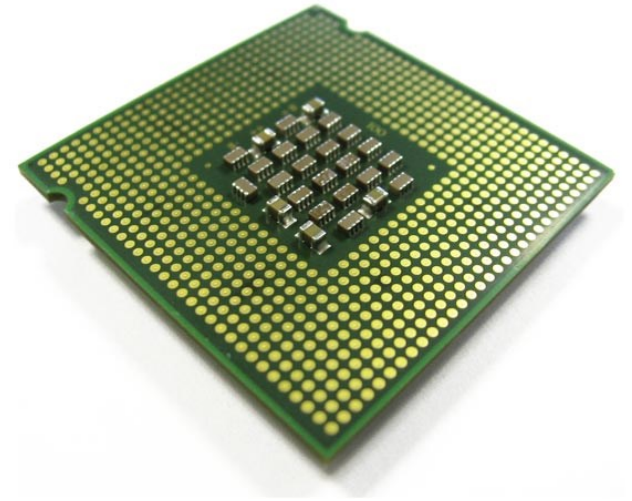
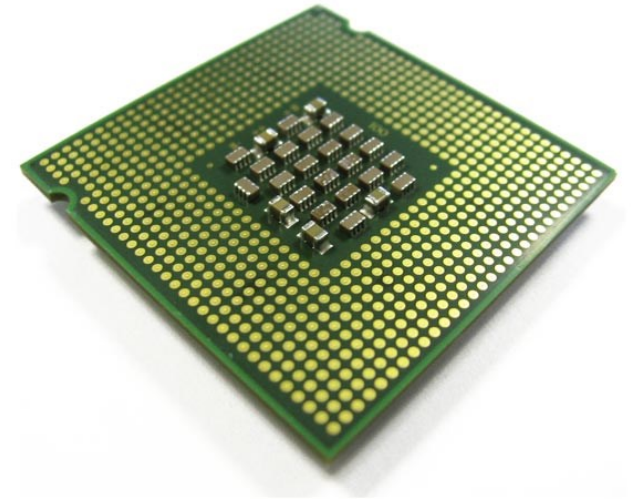


Managing GPU Concurrency in Heterogeneous Architectures



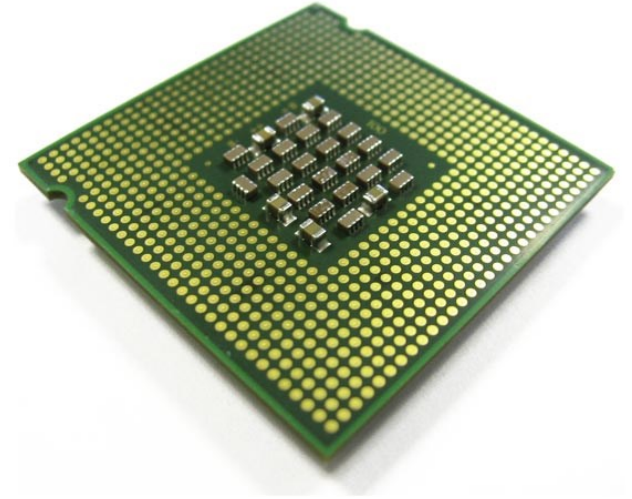
LLC
Memory **Network**
Shared Resources

Managing GPU Concurrency in Heterogeneous Architectures



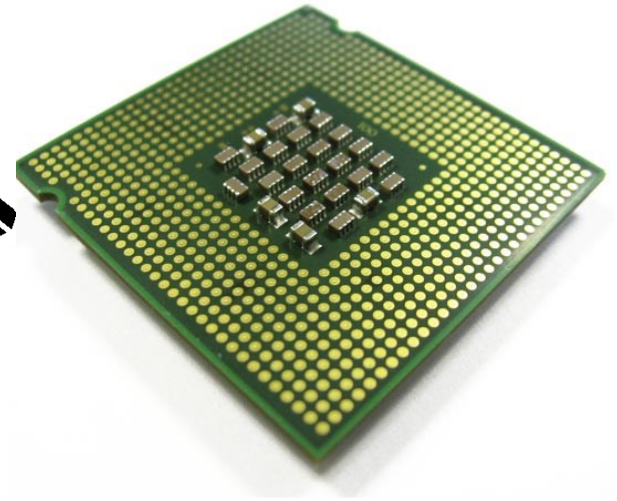
LLC
Memory Network
Shared Resources

Managing GPU Concurrency in Heterogeneous Architectures



LLC
Memory Network
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

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

Results Summary:

+7% both CPU & GPU

Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayiran¹,

Nachiappan CN¹, Adwait Jog¹, Rachata Ausavarungnirun²,

Mahmut T. Kandemir¹, Gabriel H. Loh³, Onur Mutlu², Chita R. Das¹



Carnegie Mellon



¹ Penn State

² Carnegie Mellon

³ AMD Research

Managing GPU Concurrency in Heterogeneous Architectures

Onur Kayiran¹,

Nachiappan CN¹, Adwait Jog¹, Rachata Ausavarungnirun²,

Mahmut T. Kandemir¹, Gabriel H. Loh³, Onur Mutlu², Chita R. Das¹



Carnegie Mellon



¹ Penn State

² Carnegie Mellon

³ AMD Research

Today

Session 1B – Main Auditorium

@ 3 pm