ISP Network Design

ISP Workshops

Last updated 9th October 2018

Acknowledgements

- This material originated from the Cisco ISP/IXP Workshop Programme developed by Philip Smith & Barry Greene
  - I'd like to acknowledge the input from many network operators in the ongoing development of these slides, especially Mark Tinka of SEACOM for his contributions

- Bug fixes and improvements are welcomed
  - Please email *workshop (at) bgp4all.com*

Philip Smith

ISP Network Design

- PoP Topologies and Design
- Backbone Design
- Addressing
- Routing Protocols
- Infrastructure & Routing Security
- Out of Band Management
- Test Network
- Operational Considerations

Point of Presence Topology & Design

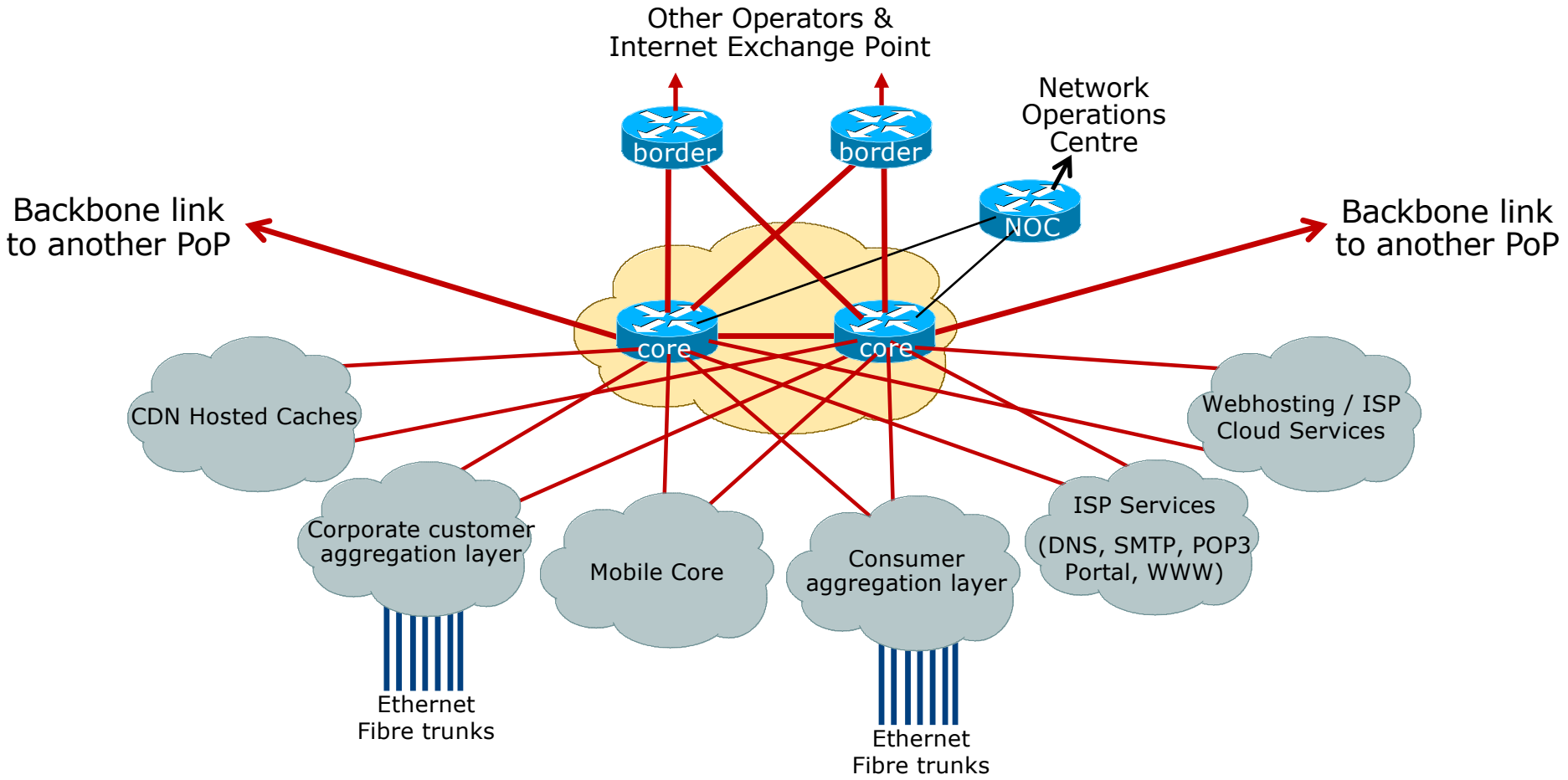
PoP Components

- Core routers
  - High speed trunk connections
- Distribution routers
  - For large networks, aggregating access to core
- Access routers
  - High port density connecting end-users
- Border routers
  - Connections to other providers
- Services routers
  - Hosting and servers
- Some functions might be handled by a single router

PoP Design

- Modular Design is essential
  - Quite often modules map on to business units in a network operator
- Aggregation Services separated according to
  - Connection speed
  - Customer service/expectations
  - Latency
  - Contention ratio
  - Technology
  - Security considerations

Modular PoP Design



Modular Routing Protocol Design

- IGP implementation
  - IS-IS is more common in larger operators
    - Entire backbone operates as ISIS Level 2
  - OSPFv2 & OSPFv3 also used
    - Backbone is in Area 0, each PoP in its own non-zero Area
- 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 Details

PoP Core

- Two dedicated high performance routers
- Technology
  - High Speed interconnect (10Gbps, 100Gbps, 400Gbps)
  - Backbone Links ONLY; no access services
  - *Do not touch them!*
- Service Profile
  - 24x7, high availability, duplicate/redundant design

PoP Core – details

- Router specification
  - High performance control plane CPU
  - Does not need a large number of interface/line cards
    - Only connecting backbone links and links to the various services
- High speed interfaces
  - Aim as high as possible
  - 10Gbps is the typical standard initial installation now
    - Price differential between 1Gbps and 10Gbps justifies the latter when looking at cost per Gbps
  - Many operators using aggregated 10Gbps links, also 100Gbps

Border Network

- Dedicated border routers to connect to other Network Operators
- Technology
  - High speed connection to core
  - Significant BGP demands, routing policy
  - DDoS front-line mitigation
  - Differentiation in use:
    - Connections to Upstream Providers (Transit links)
    - Connections to Private Peers and Internet Exchange Point
- Service Profile
  - 24x7, high availability, duplicate/redundant design

Border Network – details

- Router specification
  - High performance control plane CPU
  - Only needs a few interfaces
    - Only connecting to external operators and to the network core routers
  - Typically a 1RU or 2RU device
- High speed interfaces
  - 10Gbps standard to the core
  - 10Gbps to Internet Exchange Point
  - Ethernet towards peers (1Gbps upwards)
  - Ethernet towards transit providers (1Gbps upwards)

Border Network – details

- Router options:
  - Router dedicated to private peering and IXP connections
    - Only exchange routes originated by respective peers
      - No default, no full Internet routes
    - Control plane CPU needed for BGP routing table, applying policy, and assisting with DDoS mitigation
  - Router dedicated to transit connectivity
    - Must be separate device from private peering/IXP router
      - Usually carries full BGP table and/or default route
    - Control plane CPU needed for BGP routing table, applying policy, and assisting with DDoS mitigation
- Note: the ratio of peering traffic to transit traffic volume is around 3:1 today

Corporate Customer Aggregation

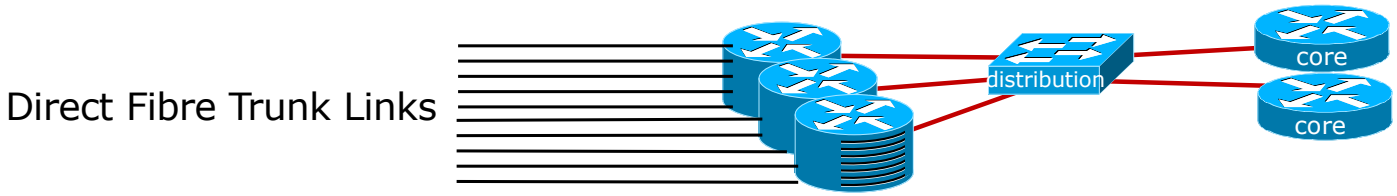
- Business customer connections
  - High value, high expectations
- Technology
  - Fibre to the premises (FTTx or GPON)
  - Aggregated within the PoP module
  - Usually managed service; customer premise router provided by the operator
- Service Profile
  - Typically demand peak performance during office hours
  - Out of hours backups to the “Cloud”

Corporate Customer Aggregation – details

- Router specification
  - Mid-performance control plane CPU
  - High interface densities
- Interface types:
  - 10Gbps uplink to core
  - Multiple 10Gbps trunks
    - Customer connections delivered per VLAN
    - Provided by intermediate ethernet switch or optical equipment

Corporate Customer Aggregation – details

- Router options:
  - Several smaller devices, aggregating multiple 1Gbps trunks to 10Gbps uplinks
    - Typically 1RU routers with 16 physical interfaces
      - 12 interfaces used for customer connections, 4 interfaces for uplinks
    - May need intermediate Distribution Layer (usually ethernet switch) to aggregate to core routers



  - One larger device, multiple aggregation interfaces, with multiple 10Gbps or single 100Gbps uplink to core
    - Typical 8RU or larger with >100 physical interfaces

Consumer Aggregation

- Home users and small business customer connections
  - Low value, high expectations
- Technology:
  - Fibre to the premises (FTTx or GPON)
  - Still find Cable, ADSL and 802.11 wireless used
  - Aggregated within the PoP module
  - Unmanaged service; with customer premise router provided by the customer
- Service Profile
  - Typically demand peak performance during evenings

Consumer Aggregation – details

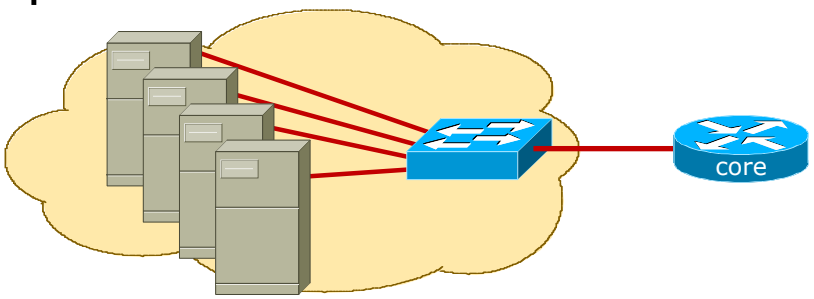
- Router specification
  - Mid-performance control plane CPU
  - High interface densities
- Interface types:
  - 10Gbps uplink to core
  - Multiple 10Gbps trunks
    - Customer connections delivered per VLAN
    - Provided by intermediate ethernet switch or optical equipment

CDN Hosted Services and Caches

- Content provider supplied infrastructure
- Technology:
  - Each CDN provides its own equipment
    - Usually a number of servers & ethernet switch, possibly a router
  - Requires direct and high bandwidth connection to the Core Network
    - Used for cache fill
    - Used to serve end-users
- Service Profile
  - High demand high availability 24x7

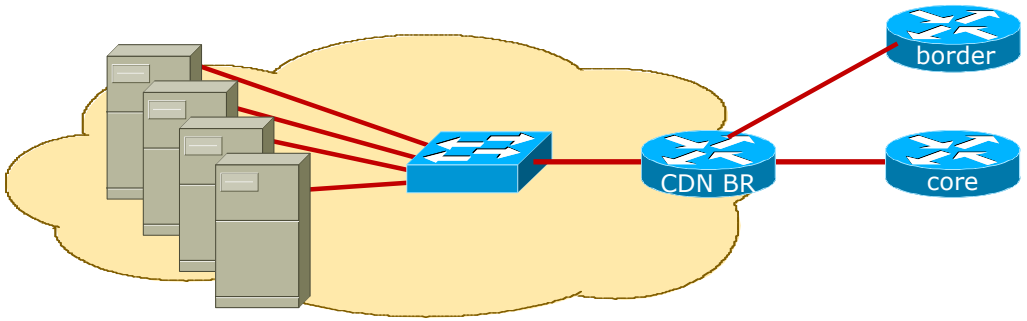
CDN Hosted Services and Caches – details

- Every CDN is different, but follow a similar pattern
  - Option 1:



Connection to Network Operator Core Router

  - Option 2:



Connection to Network Operator Border Router (Transit/Cache-Fill)

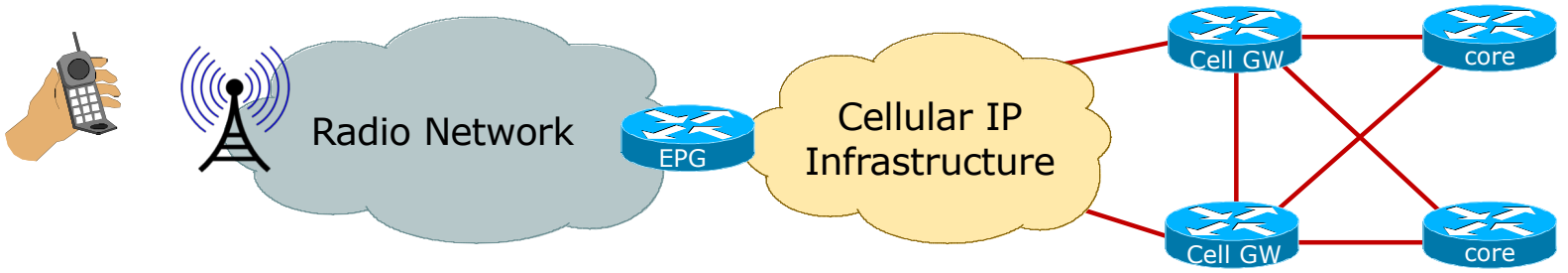
And Core Router (End-User Access)

Mobile Core

- Connection to Cellular Network infrastructure
- Technology:
  - Dedicated & redundant routers
  - Direct connection to Network Operator Core
- Service Profile
  - High demand high availability 24x7

Mobile Core – details

- Cellular network connectivity
  - Cellular infrastructure border routers (Cell GW) need to be:
    - High performance
    - High throughput
    - Able to do packet filtering as required

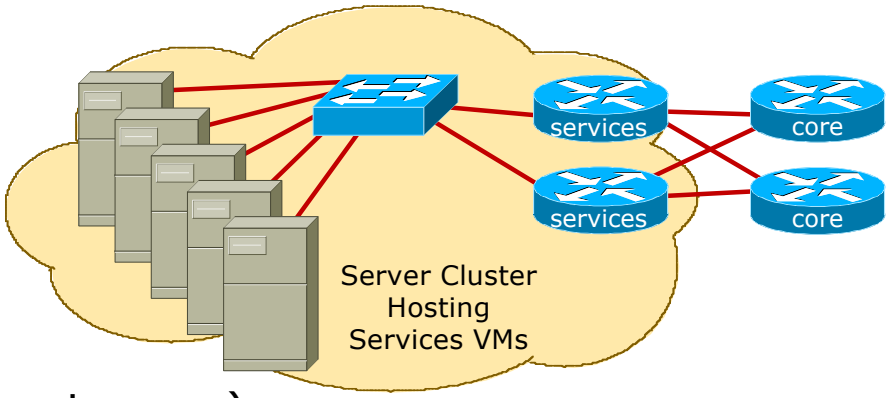


Network Operator Services

- Infrastructure / Customer services
- Technology:
  - Redundant server cluster behind two routers, hosting virtual machines
  - One virtual machine per service
- Services
  - DNS (2x cache, 2x authoritative)
  - Mail (SMTPS Relay for Customers, POP3S/IMAPS for Customers, SMTP for incoming e-mail)
  - WWW (Operator Website)
  - Portal (Customer Self-Service Portal)

Network Operator Services – details

- Infrastructure is usually multiple 1RU or 2RU servers configured into a cluster
  - Hosting Virtual Machines, one VM per Service
  - Examples:
    - WWW
    - Customer Portal
    - Authoritative DNS
    - DNS Cache (Resolver)
    - SMTP Host (incoming email)
    - SMTPS Relay (outgoing email from customers)
    - POP3S/IMAPS (Secure Mail Host for customers),

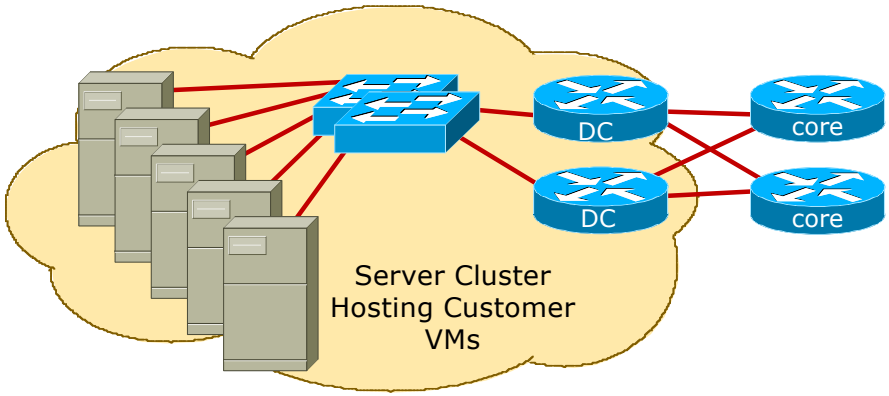


Webhosting/Cloud Module

- Hosted Services & DataCentre
  - "Cloud Computing" – or: someone else's computer!
- Technology
  - Redundant server cluster behind two routers, hosting virtual machines
  - One virtual machine per service
- Services
  - Content hosting / Websites (one VM per customer)
  - Compute Services (one VM per customer)
  - Backups (one VM per customer)

Cloud Module – details

- Infrastructure is usually multiple 1RU or 2RU servers configured into a cluster
  - Hosting Virtual Machines, one VM per Service
  - Several clusters
    - Limit the number of customers per cluster
  - Each customer gets one VM
    - Each VM in a separate private VLAN
    - Avoid exposing one customer VM to any other customer



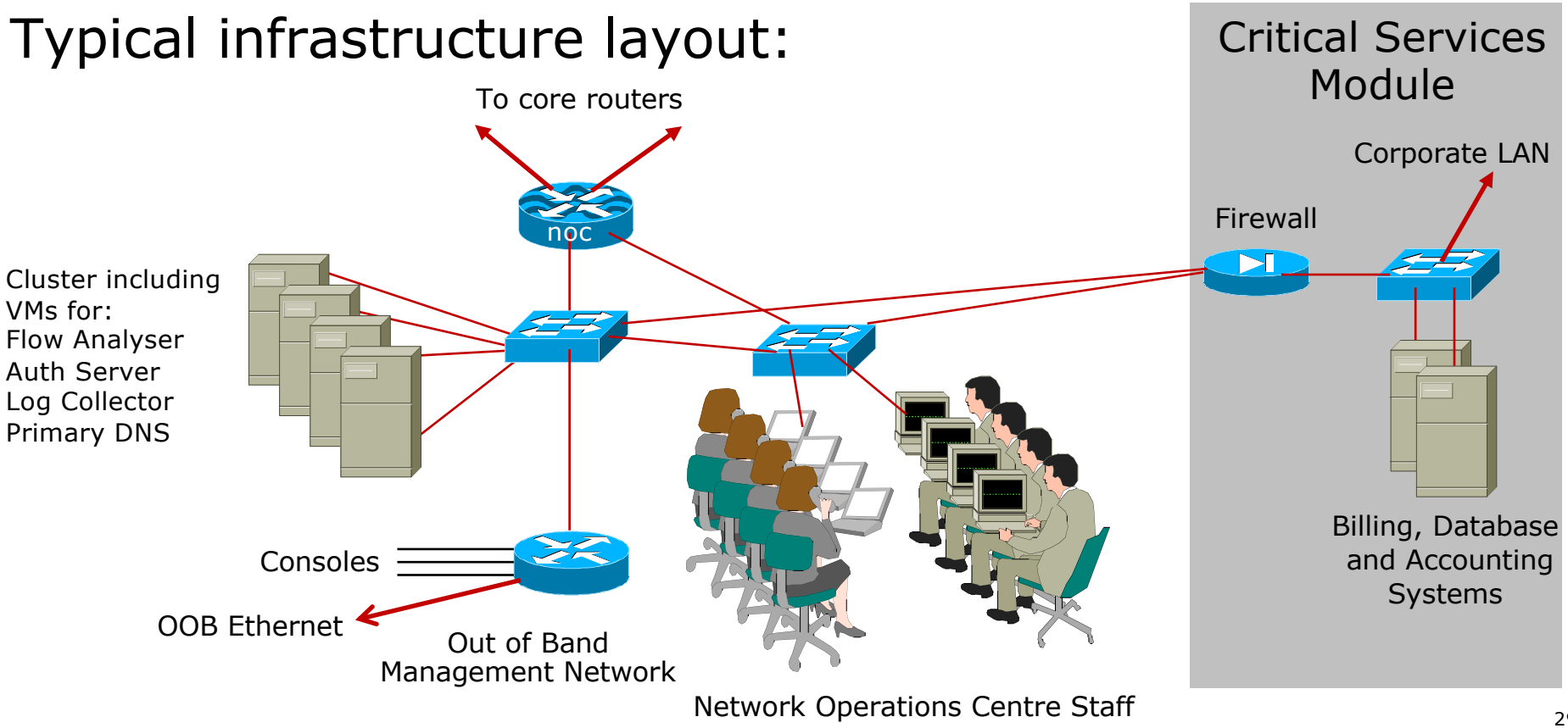
- Commercial and Open Source solutions available

Network Operations Centre

- Management of the network infrastructure
- Technology:
  - Gateway router, providing direct and secure access to the network operator core backbone infrastructure
- Services:
  - Network monitoring
  - Traffic flow monitoring and management
  - Statistics and log gathering
  - RTBH management for DDoS mitigation
  - Out of Band Management Network
    - The Network “Safety Belt”

NOC Module

- Typical infrastructure layout:



Summary

- Network Operator PoP core:
  - Modularity
  - High speed, no maintenance core
  - Direct Ethernet cross-connects
  - Two of everything
  - Rely on performance of IS-IS (or OSPF) and technologies such as BFD (Bi-directional Forwarding Detection) for rapid re-routing in case of device failure

Network Operator Backbone Infrastructure Design

Priorities

- Today's Internet is very different from 1990s
  - Back then, online content was via FTP sites, Gopher, bulletin boards, and early single location websites
- Today:
  - Dominance of content
  - Dominance of content distribution infrastructure & networks
- End user focus on social media, cloud services, and on-line videos/photos
  - i.e. Google/YouTube & Facebook accounts for 75% of traffic for an access provider
  - Access provider is merely a path between the CDN and the end-user

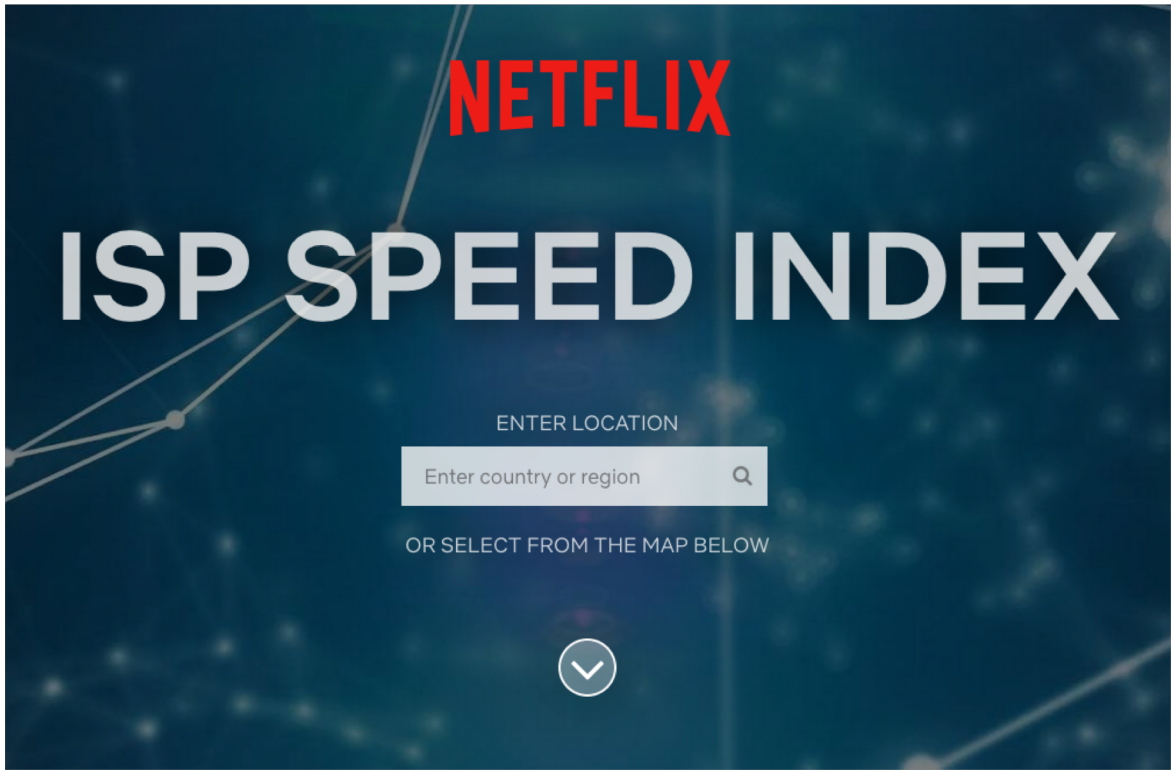
Priorities

- Priority for a service provider:
  - Providing lossless connectivity at high speed & high availability between content provider and end-user

- How:
  - Low latency backbone infrastructure
  - High bandwidth backbone infrastructure
  - Content Cache & Distribution Network Hosting
  - Interconnection with other local operators (private and IXP)
  - Optimised transit to content distribution hubs (for Cache fill)

Content delivery is competitive!

- Competition in local marketplace is all about speed and quality of content delivery
  - e.g.



These are NOT Priorities

- Last century's hierarchical transit / incumbent telco model
- Anti-competitive barriers between operators serving the same market
- Legislative barriers preventing interconnection

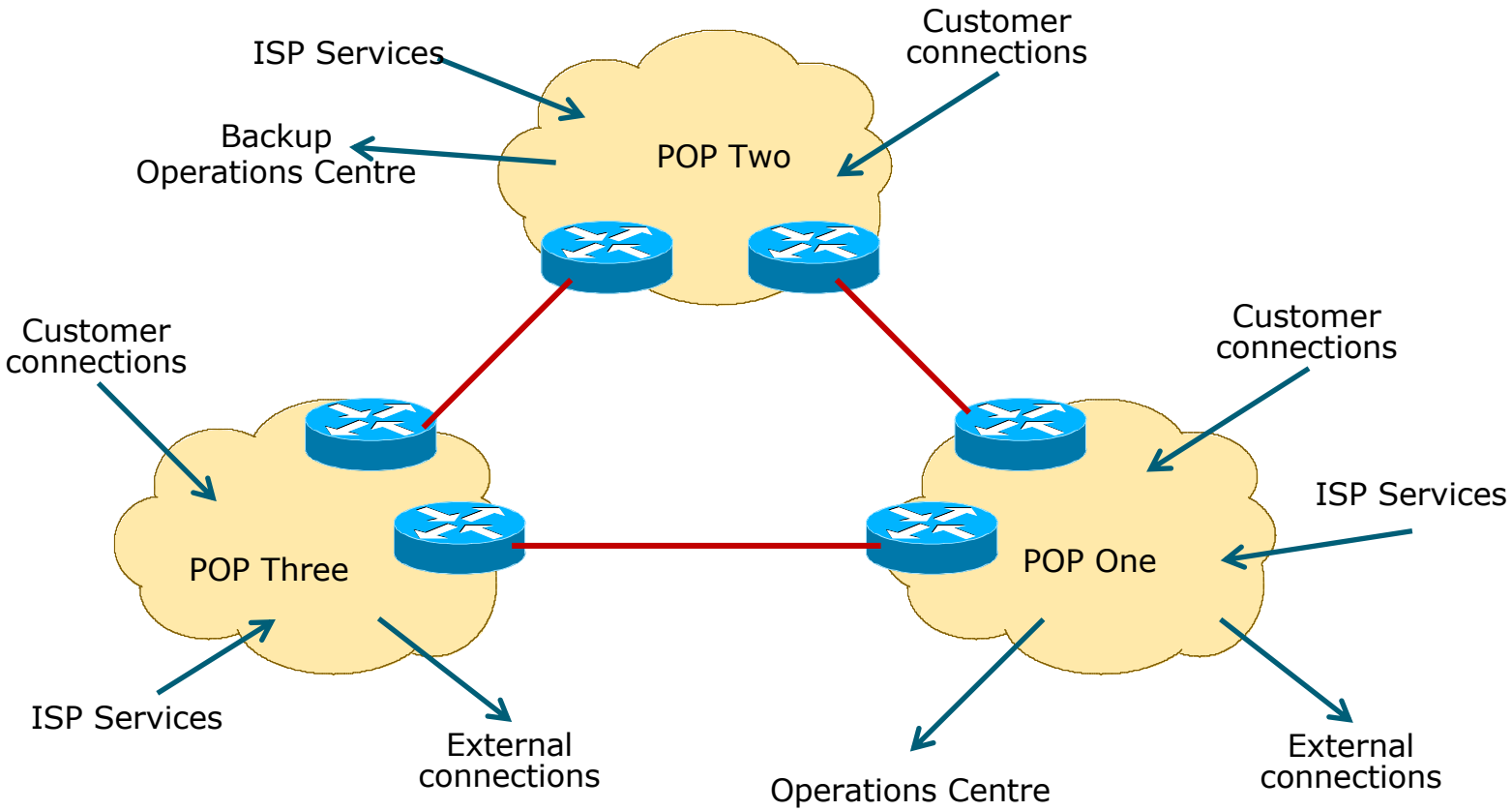
Backbone Design

- Routed Backbone
  - Some operators use MPLS for VPN service provision
- Point-to-point links using Fibre Optics
  - Ethernet (1GE, 10GE, 40GE, 100GE,…)
  - Packet over SONET (OC48, OC192, OC768)
- All other infrastructure technologies from the 90s and 00s are now obsolete
  - ATM, Frame Relay, PDH, X.25, FDDI,…

Distributed Network Design

- Important to standardise the PoP design
  - Nothing should be custom built
  - Settle on two or three standard designs (small/medium/large)
  - Using much the same hardware, same layout
  - And deploy across backbone as required
  - Maximises sparing, minimises operational complexity
- ISP essential services distributed around backbone
- NOC and “backup” NOC
- Redundant backbone links

Distributed Network Design



Backbone Links

- Fibre Optics
  - Most popular with most backbone operators today
  - Dark Fibre
    - Allows the operator to use the fibre pair as they please (implementing either CWDM or DWDM to increase the number of available channels)
    - Leased from fibre owner or purchased outright
  - Leased “lambdas”
    - Operator leases a wavelength from the fibre provider for data transmission
  - On the routers:
    - IP on Ethernet is used more and more for long haul
    - IP on SONET/SDH is more traditional long term

Fibre Optics – Brief Summary

- DWDM – Dense Wave Division Multiplexing
  - ITU-T G.694.1
  - Allows up to 96 wavelengths per fibre optic pair (transmit and receive)
  - λ: 1528 nm-1563 nm
  - 0.4 nm between channels
  - Costly, due to equipment and transceivers
- CWDM – Coarse Wave Division Multiplexing
  - ITU-T G.694.2
  - λ: 1271 nm-1611 nm
  - Allows up to 18 wavelengths per fibre optic pair (transmit and receive)
  - 20 nm between channels
  - Uses G.652.C and G.652.D specification fibre optic cables

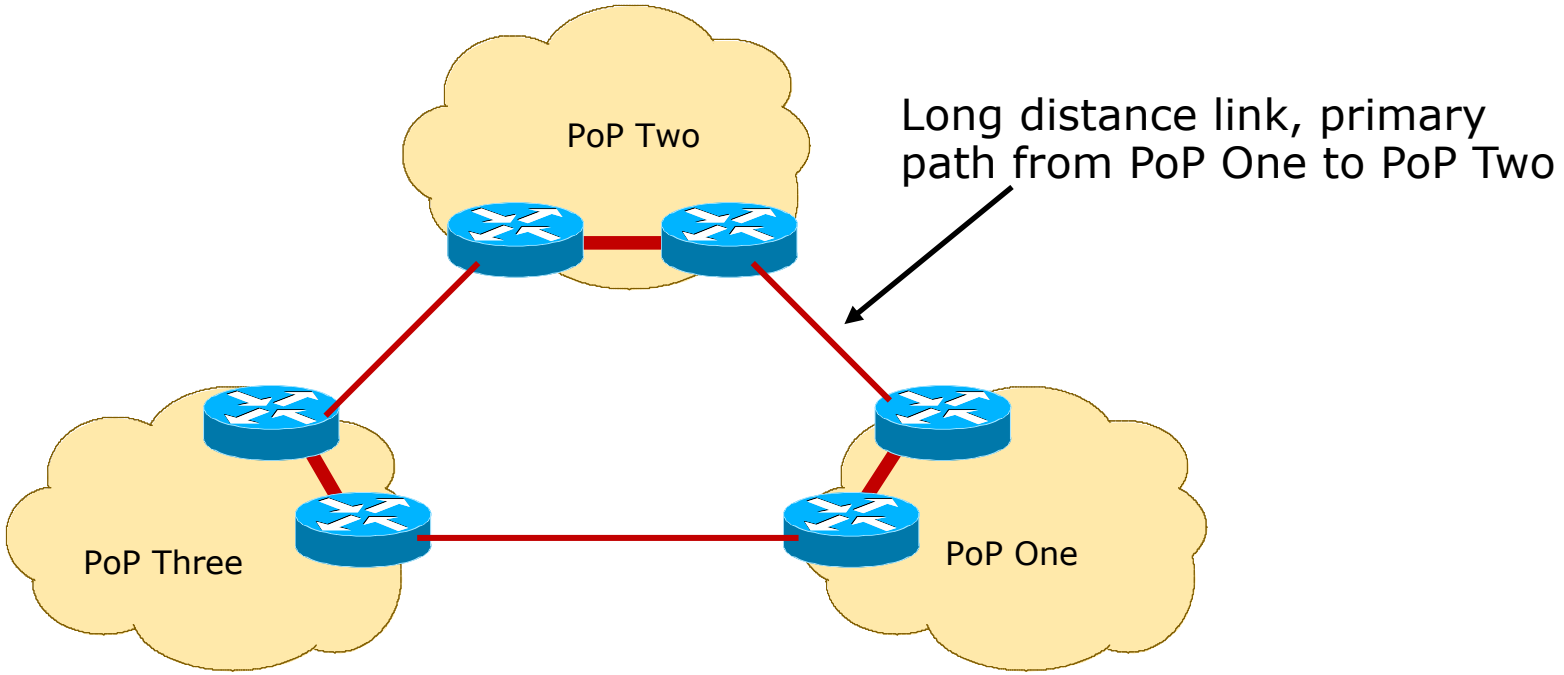
Long Distance Backbone Links

- These usually cost more if no access to Dark Fibre
  - Leasing lambdas
  - Leasing SONET/SDH circuit
- 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

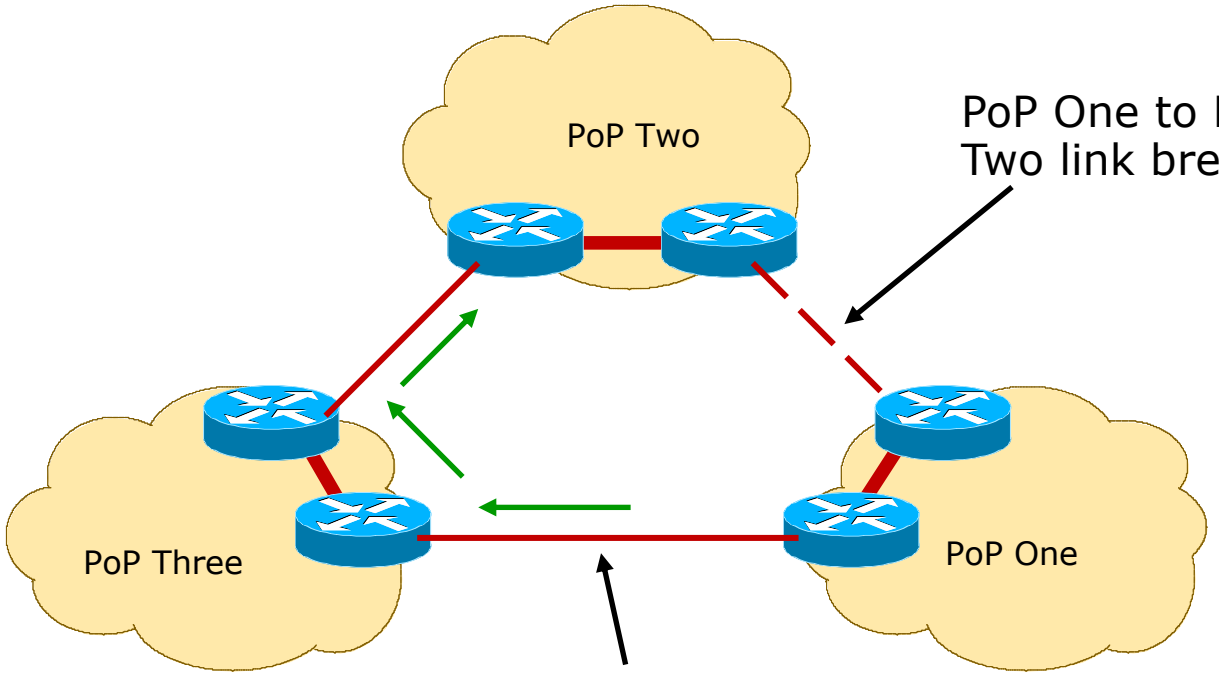
Long Distance Backbone Links

- Allow sufficient capacity on alternative paths for failure situations
- What does sufficient mean?
  - For top quality operators, this is usually at least 50% spare capacity
    - Offers "business continuity" for customers in the case of any link failure
    - Allows for unexpected traffic bursts (popular events, releases etc)
  - Lower cost operators offer 25% spare capacity
    - Leads to congestion during link failures, but still usable network
  - Some businesses choose 0%
    - Very short sighted, meaning they have no spare capacity at all!!

Long Distance Links



Long Distance Links

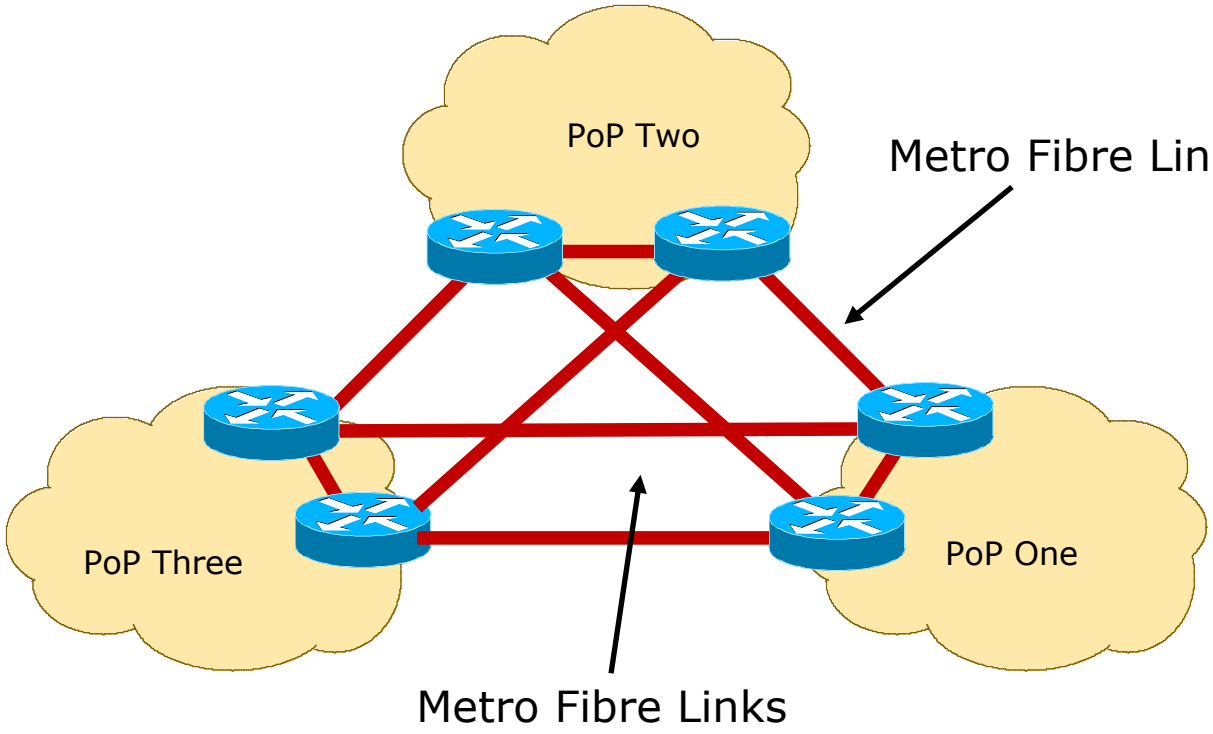


Sufficient capacity to carry traffic between PoP One and PoP Two

Metropolitan Area Backbone Links

- Tend to be cheaper
  - Circuit concentration
  - Choose from multiple suppliers
  - Existing ducts allow easy installation of new fibre
- Think big
  - More redundancy
  - Less impact of upgrades
  - Less impact of failures

Metro Area Backbone Links



Addressing

Today

- New networks are deployed using dual stack
  - The infrastructure supports both IPv6 and the legacy IPv4 addressing
  - The infrastructure runs IPv6 and IPv4 side by side
    - No interaction between IPv4 and IPv6 – independent protocols
- IPv4 address space is almost no longer available
  - Many backbones using private IPv4 address space (RFC1918 or RFC6598) and using NAT to translate to public address space
- IPv6 address space is plentiful
  - IPv6 is supported on almost every networking device available today

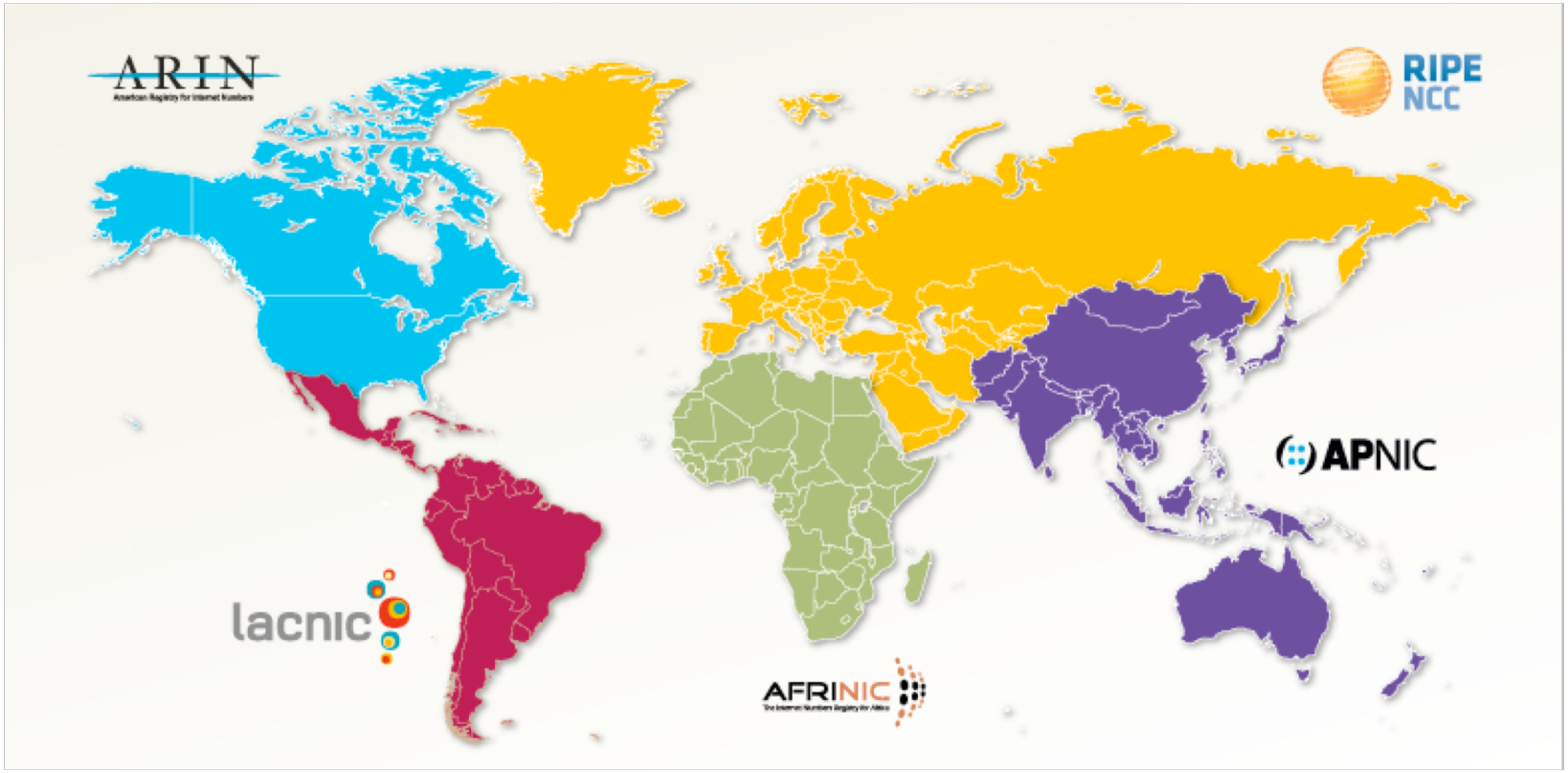
IPv4 & IPv6 dual stack operation

- IPv6 is designed to work independently of IPv4
- If a destination is available only over IPv4, IPv4 will be used
- If a destination is available over IPv4 & IPv6, Happy Eyeballs (RFC8305) ensures that the end-user uses the transport for the best user experience
- Brief summary of Happy Eyeballs for a dual stack device:
  - Application asks for IPv4 and IPv6 addresses
  - If both types are returned within 50ms of each other, application opens connection using IPv6 addresses first, followed by IPv4 addresses
    - Each attempt comes after at least 100ms delay or delay dependent on observed RTT
  - Application uses the transport which responds with a connection first

Where to get IP addresses and AS numbers

- Your upstream ISP
- Africa
  - AfriNIC – http://www.afrinic.net
- Asia and the Pacific
  - APNIC – http://www.apnic.net
- North America
  - ARIN – http://www.arin.net
- Latin America and the Caribbean
  - LACNIC – http://www.lacnic.net
- Europe and Middle East
  - RIPE NCC – http://www.ripe.net/info/ncc

Internet Registry Regions



Getting IP address space (1)

- From your Regional Internet Registry
  - Become a member of your Regional Internet Registry and get your own allocation
    - Membership open to all organisations who are operating a network
  - For IPv6:
    - Minimum allocation is a /32 (or larger if you will have more than 65k /48 assignments)
  - For IPv4:
    - APNIC & RIPE NCC have up to /22 for new members only (to aid with IPv6 deployment)
    - ARIN has nothing
    - AfriNIC and LACNIC have very limited availability – check their websites

Getting IP address space (2)

- From your upstream ISP
- For IPv4:
  - Very unlikely they will give you more than a single IPv4 address to NAT on to
    - This simply does not scale (NAT limitations)
- For IPv6:
  - Receive a /48 from upstream ISP's IPv6 address block
  - Receive more than one /48 if you have more than 65k subnets

Getting IP address space (3)

- If you need to multihome
- For IPv4:
  - Nothing available from upstream provider
  - Address block from RIR (see earlier)
- For IPv6:
  - Apply for a /48 assignment from your RIR
    - Multihoming with the provider's /48 will be operationally challenging
      - Provider policies, filters, etc

What about RFC1918 addressing?

- RFC1918 defines IPv4 addresses reserved for private Internets
  - Not to be used on Internet backbones
  - http://www.ietf.org/rfc/rfc1918.txt
- 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
  - http://www.cymru.com/Documents/bogon-list.html

What about RFC6598 addressing?

- RFC6598 defines shared IPv4 address space
  - Used for operators using Carrier Grade NAT devices
  - http://www.ietf.org/rfc/rfc65988.txt
- Commonly used within service provider backbones
  - NAT used to translate from shared internal to public external addressing
  - Allows the network operator to deploy an IPv4 infrastructure without the fear of address space used between them and their CPE conflicting with RFC1918 address space used by their customers
- Network Operators must filter RFC6598 addressing at their network edge
  - http://www.cymru.com/Documents/bogon-list.html

What about RFC1918 & RFC6598 addressing?

- There is a long list of well known problems:
  - http://www.rfc-editor.org/rfc/rfc6752.txt
- 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

Private versus Globally Routable IPv4 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 Network Operators cannot distinguish between down and broken
- Summary:
  - ALWAYS use globally routable IP addressing for ISP Infrastructure

Why not NAT? (1)

- How to scale NAT performance for large networks?
  - Limiting tcp/udp ports per user harms user experience
- CGN deployment usually requires redesign of SP network
  - Deploy in core, or access edge, or border,…?
- Breaks the end-to-end model of IP
- Breaks end-to-end network security
- Breaks non-NAT friendly applications
  - Or NAT has to be upgraded (if possible)

Why not NAT? (2)

- Limited ports for NAT:
  - Typical user device — 400 sessions
  - TCP/UDP ports per IPv4 address — 130k
  - Implies 130000/400 users — 320 users
  - One IPv4 /22 has: — 1024 addresses
  - One IPv4 /22 could support: — 320k users
- Sizing a NAT device has to be considered quite seriously

Why not NAT? (3)

- Makes fast rerouting and multihoming more difficult
  - Moving IPv4 address pools between CGNs for external traffic engineering
- Address sharing has reputation, reliability and security issues for end-users
- Layered NAT devices (double or even triple NAT)
- Mandates that the network keeps the state of the connections
- Makes the NAT device a target for miscreants due to possible impact on large numbers of users
- Makes content hosting impossible

Why not NAT? (4)

- How to support LTE & LTE-A networks?!
  - Number of users? Public IPv4 addresses for CGN?
  - Maintaining LTE performance? Throughput of CGN?
  - LTE user experience – typically 50Mbps
  - LTE-A user experience – typically 150Mbps

- How to support 5G networks?!
  - 5G promises 1Gbps to the handset with 2ms latency
  - Maintaining LTE performance? Throughput of CGN?

IPv6 Addressing Plans – Infrastructure

- All Network Operators should obtain an IPv6 /32 from their RIR
- Address block for router loop-back interfaces
  - Number all loopbacks out of **one** /64
  - /128 per loopback
- Address block for infrastructure (backbone)
  - /48 allows 65k subnets
  - /48 per region (for the largest multi-national networks)
  - /48 for whole backbone (for the majority of networks)
  - Infrastructure/backbone usually does NOT require regional/geographical addressing
  - Summarise between sites if it makes sense
- Follow a similar strategy for IPv4 address planning

IPv6 Addressing Plans – Infrastructure

- What about LANs?
  - /64 per LAN
- What about Point-to-Point links?
  - Protocol design expectation is that /64 is used
  - /127 now recommended/standardised
    - http://www.rfc-editor.org/rfc/rfc6164.txt
    - (reserve /64 for the link, but address it as a /127)
  - Other options:
    - /126s are being used (mimics IPv4 /30)
    - /112s are being used
      - Leaves final 16 bits free for node IDs
    - Some discussion about /80s, /96s and /120s too
    - Some equipment doesn't support /127s ☹

IPv6 Addressing Plans – Infrastructure

- NOC:
  - ISP NOC is "trusted" network and usually considered part of infrastructure /48
    - Contains management and monitoring systems
    - Hosts the network operations staff
    - take the last /60 (allows enough subnets)
- Critical Services:
  - Network Operator's critical services are part of the "trusted" network and should be considered part of the infrastructure /48
  - For example, Anycast DNS, SMTP, POP3/IMAP, etc
    - Take the second /64
    - (some operators use the first /64 instead)

Addressing Plans – Customer

- Customers are assigned address space according to need
  - IPv6: customer gets a single /48
  - IPv4: usually just a single IP address for them to NAT on to
- Customer address blocks should not be reserved or assigned on a per PoP basis
  - ISP iBGP carries customer nets
  - Aggregation not required and usually not desirable

IPv6 Addressing Plans – End-Site

- RFC6177/BCP157 describes assignment sizes to end-sites
  - Original (obsolete) IPv6 design specification said that end-sites get one /48
  - Operators now must recognise that end-sites need to get enough IPv6 address space (multiples of /64) to address all subnets for the foreseeable future
- In typical deployments today:
  - /64 if end-site will only ever be a LAN (not recommended!!)
  - /56 for small end-sites (e.g. home/office/small business)
  - /48 for large end-sites
- Observations:
  - RFC7084 specifies Basic Requirements for IPv6 Customer Edge Routers
    - Including ability to be able to request at least a /60 by DHCPv6-PD
  - Don't assume that a mobile end-site needs only a /64 – 3GPP Release 10 introduces DHCPv6-PD for tethering
  - Some operators are distributing /60s to their smallest customers!!

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, IS-IS,...
- EGP – Exterior Gateway Protocol
  - Carries customer prefixes and Internet routes
  - Current EGP is BGP version 4
- No interaction 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

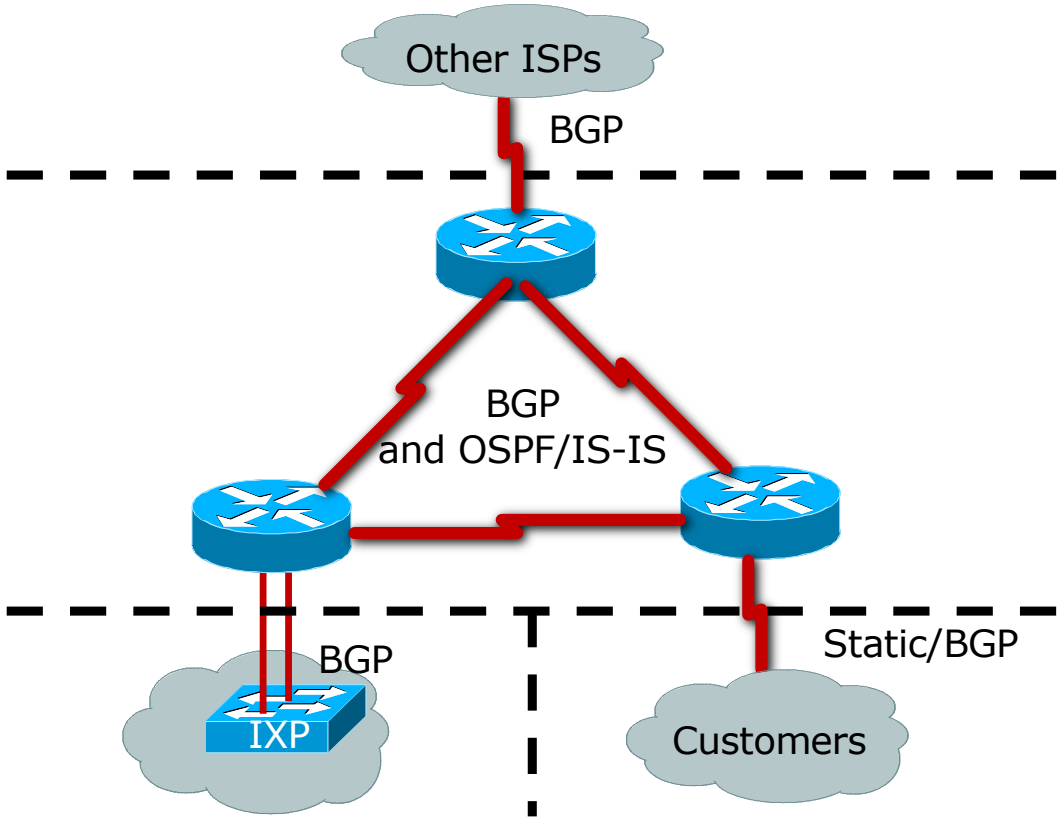
Interior versus Exterior Routing Protocols

- 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



Routing Protocols: Choosing an IGP

- OSPF and IS-IS have very similar properties
  - Review the "IS-IS vs OSPF" presentation
    - http://www.bgp4all.com/dokuwiki/_media/workshops/08-isis-vs-ospf.pdf
- Which to choose?
  - Choose which is appropriate for your operators' experience
  - In most vendor releases, both OSPF and IS-IS have sufficient "nerd knobs" to tweak/optimise the IGP's behaviour
  - OSPF runs on IP
  - IS-IS runs on infrastructure, alongside IP
  - IS-IS 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[v6] unnumbered” on customer point-to-point links – saves carrying that subnet in IGP
    - (If customer point-to-point address 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)

Infrastructure & Routing Security

Infrastructure & Routing Security

- Infrastructure security
- Routing security
- Security is not optional!
- Network Operators 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

Infrastructure Security

- Router & Switch Security
  - Use Secure Shell (SSH) for device access & management
    - Do NOT use Telnet or HTTP
  - 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

Infrastructure Security

- Remote access – JumpHost
  - 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

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

Infrastructure Security

- 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
  - **Don’t block everything: BGP, traceroute and ICMP still need to work!**

Routing System Security

- Implement the recommendations in https://www.manrs.org/
  1. Prevent propagation of incorrect routing information
     - Filter BGP peers, in & out!
  2. Prevent traffic with spoofed source addresses
     - BCP38 – Unicast Reverse Path Forwarding
  3. Facilitate communication between network operators
     - NOC to NOC Communication
  4. Facilitate validation of routing information
     - Route Origin Authorisation using RPKI

BGP Best Practices

- Industry standard is described in RFC8212
  - https://tools.ietf.org/html/rfc8212
  - External BGP (EBGP) Route Propagation Behaviour without Policies
- NB: BGP implemented by some vendors is permissive by default
  - This is contrary to industry standard and RFC8212
- Configuring BGP peering without using filters means:
  - All best paths on the local router are passed to the neighbour
  - All routes announced by the neighbour are received by the local router
  - Can have disastrous consequences (see RFC8212)

Routing System Security

- Protect network borders from "traffic which should not be on the public Internet", for example:
  - LAN protocols (eg netbios)
  - Well known exploit ports (used by worms and viruses)
  - Achieved by packet filters on border routers

- Drop mischievous traffic
  - Arriving and going to private and non-routable address space (IPv4 and IPv6)
  - Denial of Service attacks
  - Achieved by unicast reverse path forwarding and remote trigger blackhole filtering
    - **RTBH** https://tools.ietf.org/html/rfc5635 and https://tools.ietf.org/html/rfc7999
    - **uRPF** https://tools.ietf.org/html/bcp38

Routing System Security – RTBH

- 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 & RFC6598) which should not have routability across the operator's backbone
  - Dealing with DoS attacks on customers and network infrastructure

Routing System Security – RTBH

- 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
```

Routing System Security – RTBH

- Resulting routing table entries:

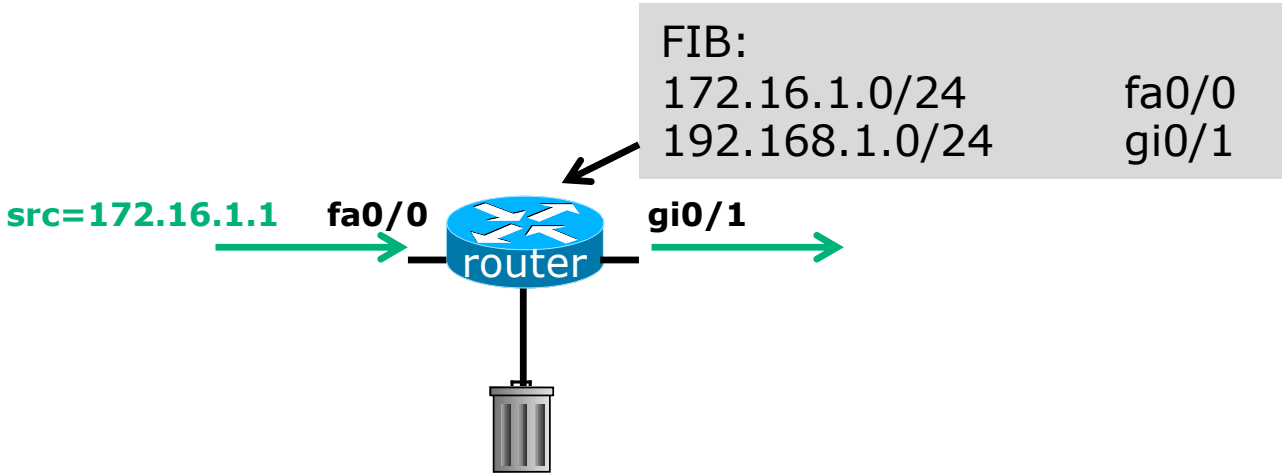
```
gw1#sh ip bgp 10.5.1.3
BGP routing table entry for 10.5.1.3/32, version 64572219
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  Local
    192.0.2.1 from 1.1.10.10 (1.1.10.10)
      Origin IGP, metric 0, localpref 1000, valid, internal, best
      Community: 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
```

Routing System Security – uRPF

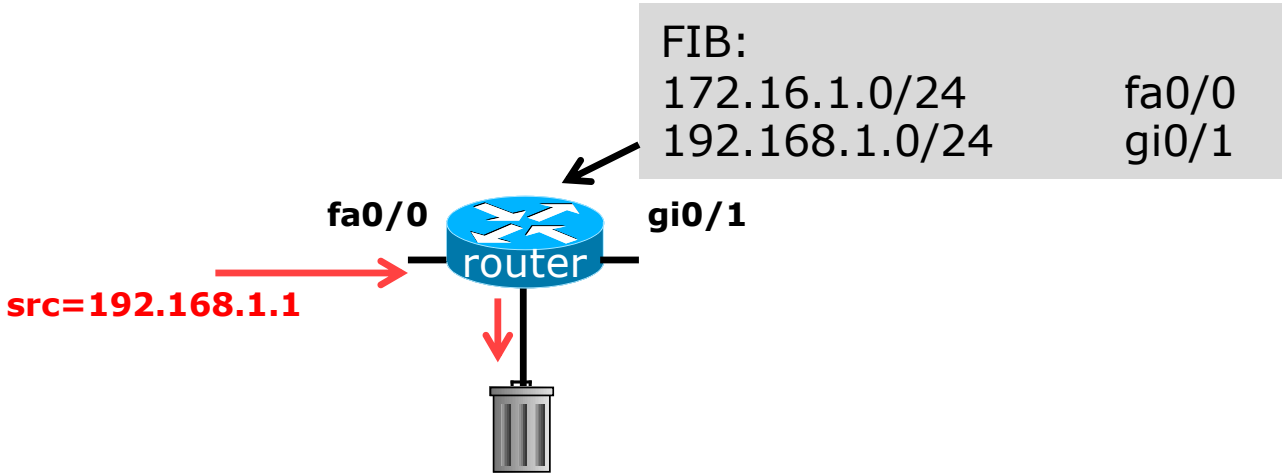
- Unicast Reverse Path Forwarding
- Strongly recommended to be used on all customer facing **static** interfaces
  - BCP 38 (https://tools.ietf.org/html/bcp38)
  - **Blocks all unroutable source addresses the customer may be using**
  - Inexpensive way of filtering customer's connection (when compared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

Aside: 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

Aside: 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

What is RPKI?

- Resource Public Key Infrastructure (RPKI)
  - RFC 6480 – An Infrastructure to Support Secure Internet Routing (Feb 2012)
  - https://tools.ietf.org/html/rfc6480
- A robust security framework for verifying the association between resource holder and their Internet resources
- Created to address the issues in RFC 4593 "Generic Threats to Routing Protocols"
- Helps to secure Internet routing by validating routes
  - Proof that prefix announcements are coming from the legitimate holder of the resource

Benefits of RPKI - Routing

- Prevents route hijacking
  - A prefix originated by an AS without authorisation
  - Reason: malicious intent
- Prevents mis-origination
  - A prefix that is mistakenly originated by an AS which does not own it
  - Also route leakage
  - Reason: configuration mistake / fat finger

Route Origin Authorisation (ROA)

- A digital object that contains a list of address prefixes and one AS number
- It is an authority created by a prefix holder to authorise an AS Number to originate one or more specific route advertisements
- Publish a ROA using your RIR member portal

Router Origin Validation

- Router must support RPKI
- Checks an RP cache / validator
- Validation returns 3 states:
  - Valid = when authorization is found for prefix X
  - Invalid = when authorization is found for prefix X but not from ASN Y
  - Unknown = when no authorization data is found

Using RPKI

- Network operators can make decisions based on RPKI state:
  - Invalid – discard the prefix
    - Several operators are doing this now
  - Not found – let it through (maybe low local preference)
  - Valid – let it through (high local preference)

- Some operators even considering making "not found" a discard event
  - But then Internet IPv4 BGP table would shrink to about 20k prefixes and the IPv6 BGP table would shrink to about 3k prefixes!

RPKI Summary

- All AS operators must consider deploying
- An important step to securing the routing system
  - Origin validation
- Doesn’t secure the path, but that’s the next hurdle to cross
- With origin validation, the opportunities for malicious or accidental mis-origination disappear

Infrastructure & Routing Security Summary

- Implement RTBF
  - Inside Operator 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
- Implement ROAs and use RPKI to validate routing updates
- Use SSH for device management access
- Use TACACS+ for device management authentication

Out of Band Management

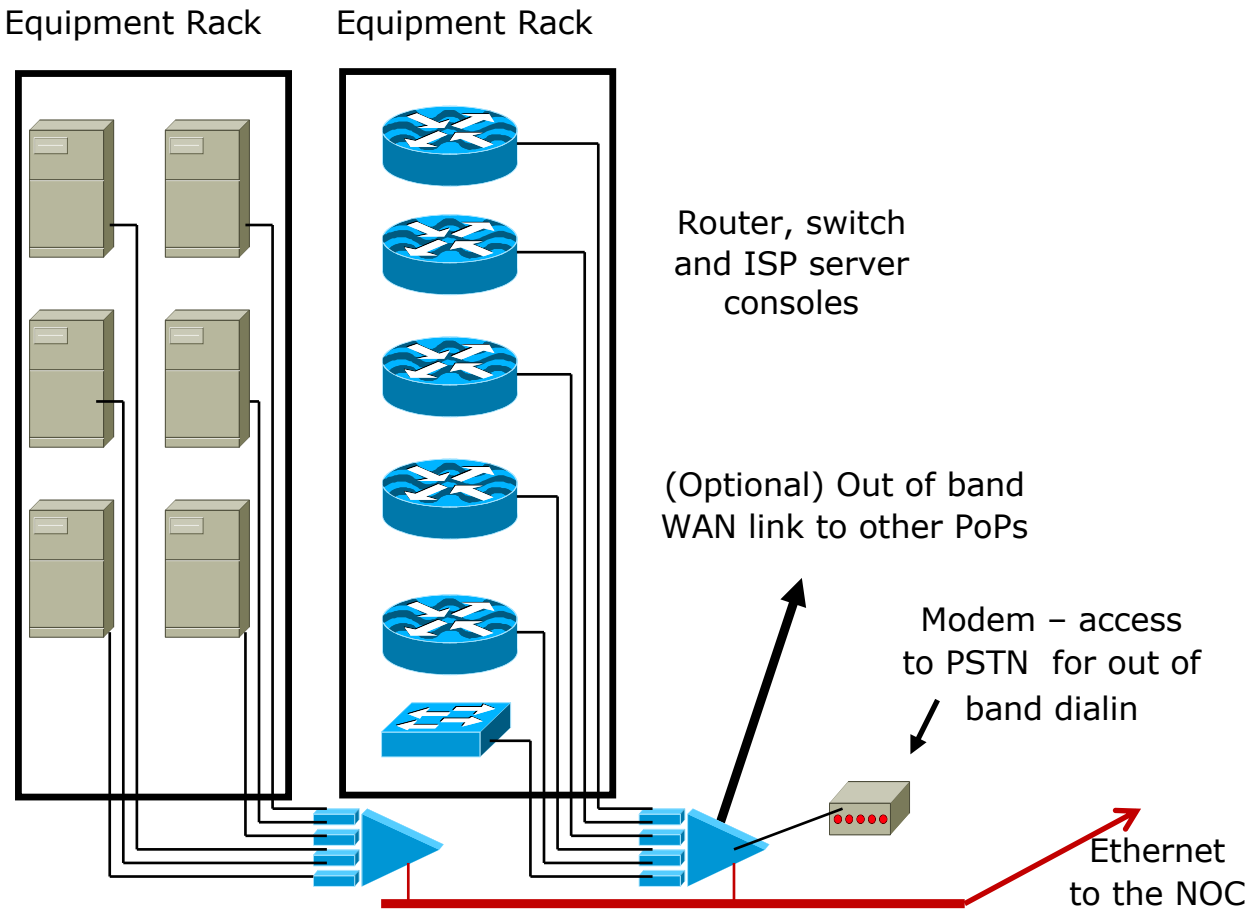
Out of Band Management

- **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



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

Test Laboratory

- 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
- Most major operators around the world have a test lab of some form
  - It's a serious consideration

Operational Considerations

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

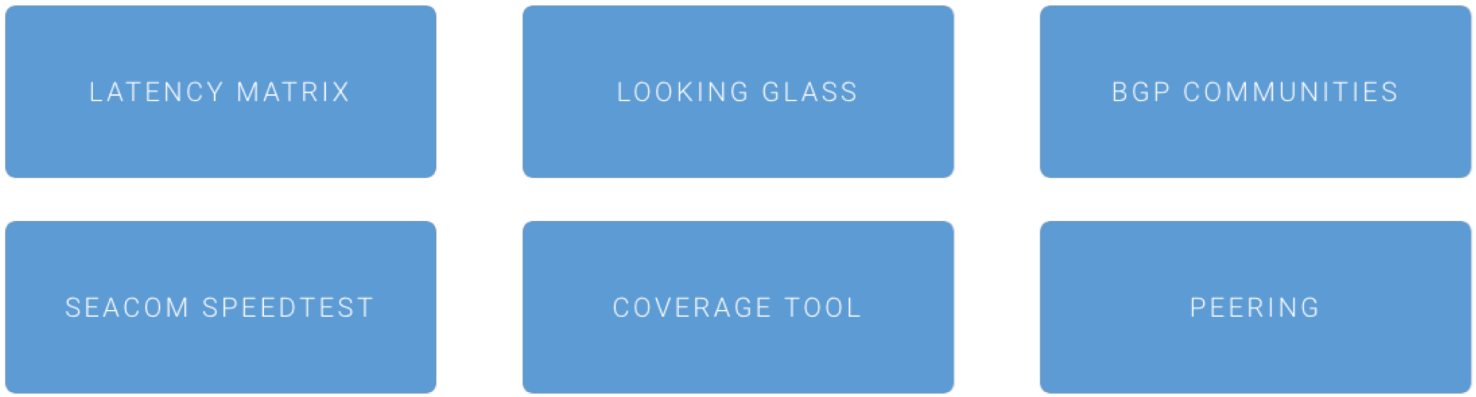
- Never work on the live network, no matter how trivial the modification may seem
  - Establish maintenance periods which your customers are aware of
    - e.g. Tuesday 4-7am, Thursday 4-7am
- Never do maintenance on the last working day before the weekend
  - Unless you want to work all weekend cleaning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

Operational Considerations
Support

- Customer Portals
  - Set up a customer self-help portal
  - For advice on:
    - CPE selection
    - CPE sample configurations
    - Frequently asked questions, frequently provided answers
  - For network status updates:
    - Outages
    - Upgrades
    - Link performance
  - The more information a customer or partner can get, the more confidence they will have in your network infrastructure & operations



Courtesy of SEACOM

Operational Considerations
NOC Communications

- NOC should know contact details for equivalent NOCs in upstream providers and peers
  - This is not "customer support" – this is network operator to network operator
- When connecting to a transit provider:
  - Make sure your NOC staff know how to contact their NOC staff directly
- When setting up a new peer connection (private or IXP):
  - Make sure your NOC staff know how to contact their NOC staff
  - In case of IXP, make sure NOC to NOC contact is well known too

ISP Network Design

Summary

ISP Design Summary

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

ISP Network Design

ISP Workshops