GridKa/KIT Service Incident Report 2010-07-10 (all times in CEST)

Type of Incident: dCache server down
Location: GridKa/KIT
Duration: 18:00 hours
Date: Monday, July 05, 20:00 to Tuesday, July 06, 14:00
Author: Xavier Mol

Description:

A server hosting central dCache services crashed because of a broken motherboard.

Impact

Complete failure of the dCache storage element for CMS.

Timeline

- July 05: 20:01: first alarm reached the on-call support
- July 05: 22:10: on-call support on site, host rebooted
- July 05: 22:37: the same machine crashed again (HW error)
- July 05: 23:00: the on-call support must wait for HW experts to exchange HW
- July 06: 08:15: beginning of the replacement of the broken machine
- July 06: 10:00: simple replacement fails due to different raid controllers/controller versions . Finding a similar machine takes some time.
- July 06: 14:00: complete dCache storage element for CMS back online and functional

Analysis

Concrete analysis of what is broken and why is taken up with the vendor. Especially the non-conclusive messages in the HW system log can be improved. The workflow of our on-call service was effective and flawless. Communication to the experiment people was prompt and detailed.
The reason why changing the machine took so long is that we were mistaken by the model of the machines. Both, the defective and the replacement, carry the same model name but differ in detail. Sorting this out and finding a workaround was time consuming.

Conclusion

CMS had bad luck that the hardware failure hit this one machine out of all others. We have no means of preventing such breakdowns and the on-call engineer cannot be instructed to handle such situations by himself.
Installing all possible kernel modules on servers could have avoided extra time needed to sort out the problem arising from the fact that different raid controllers were used by Dell in machines of the same model.
Remote management is on most of our fileservers, but in this case that would not have been of use anyway.
Since we are slowly moving towards more stable software we must reconsider the on-call procedures which are tuned for software problems. These issues will be replaced by HW issues in the near future (if indeed the software remains this stable over time).