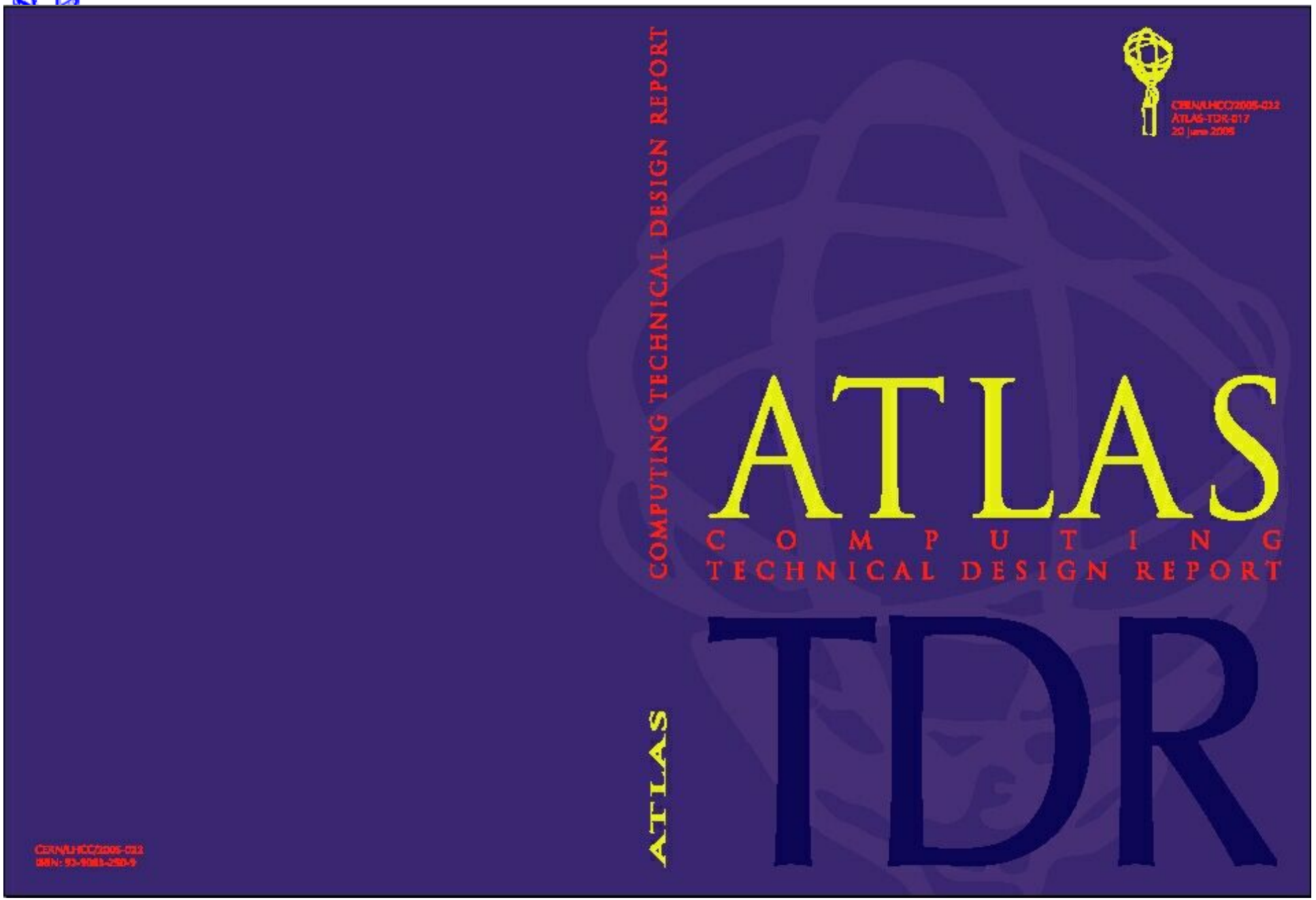
ATLAS Computing Technical Design Report

Dario Barberis

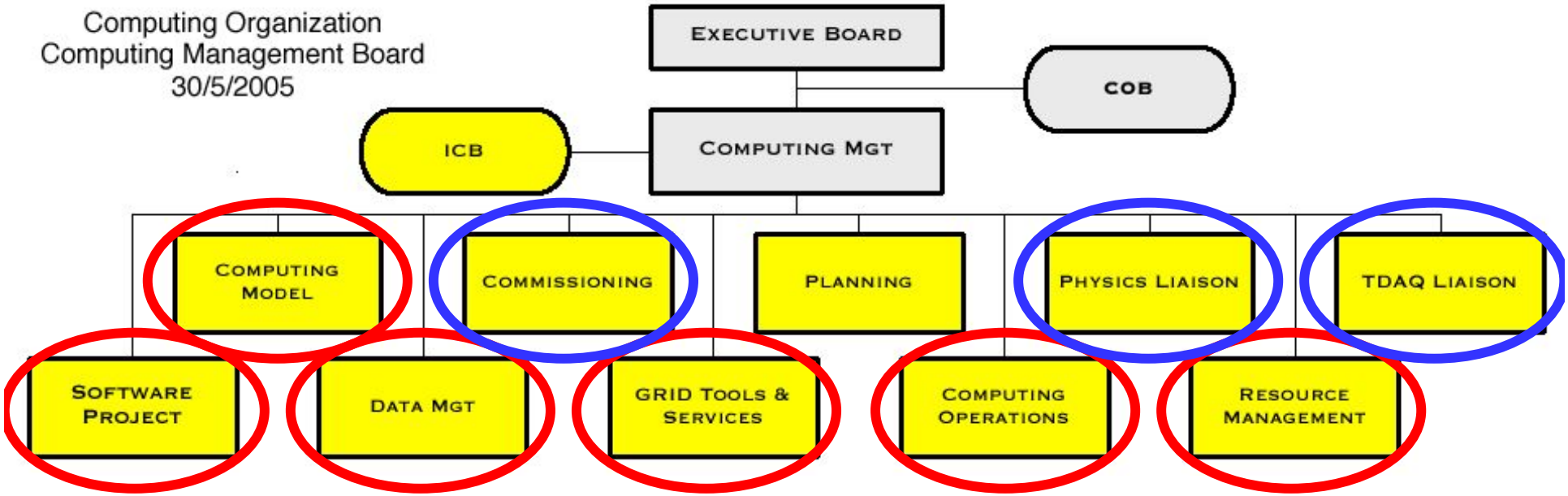
(CERN & Genoa University)

on behalf of the ATLAS Collaboration


COMPUTING TECHNICAL DESIGN REPORT
ATLAS
ATLAS
COMPUTING
TECHNICAL DESIGN REPORT
TDR

Computing TDR structure

The TDR describes the whole Software & Computing Project as defined within the ATLAS organization:



Major activity areas within the S&C Project

Liaisons to other ATLAS projects

Computing TDR structure

- The chapter structure follows closely the organization of the Software & Computing Project:
  - Chapter 1: Introduction
  - Chapter 2: Computing Model
  - Chapter 3: Offline Software
  - Chapter 4: Databases and Data Management
  - Chapter 5: Grid Tools and Services
  - Chapter 6: Computing Operations
  - Chapter 7: Resource Requirements
  - Chapter 8: Project Organization and Planning
  - Appendix: Glossary

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

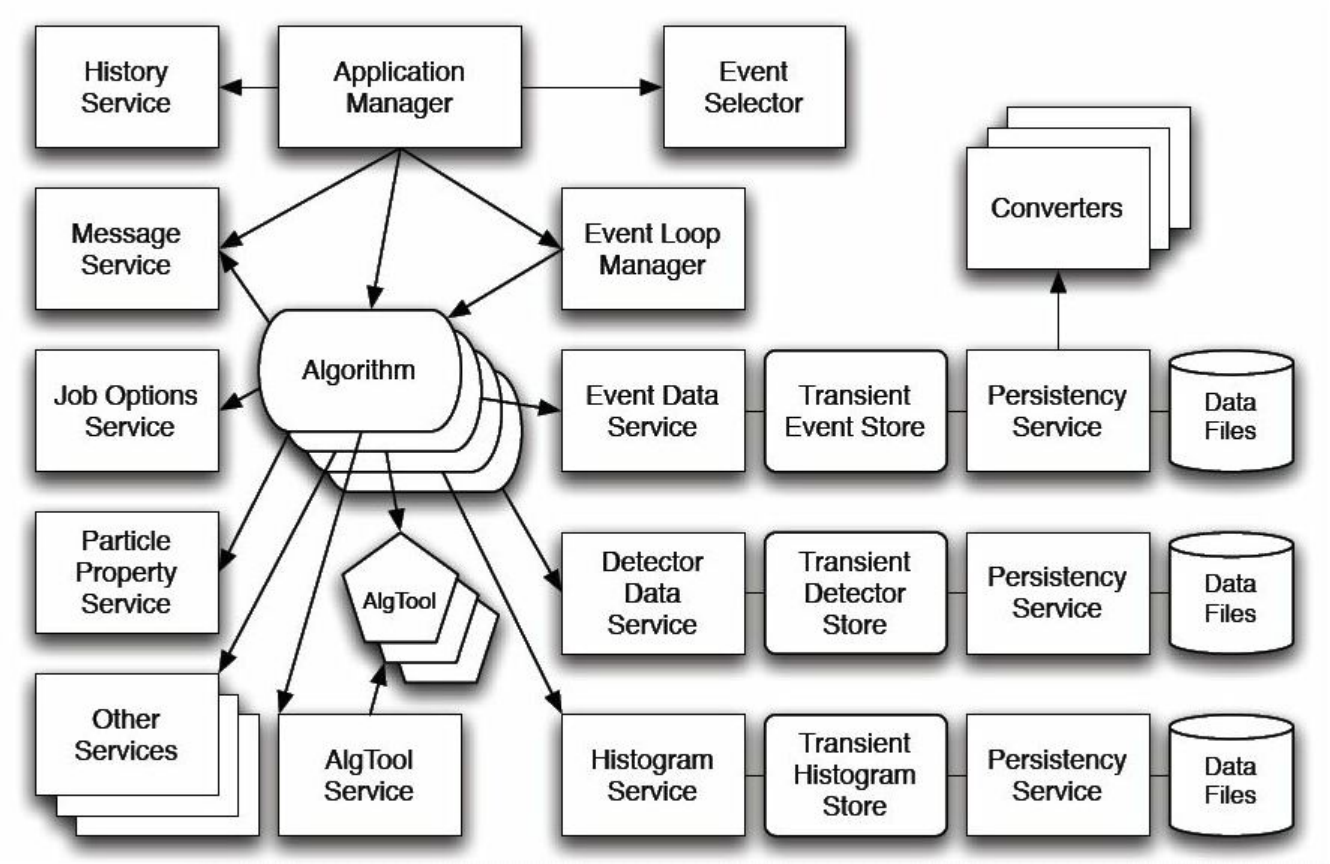
- Tier-0:
  - Copy RAW data to Castor tape for archival
  - Copy RAW data to Tier-1s for storage and reprocessing
  - Run first-pass calibration/alignment (within 24 hrs)
  - Run first-pass reconstruction (within 48 hrs)
  - Distribute reconstruction output (ESDs, AODs & TAGS) to Tier-1s
- Tier-1s:
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

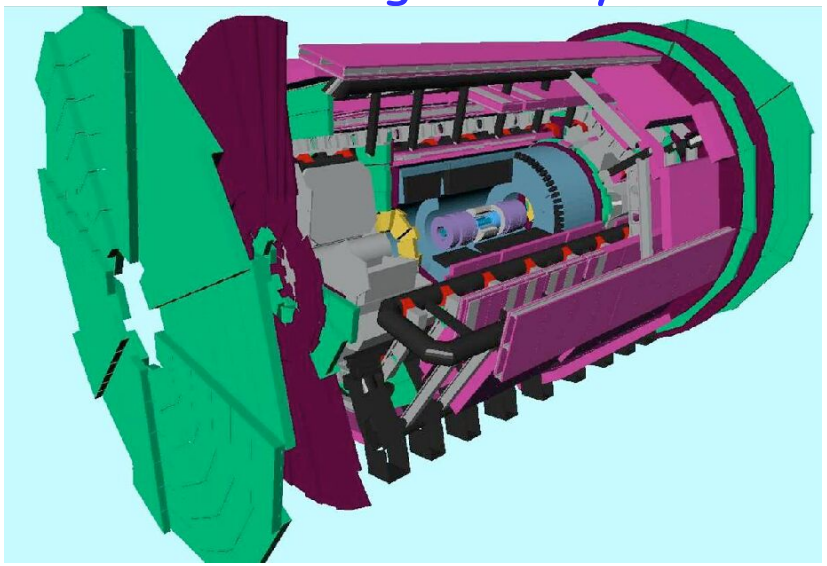
Offline Software: Architecture

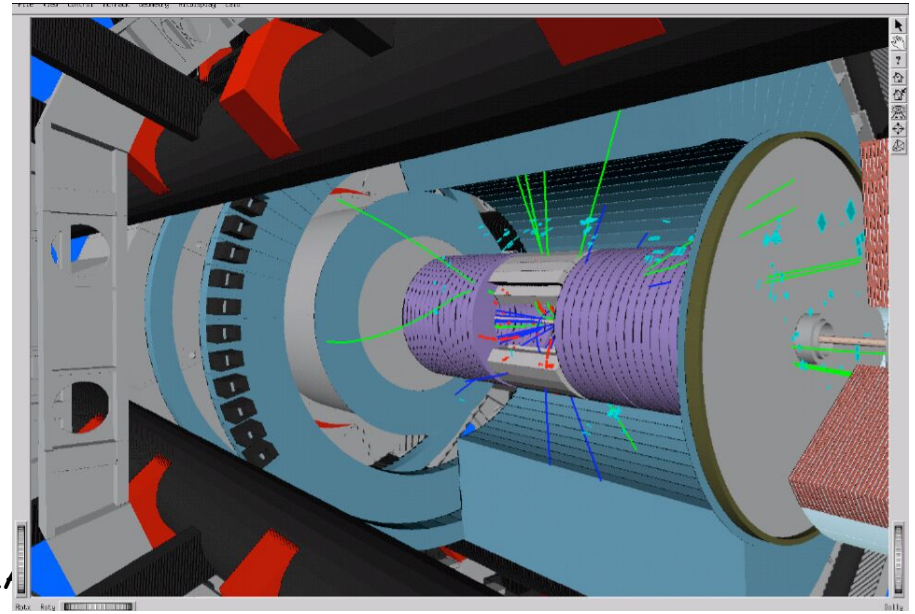
- The architecture of the Athena framework is based on Gaudi:
  - Separation of data from algorithms
  - Separation of transient (in-memory) from persistent (in-file) data
  - Extensive use of abstract interfaces to decouple the various components



Offline Software: Geometry

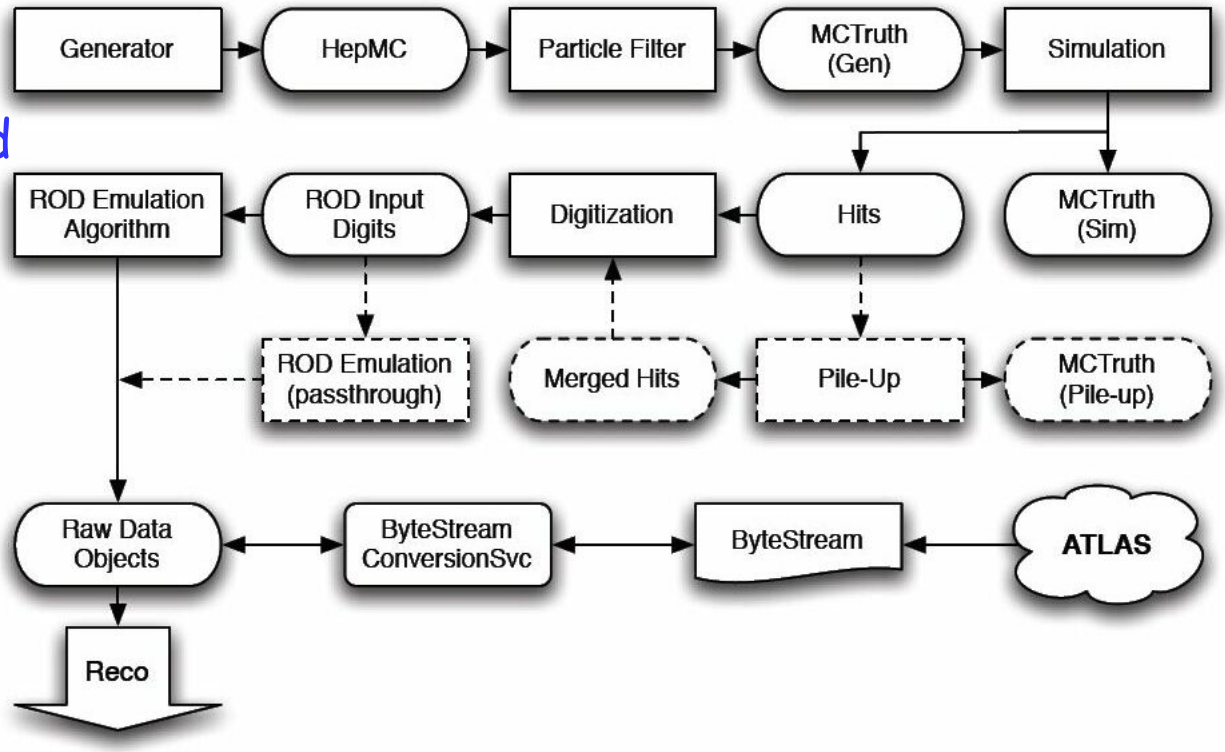
- The GeoModel detector description system provides us with an application-independent way to describe the geometry
- In this way Simulation, Reconstruction, Event Display etc. use by definition the same geometry
- Geometry data are stored in a database with a Hierarchical Versioning System
- Alignment corrections are applied with reference to a given baseline geometry





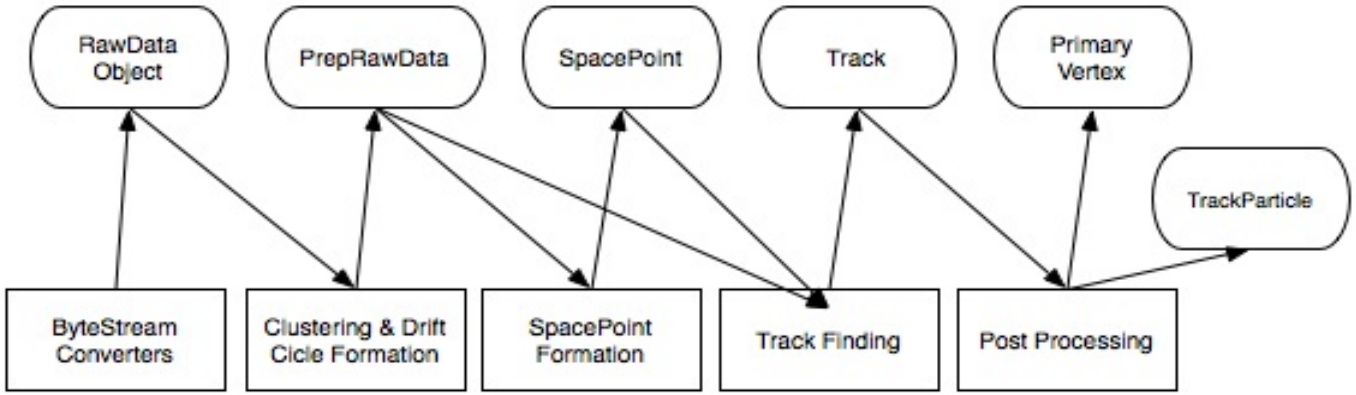
Offline Software: Simulation

- Event generator framework interfaces multiple packages
  - including the Genser distribution provided by LCG-AA
- Simulation with Geant4 since early 2004
  - automatic geometry build from GeoModel
  - >25M events fully simulated up to now since mid-2004
    - only a handful of crashes!
- Digitization tested and tuned with Test Beam
- Fast simulation also used for preliminary large-statistics (physics) background level studies

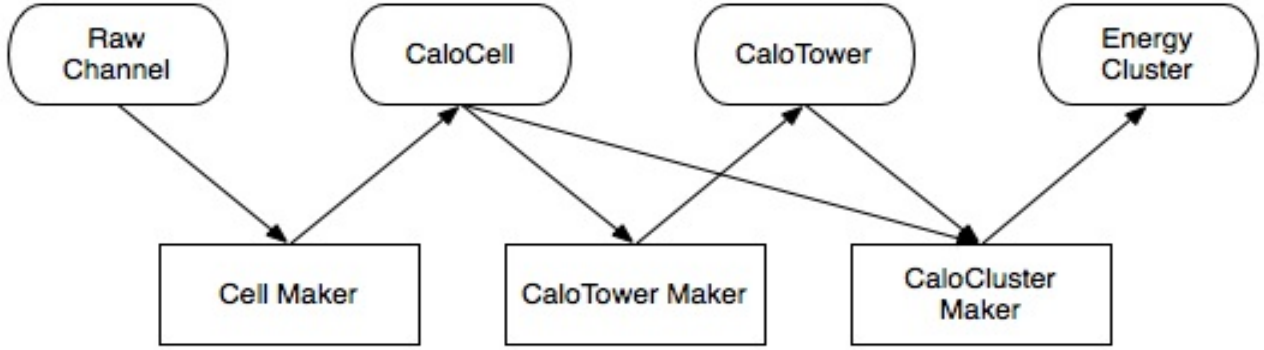


Offline Software: Reconstruction

- Separation of data and algorithms:
  - Tracking code:



  - Calorimetry code:



- Resource needs (memory and CPU) currently larger than target values
  - Optimization and performance, rather than functionality, will be the focus of developments until detector turn-on

Offline Software: Physics Analysis Tools

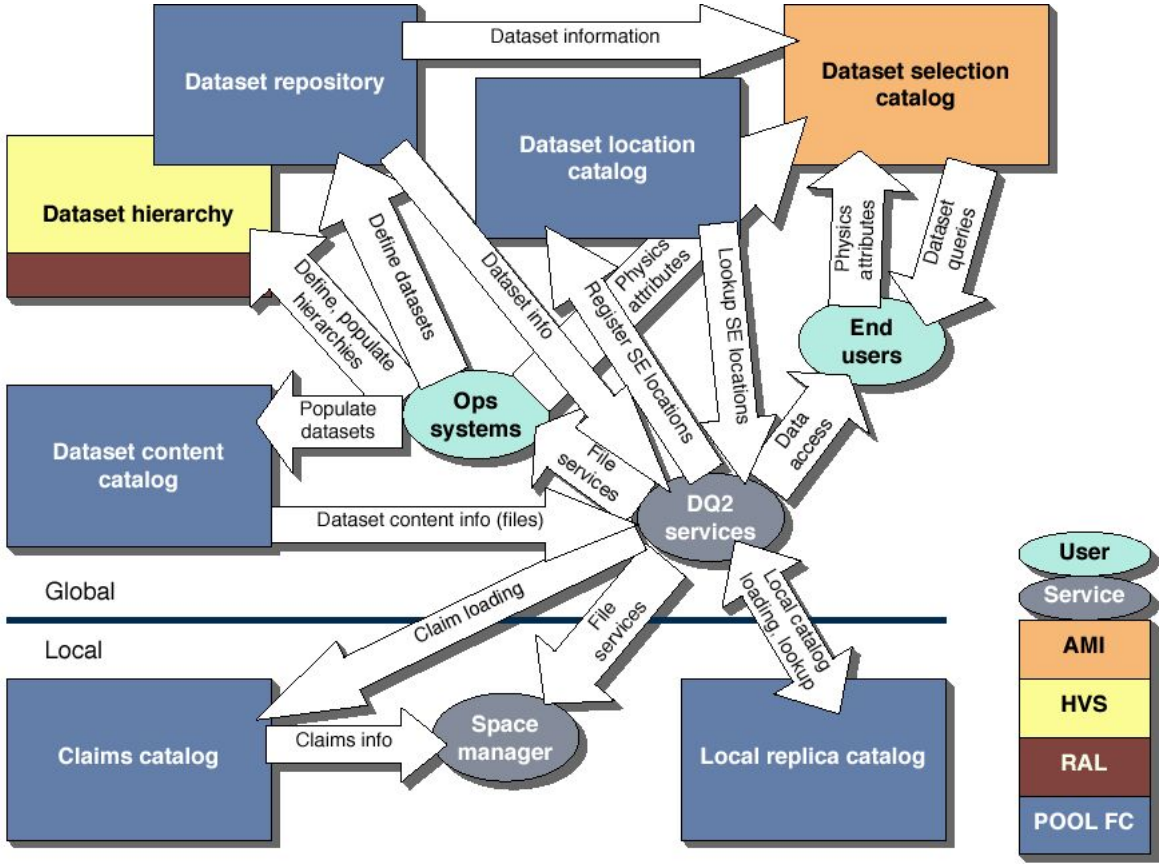
- The Physics Analysis Tools group develops common utilities for analysis based on the Athena framework
  - classes for selections, sorting, combinations etc. of data objects
  - constituent navigation (e.g. jets to clusters) and back navigation (e.g. AOD to ESD)
  - UserAnalysis package in Athena
  - interactive analysis in Athena
  - analysis in Python
  - interfaces to event displays
  - testing the concept of "Event View": a coherent list of physics objects that are mutually exclusive
    - any object appears only once in the list of reconstructed objects available for analysis

Databases and Data Management

- The DB/DM project takes care of all types of ATLAS data
- Event data:
  - file organization (using LCG POOL/ROOT data storage)
  - cataloguing
  - book-keeping
  - data access, converters, schema evolution
- Non-event data:
  - Technical Coordination, Production, Installation, Survey DBs
  - Online conditions and configuration data
  - Geometry and Calibration/alignment DB (using LCG COOL DB)
  - DB access methods, data distribution

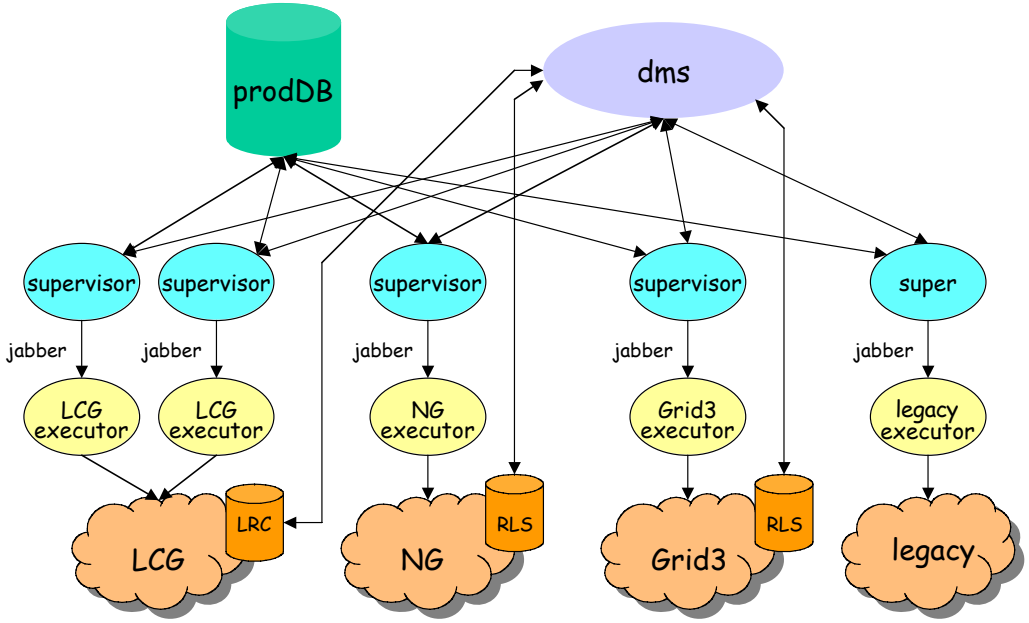
Distributed Data Management

- Accessing distributed data on the Grid is not a simple task
- Several DBs are needed centrally to hold dataset information
- "Local" catalogues hold information on local data storage
- The new DDM system (right) is under test this summer
- It will be used for all ATLAS data from October on (LCG Service Challenge 3)



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 in preparation
  - keeping the same architecture:
    - ProdDB
    - Common supervisor
    - One executor per Grid
    - Interface to DDM



Distributed Analysis System

- Several groups started independently in 2002-2003 to work on prototypes of distributed analysis systems
- LCG RTAG 11 did not produce in 2003 a common analysis system project as hoped. ATLAS therefore planned to combine the strengths of various existing prototypes:
  - GANGA to provide a user interface to the Grid for Gaudi/Athena jobs
  - DIAL to provide fast, quasi-interactive, access to large local clusters
  - The ATLAS Production System to interface to the 3 Grid flavours
- The combined system did not provide the users with the expected functionality for the Rome Physics Workshop (June '05)
- We are currently reviewing this activity in order to define a baseline for the development of a performing Distributed Analysis System
- In the meantime other projects appeared on the market:
  - DIANE to submit interactive jobs to a local computing cluster
  - LJSF to submit batch jobs to the LCG/EGEE Grid (using part of ProdSys)
- All this has to work together with the DDM system described earlier
- If we decide a baseline "now", we can have a testable system by this autumn

Computing Operations

- The Computing Operations organization has to provide for:
  - a) CERN Tier-0 operations
    - from output of EF to data distribution to Tier-1's, including calibration/alignment and reconstruction procedures
  - b) World-wide operations:
    - simulation job distribution
    - re-processing of real and simulated data at Tier-1's
    - data distribution and placement
  - c) Software distribution and installation
  - d) Site and software installation validation and monitoring
  - e) Coordination of Service Challenges in 2005-2006
  - f) User Support
- ... along the guidelines of the Computing Model
- Some of the needed components already exist
  - and have been tested during Data Challenges

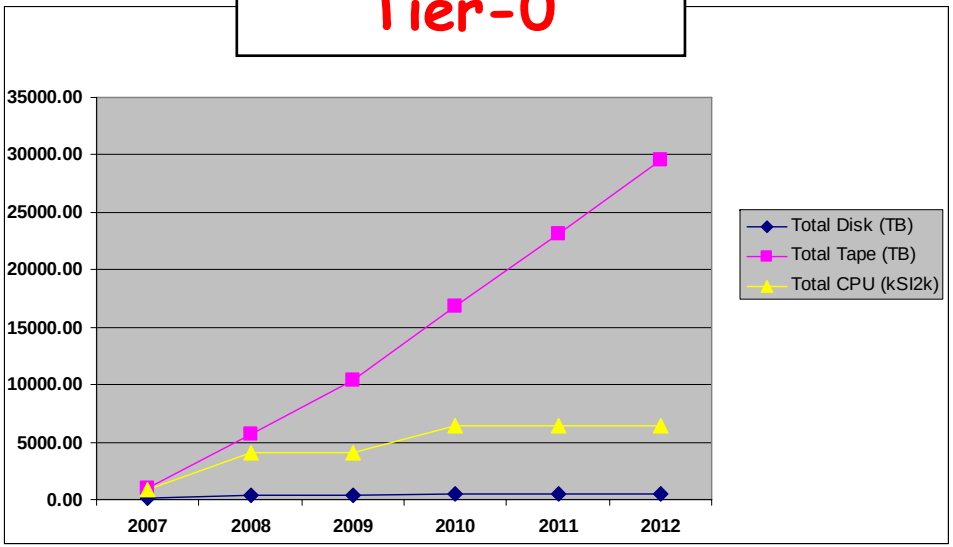
ATLAS Virtual Organization

- Right now the Grid is "free for all"
  - no CPU or storage accounting
  - no priorities
  - no storage space reservation
- Already we have seen competition for resources between "official" Rome productions and "unofficial", but organised, productions
  - B-physics, flavour tagging...
- The latest release of the VOMS (Virtual Organisation Management Service) middleware package allows the definition of user groups and roles within the ATLAS Virtual Organisation
  - and is used by all 3 Grid flavours!
- 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

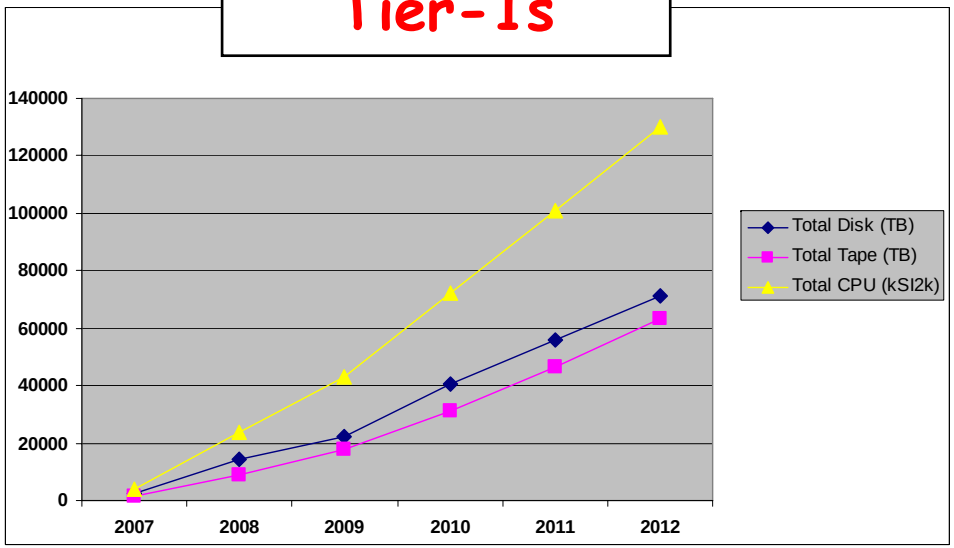
Computing Model and Resources

- The Computing Model in the TDR is basically the same as in the Computing Model document (Dec. 2004) submitted for the LHCC review in January 2005
- Main differences:
  - Inclusion of Heavy Ion runs from 2008 onwards
    - add ~10% to data storage needs
    - limit the increase in CPU needed at Tier-1s to <10% by a different model for data reprocessing
  - Decrease of the nominal run time in 2007 from 100 to 50 days
    - after agreement with LHCC referees (McBride/Forti)
- Immediate consequences of these changes on the resource needs are a slower ramp-up in 2007-2008, followed by a steeper increase as soon as Heavy Ions start (order 10% for storage everywhere and for CPU at Tier-1s)

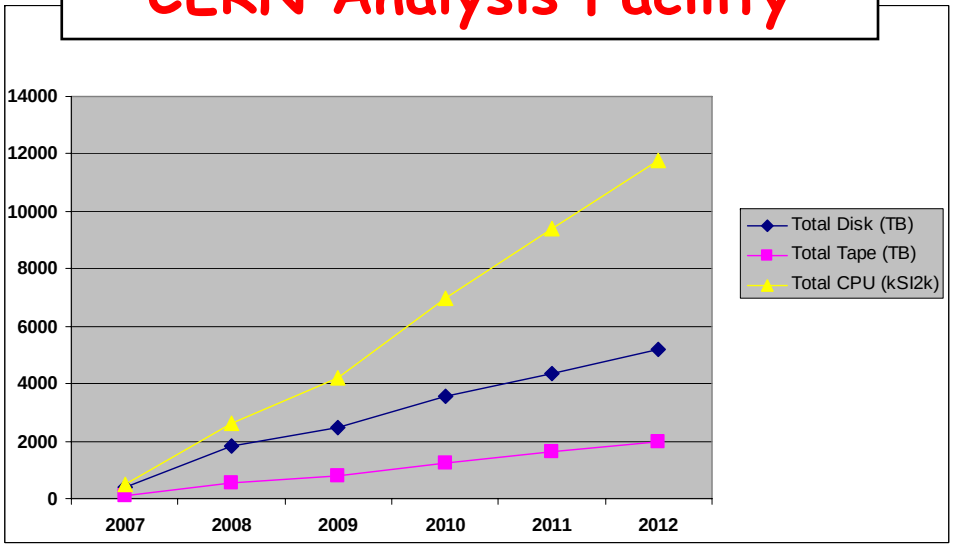
Tier-0



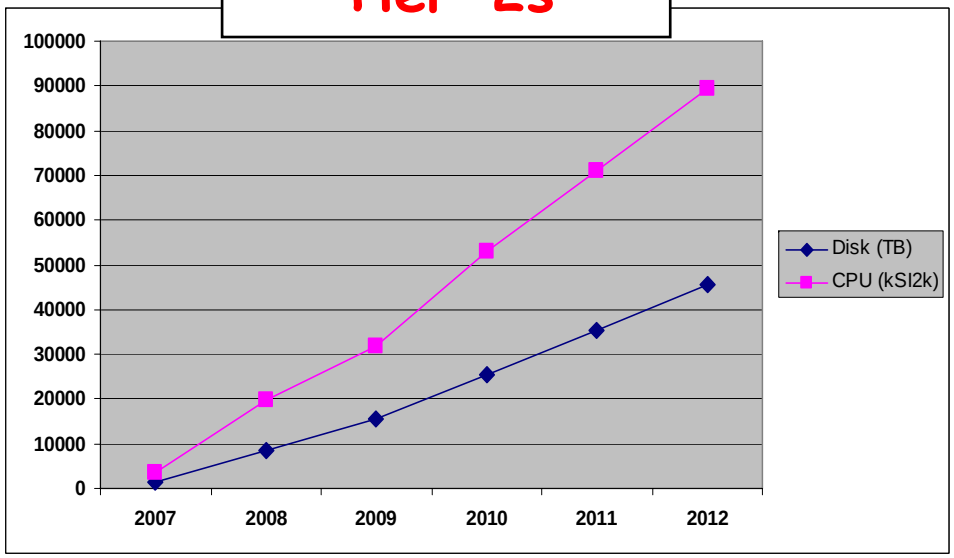
Tier-1s



CERN Analysis Facility



Tier-2s

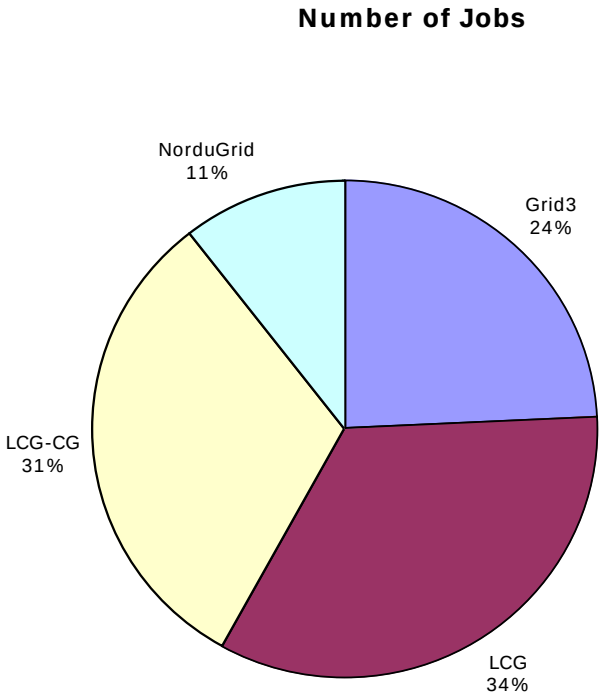


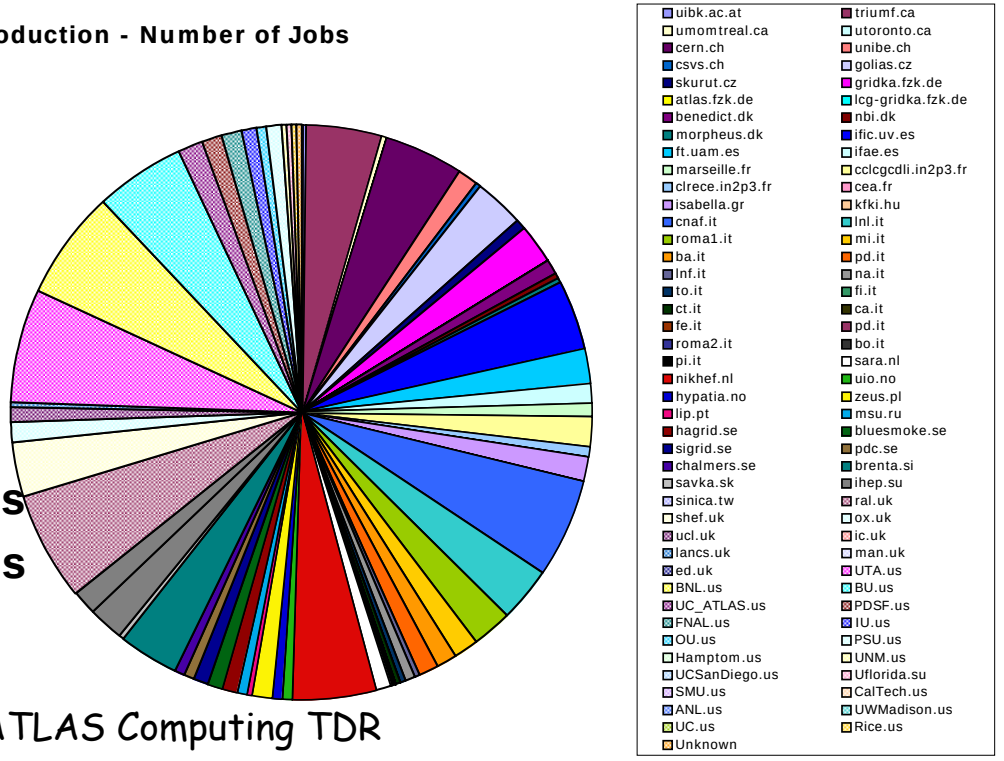
Massive productions on 3 Grids (1)

- July-September 2004: DC2 Geant4 simulation (long jobs)
  - 40% on LCG/EGEE Grid, 30% on Grid3 and 30% on NorduGrid
- October-December 2004: DC2 digitization and reconstruction (short jobs)
- February-May 2005: Rome production (mix of jobs as digitization and reconstruction was started as soon as samples had been simulated)
  - 65% on LCG/EGEE Grid, 24% on Grid3, 11% on NorduGrid

Massive productions on 3 Grids (2)

- 73 data sets containing 6.1M events simulated and reconstructed (without pile-up)
- Total simulated data: 8.5M events
- Pile-up done later (for 1.3M events done up to last week)





573315 jobs
22 countries
84 sites

Dario Barberis: ATLAS Computing TDR

Computing System Commissioning Goals

- We have recently discussed and defined the high-level goals of the Computing System Commissioning operation during the first half of 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 (summer 2006) we will have a working and operational system, ready to take data with cosmic rays at increasing rates

Computing System Commissioning Tests

- Sub-system tests with well-defined goals, preconditions, clients and quantifiable acceptance tests
  - Full Software Chain
  - Tier-0 Scaling
  - Calibration & Alignment
  - Trigger Chain & Monitoring
  - Distributed Data Management
  - Distributed Production (Simulation & Re-processing)
  - Physics Analysis
  - Integrated TDAQ/Offline (complete chain)
- Each sub-system is decomposed into components
  - E.g. Generators, Reconstruction (ESD creation)
- Goal is to minimize coupling between sub-systems and components and to perform focused and quantifiable tests
- Detailed planning being discussed now

Project Milestones

- July 2005: test the TAG and event collection infrastructure for data access for analysis (important for the Computing Model and the Event Data Model).
- July 2005: decision on the future baseline for the Distributed Analysis System.
- September 2005: new version of the Production System fully tested and deployed.
- September 2005: delivery of software and infrastructure for ATLAS Commissioning and for Computing System Commissioning; start of integration testing.
- October 2005: review of the implementation of the Event Data Model for reconstruction.
- October 2005: Distributed Data Management system ready for operation with LCG Service Challenge 3.
- November 2005: start of simulation production for Computing System Commissioning.
- January 2006: production release for Computing System Commissioning and early Cosmic Ray studies; completion of the implementation of the Event Data Model for reconstruction.
- February 2006: start of DC3 (Computing System Commissioning).
- April 2006: integration of ATLAS components with LCG Service Challenge 4.
- July 2006: production release for the main Cosmic Ray runs (starting in Autumn 2006).
- December 2006: production release for early real proton data.

Conclusions

- Data-taking for ATLAS is going to start "soon"
  - as a matter of fact, cosmic ray runs are starting now already in the pit
- The Software & Computing Project is going to provide a working system and infrastructure for the needs of first ATLAS data
  - indeed, already now for detector commissioning and cosmic ray runs
- Offline software and databases are already on a good level of functionality
  - work will concentrate, from now on, on usability, optimization and performance
- Distributed systems still need a lot of testing and running in
  - true for Distributed Productions, Distributed Data Management and Distributed Analysis
- The Computing System Commissioning activity in the first half of 2006 will test an increasingly complex and functional system and provide us in summer 2006 with an operational infrastructure