The Future of Internet Exchange Points

Richard A Steenbergen <ras@nlayer.net> nLayer Communications, Inc.
Avi Freedman <avi@servercentral.net> ServerCentral

History of Internet Exchange Points

History of Internet Exchange Points

- 0th gen: Throw a cable over the wall. PSI and Sprint conspire to bypass ANS. When the 3rd network wants to connect, thus is born MAE-East.
- 1st commercial gen: FDDI, Ethernet, multiple metro areas, head of line blocking, scaling issues.

History of Internet Exchange Points

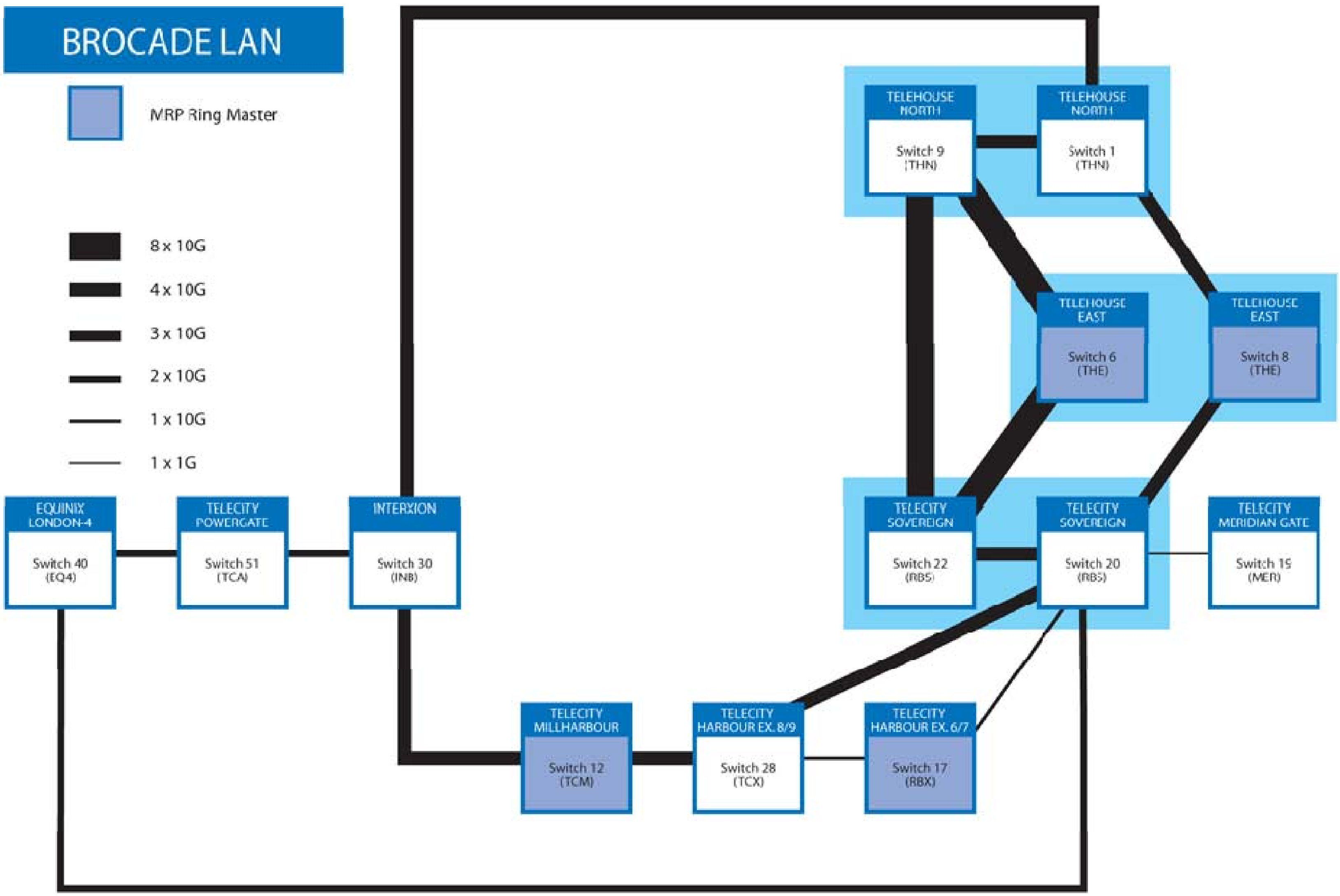
- 2nd gen: ATM exchange points, from AADS/PBNAP to the MAEs and peermaker.
- 3rd gen: GigE exchange points (mostly nonblocking internal to the switches): PAIX; the rise of Equinix, LINX, AMS-IX.
- 4th/current gen: 10Gig exchange points (upgrades and scale-out of existing IXs through 2 or 3 revs of underlying platform expansion).

How Modern Exchange Points Work

How Modern Exchange Points Work

- Almost exclusively Ethernet based today
  - Driven by the commodity price of Ethernet switches
  - ATM exchanges died at OC12, were replaced by GigE
  - 10GE and Nx10GE has been primary growth for years
- Almost exclusively flat layer 2 VLAN
  - IX will be assigned an IP block (usually a /24, or similar)
  - Each member router will be assigned an IP
  - Any member can talk to any other member via Layer 2
  - Some broadcast traffic (typically ARP) is required
    - But this is typically well policed and a small fraction of traffic.
    - Even the biggest IXs have only a few hundred members.

Example Large IX (LINX) Topology



So What’s Wrong With Existing IXs?

Fragility

- Layer 2 networks are relatively easy to disrupt
  - Forwarding loops, L2 networks have no TTL.
  - Broadcast storms, traffic flooded to all members.
    - Not only can these take down the exchange, but they can take down the members' entire router if successfully delivered.
- Today we work around these issues by:
  - Locking the port down to a single MAC address.
    - Either strictly hard-coded or via a limit of 1 dynamic MAC.
  - Only a single directly connected router port is allowed.
  - Careful monitoring of member traffic with traffic sniffers.
  - Good IX's have well trained staff for rapid response.

Accountability

- Most routers have poor or even no mechanisms for measuring traffic exchanged over a layer 2 IX.
- The options in use today are:
  - NetFlow from member router
    - Doesn't provide MAC layer information, can't handle inbound traffic.
    - Some popular platforms can't provide any useful NetFlow data at all.
  - SFlow from member router or from IX operator
    - Still sampled data, can easily be off by +/- 5% or more.
  - MAC accounting from member router
    - Not available on the vast majority of popular platforms today.
- None integrate well with provider 95^{th} % billing systems.
- IX's are a poor choice for delivering billed services.

Security

- Any member can communicate with any other member, whether this is desired or not.
  - Vulnerable to traffic injection from peers or non-peers.
  - Poor accounting options make this hard to detect.
    - When caught, “failure to set next-hop self” provides easy excuse.
  - Even less security available for selling paid transit.
- Vulnerable to Denial of Service attacks
  - Can even be delivered from the outside world if the IX IP block is announced (as is frequently the case).
- Vulnerable to traffic interception, manipulation
  - ARP/CAM manipulation is trivial on a layer 2 network.

Scalability

- Difficult to scale and debug large layer 2 networks
  - Redundancy is provided through spanning-tree or similar vendor-proprietary protocols to block loops.
  - Large portions of the network must be placed into a blocking state to provide redundancy.
    - Idle capacity wastes money, increasing the cost of service.
  - Spanning-tree or similar protocols provide poor controls over where the traffic will flow in an outage.
  - Constrained to simple ring or star topologies, hard to scale to hundreds of gigabits across multiple locations.
    - Even the largest/best IXs who have successfully built a network to handle 800Gbps+ in 8 locations would have a hard time building to 80 locations, or supporting 8000 members.

Managability

- Poor controls over traffic rates and/or QoS
  - Essentially the ports are wide open and best effort.
  - Any member can send as much traffic as they want.
  - At best traffic controls are completely voluntary.
    - “Hey $Peer, please back off you’re filling my port.”
    - Also makes it a poor choice for selling paid services.
- Difficult to manage multi-router redundancy
  - Multiple routers see the same IX /24 in multiple places.
  - Creates an “anycast” effect to the peer next-hops.
  - Can result in blackholing if there is an IX segmentation.
  - Or if there is an outage which doesn’t drop link state.

Other Issues

- Other issues
  - Inter-network Jumbo Frame support is extremely difficult.
    - No ability to negotiate a per-peer MTU today.
    - Almost impossible to find an acceptable common MTU for everyone.
  - Service is constrained to IP only, between two routers.
    - Must use IX provided IP address block.
    - Cannot use for layer 2 transport handoff.
- Summary
  - L2 IX's are an inherently fragile and unstable system.
  - We've managed to make them work for free peering traffic.
  - But they are still very poor choices for selling or buying services, delivering full transit, transport handoffs, etc.

Engineering a Better Exchange Point

Architecture of an Exchange Point

- The most common architecture of a modern exchange point is a shared broadcast domain.
  - Any member can talk to any other member.
  - Members are given a single IP from a common subnet.
  - Broadcast traffic delivered to every member.
- An alternative is using point-to-point virtual circuits
  - Essentially behaves like a private interconnection (PNI).
  - But adds additional overhead in circuit setup.

Eliminating the Shared Broadcast Domain

- So how would one do this under Ethernet?
  - Point-to-Point virtual circuits between members using 802.1q.
  - Hand off multiple virtual circuit VLANs on a single Interface.
- The concept is not a new one
  - This is how peering used to work over the old ATM exchanges.
  - But this technique was abandoned due to the significant cost and performance advantages of Ethernet over ATM.
- But it turns out, it’s not that simple
  - There are reasons why you can’t just roll out a VLAN based system for point-to-point interconnections over a traditional Ethernet switch.

The Problems with VLAN Exchanges

- The biggest issue is limited VLAN ID space
  - Ethernet VLANs are limited to 4096 possible IDs
    - 802.1q protocol can't express any more, only a 12 bit ID space.
  - "VLAN stacking" techniques are used to scale this for transport networks, but do not help in this use case.
  - With a traditional Ethernet switch, the VLAN IDs would need to be shared globally across the entire IX.
- This means a 65-member full mesh would completely exhaust all available VLAN IDs.
  - Traditional "VLAN rewrite" solutions don't help either.
    - Often they just burn both VLAN IDs, doubling exhaustion rate.
    - Also non-deterministic, due to shared resources across port ASICs.

The Problems with VLAN Exchanges

- But wait, there’s more…
  - Not only does the IX have to manage its own VLANs
  - But it has to manage what VLANs each member can use
- Most IX members today use “layer 3 switch” routers
  - Comprises essentially the entire market for “cheap” 10GE.
  - But their architecture makes VLAN IDs globally significant
    - If port 1 uses VLAN ID 123, port 2 can’t also independently use it.
    - VLAN 123 is a single global resource across the entire chassis.
  - Also, many platforms can not use the entire 4096 space.
  - And, many “routed” interfaces consume a virtual VLAN.
- Negotiating VLAN IDs would be next to impossible!

Requirements of a New IX Architecture

- Clearly a VLAN-based Ethernet switch won't work.
- A reasonable solution based on Ethernet must:
  - Expand the virtual circuit ID space significantly.
    - 2^12 = 4096 is simply not enough.
  - Decouple 802.1q VLAN IDs from the IX infrastructure.
  - Make VLAN IDs have only local, per-port significance
    - That is, allow VLAN ID reuse on the IX platform across ports.
  - Allow members to choose their own VLAN IDs per VC
    - To avoid conflicts with existing member router VLAN IDs.
    - No "negotiation" of VLAN IDs with either the IX or remote party.
- How could we possibly accomplish all these goals?

The Answer – An MPLS Based IX

- Use MPLS transport rather than Ethernet switching
  - Solves VLAN scaling problems
    - MPLS Pseudowire IDs are 32-bits – Over 4 billion virtual circuits.
    - VLAN ID is no longer carried with the packet, used only on handoff
    - VLAN IDs are no longer a shared resource across the IX or device
  - Solves VLAN ID conflict problems
    - Members could choose their own VLAN ID per VC handoff.
    - There is no requirement that the VLAN ID match the remote party.
  - Solves network scaling problems
    - Using MPLS TE is far more flexible than layer 2 protocols.
    - Allows the IX to build more complex topologies, interconnect more locations, and more efficiently utilize resources.

Advantages

- Security
  - Each virtual circuit would be isolated and secure.
  - No mechanism for a third party to inject or sniff traffic.
  - Significantly reduced potential for Denial of Service.
- Accountability
  - Most routers/L3 switches provide SNMP measurement capabilities for their VLAN interfaces/sub-interfaces.
  - Members can now accurately measure traffic on each VC, without “guestimation”, using their standard tools.
  - Capable of integration with most ISP billing systems.

Advantages

- Services
  - With proper security and accountability, delivering or buying paid services (transit, etc) becomes possible.
  - Support for “bandwidth on demand” now possible.
  - No longer constrained to IP-only or one-router-only.
    - Can be used to connect transport circuits, SANs, etc.
  - Provides the features of a metro transport solution or physical cross-connect, but with rapid provisioning.
- Others
  - Jumbo Frame negotiation across shared fabric now possible, since MTU can be configured per subint/vlan.

Getting Fancy With It

- Could interconnect with existing metro transport
  - Use Q-in-Q VLAN stacking to “extend” the network onto third party infrastructures.
  - Imagine a single IX platform able to service hundreds or thousands of buildings in a metro region.
- Could auto-negotiate VC setup using a web portal
  - Rapid provisioning using web 2.0 invite/accept model.
  - Could even auto-negotiate things like MTU and VC IPs
  - IX operator could automatically provide /30 (or /31)
    - And members could manage their DNS via the web portal
- Also functions as a transport platform.

Summary

- Existing exchange point architecture mostly works
  - With careful engineering to protect the L2 network
  - With a limited number of locations and chassis
    - Increasingly difficult to stay this way, as colocation facilities run out of space/power, and as data speeds increase.
  - With a significant amount of infrastructure overhead
  - For settlement-free services, but not paid services
  - For IP services only.
- But a new kind of exchange point would be better
  - Could transform a “peering only” platform into a new “ecosystem” model to buy and sell services too.

Send questions, comments, complaints to:

Richard A Steenbergen <ras@nlayer.net>

Avi Freedman <avi@servercentral.net>