Internet2 Post-Incident Report

Advanced Layer2 Service Maintenance Window Overshoots

Date: September 20, 2013

Overview

Several Node Insertion Maintenances overshot their allotted window due to complications. These occurred on July 3rd, 8th,13th,14th and August 30. These 5 maintenances the following issues were encountered:

- 4 out of the 5 hit a software bug related to node insertion
- 2 of of the 5 were impacted by poorly understood process for running OESS over hybrid mode interfaces.
- 2 of the 5 where impacted by poorly understood processes for setting and changing DPIDs for Juniper devices and within OESS.
- 1 of the 5 was impacted by faulty logging

Another way of looking at this is that 4 maintenances were impacted by software flaw, 4 were hit by confusion about how to perform 2 procedures and 1 was hit by a system issue.

Recommended Actions

Insert Node Bug

We need to introduce latencies in the automated testing that are similar to what we see in the field. In the specific case, we need to modify our approach to interacting with the Glimmerglass such that rather than directly going from the old to the new topology, it transitions to an intermediary topology where the path under maintenance is down for at least 15 seconds. When testing procedures by hand we need to also ensure we do not use the Glimmerglass without including such delay. Additionally we need to ensure we have good test coverage and aren't trying a poorly tested feature in production.

Estimated Completion Date: 9/22/13	Status: COMPLETE
Completion Notes: Testing is still conducted by hand for this, the test procedure was augmented to include a min 30 second delay when ink was disconnected. 30 seconds used as it was 2x the min time to detect this issue	

Hybrid Mode

We need to test any new configurations in the lab prior to deploying and ensure operations procedures for hybrid mode are documented and accessible and known to neteng. Procedures should be documented as we perform them in the lab. Any new process should be reviewed by neteng, systems and software prior to performing in production.

Estimated Completion Date: 9/22/13	Status: **COMPLETE**
Completion Notes: We now test significant config mods or version changes with system together in lab before in the field. Proceedures documented in MOPs and the MOP is lab tested.	

DPID Change

We need to test any new configuration in the lab prior to deploying in the field and ensure all procedures are documented and accessible. Procedures should be documented as we perform them in the lab. Any new process should be reviewed by neteng, systems and software prior to performing in production.

Estimated Completion Date: 9/22/13	Status: **COMPLETE**
Completion Notes: We now test new vendor revs to ensure DPID is stable from primary to backup RE.	

Logging Issue

Additional monitoring check should be added to monitor the status of the log files on the controller.

Estimated Completion Date: 9/22/13	Status: **COMPLETE**
Systems verified disk monitoring was in place and that the log rotation config was correct.	

Non-Recommended Actions

All recommended actions from the Internet2 NOC Post-Mortem are approved.

Appendix A: Vendor/Partner Post Mortem Details

The following text was authored by Internet2's vendor or partner responsible for the initial post mortem. It is provided in its entirety except in cases where it would expose sensitive information.

Post-Incident Report

Multiple AL2S Maintenance Windows exceeded Post-Mortem
Date: Fri September 20, 2013

Event Summary

Several Node Insertion Maintenances overshot their allotted window due to complications. These occurred on July 3rd, 8th,13th,14th and August 30.
These 5 maintenances the following issues were encountered:

- 4 out of the 5 hit a software bug related to node insertion
- 2 of of the 5 were impacted by poorly understood process for running OESS over hybrid mode interfaces.
- 2 of the 5 where impacted by poorly understood processes for setting and changing DPIDs for Juniper devices and within OESS.
- 1 of the 5 was impacted by faulty logging

Another way of looking at this is that 4 maintenances were impacted by software flaw, 4 were hit by confusion about how to perform 2 procedures and 1 was hit by a system issue.

Post-Mortem Trigger

Grover Browning requested post-mortem

Related Tickets or Network Owner Feedback

Systems ISSUE=7000

Parties involved

GRNOC Internal, Internet2 NOC

Impact

VLANs without a backup path would likely have been impaired until maintenance completed. VLANs with a backup path would likely have not been impaired by the maintenance. In all cases only VLANs traversing the link being split would have been impact. AL2S continued forwarding traffic on non-disrupted paths. The OESS UI was available and all services were available during the maintenance. Not all of the maintenances exceeded their window.

Root Cause(s)

Insert Node Bug

An undetected timing / logic flaw exists in OESS 1.0.12 that prevented the Insert Node in Path feature from working correctly. This was not detected in testing because the test harness was testing in a much smaller time scale that typically seen when doing a manual insertion. For bug to present, a delay of more than 15 seconds with the path down is required.

Hybrid Mode confusion

While we had tested hybrid mode in the lab and the results were documented in our testing results, this information was not located in operational documentation that was readily available to network and system engineers.

DPID Change confusion

While a document did exist that described how to handle a DPID change in OESS. The core issue was not anticipating that the DPID would change when going from the primary to backup RE based on the configuration.

Logging Issue

Logging of application logs failed due to a log rotation configuration error on the controller system.

Repair & Service Restoration action

Insert Node Bug

To address the Insert Node issue, the Software team manually set the link status to 'down' in the MySQL Database which then allowed the feature flawed feature to work correctly.

Hybrid Mode

Hybride mode fix involved adjusting the setting in the admin section of OESS.

DPID Change

DPID change required systems to manually run a query against the OESS database and update the record for the switch in question.

Logging Issue

To address the logging issue, systems restarted the apache process.

Prevention Recommendations

In general prevention relies and improved software testing, better training through practicing a procedure in the lab prior to field deployment and ensuring we have documentation to support the person performing the live maintenance. Additionally for risky / novel or relatively new maintenance procedures, we should consider having both a systems and software person on standby / on line to provide instantaneous assistance should problems arise.

Insert Node Bug

We need to introduce latencies in the automated testing that are similar to what we see in the field. In the specific case, we need to modify our approach to interacting with the Glimmerglass such that rather than directly going from the old to the new topology, it transitions to an intermediary topology where the path under maintenance is down for at least 15 seconds. When testing procedures by hand we need to also ensure we do not use the Glimmerglass without including such delay. Additionally we need to ensure we have good test coverage and aren't trying a poorly tested feature in production.

Hybrid Mode:

We need to test any new configurations in the lab prior to deploying and ensure operations procedures for hybrid mode are documented and accessible and known to neteng. Procedures should be documented as we perform them in the lab. Any new process should be reviewed by neteng, systems and software prior to performing in production.

DPID Change:

We need to test any new configuration in the lab prior to deploying in the field and ensure all procedures are documented and accessible. Procedures should be documented as we perform them in the lab. Any new process should be reviewed by neteng, systems and software prior to performing in production.

Logging Issue

Additional monitoring check should be added to monitor the status of the log files on the controller.

Impact Reduction Recommendations

Until we get a new version of OESS installed, systems needs to be trained on how to manually fix the database should we find ourselves in this state again. Additionally neteng should be trained to detect this particular bug, and future maintenances should consider having systems and software online assistance, if that is not feasible.

Timeline

All times UTC.

July 8th Maintenance Event (only major events are included)

This was impacted by the Node Insertion bug and confusion related to using OESS with hybrid interfaces.

JUL 3, 13:17 - Tom contacted AJ regarding upcoming AL2S maintenance for Node Insertion on Jul 8, requested recommended version, hybrid/non-hybrid and node insertion procedure. AJ responded with latest, doesn't matter on hybrid/non-hybrid but to let him know and the following procedure;

- Enable the interfaces (already done since the Juniper's were in static insertion)
- Don't approve the node
- Decommission existing link
- Acknowledge existing circuits and migrate

JUL 8, 01:05 - Tom initiated node insertion procedure on BOST
JUL 8, 01:27 - Repeated attempts failed, Tom recommended NOC contact CJ (Systems Oncall) and Grant or AJ (OESS Oncall)
JUL 8, 01:37 - CJ and AJ present in chat, explained actions taken and current status
JUL 8, 01:38 - AJ recommended adding discovery vlan configuration change to OESS

JUL 8, 01:45 - Tom recommended using vlan 4094 for discovery (LLDP), CJ implemented
JUL 8, 02:02 - Discovery not working, AJ recommended using vlan 100, CJ implemented, proceeded with node insertion
JUL 8, 02:05 - AJ identified JavaScript error in OESS, hot fix applied, suggested a previous maintenance may have broken JavaScript
JUL 8, 02:14 - Tom proceeded with insertion, appears to have worked, AJ said only 'halfway'
JUL 8, 02:18 - NOC notified Tom of alarms on OARnet, discovered that starting the process with inserting BOST, OESS didn't fail over.
JUL 8, 02:25 - Tom recommended restoring BOST to static mode, AJ pushed forward to fixing the 'halfway' insertion
JUL 8, 02:44 - AJ requested CJ restart OESS due to 'dbus craziness'
JUL 8, 02:50 - AJ recommend undoing and redoing the BOST node insertion
JUL 8, 02:59 - AJ finished the BOST insertion
JUL 8, 03:15 - No related Alerts, proceed with cleanup and discussion, AJ to document software bugs and diagnose insertion process
ALBA and PHOE rescheduled for July 13th

July 13th Maintenance Event (only major events are included)

This was impacted by the Node Insertion bug.

JUL 13, 01:05 - Tom, Jason (Systems) and AJ (OESS Oncall) present
JUL 13, 01:06 - Tom proceeded with PHOE insertion, starting with rebooting and then openflow configuration
JUL 13, 01:18 - Tom reported procedure failed, cannot decommission the ELPA-LOSA link
JUL 13, 01:21 - AJ reported the link had an 'unknown' status
JUL 13, 01:23 - AJ reported the insertion process was completed for PHOE
JUL 13, 01:26 - Tom proceeded with ALBA insertion, starting with rebooting and then openflow configuration
JUL 13, 01:40 - Tom reported procedure failed, cannot decommission the BOST-CLEV link
JUL 13, 01:42 - AJ recommended trying again, Tom finished the procedure
JUL 13, 01:44 - No related Alerts, proceed with cleanup and discussion, AJ to document software bugs

July 14th Maintenance Event (only major events are included)

This was impacted by a lack of understanding of how to use a new Juniper feature related to selecting the DPID to use on the switch, it was also impacted by a lack of understanding of how to change a DPID within OESS.

JUL 14, 01:01 - Tom rebooted ASHB
JUL 14, 01:04 - Jason (Systems) joined
JUL 14, 01:25 - ASBH back online, no flows from OESS, OESS identified ASHB as a new unapproved node
JUL 14, 01:37 - AJ joined, identified the DPID changed
JUL 14, 01:44 - AJ advised Jason of procedure to update DPID and advised ASHB is back online, Jason restarted OESS
JUL 14, 01:47 - Flows restored to ASHB
JUL 14, 01:49 - Tom rebooted RALE
JUL 14, 01:54 - Tom identified flow in Junos upgrade procedure in not causing port down events
JUL 14, 01:57 - Tom noticed Amazon via ASHB didn't restore
JUL 14, 02:32 - RALE back online, no flows from OESS, OESS identified RALE as new unapproved node
JUL 14, 02:37 - AJ/Jason fixed DPID issue, restarted OESS
JUL 14, 02:38 - Flows restored to RALE
JUL 14, 02:45 - Tom rebooted SEAT
JUL 14, 02:52 - Diagnosing Amazon in ASHB, AJ noticed Tom used hybrid port configuration, Amazon uses untagged interfaces
JUL 14, 02:55 - NOC extended SEAT maintenance by 1 hour
JUL 14, 02:56 - Amazon restored, configuration for untagged ports differs from hybrid on Junos
JUL 14, 03:04 - SEAT back online, no flows from OESS, OESS identified SEAT as a new unapproved node
JUL 14, 03:13 - AJ/Jason fixed DPID issue, restarted OESS
JUL 14, 03:17 - No related Alerts, proceed with cleanup and discussion

July 17th Maintenance Event (only major events are included)

This was impacted by Node Insertion bug, and confusion related to using OESS with hybrid interfaces.

JUL 17, 01:00 - Tom reloaded NEWY
JUL 17, 01:03 - NEWY restored
JUL 17, 01:10 - Tom reloaded WASH
JUL 17, 01:13 - WASH restored
JUL 17, 01:19 - Identified problems with circuits via WASH, WASH isn't receiving flows

JUL 17, 01:35 - Jason (Systems) and AJ (OESS Oncall) joined
JUL 17, 01:47 - AJ recommended Jason restart OESS
JUL 17, 01:52 - Priority to restore Amazon, then diagnose
JUL 17, 01:53 - Tom/Jason/AJ - discussion/diagnose
JUL 17, 02:22 - AJ observed that the node insertion process added a backup path to circuits that shouldn't have a backup path
JUL 17, 02:45 - Tom reloaded JACK
JUL 17, 02:48 - JACK restored
JUL 17, 02:50 - Tom reloaded CLEV
JUL 17, 02:53 - CLEV restored
JUL 17, 03:00 - AJ noticed that LOSA-PHOE and RALE-WASH links were in unknown state
JUL 17, 03:35 - AJ discovered that the discovery vlan 100 rules weren't matching causing a lack of discovery
JUL 17, 03:40 - Jason changed OESS discovery configuration to use 4094 instead of vlan 100, restarted OESS
JUL 17, 03:45 - AJ is fixing the data for circuits with incorrect backup paths
JUL 17, 03:46 - Tom reloaded ATLA
JUL 17, 03:48 - ATLA restored
JUL 17, 04:09 - AJ submitted a software ticket for OESS
JUL 17, 04:14 - AJ and Jason departed
(proceeded with other nodes)
JUL 17, 04:53 - rtr.chic RE crashed (not related to maintenance, distraction, delay)
JUL 17, 05:20 - Tom reloaded CHIC
JUL 17, 05:25 - CHIC restored
JUL 17, 05:36 - No related alerts, proceed with cleanup

August 20th Maintenance Event (only major events are included)

This was impacted by the Node Insertion bug. Troubleshooting was slowed by faulty logging.

08/30/2013 00:20: AJ Checks in with Tom and crew doing the maintenance. An optical issue is preventing the node insertion process.
08/30/2013 00:50: Andrew Lee completes the optical maintenance; both links to PITT are now up.
08/30/2013 00:50: Tom Johnson attempts to decommission the old link (ASHB-CLEV) and perform the node insertion process. However it fails.
08/30/2013 00:58: AJ notices that /var/log/httpd/ssl_error_log is not logging any messages, however other apache logs appear to be working properly. Asks the SD to contact the systems oncall.
08/30/2013 01:00: Chad joins the chatroom
08/30/2013 01:15: Chad restart apache and logging is fixed

08/30/2013 01:30: AJ determines that the link is still up in the database, and sets to down. Informs Tom that node insertion will work properly.
08/30/2013 01:32: Tom Johnson performs the node insertion, and is successful.

Document Revision History

Date	Change
1/30/14	Initial document creation
2/12/14	Added status of approved actions