Grid Computing in the ATLAS-Experiment at the LHC

Johannes Elmsheuser

Ludwig-Maximilians-Universität München

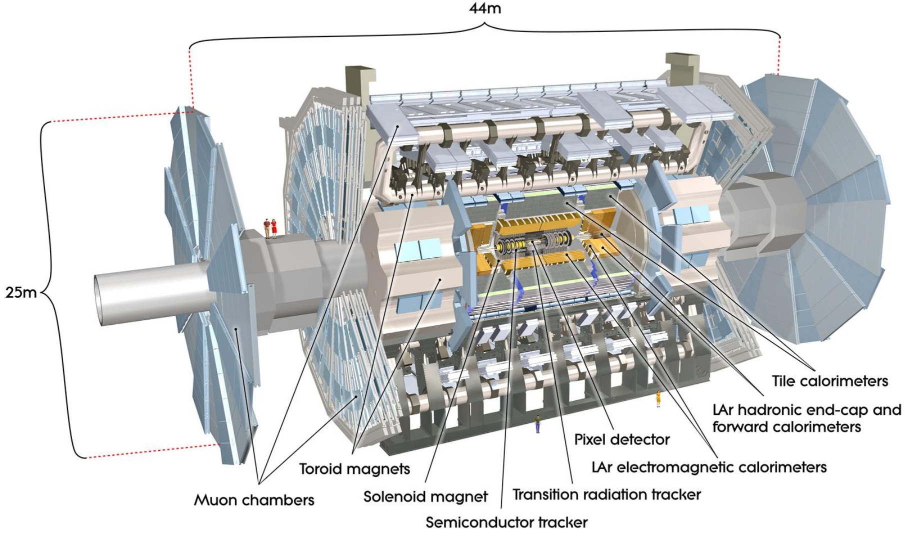
23 April 2009/BFG Workshop, Freiburg

LMU

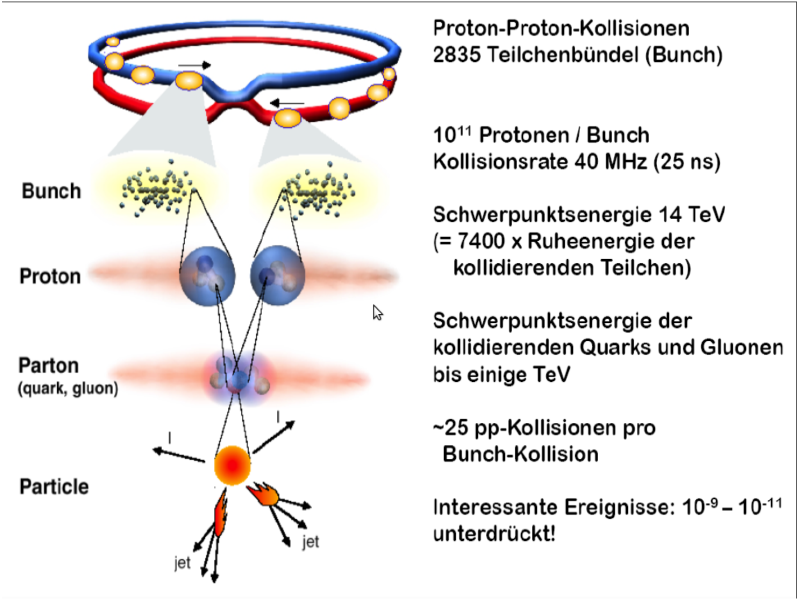
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

The LHC and Experiments

The ATLAS experiment


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Collisions at the LHC



Proton-Proton-Kollisionen
2835 Teilchenbündel (Bunch)

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

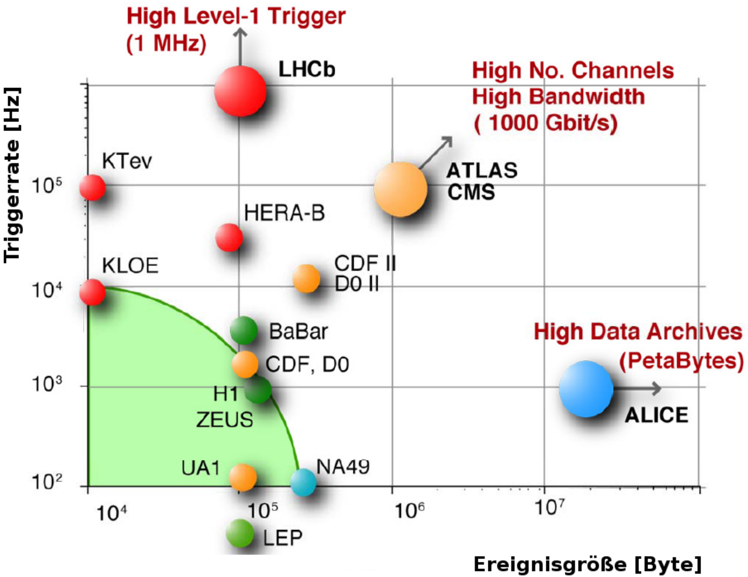
Schwerpunktsenergie 14 TeV
(= 7400 x Ruheenergie der
kollidierenden Teilchen)

Schwerpunktsenergie der
kollidierenden Quarks und Gluonen
bis einige TeV

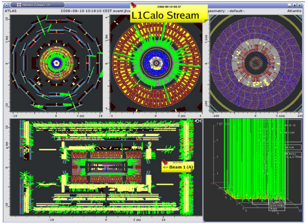
~25 pp-Kollisionen pro
Bunch-Kollision

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

Trigger and Event Sizes

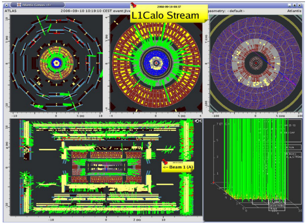


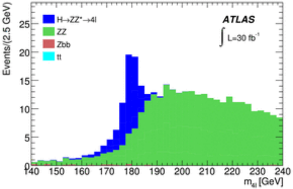
Why Grid Computing ?



Events recorded:
200 Hz (nominal)

Why Grid Computing ?

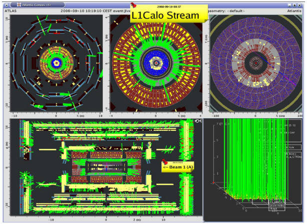




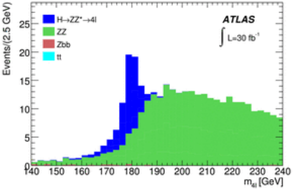
Events recorded:
200 Hz (nominal)

Statistical Analysis of
$O(10^9)$ events

Why Grid Computing ?



⇒ **Grid** ⇒

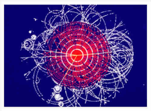


Events recorded:
200 Hz (nominal)

Statistical Analysis of
$O(10^9)$ events

Average Analysis at LHC I

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



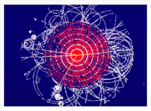
Monte Carlo events needed :

- 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 needed 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 Analysis 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)
  - Repeated 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 one
- Analysis with jets need much more CPU-time
- All given times: without additional overhead

Challenges

Data volumes

- Every experiment stores several Peta Bytes/year

CPUs

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

Software

- Every experiment has own complex software environment

Connectivity

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

ATLAS Grid Infrastructure

- Heterogeneous grid environment based on 3 grid infrastructures:

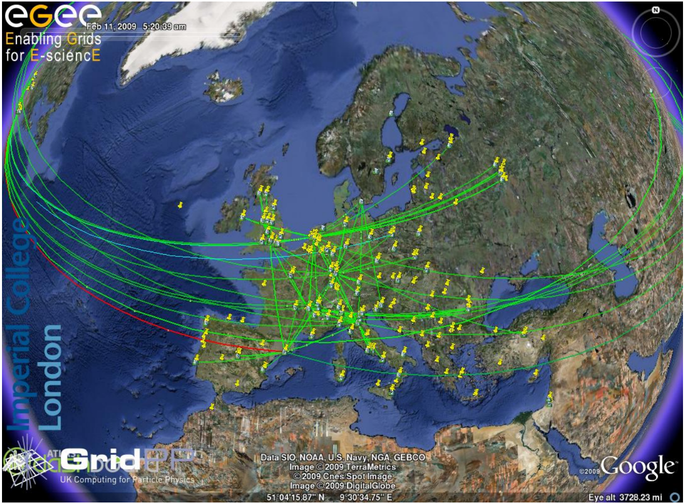
egee
Enabling Grids
for E-sciencE

Open Science Grid

NORDUGRID
Grid Solution for Wide Area
Computing and Data Handling

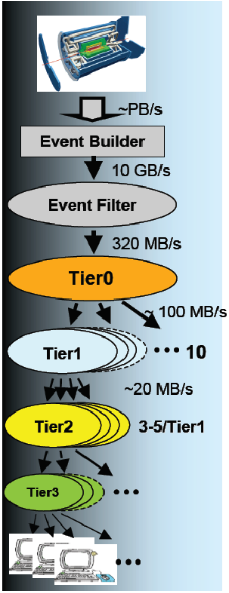
- Three major ATLAS Grid areas:
  - Production System (Panda): centralized MC simulation and Data reconstruction
  - Distributed Data Management (DDM/DQ2): centralized data movement
  - Distributed User Analysis: de-centralized individual analysis

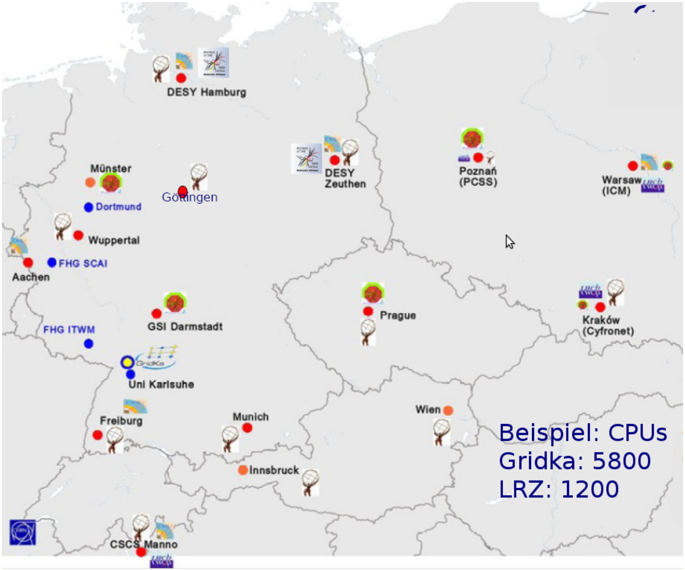
Grid Middle-ware Installations


eGee
Feb 11, 2009 5:20:39 am
Enabling Grids
for E-sciencE
Imperial College
London
ATLAS
GridPP
UK Computing for Particle Physics
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Image © 2009 TerraMetrics
© 2009 Cnes/Spot Image
Image © 2009 DigitalGlobe
51°04'15.87" N 9°30'34.75" E
©2009 Google
Eye alt 3728.23 mi
N

EGEE Real Time Monitor plug-in for Google Earth using ATLAS data

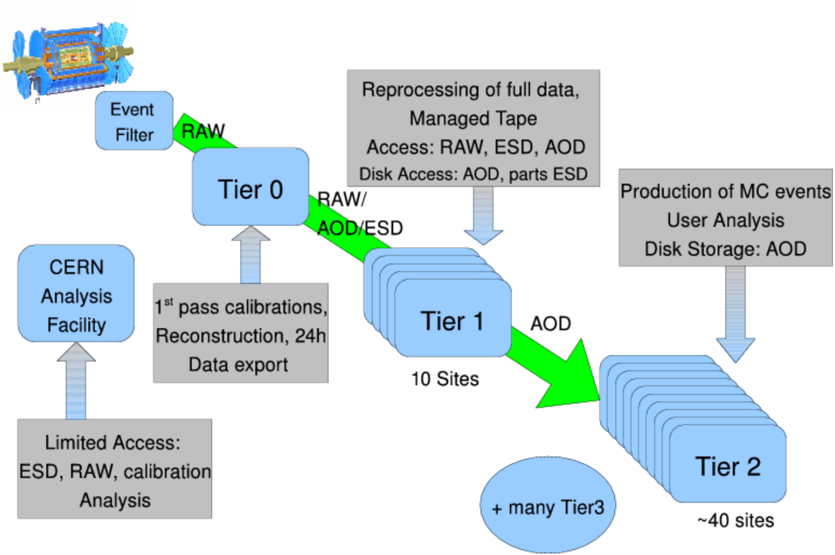
Computing center- and GridKa-Cloud





Beispiel: CPUs
Gridka: 5800
LRZ: 1200

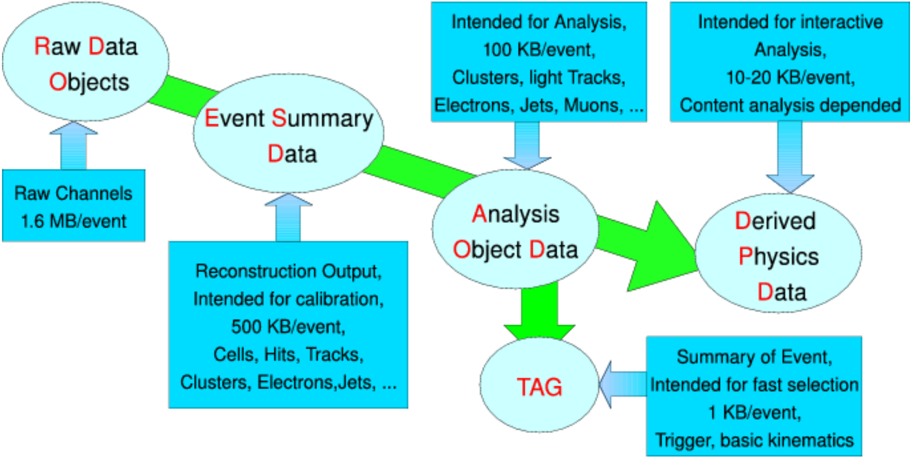
Data Distribution: ATLAS



Event Filter
RAW
Tier 0
RAW/ AOD/ESD
Reprocessing of full data, Managed Tape Access: RAW, ESD, AOD Disk Access: AOD, parts ESD
Production of MC events User Analysis Disk Storage: AOD
CERN Analysis Facility
1st pass calibrations, Reconstruction, 24h Data export
Tier 1
10 Sites
AOD
Limited Access: ESD, RAW, calibration Analysis
Tier 2
+ many Tier3
~40 sites

Event data model: ATLAS

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


Raw Data Objects
Raw Channels 1.6 MB/event
Event Summary Data
Reconstruction Output, Intended for calibration, 500 KB/event, Cells, Hits, Tracks, Clusters, Electrons,Jets, ...
Intended for Analysis, 100 KB/event, Clusters, light Tracks, Electrons, Jets, Muons, ...
Analysis Object Data
Intended for interactive Analysis, 10-20 KB/event, Content analysis depended
Derived Physics Data
TAG
Summary of Event, Intended for fast selection 1 KB/event, Trigger, basic kinematics

Grid usage in the LHC experiments

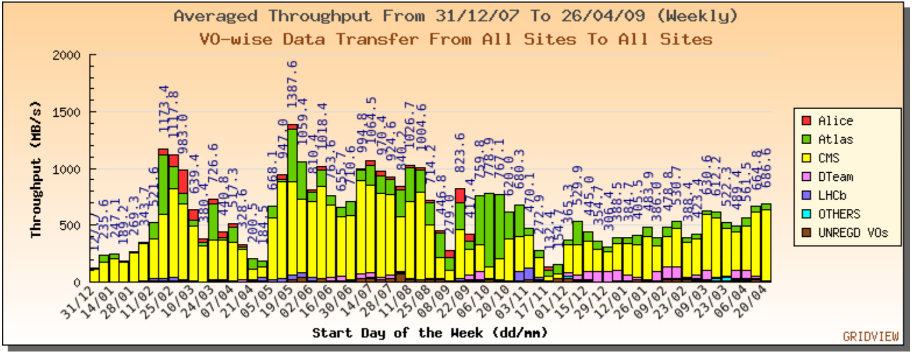
Grid has mainly been used so far for:

- Centrally organized data distribution
- Centrally organized Monte Carlo production done by a few experts
- After long learning curve: production and data distribution works at a good efficiency
- Experiments have implemented their customized DDM and WMS
- Amount of individual user not at full steam yet

Question:

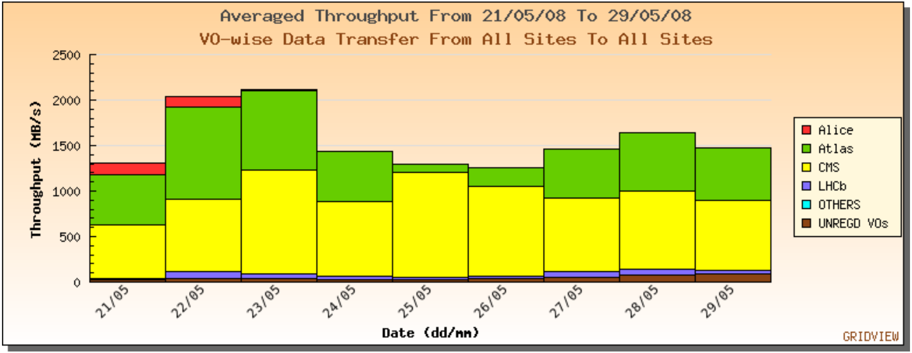
- Is the Grid infrastructure ready for user analysis of many individual users following the model:
  ,,Job to Data"

Data management and transfers



- Data Transfer rates in the last 15 Month

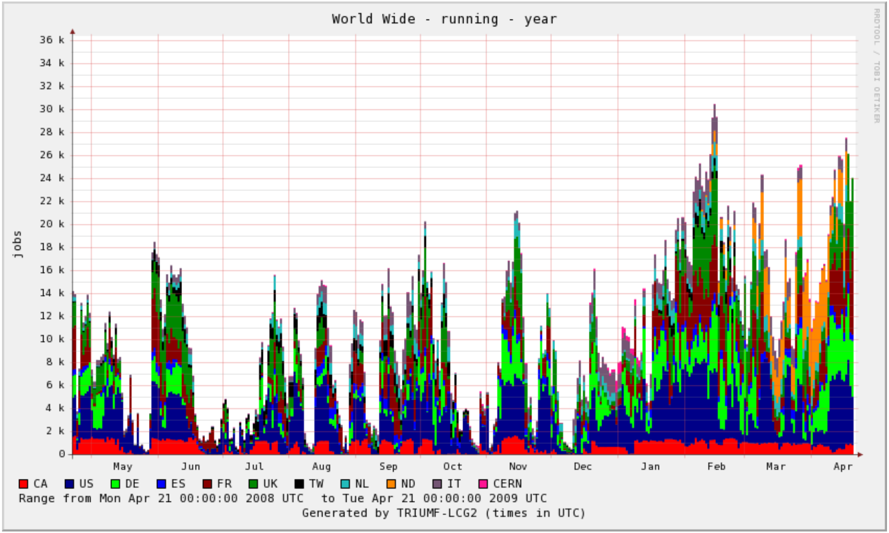
Data management and transfers



- Daily peaks at 2 GB/s (May'08 CCRC)

Number of ATLAS production Jobs

Structure caused by ATLAS software changes



Grid Job Submission

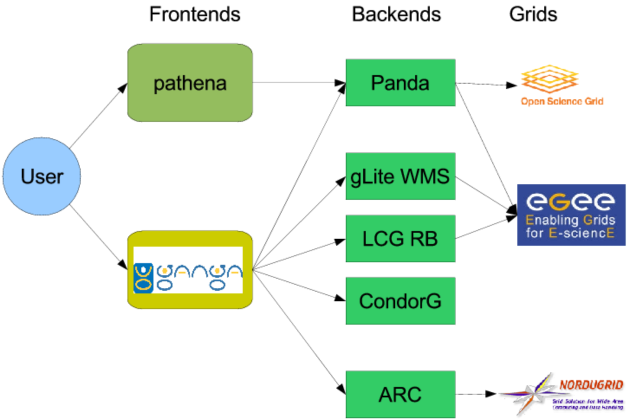
Naive assumption: Grid ≈ large batch system

- Provide complicated job configuration for WMS
- 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.

⟹ Need for automation and integration of various different components

Many ways lead into the Grid !

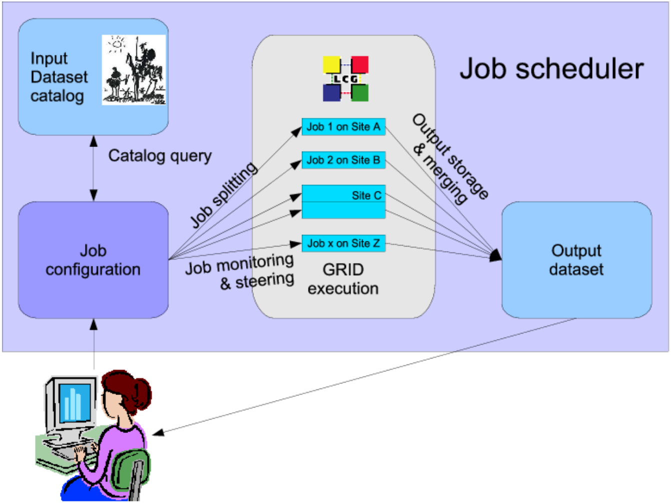
ATLAS Distributed Analysis



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

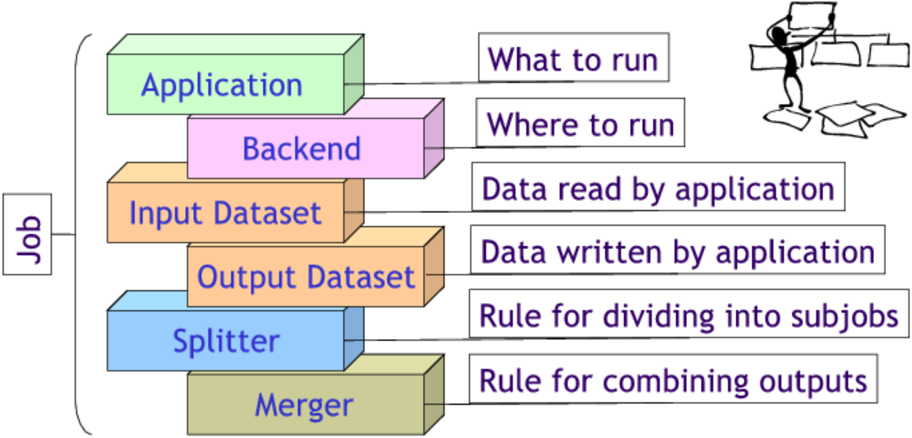
Distributed Analysis: Ganga

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



GANGA Job Abstraction

- GANGA simplifies running of ATLAS (and LHCb) applications on a variety of Grid and non-Grid back-ends



Job definition using ATLAS software

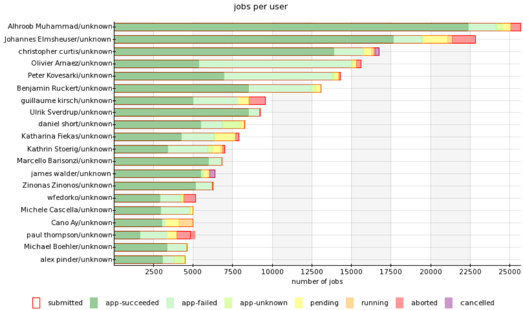
GANGA offers three ways of user interaction:

- Shell command line
- Interactive IPython shell
- Graphical User Interface

Job definition at command line for GRID submission:

```
ganga athena
  --inDS fdr08_run2.0052283.physics_Muon.merge.AOD.o3_f8_m10
  --outputdata AnalysisSkeleton.aan.root
  --split 3
  --lcg --cloud DE
  AnalysisSkeleton_topOptions.py
```

Number of ATLAS analysis Jobs



- GANGA ATLAS Jobs in EGEE Grid
- Since February >630k Jobs
- similar number on Panda

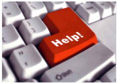
- Compare with daily ∼ 100k productions jobs
- Since beginning of the year increasing number of users - but many more expected !

Current user problems and Support

Frequently asked questions or problems:

- Where is my data ?
- There is a problem with my special code configuration
- The job had problems with accessing the input data files
- The ratio of CPU and Wall-time is largely varying btw. 10% - 100% and depends on the site and user

Support:



- Started ATLAS wide user support mailing list for DA
- Shifters in EU and US time zone
- Hoping for user2user support
- Has developed to one of the busiest mailing lists in ATLAS

Infrastructure Tests - Analysis stress tests

ATLAS is testing since several month all sites with very high automatic generated analysis load

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

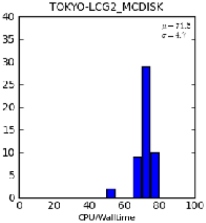
Tests of different work-flows:

- Sequential AOD analysis of MC data
- Sequential cosmics analysis with DB access at Tier1

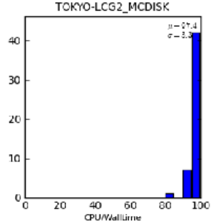
Some highlights:

- Analysis tools generally stable and reliable
- Some weak spots detected in site infrastructures

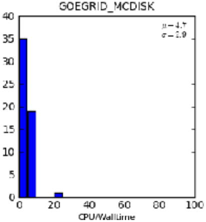
HammerCloud - Plots



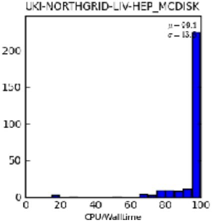
Tokyo, rfio input



Tokyo, FileStager input



data on 1 pool



Liverpool, old CPUs

Event Rate is important number

Conclusions and Summary

What is working well so far:

- MC Production
- Automatic data distribution
- Analysis:
  - At a chosen number of sites
  - Small scale MC production
  - Automatic Standard Job Configurations

What works, but needs improvement:

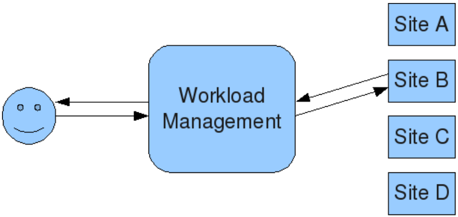
- Analysis:
  - 'Blind' job submission
  - Exotic use cases
- Site availability and Input file access

For the distributed analysis it is vital to have:

- Easy interface that does not scare off physicists
- A reliable and robust service of many components

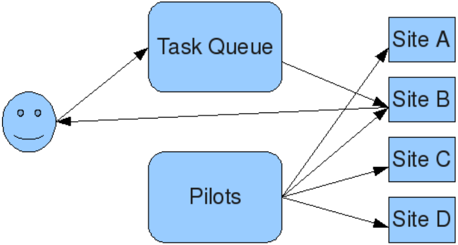
BACKUP

Job Scheduling



Job Push mode

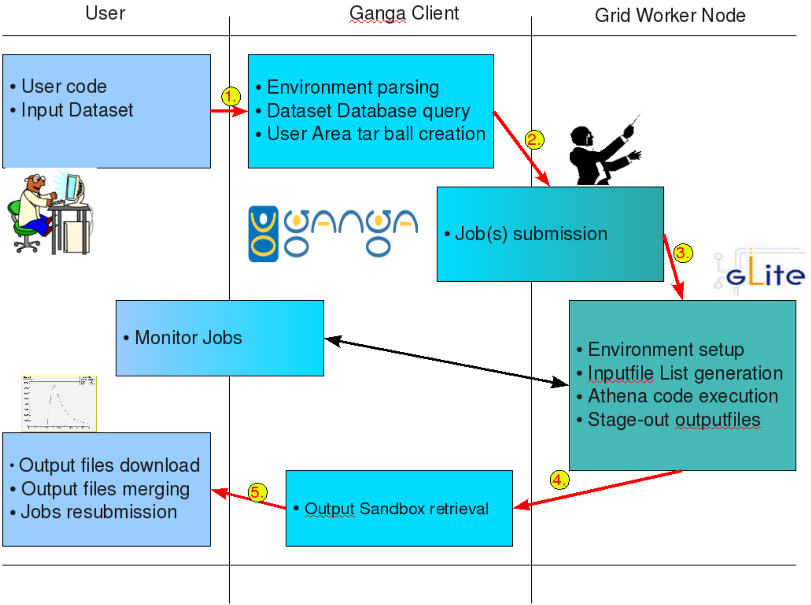
- Dependent on information system and site status
- Decentralized
- Better control of site policies



Job Pull mode

- Workarounds for many Grid problems
- Avoids ,,black holes ”, some down-times and bad site configurations
- Data pre-staging

Job work-flow: Athena on LCG back-end



User

Ganga Client

Grid Worker Node

- User code
- Input Dataset

- Environment parsing
- Dataset Database query
- User Area tar ball creation

- Job(s) submission

- Environment setup
- Inputfile List generation
- Athena code execution
- Stage-out outputfiles

- Monitor Jobs

- Output Sandbox retrieval

- Output files download
- Output files merging
- Jobs resubmission