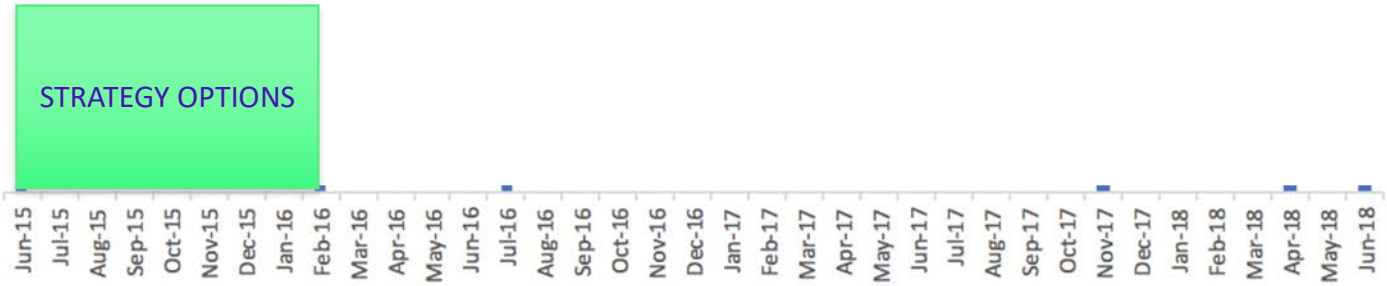
Deploying a Disaggregated Model for LINX's LON2 Network

How LINX reimagined its LON2 network architecture using EVPN routing technology

LON2 Refresh Project Background

- 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
- Were talking to existing vendor but at the time, did not fit their strategy.

Strategy Options

- Another gold plated LAN like our LON2, with traditional router vendor.
  - Costs and Partnerships were key concern
  - Use opportunity to jump to newer technologies
- Low cost layer2 solution
  - Would still be constrained by design and performance limitation of native Layer2
- Emerging Switch vendors
  - Half way between above 2
  - Not yet focused on IXP and service providers
- Same vendor as LON1
  - Cost still consideration
  - Perception of diverse solution of 2 LANs was concern
- Disaggregated Open Networking Solution
  - Promising in terms of cost and flexibility
  - But unproven in IXP and service provider space

We Looked for the Best Strategy

- Different vendors suited different strategies
- Traditional RFP, plus conversation with vendors to narrow down solution
- Selected best match for each strategy option
  - Tested solution
- 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
  - Recommendation was to be bold with LON2

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”

Why are IXPs different

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 its 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

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

The Port is the Demarcation

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

Early Steps

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 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

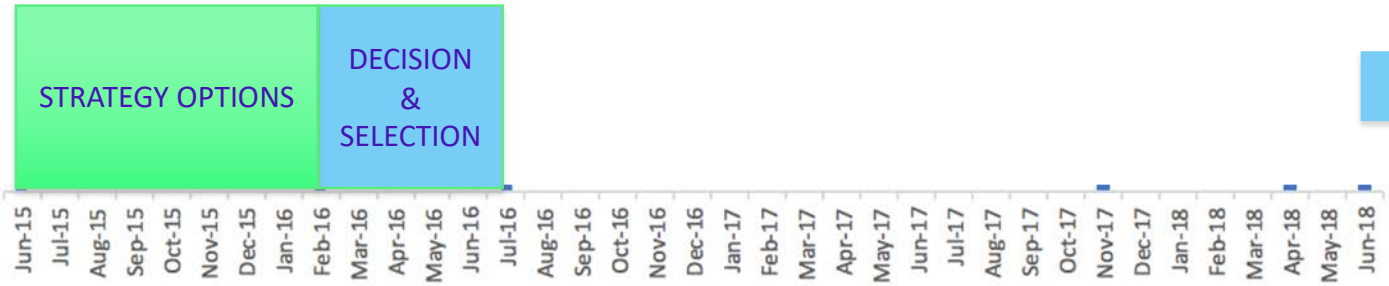
Agreed Target Solution Exchange Features

- 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

Broadcom TCAMs

State Memory on Broadcom ASIC

- If a policer is used, they use TCAM memory on the ASIC
  - Potentially upto 4 policers per member
- MAC ACL entries use TCAM entries too
- The Tomahawk only has 1024 entries for ingress traffic
  - By default - they are split into 4 buckets of 256 each
  - So only 256 Policer and 256 ACL entries by default
  - With our partner ports, we would run out of entries.
- Software can re-allocate TCAM resources by turning off capabilities and moving entries into shared features
- Pay attention to these!

Dynamic learning last resort

- There is no implementation decision on what order Broadcom performs operation
- So if dynamic learning is enabled, that happens before any ACL or rules to limit what might be learned
- If you switch on learning, and have loopback, probably have MAC Churn and dramatic drop in forwarding capacity (OUTAGE).

Broadcom StrataXGS

- Limit of how many Labels it can remove in one go
  - 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

Why Not StrataDNX?

- Alternate to Trident and Tomahawk
  - Best known as Qumran and Jericho
- When we started project, were not quite dense enough
- Buffer size was concern, but analysis was they would be enough
- External TCAM is a trade-off
  - Higher power consumption
  - Can have memory access challenges (especially for small forwarding tables)
- StrataDNX would not have been bad choice
  - They always were plan B

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

This is not a complaint about Broadcom

- They developed the ASICs that totally changed the market
- Fixed Pipeline, means fixed operations, but alternative is a lot more expensive
- Their main market is the Data Center market, so can not expect design to be optimized for our needs
- They have been very helpful and supportive
- They are working on Flexibility and Programmability

- **I hope my tone is more: pay attention to this detail**

Leaf and Spine Architecture

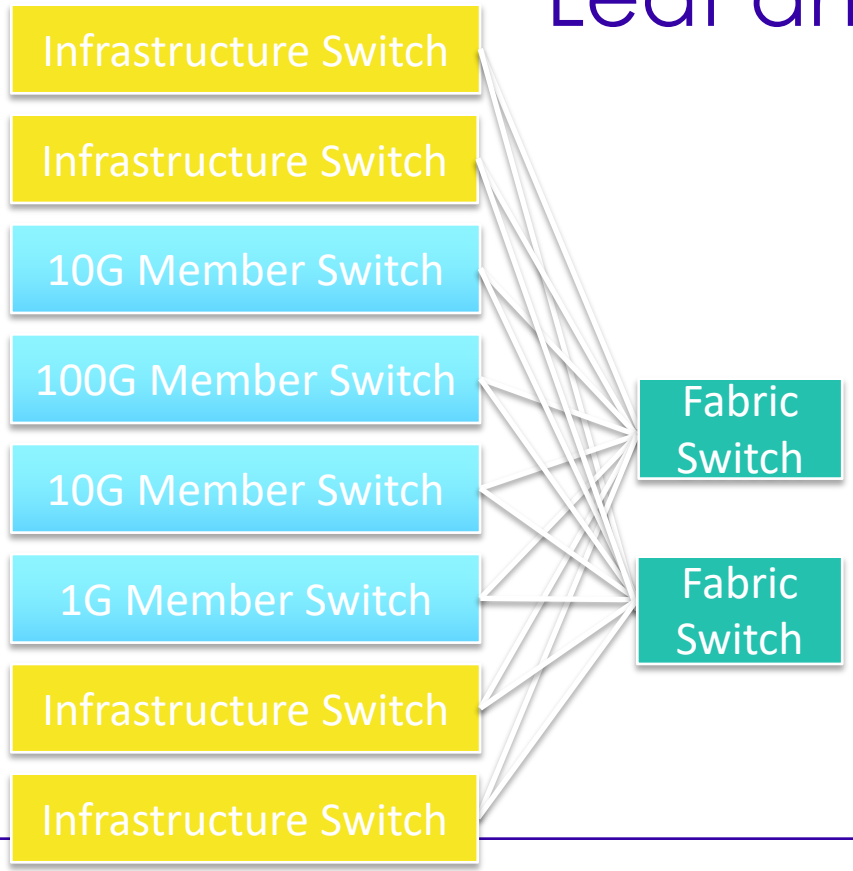
Leaf and Spine

- 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

Big Chassis

- This was the old approach
- Great if fits into chassis
- But Line Card #9 is a challenge
- Upgrading Fabric is a challenge
- All Line Cards must run same software
- **Scale up**, growth model, you scale by buying a bigger router/switch

Leaf and Spine



- Line Cards become leaf switches
- Fabric become Spine Switches
- If want more leaf switches, just install and connect to spine
- If want more fabric, can
  A. Upgrade Fabric 1 at time
  B. Add a 3rd, 4th, 6th
- **Scale out** model of growth, you re-use more of the same components

Leaf and Spine

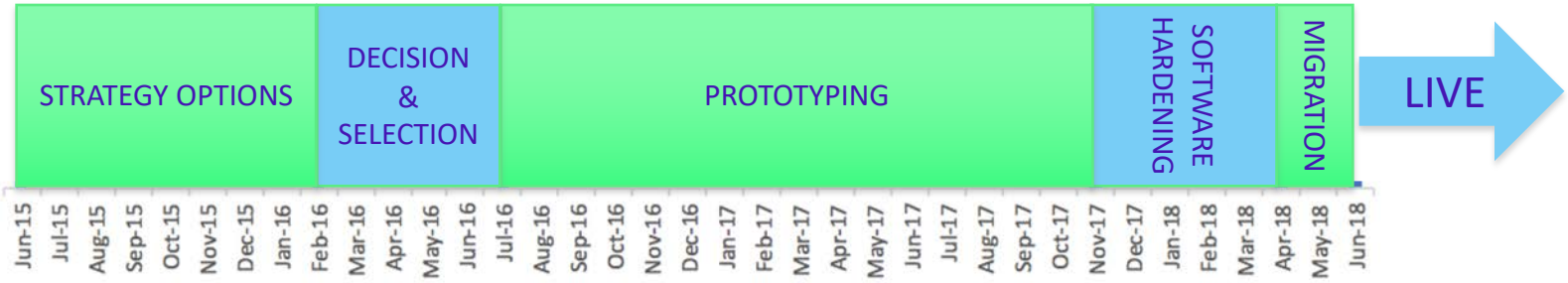
- Initially only two switch types, so easier sparing
- It also makes is easier to hold spare inventory
- Physical installation is easier
- Allows for much easier faster growth -> un-forecast orders less of challenge
- We are still at a scale IGP is not a concern

Benefits delivered for all member sizes

- Being membership based, ensuring benefits are felt across membership base is key.
- Convergence times benefit everybody
- Scalability, and faster provisioning targeted for large bandwidth members
- Lower background traffic flooding targeted for smaller bandwidth members
- Cost savings which can be passed through to members

Project Steps

Prototyping, Hardening and Migration Phases



The Network is now LIVE!

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

Questions?