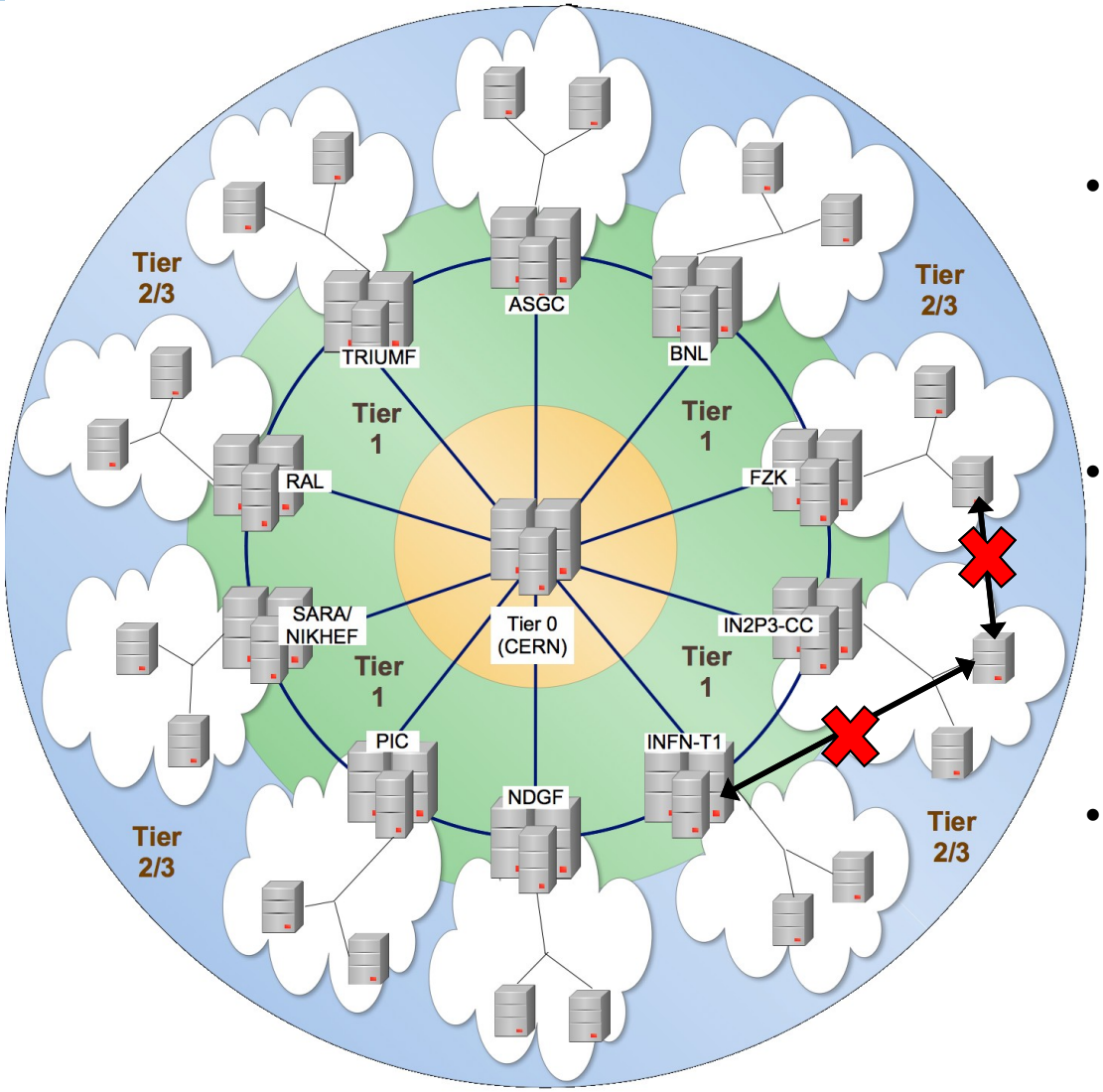
Reducing cloud boundaries in DDM Site Services

Fernando H. Barreiro Megino
CERN IT-ES-VOS

Original data distribution model



- Hierarchical tier organization according to network topology which was laid out for data distribution
- Possible communications:
  - T0-T1
  - T1-T1
  - Intra-cloud T1-T2
  - Intra-cloud T2-T2
- Restricted communications:
  - Inter-cloud T1-T2
  - Inter-cloud T2-T2

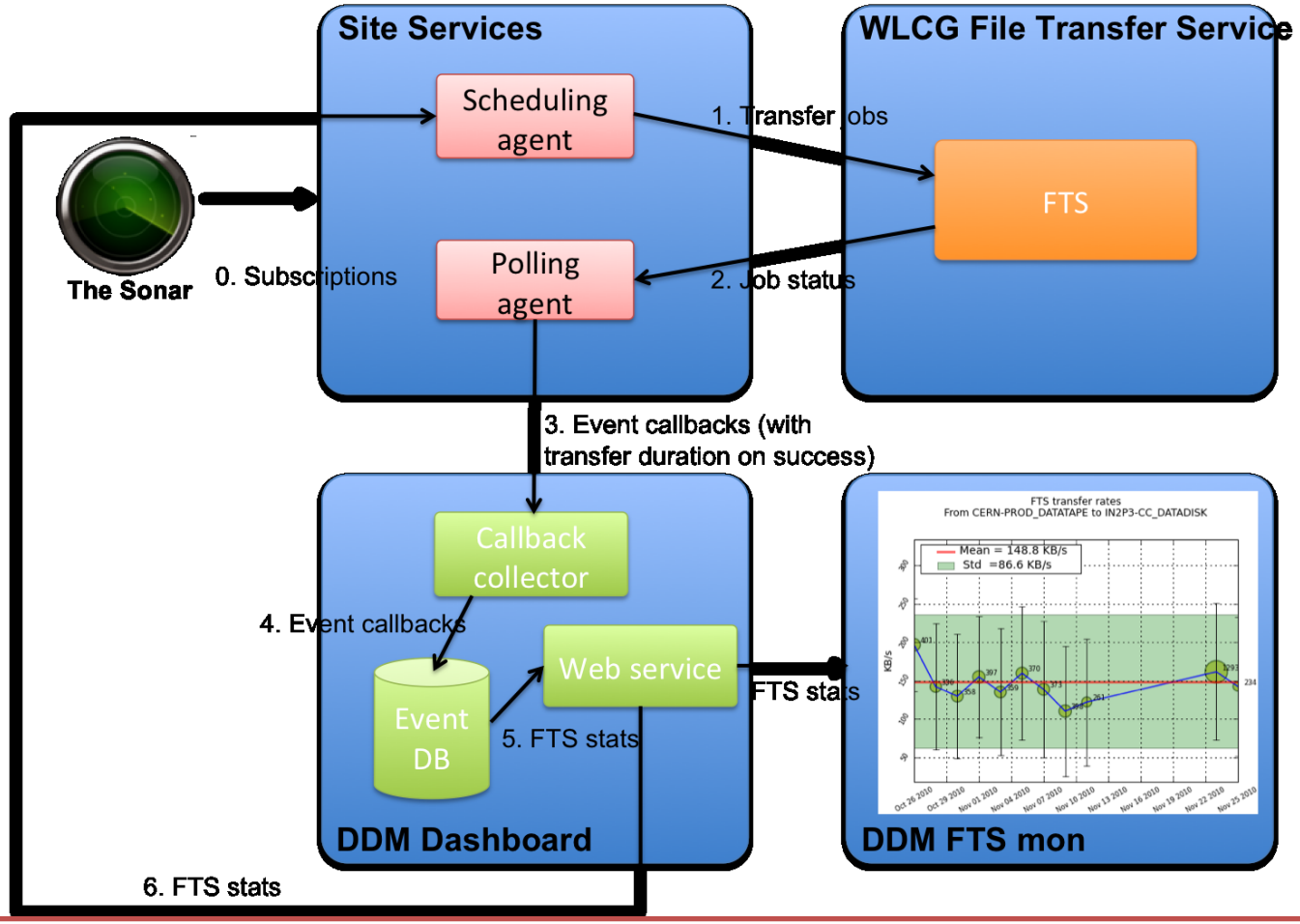
Reasons to reduce cloud boundaries

- Network bandwith has significantly increased since Computing Model was first published

- Some ADC components are currently restricted by these boundaries
  - Inflexible PD2P on one cloud
  - Consolidation of User Analysis output
  - MonteCarlo production must confine one task to one cloud

- This talk will focus on the work that has been carried out in DDM to enable cross-cloud transfers
  - Source and path optimization in data transfers
  - Tools for link commissioning: The Sonar and the FTS statistics monitor
  - Operational validation process

Machinery in place



Purpose: Generate transfer statistics for monitoring and for feedback to the system

About the FTS Statistics

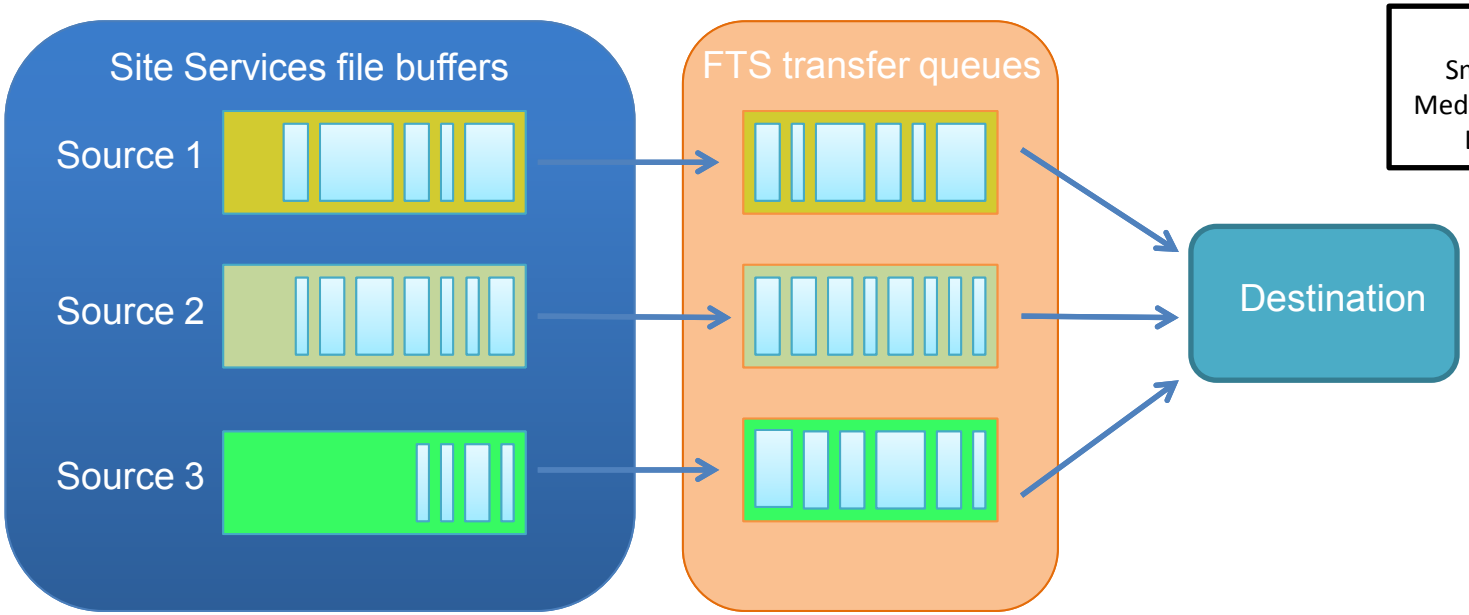
- FTS durations include SRM negotiation time
  - Transfer of small files will be dominated by SRM overhead
  - Transfer of big files will be dominated by actual transfer

➔Definition of three file-size types

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

- FTS statistics are generated by the Dashboard
  - Aggregating 10 minute bins
  - Separating by file-size types
- Available via simple API

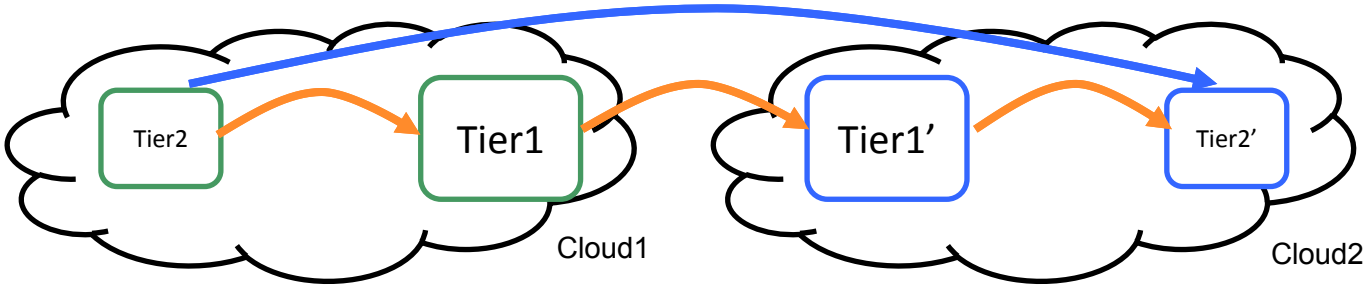
Optimized source selection



- If multiple sources are available for a file, estimate the waiting and service time of each buffer
  - Considering files of each type (S,M,L) in buffers
  - Using channel statistics of last 2 week
  - Penalizing channels without statistics
- The model is a simple approach but can be elaborated on in the future

No FTS stats means channel failed or was not tested in weekly tests

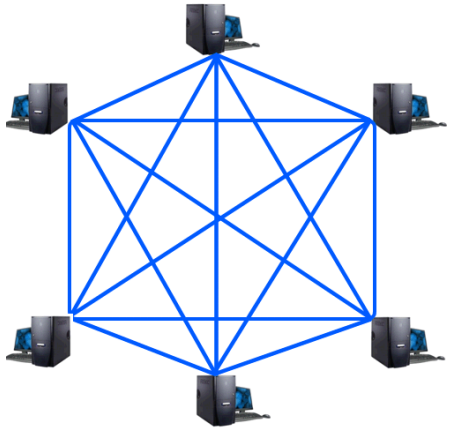
Cross-cloud transfers



- Originally cross-cloud subscriptions were submitted by DaTRI doing a multi-hop
- Now Site Services are taking care of it
  - Since mid February on Italian cloud
  - Since mid March on all clouds
- Estimate transfer time of direct transfer and multi-hop and choose most convenient one
- If no statistics available for direct link, SS will transfer by default via multi-hop
  - Probability can be tuned to be more permissive with direct transfers
- Multi-hop child subscriptions are left on T1_SCRATCHDISKs with expiration date of 15 days

Sonar tests

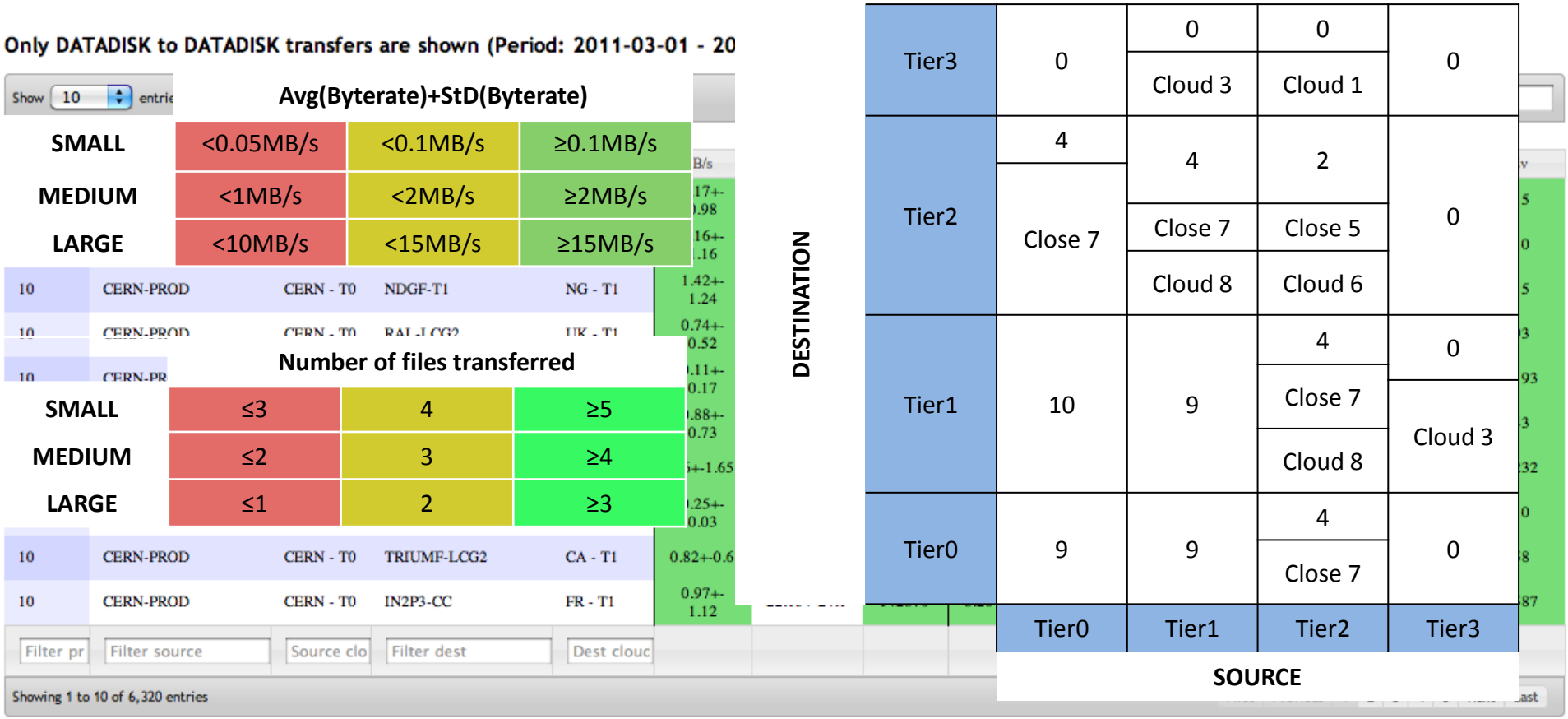
- Weekly file transfers between all ATLAS sites
- Provide recent transfer statistics for the complete Grid mesh
  - Excepting few small sites that have asked to be excluded
  - O(10.000) links
- Tests consist of 15 FTS transfers
  - 1 dataset with 5 small files: 20 MB each
  - 1 dataset with 5 medium files: 200 MB each
  - 1 dataset with 5 large files: 2 GB each
- Workflow
  - Each site owns the 3 datasets (site in name)
  - Dataset is pre-placed and stored permanently on custodial site
  - Destination replica is copied and deleted each week
- Transfers submitted gradually during the whole week
- Stopped&cleaned after 5 days



Sonar monitoring

http://bourricot.cern.ch/dq2/ftsmon/sonar_view/cached/

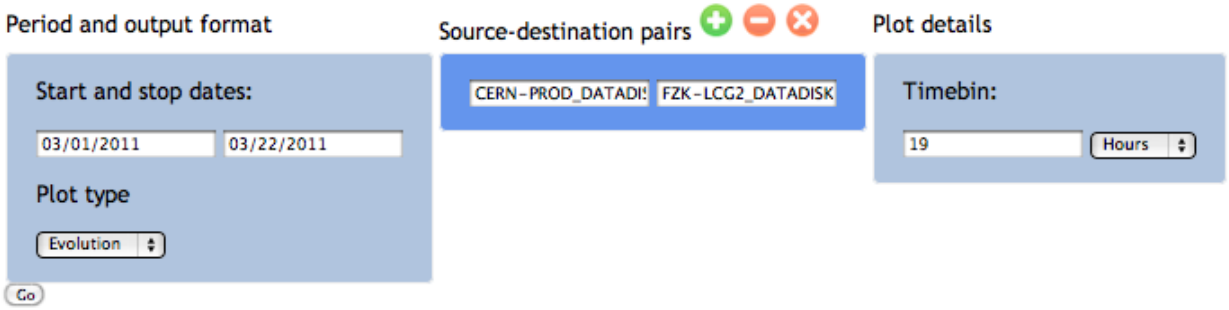
Technologies: Django, jQuery, DataTables

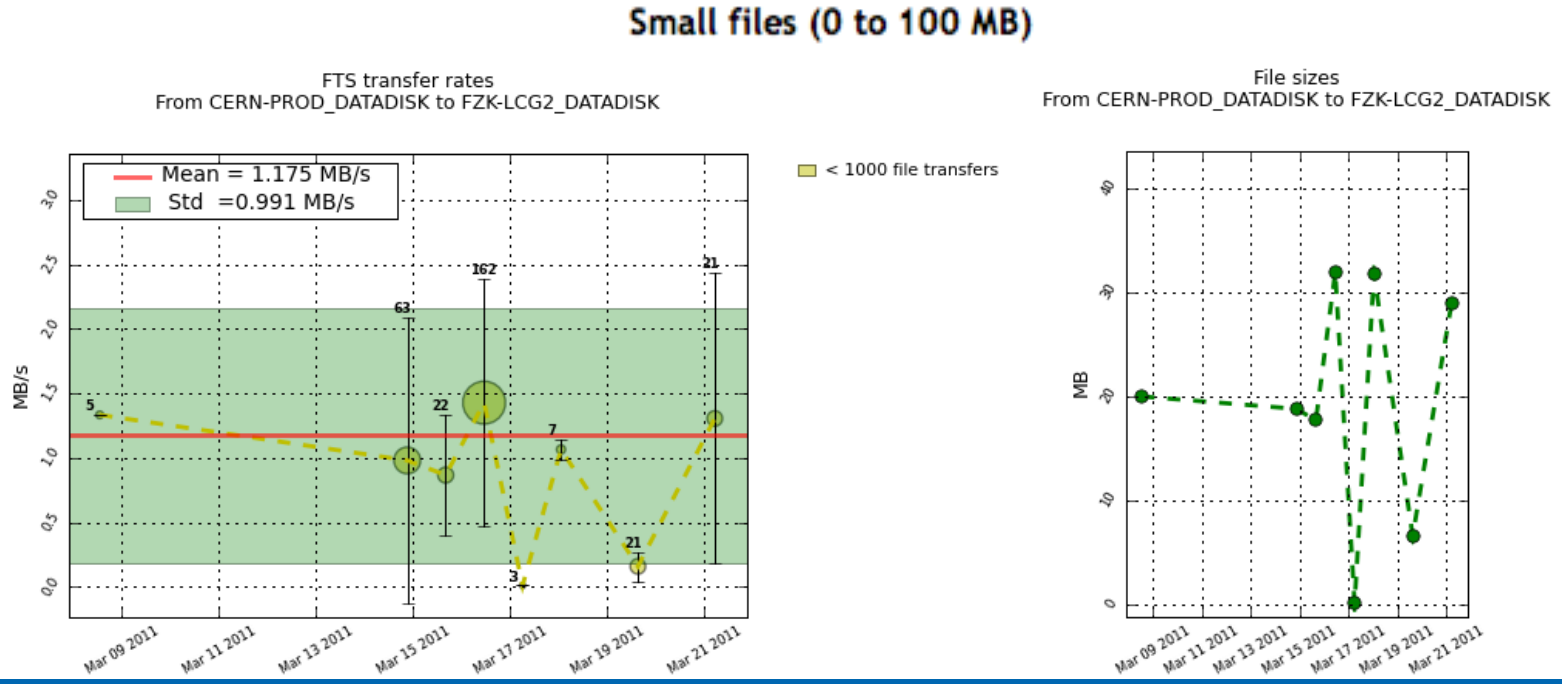


Channel history

http://bourricot.cern.ch/dq2/ftsmon/

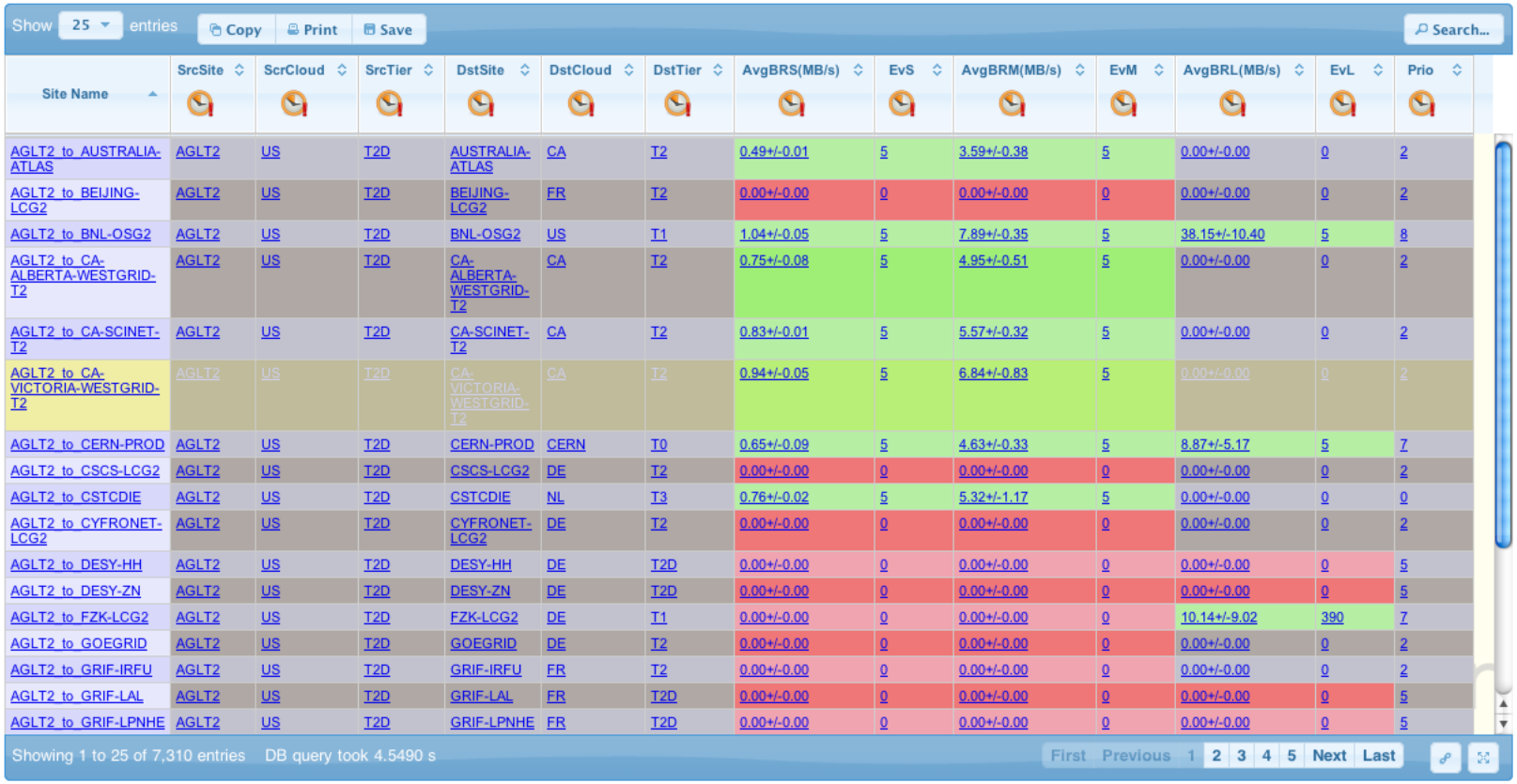
Technologies: Django, Matplotlib, jQuery





Sonar view in the Site Status Board

- http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteview?view=Sonar
- Michal Maciej and Pablo Saiz are working on implementing an improved Sonar table
  - Fix shortcomings in actual Sonar Table
  - Historic view of the table
  - Combine metrics with other information (e.g. downtime information)



Show 25 entries | Copy | Print | Save | Search...

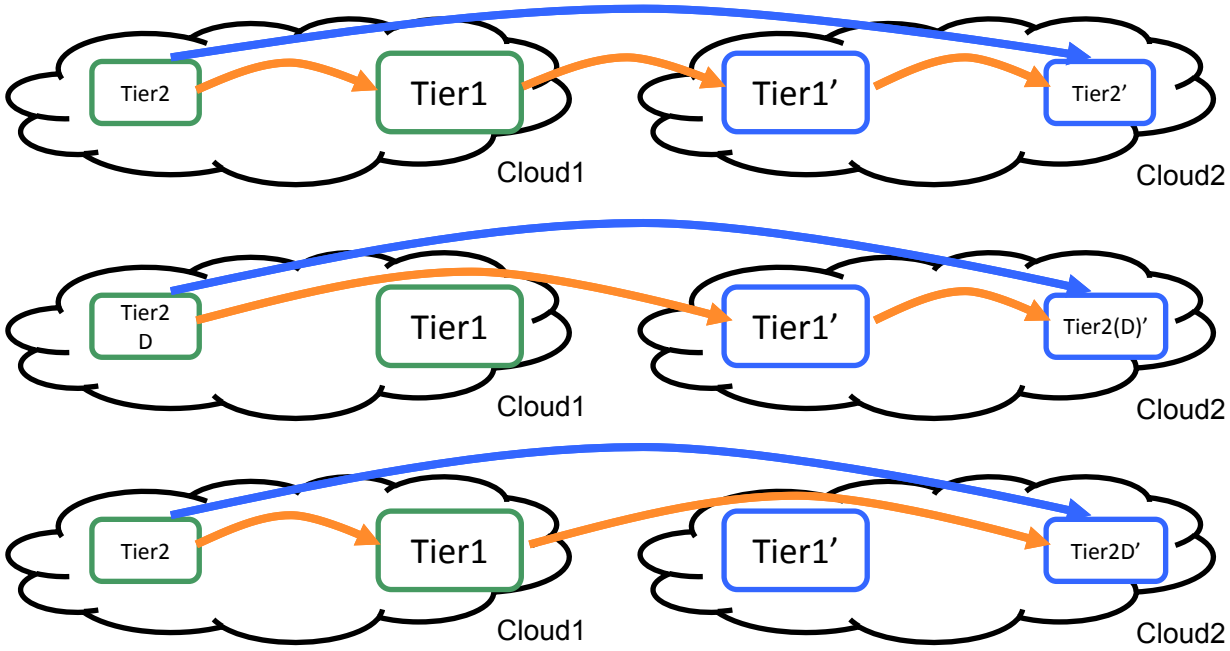
Site Name	SrcSite	ScrCloud	SrcTier	DstSite	DstCloud	DstTier	AvgBRS(MB/s)	EvS	AvgBRM(MB/s)	EvM	AvgBRL(MB/s)	EvL	Prio
AGLT2_to_AUSTRALIA-ATLAS	AGLT2	US	T2D	AUSTRALIA-ATLAS	CA	T2	0.49+/-0.01	5	3.59+/-0.38	5	0.00+/-0.00	0	2
AGLT2_to_BEIJING-LCG2	AGLT2	US	T2D	BEIJING-LCG2	FR	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
AGLT2_to_BNL-OSG2	AGLT2	US	T2D	BNL-OSG2	US	T1	1.04+/-0.05	5	7.89+/-0.35	5	38.15+/-10.40	5	8
AGLT2_to_CA-ALBERTA-WESTGRID-T2	AGLT2	US	T2D	CA-ALBERTA-WESTGRID-T2	CA	T2	0.75+/-0.08	5	4.95+/-0.51	5	0.00+/-0.00	0	2
AGLT2_to_CA-SCINET-T2	AGLT2	US	T2D	CA-SCINET-T2	CA	T2	0.83+/-0.01	5	5.57+/-0.32	5	0.00+/-0.00	0	2
AGLT2_to_CA-VICTORIA-WESTGRID-T2	AGLT2	US	T2D	CA-VICTORIA-WESTGRID-T2	CA	T2	0.94+/-0.05	5	6.84+/-0.83	5	0.00+/-0.00	0	2
AGLT2_to_CERN-PROD	AGLT2	US	T2D	CERN-PROD	CERN	T0	0.65+/-0.09	5	4.63+/-0.33	5	8.87+/-5.17	5	7
AGLT2_to_CSCS-LCG2	AGLT2	US	T2D	CSCS-LCG2	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
AGLT2_to_CSTCDIE	AGLT2	US	T2D	CSTCDIE	NL	T3	0.76+/-0.02	5	5.32+/-1.17	5	0.00+/-0.00	0	0
AGLT2_to_CYFRONET-LCG2	AGLT2	US	T2D	CYFRONET-LCG2	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
AGLT2_to_DESY-HH	AGLT2	US	T2D	DESY-HH	DE	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
AGLT2_to_DESY-ZN	AGLT2	US	T2D	DESY-ZN	DE	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
AGLT2_to_FZK-LCG2	AGLT2	US	T2D	FZK-LCG2	DE	T1	0.00+/-0.00	0	0.00+/-0.00	0	10.14+/-9.02	390	7
AGLT2_to_GOEGRID	AGLT2	US	T2D	GOEGRID	DE	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
AGLT2_to_GRIF-IRFU	AGLT2	US	T2D	GRIF-IRFU	FR	T2	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	2
AGLT2_to_GRIF-LAL	AGLT2	US	T2D	GRIF-LAL	FR	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5
AGLT2_to_GRIF-LPNHE	AGLT2	US	T2D	GRIF-LPNHE	FR	T2D	0.00+/-0.00	0	0.00+/-0.00	0	0.00+/-0.00	0	5

Showing 1 to 25 of 7,310 entries DB query took 4.5490 s | First Previous 1 2 3 4 5 Next Last

Commissioning work

- Monitoring of the links: Check Sonar test results
  - Responsibility of AMOD, shifters and cloud squads
- Create the “T2Ds” category: Directly connected T2s
  - T1s and T2Ds are topologically close in Tiers of ATLAS
  - Reduce cloud boundaries in a controlled fashion
  - Initial selection: 22 T2s in 6 clouds
  - Selection based on
    - Storage and network capabilities
    - Commitment and reliability
  - FTS service providers were asked to create 2 new channels to connect T2Ds to all T1s
    - Foreign T2Ds to local T1
    - Foreign T1s to local T2Ds
  - Monitor T1-T2D links (priority 7) actively to verify they don’t degrade

Multi-hopping to T2Ds



FTS STAR-Channel transfer
Multihop

- Site Service multi-hop already adapted to new T2D category
- Skip one hop (usually on the T2D cloud)
  - Reduce intermediate left-overs
  - Improve subscription times

Other news about Site Services

- ActiveMQ integration
  - Site Services send many different callbacks to the DDM Dashboard:
    - Dataset content
    - File events (Transferring, copied, registered, failed…)
    - Subscription events (Queued, completed, canceled, broken…)

    And, on request, a subset to analysis tools (e.g. PanDA)
  - Site Services are ready to publish callbacks to ActiveMQ
    - Feature implemented and validated in testbed
    - Not activated in production instances yet
  - Message queues will allow to send once and listen by all interested parties

Other news about Site Services (2)

- Reduced source selection policies in order to reduce the load on TAPE sites
- FTS overwrite option is being used now
  - Avoids generating infamous __DQ2-<timestamp> files
  - MWT2_UC and on Functional Test machine
- DDM is being very unlucky with the hardware lately
  - SS spare machine is being rotated frequently because of disk problems
  - However SS migrations to spare take 5 minutes now

Credits

- Simone Campana
- David Tuckett
- Andrii Thykonov
- Stephane Jezequel
- I. Ueda
- Vincent Garonne
- Martin Barisits