JBOD based archive storage project at KISTI

Heejune Han[1], Latchezar Betev[2], Eric Bonfillou[2], Sang Un Ahn[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

DOMA-QoS Workshop
7 February 2020

Motivation

- Tape Storage Market Monopoly
  - Only left a few major vendor
- High cost of High-End disk storage system
  - High-End Storage is very expensive
- Increasing storage demand
  - Every year data size and request is increase

Pre-test Equipment & Setup

- JBOD: DELL PowerVault ME484
  - Disk: 70EA (HGST 12TB 7.2k NL-SAS), 840 TB
- 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

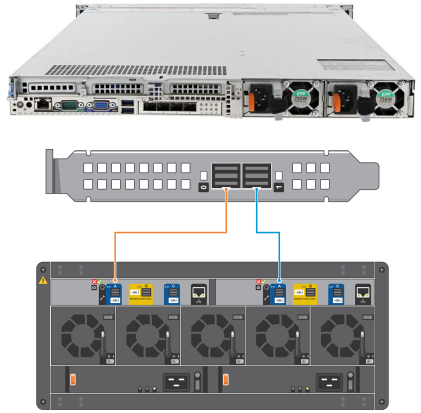
System Information

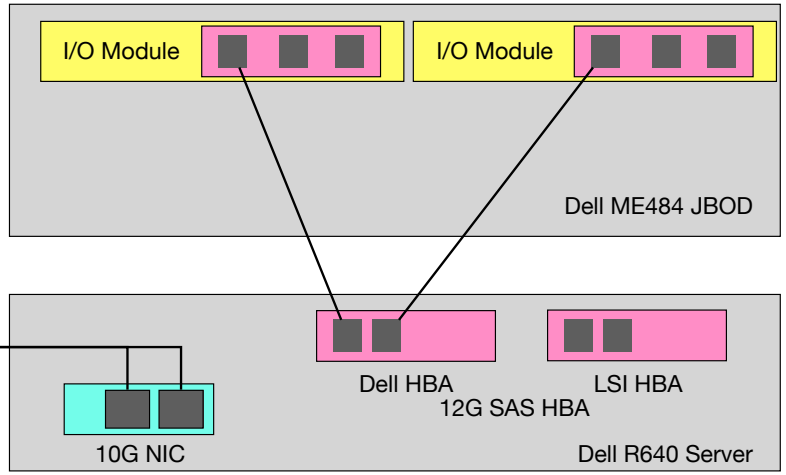
Operating System
CentOS Linux 7 (Core)

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

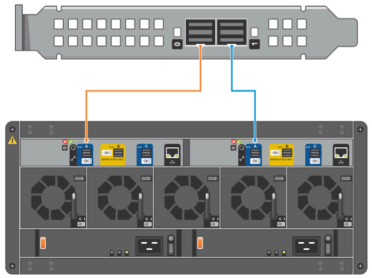
BIOS Version
1.5.6

Filesystem: XFS (Default EL7 Distribution)



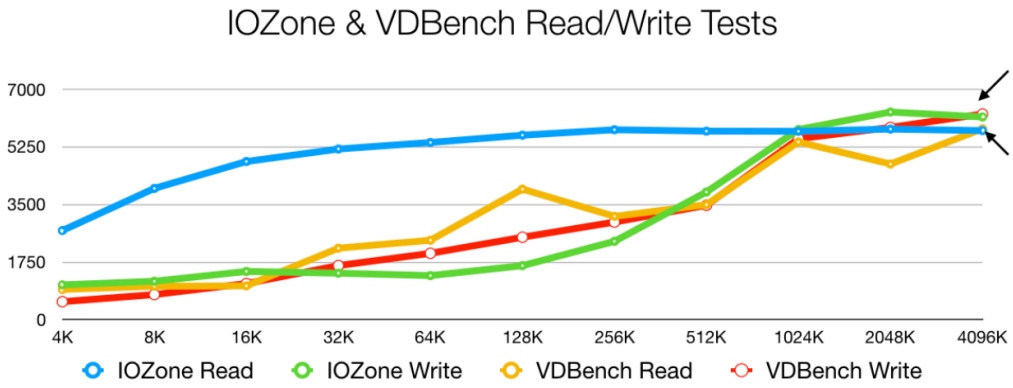


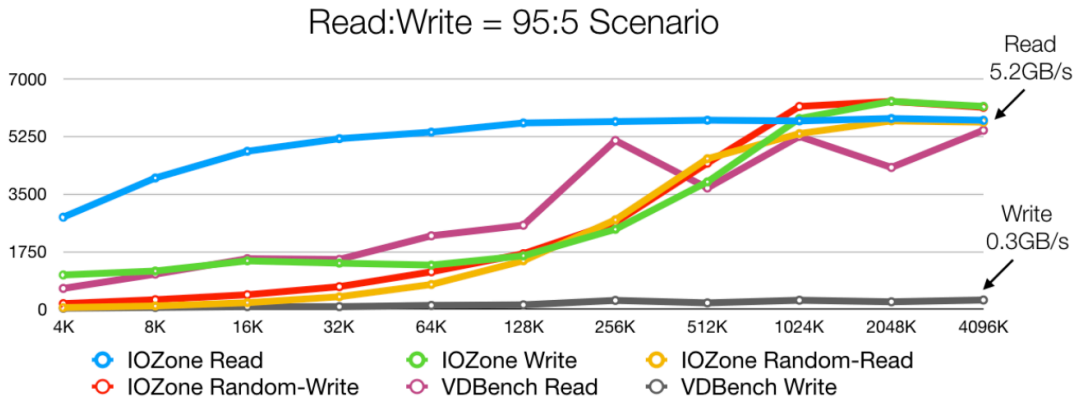
I/O Test: Read/Write



- 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

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

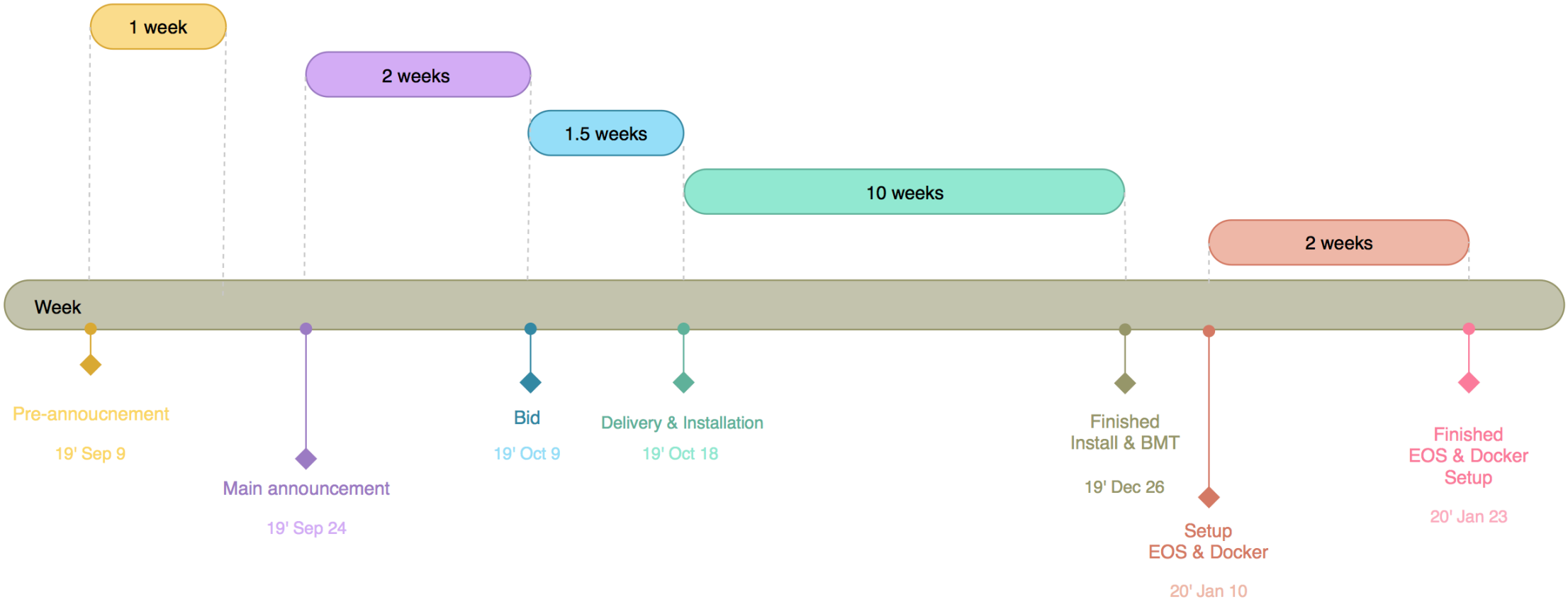
For 15k HDD (~230MB/s)

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

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

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

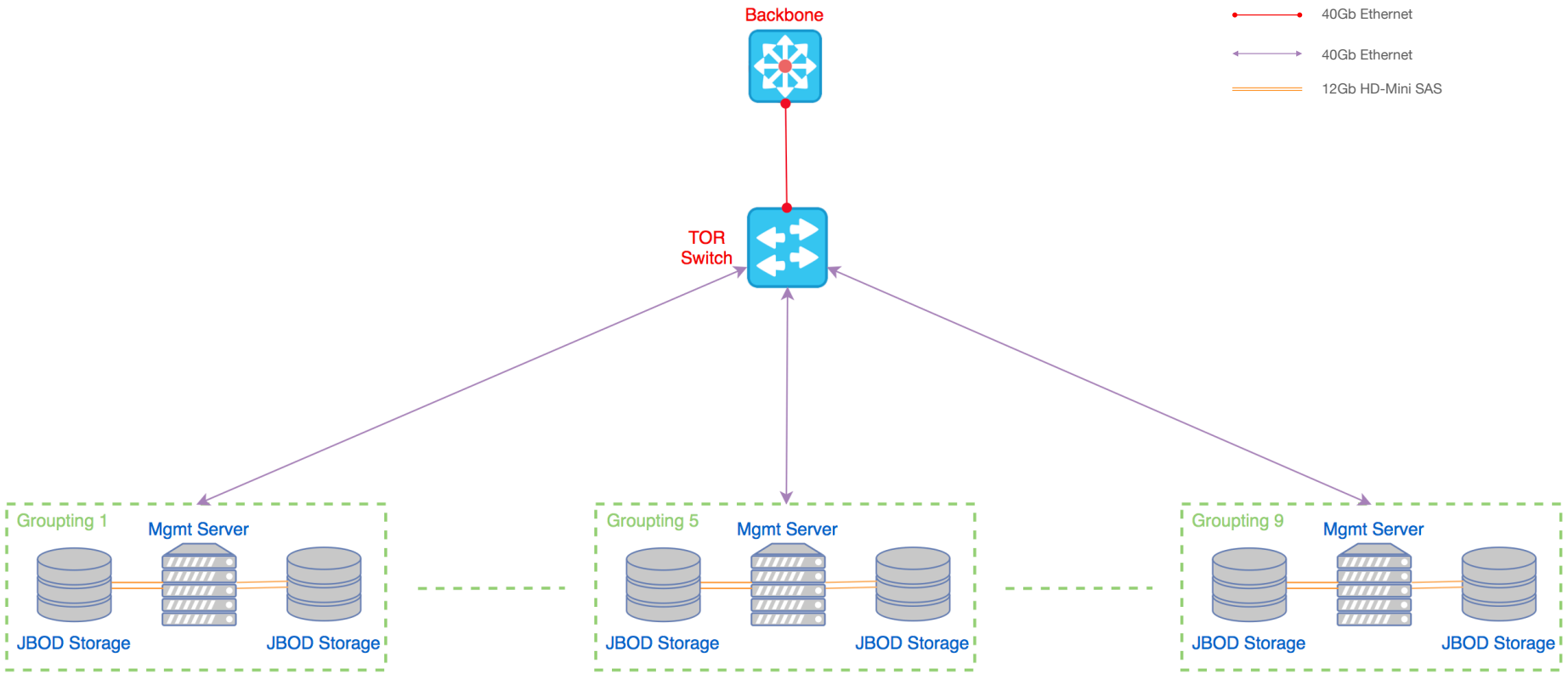
Delivery & Installation

- Dec 17th ~ 27th

Present Equipment

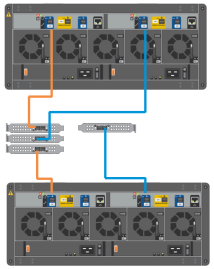
- JBOD: Dell PowerVault ME484 * 2EA
  - Disk: 84EA (HGST or Seagate 12TB 7.2K NL-SAS)
- Server: Dell PowerEdge R730
  - CPU: Intel(R) Xeon(R) CPU E5-2637 v4 @ 3.50GHz 4core * 2EA
  - Memory: DDR4 16GB 2,400MHz * 12EA
  - HBA: Dell PowerEdge 12Gbps SAS HBA * 4EA
  - NIC: MLNX 40Gb 2P ConnectX3Pro Adpt * 2EA

Present Network Diagram

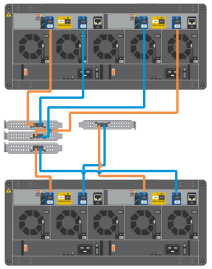


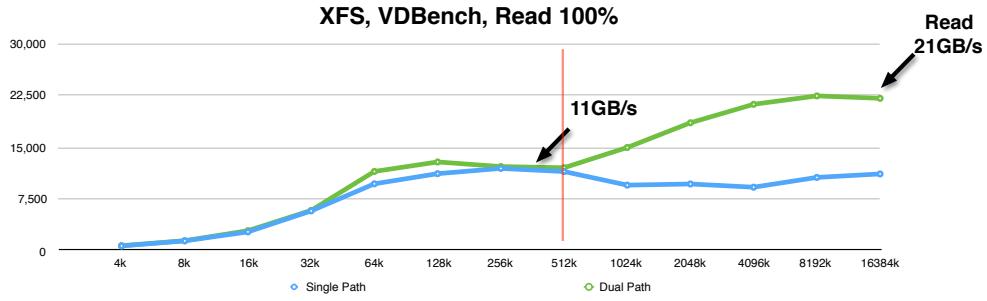
I/O Test: Single/Dual Path

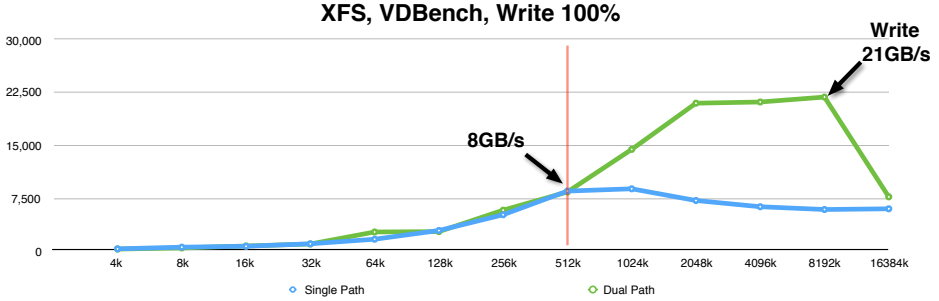
- Single Path
  - Single Path each HBA
  - Every disk has dual path

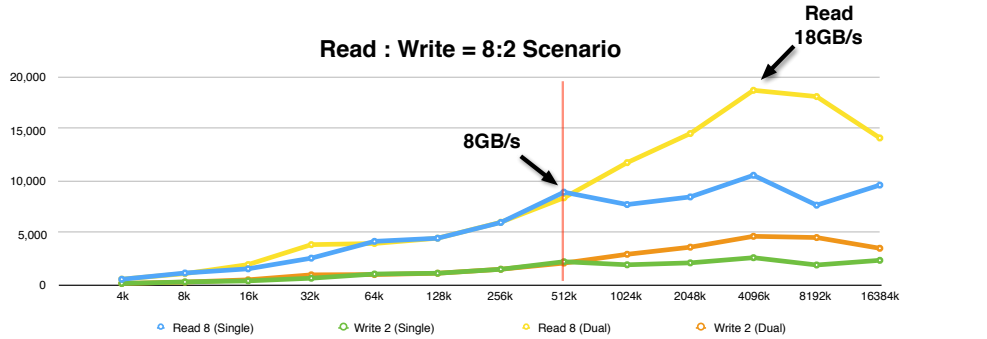


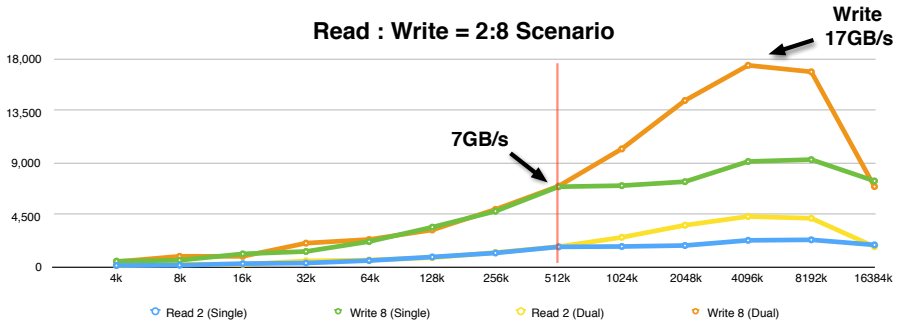
- Dual Path
  - Dual Path each HBA
  - Every disk has quad path







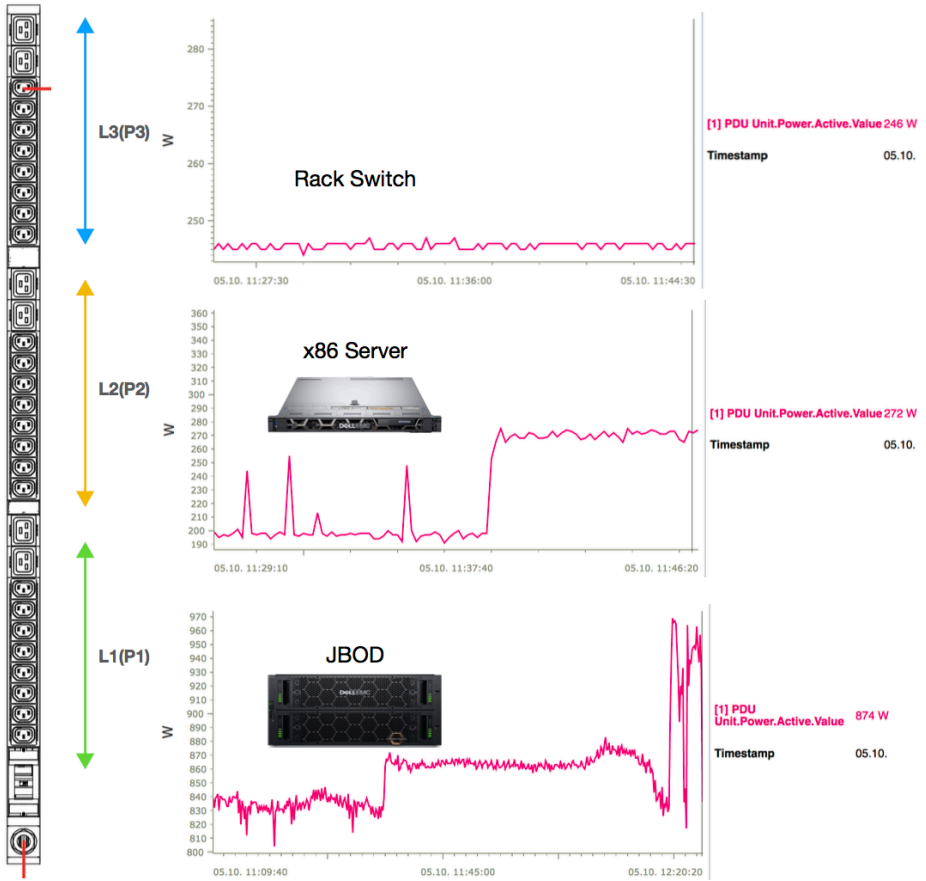




- As file transfer size exceeds 512k, performance difference appears in read and write.
- For stable operation, it is necessary to proceed with Multi path configuration.

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)**



Conclusion

- JBOD is very cheap disk storage and we can get much storage capacity.
- But hard to manage. No manage software and solutions.
- JBOD power consumption is hight than tape storage, but it is very reasonable.
- We will keep testing JBOD Storage repeatedly with Alice-EOS.

Thank you