

Is Preventive Maintenance Preventive?

October 2012

The Perils of Maintenance

Our Never Again stories are ripe with major IT failures caused by poorly performed maintenance. WestHost is a major web-hosting and domain-registration company. Its IT systems are housed in a certified SAS 70 Type II data center. Yet it was brought down by the failure of a technician to disable a fire suppression system actuator during a routine preventive maintenance test.¹ The piercing noise from the sirens damaged many disks, including some backup disks that were in the same facility. It took up to six days to restore all services, and some data was irretrievably lost.

An Availability Zone in Amazon's U.S. East Region's Elastic Compute Cloud (EC2) was taken down for four days when maintenance to increase the capacity of the Availability Zone's primary internal network capacity went terribly awry.² Rather than rerouting traffic to an alternate high-capacity primary network, traffic was rerouted to a slower secondary network. This effectively isolated the processing nodes in the Zone from their data nodes and led to a "remirroring storm" during which the processing nodes attempted to find alternate data nodes on which to back up their data.

BlackBerry subscribers lost communications for a day when their systems were upgraded to "improve cache performance" with no fall back plan.³ A major European company mistakenly took down both nodes of an active/active system while attempting to upgrade the capacity of one of the nodes.⁴ The \$38 billion Alaska Permanent Fund database was wiped out by a maintenance technician working on the disk system.⁵ Several IT systems of the State of Virginia were down for seven days, and four days of data were lost following a maintenance error by a service technician.⁶ The Three-Mile Island nuclear meltdown in 1979 was caused by the disabling of a secondary feed water pump by manual valves that were erroneously closed during preventive maintenance. And the list goes on and on.

¹ Fire Suppression Suppresses WestHost for Days, *Availability Digest*, May 2010.

http://www.availabilitydigest.com/public_articles/0505/westhost.pdf

Fire Suppressants Impact on Hard Disks, *Availability Digest*, February 2011.

http://www.availabilitydigest.com/public_articles/0602/inergen_noise.pdf

² Amazon's Cloud Downed by Fat Finger, *Availability Digest*, May 2011.

http://www.availabilitydigest.com/public_articles/0605/amazon_ebs.pdf

³ BlackBerry Gets Juiced, *Availability Digest*, May 2007.

<http://www.availabilitydigest.com/private/0205/blackberry.pdf>

⁴ Console Command Takes Down Active/Active System, *Availability Digest*, December 2006.

http://www.availabilitydigest.com/private/0103/console_command_takes_a-a_down.pdf

⁵ The Alaska Permanent Fund and the \$38 Billion Keystroke, *Availability Digest*, April 2007.

http://www.availabilitydigest.com/private/0103/console_command_takes_a-a_down.pdf

⁶ The State of Virginia – Down for Days, *Availability Digest*, October 2010.

http://www.availabilitydigest.com/public_articles/0510/virginia.pdf

What About Preventive Maintenance?

It is estimated that humans have contributed in some way or another to about 70% of all major IT outages. Maintenance functions comprise a large number of these incidents. Maintenance for repair and upgrade purposes must be done. But what about preventive maintenance? Preventive maintenance is optional, though generally considered a best practice. However, WestHost went down in a spectacular way because of a preventive maintenance error. Was preventive maintenance worth it to WestHost?

The value of preventive maintenance received attention at the 2011 Fall Conference of the 24x7 Exchange held in Phoenix, Arizona, U.S.A. The 7x24 Exchange is a forum for those who design, build, use, and maintain mission-critical enterprise information infrastructures. In his keynote address, Steve Fairfax, the president of MTechnology, addressed this issue. MTechnology provides risk-assessment services for highly available electrical power suitable for IT facilities.

In his address,⁷ Fairfax argued that a vigorous maintenance schedule can actually make a facility less reliable. "There's this mantra that more maintenance equals more reliability," he submitted. "We get the perception that lots of testing improves component reliability. It does not. The most common threat to reliability is excessive maintenance. The purpose of maintenance should be to find defects and remove them. And whenever a piece of equipment is undergoing maintenance, your data center is less reliable."

His views were echoed by other speakers. The Executive Director of Uptime Institute, Pitt Turner, said that he was in "violent agreement" with Fairfax on the risks of excessive maintenance. "Think about what preventive maintenance programs are appropriate," he said. "Do them and do them well using best practices to lower your risk." The Uptime Institute offers education, consulting, and data center tier-certification services

Others forcibly disagreed. In a comment posted to a summary of Fairfax's speech, Henry Hu of Emerson Network Power, a company providing mission-critical power and cooling solutions, referenced a detailed study done by his company⁸ on the effectiveness of preventive maintenance to maximize the availability of UPS systems and the battery systems upon which they depend. The study covered 185 million operating hours for more than 5,000 UPS units and more than 450 million operating hours for 24,000 battery installations.

The company's research found that the MTBF (mean time between failures) for units that received two preventive maintenance service events a year was 23 times greater than those that received no maintenance, provided that the maintenance was performed by a factory-trained service engineer. It found that a maintenance error by one of their service engineers caused an outage in only one out of 5,000 service events.

Emerson's analysis did show that there was a limit to the number of service events per year after which the availability of the systems decreased. If factory-trained service technicians were not used, this threshold was much lower.

The Most Critical Preventive Maintenance Testing - Failover

One very important preventive maintenance function for IT systems is the testing of failover to a disaster recovery (DR) system. This is a very complex and risky task that is often not performed completely or not at all. The result is that when a production system fails, there is a significant likelihood that the DR system will not come up. In our Never Again stories, this has been experienced by Google, American Eagle, BlackBerry, and others.

⁷ Is Maintenance Making Your Facility Less Reliable?, *Data Center Knowledge*; November 17, 2011.

⁸ The Effect of Regular, Skilled Preventive Maintenance and Remote Monitoring on Critical Power System Reliability, *Emerson Network Power White Paper*, March, 2011.

The best way to test failover is to failover and stay. Reverse the roles of the production system and DR systems periodically and stay that way until the next test. The Mayo clinic follows this procedure, switching the production/DR roles every quarter.⁹ As a result, it has a high degree of confidence that it will fail over properly should its production system fail.

Summary

Corrective maintenance is mandatory. Preventive maintenance is optional. All maintenance can cause outages. It is estimated that 70% of all IT outages have been aggravated by human actions. Maintenance errors dominate these. Preventive maintenance procedures must be carefully controlled so that an action that is supposed to improve data center availability does not instead take it down.

We invest a lot of money into our data centers to make them redundant so that any single failure can be tolerated. Why not do the same for the 70% problem - humans? If there is a critical operation that must be performed for maintenance, whether corrective or preventive, provide human redundancy. Use two people, one to define the action to be taken and one to confirm it.

So far as preventive maintenance is concerned, it seems that the consensus is that a little goes a long way. Just don't overdo it.

⁹ [Tackling Switchover Times](http://www.availabilitydigest.com/public_articles/0101/tackling_switchover_times.pdf), *Availability Digest*, October 2006.
http://www.availabilitydigest.com/public_articles/0101/tackling_switchover_times.pdf