BGP overview, the threats and how to mitigate them

Author: Siim Adamson

BGP overview

BGP definition rfc4271

The Border Gateway Protocol (BGP) is an inter-Autonomous System routing protocol.

The primary function of a BGP speaking system is to exchange network reach ability information with other BGP systems. This network reachability information includes information on the list of Autonomous Systems (ASes) that reachability information traverses. This information is sufficient for constructing a graph of AS connectivity for this reachability, from which routing loops may be pruned and, at the AS level, some policy decisions may be enforced.

BGP-4 provides a set of mechanisms for supporting Classless Inter-Domain Routing (CIDR) [RFC1518, RFC1519]. These mechanisms include support for advertising a set of destinations as an IP prefix and eliminating the concept of network "class" within BGP. BGP-4 also introduces mechanisms that allow aggregation of routes, including aggregation of AS paths. Routing information exchanged via BGP supports only the destination-based forwarding paradigm, which assumes that a router forwards a packet based solely on the destination address carried in the IP header of the packet. This, in turn, reflects the set of policy decisions that can (and cannot) be enforced using BGP. BGP can support only those policies conforming to the destination-based forwarding paradigm.[1]

BGP and other routing protocols

Protocols that run inside an enterprise are called interior gateway protocols (IGPs). Examples of IGPs include RIP versions 1 and 2, EIGRP, and OSPF.

Protocols that run outside an enterprise, or between autonomous systems, are called exterior gateway protocols (EGPs). Typically, EGPs are used to exchange routing information between Internet Service Providers (ISPs).

Since 1994, Border Gateway Protocol version 4 (BGP4) has become the core routing protocol of the Internet. All previous versions are considered obsolete. The Internet is a collection of autonomous systems that are interconnected to allow communication among them. BGP provides the routing between these autonomous systems. Most ISPs must use BGP to establish routing between one another.[2]

BGP is path-vector routing protocol. Destination IP belongs to some subnet (bunch of IP's). Subnets are collected under AS's. In general BGP gives information in which path to go to reach the AS under which destination IP belongs to. As noted in rfc4271, the classic definition of an autonomous system is a set of routers under a single technical administration, using an IGP and common metrics to route

[1] http://www.ietf.org/rfc/rfc4271, 26.09.2010

[2] Cisco CCNP Curriculum v5 Module 6 Chapter Overview, not public document, 26.09.2010

packets within the autonomous system, and using an inter-autonomous system routing protocol (also called an EGP) to determine how to route packets to other autonomous systems.

Autonomous systems can use more than one IGP, potentially with several sets of metrics. From the BGP point of view, the most important characteristic of an autonomous system is that it appears to other autonomous systems to have a single coherent interior routing plan and presents a consistent picture of reachable destinations. All parts of an autonomous system must connect to each other.

When BGP is running between routers in different autonomous systems, it is called External BGP (EBGP). When BGP is running between routers in the same autonomous system, it is called Internal BGP (IBGP).[3]

When to use and when not to use BGP

Enterprises that want to connect to the Internet do so through one or more ISPs. If an organization has only one connection to one ISP, they probably do not need to use BGP. Instead, they would use a default route. However, if they have multiple connections to one or to multiple ISPs, BGP may be appropriate because it allows them to manipulate path attributes to select the optimal path.[3]

If the routing policy that you implement in an autonomous system is consistent with the policy in the ISP autonomous system, it is not necessary or desirable to configure BGP in that autonomous system.[4]

BGP Multihoming

Multihoming is when an autonomous system has more than one connection to the Internet. Two typical reasons for multihoming are as follows:

1. **To increase the reliability of the connection to the Internet**: If one connection fails, the other connection remains available.
2. **To increase the performance of the connection**: Better paths can be used to certain destinations.

The benefits of BGP are apparent when an autonomous system has multiple EBGP connections to either single or multiple autonomous systems. Multiple connections allows an organization to have redundant connections to the Internet so that connectivity can still be maintained if a single path becomes unavailable.

An organization can be multihomed to either a single ISP or to multiple ISPs. A drawback to having all of your connections to a single ISP is that connectivity issues in that single ISP can cause your autonomous system to lose connectivity to the Internet. By having connections to multiple ISPs, an organization gains the following benefits:

1. Redundancy with the multiple connections
2. Not tied into the routing policy of a single ISP
3. More paths to the same networks for better policy manipulation

[3] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Concepts and Terminology, not public document, 26.09.2010

[4] Cisco CCNP Curriculum v5 Module 6 Chapter Features of BGP, not public document, 26.09.2010

A multihomed autonomous system can run EBGP with its external neighbors and might also run IBGP internally. If an organization wants to perform multihoming with BGP, there are three common ways to do this:

1. **Each ISP passes only a default route to the autonomous system**: The default route is passed to the internal routers. Receiving only a default route from each ISP requires the fewest resources within the autonomous system, because a default route is used to reach any external destination. The autonomous system sends all its routes to the ISPs, which process and pass them onto other autonomous systems. The limitations of this option:
   a. Path manipulation cannot be performed because only a single route is being received from each ISP.
   b. Bandwidth manipulation is extremely difficult and can be accomplished only by manipulating the IGP metric of the default route.
   c. Diverting some of the traffic from one exit point to another is challenging because all destinations are using the same default route for path selection.
2. **Each ISP passes only a default route and provider-owned specific routes to the autonomous system**: These routes may be passed to internal routers, or all internal routers in the transit path can run BGP and pass these routes between them. An enterprise running EBGP with an ISP that wants a partial routing table generally receives the networks that the ISP and its other customers own. The enterprise can also receive the routes from any other autonomous system. If the ISP passes this information to a customer that wants only a partial BGP routing table, the customer can redistribute these routes into its IGP. The internal routers of the customer (these routers are not running BGP) can then receive these routes via redistribution. They can take the nearest exit point based on the best metric of specific networks, instead of taking the nearest exit point based on the default route. Acquiring a partial BGP table from each provider is beneficial because path selection is more predictable than when using a default route.
3. **Each ISP passes all routes to the autonomous system**: All internal routers in the transit path run BGP and pass these routes between them. In the third multihoming option, all ISPs pass all routes to the autonomous system, and IBGP is run on all the routers in the transit path in this autonomous system. This option allows the internal routers of the autonomous system to take the path through the best ISP for each route. This configuration requires a lot of resources within the autonomous system, because it must process all of the external routes. The autonomous system sends all of its routes to the ISPs, which process the routes and pass them to other autonomous systems.[5]

BGP Routing Between Autonomous Systems

The main goal of BGP is to provide an interdomain routing system that guarantees loop-free exchange of routing information between autonomous systems. Routers exchange information about paths to destination networks. There are many RFCs relating to BGP4, the current version of

[5] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Concepts and Terminology, not public document, 26.09.2010

BGP, including 1772, 1773, 1774, 1930, 1966, 1997, 1998, 2042, 2385, 2439, 2545, 2547, 2796, 2858, 2918, 3065, 3107, 3392, 4223, and 4271.[6]

BGP4 and its extensions are the only acceptable versions of BGP available for use on the public-based Internet. BGP4 carries a network mask for each advertised network and supports both variable-length subnet masking (VLSM) and CIDR. BGP4 predecessors did not support these capabilities, which are currently mandatory on the Internet. When CIDR is used on a core router for a major ISP, the IP routing table, which is composed mostly of BGP routes, has more than 170,000 CIDR blocks. Not using CIDR at the Internet level would cause the IP routing table to have more than 2,000,000 entries. Using BGP4 and CIDR prevents the Internet routing table from becoming too large for interconnecting millions of users.[7]

IANA is the organization that maintains records of global IP address allocation. The Regional Internet Registry (RIR) is an organization overseeing the allocation and registration of Internet number resources within a particular region of the world. Resources include IP addresses (both IPv4 and IPv6) and autonomous system numbers.[8]

Autonomous system numbers are 16-bits, ranging from 1 to 65535. rfc1930 provides guidelines for the use of numbers. The numbers 64512 through 65535 are reserved for private use, much like private IP addresses.[9] Using an IANA-assigned autonomous system number rather than a private number is necessary only if your organization plans to use an EGP, such as BGP, and connect to a public network, such as the Internet.

Comparison with IGPs

BGP works differently than IGPs. An internal routing protocol looks for the quickest path from one point in a corporate network to another based on certain metrics. RIP uses hop counts that look to cross the fewest Layer 3 devices to reach the destination network. OSPF and EIGRP look for the best speed according to the bandwidth statement on the interface. All internal routing protocols look at the path cost to a destination.[6]

BGP - an external routing protocol, does not look at speed for the best path. Rather, BGP is a policy-based routing (PBR) protocol that allows an autonomous system to control traffic flow using multiple BGP path attributes. BGP allows a provider to fully use all its bandwidth by manipulating these path attributes.[6]

Path-Vector Functionality

Internal routing protocols announce a list of networks and the metrics to get to each network. In contrast, BGP routers exchange network reachability information, called path vectors, made up of path attributes. The path-vector information includes a list of the full path of BGP autonomous system numbers (hop by hop) necessary to reach a destination network and the networks that are reachable at the end of the path. Other attributes include the IP address to get to the next autonomous system (the next-hop attribute) and an indication of how the networks at the end of the

[6] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Routing Between Autonomous Systems, not public document, 26.09.2010

[7] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Routing Between Autonomous Systems, not public document, 26.09.2010

[8] http://www.iana.org/numbers/, 26.09.2010

[9] http://www.faqs.org/rfcs/rfc1930.html, 26.09.2010

path were introduced into BGP (the origin code attribute). This autonomous system path information is useful to construct a graph of autonomous systems based on the information exchanged between BGP neighbors. BGP views the whole internetwork as a graph, or tree, of autonomous systems. The connection between any two systems forms a path. The collection of path information is expressed as a sequence of autonomous system numbers called the AS path. This sequence forms a route to reach a specific destination. The AS path is always loop-free. A router running BGP does not accept a routing update that already includes the router autonomous system number in the path list, because the update has already passed through its autonomous system, and accepting it again would result in a routing loop.[10]

BGP Routing Policies

BGP allows routing-policy decisions at the autonomous system level to be enforced. These policies can be implemented for all networks owned by an autonomous system, for a certain CIDR block of network numbers (prefixes), or for individual networks or subnetworks. BGP specifies that a BGP router can advertise to neighboring autonomous systems only those routes that it uses itself. This rule reflects the hop-by-hop routing paradigm that the Internet generally uses. The hop-by-hop routing paradigm does not support all possible policies. For example, you cannot influence how a neighboring autonomous system routes traffic, but you can influence how your traffic gets to a neighboring autonomous system. BGP does support any policy that conforms to the hop-by-hop routing paradigm. Because the Internet currently uses the hop-by-hop routing paradigm only, and because BGP can support any policy that conforms to that paradigm, BGP is highly applicable as an inter-autonomous-system routing protocol. For example the following paths are possible for AS 64512 to reach networks in AS 64700 through AS 64520:

- 64520 64600 64700
- 64520 64600 64540 64550 64700
- 64520 64540 64600 64700
- 64520 64540 64550 64700

AS 64512 does not see all these possibilities. AS 64520 advertises to AS 64512 only its best path, 64520 64600 64700, in the same way that IGPs announce only their best least-cost routes. This path is the only path through AS 64520 that AS 64512 sees. All packets that are destined for 64700 through 64520 take this path. Even though other paths exist, AS 64512 can only use what AS 64520 advertises for the networks in AS 64700. The AS path that is advertised, 64520 64600 64700, is the AS-by-AS (hop-by-hop) path that AS 64520 uses to reach the networks in AS 64700. AS 64520 will not announce another path, such as 64520 64540 64600 64700, because it did not choose that as the best path based on the BGP routing policy in AS 64520. AS 64512 does not learn of the second-best path or any other paths from AS 64520, unless the best path of AS 64520 becomes unavailable. Even if AS 64512 were aware of another path through AS 64520 and wanted to use it, AS 64520 would not route packets along that other path because AS 64520 selected 64520 64600 64700 as its best path, and all AS 64520 routers use that path as a matter of BGP policy. BGP does not let one autonomous system send traffic to a neighboring autonomous system, intending that the traffic take a different route from that taken by traffic originating in the neighboring autonomous system. To reach the

[10] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Path-Vector Functionality, not public document, 26.09.2010

networks in AS 64700, AS 64512 can choose to use AS 64520 or it can choose to go through the path that AS 64530 is advertising. AS 64512 selects the best path to take based on its own BGP routing policies.[11]

Features of BGP

BGP is used by ISPs so that they can communicate and exchange packets. The ISPs have multiple connections to each other and agreements to exchange updates. BGP implements the agreements between two or more autonomous systems. Improper controlling and filtering of BGP updates can potentially allow an outside autonomous system to affect the traffic flow to your autonomous system. For example, if you are a customer connected to ISP-A and ISP-B (for redundancy), you want to implement a routing policy to ensure that ISP-A does not send traffic to ISP-B via your autonomous system. You do want to be able to receive traffic destined to your autonomous system through each ISP.

BGP is categorized as an advanced distance vector protocol, but it is actually a path-vector protocol. BGP is very different from standard distance vector protocols, such as RIP. BGP uses TCP as its transport protocol, which provides connection-oriented reliable delivery. BGP assumes that its communication is reliable; therefore, it does not have to implement retransmission or error recovery mechanisms. BGP uses TCP port 179. Two routers using BGP form a TCP connection with one another and exchange messages to open and confirm the connection parameters. These two BGP routers are called peer routers, or neighbors. After the connection is made, BGP peers exchange full routing tables. However, since the connection is reliable, BGP peers subsequently send only changes (incremental or triggered updates) after that. Reliable links do not require periodic routing updates; therefore, routers use triggered updates instead. BGP sends keepalive messages, similar to the hello messages sent by OSPF, IS-IS, and EIGRP. BGP is the only IP routing protocol to use TCP as its transport layer. BGP, has more than 170,000 networks (and growing) on the Internet to advertise, and it uses TCP to handle the acknowledgment function. TCP uses a dynamic window, which allows 65,576 bytes to be outstanding before it stops and waits for an acknowledgment. For example, if 1000-byte packets are being sent, BGP would stop and wait for an acknowledgment only when 65 packets had not been acknowledged, when using the maximum window size. TCP is designed to use a sliding window in which the receiver acknowledges at the halfway point of the sending window. This method allows any TCP application, such as BGP, to continue to stream packets without having to stop and wait as OSPF or EIGRP would require.[12]

BGP Databases

A router running BGP (BGP speaker) keeps several tables to store BGP information that it receives from and sends to other routers. These tables include a neighbor table, a BGP table (also called a forwarding database or topology database), and an IP routing table. For BGP to establish an adjacency, you must configure it explicitly for each neighbor. BGP uses TCP as its transport protocol (port 179). It forms a TCP connection with each of the configured neighbors and keeps track of the state of these relationships by periodically sending a BGP TCP keepalive message (every 60 secounds by default). T wo BGP speakers that form a TCP connection between one another for the purpose of exchanging routing information are referred to as neighbors or peers. To establish adjacency network layer reachability is needed (two peers can ping each other), the neighbors exchange the BGP routes

[11] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Routing Polices, not public document, 26.09.2010

[12] Cisco CCNP Curriculum v5 Module 6 Chapter Feature of BGP, not public document, 26.09.2010

that are in their IP routing table. Each router collects these routes from each neighbor that successfully establishes an adjacency and then places them in its BGP forwarding database. All routes that have been learned from each neighbor are placed into the BGP forwarding database. The best routes for each network are selected from the BGP forwarding database using the BGP route selection process and then offered to the IP routing table. Each router compares the offered BGP routes to any other possible paths to those networks, and the best route, based on administrative distance, is installed in the IP routing table. EBGP routes (BGP routes learned from an external autonomous system) have an administrative distance of 20. IBGP routes (BGP routes learned from within the autonomous system) have an administrative distance of 200.[13]

BGP Message Types

The four BGP message types are open, keepalive, update, and notification:

1. **Open message**: An open message includes the following information:
    a. **Version number**: The highest common version that both routers support is used. All BGP implementations today use BGP4.
    b. **AS number**: The AS number of the local router. The peer router verifies this information. If it is not the AS number that is expected, the BGP session is torn down.
    c. **Hold time**: Maximum number of seconds that can elapse between the successive keepalive and update messages from the sender. On receipt of an open message, the router calculates the value of the hold timer by using whichever is smaller: its configured hold time or the hold time that was received in the open message.
    d. **BGP router ID**: 32-bit field indicating the BGP ID of the sender. The BGP ID is an IP address that is assigned to that router, and it is determined at startup. The BGP router ID is chosen: the highest active IP address on the router, unless a loopback interface with an IP address exists. In this case, the router ID is the highest loopback IP address. The router ID can also be statically configured.
    e. **Optional parameters**: These parameters are Type, Length, and Value (TLV)-encoded. An example of an optional parameter is session authentication.
2. **Keepalive message**: BGP keepalive messages are exchanged between BGP peers often enough to keep the hold timer from expiring. If the negotiated hold-time interval is 0, periodic keepalive messages are not sent. A keepalive message consists of only a message header.
3. **Update message**: A BGP update message has information on one path only; multiple paths require multiple update messages. All the attributes in the update message refer to that path, and the networks are those that can be reached through it. An update message can include the following fields:
    a. **Withdrawn routes**: This list displays IP address prefixes for routes that are withdrawn from service, if any.
    b. **Path attributes**: These attributes include the AS path, origin, local preference, and so on. Each path attribute includes the attribute TLV. The attribute type consists of the attribute flags, followed by the attribute type code.
    c. **Network-layer reachability information**: This field contains a list of IP address prefixes that are reachable by this path.

[13] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Databases, not public document, 26.09.2010

4. **Notification message**: A BGP notification message is sent when an error condition is detected. The BGP connection is closed immediately after this is sent. Notification messages include an error code, an error subcode, and data related to the error.

After a TCP connection is established, the first message sent by each side is an open message. If the open message is acceptable, the side that receives the message sends a keepalive message confirming the open message. After the receiving side confirms the open message and establishes the BGP connection, the BGP peers can exchange any update, keepalive, and notification messages. BGP peers initially exchange their full BGP routing tables. Incremental updates are sent only after topology changes in the network. BGP peers send keepalive messages to ensure that the connection between the BGP peers still exists and send notification packets in response to errors or special conditions.[14]

Selecting a BGP Path

Characteristics of BGP Attributes

BGP routers send BGP update messages about destination networks to other BGP routers. The update messages contain one or more routes and a set of BGP metrics, which are called path attributes, attached to the routes. An attribute is either well-known or optional, mandatory or discretionary, and transitive or nontransitive. An attribute may also be partial. Not all combinations of these characteristics are valid. Path attributes fall into the following four categories:[15]

1. Well-known mandatory,
2. Well-known discretionary,
3. Optional transitive,
4. Optional nontransitive.

Only optional transitive attributes can be marked as partial. All BGP routers must recognize a well-known attribute and propagate it to the other BGP neighbors. Well-known attributes are either mandatory or discretionary. A well-known mandatory attribute must be present in all BGP updates. A well-known discretionary attribute does not have to be present in all BGP updates. Attributes that are not well-known are called optional. BGP routers do not have to support an optional attribute. Optional attributes are either transitive or nontransitive. The following statements apply to optional attributes: [16]

1. BGP routers that implement the optional attribute may propagate it to the other BGP neighbors, based on its meaning.
2. BGP routers that do not implement an optional transitive attribute should pass it to other BGP routers untouched and mark the attribute as partial.
3. BGP routers that do not implement an optional nontransitive attribute must delete the attribute and must not pass it to other BGP routers.

[14] Cisco CCNP Curriculum v5 Module 6 Chapter BGP Message Types, not public document, 26.09.2010
[15] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting BGP Path, not public document, 26.09.2010
[16] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting BGP Path, not public document, 26.09.2010

BGP Attributes

The following is a list of the common BGP attributes according to categories that they belong to: [17]

1. Well-known mandatory attributes,
    a. Autonomous system path,
    b. Next hop,
    c. Origin,
2. Well-known discretionary attributes,
    a. Local preference,
    b. Atomic aggregate,
3. Optional transitive attribute,
    a. Aggregator,
4. Optional nontransitive attribute,
    a. Multi-exit discriminator (MED).

Cisco defines a weight attribute for BGP. The weight is configured locally on a router and is not propagated to any other BGP routers. The atomic aggregate and aggregator attributes relate to BGP summarization (or aggregation). [18]

AS Path Attribute

The AS path is a well-known mandatory attribute. Whenever a route update passes through an autonomous system, the autonomous system number is prepended (added) to that update when it is advertised to the next EBGP neighbor. The AS path attribute is actually the list of autonomous system numbers that a route has traversed to reach a destination, with the number of the autonomous system that originated the route at the end of the list. [19]

Next-Hop Attribute

The BGP next-hop attribute is a well-known mandatory attribute that indicates the next-hop IP address that is to be used to reach a destination. BGP routes autonomous system by autonomous system, not router by router. The next-hop attribute defines the IP address of the border router that should be used as the next hop to the destination. For EBGP, the next hop is the IP address of the neighbor that sent the update. For IBGP, the protocol states that the next hop that is advertised by EBGP should be carried into IBGP. [20]

Origin Attribute

The origin attribute defines the origin of the path information. The origin attribute can be one of these three values: [21]

1. **IGP**: The route is interior to the originating autonomous system. This value normally results when the network command is used to advertise the route via BGP. An origin of IGP is indicated with an „i‟ in the BGP table.

[17] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting BGP Path, not public document, 26.09.2010
[18] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting BGP Path, not public document, 26.09.2010
[19] Cisco CCNP Curriculum v5 Module 6 Chapter AS Path Attribute, not public document, 26.09.2010
[20] Cisco CCNP Curriculum v5 Module 6 Chapter Next-Hop Attribute, not public document, 26.09.2010
[21] Cisco CCNP Curriculum v5 Module 6 Chapter Origin Attribute, not public document, 26.09.2010

2. **EGP**: The route has been learned via EGP. This value is indicated with an „e‟ in the BGP table. EGP is considered a historical routing protocol and is not supported on the Internet because it performs only classful routing and does not support classless interdomain routing.
3. **Incomplete**: The origin of the route is unknown or has been learned by some other means. This value usually results when a route is redistributed into BGP. An incomplete origin is indicated with a question mark (?) in the BGP table.

Local Preference Attribute

Local preference is a well-known discretionary attribute that provides an indication to routers in the autonomous system about which path is preferred to exit the autonomous system. A path with a higher local preference is preferred. The local preference is an attribute that is configured on a router and exchanged among routers within the same autonomous system only. The default value for local preference on a Cisco router is 100. [22]

MED Attribute

The MED attribute, also called the metric, is an optional nontransitive attribute. The MED is an indication to EBGP neighbors about the preferred path into an autonomous system. The MED attribute is a dynamic way to influence another autonomous system about which path that it should choose to reach a certain route in their autonomous system when multiple entry points exist. A lower metric is preferred. Unlike local preference, the MED is exchanged between autonomous systems. The MED is sent to EBGP peers. Those routers propagate the MED within their autonomous system, and the routers within the autonomous system use the MED but do not pass it on to the next autonomous system. When the same update is passed on to another autonomous system, the metric is set back to the default of 0. MED influences inbound traffic to an autonomous system, and local preference influences outbound traffic. By default, a router compares the MED attribute only for paths from neighbors in the same autonomous system. The MED attribute means that BGP is the only protocol that can affect how routes are sent into an autonomous system. [23]

Weight Attribute

The weight attribute is a Cisco attribute for path selection. The weight is configured locally on a router and is not propagated to any other routers. This attribute applies when you are using one router with multiple exit points in autonomous system, as opposed to the local preference attribute, which is used when two or more routers provide multiple exit points. The weight can have a value from 0 to 65535. By default, paths that the router originates have a weight of 32768, and other paths have a weight of 0. Routes with a higher weight are preferred when multiple routes exist to the same destination. [24]

Determining the BGP Path Selection

Multiple paths may exist to reach a given network. As paths for the network are evaluated, those determined not to be the best path are eliminated from the selection criteria but are kept in the BGP forwarding table (which can be displayed using the show ip bgp command) in the event that the best path becomes inaccessible. BGP is not designed to perform load balancing. Paths are chosen because of policy, not based on bandwidth. The BGP selection process eliminates any multiple paths until a

[22] Cisco CCNP Curriculum v5 Module 6 Chapter Local Preference Attribute, not public document, 26.09.2010
[23] Cisco CCNP Curriculum v5 Module 6 Chapter MED Attribute, not public document, 26.09.2010
[24] Cisco CCNP Curriculum v5 Module 6 Chapter Weight Attribute, not public document, 26.09.2010

single best path is left. The best path is submitted to the routing table manager process and is evaluated against any other routing protocols that can also reach that network. The route from the source with the lowest administrative distance is installed in the routing table. The decision process is based on the attributes described earlier. [25]

Selecting a BGP Path

After BGP receives updates about different destinations from different autonomous systems, it chooses the best path to reach a specific destination. The decision process is based on the BGP attributes. BGP considers only synchronized routes with no autonomous system loops and a valid next hop. [26]

The following process summarizes how BGP chooses the best route on a Cisco router: [27]

1. Prefer the route with the highest weight. (The weight attribute is proprietary to Cisco and is local to the router only.)
2. If multiple routes have the same weight, prefer the route with the highest local preference value. (The local preference is used within an autonomous system.)
3. If multiple routes have the same local preference, prefer the route that the local router originated. A locally originated route has a next hop of 0.0.0.0 in the BGP table.
4. If none of the routes were locally originated, prefer the route with the shortest autonomous system path.
5. If the autonomous system path length is the same, prefer the lowest origin code (IGP < EGP < incomplete).
6. If all origin codes are the same, prefer the path with the lowest MED. (The MED is exchanged between autonomous systems.) The MED comparison is made only if the neighboring autonomous system is the same for all routes considered, unless the bgp always-compare-med command is enabled. Internet Engineering Task Force (IETF) decision regarding BGP MED assigns a value of infinity to the missing MED, making the route lacking the MED variable the least preferred. The default behavior of BGP routers running Cisco IOS software is to treat routes without the MED attribute as having a MED of 0, making the route lacking the MED variable the most preferred. To configure the router to conform to the IETF standard, use the bgp bestpath missing-as-worst command.
7. If the routes have the same MED, prefer external paths to internal paths.
8. If synchronization is disabled and only internal paths remain, prefer the path through the closest IGP neighbor, which means that the router prefers the shortest internal path within the autonomous system to reach the destination (the shortest path to the BGP next hop).
9. For EBGP paths, select the oldest route to minimize the effect of routes going up and down (flapping).
10. Prefer the route with the lowest neighbor BGP router ID value.
11. If the BGP router IDs are the same, prefer the router with the lowest neighbor IP address.

[25] Cisco CCNP Curriculum v5 Module 6 Chapter Determining the BGP Path Selection, not public document, 26.09.2010

[26] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting a BGP Path, not public document, 26.09.2010

[27] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting a BGP Path, not public document, 26.09.2010

Only the best path is entered in the routing table and propagated to the BGP neighbors of the router. [28]

The threats and how to mitigate them

Summary of Principal Risks

1. Router failure or impaired performance.
2. Blackholing internet traffic to some addresses.
3. Isolation of internet networks or subnets.
4. Traffic misdirection/route hijack.
5. Eavesdropping.

All of these have cost implications for the providers, as well as Coalition for Networked Information (CNI)[29] security implications, and all also carry a risk of loss of reputation/customer confidence.[30]

Threats of BGP

May be generally categorised as accidents, insider, attack via a weaker peer or collateral damage from other activity.

1. Many of these effects can be caused by accidental mis-configuration, either in the core network, or a directly or indirectly connected network.
2. There is a threat from a subverted disaffected or malicious system administrator, who could deliberately carry out routing mis-configuration, in an effort to attack either this network, or a network in another Autonomous System (AS).
3. A privileged individual in another AS (customer or peer) could use their position to launch an attack (denial of service) against the AS or against the internet generally.
4. A network-based attack from elsewhere on the internet, for example side effects of virus attacks causing excess BGP traffic, due to increased traffic loads. A similar effect can be seen from some other IP attacks.
5. Deliberate BGP attack via a poorly configured ISP from elsewhere on the Internet.
6. Network-based attack on one or more routers within the AS (whether focussed, e.g. telnet/SNMP, or unfocussed e.g. SYN flood) is not directly a BGP vulnerability, but by creating instability within the AS may give rise to route flapping as seen from outside. This in turn may cause BGP instability. Also, e.g. flooding attack on BGP TCP port could disable BGP, though this is really a TCP/IP attack.[31]

Some examples and ideas how to attack BGP

First of all, unlike other routing protocols we have discussed previously, BGP is running over TCP. Thus, a remote intrusion into BGP routing will require guessing correct TCP numbers to insert data. Modern routers' TCP/IP stacks usually have hard-to-predict or unpredictable TCP sequence numbers,

[28] Cisco CCNP Curriculum v5 Module 6 Chapter Selecting a BGP Path, not public document, 26.09.2010
[29] http://www.cni.org/, 27.09.2010
[30] http://www.cpni.gov.uk/Docs/Border_Gateway_Protocol_v2.pdf, 26.09.2010
[31] http://www.cpni.gov.uk/Docs/Border_Gateway_Protocol_v2.pdf, 26.09.2010

eliminating such opportunities. Then, to participate in BGP routing, a rogue router must be defined in the target's BGP configuration as a neighbor with a correct netmask and AS number. Thus, a blind attacker—someone on a remote network without any opportunity to sniff the wire and run ARP (or similar) IP spoofing and TCP hijacking attacks—has little, if any, chance to hack BGP.

The most efficient way to succeed in attacking BGP routing is to take over one of the BGP peers (speakers) and reconfigure it. This is not as easy as you might expect, since backbone routers running BGP are usually looked after and reasonably well protected from external intruders. Of course, a truly massive "I'm on DShield!" type of a scan will eventually discover a few insecure BGP speakers, but this is not the kind of attack a dedicated, focused, considerate Black Hat is seeking. The second approach is to become a semi-blind attacker who seeks to "own" a host on the same network with the targeted BGP routers. This opens up a chance to inject malicious updates by combining both ARP spoofing and TCP hijacking techniques aimed at positioning the hacked host between the BGP speakers and inserting an update with a correct TCP sequence. Also, if MD5based authentication of BGP packets is in use (as it should be), the ability to sniff these packets makes MD5 cracking a more realistic task.

The attacks from the semi-blind attacker position are the main topic of this section. Internal BGP (iBGP), running within a single AS, is more susceptible to semi-blind attacks since there is a higher probability of taking over a host within that AS and close to a BGP router. External BGP (eBGP) is sometimes run over a dedicated line between two peers belonging to different autonomous systems, making such attacks impossible.

An additional factor to consider is synchronization. Unless the autonomous system in question is not a transit domain carrying BGPv4 updates between other ASs, or all routers in a transit domain AS are running BGP and are fully meshed, iBGP must be synchronized with an interior gateway protocol (IGP—such as EIGRP or OSPF). In other words, an IGP becomes responsible for routing BGP updates through the AS. By attacking this IGP, an attacker can indirectly, but strongly, influence BGP routing. And, of course, if route redistribution is in use, routes injected into the redistributed IGP will become injected into BGP. This is when the methodology of attacking becomes somewhat of an art as well as a deep understanding of the technology involved.[32]

Countermeasures Against Attacking BGPv4

1. BGP MD5-based authentication. Use MD5 authentication on peering links wherever practical, using appropriate password strength.[33]
2. Multiple forms of packet filtering from simple blocking of unauthorized hosts access to TCP port 179 with extended ACLs to long Bogon prefix filters,
3. BGP TTL hack to counter man-in-the-middle attacks,
4. Route flap dampening,
5. Layer 2 and ARP-related defenses on shared media,
6. Great but little implemented extensions to the BGP itself, including Secure BGP (S-BGP), Secure origin BGP (so-BGP), and Pretty Secure BGP (psBGP),[34] [35]

[32] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[33] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[34] http://www.ece.cmu.edu/~adrian/731-sp04/readings/KLMS-SBGP.pdf, 29.09.2010
[35] http://people.scs.carleton.ca/~paulv/papers/tissec-july07.pdf, 29.09.2010

7. Interdomain Route Validation (IRV) Service adopts a validation server as an authentication authority. BGP speaker will query IRV server for validation of routing information.[36]
8. Protecting BGP speakers from all types of intrusion.[37]
9. Have a robust process for applying manufacturers patches/workarounds for reported vulnerabilities. [38]

Other suggestions for improving BGP security

1. MD5 password strength & management issues. There are a variety of recommendations for best practice in MD5 use. It is certainly true that MD5 cracking tools are available in the hacker community. Currently, most providers take the pragmatic view that MD5 implementation (with well-chosen, non dictionary-based passwords) enhances security, even where some details of the implementation could be further improved.[39]
2. Egress filtering. Ideally, filtering should be both on ingress & egress. Providers should encourage their customers to perform egress filtering, according to these guidelines. Outbound route filtering can provide an alternative to customer egress filtering. [40]
3. BGP TTLH (time to live hack) also known as BTSH (BGP Time-to-live security Hack) – suggests that routers set TTL to be 255, and only accept BGP packets with TTL 254, as the peer is always exactly 1 hop away. This introduces extra difficulty for an attacker, compared to the default of expecting the TTL to be 1. TTLH/BTSH is principally designed to protect against DOS attacks flooding port 179. This is not widely implemented (Juniper JUNOS implements this feature). [41]
4. Filtering of incoming BGP packets on entry, before passing to the CPU, can also protect the router from DoS via port 179. Implementation is not even across vendors, feedback from providers on which implementations are most successful would be welcome. [42]
5. Graceful restart. A facility where a router preserves its forwarding state through a restart (time-limited to e.g. 30 seconds), so eliminating the need for peers to “flap” all its routes. Implemented by Juniper, Cisco (as Cisco non-stop forwarding) & Riverstone, among others. Several providers (US) suggest that the cost of implementing this feature outweighs the benefit. [43]
6. Sink-holes to protect customer networks. The ability to set these up in response to an event/attack is highly desirable, and will provide a valuable service to CNI providers in non-communications sectors. [44]
7. AS-Path filtering. In addition to prefix filtering, some providers recommend using AS-Path filtering, to drop any announcements with private AS numbers in the AS path, and setting a max-as-path-length. [45]

[36] http://web.it.kth.se/~xuan/reports/securing%20interdomain%20routing.pdf, 27.09.2010
[37] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[38] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[39] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[40] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[41] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[42] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[43] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[44] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[45] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010

8. Use of BGP MED values has been suggested as a means to enhance security, however, this can have a variety of effects. Accepting MED values from Peers may introduce a (small) extra risk. [46]
9. S-BGP aims to set up a PKI authentication scheme to authenticate BGP peering & announcements. Lack of backward compatibility with BGPv4 is a problem. [47] [48]
10. soBGP. Developed by Cisco. Aims to use PKI to authenticate the origin of BGP packets, but not the peering connections. Security increases proportionally with adoption, but it is compatible with non-secure BGPv4. [49] [50]

Border Gateway Protocol Filtering Guidelines

These guidelines identify generally accepted practices for Border Gateway Protocol (BGP) filtering. The implementation of these practices must be viewed in the context of the whole system.

Generally Accepted Practices

Deny special prefixes assigned and reserved for future use. These are described in:

1. rfc5735[51]
    a. 0.0.0.0/8 reserved for self-identification,
    b. 127.0.0.0/8 is reserved for Loopback,
    c. 169.254.0.0/16 reserved for Link Local,
    d. 198.18.0.0/15 reserved for Network Interconnect Device Benchmark Testing,
2. rfc1356 This was reserved for Public Data Networks,
3. rfc5737 [52]
    a. 192.0.2.0/24 reserved for TEST-NET-1,
    b. 198.51.100.0/24 reserved for TEST-NET-2,
    c. 203.0.113.0/24 reserved for TEST-NET-3,
4. rfc3068 192.88.99.0/24 reserved for 6to4 Relay Anycast, [53]
5. rfc5736 192.0.0.0/24 reserved for IANA IPv4 Special Purpose Address Registry, [54]
6. rfc3171 Multicast (formerly "Class D"),[55]
7. rfc1112 Reserved for future use (formerly "Class E"),[56]
8. rfc0919 and rfc0922 255.255.255.255 is reserved for "limited broadcast" destination address.[57]

Deny unallocated (grey/bogon) space. This address space has no special restrictions. It simply not jet officially allocated to anyone. Allowing and forwarding traffic from unallocated address space makes

[46] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[47] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[48] http://www.ece.cmu.edu/~adrian/731-sp04/readings/KLMS-SBGP.pdf, 29.09.2010
[49] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[50] http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_6-3/securing_bgp_sobgp.html, 29.09.2010
[51] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[52] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[53] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[54] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[55] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[56] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010
[57] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml#note2, 27.09.2010

spammers easier to send spam and makes virus and malware writer to deliver malware and viruses. Unallocated address space is massively used by spammers and malware writers. The rules how to allocate address space is described rfc1466[58] and list of Ipv4 address allocation is shown IANA website.[59]

Deny over-specific prefix lengths. This measurement helps to avoid attack which misuse routing decision-making logic – more precise route is preferred over less precise, based on subnet prefix length.[60]

1. Announcing only those networks we specifically list with neighbor <IP address> prefix-list announce out. This would also prevent the network from becoming a transit provider and is a useful countermeasure to add for nontransit Ass,
2. Block inbound Bogons with a prefix list: neighbor <IP address> prefix-list bogons in,
3. Set up appropriate route distribute lists: neighbor <ip-address | peer-group-name> distribute-list <access-list-number | name> <in | out>.[61]
4. Deny prefixes > /28, [62]
5. Deny prefixes < /6.[63]

Maintain route flap dampening default settings or set according to RIPE parameters. The route dampening prevents routers from thrashing while trying to re-calculate a large number of route updates. The overall effect is to produce a more stable routing table. BGP version 4, the BGP process assigns a penalty of 1000 to the route each time it flaps. When the penalty value exceeds the first of two limits, the route is moved into the 'historical' list of routes, dampened, and then is no longer accepted from other peers or announced to any peers. After the first limit has been exceeded, the timer which tracks the period for which the route is to be dampened is doubled for each flap.[64],[65]

As for BGP flapping route dampening, enable it even though, as mentioned earlier, the attackers may actually abuse it. However, a proper BGP dampening configuration goes well beyond the `bgp dampening` command and involves creation of route maps and a variety of prefix lists to reduce the effect of dampening on the shorter and historically more stable prefixes, as well as IP ranges that contain DNS root servers. This is done in accordance to the RIPE recommendations on flapping route dampening safety. Ready and working examples of these prefix lists and route maps can be taken directly from the Team Cymru Secure BGP Template (*http://www.cymru.com/Documents/secure-bgp-template.html*).[25]

[58] http://tools.ietf.org/html/rfc1466, 27.09.2010
[59] http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml, 27.09.2010
[60] http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094823.shtml#prefix, 27.09.2010
[61] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[62] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[63] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[64]http://www.inetdaemon.com/tutorials/internet/ip/routing/bgp/operation/bgp_route_flap_dampening.shtml 27.09.2010
[65] http://www.faqs.org/rfcs/rfc2439.html, 27.09.2010

Aggregate routes where possible. Border Gateway Protocol (BGP) allows the aggregation of specific routes into one route.[66] Set the limit on the amount of advertised prefixes to prevent de-aggregation type of attacks: neighbor <IP address> maximum-prefix 175000. Deny overspecific prefix lengths (de-aggregation, "traffic sucking").[67] Suggested: all routes more specific than /20.[68]

Use the loopback interface for iBGP announcements to increase the iBGP stability: neighbor <IP address> update-source Loopback0. [69]

Deny exchange point prefixes. For example Estonian local exchange point is TIX. The peering ISPs at the IXP exchange prefixes they originate. Sometimes they exchange prefixes from neighbouring ASNs too. Be aware that the IXP border router should carry only the prefixes you want the IXP peers to receive and the destinations you want them to be able to reach. Otherwise they could point a default route to you and unintentionally transit your backbone. If IXP router is at IX, and distant from your backbone don't originate your address block at your IXP router.[70]

Deny routes to internal IP spaces. According to rfc1918 the following subnets are assigned for local se and not forward over public Internet:

1. 10.0.0.0/8 Reserved for Private-Use Networks,
2. 172.16.0.0/12 Reserved for Private-Use Networks,
3. 192.168.0.0/16 Reserved for Private-Use Networks,

Forwarding gives intruder additional vector to attack inside network. Possible threat: routing loop, MITM-attack. [71]

Minimize BGP use with customers. Use static routes instead of BGP peering with customers (single homed customers, customers who do not themselves have downstream customers). BGP peering gives customers extra attack vector. For ISP it is additional threat. [72]

Restrict routes exchanged with customers to those concerning customers declared IP space. [73]

Restrict routes exchanged with peers, depending on relationship with peers and between peers. This may not be feasible depending on the size of the peers table and access to correct Autonomous System/route-sets among peers.[74]

Finally, do not go berserk over an accidentally lost BGP keepalive packet and turn bgp fast-external-fallover off. This can help to withstand DoS floods without flapping the routes. Also, do not

[66] http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094826.shtml#intro, 27.09.2010
[67] http://www.cisco.com/en/US/tech/tk365/technologies_tech_note09186a0080094823.shtml#prefix, 27.09.2010
[68] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[69] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[70] http://www.sanog.org/resources/sanog7/pfs-bgp-multihoming.pdf, 27.09.2010
[71] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[72] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[73] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[74] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010

forget to enable bgp log-neighbor-changes to see whether your BGP neighborhood is OK when running an occasional show logging command.[75]

Implementation of these practices should be considered in the context of existing relationships with customers and business partners and customer requirements.[76]

Other Filtering Practices

The following guidelines are presented for consideration as additional practices to improve BGP security.

Deny inappropriate length RIR allocations. [77]

Deny inappropriate announcements for legacy A/B/C space (we use CIDR). [78] IP subnets:

1. Class A 0.0.0.0-127.255.255.255/8,
2. Class B 128.0.0.0-191.255.255.255/16,
3. Class C 192.0.0.0-223.255.255.255/24.[79]

Deny over-general prefix lengths. Example do not allow supernet A class addresses with les than 8 bit subnetmask. Deny over-general prefix lengths (this is more likely to be an error than an attack).[80] Set max-prefix limits on IXP and customer peerings. [81]

Defend critical networks (backbone, specific customers) by defining different network levels, such as platinum, gold, and silver, and allowing only certain routes on each level. General consensus is that the costs/risks associated with this practice may outweigh the benefit.[82]

Deny route flap dampening on "gold" and "platinum" networks.[83] General consensus is that network operators should address route dampening. Such high level BGP peerings are stable and peers are between trustful parties (big companys Level3, Cogent, AT&T etc). [84]

Implement BGP graceful restart (e.g. Cisco non-stop forwarding or other Cisco GRIP features). General consensus is that the costs and complexity introduced by this guideline may greatly exceed the risk associated with BGP vulnerabilities. [85]

[75] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[76] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[77] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[78] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[79] http://en.wikipedia.org/wiki/Classful_network, Legacy info, classful networks not used anymore, 28.09.2010
[80] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[81] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[82] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[83] http://www.iphelp.ru/faq/35/0086.html, 27.09.2010
[84] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010
[85] http://www.cpni.gov.uk/Docs/re-20040401-00392.pdf, 29.09.2010