The ATLAS Computing Model

Alessandro De Salvo <Alessandro.DeSalvo@roma1.infn.it>
12-4-2007

Outline

- **ATLAS facilities**
- **Analysis model**
- **Distributed production and analysis tools**

LHC computing and ATLAS

- **ATLAS is one of the 4 LHC experiments and will start its operation with real data in 2007/2008**

- **LHC will provide $4 \bullet 10^7$ collisions/s in each experiment, which is reduced to ~100 events/s after the filtering**
  - **With an average event size of 1 MB this leads to a recording rate of 100 MB/s**
  - **Considering a year of data taking (10^{10} collisions) the overall storage capacity needed is about 10 PB/year**

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Computing resources

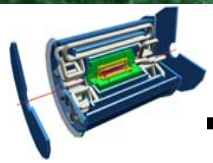
- **The ATLAS Computing model has been documented in 2005 in the Computing TDR**
  - http://doc.cern.ch//archive/electronic/cern/preprints/lhcc/public/lhcc-2005-022.pdf

- **Some components are still evolving and will not be final for some time**
  - **Calibration and alignment strategy**
  - **Physics data access patterns**
    - Could be exercised from June 2007
    - Unlikely to know the real patterns until 2007/2008!
  - **Still uncertainties on the event sizes , reconstruction time**

Data types in ATLAS

- **RAW**
  - The data coming out of the trigger is called bytestream data
  - The bytestream format is encapsulated in a C++ object called RDO (Raw Data Object)
  - An RDO object is about 1.6 MB in size
- **ESD**
  - The ESD (Event Summary Data) is a more compact format, derived from the RDO format
  - An ESD event is about 0.5 MB in size
- **AOD**
  - The AOD (Analysis Object Data) format is extracted from the ESD data and is specifically targetted to user analysis
  - An AOD event is about 10 kB in size
- **TAG**
  - TAGs are databases created to be used as an event-level metadata system
  - Supporting a fast selection of events of interest of a given analysis
  - A TAG event is about 0.1 kB in size

The ATLAS Computing Model



~Pb/sec

Event Builder

10 GB/sec

Event Filter
~159kSI2k

450 Mb/sec

- Some data for calibration and monitoring to institutes
- Calibrations flow back

Tier 0

T0 ~5MSI2k

MSS

- Calibration
- First processing

~ 300MB/s/T1 /expt

Tier 1

US Regional Centre

UK Regional Centre (RAL)

Spanish Regional Centre (PIC)

Italian Regional Centre (CNAF)

MSS

- Reprocessing
- Group analysis

≥622Mb/s

Tier 2

Tier2 Centre ~200kSI2k

Tier2 Centre ~200kSI2k

Centre ~200kSI2k

Centre ~200kSI2k

- Analysis
- Simulation

≥622Mb/s

Institute 1

titute 2

stitute 3

stitute 4

Physics data cache

100 - 1000 MB/s

Workstations

Desktop

Average Tier 2 has ~25 physicists working on one or more channels

Roughly 3 Tier 2s should have the full AOD, TAG & relevant Physics Group summary data

Tier 2 do bulk of simulation

Data flow

- **Event Filter farm → Tier-0**
  - 450 Mb/s, continuous
- **Tier-0**
  - Raw data → CERN MSS
  - Raw data → Tier-1s
- **Tier-0 → Tier-1s**
  - ESD, AOD, TAG
  - 2 copies of the full ESD set, distributed worldwide
- **Tier-0 → Tier-2**
  - Calibration data streams
- **Tier-1 → Tier-2**
  - A subset of raw data/ESD
  - Full copy of AOD and TAG
  - User/group datasets
- **Tier-2 → Tier-1**
  - MC raw, ESD, AOD, TAG
- **Tier-2 → Tier-0**
  - Calibration data processing

Network bandwidth

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

ATLAS HI
Heavy Ions data

ATLAS
pp data

Network bandwidth T0 → T1

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Network bandwidth T1 → T2

Facilities at CERN

- **Tier-0:**
  - **Prompt first pass processing on express/calibration & physics streams with old calibrations - calibration, monitoring**
  - **Calibrations tasks on prompt data**
  - **24-48 hours later, process full physics data streams with reasonable calibrations**

    → Implies large data movement from T0 →T1s

- **CERN Analysis Facility**
  - **Access to ESD and RAW/calibration data on demand**
  - **Essential for early calibration**
  - **Detector optimisation/algorithmic development**

Facilities outside CERN

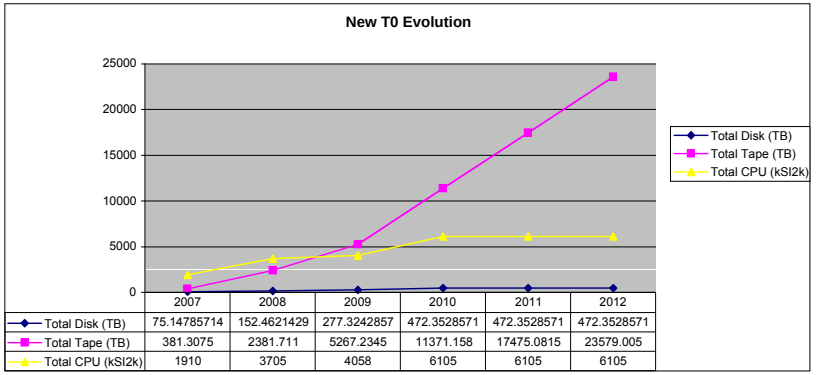
- **Tier-1**
  - **Reprocess 1-2 months after arrival with better calibrations**
  - **Reprocess all resident RAW at year end with improved calibration and software**
    → Implies large data movement from T1↔T1 and T1 → T2
- **~30 Tier 2 Centers distributed worldwide**

  Monte Carlo Simulation, producing ESD, AOD, ESD, AOD ➔ Tier 1 centers
  - **On demand user physics analysis of shared datasets**
  - **Limited access to ESD and RAW data sets**
  - **Simulation**
    → Implies ESD, AOD, ESD, AOD ➔ Tier 1 centers
- **Tier 3 Centers distributed worldwide**
  - **Physics analysis**
  - **Data private and local - summary datasets**

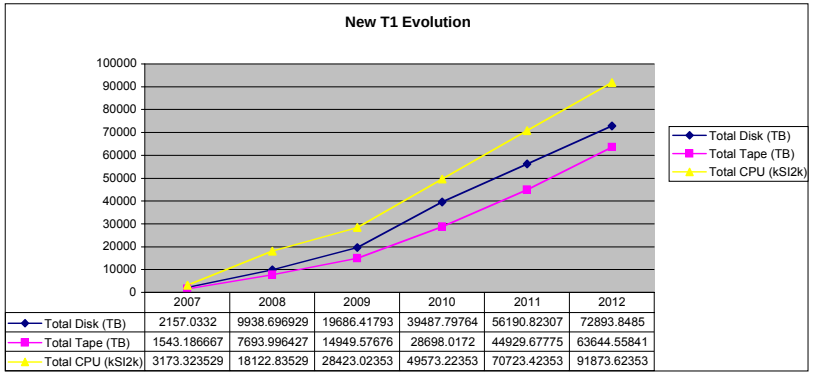
ATLAS Computing requirements (2008-2010)

	CPU (MSi2k)		Disk (PB)		Tape (PB)	
	2008	2010	2008	2010	2008	2010
Tier-0	3.7	6.1	0.15	0.5	2.4	11.4
CERN Analysis Facility	2.1	4.6	1.0	2.8	0.4	1.0
Sum of Tier-1s	18.1	50	10	40	7.7	28.7
Sum of Tier-2s	17.5	51.5	7.7	22.1		
Total	41.4	112.2	18.9	65.4	10.5	41.1

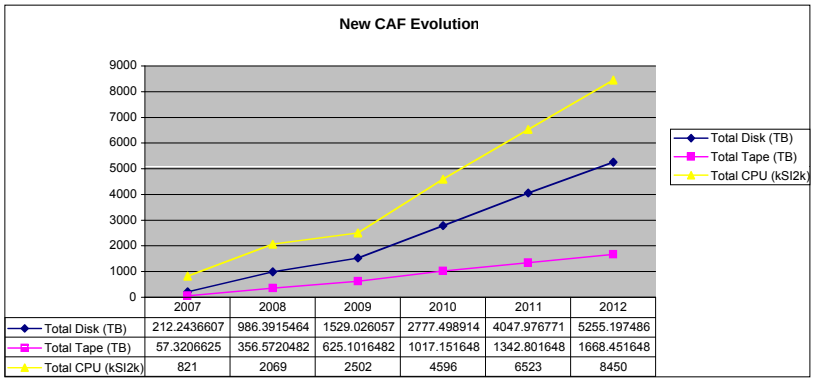
Resource evolution



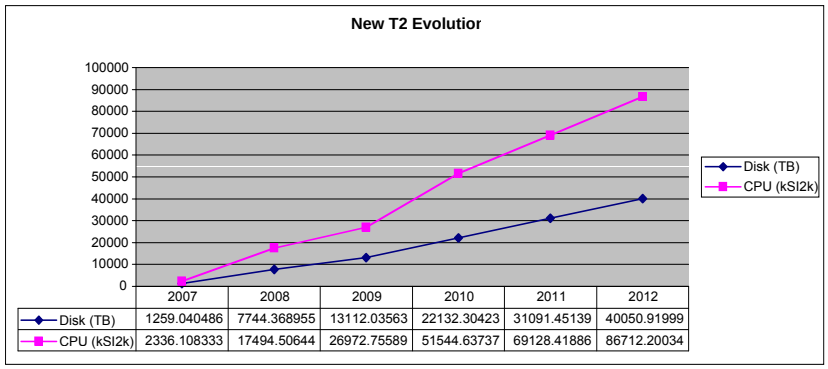
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	75.14785714	152.4621429	277.3242857	472.3528571	472.3528571	472.3528571
Total Tape (TB)	381.3075	2381.711	5267.2345	11371.158	17475.0815	23579.005
Total CPU (kSI2k)	1910	3705	4058	6105	6105	6105



	2007	2008	2009	2010	2011	2012
Total Disk (TB)	2157.0332	9938.696929	19686.41793	39487.79764	56190.82307	72893.8485
Total Tape (TB)	1543.186667	7693.996427	14949.57676	28698.0172	44929.67775	63644.55841
Total CPU (kSI2k)	3173.323529	18122.83529	28423.02353	49573.22353	70723.42353	91873.62353



	2007	2008	2009	2010	2011	2012
Total Disk (TB)	212.2436607	986.3915464	1529.026057	2777.498914	4047.976771	5255.197486
Total Tape (TB)	57.3206625	356.5720482	625.1016482	1017.151648	1342.801648	1668.451648
Total CPU (kSI2k)	821	2069	2502	4596	6523	8450

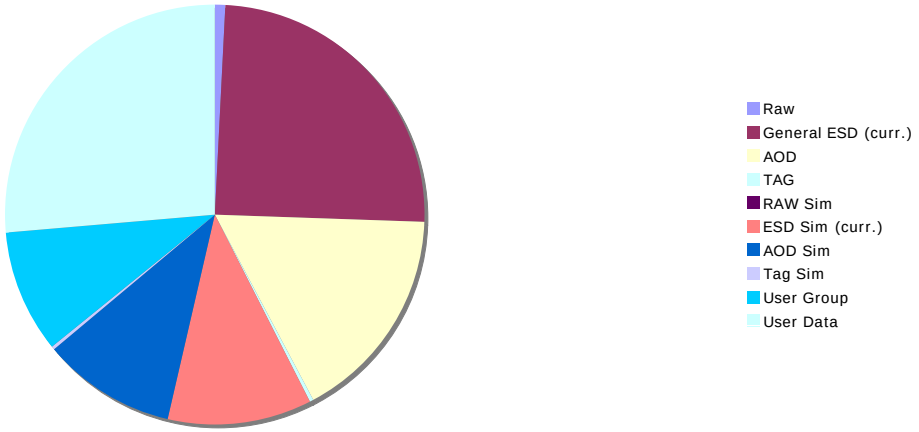


	2007	2008	2009	2010	2011	2012
Disk (TB)	1259.040486	7744.368955	13112.03563	22132.30423	31091.45139	40050.91999
CPU (kSI2k)	2336.108333	17494.50644	26972.75589	51544.63737	69128.41886	86712.20034

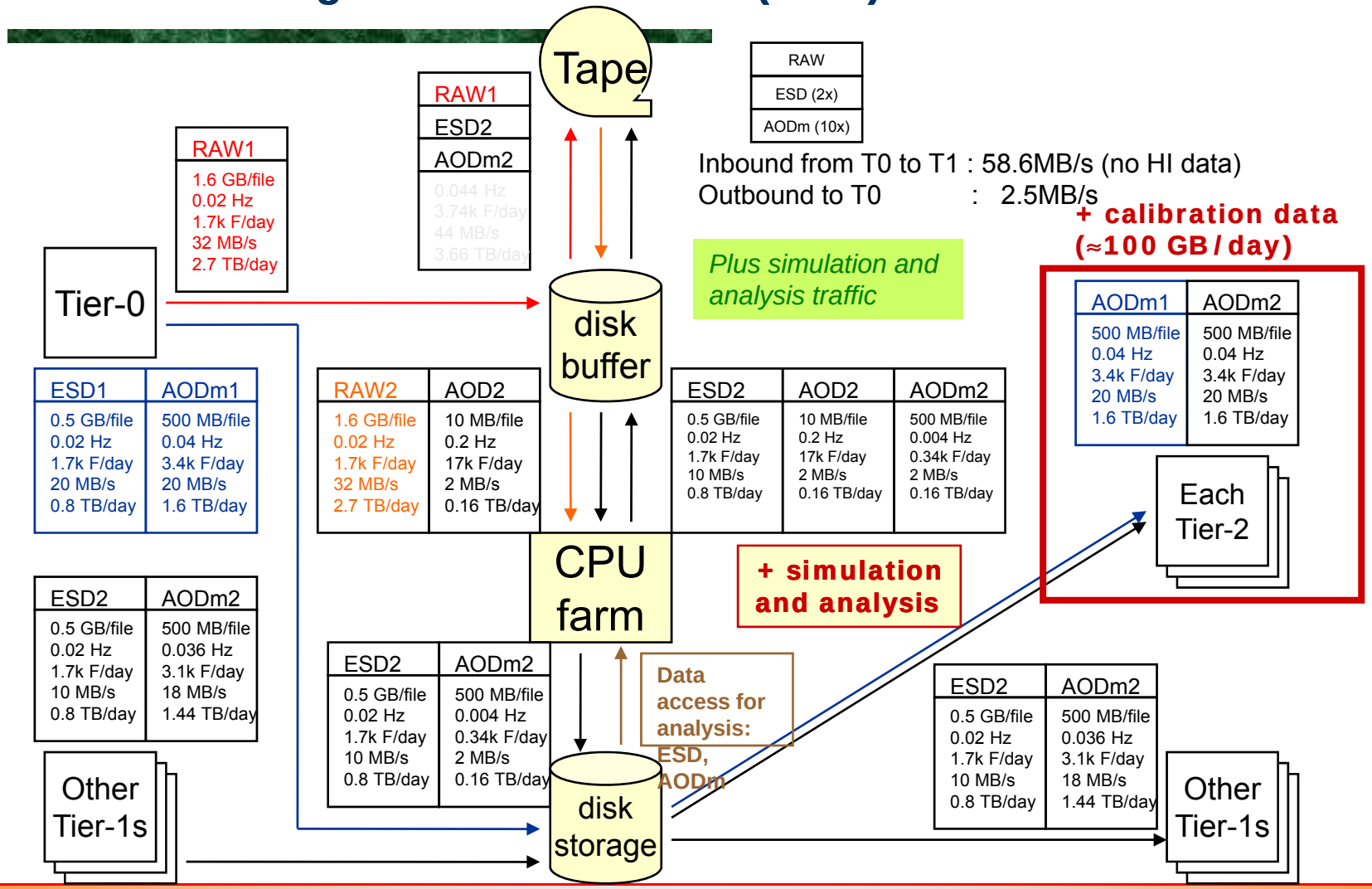
T1/T2 Disk shares (2008)

Tier-1 (2008)

Tier-2 (2008)

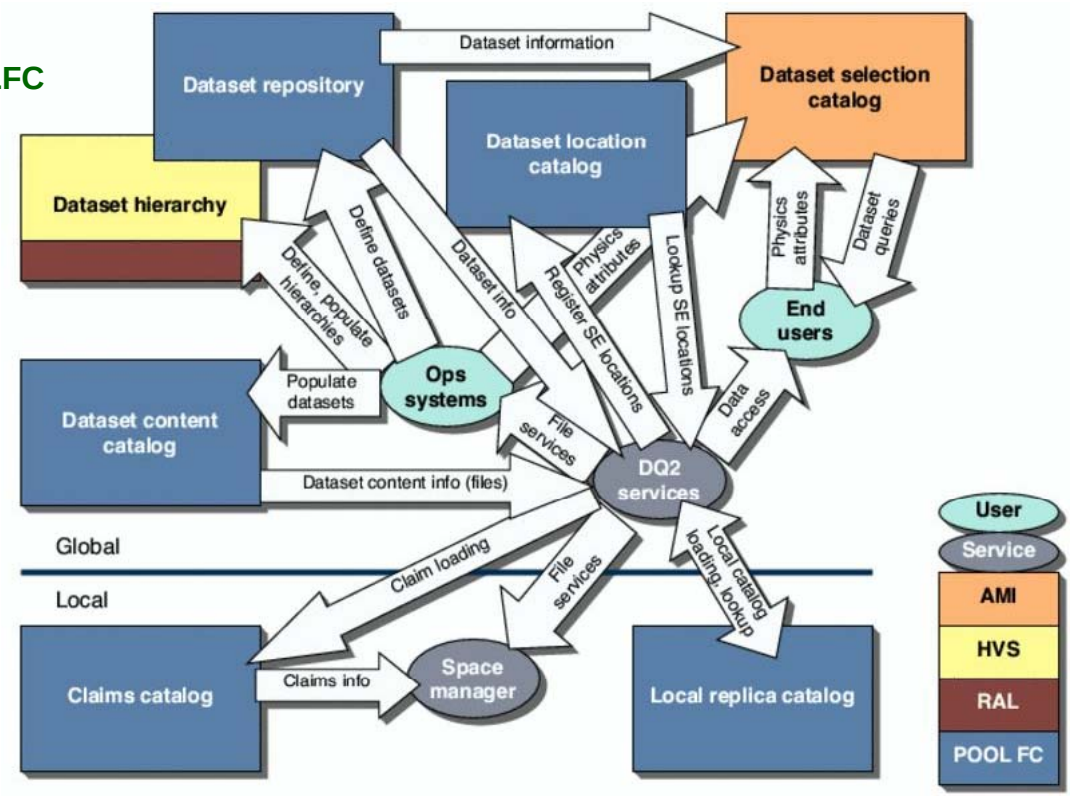


ATLAS "average" Tier-1 Data Flow (2008)

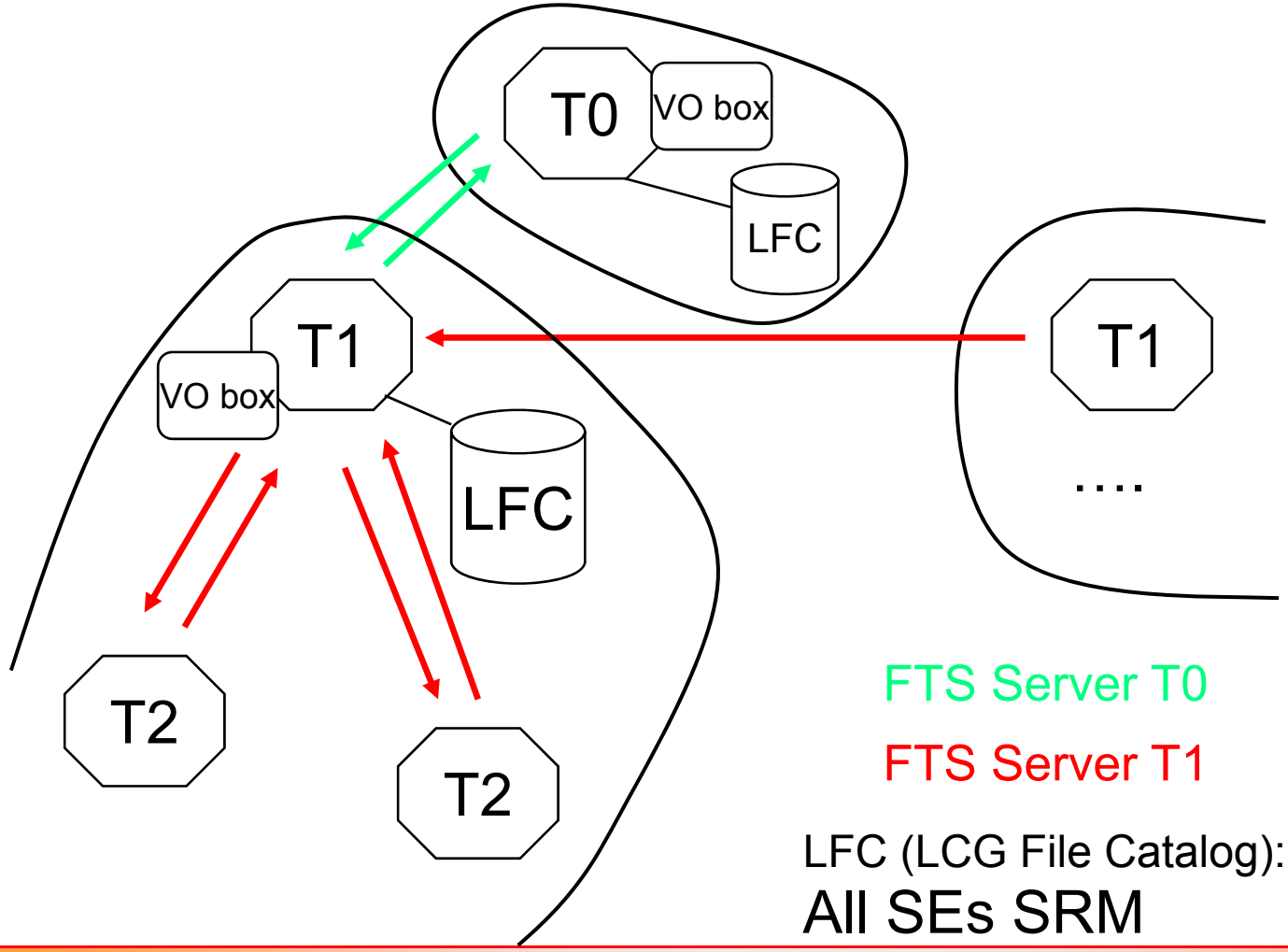


ATLAS Distributed Data Management (DQ2)

- **Files are aggregated in datasets**
  - A dataset is a defined set of files
  - Files are transferred only as part of a dataset
- **The PoolFileCatalog API is used to give uniform file access in all the grids**
  - In LCG/EGEE the replica catalog used is LFC
  - Aims to have a uniform data access in all the environments
- **ATLAS specific services and catalogs are restricted to Tier-1 centers**
- **Data transfer is operated through FTS channels**



The ATLAS cloud model



FTS Server T0

FTS Server T1

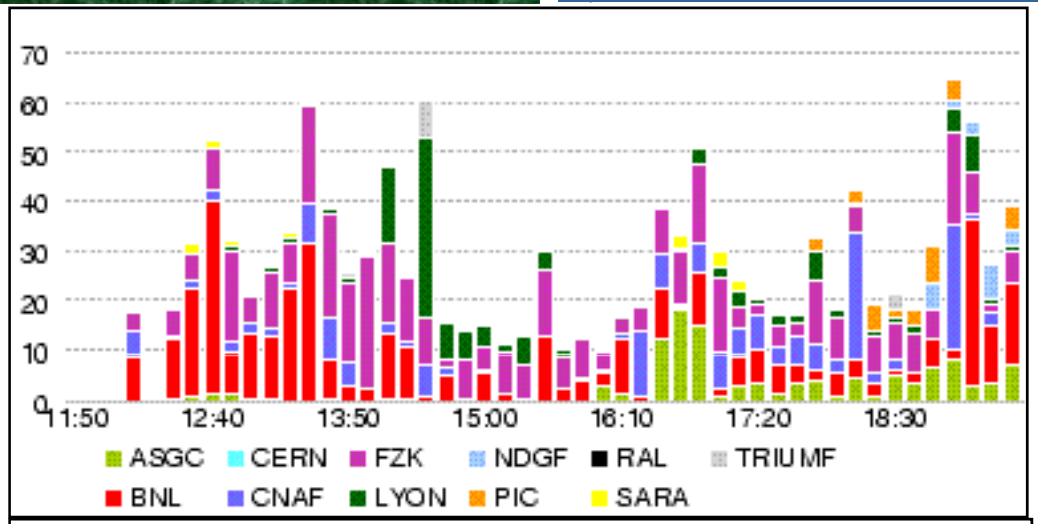
LFC (LCG File Catalog): local within 'cloud'

All SEs SRM

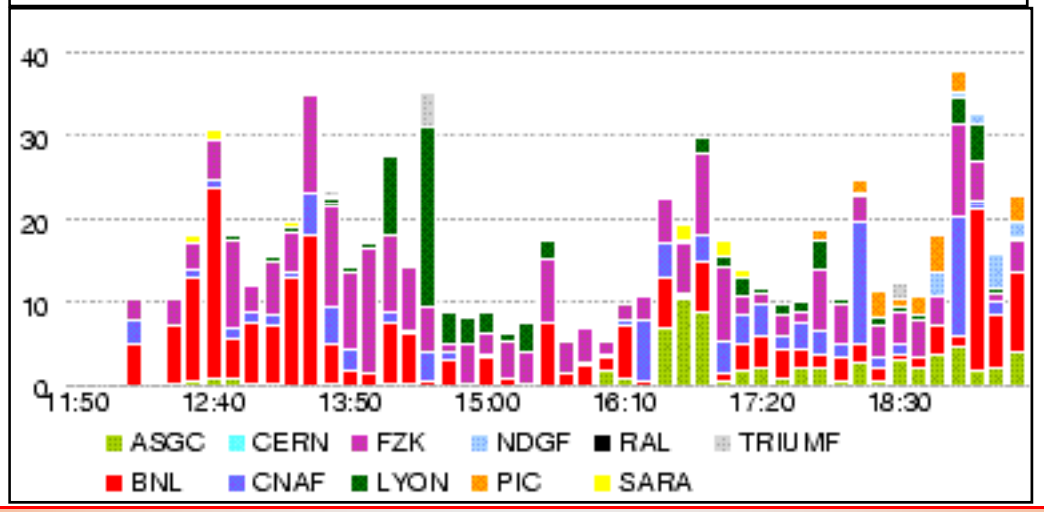
ATLAS DDM data transfers (10-4-2007)

http://dashb-atlas-data.cern.ch/dashboard/request.py/site

Throughput (MB/s)



Data Transferred (GB)



Analysis model

- **Analysis model broken into two components**
  - **@ Tier 1: Scheduled central production of augmented AOD, tuples & TAG collections from ESD**
    - → Derived files moved to other T1s and to T2s
  - **@ Tier 2: On-demand user analysis of augmented AOD streams, tuples, new selections etc and individual user simulation and CPU-bound tasks matching the official MC production**
    - → Modest job traffic between T2s
    - → Tier 2 files are not private, but may be for small sub-groups in physics/detector groups
    - → Limited individual space, copy to Tier3s

Group analysis

- **Group analysis will produce**
  - Copies of subsets
  - Dataset definitions
  - TAG selections
- **Characterised by access to full ESD and perhaps RAW**
  - Resource intensive
  - Must be a scheduled activity
  - Can back-navigate from AOD to ESD at same site
  - Can harvest small samples of ESD (and some RAW) to be sent to Tier 2s
  - Must be agreed by physics and detector groups
- **Big Trains**
  - Efficiency and scheduling gains access if analyses are blocked into a 'big train'
  - Idea around for a while, already used in e.g. heavy ions
    - Each wagon (group) has a wagon master (production manager)
    - Must ensure will not derail the train
  - Train must run often enough (every ~2 weeks?)
  - Trains can also harvest ESD and RAW samples for Tier 2s (but we should try to anticipate and place these subsets

Group analysis at Tier-1 centers

- **Per-user resources (assumed for all working groups), 1 user per group (production manager)**
  - **1000 passes through 1/10th of ESD sample (most will be on sub-streams) or 100 passes through the full ESD**

	2007	2008	2009	2010
CPU (MSI2k)	1.3	9.2	4.6	8.9
Disk (TB)	67	483	1107	2067
Events	170M	2.1B	1.6B	2.4B

On-demand analysis

- **Restricted to Tier 2s and CAF**
  - Can specialise some Tier 2s for some groups
  - ALL Tier 2s are for ATLAS-wide usage
- **Most ATLAS Tier 2 data should be 'placed' and have a lifetime of order months**
  - Job must go to the data
  - This means the Tier 2 bandwidth is lower than if you pull data to the job
- **Role and group based quotas are essential**
  - No user quotas, only group quotas
- **Data Selection**
  - Over small samples with Tier-2 file-based TAG and AMI dataset selector
  - TAG queries over larger samples by batch job to database TAG at Tier-1s/large Tier 2s
- **What data?**
  - Group-derived formats
  - Subsets of ESD and RAW
    - Pre-selected or selected via a Big Train run by working group
    - No back-navigation between sites, formats should be co-located

User analysis

- **Per-user resource (assumed with time for 700 active users)**
  - **Assume in 2007/2008, much of work done through group (to get data in shape for other work)**
  - **25 passes through user sample**

	2007	2008	2009	2010
CPU (kSI2k)	2.8	12	12	19
Disk (TB)	0.6	2.4	4.2	6.2
Events	1.7M	11M	16M	24M

Optimized access

- **RAW, ESD and AOD will be streamed to optimise access**
- **The selection and direct access to individual events is via a TAG database**
  - **TAG is a keyed list of variables/event**
  - **Overhead of file opens is acceptable in many scenarios**
  - **Works very well with pre-streamed data**
- **Two roles**
  - **Direct access to event in file via pointer**
  - **Data collection definition function**
- **Two formats, file and database**
  - **Now believe large queries require full database**
    - Multi-TB relational database
    - Restricts it to Tier1s and large Tier2s/CAF
  - **File-based TAG allows direct access to events in files (pointers)**
    - Ordinary Tier2s hold file-based primary TAG corresponding to locally-held datasets

Streaming

- **4 streams from event filter in the TDR**
  - **Primary physics, calibration, express, problem events**
- **More ESD and RAW streaming**
  - **Will explore the access improvements in large-scale exercises**
  - **Are also looking at overlaps, bookkeeping, etc**
  - **Streaming on trigger bits**
- **~10-20 streams at AOD**
  - **Stream = disjoint partition of a run, defined by the selection algorithms provided by the physics community**
  - **May split from ESD/RAW stream, but not cross parent streams**
    - Must avoid disk space wastage
      - If streams are not exclusive, must limit overlaps (<<10% total)
      - Must be very careful with bookkeeping

Test Streaming for Final Dress Reharsal (summer 2007)

Exclusive stream	fraction
Jet	22%
Electron	35%
Muon	20%
Tau/MET	6%
Photon	5%
Overlap	13%

Data hierarchy and access

- **Based on Datasets (= defined set of files)**
  - **Files will have the complete associated luminosity blocks**
- **Support for dataset of datasets**
  - **The Stream is the high level component**
  - **Subset are also datasets**
  - **Use TAG for dataset to**
    - Make logical collection of events
    - Make physical TAG collection
    - Access individual events
- **Datasets will also be defined by physics groups and detector groups**
  - **Associated data will be modified and used for detector status, calibration studies, etc.**
- **To keep track of the metadata a specific application has been developed**
  - **ATLAS Metadata Interface (AMI)**

Data processing: the ATLAS grid infrastucture

- **3 collaborating grids**
  - **LCG**
  - **NorduGrid**
  - **OSG**
- **Similar hardware but different middleware**
  - **Different teams for each grid, collaborating to aggregate the different resources for the ATLAS production**

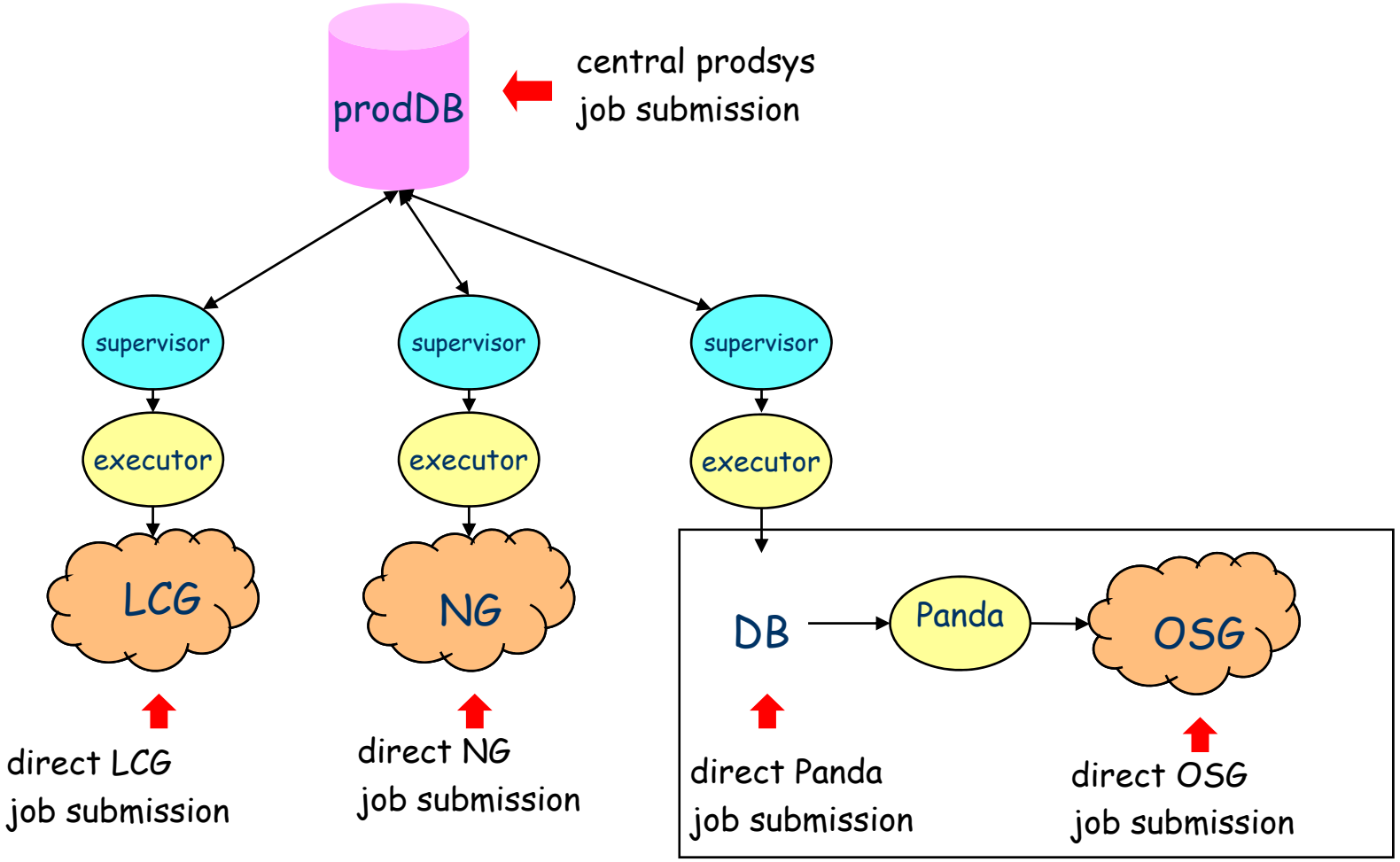
QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

- **Resource availability**
  - **The grid resources are generally open to all the ATLAS grid users (users belonging to the ATLAS VO) but must be also working in local mode**

Production on different Grids

- **Different hardware and middleware but common experiment software structure and execution method**
  - **The same execution agents (transformations) ar being used in all the grid, making the resources homogeneous**
  - **Working both in grid-enabled systems as well as in local environments**
- **Different agents and wrappers for each grid flavor, but common framework (ProdSys)**
  - **Using a global Oracle database hosted at CERN (ProdSys database)**
- **The production jobs are defined in the ProdSys DB, independently of the Grid flavor, and then dispatched to the appropriate facility**

The ATLAS Production System



WLCG/EGEE

- **Submission methods**
  - **LCG Resource Broker**
  - **Condor-G direct submission**
    - Also using glide-ins for CRONUS
- **Testing the new gLite Resource Broker (WMS)**
  - **Allows bulk submissions and other enhancements**
  - **Tested within the ATLAS LCG/EGEE Task Force**
- **http://lcg.web.cern.ch**

NorduGrid

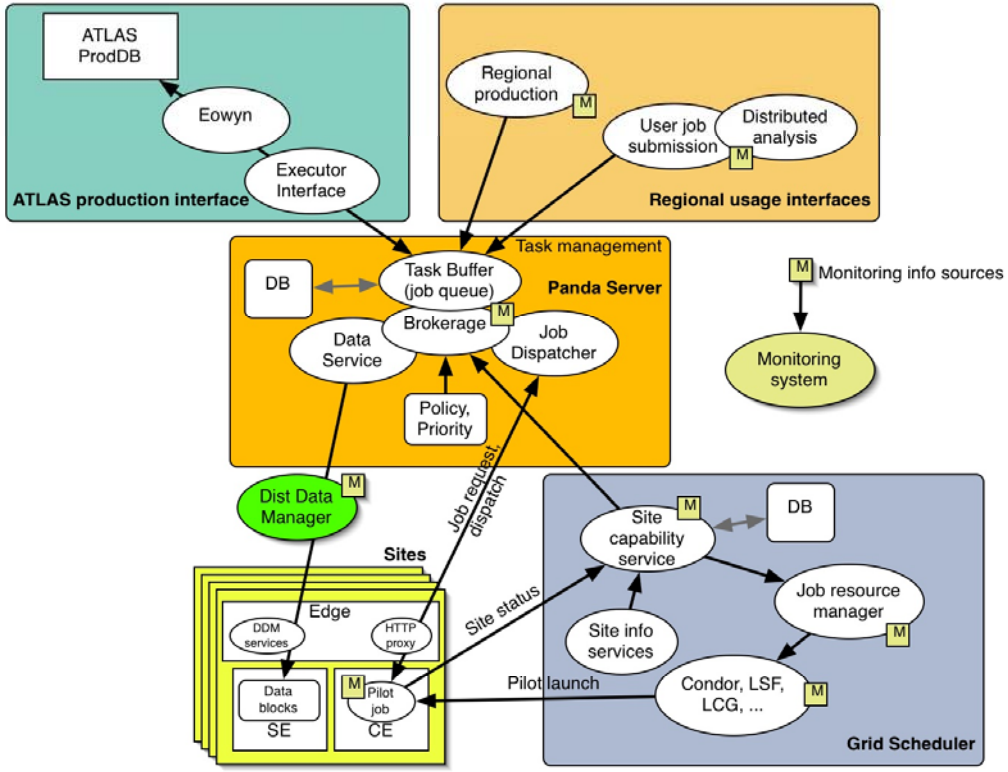
- Light middleware package (ARC, ~13 MB in total)
- Extended functionalities for the CE
  - Input/output files staging and caching
  - Controlled by XRSL (extension to globus RSL)
- Submission methods
  - The brokering is an integral part of the client software
  - Logging and bookeeping is done at the site level
- http://www.nordugrid.org

OSG/PANDA

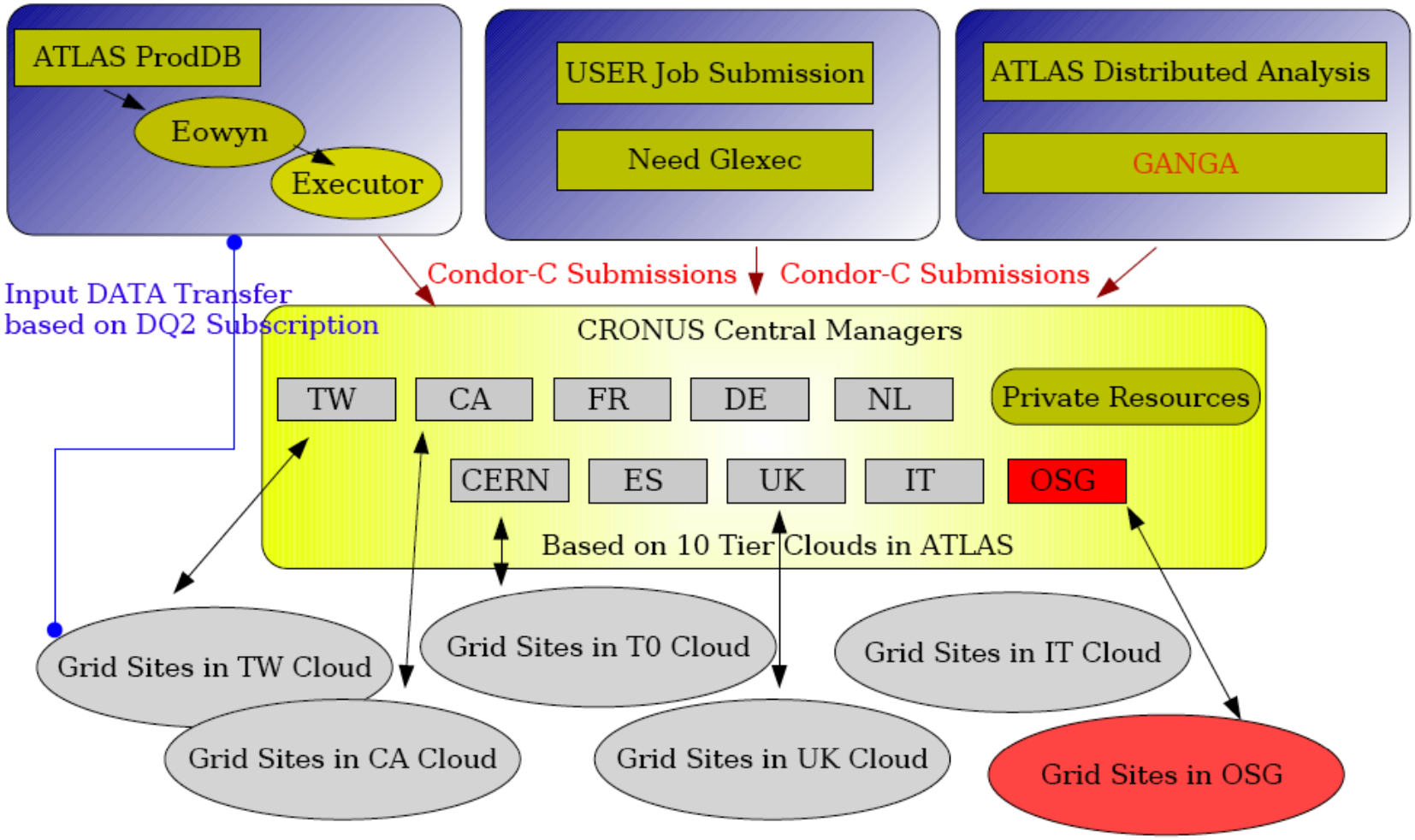
- **PANDA has been developed for OSG as an interface to ProdSys and for Distributed Analysis tasks**

- **It includes**
  - **A brokering system**
  - **Pilot jobs facility**
  - **Monitoring tools**
  - **Integration with the ATLAS DDM**

- **Direct submission of production and analysis jobs**



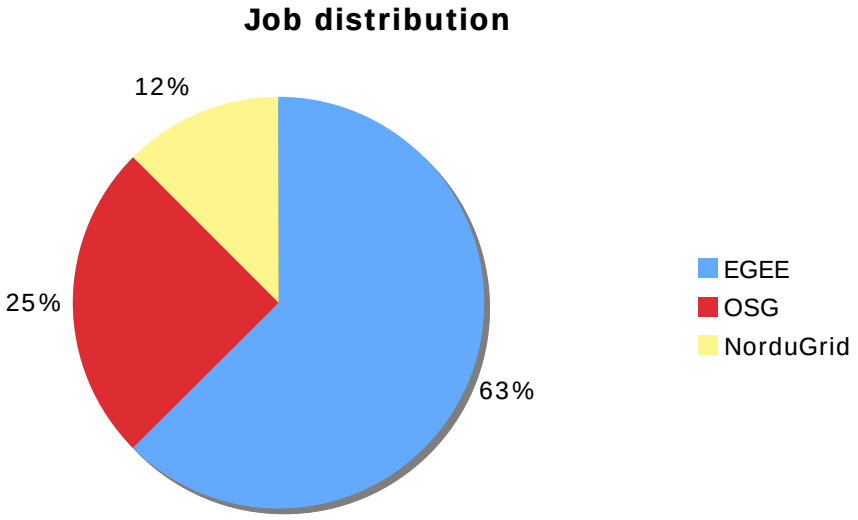
CRONUS: glide-ins based Condor-C submission



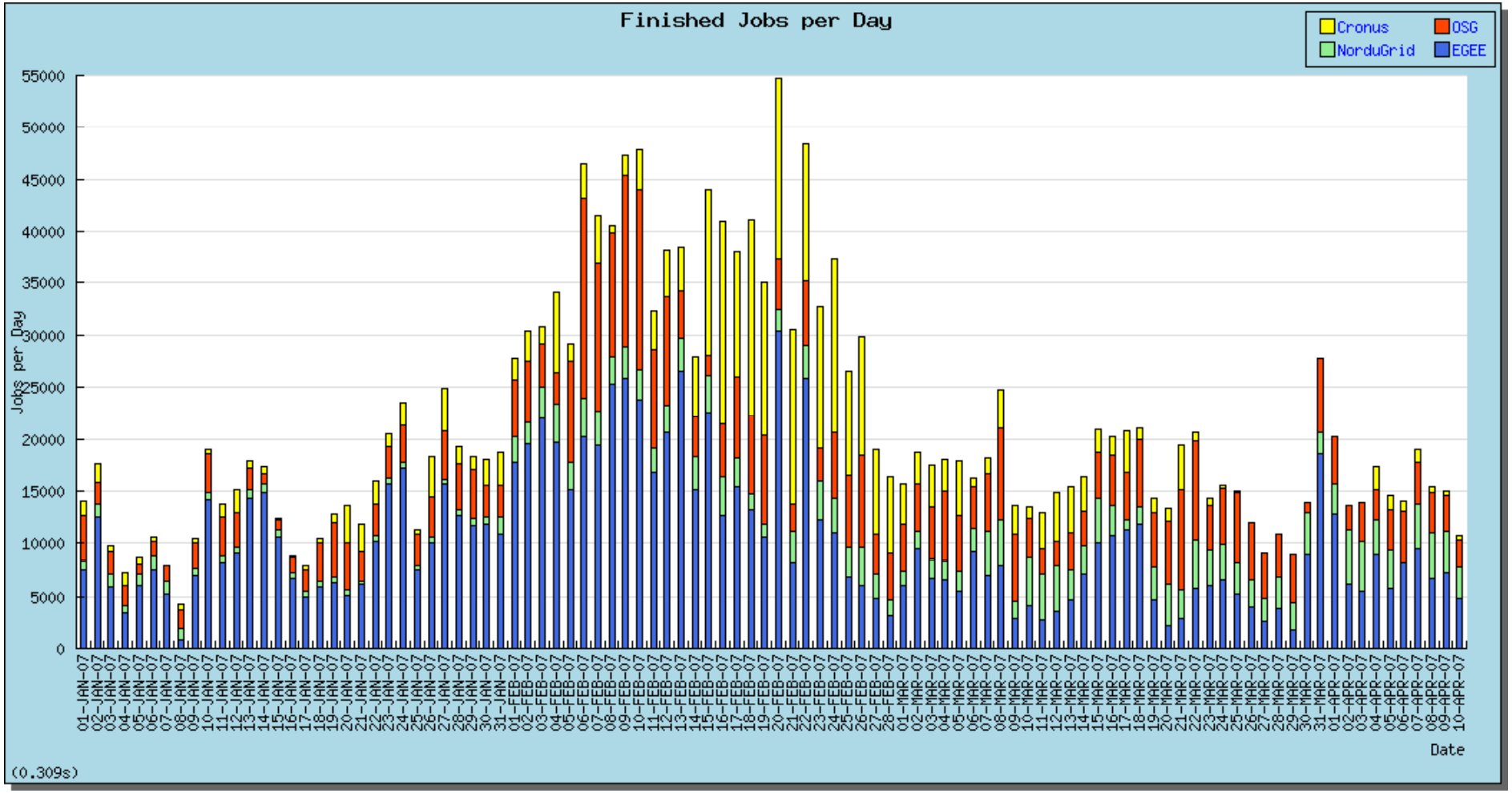
MC Production status (Computing System Commissioning)

- **Computing System Commissioning**
  - CSC started in the second half of 2006
  - Full-chain test of the ATLAS software
  - Preparation to the Full Dress Reharsal (comprehensive test of all the software components)

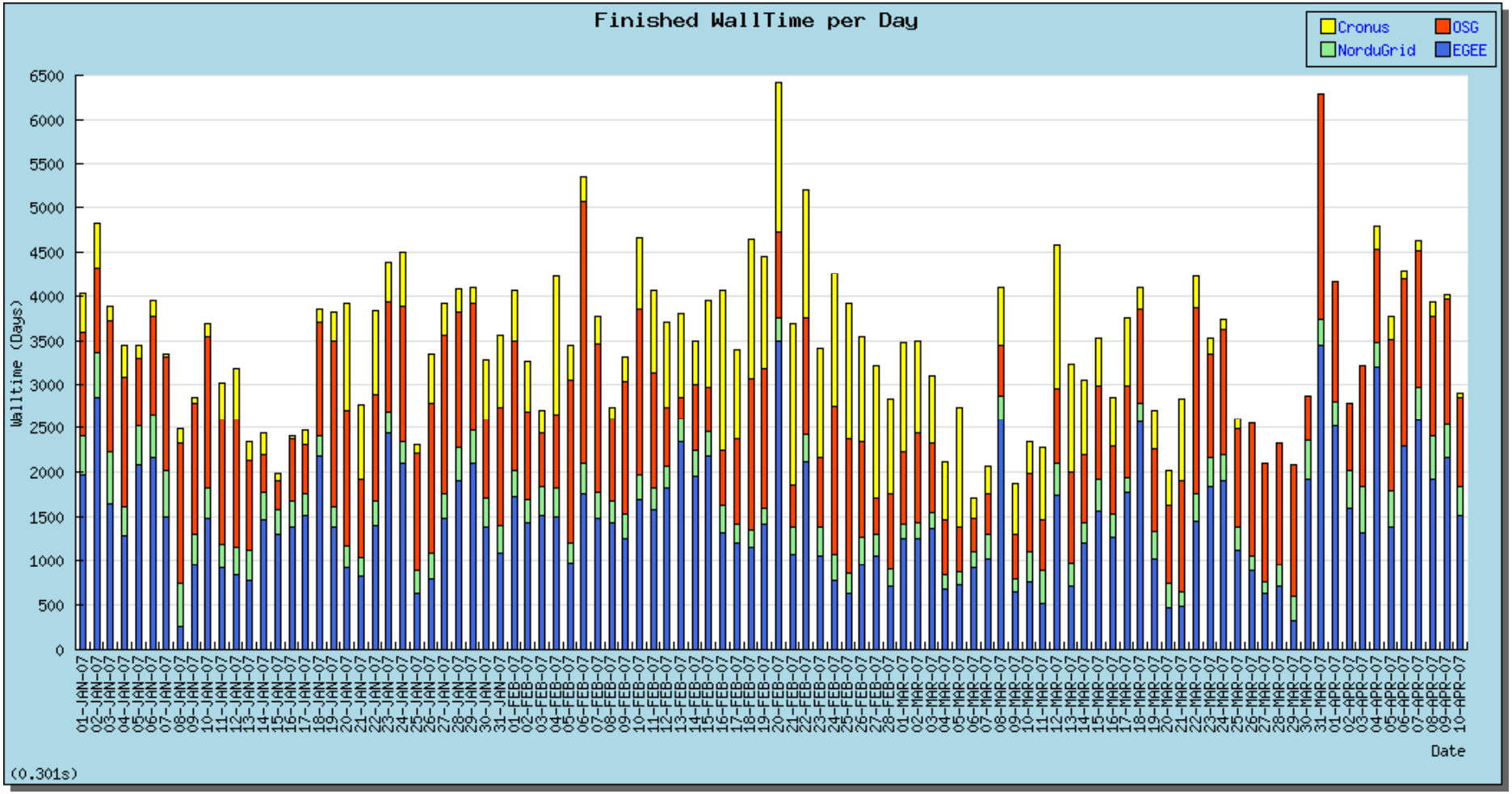
- **LCG/EGEE**
  - ~1.9M jobs, ~98M events
- **OSG**
  - ~0.8M jobs, ~39M events
- **NorduGrid (NDGF)**
  - ~0.4M jobs, ~19M events



MC production Jobs (Jan-Apr 2007)



MC production Walltime (Jan-Apr 2007)



Distributed Analysis tools

- **Tools**
  - **Configuration**
    - LCG
      - Short/Analysis queues
      - Job priorities
    - OSG
      - PANDA task queue
  - **Submission tools**
    - LCG
      - RB/WMS or direct Condor-G submissions are supported
    - OSG
      - PANDA submission tools
- **Analysis Framework**
  - **Mainly based on GANGA**
    - CLI, GUI and python interface

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Gaudi/Athena and Grid Alliance

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Conclusions

- **Computing Model Data has been well evolved to allow optimized placement of the ATLAS data at Tier Centers**
  - **Raw, ESD and AOD**
- **The analysis model is also progressing well**
  - **Most of the activites are well defined**
  - **Still we need to understand all the implications of the Physics Analysis model and the data placement/selection**
- **Several issues are being addressed during the Computing System Commissioning, started in the second half of 2006**
  - **Large-scale MC production**
  - **Data distribution and replication**
  - **…**
- **Full Dress Reharsal (comprehensive test all the Computing System components) to be performed in summer 2007**
- **Some issues will only be resolved with real data (2007/2008)**