CY-01-KIMON operational issues

Speaker: Kyriacos Neocleous

High Performance Computing systems Lab,
University Of Cyprus

SA1 workshop: Stockholm, Sweden, June 2007

Presentation outline (1)

- Tools used in grid operations
- Missing features regarding tools

- Scheduled and unscheduled site interventions
- Plans for updates deployment for uninterrupted production service provision

- Communication issues
  - ROC, other sites, VOs
  - Users

Presentation outline (2)

- **Correlation of cross-site issues**
- **Percentage of real life problems detected and reported by the COD before we are aware of them**
- **Usefulness of the following operations bodies/meetings and suggestions to improve them:**
  - COD
  - ROC support team
  - Operations meeting

Tools for daily grid operations (1)

- **Monitoring / stats**
  - Standard tools
    - Site Availability Monitoring (SAM) reports
    - Grid Statistics (GStat)
    - GridICE
    - GridView monitor / statistics
  - Additional tools
    - MoniFarm
    - **FailRank (under development)**

- **Reporting and configuration (e.g. downtime)**
  - CIC pre-report
  - CIC broadcasts
  - GOCDB

Tools for daily grid operations (2)

- **Troubleshooting**
  - SAM admin's page
  - GGUS (global) and SEE (regional) ticketing systems
  - Unix Monitoring Tools (e.g. SysStat)
    - Bash and perl scripts
    - ssh passwordless logins
    - Node-specific commands (qstat, pbsnodes, etc)
    - Machine logs

Tools for daily grid operations (3)

- **Central Logging Server**
  - aids troubleshooting, searching across nodes easier
  - non-linux machine
  - remote root access has been blocked
  - log files are read-only for the normal user accounts
- **Gateway Machine**
  - aids remote administration for site operators
  - allows accessing the machines outside of the university campus
  - protected with one-time passwords for avoiding keylogger attacks
  - remote root access has been blocked
- **Node images taken on a regular basis**
  - part of disaster recovery plan
  - decreases time between failures
  - no disruption of service to create image
    - mondo used for on-the-fly backups

Missing features (1)

- **Logs**
  - Problems identified mostly through SSCs
  - Format needs to be standardised
  - Error messages are sometimes cryptic
  - UTC / local time issues (some logs are UTC, rest local times)
  - Association of related entries across all node types
    - Unique identifier (related to jobID)
  - Perhaps a search tool would be useful
  - Binary Format? (faster)

- **SAM Admin page**
  - sometimes the load is very high (due to its usefulness)
  - suggestion to replicate the service in other ROCs

Missing features (2)

- **Middleware installation could apply some settings by default**
  - Log rotation to meet minimum requirement (90 days) where needed
  - Block SSH backdoor access
- **Firewall rules**
  - an iptables template could be provided
  - especially for core services
  - admins should still be able to over-write or amend the rules
- **Other tools**
  - e.g. indicate differences across all nodes, wrt:
    - services running
    - configuration files (O/S and MW)

Site interventions

- **Some examples of scheduled interventions**
  - Software upgrades (most common)
  - Hardware upgrades (rare)
    - Adding RAM, CPUs
  - Machines physical relocation (rare)
  - Network (rare)
- **Some examples of unscheduled interventions**
  - Disks full (fixed by setting quotas) – quite common
  - BDII overloaded (fixed via indexing) – quite common
- **Uncommon (unscheduled) interventions**
  - Network disruptions
  - Hardware failures
  - Emergency power cuts
    - Due to fire (not in the cluster room!)
    - Urgent maintenance work
  - Unfavorable room conditions
    - A/C issues
    - Site core machines crashed due to high temperatures (HDD)

Uninterrupted production (1)

1. **Planning**
   - SA3 testbed collaboration
     - Discussing and avoiding known issues
   - Monitoring and discussing in mailing lists for known issues
     - SEE-ROC tech mailing list
     - LCG-Rollout list

2. **Safety measures**
   - Backups (images) of machines
     - Minimizing downtime by restoring previous installation as fast as possible
       - *Mondo used for taking images/snapshots without shutting down (file level)*
     - Future possibilities to test
       - *LVM*
       - *VMware*

Uninterrupted production (2)

3. **Execution**
   - Incremental upgrades scheme (usual method)
     - Start by less significant WN
     - Proceed to update all WNs if nothing broke
     - Continue with core services/nodes: CE, SE, MON, UI
     - Finish off with central services (RB, BDII, WMSLB)

4. **Testing**
   - SAM Admin's page
   - UI submissions with DTEAM and regional (SEE) VO

- ***Uninterrupted production possible?***
  - *Definition? (global or per-RC)*
  - *Scheduled/unscheduled downtimes..?*

Communication improvement

- **No need for improvements:**
  - Helpdesk (GGUS and SEE regional ticketing system)
    - Excellent for operational issues
    - Excellent for allowing requests by remote users
  - E-mail exchanges between ROC (tech list and country rep list)
  - E-mail communication for local and remote users
  - Phone and face-to-face meetings for local users
  - Skype and phone calls where necessary (proved useful during routing problems) for operational issues

Communication with users

- **Remote users**
  - helpdesk: most frequent method
  - e-mail: sometimes
- **Local users**
  - phone: most frequent method
  - face-to-face meetings: frequently
  - e-mail: often
  - helpdesk: never used
  - training events and workshops: frequently
  - “Grid Clinic” day establishment (open grid lab)
  - local mailing list (CyGrid users)
  - websites: EGEE local website, and CyGrid

Cross-site issues

- ***Dealing with cross-site issues***
  - Inter-ROC communication *via SEE lists* and the *regional ticketing system* is adequate
  - ... after ROCs participate in weekly ops meetings

COD indicating site problems

- **Most problems are discovered and corrected before COD ticket**
  - CY-01-KIMON had 12 operational tickets during EGEE-II
  - most issues detected by site admins through monitoring tools
    - SAM, GStat, local tools (SysStat, custom scripts)
    - Future: nagios, ganglia, FailRank (locally)
  - on some cases, issues have been reported by local users
    - most issues concerned UI, RB, WMSLB
  - estimates
    - 70% of problems detected by site admins
    - 20% of problems detected by COD
    - 10% of problems detected by users

- **CIC tickets are very useful overall**
  - especially links provided for helping to solve the issue at hand
  - personal experience from communicating with COD is very positive

Usefulness of ops bodies/meetings

- COD (CIC-Operator-on-Duty)
- Our ROC (SEE) support team
- Operations meeting

Questions?

Thank you for your attention!

Links

- **This presentation is available at**
  - http://cygrid.org.cy/docs/EGEE-II-SA1_workshop_Stockholm.pdf
- **SysStat**
  - http://perso.orange.fr/sebastien.godard/
- **MONDO**
  - http://www.mondorescue.org/
- **MoniFarm**
  - http://www.nikhef.nl/grid/sysutils/
- **Failure management in Grids**
  - CoreGRID tech report
  - http://grid.ucy.ac.cy/Papers/CoreGRID-TR0055.pdf
- **FailRank**
  - CoreGRID'07 paper