Internet2 Post-Incident Report

February 2014 Houston AL2S Packet Loss Incident

Date: 12 February 2014

Overview

The Internet2 Network experienced a degradation of circuits passing through the AL2S Core Node in HOUH (AL2S sdn-sw.houh) affecting all layers and LEARN.

Recommended Actions

Better understand AL2S VLANs affected by outages

During this incident, getting a list of affected VLANs/customers was difficult. There is a complex manual process today to show VLANs that didn't failover when the backbone interfaces were shut down. We should ensure we have this information at our fingertips to understand the scale of service impact better.

Estimated Completion Date: pending	Status: **Pending**
Completion Notes:	

Limit amount of time allowed for vendor diagnostics, improve Service restoration guidelines

Multiple steps were taken to mitigate the service impact of the incident along the way. However, too much time was given to Brocade to run diagnostics before making the decision to reboot the switch. We should put additional guidelines for engineers in place so they will know how long a service can be impacted before escalating to more disruptive service restoration steps. It's been proposed that we target 1 hour for full service restoration, but this should be discussed with Internet2 service owners and Interne2 NOC to determine the right guidelines.

Estimated Completion Date: 2/20/14	Status: **Completed**
Completion Notes: Guidelines posted on internal NOC pages. 1 hour for restoration as a guiding principal unless overruled by Internet2 senior staff.	

Improve speed of vendor-provided diagnostic scripts

The initial data collection script Brocade has provided the Internet2 NOC took 50 minutes to complete its data collection. If we are to move to a new guideline restore

service in an hour, this should be improved. It's unreasonable to have a diagnostic script that takes this long.

Estimated Completion Date: 3/15/14	Status: **In Progress**
Completion Notes: New set of scripts was delivered by Brocade in late February and tested against the Brocade lab equipment. They don't show any material improvement in speed unless they are run local to the Brocade switch. Internet2 systems is working to verify co-existence of the troubleshooting scripts on the local performance assurance nodes to ensure that they don't disrupt the other functions of the local server.	

Improvements in software/network engineer involvement

While both software and network engineers engaged quickly in this particular case, the current SOP doesn't ensure this, because issues are passed to one or the other according to specific criteria. The Internet2 NOC should improve this, so that the first action of a software or network engineer is to engage the other group unless an issue is clearly one or the other. This will lead to faster and more frequent joint troubleshooting.

Estimated Completion Date: Pending	Status: **Pending**
Completion Notes:	

Monitor Fabric module outages

Before this incident, there were no alarms for syslog events related to switch fabric module outages. This was added already, and completed on 2/6.

Estimated Completion Date: 2/6	Status: **Completed**
Completion Notes: Alarming was added to Internet2 NOC systems on 2/6 and will continue moving forward.	

Non-Recommended Actions

All items in 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

Internet2 AL2S 1316

Date: 2/5/2014

Event Summary

Degradation of circuits passing through Core Node HOUH (AL2S sdn-sw.houh) affecting all layers and LEARN.

Postmortem Trigger

Internet2 Director of Operations and Engineering requested

Related Tickets or Network Owner Feedback

I2 IP 22312 Degraded Condition Resolved - I2 IP Backbone HOUS-KANS

I2 IP 22318 Outage Resolved - I2 IP Connector LEARN

Systems 22317 Request - I2 add a regular expression match for 'Fabric Monitoring' to GNARL

I2 TR-CPS 5002 Outage Resolved - I2 TR-CPS Various Customers

I2 TR-CPS 5007 Outage Resolved - I2 TR-CPS Customer LEARN

I2 Optical 2894 Brief Outage Resolved - I2 Optical Member ESnet Circuit ESNET-HOUS-KANS-100GE-07831

I2 Optical 2895 Outage Resolved - I2 Optical Various Circuits (HOUH/HOUS)

OneNet Ticket 6236 Stability - OneNet Peers Internet2 AL2S and TR-CPS

Parties involved

Internet2 NOC Network Engineering
Brocade TAC
Internet2 NOC Software Engineering
Internet2 NOC Service Desk

Impact

All Traffic traversing AL2S Core Node HOUH was degraded

Root Cause

The root cause is still under investigation by Brocade at the compiling of this report.

Repair & Service Restoration action

Core Node HOUH was rebooted, which cleared the immediate event, but was not thought to be a resolution or fix to the events long-term.

Prevention Recommendations

Methods to prevent the problem can't be determined until root cause is determined.

Impact Reduction Recommendations

1.) **Better understand AL2S VLANs affected by outages:** during this incident, getting a list of affected VLANs/customers was difficult. There is a complex manual process today to show VLANs that didn't failover when the backbone interfaces were shut down. We should ensure we have this information at our fingertips to understand the scale of service impact better.

2.) **Limit amount of time allowed for vendor diagnostics, improve Service restoration guidelines:** Multiple steps were taken to mitigate the service impact of the incident along the way. However, too much time was given to Brocade to run diagnostics before making the decision to reboot the switch. We should put additional guidelines for engineers in place so they will know how long a service can be impacted before escalating to more disruptive service restoration steps. It's been proposed that we target 1 hour for full service restoration, but this should be discussed with Internet2 service owners and Interne2 NOC to determine the right guidelines.
3.) **Improve speed of vendor-provided diagnostic scripts:** The initial data collection script Brocade has provided the Internet2 NOC took 50 minutes to complete its data collection. If we are to move to a new guideline restore service in an hour, this should be improved. It's unreasonable to have a diagnostic script that takes this long.
4.) **Improvements in software/network engineer involvement:** While both software and network engineers engaged quickly in this particular case, the current SOP doesn't ensure this, because issues are passed to one or the other according to specific criteria. The Internet2 NOC should improve this, so that the first action of a software or network engineer is to engage the other group unless an issue is clearly one or the other. This will lead to faster and more frequent joint troubleshooting.
5.) **Monitor Fabric module outages:** before this incident, there were no alarms for syslog events related to switch fabric module outages. This was added already, and completed on 2/6.

Timeline

Reporting/Monitoring

2/5/2014
8:39am ET (1339 UTC)

Alertmon displays Smokeping alarms related to the AL2S Houston Core node (HOUH). These alarms indicated Discards and Loss.

8:50am ET (1350 UTC)

SD Creates AL2S ticket 1316:135 to work this issue. It is opened at 1-Critical. This ticket is assigned to Net Eng. SD tech contacts the Net Eng assigned to the ticket by phone and hands the issue directly to them.

Investigation/Verification

9:10am ET (1410 UTC)

Network Engineering contacts Software Engineering about possible issues related to AL2S.

9:15am ET (1415 UTC)

SD tech advisement to SD I2 Supervisor of the issue per Critical status procedure.

9:33am ET (1433 UTC)

Software determines that OESS has received a link migration event from the HOUH switch. Neteng and Software force diff HOUH to restore proper forwarding; first OE-SS re-sync.

9:51am ET (1451 UTC)

Brocade was contacted to open a ticket opened to work with Brocade TAC.

9:58am ET (1458 UTC)

Net Eng indicates that this is likely to be a hardware issue with Brocade. This is due to the fact that they are reporting that every port is seeing discards except for ports that are in Slot 1 of the Core Node. Internet2 Director of Operations and Engineering advises that this is a Priority 1 case with the vendor, Brocade.

9:59am ET (1459 UTC)

Notification to the community is sent out regarding the issue.

10:07am ET (1507 UTC)

Systems Engineering continues to work with Network Engineers to resolve the problem. Software notices another link migration on the HOUH switch.

10:16am ET (1516 UTC)

Official start time of the Brocade ticket, created by Brocade engineers.

10:45am (1545 UTC)

Brocade TAC initiated conference call with Internet2 NOC engineers

10:53am (1553 UTC)

Internet2 NOC Network engineers disable SFM4.

11:11am ET (16:11 UTC)

Neteng and Software force the HOUH switch to diff restoring proper forwarding. Second re-sync.

11:25am ET (1625 UTC)

Software notices another link migration on the HOUH switch. Software asks permission to disable discovery until event has been resolved. Request is approved by Internet2 Director of Operations and Engineering.

11:49am ET (16:49 UTC)

Internet2 NOC network and software Engineers force the switch to diff, restoring proper forwarding temporarily. Third re-sync.

11:50am ET (1650 UTC)

Interne2 NOC enginer and Brocade believe there may be a switch-fabric issue.

12:13pm ET (1712 UTC)

Examination of the logs provides that 7 customer VLANs were affected.

12:22pm ET (1722 UTC)

Notification sent to the community with an update on the status of the issue.

Service Restoration/Resolution

Layer2 Backbone Restoration Time
10:22am ET (1522 UTC)

Internet2 IP network traffic is moved to 10G circuits. **At this point IP traffic is no longer affected by the incident. Layer2 traffic is still impacted.**

AL2S Circuits Restoration Time
12:13pm ET (1713 UTC)

With authorization of Internet2 Director of Engineering and Operations, all backbone interfaces are disabled. **At this point VLANs with backup paths are no longer impacted. VLANs that were explicitly configured without a backup path are still impacted, as well as VLANs where HOUH was the edge.**

AL2S Node Restoration Time
2:25pm ET (1925 UTC)

AL2S NPT indicated that Core Node HOUH is to be rebooted.

This reboot resolve the customer affecting nature of the issue. Once it was clear that the forwarding issues were resolved, the ports that were admin-down were brought back into service to fully resolve the outage. **At this point, all services were restored.** Problem resolution with Brocade continued.

3:26pm ET (2026 UTC)

Systems/Software re-enable discovery in OESS, and determine it is now functioning normally

5:21pm ET (2026 UTC)
Brocade TAC conference call ended, without root cause identified.

2/10/2014
8:41pm ET (0141 UTC)

Brocade requests and is provided additional Flow log data from the NOC; RFO still unknown.

```
SSH@sdn-sw.houh#show openflow flows e 3/1
Total Number of Flows associated with the port: 60

Flow ID: 7 Priority: 32768 Status: Active
        Rule:
                In Port:     e3/1
                In Vlan:     Tagged[102]
        Action: FORWARD
                Out Port:  e15/2, Tagged, Vlan: 111
        Statistics:
                Total Pkts: 0
                Total Bytes: 0
<snipped remaining>
```

Appendix B: Document Revision History

Date	Change
2/12/14	Initial Document creation