NOCTION
NETWORK INTELLIGENCE

Multihoming

A Complete Step-by-Step Guide

www.noction.com

Table of Contents

Introduction

When an e-shop's website goes down, their customers can't buy anything, so the business doesn't make any money. For most other organizations, being disconnected from the internet isn't quite that catastrophic. Or is it? A decade ago, most organizations hosted their own email servers and intranet locally within their own building. These days, more and more services are "in the cloud". So now both the servers in the datacenter and the users must have a working internet connection for the service to be used. If either of those connections goes down, organizations quickly find all kinds of functions grinding to a halt.

So how does an organization protect itself against being disconnected from the internet? An obvious first start is to buy better quality of everything: better routers, switches, cables; service with a better service level agreement (SLA). A healthy dose of contingency planning also helps a lot. For instance, basements are susceptible to flooding, so maybe that's not the best place to put equipment. Firewalls and switches can be duplicated and operated in "hot standby" mode to some degree: if one goes down, another one quickly takes over. But with all of that taken care of, there's still the physical internet connection.

In this guide we're going to discuss having more than one connection to the internet, a practice called multihoming.

Multiple physical connections to one ISP

Connecting to one ISP over multiple independent circuits offers protection against interrupted cables, and to some degree, against failing equipment. When communication moved to fiber, technologies such as SONET/SDH and FDDI allowed for fiber rings with built-in “protection” mechanisms.

Under normal circumstances, all data flows over the primary ring in one direction. When there is a cable cut, the stations on both sides of the cut reroute over the backup ring so all stations remain reachable. The downside of these fiber protection systems is that the capacity of the second ring remains unused. More modern systems, such Resilient Packet Rings (IEEE 802.17) allow for the full use of the available bandwidth.

However, today it’s much more common to use Ethernet, both within a datacenter and over longer distances.

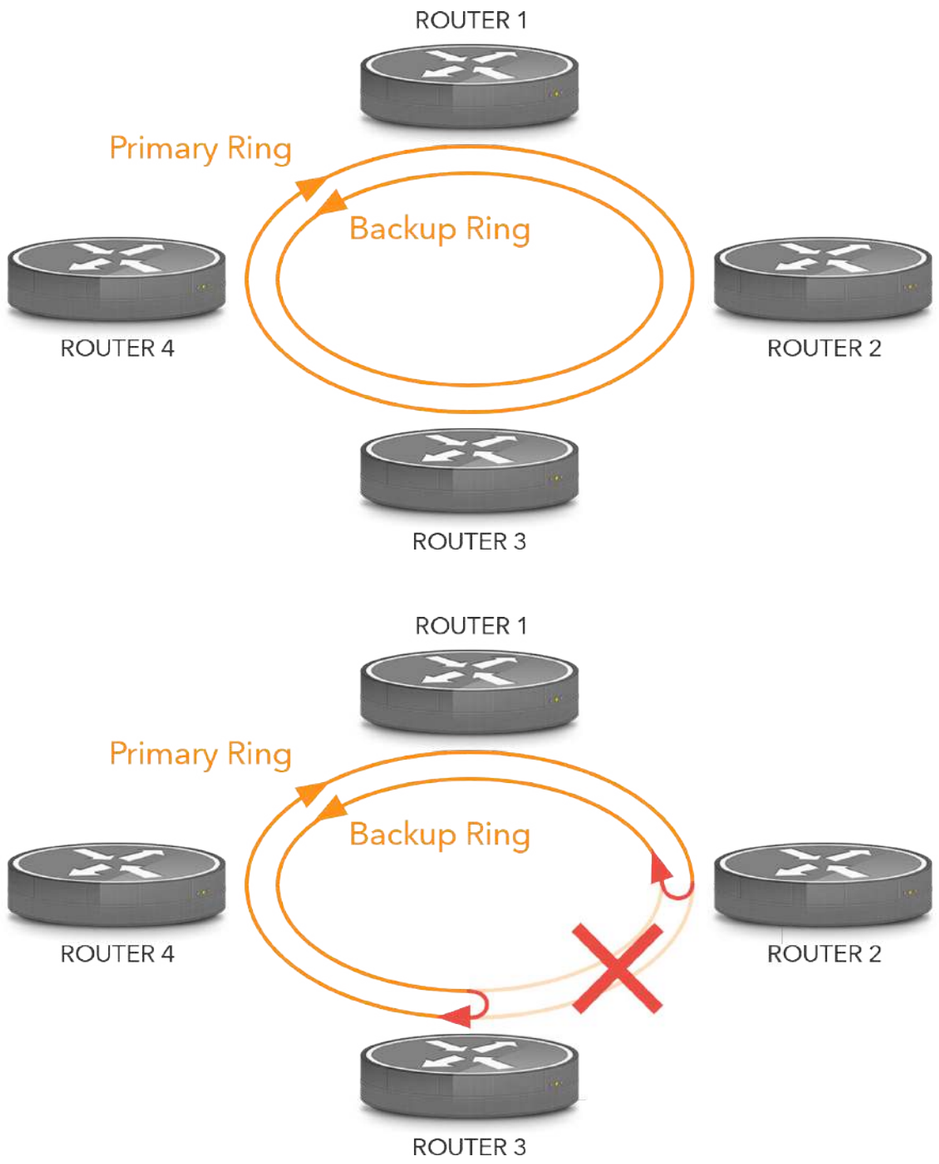

Figure 1: A fiber ring.

Figure 2 shows the simplest way to use two connections towards one ISP: simply have them both connect to the same router. This protects against cable failures, but the single router on the customer side is still a single point of failure.

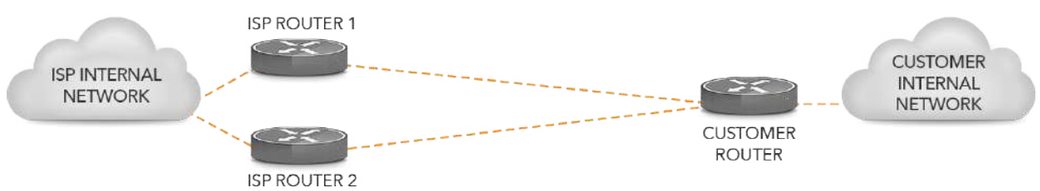

Figure 2: Two connections terminating on one router.

Even worse is the situation in ***Figure 3*** with a switch between the two connections and the router (perhaps because the router doesn't have enough high speed ports) there are now two single points of failure: the router and the switch. If either of those fails, both connections go down.

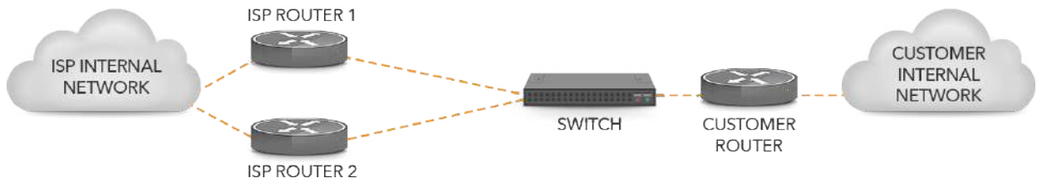

Figure 3: Two connections terminating on one router through a switch.

In ***Figure 4***, there is no longer a single point of failure: there are two routers on the ISP side as well as two routers on the customer side, with separate circuits connecting them.

Figure 4: Two connections terminating on two routers.

In the setup in ***Figure 5*** switches are put in front of the routers. Through the switch, each customer router can talk to both of the ISP routers. In this setup, there is again no single point of failure. The reason some networks use this setup is that it also provides protection against the situation where router 1t on the ISP side and router 2 on the customer side both fail at the same time. In the situation in ***Figure 4***, this would take both connections down. But in the situation in ***Figure 5***, communication is then still possible from ISP router 2 to switch 2 to customer router 1.

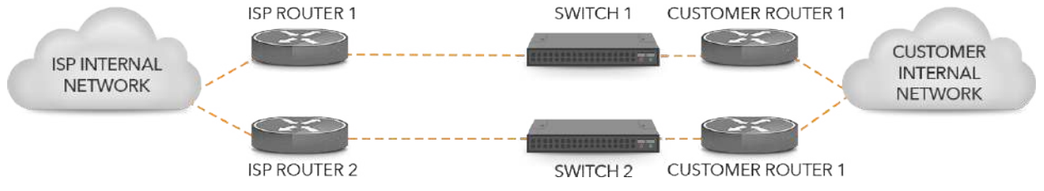

Figure 5: Two connections terminating on two routers through switches.

However, the downside of the ***Figure 5*** setup is that it isolates the customer routers from the connections. So if the circuits go down, the routers don't detect this and they will continue to send packets until the routing protocol that's used (usually BGP) determines that the connection is down. This takes much longer than simply observing a link down event on a physical circuit, and all this time packets disappear into a black hole.

NOTE: All else being equal, it's preferred to connect circuits to an ISP directly to your BGP router without switches in-between so the router can immediately reroute traffic when it sees the link go down.

Routing over multiple connections to one ISP

In the ***Figure 4*** situation, all four routers are in the position to determine if the connection is up or down, as long the connections reliably provide this feedback. For instance, this is the case with a direct Ethernet UTP or fiber link. In that situation, it's possible for the ISP to statically route the address blocks of the customer towards the interface that connects to the customer, and the customer sets a default route towards the interface that connects to the ISP. The routers on both sides then redistribute those static routes into their internal routing protocol, but those static routes will disappear if the interface in question goes down so traffic is rerouted over the other connection.

However, in most cases a routing protocol will be used between the ISP and the customer. If there is link up/down feedback available, using a routing protocol provides an extra level of protection against failures, and in many cases link up/down feedback isn't available because there are one or more switches in the path. Then, a routing protocol is necessary to detect when a connection goes down.

Because routing information is only exchanged between the ISP and the customer and doesn't propagate to the rest of the internet, any routing protocol may be used, such as RIP or OSPF. RIPv2 doesn't detect outages very quickly, so OSPF is a better choice. But in general, it's best to use BGP in this situation, as BGP is designed to be used between networks belonging to different organizations and most ISPs routinely exchange BGP routing information with some of their customers already.

However, the BGP configuration is usually slightly different from one that's used when a network connects to two or more ISPs. Often, the customer will use IP addresses from an address block that belongs to the ISP. For instance, the customer uses 10.0.16.0/22 and 10.0.20.0/24 out of the ISP's 10.0.0.0/8 block. Because the ISP already announces the 10.0.0.0/8 block, there is no need to propagate the prefixes 10.0.16.0/22 and 10.0.20.0/24 towards the rest of the world. A packet for 10.0.20.100 will flow towards the ISP because of the 10.0.0.0/8 route that the ISP advertises to the rest of the world, and then further on to the customer because of the 10.0.20.0/24 route that the customer advertises to the ISP.

Because the advertisements of the customer's prefixes aren't seen by the rest of the world, the customer can simply use a private autonomous system number rather than request a "real" AS number from ARIN, LACNIC, APNIC, AFRINIC or the RIPE NCC. Private AS numbers are the ones from 64512 to 65534. A customer should coordinate with the ISP when choosing a private AS number to avoid the situation where multiple customers use the same private AS number. Of course if a public AS number is available, that can also be used.

On the customer side, the BGP configuration is the same as one that's used towards multiple ISPs; see later in this document for examples. On the ISP side, the configuration is slightly different: the ISP has to accept the advertisements from the customer, but shouldn't let them propagate towards the rest of the world. Usually, existing prefix lists and/or AS path filter lists will take care of that.

How independent are circuits?

When connecting servers in a datacenter, the risk of physical disruption to the circuit between a customer and ISP is small. It may still be a good idea to see if it's possible to get connections routed over separate paths and/or separate cross-connects, but if that's not possible, that's unlikely to be problematic later on. However, path diversity is much more of an issue when connecting an office or other building where fiber must be brought into the building from the outside. In that situation, pay very careful attention to the routing of the circuits. Don't assume that different companies will use different paths, and when it's the same company providing multiple circuits, make sure that independent routing of the fiber paths is part of the contract.

Multihoming towards multiple ISPs

Being connected using multiple circuits to the same ISP is a lot better than having to depend on a single circuit. But depending on a single ISP still allows for several risks:

- **Physical outages.** The ISP's network may not have sufficient internal redundancy.
- **Maintenance windows.** If there is maintenance that impacts all of your connections, you'll be unreachable during the window.
- **Network management issues.** If a problematic configuration or software update is rolled out, it may affect all of your connections.
- **Routing problems.** If the ISP runs into an issue with their internal routing or BGP, this can impact your reachability.
- **Business continuity.** There have been examples of ISPs going bankrupt and their customers being disconnected. (Usually there is some lead time when this happens.)
- **Peering disputes.** Sometimes ISPs "depeer" because of peering disputes, so that customers of ISP A can no longer reach customers of ISP B, even though both are reachable from ISP C.

These risks are reason enough to connect to at least two ISPs at the same time. An additional benefit of multihoming is that once you're set up for it, it's very easy to switch ISPs, so you're in the position to negotiate for better deals. For the remainder of this guide, we'll assume multihoming towards two ISPs. However, it's entirely possible to connect to three or more ISPs at the same time.

In order to multihome towards two ISPs, you need the following:

- ***Connectivity to two BGP-capable ISPs***
- ***Your own or at least semi-independent address space***
- ***An AS number***
- ***BGP-capable routers***

In addition, you'll need to monitor the status of your BGP connectivity and you'll probably want to do at least some traffic engineering to balance incoming and/or outgoing traffic over both your ISPs.

Connectivity

Carrier-neutral datacenters are by definition served by multiple ISPs. Usually several have a router present in the datacenter, so connecting to those ISPs is relatively simple: you just need a cable within the datacenter. Almost always you'll be connected over Ethernet. Sometimes this can be UTP, but often it's done over multimode or single mode short reach fiber. Be sure to discuss this beforehand and make sure your router or switch has an interface that can accept the available cabling or fiber module options. The datacenter may charge a fee for the connection.

If you need to bring in connectivity to your office or building, things are typically more complex and more expensive. Always make sure it's possible to bring in the connection or connections beforehand. You don't want to end up in the situation where you have signed a multi-year contract with an ISP or fiber provider but the landlord, building codes or physical barriers get in the way of bringing in the connection.

Gigabit and 10 Gigabit Ethernet are the most cost effective choices for connectivity. If you are leasing dark fiber (just a fiber connection with no equipment on either end), make sure you know the optical budget so you can buy the right kind of Ethernet fiber modules. They come in many different distance ratings—typically, the longer the reach, the more expensive. You also don't want to buy longer reach than necessary because the receiver may actually receive too much power and need an attenuator to work.

Don't forget to discuss how you want to set up BGP before you sign a contract with an ISP.

Ideally, with two ISPs you'd get enough bandwidth from each to be able to run without any slowdowns if one ISP goes down. With three or more ISPs, the choice is between being able to run without slowdowns if one ISP fails or being able to run without slowdowns if all ISPs except one fail.

For instance, suppose you need 1.2 Gbps. With two ISPs, ideally you'd get at least 1.2 Gbps from each. However, most types of traffic except audio and video will slow down fairly gracefully, so if you get 1 Gbps from each ISP and one fails, you'd have to go back from 1.2 to 1 Gbps, which is probably not too problematic. As a rule of thumb, losing less than 50% of your peak bandwidth requirement is survivable for web and web-like traffic.

With three ISPs, if you get at least 0.6 Gbps from each ISP, and one fails, you still have enough bandwidth to accommodate your peak needs. If you get 1.2 Gbps from each ISP, you have enough for your peak needs even if two ISPs fail. Of course you'll also be paying for 3 x 1.2 Gbps burst capacity 100% of the time while you may only need this 0.1% of the time.

NOTE: Port capacity with the ability to burst beyond regular traffic levels may not be very expensive as long as you don't make use of that burst ability very often, and having burst capacity on the remaining ISP will serve you well if your other ISP is down.

Address space

What you need is Provider Independent (PI) address space. You'll be able to get an IPv6 /48 prefix fairly easily by becoming a "local internet registry" (LIR) at your regional internet registry (RIR). Five RIRs serve different parts of the world, see table 1 and ***Figure 6***.

RIR	REGION SERVED	IPV4 STATUS	WEBSITE
ARIN	US, Canada and some North American and Caribbean islands	None left	*www.arin.net*
LACNIC	Latin America and the Caribbean	Final /24 -/22	*www.lacnic.net*
APNIC	Asia, the Pacific and Australia	Final /22	*www.apnic.net*
AFRINIC	Africa	Available	*www.afrinic.net*
RIPE NCC	Europe, Middle East, former Soviet Union	Final /22	*www.ripe.net*

Table 1: The five Regional Internet Registries

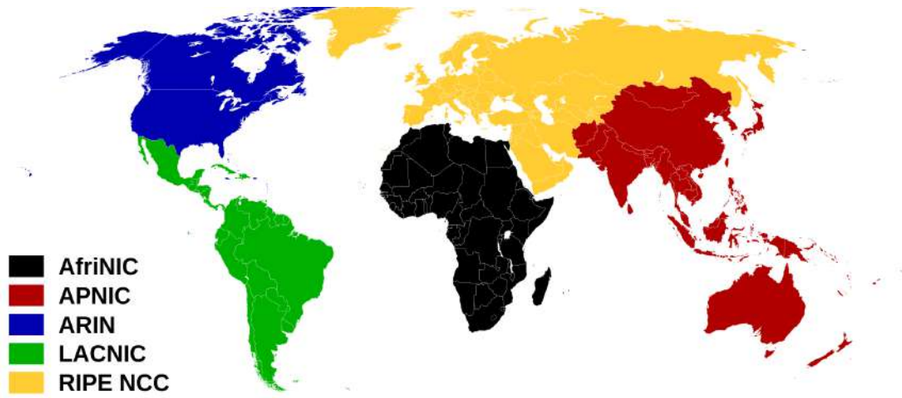

Figure 6: Parts of the world served by the RIRs. Credit: Wikmedia Commons

Becoming a LIR requires paying a one-time fee as well as a yearly fee. As a LIR, you'll be able to request IP addresses and AS numbers for yourself and your customers. You may also be able to request provider independent address space and/or an AS number without becoming a LIR, but then you'll have to go through an ISP or another intermediary that is a LIR; the RIRs don't deal directly with non-LIRs.

Becoming a LIR also qualifies you for getting IPv4 PI address space, but there is the slight snag that all RIRs except AFRINIC have effectively run out of IPv4 address space. LIRs in the RIPE NCC, LACNIC and APNIC regions can still get one last /22, but ARIN no longer has any IPv4 address space to give out. An alternative is to trade address space, i.e., buy it. See the websites of ARIN and the other RIRs to learn more about this, or use a (reputable) broker.

WARNING: There have been reports of organizations buying IPv4 address space only to find out that those addresses were still in use!

Another option is to obtain address space from an ISP or keep using address space previously obtained from an ISP. In that case, your address block or blocks will almost certainly fall within the larger address block of the ISP. You can still announce those addresses in BGP and use them much the same as provider independent addresses because of the longest match first rule. So if your ISP announces 10.0.0.0/8 and you announce 10.0.16.0/22, traffic for (for instance) 10.0.16.224 will flow towards you because even though 10.0.16.224 matches both 10.0.0.0/8 and 10.0.16.0/22, the /22 announcement is a longer match (matches a longer prefix). Using addresses in this manner is referred to as "shooting a hole" in the ISP's address block.

Being able to shoot a hole in an ISP's address block is contingent on the ISP's approval. If the ISP owning the larger address block doesn't approve, other ISPs will be reluctant to accept your advertisement. Usually, a condition for approval is that you continue to be a customer of that ISP. This of course makes sense from a business point of view, but there's also a technical reason: if networks elsewhere don't see your more specific advertisement (because it's filtered out or you have a problem with your BGP), the traffic will flow towards the ISP announcing the larger block. As such, they'll receive traffic for you so not having a connection to deliver that traffic to you (because you're no longer a customer) could be problematic.

NOTE: You need at least an IPv4 /24 prefix or an IPv6 /48 prefix to be able to multihome; many networks filter out longer prefixes.

AS number

Organizations that do their own BGP routing are called "autonomous systems" (ASes). (Organizations that don't run their own BGP are part of their ISP's AS.) Each AS is identified in BGP by an AS number. So you'll need one of those, which you can get from your RIR (through an ISP/LIR if you're not a LIR yourself). Getting an AS number is much simpler than getting address space, mostly you need to show you're going to multihome. AS numbers used to be 16-bit, but in recent years BGP has been updated to support 32-bit AS numbers.

NOTE: Make sure your router supports 32-bit AS numbers before requesting your AS number. Try *router bgp 98765* in config mode on a Cisco router; if you don't get an error message the router supports 32-bit AS numbers.

BGP-capable routers

There's the adage "nobody ever got fired for buying IBM". In the BGP space, nobody ever got fired for buying Cisco or Juniper. They both have robust BGP implementations. Slightly lesser known is Brocade, and there's also several other makers of BGP routers. Before you buy, make sure that the specific model you want to buy does have the right feature set to run BGP and any other protocols you may need.

Many routers have limitations on how many prefixes they can handle. Currently, a full IPv4 BGP table is about 600,000 prefixes. This is likely to reach a million in 2019. The IPv4 BGP table has been growing at about 16% per year, with no slowdown in recent years even though most regions are out of IPv4 addresses. The IPv6 BGP table is growing faster, but is less than 30,000 prefixes at this time.

WARNING: A router that can support a million prefixes will probably accommodate a full BGP table until some time in 2019.

Routers have a BGP RIB (routing information base) and a main routing table / RIB, which are stored in RAM. The BGP RIB holds a copy of all BGP information received from all BGP neighbors, so with two ISPs, the BGP RIB will be 1.2 million entries. The main routing table has one copy of each prefix. Then there's the FIB (forwarding information base), which is used for actually forwarding the packets. The FIB also has one copy per prefix. So the main routing table and the FIB are 600,000 prefixes each, currently. The RIBs reside in RAM, which is usually not a bottleneck. However, the FIB may have hardware constraints. Some cheap multilayer switches are able to run BGP, but only have 10,000 or so FIB entries. Until recently, routers with a FIB limit of 512,000 prefixes were used. But then the BGP table grew beyond 512,000 prefixes and those routers were no longer very useful.

It's not strictly necessary to accommodate the full BGP table in your routers, but without having full BGP feeds from each ISP, you'll have to use a default route to reach certain destinations. If that default route points to ISP A but the destination is only reachable through ISP B, this means that you won't be able to reach that destination if you don't have full BGP feeds. However, this is not something that is routinely an issue.

Router configuration

We'll assume you have two (Cisco) routers connecting to two ISPs. This means that each router speaks eBGP (external BGP) to one ISP an iBGP (internal BGP) towards your other BGP router. And they'll use OSPF to distribute the subnet prefixes used to connect to each ISP to the other router so the BGP next hop can be resolved as well as the router's loopback addresses so iBGP can be configured to/from loopback addresses and thus not depend on any particular physical interface.

```
!
interface Loopback0
 ip address10.0.19.253 255.255.255.255
!
interface GigabitEthernet0/0
 description ISP A
 ip address 10.93.194.26 255.255.255.252
!
router ospf 1
 redistribute connected subnets
 network 10.0.16.0 0.0.3.255 area 0
!
router bgp 64496
 network 10.0.16.0 mask 255.255.252.0
 timers bgp 10 30
!
! reducing the timers from default 60 / 180  so BGP will
! detect a dead neighbor in 30 second rather than 180
!
 neighbor 10.0.19.254 remote-as 64496
 neighbor 10.0.19.254 description iBGP to router 2
 neighbor 10.0.19.254 update-source Loopback0
!
! update-source makes sure we use the loopback address
! for iBGP messages
!
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 prefix-list infilter in
 neighbor 10.93.194.25 prefix-list outfilter out
 neighbor 10.93.194.25 filter-list 1 out
```

The switchover to BGP

Ideally, you'll get new IP addresses for running BGP and you'll have some time to set up BGP and test everything before these addresses are given to servers and other systems. A slightly more complex situation is the one where you'll be shooting holes in an ISP's address block. In our example, you'll be advertising 10.0.16.0/22,

while your ISP advertises 10.0.0.0/8. To switch over, two things need to happen:

1. You need to start advertising 10.0.16.0/22
2. Your ISP needs to stop statically routing 10.0.16.0/22 to you

The good thing is that both these steps can happen independently. You can set up the BGP configuration towards your ISP but without advertising your prefix (i.e., leaving out the network statement) beforehand. This shouldn't have any impact, but it's still a good idea to do this during a maintenance window outside business / busy hours. See if the BGP session comes up. Then, add the network statement and determine if your prefix propagates to the rest of the world using the monitoring tools mentioned below. If all of this works, your ISP can remove their static route, which will otherwise interfere with BGP in some situations. Again, this shouldn't have any impact, but it's best done during a maintenance window and you should be on the phone with your ISP so you can ask them to roll back the change immediately if there's any impact on your network.

The most complex situation is the one where you have a prefix that is currently advertised by your ISP, but you're going to advertise that prefix yourself. You could use the procedure discussed above, but the problem is that as long as your ISP advertises your prefix, they won't be propagating your advertisement of that same prefix. You can't have a "make before break" switchover—at least, not for the connection through that ISP. What you can do is advertise the prefix to a second ISP and then monitor if the prefix propagates to at least part of the rest of the world. (Some networks will prefer the path over your first ISP; this is normal.) Then ask your first ISP to stop advertising the prefix, and make sure that they propagate your own advertisement through them.

Monitoring BGP

Make use of the following commands to monitor BGP:

- ***show bgp ipv4 unicast summary –*** shows the status of your IPv4 BGP sessions.
- ***show bgp ipv6 unicast summary –*** shows the status of your IPv6 BGP sessions. All remote IP addresses for BGP sessions are then listed with as the last item on the line the session state or a number of prefixes received over that session for the IP version in question if the session is up. Note that state "active" means that the connection is down. (We'll leave out the IPv6 versions from now on.)
- ***show bgp ipv4 unicast -*** shows the entire BGP table.
- ***show bgp ipv4 unicast <prefix> -*** shows the information for a specific prefix.
- ***show bgp ipv4 unicast regexp <AS path regular expression> -*** shows all paths in the BGP table that match the AS path regular expression. For instance, show bgp ipv4 unicast regexp _174_ shows all AS paths with the Cogent Communications AS number in them.
- ***show bgp ipv4 unicast neighbors <address> -*** shows detailed information about a single BGP neighbor.
- ***show bgp ipv4 unicast neighbors <address> routes -*** shows all the prefixes received from the neighbor in question that are currently in the BGP table.
- ***show bgp ipv4 unicast neighbors <address> advertised-routes -*** shows all the prefixes advertised to the neighbor in question.

NOTE: You can try out the above commands on the Oregon Exchange BGP Route Viewer router, which is accessible by ***Telnet using telnet route-views.routeviews.org***. This router has BGP feeds from several dozen networks, allowing you to monitor how your prefix propagates.

It is also possible to monitor the propagation of BGP announcements and often also perform traceroutes using numerous "looking glasses" such as ***lg.he.net***. Search "BGP looking glass" to find many more.

Traffic engineering outgoing traffic

Once connections to two ISPs are operational, it is common to find that the traffic ratio between the two ISPs is suboptimal, so you may want to perform traffic engineering. Usually, there is either a lot more outgoing traffic than incoming traffic or the other way around. In networks where the majority of the traffic volume is in the outgoing direction, there is usually no need to perform traffic engineering for incoming traffic, even if incoming traffic isn't very well-balanced. For instance, if the network has 1.2 Gbps of outgoing traffic and 150 Mbps of incoming traffic, it doesn't really matter that 120 Mbps traffic arrives through ISP A and 30 Mbps through ISP B, as the 1.2 Gbps outgoing traffic is what determines what capacity the connections to each ISP need to be and how much each ISP will charge.

Traffic engineering outgoing traffic is a lot easier than traffic engineering incoming traffic for two reasons: the network has control over its own outgoing traffic, and there are 600,000 prefixes that can be manipulated for traffic in the outgoing direction, but possibly only a single prefix that can be manipulated in the incoming direction.

Suppose more outgoing traffic flows through ISP B than through ISP A, so we want a certain number of prefixes to be more attractive through ISP A. A rather blunt way to do this is to increase the local preference for certain paths/prefixes through ISP A. For instance, the following configuration increases the local preference for paths to or through Level 3 (AS 3356) over ISP A to 110, so traffic to those destinations will flow over ISP A. Note that in order to change outgoing traffic, we need to manipulate incoming BGP updates.

```
!
router bgp 64496
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 route-map traffic-eng-out-ispa in
!
ip as-path access-list 10 permit _3356_
!
route-map traffic-eng-out-ispa permit 10
 match as-path 10
 set local-preference 110
!
route-map traffic-eng-out-ispa permit 20
!
! the permit 20 clause is needed so that prefixes
```

```
! not matched by clause permit 10 are still added
! to the BGP table
!
```

Manipulating the local preference is a blunt tool because now the path over ISP A will always be preferred, even if the AS path over ISP A is much longer than the AS path over ISP B. A more subtle way to perform traffic engineering is to adjust the MED, which is only considered (with always-compare-med in effect) if the AS path is the same length. On router 2, we simply set the MED to 10 for all prefixes received from ISP B:

```
!
router bgp 64496
 bgp always-compare-med
!
! without always-compare-med MED is only compared
! between routes received from the same AS
!
 neighbor 172.31.84.17 remote-as 64511
 neighbor 172.31.84.17 description ISP B
 neighbor 172.31.84.17 route-map traffic-eng-out-ispb in
!
route-map traffic-eng-out-ispb permit 10
 set metric 10
!
```

On router 1, we now also set the MED to 10 on prefixes from ISP A, except on paths that carry the community 3356:2. Those get an MED of 5, which makes those paths preferred over the ones with MED 10 from ISP B.

```
!
router bgp 64496
 bgp always-compare-med
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 route-map traffic-eng-out-ispa in
!
ip bgp-community new-format
!
ip community-list 80 permit 3356:2
!
route-map traffic-eng-out-ispa permit 10
 match community 80
 set metric 5
!
route-map traffic-eng-out-ispa permit 20
 set metric 10
!
```

Setting up traffic engineering takes a significant amount of time and effort. Some networks have fairly stable traffic flows so traffic engineering doesn't have to adjusted often. Other

networks have less predictable traffic patterns and need frequent traffic engineering adjustments. A good alternative to managing traffic engineering manually is to use a route optimizer product such as the Noction Intelligent Routing Platform. The Noction IRP continously monitors traffic flows and how the paths over each ISP perform and automatically reroutes outgoing traffic to take advantage of the best performing paths while staying within traffic volume commitments.

Traffic engineering incoming traffic

For incoming traffic, the first option is to prepend the AS path towards the ISP that sends too much incoming traffic. This adds our own AS number one or more extra times to the AS path, making the AS path longer and thus less attractive.

```
!
router bgp 64496
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 route-map traffic-eng-in-ispa out
!
route-map traffic-eng-in-ispa permit 10
 set as-path prepend 64496
!
```

However, the AS hierarchy of the internet is very flat. This means that most networks see the same path length through ISPs A and B, and make their selection based on (rather meaningless) tie breakers. By now making the path over ISP A longer, all those paths that were previously the same length as seen by many remote ASes are now shorter through ISP B. So in most cases, a path prepend is too effective.

See ***Figure 7***. Here, ASes 100, 200, 300 and 400 see the same path length towards AS 1 through ISPs A and B. Three of them select the path through ISP A due to the BGP tie breaker rules and one selects the path through ISP B. So AS 1 receives 75% of its traffic through ISP A. Figure 8 shows what happens when AS 1 prepends the AS path towards ISP A: now the path through A is longer for all four remote ASes, so 100% of traffic arrives through ISP B in this limited example.

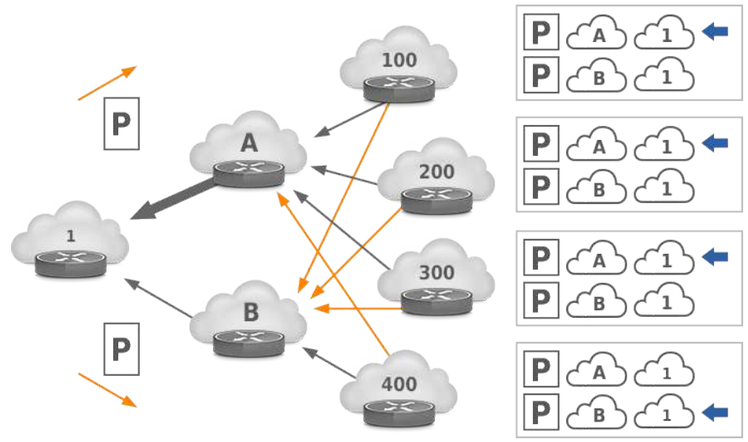

Figure 7: Without traffic engineering, most incoming traffic flows through ISP A

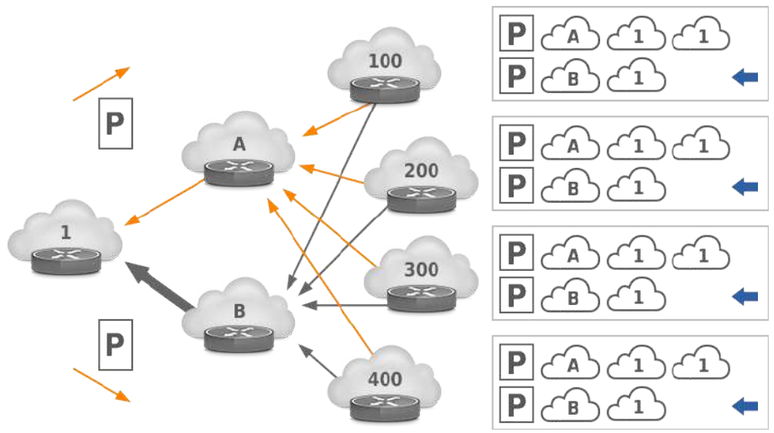

Figure 8: After an AS path prepend towards AS 10, all traffic flows through ISP B

Some ISPs offer a mechanism to selectively prepend towards some of their peers. For instance, Level 3 customers can set the following communities:

```
remarks:          ------------------------------------
remarks:          customer traffic engineering
communities - Prepending
remarks:          ------------------------------------
remarks:          65001:0 - prepend once to all peers
remarks:          65001:XXX - prepend once at peerings
to AS XXX
remarks:          65002:0 - prepend twice to all peers
remarks:          65002:XXX - prepend twice at
peerings to AS XXX
remarks:          65003:0 - prepend 3x to all peers
remarks:          65003:XXX - prepend 3x at peerings
to AS XXX
remarks:          65004:0 - prepend 4x to all peers
remarks:          65004:XXX - prepend 4x at peerings to
AS XXX
remarks:          ------------------------------------
```

So the following configuration asks Level 3 to prepend once towards Orange (AS 5511), Tata Communications (AS 6453) and AT&T (AS 7018).

```
!
router bgp 64496
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 route-map traffic-eng-in-ispa out
 neighbor 10.93.194.25 send-community
!
! the router may not send communities to BGP neighbors
! by default
!
route-map traffic-eng-in-ispa permit 10
 set community 65001:5511 65001:6453 65001:7018
!
```

Another way to limit the effect of AS path prepending is rather than advertise the entire address block as one prefix, split the block up in smaller prefixes and only prepend some of those.

```
!
router bgp 64496
 network 10.0.16.0 mask 255.255.252.0
 network 10.0.16.0 mask 255.255.254.0
 network 10.0.18.0 mask 255.255.254.0
 neighbor 10.93.194.25 remote-as 65550
 neighbor 10.93.194.25 description ISP A
 neighbor 10.93.194.25 route-map traffic-eng-in-ispa
out
 neighbor 10.93.194.25 prefix-list outfilter out
 neighbor 10.93.194.25 send-community
!
ip route 10.0.16.0 255.255.252.0 Null0
ip route 10.0.16.0 255.255.254.0 Null0
ip route 10.0.18.0 255.255.254.0 Null0


!
! null route: our own prefix needs to be in the
! routing table or it won't be advertised in BGP
!
ip as-path access-list 1 permit ^(64496_)*$
!
! regular expression that only allows AS paths
! with our AS 0 or more times
!
ip prefix-list outfilter seq 5 permit 10.0.16.0/22 le
23
!
! allow advertising subprefixes up to /23
!
ip prefix-list prepend seq 5 permit 10.0.16.0/23
!
route-map traffic-eng-in-ispa permit 10
  match ip address prefix-list prepend
  set as-path prepend 64496
!
route-map traffic-eng-in-ispa permit 20
!
```

The result is that the /22 address block is now advertised as two /23s, where 10.0.16.0/23 is prepended but 10.0.18.0/23 isn't. The entire 10.0.16.0/22 is also still advertised to make sure the network remains reachable should the /23s be filtered out.

Networks that only have an IPv4 /24 or IPv6 /48 can't use this technique effectively because prefixes longer than those aren't generally accepted by remote ASes. Also, advertising more prefixes than necessary should be avoided, as it leads to unnecessary growth of the BGP table in routers throughout the world.

NOCTION
NETWORK INTELLIGENCE

This ebook was brought to you by **Noction**.

Noction Intelligent Routing Platform enables enterprises and service providers to maximize end-to-end network performance and safely reduce infrastructure costs. The platform evaluates critical network performance metrics in real-time and responds quickly by automatically rerouting traffic through a better path to avoid outages and congestion.

Request a free trial today and see how IRP can boost your network performance.

Start a Free Trial