4/2/19 Network Maintenance Post-Action

NETS performed extensive testing of the Juniper and Cisco network systems at the Mesa Lab on 4/2. This post-action was prepared by NETS with input from the impacted network constituency. NETS met with the network constituency on Tuesday, 4/9. The summary of impacts, lessons learned, and conclusion are below. A more detailed report follows.

While the mission and goals of the testing were successfully accomplished, there were significant impacts to the network community that could have been minimized. These lessons learned will be integrated into future planning and implementation for major network maintenance.

Post-outage impact summary

- UCAR Operations’ Orion/IFAS did not come back online following the network interruption and needed to be rebooted (IFAS databases were brought down, but systems were not removed from the network).
- Notifier was unavailable as a tool for post-outage communication
- Enterprise Information Technology Operations’ (EITO) NetAPP network storage was disrupted
- SWEG’s Web service clusters were unstable
- Security cameras needed rebooting and some needed reprogramming
- CPO’s ProtectWise sensor had network issues
- Unidata’s THREDDS server was intermittently unreachable due to a core switch MAC programming issue. This was resolved by bouncing a port on ml1-core.
- Some Environmental monitors needed rebooting
- Several Infrastructure monitoring systems needed resets
- Two Wireless Access Points needed reboots
- Google activated email throttling (internal UCAR to external email addresses, e.g. Nagios or Notifier to gmail) that delayed email updates about the outage. Automated alerts triggered the throttling.

Lessons Learned

- NETS should clearly describe potential system impacts (the network was not just going down), e.g. “broadcast storms”, multiple tests, etc.
- Based on impacts from previous “reboot looping” events at ML and FL, NETS should have issued a stronger warning about the impacts of intrusive testing and advised IT staff to power off important systems to be safe. Clearer bi-directional communication would have benefitted NETS, UCAR Operations and other other groups whose systems were affected.
• In the Emergency Notification System (ENS) message, it would have been helpful to include language about the possible need to reboot. Even though it was pre-written, it could have been edited prior to sending.
  ○ There is some question about the protocols for ENS. Shawn Winkelman and Rob Reid will clarify the protocols for ENS usage outside of business continuity situations.
• In future communications announcing the end of maintenance, we should specify "The UCAR network is now functioning correctly" rather than "All network services are up"
• For future major maintenance windows NETS should identify two Network Engineers for the NOC, to provide continuous coverage during and post-maintenance.
• NETS should identify a communications lead during and post-maintenance who is not directly involved in the maintenance.
• While the NOC is the primary point of contact for network issues, we should supplement this during major maintenance with a pre-arranged Google Hangout or Slack chat channel where people could look for updates and communicate issues.
  ○ A communication manager would manage chat and real-time communication of the status of all affected systems
  ○ Send recurring updates on status and an all-clear
  ○ Keep chat infrastructure in place for ongoing issues
  ○ Test before next major changes
  ○ Have Google Doc open in parallel
• Consider turning off system email alerts during known outages to minimize email floods
• Verify emails sent are received
• Older operating systems like AIX (Orion) can be very sensitive to broadcast storms - needed a hard boot, indicating a Network Interface Card (NIC) problem. Could have pulled off the network in advance.
• Operations will update its escalation procedures to ensure timely response and escalation in the case of unforeseen issues.
• Define formal escalation process, e.g. CISL, COO, PC
• Operations IT called the NOC around midnight and NOC said status unclear and engineers had gone home.
  ○ NETS should have communicated more effectively with the NOC before, during and after outage
  ○ Staff can and should request escalation by NOC to contact engineer contact if needed
• NETS will review business continuity procedures and update and review with NCAB
• Relocating the physical security desk to CG went well
• Good test of evacuating the ML, and the process went well

Conclusion

While the work performed during this maintenance window was in-depth, intrusive, and disruptive, this was a very rare and exceptional network issue. The outage/testing clearly
identified the issue/fix required, which will be implemented during the MLDC Power Down on 4/13/19.

The packet looping problem was determined to be the result of an unfortunate coincidence of two different vendors’ interface code and a standard configuration of UCAR edge switches.

NETS apologizes to all groups that were impacted by this event and greatly appreciate everyone’s patience and cooperation. NETS acknowledges the need to strengthen overall communication before, during, and following impacting network changes. This includes identify and utilizing appropriate tools to strengthen and enable complete communications.

More details are provided below

Overview and purpose of network maintenance

On March 14th, The following message was sent seeking approval for a maintenance window, initially on 4/1. This describes the purpose of the disruptive testing on the production network. We did consider working during the 4/13/19 MLDC power down, as described below. It was also deemed infeasible to get the required Juniper staff in a timely fashion for a weekend window.

To:
Robert Reid <rreid@ucar.edu>,
David Maddy <maddy@ucar.edu>,
Mark Moore <mmoore@ucar.edu>,
Nick Wehrheim <nickw@ucar.edu>,
Anke Kamrath <anke@ucar.edu>,
Aaron Andersen <aaron@ucar.edu>,
Robert Wiley <rwiley@ucar.edu>,
Heather Harris <hharris@ucar.edu>,
Shawn Winkelman <shawnw@ucar.edu>,
Karl Werner <kwerner@ucar.edu>

NETS has been working with Juniper for many months to resolve a significant problem that occurs whenever a core switch is restarted. Unfortunately, our extensive combined efforts to replicate the issue in a lab environment have not succeeded. The only reasonable option we have left is troubleshooting in our production environment.

NETS proposed plan is to conduct a 6-hour network outage at the Mesa Lab on April 1, 2019 from 5pm to 11pm to perform disruptive, in-depth testing and data gathering for analysis and problem resolution.

We considered scheduling this during the MLDC power down on April 13 but that date does not work because key Juniper engineers are not available. Also, there will be many IT and facilities staff in the building
that may require the use of phones. Additionally, other work being done could be disrupted, and the problem being analyzed could be compromised with issues introduced during the power down.

During the proposed window, all network services at ML will be down. This includes all Cisco IP phones. Security cameras and other networked site security systems will also be unavailable. It may be prudent to require all non essential personnel to leave the building at 5pm. Networking at other UCAR facilities (including Internet access) should not be affected.

Planning and Advance Coordination

The date was shifted to 4/2 because of an evening reception in the ML lobby. IT managers with systems at ML indicated that it would be OK to proceed on 4/2. On Monday, March 18th, the following notifier was sent:

SCHEDULED OUTAGE NOTIFICATION

All network services at Mesa Lab (including ML datacenter), Fleischmann Building (FB), and the Marshall Field Site (MFS) will be down to perform disruptive, in-depth testing, data collection, analysis and problem resolution for an ongoing issue with the core routers. This outage includes all Cisco IP phones. The UCAR VPN will also be unavailable. Networking at other UCAR facilities (including Internet access) should not be affected. Cloud services including Google, Concur, Workday, and UProcure will be operational.

As a result, the ML, FB, and MFS will be closed to all personnel not directly associated with the network testing.

The building closures will be announced in Staff Notes and by direct email to ML staff.

NCAR, UCP, and UCAR management have approved this work.

NETS understands this is a large impact and inconvenient to all impacted staff, but NETS believes a dedicated testing window with the Juniper engineers is critical to resolving a long-standing and debilitating bug that is impacting NETS ability to move forward with hardware installations, upgrades, and new services. NETS appreciates your patience.

START: Tue Apr 2 2019 5:30 PM MDT
DURATION: 6 hours

SERVICES AFFECTED
UCAR Campus Network

PRIMARY CONTACT
John Hernandez
jph@ucar.edu
Building closure was deemed necessary due to the expected impact on phones and concerns over emergency response. The closure was coordinated with Dave Maddy and Rob Reid. On March 20th, Rob emailed all-ucar about the closure, and it was announced in Staff Notes on Monday, March 25th.

**Timeline of the testing**

After some preparations, at 6:08, we rebooted and triggered a broadcast loop lasting approximately from 6:14 to 6:22. This test validated/confirmed that the problem existed. The loops created about 30 million packets per second of broadcast in the core, and some fraction of this traffic reached most end systems. The effects were not entirely limited to ML. We saw evidence that high volumes of DHCP packets were relayed to other campuses. This was not anticipated, but we are unaware of any serious impacts at other campuses.

At 7:20, we restarted the packet forwarding engine, and again triggered a loop lasting from 7:24 to 7:31. This established an easier way to trigger the problem.

At 8:04, a third loop, lasting 28 minutes. During this loop, some network instability occurred at Foothills lab, presumably due to the DHCP packet leak. This interrupted our videoconference and delayed progress for about 10-15 minutes. We extended timers to provide more time to troubleshoot while the loop was occurring.

At 8:50, a fourth loop, lasting 37 minutes. This loop was somewhat lower in intensity (10 Mpps in the core due to fewer connected edge switches.) We turned off several edge switches to reduce the scope.

At 9:35, a fifth loop, lower intensity, lasting approximately 10 minutes. At this point, we had positively identified the source of the problem.
At 10:11, a sixth loop, lower intensity, that confirmed our finding.

At 10:32, we began to normalize the network.

At 11:32, we declared network service to be stable.

At 11:40, ENS message was sent indicating the end of the maintenance window (See below for ENS message text).

**Determinations**

What we discovered is the "packet looping after reboot" problem happens due to a combination of several factors:

- The configuration of our Cisco core-facing LAGs
- Inconsistencies in Juniper "hold-time down" behavior on QFX10002
- Architectural absence of spanning-tree in the EVPN core

Mitigation for the problem is being tested in a lab environment and will be deployed to the production network without causing further disruption.

**Post-outage communication and actions**

This ENS message was written in advance and sent by CASG at 11:41 when John called them:

```
Nets has completed network troubleshooting and maintenance operations. Network services at Mesa Lab have been fully restored. If you observe any problems with networking, please contact the N W S C network operations center at extension 4300.
```

At 11:45, John tried to send a Notifier message with more details, but discovered that the Notifier service was not online.

At 12:06AM 4/3, John sent the following email to nets-outage <nets-outage@ucar.edu>, nnag@ucar.edu (this email was delayed due to the queue on mail forwarders for nearly an hour)

```
Notifier has not yet been brought up by EITO/SWES yet (netapp storage needs to be up before notifier can come up), so I'm sending this update manually.

The ML network troubleshooting is complete, and we were able to successfully isolate the exact cause of our "reboot" issue! All network services have been restored at this time.

I will send a more detailed report in a day or two. Please contact the NOC x4300 if you observe any lingering effects of the network maintenance. Due to broadcast storms and
```
packet looping, some low-powered devices and embedded systems (IP cameras, Raspberry Pis, etc) at ML may need to be rebooted.

Thanks for your patience and cooperation!

Between 12:10 and 1:00AM, John and Paul worked with EITO, CPO, and Site Security to bring applications and servers back online. EITO had several machines that required reboots.

At 1:46 AM, Notifier service was back online, and the following Notifier update was sent, with more information about the intensity of the impacts:

The ML network troubleshooting is complete, and we were able to successfully isolate the exact cause of our "reboot packet looping" issue! All network services have been restored at this time.

I will send a more detailed report in a day or two. Please contact the NOC x4300 if you observe any lingering effects of the network maintenance.

Due to intense and lengthy network broadcast storms and packet looping that occurred during testing, some computers and many embedded systems (IP cameras, Raspberry Pis, etc) at ML may need to be rebooted in order to completely re-gain network functionality.

Thanks for your patience and cooperation!

Shawn W. spoke to Anke around 9am noting issues. Anke located NETS staff to work the problems. Paul, John, and Scot began arriving at 11 AM Wednesday morning. At 11 AM, Anke reported to Scot that OCOO was having trouble with IFAS. Scot called Shawn W. at 11:05am and learned that orion had been rebooted, and IFAS was back online. At 11:20, Scot sent this email to Shawn, ne@ucar.edu, and Anke:

Hey Shawn,

Thanks for talking with me on the phone about Orion just now. I'm glad that a reboot cleared it up.

As far as I know there is no ongoing network issue, but there are many knock-on effects of last nights work, such as what occurred with Orion, that will need to be addressed today.

As I explained on the phone, we were troubleshooting the cause of broadcast storms last night. Our new network gear is powerful, and when it gets in a packet loop it can generate a lot of broadcast traffic. This is exactly the thing we were troubleshooting last night, and we caused a lot of broadcast traffic in the process. Many devices' networking stacks are easily overwhelmed by this kind of storm and need to be restarted or rebooted to fix the problem.
Please let me know if there are any outstanding network issues that cannot be cleared with a reboot.