

Data Management at Tier-1 and Tier-2 Centers

Hironori Ito
Brookhaven National Laboratory
US ATLAS Tier-2/Tier-3/OSG meeting
March 2010

Outline

- FTS
 - Checksum capability
- DQ2
 - DQ2 consolidation
- Monitoring and Data management
 - SAM
 - CERN Central monitor
 - Dataset monitor at BNL
- Site naming

FTS

- FTS 2.2.3 has been deployed at BNL (and T0 and other T1s)
 - ADLER32 Checksum checking via srmls
 - If SRM does not return the checksum, it will also ask the associated grid-ftp server to return the checksum.
 - Three modes for checking:
 - Compare source and destination without the checksum information from the client
 - Compare source and destination with provided checksum from the client
 - Compare the destination with provided checksum from the client.
 - Current DQ2 uses this. But, the 2nd option should be more efficient?
 - Srmls is always called for the source and the destination files regardless of the modes.
 - SRM/SE must be able to handle a lot calls of checksum checking.

FTS

- Status of checksum capabilities at various sites.
 - dCache sites (BNL, MWT2, AGLT2)
 - dCache natively supports ADLER32 checksum.
 - Checksum is stored within the metadata of the name space.
 - BeStMan sites (SLAC, NET2, SWT2)
 - BeStMan SRM supports checksum calculation as well as an external command to calculate the checksum.
 - Since BeStMan SRM does not store the checksum internally, it is up-to the backend storage and/or external program to store the checksum value if it is desired.
 - SLAC: Checksum value is stored in extended attributes of the storage file system (GPFS)
 - SWT2: Checksum is not supported. However, the backend gridftp returns the ADLER32 checksum.
 - NET2: Checksum is not supported. It is currently investigated.

Checksum Performance

Using a 3600MB file, the following is the typical(?) breakdown of transaction time within a transfer.

	Total	Actual transfer	checksum
dCache (AGLT2)	210 seconds	200 seconds	0.3 seconds
BeStMan (SLAC)	120 seconds	100 seconds	5 seconds

Checksum option in BeStMan

- `defaultChecksumType=adler32`
- `showChecksumWhenListingFile=true`
- `hexChecksumCommand="some command"`
- `NoDirBrowsing=true`

What is the best practice?

- Storing checksum will reduce the load when it is a source of a transfer. Remember: `srmls` is issued for source as well as destination.
 - Storing in extended attributes of the file system sounds very efficient.
 - Example(s) for external command(s) to write-to and read-from SE to work with BeStMan would be very helpful?
 - Does it distribute the load when multiple files are written at the same time?
 - Transaction capability test is planned within the throughput test program.

DQ2

- New DQ2 SS, which utilizes the FTS checksum capability, is deployed at BNL.
 - Destination checksum is checked by srmls against known checksum.
 - No more US special modification DQ2 SS
 - It can simultaneously supports checksum capable and non-capable sites.
 - All US sites will be served by BNL DQ2 SS.
 - Currently, BNL, AGLT2, MWT2 and SLAC sites are being served by BNL DQ2.

Consequence of DQ2 SS consolidation

- Advantage
 - T2 administrators are no longer responsible for maintaining DQ2 SS at T2s.
- Disadvantage
 - The log is no longer available at each T2 center, eliminating the capability for T2 administrators to look at the logs directly (via grep) to debug DDM problems at own sites.
 - DQ2 log viewer should reduce (or eliminate) the need to have logs locally.
 - <http://www.usatlas.bnl.gov/dq2log/dq2log> (for T1 and T2s)
 - <http://www.usatlas.bnl.gov/dq2logt3/dq2log> (for T3)

DQ2 Log Viewer

- The method to see the content of DQ2 SS logs remotely for anyone to identify the problem.
- It uses Ferret (with Ruby on Rails), which is a full feature text search engine library. (like Lucene Java)
- Each entry in the logs are indexed and stored.
- The time stamp is stored separately to speed up the search by a time constraint
- It has found to work with at least a few hundred million entries/rows.
- The complex search is possible using “AND, OR, NOT”
- The search is very fast. (Comparing to grep, it feels much faster.)

DQ2 Log viewer

- The link to the log viewer has been in place within PANDA monitor to debug the “transferring” jobs.
 - The search can be quite complex. It is making large “OR” of the following items,
 - Dataset name
 - DUID
 - LFNs and GUIDs of all files in a given dataset ;
 - The search result is color coded to easily identify the cause of the problems from many entries. Eg. INFO is green. WARNING is yellow. Error information is red.
- A user can construct own search
- It shows the plots of all transfers within last 10 minutes as well as latest actual log entries.
 - Updated live 10 times while the page is viewed. User can request to view many more times.
 - The log is “actual” because it is the result of simple “tail” and not from the index.
- In the future, the remote HTTP API might be added if it is found to be more useful than the web page.

BNL Tier3 DQ2 logs

View and search BNL DQ2 log.

Search words(one string/word per each line. Can use * to match.)

For example, to search lfn of step09.20201003000438L.physics_C.merge.DPD_TYPE06.closed_0001_1263898879

and guid of c2e79551-2318-4818-939a-f88ecdddec96f

Type (step09.20201003000438L.physics_C.merge.DPD_TYPE06.closed_0001_1263898879' OR 'c2e79551-2318-4818-939a-f88ecdddec96f')

(Or use the Dataset name option below.)

Dataset name (just one):

Time Range

Start Time(eg. 2009/10/10 19:54) :

End time (eg. 2009/10/11 20:00) :

Max entries per page: 200

Click to search

Showing the last 100 lines of the log. Updated every 10s. Automatically stops after 10 times. You must click "Start tail" again to restart tailing if you want to see it.

[Start tail](#) [Stop tail](#)

2010-03-07 20:17:17,679 - INFO - Nothing to do for subscription mc09_7TeV.108400.PythiaB_ccmu4X.merge.AOD.e477_s624_s633_r1064_r1051_tid108166_00 for site WISC_MCDISK.

2010-03-07 20:17:17,680 - INFO - Subscription mc09_7TeV.106022.PythiaWtaunu_1Lepton.merge.AOD.e468_s624_s633_r1064_r1051_tid108139_00 for site WISC_MCDISK will query ['BNL-O8G2_MCDISK', 'AGLT2_MCDISK'] and (took off: [], blacklisted: [])

- Result of an example search “ddo.000001.Atlas.Ideal.DBRelease.v08010301” as a dataset to be searched.

```
2010-02-19 09:40:23,943 - INFO - FTS ID b656540f-1d64-11df-af78-f7b76ce7ef85 SERVER https://fts02.usatlas.bnl.gov:8443/glite-data-transfer-fts/services/FileTransfersrm://osgsv04.slac.stanford.edu/srm/v2/server?SFN=/xrootd/atlas/atlashotdisk/ddo/DBRelease/v080301/ddo.000001.Atlas.Ideal.DBRelease.v080301/DBRelease-8.3.1.tar.gz srm://osgx1.hep.uiuc.edu:8443/srm/managerv2?SFN=/pnfs/hep.uiuc.edu/data4/atlas/datadisk/ddo/DBRelease/v080301/ddo.000001.Atlas.Ideal.DBRelease.v080301/DBRelease-8.3.1.tar.gz
2010-02-19 09:40:59,301 - INFO - MATERIALIZED VIEW ddo.000001.Atlas.Ideal.DBRelease.v08010301 FOR SITE ILLINOISHEP_DATADISK
2010-03-06 17:40:39,022 - INFO - Queued 92 files missing for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK.
2010-03-06 17:40:39,970 - INFO - Queued 92 files for transfer for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 using channel BNL-OSG2_HOTDISK -> ILLINOISHEP_DATADISK
2010-03-07 00:14:24,851 - INFO - Nothing to do for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK.
2010-03-07 00:16:26,228 - INFO - Subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK will query [BNL-OSG2_HOTDISK] and (took off: [], blacklisted: [])
2010-03-07 00:16:26,230 - INFO - Nothing to do for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK.
2010-03-07 00:18:27,694 - INFO - Subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK will query [BNL-OSG2_HOTDISK] and (took off: [], blacklisted: [])
2010-03-07 00:18:27,696 - INFO - Nothing to do for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK.
2010-03-07 00:19:40,103 - INFO - FAILED GUID 0ad58976-b53f-4828-b2f6-833509ca2c01 FOR BNL-OSG2_HOTDISK->ILLINOISHEP_DATADISK [FTS State [Failed] FTS Retries [1] Reason [SOURCE error during TRANSFER_PREPARATION phase: [LOCALITY] Source file [srm://dcsrm.usatlas.bnl.gov/pnfs/usatlas.bnl.gov/HOTDISK/ddo/DBRelease/v08010301/ddo.000001.Atlas.Ideal.DBRelease.v08010301/08010301_0141562.tar.gz]: locality is UNAVAILABLE] Source Host [dcsrm.usatlas.bnl.gov]]
2010-03-07 00:20:28,931 - INFO - Subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK will query [BNL-OSG2_HOTDISK] and (took off: [], blacklisted: [])
2010-03-07 00:20:29,312 - WARNING - Skipped 1 files from BNL-OSG2_HOTDISK due to failed transfers for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for ILLINOISHEP_DATADISK.
2010-03-07 00:22:32,670 - INFO - Subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK will query [BNL-OSG2_HOTDISK] and (took off: [], blacklisted: [])
2010-03-07 00:22:33,049 - WARNING - Skipped 1 files from BNL-OSG2_HOTDISK due to failed transfers for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for ILLINOISHEP_DATADISK.
2010-03-07 00:24:36,079 - INFO - Subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 for site ILLINOISHEP_DATADISK will query [BNL-OSG2_HOTDISK] and (took off: [], blacklisted: [])
2010-03-07 00:24:36,464 - INFO - Queued 1 files for transfer for subscription ddo.000001.Atlas.Ideal.DBRelease.v08010301 using channel BNL-OSG2_HOTDISK -> ILLINOISHEP_DATADISK.
2010-03-07 00:24:45,380 - INFO - FTS ID 1111-11-11-1111-1111-1111-1111-1111 SERVER https://fts02.usatlas.bnl.gov:8443/glite-data-transfer-fts/services/FileTransfersrm://osgsv04.slac.stanford.edu/srm/v2/server?SFN=/xrootd/atlas/atlashotdisk/ddo/DBRelease/v080301/ddo.000001.Atlas.Ideal.DBRelease.v080301/DBRelease-8.3.1.tar.gz srm://osgx1.hep.uiuc.edu:8443/srm/managerv2?SFN=/pnfs/hep.uiuc.edu/data4/atlas/datadisk/ddo/DBRelease/v080301/ddo.000001.Atlas.Ideal.DBRelease.v080301/DBRelease-8.3.1.tar.gz
```

Monitoring and Data Management

- SAM (ATLAS VO)
 - Test SE, CE and FTS if available.
 - CE test is completely failing at all OSG sites. It will be modified to work with OSG sites.
 - SE is being tested by lcg-cr, lcg-cp and lcg-del in that order.
 - The site availability is evaluated after about 2 hours from the start of the test to take into the account of possible delayed results of various tests.
 - It is a part of criteria for site status to turn on/off the site automatically.
 - Grid View
 - http://gridview.cern.ch/GRIDVIEW/same_index.php
 - SAM dashboard
 - <http://dashb-atlas-sam.cern.ch/dashboard/request.py/latestresultssmry>
 - The automated warning program is already looking for failures from GridView, and send email to T2 admins when it finds failures.

SAM via Grid View

		<u>Test Status : OK</u>	
Site	: AGLT2	Node	: head01.aglt2.org
Service	: SRMv2	Test	: SRMv2-ATLAS-lcg-cp
Criticality Defining VO	: Atlas	Test VO	: ATLAS
Critical	: Y	Execution time	: 07-Mar-2010, 06:57:13
<u>Test Environment</u>			
Name	: Value		
submitterDN	: /O=dutchgrid/O=users/O=nikhef/CN=Kors Bos		
LFC_HOST	: lfc.triumf.ca		
Test Summary :	:		

Detail Result:

ATLAS specific test checking if a file can be copied back from head01.aglt2.org
launched from sam212.cern.ch

----- Copy back the lfn:SRM-lcg-cr-head01.aglt2.org-ATLASGROUPDISK-1267944912

lfn:SRM-lcg-cr-head01.aglt2.org-ATLASGROUPDISK-1267944912

lfc.aglt2.org LFC_HOST

++ pwd

+ lcg-cp -v --vo atlas -D srmv2 -T srmv2 lfn:SRM-lcg-cr-head01.aglt2.org-ATLASGROUPDISK-1267944912

file:/home/samatlas/.same/SRMv2/nodes/head01.aglt2.org/testFile.txt

Using grid catalog type: LFC

Using grid catalog : lfc.aglt2.org

VO name: atlas

Checksum type: None

Trying SURL

srm://head01.aglt2.org/pnfs/aglt2.org/atlasgroupdisk/SAM/SRM-lcg-cr-head01.aglt2.org-ATLASGROUPDISK-1267944912 ...

Source SE type: SRMv2

Source SRM Request Token: -2081536357

Source URL: /grid/atlas/dq2/SAM/SRM-lcg-cr-head01.aglt2.org-ATLASGROUPDISK-1267944912

File size: 41472

Source URL for copy:

gsiftp://msufe07.aglt2.org:2811/pnfs/aglt2.org/atlasgroupdisk/SAM/SRM-lcg-cr-head01.aglt2.org-ATLAS

Sites Abbreviations

W3C HTML 4.01

W3C CSS



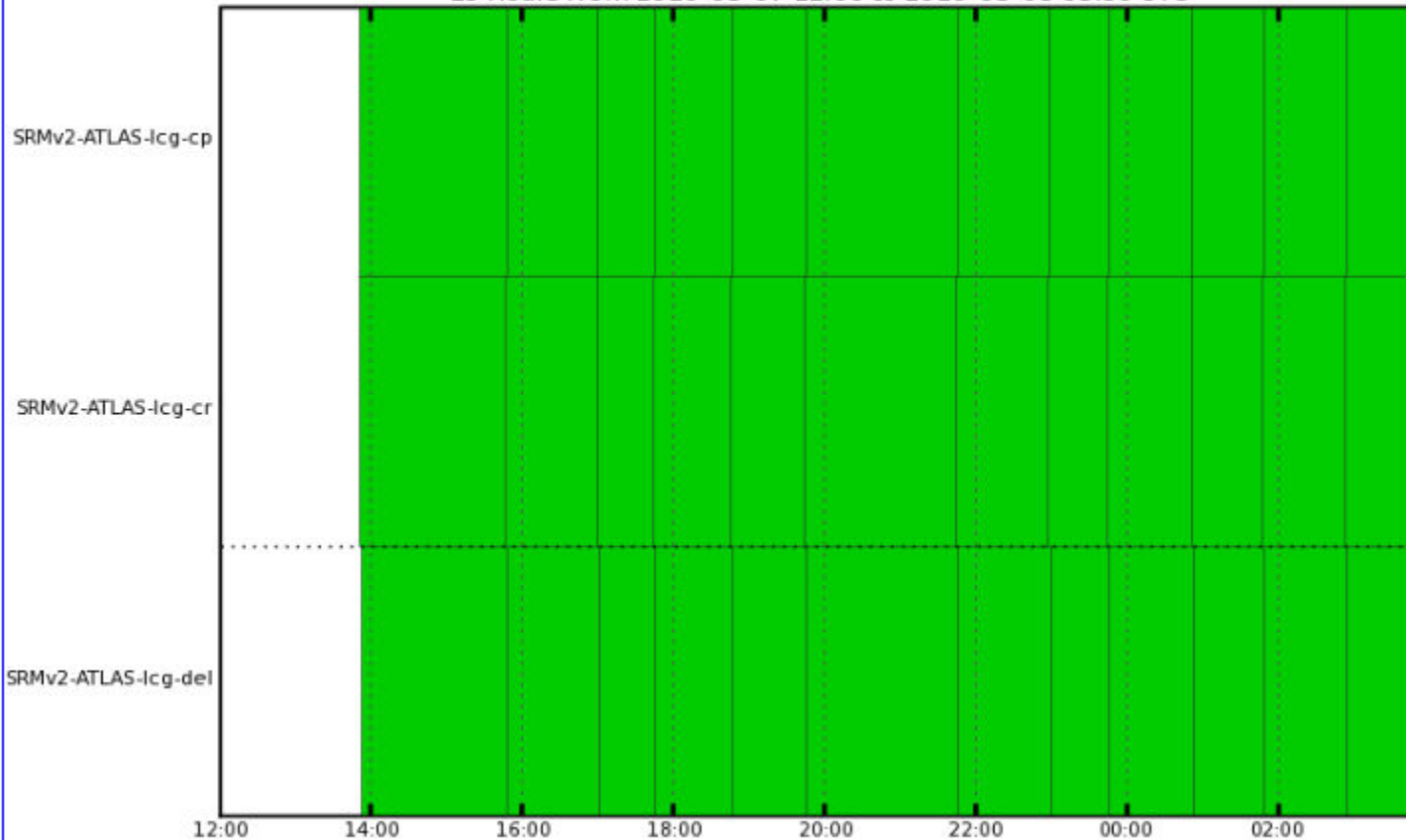
SAM via Dashboard

Previous node ()

Next node ()

Test results for dcsrm.usatlas.bnl.gov

15 Hours from 2010-03-07 12:00 to 2010-03-08 03:50 UTC



Legend:

Note: brig

Sitename

BNL-ATLAS

Algorith

Show

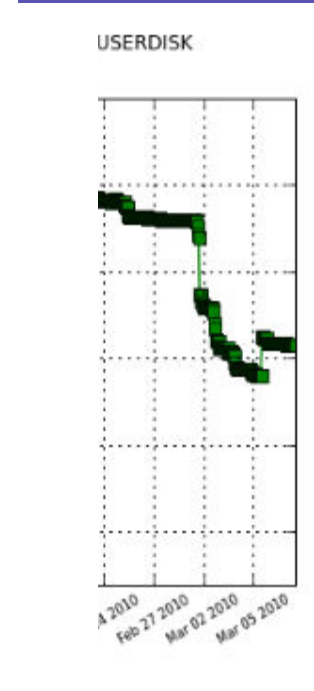
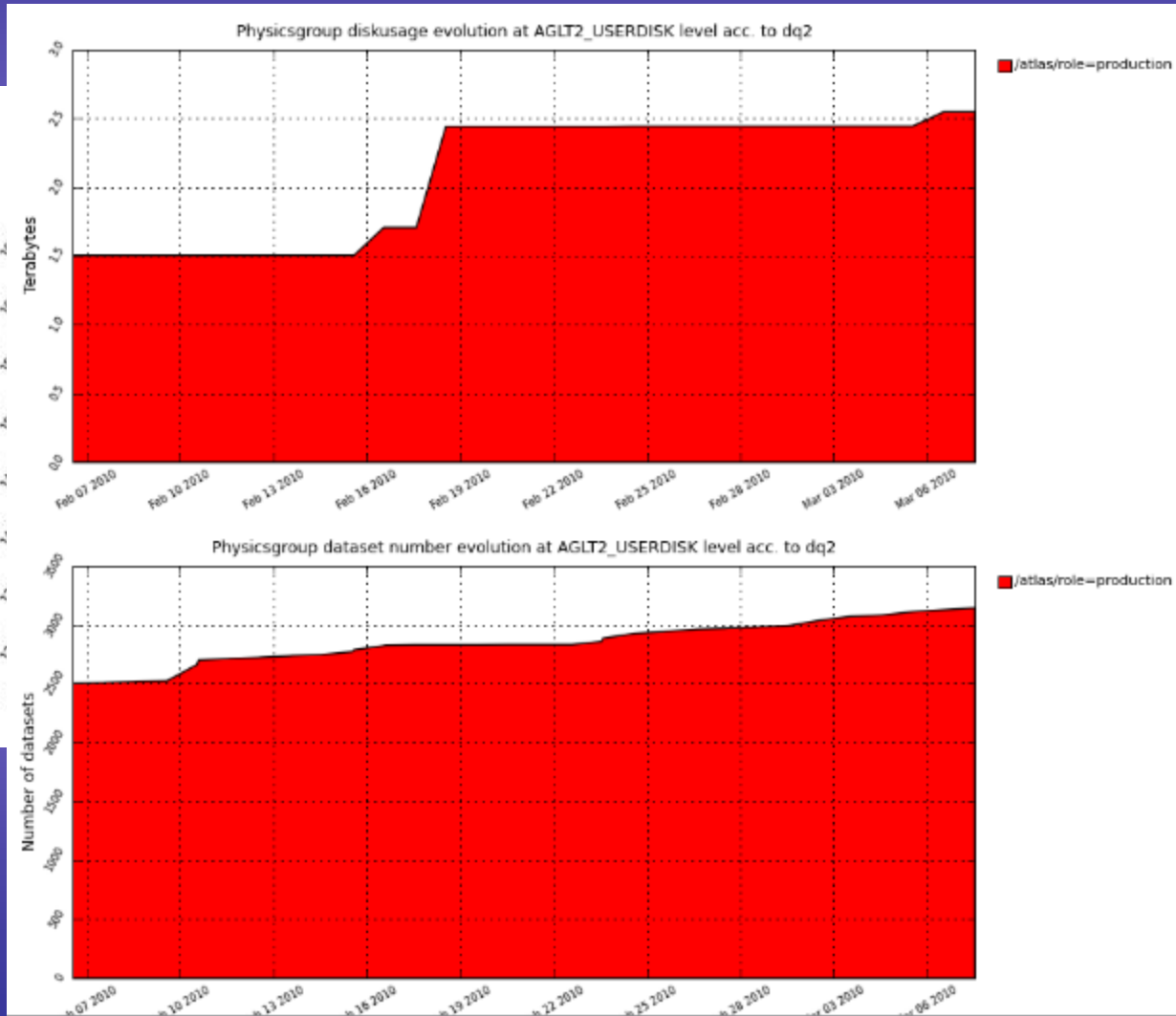
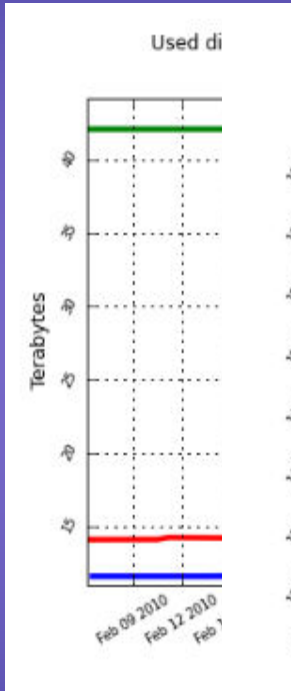
Dis

new v

Monitoring and Data Management

- Central monitoring
 - http://atlddm02.cern.ch/dq2/accounting/site_view/
 - Used space via DQ2 tracker
 - Keeping results from DQ2 subscriptions
 - Used and available space via srm
 - srm-get-space-metadata
 - Although the “available” space of a given space token area by srm is a bit “arbitrary”, the transfer will fail if it reports no space regardless of the actual, physical space in its backend storage.
 - Central Deletion
 - <http://atlddm02.cern.ch/dq2/deletion/search/>

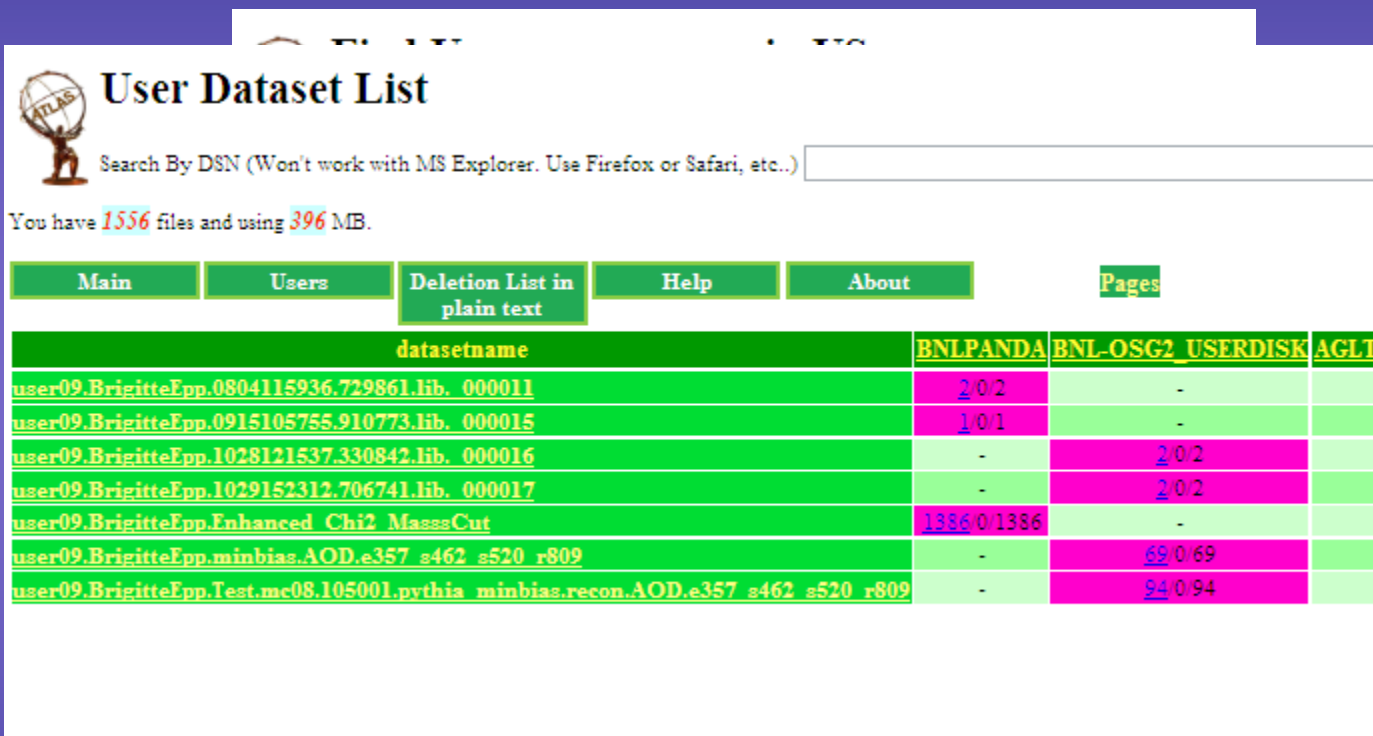
Central monitoring



Monitoring and Data Management

- Dataset monitor
 - It combines DDM central, replica and LFC catalogs in one catalog.
 - http APIs exist to access the information remotely.
 - The designed to provide the fast breakdown of the datasets information at a site.
 - Break down the space use by dataset types, project type etc...
 - Typical question: How much space is occupied by ESDs in DATADISK?
 - Use to notify users for deletion of user datasets.
 - Use for throughput testing program

Deletion Notification of user datasets



User Dataset List

Search By DSN (Won't work with MS Explorer. Use Firefox or Safari, etc.)

You have **1556** files and using **396** MB.

[Main](#) [Users](#) [Deletion List in plain text](#) [Help](#) [About](#) [Pages](#)

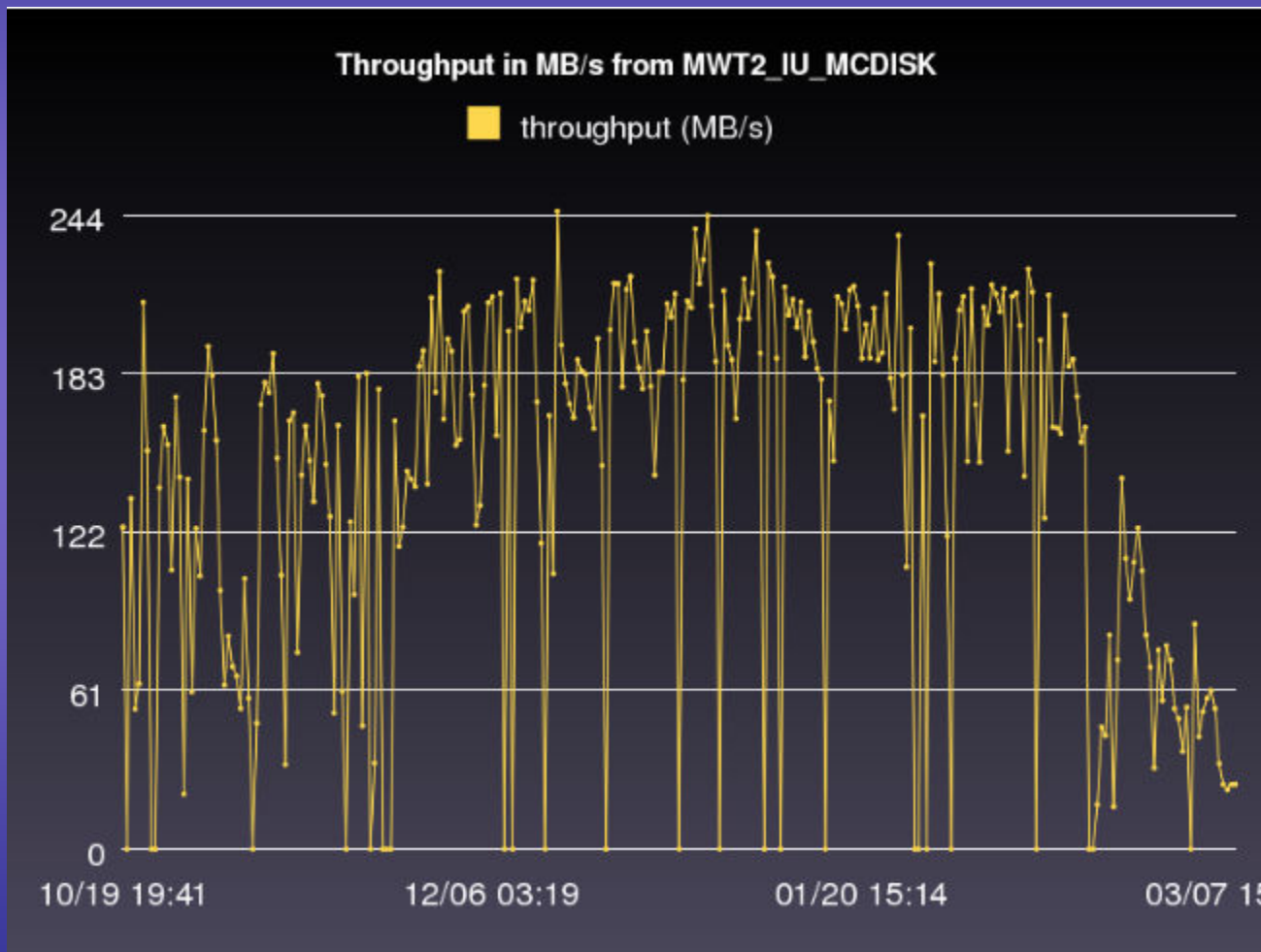
datasetname	BNLPANDA	BNL-OSG2_USERDISK	AGLT2
user09.BrigitteEpp.0804115936.729861.lib.000011	2/2	-	
user09.BrigitteEpp.0915105755.910773.lib.000015	1/1	-	
user09.BrigitteEpp.1028121537.330842.lib.000016	-	2/2	
user09.BrigitteEpp.1029152312.706741.lib.000017	-	2/2	
user09.BrigitteEpp.Enhanced_Chi2_MassCut	1386/1386	-	
user09.BrigitteEpp.minbias.AOD.e357_s462_s520_r809	-	69/69	
user09.BrigitteEpp.Test.mc08.105001.pythia_minbias.recon.AOD.e357_s462_s520_r809	-	94/94	

- [/C=AU/O=APACGrid/OU=OSYD/CN=Jason Lee](#)
- [/C=BR/O=ICPEDU/O=UFF BrGrid CA/O=UFRJ/OU=COPPE/CN=Felipe Fink Graef](#)
- [/C=CA/O=Grid/OU=lps.umontreal.ca/CN=Bertrand Brelier](#)
- [/C=CA/O=Grid/OU=lps.umontreal.ca/CN=John Idarraga](#)
- [/C=CA/O=Grid/OU=lps.umontreal.ca/CN=Jonathan Ferland](#)
- [/C=CA/O=Grid/OU=lps.umontreal.ca/CN=Privali Banerjee](#)
- [/C=CA/O=Grid/OU=phas.ubc.ca/CN=ChangWei Loh](#)
- [/C=CA/O=Grid/OU=phvs.ualberta.ca/CN=Douglas Gingrich](#)
- [/C=CA/O=Grid/OU=phvs.ualberta.ca/CN=...](#)

Throughput Test Program

- 20 files are sent from BNL to T2s
- 10 files are sent from T2s to T2s
- The size of each file is 3600MB.
- It records the number of completed transfers and their corresponding transfer times to estimate the throughput value.
- Although it does not measure the absolute maximum throughput, the relative trend seems to be useful to identify the site problem.

Typical Throughput Results



Dark Data

- Not all monitor shows the same amount of used space 😊
 - Any mismatch is a indication of dark data.
 - Many catalogs (Replica, LFC, SRM/SE name space)
 - Sum of LFC =? Sum of Replica
 - Sum of SRM =? Sum of LFC
- Cause
 - Writing multiple copies in SRM (DQ2 retry)
 - Deletion from one of the catalogs without the others (Replica, LFC, SE/namespace)
 - Central deletion have long delay between the various deletions, leading to apparent dark data.
- Fix
 - T2s: CCC.py
 - T3gs: storageManager.py (will be discussed in T3 talks.)
 - BNL: Still under consideration to find the most stable, scalable, fast way without interfering the performance of SE.
 - Use PNFSIDs to check instead of name space.
 - PNFSIDs are stored in BNL LFC and Companion table of dCache, eliminating access to the name space.

Naming Confusion

- There are different systems which hold site name: TiersOfATLAS, OSG BDII, PANDA, FTS, etc...
- Within one system, one site could have multiple names.
 - Example:
 - ToA: site name and alternate name
 - BDII: Resource name, Resource group name
- AGIS will eventually connect all sites with different names. But, it will only work if the name is following a basic rule of connecting various site in the different system.
 - Some site names probably need to be changed.

Rule of Naming

(May not be correct!)

- Rule 1:
 - “Alternate name” in ToA for a site must be the same as SE site name within CERN BDII.
- Rule 2:
 - SE site name within CERN BDII must be the same as the Resource group name.
- Rule 3:
 - CE site name within CERN BDII must be also the same as Resource group name