Site: Taiwan-LCG2 (ASGC Tier1)

Incident Date: 2010-Jan-18th

Severity: Severe

Service: All grid services, including MSS and batch pool.

Impacted: All VOs

Incident Summary:

Due to the flashover occur at local power station about 07.25 UTC on Monday 18 Jan; all computing services in data center are all affect by the power outage referring to from two main circuits connecting to Phys. Institute and also data center. Most of the grid services have been recovered within 30 minutes and critical database services extend for another two hours. And the temperature in data center reduces to normal mark at 4:17pm local time.

Type of Impact:

Incident duration: 2 day

Report date: 2010 Jan 1st,

Reported by: Jason Shih

Related URLs: https://gus.fzk.de/ws/ticket_info.php?ticket=54739 and https://gus.fzk.de/ws/ticket_info.php?ticket=54820

Incident details:

The grid services able to restore shortly, after the power cycle while we notice some server nodes and also newly deployed computing nodes didn't have netfs turn on by default. Due to this, the continue SAM job submission failures have been observed and last for 1.5Hr. Full system scanning help identifying the root cause and have restarted the service right after, the service have been enabled as part of default startup daemons.

The CASTOR service cease responding for the SRM transfer requests after confirming all services up and running. We further notice the abnormal functionality of resource manager, and two rmMaster daemons are running in the same time. The service return normal after cleaning up the message queue and restart again the scheduler and job manager. It took more than 6Hr to restore the SRM service while the direct RFIO wasn't affected after the Castor service restart.

Even though, the transfer efficiency (FTS and also RM probes carried out by experiment dashboard or SAM) have reduce to less than 65% sometime. This should have been clarified and fix the next day. We found the services are all falling back to identical instance due to the competition from all instances after cluster restore from power cycle. Have force migrating the services into correct cluster instance and restart again several daemons.

After restoring the newly setup DB cluster (CASTOR), the yum auto-update seems resetting the default kernel version before the incident. This we've notice extremely load on all the instances while the other DB clusters (grid services, and streaming) already have yum update turn off from the very beginning. We have force migrating the services running on particular instances and restart all the core services of CASTOR after confirming fully operate backend database service. Due to failure of load balancing and also the incorrect kernel release adopt after power cycle, we finally understood from cross checking the monitoring metrics and applicable to fixed it on Wednesday 20th

Two team tickets open during the reporting period are given in related URLs.