Peripherals No More

Sam H. Noh
노삼혁/盧三赫

UNIST
(Ulsan National Institute of Science & Technology)

UNIST

©GraphicMaps.com
Arctic Ocean
ARCTIC CIRCLE
ASIA
South Korea
Pacific Ocean
Atlantic Ocean
Indian Ocean
OCEANIA
AUSTRALIA
1500 mi
1500 km
NORTH KOREA
Demarcation Line
SEOUL
Inch'ŏn
Suwŏn
Wŏnju
Kangnŭng
Tonghae
Yellow Sea
Taejŏn
Kunsan
P'ohang
Taegu
Chŏnju
Ulsan
Masan
Chinhae
Kwangju
Mokp'o
Pusan
Yŏsu
Korea Strait

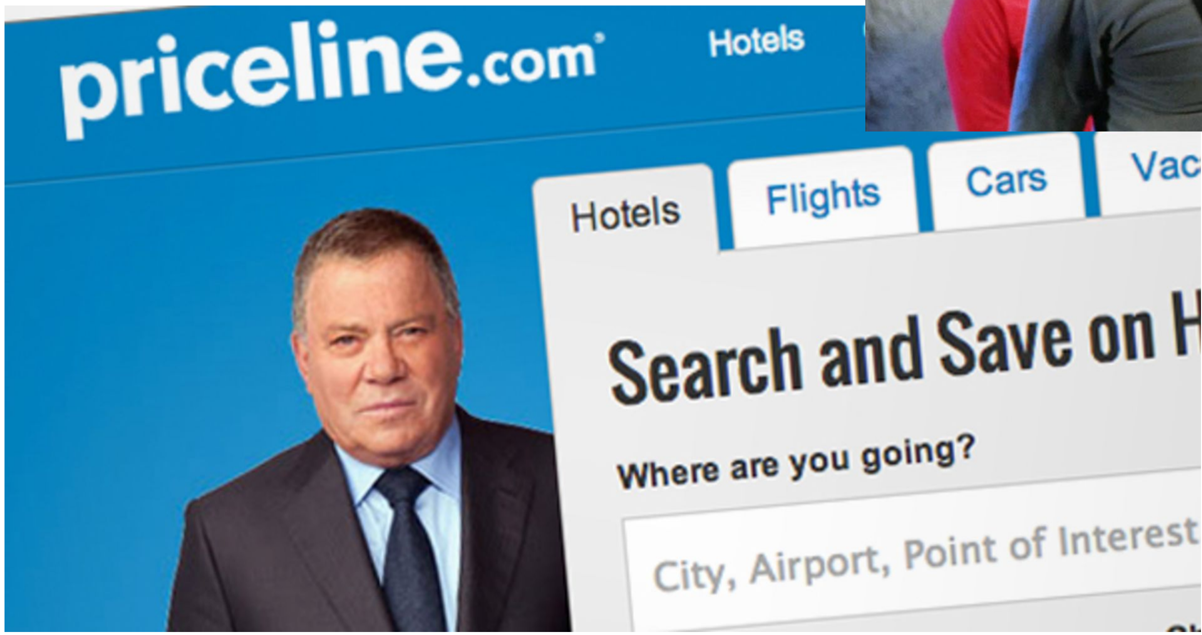

WET

Oldsmobiles

- GM
- Manufactured 1897~2004
- Went down hill in 1980’s
- [Desperation to survive](#)


MELANIE SHATNER

William Shatner




THE NEW GENERATION
OF OLDSMOBILE

NOT your father's Oldsmobile

- Fond memories
- New Generation of Oldsmobile
- Not your father's Oldsmobile

1st FISS Workshop

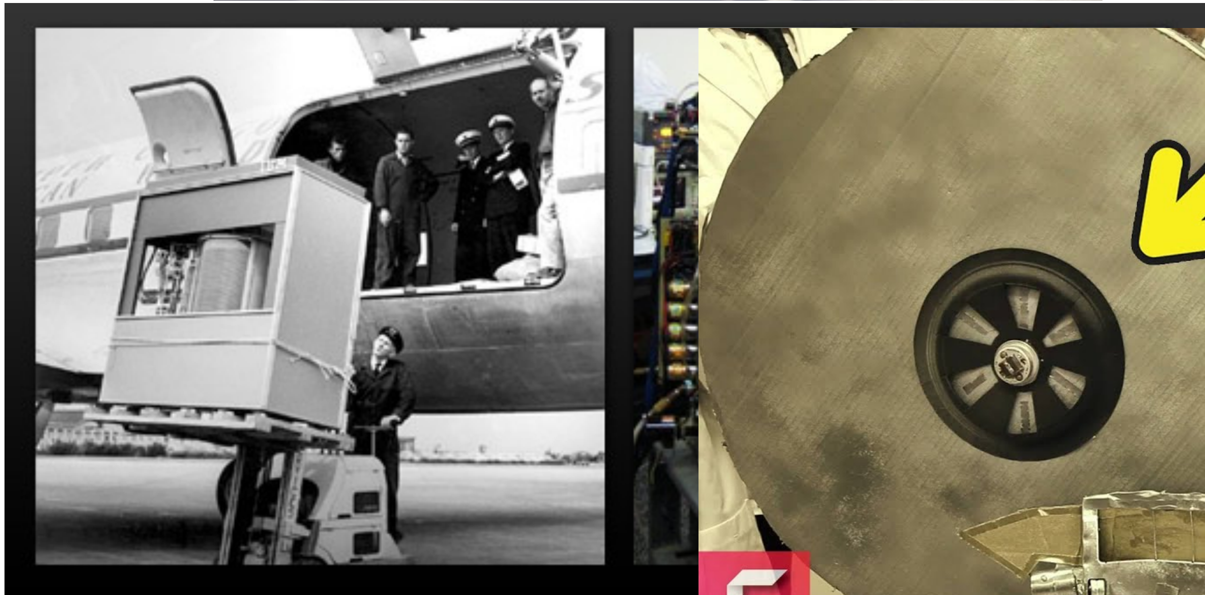
- International Workshop on File and Storage Systems
- Organized by KIISE SIGFAST

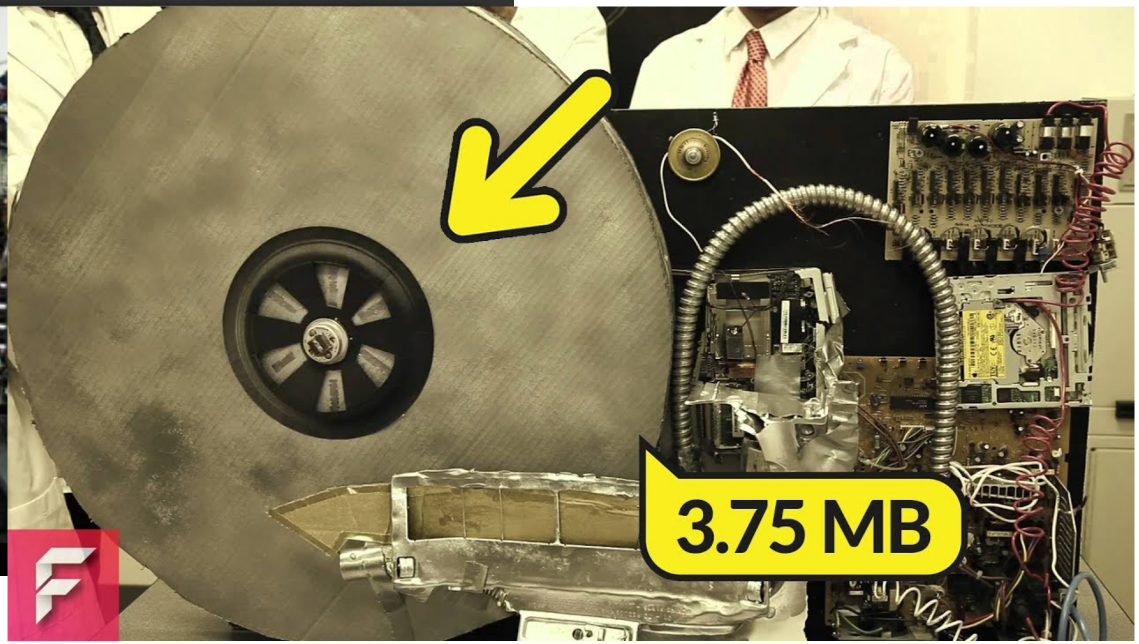
1st FISS Workshop

- International Workshop on File and Storage Systems
- Organized by KIISE SIGFAST

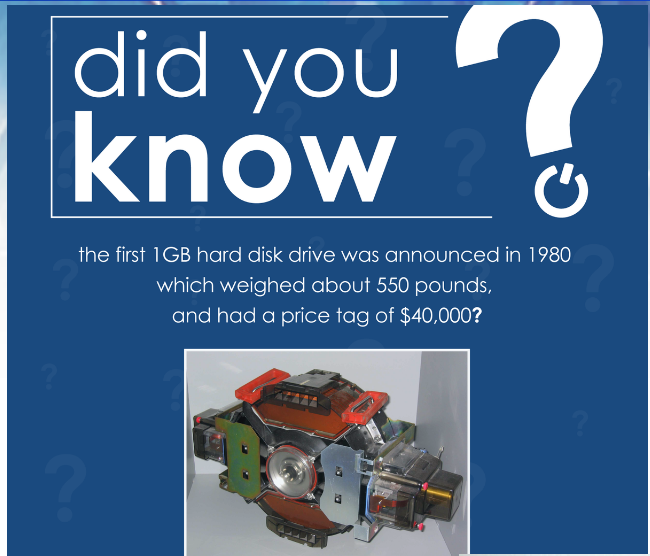
Hard Disk

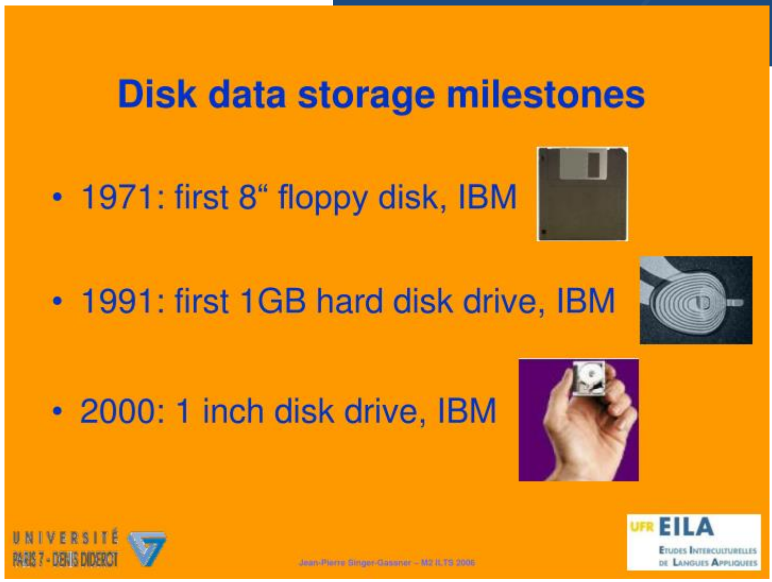
YES! your father's ~~Oldsmobile~~ Storage Device



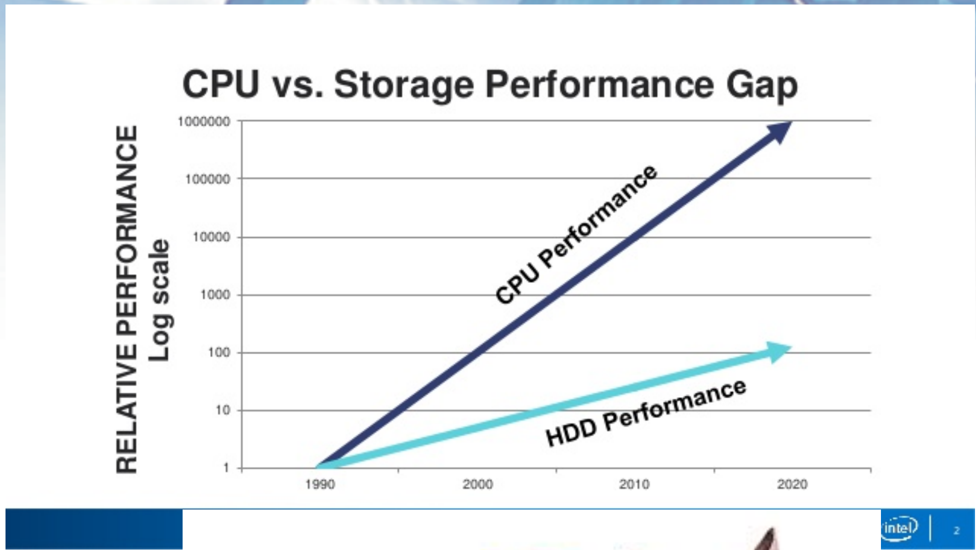


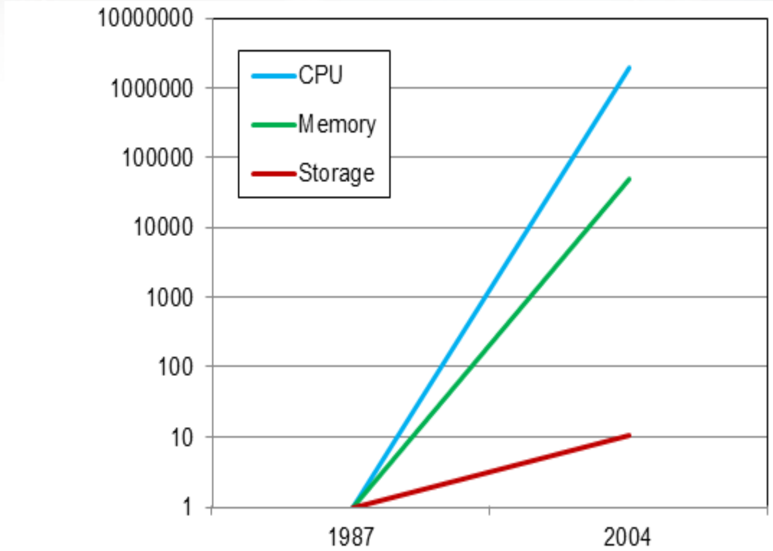
YES! your father's ~~Oldsmobile~~ Storage Device





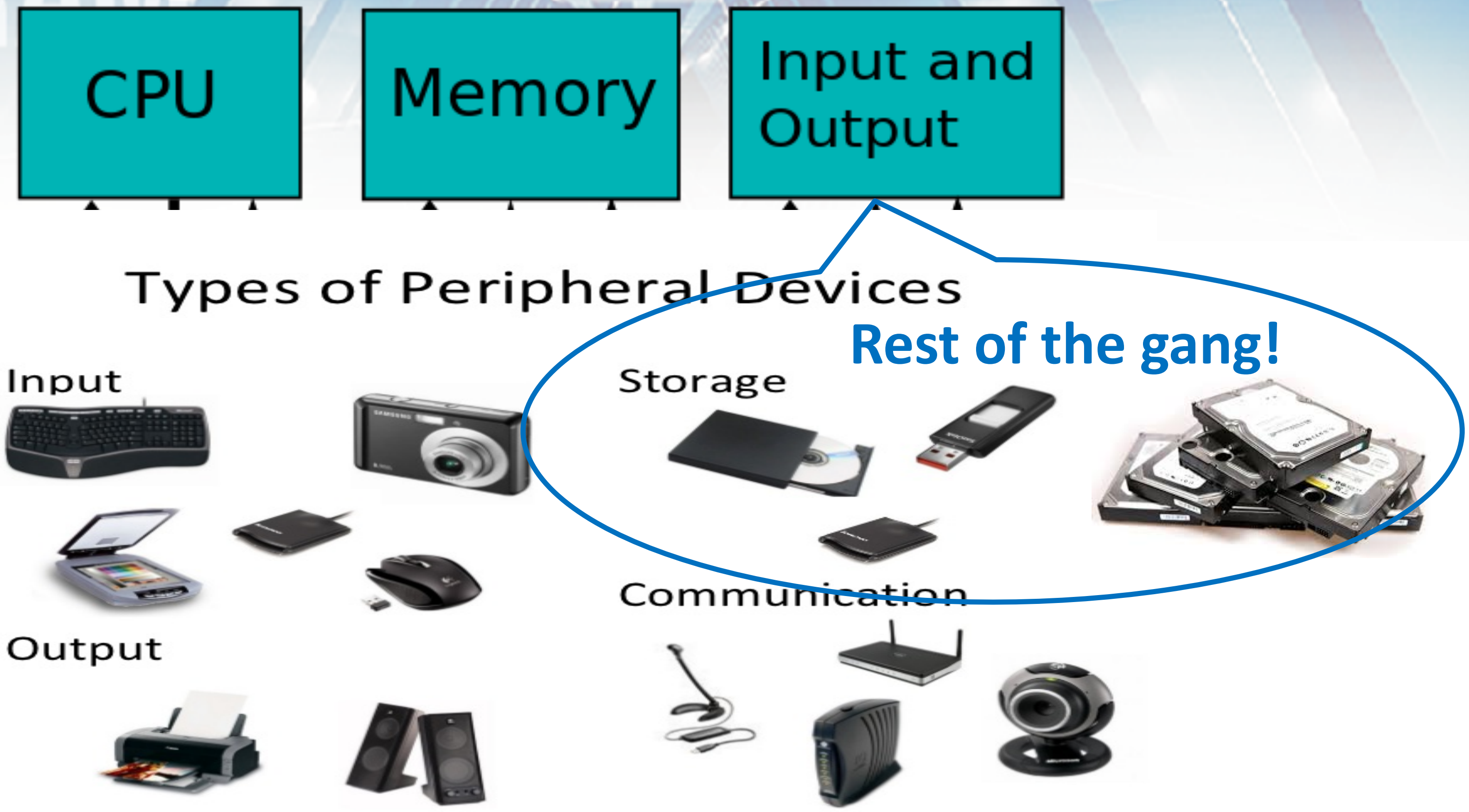
YES! your father's ~~Oldsmobile~~ Storage Device





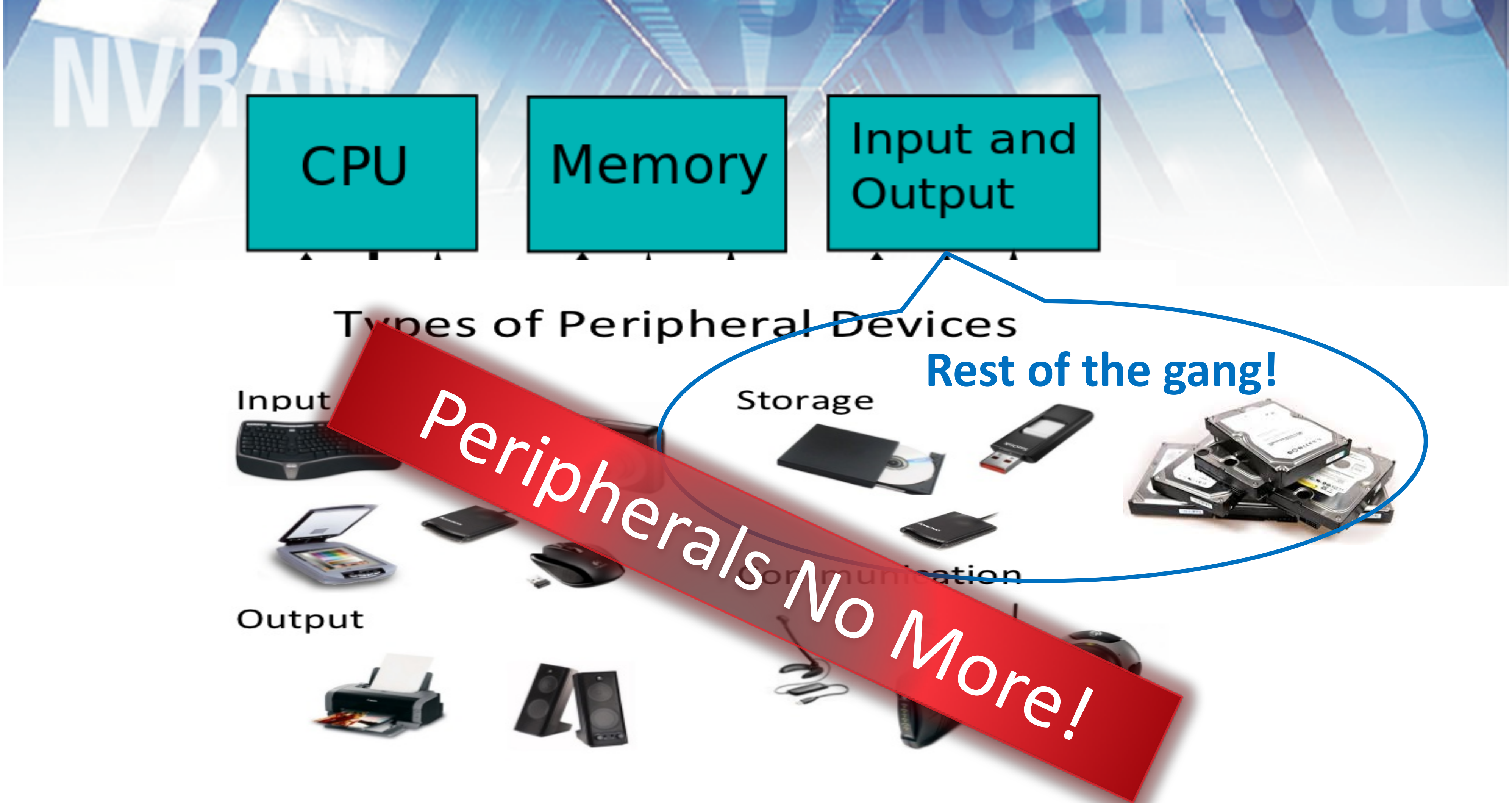
	1987	2004	Increase Multiple
CPU Performance	1 MIPS	2,000,000 MIPS	2,000,000 x
Memory Performance	100 usec	2 nsec	50,000 x
Disk Drive Performance	60 msec	5.3 msec	11x

YES! your father's ~~Oldsmobile~~ Storage Device



Peripheral: Auxiliary, Supplementary, relating to periphery

YES! your father's ~~Oldsmobile~~ Storage Device



Peripheral: Auxiliary, Supplementary, relating to periphery

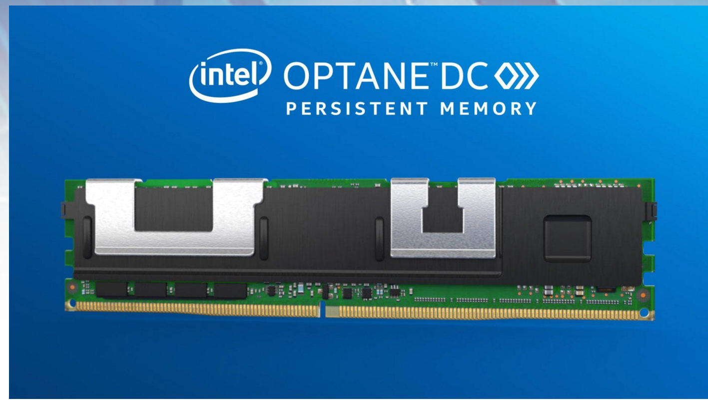
NOT your father's ~~Oldsmobile~~ Storage Device

- **New generation of storage**
  - Ultra Low Latency (ULL) drives
    - NVMe

	Samsung Z-SSD (SZ985)	**Intel Optane (P4800X)**
Technology	Z-NAND	3D Xpoint
Capacity	800GB	750GB
Sequential Read/Write (GB/s)	3.2GB/s (Both)	2.4GB/s Read 2GB/s Write
Random Read/Write (IOPS)	750K Read 170K Write	550K Read 500K Write
Random Read Latency	12-20us	10us
Random Write Latency	16us	10us

NOT your father's ~~Oldsmobile~~ Storage Device

- **New generation of storage**
  - DIMM slotted storage



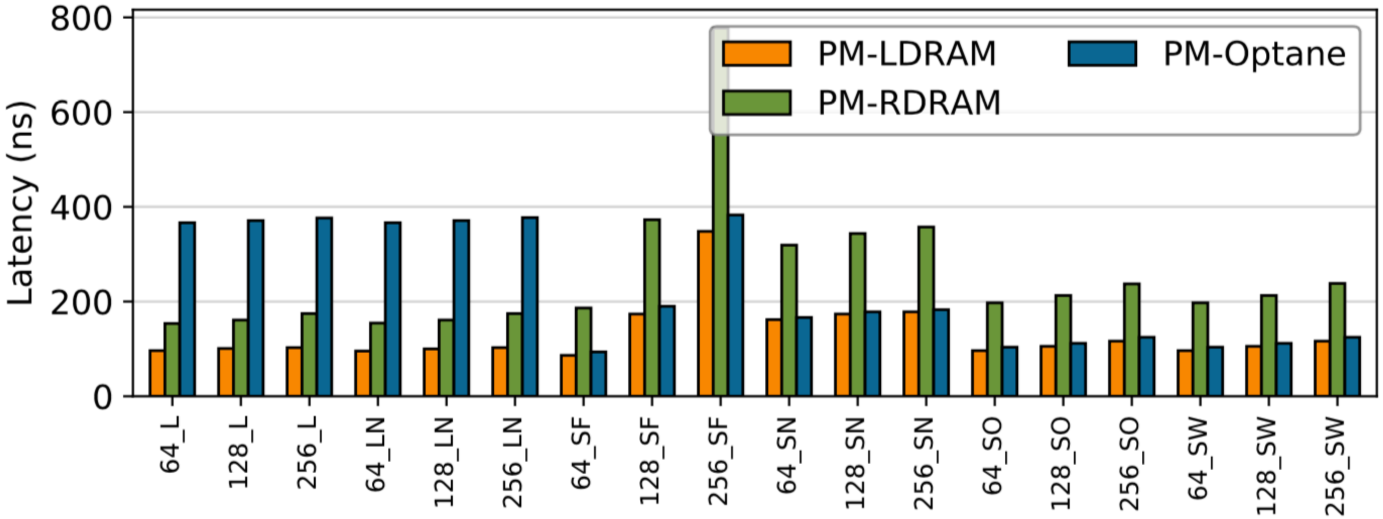
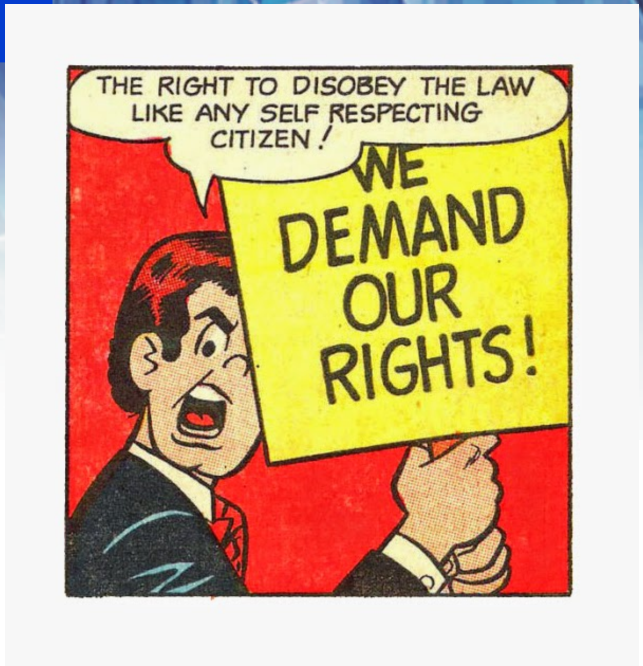


Figure 8: **Memory Instruction Latency** This graph shows the median latency for a variety of ways of accessing persistent memory. Note that for store instructions followed by flushes, there is little performance difference between PM-LDRAM and PM-3DXP, whereas the DRAM outperforms Optane DC memory for load sequences (see data in csvroot/basic/instruction_latency.csv).

Courtesy of NVSL, UCSD arXiv:1903.05714v2

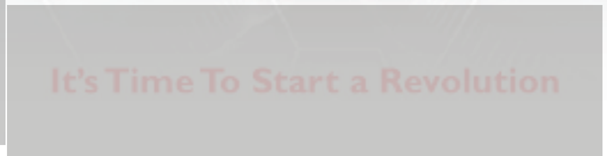

So What?

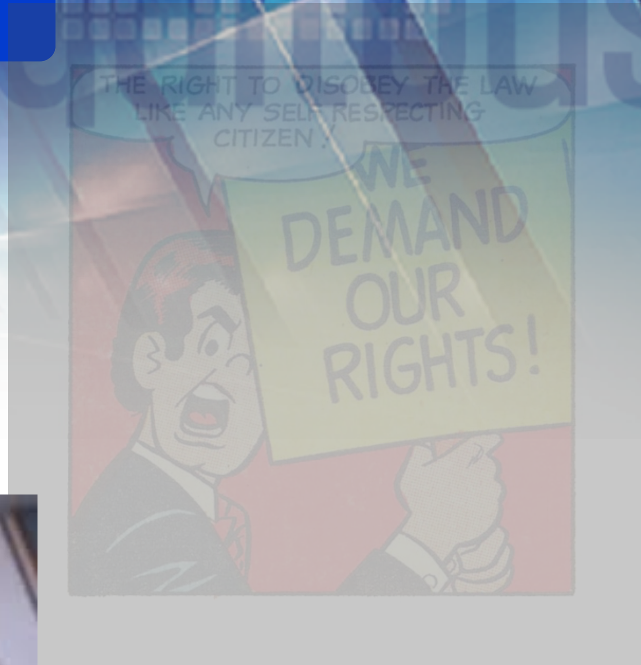
Change the world!



Peripherals No More!

Change the world!...slowly





Periphe

More!

One step at a time...

PAST storage topics of interest?

- **RAID**
  - Increase I/O bandwidth
- **Buffer Caching**
  - Improve latency
- **Swapping**
  - Improve resource sharing

Revisit & Rediscover

Take a fresh look at these old favorites.

PAST storage topics of interest?

- **RAID**
  - Increase I/O bandwidth
- **Buffer Caching**
  - Improve latency
- **Swapping**
  - Improve resource sharing

Revisit & Rediscover

Take a fresh look at these old favorites.

SWAN

It’s the network, stupid!

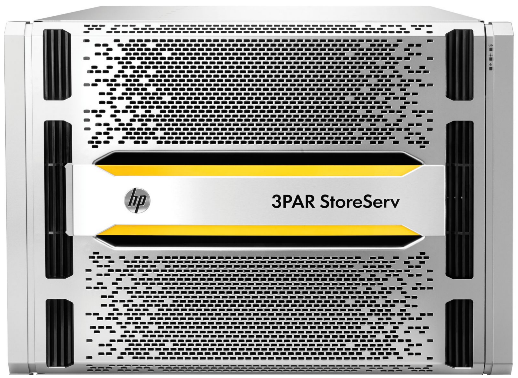
Alleviating Garbage Collection Interference through Spatial Separation in All Flash Arrays

HotStorage '17 & ATC '19

UNIST ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
ECE SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
NECSST Next-generation Embedded / Computer System Software Technology

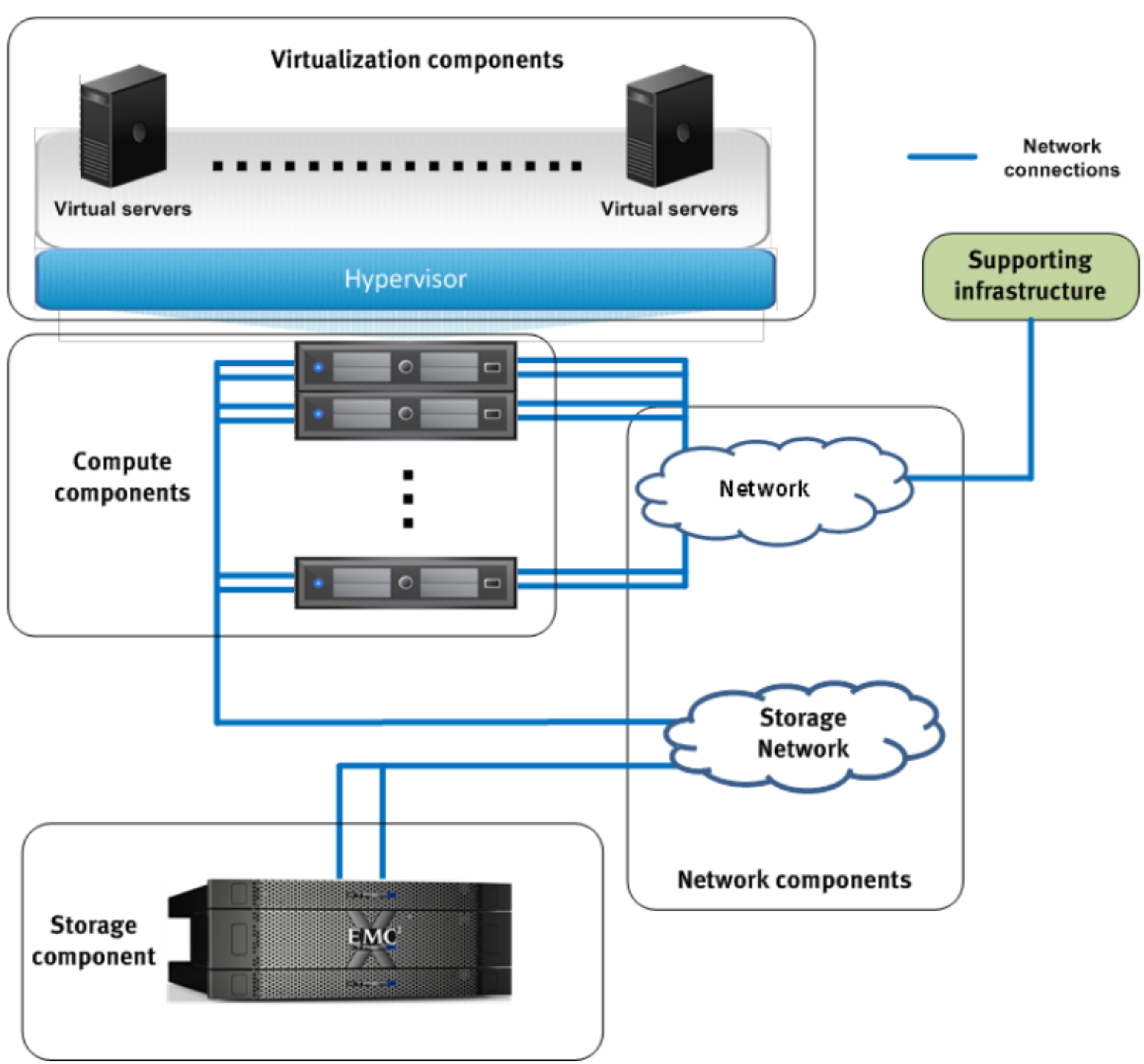
All Flash Array

- **All Flash Array (AFA)**
  - Storage infrastructure that contains only flash memory drives
    - Solid-State Array (SSA)

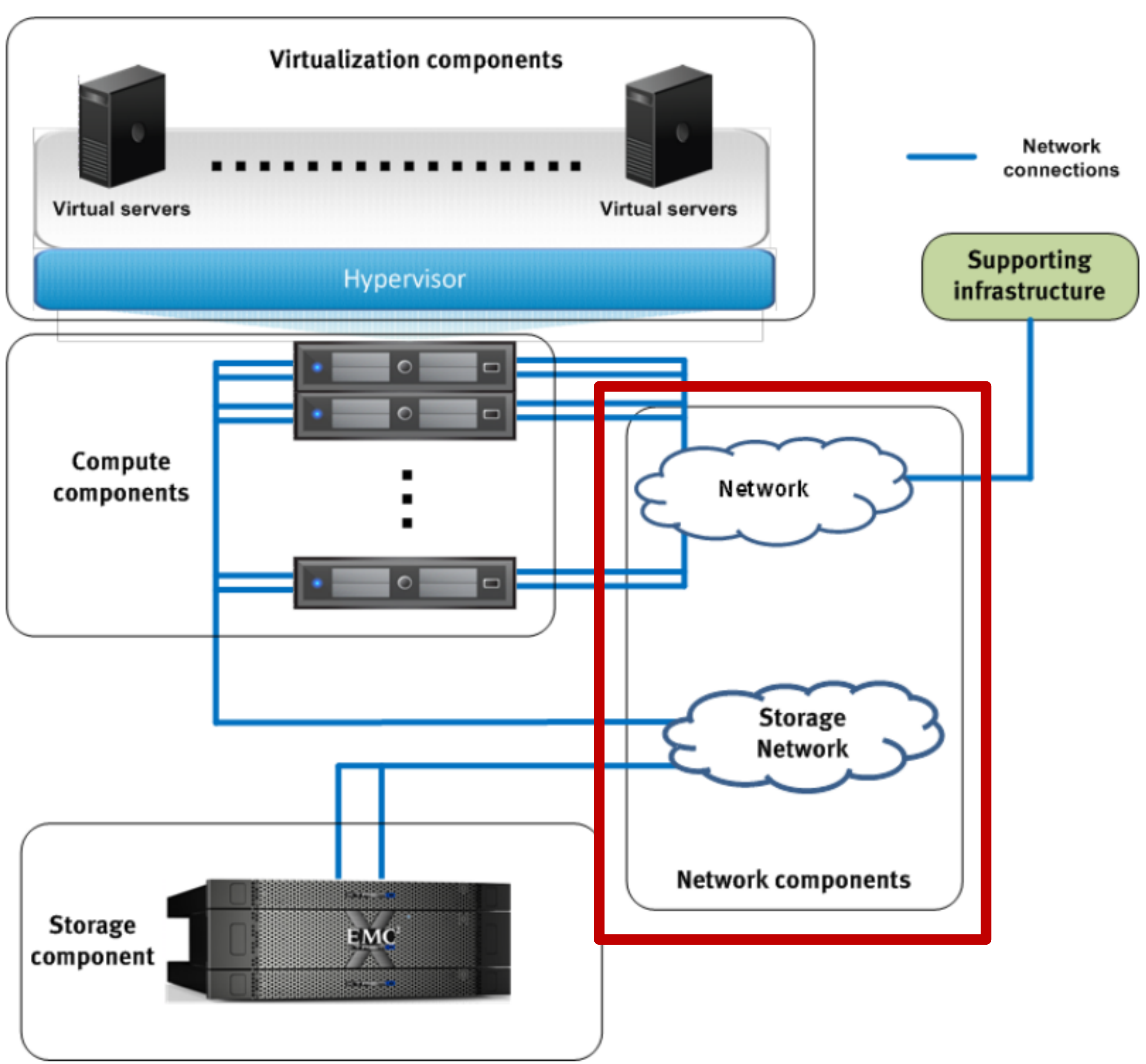


From: https://images.google.com/
https://www.purestorage.com/resources/glossary/all-flash-array.html

Architecture of All-Flash Array



Architecture of All-Flash Array

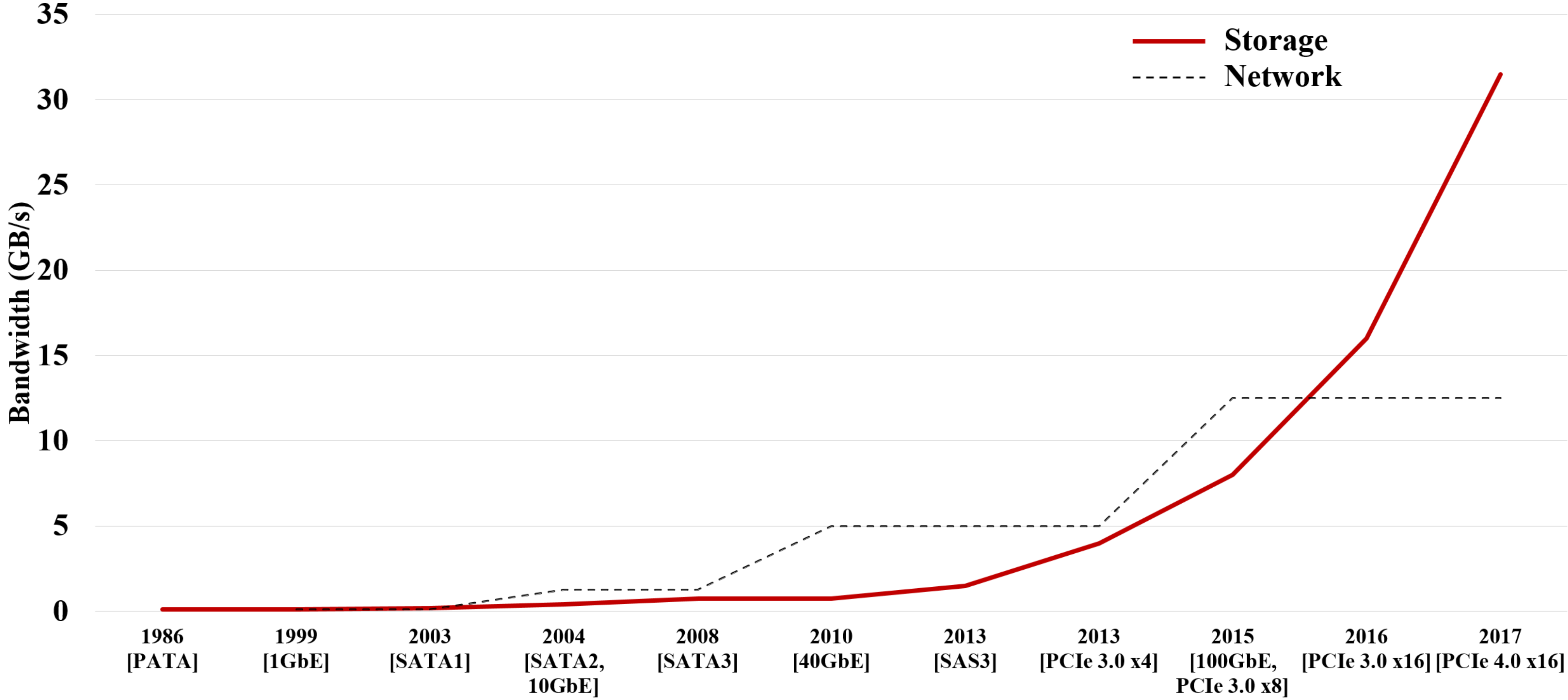


SSD Products for Data Center

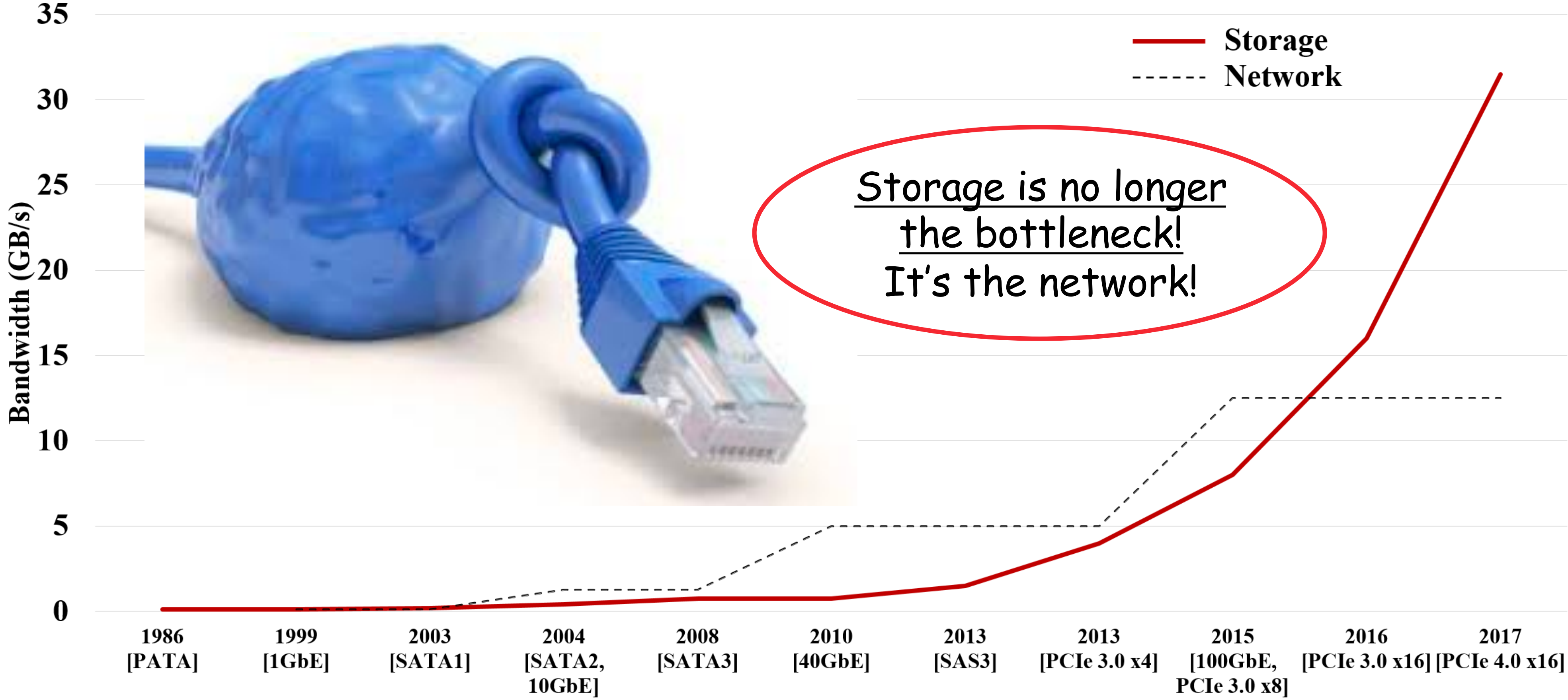
Manufacturer	Product Name	Sequential Read/Write (up to GB/s)	Random 4KB Read/Write (up to IOPS)	Interface
Intel	P3700	2.1 / 1	470K / 65K	PCIe 3 * 4
	P3520	1.7 / 1.3	370K / 26K	PCIe 3 * 4
	P3608	5 / 3	850K / 150K	PCIe 3 * 8
	S3710	0.5 / 0.5	85K / 45K	SATA 6Gb/s
Samsung	PM1725a	6.4 / 3	1M / 170K	PCIe 3 * 8
	PM963	2 / 1.2	430K / 40K	PCIe 3 * 4
	PM1633a	1.2 / 0.9	190K / 31K	SAS 3.0
	SM863	0.5 / 0.5	97K / 30K	SATA 6Gb/s

Intel: https://www-ssl.intel.com/content/www/us/en/solid-state-drives/data-center-family.html
Samsung: http://www.samsung.com/semiconductor/products/flash-storage/enterprise-ssd/

Interface Bandwidth Growth Trend



Interface Bandwidth Growth Trend

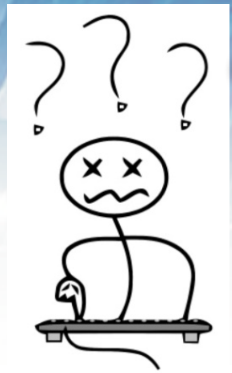


Comparison of All-flash Array

	Solid Fire (NetApp)	EMC	Pure Storage	Nimble
Model	SF19210	6X-Brick	M70	AF9000
Capacity	20TB (10 SSDs)	240TB (150 SSDs)	136TB	500TB
Performance (Random I/O)	100K	7GB (900K IOPS * 8KB)	9GB (300K IOPS * 32KB)	350K
Network	20Gb (iSCSI 10Gb * 2port)	240Gb (iSCSI 10Gb * 24port)	40Gb (iSCSI 10Gb * 4port)	40Gb (iSCSI 10Gb * 4port)
Bottleneck	Network	Storage	Network	Network

EMC: https://www.emc.com/collateral/data-sheet/h12451-xtremio-4-system-specifications-ss.pdf
Pure Storage: https://www.purestorage.com/content/dam/purestorage/pdf/datasheets/ps_ds5p_flasharraym_04.pdf
SolidFire: http://info.solidfire.com/rs/solidfire/images/SolidFire_ProductDatasheet.pdf
Nimble storage: https://www.nimblestorage.com/technology-products/all-flash-array-specifications/

Comparison of All-flash Array



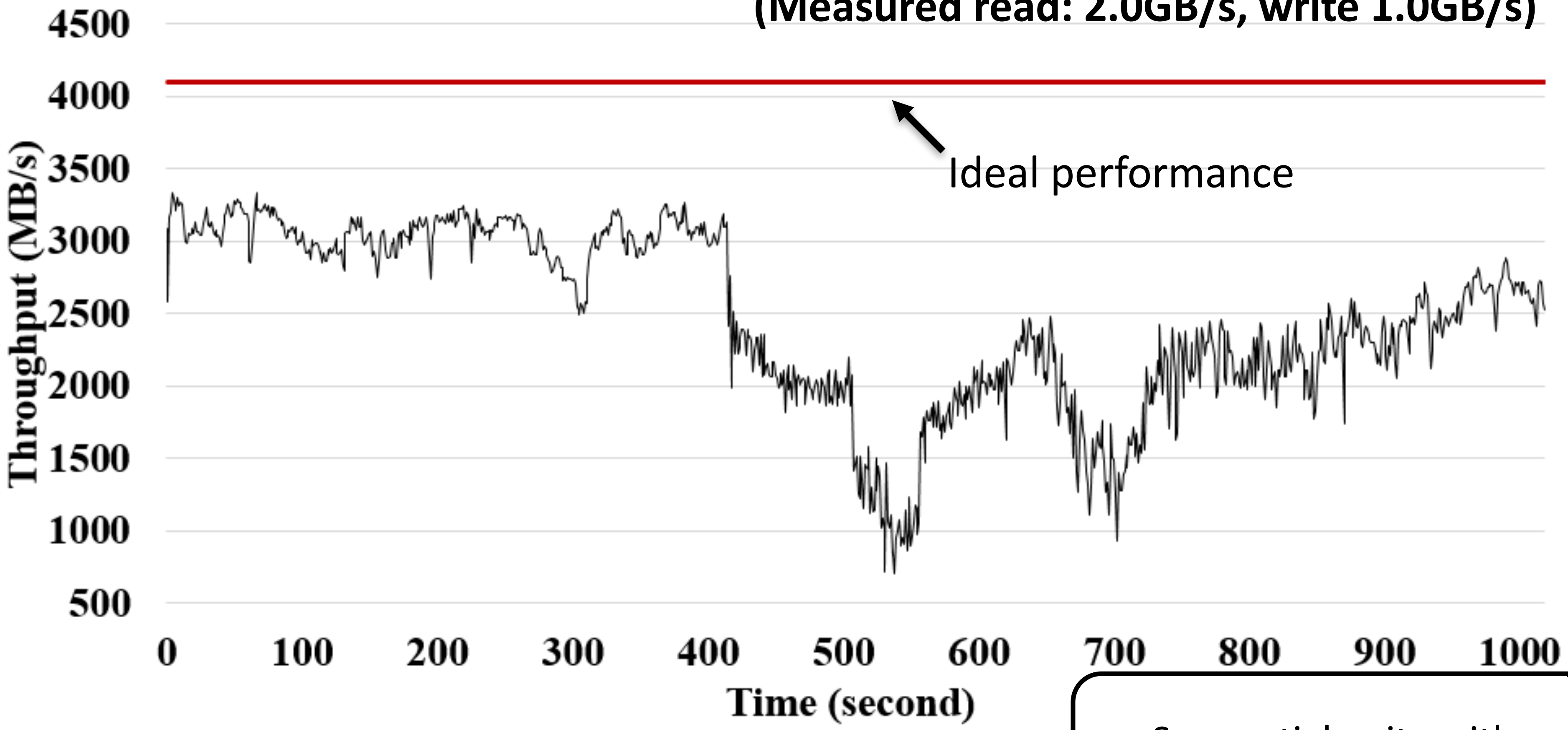
Do these many SSDs really help?

	Solid Fire (NetApp)	EMC	Pure Storage	Nimble
Model	SF19210	6X-Brick	M70	AF9000
Capacity	20TB (10 SSDs)	240TB (150 SSDs)	136TB	500TB
Performance (Random I/O)	100K	7GB (900K IOPS * 8KB)	9GB (300K IOPS * 32KB)	350K
Network	20Gb (iSCSI 10Gb * 2port)	240Gb (iSCSI 10Gb * 24port)	40Gb (iSCSI 10Gb * 4port)	40Gb (iSCSI 10Gb * 4port)
Bottleneck	Network	Storage	Network	Network

EMC: https://www.emc.com/collateral/data-sheet/h12451-xtremio-4-system-specifications-ss.pdf
Pure Storage: https://www.purestorage.com/content/dam/purestorage/pdf/datasheets/ps_ds5p_flasharraym_04.pdf
SolidFire: http://info.solidfire.com/rs/solidfire/images/SolidFire_ProductDatasheet.pdf
Nimble storage: https://www.nimblestorage.com/technology-products/all-flash-array-specifications/

Experiments with 4 SSD RAID 0

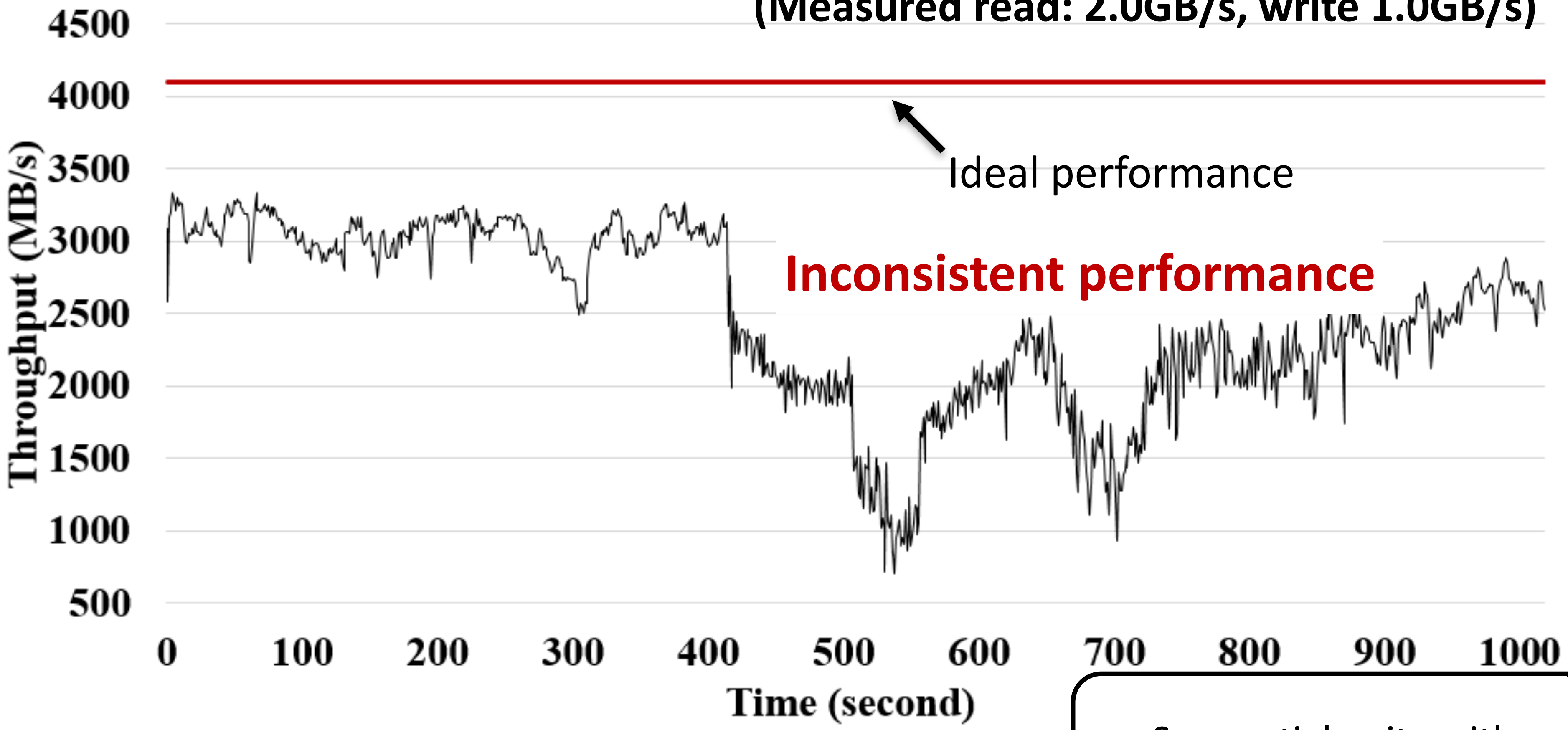
RAID 0 with 4 NVMe SSDs (spec. read: 2.4GB/s, write: 1.2GB/s)
(Measured read: 2.0GB/s, write 1.0GB/s)



Sequential write with 128KB I/O size

Experiments with 4 SSD RAID 0

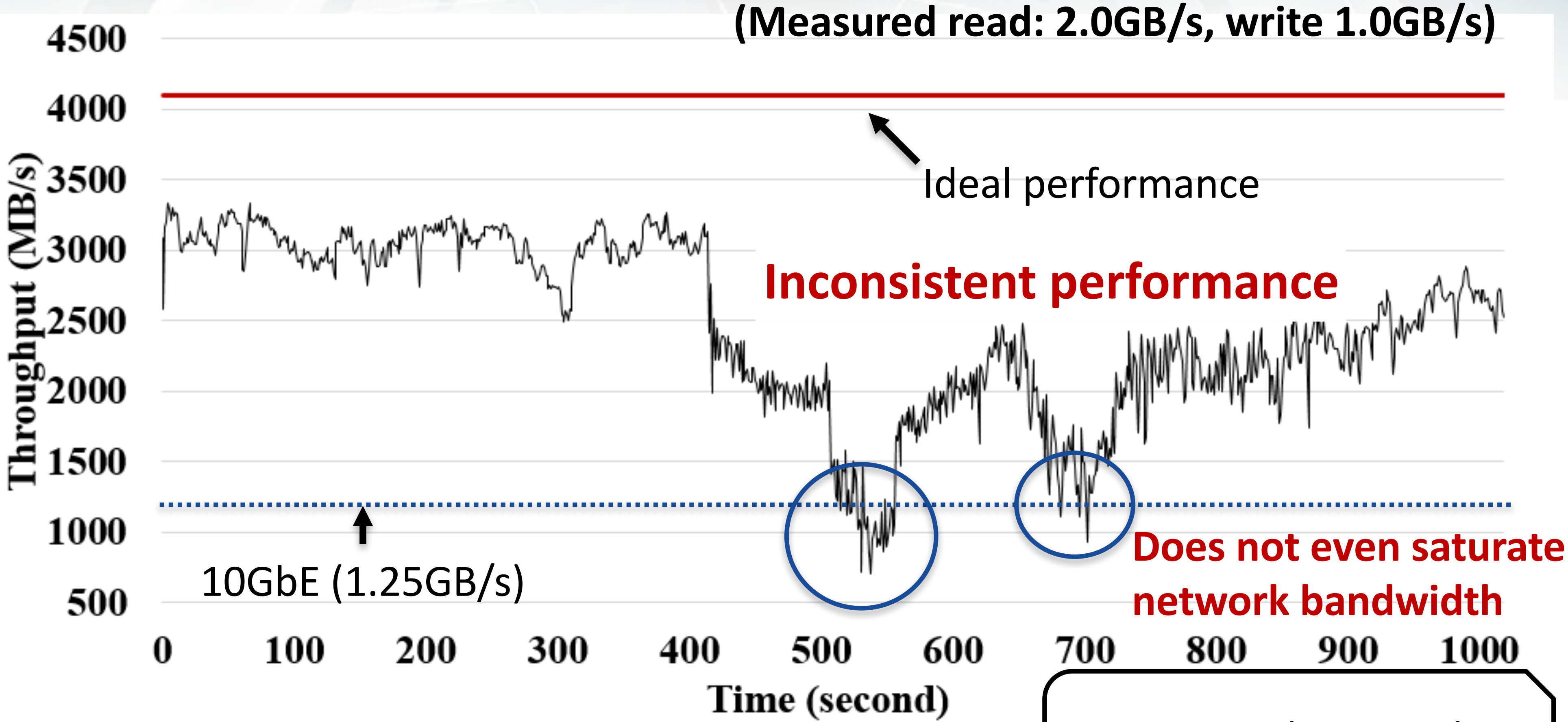
RAID 0 with 4 NVMe SSDs (spec. read: 2.4GB/s, write: 1.2GB/s)
(Measured read: 2.0GB/s, write 1.0GB/s)



Sequential write with 128KB I/O size

Experiments with 4 SSD RAID 0

RAID 0 with 4 NVMe SSDs (spec. read: 2.4GB/s, write: 1.2GB/s)
(Measured read: 2.0GB/s, write 1.0GB/s)



Sequential write with 128KB I/O size

Observations

- Inconsistent performance due to garbage collection
- Performance even limited by network bandwidth

Different approach to arrays of disks

- Inconsistent performance due to garbage collection

Get rid of garbage collection!

- Performance even limited by network bandwidth

Provide full network performance!

Our goal

Sustained, consistent
full network bandwidth performance!

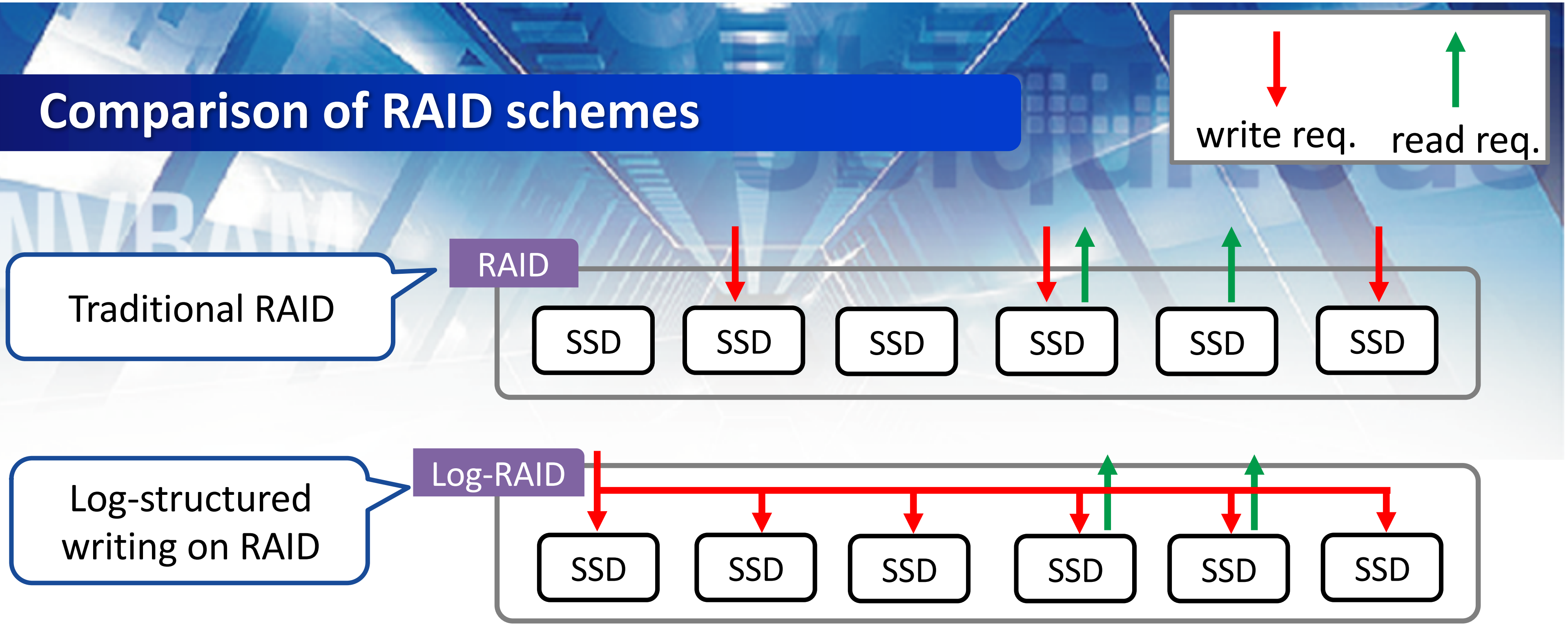
Design of SWAN

- **Our system**
  - SWAN (**S**patial separation **W**ithin an **A**rray of **S**SDs on a **N**etwork)

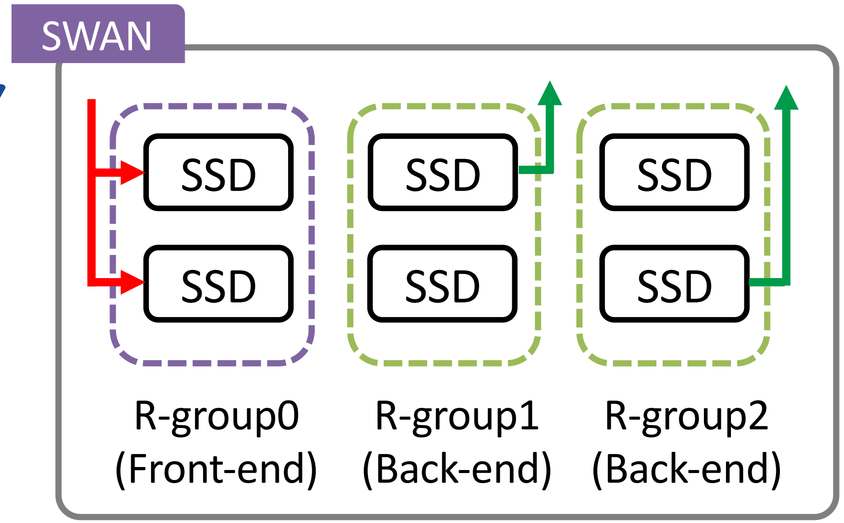
- **Goals**
  - Provide sustainable high performance for AFA
    - Alleviating GC interference at both SSD-level and AFA-level

- **Approach**
  - Spatial separation of application I/O and AFA I/O
  - Minimize GC interference by organizing SSDs into two-dimensional array

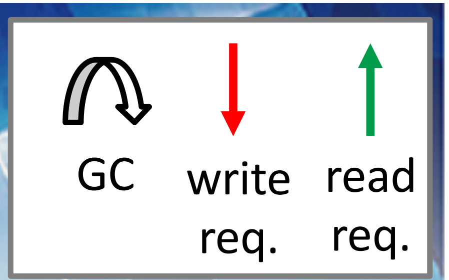
Comparison of RAID schemes



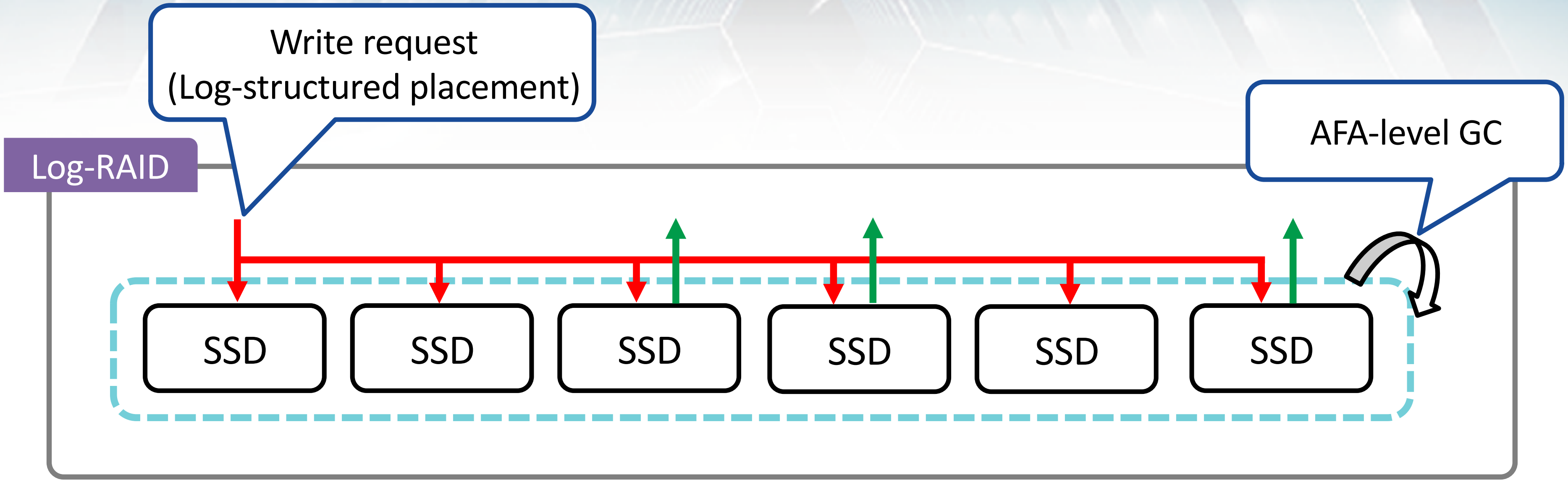
SWAN
- Two dimensional array
- Log-structured writing per R-group
- Front-end servers write requests
- Back-end is used for AFA-level GC



How Log-RAID Works



- **Key operations of Log-RAID**

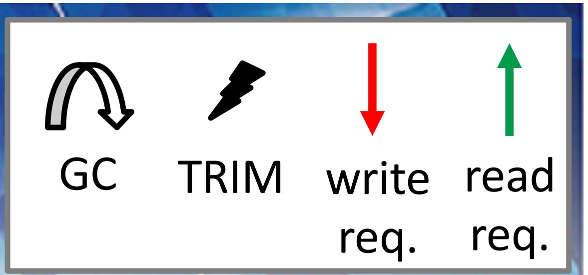


[9] CHIUEH, T.-C., TSAO, W., SUN, H.-C., CHIEN, T.-F., CHANG, A.- N., AND CHEN, C.-D. Software orchestrated flash array. In *Proceed- ings of International Conference on Systems and Storage (SYSTOR)* (2014), pp. 14:1–14:11.

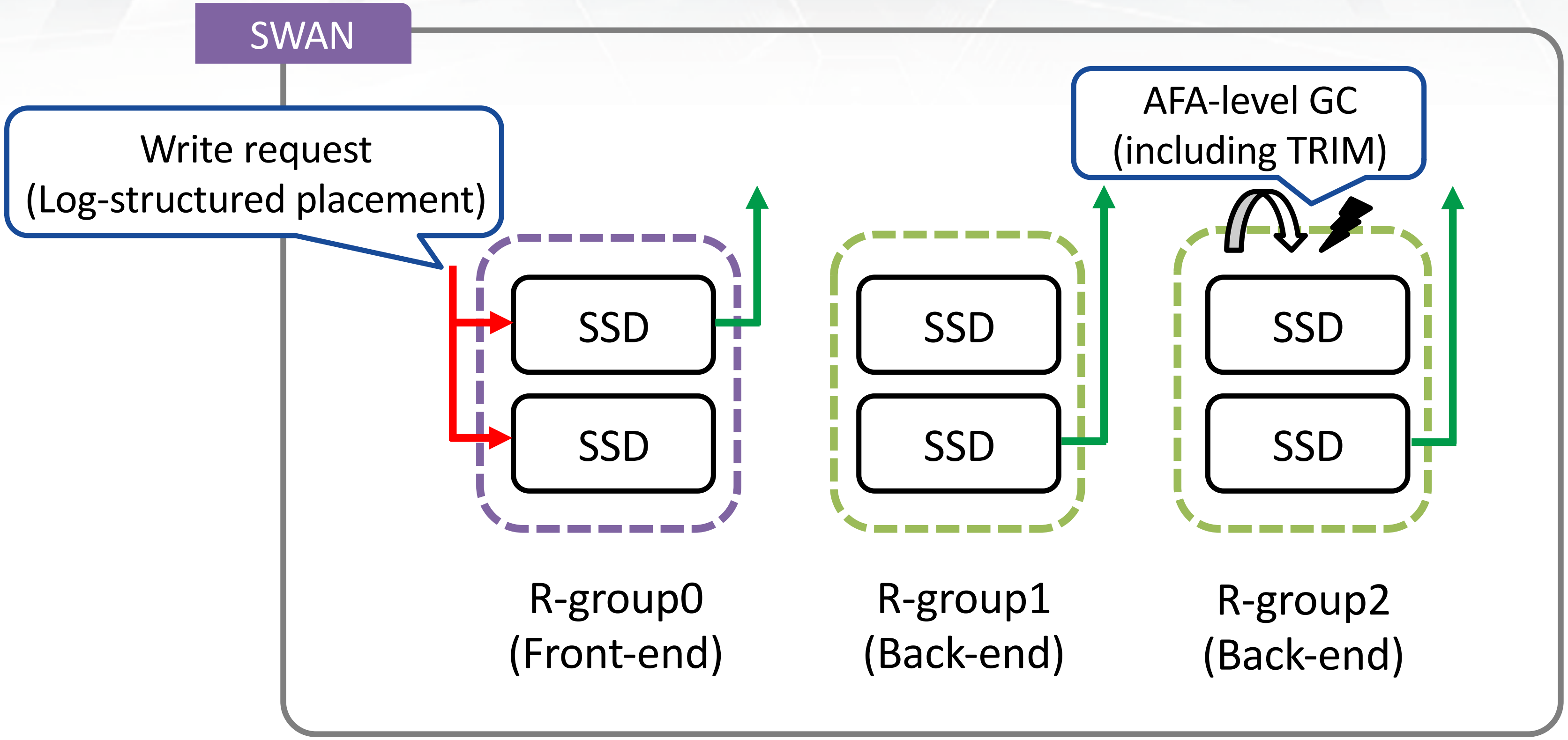
[21] IOANNOU, N., KOURTIS, K., AND KOLTSIDAS, I. Elevating com- modity storage with the SALSA host translation layer. In *Proceedings of the 26th IEEE Internationial Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MAS- COTS)* (2018), pp. 277–292.

[10] COLGROVE,J.,DAVIS,J.D.,HAYES,J.,MILLER,E.L.,SANDVIG, C., SEARS, R., TAMCHES, A., VACHHARAJANI, N., AND WANG, F. Purity: Building Fast, Highly-Available Enterprise Flash Storage from Commodity Components. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2015), pp. 1683– 1694.

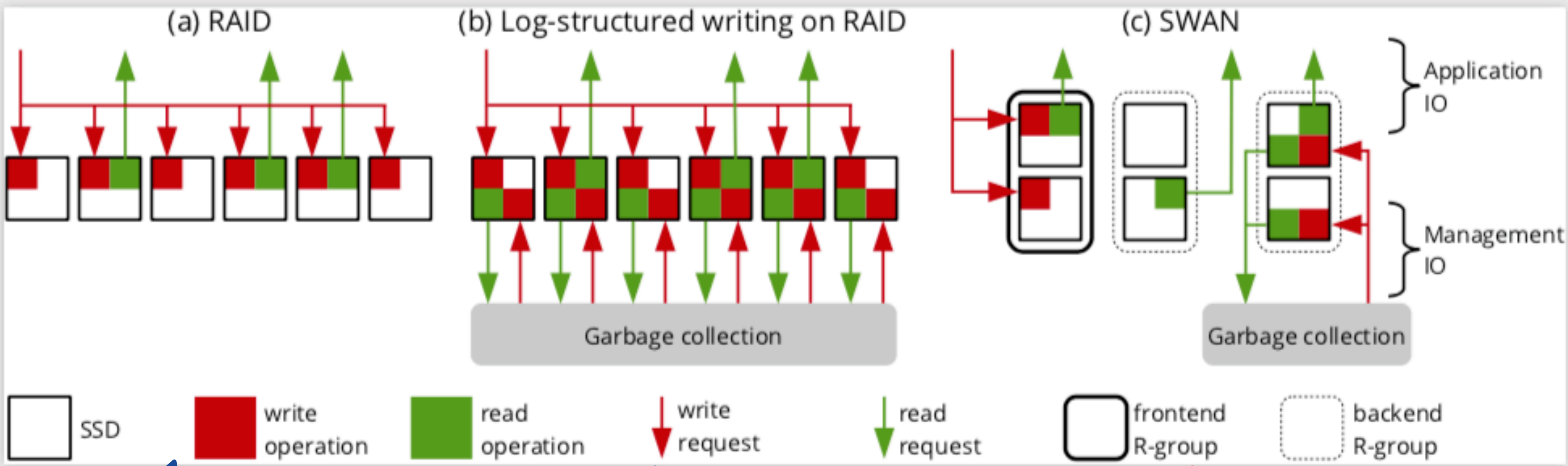
How SWAN Works



- **Key operations of SWAN**



I/O operation in All Flash Array



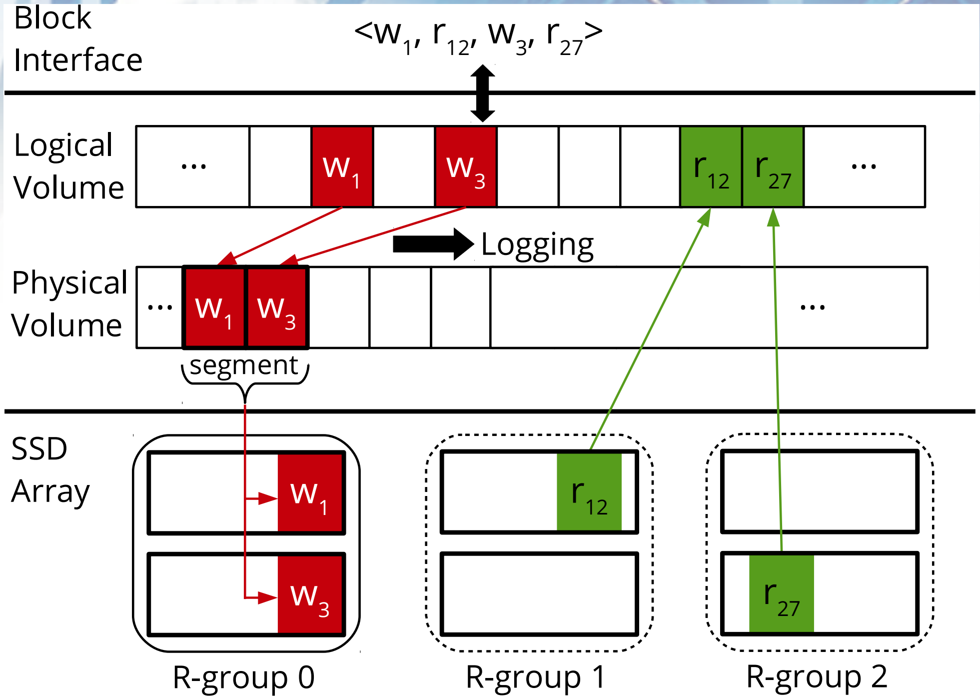
Susceptible to performance degradation due to high GC overhead inside SSD (Due to random writes)

AFA-level GC I/O may significantly interfere with application I/O

Spatial separation of application I/O and AFA-level GC I/O to minimize I/O interference

Handling Read/Write Req. in SWAN

w_n : write req. for block n
r_n : read req. for block n



Conf.
- R-group 0: Front-end
- R-group 1,2: Back-end
- Read/write req. arrives via block interface

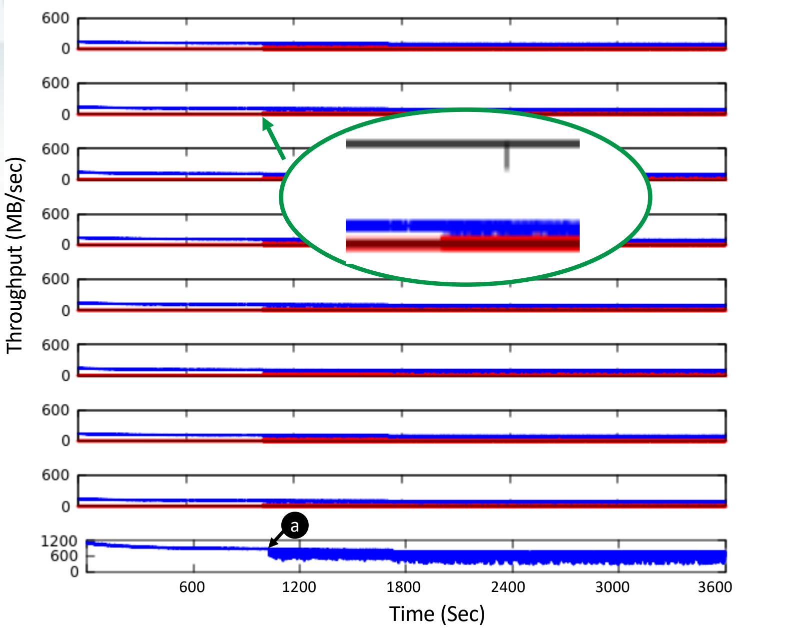
- **Operations**
  - SWAN appends write req. to the log and issues write req. to the front-end
  - Read req. will be served by any R-group holding the requested blocks

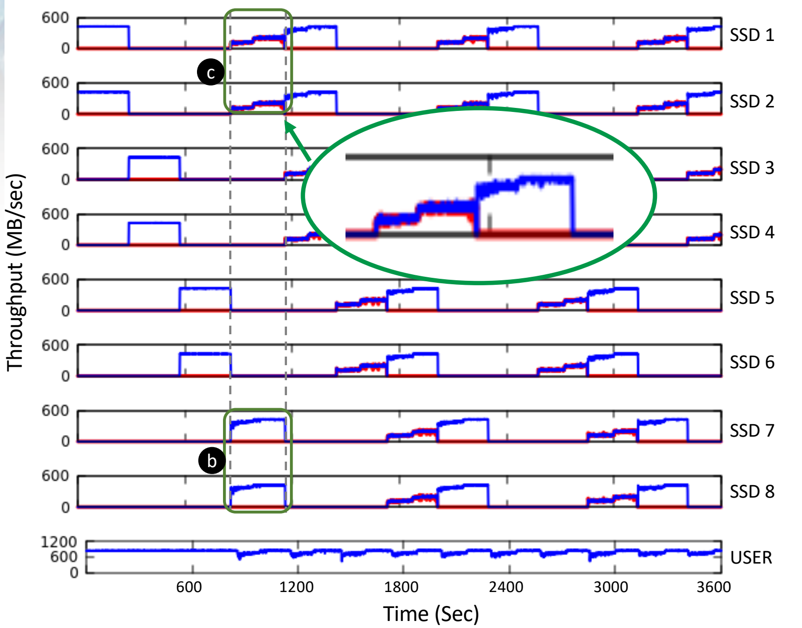
Evaluation

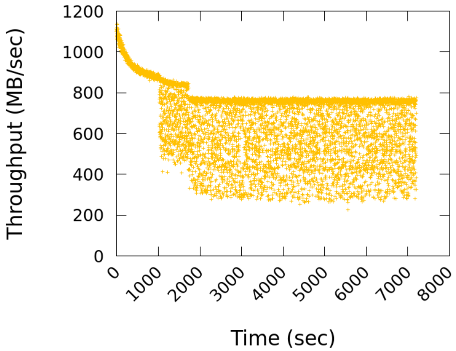
- **Environment**
  - Dell R730 server equipped with 2 Xeon CPUs and 64GB DRAM
  - Samsung 850 PRO 128GB * 9
- **Target config.**
  - RAID-0/4/5
  - Log-RAID-0/4
  - SWAN-0/4
- **Workloads**
  - Microbenchmark
  - YCSB-A, B, C, and D

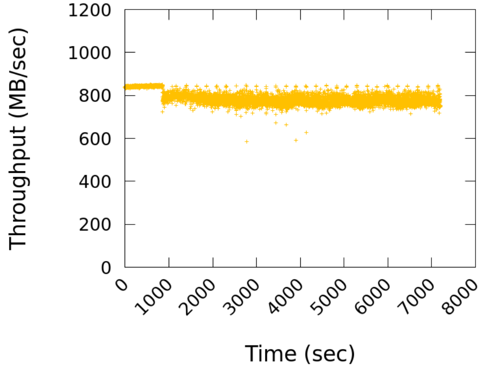
Analysis of GC Behavior

- Random write workload









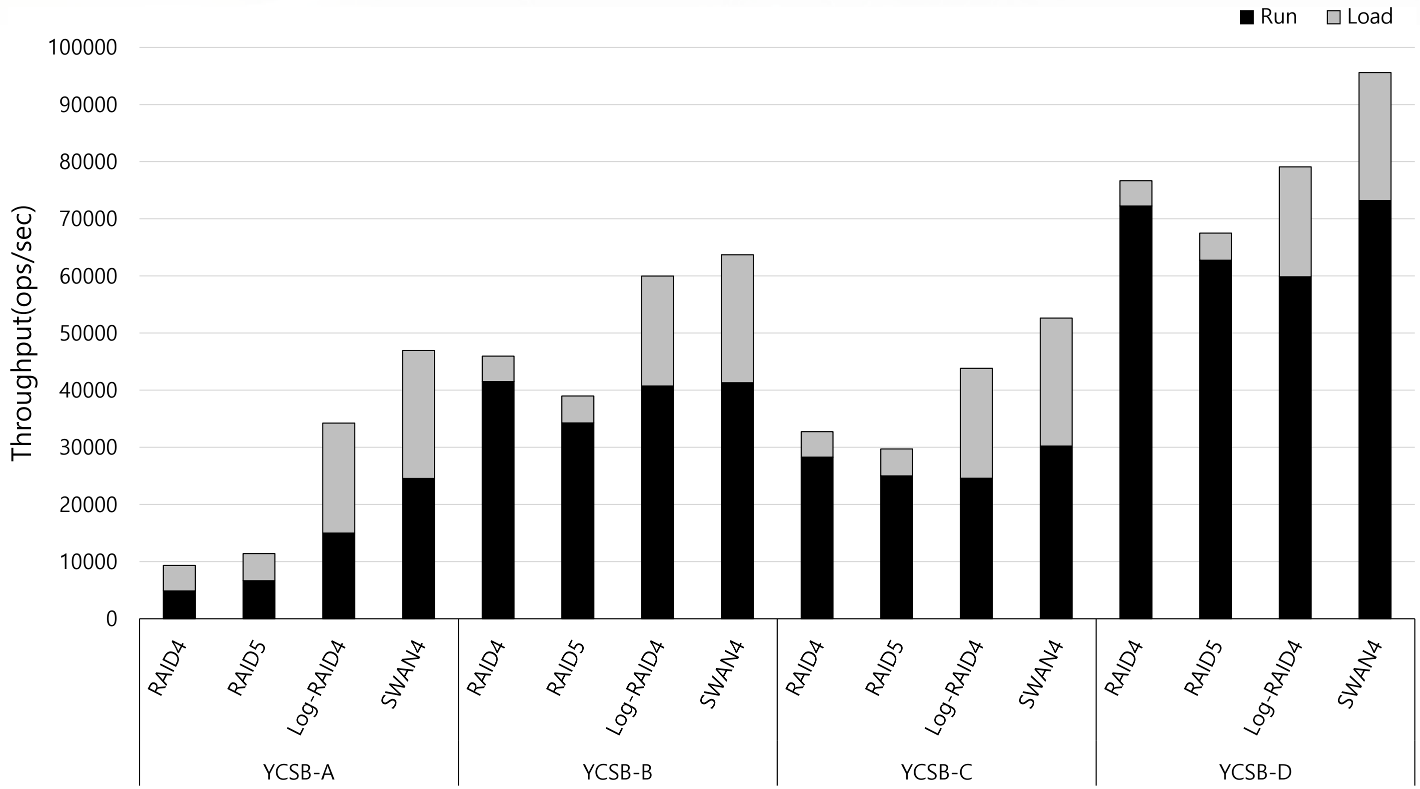
Log-RAID (8 SSDs)

SWAN (4 R-groups / 2 SSD per R-group)

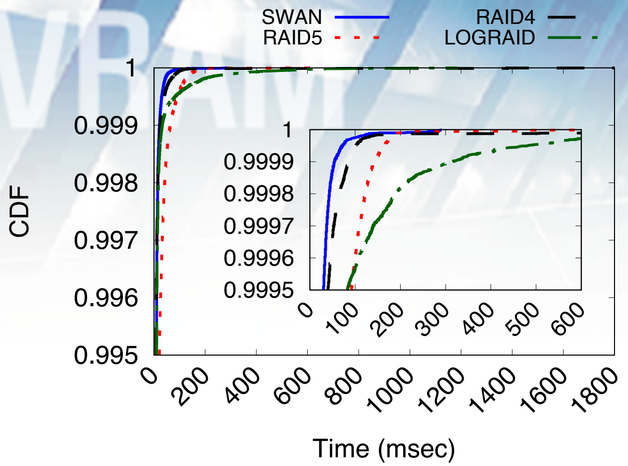
Throughput Results

Configuration

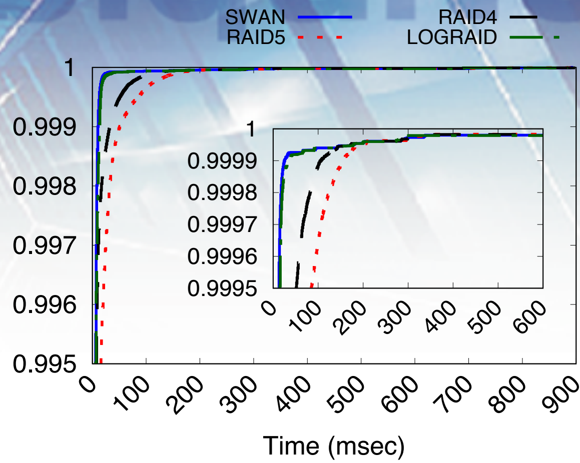
- RAID4/5: 8 data SSDs + 1 parity SSD
- Log-RAID: 8 data SSDs + 1 parity SSD
- SWAN4: 3 R-group with 2 data SSDs and 1 parity SSD per R-group



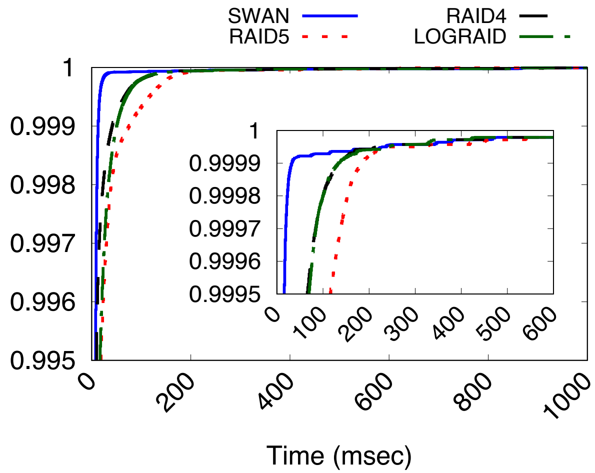
Read Latency Results (CDF)



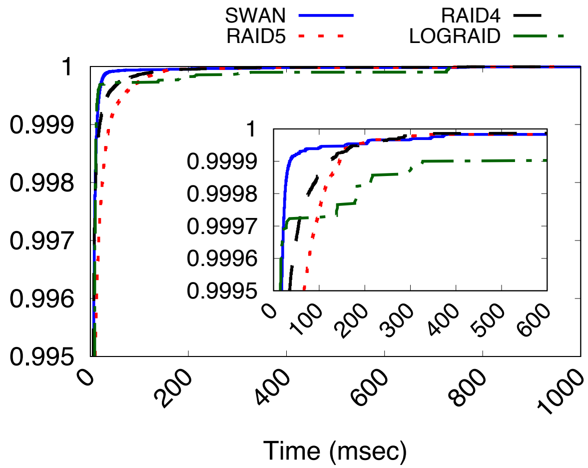
YCSB-A



YCSB-B



YCSB-C



YCSB-D

Summary

- **Proposed SWAN**
  - New management policy for All Flash Array
- **Key idea of SWAN**
  - Decouple GC I/Os from normal ones by partitioning the SSD array into 2 groups
- **Benefits of Swan**
  - SSD can be simpler

It's the network, stupid!

PAST storage topics of interest?

- RAID
  - Increase I/O bandwidth
- **Buffer Caching**
  - Improve latency
- Swapping
  - Improve resource sharing

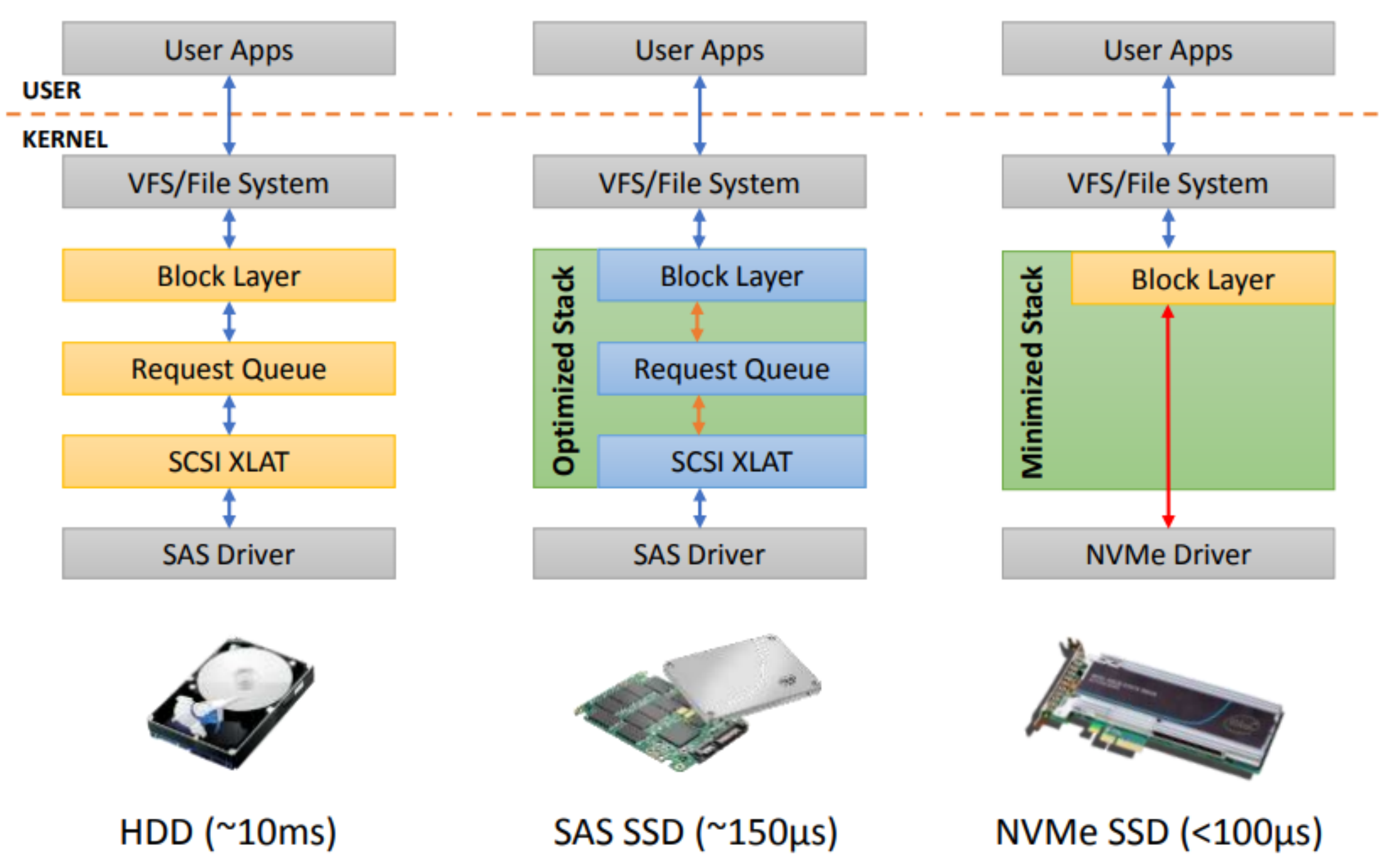
Revisit & Rediscover

Take a fresh look at these old favorites.

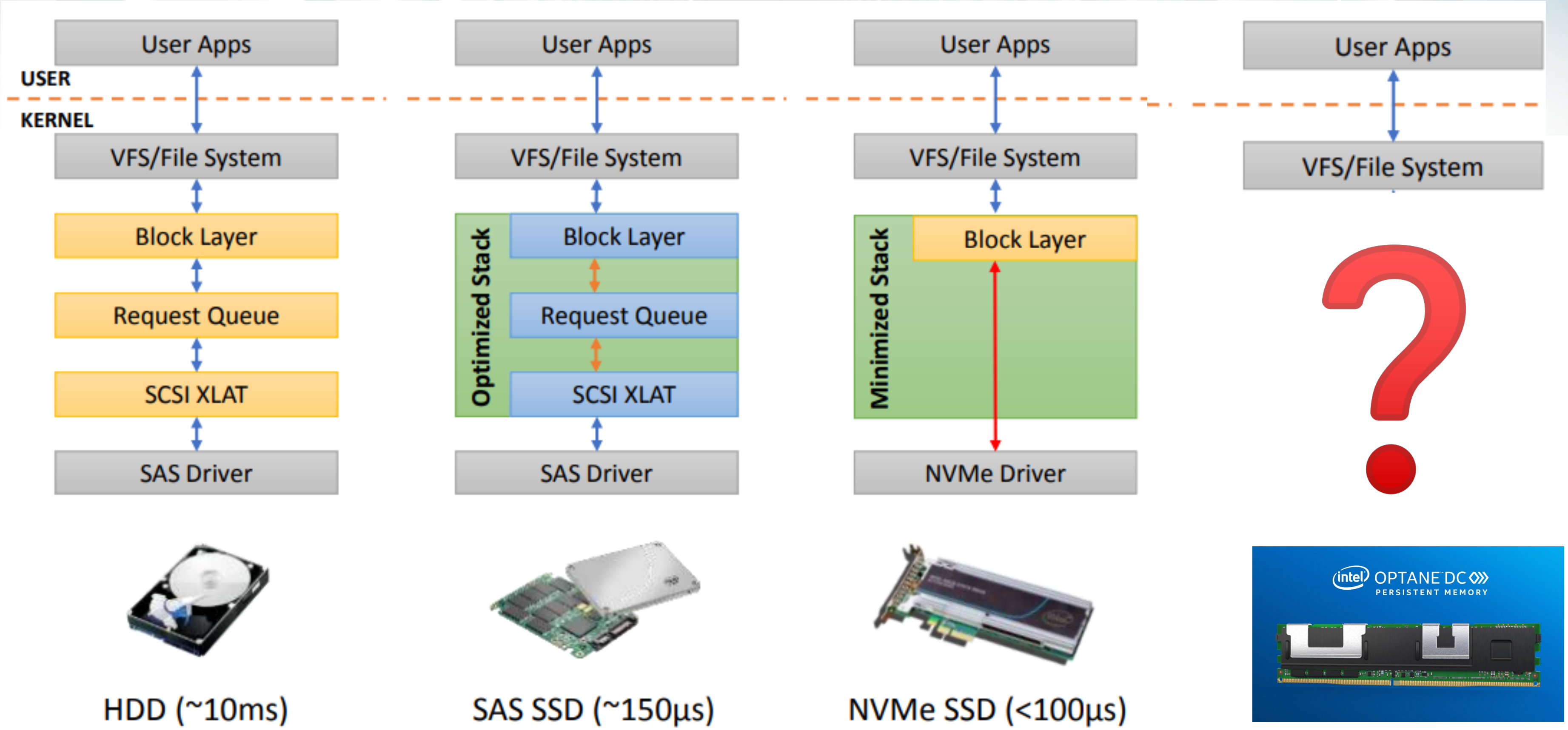
First Responder

It’s the storage stack, stupid!

Evolution of storage stack

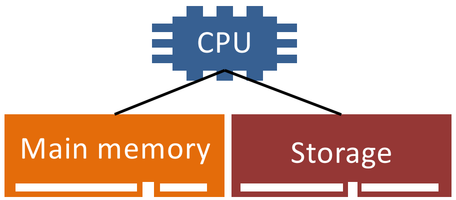


Evolution of storage stack

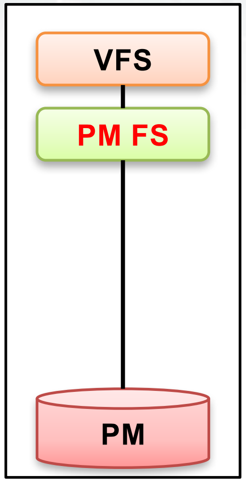


PM Targeted File Systems

- Designed to reap PM performance



PM **as Storage**

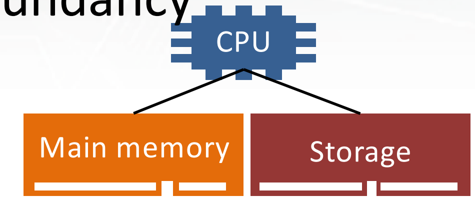


PM-aware **File System**

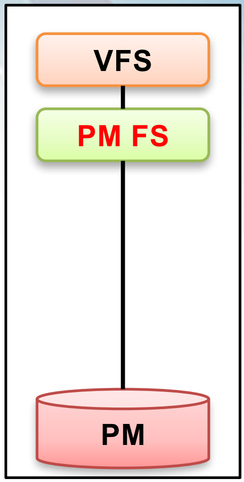
SOSP 2009	"BPFS (Better I/O Through Byte-Addressable, Persistent Memory)"
SC 2011	"SCMFS (SCMFS: A File System for Storage Class Memory)"
EuroSys 2014	"PMFS (System Software for Persistent Memory)"
EuroSys 2014	"Aerie (Aerie: Flexible File-System Interfaces to Storage-Class Memory)"
EuroSys 2016	"HiNFS (A High Performance File System for Non-Volatile Main Memory)"
FAST '16, SOSP '17	"NOVA (NOVA-Fortis: A Fault-Tolerant Non-Volatile Main Memory File System)"
SOSP 2017	"Strata (Strata: A Cross Media File System)"

BUT…

- **DAX approach**
  - Weak reliability, data integrity, redundancy
  - PM as end destination media
- **PM only**
  - Replace traditional storage?
  - Exception: Strata and Ziggurat
- **Lengthy process to maturity**
  - Ext4…still in progress
  - Wisdom with age



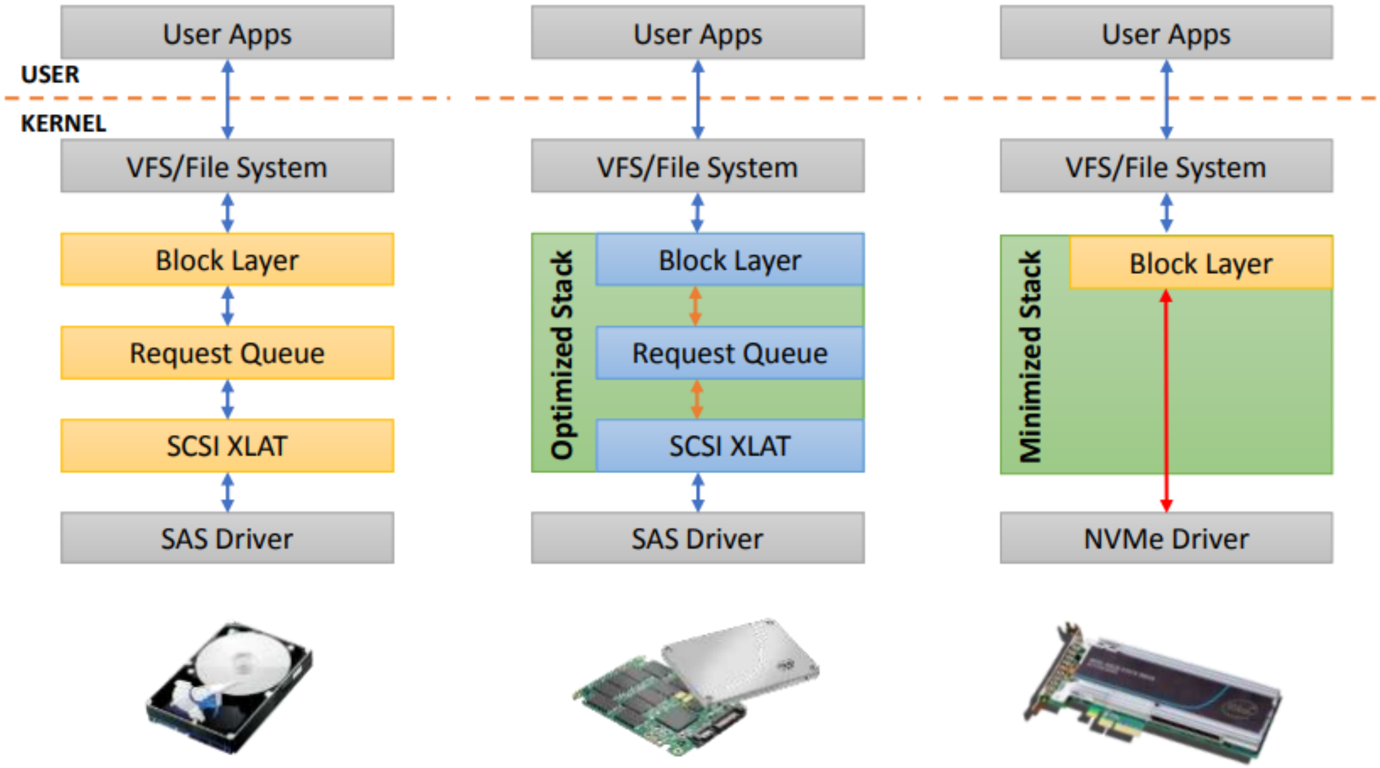
PM as Storage



PM-aware File System

Our Goal

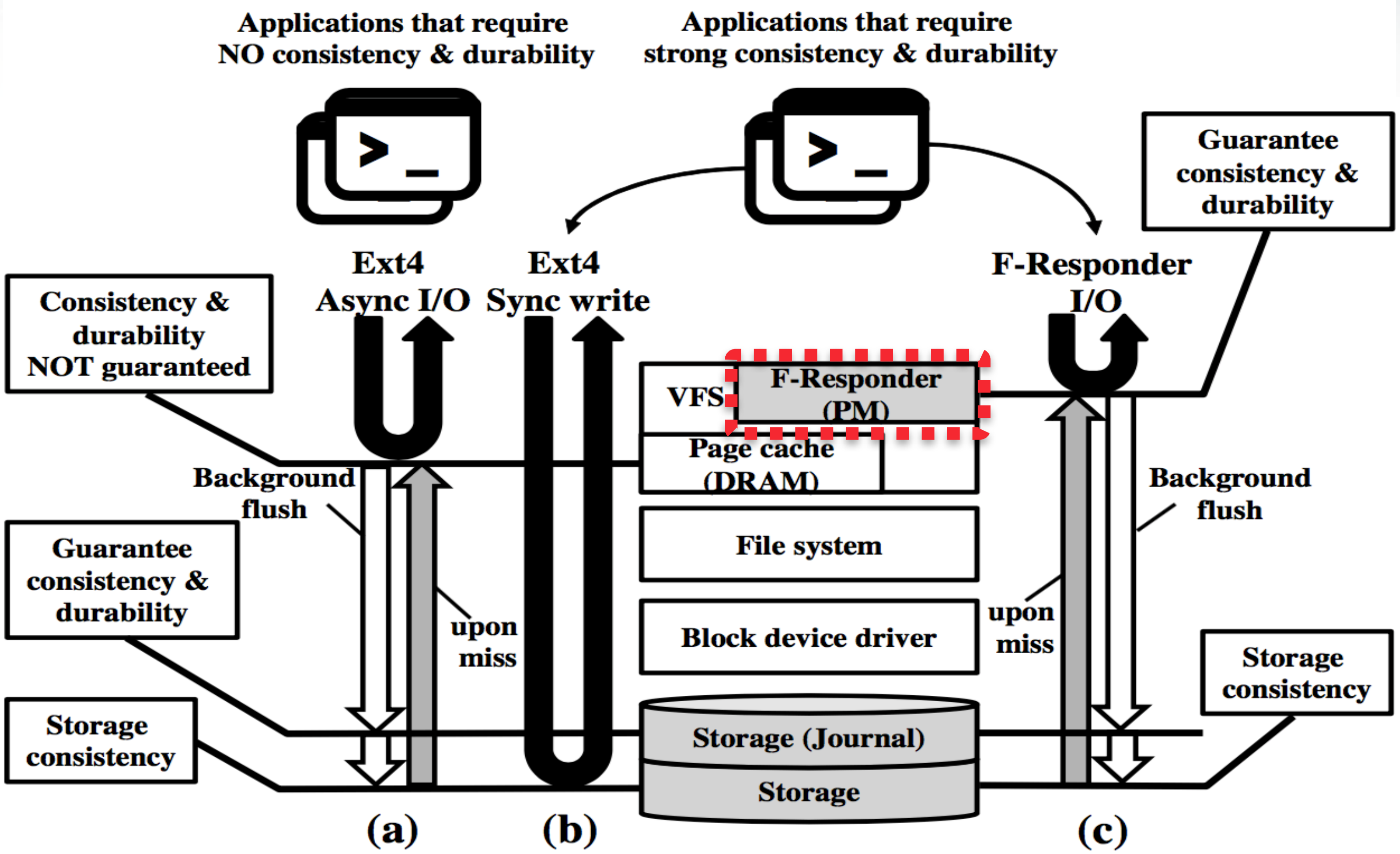
- **Keep legacy file system and storage media "as-is"**
- **Integrate PM for performance and durability/consistency**



HDD (~10ms) SAS SSD (~150µs) NVMe SSD (<100µs)

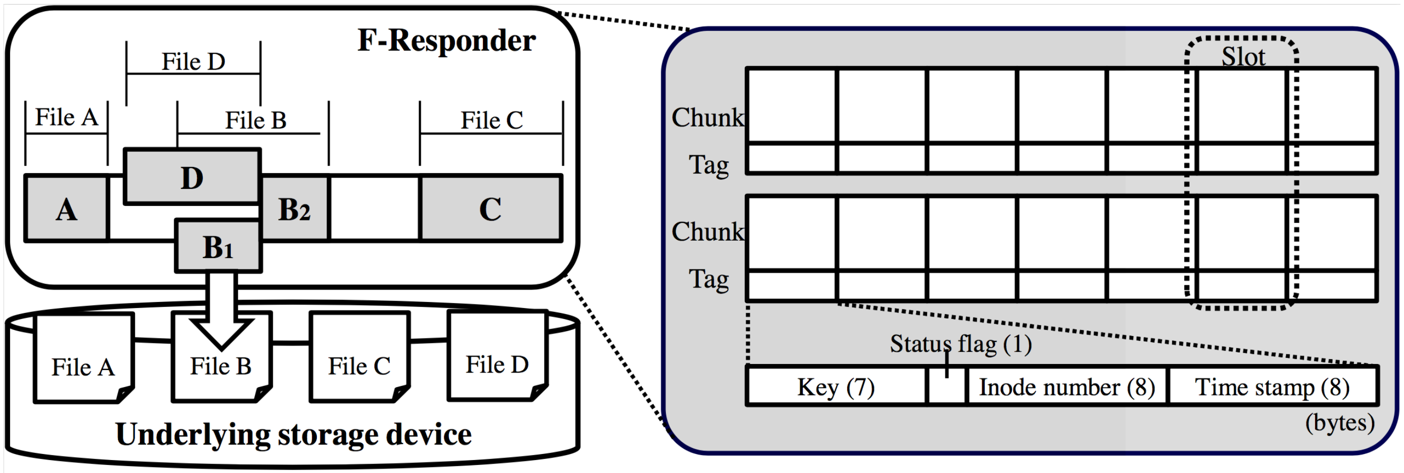
First Responder

- Overall architecture



Design

- **Static placement in "buffer cache"**
- **Sufficient large "cache"**
  - Replacement policy (almost) agnostic
- **Background flush to underlying storage device**
  - Hide storage stack overhead



(a) (b)

Performance evaluation

- **System configuration and benchmarks**

Table 1. System configuration

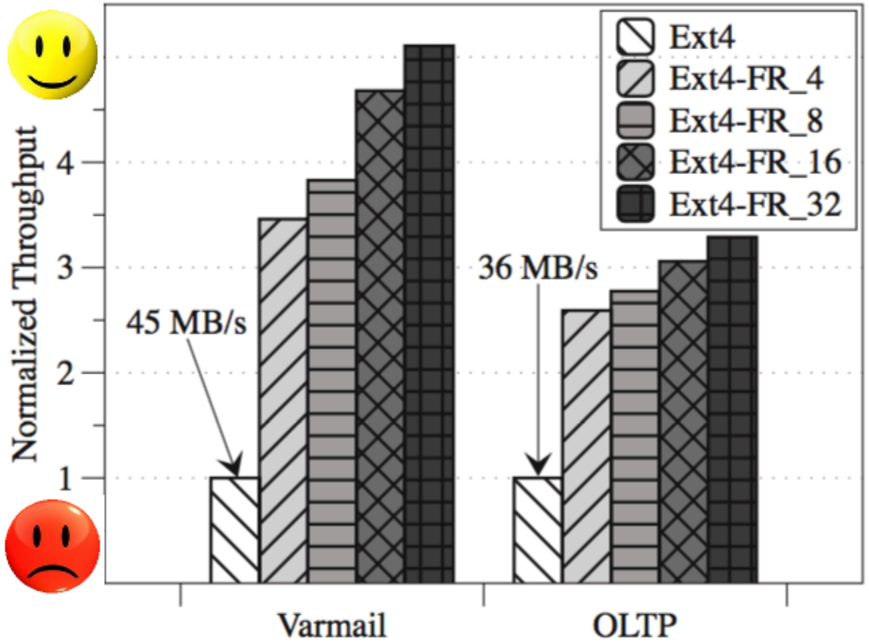
	Description
CPU	Intel Xeon E5-2620V3 (6 cores / 12 threads) × 2
Memory	Samsung DDR4 16GB PC4-17000 × 16 (256GB)
Storage	Samsung V-NAND SSD 850 PRO 256GB
OS	Linux Ubuntu 16.04 LTS (64bit) kernel v4.18

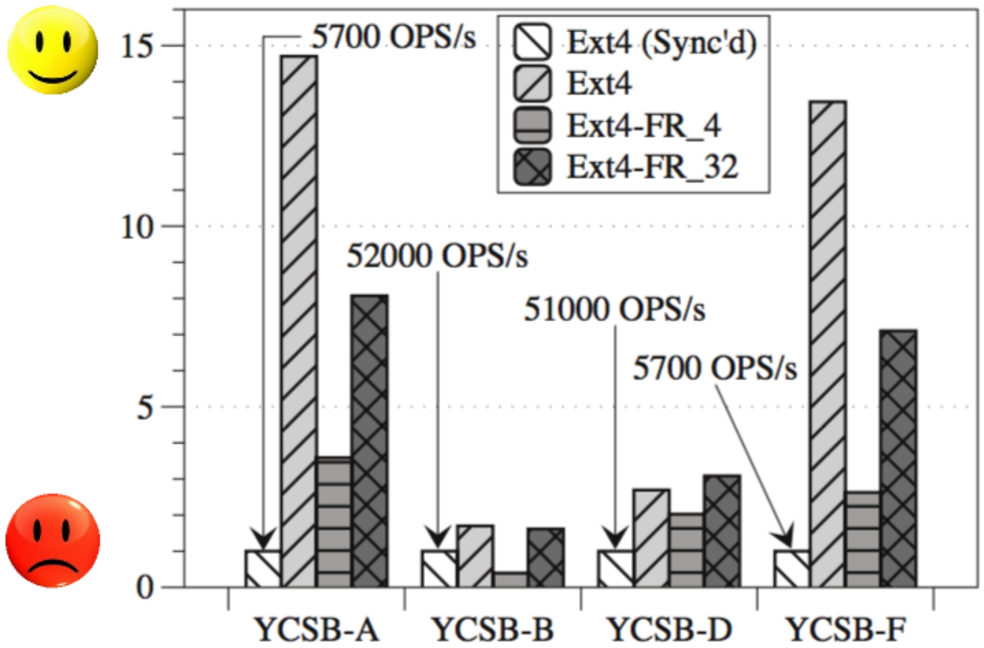
Table 2. Characteristics of benchmarks

Filebench	R:W	Mean file size	# of files	# of threads
Varmail	1:1	32KB	800K	50
OLTP	1:1	1.5GB	20	W:10 R:200
Key-value store	R:W	Record selection	Dataset size	# of threads
YCSB-A	1:1	Zipfian	12GB	20
YCSB-B	19:1	Zipfian	12GB	20
YCSB-D	19:1	Latest	12GB	20
YCSB-F	1:1	Zipfian	12GB	20

Performance evaluation

- **Overall performance**



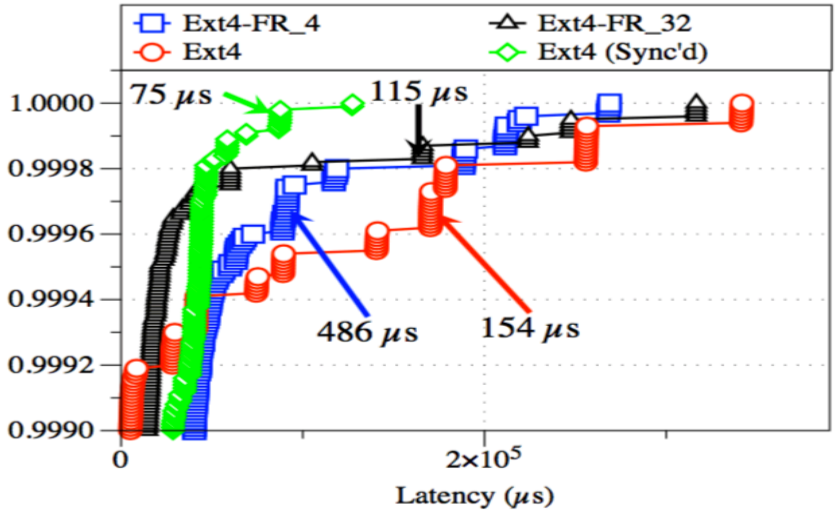


(a) Varmail and OLTP performance relative to Ext4 (async)

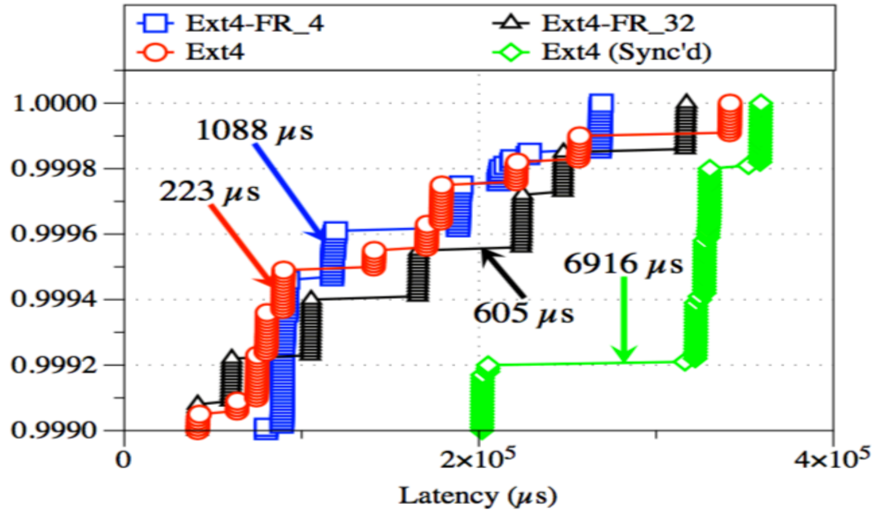
(b) YCSB (with sync mode RocksDB) performance relative to Ext4

Performance evaluation

- **YCSB-A latency results**



(a) Read

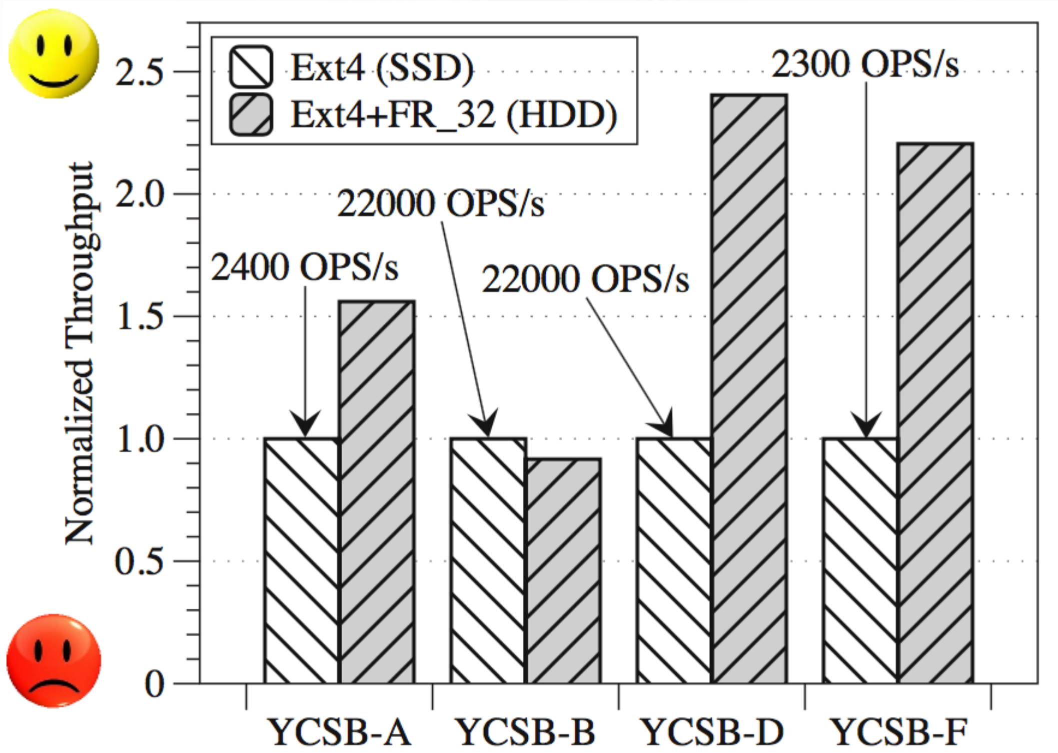


(b) Update

- **In F-Responder, consistency and durability can be guaranteed without much loss in performance**
- **Sync mode reads the average is smallest and the tail is very short**
- **F-Responder-32GB does better than sync mode on Ext4, but worse than async mode on Ext4**

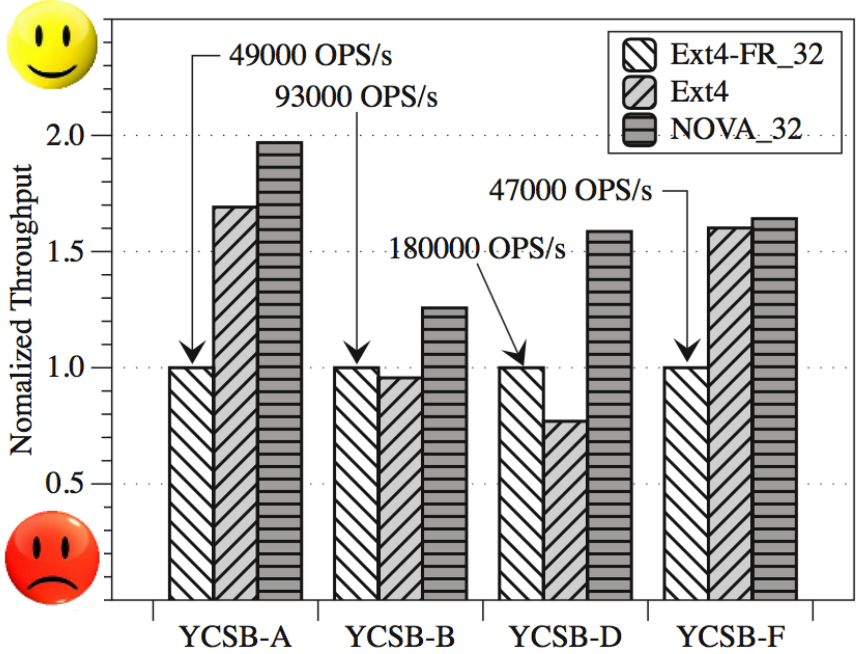
Performance evaluation

- **F-Responder with HDD**

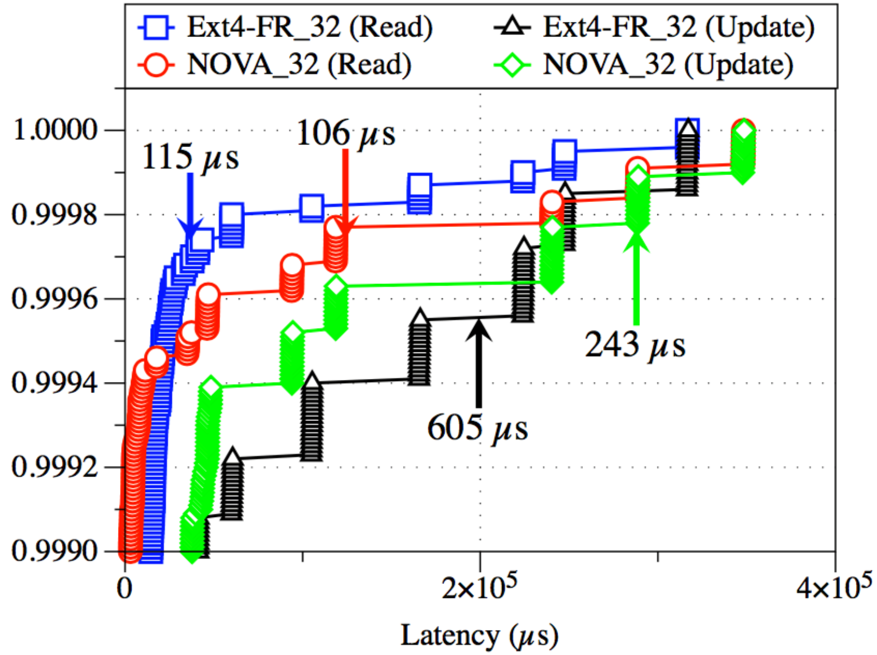


Performance evaluation

- **Comparison to NOVA-Fortis**



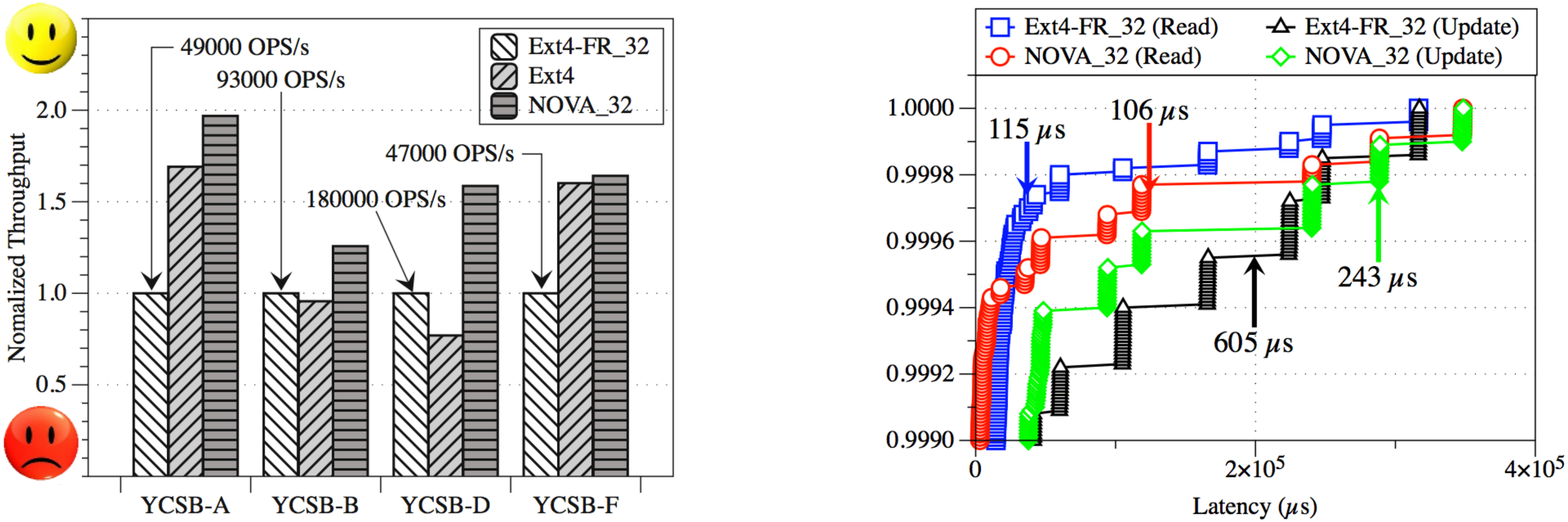
(a) YCSB Workloads



(b) YCSB-A

Performance evaluation

- **Comparison to NOVA-Fortis**



(a) YCSB Workloads (b) YCSB-A

* Issue with Linux implementation and performance reporting

- close() system call waits for background flush to complete

- even through, with F-Responder, no not need to wait

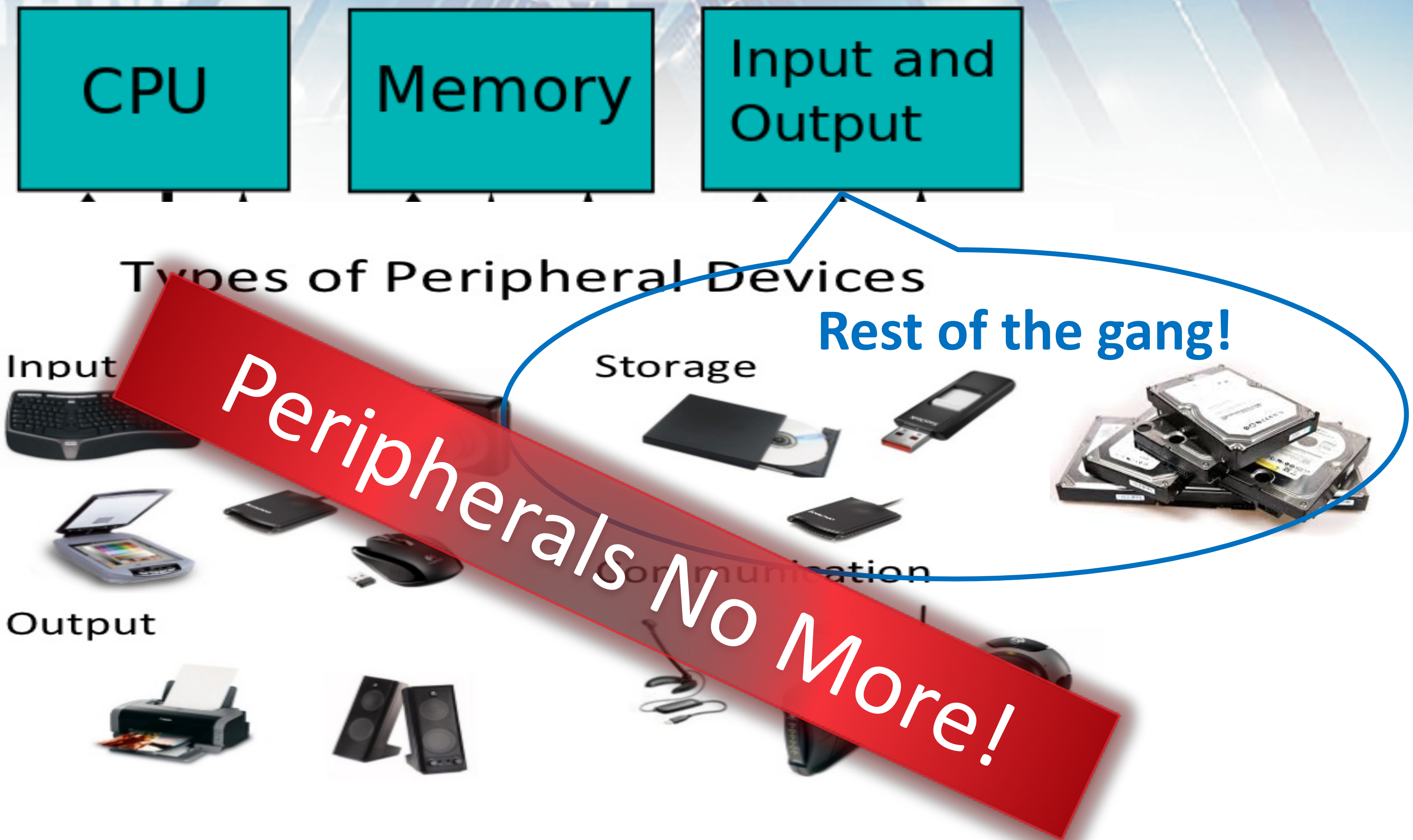
F-Responder summary

- Reap PM performance through First Responder
- Despite using legacy file system and storage
- By background trekking of critical storage stack

It's the storage stack, stupid!

Overall Summary

NOT! your father's ~~Oldsmobile~~ Storage Device



Peripheral: Auxiliary, Supplementary, relating to periphery

PAST storage topics of interest?

- **RAID**
  - Increase I/O bandwidth
- **Buffer Caching**
  - Improve latency
- **Swapping**
  - Improve resource sharing

Thank you!!!

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY

2009