Profitability-Based Power Allocation for Speculative Multithreaded Systems

Polychronis Xekalakis, Nikolas Ioannou
Salman Khan and Marcelo Cintra
University of Edinburgh
Introduction

- CMPs are here to stay
- Power and temperature limit performance
- No speedup for single thread applications
  - Use Thread Level Speculation to extract TLP
  - Energy Inefficient
Introduction

- CMPs are here to stay
- Power and temperature limit performance
- No speedup for single thread applications
  - Use Thread Level Speculation to extract TLP
  - Energy Inefficient
- Our proposal:
  - Steal power from non-profitable threads
  - Allocate it where it is useful
Contributions

- Propose power allocation based on thread profitability
- Propose a set of novel predictors to classify threads in profitable and non-profitable ones
- Our approach outperforms state-of-the-art TLS systems:
  - ED by 21.2% (up to 39.6%)
  - … while also reducing the temperature
Speculative Multithreading

- Basic Idea: Use idle cores_Contexts to speculate on future application needs
  - TLS: speculatively execute parallel threads
  - HT/RA: speculatively perform future memory operations
  - MP: speculatively execute along multiple branch targets

- When speculation fails, power inefficiency results
Outline

- Introduction
- Profitability Based Power Allocation
- Estimating Profitability
- Experimental Setup and Results
- Conclusions
Profitability Based Power Allocation

Time

P1  P2  P3  P4

T1  T2  T3  T4

Safe Thread: T1

- High Power Mode
- Normal Power Mode
- Low Power Mode

IPDPS 2010
Profitability Based Power Allocation

Time

P1  P2  P3  P4

T1  T2  T3  T4

Safe Thread: T1

T1  T2  T3  T4

Safe Thread: T1

- High Power Mode
- Normal Power Mode
- Low Power Mode
Profitability Based Power Allocation

Time

P1  P2  P3  P4

T1  T2  T3  T4

T1  T2  T3  T4

T1  T2  T3  T4

Safe Thread: T1
Safe Thread: T1
Safe Thread: T1

High Power Mode
Normal Power Mode
Low Power Mode
Outline

- Introduction
- Profitability Based Power Allocation
  - Estimating Profitability
- Experimental Setup and Results
- Conclusions
Estimating Profitability

- Benefits for TLS: TLP/ILP
  - TLP (Overlapped Execution)
  - ILP (Prefetching)
Estimating TLP

3-Bit Up/Down Counters

Memory Address

Bit-Wise XOR

5 bits from Program Counter
Estimating ILP
Power Mode Policy

- For threads that are predicted to squash:
  - Place in low power mode on first prediction
  - Place in very low power mode on third prediction
- For threads that are memory bound:
  - Place in low power mode
- If power budget allows, place safe thread in high power mode
Outline

- Introduction
- Profitability Based Power Allocation
- Estimating Profitability
  - Experimental Setup and Results
- Conclusions
Evaluation Environment

- Simulator, Compiler and Benchmarks:
  - SESC (http://sesc.sourceforge.net/)
  - POSH (Liu et al. PPoPP ‘06)
  - Spec 2000 Int.

- Architecture:
  - Four way CMP, 4-Issue cores
  - 16KB L1 Data (multi-versioned) and Instruction Caches
  - 1MB unified L2 Caches
  - Inst. window/ROB – 80/104 entries
### Power Modes Used

<table>
<thead>
<tr>
<th>Mode</th>
<th>Voltage</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Power</td>
<td>1000 mV</td>
<td>5.0 GHz</td>
</tr>
<tr>
<td>Normal Power</td>
<td>950 mV</td>
<td>4.0 GHz</td>
</tr>
<tr>
<td>Low Power</td>
<td>900 mV</td>
<td>3.0 GHz</td>
</tr>
<tr>
<td>Very Low Power</td>
<td>700 mV</td>
<td>1.0 GHz</td>
</tr>
</tbody>
</table>
Performance-Power Analysis

- **Speedup**

  - Bar chart showing speedup over normal-power mode for various applications.

  - Applications: bzip2, crafty, gap, gzip, mcf, parser, twolf, vortex, vpr, avg.

  - Speedup categories: Static VLow, Static Low, Profitability.
Performance-Power Analysis

Power
Performance-Power Analysis

Energy Delay
Thermal Analysis

Base TLS

Profitability-based Scheme
Outline

- Introduction
- Profitability Based Power Allocation
- Estimating Profitability
- Experimental Setup and Results
- Conclusions
Conclusions

- CMPs are here to stay
- Power on chip needs to be effectively utilized
- Allocating power by profitability leads to improvements
- Squash and memory boundedness predictors can estimate thread profitability
- Our approach outperforms state-of-the-art TLS systems:
  - ED by 21.2% (up to 39.6%)
  - … while also reducing the temperature
Profitability-Based Power Allocation for Speculative Multithreaded Systems

Polychronis Xekalakis, Nikolas Ioannou
Salman Khan and Marcelo Cintra
University of Edinburgh