Managing Array of SSDs When the Storage Device is No Longer the Performance Bottleneck

Byung S. Kim, Jaeho Kim, Sam H. Noh

UNIST
(Ulsan National Institute of Science & Technology)

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
CISSR Center for Innovative System Software Research
NECSST Next-generation Embedded / Computer System Software Technology

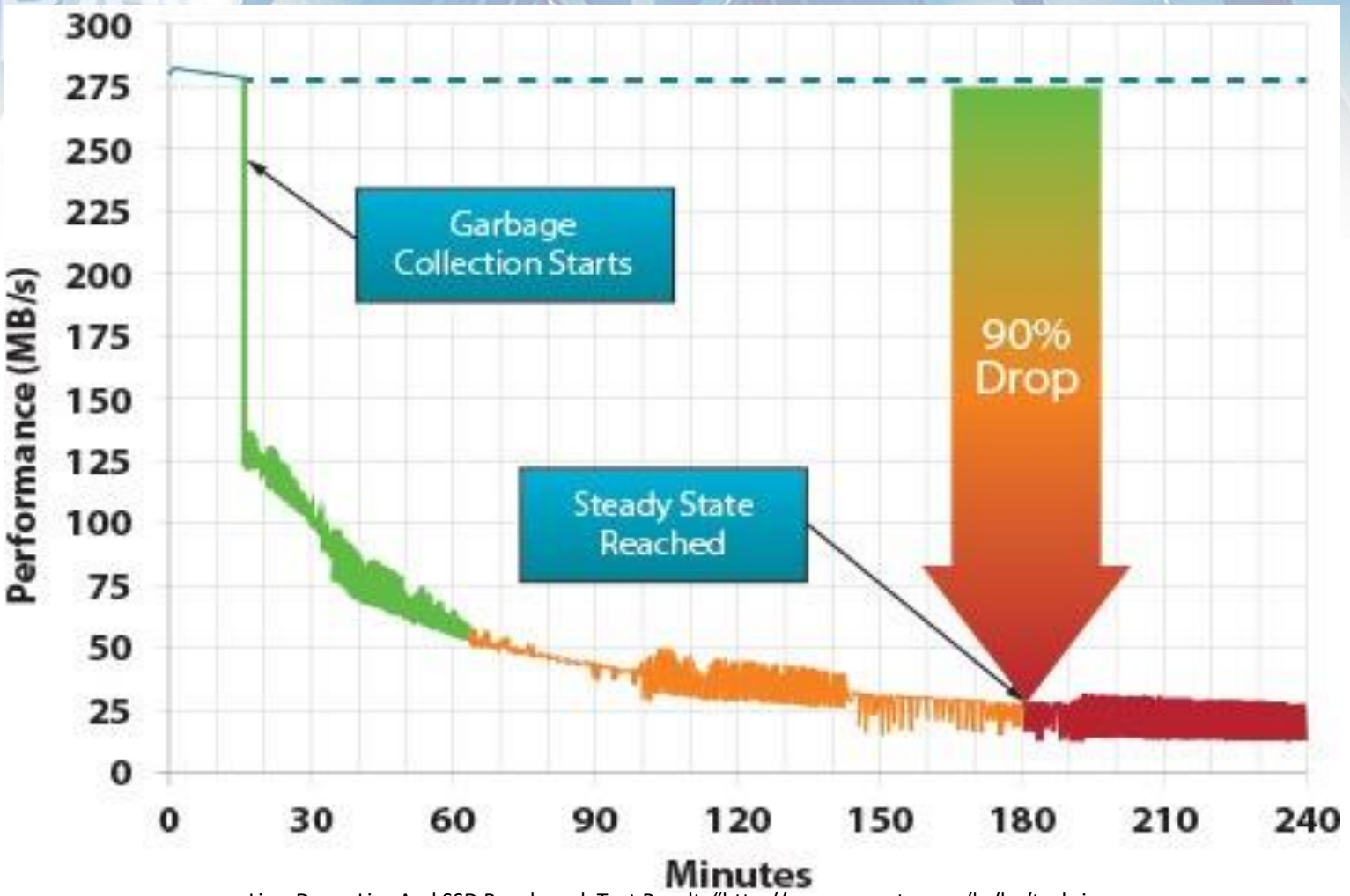
Outline

- **Motivation & Observation**
- **Our Idea**
  - Provide full network performance
  - Eliminate inconsistent performance
- **Evaluation**
  - Full network bandwidth
  - Consistent performance
- **Summary & Future work**

Outline

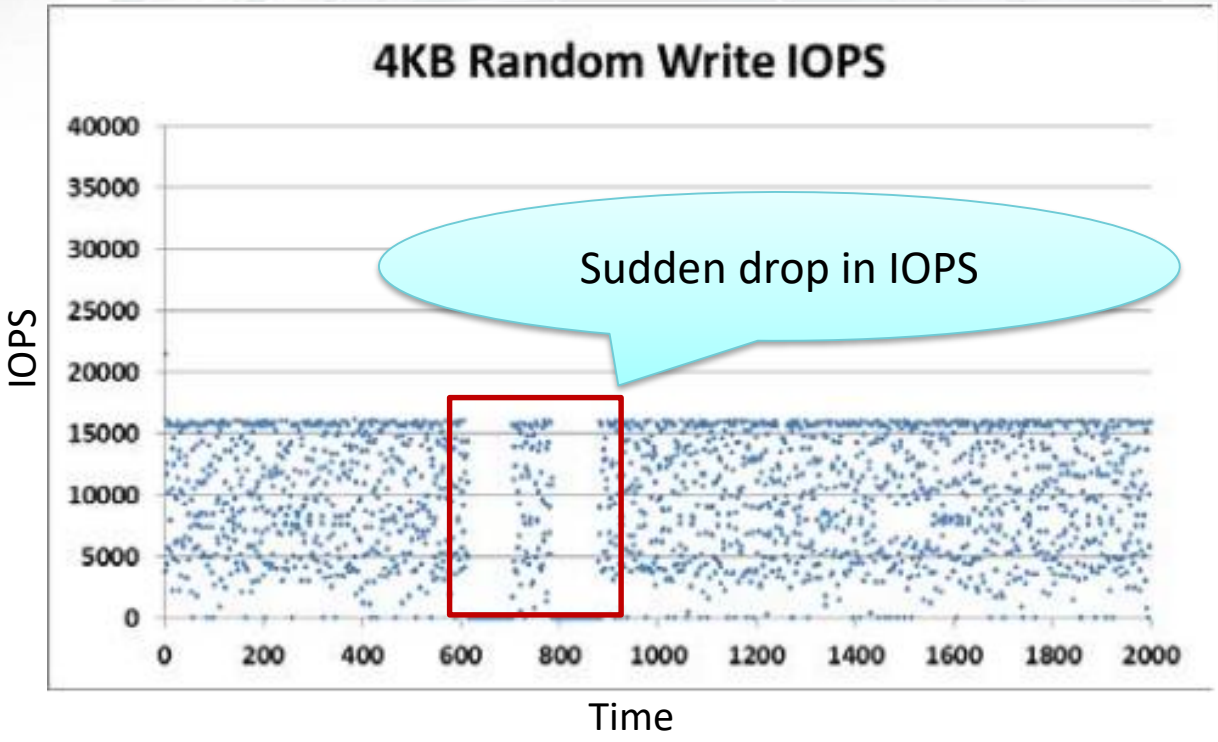
- **Motivation & Observation**
- **Our Idea**
  - Provide full network performance
  - Eliminate inconsistent performance
- **Evaluation**
  - Full network bandwidth
  - Consistent performance
- **Summary & Future work**

SSD Garbage Collection: Degraded Performance



Lies, Damn Lies And SSD Benchmark Test Result, "http://www.seagate.com/kr/ko/tech-insights/lies-damn-lies-and-ssd-benchmark-master-ti/."

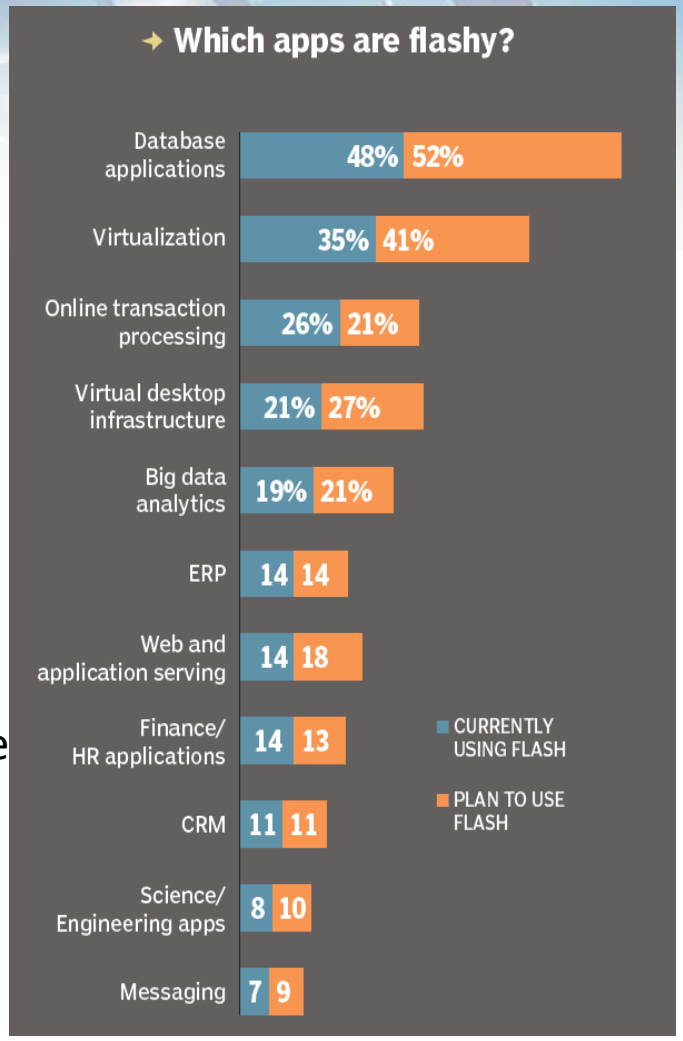
SSD Garbage Collection: Inconsistent Performance



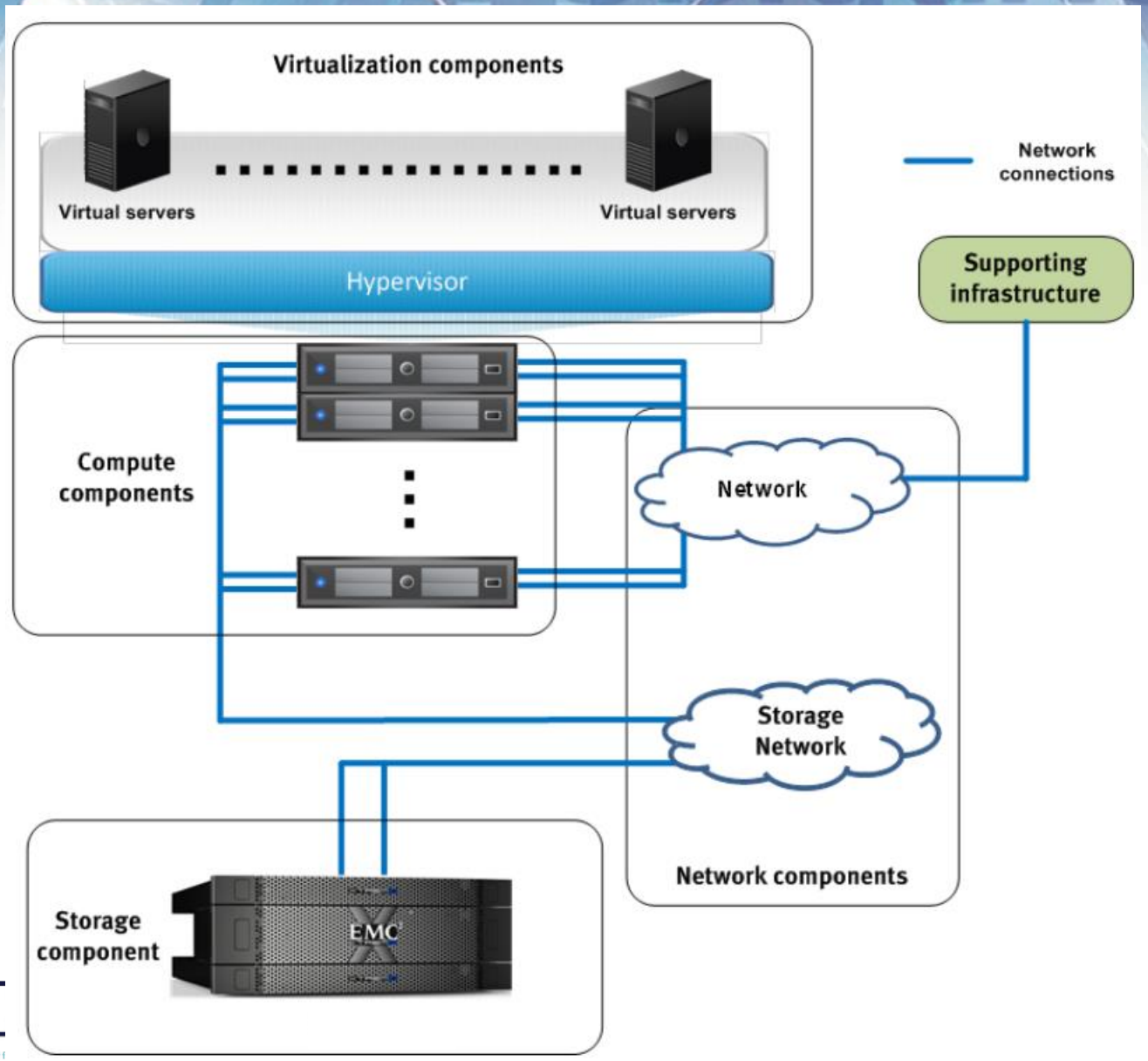
Intel Solid-State Drive DC S3700 Series – Quality of Service, "http://www.intel.com/content/dam/www/public/us/en/documents/technology-briefs/ssd-dc-s3700-quality-service-tech-brief.pdf."

Rise of All-Flash Array

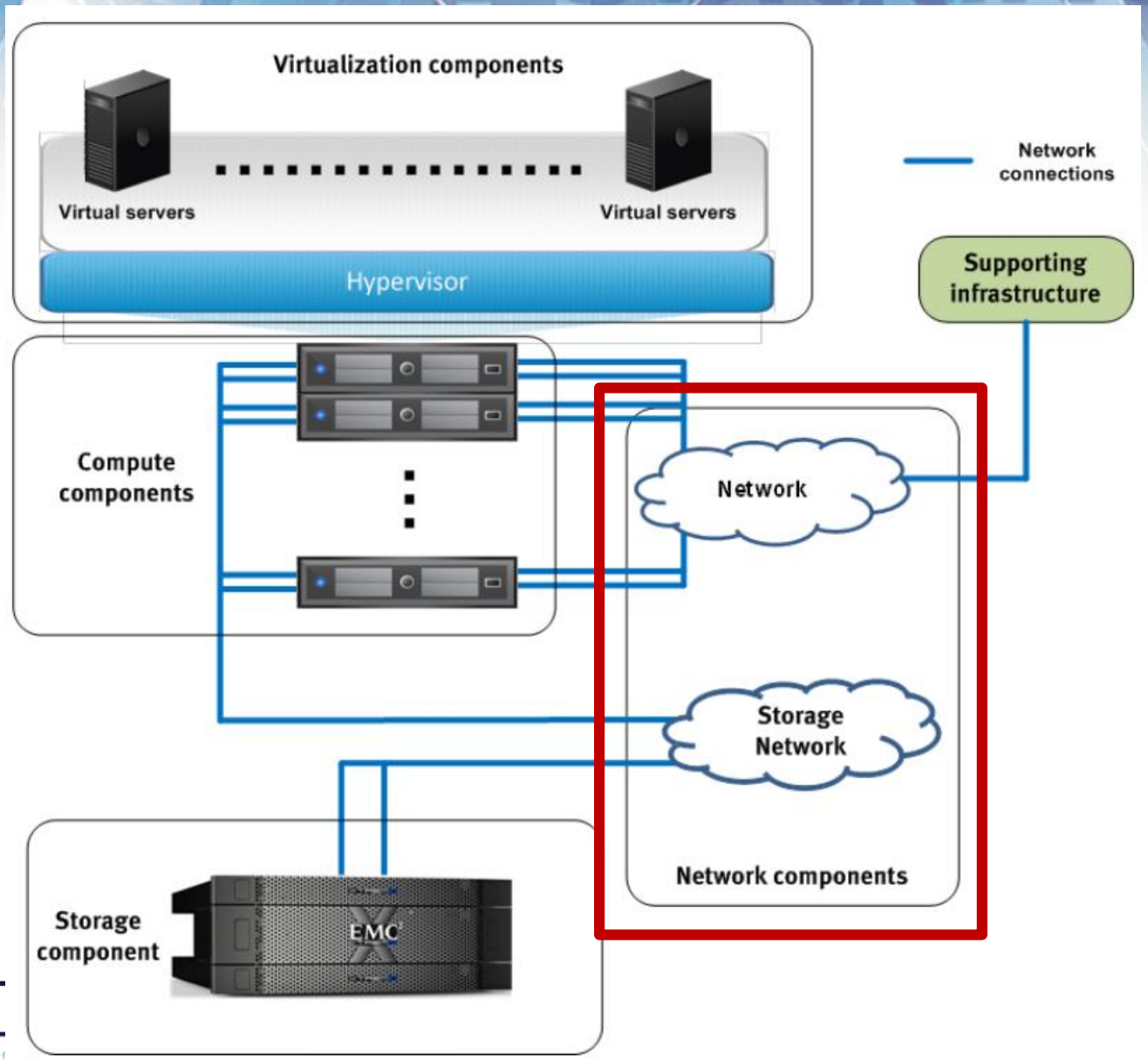
Improving performance for application request



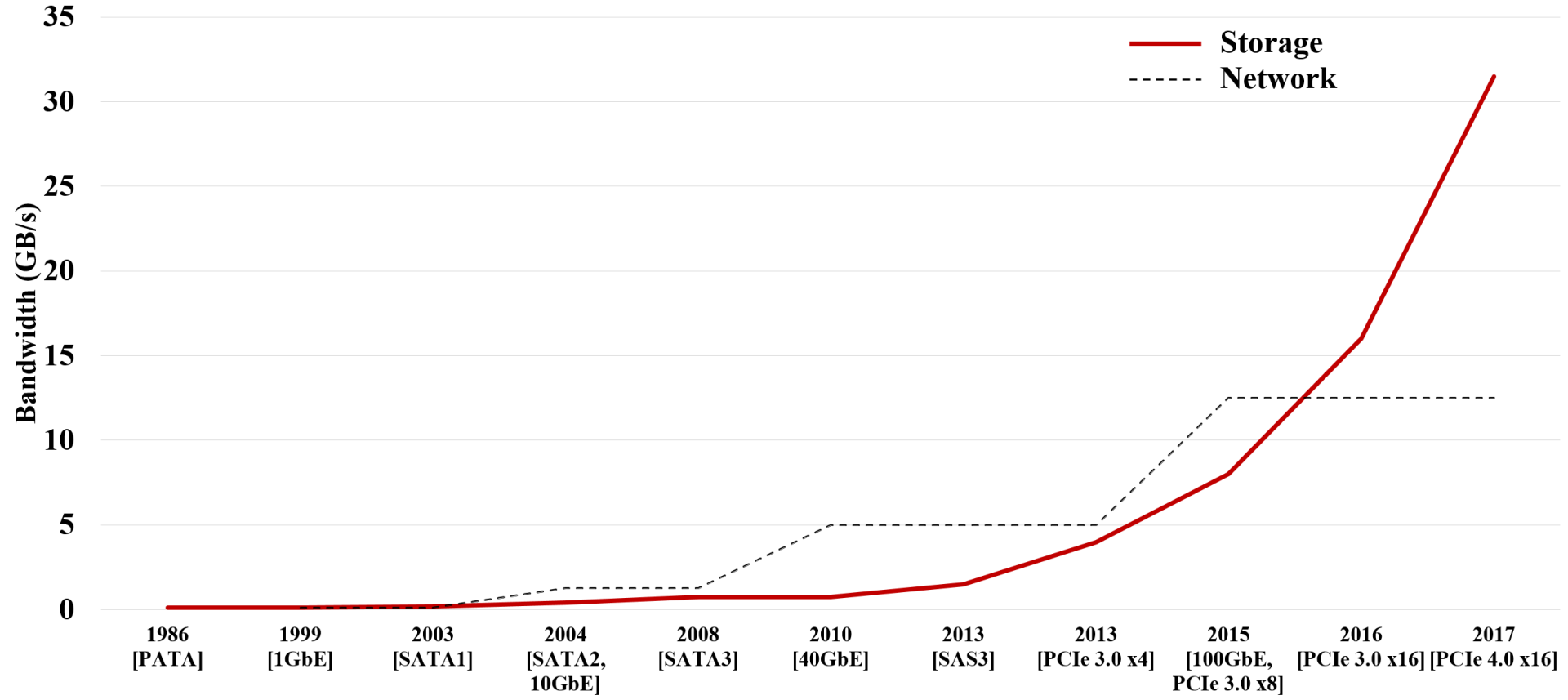
Architecture of All-Flash Array



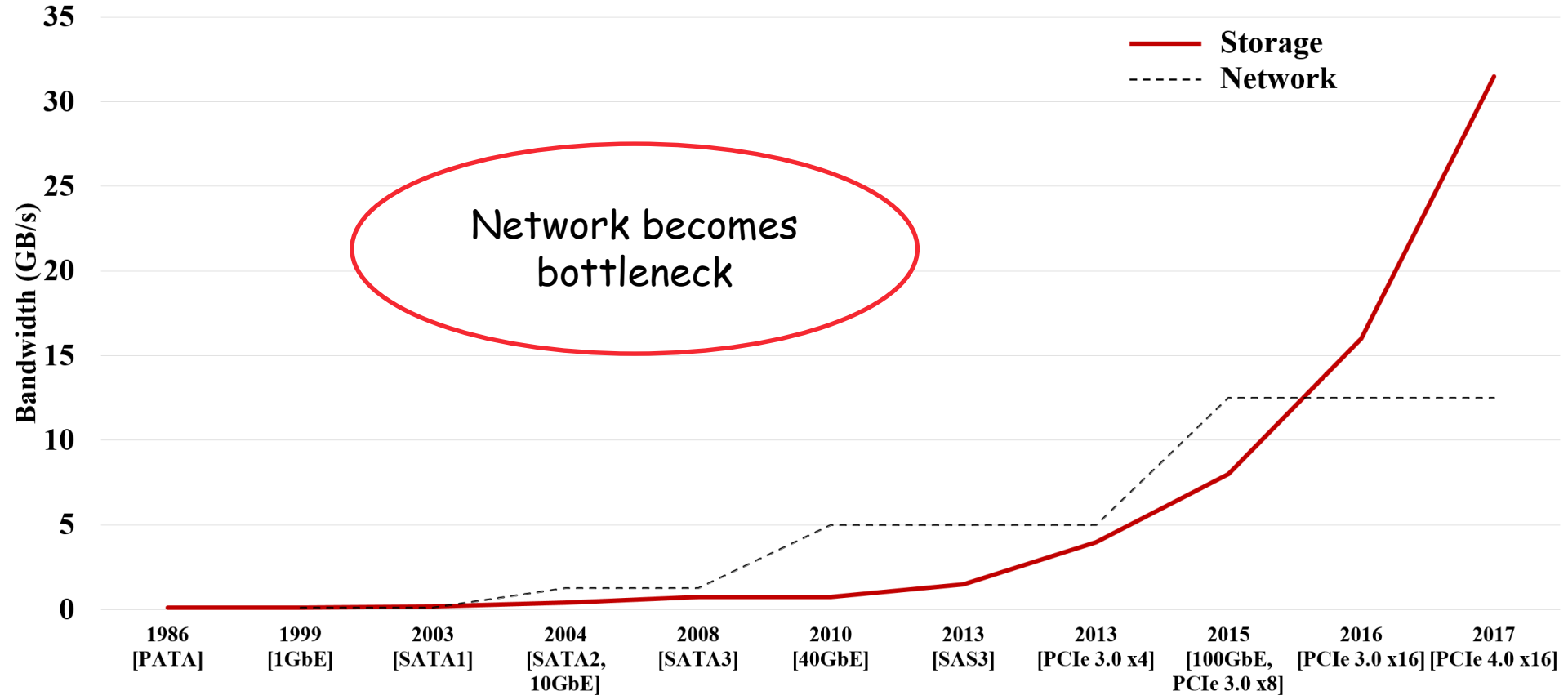
Architecture of All-Flash Array



Interface Bandwidth Growth Trend



Interface Bandwidth Growth Trend



Commercial SSD Trend

SSD product	Read performance	Write performance
Product A	6.8GB/s	4.8GB/s
Product B	3.5GB/s	2GB/s
Product C	2.7GB/s	1.5GB/s

10Gbps Ethernet: 1.25GB/s
Fibre channel: 1GB/s

Commercial All-Flash Array Trend

All-Flash Array products	# of SSDs	# of network ports
Product A	10	Up to 4
Product B	150	Up to 48
Product C	68	Up to 2
Product D	96	Up to 4

- Up to 34 SSDs per network port

Commercial All-Flash Array Trend

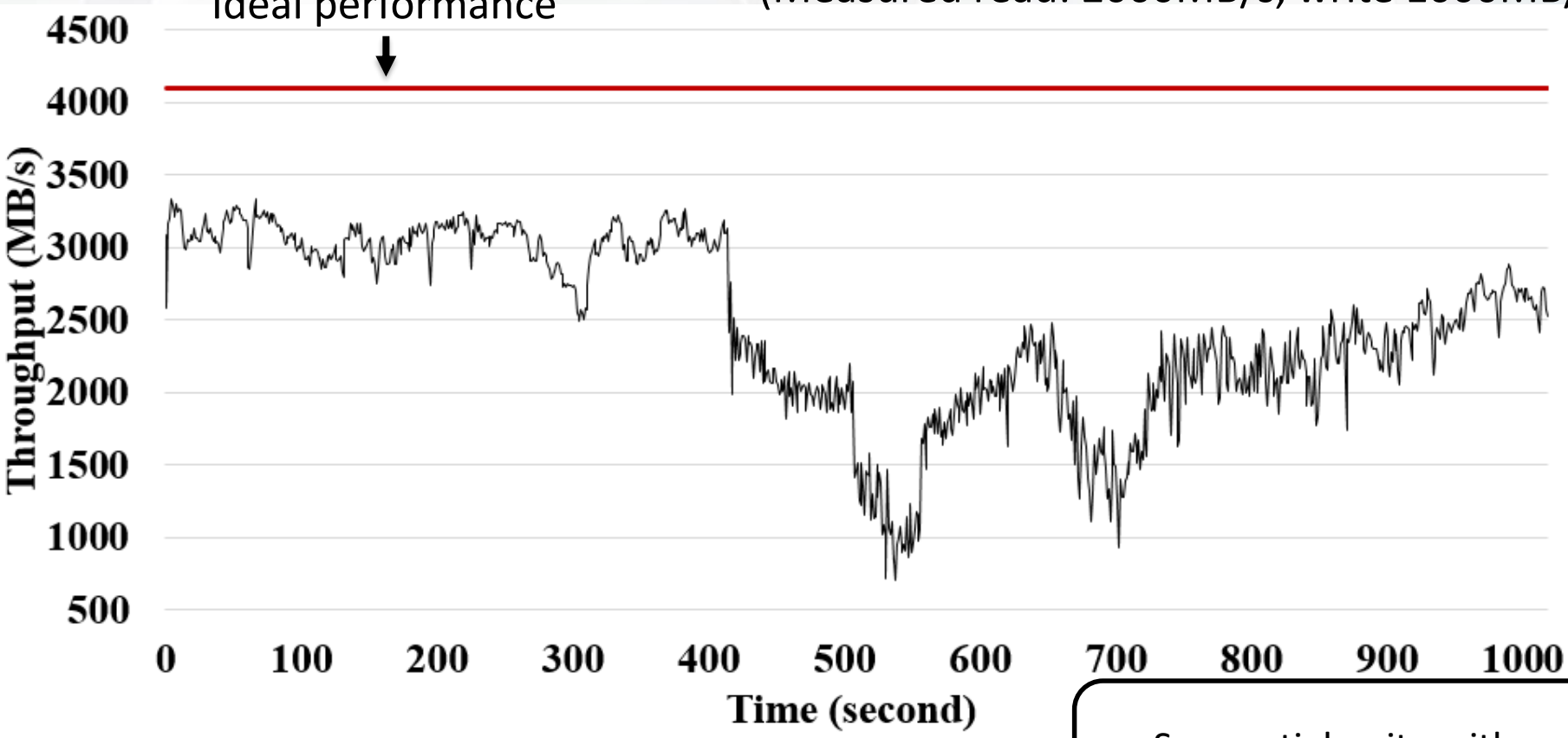
All-Flash Array products	# of SSDs	# of network ports
Product A	10	Up to 4
Product B	150	Up to 48
Product C	68	Up to 2
Product D	96	Up to 4

- Up to 34 SSDs per network port

Does these many SSDs really help?

Performance of RAID 0 with 4 SSDs

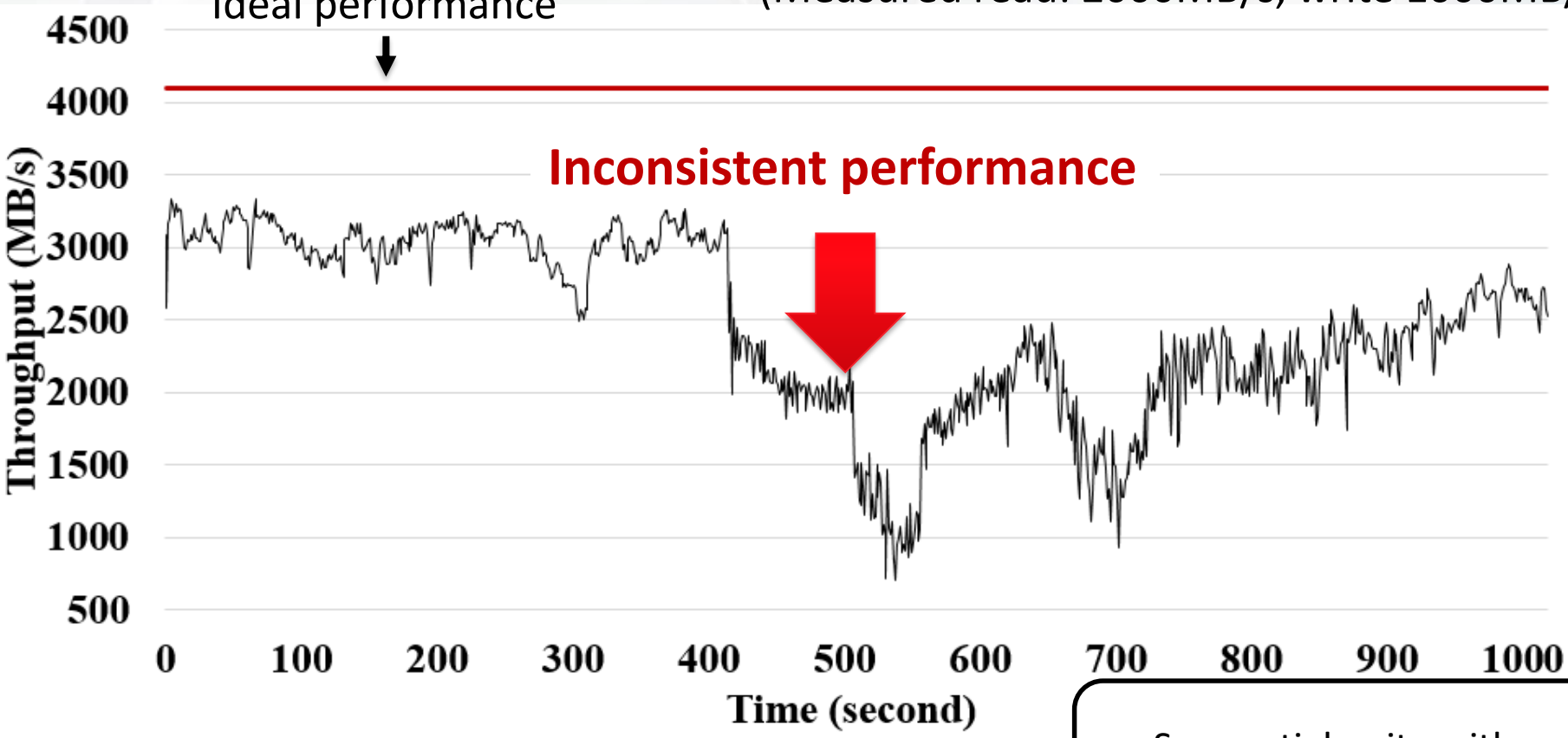
RAID 0 config. by 4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



Sequential write with
128KB I/O size

RAID 0: Inconsistent Performance

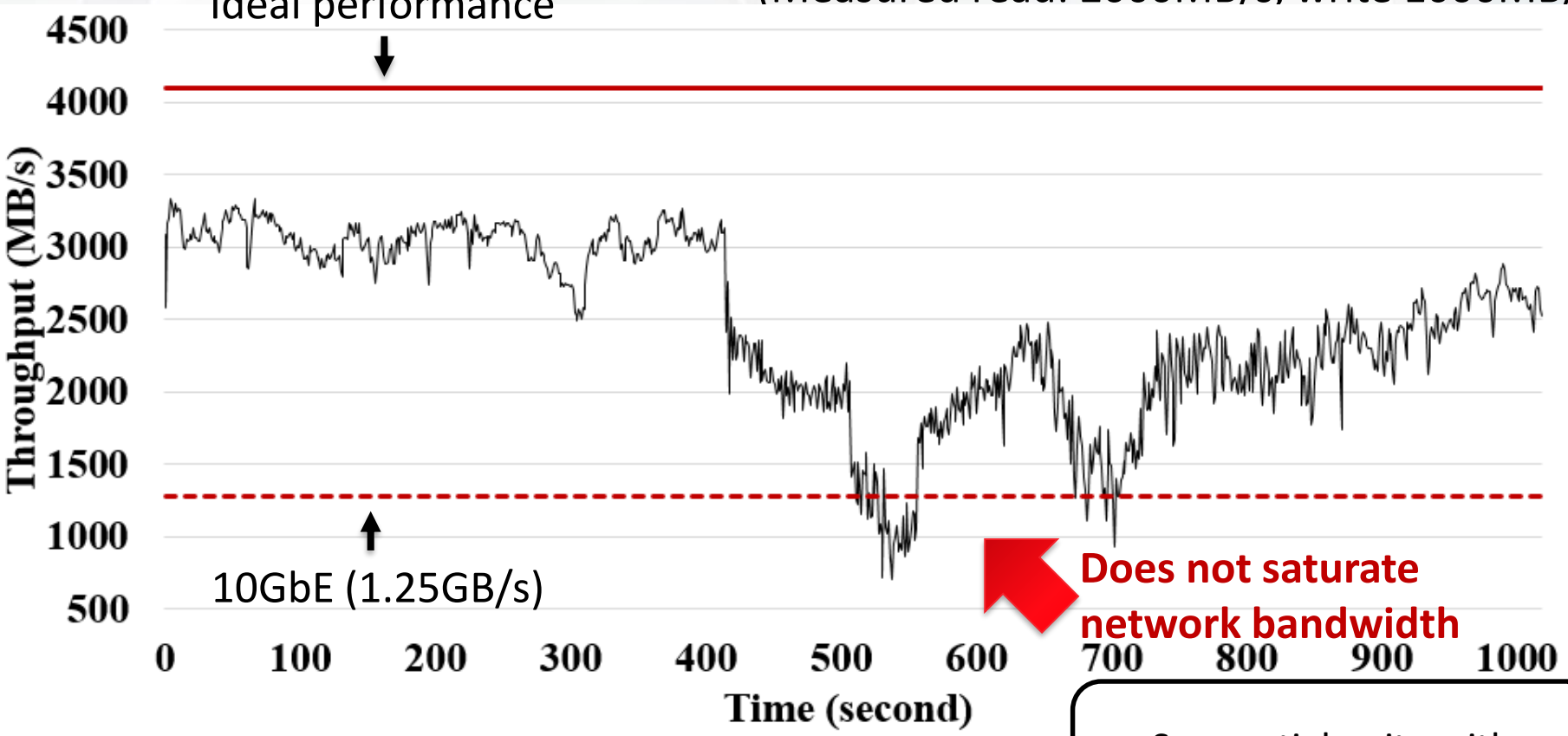
RAID 0 config. by 4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



Sequential write with
128KB I/O size

Performance of RAID 0 with 4 SSDs

RAID 0 config. by 4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



Sequential write with 128KB I/O size

Limited by Network Bandwidth

- **Performance is limited by network bandwidth**

Another motivation:
provide full network performance
under network connection

Limited by Network Bandwidth

- **Performance is limited by network bandwidth**

Another motivation:
provide full network performance
under network connection

Sustained full network performance

Problem and Our Goal

- **All-Flash Array suffers from inconsistent, limited performance**

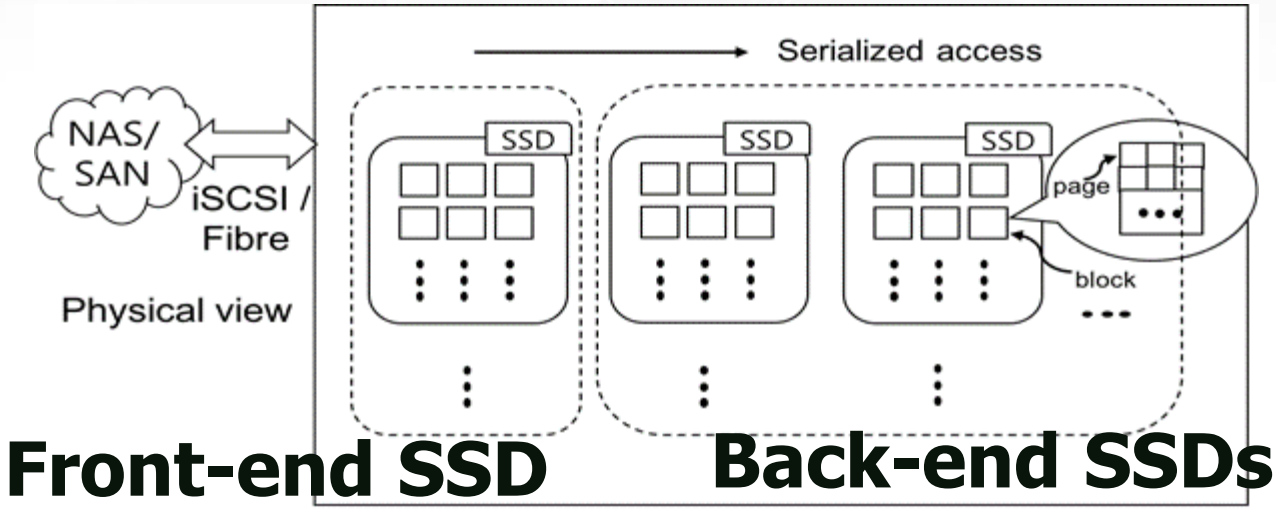
- **We want consistent, full network performance!**

Outline

- Motivation & Observation
- **Our Idea**
  - Provide full network performance
  - Eliminate inconsistent performance
- Evaluation
  - Full network bandwidth
  - Consistent performance
- Summary & Future work

Our Solution

- Serial configuration

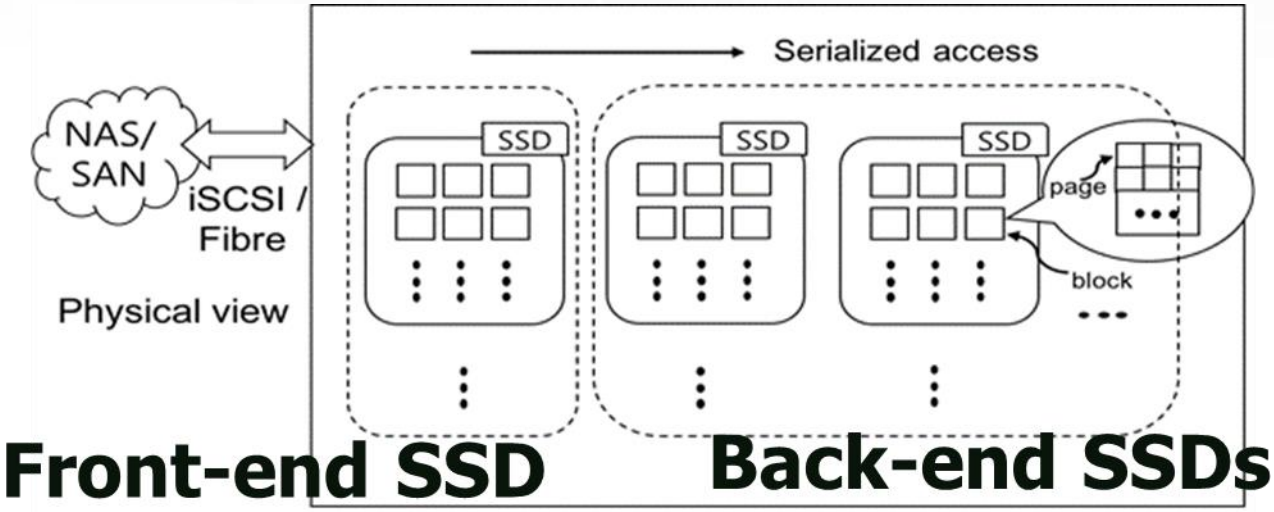


Serial Configuration Design

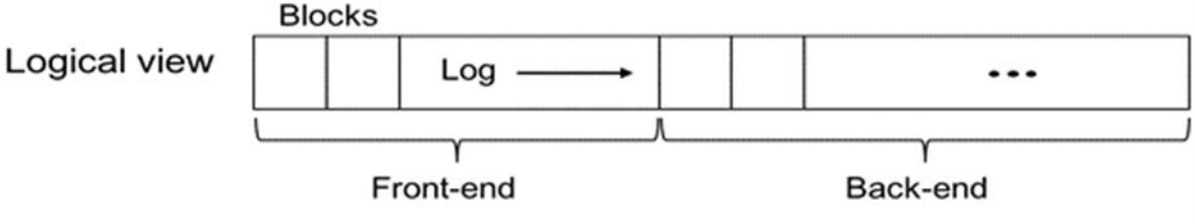
- **Provide full network performance**
  - Absorb all writes with Front-end SSD
  - Log-structured manner
- **Provide consistent performance**
  - Eliminate GC within Front-end SSD
  - Propose xGC

Full Network Performance

- **Absorb all writes with Front-end SSD**
  - Network bandwidth < Front-end SSD bandwidth



- **Log-structured manner**
  - Sequential, append only writes



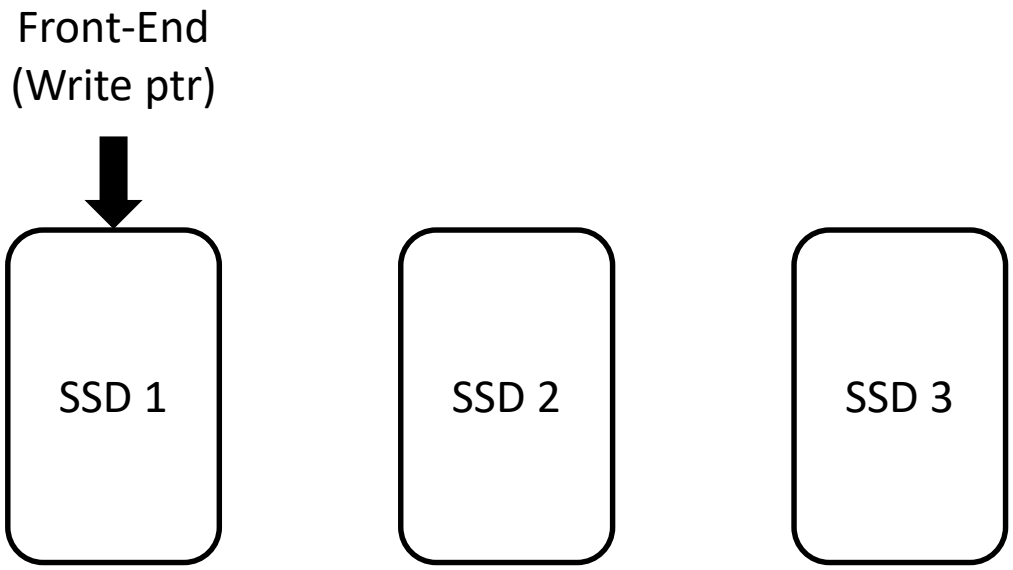
Issues in Providing High Performance

- **Front-end SSD will eventually fill up**
  - Garbage Collection?
- **Managing Front-end SSDs**
  - Selecting next Front-end
  - Making space available

Handling Garbage Collection

- **Eventually, the Front-end SSD becomes full and garbage collection is needed**

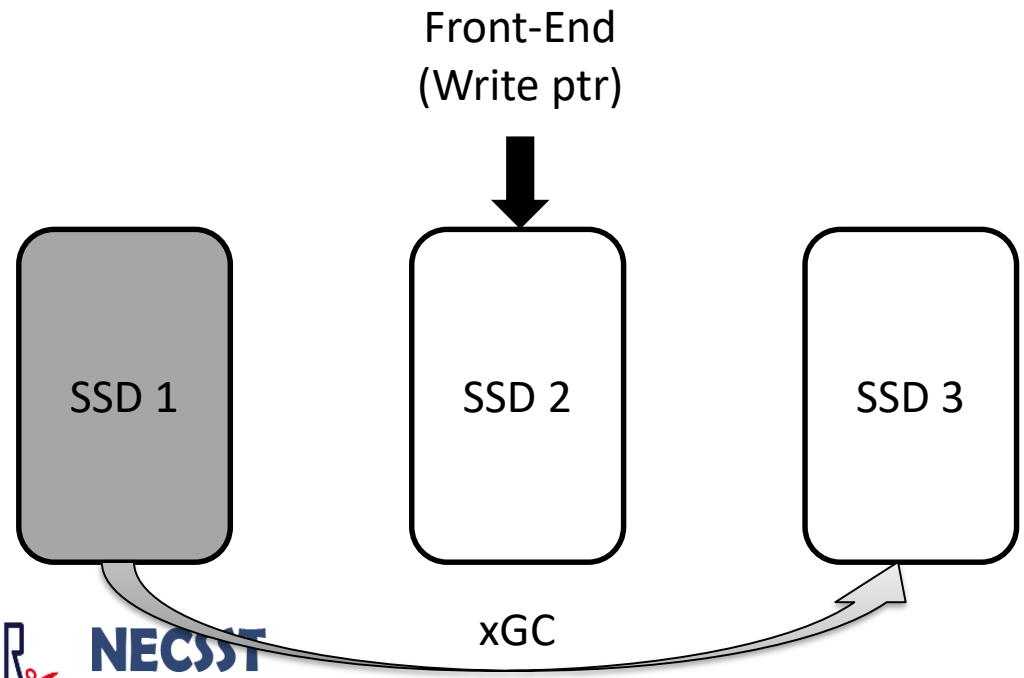
- **External GC (xGC)**
  - **Garbage collection never occurs at Front-end SSD**



Handling Garbage Collection

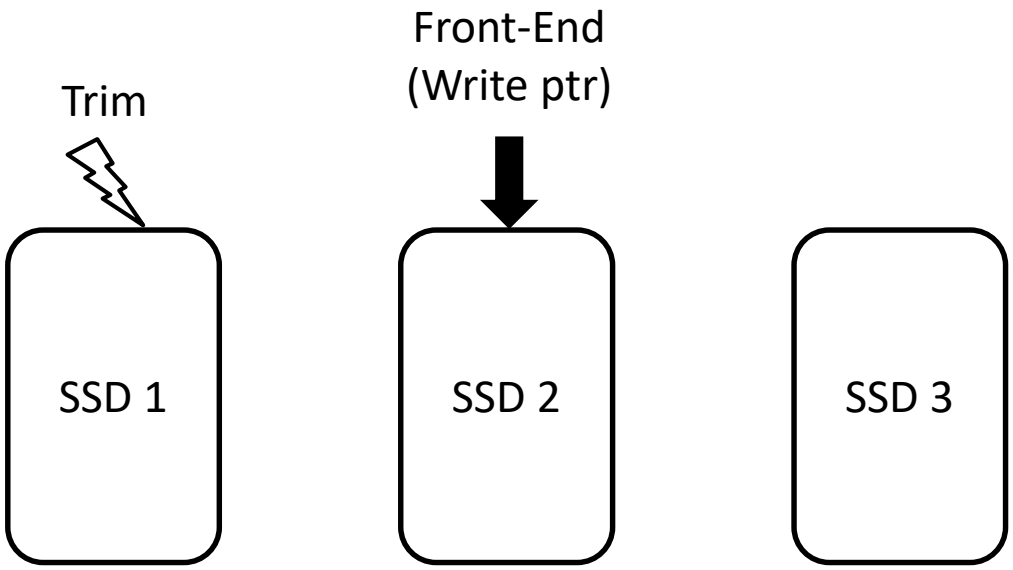
- **When the Front-end SSD fills up**
  - New Front-end is selected
  - Old Front-end SSD becomes a Back-end SSD

- **External GC (xGC) is performed between Back-end SSDs**



Handling Garbage Collection

- **When all valid data is moved**
  - Old front-end is cleaned by issuing TRIM command



Effect of xGC

- **Front-end performance is not affected by GC**

- **Front-end always provides consistent performance**

Outline

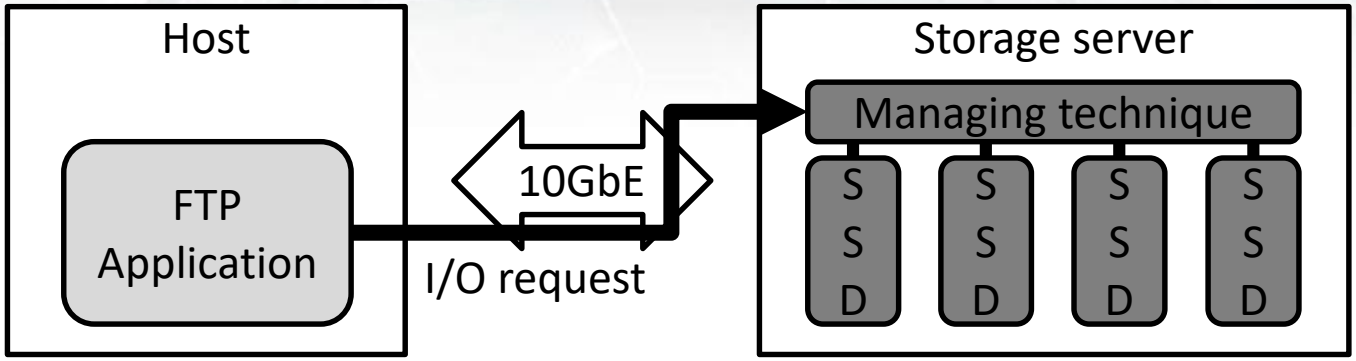
- Motivation & Observation
- Our Idea
  - Provide full network performance
  - Eliminate inconsistent performance
- **Evaluation**
  - Full network bandwidth
  - Consistent performance
- Summary & Future work

Evaluation Settings

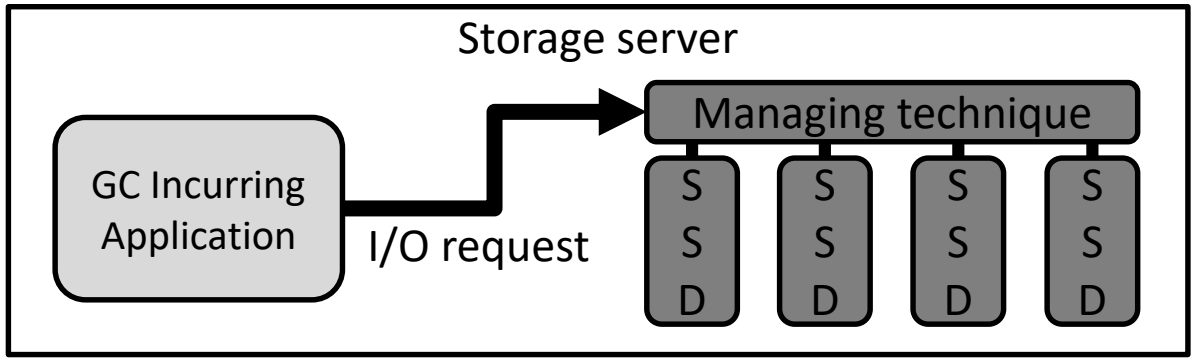
Description		
	Storage Server	Host Server
CPU	Intel Xeon E5-2609	Intel i5-6600k
RAM	64GB DRAM	16GB DRAM
Ethernet	10Gbps	
OS	Linux kernel 4.4.43	Linux kernel 4.3.3
SSD	Intel 750 400GB NVMe SSD × 4 (spec. read: 2400MB/s, write: 1200MB/s) (Measured read: 2000MB/s, write 1000MB/s)	

Evaluation Settings

- **Observe effect of network connection**



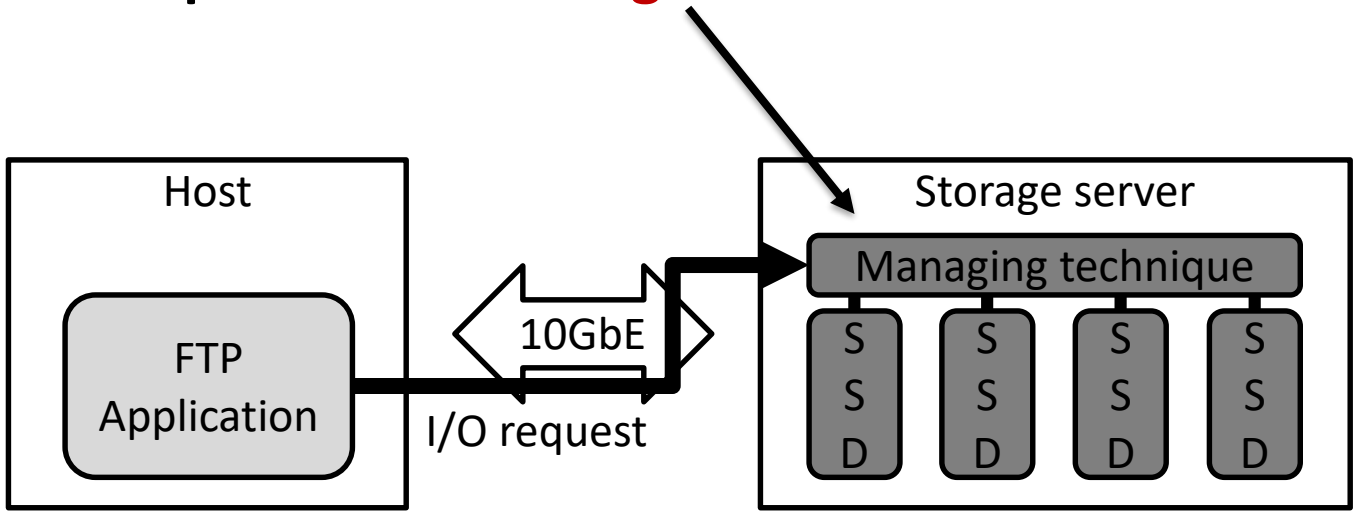
- **Observe effect of serial configuration**



Effect of Network Bandwidth

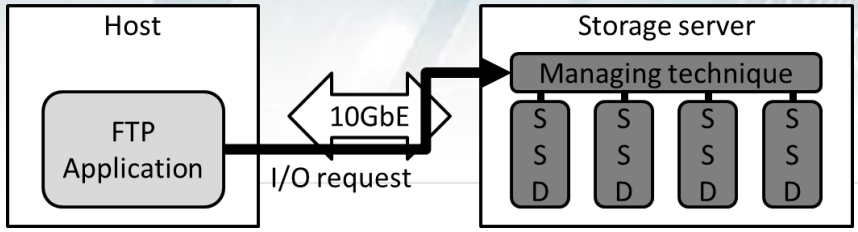
- Transfer 10 files (respectively, 10GB) with 10 threads to storage server via FTP protocol

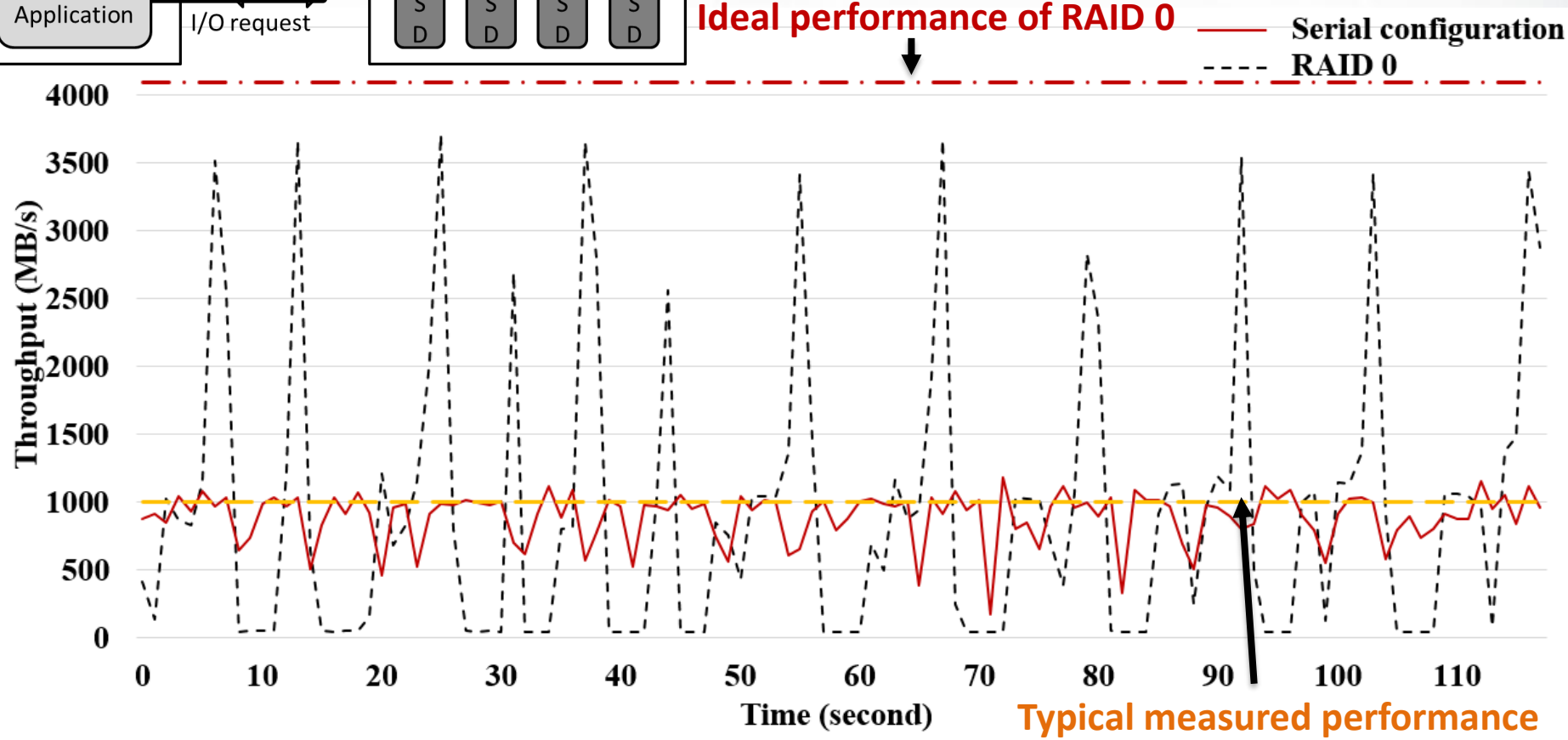
- Measurement point is the storage server



Effect of Network Bandwidth

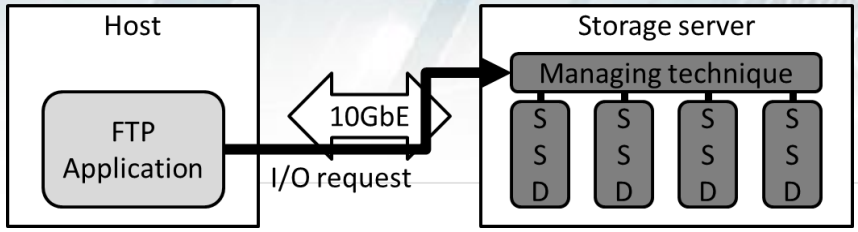
4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)

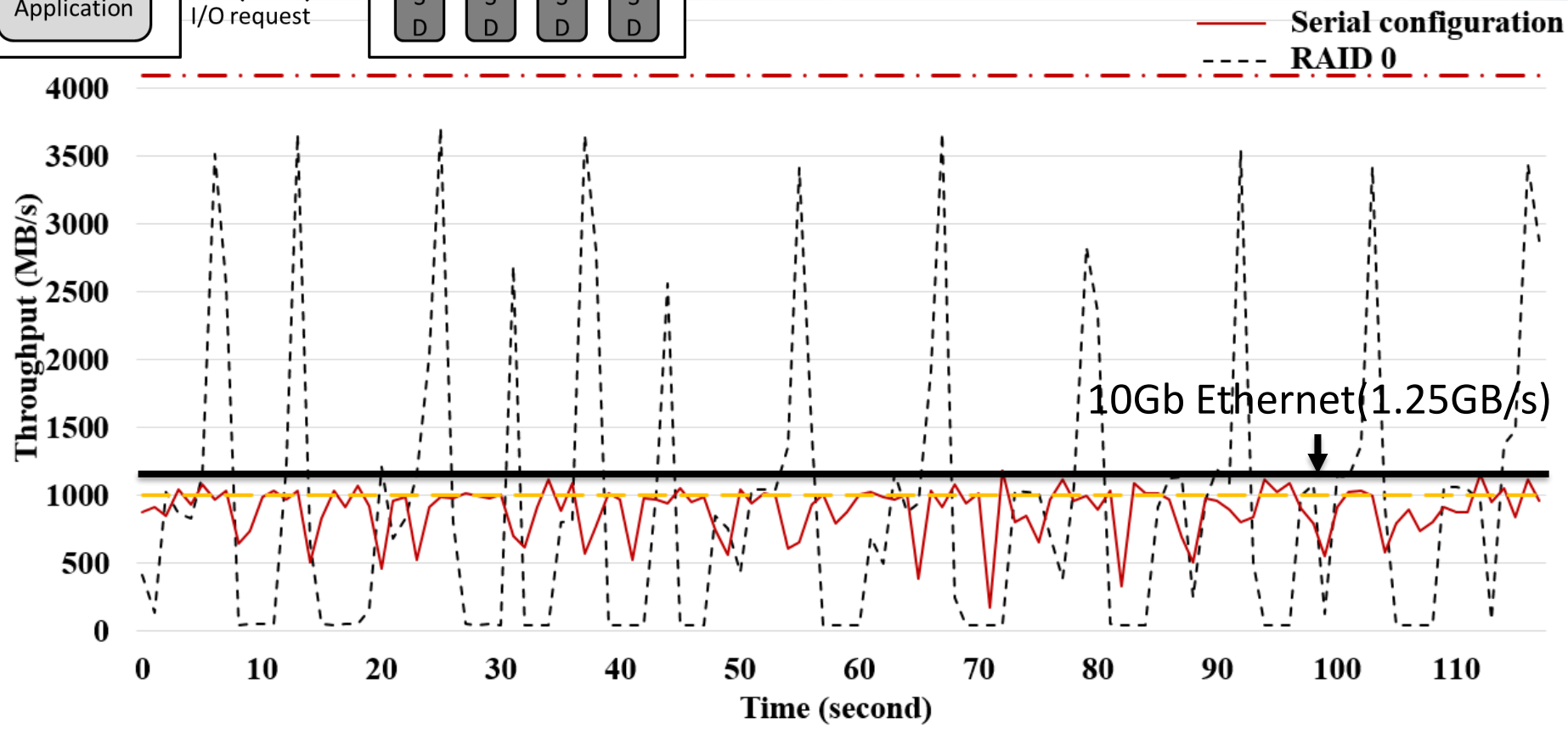




Effect of Network Bandwidth

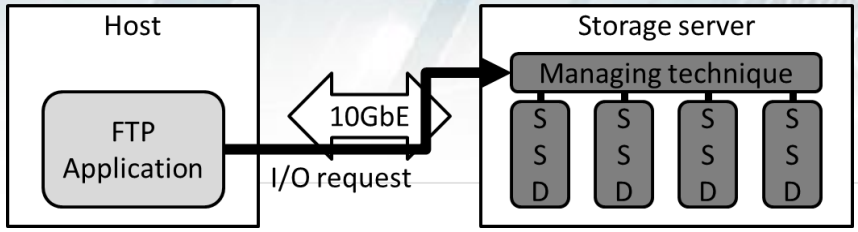
4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)

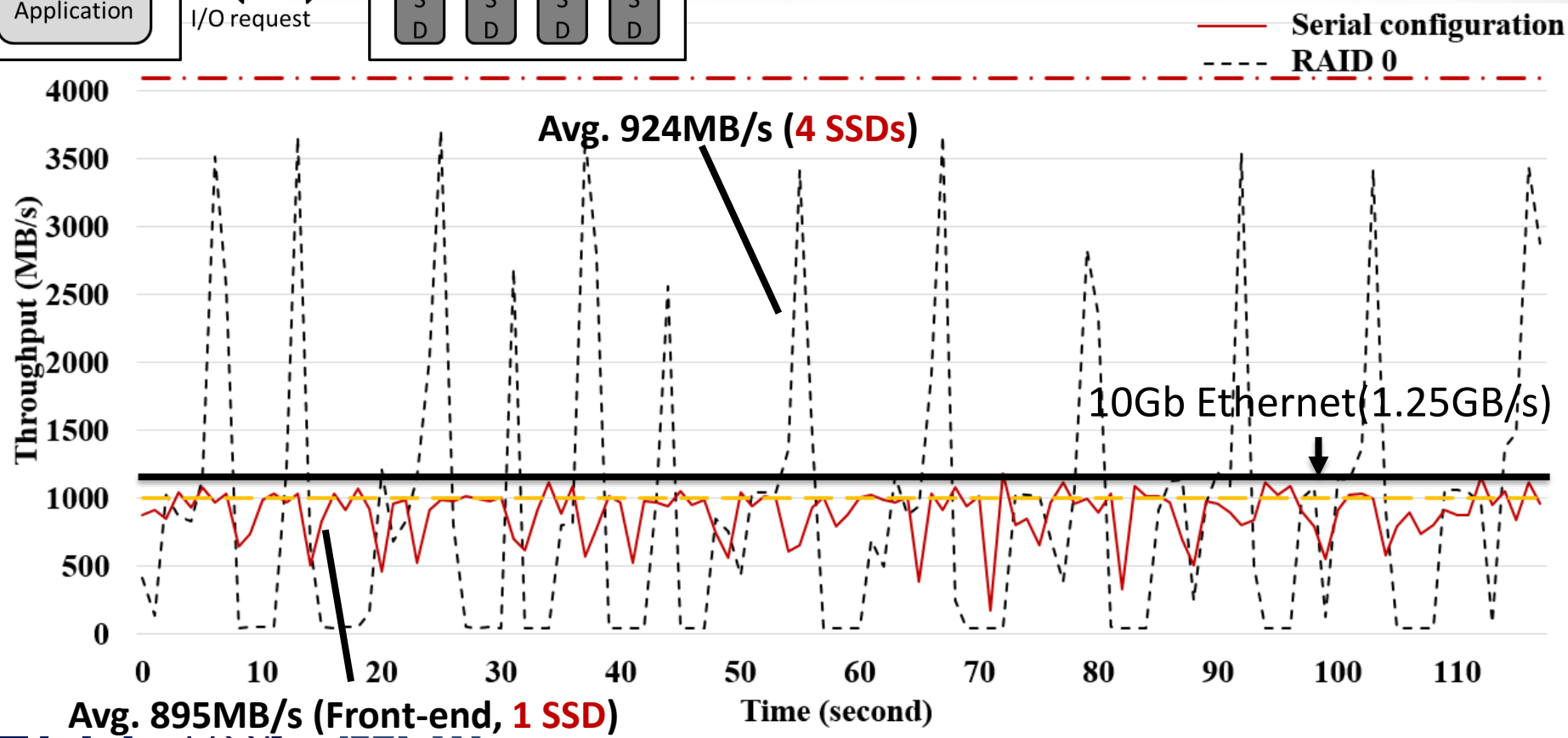




Effect of Network Bandwidth

4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



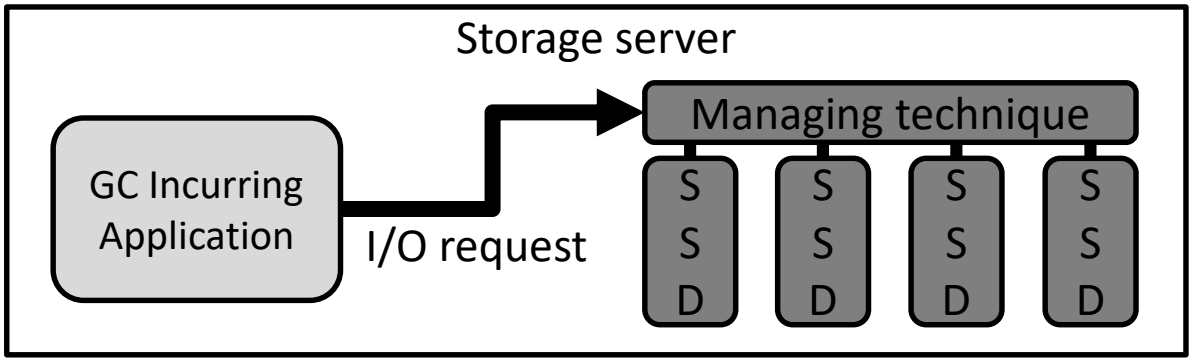


Conclusion of First Evaluation

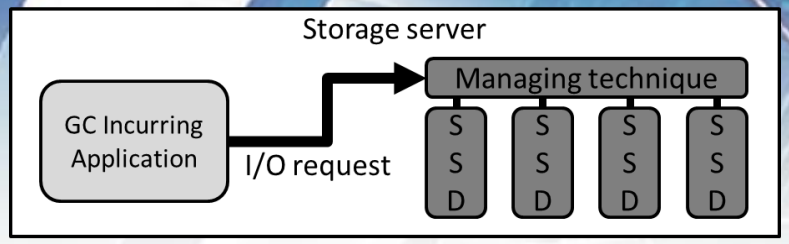
- **Performance is determined by network**
  independent of performance of storage
  - Performance of our approach is similar to that of RAID 0 with 4 SSDs

Observe Effect of Serial Configuration

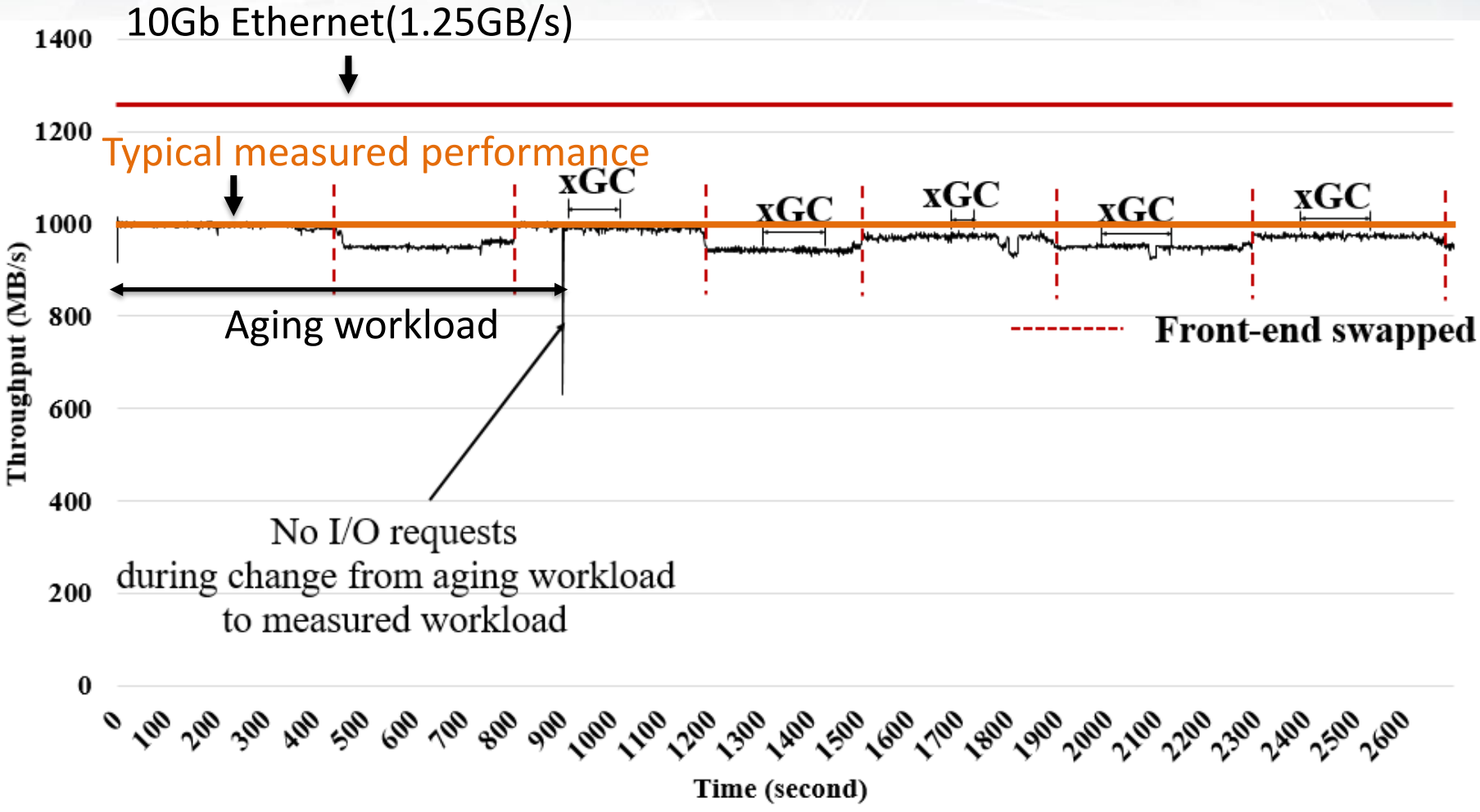
- **Performance with network effect removed**
  - Verify performance of serial configuration
- **Synthetic workload generated by FIO benchmark tool**
  - Perform I/O for 30 minutes after aging
    - 1200GB footprint
    - 256KB random writes
  - Measure performance of random write workload with 64KB I/O size



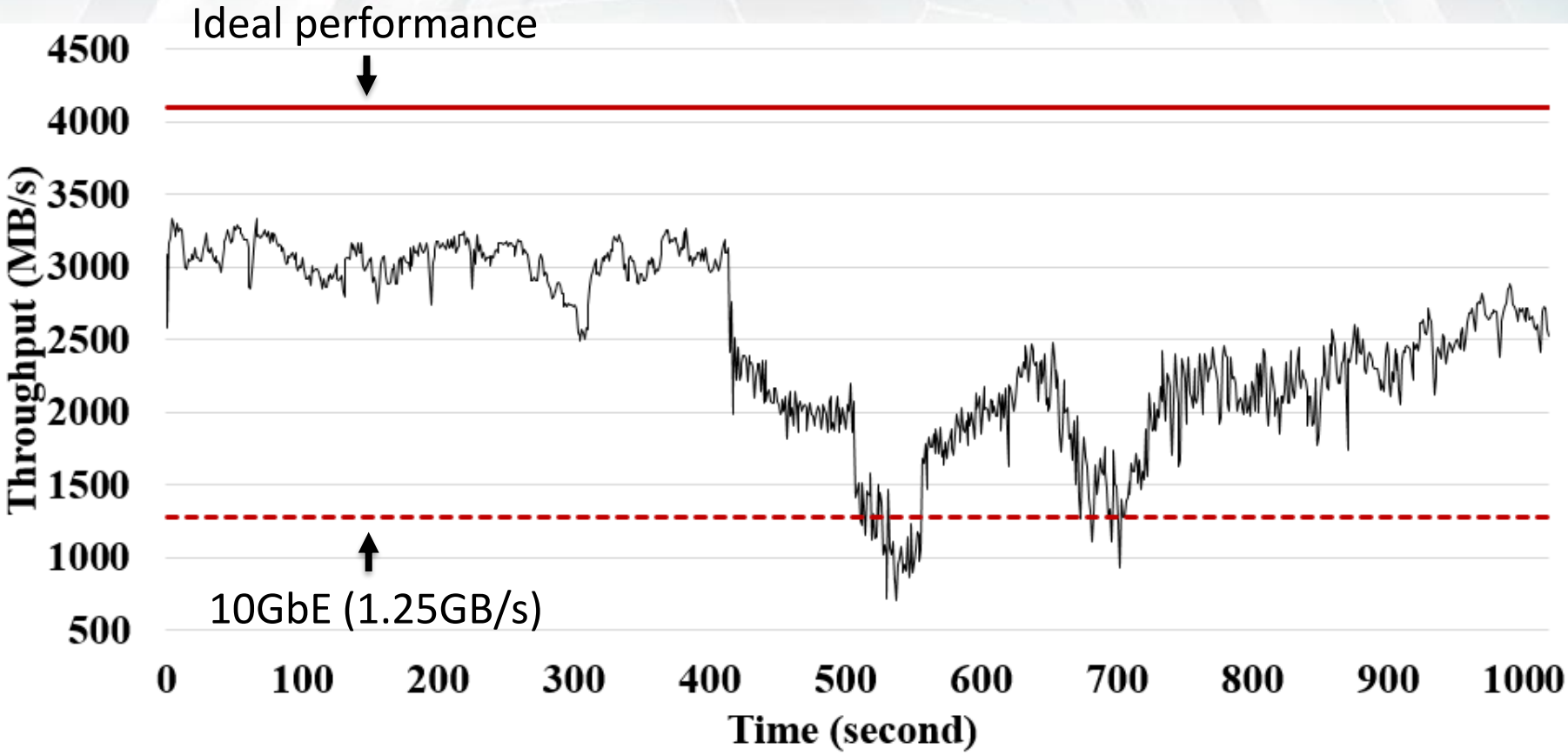
Verifying Consistent Performance



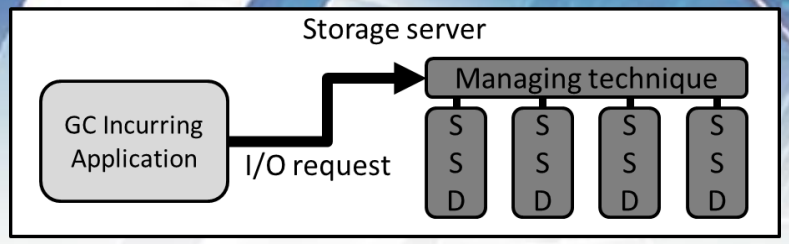
4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



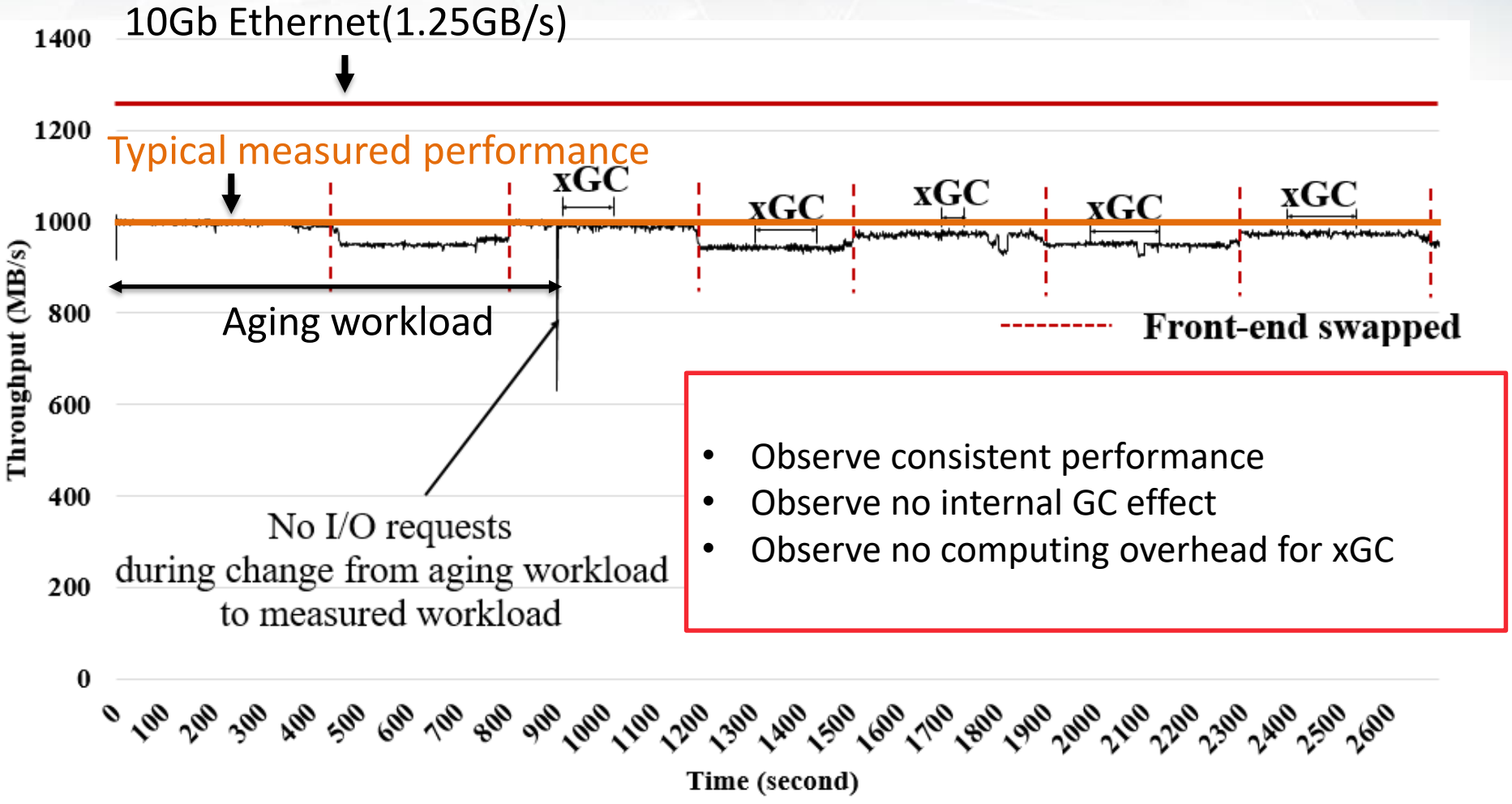
In Contrast to RAID 0 Configuration



Verifying Consistent Performance



4 NVMe SSDs (spec. read: 2400MB/s, write: 1200MB/s)
(Measured read: 2000MB/s, write 1000MB/s)



- Observe consistent performance
- Observe no internal GC effect
- Observe no computing overhead for xGC

Outline

- Motivation & Observation
- Our Idea
  - Provide full network performance
  - Eliminate inconsistent performance
- Evaluation
  - Full network bandwidth
  - Consistent performance
- **Summary & Future work**

Summary

- All-Flash Array is faster but limited by network bandwidth
- All-Flash Array suffers from inconsistent performance due to garbage collection
- Proposed technique that satisfies both full network performance and consistent performance

Future Work

- **Fault-tolerance**
- **Parallelization of serial configuration**
  - One serial configuration per network port
- **Scalability**
  - More than 4 SSDs
  - Heterogeneous SSDs
- **Metadata management**
- **Latency performance**
- **Effect on read performance**

Thank you