Enabling GPU Sharing with Address Translation

Virtual Address

GPU Core

GPU Core

GPU Core

GPU Core

Page Table Walkers

Page Table (in main memory)

App 1

App 2
Enabling GPU Sharing with Address Translation

Virtual Address

GPU Core  GPU Core  GPU Core  GPU Core

High latency page walks

Page Table Walkers

Page Table (in main memory)

App 1

App 2
State-of-the-Art Translation Support in GPUs

Virtual Address

GPU Core
  Private TLB

GPU Core
  Private TLB

GPU Core
  Private TLB

GPU Core
  Private TLB

Shared TLB

Page Table Walkers

Page Table (in main memory)

High latency page walks

Shared

Private

App 1

App 2
Three Sources of Inefficiency in Translation

- High TLB contention
- Inefficient caching
- Bypass Address translation is latency sensitive

MASK: A Translation-aware Memory Hierarchy
Three Sources of Inefficiency in Translation

High TLB contention
Three Sources of Inefficiency in Translation

High TLB contention

Inefficient caching
Three Sources of Inefficiency in Translation

High TLB contention

Inefficient caching

Address translation is latency-sensitive
Our Solution

MASK: A Translation-aware Memory Hierarchy
Three Components of MASK
Three Components of MASK

**TLB-fill Tokens**
Reduces TLB contention

**Shared TLB**
Three Components of MASK

**TLB-fill Tokens**
Reduces TLB contention

**Translation-aware L2 Bypass**
Improves L2 cache utilization

- **Shared TLB**
- **Translation Data**
- **L2 Data Cache**
Three Components of MASK

**TLB-fill Tokens**
- Reduces TLB contention

**Translation-aware L2 Bypass**
- Improves L2 cache utilization

**Address-space-aware Memory Scheduler**
- Lowers address translation latency

*Shared TLB* → *Translation Data* → *L2 Data Cache* → *Translation Data* → *Main Memory*
Three Components of MASK

- **TLB-fill Tokens**
  - Reduces TLB contention

- **Translation-aware L2 Bypass**
  - Improves L2 cache utilization

- **Address-space-aware Memory Scheduler**
  - Lowers address translation latency

**MASK improves performance by 57.8%**
MASK: Redesigning the GPU Memory Hierarchy to Support Multi-Application Concurrency

Rachata Ausavarungnirun
Vance Miller         Joshua Landgraf         Saugata Ghose
Jayneel Gandhi       Adwait Jog         Christopher J. Rossbach         Onur Mutlu

GPU 2 (Virginia EF)   Tuesday 2PM-3PM

Carnegie Mellon

ETH Zürich

Texas

vmware

SAFARI