MASK: Redesigning the GPU Memory Hierarchy to Support Multi-Application Concurrency

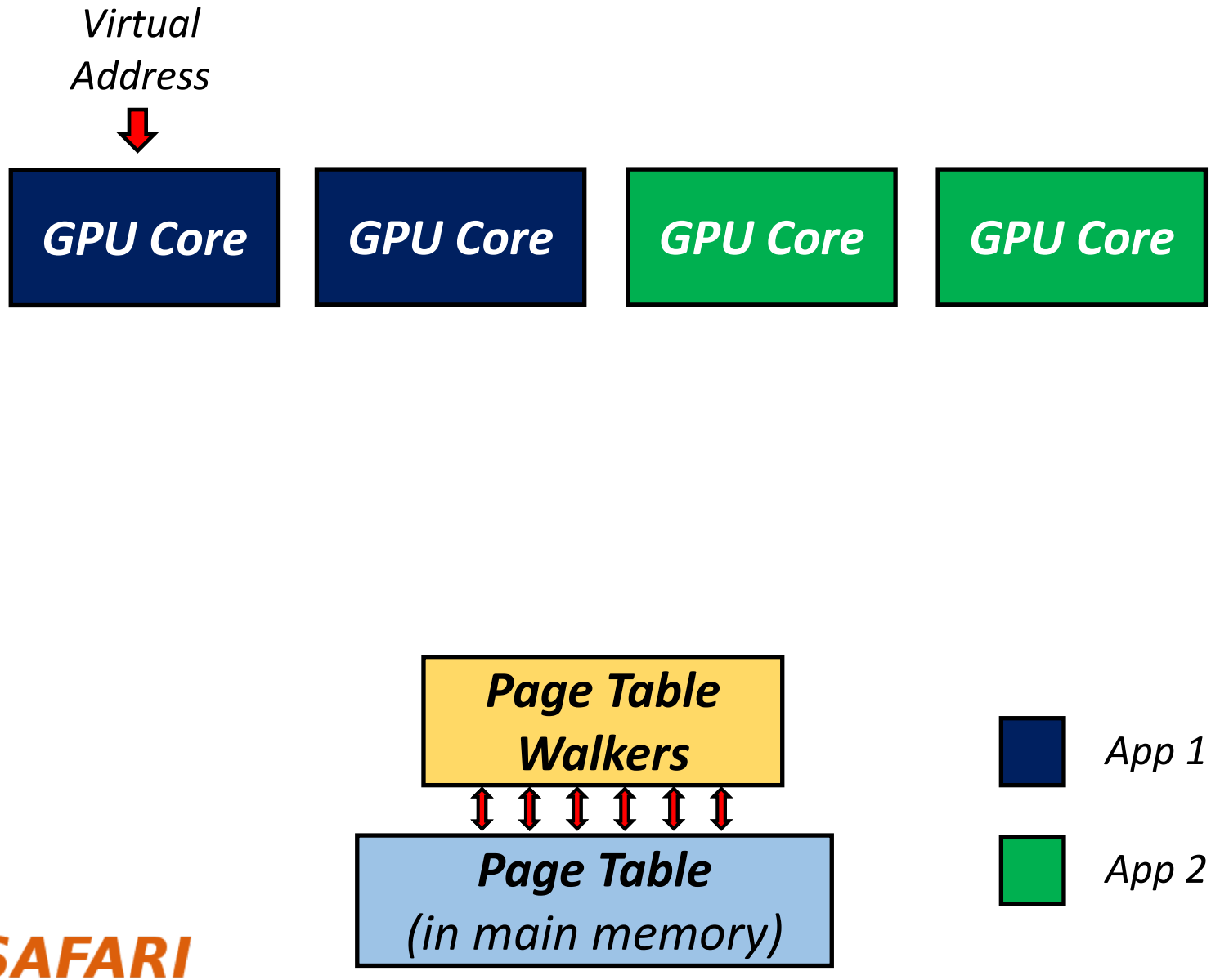
Rachata Ausavarungnirun

Vance Miller Joshua Landgraf Saugata Ghose

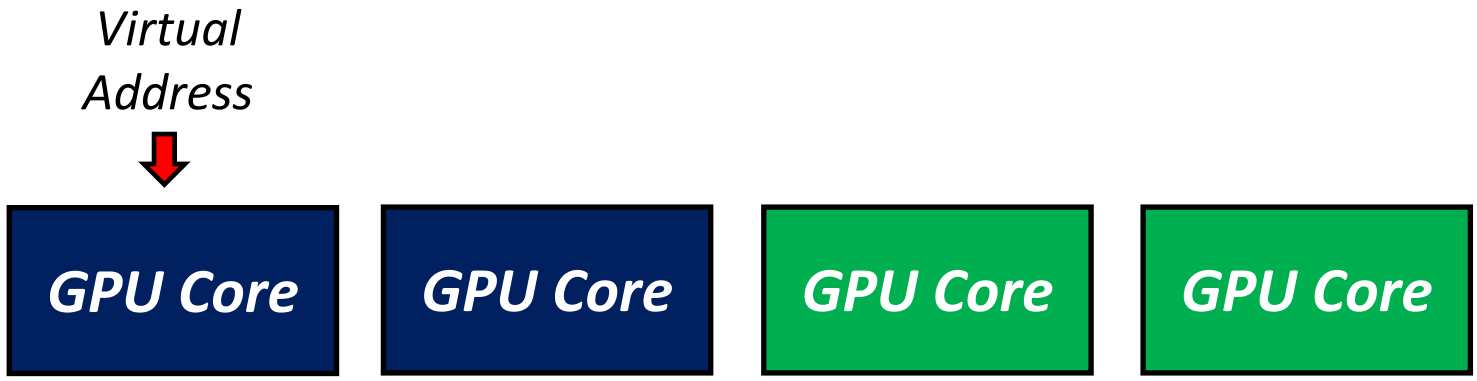
Jayneel Gandhi Adwait Jog Christopher J. Rossbach Onur Mutlu

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

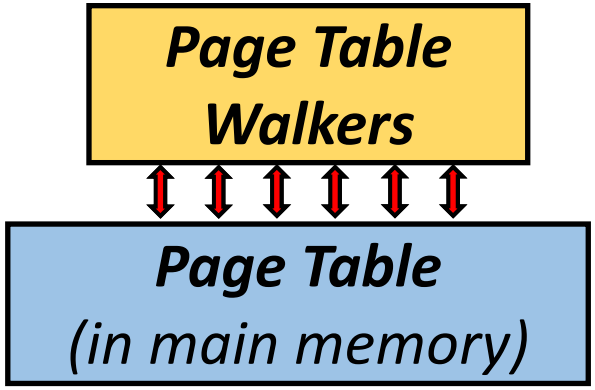
Enabling GPU Sharing with Address Translation

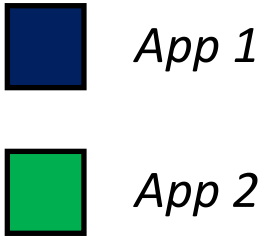


Enabling GPU Sharing with Address Translation

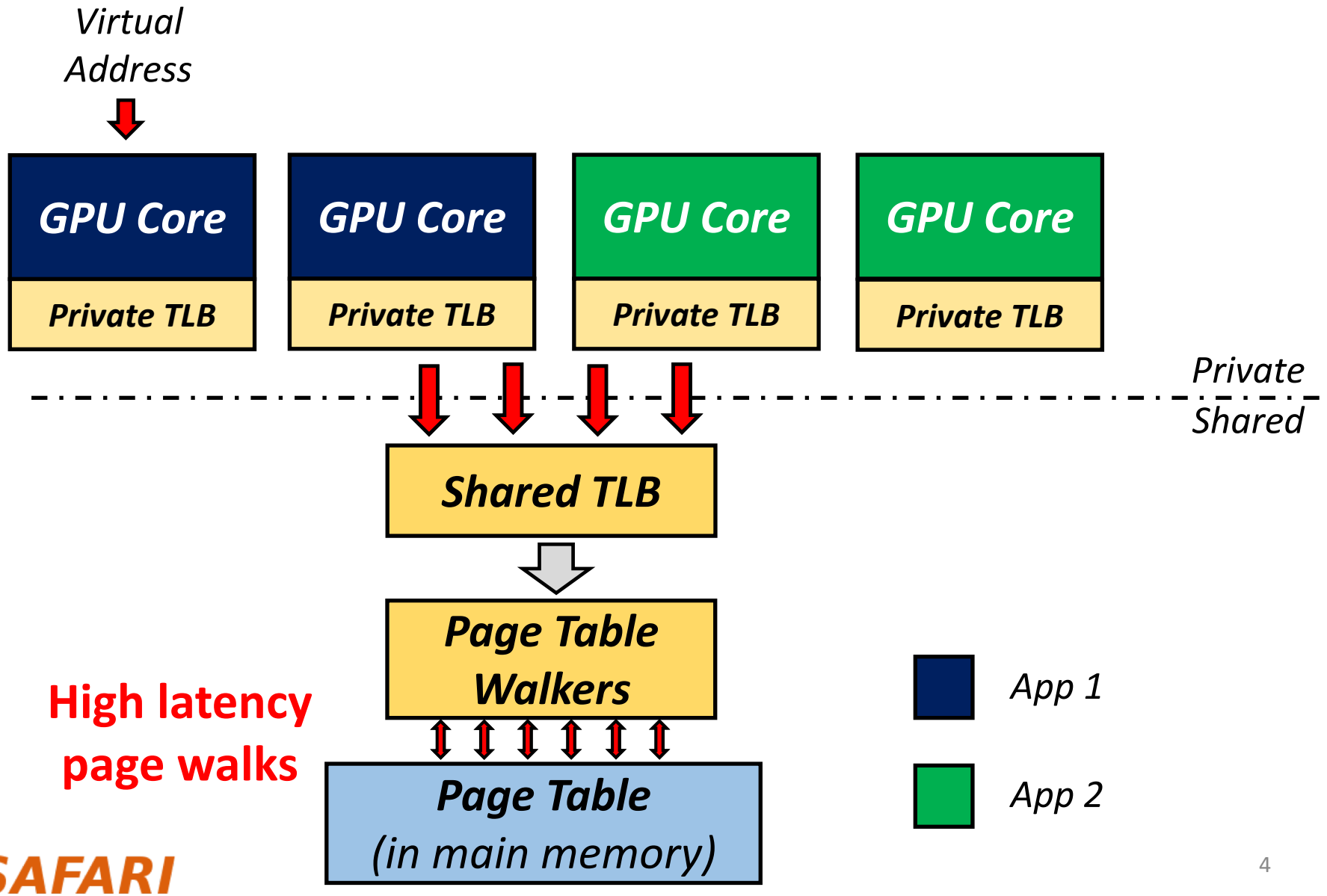


High latency page walks





State-of-the-Art Translation Support in GPUs



Three Sources of Inefficiency in Translation

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

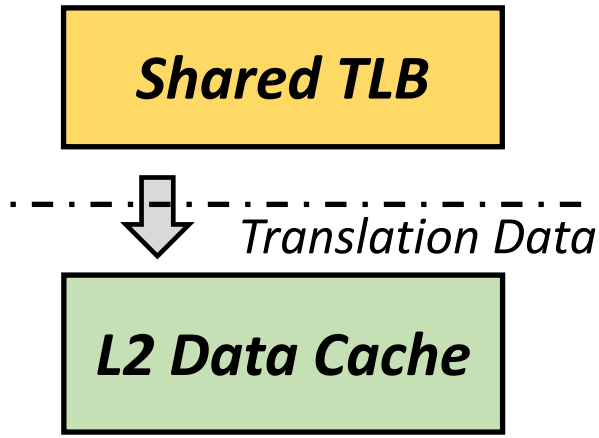
Three Components of MASK

TLB-fill Tokens

Reduces TLB contention

Translation-aware L2 Bypass

Improves L2 cache utilization



Three Components of MASK

TLB-fill Tokens

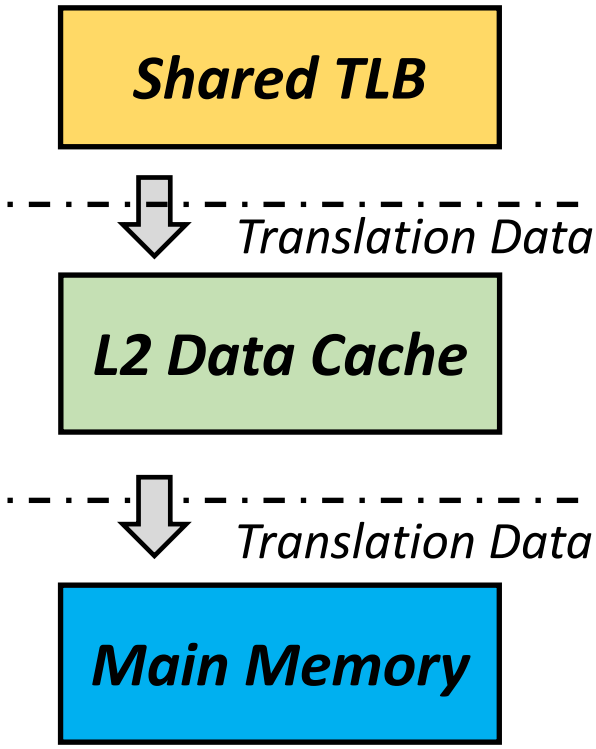
Reduces TLB contention

Translation-aware L2 Bypass

Improves L2 cache utilization

Address-space-aware Memory Scheduler

Lowers address translation latency



Three Components of MASK

TLB-fill Tokens

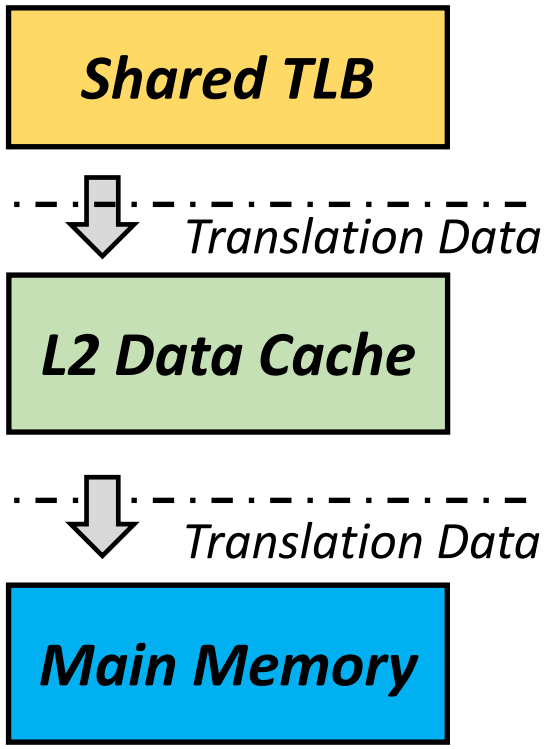
Reduces TLB contention

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**