Alleviating Garbage Collection Interference through Spatial Separation in All Flash Arrays

Sam H. Noh
노삼혁/盧三赫

UNIST
(Ulsan National Institute of Science & Technology)

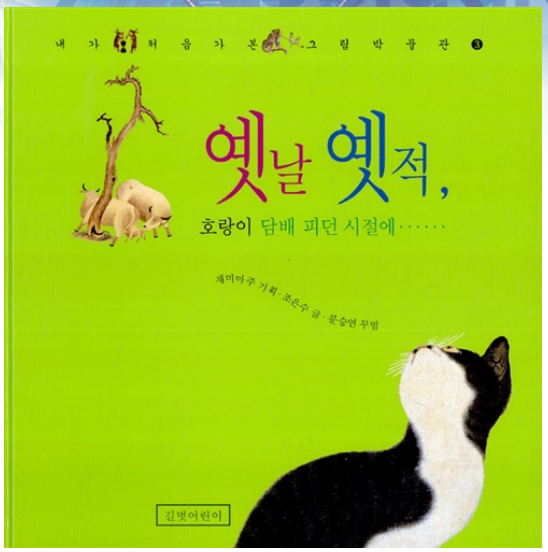
NECSST Next-generation Embedded / Computer System Software Technology

Outline

- Trend of the times
- SWAN
- Summary and Conclusion

NECSST Next-generation Embedded / Computer System Software Technology

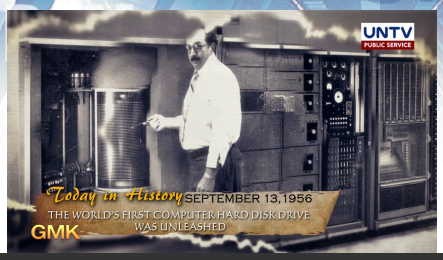
호랑이 담배 피던 시절에...

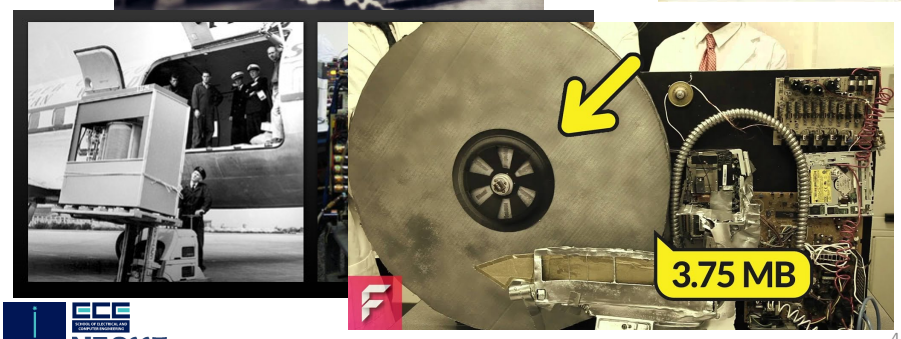


NECSST Next-generation Embedded / Computer System Software Technology

3

호랑이 담배 피던 시절에...





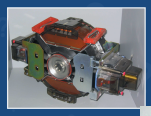
NECSST Next-generation Embedded / Computer System Software Technology

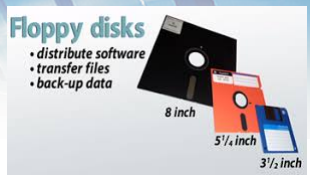
4

호랑이 담배 피던 시절에…

did you know ?

the first 1GB hard disk drive was announced in 1980 which weighed about 550 pounds, and had a price tag of $40,000?





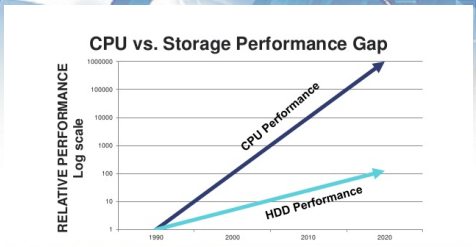
Disk data storage milestones

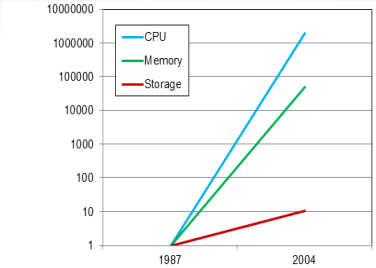
- 1971: first 8" floppy disk, IBM
- 1991: first 1GB hard disk drive, IBM
- 2000: 1 inch disk drive, IBM

NECSST Next-generation Embedded / Computer System Software Technology

5

호랑이 담배 피던 시절에…



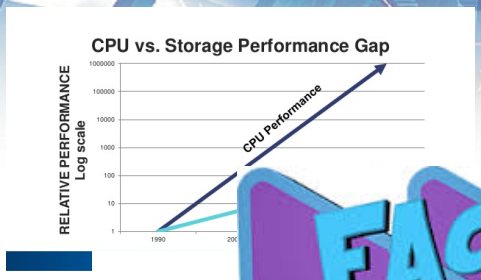


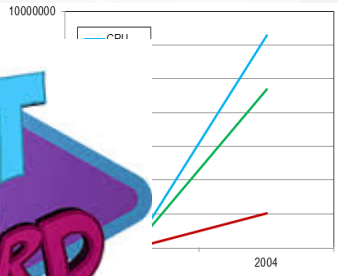
	1987	2004	Increase Multiple
CPU Performance	1 MIPS	2,000,000 MIPS	2,000,000 x
Memory Performance	100 usec	2 nsec	50,000 x
Disk Drive Performance	60 msec	5.3 msec	11x

NECSST Next-generation Embedded / Computer System Software Technology

6

호랑이 담배 피던 시절에…





		2004	Increase Multiple
Performance		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

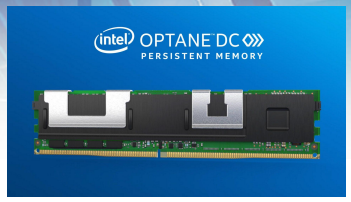
Today's New Generation of Storage Devices

- **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

Today's New Generation of Storage Devices

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



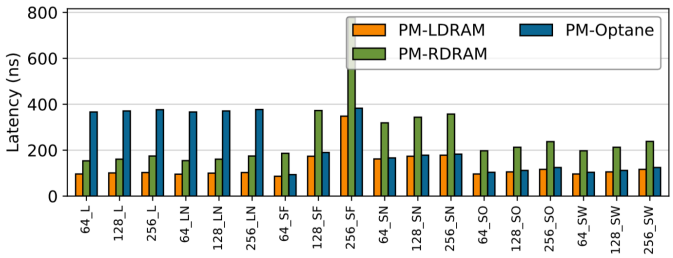


Figure 8: **Memory Instruction Latency** This graph shows the median latency for a variety of ways of accessing persistent memory. Note that for store instructions followed by flushes, there is little performance difference between PM-LDRAM and PM-3DXP, whereas the DRAM outperforms Optane DC memory for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

Courtesy of NVSL, UCSD arXiv:1903.05714v2

PAST storage topics of interest?

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

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
- **ETC**

Revisit & Rediscover
Take a fresh look at these old favorites.

Outline

- Trend of the times
- **SWAN**
- Summary and Conclusion

SWAN

It's the network, stupid!

NECSST Next-generation Embedded / Computer System Software Technology

Alleviating Garbage Collection Interference through Spatial Separation in All Flash Arrays

HotStorage '17 & ATC '19

김재호(UNIST-VT-Huawei)
김병석(UNIST), 임광현 (Cornell)
정영돈 (DGIST), 이성진 (DGIST), 민창우 (VT)

NECSST Next-generation Embedded / Computer System Software Technology

SWAN

- Background and Observations
- Design of SWAN
- Evaluation

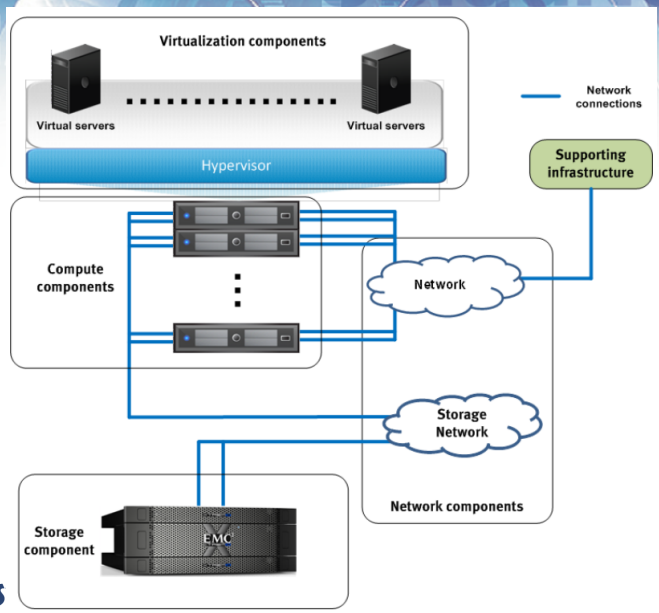
NECSST Next-generation Embedded / Computer System Software Technology

All Flash Array

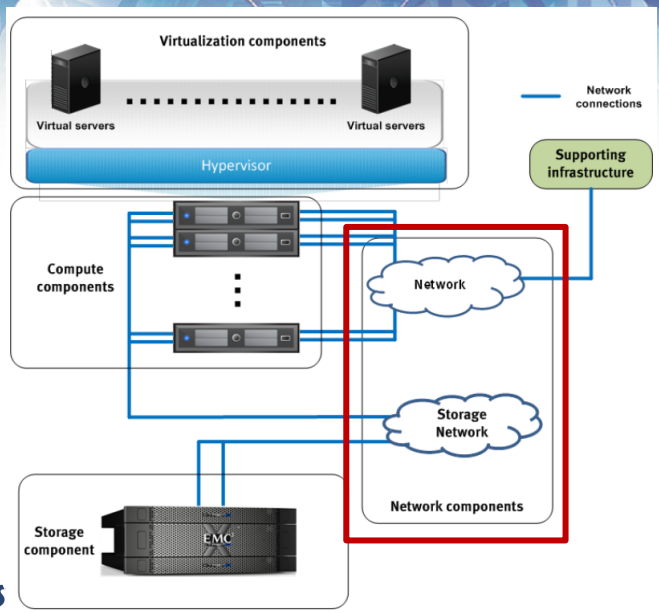
- **All Flash Array (AFA)**
  - Storage infrastructure that contains only flash memory drives
    - Solid-State Array (SSA)

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

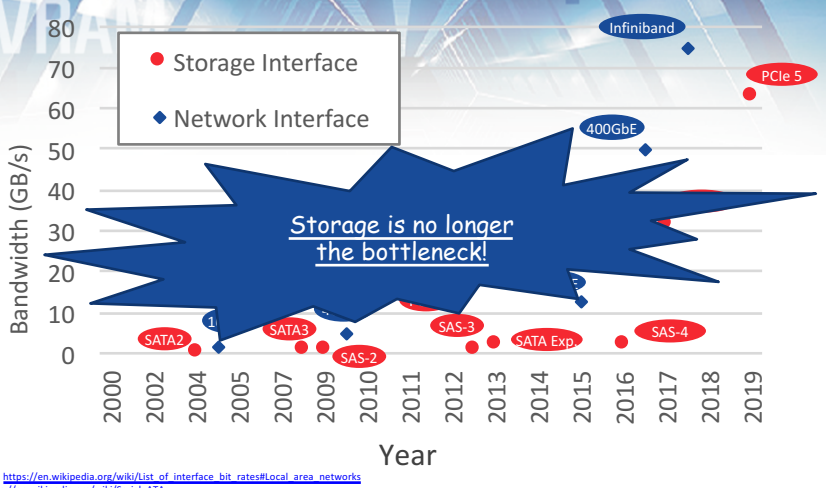


SSD Products for Data Center

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/

Bandwidth Trends



Interfaces: https://en.wikipedia.org/wiki/List_of_interface_bit_rates#Local_area_networks
SATA: https://en.wikipedia.org/wiki/Serial_ATA
PCIe: https://en.wikipedia.org/wiki/PCI_Express

NECSST Next-generation Embedded / Computer System Software Technology

21

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/

NECSST Next-generation Embedded / Computer System Software Technology

23

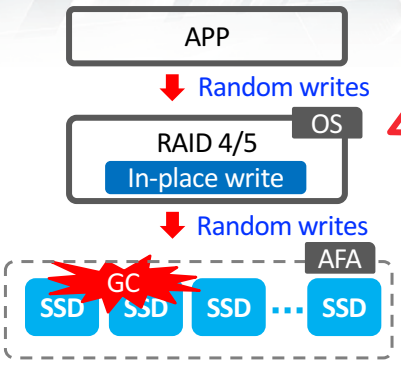
Comparison of All-flash Array

Do these many SSDs really help?
A few SSDs easily saturates network throughput!

	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/

RAID: Traditional Use of Multiple Disks



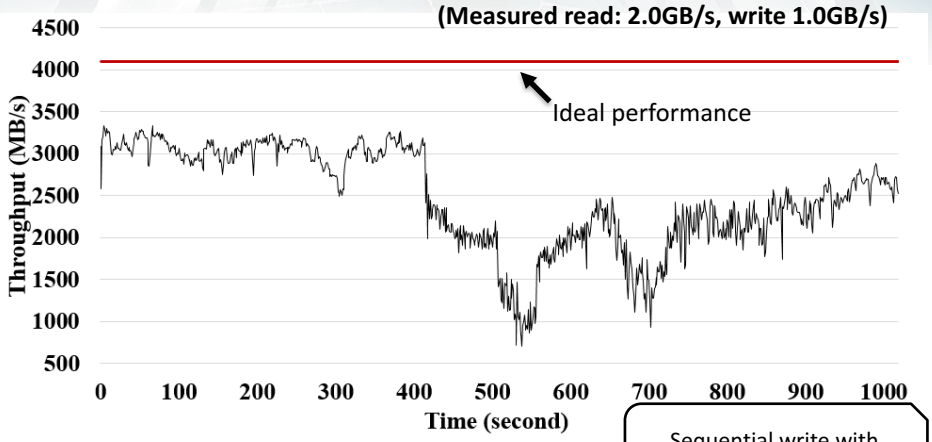
- Traditional RAID employs in-place updates to serve write requests
- High GC overhead inside SSD due to random write from the host

Limitations

- Previous solutions
  1) Harmonia [MSST'11]
  2) HPDA [TOS'12]
  3) GC-Steering [IPDPS'18]

Experiments with 4 SSD RAID 0

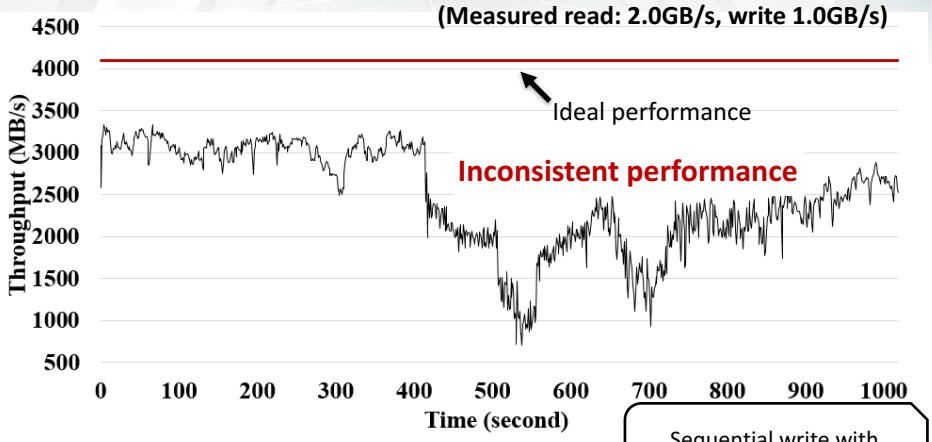
RAID 0 with 4 NVMe SSDs (spec. read: 2.4GB/s, write: 1.2GB/s)
(Measured read: 2.0GB/s, write 1.0GB/s)



Sequential write with 128KB I/O size

Experiments with 4 SSD RAID 0

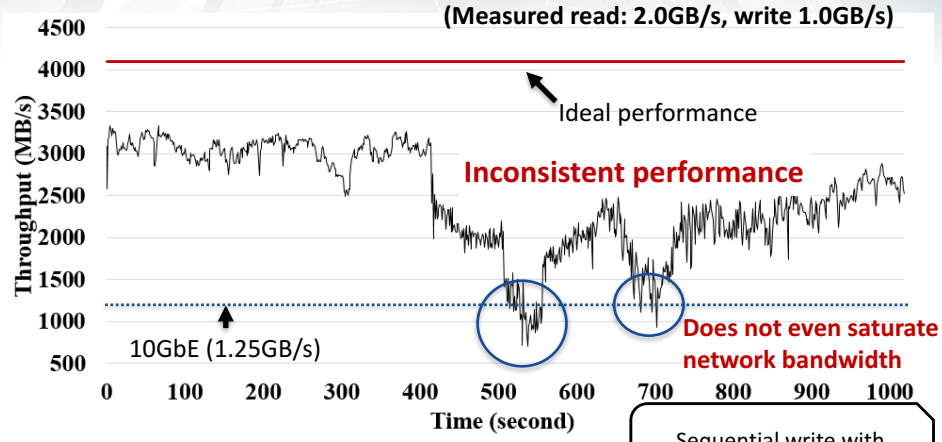
RAID 0 with 4 NVMe SSDs (spec. read: 2.4GB/s, write: 1.2GB/s)
(Measured read: 2.0GB/s, write 1.0GB/s)



Sequential write with 128KB I/O size

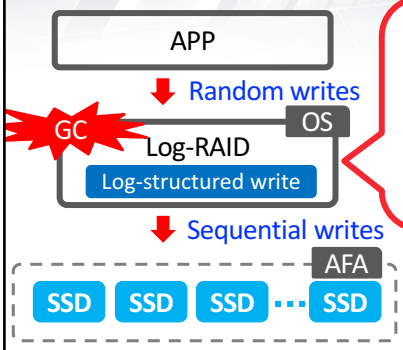
Experiments with 4 SSD RAID 0

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

Log-(based) RAID: SSD Adapted Approach



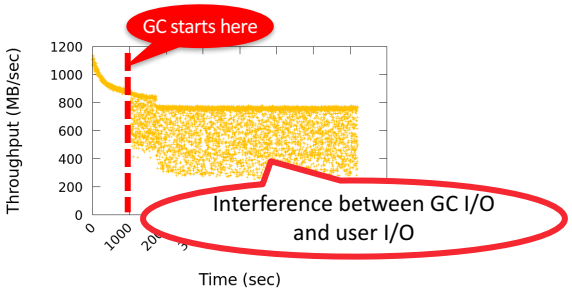
- Log-based RAID employs log-structured writes to reduce GC overhead inside SSD
- Log-structured writes involve host-level GC, which relies on idle time
- If no idle time, GC will cause performance drop

Limitations

- Previous solutions
  1) SOFA [SYSTOR'14]
  2) SRC [Middleware'15]
  3) SALSA [MASCOTS'18]

Performance of Log-based RAID

- **Configuration**
  - 8 SSDs (roughly 1TB)
- **Workload**
  - Random write requests for 2 hours



Observations

- **Inconsistent performance**
  - **due to garbage collection**
- **Performance wall**
  - **network bandwidth NOT storage**

Observations

- Inconsistent performance
  - due to garbage collection

Get rid of garbage collection!

- Performance wall
  - network bandwidth NOT storage

As best that network can support!

Our goal

Sustained, consistent
full network bandwidth performance!

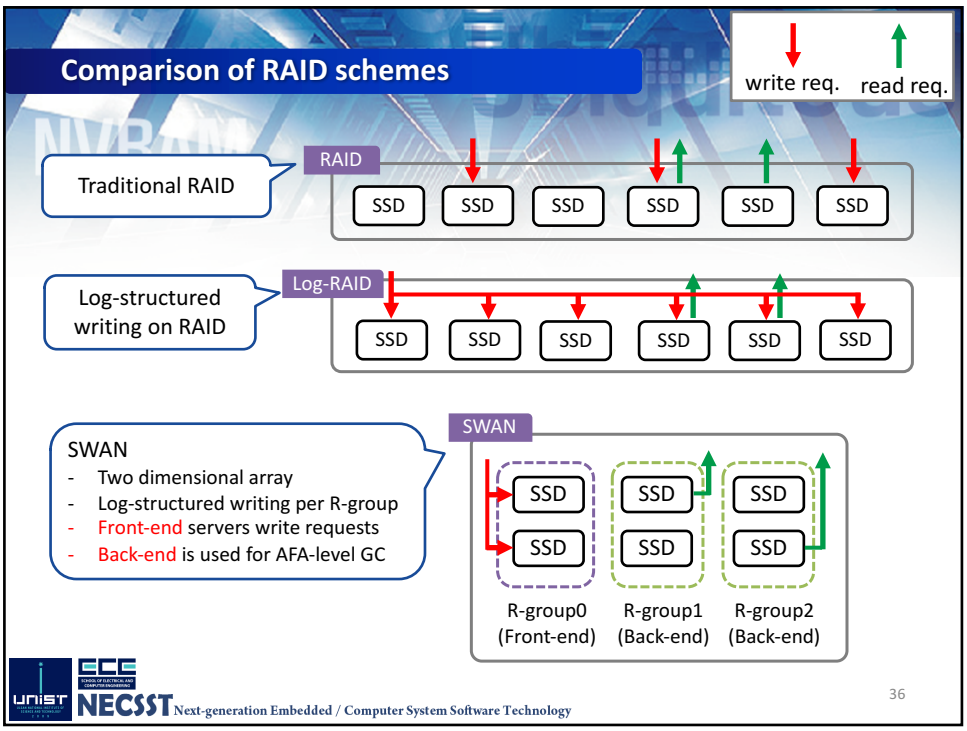
SWAN

- Background and Observations
- **Design of SWAN**
- Evaluation

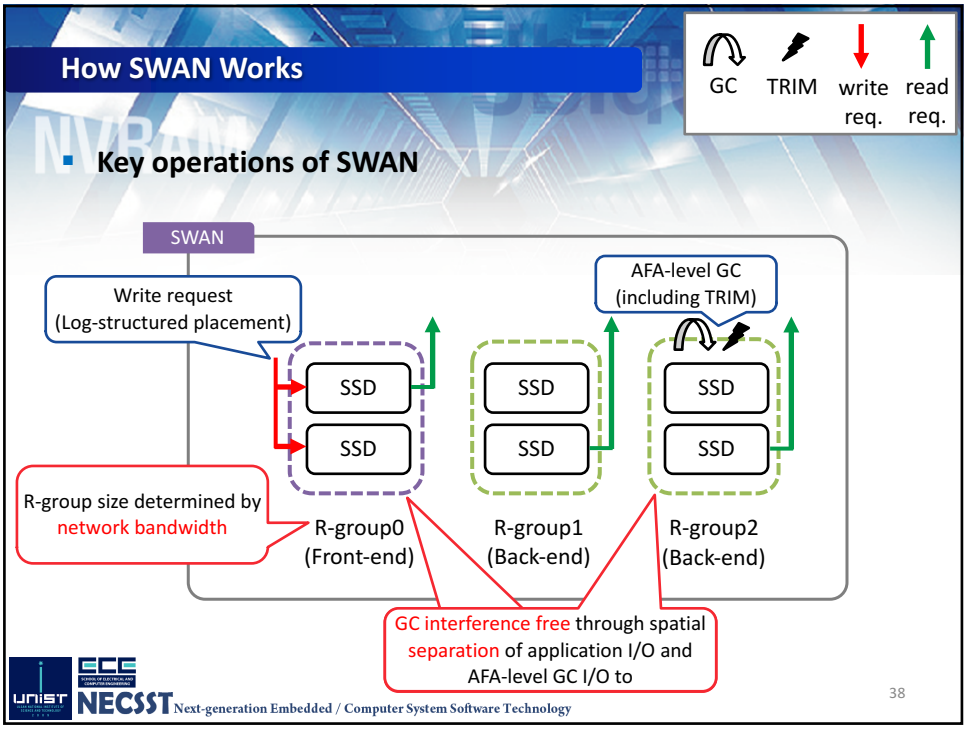
Design of SWAN

- **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

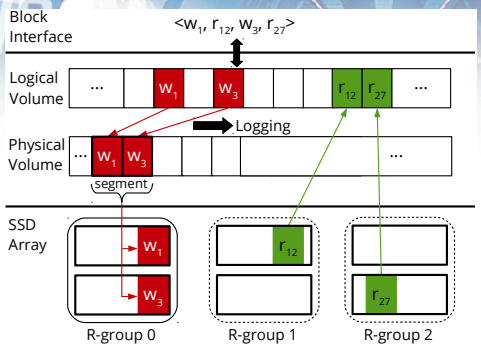


How SWAN Works



Handling Read/Write Req. in SWAN

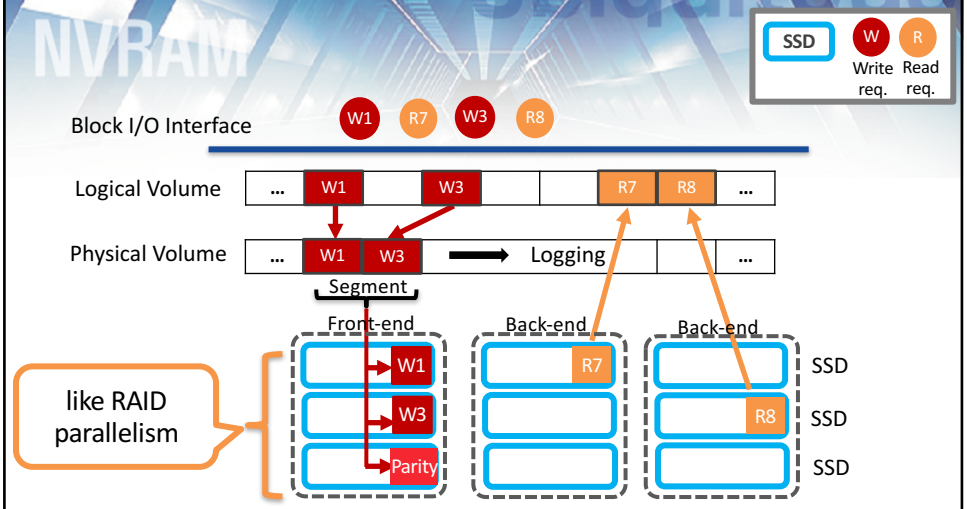
w_n : write req. for block n
r_n : read req. for block n



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

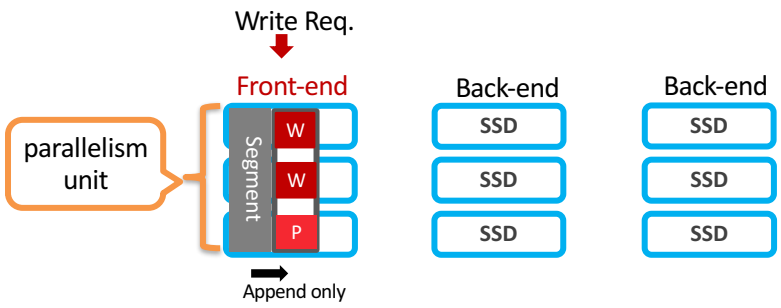
Example of Handling I/O in SWAN



Procedure of I/O Handling: Phase 1

W Write P Parity

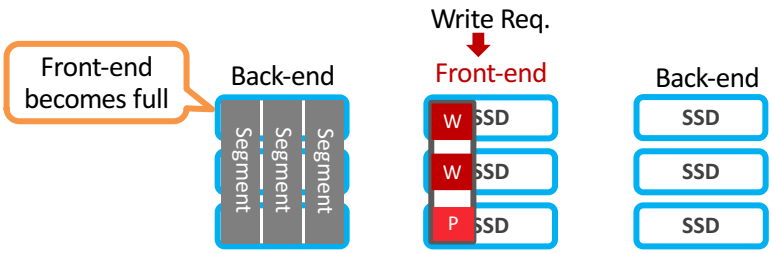
- Front-end absorbs all write requests in append-only manner
  - Exploits full performance of SSDs



Procedure of I/O Handling: Phase 2

W Write P Parity

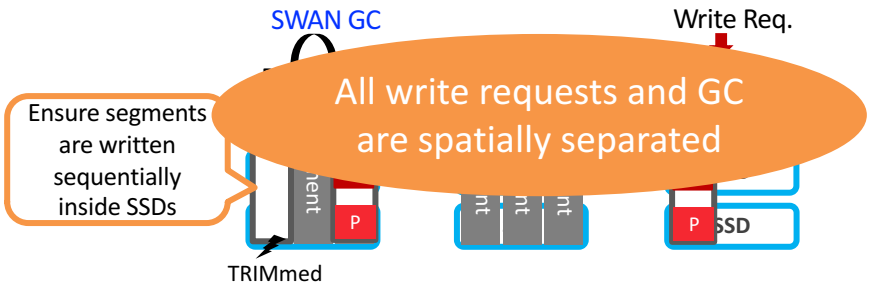
- When the front-end becomes full
  - Empty back-end becomes new front-end to serve write requests
  - Old full front-end becomes back-end
  - Again, new front-end serves write requests



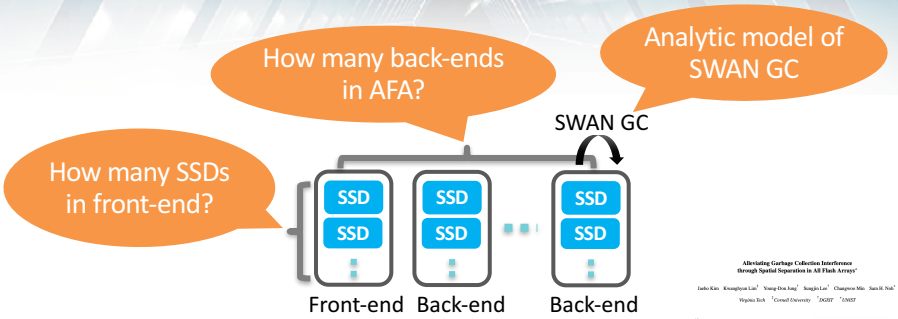
Procedure of I/O Handling: Phase 3

GC TRIM Write Parity

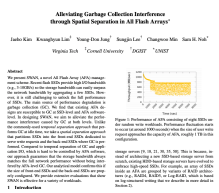
- When there is no more empty back-end
  - SWAN's GC is triggered to make free space
  - SWAN chooses a victim segment from one of the back-ends
  - SWAN writes valid blocks within the chosen back-end
  - Finally, the victim segment is trimmed



Feasibility Analysis of SWAN



Please refer to our paper for details!



SWAN

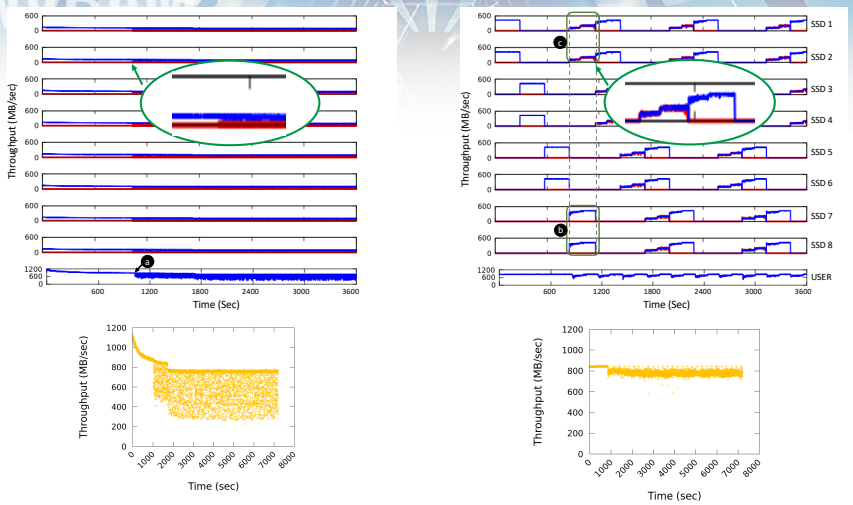
- Background and Observations
- Design of SWAN
- **Evaluation**

Evaluation

- **Environment**
  - Dell R730 server equipped with 2 Xeon CPUs and 64GB DRAM
  - Samsung 850 PRO 128GB x 9 SATA SSDs (up to 1TB capacity)
  - Open channel SSD for monitoring SSD internal activities
- **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

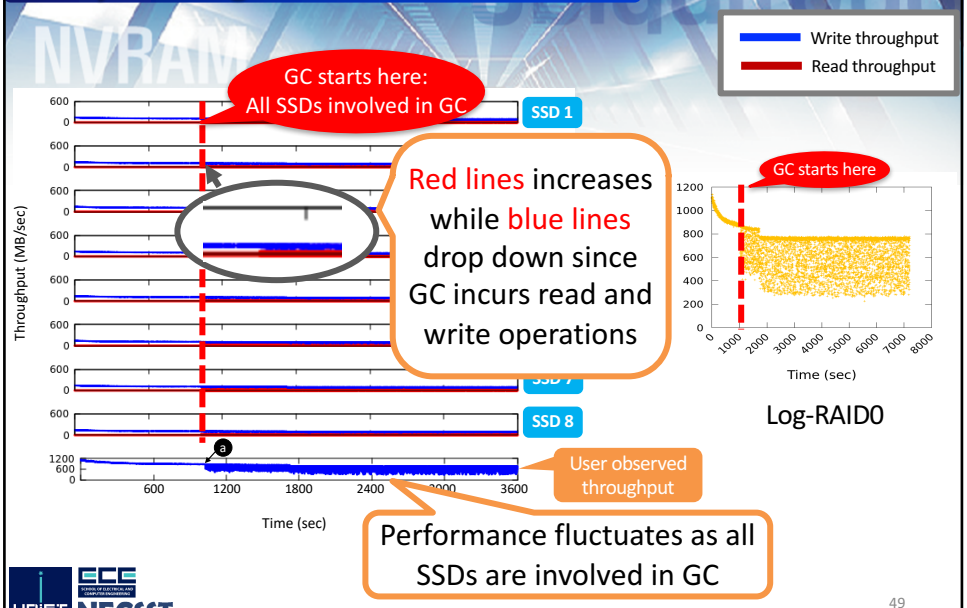
- Random write workload



Log-RAID (8 SSDs)

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

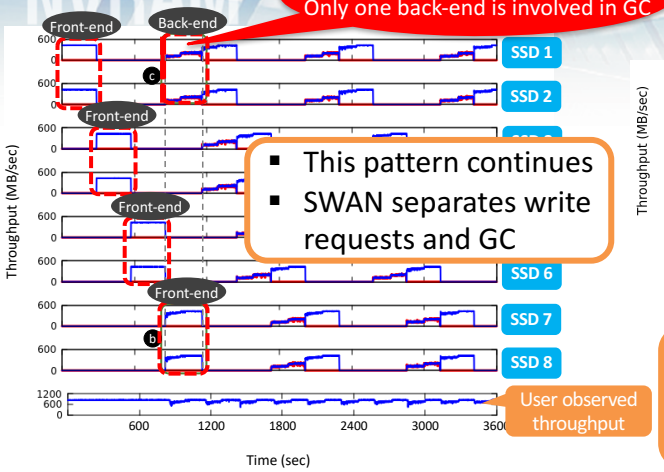
Analysis of Log-RAID Performance



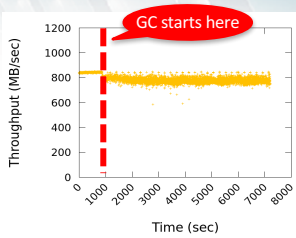
Analysis of SWAN Performance

GC starts here
Only one back-end is involved in GC

Write throughput
Read throughput



- This pattern continues
- SWAN separates write requests and GC



Configuration

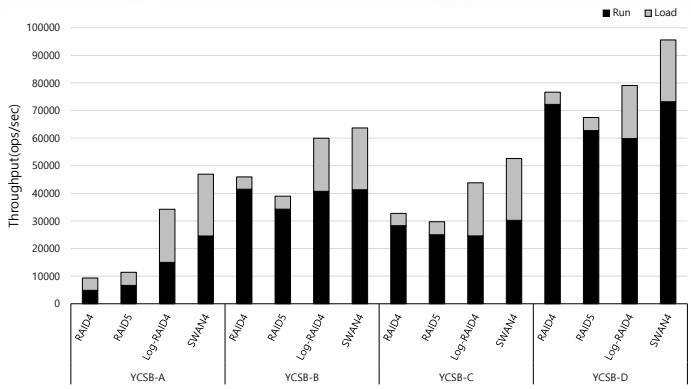
- SWAN has 1 front-end and 4 back-ends
- Front/back-ends consists of 2 SSDs

NECSST Next-generation Embedded / Computer System Software Technology

50

Throughput Results

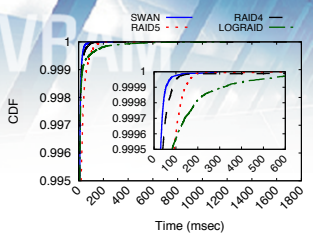
- **Configuration**
  - RAID4/5: 8 data SSDs + 1 parity SSD
  - Log-RAID: 8 data SSDs + 1 parity SSD
  - SWAN4: 3 R-group with 2 data SSDs and 1 parity SSD per R-group



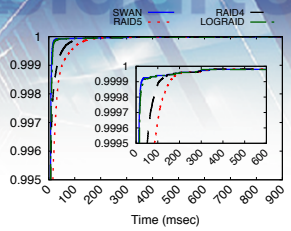
NECSST Next-generation Embedded / Computer System Software Technology

51

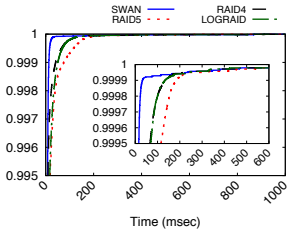
Read Latency Results (CDF)



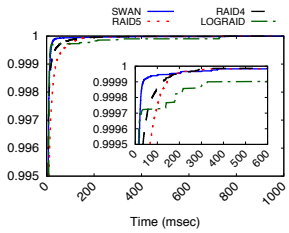
YCSB-A



YCSB-B



YCSB-C



YCSB-D

Benefits with Simpler SSDs

- SWAN can save cost and power consumption without compromising performance by adopting simpler SSDs
  1) Smaller DRAM size
  2) Smaller over-provisioning space (OPS)
  3) Block or segment level FTL instead of page-level FTL

SWAN sequentially writes data to segments and TRIMs a large chunk of data in the same segment at once

Outline

- Trend of the times
- SWAN
- **Summary and Conclusion**

Summary and Conclusion

- **Proposed SWAN**
  - New management policy for All Flash Array
- **Key idea of SWAN**
  - Performance need only be to maximum of network
  - Spatial separation
    - Decouple GC I/Os from normal ones by partitioning the SSD array into 2 groups
    - full (network bandwidth) write performance
    - "eliminate" GC effect
- **Extra benefits of SWAN**
  - SSD can be simpler

It's the network stupid!

Thank you!!!

ULSAN NATIONAL INSTITUTE OF SCIENCE AND TECHNOLOGY
2 0 0 9

NECSST Next-generation Embedded / Computer System Software Technology

56