THE CHIPS TO SYSTEMS CONFERENCE
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

$MILE : LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism

Tianyu Guo, Xuanteng Huang, Kan Wu, Xianwei Zhang, Nong Xiao

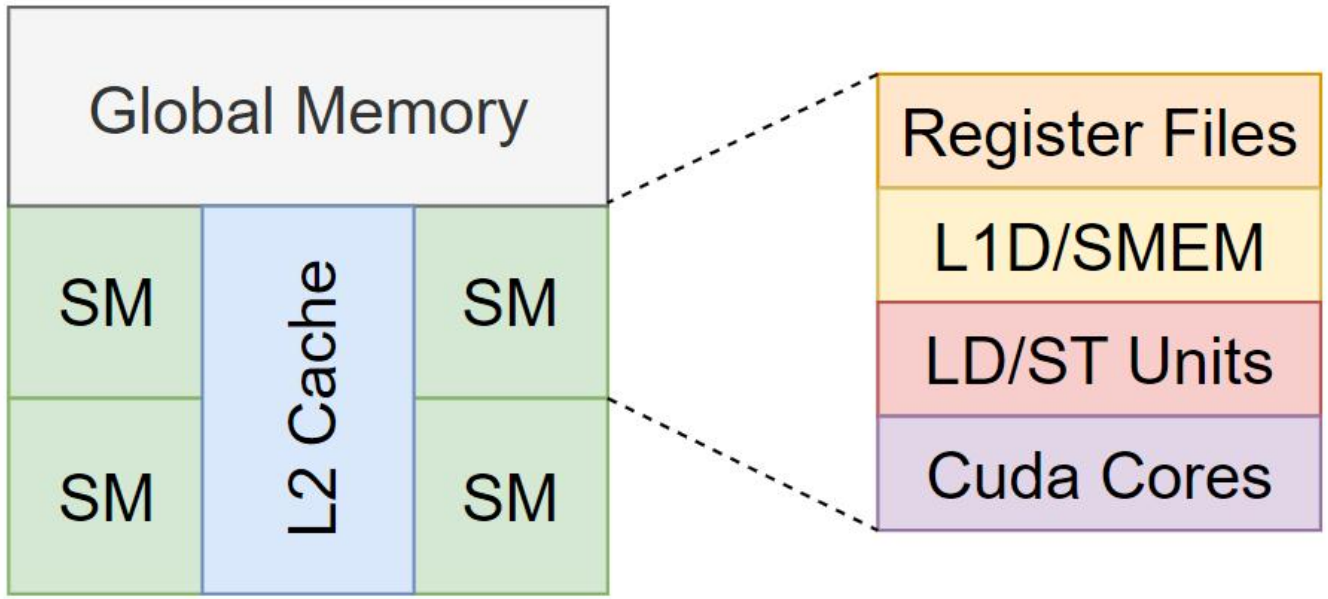
Sun Yat-sen University

Agenda

- Background
- Motivation
- Design
- Evaluation
- Summary

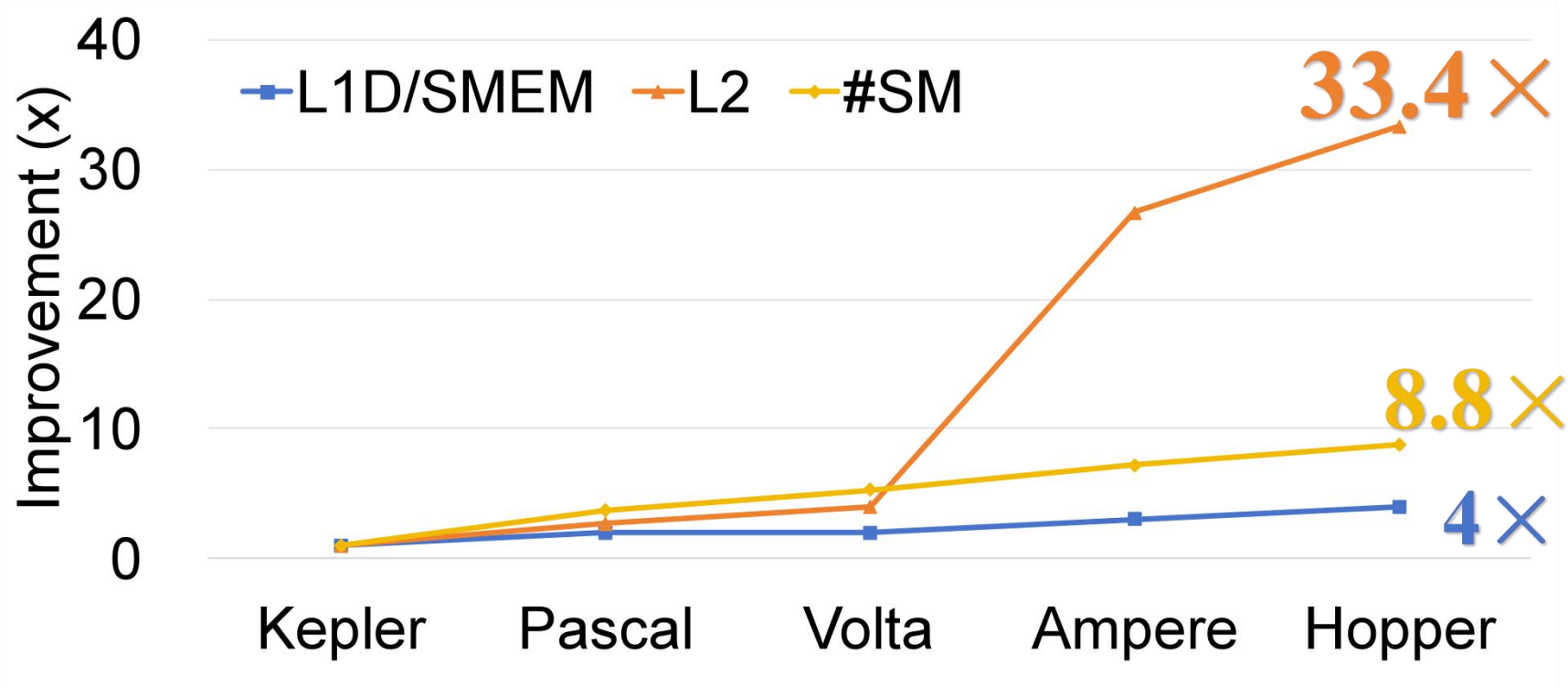
GPU Architecture

- A GPU is composed of multiple streaming multiprocessors (SMs)
- Each SM contains private L1 data cache and shared memory (L1D/SMEM)
- All the SMs share a L2 cache



GPU Evolution

- Improvement of GPU resources is out of proportion
- Since the Ampere architecture, L1D/SMEM is insufficient while LLC is abundant

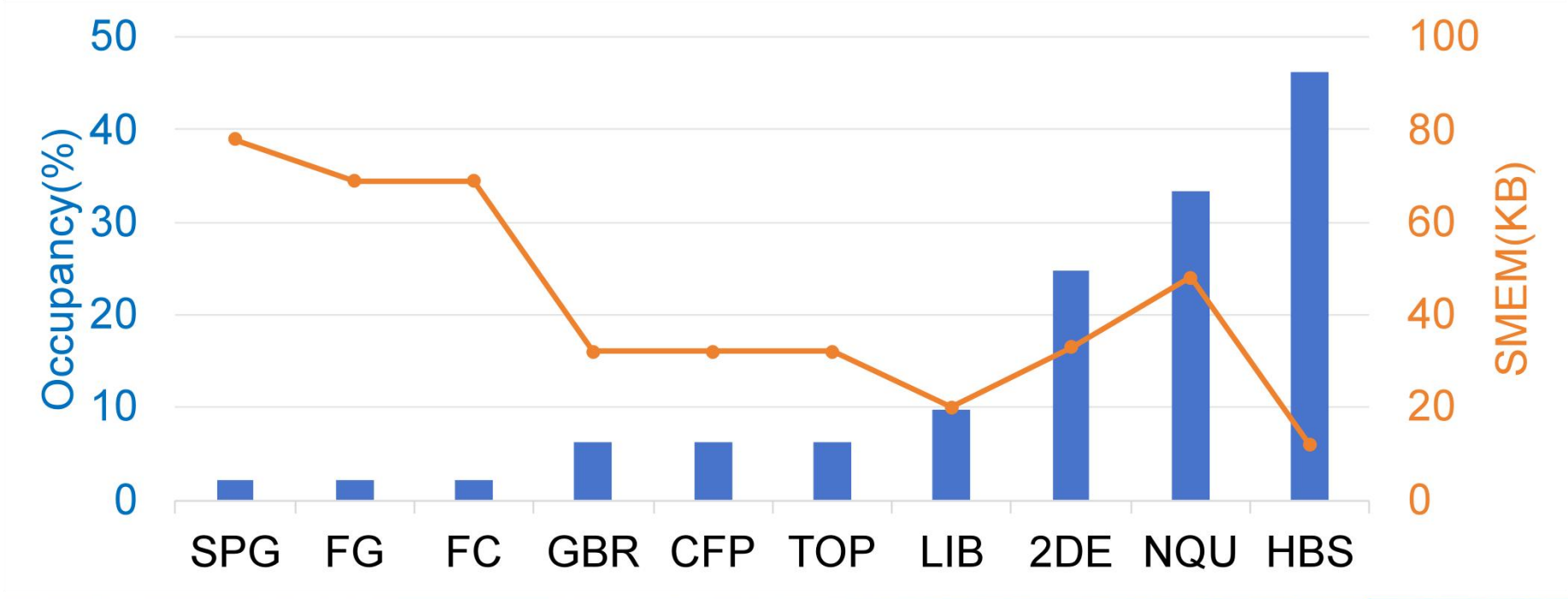


Agenda

- Background
- **Motivation**
- Design
- Evaluation
- Summary

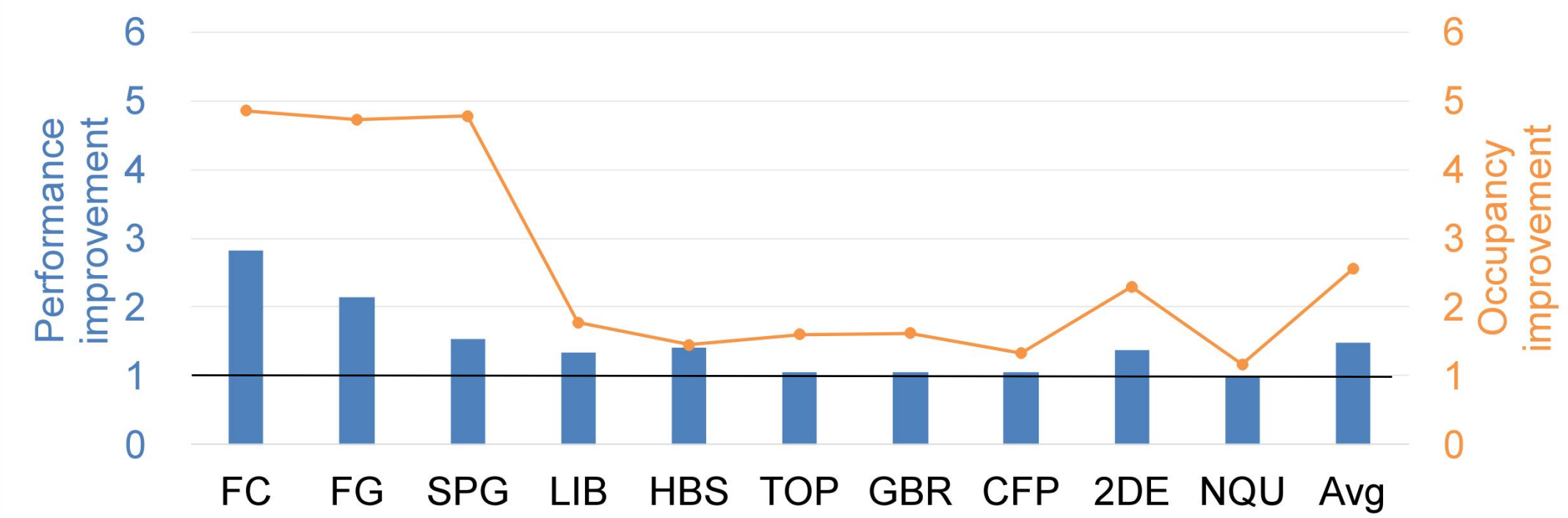
Higher SMEM Usage Causes Lower Occupancy

- Applications are exhibiting low occupancy (2% - 46%)
  - ➢ Occupancy is inversely proportional to the SMEM
- Higher SMEM usage causes less CTAs to be launched and thus lower TLP



SMEM Limits Performance

- Doubling SMEM can improve performance up to 2.8×
- SMEM can be very critical, and enlarging SMEM can be promising to improve GPU TLP and performance



Extended SMEM Supplied by Idle L2 Cache

- On-chip memory is expensive
  - ➢ Extended shared memory can be borrowed from idle L2 cache
- Huge L2 cache is in idle state

metric	bandwidth (%)	space (%)	working set(MB)
Average	8.3	71.8	4.3

Space is defined as percentage of touched cache lines

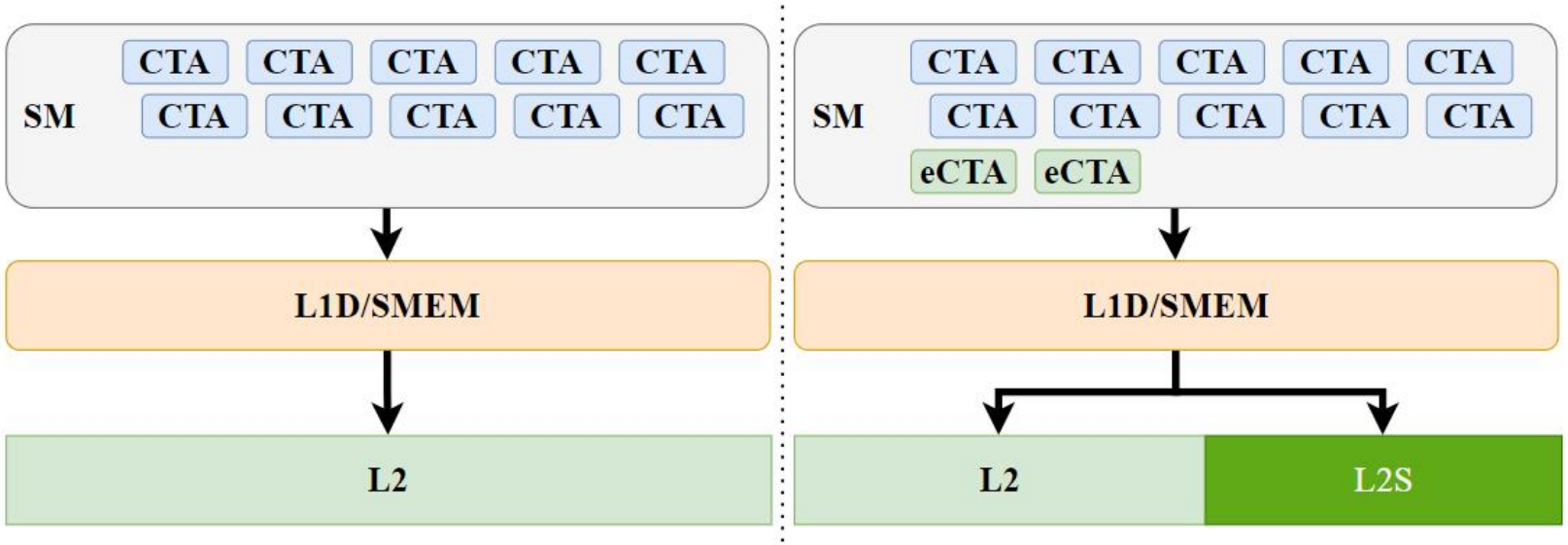
- L2 cache can be partitioned to be extended SMEM

Agenda

- Background
- Motivation
- Design
- Evaluation
- Summary

Architecture of $MILE

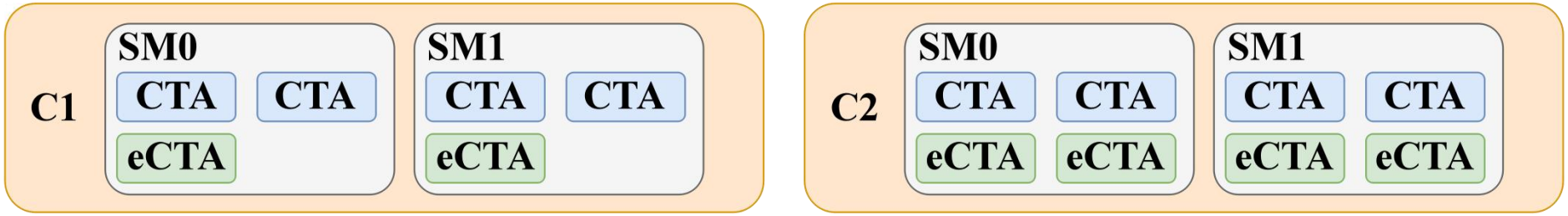
- L2 cache is partitioned to extend SMEM (L2S)
- Extra CTAs (eCTAs) are launched to each SM
- SMEM accesses of eCTAs are redirected to L2S



Baseline architecture vs. SMILE architecture

RPG — Runtime Profiling Guided Method

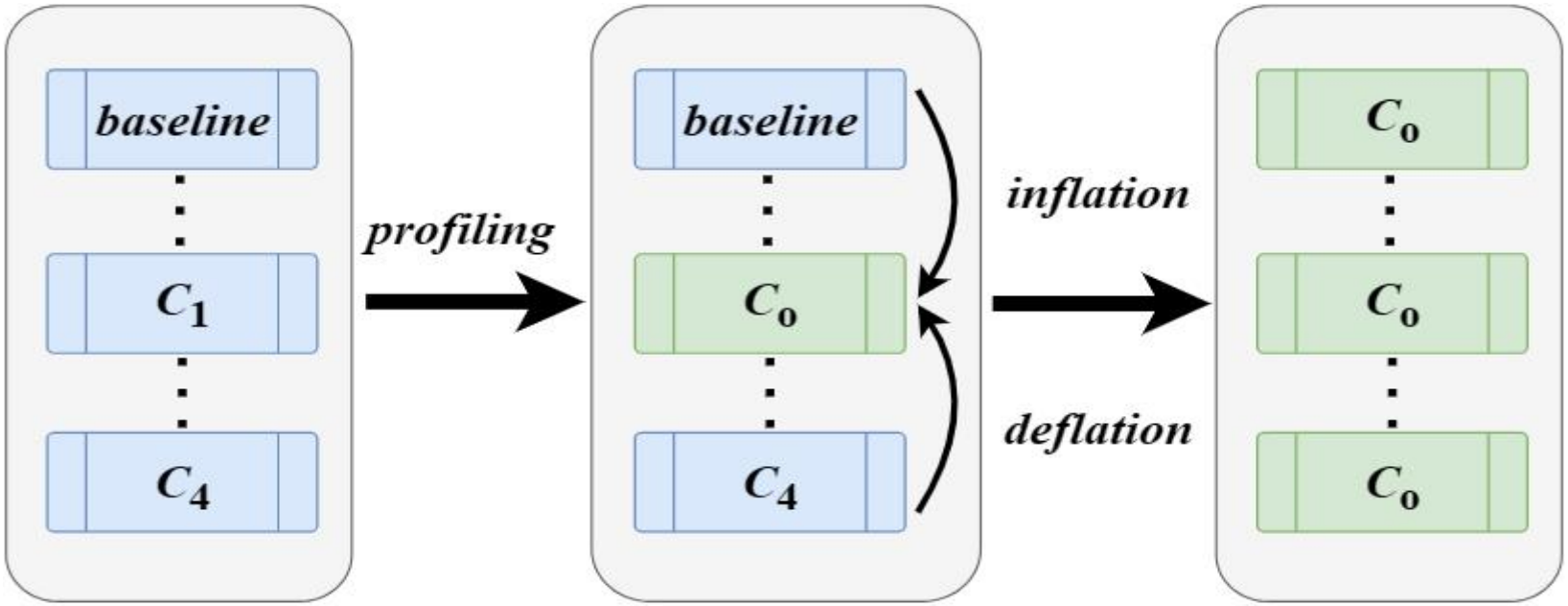
- Different number of eCTAs influence the performance of apps
- RPG is proposed to determine the best quota of eCTAs
- SMs are grouped (C1-C4) to profile



- SMs under different groups commit CTAs in varied speed
- Adjust #eCTAs to the group which commits CTAs “Fastest”

Workflow of RPG

- RPG contains two phases
  - ➢ profiling: collects the number of CTAs committed by different groups
  - ➢ alignment: adjusts the number of concurrently running eCTAs



Agenda

- Background
- Motivation
- Design
- Evaluation
- Summary

GPU Configurations, Applications and SOTA

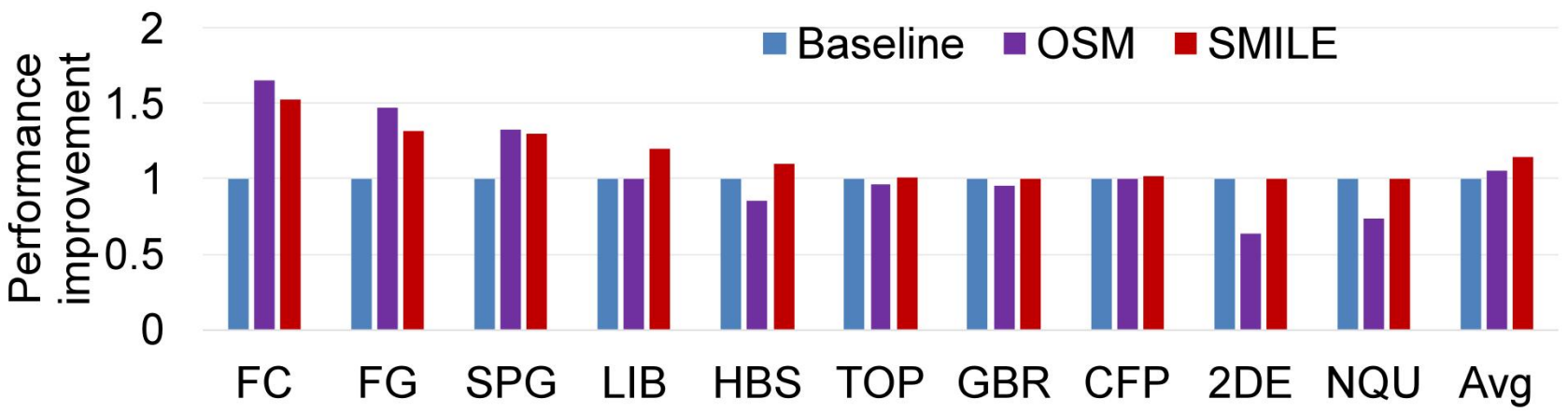
- Simulation platforms: Accel-Sim and AccelWattch to model an NVIDIA Ampere-like GPU
- Applications: involve SMEM usages and TLP are limited by SMEM
- SOTA: OSM using off-chip device memory to enlarge SMEM

Parameter	Value
#SM	80
Shared Memory Cache / SM	100KB
L1 cache / SM	28KB
L2 (or LLC)	30MB

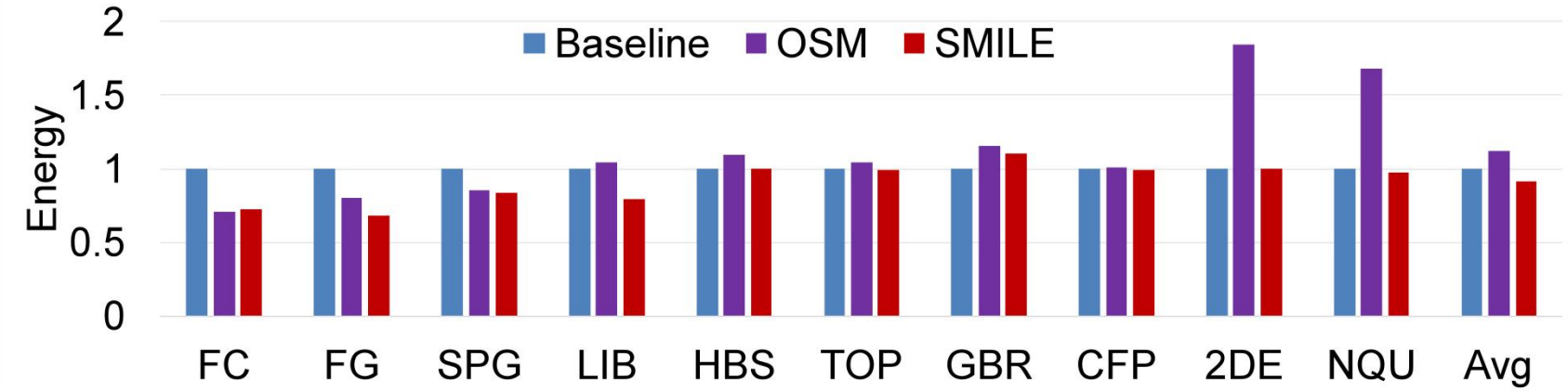
#App	benchmark
3	Rodinia
6	Cutlass
1	N/A

Performance and Energy

- Some apps are SMEM-hungry, like FC, FG and SPG
- OSM performs poorly in SMEM-insensitive apps, like 2DE and NQU



+14%



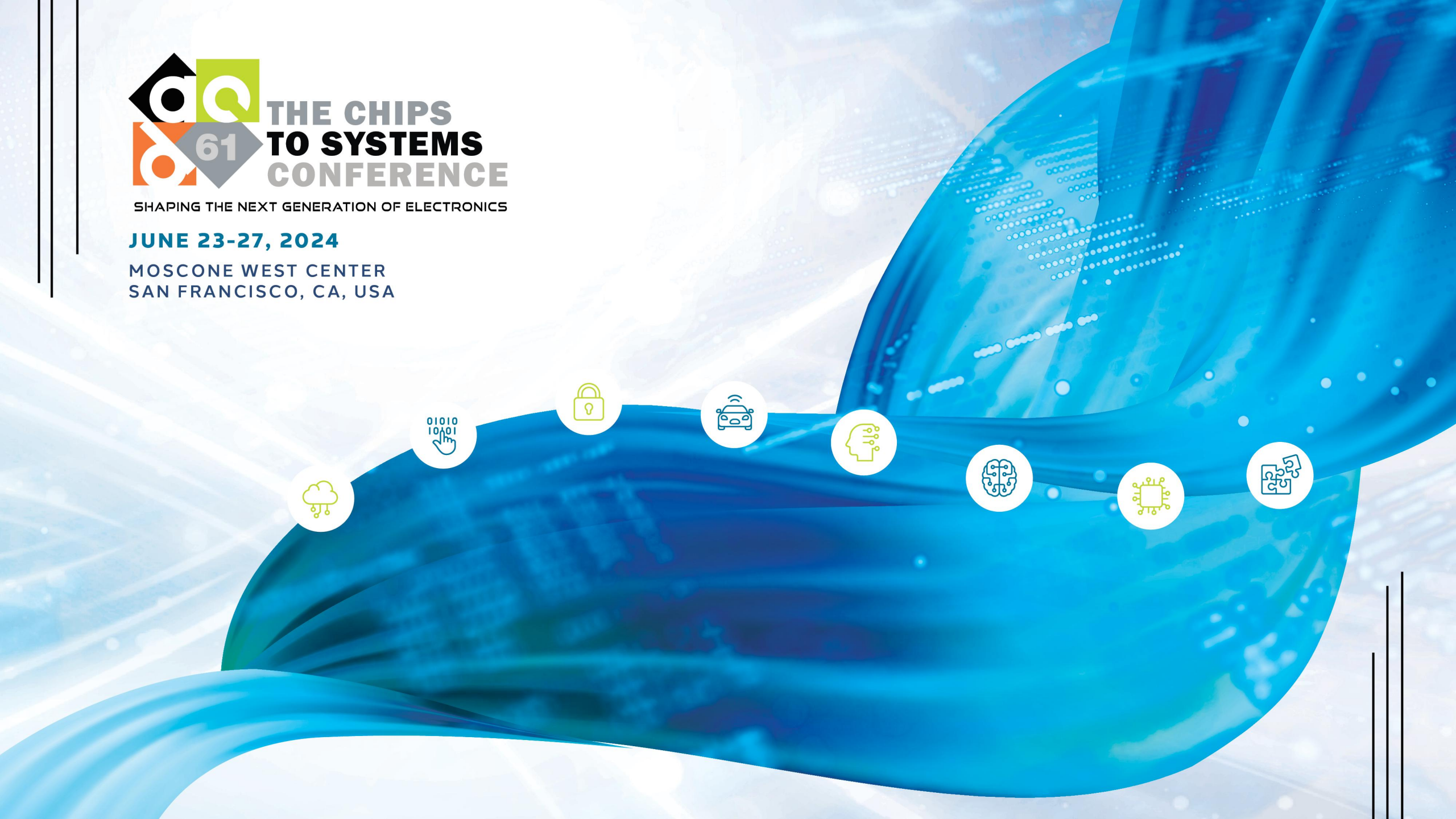
-9%

Agenda

- Background
- Motivation
- Design
- Evaluation
- Summary

Summary

- Motivation: enlarged SMEM helps boost GPU TLP, and GPUs are of huge LLC
- Design: partition LLC to expand SMEM to launch extra CTAs
  - ➢ Runtime profiling
- Result: proposed design efficiently raises TLP
  - ➢ Speedup by 14%, energy reduction by 9%


61
THE CHIPS
TO SYSTEMS
CONFERENCE
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA