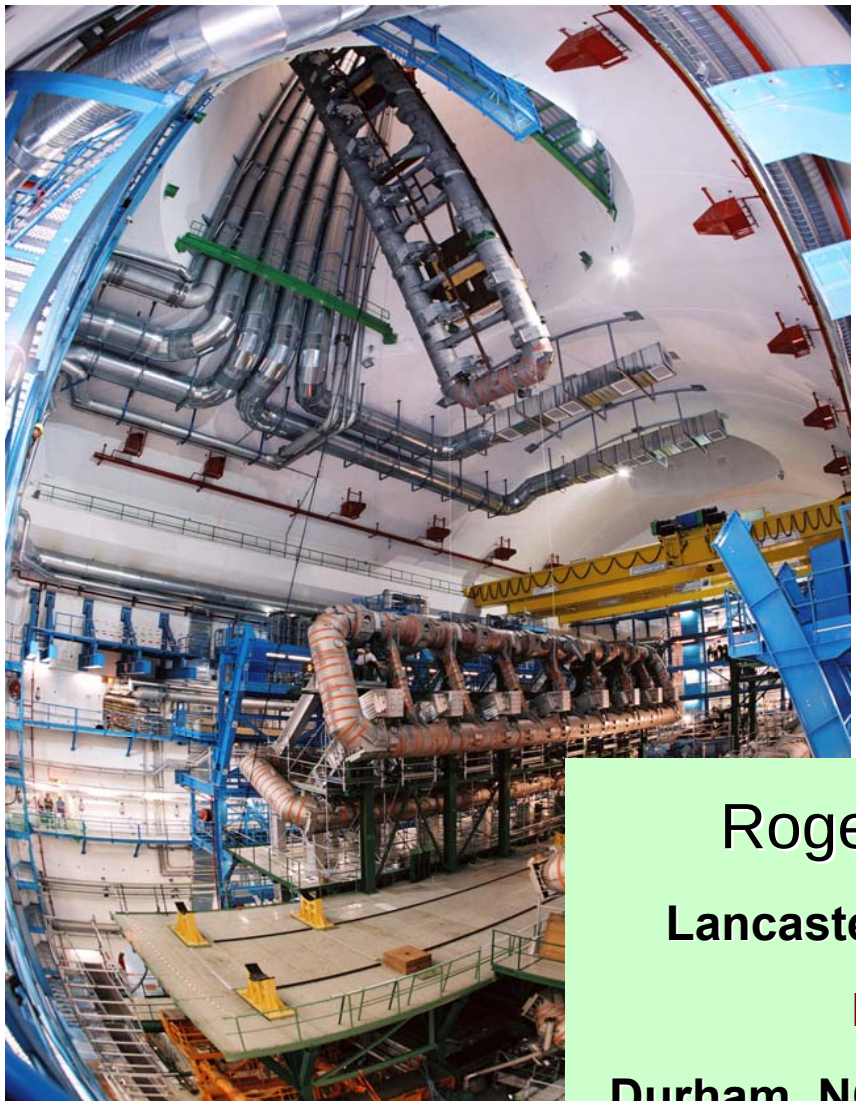
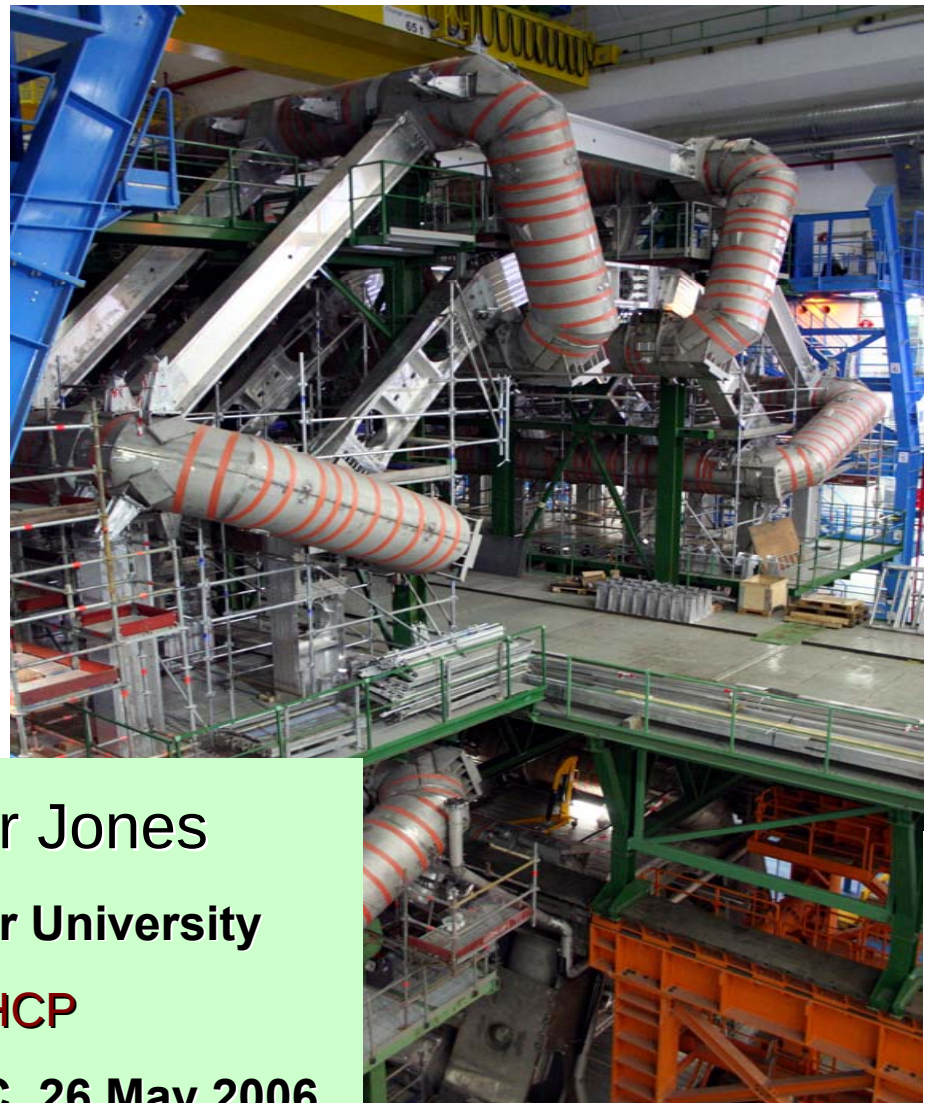
The ATLAS Computing & Analysis Model





Roger Jones

Lancaster University

HCP

Durham, NC, 26 May 2006

Overview

- Brief summary ATLAS Facilities and their roles
- Analysis modes and operations
- Data selection
- Distributed Analysis Tools
- Analysis Tools

Computing Resources

- **Computing Model fairly well evolved, documented in C-TDR**
  - **Externally reviewed**
  - **http://doc.cern.ch//archive/electronic/cern/preprints/lhcc/public/lhcc-2005-022.pdf**
- **There are (and will remain for some time) many unknowns**
  - **Calibration and alignment strategy is still evolving**
  - **Physics data access patterns MAY be exercised from June**
    - **Unlikely to know the real patterns until 2007/2008!**
  - **Still uncertainties on the event sizes , reconstruction time**
- **Lesson from the previous round of experiments at CERN (LEP, 1989-2000)**
  - **Reviews in 1988 underestimated the computing requirements by an order of magnitude!**

ATLAS Facilities

- **Event Filter Farm at CERN**
  - Located near the Experiment, assembles data into a stream to the Tier 0 Center
- **Tier 0 Center at CERN**
  - Raw data ➔ Mass storage at CERN and to Tier 1 centers
  - Swift production of Event Summary Data (ESD) and Analysis Object Data (AOD)
  - Ship ESD, AOD to Tier 1 centers ➔ Mass storage at CERN
- **Tier 1 Centers distributed worldwide (10 centers)**
  - Re-reconstruction of raw data, producing new ESD, AOD
  - Scheduled, group access to full ESD and AOD
- **Tier 2 Centers distributed worldwide (approximately 30 centers)**
  - Monte Carlo Simulation, producing ESD, AOD, ESD, AOD ➔ Tier 1 centers
  - On demand user physics analysis of shared datasets
- **CERN Analysis Facility**
  - Analysis
  - Heightened access to ESD and RAW/calibration data on demand
- **Tier 3 Centers distributed worldwide**
  - Physics analysis

Processing

- **Tier-0:**
  - Prompt first pass processing on express/calibration & physics streams
  - 24-48 hours later, process full physics data streams with reasonable calibrations
    - →Implies large data movement from T0 →T1s
- **Tier-1:**
  - Reprocess 1-2 months after arrival with better calibrations
  - Reprocess all resident RAW at year end with improved calibration and software
    - →Implies large data movement from T1↔T1 and T1 → T2

Inputs to the ATLAS Computing Model (1)

2.2.1 LHC Operational Parameters and Trigger Rates

Table 2-1 The assumed LHC operational parameters and ATLAS trigger rates

E	14 TeV (two 7 TeV proton beams)
L	$0.5*10^{33}$ $cm^{-2}s^{-1}$ in 2007
	$2*10^{33}$ $cm^{-2}s^{-1}$ in 2008 and 2009
	10^{34} $cm^{-2}s^{-1}$ (design luminosity) from 2010 onwards
σ	100 mb = 10^{-25} cm^2
Collision rate	$L \cdot \sigma = 10^9$ Hz p-p collisions at design luminosity
Trigger rate	200 Hz independent of the luminosity

Inputs to the ATLAS Computing Model (2)

Table 2-2 The assumed event data sizes for various formats, the corresponding processing times and related operational parameters.

Item	Unit	Value
Raw Data Size	MB	1.6
ESD Size	MB	0.5
AOD Size	kB	100
TAG Size	kB	1
Simulated Data Size	MB	2.0
Simulated ESD Size	MB	0.5
Time for Reconstruction (1 ev)	kSI2k-sec	15
Time for Simulation (1 ev)	kSI2k-sec	100
Time for Analysis (1 ev)	kSI2k-sec	0.5
Event rate after EF	Hz	200
Operation time	seconds/day	50000
Operation time	days/year	200
Operation time (2007)	days/year	50
Event statistics	events/day	10^7
Event statistics (from 2008 onwards)	events/year	$2 \cdot 10^9$

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

Data Flow

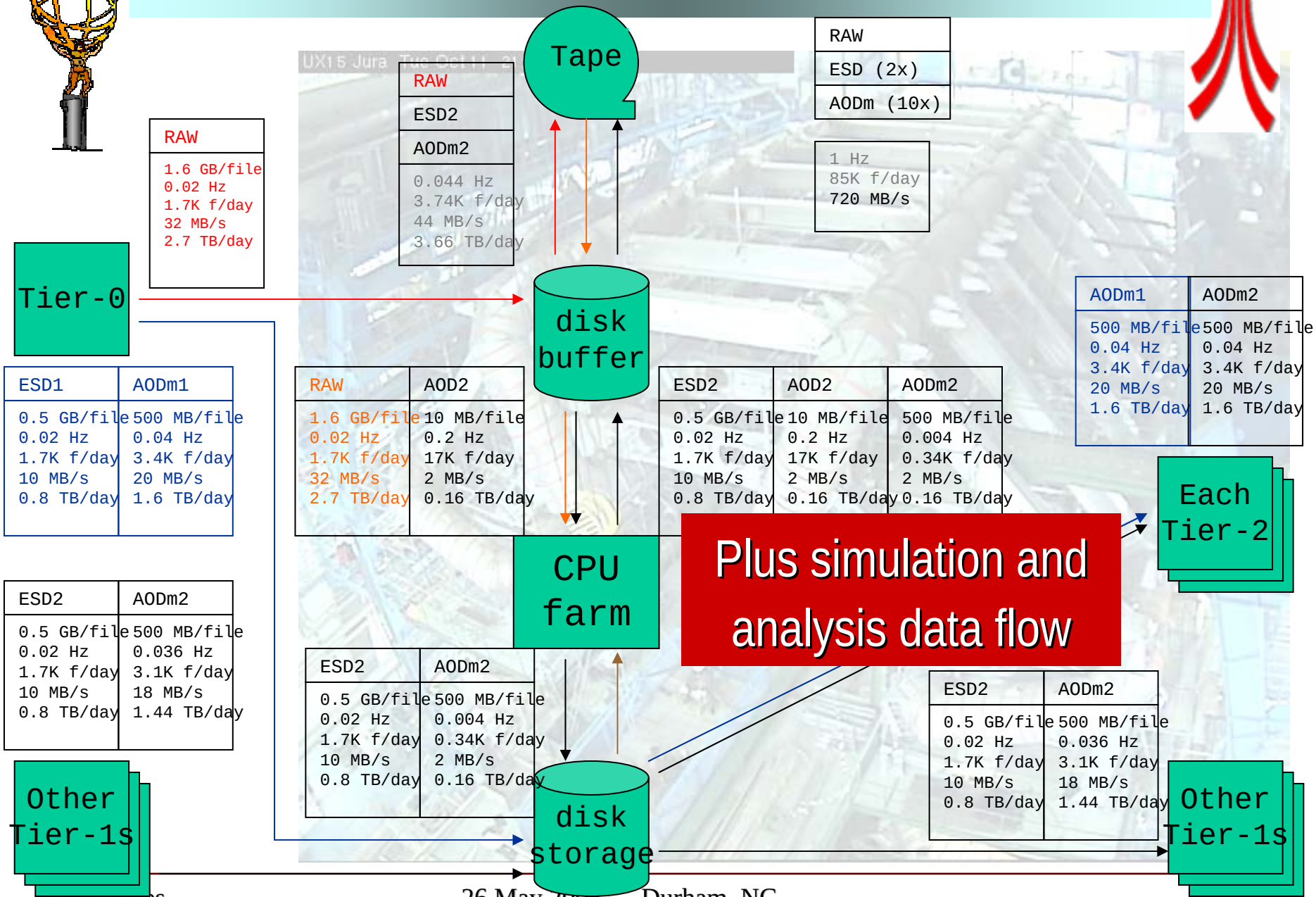
- EF farm ➔ T0
  - 320 MB/s continuous
- T0 Raw data ➔ Mass Storage at CERN
- T0 Raw data ➔ Tier 1 centers
- T0 ESD, AOD, TAG ➔ Tier 1 centers
  - 2 copies of ESD distributed worldwide

Tier 0 view

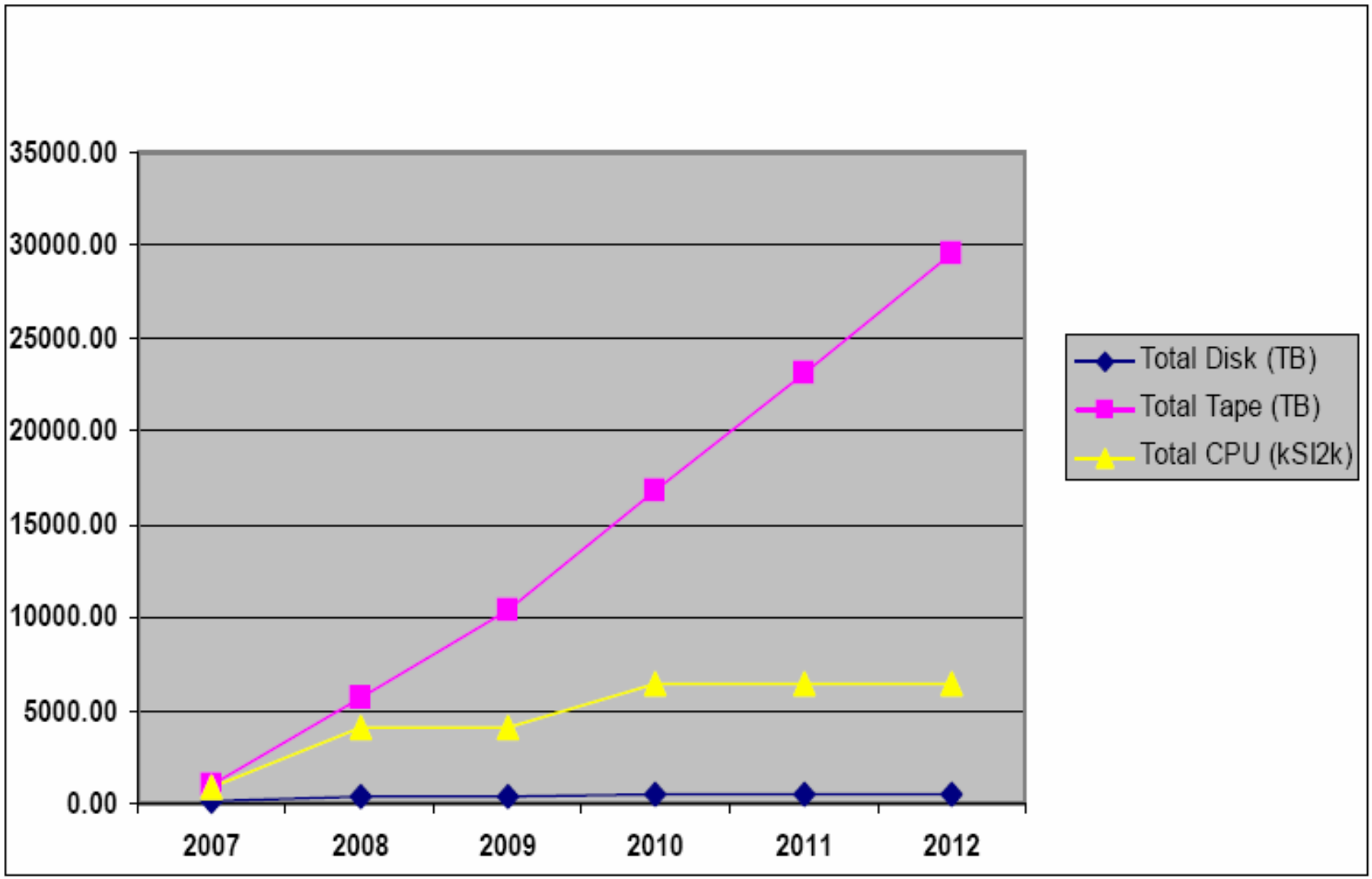
- T1 ➔ T2
  - Some RAW/ESD, All AOD, All TAG
  - Some group derived datasets
- T2 ➔ T1
  - Simulated RAW, ESD, AOD, TAG
- T0 ➔ T2 Calibration processing?

Tier 2 view

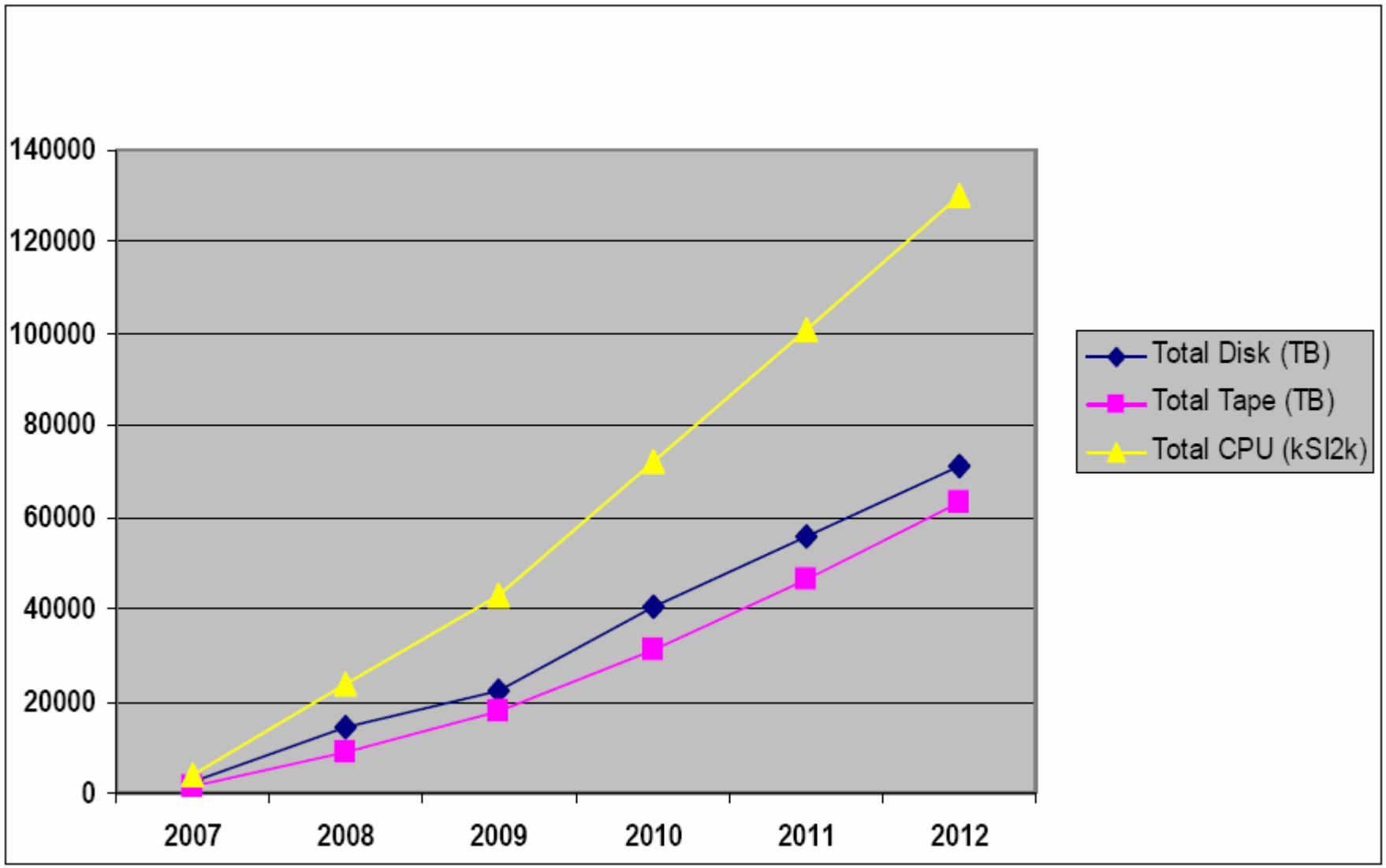
ATLAS partial &"average" T1 Data Flow (2008)



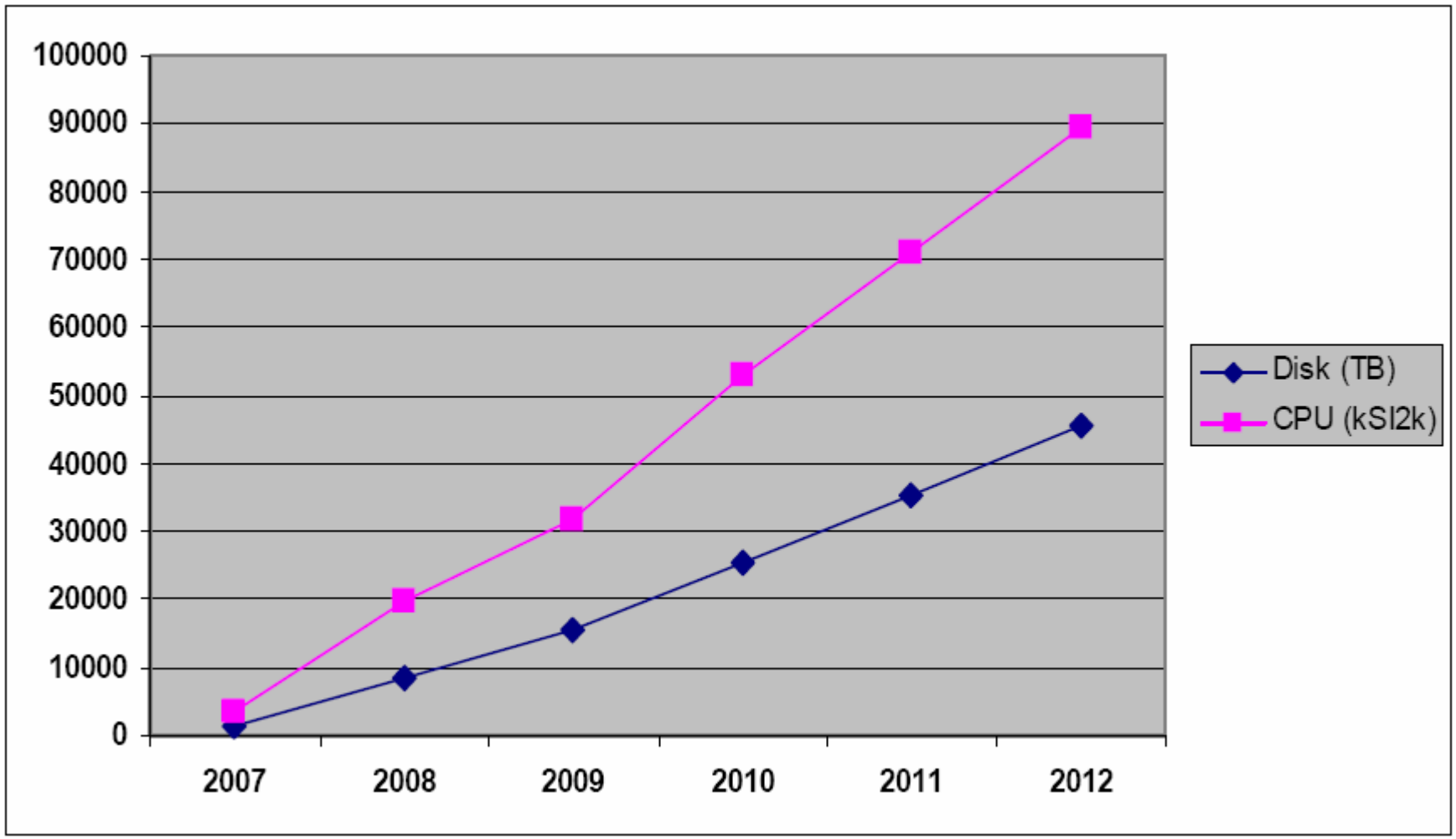
ATLAS T0 Resources



ATLAS T1 Resources



ATLAS T2 Resources



Analysis computing model

Analysis model broken into two components

- Scheduled central production of augmented AOD, tuples & TAG collections from ESD

→ Derived files moved to other T1s and to T2s

- Chaotic user analysis of augmented AOD streams, tuples, new selections etc and individual user simulation and CPU-bound tasks matching the official MC production

→ Modest job traffic between T2s

Streaming

- This is *not* a theological issue
  - All discussions are about optimisation of data access
- TDR had 4 streams from event filter
  - primary physics, calibration, express, problem events
  - Calibration stream has split at least once since!
- At AOD, envisage ~10 streams
- We are now planning ESD and RAW streaming
  - Straw man streaming schemes (trigger based) being agreed
  - Will explore the access improvements in large-scale exercises
  - Are also looking at overlaps, bookkeeping etc

TAG Access

- Streaming is not the only way to partition and access a subset
- The selection and direct access to individual events is via a TAG database
  - TAG is a keyed list of variables/event
  - Overhead of file opens is acceptable in many scenarios
  - Works very well with pre-streamed data
- Two roles
  - Direct access to event in file via pointer
  - Data collection definition function
- Two formats, file and database
  - Now believe large queries require full database
    - Multi-TB relational database
    - Restricts it to Tier1s and large Tier2s/CAF
  - File-based TAG allows direct access to events in files (pointers)
    - Ordinary Tier2s hold file-based primary TAG corresponding to locally-held datasets

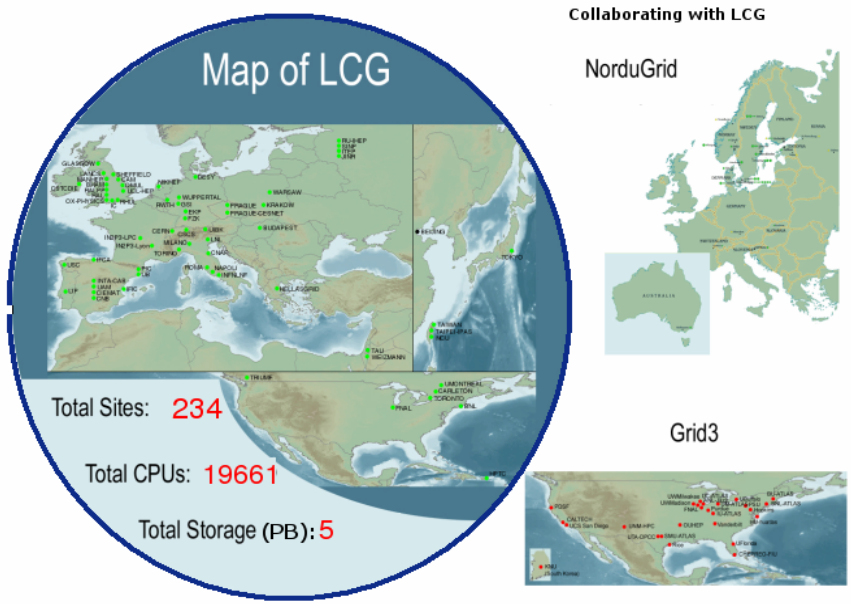
Group Analysis

- **Group analysis will produce**
  - Deep copies of subsets
  - Dataset definitions
  - TAG selections
- **Characterised by access to full ESD and perhaps RAW**
  - This is resource intensive
  - Must be a scheduled activity
  - Can back-navigate from AOD to ESD at same site
  - Can harvest small samples of ESD (and some RAW) to be sent to Tier 2s
  - Must be agreed by physics and detector groups
- **Big Trains**
  - Most efficient access if analyses are blocked into a 'big train'
  - Idea around for a while, already used in e.g. heavy ions
    - Each wagon (group) has a wagon master)production manager
    - Must ensure will not derail the train
  - Train must run often enough (every ~2 weeks?)

On-demand Analysis

- **Restricted Tier 2s and CAF**
  - Can specialise some Tier 2s for some groups
  - ALL Tier 2s are for ATLAS-wide usage
- **Role and group based quotas are essential**
  - Quotas to be determined per group not per user
- **Data Selection**
  - Over small samples with Tier-2 file-based TAG and AMI dataset selector
  - TAG queries over larger samples by batch job to database TAG at Tier-1s/large Tier 2s
- **What data?**
  - Group-derived EventViews (see later)
  - Root Trees
  - Subsets of ESD and RAW
    - Pre-selected or selected via a Big Train run by working group
- **Each user needs 14.5 kSI2k (about 12 current boxes)**
- **2.1TB 'associated' with each user on average**

ATLAS Grid Infrastructure



- **Three grids**
  - LCG
  - OSG
  - Nordugrid
- **Significant resources, but different middleware**
  - Teams working on solutions are typically associated to a grid and its middleware
- **In principle ATLAS resources are available to all ATLAS users**
  - But must also work locally

Transformations

- Common transformations is a fundamental aspect of the ATLAS strategy
- Overall no homogeneous system …. but a common transformation system allows to run the same job on all supported systems
  - All systems should support them
  - In the end the user can adapt easily to a new submission system, if he does not need to adapt his jobs
- Separation of functionality in grid dependant wrappers and grid independent execution scripts.
- A set of parameters is used to configure the specific job options
- A new implementation in terms of python is under way

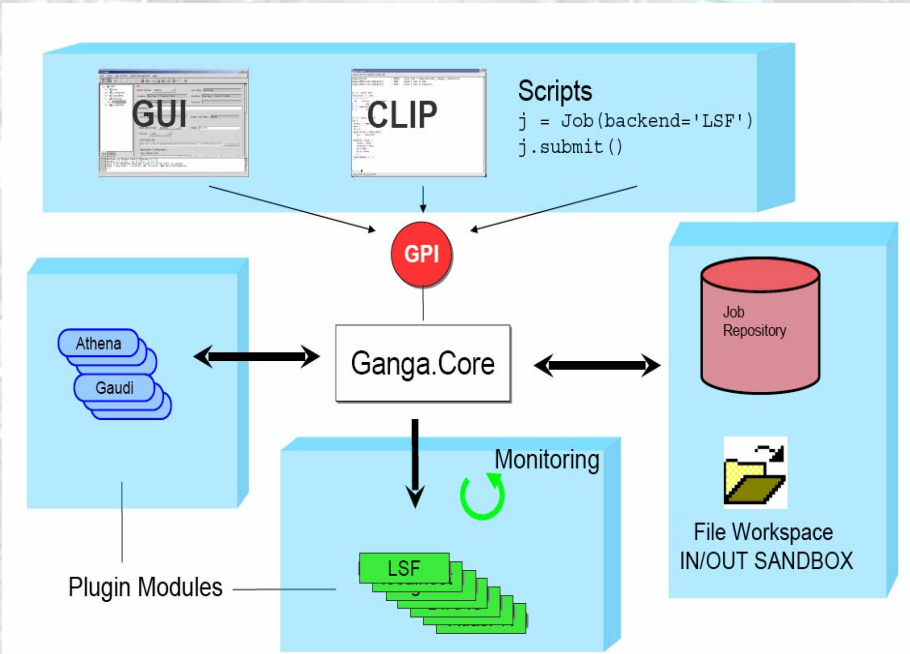
Distributed Analysis Tools

Gaudi/Athena and Grid Alliance

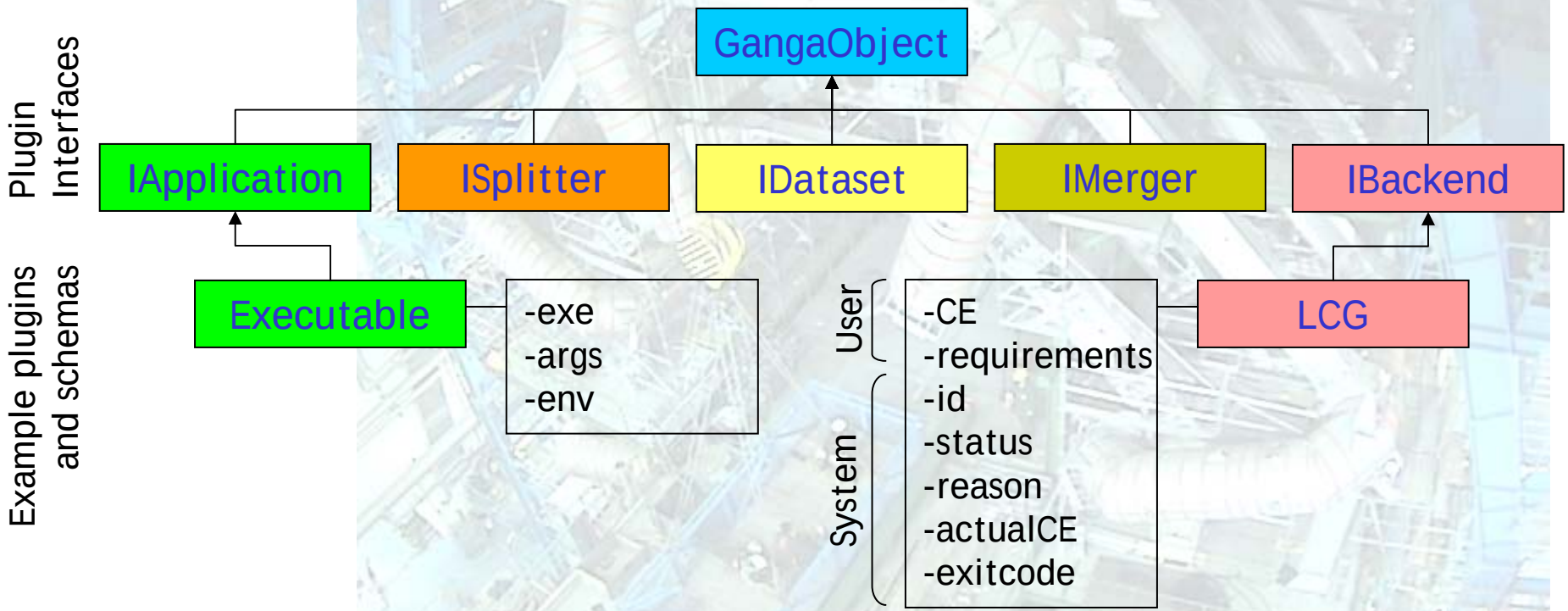
- **Distributed Analysis**
  - Data Management
    - Only now rolling-out in LCG, deployed in OSG
  - Site configuration
    - In LCG defining short/long/medium queues
    - OSG has PANDA task queue
  - Submission tools
    - In LCG use RB or Condor-G submission
    - In OSG, PANDA project provides scheduling
    - (Too?) Many possibilities here!
- **The full system design uses the GANGA framework and interface**
  - In the interim, partial solutions allow some aspects on some Grids
    - LJSF on LCG
    - ARC in NorduGrid
    - Clone of ATLAS Production system as a back-end?
      - Good for some applications, but restrictive
    - pAthena on OSG
  - GANGA provides CLI, GUI and Python scripting interface

GANGA – The GUI for the Grid

- Common project with LHCb
- Plugins allow definition of applications
  - Currently: Athena and Gaudi, ADA (DIAL)
- And backends
  - Currently: Fork, LSF, PBS, Condor, LCG, gLite, DIAL and DIRAC



GANGA


Plugin Interfaces
Example plugins and schemas
GangaObject
IApplication
ISplitter
IDataset
IMerger
IBackend
Executable
-exe
-args
-env
User
System
-CE
-requirements
-id
-status
-reason
-actualCE
-exitcode
LCG

ATLAS Data Management

- **Based on Datasets**
- **PoolFileCatalog API is used to hide grid differences**
  - On LCG, LFC acts as local replica catalog
  - Aims to provide uniform access to data on all grids
- **FTS is used to transfer data between the sites**
- **Evidently Data management is a central aspect of Distributed Analysis**
  - PANDA is closely integrated with DDM and operational
  - LCG instance was closely coupled with SC3
  - Right now we run a smaller instance for test purposes
  - Final production version will be based on new middleware for SC4 (FPS)

Dataset Access

- Collections of selected files comprise a dataset
  - Dataset will have a well defined associated luminosity (integer number of luminosity blocks)
- At present the primary source of dataset information is the simulation data from the production system
  - Production database suffices for now
- Soon (!) this will be from real data
  - Datasets will also be defined by physics groups, detector groups
  - Associated data will be modified for detector status, calibration info etc
  - → Requires a separate repository for dataset information and selection
- ATLAS Metadata Interface being developed for this
  - Keeps the production database secure
- Interaction between dataset and TAG selection being worked out

ATLAS Back-End Strategy

- **Production system**
  - Seamless access to all ATLAS grid resources
  - Not a long term solution to distributed analysis, but useful test bed and components

- **Direct submission to GRID**
  - LCG
    - LCG/gLite Resource Broker
    - CondorG
  - OSG
    - PANDA
  - Nordugrid
    - ARC Middleware

Production System

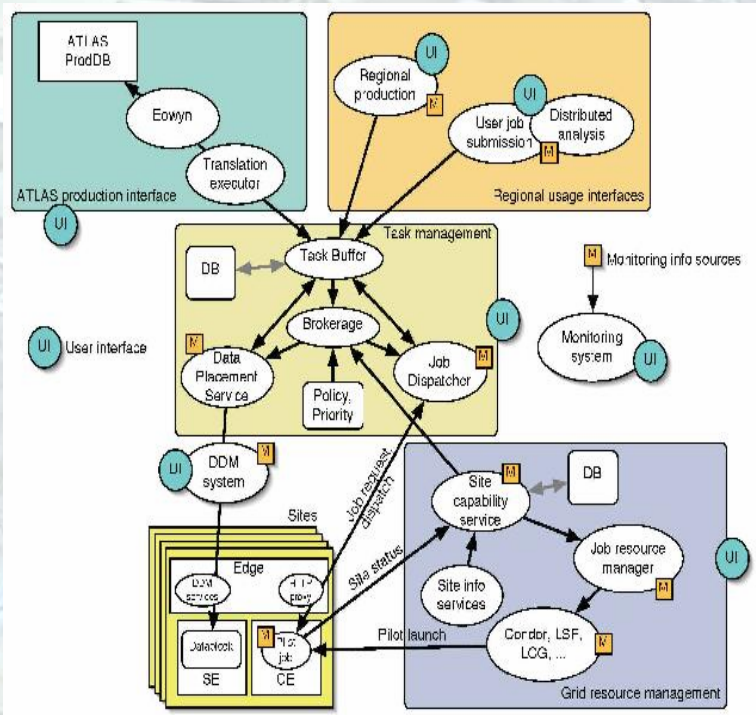
- **Provides a layer on top of the middleware**
  - Increases the robustness by the system
    - Retrials and fallback mechanism both for workload and data management
  - Our grid experience is captured in the executors
  - Jobs can be run in all systems
- **Redesign based on the experiences of last year**
  - New Supervisor - Eowyn
  - New Executors
  - Connects to new Data Management
- **Supports multiple submission mechanisms**

LCG

- **Resource Broker**
  - Scalability
  - Reliability
  - Throughput
- **Condor-G job submission**
  - Conceptually similar to LCG RB, but different architecture
  - Scaling by increasing the number of schedulers
  - No logging & bookkeeping, but a scheduler keeps track of the job
- **New gLite Resource Broker**
  - Bulk submission
  - Many other enhancements
  - Studied in ATLAS LCG/EGEE Taskforce

PANDA

- A system in itself for OSG
- Centrally, a new prodsys executor for OSG
  - Pilot jobs
  - Resource Brokering
  - Close integration with DDM
- Operational in the production since December



PANDA

- **Direct submission**
  - Regional production
  - Analysis jobs
- **Key features for analysis**
  - Analysis Transformations
  - Job-chaining
  - Easy job-submission
  - Monitoring
  - DDM end-user tool
  - Transformation repository

ARC Middleware

- **Standalone ARC client software – 13 MB Installation**
- **CE has extended functionality**
  - Input files can be staged and are cached
  - Output files can be staged
  - Controlled by XRSL, an extended version of globus RSL
- **Brokering is part of the submission in the client software**
  - Job delivery rates of 30 to 50 per min have been reported
  - Logging & bookkeeping on the site
- **Currently about 5000 CPUs, 800 available for ATLAS**

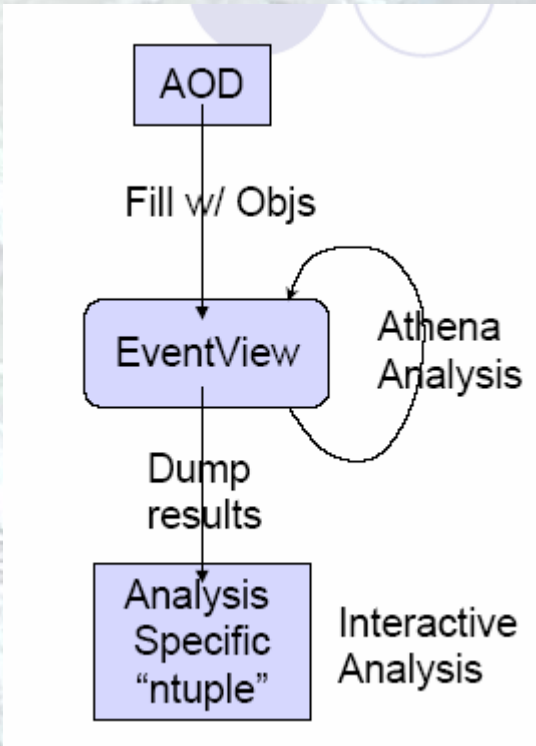
NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

Analysis Model

- Grid or non-Grid, the analyst needs software tools and an analysis model
  - First AM focus: "Modular Analysis"
    - Explore how to perform analysis in Athena
    - Basic ideas implemented. Growing user base.
    - Does modular analysis lead to faster and better analyses?
    - Now starting on interactive analysis.
- Next AM focus: "EDM".

AOD and EventViews

- General AOD 'heavy' for most analysis
- Some groups will slim AOD and some will augment it
- EventView allows group (or user) to customise the AOD info, make compound objects and persistify it
  - Avoids repeated user processing, avoids errors
  - Common interface to all particle types
  - Avoids unintentional multiple representation on the same particle
  - Provides easy way of making Athena Aware ntuple for private analysis
  - Allows consistent view of event (jet finding, vertexing, particle identification, inferred objects) to be presented and preserved.
- Two possible modes
  - *EventView for manipulating AOD objects*. EV helps organize analysis. Write your own EV tools/algs in C++.
  - *Modular analysis with EventView tools*. Most analysis task via generalized tools configured in python. Write your own tool when necessary.



Interactive Analysis

- Software framework ATHENA supports interactive analysis
- Can also run analysis from Python (e.g. PyRoot)
- GANGA/DA supports interactive ATHENA, and (being Python based) binds easily to PyRoot

Conclusions

- **Computing Model Data well evolved for placing Raw, ESD and AOD at Tiered centers**
  - **Still need to understand all the implications of Physics Analysis**
  - **Distributed Analysis and Analysis Model Progressing well**
- **SC4/Computing System Commissioning in 2006 is vital.**
- **Some issues will only be resolved with real data in 2007-8**