Understanding GPU Resource Interference One Level Deeper

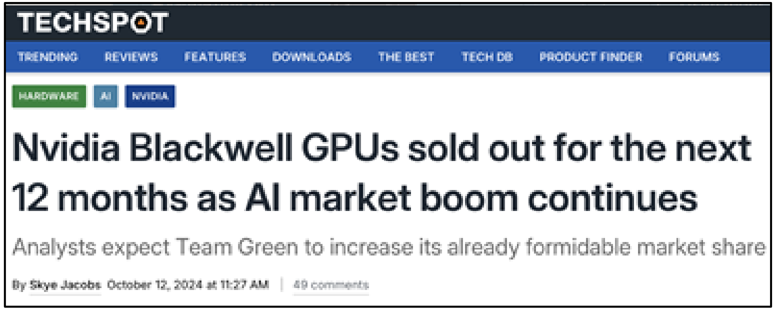
Paul Elvinger[1], Foteini Strati[1], Natalie Enright Jerger[2], Ana Klimovic[1]
[1]ETH Zurich, [2]University of Toronto

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry



TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

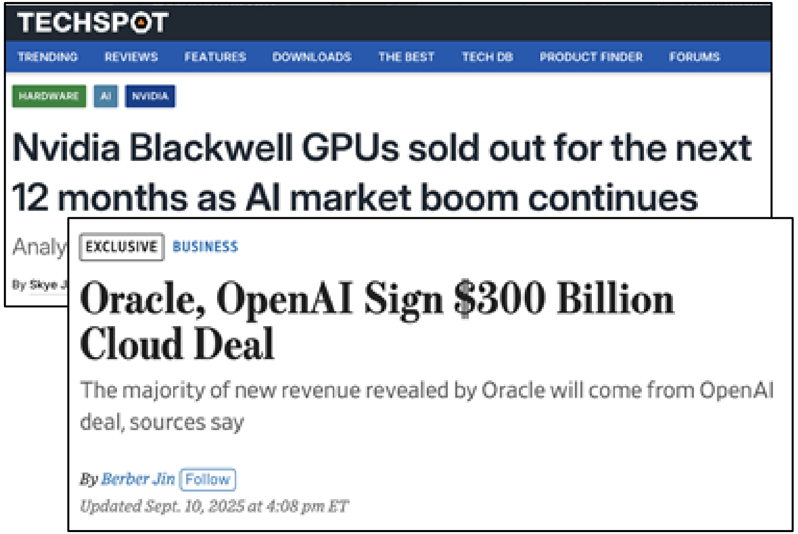
Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analysts expect Team Green to increase its already formidable market share

By Skye Jacobs October 12, 2024 at 11:27 AM | 49 comments

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry



TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analy

By Skye J

EXCLUSIVE BUSINESS

Oracle, OpenAI Sign $300 Billion Cloud Deal

The majority of new revenue revealed by Oracle will come from OpenAI deal, sources say

By *Berber Jin* Follow

Updated Sept. 10, 2025 at 4:08 pm ET

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry

TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analy

By Skye J

EXCLUSIVE BUSINESS

Oracle, OpenAI Sign $300 Billion Cloud Deal

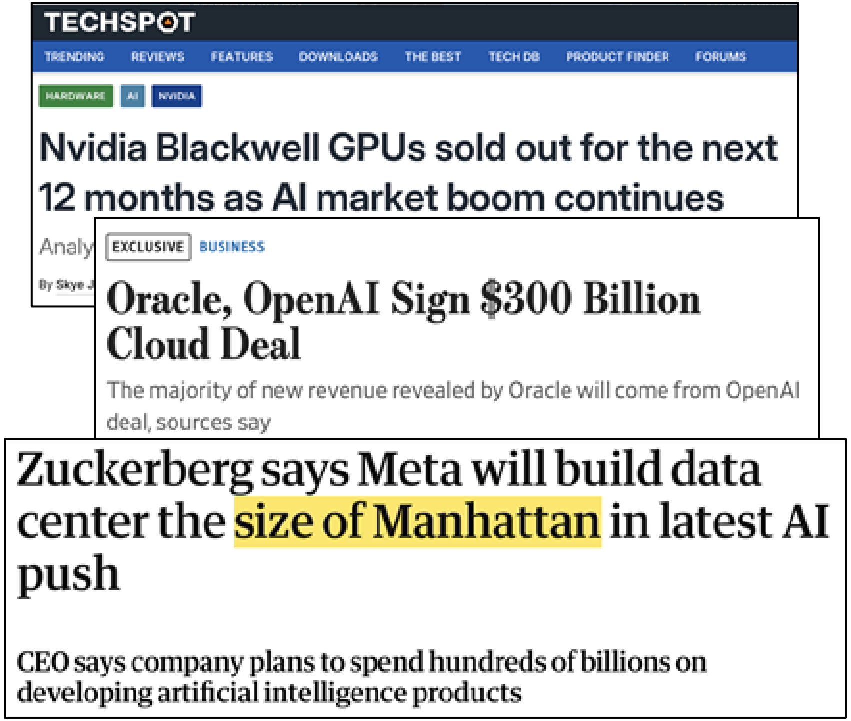
The majority of new revenue revealed by Oracle will come from OpenAI deal, sources say

Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

CEO says company plans to spend hundreds of billions on developing artificial intelligence products

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry


TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analy

By Skye J

EXCLUSIVE BUSINESS

Oracle, OpenAI Sign $300 Billion Cloud Deal

The majority of new revenue revealed by Oracle will come from OpenAI deal, sources say

Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

CEO says company plans to spend hundreds of billions on developing artificial intelligence products

But cloud providers report poor utilization…

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry

TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analy

By Skye J

EXCLUSIVE BUSINESS

Oracle, OpenAI Sign $300 Billion Cloud Deal

The majority of new revenue revealed by Oracle will come from OpenAI deal, sources say

Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

CEO says company plans to spend hundreds of billions on developing artificial intelligence products

But cloud providers report poor utilization...

An Empirical Study on Low GPU Utilization of Deep Learning Jobs

Yanjie Gao, Microsoft Research, Beijing, China, yanjga@microsoft.com

Yichen He[*], Microsoft Research, Beijing, China, v-yichenhe@microsoft.com

Xinze Li[*], Peking University, Beijing, China, panzer_marching@163.com

Bo Zhao[*], Microsoft Research, Beijing, China, v-bozha@microsoft.com

Haoxiang Lin[†], Microsoft Research, Beijing, China, haoxlin@microsoft.com

Yoyo Liang, Microsoft, Beijing, China, yoliang@microsoft.com

Jing Zhong, Microsoft, Beijing, China, jinzhong@microsoft.com

Hongyu Zhang, Chongqing University, Chongqing, China, hyzhang@cqu.edu.cn

Jingzhou Wang[*], Tsinghua University, Beijing, China, jz-wang20@mails.tsinghua.edu.cn

Yonghua Zeng, Microsoft, Beijing, China, yozen@microsoft.com

Keli Gui, Microsoft, Beijing, China, keligui@microsoft.com

Jie Tong, Microsoft, Beijing, China, jietong@microsoft.com

Mao Yang, Microsoft Research, Beijing, China, maoyang@microsoft.com

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry

TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

HARDWARE AI NVIDIA

Nvidia Blackwell GPUs sold out for the next 12 months as AI market boom continues

Analy...

By Skye J...

EXCLUSIVE BUSINESS

Oracle, OpenAI Sign $300 Billion Cloud Deal

The majority of new revenue revealed by Oracle will come from OpenAI deal, sources say

Zuckerberg says Meta will build data center the size of Manhattan in latest AI push

CEO says company plans to spend hundreds of billions on developing artificial intelligence products

But cloud providers report poor utilization…

An Empirical Study on Low GPU Utilization of Deep Learning Jobs

Yanjie Gao, Microsoft Research, Beijing, China, yanjga@microsoft.com
Yichen He*, Microsoft Research, Beijing, China, v-yichenhe@microsoft.com
Xinze Li*, Peking University, Beijing, China, panzer_marching@163.com
Bo Zhao*, Microsoft Research, Beijing, China, v-bozha@microsoft.com
Haoxiang Lin†, Microsoft Research, Beijing, China, haoxlin@microsoft.com
Yoyo Liang, Microsoft, Beijing, China, yoliang@microsoft.com
Jing Zhong, Microsoft, Beijing, China, jinzhong@microsoft.com
Hongyu Zhang, Chongqing University, Chongqing, China, hyzhang@cqu.edu.cn
Jingzhou Wang*, Tsinghua University, Beijing, China
Yonghua Zeng, Microsoft, Beijing, China
Keli Gui, Microsoft, Beijing, China
Jie Tong, Microsoft, Beijing, China

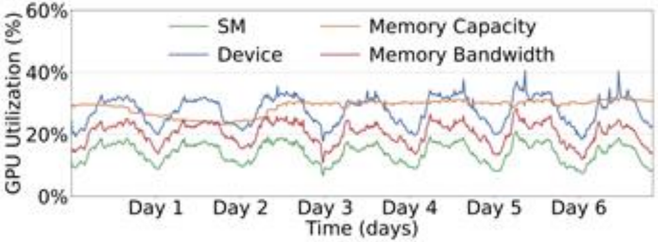


Figure 1. *GPU utilization metrics over a week in a production Ads inference service at Meta.*

The GPU Underutilization Paradox

GPUs are scarce, expensive and power-hungry

TECHSPOT

TRENDING REVIEWS FEATURES DOWNLOADS THE BEST TECH DB PRODUCT FINDER FORUMS

developing artificial intelligence products

But cloud providers report poor utilization…

An Empirical Study on Low GPU Utilization of Deep Learning Jobs

We should first operate existing clusters more efficiently!

Reasons for GPU underutilization

- **Small batch sizes** in inference due to SLOs [1]
- Input data preprocessing and **ingestion stalls** [2]
- **Communication bottlenecks** in distributed training [3]
- **Differences in resource requirements** (e.g. compute/memory) in the same workload [4,5]

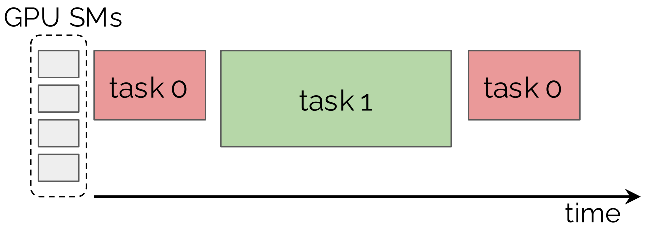
[1] Gujarati et al. Serving DNNs like Clockwork: Performance Predictability from the Bottom Up. OSDI'20
[2] Murray et al. tf.data: A Machine Learning Data Processing Framework. VLDB'21
[3] Peng et al. A generic communication scheduler for distributed DNN training acceleration. SOSP'19
[4] Strati et al. Orion: Interference-aware, Fine-grained GPU Sharing for ML Applications. EuroSys'24
[5] Kamath et al. POD-Attention: Unlocking Full Prefill-Decode Overlap for Faster LLM Inference. ASPLOS'25

Sharing GPUs across workloads as promising solution

Sharing GPUs across workloads as promising solution

Temporal Sharing

Time-slice the GPU.



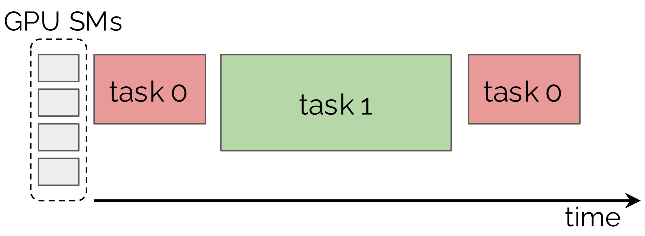
✅ **Fill idle times** with other workloads

❌ Workloads may still **not fully saturate GPU**

Sharing GPUs across workloads as promising solution

Temporal Sharing

Time-slice the GPU.

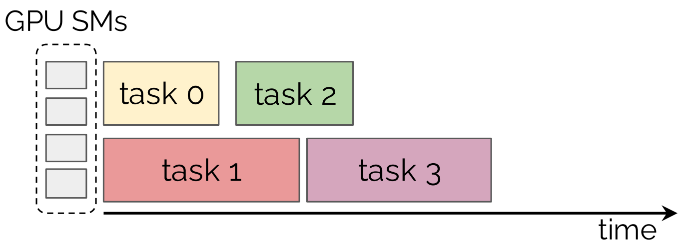


✅ **Fill idle times** with other workloads

❌ Workloads may still **not fully saturate GPU**

Spatial Sharing

Overlap kernels on the GPU (CUDA streams, MPS or MIG on NVIDIA)



✅ Better utilization

❌ Colocation **can lead to interference** and **unpredictable slowdowns** dangerous for latency critical applications

Sharing GPUs across workloads as promising solution

Temporal Sharing

Time-slice the GPU.

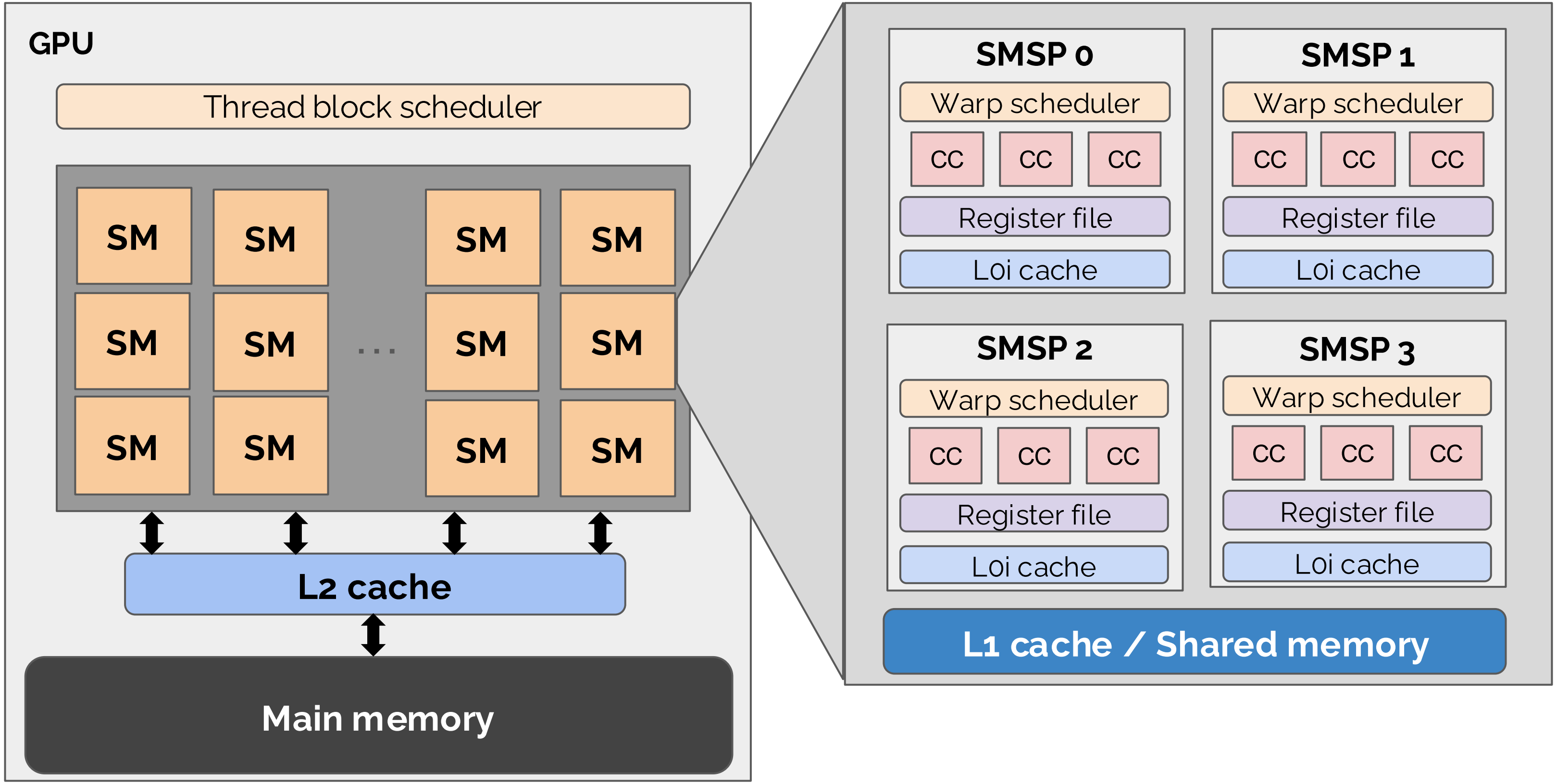
Spatial Sharing

Overlap kernels on the GPU (CUDA streams,

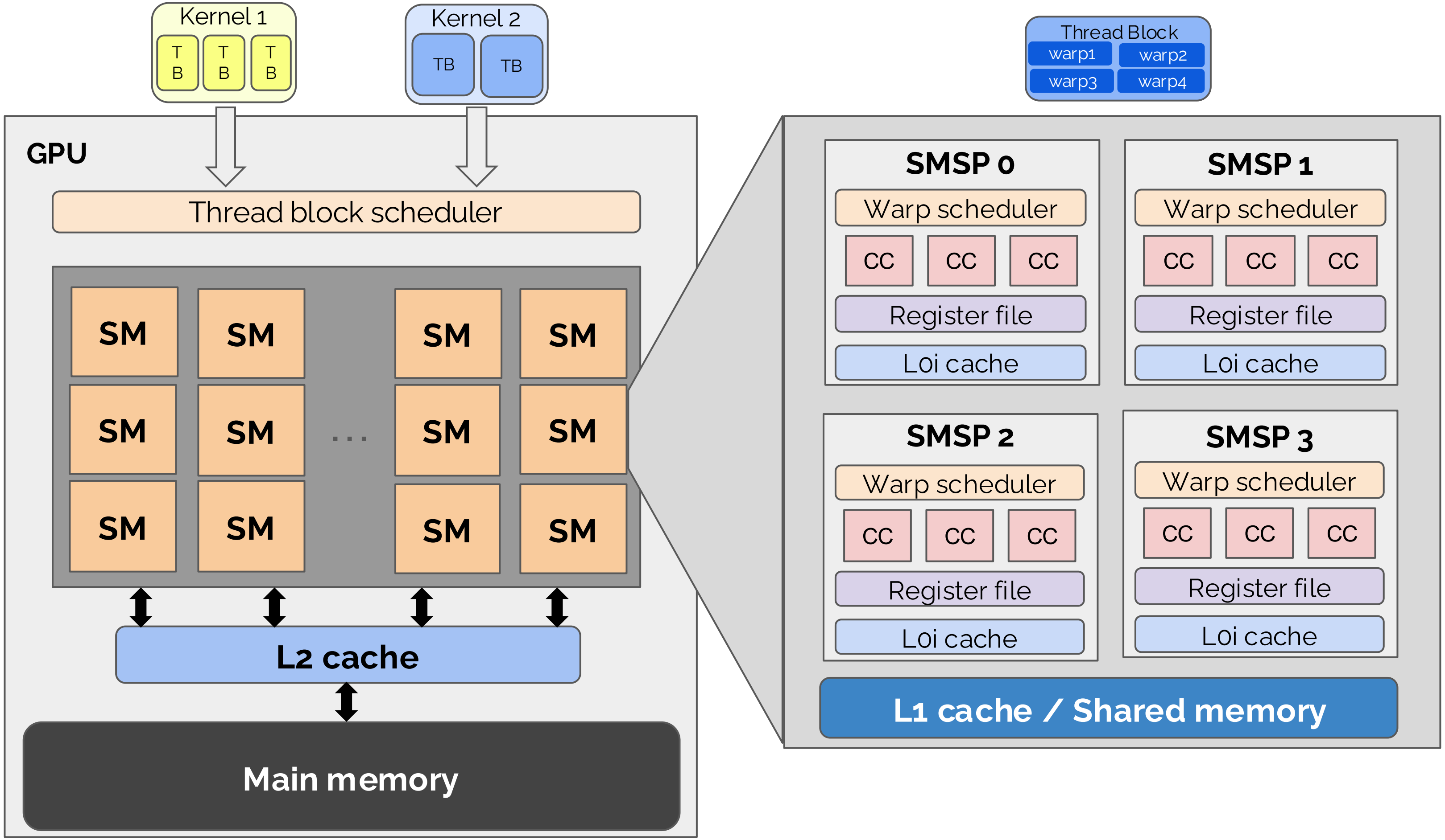
Our main problem today…

We lack a deep understanding of interference from spatial colocation.🤔

Why is predicting interference so hard?



Why is predicting interference so hard?



Why do existing approaches fall short?

Single or coarse metrics cannot capture the entire interference landscape.

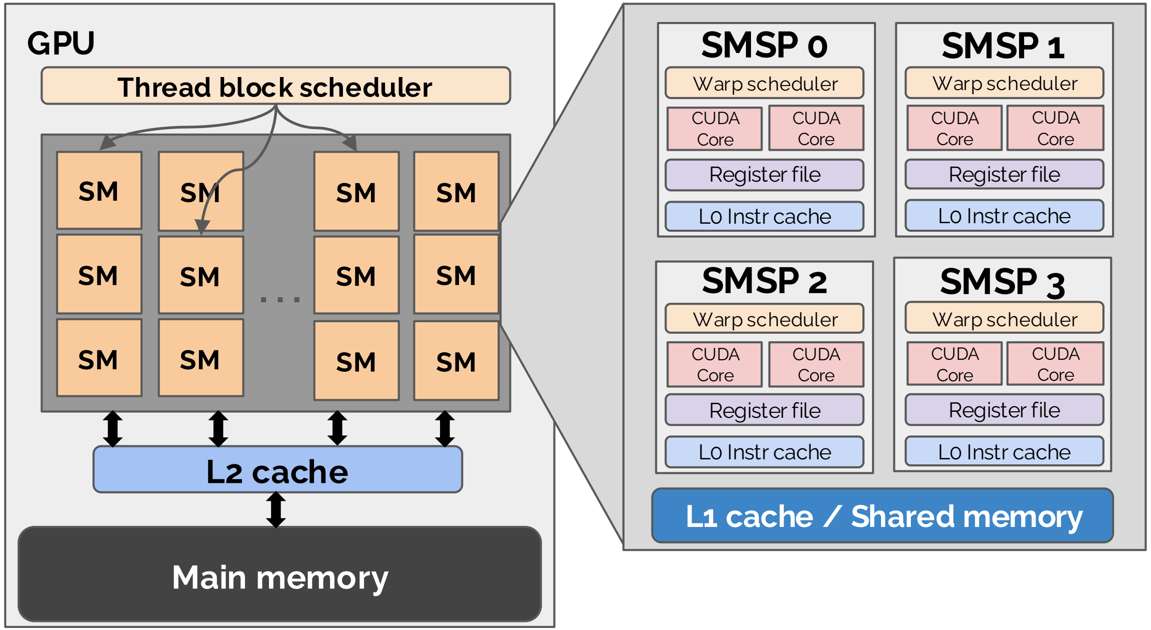
	Thread Block Scheduler	L2 Cache	Memory Bandwidth	Warp Scheduler	CUDA Cores	L1 Cache / Shared Memory
Usher [OSDI'24]	✅	❌	✅	❌	❌	❌
Orion [EuroSys'24]	✅	❌	✅	❌	✅	❌
Reef [OSDI'22]	✅	❌	❌	❌	❌	❌
iGniter [TPDS'22]	❌	✅	❌	✅	❌	❌
GPUlet [ATC'22]	❌	✅	✅	❌	❌	❌

✅ Directly or indirectly covered by the system
❌ System fails to cover this source of interference

To reason correctly about interference, we need...

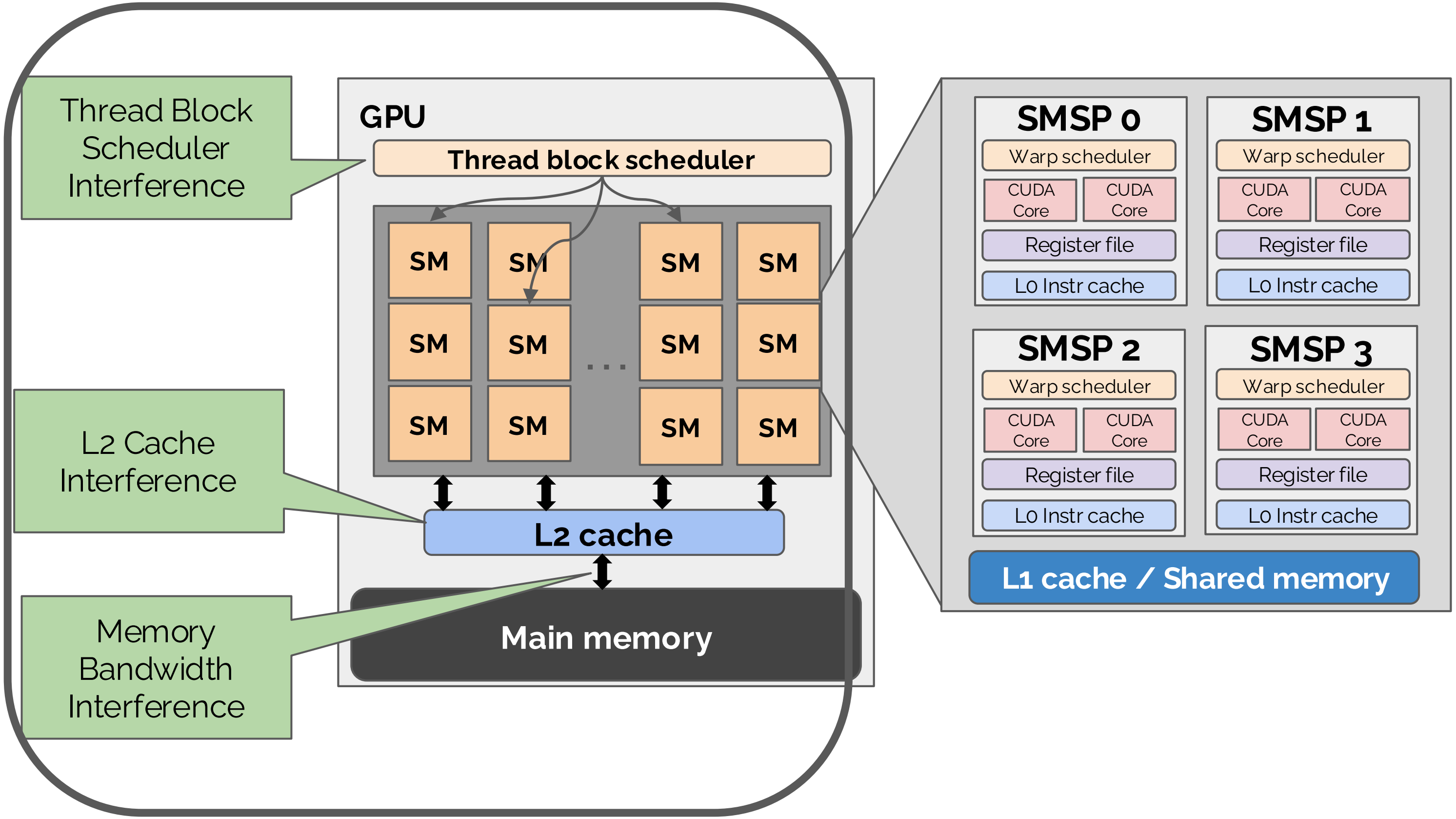
1. A complete view of all shared resources.
2. A **methodology to measure sensitivity** to each resource.

Sources of GPU interference

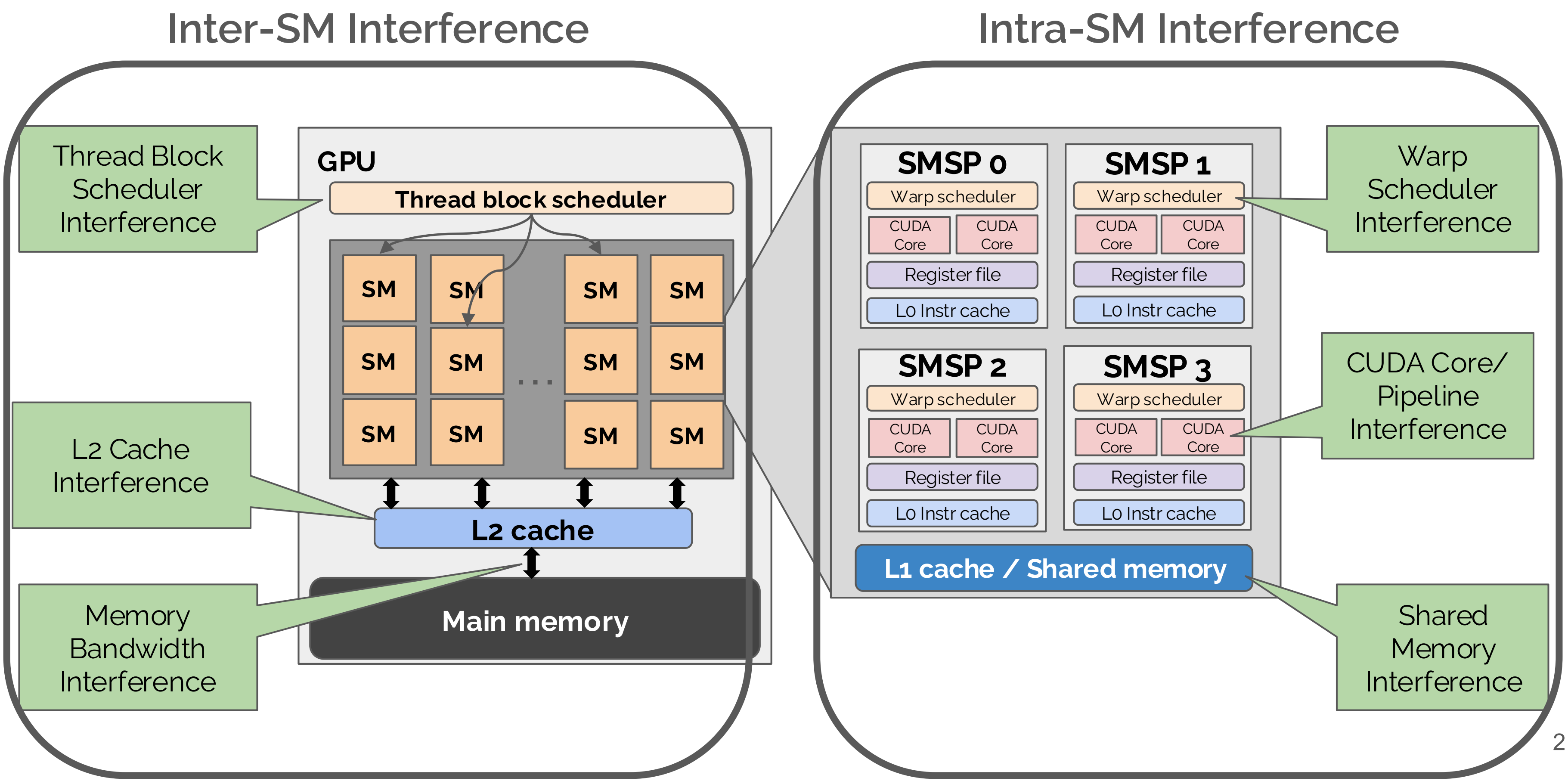


Sources of GPU interference

Inter-SM Interference



Sources of GPU interference



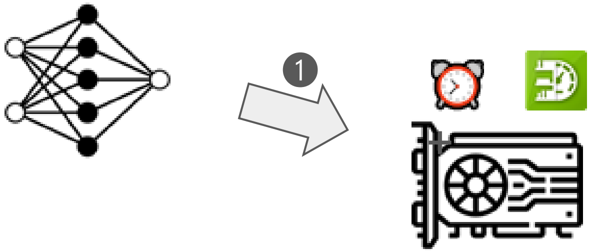
Methodology: Stressing One Resource at a Time

- **We open-source a suite of CUDA benchmarks** that each isolate and stress a single GPU resource [1,2].
- **We present a methodology for measuring workload sensitivity** by colocating workloads with these benchmarks.

[1] https://github.com/eth-easl/gpu-util-interference/tree/main
[2] https://github.com/eth-easl/vllm_profile/tree/main

Methodology: Stressing One Resource at a Time

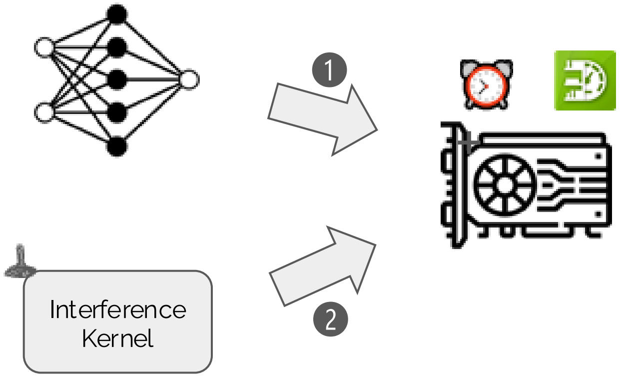
- **We open-source a suite of CUDA benchmarks** that each isolate and stress a single GPU resource [1,2].
- **We present a methodology for measuring workload sensitivity** by colocating workloads with these benchmarks.



[1] https://github.com/eth-easl/gpu-util-interference/tree/main
[2] https://github.com/eth-easl/vllm_profile/tree/main

Methodology: Stressing One Resource at a Time

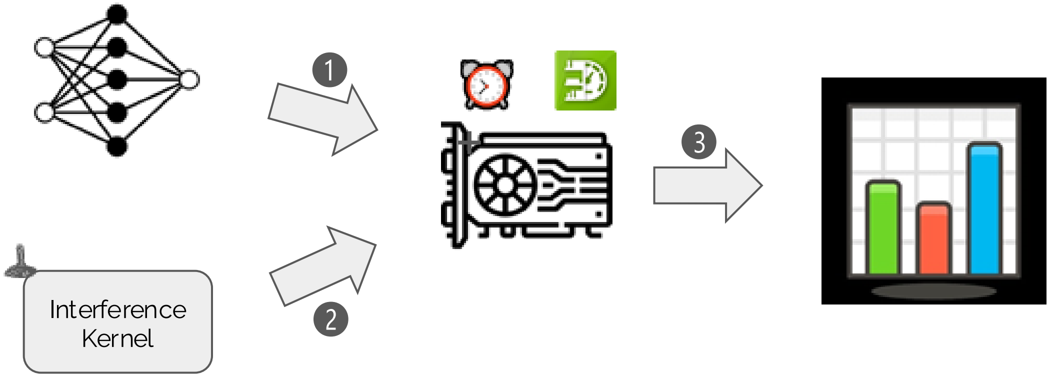
- **We open-source a suite of CUDA benchmarks** that each isolate and stress a single GPU resource [1,2].
- **We present a methodology for measuring workload sensitivity** by colocating workloads with these benchmarks.



[1] https://github.com/eth-easl/gpu-util-interference/tree/main
[2] https://github.com/eth-easl/vllm_profile/tree/main

Methodology: Stressing One Resource at a Time

- **We open-source a suite of CUDA benchmarks** that each isolate and stress a single GPU resource [1,2].
- **We present a methodology for measuring workload sensitivity** by colocating workloads with these benchmarks.



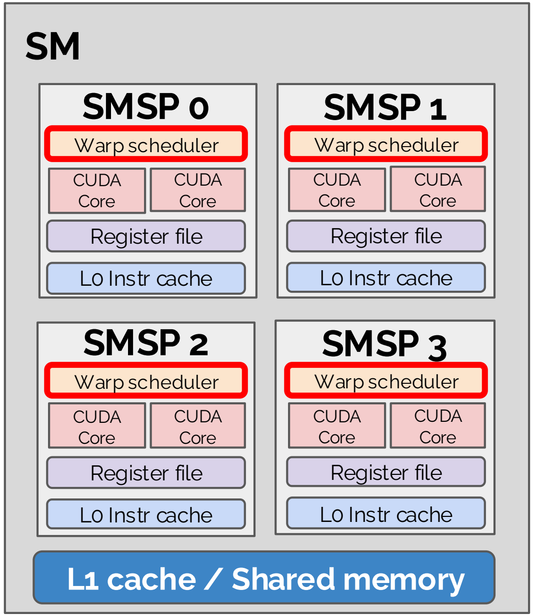
[1] https://github.com/eth-easl/gpu-util-interference/tree/main
[2] https://github.com/eth-easl/vllm_profile/tree/main

Intra-SM Interference

Interference within the Streaming Multiprocessor

Warp Scheduler Interference

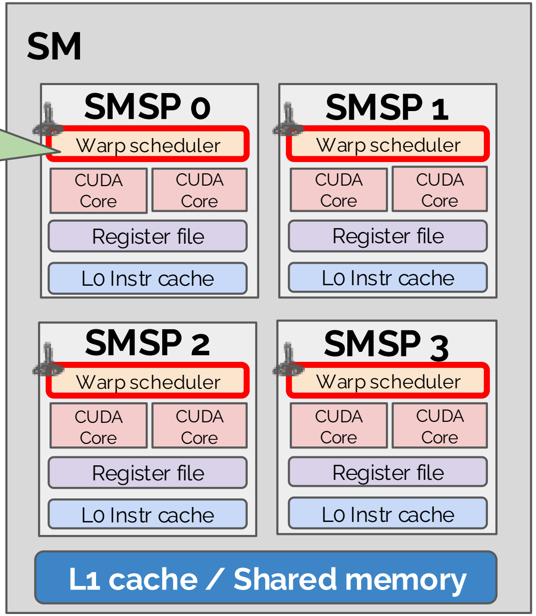
Warp scheduler schedules 1 warp (32 threads) per SMSP per cycle
=> max 4 instr/cycle/SM



Warp Scheduler Interference

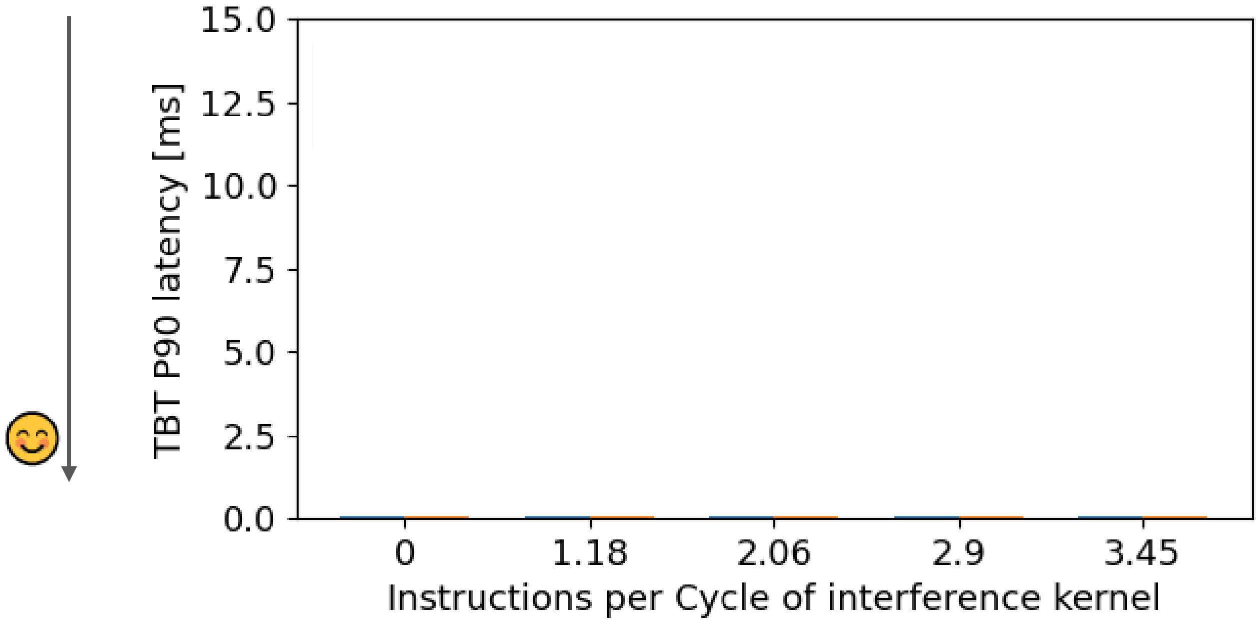
Warp scheduler schedules 1 warp (32 threads) per SMSP per cycle
=> max 4 instr/cycle/SM

Microbenchmark to **emit** a **high amount of instructions per cycle (IPC)**



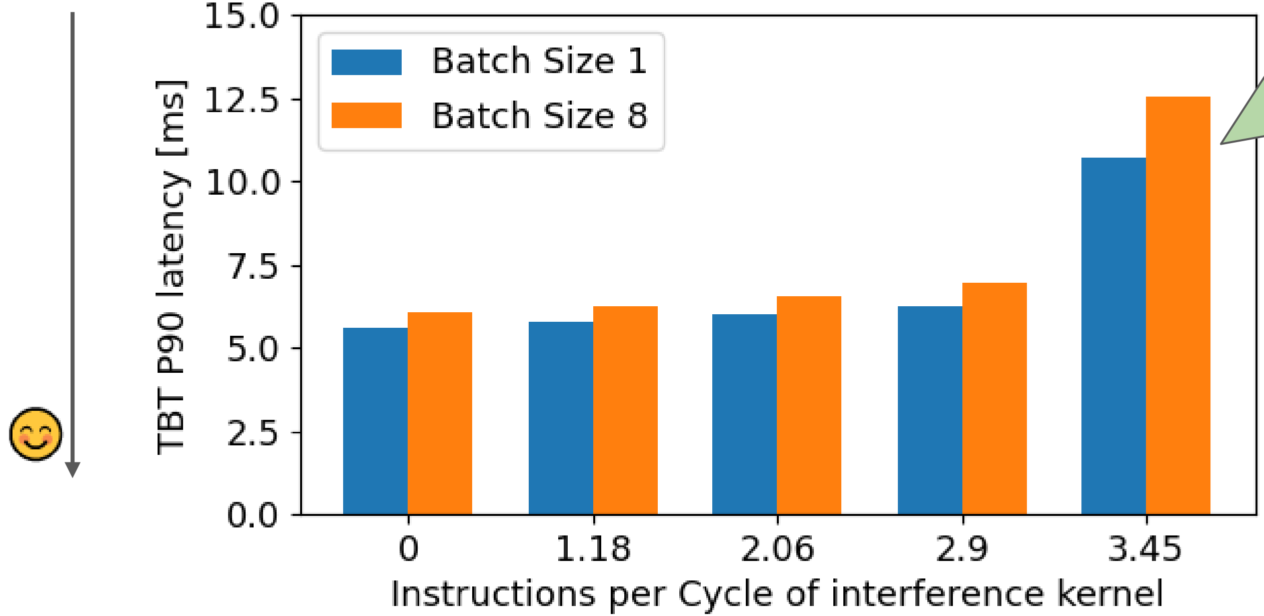
Warp Scheduler Interference

Gemma3-1B token generation with prompt size 1000 colocated with an **IPC intense microbenchmark** on a **NVIDIA RTX3090 GPU**.



Warp Scheduler Interference

Gemma3-1B token generation with prompt size 1000 colocated with an **IPC intense microbenchmark** on a **NVIDIA RTX3090 GPU**.



Colocation within an SM can be successful when handled with care

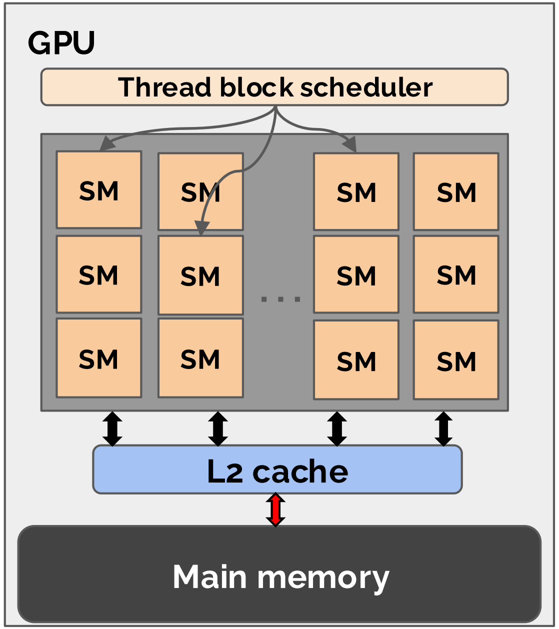
Why not just separate kernels to different SMs?

Inter-SM Interference

Interference across Streaming Multiprocessors

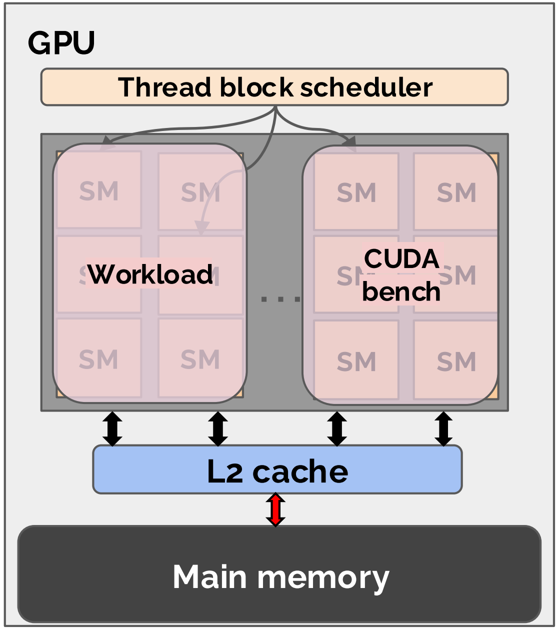
Memory/L2 Cache Bandwidth Interference

- Available **memory/l2 cache bandwidth is shared** across SMs



Memory/L2 Cache Bandwidth Interference

- Available **memory/l2 cache bandwidth is shared** across SMs
- SMs divided up into disjoint sets using CUDA Green Contexts [1] to avoid any source of interference within SM

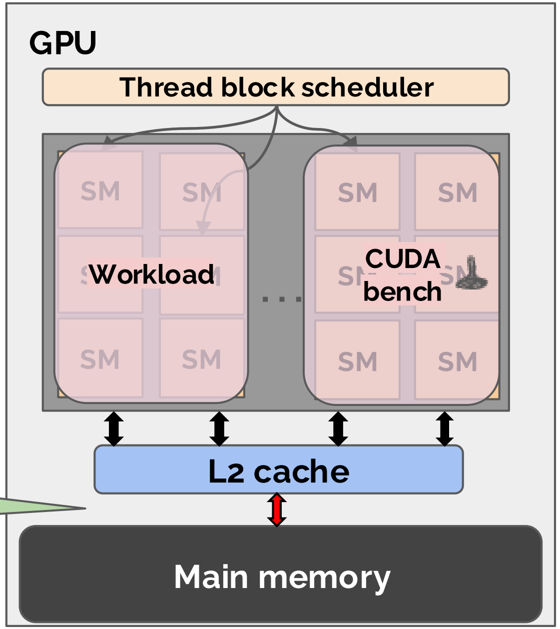


[1] CUDA Green Contexts

Memory/L2 Cache Bandwidth Interference

- Available **memory/l2 cache bandwidth is shared** across SMs
- SMs divided up into disjoint sets using CUDA Green Contexts [1] to avoid any source of interference within SM

Microbenchmark to **copy a lot of data within memory** using vectorized operations



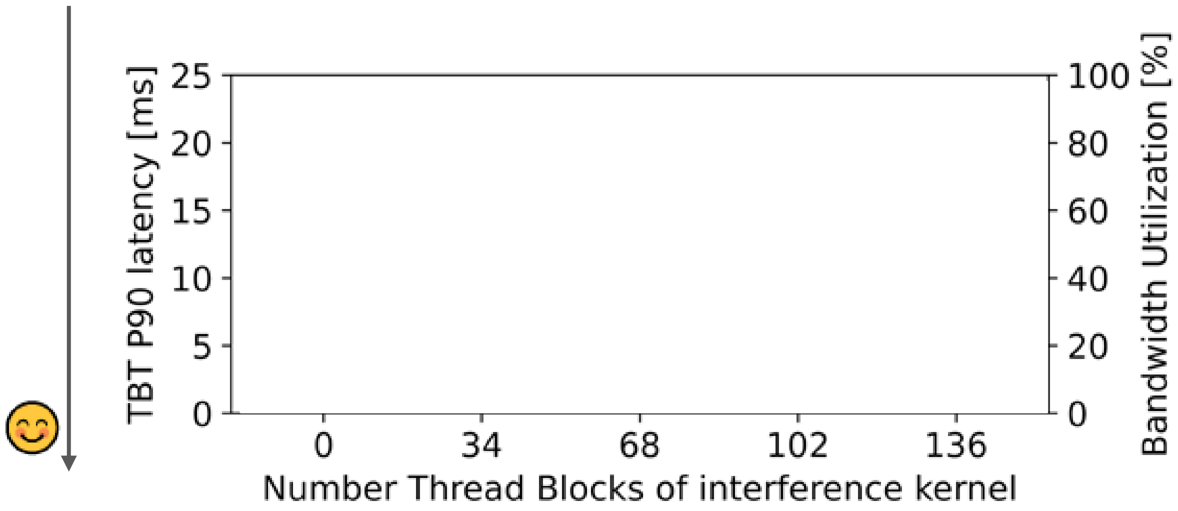
[1] CUDA Green Contexts

Memory Bandwidth Interference

Colocate **LLama3.1-8B** (BS 8, prompt size 16384) with **memory bandwidth intense microbenchmark** on a NVIDIA H100 **on disjoint SMs** (64-68 split).

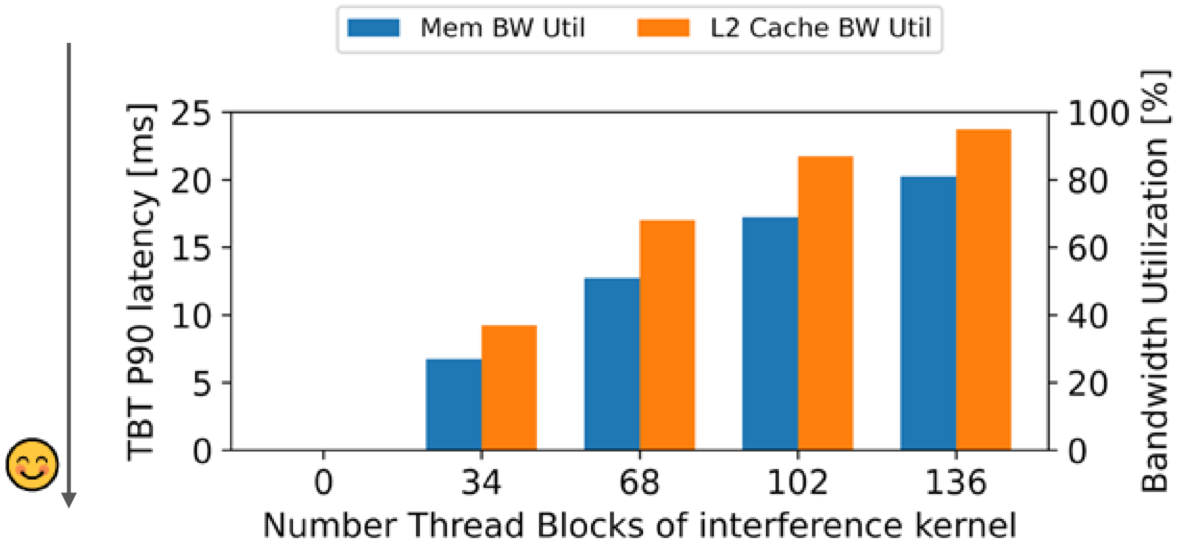
Memory Bandwidth Interference

Colocate **LLama3.1-8B** (BS 8, prompt size 16384) with **memory bandwidth intense microbenchmark** on a NVIDIA H100 **on disjoint SMs** (64-68 split).



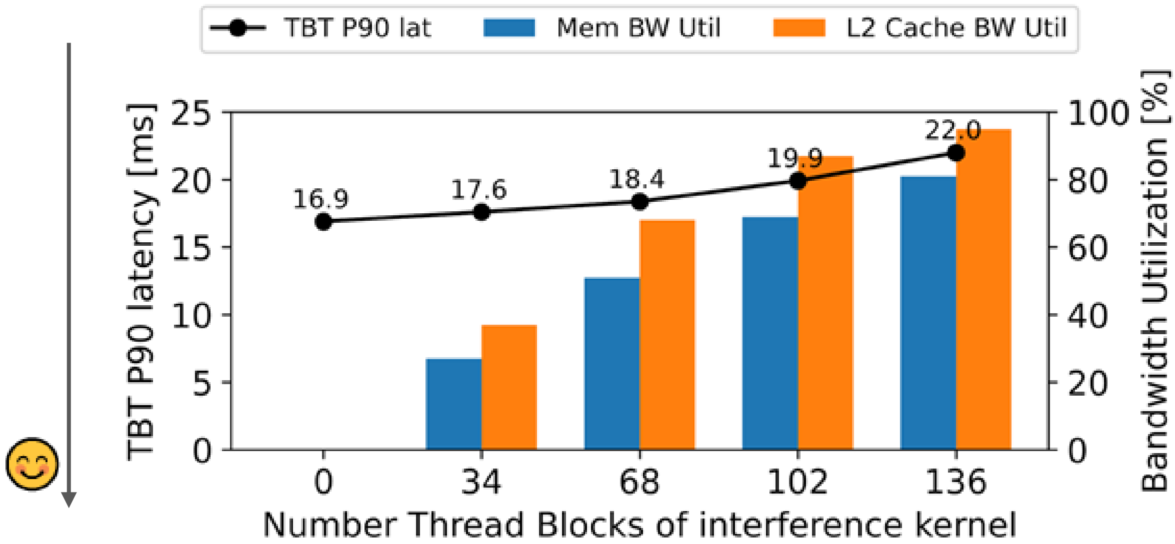
Memory Bandwidth Interference

Colocate **LLama3.1-8B** (BS 8, prompt size 16384) with **memory bandwidth intense microbenchmark** on a NVIDIA H100 **on disjoint SMs** (64-68 split).



Memory Bandwidth Interference

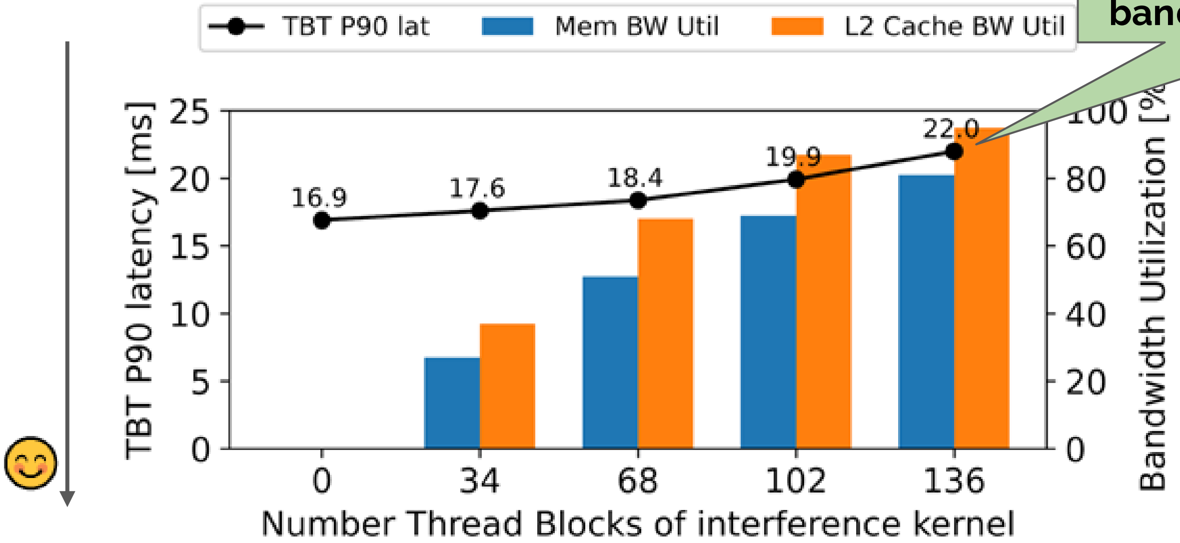
Colocate **LLama3.1-8B** (BS 8, prompt size 16384) with **memory bandwidth intense microbenchmark** on a NVIDIA H100 **on disjoint SMs** (64-68 split).



Memory Bandwidth Interference

Colocate **LLama3.1-8B** (BS 8, prompt size 16384) with **memory bandwidth intense microbenchmark** on a NVIDIA H100 **on disjoint SMs** (64-68 split).

Despite running on disjoint SMs, **resources like L2 cache and memory bandwidth remain shared**



Key learnings and future directions

1. **What have we learned?**
   a. GPUs are made up of **multiple heterogeneous resources**, each a potential source of interference.

   b. **GPU interference is multi-dimensional**. Single metrics cannot capture the entire landscape.

   c. **Colocation can be beneficial** when interference is properly modeled.

Key learnings and future directions

1. **What have we learned?**
    a. GPUs are made up of **multiple heterogeneous resources**, each a potential source of interference.

    b. **GPU interference is multi-dimensional**. Single metrics cannot capture the entire landscape.

    c. **Colocation can be beneficial** when interference is properly modeled.

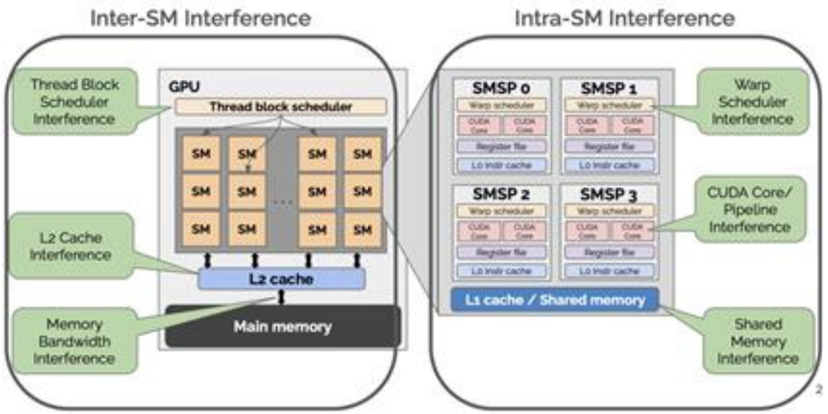
2. **Where should we go from here?**
    a. Build an **interference predictor**.

    b. **Extend the benchmark** suite to other GPU vendors.

    c. Kernel designers should start **developing kernels with colocation in mind**.
    => *“Do we really need to use 10% more resources for 2% in additional performance?”*

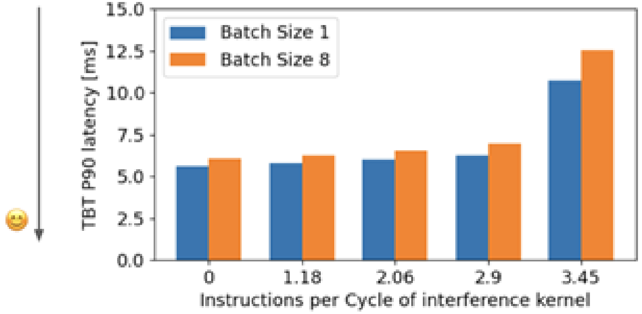
    d. Hardware manufacturers to become **more open-source** about internal functionality.

Sources of GPU interference



CUDA Benchmark suite and Methodology to isolate and stress on GPU resource at a time

Colocation can be beneficial when interference is accurately modeled along all dimensions



Github Repo

For further questions

Paul Elvinger
elvingerpa@gmail.com

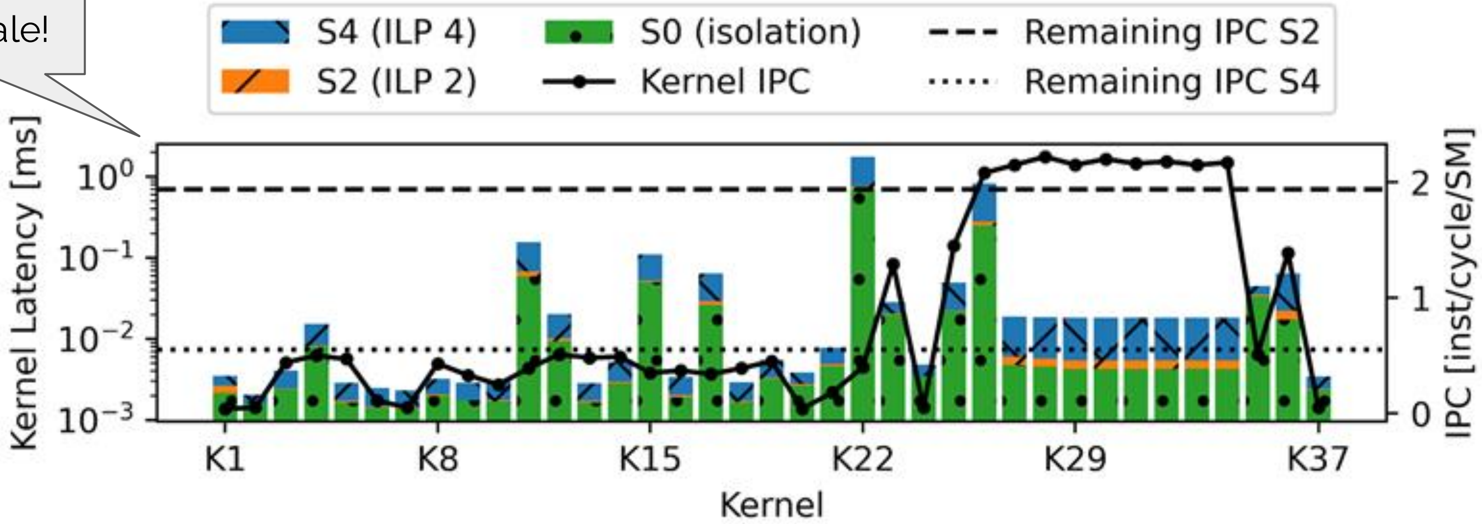
Foteini Strati
foteini.strati@inf.ethz.ch

Backup Slides

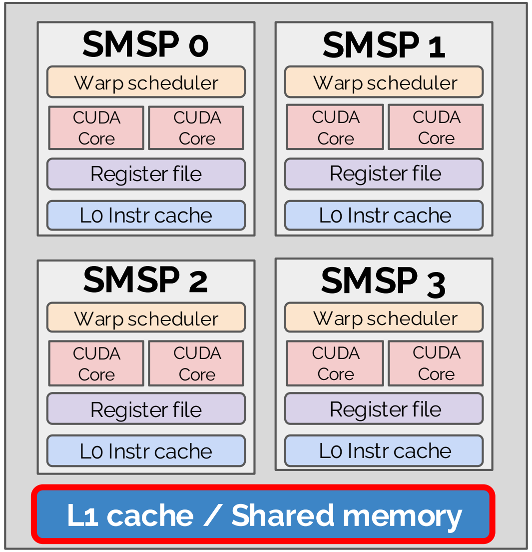
Warp Scheduler Interference - Kernel Level Impact

Kernel-level latency for **Llama3.1-8B with 1 hidden layer** (batch size 8, prompt size 1000) while colocated with an IPC intense microbenchmark.

log-scale!

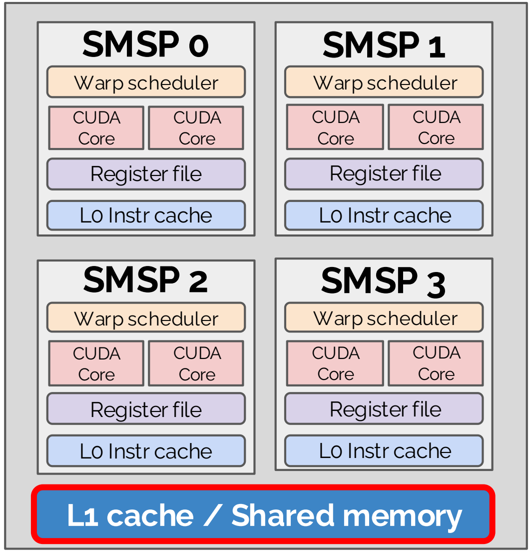


Shared Memory Interference



Shared Memory Interference

- Shared Memory is accessed over 32 banks*.
- **Bank conflict**: different addresses mapping to the same bank => **accesses are serialized**



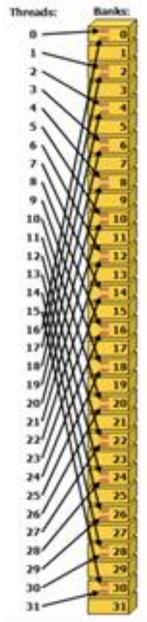
* Specific to the NVIDIA GPU architecture (CC >= 5.0)
Figures from [CUDA Programming Guide]

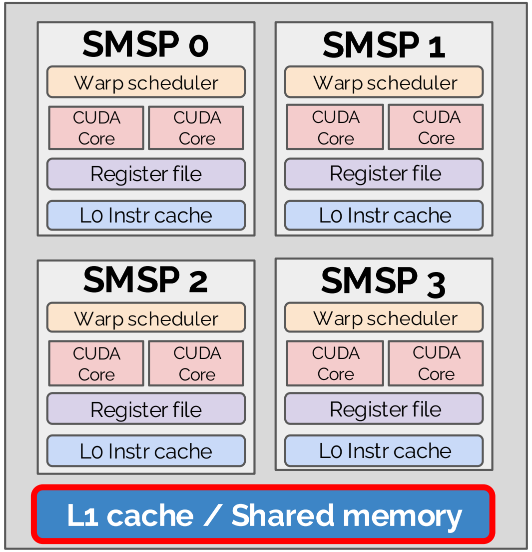
Shared Memory Interference

- Shared Memory is accessed over 32 banks*.
- **Bank conflict**: different addresses mapping to the same bank => **accesses are serialized**

conflict free

2-way conflict





* Specific to the NVIDIA GPU architecture (CC >= 5.0)

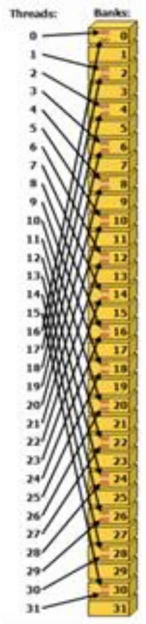
Figures from [CUDA Programming Guide]

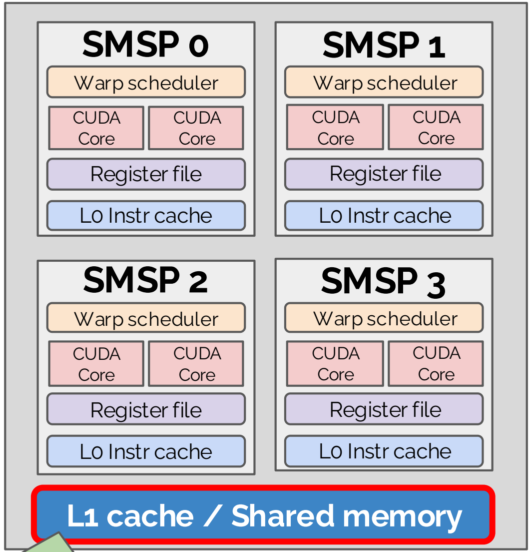
Shared Memory Interference

- Shared Memory is accessed over 32 banks*.
- **Bank conflict**: different addresses mapping to the same bank => **accesses are serialized**

conflict free

2-way conflict





Microbenchmark to **create high number of bank conflicts**

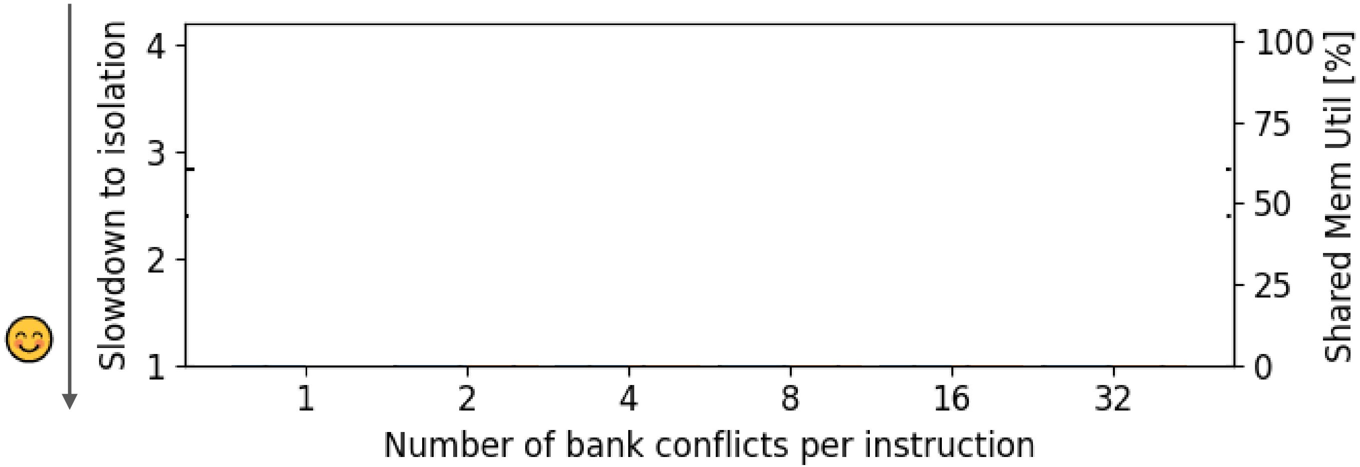
* Specific to the NVIDIA GPU architecture (CC >= 5.0)
Figures from [CUDA Programming Guide]

Shared Memory Interference

Colocate **GEMMs with a shared memory intensive microbenchmark** on NVIDIA **H100**

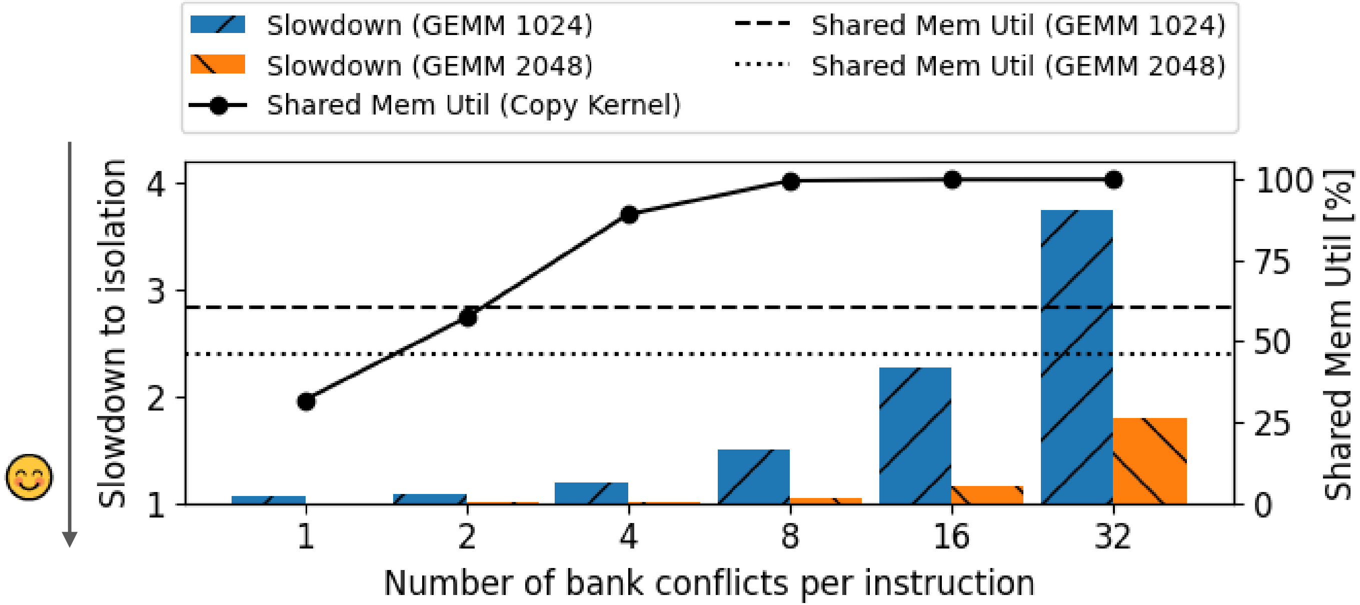
Shared Memory Interference

Colocate **GEMMs with a shared memory intensive microbenchmark** on NVIDIA **H100**



Shared Memory Interference

Colocate **GEMMs with a shared memory intensive microbenchmark** on NVIDIA **H100**



Shared Memory Interference

Colocate **GEMMs with a shared memory intensive microbenchmark** on NVIDIA **H100**

A **kernel** with non-optimal shared memory access pattern **may starve memory accesses of colocated kernels**