Internet Exchange Point "Wishlist"

Version 4.0.4
November 2013
http://www.euro-ix.net/ixp-wishlist

Mike Hughes
mike@ethernorth.net
Ethernorth Consultancy

Martin Pels
martin.pels@ams-ix.net
Amsterdam Internet eXchange

Harald Michl
harald.michl@univie.ac.at
Vienna Internet eXchange

1 Abstract

With the continued growth of video and cloud based services, Internet traffic continues to grow. Public peering via Internet Exchange points continues to grow in terms of traffic carried and the number of ports required.

The IXP operators feel that having the right tools and features implemented in the equipment they deploy will play an important part of scaling Ethernet technology to continue to meet the demands placed upon Exchange Points.

This is an informational document to outline the various features which IXPs would like to see implemented in core Ethernet (Switching) products.

In general the wishes defined in this document should be understood as logical AND wherever it makes operational sense. That means all features on the wishlist should be configurable and usable at the same time and not exclusive or.

2 Background to Version 4

The “Internet Exchange Point Switching Wishlist” exists since about 2000 and started within the RIPE EIX workinggroup. At the RIPE63 meeting in Vienna Andy Davidson (Working Group Chair) brought up the idea to update the then current version 3.0.2 from July 2005. Version 4 was then released in 2012 and after the RIPE EIX workinggroup dissolved at RIPE67 it was moved to Euro-IX. It now tries to again focus on the current demands of Internet Exchange Points, primarily in Europe, but also globally.

While in the past it was quite common to speak about Exchange Point **Switch** Infrastructure the current implementations show that the requirements for running an Exchange Point have changed.

In order to scale their business, IXP switching devices are now using functionality that lies in the realm of routers, rather than switches, such as implementing the Exchange Point using a VPLS instance. Using a VPLS service to implement the peering fabric has also allowed a concept of what some IXPs call a “virtual connection”, whereby multiple remote participants can be delivered to the IXP over a single physical port, mapping these into the same VPLS domain. This changes the “one router, one IX port” paradigm which dominated IXP provisioning to “one router per VLAN per IX port”.

Although features such as MPLS are now being used by the IXP operator, from the participants point of view the Exchange Point is still always seen as a plain Layer-2 Ethernet broadcast domain.

BGP is the standard routing protocol used at Exchange Points and two ways of routing information exchange are currently common:

- direct bilateral peerings between members
- peerings via BGP route servers operated by the Exchange Point

In both cases the Exchange Point operator does not become involved in the routing of any traffic across the Exchange, but leaves routing decisions in the hands of the participants.

3 Physical Wishes

IXPs are high-uptime environments. The equipment used in an IXP needs to be able to satisfy this requirement, in terms of inbuilt redundancy, and through use hot-swappable components.

- Hot swap of management and switch fabric cards with instantaneous failover to any installed redundancy (not rebooting onto the backup management processor or fabric)
- Hitless upgrade of software/software components without forwarding impact on the system - “nonstop forwarding”
- Full 2xN redundancy of PSUs, and hot-swap (i.e. a device should be able to operate using 50% of its PSUs). This is to allow for 2xN redundancy of incoming power feed
- Minimised booting and startup process time – preferably below one minute, the faster the better and minimise the need to completely reboot the system (e.g. graceful restarts, hitless upgrades, etc.)
- Pluggable transceivers using “state of the art” form factors for media type flexibility and easy field replacement
- "Colored" (DWDM/CWDM) optical transceiver support
- Support of tuneable optical transceivers; i.e. the wavelength/color of a physical optic WDM module should be configurable
- Support of single-fiber optics
- Vendor "lockdown" of pluggable transceivers to their own badged optics should either not be implemented, or it should be possible to switch it off in the device configuration, without “crippling” of certain features such as optical monitoring
    - Alternatively, provide a list of supported third-party pluggable transceivers
    - Consider charging a fair price for vendor badged optics
- 220-240V AC power options. Unlike most telco-managed facilities, the carrier-neutral facilities common in Europe do not provide indigenous 48V DC power. Power distribution is done using the regular utility supply voltage in that country - usually 230V AC in EU countries.
- Cable testing functionality in copper ports
- Optical ports have to support Digital Diagnostic Monitoring (DDM) / Digital Monitoring Interface (DMI)

In chassis based systems, the following are major considerations:

- Front to back cooling is preferred
- Vertical orientation of slots usually makes for easier cable management
- Cable management must be taken into account when designing the system - it should be possible, if cabled correctly, to remove one module for maintenance, without affecting the cables plugged into the adjacent card. For this, cable management brackets are preferred, and it is significantly easier if the cables are routable from the top of the chassis.

4 Management interfaces

CLI or web interfaces should support authentication using username/password or user/key pairs, to avoid the use of "password only" authentication, which implies shared passwords.

Clear text password-transportation should be avoided.

The following protocols are preferable for management access:

- SSHv2 (username/password or public key authentication) must be available
- NETCONF (RFC 6241) should be implemented
- HTTPS/SSL enabled web-access should be implemented

Should support SCP/SFTP for config copy/upload/download, as well as existing methods (TFTP/FTP).

Management interfaces should be able to perform authentication from an external source, such as TACACS, RADIUS or LDAP services, as well as providing locally held accounts which provide an authentication of last resort. This has to be retained for emergencies, if the system cannot access an authentication server due to a network incident.

It should be possible to create view-only users, as well as users that can configure only a sub-set of the device’s configuration. This would preferably be configurable per interface (allowing some users to only modify configuration related to customer ports).

All management interfaces, CLI, web and SNMP must be able to benefit from access-list control. The access lists must be able to support variable-length subnet masks.

Ability to create and disable management interfaces on a per-VLAN basis. Many IXP operators choose to configure a "management" VLAN, so that all management is done out-of-band of the core peering traffic. It is desirable to have the management interfaces to listen on the management networks only.

Devices should have dedicated management ports, preferably with a selectable interface type (copper / fiber) to allow remote-ends of a management port to be physically remote, connected over dark-fibre for instance, with no other active network element in between.

Support for both IPv4 and IPv6 for all types of access including management, SNMP and so on.

All CLI configuration options for the device should be made available through a machine-operable interface (e.g. NETCONF) to support configuration automation.

Clear command hierarchy. Use clear syntax between global context commands and port context commands. Commands made at the global context should not override commands made at the port or protocol context.

It should be possible to use one command to reset a port or linecard to its default configuration - after confirmation or only in a specific setup-mode.

5 Monitoring

It should be possible to gather information about the equipment during normal operation.

The source of this information can be any or all of:

- snmp (version 3 must be supported)
- snmp traps
- syslog

For each protocol, the secure (encrypted) counterpart should be supported.

5.1 Environmental Monitoring

There should be reasonable environmental monitoring provided:

- Temperature sensors per chassis and module (linecard, powersupply, etc.)
- Fan health sensors plus rpms
- Power supply health sensors
- Power consumption per chassis and module
- Power available

There should be exception logging via SNMP trap and syslog (as specified above) of any incidents.

It should also be possible to remotely shut down a malfunctioning element in the system (automatic, user configurable, or manual), in order to preserve system health until the defective element can be exchanged. Of course it should also be possible to explicitly reactivate the element.

For example, a power supply failing in a system could cause instability in the device. If the system could make a decision to shut that power supply down, and assuming a redundant configuration, the switch would then operate in a stable condition until such time that the power supply could be exchanged.

5.2 Operational Monitoring

The minimum set of data the equipment should provide per interface is:

- mac-addresses per (sub)interface
- interface type
- transceiver type (+ wavelength for optical ports)
- if optic, then DDM / DMI support is mandatory with an update frequency of 60 seconds or less and it should be possible to force an update by the user for the following metrics:
  - receive power
  - transmit power
  - temperature of optical module
- status
  - operational status (enabled/disabled)
  - blocked due to port security shutdown
  - error-disable
  - access-list violations

Each measured DDM / DMI value should also be classified in "good", "warning" and "critical". This classification is performed with the warning and alarm threshold of each single metric type. To define the classification by the

Switch-OS the thresholds have to be taken from the pluggable module (transceiver) itself and not programmed / coded in the Switch-OS.

5.3 Protocol Monitoring

Depending on how the forwarding plane of the exchange LAN is implemented, various support protocols such as routing protocols and loop-prevention protocols may be configured. It should be possible to monitor the operational state of these protocols.

In chapter 7 a number of topology implementations are discussed. If you imagine a VPLS setup with LSPs and an underlying OSPF infrastructure it must be possible to monitor functionality of the routing protocol.

5.4 Statistics and Accounting

As well as implementing de facto SNMP counters/RMON, also consider implementing the following:

5.4.1 Per-VLAN traffic statistics

Several implementations require different VLANs for different purposes. Examples for the use of different vlans are

- Unicast LAN
- Multicast LAN
- Quarantine LAN
- Jumbo Frame LAN
- Point-to-Point connection (pseudowire)

It is an important feature to support all statistical information not only on physical interfaces, but also for logical ones to provide the same kind of information to all connected ASes.

5.4.2 Network telemetry support (via management interface)

It is important for a modern IXP to get statistical insight into traffic patterns. Not only for the operators themselves, most IXPs provide these statistics to their connected parties. Open standard protocols to gather this statistical information are preferred. sFlow (RFC3176), IPFIX (RFC7015) and NetFlow v9 (RFC3954) seem to be the most common. The behavior of the devices with regard to network traffic telemetry should be configurable:

- support of L2 primitives (ie. MAC addresses, VLAN tag, ethertype)
- support multiple export targets
- sampling rate per collector configurable
- sampling rate per (sub)interface configurable
- and as much combinations of the 2 above as possible

In addition to network traffic records the devices should also be able to export various interface statistics typically found via SNMP MIBs (ie. bytes counts, packets counts, errors, etc.) with configurable time interval per collector.

5.4.3 Per Ethertype statistics

With the exhaustion of IPv4 and the need to adopt IPv6, IXP operators have often been asked the question “How much IPv6 traffic do you carry?”, or “What proportion of your total traffic is IPv6?”

Remember, because the IXP is not forwarding the traffic based on L3 header information, the IXP is not getting this accounting information.

The way that IXPs would prefer to collect this sort of information is through per-ethertype accounting.

This also benefits the IXPs in monitoring and enforcing “house rules”, as the majority of IXPs publish a list of permitted ethertypes, these usually being IPv4, IPv6 and ARP.

It’s recognised that maintaining accounting statistics on some 190-200 ethertypes, most of which will be zero in an IXP environment, is resource intensive, however most operators are usually only concerned with a very short list of ethertypes.

Therefore it should be possible via configuration to define a list of ethertypes which the operator is interested in (possibly using an access-list style nomenclature), on which the system should keep statistics, and the statistics should be per-packet and octet counters, ingress and egress.

Vendors should therefore define and clearly indicate in documentation how many ethertype accounting buckets are possible per system and per interface.

6 Layer 2 features

6.1 Jumbo Frame Support

All switches used in an IXP should be capable of supporting “Jumbo Frames” (frames with greater than 1500-bytes of payload) of at least 9216 bytes of maximum frame size. Furthermore, the jumbo frame size should be configurable on at least a per-system, and preferably per-VLAN or per-interface basis.

There is an effort among the Internet Peering community to standardise on 9000 bytes as the de-facto jumbo frame size for IXPs, as this is a) easy to remember, and b) works around the various different vendor implementations of a “9k” MTU.

6.2 Security and Management Features

Requirements mentioned in the following section should apply independent of the implementation at the IXP and independent of physical or logical ports.

6.2.1 Port-based access-control

Access-control of traffic entering customer interfaces is vital to protecting the Exchange Point against loops, and unwanted protocols. In addition, IXP operators generally do not desire ad-hoc extensions connected to their network. The common way of managing this is to enforce a "router-only" or "limited MAC address" rule.

Common technologies to provide access-control are Port security and Layer-2 access-lists.

Regardless of the technology chosen, the following options must be supported:

- Static configuration of allowed source MAC addresses for each VLAN on a customer port
  Typically there is one valid MAC address per each port
- Static configuration of allowed destination MAC addresses for each VLAN on a customer port
  Typically the same list of valid destination MAC addresses can be applied for all customer ports
  (it is important that in the case of defining source- and destination MAC addresses all combinations of source and destination addresses are treated as valid)
- Dynamic configuration of a number of allowed source MAC addresses for each VLAN on a customer port
- Configuration of allowed Ethertypes (typically ARP, IPv4 and IPv6)
- Blocking of link control protocol BPDUs, such as STP or LACP on non-LACP ports
- Blocking of unwanted packets from customers
  - DHCP guard
  - RA guard
- configurable action at violation: restriction of “bad traffic” (also per vlan) or shutdown of physical port
- configurable auto-recovery after port-security violations

This sort of filtering should be implemented in hardware wherever possible, and not have an effect on the forwarding performance of the system. Where this is not possible, such as the filtering being CPU-bound, it must be clearly documented.

Changes to the access-control policy must be atomic. If an access-list gets updated, there must be a seamless change from the old to the new access-list-version. Traffic that comes in while the new policy is applied must either be handled according to the old policy, or dropped.

6.2.2 Reduce the impact during maintenance

Traffic between customer routers on an IXP often passes through many different components of the IXP fabric. Due to the standard behavior of BGP an outage of one component can lead to blackholing of traffic until BGP sessions run into timeouts and traffic gets rerouted. In case of planned maintenance it should be possible to add Layer 4 access-lists to customer facing ports to block BGP port 179 as source and destination port. This would first lead to a timeout of the BGP sessions of members connected to this certain network component without interrupting the forwarding plane. Once all BGP sessions have timed out and traffic levels have reduced to zero the device can be restarted with a highly reduced impact of customer traffic. After the maintenance has been completed the Layer 4 access-lists can be removed to enable BGP sessions again.

It is necessary that these Layer 4 access-lists can be configured in addition to already existing filters or access-lists, specifically layer 2 MAC address filters.

6.3 Policy exception handling

The following events should be generated upon detection of a violating frame:

- SNMP-trap
- Configurable syslog message
    - which syslog facility to write to
    - which level should the switch start write events to syslog

And optionally:

- Copy the frame to a mirror port
- Capture the frame in a buffer on the switch, where it can be viewed.

It should be possible for the IXP operator to view the contents of the Ethernet header (source and destination MAC address and Ethertype) of the violating frame, to determine the nature of the traffic.

For frames with source MAC addresses that violate the access-control policy, the following thresholds should be configurable:

- "Forwarding" limit - the maximum number of unique source addresses for which frames will be forwarded
- "Soft" limit - the limit at which a syslog event is recorded
- "Hard" limit - the limit at which the port is automatically shut down (in addition to generation of a syslog event)

6.3.1 MAC/CAM learning

Under normal circumstances the flow of traffic between two MAC addresses on the Exchange platform is bidirectional, with BGP messages being exchanged between the two parties in both directions, and actual Internet traffic flowing in one or both directions.

With the introduction of route servers, this changed. If two parties only speak BGP with the route server, and further traffic between them is unidirectional, this may result in the Exchange platform aging out the MAC address of the receiving party from its MAC and CAM tables. This problem is aggravated by the long ARP timeouts commonly configured on Exchange Point LANs.

To remove this problem, it should be possible for the Exchange Point operator to configure the MAC and CAM aging timeout of a device. This timeout should be configurable to a value that is higher than the commonly used ARP timeout on participants routers (between 4 hours and 1 day).

6.3.2 L2 Broadcast/Multicast/Unknown Unicast control, ARP snooping

Many exchange points insist on participants using IP addresses they have assigned by the exchange operator. It is desirable for the operator to be able to monitor/restrict "off-net" ARP.

As Ethernet is a broadcast medium, broadcast storms have been known to bring exchanges to their knees, affecting the forwarding abilities of both the switches of the exchanges, and the participants' routers. Monitoring/rate limiting/control of Ethernet broadcast frames is desirable.

Most exchanges also forbid the speaking of interior routing protocols across their peering network. Since these take the form of broadcast or multicast frames on ethernet, control would help monitor this type of incidence.

Such control should be able to distinguish (through appropriate configuration) between legitimate ARP requests and genuine broadcast storms.

In addition, unknown unicast floods may also start to become an issue when there is a range of different port speeds across a single layer 2 environment - it's possible for a large flow of unknown unicast from a 10 Gbps port to saturate a 1 Gbps port. This is becoming increasingly important at the largest IXPs where fast connections are by magnitude orders larger than smaller connections to the same IXP matrix.

Being able to do hardware rate-limiting/discard of unknown unicast traffic will help maintain an uncongested port for end stations with slower access speeds.

Dynamic ARP inspection should be supported. There should be suitable configuration knobs to be able to rate limit, shut down, log exceptions, etc.

Filtering of Router Advertisements and logging if they happen (RA Guard).

6.3.3 ARP and ND optimisation

In this case a non standard behavior would be desirable:

In an environment where a lot of peering-routers are connected to a central L2 IXP infrastructure it is desirable to reduce the amount of broadcast traffic as much as possible. Filtering source-mac addresses is one way to reduce misuse – another approach would be to filter also on destination address (no unknown unicast traffic anymore) and redirect “unwanted” or unexpected packets. IXP operators generally do have knowledge of

- who is connected on which port
- the appropriate IP(v6) address expected on that port
- the appropriate MAC address expected on that port

It is not necessary to forward (broadcast) arp-request-packets to all other connected parties as “the IXP infrastructure” would know the answer already. Possible solutions to implement this could be (but not limited to)

- preload IXP hardware to answer requests autonomously
- redirect non-unicast traffic to controller (e.g. like Openflow)
- redirect non-unicast traffic to IXP management device (defined by port or different VPLS-instance)

The redirection of unexpected packets could also be used to trigger alarms or other mitigation actions to keep the IXP infrastructure clean.

6.3.4 Port mirroring

It is sometimes necessary to mirror participants' ports, either because a participant is suspected of some inappropriate activity, to help obtain information when debugging a problem, or during the installation/turn-up phase.

Not all exchange points have staff on site 24x7, and port mirroring may need to be remotely set up, without hands-on intervention on-site.

The ability to allow any port to mirror any other port with a similar or lower speed within the chassis would allow the operator to connect a traffic collector/analyser device to a monitoring port, and simply configure the switch to mirror a port as desired to the monitoring port.

This is becoming a challenge in the face of massive LAG groups - such as peering routers now attaching at speeds of greater than 16x10 Gbps, or via multiple 100 Gbps on the largest European exchanges. It is simply not viable, operationally, nor financially, to maintain a LAG group as large as your largest LAG just for port mirroring.

During mirroring operations, the IXP operator is either only interested in specific things, or just looking for a general overview (a “flavour”) of what is happening on the port under scrutiny.

The first instance would be handled by a capture filter of things which need to be copied to the mirror port. This would help if you are only looking for a specific frame type, source/dest addresses, or other specific property.

The second could be achieved using a user-configurable "sampled" mirror port - a sampling rate can be set up to control the rate of frames copied to the mirror port.

Both the latter features would permit a slower port to mirror items of interest from a faster port. It should also be possible to combine both features: define a capture filter and define a sampling rate per entry for the matching packets. In this case it would be nice to have for example all packets of a certain type, but just a sample of the rest.

Remote mirroring: A situation could occur where a mirror-destination may not necessarily be on the same device. Most IXPs consist of many devices and all of them should be able to mirror a specific port to a destination port on any of the IXPs devices. For instance, mirror output could be forwarded to a specific virtual interface attached to a VLAN, or to a VPLS service.

6.4 Performance

All physical ports within the device are expected to forward packets line rate for packets with minimal allowed packet size and all possible reasonable features turned on. If it is not possible to support line rate forwarding for reasonable features, then a detailed documentation of the limitations (such as packet or bitrate ceilings) needs to be provided.

If the packet-forwarding-performance depends on the combination of interfaces used in a device this has to be clearly documented as a limitation. E.g. X bps per slot/chassis with only 10 and 100G interfaces, but only X/2 if also 1 Gbps interfaces are used in the same box.

If there is a physical chassis- or slot-capacity limit that can never be increased during the lifetime of the product, this must be documented to give an indication how far a device can scale up in future.

6.5 Scalability and Resilience

6.5.1 Ring Restoration Protocols

There are a number of proprietary ring protocols, such as Extreme's EAPS (published as informational RFC3619), or Brocade's MRP.

They are relatively similar in operation, in that they make assumptions about the number of redundant links in a topology (i.e. only one), have a concept of master and transit nodes, use a "heartbeat" sent out by the master, and topology change messages are passed between the nodes to speed network re-convergence (by triggering FDB flushing, and backup port unblocking on the master node).

These recovery protocols may become less important as other protocols such as VPLS, TRILL and SPB become preferred for deployment in IXPs, however it is necessary that the ring-control protocols include the following behaviors:

"Non-revertive" behavior - the ring will only fail back to the "worker" state from the protect state if it is manually triggered by the operator, or by a failure elsewhere on the ring.

"Interlocked" behavior with the link state or link state protection mechanisms such as UDLD or LFN - there is a positive check that a link is up and stable before triggering a topology re-convergence.

6.5.2 Trunking and Link-Aggregation

It's become increasingly common for exchange points to become multiple device and multiple site based, and many need to deploy link aggregation to handle the volume of interdevice traffic, where it exceeds the maximum speed of a single link.

Most equipment implements load-sharing using either round-robin or address-based algorithms.

In exchange points, many pieces of equipment will have similar MAC addresses, especially the first and last bytes (corresponding to vendor and slot position on router). Load sharing based on MAC or (connected) IP-Address is not efficient.

The requirement for a modern IXP is to use a 5-tuple hashing (consisting of source MAC, destination MAC, source IP, destination IP, protocol) algorithm to determine the forwarding interface on an aggregated link. For debugging reasons there should be a possibility to predict on which link in a bundle a flow will be forwarded.

For seamless upgrades and reconfigurations it is necessary to add or remove any port from a LAG without interrupting the forwarding on the LAG.

Load-sharing of broadcasts and multicast traffic should be implemented.

IEEE 802.3ad link-aggregation "LACP" should be supported – fast and normal mode.

Port access-control features must also be applicable to trunks/link-aggregated groups, and work across that group as though the group was a single port.

It should be possible to terminate aggregated ports on multiple chassis (Multi-Chassis Link Aggregation).

6.5.3 Multicast Control and Containment

Most switches are configured with IGMP snooping for multicast control.

However, in an exchange point, with only routers attached, there is no IGMP present, only PIM and MSDP, and all multicast packets are flooded out of all ports.

An exchange point, however, is an ideal place for multicast peering to happen: inject the traffic once, and it comes out several times (as much as is needed, or in the current situation, as much as isn't needed!).

Cisco developed RGMP (Router Group Management Protocol). This is a proprietary technology whereby the router can communicate to the switch which multicast groups it wishes to see.

This remains, despite being released as an informational RFC (RFC3488), a vendor specific feature. As a wide range of platforms are present at many exchange points - both in equipment used by the operator, and the participants, these are true multi-vendor environments.

Therefore, this is not a workable solution for most exchange points, whose principles are often include "equal treatment" of participants.

While it may not solve all potential issues with multicast peering, implementing PIM-SM snooping and pruning within the switches will achieve the traffic containment requirements.

Where PIM snooping is available, this should not have a negative effect on the overall forwarding performance of the system - e.g. PIM snooping should be able to operate in concert with hardware flooding of the unicast frames. Where there is a performance impact, this and its surrounding caveats shall be clearly documented.

6.5.4 VLAN tagspace issues/overlapping

A serious emerging issue is VLAN tag space overlapping/clashing issues. Most metro transport networks can solve this by using q-in-q (tag stacking), however, this doesn't apply to shared networks like Internet Exchanges.

Current switches use a 1:1 mapping of 802.1q vlans to bridge groups, which is the way 802.1q was probably intended. This mapping should be loosened if not abandoned - nowadays there are so many ways to egress an ethernet frame from a switch that more and more often we have to resort to 'tricks' to put the right label on the right ethernet packet going out the right interface.

This problem is being exacerbated by a number of issues:

- Increased use of switch router products (e.g. Cisco 7600, Brocade MLX/XMR, AlcatelLucent 7750, etc)
- Use of switches as "channel-banks" - breaking out a single higher speed router interface to a larger number of slower/similar interfaces, using cheaper hardware
- Use of metro-ethernet, LAN extension or Ethernet over MPLS ("Martini") circuits to connect to the IXP

We think there are two (fairly similar) approaches to solving this:

- Basic VLAN tag rewrite
- Separate the tag from the virtual bridge instance

VLAN tag rewrite is, as its name suggests, being able to rewrite a dot1q tag on a specific interface to a VLAN ID on the switch. This would need to be implemented on both ingress and egress.

The other option is complete separation of VLAN ID from the virtual bridges inside the switch. You assemble a framework where you can place untagged ports, tagged ports, q-in-q tagged ports, mpls endpoints, atm vc's all together in into the same virtual bridge. Effectively a bridge group which can contain any number of these sort of entities.

6.5.5 Link failure detection

Link failure detection should be implemented, and should look like:

- UDLD - Uni-Directional Link Detection
- LFN - Link Failure Notification
- BFD - Bi-directional Forwarding Detection <http://www.ietf.org/ids.by.wg/bfd.html>
- IEEE 802.1ag compliant behavior

This avoids the risk of an ethernet link going "one-way" and fooling the restoration protocols that the link is working, when really it isn't.

The switch should also provide user configurable options for link aggregated (trunked) ports - the option may be to shut down the entire link-aggregate group, or keep operating on the remaining ports in the group.

7 Topological features

In cases where IXPs consist of more than two pops a standard implementation with spanning tree protocol loop prevention leads to unused/blocked links. Independently of the technology a vendor uses to prevent layer 2 loops it’s always an advantage to

- use all links within the network
- select the shortest path for the destination

The most important features of common technologies are listed here:

7.1 Virtual Private LAN Service

The current common state of the art way for providing a loop-free topology is Virtual Private LAN Services (VPLS), as defined in RFC4761 (VPLS using BGP signaling) and RFC4762 (VPLS using LDP signaling). VPLS works by creating an Ethernet broadcast domain on top of a mesh of Label Switched Paths (LSPs) in an MPLS network. In addition to providing a loop-free topology, VPLS also brings the possibility to balance traffic over multiple distinct paths in the network, so that redundant links are always used simultaneously.

In order to allow for deploying an IXP peering LAN as a VPLS service, VPLS implementations should support the following features:

- Load-sharing of traffic over multiple LSPs based on src/dst MAC address, IP address and TCP/UDP port number (for both regular and Spoke pseudowires)

- Load-sharing of traffic over multiple physical links, based on src/dst MAC address, IP address and TCP/UDP port number. This fine-grained load-sharing should be available both for access devices (PE) and core devices (P)
- Configuring a port in multiple VPLS instances
- Hair-pinning of traffic: The mapping of multiple VLANs on the same port to the same VPLS instance. This allows traffic coming in on a port on one VLAN to exit the access node on the same port using a different VLAN tag. In addition, broadcast, multicast and unknown-unicast traffic is replicated for each VLAN.
- Standard features for L2 operation mentioned in this document (access-lists, rate-limiting of broadcast, multicast and unknown-unicast, etc.) must also be supported inside VPLS

In addition, the following MPLS features should be available:

- Statically defined LSPs, allowing an administrator to specify some or all LSRs which an LSP must cross
- Active standby paths for LSPs, allowing sub-second failovers
- A mechanism to prevent traffic oscillations upon rapid flapping of LSPs (for example, a function whereby an LSP that failed over to its standby path will never automatically move back to the primary path, unless the standby path fails)
- Support for RFC6790
- On P-routers: load-balancing of traffic based on L3 header information in the MPLS payload (for backwards compatibility with PEs that do not support RFC6790)

7.1.1 Monitoring

For VPLS-enabled PEs it should be possible to retrieve the operational status of the following items

- LSPs
- Signaling protocol (e.g. RSVP, LDP) sessions
- VPLS instances and peers

7.2 Spanning Tree

Spanning Tree is an old technology, but still the only cross-platform dynamic solution available to operators of exchange points for dynamically managing multiple redundant links in their architecture.

There are a number of problems with Spanning Tree:

- Slow convergence
  - especially in cases of root bridge re-election
- Wasteful of resilient/redundant resources
  - redundant links are switched off
  - no traffic sharing
- Security concerns (highlighted above)

As the routes collected at an Exchange Point can be routed all over the world, any routing instability can act like dropping a pebble in a pond, and will spread around the Internet.

It's desirable to maintain stable routing sessions across Exchange Point LANs to minimise these routing flaps, because of load it places on routers, and the effects of route dampening penalties.

We believe that being able to declare ports as "end-stations" should avoid them being counted in the STP calculation, enable these ports to start forwarding more rapidly, and speed overall STP convergence time.

Rapid spanning tree (IEEE 802.1w) should be implemented (http://www.ieee802.org/1/pages/802.1w.html), and results from testing RSTP on certain platforms show that for simple topologies with few redundant links, sub-second failover and re-convergence is achievable with minimal tuning or additional configuration.

7.3 TRILL

TRILL (Transparent Interconnect of Lots of Links, RFC6325) is another approach to optimize traffic flows in switched Layer2 environments.

Much like a VPLS-based topology, TRILL provides an optimal forwarding path through the network for unicast traffic in an “all links active” topology. One of the advantages of TRILL is that it does not require overlaying the L2 service onto an IP substrate.

8 Standards

IXPs are very heterogeneous installations. Many different network operators connect with different kind of equipment (in terms of vendor and functionality) and all expect the same behavior of the IXP in the middle. It is therefore very important that IXP equipment behaves according to international accepted standards, especially for customer-facing interfaces.

If a product is declared to support a standard (e.g. TRILL) it is important that all features of this standard are supported. This is especially important for interoperability and future upgrades.

Since IXPs are often on the very edge of hardware development it could be that functionality is asked for which is not yet a standard or even supported with a certain proprietary. While standard behavior always gets preferred, non-standard solutions are possible as long as they only affect intra-IXP-links and devices.

9 Acknowledgements and Thanks

Thanks are due to the Euro-IX and wider IXP community for feedback and contribution of content.