

Distributed Computing for the ATLAS Experiment

Dario Barberis

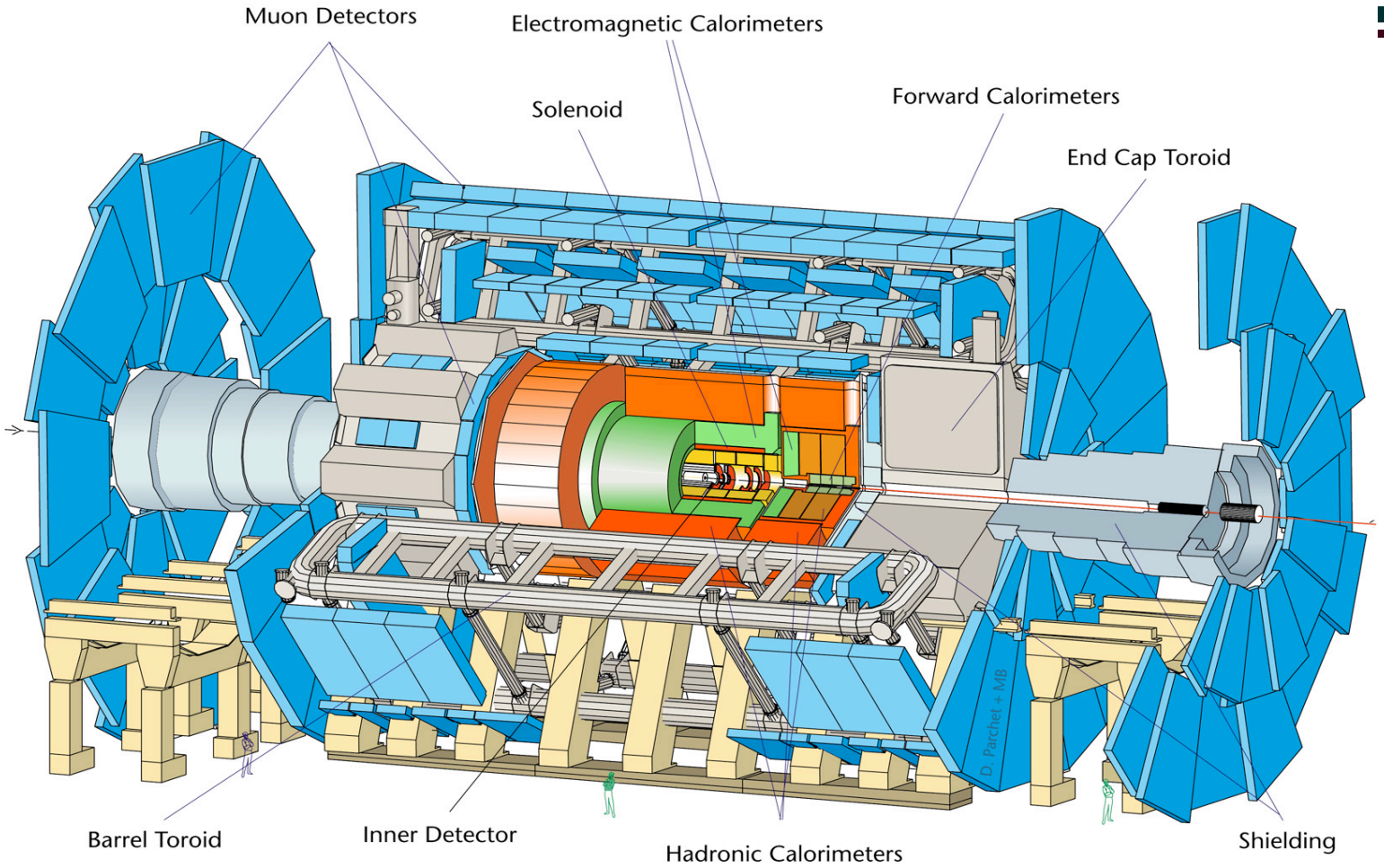
CERN & Genoa University/INFN



Outline

- The ATLAS experiment
- Computing and data model
- Data access for analysis
- Grid tools
- Analysis workflow
- Local facilities for analysis
- Reality checks
- Conclusions (work in progress)

Main detector systems



Diameter	***25 m***
Barrel toroid length	***26 m***
Endcap end-wall chamber span	***46 m***
Overall weight	***7000 Tons***

ATLAS Collaboration

IFI

(Status October 2007)

6 Continents
37 Countries
167 Institutions
2000 Scientific Authors total
(1600 PhD physicists, 400 students)
1000 Technical and support staff



Albany, Alberta, NIKHEF Amsterdam, Ankara, LAPP Annecy, Argonne NL, Arizona, UT Arlington, Athens, NTU Athens, Baku, IFAE Barcelona, Belgrade, Bergen, Berkeley LBL and UC, HU Berlin, Bern, Birmingham, Bogota, Bologna, Bonn, Boston, Brandeis, Bratislava/SAS Kosice, Brookhaven NL, Buenos Aires, Bucharest, Cambridge, Carleton, Casablanca/Rabat, CERN, Chinese Cluster, Chicago, Chilean Cluster (Santiago+Valparaiso), Clermont-Ferrand, Columbia, NBI Copenhagen, Cosenza, AGH UST Cracow, IFJ PAN Cracow, DESY, Dortmund, TU Dresden, JINR Dubna, Duke, Frascati, Freiburg, Geneva, Genoa, Giessen, Glasgow, Göttingen, 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, Nagoya, Naples, New Mexico, New York, 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, Regina, Ritsumeikan, UFRJ Rio de Janeiro, Rome I, Rome II, Rome III, Rutherford Appleton Laboratory, DAPNIA Saclay, Santa Cruz UC, Sheffield, Shinshu, Siegen, Simon Fraser Burnaby, SLAC, Southern Methodist Dallas, NPI Petersburg, Stockholm, KTH Stockholm, Stony Brook, Sydney, AS Taipei, Tbilisi, Tel Aviv, Thessaloniki, Tokyo ICEPP, Tokyo MU, Toronto, TRIUMF, Tsukuba, Tufts, Udine/ICTP, Uppsala, Urbana UI, Valencia, UBC Vancouver, Victoria, Washington, Weizmann Rehovot, FH Wiener Neustadt, Wisconsin, Wuppertal, Yale, Yerevan

ATLAS Data Collection

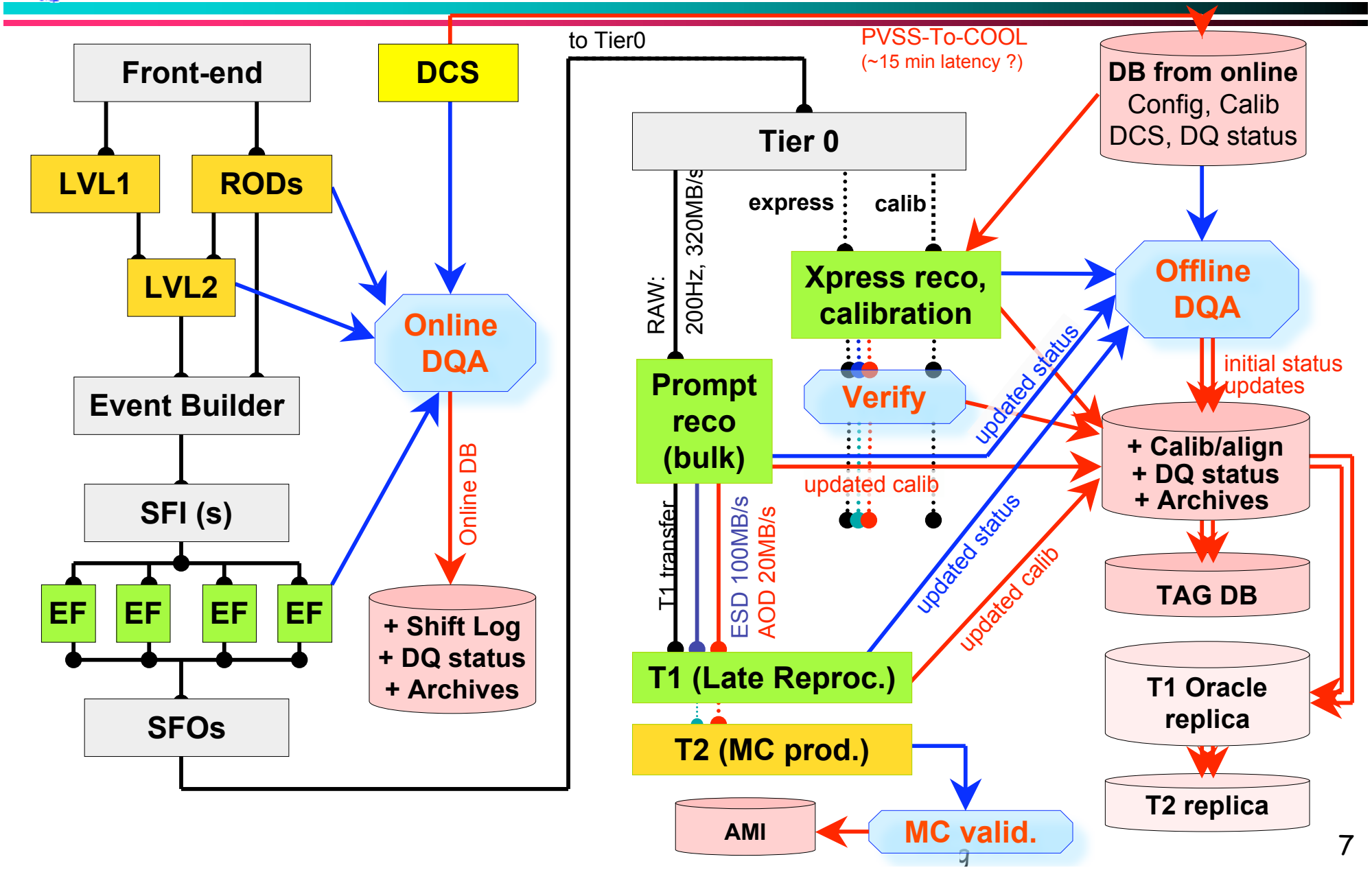
- Protons flying in opposite directions will collide with a centre-of-mass energy of 14 TeV (~14000 times the proton rest mass) in the centre of the ATLAS detector
- Each such collision produces several hundreds of particles that are absorbed and detected by the ATLAS detector
- The ensemble of the electronic 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 flow from online to offline

- 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 to be processed immediately
    - ~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 transferred to the Tier-0 input buffer at 320 MB/s (average)

Data flow from DAQ to offline



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 1 MB/event initially, to decrease as the understanding of the detector improves
    - Compromise between "being able to do everything on the ESD" and "not enough disk space to store too large events"
- AOD (Analysis Object Data):
  - Summary of event reconstruction with "physics" (POOL/ROOT) objects:
    - electrons, muons, jets, etc.
  - Nominal size 100 kB/event (now 200 kB/event including MC truth)
- DPD (Derived Physics Data):
  - Skimmed/slimmed/thinned events + other useful "user" data derived from AODs and conditions data
  - Nominally 10 kB/event on average
    - Large variations depending on physics channels
- TAG:
  - Database (or ROOT files) used to quickly select events in AOD and/or ESD files

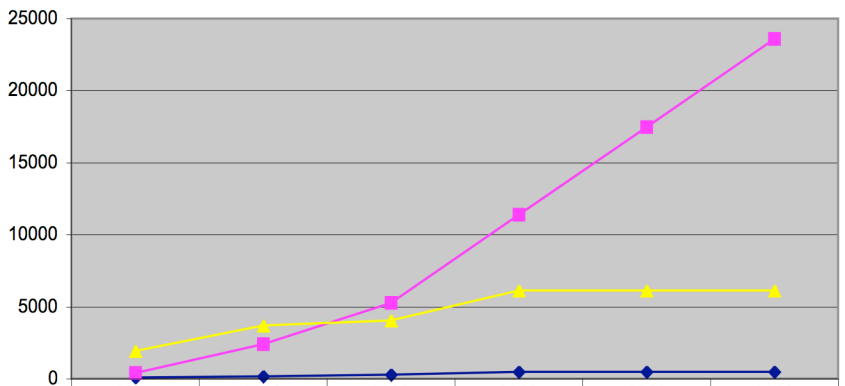


Necessity of Distributed Computing?

- We are going to collect raw data at 320 MB/s for 50k seconds/day and ~100 days/year
  - RAW data: 1.6 PB/year
- Processing (and re-processing) these events will require ~10k CPUs full time the first year of data-taking, and a lot more in the future as data accumulate
- Reconstructed events will also be large, as people want to study detector performance as well as do physics analysis using the output data
  - ESD data: 1.0 PB/year, AOD data: 0.1 PB/year
- At least 10k CPUs are also needed for continuous simulation production of at least 30% of the real data rate and for analysis
- There is no way to concentrate all needed computing power and storage capacity at CERN
  - The LEP model will not scale to this level
- The idea of distributed computing, and later of the computing grid, became fashionable at the turn of the century and looked promising when applied to HEP experiments' computing needs

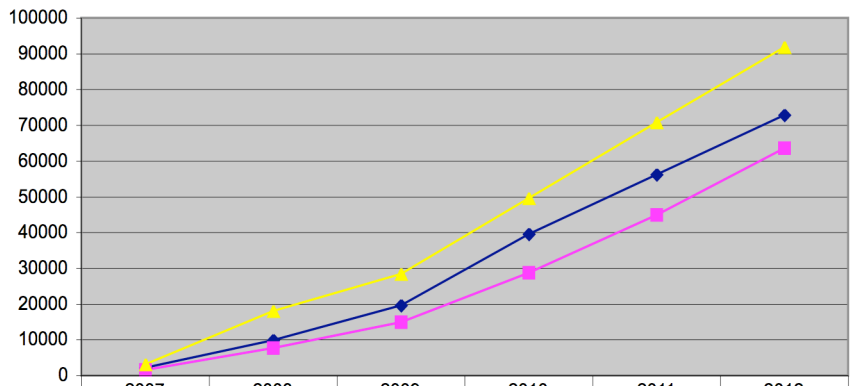
Resource Need Evolution

T0 Evolution



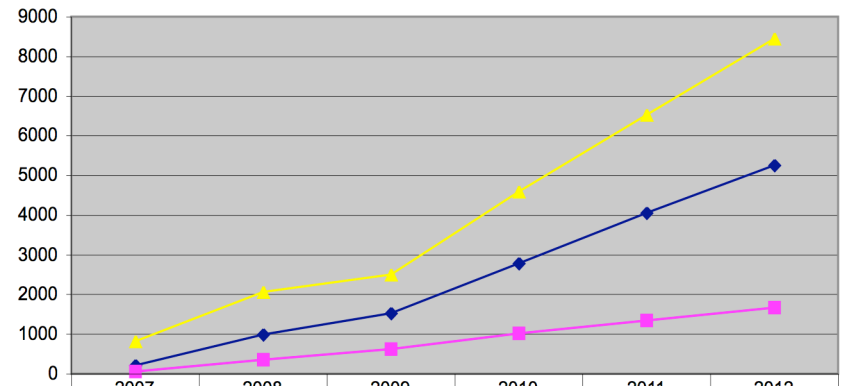
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	75.14785714	152.4621429	277.3242857	472.3528571	472.3528571	472.3528571
Total Tape (TB)	381.3075	2381.711	5267.2345	11371.158	17475.0815	23579.005
Total CPU (kSI2k)	1910	3705	4058	6105	6105	6105

T1 Evolution



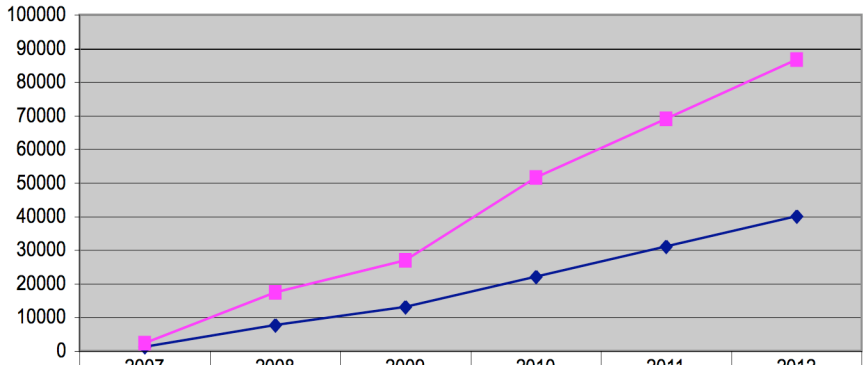
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	2157.0332	9938.696929	19686.41793	39487.79764	56190.82307	72893.8485
Total Tape (TB)	1543.186667	7693.996427	14949.57676	28698.0172	44929.67775	63644.55841
Total CPU (kSI2k)	3173.323529	18122.83529	28423.02353	49573.22353	70723.42353	91873.62353

CAF Evolution



	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

T2 Evolution



	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

Resources for Analysis (2008)

CPU share	Tier-1s	Tier-2s	CAF
Simulation	20%	33%	-
Reprocessing	20%	-	10%
Analysis	60%	67%	90%

DISK share	Tier-1s	Tier-2s	CAF
RAW	10%	1%	25%
ESD	55%	35%	30%
AOD	25%	25%	20%
DPD	10%	39%	25%

A Hierarchical Model

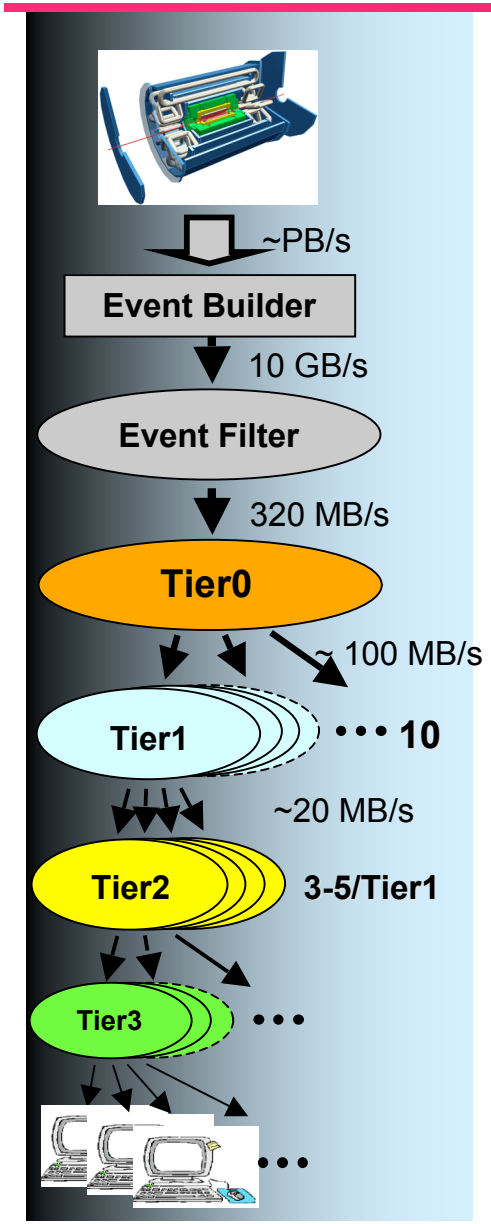
- Even before defining exactly what the Grid is, and what we can do with it, we can define a hierarchical computing model that optimises the use of our resources
  - Not all computing centres are of equal size nor offer the same service levels
  - We need to distribute RAW data to have 2 safely archived copies
  - We must distribute data for analysis and also for reprocessing
  - We must produce simulated data all the time
  - We must replicate the most popular data formats in order to make access for analysis as easy as possible for all members of the Collaboration
- The ATLAS Distributed Computing hierarchy:
  - 1 Tier-0 centre: CERN
  - 10 Tier-1 centres:

    BNL (Brookhaven, US), NIKHEF/SARA (Amsterdam, NL),
    CC-IN2P3 (Lyon, FR), FZK (Karlsruhe, DE), RAL (Chilton, UK),
    PIC (Barcelona, ES), CNAF (Bologna, IT), NDGF (DK/SE/NO),
    TRIUMF (Vancouver, CA), ASGC (Taipei, TW)
  - 35 Tier-2 centres, some of them distributed, in almost all participating countries
  - Tier-3 centres in all participating institutions

Computing Model: main operations

- Tier-0:
  - Copy RAW data to CERN Castor Mass Storage System tape for archival
  - Copy RAW data to Tier-1s for storage and subsequent 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
  - Run large-scale event selection and analysis jobs
- Tier-2s:
  - Run simulation (and calibration/alignment when/where appropriate)
  - Keep current versions of AODs on disk for analysis
  - Run analysis jobs
- Tier-3s:
  - Provide access to Grid resources and local storage for end-user data
  - Contribute CPU cycles for simulation and analysis

Data replication and distribution



In order to provide a reasonable level of data access for analysis, it is necessary to replicate the ESD, AOD and TAGs to Tier-1s and Tier-2s.

RAW:

- Original data at Tier-0
- Complete replica distributed among all Tier-1
  - Randomized datasets to make reprocessing more efficient

ESD:

- ESDs produced by primary reconstruction reside at Tier-0 and are exported to 2 Tier-1s
- Subsequent versions of ESDs, produced at Tier-1s (each one processing its own RAW), are stored locally and replicated to another Tier-1, to have globally 2 copies on disk

AOD:

- Completely replicated at each Tier-1
- Partially replicated to Tier-2s (~1/3 – 1/4 in each Tier-2) so as to have at least a complete set in the Tier-2s associated to each Tier-1
  - Every Tier-2 specifies which datasets are most interesting for their reference community; the rest are distributed according to capacity

TAG:

- TAG databases are replicated to all Tier-1s (Oracle and ROOT files)
- Partial replicas of the TAG will be distributed to Tier-2 as ROOT files
  - Each Tier-2 will have at least all ROOT files of the TAGs that correspond to the AODs stored there

Samples of events of all types can be stored anywhere, compatibly with available disk capacity, for particular analysis studies or for software (algorithm) development.

Pre-Grid: LHC Computing Models

- In 1999-2000 the "LHC Computing Review" analyzed the computing needs of the LHC experiments and built a hierarchical structure of computing centres: Tier-0, Tier-1, Tier-2s, Tier-3s…
  - Every centre would have been connected rigidly only to its reference higher Tier and its dependent lower Tiers
  - Users would have had login rights only to "their" computing centres, plus some limited access to higher Tiers in the same hierarchical line
  - Data would have been distributed in a rigid way, with a high level of progressive information reduction along the chain
- This model could have worked, although with major disparities between members of the same Collaboration depending on their geographical location
- The advent of Grid projects in 2000-2001 changed this picture substantially
  - The possibility of sharing resources (data storage and CPU capacity) blurred the boundaries between the Tiers and removed geographical disparities
  - The computing models of the LHC experiments were revised to take these new possibilities into account



Pre-Grid: HEP Work Models

- The work model of most HEP physicists did not evolve much during the last 20 years:
  - Log into a large computing centre where you have access
  - Use the local batch facility for bulk analysis
  - Keep your program files on a distributed file system (usually AFS or NFS)
  - Have a sample of data on group/project space on disk (also on AFS or NFS)
  - Access the bulk of the data in a mass storage system ("tape") through a staging front-end disk cache
- Therefore the initial expectations for a Grid system were rather simple:
  - Have a "Grid login" to gain access to all facilities from the home computer
  - Have a simple job submission system ("gsub" instead of "bsub"…)
  - List, read, write files anywhere using a Grid file system (seen as an extension of AFS)
- As we all know, all this turned out to be much easier said than done!
  - E.g., nobody in those times even thought of asking questions such as "what is my job success probability?" or "shall I be able to get my file back?"…



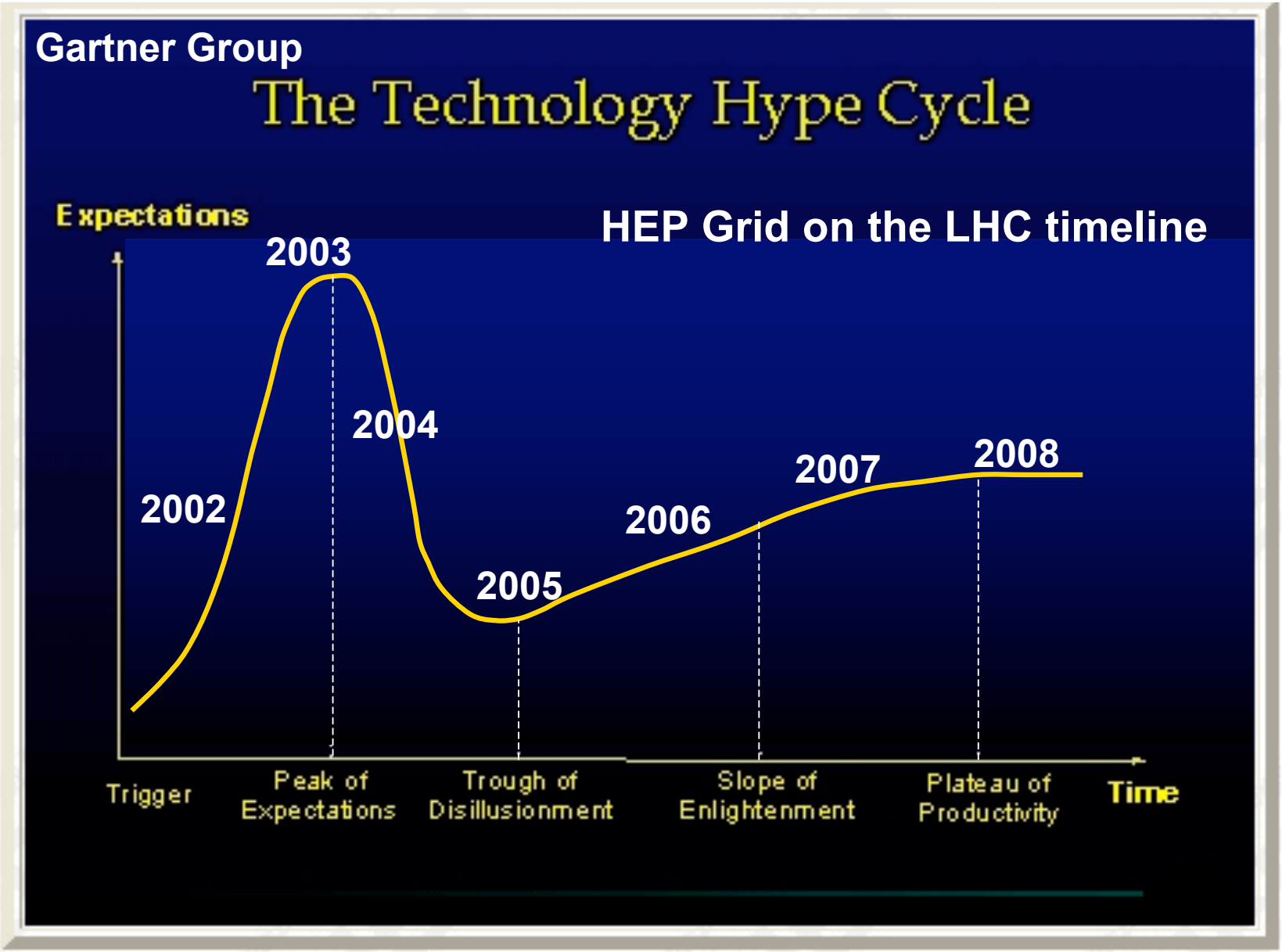
First Grid Deployments

- In 2003-2004, the first Grid middleware suites were deployed on computing facilities available to HEP (LHC) experiments
  - NorduGrid (ARC) in Scandinavia and a few other countries
  - Grid3 (VDT) in the US
  - LCG (EDG) in most of Europe and elsewhere (Taiwan, Canada…)
- The LHC experiments were immediately confronted with the multiplicity of m/w stacks to work with, and had to design their own interface layers on top of them
  - Some experiments (ALICE, LHCb) chose to build a thick layer that uses only the lower-level services of the Grid m/w
  - ATLAS chose to build a thin layer that made maximal use of all provided Grid services (and provided for them where they were missing, e.g. job distribution in Grid3)

Communication Problems?

- Clearly both the functionality and performance of first Grid deployments fell rather short of the expectations:
  - VO Management:
    - Once a person has a Grid certificate and is a member of a VO, he/she can use ALL available processing and storage resources
      - And it is even difficult *a posteriori* to find out who did it!
    - No job priorities, no fair share, no storage allocations, no user/group accounting
    - Even VO accounting was unreliable (when existing)
  - Data Management:
    - No assured disk storage space
    - Unreliable file transfer utilities
    - No global file system, but central catalogues on top of existing ones (with obvious synchronization and performance problems…)
  - Job Management:
    - No assurance on job execution, incomplete monitoring tools, no connection to data management
    - For the EDG/LCG Resource Broker (the most ambitious job distribution tool), very high dependence the correctness of ALL site configurations

Disillusionment?





Realism

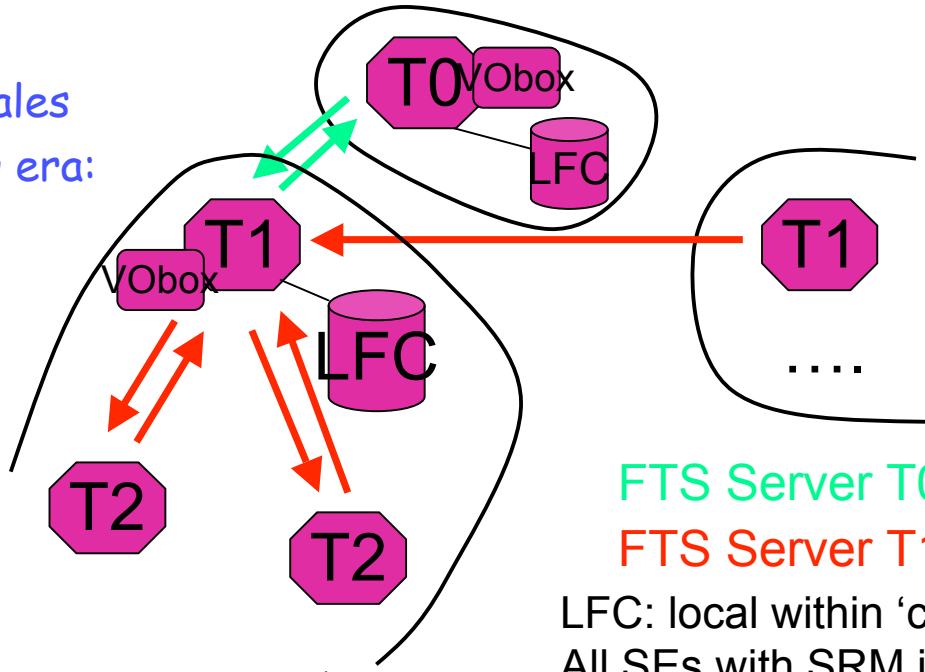
- After the initial experiences, all experiments had to re-think their approach to Grid systems
  - Reduce expectations
  - Concentrate on the absolutely necessary components
  - Build the experiment layer on top of those
  - Introduce extra functionality only after thorough testing of new code
- The LCG Baseline Services Working Group in 2005 defined the list of high-priority, essential components of the Grid system for HEP (LHC) experiments
  - VO management
  - Data management system
    - Uniform definitions for the types of storage
    - Common interfaces
    - Data catalogues
    - Reliable file transfer system

ATLAS Grid Architecture

- The ATLAS Grid architecture is based on 4 main components:
  - Distributed Data Management (DDM)
  - Distributed Production System (ProdSys)
  - Distributed Analysis (DA)
  - Monitoring and Accounting
- DDM is the central link between all components
  - As data access is needed for any processing and analysis step!
- In 2005 there was a global re-design of ProdSys and DDM to address the shortcomings of the Grid m/w, and allow easier access to the data for distributed analysis
  - At the same time, the first implementations of DA tools were developed
- The new DDM design is 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)

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





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 late 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 * 0.3 * 200 Hz * 2.6 MB = 4.7 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

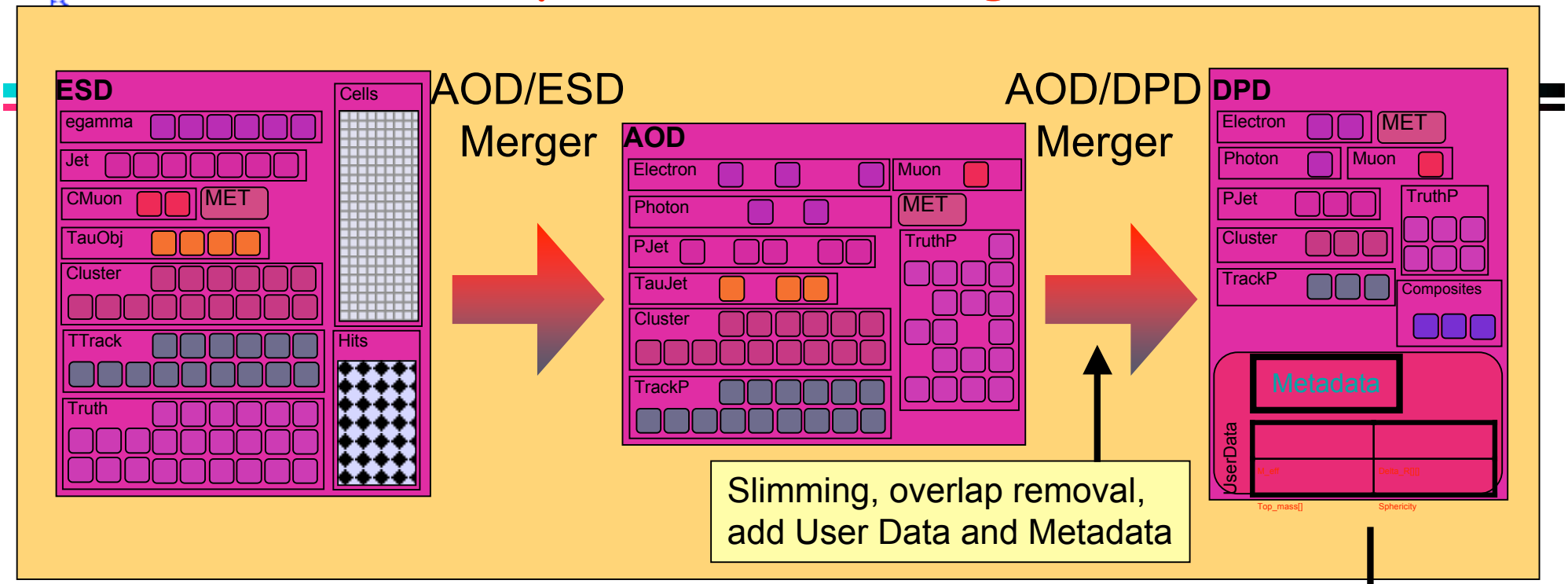
- Once we have data distributed in the correct way (rather than sometimes hidden in the guts of automatic mass storage systems), we can rework the distributed production system to optimise job distribution, by sending jobs to the data (or as close as possible to them)
  - This was not the case previously, 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
- Further improvements (end 2007 – early 2008): use pilot jobs to decrease the dependence on misconfigured sites or worker nodes
  - Pilot jobs check the local environment before pulling in the payload



Analysis Data Formats

- Evolving view of what Derived Physics Datasets (DPD) are
  - In the Computing TDR (2005), they used to represent many derivations
    - Skimmed AOD, data collections, augmented AOD, other formats (Athena-aware Ntuples, root-tuples)
  - Much effort was invested to see if one format can cover most needs
    - Saves resources
    - But diversity will remain
    - 'Everyone ends-up with a flat n-tuple'?
  - In each case, the aim is to be faster, smaller and more portable
- Group-level DPDs have to be produced in scheduled activity at Tier 1s
  - 'Big trains' becomes small meetings!
  - Overall co-ordinator and production people in each group
- User-level DPDs can be produced at Tier-2s
  - And brought "home" to Tier-3s or desk/lap-tops if small enough
- The conclusion of many discussions last year in the context of the Analysis Forum is that DPDs will consist (for most analyses) of skimmed/slimmed/thinned AODs plus relevant blocks of computed quantities (such as invariant masses)
  - Stored in the same format as ESD and AOD
  - Therefore readable both from Athena and from ROOT (using the AthenaRootAccess library)

The Analysis Model is being finalized



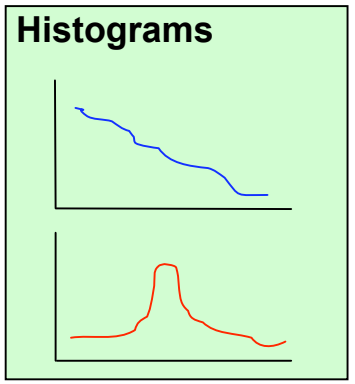
This model is very elegant and "clean" since it allows:

- the same code to run at all levels, from primary AOD to DPD
- seamless code porting Athena ⇔ ROOT
- the same analysis to be performed in different frameworks (Athena batch, Athena interactive, ROOT)

However, it is a viable solution only if:

- no big CPU penalty compared to e.g. a flat nutple
- adequate/positive user feedback

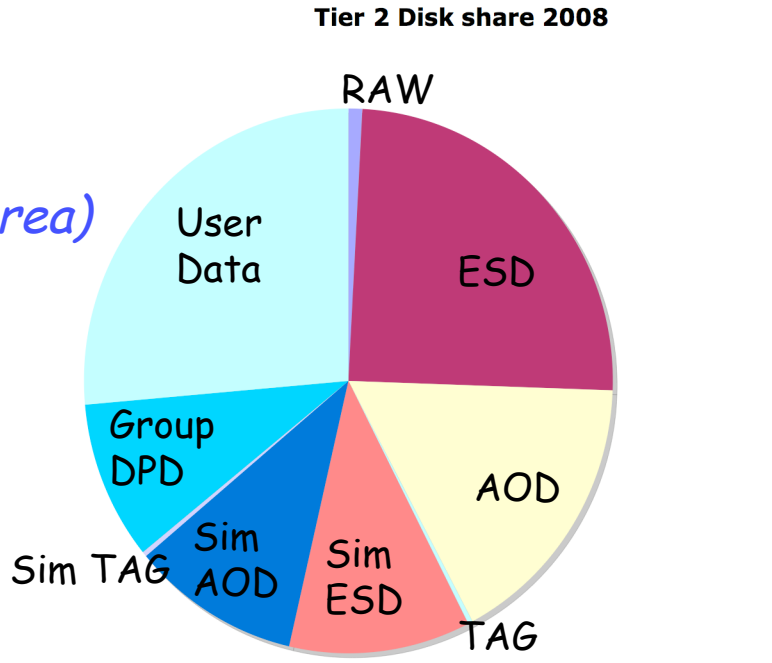
New DPDs will be part of forthcoming productions; people will be strongly 'pushed' to use them



Tier-2 Data on Disk

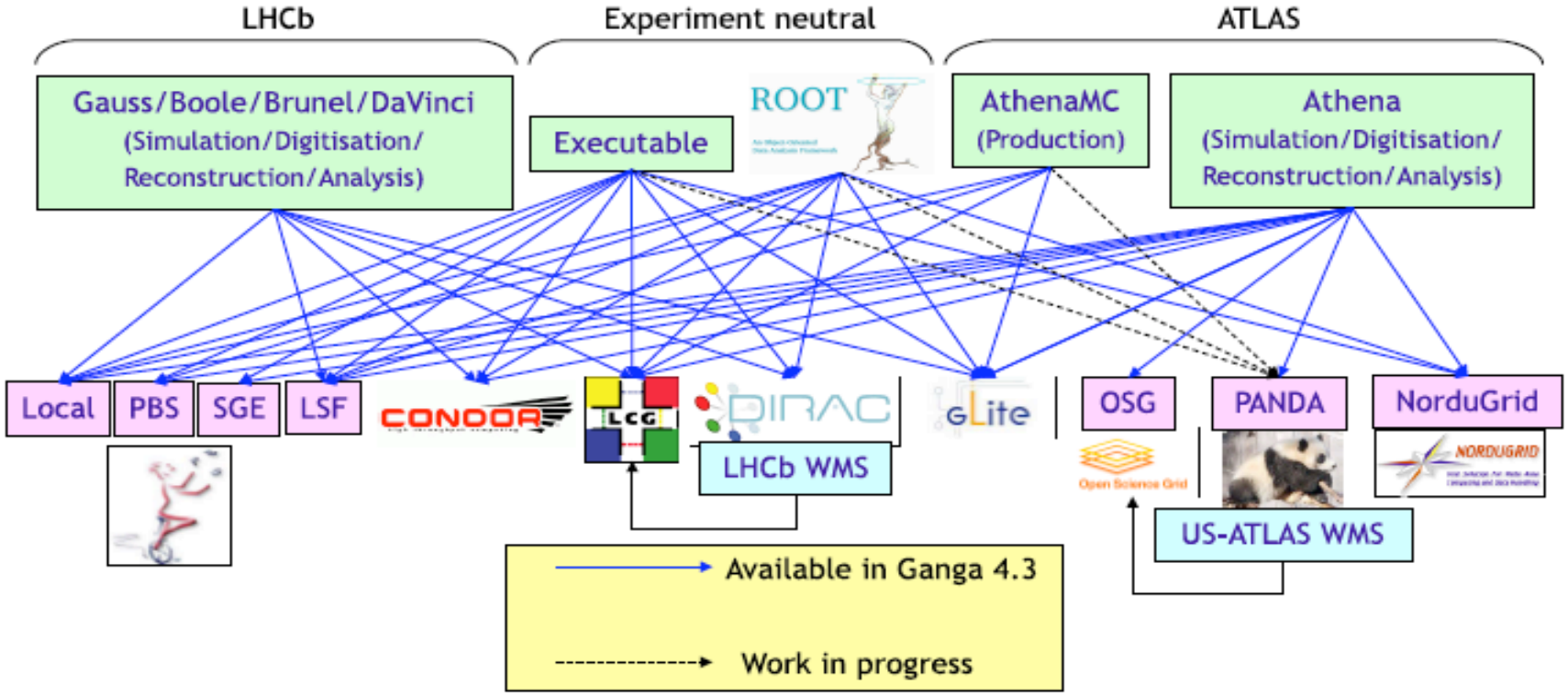
~35 Tier-2 sites of very, very different size contain:

- Some fraction of ESD and RAW
  - In 2008: 30% of RAW and 150% of ESD in Tier-2 cloud
  - In 2009 and after: 10% of RAW and 30% of ESD in Tier-2 cloud
  - This will largely be 'pre-placed' in early running
  - Recall of small samples through the group production at T1
  - Additional access to ESD and RAW in CAF
    - 1/18 RAW and 10% ESD
- 10 copies of full AOD on disk
- A full set of official group DPD *(in production area)*
- Lots of small group DPD *(in production area)*
- User data
  - Access is 'on demand'

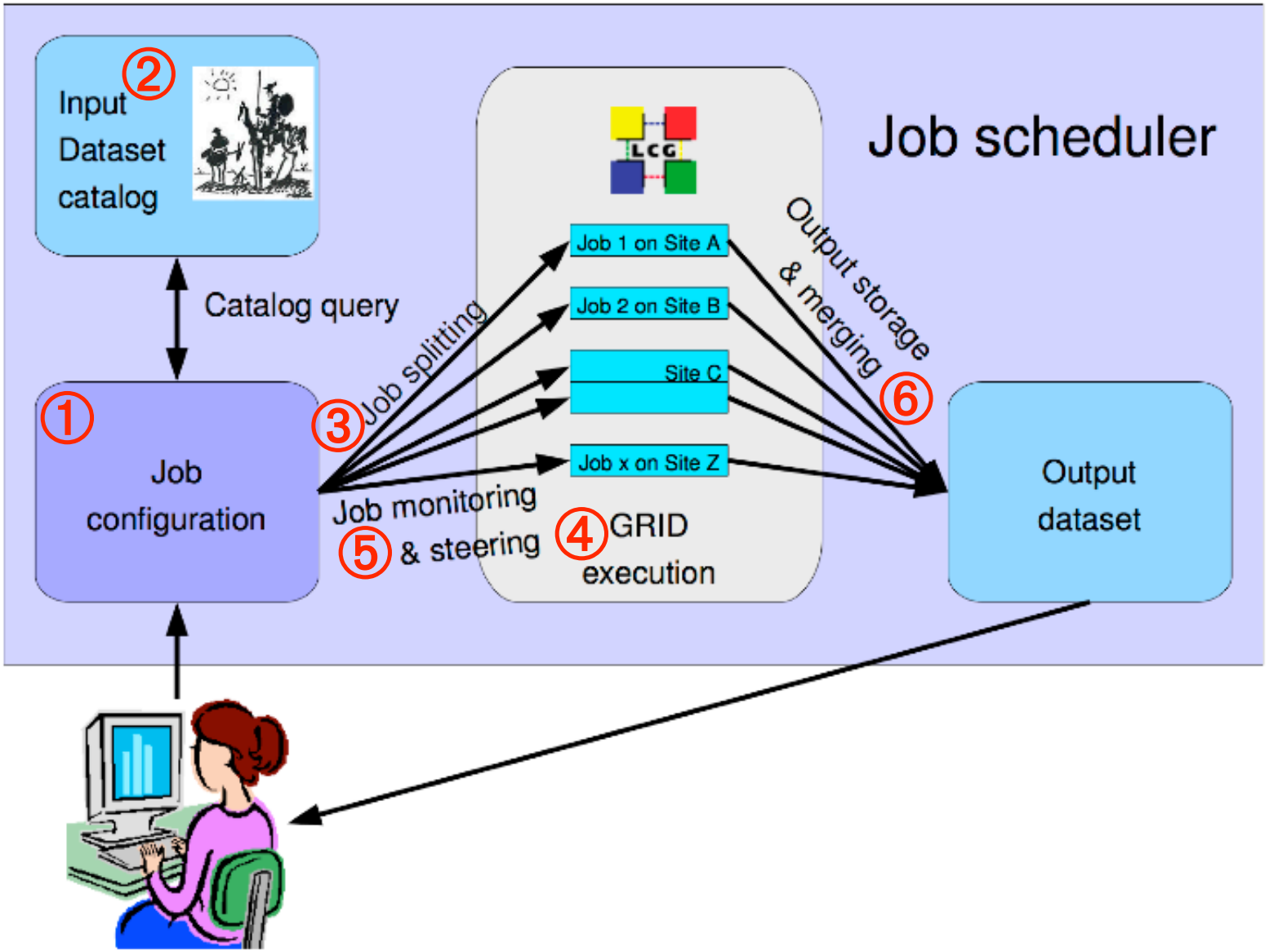


Distributed Analysis with Ganga

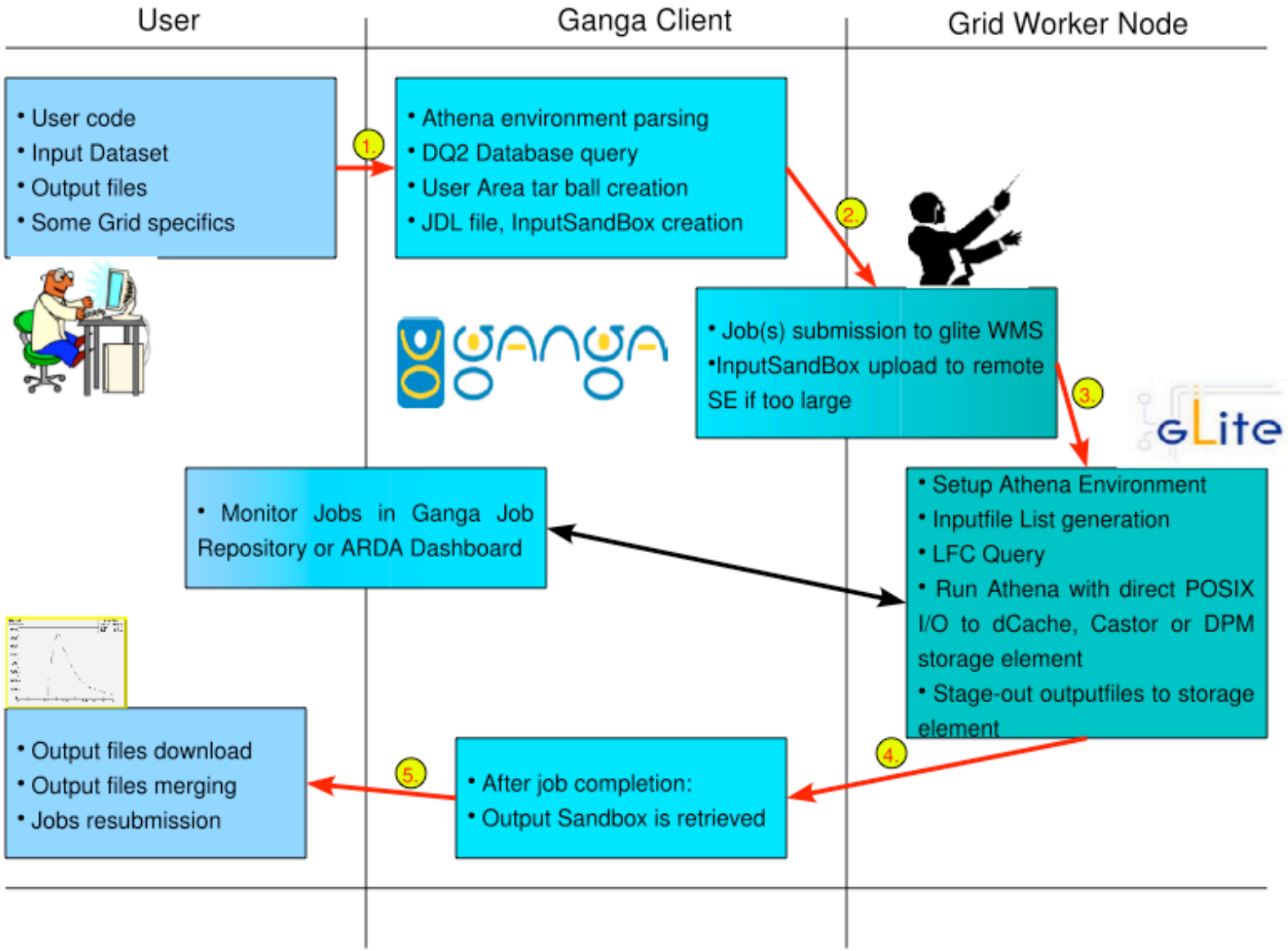
- The GANGA (Gaudi Athena aNd Grid Alliance) framework simplifies running ATLAS (and LHCb) applications on a variety of Grid and non-Grid back-ends
  - Uniform job submission to EGEE/gLite, OSG/PanDA and NorduGrid/ARC facilities is of particular interest to ATLAS collaborators
  - ATLAS production transformations, as well as analysis jobs with user code to be compiled, or ROOT-based jobs, can be submitted to the Grid after being tested on a smaller scale on local facilities



Ganga user actions

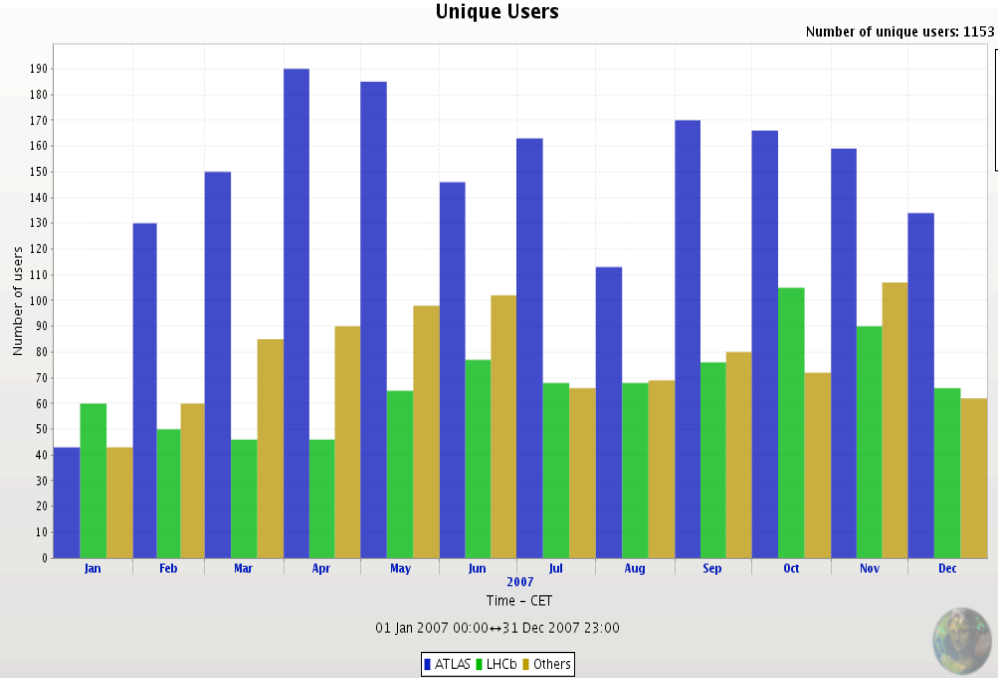


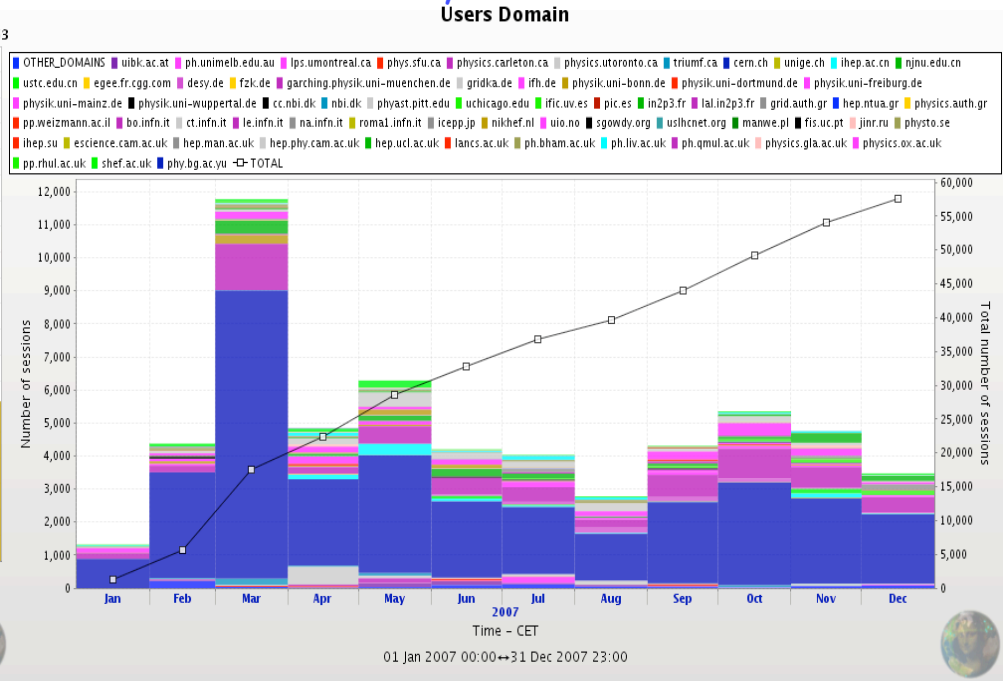
Ganga workflow



Ganga usage in ATLAS

- ATLAS end users are finally learning to use Ganga to send jobs to their input data
  - Rather than copying files to their local computing clusters and running locally, which turns out to be close to impossible
- A major improvement in job splitting and brokering has been recently implemented:
  - Use the information from DDM catalogues to split jobs along the boundaries of parts of datasets present at different sites, then direct each subjob to the location of all its input files
- More than 1000 people tried Ganga in 2007
  - >600 ATLAS, the rest LHCb and others
- ~150 regular ATLAS users
  - A sizeable fraction of the physics analysis activities
- Many people still use it from lxplus.cern.ch
  - ~50% of the ATLAS sessions since Summer 2007
  - Decreasing fraction with time
- People should really download the client and install it locally, then submit to all Grid resources from their "Tier-3" facility





Tier-3s

- These have many forms
- Basically represent resources not for general ATLAS usage
  - Some fraction of T1/T2 resources
  - Local University clusters
  - Desktop/laptop machines
  - Tier-3 task force will provide soon recommended solutions (plural!)
- Concern over the apparent belief that Tier-3s can host large samples
  - Required storage and effort, network and server loads at Tier-2s
- Network access
  - ATLAS policy in outline:
    - O(10GB/day/user) who cares?
    - O(50GB/day/user) rate throttled
    - O(10TB/day/user) user throttled!
    - Planned large movements are possible if negotiated



Minimal Tier-3 requirements

- The ATLAS software environment, as well as the ATLAS and grid middleware tools, allow us to build a work model for collaborators who are located at sites with low network bandwidth to Europe or North America.
- The minimal requirement is on local installations, which should be configured with a Tier-3 functionality:
  - A Computing Element known to the Grid, in order to benefit from the automatic distribution of ATLAS software releases
  - A SRM-based Storage Element, in order to be able to transfer data from the Grid to the local storage, and vice versa
- The local cluster should have the installation of:
  - A Grid User Interface suite, to allow job submission to the Grid
  - ATLAS DDM client tools, to permit access to the DDM data catalogues and data transfer utilities
  - The Ganga client, to allow the submission of analysis jobs to all ATLAS computing resources



Computing System Commissioning tests

- We started at the turn of the century to run "data challenges" of increasing complexity
  - Initially based on distributed simulation production
  - Using all Grid technology that was available at any point in time
    - And helping debug many of the Grid tools
- Since 2005 we set up a series of system tests that were designed to check the functionality of basic component blocks
  - Such as the software chain, distributed simulation production, data export from CERN, calibration loop, and many others
  - Collectively known as "Computing System Commissioning" (CSC) tests
- The logical continuation of the CSC tests is the complete integration test of the software and production operation tools: the FDR (Full Dress Rehearsal)
  - Next slide…

Full Dress Rehearsal and CCRC'08

- The FDR tests will take place in 2 phases, early February and May:
  - Simulated data in RAW data format are pre-loaded on the output buffers of the online computing farm and transmitted to the Tier-0 farm at nominal rate (200 Hz, 320 MB/s), mimicking the LHC operation cycle
  - Data are calibrated/aligned/reconstructed at Tier-0 and distributed to Tier-1 and Tier-2 centres, following the computing model
  - At the same time, distributed simulation production and distributed analysis activities continue, providing a constant background load
  - Reprocessing at Tier-1s will also be tested in earnest for the first time
- The February tests are the first time these operations are all tried concurrently
  - The probability that something will fail is high
- The May tests should give us the confidence that all major problems have been identified and solved
- The Common Computing Readiness Challenge (CCRC) will take place also in February and May, following the FDR tests, with all LHC experiments at the same time
  - This is mostly a load test for CERN, Tier-1s and the network

Is everything ready then?

- Unfortunately not yet: a lot of work remains
  - Thorough testing of existing software and tools
  - Optimisation of CPU usage, memory consumption, I/O rates and event size on disk
  - Completion of the data management tools (including disk space management)
  - Completion of the accounting tools (both for CPU and storage)
- Just one example (but there are many!):
  - In the computing model we foresee distributing a full copy of AOD data to each Tier-1, and an additional full copy distributed amongst all Tier-2s of a given Tier-1 "cloud"
    - In total, >20 copies around the world, as some large Tier-2s want a full set
    - This model is based on general principles to make AOD data easily accessible to everyone for analysis
  - In reality, we don't know how many concurrent analysis jobs a data server can support
    - Tests could be made submitting large numbers of grid jobs to read from the same data server
      - Results will be functions of the server type (hardware, connectivity to the CPU farm, local file system, Grid data interface) but also access pattern (all events vs sparse data in a file)
  - If we can reduce the number of AOD copies, we can increase the amount of other data samples (RAW, ESD, simulation) on disk



Conclusions

- The ATLAS Collaboration has developed a set of software and middleware tools that enable access to data for physics analysis purposes to all members of the collaboration, independently of their geographical location.
- Main building blocks of this infrastructure are
  - the Athena software framework, with its associated modular structure of the event data model;
  - the Distributed Data Management system;
  - the Ganga framework for distributed analysis on the Grid.
- Almost all required functionalities are already provided; an extensive program of field tests using simulated, as well as real data from cosmic ray events, is in place to test and commission the full system.
  - Help to complete this program before the beginning of LHC operation is very much appreciated!



Thank you!