GSI-Computing Users Meeting
date: September 03 2012

Participants:

SC	Kilian Schwarz (KS)
SC	Carsten Preuss (CP)
HPC	Jan Trautmann
HPC	Thomas Stibor
HPC	Christopher Huhn (CH)
HPC	Victor Penso (VP)
HPC	Thomas Roth (TR)
CBM/IT	Florian Uhlig (FU)
CBM	Volker Friese (VF)
ALICE	Jochen Thaeder (JT)
HADES	Jochen Markert (JM)
Beschleunigerphysik	Paul Görgen
PANDA	Klaus Goetzen

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.
- 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.
- 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.
- 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.
- 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.
- 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)

- Squeeze desktops have a resolution problem with 16:9 monitors. A solution has to be found (HPC)
- a tool should be written to enable black hole discovery in (S)GE (CP)
  - CP: this is a one-liner effort
- lxalitransfer1 – 3 should be moved from /lustre to /hera (KS)
- 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)

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)
- CVMFS muss auf Squeeze-Desktops getestet werden (vom 06.02.12: JT, IT)
  - Because of heavy workload CVMFS on squeeze desktops does not exist so far. Currently this task does not have highest priority for the HPC group.
  - This AI shall be moved to “pending”.
  - FairRoot would like to move completely to Squeeze. In that case no support for older OS would be given anymore. Precondition would be to have running CVMFS on all desk top machines
  - CVMFS should run on all Squeeze desk tops
  - This AI can be closed.
- 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
- Cleanup of GridEngine jobs that are not scheduled during a given interval (HPC)
  - CP is not convinced.
  - AI closed.
- 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))
- Migrate submit nodes (pro.hpc.gsi.de) to a sub-net with internet and NFS access
    - new submit nodes have been put into operation (lxsub5-14)
    - corresponding information has been sent via the hpc-info list.
    - This AI has been closed.
- 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.
- Check if version control can serve as a backup replacement for ALICE software development (JT)
    - this AI will be removed.
- Increase/remove job limits on Prometheus for hadesdst account (HPC)
    - this has been done by HPC already
    - should be tested by HADES (JM)
    - This AI has been closed.
- names of the various IT Cluster and Storage systems should be published (IT)
    - see http://wiki.gsi.de/cgi-bin/view/Linux/BatchFarm
    - This AI has been closed.
- the way GridEngine determines a suitable queue for a job needs to be optimised (JW, IT).
    - Queue priorities have been introduced.
    - This AI has been closed.

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)

- Call for tender for 30 new Lustre OSSs with > 3 PB net is in preparation
- JM: “ls” on /lustre hurts.
    - JT: also on /hera one is happy when the prompt returns. ALICE deleted already a lot of data.
    - CH: it is still unclear why /hera is slow sometimes.
- JT would like to delete data, but data location to some extend is unknown. Therefore a new “Robin Hood” run would be nice to have.
- Delivery of Melanox Infiniband equipment arrived. But the racks are still without electricity.
    - TR: 8 emptied boxes with 27 TB each are waiting to be moved to TH as soon as the rack infrastructure is ready.
- JM: how robust is /lustre if you have many open files. Currently on 1 server 10000 open file handles have been seen.
    - VP: executables should be copied from /hera to /tmp. Every job has a /tmp dir where files can be copied to. Input files should go to CVMFS. Also in ALICE we

have a problem with open file handles. 6000 jobs produce 90000 open files. Currently a manual how to use Lustre properly is being written.
    - JM: parameter files in CVMFS would need too much processing time.
    - JT: a solution has to be found by working together (AI).
- JT: monitor resolution on Squeeze desk tops is still a problem.
    - CH: Stefan Haller is doing this, but he is on holidays. A problem are 16:9 monitors. Eventually are the drivers too old.
- JM: is there a switch for enabling black hole detection in (S)GE ?
    - CH: eventually should a job description contain an ETA. If this does not fit …
    - CP: will enable this capability by writing a small tool (AI).
- lxp025 – lxp029 have been switched off

TOP2: Plans for next weeks

- ALICE:
    - normal operation.
    - More simulation
    - more user shall be moved to /hera. But for this more space is needed and therefore the output of "du".
- CBM:
    - nothing unusual
- HADES:
    - DST production.
        - This will be no problem if there is space on /hera.
        - Start will be next week.
        - The idea is to run on Prometheus.
    - It will need some time to move users from /lustre to /hera.
- Beschleunigerphysik:
    - a work around for the MPI problem needs to be found
    - no other issues
- HPC:
    - Lustre is being moved
    - more Lustre file servers are being bought
    - half of HPC and IT-security is busy doing the Windows domain restructuring.
    - Thomas Stibor is a new member who will work especially on Lustre improvements.
- PANDA:
    - no news
- SC:
    - no news

TOP3: next meeting

- usual date and usual place
    - Monday, October 1, 2pm-3pm in KBW building

TOP4: AOB

- lxalitransfer1 – 3 should be moved from /lustre to /hera
- transfers via 10 Gb link should be activated.

TOP5: LSDMA

- a short introduction to LSDMA (Large Scale Data Management and Analysis) has been given.
- LSDMA is organised in 5 Data Life Cycle Labs (key technologies, energy, earth and environment, health, structure of matter) and the Data Service Integration Team (DSIT).
- GSI is actively taking part in the DLCL “structure of matter” (FAIR) and DSIT. In total GSI IT will be able to hire about 4 persons for LSDMA.
- DSIT is supposed to take care of IT topics which are of interest for more than 1 single DLCL. The identified topics of common interest are:
  - Federated Identity Management
  - Federated Data Access
  - Meta Data
  - Archive Service
  - Monitoring
  - Data Life Cycle
  - Distributed Computing
- The main DSIT topics as well as the topics of the DLCL “structure of matter” needs to be discussed with the FAIR community.
  - VF suggests that a set of concrete questions should be formulated beforehand and distributed to the experiments (AI)
- The LSDMA gap analysis and application document should be sent to this list (AI)