Fable Of Contents

ISP TECH TALK

by Avi Freedman

WHERE WE ARE IN THE "ROUTING COLUMNS" SERIES

This month, we'll talk more about BGP and introduce some concepts and details that you'll need to have before we can move on to showing you how to actually configure a router to speak BGP to your provider(s).

By the end of next month's column you should understand enough about BGP to set it up in a simple and safe configuration if you're single or multi-homed.

In future columns we'll be talking about things like RIP, OSPF, and other internal routing protocols. Also in future months we'll talk about more exotic BGP features, but we're going to keep it pretty simple for now.

BGP is a routing protocol-it's used to tell people outside your network (upstream providers or "peers") about the routes (or portions of the IP address space) you know how to get to inside your network. The primary purpose of BGP4 (as we're studying it here) is to advertise routes to other networks (Autonomous Systems). The secondary purpose of BGP4 is to get routes from providers and peers so that you can make more intelligent decisions about how to send traffic out of your network. BGP stands for Border Gateway Protocol. The popular BGP protocol that people speak of when they ask, "can a Cisco 2501 speak BGP?" is actually BGP4. BGP as we're studying it here is strictly an EGP, or Exterior Gateway protocol.

BGP TERMINOLOGY PART I

An AS, or Autonomous System, is a way of referring to someone's network. That network could be yours; a friend's; MCI's; Sprintlink's; or anyone's. Normally an AS will have someone or ones responsible for it (a point of contact, typically called a NOC, or Network Operations Center) and one or multiple border routers (where routers in that AS peer and exchange routes with other ASs), as well as a simple or complicated internal routing scheme so that every router in that AS knows how to get to every other router and destination within that AS.

When you "advertise" routes to other entities, one way of thinking of those route advertisements is as promises to carry data to the IP space represented in the route being advertised. For example, if you advertise 192.204.4.0/24 (the Class C starting at 192 .204.4.0 and ending at 192.204.4.255), you promise that you know how to carry that data destined for any address in 192.204.4.0/24 to its ultimate destination. The cardinal sin of BGP routing is advertising routes that you don't know how to get to. This is called black-holing someone. If you advertise some part of the IP space that is owned by someone else, and that advertisement is more specific than the one made by the owner of that IP space, then all of the data on the Internet destined for the black-holed IP space will flow to your border router. Needless to say, this makes that address space disconnected from the Net and makes many people unhappy. The second most heinous sin of BGP routing is not having strict enough filters on the routes you advertise. The bottom line: Test your configs and watch out for typos. Think through everything that you do in terms of how it could screw things up.

Also, one terminology note: Classless routes are sometimes called prefixes. When someone talks about a prefix they're talking about a route with a particular starting point and a particular specificity (length)-So 207.8.96.0/24 and 207.8.96.0/20 are not the same prefix (route), because they are of a different specificity-a /24 is as big as one Class C and /20 is as big as 16 Class Cs.

BEING "CONNECTED" TO THE INTERNET

Throughout this discussion it's critical to think about what it means to be connected to the Internet. To be connected to the Internet, for each host that is on the Internet, you need to be able to send a packet out a path that will ultimately wind up at that host. Furthermore, that host has to have a path back to you. This means that whoever provides Internet connectivity to that host has to have a path to you. This ultimately means that the provider must hear a route that covers the section of the IP space that you are using, or you will not have connectivity to the host in question.

Every IP address that you can get to on the Internet is reachable because someone, somewhere, has advertised a route that "covers" it. Similarly, if there is not a generally-advertised route to cover an IP address, then no one on the Internet will be able to reach it.

HARDWARE AND SOFTWARE FOR SPEAKING BGP

The most commonly used implementations of BGP are Cisco routers; Bay routers; and PC clones running Linux, BSD, or some other UNIX variant and a program called gated to manage BGP. We're going to talk about configuring Ciscos for BGP now, but you can use Bays or Gated boxes. For now, I'd advise using Ciscos over Bays for BGP since there are many more Cisco-literate people, and since the Bay command- line interface (BCC, or Blatant Cisco Clone) isn't out yet. Gated is sort of a "BGP assembly language," and I'd advise against using it unless you have a very good low-level understanding of routing and of BGP.

A full view of the Internet is basically all the BGP routes that your upstream providers hear. You'll need a Cisco 4500 or 4700 with 32 MB; or 36xx, 720x, 70x0, or 75xx with 64 MB to do this. Currently, you can take at least one or two views of the

45,000 routes or so that are the Global Internet Routing Table. A view is another way to get to a certain prefix. So both routes, or views have the same starting IP and prefix length, but each route comes from a different source.

But a Cisco 2501 will do fine to announce your routes to the world. It should be able to handle 1.5 loaded T1s worth of traffic as well, and take 5,000 to 15,000 BGP routes from your providers. This allows you to do some tuning of your outbound traffic flow.

TERMINOLOGY PART II:
PEERING SESSIONS AND ASNs

There's a bunch of terminology associated with BGP. We already talked about Autonomous Systems. An ASN is an Autonomous System Number used to represent an Autonomous System to the world. That number "identifies" your network to the world. Except for Sprintlink, most networks use (or at least show to the world) only one ASN.

BGP-speaking routers exchange routes with other BGP-speaking routers via peering sessions. At a technical level, this is what it means for one network or organization to peer with another. Here's a simplified Cisco code snippet of a router bgp clause:

```
router bgp 64512
 &ltomitted lines>
 neighbor 207.106.127.122 remote-as 701
 &ltomitted lines>
 neighbor 137.39.10.46 remote-as 4969
 &ltomitted lines>
```

The clause starts out by saying router bgp 64512. This means "What follows is a list of commands that describe how to speak BGP on behalf of ASN 64512." (We're using 64512 in our examples because it's not a live ASN, so if anyone uses a configuration straight from this column and leave the fake ASN, automated route-examination programs will detect it). ASNs go from 1-65535, and 64512 is in the reserved range.

To bring up a peering session, all you need to do is have that one neighbor line under the router bgp clause. In this example, 137.39.10.46 is the remote IP address of a UUNET router

(UUNET is ASN 701). Remote, that is, with respect to the customer's router. 207.106.127.122 is the remote IP address of a Net Access router (Net Access is ASN 4969).

In practice, however, you almost always use more than that one line to tell BGP how to exchange routes with that neighbor via that peering session. A typical neighbor clause is:

```
router bgp 64512
 &ltomitted lines>
 neighbor 207.106.127.122 remote-as 4969
 neighbor 207.106.127.122 next-hop-self
 neighbor 207.106.127.122 send-communities
 neighbor 207.106.127.122 route-map prepend-once out
 neighbor 207.106.127.122 filter-list 2 in
 &ltomitted lines>
```

WHAT DO YOU DO WITH BGP?

Speaking BGP to your providers and/or peers lets you do two things. The first is to make semi-intelligent routing decisions concerning the best path for a particular route to take outbound from your network. Otherwise you would simply set a default route from your border routers into your providers. Second, and more importantly, you can announce your routes to those providers, for them to in turn to announce to others (transit) or just use internally (in the case of peers).

PEERING SESSIONS

The purpose of the neighbor clauses is to bring up peering sessions with neighbors. For now, assume that all neighbors must be either on the other end of a leased-line from you, or on a LAN interface (Ethernet, Fast Ethernet, FDDI). It is possible to have BGP peering sessions that go over multiple hops-but eBGP multihop is a more advanced topic and has many potential pitfalls.

Every time a neighbor session comes up, each router will evaluate every BGP route it has by running it through any filters you specificity in the neighbor clause. Any routes that pass the filter are sent to the remote end. This filtering is a critical process-the most dangerous element of BGP is the risk that your filtering will go awry and you'll announce routes that you shouldn't to your upstream providers.

While the session is up, BGP updates will be sent from one router to the other each time one of the routers knows about a new BGP route or needs to withdraw a previous announcement.

The

```
sho ip bgp summ
```

Cisco command will show you a list of all peering sessions:

```
brain.netaxs.com#sho ip bgp summ
BGP table version is 1159873, main routing table version 1159873
44796 network entries (98292/144814 paths) using 9596344 bytes of memory
16308 BGP path attribute entries using 2075736 bytes of memory
12967 BGP route-map cache entries using 207472 bytes of memory
16200 BGP filter-list cache entries using 259200 bytes of memory

Neighbor       V AS    MsgRcvd MsgSent TblVer  InQ OutQ Up/Down  State
205.160.5.1    4 6313  0       0       0       0   0    never    Active
207.106.90.1   4 64514 1145670 237369  1159873 0   0    4d03h
207.106.91.5   4 64515 6078    5960    1159869 0   0    4d03h
207.106.92.16  4 64512 6128    6782    1159870 0   0    4d03h
207.106.92.17  4 64512 5962    6894    1159870 0   0    10:08:46
206.245.159.17 4 4231  161072  276660  1159870 0   0    2d05h
207.44.7.25    4 3564  6109    310292  1159867 0   0    22:40:50
207.106.33.3   4 64513 164708  724571  1159866 0   0    3d23h
207.106.33.4   4 3564  6086    274182  1159853 0   0    4d03h
207.106.127.6  4 6078  5793    310011  1159869 0   0    2d03h
```

Most of it is pretty self-explanatory; briefly:

- The V column is the BGP version number. If it is not 4, something is very wrong! BGP version 3 doesn't understand about Classless (CIDR) routing and is thus dangerous.
- The AS column is the remote ASN.
- InQ is the number of routes left to be sent to us.
- OutQ is the number of routes left to be sent to the other side.

- The Up/Down column is the time that the session has been up (if nothing is in the State column) or down (if something is).
- Anything in the State column indicates that the session is not up. A State of Active means that the session is inactive. Just one of the nomenclature flaws of BGP.

This is a session summary from one of Net Access's core routers. The 6451X Autonomous Systems are BGP sessions to other Net Access routers-those ASNs are not shown to the world. The 205.160.5.1 session is a session that is down, and the sessions where the remote Autonomous Systems are 4231, 3564, and 6078 are external peering sessions. (This router peers with our Pennsauken, MAE East, and MAE West routers via the 645xx sessions.)

eBGP vs. iBGP: A Note

We're talking about eBGP in this document. eBGP and iBGP share the same low-level protocol for exchanging routes, and also share some of the algorithms, but eBGP is used to exchange routes between different Autonomous Systems, while iBGP is used to exchange routes between the same Autonomous System. In fact, iBGP is one of the interior routing protocols that you can use to do active routing inside your network. We'll talk more about iBGP in future columns when we cover all of the major interior routing protocols: OSPF, iBGP, IS-IS, RIP, and RIPv2.

The major difference between eBGP and iBGP is that eBGP tries like crazy to advertise all known BGP routes to everyone-you have to put filters in place to stop it from doing so. iBGP is actually pretty difficult to get working because it tries like crazy not to redistribute routes. In fact, all iBGP-speakers inside your network have to peer with all other iBGP speakers to make it work. This is called a routing mesh and, as you can imagine, is quite a mess. If you have 20 routers, each router has to peer with every other router. The solution to this is BGP confederations-a topic for a future column.

BGP AND THE SINGLE-HOMED

When you have one upstream provider, it is rarely desirable to speak BGP to them. Why? Well, you only have one path out of your network. So filling your router with 45,000 BGP routes isn't going to do you any good, since all of those routes point to the same place (your one upstream provider).

And if you have one upstream provider, it's almost guaranteed that you are using sub-allocations (CIDR delegations, to be precise) of their larger IP blocks (aggregates). In this case your provider is not going to advertise your more specific routes-see May 1997's column for an explanation.

ROUTE WITHDRAWLS

If you're single-homed, the whole Internet doesn't need to know if you lost connectivity to your provider, since there's no other path to get to you.

Similarly, in case it's not obvious, if you are multi-homed and your providers are announcing more specific prefixes for you, then they should stop announcing you when they don't know how to get to you any more.

The beauty of speaking BGP to your providers is that when you lose connectivity to them (a T-1 or 56k or what you have goes down), the BGP session will go down as well and all of those route advertisements will be automatically withdrawn.

AS-PATHS

Every time a route is advertised via BGP, it is stamped with the ASN of the router doing the advertising. As a route moves from Autonomous System to Autonomous System (network to network), it builds up an AS-PATH. Each route starts out with a null AS-PATH, represented by the regular expression ^$.

The AS-PATH is useful for the following reasons:

- It provides a diagnostic trace of routing on the Net. If you have full routes in one of your routers, or have query access to a router that does (such as telnet://route-server.cerf.net), you can find the route that encompasses a particular IP address and see which ASNs have advertised it. If you do some poking around, you can even see how a provider is actually connected (as opposed to what they might claim...)

- It is one of a number of metrics that determine how routes heard via BGP are inserted into the actual IP routing table. We'll be talking more about metrics in the future.

- It is something that allows you to do policy routing of sorts, although policy routing has many different definitions, so watch out. You use the AS-PATH to filter routes. Why would you want to do this? Perhaps you only want to take UUNET, MCI, and ANS route from one provider (because of limited memory in your router). Or perhaps you want to make sure you only send routes originating in your network. There are many reasons why you'd want to filter based on the AS-PATH. These reasons will become clear as you read on. While it's true that most filtering is now done with communities (a community is another number that you can stamp on a route heard or to be announced via), AS-PATH filtering is the best first step that you can work with to get comfortable with filtering routes. And if your network is fairly simple (as 90 percent of the networks are), then you won't need anything fancier for quite some time. We'll talk about doing this in a few months, when we focus on tuning outbound traffic flow.

For routes of the same specificity, as-path length is going to be the deciding factor for choosing which of multiple routes gets used by the router (i.e. put into the IP routing table) when you're just starting out.

BGP METRICS (ATTRIBUTES) AND ROUTE SELECTION: INTRODUCTION

We'll shortly go into BGP metrics and attributes-parameters associated with BGP routes that allow you to select and change the selection of the best BGP route for a certain destination.

For now, keep in mind that unless you do any tuning on your own:

- The most specific route always wins. Whether it's a BGP route or a static internal route, the most specific route always wins.
- If there is a choice between multiple BGP routes, then the one with the shortest AS_PATH wins.

WHAT IS ROUTE FLAP AND WHY IS IT BAD?

When you assert a route-saying "I know how to get to 192.204.4.0/24" based on some internal knowledge that you actually do know how to get to 192.204.4/0- the natural (and previously-thought-to- be-correct-thing-to-do) is to withdraw that assertion if you in fact no longer know how to get to that route.

But look at what happens when you withdraw that assertion. Your providers must then also withdraw that assertion. And then their providers and peers must do the same. All in all, thousands of routers around the world now have to look at that route and decide if they have a next-best path in their BGP table, and insert it as the current best path in their IP routing table. This consumes many CPU-seconds on routers that are sometimes very busy.

In fact, it was consuming so much CPU time a few years ago that Sean Doran of Sprintlink said, "this must stop." Several people came up with an idea, which Cisco implemented in record time, to dampen the route flaps. You'll hear people say "damp" and "dampen." There's no real consensus about which is the correct term.

What this means in practice today is that if your routes flap more than one or two complete up-down-up cycles, then you will be dampened by many providers for at least an hour or so. So even if you're only single-homed, you will be dampened if your provider withdraws your routes every time your T-1 flips up and down a few times because some Bell guy tripped over a wire.

So do not ask your upstream provider to announce you unless it makes a difference. The benefit of being multiply-announced outweighs the possible negative effects of being dampened due to instability in either your or your provider's network.

WHAT TO KEEP IN MIND WHEN CONFIGURING BGP

When you're bringing up a new BGP session, or considering how to do BGP in general, the things to keep in mind for each peer are:

- What routes do you want them to hear? The most important thing is to ensure that you do not redistribute routes to which you are not providing Internet connectivity.
- What do you want to do with the routes that you hear via the session? Do you want to tune them? Only take some? Take them all?
- Do you have the memory and CPU in your router to really do what you want?

INTERNET CONNECTIVITY WITHOUT BGP

Let's review what happens when you are connected to the Internet without speaking BGP to your provider. Last month's column explained this in more detail.

- You create a default route toward your upstream provider, and all non-local packets go out the interface specified by the route
- Your provider probably put static routes toward you on their side, and redistributes those static routes into their IGP. Then they probably redistribute their IGP into BGP -unless all of its BGP is done statically.

If you have any address space inside of your provider's larger netblock or aggregate, then you won't be advertised to the outside world specifically-your provider will just advertise their larger block. If you have any other networks (an old Class C; customers with address space; etc.), then your provider will just statically announce those routes to the world and statically route them inside their network to your leased- line/ router interface(s).

With BGP, your provider gives you all of the routes they have (the easy part), and listens to your route announcements and then redistributes some or all of those to their peers and customers. This is the hard part for them. The net difference is just that they may start advertising a more specific route (no mean task in a complicated network designed, as most networks are, to prevent the accidental leaking of more specific routes) or that the routes that they normally advertise for you under just their ASN will now have your ASN attached as well.

BGP AND THE MULTI-HOMED

OK, so you're multi-homed. What is the most important thing about BGP to you? The ability to have it announce routes. Getting full or partial routes from your providers is cool-and may even be useful-but you can do almost as well by just load-balancing all outgoing traffic in either a round-robin or route- caching manner.

So, the most important thing about being multi-homed is the ability to have your routes advertised to your providers -and by them to their providers and peers (i.e. to the rest of the Internet). Doing this basic level of route advertisement is not hard. You just have to do it in a paranoid way.

If you screw up BGP routing, then you may get slapped down pretty hard. Screw ups with BGP route advertisements can be felt all over the Internet. To repeat: Screw ups with BGP route advertisements can be felt all over the Internet. If your provider is smart, then they will also implement filters to prevent you from screwing them and the Internet up. But don't count on it.

If you were to announce a route that was more specific than, for example, the otherwise-best route for Yahoo's web servers, you would black-hole Yahoo for a period of time. Needless to say, they would not be very happy with you. The solution is to do good filtering on your end-and for your provider to also do excellent filtering wherever possible.

Before you start playing with BGP, you might really want to wait and read the "Configuring a Cisco Router" column (coming out in the next few months). If you do go ahead and are implementing BGP for the first time, then get a friend or another provider to review your proposed configs for you before implementing them.

MULTI-HOMING AND LOAD-BALANCING

Generally, the goal of multi-homing is to use both connections in a sane manner and load-balance them somehow. Ideally, you'd like roughly half the traffic to go in and out of each connection. You'd also like fail-over routing, where if one connection goes down, then the other one keeps you connected to the Internet. In an ideal network, you'd be able to have any one of your connections to the Net go down and still maintain connectivity and speed.

Incoming traffic is controlled by how you announce your routes to the world. Packets will flow into your network because someone heard and is using a route announcement. Outgoing traffic is controlled by the routes that you allow to flow into your border router(s)-and is thus much easier to control and tune.

APOLOGIES

Sorry if this seems like a teaser-we'll get to the actual BGP configuration process and examples next month. If you're impatient, a somewhat long document is available at www.netaxs.com/~freedman/bgp/bgp.html.

Colorado Offices
13949 W Colfax Ave Suite 250, Golden, CO 80401
Voice: 303-235-9510; Fax: 303-235-9502

Fable Of Contents