Summary of ATLAS-D Computing Questionnaire: Top

U. Husemann, July 8, 2010

General Remarks

- Received answers from 8 institutions: Berlin, Bonn, DESY, Dortmund, Dresden, Göttingen, MPI, Wuppertal -> thanks!
- Answers come from people with very different levels of experience and grid computing knowledge, from “power users” to beginners

Executive Summary on Questions

Data Access

- everybody is using MC AODs from the grid (any site)
- most people also use data AODs
- private ntuples: mainly local storage (2 also on LOCALGROUPDISK)

Data Transfer

- Groups with central ntuple production: a few TB per group
- Individuals: several 10s of GBs
- Mostly happy with transfer rate

Data Processing

- Any possible way, depending on the problem
- Half of the groups use local processing, mainly for ntuple-based analysis
- If restricted number of grid sites: mostly German cloud, sometimes also US and France

Processing Experience

- Generally very good processing speed and reasonable I/O speed
- Variable site availability (good e.g. Wuppertal, bad e.g. MPPMU), but downtimes not always well documented and announced
- Low job failure rate, but sometimes waiting in queues for hours

NAF

- (Only?) used by 50% of the groups
- Criticisms: Lustre problems, load balancing on login nodes, frequent down-times, installed software not well documented, Athena CMT problem

Optimization of Turn-Around Time

- Received answers on very different levels: from grid jobs to user level analysis (all with different solutions)
- Grid: dataset splitting, centralized production, veto of unstable sites, …

Group Disk

- Concept not well known: 25% of the groups understood “local disk storage”
- Not used by 50% of the groups

Individual Answers to Questions

Which data have you accessed, and where were they physically stored?

Berlin:
- AOD, ESD at Tier1+2 (German cloud)

Bonn:
- Latest Monte-Carlo productions and data AOD (mc09, data10_7TeV). They were replicated on the GRID

DESY:
1) I have accessed a bunch of SM-jet D3PD's, both regular and slimmed. In addition, I have accessed several MC08 AOD samples and SYNTmaker ntuples.
2) mc09 and data10 AODs (on any grid site containing them), ntuples created on those sites and on LOCALGROUPDISK

Dortmund:
- ATLAS AODs: all over the world, private ntuples: local storage
- the samples included Collisions, Minbias MC, DiJet MC, as b-tagging ntuples, from grid and from Wuppertal group space
- Top MC, somewhere on the Grid (German as well as non-German sites);
- private produced SUSY MC (FZK,DESY-HH,UNI-Dortmund), SUSY, tT, single top, Wbb+jets, W+light jets, WW+jets, WZ+jets, ZZ+jets, QCD+jets

Dresden:
- Über das Grid habe ich auf Daten zugegriffen, die von der HU Berlin zur Verfügung gestellt werden. Diese liegen zur Analyse auf einer Platte am Rechenzentrum der TU in Dresden, die dem Institut an dem ich arbeite zur Verfügung gestellt wird.

MPI:
- mc08, mc09 AODs. In general datasets were replicated to the German Cloud.

Göttingen:
1) ttbar AODs and W+jet AODs (German and French group disks), pixel hit samples (?), PAU ntuples (Lyon)
2) data10 events, mc08, mc09 stored on Tier1 and Tier2 in and outside of Germany
3) all kinds of actual data + MC from top and backgrounds, Z->ll, QCD, minbias

Wuppertal:
- AOD, TopPhys D2PD, SM D3P

How much data did you transfer to Germany, and how was the bandwidth?

Berlin:

- Only data stored inside Germany has been used - no transfer from abroad.

Bonn:

- Approximately 5TB of ntuples.

DESY:

1) Several GB. The bandwidth has been up and down, I must say. Recently it has been OK, but around Easter it was incredibly slow when downloading stuff from the grid. I don't remember any numbers though.
2) 5 TB, don't know about bandwidth as I used DaTRI

Dortmund:

- During the last year different people were involved in data transfer for analysis. Each person transferred some 100 GB, for our total group we roughly estimate a few TB in total. We have no exact speed measurement, however speed was bearable but not great

Dresden:

- Ca. 50 Gigabyte. Transferprobleme gab es nur einmal beim Zugriff auf Daten, die in Prag hinterlegt waren. Die Daten sind vergleichsweise schnell da, für einen 25 Gigabyte-Transfer kann man 1 bis 2 Stunden rechnen, was ich für ok befinde.

Göttingen:

1) ?
2) Have not transferred data to German sites - only finished jobs.
3) none

MPI:

- Output datasets were already in the German cloud, so no external datasets have been transferred in Germany

Wuppertal:

- total: O(10 TB), transfer times reasonable (1 day per dataset)

Where did you process the data (e.g. locally, German Tier 2, GridKa, NAF, CERN)?

Berlin:
- German Tier2

Bonn:
- GRID

DESY:
1) I have processed data using the normal grid (mostly at BNL I think), at local Stockholm machines and on the NAF machines. The local jobs are of course nothing major, like ntuple making, but rather just final analysis on the ntuples.
2) any available grid site for AODs, test jobs at NAF, ntuples locally

Dortmund:
- locally, GridKa, CERN and other German as well as non-German sites

Dresden:
- Bislang sind alle Analysen lokal auf der Rechnerfarm des Institutes an dem ich arbeite geschehen.

Göttingen:
1) German and French cloud, CERN batch, locally, NAF
2) Process data on Tier2s in Germany and many at LBNL since the data was not available here.
3) grid or locally at CERN

MPI:
- German Cloud and RZG Tier2 (MPPMU), Local

Wuppertal:
- All of the above

What was you experience processing the data (e.g. site availability, processing speed, I/O speed)?

Berlin:
- Processing speed always good.
- Some sites are often not available, eg MPPMU, CSCS-LCG2, CYFRONT-LCG2. Information about broken sites often not updated. Sites are even accepting jobs then.
- Important information (updates, downtimes etc.) should always be spread on the Atlas distributed analysis hypernews forum by GRID admins.
- Some nodes seem to be incorrectly installed.
- I/O speed okay.
- At many sites the pre-staging is not working which means the autoconfiguration tool cannot be used at all. Here, a quick solution is desired.

Bonn:
- OK

DESY:
1) I don't think this question is relevant for me given the previous answer...
2) site availability varies, also speed

Dortmund:
- There is a significant spread in the reliability of different grid sites.
- Partially we had to exclude some sites due to permanent problems, general speed was fine, also I/O was fine most of the time; especially so far our experience with the Wuppertal Tier2 is positive

Dresden:
- Transfer von anderen Orten nach Dresden funktioniert sehr gut, alles andere habe ich noch nicht ausprobiert.

Göttingen:
1) all reasonably fine
2) Find several transfer problems on occasion.
3) no problems

MPI:
- Data processing went fine, job failure rates order 1%. Job execution sometime (very) slow: job waiting in the queues for hours.

Wuppertal:
- With a few exceptions, German T2 grid sites are generally very reliable and fast.

Have you used the NAF, how was your experience there?

Berlin:
- NAF is slow due to high user load. The automatic host dispatcher (via gssh) for an efficient distribution of ressources does not work.
- ATHENA way too slow (CMT?).

Bonn:
- No

DESY:
1) Yes. I have found NAF useful mostly for storing large amounts of data. However, the frequent down-time periods are a problem.
2) Yes. speed has improved though sometimes still very slow. Sometimes database or athena releases missing, but were installed right-away on request.

Dortmund:
- Only partially, e.g. for software tutorials
- We see limitations in documentations e.g. about installed s/w and limited storage; it was usually easier and faster to get results on a small local cluster

Dresden:
- Noch nicht

Göttingen:
1) just sometimes, worked fine if data once data is available on disk
2) I have used it for one project - overall it was quite efficient and I had no problems.
3) no

MPI:
- No experience there

Wuppertal:
- I use it regularly, positive experience though Lustre problems sometimes affect the site

Have you tried to optimize the turn-around time for your jobs? How?

Berlin:
- No.

Bonn:
- Yes. By centralizing ntuple production and standardizing jobs.

DESY:
1) No, except by moving to slimmed ntuples.
2) veto unstable grid sites, automatically request datasets to localgroupdisk

Dortmund:
- Some of us manually split the dataset or used torque to do so, others did not tried any optimization.

Dresden:
- Meine derzeitige Analyse umfasst 31 verschiedene "Arbeitspunkte" (Sets von Einstellungen verschiedener Input Parameter). Für jeden dieser Arbeitspunkte benötigt die Analyse 3 Stunden. Über Optimierungen habe ich nachgedacht, z.B. über das Erstellen von Sub-Samples in denen bestimmte Cuts schon durchgeführt wurden und damit die Anzahl der Ereignisse verringert wurde oder die Samples so aufbereiten, dass z.B. ein Overlap Removal standardmäßig durchgeführt ist und dies nicht bei jeder Analyse wieder ausgeführt werden muss.

Göttingen:
1) choose grid/CERN batch/local cluster for different expected run times
2) No
3) use pathena instead of ganga and don't limit to any cloud. Use --noBuild.

MPI:
- Adjusted the number of subjobs and input files per subjobs.

Wuppertal:
- I am developing my own nanoparticle-based quantum-computing multiprocessor mainframe (Not)

Are you using a group disk? If yes: how are you using it?

Berlin:
- We do not use a group disk.

Bonn:
- No, but we are interested to learn about that.

DESY:
1) What do you mean by group disk?
2) No

Dortmund:
- yes, at TU-Dortmund and DESY-HH for permanent storage of private MC productions and for access to the b-tagging ntuples

Dresden:
- Jeder Nutzer am Institut bekommt Speicherplatz auf einem Raidsystem, das wie eine normale Festplatte mit Backup funktioniert und hat Schreibrechte für seinen eigenen Ordner.

Göttingen:
1) no
2) yes: Storing finished jobs locally
3) no

MPI:
- Only accessing data there stored.

Wuppertal:
- I am group space manager for TopPhys and I replicate and store all produced datasets on group disks.