VIFI-CMP:
VARIABILITY-TOLERANT CHIP-MULTIPROCESSORS FOR THROUGHPUT & POWER

WAN-YU LEE
IRIS H.-R. JIANG
GLSVLSI-09
IRIS Lab @ National Chiao Tung Univ.
Outline

- Introduction
- Chip-multiprocessor architectures
- Models
- Monte Carlo analysis
- Experimental results
- Conclusion
Introduction
Process Variability

- **Cause: Transistor level**
  - Unexpected deviations on
    - Threshold voltage $V_{th}$
    - Effective channel length $L_{eff}$

- **Effect: Architectural level**
  - Unwanted degradation on
    - Throughput
    - Power

- **Tackle process variability at high level instead of low level to gain global benefits**

<table>
<thead>
<tr>
<th>Abstraction level</th>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architectural level</td>
<td>Design space exploration</td>
<td>Global benefit</td>
</tr>
<tr>
<td>Layout level</td>
<td>Post-layout optimization</td>
<td>Local refinement</td>
</tr>
</tbody>
</table>
Chip-Multiprocessors (CMPs)

- **Chip-multiprocessors**
  - Integrate *multiple* microprocessors (cores) into a single chip
    - The # of cores is increasing as technology scaling
  - Multi-threaded multi-core design

- **Process variability in CMPs?**
  - Is restricted the impact within each core
  - Is viewed as *core-to-core* variations
    - Variant $V_{th} / L_{eff} \Rightarrow$ Delay $\Rightarrow$ Frequency $\Rightarrow$ Throughput & power
Tasks & Contributions

- **Tasks**
  1. Characterize process variability in CMPs
     - Adopt an analytical model
     - Quantify the impact by Monte Carlo analysis
  2. Mitigate the impact of process variability

- **Contributions**:
  1. Enable fine-grained throughput & power management
  2. Characterize without test chips (for new process generations)

<table>
<thead>
<tr>
<th>Abstraction level</th>
<th>Method</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Architectural level</td>
<td>Variant CMP architectures</td>
<td>Global benefit</td>
</tr>
<tr>
<td>Layout level</td>
<td>Post-layout optimization</td>
<td>Local refinement</td>
</tr>
</tbody>
</table>

- Propose variability-tolerant CMP architectures
- Reduce **throughput degradation** & **dynamic power consumption**
## Variability-Tolerant CMP Architectures

<table>
<thead>
<tr>
<th>Architectures</th>
<th>Globally-clocked (GC-CMP)</th>
<th>Frequency island (FI-CMP)</th>
<th>Voltage island (VI-CMP)</th>
<th>Voltage island frequency island (VIFI-CMP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISLPED-2007</td>
<td>DAC-2008</td>
<td></td>
<td></td>
<td>Ours</td>
</tr>
<tr>
<td></td>
<td>(best paper nominate)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Globally-Clocked
-- GC-CMP

- Use a general clock for all cores
- Clock frequency is dominated by the slowest core
Frequency Island (1/2)
-- FI-CMP

- Adding **frequency island** to each core
- Minimizing throughput degradation due to process variation

![Diagram showing frequency islands with different clock frequencies: 4 GHz, 3.6 GHz, 3.2 GHz, 2.8 GHz, 2.4 GHz.](image)
The granularity of frequency bins is 10% of the nominal value.

- Feasible frequency (FF)
  - Original frequency of a core
- Operating frequency (OF)
  - Core frequency with frequency island

OF ≤ FF

Example: a 3.7GHz core with nominal value of 4GHz will be assigned a frequency island 3.6GHz
Voltage Island (1/2)
-- VI-CMP

- Adding **voltage island** to each core
- Meeting the frequency of nominal design by voltage scaling
  - Nominal design (Nom): design without process variation
The granularity of voltage bins is 10% of the nominal value.

Example: a 3.7GHz core to reach its nominal value 4GHz → adding VI to change supply voltage from 0.9V to 0.99V

Huge power penalty
Adding both voltage and frequency island

Voltage island is used for power saving
Taking advantages of slack between FF and OF of each core
- If FF > OF, the supply voltage can be lower.
- Naturally collaborating with dynamic voltage and frequency scaling (DVFS)
Models

Process variations
Transistor delay model
Throughput model
Power model
Process Variations (1/2) [Herbert 2008]

- **Within-die (WID)**
  - Systematic components
    - Caused by lithography issue
    - Modeled by normal distribution with spherical spatial correlation
  - Random components
    - Random dopant fluctuation
    - Modeled by normal distribution
- **Die-to-die (D2D)**
  - Offset of within-die variation
Process Variations (2/2)

- **Spatial correlation**
  - Chip is divided into $n \times m$ grids

  \[ \rho(r) = \begin{cases} 
  1 - \frac{3r}{2\phi} + \frac{r^3}{2\phi^3}, & \text{if } r \leq \phi \\
  0, & \text{otherwise.} 
  \end{cases} \]

- **ITRS projected the variation on $L_{eff}$ is as half of $V_{th}$**

  \[ L_{eff} = L_{eff}^0 \left( 1 + \frac{V_{th} - V_{th}^0}{2V_{th}^0} \right) \]

  - $L_{eff}$ can be obtained from $V_{th}$
Transistor Delay Model [Sarangi 2008]

- **Alpha-power law model [Sakurai 1990]**
- **Drain-situation current**
  \[
  I_d = I_d^0 \left( \frac{V_{gs} - V_{th}}{V_{DD} - V_{th}} \right)^{\alpha} = \frac{W}{L_{eff}} P_c (V_{gs} - V_{th})^{\alpha}, \text{ if } V_{ds} \geq V_d^0
  \]
  \[
  V_d^0 = P_v (V_{gs} - V_{th})^{\alpha/2}
  \]
- **Transistor delay**
  - The transistor delay is related to \( V_{th} \) and \( L_{eff} \)
  \[
  T_g \propto \frac{L_{eff} V_{DD}}{(V_{DD} - V_{th})^{\alpha}}
  \]
Throughput Model [Herbert 2008]

\[
TP(1) = \frac{1}{\frac{CPI_{com}}{F_{clk_i}} + t_{dram} + t_{link_i}(U)}
\]

\[
\sum_{i=1}^{N_{threads}} TP_i = \frac{U}{M_{L2}(S_{L2}) \cdot L_{s_i}}
\]

\[
0 = \sum_{i=1}^{N_{threads}} \frac{1}{\frac{CPI_{com}}{F_{clk_i}} + t_{dram} + t_{link_i}(U)} - \frac{U}{M_{L2}(S_{L2}) \cdot L_{s_i}}
\]

- \(TP(N_{threads})\): total system throughput
- \(U\): utilization
- \(M_{L2}(S_{L2})\): the miss rate of L2 cache
- \(S_{L2}\): the effective L2 cache size
- \(L_{s}\): the unloaded service latency of the L2 miss path
- \(F_{clk_i}\): the frequency of thread \(i\)
- \(t_{link_i}\): stall time of link component
- \(t_{dram}\): stall time of DRAM
Power Model

\[ P_{\text{dyn}}(N_{\text{threads}}) = \sum_{i=1}^{N_{\text{threads}}} P_{\text{com}_i} + P_{\text{dram}} + P_{\text{link}_i}(U) \]

- \( P_{\text{com}_i} \propto c_i \cdot V_{DD_i} \cdot F_{\text{clk}_i} \)
- \( P_{\text{dram}} = M_{L2}(S_{L2}) \cdot S_{L2} \cdot P_{\text{cell}} \)
- \( P_{\text{link}_i}(U) = M_{L2}(S_{L2}) \cdot P_{\text{lki}} \)

- \( P_{\text{dyn}}(N_{\text{threads}}) \): total dynamic power
- \( P_{\text{com}_i} \): dynamic power of a computation component
- \( p_{\text{dram}} \) and \( P_{\text{link}_i} \): power penalty when L2 cache miss occurs
- \( c_i \): self loading capacitance
Monte Carlo Analysis
Monte Carlo Methods

- **Computational algorithms**
  - Used when it is infeasible or impossible to compute an exact result with a deterministic algorithm
  - Rely on *repeated random sampling* to compute their results
  - Suited to calculation by a computer
    - Rely on repeated computation and random or pseudo-random numbers
Box Muller Transform [Box, 1958]

- Generating normal random distributions by uniform distributions
- Taking two samples from the uniform distribution on the interval $(0, 1]$
  - One is used for radius, the other is for angle
- Generating two independent normal random distributions
  - $x = R \cos \theta$, $y = R \sin \theta$
Monte Carlo Analysis [Herbert, 2008]

- **Generating 1000 variation maps of $V_{th}$ and $L_{eff}$**

- **Flow**
  - Using Box Muller transform to generate $n \times m$ normal random numbers of $\mu = V_{th}^0$ and $\sigma_{V_{th}}^{sys} / \mu = 6.4\%$
  - Adding spatial correlation
  - By $V_{th}$ and $L_{eff}$, $T_g$ of each grid can be obtained
  - Partitioning grids into corresponding #cores
  - FF of each core is the inverse of its slowest delay
Experimental Results
Experimental Setting (1/2)

- **Benchmark**
  - Obtained from [Bowman 2007]
  - 4 cores: Small (S), Medium (Me), Large (L), Monolithic (Mo)
  - 16MB shared L2 cache
  - $N_{threads}$ ranges from 1 to 70

<table>
<thead>
<tr>
<th>Core type</th>
<th>Area (mm$^2$)</th>
<th>Cores per chip</th>
<th>Core power (W)</th>
<th>$F_{clk}$ (GHz)</th>
<th>$S_{L2}$ (MB)</th>
<th>$M_{L2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>0.5</td>
<td>70</td>
<td>1.3</td>
<td>3.0</td>
<td>16</td>
<td>1.0%</td>
</tr>
<tr>
<td>Medium</td>
<td>1.5</td>
<td>23</td>
<td>3.5</td>
<td>4.0</td>
<td>16</td>
<td>1.0%</td>
</tr>
<tr>
<td>Large</td>
<td>5.0</td>
<td>7</td>
<td>9.6</td>
<td>4.0</td>
<td>16</td>
<td>1.0%</td>
</tr>
<tr>
<td>Monolithic</td>
<td>35.0</td>
<td>1</td>
<td>75.0</td>
<td>4.0</td>
<td>16</td>
<td>1.0%</td>
</tr>
</tbody>
</table>
Experimental Setting (2/2)

- **Technology**: 22nm
- **Operating point**
  - $\alpha = 1.3$
  - $L_{\text{eff}}^0 = 22\text{nm}$, $V_{DD}^0 = 0.9\text{V}$, $V_{th}^0 = 0.094\text{V}$
- **Granularity**
  - $\Delta F_{\text{clk}} = 10\% F_{\text{clk}}$
  - $\Delta V_{DD} = 10\% V_{DD}$
- **Variability**
  - $\sigma_{V_{th}}^{\text{sys}} = 6.4\%$, $\sigma_{V_{th}}^{D2D} = 0.0\%$
  - $\Phi = 0.5$, $n \times m = 100 \times 100$
Experimental Results—$V_{th}$ & $F_{clk}$

- The distributions of $V_{th}$ and $F_{clk}$ are generated by Monte Carlo analysis
  - Over 1000 variation maps
Experimental Results—Throughput (1/2)

(a) Nthreads=1

(b) Nthreads=7

(c) Nthreads=23

(d) Nthreads=70
Experimental Results—Throughput (2/2)

- Saturated when $N_{\text{threads}} > \#\text{cores}$
- Throughput degradation
  - $\text{Mo} > \text{L} > \text{Me} > \text{S}$
  - Small core has the best variability-tolerance
Experimental Results—Power

- **Core type:** Mo > L > Me > S
  - Small core has the best power efficiency.
- **CMP:** VI > Nom > GC > FI > VIFI
  - VIFI-CMP has the best power efficiency.
Best tradeoff between throughput and power occurs when $N_{\text{threads}} = \#\text{cores}$
- $N_{\text{threads}}=70$, S/VIFI has 0.06% throughput degradation and saves 36.27% dynamic power

VIFI-CMP has the best variability-tolerance
- Same throughput as FI-CMP
- Lower power consumption than Nom
Conclusion

- This work characterized process variability on throughput and power for CMPs
- Results coincided small core has the best variability-tolerance
- Without considering variations, GC-CMP delivered a lower throughput
- VIFI-CMP
  - Has the best variability-tolerance on throughput and power and can naturally collaborate with DVFS
  - Has lower power consumption than the nominal design
References


