LHC ATLAS users analysing data on the Grid

M. Lamanna, D. Van Der Ster (CERN IT – ES group) and J. Elmsheuser (LMU Munich)

for the ATLAS Distr. Computing project

ATLAS EXPERIMENT http://atlas.ch

www.eu-egee.org

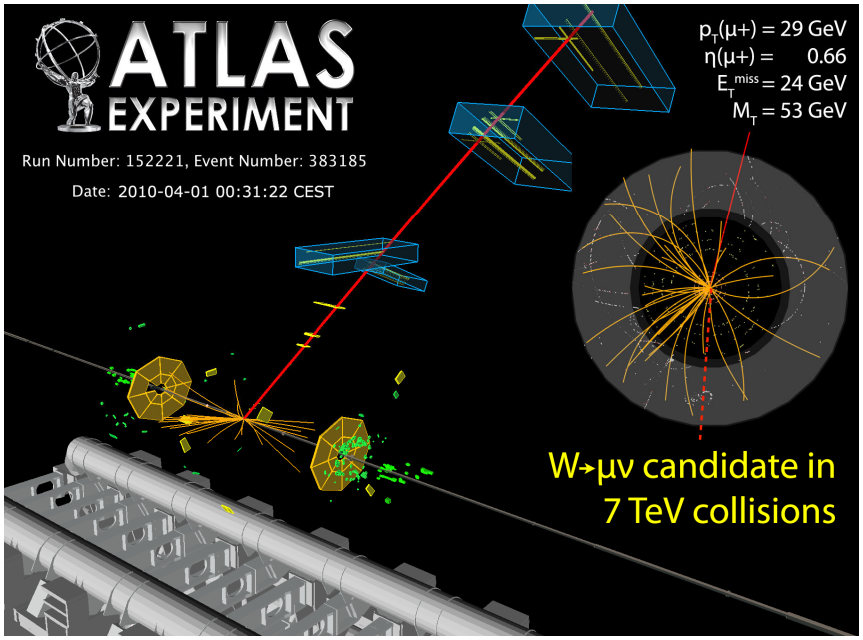
e-infrastructure

Objective of the talk

- **Share ATLAS experience in dealing with large data samples (data handling and data access) with other grid users**
  - This is one of the goal of the User Forum
- **Describe the commissioning procedure to match the requirements for analysis (I/O intensive activities)**
  - Training
  - Site commissioning
  - User support
- **Discuss the current status of the evolution of the system**

ATLAS

Examples of pre-analysis activities



- **LHC pp-collisions in the ATLAS detectors**
  - Now on at 3.5 TeV + 3.5 TeV
- **Data distribution (DDM)**
  - ATLAS distr. Data management
- **Data reconstruction (PanDA)**
  - ATLAS production system
- **All must run on EGEE, OSG and NDGF infrastructures (WLCG)**

Analysis activities

- **At this point, user analysis kicks in**
  - For small data sets, download and study them "at home" (more later in the talk)
  - Make use of the Grid infrastructure (PanDA/WMS+DDM on top of EGEE/OSG/NDGF)
- **Analysis is an "individual" activity**
  - Many users
- **Analysis is about understanding complex data**
  - Iterative
  - Interactive
  - Task-completion latency more important that global throughput
  - I/O performances absolutely critical
  - Well-established practices (on batch and personal workstations)
- **Grid is essential (not an optional nice-to-have tool) but can be complemented by the use of other resources (for example to development and tune up algorithms, debugging…)**

ATLAS users accessing the data

Data type	RAW	ESD	AOD	dESD
LHC data 2009	102	455	527	725
Cosmics 2009	33	87	30	21
MC09	–	523	590	–

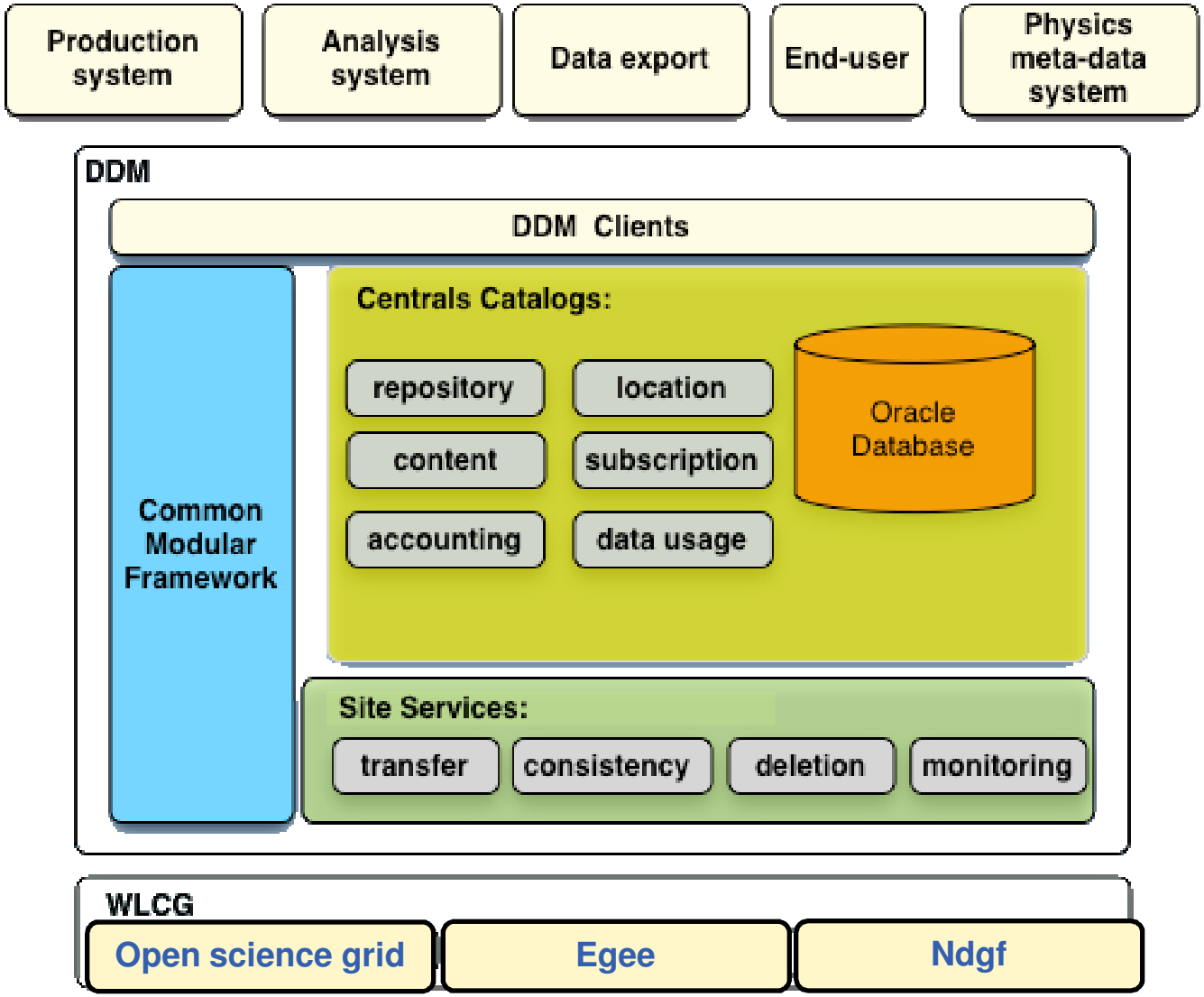
Number of distinct users that accessed ATLAS data on the Grid (by data type) in the period between 20 November 2009 and 23 February 2010. This includes job access and data "download".

Simulated HITS (the equivalent of RAW for real data) are kept on tape and not made generally available for analysis. No dESDs were produced for simulated events.

ATLAS Distributed Data Management

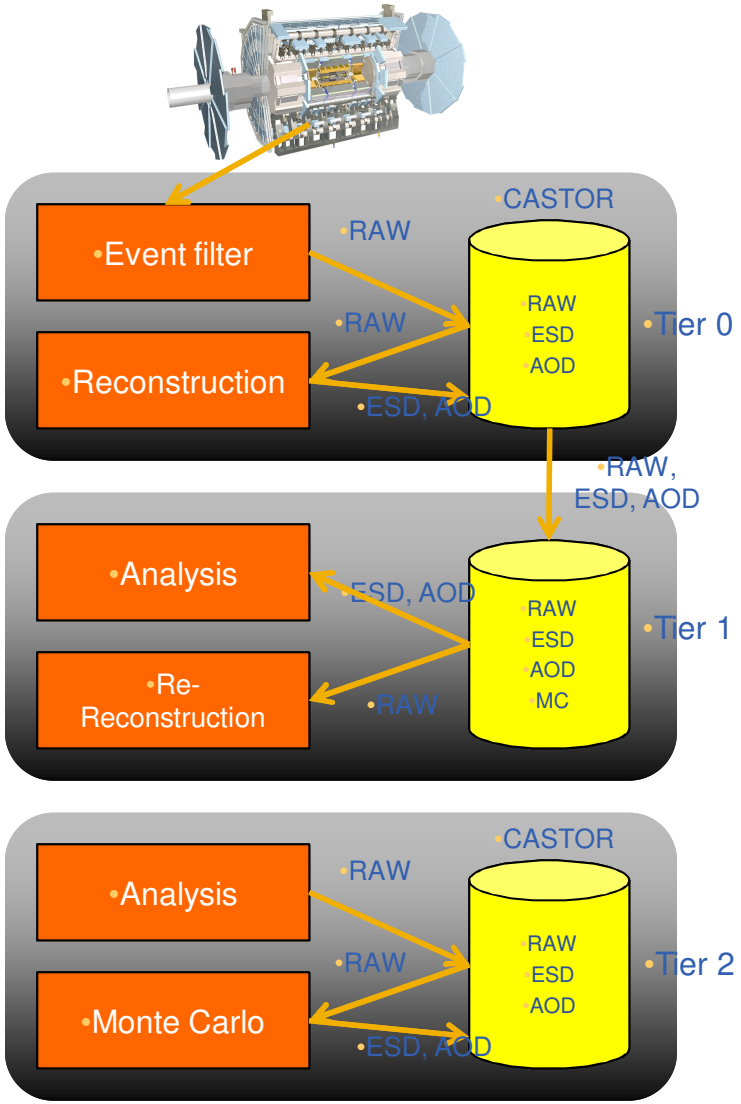
- **LHC experiments collects data in the the 100-1000 MB/s range**
- **Integrated values (RAW data) are in the PB/year range**
- **Additional data have to handled**
  - Simulation
  - Derivated types (RAW reconstruction and analysis stages)
  - Data are replicated
    - Access efficiency, data safety
- **The data management system is a key component of the ATLAS distributed computing**
  - Example of a complex high-level service
    - It uses LFC, FTS and SRM
    - Complex software product (development and operations)

DDM software stack

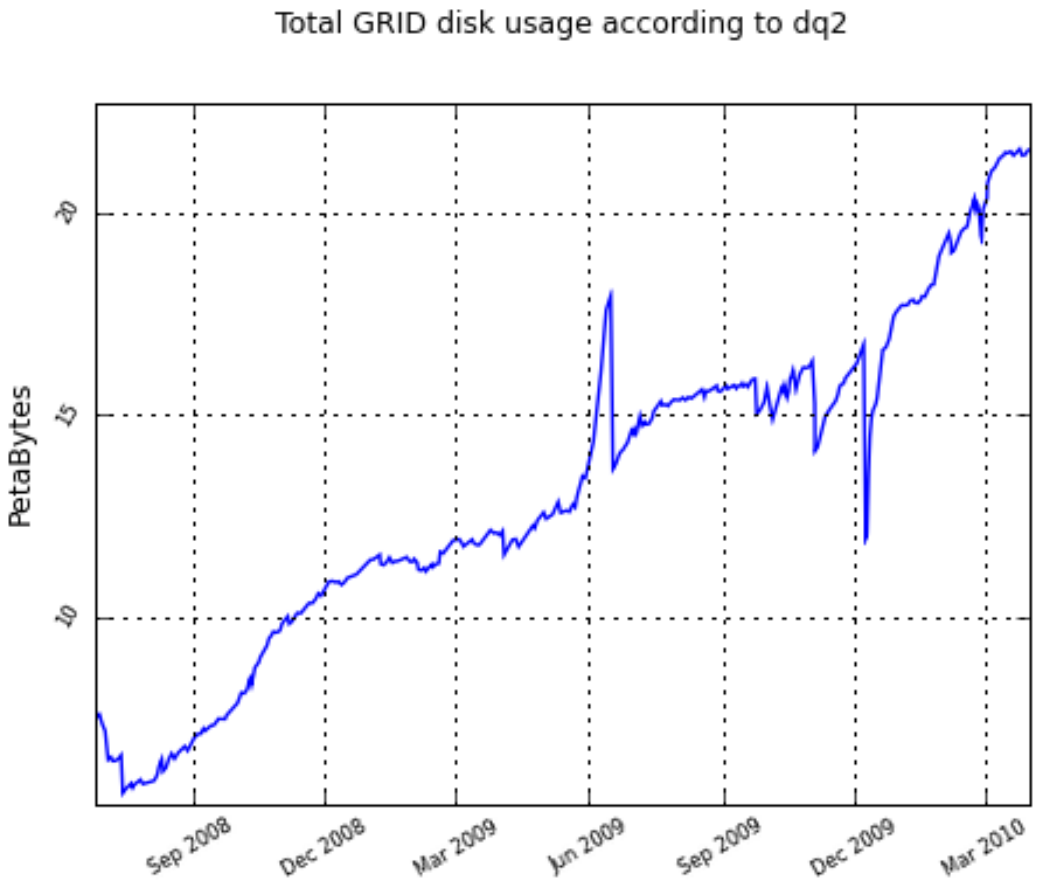


ATLAS Computing Model: Hierarchical Grid Organization

- **Tier-0 facility (CERN):**
  - Custodial and distribution of primary RAW detector data
  - First pass reprocessing of the primary event stream
  - Distribute the derived datasets to the Tier-1s
- **10 Tier-1 data centers:**
  - Custodial and long-term access to a subset of the RAW data
  - Access to derived datasets
- **~100 Tier-2 institutes:**
  - Analysis capacity for users & physics-groups
  - Monte-Carlo simulation

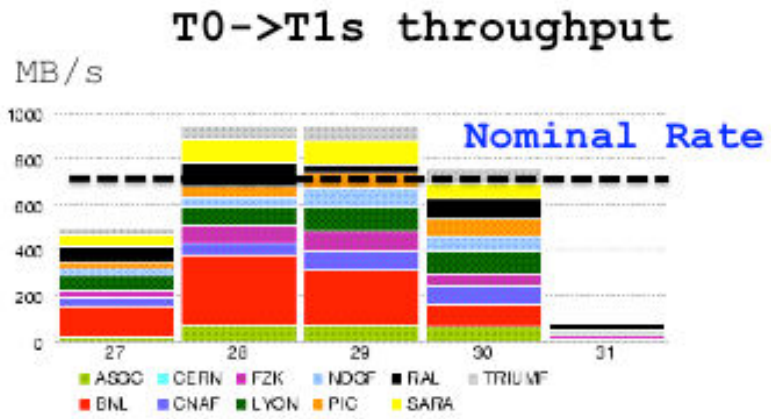


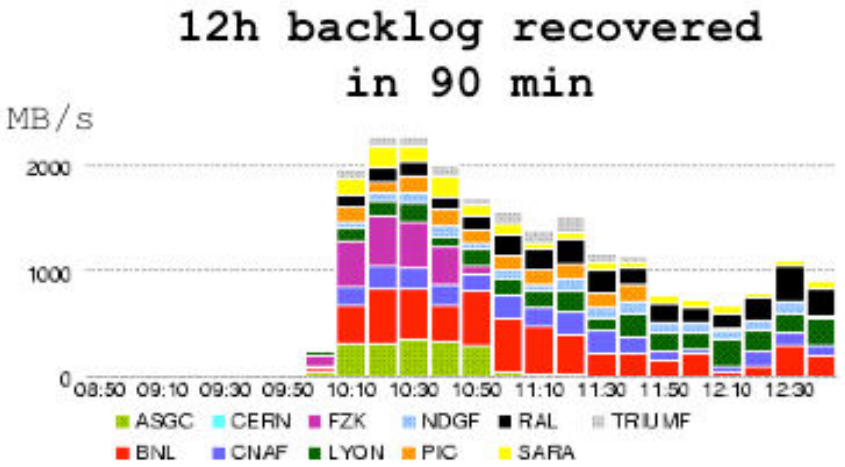
Baseline services for ATLAS data

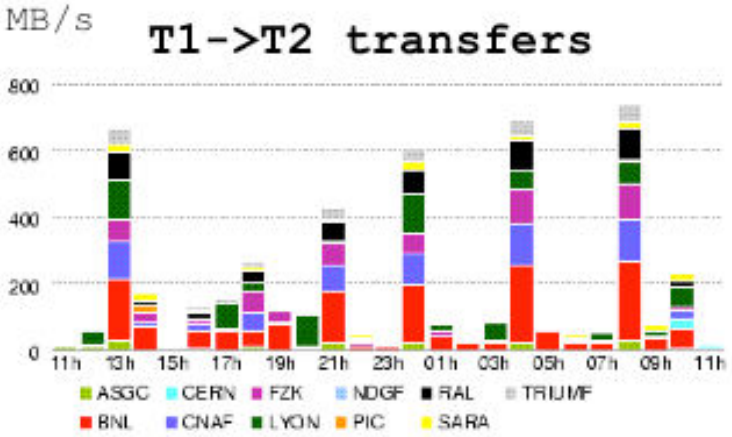


- DDM: organization, access, placement and deletion
- 1000+ users
- 500+ endpoints
- 3 grids
- ~2,000,000 dataset replicas
- ~80,000,000 file replicas

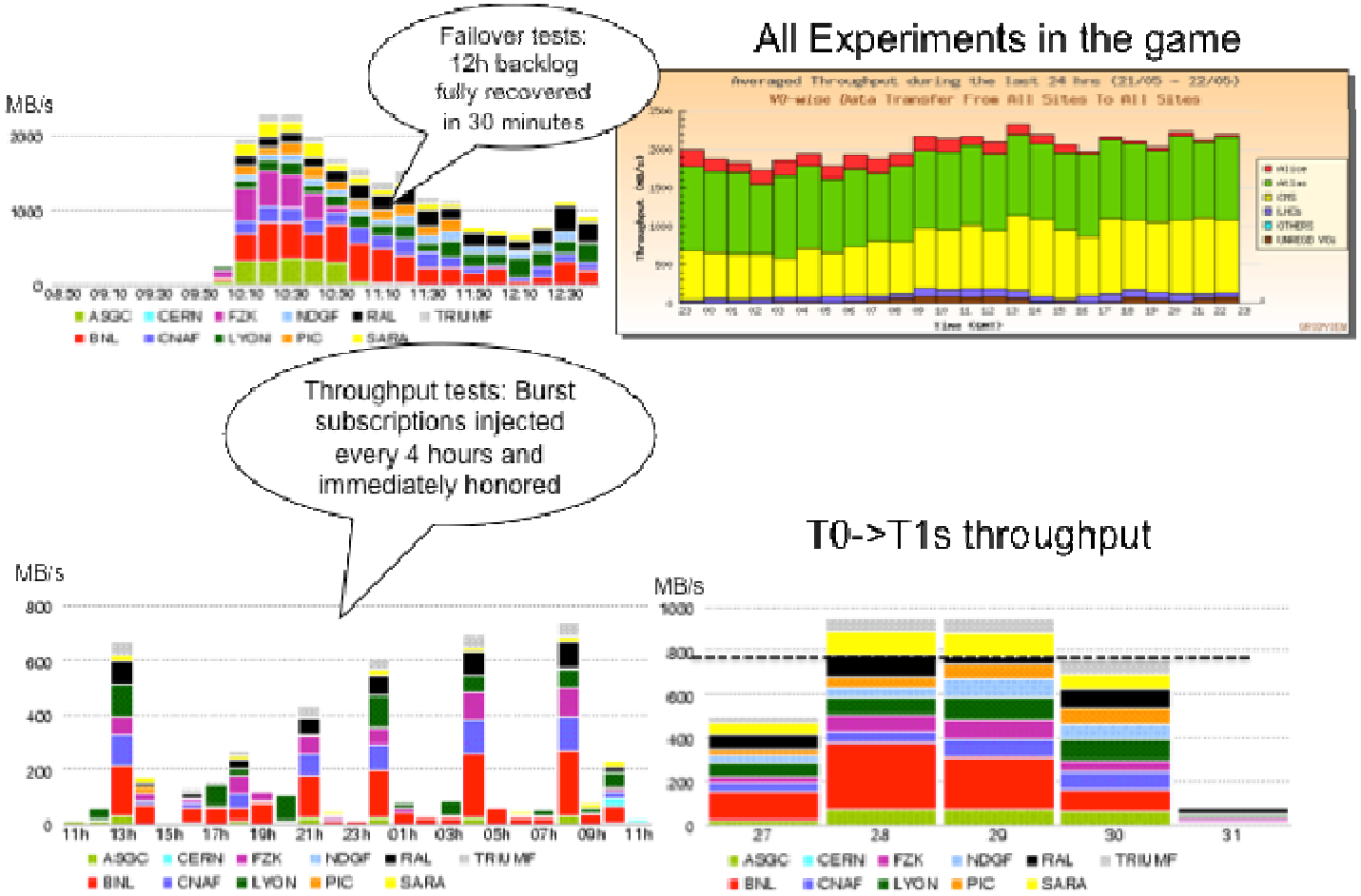
Performances 1







Performances 2



Users

- **What ATLAS users should do**
  - They use the ATLAS framework on their desktop (Athena)
  - They use Ganga or pathena to submit their Athena job
    - Ganga has been developed also in EGEE.
    - Pathena is a CLI built on the PanDA API (used also by Ganga)
    - Ganga allows native access to EGEE and NDGF infrastructures (+batch resources)
  - They run their jobs and are happy ☺
- **What does it take it to have users happy?**
  A. Teaching the tools
  B. Commissioning the sites where jobs run
  C. Support users in case of problems

Important:

500+ users (steadily growing)

100+ sites (rather stable)

Teaching the tools

- **Material generated by the tools developers and power users**
  - Documentation
  - Training material (part of the Ganga distribution)
- **A lot of "peer-to-peer" help**
  - Learning from your colleague in an informal way
- **Regular programme of tutorial**
  - By now ran by members of the experiments.
  - Regularly ran every 6 weeks for about 2 years
    - **This is part of general ATLAS training (including using Athena)**

Commissioning the sites (1)

- **In the past major efforts to improve "grid reliability"**
  - Out of all jobs, compute (1 – #failure/#total)
  - Initially essentially reconstruction jobs
- **Evidence that good sites for simulation activities might show poor behaviour for analysis**
  - Low reliability (many failures)
    - Data access? Software distribution? User code problems? …
  - Low efficiency (wasting resources)
    - Not enough bandwidth to the storage
      - *These jobs are often IO-limited*
      - *For a set of job we measure the mean value of the distribution CPU_time/Elapsed_time*
      - *Often long elapsed time indicate possible problems in reading data*
  - Of course they can be fixed!

Commissioning the sites (2)

1. **Reproducible benchmark (on top of Ganga)**
   - Define a set of datasets (commonly used ones by users)
   - Define a set of user jobs (*frozen* applications from representative users). We call this the *job cocktail*.
   - Define a procedure to load centres (submit jobs)
2. **Submit jobs, retrieve results and compute reliability and efficiency**
   - Tests can be defined both centrally and by individual sites
     - You can test your site during troubleshooting or after an upgrade
     - You can test software (e.g. ATLAS Munich developed improved dCache libraries and tested with hammerCloud)
3. **Publish the results**
   - Proved to be very important!
4. ***And keep an eye on it…***
   - *Small scale, regular, centralised-only jobs*
   - *Results are published in SAM*

1,2,3: in use for about one year:

HammerCloud

Running for a few years now:

GangaRobot

HammerCloud

Hammercloud

Home | Clouds | Tests | Last Tests | Time | HC Stats | Administration

all the tests

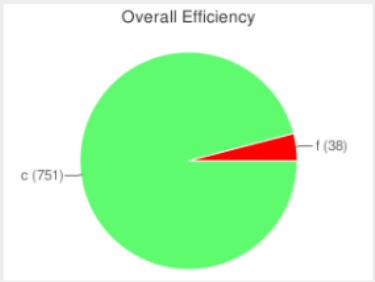
state	id	host	clouds	start time (CET)	end time (CET)	sites	submitted jobs
scheduled	1284	voatlas49.cern.ch	ES, NL, IT, 1				
running	1283	voatlas49.cern.ch	ES_PANDA, NL				
completed	1282	voatlas49.cern.ch	US				
completed	1281	voatlas73.cern.ch	UK				
completed	1280	voatlas73.cern.ch	UK				
completed	1279	voatlas73.cern.ch	UK				
completed	1278	voatlas49.cern.ch	ES, NL, IT, 1				
completed	1277	voatlas49.cern.ch	ES				
completed	1276	voatlas49.cern.ch	ES				
completed	1275	voatlas49.cern.ch	DE				
completed	1273	voatlas49.cern.ch	ES_PANI				
completed	1271	voatlas49.cern.ch	DE				
completed	1270	voatlas49.cern.ch	US, ES_PANDA, NL_P				
completed	1269	voatlas73.cern.ch	UK				
completed	1268	voatlas49.cern.ch	ES_PANI				
draft	1264	voatlas73.cern.ch	UK_PANI				
completed	1263	voatlas73.cern.ch	UK				

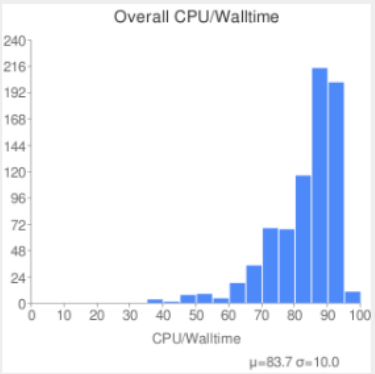
Summary

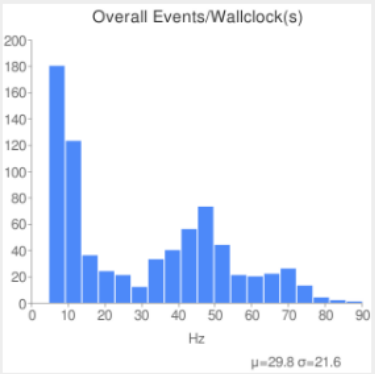
state	id	host	clouds	start time (CET)	end time (CET)	submitted jobs
completed	1236	voatlas73.cern.ch	DE	2010-03-26 15:30:00	2010-03-26 21:31:33	2828

Input type: DQ2_LOCAL
Output DS: user09.JohannesElmsheuser.ganga.sitetest.DEWMS.20100326.1.[sitename]
Input DS Patterns: mc*merge.AOD*_r11*_r*
Ganga Job Template: /data/gangarobot/hammercloud/inputfiles/D3PD1566/D3PD_LCG_v1566.tpl
Athena User Area: /data/gangarobot/hammercloud/inputfiles/D3PD1566/D3PDMaker_v1566.tar.gz
Athena Option file: /data/gangarobot/hammercloud/inputfiles/D3PD1566/AODToPhysicsD3PD_v1566.py

View Test Directory (for debugging)



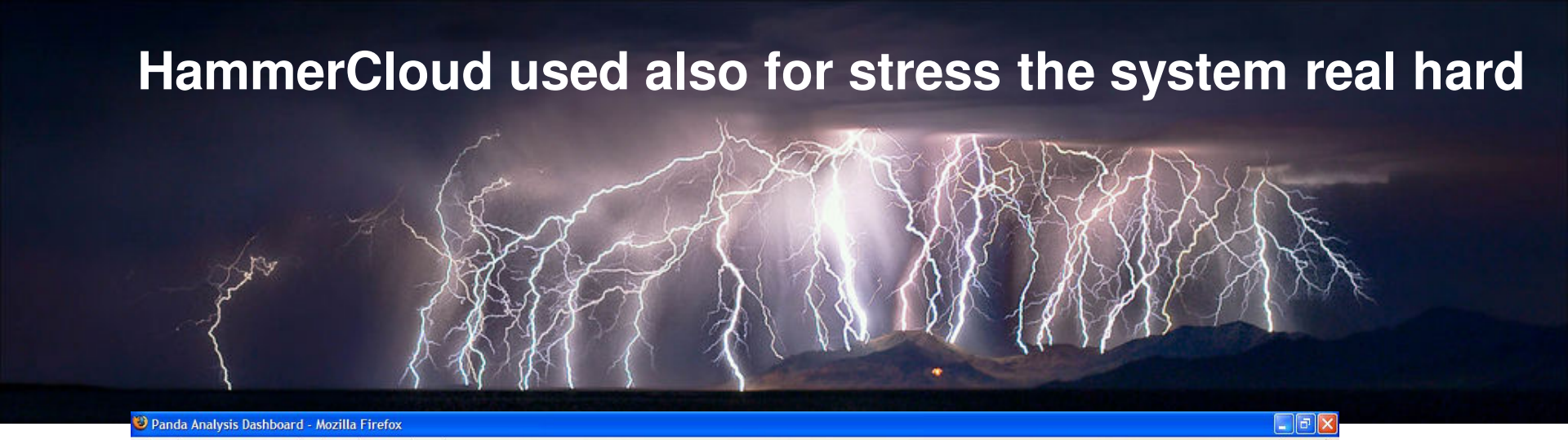




Overall

Collaboration with CMS to extend it to CRAB jobs

HammerCloud used also for stress the system real hard



Panda Analysis Dashboard - Mozilla Firefox

http://panda.cern.ch:25980/server/pandamon/query?dash=analysis

Summaries
Blocks: days
Errors: days
Nodes: days
Usage 1, 3 days

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
New Tag
Bug Report
Task overview query

Datasets - search
Popular datasets
Aborted datasets
Datasets Browser

Datasets Distribution
DaTRI
DaTRI: Req list
AODs
EVNTs
Conditions DS
DB Releases
SIT pacballs
Validation Samples
Functional Tests
ATLAS Data
Reprocessed_Datasets

Sites

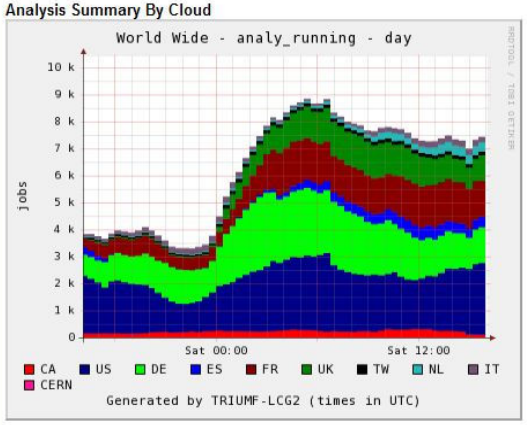
Logging monitor

run a Panda job); if you 'log in' you'll get easier access to your page from a new menu at the top of the page.

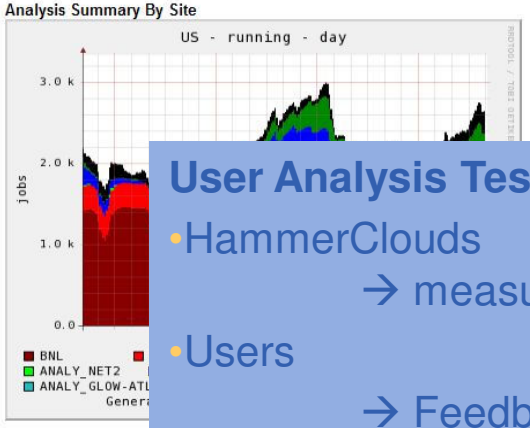
Groups: Groups are supported to organize users by role, physics working groups etc. and support collaborative work, accounting rights etc. (Not much used yet.)

Data access: See the physics data page linked above for infomation on data location, requesting replication of data, and staging data from tape to disk.

Analysis Summary By Cloud



Analysis Summary By Site



Analysis job summary, last 24 hours (Details: errors, nodes) pathena analysis queue status

Cloud Information	Job Nodes	Jobs	Latest	Pilot Nodes	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed tot trf other
Overall Analysis	7792	21367	10-31 16:02	7144	23886 / 0	4 / 0	0	77339 / 0	1 / 0	7712 / 0	2657 / 0	14 / 0	71771 / 0	21367 / 0 23% 3% 20%

User Analysis Test (October 2009)

- HammerClouds
  → measurements
- Users
  → Feedback (formulaire)

GangaRobot

- **Based on Ganga**
  - Inspired from the CMS JobRobot
- **Testing site since early 2008 ☺**
  - Small volume of jobs
  - Run under user credentials
  - Cocktail of applications
- **It is also an active user protection:**
  - “bad sites” are automagically blacklisted
    - This avoid unnecessary suffering on the user side
- **Increasing the integration with HammerCloud**

Support users in case of problems

- **This is not enough…**
  - Users need support
  - Expect high pressure from users with first LHC data
    - Less expert users
    - More pressure on the system, hence more hickups
- **We decided for a shift system**
  - Good examples in ATLAS for example in running production activities
    - X. Espinal (PIC) and K. De (UTA)
  - Experts taking shifts from their home institute
    - Benefit from a worldwide collaboration!
- **What is special here:**
  - More diverse community to be supported
  - "**Protect the developers**" mantra

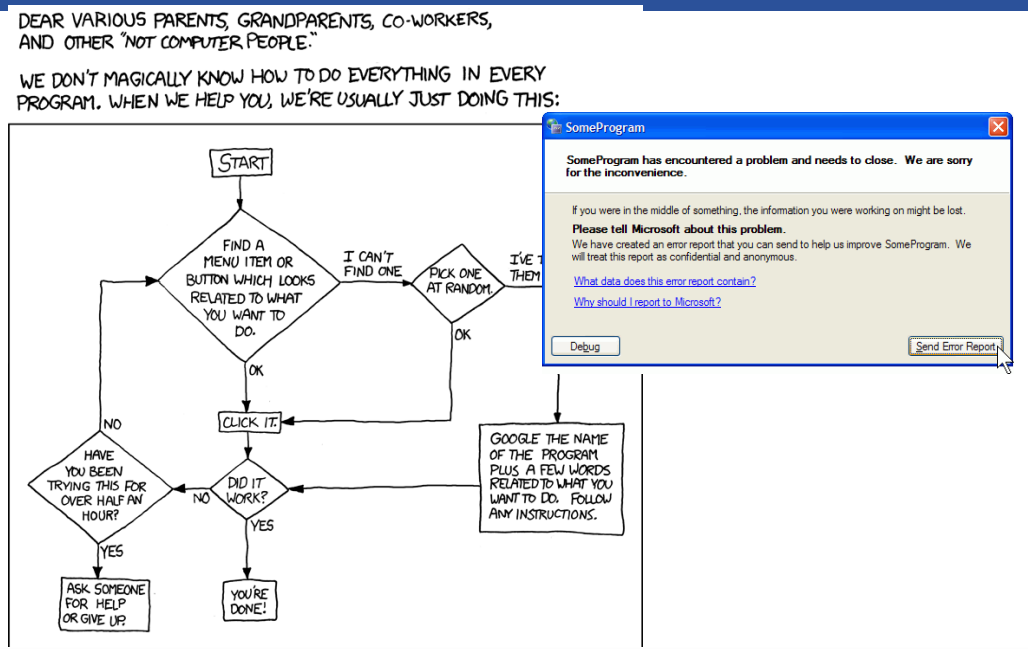
Distributed Analysis Shift Team (DAST)

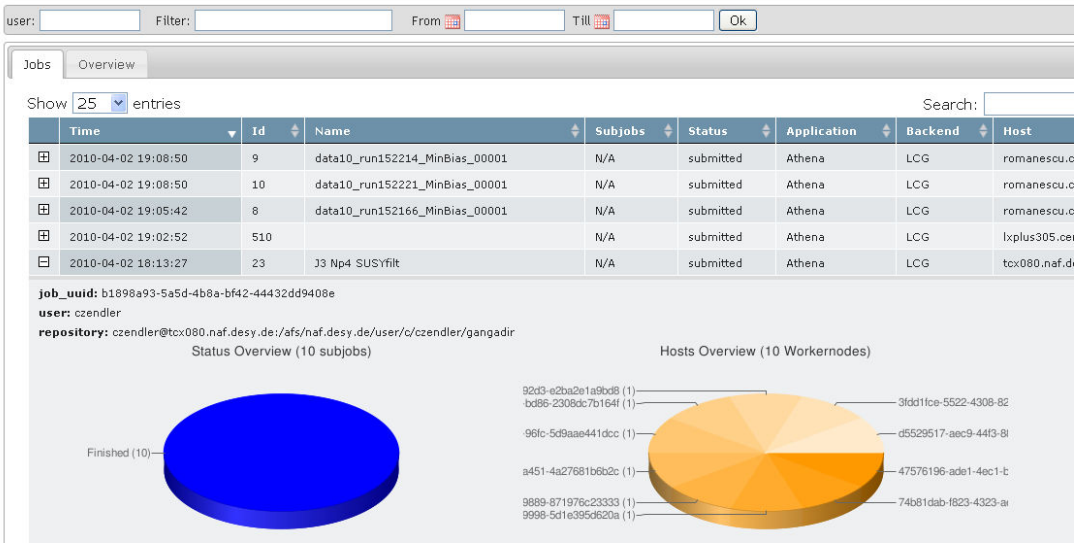
- **Providing support to Distributed Analysis users since September 2008**
  - https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasDAST

- **First contact point with DA users via a mail forum:**
  - hn-atlas-dist-analysis-help@cern.ch
  - Less formal that a ticketing system
    - Encouraging user2user support
      - *Which is important and we would like to stimulte it even more*
  - Back-office
    - Simple (yet sophisticated) procedures: shifters are not sitting in a “control room”. They need to privately communicate and annotate issues
      - *“I am working on this issue”, “I have already asked the 2nd line expert”*
    - Prototype using gmail and gcalendar

More ATLAS users are joining DAST

- **DAST team**
  - Train the shifters
    - Formal tutorial (sometimes videotaped)
    - Junior shifter paired with senior ones
- **Empower the shifters and the users**
  - Bugs are then fed in the proper tracking system
    - *To developers*
  - Monitoring!
  - Allow users to create snapshots of their environment for the support group



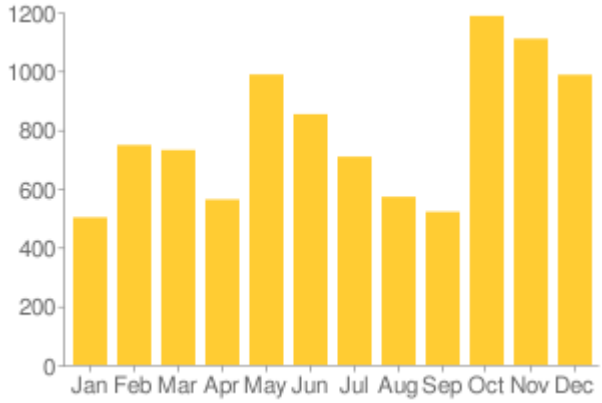


Experience with DAST

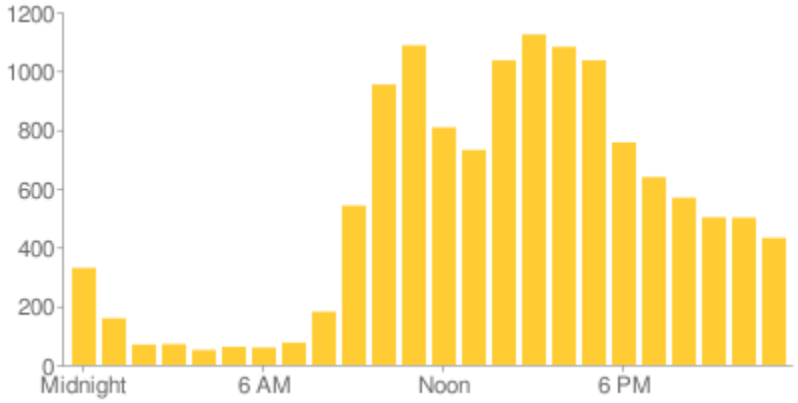
- **Catch all of "all" user problems:**
  - Conditions database access, Athena, physics analysis tools, site problems, dq2-* tools, data access at sites, data replication

- **Shifters: 8-hour day shifts to cover around 15h per day (5 days per week)**
  - 8 (+4 junior) in the EU time zones
  - 5 (+3 junior) in the NA time zone

Stable on ~ 800 messages a month



Time structure (as UTC+1)



What's next

- **It is difficult to make predictions, especially about the future**
  - ☺ http://www.larry.denenberg.com/predictions.html
- **Trends from ATLAS analysis**
  - Will these trends be confirmed?
  - Are they relevant for ATLAS only?
    - HEP only?
    - Scientific computing only?
- **Where do we stand now?**
  - Powerful federation of data-handling infrastructures
    - Largely based on Grid infrastructures, notably EGEE
- **Near future**
  - ATLAS (and HEP): complement the infrastructure with Tier3 facilities
    - "Analysis facilities"
      - *Small ones ("Tier3 proper")*
      - *Large ones: National Analysis Facilies (NAF@DESY, LAF@IN2P3CC, …)*

ATLAS Tier3 key points

- **Operations**
  - Must be simple (for the **local** team)
  - Must not affect the rest of the system (hence **central** operations)
- **Data management**
  - Again simplicity
  - Different access pattern (analysis)
    - *I/O bound, iterative/interactive*
    - *More ROOT-based analysis (PROOF?)*
    - *Truly local usage*
  - "Performances"
    - *Reliability (successful jobs / total)*
    - *Efficiency (CPU/elapsed) → events read per second*
- **Client-based grid sites**
  - Data are downloaded, not replicated
  - Data are accessed via native protocols
  - No catalogues (the storage *is* the catalogue, e.g. file systems)
  - Other data (conditions, in some cases even the software) are cached on demand

ATLAS Tier3 working groups

- **DDM-Tier3 link**
  - How to “download” data...
    - S. Campana (CERN)
- **Software / Conditions data**
  - Data distribution access and caching of auxiliary data
    - A. de Salvo (Roma) and A. da Silva (Victoria)
- **Data access (Lustre/Xrootd)**
  - Main data access via file system or file-system like
    - S. Gonzalez La Hoz (Valencia) and R. Gardner (Chicago and OSG)
      - ***Also creating an inventory and a knowledge base!***
- **Tier 3 Support**
  - Tools/infrastructure: HammerCloud, DAST, docs...
    - D. Van der Ster (CERN)
- **PROOF Working Group**
  - Parallel ntuple scan
    - Neng Xu (Wisconsin) and W. Ehrenfeld (DESY)
- **Virtualization**
    - Yushu Wu (LBNL)

Interim reports: due in a week (next S&CW

At CERN)

Conclusions

- **Collisions are the real new thing**
  - Get analysis activities done!
- **Go for simple effective solutions**
  - Keep open to new ideas and inevitable evolution on all fronts
- **Trends/desires**
  - Simplify the sites
    - less persistent services, e.g. catalogues
    - PhD should do more physics less system administration ☺
  - Simplify the global system
    - Privilege client-based sites and on-demand replication (cache)
  - Keep on investing in user support
    - It will be never to much

Credits: material from Dan Van Der Ster (CERN), Johannes Elmsheuser (LMU), Vincent Garonne (CERN) and Nurcan Ozturk (UTA)