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

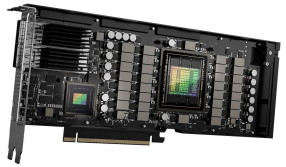
$MILE : LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism

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

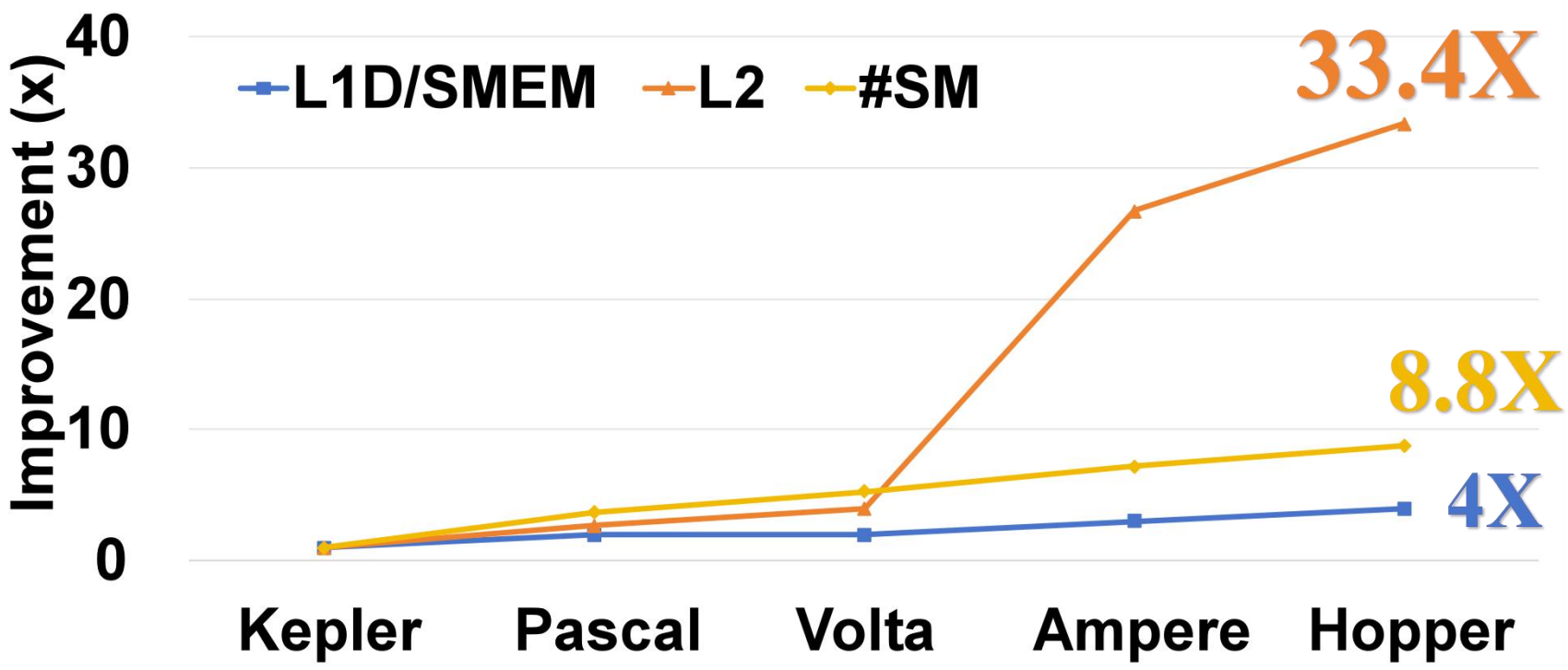
Sun Yat-sen University

Catalogue

- GPU Evolution
- Motivation
- Design
- Evaulation
- Summary

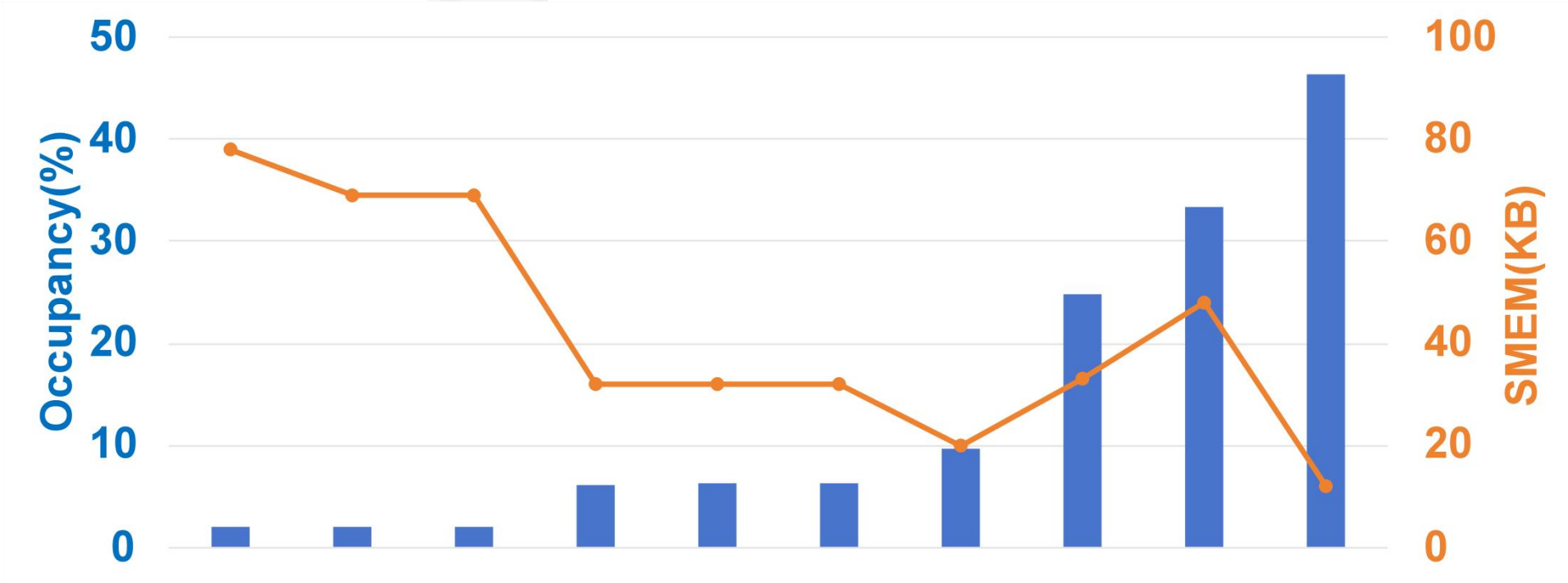


GPU Evolution



- Improvement of L1D/SMEM, L2 and the number of SMs are out of proportion
- L1D/SMEM is insufficient while LLC is abundant

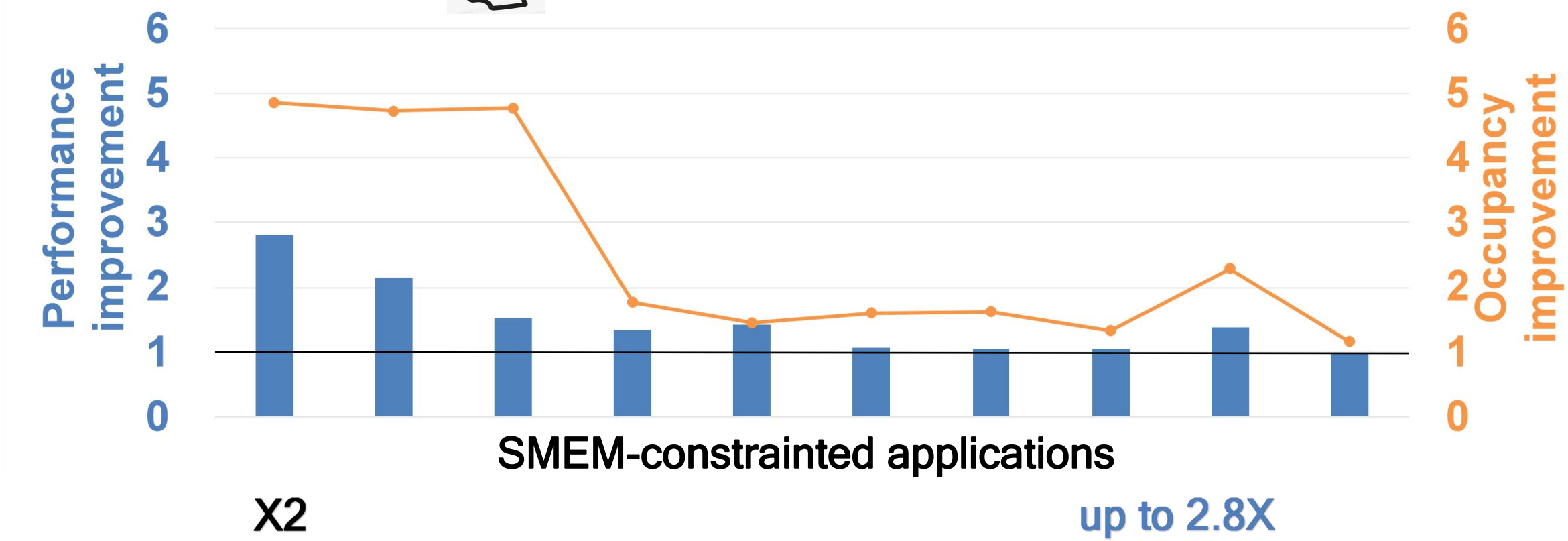
Motivation



SMEM-constrainted applications

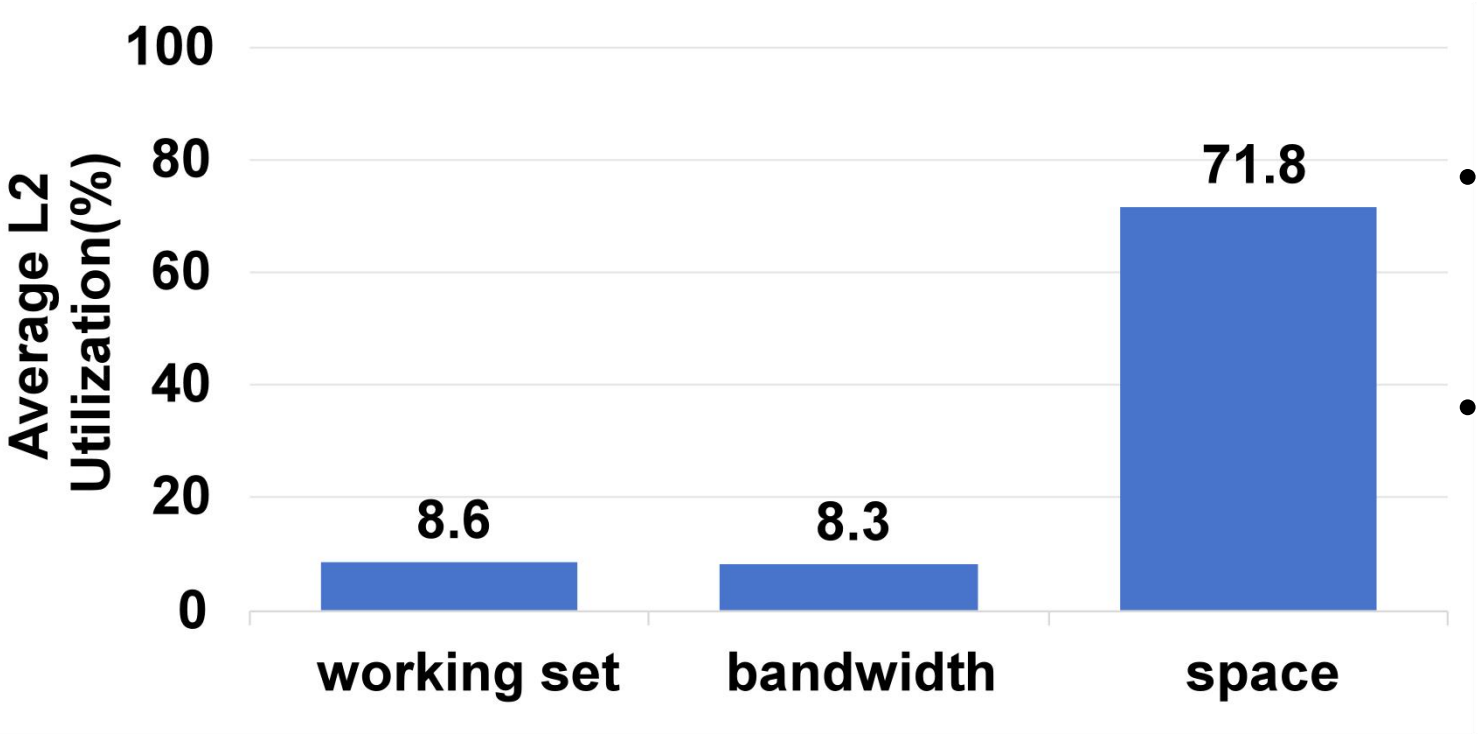
- Applications are exhibiting low occupancy (2% - 46%), which is inversely proportional to the SMEM
- Higher SMEM usage causes less CTAs to be launched and thus lower TLP

Motivation



- Doubling Shared Memory can Vastly Improve Performance
- SMEM can be very critical, and enlarging SMEM can be promising to improve GPU TLP and performance.

Motivation

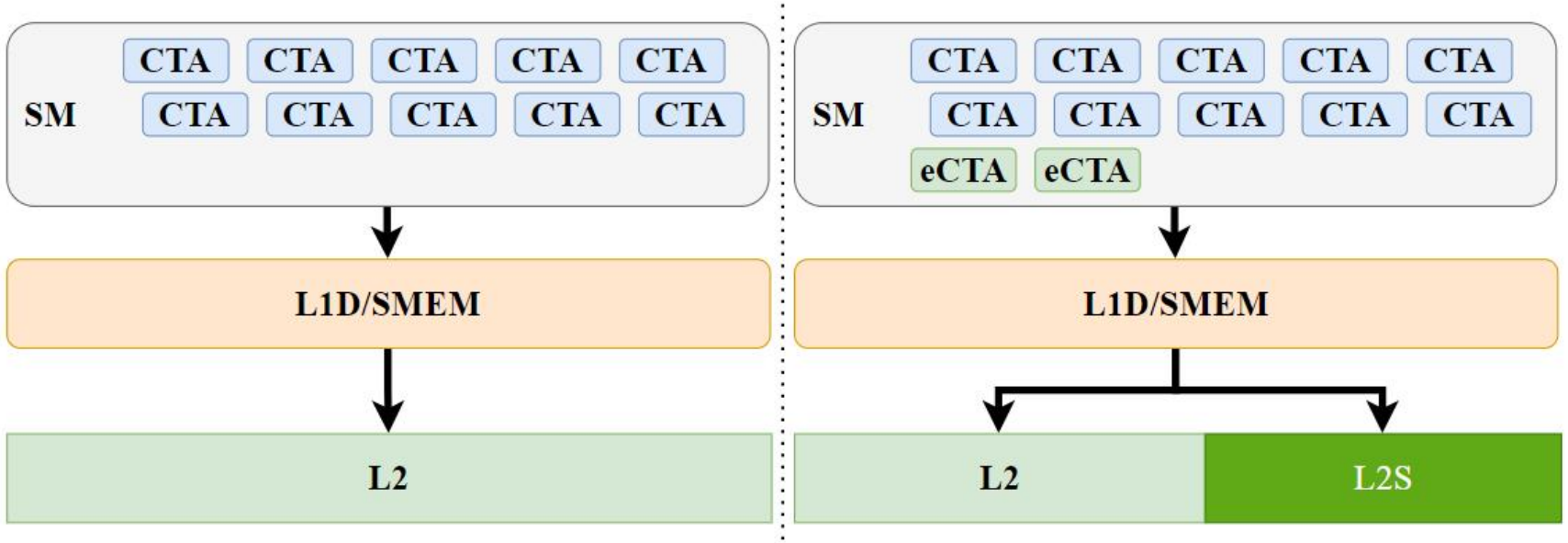


- Working set is compared to Hopper's L2 (50MB)
- Space is defined as percentage of touched cache lines

- L2 cache is in idling state
- Combine with SMEM is helpful to solve GPU low occupancy, it is thus natural to integrate both to improve TLP and reduce L2 idling simultaneously

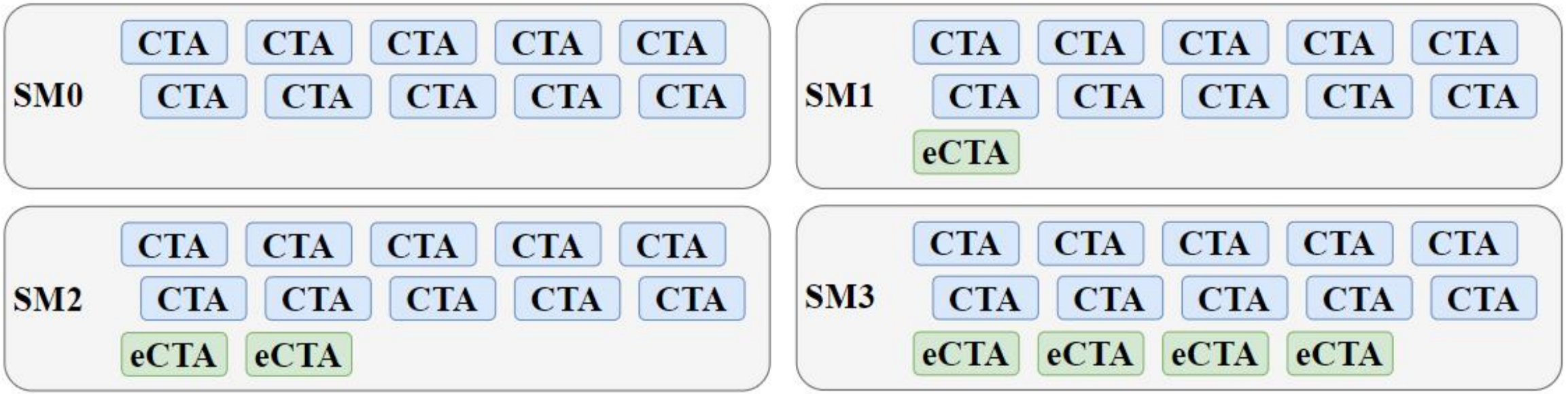
Design– $MILE

Baseline architecture VS SMILE architecture



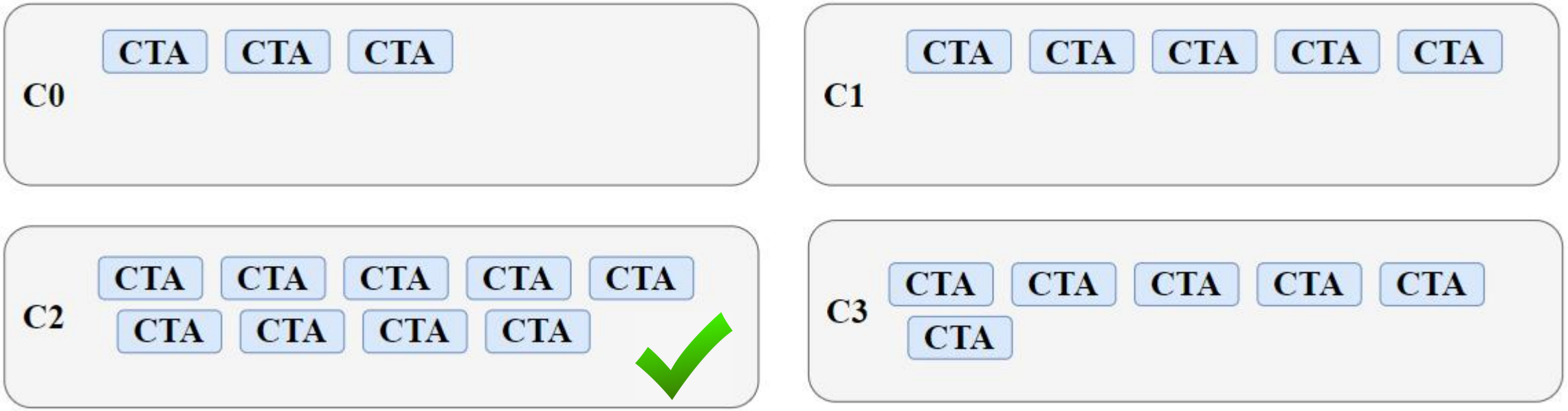
- Extra CTAs (eCTAs) are launched to each SM
- L2 cache is partitioned as extended SMEM (L2S)
- SMILE redirect eCTA's SMEM accesses to L2S

Design– $MILE



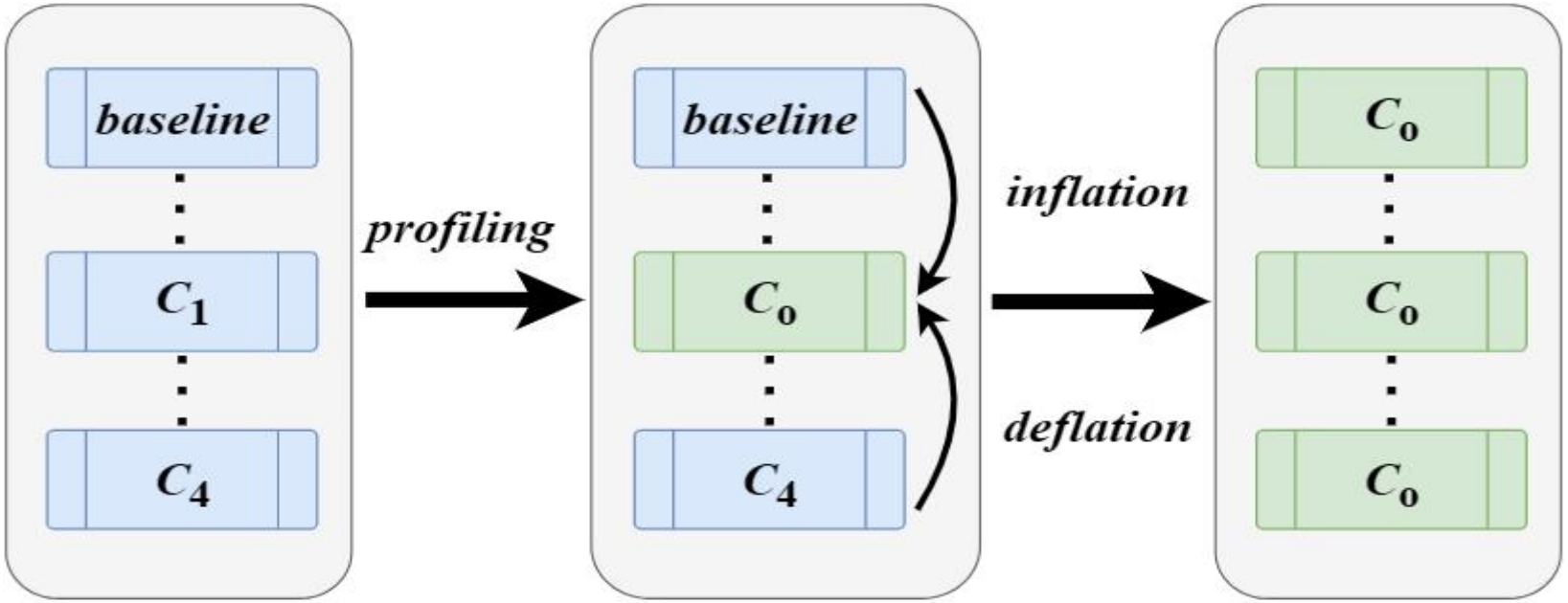
- The best amount of eCTA is application-aware
- To determine the best quota of eCTA, we propose Runtime Profiling Guided (RPG) where SMs are grouped (C0-C4) to profile

Design– $MILE



- SMs under different configs commit CTAs in varied speed
- Select Config which commit CTAs “Fastest”

Design– $MILE



- RPG encompasses profiling phase (collects the number of CTAs committed by different groups) and alignment phase (adjust the number of concurrently running eCTAs)

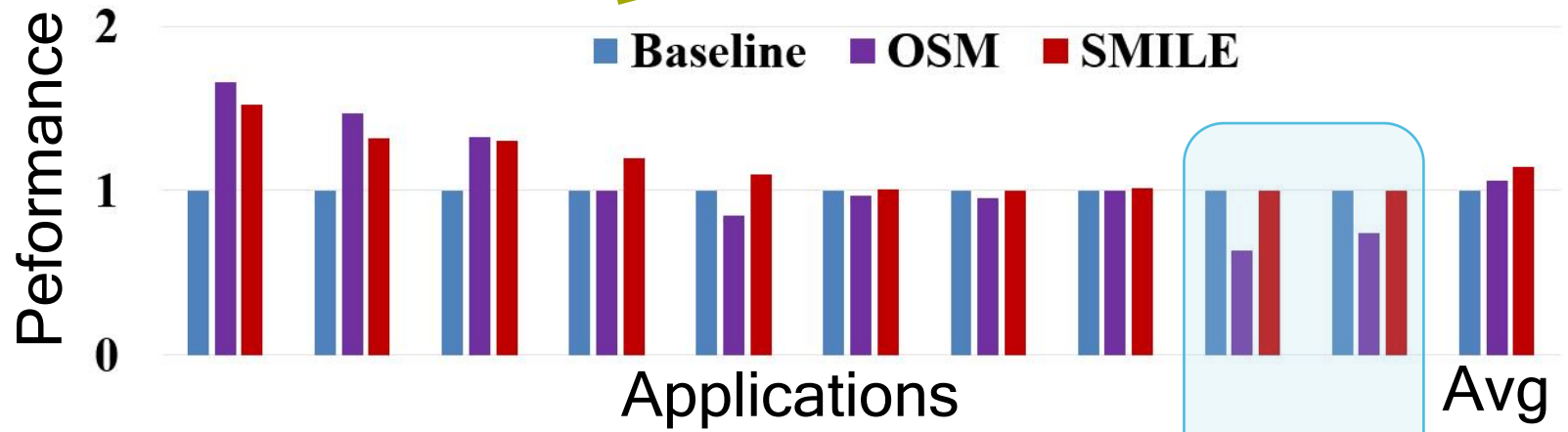
Evaluation

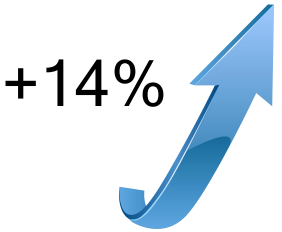
GPU configurations and applications

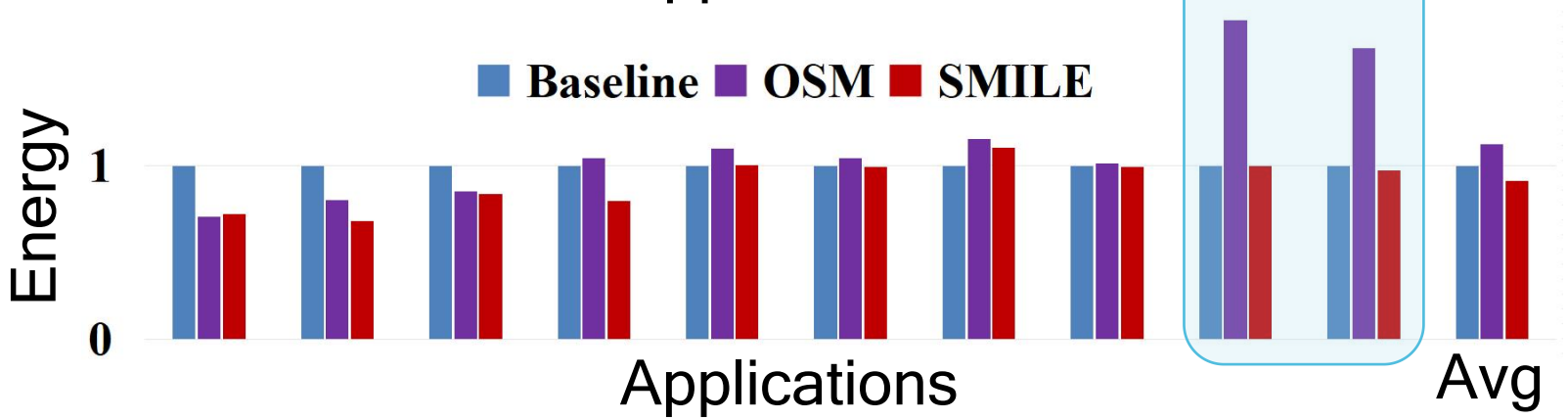
Parameter	Value
#SM	80
Scheduler	LRR
Register File Size / SM	256KB
Shared Memory Cache / SM	100KB
Core clock	1132MHz
Schedulers / SM	4
L1 cache / SM	28KB
L2 (or LLC)	30MB

App	benchmark
N-queen	Rodinia
LIBOR	
Hybridsort	
Fused gemms	Cutlass
Fused convs	
Gemm bias relu	
Conv2dfprop	
Tensorop	
Sparse gemm	
2Dentropy	NA

Evaluation

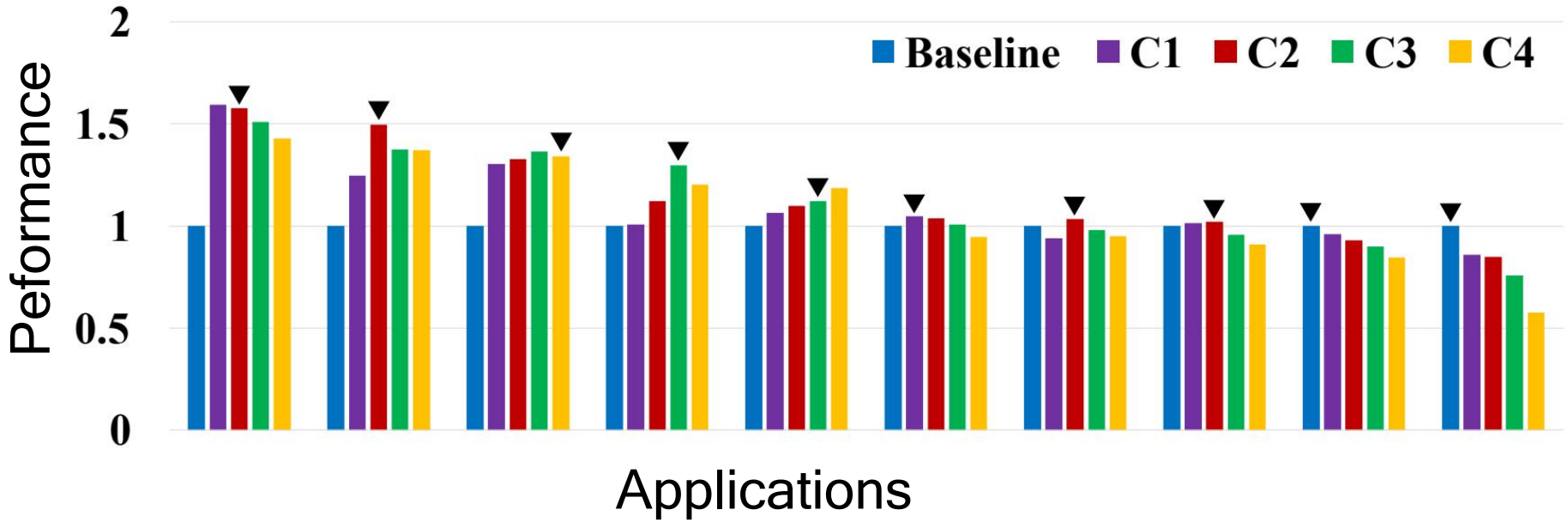






- SMILE outperforms OSM in critical apps by avoiding great performance reduction and energy consumption

Evaluation



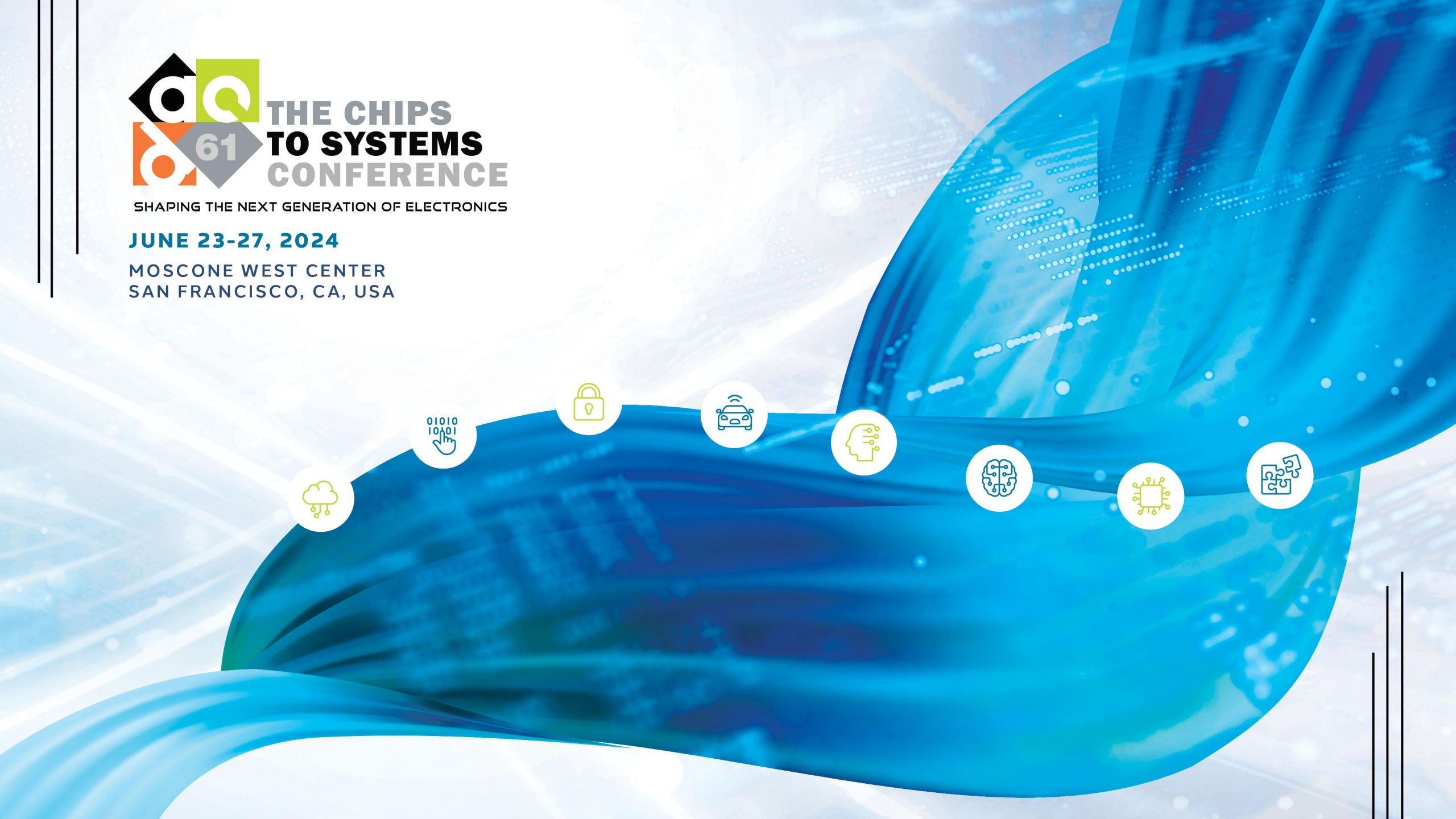
- Profiling Quality of RPG: the choices of SMILE mostly locate on the highest speedup bar, confirming the online profiling accuracy.

Summary

- GPU TLP can be bounded by the deficient SMEM, and are motivated to divide the increasingly large LLC for SMEM expansion

- Through light-weight runtime profiling, SMILE is capable to decide the reasonable partition ratio, and effectively enables extra CTAs to be launched

- SMILE remarkably raises the TLP and accelerates the executions, effectively outperforming the state-of-the-arts


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