Deep Storage for Scientific Data

NEW YORK CITY

August 15, 2016

Guangwei Che, Tim Chou,
Ognian Novakov, David Yu
Shigeki Misawa

BROOKHAVEN NATIONAL LABORATORY
a passion for discovery

U.S. DEPARTMENT OF ENERGY | Office of Science

Deep Storage for Scientific Data

Scientific Data

BIG data

Bigger and bigger

Non-compressible

Preserve for decades

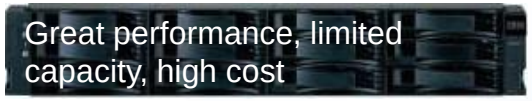
As the precision, energy, and output of scientific instruments such as particle colliders increase, so does the volume of generated data.

Scientific experiments produce huge amount of data which is usually compressed before being stored by efficiently utilizing the network throughput and storage infrastructure.

As the amount of our scientific experiments data has increased rapidly, there will be a serious need to provide a long term and efficient data storage.

Storage Technologies

Data Storage Solutions

- SSDs  Extreme performance, limited capacity, high cost
- Enterprise Disk Arrays FC, SAS  
- Midrange Disk Arrays SATA   capacity, value, performance
- Automated Tape Libraries  Giant capacity, low cost per GB and great power efficiency

Long Term Storage

Keeping 100+ PB data for decades!

Long Term Storage

Long Term Storage Keeping 100+ PB data for a long time

- Life Expectancy For archiving BIG data, only disk or tape

Disk: 1- 5 years

Tape: 30+ years

- Reliability Bit Error Rate (BER)

Enterprise SATA
1×10^{15}

Enterprise SAS
1×10^{16}

LTO-6
1×10^{17}

LTO-7, T10K
1×10^{19}

- Performance For *archiving* BIG data, sequential write

15K RPM SAS
160 - 233 MB/s

LTO-6
~160 MB/s

LTO-7
~300 MB/s

T10K
252 MB/s

- Scalability

Complicate

Just add tapes

Performance and Preservation

Storage for performance

New or frequently accessed data is typically supported on **disk storage** resources due to performance requirements.

Storage for Deep Archive

Deep Archive: high reliability storage for inactive data that is rarely used or accessed.

As the volume of data collected by scientific experiments explodes year after year, there will be a need to provide a deep archive storage to economically store the data; the obvious candidate for providing this service is…

BNL Scientific Data and Computing Center

Mass Storage as a Service

The Tape Storage System at the **RHIC and Atlas Computing Facility (RACF)** at BNL has been providing mass storage services to the scientific experiments of RHIC and LHC (CERN, Geneva)

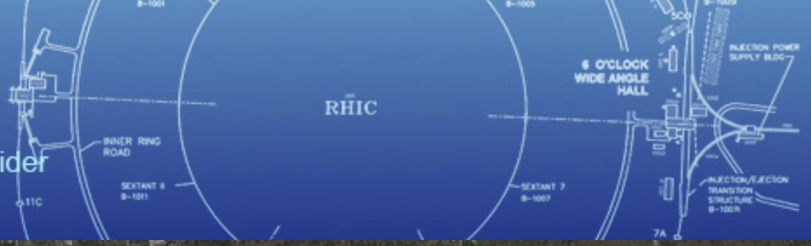
- RHIC - Relativistic Heavy Ion Collider.
  - Data storage for all experiments data and esp. STAR and PHENIX.
- U.S. ATLAS Tier 1 Facility - Large Hadron Collider (LHC) at CERN.
  - Secondary data store for a fraction of ATLAS data (23% - 25%).
  - Primary US site for data storage, processing and distribution.
- Other BNL Facilities

Designed to provide reliable and high throughput parallel archiving and retrieval of large amount of data on a 24x7, year-round basis.

Currently, ~90 PB of data in BNL RACF Tape Storage System

RHIC

Brookhaven National Laboratory's Relativistic Heavy Ion Collider



PHENIX

RHIC

STAR

RACF

RHIC

Brookhaven National Laboratory's Relativistic Heavy Ion Collider

RHIC is a world-class particle accelerator at Brookhaven National Laboratory where physicists are exploring the most fundamental forces and properties of matter and the early universe.

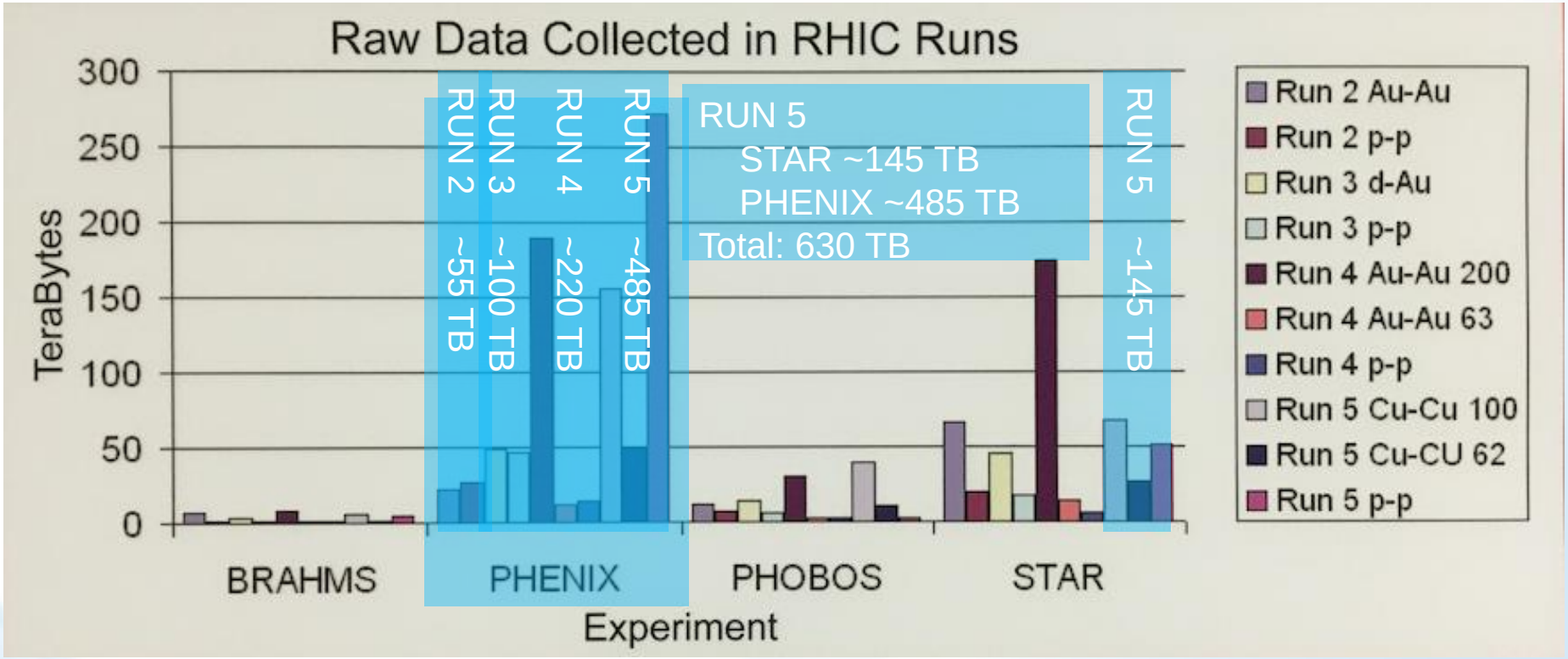
RHIC accelerates beams of particles to nearly the speed of light, and smashes them together to recreate a state of matter thought to have existed immediately after the Big Bang some 13.8 billion years ago.

The Large Hadron Collider

The Large Hadron Collider (LHC) is the world's largest and most powerful particle accelerator. It first started up on 10 September 2008, and remains the latest addition to CERN's accelerator complex. The LHC consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way.

10 Years Ago

RHIC Run 2 – Run 5

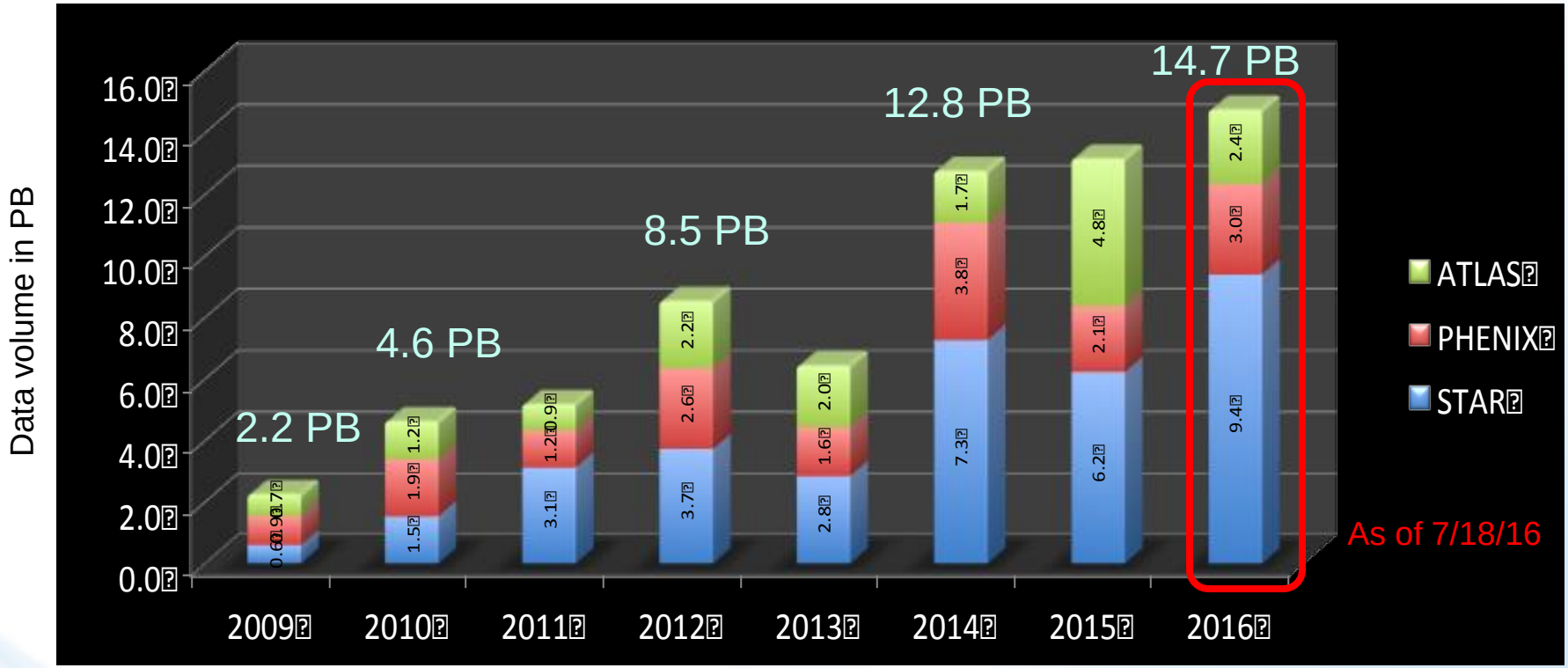


Data received by year from 2002 to 2005

Data Received In Recent Years

Mass Storage as a Service

The amount of our science experiments data has increased rapidly.

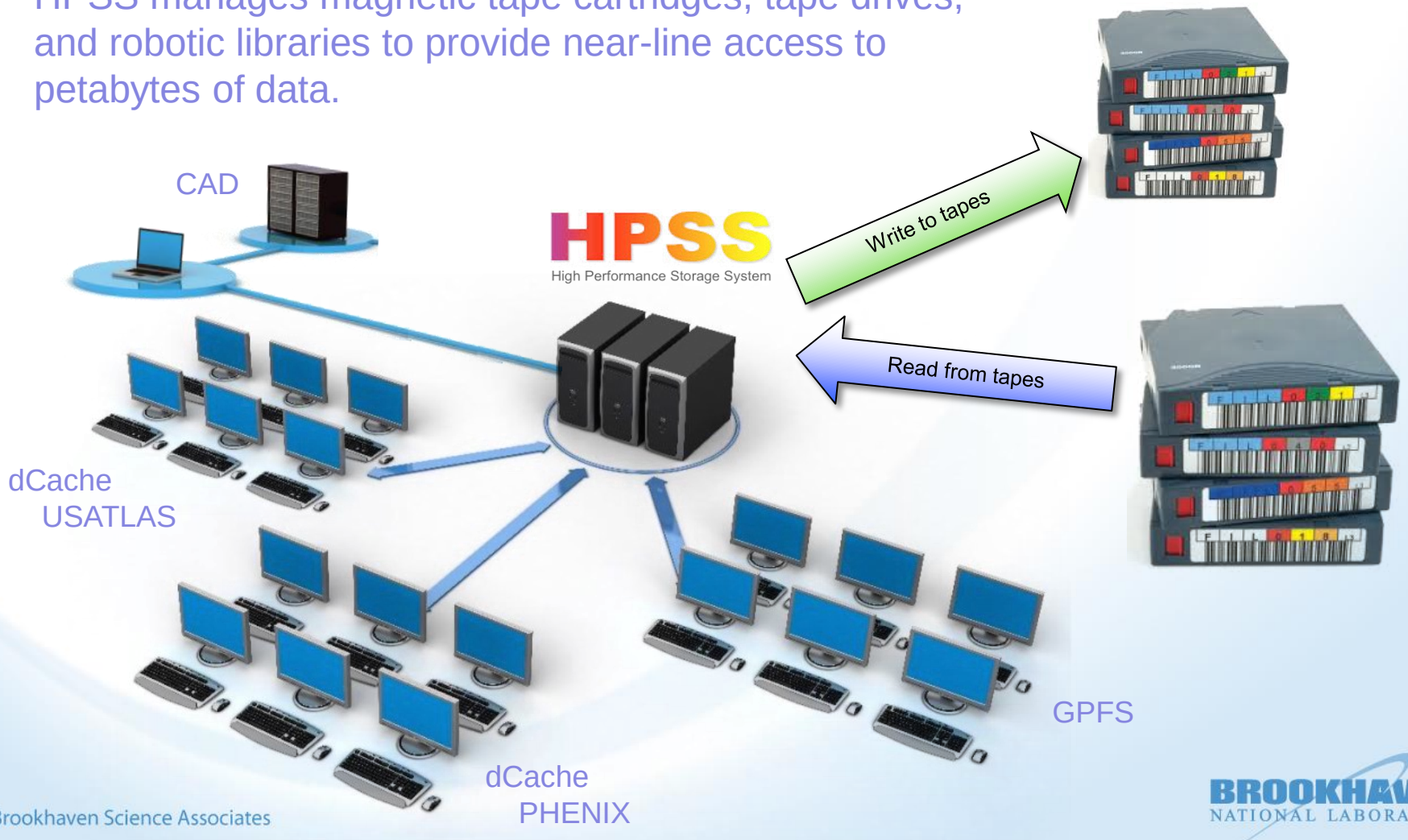


Data received by year since 2009

Storage Management

High Performance Storage System (HPSS)

HPSS manages magnetic tape cartridges, tape drives, and robotic libraries to provide near-line access to petabytes of data.



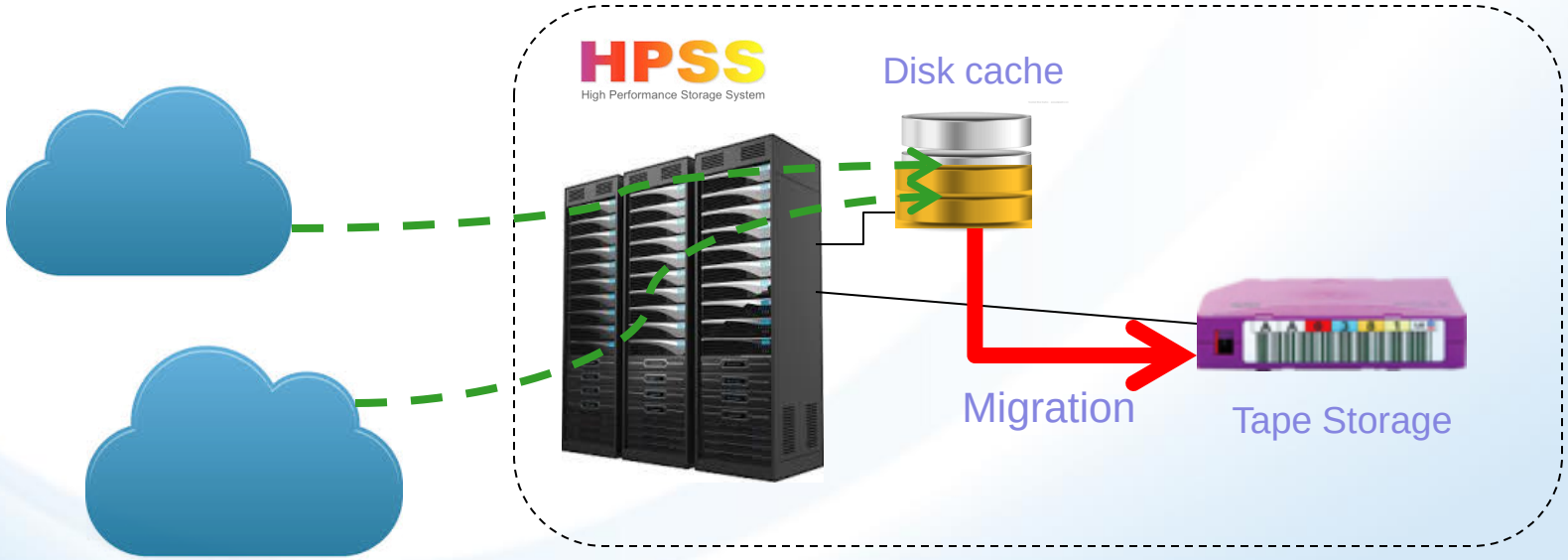
Data in HPSS

HPSS
High Performance Storage System

~90 PB of data on tapes

~60K+ tapes, mix of LTO 4,5,6 and T10KD technologies

~900 TB total disk cache



Access Files From Tape Storage

Automated Tape Library

SL8500

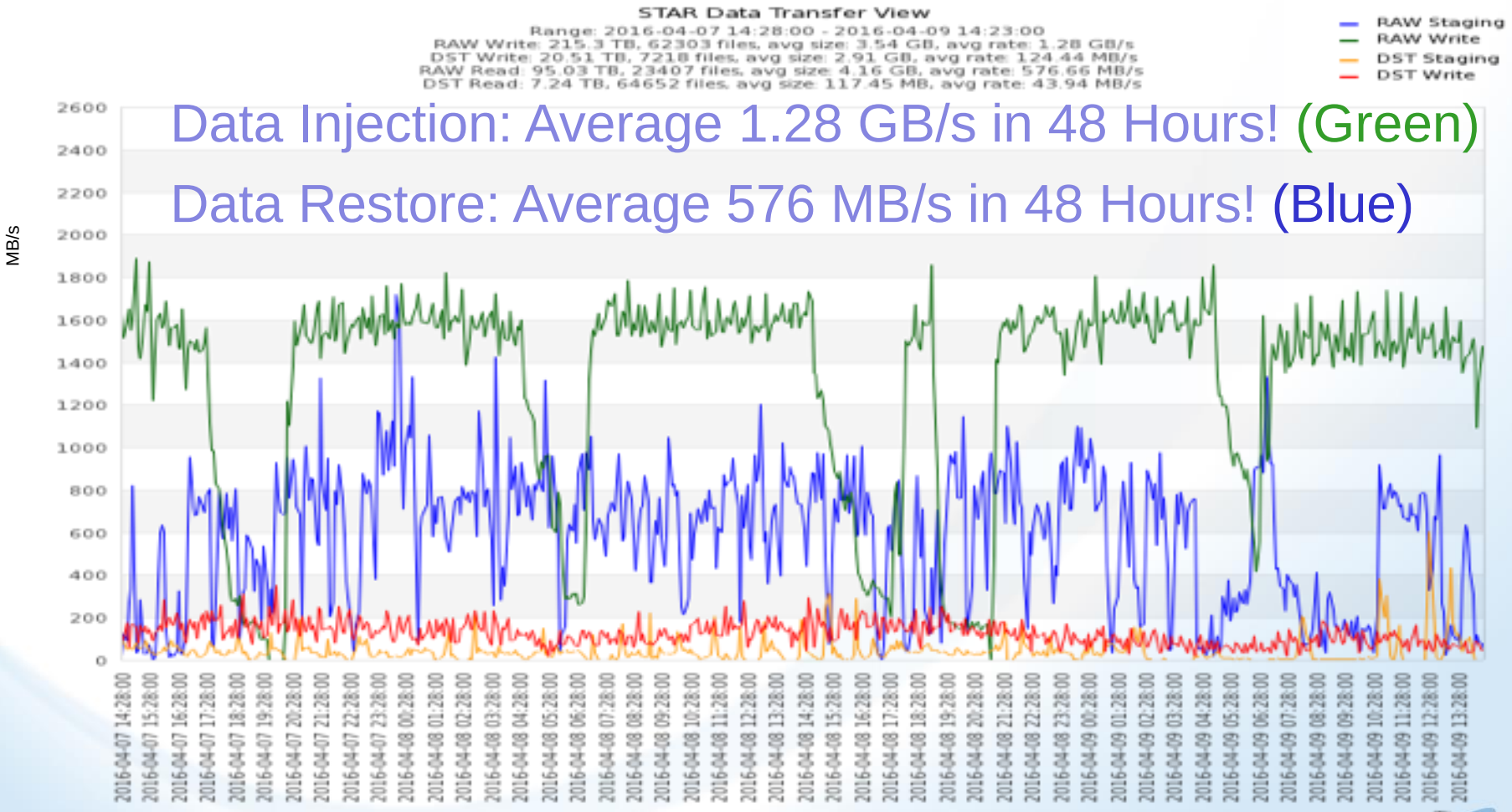
Capacity: 6~10K cartridges

LTO-6: 24 PB

LTO-7: 58 PB

High Throughput Parallel Archiving

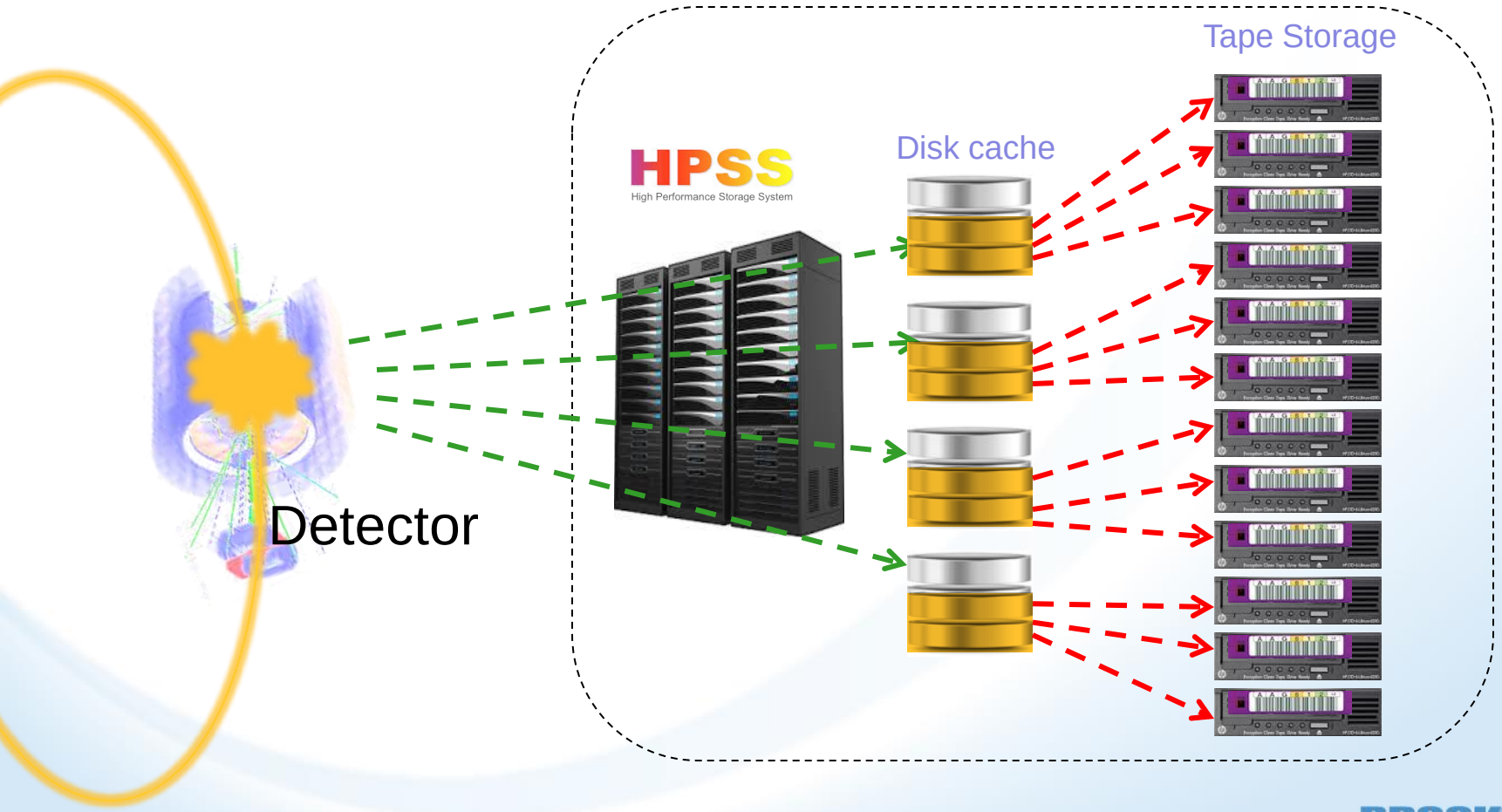
RHIC RUN 16 - STAR



High Throughput Parallel Archiving

Each LTO-6 can write about 110~140 MB/s

Allocated 12 Drives to migrate data to tapes

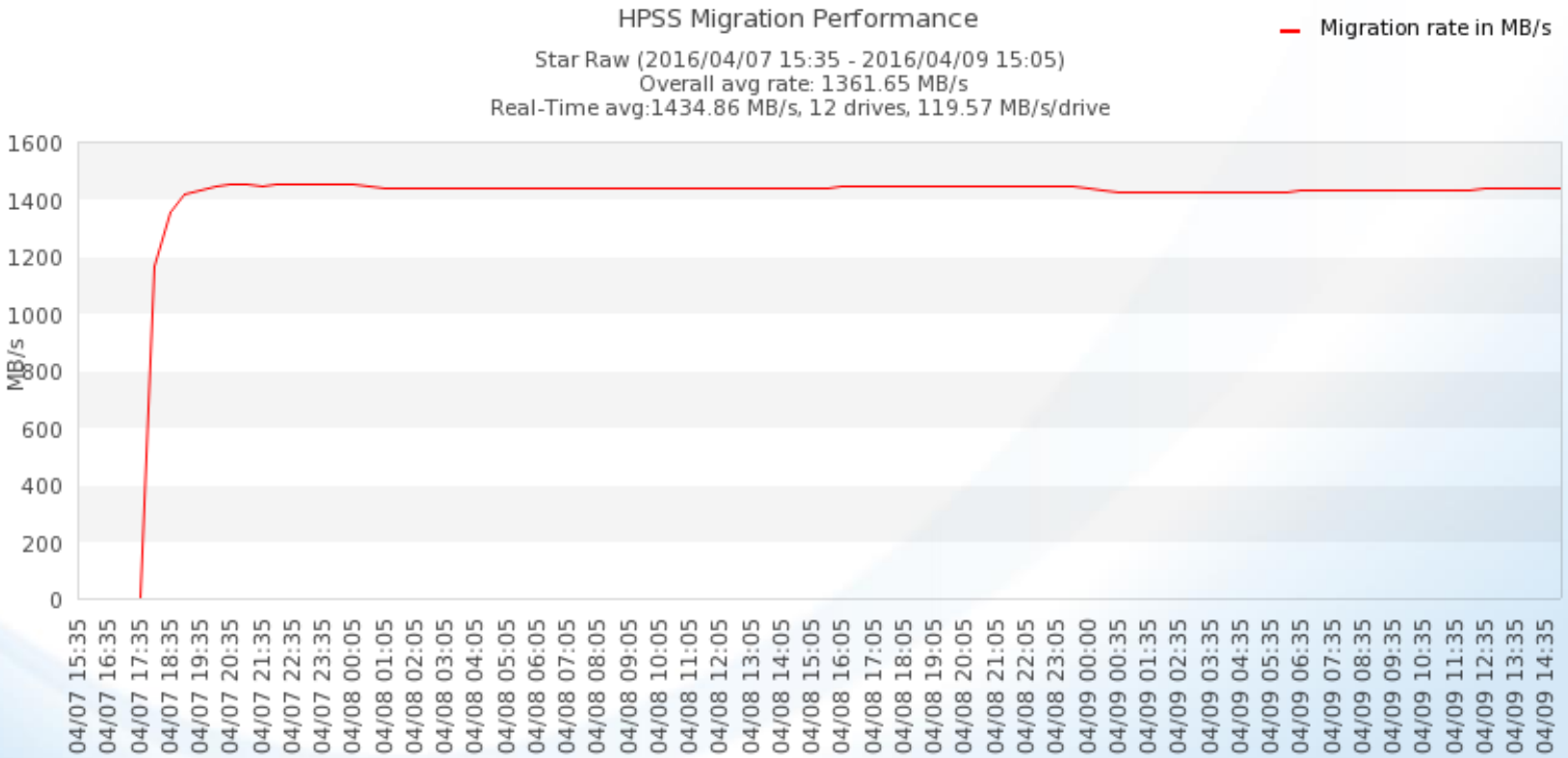


Tape Write Performance

Tape Drive Performance

With12 Drives to catch up the data injection rate (1.28 GB/s)

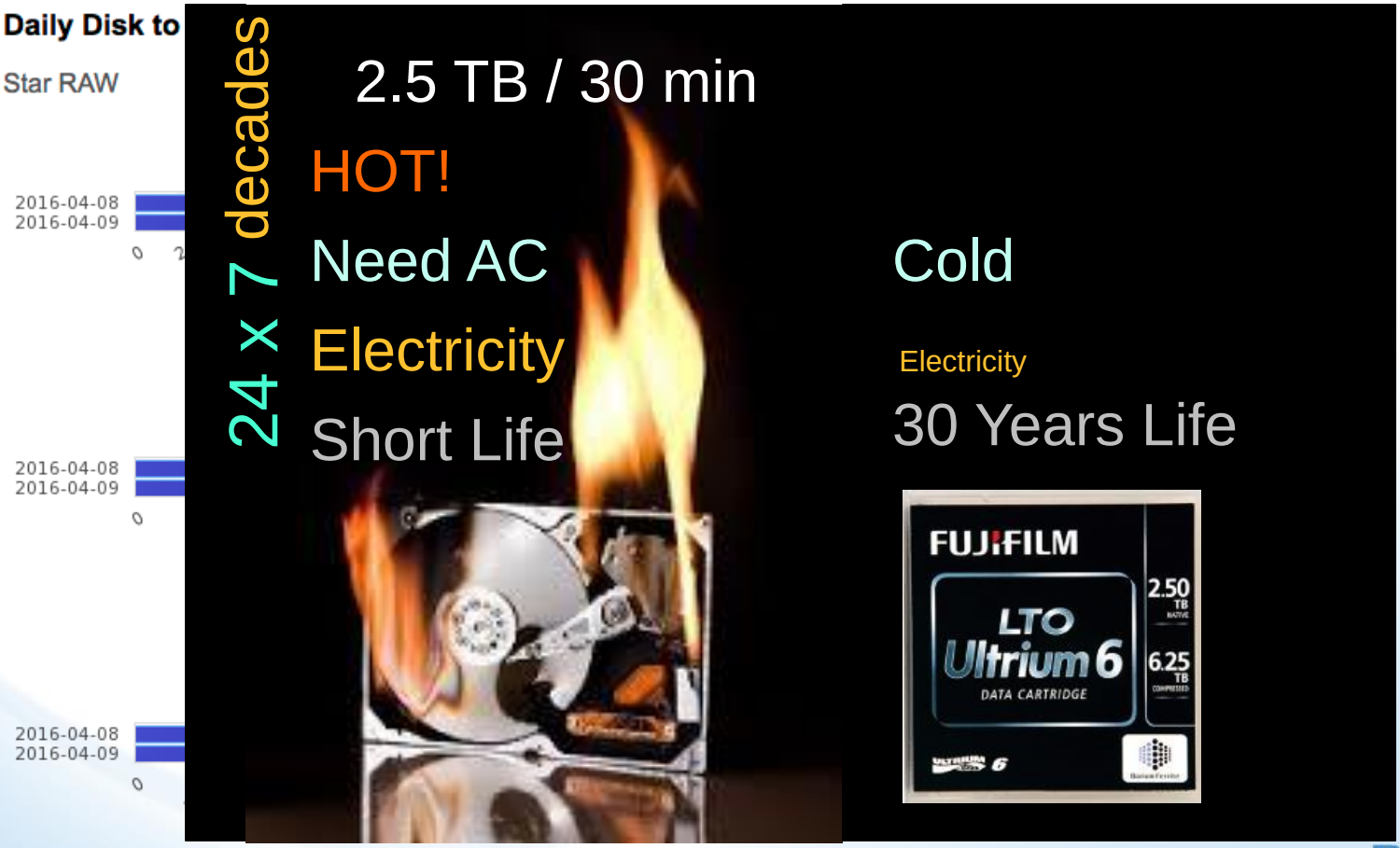
Real-time Tape Write speed at 1.4 GB/s (1434.86 MB/s)



Pipeline to Drain Data Swamp

Tape Drive Performance

In 48 hours, we migrated 238 TB to 97 LTO-6 tapes.



Resource Management

Resource Optimization

Balance the tape drive resources for read and write

- Allocate just enough resources for read and write
- Allocate resources based on priority
- Able to adjust the resources dynamically.

Tape mount and dismount degrades the performance

Minimize the tape mounts

- Write: Accumulate enough data, mount once, write a full tape and then dismount.
- Read: Mount once, read all requested files, then dismount.

Flexible Policies

Multiple Users, multiple policies

Customized write policy per storage class

Adjust write policy on demand

STAR Raw
12 Streams

STAR DST
3 Streams

PHENIX Raw
6 Streams

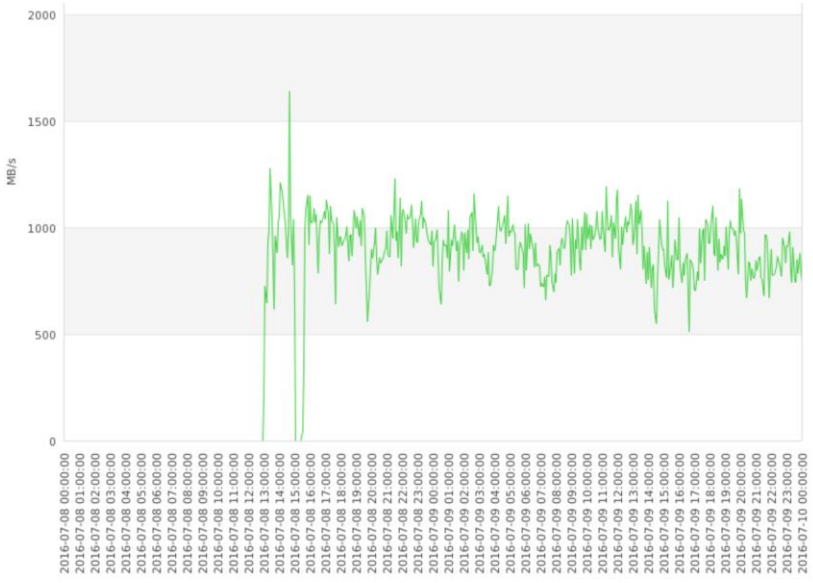
PHENIX DST
10 Streams

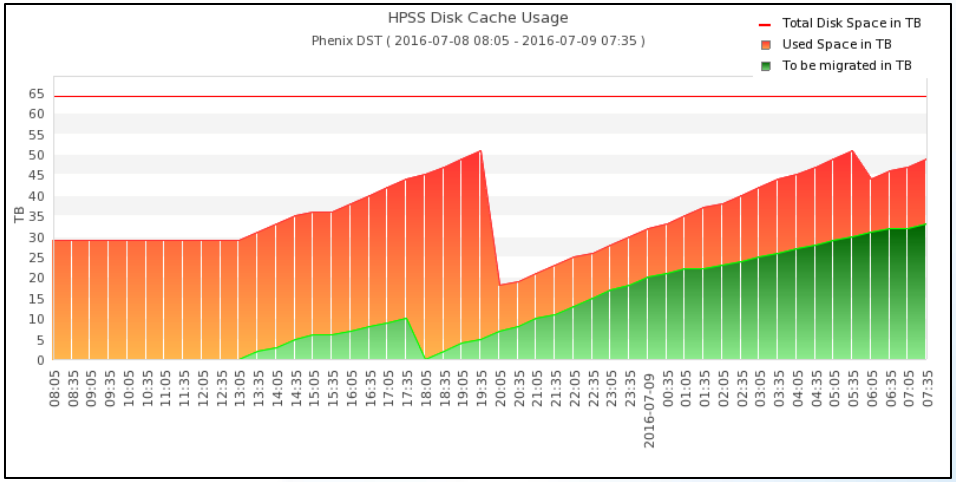
USATLAS
4 Streams

Operating the Data Warehouse

Write Optimization

Surprisingly high data injection rate





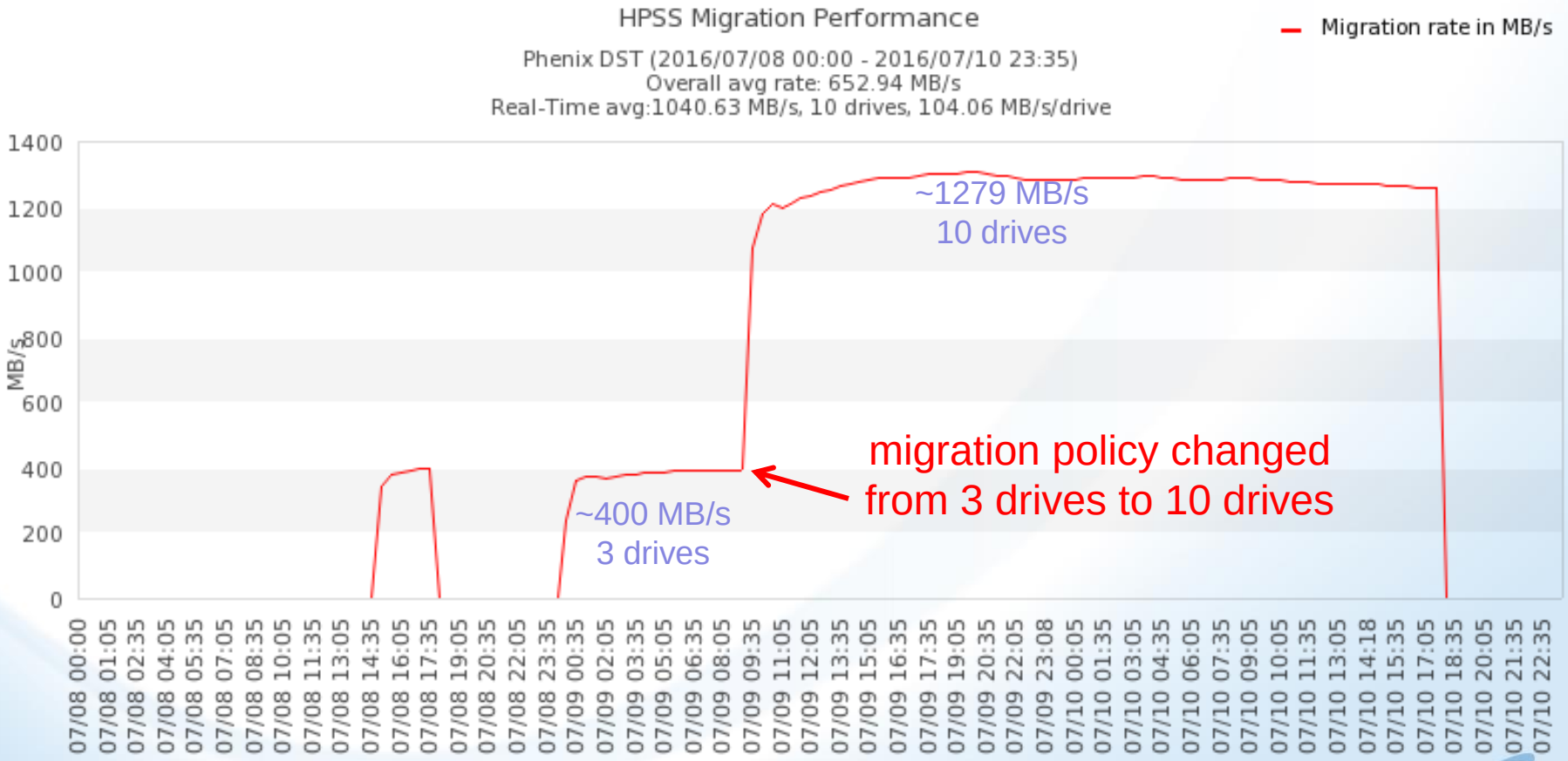
Disk usage going up (green)!

Migration cannot keep up!

Operating the Data Warehouse

Write Optimization

Adjust write streams.



Accessing Archives

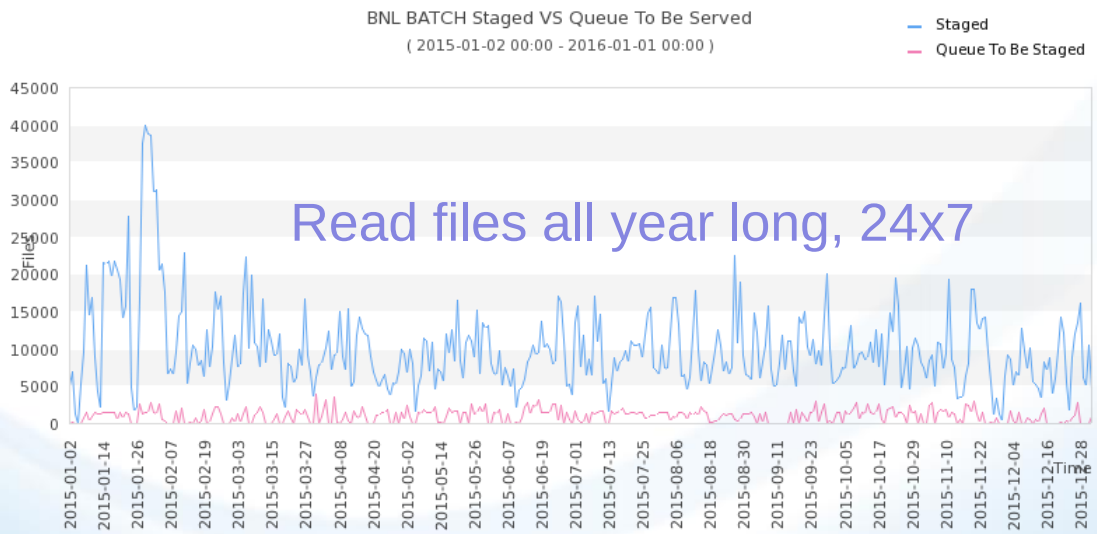
Restore Files

Old data are being archived as Deep Archive

New data are being actively accessed

Data restored from RACF tape storage in 2015:
10,720,308 files, 17 PB!

That's average 29,370 files/day, 47 TB/day.



Efficient Retrieval and Access to Data Archived on Tape

Read back is optimized

Tape is sequential access, good for data archiving

Restoring from tape is a challenge, as data may end up being spread over multiple tapes

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.

* Based on Oak Ridge National Lab's "Batch", developed in early 2000s

Mount once, read all the needed files sequentially

Restore Files Optimization

- Files are pre-sorted by tape id, position number and offset (for small file aggregation).

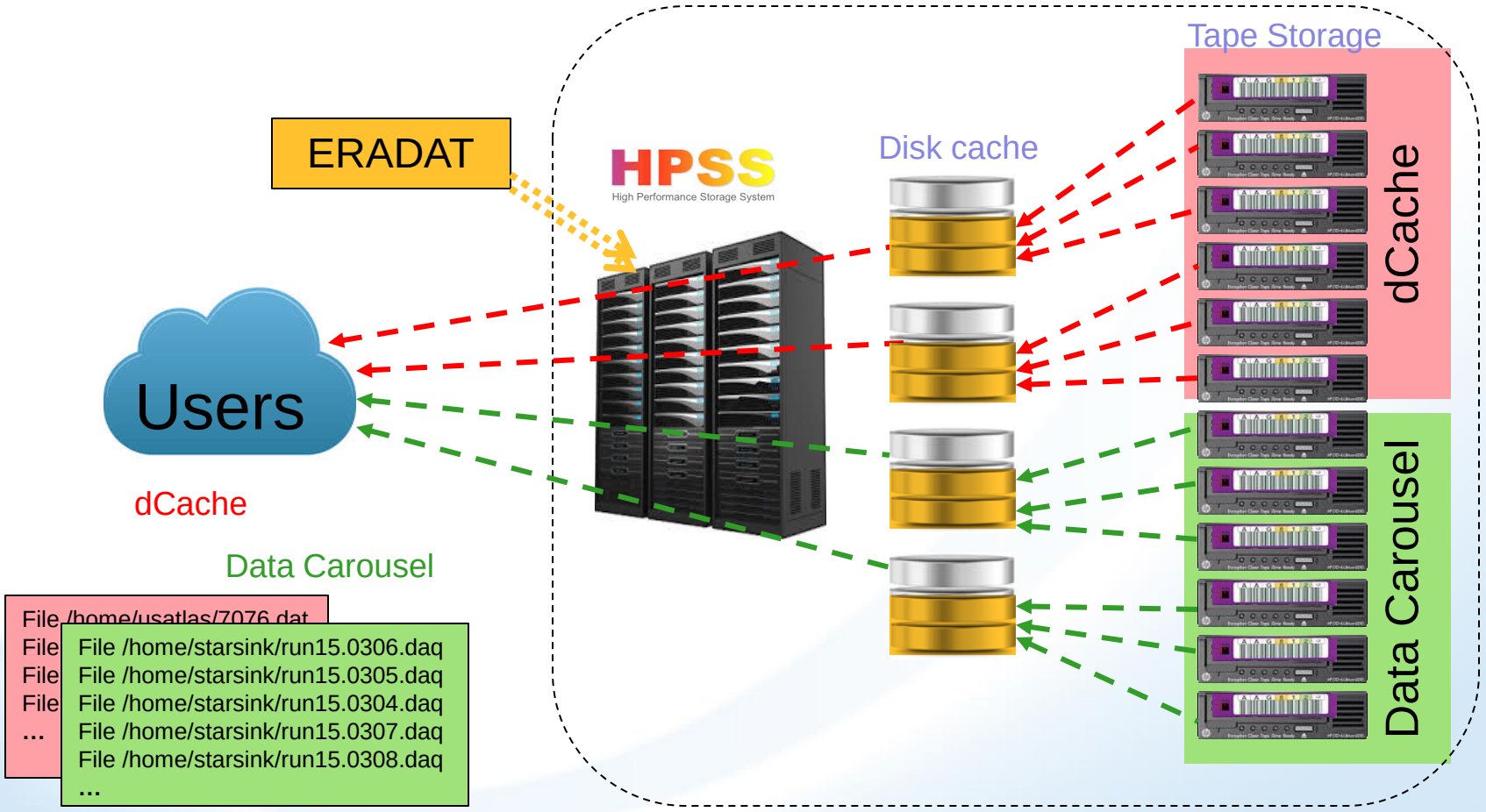
  Aggregate all files within the same tape together, and read them sequentially from the beginning of the tape towards the end of the tape, minimize unnecessary rewinds.
- Tape selection order: FIFO, LIFO, and “By Demand”
- Priority Staging

Observed good read performance as high as 126 MB/s* (single stream) for LTO-6.

*USATLAS experiment, 62 x avg ~6 GB files in random position. Included overheads such as tape-mount (and previous tape’s dismount), forward-position and rewinds.

High Throughput Parallel Retrieval

Drives Resources are pre-allocated. “Resource Guaranteed”

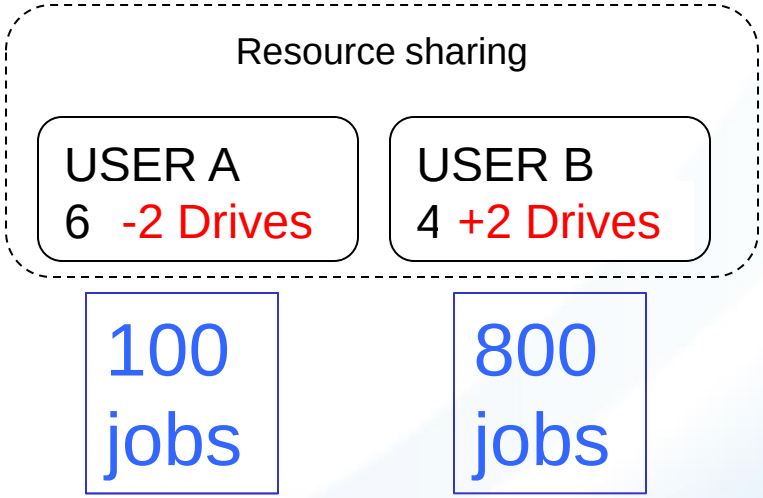


Resource Sharing

Multiple Users, Multiple Policies

Customized resource allocation for each user

Adjust resource allocation on demand,
no service interruption



New Technology

LTO-7

We have done some evaluations using HPSS with real data.

- 587 x 10 GB files, 5.73 TB
- Sequential write: ~258 MB/s/tape
- Sequential read: ~253* MB/s/tape

* Included tape mount time

We plan to convert most of the LTO-4 tapes to LTO-7 starting from the end of this year.

LTO-4 (800 GB) → LTO-7 (6 TB)

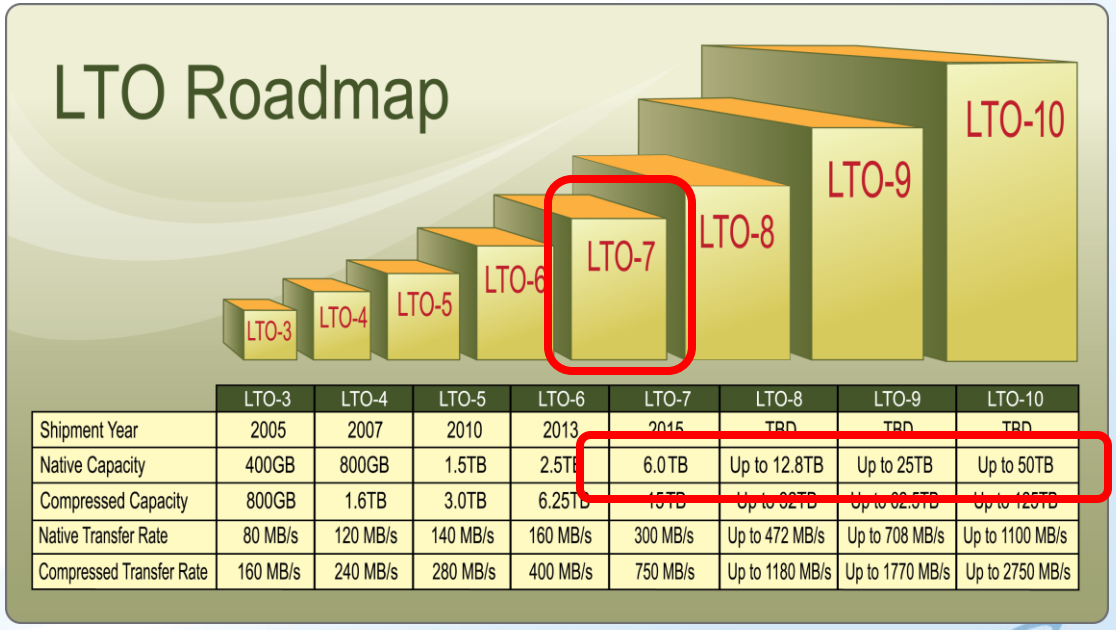
To free up more storage space inside the automated tape library.

New Technology

The Future of Tape Storage

Linear Tape Open (LTO) Ultrium is a high-capacity, single-reel tape storage solution developed and continually enhanced by Hewlett Packard Enterprise, IBM and Quantum and promoted by the LTO Program. It's a powerful, scalable and adaptable tape format that helps address the growing demands of data protection.

It's also an open format, licensed by some of the most prominent names in the storage industry to ensure a broad range of compatible tape drives and cartridges.



	LTO-3	LTO-4	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO-10
Shipment Year	2005	2007	2010	2013	2015	TBD	TBD	TBD
Native Capacity	400GB	800GB	1.5TB	2.5TB	6.0TB	Up to 12.8TB	Up to 25TB	Up to 50TB
Compressed Capacity	800GB	1.6TB	3.0TB	6.25TB	15TB	Up to 32TB	Up to 62.5TB	Up to 125TB
Native Transfer Rate	80 MB/s	120 MB/s	140 MB/s	160 MB/s	300 MB/s	Up to 472 MB/s	Up to 708 MB/s	Up to 1100 MB/s
Compressed Transfer Rate	160 MB/s	240 MB/s	280 MB/s	400 MB/s	750 MB/s	Up to 1180 MB/s	Up to 1770 MB/s	Up to 2750 MB/s

http://www.lto.org/technology/what-is-lto-technology/

https://www.spectralogic.com/features/lto-7/

Future Archive Technology

IBM, Fujifilm are still researching

Tape is already the least expensive storage medium per bit, easily beating spinning hard disks or solid-state drives. The trade-off is slower retrieval time, but this makes tape perfect for archiving large amounts of infrequently used data.

IBM and Fujifilm have figured out how to fit **220 TB** data on a standard-size tape that fits in your hand, flexing the technology's strengths as a long-term storage medium.

Source: http://www.networkworld.com/article/2908654/data-center/ibm-fujifilm-show-tape-storage-still-has-a-long-future.html

Conclusions

Conclusions

- We have implemented an active tape-storage data archive currently holding 88 PB of scientific data and serving scientists from multiple collaborations worldwide
- The archive and the underlying tape storage complex are managed in a very efficient way by a suite of in-house developed optimization and monitoring tools
- The data archive volume can grow infinitely with the addition of relevant tape storage resources (media, libraries and tape drives)
- Tape is still the leader in cost per GB and power efficiency. The implementation concept is based on the most cost-effective and energy efficient (green) memory model available today

Deep Storage for Scientific Data

Thank you!

Questions?