Analyse dans le nuage DE

Journées LCG France

Cédric Serfon

Ludwig-Maximilians Universität, München

27 Novembre 2008

Plan

- Introduction
- Analysis tools
- Ganga stress tests
- NAF
- Open issues
- Conclusion

Introduction

Presentation of the cloud

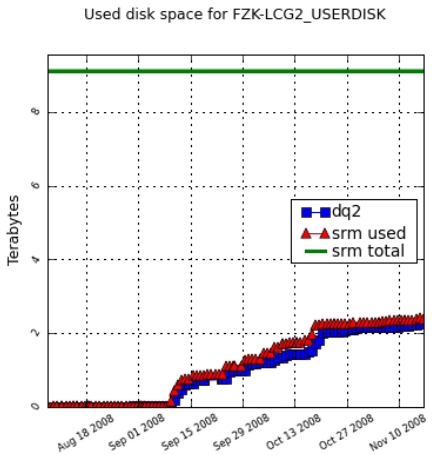
- As for the FR cloud, DE cloud a multinational cloud : DE+CZ+PL+A+CH.
- T1 (GridKa/FZK) + 12 T2s + N T3s (N$\sim$10, not in TiersOfAtlas).
- 2 sites (Desy Hamburg and Desy Zeuthen) have a special role : NAF (National Analysis Facility).
- Very different size for the T2s :
  - From $\sim$100 to $\sim$1000 CPU.
  - From $\sim$10 TB to $>$200 TB.
- Mainly dCache sites (9/12).

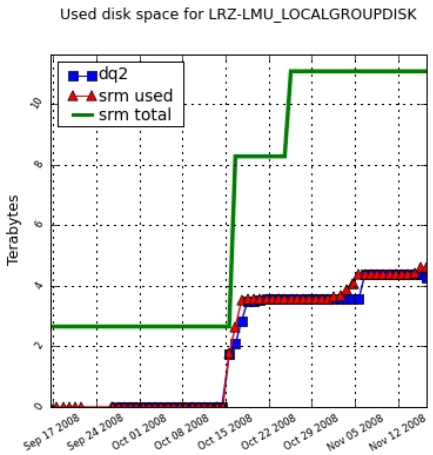
Data distribution over the DE cloud

- In DE cloud, at least 3 copies of AODs available (Computing model asks only for 2 AOD copies):
  - 1 at GridKa.
  - 1 at the Desy sites.
  - 1 splitted in the other 6 T2s (each of them get 16.6%).
  - 3 T2s get a smaller share (between 5 and 17%) of random AODs.
- Discussion is on-going to see if we cannot get 100% of ESD during the LHC start-up at GridKa.
- Additional request for RDOs from people working on calibration studies (HEC, muons) will also be served.

Analysis in the cloud

- Since long time many user analysis has been run in the cloud. Right now 21.5 TB !!! of datasets on USERDISK and LOCALGROUPDISK





- Many reasons for that :
  - Always a least one copy of all AOD datasets was available in the cloud.
  - Many Ganga tutorials in Germany to teach people how to run Distributed Analysis. Now new students are being teached by old ones.
  - Cloud support available through hypernews.
- Right now only Ganga is available in the cloud. Panda queues not foreseen on the short term.

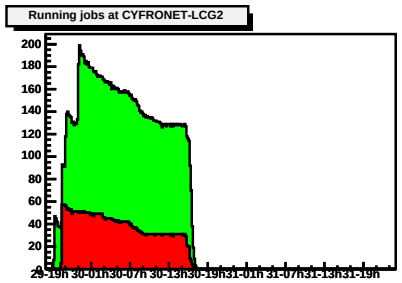
Distributed Analysis stress test

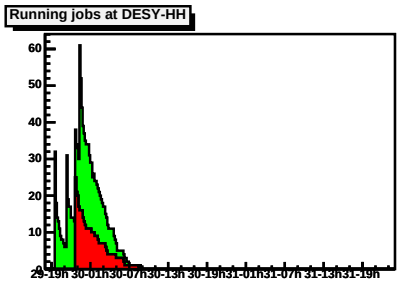
Ganga stress test : Presentation

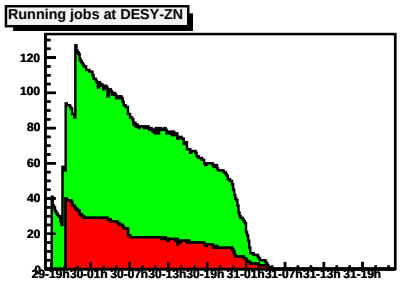
- "Standard" real user (from M. Biglietti) analysis using recent datasets (mc08.*) and release 14.2.20 : read files with local protocol (rfio, dcap), compute some values, write into ATLASUSERDISK.
- Already 2 runs in the DE cloud. These runs allowed to identified major problem (missing software, information not published in the bdii).
- 9 sites tested (out of 12 for the DE cloud).
- Between 200 to 400 jobs submitted by sites.
- At the time of these tests, no Production was running on any site.

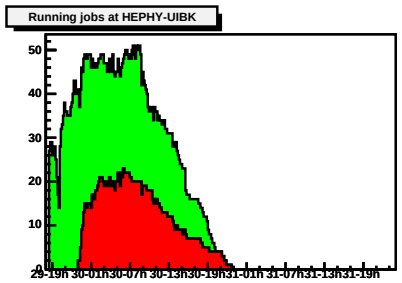
Ganga stress test : Results of run 1

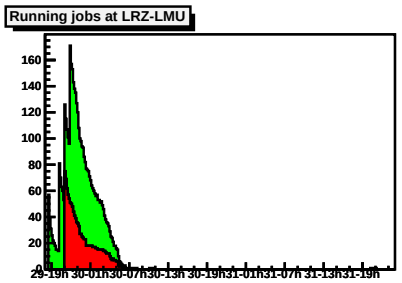
Running jobs on the sites vs time (2 days slice)

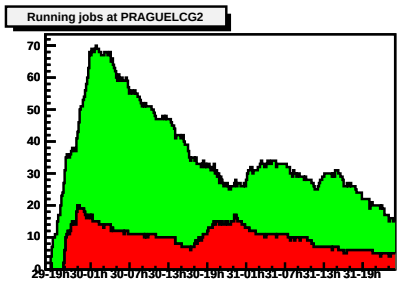


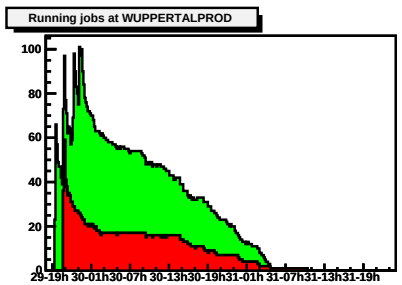












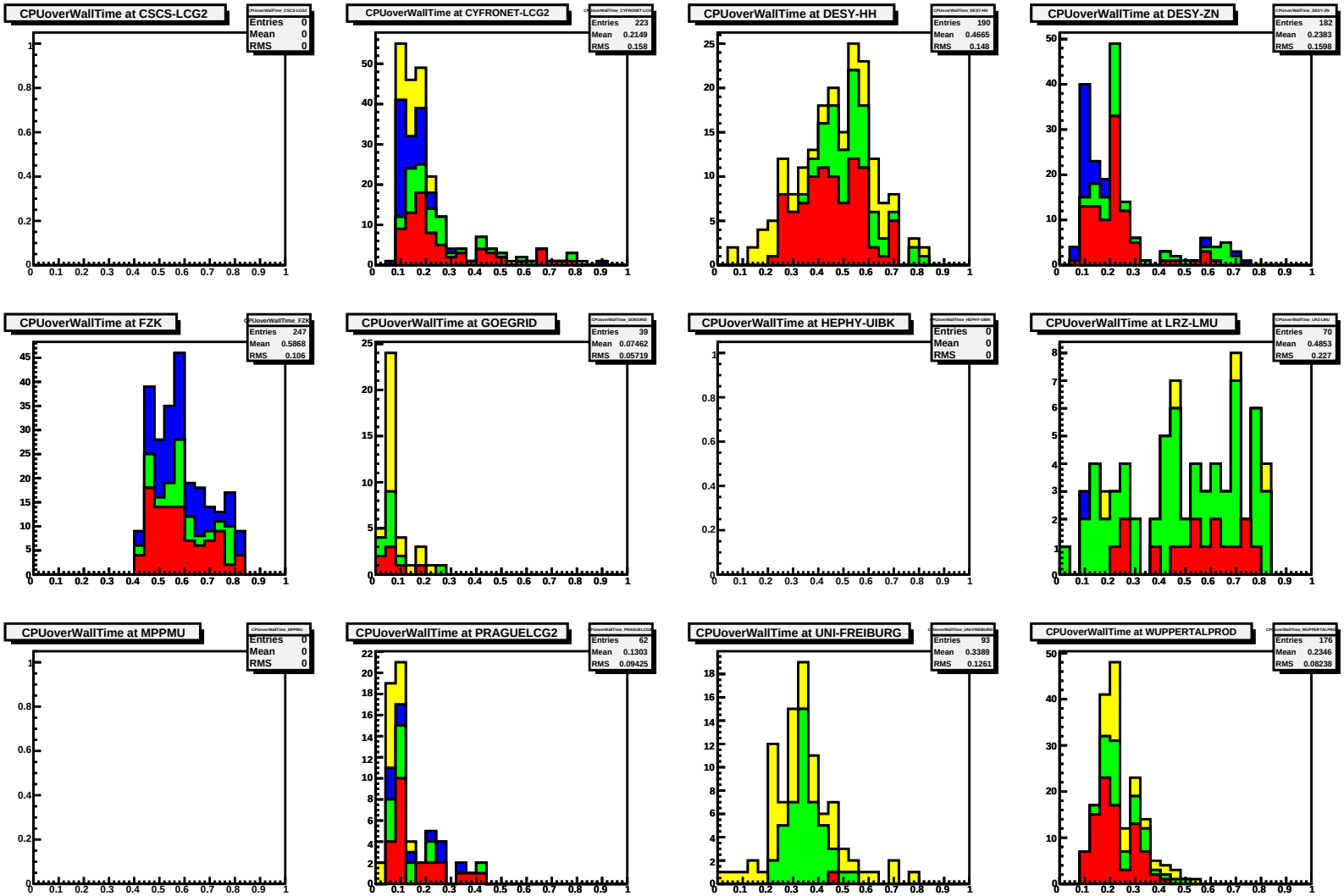
- Color represent differents users
- Sites have very different size : Smaller O(40), bigger $>1000 \rightarrow$ Big differences to process the same number of jobs.

Ganga stress test : Results of run 1

Site	Submited	Running	Completed	Failed	Total
CYFRONET-LCG2_MCDISK	0	0	391	9	400
DESY-HH_MCDISK	0	0	296	4	300
DESY-ZN_MCDISK	0	0	380	20	400
HEPHY-UIBK_MCDISK	0	0	185	100	285
LRZ-LMU_MCDISK	0	0	369	31	400
PRAGUELCG2_MCDISK	0	0	164	136	300
WUPPERTALPROD_MCDISK	0	0	213	87	300
UNI-FREIBURG_MCDISK	0	0	17	14	31

- For all sites more than 50% of the jobs succeeded. For 4 sites even better than 90%.
- Most of the error due to failure to determine TURL.
- Other errors : application errors, expired proxy.

Ganga stress test : Results of run 2

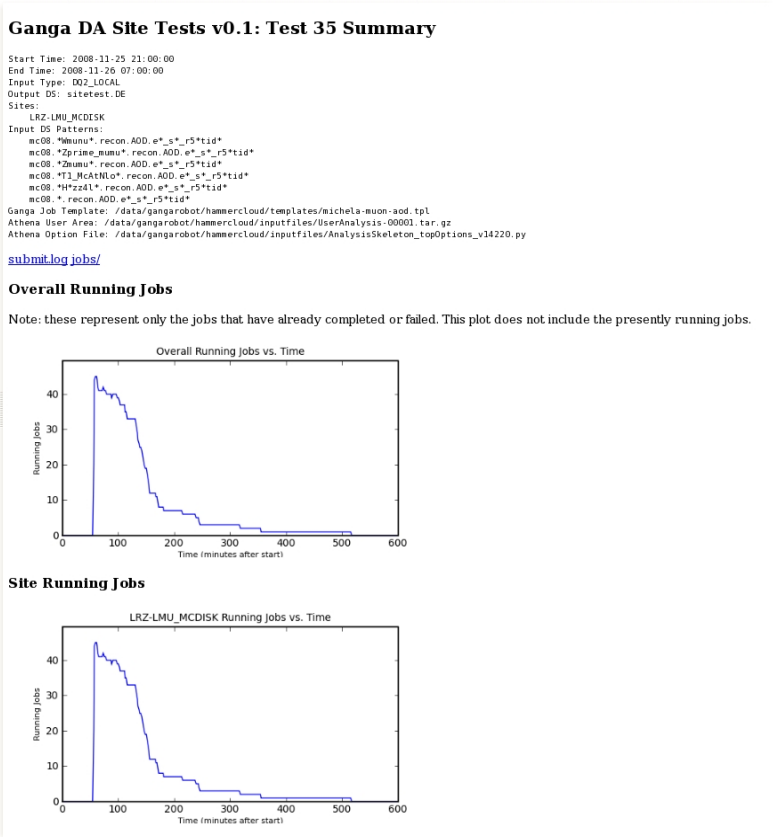


Ganga stress test : Results of run 2

- 2 sites (CYFRONET, PRAGUE) had a bad CPU/Walltime due to network limitation : only 1Gbps links to the pools.
- For in DESY-ZN, only 1 pool used.
- Best performance for bigger sites (FZK, DESY-HH). Probably due to better spread of the data on the pools.
- For LRZ-LMU, some saturation observed on ganglia plots (pools throughput).
- No feed-back yet from other sites with bad CPU/Walltime.

Next steps

- New stress tests are being submitted automatically (like ganga robot) and results available on `http://gangarobot.cern.ch/st/`
- Need to test other features (copy of the files to the WNs, new settings for dCache access...).
- Results can only be interpreted with help of sites !



Ganga DA Site Tests v0.1: Test 35 Summary

```
Start Time: 2008-11-25 21:00:00
End Time: 2008-11-26 07:00:00
Input Type: DQ2_LOCAL
Output DS: sitetest.DE
Sites:
    LRZ-LMU_MCDISK
Input DS Patterns:
    mc08.*Wmunu*.recon.AOD.e*_s*_r5*tid*
    mc08.*Zprime_mumu*.recon.AOD.e*_s*_r5*tid*
    mc08.*Zmumu*.recon.AOD.e*_s*_r5*tid*
    mc08.*T1_McAtNlo*.recon.AOD.e*_s*_r5*tid*
    mc08.*H*zz4l*.recon.AOD.e*_s*_r5*tid*
    mc08.*.recon.AOD.e*_s*_r5*tid*
Ganga Job Template: /data/gangarobot/hammercloud/templates/michela-muon-aod.tpl
Athena User Area: /data/gangarobot/hammercloud/inputfiles/UserAnalysis-00001.tar.gz
Athena Option File: /data/gangarobot/hammercloud/inputfiles/AnalysisSkeleton_topOptions_v14220.py
```

submit.log jobs/

Overall Running Jobs

Note: these represent only the jobs that have already completed or failed. This plot does not include the presently running jobs.

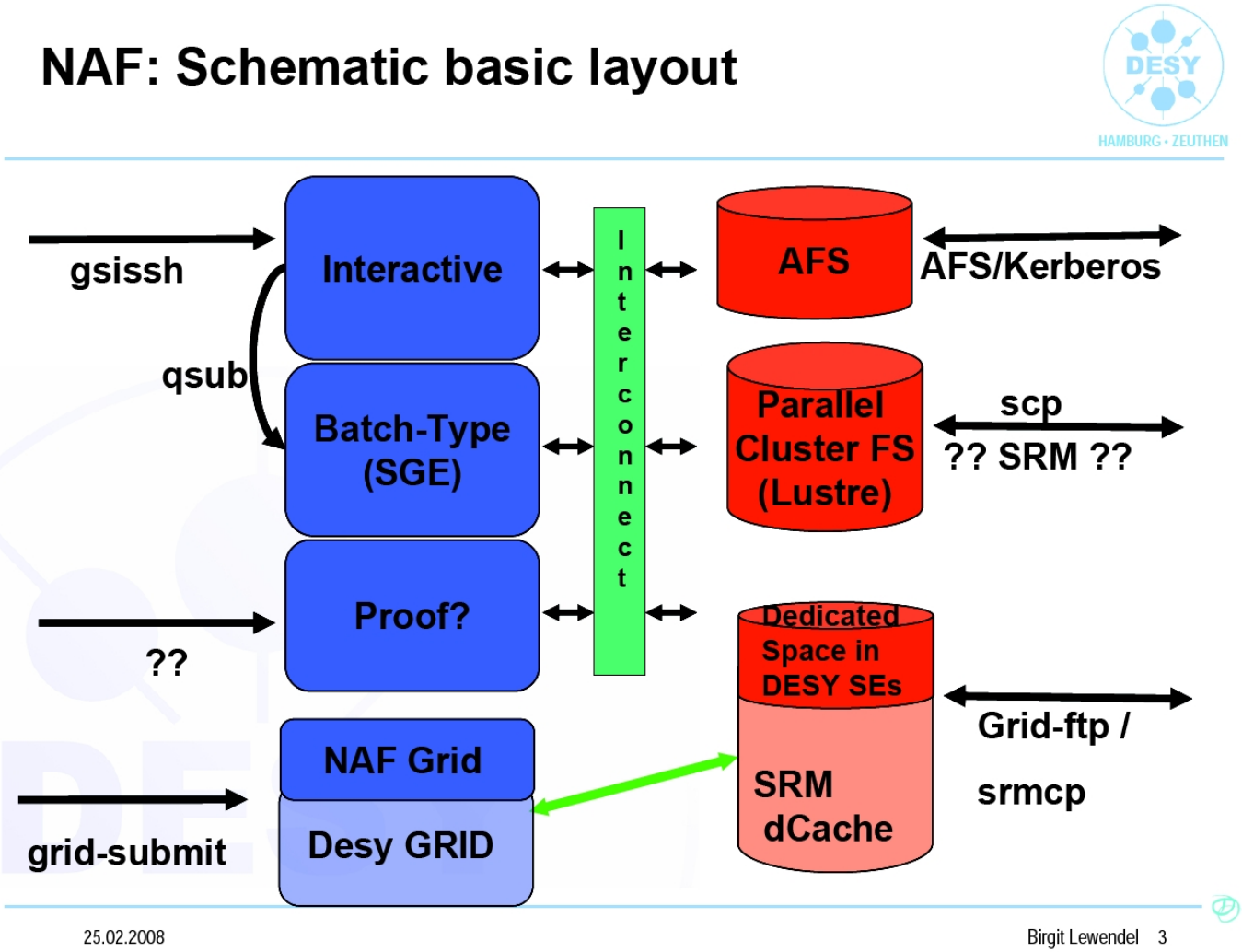
Site Running Jobs

NAF

What is the NAF ?

- NAF for National Analysis Facility (`http://naf.desy.de`). 2 sites :
  - Desy Hamburg.
  - Desy Zeuthen.
- Provides :
  - Additional Grid ressources.
  - Interactive ressources.
- Access via gsissh restricted to Role /atlas/de

NAF layout


NAF: Schematic basic layout
DESY
HAMBURG • ZEUTHEN
gsissh
Interactive
qsub
Batch-Type (SGE)
??
Proof?
grid-submit
NAF Grid
Desy GRID
Interconnect
AFS
AFS/Kerberos
Parallel Cluster FS (Lustre)
scp
?? SRM ??
Dedicated Space in DESY SEs
SRM dCache
Grid-ftp /
srmcp
25.02.2008
Birgit Lewendel 3

Software available

- /afs mounted with 500 MB home directory.
- Lustre space (large bandwith)
- No dedicated queues for German users, but people using ATLAS-D role (/atlas/de) have higher priority.
- Batch system used : SGE.
- Software installed
  - General software : (root, UI).
  - Atlas Software : Athena, EventView grouparea, Ganga

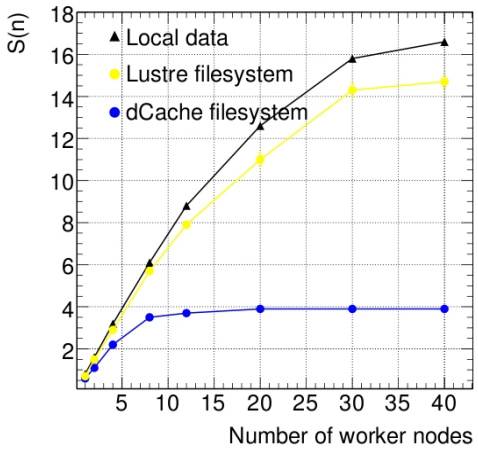
Miscellaneous

Extra resources for German users

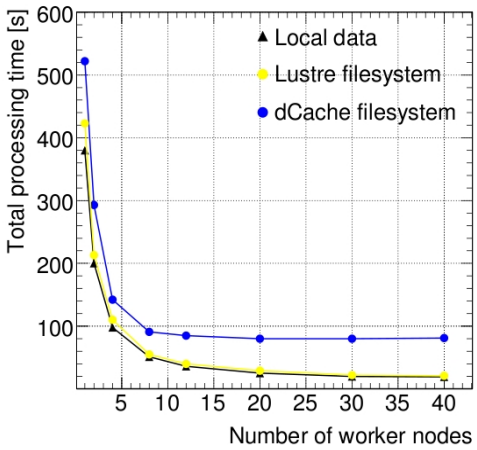
- Extra resources (CPU+Disk) available at GridKa and on some T2s.
  - 400 TB/700 kSi2k at GridKa
  - 200-400 TB/1500-2000 kSi2k spread over 4 T2s
- Computing Resource Board to decide how these resources (Disk) will be used. Plan is to have extra ESD/RDOs/RAW to perform calibration studies.

PROOF - Storage studies

- PROOF cluster installed at LRZ-LMU.
- Test being conducted to determine performance from different storage/file system, using real analysis on DPD.



(a) Speedup factor S(n)



(b) Total processing time

- Speed-up factor scales almost linearly with number of workers (up to 20) for local data and Lustre. Quick saturation with dCache.

Open issues

- T3s. How they fit in the current schema ?
- Panda queues in the cloud ?
- Data management on LOCALGROUPDISK : how to control user space ?
- How to implement share for user/prod ?
- Tag analysis.

Conclusion

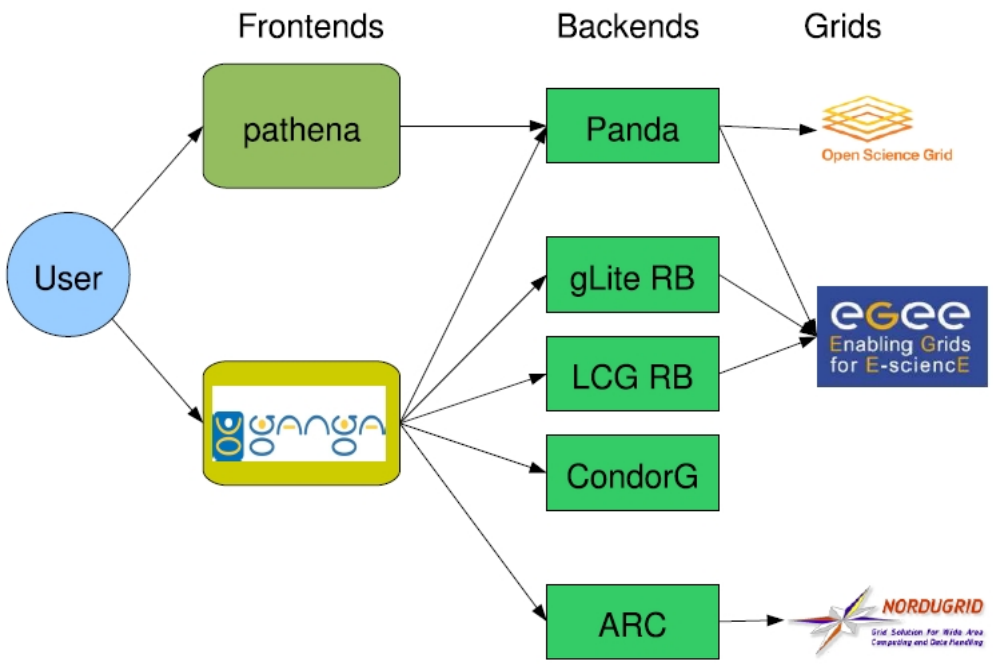
- Many users are now running Distributed Analysis in the DE cloud and most of them are happy :-)
- Need to increase stress test to see if we can handle a higher load of user analysis jobs. Weak point now seems to be Storage Element access.
- NAF is available and used by many German users.
- Still open issues that will keep us busy till the first collisions.

Questions

(à moins que tout le monde ne veuille aller manger)

Backup

Distributed analysis in tools (for non ATLAS people)



- 2 tools available : pathena (ATLAS specific) and Ganga (ATLAS/LHCb)

User analysis : use cases

1st use case : AOD/DPD analysis

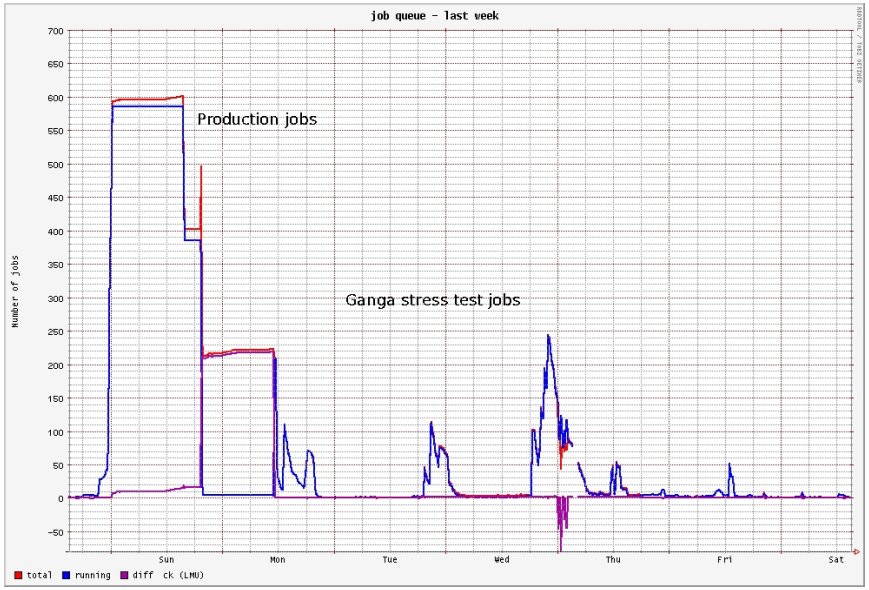
- Run over an AOD or D1PD or D2PD.
- Compute some variable and dump the content into a ntuple (D3PD) using ATLAS tools (EventView, DPDMaker...).
- Retrieve the data locally (dq2-get) and perform analysis and plots histos.

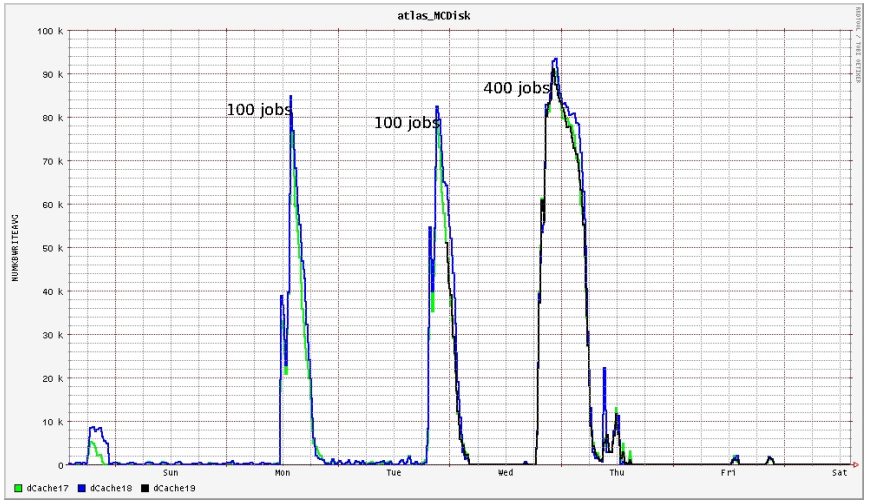
2nd use case : Small MC Production

- Local production of evgen uploaded on the Grid via dq2-put
- Running MC production (ATLFAST2 or full sim).
- Storing the output on LOCALGROUPDISK.

Despite large number of users already using the sites, stress tests need to be performed to see what are the current limitations.

Pools saturation



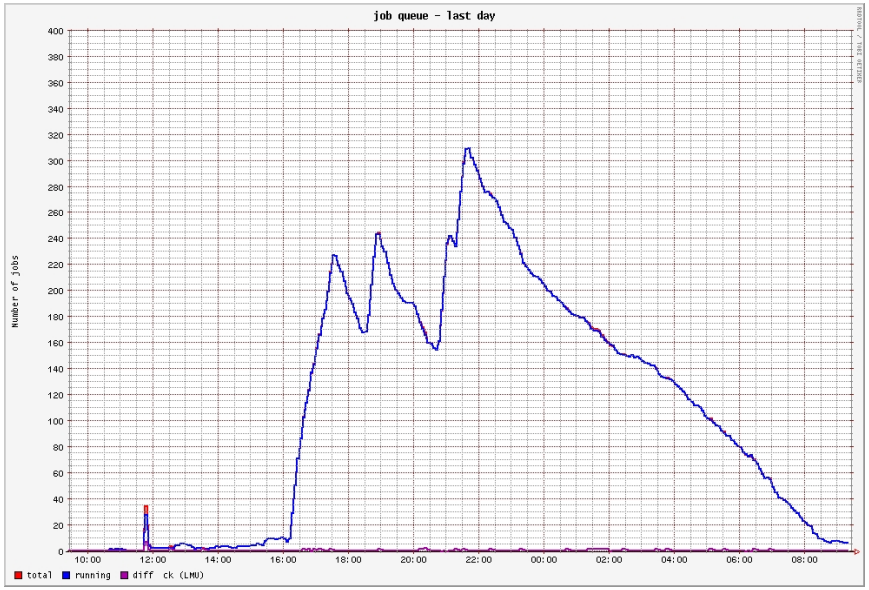


Jobs running at LRZ-LMU

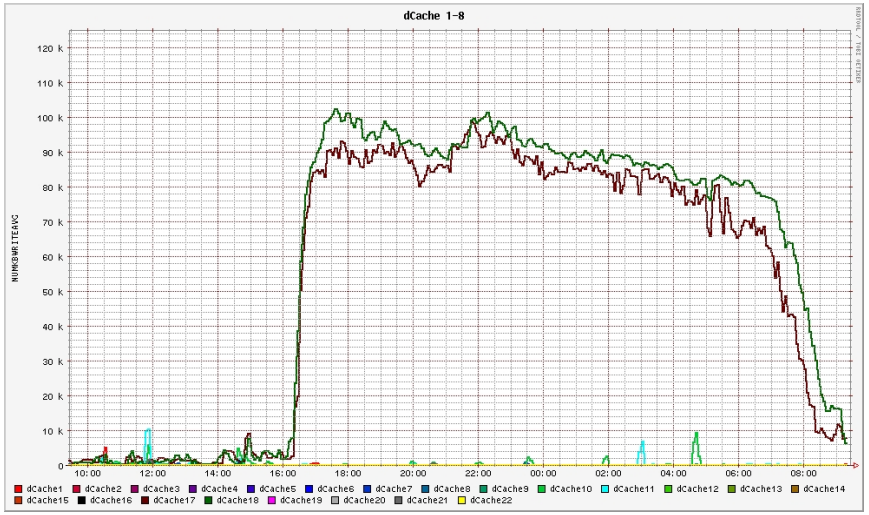
Output from the pool-nodes used

- Increasing the number of jobs from 100 to 300 doesn't increase the outbound traffic from the pools which saturate a bit bellow 100 Mbps.

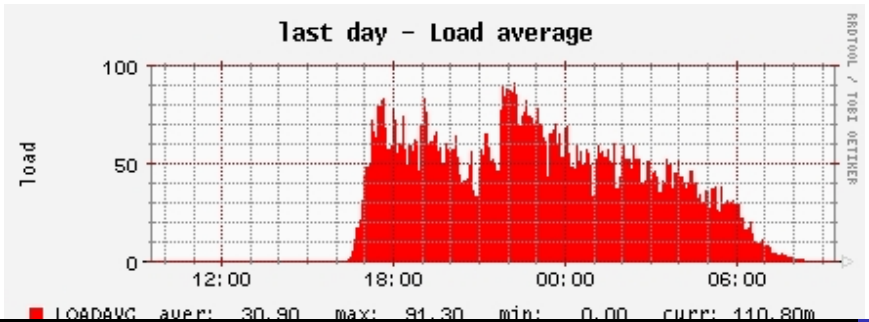
Pools saturation

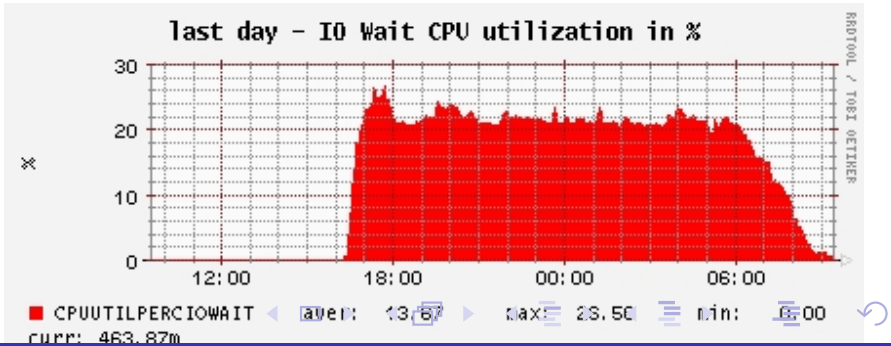


Jobs running at LRZ-LMU



Output from the pool-nodes used





NAF

Support and Documentation

As always this is the hard part!

Support:

- non-experiment specific: `naf-helpdesk@desy.de`
- ATLAS specific:
  - HN: `gridkaCloudUserSupport`
  - `naf-atlas-support@desy.de`

User Communication:

- NAF User Committee: `http://naf.desy.de/nuc`
  Jan Erik Sundermann, Wolfgang Ehrenfeld

Documentation (feel free to contribute):

- general NAF: `http://naf.desy.de`
- ATLAS@NAF: `http://naf.desy.de/atlas`

(W. Ehrenfeld)