Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

HPSS Operation Updates Spring, 2023

Tim Chou

tchou@bnl.gov

May 4th, 2023

SDCC, BNL

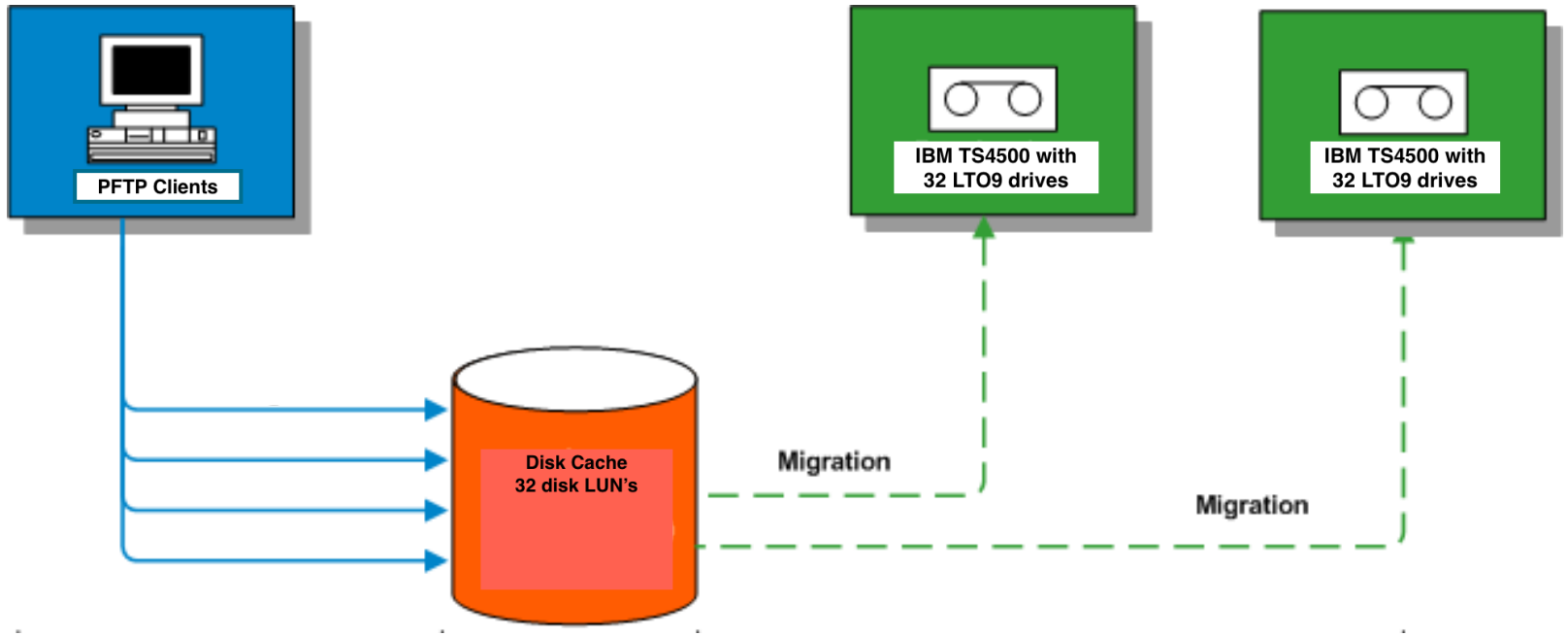
@BrookhavenLab

HPSS Upgrade to 8.3.20 from 8.3.10

- LTO9 support
- Movers
- Batch servers
- Lustre HPSS CopyTool
- Gateway servers
- Tape library controller server
- Core server
- PFTP/HSI clients

sPhenix Tape Storage Configurations

- Pftp and HSI clients.
- Data movers
- Disk cache
- Tape libraries
- ✓ Sustain10GB/sec



Tape Achive considerations

- Tape libraries
  - Tape slot count
  - Number of tape drives
  - Robotics mount speed
  - Floor footprint, number of silo frames
  - Robotics redundancy with fail-over features
- Disk cache
  - SSD, Disk array with controllers or JBOD
  - Data I/O throughput and capacity
  - Data path redundancy
- Movers
  - Network connections and redundancy
  - Fiber Channel connections and multipath

LTO9 Support on HPSS

- All LTO9 tape requires initialization, each takes 25 ~ 50 minutes
- Tape initializations must be done using LTO9 drives
- BNL HPSS team developed software to automate LTO9 tape initialization
  - With 64 LTO9 drives, 1,200 tapes can be initialized in 24 hours

Tape Libraries

Two units of IBM TS4500, Each unit has…

- Two robotic accessors with fail-over features
  - Each accessor has two grippers
- 8-frame, 8,806 tape slots
- 32 LTO9 tape drives
- To ensure equal usage on two libraries
  - Each tape volume creation is alternated across two libraries one by one, so data injections will be evenly distributed on two libraries

Tape Libraries - continued

- Two library units double the mount counts
- 565.2 PB requires 32,000 slots
- Two more library units to be added in 2024

Disk Cache for sPhenix

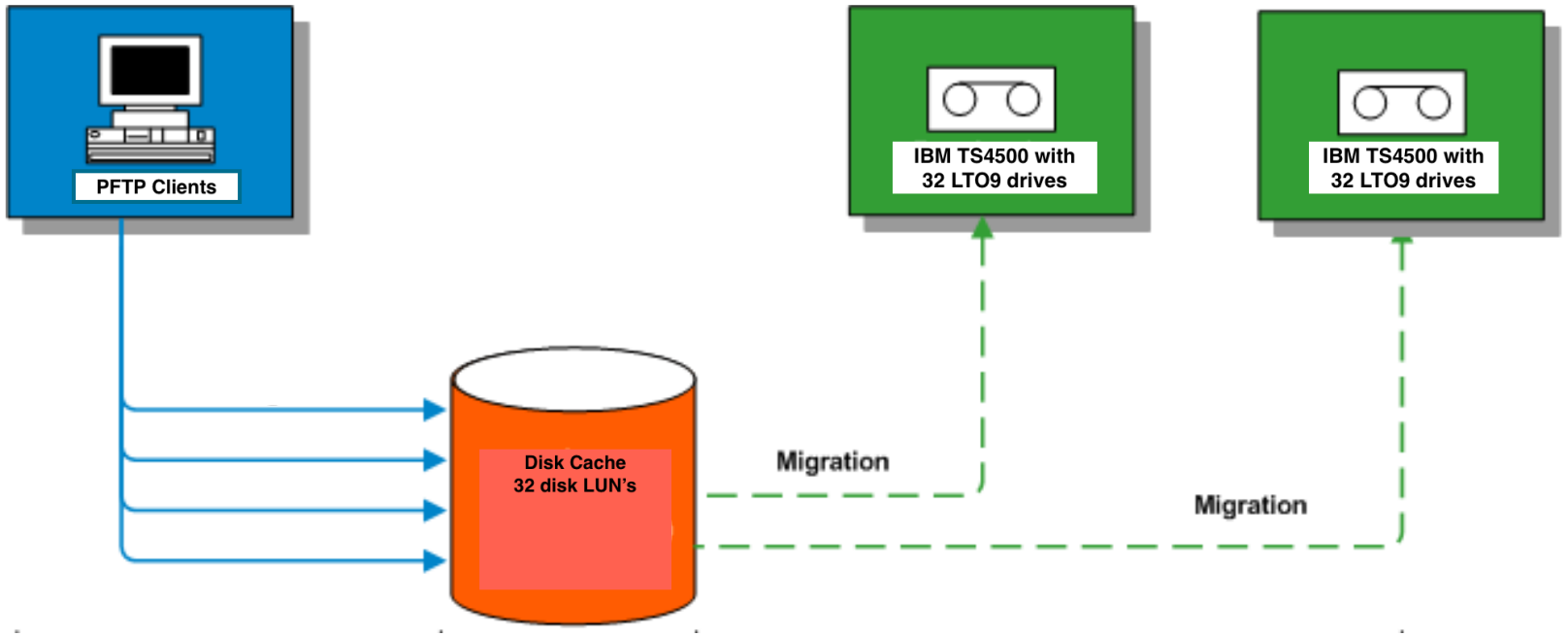
To meet 10 GB/sec, Disk Cache needs 20GB/sec throughput

- Three units of NetApp HDD array
  - SSD price is high
  - JBOD is slow after disabling buffering (required by application)
  - Cache size should hold multiple data sets for better data colocations on tape
- To maximize controller throughput, 114 HDD drives are needed
  - 11 RAID-6 LUN's (8+2) plus 4 global hot spares
- 33 disk LUN's total
  - 1,984TB total disk cache

Movers

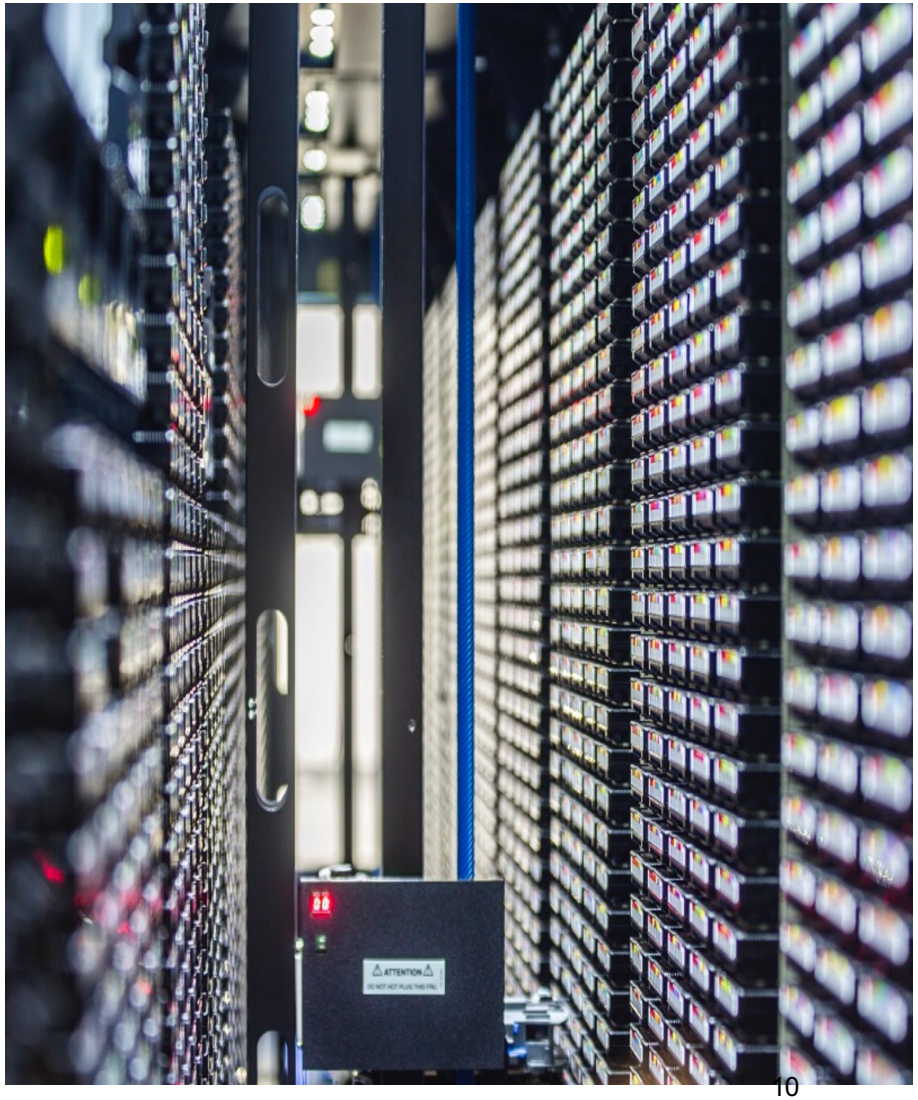
Four data movers, each connects…

- 200GbE (100GbE x 2 LACP)
  - 8 100Gb connections total
- LTO9 drives x 16
  - 64 LTO9 drives total
- Disk LUN's x 8
  - 32 disk LUN's total
  - 1,984TB total disk cache



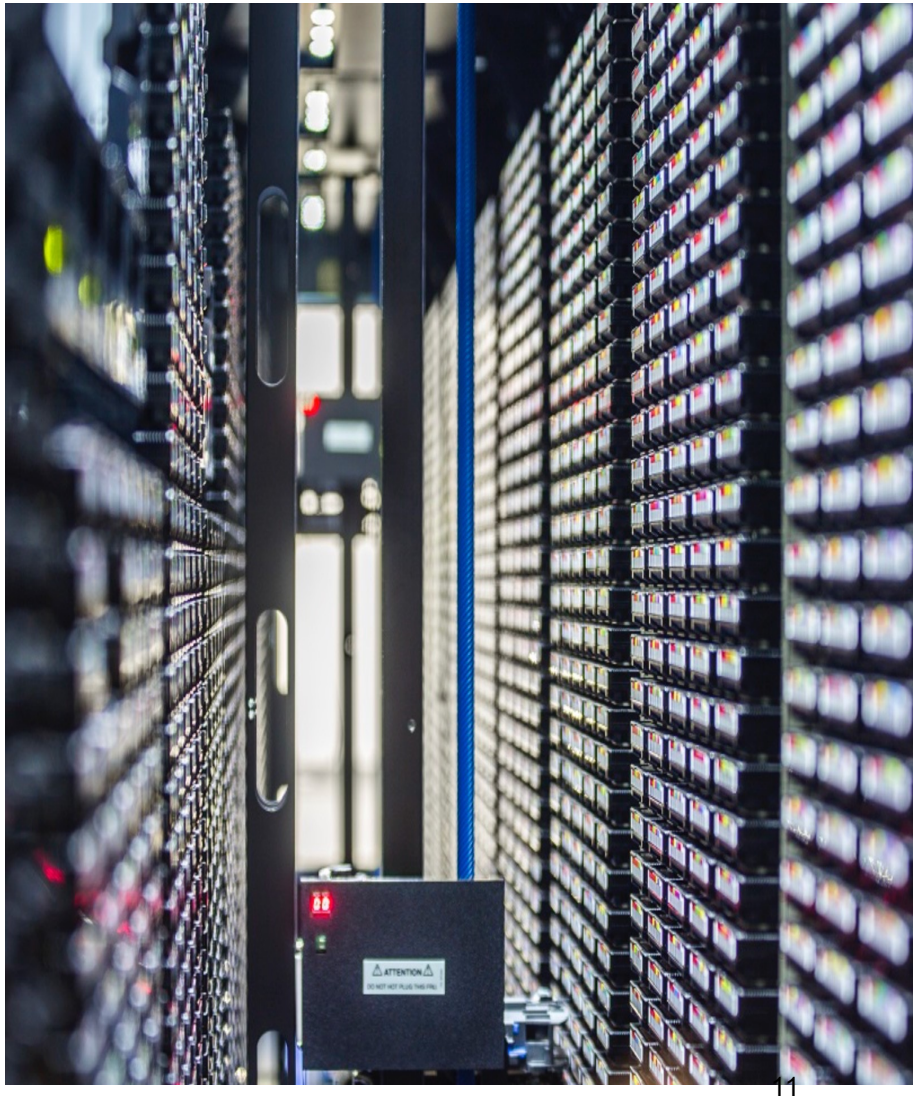
Tape Mount Testing

- Mount 32 drives, 151 sec (4.72 sec/mount)
  - 762 mounts/hour on each library
- Dismount 32 drives, 168 sec (5.25 sec/dismount)
  - 640 dismounts/hour on each library
  - TS4500 automatically remap the home slot address of a mounted tape to a nearest physical slot.
    - This expedites the subsequent mounts of this loaded tape.
- 361 tapes can be swapped each hour
  - Dismount + Mount = Swap tapes
  - The highest mount rate observed in Atlas is 285/hour
- When tapes go to deeper tiers, it gets slower



Tape Mount Testing - continued

- Each robot has two grippers, fast tape access to the first two tiers
  - 7,044 out of 18,000 slots (126.8PB) are on the first two tiers in the two libraries
  - With our projected data patterns, the hot tapes are likely all in the fast tiers
  - Tapes with cold data will gradually move to deeper tiers



Tape Data Injection Testing

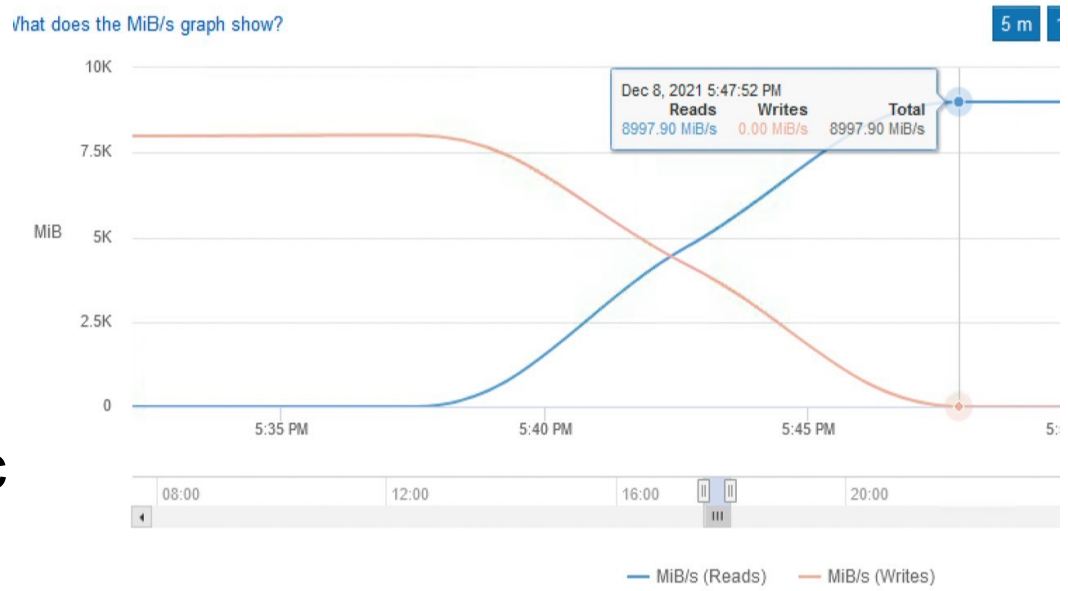
Concurrent injections to 64 LTO9 drives

- 64-drive total throughput 23,728.6 MiB/sec
- Average drive throughput 371.0 MiB/sec/drive
  - LTO9 drive spec is 400MB = 390MiB/sec/drive
  - All tapes are mounted and positioned before writes
  - File size is 20GB per file

Disk Cache Testing

32 disk LUN’s on 4 data movers

- 100% write throughput 24,454.5 MiB/sec
  - Data injection to all 32 LUN’s concurrently
  - 764.2MiB/sec per LUN (100% write)
  - Each LUN contains 8+2 HDD’s (RAID-6)
  - 32 Gb/sec FC connections
- 50% Read and 50% Write, 26,740 MiB/sec
- 100% Read, 25,780 MiB/sec



Data Mover Network Testing

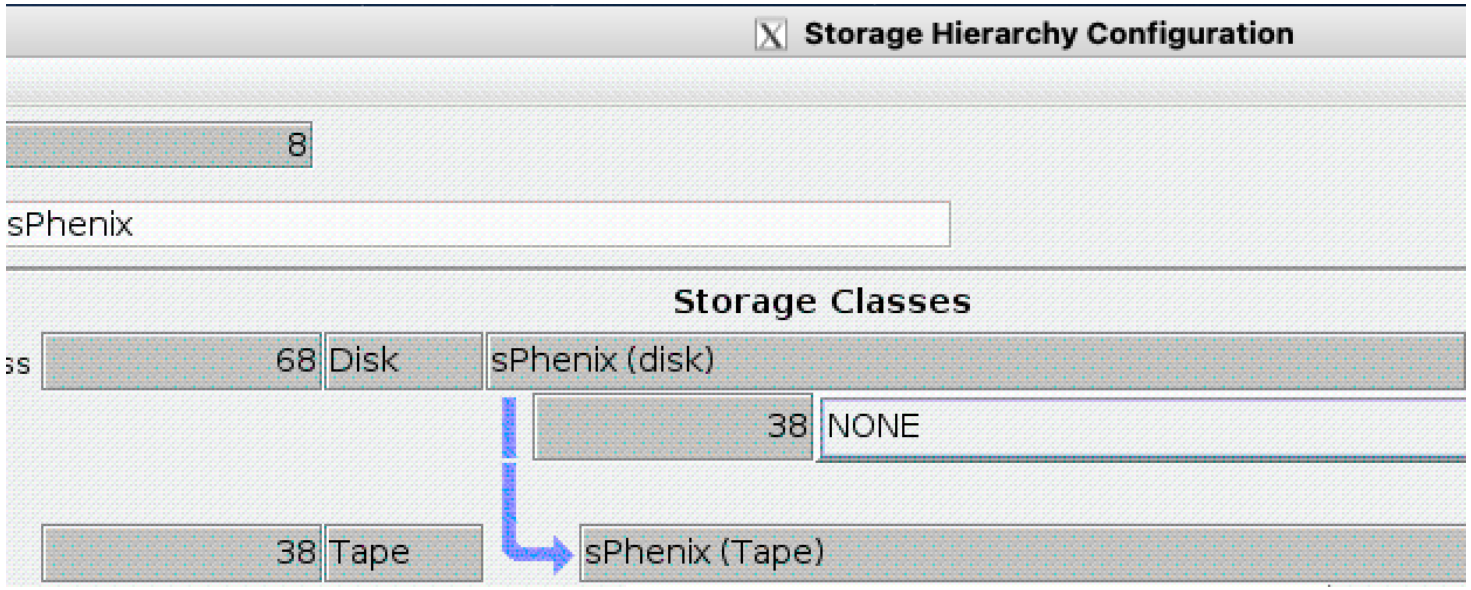
Each mover has two 100Gb ethernet connections(LACP), 4 Movers

- Total network throughput on 4 movers: 48,845 MiB/sec
  - Each mover transfer data to the other three movers with 10 connections
  - No disk and tape I/O involved

Tape service configuration

A dedicated Class Of Service for sPhenix

- Disk Storage Class
  - 1,984TB, 32 LUN's
- Tape Storage Class
  - LTO9 (18TB per tape)
  - Two IBM tape libraries
  - 17,612 slots
- pftp and HSI clients



Batch for Tape Staging

Staging requests are submitted to Batch application

- Staging requests are grouped by tapes
  - To minimize tape mounts
- Files on the tapes are read in tape position order
  - To minimize tape repositioning
- Call-back mechanism provided when a request completes

FILE STAGING REAL-TIME STATUS

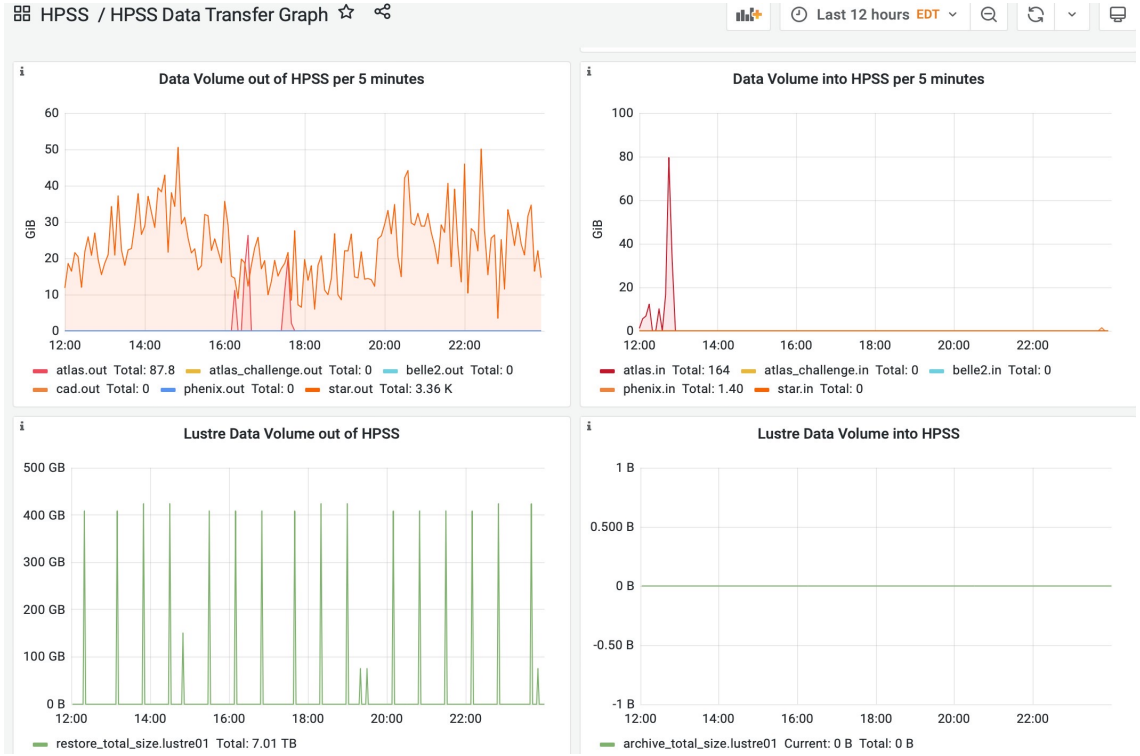
atlasdat

Tape Info	Tape ID	Files	Avg size	Status	Files staged	GB Staged	Avg MB/s	Files failed	Last staged	Mount Time	Drv Addr	Drv Type
Atlas Large LTO-7	A70574	2 / 268	475,889,685	Reading	1459	651.17	201.15		3-15 20:56:32	3-15 20:01:43 (00:54:58)	2,0,1,2	IBM LTO7
Atlas Large LTO-7	A70575	2 / 220	484,991,224	Reading	1491	665.99	207.92		3-15 20:55:57	3-15 20:01:51 (00:54:50)	2,3,1,0	IBM LTO7
Atlas Large LTO-7	A70578	2 / 280	465,058,091	Reading	1375	614.50	191.38		3-15 20:56:05	3-15 20:01:59 (00:54:42)	2,0,1,14	IBM LTO7
TOTAL:	3 Tapes	6 Files		Avg 200.15 MB/s/dr								

System Monitoring

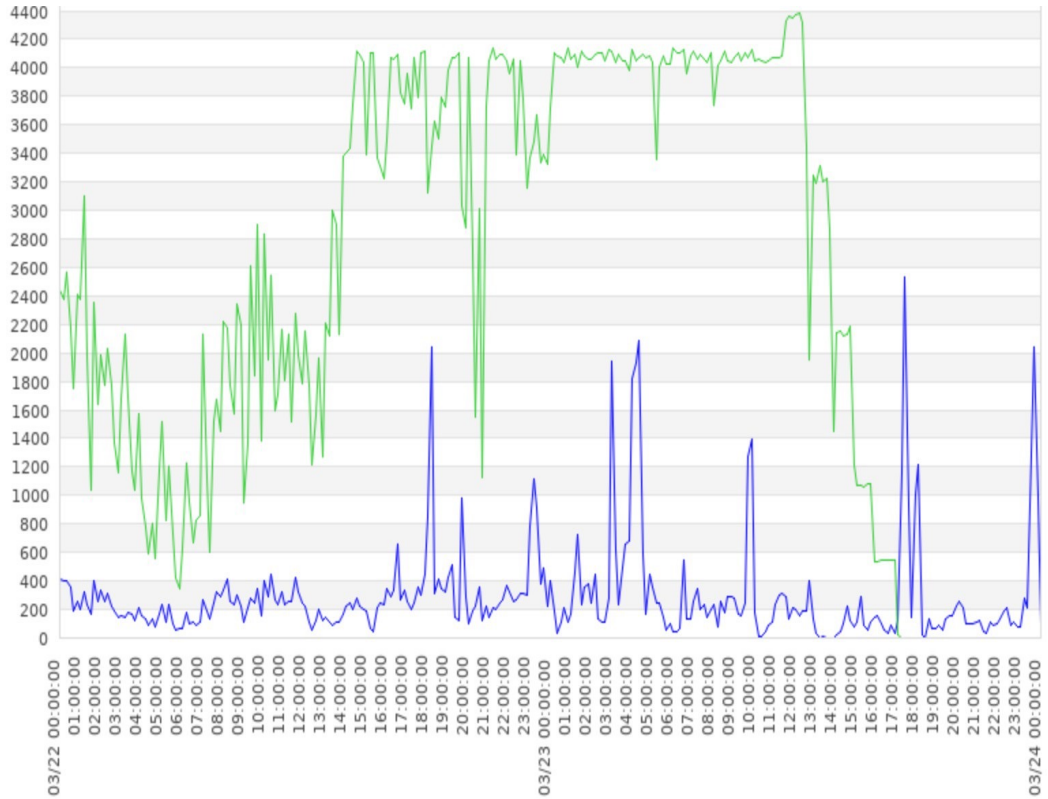
Grafana and MySQL DB

- Operational numbers such as network traffic, tape mounts, disk and tape usage … etc are monitored and recorded,
- Recorded numbers are displayed on Grafana



Star MDC 2023 (03/22 ~ 03/24)

- 48 hours of continuous data injection
- Injection rate stabilized at 4.1 GB/sec
- 557.5TB injected
- 14 LTO8 drives migrating concurrently
- All used LTO8 tapes reclaimed after MDC
- MDC has met the 4.0GB/sec requirement



Brookhaven National Laboratory

Thank you!

Q & A…