ISP & IXP Design

Philip Smith
MENOG 11
Amman
30th September – 9th October 2012
ISP & IXP Network Design

- PoP Topologies and Design
- Backbone Design
- Upstream Connectivity & Peering
- Addressing
- Routing Protocols
- Out of Band Management
- Operational Considerations
- Internet Exchange Points
Point of Presence
Topologies
PoP Topologies

- **Core** routers – high speed trunk connections
- **Distribution** routers and **Access** routers – high port density
- **Border** routers – connections to other providers
- **Service** routers – hosting and servers
- Some functions might be handled by a single router
PoP Design

- Modular Design
- Aggregation Services separated according to
  - connection speed
  - customer service
  - contention ratio
  - security considerations
Modular PoP Design

- Backbone link to another PoP
- Business customer aggregation layer
  - Channelised circuits for leased line circuit delivery
- MetroE customer aggregation layer
  - GigE fibre trunks for MetroE circuit delivery
- Consumer Dial Access
- Network Core
- Network Operations Centre
- Other ISPs
- Hosted Services & Datacentre
- ISP Services (DNS, Mail, News, FTP, WWW)
- Web Cache
- Consumer cable, xDSL and wireless Access
- ISP Services (DNS, Mail, News, FTP, WWW)
- MetroE customer aggregation layer
  - GigE fibre trunks for MetroE circuit delivery
- ISP Services (DNS, Mail, News, FTP, WWW)
- Web Cache
- Consumer cable, xDSL and wireless Access
Modular Routing Protocol Design

- Modular IGP implementation
  - IGP “area” per PoP
  - Core routers in backbone area (Area 0/L2)
  - Aggregation/summarisation where possible into the core

- Modular iBGP implementation
  - BGP route reflector cluster
  - **Core routers** are the route-reflectors
  - Remaining routers are clients & peer with route-reflectors only
Point of Presence Design
PoP Modules

- **Low Speed customer connections**
  - PSTN/ISDN dialup
  - Low bandwidth needs
  - Low revenue, large numbers

- **Leased line customer connections**
  - E1/T1 speed range
  - Delivery over channelised media
  - Medium bandwidth needs
  - Medium revenue, medium numbers
PoP Modules

- Broad Band customer connections
  - xDSL, Cable and Wireless
  - High bandwidth needs
  - Low revenue, large numbers

- MetroE & Highband customer connections
  - Trunk onto GigE or 10GigE of 10Mbps and higher
  - Channelised OC3/12 delivery of E3/T3 and higher
  - High bandwidth needs
  - High revenue, low numbers
PoP Modules

- PoP Core
  - Two dedicated routers
  - High Speed interconnect
  - Backbone Links ONLY
  - Do not touch them!

- Border Network
  - Dedicated border router to other ISPs
  - The ISP’s “front” door
  - Transparent web caching?
  - Two in backbone is minimum guarantee for redundancy
PoP Modules

- ISP Services
  - DNS (cache, secondary)
  - News (still relevant?)
  - Mail (POP3, Relay, Anti-virus/anti-spam)
  - WWW (server, proxy, cache)

- Hosted Services/DataCentres
  - Virtual Web, WWW (server, proxy, cache)
  - Information/Content Services
  - Electronic Commerce
PoP Modules

- Network Operations Centre
  - Consider primary and backup locations
  - Network monitoring
  - Statistics and log gathering
  - Direct but secure access

- Out of Band Management Network
  - The ISP Network “Safety Belt”
Low Speed Access Module

- Primary Rate T1/E1
- PSTN lines to modem bank
- PSTN lines to built-in modems
- TACACS+/Radius proxy, DNS resolver, Content
- Access Servers
- Access Network Gateway Routers
- Web Cache
- To Core Routers
Medium Speed Access Module

- Channelised T1/E1
- 64K and nx64K circuits
- Mixture of channelised T1/E1, 56/64K and nx64K circuits

Aggregation Edge

To Core Routers
High Speed Access Module

Aggregation Edge

- Metro Ethernet
- Channelised T3/E3
- Channelised OC3/OC12

To Core Routers
Broadband Access Module

Telephone Network

The cable system

DSLAM

BRAS

Web Cache

Access Network Gateway Routers

Cable RAS

SSG, DHCP, TACACS+
or Radius Servers/Proxies,
DNS resolver, Content

To Core Routers
ISP Services Module

- DNS cache
- DNS secondary
- POP3 Mail Relay
- NEWS
- DNS cache

To core routers

Service Network Gateway Routers

WWW cache
Hosted Services Module

To core routers

Hosted Network Gateway Routers

Customer 1
Customer 2
Customer 3
Customer 4
Customer 5
Customer 6
Customer 7

vlan11 vlan12 vlan13 vlan14 vlan15 vlan16 vlan17
Border Module

**NB:** router has no default route + local AS routing table only
NOC Module

Out of Band Management Network

2811/32async

NetFlow Analyser
TACACS+ server
SYSLOG server
Primary DNS

Hosted Network Gateway Routers

Network Operations Centre Staff

To core routers

Critical Services Module

Corporate LAN

Firewall

Billing, Database and Accounting Systems

Critical Services Module

Network Operations Centre Staff

21
Out of Band Network

Out of Band Management Network

Router consoles

Terminal server

To the NOC

NetFlow Collector

Out of Band Ethernet

NetFlow enabled routers
Backbone Network Design
Backbone Design

- Routed Backbone
- Switched Backbone
  - ATM/Frame Relay core network
  - Now obsolete
- Point-to-point circuits
  - nx64K, T1/E1, T3/E3, OC3, OC12, GigE, OC48, 10GigE, OC192, OC768
- ATM/Frame Relay service from telco
  - T3, OC3, OC12, ... delivery
  - Easily upgradeable bandwidth (CIR)
  - Almost vanished in availability now
Distributed Network Design

- PoP design “standardised”
  - operational scalability and simplicity
- ISP essential services distributed around backbone
- NOC and “backup” NOC
- Redundant backbone links
Distributed Network Design

POP One

POP Two

POP Three

ISP Services

Customer connections

Backup Operations Centre

Operations Centre

External connections

External connections

ISP Services

Customer connections

Customer connections
Backbone Links

- **ATM/Frame Relay**
  - Virtually disappeared due to overhead, extra equipment, and shared with other customers of the telco
  - MPLS has replaced ATM & FR as the telco favourite

- **Leased Line/Circuit**
  - Most popular with backbone providers
  - IP over Optics and Metro Ethernet very common in many parts of the world
Long Distance Backbone Links

- These usually cost more
- Important to plan for the future
  - This means at least two years ahead
  - Stay in budget, stay realistic
  - Unplanned “emergency” upgrades will be disruptive without redundancy in the network infrastructure
Long Distance Backbone Links

- Allow sufficient capacity on alternative paths for failure situations
  - Sufficient can depend on the business strategy
  - Sufficient can be as little as 20%
  - Sufficient is usually over 50% as this offers “business continuity” for customers in the case of link failure
  - Some businesses choose 0%
    - Very short sighted, meaning they have no spare capacity at all!!
Long Distance Links
Metropolitan Area Backbone Links

- Tend to be cheaper
  - Circuit concentration
  - Choose from multiple suppliers

- Think big
  - More redundancy
  - Less impact of upgrades
  - Less impact of failures
Metropolitan Area Backbone Links

Traditional Point to Point Links
Upstream Connectivity and Peering
Transits

- Transit provider is another autonomous system which is used to provide the local network with access to other networks
  - Might be local or regional only
  - But more usually the whole Internet

- Transit providers need to be chosen wisely:
  - Only one
    - no redundancy
  - Too many
    - more difficult to load balance
    - no economy of scale (costs more per Mbps)
    - hard to provide service quality

- **Recommendation:** at least two, no more than three
Common Mistakes

- ISPs sign up with too many transit providers
  - Lots of small circuits (cost more per Mbps than larger ones)
  - Transit rates per Mbps reduce with increasing transit bandwidth purchased
  - Hard to implement reliable traffic engineering that doesn’t need daily fine tuning depending on customer activities

- No diversity
  - Chosen transit providers all reached over same satellite or same submarine cable
  - Chosen transit providers have poor onward transit and peering
Peers

- A peer is another autonomous system with which the local network has agreed to exchange locally sourced routes and traffic

- Private peer
  - Private link between two providers for the purpose of interconnecting

- Public peer
  - Internet Exchange Point, where providers meet and freely decide who they will interconnect with

- Recommendation: peer as much as possible!
Common Mistakes

- Mistaking a transit provider’s “Exchange” business for a no-cost public peering point
- Not working hard to get as much peering as possible
  - Physically near a peering point (IXP) but not present at it
  - (Transit is rarely cheaper than peering!!)
- Ignoring/avoiding competitors because they are competition
  - Even though potentially valuable peering partner to give customers a better experience
Private Interconnection

- Two service providers agree to interconnect their networks
- They exchange prefixes they originate into the routing system (usually their aggregated address blocks)
- They share the cost of the infrastructure to interconnect
  - Typically each paying half the cost of the link (be it circuit, satellite, microwave, fibre,...)
  - Connected to their respective peering routers
- Peering routers only carry domestic prefixes
Private Interconnection

- PR = peering router
  - Runs iBGP (internal) and eBGP (with peer)
  - No default route
  - No “full BGP table”
  - Domestic prefixes only
- Peering router used for all private interconnects
Public Interconnection

- Service provider participates in an Internet Exchange Point
  - It exchanges prefixes it originates into the routing system with the participants of the IXP
  - It chooses who to peer with at the IXP
    - Bi-lateral peering (like private interconnect)
    - Multi-lateral peering (via IXP’s route server)
  - It provides the router at the IXP and provides the connectivity from their PoP to the IXP
  - The IXP router carries only domestic prefixes
ISP1-PR = peering router of our ISP
- Runs iBGP (internal) and eBGP (with IXP peers)
- No default route
- No “full BGP table”
- Domestic prefixes only

Physically located at the IXP
Public Interconnection

- The ISP’s router IXP peering router needs careful configuration:
  - It is remote from the domestic backbone
  - Should not originate any domestic prefixes
  - (As well as no default route, no full BGP table)
  - Filtering of BGP announcements from IXP peers (in and out)
Upstream/Transit Connection

Two scenarios:

- Transit provider is in the locality
  - Which means bandwidth is cheap, plentiful, easy to provision, and easily upgraded
- Transit provider is a long distance away
  - Over undersea cable, satellite, long-haul cross country fibre, etc

Both scenarios have different requirements which need to be considered
Local Transit Provider

- BR = ISP’s Border Router
  - Runs iBGP (internal) and eBGP (with transit)
  - Either receives default route or the full BGP table from upstream
  - BGP policies are implemented here (depending on connectivity)
  - Packet filtering is implemented here (as required)
BR = ISP’s Border Router
- Co-located in a co-lo centre (typical) or in the upstream provider’s premises
- Runs iBGP with rest of ISP1 backbone
- Runs eBGP with transit provider router(s)
- Implements BGP policies, packet filtering, etc
- Does not originate any domestic prefixes
Distant Transit Provider

- Positioning a router close to the Transit Provider’s infrastructure is strongly encouraged:
  - Long haul circuits are expensive, so the router allows the ISP to implement appropriate filtering first
  - Moves the buffering problem away from the Transit provider
  - Remote co-lo allows the ISP to choose another transit provider and migrate connections with minimum downtime
Distant Transit Provider

- Other points to consider:
  - Does require remote hands support
  - (Remote hands would plug or unplug cables, power cycle equipment, replace equipment, etc as instructed)
  - Appropriate support contract from equipment vendor(s)
  - Sensible to consider two routers and two long-haul links for redundancy
Summary

- Design considerations for:
  - Private interconnects
    - Simple private peering
  - Public interconnects
    - Router co-lo at an IXP
  - Local transit provider
    - Simple upstream interconnect
  - Long distance transit provider
    - Router remote co-lo at datacentre or Transit premises
Addressing
Getting IPv4 & IPv6 address space

- Take part of upstream ISP’s PA space
  
or
- Become a member of your Regional Internet Registry and get your own allocation
  - Require a plan for a year ahead
  - General policies are outlined in RFC2050, more specific details are on the individual RIR website
- There is no more IPv4 address space at IANA
  - APNIC & RIPE NCC are now in their “final /8” IPv4 delegation policy phase
  - Limited IPv4 available
  - IPv6 allocations are simple to get in most RIR regions
What about RFC1918 addressing?

- RFC1918 defines IPv4 addresses reserved for private Internets
  - Not to be used on Internet backbones
- Commonly used within end-user networks
  - NAT used to translate from private internal to public external addressing
  - Allows the end-user network to migrate ISPs without a major internal renumbering exercise
- ISPs must filter RFC1918 addressing at their network edge
What about RFC1918 addressing?

- There is a long list of well known problems:
- Including:
  - False belief it conserves address space
  - Adverse effects on Traceroute
  - Effects on Path MTU Discovery
  - Unexpected interactions with some NAT implementations
  - Interactions with edge anti-spoofing techniques
  - Peering using loopbacks
  - Adverse DNS Interaction
  - Serious Operational and Troubleshooting issues
  - Security Issues
    - false sense of security, defeating existing security techniques
What about RFC1918 addressing?

- Infrastructure Security: not improved by using private addressing
  - Still can be attacked from inside, or from customers, or by reflection techniques from the outside

- Troubleshooting: made an order of magnitude harder
  - No Internet view from routers
  - Other ISPs cannot distinguish between down and broken

- Summary:
  - ALWAYS use globally routable IP addressing for ISP Infrastructure
Addressing Plans – ISP
Infrastructure

- Address block for router loop-back interfaces
- Address block for infrastructure
  - Per PoP or whole backbone
  - Summarise between sites if it makes sense
  - Allocate according to genuine requirements, not historic classful boundaries
- Similar allocation policies should be used for IPv6 as well
  - ISPs just get a substantially larger block (relatively) so assignments within the backbone are easier to make
Addressing Plans – Customer

- Customers are assigned address space according to need
- Should not be reserved or assigned on a per PoP basis
  - ISP iBGP carries customer nets
  - Aggregation not required and usually not desirable
Addressing Plans (contd)

- Document infrastructure allocation
  - Eases operation, debugging and management

- Document customer allocation
  - Contained in iBGP
  - Eases operation, debugging and management
  - Submit network object to RIR Database
Routing Protocols
Routing Protocols

- **IGP – Interior Gateway Protocol**
  - Carries infrastructure addresses, point-to-point links
  - Examples are OSPF, ISIS,…

- **EGP – Exterior Gateway Protocol**
  - Carries customer prefixes and Internet routes
  - Current EGP is BGP version 4

- No connection between IGP and EGP
Why Do We Need an IGP?

- ISP backbone scaling
  - Hierarchy
  - Modular infrastructure construction
  - Limiting scope of failure
  - Healing of infrastructure faults using dynamic routing with fast convergence
Why Do We Need an EGP?

- Scaling to large network
  - Hierarchy
  - Limit scope of failure
- Policy
  - Control reachability to prefixes
  - Merge separate organizations
  - Connect multiple IGPs
Interior versus Exterior Routing Protocols

- **Interior**
  - Automatic neighbour discovery
  - Generally trust your IGP routers
  - Prefixes go to all IGP routers
  - Binds routers in one AS together

- **Exterior**
  - Specifically configured peers
  - Connecting with outside networks
  - Set administrative boundaries
  - Binds AS’s together
Interior versus Exterior Routing Protocols

- **Interior**
  - Carries ISP infrastructure addresses only
  - ISPs aim to keep the IGP small for efficiency and scalability

- **Exterior**
  - Carries customer prefixes
  - Carries Internet prefixes
  - EGPs are independent of ISP network topology
Hierarchy of Routing Protocols

- Other ISPs
- BGP4
- BGP4 and OSPF/ISIS
- BGP4
- Static/BGP4
- IXP
- Customers
Routing Protocols: Choosing an IGP

- OSPF and ISIS have very similar properties
- Which to choose?
  - Choose which is appropriate for your operators’ experience
  - In most vendor releases, both OSPF and ISIS have sufficient “nerd knobs” to tweak the IGP’s behaviour
  - OSPF runs on IP
  - ISIS runs on infrastructure, alongside IP
  - ISIS supports both IPv4 and IPv6
  - OSPFv2 (IPv4) plus OSPFv3 (IPv6)
Routing Protocols: IGP Recommendations

- Keep the IGP routing table as small as possible
  - If you can count the routers and the point-to-point links in the backbone, that total is the number of IGP entries you should see

- IGP details:
  - Should only have router loopbacks, backbone WAN point-to-point link addresses, and network addresses of any LANs having an IGP running on them
  - Strongly recommended to use inter-router authentication
  - Use inter-area summarisation if possible
Routing Protocols: More IGP recommendations

- To fine tune IGP table size more, consider:
  - Using “ip unnumbered” on customer point-to-point links – saves carrying that /30 in IGP
    - (If customer point-to-point /30 is required for monitoring purposes, then put this in iBGP)
  - Use contiguous addresses for backbone WAN links in each area – then summarise into backbone area
  - Don’t summarise router loopback addresses – as iBGP needs those (for next-hop)
  - Use iBGP for carrying anything which does not contribute to the IGP Routing process
Routing Protocols: iBGP Recommendations

- iBGP should carry everything which doesn’t contribute to the IGP routing process
  - Internet routing table
  - Customer assigned addresses
  - Customer point-to-point links
  - Access network dynamic address pools, passive LANs, etc
Routing Protocols:
More iBGP Recommendations

- Scalable iBGP features:
  - Use neighbour authentication
  - Use peer-groups to speed update process and for configuration efficiency
  - Use communities for ease of filtering
  - Use route-reflector hierarchy
    - Route reflector pair per PoP (overlaid clusters)
Security
Security

- ISP Infrastructure security
- ISP Network security
- Security is not optional!

ISPs need to:
  - Protect themselves
  - Help protect their customers from the Internet
  - Protect the Internet from their customers

The following slides are general recommendations
  - Do more research on security before deploying any network
ISP Infrastructure Security

- **Router & Switch Security**
  - Use Secure Shell (SSH) for device access & management
    - Do NOT use Telnet
  - Device management access filters should only allow NOC and device-to-device access
    - Do NOT allow external access
  - Use TACACS+ for user authentication and authorisation
    - Do NOT create user accounts on routers/switches
Remote access

- For Operations Engineers who need access while not in the NOC
- Create an SSH server host (this is all it does)
  - Or a Secure VPN access server
- Ops Engineers connect here, and then they can access the NOC and network devices
ISP Infrastructure Security

- Other network devices?
  - These probably do not have sophisticated security techniques like routers or switches do
  - Protect them at the LAN or point-to-point ingress (on router)

- Servers and Services?
  - Protect servers on the LAN interface on the router
  - Consider using iptables &c on the servers too

- SNMP
  - Apply access-list to the SNMP ports
  - Should only be accessible by management system, not the world
General Advice:

- Routers, Switches and other network devices should not be contactable from outside the AS
- Achieved by blocking typical management access protocols for the infrastructure address block at the network perimeter
  - E.g. ssh, telnet, http, snmp,…
- Use the ICSI Netalyser to check access levels:
  - [http://netalyzr.icsi.berkeley.edu](http://netalyzr.icsi.berkeley.edu)
- Don’t block everything: BGP, traceroute and ICMP still need to work!
Effective filtering

- Protect network borders from “traffic which should not be on the public Internet”, for example:
  - LAN protocols (e.g., netbios)
  - Well known exploit ports (used by worms and viruses)
  - Drop traffic arriving and going to private and non-routable address space (IPv4 and IPv6)

- Achieved by packet filters on border routers
  - Remote trigger blackhole filtering
ISP Network Security – RTBF

- Remote trigger blackhole filtering
  - ISP NOC injects prefixes which should not be accessible across the AS into the iBGP
  - Prefixes have next hop pointing to a blackhole address
  - All iBGP speaking backbone routers configured to point the blackhole address to the null interface
  - Traffic destined to these blackhole prefixes are dropped by the first router they reach

- Application:
  - Any prefixes (including RFC1918) which should not have routability across the ISP backbone
Remote trigger blackhole filtering example:

- Origin router:
  
  ```
  router bgp 64509
  
  redistribute static route-map black-hole-trigger
  
  ip route 10.5.1.3 255.255.255.255 Null0 tag 66
  
  route-map black-hole-trigger permit 10
  
  match tag 66
  
  set local-preference 1000
  
  set community no-export
  
  set ip next-hop 192.0.2.1
  
  ```

- iBGP speaking backbone router:
  
  ```
  ip route 192.0.2.1 255.255.255.255 null0
  ```
Resulting routing table entries:

gw1#sh ip bgp 10.5.1.3
BGP routing table entry for 10.5.1.3/32, version 64572219
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  Local
    192.0.2.1 from 1.1.10.10 (1.1.10.10)
      Origin IGP, metric 0, localpref 1000, valid, internal, best
      Community: no-export

gw1#sh ip route 10.5.1.3
Routing entry for 10.5.1.3/32
  Known via "bgp 64509", distance 200, metric 0, type internal
  Last update from 192.0.2.1 00:04:52 ago
  Routing Descriptor Blocks:
    * 192.0.2.1, from 1.1.10.10, 00:04:52 ago
      Route metric is 0, traffic share count is 1
      AS Hops 0
ISP Network Security – uRPF

- Unicast Reverse Path Forwarding
- Strongly recommended to be used on all customer facing static interfaces
  - BCP 38 ([tools.ietf.org/html/bcp38](https://tools.ietf.org/html/bcp38))
  - Blocks all unroutable source addresses the customer may be using
  - Inexpensive way of filtering customer’s connection (when compared with packet filters)
- Can be used for multihomed connections too, but extreme care required
What is uRPF?

- Router compares source address of incoming packet with FIB entry
  - If FIB entry interface matches incoming interface, the packet is forwarded
  - If FIB entry interface does not match incoming interface, the packet is dropped
Security Summary

- Implement RTBF
  - Inside ISP backbone
  - Make it available to BGP customers too
    - They can send you the prefix you need to block with a special community attached
    - You match on that community, and set the next-hop to the null address
- Implement uRPF
  - For all static customers
- Use SSH for device access
- Use TACACS+ for authentication
Out of Band Management
Out of Band Management

- Not optional!
- Allows access to network equipment in times of failure
- Ensures quality of service to customers
  - Minimises downtime
  - Minimises repair time
  - Eases diagnostics and debugging
Out of Band Management

- OoB Example – Access server:
  - modem attached to allow NOC dial in
  - console ports of all network equipment connected to serial ports
  - LAN and/or WAN link connects to network core, or via separate management link to NOC

- Full remote control access under all circumstances
Out of Band Network

Router, switch and ISP server consoles

(Optional) Out of band WAN link to other PoPs

Modem – access to PSTN for out of band dialin

Ethernet to the NOC
Out of Band Management

- OoB Example – Statistics gathering:
  - Routers are NetFlow and syslog enabled
  - Management data is congestion/failure sensitive
  - Ensures management data integrity in case of failure
- Full remote information under all circumstances
Test Laboratory
Test Laboratory

- Designed to look like a typical PoP
  - Operated like a typical PoP
- Used to trial new services or new software under realistic conditions
- Allows discovery and fixing of potential problems before they are introduced to the network
Test Laboratory

- Some ISPs dedicate equipment to the lab.
- Other ISPs “purchase ahead” so that today’s lab equipment becomes tomorrow’s PoP equipment.
- Other ISPs use lab equipment for “hot spares” in the event of hardware failure.
Can’t afford a test lab?
- Set aside one spare router and server to trial new services
- Never ever try out new hardware, software or services on the live network

Every major ISP in the US and Europe has a test lab
- It’s a serious consideration
Operational Considerations
Why design the world’s best network when you have not thought about what operational good practices should be implemented?
Operational Considerations

Maintenance

- Never work on the live network, no matter how trivial the modification may seem
  - Establish maintenance periods which your customers are aware of
    - e.g. Tuesday 4-7am, Thursday 4-7am
- Never do maintenance on the last working day before the weekend
  - Unless you want to work all weekend cleaning up
- Never do maintenance on the first working day after the weekend
  - Unless you want to work all weekend preparing
Operational Considerations

Support

- Differentiate between customer support and the Network Operations Centre
  - Customer support fixes customer problems
  - NOC deals with and fixes backbone and Internet related problems

- Network Engineering team is last resort
  - They design the next generation network, improve the routing design, implement new services, etc
  - They do not and should not be doing support!
NOC Communications

- NOC should know contact details for equivalent NOCs in upstream providers and peers
ISP Network Design

Summary
ISP Design Summary

- KEEP IT SIMPLE & STUPID ! (KISS)
- Simple is elegant is is scalable
- Use Redundancy, Security, and Technology to make life easier for yourself
- Above all, ensure quality of service for your customers
Why an Internet Exchange Point?

Saving money, improving QoS, encouraging 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?

- 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
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
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
Exchange Point Design
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

IXP Services:
- Root & TLD DNS
- Routing Registry
- Looking Glass, etc
Layer 2 Exchange

IXP Services:
Root & TLD DNS,
Routing Registry
Looking Glass, etc
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 & TLD, NTP servers
  - Routing Registry and Looking Glass
Layer 2 Exchange

- Requires neutral IXP management
  - Usually funded equally by IXP participants
  - 24x7 cover, support, value add services
- Secure and neutral location
- Configuration
  - IPv4 /24 and IPv6 /64 for IXP LAN
  - ISPs require AS, basic IXP does not
Layer 2 Exchange

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 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
- 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

- Some 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
Charging

- IXPs should be run at minimal cost to participants
- Examples:
  - Datacentre hosts IX for free
    - Because ISP participants then use data centre for co-lo services, and the datacentre benefits long term
  - IX operates cost recovery
    - Each member pays a flat fee towards the cost of the switch, hosting, power & management
  - Different pricing for different ports
    - One slot may handle 24 10GE ports
    - Or one slot 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
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 MRTG/Cacti 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

- Looking Glass
  - One way of making the Route Collector routes available for global view (e.g. www.traceroute.org)
  - Public or members only access
  - Useful for members to check BGP filters
  - Useful for everyone to check route availability at the IX
Services to Offer

- **Route Server**
  - A Route Collector that also sends the prefixes it has collected to its peers
  - Like a Route Collector, usually a router or Unix based system running BGP
  - Does **not** forward packets
  - Useful for scaling eBGP sessions for larger IXPs
  - Participation needs to be optional
    - And will be used by ISPs who have open peering policies
Services to Offer

- Content Redistribution/Caching
  - For example, Akamised update distribution service

- 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
What can go wrong?

- High annual fees
  - Should be cost recovery
- Charging for traffic between participants
  - Competes with commercial transit services
- Competing IXPs
  - Too expensive for ISPs to connect to all
- Too many rules & restrictions
  - Want all network operators to participate
- Mandatory Multi-Lateral Peering
  - Has no history of success
- Interconnected IXPs
  - Who pays for the interconnection?
- Etc...
Conclusion

- IXPs are technically very simple to set up

- Little more than:
  - An ethernet switch
  - Neutral secure reliable location
  - Consortium of members to operate it

- Political aspects can be more challenging:
  - Competition between ISP members
  - “ownership” or influence by outside parties