ATLAS Computing Model and Grid Operations

Dario Barberis

(CERN & Genoa University)

ATLAS Computing Co-ordinator

19 September 2006

The ATLAS Collaboration

(As of September 2006)

35 Countries
161 Institutions
1650 Scientific Authors total
(1300 with a PhD, for M&O share)

Albany, Alberta, NIKHEF Amsterdam, Ankara, LAPP Annecy, Argonne NL, Arizona, UT Arlington, Athens, NTU Athens, Baku,
IFAE Barcelona, Belgrade, Bergen, Berkeley LBL and UC, Humboldt U Berlin, Bern, Birmingham, Bologna, Bonn, Boston, Brandeis,
Bratislava/SAS Kosice, Brookhaven NL, Buenos Aires, Bucharest, Cambridge, Carleton, Casablanca/Rabat, CERN, Chinese Cluster, Chicago, Clermont-Ferrand, Columbia, NBI
Copenhagen, Cosenza, INP Cracow, FPNT Cracow, DESY, Dortmund, TU Dresden,
JINR Dubna, Duke, Frascati, Freiburg, Geneva, Genoa, Giessen, Glasgow, LPSC Grenoble, Technion Haifa, Hampton, Harvard, Heidelberg, Hiroshima, Hiroshima IT, Indiana, Innsbruck,
Iowa SU, Irvine UC, Istanbul Bogazici, KEK, Kobe, Kyoto, Kyoto UE, Lancaster, UN La Plata, Lecce, Lisbon LIP, Liverpool, Ljubljana, QMW London, RHBNC London, UC London, Lund,
UA Madrid, Mainz, Manchester, Mannheim, CPPM Marseille, Massachusetts, MIT, Melbourne, Michigan, Michigan SU, Milano, Minsk NAS, Minsk NCPHEP, Montreal, McGill Montreal,
FIAN Moscow, ITEP Moscow, MEPhI Moscow, MSU Moscow, Munich LMU,
MPI Munich, Nagasaki IAS, Naples, New Mexico, New York U, Nijmegen, BINP Novosibirsk, Ohio SU, Okayama, Oklahoma, Oklahoma SU, Oregon, LAL Orsay, Osaka, Oslo, Oxford,
Paris VI and VII, Pavia, Pennsylvania, Pisa, Pittsburgh, CAS Prague,
CU Prague, TU Prague, IHEP Protvino, Ritsumeikan, UFRJ Rio de Janeiro, Rochester, Rome I, Rome II, Rome III,
Rutherford Appleton Laboratory, DAPNIA Saclay, Santa Cruz UC, Sheffield, Shinshu, Siegen, Simon Fraser Burnaby,
Southern Methodist Dallas, NPI Petersburg, SLAC, Stockholm, KTH Stockholm, Stony Brook, Sydney, AS Taipei, Tbilisi, Tel Aviv, Thessaloniki, Tokyo ICEPP, Tokyo MU, Toronto,
TRIUMF, Tsukuba, Tufts, Udine, Uppsala, Urbana UI, Valencia, UBC Vancouver, Victoria, Washington, Weizmann Rehovot, Wisconsin, Wuppertal, Yale, Yerevan

ATLAS Data Collection

- Protons flying in opposite directions will collide with a centre-of-mass energy of 14 TeV (~14000 times the proton rest mass) in the centre of the ATLAS detector
- Each such collision produces several (tens of) particles that are absorbed and detected by the ATLAS detector
- The ensemble of the electronic signals produced in all detector components by a single collision is called an "event"
- Events can take place at rates up to 40 MHz, but "interesting" ones will occur much more rarely (100-1000 Hz)
- The online data acquisition system will collect together all signals that belong to the same event and select "interesting" ones (max. rate 200 Hz, limited by bandwidth and offline processing)
- These events are sent to the CERN computing centre (Tier-0) for processing and distribution

Event Data Model

- RAW:
  - "ByteStream" format, ~1.6 MB/event
- ESD (Event Summary Data):
  - Full output of reconstruction in object (POOL/ROOT) format:
    - Tracks (and their hits), Calo Clusters, Calo Cells, combined reconstruction objects etc.
  - Nominal size 500 kB/event
    - Currently 2-3 times larger: contents and technology under revision, following feedback on the first prototype implementation
    - Assumed to decrease in size as we improve our understanding of the detector
- AOD (Analysis Object Data):
  - Summary of event reconstruction with "physics" (POOL/ROOT) objects:
    - electrons, muons, jets, etc.
  - Nominal size 100 kB/event
- TAG:
  - Database used to quickly select events in AOD and/or ESD files

Computing Model: event data flow from EF

- Events are written in "ByteStream" format by the Event Filter farm in ~2 GB files
  - ~1000 events/file (nominal size is 1.6 MB/event)
  - 200 Hz trigger rate (independent of luminosity)
  - Currently several streams are foreseen:
    - Express stream with "most interesting" events
    - ~5 event streams, separated by trigger signature (e.g. muons, electromagnetic, hadronic jets, taus, minimum bias)
    - Calibration events
    - "Trouble maker" events (for debugging)
  - One 2-GB file every 5 seconds will be available from the Event Filter
  - Data will be transfered to the Tier-0 input buffer at 320 MB/s (average)

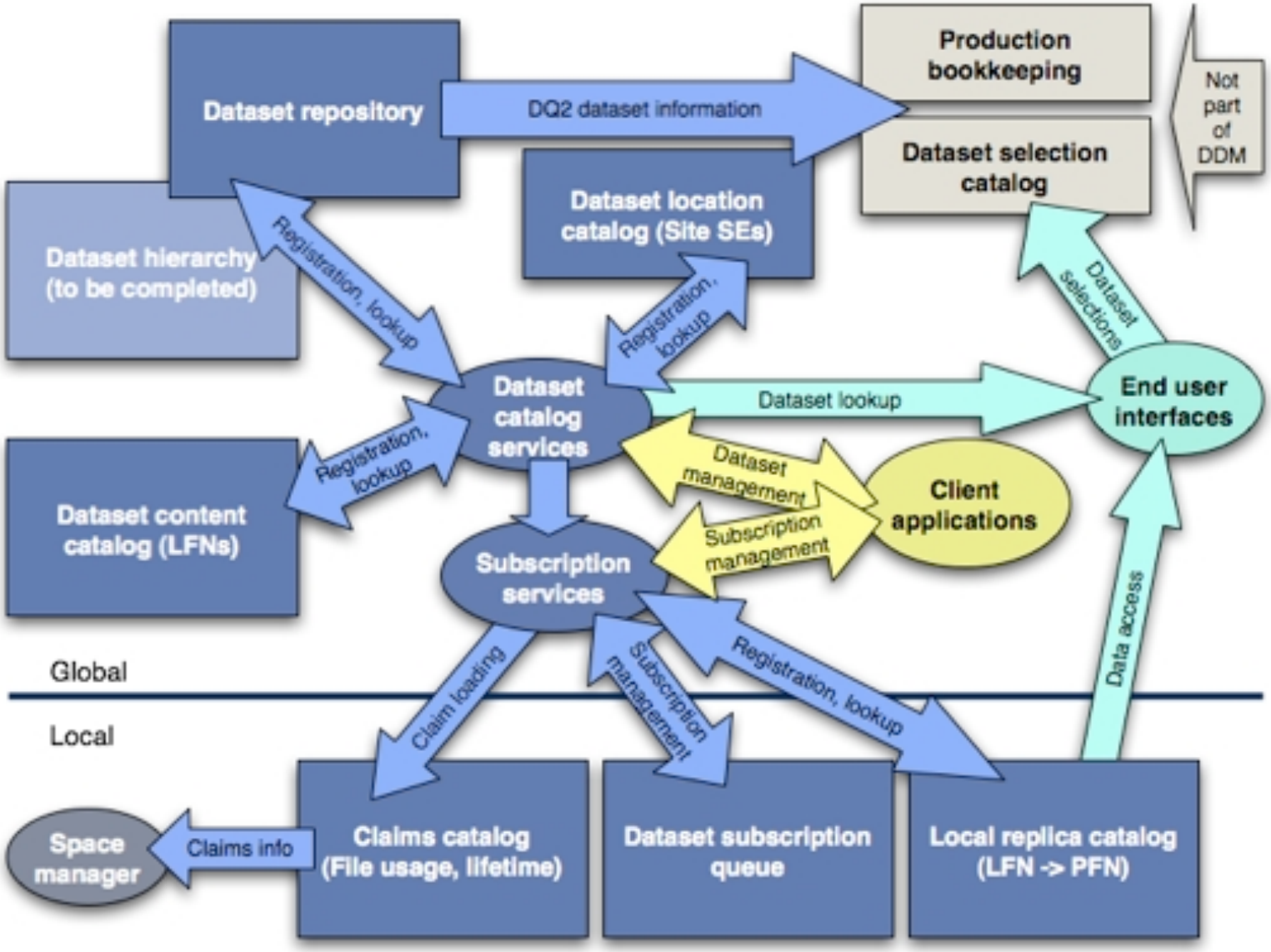
Computing Model: central operations

- 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 (forever)
  - 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 (and calibration/alignment when appropriate)
  - Keep current versions of AODs on disk for analysis

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 WLCG 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

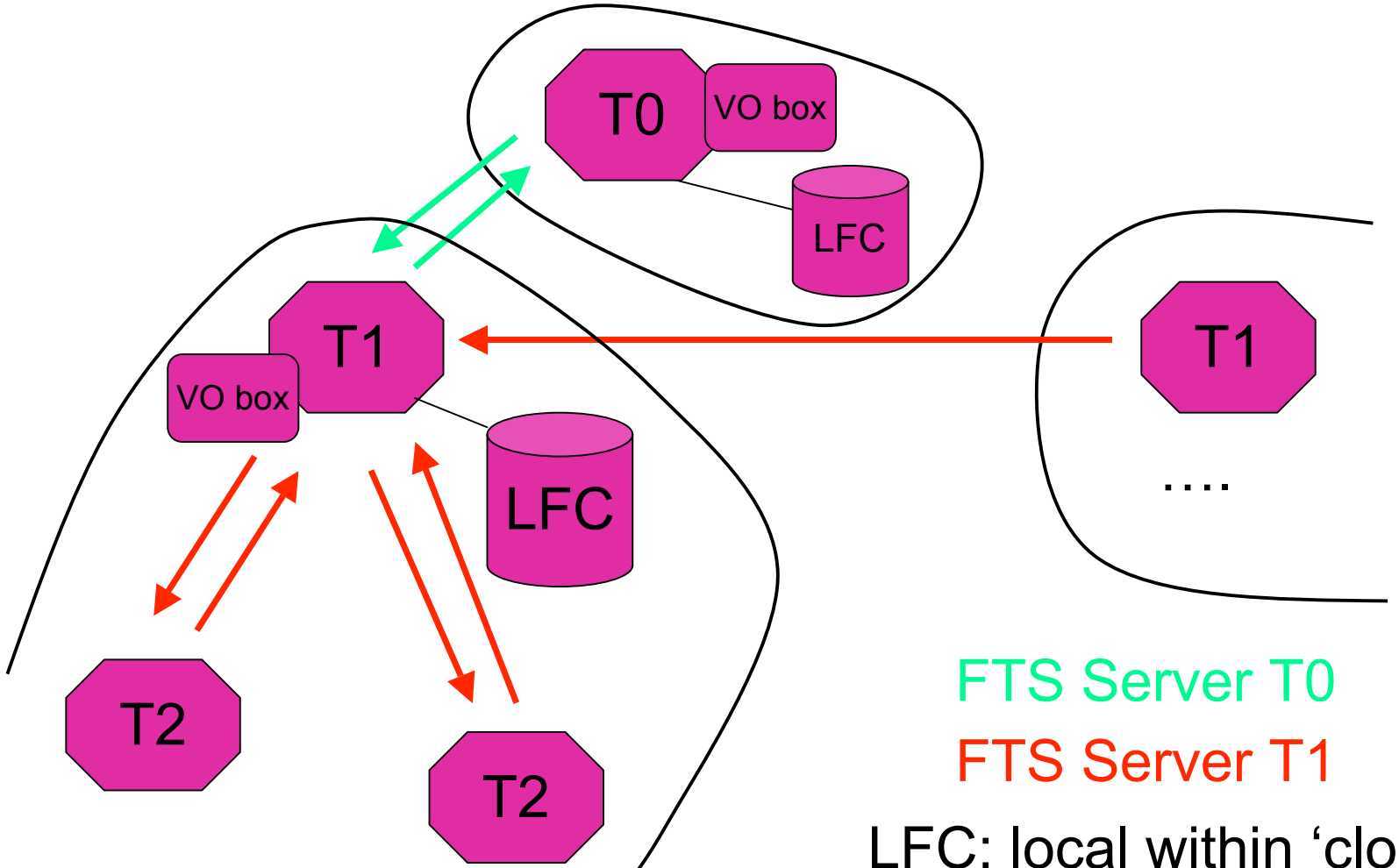
ATLAS DDM Organization



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

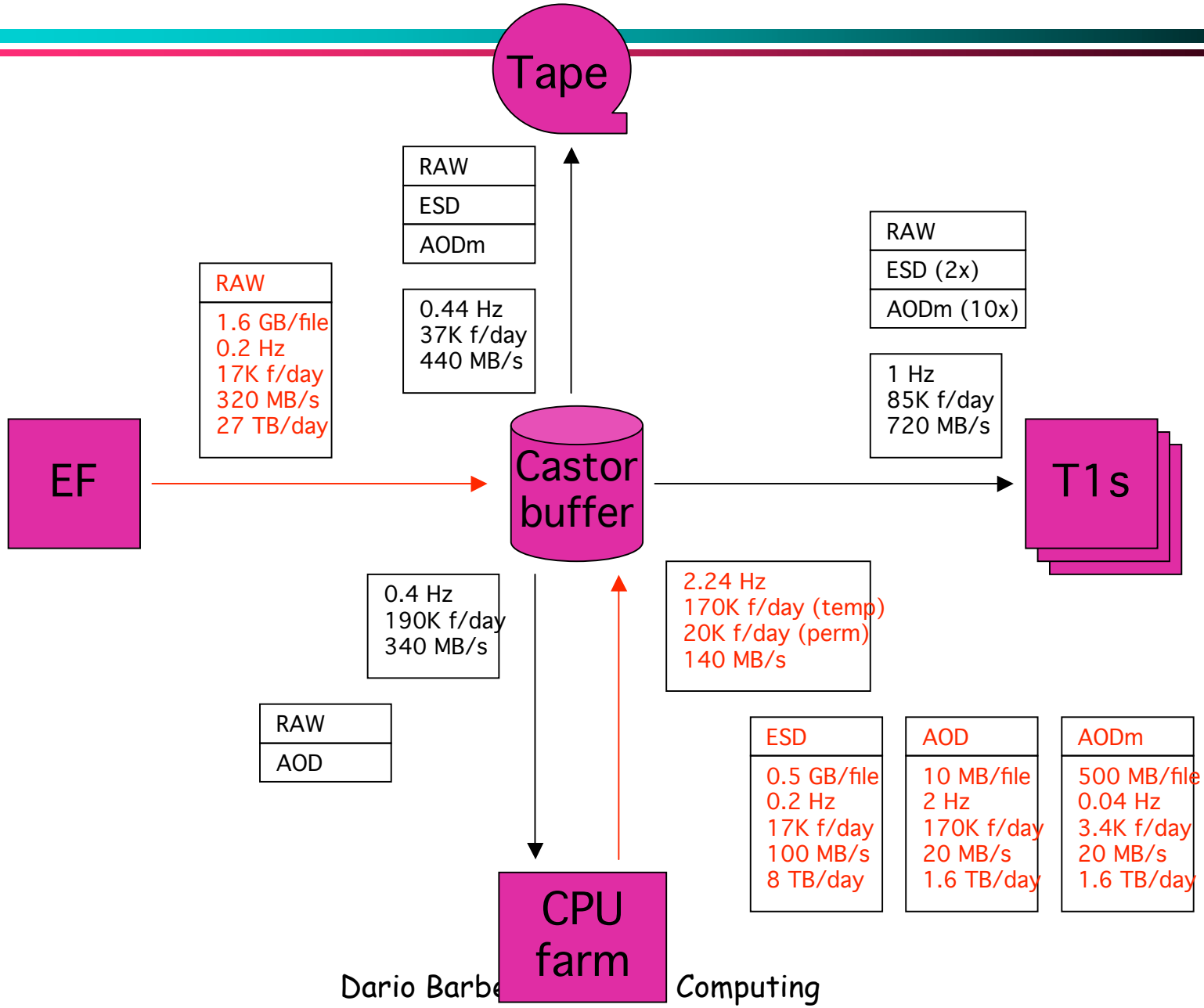
Tiers of ATLAS



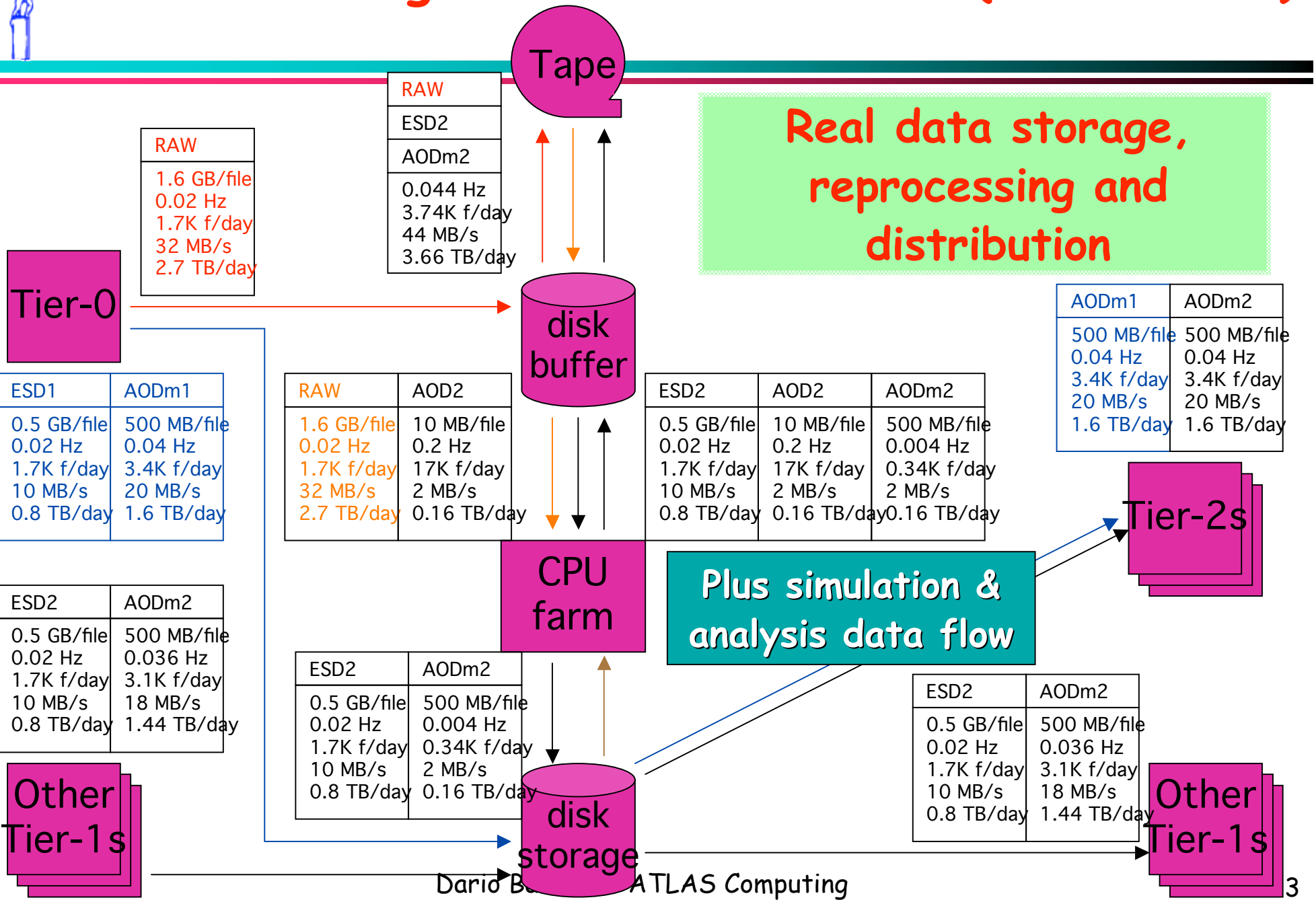
ATLAS Data Management Model

- 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)

ATLAS Tier-0 Data Flow



ATLAS "average" Tier-1 Data Flow (2008-2009)



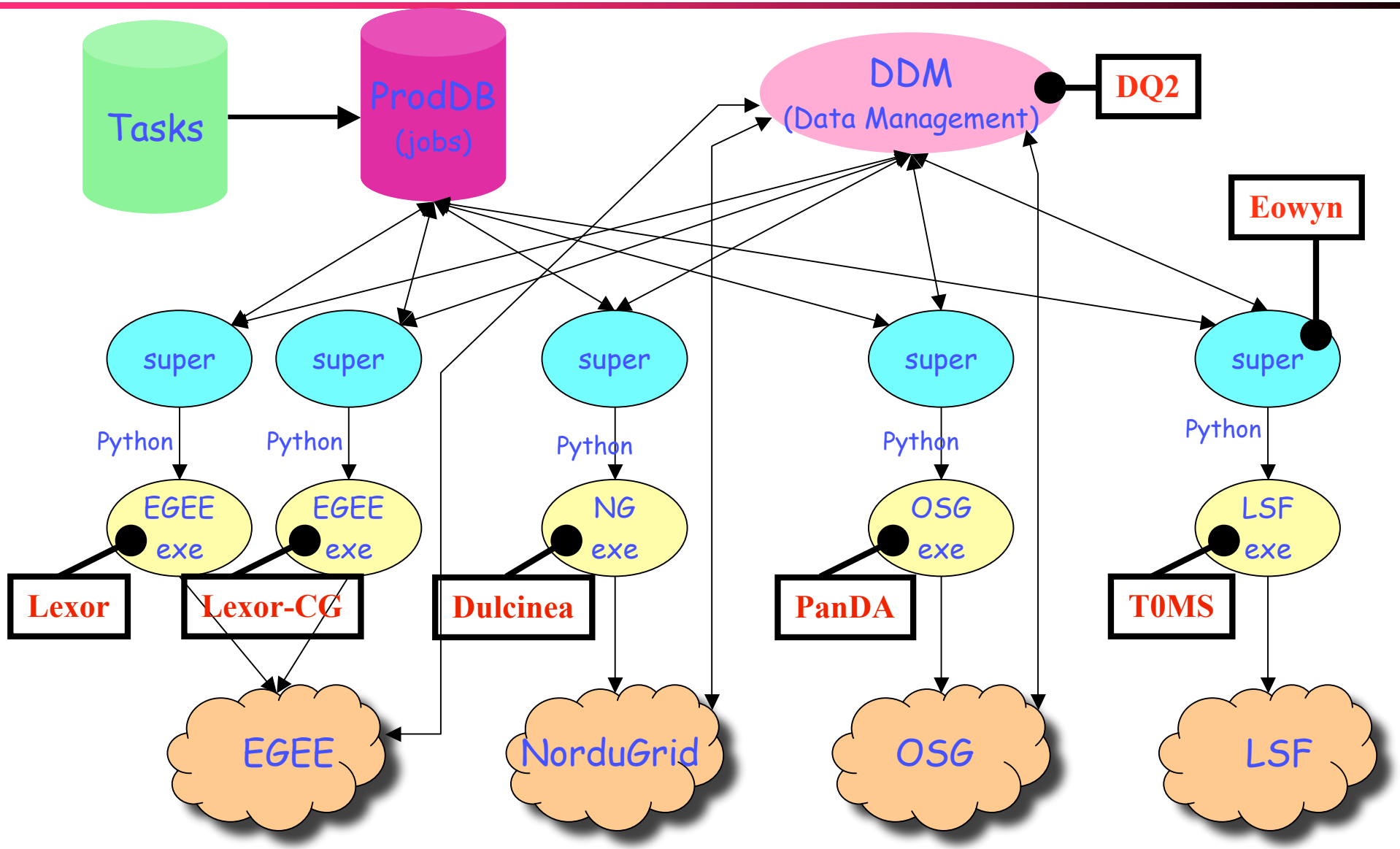
Data Management Considerations

- 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 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

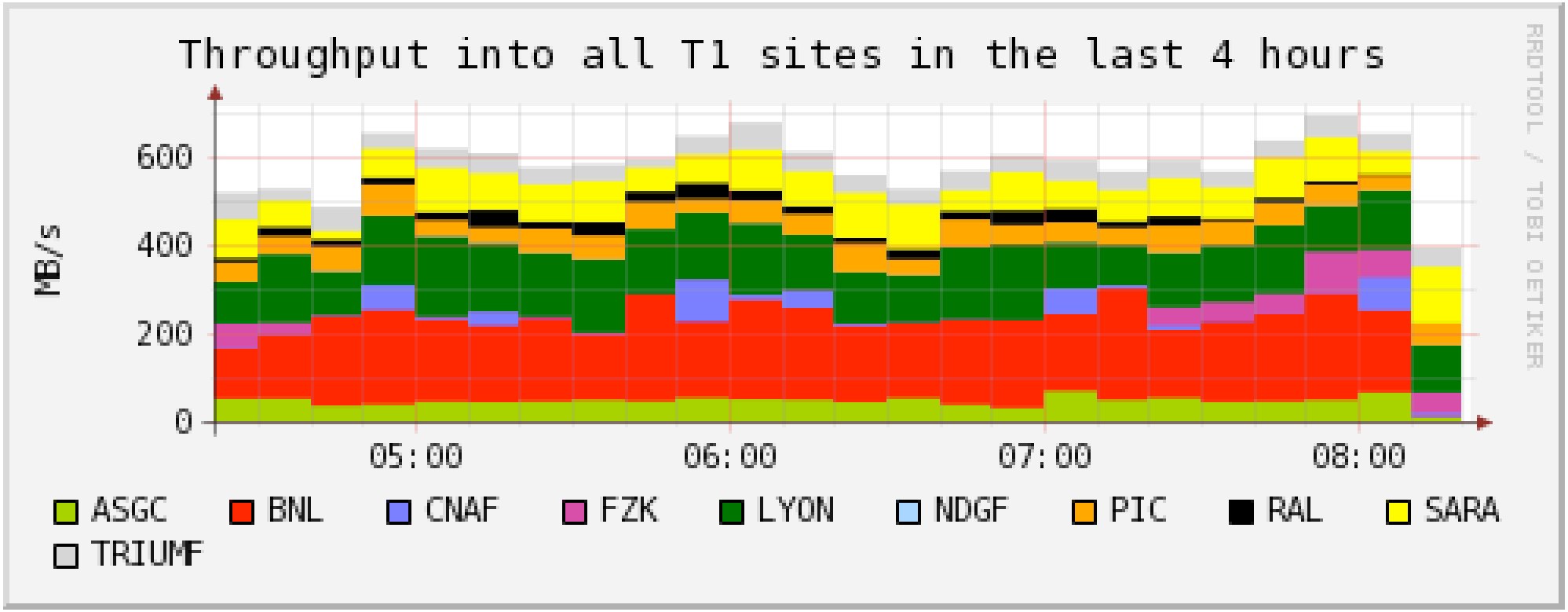
- Analysis jobs must run where the input data files are
  - As transfering data files from other sites may take longer than actually running the job
- Most analysis jobs will take AODs as input for complex calculations and event selections
  - And most likely will output Athena-Aware Ntuples (AAN, to 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

ATLAS SC4 tests

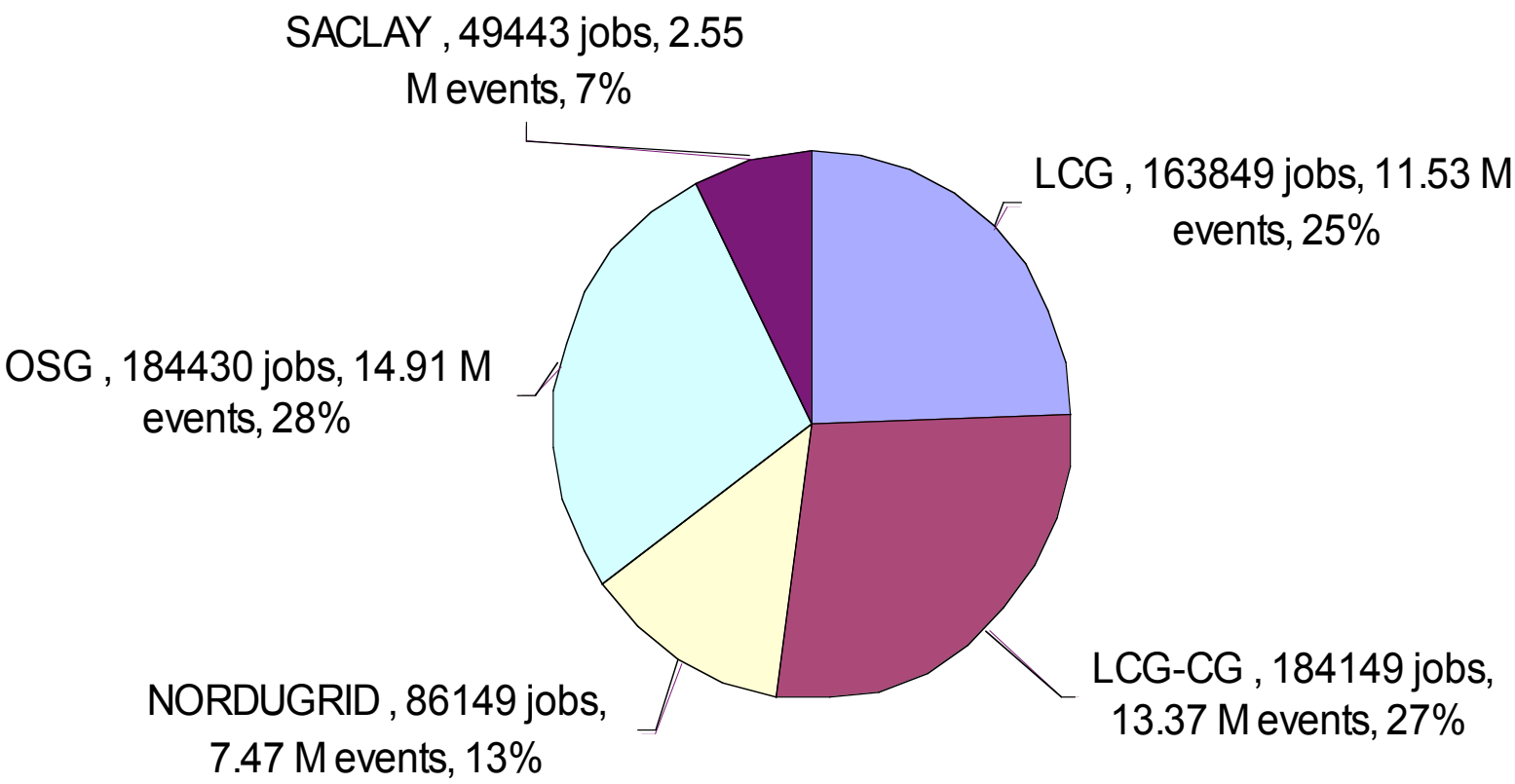
- March-April (pre-SC4): 3-4 weeks in for internal Tier-0 tests (Phase 0)
- April-May (pre-SC4): tests of distributed operations on a "small" testbed
- Last 3 weeks of June: Tier-0 test (Phase 1) with data distribution to Tier-1s
- 3 weeks in July: distributed processing tests (Part 1)
- 2 weeks in July-August: distributed analysis tests (Part 1)
- 3-4 weeks in September-October: Tier-0 test (Phase 2) with data to Tier-2s
- 3 weeks in October: distributed processing tests (Part 2)
- 3-4 weeks in November: distributed analysis tests (Part 2)

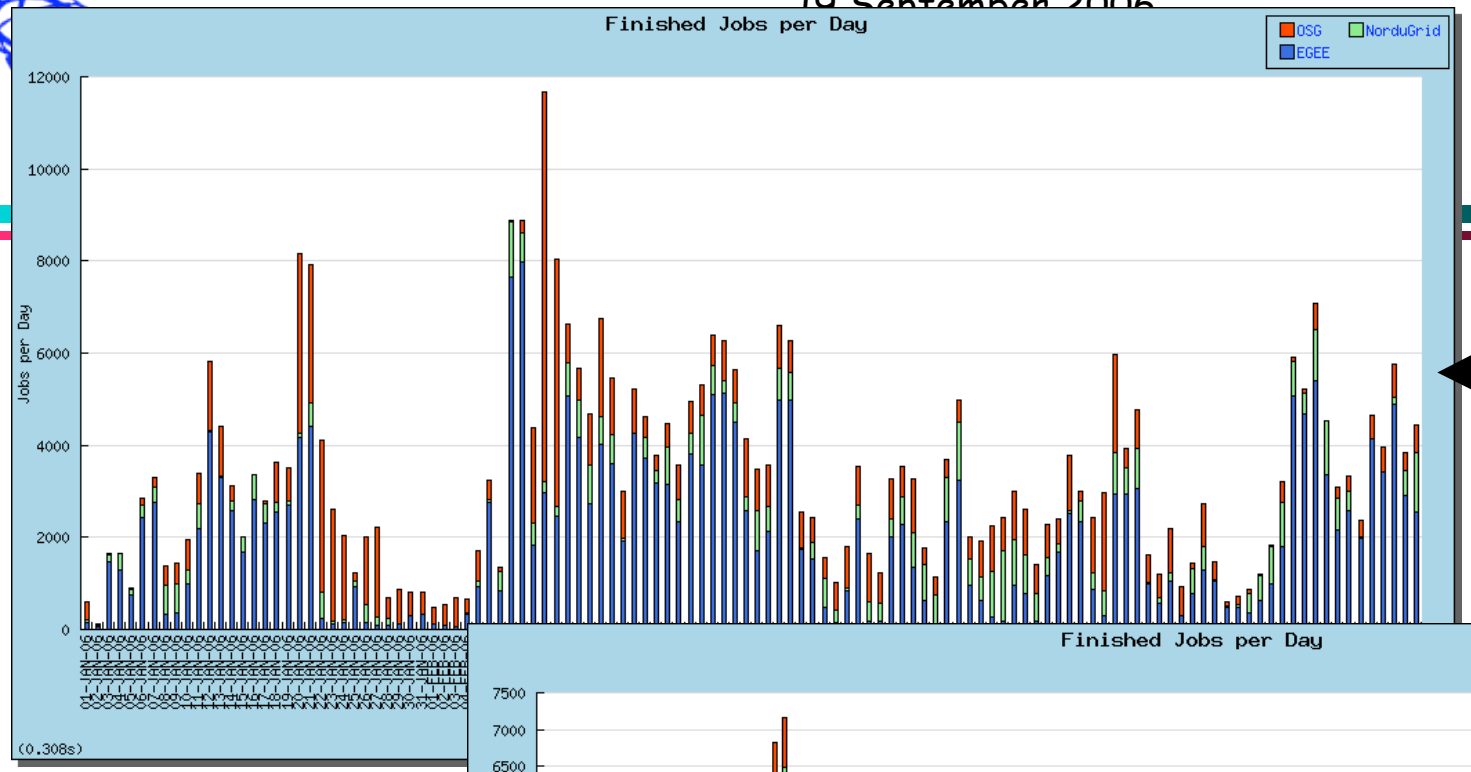
SC4 Tier-0 Data Distribution Tests

- Run a full-scale exercise, from EF, reconstruction farm, T1 export, T2 export
  - Realistic data sizes, complete flow
  - On a good day:

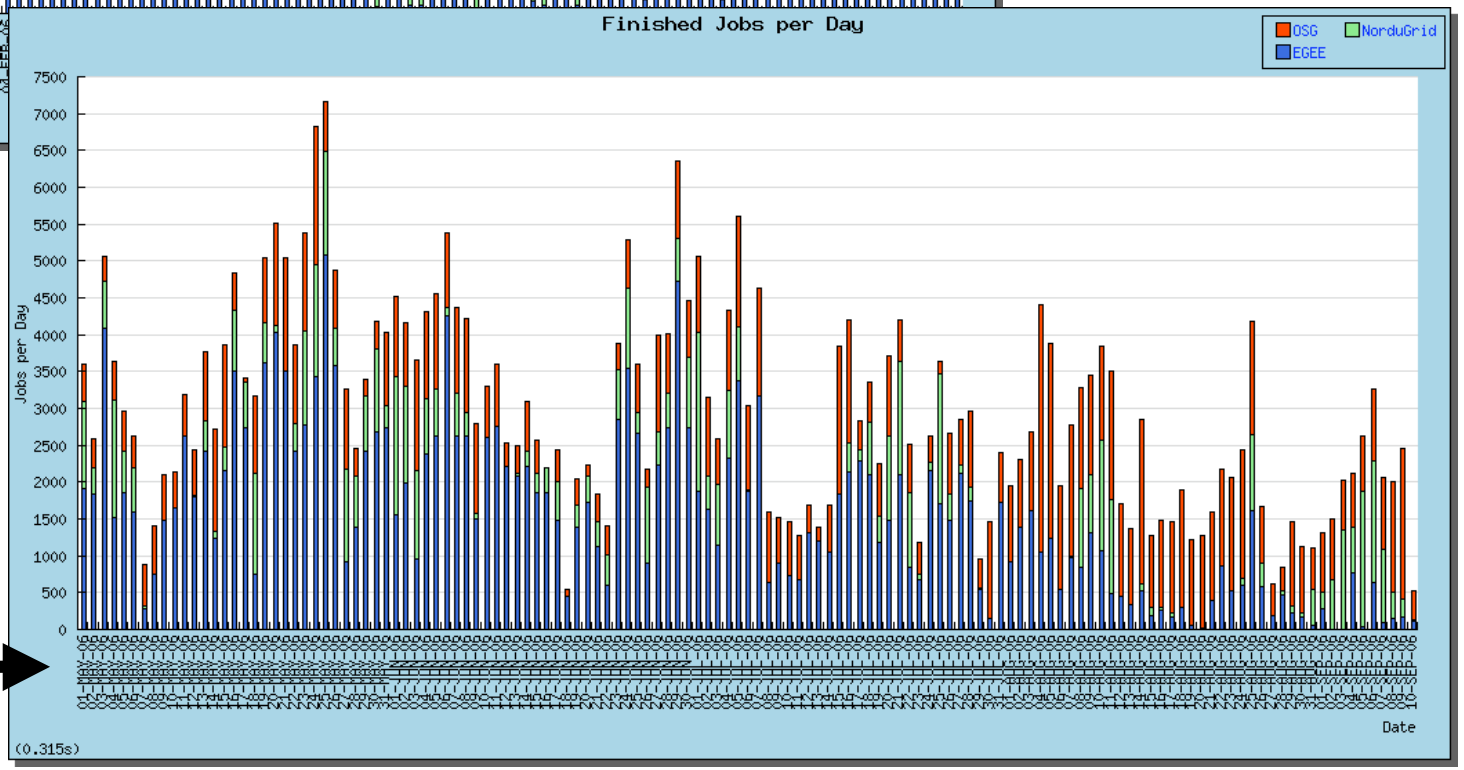


Number of successful jobs (Release 11.0.x)





Jan 1-April 30
(max 12k jobs
Reached in Feb)



May 1-now

Productions in Q4-2006

MC production for 10 M events per month				Summary						12-Sep-06	
Country	Site	Fraction at T1	Fraction at T2	CPU at T1	CPU at T2	Total	Disk	Tape	Incoming Rate from T1	from T2	Total
		%	%	kSI2k.months			TB	TB		MB/s	
Canada	TRIUMF	5.3	4.9	69	95	164	3.5	1.5	0.7	0.4	1.1
France	CC-IN2P3	13.5	14.5	175	282	458	7.7	4.2	1.3	1	2.3
Germany	FZK/GridKA	10.5	19.4	136	378	514	7.6	4.9	1.1	1.4	2.5
Italy	CNAF/INFN	7.5	8.2	97	160	257	4.7	2.4	0.9	0.6	1.5
Nordic	NDGF	5.5	1.5	71	29	101	2.9	0.9	0.8	0.1	0.9
Netherlands	SARA/NIKHEF	13	7.4	169	144	313	6.2	2.9	1.3	0.5	1.8
Spain	PIC	5.5	6.3	71	123	194	3.8	1.8	0.7	0.5	1.2
Taiwan	ASGC	7.7	2.5	100	49	149	3.8	3.8	0.9	0.2	1.1
UK	RAL	7.5	15.4	97	300	397	6	3.7	0.9	1.1	2
USA	BNL	24	19.9	312	388	699	11.8	6.5	2	1.4	3.4
Total		100	100	1298	1948	3246	58	32.6	10.6	7.2	17.8

40%	at Tier-1	
60%	at Tier-2	

	MB	kSI2k-s
Hits	2	800
ESD	1	40

Grid Interoperability

- Of course we assume here that all Grids recognise the ATLAS VO as defined in the VOMS database and ancillary tools, therefore all members of the ATLAS VO can submit jobs to all available resources, within the shares defined by internal ATLAS policies.
- The information system is clearly at the base of any interoperability possibility. If the ISs are not compatible between Grids, there is no way for any service discovery mechanism to work in an automatic way.
- We have different strategies for Production and Analysis procedures on the different Grids. Our production system is providing an additional layer which does the abstraction of different Grid infrastructures. Also we have several ways to submit analysis jobs (Ganga, Panda). Therefore interoperability in the sense that we can cross-submit jobs from one Grid to the next is for us nice to have in the medium term (mainly for analysis), but not an issue with high priority. It is important instead to have efficient plugins for ProdSys, Ganga and Panda.
- Better interoperability in terms of (CPU and storage) resource allocation, monitoring and accounting is instead a real necessity. There is at the moment also no consistent way to allocate job priorities or storage areas to different groups/roles within the VO. Compatible accounting is essential.
- We have a strong need for interoperability on the data management level. This includes components as Storage Elements with SRM interfaces, data catalogues, FTS and the like. Here interoperability is fundamental for us. We have in particular to be able to transfer data from our production to the sites where we want to analyze them. There are different issues in different Grids, but these items have to be followed up with high priority.
- SRM: work is ongoing in this area with contributions from the various Grid providers so we do not expect any major problems. Nonetheless, it is important that the deployment of SRM-enabled storages on all sites proceeds as fast as possible.
- FTS: ATLAS is deploying FTS on both EGEE and OSG (that is, US Tier-2s and Tier-1 have their own FTS server and channels defined just like EGEE sites). We are not sure this is actually realised by others. The issue with FTS interoperability is on the information system. FTS has a 'neutral' plugin for information systems, but more advanced FTS functionality (e.g. service discovery) requires a compatible information system.

Conclusions

- NorduGrid/NDGF has been since several years a very important component of the ATLAS distributed computing infrastructure
  - Its contribution to scheduled production so far always exceeded its nominal share
- The NDGF Tier-1 will contribute to the global ATLAS computing capacity and to the efficiency of data access and analysis
- It is essential that data transfer tests be set up soon between CERN and NDGF, and between NDGF and the other Tier-1s
- It is also essential that the ARC middleware be made to interoperate seamlessly with the other Grid infrastructures
  - Especially for authentication/authorization and data management (catalogues and data transfer services)
  - But also for job submission
- We are looking forward to a continued very close relationship between ATLAS and NDGF!