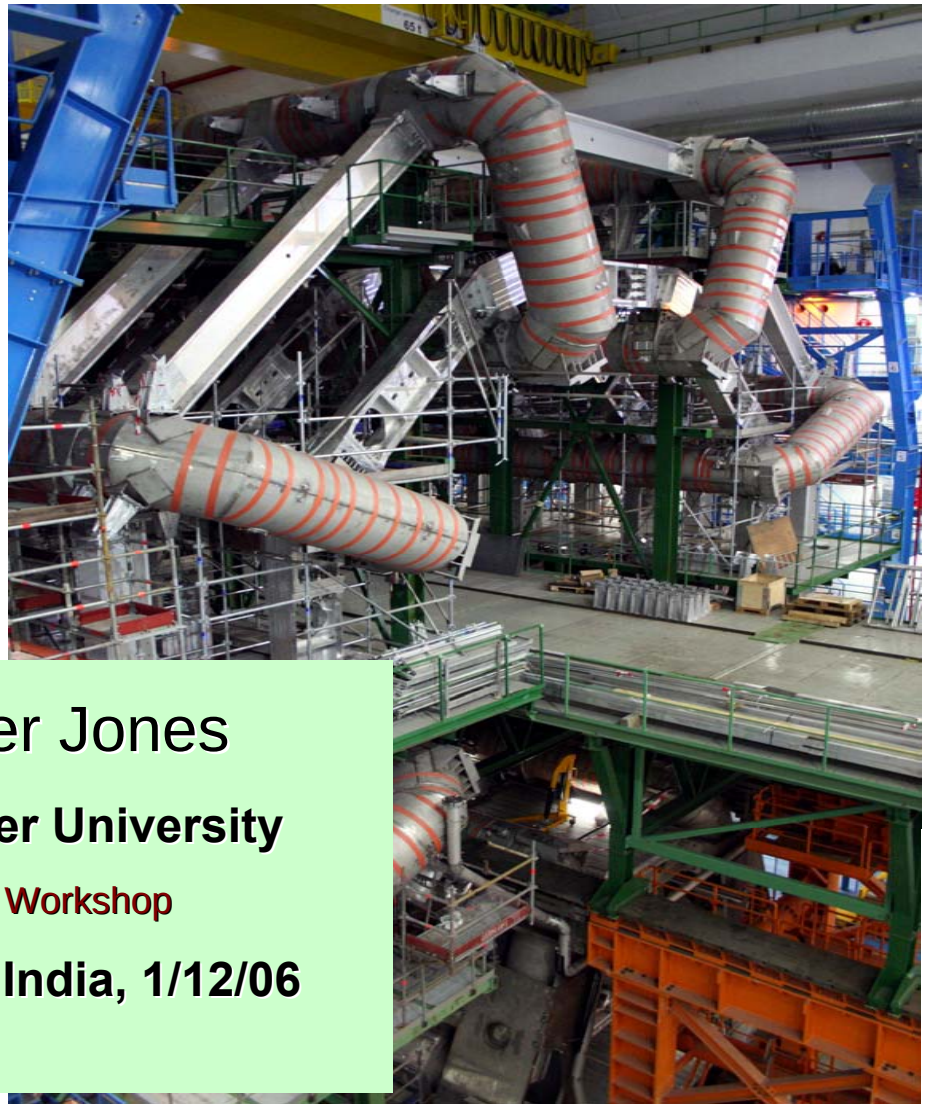
The ATLAS Computing Model



Roger Jones

Lancaster University

Grid Workshop

Mumbai, India, 1/12/06

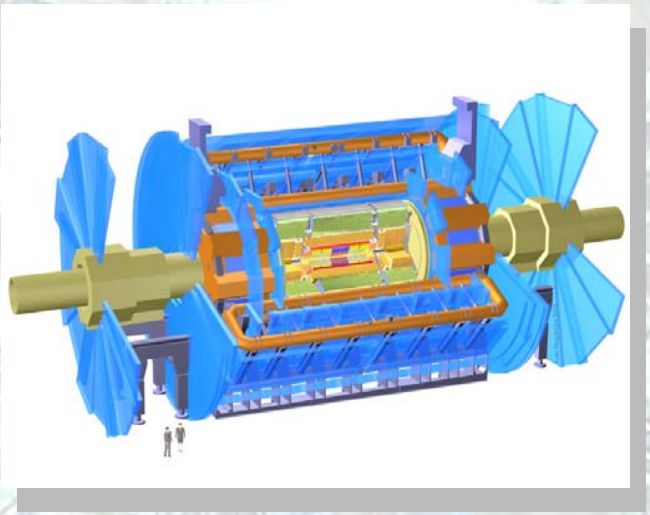
Overview

- Brief summary ATLAS Facilities and their roles
- Analysis modes and operations (most relevant to Tier 2s)
- Data selection
- Distributed Analysis Tools

Computing for ATLAS

SUSY

B Physics



Heavy Ions

Standard Model

Higgs

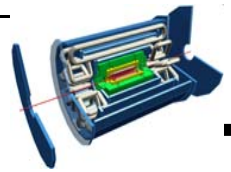
Exotics

Top quark

Computing Resources

- **Computing Model is well evolved, documented in C-TDR, but still evolves**
  - 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 just starting to be tested
    - Unlikely to know the real patterns until 2007/2008!
  - Still uncertainties on the event sizes , reconstruction time
  - Data access is being optimised
- **Lesson from the previous round of experiments at CERN (LEP, 1989-2000)**
  - Reviews in 1988 underestimated the computing requirements by an order of magnitude!

The Computing Model



~Pb/sec

Event Builder

10 GB/sec

Event Filter
~159kSI2k

- Some data for calibration and monitoring to institutes
- Calibrations flow back

450 Mb/sec

Tier 0

T0 ~5MSI2k

MSS

- Calibration
- First processing

~ 300MB/s/T1 /expt

Tier 1

US Regional Centre

Italian Regional Centre

Spanish Regional Centre (PIC)

UK Regional Centre (RAL)

MSS

- Reprocessing
- Group analysis

≥622Mb/s

Tier 2

Northern Tier ~200kSI2k

Tier2 Centre ~200kSI2k

Centre ~200kSI2k

Centre ~200kSI2k

- Analysis
- Simulation

≥622Mb/s

Lancaster ~0.25TIPS

pool

anchest er

Sheffield

Physics data cache

100 - 1000 MB/s

Desktop

Workstations

Average Tier 2 has ~25 physicists working on one or more channels

Roughly 3 Tier 2s should have the full AOD, TAG & relevant Physics Group summary data

Tier 2 do bulk of simulation

Facilities at CERN

- **Tier-0:**
  - Prompt first pass processing on express/calibration & physics streams with old calibrations - calibration, monitoring
  - Calibration tasks on prompt data
  - 24-48 hours later, process full physics data streams with reasonable calibrations
    → Implies large data movement from T0 →T1s
- **CERN Analysis Facility**
  - Access to ESD and RAW/calibration data on demand
  - Essential for early calibration
  - Detector optimisation/algorithmic development

Facilities Away from CERN

- **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
  - → Also Group Analysis - see later
- **~30 Tier 2 Centers distributed worldwide** Monte Carlo Simulation, producing ESD, AOD, ESD, AOD ➔ Tier 1 centers
  - On demand user physics analysis of shared datasets
  - Limited access to ESD and RAW data sets
  - Simulation
    - → Implies ESD, AOD, ESD, AOD ➔ Tier 1 centers
- **Tier 3 Centers distributed worldwide**
  - Physics analysis
  - Data private and local - summary datasets

Straw Man Profile

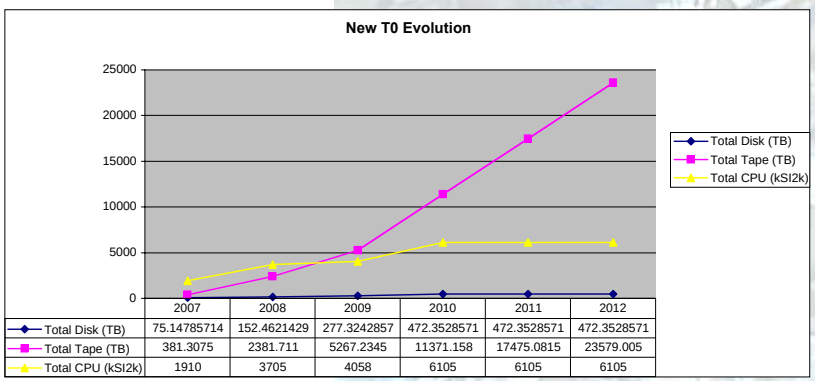
year	*energy*	*luminosity*	*physics beam time*
2007	450+450 GeV	$5x10^{30}$	protons - 26 days at 30% overall efficiency → $0.7*10^6$ seconds
2008	7+7 TeV	$0.5x10^{32}$	protons - starting beginning July $4*10^6$ seconds ions - end of run - 5 days at 50% overall efficiency → $0.2*10^6$ seconds
2009	7+7 TeV	$1x10^{33}$	protons:50% better than 2008 → $6*10^6$ seconds ions: 20 days of beam at 50% efficiency
2010	7+7 TeV	$1x10^{34}$	TDR targets: protons: → 10^7 seconds ions: → $2*10^6$ seconds

- This changes requirements from those in Technical Design Report
- We also have a better idea of:
  - Processing requirements
  - Event sizes for first data
  - Calibration requirements
- We are learning from the Computing System Commissioning

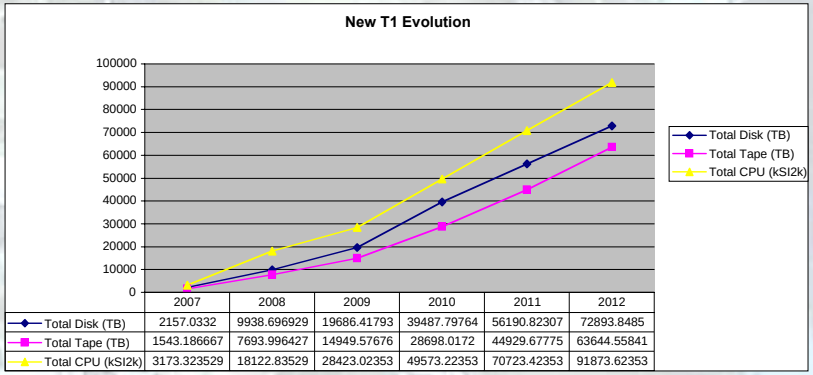
ATLAS Requirements start 2008, 2010

	CPU (MSi2k)		Disk (PB)		Tape (PB)	
	2008	2010	2008	2010	2008	2010
Tier-0	3.7	6.1	0.15	0.5	2.4	11.4
CERN Analysis Facility	2.1	4.6	1.0	2.8	0.4	1.0
Sum of Tier-1s	18.1	50	10	40	7.7	28.7
Sum of Tier-2s	17.5	51.5	7.7	22.1		
Total	41.4	112.2	18.9	65.4	10.5	41.1

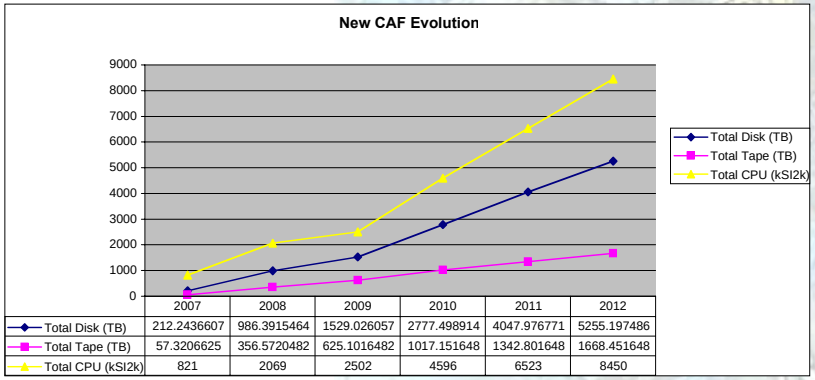
Evolution



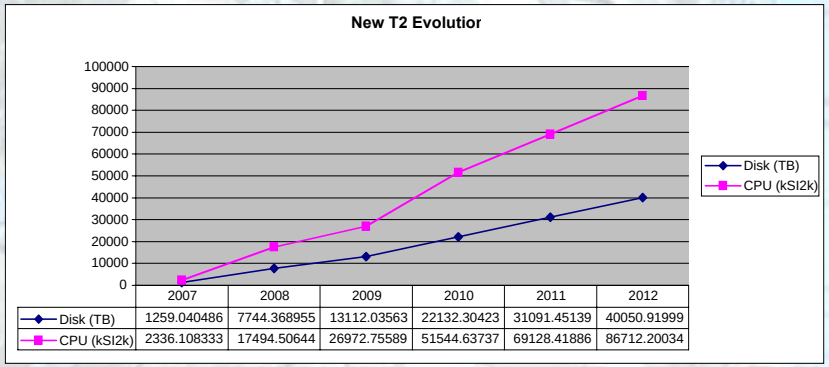
	2007	2008	2009	2010	2011	2012
Total Disk (TB)	75.14785714	152.4621429	277.3242857	472.3528571	472.3528571	472.3528571
Total Tape (TB)	381.3075	2381.711	5267.2345	11371.158	17475.0815	23579.005
Total CPU (kSI2k)	1910	3705	4058	6105	6105	6105



	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



	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



	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

T1/T2 Group

- This has been trying to describe:
  - Network traffic to T1s and T2s at each specific site
  - Required T2 storage at associated T1s

➢ **Note: this is also evolving**

- ➢ The new schedule is included
- ➢ We also know that some pledges will change
- ➢ The sharing of the Tier 1 load is still under discussion (but the one in the current megatable will change)

Observations

- **The wide range of T1 sizes introduces some inefficiencies compared with the ideal case**
  - **Some T1s will have a large load because of their chosen T2s**
  - **Some are underused and we continue to negotiate better balance**
- **The T2s tend to have too high a cpu/disk ratio**
  - **Optimal use of the T2 resources delivers lots of simulation with network and T1 disk consequences (although the higher cpu/event will reduce this)**
  - **The T2 disk only allows about ~60% of the required analysis**
  - **Other models would seriously increase network traffic**
- **BNL full ESD copy has network implications elsewhere**

Data Flow

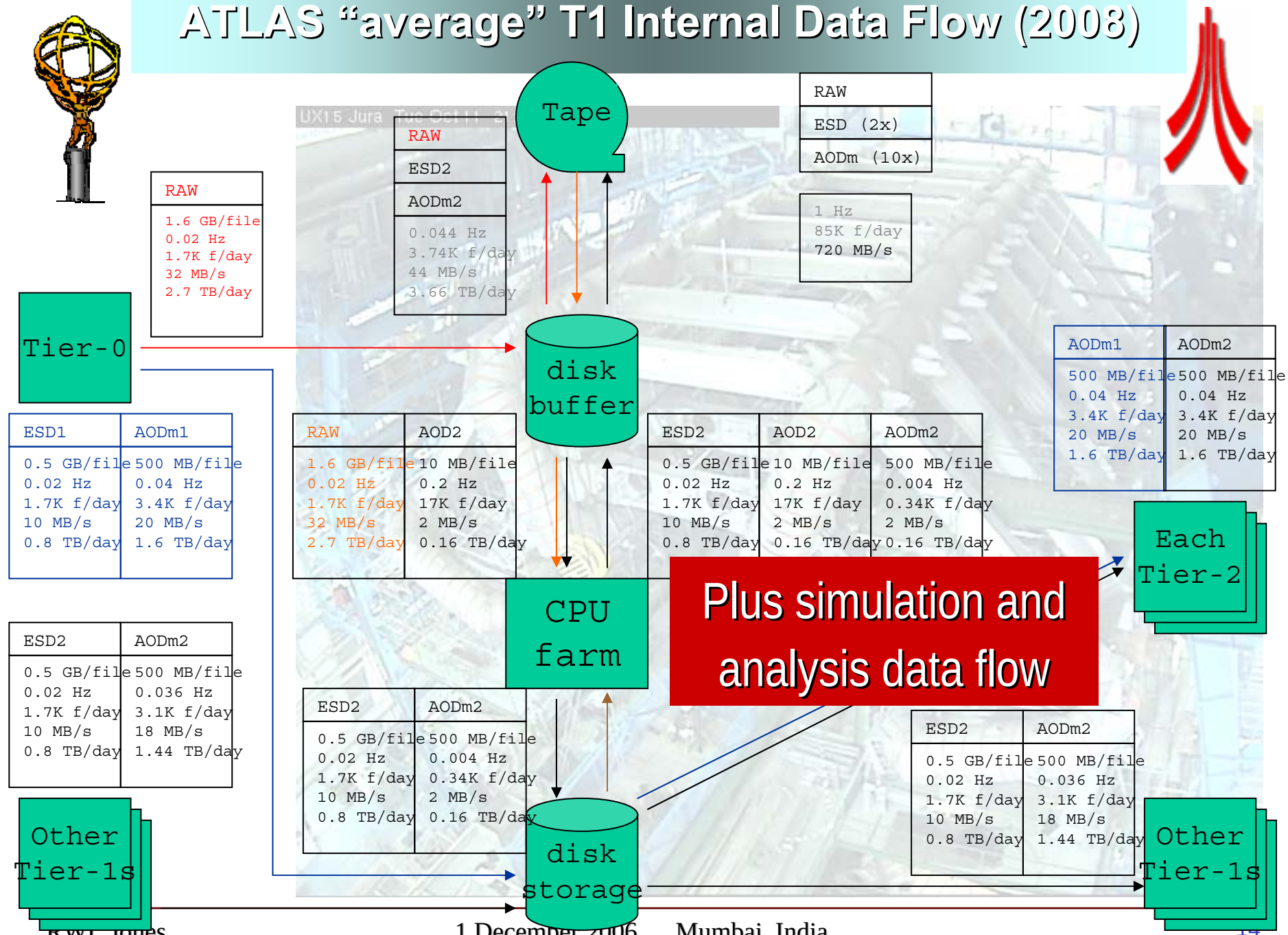
- EF farm ➔ T0
  - 320 MB/s continuous
- T0 Raw data ➔ Mass Storage at CERN
- T0 Raw data ➔ Tier 1 centers
- T0 ESD, AOD, TAG ➔ Tier 1 centers
  - 2 copies of ESD distributed worldwide

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 "average" T1 Internal Data Flow (2008)



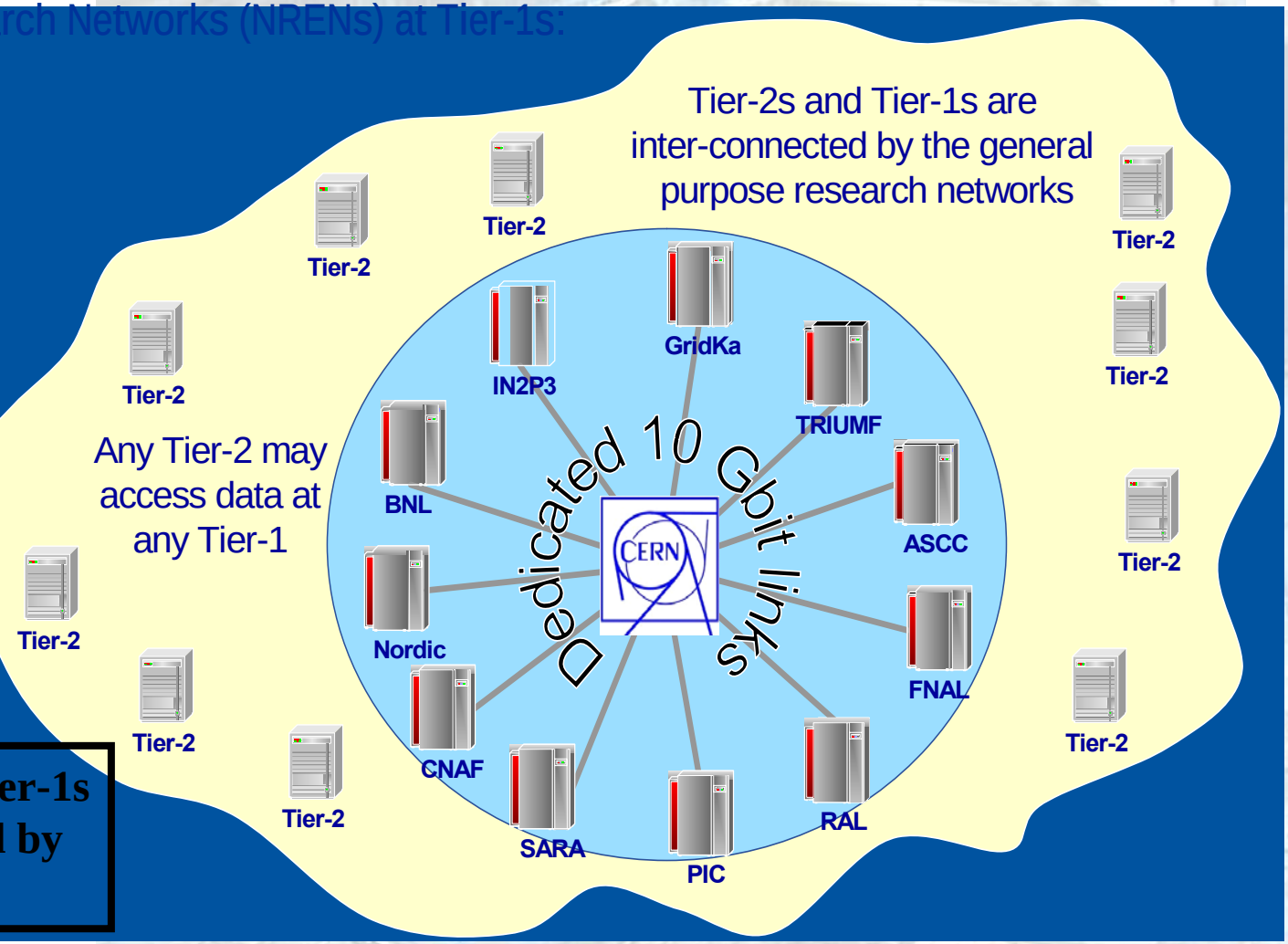
WLCG: Tier-1s

Tier-1 Centre	Experiments served with priority			
	ALICE	ATLAS	CMS	LHCb
TRIUMF, Canada		X		
GridKA, Germany	X	X	X	X
CC, IN2P3, France	X	X	X	X
CNAF, Italy	X	X	X	X
SARA/NIKHEF, NL	X	X		X
Nordic Data Grid Facility (NDGF)	X	X	X	
ASCC, Taipei		X	X	
RAL, UK	X	X	X	X
BNL, US		X		
FNAL, US			X	
PIC, Spain		X	X	X

Tier-0 -1 -2 Connectivity

Research Networks (NRENs) at Tier-1s:

- ASnet
- LHCnet/ESnet
- GARR
- LHCnet/ESnet
- RENATER
- DFN
- SURFnet6
- NORDUnet
- RedIRIS
- UKERNA
- CANARIE



Tier-0 and Tier-1s are connected by LHCOPN

Data Location

- The model assumes that most data is placed
- Jobs go to the data, not data to the jobs
  - Tier 2 capacity is collective, although some regional specialisation for calibration, some physics groups
- On average, 3 nearby Tier 2s hold the full AOD
  - There should be very little long-distance T2-T2 traffic
- Over half of the RAW and ESD in the Tier 2s (and on disk at the Tier 1) should be pre-decided
  - The rest should be requested via production manager of physics/detector group
  - Tape access will be carefully controlled and optimised
  - Data from disk in a few hours, data from tape in ~ 1 week

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

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
- **Most ATLAS Tier 2 data should be 'placed' and have a lifetime of order months**
  - Job must go to the data
  - This means the Tier 2 bandwidth is lower than if you pull data to the job
- **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/SAN/pAOD
  - Root Trees
  - Subsets of ESD and RAW
    - Pre-selected or selected via a Big Train run by working group

Optimised Access

- RAW, ESD and AOD will be streamed to optimise access
- 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

Streaming

- 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

ATLAS Data Management

- Based on Datasets = defined set of files (see David's talk about our Data Management)
- 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
- Catalogues and ATLAS-specific services are restricted to associated Tier 1s
- FTS is used to transfer data between the sites
  - Tier 2 must define endpoints and also install end-user tools
- 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)

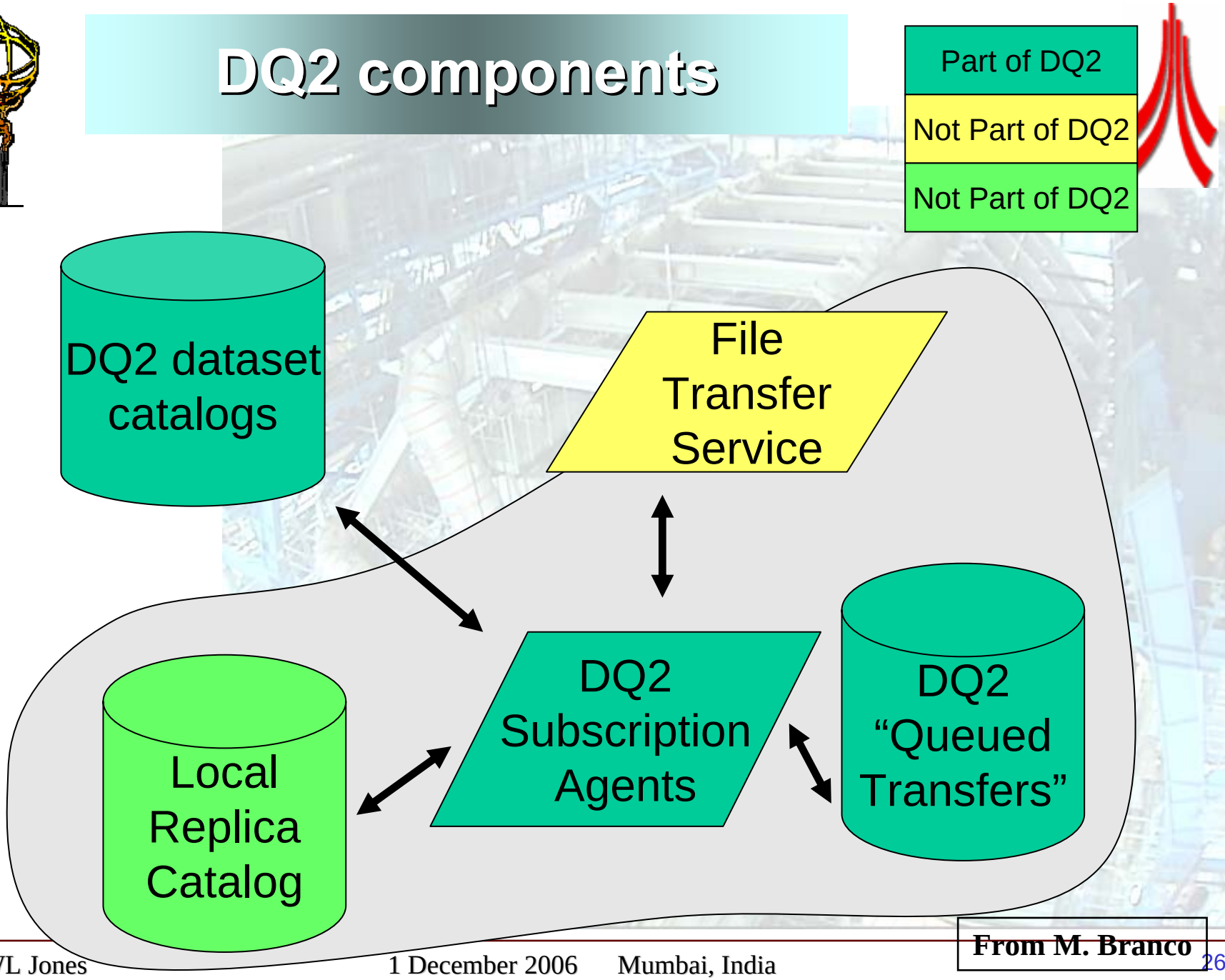
Dataset Access

- 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

DQ2: ATLAS Distributed Management system

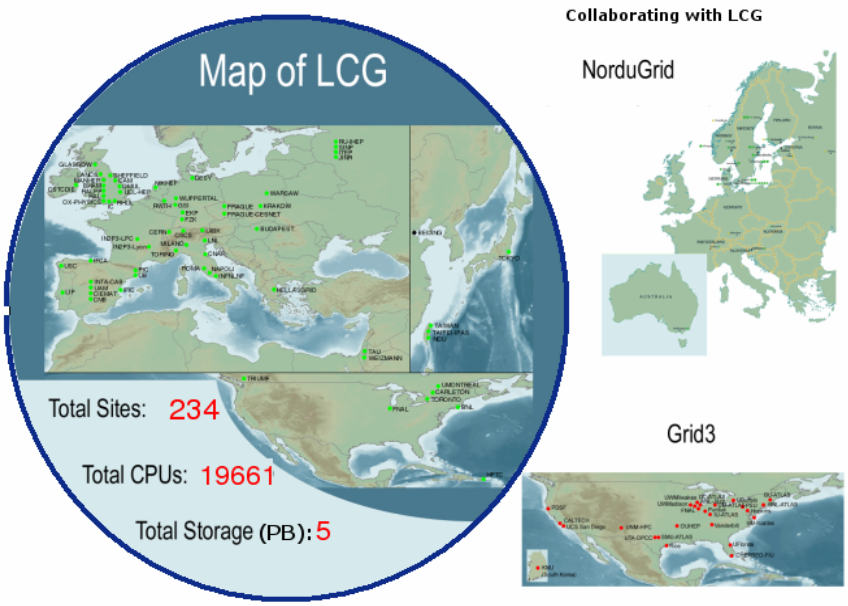
- **DQ2, is built on top of Grid data transfer tools, is based on:**
  - **Hierarchical definition of files and datasets**
    - **Through dataset catalogs**
  - **Datasets as the unit of file storage and replication**
    - **Supporting dataset versions**
  - **Distributed file catalogues at each site**
  - **Automatic data transfer mechanisms using distributed site services**
    - **Dataset subscription system**
- **DQ2 allows the implementation of the basic ATLAS Computing Model needs:**
  - **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**

DQ2 components



From M. Branco

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 (now out of use)
    - ARC in NorduGrid
    - Clone of ATLAS Production system as a back-end?
      - Good for some applications, but restrictive
    - pAthena on OSG (proof of principle on LCG also)
  - GANGA provides CLI, GUI and Python scripting interface

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

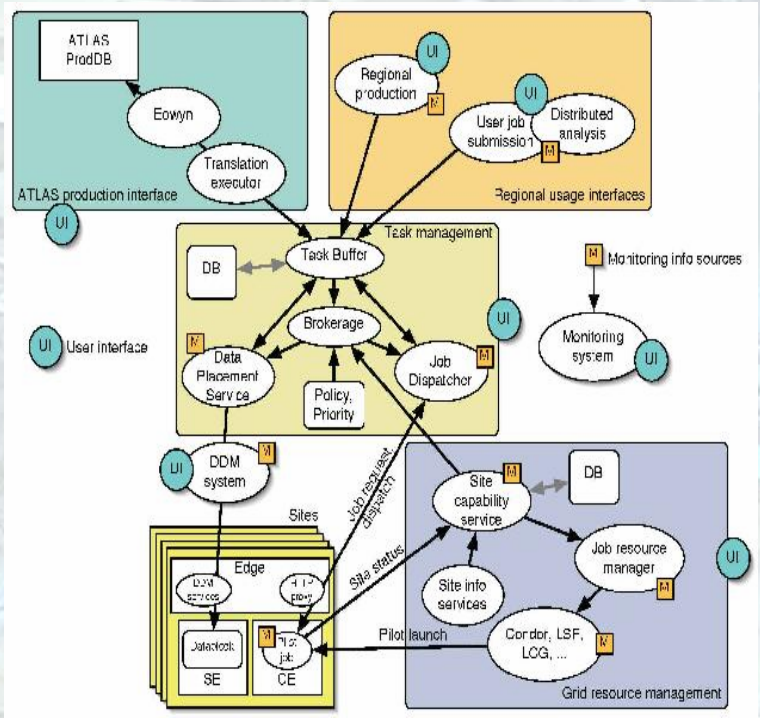
- **Provides a layer on top of the middleware**
  - Increases the robustness by the system
    - Retrials and fallback mechanism both for workload and data management
  - Our grid experience is captured in the executors
  - Jobs can be run in all systems
- **Redesign based on the experiences of last 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**

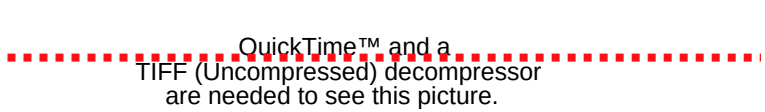
NORDUGRID
Grid Solution for Wide Area Computing and Data Handling

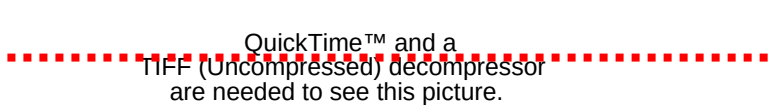
Tier-0 Scaling test (October 2006)

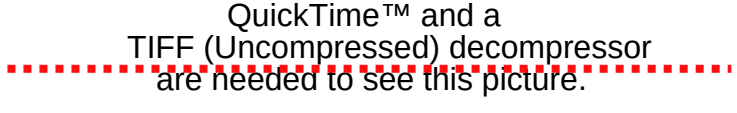
- Put in place monitoring system allowing sites to see their rates (disk/tape areas), data assignments, errors in the last hours, per file, dataset, …
- FTS channels in place between T0 and T1 and now progressing between T1 and T2s
  - By ‘pressure’ of regional contacts
- Start of the exercise marked by deployment of new DQ2 version (LCG and OSG sites)
  - Hopefully this is last major new release for near future
    - Many improvements to the handling of FTS requests
- Tier-2s participate on a “voluntary basis”.

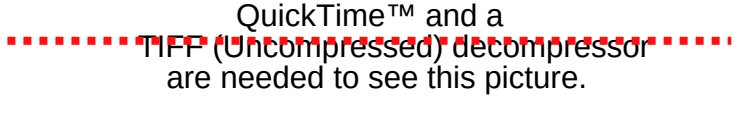
Tier-0 Internal Transfers (Oct)

- Data flows and operations can be maintained for ~ week

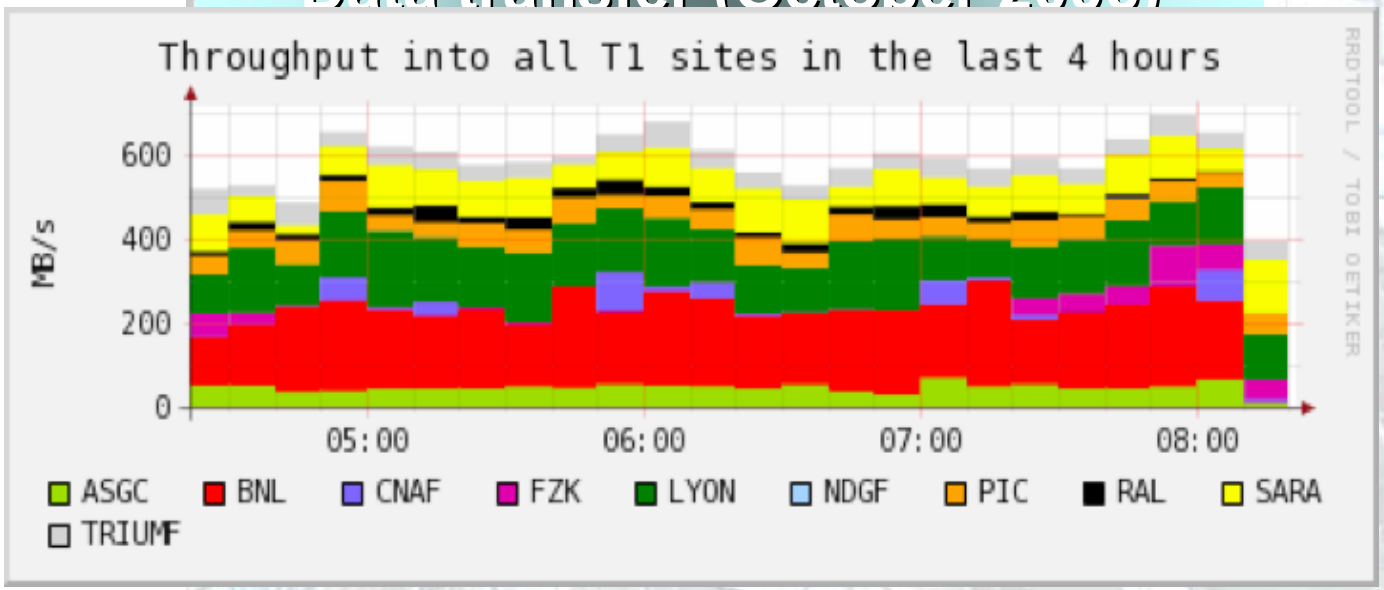


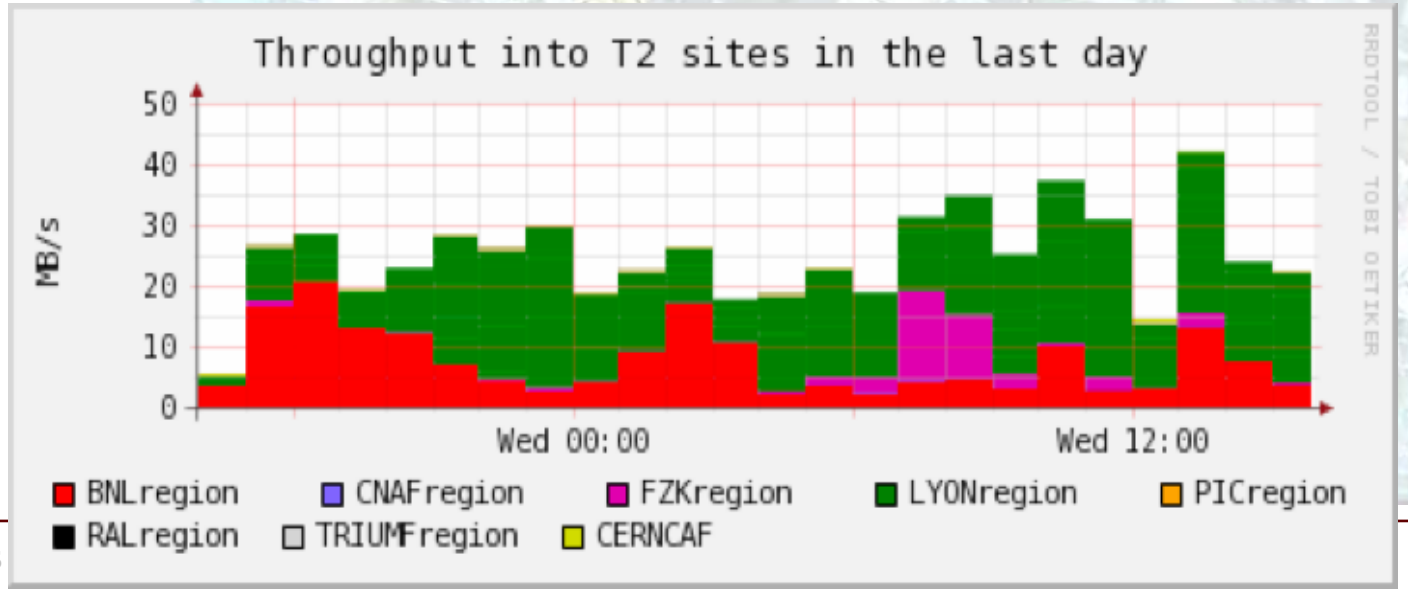






Data transfer (October 2006)





Computing System Commissioning (2006)

- The high-level goals of the Computing System Commissioning operation during 2006
  - A running-in of continuous operation not a stand-alone challenge
- Main aim of CSC is 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
  - Event reconstruction and data distribution
  - Distributed access to the data for analysis
- 60 M events have already been produced; new production of 10M events will be done from now until the end of the year.
- At the end of 2006 we will have a working and operational system, ready to take data with cosmic rays at increasing rates

The Calibration Data Challenge (CDC)

Fully simulate ~ 20M events (mainly SM processes: $Z \rightarrow ll$, QCD di-jets, etc.) with "realistic" detector

"Realistic" ≡

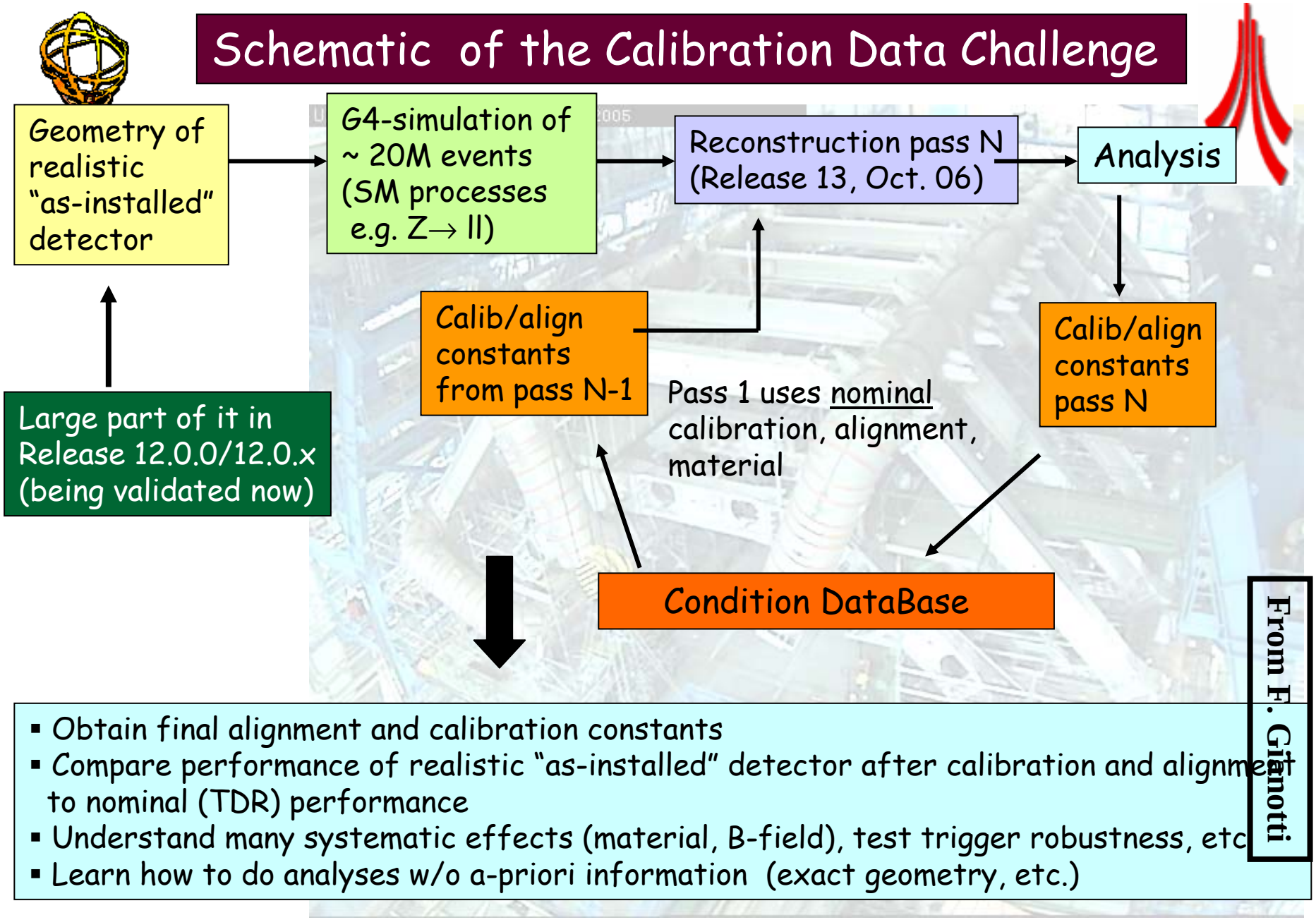
1) As installed in the pit : already-installed detector components positioned in the software according to survey measurements
2) Mis-calibrated (e.g. calo cells, R-t relations) and mis-aligned (e.g. SCT modules, muon chambers); include also chamber/module deformations, wire sagging, HV imperfections, etc.

Use the above samples and calibration/alignment algorithms to calibrate and align the detector and recover the nominal ("TDR") performance.
Useful also to understand the trigger performance in more realistic conditions.

Includes exercise of (distributed) infrastructure: Condition DB, bookkeeping, etc.

Scheduled for Spring 2007; needs ATLAS Release 13 (February 2007)

Schematic of the Calibration Data Challenge



- Obtain final alignment and calibration constants
- Compare performance of realistic "as-installed" detector after calibration and alignment to nominal (TDR) performance
- Understand many systematic effects (material, B-field), test trigger robustness, etc
- Learn how to do analyses w/o a-priori information (exact geometry, etc.)

From F. Gianotti

"The Dress rehearsal"

A complete exercise of the full chain from trigger to (distributed) analysis, to be performed in 2007, a few months before data taking starts

- Generate O(10^7) evts: few days of data taking, ~1 pb^{-1} at L=10^{31} cm^{-2} s^{-1}
- Filter events at MC generator level to get physics spectrum expected at HLT output
- Pass events through G4 simulation (realistic "as installed" detector geometry)
- Mix events from various physics channels to reproduce HLT physics output
- Run LVL1 simulation (flag mode)
- Produce byte streams → emulate the raw data
- Send raw data to Point 1, pass through HLT nodes (flag mode) and SFO, write out events by streams, closing files at boundary of luminosity blocks.
- Send events from Point 1 to Tier0
- Perform calibration & alignment at Tier0 (also outside ?)
- Run reconstruction at Tier0 (and maybe Tier1s ?) → produce ESD, AOD, TAGs
- Distribute ESD, AOD, TAGs to Tier1s and Tier2s
- Perform distributed analysis (possibly at Tier2s) using TAGs
- MCTruth propagated down to ESD only (no truth in AOD or TAGs)

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**