ERADAT (BNL Batch) & the Data Carousel

David Yu, Jérôme Lauret

Brookhaven National Laboratory

ACAT 2010

Outline

- Introduction
- ERADAT (BNL Batch)
- The Data Carousel
- Case studies
  - Data mining at RHIC
  - ESD re-processing at US-Atlas
  - Data Carousel restore of analysis files in Xrootd namespace
- Conclusion

About BNL

- **Brookhaven National Laboratory**
  - Established in 1947 on Long Island, New York
  - A multi-program national laboratory
  - Approximately 3,000 scientists, engineers, technicians and support staff and over 4,000 guest researchers annually
- **RHIC and ATLAS Computing Facility**
  The facility provides computing services for
  - the experiments at the Relativistic Heavy Ion Collider (RHIC) at BNL
  - the US-based collaborators in the ATLAS experiment at the Large Hadron Collider (LHC) at CERN
  - RACF is
    - The Tier 0 Facility for RHIC
    - A tier 1 Facility for US-ATLAS

Tape storage & problem statement

- Hardware:
  - 6 Sun/STK SL8500, each can hold ~ 5 PB data, managed by IBM High Performance Storage System (HPSS)
  - BNL's tape storage holds over 13 PB of data
- Problematic
  - Data production in time sequence for submission + different data ⇔ stochastic file saving to tape from data mining workflow
  - User may be staging any number of files out of any random tape
    - Reading back by "group" (production series, collision, year, …)
  - May have thousands of reading demands, 24 x7
  - HPSS is designed for archiving, not optimized for reading

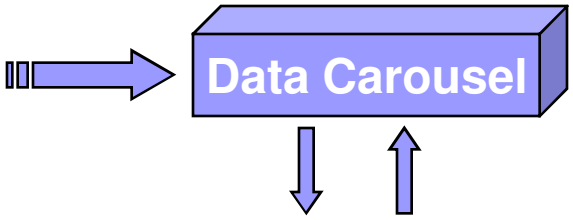
Workflow + usage pattern = great potential for chaos

Reading files randomly placed back from tape is definitely not so effective - latencies

Tape Technology

- Tape is sequential access.
  - Reading random files back from tape is definitely not effective
- File Access latency
  - Tape transport inside the library
  - Mounting time
  - Tape position seeking time
  - Rewind and dismounting time
  - These latency may take at least 140 seconds.
  - Tape condition, tape marks.

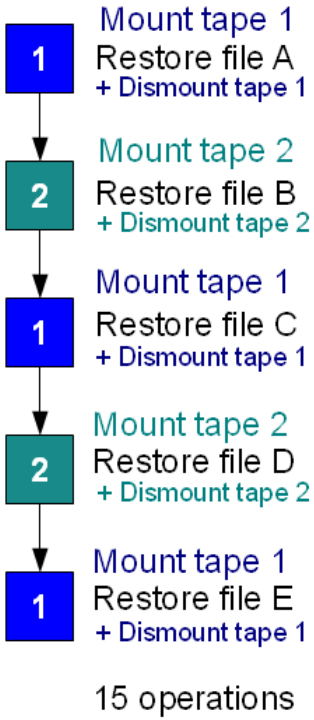
A high level resource usage policy handler

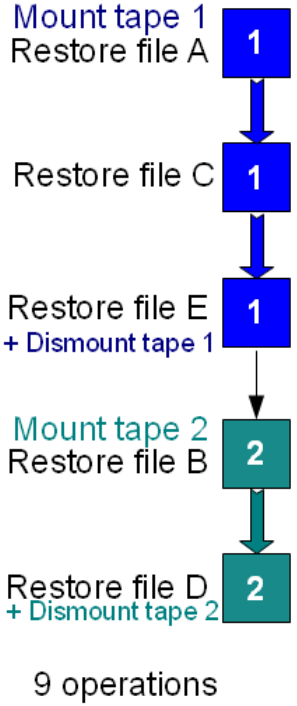


ERADAT = Efficient Retrieval and Access to Data Archived on Tape

A tape queuing system

Requested: A, B, C, D E
Tape 1: A, C, E
Tape 2: B, D





Timeline …

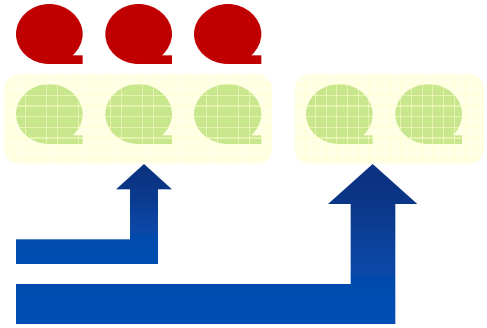
- Order files by tape access as much as possibly achievable
  - ORNL batch to **BNL Batch** (RICH data production) — 2000
- Multi-user considerations
  - One user could still bring the (prod) system to a stall
  - Policy driven needed -> Data Carousel (treat by "ground" with share) — 2001
- Try optimizing for throughput
  - Biggest request queue first – **ERADAT**
  - Use ***Data Carousel*** for data management (Xrootd file request) — 2005
- Further fairshare considerations
  - Across users, group shares
  - Multiple-policies — 2007
- Now – more monitoring and controls, … — 2009

ERADAT (BNL Batch)

- Is a “file retrieving scheduler” for IBM High Performance Storage System (HPSS).
- Is based on Oak Ridge Batch, customized to BNL’s requirements and improvements:
  - Dynamic drive usage allocation, supports multi-projects, multi-technologies, and multi-users.
  - Keeps all transaction history for performance reports, and fine tuning the configuration, as well as altering file submission mechanism.
  - Web-based monitoring system.

ERADAT (BNL Batch)

- *Dynamic drive usage allocation, supports multi users.* Configuration can be altered in real-time
  - Reserving N drives for Writing
  - Reserving M drives for Reading
  - Reserving P drives for user A
  - Reserving Q drives for user B
  - ...
- *Supports multiple hardware technologies*
  - Each drive-type has it's own drive usage allocation
- *Supports multiple groups*
  - Each group has it's own drive allocations



- Example
  - 9940B: 4 for Write, 8 for Read (n for user A, m for user B, ..., t for user H)
  - LTO-3: 6 for Write, 12 for Read
  - LTO-4: n for Write, m for Read (...)
- Example
  - Group A: 9940B only (n for W, m for R)
  - Group B: 9940B + LTO-3 + LTO-4
  - Group C: LTO-3 + LTO-4

How ERADAT works?

- If the file is still on disk cache, return immediately
- If the tape is locked, return error immediately
- Sort the files by Tape ID and position
- Giving the option of tape selecting
  - Process the most high demanded tape first
  - Process the tape based on FIFO (useful for handling of external complex policies)
- Provided “Priority Staging” - Process this tape in next available drive

How ERADAT works?

- Optimization Options
  - Mount tape based on number of files
    This is recommended when user requests are completely un-organized.
    Ex: restoring files from archive.

- Process the tape based on FIFO
  - Ordering provided by external algorithm
  - The processing will be “First in, first serve”
  - Ex: Data Carrousel

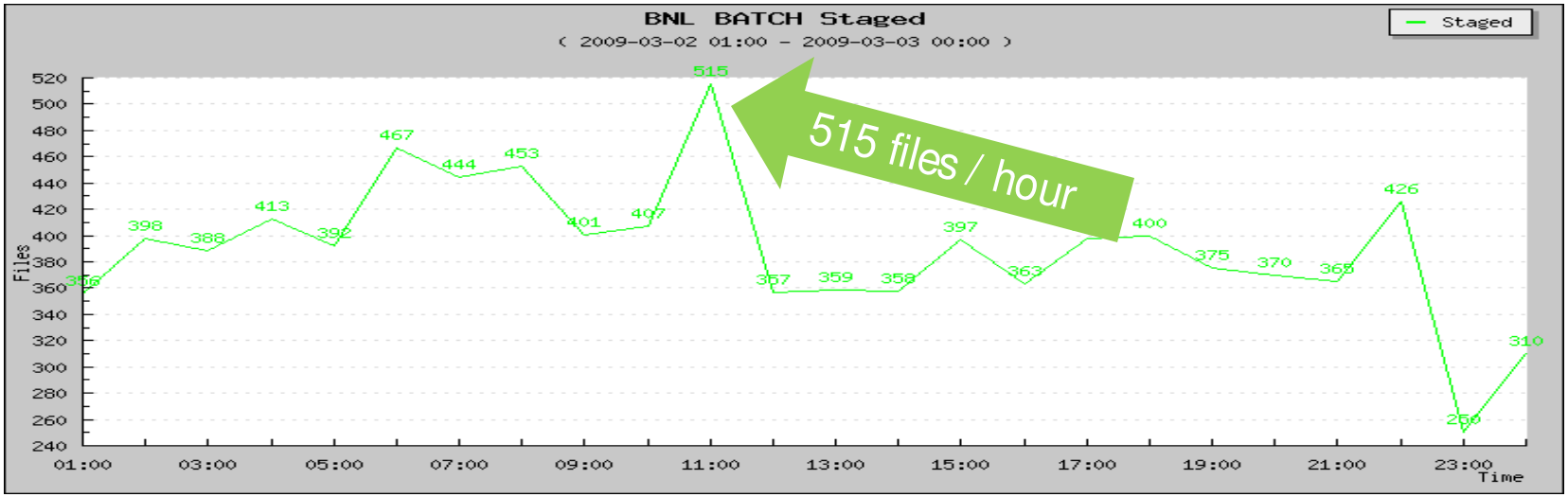
The Data Carousel

- An extendable fault tolerant policy driven framework and API
  - User make requests, asynchronous restore
  - Server handles the requests and execute restores from HPSS cache to location on behalf of user
- "Server"
  - *Applies policy*: FIFO, user share, group share, mixed, weighted faire queuing
    - P. Jakl et al., CHEP 2009 proceedings **Fair-share scheduling algorithm for a tertiary storage system**
  - *May consider "files on the same tape" within time interval (Time slicing)*
    - Avoids resource starvation – single file on a single tape will be satisfied
  - *Delegate restore to ERADAT* → call back
- Other features
  - Monitors (client command line reporting of progress, possibility to "see" what the server does from command line, Web interface & graphs)
  - Ability to retry on errors (all transient errors successfully treated, some are self-repairing, every request leading to HPSS errors can be re-queued N times)

How does it perform?

RHIC/STAR data mining performance

- Using default optimization option
  RHIC/STAR CRS Job Processing (on demand)
  - 18 LTO-3
  - Max 515 files, 189 GB (avg filesize: 376MB) per hour



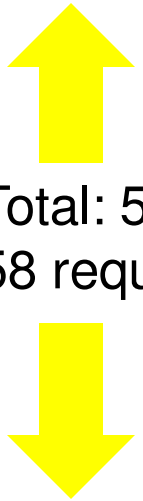
Statistics based on STAR CRS Jobs 03/02/2009

RHIC/STAR data mining performance

- Case Study
  On 03/02/2009, between 9:40 and 10:46
  Received 575 requests (involved 15 tapes)
  Tape #409167, a LTO-3 tape, only mounted 2 times
  Staged 58 files, 25.5 GB of data. Avg file size: 451 MB
  Sample:

DATE	TIME	Tape#	# Files
2009-03-02	09:40:20	409167	1
2009-03-02	09:41:17	409167	1
2009-03-02	09:44:17	409167	2
2009-03-02	09:45:17	409167	1
2009-03-02	09:48:17	409167	3
2009-03-02	09:49:17	409167	1
2009-03-02	09:50:17	409167	1
. . .			
2009-03-02	10:43:17	409167	2
2009-03-02	10:44:17	409167	2
2009-03-02	10:46:17	409167	6

  58 files associated with Tape #409167, arrived in 32 bundles
  That means average 1.8 files / bundle



Total: 575 requests
58 requests from #409167

RHIC/STAR data mining performance

- Case Study
  - All 58 files arrived in 32 bundles (consecutive tape mounts)
  - Average 1.8 files / bundle
  - If FIFO – No optimization, we would have 32 mounts!

How long would it take?

According to HP's webpage

HP StorageWorks LTO-4 Ultrium 1840 Tape Drive - Specifications

Component Specifications

Specifications	Component	LTO-4 Ultrium 1840	LTO-3 Ultrium 960
Performance (1 Kb = 1,000 bytes)	**Native Data transfer rate**	120 MB/s with LTO 4 media	80 MB/s with LTO 3 media
	Data rate matching range	40-120 MB/s	27-80 MB/s
	Data access time (from BOT)	62 seconds typical for LTO 4 media	53 seconds typical for LTO 3 media
	Rewind time from EOT	< 124 seconds for LTO 4 media	< 98 seconds for LTO 3 media
	Rewind tape speed	7 m/s for (LTO 3 and LTO 2 media)	7 m/s for (LTO 3 and LTO 2 media)
	Average load time	< 19 seconds (RW)	< 19 seconds (RW)
	Average unload time	< 19 seconds (RW)	< 19 seconds (RW)

RHIC/STAR data mining performance

Case Study

If process with FIFO (without optimization): Ave 1.8 files/mount
Tape delivery time: 5 sec
Mounting (loading): 19 sec
Set position (assume in the middle): 53 / 2 = 26 sec
Data transfer: **1.8 x 441 MB** / **80 MB/s** = 9.9 sec
Rewind tape: 98 / 2 = 49 sec
Dismount (Unload) : 19 sec
Place the tape back: 5 sec
Total: 132.9 seconds / mount

32 mounts = 71 minutes! For 25.5 GB => 6 MB / sec!

This is calculated based on theory, actual performance should also factor in the latency caused by tape marks and file size effects

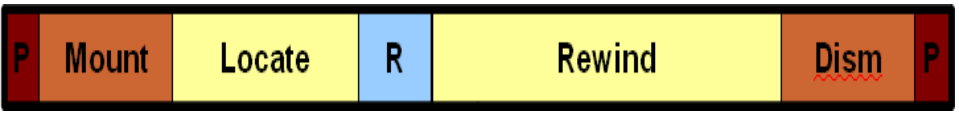
RHIC/STAR data mining performance

Case Study:

ERADAT with optimization: Ave 29 files / tape
Tape delivery time: 5 sec
Mounting (loading): 19 sec
Set position (assume in the middle): 53 / 2 = 26 sec
Data transfer: **29** x 441 MB / 80 MB/s = 160 sec
Rewind tape: 98 / 2 = 49 sec
Dismount (Unload) : 19 sec
Place the tape back: 5 sec
Total: 283 seconds / mount

58 files => 2 mounts ~ 10 minutes! Average about 46.16 MB / sec!

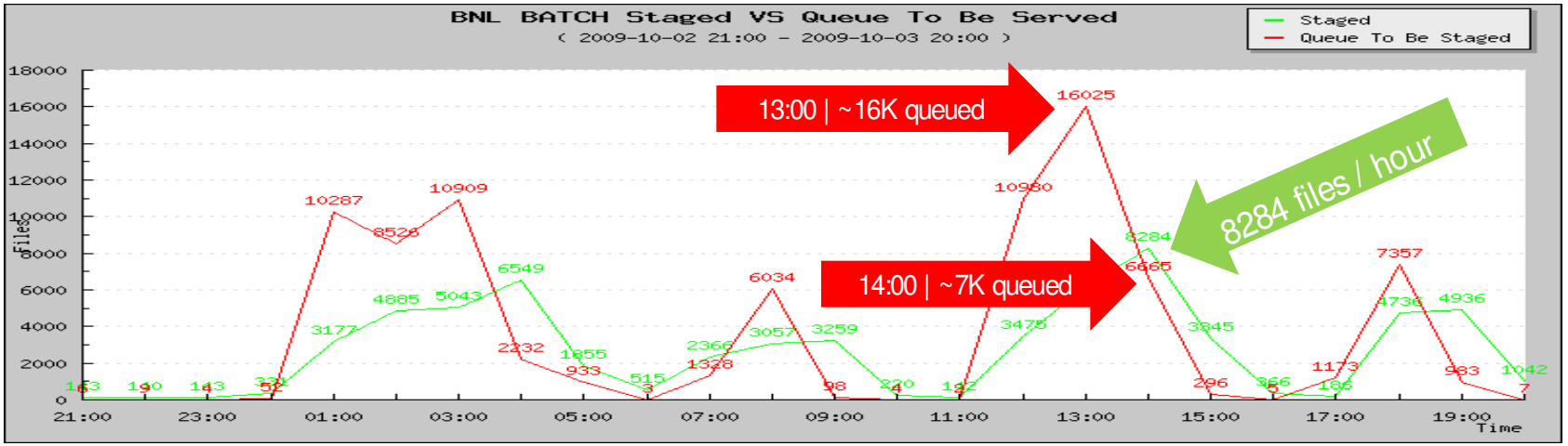
Statistics based on RHIC STAR CRS Job Processing 03/02/2009



6 MB/sec → 46.2 MB/sec

ESD processing at US-Atlas performance

- Using default optimization option:
  US Atlas ESD Reprocessing
  - 10 LTO-3 + 17 LTO-4 drives
  - Max 8284 files, 225 G (avg filesize: 27M) per hour



Statistics based on LHC Atlas ESD Reprocessing 10/03/09

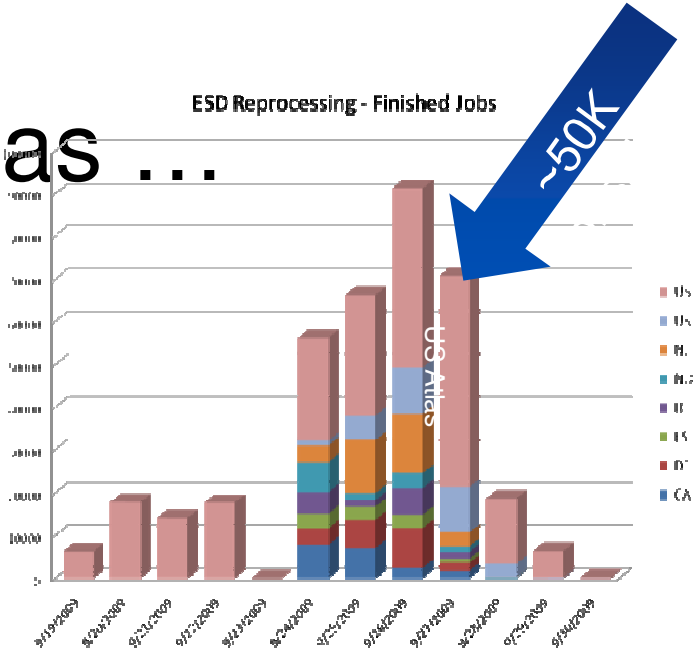
ESD processing at US-Atlas performance

- Case Study
  On 10/3/2009, between 2:03 and 13:21
  Received 73706 requests (involved 270 tapes)
  Tape #500425, a LTO-4 tape, only mounted 3 times
  Staged 2279 files, 77 GB of data. Avg file size: 34 MB

- 2279 files associated with Tape #500425, arrived in 530 bundles.
  That means average 4.3 files / bundle
  If FIFO – No optimization, we would do 530 mounts

  How long would it take?

ESD processing at US-Atlas …



Case Study

If process with FIFO (without optimization)

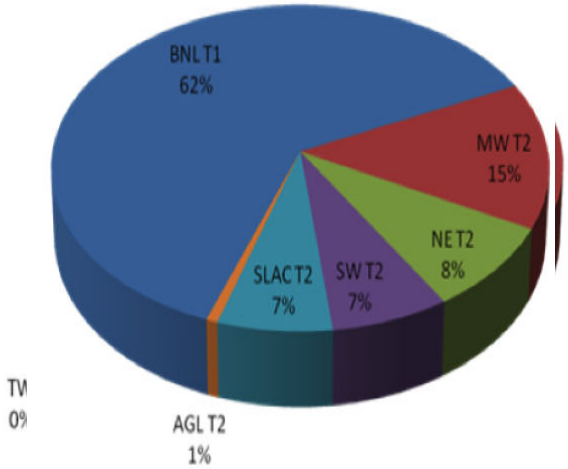
- Tape delivery time: 5 sec
- Mounting (loading): 19 sec
- Set position (assume in the middle): 62 / 2 = 31 sec
- Data transfer: 4.3 x 34 MB / 120 MB/s = 1.22 sec
- Rewind tape: 124 / 2 = 62 sec
- Dismount (Unload) : 19 sec
- Place the tape back: 5 sec
- Total: 142.21 seconds / mount

530 mounts = 21 Hours! About 1 MB / sec!

BNL Batch with optimization: Average 760 files / mount

- Tape delivery time: 5 sec
- Mounting (loading): 19 sec
- Set position (assume in the middle): 62 / 2 = 31 sec
- Data transfer: 759 x 34 MB / 120 MB/s = 215 sec
- Rewind tape: 124 / 2 = 62 sec
- Dismount (Unload) : 19 sec
- Place the tape back: 5 sec
- Total: 356 seconds / mount

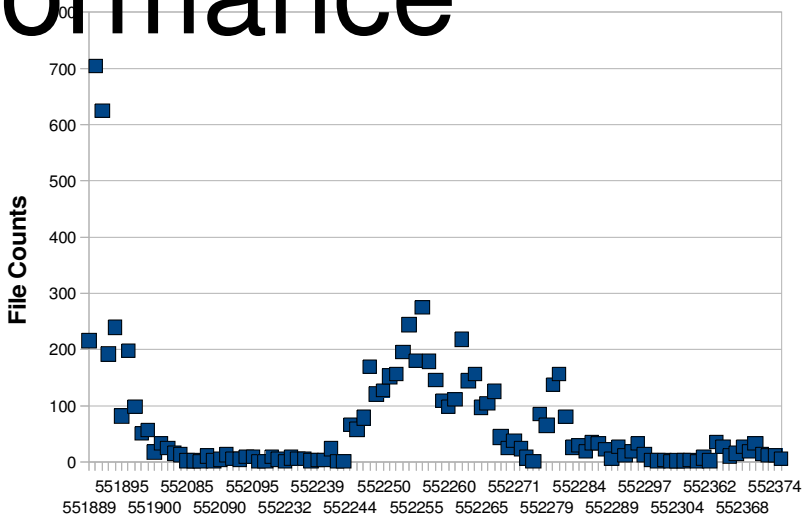
3 mounts = 18 minutes! Average about 73 MB / sec!

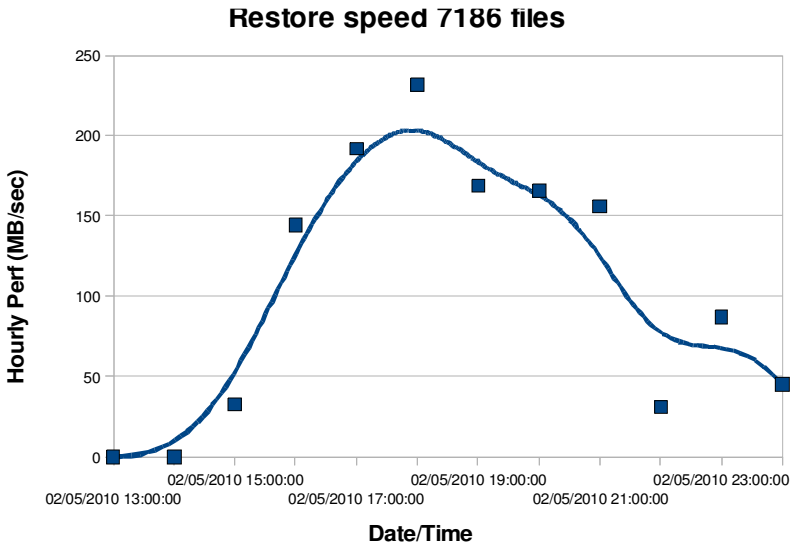


1 MB/sec → 72 MB/sec

Data Carousel performance

- Using user's own optimization option
  - ERADAT set to FIFO no conflicts
  - Carousel handles ordering and sorting by TapeID all tapes expected to be mounted once only
- Statistics based on RHIC/STAR Experiment February 5th 2010
  - 15 LTO-3 drives
  - 7187 files restored over 106 tapes, <size>=628 MB, total 4.4 TB
  - **All tapes 1.21 times**
  - So, why not 1?
    - Competition with other restore – HPSS competition for drives may make the low level kick out a tape to satisfy "the other guy's request"
    - HPSS has a mind of his own





Conclusion

- Tap access optimization is crucial – random access destroys your efficiency
- BNL has developed tools to optimize access –
  - One to two order of magnitude improvements
  - ERADAT (BNL Batch) has been developed in the RHIC data processing era
    - It has demonstrated a great performance for RHIC experiment (multi-context)
    - It has now been adopted by LHC/US-Atlas helping with data processing
  - DataCarousel also developed in house @ BNL
    - Out performs default BNL Batch; test bench for testing what would move “down” to batch
    - Best when faireshare in mind
- … Not the end of the story. In 2009, BNL Batch has been adapted into CCIN2P3 (called TReqS), and had a success story (HEPiX October 2009)
    - From the few month of our experience with TReqS:
    - Better resources usage (less mounting, more reading)
    - Sharing resources between experiments, ability to guarantee a minimum of drives used
    - Quicker file access implies less slow jobs
    - HPSS experts less stressed (shiny hairs, shiny smiles, lovely people)
- Future
  - Always better improvements … always faster …

DataCarousel & High demand

- Performance drops
- 1.98 mount / tape

starrdat — hpssbat.rcf.bnl.gov

PVR	Tape ID	# of Files	LSM	LSM Status	Status
Star Raw LTO-3	551889	398	1,8	ONLINE	MOUNTED
Star Raw LTO-3	551890	367	1,1	ONLINE	MOUNTED
Star Raw LTO-3	551893	318	1,1	ONLINE	MOUNTED
Star Raw LTO-3	551892	194	1,1	ONLINE	MOUNTED
Star Raw LTO-3	551899	108	1,8	ONLINE	MOUNTED
Star Raw LTO-3	551897	103	1,1	ONLINE	MOUNTED
Star Raw LTO-3	551900	100	1,1	ONLINE	MOUNTED
Star Raw LTO-3	552081	98	1,9	ONLINE	MOUNTED
Star Raw LTO-3	551895	71	1,9	ONLINE	MOUNTED
Star Raw LTO-3	551898	42	1,8	ONLINE	MOUNTED
Star Raw LTO-3	551894	38	1,9	ONLINE	
Star Raw LTO-3	551896	36	1,1	ONLINE	MOUNTED
Star Raw LTO-3	552082	26	1,10	ONLINE	
	TOTAL:	1899 files (13) tapes			

starrdat — hpssbat.rcf.bnl.gov

PVR	Tape ID	# of Files	LSM	LSM Status	Status
Star Raw LTO-3	551880	500	1,8	ONLINE	MOUNTED
Star Raw LTO-3	551903	630	1,1	ONLINE	MOUNTED
	551900	366	1,1	ONLINE	MOUNTED
	552092	156	1,10	ONLINE	
	552201	146	1,9	ONLINE	
	551092	145	1,1	ONLINE	MOUNTED
	552200	119	1,9	ONLINE	
	551894	105	1,9	ONLINE	
	551897	89	1,1	ONLINE	MOUNTED
	552084	81	1,10	ONLINE	
	552278	80	1,10	ONLINE	
	552282	74	1,10	ONLINE	
	552269	71	1,8	ONLINE	
	552271	63	1,8	ONLINE	
	552279	59	1,10	ONLINE	
	552270	51	1,10	ONLINE	
	552272	35	1,10	ONLINE	
	552268	33	1,9	ONLINE	
	551896	29	1,1	ONLINE	MOUNTED
	551895	27	1,9	ONLINE	MOUNTED
Star Raw LTO-3	551899	26	1,8	ONLINE	MOUNTED
Star Raw LTO-3	552262	26	1,10	ONLINE	
Star Raw LTO-3	552256	22	1,10	ONLINE	
Star Raw LTO-3	552259	21	1,9	ONLINE	
Star Raw LTO-3	552258	17	1,8	ONLINE	
Star Raw LTO-3	552263	12	1,10	ONLINE	
Star Raw LTO-3	552083	12	1,9	ONLINE	
Star Raw LTO-3	551900	10	1,1	ONLINE	MOUNTED
Star Raw LTO-3	552261	7	1,8	ONLINE	
Star Raw LTO-3	552265	5	1,11	ONLINE	
Star Raw LTO-3	552233	5	1,10	ONLINE