CERN Tape Status

Tape Operations Team
IT/FIO
CERN

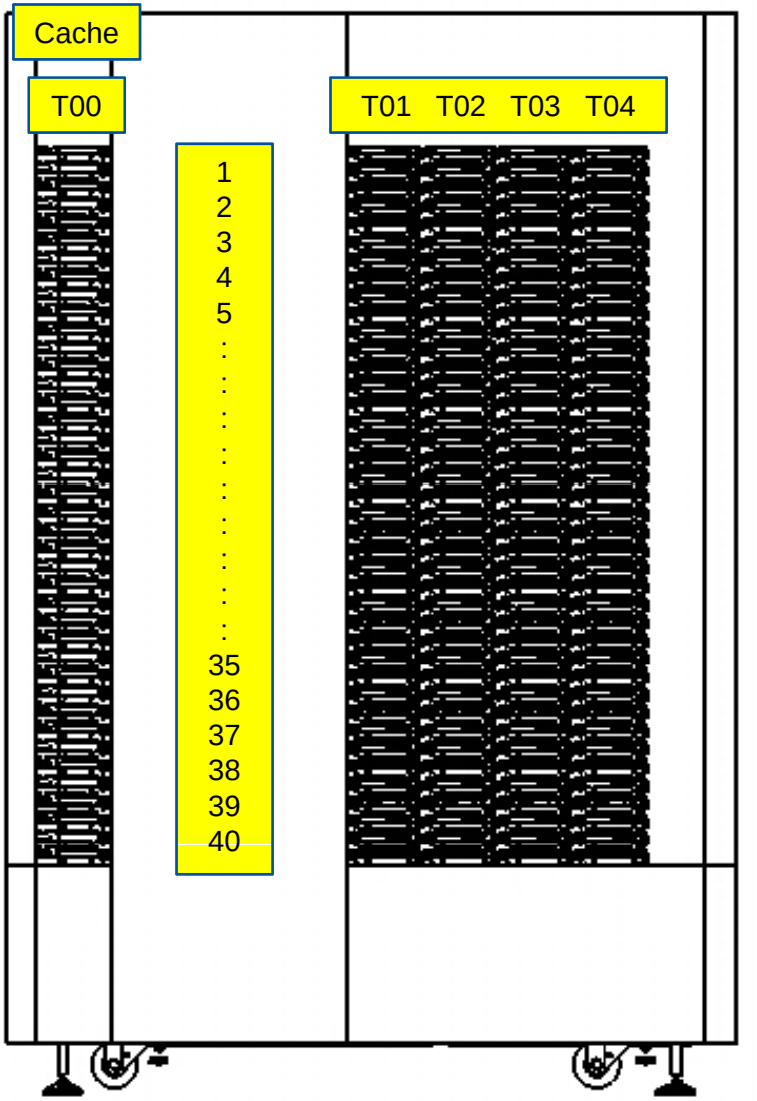
Agenda

- Hardware
- Low Level Tape Software
- Tape Logging and Metrics
- Repack Experience and Tools
- Some Thoughts on the Future

Hardware

- Migration of all drives to 1TB to be completed by end Q1 2009
  - 60 IBM Jaguar3 (700GB to 1TB @ 160MB/s)
  - 70 T10K B (500GB to 1TB @ 130MB/s)
- IBM High Density Robot in production
  - Improve GB/m^2 by a factor of 3
- Move to blade based tape servers
  - Improve power efficiency
  - Reduce unused memory and CPU
- Visited IBM and Sun US labs in Q4/2008
  - Tape market is strong especially with new regulatory requirements

IBM HD Frame



CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Low Level Tape Changes

- During past 6 months, the following changes have been included into the 2.1.8.X tape server.
  - blank tape detection has been improved and CERN now uses tplabel
  - st/sg mapping has been fixed to cope with driver removal and reinsertion (which changed the order in /proc/scsi/scsi)
  - large messages from network security scans do not crash the SCSI media changer rmc daemon anymore
  - added an option to ignore errors on unload
  - **added an option to detect and abort too long position (where driver return status OK, but positions the tape incorrectly and are hence suspected to overwrite data)**
  - added support for the 1000GC tapes
  - added central syslog logging to rmc daemon for central tape logging
- CERN moved tape servers to 2.1.8-5 last week
  - Running well with Castor 2.1.7 stagers

Tape Log Database

- Provides a central log of all tape related messages and performance
  - LHCC Metrics to SLS
    - Number of drives / VO
    - File sizes
    - ...
  - Problem investigations
    - When was this tape mounted recently ?
    - Has this drive reported errors ?
    - ...
  - Automated problem reporting and action
    - Library failure
    - Tape or drive disable
    - ...
  - GUI for data visualisation (Work In Progress)
    - Graphs
    - Annotate comments such as ‘sent tape for repair’
- **Note:** This is a CERN tape operations tool rather than part of the Castor development deliverable. The source code is available if other sites want to use it for inspiration but no support is available.

Service Level Status overview - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://sls.cern.ch/sls/service.php?id=CASTORTapeInfrastructure

TsiSection < FIOgroup... | CERN TSM Manageme... | Lemon Monitoring We... | c2repack | Service Level Status ov... | CASTOR - Bugs: Brows... | CERN Problem Report ...

Service Level Status overview

CASTOR Tape Infrastructure

6 Feb 2009 Fri 13:27:49

Service information

full name: **CASTOR Tape Infrastructure**

group: IT-FIO

email: **Tape.Support@cern.ch**

web site: http://cern.ch/it-dep-fio-ds/Documentation/tapedrive/We...

service managers: Vlado Bahyl, Giuseppe Lo Re, Gordon Lee

Service availability (more)

availability:

percentage: 92%

availability info: 156 tape drives in total, 145 fully operational

status: **available**

last update: 13:21:53, 6 Feb 2009 (6 minutes ago)

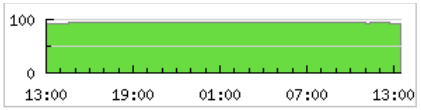
expires after: 30 minutes

rss feed with status changes

how is availability measured or estimated:

CASTOR Tape Service is fully available if > 80% of drives are in production. Drive status in each library (DGN) is shown in additional service information.

availability in the last 24 hours (more):



Service performance

Key Performance Indicators:

Reserve of free supply tapes in days:

Additional service information (more)

service notes:

Availablity of drives in each tape library (DGN), Tape Supply Pool status and Tape Space Usage by VO

Drive Availability in DGN T10KR1 [14 drives]:	13
Drive Availability in DGN T10K60 [16 drives]:	16
Drive Availability in DGN T10KB5 [33 drives]:	27
Drive Availability in DGN T10KB6 [32 drives]:	29
Drive Availability in DGN 3592B1 [17 drives]:	16
Drive Availability in DGN 3592B2 [24 drives]:	24
Drive Availability in DGN 3592B3 [20 drives]:	20
Free 0GB tapes in pool supply_3592B1 :	0
Free 0GB tapes in pool supply_3592B2 :	0
Free 1000GB tapes in pool supply_3592B3 :	1405
Free 1000GB tapes in pool supply_JB123 :	446
Free 1000GB tapes in pool supply_T10K56 :	39
Free 1000GB tapes in pool supply_T10K60 :	3428
Free 1000GB tapes in pool supply_T10KR1 :	3208
Total Free Tape Capacity TB all pools:	8528
Rate of tape consumption in GB per day:	9266

CGI time 3 secs

Part of (subservice of):

Services for physics

IT/FIO services

Subservices

CASTOR Tape ATLAS

CASTOR Tape REPACK

CASTOR Tape LHCB

CASTOR Tape ALICE

CASTOR Tape NA48

CASTOR Tape NTOF

CASTOR Tape COMPASS

CASTOR Tape CMS

Clusters, subclusters and nodes

cluster tapeserver

subcluster tapeserver / T10KR1

subcluster tapeserver / T10K60

subcluster tapeserver / T10KB5

subcluster tapeserver / T10KB6

subcluster tapeserver / 3592B1

subcluster tapeserver / 3592B2

subcluster tapeserver / 3592B3

Depends on

this service uses:

CASTOR Service

Depended on by

none / not declared

Find: Next Previous Highlight all Match case

Done

Service Level Status overview - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://sls.cern.ch/sls/service.php?id=CASTORTapeATLAS

TsiSection < FIOgroup... | CERN TSM Manageme... | Lemon Monitoring We... | c2repack | Service Level Status ov... | CASTOR - Bugs: Brows... | CERN Problem Report ...

Google News | Linux Today - Li... | EDH - Tasks (Ti... | Lemon Monitori... | TsiMarsSchedul... | MARS Overview | Appraisal Report... | DiskRefAdaptec ... | Service Lev...

Service Level Status overview

Service information

full name: **CASTOR Tape ATLAS**

group: IT-FIO

email: **Tape.Support@cern.ch**

web site: http://cern.ch/it-dep-fio-ds/Documentation/tapedrive/We...

service managers: Vlado Bahyl, Giuseppe Lo Re, Gordon Lee

Part of (subservice of):

CASTOR Tape Infrastructure

Subservices

none / not declared

Clusters, subclusters and nodes

none / not declared

Depends on

this service uses:

CASTOR Service

Depended on by

none / not declared

Service availability (more)

availability:

percentage: 90%

availability info: Based on average wait times for file read/write

status: **available**

last update: 13:21:44, 6 Feb 2009 (9 minutes ago)

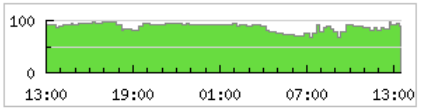
expires after: 30 minutes

rss feed with status changes

how is availability measured or estimated:

ATLAS Performance Metrics in the use of the Castor Tape Services. Data volumes in GB and transfer rates in MB/sec.

availability in the last 24 hours (more):



Additional service information (more)

DATA READ: data volume read in GB:	968
DATA READ: number of files transferred:	992
DATA READ: average filesize in MB:	976
DATA WRITE: data volume written in GB:	3036
DATA WRITE: number of files transferred:	2314
DATA WRITE: average filesize in MB:	1312
RATE READ: read transfer rate inc drive overhead MB/sec:	34
RATE READ: drive read transfer rate MB/sec:	60
RATE WRITE: write transfer rate inc drive overhead MB/sec:	48
RATE WRITE: drive write transfer rate MB/sec:	60
TAPE MOUNT: successful mounts in last 4hrs :	151
TAPE MOUNT: failed mounts in last 4hrs :	3
TAPE MOUNT: read mounts in last 4hrs :	62
TAPE MOUNT: write mounts in last 4hrs :	92
TAPE MOUNT: average tape mount time in secs:	58
TAPE QUEUES: average wait for read in secs :	6086
TAPE QUEUES: average wait for write in secs :	711
FILES PER MOUNT: read average :	5.7
FILES PER MOUNT: write average :	28.8
TAPE VOLUME for ATLAS in TB [capacity 2840] :	2452
TAPE REPEAT MOUNT: read average last 24hrs:	2.15
TAPE REPEAT MOUNT: write average last 24hrs:	6.58
TAPE FRAGMENTATION: Percentage full level for ATLAS tapes:	78.6

Values are averages over the last 4 hrs

Data Vol in GB, File Sizes in MB and Rates in MB/sec

Find: rdma | Next | Previous | Highlight all | Match case

Done

History view - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apex.cern.ch/pls/htmldb_tapelog/f?p=105:14:222170607723698:P14_GO

a screen shout with kde

TapelogWeb

GLORE Logout

Menu

- Home
  - Accounting
    - Graphs
    - Tables
    - Details
  - Errors
    - Top 10s
    - Details
  - History
    - View
    - Modify

Start 30-JAN-2009 10:38:23 End 04-FEB-2009 10:38:23 Tape T03255 Operator - All -

Go

TIME	TAPE	FIELD	OLDVALUE	NEWVALUE	OPERATOR	WORKLOG
2009/02/03 08:12:26	T03255	REPACKSTATUS	FAILED	ONGOING	CASTOR	
2009/02/03 08:41:15	T03255	REPACKSTATUS	ONGOING	DONE	CASTOR	
2009/02/03 08:42:30	T03255	DENSITY	500GC	1000GC	CASTOR	
2009/02/03 08:42:30	T03255	LIBRARY	SL8600_0	SL8600_1	CASTOR	
2009/02/03 08:42:30	T03255	STATUS	FULL		CASTOR	
2009/02/03 08:42:31	T03255	TAPEPOOL	r_lep_data	tolabel_T10K60	CASTOR	
2009/02/03 16:28:24	T03255				GLORE	This morning I have disabled the segment of file /castor/cern.ch /chorus/tape/SR0001/SR0001.72.sl [root@c2repacksrv101 ~]# nssetsegment -c 1 -s 1 -d /castor/cern.ch/chorus/tape/SR0001 /SR0001.72.sl

1-7

Done

Accounting graphs - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apex.cern.ch/pls/htmldb_tapelog/f?p=105:24:222170607723698:P24_GO

a screen shout with kde

TapelogWeb

GLORE Logout

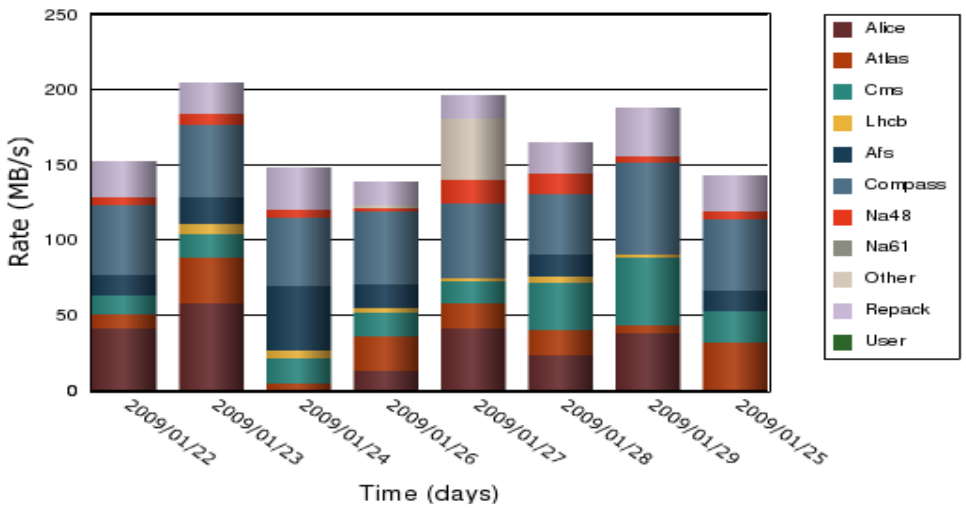
Menu

- Home
  - Accounting
    - Graphs
    - Tables
    - Details
  - Errors
    - Top 10s
    - Details
  - History
    - View
    - Modify

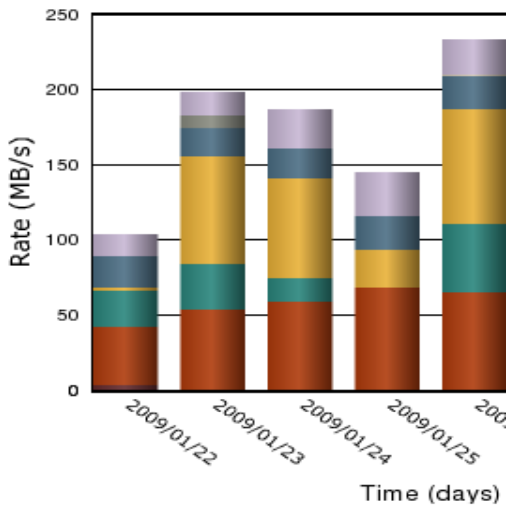
Start 22-JAN-2009 10:40:00 Start 04-FEB-2009 10:40:37 Graph Type Rate vs VO & Time

Go

Read rate vs VO & Time



Write rate vs VO & Time



Transferring data from apex.cern.ch...

Accounting graphs - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://apex.cern.ch/pls/htmldb_tapelog/f?p=105:24:222170607723698:P24_GO

a screen shout with kde

TapelogWeb

GLORE Logout

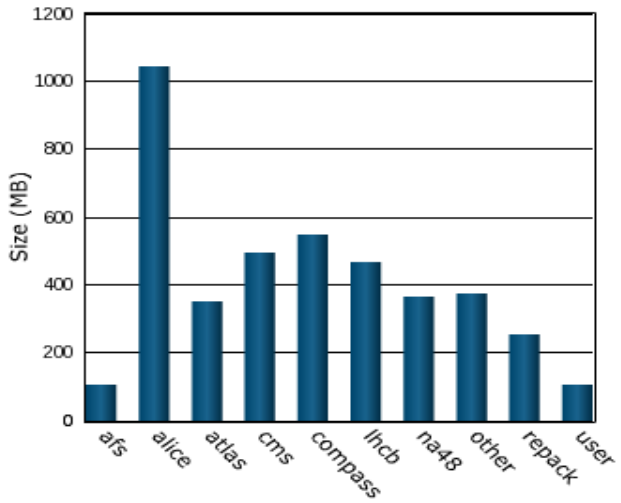
Menu

- Home
  - Accounting
    - Graphs
    - Tables
    - Details
  - Errors
    - Top 10s
    - Details
  - History
    - View
    - Modify

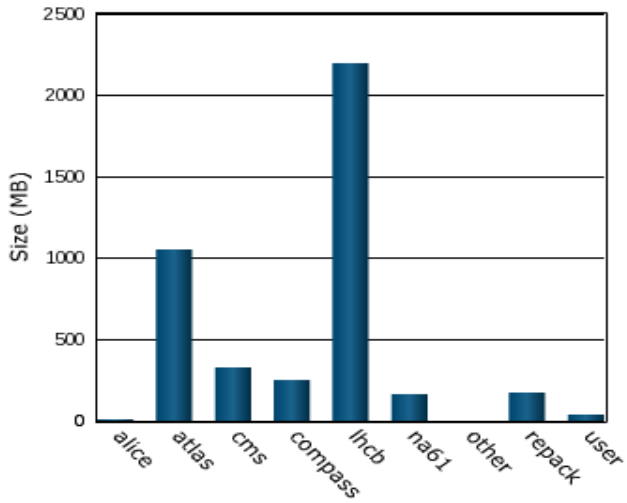
Start 22-JAN-2009 10:40:00 Start 04-FEB-2009 10:40:37 Graph Type File size

Go

Average file size (read) vs VO



Average file size (write) vs VO



Average file size (read) vs DGN

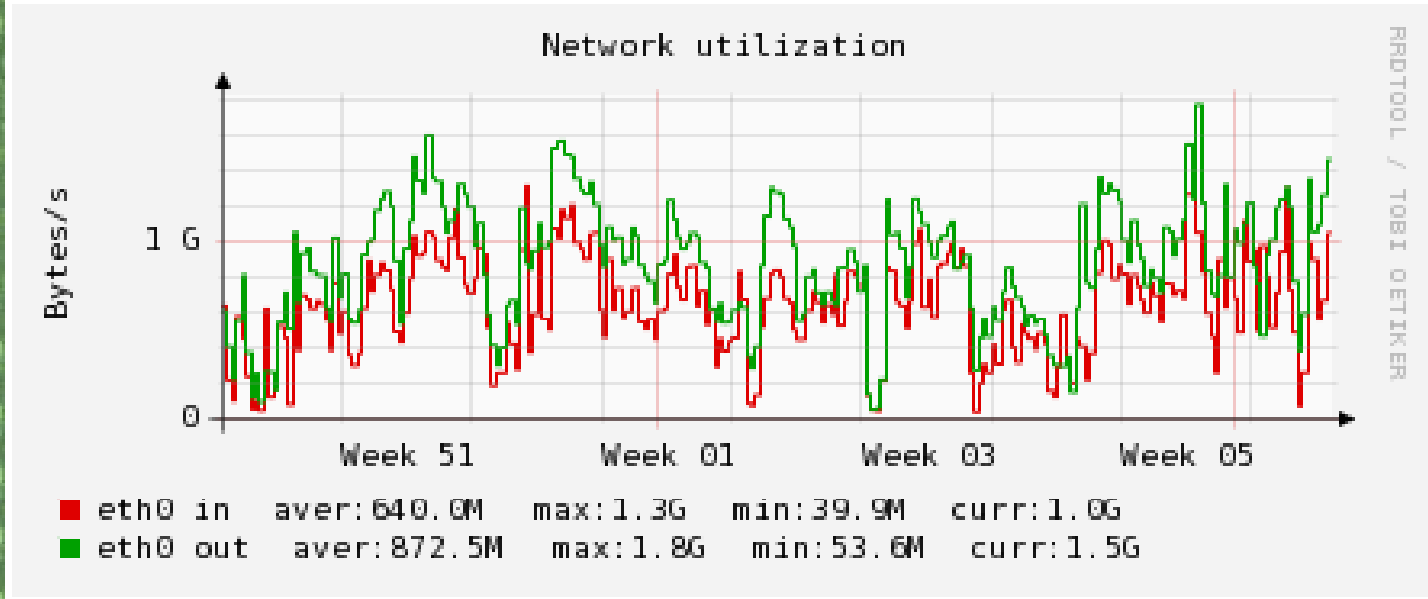
Average file size (write) vs DGN

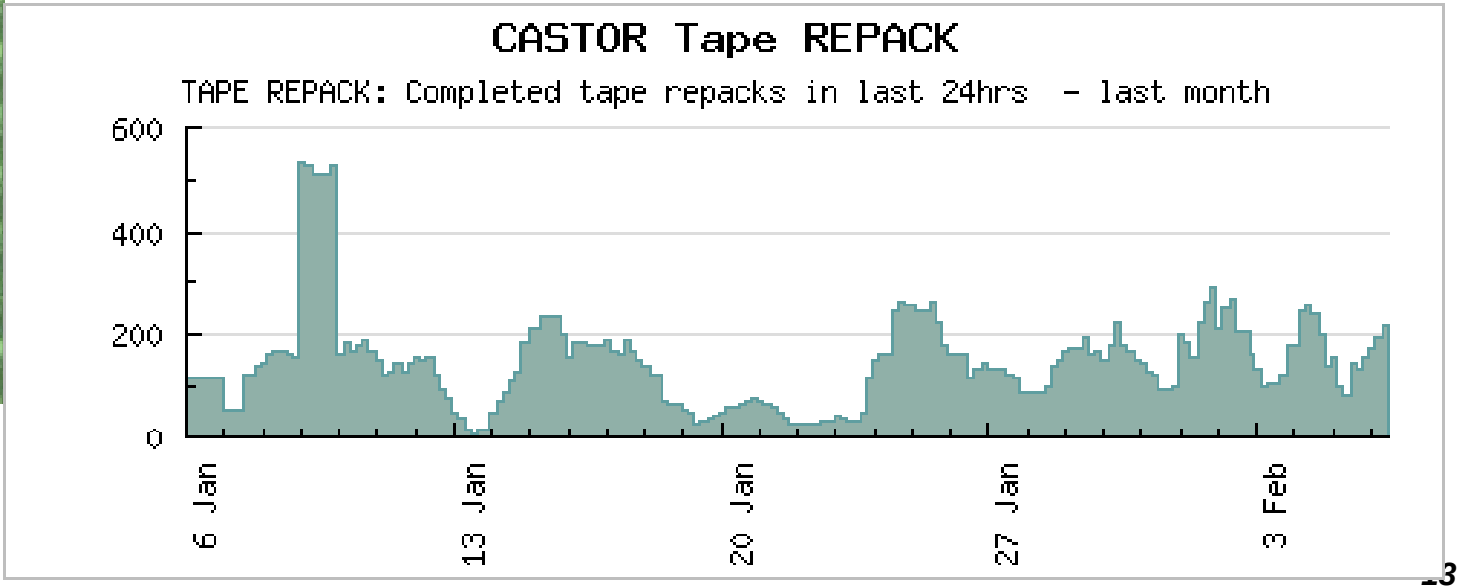
Done

Repack – Configuration

- Dedicated castor instance for repack
  - Isolated load from other stagers following experience sharing with public instance
  - Simplified debugging
  - Allowed easier intervention planning and upgraded
- Configuration
  - 20 dedicated disk servers (12 TB raw each)
  - 150 service classes
  - Single headnode
  - No LSF
  - Dedicated policies, service classes and tape pools for repack to ensure no mixing of data between pools
- 22,000 tapes at start of intensive run in December 2008 with Castor 2.1.8-3

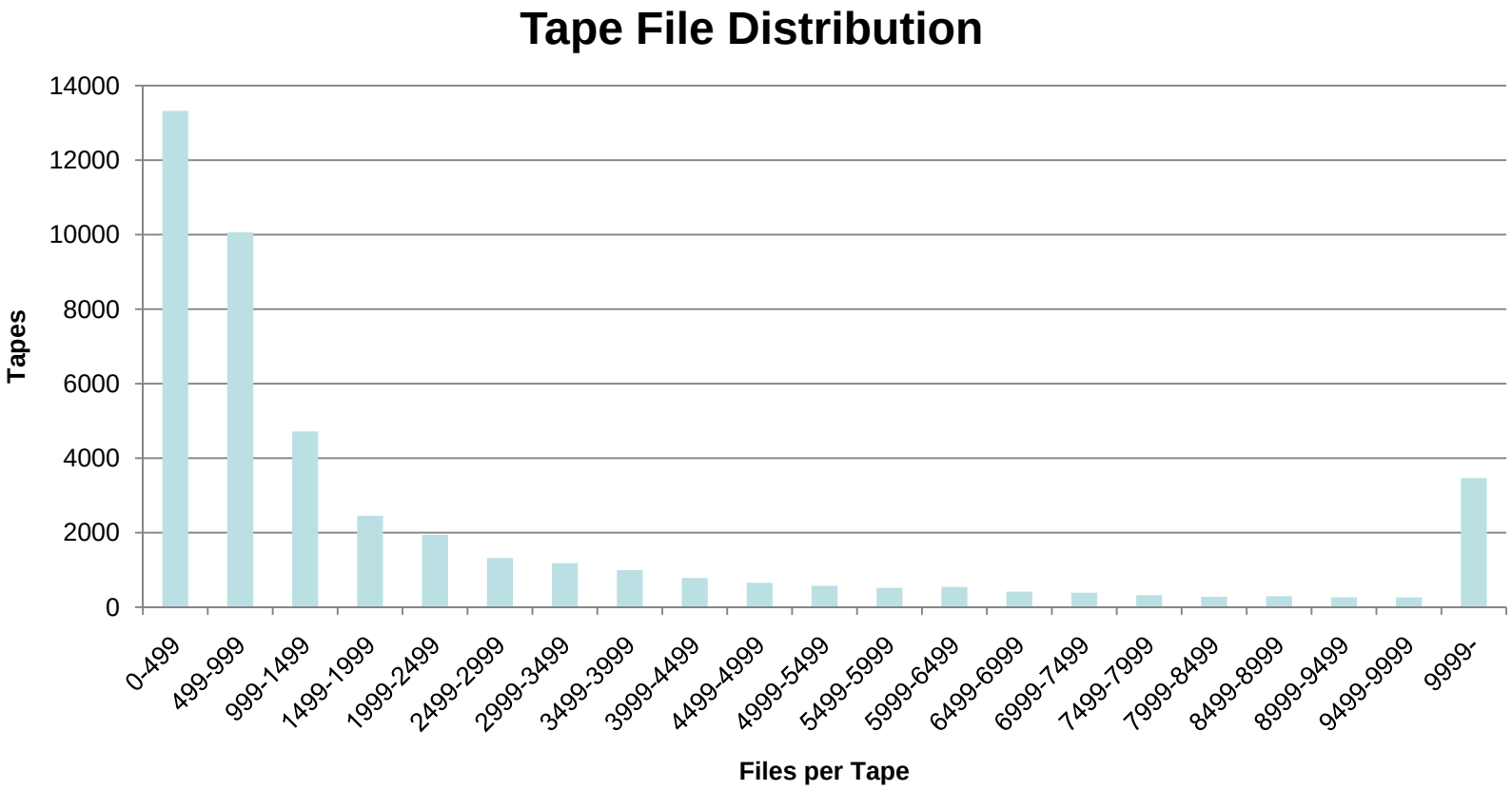
Repack – At Work





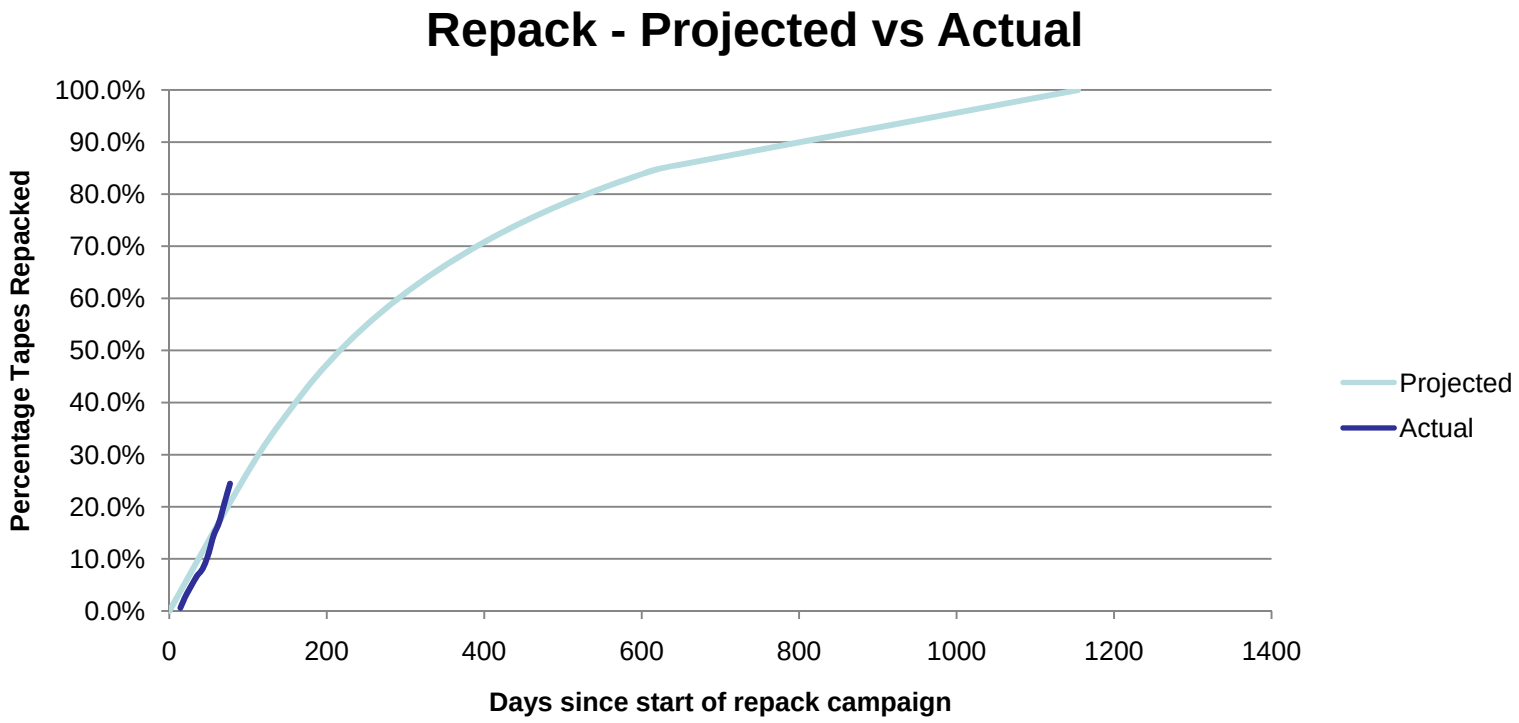
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Repack – File Sizes



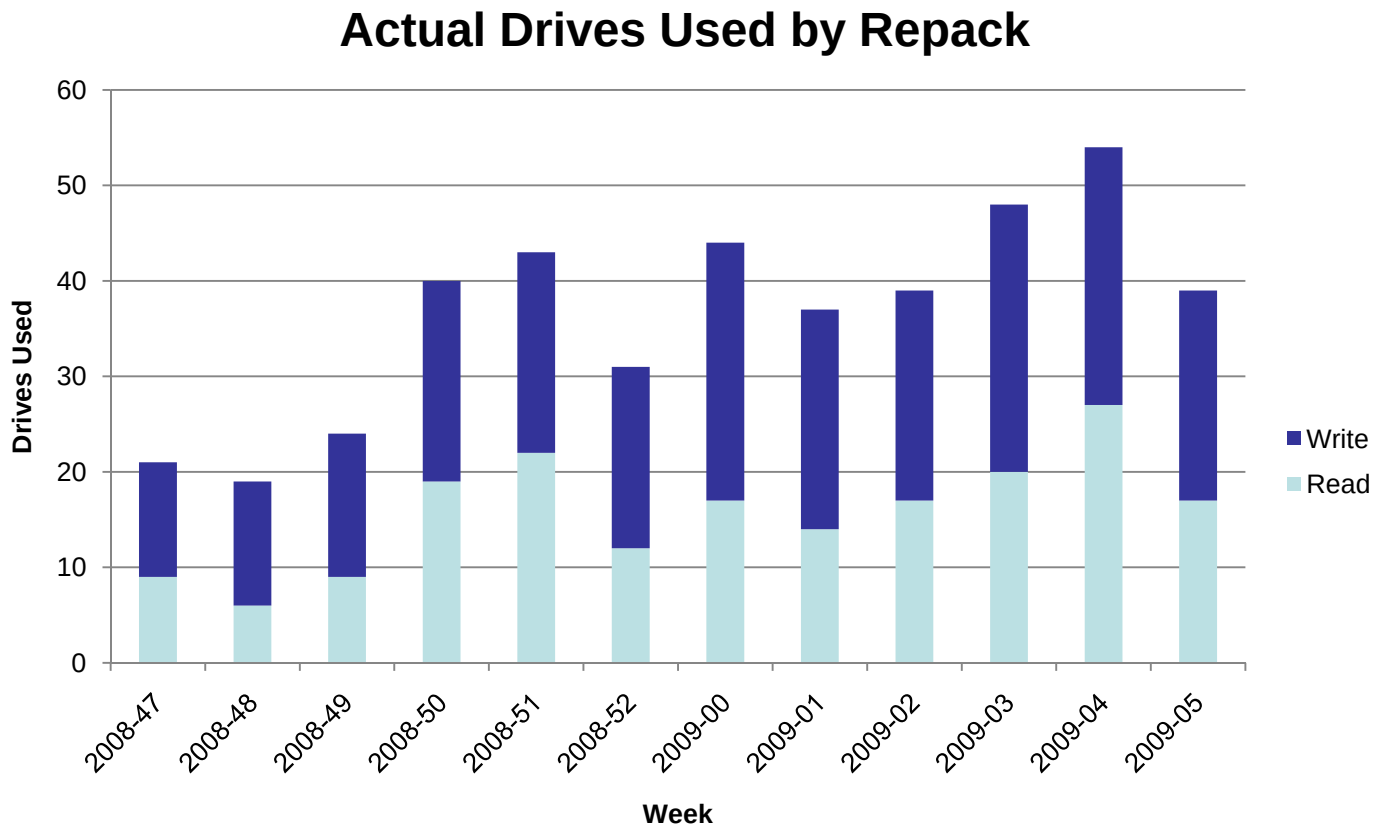
- Outlook is better than we thought in 2008 since no data taking so far
- Basic problem remains that file sizes, especially legacy files, are small so slow write performance

Repack – Progress So Far



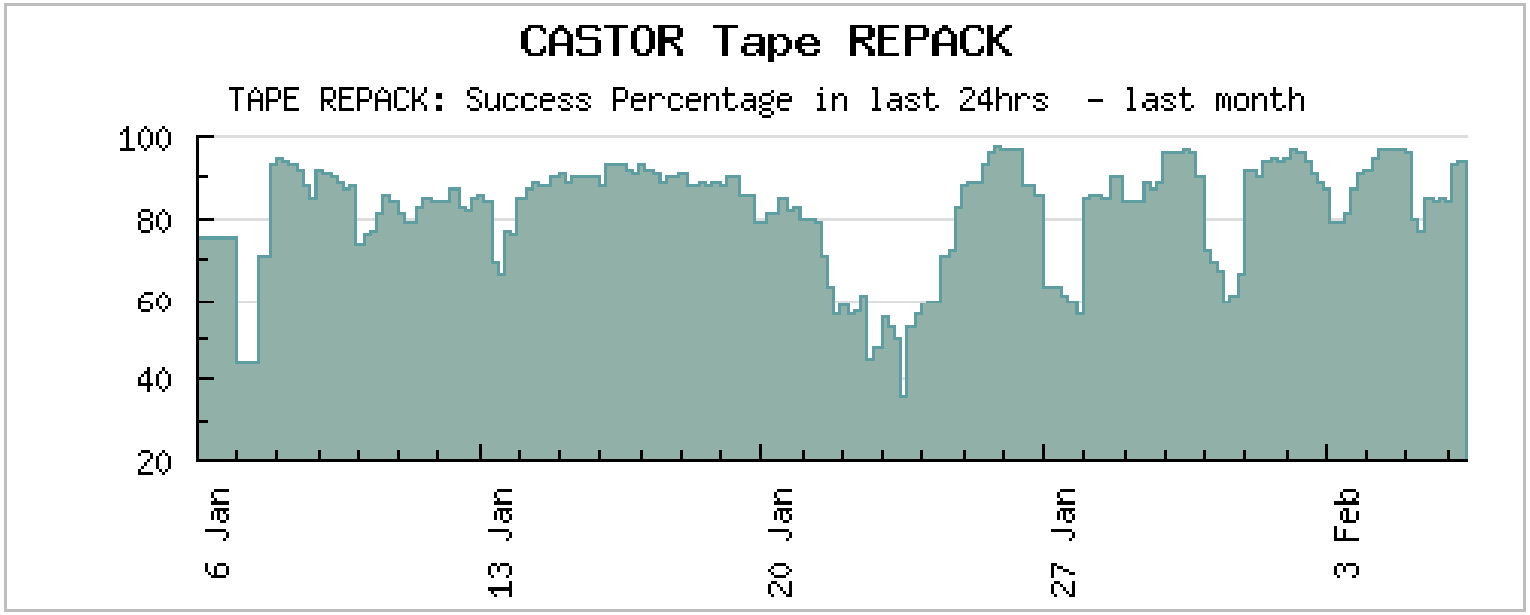
- Projections based on 20 drives used as this was the most we felt we could use while also doing data taking
  - Repack 60% of the tapes would take a year
  - Completion unlikely before next tape drive model
- Actual repack progress is close to projected
  - but.....

Repack – Drive Usage



- Twice as many drives as planned for projected rate
  - Not clear why data rate is slower
- Luckily, demand is low currently but expected to increase soon

Repack – Success Rate



- Success percentage is defined as the number of tapes which can be repacked 1st go without any human intervention
- Around 1 in 5 tapes fails to repack completely
  - Stalled repacks when streams became blocked
  - Multiple copy tapes
  - Bad name server file sizes compared to tape segments
  - Media/tape errors are occasional

Repack – Difficult Case Example

- Repack of tape gives on 1 failed file
- Check logs which indicates a difference between name server and tape segment file size
- Recall the file from tape using tpread and check size and checksum
- Set the file size in name server and clear the checksum .. Is it the right file/owner ?
- Repack the tape
- Stager reports bad file size
- Fix the file size in the stager and remove staged copy using SQL scripts
- Repack the tape, file still does not migrate
- Report issue to development sr#107802
- Manually stage file completed OK ...
  - 5 hours work to recover one file ...

Repack – Thoughts

- Repack has been an effective stress test for Castor, finding many issues in the name server, stager and tape stream handling
- The basic repack engine requires regular surveillance to keep busy but not too over-loaded.
  - 5K LOC of scripts to select, submit, reclaim and try to guess the failure causes. A small part of this function is now included in the repack server
  - Keeping streams balanced across pools to avoid long queues, device group hot spots and user starvation
  - 1 FTE required to tweak the load, analyse the problems and clean up failed repacks
- Selecting the large file tapes first has helped
  - Larger files to get good data rates
  - A 10,000 file tape can take several hours to get started
- We've been able to benefit from the delayed start up but it is not likely to continue so quietly in the next months

Crystal Ball – 2009/10...

- Main Goals for the year
  - 1TB everywhere
  - Repack as much as we can
  - SLC5
- Finish phase out
  - Sun 9940B infrastructure
  - Sun T10000A drives
  - IBM Jaguar 2 (upgrade to Jaguar 3)
- Extended run is planned to produce 30PB
  - Need 15,000 new library slots+media
  - Installation must be non disruptive if after September
- Tape Operations team will be reduced to 2 FTE in 2010
  - Was 4 FTE in 2008

Further out … 2011 / 2012

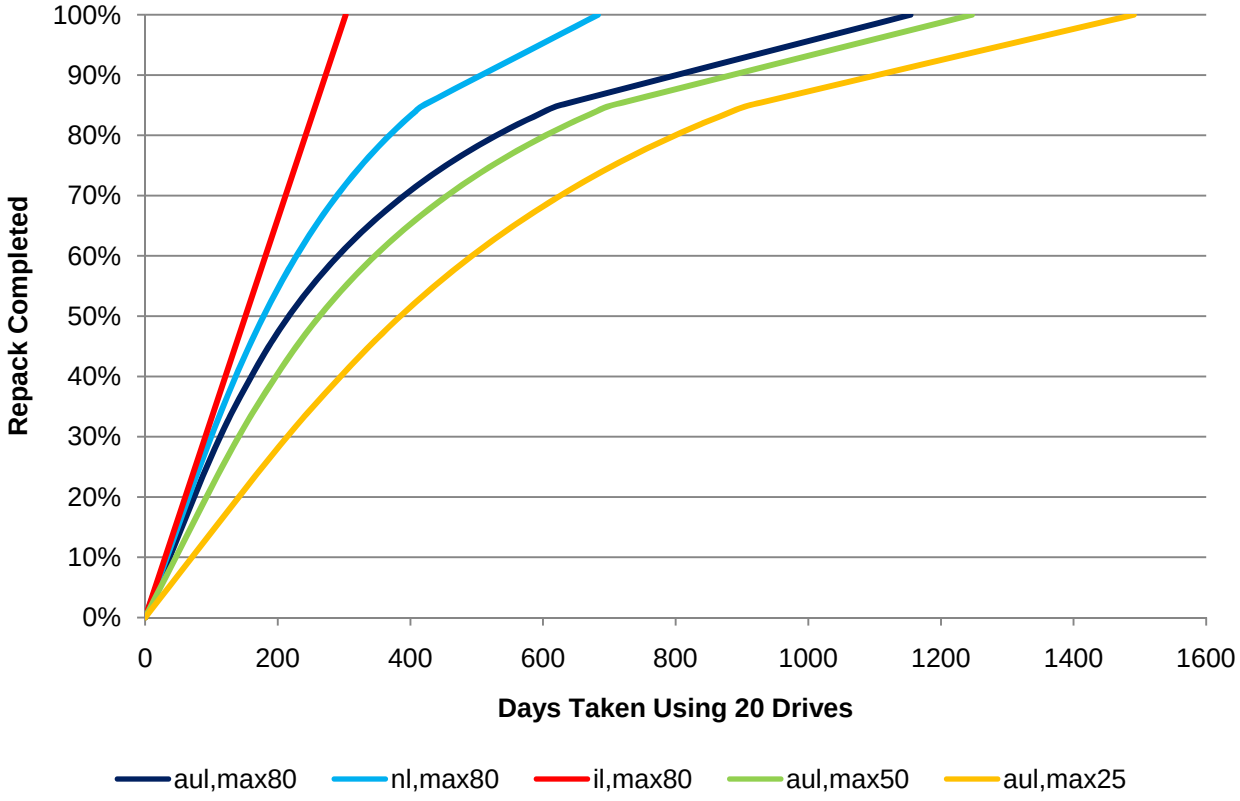
- Expecting 15 PB/year
- Upgrade remaining IBM libraries to HD frames
  - How to do it while still full of tapes ?
- New contract for tape robotics
- New computer centre
- New drives, new media, same libraries ?
  - LTO-5 ? Media Re-use ?
- Repack 50,000 cartridges (and re-sell if possible) ?
  - Or buy more libraries ?

Conclusions

- 2008 saw major changes in drive technology which are now completed
- CERN tape infrastructure is ready for data taking this year
- Getting bulk repack into production has been hard work but should benefit overall Castor 2.1.8 stability
- Repacking to completion seems very unlikely during 2009 and will have to compete with experiments for drive time
- Continued pressure on staffing forces continued investment in automation

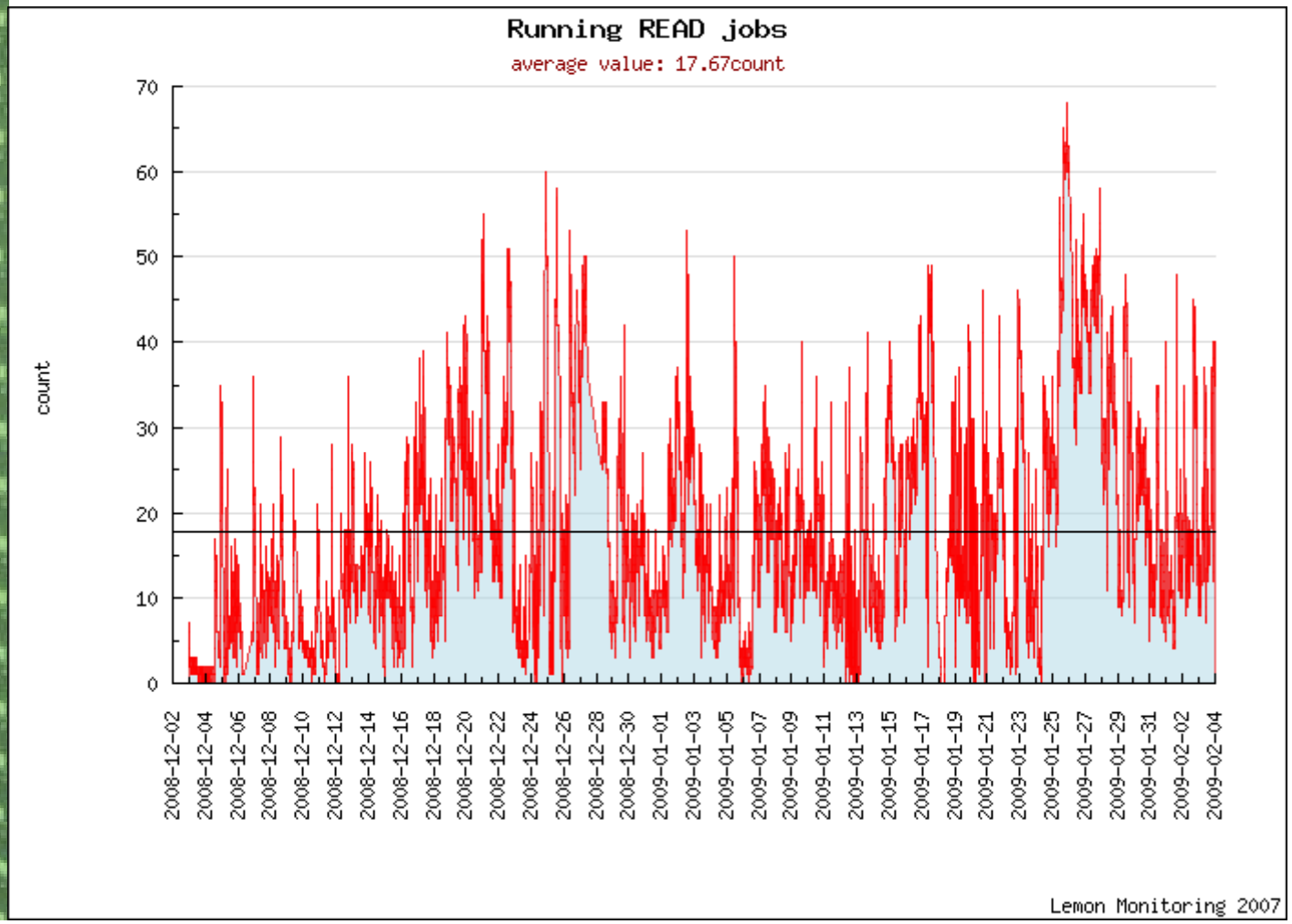
Backup Slides

Repack – 2008 theory

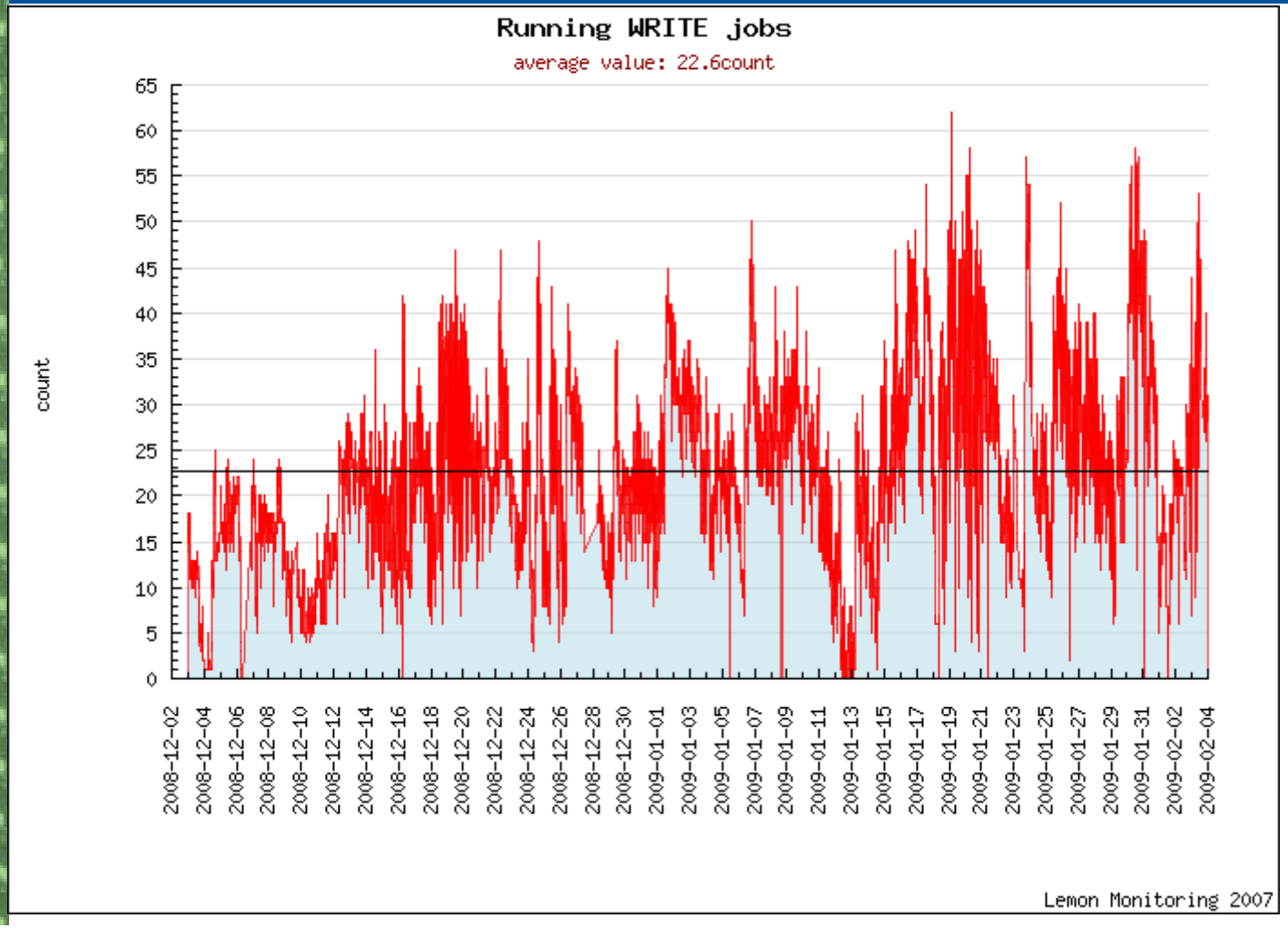


- aul,max80 corresponds to AUL label format with 80MB/s read rate, around 3 years to complete

Repack – Read Drive Usage



Repack – Write Drive Usage



Free Space Trends