Options for Implementing Internet2 IP Network Atop AL2S

Overview & Scope

In February, 2013, the Internet2 technical community worked with Internet2 staff to define a topology that leverages the investment in the 100G AL2S equipment to provide enhanced bandwidth for the IP (AL3S) Network. At a high level, the AL2S network will provide backbone and edge connectivity to the IP routers. This document will focus on the logical organization of the IP network as it relates to the AL2S network.

A full description of the physical topology can be found in the NTAC documents circulated in February. Briefly, each of the 10 Internet2 IP routers will be interconnected with the resident AL2S equipment via two 100GigE circuits, spread across diverse linecards. Internet2 Members may have connections to AL2S, IP Network, or both, with backups to other nodes possible.

The Community dialogue resulted in the utilization of the following three types of logical interconnects between the IP routers, atop the AL2S network:

- SDN signalled - using the OESS software to create an SDN-signalled VLAN between router nodes.
- Native signalled - using the hybrid capabilities of the AL2S equipment to configure a static VLAN between router nodes, using the vendor CLI.
- 10G LAGs - using the existing Nx10G circuit link aggregation groups (LAGs) that are in place today between the IP routers.

Edge connections will have a similar set of SDN- and native-signalled VLANs. Each physical interconnect to the AL2S network will be configured with sdn- and native-signalled VLANs to two physically diverse IP routers. In addition, Internet2 members will be able to maintain the their pre-existing set of 10G interconnects to the routers through at least the end of 2013.

The flexibility of the lower-layer topologies provides the Internet2 community with several different options to organize the Layer3 network atop the Layer2 topology. The NTAC discussed several approaches, but the final decision was left to a more in-depth analysis of the different options. This document attempts to collect the different options along with a high level analysis of the potential advantages or disadvantages to each approach.

There are a few areas that are currently out of scope for the document. The hybrid capabilities of the AL2S equipment may allow for utilization of native routing and Layer 2.5 technologies like

MPLS. While those may ultimately provide us with additional knobs and ways to optimize data, the February architecture discussion limited itself to considering native and SDN ethernet switching. This document adheres to those principals and leaves the discussion of a more integrated routing approach for a future exercise. All routing and restoration will continue on the existing Juniper MX960 IP routers.

IP to AL2S Interconnects

Each IP router will be attached to the AL2S network with two 100G ethernet circuits. These will be spread across physically diverse linecards on the routers and Layer2 equipment. Unfortunately, the Openflow 1.0 standard does not support link aggregation, so the network will need to be engineered so that specific VLANs are steered across the individual links. This provides several options:

SDN/Native separation

The traffic could be separated such that one 100G interconnect carries all the SDN-signalled VLANs and the other carries all the native-signalled VLANs. This has the advantage of being simple to understand, but will likely provide a poor distribution of traffic across the two 100G circuits. It also carries a higher risk of completely cutting off one interconnect technology, should an individual linecard fail and take down one of the two circuits (see discussion below). However, this may be preferable from a Member access circuit perspective since one of the two member access methodologies will remain up in the event of a single 100G circuit failure.

100G Link #1	100G Link #2
Chicago Backbone (SDN)	Chicago Backbone (Native)
Houston Backbone (SDN)	Houston Backbone (Native)
Salt Lake City Backbone (SDN)	Salt Lake City Backbone (Native)
LEARN (SDN)	LEARN (Native)
GPN (SDN)	GPN (Native)

Backbone Split/Member Access Split

This scenario spreads the bandwidth between the two links, placing some backbone interconnects and some member access interconnects on one link, and the remainder on the second. This has the advantage of providing more flexibility to balance traffic, but

it will be a manual process that will require constant attention. It also could mean the potential total outage of a member access path, should the link they be configured on go down.

100G Link #1	100G Link #2
Chicago Backbone (SDN)	Houston Backbone (SDN)
Chicago Backbone (Native)	Houston Backbone (Native)
Salt Lake City Backbone (SDN)	Salt Lake City Backbone (SDN)
LEARN (SDN)	Salt Lake City Backbone (Native)
LEARN (Native)	GPN (SDN)
	GPN (Native)

Backbone Split/Member Access Split with methodology split

This scenario spreads the bandwidth between the two links by placing some backbone interconnects and some member access interconnects on one link, and the remainder on the second. It differs from the above in that for each SDN-signalled path on one interconnect, the twin native-signalled path is configured on the other link. This has the advantage of providing more flexibility to balance traffic, but it will be a manual process that will require constant attention. It also provides the ability for one of the two signalled path methodologies to remain up, if any one of the two interconnects goes down.

100G Link #1	100G Link #2
Chicago Backbone (SDN)	Chicago Backbone (Native)
Houston Backbone (Native)	Houston Backbone (SDN)
Salt Lake City Backbone (SDN)	Salt Lake City Backbone (Native)
LEARN (SDN)	GPN (SDN)
GPN (Native)	LEARN (Native)

Outage Scenarios

Given the physical topology, IP over AL2S will introduce some new outage scenarios to contemplate. The list below is partial, but illustrative of some of the types of scenarios that would need to be considered when designing the logical organization of the IP network.

- AL2S Chassis Failure
    - Backbone
        - Both SDN-signalled and native-signalled backbone paths across the 100G SDN backbone will go down. 10G LAG interconnects will remain up.
        - IP router will receive link down alert and *could potentially* be configured to react to event without waiting for IGP timers to time out
    - Edge
        - Both SDN-signalled and native-signalled member access VLANs will go down to the IP router in the same city
        - Members that have multiple AL2S connections will maintain their connectivity to their backup set of routers via their other SDN- and native-signalled VLANs
        - 10G edge connections to the router in the same city will remain up
- AL2S Line Card Failure (100G Backbone)
    - Backbone
        - A subset of SDN and native signaled backbone VLAN paths will go down.
        - 10G LAG interconnects will remain up.
        - IP router will not receive a link-down alert and IGP timers will be invoked to detect the outage
    - Edge
        - A subset of Non-local SDN- and native-signalled member access VLANs will go down. Those members will fail over to their backup Layer3 cities
- AL2S Line Card Failure (AL2S-IP interconnect)
    - Backbone
        - A subset of SDN and native signaled backbone VLAN paths will go down.
        - 10G LAG interconnects will remain up.
        - IP router will receive link down alert and *could potentially* be configured to react to event without waiting for IGP timers to time out

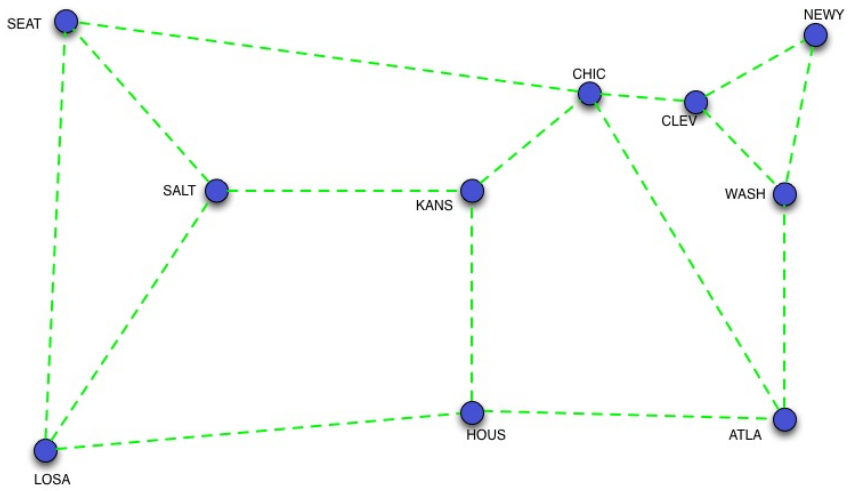
    - Edge
        - See above discussion of different edge balance methodologies
- Switch firmware upgrade
    - Backbone
        - Both SDN-signalled and native-signalled backbone paths across the 100G SDN backbone will go down. 10G LAG interconnects will remain up.
        - IP router will receive link down alert and *could potentially* be configured to react to event without waiting for IGP timers to time out
        - IP traffic *could be* manually re-routed around the AL2S device prior to the switch maintenance.
    - Edge
        - Both SDN-signalled and native-signalled member access VLANs will go down to the IP router in the same city
        - Members that have multiple AL2S connections will maintain their connectivity to their backup set of routers via their other SDN- and native-signalled VLANs
        - 10G edge connections to the router in the same city will remain up
        - IP traffic *could be* manually re-routed around the AL2S device prior to the switch maintenance.
- Fiber cuts and Layer1 segment-wide outages
    - Backbone
        - All three backbone interconnect options on a given segment will go down.
        - IP router will receive link-down on direct Nx10G
        - IP router will not receive a link-down alert on SDN- and Native-signalled 100G paths
        - IGP timers will be invoked to detect the outage
    - Edge
        - Locally switched member access circuits (those physically interconnected to the local AL2S switch) will remain up
        - Remote switch member access circuits that are configured over the affected Layer1 segment will go down.
- Layer1 transponder failures
    - Individual SDN, native signalled or 10G LAG interconnects will go down
    - Restoration will happen as described above, depending on the link
    - Remaining connections will continue to pass traffic in the designated priority

Upcoming Topology and Architecture

The February architecture discussion explored two different ways of logically organizing the IP network atop the Layer2 network substrate. This document discusses both of those options and a hybrid of the two.

Mirror Legacy Network

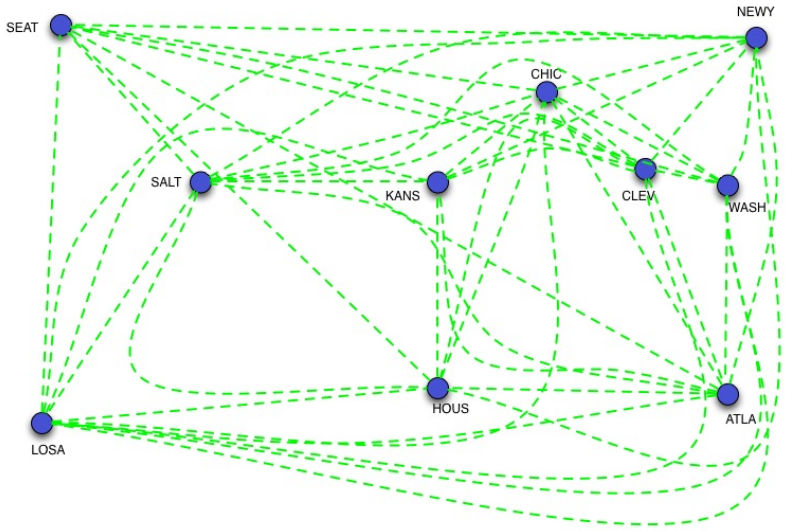
In this topology, the SDN- and native-signalled inter-IP VLANs are created in a topology that closely mirrors the existing IP network topology as of early 2013. Cross-country traffic will “pop” out of the AL2S network into intermediate IP nodes, then be “pushed” back into the AL2S network on its way to the final destination. For example, a flow between Los Angeles and Chicago may find itself routing through the Salt Lake City and Kansas City routers on the journey to Chicago.



- Pro’s
  - Provides greater visibility into network path across the network using traditional tools (eg, traceroute)
  - simplifies configuration and troubleshooting
  - provides the means for a straightforward comparison of historical availability data with the new topology
- Con’s
  - requires IP-switching of transit flows at intermediate hops via the AL2S network (router-on-a-stick).
  - Carries the highest utilization of AL2S/IP interconnects over time.
  - Greatest hop count

Full mesh

The full mesh would connect a native 100G VLAN, and a 100G SDN signaled VLAN between every IP router, resulting in N*(N-1) Layer-2 paths across AL2S for both SDN- and native-signalled VLANs. The VLANs would generally be configured to follow the shortest path between two nodes as determined by fiber distance (though there could be a scenario where traffic is routed based on traffic load). Because of this, traffic would follow the same geographical path it would have historically.



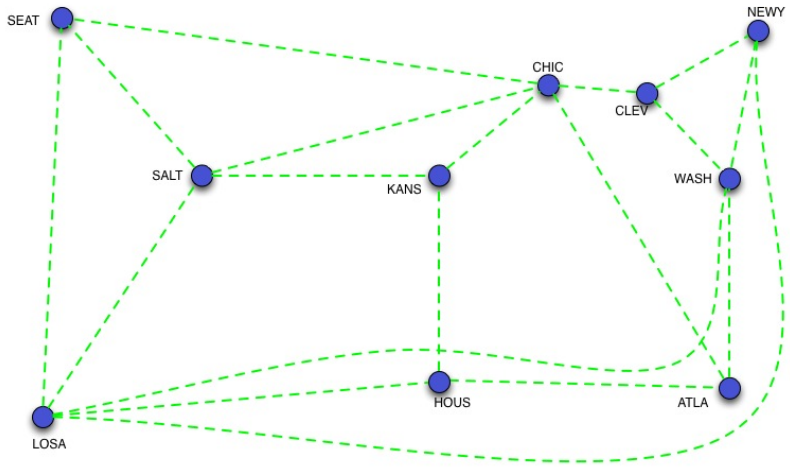
- Pro's
  - Lowest hop count
  - Avoids loop of traffic between AL2S and IP equipment at intermediate nodes on end-to-end path; more efficient use of 2x100G interconnects
  - Minimized hop latency
  - Measuring traffic between endpoints may be easier

- Con's
  - Complex multi-protocol, multi-layer restoration
  - Greatest complexity in configuration (both L2 and L3)
  - Convergence times may suffer in case of a failure/outage, due to the increased number and complexity of interconnects, and may require additional tuning.
  - Use of BFD (see below) for link failure detection would be dependent upon propagation delay and require a great deal of tuning
  - Makes comparison of uptime data to historical uptime data very difficult given the different topology

- IP nodes will have no knowledge of the underlying AL2S topology. This will require manual correlation when configuring AL2S paths, and tuning IGP metrics or certain protection mechanisms such as MPLS Fast Reroute (see below).
- The flow space on the Juniper hardware is currently limited (as of May 2013), so we may bump up against the available limit, leaving little or no room for member-signalled circuits

Partial Mesh

Like above, native and signaled VLANs are created in a topology that mirrors the existing physical 10G topology. Additionally, when higher traffic paths between routers are identified, AL2S paths are constructed between the two IP nodes.



- Pro's
  - Provides a phased migration strategy that enables up front analysis of new topology against historical topology data
  - Added adjacencies based upon demand
  - Mitigates some of the router-on-a-stick issues
- Con's
  - Requires constant monitoring and path loading policies that trigger direct adjacency over the IP network
  - Create topology churn on a continual basis that may be confusing to the user base
  - Greater complexity in configuration (both L2 and L3)
  - Diverges from the legacy topology over time and makes statistical comparison difficult.

- Some full mesh concerns (see above) apply

Current Link Restoration Method

The IP Network currently uses IS-IS with default settings as it's IGP. There is no methodology currently configured that provides for more immediate restoration of link-down events, so all circuit outages will result in at least a 5 second delay before the SPF algorithm processes the topology change.

For non-local link failures, IS-IS relies on the loss of hello PDUs to detect a link failure. The hello-interval is set at three seconds, with a hold-time of nine seconds. Three hello-intervals can be missed before IS-IS will consider the link down and start processing a topology change. SPF will wait an additional 200 ms after a topology change is detected. Therefore, the current network could result in times between 5-27 seconds to restore traffic around a disruption, depending on where the event occurred relative to the timers.

Possible Link Restoration Methods

Given the long traffic restoration times, and the vast increase in network paths between each router node, the following potential restoration options are posed for discussion.

Tune IS-IS:

By modifying the default IS-IS timers, for local link failures, the hold-down timer can be lowered to 2000 ms after a topology change is detected. For non-local failures, the lower bound of the hello-interval is one second. To avoid false positives, an option is to implement a hold-time of up to 3x the hello-interval, or three seconds. This could bring restoration down to 3-9 seconds.

IS-IS Loop Free Alternative (LFA) and Node-Link Protection:

Juniper claims restoration time of 50-100 ms is possible with LFA. LFA precomputes loop-free backup routes for all IS-IS routes. These backup routes are pre-installed in the Packet Forwarding Engine, which performs a local repair and implements the backup path when the link for a primary next hop for a particular route is no longer available. With local repair, the Packet Forwarding Engine can correct a path failure before it receives recomputed paths from the Routing Engine.

Enhance IS-IS with Bidirectional Forward Detection (BFD):

By implementing BFD with IS-IS, detections of link failure or topology changes can be greatly reduced. This is important, as the AL2S inter-node link-state is not relayed to the IP router. (see outage scenarios discussed above) BFD will allow for sub-second failure detection, allowing IS-IS to quickly process topology changes. Best current operational practices suggest setting intervals at 300ms and a multiplier of 3. This would

give us sub second restoration.

MPLS Fast Reroute (FRR):

FRR relies on pre-calculated, alternate Label Switch Paths (LSP) to follow when the primary path is impaired. FRR offers the best restoration time, and requires no additional hardware. Utilizing FRR, restoration time down into the sub-100ms range is possible, with exceptions for propagation delay. Normally, FRR relies on link-detection to rapidly identify and signal re-route events. This can be sped up by coupling FRR with BFD. Another potential challenge is the signalling of the backup paths in any scenario where several router interconnects are using the same Layer 2 physical path between two cities without knowledge of the underlying topology (as with a partial or full mesh topology).