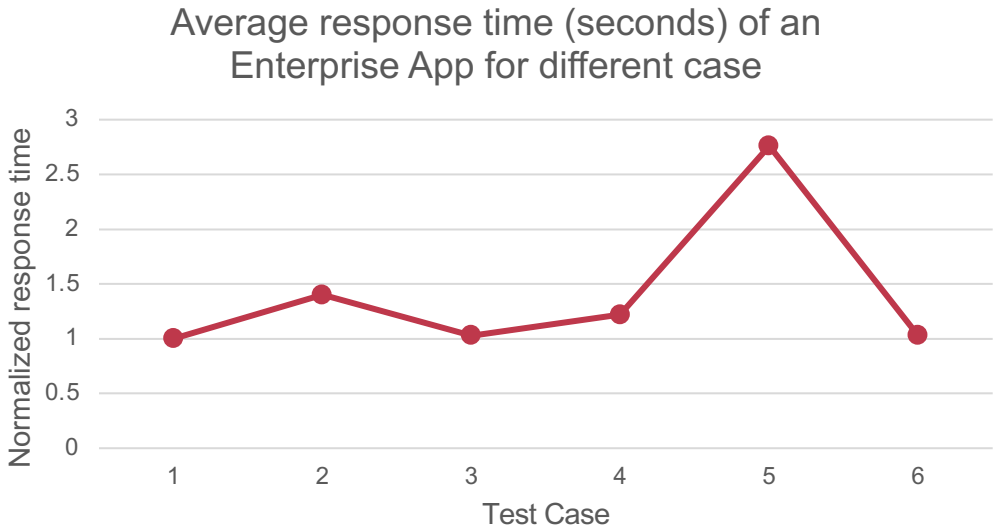
MT^2: Memory Bandwidth Regulation on Hybrid NVM/DRAM Platforms

Jifei Yi, Benchao Dong, Mingkai Dong, Ruizhe Tong, Haibo Chen

Institute of Parallel and Distributed Systems, Shanghai Jiao Tong University
Engineering Research Center for Domain-specific Operating Systems, Ministry of Education, China

Noisy Neighbors in Data Center

- Noisy neighbors bother all applications on the same platforms
  - **Per-thread** bandwidth monitoring is the key to identifying the noisy neighbor



Ref: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/intel-rdt-infrastructure-paper.pdf

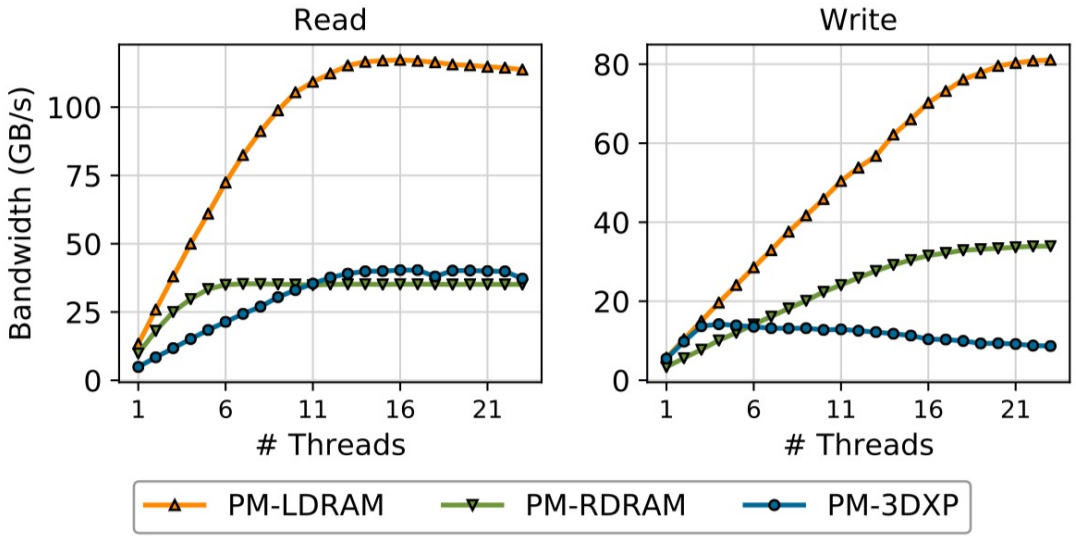
Noisy Neighbor in Hybrid Platforms

- **Non-Volatile Memory (NVM)**
  - Fast
  - Byte-addressable
  - Non-volatile

- **Setup**
  - Each channel has one DRAM and one NVM
- **Different types of memory traffic mix together**
  - Make the interference model more complex

Challenges in Hybrid Platforms

- **Memory bandwidth asymmetry**
  - NVM bandwidth is much smaller than DRAM



Ref: Joseph Izraelevitz, et al. 2019. Basic Performance Measurements of the Intel Optane DC Persistent Memory Module. http://arxiv.org/abs/1903.05714

Challenges in Hybrid Platforms

- **Memory bandwidth asymmetry**
- **Hard to distinguish access to NVM and DRAM**
  - NVM traffic and DRAM traffic are inevitably mixed and difficult to separate

Challenges in Hybrid Platforms

- **Memory bandwidth asymmetry**
- **Hard to distinguish access to NVM and DRAM**
- **Inadequate hardware and software mechanisms**
  - Hardware mechanism
    - Intel MBA can just slow down the memory access instructions
    - It has no effect for NVM
  - Software mechanisms
    - Frequency scaling and CPU scheduling
    - Slow down both computation and memory accesses

Outline

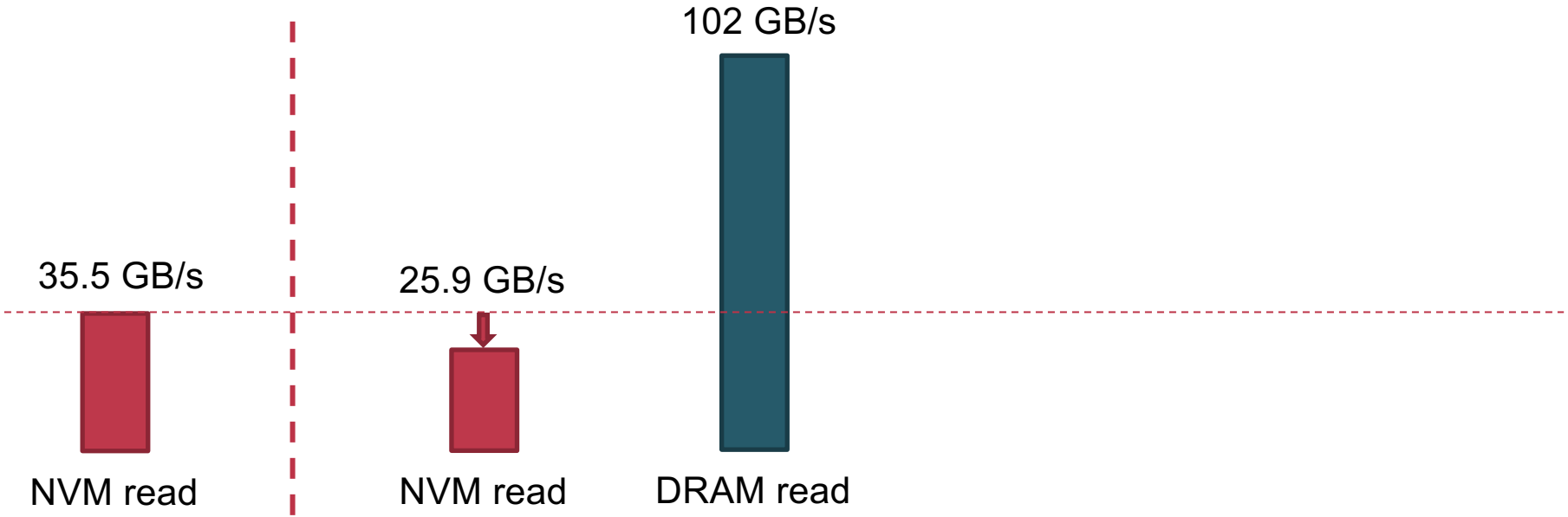
- Background
- **Analysis and Observation**
- Design
- Evaluation
- Discussion

Analysis and Observations

- The impact of memory interference is closely related to the type of memory access

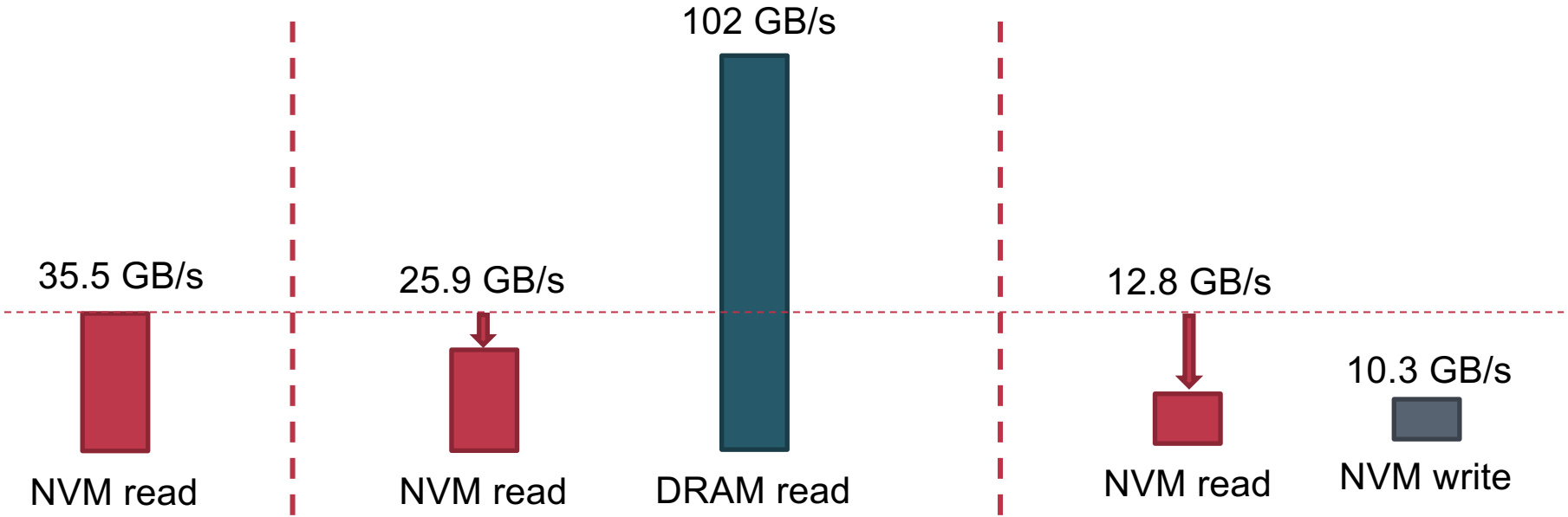
Analysis and Observations

- The impact of memory interference is closely related to the type of memory access



Analysis and Observations

- The impact of memory interference is closely related to the type of memory access



Analysis and Observations

- The impact of memory interference is closely related to the type of memory access
- Bandwidth interference level cannot be represented by the total bandwidth (absolute value is not enough)

Analysis and Observations

- The impact of memory interference is closely related to the type of memory access
- Bandwidth interference level cannot be represented by the total bandwidth (absolute value is not enough)
- NVM accesses of the same bandwidth have a more severe impact on other tasks than DRAM accesses

Outline

- Background
- Analysis and Observation
- **Design**
- Evaluation
- Discussion

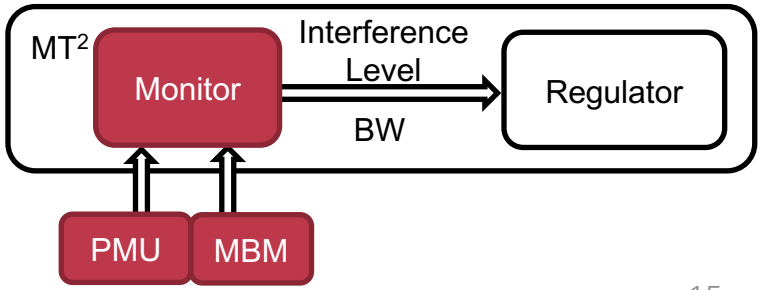
MT^2 (Memory Traffic Throttle)

- **Architecture**
  - Monitor
  - Regulator
- **Use cases**
  - Noisy neighbor suppression
  - Memory bandwidth allocation
  - Cloud SLO guarantee

Monitor: Bandwidth

Total bandwidth (Intel MBM)

- NVM read: PMU (ocr.all_data_rd.pmm_hit_local_pmm.any_snoop)
- NVM write: ?
- DRAM read: PMU (ocr.all_data_rd.l3_miss_local_dram.any_snoop)
- DRAM write: ?



Monitor: Bandwidth

Total bandwidth (Intel MBM)

- NVM read: PMU (ocr.all_data_rd.pmm_hit_local_pmm.any_snoop)
- NVM write: ?
- DRAM read: PMU (ocr.all_data_rd.l3_miss_local_dram.any_snoop)
- DRAM write: ?

It is hard to distinguish between NVM write bandwidth and DRAM write bandwidth of a thread

Monitor: Bandwidth

Total bandwidth (Intel MBM)

- NVM read: PMU (ocr.all_data_rd.pmm_hit_local_pmm.any_snoop)
- NVM write: **Software tracking in libraries (such as PMDK)**
- DRAM read: PMU (ocr.all_data_rd.l3_miss_local_dram.any_snoop)
- DRAM write: **Total bandwidth - others**

- **If we can trust the applications, we can rely on the applications to report their NVM write bandwidth faithfully**

Monitor: Bandwidth

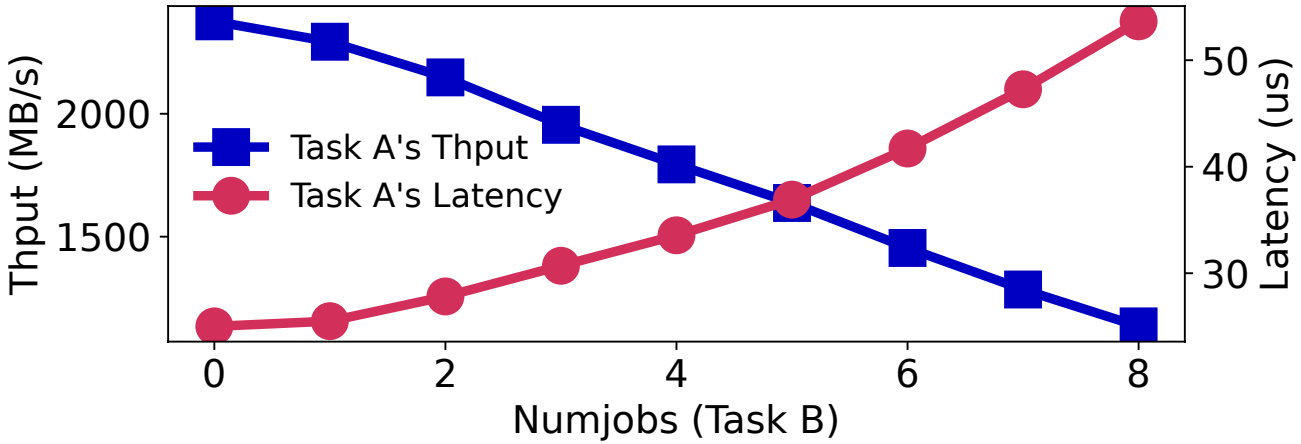
Total bandwidth (Intel MBM)

- NVM read: PMU (ocr.all_data_rd.pmm_hit_local_pmm.any_snoop)
- NVM write: **PEBS**
- DRAM read: PMU (ocr.all_data_rd.l3_miss_local_dram.any_snoop)
- DRAM write: Total bandwidth - others

- **Untrusted environment: use PEBS to estimate the write bandwidth roughly (detail can be found in the paper)**

Monitor: Interference Level

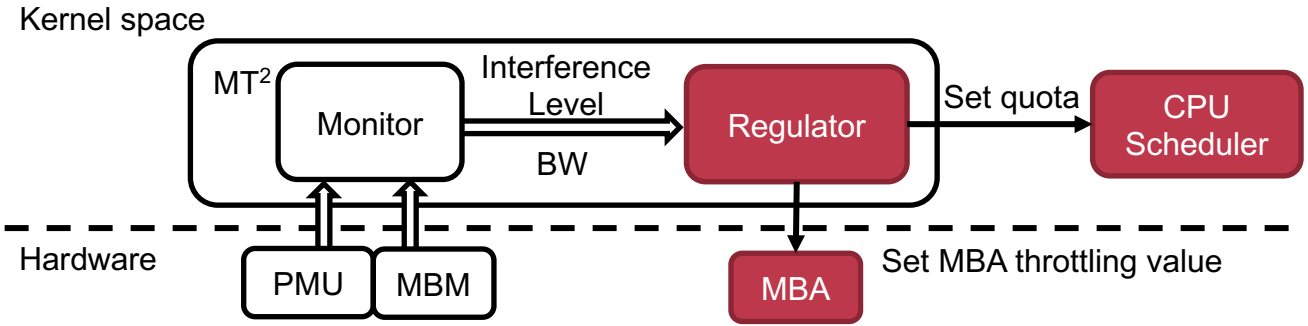
- The memory access latency is negatively correlated to the bandwidth



Measure memory access latency to indicate the bandwidth interference level

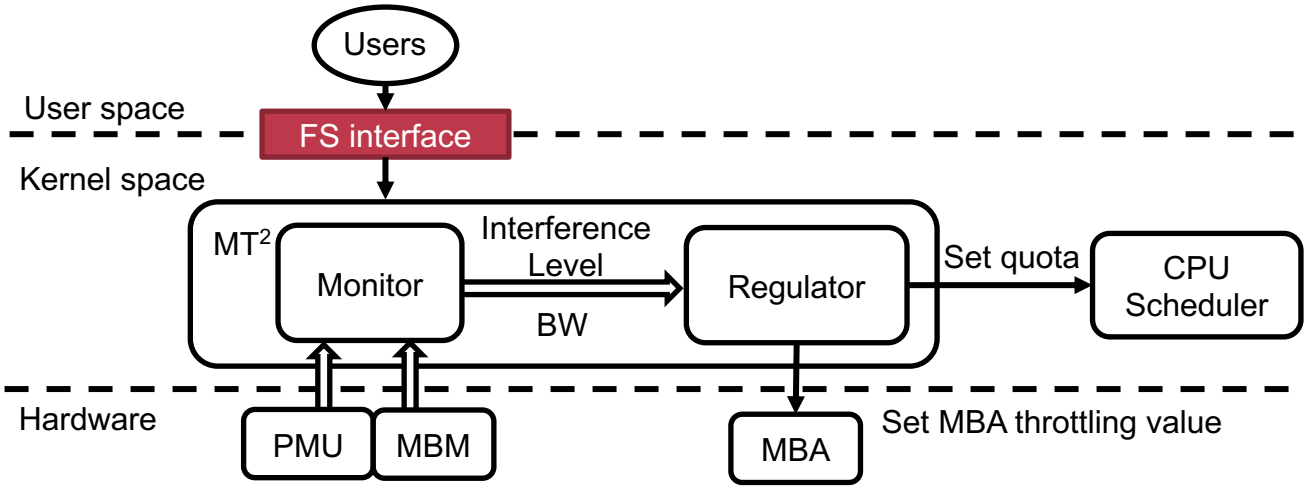
Bandwidth Regulator Mechanisms

- **Intel MBA**
  - Intel MBA can inject delay in memory accesses to reduce the memory bandwidth of a core
- **CPU quota**
  - Completely Fair Scheduler (CFS) allows to specify a cap of CPU time for a cgroup in a period

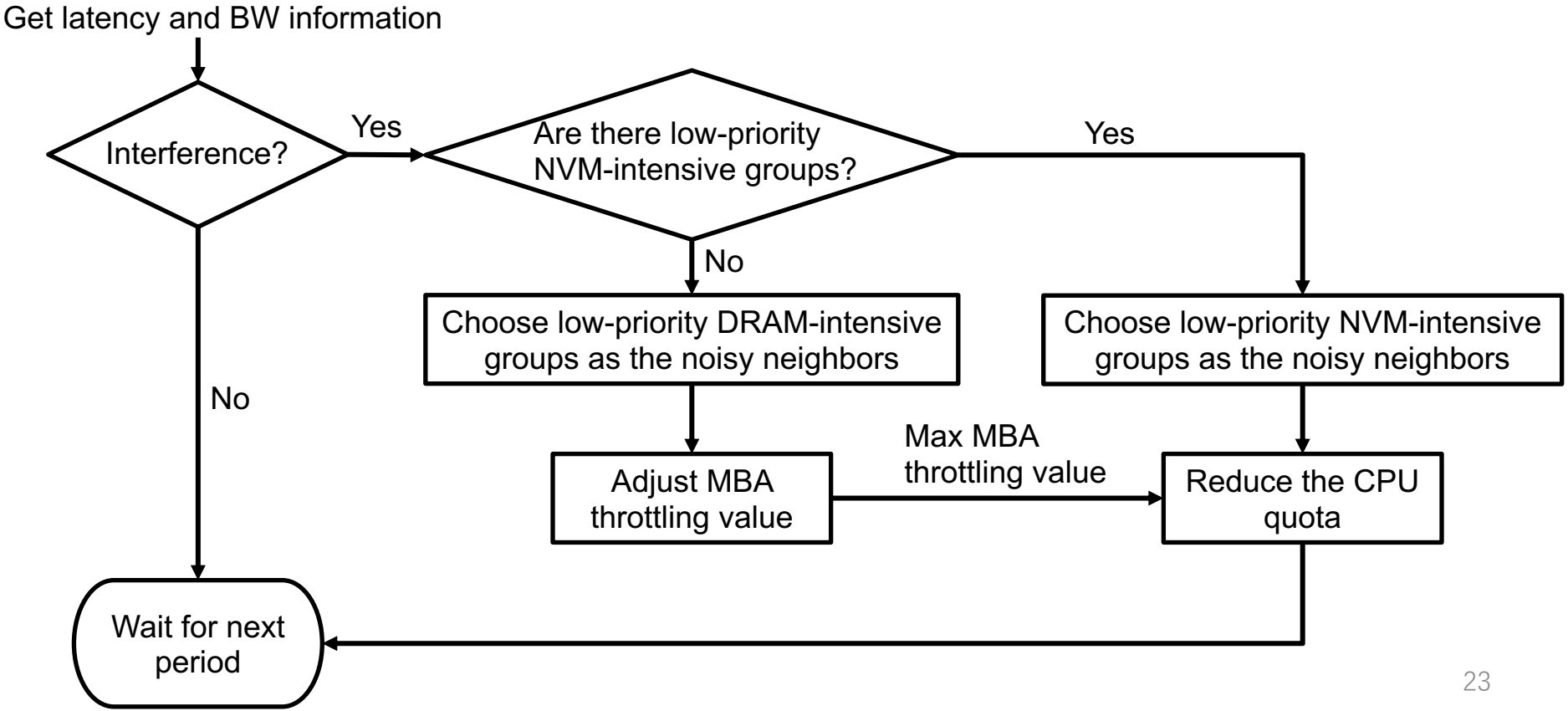


Interface

File Name	Permission	Description
priority	read/write	Get/set the priority of a group
bandwidth	read only	Get the bandwidth of a group for the last second
limit	read/write	Get and set the absolute bandwidth limit of a group



Use Case: Noisy Neighbors Suppression



Outline

- Background
- Observation
- Design
- **Evaluation**
- Discussion

Evaluation

- **Evaluation setup**
  - Two NUMA nodes, each has
    - Intel Xeon Gold 6238R CPU (28 cores)
      - Hyper-threading: disabled
    - 6 * 32GB DDR4 DRAM
    - 6 * 128GB Intel Optane Persistent Memory

Accuracy

- Use four FIO tasks to generate mixed bandwidth
- PCM monitors system-wide memory bandwidth

Bandwidth(GB/s)	DR	DW	NR	NW
MT^2	10.51	4.19	3.84	2.79
PCM	10.69	4.22	3.89	2.81
Deviation	**1.68%**	**0.71%**	**0.13%**	**0.71%**

The results reported by MT^2 are very close to the ground-truth bandwidth (<2%)

Performance Overhead

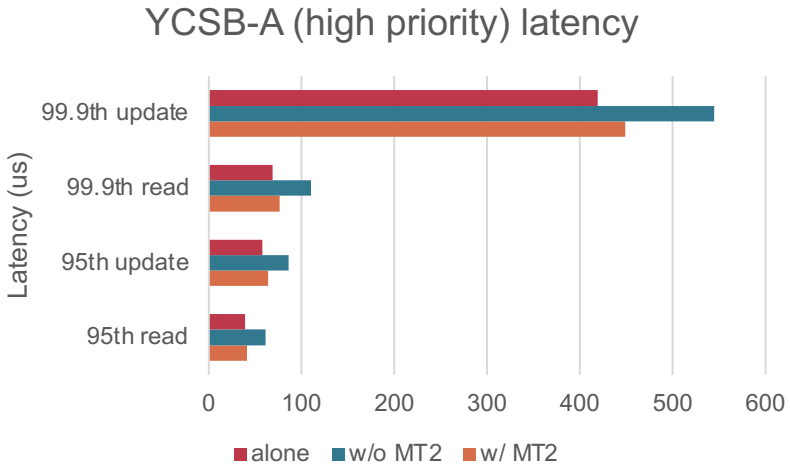
- Run applications without/with MT^2
  - FIO, Graphchi, Hadoop, and RocksDB

Throughput/Time	w/o MT^2	w/ MT^2	Overhead
FIO	31505 MB/s	31507 MB/s	< 0.01%
Graphchi	321.64 s	321.55 s	< 0.01%
Hadoop	54.93 s	54.93 s	< 0.01%
RocksDB	37770 ops/s	37767 ops/s	< 0.01%

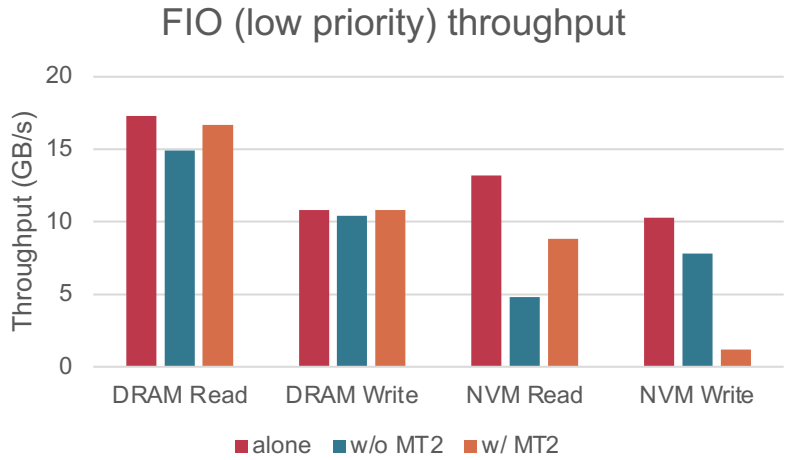
Introduced overhead is negligible (less than 0.01%)

Noisy Neighbors Suppression

- High priority: YCSB
  - Tail latency: lower is better



- Low priority: four FIO tasks
  - Throughput: higher is better



MT^2 optimizes the performance of the high-priority applications and some low-priority applications by restricting the noisiest low-priority application

Discussion and Future Works

- **Hybrid memory bandwidth allocation**
  - Not NUMA-aware: NUMA-and-NVM-aware scheduler
  - More accurate and generic bandwidth monitoring mechanism
  - More sophisticated restriction policies
- **Suggestions for NVM-aware data structure and system software design**
  - Minimize memory footprint, especially for NVM writes, which can lead to poor scalability

Conclusion

- Memory interference become more severe on the hybrid NVM/DRAM platforms
- Per-thread bandwidth monitoring is the key to reduce memory interference
- MT^2 is the first comprehensive system to monitor and regulate memory bandwidth on the hybrid platforms with thread granularity
- MT^2 can effectively regulate the bandwidth among applications with nearly zero performance overhead
- High-performance NVM data structure/software design needs to take bandwidth interference into consideration

IPADS
INSTITUTE OF PARALLEL AND DISTRIBUTED SYSTEMS

Thanks!