Managing GPU Concurrency in Heterogeneous Architectures

Shared Resources

Memory

LLC

Network

Shared Resources
Managing GPU Concurrency in Heterogeneous Architectures
Managing GPU Concurrency in Heterogeneous Architectures
Managing GPU Concurrency in Heterogeneous Architectures
Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism
## Our Proposal

**Warp Scheduler**
Controls GPU Thread-Level Parallelism

<table>
<thead>
<tr>
<th>CPU-centric Strategy</th>
<th>Improved GPU performance</th>
<th>Improved CPU performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>×</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Our Proposal

**Warp Scheduler**
- Controls GPU Thread-Level Parallelism

<table>
<thead>
<tr>
<th></th>
<th>Improved GPU performance</th>
<th>Improved CPU performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU-centric Strategy</strong></td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td><strong>CPU-GPU Balanced Strategy</strong></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism

<table>
<thead>
<tr>
<th></th>
<th>Improved GPU performance</th>
<th>Improved CPU performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU-centric</strong></td>
<td>×</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Strategy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CPU-GPU</strong></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Balanced</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Strategy</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Control the trade-off
Our Proposal

CPU-centric Strategy

Memory Congestion ↑
CPU Performance ↓
Our Proposal

CPU-centric Strategy

Memory Congestion ↑
CPU Performance ↓

IF Memory Congestion ↑
GPU TLP ↓
Our Proposal

**CPU-centric Strategy**

- Memory Congestion ➡️
- CPU Performance ➡️

**IF Memory Congestion ➡️**

- GPU TLP ➡️

**Results Summary:**

+24% CPU & -11% GPU
Our Proposal

CPU-centric Strategy

- Memory Congestion
- CPU Performance

CPU-GPU Balanced Strategy

- GPU TLP
- GPU Latency Tolerance

Results Summary:

+24% CPU & -11% GPU
Our Proposal

**CPU-centric Strategy**
- Memory Congestion
- CPU Performance

**CPU-GPU Balanced Strategy**
- GPU TLP
- GPU Latency Tolerance

**Results Summary:**
+24% CPU & -11% GPU
Our Proposal

<table>
<thead>
<tr>
<th>CPU-centric Strategy</th>
<th>CPU-GPU Balanced Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory Congestion</td>
<td>GPU TLP</td>
</tr>
<tr>
<td>CPU Performance</td>
<td>GPU Latency Tolerance</td>
</tr>
</tbody>
</table>

Results Summary:

CPU-centric Strategy: +24% CPU & -11% GPU

CPU-GPU Balanced Strategy: +7% both CPU & GPU
Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayıran¹,
Nachiappan CN¹, Adwait Jog¹, Rachata Ausavarungnirun²,
Mahmut T. Kandemir¹, Gabriel H. Loh³, Onur Mutlu², Chita R. Das¹

¹ Penn State
² Carnegie Mellon
³ AMD Research
Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayıran¹, Nachiappan CN¹, Adwait Jog¹, Rachata Ausavarungnirun², Mahmut T. Kandemir¹, Gabriel H. Loh³, Onur Mutlu², Chita R. Das¹

¹ Penn State ² Carnegie Mellon ³ AMD Research

Today
Session 1B – Main Auditorium  @ 3 pm