GridKa/KIT Service Incident Report 2011-01-29 (all times in UTC)

Type of Incident: **Batch system unavailable / degraded**
Location: GridKa/KIT
Duration: several days
Date: Friday, Jan 29, approx. 17:00 to February 02, evening
Authors: Andreas Heiss, Manfred Alef

Description:
After upgrading PBSPro to version 11.0, the server occasionally failed to start queued jobs. A workaround to restart the pbs_server process was working temporarily.
After several successful restarts, the server crashed immediately after each further restart and the system was unavailable.

Impact:
One of two GridKa subclusters was unavailable or degraded for several days.

Timeline:

- Wed, Jan 26: PBSPro was upgraded to v11.0.0 on the production system during the full-day maintenance window of GridKa. (Before this upgrade, v11.0 was running stable for 2 weeks in our test environment.)
- Fri, Jan 28, ~17:00: the PBS server stopped scheduling. The number of running jobs is decreasing.
- Sat, Jan 29: 02:41: the problem is detected by the monitoring system. The on-call service gets a notification.
- Sat, Jan 29: 04:00: first analysis of on-call engineer: error messages concerning the licence server found in the log file of the pbs_server.
- Sat, Jan 29: 09:00: local PBS expert cannot fix the problem. Vendor support is notified. Restarting the pbs process seems to help for a short time.
- Sun, Jan 30: 10:00: PBS restart does not help any more. The server process starts crashing hard after a short run time of few seconds without printing any helpful error message or log entry. The affected subcluster is unavailable. Vendor is informed about the new situation and the problem ist escalated to the highest priority. Log files, traces and core dumps are sent to vendor.
- Sun, Jan 30: 16:30: Vendor provides an extended license file for the second, not affected subcluster. Workernodes forseen for milestone April 1st, which were then running burn-in tests, are taken into production in the second subcluster. About 5000 job slots and more than 50% of HEPSPECs are available.
- Tue, Feb 1, 08:20: Vendor support suspects corrupted database to be the reason for the server crashes and suggests to reinstall the system.
- Tue, Feb 1, noon: system reinstalled and now seems to work.
- Tue, Feb 1, 22:36: The server again stops scheduling new jobs.
- Wed, Feb 2, 13:20: Vendor support engineer confirms license problem. The license problem seems to occur, when the number of used licenses exactly matches the number of job slots. The vendor is asked to provide a license file for more job slots as a temporary workaround.
- Web, Feb 2, 18:43: Vendor provides new license file for more job slots. After installing this license file, the system runs stable.

Analysis:
KIT was suffering from limited batch server performance since more than one year. The new version of PBSPro promised to solve this performance issues.
After successfully testing the new version 11.0 for two weeks in a test environment, we installed this version during our full-day maintenance window on Wednesday, January 26.
Two days later, the PBS server stopped scheduling jobs because of a meanwhile confirmed bug which was introduced with a new license server system in version 11.0.
(The vendor told us, that KIT was the first customer to run into this problem but other customers meanwhile got the same issue.)

Conclusion:
KIT hit a bug introduced in the latest version of PBSPro which uses a new license system.
The problem analysis at the weekend of January 29 and 30 focussed on problems known from earlier versions of PBSPro and was not successful. However, only one of two batch servers was affected and additional compute nodes were added to the second, working batch farm to attenuate the situation.
The vendor needed two work days to identify the problem. As a workaround, a modified license file was provided.
It was unfortune that the bug did not appear within the 2-week long testing period of the version 11.0 at KIT.