ATLAS Activities at Tier-2s

Dario Barberis

CERN & Genoa University

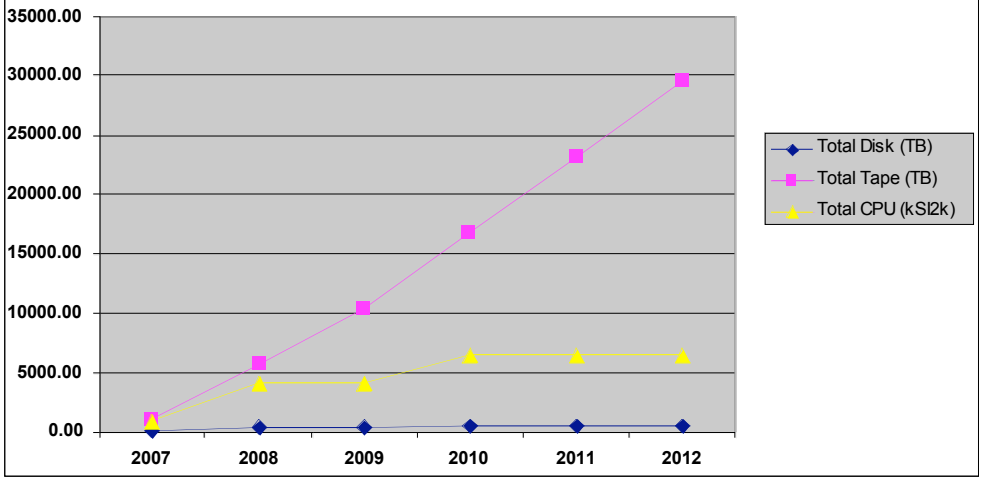
Computing Model: central operations

- Tier-0:
  - Copy RAW data to Castor tape for archival
  - Copy RAW data to Tier-1s for storage and reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs & TAGS) to Tier-1s
- Tier-1s:
  - Store and take care of a fraction of RAW data
  - Run "slow" calibration/alignment procedures
  - Rerun reconstruction with better calib/align and/or algorithms
  - Distribute reconstruction output to Tier-2s
  - Keep current versions of ESDs and AODs on disk for analysis
- Tier-2s:
  - Run simulation
  - Run calibration/alignment procedures
  - Keep current versions of AODs on disk for analysis
  - Run user analysis jobs

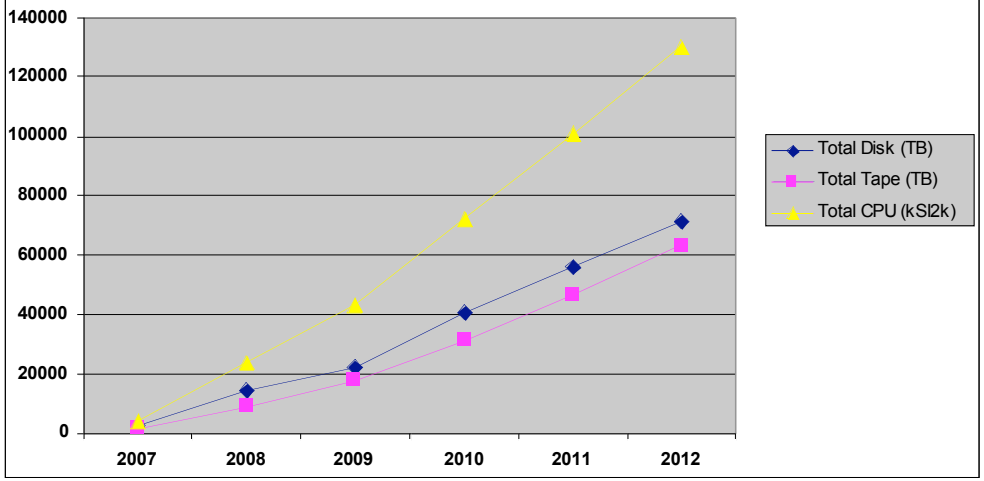
Computing Model and Resources

- The ATLAS Computing Model is still the same as in the Computing TDR (June 2005) and basically the same as in the Computing Model document (Dec. 2004) submitted for the LHCC review in January 2005
- The sum of 30-35 Tier-2s will provide ~40% of the total ATLAS computing and disk storage capacity
  - CPUs for full simulation productions and user analysis jobs
    - On average 1:2 for central simulation and analysis jobs
  - Disk for AODs, samples of ESDs and RAW data, and most importantly for selected event samples for physics analysis
- We do not ask Tier-2s to run any particular service for ATLAS in addition to providing the Grid infrastructure (CE, SE, etc.)
  - All data management services (catalogues and transfers) are run from Tier-1s
- Some "larger" Tier-2s may choose to run their own services, instead of depending on a Tier-1
  - In this case, they should contact us directly
- Depending on local expertise, some Tier-2s will specialise in one particular task
  - Such as calibrating a very complex detector that needs special access to particular datasets

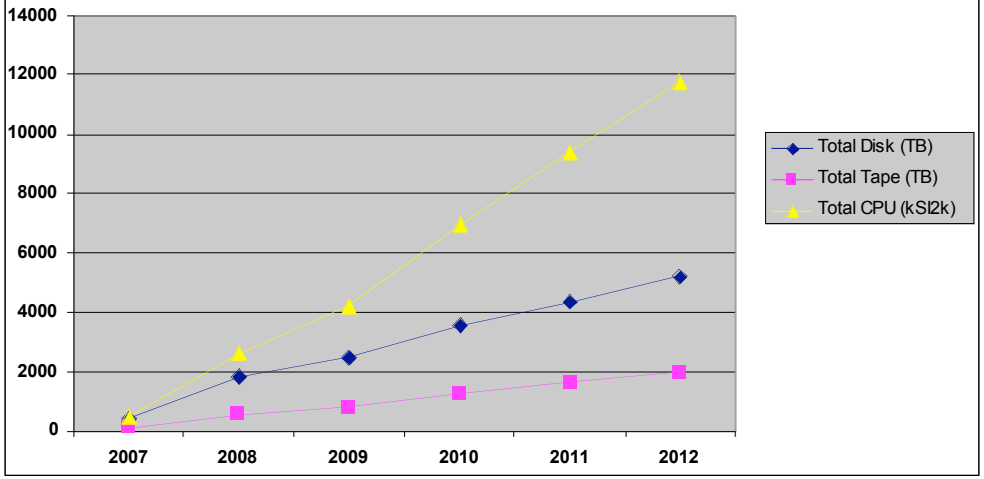
Tier-0



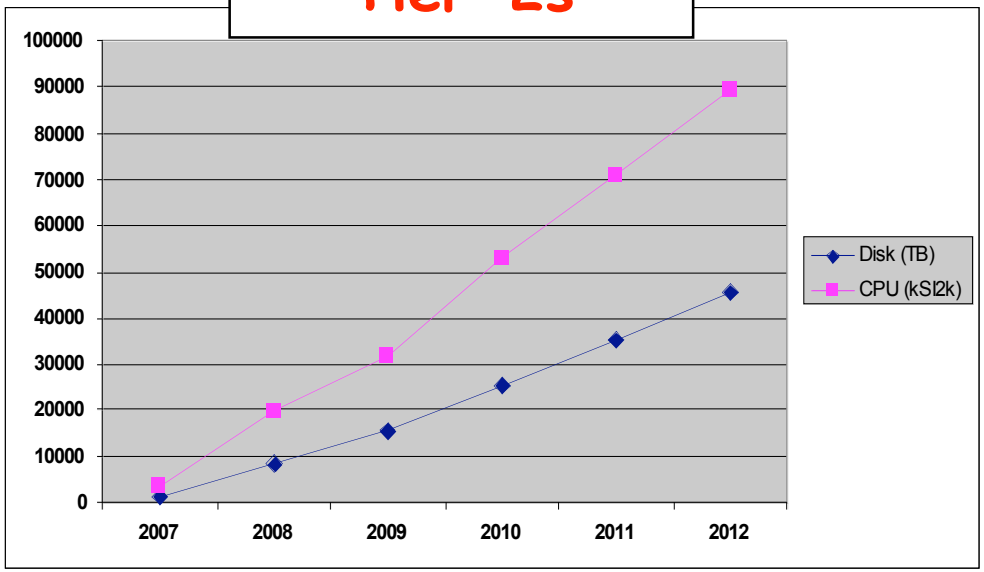
Tier-1s



CERN Analysis Facility



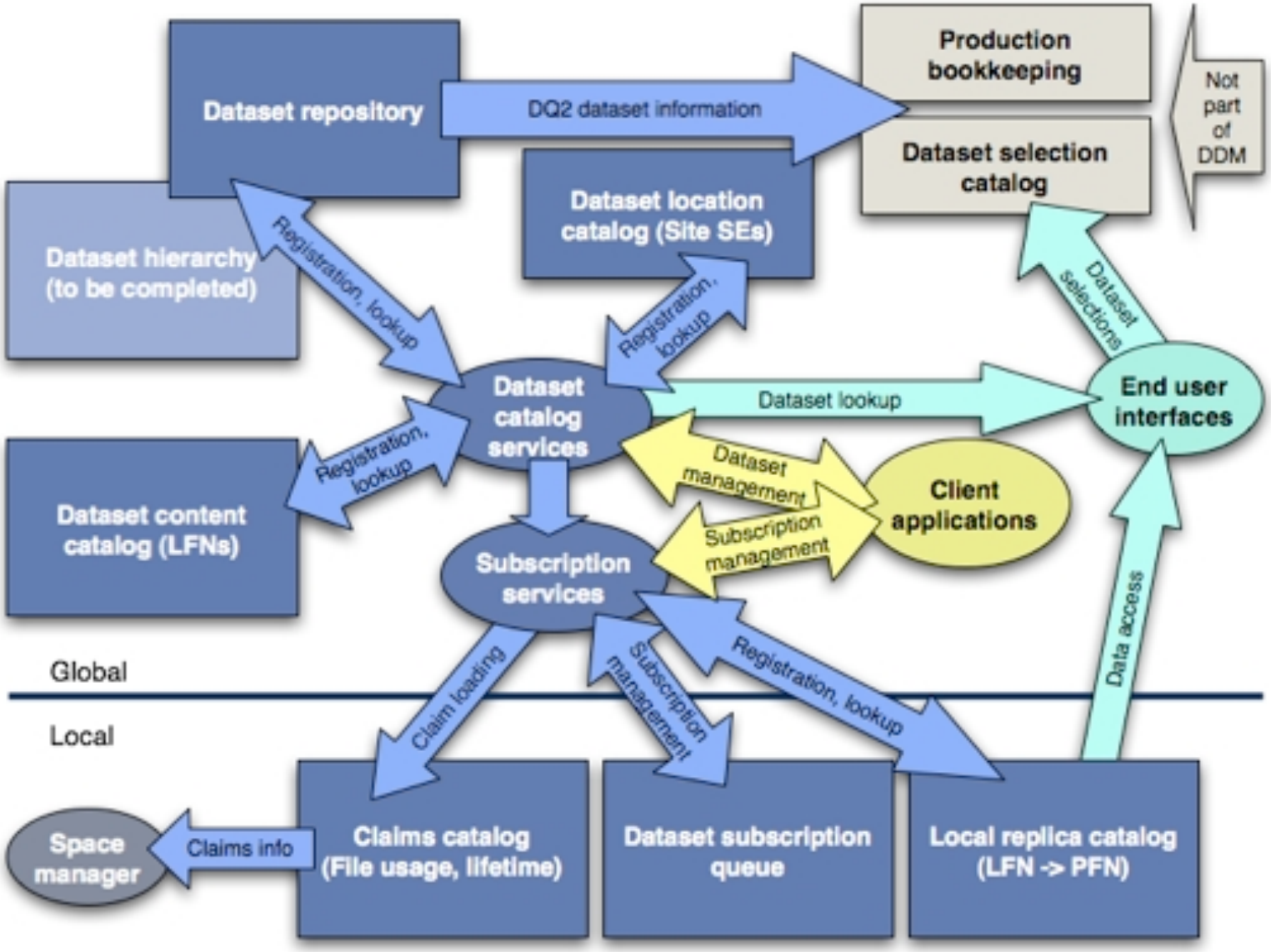
Tier-2s



ATLAS Distributed Data Management

- ATLAS reviewed all its own Grid distributed systems (data management, production, analysis) during the first half of 2005
  - In parallel with the LCG BSWG activity
- A new Distributed Data Management System (DDM) was designed, based on:
  - A hierarchical definition of datasets
  - Central dataset catalogues
  - Data blocks as units of file storage and replication
  - Distributed file catalogues
  - Automatic data transfer mechanisms using distributed services (dataset subscription system)
- The DDM system allows the implementation of the basic ATLAS Computing Model concepts, as described in the Computing TDR (June 2005):
  - Distribution of raw and reconstructed data from CERN to the Tier-1s
  - Distribution of AODs (Analysis Object Data) to Tier-2 centres for analysis
  - Storage of simulated data (produced by Tier-2s) at Tier-1 centres for further distribution and/or processing

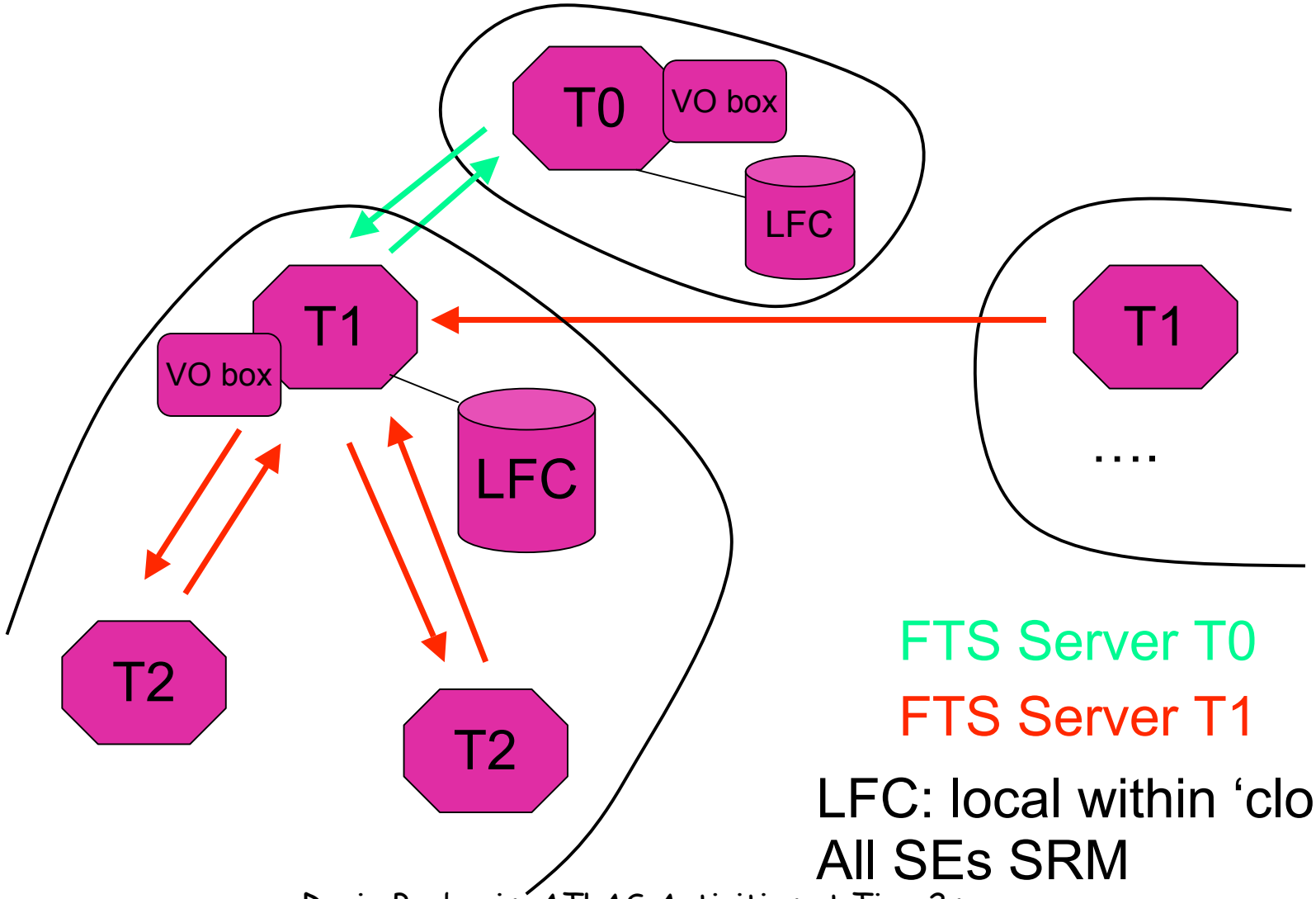
ATLAS DDM Organization



Central vs Local Services

- The DDM system has now a central role with respect to ATLAS Grid tools
- One fundamental feature is the presence of distributed file catalogues and (above all) auxiliary services
  - Clearly we cannot ask every single Grid centre to install ATLAS services
  - We decided to install "local" catalogues and services at Tier-1 centres
  - Then we defined "regions" which consist of a Tier-1 and all other Grid computing centres that:
    - Are well (network) connected to this Tier-1
    - Depend on this Tier-1 for ATLAS services (including the file catalogue)
- We believe that this architecture scales to our needs for the LHC data-taking era:
  - Moving several 10000s files/day
  - Supporting up to 100000 organized production jobs/day
  - Supporting the analysis work of >1000 active ATLAS physicists

Tiers of ATLAS



ATLAS Data Management Model

- Tier-1s send AOD data to Tier-2s
- Tier-2s produce simulated data and send them to Tier-1s
- In the ideal world (perfect network communication hardware and software) we would not need to define default Tier-1—Tier-2 associations
- In practice, it turns out to be convenient (robust?) to partition the Grid so that there are default (not compulsory) data paths between Tier-1s and Tier-2s
  - FTS channels are installed for these data paths for production use
  - All other data transfers go through normal network routes
- In this model, a number of data management services are installed only at Tier-1s and act also on their "associated" Tier-2s:
  - VO Box
  - FTS channel server (both directions)
  - Local file catalogue (part of DDM/DQ2)

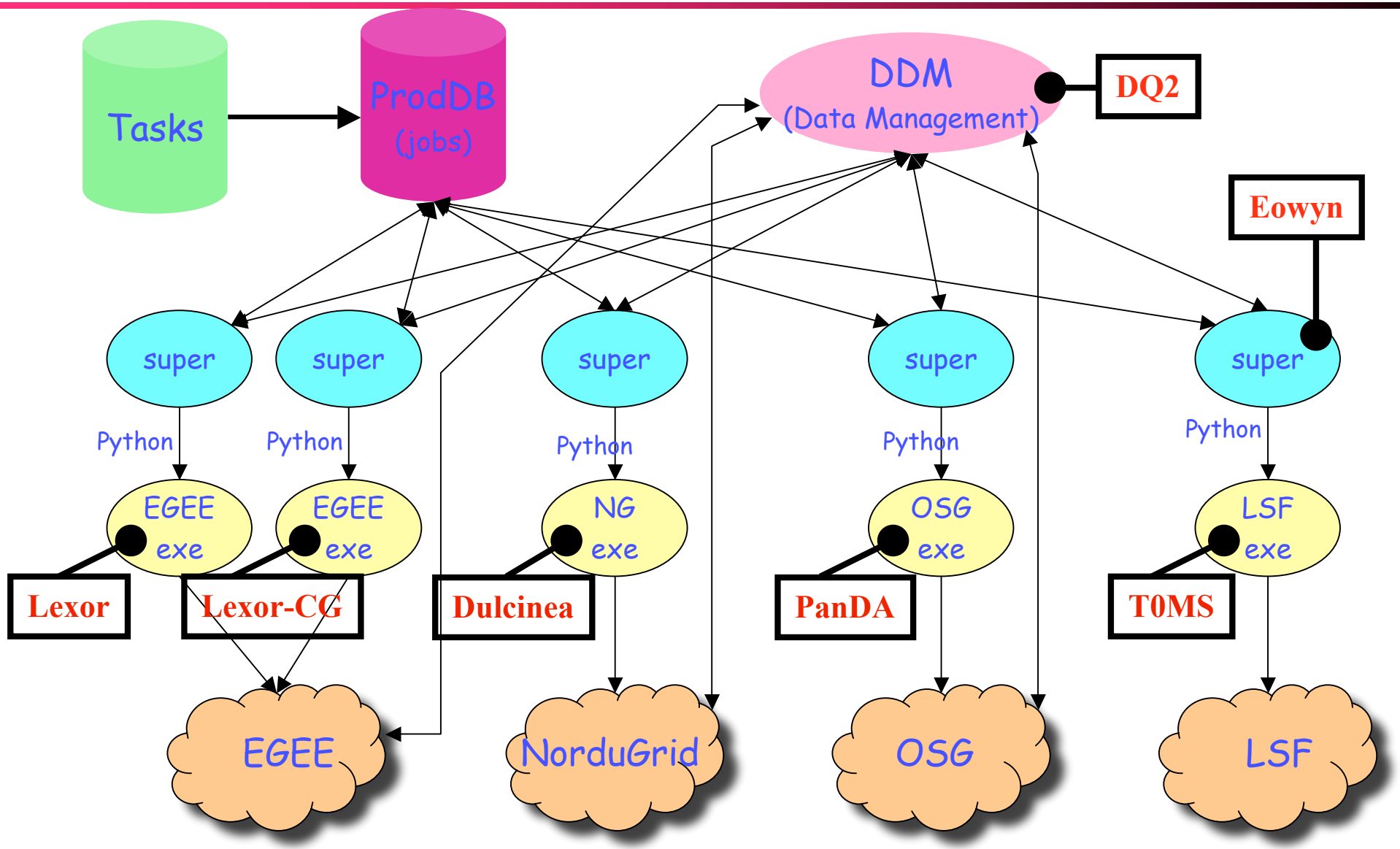
Data Management Considerations

- It is therefore "obvious" that the association must be between computing centres that are "close" from the point of view of:
  - network connectivity (robustness of the infrastructure)
  - geographical location (round-trip time)
- Rates are not a problem:
  - AOD rates (for a full set) from a Tier-1 to a Tier-2 are nominally:
    - 20 MB/s for primary production during data-taking
    - plus the same again for reprocessing from 2008 onwards
    - more later on as there will be more accumulated data to reprocess
  - Upload of simulated data for an "average" Tier-2 (3% of ATLAS Tier-2 capacity) is constant:
    - 0.03 * 0.2 * 200 Hz * 2.6 MB = 3.2 MB/s continuously
- Total storage (and reprocessing!) capacity for simulated data is a concern
  - The Tier-1s must store and reprocess simulated data that match their overall share of ATLAS
    - Some optimization is always possible between real and simulated data, but only within a small range of variations

Job Management: Productions

- Once we have data distributed in the correct way (rather than sometimes hidden in the guts of automatic mass storage systems), we can rework the distributed production system to optimise job distribution, by sending jobs to the data (or as close as possible to them)
  - This was not the case previously, as jobs were sent to free CPUs and had to copy the input file(s) to the local WN, from wherever in the world the data happened to be
- Next: make better use of the task and dataset concepts
  - A "task" acts on a dataset and produces more datasets
  - Use bulk submission functionality to send all jobs of a given task to the location of their input datasets
  - Minimise the dependence on file transfers and the waiting time before execution
  - Collect output files belonging to the same dataset to the same SE and transfer them asynchronously to their final locations

ATLAS Production System (2006)



Job Management: Analysis

- A system based on a central database (job queue) is good for scheduled productions (as it allows proper priority settings), but too heavy for user tasks such as analysis
- Lacking a global way to submit jobs, a few tools have been developed to submit Grid jobs in the meantime:
  - LJSF (Lightweight Job Submission framework) can submit ATLAS jobs to the LCG/EGEE Grid
    - It was derived initially from the framework developed to install ATLAS software at EDG Grid sites
  - Pathena can generate ATLAS jobs that act on a dataset and submits them to PanDA on the OSG Grid
- The ATLAS baseline tool to help users to submit Grid jobs is Ganga
  - Job splitting and bookkeeping
  - Several submission possibilities
  - Collection of output files

ATLAS Analysis Work Model

1. Job preparation:

Local system (shell)

Prepare JobOptions → Run Athena (interactive or batch) → Get Output

2. Medium-scale testing:

Local system (Ganga)

Prepare JobOptions
Find dataset from DDM
Generate & submit jobs

→ **Grid** Run Athena →

Local system (Ganga)

Job book-keeping
Get Output

3. Large-scale running:

Local system (Ganga)

Prepare JobOptions
Find dataset from DDM
Generate & submit jobs

→ **ProdSys** Run Athena on Grid
Store o/p on Grid →

Local system (Ganga)

Job book-keeping
Access output from Grid
Merge results

Analysis Jobs at Tier-2s

- Analysis jobs must run where the input data files are
  - As transfering data files from other sites may take longer than actually running the job
- Most analysis jobs will take AODs as input for complex calculations and event selections
  - And most likely will output Athena-Aware Ntuples (AAN, to be stored on some close SE) and histograms (to be sent back to the user)
- We assume that people will develop their analyses and run them on reduced samples many many times before launching runs on a complete dataset
  - There will be a large number of failures due to people's code!
- In order to assure execution of analysis jobs with a reasonable turn-around time, we have to set up a priority system that separates centrally organised productions from analysis tasks
  - More on this in D.Liko's talk tomorrow afternoon

Conclusions

- ATLAS operations at Tier-2s are well defined in the Computing Model
- We thank all Tier-2 managers and funding authorities for providing ATLAS with the much needed capacity
- We are trying not to impose any particular load on Tier-2 managers by running distributed services at Tier-1s
  - Although this concept breaks the symmetry and forces us to set up default Tier-1–Tier-2 associations
- All that is required of Tier-2s is to set up the Grid environment
  - Including whichever job queue priority scheme will be found most useful
  - And SRM Storage Elements with (when available) a correct implementation of the space reservation and accounting system