Mosaic: A GPU Memory Manager with Application-Transparent Support for Multiple Page Sizes

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

Session 2-A, 2PM-4PM

Carnegie Mellon, ETH Zürich, TEXAS, The University of Texas at Austin, VMware®

SAFARI
Bottlenecks of GPU Virtual Memory

- GPU Core
  - Private TLB
- GPU Core
  - Private TLB
- GPU Core
  - Private TLB
- GPU Core
  - Private TLB

Shared TLB

Page Table Walkers

Page Table (Main memory)

Data (Main Memory)

CPU Memory

GPU-side memory

CPU-side memory

Private
Shared

2
Bottlenecks of GPU Virtual Memory

- **Shared TLB**
  - Limited TLB reach

- **Page Table Walkers**
  - High latency page walks

- **Page Table (Main memory)**

- **Data (Main Memory)**

- **GPU Core**
  - Private TLB

- **CPU Memory**

High latency I/O
Key Page Size Tradeoffs

Larger pages: Better TLB reach
High demand paging latency
Key Page Size Tradeoffs

Larger pages: Better TLB reach
High demand paging latency

Smaller pages: Lower demand paging latency
Limited TLB reach
Key Page Size Tradeoffs

Larger pages: Better TLB reach
High demand paging latency

Smaller pages: Lower demand paging latency
Limited TLB reach

Mosaic enables application-transparent use of both page sizes
Key Challenge with Multiple Page Sizes

*State-of-the-art*

Large Page Frame 1

Unallocated  App 1  App 2

Large Page Frame 2

*Cannot coalesce pages*
Key Idea of Mosaic

State-of-the-art

Cannot coalesce pages

With Mosaic

In-Place Coalescing
Mosaic

GPU Runtime

Hardware
Mosaic

GPU Runtime

Contiguity-Conserving Allocation

Hardware
Mosaic

**GPU Runtime**

- Contiguity-Conserving Allocation
- In-Place Coalescer

**Hardware**
Mosaic

- Contiguity-Conserving Allocation
- In-Place Coalescer
- Contiguity-Aware Compaction

**GPU Runtime**

**Hardware**
Benefits

High TLB reach

Low demand paging latency

Application-transparent

55% higher average performance
Mosaic: A GPU Memory Manager with Application-Transparent Support for Multiple Page Sizes

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

Session 2-A  2PM-4PM

Carnegie Mellon
ETH Zürich
TEXAS
vmware

SAFARI