Management of Cost Effective Mass Storage Environments

BROOKHAVEN NATIONAL LABORATORY | Scientific Data and Computing Center

70 YEARS OF DISCOVERY

A CENTURY OF SERVICE

March 21st, 2018

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

Scientific Data

Exponential Growth

Preserve for decades

Non-compressible

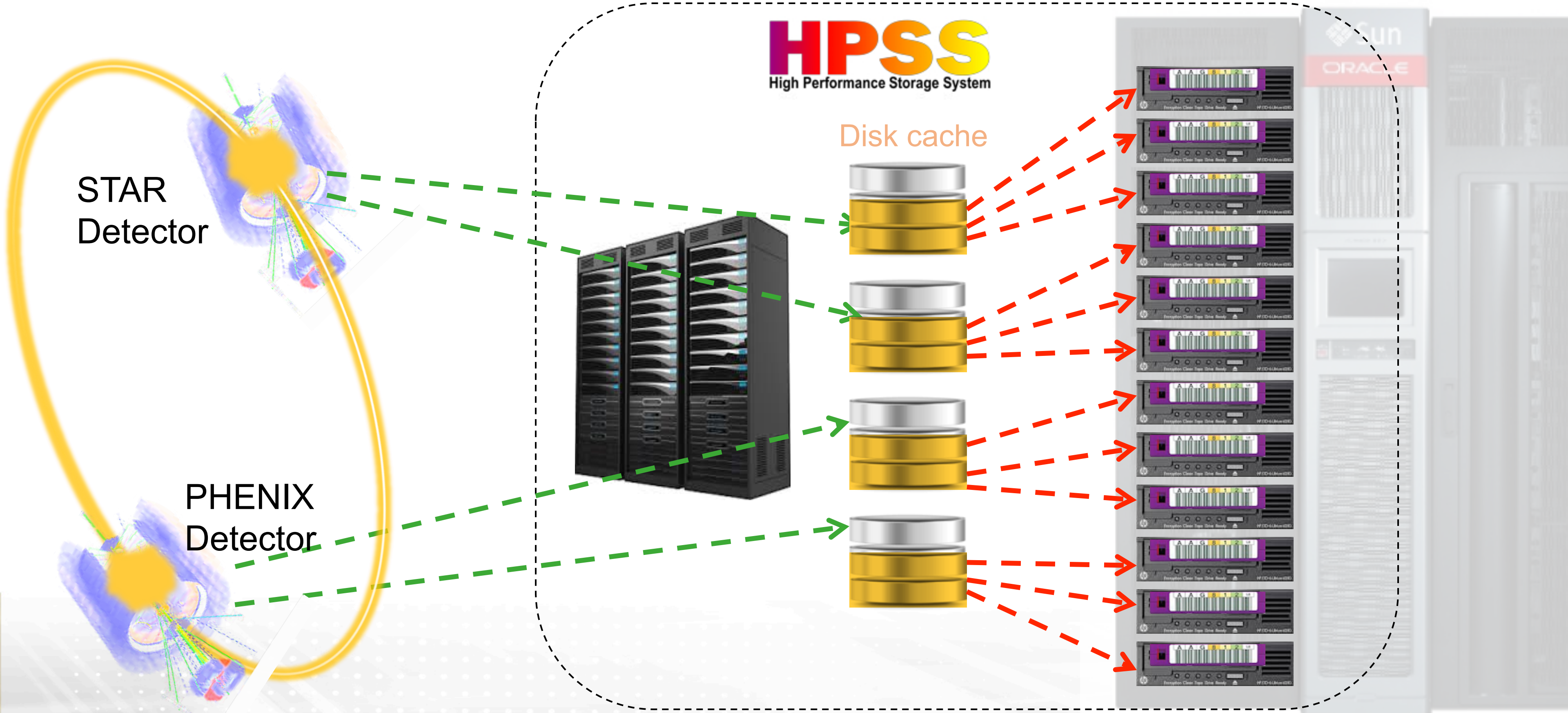
Retention Policy

Current practice is to retain the data, and the ability to retrieve them indefinitely.

High Throughput Data Archiving

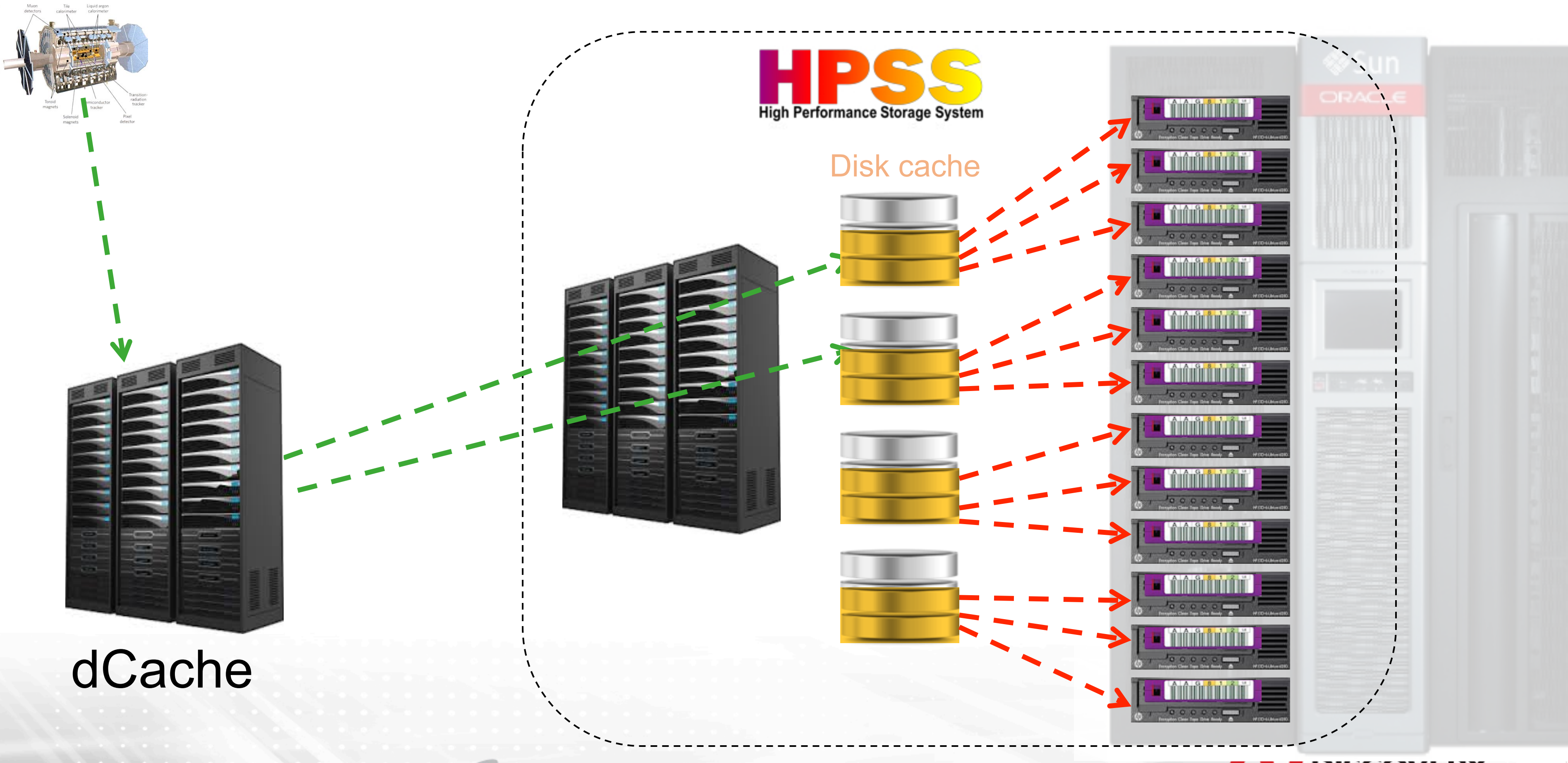
RHIC Experiment data directly go to tape storage (primary)

RHIC detectors:



High Throughput Data Archiving

ATLAS Experiment data goes to dCache and then send to tape storage.



Archival Storage

Provides permanent data storage for all RHIC experiment

- STAR, PHENIX,PHOBOS and BRAHMS
- RAW and DST
- User Data (No Personal data, no PII allowed)

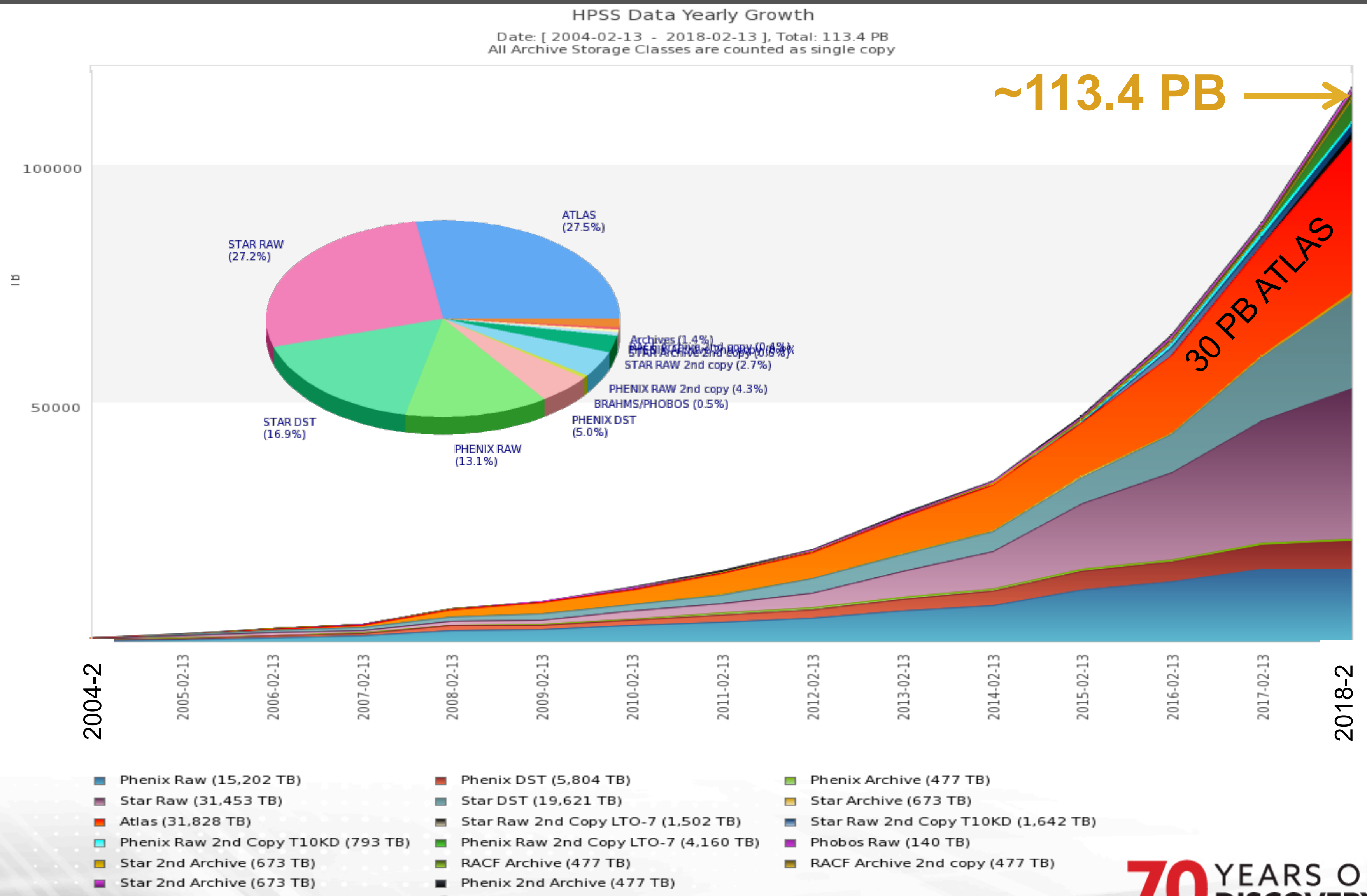
Archival storage for C-AD Operational Logger Data

Serves as LHC ATLAS Tier-1 for the US

- Secondary data storage for fraction of data (~23%).

Serves as Belle-2 Tier-1 (New)

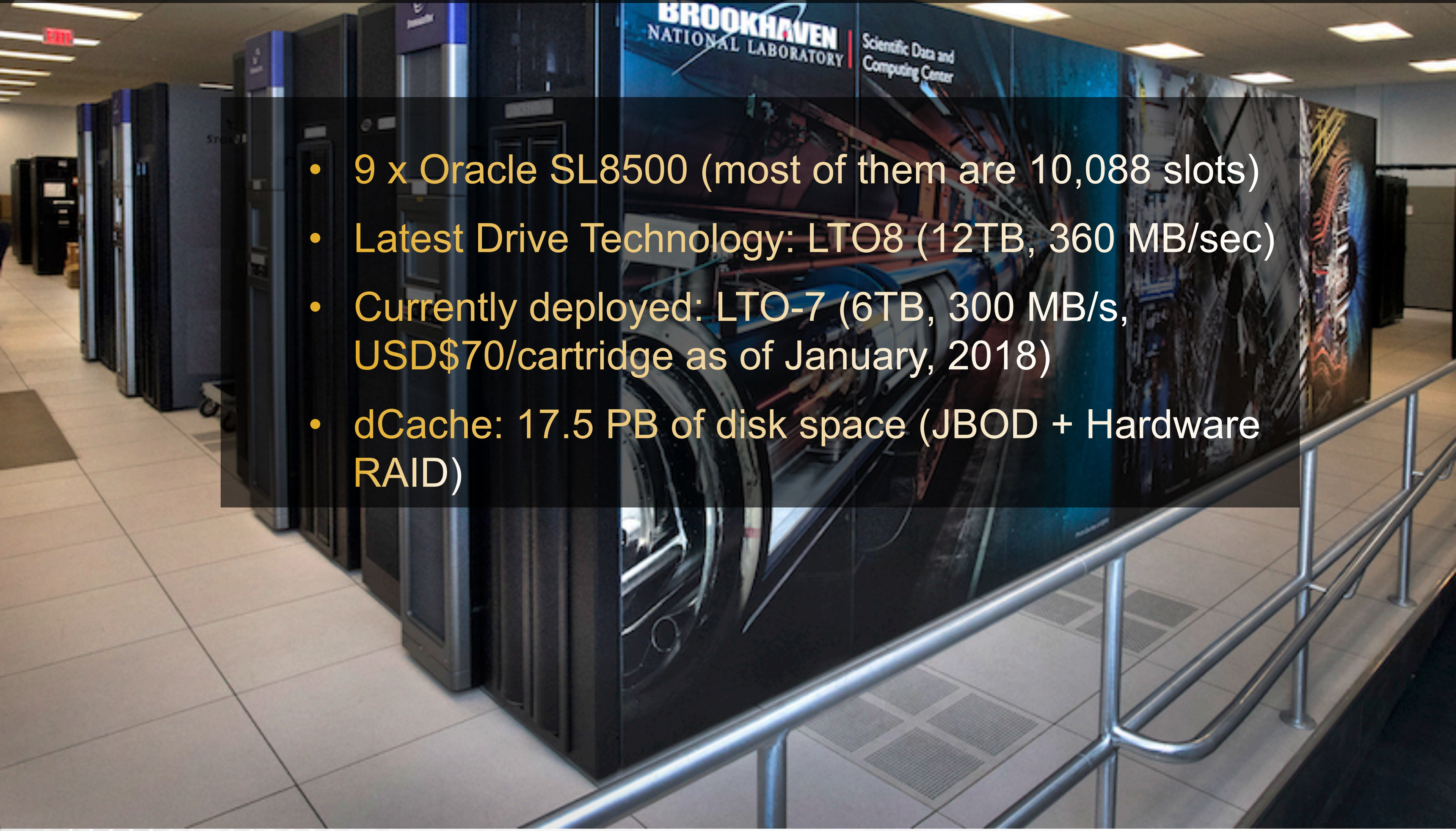
Archived Data



~113.4 PB →

30 PB ATLAS

Mass Storage on Tape

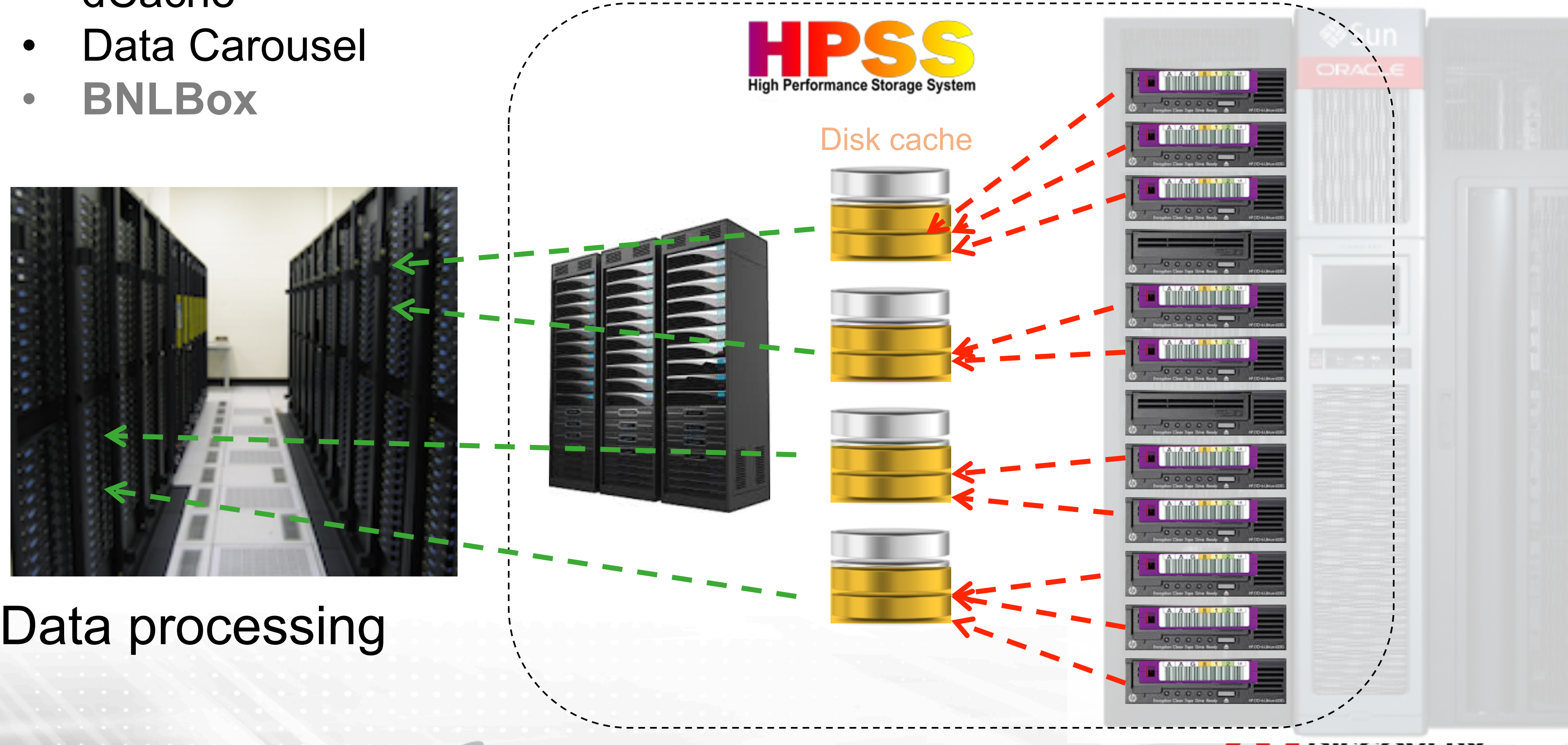

EXIT
BROOKHAVEN NATIONAL LABORATORY
Scientific Data and Computing Center

- 9 x Oracle SL8500 (most of them are 10,088 slots)
- Latest Drive Technology: LTO8 (12TB, 360 MB/sec)
- Currently deployed: LTO-7 (6TB, 300 MB/s, USD$70/cartridge as of January, 2018)
- dCache: 17.5 PB of disk space (JBOD + Hardware RAID)

High Throughput Data Archiving

Retrieving data on demand

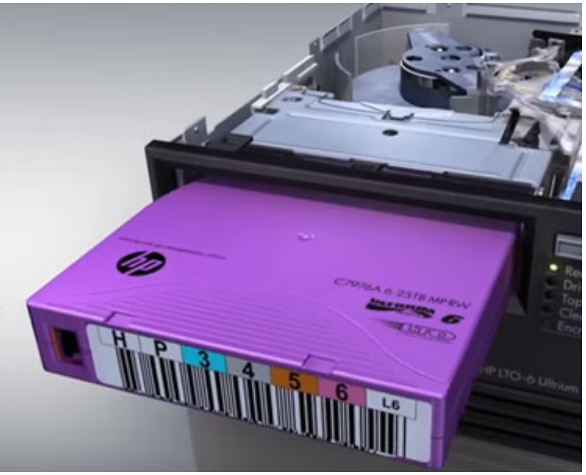
- dCache
- Data Carousel
- **BNLBox**



Tape Storage - Usage

Tape Usages

In 2017

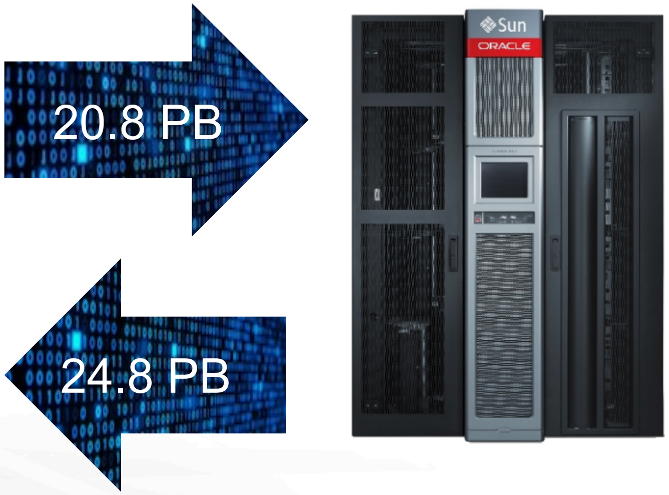


Archived to tape:

19,412,702 files – Average 53,185 files/day
20.8 PB – Average 58.4 TB / day

Restored from tape:

11,693,141 files - Average 32,036 files/day,
24.8 PB - Average 69.5 TB/day



20.8 PB

24.8 PB

JBOD Management

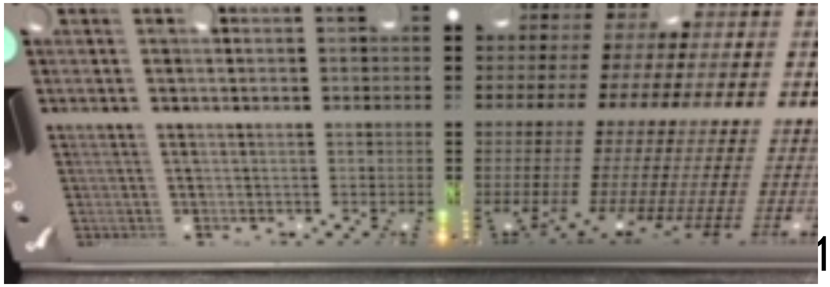
- Instead of hardware RAID’s, we have deployed many SAS JBOD systems that cost approximately 50% less than hardware RAID’s.
- The JBOD’s were configured as RAID-6 using MDADM (RedHat 7 or RedHat 6).
- The JBOD’s were configured with redundant SAS HBA connections (12 Gbit X 4 channels) using Multipath drivers for failover.

JBOD Management

Drive failure…

- Control LED on slot 28

```
  sg_ses --index 28 --set 2:1:1  /dev/sg6  // enalbe LED
  sg_ses –index -1 –clear 2:1:1 /dev/sg6 // disable all LEDs
  sg_ses -ee|grep “slot”  //3:5:1 amber LED ; 2:1:1 flashy blue LED
          fault  [Device slot] [3:5:1]  //solid amber LED
          ident  [Device slot] [2:1:1]      //flashing blue LED
```



1.

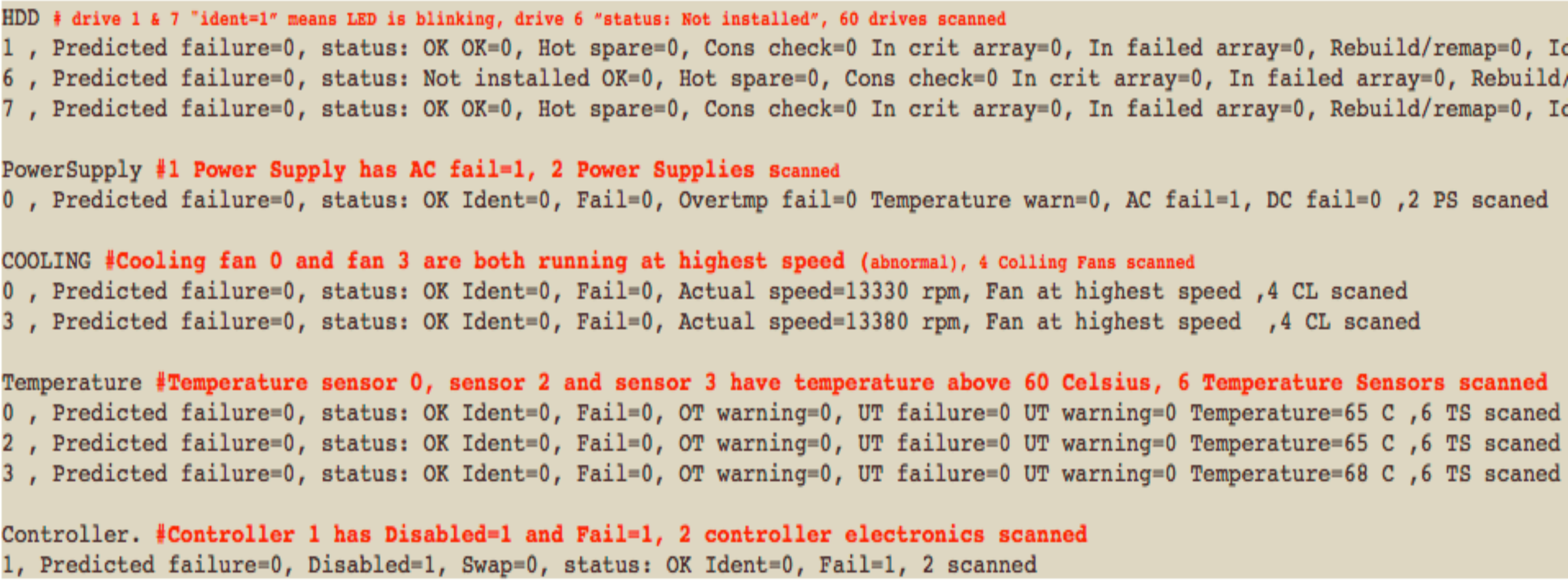
The slot numbers on sg_ses start from 0.

- Use MDADM commands to remove, add and rebuild the disk array

Monitor components on chassis

Besides disk drives, all components on JBOD's can be monitored…

- sg_ses -p0x2 /dev/sg6 (enclosure status/control)
  - Query JBOD enclosure for status and control settings



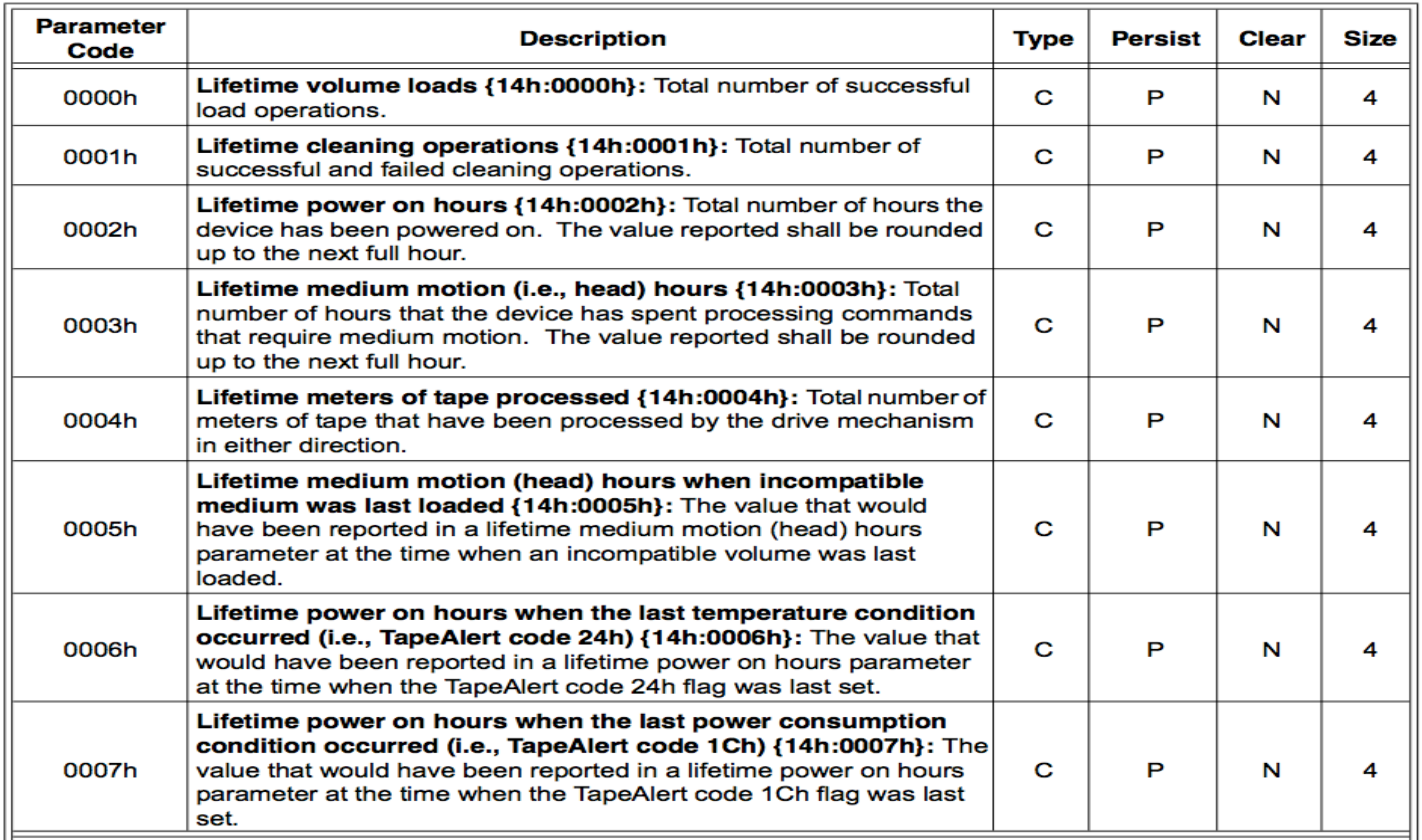
```
HDD # drive 1 & 7 "ident=1" means LED is blinking, drive 6 "status: Not installed", 60 drives scanned
1 , Predicted failure=0, status: OK OK=0, Hot spare=0, Cons check=0 In crit array=0, In failed array=0, Rebuild/remap=0, Id
6 , Predicted failure=0, status: Not installed OK=0, Hot spare=0, Cons check=0 In crit array=0, In failed array=0, Rebuild/
7 , Predicted failure=0, status: OK OK=0, Hot spare=0, Cons check=0 In crit array=0, In failed array=0, Rebuild/remap=0, Id

PowerSupply #1 Power Supply has AC fail=1, 2 Power Supplies scanned
0 , Predicted failure=0, status: OK Ident=0, Fail=0, Overtmp fail=0 Temperature warn=0, AC fail=1, DC fail=0 ,2 PS scaned

COOLING #Cooling fan 0 and fan 3 are both running at highest speed (abnormal), 4 Colling Fans scanned
0 , Predicted failure=0, status: OK Ident=0, Fail=0, Actual speed=13330 rpm, Fan at highest speed ,4 CL scaned
3 , Predicted failure=0, status: OK Ident=0, Fail=0, Actual speed=13380 rpm, Fan at highest speed  ,4 CL scaned

Temperature #Temperature sensor 0, sensor 2 and sensor 3 have temperature above 60 Celsius, 6 Temperature Sensors scanned
0 , Predicted failure=0, status: OK Ident=0, Fail=0, OT warning=0, UT failure=0 UT warning=0 Temperature=65 C ,6 TS scaned
2 , Predicted failure=0, status: OK Ident=0, Fail=0, OT warning=0, UT failure=0 UT warning=0 Temperature=65 C ,6 TS scaned
3 , Predicted failure=0, status: OK Ident=0, Fail=0, OT warning=0, UT failure=0 UT warning=0 Temperature=68 C ,6 TS scaned

Controller. #Controller 1 has Disabled=1 and Fail=1, 2 controller electronics scanned
1, Predicted failure=0, Disabled=1, Swap=0, status: OK Ident=0, Fail=1, 2 scanned
```

Tape device monitoring…

Query Tape Device Usage and errors

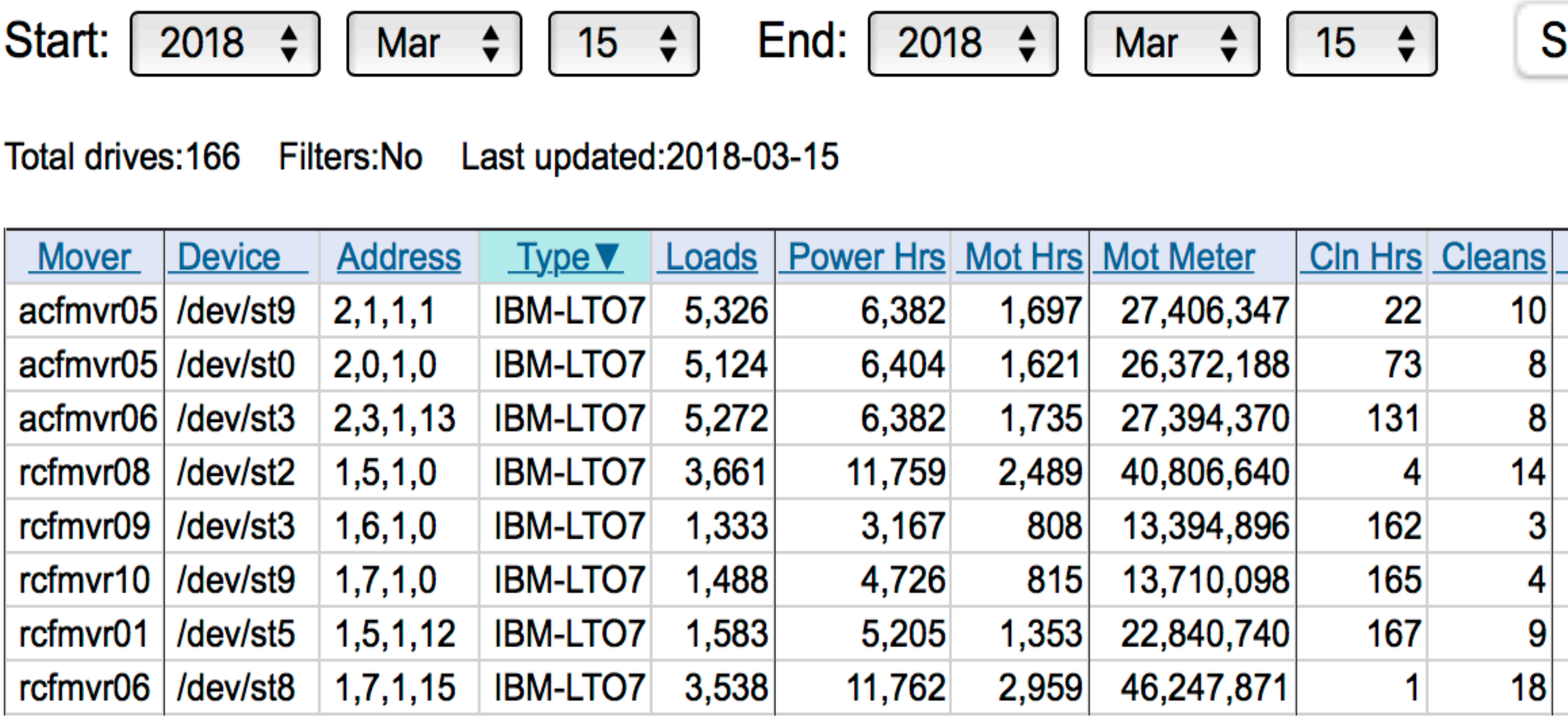
- Use SCSI command “Log Sense, page 0x14”

Table 171 — LP14h: Device Statistics log parameter codes (part 1 of 4)



Parameter Code	Description	Type	Persist	Clear	Size
0000h	**Lifetime volume loads {14h:0000h}:** Total number of successful load operations.	C	P	N	4
0001h	**Lifetime cleaning operations {14h:0001h}:** Total number of successful and failed cleaning operations.	C	P	N	4
0002h	**Lifetime power on hours {14h:0002h}:** Total number of hours the device has been powered on. The value reported shall be rounded up to the next full hour.	C	P	N	4
0003h	**Lifetime medium motion (i.e., head) hours {14h:0003h}:** Total number of hours that the device has spent processing commands that require medium motion. The value reported shall be rounded up to the next full hour.	C	P	N	4
0004h	**Lifetime meters of tape processed {14h:0004h}:** Total number of meters of tape that have been processed by the drive mechanism in either direction.	C	P	N	4
0005h	**Lifetime medium motion (head) hours when incompatible medium was last loaded {14h:0005h}:** The value that would have been reported in a lifetime medium motion (head) hours parameter at the time when an incompatible volume was last loaded.	C	P	N	4
0006h	**Lifetime power on hours when the last temperature condition occurred (i.e., TapeAlert code 24h) {14h:0006h}:** The value that would have been reported in a lifetime power on hours parameter at the time when the TapeAlert code 24h flag was last set.	C	P	N	4
0007h	**Lifetime power on hours when the last power consumption condition occurred (i.e., TapeAlert code 1Ch) {14h:0007h}:** The value that would have been reported in a lifetime power on hours parameter at the time when the TapeAlert code 1Ch flag was last set.	C	P	N	4

Display tape drive Status



Start: 2018 ▲▼ Mar ▲▼ 15 ▲▼ End: 2018 ▲▼ Mar ▲▼ 15 ▲▼ S

Total drives:166 Filters:No Last updated:2018-03-15

Mover	Device	Address	Type▼	Loads	Power Hrs	Mot Hrs	Mot Meter	Cln Hrs	Cleans
acfmvr05	/dev/st9	2,1,1,1	IBM-LTO7	5,326	6,382	1,697	27,406,347	22	10
acfmvr05	/dev/st0	2,0,1,0	IBM-LTO7	5,124	6,404	1,621	26,372,188	73	8
acfmvr06	/dev/st3	2,3,1,13	IBM-LTO7	5,272	6,382	1,735	27,394,370	131	8
rcfmvr08	/dev/st2	1,5,1,0	IBM-LTO7	3,661	11,759	2,489	40,806,640	4	14
rcfmvr09	/dev/st3	1,6,1,0	IBM-LTO7	1,333	3,167	808	13,394,896	162	3
rcfmvr10	/dev/st9	1,7,1,0	IBM-LTO7	1,488	4,726	815	13,710,098	165	4
rcfmvr01	/dev/st5	1,5,1,12	IBM-LTO7	1,583	5,205	1,353	22,840,740	167	9
rcfmvr06	/dev/st8	1,7,1,15	IBM-LTO7	3,538	11,762	2,959	46,247,871	1	18

More discussions?

We can have further discussion if necessary…

- Why use Tape?
  - Reliability, life expectancy, cost…
  - Advantages and Disadvantages
- Why use JBOD?
  - Cost, scalability, monitoring…
  - Advantages and Disadvantage
- In-house storage VS Cloud
- The future of archival storages

EMAIL tchou@bnl.gov

Thank
YOU