ATLAS plans for batch system use

- Grid and local
- Steady state and startup

Layout

- The ATLAS computing model
  - The role of the Tier's
  - The usage of the CPU
  - The CPU resources
- Scenario for batch use
  - Steady state and startup
- Requirements for the batch system
  - Grid and local
- Caveat

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 4 streams are foreseen:
    - Express stream with "most interesting" events
    - Calibration events (including some physics streams, such as inclusive leptons)
    - "Trouble maker" events (for debugging)
    - Full (undivided) event stream
  - 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)
- The Tier-0 input buffer will have to hold raw data waiting for processing
  - And also cope with possible backlogs
  - ~125 TB will be sufficient to hold 5 days of raw data on disk

Computing Model: central operations

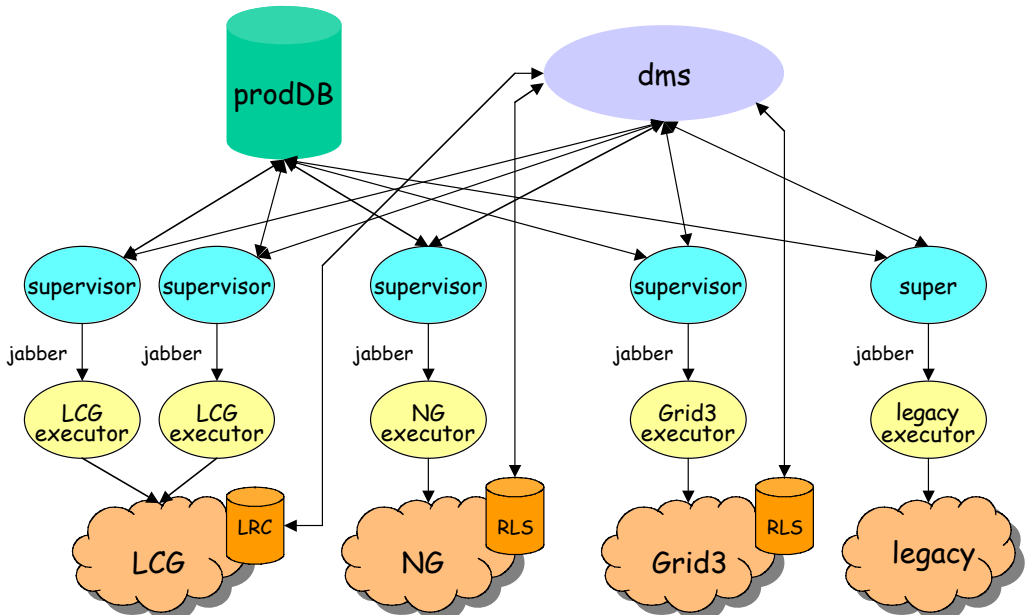
- Tier-0:
  - Copy RAW data to Castor tape for archival
  - Copy RAW data to Tier-1s for storage and reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs & TAGS) to Tier-1s
- Tier-1s:
  - 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
  - Keep current versions of AODs on disk for analysis

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 2.5 times larger: contents and technology under revision, following feedback on the first prototype implementation
- AOD (Analysis Object Data):
  - Summary of event reconstruction with "physics" (POOL/ROOT) objects:
    - electrons, muons, jets, etc.
  - Nominal size 100 kB/event
    - currently 70% of that: contents and technology under revision, following feedback on the first prototype implementation
- TAG:
  - Database used to quickly select events in AOD and/or ESD files

Distributed Production System

- ATLAS has run already several large-scale distributed production exercises
  - DC1 in 2002-2003, DC2 in 2004, "Rome Production" in 2005
  - Several tens of millions of events fully simulated and reconstructed
- It has not been an easy task, despite the availability of 3 Grids
  - DC2 and Rome prod. were run entirely on Grids (LCG/EGEE, Grid3/OSG, NorduGrid)
- A 2nd version of the distributed ProdSys is ready and being
- interfaced to DQ2
  - keeping the same architecture (Fig. V1)
    - ProdDB
    - Common supervisor
    - One executor per Grid
    - Interface to DDM (DQ2)



Distributed Analysis

- At this point emphasis on batch model to implement the ATLAS Computing model
  - Analysis = user jobs without central coordination (from test algorithms to group coordinated analysis)
- We expect our users to send large batches of short jobs to optimize their turnaround
  - Scalability
  - Data Access
- Analysis in parallel to production
  - Job Priorities, short queues, fair share… see later
- Data for analysis will be available distributed on all Tier-1 and Tier-2 centers
  - AOD & ESD: full AOD 1 year = 200 TB
  - T1 & T2 are open for analysis jobs
  - The computing model foresees about 50 % of grid resources to be allocated for analysis
- Users will send jobs to the data and extract relevant data
  - typically NTuples or similar
  - then local interactive use of these samples

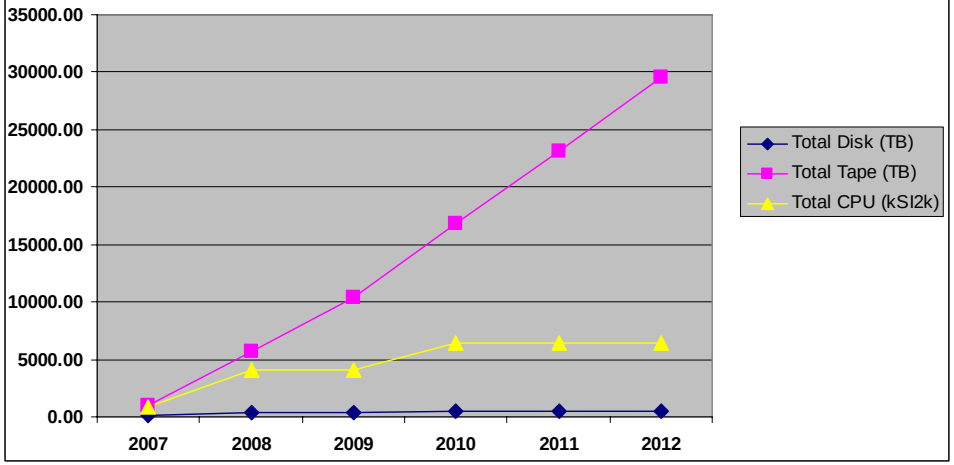
CPU batch usage

- At Tier2 about half CPU devoted to simulation and half to analysis
  - Simulation is planned just for Tier2's and in this scenario we plan to be able to simulate 20% of the events we take (i.e. we plan to simulate 40 Hertz)
  - Simulation jobs will last 10-20 hours each
  - Analysis jobs will be typically shorter
    - Not shorter than 1/2 hour, maybe 1 hour
    - We expect too short jobs make less efficient use of the Grid
- Tier1 will be shared between recosntruction, calibration, analysis
  - Share may be different at startup from steady state
  - Startup will require more calibration and more ESD analysis (and ESD are in Tier1)
  - Length in time of reconstruction jobs still to be optimized
    - 1-2 hours for simulation reconstruction, probably more for raw data reconstruction
- We assume available resources will be saturated or nearly so

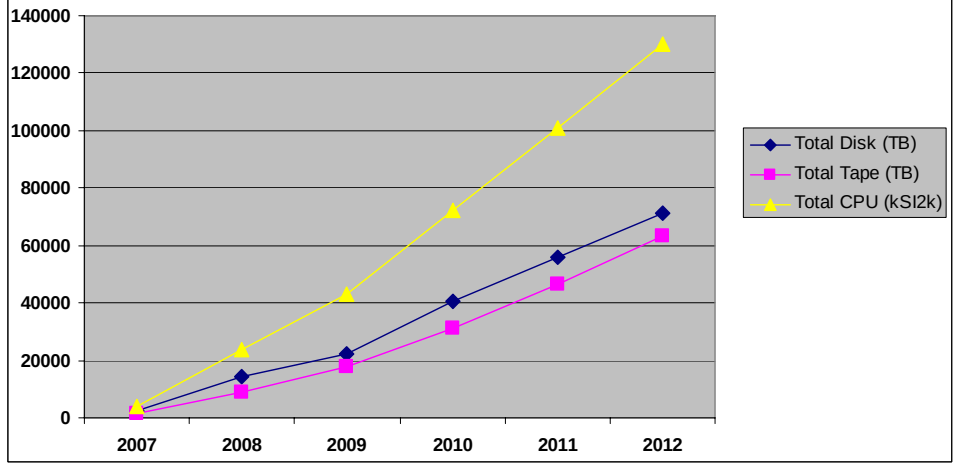
CPU planned Resources

- In the next slides the planned resources for ATLAS are shown
  - As in the Computing TDR
- Assuming nearly saturation one can easily extract job numbers etc. in the different Tiers…

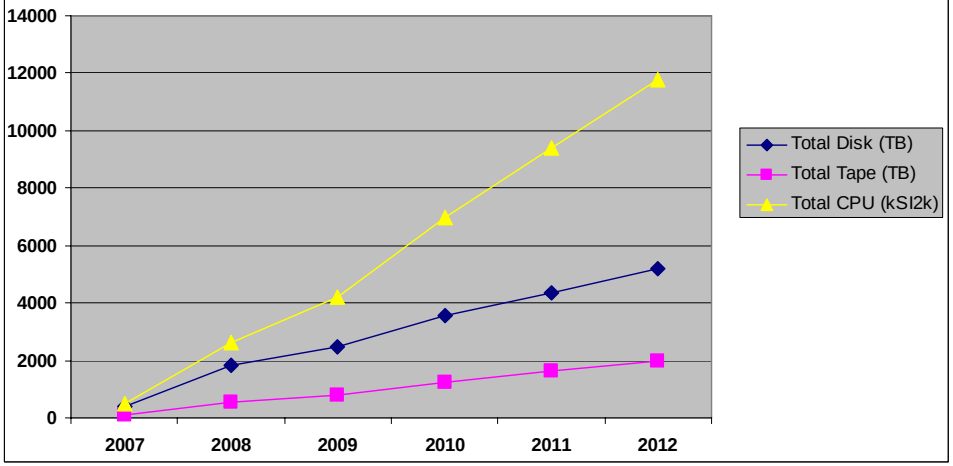
Tier-0



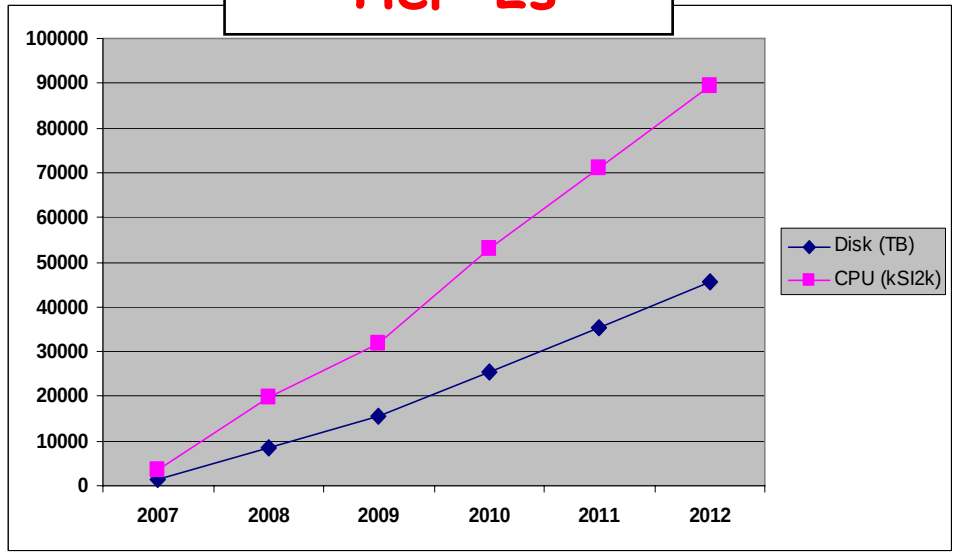
Tier-1s



CERN Analysis Facility



Tier-2s



Total ATLAS Requirements in for 2008

Table 7-2 The projected total resources required at the start of 2008 for the case when 20% of the data rate is fully simulated.

	CPU (MSI2k)	Tape (PB)	Disk (PB)
Tier-0	4.1	5.7	0.39
CERN AF	2.7	0.5	1.9
Sum of Tier-1s	24.0	9.0	14.4
Sum of Tier-2s	19.9	0.0	8.7
Total	50.6	16.9	25.4

Grid usage requirements

- Most members of the Collaboration have not been confronted yet with the current Grid middleware
  - They expect a simple extension of the common batch systems (such as LSF @ CERN)
    - User disk space
    - Project (group) space
    - Fair share job submission
- VOMS is the tool we have for making a step forward wrt the "free for all" current situation
  - But the consistent implementation of all client tools is needed NOW
  - As we need to experiment for being able to setup in time an efficient system
- Next slides taken (with some mod.s) from ATLAS presentation at the March GDB

ATLAS Grid Requirements

- ATLAS is a very large VO (the largest?) and consists of several "activity groups" that will compete for computing resources
- Assume we have defined VOMS groups and roles and registered all ATLAS VO members accordingly
- Naively we would like to use this information for:
  - Monitoring & accounting
  - Assigning job priorities (and fair share)
  - Allocating disk storage space
  - Selecting the machines that have enough memory (Reconstruction needs 2 GB of memory…). Other attributes may be useful for selection…
- We would also expect to be able to specify requirements and use the information at user, group and VO level
- We have therefore to be able to assign resources to activity groups and get accurate monitoring and accounting reports

3 Dimensions

- Roles:
  - Grid software administrators (who install software and manage the resources)
  - Production managers for official productions
  - Normal users
- Groups:
  - Physics groups
  - Combined performance groups
  - Detectors & trigger
  - Computing & central productions
- Funding:
  - Countries and funding agencies

Group list

- phys-beauty phys-top phys-sm
- phys-higgs phys-susy phys-exotics
- phys-hi phys-gener phys-lumin
- perf-egamma perf-jets perf-flavtag
- perf-muons perf-tau trig-pesa
- det-indet det-larg det-tile
- det-muon soft-test soft-valid
- soft-prod soft-admin gen-user

- It is foreseen that initially only group production managers would belong to most of those groups
  - All Collaboration members would be, at least initially, in "gen-user"
  - Software installers would be in soft-admin
- The matrix would therefore be diagonal
  - Only ~25 group/role combinations would be populated

Job Priorities (and fair share)

- Once groups and roles are set up, we have to use this information
- Relative priorities are easy to enforce if all jobs go through the same queue (or database)
- In case of a distributed submission system, it is up to the resource providers to:
  - agree the policies of each site with ATLAS
  - publish and enforce the agreed policies
- The jobs submission systems must take these policies into account to distribute jobs correctly
  - the priority of each job is different on each site
- Developments are in progress in both OSG and EGEE in this direction
  - But we do not have any complete solution to this problem yet

Storage allocation

- The bulk of ATLAS disk storage will be used by central productions and organized group activities
- Disk storage has to be managed according to VOMS groups on all SE's available to ATLAS
- In addition, individual users will have data they want to share with their colleagues on the Grid
  - Similarly to the way (ATLAS) people use public directories and project space on lxplus at CERN
- Therefore, again, we need resource allocation and accounting at user, group and VO level

Caveat

- The scenario I described is only partially tested yet
  - We still have to perform our Computing System Commissioning runs
- As far as VO Grid requirements we plan to "start simple" in SC4 and see then how to escalate to more complexity
- Next slides elaborate a bit more on these points

Computing System Commissioning Goals

- We have defined the high-level goals of the Computing System Commissioning operation during 2006
  - Formerly called "DC3"
  - More a running-in of continuous operation than a stand-alone challenge
- Main aim of Computing System Commissioning will be to test the software and computing infrastructure that we will need at the beginning of 2007:
  - Calibration and alignment procedures and conditions DB
  - Full trigger chain
  - Tier-0 reconstruction and data distribution
  - Distributed access to the data for analysis
- At the end we will have a working and operational system, ready to take data with cosmic rays at increasing rates

Start simple with VO requirements

- Work is going on in the "Job priority WG"
- We need something simple to start with in SC4
- The minimal requirements to satisfy are
  - Job characterized by quartet
    - DN, GROUP, ROLE, QUEUE
  - Production has a guaranteed share
  - Short jobs do not have to wait too long
  - Limits within a local fair share must apply
- At each site 4 "queues" (or "attributes") may be enough and they can be then mapped to our groups and roles
  - Coordinated deployment in all the sites absolutely needed
- However ATLAS estimates that fine grained CPU accounting is needed from start

Conclusions

- ATLAS plans for a massive amount of CPU usage
  - In Tier0, 1, 2
  - Mostly in batch mode and via GRID
- We definitely want to start everywhere with simple solutions and increase the complexity step by step
- Still we expect we will need "soon" tools that allow the ATLAS user to use the GRID in a way somewhat comparable to a local batch system
  - Or at least not unbearably worse than that….
  - Relevant deployment and support work required also from the sites