Incident Report - June 6, 2016

Incident #2016-153

New Medical Building Outage

Summary

A network switch (device NEWMED-023-D01) was part of a scheduled maintenance activity (iTrack #8675), including the refresh of the operating code on approximately 100 devices. The update was performed but upon reboot it was found unresponsive.

Some confusion regarding building access caused a delay in responding to the issue. After on-site examination, the issue was escalated to Level 2 Networks with a preliminary diagnosis of hardware fault. Partial service was restored (at approximately 9:04 am), and was then escalated to Level 3 Networks (9:10 am).

Networks Level 3 found the device deployed as a switch "stack" (atypical for locations outside the datacenter environment).

The stack master was active; however, the stack subordinate member was not active, causing the issue.

Impact

Reporting users: Andrew DosSantos, Matt Simpson, Jason Palmer, Svetlana Rytchkova.

Unscheduled outage from 7:31 am through 9:03 am, partial service only 9:04 am through 10: 46 am, full service restored at 10:47 am.

Outage length was intensified due to the inordinate time required at reboot for FPGA to apply 9 years of microcode patches, common to the 3750x family of hardware (regular o/s version reboot on this platform takes approximately five minutes or less. The reboot including required run-once FPGA microcode took approximately 35 minutes.

Root Cause

Services were harshly impacted rather than simply degraded or impaired.

Without being fully attached across the master and its member, a fault event could not be sustained without loss of service.

Networks Level 3 observed a defect in that the stack master on its original 9-year-old code that had attempted but did not successfully complete auto-replication of the new (refresh) code to its member.

Resolution

Standard methods to manually replicate were unsuccessful.

Once a suitable method to force replication was devised and applied, the member was rebooted and once online, transitioned to an active state successfully.

Communications (Internal)

Email, cellphone, SMS and IM amongst ITS personnel.

ITSPP Communications (External)

A notice was posted through the ITS Notification Tool, emails and phone calls to reporting clients in order to provide updates regarding progress, as well as to obtain confirmation at resolution.

Lessons Learned

Campus areas with restricted physical access are also effective in restricting physical access to ITS equipment by ITS service personnel. Review of procedures to adjust for these areas is needed.

Services to which other than best-effort SLA are associated with should be provisioned in a fashion such that the impact of failure events is one of degradation or impairment rather than loss, especially for public-facing services (e.g., websites), and resiliency should be declared mandatory.

Stacked devices should contain a suitable name post-fix in order to serve as a flag for any special operating requirements, *especially* where they are deployed irregularly outside the datacenter environment.

Reboot times at refresh may be reduced to more acceptable levels by keeping interim updates more current.