L2 Internet eXchange Point (IXP) using a BGP Route Reflector

Technical Design, Configuration, and General Advice about IXPs

Barry Raveendran Greene
bgreene@cisco.com

Draft – Version 0.4

Apologies if parts of this paper is rough or incomplete. This whitepaper is a draft of a living document that will frequently updated. Questions, comments, and suggestions from ISPs are welcome. This sort of private dialog is what drives the materials in this whitepaper. Please post them directly to the author. Updated versions of the document can be found in http://www.cisco.com/public/cons/isp/ixp/.

Preface

This whitepaper is targeted for ISPs outside the United States. The context of the information, recommendations, and configurations are tooled specifically for ISPs who have the dual function of providing international connectivity to the entire Internet and domestic connectivity to their country. While suggests and techniques can be applied to IXP operations any where, mindfulness is required to view the information presented here in it's proper context.

INTRODUCTION

Internet eXchange Points and Peering

Internet eXchange Points (IXPs) are the most critical part of the Internet's Infrastructure. It is the meeting point where ISPs interconnect with one another. With out IXPs, there would be no Internet.[1] Interconnecting with other networks is the essence of the Internet. ISPs must interconnection with other networks to provide Internet services. Yet, for the majority of countries in the world, there are no local interconnections between ISPs. This whitepaper will focus on the scalable Layer 2 BGP Route Reflector based IXP. This architecture has proven to provide a low cost solution with a clear scaling path for future growth. The whitepaper will also cover the rolls IXPs play in the Internet, reasons why ISPs must interconnect, types of IXPs, design techniques, peering techniques, IXP services, and some IXP case studies.

[1] Private and Bi-Lateral Peering are considered to be a type of IXP.

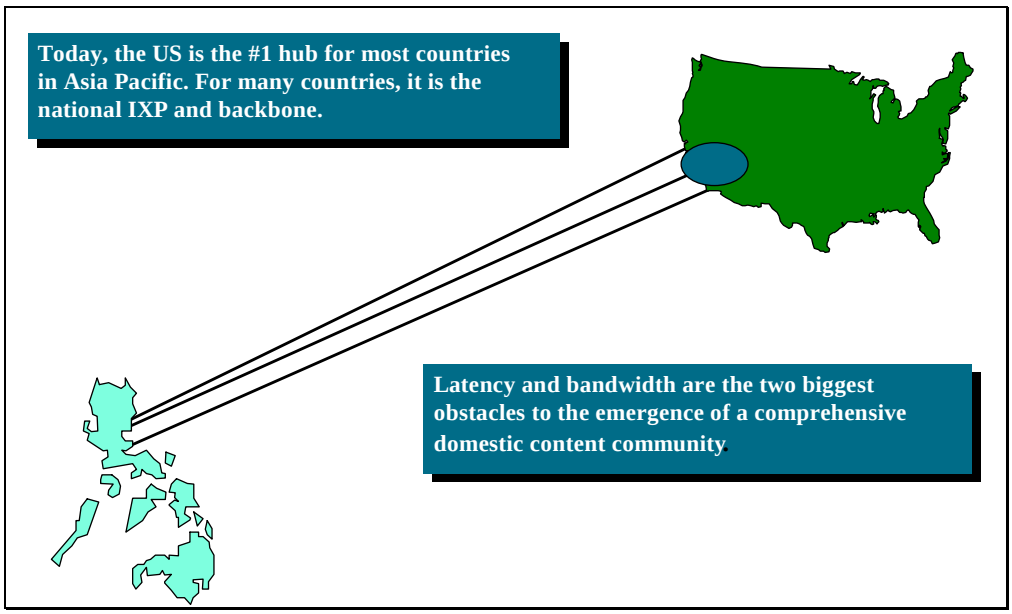

Figure 1 - Scope of the problem - too many countries using the US as their IXP

Why IXPs are so critical to ISP's Survival and Business Success?

IXPs can be considered the ***keystones*** of the Internet. Keystones are the critical stone in an arch that holds the entire arch in place. It is held in place by the collective pressure of the other stones in the arch. These stones depend on the keystone to hold everyone together with a specific shape and function. With out the keystone, the entire arch will collapse. IXPs play the same roll as the arch's keystone. IXPs hold together a collection of ISPs with a locality. The IXP is the glue that allows local traffic to stay local. It provides structure to the local ISP community, allowing local E-Communities, E-Commerce, and the new Internet Economy to thrive.

With out the IXP, the local ISP community will interconnect in another country – placing a major obstacle to the development of local E-Communities and E-Commerce. International transit traffic is mixed with domestic traffic. The result for the ISPs is a business case built on international connectivity to the Internet with little revenues based on interaction in the local Internet community. Customers who are more interested in communicating with people and organizations within their own country would find the experience frustrating and discouraging. With an IXP, customers looking for the local exchange of information would find the experience fast and efficient. The result is that more customers us the domestic services hosted by all the ISPs - increasing the domestic bandwidth demand – prompting a cycle where there is more bandwidth dedicated to the local interconnection then to the international links. Since domestic bandwidth is always cheaper then international bandwidth, new domestic oriented business cases emerge. These new domestic models for Internet traffic flows are often the key factor between a profitable ISP and an ISP in danger of failing.

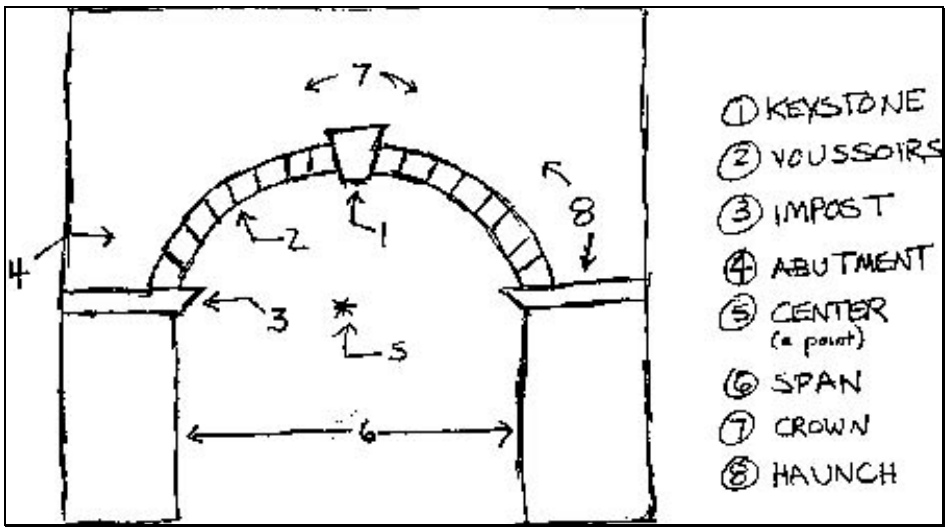

Figure 2 - Diagram of an Arch with the Keystone – no Keystone – no Arch[2]

The Force that Drives ISPs to Interconnect

Many people - including the ISP - perceive the key benefit for an ISP to connect to an IXP is the bandwidth savings. True, there are bandwidth savings. Bandwidth that would otherwise traverse their upstream links would now traverse their domestic link to the IXP. The local links and bandwidth savings improve and enhance service quality for the customers. Yet, *bandwidth savings* and *service quality* are not the most critical benefit ISP derive from the IXP connection. The most critical benefit to the ISP is the ***new revenue opportunities***.

Before the IXP, the ISP business case was dominated by their international link to the rest of the Internet. Domestic service (two local companies linked to each other over an Internet Extranet) and international services (access to the entire global Internet) were both depended on the upstream connection to the Internet. The IXP changes that linkage. Through the IXP connection the ISP can create new business models that drive the domestic link. E-Banking, E-Commerce, E-Government, VPNs, Content Hosting, and many other services are severely hampered by the high latency/high congestion on the upstream links. So with no local interconnection, these services are extremely difficult to work effectively. With the IXP, a new range of services can be offered by the ISPs and Internet businesses that connect to the ISPs. In essence, the IXP facilitate new business opportunities that were not feasible before.

So the real benefit for ISP to connection to the IXP is not the bandwidth savings. The force that drives ISPs to interconnect is the potential for new revenues. For example, think about national Quake Tournaments run by all the in a city Internet Cafes - with the entire volume of Quake traffic going over the IXP. Low latency, un-congestion bandwidth, and a fast computer are key to winning on-line games. If this on-line game thrives, there is then a revenue opportunity for an ISP to offer on-line gaming services. The interesting part of a on-line gaming service is that the potential customer base consist all of the internet consumers connected to all the ISPs connected to the IXP. In other words, you ISP competitor's customers are you potential on-line gaming service customers. Yet, with out the IXP, this new revenue stream of on-line gaming would not be feasible. It is the connection to the IXP that makes it possible to view all the Internet consumers on all the ISPs as one large potential market.

[2] Found on the Net at http://www.cmhpf.org/kids/dictionary/arch.html

Cooperation with the Competition – the Factor that makes the Internet the Internet

Co-Locating Transit on a IXP

From the early days of IXPs on the Internet, selling transit over an IXP has been a taboo. IXPs are supposed to be *neutral* places for ISPs and NSPs to come and exchange traffic as ***peers***[3]. Selling transit has been considered to be a customer provider relationship, not a peer relationship. Hence, customer provider transit connections at IXPs were either discouraged through peer pressure or not contractually allowed in the terms and conditions of the IXP. So, most transit connections to the Internet were done with a separate lease line connection to a Network Service Provider (NSP)[4] (see Figure 3).

[3] Peering on a IXP usually means that two ISPs will exchange traffic their respective autonomous system numbers (ASN).
[4] For the purposes of this document, we are define Network Service Providers (NSPs) as those Internet Service Providers (ISPs) who are *default free*. Meaning that they do not have an upstream ISP where they send their traffic if it is not in their routing tables.

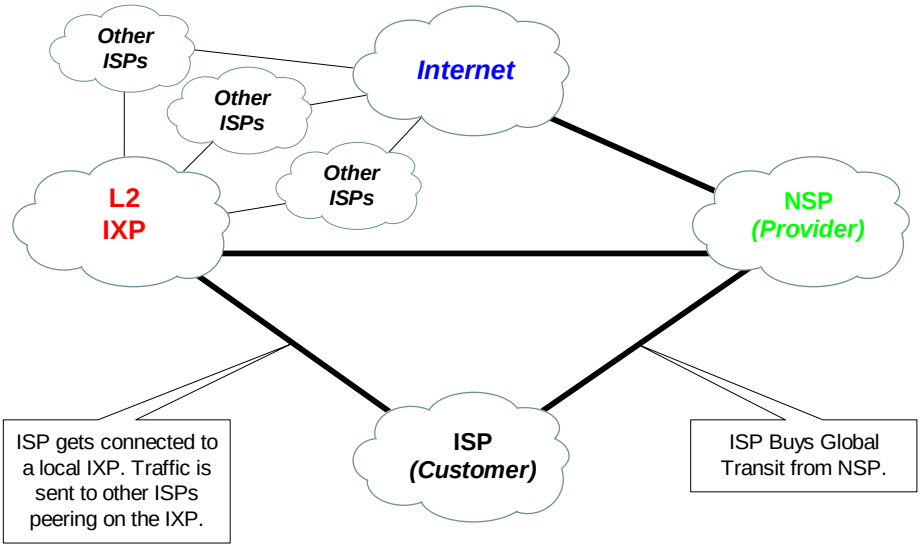

Figure 3 - Typical Way an ISP Connects to the Internet

Yet, the Internet is built upon the themes of open communication and interconnectivity. So, if someone really wants to interconnect, ways are found to do it. Bypassing barriers of communication is a core cultural value of the Internet. Over time, ways were found to sell transit on some IXPs around the world. Three scenarios for selling transit on an IXP developed:

- Private Interconnects at L2 IXPs
- Regional IXPs Hubs and National Gateways
- Co-Locating Transit Relationships on an IXP

Private Interconnects at L2 IXPs

Some providers at L2 IXPs interconnected with each other privately. That is, while they were interconnected at the L2 IXP's interconnect medium[5], they created a second *private* connection between each other. This is done via back-to-back serial connections, ethernet, FDDI, or higher interconnect speeds[6]. At first, these *private interconnects* were used to off load traffic from the L2 IXP medium. Two ISPs that would have the majority of traffic on the IXP would create a private interconnect to off load this traffic through a private connection. This helped to keep the IXP healthy, minimizing traffic overload on the L2 IXP medium.

[5] *Interconnect medium* on an IXP is the ethernet, FDDI, ATM, SONET, or other technology used to interconnect router/switches the belonging to the members of the IXP.
[6] Just recently back-to-back PPP over Sonet/SDH (POSIP) interconnects were installed between ISP on a L2 IXP. This provides the two ISPs near optical line speeds in the amount of traffic over the link - minimizing the effects of *cell tax* felt on ATM.

In time, ISPs who were connected to the IXP sought to use private interconnects as a way to buy transit from another ISP on the L2 IXP.

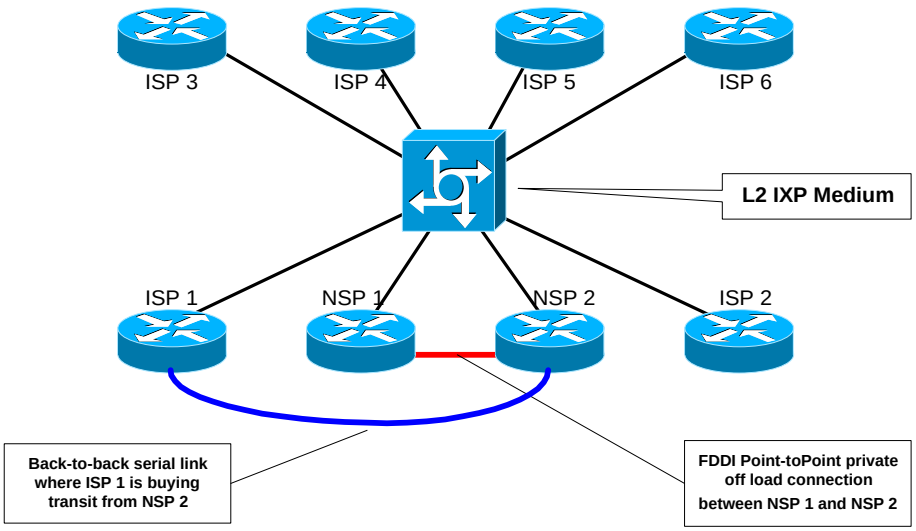

Figure 4 - L2 IXP with Private Off-Load and Transit Connections

One question begs to be asked - *why not just point default across the L2 IXP medium at the ISP you are buying transit from*? There are several reasons why ISP do not do this even through, technically, it is totally possible. First, most L2 IXPs includes clauses in the terms and conditions of the IXP that forbid transit across the IXP medium. This places a contractual bind preventing a customer provider from establishing itself over the IXP fabric. The objective is to maintain the neutrality of the IXP as a place of *Internet peering*.

Second, *peer pressure* and *doing the right thing* have an inhibiting effect for any customer provider relationships from developing on L2 IXPs. Most L2 IXPs monitor traffic on the IXP medium. This data collection and analysis is usually open to the members of the IXP for traffic engineer improvements. Customer provider traffic flows are easy to spot with this data, allowing the IXP membership at large to apply pressure to have the customer provider relationship pulled off the IXP medium through another connections avenue.

Finally, even through it is technically very easy to establish a customer provider relationship over the IXP medium (point default from the customer and a static route from the provider), there is no guarantee that the customer will be able to receive any specified throughput over the IXP medium. The IXP exchange medium is not controlled by the customer nor the provider. Hence neither can trust that the bandwidth will be available when they really need it.

For example, if the customer was looking for a 2 Mbps symmetrical backup link from a provider at an IXP, the provider would not be able to guarantee 2 Mbps over the IXP medium. Since the IXP medium is under the control of the L2 IXP operator and is shared by all the members of the L2 IXP, predictable guarantees would not be possible - even with the new waive of QoS features being deployed in Cisco's IOS.

Regional IXPs, Hubs, and National Gateways

One of the key shifts of selling transit over an IXP started in 1995. This is when several Telco/ISPs start L3 IXPs with the additional feature of transit to the rest of the Internet[7]. While the goal of the original IXPs was to interconnect ISPs with each other, these L3 IXPs were primarily interested in selling global/regional Internet transit, International Lease Circuits (ILC), and VSAT services to ISPs within their region (see Figure 5). They worked because the cost of connecting to a to a ISP with in the region was cheaper than connecting directly to the US.

From some, L3 IXPs that sell transit are not considered to be real *Internet eXchange Points*. Yet, the fact that ISPs who connect to these L3 IXPs can get access to other ISPs within the region without going all the way to the US, would give some weight of their impact on the Global Internet.

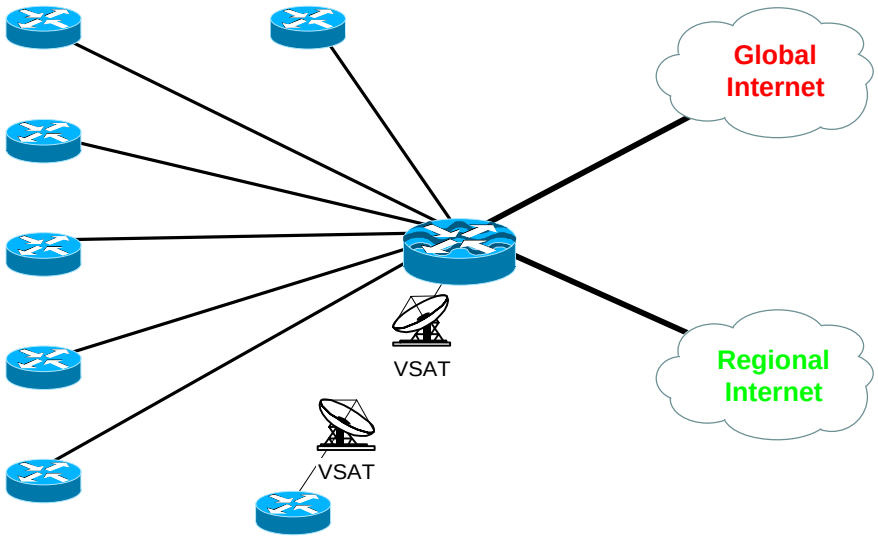

Figure 5 - L3 IXP Selling Transit Services

Are L3 IXPs (National Gateways) Really an *Internet eXchange Point*?

No, even through Commercial L3 *exchange/gateway* services do over local interconnection, there are not really considered a true IXP. There is some controversy on whether a router can be a viable IXP architecture. True, in the past routers have been used to establish a Layer 3 (L3) IXP. The Commercial Internet eXchange (CIX) was one of the first examples of an IXP that used a router as an IXP. Regional Transit services has also used the L3 IXP model for their service. STIX, SingTel IX, HKT NetPlus, AIH, and Telstra's Big Pond are all examples of this type of L3 IXP architecture. This key issues is that the regional L3 *Transit Services,* while offering regional connectivity, do not scale when used as a nation's primary IXP architecture.

[7] Three major L3 IXPs which sold transit included Asia Internet Holding (AIH) together with IIJ, Singapore Telecom Internet eXchange (STIX), and Hong Kong Telecom's NetPlus service.

The two key factors that shape this perception: the limitation of an ISP's autonomy and inability to separate the internationals and domestic business cases.

Why an ISP's Autonomy is Critical for an IXP to Succeed

When an ISP connects to an IXP, they are looking for a way to benefit from the interconnection while maintaining their autonomy. An ISP want to control who they interconnection, how they interconnect, apply policy controls, apply security controls, and be able make changes at any time of the day. While the ISPs have control over their router, they do not control the L3 IXP Router. So any change that need to be done much be coordinated with the person maintaining the L3 IXP Router. So if one ISP wants to change their peering policy on the IXP, they must adjust their filters, then get the person maintaining the L3 IXP Router to change their filters. This requirement to have two parties to synchronize to commit a policy change is perceived to be a major limitation of an ISP's autonomy.

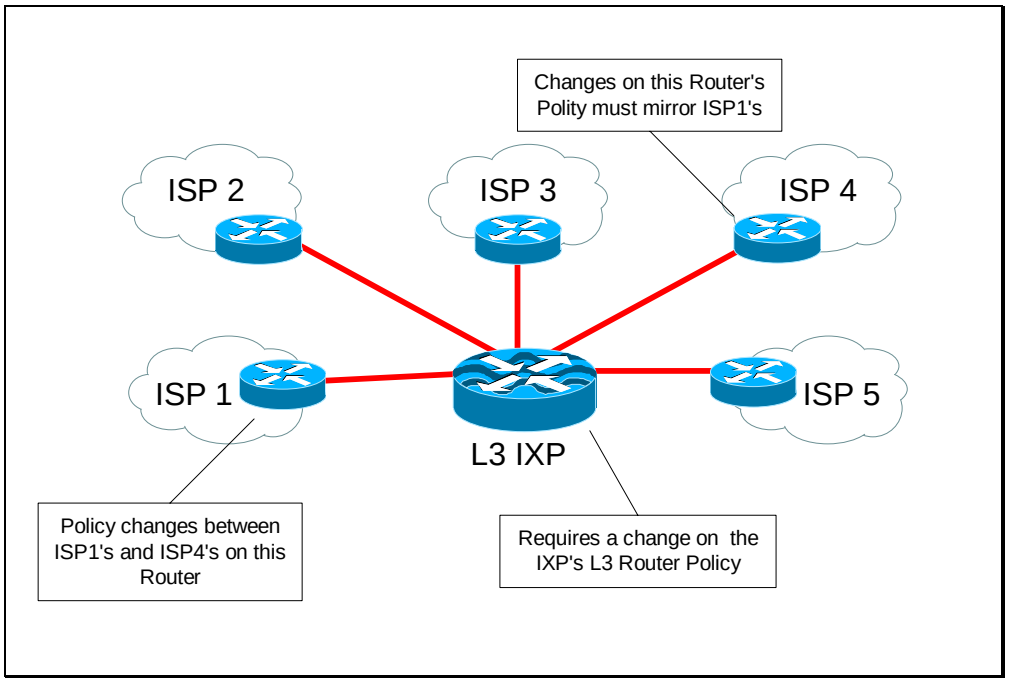

Figure 6 - L3 IXPs Limit an ISP's Autonomy

Why is the Separation of the International and Domestic Bandwidth Critical?

Many people - including the many ISP - think the key benefit for an ISP to connect to an IXP is the bandwidth savings. Indeed, any domestic traffic that can traverse a local link frees up capacity for more international traffic on over the upstream link. But this *bandwidth saved* is not the ISP's key benefit. The most critical benefit to the ISP is the new **revenue opportunities**. Before the IXP, the ISP business case was dominated by their international link to the rest of the Internet. After the IXP, the ISP can create new business models over the domestic link to the IXP. So the real benefit for ISP to connection to the IXP is not the bandwidth savings, it the potential for new revenues.

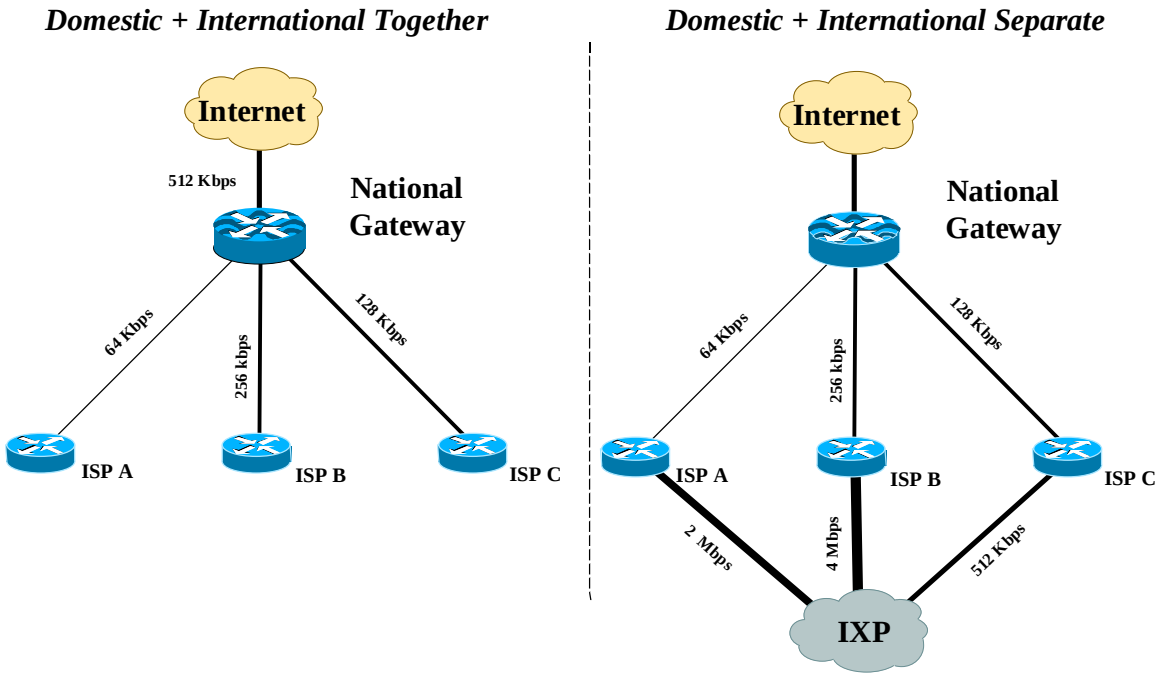

Figure 7 - Growing the Domestic Bandwidth independently from the International Bandwidth

There is a perception propagated by several *development* agencies that *National Internet Gateways* are IXPs. They are not. National Internet Gateways are Network Service Providers (NSP) – and ISP that provides Internet connectivity to the rest of the Internet. An NSP's customers are other ISPs. They are not IXPs. If an ISP wants to gain revenue benefits from separate internationals and domestic traffic flows, then these two traffic flows must be separate. National Internet Gateways combine these two traffic flows, making it more difficult grow the two revenue streams separately. For example, an ISP with 128Kbps of international bandwidth could easily have 4 Mbps of domestic bandwidth - each of them generating revenue. If the domestic services increase, then the ISP should be able to upgrade the domestic bandwidth separate from the international bandwidth. Hence, National Internet Gateways, while excellent for internationals services, are not optimal for domestic interconnections. National Internet Gateways with one or more separate IXPs (one in each major city) have proven to be the optimal mode for domestic interconnections."

Co-Locating Transit Relationships on an IXP

As mentioned previously, transit on an IXP is very controversial topic. Most of the ISP Engineers who have experience interconnecting to IXPs would strongly discourage any sort of talks of a transit relationships happening over an IXP. Yet, given the core design of many IXPs located outside of the US, there are some options available to allow some transit relationships. The crux of the idea is to have any Network Service Provider who wishes to sell transit to co-locate a router at the IXP. If a smaller ISP wants to connect to the IXP ***and*** buy transit, they would either connect to the NSP's router for a L3 connection to the IXP or co-locate their own router and add a back to back connection to the NSP's router.

For example, an IXP has two NSP customers who wish to sell transit to small ISP customers connecting to the IXP (see Figure 8). Each NSP would co-locate a router at the IXP and peer with other ISPs and NSPs. ISP who wish to buy transit from NSP1 would connect directly to NSP1's router. ISP who wish to buy transit from NSP2's would connect directly to NSP2's router. ISP who just want to peer would connect to the L3 router.

Figure 8 shows how ISPs 1 & 2 connect directly to R3 - NSP1's router. ISPs 3 & 4 only wish to peer with ISPs and NSPs on the IXP so they connect directly to R1. ISPs 5 & 6 wish to buy transit from NSP2, so they connect to R2 - NSP2's router.

ISP7 is a special case. They wish to buy transit from both NSP 1 & 2. Hence, they co-locate their own router at the IXP. This router can be a small router with three serial ports (i.e. a Cisco 3620 in this case). ISP7;s lease line terminates in the their router - R4. R4 has back-to-back DTE-DCE connections to router R2 and R3 (the routers for NSP 1 & 2). The back-to-back serial connections are *clocked* from the routers R2 and R3 - determining the speed of the transit connection to ISP7. ISPs and NSPs on the IXP would send traffic to either R2 or R3 - depending on shortest path rules through BGP4.

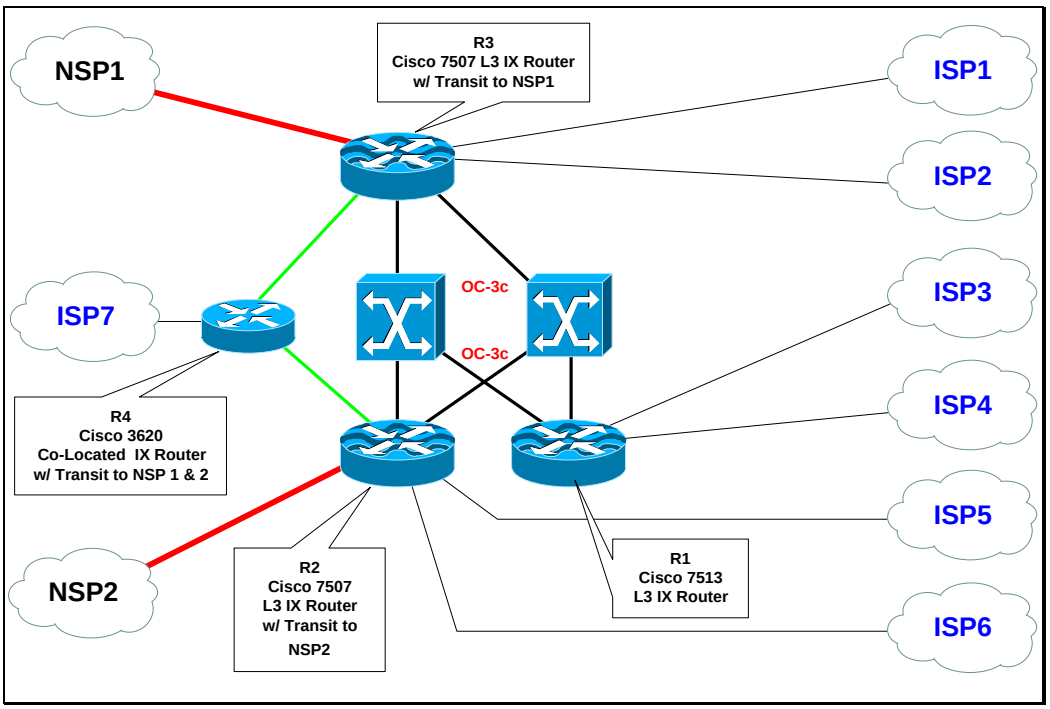

Figure 8 - L3 IXP Transit Options

Keeping transit traffic off the exchange medium simplifies things on IXP while offering ISPs a wider choice. For instance, let say ISP1 is not happy with NSP1's transit service. ISP1 can use the same lease line into the IXP, unplug from NSP1's router and connect to NSP2's transit service instead. ISP3 could decide that connecting to the IXP via the L3 service is too limiting. Hence, they co-locate to the L2 IXP medium with their own router.

Services Offered by an IXP

Maintenance - Contractual Level 1 and Level 2 maintenance contracts for members of an IXP.
IXP's Web Pages - What information should be included.
Route Server and Router Reflector. What do they do? How do they work? Which one is best fit?
WWW Caching and Replication Services.
Multicast Server
Content Co-location
Traffic Analysis Tools
Looking Glass
SNMP, Netflow, and others tools
What data is confidential, and what is public?

Services and Business Strategies ISPs should consider gain the most benefit from the IXP.

IXP Businesses – is there an IXP Market?

There is not really an *IXP Business* in the traditional sense of making large profit margins on IXP services. Mainly cause there is no "market" for IXP. IXPs are created as a tool for the ISPs to peer with each other. The ISPs will find the most cost efficient mode of interconnecting. Usually that means the collectively find a way to create an non profit IXP through an ISP Association.

Commercial operators who run IXP do so for two reasons. First, companies will run an IXP as a service to the industry (i.e. like the MAEs). These companies price the service so they do not loose any money. That way they gain some PR value from providing the service to the industry. Core revenues are gained from the infrastructure sold into the IXP (i.e. lease lines). The second reason is that companies off an IXP service to add value to a co-location business. These providers - like AboveNet and Eron - offer the IXP service in their facilities at cost. Their gain is that they can use the IXP as an enticement for companies to move their equipment into their co-locations facilities. For example, AboveNet worked really hard to get the top two ISP from every major country in Asia co-located and peering with each other in AboveNet. Now they can specialize in the "Asian niche." If you want to reach all of Asia, then co-location at AboveNet is the place to be.

So we you, there is no real "IXP business," hence no "market." If companies tried to make real profit from an IXP business, the ISPs would not buy it. The services would have to be priced too high to make a significant margin. ISPs would look for others ways to peer with each other.

General Services that are extremely limited by no local interconnection (E-Commerce, E-Banking, VPNs, etc.)

The entire local Internet Industry benefits from the opening of new services that can be deployed over an IXP. E-Banking, E-Commerce, E-Government, VPNs, Content Hosting, and many other services are not possible when there is no local interconnection between the ISPs. With the IXP, a new range of services can be offered by the ISPs and Internet businesses that connect to the ISPs. In essence, the IXP facilitate new business opportunities that were not feasible before.

Two Speed Local Lease Line Access

Good Traffic vs Bad Traffic

Fostering the creation of a local content production industry

TECHNICAL OVERVIEW

The L2 BGP Route Reflector (RR) Internet eXchange Point (IXP) is an IXP architecture pioneered through the work at the Hong Kong Internet eXchange (HKIX)[8]. The L2 RR IXP uses routers as dedicated "BGP Route Reflectors" to minimize the number of peering sessions a member router have to configure. Since the number of BGP peering sessions between ISPs are less, smaller routers can be used on the L2 RR IXP, reducing the cost of entry for the IXP. At the same time, this architecture allows for normal point to point BGP peering between ISPs or a Route Server. The result is a low cost of entry for ISPs connecting to the IXP while the IXP enjoys a proven scaling path – allow for growth as more ISP connect and peer to the IXP.

Some other advantages the L2 RR IXP offers countries that are in the early stages of IXP development include:

Lower Entry Cost for IXP Participants. There are two major factors that effect the selection of a router for an IXP. One is the Packets Per Second (PPS) performance of the router. The second is the CPU and memory needed to process the BGP routing updates. Since the L2 RR IXP has at a maximum two BGP Peer connections per ISP router[9], the routers needed will not require high end CPU and lots of memory. The HKIX experience demonstrates that L2 RR IXPs can successful scale with Cisco 2501 routers for each IXP member. This reduces the cost of entry for IXP members – allowing for more ISPs to join the IXP.

Simplicity of Design. The L2 RR IXP offers a simple design while providing proven reliability.

Scales to +60 IXP Members. HKIX has prove the scalability if this design with over +60 ISPs members with the vast majority of them using the BGP Route Reflectors.

Proven Upgrade Path. A L2 RR based IXP can *upgrade* to either a fully meshed L2 IXP or Router Server based L2 IXP. As the IXP grows, the L2 RR IXP can add mesh of BGP connections, a BGP Router Server, or a hybrid of all of the L2 options. This minimizes future redesign.

Mandatory Multi-Lateral Peering Agreement (MPLA). Bi-Lateral Peering agreements are difficult to implement on a L2 RR IXP. Hence, a multilateral agreement is required. For new IXPs, this is a benefit – eliminating on of the contentious issue with ISP interconnections on IXPs.

Layer 2 IXP using the BGP Router Reflector – Basic Theory

This flavor of L2 IXPs relies on the BGP Route Reflector technology to enable peering between ISPs. BGP Route Reflectors were originally designed to allow iBGP meshes to scale by added hierarchy. RFC 1966 ***BGP Route Reflection An alternative to full mesh IBGP***. by T. Bates & R. Chandrasekeran. (June 1996) provides the details specification of how BGP Route Reflectors work. Addendum 1 provides additional details.

[8] HKIX's URL is www.hkix.net
[9] One will be for the primary route reflector, the second will be for the back-up route reflector.

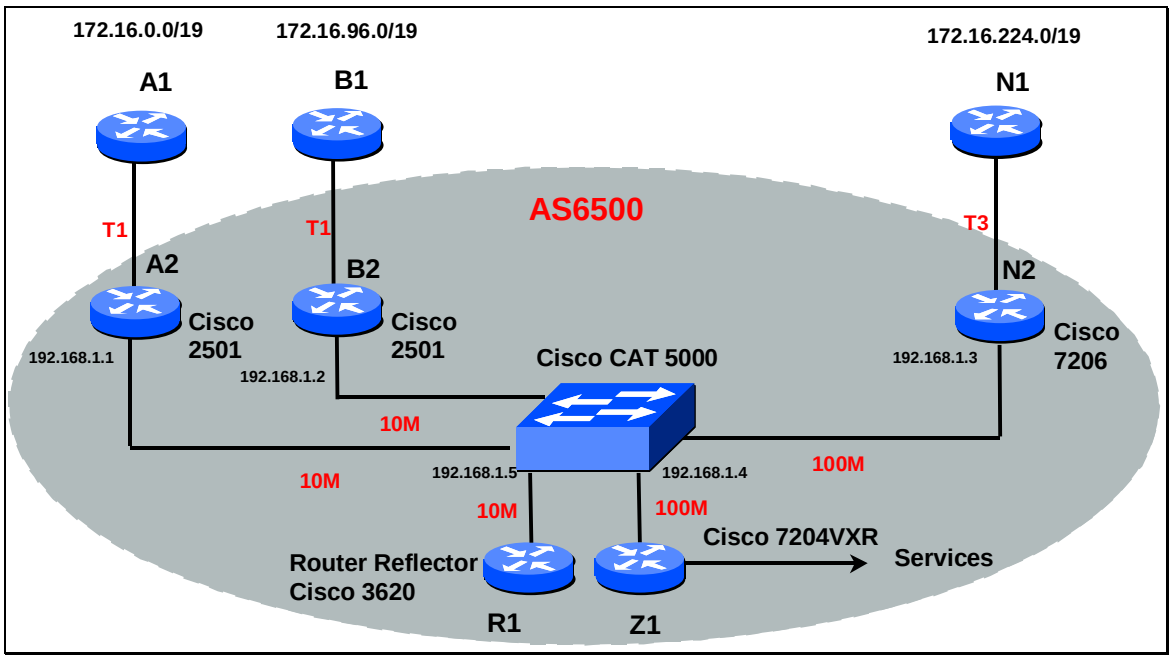

Figure 9 - Basic L2 Route Reflector IXP

In a basic L2 RR IXP, the IXP provides the interconnection medium (usually switched ethernet), the router reflector router, and another router to the IXP services. The ISP members provide their own routers and connections to the IXP. The IXP management is responsible for the configuration and operations of the interconnection medium, the router reflector, and services on the IXP. The ISPs are responsible for the configuration of their routers. Figure 9 provides a basic template of how a L2 RR IXP is interconnected.

NOTE: The IP addresses in this example are used for illustration only. Some are valid IP addresses and should not be used on production systems.

The L2 RR IXP needs it's own BGP Autonomous System number. Since it is strongly discouraged to not have any transit services on the IXP, private AS numbers may be used on the IXP (see the section on *Co-Locating Transit Services on an IXP*). The private AS number 65000 is used in this example. Three ISPs – ISP A, ISP B, and ISP N – shown used in this example. Each of these ISPs has an allocated block of IP addresses from their Regional IP Registry for IPv4 addresses[10].

[10] Active Regional Internet Registries (RIRs) include APNIC (http://www.apnic.net), ARIN (http://www.arin.net), and RIPE-NCC (http://www.ripe.net). RIRs working towards certification include AFRINIC (http://www.afrinic.org).

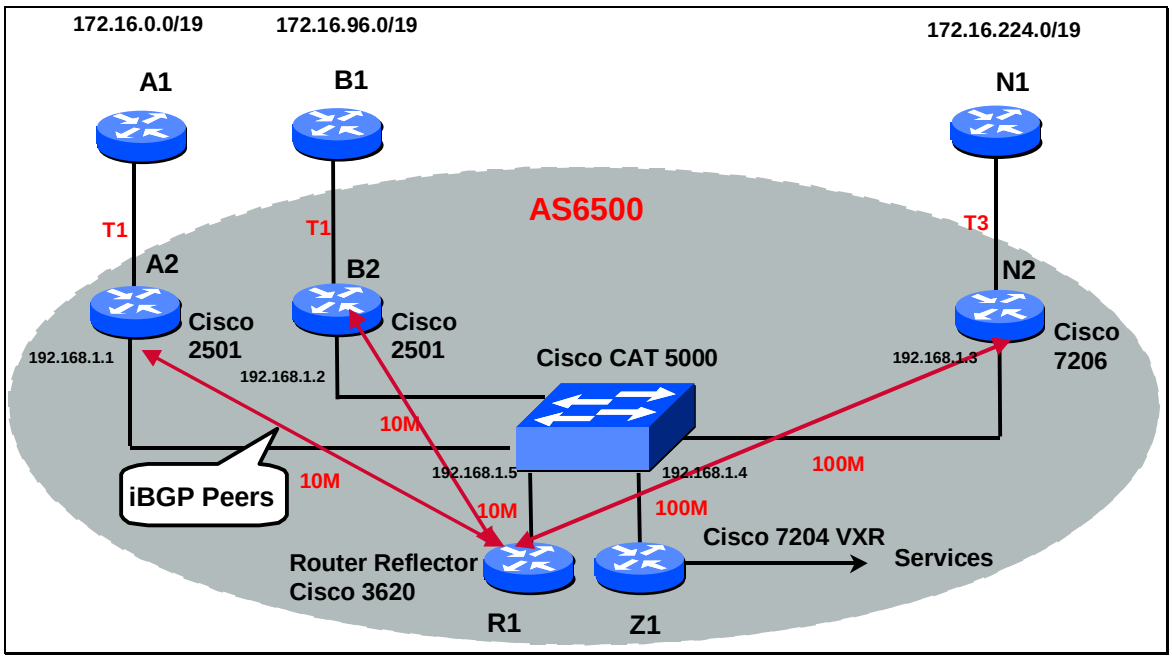

Figure 10 - How ISP Member Peer with the BGP Router Reflector

Each ISP router creates one iBGP session with the IXP's router reflector. The ISP's router will either originate or redistribute their routing information into AS 65000. The IXP's router reflector will reflect these advertisements to the other routers on the IXP. Figure 10 illustrates this configuration.

The BGP Route Reflector (Router R1 in Figure 10) will use iBGP *Peer Groups* to minimize the CPU processing load from the number of BGP peers connected to it.

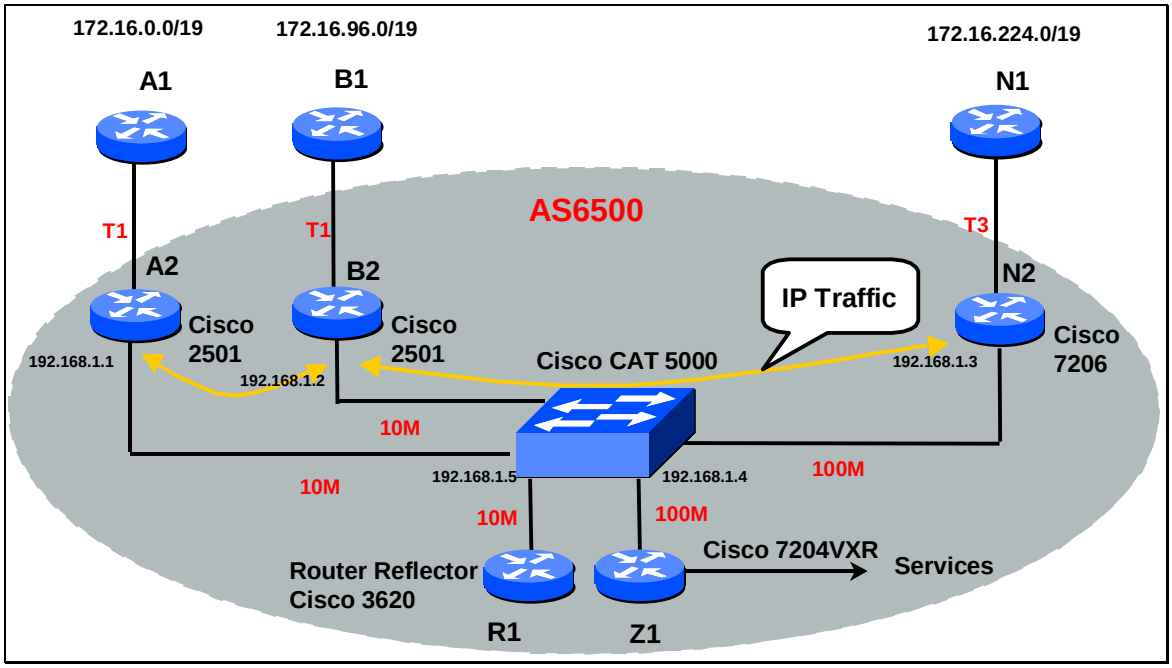

Figure 11 - Traffic between the ISPs will flow to each router - *not* through the Router Reflector

While BGP Routing information will flow between the ISP's router and the route reflector, traffic between the ISP's routers will not flow through the router reflector. When the BGP route is advertised by an ISP in to the AS 65000, the router reflector will preserve the *BGP Next Hop* of the prefix. Hence, when 172.16.0.0/19 is advertised into AS 65000, router N2 will see a next hop of 192.168.1.1. All traffic from N1 destined for ISP A will go directly to router A2.

Figure 11 provides an illustration of the traffic flow – separate from the BGP Routing flow.

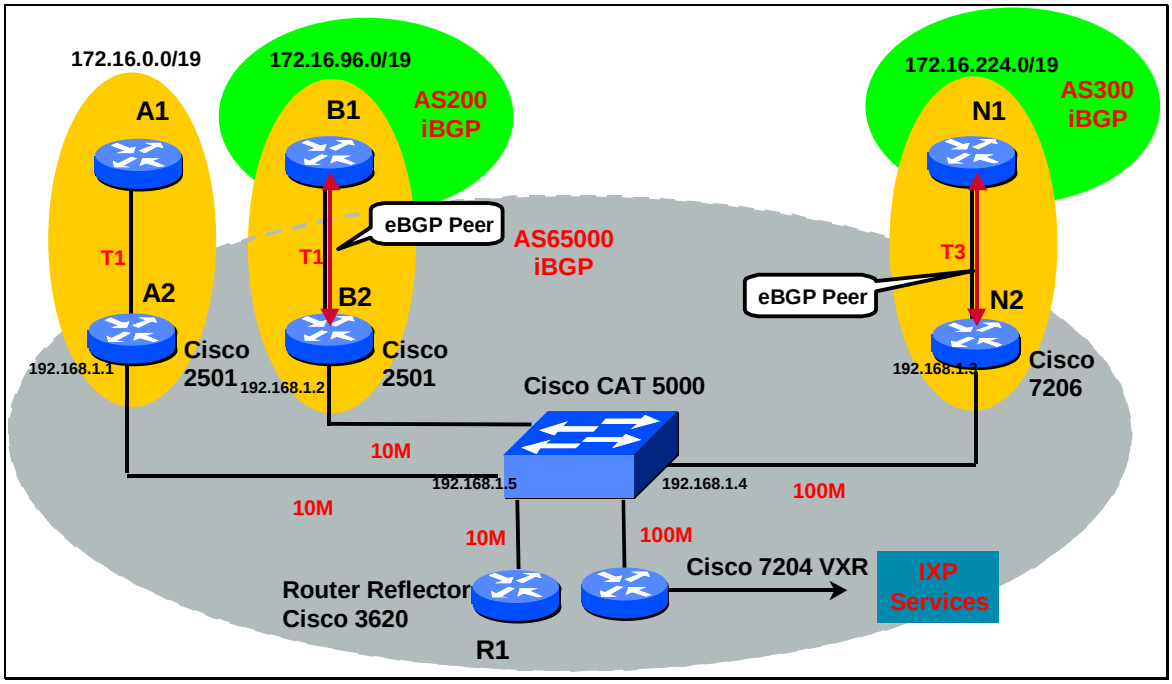

Figure 12 - ISPs with their own BGP mesh will use eBGP to peer with their IXP Router

Each ISP is responsible for their own router at the IXP. They will need to configure the router to advertise their IPv4 address block to the IXP while taking the advertisements from all the other ISPs and communicate it through out their network. The advertisement happens on the router inside the ISP's own AS number. For example in Figure 12, ISP A would advertise an aggregate of their CIDR block from Router A1 via an eBGP session to router A2. Since ISP A controls routers A1 and A2, they place extra route filters to insure only the aggregates of the CIDR blocks go out to the IXP.

The link between router A1 and A2 is provisioned, managed, and controlled by ISP A. Hence, the IP addresses on that link would be a /30 allocated from ISP A. ISP A would insert this /30 into their IGP (either OSPF or ISIS). Router A1 would have the IGP configured, but ***router A2 would not have an IGP configured.*** Router A2 would only have static routes and BGP running to pass information into its forward table. This would insure that there would be no leakages of ISP A's IGP to any other ISP on the IXP. Router A1 and A2 would have eBGP configured between them. Each router would use the link IP address for their peering, not the loopback interface.[11]

Finally, Router A2 would not have a default route configured. Configuring a default route on a router peering on an IXP opens the door for abuse by another ISP. Packets can be forward by another ISP into a router with default turned on. That router would then pass the packet to its default router – to be forward through the network out to the Internet. This has resulted in cases where one ISP "hijacks" bandwidth from another ISP's backbone. This potential problem is eliminated by not including any default route on the router connected to the IXP. In fact, all router A2 should know is how to get to ISP A's CIDR block(s) and to the CIDR blocks advertised from the other peers on the IXP. If this is the case, then the problem of bandwidth hijacking is eliminated.

[11] Unless the ISP have parallel links to the IXP. In that case, the ISP would use the eBGP Multihop technique to load balance between the links.

Scaling Paths - Directions of Growth

A Layer 2 Route Reflector IXP has several directions of growth – providing the participants with a foundation that will be growing into obsolescence. For example, an IXP can start down their scaling path by replacing the early IXP Ethernet Switch with one that is more powerful. They then upgrade their BGP Route Reflector – allow it to handle more IXP member and routes. Larger ISP customers are added to the IXP – peering with each other via eBGP – peering with the BGP Route Reflector to get to the smaller ISPs. Finally, a Route Server is added to the IXP to allow for eBGP peering directly – allowing for the larger ISPs to have the same benefit of a Route Reflector with out the extra BGP AS hop. The end result is an IXP identical to the largest IXPs in the US, Europe, and Japan – built on the foundation of a L2 Route Reflector architecture.

Upgrading the IXP Switch

Upgrading the IXP switch is one of the most cost-effective means of enhancing the IXP. A simple host swap between switches could offer instant performance improvements. Switching speeds vary among ethernet switches. A while a Catalyst 2900XL would do fine handling multiple 64 Kbps to 8 Mbps flows from the ISP members, it would not be able to handle multiple 45 Mbps flows from ISP on a larger IXP. Table 1 lists the switching capabilities of the various Cisco ethernet switches. Small IXPs in which ISP are connection in speeds ranging in 64 Kbps to 2 Mbps are more than sufficiently served with a Catalyst 2900XL.

Table 1 - Ethernet Switching Capabilities for IXP Mediums

Switch	Backplane	Forwarding Rate	Port Speeds
Cat 2900XL	3.2 Gbps	3.0 MPPS	10/100 Ethernet & EtherChannel
Cat 3500XL	10.8 Gbps	8.0 MPPS	10/100 Ethernet & EtherChannel
Cat 4000	24 Gbps	18 MPPS	10/100/1000 Ethernet & EtherChannel
Cat 5000	50 Gbps	36 MPPS	10/100/1000 Ethernet & EtherChannel
Cat 6000	32 Gbps	15 MPPS	10/100/1000 Ethernet & EtherChannel
Cat 6500	256 Gbps	+100MPPS	10/100/1000 Ethernet & EtherChannel

Direct Peering between Peers and the Router Reflector

As the IXP grows and matures, situations will arise where an ISP will want to peer with the IXP, gain the benefits form the IXP, but does not want to use the BGP Router Reflector. Instead, they wish to have direct peering via eBGP and their own AS number. They will understand that the router requirements and have made their decision. In this case, these ISPs will connect to the IXP, but use eBGP peering with the Router Reflector and all other peers who are using the same technique.

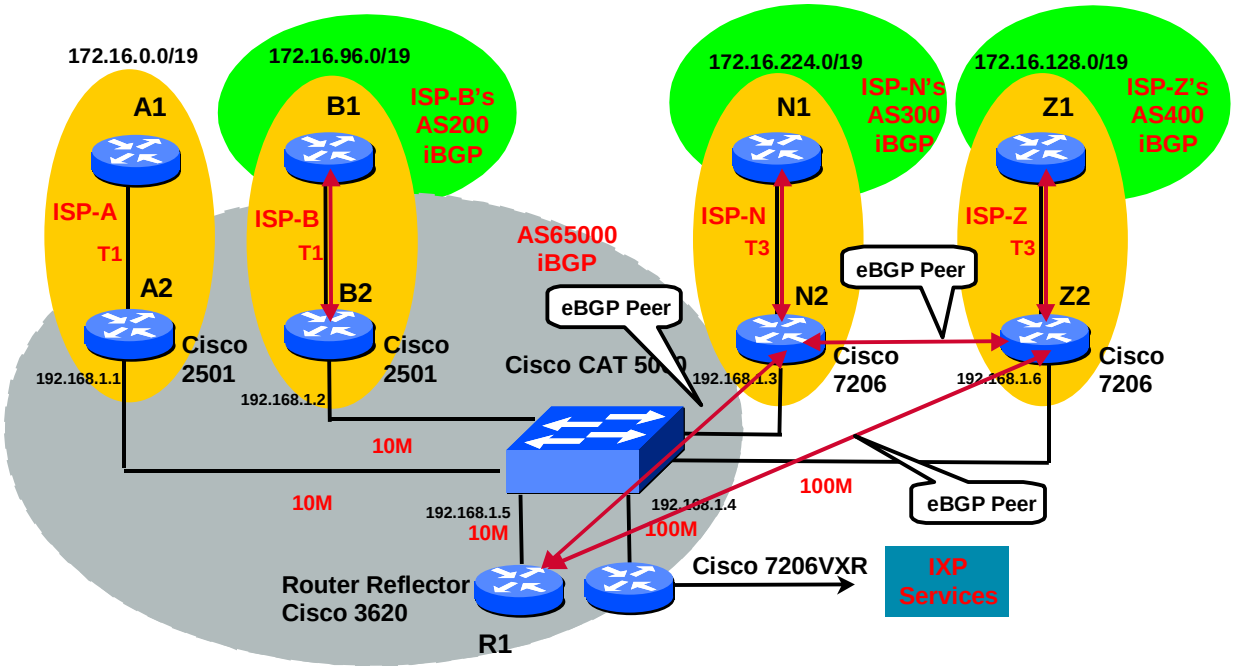

Figure 13 - Scaling - ISPs with direct eBGP Peering

Figure 13 illustrates two ISPs who have selected direct peering – ISP-N and ISP-Z. To connect to the IXP, each ISP must have a eBGP session with the Router Reflector and to each other. The connection to the Router Reflector provides the routes from the other IXP members. Since all connections are on the same shared IXP medium, the next-hop from the iBGP peer is used on the routes sent routers to N2 and Z2.
This option to have ISP connect to the IXP, peer via eBGP, and not disrupt the existing Router Reflector based IXP is an important scaling option. It provide the flexibility needed to insure the theme of ***interconnection*** is not limited by any limitations of the L2 Router Reflector based IXP. It allows small ISP limited on capital to use a Cisco 1600-M connect via the Route Reflector and community with another ISP with a Cisco 7204VXR who is peering directly via eBGP.

Transition to a Router Server

While the BGP Route Reflector architecture has proven to work, some IXPs may wish to transition to a full *Route Server* implementation. The Router Server project was started in 1994 as a way to scale the BGP session on an IXP. Some of the objectives between the Router Reflector and Router Server are the same. For example, both solutions take care of the memory and CPU scaling issue of N-1 BGP peering sessions in the router. The big difference is that the Router Server uses eBGP to for each of the peers while the Router Reflector uses iBGP.

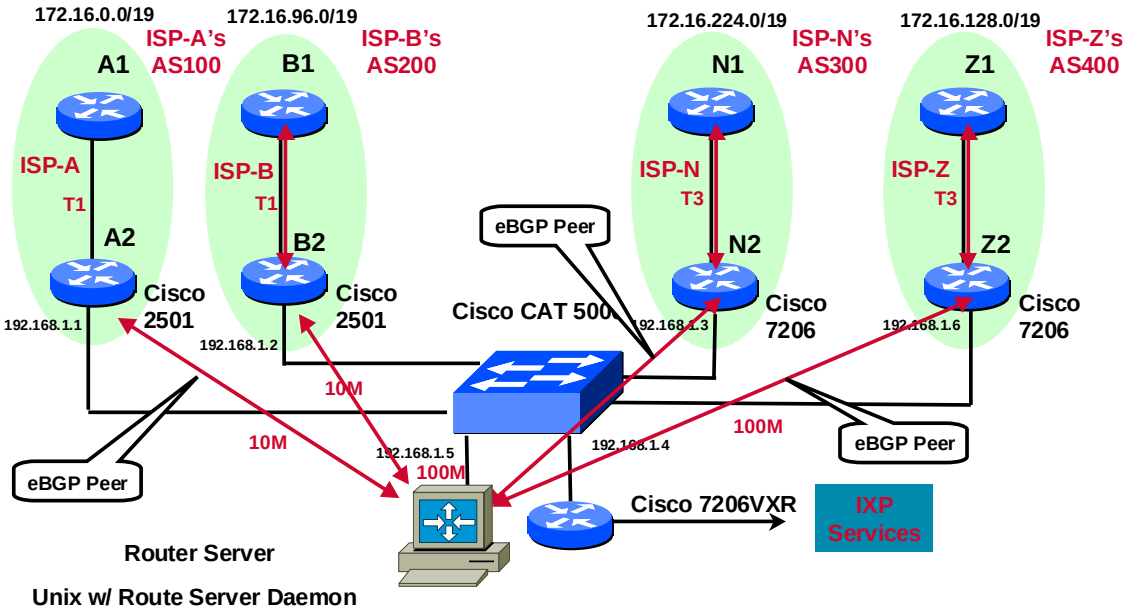

Figure 14 - Scaling - Migrating to a Router Server

The real value the Router Server adds is a way to implement route announcement policies through the Router Server. Details of how this works is beyond the scope of this paper. Further deepening is suggested at the following sites:

Merit's Global Routing & Operations Web Site:
http://www.merit.edu/internet/
Merit's Route Server Next Generation Project
http://www.rsng.net/
The Routing Arbiter Project
http://www.ra.net/

The key point from the context of a L2 Router Reflector based IXP, is that a Router Server is a future scaling option. IXP operators are *not* locked into one way of peering across the IXP. In fact, an IXP could easily keep its Router Reflector, allow for point to point peering, and insert a Router Server. While the complexity of the IXP increases, the option is there (although not recommended).

Routers Options for the L2 RR IXP Architecture

The core advantage of the use of a BGP Route Reflector on an IXP is the ability to use smaller routers. Since each router peering with the Route Reflector will only have one (or two) peer sessions, the amount of memory consumed to handle each peer session is saved. Since an IXP would not carry the full Internet route table, huge amounts of memory are not required for the router the ISP brings to the IXP[12]. Experience with HKIX has proven that a L2 RR based IXP will work with Cisco 2501 routers as the ISP router. The Cisco 2501 will have its memory and flash maxed out (16M Ram/ 16M Flash). Using a Cisco 2501with 16M of memory vs a Cisco 7206 with 128M of memory significantly reduces the cost for a small ISP to join and benefit from the IXP.

[12] As of August 28, 2000, the Internet Route table is over 90000 routes. This means an average Cisco router would need more than 64M of memory to handle the full table.

Some factors ISPs need to consider when selecting a router for an IXP:

- **Physical IXP Interconnection** - What is the physical interconnect medium? The most common IXP mediums are ethernet based (include Gigabit ethernet). Other IXP mediums in use today include Switched FDDI and ATM. Future IXP mediums will be optimal based (SDH/SONET).
- **Traffic** - How much traffic will be sent across the IXP? The volume of traffic across the IXP Router will heavily influence the selection. Packet Per Second (PPS) capability is the key factor.
- **Routes** - The number of routes advertised to the IXP peers, the number received form the IXP peers, and the size of the routing tables all effect the amount of memory needed on the router.
- **Peer Connections** - How many total BGP peers will there be on the IXP? Each BGP peer session takes up memory. Hence, it will effect the amount of memory required on the IXP router. BGP Route Reflector and Route Servers reduce this factor by allowing one BGP peer session to work with all.
- **Policy Enforcement** - How large do the route filters need to be to enforce peering policies? Route Filters take up CPU time. Effecting the capability of the CPU needed on the IXP Router.
- **Dampening** - Can your router withstand route flapping on the IXP?
- **Network Ingress and RFC 1918 Filtering** - Minimizing the effects your network will have on the Internet. Some networks need to be to be before they enter the ISP's network While BCP 38/RFC2827 strongly advices to perform the ingress filtering on the customer edge.
- **Security** - Filters needed to protect your router and network from attack.
- **Internal Network** - How is your network designed - physical, routing protocols, etc.
- **Statistics Tools** - Do you want to turn NetFlow on for gathering traffic statistics on the IXP edge router.
- **Multicast Support** - Do you want to have Mbone distributed through the IXP Router? Do you want a inter-domain multicast service through your IXP router (i.e. with Multicast BGP)?
- **Accounting** – Will you be using MAC, CEF, or Prefix accounting on the IXP router?
- **CDN Redirection** – Will you redirect packets into you Content Distribution Network (CDN)?

When HKIX was created in April of 1995, the Cisco 2501 was the *low end* router of choice. Today, while the Cisco 2501 still works and maintains its popularity, there are other options for ISPs to consider. The core requirements for the ISP's selection of the "IXP Router" vary. The interface types, packets per second (PPS) capability, security considerations, ability to handle large router tables, and other features will determine which router is the *right fit* at any give point of a ISP's life cycle. Table 2 provides a brief comparison for ISPs to begin their evaluation[13].

[13] ISPs should consult the on-line product catalog – www.cisco.com – for details and up to date information. ISPs can also submit a RFI (Request For Information) to their Cisco Partners and/or Cisco directly to get pricing details.

Cisco Router	CPU	Max Memory	Max Flash	Flash Card	Rack Mount	Serials	HSSI	Ethernet	Fast Ethernet	Packet Per Second (PPS)
1605M[14]	Motorola 68360 at 33 MHz	24 MB	16MB	✔		✔		✔		4 kpps fast switched
1750	Motorola MPC860T PowerQUICC @ 48 MHz	48MB	16MB			✔		✔	✔	8 kpps fast switched
2501	20 MHz 68030	16MB	16MB		✔	✔		✔		
2610	40 MHz RISC	64MB	16MB		✔	✔		✔		15 kpps fast switched
2620	50 MHz RISC	64MB	16MB		✔	✔		✔	✔	25 kpps fast switched
2650	80 MHz RISC	128MB	32MB		✔	✔		✔	✔	37 kpps fast switched
3620	80-MHz IDT R4700 RISC	64MB	32MB	✔	✔	✔	✔	✔	✔	20-40 kpps fast switched
3640	100-MHz IDT R4700 RISC	128MB	32MB	✔	✔	✔	✔	✔	✔	50-70 kpps fast switched
7206w/ NPE 200	RISC R5000 @200MHz	128MB	128MB	✔	✔	✔	✔	✔	✔	600 Mbps backplane + 200 kpps fast switched
7206 w/ NPE 250	RISC R5271 @263MHz	128MB	128MB	✔	✔	✔	✔	✔	✔	600Mpbs backplane + 225 kpps CEF switched
7200 VXR w/ NPE 300	RISC RM7000 @263MHz	256MB	128MB	✔	✔	✔	✔	✔	✔	1 Gbps backplane + 300-kpps CEF switched
7200 VXR w/ NSE-1	PXF processor + RISC RM7000 @263MHz	256MB	128MB	✔	✔	✔	✔	✔	✔	1 Gbps backplane + 300-kpps CEF switched

Table 2 - Cisco Router Comparison - What is the best fit for an ISP's router on the IXP?

Example of a BGP Route Reflector IXP

Cisco Systems is providing the core IXP equipment for ISP who collaborate to create an IXP. The core equipment includes two routers for the route reflectors, two 10/100 ethernet switches, and a 2511-RJ for out of band access. Figure 15 illustrates one example of a BGP RR based IXP. Two Cisco 3620s with 64 MB or memory comprise the BGP route reflectors. One Catalyst 2924XL operates as the primary IXP fabric. The second Catalyst 2924XL is used for the management network and can be used as a backup if the primary switch has a catastrophic failure. The Catalyst 2924XL, with 3.2-Gbps switching fabric and 3.0 million packets-per-second forwarding rate is an cost efficient switch for IXPs that will have ISPs pulling 64 Kbps to 8 Mbps links into the exchange.

[14] The Cisco 1600 requires the IOS Plus Feature set for IXP operation. The IOS Plus Feature set contains BGP.

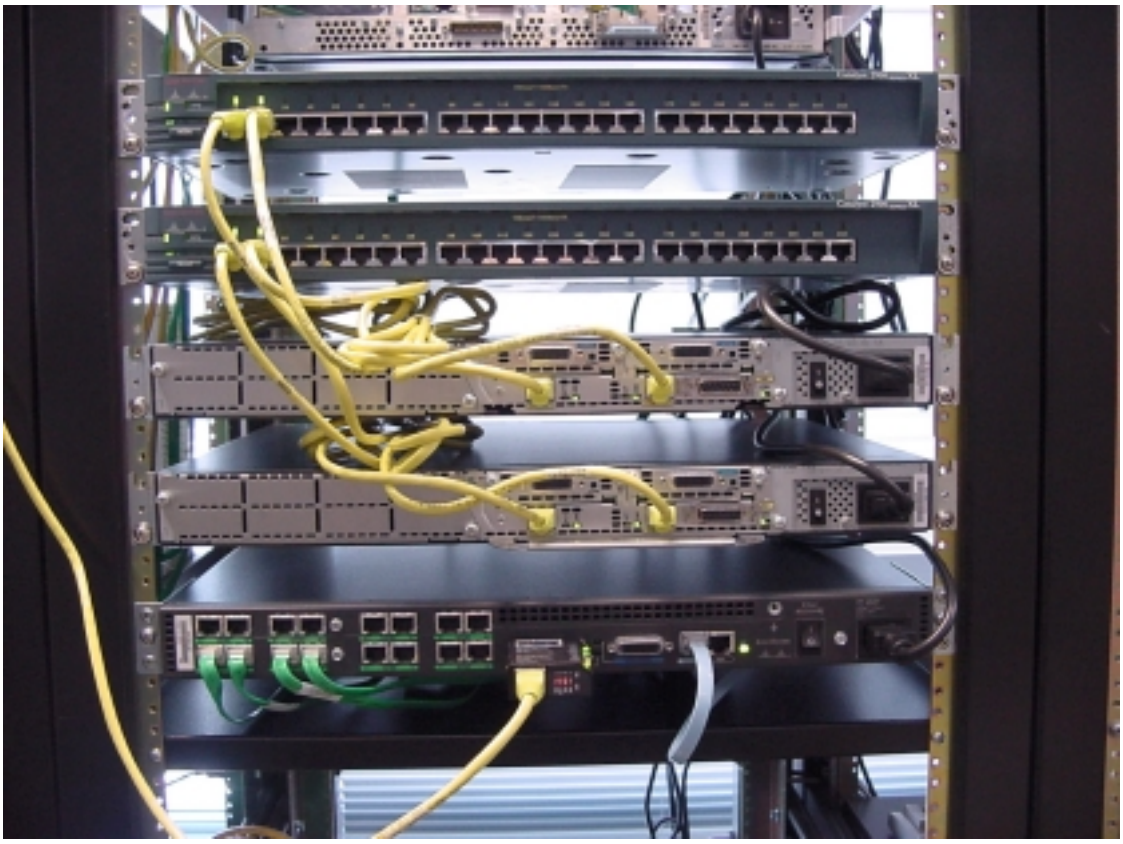

Figure 15 - Core IXP Equipment Example

Technical Design Details

Where does the IXP get its IPv4 Addresses?

Over the history of IXP development, the actually IPv4 address allocated for the IXP interconnection medium have come from a variety of resources. With the advent of RFC2050, the Regional Internet Registries (RIRs) defaulted to a minimal allocation block of a /19 (it is now a /20 as of Nov 2000). A /19 or /20 is way to large for an IXP. A /24 is more than sufficient for the requirements of an IXP. As a path to the RIR's allocation limitations, a special allocation was provided to ISI (a.k.a Bill Manning at bmanning@isi.edu or www.ep.net). This *IXP Registry* allowed IXPs to get unique IPv4 address blocks in allocations of /24. This was acceptable to the ISP community since IXP blocks should not be globally advertised on the Internet.[15] Since the early part of 2000, each of the RIR's has elected to create a *micro allocation* policy.

[15] The IXP's /24 should remain inside each of the ISP members routing domains. They should not be globally advertised to the rest of the Internet. The core reasons are for security (minimizing risk) and minimizing the growth of the IPv4 Internet Routing Table.

This new allocation policy allows each RIR to allocate /24 to IXPs and other infrastructure critical to the Internet (i.e. segments that have root domain name servers).

As of November 2000, new IXPs should go to their RIRs to get their unique IPv4 micro allocation.

The special ISI Registry for IXPs still exist, but special emphasis will be made to have new IXPs go to the RIRs for address space.

Autonomous Systems (AS) Number

Autonomous system (AS) numbers are used by BGP to describe administrative boundaries on the Internet. A BGP base Route Reflector IXP needs an AS number all it's own. As to which AS number is used (public or private) depending on the IXP operations. It is recommended that the IXP gets a public AS number for their RIR.

The IXP would get it public AS number from one of the RIR's or from the special ISI registry set of for IXPs. The preferred path would be for the IXP to get their AS number from their RIR. The core issue for the IXPs is who will pay for the AS number allocation. That is an issue the ISPs must collectively work out.

Does an ISP need a Unique AS Number to peer with the IXP?

It is strongly recommended that all ISPs get a Globally Unique AS Number. While there are ways to have an ISP connect and gain the benefits of a L2 IXP with out a Globally Unique AS number, an ISP with their own AS number has advantages over ISPs who do not. The core requirement for an ISP to qualify for a Globally Unique AS number is that they be *multihomed.* If an ISP is connected to an upstream provider and peering over a domestic IXP, then they qualify for a globally unique AS number.

Once the ISP qualifies for the AS number, they must apply to their RIR and pay the setup and annual fees to maintain the AS number.

How to ISPs connecting to the IXP gets a unique AS Number?

Public AS numbers are allocated by one of the RIRs. There are currently three active RIRs providing Autonomous Systems Numbers (AS number). These are ARIN (www.arin.net), APNIC (www.apnic.net) , and RIPE-NCC (www.ripe.net). The basic requirement to justify a unique AS number is that the requestor be multihomed. Since IXPs are all about ISPs multihoming[16] ISPs connecting to a IXP would have no problem justifying an AS number allocation. The key issue the ISP will need to resolve is the membership to the RIR and payment for the allocations.

Example of a AS allocation from Asia & Pacific Network Information Center (APNIC)

Organizations in the Asia & Pacific region get their Autonomous Systems Number (AS number) from APNIC (http://www.apnic.net). To apply to get your own AS number, you first must be either a member of APNIC or you will need to pay a one time charge and annual maintenance fee for the AS number. Current Internet policies regarding ASN allocations are documented in RFC 2050 (ftp://ds.internic.net/rfc/rfc2050.txt). APNIC-048 template (ftp://ftp.apnic.net/apnic/docs) will need to be filled in and submitted.

[16] At a minimum the ISP will have one connection to their upstream provider and a second connection to the IXP.

The following is an example from the AS forms submitted by ISPs connected to the Indonesian Internet eXchange (IIX – www.iix.net.id). Each ISP would fill in the as-in as-out as with the AS number from the IXP (in this case 7597):

```
as-in:     7597  ANY
as-out:    THIS-AS ANY
```

Explanation:

```
as-in:     7597  ANY     !  receive any routing information from AS 7597 ( IIX AS number)
as-out:  THIS-AS ANY     !  send any routing information from your as number.
```

After completing this form, the ISP would submit it via email to:

as-request@rs.apnic.net

Check each RIR's web site for more details on these the AS number applications and requirements.

Can an IXP or ISPs on the IXP use Private AS Numbers?

Router Reflector based L2 IXPs can use either a unique AS number or a private AS numbers.[17] Either would work, but a public AS number is recommended. The primary concern is the added complexity. Private AS number must be kept off the public Internet. So special filters must be put in place to insure that the private AS numbers are removed from any advertisements. The other concern is over-lapping private AS numbers. The private AS number must be one that all ISPs agree that will be used exclusively in their country for the IXP. If there is some doubt as to how use of the private AS number on the IXP will interfere with current ISP operations, then it would be recommended to get a publicly registered AS number from the RIR.

If private AS numbers are used either as the IXP's AS number of by any ISP connected to the IXP, it would be Best Common Practice to use the BGP *neighbor xxx.xxx.xxx.xxx remove-private-as* command to remove private AS numbers from the IXP.

```
router bgp 109
 neighbor 222.222.10.17 remote-as 110
 neighbor 222.222.10.17 remove-private-AS
 neighbor 222.222.10.17 prefix-list Customers out
```

For example in Figure 16, ISP 'B' does not yet have a public AS number. So get connected quickly, they have used a private AS number 65000. ISP 'N' has a public AS number. They need to insure that this private AS number does not interfere with their operations and does not get mistakenly advertised on the Internet. The BGP *neighbor xxx.xxx.xxx.xxx remove-private-as* applied on the eBGP peering connection on router N2 going to router N1. This will ensure that the private AS number will not interfere with ISP 'N's' internal operations. Since Router N2 should have the BGP command *next-hop-self*, removal of the private AS will not interfere with the packet flow between ISPs B and N.

As a second stage *Murphy's Law* filter, ISP N also put the BGP *neighbor xxx.xxx.xxx.xxx remove-private-as* on the gateway router to their upstream connection N-GW. This insures that if for some reason the filter on Router N2 breaks, that the private AS number will not get leaked out to the Internet. It also put an internal check in case of any of ISP N's customers using private AS numbers are removed form the upstream advertisements. Also note that router N-GW should

[17] Private AS numbers are reserved for internal use and are not to be advertised on the Internet. The include a range of 64512 to 65535.

not be advertising any of ISP B's routes. So the if the remove-private-as command breaks on router N2, then the BGP Policy filter that removes ISP B's routes from ISP Ns advertisements should do the primary filtering.[18]

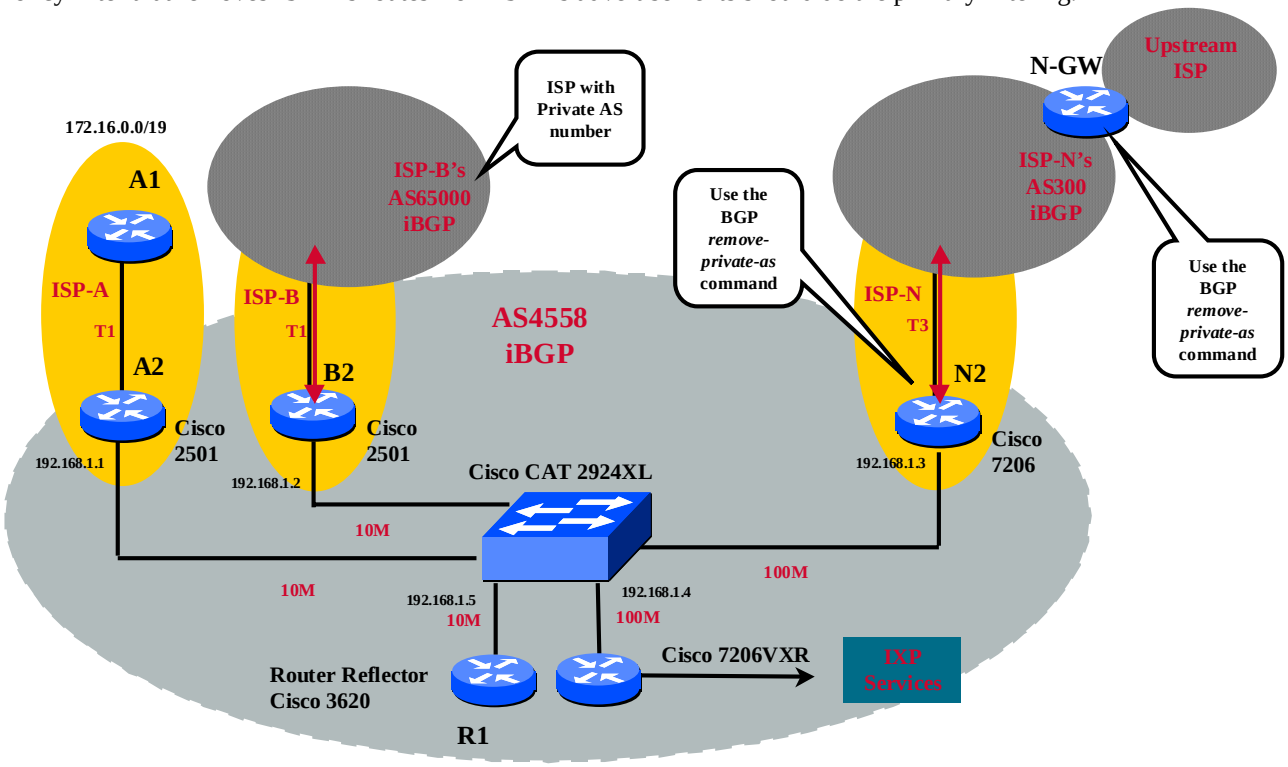

Figure 16 - Removing Private AS numbers from the IXP Peering information

How does the IXP get Transit for IXP services?

IXPs need to have some sort of global transit arrangement to the rest of the Internet. The IXP's web servers and other publicly accessibly services need to be reachable via the entire Internet. There is several ways that this can be done. Each IXP will work out a solution that is most appropriate for them. The one thing that cannot be assumed by the IXP management is *automatic* transit. Just because lots of ISP are peering on a IXP does not guarantee the IXP's accessibility through those ISPs to the rest of the Internet. In essence, the IXP management must enter some sort of customer relationship with one of their IXP customers to get access to the Internet.

Route Reflector Configuration

[18] ISP B's routes can be filtered from ISP N's upstream advertisements by using BGP Community, AS-Path, Prefix List, or Distributed List filters. The selection of the filter type depends on how ISP N configures their BGP and the personal preference of the ISP N's Network Engineer.

The core configuration of the BGP Route reflector is very simple. If we were to use the example presented in Figure 17, the BGP configuration one the BGP route reflector (router R1) would look like:

```
! Router R1 – IXP Route Reflector
  router bgp 65000
  no synchronization
  bgp cluster-id 1000
  bgp log-neighbor-changes
  neighbor IXP-Peers peer-group
  neighbor IXP-Peers remote-as 65000
  neighbor IXP-Peers route-reflector-client
  neighbor IXP-Peers send-community
  neighbor IXP-Peers version 4
  neighbor IXP-Peers soft-reconfiguration inbound
  neighbor IXP-Peers distribute-list 150 in
  neighbor IXP-Peers distribute-list 150 out
  neighbor IXP-Peers route-map Murphy-No-Export out
  neighbor IXP-Peers password 7 072404627728303D
  neighbor IXP-Peers maximum-prefix 1000 500
  neighbor 192.168.1.1 peer-group IXP-Peers
  neighbor 192.168.1.1 description ISP ‘A’
  neighbor 192.168.1.2 peer-group IXP-Peers
  neighbor 192.168.1.2 description ISP ‘B’
  neighbor 192.168.1.3 peer-group IXP-Peers
  neighbor 192.168.1.3 description ISP ‘N’
  no auto-summary
!
access-list 150 deny   ip host 0.0.0.0 any
access-list 150 deny   ip 10.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 150 deny   ip 127.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 150 deny   ip 169.254.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 150 deny   ip 172.16.0.0 0.15.255.255 255.240.0.0 0.15.255.255
access-list 150 deny   ip 192.0.2.0 0.0.0.255 255.255.255.0 0.0.0.255
access-list 150 deny   ip 192.168.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 150 deny   ip 224.0.0.0 31.255.255.255 224.0.0.0 31.255.255.255
access-list 150 permit ip any any
!
route-map Murphy-No-Export permit 10
   set community no-export
!
route-map Murphy-No-Export permit 20
!
```

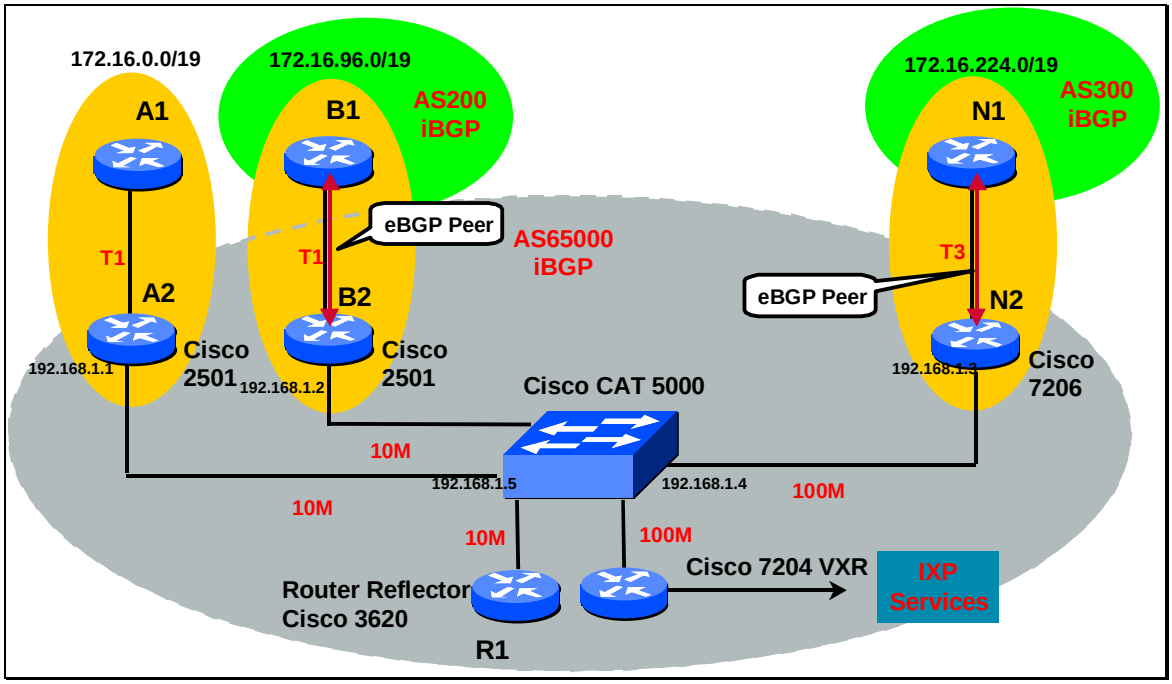

Figure 17 - L2 Router Reflector IXP Example

Each of these commands are have an important function in a ISP' network. The reasons and functions are described in detail.

`no synchronization` - Should be configured for all BGP speaking routings on the Internet. Synchronization will try to get the IGP and BGP in sync – which will never happen on a ISP/IXP router.

`bgp cluster-id 1000` - Used when there are two BGP Route Reflectors on the IXP (for redundancy). Keeps the two in sync.

`bgp log-neighbor-changes` – Logs all BGP status and state changes to the router log file – which could then be exported to a syslog server for trends/fault analysis.

`neighbor IXP-Peers peer-group` – Peer Groups are used for easier and consistent configuration management. As seen further in the config, each config for the ISP router is standard two lines.

`neighbor IXP-Peers remote-as 65000` – AS number of the IXP.

`neighbor IXP-Peers route-reflector-client` – Used to turn the session into a BGP Route Reflector session. The router reflector router determine is the peering relationship is a standard iBGP mesh or a iBGP route reflector.

`neighbor IXP-Peers send-community` – Send the BGP Community attribute across the IXP. That allows each ISP to use BGP communities for their policy implementation.

`neighbor IXP-Peers version 4` – The vast majority of BGP routers are BGP version 4. This command locks the session into a BGP version 4 only. If an old router with BGP v3 were to try to negotiate a session, it would be dropped.

neighbor IXP-Peers soft-reconfiguration inbound – Not all routers can use the new route refresh capabilities. Soft reconfiguration will keep all the BGP advertisements in the BGP table so they can be reapplied to filters with out resetting the BGP session. Resetting the BGP session interrupts the traffic flow – which causes a network outage.

neighbor IXP-Peers distribute-list 150 in
neighbor IXP-Peers distribute-list 150 out – Distribute list 150 are all the RFC1918 and martin routes that should not be advertised outside of an ISP. Each ISP should be doing this filter on their ingress/egress route filters. This filter double-checks the ISP's filters.

neighbor IXP-Peers route-map Murphy-No-Export out – See the details explaination below in the route map section. This route-map will have all the advertised prefixes set with a BGP community of "no-export." That will inform each ISP's BGP protocol to not export the prefixes received over the IXP to their upstream peers.

neighbor IXP-Peers password 7 072404627728303D – Puts a MD5 checksum on all BGP packets between the IXP's route reflector and the ISP routers. This is a critical security feature. The ISP routes must have a matching password for the BGP session to become active. The IXP operator must coordinate the passwords with the ISPs.

neighbor IXP-Peers maximum-prefix 1000 500 – There have and will be cases where an ISP's router starts to advertise more routes than it should. One danger on the IXP with small routers is an ISP mistakenly advertising the full Internet Route Table. This would take down the IXP. Maximum-prefix puts a limit on the number of prefixs advertised. If the prefix limits are exceeded, the BGP session to that router is shut down. ISPs would be encouraged to do the same on their router. This config is set to limit the number of prefixes to 1000 and provide a warning to the routers log file at 500. If a ISP advertised more than 1000 prefixes, BGP would shutdown the connection – protecting everyone from a flood of prefixes.

```
neighbor 192.168.1.1 peer-group IXP-Peers
neighbor 192.168.1.1 description ISP 'A'
neighbor 192.168.1.2 peer-group IXP-Peers
neighbor 192.168.1.2 description ISP 'B'
neighbor 192.168.1.3 peer-group IXP-Peers
neighbor 192.168.1.3 description ISP 'N'
```

ISP Peering sessions created for each of the routers on the IXP. The description command helps to tell which ISP owns the session.

no auto-summary – Required on today's Internet. With out this, allocations from the old Class A space would get aggregated into a class /8. For example, a 64.1.0.0/19 would be aggregated as a 64.1.0.0/8.

```
access-list 150 deny   ip host 0.0.0.0 any
access-list 150 deny   ip 10.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 150 deny   ip 127.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 150 deny   ip 169.254.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 150 deny   ip 172.16.0.0 0.15.255.255 255.240.0.0 0.15.255.255
access-list 150 deny   ip 192.0.2.0 0.0.0.255 255.255.255.0 0.0.0.255
access-list 150 deny   ip 192.168.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 150 deny   ip 224.0.0.0 31.255.255.255 224.0.0.0 31.255.255.255
access-list 150 permit ip any any
```

List of routers that should not be advertised on the Internet. This filter on the IXP Router Reflectors reinforces filters that should be placed on the advertisements from each ISP's router. Distribute List are used in this configuration. They are the

most commonly used form of route filtering. Prefix List can also be used and are gaining in popularity. See the second on BGP Route Filtering later in this document for more details.

```
route-map Murphy-No-Export permit 10
   set community no-export
!
route-map Murphy-No-Export permit 20
```

This route-map sets any prefix advertisement going out through the route reflector to have the BGP community *no-export* set. This community allows the advertisement to go one AS hop away (i.e. the adjacent neighbor). The neighboring ISP who receives this prefix will not advertise it to their other BGP peer connections. Figure 18 provides an illustration on how this works. ISP A advertises their assigned address block 170.10.0.0/20 to the IXP. The IXP's Router Reflector receives this route and will advertise it to all the other IXP routers. Before that happens, the route refection will set the prefix to have a BGP community of *no-export*. When ISP B receives the prefix, it will propagate the advertisement inside ISP B's network. Since ISP B's BGP Gateway routers see the *no-export* command, it will filter the prefix from the upstream #2's connection.

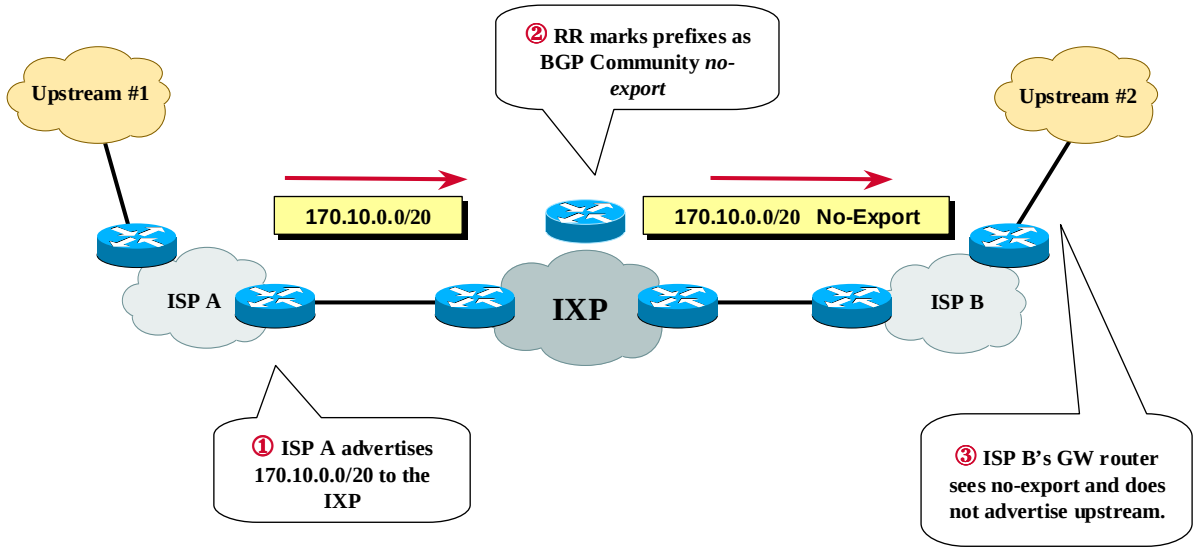

Figure 18 - Using BGP Community *no-export* as a safe guard.

Connecting the ISP to the L2 Router Reflector IXP

Preparing the ISP to connect to the IXP

ISPs who connect to IXPs and wish to gain significant benefits from the local interconnect must insure that their network is ready. ISPs who have grown with simple static routes as their "layer 3 glue" must deploy an Interior Routing Protocol (IGP). ISPs who have not dutiful considered how they will interact with the largest network on the planet (i.e. the Internet) needs to create their interconnection policies. Some ISPs may need to redesign their entire network to be truly integrated

with the benefits of the connection to the IXP. The good news is that ISPs who are growing from small one to two router operations must do this some day – the IXP just becomes the excuse. The bad news is that this is a bit of work to be done. It is recommended that ISPs spend extra investment of time to insure their network is optimized not only for the benefits of the IXP, but so the network can scale to meet the demands of growth.

> **No amount of magic knobs will save a sloppily designed network**
>
> **Paul Ferguson—Consulting Engineer, Cisco Systems**

There are not short cuts to good network design. Many young ISPs get away with sloppy network designs because their growth rates a low. Yet, one an ISP interconnects to the new revenue opportunities of an IXP, the growth rate forces the issue. Sloppy network designs fail under growth. When they fail, it is usually spectacular and revenue impacting. Proactive efforts to redesign the network are the preventive and prescribed remedy.

A wealth of materials focused on ISP network design, routing protocols and other information can be found on the CTO Corporate Consulting archive at http://www.cisco.com/public/cons/.

ISP Router's Configuration to the IXP Router Reflector

The ISP's router on the IXP is a critical point where route advertisement routing filter, packet filtering, traffic analysis, and security tools all have to work together to provide effective IXP connectivity. Each of these requirements will take some efforts to learn, design, and implement. This section will briefly review a working configuration. Each command will behave a short description with references to where more detailed information can be obtained.

Figure 19 illustrates how ISP N is connected to the IXP. The specific router we are addressing in this section is N2. The ISP's router on a L2 BGP Router Reflector IXP is in an interesting role of having control of the ingress/egress flows, but not the BGP routing protocol. In the example, router N2 is part of the IXP's Autonomous System number – AS 4558.[19] Router N2 is also part of ISP N's OSPF administrative domain (OSPF is used for ISP N's IGP). EBGP is used between routers N1 and N2, providing a link of the external prefixes into ISP N. Finally, router N2 is totally owned and controlled by ISP N. No one on the IXP – even the IXP operator – has access to router N2.

[19] AS 4558 is the AS number for the Kenyan IX.

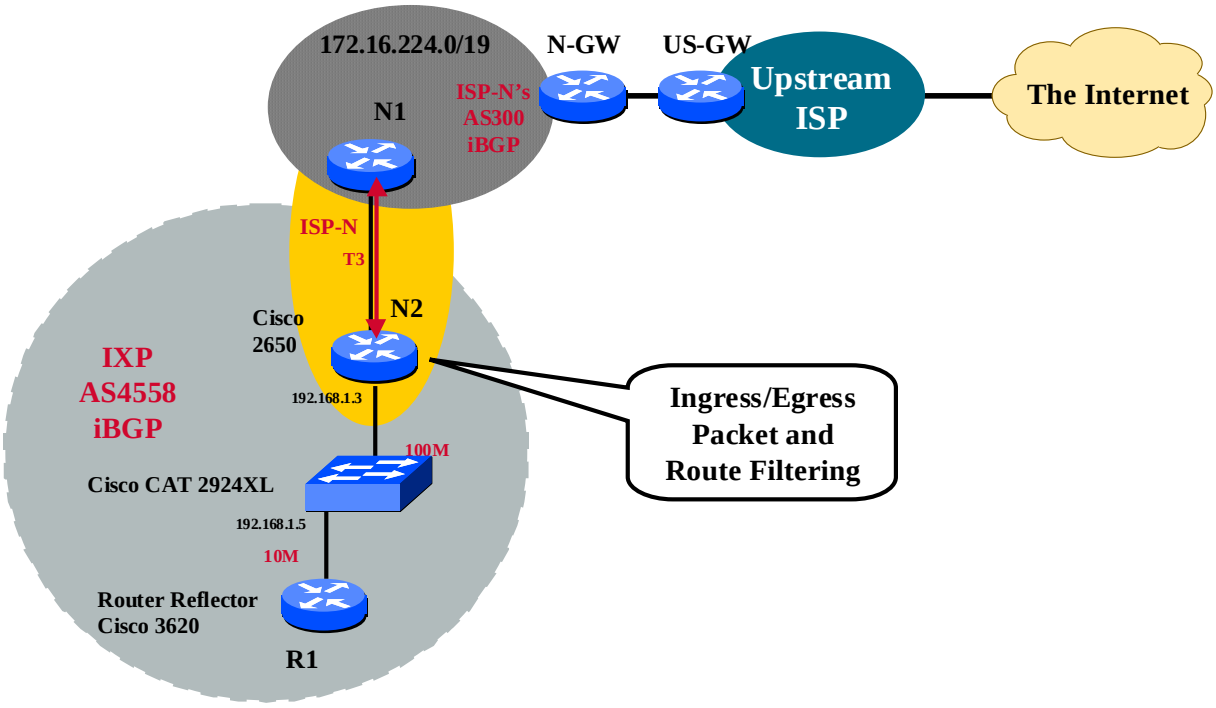

Figure 19 - ISP's Routers on the IXP

ISP Router's Features That Need to be Configured

- CEF Switching. The switching/forwarding path of a router on the IXP should use the most effective path available. Switching/forwarding path's effectiveness is measured in Packets Per Second (PPS). For the Cisco router, that is CEF switching. For many routers and IOS versions, CEF is turned off by default. Hence the ISP should insure that "ip cef" or "ip cef distributed" included in the config.

- ISP's Interior Gateway Protocol (IGP). OSPF is used in this example for the ISP's IGP.[20] OSPF will carry the BGP next-hop prefix through ISP N's network.

- Ingress Packet Filtering. Filter packets coming into the network. Unicast RPF is used in this example.

- Egress Packet Filtering. Filter packets leaving the network – insuring the source IP address matches the address block of ISP N (172.16.224.0/19 in this example).

- Ingress Packet Coloring. No packet entering ISP N's network should have a IP Precedent in the ToS Field set to 6 or 7. Both are reserved for network use. Some people on the Internet will set their packets to type 7 in an effort to have their packets get preferred treatment.

- General Security and Management Services. Access, network management, Network Time Protocol, syslogging, and other services need to be configured for effective management of the router.

[20] The most popular IGP for ISPs is OSPF followed by ISIS and EIGRP. RIPv2 is not recommended as a ISP's IGP.

- BGP with the IXP and the Internal Network. IBPG is used with the IXP while eBGP is used with the ISP's AS number.

```
! Router N2 – a Cisco 2650

ip cef
!
interface Ethernet 0/0
!
router bgp 4558
  neighbor 192.168.1.5  remote-as 4558
  neighbor 192.168.1.5  description IXP Route Reflector
  neighbor 192.168.1.5  send-community
  neighbor 192.168.1.5  next-hop-self
  neighbor 192.168.1.5  version 4
  neighbor 192.168.1.5  soft-reconfiguration inbound
  neighbor 192.168.1.5  distribute-list 150 in
  neighbor 192.168.1.5  distribute-list 150 out
  neighbor 192.168.1.5  password 7 072404627728303D
  neighbor 192.168.1.5  maximum-prefix 1000 50
```

IXP Router's Configuration to the ISP's Backbone

ISP's Gateway Router to their Upstream Connection

BGP Route Filtering and IXPs

Security and Policy enforcement is a requirement between ISPs. Each ISP is responsible for what they transmit/advertise to other ISPs and what other ISPs transmit/advertise to them. Filtering on the packets and routing protocols are the two key ways ISP accomplish policy and security filters. Policy filters enforce agreements.

Filtering routes via BGP is necessary to enforce the IXP's peering policy and to protect the members of the IXP from unwanted routes. Two filtering techniques are used in this configuration: *AS Path Filters* and *Distribute Lists Filters*.

AS Path Filters

AS Path Filters will be used to enforce the policy for the IXP. Only the ASNs with belong to the members of the IXP will be distributed to the IXP members. This *explicit permit* rule will, by default, deny ASNs which are not explicitly permitted in the AS Path Filter. This prevents routes leaking from one provider from being propagated across the IXP. For example, if provider A was connected to Internet MCI (AS3561) in the US, and for some reason Internet MCI's routes *leaked* through provider A, then the AS Path Filters would block AS3561 since they are not explicitly in the AS Path list.

In addition, the AS Path filters will use a format that will only accept routes that originate from the ISP peering to the IXPs. For example, _7597$ will only allow routes which originate in AS7597. Any routes that originate outside of 7597 (for example Internet MCI) will be denied.

Lastly, the AS Path filters will be applied on all ***outbound*** BGP peers for both the IBGP route reflector connections (for those ISPs with out their own ASN) and the EBGP peer connections (for those ISP with their ASN). AS Path filters will also be applied on all EBGP ***inbound*** connections.

Distribute List Filters

Distribute List filters will be used to remove RFC 1918 (Private Address Space), multicast, and other routes that should not be propagated on the Internet routing table (i.e. like 127.0.0.0/16). This safe guard keeps networks link 10.0.0.0/8 from getting leaked out across the Internet.

The following is the BGP ingress/egress route recommended filter:

```
access-list 180 deny   ip host 0.0.0.0 any
access-list 180 deny   ip 127.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 180 deny   ip 10.0.0.0 0.255.255.255 255.0.0.0 0.255.255.255
access-list 180 deny   ip 172.16.0.0 0.15.255.255 255.240.0.0 0.15.255.255
access-list 180 deny   ip 192.168.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 180 deny   ip 192.0.2.0 0.0.0.255 255.255.255.0 0.0.0.255
access-list 180 deny   ip 128.0.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 180 deny   ip 191.255.0.0 0.0.255.255 255.255.0.0 0.0.255.255
access-list 180 deny   ip 192.0.0.0 0.0.0.255 255.255.255.0 0.0.0.255
access-list 180 deny   ip 223.255.255.0 0.0.0.255 255.255.255.0 0.0.0.255
access-list 180 deny   ip 224.0.0.0 31.255.255.255 224.0.0.0 31.255.255.255
access-list 180 permit ip any any
```

Community Filters

BGP Communities are used for a wide variety of reasons. The most common BGP Community filter functions are to enforce what is advertised to the IXP members and to localize an ISP's advertisements on an IXP.

```
ip bgp-community new-format
! Needed so that a community is treated in 16-bit:16-bit format
! rather than one 32-bit integer.
!
      route-map IXP-Routes-In permit 10
       set community 220:99
      !
```

```
      route-map IXP-Routes-Out permit 10
       match community 10
!
  ip community-list 10 deny 220:1
!
      !
      router bgp 220
       neighbor x.x.x.x remote-as ASN
       neighbor x.x.x.x route-map IXP-Routes-In in
       neighbor x.x.x.x route-map No-Upstream-Routes out
       neighbor x.x.x.x send-community



route-map Upstream permit 10
 match community <community-list-no>
!
ip community-list <community-list-no> permit 200:1
!
router bgp 222
 neighbor 200.200.6.1 route-map infilter in
```

For example, a large national ISP has connections to several IXPs with the country – each serving a separate state. The ISP wants to gain the benefits of local exchange of traffic over the IXPs, but they do not want to provide free national transit to small ISPs who only cover the specific state.

To meet it's objective, the ISP uses BGP communities to govern what routes are advertised over the specific IXP. For example, an ISP with a connection to two different IXPs wishes to get the best value from the IXP connection. On IXP is on the East Side of the country and the other on the West Side of the country (see Figure 20). Some peers on the IXPs are true peers who have equivalent national backbones (see Peer A and Peer B in the figure). So the ISP can advertise all their customer routes to these peers.

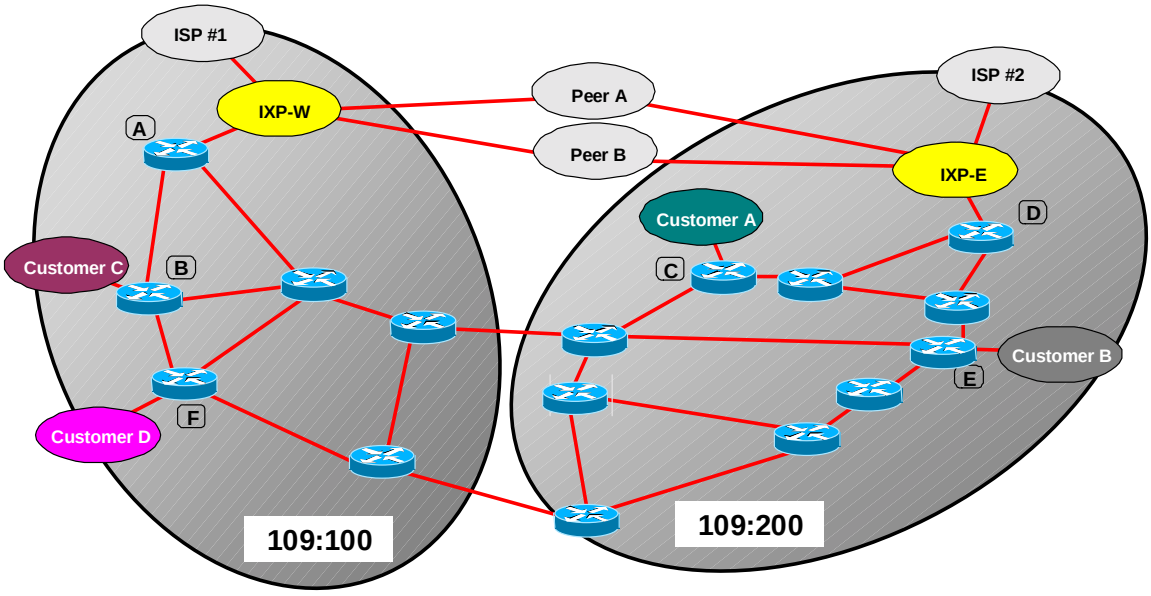

Figure 20 - Using BGP community filtering to have regional peering.

Prefix-List Filters

Packet Filtering and IXPs

IP filtering is an additional policy enforcement tool. Essentially, you place a egress and ingress ip packet filter on the peer connections connecting to the IXP. Every packet coming into or out of the IXP will get checked. Some filters are designed to insure that all packets that leave a ISP belong to that ISP. This called ***ingress filtering***. Other filters check packets coming into a ISP are not "rouge," private, or spoofed addresses. This is called ***egress filtering***. ISPs who connect to an IXP are encouraged to do both.

The following is an example of ISP connecting to an IXP. Egress and ingress ip packet filters are applied. The ingress filters insure that only packets with source address in the ISP's CIDR block(s) are advertised out. The egress filters prevent spoofing, rouge addresses, and private addresses from coming into the ISP.

NOTE: There could be a performance impact on the forwarding speed of the router when a lot of filters are applied. Cisco's newer switching technologies minimize the performance impact of IP filtering. For example, Netflow switching is extremely efficient minimizing the performance impact of very long IP access list. Due care and consideration should be taken when ever the access list start getting beyond 50 entries. Though, it should be stated that trading a few microsecond of IP forwarding vs opening yourself to certain denial of service attacks may prove to be worth it the few microseconds.

Ingress Packet Filtering - Preventing Transmission of Invalid IP Addresses

By filtering packets on your routers that connect your network to the Internet, you can permit only packets with valid source IP addresses to leave your network and get into the Internet. For example, if your network consists of network 165.21.0.0, and your router connects to your ISP using a serial 0/1 interface, you can apply the access-list as follows:

```
access-list 110 permit ip 165.21.0.0 0.0.255.255 any
access-list 110 deny ip any any log

interface serial 0/1
ip access-group 110 out
```

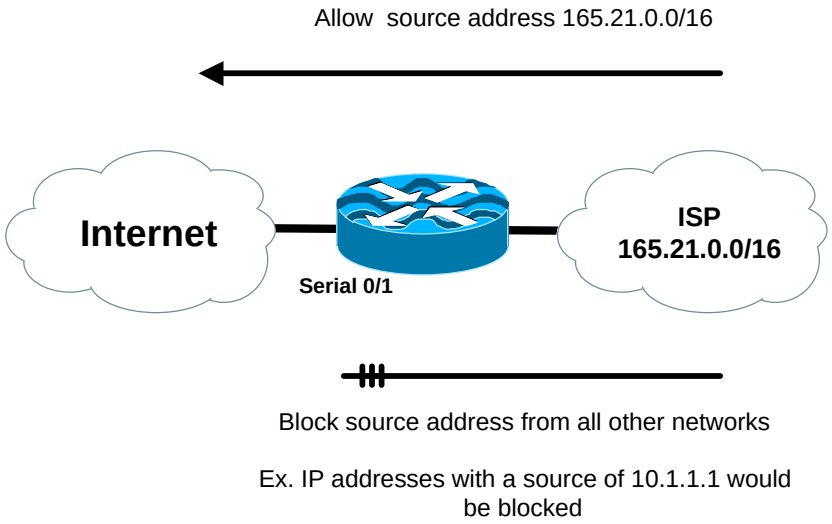

Figure 21 - Ingress Filtering

The last line of the access-list determines if there is any traffic with an invalid source address entering the Internet. If there are any matches, they will be logged. It is not crucial to have this line, but it will help locate the source and extent of the possible attacks.

Egress Packet Filtering - Preventing Reception of Invalid IP Addresses

For ISPs who provide service to end networks, we highly recommend the validation of incoming packets from your clients. This can be accomplished by the use of inbound packet filters on your border routers. For example, if your clients has a network number of 165.21.0.0/16, your should not seen any packets coming into your network with 165.21.0.0 in the source. These packets are attempts at spoofing and should be dropped. The following example shows a sample filter for network 165.21.0.0 with filters for private and rouge routes:

```
access-list 111 deny ip host 0.0.0.0 any log
access-list 111 deny ip 127.0.0.0 0.255.255.255 any log
access-list 111 deny ip 10.0.0.0 0.255.255.255 any log
access-list 111 deny ip 172.16.0.0 0.15.255.255 any log
access-list 111 deny ip 192.168.0.0 0.0.255.255 any log
access-list 111 deny ip 165.21.0.0 0.0.255.255 any log
access-list 111 permit ip any any

interface serial 1/0
ip access-group 111 in
```

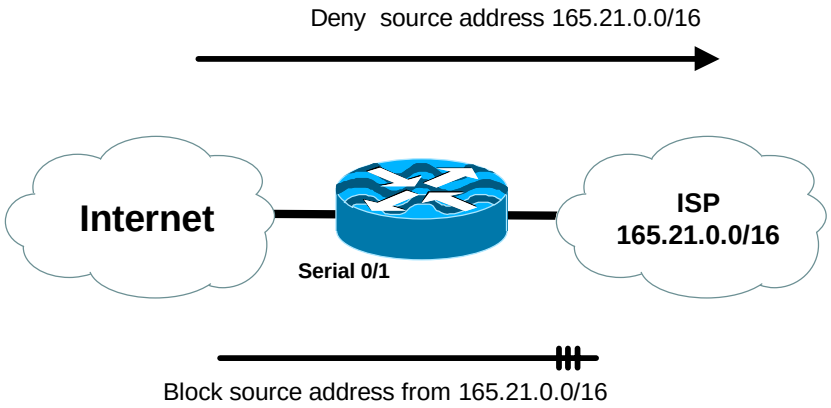

Figure 22 - Egress Filtering

All the "anti spoof," private address, and rouge filters have *log* any matches. If there are any, they would be logged. It is not crucial to have this line, but it will help locate the source and extent of the possible probes or attacks.

Unicast RPF

Standard and Extended ACLs

Turbo ACLs

Committed Access Rate

Putting it all together.

The following is a ISP allocated CIDR block 165.21.0.0/16 with both filters on the Interface:

```
interface Serial 0
 description Connection to Upstream ISP
 ip address XXX.XXX.XXX.XXX 255.255.255.252
 no ip redirects
 no ip directed-broadcast
 no ip proxy-arp
 ip route-cache same-interface
 ip access-group 111 in
 ip access-group 110 out

access-list 110 permit ip 165.21.0.0 0.0.255.255 any
access-list 110 deny ip any any log
access-list 111 deny ip host 0.0.0.0 any log
access-list 111 deny ip 127.0.0.0 0.255.255.255 any log
access-list 111 deny ip 10.0.0.0 0.255.255.255 any log
access-list 111 deny ip 172.16.0.0 0.15.255.255 any log
access-list 111 deny ip 192.168.0.0 0.0.255.255 any log
access-list 111 deny ip 165.21.0.0 0.0.255.255 any log
access-list 111 permit ip any any
```

Where to apply Packet Filtering in a L2 RR IXP

Classification & Coloring Ingress Packets

Network Integration – How to Integration an ISP's Routing Architecture with the L2 Router Reflector IXP

L2 Router Reflector IXPs will have several types of ISPs connecting to the IXP fabric. Some ISPs will have their own Autonomous System Number (AS number) and will peer directly to the IXP's route reflector. Other ISPs will peer directly, but have yet obtained an AS number. These will connect via the IXP's route reflectors but only use the IXP's AS

number on the ISP router – not anywhere else in their network. All of these ISPs will have to integrate BGP into their internal network routing topology. This section's goal is to highlight some of the design and integration options available to the ISP connecting to a L2 Router Reflector IXP.

ISPs with no AS Number

Some ISPs will not have their own ASN. The are connected to the Internet via a lease line to a Network Service Provider (NSP) in the US. Their internal routing could be built with static routes, OSPF, IS-IS, EIGRP, or RIPv2[21]. What follows is one example of a way a ISP can configure their routing protocols.

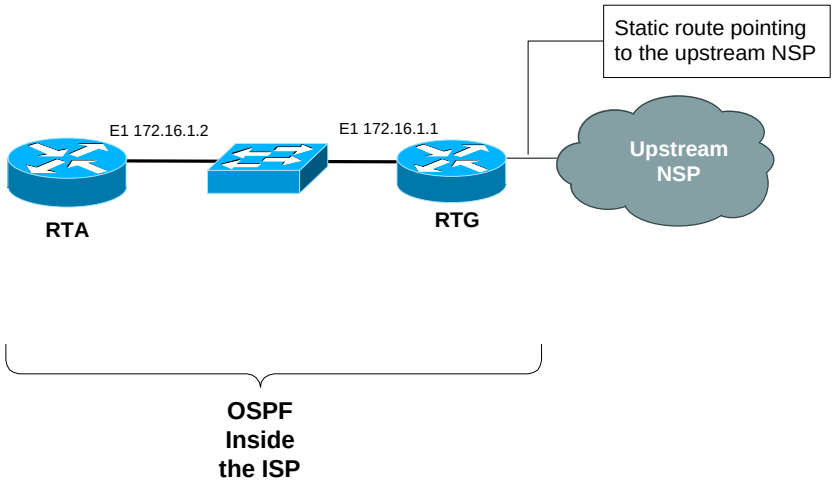

Figure 23 - ISP with no ASN

ISP1 is a typical ISP in their early deployment. They get one connection to an upstream NSP (usually in the US), get allocated a block of addresses - 171.16.0.0/19 from their IPv4 registry, and use OSPF in their internal backbone routing protocol. In this example, the internal backbone is a ethernet switch with 171.16.1.0/24 assigned for devices connecting to the backbone.. A static route pointing the default to their upstream provider is used to provide proper routing within their network. The NSP take their CIDR block and advertises it to the rest of the Internet. The configs on routers RTA and RTG might look something like the following:

RTA

```
interface Ethernet 0
 description Internal Backbone link to RTG

router ospf 1
 network 171.16.1.0 0.0.0.3 area 0
 redistribute static subnets
```

[21] Hopefully no one is still using RIPv1. RIP version 1 is a historical protocol that was replaced by RIP version 2. Since the Internet requires CIDR. All ISPs should be using a CIDR compliant routing protocol.

RTG

```
interface Serial 0                                    router ospf 1
 description Link to Upstream NSP                      network 171.16.1.0 0.0.0.3 area 0
                                                       redistribute static subnets
interface ethernet 0
 description Internal backbone link to RTA            ip     route     0.0.0.0     0.0.0.0     Serial0
```

Config Example 1 - ISP with out an ASN using OSPF as IGP

This simple OSPF config demonstrates how RTG is inserting a static default into OSPF for the rest of ISP1's backbone. If the link to the upstream provider is lost, then the static will drop and ISP1 will loose their default. Traffic will still be routed within their network.

How does this connect to a IXP link iIX? Especially since ISP1 does not have their ASN? Simple. You connect RTA to the IXP and include it in the ASN for iIX. While not optimal, this works until ISP1 gets their own ASN.

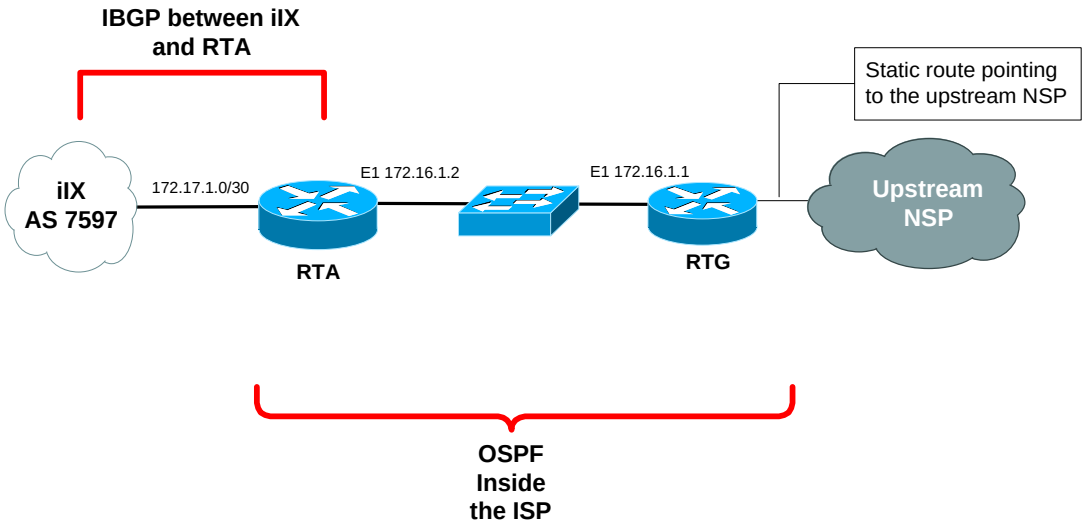

Figure 24 - ISP using Route Reflector to Connect to IXP

In this example, RTA's Serial0 interface is connected to the iIX via a lease line. RTA becomes a member of the iIX ASN, 7597. Routed received from AS7597 are redistributed into OSPF. All other routers in ISP1 would get a list of all the routes from iIX. Hence, all traffic bound for a member of iIX would go out that iIX ↔ RTA link. All other routes would be forward to default.

Here is an example config:

RTA

```
interface serial 0
 Description link to iIX
 p address 172.17.1.1 255.255.255.252

interface Ethernet 0
 description Internal Backbone link to RTG
 ip address 172.16.1.2 255.255.255.0

router ospf 1
 network 171.16.1.0 0.0.0.255 area 0
 network 171.17.1.0 0.0.0.3 area 0
 redistribute static subnets
 redistribute bgp

router bgp 7597
 no synchronization
 bgp dampening
 network 172.16.0.0 mask 255.255.192.0
 neighbor 172.17.1.2 remote-as 7597
 neighbor 172.17.1.2 next-hop-self
 neighbor 172.17.1.2 version 4
 neighbor 172.17.1.2 distribute-list 180 in
 neighbor 172.17.1.2 distribute-list 180 out
 neighbor 172.17.1.2 route-map IIX-IN in
 no auto-summary

ip as-path access-list 1 permit _7597_

route-map IIX-IN permit 10
 match as-path 1

ip route 172.16.0.0 255.255.192.0 Null0
```

RTG

```
interface Serial 0
 description Link to Upstream NSP

interface ethernet 0
 description Internal backbone link to RTA
 ip address 172.16.1.1 255.255.255.0

router ospf 1
 network 171.16.1.0 0.0.0.255 area 0
 network 171.16.0.0 0.0.31.255 area 0
 default-information originate
 redistribute static subnets

ip route 0.0.0.0 0.0.0.0 Serial0
```

Config Example 2 - IBGP (ISP without an ASN)

Locking your BGP Network Advertisements *UP*

Some configuration techniques on the Internet have been developed over time to insure ISPs are being good Internet citizens and minimizing the effect their network has on the rest of the Internet. Locking BGP network advertisements *up* is one of these configuration techniques. Normally, BGP would be dynamic, sensing if a network is up or down and sending update accordingly. Hence, internal changes in your network would be propagate throughout the Internet. This is commonly referred to as route flapping and is not a good thing for the Internet[22].

The commonly used BGP config technique is to *lock* your BGP CIDR advertisements up is to include a status route for the CIRD block pointing to Null0. For example, if you have a BGP network advertisement like:

[22] Route flapping got so bad that a new BGP technique has to be developed to protect - route flap damping. Cisco's IOS has a BGP feature - *bgp dampening*. See the Cisco documents for more details.

```
router bgp 7597
 network 172.16.0.0 mask 255.255.192.0
```

then you would *lock* the route up with a static:

```
ip route 172.16.0.0 255.255.192.0 Null0
```

ISPs with their own AS Number (OSPF as the IGP example)

ISPs who are already using BGP to connect to the Internet should not have any problem connecting to iIX. Due consideration must be paid to the filters used. ISPs should insure that iIX routes are not leaked to the rest of the Internet and routes from the Internet are not leaked into iIX. Figure 25 is a diagram of an ISP (AS 100) with connections to two upstream NSPs (AS 200 & AS 300). EBGP is used to advertise the ISP's route objects to iIX, NSP2 and NSP3. The full Internet routing table is pulled from the two NSPs.

Proactive filtering is a must in this situation. Config Example 3 is one example of BGP filtering that would satisfy the requirements the filtering requirements. Note that the *as-path access-list 3* filters routes that originate in AS 100. This is to prevent routes advertised out one NSP from coming back through the other NSP.

Other filter rules will also work. Refer to the documentation on BGP Regular Expression filters.

RTA

```
interface serial 0
 Description link to iIX
 ip address 172.17.1.1 255.255.255.252

interface Ethernet 0
 description Internal Backbone link to RTG
 ip address 172.16.1.2 255.255.255.0

router ospf 1
 network 171.16.1.0 0.0.0.255 area 0
 network 171.17.1.0 0.0.0.3 area 0
 redistribute static subnets

router bgp 100
 no synchronization
 bgp dampening
 network 172.16.0.0 mask 255.255.192.0
 neighbor 172.16.1.1 remote-as 100
 neighbor 172.16.1.1 version 4
 neighbor 172.17.1.2 remote-as 7597
 neighbor 172.17.1.2 version 4
 neighbor 172.17.1.2 distribute-list 180 in
 neighbor 172.17.1.2 distribute-list 180 out
 neighbor 172.17.1.2 route-map IIX-IN in
 neighbor 172.17.1.2 route-map IIX-OUT out
 no auto-summary

ip as-path access-list 1 permit _7597_
ip as-path access-list 2 permit _100$
ip route 172.16.0.0 255.255.192.0 Null0

route-map IIX-IN permit 10
 match as-path 1

route-map IIX-OUT permit 10
 match as-path 2
```

RTG

```
interface Serial 0
 description Link to Upstream NSP #1
 ip address 192.68.1.1 255.255.255.252

interface Serial 1
 description Link to Upstream NSP #2

interface ethernet 0
 description Internal backbone link to RTA
 ip address 172.16.1.1 255.255.255.0

router ospf 1
 network 171.16.1.0 0.0.0.255 area 0
 network 171.16.0.0 0.0.31.255 area 0
 default-information originate
 redistribute static subnets

router bgp 100
 no synchronization
 bgp dampening
 network 172.16.0.0 mask 255.255.192.0
 neighbor 172.16.1.2 remote-as 100
 neighbor 172.16.1.2 version 4
 neighbor 192.68.1.2 remote-as 200
 neighbor 192.68.1.2  version 4
 neighbor 192.68.1.2 distribute-list 180 in
 neighbor 192.68.1.2 distribute-list 180 out
 neighbor 192.68.1.2 route-map NSP1-IN in
 neighbor 192.68.1.2 route-map NSP1-OUT out
 neighbor 192.68.1.6 remote-as 300
 neighbor 192.68.1.6  version 4
 neighbor 192.68.1.6 distribute-list 180 in
 neighbor 192.68.1.6 distribute-list 180 out
 neighbor 192.68.1.6 route-map NSP2-IN in
 neighbor 192.68.1.6 route-map NSP2-OUT out

 no auto-summary

ip as-path access-list 3 deny _100$
ip as-path access-list 3 permit *.
ip as-path access-list 4 deny _7597_
ip as-path access-list 4 permit _100$
ip route 172.16.0.0 255.255.192.0 Null0

route-map NSP1-IN permit 10
 match as-path 3

route-map NSP1-OUT permit 10
 match as-path 4

route-map NSP2-IN permit 10
 match as-path 3

route-map NSP2-OUT permit 10
 match                as-path                4
```

Config Example 3 – ISPs with their own AS Number

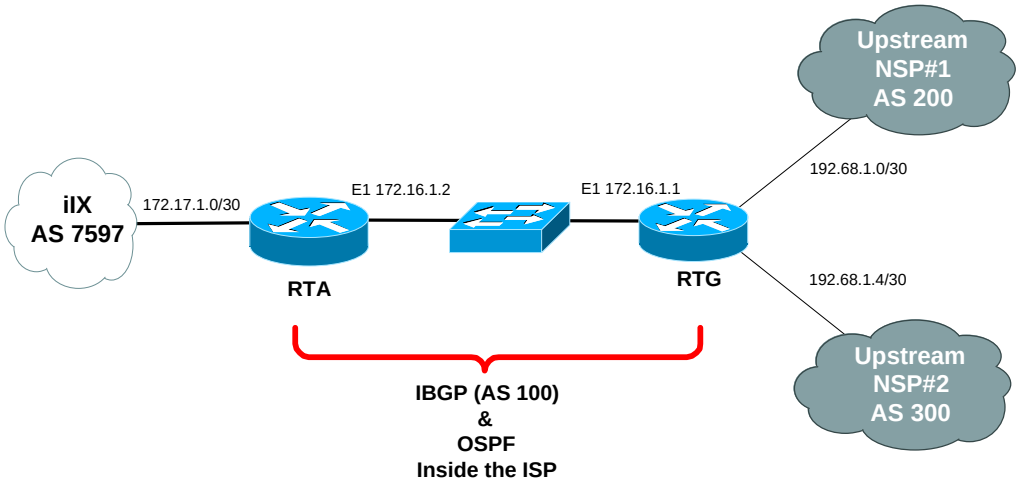

Figure 25 - ISP with eBGP to IXP

FURTHER READING AND REFERENCES

Here is some suggested reading for further reading:

- "Internet Routing Architectures" from New Riders Publishing (Cisco Press). ISBN 1-56205-652-2. Author: Bassam Halabi.
- Using the Border Gateway Protocol for Interdomain Routing, Cisco Connection On-line (CCO) Web site: http://www.cisco.com/univercd/data/doc/cintrnet/ics/icsbgp4.htm

Addendum 1 – BGP Route Reflectors

BGP requires that all of the iBGP speakers be fully meshed. However, this requirement does not scale when there are many iBGP speakers. As the number of iBGP speakers increase, N-1 mesh increases a way to reduce the load and complexity of a iBGP mesh is to configure a route reflector.

Figure 26 illustrates a simple iBGP configuration with three iBGP speakers (Routers A, B, and C). Without route reflectors, when Router A receives a route from an external neighbor, it must advertise it to both Routers B and C. Routers B and C do not readvertise the iBGP learned prefix to other iBGP speakers because the routers do not pass routes learned from internal neighbors on to other internal neighbors, thus preventing a routing information loop.

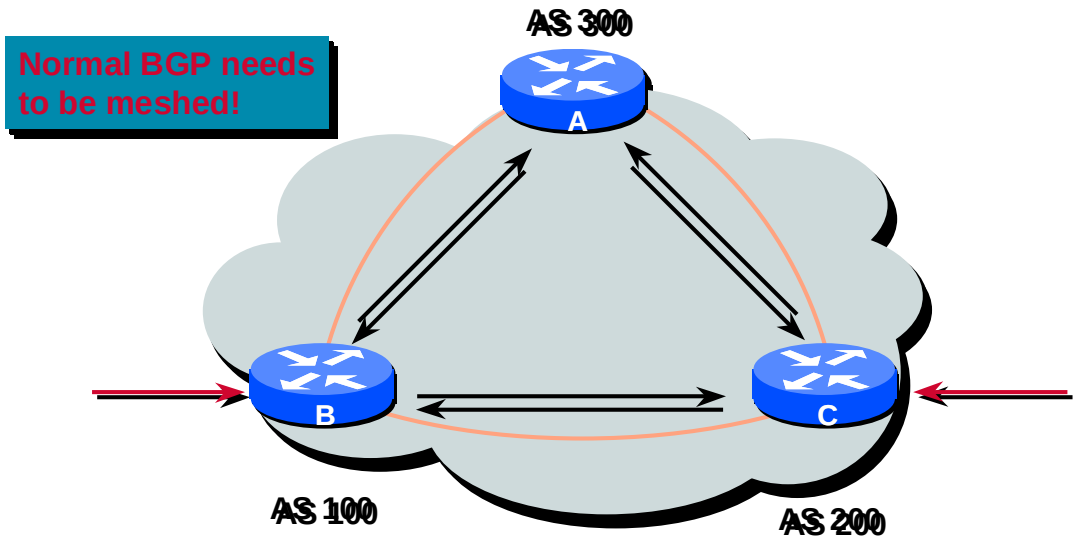

Figure 26 – Three fully meshed iBGP Speakers.

With route reflectors, all iBGP speakers need not be fully meshed because there is a method to pass learned routes to neighbors. In this model, an internal BGP peer is configured to be a *route reflector* responsible for passing iBGP learned route to a set of iBGP neighbors configured as *route reflector clients*. In Figure 27, Router B is configured as a route reflector. When the route reflector receives routes advertised from Router A, it advertises them to Router C, and vice versa. This scheme eliminates the need for the iBGP session between Routers A and C.

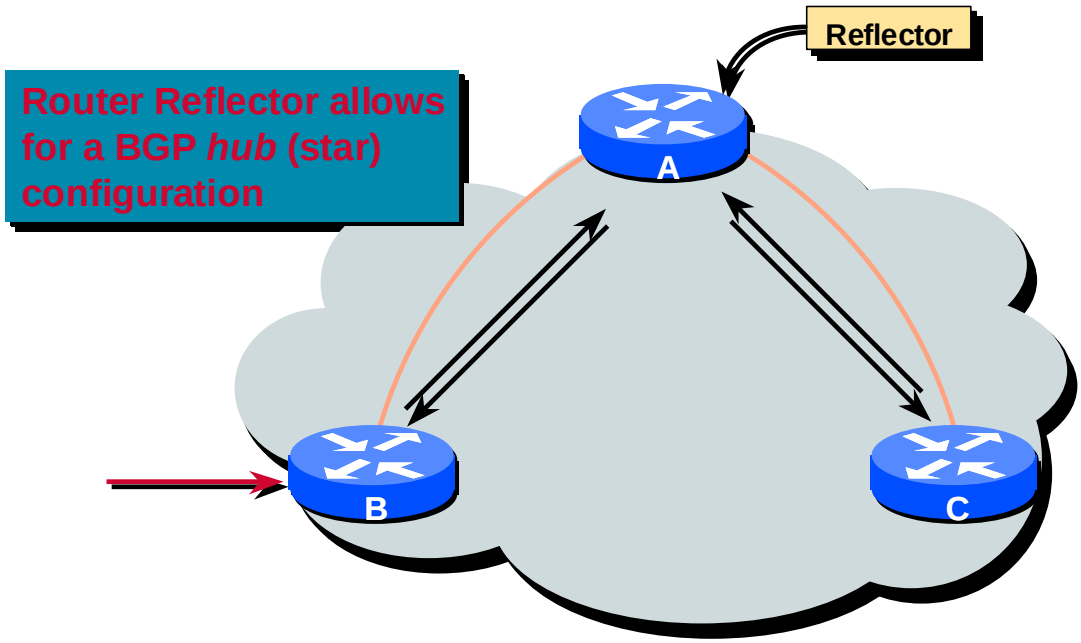

Figure 27 - Simple iBGP Model with Router Reflector

The internal peers of the route reflector are divided into two groups: *route reflector clients* and all the other iBGP speakers in the autonomous system (non-clients). A route reflector reflects routes between these two groups. The route reflector and its clients form a *route reflector cluster*. The non-clients must be fully meshed with each other, but the route reflector clients need not be fully meshed. The clients in the cluster do not communicate with iBGP speakers outside their cluster.

Figure 28 illustrates a more complex route reflector scheme. Routers A, B, and C a fully meshed in the *iBGP core*. Each are route reflectors with three router reflector clients in their clusters. This technique allows for logical hierarchy to be built in the iBGP topology.

When the route reflector receives an advertised route, depending on the neighbor, it does the following:

- Routes from an external eBGP or iBGP speaker are advertised to all route reflector clients and iBGP peers.
- Routes from an iBGP peer are advertised to all route reflector clients.
- Routes from a route reflector client are advertised to all clients and iBGP peers. Hence, the clients need not be fully meshed.

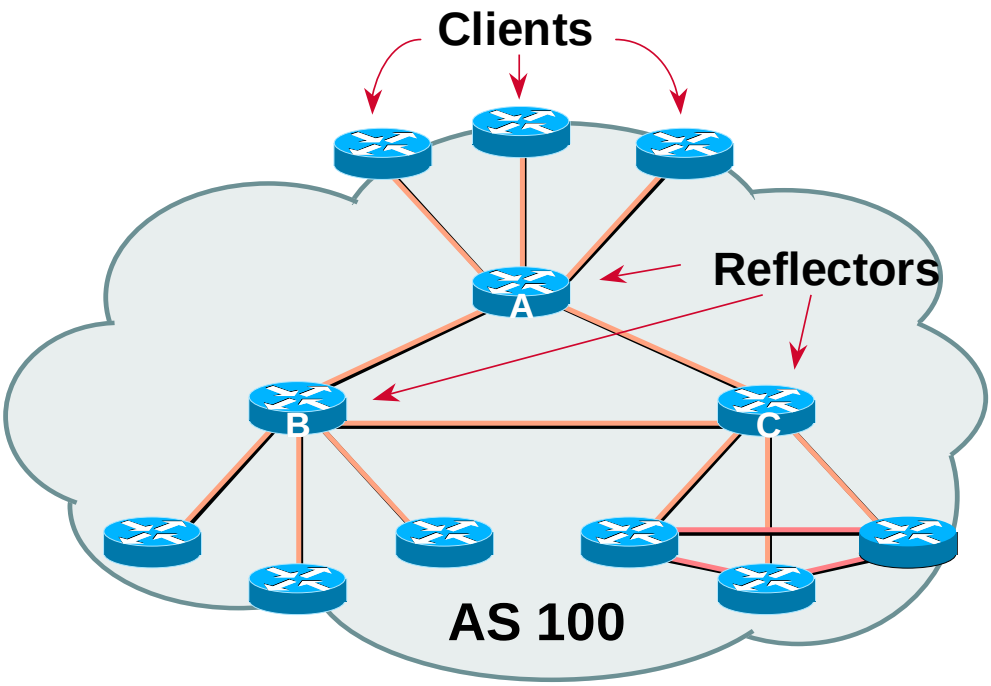

Figure 28 - Router Reflectors Build Hierarchy

To configure a route reflector and its clients, use the following command in router configuration mode:

Command	**Purpose**
neighbor *ip-address* route-reflector-client	Configure the local router as a BGP route reflector and the specified neighbor as a client.

Usually a route reflector cluster will have a single route reflector. In that case, the router ID of the route reflector identifies the cluster. To increase redundancy and avoid a single point of failure, a cluster might have more than one route reflector. In this case, all route reflectors in the cluster must be configured with the 4-byte cluster ID so that a route reflector can recognize updates from route reflectors in the same cluster. All the route reflectors serving a cluster should be fully meshed and all of them should have identical sets of route reflector clients and non-client iBGP peers.

If the cluster has more than one route reflector, configure the cluster ID by using the following command in router configuration mode:

Command	**Purpose**
bgp cluster-id *cluster-id*	Configure the router reflector cluster ID

Use the show ip bgp command to display the originator ID and the cluster-list attributes.

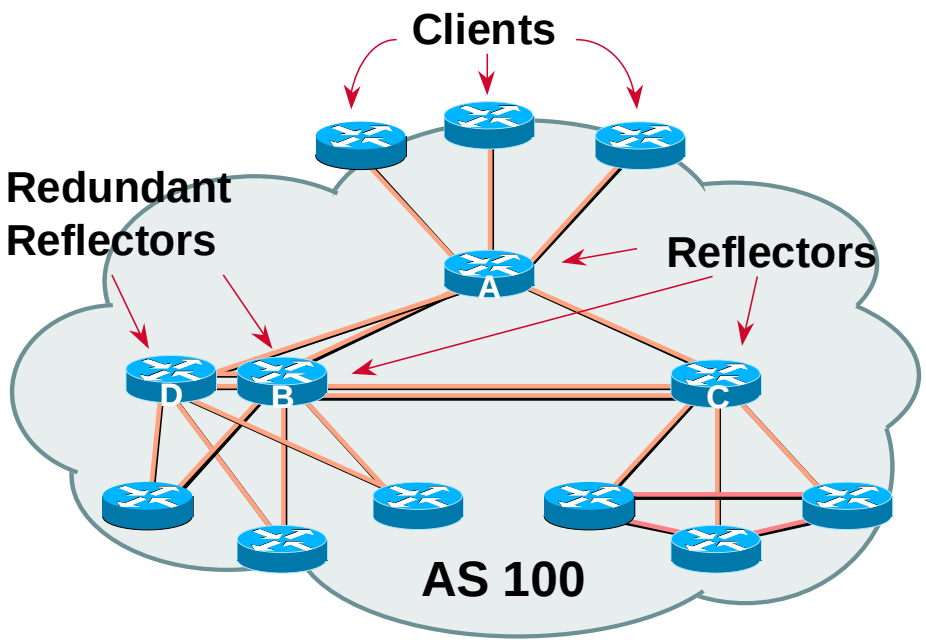

Figure 29 - Router Reflector Redundancy

By default, the clients of a route reflector are not required to be fully meshed and the routes from a client are reflected to other clients. However, if the clients are fully meshed, the route reflector does not need to reflect routes to clients. To disable client-to-client route reflection, use the following command in router configuration mode:

Command	Purpose
no bgp client-to-client reflection	Disable client-to-client route reflection in a fully meshed router reflector cluster.

Note: If client-to-client reflection is enabled, the clients of a route reflector cannot be members of a peer group.

As the iBGP learned routes are reflected, it is possible for routing information to loop. The route reflector model has the following mechanisms to avoid routing loops:

- *Originator-ID* is an optional, nontransitive BGP attribute. This is a 4-byte attributed created by a route reflector. The attribute carries the router ID of the originator of the route in the local autonomous system. Therefore, if a misconfiguration causes routing information to come back to the originator, the information is ignored.

- *Cluster-list* is an optional, nontransitive BGP attribute. It is a sequence of cluster IDs that the route has passed. When a route reflector reflects a route from its clients to non-client iBGP peers, it appends the local cluster ID to the cluster-list. If the cluster-list is empty, it creates a new one. Using this attribute, a route reflector can identify if routing information is looped back to the same cluster due to misconfiguration. If the local cluster ID is found in the cluster-list, the advertisement is ignored.

- Using set clauses in outbound route maps modifies attributes, possibly creating routing loops. To avoid this, set clauses of outbound route maps are ignored for routes reflected to iBGP peers.

Addendum 2 – Case Study of Hong Kong Internet eXchange (HKIX)

This addendum is writing in cooperation with Mr. Che-Hoo Cheng [chcheng@cuhk.edu.hk]. It is an update to a paper submitted to INET 96. Specific questions on HKIX should be forwarded to HKIX – http://www.hkix.net.

Internet eXchange for Local Traffic: Hong Kong's Experience

Che-Hoo Cheng <chcheng@cuhk.edu.hk>
Head, Data Communications and Networking Section
Information Technology Service Unit
Chinese University of Hong Kong
Shatin, N.T., Hong Kong
Tel: +852 2609-8848
Fax: +852 2603-5001
http://www.cuhk.edu.hk/hkix/
http://www.hkix.net

Abstract

Most people think that the United States is the "backbone" of the Internet. Many ISPs in other countries choose to connect to the Internet by leasing circuits to the United States. With this "star topology," local traffic within a country or a city may have to be routed through the United States if there are no local connections among local ISPs. This is highly undesirable because the long-distance circuits are very expensive and are often of relatively slow speed.

The situation in Hong Kong may not be unique. There are more than 40 ISPs in this small city, more than 10 of which have their own links to overseas, mainly the United States. On the other hand, local circuits are quite affordable because of the short distances involved, and this can help the development of local high-speed connections. In view of this, The Chinese University of Hong Kong (CUHK) made use of its own resources to set up a neutral interconnection point called Hong Kong Internet eXchange (HKIX), mainly for the routing of intra-Hong Kong traffic. This has proved to be very successful. As of late March 1996, 26 commercial ISPs are connected. Among the 26, there are even regional or global ISPs such as AT&T, IBM Global Network and Global SprintLink.

In this paper, the key reasons for the success of HKIX are presented. Other issues, including the history, the technical aspects, the problems, the funding issue and the future of HKIX, are also addressed.

Introduction

There is no doubt that the Internet is blooming in a very fast pace around the world. Many organizations and individuals are rushing to the Internet. It is now very difficult to find a single well-known multinational company that does not have a presence on the Internet.

Although the Internet is a worldwide computer network, many people still think that the United States is the "backbone" of the Internet. In fact, more than 60 percent of the users and nodes of the Internet are located inside the United States. And the United States carries most of the intercontinental traffic. Because the Internet is a loosely controlled network, the connections among different networks are often arbitrary, at least from an outsiders' point of view. In fact, nobody in the world can have the exact configuration of who is connecting to whom. Many ISPs in other countries choose to connect to the Internet by leasing circuits to the United States. Although intraregional network connections are gradually being set up, this has not occurred to much extent in the Asia Pacific region. Of course, the main reason for this is because of the high cost of intraregional leased circuits and the low volume of traffic among the countries within the region. But with this "star topology," local traffic within a country or a city may have to be routed through the United States if there are no local connections among local ISPs. This is highly undesirable because the long-distance circuits are very expensive. Therefore, setting up Internet eXchange (IX) for the routing of local traffic surely can benefit all parties involved. But how can arrangements for interconnecting local ISPs be made, especially when there are many ISPs involved? Many unsuccessful stories have been heard in some regions and countries. The reasons are varied, but it is believed that a lot of politics and competition are involved in most of the cases.

Internet development in Hong Kong

The first 64kbps Internet link in Hong Kong was set up in September 1991 by The Chinese University of Hong Kong (CUHK). The link started to be shared by all higher educational institutions in Hong Kong in 1992. The network linking up all institutions is called HARNET (Hongkong Academic and Research NETwork). In September 1993, the link was upgraded to 128kbps after some delay. Before then, there were very few commercial elements within the Internet community in Hong Kong. In late 1993, two small commercial Internet service providers (ISPs) were set up with their own 64kbps links to the United States, and they started to change the status quo. As one of them (HK Supernet) was a spin-off of a university (The Hong Kong University of Science and Technology), it had a direct connection to HARNET from the very beginning. The other one (HKIGS) was a small independent ISP, so it had no connections to the other two networks. At that time, there was very little need to have full interconnections because the gold mine was at the other side of the ocean and there was little value to connect to local sites. So people could live with that situation then.

In 1994, Internet on the nonacademic side continued to grow. A number of small ISPs were set up to serve mostly individual subscribers. The new ones did not have their own links to the United States. Instead, they were just piggybacked to HKIGS. But the growth rate was still not very high. The blooming of Internet development in Hong Kong in fact happened in 1995. In February and March of that year, two incidences (mostly related to the licensing requirements of operating an ISP in Hong Kong) caused the close-down of quite a number of ISPs, but at the same time awakened the general public to the Internet. After the licensing issue was clarified, most ISPs were back in operation. At the same time, many ISPs were entering the market. Some were even established by listed companies or large multinational companies. Many companies in Hong Kong, no matter how large or how small, started to consider having a presence on the Internet. On the other hand, students, professionals and computer hobbyists are rushing to join the Internet. This started to create more and more local content and importance, and intra-Hong Kong communications became more and more needed.

At the end of 1995, more than 40 ISPs were doing business. Quite a number of them, especially the larger ones, have direct links to overseas. Others were just piggyback ISPs relying on other ISPs that had local presence to do transit for them.

Timeline of Internet Development in Hong Kong

Sep 1991:	CUHK set up a 64Kbps Internet link to US.
Early 1992:	Other Universities joined.
Jul 1992:	JUCC/HARNET took up the management.
Late 1992:	HARNET T1-Ring Backbone was set up.
Sep 1993:	HARNET-Internet link upgraded to 128Kbps
Late 1993:	2 commercial ISPs (HK Supernet and HKIGS) were set up with their own 64Kbps links to US.
1994:	A few piggy-back ISPs were set up under
1995:	More ISPs were being set up. Some had their own links to US.
Feb/Mar 1995:	Two incidences caused temporary shutdown of several ISPs.
Apr 1995:	**HKIX was set up by CSC of CUHK.**
Sep 1995:	HARNET-Internet link upgraded to T1
Oct 1995:	HARNET T1-Ring Backbone converted to T1-Star
1996:	More ISPs being set up
Mar 1997:	HARNET-Internet link upgraded to E1
Dec 1997:	First T3 link from HK ISP to US; HARNET Backbone Converted to ATM; HARNET-Internet link upgraded to 6Mbps
May 1998:	Second T3 link from HK ISP to US
Sep 1998:	HARNET-Internet link upgraded to 12Mbps
Feb 2000:	First 155Mbps link from HK ISP to US
Mar 2000:	**159 PNETS-ISP licensees; >1.7M dial-up users; >7.5K leased-line users; Total bandwidth to overseas > 750Mbps**

Setting up of HKIX by CUHK

As mentioned above, there was only one interconnection initially. In September 1994, HKIGS set up a local T1 circuit to CUHK allowing their customers including those of their downstream ISPs to have more direct and faster communications with HARNET. HKIGS was in charge of the rental of the local T1 circuit (around the United States, $1,200/month then), and CUHK provided the router port for the connection. As can be seen, all parties involved gained benefits. Still, HK Supernet and HKIGS were not connected locally in any way because HARNET could not do transit for them.

Things started to change dramatically in 1995. With many more ISPs entering the business, they needed efficient network infrastructure very much in order to lower the cost of operations for better competitiveness. They could not afford to route intra-Hong Kong traffic overseas because their overseas links were expensive and relatively slow. Although all of them wanted to set up interconnections, they could not do it easily, mostly because they are all competitors to one another and it was impossible to have most of them come together and discuss interconnection. Having full-mesh interconnections among them was out of the question at that time. In view of this, and having the precedence of connecting to HKIGS, CUHK saw the need to do something for Internet development in Hong Kong again and set up the framework of HKIX (see Figure 1). CUHK started to negotiate with newly established local ISPs that had direct links to overseas. Most of them agreed with the idea very much and committed to order circuits to CUHK immediately.

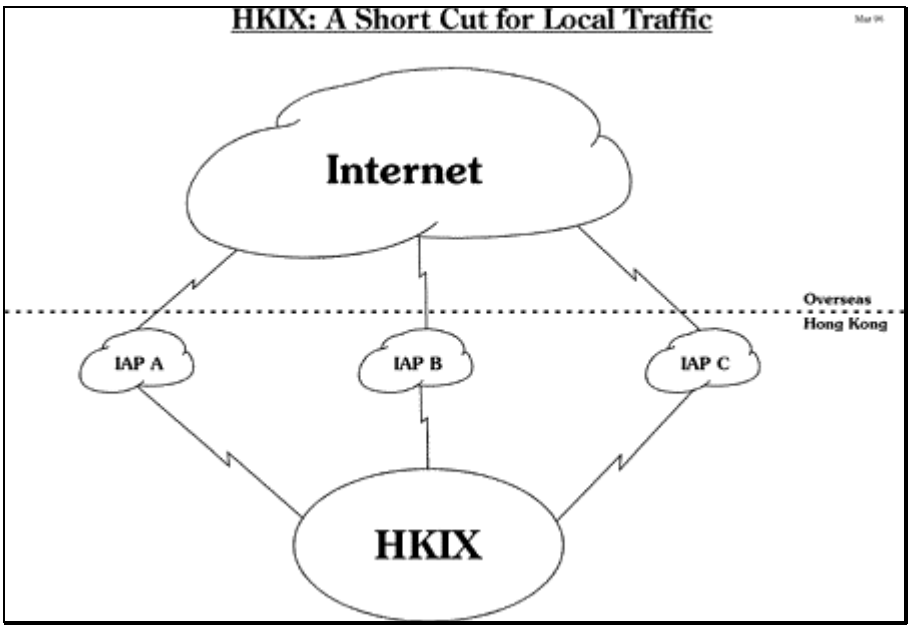

Figure 30 - Macro View of HKIX

After all the hard work, the first two HKIX connections (LinkAGE Online and Global Link) were set up in April 1995. The new arrangement was that the leased circuits to CUHK and the routers placed here were provided by the participating ISPs. And CUHK was responsible for providing space, electricity, air conditioning and a simple Ethernet network to connecting all routers of the participants. HKIGS was migrated to this new HKIX connection a little bit later after its router was delivered to CUHK. HK Supernet was connected to HKIX in October 1995. (For the current list of participants, please refer to http://www.hkix.net/hkix/connected.html.) As of late May 2000 1996, over 60 commercial ISPs were connected (including some regional and global ISPs such as GlobalOne, AT&T UUNET) and four Fixed Telecommunication Network Services (FTNS) operators which have fiber co-located to the site of HKIX.

Technical aspects of HKIX

HKIX is a layer 2 interconnection point. The physical part of the HKIX is very simple. Each participant leases a dedicated circuit to CUHK and places a router here. CUHK provides the Ethernet connections among the routers. The minimum speed of the leased circuit is 512kbps. By July 1996, the minimum speed will be T1 (1.536Mbps). Initially, the HKIX Ethernet was just a coaxial cable segment. After more than 10 ISPs were connected, the shared Ethernet became quite congested, so the coaxial cable was replaced by a Cisco/Kalpana EtherSwitch PRO16 in December 1995 (see Figure 31).

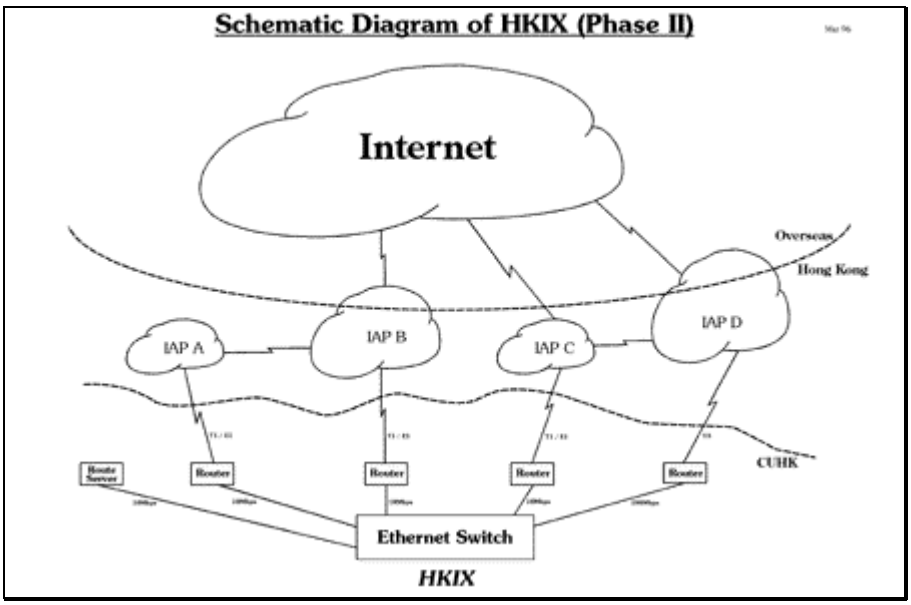

Figure 31 - HKIX Phase II

As of late March 1996, the highest speed to HKIX is dual T1. One T3 connection is on order. In order to cater for better support of direct ATM connections, Cisco Catalyst 5000 will be used to replace the PRO16 in April 1996. Initially, the first T3 connection will be supported by the 100Mbps Ethernet port of the switch. Later when ATM switch is installed, this T3 connection and other new T3 connections will be migrated to ATM (see Figure 32).

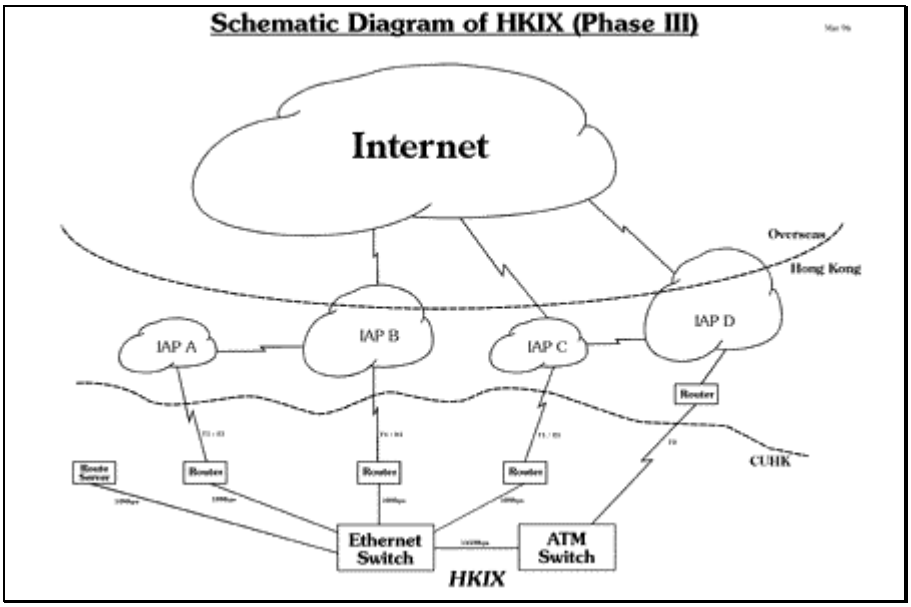

Figure 32 - HKIX Phase III

The initial routing setup is also simple if the participants have enough knowledge of Internet routing, especially BGP4. As it is desirable to have mandatory multilateral peering agreement to ensure greatest possible benefits to all, a router server (Cisco 2501) is used to provide a single view of routing for all participants. Each router on HKIX belongs to the same

autonomous system (AS) as the corresponding ISP. It must peer with the route server using BGP4 via the HKIX Ethernet and announce all routes of the internal networks and downstreams. It must also accept all routes distributed by the route server.

This arrangement aligns with the philosophy that everyone on HKIX is equal and will be treated fairly. It also makes sure that intra-Hong Kong traffic is routed through the fastest possible path and all participants can gain the greatest possible benefits.

The current way to control routing is by the IP network access lists on the route server, which filter incoming BGP routes. As more participants join HKIX, this method does not scale. Materials of Routing Arbiter Project are being studied to see if RA database and programs suit the purpose and are easily implemented and managed.

As for fully utilizing the connections to HKIX, all participants are encouraged to have more cooperation through the HKIX, e.g., news exchange and Domain Name System (DNS) backup. Although not much has been done so far except news exchange, it is seen that cooperation among ISPs is gradually increasing, although slowly.

Another important philosophy is that all participants must have global Internet connectivity independent of HKIX facilities. As the initial setup was very simple and the lowest possible investment was made for earliest possible establishment, this philosophy was thought out to make sure that HKIX was not used as the sole connection to the outside world for better reliability. Although the philosophy is still valid, it is encouraged that participants seek backup global Internet connectivity via HKIX.

Figure 33 - View of HKIX's Equipment Racks

The success of HKIX

There is no doubt that the HKIX project is extremely successful. As a matter of fact, nobody could imagine this at the very beginning. To look back, the reasons may include the following:

Operated by a relatively neutral - not for profit party. CUHK is perceived as a neutral party because it does not compete with the participants.
Low setup cost and simple configuration. The investment of each participant on HKIX is only a little when compared with its other investment, but the benefits can be a lot.
Mandatory multilateral peering agreement and no settlement for routing of local traffic. This policy makes all parties have the greatest possible benefits from the connections.
Need of highly efficient network infrastructure. All ISPs want to lower their cost of operations in order to survive in this highly competitive industry. This HKIX project has helped them a lot on this.
Technical competence of CUHK. Because CUHK set up the first Internet link and has led many major Internet developments in Hong Kong, its technical skill is trusted by the participants. Fortunately, CUHK has not let them down.

Requirements for ISPs to join HKIX

The requirements for ISPs to join HKIX are as follows:

1. Internet Service Providers (ISPs). They must be offering Internet access services to their customers.
2. They must has a PNETS-ISP License.
3. They must have global Internet connectivity independent of HKIX facilities.
4. Self sufficient
5. They must be self-sufficient. For example, they must have their own primary DNS, e-mail, WWW and news servers.
6. Run BGPv4 to exchange routing information.
7. Have a globally unique Autonomous System Number (AS Number)
8. Have an IPv4 address block of a /24 or higher
9. T1 or above to HKIX
10. Provide necessary router, CSU/DSU, and circuit.

ISPs must fulfill all these requirements in order to join HKIX.

Problems

Of course, HKIX is not without problems. When a customer of ISP A switches to ISP B, we must be very careful about the modification of access lists because this will affect the reachability of that customer on HKIX. We sometimes need to act as a mediator when the customer holds an IP network address sub-block of ISP A and wants to use it from ISP B. Fortunately, all such incidences so far have been handled without many difficulties.

Another problem is that many network engineers of local ISPs do not have experience with Internet routing. They often use RIP as their interior gateway protocol (IGP) and do not know CIDR. We have spent significant amounts of time to deal with them when setting up the HKIX routing using BGP4. But with more books and documents available to help them to understand Internet routing, CIDR and BGP4, they seem to have caught up quite a bit.

The most critical problem is that HKIX is still a project of CUHK starting from the very beginning. The project was initiated by CUHK purely as a community service to Hong Kong. Everything is done with the goodwill of CUHK. The policy is set up by CUHK with reference to the policies of other exchanges and after consultation with participants and other experts in this area. Anyway, the final say is still in the hands of CUHK. Although we have tried our best to act fairly and make decisions largely based on the amount of benefits to the whole community, not just to individual parties, complaints and challenges from some ISPs are received from time to time. It seems that whenever they see a potential threat to their business, they will stand up and fight against it. They consider HKIX as a monopoly and try to hold control of it because they fear that HKIX will turn into another major competitor of theirs and grab significant market share very easily because of the name of HKIX and CUHK. But at the same time, some other ISPs want us to maintain the control because they think we are more neutral than anyone else. So we are somehow facing pressure from both sides.

With the issue of whether to support the piggyback ISPs, we faced a lot of challenges from some first-tier ISPs that have they own links to overseas. (Initially, HKIX served first-tier ISPs only.) They feared that if HKIX supports those piggyback ISPs, HKIX would expand to take a too important role in Hong Kong and could not "resist the temptation" to turn into a major transit provider that would immediately have those piggyback ISPs as customers. On the other hand, we were facing pressure from some newly established piggyback ISPs and some large transit providers to open up the HKIX for those piggyback ISPs. Finally, after an unofficial opinion poll and consultation with some neutral experts and individuals, it was decided to "open up" the HKIX. Still, other requirements for ISPs to join HKIX were not changed.

In the long run, a more formal organization has to be set up to oversee the management of HKIX. Of course, the membership should include all the HKIX participants so as to ensure that the interests of all parties can be taken care of. Besides being in charge of the operations of HKIX, CUHK probably will still play a major role on the management side, especially when there are major arguments and issues that cannot be resolved among the members. CUHK will still care moreabout the benefits of the whole community.

Funding issue

Another problem of HKIX is funding. HKIX is totally "funded" by CUHK so far. The resources used are all provided by CUHK, including staffing (part-time work only), network equipment, Ethernet cables, space, electricity and air conditioning. Although the participants need to provide their own leased circuits and router equipment, they do not need to pay any charges to CUHK for the service. In return, they receive only best-effort service from us. As the scale becomes larger and larger and HKIX becomes more and more important, we need dedicated staff to handle all the work. If we cannot have dedicated staff, the reliability of HKIX will become questionable. On the other hand, as HKIX migrates to more sophisticated infrastructure being built up with ATM switch and accompanying Ethernet switch, more resources are needed for network equipment, the expenses of which cannot easily be absorbed by CUHK.

The most logical way to obtain funding is to implement charging for services to cover the cost of operations. The other way is to request support from the government. As of late March 1996, the chances of getting funding from the government are quite high. If the funding is approved, the cost of operations of HKIX can be covered for two more years. After these two years, HKIX probably will need to implement charging in order to sustain operations.

Conclusion

When Internet grows to a state that is out of control of everybody, it may be time to do something to make it more controllable so that the growth can be sustained. Setting up one and only one local IX for the routing of local traffic is one thing that should be done in every major country and city that has a significant amount of Internet users and information content. If local IX is set up, the growth of long-distance bandwidth can be slowed down a little bit. It surely can help the Internet to develop more healthily.

But in a highly competitive community such as Hong Kong, setting up such important infrastructure as HKIX can hardly be achieved by getting all parties involved together and having everything agreed upon by all before setting it up. Everyone is everyone's competitor, so total agreement can hardly be reached. Doing it with goodwill by a relative neutral party may be the most effective way to implement it. After everything is built up and running smoothly, it may be time to hand the management over to the participants. But still, the local IX should be operated at a neutral point in order for it to survive.