KU-The Future

최신 GPU 아키텍처의 데이터 캐시 성능 분석

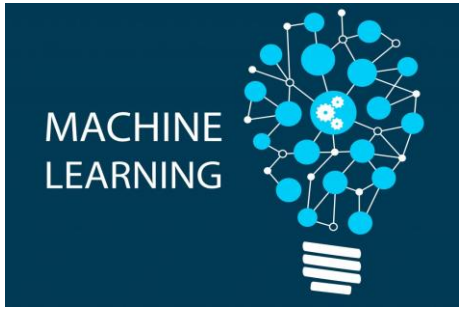
Analyzing Data Cache Performance of Modern GPU Architecture

정종현O*, 오윤호+, 구건재*
*고려대학교 컴퓨터학과
+성균관대학교 전자전기공학부

Outline

- **Motivation**
- **Background**
- **Evaluation**
  - Configuration
  - Workloads
  - Simulation Results
- **Conclusion**

Motivation

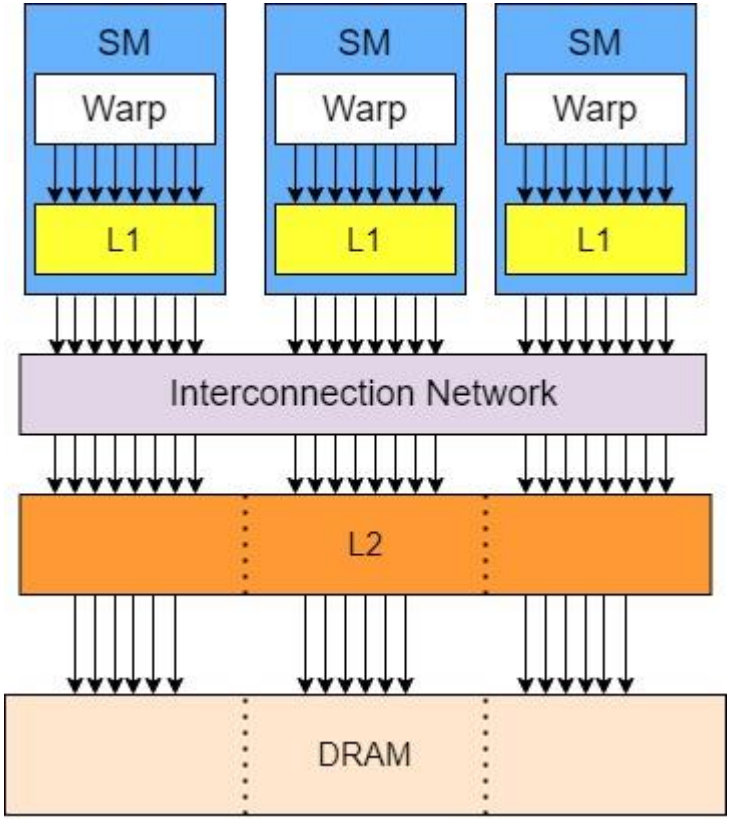


AI / ML

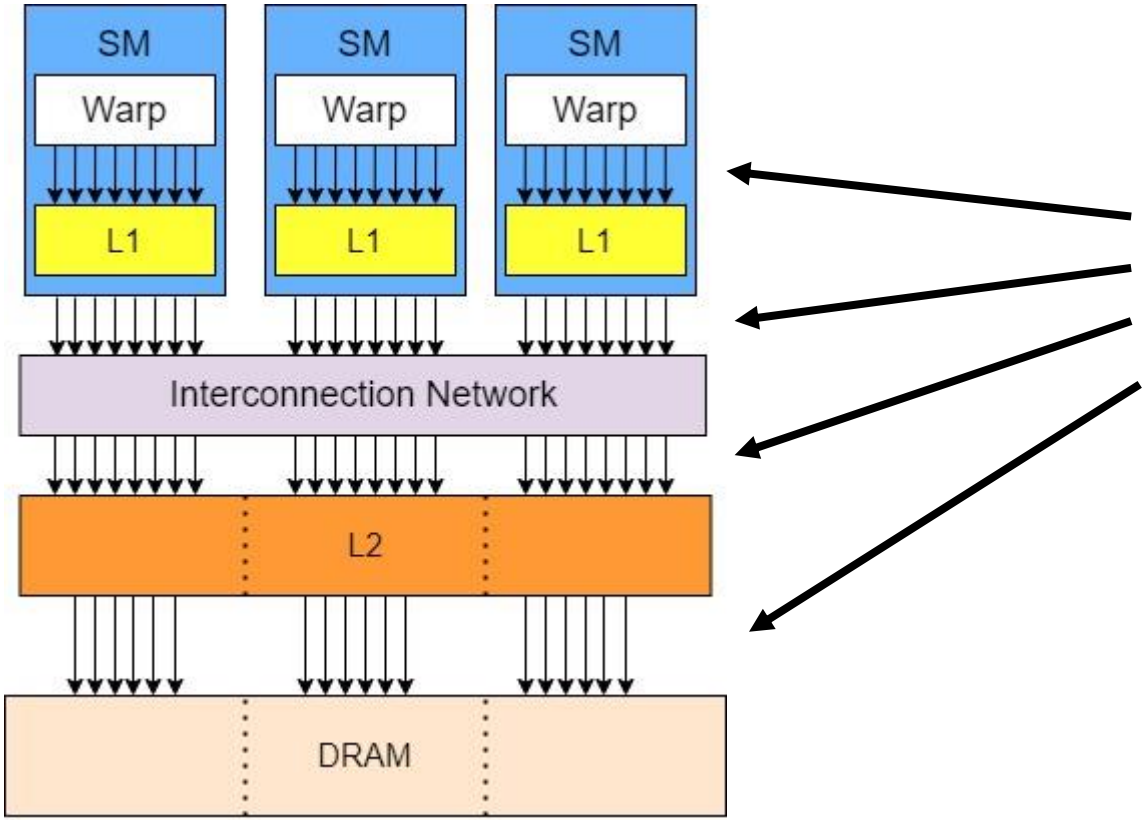
Data Analytics

HPC

Motivation


SM
Warp
L1
SM
Warp
L1
SM
Warp
L1
Interconnection Network
L2
DRAM

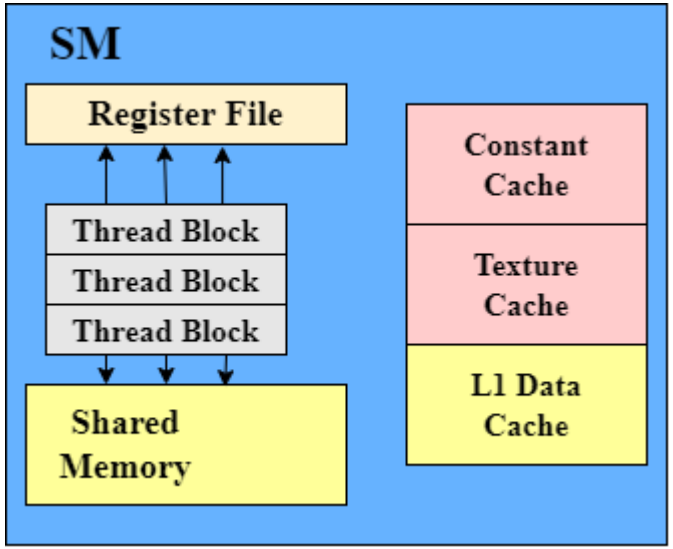
Motivation



Bottleneck!!!

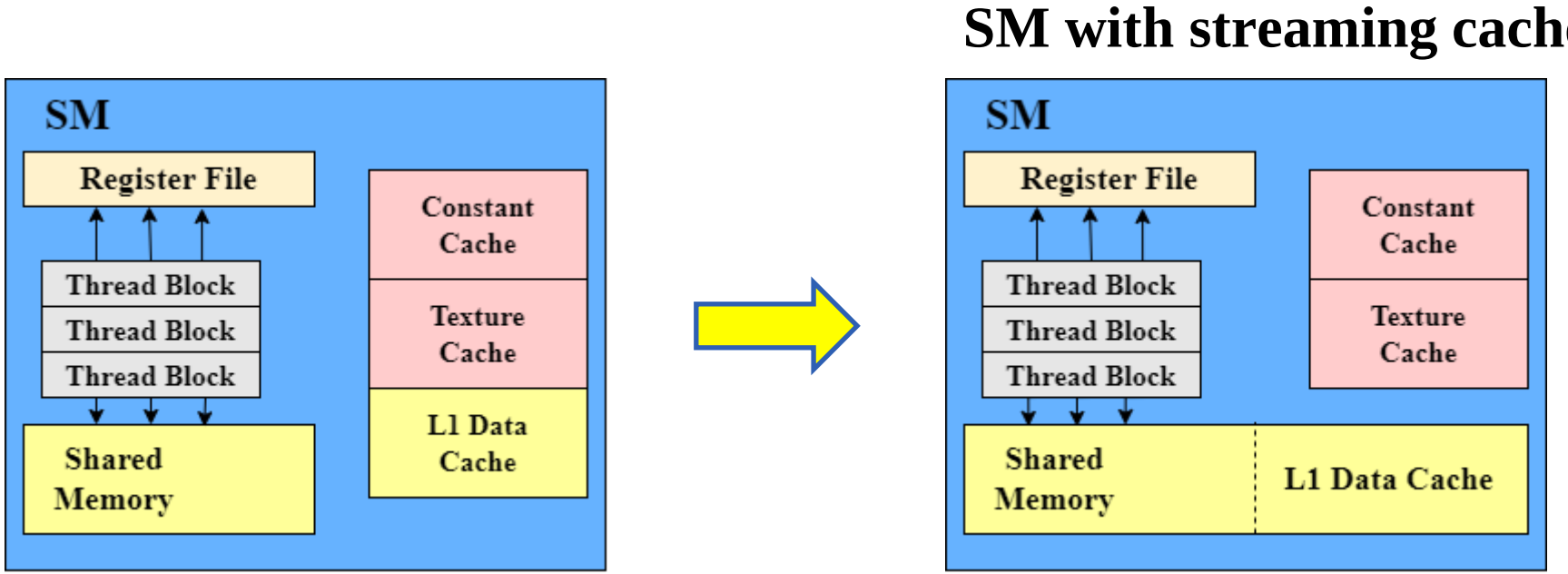
Background

Previous SM architecture



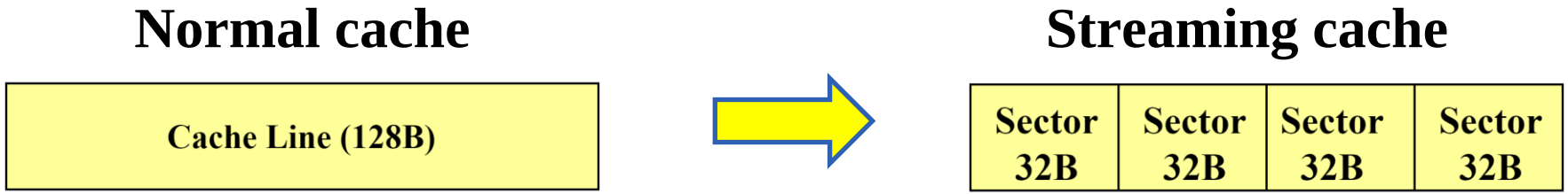
Data cache and shared memory are separated.

Background



Streaming cache integrates data cache and shared memory space
→ Access latency of data cache decreases

Background



Sector cache design (memory access granularity: 128B → 32B)

Background

MSHR : Miss Status Holding Register

- **Expanded MSHRs in Streaming cache**
  - Individual MSHR entry for all sectors in data cache
  - The maximum number of merged outstanding requests is equals to the number of threads per warp

Configuration

Simulator : GPGPU-Sim 4.0
GPU architecture : Turing, RTX2060

Common configuration

Core	30 SMs, 64 CUDA cores / SM
Warp	32 Warps / SM
CTA	32 CTAs / SM
Register file	64KB / SM
Shared memory	64KB / SM
L2 cache	3MB, 128B line, 16 way, 192 MSHR entries
Dram	GDDR6
ICNT topology	52 × 1 butterfly
ICNT peak BW	786.24GB/s

Data cache configuration

	Normal Cache	Streaming Cache
L1D cache	64KB	64KB
L1D latency	27 cycle	20 cycle
MSHR entry	256	2048
MSHR max merge	8	32
Allocation policy	On fill	On fill

Configuration

Simulator : GPGPU-Sim 4.0
GPU architecture : Turing, RTX2060

Common configuration

Core	30 SMs, 64 CUDA cores / SM
Warp	32 Warps / SM
CTA	32 CTAs / SM
Register file	64KB / SM
Shared memory	64KB / SM
L2 cache	3MB, 128B line, 16 way, 192 MSHR entries
Dram	GDDR6
ICNT topology	52 × 1 butterfly
ICNT peak BW	786.24GB/s

Data cache configuration

	Normal Cache	Streaming Cache
L1D cache	64KB	64KB
L1D latency	27 cycle	20 cycle
MSHR entry	256	2048
MSHR max merge	8	32
Allocation policy	On fill	On fill

Configuration

Simulator : GPGPU-Sim 4.0
GPU architecture : Turing, RTX2060

Common configuration

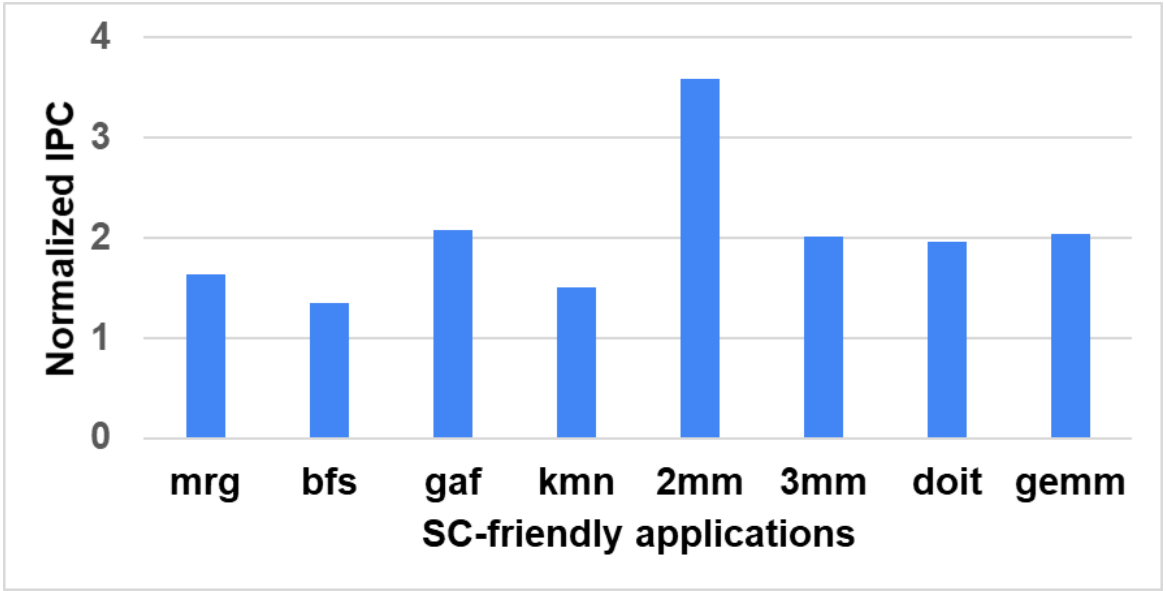
Core	30 SMs, 64 CUDA cores / SM
Warp	32 Warps / SM
CTA	32 CTAs / SM
Register file	64KB / SM
Shared memory	64KB / SM
L2 cache	3MB, 128B line, 16 way, 192 MSHR entries
Dram	GDDR6
ICNT topology	52 × 1 butterfly
ICNT peak BW	786.24GB/s

Data cache configuration

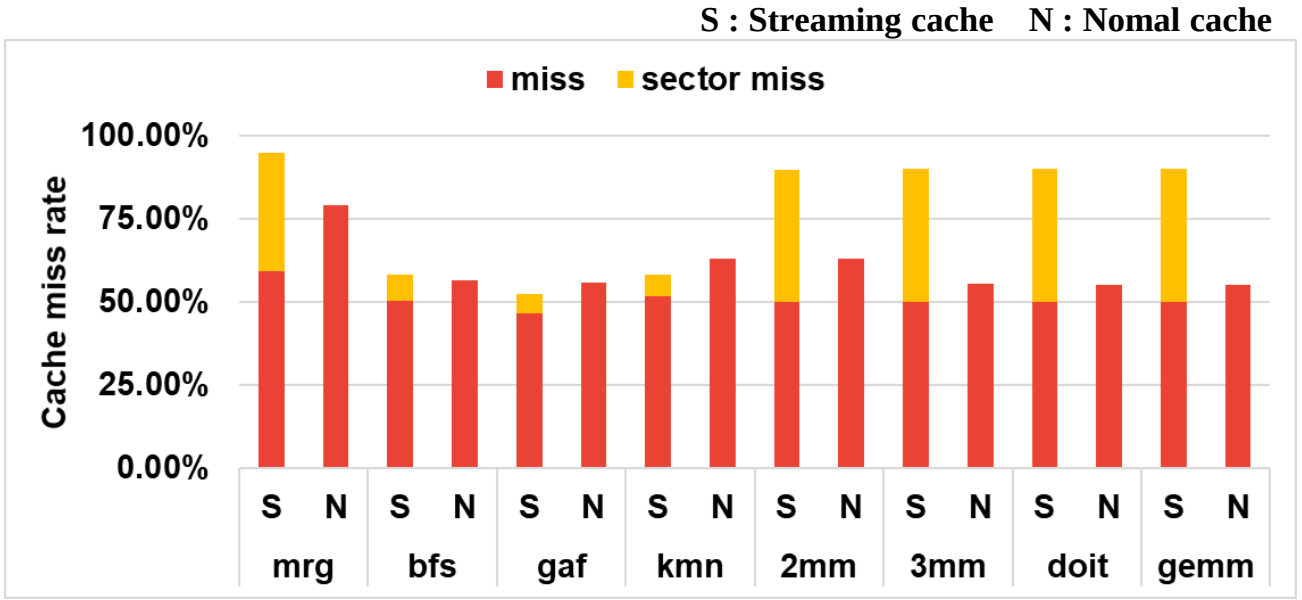
	Normal Cache	Streaming Cache
L1D cache	64KB	64KB
L1D latency	27 cycle	20 cycle
MSHR entry	256	2048
MSHR max merge	8	32
Allocation policy	On fill	On fill

Workloads

- **CUDA_SDK, Parboil, PolyBench, Rodinia3.0** - 41 applications
- **Streaming cache friendly application** - mrg, bfs, gaf, kmn, 2mm, 3mm, doit, gemm



Data Cache Miss Rate



- Cache miss rate increases for most applications
- Most of the increased misses are sector misses

Reservation Fail Statistics

	Normal cache	Normal cache + larger MSHR	Streaming cache
ICNT injection queue full	30,151 (0.001%)	297,013 (3.72%)	461,462 (100%)
MSHR entry full	96,771,331 (0.18%)		
MSHR merge full	422,051,937 (99.82%)	7,678,751 (96.28%)	0
Fail per access	1.71	0.03	0.0017

MSHR resource shortage

Reservation Fail Statistics

	Normal cache	Normal cache + larger MSHR	Streaming cache
ICNT injection queue full	30,151 (0.001%)	297,013 (3.72%)	461,462 (100%)
MSHR entry full	96,771,331 (0.18%)	0	0
MSHR merge full	422,051,937 (99.82%)	7,678,751 (96.28%)	0
Fail per access	1.71	0.03	0.0017

Still MSHR resource shortage

Reservation Fail Statistics

	Normal cache	Normal cache + larger MSHR	Streaming cache
ICNT injection queue full	30,151 (0.001%)	297,013 (3.72%)	461,462 (100%)
MSHR entry full	96,771,331 (0.18%)	0	
MSHR merge full	422,051,937 (99.82%)	7,678,751 (96.28%)	0
Fail per access	1.71	0.03	0.0017

Memory access granularity decrease

Reservation Fail Statistics

	Normal cache	Normal cache + larger MSHR	Streaming cache
ICNT injection queue full	30,151 (0.001%)	297,013 (3.72%)	461,462 (100%)
MSHR entry full	96,771,3 (0.18%)		
MSHR merge full	422,051,937 (99.82%)	7,678,751 (96.28%)	0
Fail per access	1.71	0.03	0.0017

×15

ICNT congestion increases

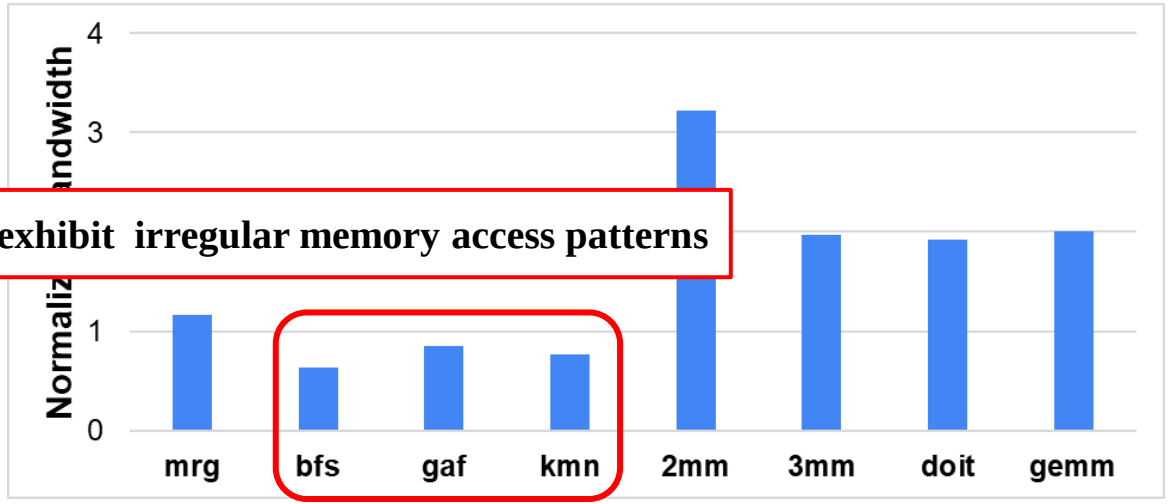
Reservation Fail Statistics

	Normal cache	Normal cache + larger MSHR	Streaming cache
ICNT injection queue full	30,151 (0.001%)	297,013 (3.72%)	461,462 (100%)
MSHR entry full	96,771,331 (0.18%)	0	0
MSHR merge full	422,051,9 (99.82%)	(96.28%)	
Fail per access	1.71	0.03	0.0017

Reservation fail per memory access decrease

-99.9%

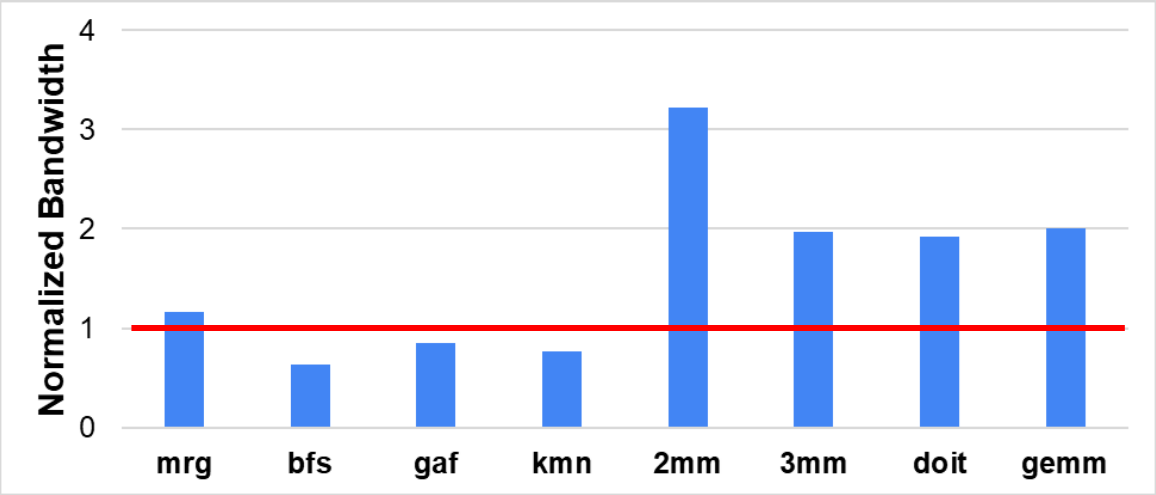
Interconnection Network Bandwidth



Applications that exhibit irregular memory access patterns

Irregular memory accesses : memory requests in a warp are not coalesced well

Interconnection Network Bandwidth



Outstanding requests increase → bandwidth increases except applications that exhibit irregular memory access patterns

Conclusion

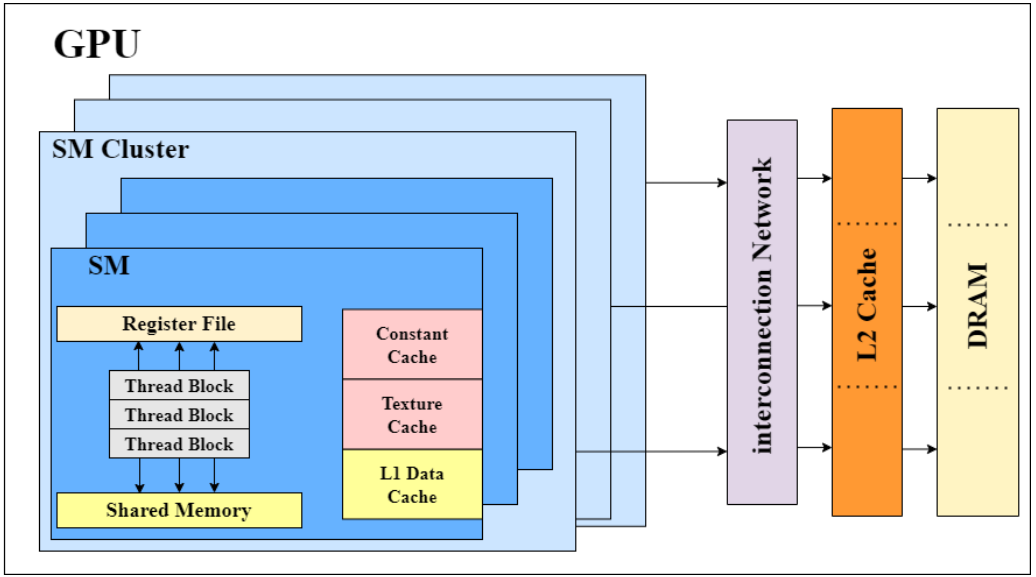
- Streaming cache is advantageous for applications that exhibit irregular memory access patterns.
- Data cache performance improvement by resolving cache congestions.
- Memory congestions move from data cache to interconnection networks.
- Further work will be an analysis of performance bottleneck observed in the interconnection networks, L2 cache, and DRAM.

Thank you

Backup

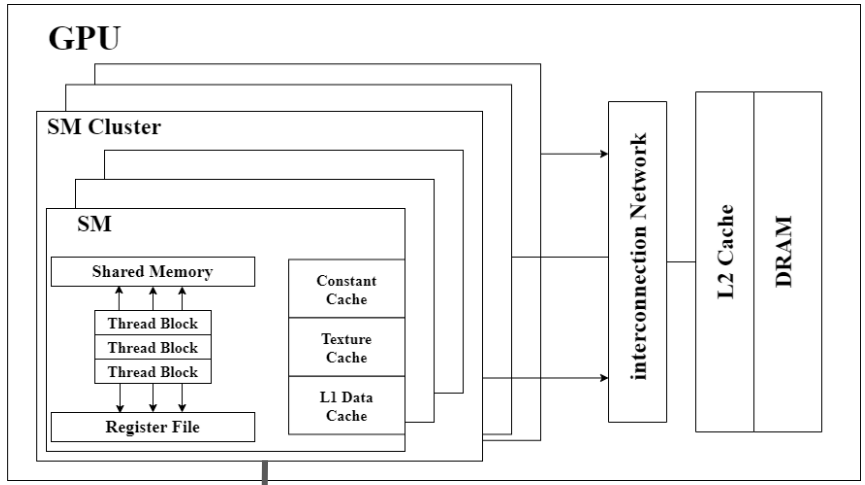
Background

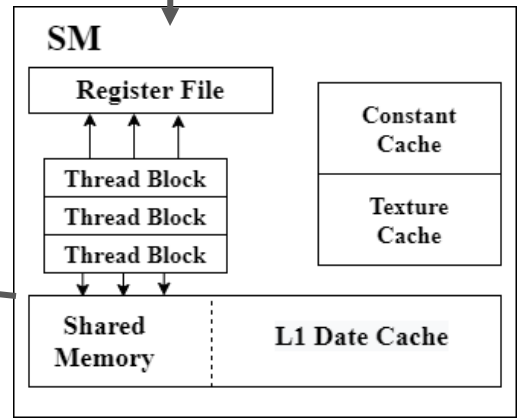
Conventional GPU architecture

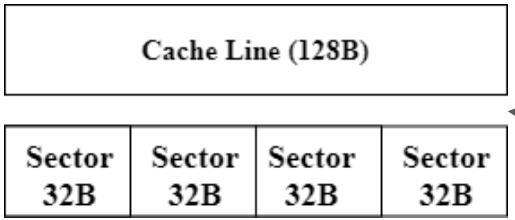


Background

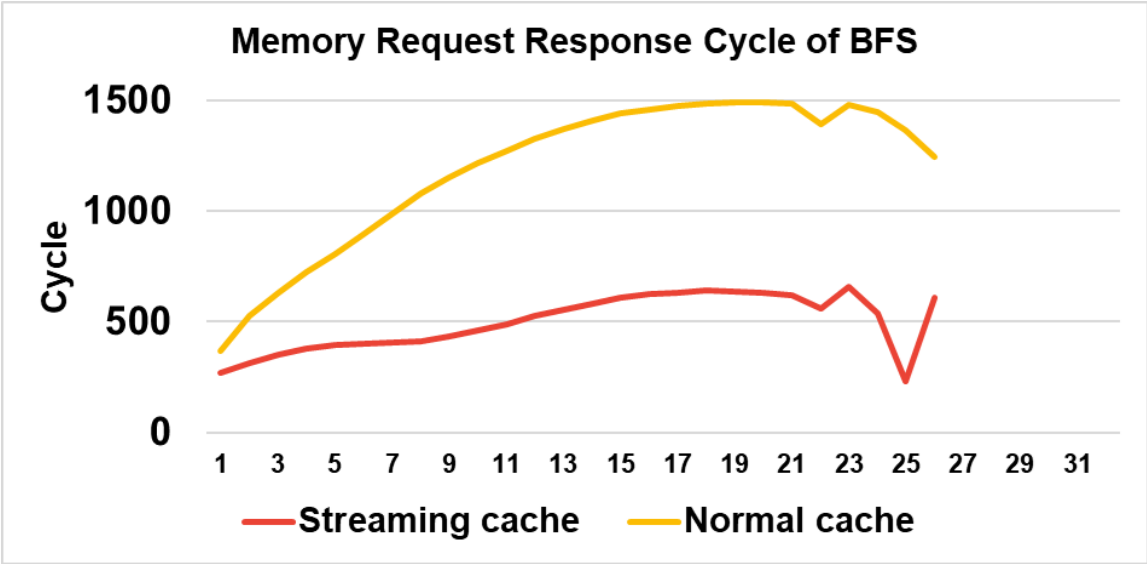
- **What's different in Streaming cache**
  - **sector** cache design (memory access granularity 128B → 32B)
  - **expanded MSHR** (individual MSHR entry for each sector, max merge of outstanding request equals to the number of threads per warp)
  - unified storage : **low data cache latency**



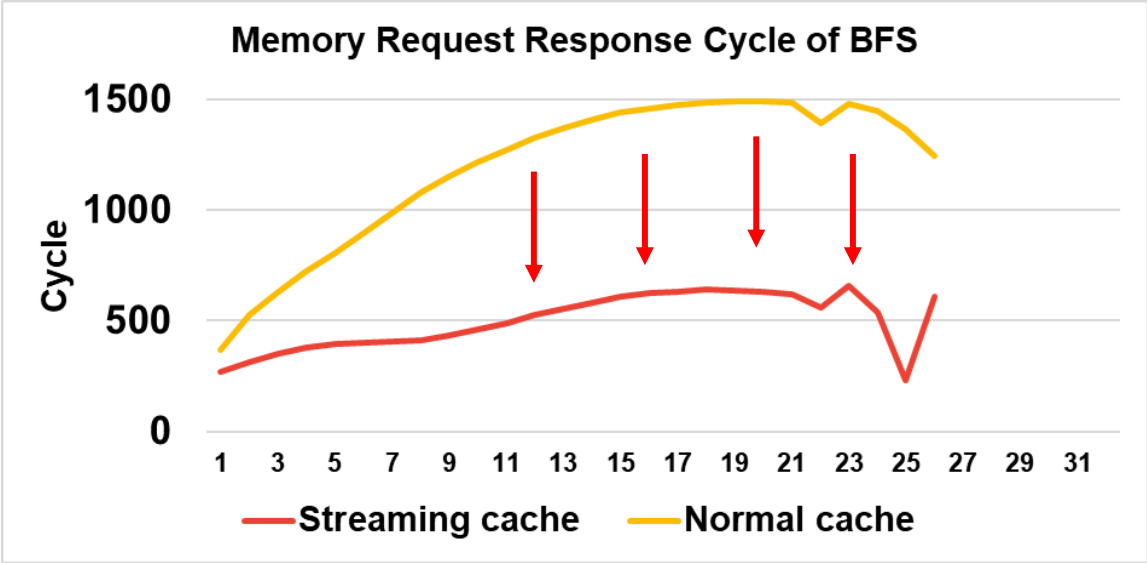




Simulation Result



Simulation Result

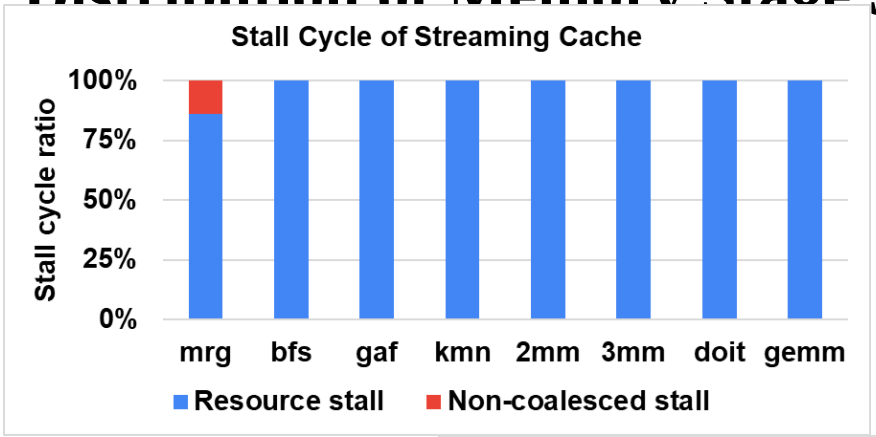


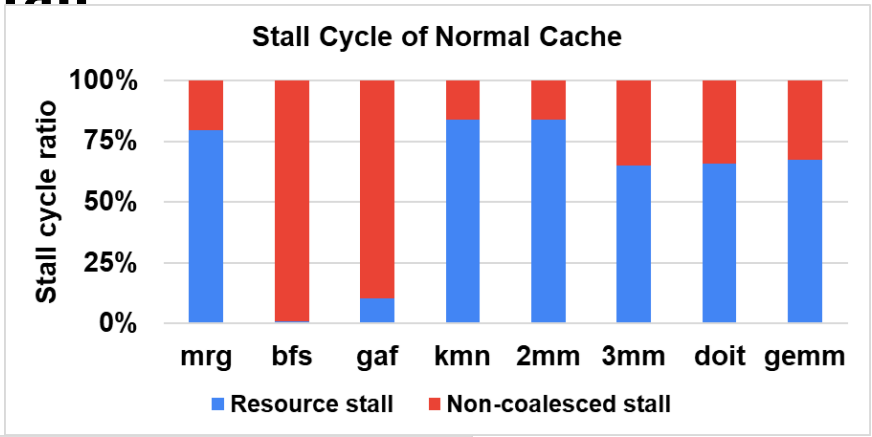
Smaller memory access granularity → Lower response cycle

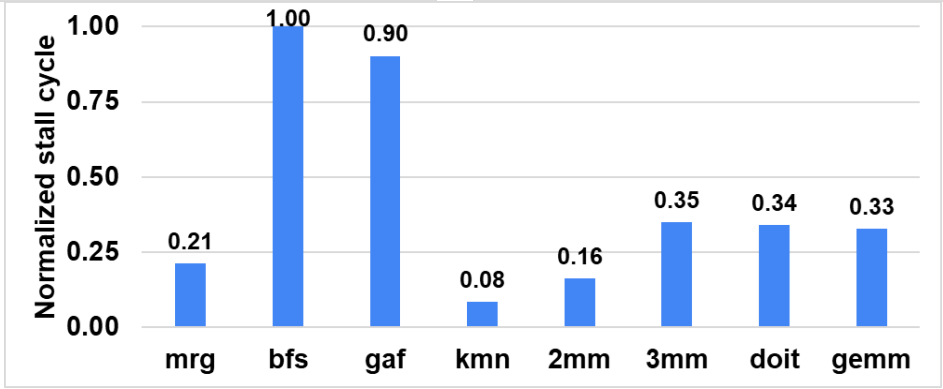
→ better performance of applications with irregular memory access pattern

Simulation Result

Distribution of Memory Stage Stall

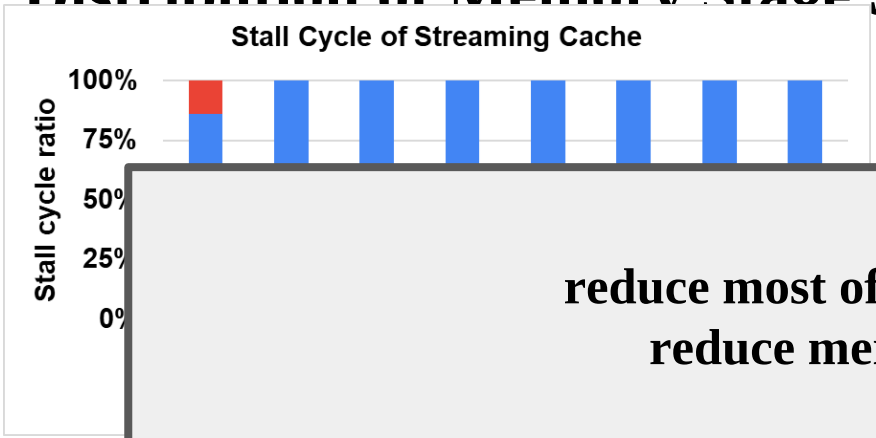


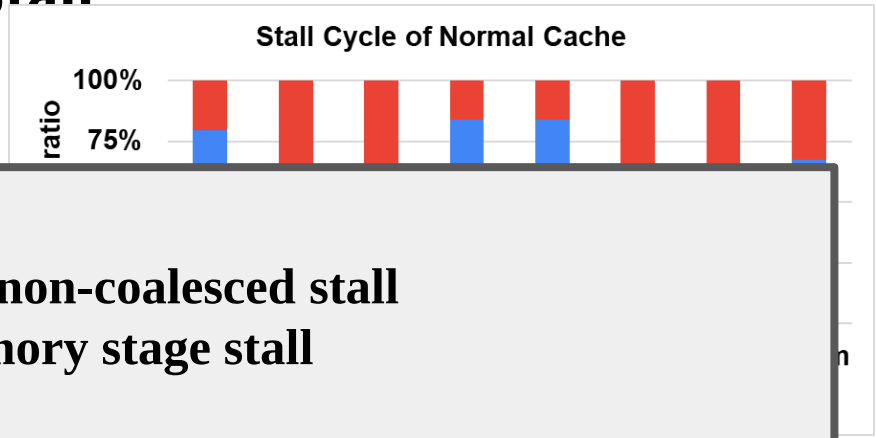




Simulation Result

Distribution of Memory Stage Stall





reduce most of non-coalesced stall
reduce memory stage stall