goal I: Vary all parameters simultaneously; study the statistical variability of power and delay.

Variable parameters:
- $V_{th}$, $L_{poly}$, $W$, $t_{ox}$, $V_{dd}$: 1V (mean value)
- Temperature held at 85°C
- Interdependencies between parameters reconciled within the simulation

- $N = 200$ for adders, $N = 1000$ for NANDs

The spatial correlation coefficient defines parameter matching between adjacent transistors
- Each parameter is assigned identically to all transistors within each circuit instance
- $\rho$ is set to 1, indicates perfect correlation (worst-case)
The operating value of $V_{th}$ is composed of its long channel $V_{th0}$ value modified by $\Delta V_{th}$ factors (BSIMSOI Model):

$$V_{th,\text{OPERATING}} \approx V_{th0}(N_{\text{channel}}, \Phi_M, \Phi_S, I_{ox}) + \Delta V_{th,\text{HALO}}(V_{th}) - \Delta V_{th,\text{DIBL}}(V_{ds}, L) + \Delta V_{th,\text{BIAS}}(N_{halo}, L) + \Delta V_{th,\text{NarrowWidth}}(W)$$

Interdependencies between parameters are reconciled within each simulation by separating $V_{th,\text{OPERATING}}$ into independent and dependent components.

Monte Carlo Simulation II

Goal II: Isolate individual parameter contributions to overall power/delay variability

Parameter distributions same as in previous setup

Again, perfect spatial correlation of parameters is assumed ($\rho = 1$)
NAND Chains (6-stages)

- Static capacitive load, $C_L = 10\text{fF}$
- Active, FO3 load (value varies with parameter fluctuations)

Adders

- Ripple carry with Manchester carry chain (passgate-based)
- Carry select, logarithmic configuration

Static CMOS

Static Passgate (LEAP)

Pulsed Static

Dynamic Domino

Static capacitive load, $C_L = 10\text{fF}$

Active, FO3 load (value varies with parameter fluctuations)

Ripple carry with Manchester carry chain (passgate-based)

Carry select, logarithmic configuration

Static

Dynamic

Bit level $C_{out}$ and Sum selection
Block level $C_{out}$ selection
Cout generation

Static, Dynamic Domino, Passgate
Adders: CLA Trees

- **Kogge Stone, Radix 2**
- **Kogge Stone, Radix 4**
- **Brent-Kung**
  - Large intermediate load capacitance along critical path (Sum07 node)
- **Han-Carlson**

Delay, Power Variability: NAND chains

- The static CMOS implementation is the most robust to process parameter variations
- The passgate style (LEAP) displays the highest levels of delay and power variability (30% higher than static)
Delay Variability: Adders

- Static carry select is the most robust
- The three most variable are passgate-based, between 31% - 67% more spread than static carry select

Power Variability: Adders

- Most robust: static ripple with Manchester carry chain
- The least robust: designs with large/irregular intermediate load capacitance along critical paths (radix 4 Kogge Stone, Brent Kung)
Single Parameter Breakdown: NAND Chains

- Results vary depending on final loading stage (static vs. FO3)
- \( V_{th} \) is most significant contributor in all cases
- For active, FO3 loads:
  - Passgate design is most sensitive to \( V_{th} \) variations
  - Increased significance of \( L \) variations

Single Parameter Breakdown: Adders (Delay)

- \( V_{th} \) is most significant contributor (33% average)
- Passgate designs are the most sensitive to \( V_{th} \) variations
- \( L \) is nearly as significant (28% average)
Conclusions

- Static CMOS implementations are generally the most robust to parameter variations, for both delay and power.
- Passgate designs display the least amount of robustness:
  - Suffer spreads in delay and power variability between 30% – 70% higher than static designs.
  - Tend to display highest sensitivity to $V_{th}$ variations.
  - These are worst-case results, due to the assumption of perfect parameter correlation.
- $V_{th}$ variations account for 35% - 40% of delay variability.
- Power variability trends suggest a dependence upon large or irregular intermediate load capacitances.
- $V_{dd}$, L, and $V_{th}$ are consistently the highest contributors to both delay (85%) and power (80%) variation.
A Self-adapting Approach

**Motivation:** Most timing variations are systematic, and can be adjusted for at start-up time using one-time calibration!

- Relevant parameters: $T_{\text{clock}}$, $V_{\text{dd}}$, $V_{\text{th}}$
- $V_{\text{th}}$ control — the most effective and efficient at low voltages
- Can be easily extended to include leakage-reduction and power-down in standby

- Achieves the maximum power saving under technology limit
- Inherently improves the robustness of design timing
- Minimum design overhead required over the traditional design methodology

---

**$V_{\text{th}}$ Tuning via Body Bias**

- Less design cost than $V_{\text{dd}}$ tuning
- $V_{\text{th}}$ tunable range: >150mV for a 90nm Technology
Power and Timing Tradeoffs

V_{th} tuning can effectively gain performance back

Adaptive Body Bias--Experiment

Die frequency: Min(F_{1}..F_{21})
Die power: Sum(P_{1}..P_{21})
Adaptive Body Bias--Results

- Adaptation based on variations
- Yield increase with ABB
- 97% highest bin with ABB for within-die variability

Adaptive Approach for Dealing with Variations

- Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV)

Source: Sam Naffziger, HP