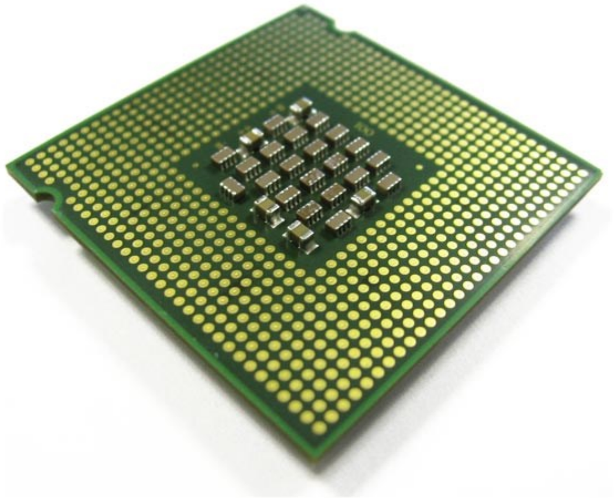
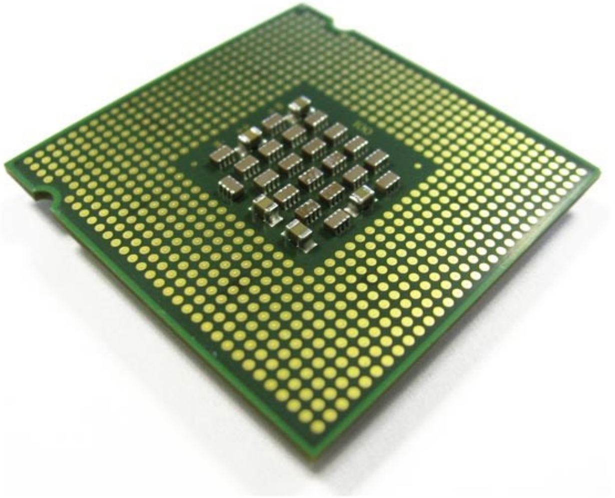
Managing GPU Concurrency in Heterogeneous Architectures



Shared Resources

Managing GPU Concurrency in Heterogeneous Architectures



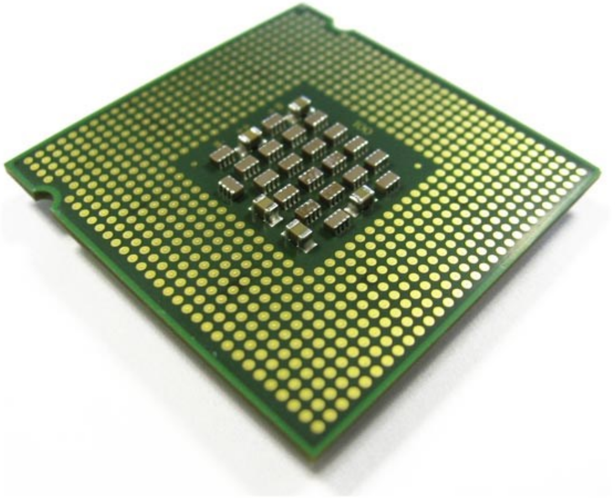
LLC

Memory

Network

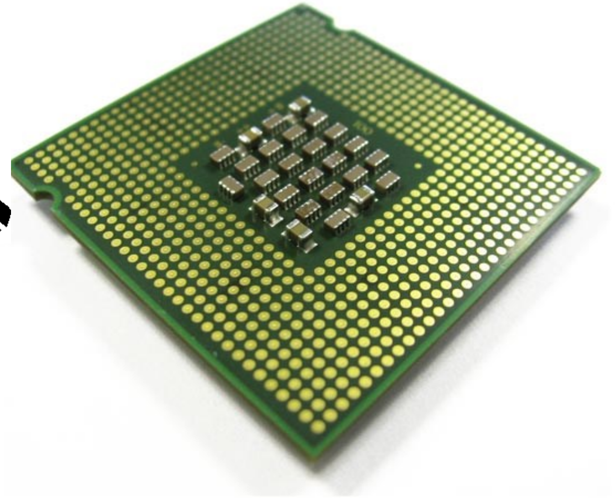
Shared Resources

Managing GPU Concurrency in Heterogeneous Architectures



Shared Resources

Managing GPU Concurrency in Heterogeneous Architectures



LLC

Memory

Network

Shared Resources

Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism

Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism

	Improved GPU performance	Improved CPU performance
CPU-centric Strategy	×	☑

Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism

	Improved GPU performance	Improved CPU performance
CPU-centric Strategy	×	☑
CPU-GPU Balanced Strategy	☑	☑

Our Proposal

Warp Scheduler
Controls GPU Thread-Level Parallelism

	Improved GPU performance	Improved CPU performance
CPU-centric Strategy	×	☑
CPU-GPU Balanced Strategy	☑	☑

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 ⬇

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 ⬇

IF Memory Congestion ⬆

⬇ GPU TLP

IF Latency Tolerance ⬇

⬆ GPU TLP

Results Summary:

+24% CPU & -11% GPU

Our Proposal

CPU-centric Strategy	**CPU-GPU Balanced Strategy**
Memory Congestion ⬆	GPU TLP ⬇⬇⬇
CPU Performance ⬇	GPU Latency Tolerance ⬇
IF Memory Congestion ⬆ ⬇ GPU TLP	IF Latency Tolerance ⬇ ⬆ GPU TLP
Results Summary: +24% CPU & -11% GPU	Results Summary: +7% both CPU & GPU

Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayıran[1],

Nachiappan CN[1], Adwait Jog[1], Rachata Ausavarungnirun[2],

Mahmut T. Kandemir[1], Gabriel H. Loh[3], Onur Mutlu[2], Chita R. Das[1]

[1] Penn State

[2] Carnegie Mellon

[3] AMD Research

Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayıran[1],

Nachiappan CN[1], Adwait Jog[1], Rachata Ausavarungnirun[2],

Mahmut T. Kandemir[1], Gabriel H. Loh[3], Onur Mutlu[2], Chita R. Das[1]

[1] Penn State

[2] Carnegie Mellon

[3] AMD Research

Today

Session 1B – Main Auditorium

@ 3 pm