GSI-Computing Users Meeting
date: October 01 2012

Participants:

SC	Kilian Schwarz (KS)
SC	Carsten Preuss (CP)
SC	Peter Malzacher (PM)
HPC	Bastian Neuburger
HPC	Jan Trautmann
HPC	Thomas Stibor
HPC	Christopher Huhn (CH)
HPC	Victor Penso (VP)
ALICE	Jochen Thaeder (JT)
HADES	Jochen Markert (JM)
HADES	Tetyana Galatyuk
Theorie	Thomas Neff
FOPI	Yvonne Leifels (YL)

Action Items (AI):

- Create up-to-date fill stats for the Lustre file-systems (TR)
  - this has been done already.
  - A next “du” run will take place after the Quark Matter.
  - /hera and /lustre check is currently running. Will be finished within 1 week.
  - The Robin Hood Runs should be done on monthly basis and the results should be published on a Web Site.
  - This AI will be closed.
- Prepare interactive Wheezy test nodes (HPC)
  - CH upgraded his desktop and it works
  - a test node for general testing will be created
    - CH has done successful testing.
  - requests should go to BN
    - requests came from FU and JT. BN is working on them.
    - So far nothing has been done, also because Lustre modules for Wheezy are not yet existing.
    - It should be possible, though, to create Wheezy test nodes without Lustre or with Lustre being mounted via NFS
- Investigate the implementation of GridEngine group admins via “sudo -u group_member [qdel|qmod|...]” (HPC)
  - HPC will check if the implementation has been done already.
  - This is needed by ALICE and HADES.
  - HPC will test up to next meeting.
  - There are problems envolved connected to the group membership of the experiment members.
    - In principle it can be possible this way that HADES eliminates ALICE jobs and vice versa (in such a case “Rufbereitschaft” should not be called).
    - One mainly responsible person should be identified for each experiment and communicated to BN.

        - Eventually the group KP1 for ALICE members can be removed completely.
- Status of Hera/gStore connection (TR, Horst Goeringer (HG))
    - HG tested the setup with 200-300 MB/s via Lnet router.
        - Opening to public will be done at the beginning of September after the current beam time
        - HADES should test this (JM)
            - tests have not been done yet.
        - The feature is available and has been tested by Horst Goeringer.
            - More bandwidth (more data mover) will be available after the HADES beam time (October 2 up to the coming Monday).
- representatives from GSI and Frankfurt should sit together to identify how the new /hera mount at Frankfurt can be used most efficiently. Connected topics are e.g. CVMFS mirror for software (IT, ALICE)
    - a CVMFS mirror from GSI to Frankfurt might be problematic due to different operating systems envolved.
    - According to HPC the system is currently not in a usable state. The problem is that /hera is currently mounted only from a small number of clients. Big data throughput is not possible yet.
    - PM suggests that one should sit together with representatives from the Frankfurt side. The timeline until production state should be defined. Currentlly GSI has a bandwidth of 80 Gb to Loewe CSC.
- HADES and ALICE jobs suffer from too many open file handles. Since this is a problem for Lustre a solution has to be found (HADES, ALICE, IT)
    - the problem did not appear anymore because ALICE did not run simulation jobs.
    - JM mentions that the scheduler is unstable
    - JT wants to test once more.
    - This AI will be removed.
- Squeeze desktops have a resolution problem with 16:9 monitors. A solution has to be found (HPC)
    - Stefan Haller (SH) did not know that he was supposed to fix the problem.
    - CH started to look for a solution. When using the correct parameters booting with proper graphic resolution is possible.
    - CH and SH will drop by and solve the monitor resolution related problems of ALICE and HADES.
- a tool should be written to enable black hole discovery in (S)GE (CP)
    - CP: this is a one-liner effort
    - CP wants to discuss if and what actions the script should do, e.g. it needs to be defined what a black whole is and what is to be done with that. Does it affect only one user or all users ? Is only 1 node affected or all ?
    - The idea is that the script should discover black wholes and if one is discovered the node should be removed from the queue and a ticket should be created automatically.
- lxalitransfer1 – 3 should be moved from /lustre to /hera (KS)
    - a discussions started if the boxes should be moved to different hardware and to a different location. This should be clarified within the IT.
- transfers via 10 Gb link should be activated. (KS)
- The LSDMA topics need to be discussed with the FAIR experiments. Therefore a set of concrete questions should be formulated beforehand and circulated (KS, CH)
- The LSDMA gap analysis and application document should be sent to this list (KS)

    - information about LSDMA has been sent to the cluster-mailing list. This AI will be closed.
- The protocoll should be sent around shortly after the meeting (KS)
- AI: within the technical Computing meeting shall be presented on a regular basis how many computing resources each experiment consumed within the last 4 weeks and also how much space is left on the Lustre storage (KS, CH).

Closed and PENDING Action Items:

- Finden einer Lösung für problematische Maschinen, die nicht automatisch aus der Produktion ausgeschlossen werden und somit zu erheblichen Problemen führen können (vom 12.9.11: IT)
    - So lange nicht klar unterschieden werden kann, ob die Probleme von den Jobs oder von der Maschine kommen, ist eine vollautomatische Lösung schwierig.
    - Da ALICE hierunter nicht akut leidet wird das Problem auf PENDING verschoben.
- Um die Storage-Elemente von PANDA und CBM, die derzeit unter Etch laufen, auf Lenny oder Squeeze umziehen zu können, muss hierzu vorher geklärt werden, was von Grid-Seite aus vorbereitet werden muss (vom 16.01.12: KS)
- CVMFSServer for the test Cluster (B) runs on old hardware. Money for proper service deployment is missing. WS should investigate in the next IT coordination meeting who in principle is responsible for buying hardware for server in the Minicube (CVMFS server, Build server). VP mentions that hardware should be bought by the experiments. JM asks VP to specify what hardware would be needed. VP thinks that a server including storage for 3000 Euro in total would be sufficient. VP will make a concrete suggestion concerning the necessary hardware. WS will bring the topic to the IT coordination meeting (05.03.12: VP, WS)
    - together with the last hardware order a server for CVMFS has been bought
- Plan the LSF decommissioning and hardware recycling (HPC)
    - Shutdown of Lenny farm scheduled for the end of 2012 (14th of December 2012)
    - SGE test cluster and old LSF/SGE cluster shall be switched off. A list of computers to be switched off will be created (SC) and sent to the experiments for approval (KS). Affected will be Lenny and Etch boxes, e.g. some lxi machines.
    - work is in progress.
- MPI jobs never scheduled outside the default queue
    - a patch for GridEngine will be developed (SC)
    - work is in progress (Anar Manafov (AM))
- Sketch lustre-hera migration time line (TR)
    - estimate of exact time lines won't be possible before Decembre.
    - This AI has been moved to PENDING.

Minutes of last meeting have been accepted. The minutes will be written in English. The meetings will be done in German language.
The minutes should be distributed shortly after the meeting.

TOP1: Status report HPC (CH)

- everything like always
- instabilities with Prometheus experienced
  - VP: changes have been done with the scheduling. Job arrays should be used. CP: This leads to less information overhead and scheduling is faster.
  - VP: large cluster for GE according to documentation is 256 nodes. Starting with 1000 nodes and 40000 cores one runs against the limit.
  - Does and hardware upgrade of the master make sense ?
    - CP: the GE master dies because of lack of memory and is moreover too slow. Currently the machine has 8 GB RAM.
- On the 10th of October /lustre will be switched off because the UPS will be moved from C25 to the Testing Hall. For this also cabel connection work is needed.
  - The UPS plugs are already there. Infiniband switches will be connected to standard electricity supply.
  - 3 machines have been moved already. They only have to be formatted and mounted. But then they might be overloaded quite fast.
- Thomas Roth (TR) will not be around from Wednesday up to October 22nd. TR is the only person who has a complete knowledge of the set up. Know how transfer is taking place, though.
- 30 TB more can be moved into the new storage cluster during the week.
- Call for tender for new Lustre file servers has been written. Money is available and a Banf has been created. The new machines will be usable probably in April 2013.
  - YL: if the money will be spent after February 28th 2013 an extra application has to be written.
- JT: 93% of /hera is full. A better coordination with HADES is needed.
  - JM announced the HADES activities in the last Technical Computing Meeting.
  - Large scale experiment activities should be announced additionally by the experiment via e-mail.
  - HADES: eventually /hera needs also to be cleaned up.
  - JT: for this new Robin Hood runs are needed.
  - CH would like to know if the performance becomes worse if the file system is such heavily filled
    - JM: the write performance does not. HADES was happy with the performance during the DST creation.
  - CH: HPC will try to create more space on /hera.
- It does not make sense to clean up more data in /lustre if the moving to /hera is delaying.
  - Currenly /lustre is about 50% full. HPC things that this should be fine for the time being. No more data need to be deleted in very near future. But to be on the safe side TR shall be asked.

TOP2: Plans for next weeks

- HADES:
  - beam time up to Monday.
  - Not many data and no big load will be created.
- ALICE:
  - normal operation.
- Theory:

    - no large activities foreseen.

TOP3: preparation FAIR computing meeting.

- The agenda of the upcoming FAIR computing meeting is being presented
    - Dokumentmanagement
    - Oracle
    - Dark Fibre
    - Domainstruktur
    - Neue Massenspeicher
    - Planungsstand Green Cube
    - projizierte Leistungsaufnahme
- YL: "dark fibre" is the link to Frankfurt, "neue Massenspeicher" is the upgrade of the GSI Lustre storage capacity, "DMS" is the move from "OnTeam" to a new system. Beta testing should be done by CBM since they have the largest number of documents in the system. "Oracle" has been put on the agenda according to the request of users.
    - CH: hardware for the new Posgres servers has been bought already. The OS has been installed, the DB packages will be. DB administration will be done by Core IT and Admin IT. Who will move applications from Oracle is unclear.
    - PM: this should be discussed tomorrow.
    - CH: so far this did not work out perfectly.
    - CH: concerning "Domainstruktur" - it affects the Windows domains: in the context of the strong separation of Core IT and Admin IT also the Windows infrastructure should be separated. A high security area will be created which includes 2 factor authorisation. The target for this is the buying department, the personal department and the law department.
        - This includes also a new exchange server for e-mails with 2 factor authorisation. This means that every GSI member will need a smart card.
        - JM thinks that this might be over the target.
        - CH mentions that some of those things GSI is required to do by law.
    - CH: "Planungsstand Green Cube": the company which was responsible for planning went bankrupt. Now a new company is needed. The delay will be ½ year. The new planning company will be a professional computing centre planning company.
    - PM: "Leistungsaufnahme" - FAIR GmbH did an evaluation for the running cost of FAIR computing. These numbers went without discussing them beforehand directly to the Ministry. The numbers may have been too low. Volker Lindenstruth will therefore present his own numbers in the meeting tomorrow.

TOP4: AOB

- PM suggests to present in this meeting what the experiment computed within the last 4 weeks on a regular basis and also what space is left in Lustre. This shall be made an AI.
- The experiments have to agree on how to distribute the compute cluster among themselves. So far there was no delay component in the fair share algorithm so that all ALICE activities done at the beginning of the cluster life time still counted up to now.
    - PM: if the experiments want to discuss that on a political level they should announce this topic in the FAIR computing meeting.

    - CH: share have to be distributed on project level. We can make all projects equal in terms of size. One could also create some kind of dependency to the money invested.

TOP5: next meeting

- the next meeting will take place at the usual time (Monday, November 5, 2012, from 2 pm to 3 pm) in the usual location.