

Seeking an alternative to tape-based custodial storage

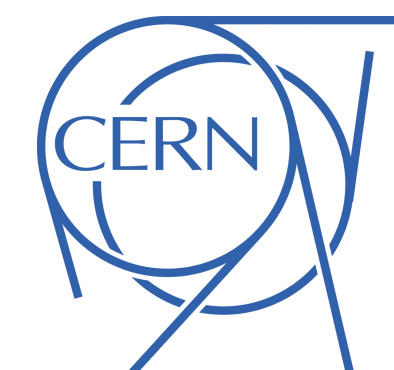
Sang Un Ahn¹, Latchezar Betev², Eric Bonfillou², Heejune Han¹, Jeongheon Kim¹, Seung Hee Lee¹, Bernd Panzer-Steindel², Andreas Joachim Peters², Heejun Yoon¹

¹KISTI, Daejeon, South Korea

²CERN, Geneva, Switzerland

*24th International Conference on Computing in High Energy and Nuclear Physics
4 - 8 November 2019*

Adelaide Convention Centre
Adelaide, Australia



Contents

- Motivation
- Technology Search
- Initial Design
- Read/Write Test Result & Power Consumption
- Schedule
- Plan




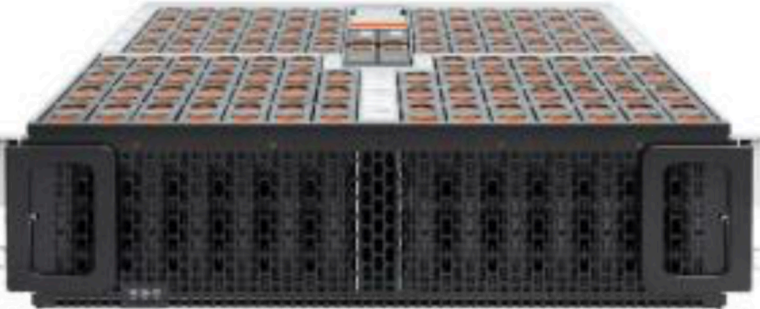


Motivation

- Why we do?
 - Shrinking market: Tape technology mono(or bi-)poly
 - ▶ One enterprise tape drive manufacturer; Two tape cartridge manufacturers
 - High cost of operating HSM for tape storage
 - ▶ Commercial licenses for Spectrum Protect (TSM) and Spectrum Scale (GPFS)
 - ▶ Expensive to update or upgrade - .5 Million USD @ KISTI
 - Tape operation requires own experts, not easy to find and train
- Goal
 - Replace the existing 3+ PB tape archive system with equally data-secure alternative
 - Use cheap off-the-shelf equipments and open-source storage solution

ATAS Project

- A proposal on seeking an alternative to tape archive system approved by WLCG Overview Board (30 Nov 2018) and endorsed by ALICE
- Expert meetings in mid of February @ KISTI and in end of March @ CERN
 - Focus on design of disk-based custodial storage system
 - ▶ Latest model JBODs with high density (up to 102 HDDs), 12Gb/s SAS HBAs
 - ▶ Storage management through EOS
 - ▶ Data protection through erasure coding RAIN
 - ▶ Project budget ~1M USD

High Density JBOD Products

Image						
Model	Dell EMC PowerVault ME484	HPE D6020	QCT JB4602 JB9T	WD Ultrastar Data102 H4102-J SE4U102-102	WD Ultrastar Data60 H4060-J SE4U60-60	Promise VTrak J5800S
Unit	5U	5U	4U	4U	4U	4U
Disk	12TB	12TB	12TB	12TB	12TB	12TB
# Disks	84	70	60	102	60	24

- Note that each JBOD enclosure has different dimensions depending on its unit and the number of disk drives to mount
- Proprietary SAS HBA cards shipped with x86 server may not provide enough compatibility to other JBOD products
- JBOD enclosures with RAID controller to provide hardware-level data protection are available in the market

State-of-the-art SAS HBA

3rd Generation

- Broadcom (Avago, LSI) SAS 9300 16(8)-port 12Gb/s SAS HBA
 - IO Controller: Two I/O controller
 - PCI Data Burst Transfer Rates: Half Duplex, 19200MB/s
 - Device support: 1024 non-RAID devices

In case of 4 ports



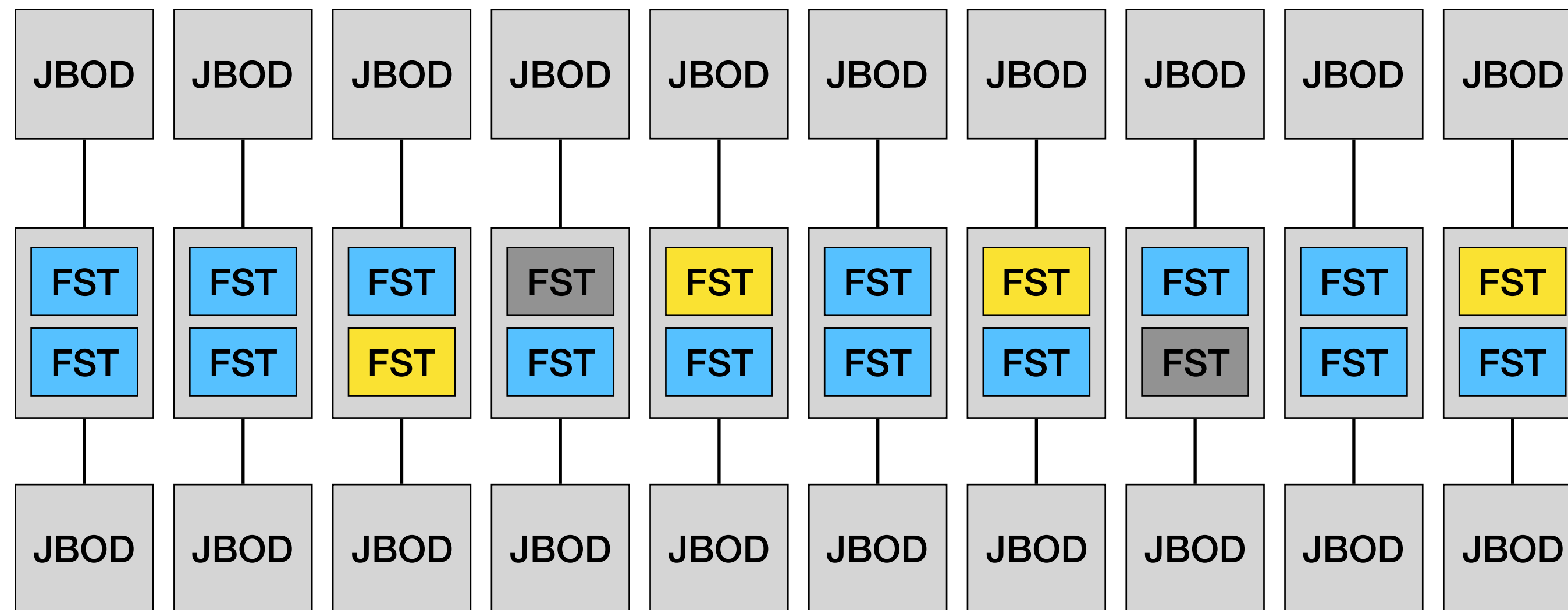
Allowing the transmission of signals in both directions but not simultaneously

Initial System Design

- 10 EOS front-end node, each hosts 2 EOS FSTs, each EOS FST serves 1 JBOD box
 - EOS EC (M, K) = (14, 4) to balance between usable space (77.7% of physical capacity) and data security
 - data loss probability ~ 0.000000005% (acceptable for ALICE)
- Each front-end node equipped with 2 SAS HBA cards (2 ports for each)
 - 1 HBA = 1 JBOD, SAS multi-path configuration to be tested for HA

← M = data node
K = parity node

EOS RAIN6 (14,4)



- (x2) EOS FSTs based on Docker container
- EOS decides where to store data fragments across FST nodes randomly (no fixed scheme)

Design Limitation Study

- In case of direct attached storage, PCIe 3.0 is the bottleneck
 - Third generation 12Gb/s SAS
 - Typical HDD transfer rate : 230MB/s for 15k, 100MB/s ~ 170MB/s for slower
 - Theoretical burst of PCIe 3.0 is about 8000MB/s while typical number is 6400MB/s (80% efficiency)

SAS Two Ports
 4 Lane each port
 1 Lane = 12Gb/s
 ∴ 48Gb/s or 4800MB/s (per port)
 Total bandwidth = 9600MB/s

Configuration	Bottleneck (MB/s)	# of HDDs	# of SSDs
6Gb/s SAS x4 / PCIe 2.x	SAS (2200)	9	4
6Gb/s SAS x8 / PCIe 2.x	PCIe (3200)	14	6
12Gb/s SAS x4 / PCIe 2.x	PCIe (3200)	14	6
12Gb/s SAS x4 / PCIe 3.0	SAS (4400)	19	8
12Gb/s SAS x8 / PCIe 3.0	PCIe (6400)	28	12

For 15k HDD (~230MB/s)
 56 slower disks can fulfill
 the bandwidth provided by
 Two port 12Gb/s SAS HBA card
 connected to a PCIe 3.0 slot

Table 4 – Sample storage configurations showing each one’s bottleneck and the number of drives supported at their peak throughput

Test Equipment & Setup

- JBOD: DELL PowerVault ME484
 - Disk: 70EA (HGST 12TB 7.2k NL-SAS), 840 TB
- Front-end Server: DELL PowerEdge R640
 - CPU: Intel Xeon Scalable 6150 2.7GHz 18 core * 2EA
 - Memory: DDR4 16GB 2666MHz * 24EA
 - HBA: DELL PowerEdge 12Gbps SAS HBA (FW version: 16.17.00.03)
 - NIC: QLogic 4x10GE QL41164HMCU CNA

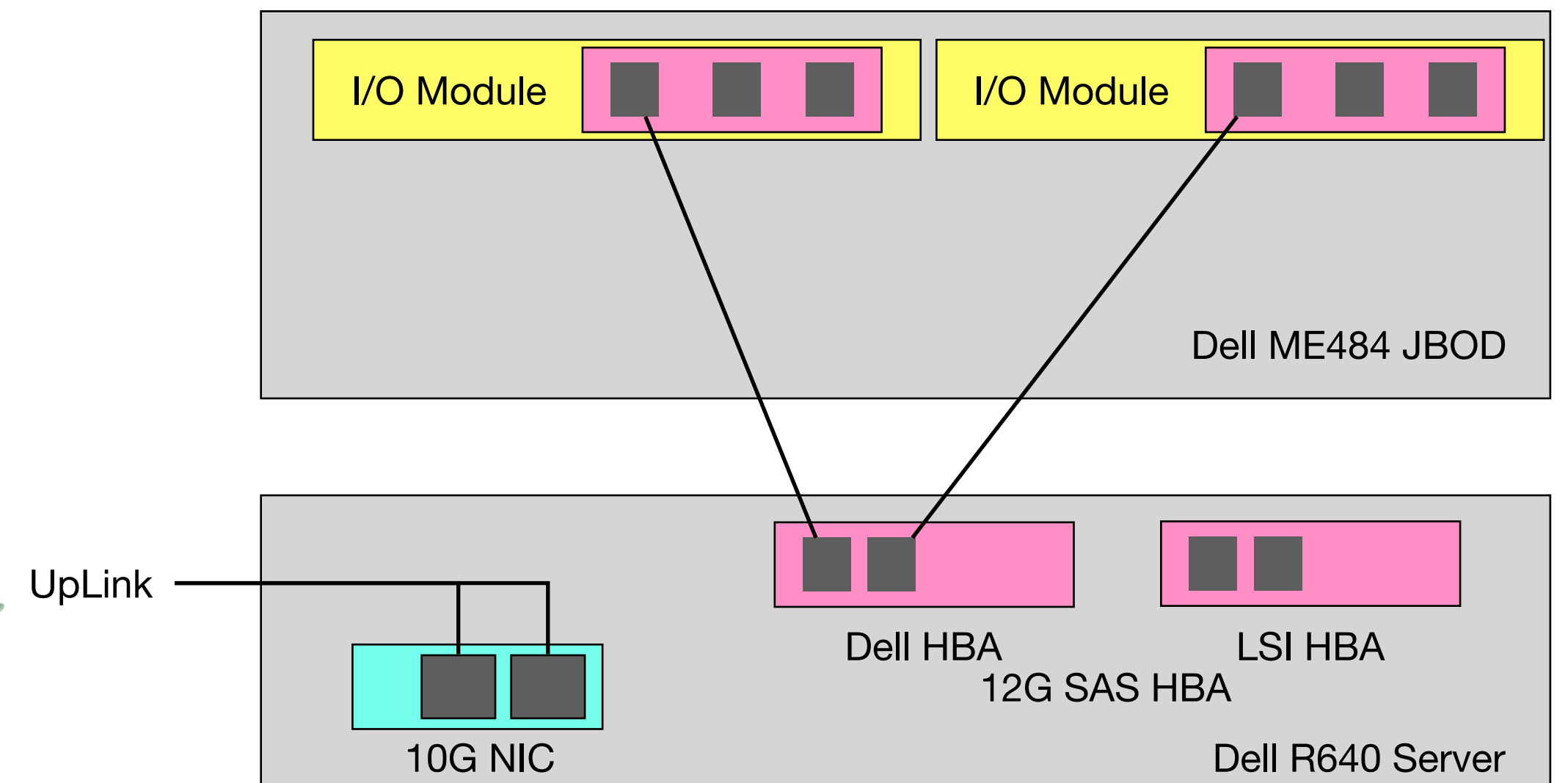
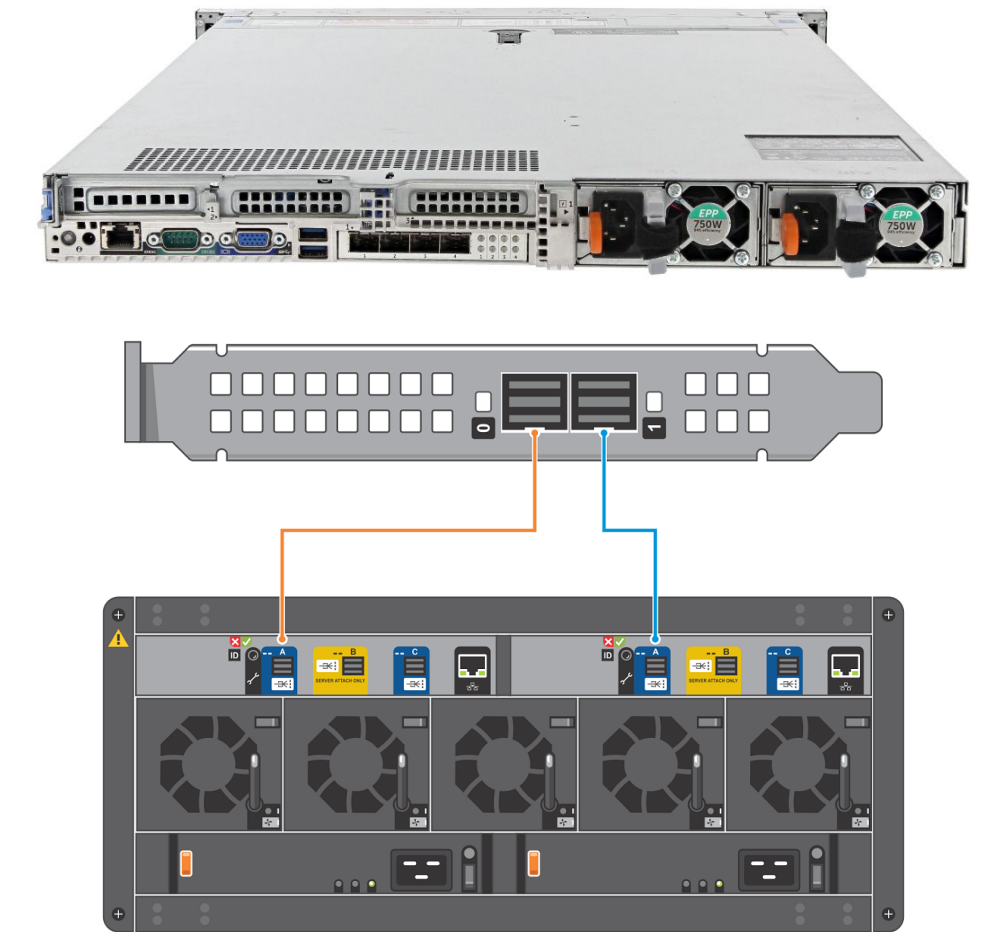
System Information

Operating System
CentOS Linux 7 (Core)

Operating System Kernel Version
7 (Core) Kernel 3.10.0.-957.el7.x86_64

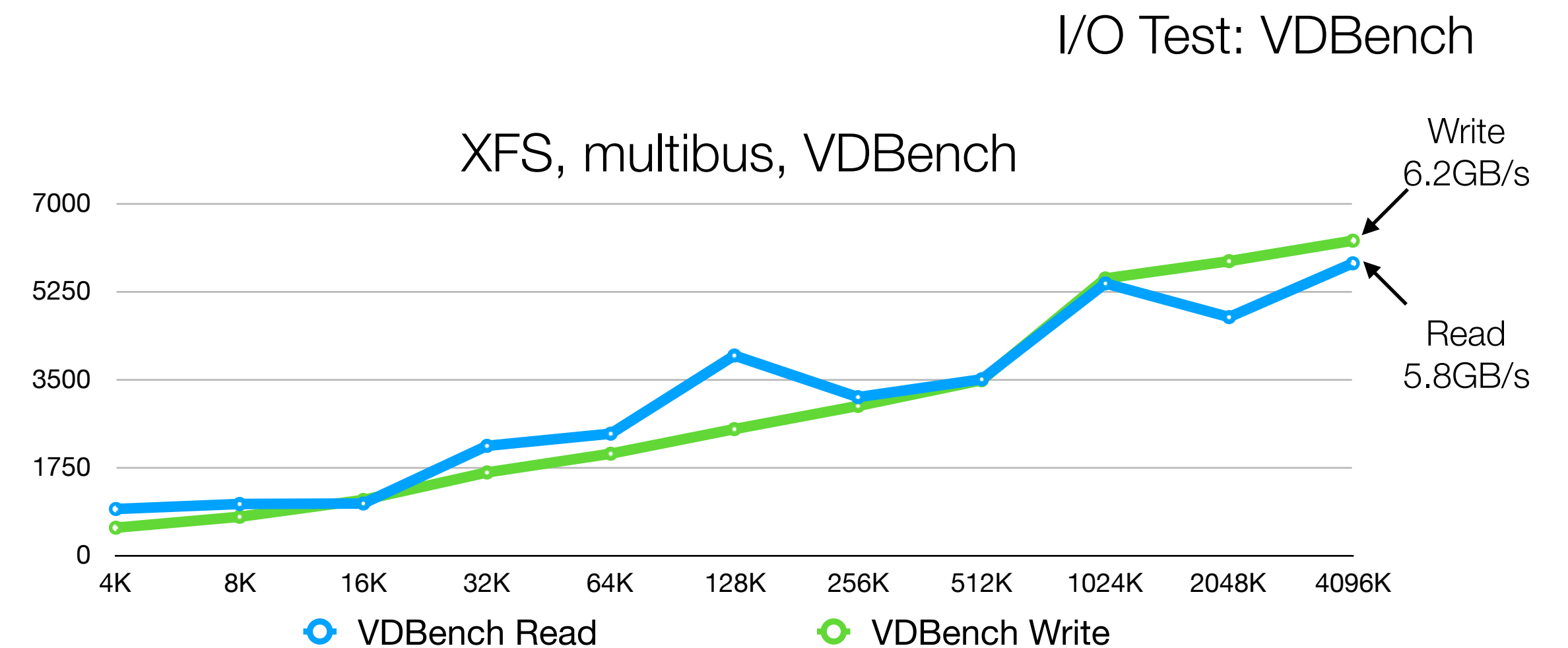
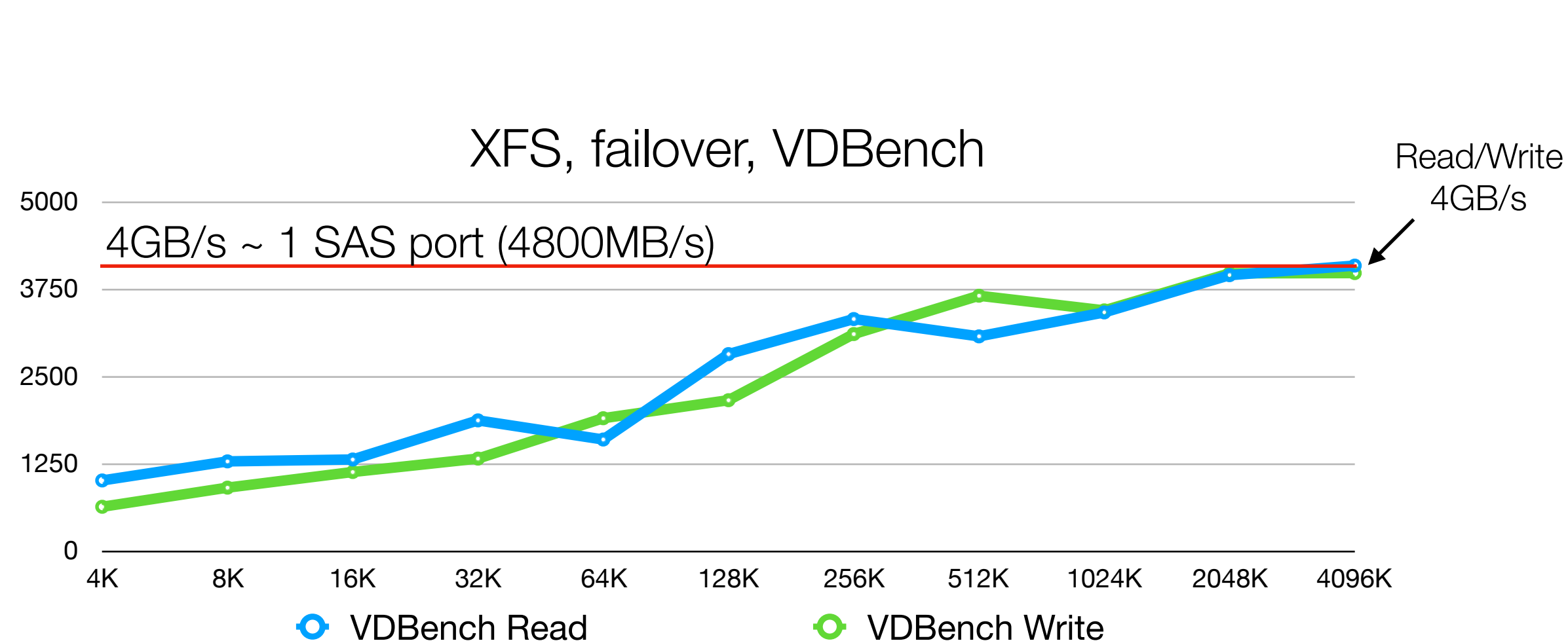
BIOS Version
1.5.6

Filesystem: XFS (Default EL7 Distribution)



I/O Test: Multipath Mode

- Multipath mode: failover (active-standby) vs. multibus (active-active)
 - **multibus** mode showed the maximum I/O speed up to 6GB/s for read/write
 - ▶ Bottleneck on PCIe 3.0 (6400MB/s)
 - **failover** could not fulfill the available bandwidth, limited under 1 SAS port (48Gb) pipe

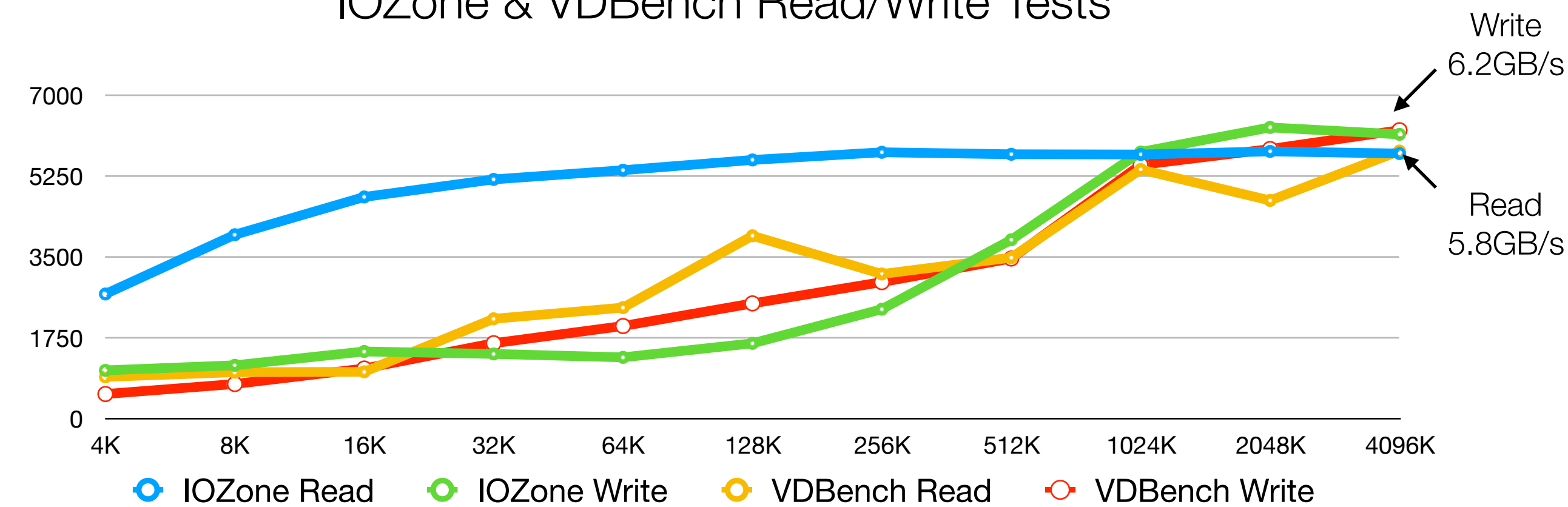


I/O Test: Read/Write

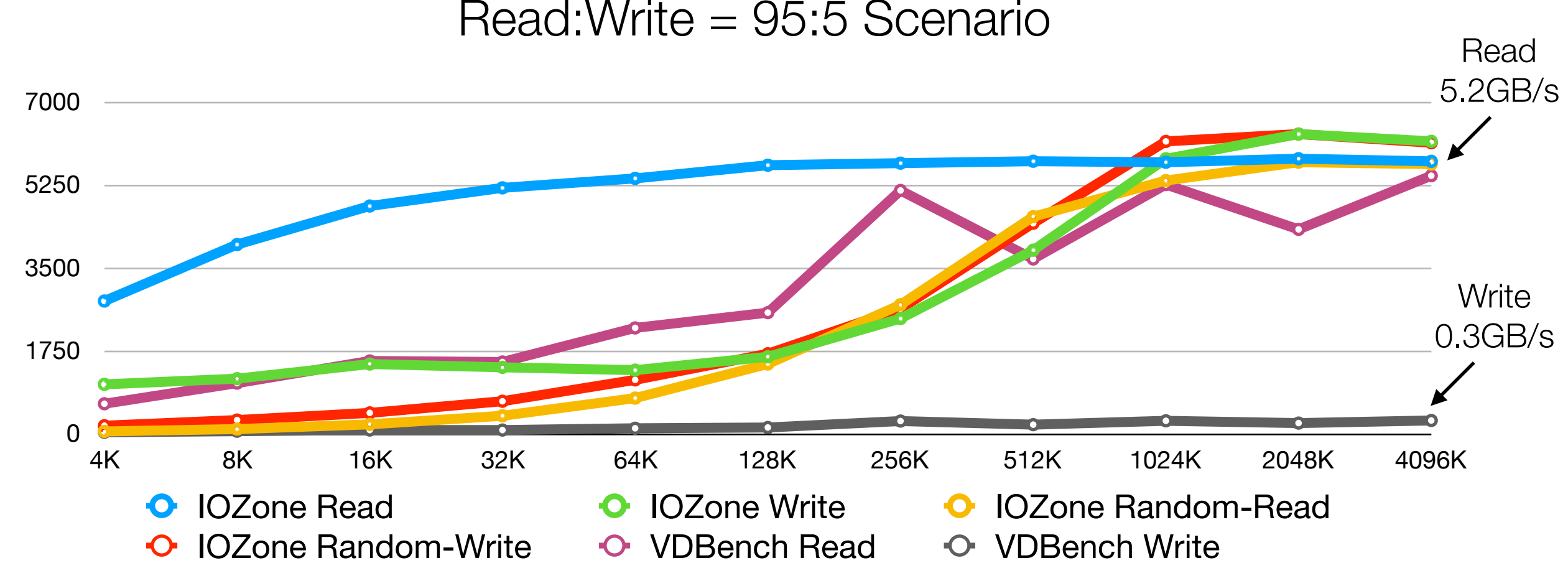
- XFS read/write performance (simultaneous read and/or write from 70 disks)
 - **VDBench** shows full read/write transfer performance @ transfer size $\geq 2048k$ (6GB/s)
 - **IOZone** shows full read/write transfer performance @ transfer size $\sim 2048k$ (6GB/s)

Disk: 70EA
Filesize: 2GB

IOZone & VDBench Read/Write Tests



Read:Write = 95:5 Scenario



* IOZone tests with different Read/Write ratio Scenario did not much affect on the performance

Power Consumption

- JBOD Test Equipment (70 Disks)

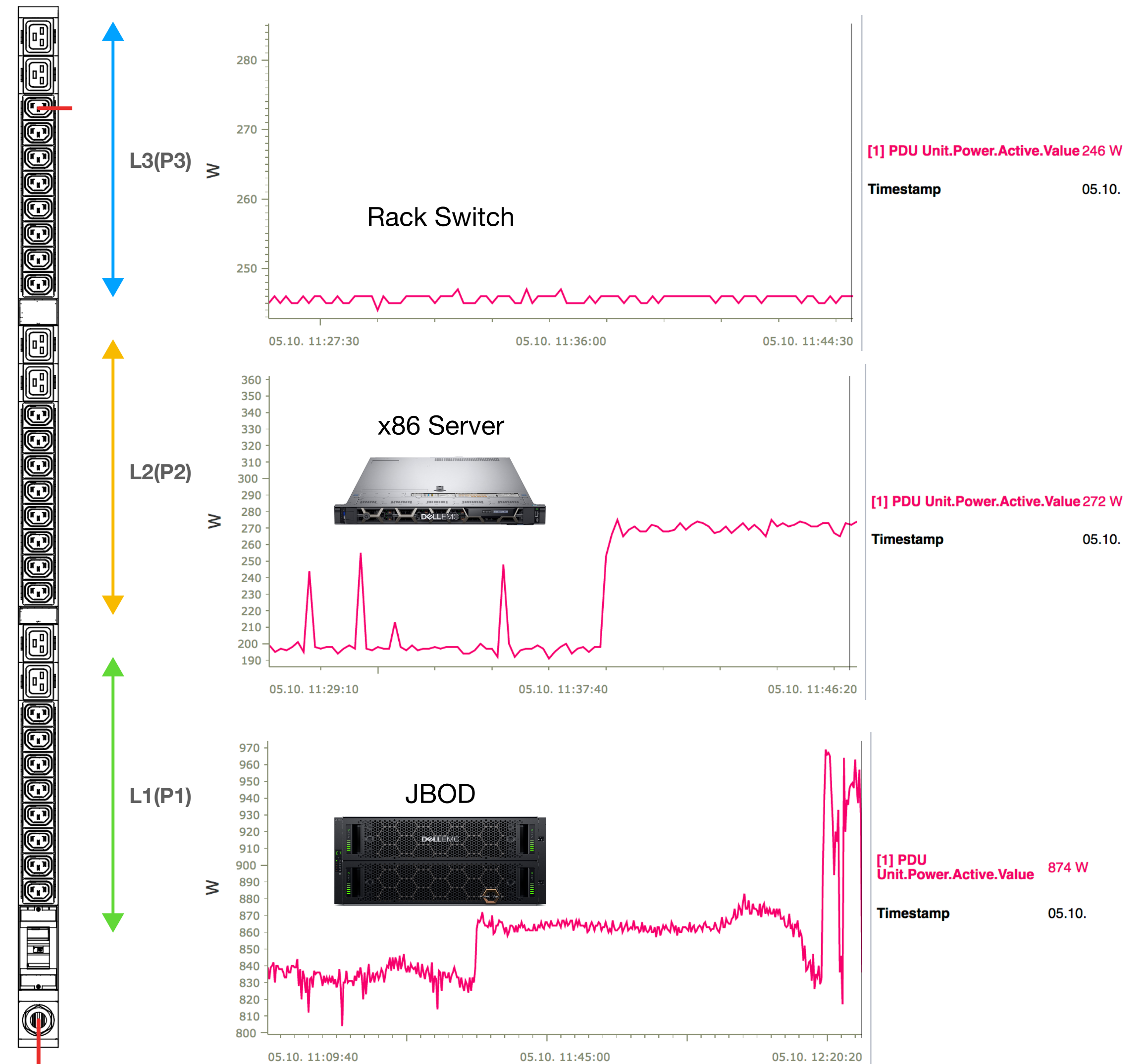
- JBOD (DELL ME484): idle = 830W; load = 860W (Max 960) **(1.12W/TB)**
- Server: idle = 200W; load = 270W
- Switch: idle = 246W; load = 246W
- **1.75W/TB** including JBOD, Server and Switch

- Disk Storages (Full Load)

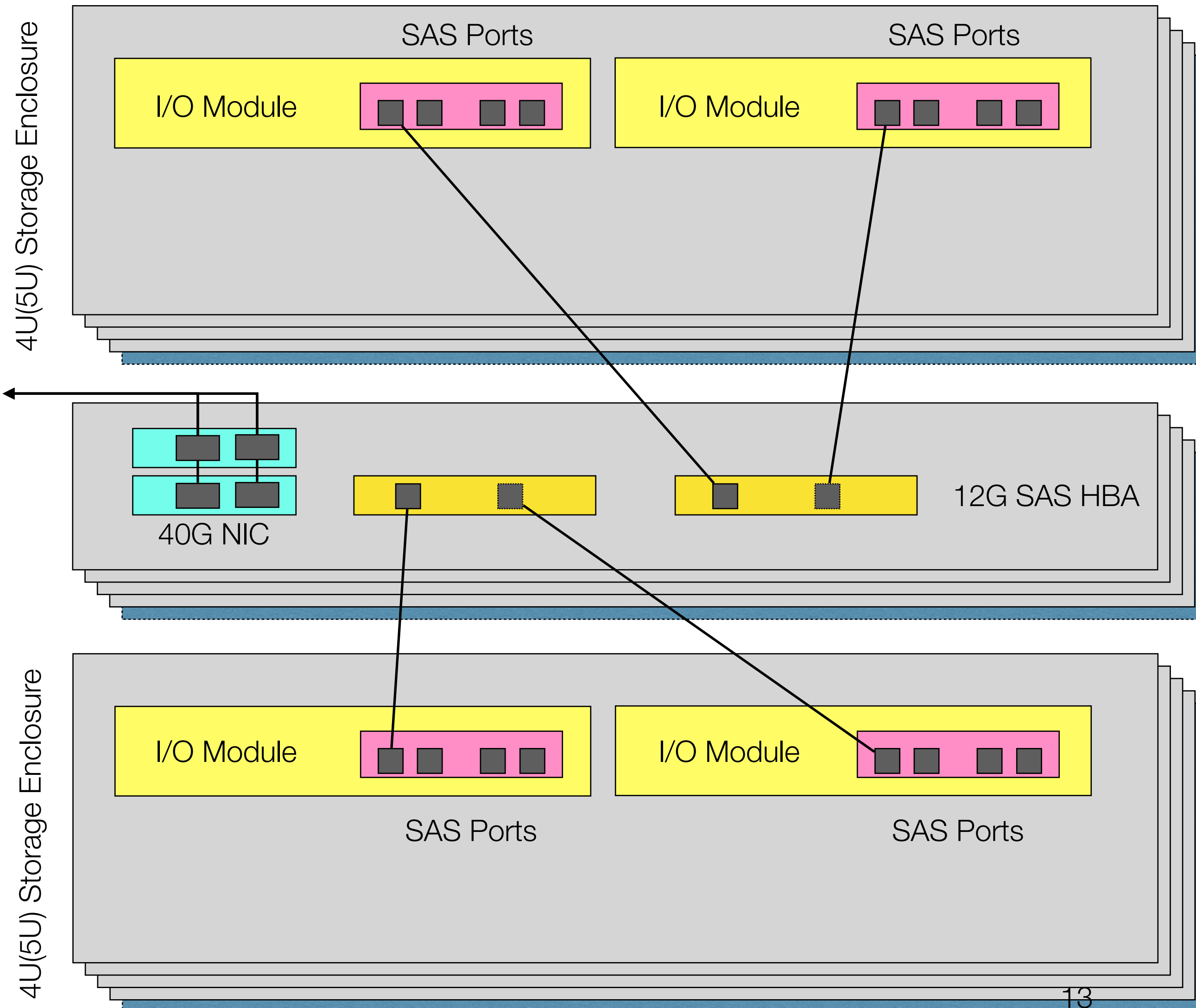
- DellEMC SC7020, 2.5PB - 12,120W **(4.8W/TB)**
- EMC Isilon, 16 Nodes, 2.95 PB- 13,730W **(4.6W/TB)**
- EMC VNX, 12 Nodes, 2.36 PB - 5,100W **(2.2W/TB)**
- HITACHI VSP, 2 PB - 18,300W **(9.15W/TB)**
- EMC Isilon, 15 Nodes, 1.43 PB - 12,880W **(9W/TB)**
- EMC CX4-960, 1.5PB - 14,900W **(9.9W/TB)**

- Tape Library (Full Load)

- **IBM TS3500 5-Frame (3.2PB) - 1,600W (0.5W/TB)**



Deployment Setup



- This is a setup similar in all aspects to the CERN EOS current/future deployment

x10 EOS Nodes hosting 2 EOS FSTs for each

Specifications

- x10 2U x86 servers
 - x2 40G NICs, x2(x4) 12G SAS HBA cards
- x2 40G network switches
- Even number of JBOD boxes filled up to 18PB

Schedule

Tasks	1	2	3	4	5	6	7	8	9	10	11	12
Technology Search	█											
Product Survey		← KISTI-CERN Expert Meeting @ KISTI										
Architecture Design and Specification			← KISTI-CERN Expert Meeting @ CERN									
Testing					█	█	█					
Procurement					█	█	█	█	█	█	█	█
Implementation											█	█
Validation											█	█

Call for tender (delayed) →

Delivery

Today ↑

- Change of procurement planning had been approved in May by National Facility & Equipment Committee
- Call for tender delayed due to change of procurement procedure (technical pre-estimation included)
- Tested QCT Intel x86 server - IPMI management, Remote KVM, BIOS etc.
- Tested DELL ME484 (likely a production system choice) - 12G SAS performance, Power consumption, Management

Conclusions

- We are investigating a disk-based storage, using standard JBODs and EOS with erasure coding, as an alternative to tape-based custodial storage
- Obvious benefits: avoid single-vendor dependency, common expertise for all storage systems across the computing centre
- A final system unit I/O tests show ~6GB/s read/write performance, as expected from the limits of the PCIe 3.0 and SAS 12Gb/s HBA
 - Much higher than the current tape library
- Power consumption is shown to be 1.75W/TB, not uncomfortably higher than a tape library
- Procurement has finished and delivery of systems is expected in November, EOS deployment with RAIN configuration will be started as early as in December
- In 2020, the disk-based custodial storage will be tested and verified with ALICE
- Upon the succeed of implementation of JBOD based archive storage, we may apply it for different level of services (QoS)
 - E.g. lesser data protection with larger capacity : ordinary disk storage

Questions?



Backup

Abstract

- In November 2018, the KISTI Tier-1 centre started a project to design, develop and deploy a disk-based custodial storage with error rate and reliability compatible with a tape-based storage.
- This project has been conducted in the collaboration between KISTI and CERN, especially the initial system design was laid out from the intensive discussion with CERN IT and ALICE.
- The initial system design of the disk-based custodial storage accommodated high density JBOD enclosures and the erasure coding implemented in EOS, the open-source storage management developed at CERN.
- In order to balance among system reliability, data security and I/O performance, we investigated the possible SAS connections of JBOD enclosures to the front-end node managed by EOS and the technology constraints of interconnections in terms of throughput to deal with the large number of disks.
- This project targets to have a production system before the start of LHC RUN3 in 2021.
- This year we will procure and deploy the disk-based custodial storage with the hardware specification derived from the initial system design.
- In this paper we present the detailed description on the initial system design, the brief results of test equipments for the procurement, the deployment of the system and the further plan of the project.

Concerns about Tape Market

- One enterprise tape drive manufacturer, two tape cartridge manufacturers
- Oracle enterprise tape drive
 - https://www.theregister.co.uk/2017/02/17/oracle_streamline_tape_library_future/
- Concerning steady tape cartridge supply, tape suppliers shrunk over the past three years from six to two - Sony, Fujifilm
 - <https://www.bloomberg.com/news/articles/2018-10-17/the-future-of-the-cloud-depends-on-magnetic-tape>
- Patent dispute between Sony and Fujifilm => No LTO-8 supply available globally
 - https://www.theregister.co.uk/2019/05/31/lto_patent_case_hits_lto8_supply/
 - https://www.theregister.co.uk/2019/08/06/sony_fujifilm_storage_patent_lawsuit_settled/
- Sony, Fujifilm stopped patent dispute (however not officially announced from both sides) at the end of July, starting production of LTO-8 media
- Disk = \$25/TB, Tape = \$10/TB, SSD \$100/TB (QLC), SpectraLogic 2019 Report

Data Loss Probability

Data loss probability $p = e^{-\lambda} \frac{\lambda^k}{k!}$

where $\lambda = \frac{AFR \times (\text{Number of Disks})}{365 \times 24 \div MTTR}$

MTTR = Mean Time To Repair
AFR = Annualized Failure Rate

Assuming 1680 disks, 2% of AFR and 24h of MTTR, one can have $\lambda = 0.046$ so with 4 parity disks the data loss probability p gives,

$$p = e^{-0.092} \frac{0.092^5}{5!} = 0.000000050242575 = 5.02 \times 10^{-9}$$