ATLAS Bulk Pre-stageing Tests

Graeme Stewart
University of Glasgow

Overview

- ATLAS computing model always had most RAW data on tape
- T1s need to reprocess this data about 3 times a year
  - This will happen more often in early running
  - But in early running there is less data
- We want to be able to reprocess within a month
- So this is about ten times faster than we take data

Some Numbers

- A RAW event is 1.6 MB
- We take data at 200 Hz for 50,000 sec/day
- So we need to write 16 TB/day to tape at CERN
- Each T1 also needs to write its share to tape
- But it may take 86,000 sec/day to do so.
- So a 10% T1 needs to write at:
- 0.1 * 16 * 1,000,000 / 86,400 = 18.6 MB/sec
- And needs to be able to read at 186 MB/sec

Reprocessing Target Rates

Tier-1	ATLAS Share %	Rate to Tape MB/s	Reprocessing Rate MB/s
BNL	25	47	465
IN2P3	15	28	279
SARA	15	28	279
RAL	10	19	186
FZK	10	19	186
PIC	5	9	93
TRIUMF	5	9	93
CNAF	5	9	93
ASGC	5	9	93
NDGF	5	9	93

DDM Pre-Stage Service

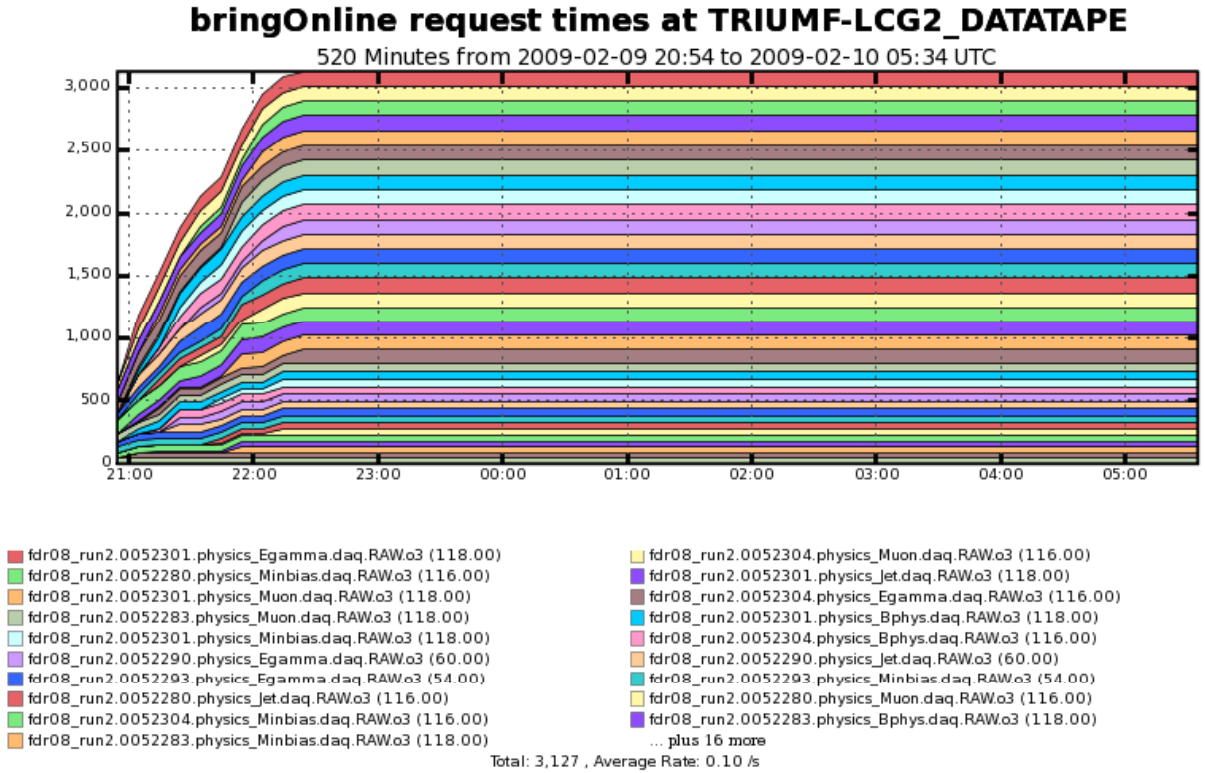
- Pre-staging is a service provided by ATLAS Distributed Data Management (DDM)
- When a prestage request for a dataset arrives DDM
  - Checks there is room for the dataset on the stage buffer (CASTOR – soft pins!)
  - Issues srmBringOnline for the files of that dataset
  - Monitors progress using srmLs to see if files are ONLINE or NEARLINE

Bulk Pre-stage Tests

- At most Tier-1s we do a test where we ask for 3127 files (in 36 datasets) which total 9371TB
- And we see what happens, which should be this…

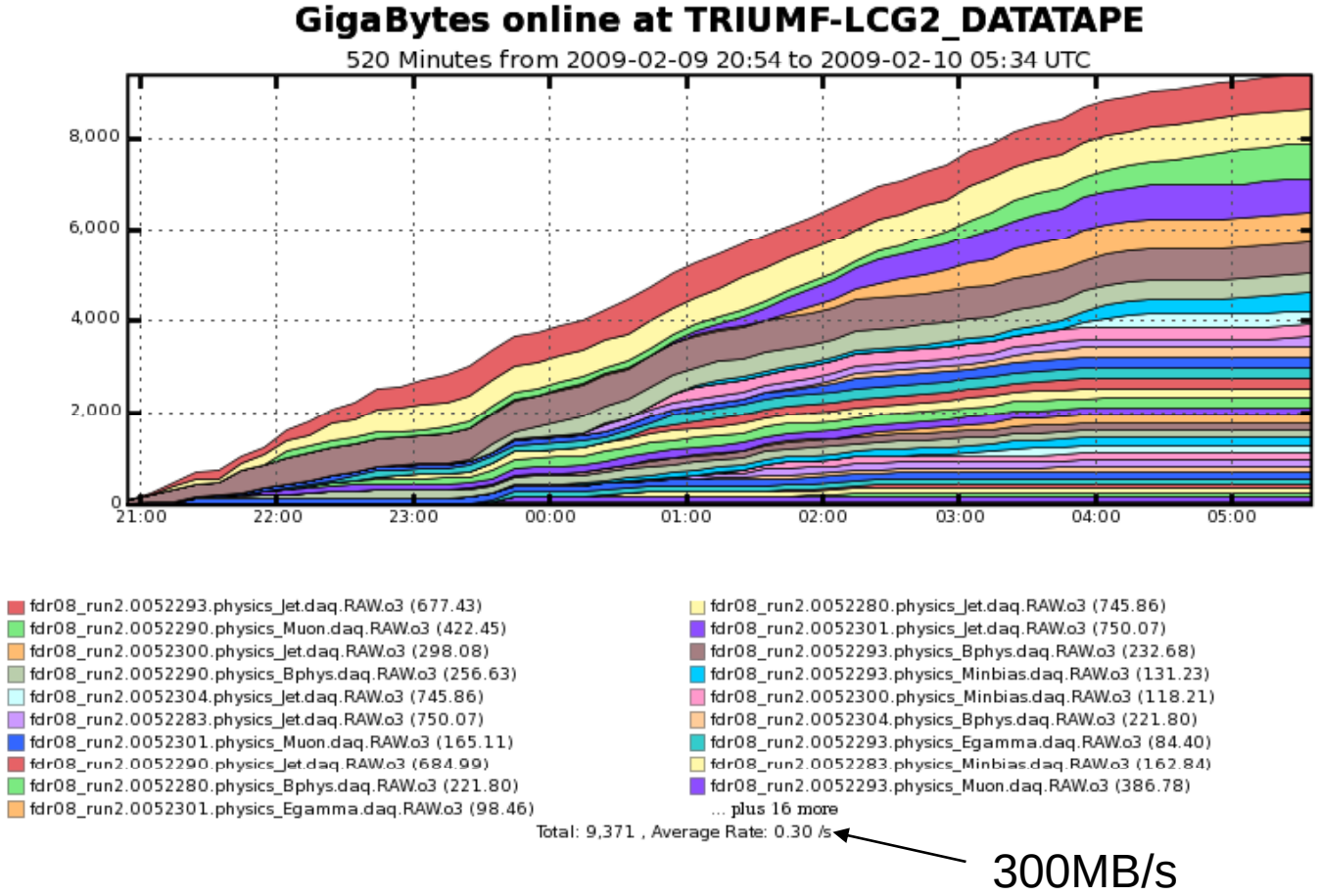
DDM Stage Request Times

- This measures the rate at which DDM is able to issue the srmBringOnline requests:

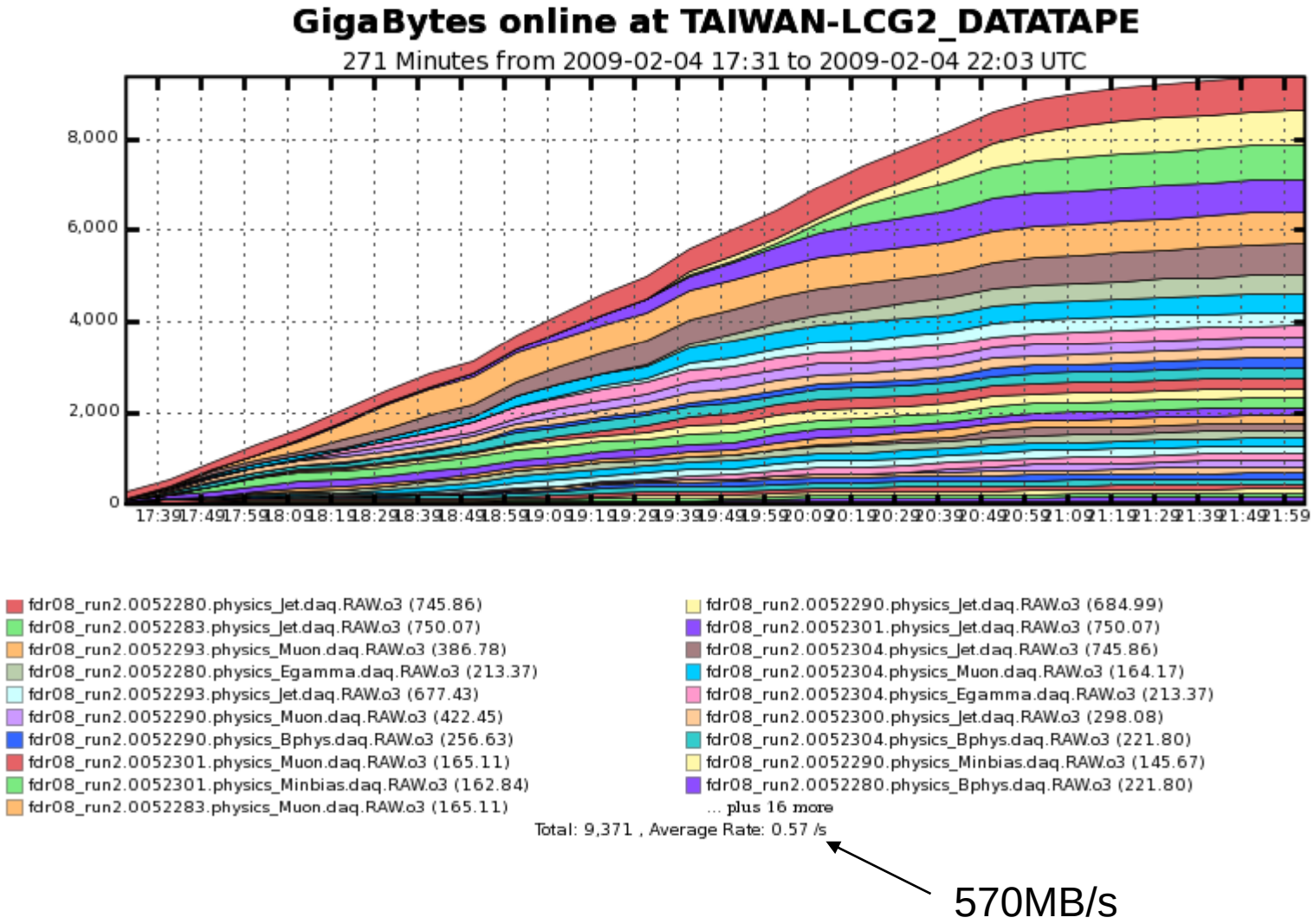


Site Stage Rates

- Site then stages files…

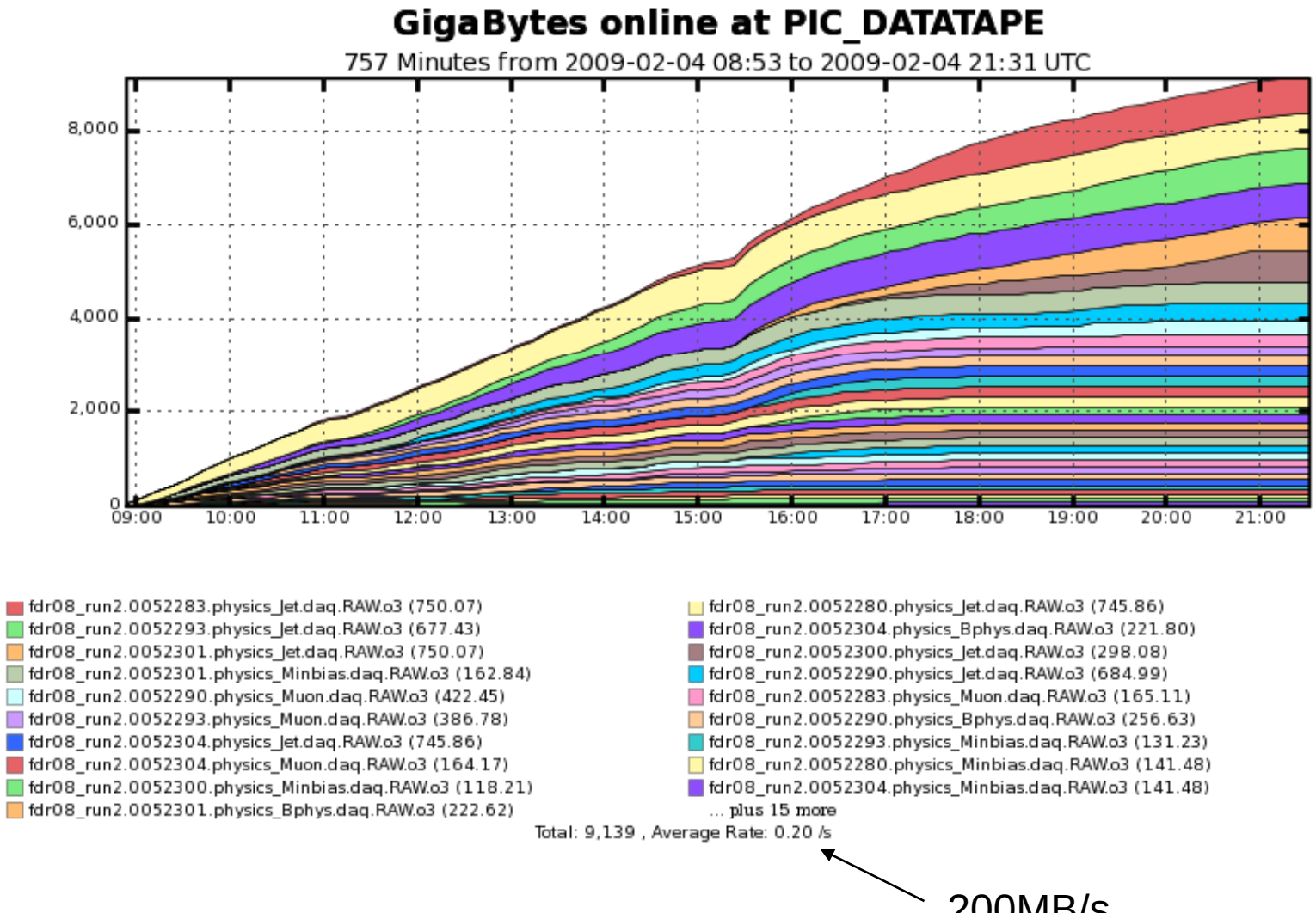


ASGC


GigaBytes online at TAIWAN-LCG2_DATATAPE
271 Minutes from 2009-02-04 17:31 to 2009-02-04 22:03 UTC
8,000
6,000
4,000
2,000
0
17:39 17:49 17:59 18:09 18:19 18:29 18:39 18:49 18:59 19:09 19:19 19:29 19:39 19:49 19:59 20:09 20:19 20:29 20:39 20:49 20:59 21:09 21:19 21:29 21:39 21:49 21:59
fdr08_run2.0052280.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052283.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052293.physics_Muon.daq.RAW.o3 (386.78)
fdr08_run2.0052280.physics_Egamma.daq.RAW.o3 (213.37)
fdr08_run2.0052293.physics_Jet.daq.RAW.o3 (677.43)
fdr08_run2.0052290.physics_Muon.daq.RAW.o3 (422.45)
fdr08_run2.0052290.physics_Bphys.daq.RAW.o3 (256.63)
fdr08_run2.0052301.physics_Muon.daq.RAW.o3 (165.11)
fdr08_run2.0052301.physics_Minbias.daq.RAW.o3 (162.84)
fdr08_run2.0052283.physics_Muon.daq.RAW.o3 (165.11)
fdr08_run2.0052290.physics_Jet.daq.RAW.o3 (684.99)
fdr08_run2.0052301.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052304.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052304.physics_Muon.daq.RAW.o3 (164.17)
fdr08_run2.0052304.physics_Egamma.daq.RAW.o3 (213.37)
fdr08_run2.0052300.physics_Jet.daq.RAW.o3 (298.08)
fdr08_run2.0052304.physics_Bphys.daq.RAW.o3 (221.80)
fdr08_run2.0052290.physics_Minbias.daq.RAW.o3 (145.67)
fdr08_run2.0052280.physics_Bphys.daq.RAW.o3 (221.80)
... plus 16 more
Total: 9,371 , Average Rate: 0.57 /s

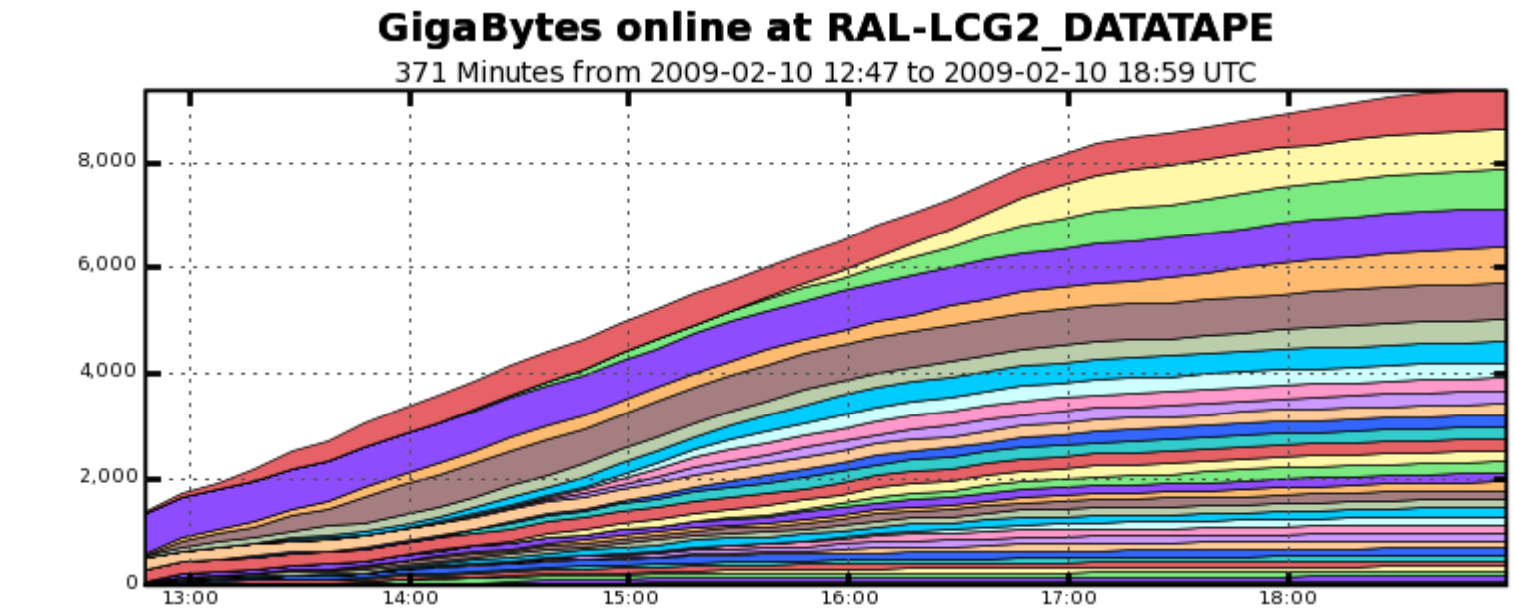
570MB/s

PIC


GigaBytes online at PIC_DATATAPE
757 Minutes from 2009-02-04 08:53 to 2009-02-04 21:31 UTC
8,000
6,000
4,000
2,000
0
09:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
fdr08_run2.0052283.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052293.physics_Jet.daq.RAW.o3 (677.43)
fdr08_run2.0052301.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052301.physics_Minbias.daq.RAW.o3 (162.84)
fdr08_run2.0052290.physics_Muon.daq.RAW.o3 (422.45)
fdr08_run2.0052293.physics_Muon.daq.RAW.o3 (386.78)
fdr08_run2.0052304.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052304.physics_Muon.daq.RAW.o3 (164.17)
fdr08_run2.0052300.physics_Minbias.daq.RAW.o3 (118.21)
fdr08_run2.0052301.physics_Bphys.daq.RAW.o3 (222.62)
fdr08_run2.0052280.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052304.physics_Bphys.daq.RAW.o3 (221.80)
fdr08_run2.0052300.physics_Jet.daq.RAW.o3 (298.08)
fdr08_run2.0052290.physics_Jet.daq.RAW.o3 (684.99)
fdr08_run2.0052283.physics_Muon.daq.RAW.o3 (165.11)
fdr08_run2.0052290.physics_Bphys.daq.RAW.o3 (256.63)
fdr08_run2.0052293.physics_Minbias.daq.RAW.o3 (131.23)
fdr08_run2.0052280.physics_Minbias.daq.RAW.o3 (141.48)
fdr08_run2.0052304.physics_Minbias.daq.RAW.o3 (141.48)
... plus 15 more
Total: 9,139 , Average Rate: 0.20 /s

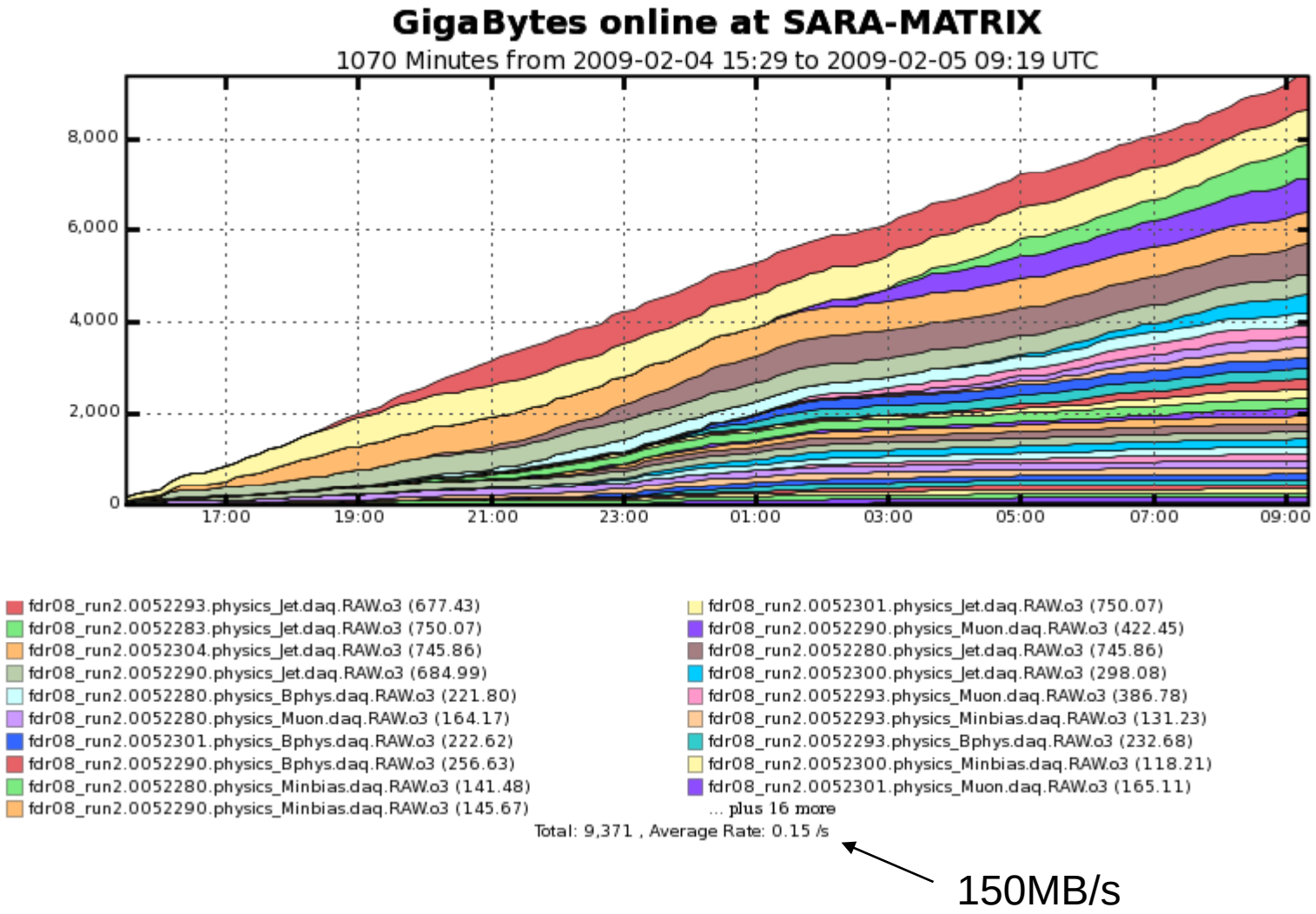
200MB/s

RAL



420MB/s but small head start

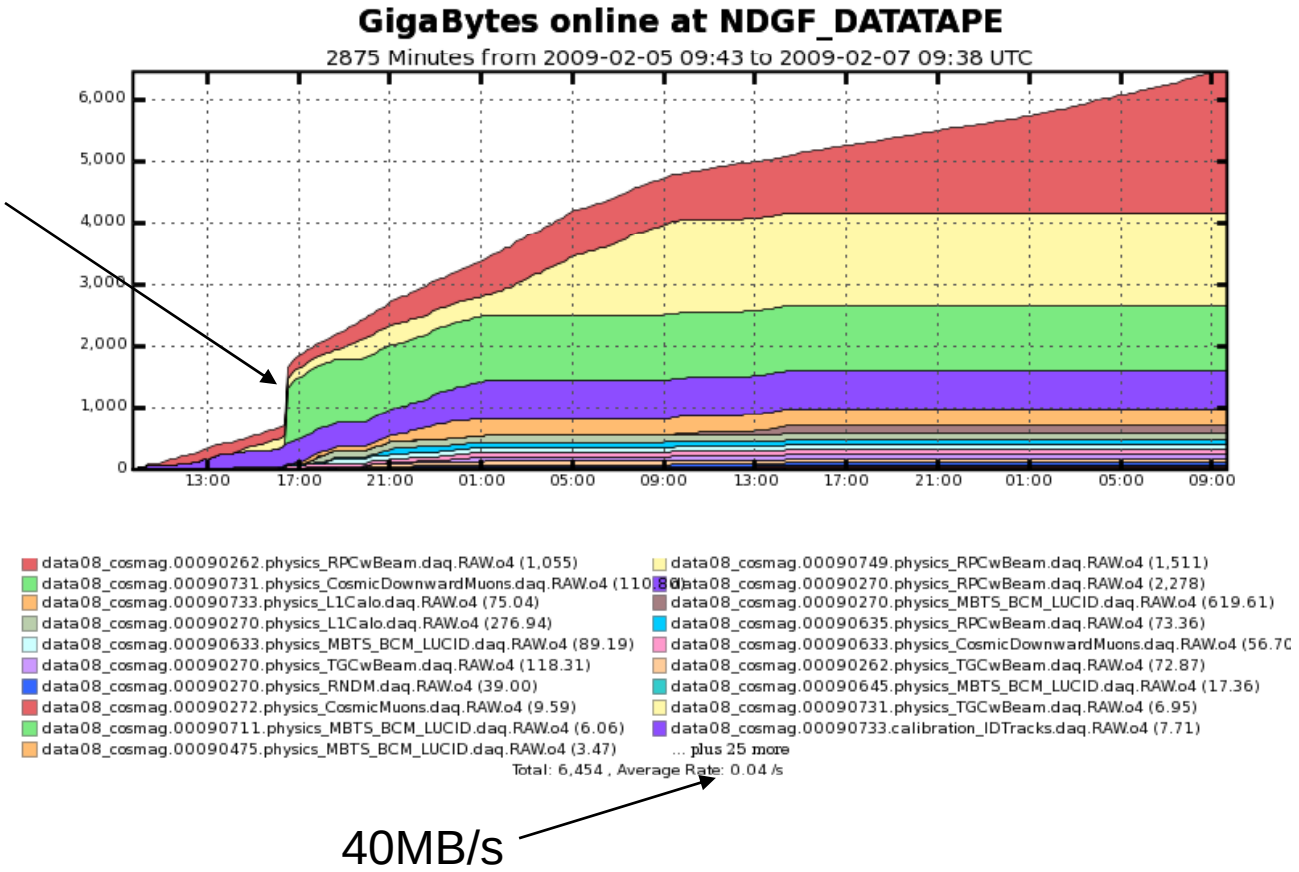
SARA


GigaBytes online at SARA-MATRIX
1070 Minutes from 2009-02-04 15:29 to 2009-02-05 09:19 UTC
8,000
6,000
4,000
2,000
0
17:00
19:00
21:00
23:00
01:00
03:00
05:00
07:00
09:00
fdr08_run2.0052293.physics_Jet.daq.RAW.o3 (677.43)
fdr08_run2.0052283.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052304.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052290.physics_Jet.daq.RAW.o3 (684.99)
fdr08_run2.0052280.physics_Bphys.daq.RAW.o3 (221.80)
fdr08_run2.0052280.physics_Muon.daq.RAW.o3 (164.17)
fdr08_run2.0052301.physics_Bphys.daq.RAW.o3 (222.62)
fdr08_run2.0052290.physics_Bphys.daq.RAW.o3 (256.63)
fdr08_run2.0052280.physics_Minbias.daq.RAW.o3 (141.48)
fdr08_run2.0052290.physics_Minbias.daq.RAW.o3 (145.67)
fdr08_run2.0052301.physics_Jet.daq.RAW.o3 (750.07)
fdr08_run2.0052290.physics_Muon.daq.RAW.o3 (422.45)
fdr08_run2.0052280.physics_Jet.daq.RAW.o3 (745.86)
fdr08_run2.0052300.physics_Jet.daq.RAW.o3 (298.08)
fdr08_run2.0052293.physics_Muon.daq.RAW.o3 (386.78)
fdr08_run2.0052293.physics_Minbias.daq.RAW.o3 (131.23)
fdr08_run2.0052293.physics_Bphys.daq.RAW.o3 (232.68)
fdr08_run2.0052300.physics_Minbias.daq.RAW.o3 (118.21)
fdr08_run2.0052301.physics_Muon.daq.RAW.o3 (165.11)
... plus 16 more
Total: 9,371 , Average Rate: 0.15 /s

150MB/s

NDGF

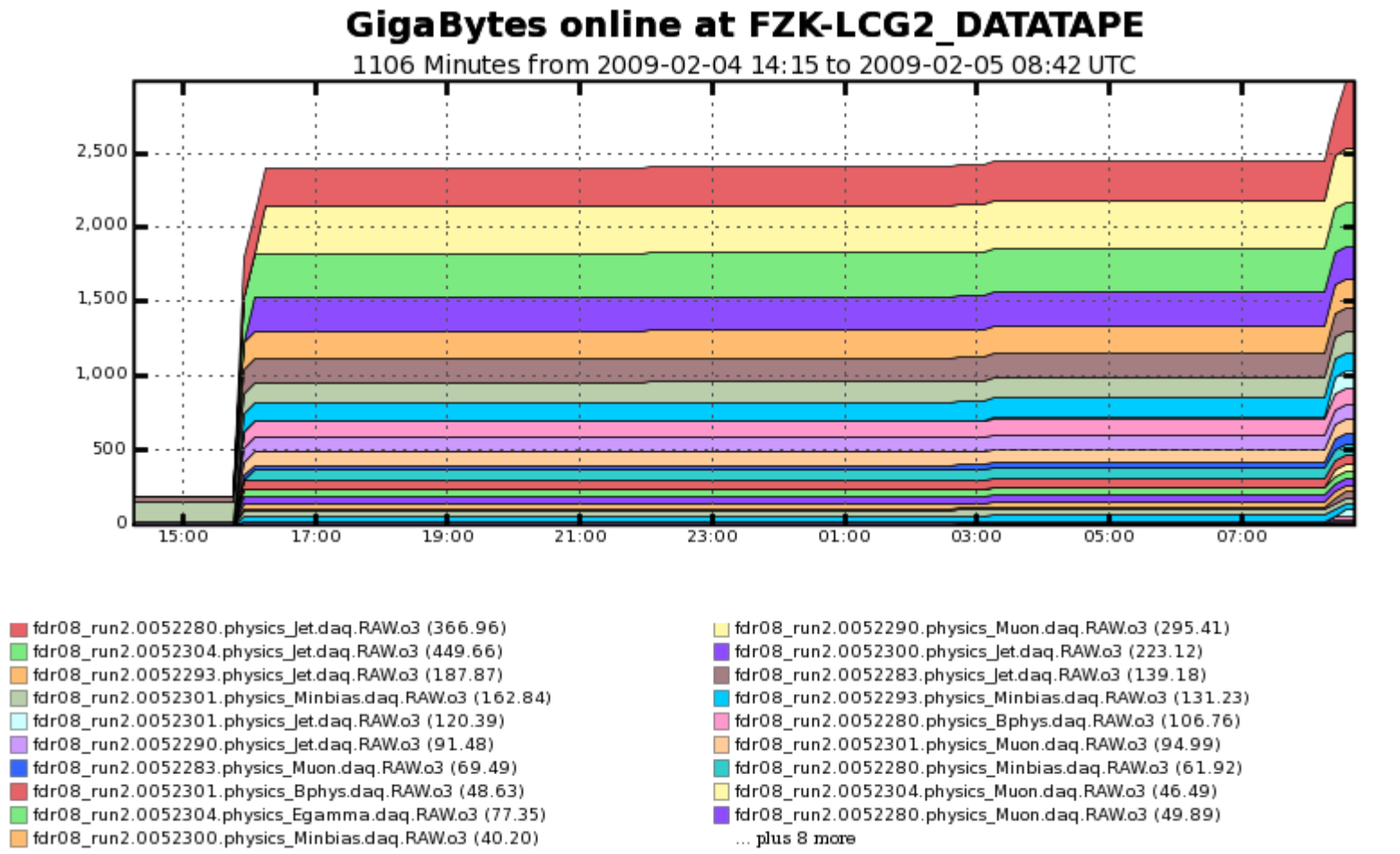
Problems polling SRM
(connection timeouts)



40MB/s

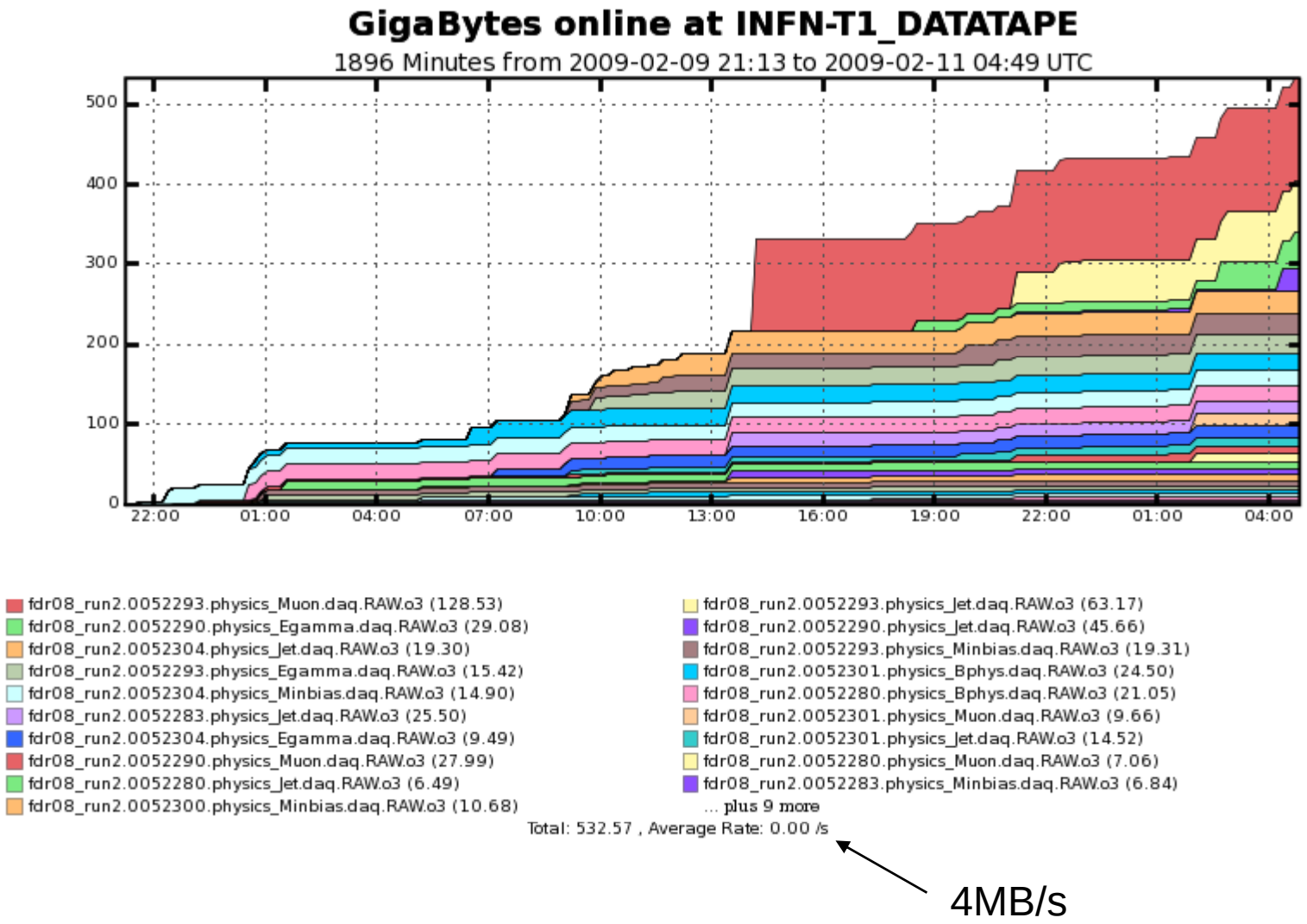
- NDGF used different datasets to ensure we used their new tape library

FZK



40MB/s, only 3k files

CNAF


GigaBytes online at INFN-T1_DATATAPE
1896 Minutes from 2009-02-09 21:13 to 2009-02-11 04:49 UTC
500
400
300
200
100
0
22:00
01:00
04:00
07:00
10:00
13:00
16:00
19:00
22:00
01:00
04:00
fdr08_run2.0052293.physics_Muon.daq.RAW.o3 (128.53)
fdr08_run2.0052290.physics_Egamma.daq.RAW.o3 (29.08)
fdr08_run2.0052304.physics_Jet.daq.RAW.o3 (19.30)
fdr08_run2.0052293.physics_Egamma.daq.RAW.o3 (15.42)
fdr08_run2.0052304.physics_Minbias.daq.RAW.o3 (14.90)
fdr08_run2.0052283.physics_Jet.daq.RAW.o3 (25.50)
fdr08_run2.0052304.physics_Egamma.daq.RAW.o3 (9.49)
fdr08_run2.0052290.physics_Muon.daq.RAW.o3 (27.99)
fdr08_run2.0052280.physics_Jet.daq.RAW.o3 (6.49)
fdr08_run2.0052300.physics_Minbias.daq.RAW.o3 (10.68)
fdr08_run2.0052293.physics_Jet.daq.RAW.o3 (63.17)
fdr08_run2.0052290.physics_Jet.daq.RAW.o3 (45.66)
fdr08_run2.0052293.physics_Minbias.daq.RAW.o3 (19.31)
fdr08_run2.0052301.physics_Bphys.daq.RAW.o3 (24.50)
fdr08_run2.0052280.physics_Bphys.daq.RAW.o3 (21.05)
fdr08_run2.0052301.physics_Muon.daq.RAW.o3 (9.66)
fdr08_run2.0052301.physics_Jet.daq.RAW.o3 (14.52)
fdr08_run2.0052280.physics_Muon.daq.RAW.o3 (7.06)
fdr08_run2.0052283.physics_Minbias.daq.RAW.o3 (6.84)
... plus 9 more
Total: 532.57 , Average Rate: 0.00 /s

4MB/s

BNL and LYON

- BNL will test at the end of next week after a scheduled intervention on their dCache
- LYON know they have a serious problem in the interface between HPSS and dCache
  - We understand they are working on a new interface, but at the moment we know we would get a very poor rate
  - They have the ATLAS test infrastructure and are able to run this themselves

The Story So Far

Tier-1	Target Rate MB/s	Measured Rate MB/s	Notes
BNL	465	-	
IN2P3	279	-	We are worried
SARA	279	150	Bottleneck understood – waiting for more DMF hardware
RAL	186	360	1 file MIA
FZK	186	40	Suffering from dCache pnfs overload
PIC	93	200	Some 10s of files MIA
TRIUMF	93	300	
CNAF	93	4	Much worse than last year – experts investigating
ASGC	93	570	
NDGF	93	40	Have applied a firmware fix to one tape library – should retest

Other Considerations

- At sites where we made our target rate we were the only experiment using the tape system and this is a shared infrastructure
- Staging from tape is only the first step in reprocessing – this data has to be copied to the farm at the same time, run over, results stored
- Missing files are still a problem – no clear error messages
- New version of site services will use statusOfBringOnline, which is kinder to the SRM
- We would like to try to re-reprocess the combined cosmics runs from tape next month
- Twiki has latest news:
  https://twiki.cern.ch/twiki/bin/view/Atlas/PreStageTests