Distributed data management on the petabyte-scale with DQ2

Mario Lassnig
on behalf of ATLAS DDM

3rd EGEE User Forum
Clermont-Ferrand, France

CERN, Switzerland
University of Innsbruck, Austria

Outline

- **System overview**

  - *Concepts and principles*

- **Architecture**

  - *Interconnecting grids*
  - *Implementation details*

- **Conclusions**

  “*The throughput peaked at 200MB/s for 2 hours at the end of the exercise, our largest average daily rate was over 90MB/s.*”

  -CHEP2006

Scope

- **Responsibilities of ATLAS Distributed Data Management**
  - bookkeeping of all ATLAS file-based experiment and user data
  - managing movement of data across sites and for endusers
  - enforcing access control and quotas

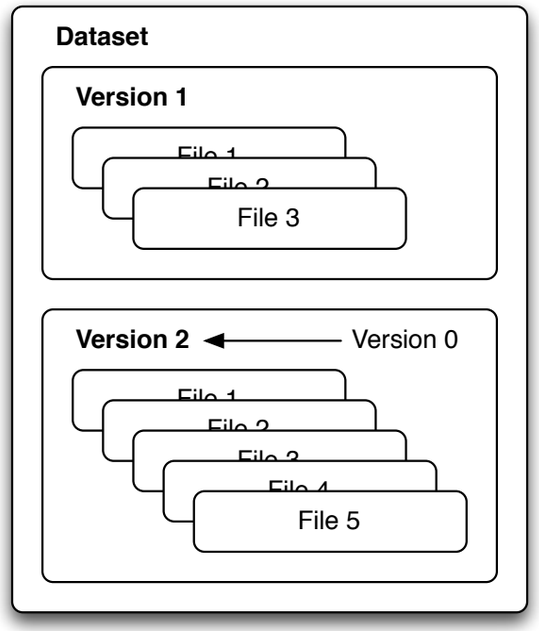
- **Objective of ATLAS Distributed Data Management**
  - manage the ATLAS dataflow
  - according to the ATLAS Computing Model
  - with a single entry point to all distributed data

The ATLAS Experiment

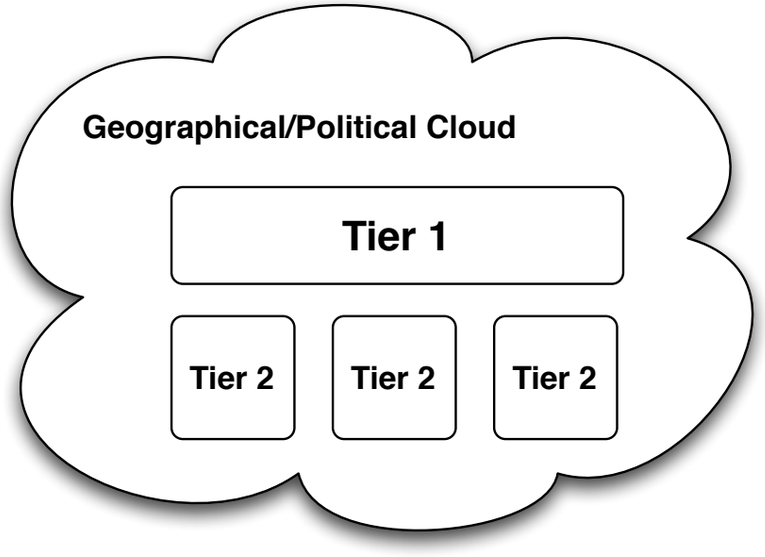
- **2000 users**
- **200 sites**, distributed globally in different grid infrastructures
- **75.000 computing jobs** per day
- **27.000 GB of new data** per day
- **10 PB of new data** per year
- **23 million files** already with **60 million replicas**
- **and increasing...**

Basic Concepts

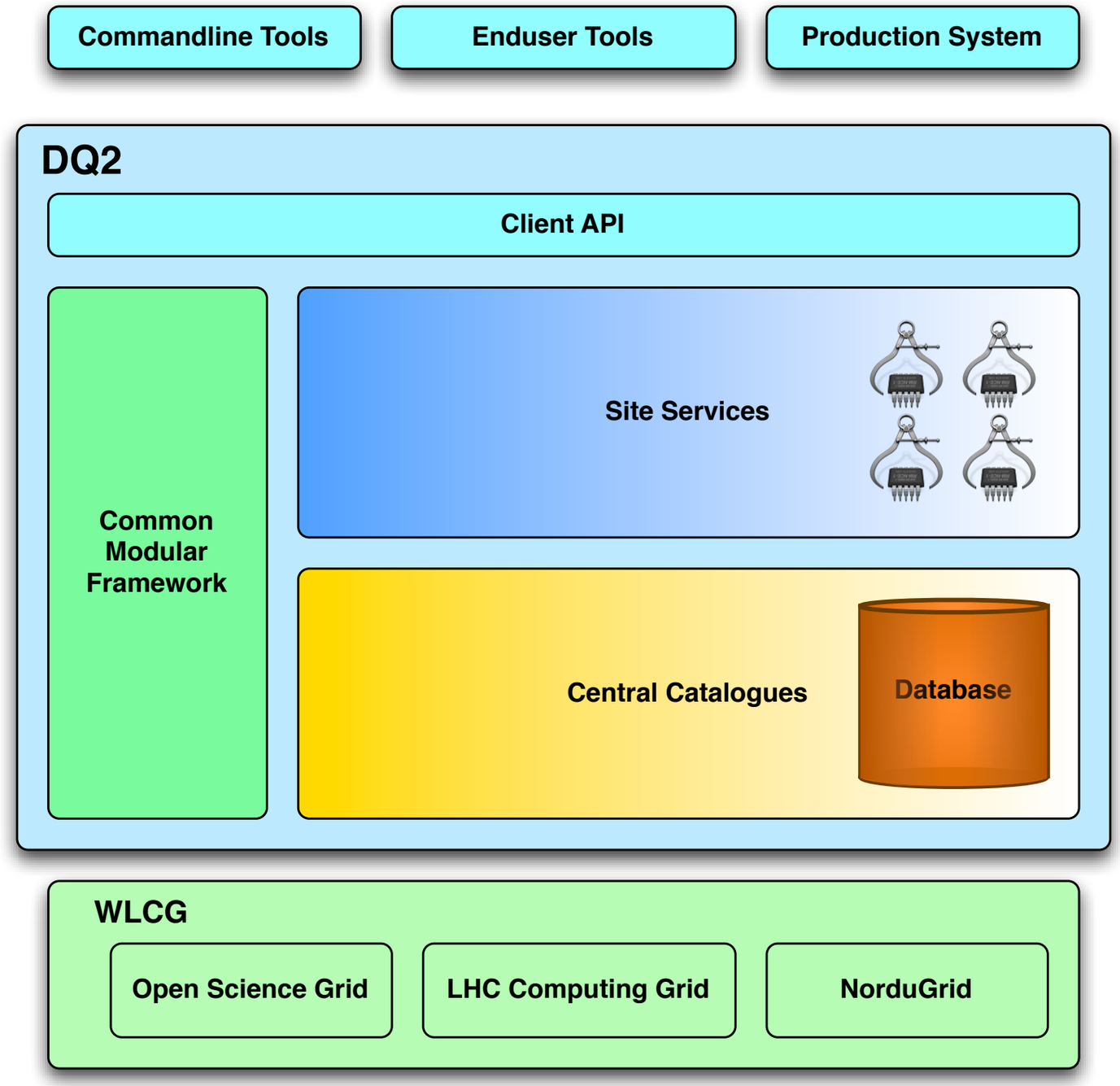
- **Experiment data**
  - Files are grouped in datasets
  - Datasets provide metadata for grouped files
  - Extended features (versions, immutability, overlapping)
  - Cloud-based architecture provides sites organised in hierarchical Tiers



- **Transferring data**
  - Pull-based Subscription model
  - Sites subscribe to a dataset (version)
  - Local site service agents satisfy the subscriptions



System Architecture Overview



- **Common Framework**
  - Python
  - EGEE, OSG, NG

- **Client API**
  - UI Tools
  - Data Acquisition

- **Local Site Services**
  - Autonomous agents

- **Central Catalogues**
  - Locations
  - Subscriptions
  - Repositories
  - Contents

Interconnecting datagrids

- **gLite File Transfer Service (FTS v1, v2)**

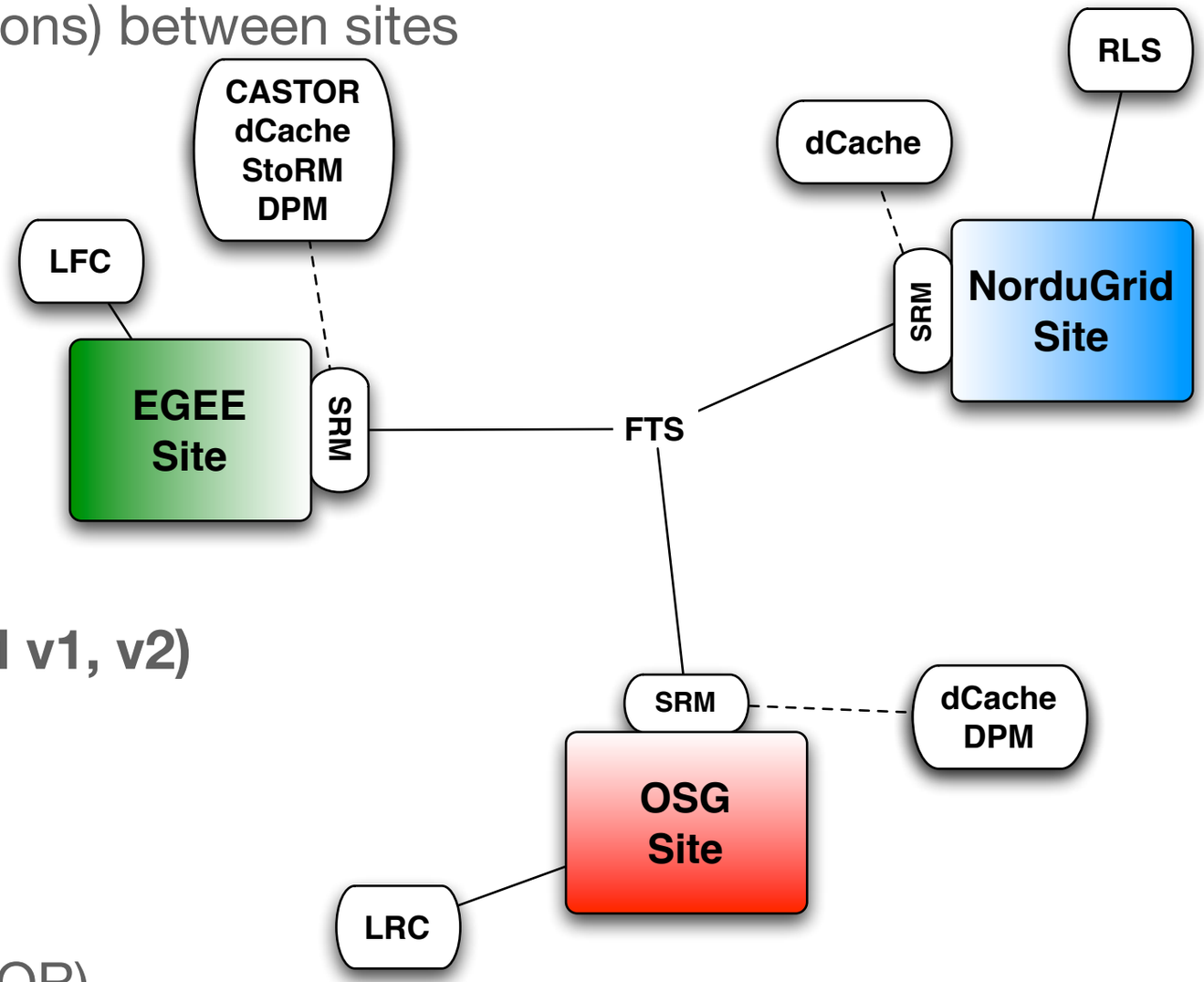
  - directed data transfer (subscriptions) between sites

- **File catalogues**

  - gLite File Catalogue
  - Open Science Grid LRC
  - NorduGrid RLS

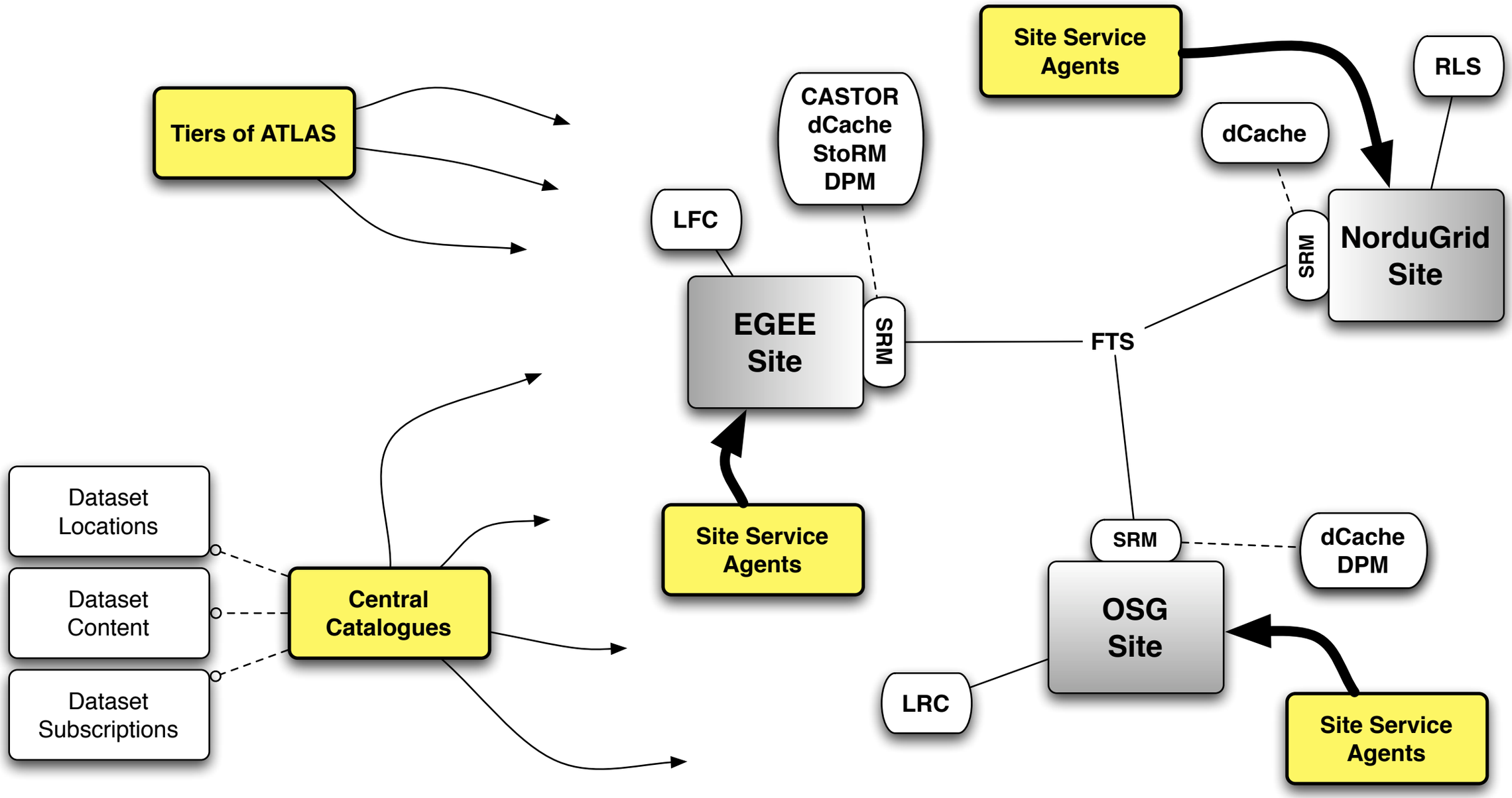
- **Storage Resource Manager (SRM v1, v2)**

  - Disk Pool Manager (DPM)
  - dCache
  - StoRM
  - CERN Advanced Storage (CASTOR)



Interconnecting datagrids

- **DQ2 is like the force - it binds the universe together**



Site Services

- **Parallel queue-based architecture**

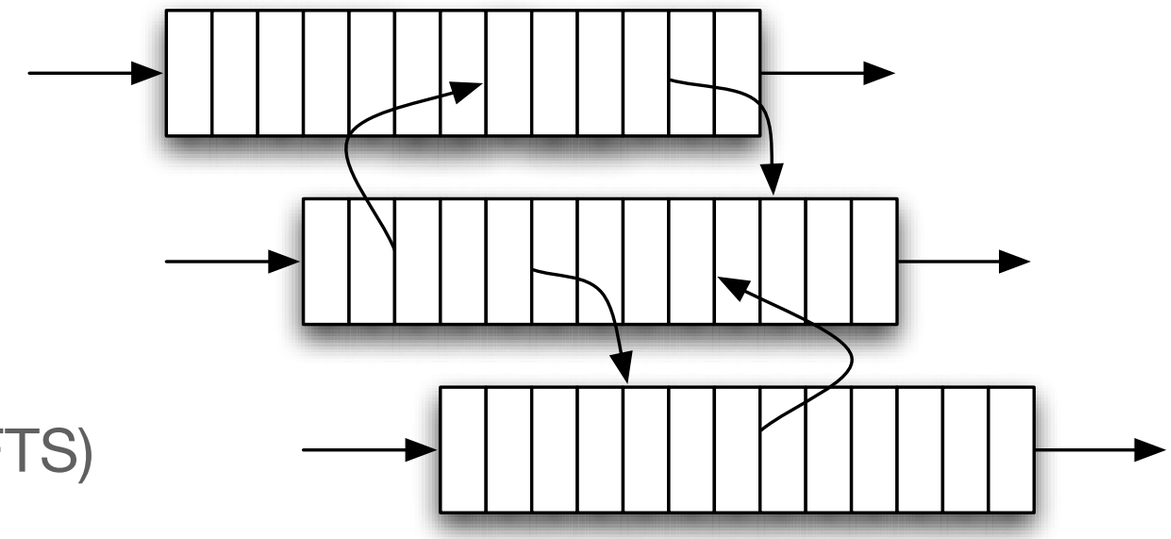
Queues of subscriptions (of datasets)

Queues of files (in a dataset)

Queues of file replicas (to lookup)

Queues for FTS channels (in front of FTS)

Queue of files and datasets to register (in central catalogues)



- **Non-hierarchical Fairshare Algorithm**

  - Divides available slots according to file transfers (not filesize) into shares

  - Measure state of active transfers against queue for FTS channels

How to use DQ2?

- **DQ2 is non-intrusive - just setup the python libraries and grid dependencies**

- **Datasets directly usable by GANGA and PanDA**

- **Job/Enduser-tools can directly work with datasets**

- **dq2-ls**
  - list datasets
  - query all properties of datasets

- **dq2-get**
  - copy datasets to your local system

- **dq2-put**
  - make files accessible to DQ2

How to really use DQ2

```
$ dq2-ls *mlassnig*
dq2-ls 1.14
tutorial.dps.mlassnig.001
user.mlassnig.dataset.1
user.mlassnig.td4
user.mlassnig.test.file.1


$ dq2-ls -r user.mlassnig.dataset.1
dq2-ls 1.14

user.mlassnig.dataset.1
 INCOMPLETE:
   AGLT2
 COMPLETE:
   ASGCDISK_V2
   TRIUMFDISK
   BNLDISK
   PICDISK
   CNAFDISK
   RALDISK
   NDGFT1DISK
   LYONDISK
   FZKDISK
   SARADISK
   NDGFT1TAPE
   BNLTAPE
   TOKYO
   GLASGOW
   CYF
   UVIC
   HEPHY-UIBK
   BU
   CERNPROD
```

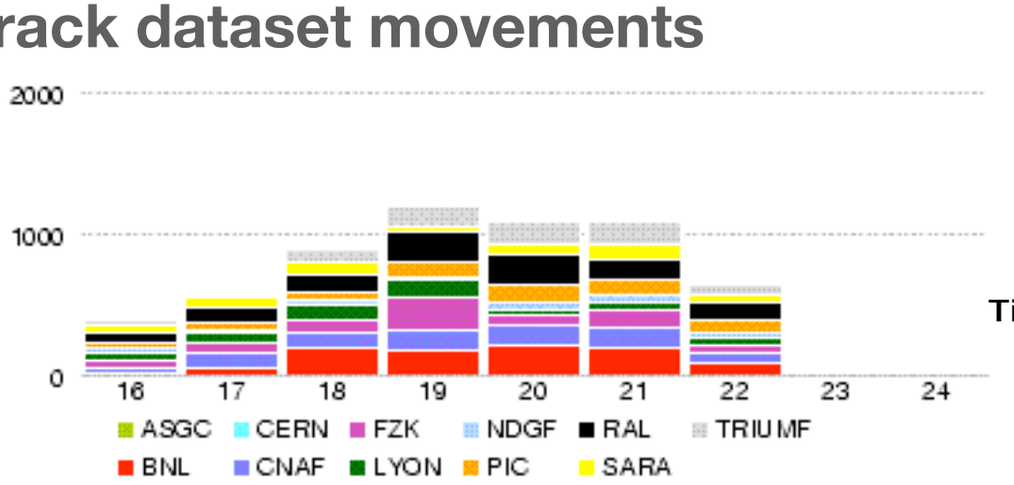
```
$ dq2-ls -f user.mlassnig.dataset.1
dq2-ls 1.14

user.mlassnig.dataset.1
[X]     dummyfile1      35aeb84a-aeef-41f4-bd2f-616972f4cd69    md5:bf7700bd815231d79ec96613f35e175c    52428800
[X]     dummyfile2      0c752981-0432-48e6-84b4-54baa245a61a    md5:6b2c90fa80e7160b1ff615c606796acb    52428800
[X]     dummyfile3      b712f157-e5da-4946-983e-536a97473652    md5:8e46d1a0ebd2de342df654e18737f849    52428800
total files: 3
local files: 3
total size: 157286400
date: 2007-10-11 12:23:39

$ dq2-get user.mlassnig.dataset.1
dq2-get 1.12
Querying DQ2 central catalogues to resolve datasetname user.mlassnig.dataset.1
Datasets found: 1
user.mlassnig.dataset.1: Querying DQ2 central catalogues for replicas...
Querying DQ2 central catalogues for files in dataset...
user.mlassnig.dataset.1: Trying site CERNPROD
user.mlassnig.dataset.1: Checking availability of transfer tools
user.mlassnig.dataset.1: Querying local file catalogue of site CERNPROD...
user.mlassnig.dataset.1/dummyfile3: Getting srm metadata for srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user/user.mlassnig.
user.mlassnig.dataset.1/dummyfile2: Getting srm metadata for srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user/user.mlassnig.
user.mlassnig.dataset.1/dummyfile1: Getting srm metadata for srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user/user.mlassnig.
user.mlassnig.dataset.1/dummyfile2: is cached at source.
user.mlassnig.dataset.1/dummyfile2: Starting transfer: lcg-cp -v --vo atlas srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user
user.mlassnig.dataset.1/dummyfile3: is cached at source.
user.mlassnig.dataset.1/dummyfile3: Starting transfer: lcg-cp -v --vo atlas srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user
user.mlassnig.dataset.1/dummyfile1: is cached at source.
user.mlassnig.dataset.1/dummyfile1: Starting transfer: lcg-cp -v --vo atlas srm://srm.cern.ch:8443/castor/cern.ch/grid/atlas/dq2/user
user.mlassnig.dataset.1/dummyfile2: 0/52428800 transferred
user.mlassnig.dataset.1/dummyfile3: 0/52428800 transferred
user.mlassnig.dataset.1/dummyfile1: 0/52428800 transferred
user.mlassnig.dataset.1/dummyfile2: 1048576/52428800 transferred
user.mlassnig.dataset.1/dummyfile2: 27262976/52428800 transferred
user.mlassnig.dataset.1/dummyfile3: 18874368/52428800 transferred
user.mlassnig.dataset.1/dummyfile1: 25165824/52428800 transferred
user.mlassnig.dataset.1/dummyfile2: 52428800/52428800 transferred
user.mlassnig.dataset.1/dummyfile3: 52428800/52428800 transferred
user.mlassnig.dataset.1/dummyfile1: 52428800/52428800 transferred
user.mlassnig.dataset.1/dummyfile2: validated
user.mlassnig.dataset.1/dummyfile1: validated
user.mlassnig.dataset.1/dummyfile3: validated
```

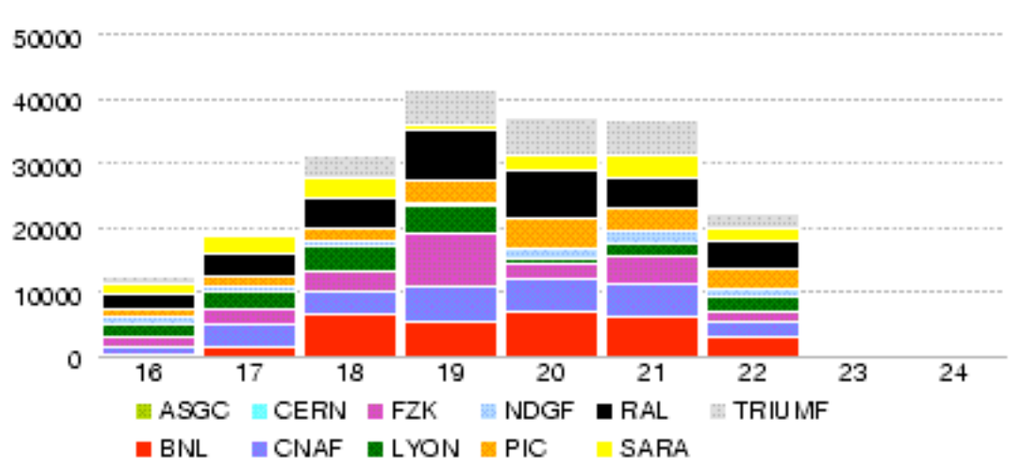
Monitoring the DDM

- **Experiment dashboard by the ARDA team**
- **Decoupled infrastructure based on event callbacks**
- **Track dataset movements**



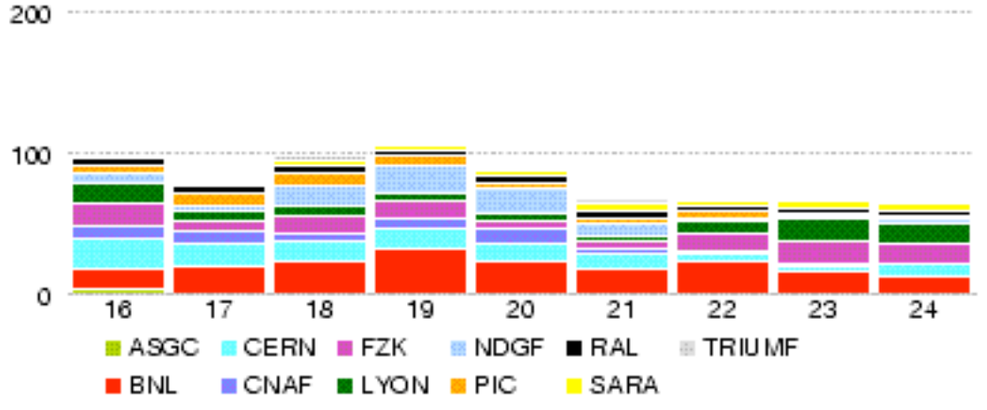
Throughput MB/sec

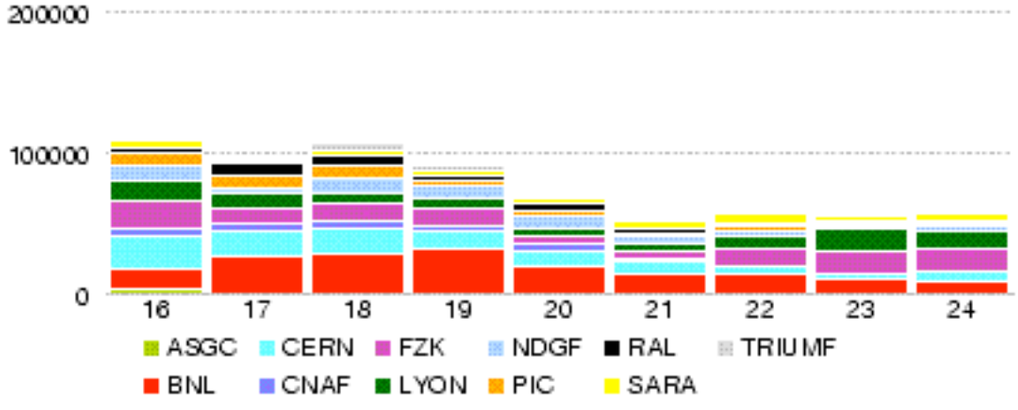
Tier0 Export



Completed file transfers

Production





Issues

- **Deletion of data will become a problem very soon**
  - not only because of overlapping datasets...

- **Better integration with user-applications**
  - full API available, yet still possible for “clever” users to DoS
  - tighter bindings to PanDA and GANGA

- **Better end-user “experience”**
  - virtual file system?
  - we need some feedback on that...

- **Consolidation of grid reliability features**
  - what do the errors really mean?
  - e.g. are errors from NorduGrid related to errors in OSG or EGEE?
  - no common error “interface” except SRM/FTS response

Conclusions and future prospects

- **Successfully use three different grid infrastructures**

- **Performance improvements**
  - peaks increased by factor 6 to over 1200 MB/sec
  - largest average daily increased by factor 8 to over 700 MB/sec
  - number of files transferred increased by factor 2 to over 100.000 files daily

- **Operational improvements**

- **Gathering usage information for future system improvements**

Ready for full-scale LHC data-taking!

Distributed data management on the petabyte-scale with DQ2

Mario Lassnig
on behalf of ATLAS DDM

3rd EGEE User Forum
Clermont-Ferrand, France

CERN, Switzerland
University of Innsbruck, Austria