Report sent on Mai 3rd 2010 to: wlcg-scod@cern.ch

Type of Incident: Batch system outage
Location: IN2P3-CC
Duration: 17 hours
Date: April 24th 23:45 to April 25th 16:45
Author: Rolf Rumler

Description

The services location service (named Saphir) of the IN2P3-CC stopped responding to requests, blocking most batch system commands and various other services.

Timeline

- April 24th, about 23:30 : the engineer on duty made his last standard verifications for the day.
- April 24th, 23:47 : NAGIOS detected that both servers of the services location service, Saphir, became unresponsive. A lot of services including the batch system could not find their master servers any longer; as at least from time to time one of the Saphir servers responded, various services reworked intermittently, including batch.
- 1:30 : a user signals a problem with an Oracle database via the problem ticketing system.
- April 25th, 3:40 : NAGIOS signalled that IN2P3-CC-T1 appeared in FCR for CMS
- 4:00 : batch job submission stopped working definitely. Most other batch commands did no longer function neither. Queued jobs still got started.
- 9:00 : the engineer on duty detected the situation during his standard verifications.
- 9:09 : the main batch system developer (IN2P3-CC uses its own system, BQS) made a first analysis and put in place the controls to block job initiation. This normally would have been done by batch system administration commands but those were unavailable due to the type of incident.
- 9:10 : IN2P3-CC-T2 appears in the FCR for CMS
- 9:18 : an unscheduled downtime was declared with a start time of April 25th, 0:00 and an estimated end time of April 26th, 12:00.
- Between 9:22 and 9:45 : various attempts to connect to the failing servers which still responded to ping were unsuccessful, using accounts with different privilege levels.
- 9:50 : system administrators alerted by the engineer on duty start analysing the servers. They find that the saphir demons on both servers were squeezed out by processes of other services, Symod and ActiveMQ, inducing a very high swap rate. Restarting all services (not the machines) helped.
- Somewhere around 10:10, the failing Oracle database gets restarted (see another SIR, SIR-IN2P3-CC-OperationsPortal-2010-04-22v2, for the reasons of the Oracle problem).
- Between 10:10 and 12:30 various services resume. The batch system does not, as a consequence of a stopped daemon. Saphir is still failing intermittently but only for short periods of time.
- 12:30 : the batch system developer restarts the missing daemon and
- 12:40 : the system is opened for job initiation. No new grid jobs are accepted though.
- 12:45 : IN2P3-CC-T1 gets out of the FCR.
- 13:21 : IN2P3-CC-T2 gets out of the FCR.
- 15:15 : the processes of the Symod service are definitely stopped on the Saphir servers. ActiveMQ is left running despite of being suspected of memory leaks.

- After 30 minutes of Saphir running without any failure, the downtime is shortened and the site re-opened to new grid jobs.

Analysis

The Saphir service is addressed by all distributed services of the IN2P3-CC site to locate the various master servers, for example the batch system's master server which is needed to accept job queueing or job status commands. Saphir also controls authorisation levels for users, groups, and services. As a fundamental service of the site, it is itself doubled and uses a self contained, LDAP based database. Both used servers have only 2 GB of RAM because Saphir doesn't use much resources.

The Symod service running also on those machines files statistics data and centralises syslog and other messages from the machine park. It is using ActiveMQ for messaging and an Oracle database for storing the data. Since about April 20th it was found that the Oracle database sometimes accepted data at a slower rate than they arrived. During the night of April 24th to April 25th the arriving data apparently could no longer be stored on the disk cache and started to be kept in memory, which triggered the observed swapping. In addition, the log space of Symod / Oracle got full, probably also because of the otherwise unrelated Oracle problem reported in the already mentioned SIR.

The impossibility for the engineer on duty to login to the servers slowed down the process.

The fact that the failure of this essential service was not signalled to the engineer on duty by an automated SMS probably accounted for about half of the duration of the outage.

Once started the incident handling didn't show any significant problems.

Impact

No job submission was possible from the beginning of the failure of the Saphir service until just before 13:00. Running jobs generally were not impacted, except that the CE's job status requests failed. Job initiation stopped at about 10:00 and resumed at the same time as the job submission. Grid jobs were accepted until the declaration of the downtime but filed up in the queues because submission to the local batch system was impossible.

Corrective actions

The Saphir service will be run on dedicated machines.

A persistent outage of Saphir will be signalled by a SMS to the engineer on duty.