GSI-Computing Users Meeting
Datum: 06.02.2012

Teilnehmer:

SC	Kilian Schwarz (KS)
SC	Carsten Preuß (CP)
SC	Florian Uhlig (FU)
SC	Mohammad Al-Turany (MA)
HPC	Christopher Huhn (CH)
HPC	Thomas Roth (TR)
Core-IT	Horst Göringer
ALICE	Silvia Masciocchi
ALICE	Jochen Thäder (JT)
HADES	Jochen Markert (JM)
HADES/EE	Michael Traxler (MT)
HADES	Tetyana Galatyuk
Theorie	Thomas Neff (TN)
PANDA	Klaus Götzen
SIST	Torsten Radon

Action Items (AI):

- Finden einer Lösung für die Probleme von SGE, Memory korrekt zu Accounten (vom 12.9.11: IT)
    - Die auf realen Speicher bezogenen Limits werden nicht eingehalten, was aber ein Linux-Problem ist. Limits bezüglich virtuellem Speicher funktionieren.
    - An einem Patch wird seitens der IT gearbeitet. Hierbei wird es sich vermutlich um einen Cron-Job handeln, der sich die Speicherlimits der Jobs anschauen soll und auf den Batchfarmknoten laufen wird.
        - Der Cron-Job läuft bereits in der Testphase und wird in ein bis zwei Wochen auf dem „neuen Cluster“ installiert werden.
    - Anbei erfolgt eine Definition der verschiedenen GSI-Computer-Cluster und deren Bezeichnung:
        - A) „alte Batchfarm“
            - 2200 Kerne
            - Standort: RZ2
            - Schedulung-Systeme: LSF und SGE
            - Features: shared $HOME
        - B) „Test-Cluster“
            - 1600 Kerne
            - Standort: RZ2
            - Scheduling-System: SGE
            - Features: kein shared $HOME, Software auf CVMFS
        - C) „Minicube-Cluster“
            - Ca. 10000 Kerne
            - Standort: Testinghalle
            - Schedulung-System: SGE
            - Features: siehe B)

- Installieren von OpenMPI mit SGE-Support auf der neuen BatchFarm (vom 12.9.11: HPC)
    - Die sichere Kommunikation der MPI-Jobs im User-Space bereitet noch Probleme.
    - Um das Problem zu lösen sollten sich TN und IT zusammen setzen.
    - MPI-Jobs laufen und wurden getestet (TN)
        - u.U. Probleme beim Accounting (CP)
        - grundsätzlich laufen MPI-Jobs, aber nicht außerhalb der Default-Queue.
        - Tests von CP haben ergeben, dass im „alten SGE-Cluster“ MPI-Jobs sowohl bei Queue-spezifischem Scheduling als auch bei Ressourcen-spezifischem Scheduling funktionieren. Nun muss gefunden werden, wo der Unterschied zum „neuen SGE-Cluster“ ist.
            - Bisher ist es noch unklar, warum es auf einem Cluster funktioniert, auf dem anderen aber nicht. Eventuell muss man die „Resource Quotas“ weg lassen und dann schauen, ob es auf auch auf dem „neuen SGE-Cluster“ läuft.
        - TN: wenn die Default-Queue auf 24 h gesetzt wird, dann wären die Anforderungen aus Theorie-Sicht erfüllt.
        - Wenn „qlogin“ von SGE verwendet wird, funktioniert Accounting, aber kein ssh mit X-Windows. Aktuell werden ssh-Keys verwendet.
- Diskutieren der Laufzeit von Grid-Jobs mit ALICE Offline (vom 12.9.11: KS)
- Seminarraum im KBW-Gebäude muss für ½ Jahr im Vorraus gebucht werden (vom 07.11.11: KS)
    - Der Seminarraum im KBW-Gebäude KBW 5.032 ist für jeden ersten Montag im Monat gebucht worden. Also für den 6. Februar, den 5. März, den 2. April, den 7. Mai, den 4. Juni und den 2. Juli.
    - Dieses AI wurde erfolgreich bearbeitet und wird nun geschlossen.
- Mit FOPI muss diskutiert werden, ob Etch noch benötigt wird (vom 07.11.11: KS)
    - Stellt noch benötigte 32-bit-Software ein Problem dar ?
- Probleme von Fluka mit gFortran unter neueren Betriebsystemen müssen untersucht werden bevor Etch abgeschaltet wird (vom 05.12.11: DB, VP)
    - Noch läuft Fluka nicht. Fluka muss auf dem entsprechenden Build-Server installiert und anschließend getestet werden (VP, SIST)
    - Peter Malzacher merkt via E-Mail an, dass bisher Fluka immer ohne Schwerionentransport-Code getestet wurde. Das RQMD-Problem muss vom Fluka-Team gelöst werden. Bis es dafür eine Lösung gibt, muss SiSt weiter auf Etch rechnen. Es sollte heute geklärt werden, wie viele Ressourcen SiSt in nächster Zeit unter Etch benötigt. Mittelfristig hilft ev. Virtualisierung oder g77 auf moderneren Betriebssystemen.
    - Vermutlich wird nur ein Etch32-Build-Host benötigt und die hier kompilierten Programme sollten unter Lenny32 laufen. SiSt probiert, ob unter Etch kompilierte Software unter Lenny funktioniert. Hierfür wird libg2c benötigt und muss auf Lenny installiert werden.
    - Mittelfristig sollten virtuelle Etch32-Rechner auch über SCLAB2 angeboten werden können, aber frühestens in einem Monat.
    - FU sagt, dass gfortran 4.4 Minimum sei, Version 4.6 sei installiert, es kracht aber im HI-Code. Eventuell kann man g77 auf Squeeze installieren.

    - SiSt benötigt mehr als 80 Job-Slots für Fluka-Rechnungen. Es laufen noch 10 Etch-Rechner in der Batchfarm, die hierfür verwendet werden können.
    - Von Fluka werden Binaries verteilt. Der Source-Code ist nicht verügbar.
- CVMFS: die „forward Time“ für alice.gsi.de soll von 1 Stunde auf 0,5 Stunden gesenkt werden (vom 05.12.11: VP/JT)
    - Die Umstellung ist bereits erfolgt. JT muss testen.
- Um die interaktiven ALICE-Rechner lxir35-38 auf Squeeze umziehen zu können, soll JT testen, ob die unter Squeeze kompilierte ALICE-Software auf Lenny läuft (vom 16.01.12: JT)
    - Die interaktiven ALICE-Rechner sind auf Lenny umgezogen worden.
- Ein Rechner von MA soll ebenfalls auf Squeeze upgegraded werden (vom 16.01.12: HPC)
    - MA sollte ein entsprechendes Ticket aufmachen, indem die umzuziehenden Rechner namentlich aufgeführt sind.
- 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 weren muss (vom 16.01.12: KS)
- Eine Etch-Maschine für Backup muss für HADES gefunden werden (vom 16.01.12: IT)
    - Dieses AI kann geschlossen werden.
- Da HPC vorrangig nur noch 64-bit-Umgebungen unterstützen möchte, sollten alle Experimente klären, ob noch benötigte 32-bit-Software ebenfalls unter 64-bit läuft (vom 16.01.12: FOPI, Experimente)
    - ALICE hat keine Probleme hiermit
    - FU und MA haben Software, die nur unter 32-bit läuft
    - Es werden bis auf weiteres auch 32-Bit-Maschinen zum Bau von Software zur Verfügung gestellt werden.
- 30 Batchfarm-Rechner sind bereits auf Lenny umgezogen worden. Weitere 70 Rechner müssen noch umgezogen werden (vom 16.01.12: HPC)
    - Alle Rechner bis auf 10 sind bereits umgezogen worden.
    - Diese 10 Etch-Rechner sollen bis auf weiteres auch weiter laufen, um z.B. SiSt für Fluka-Rechnungen dienen zu können.
- Eine bessere Methode zur Ankündigung von z.B. Wartungsarbeiten an alle Batchfarm-Nutzer muss gefunden werden. Vermutlich wird eine neue Mailingsliste eingerichtet (vom 16.01.12: IT)
    - Eine ClusterInfo-Mailingliste wurde eingerichtet, auf die man sich einschreiben kann. Sublisten sind möglich. Eine Anleitung findet sich im Wiki. Eingerichtet hat dies Victor Penso.
- Lustre-Pfade erscheinen falsch bei Verwenden eines Perl-Moduls namens „fish“. (vom 06.02.12: JM, IT)
- CVMFS muss auf Squeeze-Desktops getestet werden (vom 06.02.12: JT, IT)
- SVN Clients auf Lenny und Squeeze sollten angeglichen werden und auf beiden Betriebssystemen in Version 1.6 zur Verfügung stehen (vom 06.02.12: JT, IT)
- „Modules“ sollte auf der „alten Batch-Farm“ unter Lenny zur Verfügung stehen (vom 06.02.12: JT, IT)

Geschlossene und 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.

Protokoll letzte Sitzung: allgemein akkzeptiert, sollte in Zukunft aber auf Englisch geschrieben werden. Die Sitzungen selbst können weiterhin auf Deutsch abgehalten werden.

TOP1: Bericht vom GSI-FAIR-Computing-Meeting (KS):

- Ein detaillierter Bericht vom letzten GSI-FAIR-Computing-Meeting am 17. Januar 2012 wurde von KS gegeben.
- Ein ausführliches Protokoll vom GSI-FAIR-Computing-Meeting wird demnächst an die entsprechende Verteilerliste geschickt werden.

TOP2: Status-Bericht Minicube (CH):

- Die Temperatur im Minicube stieg auf 70 Grad und die Feuerwehr wurde durch den Rauchmelder alarmiert, der u.A. auch die Raumtemperatur misst.
- Grund war die eingefrorene Nachspeiseleitung, weshalb die Kühlung nicht mehr funktioniert hat.
- Steckdosenleisten haben nicht abgeschaltet, die entsprechende Funktionalität wurde noch nicht konfiguriert.
- Die Rechner selbst haben auch nicht abgeschaltet. Die entsprechende Funktionalität wird erst bei der Installation des GSI-Linux aktiviert, die zur Installation nötige Infrastruktur stand im Minicube jedoch noch nicht zur Verfügung.
- Im Anschluß an das Meeting wird eine Führung durch die Testinghalle gegeben (TR).
- Notfallmaßnahmen wegen Überhitzung haben Vorrang vor der Inbetriebnahme des neuen Clusters.

TOP3: Status BatchFarm-Betrieb

- Die Stabilität des Testclusters (Farm B) ist nicht gut. Ein Betriebsausfall von 24 h ist nicht tolerierbar (JT).
- ALICE hat Lustre bis auf 76% aufgeräumt
- TR ist mit dem Dateisystem durch. Das Ergebnis (Lustre-Belegung) ist in Listen verfügbar.
- CH sieht auch Instabilitäten. So waren letzte Woche alle SNMP-Daemonen tot. Der Grund hierfür ist unklar.

TOP4: Pläne für die nächsten Wochen:

- HADES (JM):
  - Alles muss für die Strahlzeit bereit sein. Für die HADES DAQ-Server wird Unterstützung seitens der IT gewünscht (MT)
  - „Netzwerk"-Schluckauf wirkt störend. Ein ssh-Login ist manchmal schnell, manchmel weniger. Eventuell dauert es von Mac besonders lang. Von Windows aus geht es schneller. Ev. gibt es ein Problem mit Keys und NFS.
  - JT vermutet auch eher NFS-Probleme als Netzwerk.
  - CH sagt, dass es zu wenig Manpower in der Netzwerktruppe gibt. Gruppenserver für die Verteilung des Betriebssystems sind teilweise auch überaltert.
  - Die HADES-Strahlzeit beginnt am 2. April 2012.
  - Bänder wurden gekauft und sind bereit, Laufwerke ebenfalls. Mitte März wird zunächst mit Cosmics begonnen.
- ALICE
  - Die neue Datennahme beginnt im März.
- Minicube-Cluster
  - Das Routing zwischen Infiniband und Ethernet erfolgt via Linux-Gateways. Das alte Lustre und das neue Lustre in der Testinghalle werden getrennt sein. Überkreuz wird es eine maximale Bandbreite von 40 Gb zum Umkopieren geben.
- Theorie
  - Der Request für High-Mem-Maschinen wird auf das nächste Meeting verschoben.

TOP5: neue Nutzer für die Farm

- weitere Nutzer werden dringend benötigt. Die IT bietet bei Bedarf auch Schulungen an.

TOP6: nächster Termin

- das nächste Treffen findet am 05.03. 2012 statt. Ausnahmsweise wegen Raumänderung im Seminarraum 1 (Theorie) in KBW 2.27.