Deploying a Disaggregated Model for LINX's LON2 Network

How LINX reimagined it's LON2 network architecture using EVPN routing technology

The first IXP in the World to do so.

About LINX

LINX is one of the largest Internet exchanges in the world

Connecting over

825 members from

80+ countries around the globe.

LINX members are able to **reach**

80% **of the total global Internet** with **traffic peaks of over 4 Tb/sec** on their public peering platform alone.

LINX operates a dual-LAN infrastructure in London. It also operates regional exchanges in **Manchester**, **Scotland** and **Wales** and in the Ashburn metro area in the US just outside **Washington DC**.

Dual LAN Platform in London

LINX's two London networks span in excess of

65Km

12 different locations, operated by four different data centre partners.

Digital Realty
Equinix
Interxion
Telehouse

Why did we require at network refresh?

2015 had seen a huge take-off in 100G orders

- We could see that we were going to outgrow existing chassis
- Core growth also would require reasonable investment

We needed to respond and be ready to meet the current and future needs of our members

What LINX wanted to Achieve

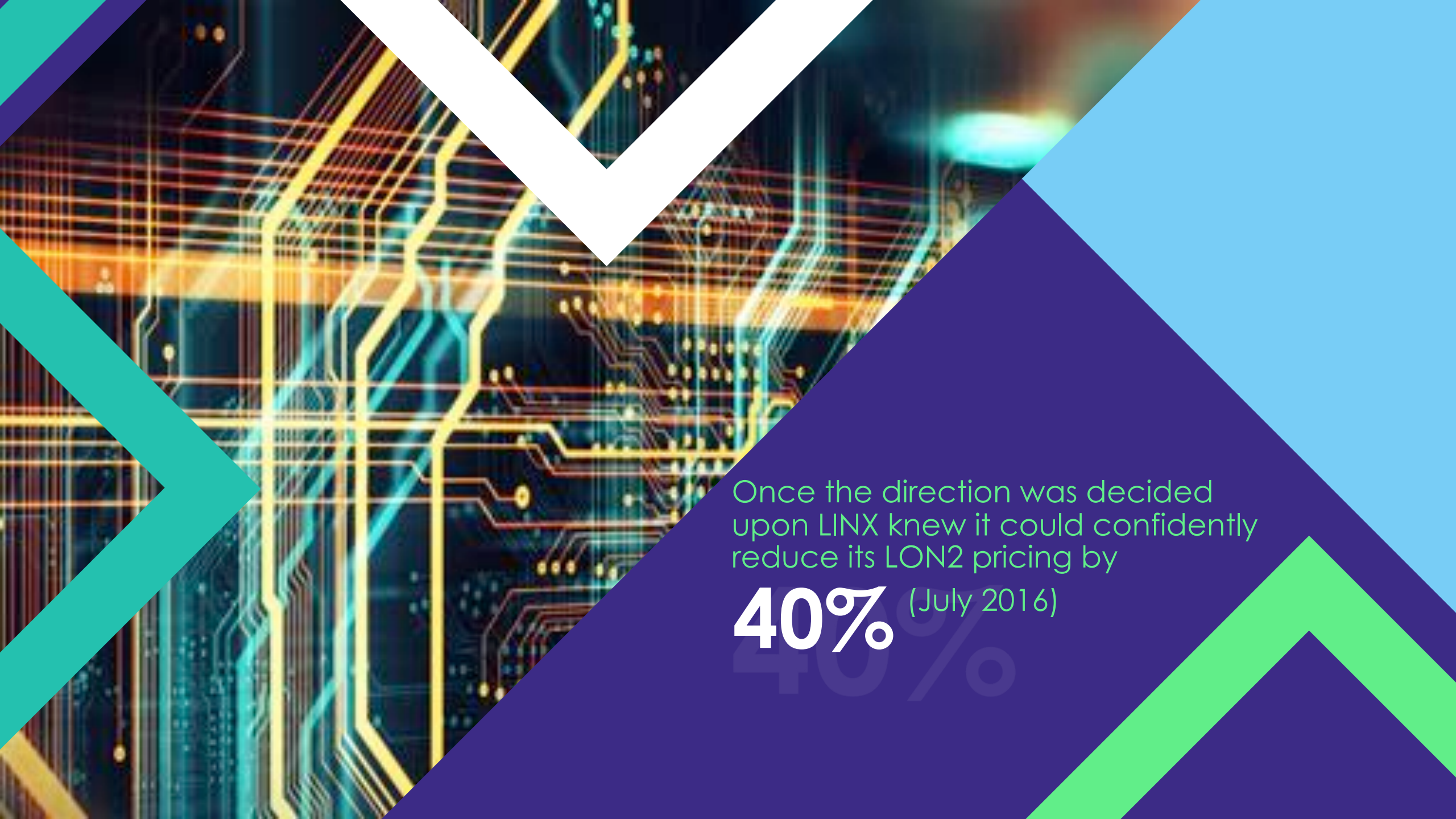
A network solution that offered more reliable delivery of traffic, faster convergence, less background flooded traffic plus opportunities for future developments

LON2 Infrastructure Review

The Background

LINX wanted a new architecture that offered choice, resilience and robustness for its

700+ membership (now 825+)

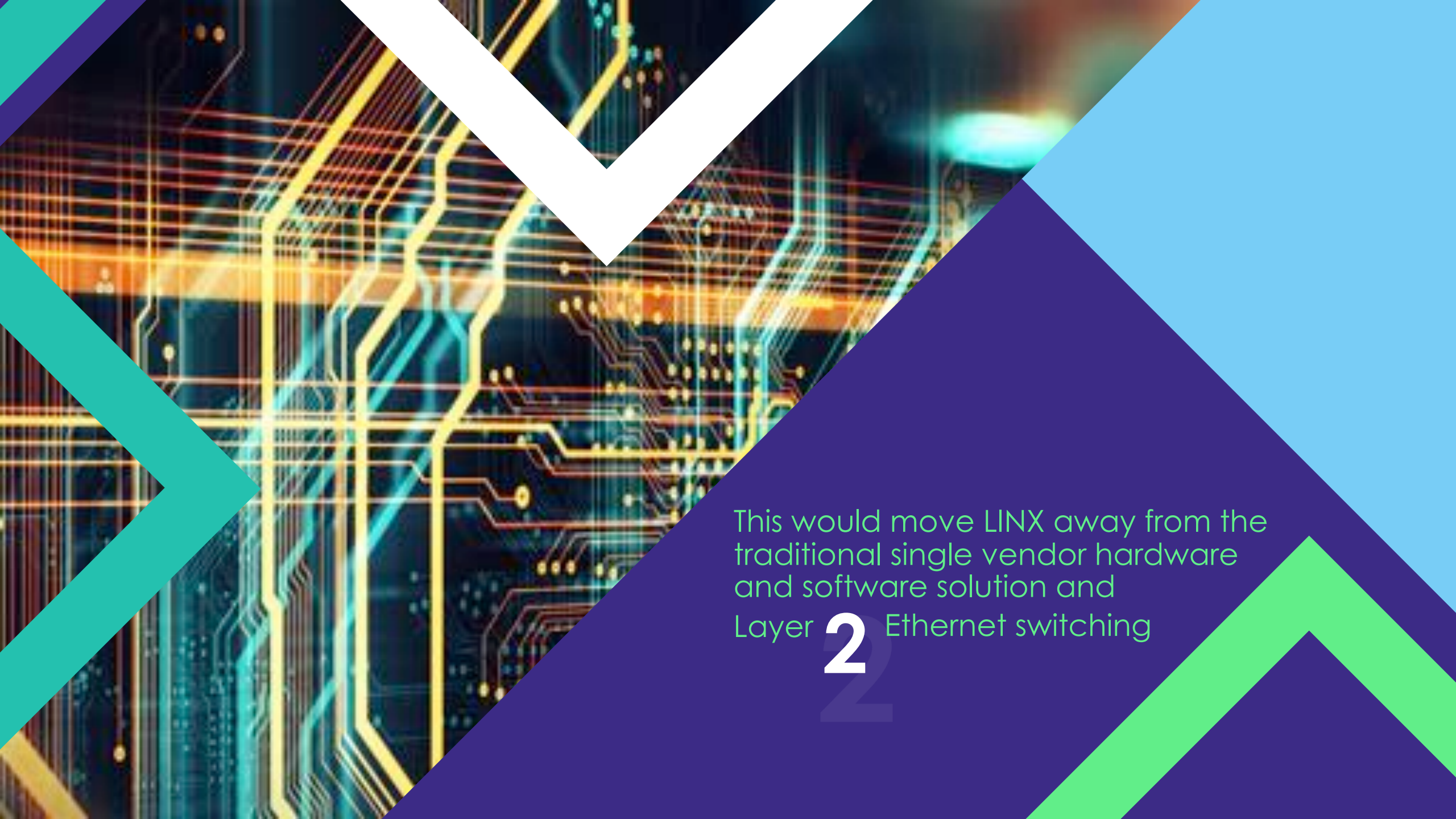


Once the direction was decided upon LINX knew it could confidently reduce its LON2 pricing by

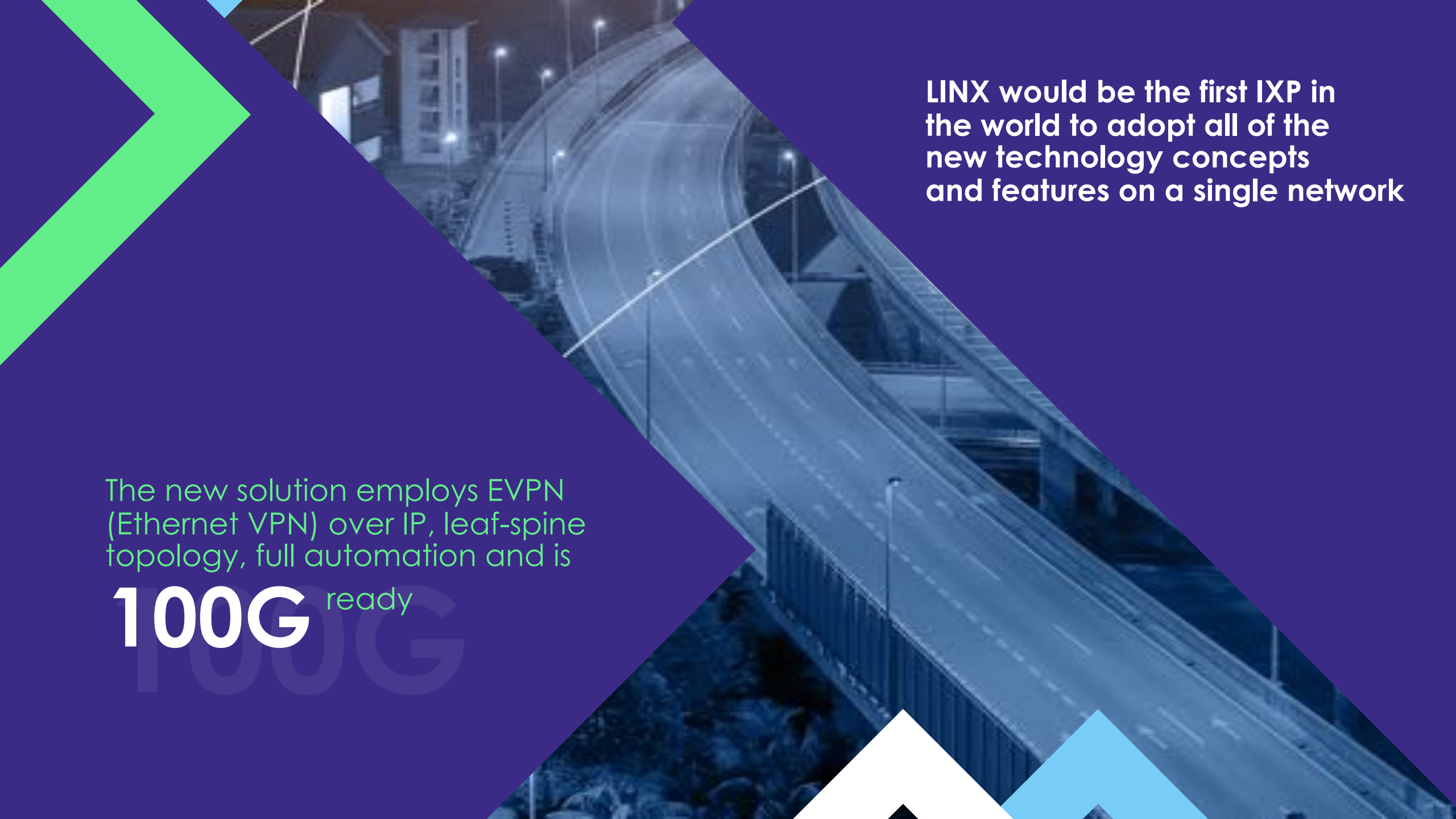
40% (July 2016)

After a vendor testing process an **improved technical solution** was found at a **significantly lower cost**

An extensive review of LINX's LON2 infrastructure began in **November 2015** in preparation for a major network upgrade



This would move LINX away from the traditional single vendor hardware and software solution and Layer 2 Ethernet switching



LINX would be the first IXP in the world to adopt all of the new technology concepts and features on a single network

The new solution employs EVPN (Ethernet VPN) over IP, leaf-spine topology, full automation and is

100G ready

Project Partners

Collaborative Process

LINX decided to adopt hardware from **Edgecore Networks**, owned by Accton Technology Group, as well as software from **IP Infusion**

Project Partners

Edgecore Networks delivers wired and wireless networking products and solutions through channel partners and system integrators worldwide for data centre, service provider, enterprise, and SMB customers.

Project Partners

IP Infusion was founded in 1999 by Kunihiro Ishiguro and Yoshinari Yoshikawa as commercial-grade, hardware-independent network software for IPv4 and IPv6

Disaggregation Explained

Disagregated in the router/switch context is a model where a operator selects a **generic switch** from one source, then **selects independently software** to run on that switch.

Disaggregation Explained

The traditional model involved buying fully both the switch/router hardware and software from a single supplier, the two tightly bound.

But the **server space has long demonstrated that need not be the case**, where you purchase the hardware from one supplier, and the software from a different source - allowing individual companies to focus on their strengths.

Disaggregation Explained

The approach allows an operator to **independently select suppliers to best meet their needs**. They might prefer the form factor or density from one hardware manufacturer, but the features from a different software vendor. And can **review independently** the choices as their requirements evolve.

Disaggregation Explained

By introducing a disaggregated platform, LINX members will benefit from **increased flexibility** plus continued value from their investment.

Disaggregation Explained

The disaggregated platform with Ethernet EVPN **allows LINX to play the long game** with the aim of delivering long-term innovative technology to members.

The New Technologies

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

Proxy-ARP

MAC hold-down

Faster Reconvergence

This is the topology that is emerging from hyper-scale data-centres

By having generic building blocks, and a "scale-out" model, we can build the networks with more headroom and scale it faster

The are therefore able to be more responsive to unforecast requests from our members

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

Proxy-ARP

MAC hold-down

Faster Reconvergence

In traditional switched networks, each switch had to snoop on the traffic to learn where to MAC addresses are located

Basic assumption is that if traffic from MAC address A is seen on port X, then return traffic to MAC address A must be sent to port X

- Behaves badly if there is asymmetrical traffic
- Behaves even worse in case of loop
- There is no mechanism to unlearn a MAC address other than waiting for long enough to be confident it no longer is at old location

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

With EVPN, MAC addresses, once learned or configured at the entry to the network are propagated via BGP,the same way as IP routes are propagated

Proxy-ARP

This ensures that all switches have a synchronised MAC table

MAC hold-down

If a MAC address is unknown, the switch can discard it as unknown everywhere, instead of flooding it just in case

Faster Reconvergence

It is possible withdraw a MAC address if a port goes down (or it's unconfigured)

The net impact is more reliable delivery of traffic, faster convergence, and less background flooded traffic

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

The same EVPN mechanism to propagate MAC information, and be used to propagate MAC to IP mappings (ARP resolutions)

Proxy-ARP

An mapping can be learned (snooped) or configured at the entry to the network and synchronised

MAC hold-down

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, and reducing background traffic

Faster Reconvergence

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

Proxy-ARP

MAC hold-down

Faster Reconvergence

If a port goes down, the normal behaviour is the local switch removes its local MAC forwarding entry, then propagates it via BGP

So until BGP has converged, the network is out of synch, where the MAC is known at the entry of the network (the remote switch),but not at exit (the switch with the port that went down)

But most networks implement rate limit of unknown traffic at the entry switch (which still knows the MAC), so the rate limiting of the unknown traffic does not occur

So at that point traffic is flooded to all ports on the same LAN on that exit switch - without any rate limit

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

EVPN makes this a lot faster, as it is a BGP withdraw (1-2 seconds), not a MAC time-out (of the order of minutes)

Proxy-ARP

MAC hold-down is a further optimisation on the exit switch, where instead of deleting the entry for the port that went down, it replaces it with an instruction to discard traffic to that MAC address (and hence not flood it)

MAC hold-down

That is kept in place long enough to ensure BGP has time to converge

Faster Reconvergence

No traffic flooded on local switch when port goes down

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

Proxy-ARP

MAC hold-down

Has more to do with the impact of a number of design decisions to make sure the network converges as fast as possible

We have fine-tuned BFD on links so failure is single links and dark fibers carrying multiple links are detected within 3 milliseconds

Faster Reconvergence

We have optimised OSPF to get it to reconverge as fast, without pushing it to the limit of stability

New Technology Key Features

IP Fabric leaf/spine architecture

EVPN control plane

Proxy-ARP

MAC hold-down

The topology is selected so the most likely failure modes trigger the fast reconvergence of the underlying Broadcom ASIC

The software is optimised to minimise traffic loss on link restoration

And there are mechanisms (eg link-flap dampening) to detect network churn, and lock down topology

Faster Reconvergence

Events such as failures and fibre cuts are much less visible to member networks

Future New Features

Future New Features

Multi-homing ports

It already is possible to have multiple ports treated as a single connection, but without multi-homing, they must be connected to a single switch or router.

How do multi-homing ports work?

EVPN Multi-homing is a standards based mechanism to extend link-aggregation bundles across several switches

The link-aggregation is allocated a **logical reference** that is shared between all ports participating in the link aggregation

The Logical Reference

The logical reference is used to communicate MAC address information.

So any MAC or equivalent entry learned on one switch is propagated to all other switches – **bound to that logical reference**

The Logical Reference

That logical reference is also used to communicate the status of the ports associated with that connection. **A port going down is removed from the logical reference**

The Logical Reference

By linking the two together, **all switches know all the locations of all MAC addresses**. But without the end switches being in tight synchronization.

The Logical Reference

Finally, the logical reference is used for handling of **broadcast and multicast** traffic to ensure that it is **not looped**

This allows the multiple switches to look as a single bundle to protocols such as LACP

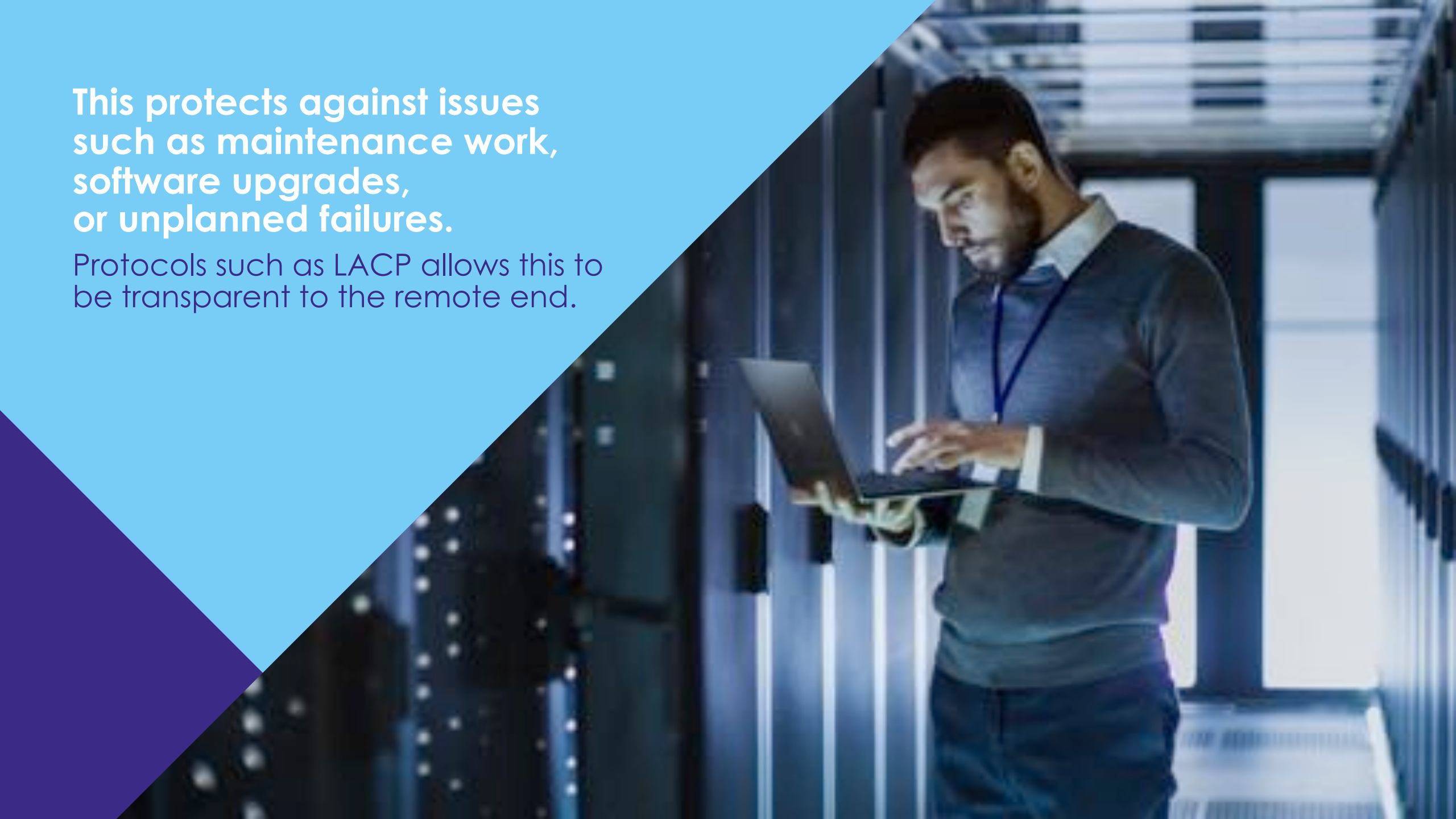
In case of failure, or maintenance, ports are removed from the link-aggregation, but state learned from those ports is maintained.

The load-sharing across the network is adjusted to the surviving topology.

At worse that would represent a drop to half the total capacity.

This protects against issues such as maintenance work, software upgrades, or unplanned failures.

Protocols such as LACP allows this to be transparent to the remote end.



The Process and Challenges

The Process

The LON2 migration process has taken two years but was broken down into phases

Demonstrator Phase (2016)

This was at the end of the vendor selection, where they demonstrated they could achieve our goals

The Process

The LON2 migration process has taken two years but was broken down into phases

Prototyping Phase
(late 2016 through 2017)

Iterative development where we incrementally test new features, and fine tune the requirements

The Process

The LON2 deployment and migration phases

Hardening Phase
(late 2017 through early 2018)

Finding and fixing the last remaining bugs

The Process

The LON2 deployment and migration phases

Deployment Phase
(early 2018) [Parallel to hardening]

Where we deployed the new network ready for migration

The Process

The LON2 deployment and migration phases

Migration Phase
(April-May 2018)

Made network live, and moved members across

The Process

The LON2 deployment and migration phases

Fully operational in June 2018

The Process

The LON2 deployment and migration phases

Enhancement Phase including new software releases
(late 2018 and beyond)

The Challenges Faced

Automation

Already adopted a fully automated configuration platform for LON1 and we wanted the same for LON2

The Challenges Faced

Knowing what we wanted and what was possible

The Challenges Faced

Testing

Making sure we eliminated all the bugs in the testing phase in readiness for the completion of member migration

The Challenges Faced

Operational Model

Moving from a development and PoC thought process to an operational mindset

What does this mean in the Market Place?

What does this mean in the Market Place?

LINX is one of the **Top 3** exchanges in Europe/world

While **LON2** is smaller than the LINX LON1 network, it is still **larger and more complex than many other European IXPs**

Having dual LANs in London **enabled LINX to be bold** in trying something new

All members will benefit from the new infrastructure

Smaller networks will see **background traffic reduced** on their ports and thus offering more value

All members will benefit from the new infrastructure

Larger networks will see **more flexibility and scalability** and be able to deliver higher capacity at lower prices

All members will benefit from the new infrastructure

Solution designed with **Automation in mind**

The Net Impact of Employing a Disaggregated Model

A solution that offers more reliable delivery of traffic, faster convergence, and less background flooded traffic

Questions

Thank you

Marketing@linx.net

07133 207705

Facebook.com/LondonInternetExchange

Twitter.com/linx_network

Linkedin.com/company/linx