Impact of Process Variations on Low Power Cache Design

Mahmoud Bennaser and Csaba Andras Moritz
Department of Electrical and Computer Engineering
University of Massachusetts, Amherst
February 3, 2006
Introduction

- Process variations increase as the feature reduces due to the difficulty of fabricating small structures consistently across a die or wafer.
- In order to analyze the delay and power consumption of a cache under process variation, we must consider both inter-die and intra-die variation
  - **Intra-die variations** are the variations in device parameters within a single chip, which means different devices at different locations on a single die may have different device features
  - **Inter-die variations** are the variations that occur from one die to the other, from wafer to wafer, and from wafer lot to wafer lot

- Two main sources of variation:
  - Physical factors
  - Environmental factors
Introduction

- The physical factors are permanent and result from limitations in the fabrication process
  - **Effective Channel Length** (Geometric Variations):
    - Imperfections in photolithography
    - Variations in $L_{eff}$ can be as high as 50% within a die
  - **Threshold Voltage** (Electrical Parameter Variation):
    - Variation in device geometry
    - Variations in $V_{th}$ can be modeled as 10% of $V_{th}$ of the smallest device in a given technology [A. Chandrakasan et al., IEEE press 2001]

- The environmental factors depend on the operation of the system and include variations in:
  - **Temperature, Power Supply, Switching Activity**
Impact of Process Variations on Caches

- The parameter variations are random in nature and are expected to be more pronounced in minimum geometry transistors commonly used in memories.
  - Caches in processors like UltraSPARC III, Itanium 2, StrongARM110, and Alpha 21164 can occupy more than 50% of die area.

- Process variations impact the components of a memory subsystem:
  - SRAM Cell
  - Sense Amplifier
  - Address Decoder

- Can cause failure in data access
  - E.g., due to incorrect sensing or slow cell access
Effect of Process Variations on Delay Accessing 1-bit in SRAM Column of 32 Bit Height

The delay can increase as such as 16% per cell.

The Threshold voltage (Vth) variation can impact the delay by 30% per cell access.
Worst-case Delay

The delay can increase as such as 50% combining the effects of Vth and Leff.
Effect of Process Variations on Power Consumption of 1KB SRAM

A small variation in the $L_{eff}$ value causes a change in the leakage power by as such as 40X from the nominal value.

The Threshold voltage ($V_{th}$) variation can impact the power consumption by 65X

[HSPICE simulation]
Cache Access Failure?

- A failure in a cell can occur due to:
  - Access Time Failure (due to increase in the access time)
  - Read Stability Failure
  - Write Stability Failure
  - Hold Failure

- Failure Probability of a Read
  - E.g., the minimum differential voltage required for correct sensing (Taccess in figure) needs to be < Tmax for a correct read
    - Threshold voltage distributions are approximated as Gaussian

[Classification is take from S. Mukhopadhyay, et al. Symposium on VLSI Circuits, June 2004 ]
Failure in Sense Amplifiers

- Circuits like differential sense amplifiers are affected
  - Changing offset voltage may lead to erroneous behavior (e.g., due to access Transistors MN3 and MN3B).
What About Application Performance?

- To account for the worst case scenario we might need to increase the cache access time.
- Performance impact as much as 30-40% in the example on the left.

![Performance Of four instructions issue machine](chart.png)

SPEC2000 Benchmarks

- bzip
- mcf
- gcc
- vpr
- ammp00
- art
- equake
Possible Architectural Directions

- How do we design caches that work in face of these problems?
- We can select a cache design using worst case assumptions
  - ALL VARIATIONS and ALL COMPONENTS on the critical path

- Alternatively, we need to design circuits and architectures that would work *adaptively* depending on actual delay
  - Process variation resilient design
  - Resilience against delays in different parts of the cache
Our Choice: An Adaptive Process Resilient Cache Architecture

- Two phases of operation: classifying and execution
- Classifying phase
  - The cache is equipped with a built-in-self-test (BIST) to detect speed difference due to process variation.
  - Each cache line is tested using BIST when the test mode signal is on. A block is considered fast, medium, or slow (this is for the sake of an example).
An Adaptive Process Resilient Cache Architecture

- Since the speed of the accessed cells (cache lines) changes depending on operating condition (e.g., supply voltage, frequency), such tests are conducted whenever there is a change in operating condition.
  - BIST feeds this information into the delay storage.

- Execution phase
  - The speed information stored in the delay storage is used to control sense amplifiers during regular operations of the circuit.
Circuit Level Support: Double Sensing

- We need a mechanism to avoid sensing prematurely.
- The basic idea of double sensing is to have parallel sense amplifiers to sample the bitline twice during a read cycle. This is required in an adaptive cache design with different cache line latencies.
- The first sensing is performed as the conventional one. The second sensing is delayed and has to be fired as late as required.

[K. Roy, et al. VLSI Test Symposium, May, 2005]
Preliminary Results

Baseline: 3 cycle D-cache. Out of order issue. Adaptive caching scheme: e.g.,
3% 3 cycle, 12% 2 cycle. 85% 1 cycle cache line access. Results below show performance is improved by 13% to 29%!

![Performance of Four Instructions Issue Machine](chart.png)

- **Baseline**
- **Adaptive**

**SPEC2000 Benchmarks**

- mcf
- parser
- vpr
- ammp
- art
- equake

**IPCs**

- 0
- 0.5
- 1
- 1.5
- 2
- 2.5
Conclusion

- Parameter variations will become worse with technology scaling.
- Robust variation tolerant circuits and architectures needed.
- We have shown that process variation can have a significant impact on delay (expected > 2X with all factors included), and in worst-case leads to timing violations.
- In addition, power dissipation, especially leakage power has been shown to be significantly affected (>60X) by the parameter variations.
- Shown new resilient cache architecture