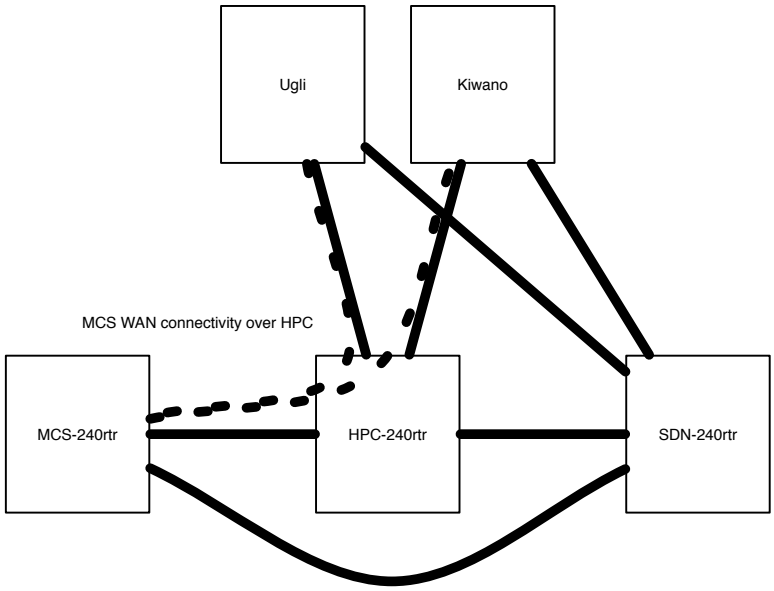
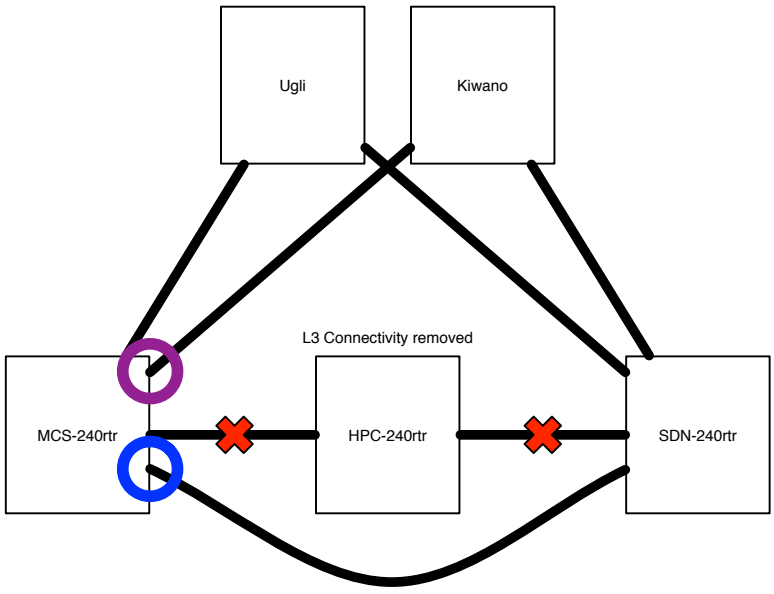
MCS Network Outage Digest

Background Details

On Tuesday, March 4, 2014, work was undertaken to remove the HPC router from the network. The initial state can be seen in this drawing:



Two things of note in this drawing is the reliance on HPC by MCS for access to the border routers. Additionally, the full routing mesh between the three devices: MCS, HPC and SDN. During the conversion process, the links from the border routers that terminated on HPC were migrated to MCS. This was done one at a time, allowing for the network to gracefully move traffic from one interface to the other. The network components were reused. This included cables and optics that had been in production since we moved into the space. Additionally, the virtual components, VLANs, networks, etc were also reused to make the transition seamless. Finally, all Layer 3 capabilities in HPC were disabled allowing the box to play a role as a simple switch, thereby simplifying the infrastructure.
The final state of the network can be seen in this drawing:



Two points of note here. The first is the blue circle. This represents the first outage affecting traffic travelling from MCS connected networks, i.e. routed on mcs-240rtr, particularly ALCF desktop net, towards SDN. The second is represented by the purple circle. This represents the second outage affecting traffic flowing in or out of the MCS connected networks, including traffic between 240 Server net and 221 Server net. These outages will be dissected in the following sections.

First outage

Notified by ALCF staff at 1210hrs that there was impact on the path from LCF desktop net to the Mira network. The root cause of this outage was a bad optic on the MCS-240rtr facing SDN-240rtr. Under load, this interface would "flap" causing a twofold effect. The first was a 30 second penalty imposed by spanning-tree as the link came back up. The second was the holdtime expiration on the OSPF peering between MCS and SDN. This is generally a survivable as there are multiple paths as can be seen in the drawing that would have carried the traffic without incident. There was another component to this failure which manifested the interruption symptoms. The dmz-incoming ACL implemented on the uplinks of MCS-240rtr control spoofing of traffic. Long ago, when the HPC and MCS infrastructure was split a decision was made to implement spoofing on the MCS side as assuming all of 140.221.0.0/17 lived on the box. All exceptions are implemented in the ACL as living elsewhere, on HPC, SDN and so on.
Traffic from Mira can flow through the border routers (Kiwano and Ugli) toward MCS, but the missing ACL entry for 140.221.68.0/23 meant that the traffic was dropped at the edge. Because we engineer traffic to use the short path between SDN and MCS, this oversight was not identified until the traffic was forced to transit this path. The remediation process began with the addition of this rule in the ACL allowing traffic to flow over the suboptimal path. Once this was done, the optic was replaced, link was tested with load, and OSPF was brought up with high costs (to ensure convergence first before traffic flowed). Once everything was tested, coordination with ALCF was completed and we moved traffic onto the link. Services restored and we moved forward.

Next Steps from First Outage

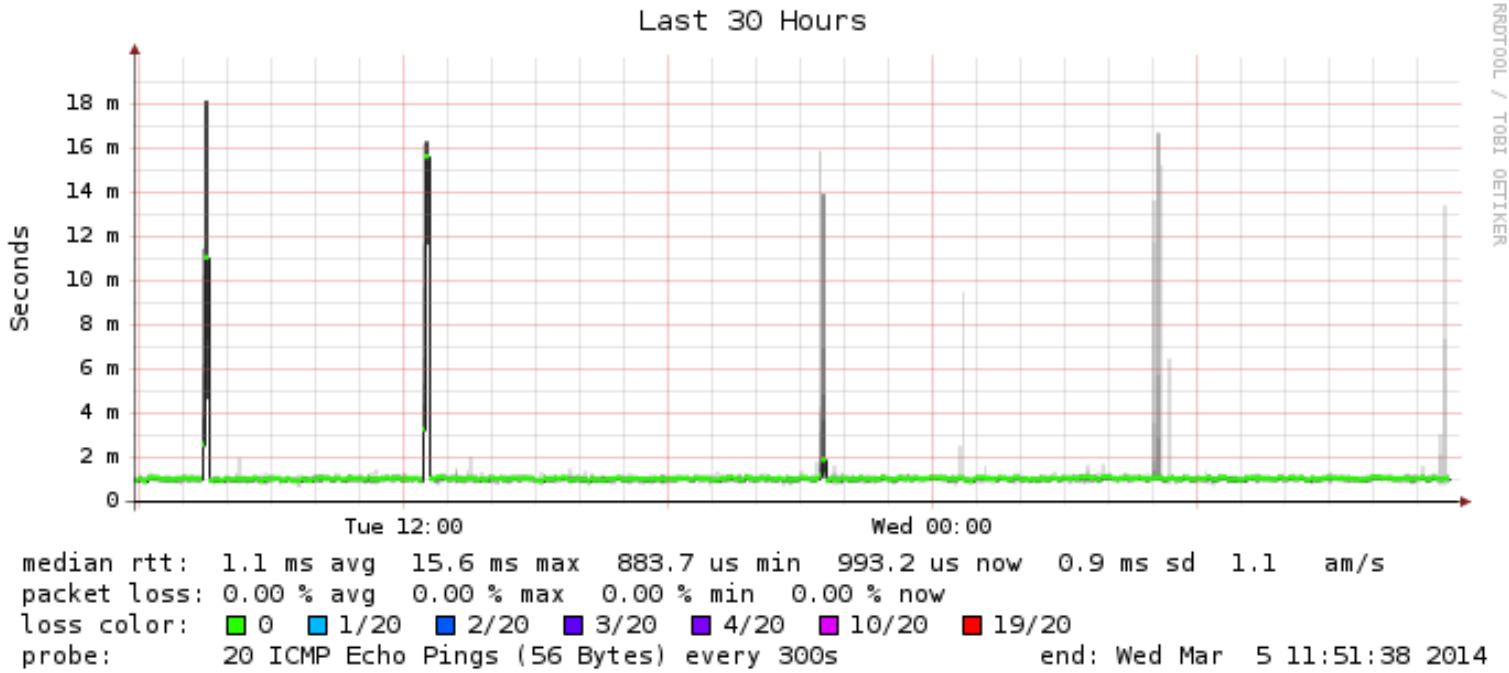
Two key areas were identified for review and action:

1. Full review of the DMZ ACLs to ensure they implement the appropriate protections and correct exceptions
2. Testing methodology change to allow detection of loss and impact on both WAN links as well as internal links

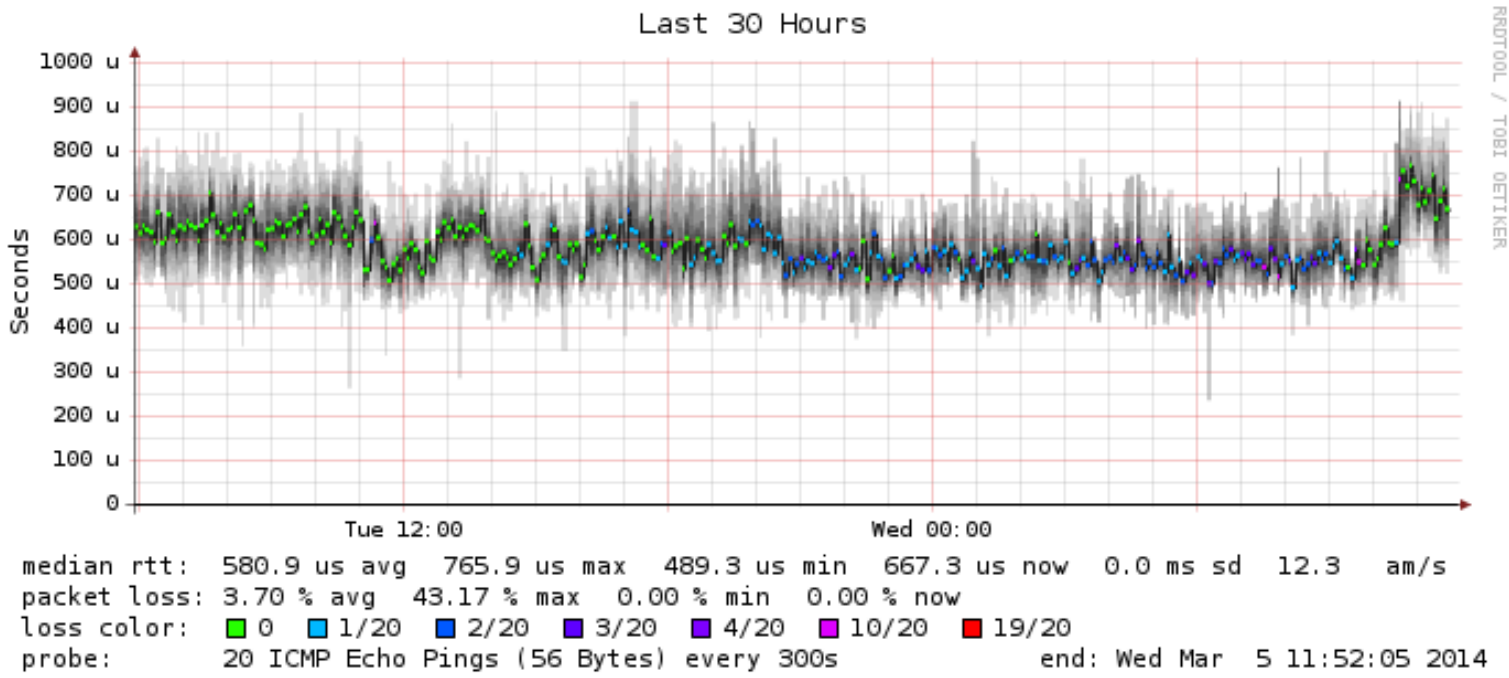
Additional work in this area may be warranted, but careful study of 1 and 2 will lead to actionable items.

Second Outage

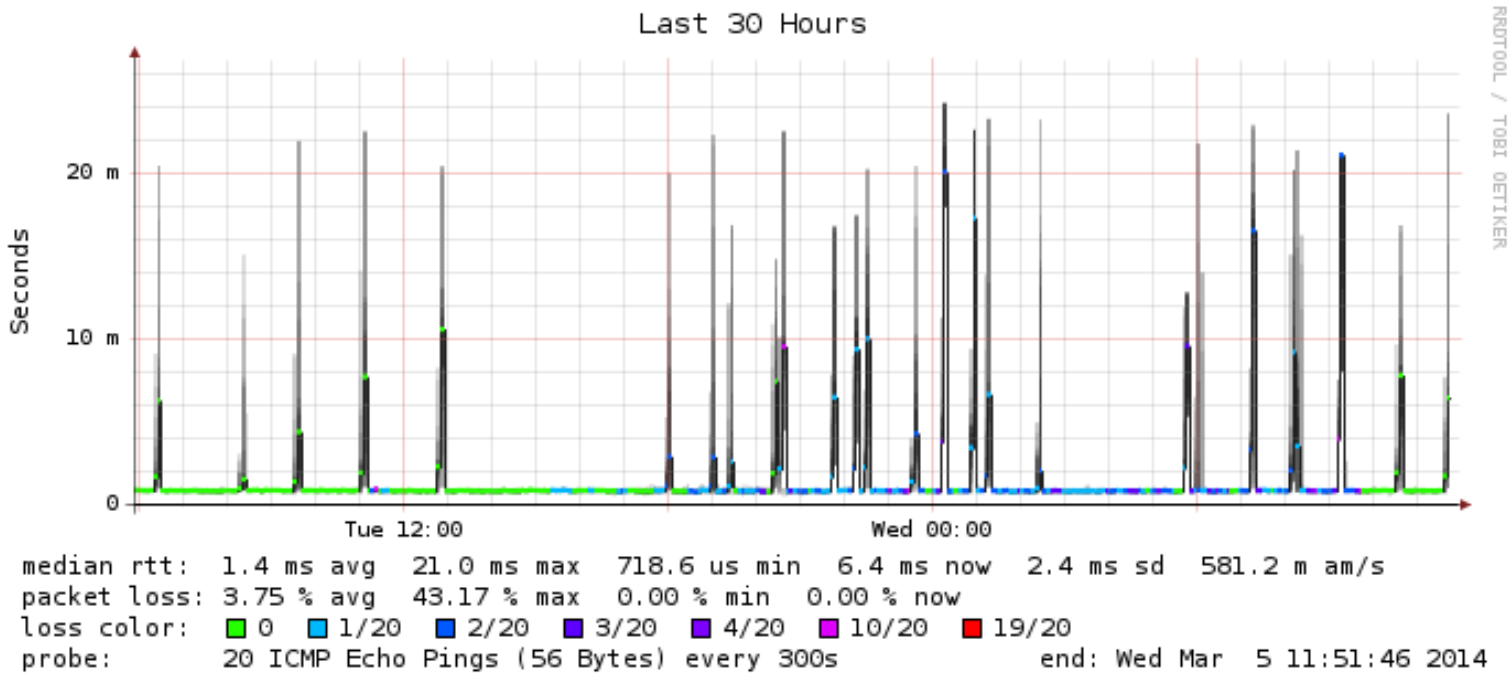
Notified by CELS Systems staff at 1721hrs that there were issues with infrastructure in 221. The root cause of this outage was a bad optic on the MCS_240rtr facing Kiwano. While the first outage was easy to identify the failed optic due to impacts on routing protocols, this failure had two key areas of mystery. First, the loss was in the 30-40% range, enough to allow communication, but forcing multiple retransmissions. This accounts for the reports of "laggy" response times. Second, the interface involved did not clock errors in the counters, and the far side of the link also did not record any errors in flight. The interface was effectively silently dropping packets.
During the outage, smokeping was reference to identify if there was loss or incurred delay inconsistent with the historical record. Looking at the graphs for the main MCS web server showed no issues or changes



At this point, we moved on to other testing points and troubleshooting steps. In doing so, we missed graphs like this (showing performance of one of the login servers):



There is no indication in change of latency, but there is a 30-50% packet loss event starting in the correct time window and lasting until the remediation steps were put in place. Another example:



Through much testing, config change reviews, troubleshooting steps, network isolation and a multitude of other attempts to correct the problem, we could not identify the source of the problem. A report from another source pointed us at performance issues interfacing with MCS networks. Testing on the network where power controllers are placed showed the problem, but there was no indication of where it lived in the network.
In looking at the MCS-240rtr, we noticed receive light levels on the MCS to Ugli link due to the placement of a hardware attenuator. We were set to remove this item when we noticed we were not getting light level diagnostics from the optic between MCS and Kiwano. Replacing that optic rectified the issue. Removing the attenuator normalized light levels, but this was not part of the operational impact.

Next Steps from Second Outage

Four key areas were identified for review and action:

1. Software update to GD release on the MCS-240rtr to allow full OAM (optical attribute monitoring) during the next maintenance window

2. Testing methodology change to allow detection and alerting of loss and impact on both on services from both inside and outside of the infrastructure providing a more comprehensive view of service delivery health
3. Implement full segmentation of the 221 and 240 server nets, they currently share vlan-id 808 which is plumbed between the two datacenters. This helps remove spanning-tree as a culprit and provides for better resiliency and removes a dependency, which is unnecessary.
4. Update drawings and documentation to better understand the placement and dependencies of services to help us diagnose and detect problems and do correlation. Without this we are essentially flying blind and grasping at straws. This may have sped up the process, but it helps the network folks understand what is being supported.

Additional work in this area may be warranted, but careful study of the items above, as well as feedback from the CELS Systems group will lead to additional actionable items.