Internet Exchange Point Design

Philip Smith
<philip@apnic.net>

MENOG 13

Kuwait

24th September 2013

Last updated 20 September 2013

Presentation Slides

- Available on
  - http://thyme.apnic.net/ftp/seminars/MENOG13-IXP-Network-Design.pdf
  - And on the MENOG 13 website
- Feel free to ask questions any time

IXP Design

- Background
- Why set up an IXP?
- Layer 2 Exchange Point
- Layer 3 “Exchange Point”
- Design Considerations
- Route Collectors & Servers
- What can go wrong?

A bit of history

Where did the IX concept come from?

A Bit of History…

- NSFnet – one major backbone
  - US "National Science Foundation" funded
  - Connected academic & research institutions
  - Also connected "private company" networks, with acceptable use policy
    - **AUP: No commercial activity**
  - Three Network Access Points (NAPs): Chicago, New York, San Francisco
- Private companies needed to interconnect their networks
  - Requirement to send "commercial traffic"
  - Could not cross NSFnet
  - Resulted in the early "commercial Internet Exchanges"

More History…

- Early Internet Exchanges created in early 90s
  - CIX-West – west coast USA
  - MAE-East – east coast USA
  - D-GIX – Stockholm
- End of the NSFnet in 1995:
  - Meant move towards commercial Internet
  - Private companies selling their bandwidth
  - ANS (operator of the late NSFnet) had to join IXes
- Routing Arbiter project helped with coordination of routing exchange between providers
  - Traffic from ISP A needs to get to ISP B

More History still…

- The NAPs established late in NSFnet life were some of the original “exchange points”
  - NAP operators supported commercial activities as well
  - (Sprint: NY, PacBell: SF, Ameritech: Chicago, MFS: Vienna/VA)
- The NAPs replaced by IXPs:
  - NAPs didn’t succeed (operated by ISPs), replaced by more neutral IXPs
  - E.g. Virginia NAP replaced by MAE-East (by MFS)
- Mid 90s saw rapid Internet growth, with major providers connecting…

Even more History

- D-GIX formed in Stockholm in 1992
  - Three major ISPs interconnected
  - Latency reduction, performance gains
  - Local traffic stays local
- LINX formed in London in 1994
  - Five UK operators interconnected
  - Latency reduction, performance gains
  - Local traffic stays local
- HKIX formed in Hong Kong in 1995
  - Vibrant Internet community, many small operators
  - Latency, performance, and local traffic benefits
- Also AMS-IX in Amsterdam in 1994
  - Same reasons as others

Internet Exchange Point

- What:
  - **A neutral location where network operators freely interconnect their networks to exchange traffic**
- What is the physical IX:
  - An ethernet switch in a neutral location
- How does it work:
  - IX Operator provides the switch and rack space
  - Network Operators bring routers, and interconnect them via the IX fabric
- Very simple concept – any place where providers meet to exchange traffic

Internet Exchange Point

- Layer 2 exchange point
  - Ethernet (100Gbps/10Gbps/1Gbps/100Mbps)
  - Older technologies used in the past included ATM, Frame Relay, SRP, FDDI and SMDS
- Layer 3 exchange point
  - Has historical status now
  - Router based
    - Best known example was CIX-West
    - Router very quickly overwhelmed by the rapid growth of the Internet

Why an Internet Exchange Point?

Saving money, improving QoS,
Generating a local Internet economy

Internet Exchange Point
Why peer?

- Consider a region with one ISP
  - They provide internet connectivity to their customers
  - They have one or two international connections
- Internet grows, another ISP sets up in competition
  - They provide internet connectivity to their customers
  - They have one or two international connections
- How does traffic from customer of one ISP get to customer of the other ISP?
  - Via the international connections

Internet Exchange Point
Why peer?

- Yes, International Connections…
  - If satellite, RTT is around 550ms per hop
  - So local traffic takes over 1s round trip
- International bandwidth
  - Costs significantly more than domestic bandwidth
  - Congested with local traffic
  - Wastes money, harms performance

Internet Exchange Point
Why peer?

- Solution:
  - Two competing ISPs peer with each other
- Result:
  - Both save money
  - Local traffic stays local
  - Better network performance, better QoS,…
  - More international bandwidth for expensive international traffic
  - Everyone is happy

Internet Exchange Point
Why peer?

- A third ISP enters the equation
  - Becomes a significant player in the region
  - Local and international traffic goes over their international connections
- They agree to peer with the two other ISPs
  - To save money
  - To keep local traffic local
  - To improve network performance, QoS,…

Internet Exchange Point
Why peer?

- Private peering means that the three ISPs have to buy circuits between each other
  - Works for three ISPs, but adding a fourth or a fifth means this does not scale
- Solution:
  - Internet Exchange Point

Internet Exchange Point

- Every participant has to buy just one whole circuit
  - From their premises to the IXP
- Rather than N-1 half circuits to connect to the N-1 other ISPs
  - 5 ISPs have to buy 4 half circuits = 2 whole circuits → already twice the cost of the IXP connection

Internet Exchange Point

- Solution
  - Every ISP participates in the IXP
  - Cost is minimal – one local circuit covers all domestic traffic
  - International circuits are used for just international traffic – and backing up domestic links in case the IXP fails
- Result:
  - Local traffic stays local
  - QoS considerations for local traffic is not an issue
  - RTTs are typically sub 10ms
  - Customers enjoy the Internet experience
  - Local Internet economy grows rapidly

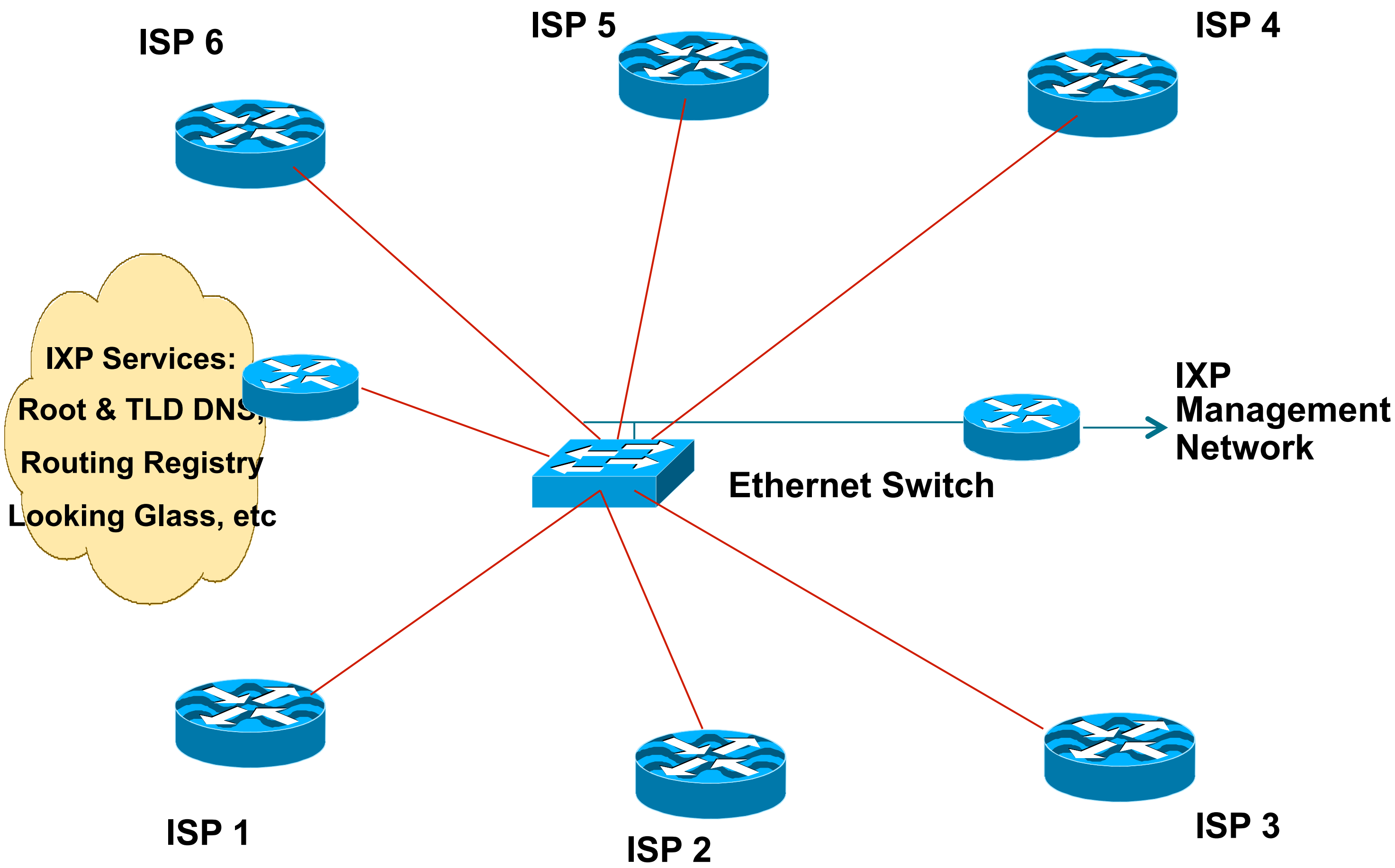
Layer 2 Exchange

The traditional IXP

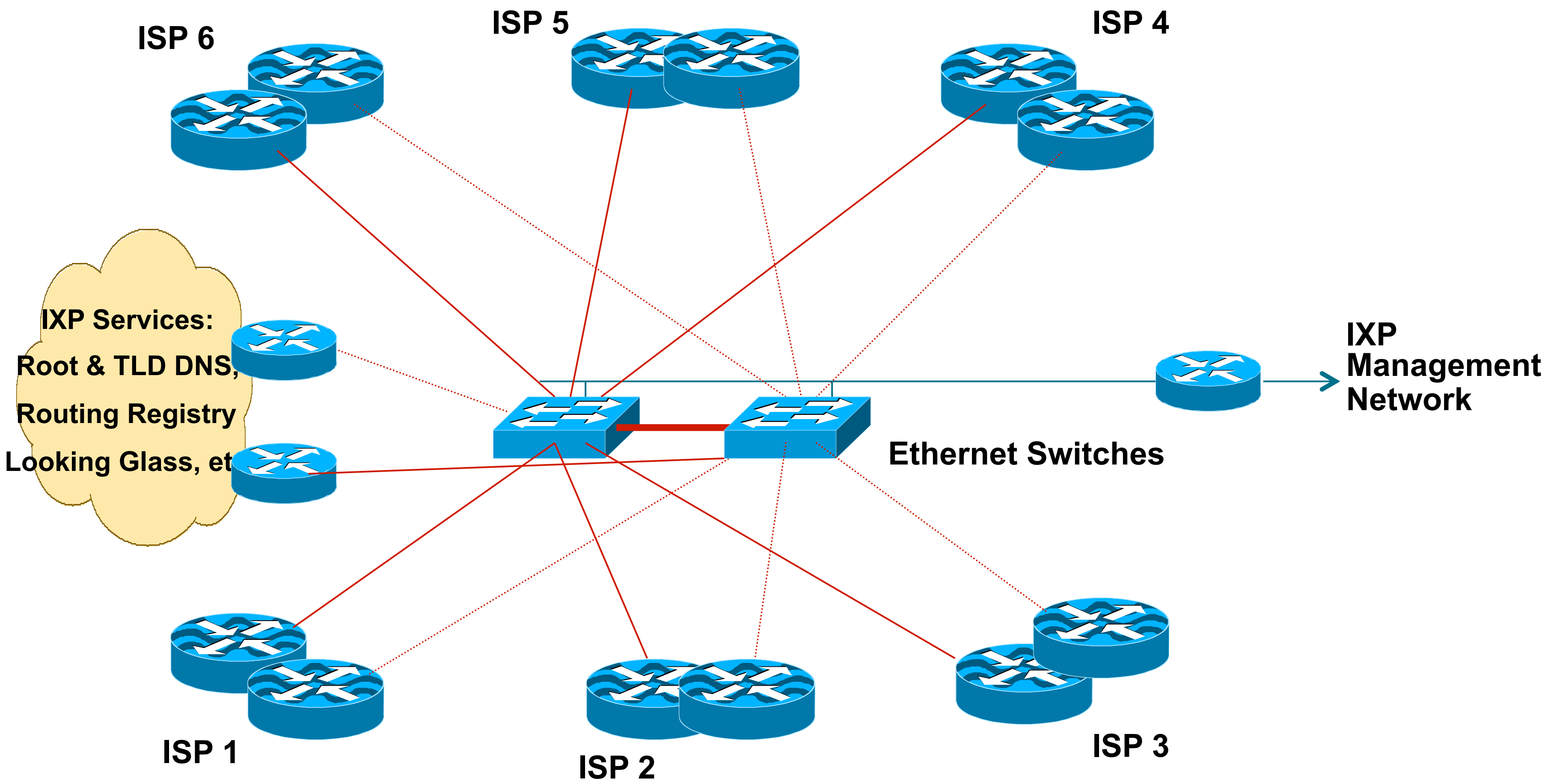
IXP Design

- Very simple concept:
  - Ethernet switch is the interconnection media
    - IXP is one LAN
  - Each ISP brings a router, connects it to the ethernet switch provided at the IXP
  - Each ISP peers with other participants at the IXP using BGP
- Scaling this simple concept is the challenge for the larger IXPs

Layer 2 Exchange



Layer 2 Exchange



Layer 2 Exchange

- Two switches for redundancy
- ISPs use dual routers for redundancy or loadsharing
- Offer services for the "common good"
  - Internet portals and search engines
  - DNS Root & TLDs, NTP servers
  - Routing Registry and Looking Glass

Layer 2 Exchange

- Neutral location
  - Anyone can install fibre or other connectivity media to access the IXP
    - Without cost or regulations imposed by location
- Secure location
  - Thorough security, like any other network data centre
- Accessible location
  - Easy/convenient for all operators to access
- Expandable location
  - IXPs result in Internet growth, and increasing space requirements

Layer 2 Exchange

- Requires neutral IXP management
  - “Consortium”
    - Representing all participants
    - “Management Board” etc
  - Usually funded equally by IXP participants
  - 24x7 cover provided by hosting location
    - Managed by the consortium

Layer 2 Exchange

- Configuration
  - Private address space if non-transit and no value add services
  - Otherwise public IPv4 (/24) and IPv6 (/64)
  - ISPs require AS, basic IXP does not
- Network Security Considerations
  - LAN switch needs to be securely configured
  - Management routers require TACACS+ authentication, vty security
  - IXP services must be behind router(s) with strong filters

“Layer 3 IXP”

- Layer 3 IXP today is marketing concept used by Transit ISPs
- Real Internet Exchange Points are only Layer 2

IXP Design Considerations

Exchange Point Design

- The IXP Core is an Ethernet switch
  - It must be a managed switch
  - It must have reasonable security features
  - http://www.ripe.net/ripe/groups/wg/eix/ixp-wishlist has more details
- Has superseded all other types of network devices for an IXP
  - From the cheapest and smallest managed 12 or 24 port 10/100 switch
  - To the largest switches now handling high densities of 10GE and 100GE interfaces

Exchange Point Design

- Each ISP participating in the IXP brings a router to the IXP location
- Router needs:
  - One Ethernet port to connect to IXP switch
  - One WAN port to connect to the WAN media leading back to the ISP backbone
  - To be able to run BGP

Exchange Point Design

- IXP switch located in one equipment rack dedicated to IXP
  - Also includes other IXP operational equipment
- Routers from participant ISPs located in neighbouring/adjacent rack(s)
- Copper (UTP) connections made for 10Mbps, 100Mbps or 1Gbps connections
- Fibre used for 1Gbps, 10Gbps, 40Gbps or 100Gbps connections

Peering

- Each participant needs to run BGP
  - They need their own AS number
  - **Public** ASN, **NOT** private ASN
- Each participant configures external BGP directly with the other participants in the IXP
  - Peering with all participants

    or
  - Peering with a subset of participants

Peering (more)

- Mandatory Multi-Lateral Peering (MMLP)
  - Each participant is forced to peer with every other participant as part of their IXP membership
  - **Has no history of success** — the practice is strongly discouraged
- Multi-Lateral Peering (MLP)
  - Each participant peers with every other participant (usually via a Route Server)
- Bi-Lateral Peering
  - Participants set up peering with each other according to their own requirements and business relationships
  - This is the most common situation at IXPs today

Routing

- ISP border routers at the IXP must NOT be configured with a default route or carry the full Internet routing table
  - Carrying default or full table means that this router and the ISP network is open to abuse by non-peering IXP members
  - Correct configuration is only to carry routes offered to IXP peers on the IXP peering router
- Note: Some ISPs offer transit across IX fabrics
  - They do so at their own risk – see above

Routing (more)

- ISP border routers at the IXP should not be configured to carry the IXP LAN network within the IGP or iBGP
  - Use next-hop-self BGP concept
- Don't generate ISP prefix aggregates on IXP peering router
  - If connection from backbone to IXP router goes down, normal BGP failover will then be successful

Address Space

- Some IXPs use private addresses for the IX LAN
  - Public address space means IXP network could be leaked to Internet which may be undesirable
  - Because most ISPs filter RFC1918 address space, this avoids the problem
- Most IXPs use public addresses for the IX LAN
  - Address space available from the RIRs
  - IXP terms of participation often forbid the IX LAN to be carried in the ISP member backbone
- Typically IXPs now provide both IPv6 and IPv4 support on IX LANs

Autonomous System Numbers

- IXPs by themselves do not require ASNs
  - Ethernet switch is L2 device, and does not run BGP
- Some IXPs have a Route Collector
  - This usually runs in a private ASN
- Some IXPs have a Route Server
  - This usually runs in a private ASN
- Some IXPs have "common good services"
  - These usually require Internet transit
  - Meaning the IXP requires a transit router
    - IXP arranges transit for services with a couple of providers
  - And this transit router requires a Public ASN and Public Address space

Hardware

- Try not to mix port speeds
  - If 10Mbps and 100Mbps connections available, terminate on different switches (L2 IXP)
- Don't mix transports
  - If terminating ATM PVCs and G/F/Ethernet, terminate on different devices
- Insist that IXP participants bring their own router
  - Moves buffering problem off the IXP
  - Security is responsibility of the ISP, not the IXP

Charging

- IXPs needs to be run at minimal cost to its member participants
- Common examples:
  - Datacentre hosts IX for free
  - IX operates cost recovery
  - Different pricing for different ports
- IXes do **NOT** charge for traffic crossing the switch fabric
  - They are a peering enabler, encouraging as much traffic as possible between members

Charging:
Datacentre hosts IX for free

- Datacentre covers all costs relating to the IX
  - They provide the switch and supporting infrastructure
  - They provide the operator cover
  - They benefit from the business the IX members and their customers bring to the DC
  - They benefit from the "prestige" of hosting the IX and its ancillary services
- The IX does not charge members for anything at all
  - Example: Seattle IX

Charging: IX Members pay flat fee

- Each member pays a flat annual fee towards their IX membership
- How it works:
  - Cost of switch and ports
  - Cost of operator support
  - Datacentre cost: power, air-conditioning, etc
  - Cost of IX membership association
  - Contingency needed for new equipment and upgrades
- Total annual cost shared equally amongst members
  - The more members, potentially the lower the costs to each

Charging: Differential pricing by port

- IXP Member pays according to the port speed they require
  - One linecard may handle 4 100GE ports
  - Or one linecard may handle 24 10GE ports
  - Or one linecard may handle 96 1GE ports
  - 96 port 1GE card is tenth price of 24 port 10GE card
  - Relative port cost is passed on to participants
  - Plus share in the cost of the switch
  - Plus all the costs mentioned in the flat-fee model
- IX members pay according to the expense of provision of the infrastructure they use
  - Example: Netnod IXes in Sweden

Services Offered

- Services offered should not compete with member ISPs (basic IXP)
  - e.g. web hosting at an IXP is a bad idea unless all members agree to it
- IXP operations should make performance and throughput statistics available to members
  - Use tools such as Cacti or Observium to produce IX throughput graphs for member (or public) information

Services to Offer

- ccTLD DNS
  - The country IXP could host the country's top level DNS
  - e.g. "SE." TLD is hosted at Netnod IXes in Sweden
  - Offer back up of other country ccTLD DNS
- Root server
  - Anycast instances of I.root-servers.net, F.root-servers.net etc are present at many IXes
- Usenet News
  - Usenet News is high volume
  - Could save bandwidth to all IXP members

Services to Offer

- Route Collector
  - Route collector shows the reachability information available at the exchange
  - Technical detail covered later on
- Route Server
  - Helps scale large IXes by providing easier BGP configuration & operation for participants
  - Technical detail covered later on
- Looking Glass
  - One way of making the Route Collector routes available for global view (e.g. www.traceroute.org)
  - Public or members only access

Services to Offer

- Content Redistribution/Caching
  - Google Global Cache
  - Akamai update distribution service
  - Broadcast media
- Network Time Protocol
  - Locate a stratum 1 time source (GPS receiver, atomic clock, etc) at IXP
- Routing Registry
  - Used to register the routing policy of the IXP membership (more later)

Introduction to Route Collectors

What routes are available at the IXP?

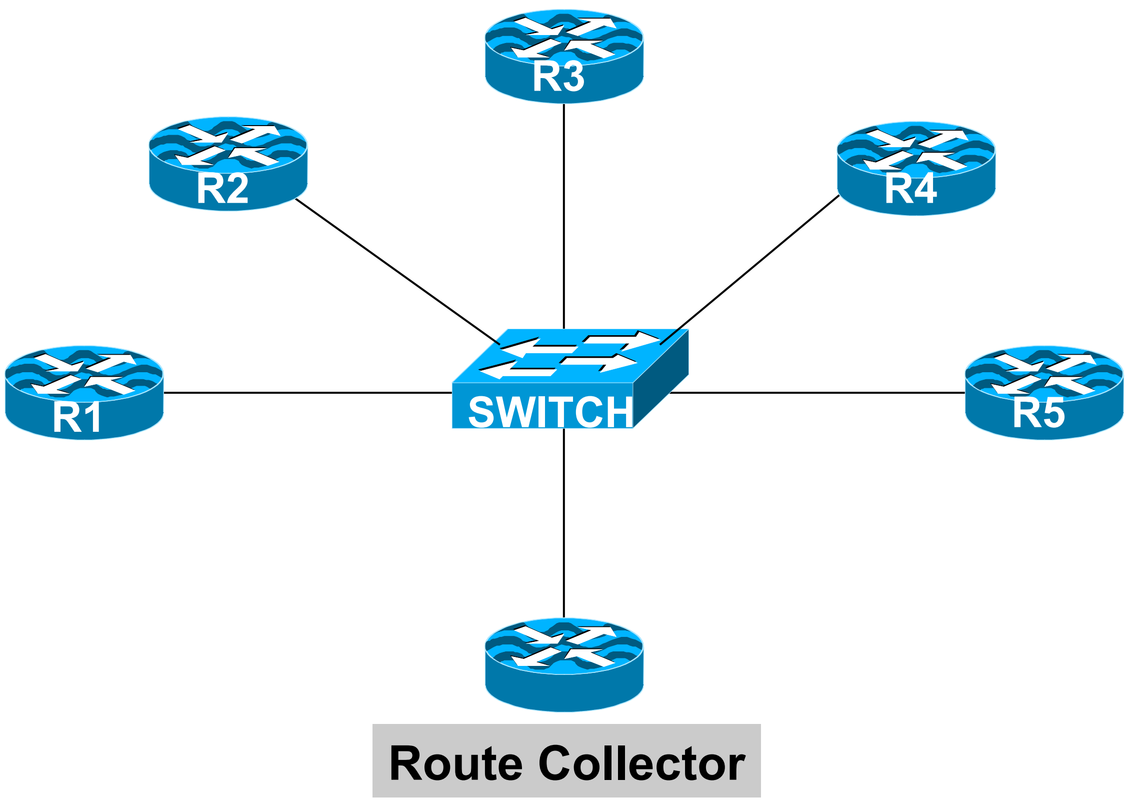
What is a Route Collector?

- Usually a router or Unix system running BGP
- Gathers routing information from service provider routers at an IXP
  - Peers with each ISP using BGP
- Does not forward packets
- Does not announce any prefixes to ISPs

Purpose of a Route Collector

- To provide a public view of the Routing Information available at the IXP
  - Useful for existing members to check functionality of BGP filters
  - Useful for prospective members to check value of joining the IXP
  - Useful for the Internet Operations community for troubleshooting purposes
    - E.g. **www.traceroute.org**

Route Collector at an IXP



Route Collector Requirements

- Router or Unix system running BGP
  - Minimal memory requirements – only holds IXP routes
  - Minimal packet forwarding requirements – doesn't forward any packets
- Peers eBGP with every IXP member
  - Accepts everything; Gives nothing
  - Uses a private ASN
  - Connects to IXP Transit LAN
- "Back end" connection
  - Second Ethernet globally routed
  - Connection to IXP Website for public access

Route Collector Implementation

- Most IXPs now implement some form of Route Collector
- Benefits already mentioned
- Great public relations tool
- Unsophisticated requirements
  - Just runs BGP

Introduction to Route Servers

How to scale IXPs

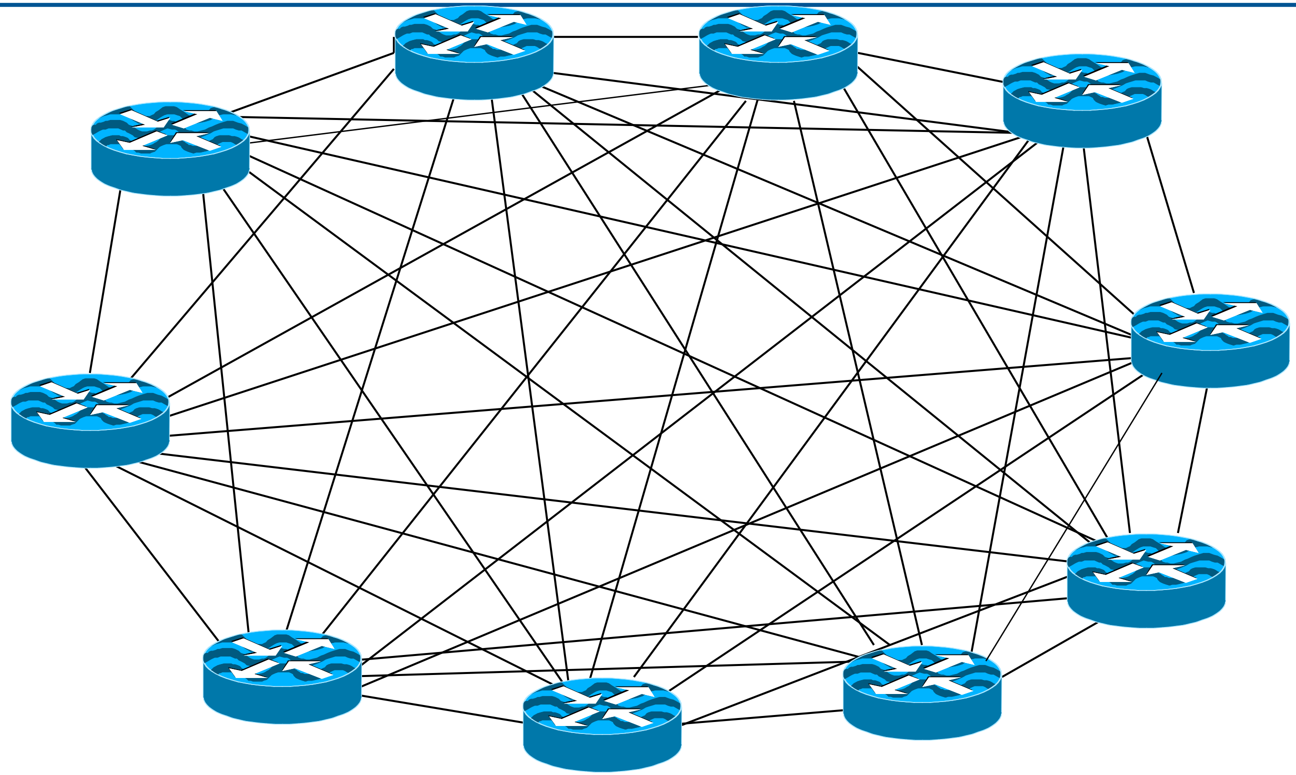
What is a Route Server?

- Has all the features of a Route Collector
- But also:
  - Announces routes to participating IXP members according to their routing policy definitions
- Implemented using the same specification as for a Route Collector

Features of a Route Server

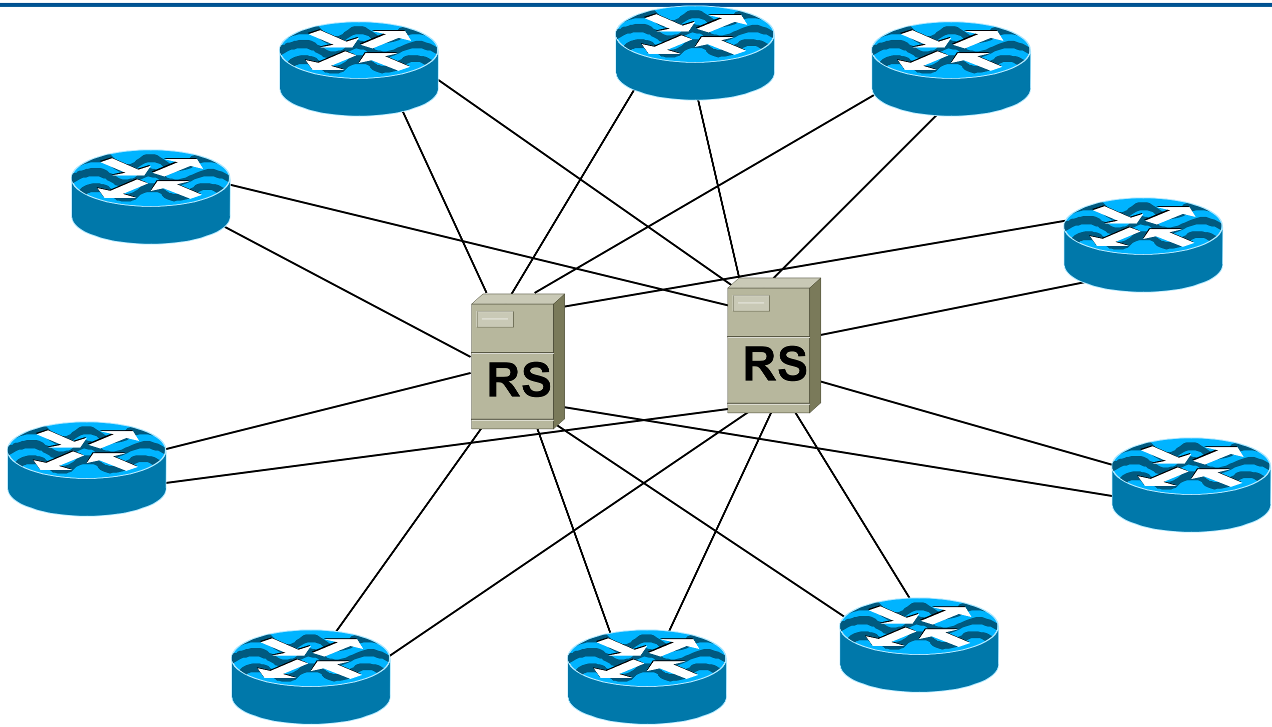
- Helps scale routing for large IXPs
  - Forward of packets is unaffected
- Simplifies Routing Processes on ISP Routers
- Optional participation
  - Provided as service, is NOT mandatory
- If traditional router used, will result in insertion of RS Autonomous System Number in the Routing Path
- Optionally uses Policy registered in IRR

Diagram of N-squared Peering Mesh



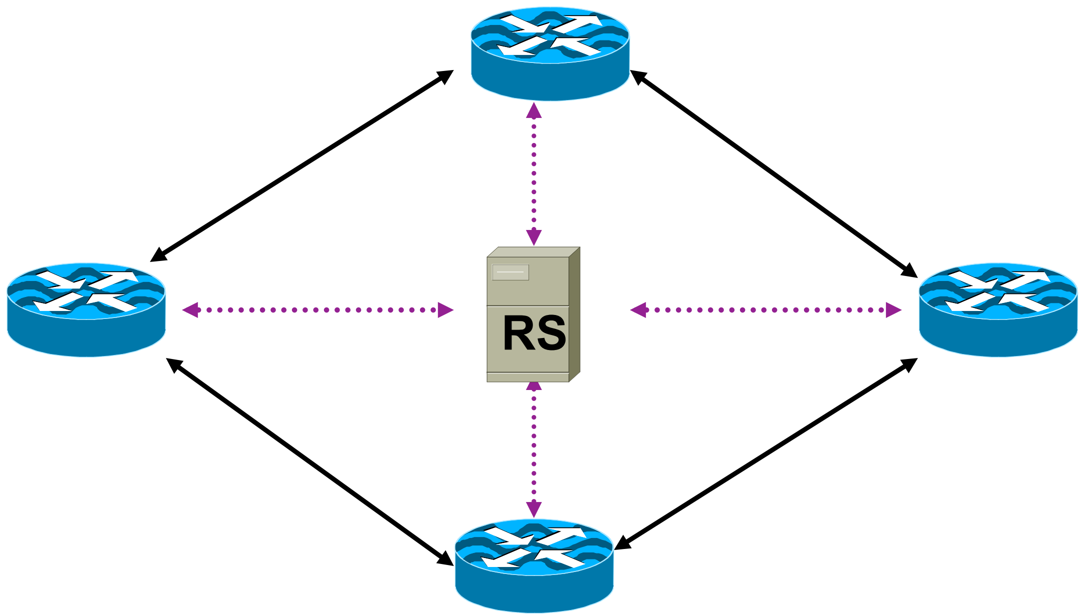
- For large IXPs (dozens for participants) maintaining a larger peering mesh becomes cumbersome and often too hard

Peering Mesh with Route Servers



- ISP routers peer with the Route Servers
  - Only need to have two eBGP sessions rather than N

RS based Exchange Point Routing Flow



TRAFFIC FLOW

ROUTING INFORMATION FLOW

Advantages of Using a Route Server

- Advantageous for large IXPs
  - Helps scale eBGP mesh
  - Helps scale prefix distribution
- Separation of Routing and Forwarding
- Simplifies BGP Configuration Management on ISP routers

Disadvantages of using a Route Server

- ISPs can lose direct policy control
  - If RS is only peer, ISPs have no control over who their prefixes are distributed to
- Completely dependent on 3rd party
  - Configuration, troubleshooting, etc…
- Possible insertion of RS ASN into routing path
  - (If using a router rather than a dedicated route-server BGP implementation)
  - Traffic engineering/multihoming needs more care

Typical usage of a Route Server

- Route Servers may be provided as an **OPTIONAL** service
  - Most common at large IXPs (>50 participants)
  - Examples: LINX, TorIX, AMS-IX, etc
- ISPs peer:
  - Directly with significant peers

    -and-
  - With Route Server for the rest

Route Server implementations

- Linux/FreeBSD server:
  - BIRD, Quagga, OpenBGPd
- Router:
  - Any router (but has RS AS in the AS-path)
  - Cisco IOS 15.2 and IOS XE 3.7 onwards has route-server-client configuration:

```
neighbor 172.16.1.1 route-server-client
```

Things to think about...

- Would using a route server benefit you?
  - Helpful when BGP knowledge is limited (but is NOT an excuse not to learn BGP)
  - Avoids having to maintain a large number of eBGP peers
  - But can you afford to lose policy control? (An ISP not in control of their routing policy is what?)

What can go wrong…

The different ways IXP operators harm their IXP…

What can go wrong? Concept

- Some Service Providers attempt to cash in on the reputation of IXPs
- Market their Internet transit services as "Internet Exchange Point"
  - "We are exchanging packets with other ISPs, so we are an Internet Exchange Point!"
  - So-called Layer-3 Exchanges — they really are Internet Transit Providers
  - Router(s) used rather than a Switch
  - Most famous example: SingTelIX

What can go wrong? Financial

- Some IXPs price the IX out of the means of most providers
  - IXP is intended to encourage local peering
  - Acceptable charging model is minimally cost-recovery only
- Some IXPs charge for port traffic
  - IXPs are not a transit service, charging for traffic puts the IX in competition with members
  - (There is nothing wrong with charging different flat fees for 100Mbps, 1Gbps, 10Gbps etc ports as they all have different hardware costs on the switch.)

What can go wrong? Competition

- Too many exchange points in one locale
  - Competing exchanges defeats the purpose
- Becomes expensive for ISPs to connect to all of them
  - So they don’t, or won’t, and local traffic suffers, defeating the viability of the IXPs
- An IXP:
  - is **NOT** a competition
  - is **NOT** a profit making business

What can go wrong?
Rules and Restrictions

- IXP tries to compete with their membership
  - Offering services that ISPs would/do offer their customers
  - **In reality, IXPs are operated by the members for the members**
- IXP is run as a closed privileged club e.g.:
  - Restrictive membership criteria
  - **In reality, a participant needs to have an ASN and their own independent address space**
- IXP located in a data centre with restricted physical/transmission access
  - **IXP must be a neutral interconnect in a neutral location**

What can go wrong? Rules and Restrictions

- IXP charges for traffic
  - So do transit providers – **charging for traffic is a sure way of ending the viability of the IXP**
- IXPs providing access to end users rather than just Network Operators & Service Providers
  - **A participant at an IXP needs to have their own address space, their own ASN, and their own transit arrangements**
- IXPs interfering with member business decisions
  - **The most common error: Mandatory Multi-Lateral Peering**

What can go wrong? Technical Design Errors

- Interconnected IXPs
  - IXP in one location believes it should connect directly to the IXP in another location
  - Who pays for the interconnect?
  - How is traffic metered?
  - Competes with the ISPs who already provide transit between the two locations (who then refuse to join IX, harming the viability of the IX)
  - Metro interconnections work ok

What can go wrong? Technical Design Errors

- ISPs bridge the IXP LAN back to their offices
  - "We are poor, we can't afford a router"
  - Financial benefits of connecting to an IXP far outweigh the cost of a router
  - In reality it allows the ISP to connect any devices to the IXP LAN — with disastrous consequences for the security, integrity and reliability of the IXP

What can go wrong? Routing Design Errors

- Route Server implemented from Day One
  - ISPs have no incentive to learn BGP
  - Therefore have no incentive to understand peering relationships, peering policies, &c
  - Entirely dependent on operator of RS for troubleshooting, configuration, reliability
    - RS can't be run by committee!
- Route Server is to help scale peering at LARGE IXPs

What can go wrong? Routing Design Errors (cont)

- iBGP Route Reflector used to distribute prefixes between IXP participants
- Claimed Advantage (1):
  - Participants don't need to know about or run BGP
- Actually a Disadvantage
  - IXP Operator has to know BGP
  - ISP not knowing BGP is big commercial disadvantage
  - ISPs who would like to have a growing successful business need to be able to multi-home, peer with other ISPs, etc — these activities require BGP

What can go wrong? Routing Design Errors (cont)

- Route Reflector Claimed Advantage (2):
  - Allows an IXP to be started very quickly
- Fact:
  - IXP is only an Ethernet switch — setting up an iBGP Route Reflector mesh with participants is no quicker than setting up an eBGP Route Server mesh
    - But the latter scales, and works

What can go wrong? Routing Design Errors (cont)

- Route Reflector Claimed Advantage (3):
  - IXP operator has full control over IXP activities
- Actually a Disadvantage
  - ISP participants surrender control of:
    - Their border router; it is located in IXP's AS
    - Their routing and peering policy
  - IXP operator is single point of failure
    - If they aren't available 24x7, then neither is the IXP
    - BGP configuration errors by IXP operator have real impacts on ISP operations

What can go wrong? Routing Design Errors (cont)

- Route Reflector Disadvantage (4):
  - Migration from Route Reflector to "correct" routing configuration is highly non-trivial
  - ISP router is in IXP's ASN
    - Need to move ISP router from IXP's ASN to the ISP's ASN
    - Need to reconfigure BGP on ISP router, add to ISP's IGP and iBGP mesh, and set up eBGP with IXP participants and/ or the IXP Route Server

More Information

Exchange Point Policies & Politics

- AUPs
  - Acceptable Use Policy
  - Minimal rules for connection
- Fees?
  - Some IXPs charge no fee
  - Other IXPs charge cost recovery
  - A few IXPs are commercial
- Nobody is obliged to peer
  - Agreements left to ISPs, not mandated by IXP

Exchange Point etiquette

- Don’t point default route at another IXP participant
- Be aware of third-party next-hop
- Only announce your aggregate routes
  - Read RIPE-399 and RIPE-532 first

    www.ripe.net/ripe/docs/ripe-399

    www.ripe.net/ripe/docs/ripe-532
- Filter! Filter! Filter!

Exchange Point Examples

- LINX in London, UK
- TorIX in Toronto, Canada
- AMS-IX in Amsterdam, Netherlands
- SIX in Seattle, Washington, US
- PA-IX in Palo Alto, California, US
- JPNAP in Tokyo, Japan
- DE-CIX in Frankfurt, Germany
- HK-IX in Hong Kong

…

- All use Ethernet Switches

Features of IXPs (1)

- Redundancy & Reliability
  - Multiple switches, UPS/Generator
- Support
  - NOC to provide 24x7 support for problems at the exchange
- DNS, Route Collector/Server, Content Caches & NTP servers
  - ccTLD & root servers
  - Content caches such as Google Global Cache
  - Content redistribution systems such as Akamai
  - Route Collector – Routing Table view

Features of IXPs (2)

- Location
  - Neutral, secure & accessible co-location facilities
- Address space
  - Public address for Peering LAN
  - Public address for IXP Services LAN
- AS Number
  - Private ASN needed for Route Collector/Server
  - Public ASN needed for IXP Services
- Route servers (for larger IXPs)
- Statistics
  - Traffic data – for membership

More info about IXPs

- Euro-IX
  - European Internet Exchange consortium
  - All the information needed to start an IXP
  - https://www.euro-ix.net/starting-an-ixp
- PCH
  - https://www.pch.net/resources/papers.php
  - Excellent collection of IXP locations, discussion papers, IXP statistics, etc
- Telegeography
  - http://www.telegeography.com/telecom-resources/internet-exchange-map/
  - A collection of IXPs and interconnect points for ISPs

Summary

- L2 IXP – universally deployed
  - The core is an ethernet switch
  - ATM and other old technologies are obsolete
- L3 IXP – is just a marketing concept used by wholesale ISPs
  - Does not offer the same flexibility, neutrality or effectiveness as L2
  - Not recommended unless there are overriding regulatory or political reasons to do so
  - **Avoid!**

Internet Exchange Point Design

End of Tutorial