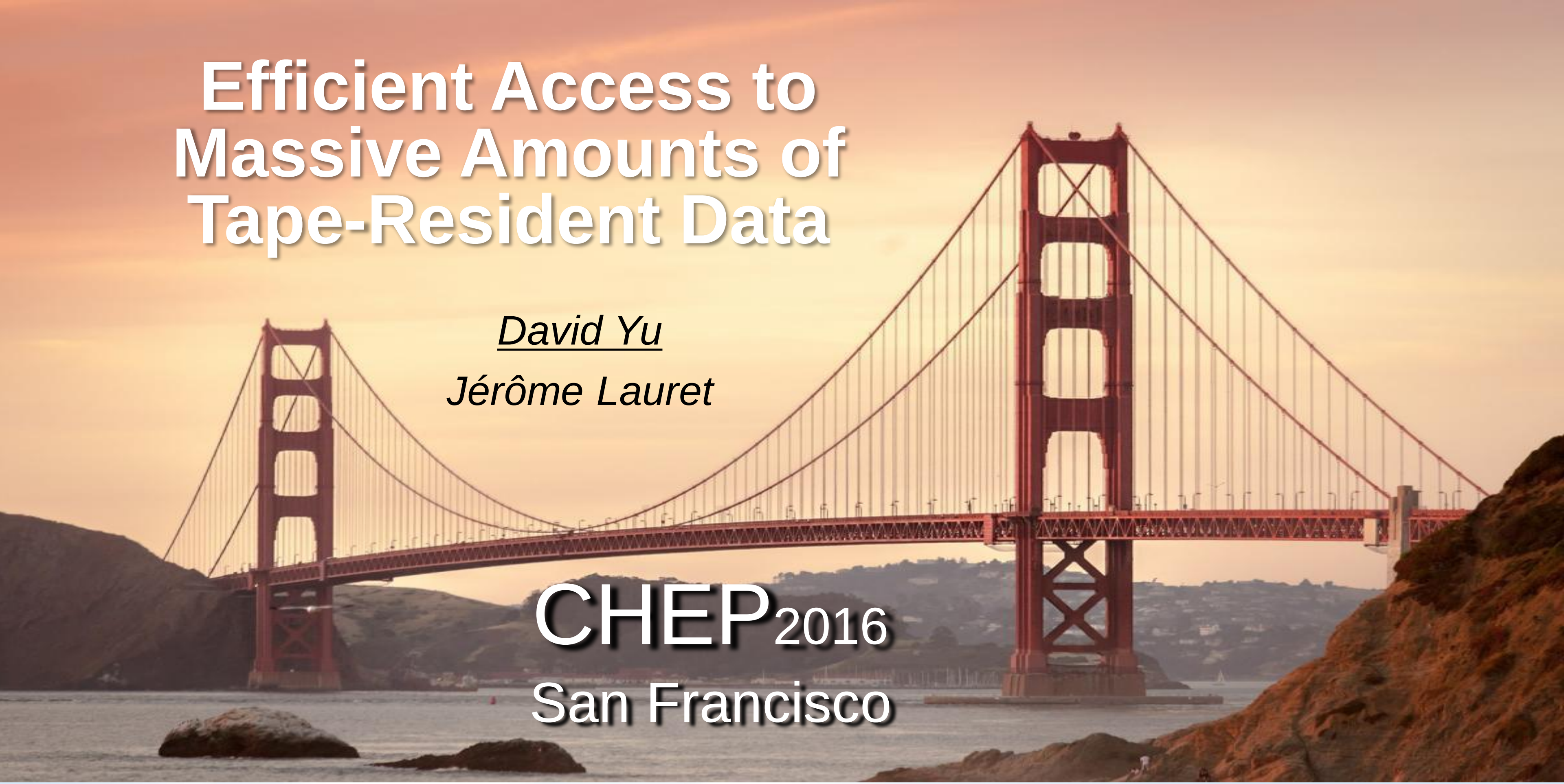


Efficient Access to Massive Amounts of Tape-Resident Data

David Yu

Jérôme Lauret

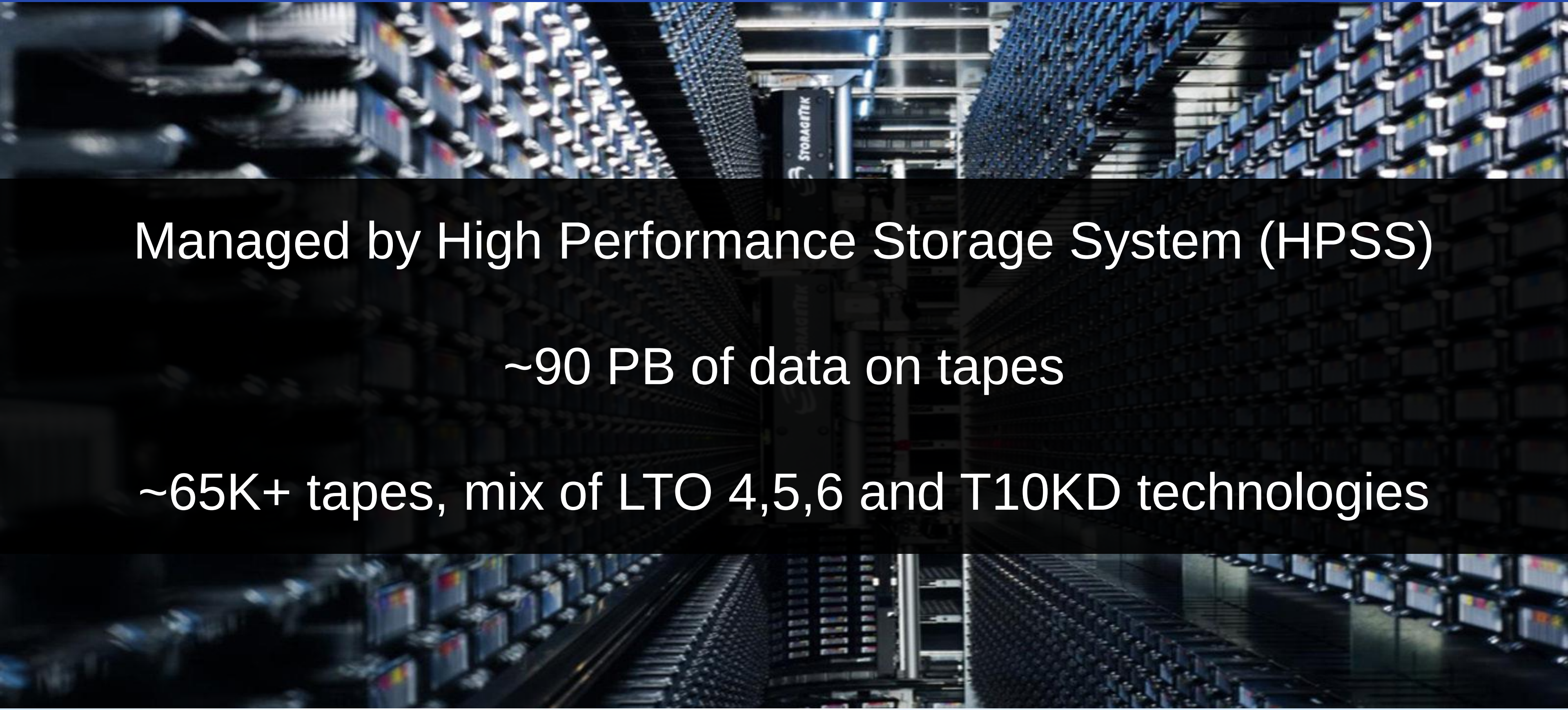
CHEP2016

San Francisco

Brookhaven Science Associates

BROOKHAVEN NATIONAL LABORATORY

BNL Mass Storage



Managed by High Performance Storage System (HPSS)

~90 PB of data on tapes

~65K+ tapes, mix of LTO 4,5,6 and T10KD technologies

Tape Technology

Why are we using tapes?

Tape is reliable, energy efficient, low cost (per GB), and fast!

One LTO-7 can reach 300 MB/s, Bit Error Rate 1 x 10^{19}, last for 30 years

We just restored some 15 years tapes (9940B), 100% successful.

Tape is great for data archiving, but it's sequential access!

Randomly restoring files from massive amount of tapes degrades the read performance primarily due to frequent tape mounts, forwards and rewinds

We have an in house developed system, called ERADAT, to optimize the tape mounts, tape reads, and resource control.
It also provides performance monitoring as well as statistics.

ERADAT

- A scheduler system, originally based on a software from Oak Ridge National Lab, developed in the early 2000s.
- After some major modifications and enhancements, ERADAT now provides advanced HPSS resource management, priority queuing, resource sharing, web-browser visibility of real-time staging activities and advanced real-time statistics and graphs.
- An interface between HPSS and other applications such as the locally developed Data Carousel providing fair resource-sharing policies and related capabilities.

Unmanaged Resources (Tape and Drives)

- Tape drives are first come first serve…
- Users will fight for tape drives
- Waiting time becomes unpredictable
- Unable to prioritize tasks

ERADAT

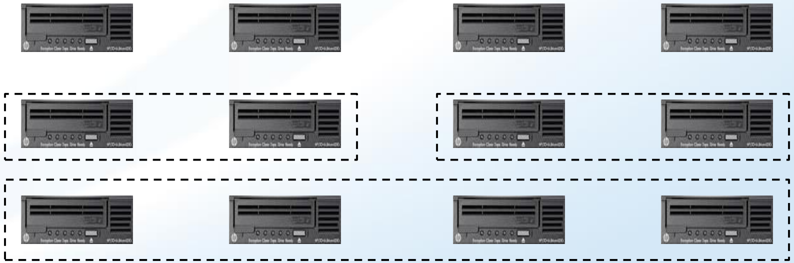
How to address the problems:
Repeat mounts, forwards, rewinds…

Optimize the performance

- Reduce Tape Mounts
- Reduce Rewinds
- Processing Order (FIFO…)

Resource Management

- Resource Allocations
- Dedicated resources
- Shared resources



ERADAT Features

- Staging optimization
- Tape selection orders:

  FIFO, LIFO, and “By Demand”
- Priority Staging
- Resource Management
  - Resource guaranteed
  - Resource sharing
- Resource Allocation Oversubscription
- Drive-generation Oversubscription

- Multi-level real-time debug log on/off switch
- Sync or async (callback) option
- Real-time configurable auto-retry
- Advanced Thread control

ERADAT Features (2)

Real-time Monitoring Tools:

- Web-based Control Panel
- Performance Graphs/Reports
  - Staging Activity Graph
  - Tape staging performance report
  - Drive staging performance report
- Staging suspension/resume control
  - Drive-generation level
  - Global level (lock all)
- Resource Lock: tape and drive (HPSS level)
- Auto-detect LSM down, bypassing offline LSM

ERADAT Features (3)

All interfaces, graphs and reports are HTML based, works on any web-browsers (Any OS, any newer web-browsers)

Tested Environment:

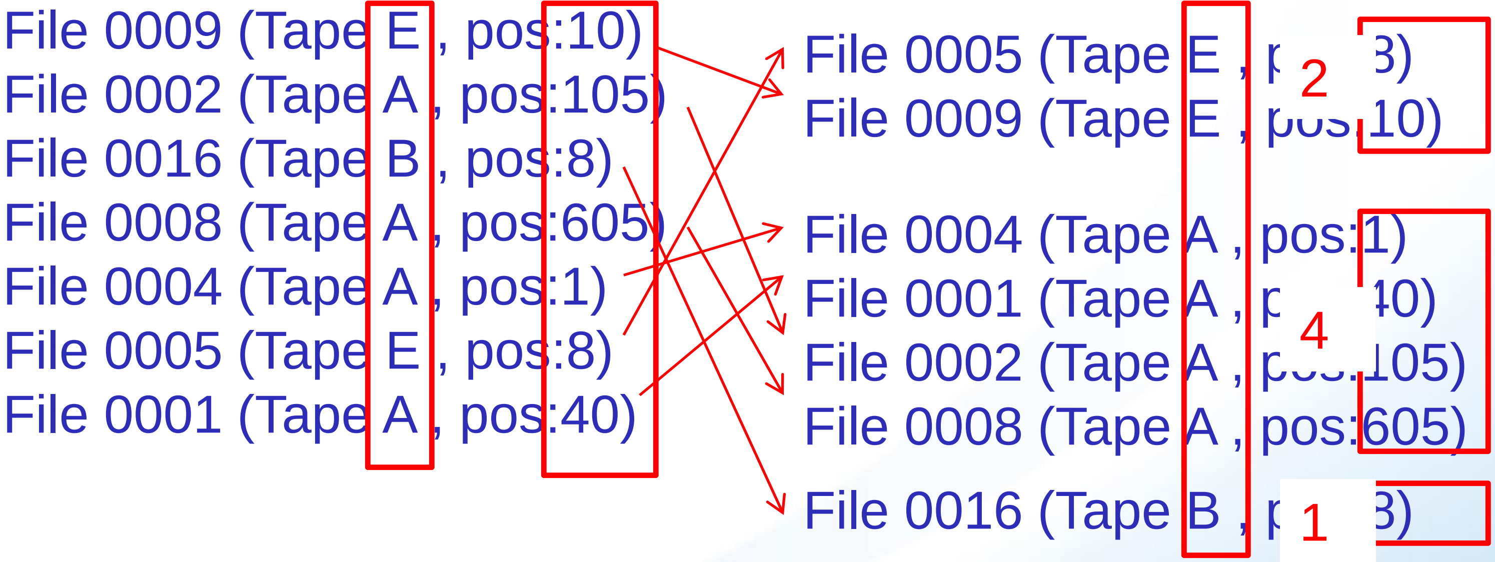
- Mac OSX
- Linux
- Windows
- IOS
- Android

Tested Web-Browsers:

- Safari
- Firefox
- Chrome
- IE (PC only)

Staging Optimization

Requests are first aggregated by tape-id – reduce mounts
Sorted by position on the tape – reduce rewinds.
If “By Demand”, sort the tape list by # of reqs.



File 0009 (Tape E , pos:10)
File 0002 (Tape A , pos:105)
File 0016 (Tape B , pos:8)
File 0008 (Tape A , pos:605)
File 0004 (Tape A , pos:1)
File 0005 (Tape E , pos:8)
File 0001 (Tape A , pos:40)

File 0005 (Tape E , p 2 3)
File 0009 (Tape E , pos:10)

File 0004 (Tape A , pos:1)
File 0001 (Tape A , p 4 40)
File 0002 (Tape A , pos:105)
File 0008 (Tape A , pos:605)

File 0016 (Tape B , p 1 3)

By FIFO: Tape mount order: **E, A, B**
By Demand: Tape mount order: **A, E, B**

Staging Optimization

FIFO vs “By Demand”

Which one is more efficient?

Priority Staging

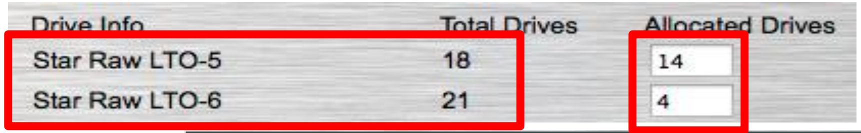
File 0009 (Tape E , pos:10)
File 0002 (Tape A , pos:105)
File 0016 (Tape B , pos:8)
File 0008 (Tape A , pos:605)
File 0004 (Tape A , pos:1)
File 0005 (Tape E , pos:8)
File 0001 (Tape A , pos:40)

File 0022 (Tape C , pos:48)
File 0021 (Tape C , pos:40)

By Demand: Tape mount order: **C**, **A, E, B**
By FIFO: Tape mount order: **C**, **E, A, B**

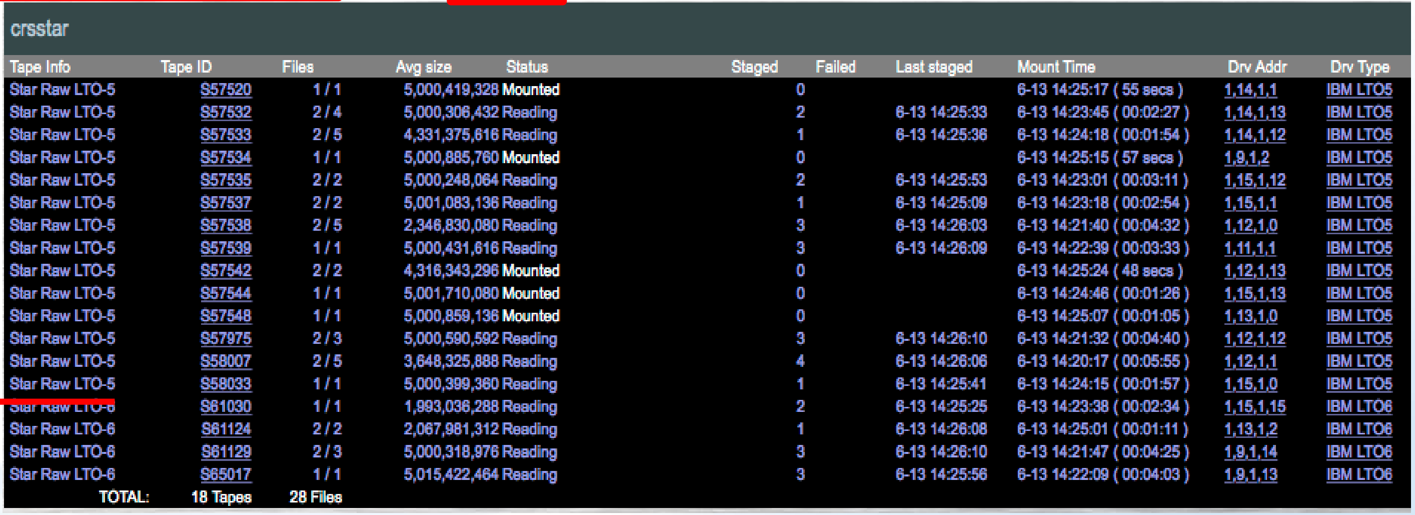
Resource Management

Tape drive resource allocation should be under total control, to avoid a service interruption from drive being taken by other process.



Drive Info	Total Drives	Allocated Drives
Star Raw LTO-5	18	14
Star Raw LTO-6	21	4

Resource guaranteed



crsstar

Tape Info	Tape ID	Files	Avg size	Status	Staged	Failed	Last staged	Mount Time	Drv Addr	Drv Type
Star Raw LTO-5	S57520	1 / 1	5,000,419,328	Mounted	0			6-13 14:25:17 (55 secs)	1,14,1,1	IBM LTO5
Star Raw LTO-5	S57532	2 / 4	5,000,306,432	Reading	2		6-13 14:25:33	6-13 14:23:45 (00:02:27)	1,14,1,13	IBM LTO5
Star Raw LTO-5	S57533	2 / 5	4,331,375,616	Reading	1		6-13 14:25:36	6-13 14:24:18 (00:01:54)	1,14,1,12	IBM LTO5
Star Raw LTO-5	S57534	1 / 1	5,000,885,760	Mounted	0			6-13 14:25:15 (57 secs)	1,9,1,2	IBM LTO5
Star Raw LTO-5	S57535	2 / 2	5,000,248,064	Reading	2		6-13 14:25:53	6-13 14:23:01 (00:03:11)	1,15,1,12	IBM LTO5
Star Raw LTO-5	S57537	2 / 2	5,001,083,136	Reading	1		6-13 14:25:09	6-13 14:23:18 (00:02:54)	1,15,1,1	IBM LTO5
Star Raw LTO-5	S57538	2 / 5	2,346,830,080	Reading	3		6-13 14:26:03	6-13 14:21:40 (00:04:32)	1,12,1,0	IBM LTO5
Star Raw LTO-5	S57539	1 / 1	5,000,431,616	Reading	3		6-13 14:26:09	6-13 14:22:39 (00:03:33)	1,11,1,1	IBM LTO5
Star Raw LTO-5	S57542	2 / 2	4,316,343,296	Mounted	0			6-13 14:25:24 (48 secs)	1,12,1,13	IBM LTO5
Star Raw LTO-5	S57544	1 / 1	5,001,710,080	Mounted	0			6-13 14:24:46 (00:01:26)	1,15,1,13	IBM LTO5
Star Raw LTO-5	S57548	1 / 1	5,000,859,136	Mounted	0			6-13 14:25:07 (00:01:05)	1,13,1,0	IBM LTO5
Star Raw LTO-5	S57975	2 / 3	5,000,590,592	Reading	3		6-13 14:26:10	6-13 14:21:32 (00:04:40)	1,12,1,12	IBM LTO5
Star Raw LTO-5	S58007	2 / 5	3,648,325,888	Reading	4		6-13 14:26:06	6-13 14:20:17 (00:05:55)	1,12,1,1	IBM LTO5
Star Raw LTO-5	S58033	1 / 1	5,000,399,360	Reading	1		6-13 14:25:41	6-13 14:24:15 (00:01:57)	1,15,1,0	IBM LTO5
Star Raw LTO-6	S61030	1 / 1	1,993,036,288	Reading	2		6-13 14:25:25	6-13 14:23:38 (00:02:34)	1,15,1,15	IBM LTO6
Star Raw LTO-6	S61124	2 / 2	2,067,981,312	Reading	1		6-13 14:26:08	6-13 14:25:01 (00:01:11)	1,13,1,2	IBM LTO6
Star Raw LTO-6	S61129	2 / 3	5,000,318,976	Reading	3		6-13 14:26:10	6-13 14:21:47 (00:04:25)	1,9,1,14	IBM LTO6
Star Raw LTO-6	S65017	1 / 1	5,015,422,464	Reading	3		6-13 14:25:56	6-13 14:22:09 (00:04:03)	1,9,1,13	IBM LTO6
TOTAL:	18 Tapes	28 Files								

14 LTO-5 drives

4 LTO-6

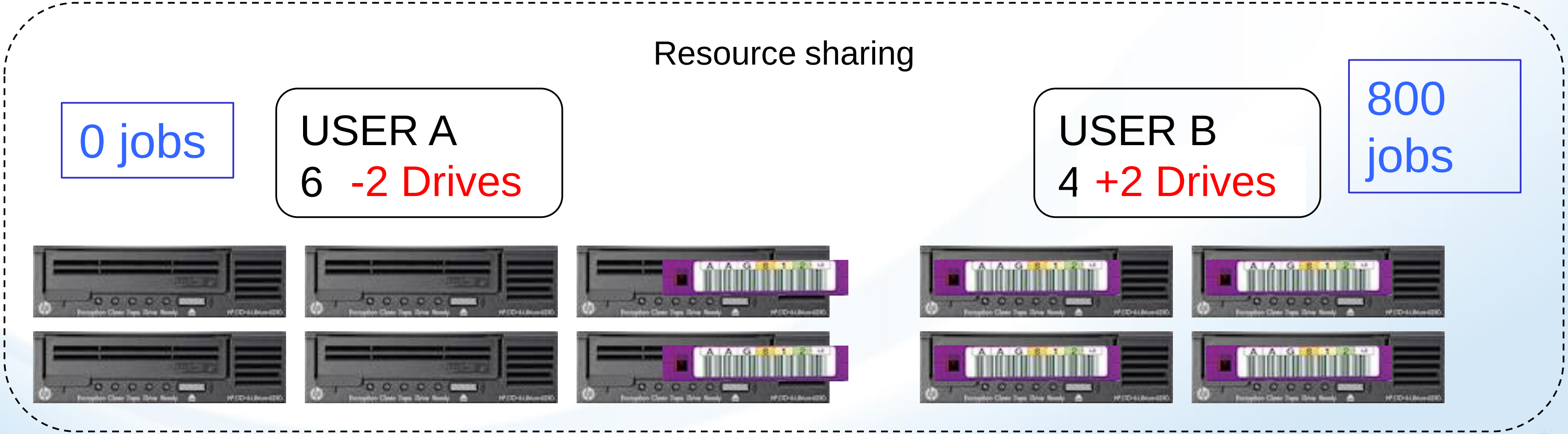
Efficient Access to Massive Amounts of Tape-Resident Data

Multiple Users, multiple policies

Customized resource allocation for each user

Resource sharing:

Adjust resource allocation on demand.



Resource sharing

0 jobs

USER A
6 -2 Drives

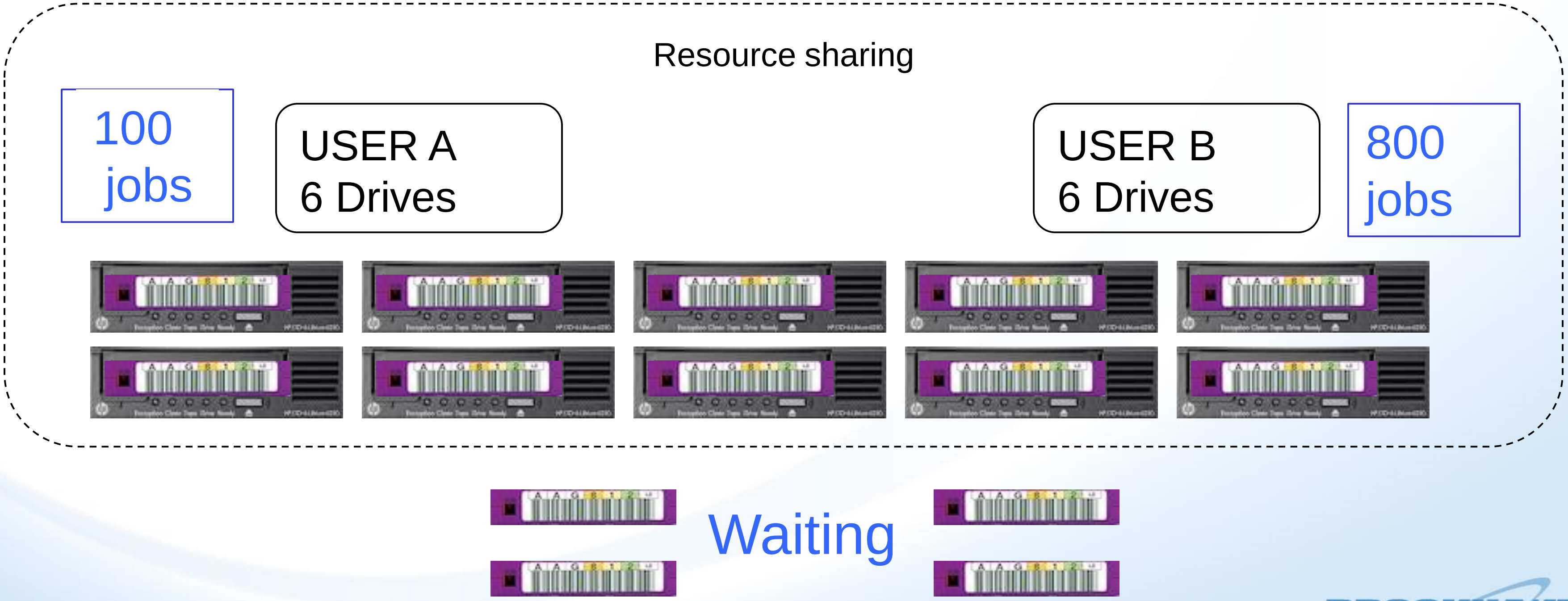
USER B
4 +2 Drives

800 jobs

Resource Allocation Oversubscription

Resource Allocation Oversubscription

Let user borrow (fight for) drives on demand



Resource sharing

100 jobs

USER A
6 Drives

USER B
6 Drives

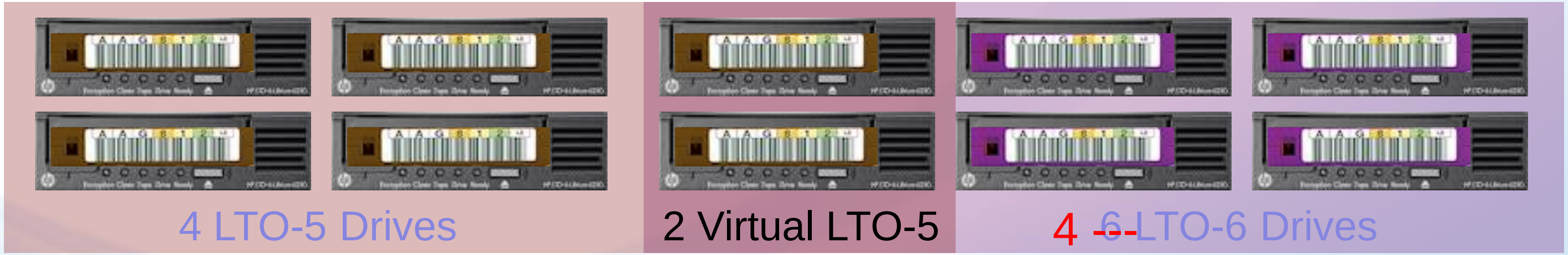
800 jobs

Waiting

Drive-generation Oversubscription

Use later gen drive to read prev gen tapes

Use 2 LTO-6 drives (virtual LTO-5) to read LTO-5 tapes



800 LTO-5 Reqs

20 LTO-6 Reqs

USER A
6 ~~4~~ LTO-5 Drives
4 ~~6~~ LTO-6 Drives

4 LTO-5 Drives

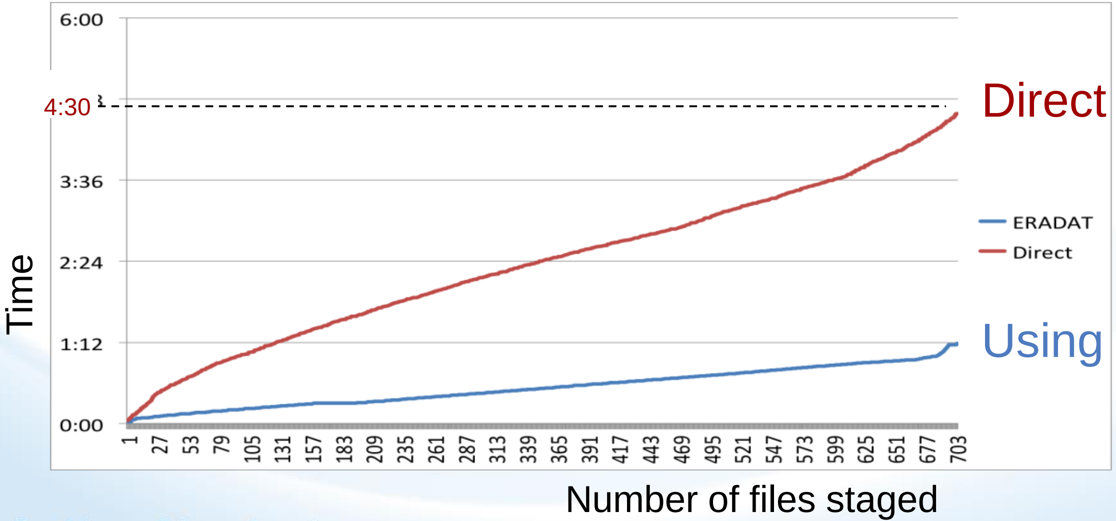
2 Virtual LTO-5

4 ~~6~~ LTO-6 Drives

Use case (test)

Randomly restoring 704 x 10 GB files out of 21 tapes, with 15 available drives.

- Direct submission: Using 15 job-queues, it took 270 mins to complete. Average ~444 MB/s. Used 34 mounts.
- Using ERADAT: Using 15 job-queues, it took only 70 mins to complete. Average ~1.7 GB/s. Used 21 mounts.



Direct submit, 4.5 hours, 34 mounts

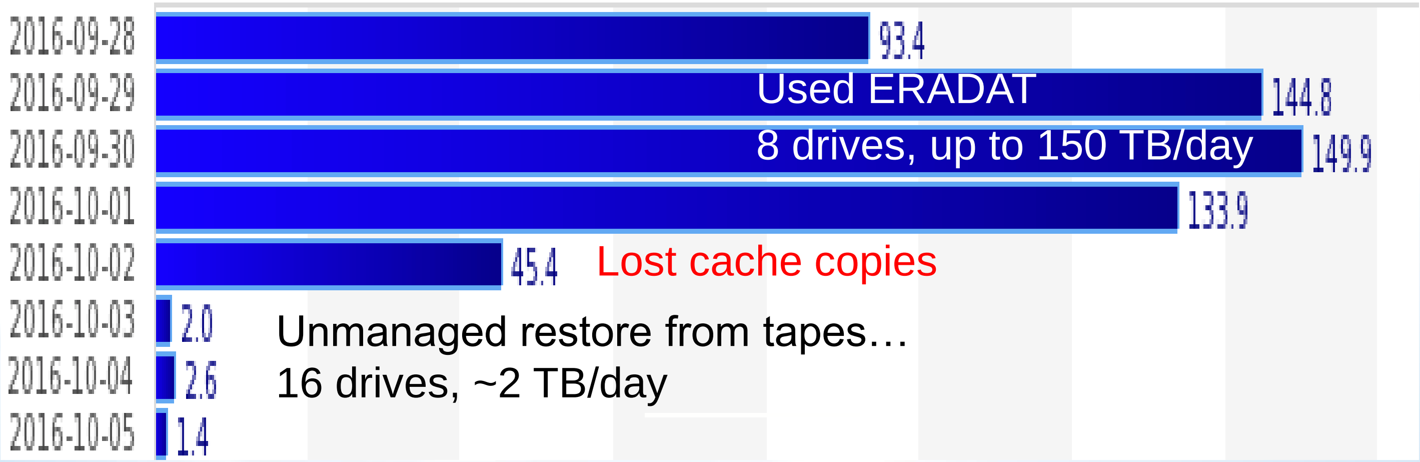
Using ERADAT, 70 min, 21 mounts

Use case (real)

In production system, used ERADAT to stage files using 8 LTO-6 drives

- But something happened, we lost the disk cache copies (purged)
- All users were pulling files directly from tapes in random order… used all 16 drives unmanaged!

Daily Data Transfer Volume (TB) into HPSS in last 8 days



Used ERADAT
8 drives, up to 150 TB/day

Lost cache copies

Unmanaged restore from tapes…
16 drives, ~2 TB/day

Date	Mounts
09-28	67
09-29	92
09-30	68
10-01	140
10-02	507
10-03	393
10-04	518
10-05	709

Staging Performance

Massive Staging

Tapes: LTO-5

Avg File Size: 10 GB

16 LTO-5 Drives

Delivered ~2.0 GB/s

Or **128 MB/s/drive**

All overhead included

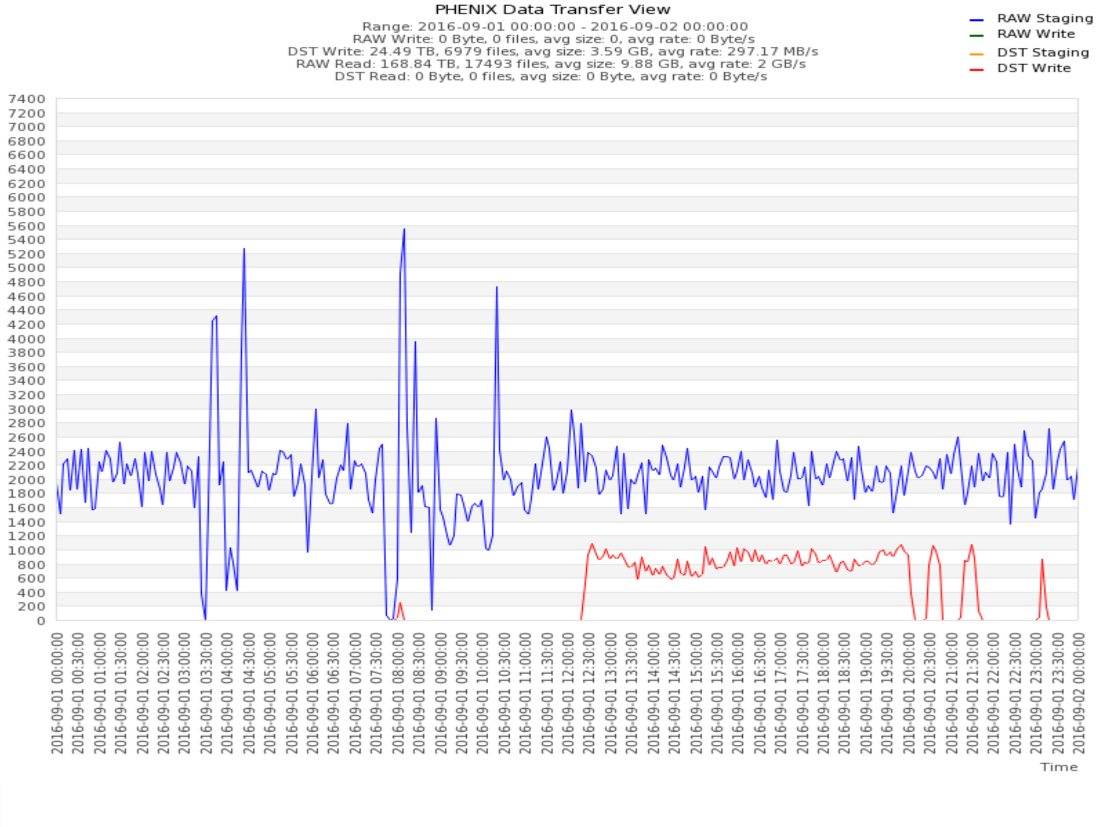
LTO-5 max transfer rate: 140 MB/s

8/30 – 9/16 (18 days)

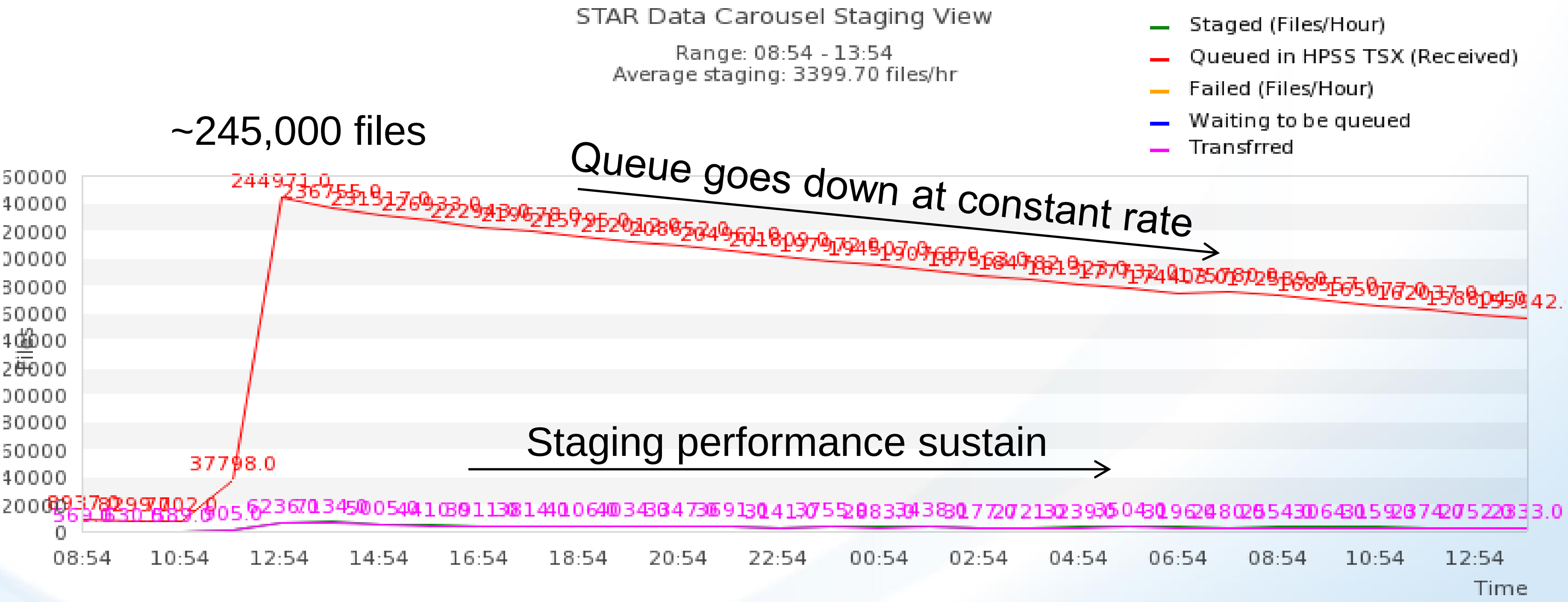
242,679 file

2.3 PB

1,708 LTO-5



ERADAT Staging Activity Graph

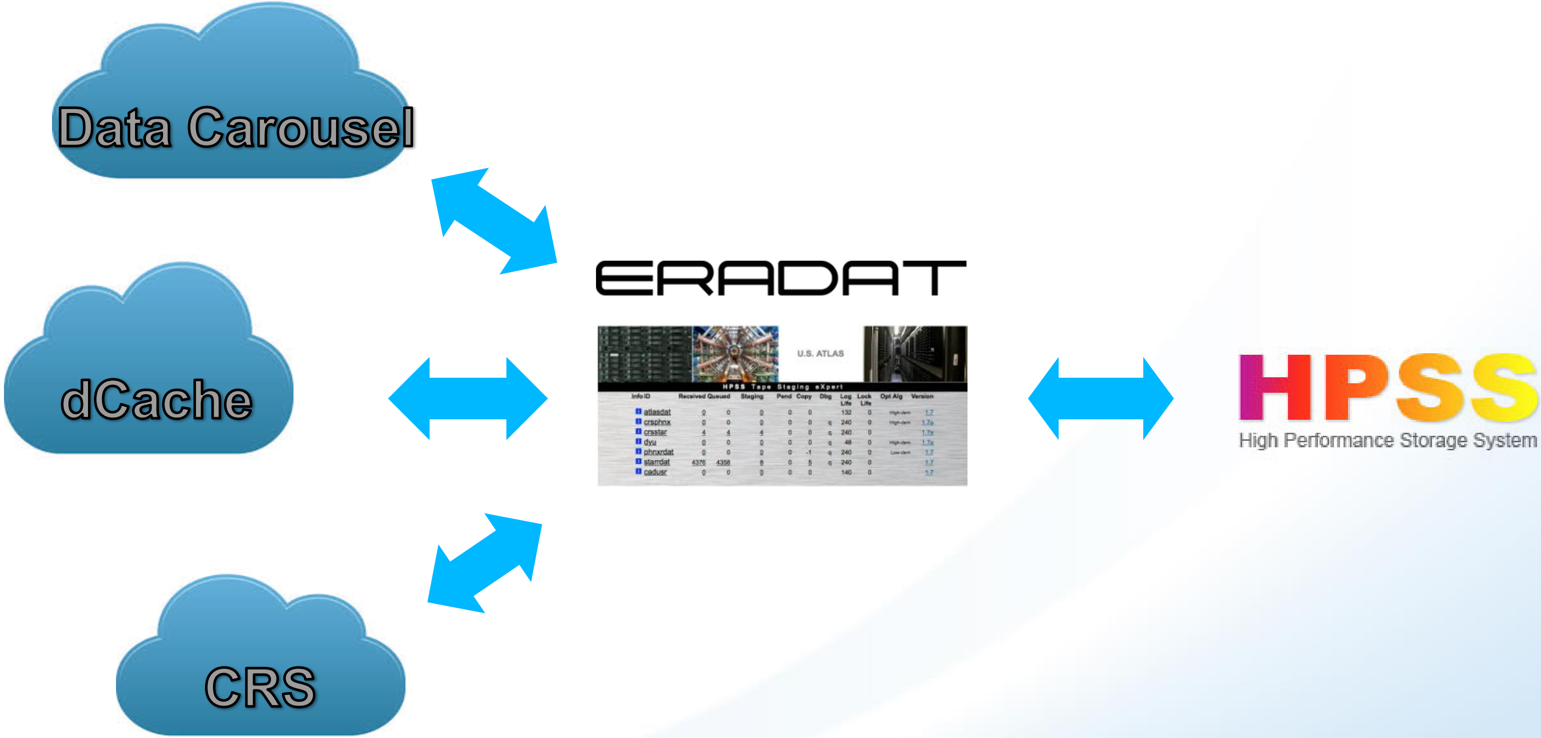


~245,000 files

Queue goes down at constant rate

Staging performance sustain

ERADAT and other systems



Data Carousel

dCache

CRS

ERADAT

U.S. ATLAS

HPSS Tape Staging eXpert

Info ID	Received	Queued	Staging	Pend	Copy	Dbg	Log Life	Lock Life	Opt Alg	Version
atlasdat	0	0	0	0	0		132	0	High-dem	1.7
crsphnx	0	0	0	0	0	q	240	0	High-dem	1.7a
crsstar	4	4	4	0	0	q	240	0		1.7a
dyu	0	0	0	0	0	q	48	0	High-dem	1.7a
phnxrdat	0	0	0	0	-1	q	240	0	Low-dem	1.7
starrdat	4376	4358	8	0	5	q	240	0		1.7
cadusr	0	0	0	0	0		140	0		1.7

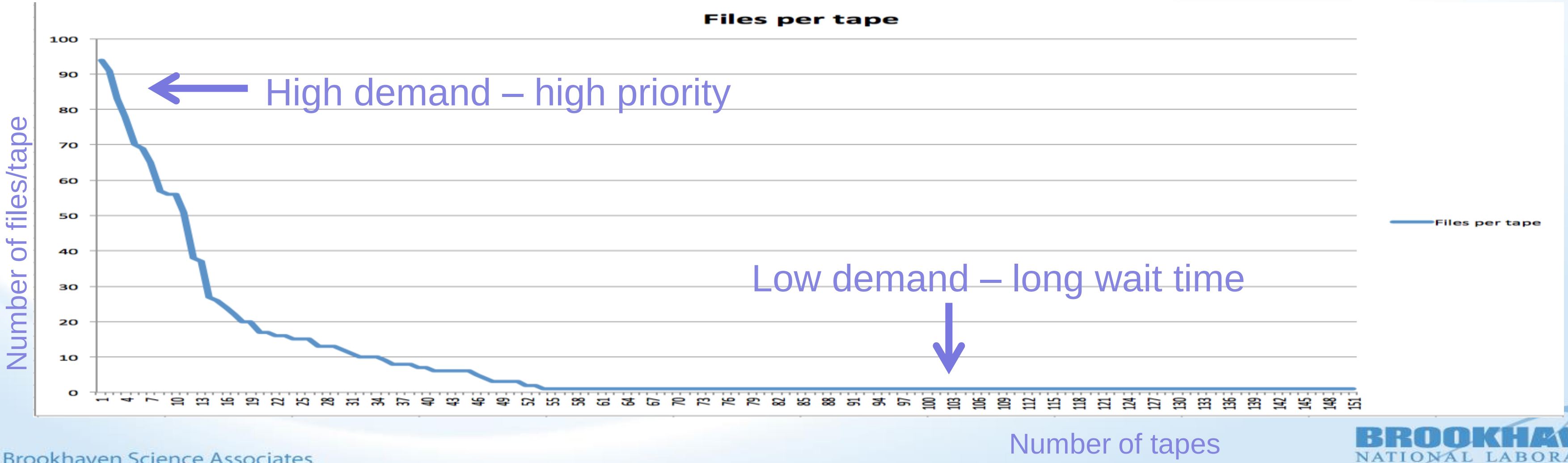
HPSS
High Performance Storage System

Data Carousel

Why the carousel

In a sustained request and multi-user environment, ultimate performance may not be the best choice - 1 user requesting 1 file from 1 tape may wait "forever" if ERADAT's restore policy is strictly "high demand”

The long tail

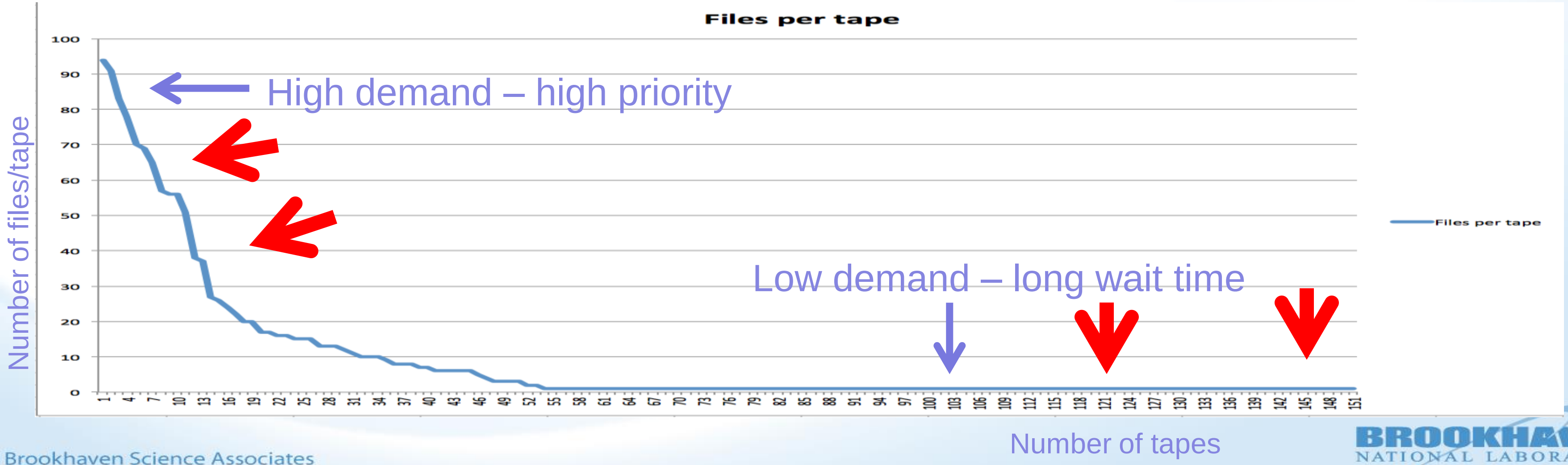


High demand – high priority

Low demand – long wait time

Data Carousel

The Carousel provides ways to achieve fair-shareness by switching between ERADAT policies (high demand, FIFO, low demand) and allows sparse requests to be satisfied in a reasonable time.



High demand – high priority

Low demand – long wait time

Data Carousel

Implements:

- SHARE: user and group based sharing policies
  EQUAL all users get equal share
  GROUP all groups get equal share
  GRPW group are weighted, equal share within group
- ORDER: sorting of requests ahead of ERADAT
  * By time files were requested
  * By tapeID (strict tape ordering, or hybrid approach i.e. submits all files requested from a given tape at the same time to balance fair-share and optimization)

The system balances between sharing (the bandwidth) and sorting (optimization) and and switches between ERADAT high demand and FIFO to achieve goals.

Flexible framework - can be extended by any custom SHARE or ORDER policies of your own.

Summary

ERADAT is a file retrieval scheduler for general use, it is designed to optimize the tape mount and read, and provides resource management for multi-user and multi-purpose use.

Data Carousel is designed for further optimization for STAR's environment.

- The DataCarousel is an extendable and fault tolerant policy driven framework in a multi-user environment
- For collaboration to make file retrieval requests to ERADAT
- The DataCarousel allows extending the SHARE policies using a simplistic yet very flexible mechanism.

Summary

Questions?

Thank you!