

**Internet2 AL2s
Ternary Report
4-1-13 through 7-31-13**

Introduction

This is the second of a series of periodic reviews of the Internet2 AL2s infrastructure. The goal is to examine what is working and what isn't with an eye toward improvements and a focus on information important to the community for their decision-making process.

AL2S is envisioned as a ground-breaking new infrastructure for the Internet2 community that reintroduces concepts of bandwidth abundance, expands support for data intensive science and introduces for the first time broad programmability features inherent in SDN. The network is intended to assure both the Internet2 community is "way out in front" of commercial alternatives in its capabilities *and* that the community has a rock solid production platform for its critical science. To date, the SDN elements have been very strong while the infrastructure has been impacted by issues that should be trivial: failed hardware and core switching services.

The goal date for this report was the second week of August, 2013. There were issues with the preparation and development of the report. Given the timing it is difficult to examine the reporting period without also including issues and commentary related to the Noisy Backplane/FEC issues that hit the Brocade immediately after the reporting period.

Given the extended period of this report, it has become important to add commentary about repeated issues with the Brocade nodes of the network. The Brocades have been hit by at least two hardware bugs. The first involved an initialization sequence on the LR10 optics that caused flapping of the optical interconnects between nodes. This was resolved with a code update. The second was a data corruption bug, discovered by IU Research, which corrupted certain data profiles that traversed the backplane of Brocade nodes. This was again solved through a code update. There are three other issues with the Brocades which are more ambiguous. The first was a flapping issue between two nodes in Chicago: CHIC and STAR. The circuit flapped over 50 times over a period of a few days. This turned out to be a hardware issue with a failing line card in the Brocades. This occurred in the midst of a larger Brocade flapping issue where several BROCADE-BROCADE and BROCADE-JUNIPER circuits, all with underlying Ciena transport, have flapped for less than 600ms. These 600 ms "micro-flaps" remain an issue. It is unclear if Brocade support had trouble identifying the CHIC and STAR hardware failure because of simultaneous work to identify the micro-flap issue. Finally there was a severe packet loss issue that effected major peer AMAZON where the Brocade in WASH was silently dropping frames until rebooted. The reboot fixed the packet loss

issue though. A fabric module was replaced in WASH, adjacent to the Amazon peering at Brocade's recommendation. A failure analysis has not been completed on that card yet but it's believed that the failing fabric module was the root cause of the issue. Finally, the packet-loss issue repeated itself again in on links between Colorado and Chicago with a Big Data project having dropped frames on the backbone. To correct this the node was rebooted, which fixed the problem. In both the WASH and DENV case FEC was enabled at the suggestion of Brocade but Brocade is NOT committing that FEC solved the second frame drop issue or even that the MLX-16 chassis needs FEC. As of October 1st, the messaging from Brocade was quite the opposite. The FEC feature is currently being recommended as a Good Idea to have enabled, but it not required. Currently it is not clear that Brocade understands why the second frame drop problem occurred. To enhance monitoring for future events, Internet2 now has active performance monitoring turned up to watch for the silent drop issue, is monitoring several new counters on the Brocades, has FEC enabled on every Brocade backplane, and will be much more actively managing Brocade in the future.

The raw data that most of the report is based off of comes from the AL2s Weekly reports, compiled in to spreadsheet form.

Report Overview:

- Upgrade Details*
- Unscheduled Outages/Incident Management*
- Scheduled Outages/Change Management*
- Availability Information*
- Bandwidth Utilization*
- OESS Utilization*
- Progress from Last Report Deliverables*
- Future Plans for Layer 2*
- Timeline of Installs/Upgrades*

Upgrade Details

A number of upgrades have been done to the core node software as well as the OE-SS application that provisions services on AL2S. Feature additions and bug fixes are the primary drivers for upgrades, although the DR Failover test is indicative of the sorts of proactive work that the is increasing.

Firmware upgrades;

2013-7-14 - Upgrade Juniper's to 13.2-20130608

Junos 13.2-20130608 improved OpenFlow flow mod processing time, all Junipers were removed from static insertion and put into production OpenFlow mode at this time

2013-7-17 - Upgrade Brocade's to 5.4.0d

5.4.0d fixed two bugs, the LR10 Initialization State and the Payload Corruption

Software upgrades:

2013-3-31 - Upgraded OESS to 1.0.7-1

2013-4-22 - Upgraded OESS to 1.0.7a-1

Emergency release to fix a bug in fwdctl that causes flow_mods to be counted multiple times on device reconnections

2013-4-23 - Upgraded OESS Database components to 1.0.7a-1

2013-4-23 - Upgraded OESS Database components to 1.0.7b-1

2013-5-19 - Upgraded OESS to 1.0.8-1

2013-5-24 - Upgrades OESS core to 1.0.8a

Critical issues in fwdctl

2013-7-2 - Upgraded OESS to 1.0.10 (including 1.0.9)

2013-7-20 - OESS to 1.0.12 (including 1.0.11)

Unscheduled Outages/Incident Management (Data in Appendix E)

An examination of the graphs in Appendix E shows some interesting facts. It should be noted that the most recent recording period, September, is represented to the far right and the most distance, April 2013, is represented to the far left; a labeling error on the X-axis occurred.

The largest impact on unavailable minutes for AL2S remains maintenance or fiber cuts in the underlying L1 circuit infrastructure. The ability to offer resiliency, either at L2 or at higher layers, remains extremely important in order to mitigate particularly the uncontrollable external influence of long duration fiber cuts.

The “Brocade mystery flaps” along with “suspect fiber bumps”, lumped together as a category “Undetermined” represent a large number of incidents but a relatively small amount of actual unavailable minutes on AL2S. It would be interesting to understand the degree of churn in both the higher-layer resiliency methods (IE: IGP/ISIS) as well as the churn in the AL2S ‘failover circuit’ resiliency methods. Is it worth examining at what point L2 churn/L3 churn becomes a problem? The L3 topic area is well explored.

Note the average duration of the hardware impact: generally less than 2.5 hours. Not bad. Note also the large spike in Software in April. This is because of code/OS maintenance on the nodes; software loads in both cases.

As compared to the last ternary, Layer 1 is improving. There were 34 incidents in the previous report and only 12 in this report and that’s with the assumption that “Undetermined” is L1 and the busy summer road construction season that often affects Layer 1 availability. Total number of incidents is also dropping, from 45 in the first report to 22 in this one. It’s unclear why, although there may have been false positives in the first report; the collection method is now weekly instead of every four months, which helps the integrity of the data.

Scheduled Outages/Change Management (Data in Appendix F)

An examination of the graphs in Appendix F shows some interesting facts. It should be noted that the most recent recording period, September, is represented to the far left and the most distance, April 2013, is represented to the far right.

We began tracking Change better during this period, based on some industry standard ideas around Service Metrics, KPI's, and Critical Success Factors, although the process is still in it's infancy. We're looking at Emergency vs. Non-Emergency change, failed changes (0 so far), Changes Causing Incidents (1 so far), Changes with unexpected results, and breakdowns by change type.

The "Change causing Incidents" event was a change that had problems and ran over its maintenance window. The weekly after-action of Changes on AL2s triggered a post-mortem analysis that should be released Any Day Now .It also caused us to now track "Change Causing Incidents" and "Changes delivering unexpected Results", the later basically meaning that the change worked but did not go 100% smoothly.

The graphs reveal us getting better at Change, as revealed by the total Duration of Change Type as well as the Average Duration of Change Type, although they also reveal a general trend of more change. The challenge here is to retain the nimbleness required of an advanced network while minimizing the risks and impact inherent in Change. Our regular examination of Change is part of minimizing risk while reducing the impact points once again towards the need for a more comprehensive view of Availability and an effort at 100% availability for members.

Finally, the rate of software change may be misleading. We may need to break this out to differentiate OS Upgrades on the Switches vs non-OS software changes, such as controllers, etc. Or it could be data collection for the sake of data collection.

The number of changes as compared to the last report is also down 50%: 83 vs 31. L1 fell from 27 to 15 and software changes fell dramatically also. Again, this is probably relates to changes in how the data is recorded, completion of the BTOP build-out work, etc rather than a real improvement.

Availability

An examination of the data in Appendix D (Availability Graphs) reveals some interesting data points. When looking at the graphs it must be noted that in August the ability to differentiate between Schedule & Unscheduled outages became available. IE: The blue line can be misleading.

Node availability remains very good however that in and of itself is a problem. The packet drop issues related to data corruption and noisy backplanes don't show in these reports because they don't represent the node or circuit being 'down' in the current definitions. This is clearly a major problem. The purpose of the graphs to is display problems so they can be addressed and the purpose of the network is to transfer data reliably to end connectors, not keep a 100% availability figure on core nodes. The definitions for what is 'Available' clearly need to change. Link-up/down, etc are just one symptom of an unavailable network. Another fallout from the FEC/noisy backplane issue must be in the way we report availability. We must move to a new model in which the ability to move data error-free is the measure to which we hold our service commitments. IE: matching what our member community is actually trying to do. As Internet2 and the Internet2 NOC move to implement a more comprehensive operational Performance Assurance/Masurement infrastructure it must be the case that the Availability measurement move along with that. Which still leaves open the question of 'how many packets can the system drop before we consider it down? And/or is there a more meaningful measurement of availability than that?'

The report is also hard to read because of the inability to remove Layer-1 issues. The weekly operational report has to be examined for the time periods in question to determine if the Availability drop was actually L1 or L2 induced ... and while they both result in downtime we have to keep the scope of this effort scale to I2. It's currently a manual process.

Bandwidth Utilization

An examination of the data in Appendix C (Bandwidth Graphs) reveals some interesting data points.

The transition of certain AL3s network segments from native circuits to using AL2s as transport shows up readily on the graphs. Usage on the western US remains light however that should pick up as CENIC brings their 100g circuits on line.

Another observation requires more thought. The first is the existence of large research flows on the network. This shows up readily on circuits such as ELPA-HOUS or ATLA-JACK, among others. The circuit runs flat and stable until a large spike of data is observed and then returns to normal. This would be indicative of bandwidth tests, and some inquiries confirm this: several large flow projects are beginning to ramp up their usage and the spikes represent testing that are the first signs. IE: Exactly what should be happening. This does however bring to light the discussions over the headroom practices that have been going on since at least December 2012. This was recently brought up again as a major discussion area in the NTAC face-to-face meeting in Dallas. It doesn't matter if the system is dropping as a result of hardware issues or because the links are oversaturated because the results are the same: higher-layer services are impacted. The Rapid Provisioning Pool allows Internet2 to augment 100gig circuits within 2 weeks however that's not good enough in a system supporting higher-layer services. QOS must be deployed to protect the Layer-3 services (R&E, TR-CPS, etc) and finalizing the headroom practice must be given priority. In real terms this means deploying the existing QOS solution on the Brocades and pushing Juniper harder on solution to queuing. QOS solves the higher-layer services issues and the rapid provisioning pool in combination with a new headroom practice solves the issue of the research/big-data flows not getting the bandwidth they need.

OESS Utilization

OESS is the provisioning tool which allows users of AL2s to configure vlans across the infrastructure. Typically this would be through the user/member using the web interface to set up the VLAN although an API service is available also.

There are currently 42 different workgroups configured for use in OESS, with 153 users defined among those groups, 177 vlans in use, 174 ports available for configurations, and 166 circuits events just in the month from July through August.

Only about 25% of the users are GlobalNOC/Internet2 staff, which would seem to indicate a decent number of users in the community. The remaining users tend to be clustered among four other major users of the system with most other sites only have 3-8 users. It would seem that outreach to those sites to expand the provisioning capability could be in order.

Other than the backbone proper there are only two other major users of the system. OSCARS, a software IDC package/protocol that enables the stitching together of L2 VLANs along an inter-domain end-to-end path, ranks high as does a particular big science/data project. Other sites generally have only a few vlans each configured.

VLANs tend to be long-lived, as illustrated by the quantity of circuit events in July. Of the 166 events 102 were OSCARS events. The nature of OSCARS makes this interesting: these are all VLAN requests from outside the I2 backbone domain.

Ensuring that new AL2s connectors promptly get a login to the system and they get the data/training they need to use it should be a goal within the next reporting period.

Progress on previous report deliverables (From the October-March report)

Little to no progress was made on the suggested deliverables from the last report. The transition from the first report to this one was more difficult than anticipated. See

- The reporting tools clearly need the ability to differentiate between Scheduled & Unscheduled outages to bring I2 in to line with best practices.
 - **No progress in the reporting period, although the change was implemented in August 2013, the currently active work period.**
- Availability reporting should shift from an individual elements/nod/circuit model to a Services based model.
 - **No deliverable during this reporting period although the backend systems have begun modifications to shift from a node/circuit model to a Services based model**
- Another tracking category is needed. "Impairments", representing something other than a binary "Up/Down" status.
 - **No progress during the reporting period, although progress should be reported during the next report as as result of actions taken as part of the Performance Assurance portion of the "The Amazon/FEC" post-mortem.**
- More granularity is needed in the reporting system to differentiate the layer of the service impact. For example,L2 outages caused by L1 circuit outages.
 - **No progress during this reporting period, however in the currently active reporting period it is the case that better records are being kept and there is a regular weekly process to update those records based on the weekly reports.**
- Further support training is needed for the engineering and systems support staff.
 - **Staff have been engaged in a directed learning program to increase their knowledge of the systems involved. This area will be targeted again in the next quarter by the AL2s NPT with a group training exercise scheduled, include the SD SST.**
- Better tools are needed to identify and localize problems. The Layer 2 environment is sufficiently different to require reexamination of the troubleshooting process.
 - **No progress during this reporting period. This gap was highlighted again during the NTAC face-to-face meeting in Dallas.**
- A better tracking system for major project events (upgrades, major bugs, etc) is under way.
 - **No progress during this reporting period, however in the currently active reporting period it is the case that better records are being kept and there is a regular weekly process to update those records based on the weekly reports.**
- The community groups are working on a set of metrics to better measure & compare the Layer 2 system.

- **No progress during this reporting period. This gap was highlighted again during the NTAC face-to-face meeting in Dallas.**
- The backbone upgrade policy is vague. The only current policy is based on a Layer 3 IP network running at 10gigabit speeds and handling generalized R&E traffic.
 - **No progress during this reporting period. This gap was highlighted again during the NTAC face-to-face meeting in Dallas.**
- The Layer 3 backbone will begin using Layer 2 (via vlans and SDN-signalled circuits) for 100g transport.
 - **No progress during this reporting period, however additional progress was made during the currently active reporting period.**
- Openflow 1.3 support is heavily examined on OE-SS, Flowvisor, and the backbone switching nodes, with a desire to move to it quickly to support several gaps in the 1.0 standard.
 - **The beginnings of a justification paper have been made, with Mpt-Mpt, QinQ, and QOS being heavily referenced with member use cases.**

Future AL2s Plans (August 2013-November 2013)

Based on this ternary observations:

- A Service owner for AL2s to drive the service from more than a technical viewpoint. While this report focuses on the technical it is clear that's not the goal for AL2s. AL2s exists to enable research and higher-layer services. More focus on that can be brought through quarterly goal setting. We need to develop targets for inclusion on the next report that are more than just technical.
- More assertive management of the deliverables. The deliverables must get on to the development roadmap faster and be fleshed out with dates assigned. It is critical that we be able to deliver on commitments.
- Better management of Brocade. We must be much more focused in our dealings with Brocade. Cases must be opened and aggressively engaged on every incident and we must strictly follow their escalation policies for each one. In addition, a call should be arranged between the community and Brocade to help explore the issues around the boxes.
- As indicated in the Unscheduled Outage section, we should explore the impact of SDN circuit failover churn on L2. This should probably start with measurement.
- In the Change reports differentiate between switch software and non-switch software loads.
- Emergency Change was high in the last two months (outside of this ternary report but in the next one) because of the reboots to enable FEC. It was a risk management decision: do we perform an emergency change with the risk that a shortened time window implies ... and risk the 'silent frame drop' issue reappearing or do we perform a normal Change after a longer notification window? Better understanding of what causes us to perform an Emergency Change would help us make those decisions quicker and better.
- We should expand deployment of the active performance-monitoring infrastructure. This acknowledges the gap in our systems and the real goal not being "all links/nodes up" but rather "the community can transfer data error free." This would mirror the shift from thinking of AL2s-as-a-network to AL2s-as-a-service.
- Similarly, we need to look at the way we measure Service Availability on AL2s. While link up/down may play a part in network availability the service availability should probably transition to something dealing with active performance monitoring. We should figure out a position on this in the next ternary, at a minimum.
- While not directly related to AL2s, it is worth mentioning that we need to better understand what we do in relation to our connectors having 100% availability. At a minimum we should examine all of the L2 & L3 connectors and determine, in conjunction with them, if they have the physical infrastructure/connections in place to support 100% availability of services and what we can do to help beyond current incentives for dual homing and regional collaboration. Perhaps a goal for the ternary after this could be

examining our processes to ensure we don't step on backups during Scheduled/Unscheduled events.

- We must finalize the headroom policy at L2. The existence of large research flows must be taken in to account.
- We must protect higher-layer services on L2, be they TR-CPS, member backhaul to L3, or L3 backbones. We need to come up with a plan and deploy it in the next ternary.
- We should embed an engineer in to the weekly/bi-weekly technical calls of our largest Big Data users. This should allow us to provide a higher level of service to these critical users and also to get ahead of problems that tend to show themselves first in this cohort.

Improvements made in the last ternary:

- Better change tracking.
- Better incident tracking.
- A better post-mortem process.
- Q/A process implemented for weekly report data for scheduled & unscheduled outages
- Network Availability differentiated Scheduled/Unscheduled.

AL2s Service Roadmap, next Ternary

- GENI API support in OESS
FOAM is an aggregate manager used on GENI today to manage slice creation and approval for OpenFlow controllers being virtualized with FlowVisor. This allows users to manage GENI slivers through the GENI Aggregate Manager API. Internet2 and GlobalNOC are working with the GENI Project Office and the FOAM authors to extend FOAM to interoperate with OESS. This will allow OESS circuit creation and management on AL2S using the GENI API.
- FlowSpace Firewall
FlowSpace Firewall provides the ability to run multiple OpenFlow applications/controllers on the same switches providing a form of network multi-tenancy. It operates as a proxying OpenFlow firewall, restricting which part of the flow space a controller can manipulate. It provides the ability to enforce VLAN Tag based flow space restrictions and provides control channel rate limiting.

AL2s Timeline

2013-3-31 - Upgraded OESS to 1.0.7-1

2013-4-17 – Installation of the Ashburn AL2s node.

2013-4-22 - Upgraded OESS to 1.0.7a-1

2013-4-23 - Upgraded OESS Database components to 1.0.7a-1

2013-4-23 - Upgraded OESS Database components to 1.0.7b-1

2013-5-19 - Upgraded OESS to 1.0.8-1

2013-5-24 - Upgrades OESS core to 1.0.8a

2013-7-2 - Upgraded OESS to 1.0.10 (including 1.0.9)

2013-7-14 - Upgrade Juniper's to 13.2-20130608

2013-7-17 - Upgrade Brocade's to 5.4.0d (data corruption & Lr10 bugs fixed)

2013-7-20 - OESS to 1.0.12 (including 1.0.11)

2013-7-25 – Installation of the Minneapolis AL2s node.

2013-7-27 – Installation of the Phoenix AL2s node.

2013-8-1 – Change, Availability, and Post-Mortem tracking improvements

2013-8-11 – Tracking of Scheduled vs. Unscheduled outages implemented

2013-8-18 – Micro-flaps noticed

2013-8-26 – CLEV-CHIC flapping fixed/hardware RMA'd

2013-8-30- Installation of the Pittsburgh AL2s node.

2013-9-2 – “The Amazon Issue”

2013-9-4 – Installation of the Portland AL2s node.

2013-9-15 – FEC enabled on Brocades

Appendix B – OESS Feature/Bug Details

More release details at: <http://globalnoc.iu.edu/sdn/oess/revision-history.html>

Mar 31st 2013 - Upgraded OESS to 1.0.7-1

- Add data refresh to discovery data tables.
- OESS usage graph show interface description
- OESS ability to browse list of interfaces and their descriptions
- Show DPIDs in HEX in logging
- Add circuit CLR To UI
- Add ability to force "re-provision" circuit
- Dynamic sizing for workgroup selector box (along with the next buttons)
- Add ability to search for workgroup or user
- Remove internal circuit 'identifier' from UI as it is not useful to users
- Properly Display nodes as having no endpoints in circuit provisioning if user does not have access to any endpoints on that node.
- OESS UI Refresh: Users Tab, Actions Section
- OESS Edit circuit (with a backup path) leaves primary path in deploying state
- OESS allows for link decommissioning even with backup paths on that link
- Add OESS Network Status Page
- Add the available resources tab
- Make feedback button mailto configurable
- SNAPP config gen for OESS bombing with undefined host_id
- Add ability for OESS to manage edge port vlan restrictions
- Fix bug where fwdctl on device connect does not reset flowmod count to 0
- Resolve bug where multi-point circuits in OESS have too many actions causing dup packets
- Cleanup Internal errors in nox
- Resolve bug which results in handling of OESS OSCARS topology submission with duplicate entries
- Resolve bug where OESS is not counting existing flows against the max-flow-mods during diff
- Resolve bug where OESS rule diffing breaks vlans with untagged endpoint
- Resolve Bug where port down events firing too fast causing "double" fail over
- Resolve OESS UI Bug: Network Status / Available Resources page does not clear the Map Session Data
- Resolve OESS Bug: Scheduler reporting
- Resolved bug where OESS topology submission not including edge interfaces

Apr 22nd 2013 - Upgraded OESS to 1.0.7a-1

- Emergency release to fix a bug in fwdctl that causes flow_mods to be counted multiple times on device reconnections

Apr 23rd 2013 - Upgraded OESS Database components to 1.0.7a-1

- Fix for admin section problem

Apr 23rd 2013 - Upgraded OESS Database components to 1.0.7b-1

- fix for internal vlan ids not returning proper values.

May 19th 2013 - Upgraded OESS to 1.0.8-1

- added configurable restore to primary behavior after hold timer
- added email notifications for network events and user events
- added a raw (flow rule) view of circuits
- added the ability to check for failover events on device re-sync
- generation of OESS topology for IDC purposes was including multiple copies of decom nodes
- improved barrier detection
- improved logging in oess_scheduler
- improved logging in fwdctl
- improved logging in NOX
- resolved an issue that prevented the network status map from properly centering
- resolved an issue that prevented non-endpoint nodes from showing data for circuits
- resolved an issue where large numbers of flow mods caused vlan_stats to process flows slowly, and would back up

all other process (fwdctl, topo, etc)

May 24th 2013 - Upgrades OESS core to 1.0.8a

- Critical issues in fwdctl

Jul 2nd 2013 - Upgraded OESS to 1.0.10 (including 1.0.9)

Features:

- Added the ability to send barriers to each device once instead of after every flow mod when doing failovers added dpid to oess admin network elements popup
- graceful trunk interface swaps
- trunk metrics defined in admin UI and used in path calculations
- support Discovery when using hybrid mode on switches that do not support untagged frames.
- modest performance improvements in the restoration code path
- Added the ability to detect link port moves and adjust flow rules accordingly

Bugs:

- fixed a possible synchronization issue between topo and fwdctl
- fixed OESS OSCARS integration when interfaces have names with spaces in the topology
- several fixes related to total flow_mods per switch tracking
- several fixes related to email notification based on network event
- fixed issue where we should have been interface diff but were doing a full node diff
- resolved internal event handling race condition discovered as a result of improvements in restoration code path, would have impacted restoration triggered by port_down events.

Jul 20th 2013 - OESS to 1.0.12 (including 1.0.11)

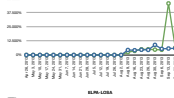
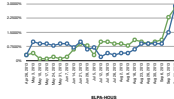
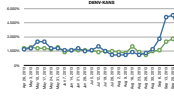
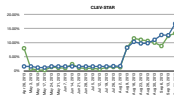
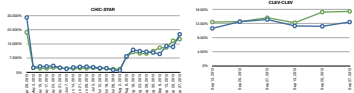
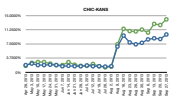
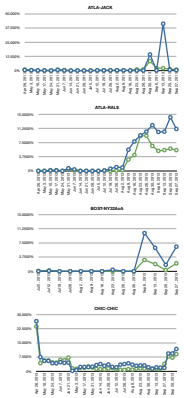
Features:

- OSCARS circuits created through the OESS UI are now owned by the OESS workgroup

Bugs:

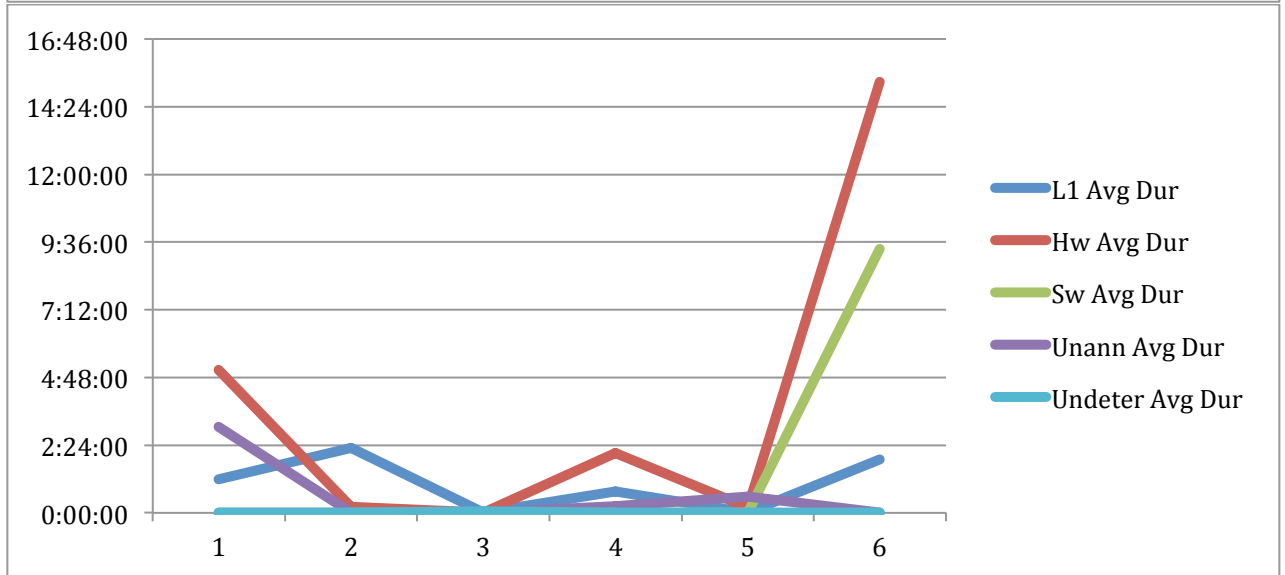
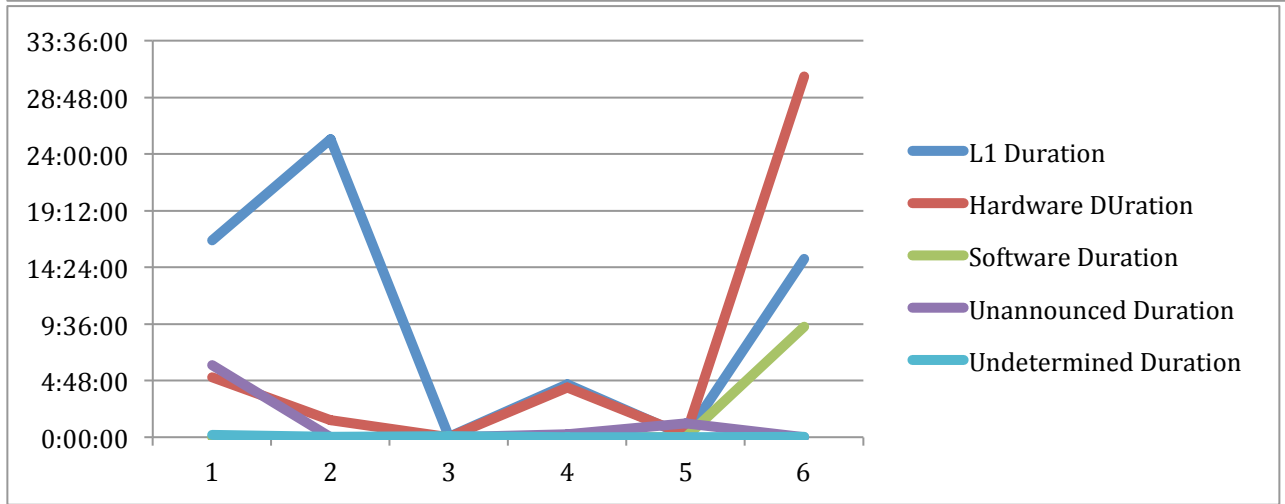
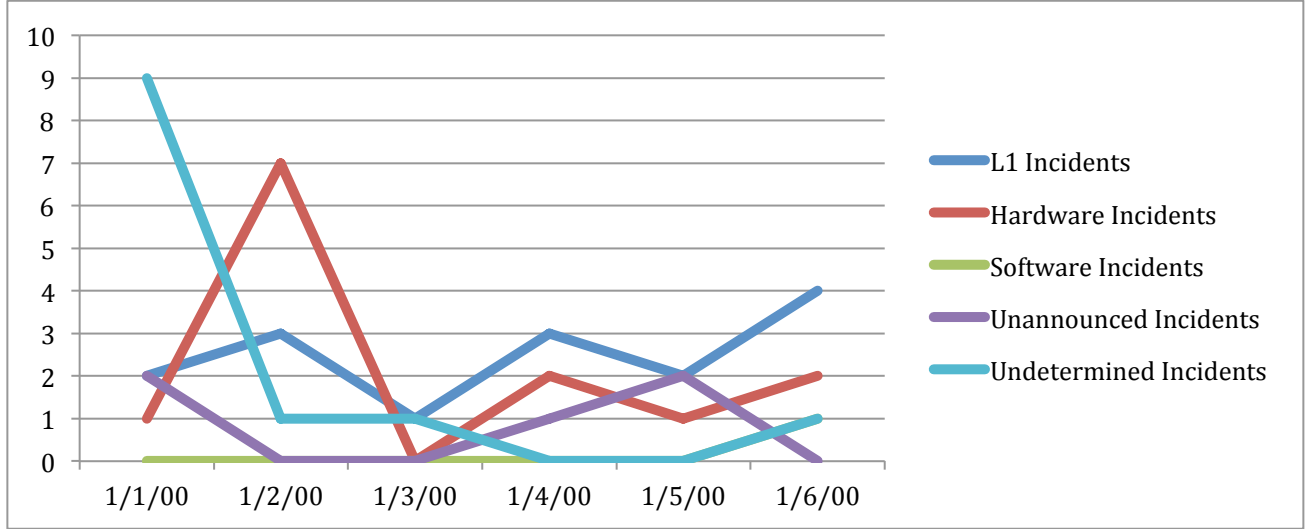
- OESS Circuits without backup paths would end up with empty backup paths, and would fail over and not work
- OESS could not schedule a removal time on circuit creation
- new links on insertion do not have a status
- OESS CLR not showing minutes < 10 properly
- approval of node not honoring bulk barrier flag
- fixed problems related to inserting a node in the middle of the path

Appendix C - Bandwidth Graphs



Appendix D – Availability Graphs

Appendix E: Unscheduled Outages/Incidents



Appendix F - Scheduled Maintenance/Change Management

