

Active/Active Save #1: Coffee Pot Takes Down Node

November 2006

If it hadn't been running in an active/active configuration, this mission-critical system would have been taken down by, of all things, a coffee pot. As it turned out, only one of its nodes was sensitive to caffeine. The system survived because the users at the failed node were switched quickly to surviving nodes.

However, as most catastrophic faults go, there was a chain of events that led to this failure. Breaking any link in this chain would have prevented the failure.

Here is the story.

The Rolling Upgrade

After running its active/active network successfully for several years on its existing equipment, the company decided to upgrade to the next version of the system that it was using for its nodes. This was a major upgrade involving new hardware and a new operating system. The company had successfully applied rolling upgrades to its nodes in the past by taking down one node at a time, upgrading it, and reintroducing it into the system.

Event 1 – Not Enough Power Connectors

As best practices dictate, the system at each node was powered by a separate circuit protected by an uninterruptible power supply (UPS). When the new system was rolled in at one of the nodes, it was found that all of the UPS power connectors were being used. There was not one available for the new system. As a consequence, the new system could not be powered up.

So as not to delay the upgrade, the node was temporarily connected to the facility's unprotected power. Though this power source was not protected by a UPS, the plan was to correct this problem in short order by adding an additional connector to the UPS output.

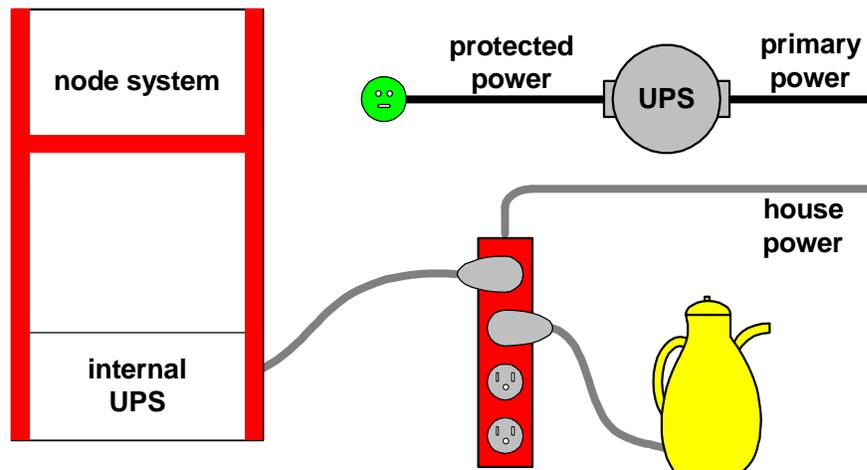
Event 2 – Forgotten Task

Evidently, no record was made of this issue on any task list. The required power connector change was forgotten; and the upgraded system continued to run successfully for quite a while on the unprotected power source.

Event 3 – The Coffee Pot

As time went on, the load on the unprotected circuits gradually increased as the company grew. More and more people meant more lighting, more heating, more air conditioning, and more workstations.

One fateful day, an employee performed a normal, everyday task. He or she plugged in the coffee pot to make fresh coffee. This was the straw that broke the camel's back. The coffee pot blew the circuit breaker, taking down everything that was on that circuit. This included dropping primary power to the upgraded system, which had never been moved to the UPS circuit.



The Coffee Connection

Event 4 – The System's Internal UPS

The node, however, kept on running for a while. It was supported by an internal UPS that kept it operating long enough to save its state and to shut it down gracefully following a primary power failure.

Fast action on the part of the staff at the site restored the primary power in just 35 seconds – an admirable feat. Unfortunately, the system's internal UPS only lasted for 30 seconds. The node shut down and suffered a 30-minute outage until it was brought back online.

The Active/Active Save

If the system had not been an active/active system, its users would have been denied service for a half hour. However, as it turned out, the users assigned to the failed node were quickly switched over to surviving nodes and suffered no apparent outage.

Once the node was brought back up, its users were switched back to it; and the active/active system was returned to normal service.

Lessons Learned

As is the case with most failures of fault-tolerant systems, a chain of errors led to this failure. If any link in this chain had been broken, the failure would not have occurred.

Proper installation planning would have ensured that the correct power connector was available. Given this error, the installation could have been delayed until the proper connector was installed. Given that this was not done, an effective task list of "to-dos" would have ensured that the situation would not have been forgotten and that the connector would have been installed shortly after the upgrade.

Every one of these problems was a link in the error chain. With these links intact, the rest of the story was inevitable. No one can fault the person who plugged in the coffee pot, nor can the system be faulted because it had only a 30-second internal UPS. The stage was set for disaster.

Only the company's active/active architecture saved it from a perhaps costly and embarrassing total system failure.