BCW: Buffer-Controlled Writes to HDDs for SSD-HDD Hybrid Storage Server

Shucheng Wang[1], Ziyi Lu[1], Qiang Cao[1], Hong Jiang[2], Jie Yao[1], Yuanyuan Dong[3] and Puyuan Yang[3]

[1]Huazhong University of Science and Technology,

[2]University of Texas at Arlington, [3]Alibaba Group

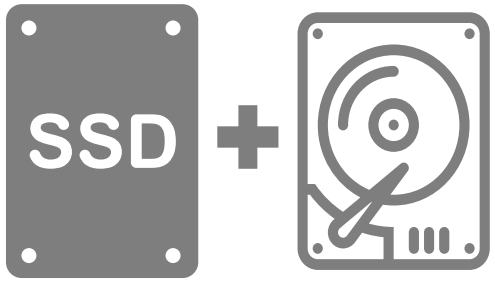
FAST 2020

Background

- SSD
  - High-speed performance & Low latency
  - Expensive & DWPD (Drive Writes Per Day) limitation

- HDD
  - Large capacity & Low cost & No writes limitation
  - Low-speed performance & High latency

- **SSD-HDD Hybrid Storage**
  - Cost-effectiveness
  - High-speed & low-latency



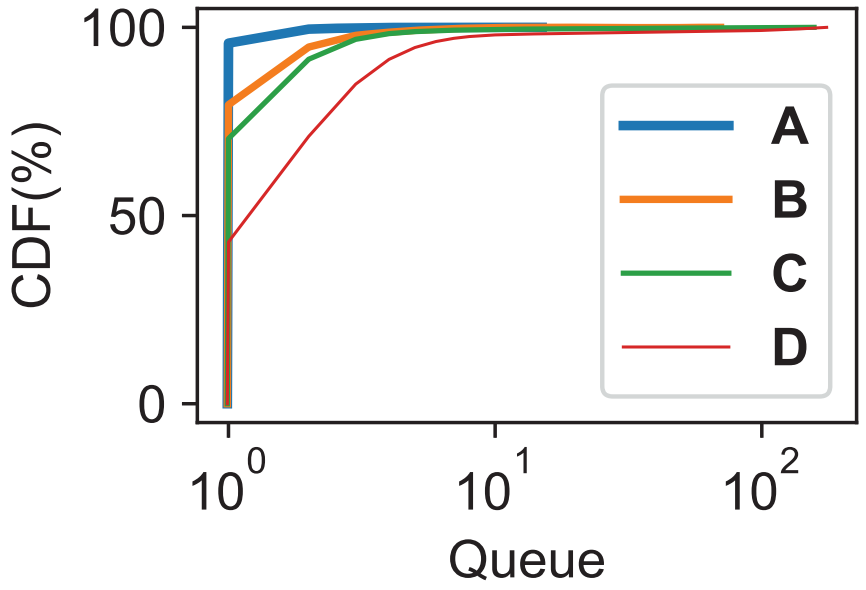
Write-intensive Workload

- Workload characteristics

Workload Types	A	B	C	D
Business	Cloud Computing	Cloud Storage	Structured Storage	Structured Storage
SSD Writes (GB)	14.7	**61.2**	7.2	7.5
SSD Write Requests (millions)	**0.43**	4.4	4.8	4.7
Note	**Lowest IO intensity**	**Most written data**		**Peak: 11KRPS**

Write-intensive Workload

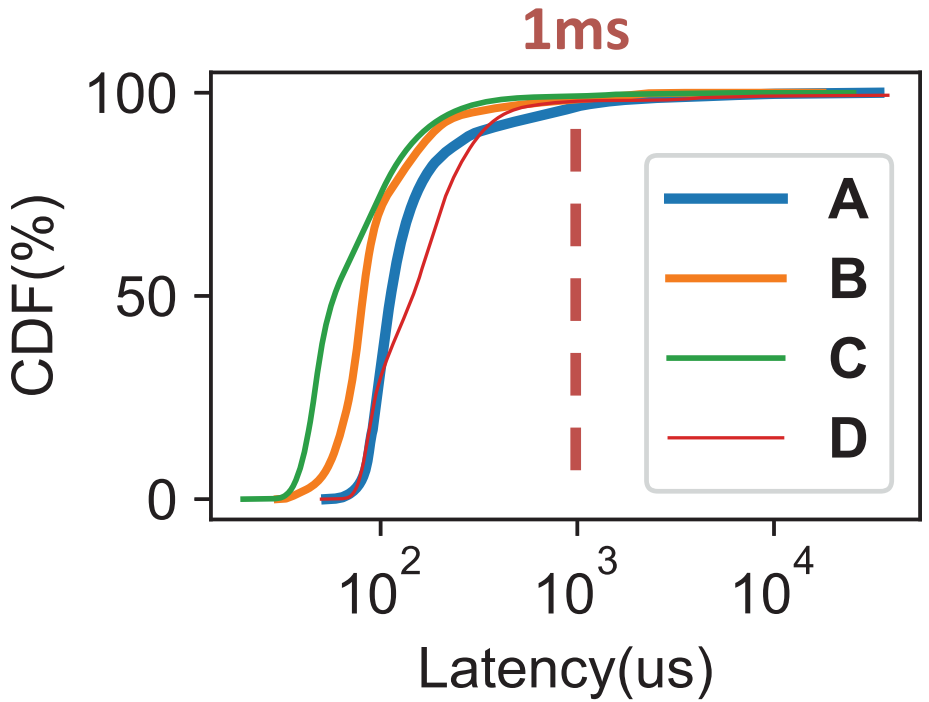
- SSD performs badly under heavy write workload
  - Peak write requests per second > 10KRPS
  - Daily write per day > 3TB → *Reach limited DWPD of SSD*



- Large queue length
  - Large writes (e.g., 1MB)
  - Frequent GC caused by high write intensity

Write-intensive Workload

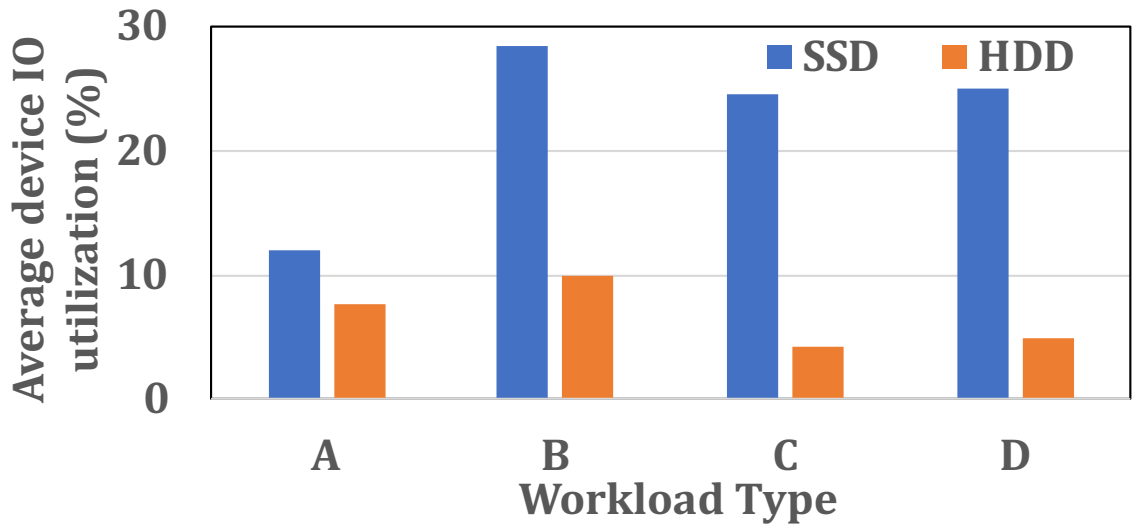
- SSD performs badly under heavy write workload
  - Peak write requests per second > 10KRPS
  - Daily write per day > 3TB → *Reach limited DWPD of SSD*



- Long tail latency
  - 99th percentile latency is *10ms*
  - 99.9th percentile latency is *50ms*

HDD Underutilization

- HDD is underutilized
  - **1/5 – 1/2** of SSD utilizations
  - **90% - 95%** of times are in idle state (*less than 10% utilization*)



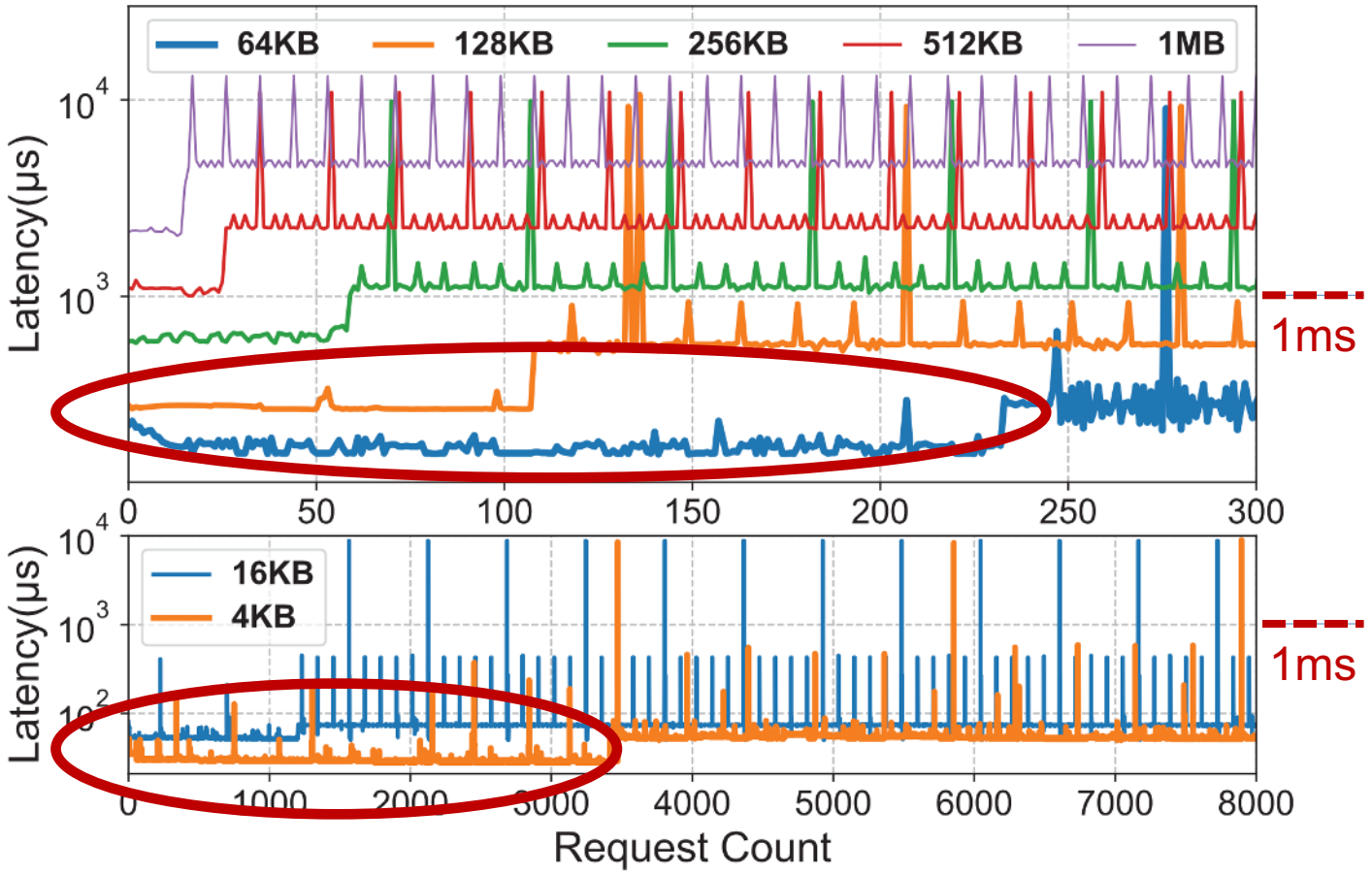
Motivation

- SSD suffers from heavy-write pressure
  - Long tail latency encountered
  - Large queue length
- HDD is underutilized (less than 10%)

- ***Can we utilize HDD properly and maintain acceptable latency?***

Average Latency of HDD < **Tail** Latency of SSD ✓

HDD Write Behaviors

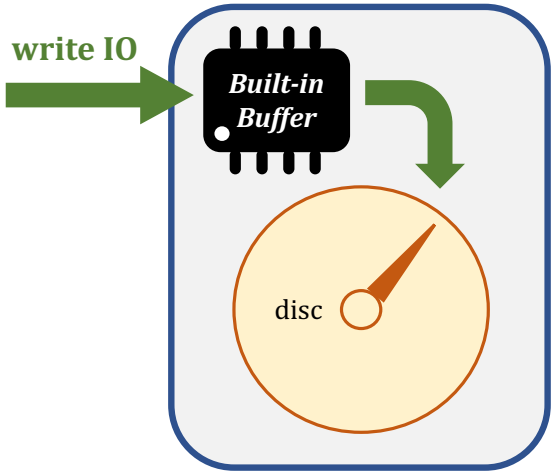


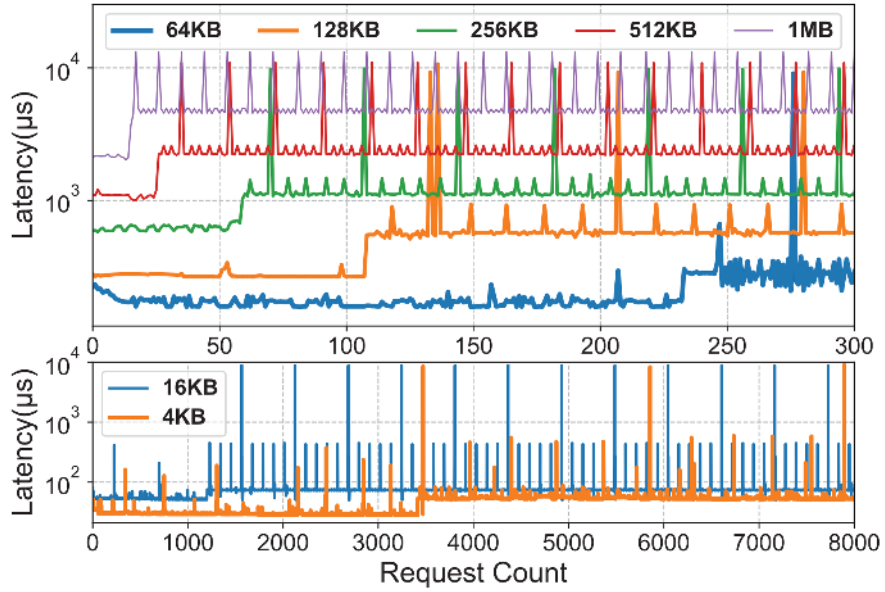
10TB West Digital HDD

- HDD can achieve *µs-level* write latency
  - 66µs for 16KB writes
- *ms-level* latency spikes
- Fixed periodic write pattern
  - Fast (low latency)
  - Mid (mid latency)
  - Slow (high latency spikes)

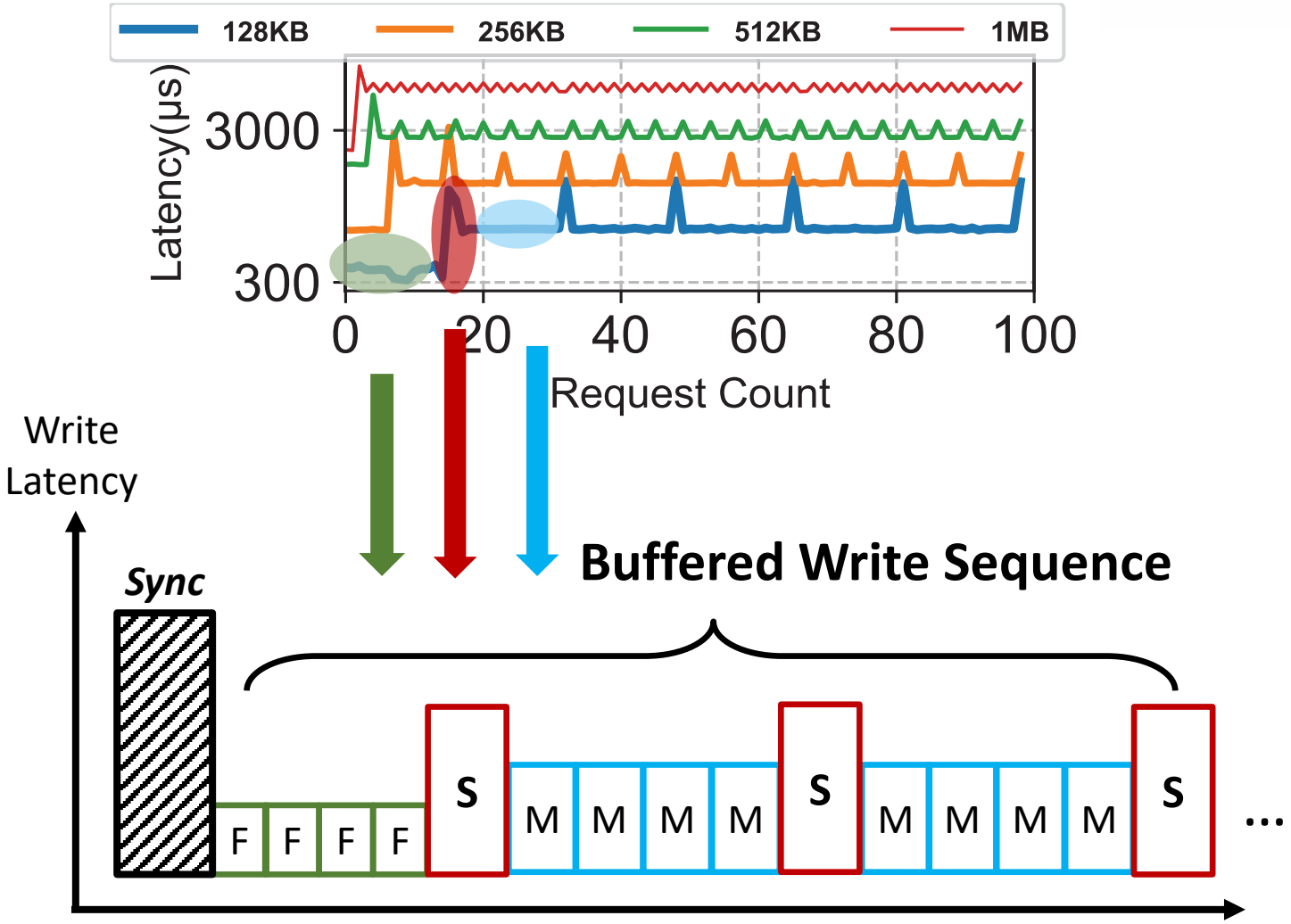
HDD Built-in Buffer

- Built-in buffer causes this write behaviors
  - Part of buffer can be used to buffer incoming write IOs (16MB in this tested WD 10TB HDD)
  - Remaining capacity is used by HDD itself



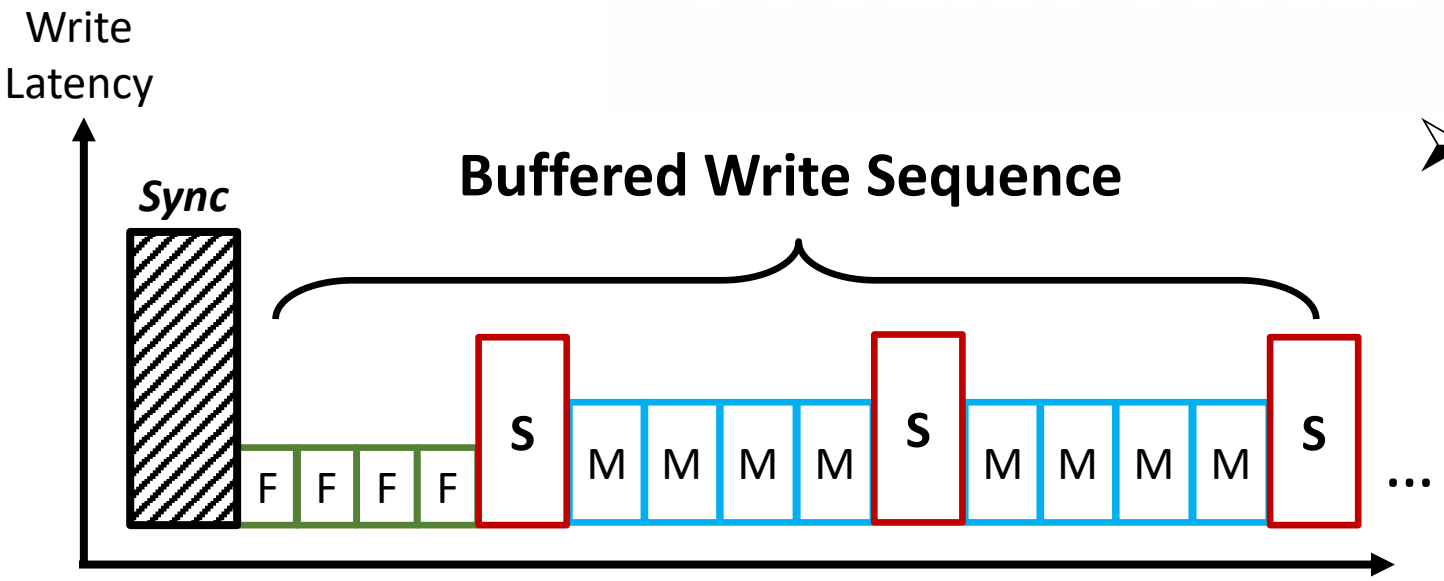


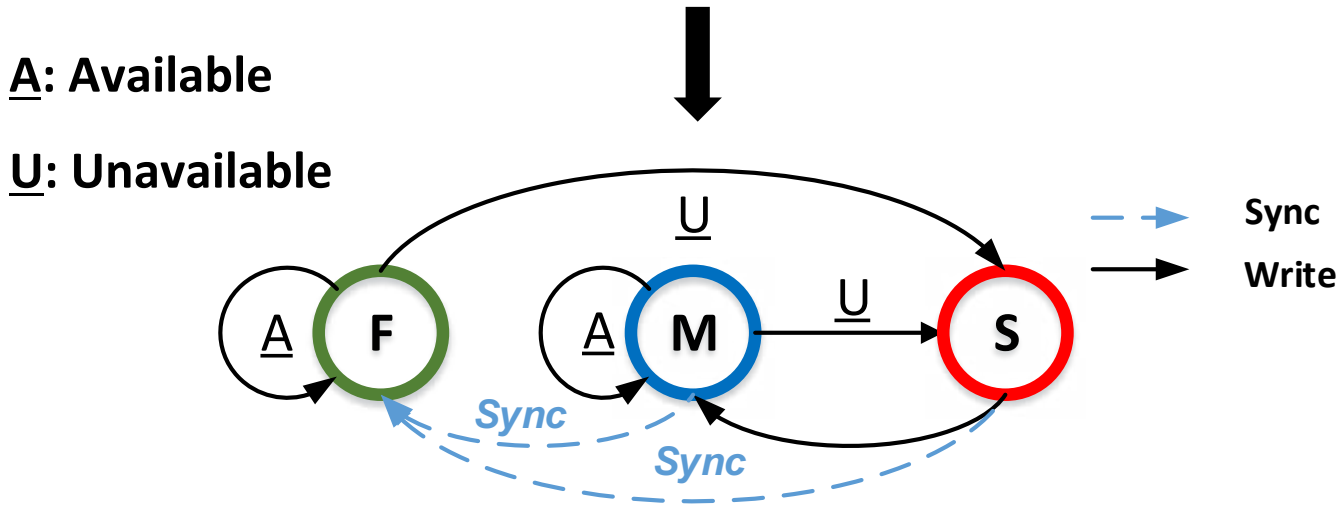
HDD Buffered-Write Model



- Three types of HDD buffered write
  - **F**ast write (low-latency)
  - **M**id write (mid-latency)
  - **S**low write (high-latency spike)
- Buffered write sequence
  - Start with Fast write
  - Followed by Slow-and-Mid write pair
  - After *sync()*

Predict Write State



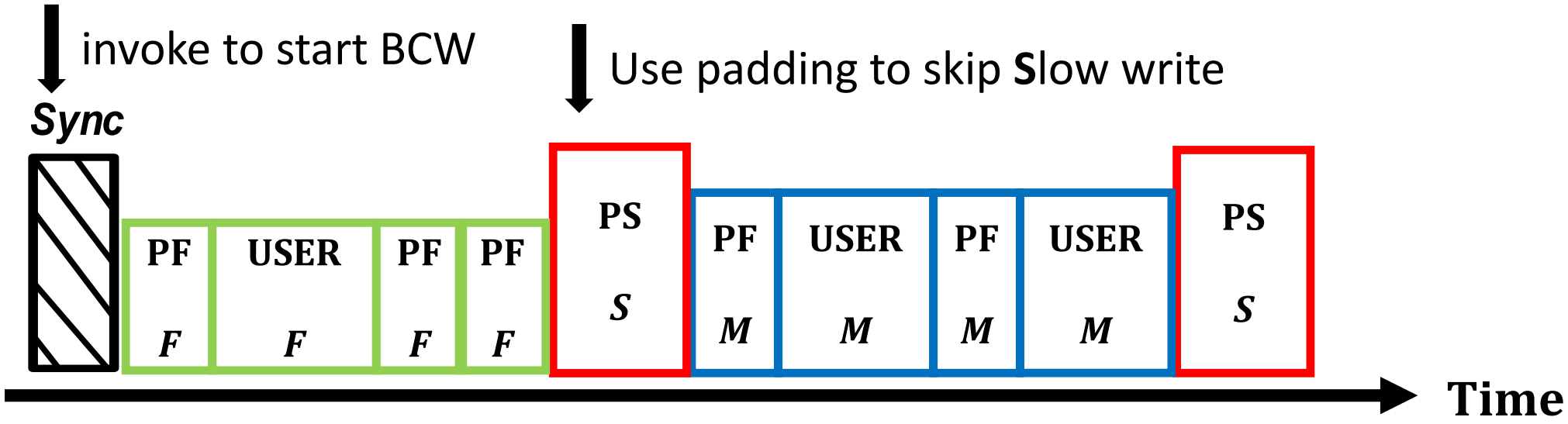


- Three types of HDD buffered write
  - **F**ast write (low-latency)
  - **M**id write (mid-latency)
  - **S**low write (high-latency spike)

- Prefer to predict **S**low write
  - Avoid penalty from misprediction

Buffer-Controlled Writes (BCW)

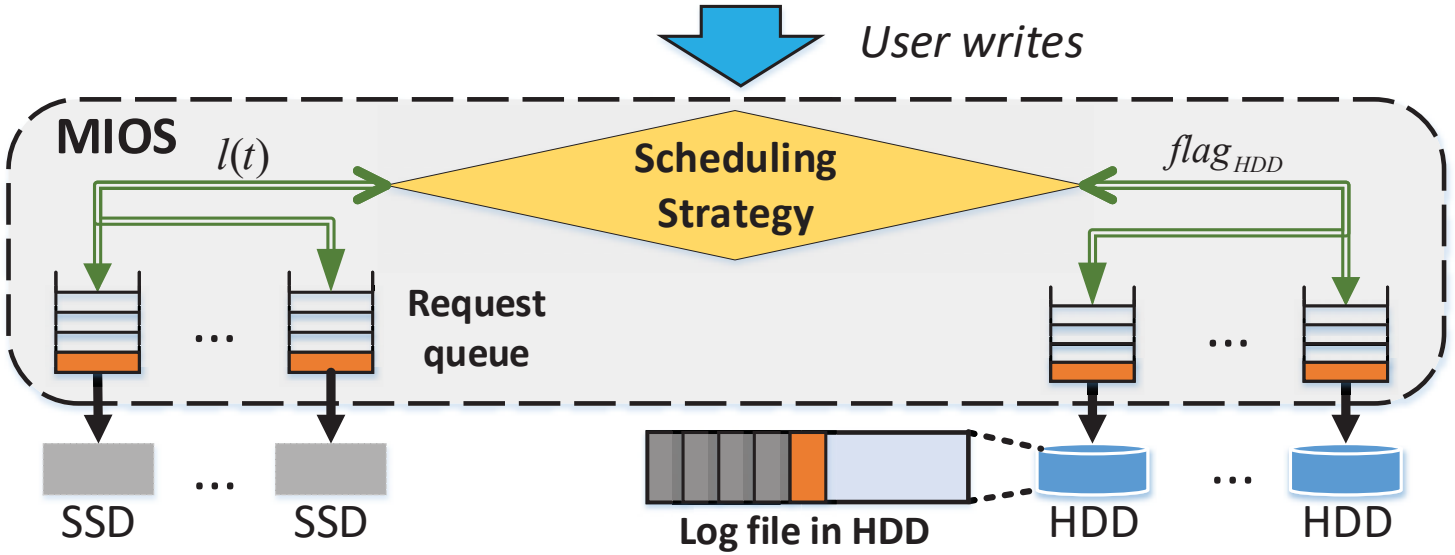
- Actively pad non-user data to HDD when there are no user requests to keep predictor consistent
  - **PF**: padding for ***F*** and ***M*** stage (4KB)
  - **PS**: padding for ***S*** stage (64KB)



- *Profiling* process is performed to determine the key parameters for predictor

Mixed IO scheduler (MIOS)

- Monitor the request queue length of SSD and HDD
  - Redirect writes to HDD when latency in SSD is larger than ***F*** or ***M*** states in HDD
  - Log file in append-manner in HDD

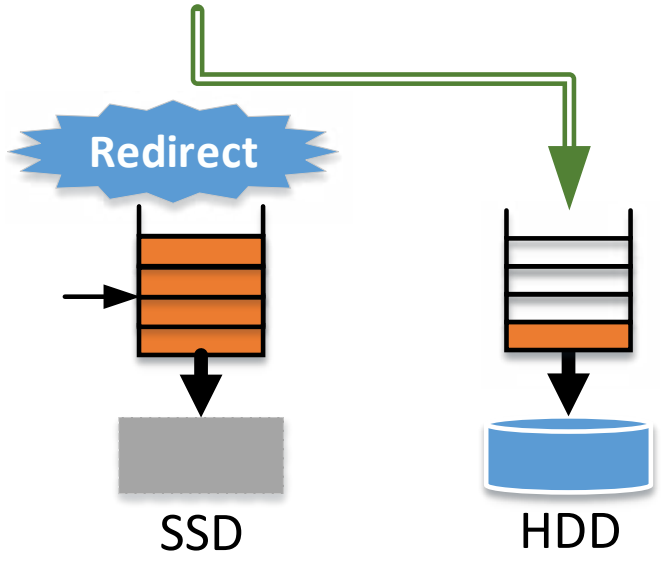


- No redirection to HDD when in **S** stage

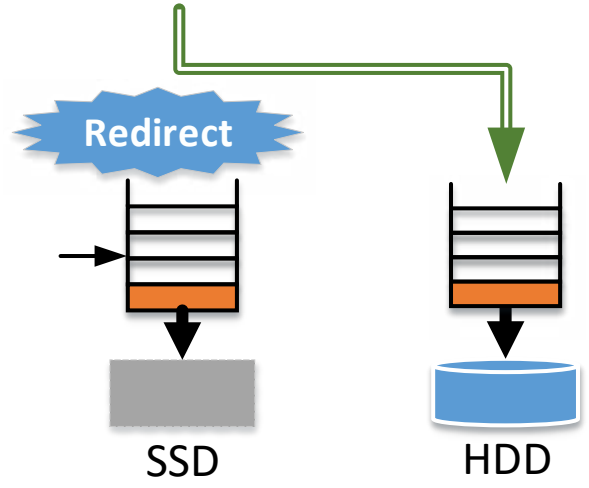
- Mixed IO scheduler runs atop of file system level

Strategies in MIOS

- Enable or disable BCW when the queue length is lower than the threshold
  - **MIOS_D** ✗
  - **MIOS_E** ✓



MIOS_D



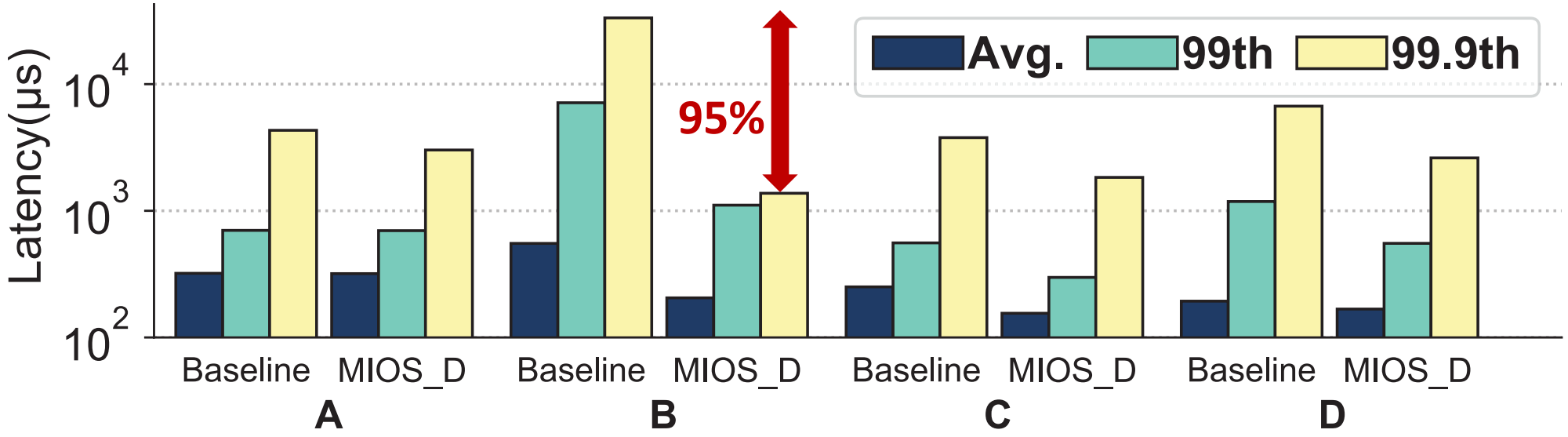
MIOS_E

Experimental Setup

- Datasets:
  - Real workload trace (A, B, C, D)
  - A: Lowest IO intensity
  - B: Most written data (61.2GB)
- Comparison:
  - **Baseline**: Pangu workload replay (writing all data into SSDs)
  - **MIOS_D**
  - **MIOS_E**

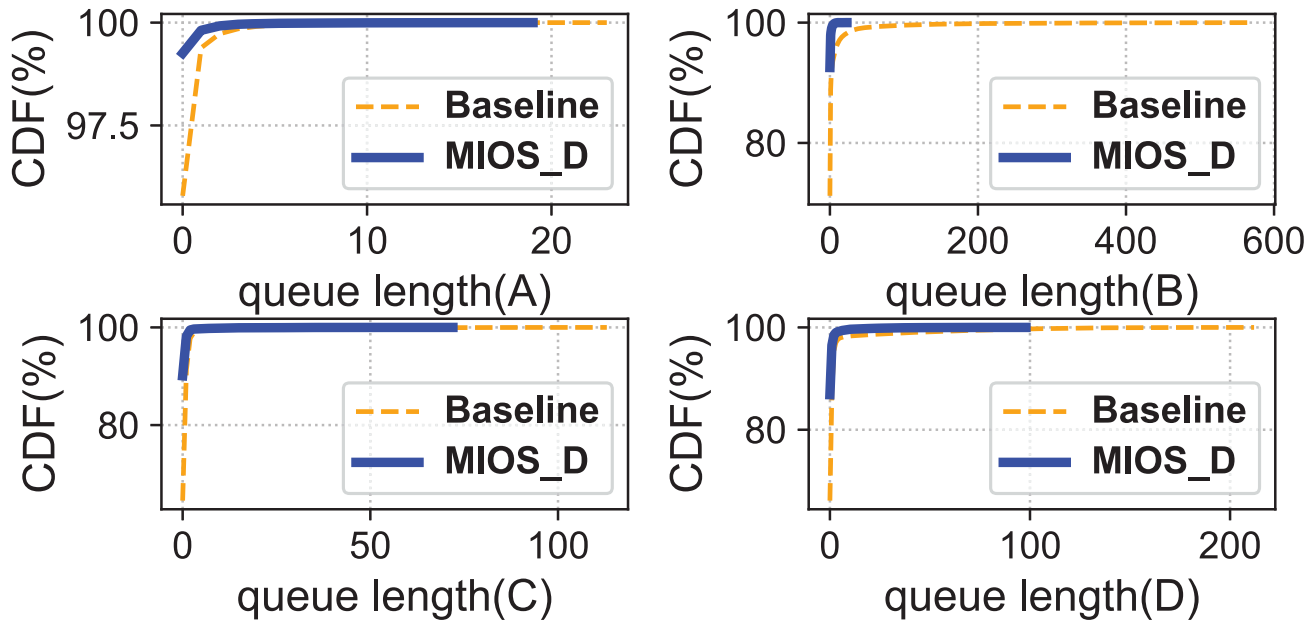
System	Linux version 4.15.0-52-generic
CPU	Intel Xeon E5-2696 v4 (2.20 GHz, 22 CPUs)
Memory	128 GB
HDDs	West Digital 10TB (default)
	West Digital 4TB
	Seagate 4TB
SSDs	Samsung 960EVO 256GB(NVMe, 2000MB/s)

Write Performance



- For average, 99th and 99.9th-pencentile latency
  - 65%, 85%, 95% latency reduction for workload B (*most intensive*)
  - 2%, 3.5%, 30% reduction for A (*less intensive and less queue blocking*)

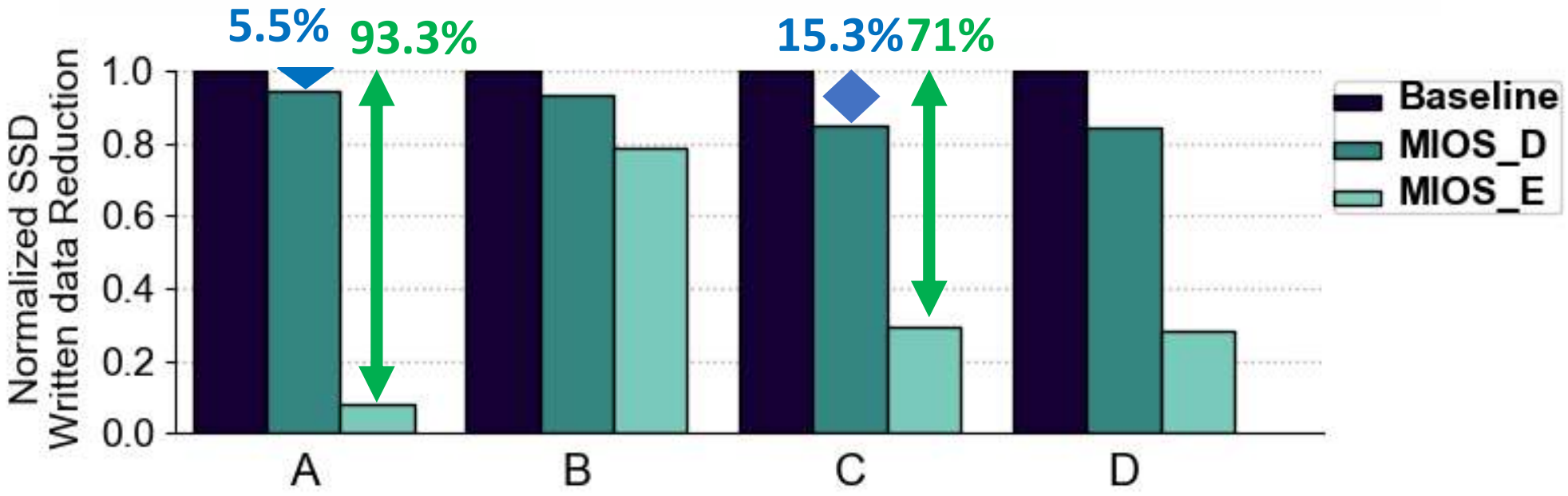
Queue Length Elimination



Note: The CDF of SSD queue length

- **MIOS_D** significantly shortens queue lengths compared to **Baseline**
  - B has 95% reduction since B is the most intensive workload
  - A has only 15% reduction since MIOS_D is only triggered when the queue length is high

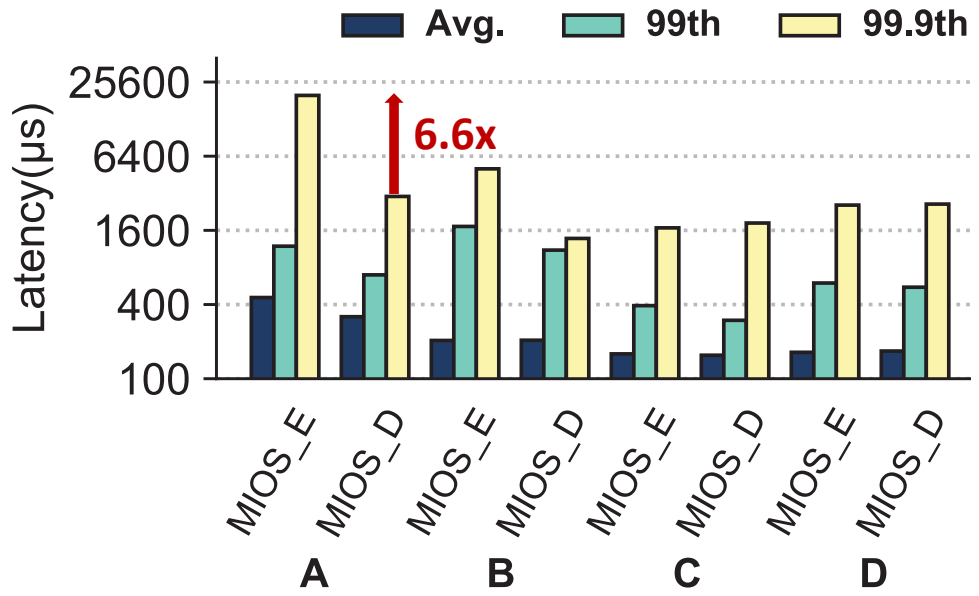
SSD Written Data Reduction



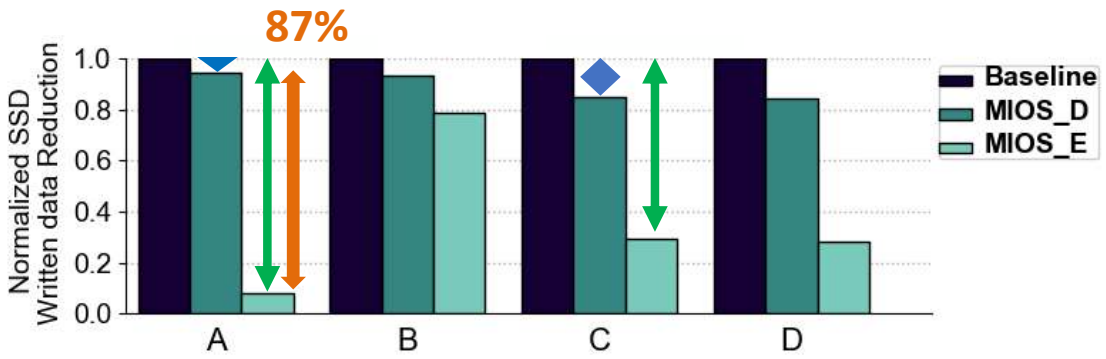
Note: Data reduction normalized to baseline

- SSD written data saving benefits from **MIOS_E** significantly
  - More changes to actively redirect requests to HDD with **MIOS_E** in A (>90%)

MIOS_D vs MIOS_E



Note: **MIOS_E** Latencies normalized to **MIOS_D**



Note: **MIOS_E** write reduction normalized to **MIOS_D**

- **MIOS_E** has *worse latency* compared with **MIOS_D**
- *87%* of SSD write data redirection compared with **MIOS_D**

HDD Utilization

Node Type	Duration(s)	Baseline	Net Util. MIOS_D	Net Util. MIOS_E	Gross Util. MIOS_E
A	2700	7.6%	7.9%	11.9%	27.9%
B	1800	9.8%	**18.2%**	**26.8%**	**56.9%**
C	1800	4.1%	10.7%	16.2%	35.8%
D	1560	4.8%	12.3%	17.3%	39.5%

Note: **Net utilization** is only about **redirected data write** to HDD

- **B** with **MIOS_E** has the highest net utilization improvement over Baseline, by *2.7x*, while *1.8x* under **MIOS_D**

Conclusion

- Motivation
  - SSD suffers from heavy write pressure
  - HDD is underutilized
- Buffer-Controlled Write (**BCW**)
  - Use **BCW** model to predict the future write stage
  - Actively prefer ***S*** stage
- Mixed IO scheduler (**MIOS**)
  - Redirect write requests by monitoring the queue length