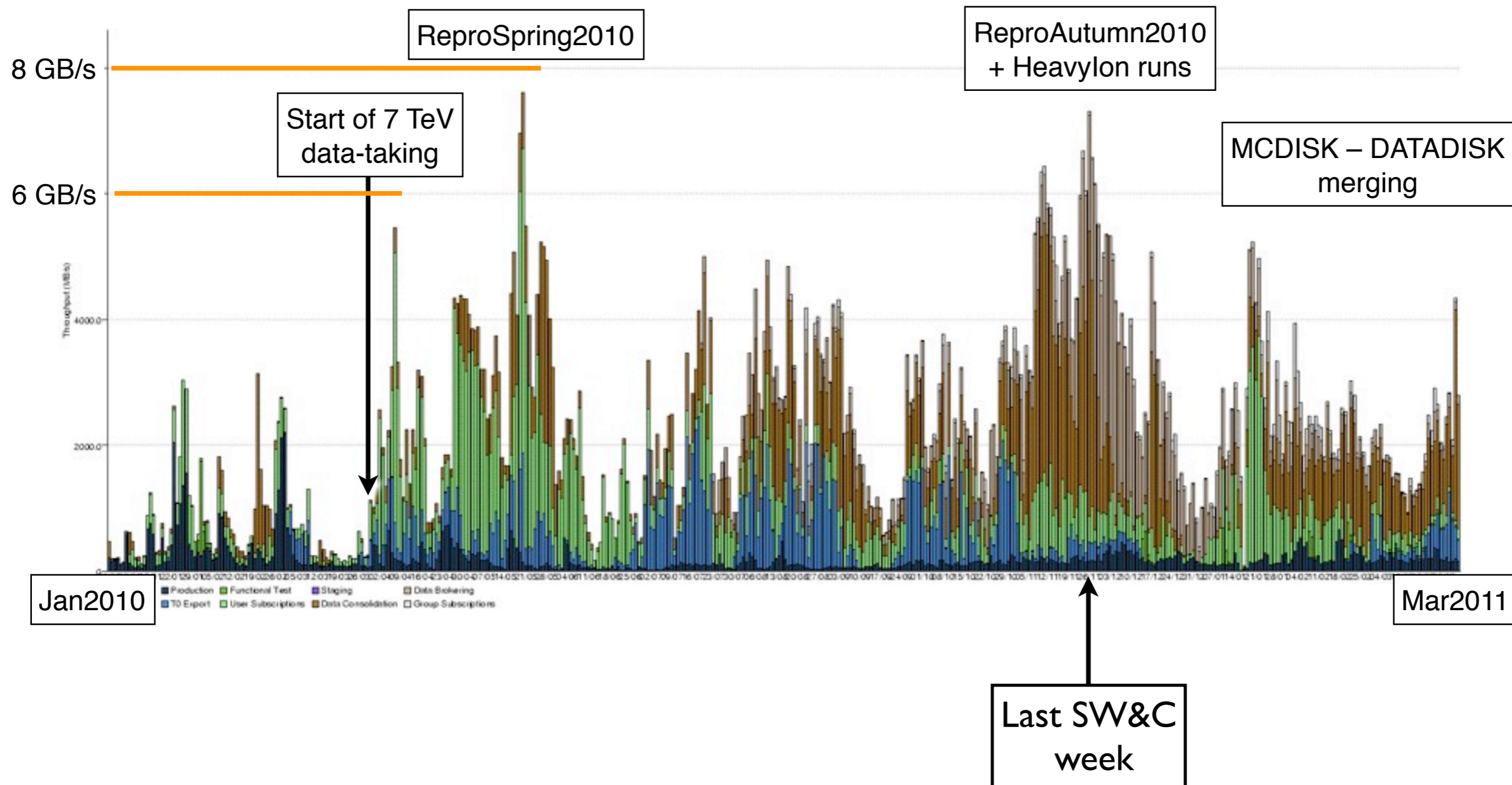


DDM operations

Ueda I.
ATLAS S&C workshop
2011.04.04.

DDM Transfer Activities



Overview

Operation work since last SW&C

- Merging MCDISK into DATADISK
 - ▶ As has been discussed at the last SW&C, and approved by CREM

New implementations

- Data Distribution 2011
 - ▶ also in the plenary session this morning
- Inter-Cloud Transfers
 - ▶ also in the plenary session this morning

Recent issues

- Missing Input Files
 - ▶ sounds like an old issue, but is a new one after solving the old one
- Deletion backlog

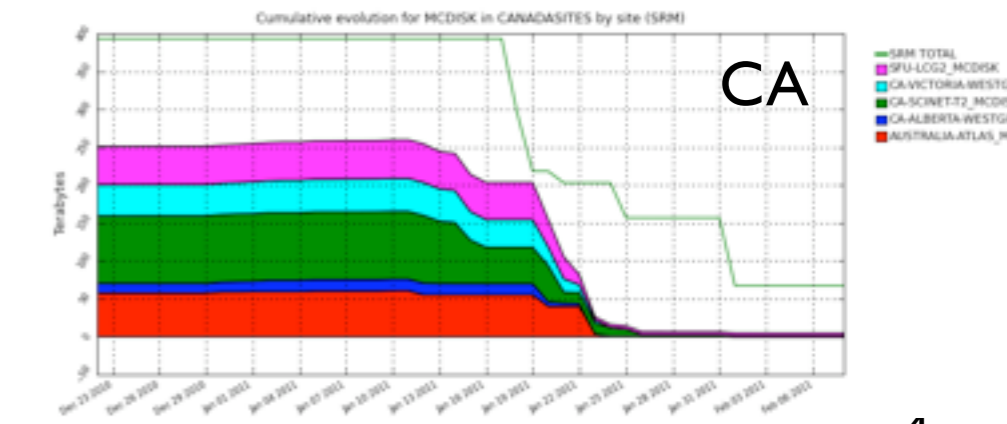
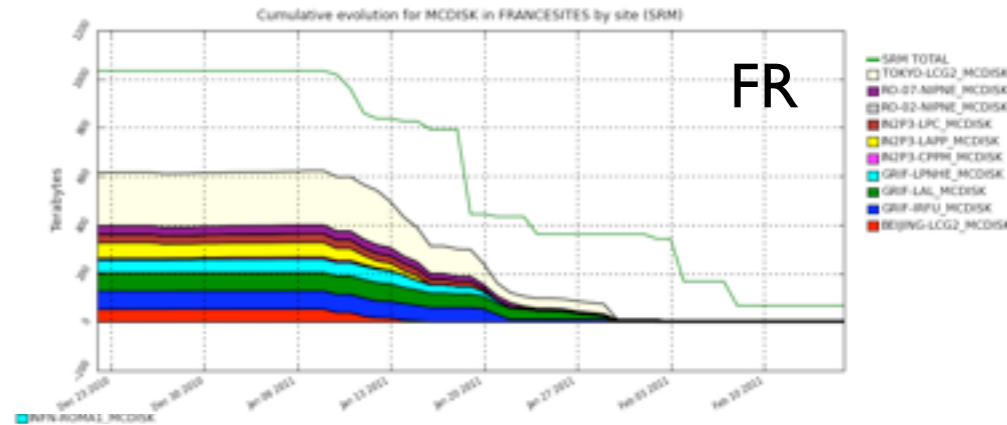
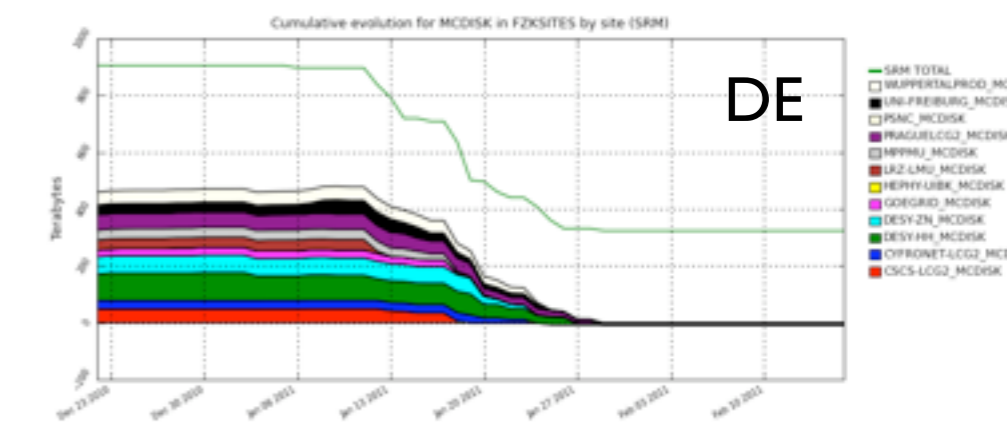
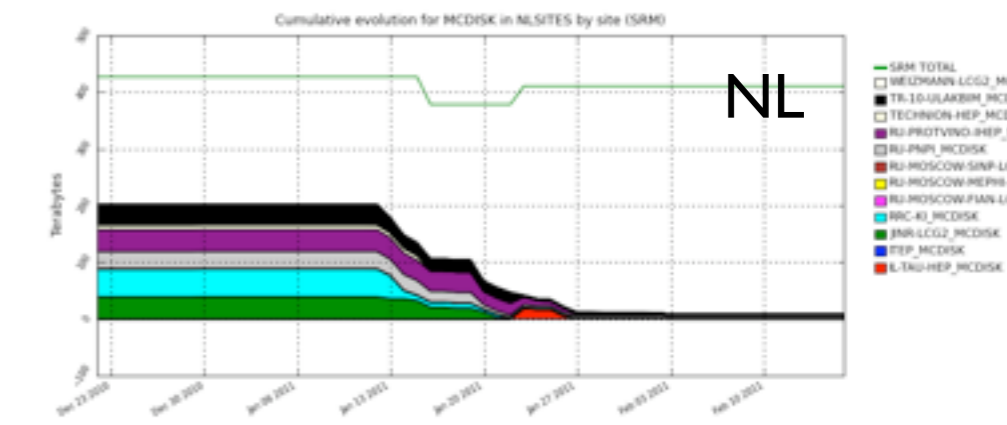
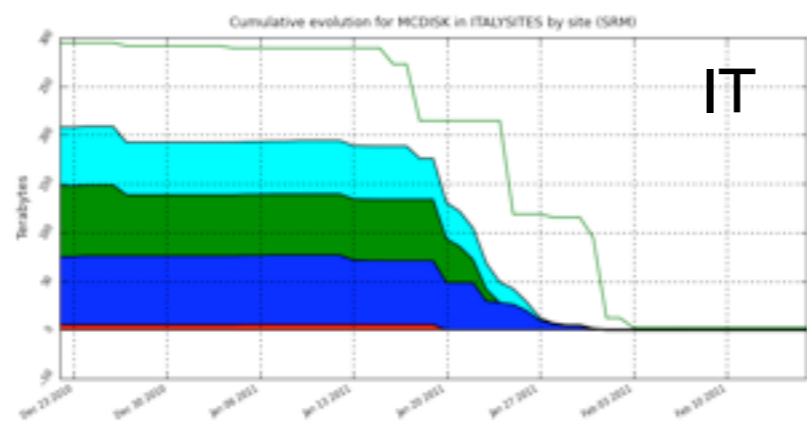
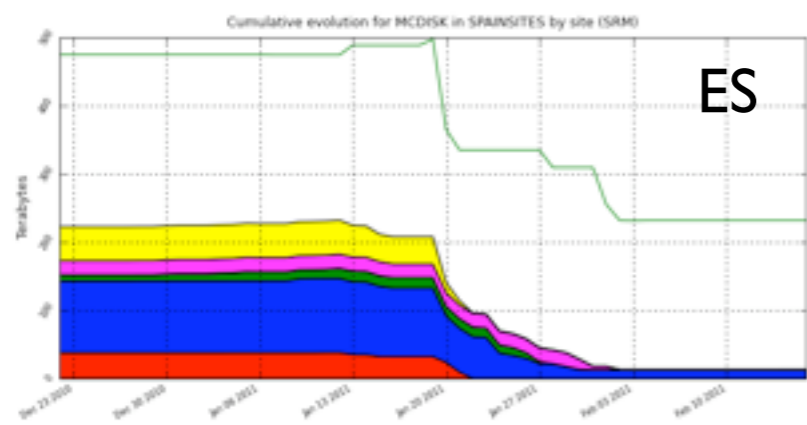
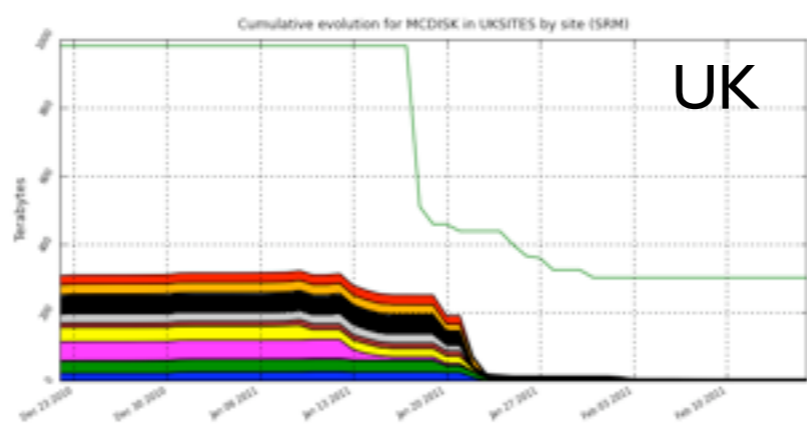
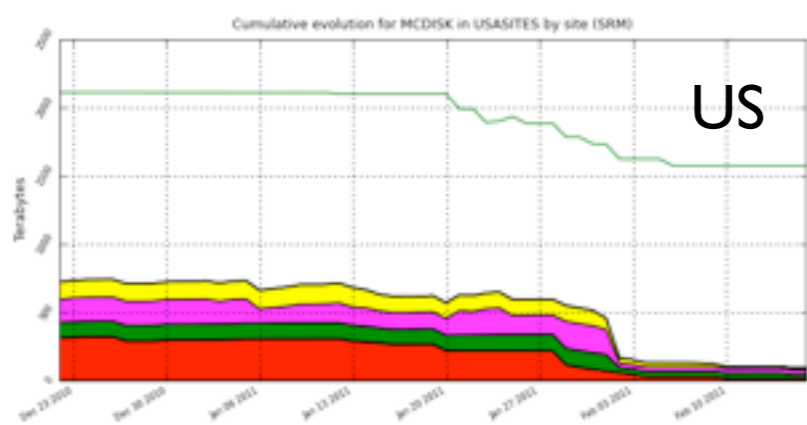
On-going project

- LFC consolidation
 - ▶ As has been discussed at the last SW&C, and at the Napoli workshop

Merging MCDISK into DATADISK (T2s)

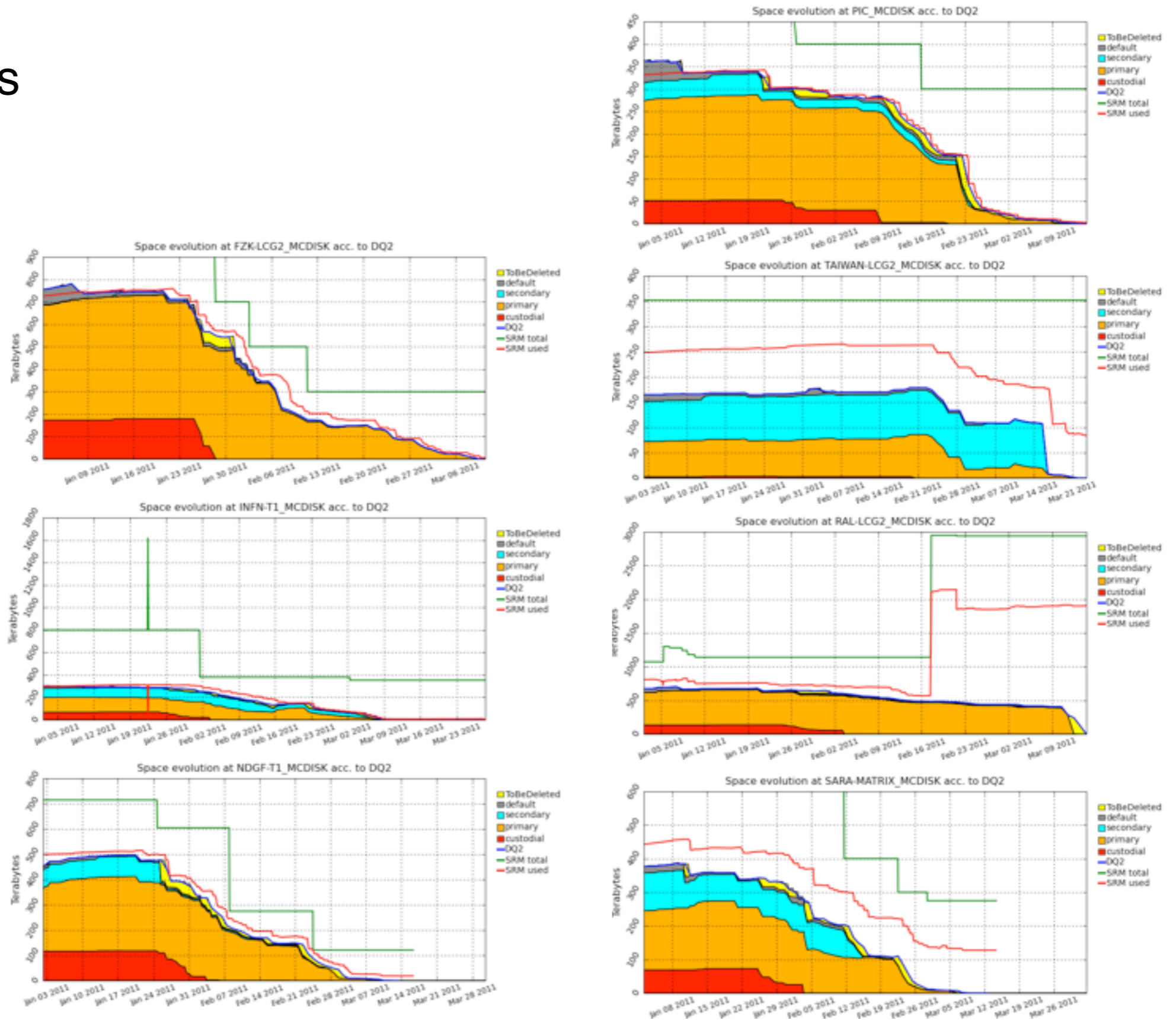


MCDISK has been merged into DATADISK at all the T2(3)s

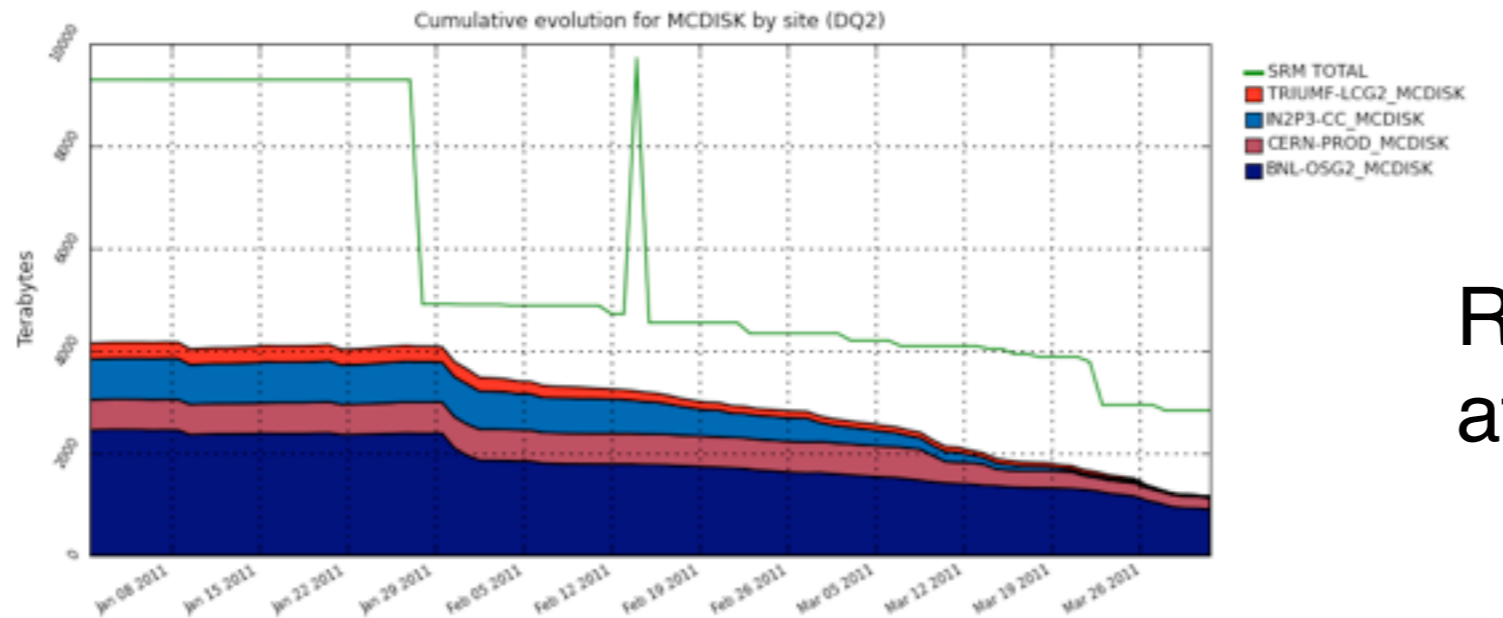


Merging MCDISK into DATADISK (T1s)

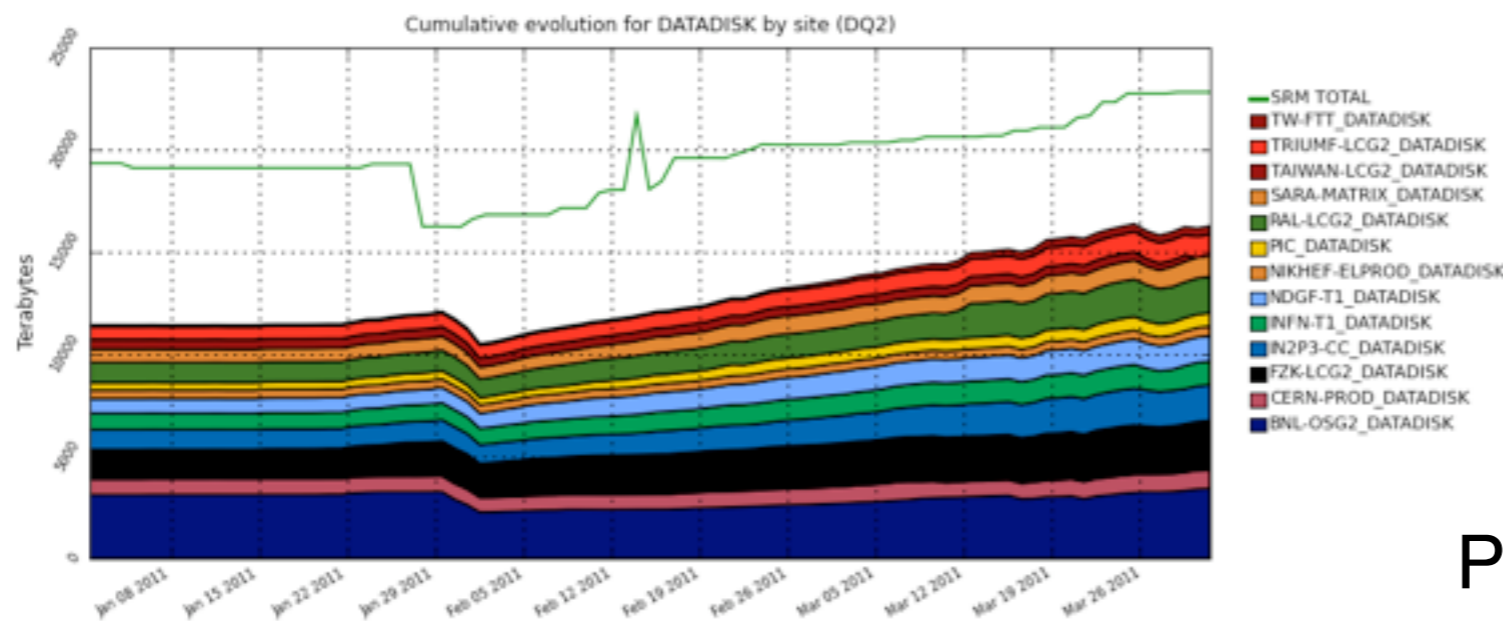
and 7 T1s



Merging MCDISK into DATADISK (T1s)



Remaining MCDISK
at 3 T1s and CERN



Procedures and
Issues in Cedric's talk

Data Distribution 2011

DAQ rate ~ 400 HZ

- ▶ impossible to do the same distribution as in 2010

“Life without ESD”

- ▶ Data Preparation Coordination (20 January 2011)
 - <http://indico.cern.ch/conferenceDisplay.py?confId=121964>
- ▶ ATLAS Weekly (08 February 2011)
 - <http://indico.cern.ch/contributionDisplay.py?contribId=3&confId=119624>
- ▶ ATLAS Week (01 March 2011)
 - <http://indico.cern.ch/contributionDisplay.py?contribId=32&sessionId=2&confId=105534>
- ▶ **and this morning in the plenary session**
- No full ESD on disk, No ESD on tape
- Replication to T1_DATADISK only for limited use cases (pre-defined)
 - ▶ “bulk” streams stay on T1 disk for limited period (6–8 weeks)
 - No replication to T2_DATADISK (pre-defined, PD2P, DaTRI) expected
 - ▶ small streams stay on disk with $v(N) + v(N-1)$
- DaTRI requests to groupdisk, localgroupdisk possible

RAW data compression at Tier-0 (factor ~2)

- ▶ under preparation — **to be discussed on Friday in the plenary session**
- RAW data on T1_DATADISK (for “discovery”, and also for “expired” ESD)

Data Distribution 2011

Tier-0 exports

- 2 copies of RAW (disk and tape at different T1s)
- 1 copy of ESD distributed between T1s
- 2 copies of AOD distributed between T1s
- 2 copies of DESD distributed between T1s

T1-T1 replication

- adds 1 copy for each of ESD , AOD and DESD on T1s
- extra copies with PD2P

T2 replication

- No pre-placement on Tier-2s
- AOD and DESD will be replicated with PD2P
- ('T2Ds' could be treated differently — see later slides)

Inter-Cloud Transfers

a la Cloud Model

- T2a-T1a-T1b-T2b = 'commissioned' links
 - ▶ puts a pressure on T1s (SRM access, buffer space)
 - ▶ more steps = more overhead = more delays, especially for small files
 - ▶ more steps = more points of failures

DaTRI transfers

- T2a-T1b-T2b = via T1 in the destination cloud
 - ▶ less step, less pressure on T1s
- subscription and deletion per request
 - ▶ Sometimes filling up the T1b_SCRATCHDISK

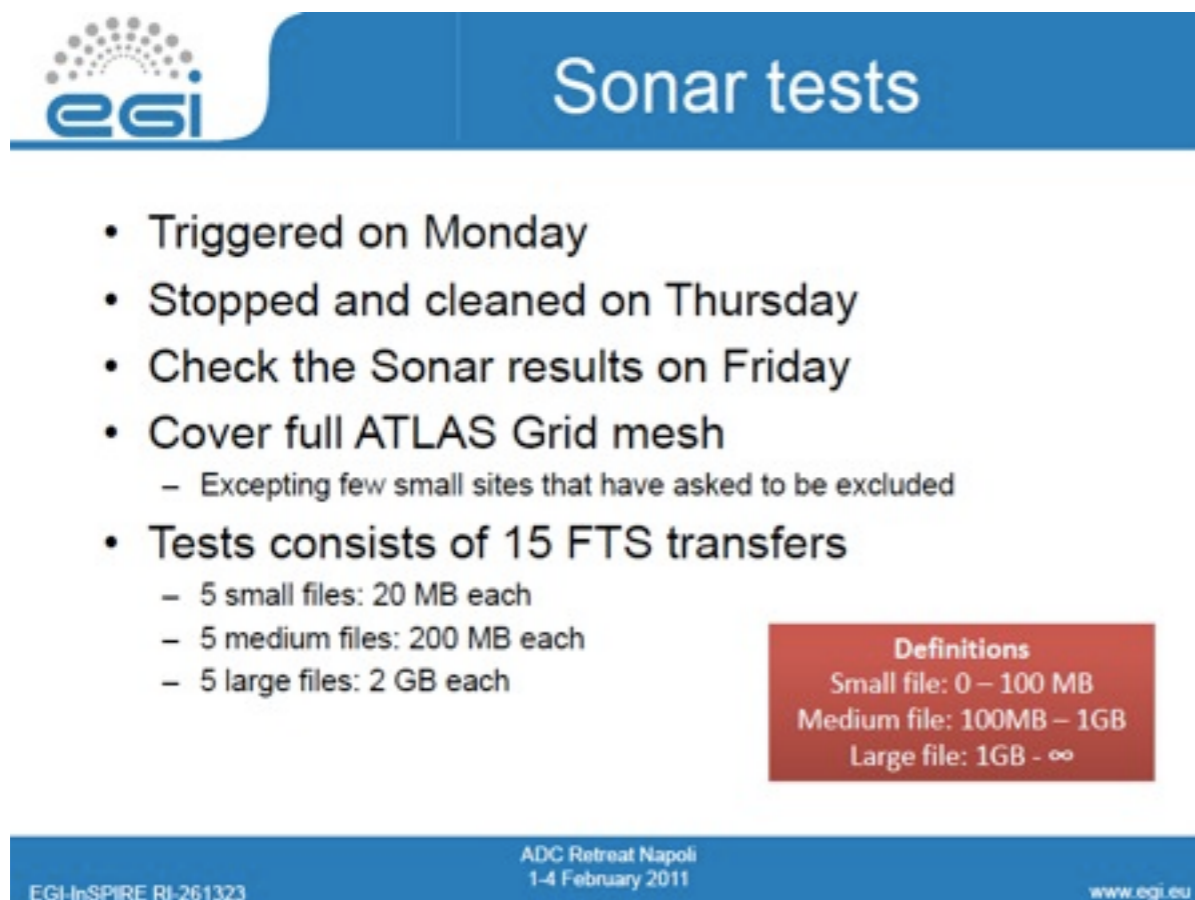
Why not direct T2a-T2b?

- many combinations of T2(3)s would be capable of direct transfers
 - ▶ especially for small files
- but not all of them, need to commission

Inter-Cloud Transfers

Commissioning inter-cloud links

- ‘Sonar’ tests to inject transfers to measure link performance
 - ▶ from every site to every site (with some exceptions)
 - ▶ Transfer submission smeared from Monday to Friday
- Transfer statistics monitoring

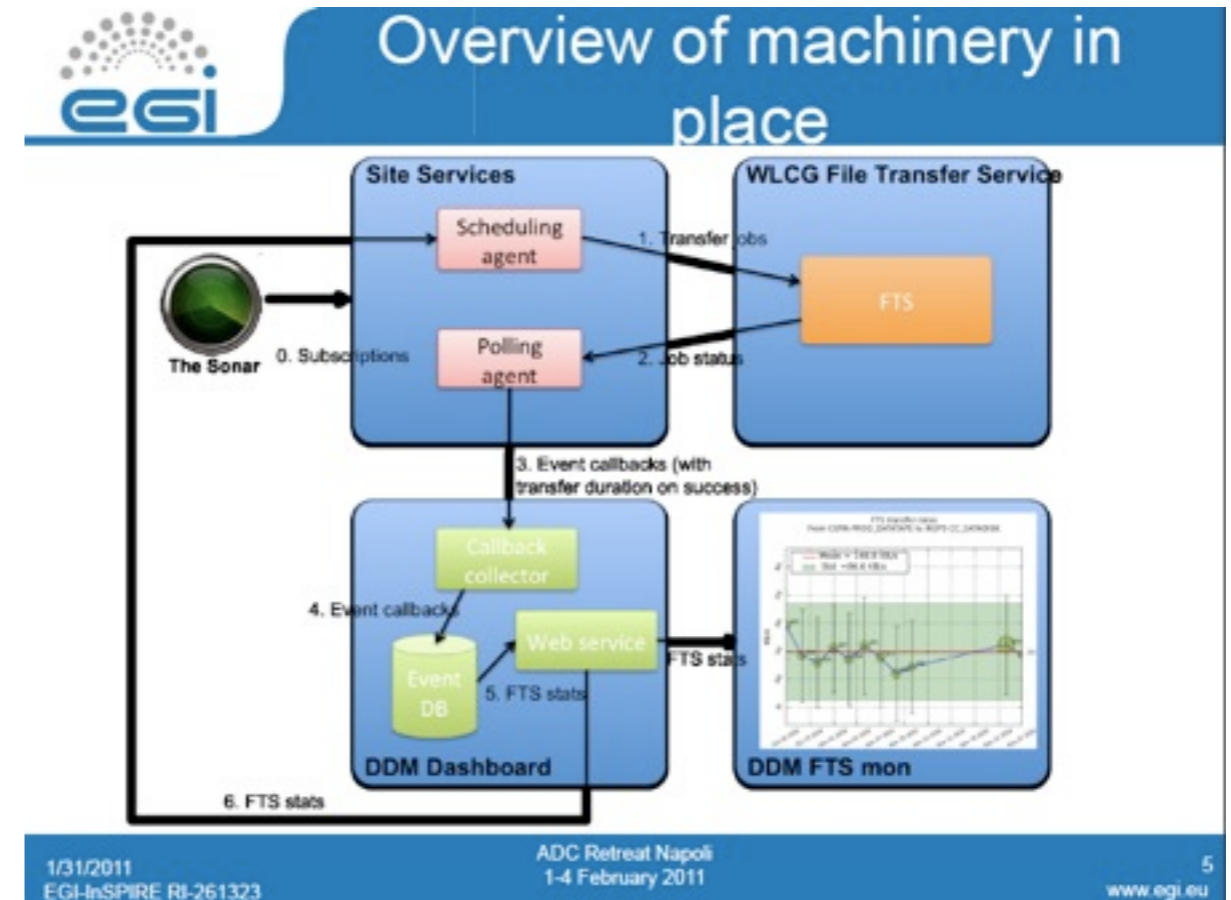


esi Sonar tests

- Triggered on Monday
- Stopped and cleaned on Thursday
- Check the Sonar results on Friday
- Cover full ATLAS Grid mesh
 - Excepting few small sites that have asked to be excluded
- Tests consists of 15 FTS transfers
 - 5 small files: 20 MB each
 - 5 medium files: 200 MB each
 - 5 large files: 2 GB each

Definitions
 Small file: 0 – 100 MB
 Medium file: 100MB – 1GB
 Large file: 1GB - ∞

EGE-InSPIRE RI-261323 ADC Retreat Napoli 1-4 February 2011 www.esi.eu



F. Barreiro 02-Feb-2011 | Napoli <http://indico.cern.ch/contributionDisplay.py?contribId=6&sessionId=14&confId=111538>

Transfer statistics monitoring

Average transfer rate per file (Small, Medium, Large)

- ▶ <http://bourricot.cern.ch/dq2/ftsmon/>
- ▶ the “Sonar table” is a quick client-side implementation
 - slow and limited in functions
- Integration into SSB on-going
 - ▶ <http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview?view=Sonar>

Only DATADISK to DATADISK transfers are shown (Period: 2011-03-11 - 2011-04-01)

| Show <input type="text" value="25"/> entries | | Search all columns: <input type="text"/> | | | | | | | | | | | | | | |
|--|-------------------|--|-------------|---------|---------------|---------------|------|---------------|---------------|-----|-------------|---------|-----|--|--|--|
| Prio | Source | SCloud | Destination | DCloud | SMALL FILES | | | MEDIUM FILES | | | LARGE FILES | | | | | |
| | | | | | MB/s | MB | #Ev | MB/s | MB | #Ev | MB/s | GB | #Ev | | | |
| 9 | INFN-T1 | IT - T1 | BNL-OSG2 | US - T1 | 0.03+ 0.05 | 1.34+ 2.11 | 1190 | 3.3+2.15 | 200.0+ 0.0 | 15 | 28.73+4.3 | 2.0+0.0 | 15 | | | |
| 7 | INFN-T1 | IT - T1 | SLACXRD | US - T2 | 0.57+ 0.03 | 20.0+0.0 | 15 | 4.29+ 0.41 | 200.0+ 0.0 | 15 | 13.9+3.27 | 2.0+0.0 | 15 | | | |
| 7 | INFN-T1 | IT - T1 | MWT2 | US - T2 | 0.47+ 0.13 | 20.0+0.0 | 15 | 5.62+ 1.33 | 200.0+ 0.0 | 15 | - | - | - | | | |
| 7 | INFN-T1 | IT - T1 | AGLT2 | US - T2 | 0.53+ 0.15 | 20.0+0.0 | 15 | 3.47+ 0.67 | 200.0+ 0.0 | 15 | 11.5+6.95 | 2.0+0.0 | 15 | | | |
| 7 | INFN-NAPOLI-ATLAS | IT - T2 | BNL-OSG2 | US - T1 | 0.5+ 0.25 | 20.0+0.0 | 15 | 1.64+0.6 | 200.0+ 0.0 | 15 | 6.51+7.65 | 2.0+0.0 | 15 | | | |
| 7 | INFN-ROMA1 | IT - T2 | BNL-OSG2 | US - T1 | 0.4+ 0.22 | 20.0+0.0 | 15 | 1.48+ 0.69 | 200.0+ 0.0 | 15 | 3.8+1.06 | 2.0+0.0 | 15 | | | |
| 4 | INFN-T1 | IT - T1 | NET2 | US - T2 | 0.7+ 0.15 | 20.0+0.0 | 15 | 3.72+0.9 | 200.0+ 0.0 | 15 | 7.74+0.91 | 2.0+0.0 | 15 | | | |
| 4 | INFN-T1 | IT - T1 | SWT2CPB | US - T2 | 0.52+ 0.06 | 20.0+0.0 | 10 | 1.54+ 0.17 | 200.0+ 0.0 | 10 | 1.38+0.33 | 2.0+0.0 | 10 | | | |

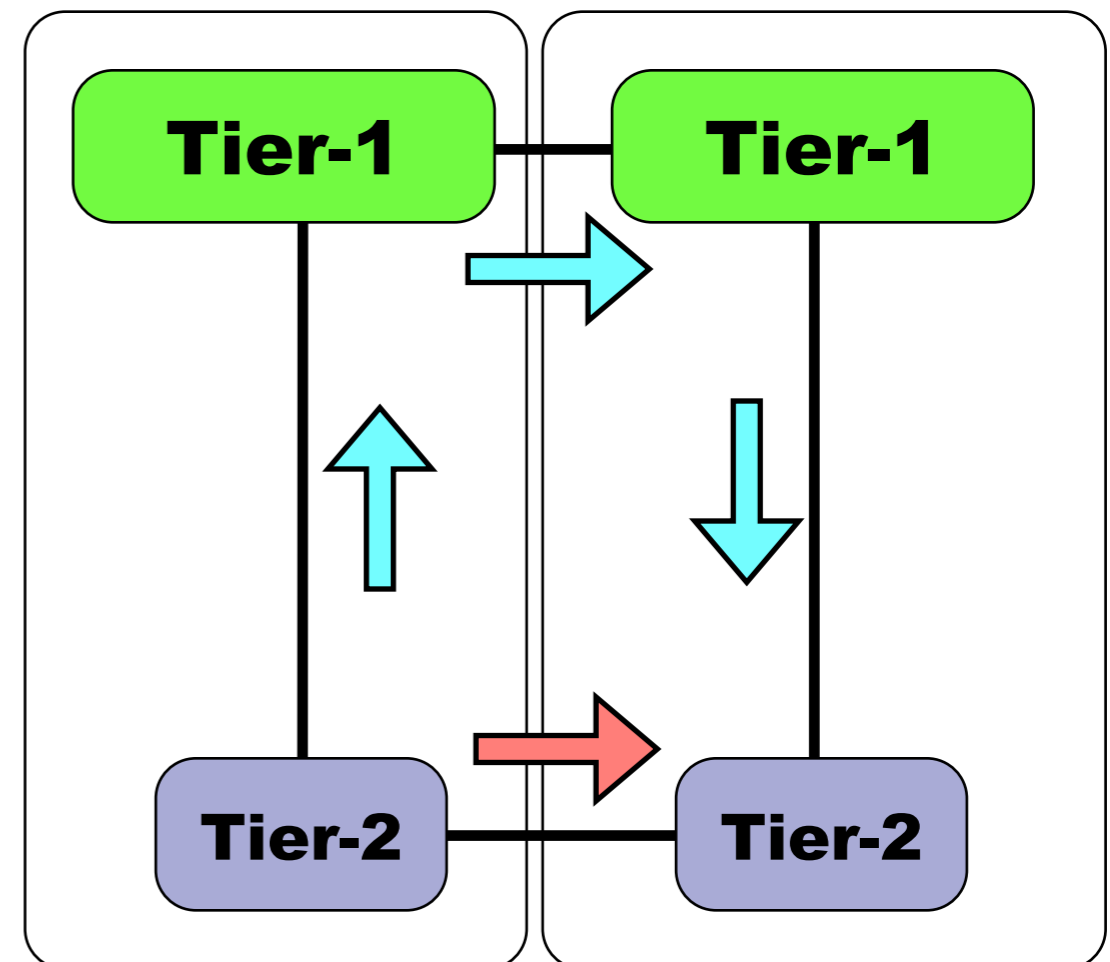
Inter-Cloud Transfers

DDM New Feature

- DDM decides the routing **based on the measurements**
 - take the route with less expected total transfer time
 - ▶ T2a-T1a-T1b-T2b (multi-hop)
 - DDM can now make ‘multi-hop’ transfers by itself
 - ▶ T2a-T2b (direct)

DaTRI now uses this new DDM feature

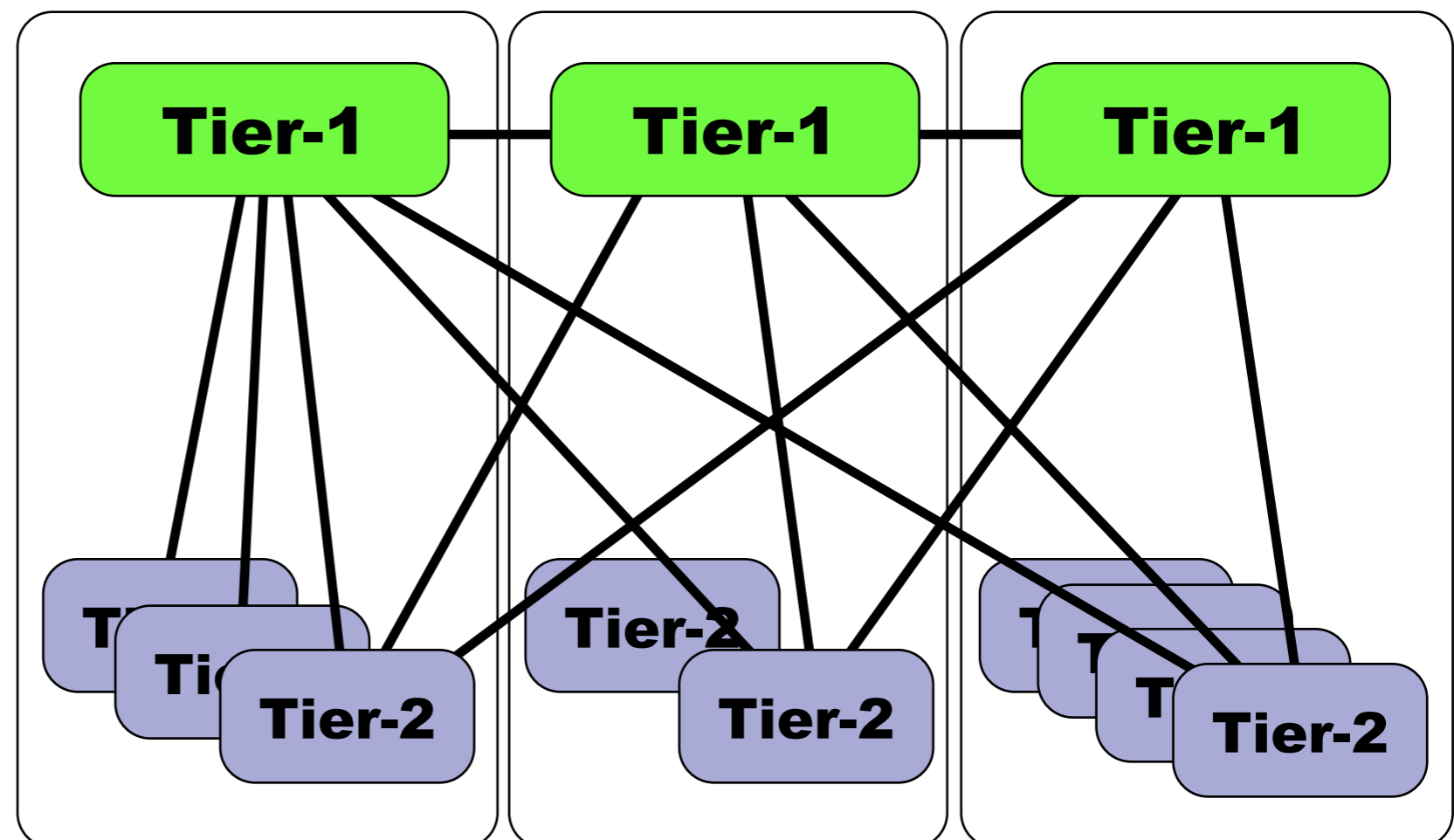
- Putting a single subscription
- Letting DDM decide the routing
- Instead of making multi-subscriptions by itself
- **US ddm-ss needs upgrade**



Inter-cloud T1-T2 transfers

T2Ds – Tier-2s that can be **directly** connected to Tier-1s outside the cloud

- We would like to have as many T2D sites as possible
- probation on-going using the sonar tests results
- T2Ds can contribute to disk space for primary data
- T2Ds can contribute to production in other clouds



See the dedicated talk in the plenary session this morning

Recent issues – Missing Input Files



panda creates temporary ‘dispatch’ datasets with input files and puts them on T2_PRODDISK

- No more ‘incoherent’ deletion such as;
 - ▶ DDM deleting after a predefined period
 - ▶ panda re-using ‘_dis’ datasets it had created for previous jobs
 - ▶ see the ddm-ops report at the last SW&C
- Panda creates ‘_dis’ datasets
 - ▶ everytime anew (no more re-using)
 - ▶ Files can belong to multiple ‘_dis’ datasets
 - for different jobs
 - ▶ Files can belong to ‘_dis’ and ‘_sub’ datasets
 - The “input” files can be a part of the “output” of the previous production step produced at the same site
 - ▶ with lifetime — automatically deleted after its expiration

Deletion of ‘overlapping’ replicas (same files in multiple datasets)

- Deletion services skips files in other ‘valid’ datasets
 - ▶ assuring input files not to be deleted while in use

Recent issues – Missing Input Files



Race condition:

- Dataset replicas are logically removed with lifetime expiration
- Files are put into **deletion queue** for LFC + SRM deletion
 - ▶ The problem happens with a **backlog** in the deletion queues
- New ‘dispatch’ datasets defined **after replica removal** and **before LFC deletion**
 - ▶ Missing input file error

DDM-dev provides a short-term solution:

- ▶ **shorten** the deletion queue depth
- ▶ **re-check** the files if they are in other ‘valid’ datasets, just before queueing
- and reconsider the workflow in the middle term

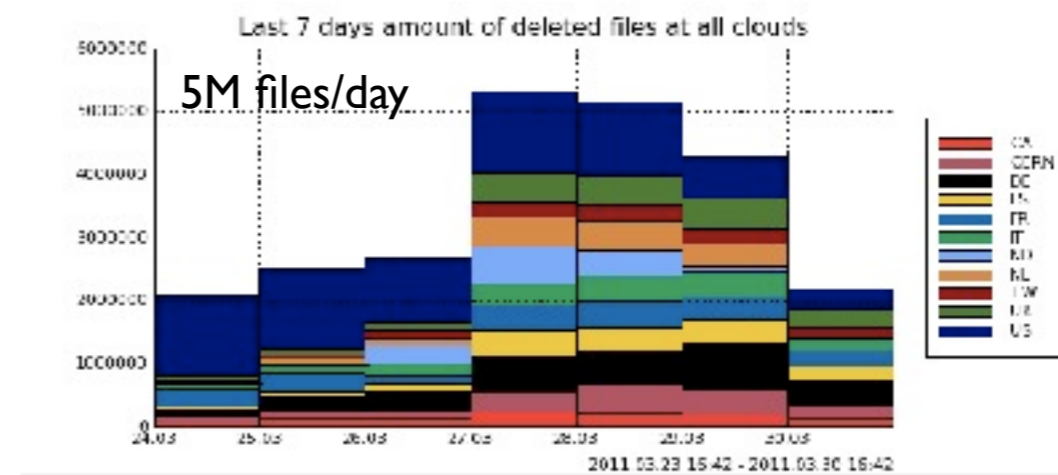
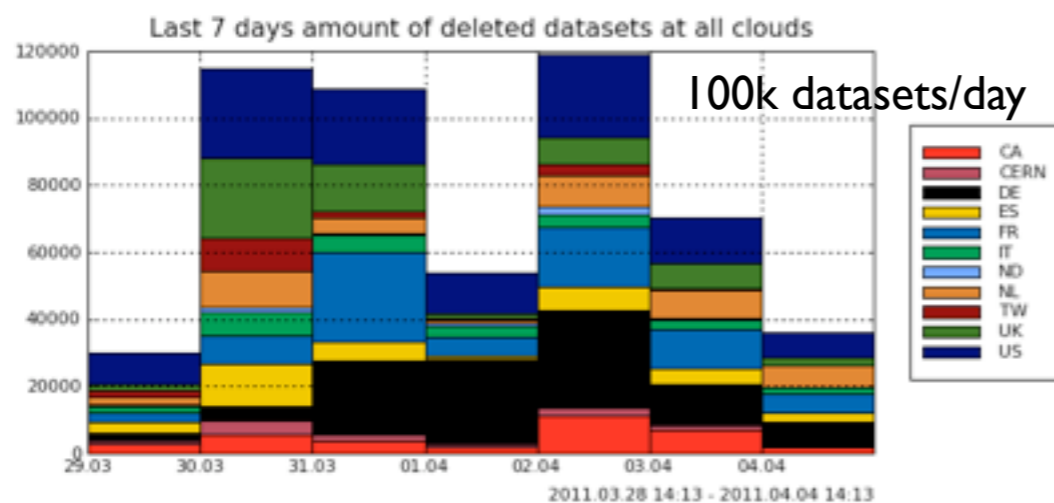
Recent issues – Deletion Backlog

Now we delete a huge number of files everyday and often observe 'backlog' in deletion

- SRM deletion for some sites
- LFC deletion for some clouds
- Central Catalog query/update

improvements have been put into the deletion service, but we are at the limit

- to be addressed in the DDM session



LFC consolidation

As has been discussed and agreed at the last SW&C, and at the Napoli workshop;

- All the “regional” LFCs are to be merged into one single “central” LFC at CERN
 - ▶ and a backup at BNL
- the study about feasibility of consolidation has started
 - ▶ with the first candidate for merge = CERN LFC + NL-T1 LFC
 - ▶ consistency check of the entries
 - ▶ merging policies for ACL and directory structure
- the second candidates NDGF-T1, TW-T1 (May)
- more to be included one-by-one during the year
- aiming to finish by the end of the year

More details

- to be discussed in DDM session
- <https://svnweb.cern.ch/trac/lcgdm/wiki/Lfc/Dev/CentralLfc>

Next Talk

The DDM coordinator for Operations (ddm-ops) this year (starting from March) is C. Serfon

- who has been working as an active ddm-ops
 - ▶ will talk about the major activity he looked after in the last months
 - ▶ will talk about the current issues to be solved in the coming months

Farewell, and bon courage