The alternative to tape based archive storage at KISTI

Sang Un Ahn[1], Latchezar Betev[2], Eric Bonfillou[2], Heejune Han[1], Jeongheon Kim[1], Seung Hee Lee[1], Bernd Panzer-Steindel[2], Andreas Joachim Peters[2], Heejun Yoon[1]

[1]KISTI, Daejeon, South Korea
[2]CERN, Geneva, Switzerland

WLCG GDB
11 March 2020

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

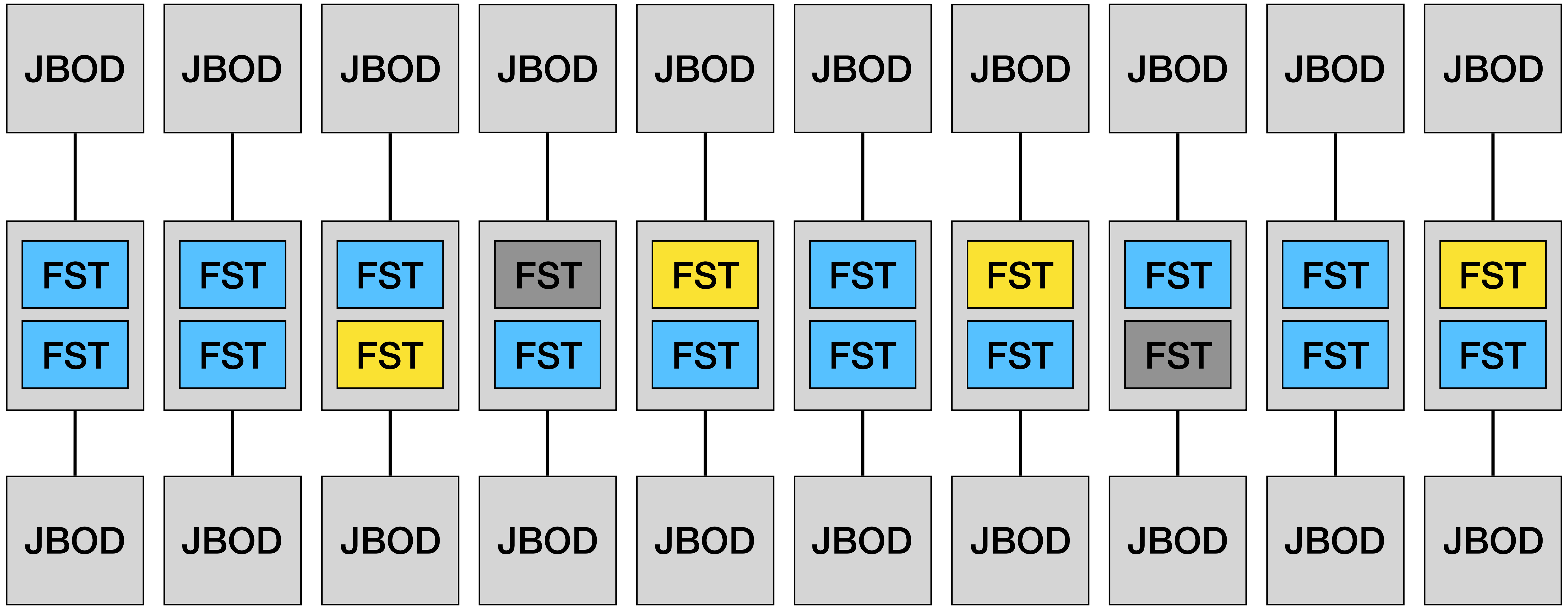
ATAS Project

- 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

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 ← M = data node K = parity node
  - 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

EOS RAIN6 (14,4)



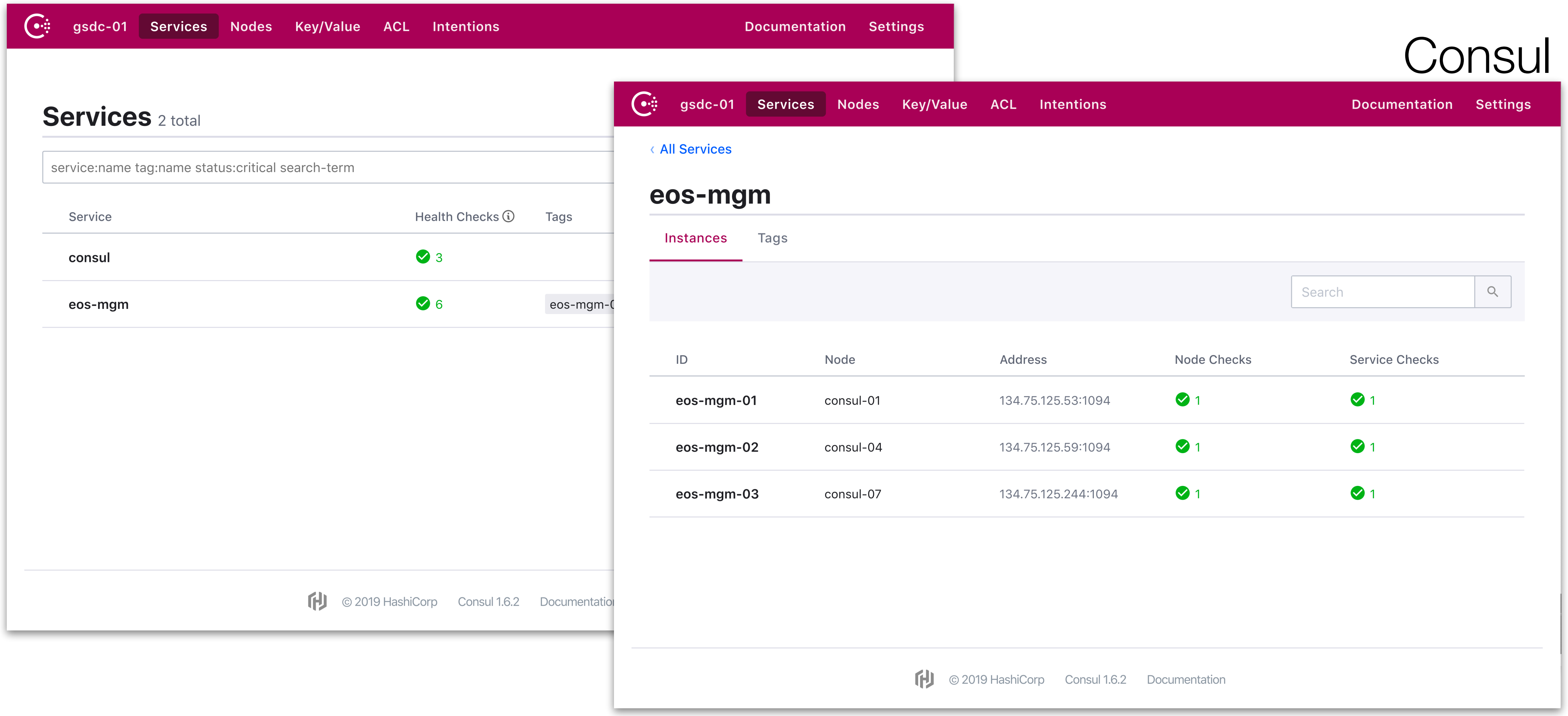
- (x2) EOS FSTs based on Docker container
- EOS decides where to store data fragments across FST nodes randomly (no fixed scheme)

EOS-Docker (1/2)

- Based on CERN EOS Docker Project + AARNet EOS Docker
  - Dumb-init + bash-script-with-exec to forward properly SIGTERM signal when docker stop command issued
- Ansible configuration in "2-pass"
  - Pass 1: Pre run to auto-generate inventory, group_vars, host_vars from config.yaml
  - Pass 2: Full run using the auto-generated configs
- Not ready for orchestration (e.g. Kubernetes); optimal network equipments and configurations are required
- Host container network to share the host IP among containers
- Available on GitHub
  - https://github.com/jeongheon81/gsdc-eos-docker

EOS-Docker (2/2)

- Consul : health check and DNS service, event trigger (experimental)
- Simple log collection : Grafana + Loki + Promtail (all-in-one setup)
- Simple monitoring : Cockpit (cluster mode)
- Others : chrony, firewalld, network (nmcli)



Consul

Services 2 total

service:name tag:name status:critical search-term

Service	Health Checks	Tags
consul	✓ 3	
eos-mgm	✓ 6	eos-mgm-(

© 2019 HashiCorp Consul 1.6.2 Documentatio

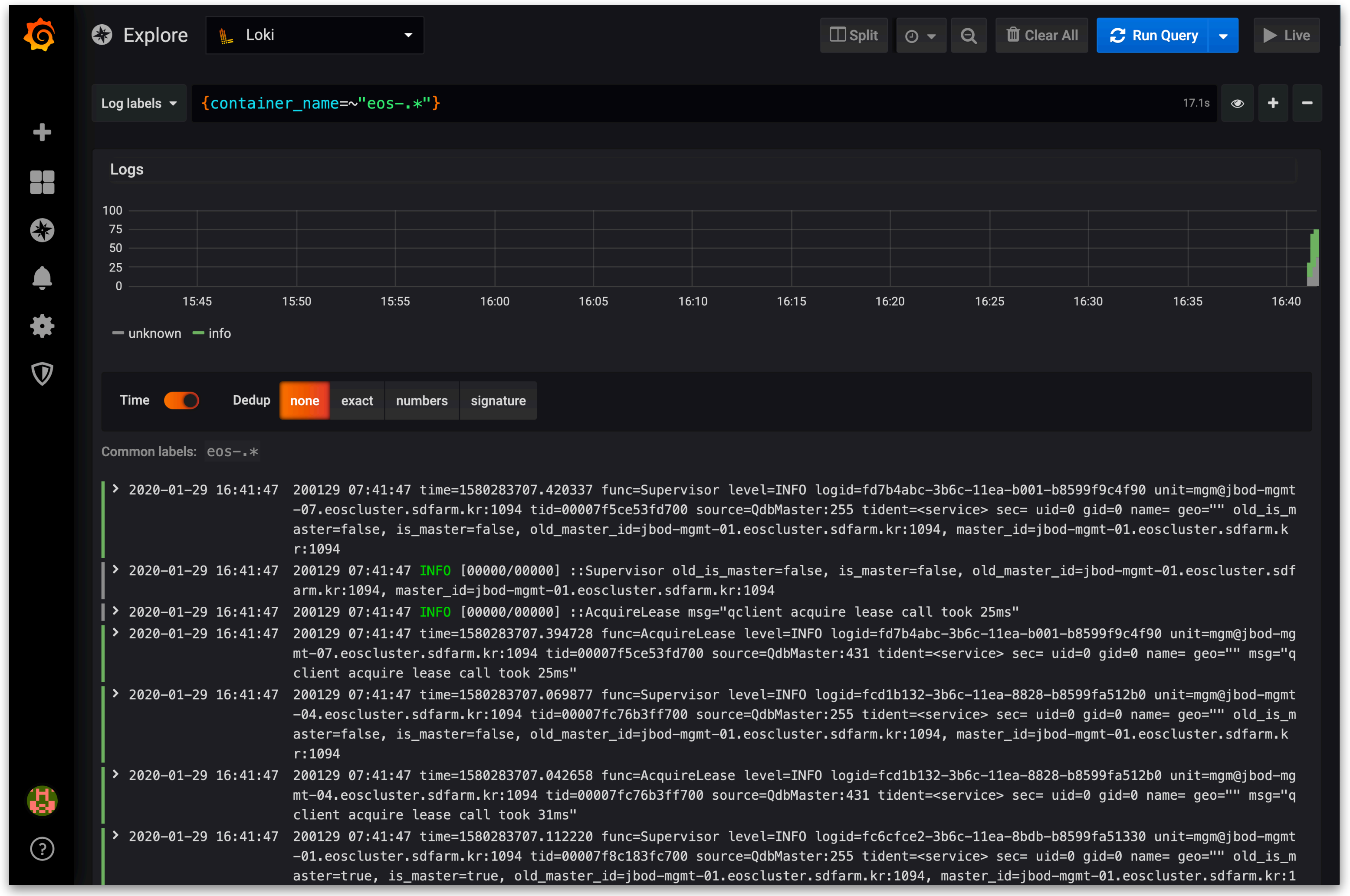
‹ All Services

eos-mgm

ID	Node	Address	Node Checks	Service Checks
eos-mgm-01	consul-01	134.75.125.53:1094	✓ 1	✓ 1
eos-mgm-02	consul-04	134.75.125.59:1094	✓ 1	✓ 1
eos-mgm-03	consul-07	134.75.125.244:1094	✓ 1	✓ 1

© 2019 HashiCorp Consul 1.6.2 Documentation

Grafana + Loki

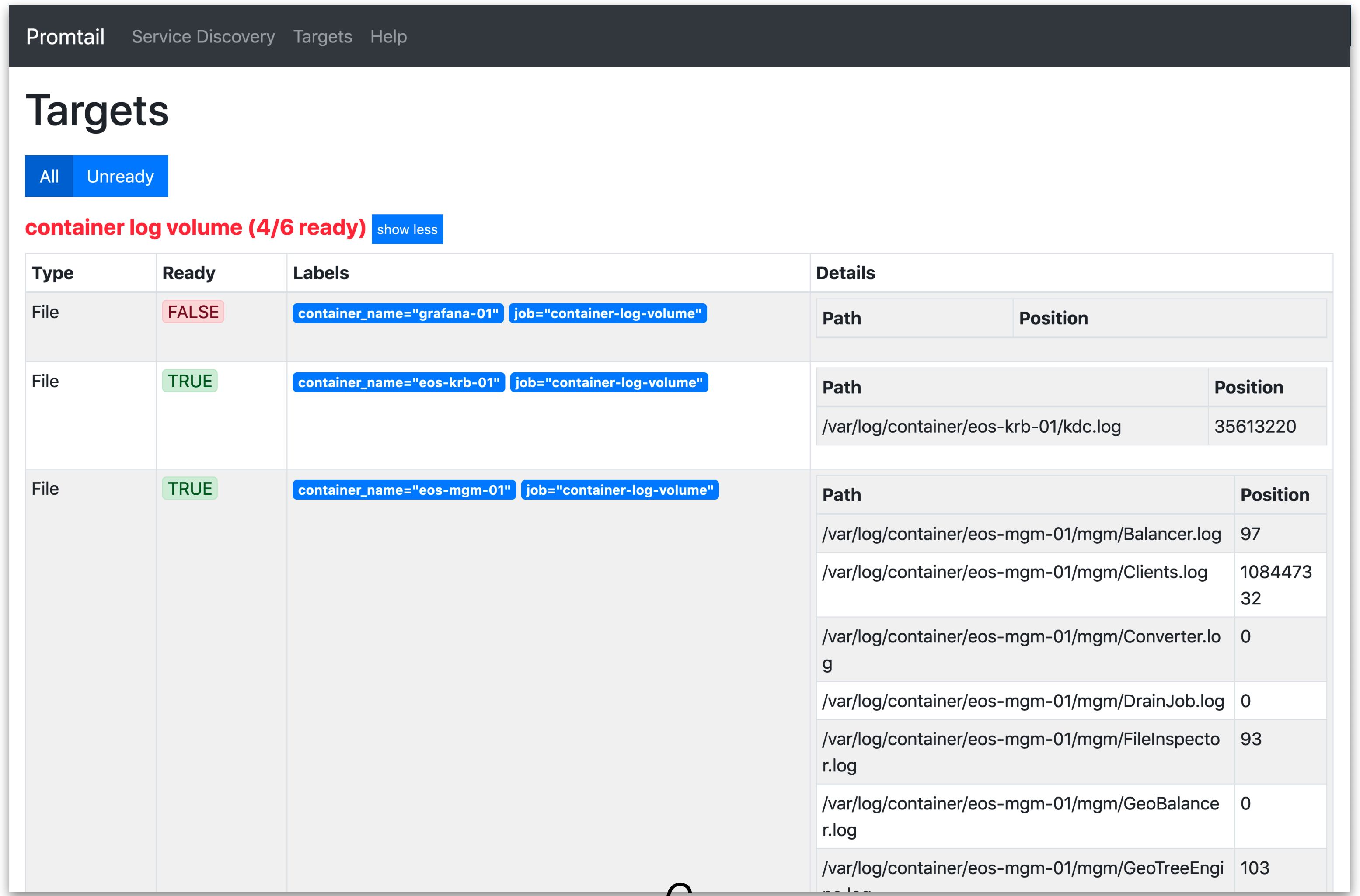


```
{container_name=~"eos-.*"}
```

Common labels: eos-.*

```
> 2020-01-29 16:41:47  200129 07:41:47 time=1580283707.420337 func=Supervisor level=INFO logid=fd7b4abc-3b6c-11ea-b001-b8599f9c4f90 unit=mgm@jbod-mgmt
                       -07.eoscluster.sdfarm.kr:1094 tid=00007f5ce53fd700 source=QdbMaster:255 tident=<service> sec= uid=0 gid=0 name= geo="" old_is_m
                       aster=false, is_master=false, old_master_id=jbod-mgmt-01.eoscluster.sdfarm.kr:1094, master_id=jbod-mgmt-01.eoscluster.sdfarm.k
                       r:1094
> 2020-01-29 16:41:47  200129 07:41:47 INFO [00000/00000] ::Supervisor old_is_master=false, is_master=false, old_master_id=jbod-mgmt-01.eoscluster.sdf
                       arm.kr:1094, master_id=jbod-mgmt-01.eoscluster.sdfarm.kr:1094
> 2020-01-29 16:41:47  200129 07:41:47 INFO [00000/00000] ::AcquireLease msg="qclient acquire lease call took 25ms"
> 2020-01-29 16:41:47  200129 07:41:47 time=1580283707.394728 func=AcquireLease level=INFO logid=fd7b4abc-3b6c-11ea-b001-b8599f9c4f90 unit=mgm@jbod-mg
                       mt-07.eoscluster.sdfarm.kr:1094 tid=00007f5ce53fd700 source=QdbMaster:431 tident=<service> sec= uid=0 gid=0 name= geo="" msg="q
                       client acquire lease call took 25ms"
> 2020-01-29 16:41:47  200129 07:41:47 time=1580283707.069877 func=Supervisor level=INFO logid=fcd1b132-3b6c-11ea-8828-b8599fa512b0 unit=mgm@jbod-mgmt
                       -04.eoscluster.sdfarm.kr:1094 tid=00007fc76b3ff700 source=QdbMaster:255 tident=<service> sec= uid=0 gid=0 name= geo="" old_is_m
                       aster=false, is_master=false, old_master_id=jbod-mgmt-01.eoscluster.sdfarm.kr:1094, master_id=jbod-mgmt-01.eoscluster.sdfarm.k
                       r:1094
> 2020-01-29 16:41:47  200129 07:41:47 time=1580283707.042658 func=AcquireLease level=INFO logid=fcd1b132-3b6c-11ea-8828-b8599fa512b0 unit=mgm@jbod-mg
                       mt-04.eoscluster.sdfarm.kr:1094 tid=00007fc76b3ff700 source=QdbMaster:431 tident=<service> sec= uid=0 gid=0 name= geo="" msg="q
                       client acquire lease call took 31ms"
> 2020-01-29 16:41:47  200129 07:41:47 time=1580283707.112220 func=Supervisor level=INFO logid=fc6cfce2-3b6c-11ea-8bdb-b8599fa51330 unit=mgm@jbod-mgmt
                       -01.eoscluster.sdfarm.kr:1094 tid=00007f8c183fc700 source=QdbMaster:255 tident=<service> sec= uid=0 gid=0 name= geo="" old_is_m
                       aster=true, is_master=true, old_master_id=jbod-mgmt-01.eoscluster.sdfarm.kr:1094, master_id=jbod-mgmt-01.eoscluster.sdfarm.kr:1
```

Promtail

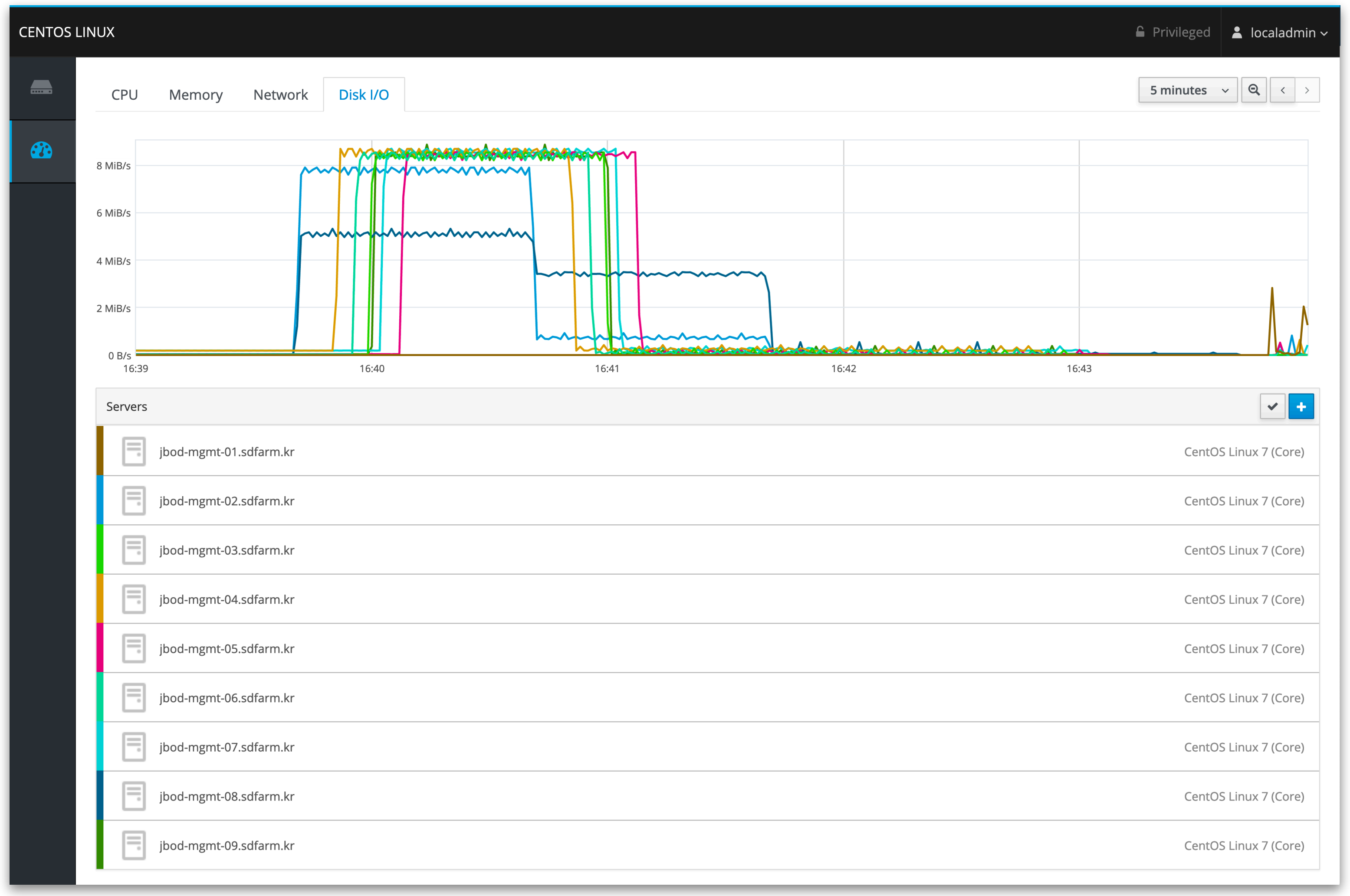


Targets

container log volume (4/6 ready)

Type	Ready	Labels	Details	
File	FALSE	container_name="grafana-01" job="container-log-volume"	**Path**	**Position**
File	TRUE	container_name="eos-krb-01" job="container-log-volume"	**Path**	**Position**
			/var/log/container/eos-krb-01/kdc.log	35613220
File	TRUE	container_name="eos-mgm-01" job="container-log-volume"	**Path**	**Position**
			/var/log/container/eos-mgm-01/mgm/Balancer.log	97
			/var/log/container/eos-mgm-01/mgm/Clients.log	108447332
			/var/log/container/eos-mgm-01/mgm/Converter.log	0
			/var/log/container/eos-mgm-01/mgm/DrainJob.log	0
			/var/log/container/eos-mgm-01/mgm/FileInspector.log	93
			/var/log/container/eos-mgm-01/mgm/GeoBalancer.log	0
			/var/log/container/eos-mgm-01/mgm/GeoTreeEngi	103

Cockpit



Servers	
jbod-mgmt-01.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-02.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-03.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-04.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-05.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-06.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-07.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-08.sdfarm.kr	CentOS Linux 7 (Core)
jbod-mgmt-09.sdfarm.kr	CentOS Linux 7 (Core)

DEMO 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

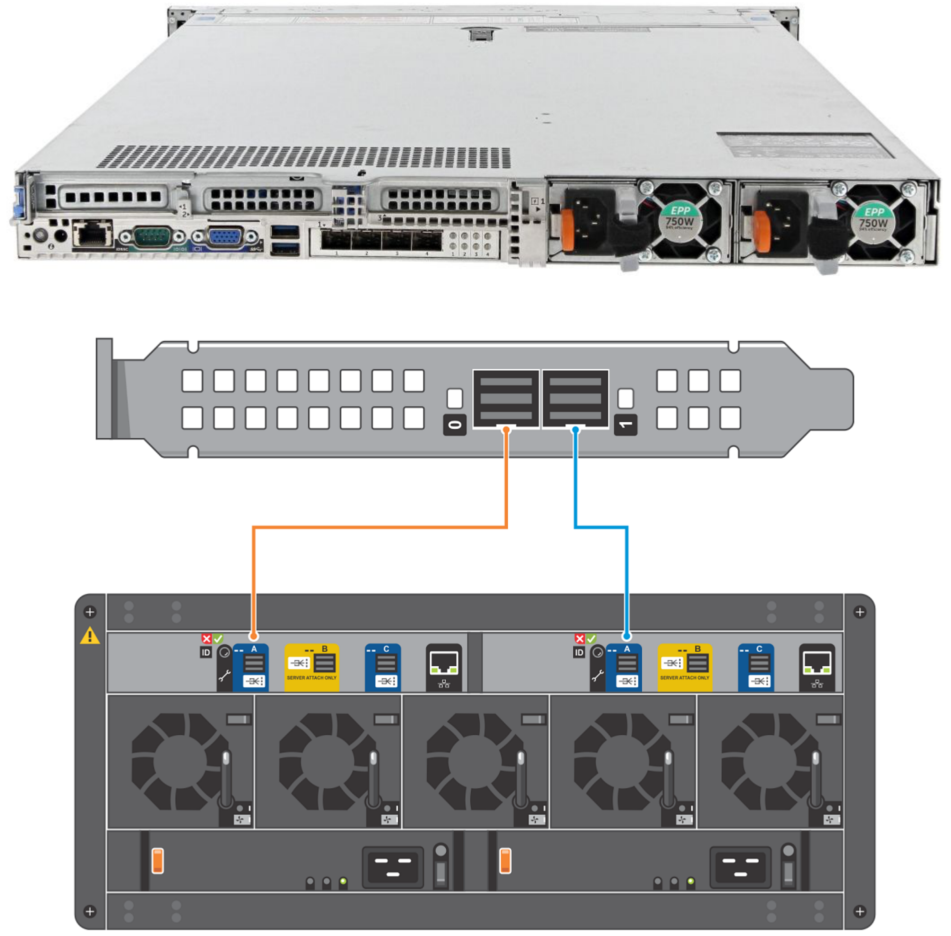
Operating System
CentOS Linux 7 (Core)

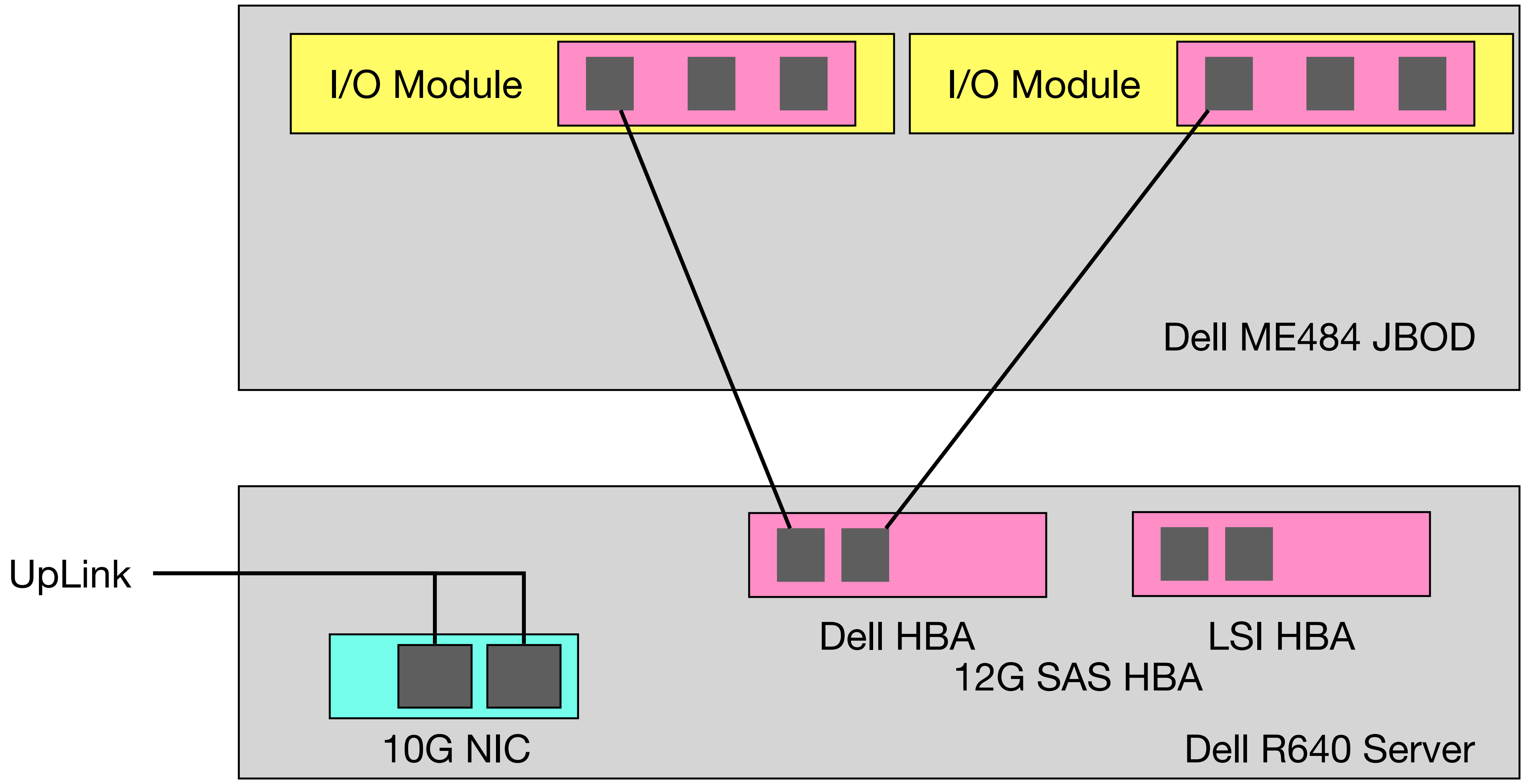
Operating System Kernel Version
7 (Core) Kernel 3.10.0.-957.el7.x86_64

BIOS Version
1.5.6

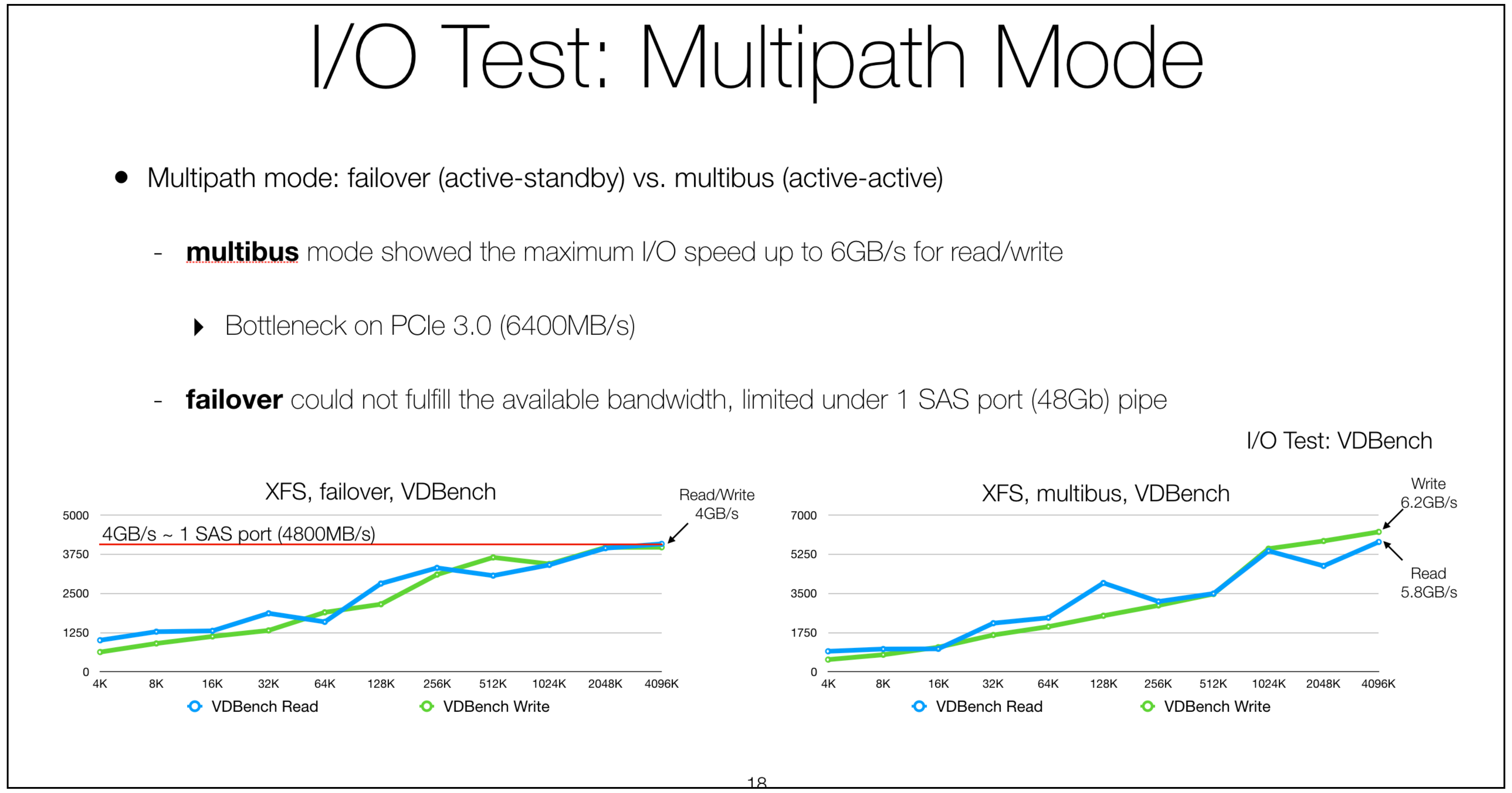
Filesystem: XFS (Default EL7 Distribution)

Powered by CentOS





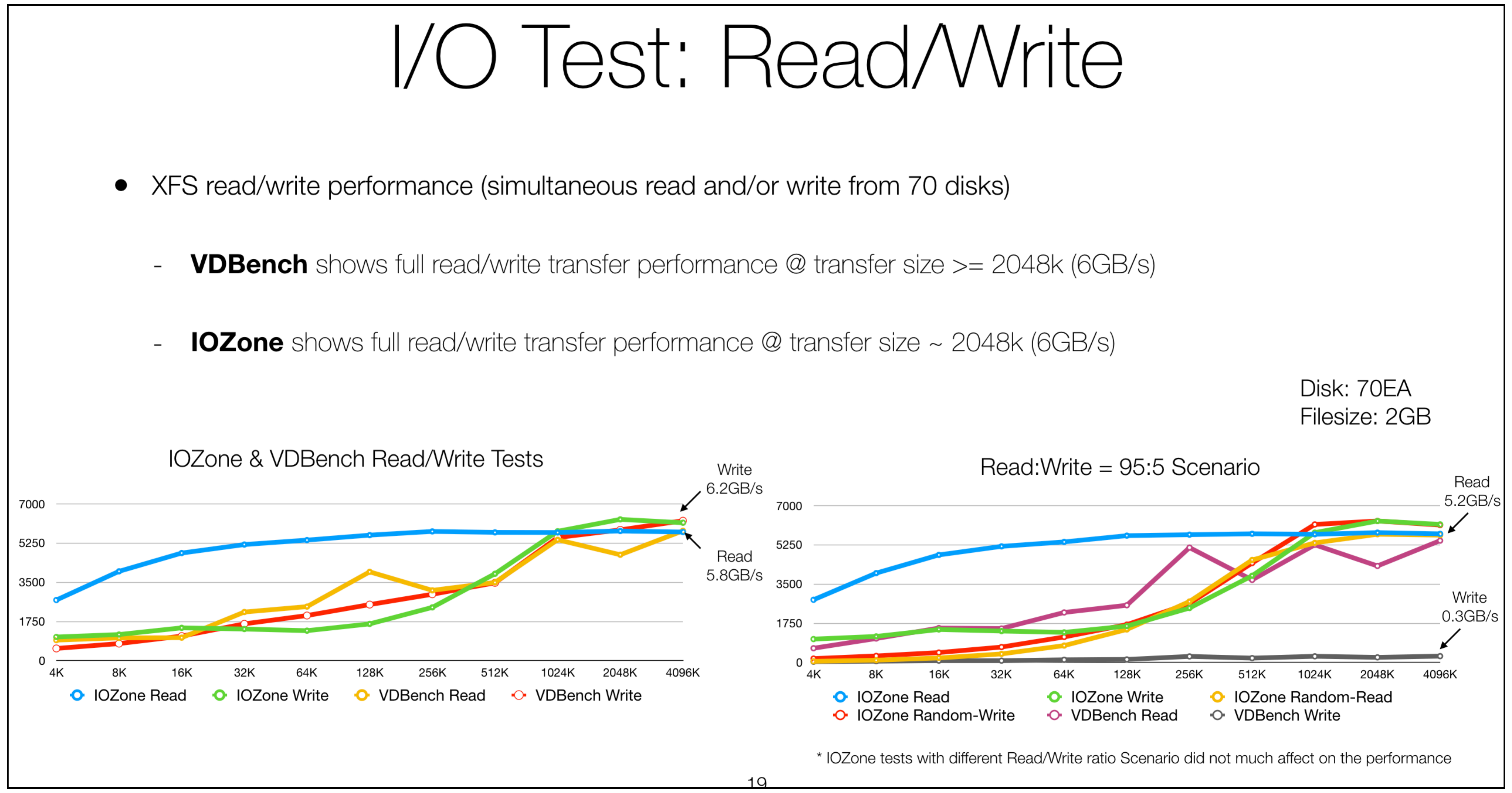
Performance Test Results



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: VDBench

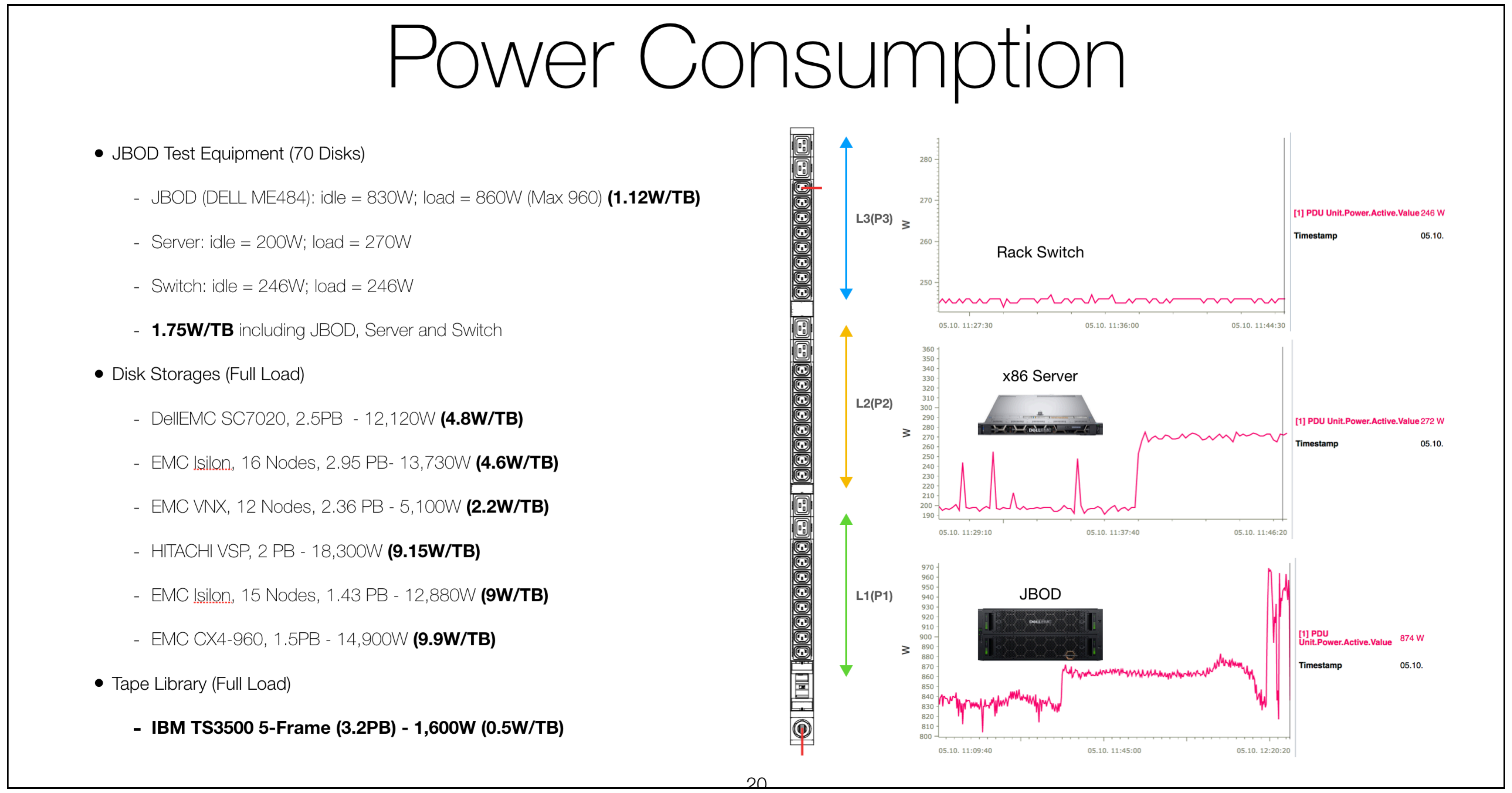


I/O Test: Read/Write

- XFS read/write performance (simultaneous read and/or write from 70 disks)
  - **VDBench** shows full read/write transfer performance @ transfer size >= 2048k (6GB/s)
  - **IOZone** shows full read/write transfer performance @ transfer size ~ 2048k (6GB/s)

Disk: 70EA
Filesize: 2GB

* 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)**

- Confirmed the upper cap of read/write performance ~ 6GB/s (intrinsic limit by PCIe 3.0)
- Power consumption shown ~ 1.75W/TB, not uncomfortably higher than Tape (0.5W/TB)
  - High-end Enterprise Disk Storage 5 ~ 9W/TB

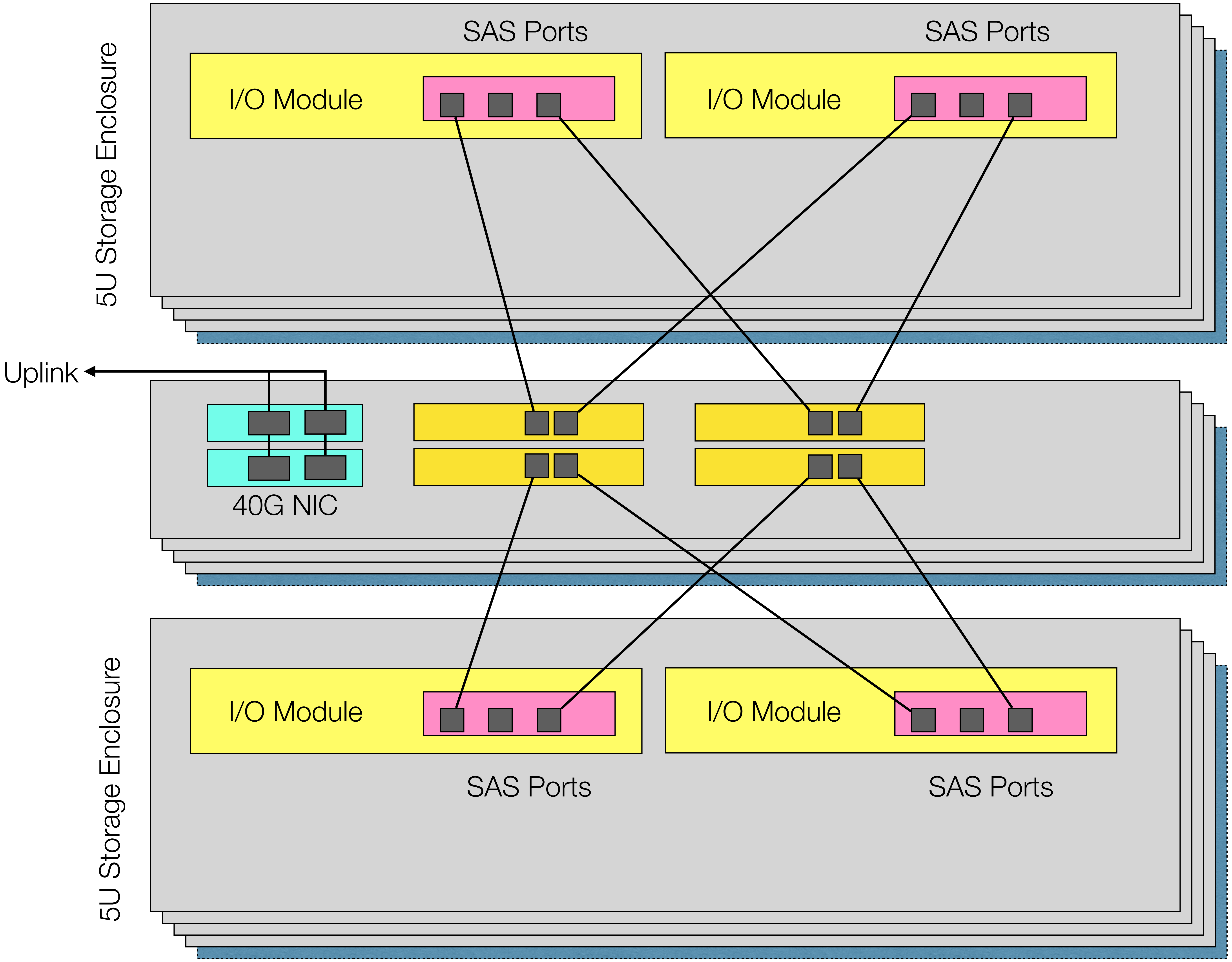
Procurement Schedule

1 week
2 weeks
1.5 weeks
10 weeks
2 weeks
Week
Pre-annoucnement
19' Sep 9
Main announcement
19' Sep 24
Bid
19' Oct 9
Delivery & Installation
19' Oct 18
Finished
Install & BMT
19' Dec 26
Setup
EOS & Docker
20' Jan 10
Finished
EOS & Docker
Setup
20' Jan 23

Delivery & Installation

- Dec 17th ~ 27th

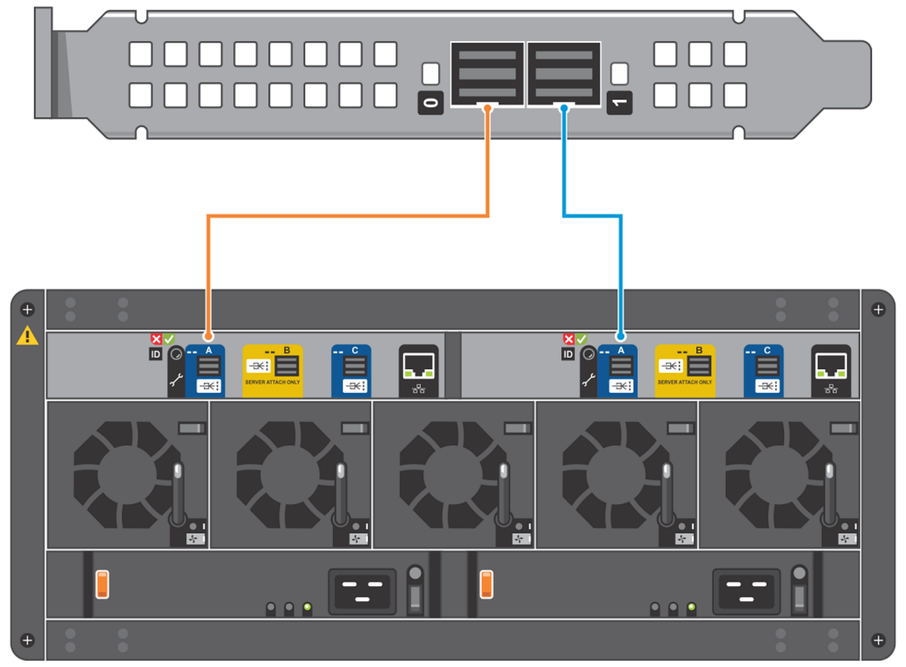
Delivered Systems



Specifications

- x9 Servers: Dell PowerEdge R730
  - Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50GHz * 2EA
  - DDR4 16GB 2,400MHz * 12EA (192GB)
  - Dell PowerEdge 12Gbps SAS HBA * 4EA
  - MLNX 40Gb 2P ConnectX3Pro Adpt * 4EA
- x18 JBOD: Dell PowerVault ME484
  - 84EA HGST or Seagate 12TB 7.2K NL-SAS)
- x2 40G network switches

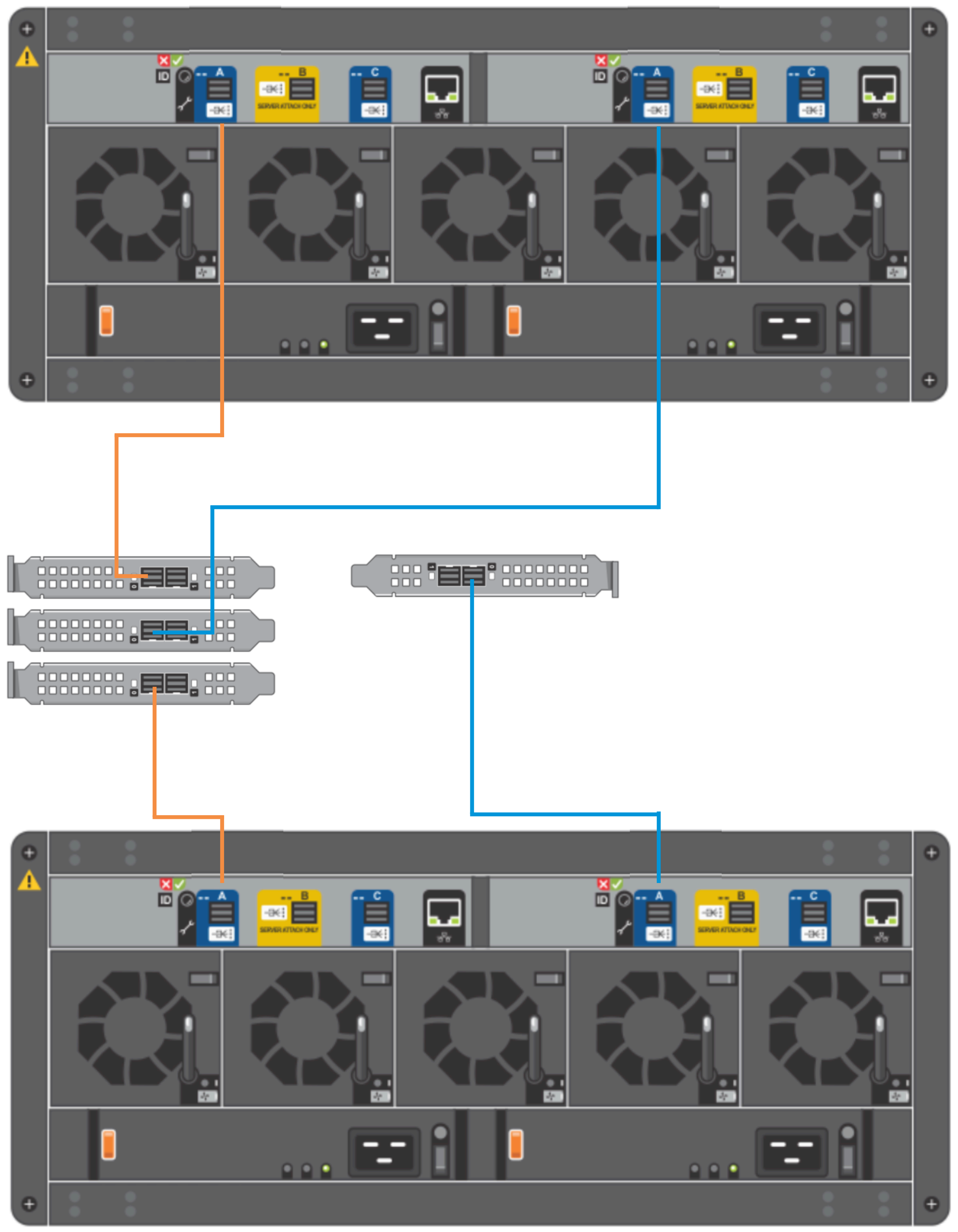
ATAS Storage
18PB (RAW)
14PB (Usable)



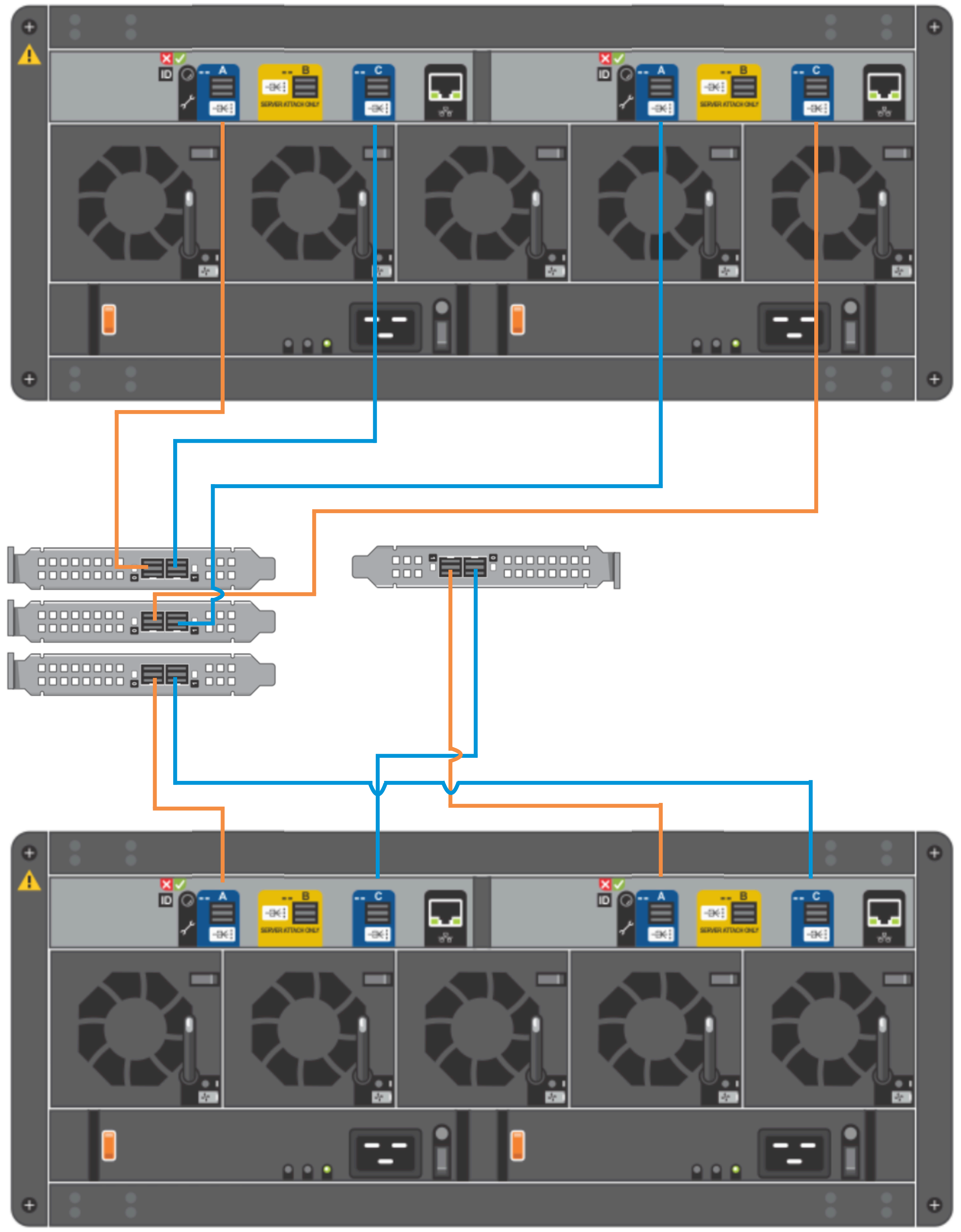
- Confirmed the upper cap of read/write performance ~ 6GB/s (intrinsic limit by PCIe 3.0)
- Power consumption will be measured throughout the whole testing & commissioning periods

JBOD Cabling

- Recommended
  - One Server / Four HBA per server / Single path

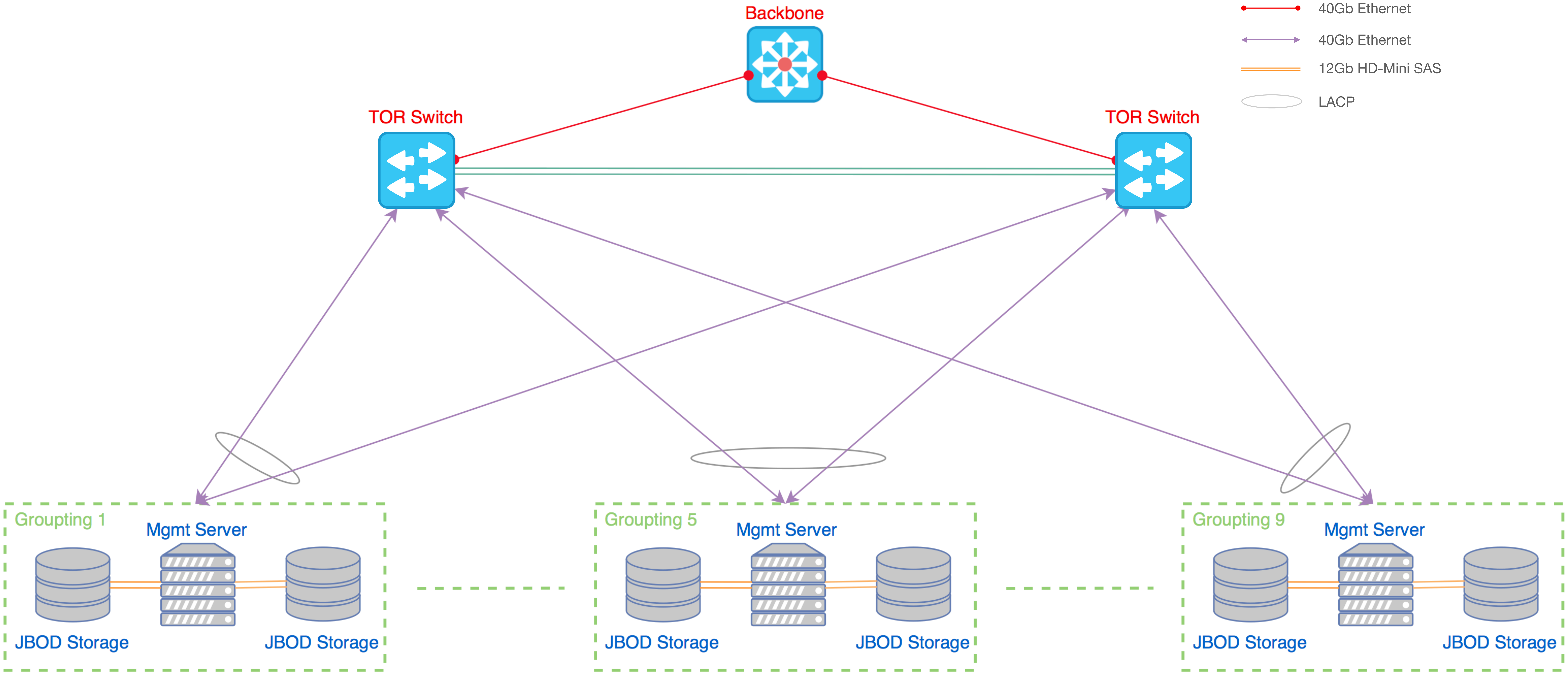


- Target
  - One Server / Four HBA per server / Dual path



- Disk Access, Recognition Test via Multipath
- Data Consistency, Corruption Test
- Read/Write Performance Test

Network Topology



Schedule

	2019												2020		
Tasks	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3
Technology Search	▨														
Product Survey		▨ ←	KISTI-CERN Expert Meeting @ KISTI												
Design & Specification			▨	▨ ←	KISTI-CERN Expert Meeting @ CERN										
Testing					▨	▨	▨					EOS Workshop @ CERN			
Procurement					▨	▨ 🟨	▨ Call for tender (delayed) →	▨	▨	▨ 🟨	🟨 Delivery			↓	Today ↓
Installation & Validation											▨	▨			
EOS Deployment Test													▨	▨ RAIN Layout	▨

- 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)
- Commissioning and production test in 2020 targeting the production service before RUN3

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
- Power consumption is shown to be 1.75W/TB, not uncomfortably higher than a tape library
- Procurement and system deliveries finished in November 2019, installation and validation on delivered systems finished in January 2020
- EOS deployment with RAIN layout is being applied and tested after the recent EOS workshop in February, special thanks to EOS developers!!
- During the whole year of 2020, this disk-based custodial storage will be tested and verified repeatedly with ALICE targeting the production service before the start of RUN3

Questions?

TORRENS
ROWING
CLUB
RTC
MENS SANA IN CORPORE SANO
EST. 1903
All Enquiries Welcome
secretary@torrensrowingclub.com.au www.torrensrowingclub.com.au
PR44019B
DELIVERIES

Backup

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

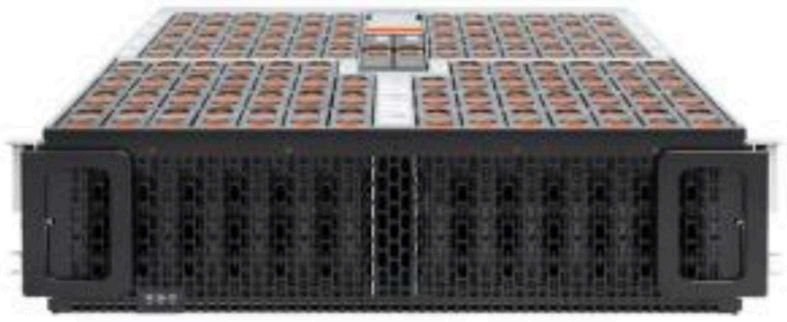
MTTR = Mean Time To Repair
AFR = Annualized Failure Rate

Assuming 1680 disks, 2% of AFR and 24h of MTTR, one can have λ = 0.092 so with 4 parity disks the data loss probability p gives,

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

https://blog.synology.com/data-durability/

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

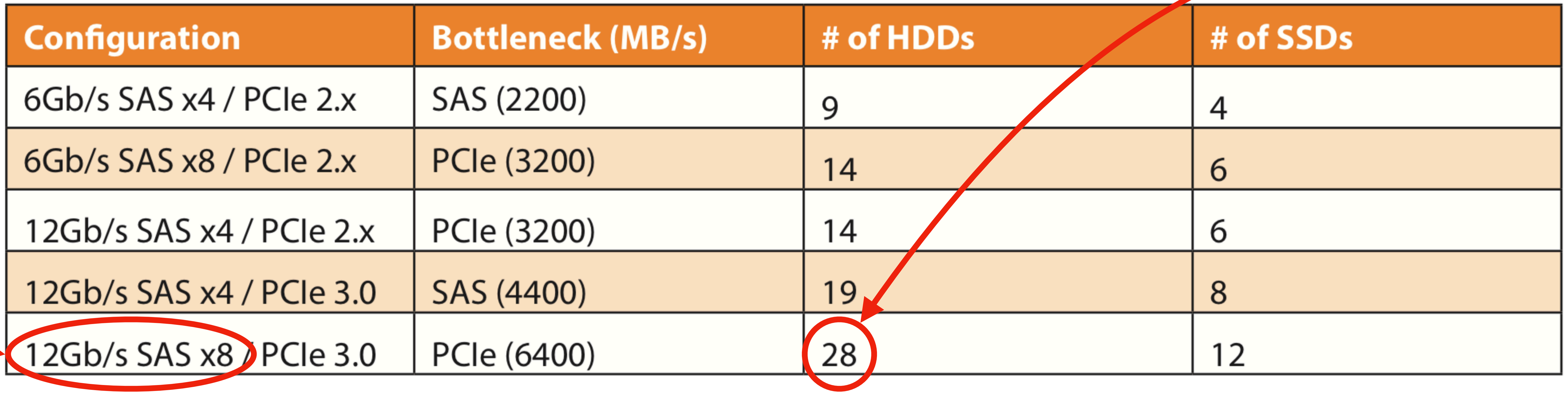
In case of 4 ports

Allowing the transmission of signals in both directions but not simultaneously

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

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

For 15k HDD (~230MB/s)

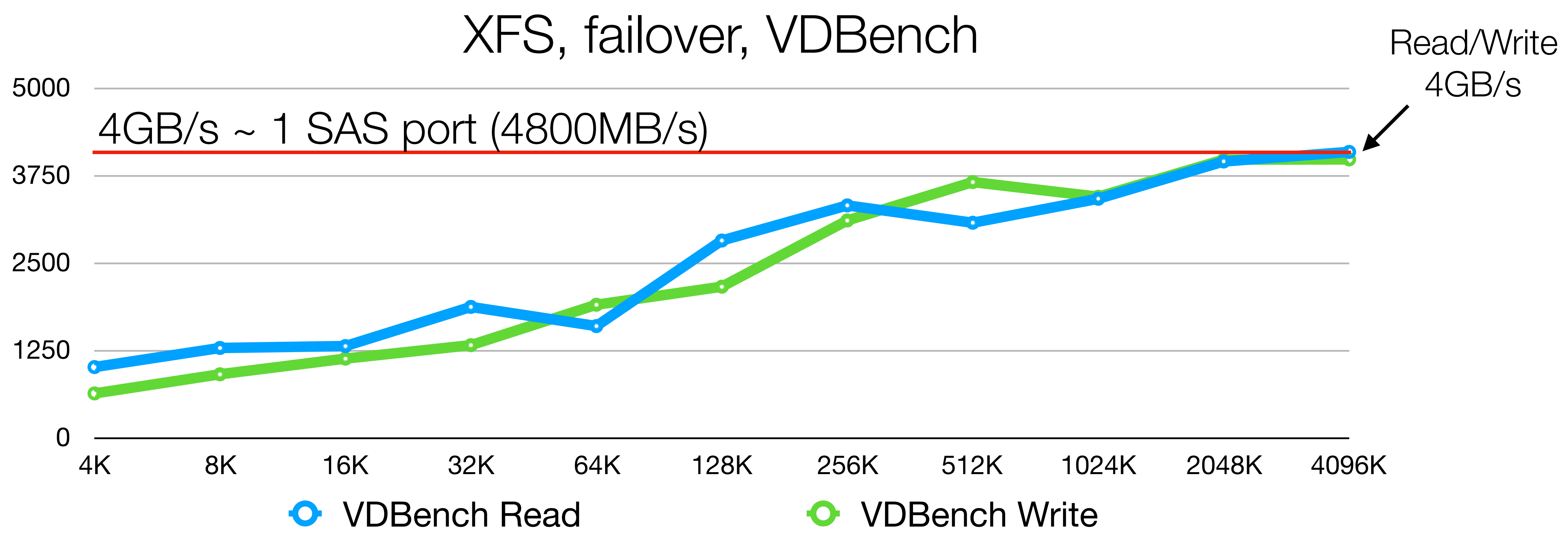
56 slower disks can fulfill
the bandwidth provided by
Two port 12Gb/s SAS HBA card
connected to a PCIe 3.0 slot

https://docs.broadcom.com/docs/12353459

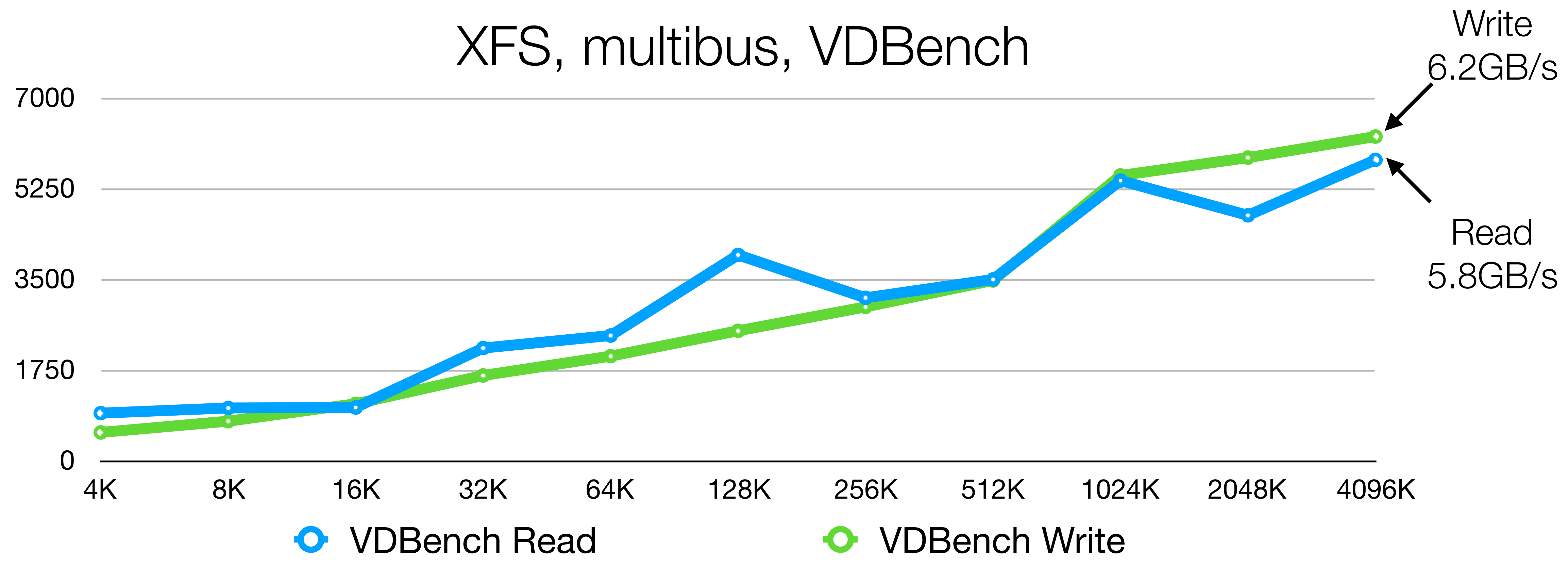
I/O Test: Multipath Mode

- 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: VDBench


XFS, failover, VDBench

Read/Write 4GB/s


XFS, multibus, VDBench

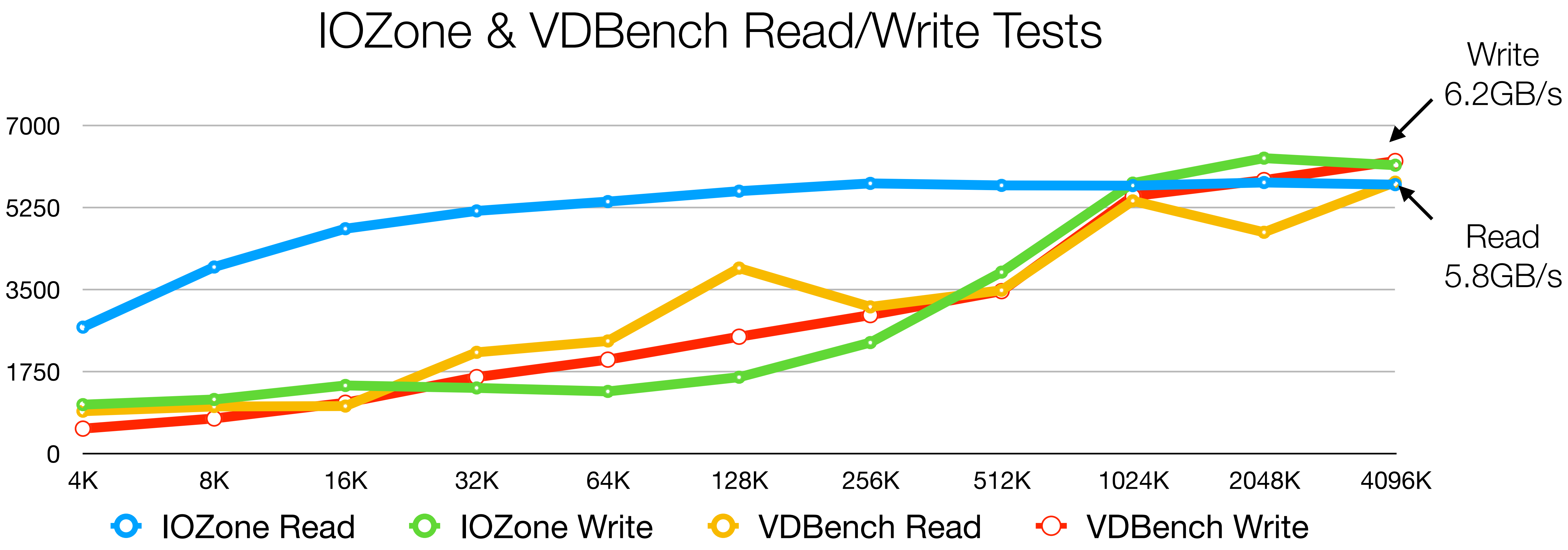
Write 6.2GB/s

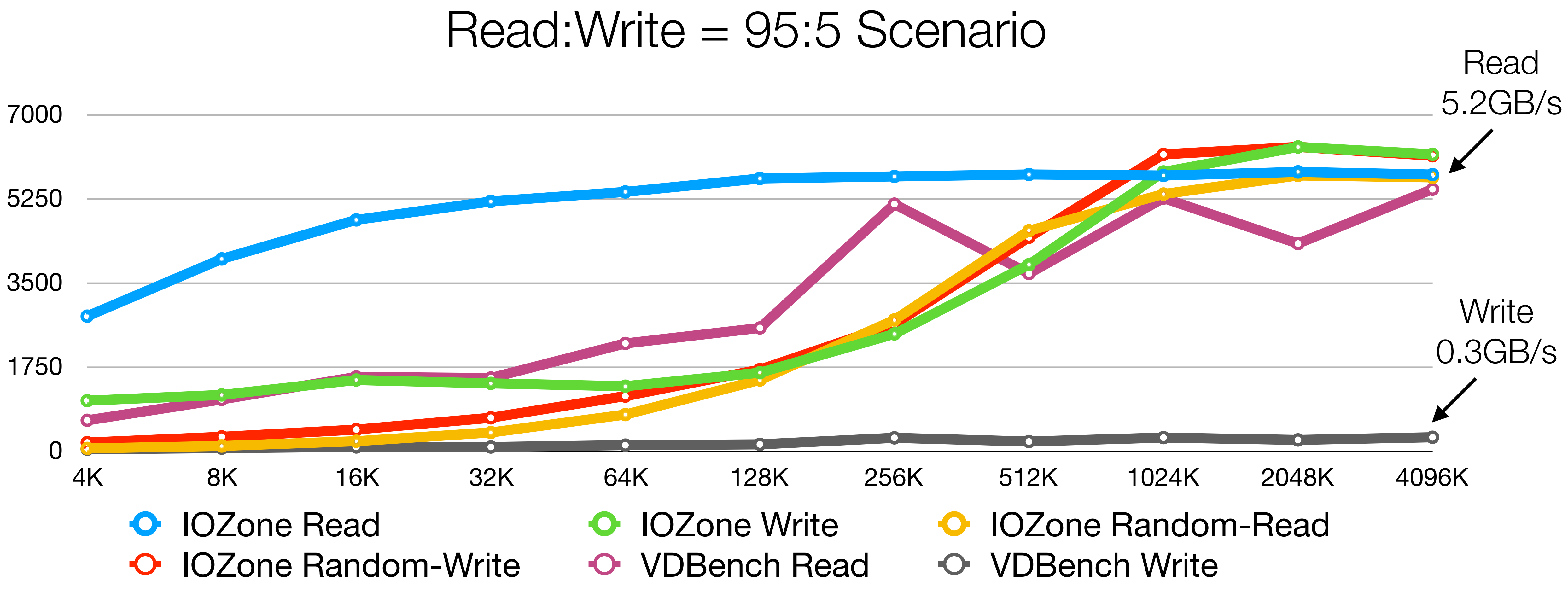
Read 5.8GB/s

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 >= 2048k (6GB/s)
  - **IOZone** shows full read/write transfer performance @ transfer size ~ 2048k (6GB/s)

Disk: 70EA
Filesize: 2GB

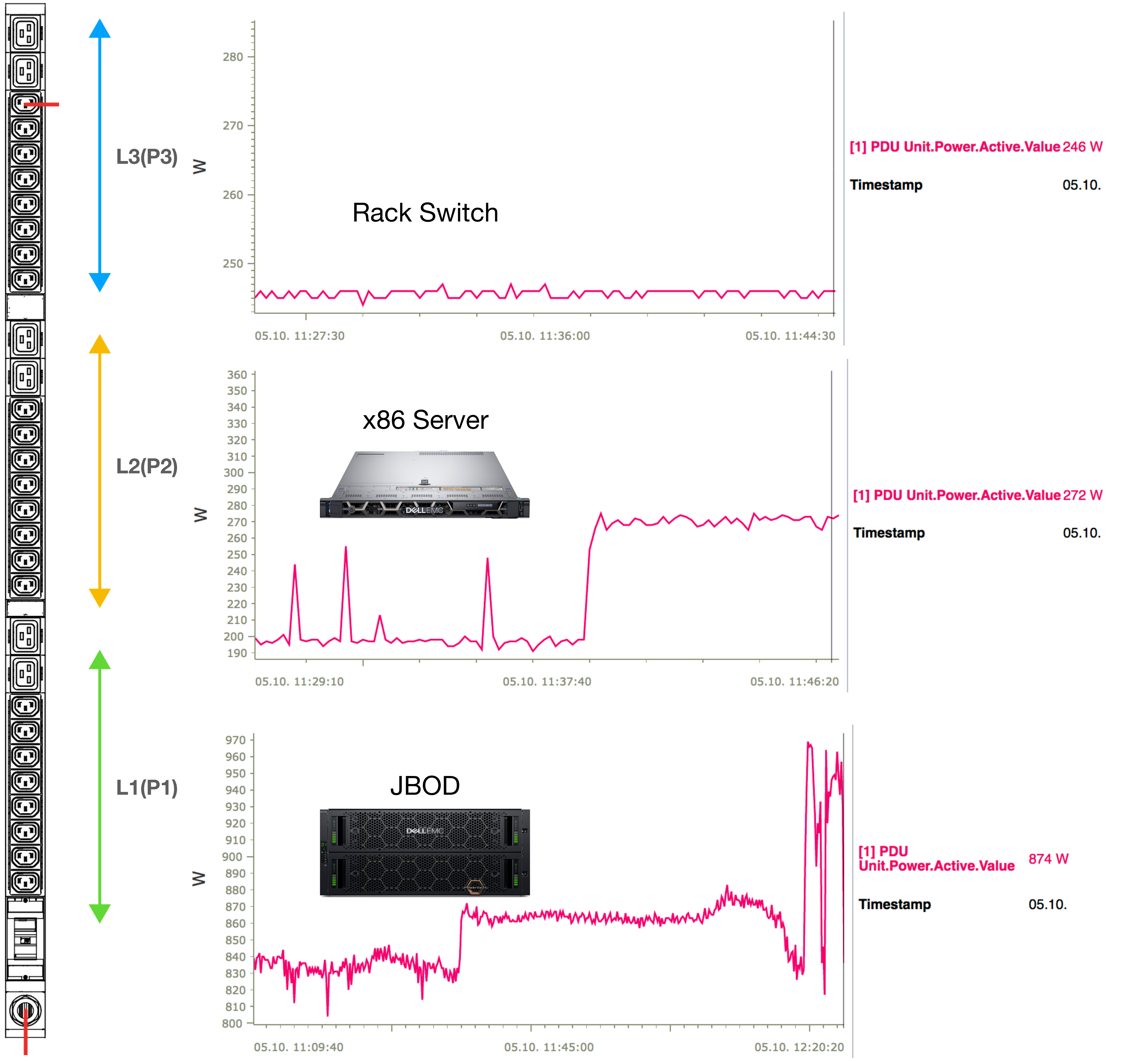




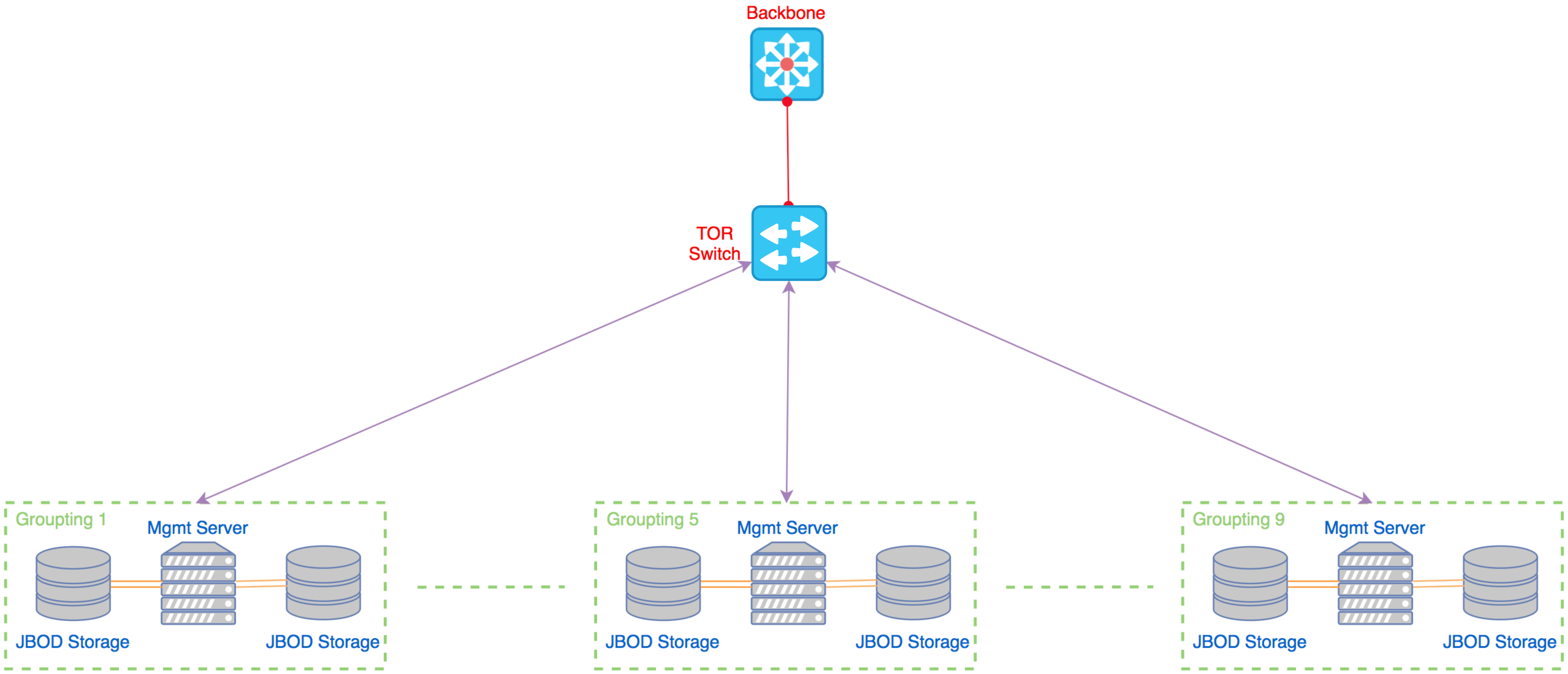
* IOZone tests with different Read/Write ratio Scenario did not much affect on the performance

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)**



Present Network Diagram


Backbone
TOR
Switch
Groupting 1
Mgmt Server
JBOD Storage
JBOD Storage
Groupting 5
Mgmt Server
JBOD Storage
JBOD Storage
Groupting 9
Mgmt Server
JBOD Storage
JBOD Storage