Building an IXP

MENOG 7, Istanbul, Turkey
Andy Davidson

21st October 2010
NetSumo / LONAP / IXLeeds

Agenda

- Before you buy a switch
- After installation, before customers
- Turn up customers!

Draw **Plans**

Photo courtesy
Orange County Archives

The most important thing a technical manager at a start-up IXP must do, is plan the project. An Internet Exchange point is simple to build from a technical point of view, but in order to be successful, scalable, and maintain a good reputation, you must plan every aspect of the technology and service.

This presentation covers many of the technology requirements that you must consider when starting and building a new or young Internet Exchange Point.

Define your offering

Jason Jones - http://www.flickr.com/photos/jjay69/4050881930/

- Peering LAN
- Closed User Groups?
- Interlink own network?
- Transit OK?

Define the offering carefully, as this will have consequences for your configuration. You should do this before you buy a switch so that you know which features you need to support

Will you offer a single peering LAN, or will you offer peers the opportunity to have private VLANs between other customers ? If you offer multiple VLANs then you should make sure that your switch can support enough VLANs for your expected growth. If you do offer multiple VLANs, then will you allow your customers to interlink their own infrastructure? This would mean that - if you had multiple locations - you were filling up your inter-site links with non-peering traffic. It also means that you would be offering a transport service - does this compete with your intended customers ? Is it ok for your customers to sell each other transit via the exchange ?

There is no right or wrong answer to any of these questions, you should just decide what the answers to them at the beginning are (allowing customer to use your interlink, or sell transit could improve reliability of internet in your area ?). You can also change your mind in the future (but it is easier to start to offer something you did not at the start, than take something away in the future.)

If you do offer multiple VLANs, make sure that you can support all of the features you need (e.g. port-security - explanation coming up soon) on trunk ports as well as access ports. Or will you have a one vlan per port rule ?

Jumbo Frames

> 1500/1522

Peering LAN / CUG

ISLs

Also an important part of the service offering is whether you will offer Jumbo frames transport at the exchange.

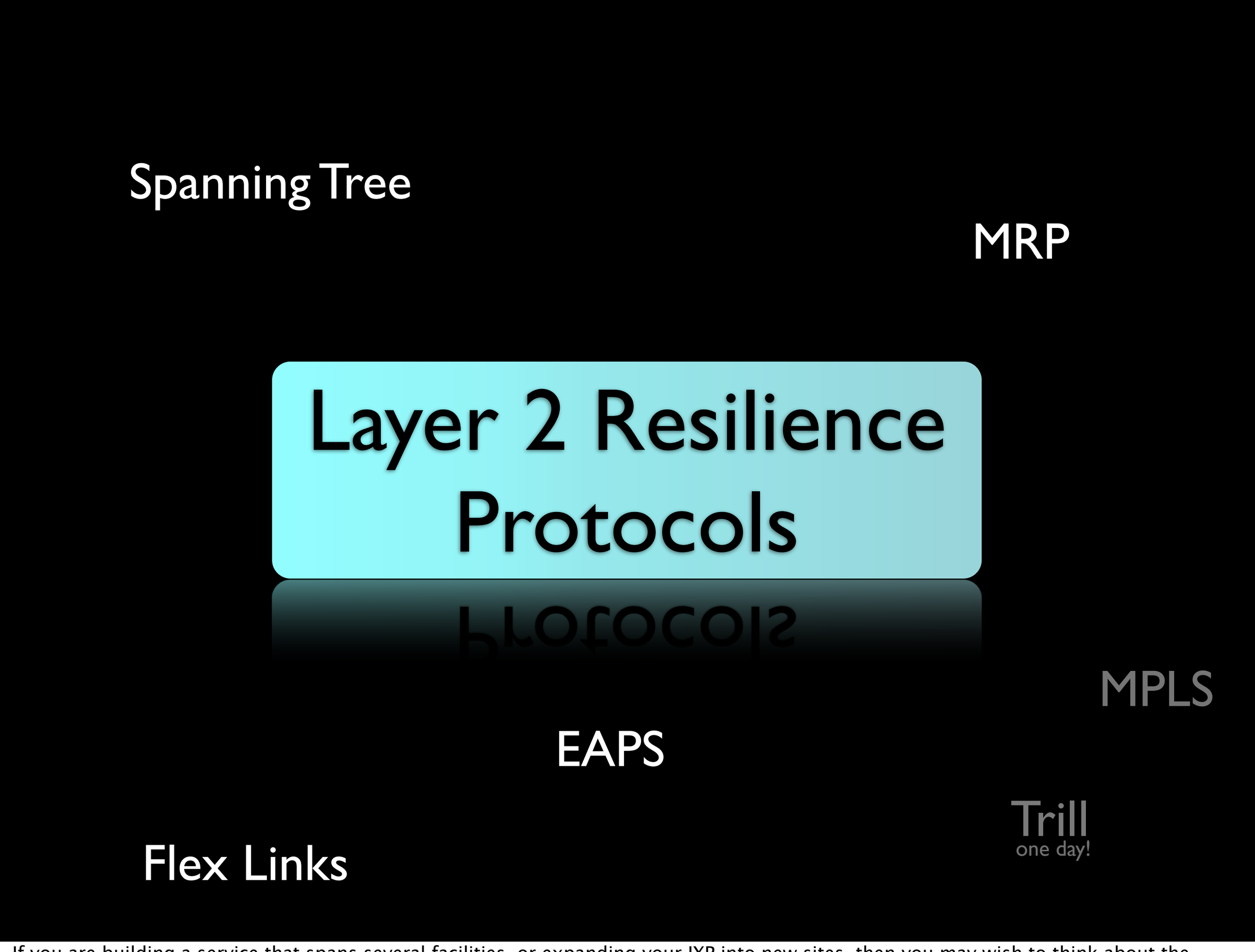
This is popular because it enables services between members, such as :
- storage
- dsl aggregation

But you need to consider whether to allow it on the peering LAN, or insist that users use a closed user group for Jumbo features.

You should check whether jumbo frames can be configured on a per-port, or whether it is a global setting. If you offer Jumbo for any service, then all inter-switch links need to be Jumbo, which means that if you use an ethernet based service to glue the exchange together, this must support Jumbo too.

Recommendation: set peering VLAN to 1500/1522, but permit members to exchange larger frames on CUG, set ISL to 9000.

See http://www.nanog.org/meetings/nanog42/abstracts.php?pt=MTI1Jm5hbm9nNDI=&nm=nanog42



If you are building a service that spans several facilities, or expanding your IXP into new sites, then you may wish to think about the resilient design of your peering LAN. Typically, exchanges build loops and then configure a layer two resilience protocol to shut down a single link in that loop to prevent an ethernet storm.

You may need to build a complex topology due to the location of your facility (and fibre availability), or in order to make use of resilient fibre providers in a region. In this scenario, you may wish to consider a resilience protocol like MRP which can have multiple/overlapping rings configured.

There are no significant differences between how you would design a layer 2 IXP network, compared with how you would design a resilient ISP network - inspect the maps of your providers, ideally select different providers, certainly select different paths/ducts/entry points to the building, run interlinks through different ducts inside your colo, obtain a contract (even if the service is provided to the IXP for free).

See AMS-IX presentations on the use of MPLS forwarding rather than a traditional flat VLAN.

Port Security

‘n’ MACs per port

Storm Prevention

Leaky ISP Layer 2

Mike Zebble - http://www.flickr.com/photos/zebble/6080622/

Port security is essential to the health of your new internet exchange point.

It is simple for ISPs to accidentally create a loop facing the Internet Exchange point. Where an ISP transports internet exchange points from different regions to their router in a certain town, it is also simple for them to bridge together multiple exchange points. Further, it is easy for the ISP to mistakenly expose their own layer-2 to the exchange. Easy for telco to loop fibre to exchange.

Port-security guards against all of these failure modes by limiting the number of MAC addresses that a switch will add to the CAM table, on any particular port.

- Only configured on customer facing ports, not inter-switch link
- Typical to see ‘1’ MAC only, but some exchanges permit 2, to allow smooth customer router-swap, or layer 2 keepalive protocol (urgh!)
- Exceed modes can be ‘drop frame’ (polite, but can be harder to debug), or ‘shut port’ (firm and effective..., make sure you monitor for this)

Recommended configuration: enable on all customer ports, with dynamic (learning) MAC mode, shutdown on any violations. **This will be a support burden**, but a broken ixp is a larger burden!

Also normal to see DHCP guard (DHCP is often prohibited traffic at an exchange anyway). RA Guard in standardisation process. Spanning tree/BPDU filter useful.

Make sure that your switch vendor allows you to run the port security features you need - example problems :

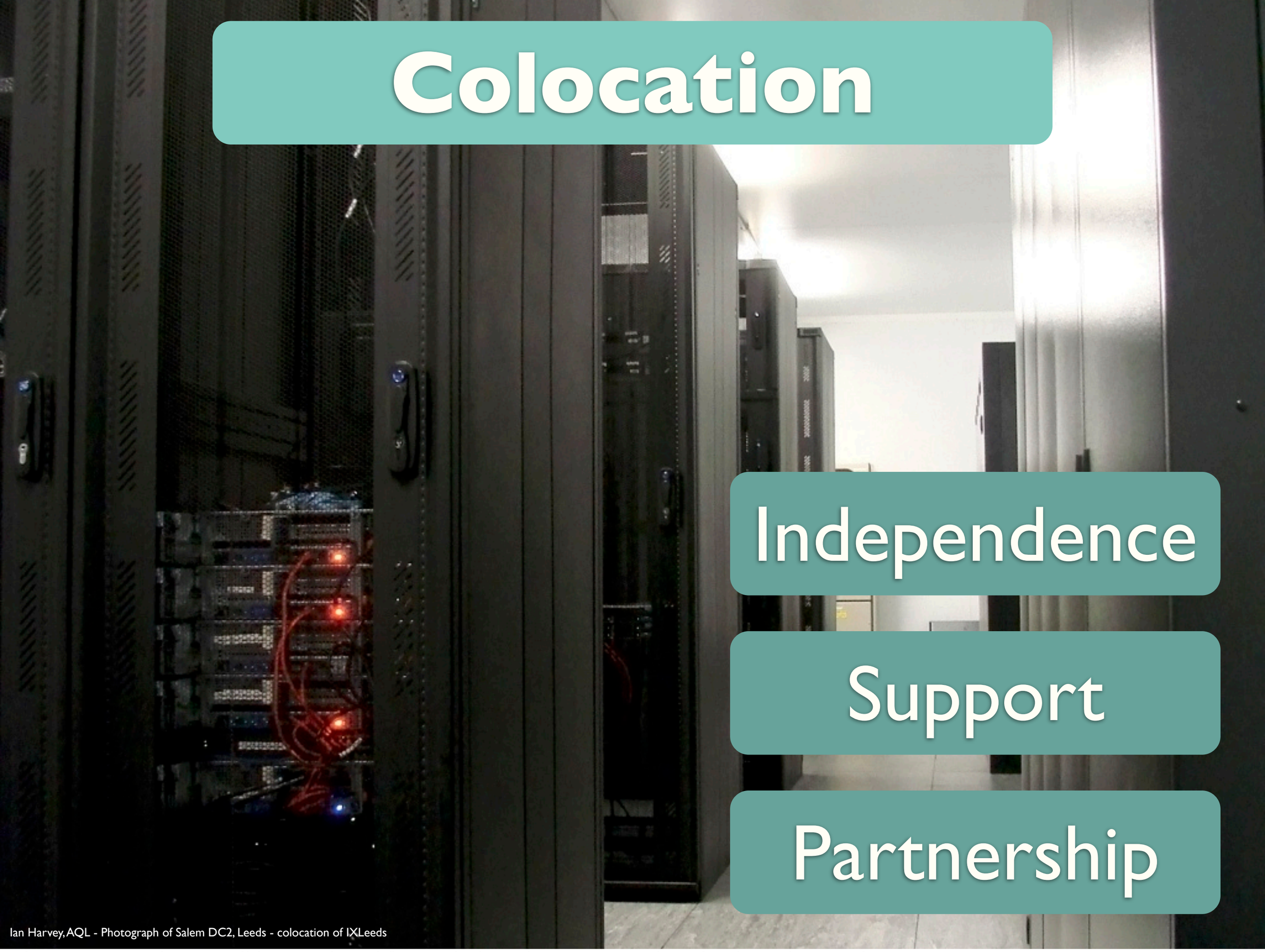
- no shutdown mode out of the box on extreme
- Juniper EX can only do port-security on access, not trunk, port.
- Juniper EX chassis can not do Port Sec (on roadmap)
- Cisco 6500/Sup720 can not do Port Security on etherchannel

Other Switch Features

Here's a photo of LONAP's 6500 - not necessarily advocated as an exchange switch - but has done a good job for us for some time.

Cut through vs Store and Forward - Mixing port speeds on cut-through switching can cause buffering problems.
Mac learning - Dynamic, and configured
DHCP / Spanning Tree filter (BPDU Guard)
Layer 2 / 3 inspection ?
SNMP and Management features
Port mirroring
Upcoming 40GE / 100GE support
Resilient core components - power supplies, supervisor
sFlow
Unknown Unicast control
Link Aggregation
Optic Locking - Support for WDM optics
Vlan Translation
DOM / optical monitoring - avoid DC trips to test cables!
PORT COST - 1GE / 10GE - related to port density

Colocation



Ian Harvey, AQL - Photograph of Salem DC2, Leeds - colocation of IXLeeds

Independence

Support

Partnership

The internet exchange will have a shared fate with the colocation providers that host the switch :

- Successful IXP makes a colocation desirable
- IXP will rely on excellent colocation, so that service is reliable and customers are happy

European IXP model is usually one of independence from colo - very successful. However, this relies on the good **partnership** between IXP and Colocation. Must stress value of good partnership, ensure they understand what you **do**.

It is easier to set expectations with the colocation at the beginning of the project rather than when the project is rolled out.

Examples of things which make partnership with colocation good :
- Cheap cross-connects for IXP participation
- Donated or cheap rack space, power
- Marketing support

Challenges / strain on partnership with colo :
- Change in staff = replacements see you as any other customer
- Density of cabling to IXP rack
- Treating colocations evenly when you are in many sites. What information do you publish ?

Collector
(BGP Router)

Debug

Monitor

Support

Install a collector at your Internet Exchange Point.

A collector is a BGP speaking router which :
- Announces the internet exchange's own routes
- Has a unique, public ASN
- Peers with all of your customers, mandatory

This allows you to test your customers' BGP configuration before putting them in the 'wild' for the first time, and it also gives you a perspective to monitor the health of your customers (and by extension, the exchange), and a perspective to understand your customers' issues when trying to support them. You can run other monitoring systems, e.g. test for prohibited traffic, via the collector (more on this topic soon).

On a brand new build, make sure the collector BGP software supports ASN32.

You can use a Linux/BSD based BGP implementation for this, since the forwarding requirement is very small.

Monitoring

Availability

Performance

Ethernet OAM?

Monitoring is important. I recommend that you monitor :

- Availability of switches, collector
- Availability of member ports, bgp sessions
- Route servers (if offered) [Example - Nagios]

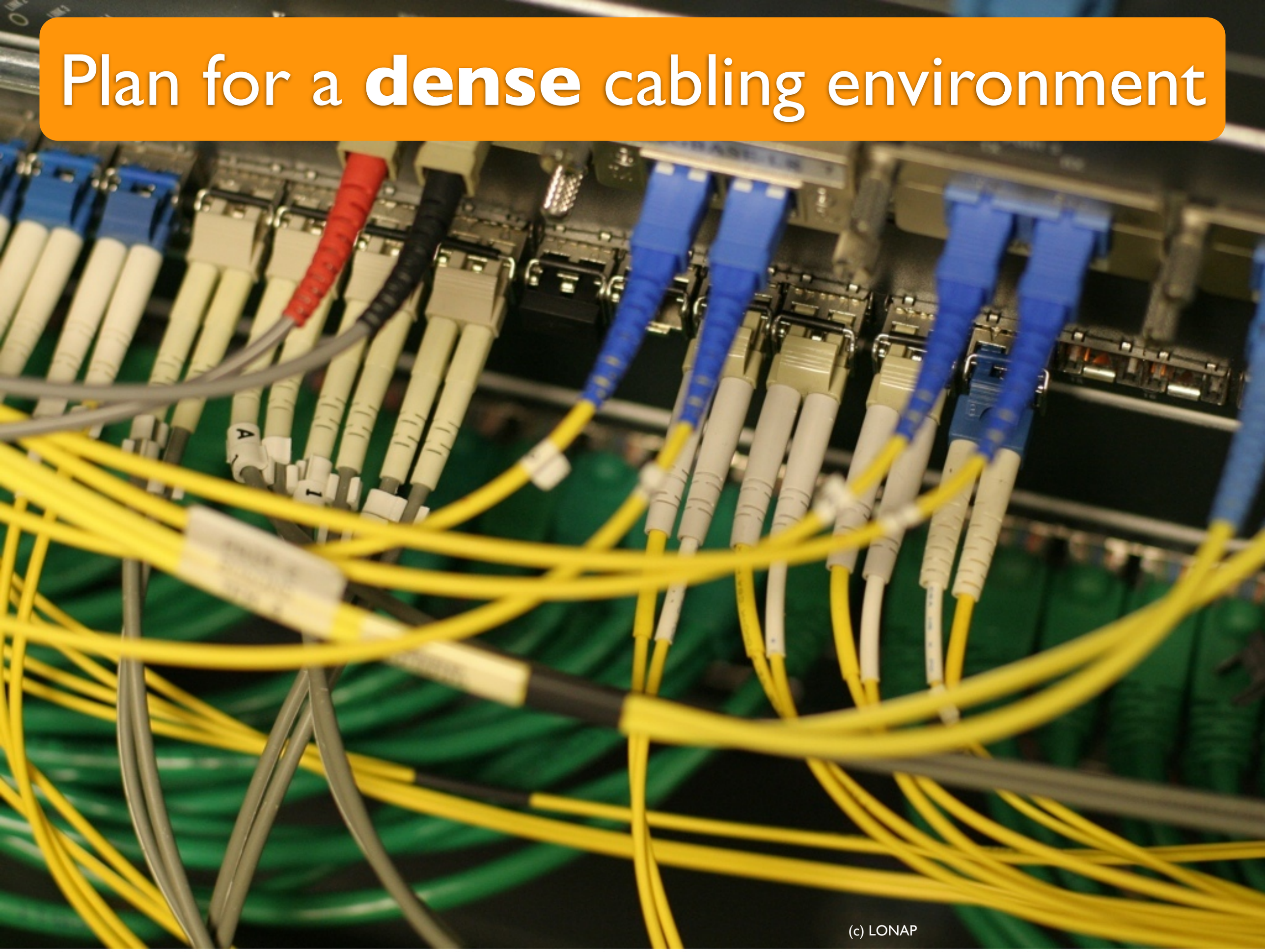
- Traffic on exchange LAN (excellent for marketing purposes!)
- Traffic on customer port
- Dropped frames / congestion on customer port
- Dropped frames / congestion on ISL [Example - Cacti]

- Latency between exchange infrastructure and customer/member port [Example - Smokeping]

Feed this data into exchange support platform - exchange staff irc channel or similar !

Ethernet OAM is a new protocol for monitoring and doing remote fault detection.

Plan for a **dense** cabling environment



(c) LONAP

Plan for a dense cabling environment. A successful exchange is the aim of everyone here, but what will you do if you are managing 200, 300, or 400 cables ? Or more ?

Pre-cable what can be done in the rack, and plan a cable database. Will you manage the cabling or the end users ? If you manage the cabling, then you have the power to resolve faults quickly. This may cost a lot of money, however.

This is not a lesson in keeping cables tidy, the message is to plan for a big environment, before you have a problem with scale.

SFlow

Customers

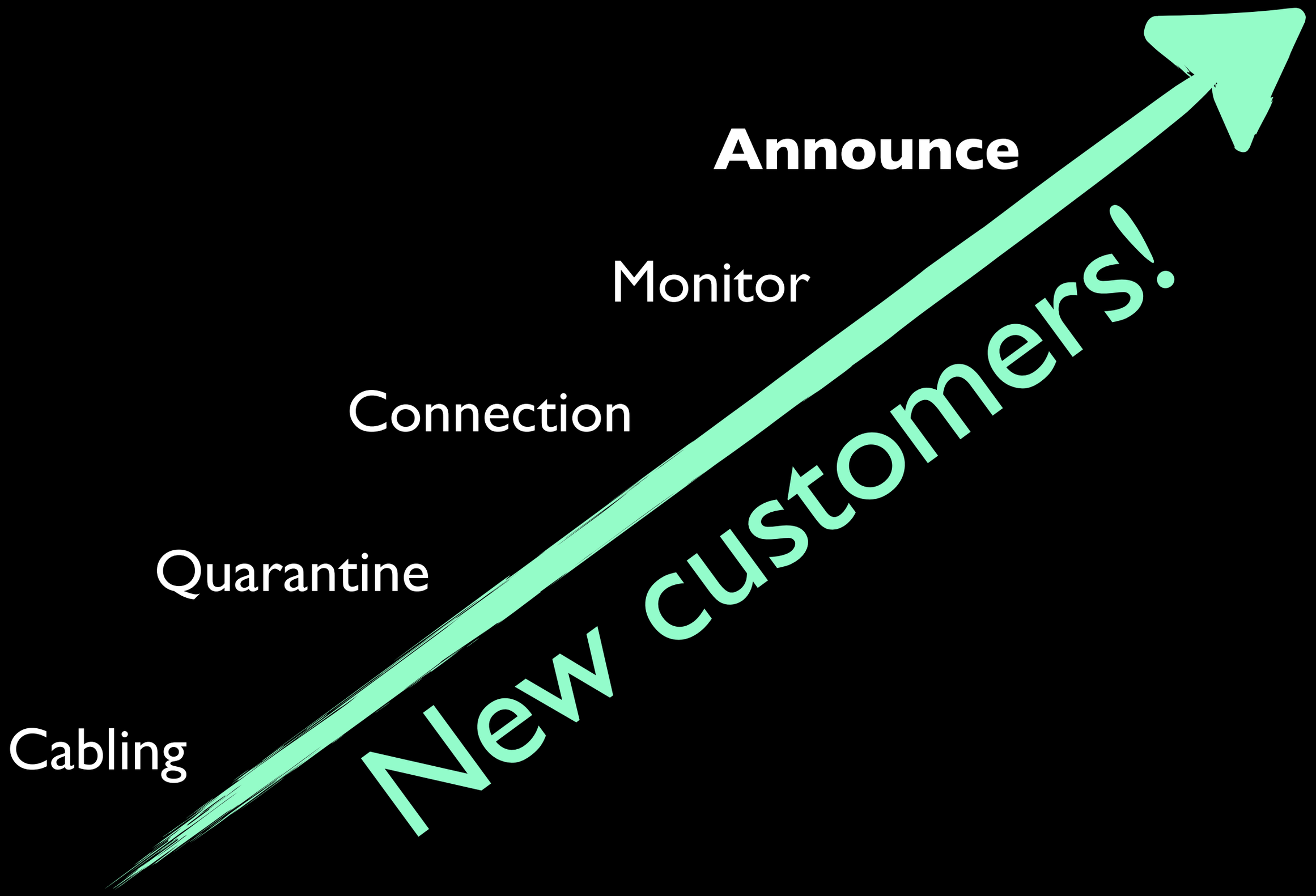
Top Peers

Exchange

Marketing
Resilience
Missed Ops

Internet Exchange users adore s-flow, because it gives them insight into their top peers at an internet exchange (and possibly move peers to PNI, backup exchanges).

Running an s-flow service is also of benefit to the internet exchange, because it gives you an insight into who is peering with who for marketing reasons, also allows you to ensure that the main flows are on protected Interswitch links, and also look for missed peering opportunities.



This is the rough process which I recommend for turning up new customers/members.

Cabling – who is responsible for cabling ? Customer / Exchange / Colo ?
When connected, add the port to the Quarantine VLAN (more in a moment), so that tests can occur
When quarantine is complete, connect the customer to the main peering LAN
As soon as connection is complete, add to monitoring systems
Make sure you ANNOUNCE new connections, so that other participants know to turn up peers.

Quarantine Process

New customers

Naughty customers

Prohibited Traffic

BGP Clue

Magnus http://www.flickr.com/photos/magh/2159613408/

Install a Quarantine VLAN. I recommend that you put the collector on this VLAN (as well as the production VLAN too), so that you can sanity check every single new customer's behaviour :

- They do not trip port-security
- They are not emitting banned link state protocols, e.g. OSPF
- They are not emitting banned layer 2 traffic, like CDP/decnet...
- They have to speak to you when service commences (check BGP in real time)

Get them to configure the collector BGP session when they are in the quarantine VLAN. It will not establish, but this will have the effect of putting some real traffic onto the interface, so that you can check for good compliance against your technical policy.

The purpose of the quarantine is to help customers with the - sometimes daunting - process of connecting to an exchange and peering, rather than punish them. Be supportive when your new customers fail the quarantine process.

You can drop customers who start to emit prohibited frames into this quarantine as well, so that you can check for compliance before letting them out. (In practice, friendly hints are going to be a more polite first stage!)

1 2 3 4 5 **6**

Offer **IPv6**

From day one

Same service level

It is extremely easy to offer IPv6 as an internet exchange point.

- Get resources
- Add to collector
- Encourage customers to peer

You can derive customer address space from as number or customer id number, e.g. 2001:7f8:xx::[id]:1 - id can be AS or other identifier. xx is assigned by RIPE (Internet exchange assignments come from this /32, each IXP gets a /48 each.)

“Same service level” means you monitor the same things as you do for v4 (sessions, etc.)

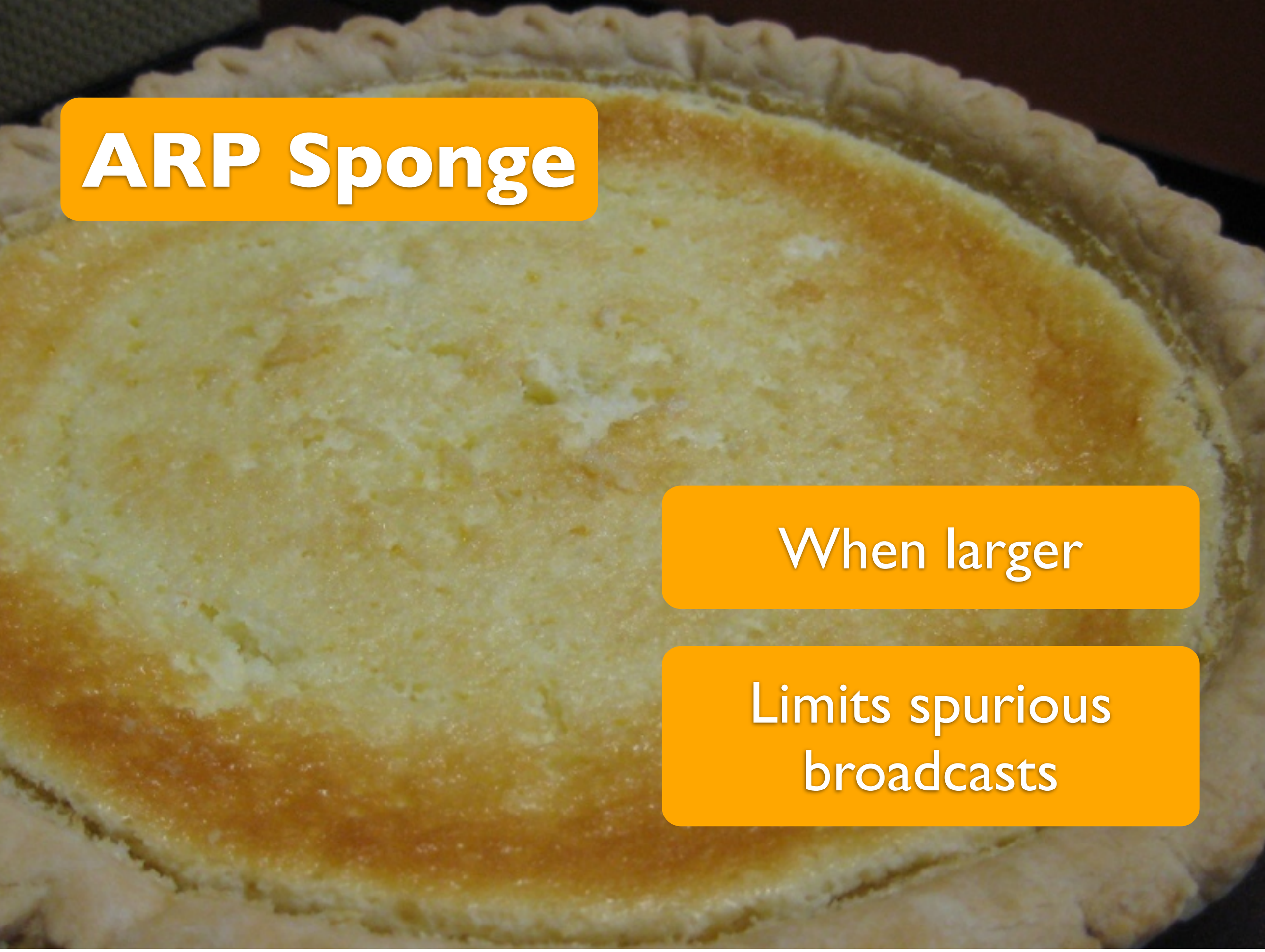
Used to be trendy to deploy v4 and v6 peering in different VLANs, so that it was easier for people to monitor v4/6 traffic - recommend against this, since it led to poor adoption in places where tried (e.g. LONAP - take up was much higher when this restriction was removed.)

Congestion is the exchange point's enemy. It manifests as dropped frames. It can normally occur in two places :

 - Member ports - give them tools to monitor their port utilisation, and be proactive about contacting customers about upgrade paths. Offer affordable upgrade paths (Nx1GE, fractional 10GE, as well as 1GE->10GE). Congestion can also appear in the customer's backbone, if they are transporting the peering port from one city to another, or using a third party metro ethernet carrier to deliver the port.

The way that your member reacts (upgrade, move traffic away, shout) will depend on the local culture - i.e. do they even care that they are congesting ? Is a congested exchange port better than a severely congested transit port ? In an ideal environment, member ports will be congestion free.

 - ISLs - try to never allow these to congest - talk to your members to understand the flow of traffic, where they need to deliver traffic to in a hurry, what they are using the exchange for, so that you can upgrade the ISLs in plenty of time. Try to buy Dark Fibre/Wavelength products, rather than carrier ethernet products to join your facilities together, so that you are responsible for the delivery of all frames, and not reliant on some other third party switch vendor.



ARP Sponge

When larger

Limits spurious broadcasts

When your exchange starts to scale, you may wish to look at installing an ARP sponge.

This is an automated software tool which catches ARP packets for IP addresses that are offline. The AMS-IX have an open source arp-sponge which is based on the number of requests for a hwaddress in a given period.

Any Questions?

Thank you to :
Will Hargrave
Tom Bird
Sebastien Lahtinen
of LONAP for their editorial support!

andy.davidson@netsumo.com
+44 20 7993 1700
www.netsumo.com / www.andyd.net
twitter: @andyd

Thank you for listening / reading. :-)