Root Cause Analysis

QS Outage

The following is a detailed accounting of the service impairment that Agile Central users experienced on August 15, 2016 at 11:55am.

Root Cause Analysis Summary

Event Details (AM issue):

Event Date	8/15/16
Event Start	11:55am
Downtime Start	Intermittent, degraded over time
Time Detected	11:55am
Time Resolved	14:53pm
Downtime End Time	14:53pm
Event End Time	14:53pm
Root Cause	After analysis, we have identified an issue that caused our primary and secondary active directory controllers, which also provide internal DNS, to go into a reboot cycle. This rebooting caused intermittent internal DNS issues which prevented our internal services and SSO users from being able to communicate properly in the data center.
Customer Impact	Users intermittently unable to login to the application
Duration	• 11:42 - 11:43 **1 minute** • 12:02 - 12:03 **1 minute** • 12:09 - 12:12 **3 minutes** • 12:17 - 12:18 **1 minute** • 12:24 - 12:31 **7 minutes** • 12:33 - 12:37 **4 minutes** • 12:44 - 12:45 **1 minute** • 12:46 - 12:47 **1 minute** • 12:56 - 12:58 **2 minutes** • 13:00 - 13:17 **17 minutes** • 13:21 - 13:30 **9 minutes** • 13:39 - 13:50 **11 minutes** • 14:00 - 14:11 **11 minutes** • 14:21 - 14:23 **1 minute** **Total of outages**: 70 minutes

Future Preventative Measures

Actions that should be taken to prevent this event in the future.

Actions	Description
Add redundancy and alerting	To prevent future occurrences of this scenario, we are implementing work to create an extra layer of redundancy for our internal DNS along with clearer and quicker ways to detect and remediate the issue which occurred.
Add additional monitors for services to Pingdom Status Page	Could add a monitor that logs into Agile Central using SSO, which would test ALM and SSO.
Add non-Windows nameservers to mix in prod	Currently if we have a cascading domain controller failure our name service goes with it. Adding a secondary Linux nameserver is easy and would help maintain DNS during a Windows outage.
Fix resolv.conf in qd to not point at QS	Chef / puppet story to fix resolv.conf across DCs
Fix Nagios to withhold CHECK_NRPE alerts when nameservice is lost.	The on-call team was flooded and distracted with bogus CHECK_NRPE alerts from Nagios when Nagios could no longer resolve the hosts it was monitoring. Nagios should suppress these alerts during a loss of nameservice. Alternative is to just put VO in maintenance mode sooner
Update Ops Docs	Update Ops Docs about how to shutdown tunnel between DCs