Tape Performance at T1 Sites

John Gordon, STFC-RAL

GDB meeting @CERN February 11th 2009

Science & Technology Facilities Council

Outline

- I do not believe that WLCG has yet shown that it can deliver the tape performance required by the LHC experiments for their custodial data.
- Discuss today, seek more information
- Present status and plans to LHCC mini-Review 16th February

- This is next week. Do we have any plans?

I asked the sites.

a) do you feel your tape system has been stressed by the LHC experiments yet?

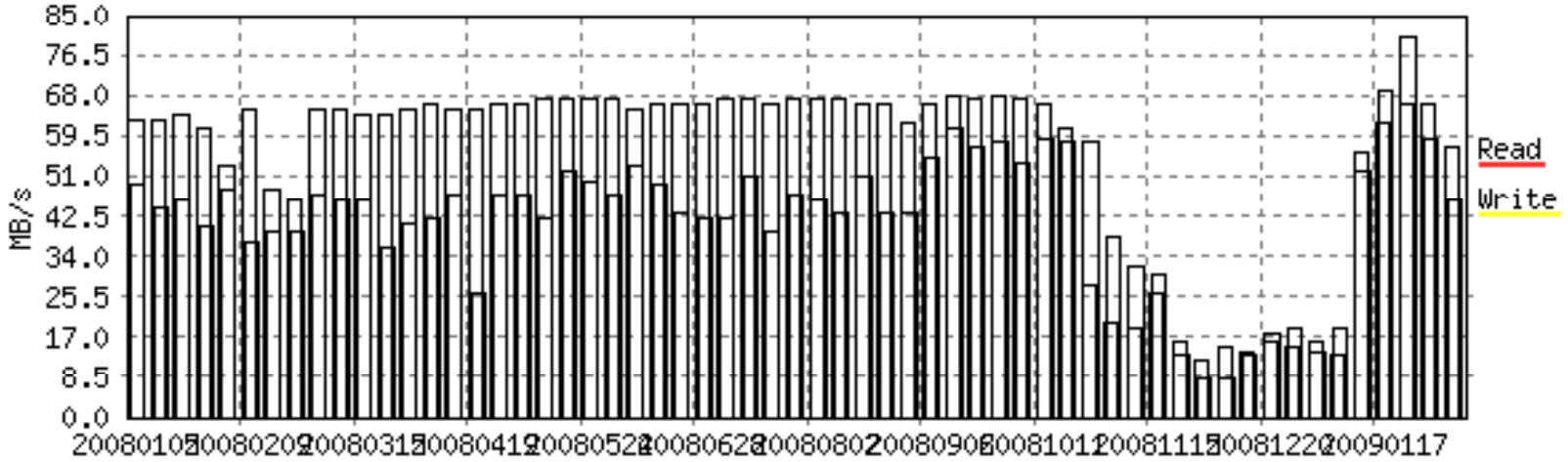
b) do you have a method of dynamically seeing the performance of your service to tape?

FNAL (new)

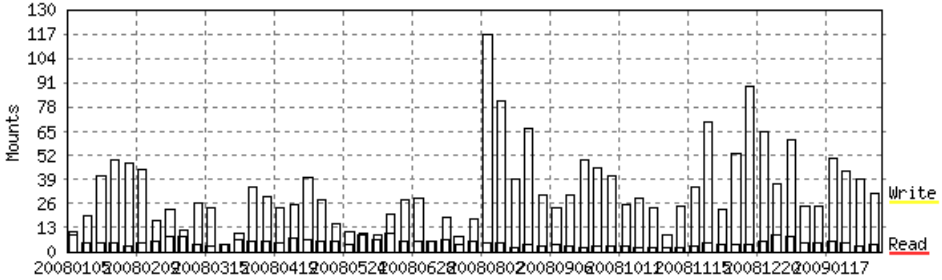
- A) Yes
- B) Yes

GDB

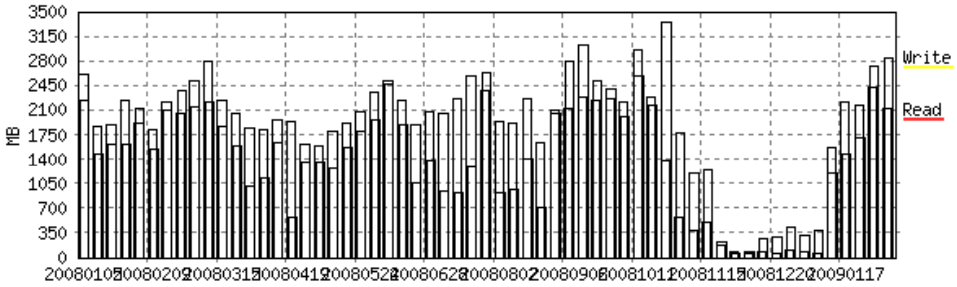
Total Data Rate for CMS



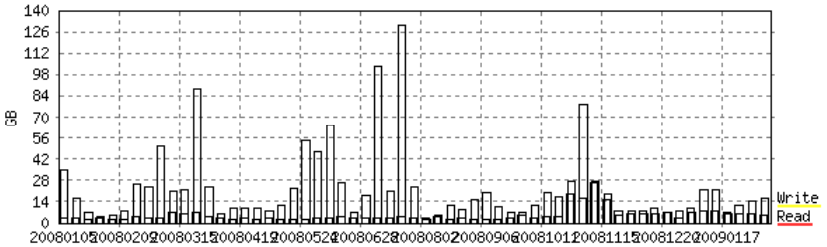
Repeat Mounts per Tape for CMS



Average File Size for CMS



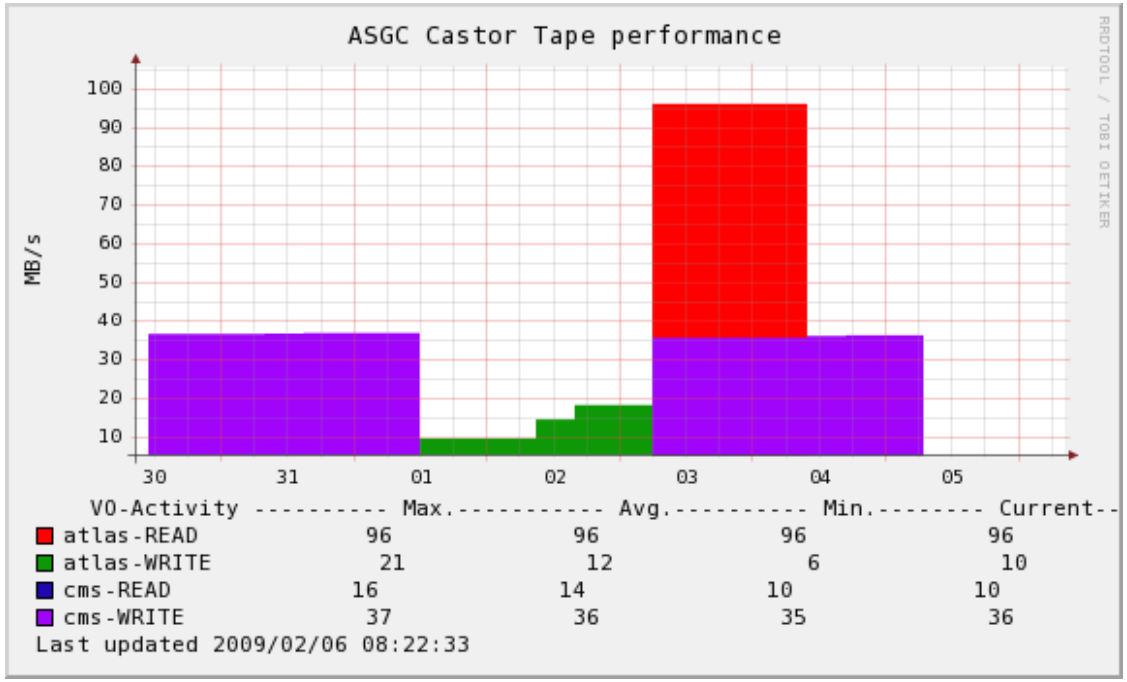
Data Transferred per Mount for CMS



ASGC (new)

Table I: tape performance metrics wrt to VO, the tape pools:

vo	data-vol	files	wall-time	drive-time	mount-time	pos-time	trans-time	umount-time
atlas	494915	3394	189999	122911	5031	19111	88560	13370
cms	2.93E+06	6793	23946	241741	17174	45571	140484	53920
atlasPrdtp	494912	3338	188631	120432	3782	18727	88119	11863
atlasMCtp	2.24	56	1368	2479	1249	384	441	1507
cmsCSAtp	2.93E+06	6793	23946	241741	17174	45571	140484	53920



Science & Technology Facilities Council

RAL (new)

- A) Have the experiments tested you at the required rates?
- No. certainly I suspect that ATLAS and CMS have not fully tested the tape system in realistic situations and load. Worse I do not believe that we have seen combined running of all VOs running a realistic mix of work.
- B) Can you monitor the rates you achieve to tape (i) in aggregate and (ii) for a particular experiment?
- Yes – although not realtime. See:
- https://twiki.cern.ch/twiki/bin/view/LCG/MssEfficiencyUK-T1-RAL

Science & Technology Facilities Council

RAL (2)

- It remains very hard for a Tier-1 to realistically project tape drive demand into the future. Our projections need to be 3-5 years in order to properly plan financial profiles and technology roadmaps but I find it hard even to fully determine demand expected for the first full year of data taking.

PIC (new)

- a) Have the experiments tested you at the required rates?
- I think yes. But I think that, even if the rate was nominal, I have the feeling the tests were not realistic because:

  1-the access pattern was not realistic: the tests for reads for instance always launch massive prestages of quite "dummy" data, and up to now the exercise finished there. There were no reconstruction jobs consuming the data at the other end, so the management of the disk cache was not tested at all (which is the difficult bit I think).

  2-these tests never coincided in time for the various VOs. So, each VO had effectively almost all the tape drives available for them when they launched their tests. The vo-sharing of the tape drives has not been tested at all.

PIC 2

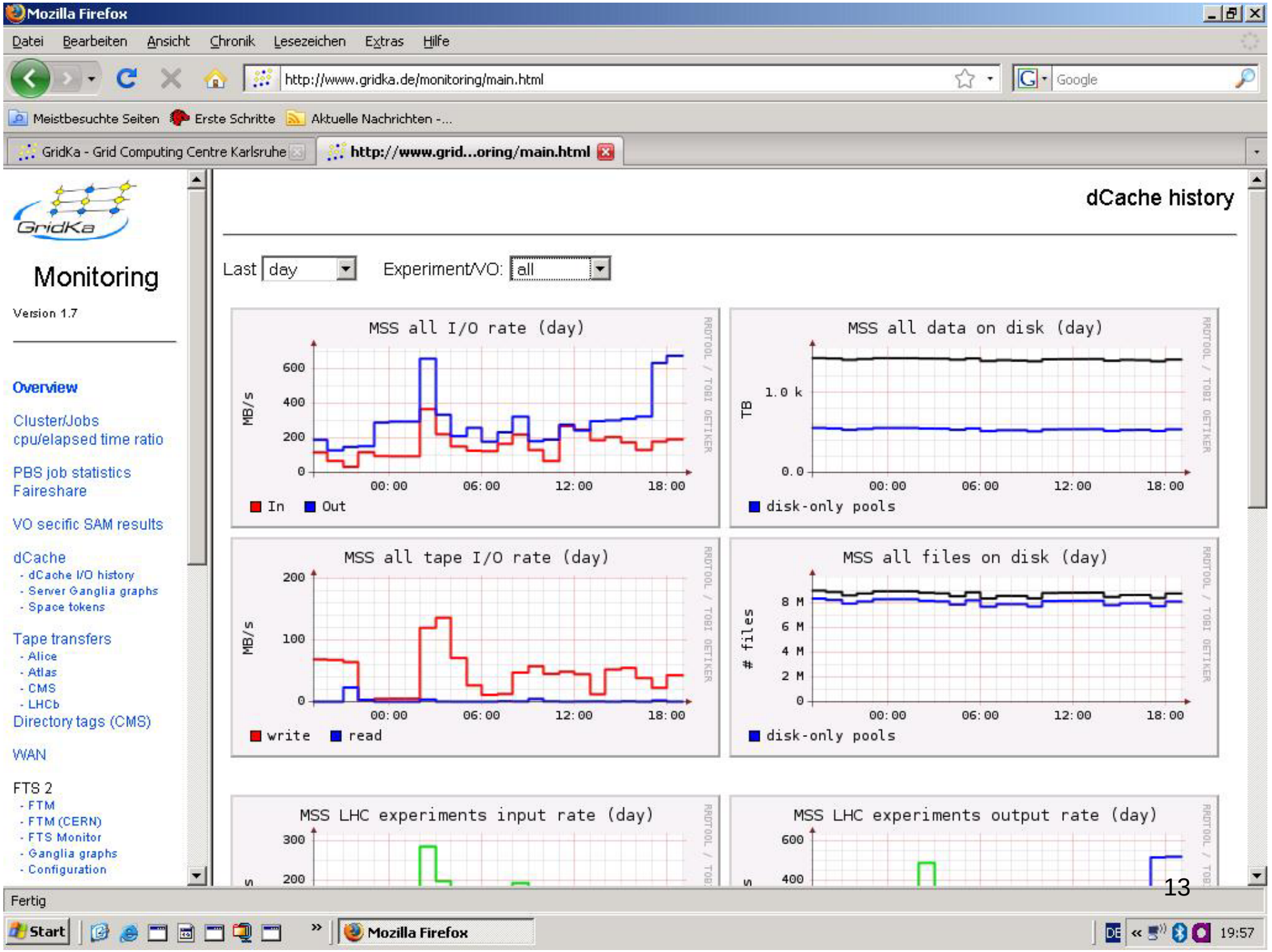
- b) Can you monitor the rates you achieve to tape (i) in aggregate and (ii) for a particular experiment?
-
- Yes. These are the numbers we provide in the MB wiki with the 4 agreed tape metrics.

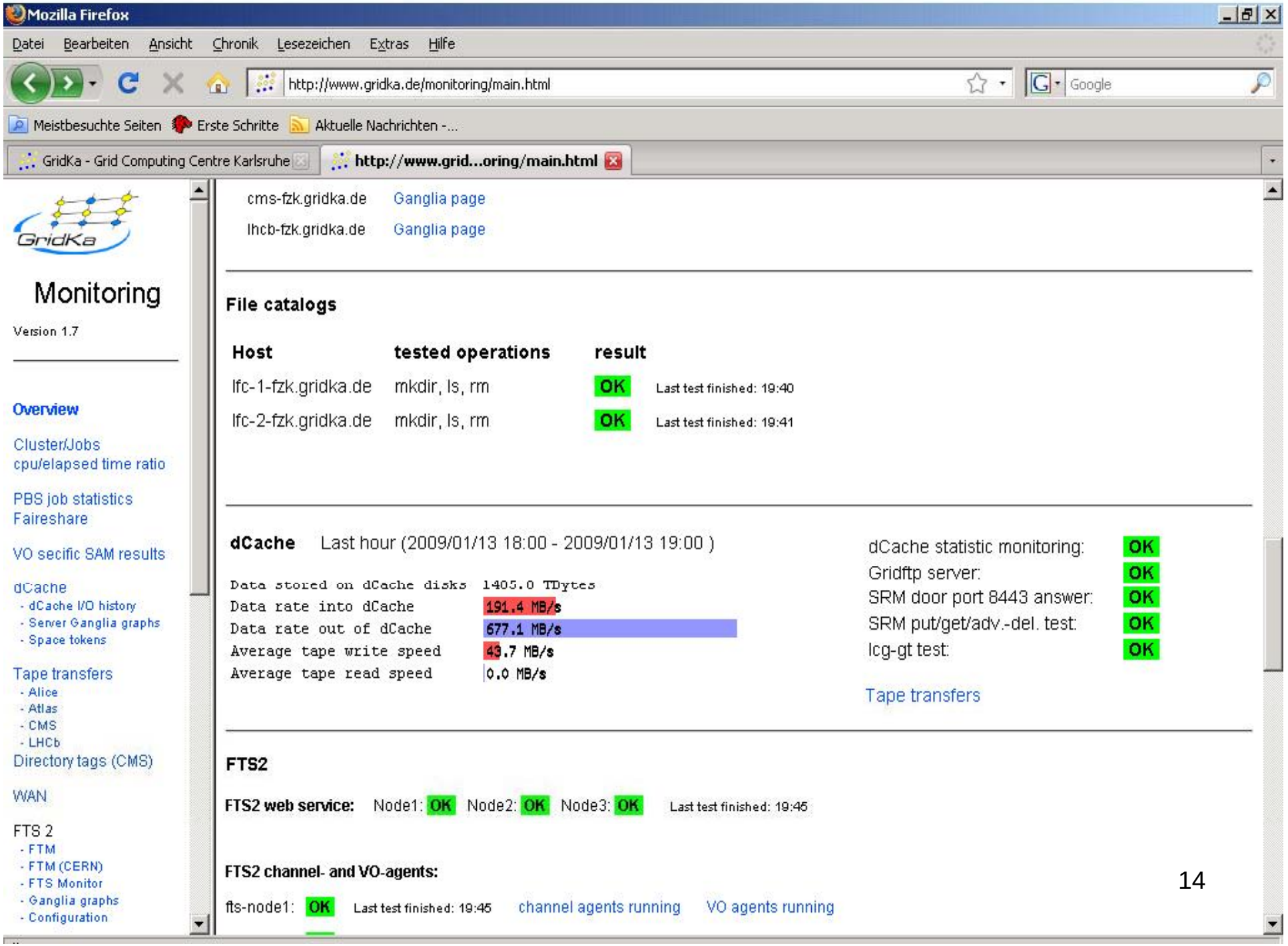
BNL

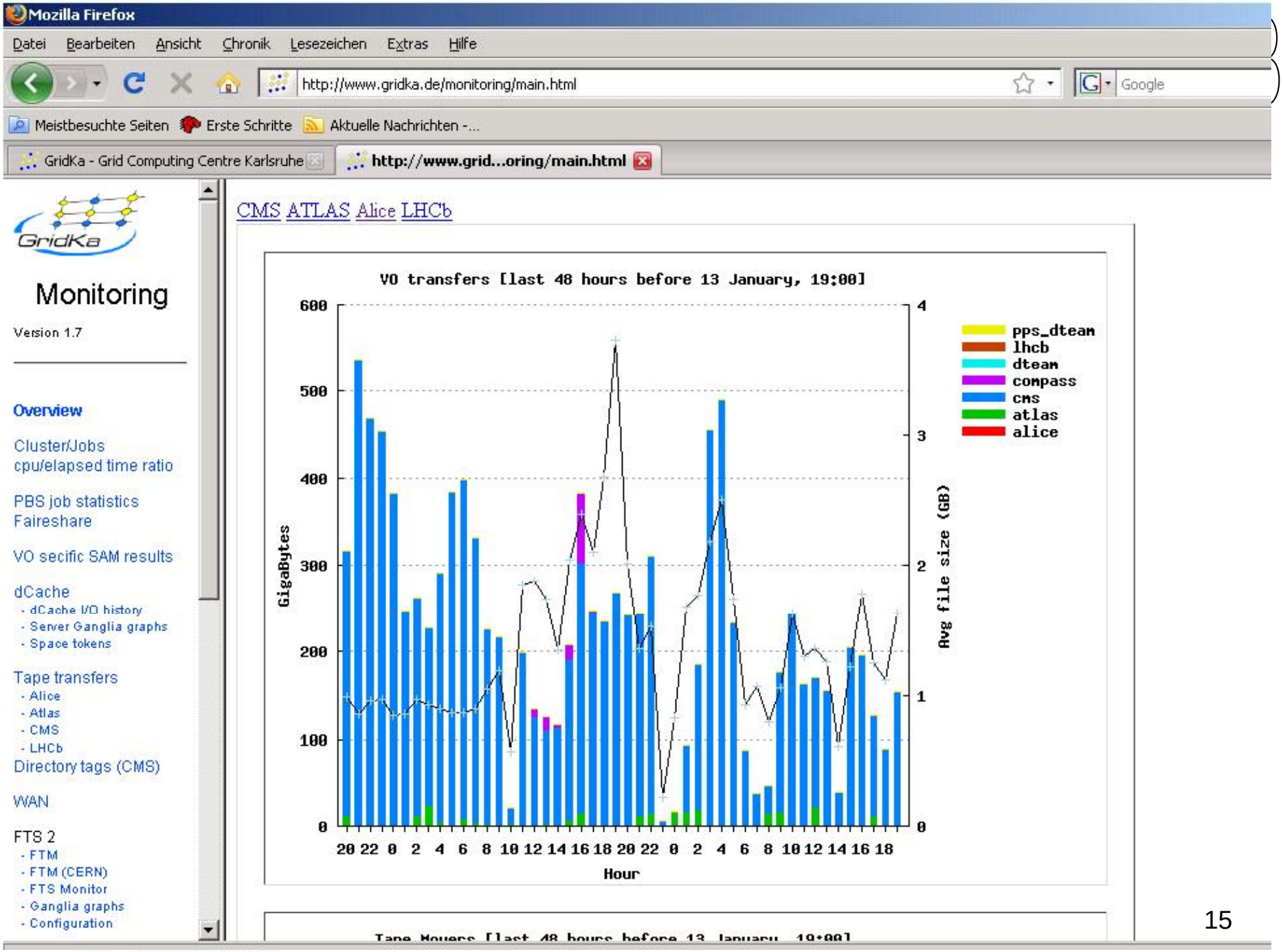
- a) Yes and no. While the system is, due to the vast majority of small ATLAS files (10-500MB), stressed in terms of tape mounts the aggregate bandwidth delivered is relatively low. ATLAS has addressed the issue and has implemented file merging for a variety of data categories.
- b) We maintain a set of dynamically updated plots available at
- https://www.racf.bnl.gov/Facility/HPSS/Monitoring/forUsers/atlas_generalstats.html
- There are 4 graphs for each of 9940B and LTO tape drives (Usage","More","Flow","Mounts").

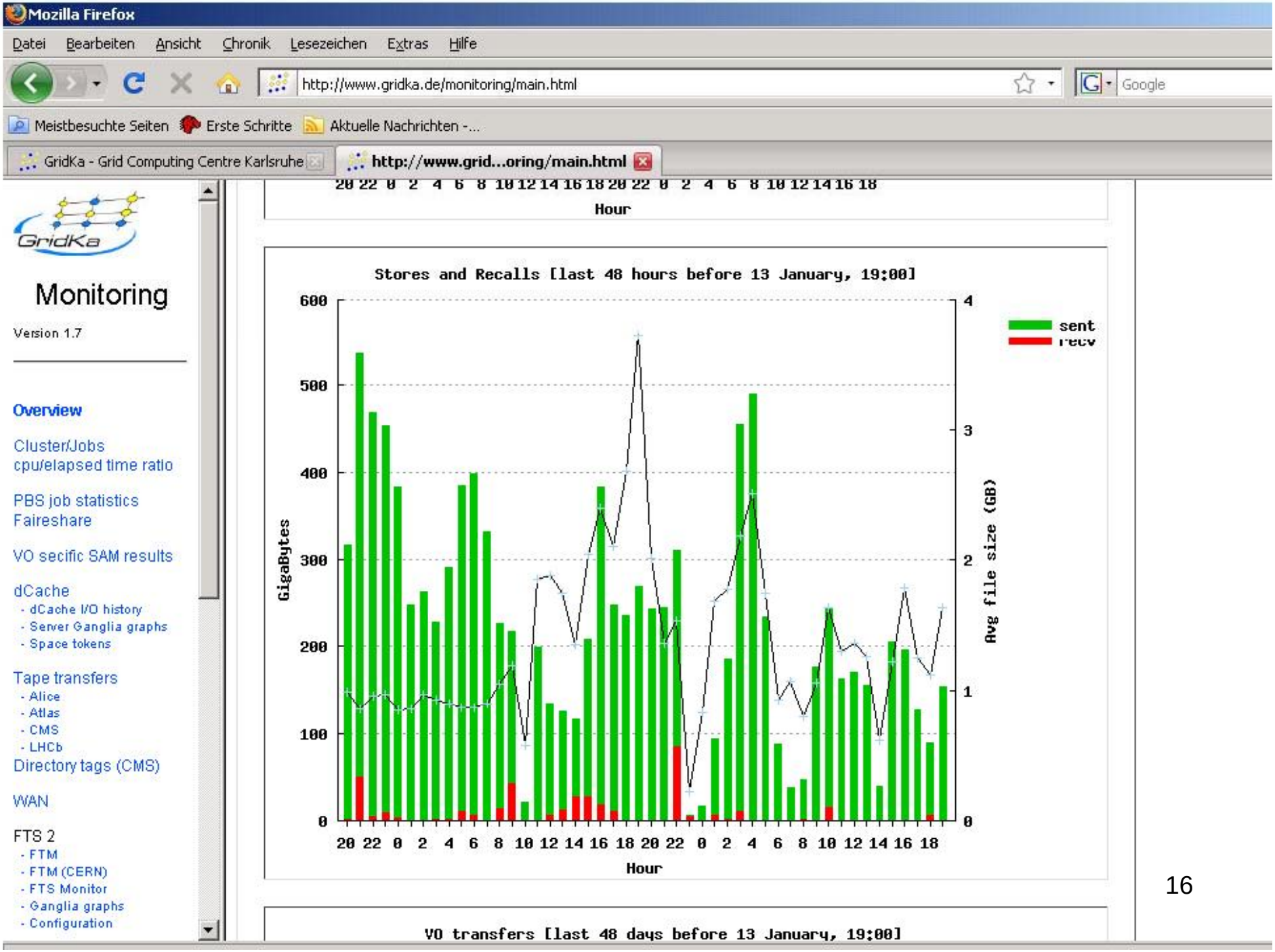
FZK

- Use custom built tool to display data logged by dcache


Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.gridka.de/monitoring/main.html
Google
Meistbesuchte Seiten
Erste Schritte
Aktuelle Nachrichten -...
GridKa - Grid Computing Centre Karlsruhe
http://www.grid...oring/main.html
GridKa
Monitoring
Version 1.7
Overview
Cluster/Jobs cpu/elapsed time ratio
PBS job statistics Faireshare
VO secific SAM results
dCache
- dCache I/O history
- Server Ganglia graphs
- Space tokens
Tape transfers
- Alice
- Atlas
- CMS
- LHCb
Directory tags (CMS)
WAN
FTS 2
- FTM
- FTM (CERN)
- FTS Monitor
- Ganglia graphs
- Configuration
dCache history
Last day
Experiment/VO: all
MSS all I/O rate (day)
MB/s
600
400
200
0
00:00
06:00
12:00
18:00
In
Out
MSS all data on disk (day)
TB
1.0 k
0.0
disk-only pools
MSS all tape I/O rate (day)
write
read
MSS all files on disk (day)
files
8 M
6 M
4 M
2 M
MSS LHC experiments input rate (day)
MSS LHC experiments output rate (day)
Fertig
Start
Mozilla Firefox
19:57


Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.gridka.de/monitoring/main.html
Google
Meistbesuchte Seiten Erste Schritte Aktuelle Nachrichten -...
GridKa - Grid Computing Centre Karlsruhe
http://www.grid...oring/main.html
GridKa
Monitoring
Version 1.7
Overview
Cluster/Jobs cpu/elapsed time ratio
PBS job statistics Faireshare
VO secific SAM results
dCache
- dCache I/O history
- Server Ganglia graphs
- Space tokens
Tape transfers
- Alice
- Atlas
- CMS
- LHCb
Directory tags (CMS)
WAN
FTS 2
- FTM
- FTM (CERN)
- FTS Monitor
- Ganglia graphs
- Configuration
cms-fzk.gridka.de Ganglia page
lhcb-fzk.gridka.de Ganglia page
File catalogs
Host tested operations result
lfc-1-fzk.gridka.de mkdir, ls, rm OK Last test finished: 19:40
lfc-2-fzk.gridka.de mkdir, ls, rm OK Last test finished: 19:41
dCache Last hour (2009/01/13 18:00 - 2009/01/13 19:00)
Data stored on dCache disks 1405.0 TBytes
Data rate into dCache 191.4 MB/s
Data rate out of dCache 677.1 MB/s
Average tape write speed 43.7 MB/s
Average tape read speed 0.0 MB/s
dCache statistic monitoring: OK
Gridftp server: OK
SRM door port 8443 answer: OK
SRM put/get/adv.-del. test: OK
lcg-gt test: OK
Tape transfers
FTS2
FTS2 web service: Node1: OK Node2: OK Node3: OK Last test finished: 19:45
FTS2 channel- and VO-agents:
fts-node1: OK Last test finished: 19:45 channel agents running VO agents running


Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.gridka.de/monitoring/main.html
Google
Meistbesuchte Seiten
Erste Schritte
Aktuelle Nachrichten -...
GridKa - Grid Computing Centre Karlsruhe
http://www.grid...oring/main.html
GridKa
Monitoring
Version 1.7
Overview
Cluster/Jobs
cpu/elapsed time ratio
PBS job statistics
Faireshare
VO secific SAM results
dCache
- dCache I/O history
- Server Ganglia graphs
- Space tokens
Tape transfers
- Alice
- Atlas
- CMS
- LHCb
Directory tags (CMS)
WAN
FTS 2
- FTM
- FTM (CERN)
- FTS Monitor
- Ganglia graphs
- Configuration
CMS ATLAS Alice LHCb
VO transfers [last 48 hours before 13 January, 19:00]
GigaBytes
Avg file size (GB)
Hour
pps_dteam
lhcb
dteam
compass
cms
atlas
alice
Tape Movers [last 48 hours before 13 January, 19:00]


Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.gridka.de/monitoring/main.html
Google
Meistbesuchte Seiten
Erste Schritte
Aktuelle Nachrichten -...
GridKa - Grid Computing Centre Karlsruhe
http://www.grid...oring/main.html
GridKa
Monitoring
Version 1.7
Overview
Cluster/Jobs
cpu/elapsed time ratio
PBS job statistics
Faireshare
VO secific SAM results
dCache
- dCache I/O history
- Server Ganglia graphs
- Space tokens
Tape transfers
- Alice
- Atlas
- CMS
- LHCb
Directory tags (CMS)
WAN
FTS 2
- FTM
- FTM (CERN)
- FTS Monitor
- Ganglia graphs
- Configuration
Hour
Stores and Recalls [last 48 hours before 13 January, 19:00]
GigaBytes
Avg file size (GB)
sent
recv
Hour
VO transfers [last 48 days before 13 January, 19:00]

IN2P3

A) CMS and ATLAS have done some tests of reprocessing data stored on tape in our site. The results of those tests are that the throughput we observed are under the expectations of the experiments. For instance, for the latest Atlas exercice in october we did not meet the target.

IN2P3

b) Below are the data we currently collect:

- Throughput between dCache and HPSS (as perceived by dCache), for reading & writing data, in MB/sec
  - both agregated for the 4 LHC experiments, and per experiment
- Tape drive usage, per drive type. For LHC experiments, only the usage of T10.000 tape drives is relevant. We measure:
  - the number of drives in use
  - the number of copy requests waiting for a cartridge to be available
  - the number of copy waiting for a cartridge to be mounted
  - the number of copy waiting for a tape drive to be free
- Maximum mount time per drive type (in minutes)
- Network bandwidth used by HPSS tape servers and disk servers: as the tape and disk servers are not dedicated per experiment, we currently don't have a means to get this information per experiment

IN2P3

- What currently don't have (at least, not systematically) is the distribution of the sizes of files on tape, per experiment, nor the number of files read or written per tape mount.
- We don't have a means to easily correlate the activity of the experiments (as shown in their dashboards) to the activity on HPSS.
- Looking at all the plots of the data we collect, our impression is that the problem we have is that the data on tape is not organized correctly so that the retrieval of them is optimized. We are currently exploring ways to improve the interaction of dCache and HPSS in order to organize the writing of the data on tape and optimizing the reading of those data (for instance, by making sure that dCache requests HPSS the copy of several files on the same tape, so that the tape is mounted only once).

NDGF

a) Yes, ATLAS has stressed it, but we haven't had a stress test since the last time we increased capacity.

b) Yes, on multiple levels. There may be some good plots around, but since NDGF has many tape systems, the overview plots are kind of bland and the detailed ones are per-tape system.

What Next?

- Present conclusions to LHCC Mini-Review in February.
- Main conclusions that
  - All(?) sites have achieved required throughput for RAW and distribution of ESD and simulated data for each experiment.
  - No confidence that they can do this for all experiments simultaneously
  - Mixed results on recalling RAW for reprocessing.
  - Effect of chaotic analysis unknown
- Discuss!