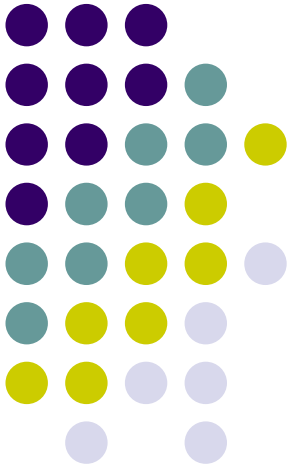
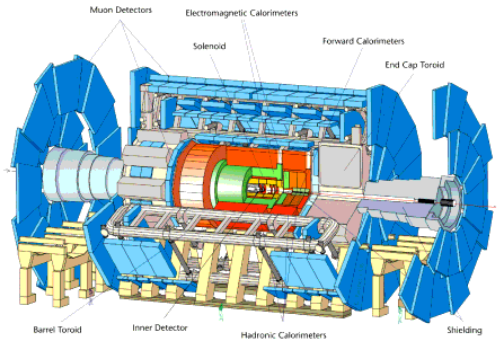
FTS and ATLAS DDM

David Cameron
CERN
FTS Administrators Workshop
Amsterdam, 18/10/06



The ATLAS Experiment Data Flow



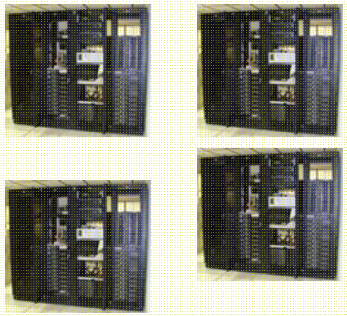
Detector

RAW data

CERN Computer Centre + Tier 0

Reconstructed + RAW data

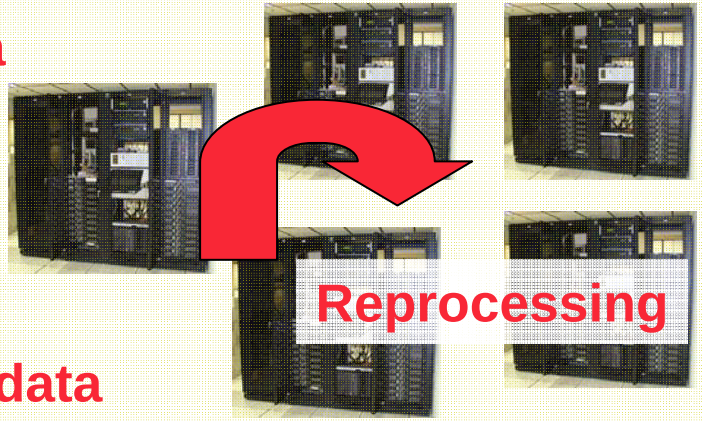
GRID



Tier 2 centres

Small data products

Simulated data



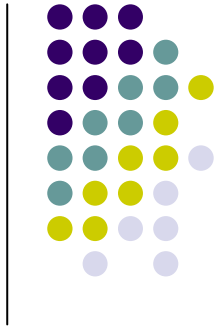
Reprocessing

Tier 1 centres

Don Quijote 2

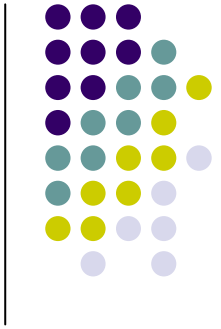
- Our software is called Don Quijote 2 (DQ2)
- Data is organised into **datasets** which are the unit of data movement
- To enable data movement we have a set of distributed ‘site services’ which use a subscription mechanism to pull data to a site
  - As content is added to a dataset, the site services copy it to subscribed sites
- The goal is to manage data flow as described in the computing model and provide a single entry point to all distributed ATLAS data
  - Distribution of raw and reconstructed data from CERN to the Tier-1s
  - Distribution of AODs (Analysis Object Data) to Tier-2 centres for analysis
  - Storage of simulated data (produced by Tier-2s) at Tier-1 centres for further distribution and/or processing

Site Services



- Site services are deployed on VOBOXes
  - On LCG, there is one VOBOX per Tier 1 site and the site services here serve the associated Tier 2 sites
  - On OSG, there is one VOBOX per Tier 1 site and one per Tier 2 site
- The site services work as a state machine
- A set of agents pick up requests and process from one state to the next state
- A local database on the VOBOX stores the files' states
  - With the advantage that this database can be lost and recreated from central and local catalog information

Tiers Of ATLAS

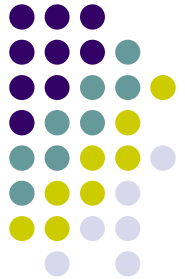


- Tiers of ATLAS is the ATLAS data management information system
- Defines T1-T2 association and FTS topology
- Idea of disk/tape sites
- Used to find FTS server given source and destination

TiersOfATLASCache.py

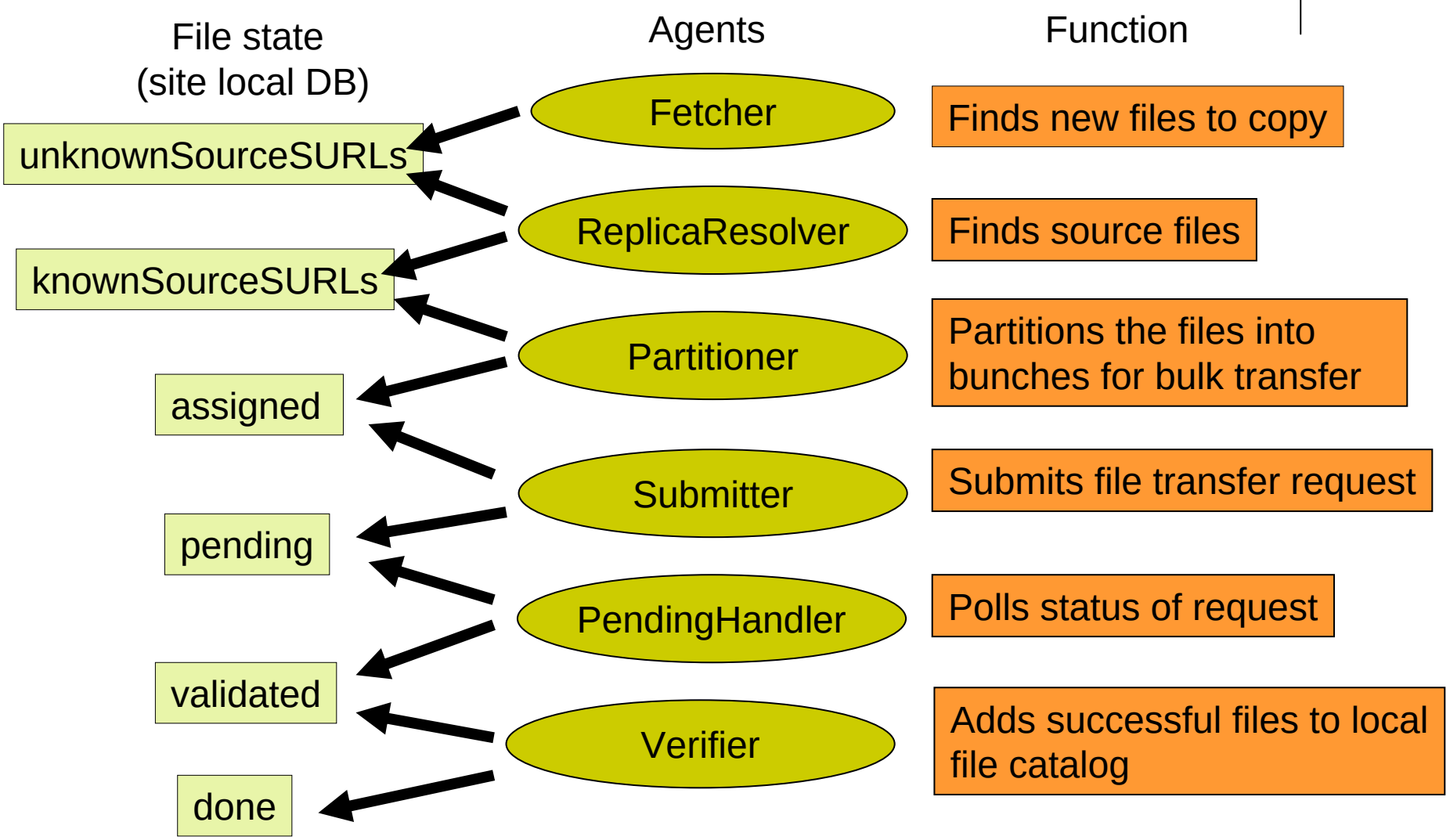
```
'LYONTAPE':
 {
    'state': [SE_READABLE, SE_WRITEABLE, CE_USEABLE],
    'istape': True,
    'email': 'ddm-support@in2p3.fr',
    'domain': '.*in2p3.fr.*',
    'toolAssigner': 'lcg',
    # LCG tool toolAssigner attributes
    'srm': 'srm://ccsrm.in2p3.fr/pnfs/in2p3.fr/data/atlas/dq2/',
    'srmsc4': 'srm://ccsrm.in2p3.fr/pnfs/in2p3.fr/data/atlas/tape/sc4',
    # LCG executor attributes
    'ce': [ '' ],
    },
....

 LYONFTS: {    'srm://ccsrm.in2p3.fr': [ '*' ],
            # LYON -> Tier2s
            'srm://clrlcgse03.in2p3.fr':    [ 'srm://ccsrm.in2p3.fr' ],
            'srm://grid05.lal.in2p3.fr':    [ 'srm://ccsrm.in2p3.fr' ],
            'srm://node12.datagrid.cea.fr': [ 'srm://ccsrm.in2p3.fr' ],
            'srm://lpnse1.in2p3.fr':        [ 'srm://ccsrm.in2p3.fr' ],
            'srm://lapp-se01.in2p3.fr':     [ 'srm://ccsrm.in2p3.fr' ],
            'srm://sedpm.mrs.grid.cnrs.fr': [ 'srm://ccsrm.in2p3.fr' ],
...
```

FTS Channels

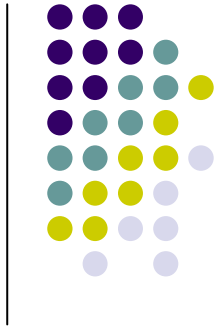


- Channels
  - Dedicated channels (these fit all cases of comp model)
    - CERN: T0-T1
    - T1s: T1-T1 (at dest), T1 - assoc. T2s
  - **Non-info system dependent** STAR channels (rare cases outside comp model, private data movement, non-LCG sites)
    - CERN: T2 - T0
    - T1s: T2s outside local cloud - T1

Site Services Workflow



Using FTS



- We call command line interfaces from our python code
  - glite-transfer-submit, glite-transfer-status
- Parse the output to determine status, error/success of files and error reasons
- Recently added pre and post transfer (after failure) ‘deleting’
  - To avoid ‘file exists’ errors
  - glite-srm-delete
  - CERN specific tools (stager_rm, nsrm etc)

Monitoring Transfers

This page shows information on files currently being processed or processed recently. Files in this
to appear here. Also, after one week successful transfers are deleted from the monitoring databa
Use the dq2 command line client for full information on the content of a dataset.

286 files found

LFN	State
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00001.pool.root.1	HOLD_NO_REPLICAS
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00002.pool.root.1	FILE_DONE
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00003.pool.root.1	FILE_DONE
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00004.pool.root.1	ASSIGNED
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00005.pool.root.1	FILE_DONE
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00006.pool.root.1	ASSIGNED
(info) csc11.005013.J4_pythia_jetjet.recon.AOD.v11004205._00007.pool.root.1	HOLD_NO_REPLICAS

File Attributes

Attribute	
state	FILE_DONE
guid	8A662FD2-3E51-DB11-A596-001125754D56
lfn	calibg_csc11.007061.singlepart_e_E100.digit.RDO.v12000301_tid003191._00018.pool.root.1
dsn	calibg_csc11.007061.singlepart_e_E100.digit.RDO.v12000301_tid003191_dis422698
duid	377e17b3-a2fd-41d6-8009-83a6aa94ea74
vuid	355f2a78-5884-41b7-81e1-33a456883b3d
uid	377e17b3-a2fd-41d6-8009-83a6aa94ea74
version	1
site	UTA_SWT2
creation_datetime	1161005517.14 (Mon Oct 16 13:31:57 2006 UTC)
modified_datetime	1161006113.0 (Mon Oct 16 13:41:53 2006 UTC)
fsize	85581874
md5	7912a7de711b4b017bfda99074c93498
src_surls	['srm://dcsrm.usatlas.bnl.gov/pnfs/usatlas.bnl.gov/others01/2006/39/calibg_csc11.007061.singlepart_e
src_surl	srm://dcsrm.usatlas.bnl.gov/pnfs/usatlas.bnl.gov/others01/2006/39/calibg_csc11.007061.singlepart_e_
dest_surl	gsiftp://gk01.swt2.uta.edu/ifs1/dq2_cache/storageA/calibg_csc11/calibg_csc11.007061.singlepart_e_E
tool_id	fts
transfer_channel	https://fts.usatlas.bnl.gov:8443/glite-data-transfer-fts/services/FileTransfer
transfer_id	fbc4b07d-5d11-11db-a2b5-c5f87d204b43
	State from FTS: Waiting; Retries: 1; Reason: TRANSFER Operation Timed out.

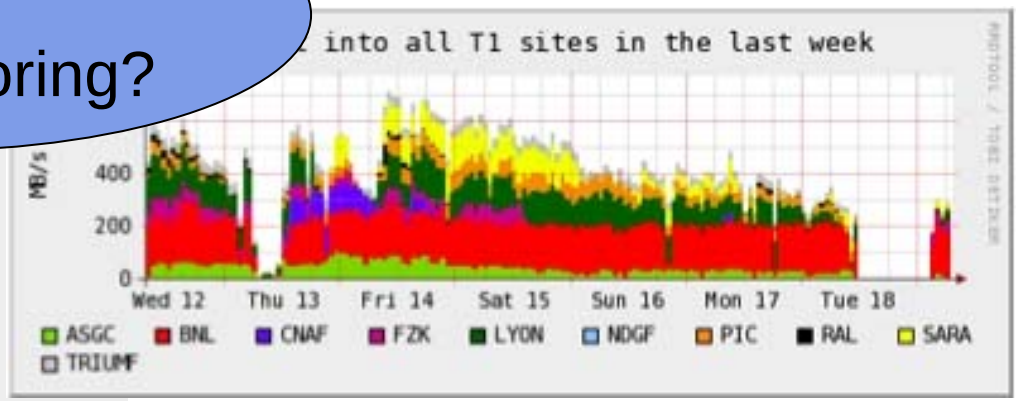
This page shows datasets currently being processed. If you have just added a subscription it may take a few minutes to appear here.
fully processed more than a few days ago will not appear here. Use the dq2 command line client to get full information on datasets s

807 results found

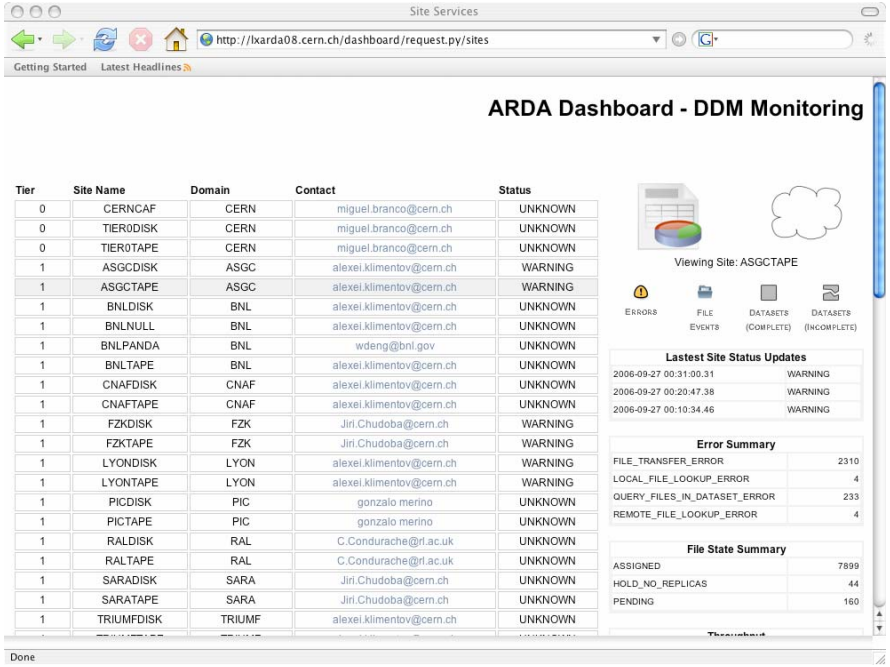
Dataset

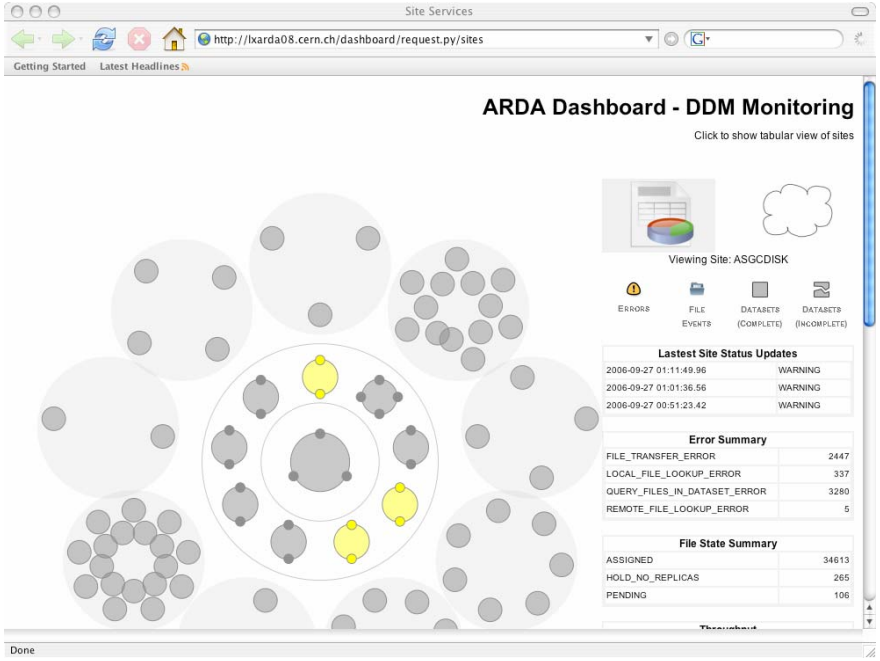
acermc.005177.Zbb4l
csc11.005009.J0_pythia_jetjet.recon
csc11.005010.J1_pythia_jetjet
csc11.005011.J2_pythia_jetjet
csc11.005012.J3_pythia_jetjet
csc11.005013.J4_pythia_jetjet
csc11.005014.J5_pythia_jetjet
csc11.005015.J6_pythia_jetjet.recon
csc11.005015.J6_pythia_jetjet.recon.EM
csc11.005015.J6_pythia_jetjet.recon.log.v1100
csc11.005015.J6_pythia_jetjet.recon.log.v11004210_tid002
csc11.005015.J6_pythia_jetjet.recon.log.v11004210_tid002947_sub347
csc11.005015.J6_pythia_jetjet.recon.log.v11004210_tid002947_sub349
csc11.005015.J6_pythia_jetjet.recotrig.AOD.v11000505
csc11.005015.J6_pythia_jetjet.recotrig.CBNT.v11000505_tid002946_sub526
csc11.005015.J6_pythia_jetjet.recotrig.CBNT.v11000505_tid002946_sub544
csc11.005015.J6_pythia_jetjet.recotrig.ESD.v11000505_tid002946_sub524
csc11.005015.J6_pythia_jetjet.recotrig.ESD.v11000505_tid002946_sub542
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub464
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub476
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub482
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub505
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub509
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub512
csc11.005015.J6_pythia_jetjet.recotrig.log.v11000505_tid002946_sub514

Click here to go to FTS monitoring?

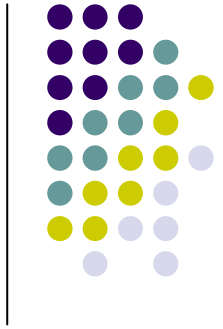


ARDA Dashboard Monitoring

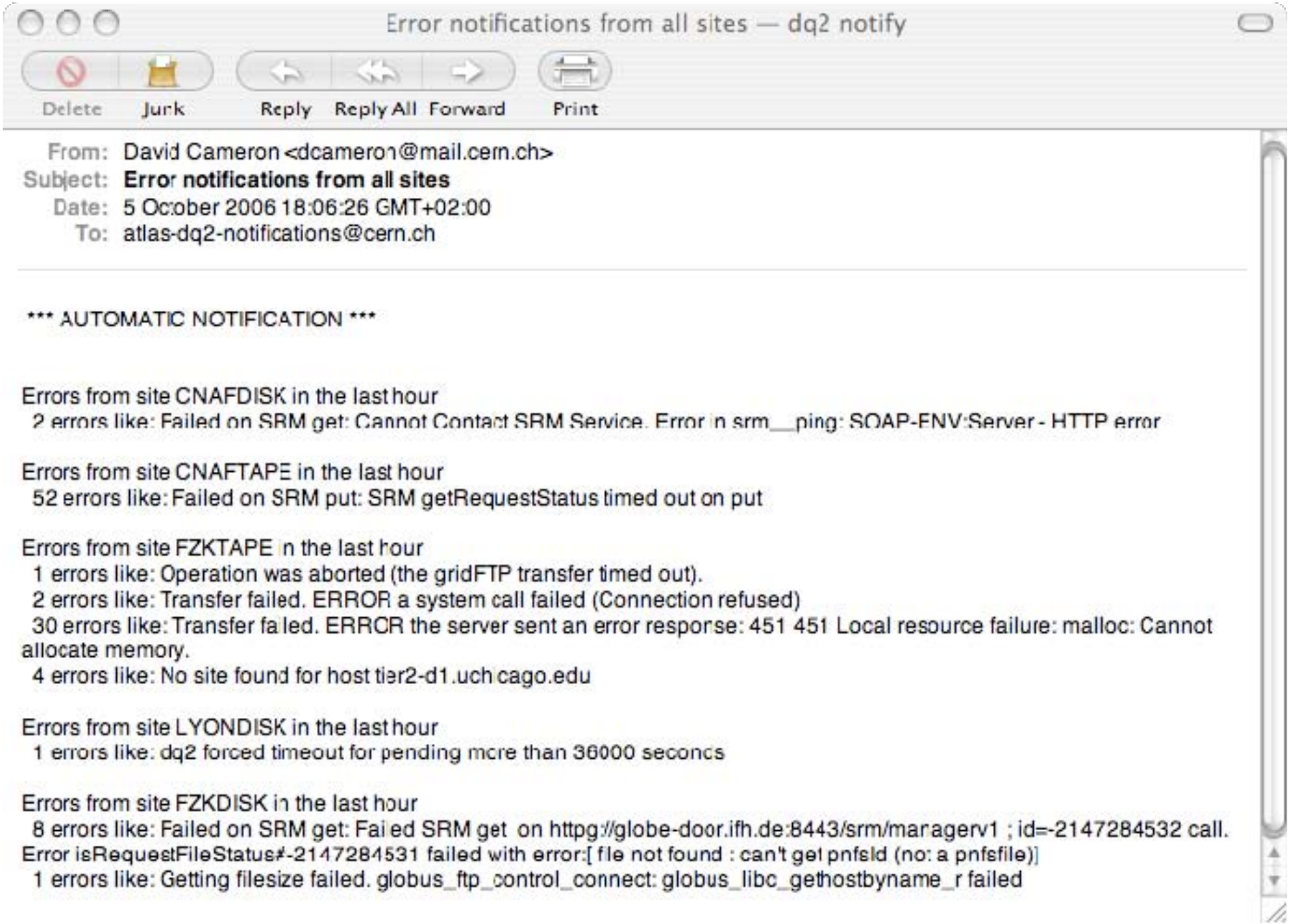




Issues/Problems



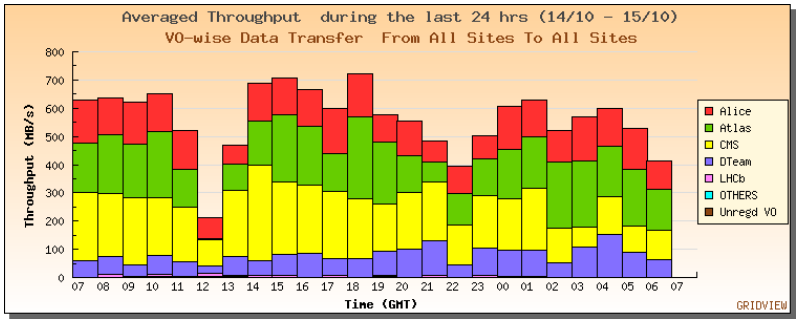
- The FTS software seems very stable
- The service can cause problems
  - Usually due to bad configuration
  - Errors reported from the service are not always clear!
  - ‘Not authorised to query request’ is confusing
- Almost all errors on file transfers are due to underlying storage software (see next slide)
  - Not clear what exactly is the root problem from the error message
  - And which end the error comes from


Error notifications from all sites — dq2 notify
Delete
Junk
Reply
Reply All
Forward
Print
From: David Cameron <dcameron@mail.cern.ch>
Subject: Error notifications from all sites
Date: 5 October 2006 18:06:26 GMT+02:00
To: atlas-dq2-notifications@cern.ch
*** AUTOMATIC NOTIFICATION ***
Errors from site CNAFDISK in the last hour
2 errors like: Failed on SRM get: Cannot Contact SRM Service. Error in srm__ping: SOAP-ENV:Server - HTTP error
Errors from site CNAFTAPE in the last hour
52 errors like: Failed on SRM put: SRM getRequestStatus timed out on put
Errors from site FZKTAPE in the last hour
1 errors like: Operation was aborted (the gridFTP transfer timed out).
2 errors like: Transfer failed. ERROR a system call failed (Connection refused)
30 errors like: Transfer failed. ERROR the server sent an error response: 451 451 Local resource failure: malloc: Cannot allocate memory.
4 errors like: No site found for host tier2-d1.uchicago.edu
Errors from site LYONDISK in the last hour
1 errors like: dq2 forced timeout for pending more than 36000 seconds
Errors from site FZKDISK in the last hour
8 errors like: Failed on SRM get: Failed SRM get on httpg://globe-door.ifh.de:8443/srm/managerv1 ; id=-2147284532 call. Error isRequestFileStatus#-2147284531 failed with error:[file not found : can't get pnfsId (no: a pnfsfile)]
1 errors like: Getting filesize failed. globus_ftp_control_connect: globus_libc_gethostbyname_r failed

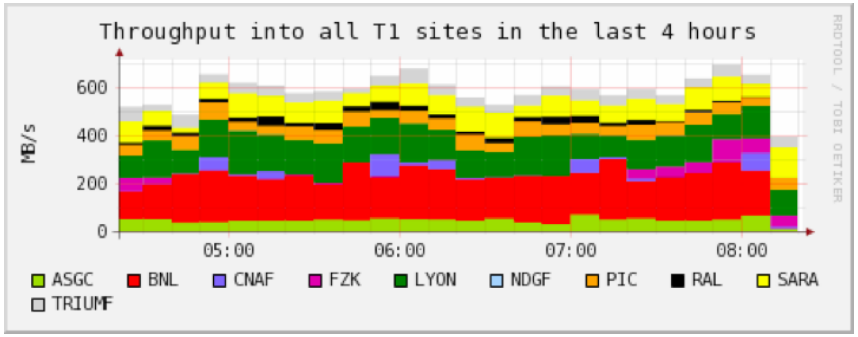
Issues/Problems

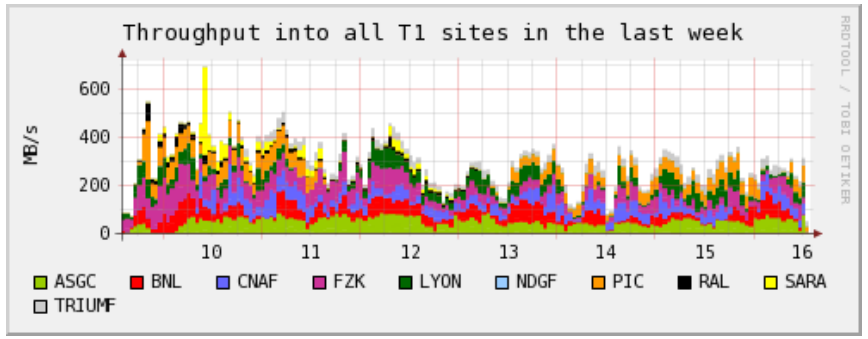
- We cannot reach stable nominal rate (780 MB/s) to all T1 sites
  - Even though we keep channels full (50 pending files/channel)
  - Errors (esp timeouts) reduce throughput **a lot**
- Other VOs running at the same time reduces performance?

Last week

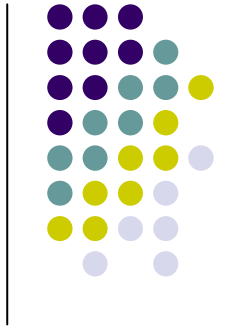


From July



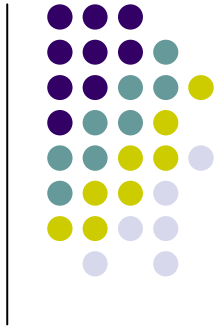


Other issues



- We have a lot of complaints of zero length files left by FTS
  - Some it seems are reported as a success by FTS
  - What are the integrity checks done by FTS at the destination storage?
- Hanging submits which are killed by us but get through eventually
  - Need to understand FTS timeouts better
- Info system dependency
  - We have asked all T1s to configure non info system star - self channels since we copy from non LCG sites
- How to optimise performance with the number of pending requests in the queue
  - We know many (~50) files in a request is good
  - But our agents work per dataset
  - We want to reduce the load from polling requests

Wishlist



- SRM 2.2 integration..
- Parsing command line outputs is nasty - how else to do it?
  - Callbacks from FTS
  - Direct WSDL interface
    - Does this change regularly?
  - Python client? ;)
- Be able to specify “do not stage” in the submit
  - i.e. don’t stage from tape if we don’t want to