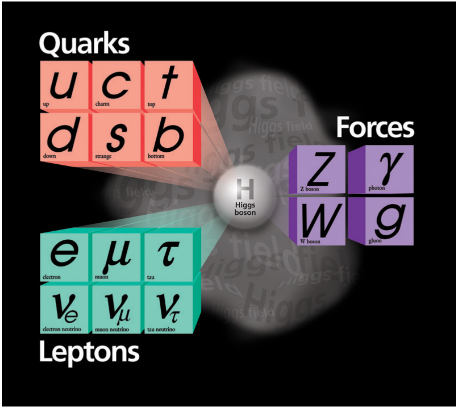
Grid Computing for LHC

Johannes Elmsheuser

Ludwig-Maximilians-Universität München

08 September 2010/Karlsruhe

Standard Model of Particle Physics

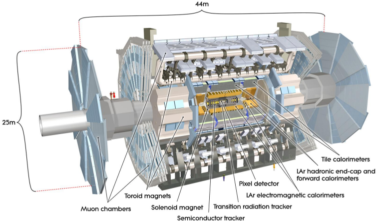


- Building blocks of matter and their interactions - describe well current observations, but missing pieces
- Higher energy: Reproduce conditions of early Universe
- TeV energy scale: Expect breakdown of current calculations unless a new interaction or phenomenon appears
- Many theories, but need data to distinguish between them

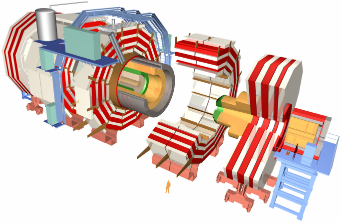
The LHC and Experiments

4 LHC Experiments

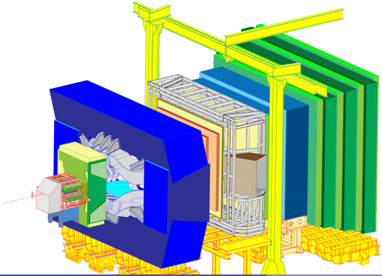
ATLAS



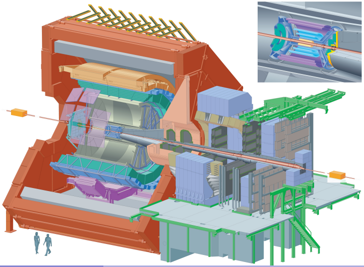
CMS



LHCb



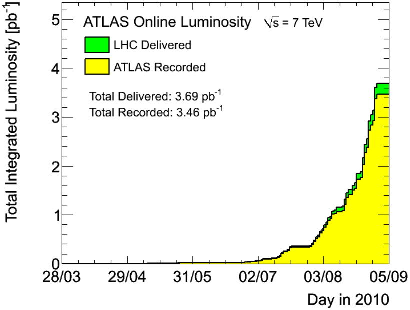
ALICE

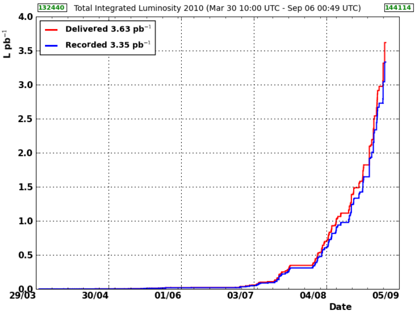


Detectors built and operated by a large team

Worldwide Collaboration of over 3000 physicists and engineers in ATLAS and CMS each + similar in LHCb and ALICE

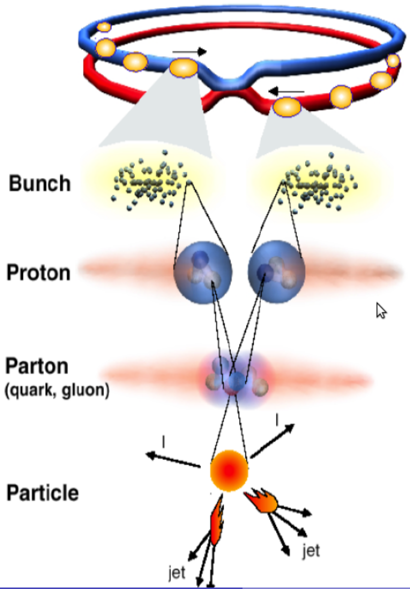
Recorded Luminosity so far 2010





- 2010: 30-50 pb^{-1}, ,,Re-discover" Standard Model: J/ψ, W, Z, top
- 2011: up to 1 fb^{-1} at $\sqrt{s} = 7(8)$ TeV

Collisions at the LHC



Proton-Proton-Kollisionen
2835 Teilchenbündel (Bunch)

10^{11} Protonen / Bunch
Kollisionsrate 40 MHz (25 ns)

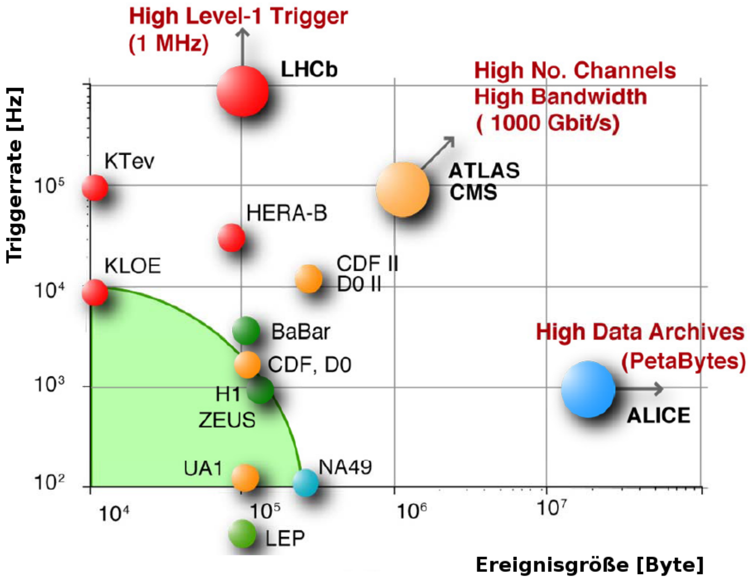
Schwerpunktsenergie 14 TeV
(= 7400 x Ruheenergie der kollidierenden Teilchen)

Schwerpunktsenergie der kollidierenden Quarks und Gluonen bis einige TeV

~25 pp-Kollisionen pro Bunch-Kollision

Interessante Ereignisse: $10^{-9} – 10^{-11}$ unterdrückt!

Trigger and Eventsizes



Challenges in Data Analysis

Data volumes

- LHC experiments produce and store several PetaBytes/year

CPUs

- Event complexity (large number of channels) and number of users demands: at least 100000 fast CPUs based on computing model

Software

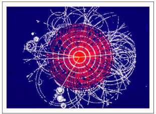
- The experiments have complex software environment and framework

Connectivity

- Data should be available 24/7 at a high bandwidth

Average Analysis at LHC I

Higgs-Search: $\mathrm{H} \rightarrow \mathrm{WW}^{(*)} \rightarrow \mu^+ \nu_\mu \mu^- \bar{\nu}_\mu$ für $1\,\mathrm{fb}^{-1}$



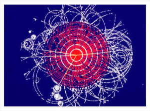
Monte Carlo events needed :

- 4 mass points: $m_H = 130 - 190\,\mathrm{GeV}$: 100k + 500k Systematic studies
- Background: Z/γ^*: 2M, $t\bar{t}$: 500k, WW+WZ+ZZ: 200k, W+jets: 1M
- Total: 4.3M
- Time for simulation: 200h @ 10000 CPUs with 0.5h/event (no overhead)

Data:

- 10^9 Events/year
- $\approx$ 50d time for reconstruction @ 10000 CPUs with 45s/event

Average Analyse at LHC II



Analysis:

- 10^6 data events from trigger and skim pre-selection
- Estimated time:
  - 1 week MC+data at 1 CPU with 10Hz
  - 4h MC+data at 1000 CPUs (Tier2-share)
  - Optimization of analysis demands much more time

Scaling up:

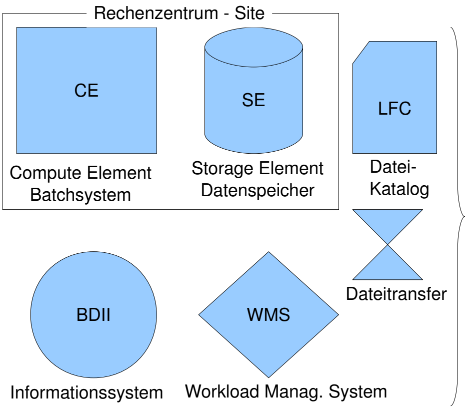
- Assume 2000 physicist with same analysis
- Time: 3h at 100000 CPUs
- Shown analysis is not the most time consuming
- Analysis with jets need much more CPU-time
- All given time: without additional overhead

Grid Infrastuctures

- Heterogeneous grid environment based on 3 grid infrastructures:

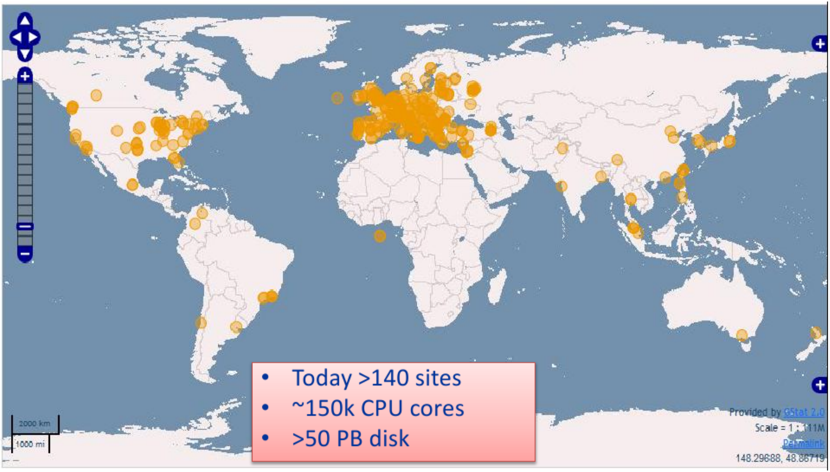
- e.g. 3 major ATLAS Grid areas:
  - Production System (Panda): centralized MC simulation and Data reconstruction
  - Distributed Data Managment (DQ2): centralized data movement
  - Distributed User Analysis: de-centralized individual analysis

What is needed - some grid components:



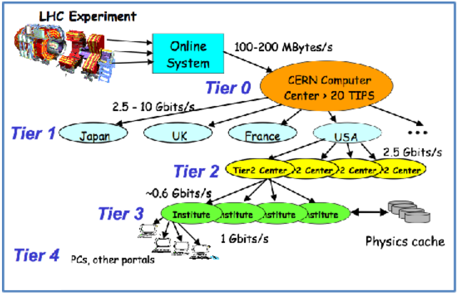
Grid Middleware

Worldwide resources

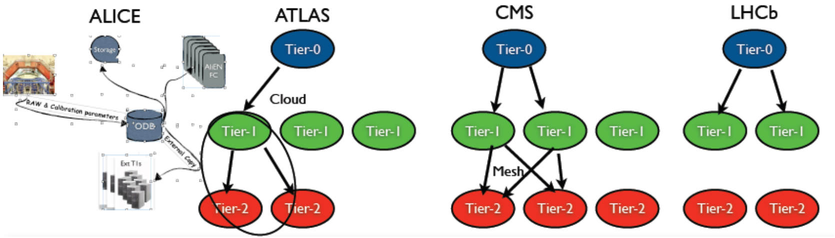


- Today >140 sites
- ~150k CPU cores
- >50 PB disk

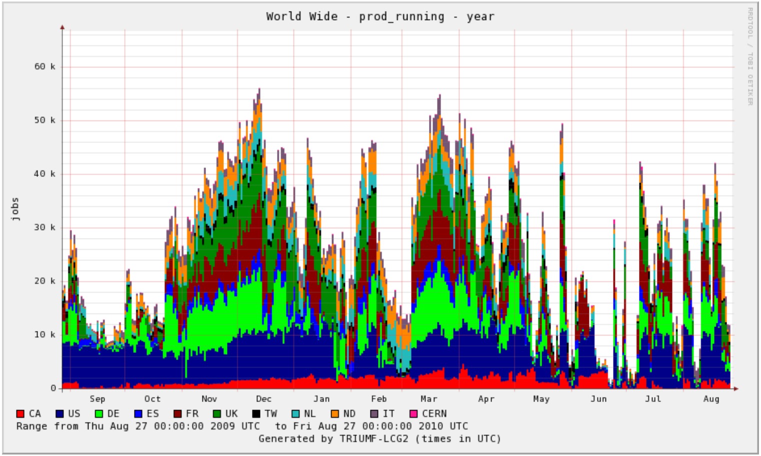
Experiment Models and Tier Structure



- Models all based on the MONARC tiered model of 10 years ago
- Several significant variations, however



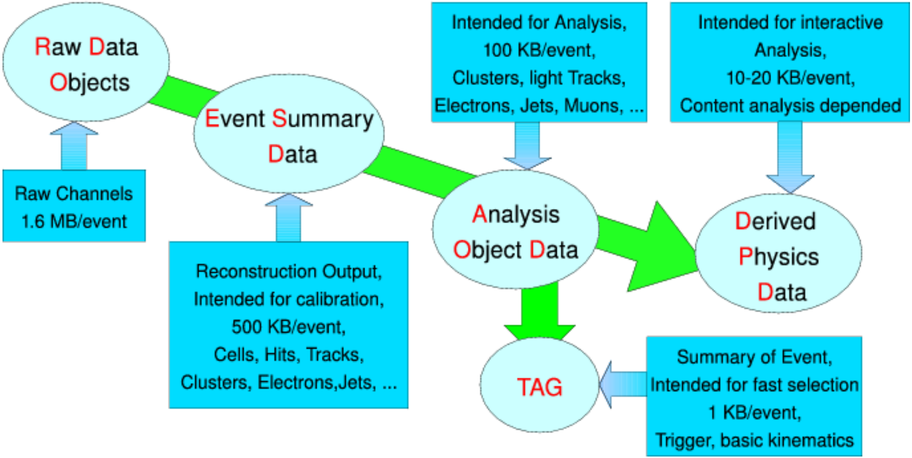
ATLAS Production System Jobs - last year



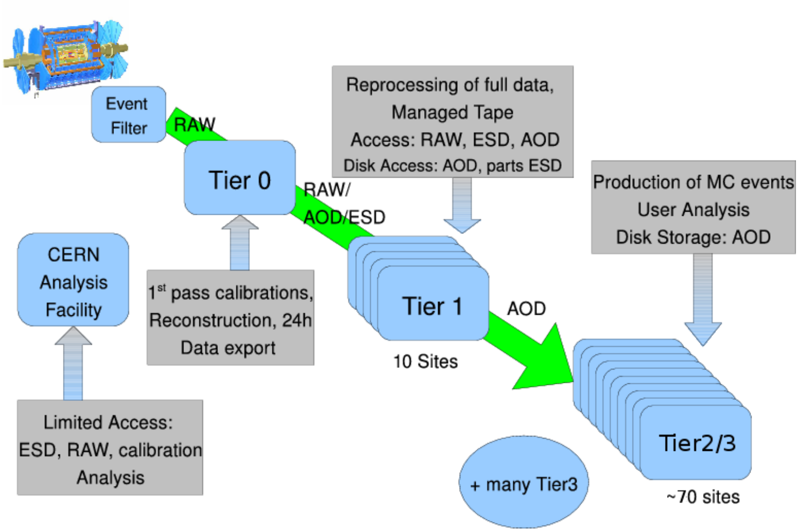
Up to 50k simulaneous jobs - structure related to SW releases and simulation campaigns

Event data model: ATLAS

Refining the data by: Add higher level info, Skin, Thin, Slim



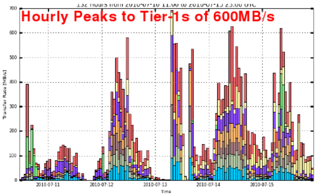
Data Distribution: ATLAS

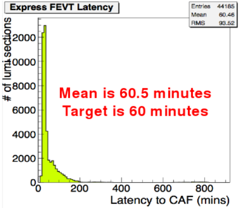


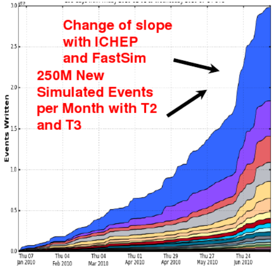
≈ 80 Tier1/2/3 sites managed by DQ2 right now

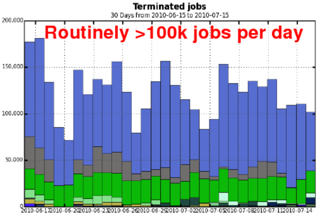
Data Processing, Transfer and Analysis Activities

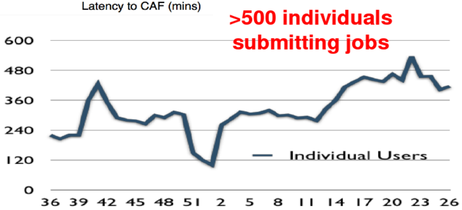
Excellent experience so far: the whole offline and computing organization + GRID infrastructure performing very well.





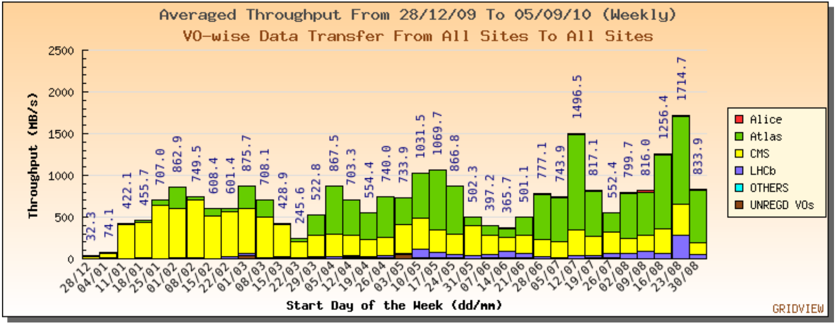






Data transfers 2010

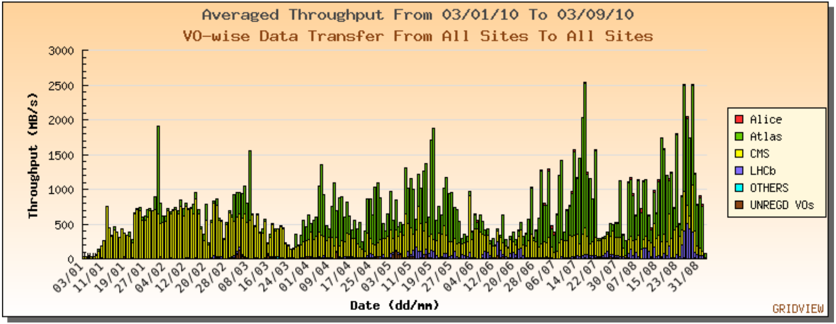
Data transfer capability today able to manage much higher bandwidths than expected and planned



- Data transfer rates per week in 2010

Data transfers 2010

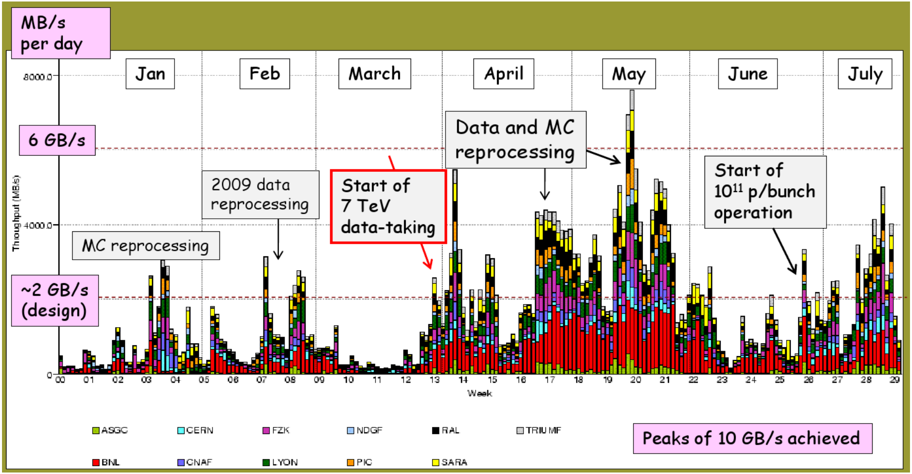
Data transfer capability today able to manage much higher bandwidths than expected and planned



- Data transfer rates per day in 2010

ATLAS Data Transfers

Total throughput of ATLAS data through the Grid: 1 Jan - 31 July 2010



Grid Job Submission

Naive assumption: Grid $\approx$ large batch system

- Provide complicated job configuration for Workload Management System
- Find suitable experiment software, installed in the Grid (100 CEs, 30 Software versions)
- Locate the data on different storage elements
- Job splitting, monitoring and book-keeping
- etc.

$\Longrightarrow$ Need for automation and integration of various different components

Several ways lead into the Grid !

Grid Software in the LHC experiments

Every experiment has built own system on top of grid middleware:

- Grid infrastructure middleware - different workflows
- work-arounds for grid middleware problems
- Often batch-like analysis, Alice uses PROOF in addition

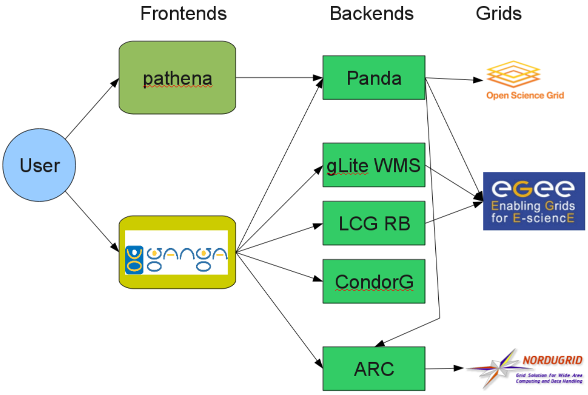
Similar SW stack in experiments:

- SW environement in C/C++ and Root
- Analysis-Grid-Tools in script language (Python)
- Grid data transfers (SRM, FTS)
- Workload Management (glite WMS)

Similar Ansatz, but experiment dependent:

- Crab (CMS), Ganga (LHCb/ATLAS)
- Various monitoring packages
- Pilot Job Workload Management:
  - e.g. Dirac (LHCb), Panda (ATLAS), Alien (Alice))
- Data managment:
  - e.g. Phedex (CMS), DQ2 (ATLAS)

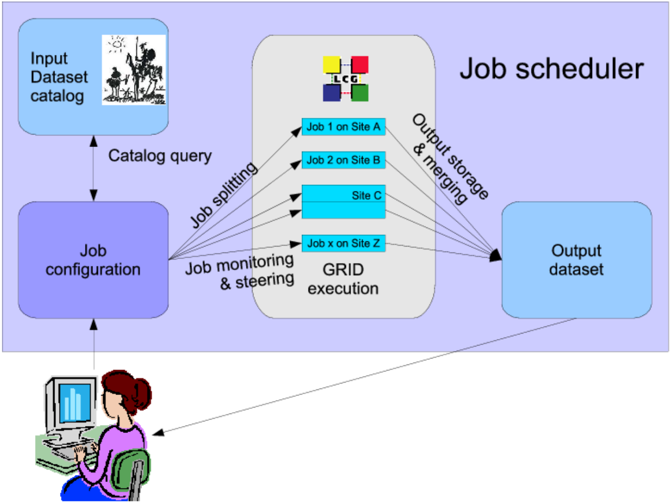
ATLAS Distributed Analysis



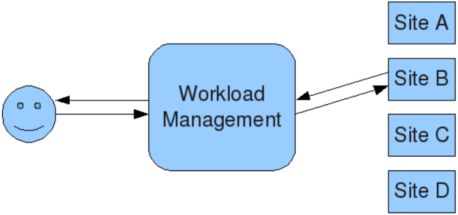
Data is centrally being distributed by DQ2 - Jobs go to data

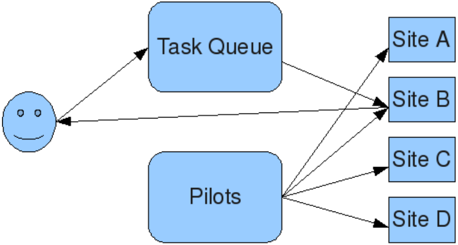
Distributed Analysis: Ganga

How to combine all different components: Job scheduler/manager: GANGA



Job Scheduling





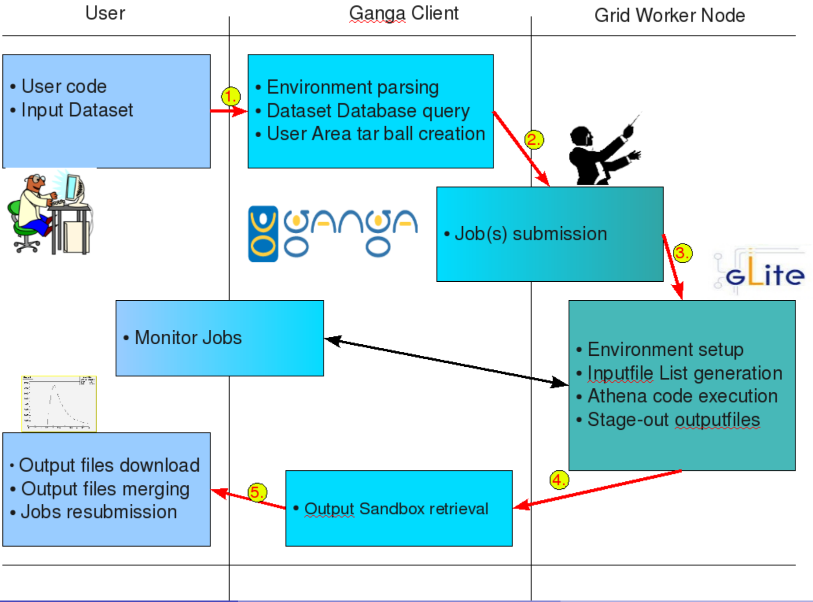
Job Push mode

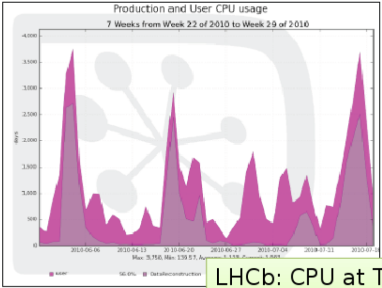
- Dependent on information system and site status
- Decentralized
- Better control of site policies
- Ganga: LCG and NG backend

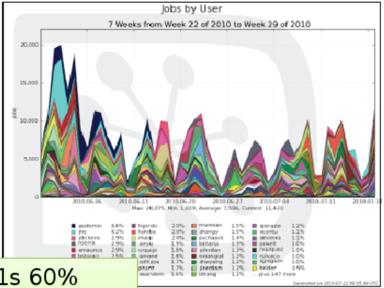
Job Pull mode

- Workarounds for some Grid problems
- Data pre-staging
- Panda clients or Ganga Panda backend

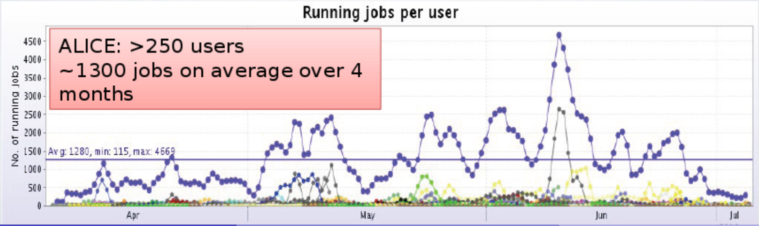
Example job workflow







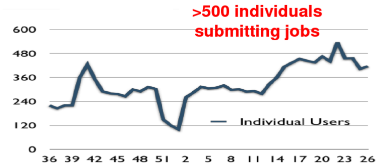
LHCb: CPU at Tier 1s 60% user and 40% reconstruction; ∅200 users 30k jobs/day



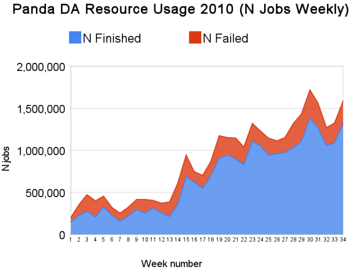
ALICE: >250 users
~1300 jobs on average over 4 months

Number of analysis jobs II

CMS:



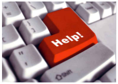
ATLAS:



- Compare ATLAS number with daily $\sim$ 50-100k production jobs
- Since start of 7 TeV collisions large increase of jobs and users

Current user problems and Support

User support is very important but time consuming



Central ticketing system for site or grid middleware probleme: GGUS

- Site or experiment experts try to solve problems
- Often ,,one-way" communication

Support mailing list for analysis tools

- Central discussion board for ,,all" problems
- Dicussion of several people
- E.g. in ATLAS and LHCb:
  - Before: only developers as experts - very time consuming
  - Now: experiment shift teams with shift credits
  - Very busy mailing list
  - Hope: user-to-user support similar to open-source projekts
- Sites are more stable but still day to day glitches

Infrastructure Tests - Analysis stress tests

ATLAS is/has been testing sites with very high automatic generated analysis load: HammerCloud
http://hammercloud.cern.ch/

Now also available of CMS and soon for LHCb

Differences Analysis vs. MC Production:

- ,,unorganized" user analysis vs. ,,organized" MC production
- User Analysis puts much higher load on SE compared to CPU dominated simulation

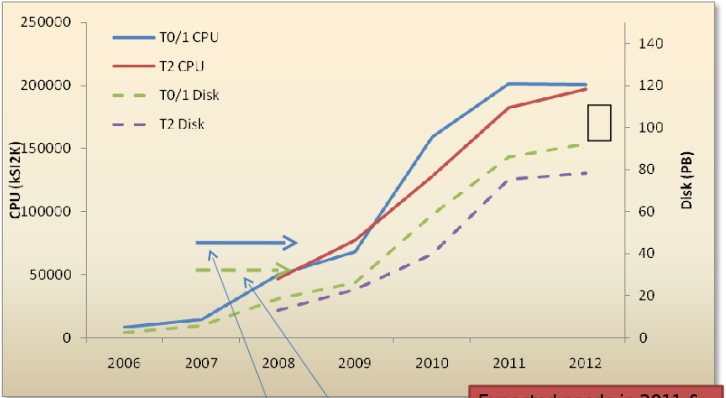
Tests of different work-flows:

- Sequential AOD analysis of MC data
- Sequential cosmics analysis with DB/Frontier/Squid access

Some highlights:

- Analysis tools generally stable and reliable
- Some weak spots detected in site infrastructures, especially in input file access mode lots of tuning potential

Resource Evolution



Expected needs in 2011 & 2012

Need foreseen @ TDR for T0+1 CPU and Disk for 1st nominal year

From: Ian Bird

NB. In 2005 only 10% of 2008 requirement was available. The ramp-up has been enormous!

Prospects and Evolutions

- Infrastructure demonstrated to be able to support LHC data processing and analysis
- Spin off in different areas
- A reliable and robust service of many components neccessary
- Significant operational infrastructure behind it
- Adapt to future technologies:
  - Improve data storage and data access
  - multi-core CPUs
  - Virtualisation
- Network is much better than initially anticipated
  - Rethink data access models
- Experiments have truly distributed models