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

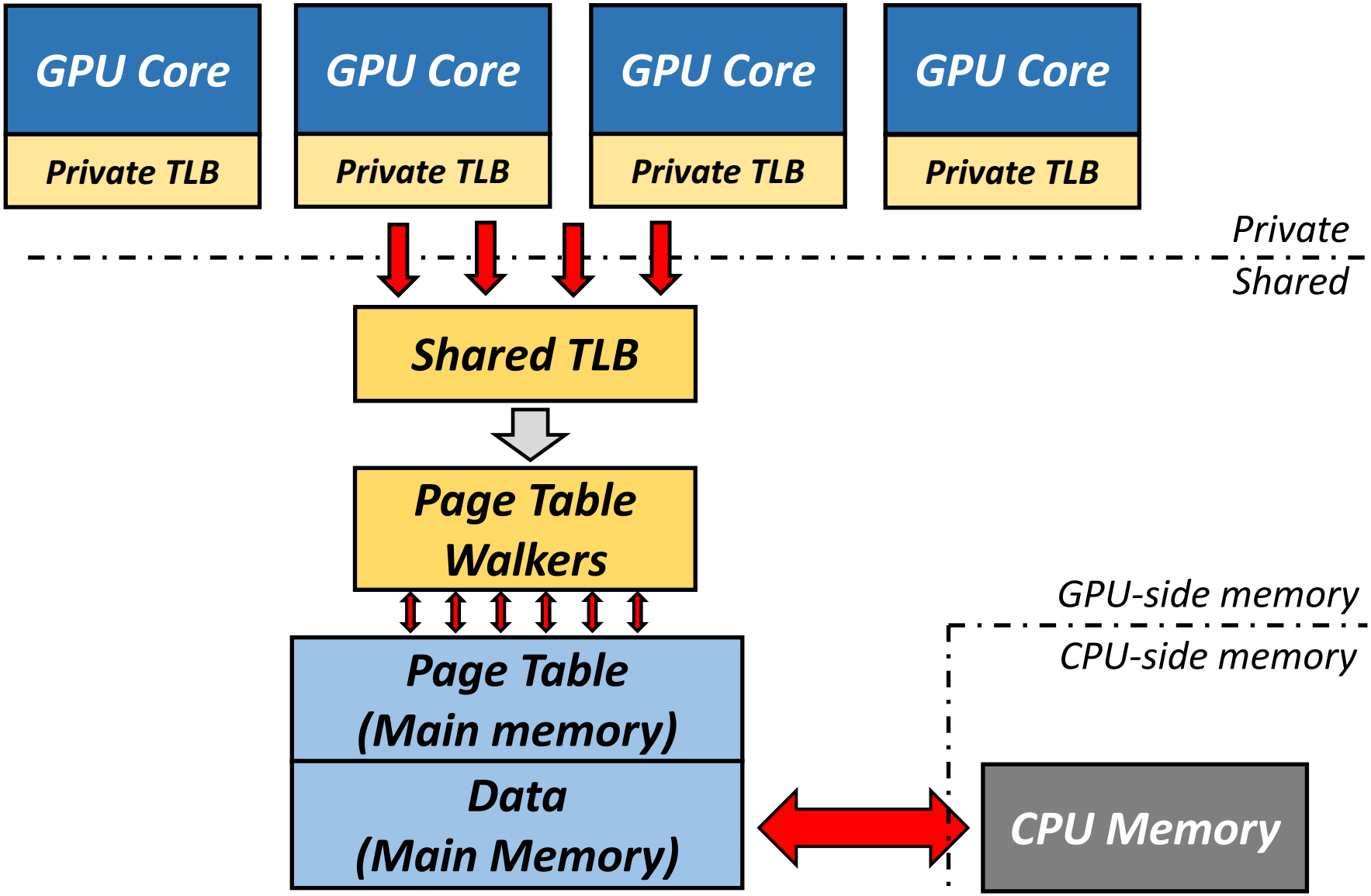
ETH Zürich

TEXAS
The University of Texas at Austin

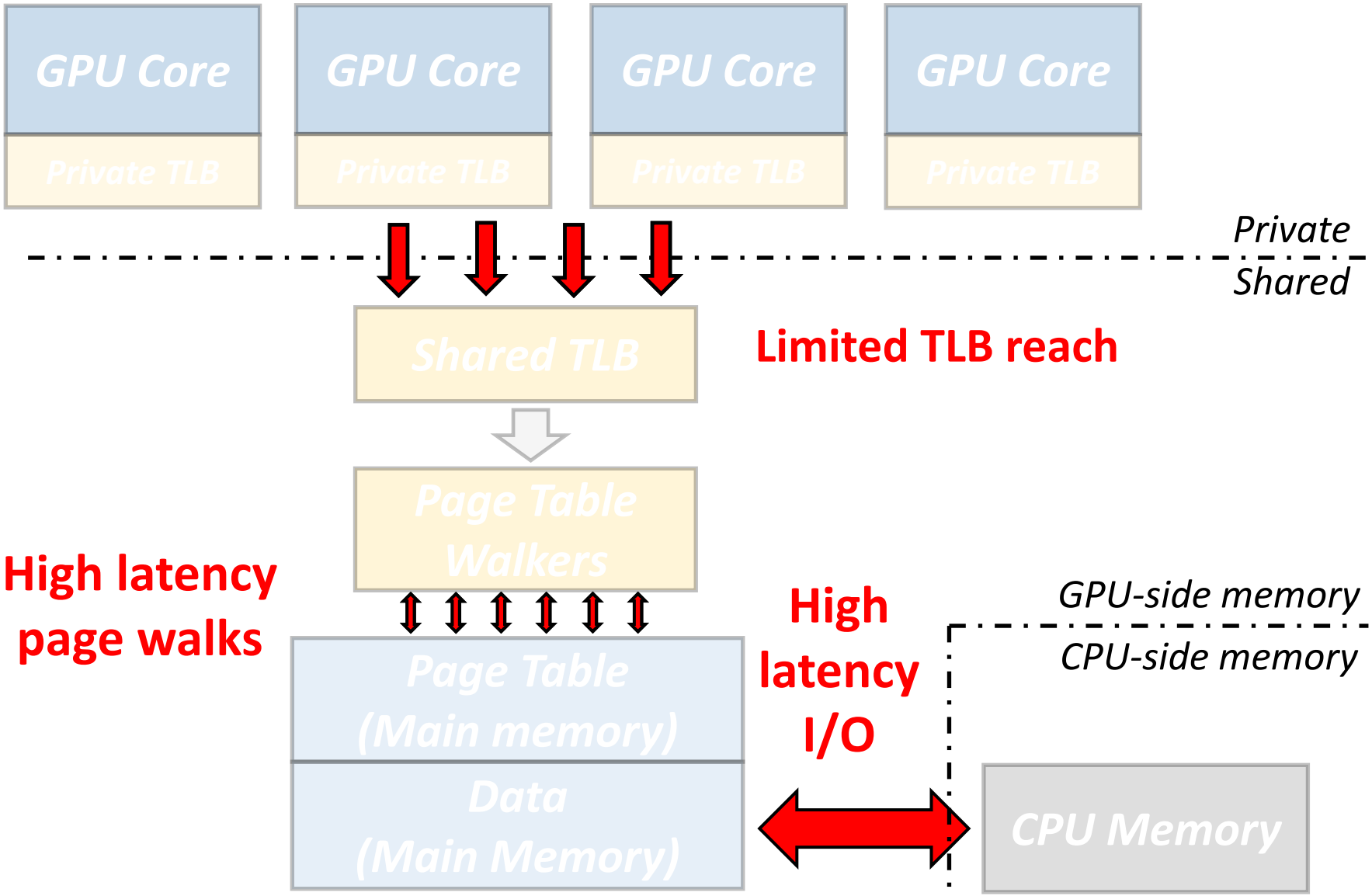
vmware®

SAFARI

Bottlenecks of GPU Virtual Memory



Bottlenecks of GPU Virtual Memory



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

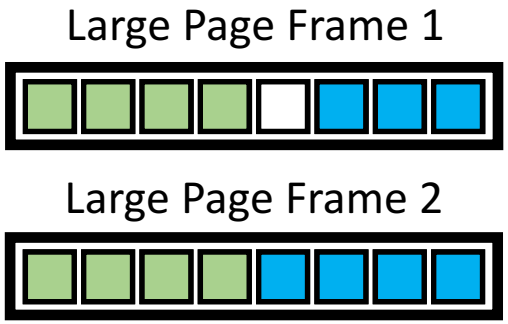
Smaller pages: Lower demand paging latency

Limited TLB reach

Mosaic enables application-transparent use of both page sizes

Key Challenge with Multiple Page Sizes

State-of-the-art



Cannot coalesce pages

Unallocated **App 1** **App 2**

Key Idea of Mosaic

State-of-the-art

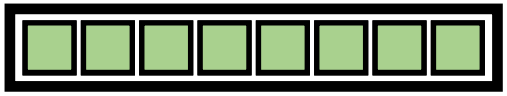
Large Page Frame 1

Large Page Frame 2

Cannot coalesce pages

With Mosaic

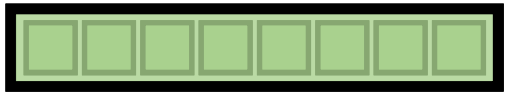
Large Page Frame 1



Large Page Frame 2

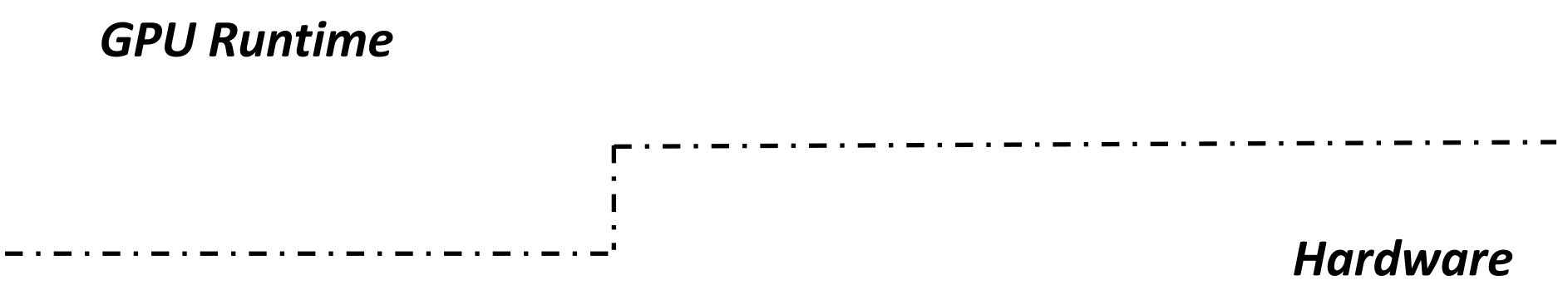
In-Place Coalescing

Large Page Frame 1

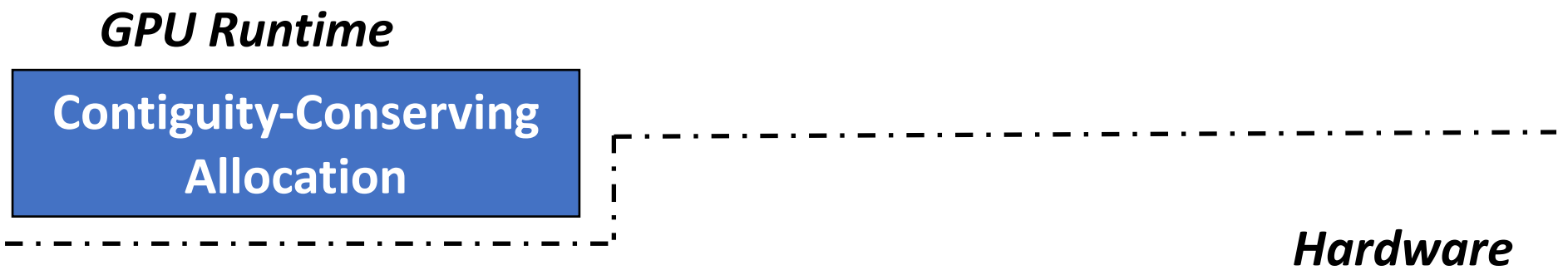


Large Page Frame 2

Mosaic



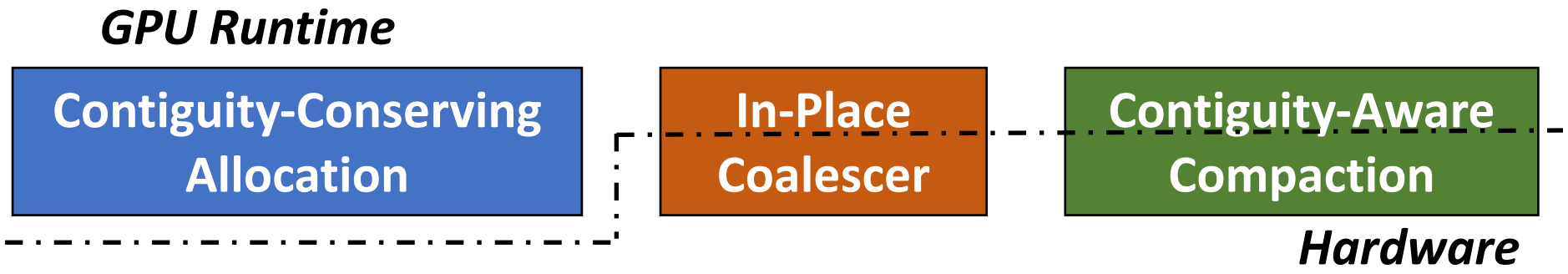
Mosaic



Mosaic


GPU Runtime
Contiguity-Conserving Allocation
In-Place Coalescer
Hardware

Mosaic


GPU Runtime
Contiguity-Conserving Allocation
In-Place Coalescer
Contiguity-Aware Compaction
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
The University of Texas at Austin

vmware®

SAFARI