Analytical Latency-Throughput Model of Future Power Constrained Multicore Processors

Amanda Chih-Ning Tseng and David Brooks
Harvard University

WEED 2012    June 9th, 2012
For Today’s Multicore Processors, the Trend Towards Maximizing **Throughput** Due to Power Constraint

What about Single-Thread Performance?
Most popular desktop and mobile applications benefit less from multicore processors

Processor performance increased 22%/year for the past decade compared to the historical 52%**

** Hennessy and Patterson, Computer Architecture, A Quantitative Approach, Ed.5

---

<table>
<thead>
<tr>
<th>App Type</th>
<th>Office</th>
<th>Media Playback</th>
<th>Game</th>
<th>Web Browsing</th>
<th>Image</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. TLP</td>
<td>1.2</td>
<td>1.5</td>
<td>1.6</td>
<td>2.0</td>
<td>2.1</td>
<td>7.4</td>
</tr>
</tbody>
</table>

* Blake et. al., Evolution of Thread-Level Parallelism in Desktop Applications, ISCA’10
To study the tradeoff, a model that analyzes the performance relation is necessary.
Model Overview

- Single Core Info
  - Core Power
    - Power Constraint
  - Core STP
    - Arch Type

Generate Multicore Configuration

Performance Computation

- Workload Characteristic
  - STP, NSTP
  - Arch Type

Output Results

- Best?
  - NO
  - YES
  - Optimal Core Configuration
- F
  - Latency Performance
- M
  - Throughput Performance

Results
Model Overview

Model Overview

Thermal Design Power (TDP) :  
Maximum power that the cooling system can dissipate  
Decided at design time

Architecture Type :  
Symmetric or asymmetric multicore  
(Currently implemented)
Model Overview

A Task:

Sequential Portion

Parallel Portion

Workload Characteristic

M

1-F

F
Related Works

  - Augmenting Amdahl’s low to examine theoretical speedups for symmetric, asymmetric, and dynamic multicore chips.

• “Dark Silicon and the End of Multicore Scaling,” H. Esmaeilzadeh et. al., ISCA ’11
  - Using technology scaling and performance model to predict speedup and % of dark silicon of future computing systems
Single Task: Symmetric Multicore

Amdahl’s Law: \[
\text{Speedup} = \frac{1}{(1-F) + \frac{F}{N}}
\]

Power Constraint:
\[
N_S = \frac{\text{TDP}}{P_s}
\]

Performance:
\[
\text{STP} = \text{STP}_s
\]
\[
\text{THP} = \text{STP}_s \times \frac{1}{(1-F) + \frac{F}{N_S}} = \frac{1}{\frac{(1-F)}{\text{STP}_s} + \frac{F}{N_S \times \text{STP}_s}}
\]
Single Task: Asymmetric Multicore

Amdahl's Law: \[ \text{Speedup} = \frac{1}{(1-F) + F/N} \]

Power Constraint:
\[ N_S = \frac{(\text{TDP} - P_B)}{P_S} \]

Performance:
\[ \text{STP} = \text{STP}_B \]
\[ \text{THP} = \frac{1}{\frac{(1-F)}{\text{STP}_B} + \frac{F}{\text{STP}_B + N_S \times \text{STP}_S}} \]
Multi-Thread Throughput v.s. Single Thread Performance

- Max THP
- Better Tradeoff btw THP and STP
- Sacrificed STP When Maximize THP
- Max STP
Multi-Task Model

- **Single Task Performance**: Throughput performance for single task when using all cores
- **Multi-Task Throughput Performance**: Average time to finish a task when executing multiple tasks
- **Resource Group**: One or more cores exclusively assigned to a particular task
Throughput Generalization

\[ \text{THP}_{\text{Symm}} = \frac{1}{(1-F) + \frac{F}{N_S \times \text{STP}_S}} \]

\[ \text{THP}_{\text{Asym}} = \frac{1}{(1-F) + \frac{F}{\text{STP}_B + N_S \times \text{STP}_S}} \]

\[ T(n_B, n_S, \text{stp}_{\text{seq}}) = \frac{(1-F)}{\text{stp}_{\text{seq}}} + \frac{F}{(n_B \times \text{STP}_B) + (n_S \times \text{STP}_S)} \]

\[ \text{THP}_{\text{Symm}} = \frac{1}{T(0, N_S, \text{STP}_S)} \]

\[ \text{THP}_{\text{Asym}} = \frac{1}{T(1, N_S, \text{STP}_B)} \]
Multi-Task Symmetric Model

\[ t = 0 \]

\[ t = T(0, 1, STP_S) \]

\[ t = 2 \times T(0, 1, STP_S) \]

\[ t = T_{\text{last}} = 2 \times T(0, 1, STP_S) + T(0, 2, STP_S) \]
Multi-Task Asymmetric Model

\[ t = 0 \]
\[ T(n_B, n_S, \text{stp}_{\text{seq}}) \]

\[ t = T(1, 2, \text{STP}_B) \]

\[ t = 2 \times T(1, 2, \text{STP}_B) \]

\[ t = T(0, 8, \text{STP}_S) \]

\[ t = T_{\text{last}} = 2 \times T(1, 2, \text{STP}_B) + T(2, 4, \text{STP}_B) \]
Multi-Task Throughput v.s. Single Task Performance, $F = 0.75$

- **Max Task THP**
- **Max Performance Product**
- **Max Single Task Performance**
Asymmetric Core Configuration, $F = 0.75$

Core Count

Core Performance
Takeaways

• Asymmetric design benefits more when power budget is higher

• Core configurations varied significantly for different performance metrics

• Almost every size of single cores is selected to form an optimized configuration
For future processors, improving both single thread performance and throughput performance is significant.

We construct an analytical model that computes both type of performance under a power constraint for different multicore architecture.

Future work: considering overheads like cache sharing and NoC, extending the model for heterogeneous multi-task, and analyzing QoS metrics.
Questions?