From: Jeff Templon
To: Jamie Shiers;
Subject: Nikhef cooling incident post mortem.
Date: Wednesday, March 04, 2009 15:22:52

Around 13:30 on 3 March, the facilities people at Nikhef did a routine test of the emergency power system. This power system is backing the critical grid systems at Nikhef, as well as critical systems in the Amsterdam Internet Exchange.

The test is realistic : it is done by cutting the primary power to the chain, just as would happen in a real power failure. The way the power chain is configured, implies that the power is also removed from the cooling system (the part that produces the cold water). This is normally not a problem, as there is by design a large cold-water buffer that supplies cooling for some period of time.

During the test yesterday, a mixing valve got stuck in the cooling system, resulting in no new cold water injected into the cooler for our server room (most of the valves for other rooms worked correctly) during the test. The result was that in a period of a few minutes, the room temperature went from about 23 degrees to almost 32 degrees. Numerous machines raised temperature alarms. During the temperature excursion, we had no idea what the cause was, so we turned off the bulk of the worker nodes. As the temperature began to recover yesterday afternoon, the cause of the problem was still not clear (nor was there any temperature history available, so we could not be sure that things had returned to normal), we left the bulk of the worker nodes off for the night.

Today we are slowly turning machines back on, and monitoring the temperature as we do so.

Positive consequences of the incident are that we have improved documentation for the remote management interfaces on site (the remote management gurus were both absent during the event) and we plan to add temperature sensors and alarms to our Nagios/Ganglia system, as we discovered the temperature problem almost by accident yesterday.