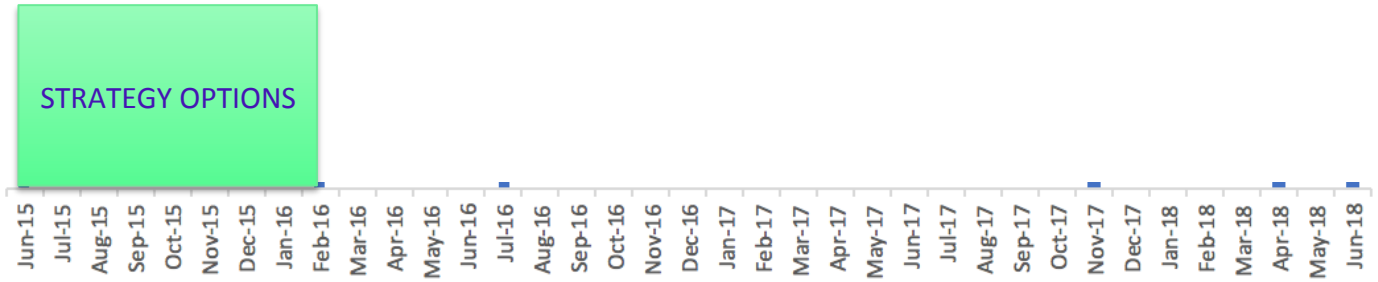

linx
Deploying a Disaggregated Model for LINX's LON2 Network
How LINX reimagined its LON2 network architecture using EVPN routing technology

LON2 Refresh Project Background

The Network

- LINX runs two exchange fabrics in London
  - LON1 being the larger LAN running VPLS using traditional Router Equipment
  - LON2 was running native layer-2 using switching equipment
- We had been attempting to move to VPLS on LON2, but not successfully
- 2015 saw huge take-off in 100G orders,
  - Could see we were going to outgrow existing chassis
  - Core growth also would require reasonable investment

New Strategy



- Even if we did not change vendor, a significant refresh was needed
- Started talking to equipment suppliers
  - Traditional router vendors at one end of spectrum
  - Open Networking solutions at the other end
- Instead of just comparing vendors, we looked at potential strategies for LON2
- At meeting coinciding with NANOG64, existing vendor stepped back

Looked for best strategy

- Different vendors suited different strategies
- Traditional RFP, plus conversation with vendors to narrow down solution
- Selected best match for each strategy option
- However, IXPs have requirements that were new for several vendors
  - Worked with vendors on how to address those
- Consulted with membership on their preferences
  - Strategy, not vendor
  - Took advantage of NANOG66 to meet face to face with US based members

Why are IXPs different

The port is the demarcation

We need to monitor, diagnose and fault-find based on only seeing one end of the link

Large range of port speeds

- Larger Members are multiple 100G, smallest GE.
- Limited control of location of various speeds –
  - ports all over the place
- Background flooding is significant issue for smaller members
- All on one big layer2 broadcast domain
  - Can't logically separate big ports from small

MAC Security

- Controlling exactly what MAC addresses come from what port is key to an IXP.
- MAC Learning is not always a good thing.
  - Broadcom learns before MAC ACL

Partner Ports

- Like most exchanges, LINX has a partner program
- It allows 3rd party partners to manage connectivity from the member to the exchange
- Member is now a VLAN
  - Partner connects with single port (or LAG)
  - Each member delivered on their own VLAN on that port
  - The bandwidth of the partner port is shared between the members
  - All Member features are now per VLAN
- Multiple VLAN tags on same port mapping to a common VLAN is a very unusual feature for a layer2 switch.

Early Steps

First Found Hardware Partner

- Edgecore Networks
  - Hardware provider
  - Part of Accton, one of the largest more respected OEMs/ODMs
  - 30 Years Experience, many established customers
- First attempt at testing was a failure
  - Wrong NOS (Software) for our needs
  - Exchange features were “Fragile”
  - Called POC off early
- Edgecore team used experience to really understand our requirements
  - Last day of POC was just a dialogue on requirements

Edgecore introduced us to IP Infusion

- IP Infusion
  - Original developers of Zebra, became specialist stack vendors
  - Investing heavily in NOS Ecosystem
- Worked with Edgecore to build an initial demo (not quite full POC)
- As we did not know IP Infusion, we also got 3rd party references
- IP Infusion had ambitious plans for their NOS
  - If successful, would be not only low cost, but high featured
- Edgecore Networks and IP Infusion seemed committed to invest significantly in the project to make is a success
- Our conclusion was: “If it works, it’s the right choice”.

Agreed target solution EVPN

- All switches have a common MAC table – synchronized by BGP
  - Don't need to worry about one-way traffic flows
  - Less likely to run into data-plane learning Bugs
    - A MAC address is a BGP learned route populated into a forwarding table, just like IP
- Traffic is tunneled through network, so no MAC-Flush re-convergence
- Much better at controlling flooded traffic
  - Can manually configure a MAC address, and rely on BGP for its propagation to other switches
  - If switch does not know about the location of a MAC address, it is not reachable, no need to flood.
- Has option of multi-homing

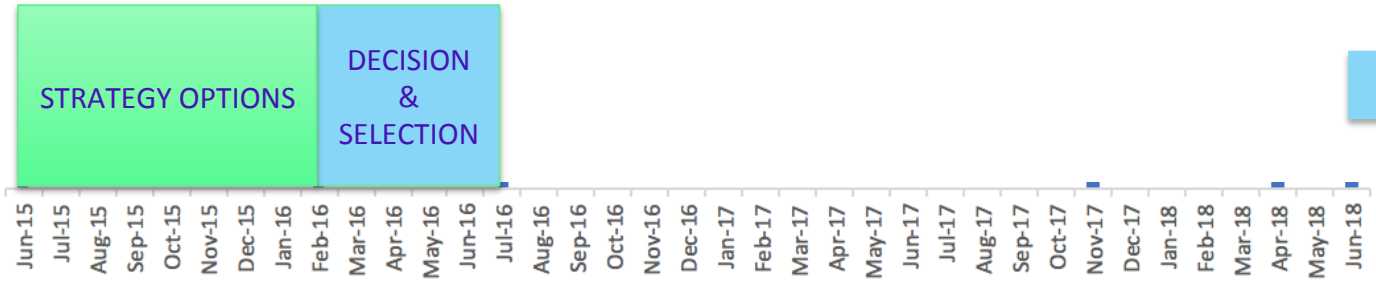
Agreed target solution Exchange features

- MAC ACLs
- Many to one VLAN mapping
- Per VLAN traffic policers on single port
- Per VLAN allowance for ARP and IPv6 ND traffic
- Disabled MAC Learning and statically configured MAC addresses
  - With option to fallback
- Proxy-ARP and Proxy-ND to reduce background traffic
  - With option to fallback
- Limit traffic to traffic types legal on Exchange
  - Want to see everything if in Quarantine

No Central Controller

- LINX had wrong DNA
  - In those days, our technical team was primarily network engineers
  - Our software platform team were primarily focused on non-mission critical infrastructure
- We had ambitions on Automation, but did not want to overstretch a developing team
- Control-plane based re-convergence is faster than controller based

Start of the real work



- And yes, that was a bigger gap than expected or hoped
- We were sweating existing assets in the mean time

Reality

MPLS Labels

Broadcom StrataXGS

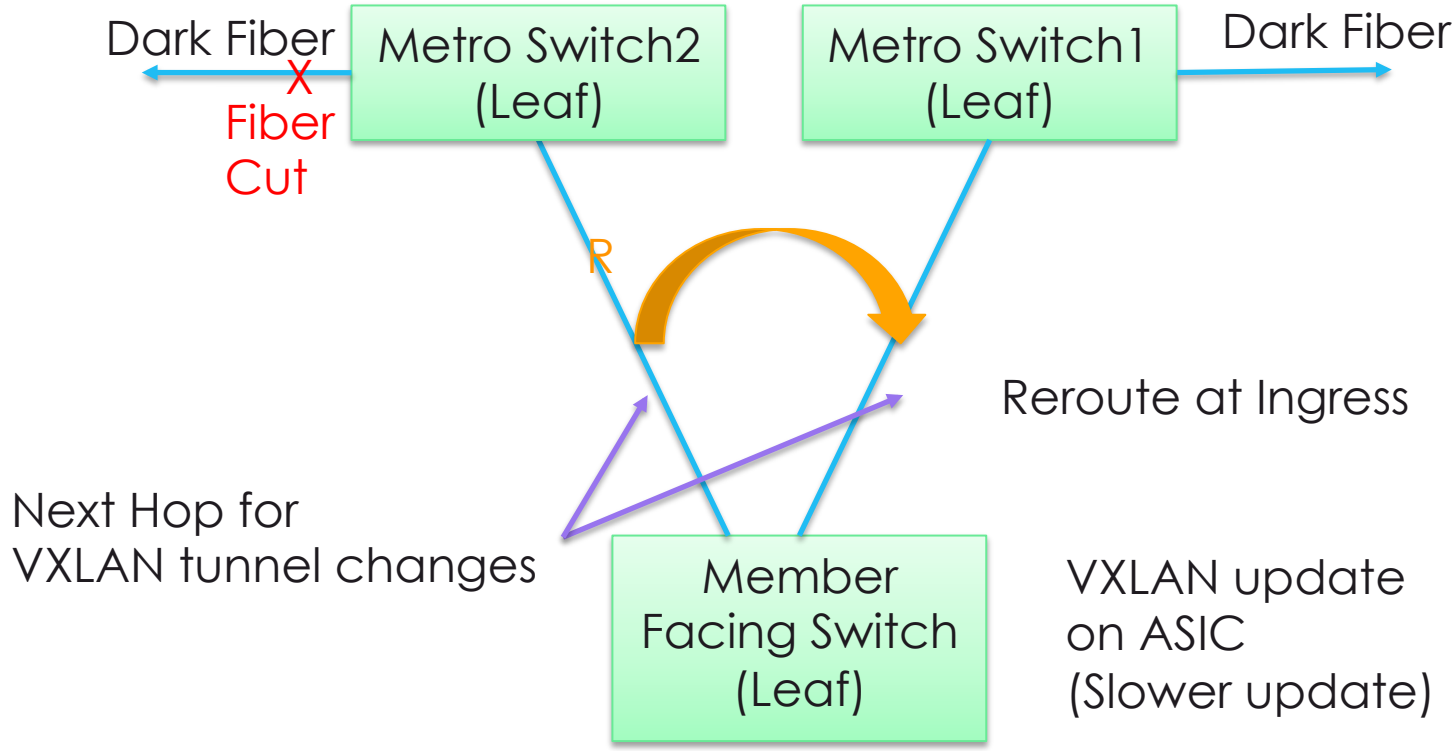
- Limit of how many labels it can remove in one go
  - Therefore Entropy Label not an option, multiple end to end LSPs needed
  - ESI label for Multi-homing a real push, would need to violate RFC
  - Could go through pipeline twice, but that is half the bandwidth lost
- Designed for VPLS, so EVPN pseudowire-less operation a real concern
- Each LSP consumes an entry in interface-table
  - We were likely to run out of entries at the core of the network (N-squared scaling with the number of edges).
- Broadcom were very supportive, but in the end too high a risk

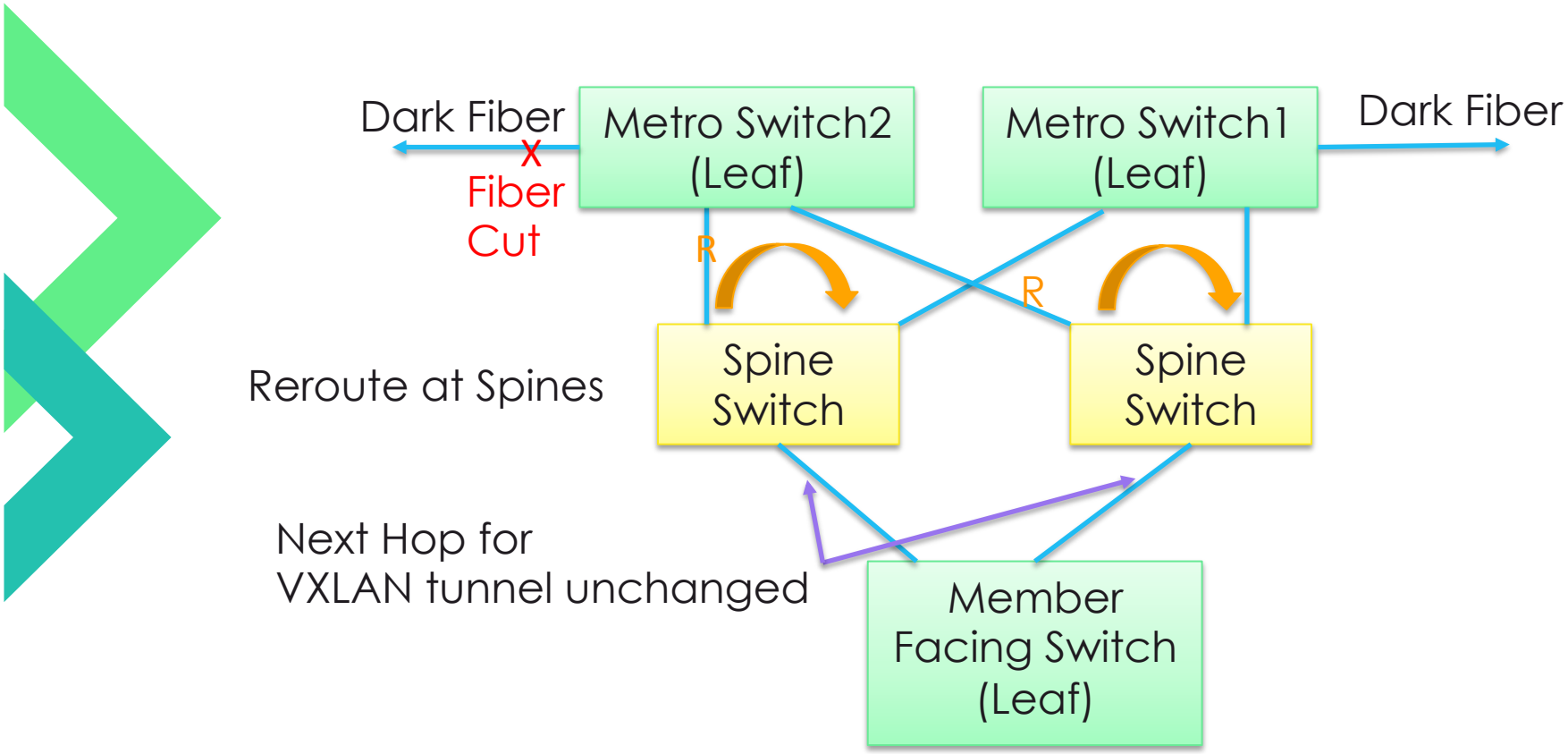
New target solution VXLAN

- Alternative way to carry EVPN signaled Ethernet
- IP Infusion already working on this with other customers – but without exchange features
  - Those could be ported
  - All the work on EVPN re-usable
- Avoided many of the challenges of MPLS
  - Use UDP source port instead of Entropy Label
  - No ESI label requirement for Multihoming
- We could work around the limitations
  - Tunnel statistics good enough for traffic planning
  - Convergence was worse than MPLS, but expected to be good enough

VXLAN Convergence

- MPLS with Fast-Reroute can reconverge at any point to a pre-computed alternate path – sub-50ms
- For VXLAN (IP) Topology is key!
- Need to both re-compute destination and reprogram ASIC
  - Worse case is if re-route occurs at the entry to the network, and flips from next-hop A to next-hop B.
    - VXLAN has more state to update.
    - 300 to 600ms full reconvergence
  - Better if the re-route is not at the first hop.
    - 150ms to 300ms full reconvergence


Dark Fiber
X
Fiber
Cut
Metro Switch2
(Leaf)
Metro Switch1
(Leaf)
Dark Fiber
R
Reroute at Ingress
Next Hop for
VXLAN tunnel changes
Member
Facing Switch
(Leaf)
VXLAN update
on ASIC
(Slower update)


Dark Fiber
X
Fiber
Cut
Metro Switch2
(Leaf)
Metro Switch1
(Leaf)
Dark Fiber
R
R
Reroute at Spines
Spine
Switch
Spine
Switch
Next Hop for
VXLAN tunnel unchanged
Member
Facing Switch
(Leaf)

VXLAN Convergence (cont)

- Significantly better if re-route is just loosing options from ECMP (load-sharing) - So from 2 next-hops: A+B, to just one next-hop: B
  - About 50-100ms if at entry (eg loosing a spine)
  - About 50ms if not at 1st switch
  - Loosing links from LACP well below 50ms – usually sub 10ms
- Don't have hold-time at repair, so small hit there if not load-sharing
  - But still sub 50ms
  - Added requirement to deal with flapping interfaces
- Better than previous LON2 convergence times

The Technical Solution

Leaf and Spine

- Design methodology emerged from hyper-scale data-centers
- We chose it due to easy and predictable scaling
  - Common simple building blocks means fast deployment
  - Made convergence simpler and faster

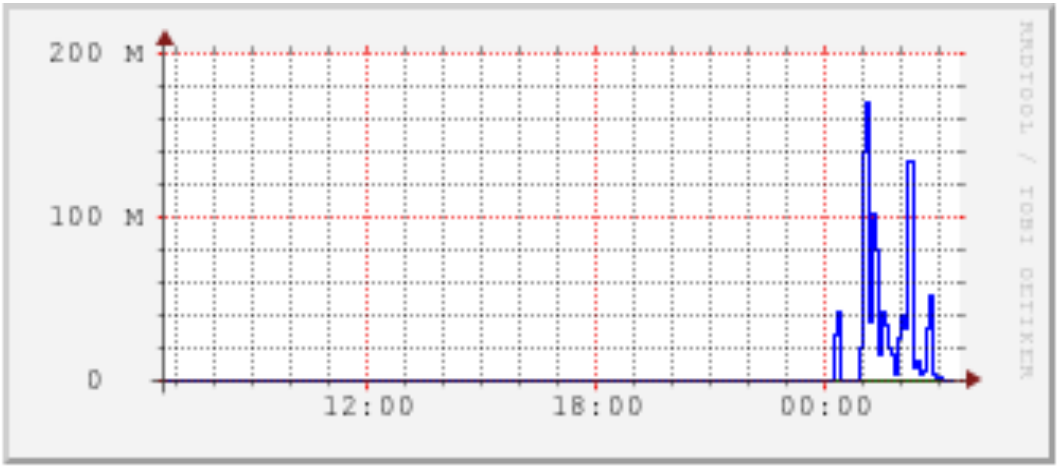
EVPN + Proxy-ARP

- EVPN, MAC state is BGP propagated, better than dataplane learning
- We saw benefits in reducing flooded traffic
  - With MAC/IP mapping known on all edge switches, they can be configured to proxy-respond to ARP requests, eliminating the need to flood the request, further reducing background traffic
  - For IPv6, we allocated address ranges based on AS number, not individual addresses so a bit more work required before making live.

MAC hold-down

- If a port goes down, the normal behaviour is the local switch removes its local MAC forwarding entry, then propagates it via BGP
- This is because the next hop interface in FDB has gone down.
- Until BGP has converged, the network is out of synch,
  - MAC is known at the entry of the network (the remote switch),
  - But not at exit (the switch with the port that went down)
- We implement rate limit of unknown traffic at the entry switch
  - That switch still knows the MAC, so does not rate limiting the unknown traffic
  - Still a lot better than pre-EVPN as its BGP convergence, not time-out
- Solution is to temporarily route MAC to /dev/null
  - Traffic to MAC is discarded for long enough to converge BGP

MAC hold-down GigE Members



- Graph on old LON2 during from Member migration
- Taken from 1Gig network monitoring port on old network
- This is 5 minute average. Initially went to line-rate
- Essentially short outage to GigE Members

Faster Reconvergence

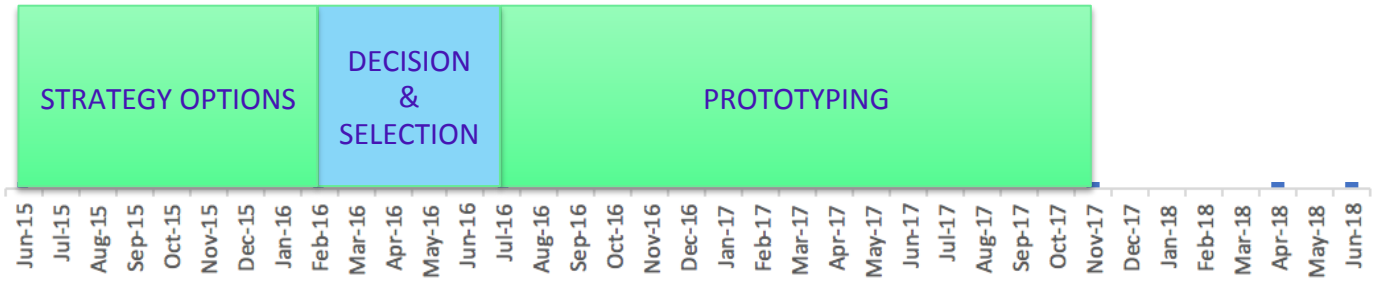
- Micro BFD run natively on ASIC
  - Maximum 4ms detection of failure, even single link in LAG
- OSPF timers tuned near to the limit, but not beyond
  - Extensive testing
- Software, and Topology designed to optimize push from control plane to data plane
- Mechanisms added (e.g. link-flap dampening) to detect network churn, and lock down topology

Benefits for all member sizes

- Convergence times benefit everybody
- Scalability, and faster provisioning targeted for large members
- Lower background traffic flooding targeted for smaller members
- Cost savings which can be passed through to members

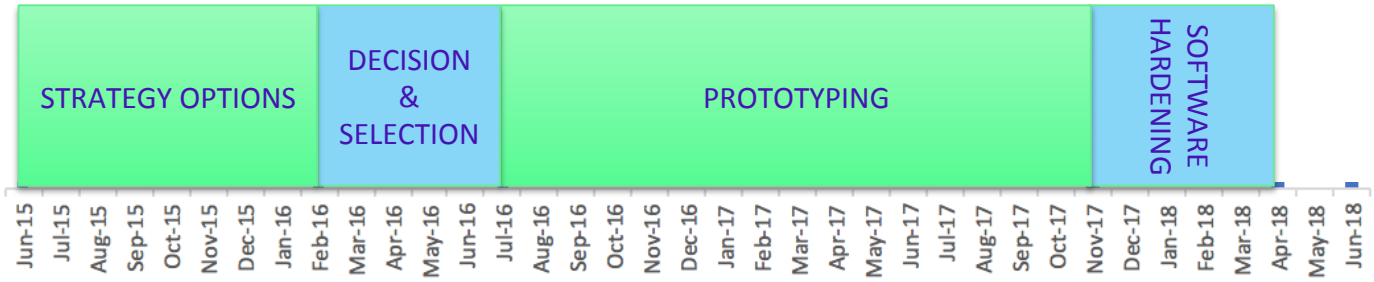
Project Steps

Prototyping phase



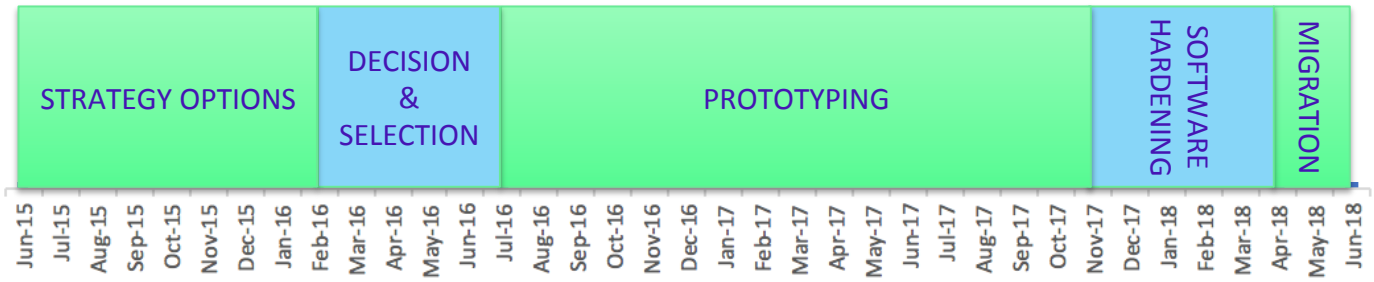
- Testbed provided in Taiwan by Edgecore
  - Features progressively added to solution
  - Different features at different level of maturity in each drop
  - Time difference meant that significant portion of mornings in 2017 spent on testing videoconferences
- Design evolved during this phase
- Significant focus on usability, manageability of features
- Very useful for developing test methodologies

Hardening Phase



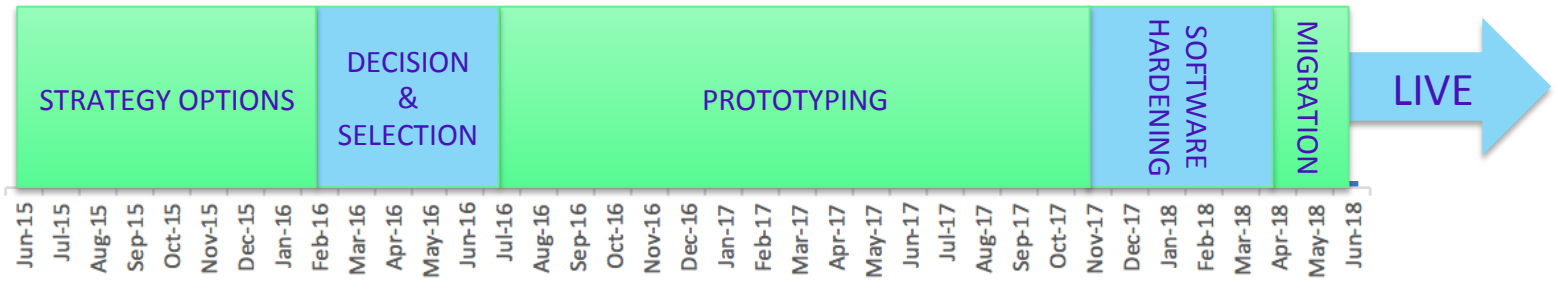
- Equipment purchased for project used to form Lab in London modelling target topology and configurations
  - Complementing multiple development labs in Taiwan and Bangalore
- Whenever bug was found, question asked was always how did that miss earlier testing, and where else where the same assumptions made
- Test plan instructions often deliberately vague methodology, and various iterations performed by different engineers
- If weakness found, tested in greater detail at next round of testing

Migration



- Started with one prototype site
  - Found one packet type not tested for that got mis-interpreted as control plane traffic
  - Caused two short network events
  - Luckily work-around could be directly programmed on Broadcom ASIC without change to Software.
- Also ran into bugs in old LON2 equipment
  - During the migration, it was in a slightly different state than before
  - These were more of a challenge than bugs on new network

Network LIVE



- Running, if anything better than hoped
- One software update to make temporary fixes permanent


It is now Live


linx
Questions?