SeRW:

Adaptively Separating Read and Write upon SSDs of Hybrid Storage Server in Clouds

[1]Fan Deng, [1]Qiang Cao, [1]Shucheng Wang, [1]Shuyang Liu, [1]Jie Yao,
[2]Yuanyuan Dong, and [2]Puyuan Yang

[1]*Huazhong University of Science and Technology*

[2]*Alibaba*

Outline

- ✓ Introduction
- ✓ Background
- ✓ Analysis and Motivation
- ✓ Design of SeRW
  - ➢ Redirecting Strategy
  - ➢ Log Machanism
- ✓ Evaluation
- ✓ Conclusion

Introduction

- SSD-HDD hybrid storage in clouds.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
- We present a SeRW scheduling approach.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.

Introduction

- SSD-HDD hybrid storage in clouds.
  - ✓ SSDs as the primary storage directly serving requests from front-end applications.
  - ✓ HDDs as the secondary storage to provide sufficient storage capacity.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
- We present a SeRW scheduling approach.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.

Introduction

- SSD-HDD hybrid storage in clouds.
  - SSDs as the primary storage directly serving requests from front-end applications.
  - HDDs as the secondary storage to provide sufficient storage capacity.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
  - These long read latencies are primarily caused by (1) write-induced-blocking and (2) write-induced-garbage-collection (GC).
- We present a SeRW scheduling approach.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.

Introduction

- SSD-HDD hybrid storage in clouds.
  - SSDs as the primary storage directly serving requests from front-end applications.
  - HDDs as the secondary storage to provide sufficient storage capacity.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
  - These long read latencies are primarily caused by (1) write-induced-blocking and (2) write-induced-garbage-collection (GC).
- We present a SeRW scheduling approach.
  - The main idea is to adaptively steers some SSD-writes to idle HDDs in running time.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.

Introduction

- SSD-HDD hybrid storage in clouds.
  - ✓ SSDs as the primary storage directly serving requests from front-end applications.
  - ✓ HDDs as the secondary storage to provide sufficient storage capacity.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
  - ✓ These long read latencies are primarily caused by (1) write-induced-blocking and (2) write-induced-garbage-collection (GC).
- We present a SeRW scheduling approach.
  - ✓ The main idea is to adaptively steers some SSD-writes to idle HDDs in running time.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.
  - ✓ SeRW decreases the average, 99^{th} , 99.9^{th}, 99.99^{th}-percentile latencies of reads by up to 2.07x, 1.48x, 4.29x, and 4.24x, respectively.
  - ✓ Reducing the amount of data written to SSDs by up to 37.5%.

Outline

- ✓ Introduction
- ✓ Background
- ✓ Analysis and Motivation
- ✓ Design of SeRW
  - ➢ Redirecting Strategy
  - ➢ Log Machanism
- ✓ Evaluation
- ✓ Conclusion

Primary Storage

The performance characteristics of commodity SSDs and HDDs

Disk Type	SSD			HHD
Interface	**PCIe NVMe**	**PCIe AHCI**	**SATA AHCI**	**SATA AHCI**
Cost ($/GB)	1.2-2.6	0.6-1.1	0.5-1.0	0.2-0.45
Avg. write latency (us)	20-100	30-200	30-200	10k-30k
Avg. read latency (us)	20-100	30-200	30-200	10k-30k
Max. throughput (GB/s)	3	0.52	0.52	0.2

Primary Storage

The performance characteristics of commodity SSDs and HDDs

Disk Type	SSD			HHD
Interface	PCIe NVMe	PCIe AHCI	SATA AHCI	SATA AHCI
Cost ($/GB)	1.2-2.6	0.6-1.1	0.5-1.0	0.2-0.45
Avg. write latency (us)	20-100	30-200	30-200	10k-30k
Avg. read latency (us)	20-100	30-200	30-200	10k-30k
Max. throughput (GB/s)	3	0.52	0.52	0.2

- ✓ HDD
  - ➢ high capacity/cost ratio
  - ➢ limited peak throughput (e.g., 180MB/s), and a notorious random IO performance (e.g., 200 IOPS)

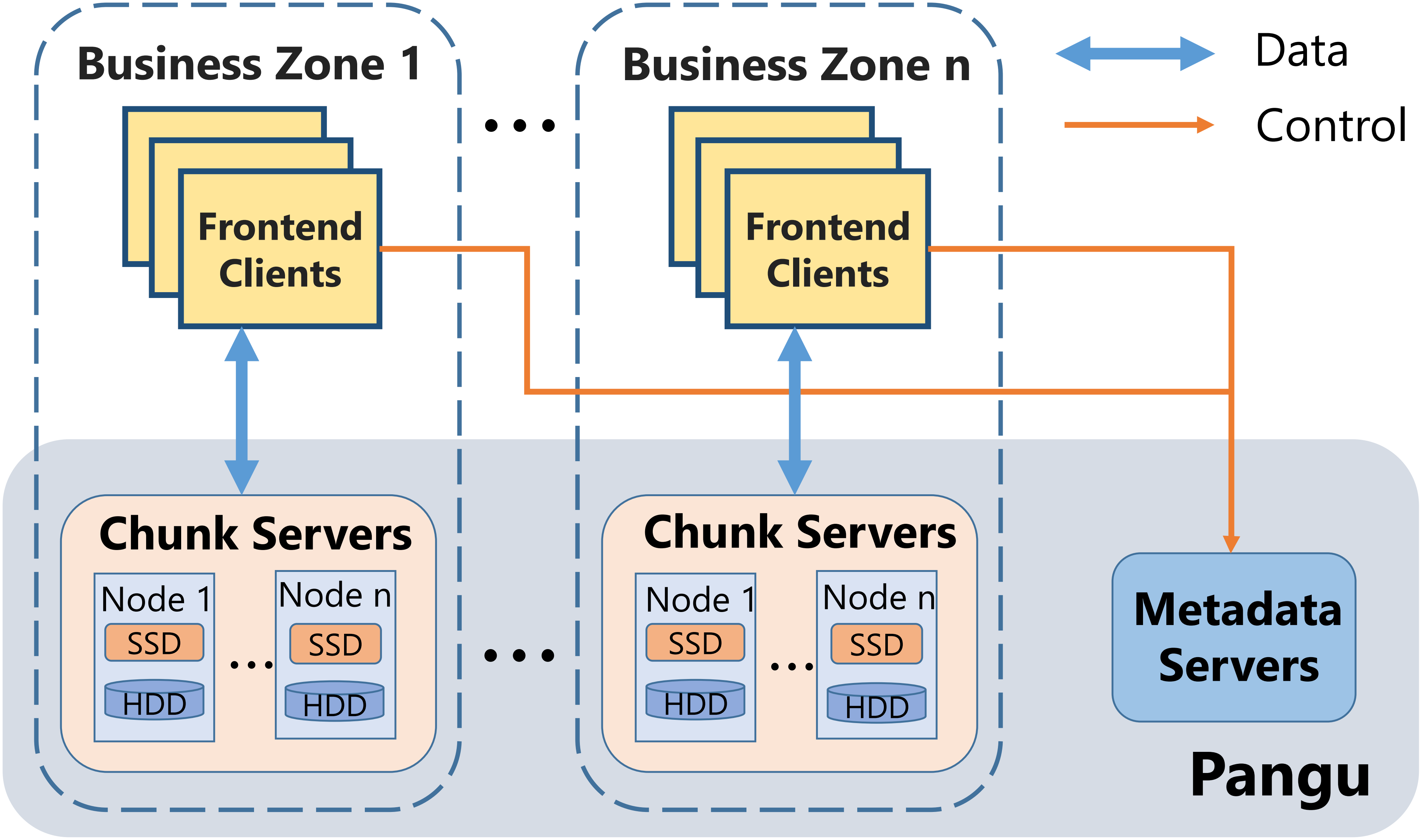
Primary Storage

The performance characteristics of commodity SSDs and HDDs

Disk Type	SSD			HHD
Interface	**PCIe NVMe**	**PCIe AHCI**	**SATA AHCI**	**SATA AHCI**
Cost ($/GB)	1.2-2.6	0.6-1.1	0.5-1.0	0.2-0.45
Avg. write latency (us)	20-100	30-200	30-200	10k-30k
Avg. read latency (us)	20-100	30-200	30-200	10k-30k
Max. throughput (GB/s)	3	0.52	0.52	0.2

- ✓ HDD
  - ➢ high capacity/cost ratio
  - ➢ limited peak throughput (e.g., 180MB/s), and a notorious random IO performance (e.g., 200 IOPS)
- ✓ SSD
  - ➢ high throughput, low IO delay, high internal-parallelism
  - ➢ write penalty and GC penalty

Pangu



Business Zone 1
Frontend Clients
Business Zone n
Frontend Clients
Data
Control
Chunk Servers
Node 1
SSD
HDD
Node n
SSD
HDD
Chunk Servers
Node 1
SSD
HDD
Node n
SSD
HDD
Metadata Servers
Pangu

Outline

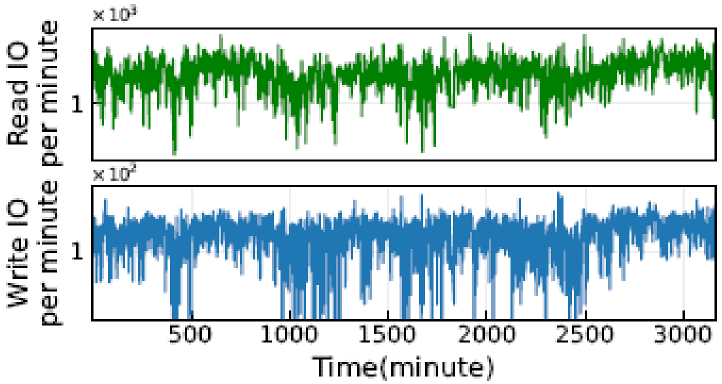
Pangu Workload

- Pangu workload traces
  - A1 and A2 : read-dominated nodes with SSD only from business I.
  - B1 and B2 : read/write mixed nodes combining SSDs and HDDs from business II.

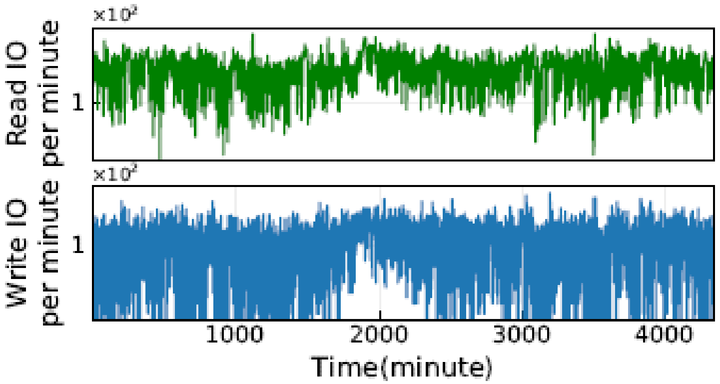
Pangu Workload

- Pangu workload traces
  - A1 and A2 : read-dominated nodes with SSD only from business I.
  - B1 and B2 : read/write mixed nodes combining SSDs and HDDs from business II.
- Load balance
  - Pangu achieves good load balancing and schedules across nodes.

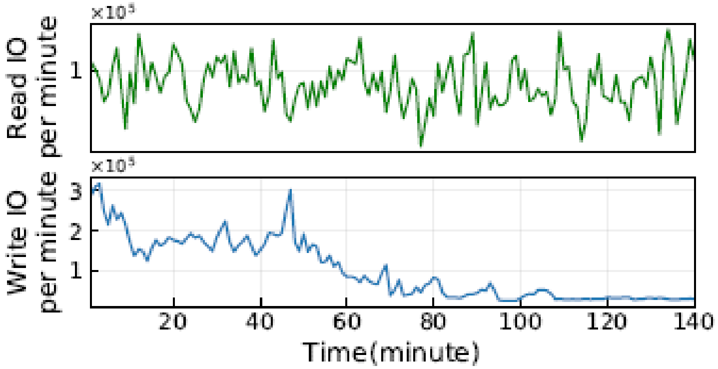
Pangu Workload



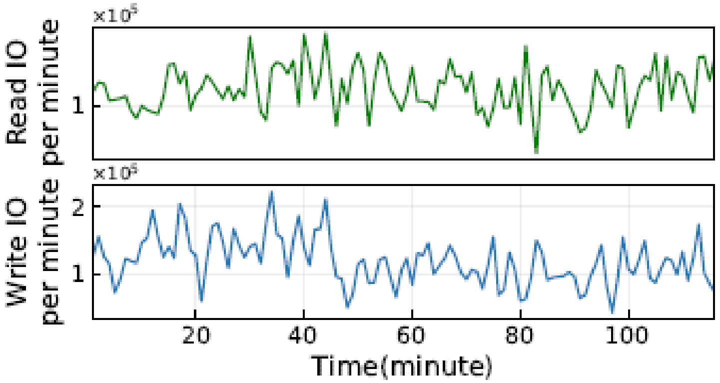
(a) A1



(b) A2



(c) B1



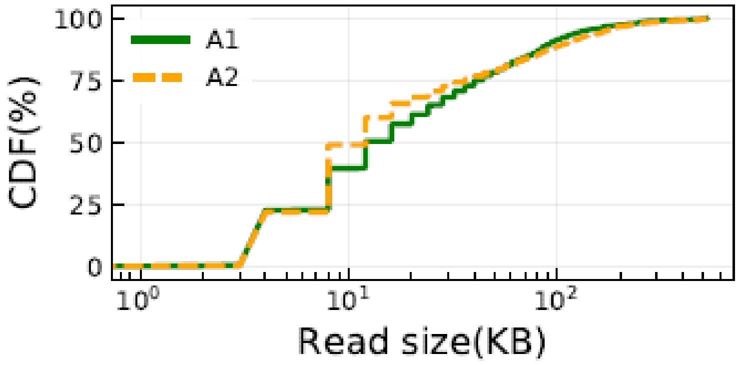
(d) B2

- Pangu achieves good load balancing and schedules across nodes.

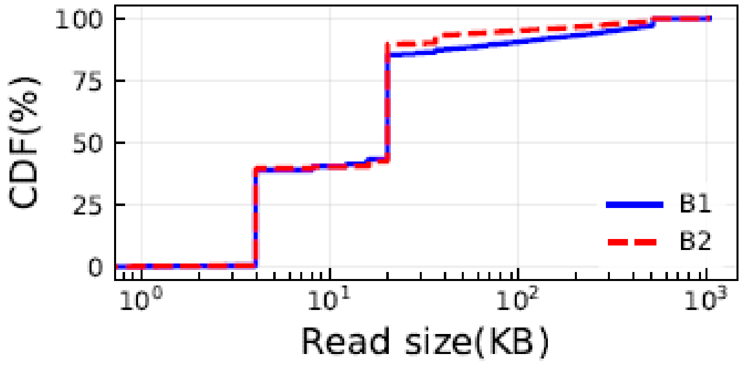
Pangu Workload

- Pangu workload traces
  - A1 and A2 : read-dominated nodes with SSD only from business I.
  - B1 and B2 : read/write mixed nodes combining SSDs and HDDs from business II.
- Load balance
  - Pangu achieves good load balancing and schedules across nodes.
- Read and write request size
  - The IO sizes for 93% of writes exceed 500KB in A nodes while the IO sizes for 95% of writes are smaller than 1KB in B nodes.
  - All four nodes have a wide range distribution of read request sizes.

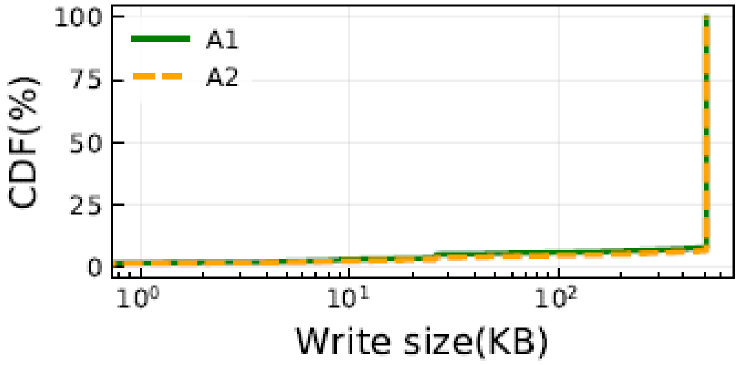
Pangu Workload



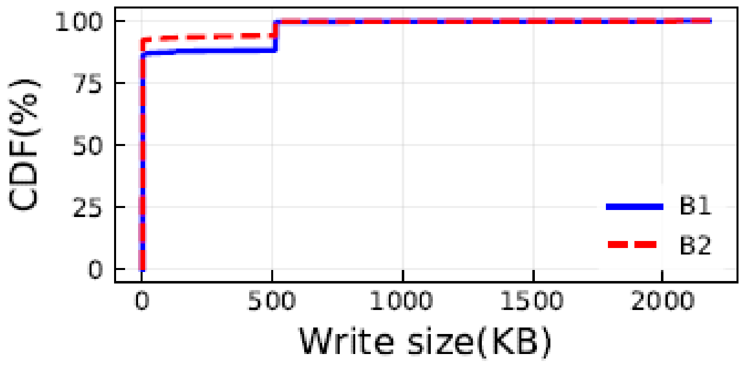
(a) Read request size of A nodes



(b) Read request size of B nodes



(c) Write request size of A nodes



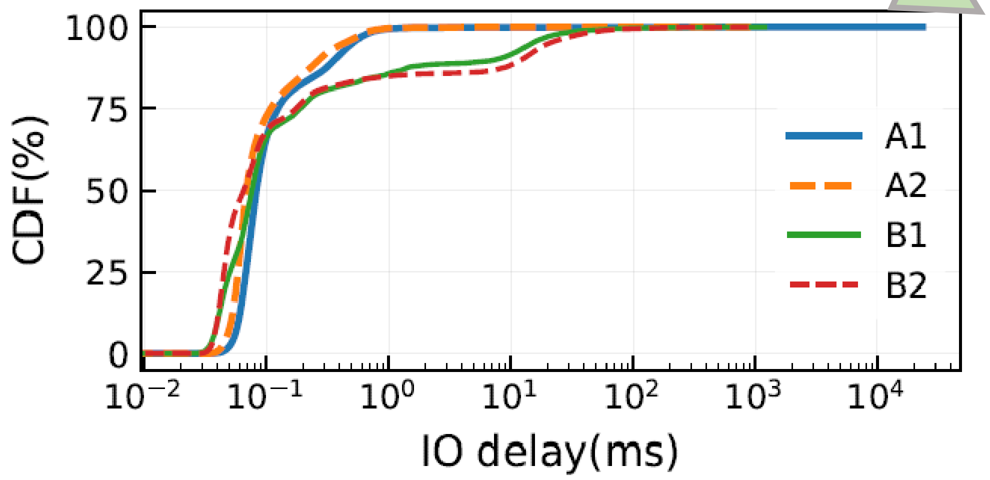
(d) Write request size of B nodes

- The IO sizes for 93% of writes exceed 500KB in A nodes while the IO sizes for 95% of writes are smaller than 1KB in B nodes.
- All four nodes have a wide range distribution of read request sizes.

Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.

long tail



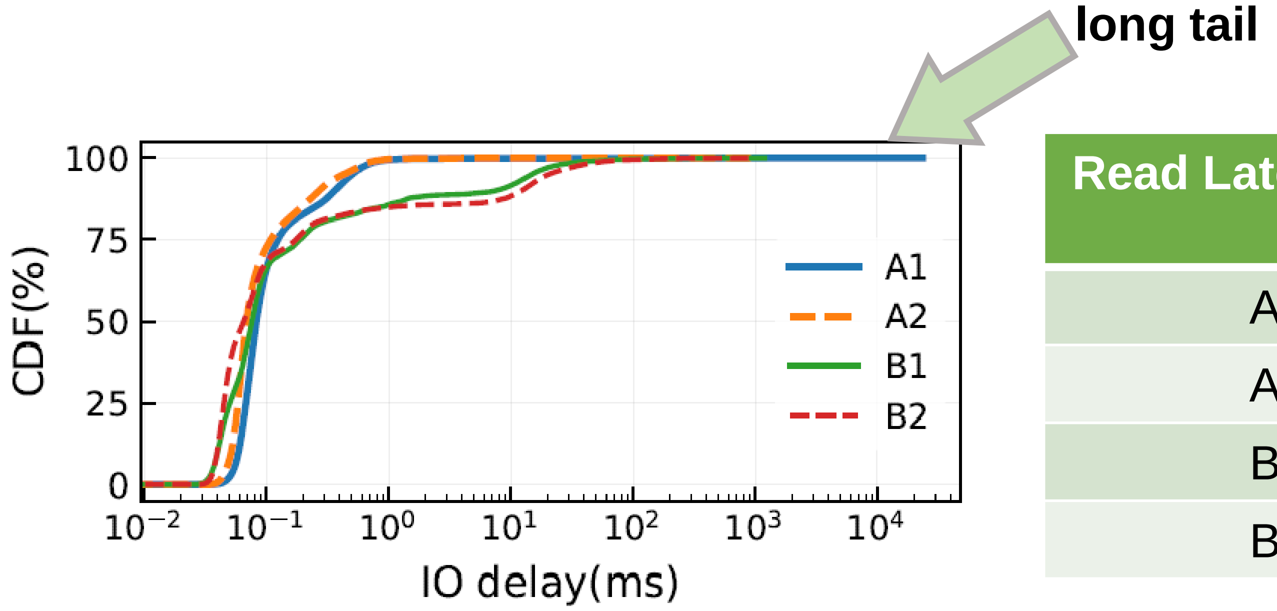
SSD-read latency CDF

Read Latency (us)	Avg.	90th	99th	99.9th	99.99th
A1	425	352	728	2950	180294
A2	127	260	658	1488	3620
B1	2651	6533	31347	156473	352759
B2	3826	11745	47523	182325	396168

Average and tail latencies of read requests

Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.



SSD-read latency CDF

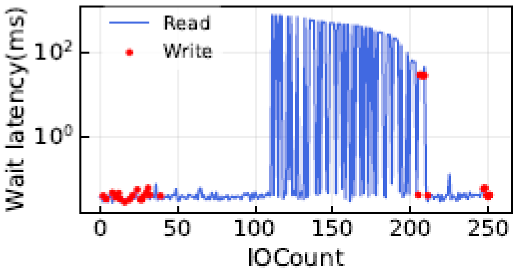
Read Latency (us)	Avg.	90th	99th	99.9th	99.99th
A1	425	352	728	2950	180294
A2	127	260	658	1488	3620
B1	2651	6533	31347	156473	352759
B2	3826	11745	47523	182325	396168

Average and tail latencies of read requests

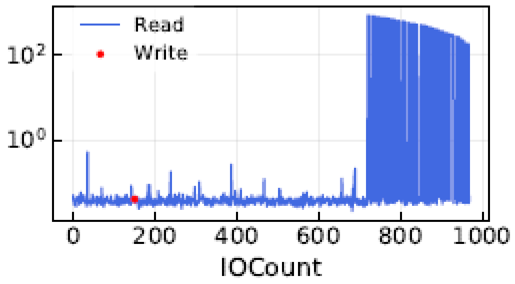
What causes the long tail latency of SSD-reads?

Motivation

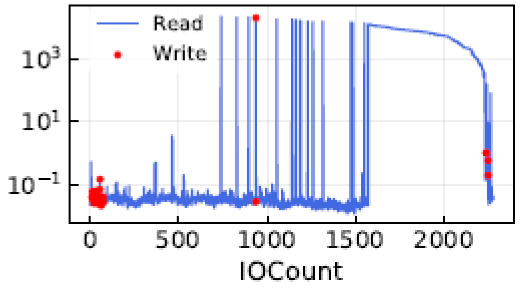
- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-GC and write-induced-blocking heavily worsen reads.



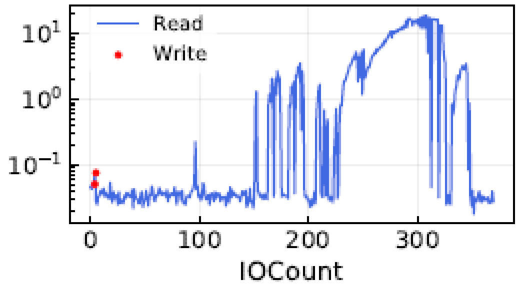
(a) Wait delay in A1 (Seq.1)



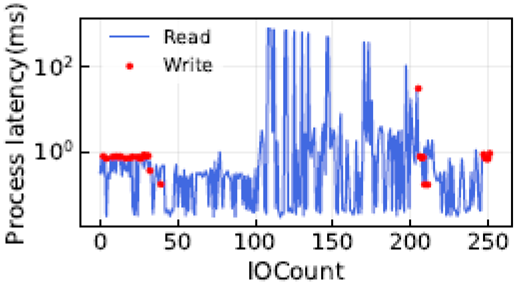
(b) Wait delay in A1 (Seq.2)



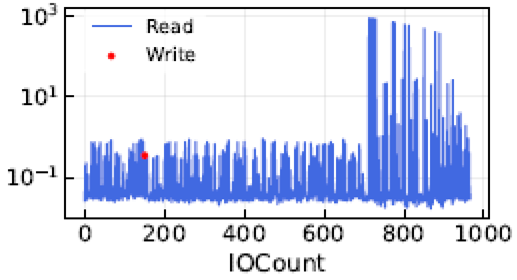
(c) Wait delay in A1 (Seq.3)



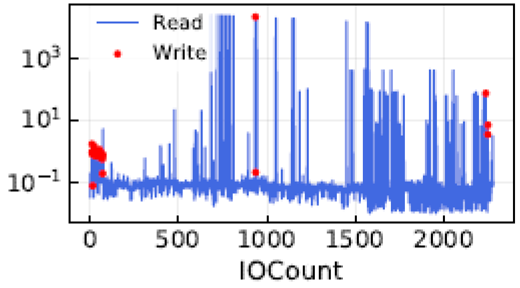
(d) Wait delay in A2 (Seq.4)



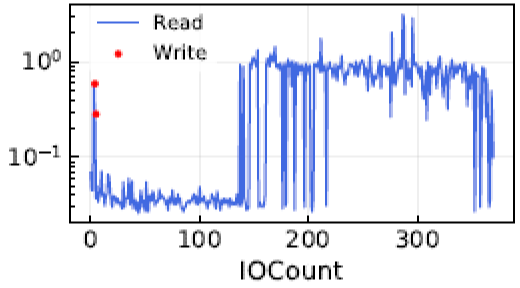
(e) Process delay in A1 (Seq.1)



(f) Process delay in A1 (Seq.2)



(g) Process delay in A1 (Seq.3)

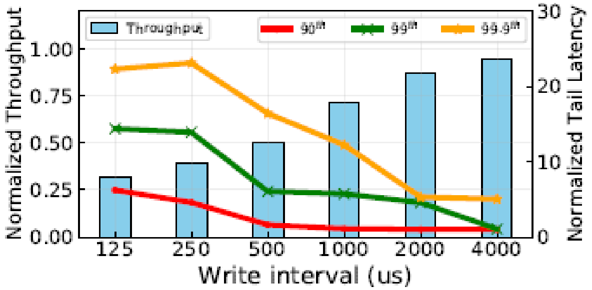


(h) Process delay in A2 (Seq.4)

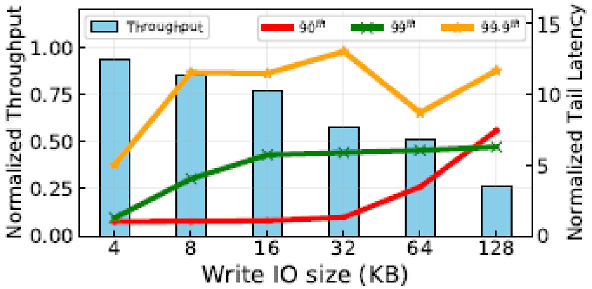
Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-GC and write-induced-blocking heavily worsen reads.
- We conduct a set of read/write mixed experiments on SSDs to effectively validate and understand the read/write contention on SSDs besides of Pangu.

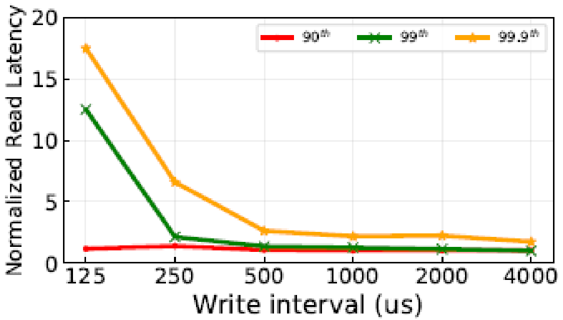
Motivation



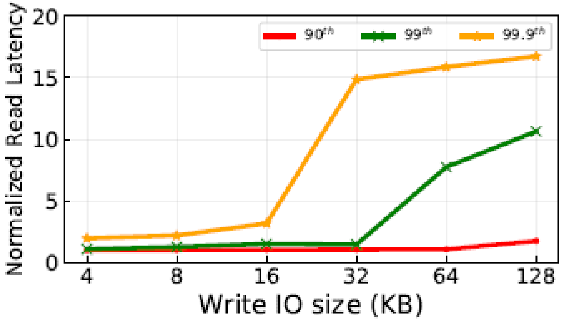
(a) High-intensity reads vs. writes interval



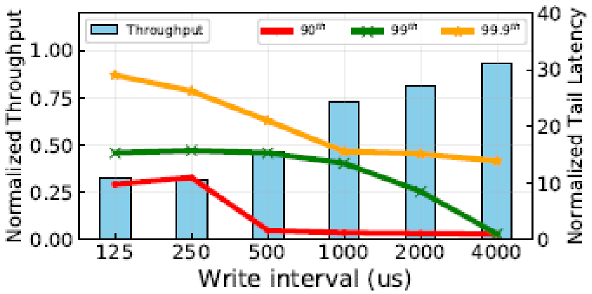
(b) High-intensity reads vs. writes IO sizes



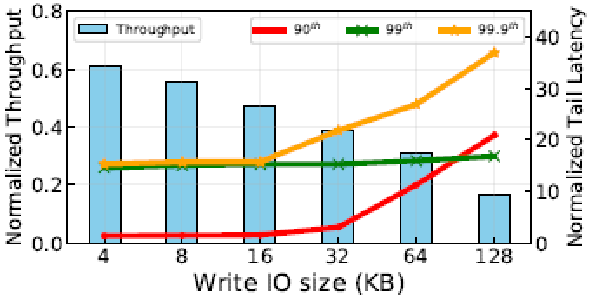
(e) Low-intensity reads vs. writes interval



(f) Low-intensity reads vs. writes IO sizes



(c) Mid-intensity reads vs. writes interval

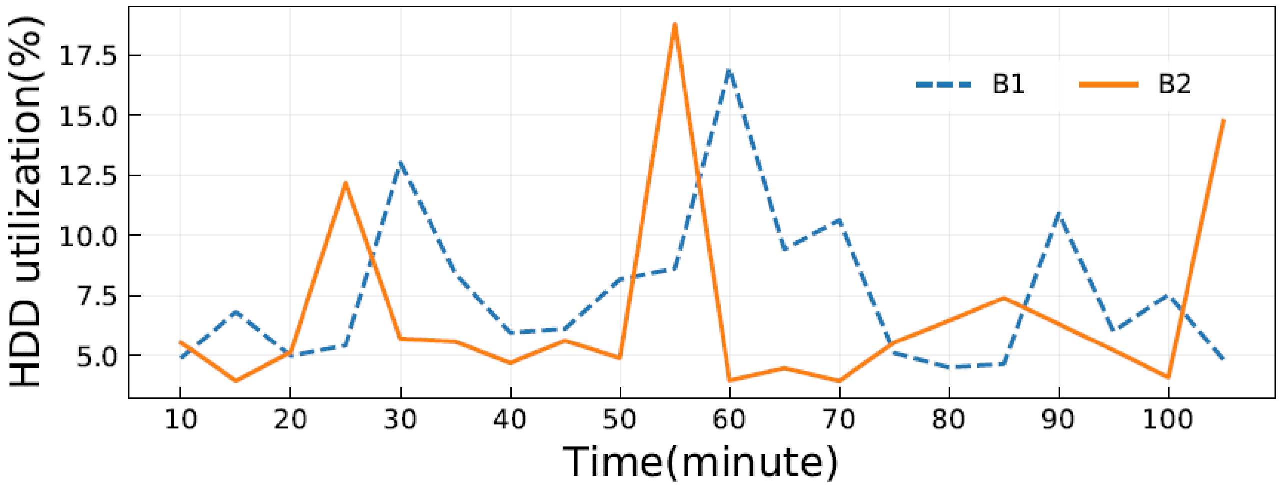


(d) Mid-intensity reads vs. writes IO sizes

- The read performance of FIO under concurrent writing is significantly lower than a read-only FIO.
- Even small and discrete write requests could cause high tail latency for SSD reads.
- Large write IOs take more time and hardware channels, resulting in severe blockage.
- An light-load writing can remarkably impact reads.
- The performance slowdown on the mid-intensity case is even higher than the high-intensity case.

Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-GC and write-induced-blocking heavily worsen reads.
- We conduct a set of read/write mixed experiments on SSDs to effectively validate and understand the read/write contention on SSDs besides of Pangu.
- The traditional SFL mode makes SSDs heavily loaded, while the HDDs are always underutilized due to its role as the secondary storage.



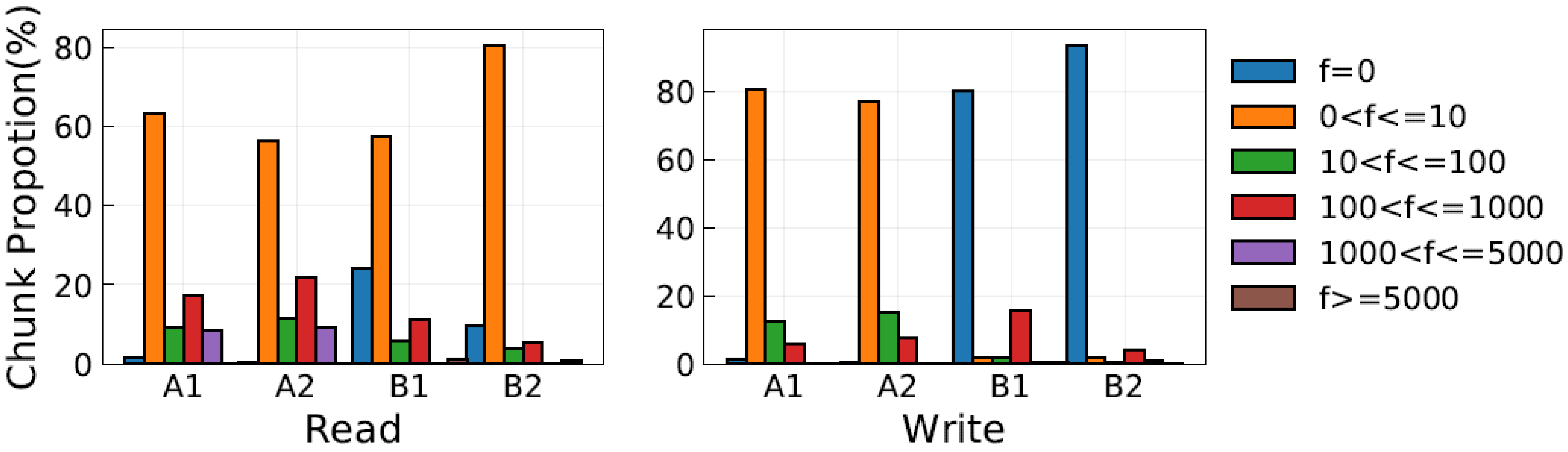
- ➢ The HDD utilization in B nodes is less than 10% on average.

Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-GC and write-induced-blocking heavily worsen reads.
- We conduct a set of read/write mixed experiments on SSDs to effectively validate and understand the read/write contention on SSDs besides of Pangu.
- The traditional SFL mode makes SSDs heavily loaded, while the HDDs are always underutilized due to its role as the secondary storage.
- The Chunk Accessing Behavior reveals that a fixed-size SSD space allocated to a large read cache can gain more than giving it to a large write buffer.

Motivation

Proportion of chunks to all accessed chunks under different frequency ranges



- For A nodes, more than 60% chunks and more than 80% chunks are **read and written less than 10 times**. For B nodes, about 80% chunks are **read less than 10 times** while 80% chunks are **never written**.

Motivation

- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-G
- W nd u
- The traditional SFL mode makes SSDs heavily loaded, while the HDDs are always underutilized due to its role as the secondary storage.
- The Chunk Accessing Behavior reveals that a fixed-size SSD space allocated to a large read cache can gain more than giving it to a large write buffer.

How to exploit the **underutilized HDD** to relieve the pressure of SSDs in hybrid storage nodes?

Motivation

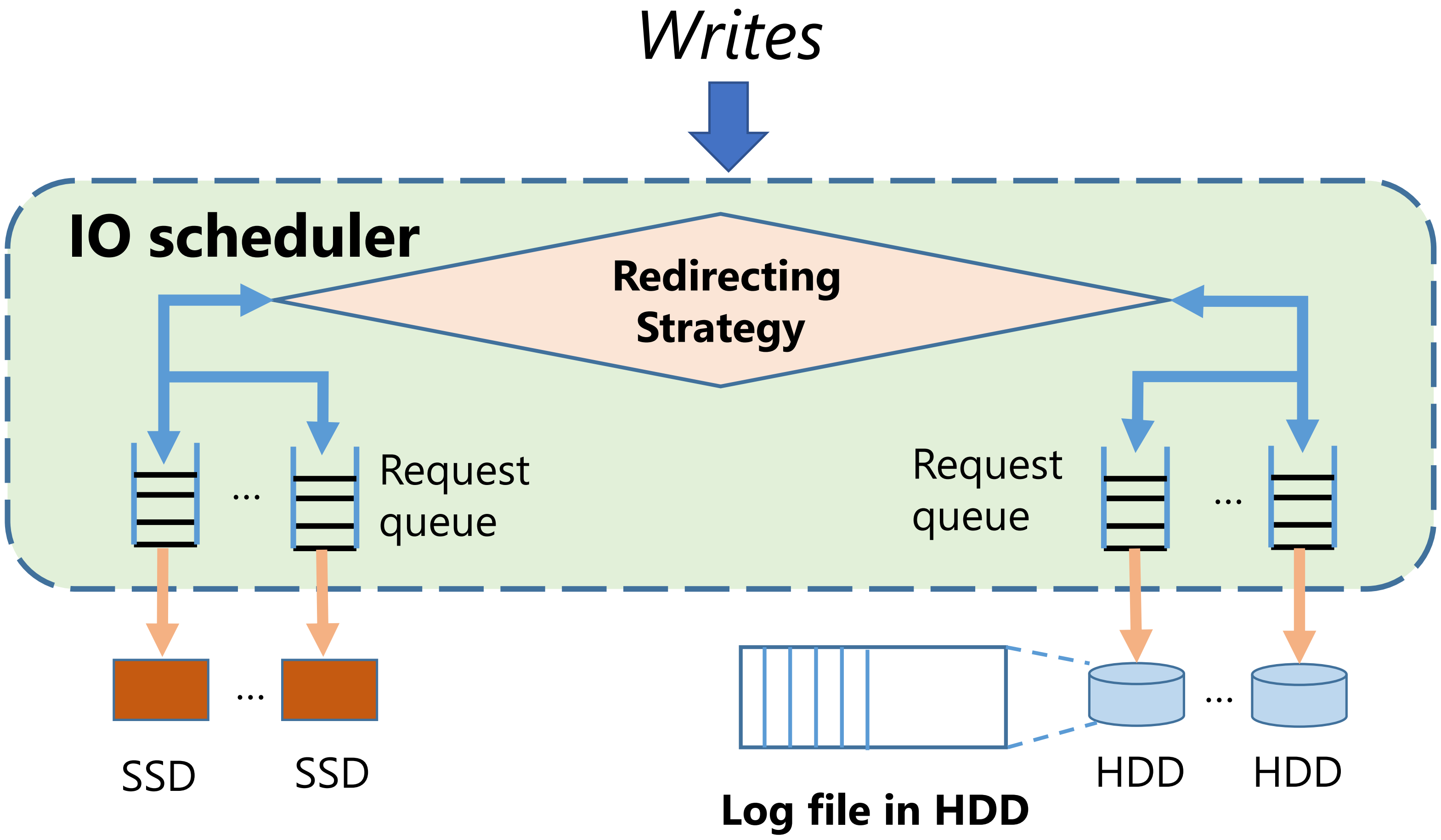
- SSD reads with an IO delay of less than 50 μs often suffer from long ms-level read latency in term of both average and tail.
- The typical IO sequences from the traces confirm the phenomena where the write-induced-G
- W nd u
- The traditional SFL mode makes SSDs heavily loaded, while the HDDs are always underutilized due to its role as the
- The Chunk Accessing Behavio ize SSD space allocated to a large read cache can gain more tha e write buffer.

How to exploit the **underutilized HDD** to relieve the pressure of SSDs in hybrid storage nodes?

SeRW

Outline

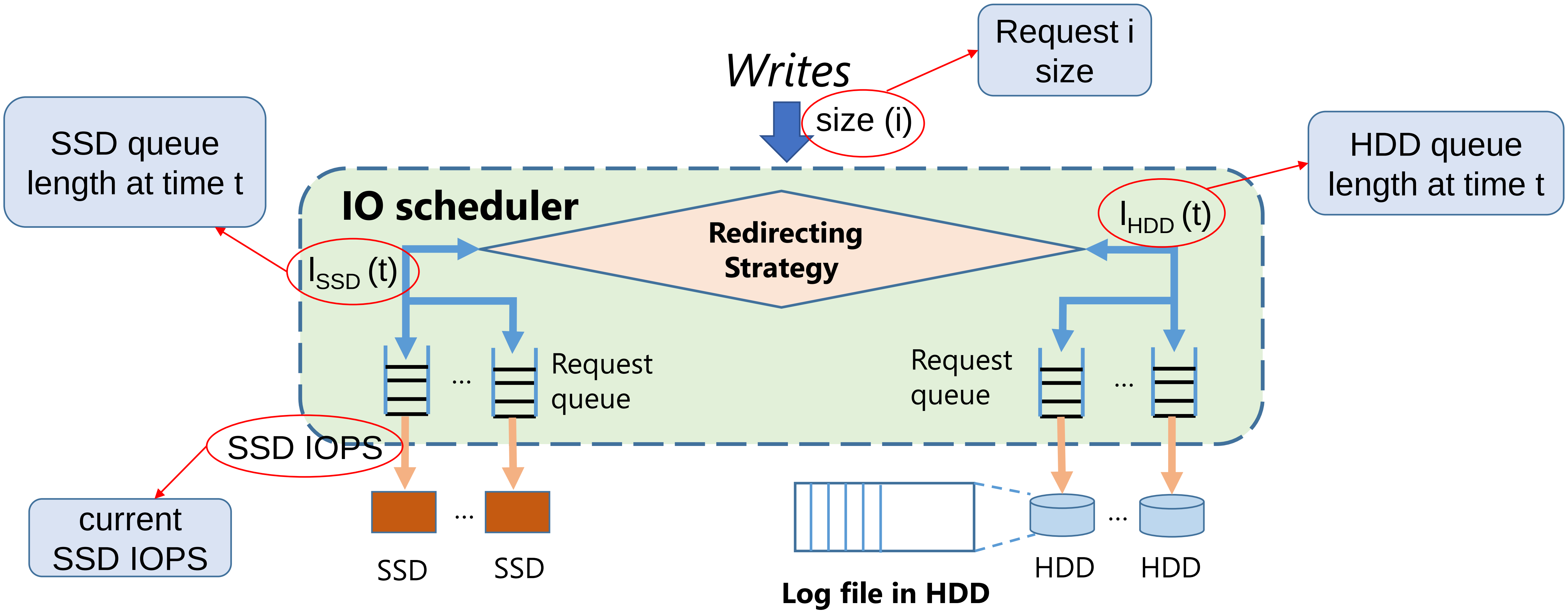
IO Scheduler (SeRW)



- An adaptive IO scheduler to separate read and write upon SSDs of hybrid storage servers at runtime.
- Architecture
  - A redirecting scheduler monitoring all request queues of SSDs and HDDs at runtime.
  - Log file in each HDD.

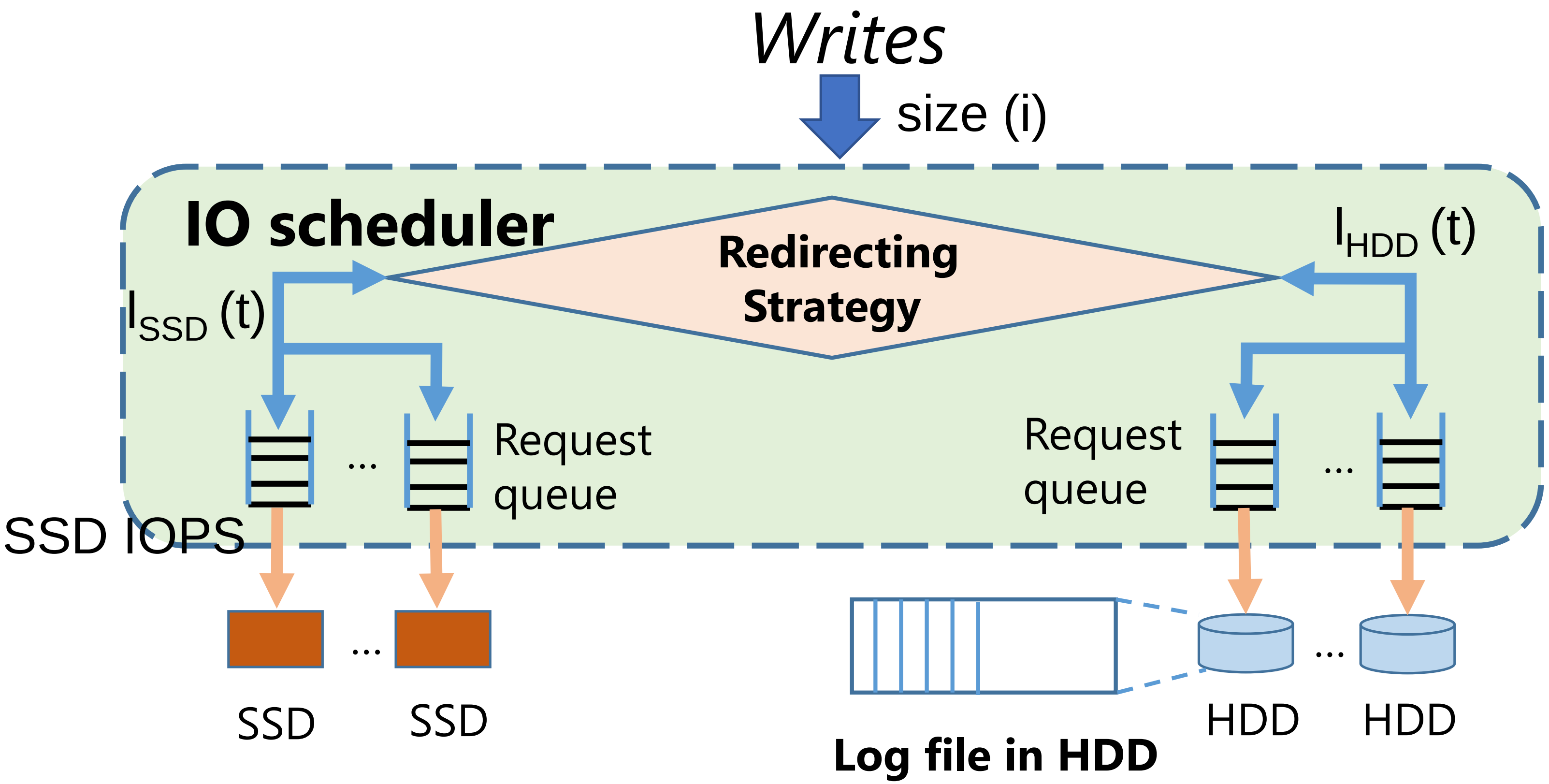
IO Scheduler (SeRW)

- Four key parameters



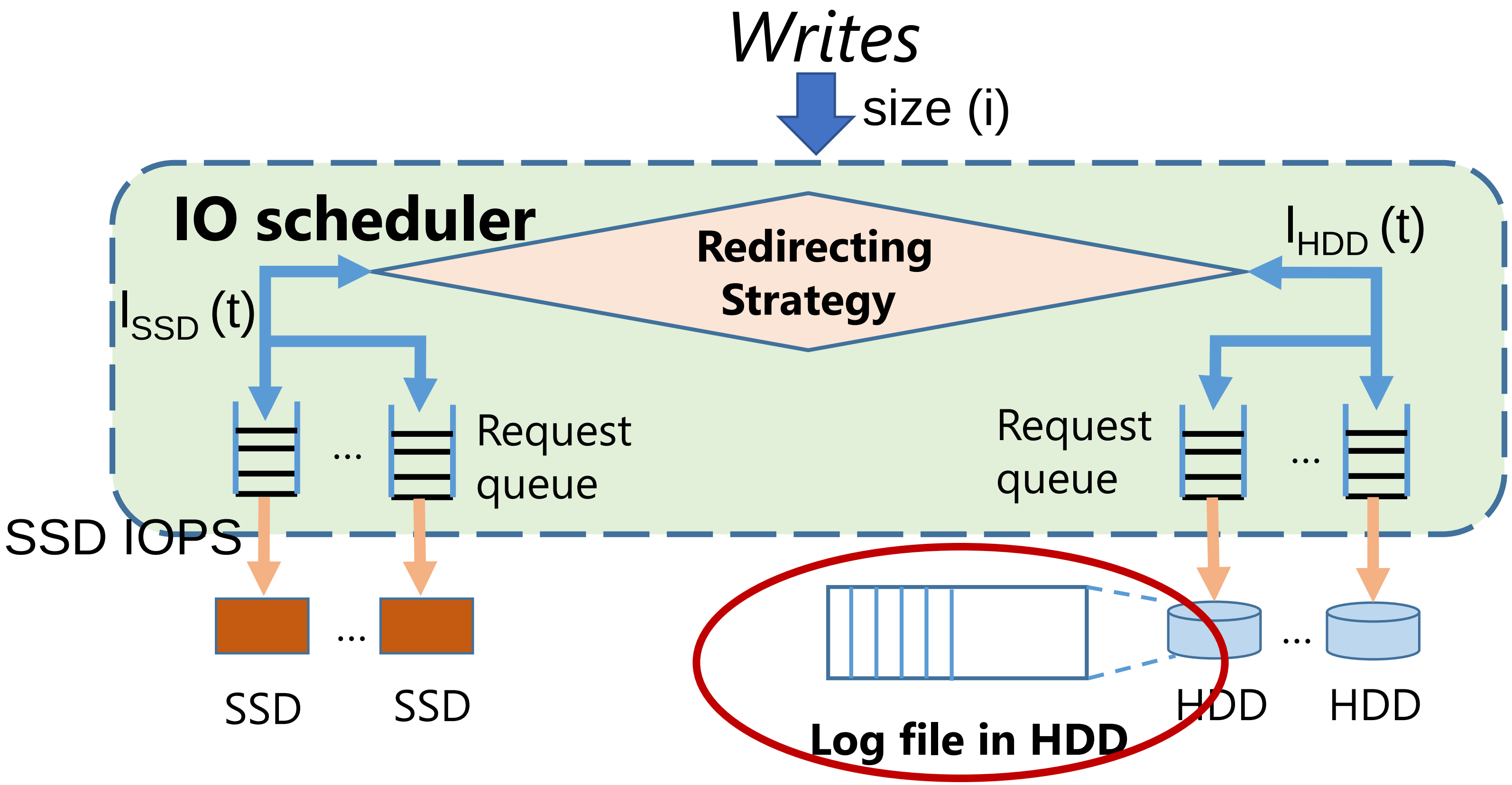
Redirecting Strategy

- Redirect SSD writes to idle HDDs when:
  - The IOPS of an SSD is higher than a threshold I.
  - I_{SSD} (t) is larger than a threshold L or size (i) is larger than a size threshold S.



Log Mechanism

- To take full advantage of HDD sequential-write performance, SeRW writes redirected data into a log file in an append-only way. The DIRECT_IO mode is turned on to accelerate the data persistence process.



Outline

- ✓ Introduction
- ✓ Background
- ✓ Analysis and Motivation
- ✓ Design of SeRW
  - ➢ Redirecting Strategy
  - ➢ Log Machanism
- ✓ **Evaluation**
- ✓ Conclusion

Experimental Setups

- Comparisons
  - **Baseline**: Pangu workload replay (**SFL**)
  - **SeRW**

Experimental Setups

- Comparisons
  - **Baseline**: Pangu workload replay (**SFL**)
  - **SeRW**
- Evaluation environment

System	Linux server
CPU	Intel Xeon E5-2696 v4 (2.20 GHz, 22 CPUs)
Memory	DDR3 DRAM 64GB
SSD	Samsung SM961 256GB (NVMe, 2.8GB/s read and 1.2 GB/s write at peak)
HDD	West Digital WD40EZRZ 4TB (180 MB/s peak throughput)

Experimental Setups

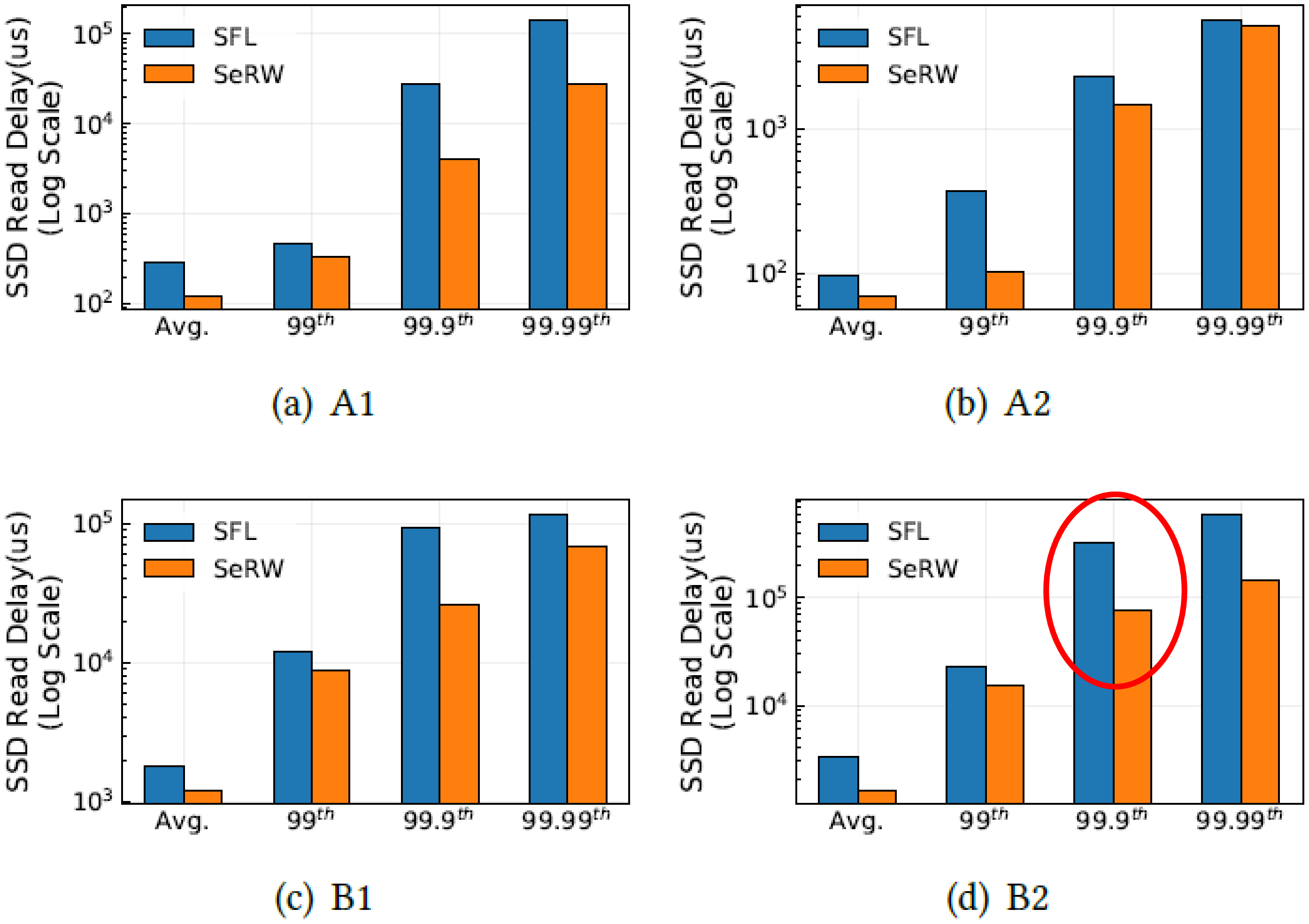
- Comparisons
  - **Baseline**: Pangu workload replay (**SFL**)
  - **SeRW**

- Evaluation environment

System	Linux server
CPU	Intel Xeon E5-2696 v4 (2.20 GHz, 22 CPUs)
Memory	DDR3 DRAM 64GB
SSD	Samsung SM961 256GB (NVMe, 2.8GB/s read and 1.2 GB/s write at peak)
HDD	West Digital WD40EZRZ 4TB (180 MB/s peak throughput)

- Threshold selection
  - The redirected write size threshold S : the 50^{th}-percentile write size of all writes
  - The mid/high IOPS threshold I : the 50^{th}-percentile of IOPS
  - The SSD queue length threshold L : 3
  - The HDD queue length $l_{HDD}(t)$: 0

Read Performance



Average and tail read latency with SeRW and SFL

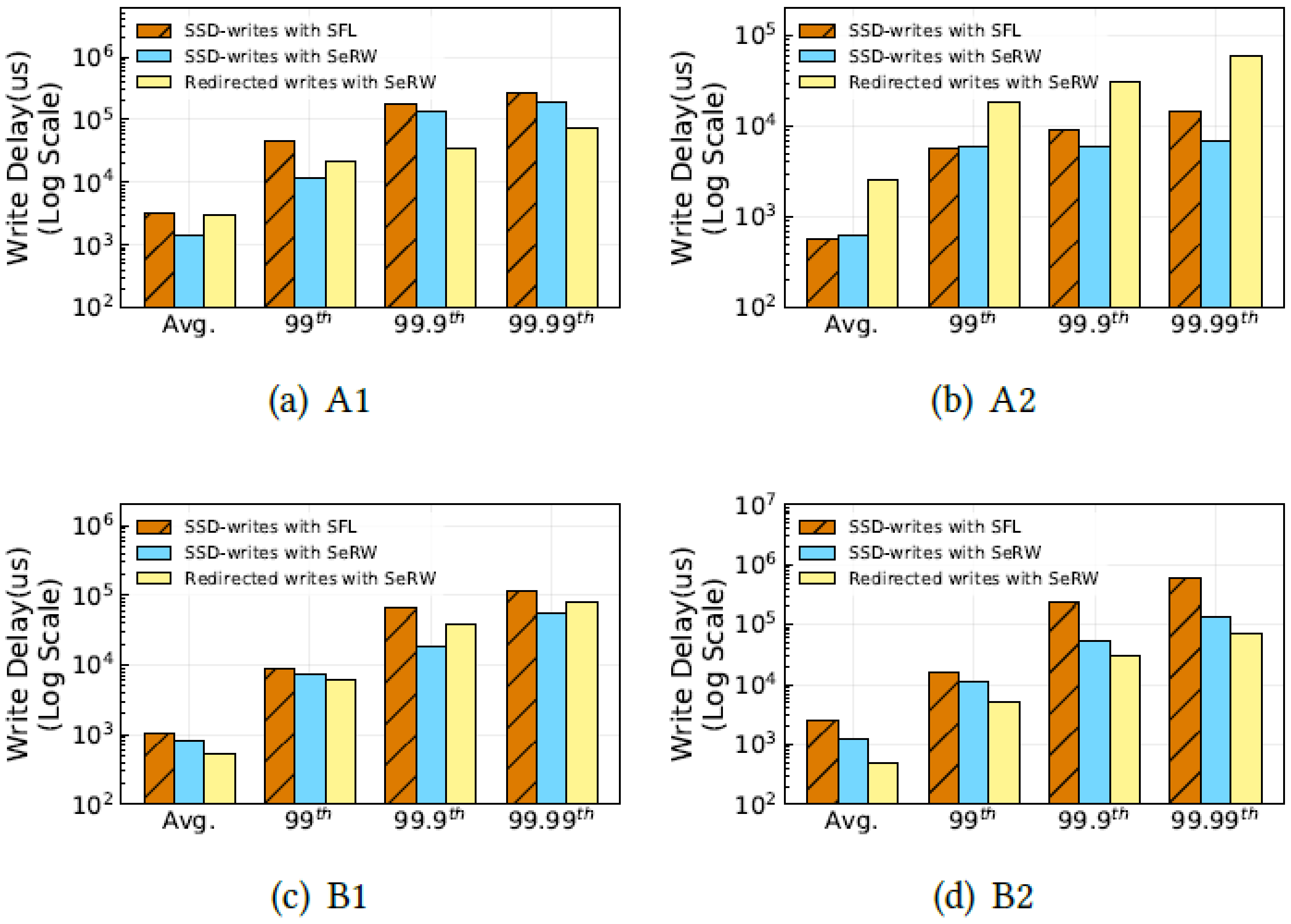
- SeRW significantly and consistently reduces the average and tail latency in all nodes, especially for A1, B1, and B2 with mid/high intensity.
- B2 node gains the most benefit. Its 99th, 99.9th, 99.99th-percentile latency reduces by 32.2%, 76.7%, and 76.4%.

SSD Written Data Reduction

Node Type	A1	A2	B1	B2
SSD data written with SFL (GB)	34.9	26.6	44.9	46.9
SSD data written with SeRW (GB)	28.4	16.6	40.9	42.1
Redirected write requests (%)	17.4	35.6	1.6	2.7

- SeRW effectively reduces the amount of data written to SSD by 18.5% in A1, 37.5% in A2, 8.8% in B1, and 10.2% in B2.
- The SSD-write reduction also means that SeRW mitigates SSD wearout, increasing the lifetime of SSD relative to SFL.

Write Performance

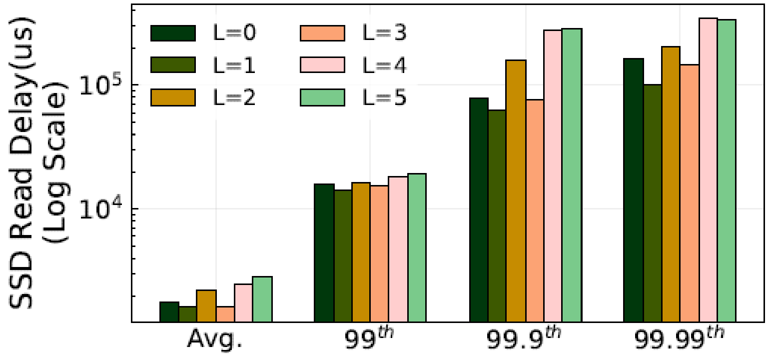


Average and tail write latency with SFL and SeRW

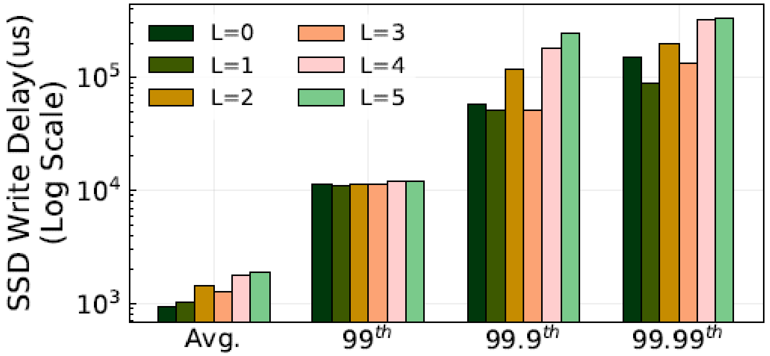
- SeRW does not significantly increase the overall average and tail latencies combining SSD-writes and HDDwrites.
- For A1 and B1 with high intensity, the latency of HDD-writes is even better than that of SSD-writes.

Impact of Thresholds

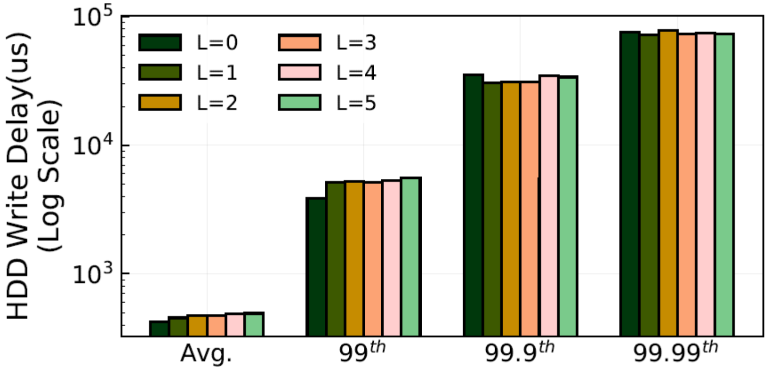
- **Queue length threshold L**



(a) SSD read delay



(b) SSD write delay



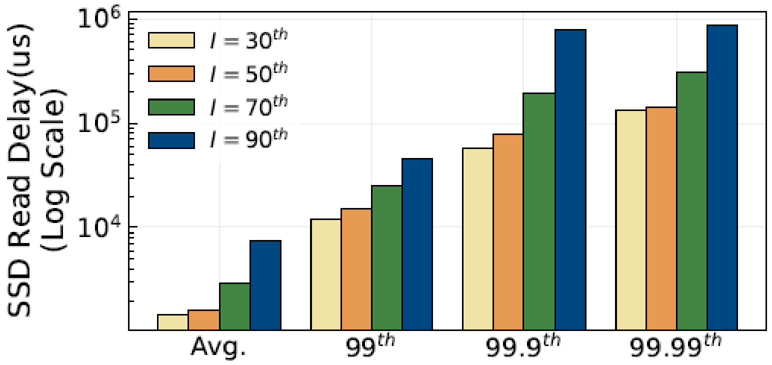
(c) HDD write delay

➢ With a higher L, only the fewer burst cases could trigger redirecting writes. As a result, SeRW has to execute more SSD-writes and is more likely to suffer the SSD queueing blockage.

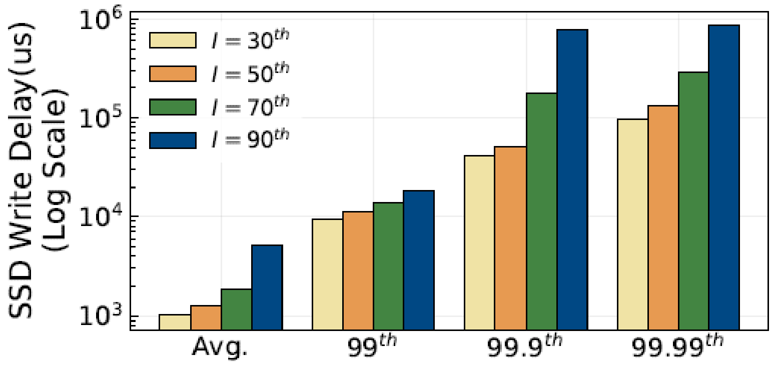
➢ The L value has no remarkable impact on HDD-writes performance, as well as the write amount reduction within 0.5%.

Impact of Thresholds

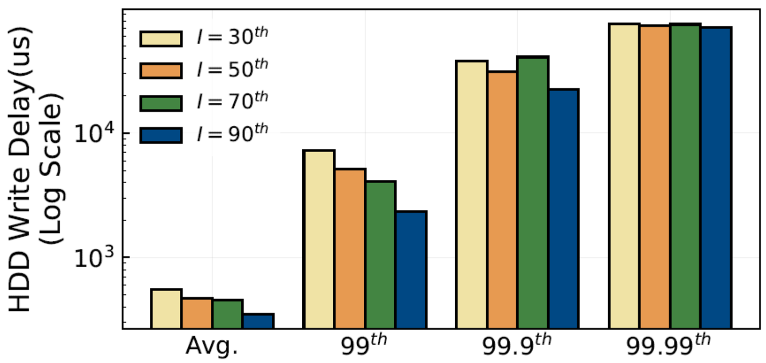
- **Workload Intensity Threshold I**



(a) SSD read delay



(b) SSD write delay



(c) HDD write delay

- The average and tail SSD read/write latency are significantly increased with higher I value.

- The average and 99^{th}-percentile latencies of HDD-writes are significantly increased with an increase of I value but the 99.99^{th}-percentile latency of HDD-writes is almost unchanged in these five cases.

Outline

Conclusion

- SSD-HDD hybrid storage in clouds.
  - ✓ SSDs as the primary storage directly serving requests from front-end applications.
  - ✓ HDDs as the secondary storage to provide sufficient storage capacity.
- Writes mixed with mid/high intensive reads upon SSDs dramatically increase read-latency, especially for tail latency.
  - ✓ These long read latencies are primarily caused by (1) write-induced-blocking and (2) write-induced-garbage-collection (GC).
- We present a SeRW scheduling approach.
  - ✓ The main idea is to adaptively steers some SSD-writes to idle HDDs in running time.
- SeRW relieves the write-blocking read delay on SSDs at mid/high load and reduces the amount of data written into SSDs.
  - ✓ SeRW decreases the average, 99^{th} , 99.9^{th}, 99.99^{th}-percentile latencies of reads by up to 2.07x, 1.48x, 4.29x, and 4.24x, respectively.
  - ✓ Reducing the amount of data written to SSDs by up to 37.5%.

Thank you!