# Border Gateway Protocol (BGP)

# 1 What is Border Gateway Protocol (BGP)?

Border Gateway Protocol (BGP) is a routing protocol used on the edge of autonomous systems (AS). It is an exterior routing protocol and calculates loop-free paths across the Internet. It is considered to use a path-vector routing algorithm. This means it tracks the path in terms of which AS it passes through, and does NOT track the 'route' through individual routers within an AS, and is not specifically capable of performing load balancing or packet forwarding itself. BGP is the routing protocol of choice and is used by all the Network Service Providers (NSPs) such as UUNet, Sprint, Cable & Wireless, Level3, Qwest etc. It is dynamic and handles outages and link failures fairly gracefully. To use BGP, you must have a router that supports BGP; register an AS Number and contact your provider to set up a BGP session. See the requirements page for more information.

BGP has gone through three revisions. The current version in use is BGP4 and is supported by most router manufacturers including Cisco, Lucent/Bay, Juniper and many others, as well as by Unix and Linux programs such as Zebra.

BGP uses a TCP connection to send routing updates using TCP port 179. BGP is therefore by definition a 'reliable' protocol. While BGP version 3 provides for the dynamic learning of routes, BGP 4 adds additional route dampening functionality, communities, MD5 and multicasting capability.

## 1.1 External vs. Internal Peers (eBGP vs iBGP)

Peering is when you exchange routes with another BGP speaking device. There are two types of peering sessions:

# INTERNAL PEERS (iBGP)

An Internal peer is a BGP speaking neighbor who has the *same* Autonomous System (AS) number as you do. An internal peer will only pass on the best routes it knows from it's own connections.

# EXTERNAL PEERS (eBGP)

External peers have *different* Autonomous System (AS) numbers. An external peer will pass on all the best routes it knows or has learned from any other peer to all other directly connected external peers. This is what I mean when I say eBGP is a 'gossipy' protocol. Routers speaking eBGP gab everything they know to their neighbors unless you install a gag (a route filter).

## 2 BGP and Autonomous Systems

An autonomous system is one network or sets of networks under a single administrative control. An autonomous system might be the set of all computer networks owned by a company, or a college. Companies and organizations might own more than one autonomous system, but the idea is that each autonomous system is managed independently with respect to BGP. An autonomous system is often referred to as an 'AS'.

A good example is UUNet, who uses one autonomous system as their Eurpoean network, and a separate autonomous system for their domestic networks in the Americas.

# AUTONOMOUS SYSTEM NUMBERS

The American Registry for Internet Numbers (ARIN) defines Autonomsous System Numbers as:

"*Autonomous System Numbers (ASNs) are globally unique numbers that are used to identify autonomous systems (ASes) and which enable an AS to exchange exterior routing information between neighboring ASes. An AS is a connected group of IP networks that adhere to a single and clearly defined routing policy.*"

To identify each autonomous system, a 'globally unique' number is assigned to them from a centralized authority (ARIN) so that there are no duplicate numbers. Globally Unique means exactly that. Within the entire Internet all around the globe, the AS number should be unique. The AS number will be from 1 to 64511, and the next highest unused number is what is generally assigned. These numbers are referred to as 'AS numbers'. The American Registry for Internet Numbers (ARIN) is the authority responsible for tracking and assigning these numbers as well as managing IP address allocations and assignments. ARIN charges a fee to organizations wishing to obtain an AS number to cover the administrative costs associated with managing AS number registrations and assignments.

To receive an AS number from ARIN, you must be able to prove you are 'dual homed' to the Internet, which means that you have more than one Internet provider with which you plan to run BGP. You must also have a 'unique routing policy' that differs from your BGP peers. Some companies have difficulty getting an AS number.

Here is a short list of the top (per Caida's Skitter Plot April 2003) ISP's system numbers. You can always go to ARIN's website to look them up.

| AS # | Provider |
|---|---|
| 701 | UUnet (U.S. domestic) (AS 701-705) |
| 1239 | Sprintlink U.S. Domestic |
| 3356 | Level 3 |
| 7018 | AT&T WorldNet |
| 209 | Qwest |
| 3561 | Cable and Wireless (aq'd by SAVVIS) |
| 3549 | Global Crossing |

| | |
|---|---|
| 2914 | Verio |
| 6461 | AboveNet |
| 702 | UUnet (International) |
| 1299 | TeliaNet |
| 5511 | OpenTransit |
| 5459 | LINX |
| 16631 | Cogent |
| 6453 | Teleglobe |

# PRIVATE AS NUMBERS (64512 - 65535)

If it is not necessary to connect to the Internet, or you are part of a special type of BGP configuration you can use any of the AS numbers 64512 through 65535. However, these numbers should NOT be seen on the global Internet. One example of when you might use private AS numbers is in BGP confederations. The confederation AS number should not be seen on the global Internet.

# AS NUMBERS AND BGP

BGP learns and exchanges path information regarding the route to a given destination network by keeping lists of AS numbers and associating them with destination networks. This is why AS numbers should be unique. BGP makes certain that an AS number does not appear in a path more than once, thereby preventing routing loops.

# 3 BGP and Classless Inter-Domain Routing (CIDR)

For years, the Internet has been growing at an alarming pace. The proliferation of the 'dot com' and e-commerce companies caused a huge surge in the use of Internet Protocol (IP) addresses, and an increase in the number of destinations on the Internet. Each destination has a range of IP addresses associated with it (a 'block of IP's' in *NetSpeak*).

Routers use routing protocols such as BGP exchange information about how to reach these destinations. As the number of destinations grew, so did the number of routes (paths) to reach these networks. Soon, it became clear that the routers couldn't store the growing number of routes, they also couldn't handle the optimum path calclulations much longer, and the IP address space was being handed out far too quickly because it was carved into large classful blocks.

**CIDR: SUBNETTING, SUPERNETTING AND ROUTE AGGREGATION**

The soluton to the depletion of IP addresses and the proliferation of the number of routes was two fold. ARIN used a three-pronged solution to this problem.

To reduce the number of routes in the Internet backbone, *supernetting* was used to aggregate destinations together into larger blocks of IP's. Larger blocks of IP's would only be allocated to the ISP's. The ISP's were, in turn, expected to aggregate their routes so that routing protocols could make fewer announcements of larger blocks. Of course, this created other problems for customers who were wise enough to not count on a single ISP for their Internet access...

**ROUTE AGGREGATION**

ARIN recommended that all backbone Internet providers combine or 'aggregate' their routes to reduce the total number of routes being advertised. ARIN further recommended that all companies and organizations wishing to connect to the Internet request IP addresses from their Internet provider instead of ARIN.

If a company gets a small block of IP addresses from the larger block of IP addresses owned by their upstream provider, their provider can more easilly aggregate those routes because the customer's IP addresses originally came from the larger block of IP addresses owned by the Internet provider.

# 4 BGP and Classless Inter-Domain Routing Notation (CIDR Notation)

BGP uses Classless Inter-Domain Routing (CIDR) notation for masks. When advertising routes, BGP will include prefixes in it's advertisements. A prefix is the network IP address plus the mask in CIDR notation.

Below are tables showing how IP masks and CIDR masks are related. In every IP address, certain bits are used to identify the network, and certain bits are used for host. The mask allows the receiver to tell which bits are network, and which are host. Ones are used to mark the Network bits, and zeroes are used to mark the Host bits.

| | CLASS 'A' NETWORKS |
|---|---|
| BINARY | 11111111.00000000.00000000.00000000 |
| DECIMAL | 255.0.0.0 |
| CIDR | /8 |

| | CLASS 'B' NETWORKS |
|---|---|
| BINARY | 11111111.11111111.00000000.00000000 |
| DECIMAL | 255.255.0.0 |
| CIDR | /16 |

| | CLASS 'C' NETWORKS |
|---|---|
| BINARY | 11111111.11111111.11111111.00000000 |
| DECIMAL | 255.255.255.0 |
| CIDR | /24 |

You will note that the number in the CIDR mask notation is equal to the number of 1's in the binary mask. The same holds true for subnets:

| | 1/2 CLASS 'C' NETWORK |
|---|---|
| BINARY | 11111111.11111111.11111111.10000000 |
| DECIMAL | 255.255.255.128 |
| CIDR | /25 |

And for supernets as well:

| | 2 CLASS 'C' NETWORKS |
|---|---|
| BINARY | 11111111.11111111.11111110.00000000 |
| DECIMAL | 255.255.254.0 |

| | |
|---|---|
| CIDR | /23 |

## 5 BGP Peering

Peering is the term used for connections between BGP speakers. Technically, any two routers running BGP which have different AS numbers are peering. However, in the Telecommunications and Internet provider arena, the term 'peering' has taken on a politically-loaded meaning that is unrelated to the technical process and more related to the politics of which ISP is paying whom for the connection.

# PUBLIC PEERING

Originally, four main locations were established where anyone who could afford to pay the local telephone company for a data circuit could connect. In those days there were so few connection points between networks that everyone who connected to the public peering points set up a BGP session to everyone else connected there. This increased access between networks dramatically, but this created other problems, such as route flaps and congestion.

# PRIVATE PEERING

Now here's where the politics comes in. Today, the majority of peering connections are private, and are paid connections. These connections carry public Internet user's traffic; however, they become overutilized over time. The connections are only upgraded when the company that is purchasing the connection pays for a bigger circuit.

The problem is, each ISP will always blame the other, leaving the customer in the middle to sort it out himself.

# THE POLITICAL PROBLEM

People don't pay to use an Internet Provider's network, they pay for access to Internet destinations. A provider who won't peer isn't willing to connect to destinations such as Yahoo, Google, Microsoft and AOL. **The Internet belongs to noone. The Internet is not the property of any ISP.** The Internet is the combination of all the public networks on the planet. Therefore, connectivity between one provider's network and everyone else's is the prime capital investment any Internet company can make. Unfortunately, the managers of these companies are not, and never have been technical people. They don't understand this basic fact and go out of their way to shut down connections to outside networks because they are 'unprofitable', even when the connection is matched by an equal connection installed by the other provider to handle traffic in the opposite direction.

# 6 BGP Operation

1. Finite State Model
    1. Idle
    2. Connect
    3. Active
    4. Open Sent
    5. Open Confirm
    6. Established

# 6.1 BGP Finite State Model

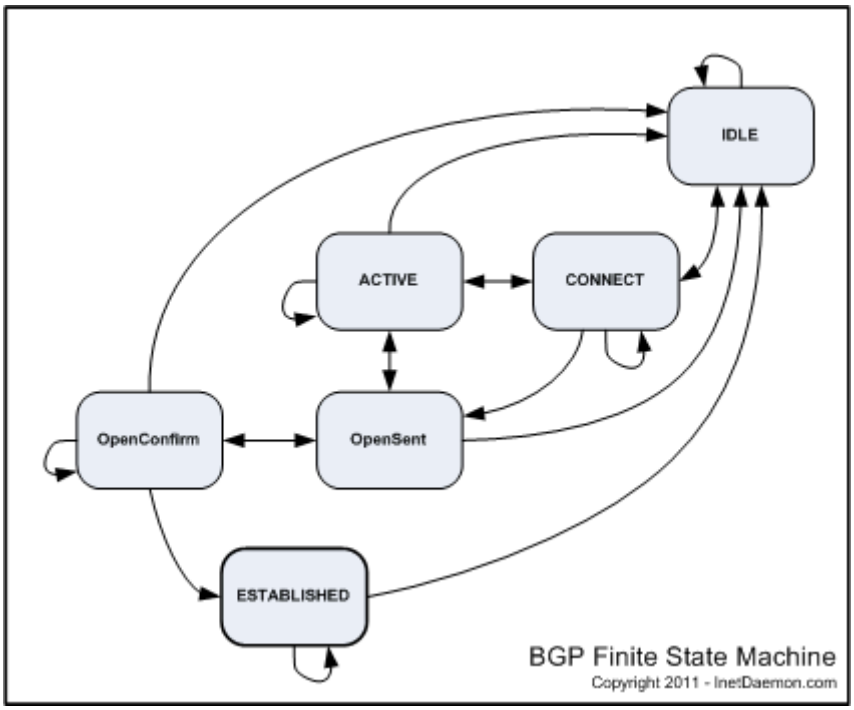
Last Updated: Wednesday, 03-Apr-2013 12:52:14 MDT | By InetDaemon

Border Gateway Protocol as defined in RFC 4271 (obsoletes RFC 1771) defines what is called a "finite state model" which describes BGP's behavior at routing engine startup and during the establishment of BGP neighbor sessions. The finite-state-machine is a description of what actions should be taken by the BGP routing engine and when. There are *six states* in the model, and there are specific conditions under which each BGPstate will transition to the next during the process of establishing first a TCP connection, and then a BGP session. Each step indicates a different state in the BGP session. For the purpose of this discussion, a router is any device running BGP.

From the OSI Model's perspective, BGP is simply a networking application running on top of the the Session layer and everything below it. Thus, an ESTABLISHED BGP SESSION is required for BGP to begin exchanging routes.

**NOTE:** A valid Transport session via a reliable protocol is required in order to establish a BGP peering session between two neighbors.

BGP will fail to negotiate a peering session if the underlying communications layers fail. Troubleshoot the physical, datalink and network layers first, if the network interfaces are up/down or down/down.




*BGP Finite State Machine*

## 6.2 IDLE

The IDLE state is the initial state of the BGP Finite State Machine on startup. A BGP speaking router inthe IDLE state is awaiting a session it sits in the IDLE state awaiting the ManualStart event or the AutomaticStart event. When either start event is received BGP performs the following:

- Iinitializes all resources for the peer connection
- Sets ConnectRetryCounter to zero
- Starts the ConnectRetryTimer with the initial value
- Initiates a TCP connection to the other BGP peer
- Listens for a connection that may be initiated by the remote BGP peer
- Changes its state to CONNECT.

The BGP router will not start a BGP session until either start event occurs. Cisco classifies initial configuration or clearing of a BGP peering session as a start event and the system transitions to the CONNECT state. Whenever a BGP peering session is shut down because of an error, it returns to the IDLE state. NOTIFICATION messages used to signal connection errors return the router to the IDLE state.

## 6.3 CONNECT

Once the BGP software and it's environment have been initialized, BGP initiates a TCP connection to the remote neighbor IP address. The CONNECT state indicates the router has awaiting the completion of a TCP connection between itself and another BGP speaking peer. The BGP Finite State Machine remains in CONNECT until the TCP three-way handshake completes.

It is assumed that both sides of the connection will attempt to initiate a BGP session with the peer. The peer with the higher router ID (highest IP address) becomes the router that manages the BGP session and the connection attempted by the other router is abandoned.

## 6.4 ACTIVE

The router has started the first phase of establishing a BGP session by initializing a new TCP three-way handshake to the remote router (peer) because the initial connect failed. Typically, you only see this state if you failed the initial connect. From the ACTIVE state, BGP will attempt to send another OPEN message to negotiate a BGP session. If the second attempt fails, the state falls back to CONNECT.

If you check the state of BGP, and it shows ACTIVE, you do NOT have a functional BGP session. The Finite State Machine passes through ACTIVE only when the CONNECT phase fails.

## 6.5 OPEN SENT

At this stage, a TCP connection should be open ( TCP three-way handshake completed) and an OPEN message successfully transmitted by both routers. The BGP OPEN message contains:

- The BGP Version number (Binary: 00000100, Decimal: 4)
- The AS Number
- The Hold Down Time value
- The BGP Identifier (management IP address of the router) and Optional Parameters.

## 6.6 OPEN CONFIRM

BGP confirms that the OPEN message was received, a KEEPALIVE message is transmitted and the BGP state transitions to ESTABLISHED.

## 6.7 ESTABLISHED

After the BGP session parameter negotiation is completed, the routers begin exchanging BGP routes.

**ESTABLISHED IS THE ONLY STATE THAT COUNTS FOLKS!** This is the only state in which BGP will actually exchange routes. If you have any other state, you have a non-functional BGP session (and possibly a broken physical link if it refuses to establish the connection). On a Cisco router, you cannot have an ESTABLISHED BGP session if the network interface is Line Protocol Up/Network Protocol Down.

2. Metrics
3. Timers
4. Advertisements

5. Messages
    1. Open
    2. Keepalive
    3. Notification
    4. Update
        1. Network Layer Reachability Information (NLRI)
        2. AS-Path Attributes
            1. WEIGHT
            2. LOCAL PREFERENCE
            3. AS_PATH
            4. ORIGIN
            5. NEXTHOP
            6. METRIC
            7. MULTI_EXIT_DISC
            8. COMMUNITY
            9. CLUSTER_ID
6. BGP Best Path Selection Algorithm
7. Route Flaps and Route Dampening
8. Cisco Nonstop Forwarding
    1. Routing Information Base
    2. Forwarding Informaton Base

# BGP Finite State Model

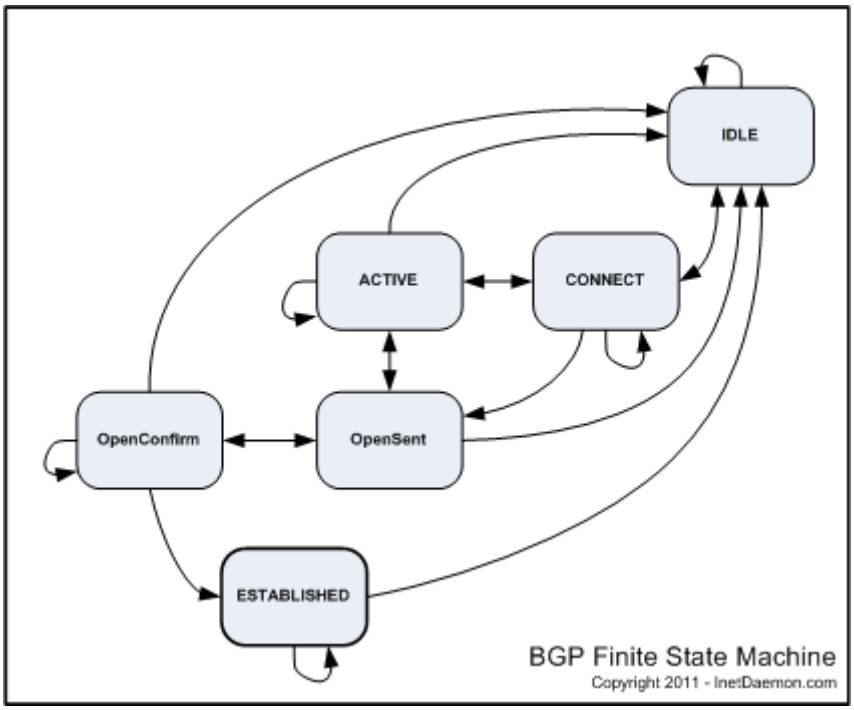
Last Updated: Wednesday, 03-Apr-2013 12:52:14 MDT | By InetDaemon

Border Gateway Protocol as defined in RFC 4271 (obsoletes RFC 1771) defines what is called a "finite state model" which describes BGP's behavior at routing engine startup and during the establishment of BGP neighbor sessions. The finite-state-machine is a description of what actions should be taken by the BGP routing engine and when. There are *six states* in the model, and there are specific conditions under which each BGPstate will transition to the next during the process of establishing first a TCP connection, and then a BGP session. Each step indicates a different state in the BGP session. For the purpose of this discussion, a router is any device running BGP.

From the OSI Model's perspective, BGP is simply a networking application running on top of the the Session layer and everything below it. Thus, an ESTABLISHED BGP SESSION is required for BGP to begin exchanging routes.

**NOTE:** A valid Transport session via a reliable protocol is required in order to establish a BGP peering session between two neighbors.

BGP will fail to negotiate a peering session if the underlying communications layers fail. Troubleshoot the physical, datalink and network layers first, if the network interfaces are up/down or down/down.




*BGP Finite State Machine*

# IDLE

The IDLE state is the initial state of the BGP Finite State Machine on startup. A BGP speaking router inthe IDLE state is awaiting a session it sits in the IDLE state awaiting the

ManualStart event or the AutomaticStart event. When either start event is received BGP performs the following:

- Iinitializes all resources for the peer connection
- Sets ConnectRetryCounter to zero
- Starts the ConnectRetryTimer with the initial value
- Initiates a TCP connection to the other BGP peer
- Listens for a connection that may be initiated by the remote BGP peer
- Changes its state to CONNECT.

The BGP router will not start a BGP session until either start event occurs. Cisco classifies initial configuration or clearing of a BGP peering session as a start event and the system transitions to the CONNECT state. Whenever a BGP peering session is shut down because of an error, it returns to the IDLE state. NOTIFICATION messages used to signal connection errors return the router to the IDLE state.

# CONNECT

Once the BGP software and it's environment have been initialized, BGP initiates a TCP connection to the remote neighbor IP address. The CONNECT state indicates the router has awaiting the completion of a TCP connection between itself and another BGP speaking peer. The BGP Finite State Machine remains in CONNECT until the TCP three-way handshake completes.

It is assumed that both sides of the connection will attempt to initiate a BGP session with the peer. The peer with the higher router ID (highest IP address) becomes the router that manages the BGP session and the connection attempted by the other router is abandoned.

# ACTIVE

The router has started the first phase of establishing a BGP session by initializing a new TCP three-way handshake to the remote router (peer) because the initial connect failed. Typically, you only see this state if you failed the initial connect. From the ACTIVE state, BGP will attempt to send another OPEN message to negotiate a BGP session. If the second attempt fails, the state falls back to CONNECT.

If you check the state of BGP, and it shows ACTIVE, you do NOT have a functional BGP session. The Finite State Machine passes through ACTIVE only when the CONNECT phase fails.

# OPEN SENT

At this stage, a TCP connection should be open ( TCP three-way handshake completed) and an OPEN message successfully transmitted by both routers. The BGP OPEN message contains:

- The BGP Version number (Binary: 00000100, Decimal: 4)
- The AS Number

- The Hold Down Time value
- The BGP Identifier (management IP address of the router) and Optional Parameters.

# OPEN CONFIRM

BGP confirms that the OPEN message was received, a KEEPALIVE message is transmitted and the BGP state transitions to ESTABLISHED.

# ESTABLISHED

After the BGP session parameter negotiation is completed, the routers begin exchanging BGP routes.

**ESTABLISHED IS THE ONLY STATE THAT COUNTS FOLKS!** This is the only state in which BGP will actually exchange routes. If you have any other state, you have a non-functional BGP session (and possibly a broken physical link if it refuses to establish the connection). On a Cisco router, you cannot have an ESTABLISHED BGP session if the network interface is Line Protocol Up/Network Protocol Down.

# 6.8 BGP Session Timers

Last Updated: Wednesday, 03-Apr-2013 12:52:19 MDT | By InetDaemon

There are two primary timers in BGP. The first is the Hold Down timer, the other is the Keepalive Interval.

# 6.9 HOLD DOWN TIMER

Cisco default setting: 180 seconds = 3x Keepalive

The Hold Down Timer indicates how long a router will wait between hearing messages from it's neighbor. The Hold Down Timer defaults to 180 seconds on a Cisco router, but can be reconfigured. The timer starts at zero and counts it's way up to the Hold Down Timer value. If either a Keepalive or Update message is not received in that time, then the router declares the peering session dead, places all routes learned from that peer into a 'dampened' state and attempts to reset the session.

# 6.10 KEEPALIVE INTERVAL

Cisco default setting: 60 seconds

To be certain that a BGP session stays up and functional, Keepalive messages are exchanged. The Keepalive Interval counts down to zero and then sends out another Keepalive. There is no timer for route updates, as updates happen dynamically on an incremental basis.

# BGP Route Advertisements

Route advertisements are not sent out on regular intervals as in other protocols. A full table exchange is sent out when BGP is first started, and then only incremental updates are sent when changes occur in topology.

BGP is a PATH VECTOR protocol, which means that it does not keep track of internal routing within the AS, but rather keeps track of paths through other autonomous systems to reach destination networks. Any network that is not being advertised, cannot be reached.

BGP uses the UPDATE message to advertise routes. It is important to remember that BGP routers listen for routes. A BGP router cannot forward a packet if it has not heard a route. By the same token, if a route isn't being advertised, it is also not possible to forward packets.

The UPDATE message contains:

I. Network Layer Reachability Information (NLRI)
   A. Prefix
   B. Length
II. AS_Path
III. AS_Path Attributes

When a BGP peer becomes unreachable, BGP sends UPDATE messages that contain 'withdrawal' information, requesting that other BGP routers remove those routes from their tables.

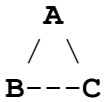
eBGP peers will advertise all known eBGP routes to all other eBGP peers. iBGP peers will only advertise their own internal routes to other iBGP peers. A BGP speaking router will never advertise another iBGP peer's routes to any other iBGP peer.

To clarify, I will use an example. Picture a string of three routers:

A = B = C

1. Routers A and C are iBGP peers of B.
2. Router A tells Router B about it's directly connected networks.
3. Router C tells router B about it's directly connected networks.
4. Router B will never tell A about C's routes, and will never tell C about A's routes.

If it is desirable for A and C to know each other's routes, then either use an Interior Gateway Protocol of some kind (RIP, OSPF, IS-IS, IGRP) to pass this information, or directly connect A and C so that you will have built a *fully meshed* BGP network like so:




1. Routers A and C are iBGP peers of B ( A <-> B <-> C )
2. Router A tells Router B about it's directly connected networks.
3. Router C tells router B about it's directly connected networks.

4. Router B will never tell A about C's routes, and will never tell C about A's routes.

This is a loop-prevention algorithm. The path between A and C should be managed by an interior gateway protocol.

# BGP Messages

BGP messages exchange information and help maintain state between the two routers in the peering session. I will more clearly define these messages later.

## KEEPALIVE

This is the packet used to keep the session running when there are no updates. Keepalives are sent between BGP speakers to let each other know they are still there. When a BGP router fails to hear a Keepalive message, it removes all routes heard from that peer from it's forwarding information base (FIB).

## NOTIFICATION

Notifications are used to send error messages when an update is received but is corrupt, or when the router needs to turn down the session unexpectedly.

## OPEN

Open messages are used to start a BGP session by requesting that a BGP session be opened over an existing TCP/IP session.

## UPDATE

This message type contains the actual route updates. The route updates are composed of the following:

1. Network Layer Reachability Information
2. AS-Path
3. AS-Path Attributes

Updates received are placed in the Routing Information Base (RIB). If a route in an Update message is better than all other routes in the RIB, then that route is placed in the Forwarding Information Base (FIB).

## 6.11 BGP Network Layer Reachability Information (NLRI)

The Network Layer Reachability Information (NLRI) is exchanged between BGP routers using UPDATE messages. An NLRI is composed of a LENGTH and a PREFIX. The length is a network mask in CIDR notation (eg. /25) specifying the number of network bits, and the prefix is the Network address for that subnet.

The NLRI is unique to BGP version 4 and allows BGP to carry supernetting information, as well as perform aggregation.

The NLRI would look something like one of these:

/25, 204.149.16.128
/23, 206.134.32
/8, 10

Only one NLRI is included in an UPDATE Message, though there may be multiple AS-paths and AS-path attributes.

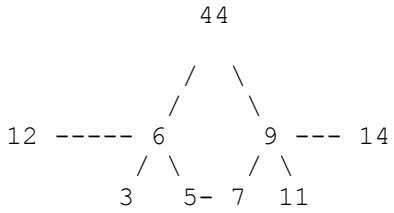
# 6.12 BGP Autonomous System Path (AS-path)

Last Updated: Wednesday, 03-Apr-2013 12:51:57 MDT | By InetDaemon

An Autnonomous System path is a list of all the autonomous systems that a specific route passes through to reach one destination. The AS path is displayed as a series of autonomous system (AS) numbers separated by spaces, with the originator's AS number at the end of the path, and the next AS hop from the current router's location in the beginning of the path.

AS-paths are created when a BGP router receives an announcement from an exterior neighbor. When the router receives the route, it adds the AS number of the exterior neighbor to the AS-Path. As the route announcement passes from autonomous system to autonomous system, the path grows longer with each receiver adding the neighbor's AS to the path.

Here is an illustration. Given the following diagram showing the relationship between several AS you can trace the path between them by listing their AS numbers:




For AS 12 to reach AS 14, it would need to see an AS-Paths containing the following AS numbers tracing the actual path to vector through to reach back through the Internet to the originator of Network Layer Reachability Information for a CIDR block of addresses:

```
6  44  9  14
6   5  7   9  14
```

If we recall that *by default, BGP selects only 1 'best' path to a destination*, as long as both paths are valid (reachable), the route with the fewest hops might be the preferred route. Since the first path has only 4 hops, it would be the preferred route, but only if the AS hop count attribute were the tie breaker between the routes. As a point of fact, the selection of the best path rarely comes down to AS-path hop count as it is fairly low in the list of 'tie breaker' decisions in the BGP Best Path Algorithm. Local Preference tends to be the tie breaker these days since most ISP's offer a means to control their local preference by the use of advertised communities.

# 6.13 BGP AS-Path Attribute

Last Updated: Wednesday, 03-Apr-2013 12:51:59 MDT | By InetDaemon

AS-path attributes are used to provide route *metrics*. Along with the NLRI information are path attributes. Path Attributes allow BGP to make determinations of what is the best path. There are several categories that Path Attributes fall into:

# 6.14 AS-PATH ATTRIBUTE CATEGORIES

- Well-known, mandatory
- Well-known, discretionary
- Optional, Transitive
- Optional, Non-transitive

### 6.14.1 WELL KNOWN, MANDATORY

This attribute MUST appear in every UPDATE message. It must be supported by all BGP software implementations. If a well-known, mandatory attribute is missing from an UPDATE message, a NOTIFICATION message must be sent to the peer.

Examples:

- AS_path
- ORIGIN
- NEXT_HOP

### 6.14.2 WELL KNOWN, DISCRETIONARY

This attribute may or may not appear in an UPDATE message, but it MUST be supported by any BGP software implmentation.

Examples:

- LOCAL_PREF
- ATOMIC_AGGREGATE

### 6.14.3 OPTIONAL, TRANSITIVE

These attributes may or may not be supported in all BGP implementations. If it is sent in an UPDATE message, but not recognized by the receiver, it should be passed on to the next AS.

Examples:

- AGGREGATOR
- COMMUNITY

### 6.14.4 OPTIONAL, NON-TRANSITIVE

May or may not be supported, but if received, it is not required that the router pass it on. It may safely and quietly ignore the optional attribute.

Examples:

- MULTI_EXIT_DISC
- ORIGINATOR_ID
- Cluster List

# BGP Best Path Selection Algorithm

BGP selects a **single** best path to a destination, and inserts it in the IP routing table. IP datagrams are only forwarded based on routes in the IP table, NOT by the routes in the BGP table.

Since each distinct prefix is a unique destination, BGP will select and advertise only a single best path. To decide which is the 'best path', BGP uses an extensive tie-breaking algorithm. At each step, BGP seeks to break a tie between metrics. The selection process ends at the point where a tie between routes is broken by a better metric. The metrics used are as follows:

# BEST PATH ALGORITHM

1. The first path received is automatically the 'best path'. Any further paths received are compared to this path to determine if the new path is 'best'.
2. Is the route **VALID**? To be valid:
    - The route must be synchronized with the Interior Gateway Protocol (*unless synchronization is turned off*).
    - The route must appear in the IP routing table (see previous bullet point).
    - The NEXT_HOP must be *reachable*.
    - The AS_PATHs received from an external AS *must not contain the local AS*, or they will be discarded.
    - The local routing policy must permit the route. If the neighbor is filtering the route, they won't use it.
3. Highest **WEIGHT**
   Weight is a Cisco-proprietary setting and only exists on the router on which it is configured. It is otherwise useless throughout an AS.
4. Highest **LOCAL_PREF** (used within an AS)
5. Prefer **LOCAL**LY **ORIGIN**ATED route (originated from this router)
6. **Shortest AS-PATH**
7. **Lowest ORIGIN type:**
   **IGP** < **EGP** < **INCOMPLETE**
8. Lowest **MULTI-EXIT-DISCRIMINATOR (MED)**
9. **Prefer eBGP** route **over iBGP** route
10. Lowest IGP metric to **BGP NEXT_HOP**
11. Prefer **FIRST RECEIVED EXTERNAL ROUTE** (prefer the OLDEST external PATH)
12. Prefer **lowest router ID**
    The Cisco router ID is *IP address of the router*, which in turn is the *highest IP address* on the router, **or** its *Loopback* Interface if it has one.
13. Minimum **CLUSTER_ID** length
14. **Lowest neighbor address**

From this list, you can see it is *impossible* for two BGP routes to TIE each other and become equally preferred. As has been stated elsewhere in this site, BGP contains only a *single best path* to any given destination. BGP runs across the entire Internet, therefore it must manage to reduce the number of advertised routes in order to prevent the Internet from becoming flooded with route advertisement traffic. Thus, this algorithm is designed to eliminate all but 1 route to a destination.

# BGP Route Flap Dampening (RFC 2439)

While reading this tutorial, remember that BGP is a routing protocol. That means BGP learns and selects the best route to a given destination. Anything that makes a route less preferred causes the router to stop adding the route to the IP routing table and to stop advertising the route to neighbors.

# What is a Route Flap?

BGP peers exchange routes and send updates not faster than every 90 seconds by default. When a route is repeatedly advertised and withdrawn, it is considered to be 'flapping'. Flapping routes cause instability in the Internet routing table and Cisco routers running BGP contain an optional mechanism designed to dampen the destabilizing effect of flapping routes. When a Cisco router running BGP detects a flapping route it automatically *dampens* that route. 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 routes can remain in the routing table for *months.*

The term *route flap* is used when a previously advertised route is withdrawn and then readvertised. Cisco IOS later than 11.0 has a route dampening function built into it. If you are running 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.

The suppression half-life is 15 minutes. The maximum suppress limit is four times the half-life; thus, one hour is the default. The suppression penalty decays at half the half life (7.5 minutes). So:

1. First flap, penalty 1000 assigned, route placed in 'historical' category and becomes less preferred.
2. Second flap, route has met the suppression limit of 2000 (a Cisco default). The route is *dampened* and no longer advertised to neighbors or accepted from neighbors.
3. If route does not flap any further the penalty is decayed. The decay process begins 7.5 minutes after the route stabilized and decays exponentially every 5 seconds thereafter.
4. Once the suppression penalty decays below 750 (the default value for the reuse threshold), the route is removed from dampened state and reused. The router parses the historical routes list every 10 seconds for reusable routes.

# Checking for Dampened BGP Paths

You can check for dampened paths by issuing the following command at the command prompt:

router# **show ip bgp dampened-paths**

This works on Cisco, Juniper, Avici and HP routers.

# 6.15 Routing Information Base (RIB)

The Routing Information Base RIB is the location in which all IP Routing information is stored. The RIB is not specific to *any* routing protocol, rather, it is the repository where all the routing protocols place *all* of their routes. Routes are inserted into the RIB whenever a routing protocol running on the router learns a new route. When a destination becomes unreachable, the route is first marked unusable and later removed from the RIB as per the specifications of the routing protocol they were learned from. It is important to note that the RIB is NOT used for forwarding IP datagrams, nor is it advertised to the rest of the networks to which the router is attached.

A Cisco router's RIB will contain filtered routes; however, these will never make it to the "Forwarding Information Base", which contains yet a different set of routes.

# 6.16 Forwarding Information Base (FIB)

Last Updated: Wednesday, 03-Apr-2013 12:52:57 MDT | By InetDaemon

Cisco routers build a Forwarding Information Base (FIB) which contains all the routes that could potentially be advertised to all neighboring routers within the next set of announcements. This is also the same set of routes used to forward IP datagrams. On Cisco routers with a distributed forwarding architecture and which have Distributed Cisco Express Forwarding (DCEF) enabled, a copy of the FIB will be 'compiled' and downloaded to the applicable line-cards or modules. This offloads a large portion of the routing load from the main CPU and increases the overall traffic load that can be sustained by the router.

# 7 BGP Configuration

InetDaemon spent four and a half years setting up 10-15 BGP sessions per day. During that time, he also managed peering between the big carriers such as Sprint, WorldCom/UUNet, Qwest, Level3 and more than 40 other organizations including NASA and the public network side of the AT&T DISA connection.

Setting up BGP requires meeting a few requirements first, and InetDaemon can walk you through it. BGP is a dynamic protocol and can provide for automatic failover and load balancing in the right configuration.

**Basic BGP Configuration**

I. Requirements
II. Standard Configuration
III. Configuration Options
    A. Default Originate
    B. eBGP Multi-Hop
    C. Maximum-Prefix
    D. Auto-Summary

**Advanced BGP Configuration**

I. Routing Policies
    A. Distribute Lists
    B. Filter Lists
    C. Prefix Lists
    D. Route Maps
II. Redundancy and Fail-Over
III. Load Balancing
IV. Communities
V. Community String Controlled Local Preference
VI. Confederations
VII. Route Reflectors
VIII. Route Reflector Clusters

**Basic BGP Configuration**

# Requirements for Running BGP

To run BGP you are required to have the following:

1. An AS Number
2. Multi-homed to the Internet
3. BGP4 Capable Router
4. Sufficient Router Memory
5. Fully Functional IGP
6. Qualified Internet Engineer

**AS NUMBER**

Tutorial: AS Numbers

Originally, an AS number was only available through ARIN. An administrative agreement was worked out so that regional registrars can assign ranges of AS numbers. An AS number can ONLY be purchased from one of the Regional Internet Registries (RIR's). Only certain kinds of organizations qualify to obtain an ASN, such as goverments, global corporations, Internet service providers, telecommunications companies and so forth. An ASN (AS Number) can be requested from one of the registries by filling out an *ASN request form* (sometimes called the *ASN request template*) and submitting it to the Which registry you obtain your AS number from is based upon where in the world your network resides physically and will be connecting. This list has expanded to the following registrars:

- AMERICAS - American Registry for Internet Numbers (ARIN)
- AFRICA - African Network Information Center (AFRINIC) - ASN Request Form
- EUROPE - Reseaux IP Europeens (RIPE) - ASN Template
- LATIN AMERICA - Latin American Network Information Center (LACNIC) - ASN Template
- ASIA - Asian Pacific Network Information Center (APNIC)

**MULTI-HOMED**

To receive an AS number from ARIN, RIPE, or APNIC, you must be able to prove your network is connected to more than one Internet Provider running BGP by providing the contact phone number. This is called being 'multi-homed'.

**BGP4 CAPABLE ROUTER**

USE A BRAND OF ROUTER YOUR PROVIDER SUPPORTS. This reduces the chances of incompatibility issues and allows your provider to give you better support, as they will have experience with the equipment already.

Cisco routers must be running version 10.3 of the IOS or later to support BGP version 4.

**ROUTER MEMORY**

Your router will need sufficient memory to process the BGP routes your providers will be sending you. The table below gives a GENERAL outline of how much RAM will be required. Most providers support most of the following ranges of routes.

| RECEIVING | # | TOTAL RAM REQ'D |
|---|---|---|
| FULL ROUTES (entire Internet routing table) | ~ 135 K | 128 MB |
| PARTIAL ROUTES | 45 K+ | 64 MB - 128 MB |
| BACKBONE ONLY | 10 - 2K | 32 MB - 64 MB |
| NO ROUTES* | 0 or 1 | -- |

* If you are receiving NO ROUTES from your provider, you will either need a static default route, or ask your provider to send you the default route via BGP. If your ISP uses a Cisco router, your ISP can install the 'neighbor x.x.x.x default-originate' command in their neighbor statements for your BGP session.

**FULLY FUNCTIONAL IGP**

Your Interior Gateway Protocol (Static routes, RIP, OSPF, EIGRP) should be completely configured and functionong correctly. ALL internal networks should be completely installed, powered on, and routing correctly internally, as well as having the correct default routes pointing to your soon-to-be-installed Internet BGP4 gateway. Unless these are complete, BGP will NEVER advertise your route unless you take extreme measures, and even then your connectivity to the Internet will likely STILL not work. By default, BGP DOES NOT advertise networks it cannot reach. Thus, an interior IP address range must be fully synchronized with the IP route table before BGP will advertise it to the Internet. You can of course set a Cisco router to use the 'no-synchronization' command, but all that will happen is that traffic will be sent to your Internet router, but your traffic will die right there on the spot unless your Internet router is also the ONLY router on your network and it is directly connected to ALL your networks.

**QUALIFIED INTERNET ENGINEER**

If your company's engineer cannot answer the following questions, have someone who CAN answer these questions configure your BGP:

1. What is your AS number?
2. What is your router's PUBLIC and ROUTABLE IP address?
3. Who are your neighbors and what are their IP addresses?
4. What CIDR prefixes will you advertise?
5. Will you be aggregating?
6. What is your routing policy?
7. Will you be accepting full, partial or no routes from your provider?

# 7.1 Standard BGP Configuration

Last Updated: Wednesday, 03-Apr-2013 12:51:55 MDT | By InetDaemon

To configure BGP on a Cisco router, you must use some or all of the following commands:

```
router bgp [your AS]
 network x.x.x.x [ mask x.x.x.x ]
 neighbor n.n.n.n remote-as NNNN
 neighbor n.n.n.n version 4
 neighbor n.n.n.n distribute-list in|out
 neighbor n.n.n.n filter-list in|out
 neighbor n.n.n.n route-map NAME in|out
 neighbor n.n.n.n ebgp multi-hop <hop-count>
```

Let's break these commands down one by one.

**router bgp [your AS]**

This statement is REQUIRED. This command enables BGP on the router. Entering this command in global configuration mode places all configured neighbors in the "Active" state and changes the command prompt to read (config-router)#. If BGP is unable to initiate a TCP connection for a neighbor session you have configured, the state of that session will be 'Idle'.

**network x.x.x.x [ mask x.x.x.x ]**

This statement is REQUIRED. The network statement informs BGP what prefixes it is permitted to announce network X.X.X.X. The optional mask statement will cause aggregation of all CIDR blocks smaller than the network/mask combination into the larger supernet. All routes in the IP table that fall within this range will be advertised as originating from that router.

**neighbor n.n.n.n remote-as NNNN**

This statement is REQUIRED to set up a session to a BGP speaking neighbor. This is the first statement in any block of neighbor statements. It iinforms BGP which IP address to peer with, and whether the session will be an iBGP or eBGP session based on the AS number. Neighbors with the same AS number will establish an **i**BGP session. Neighbors with different AS numbers will establish an **e**BGP session. The IP address in each of the neighbor statements serves to inform the router which statements apply to a specific neighbor.

**neighbor n.n.n.n version 4**

This statement is not required, but there are issues with BGP3. Older routers might default to BGP version 3 or may not support BGP version 4 at all. Using this statement guarantees that your router will only establish the session if the neighbor can 'speak' BGP version 4. This statement forces BGP to run as version 4 or not at all. Version 4 supports CIDR and route dampening.

**neighbor n.n.n.n distribute-list xxx in|out**
**neighbor n.n.n.n filter-list xxx in|out**
**neighbor n.n.n.n prefix-list xxx in|out**

If you are not an ISP, these are REQUIRED or your private network will become a transit network and traffic will start flowing between your ISP's *through* your network. Both prefix-lists and distribute-lists filter routes by ranges of IP addresses. Prefix-lists are simply another way to write distribute-lists. Filter-lists filter by AS-path. Therefore, you cannot use them together on the same neighbor session to filter routes in the same direction. You *can* combine a filter-list with either in the same direction. You may use any two of these you wish, so long as it is in opposite directions."

**neighbor n.n.n.n route-map NAME in|out**

A Route Map is a more effective means of implementing a route policy, and allows the administrator to implement not only a more complex filter, but to adjust routing metrics using the MATCH and SET statements.

**neighbor n.n.n.n ebgp multi-hop <hop-count>**

This command is optional and not often used but it can allow two BGP speaking routers not directly connected to establish a BGP session. As it's name implies, eBGP multi-hop cannot be used for iBGP sessions.

Satellite Internet customers who have a simplex connection will use another connection to provide the physical layer return path for BGP. eBGP MULTI-HOP provides the means to allow such network environments to function using BGP, albeit with rediculously high hop counts (20+).

# BGP Configuration Options

There are a number of options on a Cisco router for configuring BGP.

- Default Originate
- eBGP Multi-Hop
- Maximum-Prefix
- Auto-summary

### DEFAULT ORIGINATE

You can allow a BGP-speaking router to originate a default route. Typically, this is used by Internet Service Providers wishing to provide their customers with a default route into their network. This option is most frequently used in BGP sessions where the receiver has almost no CPU capacity and/or RAM, and therefore cannot receive a full BGP table.

In this case, an ISP will configure the default-originate option on their side of the BGP session as shown here:

```
neighbor x.x.x.x remote-as <as-number>
neighbor x.x.x.x default-originate
neighbor x.x.x.x distribute-list CUSTOMER in
```

### EBGP MULTI-HOP

Multi-hop is used to allow two routers that do not share a direct physical connection to establish a BGP peering sesson. The command must appear in the configuration of BOTH SIDES of the BGP session. The Cisco command to enable multi-hop is:

```
neighbor x.x.x.x remote-as <as-number>
neighbor x.x.x.x ebgp-multihop <hop count>
```

### MAXIMUM PREFIX

Sometimes the administrator of a remote AS will incorrectly configure their BGP session and will begin leaking routes into your network. Since a peer is normally connected to the Internet, this can be a VERY large number of routes. That large a change in the routing table can frequently overwhelm a router that is running on a thin margin of RAM or CPU load. To prevent this occurance, you can add a 'safety fuse' to the BGP session using the 'maximum-prefix' command.

```
neighbor x.x.x.x remote-as <as-number>
neighbor x.x.x.x maximum-prefix <threshold>
```

Typically, you will wish to set the maximum prefix threshhold approximately 20% over the usual number of prefixes received. To see the current number of prefixes received, run the 'show ip bgp sum' command.

### AUTO SUMMARY

BGP normally sumarrizes route announcements along classful boundaries. Using the NO AUTO-SUMMARY command turns this off. This command is not part of a neighbor session, but rather is turned off at the router configuration level (the prompt looks like this: router-name(config-router)# ). Once this is configured, this command will appear at the end of all the neighbor configurations.

```
router-name(config)# router bgp <as-number>
router-name(config-router)# no auto-summary
...
router-name# show run
...
router bgp <as-number>
 network x.x.x.x
 network y.y.y.y
 neighbor n.n.n.n remote-as <as-number>
 neighbor n.n.n.n version 4
 neighbor n.n.n.n no auto-summary
```

**Advanced BGP Configuration**

# Creating a BGP Routing Policy

## WHAT IS A ROUTING POLICY?

What is a routing policy? A routing policy allows an administrative authority to control the routing behavior of an autonomous system. A routing policy can be applied to all external and internal routes, and can control the propagation of routes, vastly improving the stability of the network. A well-planned routing policy will help the administrator control network behavior automatically, reducing the administrative overhead associated with maintaining the network.

Routing policies are created by applying a series of route filters and route maps. Route filters control which routes are advertised and received. Route maps control the metrics on those routes to determine which routes should be received and used normally, and which routes should be administratively adjusted to change traffic flow and routing behavior.

Also, your route policy can be communicated to the Internet by use of the Route Arbiter Data Base (RADB), allowing others to adjust their routing policies to work in harmony with yours. However, keep in mind that the top five ISP's (UUNet/WorldCom, Sprint, Teleglobe, Cable & Wireless and AT&T) DO NOT PARTICIPATE in the RADB, which renders it rather meaningless since those providers supply access to every Internet destination worldwide.

## THE ISP'S INBOUND ROUTE POLICY

Most service providers will implement inbound route filters of SOME type on their session with you. Keep this in mind when announcing new blocks. Never assume that if your provider is smart enough to keep track of what IP addresses they have assigned you, it means they have permitted it through their filters already. Most ISP's simply aren't that organized. ALWAYS request that they open the filters after you receive a new IP address block you intend to announce. Also, make sure you identify the IP address with which you peer and your AS number whenever communicating with your ISP regarding ANY BGP problem or issue.

## BACKBONE PROVIDERS PEERING FILTERS

At one time, there was a call to the Internet by ARIN requesting that ISP's begin aggregating all routes into supernets. The explosion in the number of routes was threatening to overpower the hardware on the routing equipment and bog down the Internet. ARIN requested that backbone Internet providers restrict which routes were advertised out of their networks into other Backbone provider's networks to reduce the size of the Internet's backbone route table.

Such route policies looked something like this:

| CLASS | IP Range | | FILTER BLOCKS SMALLER THAN |
|---|---|---|---|
| | START | END | |
| A | 0.0.0.0 | 127.255.255.255 | /16 |
| B | 128.0.0.0 | 191.255.255.255 | /18 |
| C | 192.0.0.0 | 223.255.255.255 | /19 |

Some major ISP's still implement such a routing policy but with customers screaming about their need for redundancy and load balancing; with ARIN assigning ever smaller subnets, this policy is vanishing. The routers in use today are now fast enough and the cost of memory is cheap enough that many more routes can be handled today than in the past.The original hardware issues are no longer such a primary concern.

If you plan to run BGP, you need to find out if your provider still implements such a routing policy and what they will accept. Their policy will determine where your traffic will tend to flow.

If your provider still implements such a policy a variety of odd routing problems will occur depending upon what happens to the routes you announce. If you have your own /24 from ARIN, your provider may not announce it, thus leaving you dead in the water. route will not propagate as a /24, and will be preferred on the provider network that advertises it as /24 instead of aggregating it. The end result is often that routes for blocks belonging to provider A, prefer provider B's network to reach you. This is because provider B cannot aggregate your IP block that was assigned to you out of provider A's address space. The prefix announced out of provider B (/24) is more specific, and therefore preferred than provider A's announcment (/16 for example).

**IMPLEMENTING YOUR OWN ROUTE POLICIES**

There are four Cisco supported methods for configuring a routing policy:

1. Distribute lists
2. Prefix-lists
3. AS-path filter lists
4. Route Maps (combines both Prefix and AS-path filters)

# 7.2 BGP Distribute Lists

A distribute list filters routes based on the IP of the destination and are therefore more effective than filter lists because they focus on the prefixes, instead of the AS-paths.

To set up a distribute list, you must create an access list. The access-list must permit all blocks you wish to allow. Cisco routers use an inverse or 'wildcard' mask for access lists.

For example if the IP address range you own runs from 204.134.12.0 through 204.124.24.255 then you would use the following access list (the list number is arbitrary and used to group the statements together).

```
access-list 21 permit 204.134.12.0 0.0.3.255
access-list 21 permit 204.134.16.0 0.0.7.255
access-list 21 permit 204.134.24.0 0.0.1.255
```

Next, you must apply this distribute list to the correct BGP neighbor session as an inbound our outbound list. Outbound distribute lists filter your announcements to your peer. Inbound announcements filter what routes you will accept from your peer.

```
router bgp <your AS>
 ...
 neighbor x.x.x.x remote-as NN
 neighbor x.x.x.x version 4
 neighbor x.x.x.x distribute-list 21 out
```

**OUTBOUND DISTRIBUTE LIST**

An outbound distribute list assures that you do not announce routes heard from one of your peers to another peer. An outbound list restricts your announcements to only those routes you own and can reach.

**INBOUND DISTRIBUTE LIST**

If you are an Internet Service Provider, you will also need to restrict the routes your downstream customers and peers announce to you. You will need a complete list of routes from that customer to apply to the inbound routes announced by your customer. This is the most common reason for an inbound distribute list, but you should always apply one for 'sanity checking' to block private and non-routable IP addresses (such as 192.168.0.0 or 127.0.0.1).

# 7.3 BGP Prefix-Lists

Prefix lists are more sophisticated forms that Cisco provides for filtering BGP route advertisements. They filter on IP address just as distribute-lists do, however they are easier to read, and require fewer commands to configure. The other advantage to a distribute list is that it is easeir to add, remove and organize the statements in the manner you chose.

For example:

```
prefix-list xx seq 10 permit 204.134.12.0/22
prefix-list xx seq 20 permit 204.134.16.0/21
prefix-list xx seq 30 permit 204.134.24.0/24
```

While this configuration requires the same number of statements as the distribute list example, you have the option of adding **ge**, or **le** to make statements more flexible as to how you will permit blocks in that range.

For example:

```
prefix-list xx seq 10 permit 63.1.0.0/16 ge 18
```

The statement above allows any route announcement in the range of 63.1.0.0 - 63.1.255.255 but that announcement must have a length greater than 18 bits in the mask. This permits you to allow announcements in the range, but not an announcement equalling the entire range (/16), or even announcements of half the range (/17). Only announcments with a length "greater than or equal to" /18 will be permitted.

If this is the power of **ge**, what could you do with **le**?

# 7.4 BGP AS-Path Filter Lists

A filter list is a form of route policy that restricts the routes that will be advertised or accepted based on the AS-Path of the route. To configure a filter list, you must first create an AS-path access list based on the known paths you wish to permit.

```
as-path access-list xx permit 701
as-path access-list xx permit 701 6461
as-path access-list xx permit 701 6461 3
```

The list above will permit the following AS-paths:

701
701 6461
701 6461 3

To appy this list to a BGP session, use the following command:

**neighbor <IP address> filter-list xx in|out**

The list can be applied either to the route received (inbound) or the routes advertised (outbound). Now let us suppose that to adjust the routing, an administrator at MIT used AS-path-prepending to make routes to one provider more preferred over another. This new prepended AS-path would look like this:

701 6461 3 3

This path would never be permitted through the AS-path filter because AS 3 appears twice. Worse, suppose that after the filter was changed to match this, the administrator at MIT decided to go back to a standard announcement, or decided to prepend twice. This would mean a headache for the person maintaining the filter and delay needed changes.

To make the list more flexible, Cisco has enabled the use of *regular expressions* in an as-path filter list. The same list above could be rewritten to permit prepends from all of the providers in the AS path, and even shorten the list:

**as-path access-list xx permit ^(_701)+(_6461)*(_3)$**

The filter list above whould permit the following AS-paths:

701
701 701
701 6461
701 3
701 6461 3
701 6461 6461 3 3 3

Clearly this second list is shorter, and much more flexible. The characters that are used above are as follows:

| Char. | Meaning |
|---|---|
| ^ | Beginning of character string |
| _ | Any whitespace |
| ( ) | Brackets are used to group items together |
| NNN | The numbers represent the number patterns of the AS numbers. |
| * | Zero or more of the previous object |
| + | One or more of the previous object |

The list above will permit the following AS-paths:

701
701 6461
701 6461 3

To appy this list to a BGP session, use the following command:

**neighbor <IP address> filter-list xx in|out**

The list can be applied either to the route received (inbound) or the routes advertised (outbound). Now let us suppose that to adjust the routing, an administrator at MIT used as-path-prepending to make routes to one provider more preferred over another. This new prepended AS path would look like this:

701 6461 3 3

This path would never be permitted throught the AS-path filter because AS 3 appears twice. Worse, suppose that after the filter was changed to match this, the administrator at MIT decided to go back to a standard announcement, or decided to prepend twice. This would mean a headache for the person maintaining the filter and delay needed changes.

To make the list more flexible, Cisco has enabled the use of regular expressions in an as-path filter list. The same list above could be rewritten to permit prepends from all of the providers in the AS path, and even shorten the list:

**as-path access-list xx permit ^(_701)+(_6461)*(_3)$**

The filter list above whould permit the following AS-paths:

701
701 701

701 6461
701 3
701 6461 3
701 6461 6461 3 3 3

Clearly this second list is shorter, and much more flexible. The characters that are used above are as follows:

| Char. | Meaning |
|---|---|
| ^ | Beginning of character string |
| _ | Any whitespace |
| ( ) | Brackets are used to group items together |
| NNN | The numbers represent the number patterns of the AS numbers. |
| * | Zero or more of the previous object |
| + | One or more of the previous object |

# BGP Route Maps

Cisco routers implement a route policy using Route Maps. A route map can utilize access-lists, prefix-lists, as-path access lists, and community lists to create an effective route policy.

A route map consists of a series of statements that check to see if a route matches the policy, to permit or deny the route, and then possibly an additional series of commands to adjust the atrributes or metrics of those routes.

AS-path prepending is an example of one such use of route maps, as is the implementation of community string controlled local preference. Using a route map, you can lable routes you receive with special community strings so that you can modify the metrics, or filter the routes before announcing them.

A route map consists of the route map statement permitting or denying all routes matching the list it calls. Each route map statement contains a number. These numbers are used to place the steps of the route map in order.

For example:

```
route-map NAME permit 10

 match access-list 22
 set community 701:666

!
route-map NAME permit 20
 match prefix-list NO-GO
 set metric 20000000

!
route-map NAME deny 30
 match community 41
```

**WHAT YOU CAN MATCH**

You can match on any of the following list types:

| LIST TYPE | COMMAND | MATCHES BY |
|---|---|---|
| access-list | match ip address | IP address |
| prefix-list | match prefix | IP address |
| as-path-access-list | match as-path | AS-path |
| community-list | match community | Community String |

You can also match on the following:

- Interface
- NEXT_HOP
- route-source
- metric

- route-type
- tag

**WHAT YOU CAN SET**

You can set any of the following metrics and attributes:

- LOCAL_PREF (affects routes within the AS)
- NEXT_HOP
- AS_PATH (prepend routes or modify the path)
- Multi Exit Discriminator (MED) (Affect the route an external AS uses to enter your network)
- Community strings (lable a route)

**HOW A ROUTE MAP WORKS**

ROUTE MAP <NAME> **PERMIT** nn

A route that meets the route map's MATCH criterion will have all SET commands applied to the route's metrics or attributes. The lists called by the MATCH statements can have PERMIT or DENY commands. Items matching the PERMIT statement will be SET, items matching a DENY will not be SET.

ROUTE MAP <NAME> **DENY** nn

A route not matching some line in the lists this route map's MATCH statements call will be permitted. The route map will exit to begin again and evaluate the next route.

Redundancy and Fail-Over

# 7.5 BGP Load Balancing

BGP does not perform load balancing. BGP is specifically designed to select the single best route to a destination, therefore it is not able to perform load balancing itself. However, if you can configure BGP over a virtual connection, you can get a Cisco router to balance traffic using up to six static routes.

### USING EQUAL COST PATHS

To do so, you set up more than one equal cost path. Using static routes, you can easilly accomplish this. Each static route is configured to reach a matching prefix, and because they are static routes, have equal administrative distance. Most often, static routes are used between two points. BGP is then configured to run over that set of connections using the 'ebgp-multihop' command.

### USING LOOPBACK INTERFACES

By setting up connections between two loopback interfaces, the routers can take advantage of the fact that the loopback interface is never down, and will be 'up' so long as it is reachable via the configured static routes.

### CONFIGURING BGP 'LOAD BALANCING'

**ROUTER A**

```
interface serial 0
 ip address 1.1.2.1 255.255.255.252
interface serial 1
 ip address 1.1.3.1 255.255.255.252
interface loopback 0
 ip address 1.1.1.1 255.255.255.255

router bgp 1111
 network 1.2.3.0
 neighbor 2.2.2.2 version 4
 neighbor 2.2.2.2 ebgp-multihop 2
 neighbor 2.2.2.2 update-source loopback 0

ip route 1.1.1.1 255.255.255.255 1.1.2.2
ip route 1.1.1.1 255.255.255.255 1.1.3.2
```

**ROUTER B**

```
interface serial 0
 ip address 1.1.2.2 255.255.255.252
interface serial 1
 ip address 1.1.3.2 255.255.255.252
interface loopback 0
 ip address 2.2.2.2 255.255.255.255

router bgp 2222
 network 5.4.3.0
 neighbor 1.1.1.1 version 4
 neighbor 1.1.1.1 ebgp-multihop 2
 neighbor 1.1.1.1 update-source loopback 0
```

```
ip route 1.1.1.1 255.255.255.255 1.1.2.1
ip route 1.1.1.1 255.255.255.255 1.1.3.1
```

- Communities
- Community String Controlled Local Preference
- Confederations
- Route Reflectors
- Route Reflector Clusters

## 8 Security and BGP

Why would you need security in BGP?

To protect yourself against man-in-the-middle attacks and prevent illicit routes from being artificially inserted into your tables for nefarious purposes. By using several security features added to BGP, you can greatly reduce the vulnerability of BGP to attack by malicious attackers.

- **Challenge-Handshake Authentication Protocol**
  This is your last alternative. Use Message Digest 5 if you canand combine that with effective route policies. This uses a pre-arranged value on both routers to authenticate the sender and receiver when establishing a connection. However, CHAP is sent in cleartext. It won't stop a man-in-the-middle attack because they can see the CHAP exchange.
- **MESSAGE DIGEST v5**
  Configuring a Message Digest key on neighboring routers enables you to authenticate the BGP messages as coming from a valid source. An MD5 hash key is difficult to break and nearly impossible to forge in real time with current technology. Enable MD5 on your BGP route announcements. Routers must be pre-configured with the correct hash and key values, however this makes a network far more safe against malicious route insertion as the keys are never passed over the network.
- **Establish an Effective Route Policy**
  Establishing a route policy goes a long way towards defeating external threats in addition to preventing congestion, routing loops and other network anomolies.
    - Build good ingress and egress filters (especially if you are an ISP)
    - Block non-routable addresses (Loopback and private addresses)
    - Block the 'special/experimental use' IP addresses
    - Blockcertain known troublemakers
        - Block any packets sourcing your IP addresses from outside your network
        - Residential dial-up, cable and DSL modem addresses
        - Some foreign addresses in certain countries
    - Configure null routes within your network that dump traffic bound for unreachable destinations into the bit bucket.

# Best Option

Your best option to protect your BGP session is to:

1. Configure an effective Route Policy (e.g.: Cisco route maps)
    1. Sane inbound and outbound prefix lists and route filters
    2. Set a limit on the maximum number of BGP prefixes accepted from each peer
2. Use MD5 keys
3. Enable route dampening properly
4. Set a limit on the maximum number of prefixes accepted from each peer.
5. Place null routes in strategically distributed locations so that bad traffic is dumped as soon as it enters the network.

# 9 Troubleshooting BGP

Last Updated: Wednesday, 03-Apr-2013 12:52:25 MDT | By InetDaemon

Troubleshooting BGP is not an art, nor a science, it's straightforward methodical verification of each layer of functionality. Skip something and you'll be sitting on the phone for hours with tech support.

I. Learn TCP/IP and know it COLD
II. Learn BGP and know it COLD
III. MEMORIZE the BGP Best Path Selection Algorithm
IV. **Know the common causes of Problems**
    - Impatient BGP administrator
    - No TCP/IP connection
    - Bad Access Lists
    - Bad Route Filters
    - Advertised Network Unreachable
    - Synchronization
V. **Troubleshooting Techniques**
    - Use the appropriate BGP Commands for troubleshooting
        - show ip route
        - show ip bgp (regex)
        - show ip bgp longer-prefixes
        - show ip bgp neighbor
    - Route Servers
        - Memorize a few addreses
        - Use them to check your routes as heard outside your ISP's network.

- Learn TCP/IP and know it COLD
- Learn BGP and know it COLD

# BGP Best Path Algorithm

Last Updated: Wednesday, 03-Apr-2013 12:52:23 MDT | By InetDaemon



# Introduction

The 'best path algorithm' is used to narrow a list of routes down to 1 and only 1 *best path.* The list is composed of a set of criterion that are used for breaking ties between routes with equal costs.

# The Algorithm (simplified Cisco)

I. Largest Weight
    A. Set by the administrator
    B. Proprietary to Cisco
    C. Not passed to other routers (local metric)
II. Highest LOCAL_PREF (defaults to 100)
III. Prefer routes aggregated with *network* or *aggregate-address* commands. Prefer *network* over *aggregate-address.*
IV. Shortest AS_PATH.
V. Lowest ORIGIN type: IGP > EGP > INCOMPLETE

## 10 Know the common causes of Problems

## 11 The Impatient BGP Administrator

**THE** BIGGEST AND **MOST COMMON** REASON FOR PROBLEMS WHEN TROUBLESHOOTING OR CONFIGURING BGP? **IMPATIENCE!**

Due to his impatience the local administrator begins changing his BGP configuration as fast as he can type which is about every thirty seconds. The changes to the BGP session cause it to repeatedly reset, making the BGP entries in the routing tables across the Internet flap (but not if the ISP has properly tuned their own BGP). These route flaps trigger the hold down timer, cause route dampening and ultimately cause his traffic to drop off to ZERO even though the local BGP session is up and exchanging routes.

*Don't fiddle around with the BGP session unless you have no other choice.*
*Figure out why it broke before meddling with things you don't understand.*
*Get second or third level support personnel at your ISP on the horn before doing any serious BGP configuration.*
***--InetD***

By default, BGP sends messages NOT FASTER than 90 seconds apart to prevent the network from being flooded with updates. **It takes AT LEAST TWO UPDATES or THREE MINUTES to get the BGP state to stabilize**. The changes need to propagate through the Internet and it takes three minutes or longer per router to clear the routing issues. Every time your route flaps, the hold down timer DOUBLES. If your routes are dampened, your BGP session could theoretically be down for DAYS depending upon how your ISP configured their timer settings!

# 11.1 No TCP/IP Connection

## 12 No TCP/ IP Connection

**The number one *TECHNICAL* cause of BGP problems is a failed TCP session.**

**** TROUBLESHOOT THE NETWORK CONNECTIONS FIRST ****

**You CANNOT troubleshoot BGP if you don't have a stable physical link and TCP cannot establish a connection.**

Why is all this in big bold letters?

1. You must have a green light on your CSU/DSU (or DSL, FRAD or ISDN modem etc.). The green light only means the physical layer and data link layer are working.
2. IP is a Network Layer protocol. IP can't run if the physical link isn't working and PING will fail if IP is not properly configured.
    1. You need an IP address
    2. You need to make sure the mask is the same on both sides of the circuit

    3. You need to be sure that your router knows how to reach the IP address at the far end of the circuit.
3. TCP is a transport layer protocol and runs over IP. If you can't establish a TCP connection you can't communicate TCP transported data.
4. BGP is effectively an Application layer protocol and runs on TCP over IP. No TCP, IP or physical/data link layer connectivity equals no BGP communication.
5. ICMP PING *does not* need BGP to ping your ISP's router on the other side of an IP connection. If you cannot PING, you have bigger problems than BGP and need to fix the link itself.

# 12.1 BGP Blocked by Access Lists

The SECOND most common cause is an ACCESS-LIST that does not explicitly permit TCP port 179 (BGP). Access lists on Cisco routers are built to perform what is called *implicit deny*. That means that unless you have an explicit *permit* statement for the protocol or you have a global 'permit ip any any' , BGP gets blocked by default. What makes this tricky is that when you remove the ACL, BGP seems to be working. When you add the ACL, it STILL seems to be working. BGP has a *timeout*. When you apply the access list, BGP must first time out the connection before it will show the session as down, so for the first three minutes after you add the access list, it appears to still be up, but you will see no routes.

# 12.2 BGP Routes Filtered

If the route filters aren't implemented correctly, the BGP session won't function properly. The session might remain up, but it will not be able to exchange routes or critical routes will not get advertised. If you haven't done so already, look into the route filters section of this tutorial.

If you suspect that your route filters are blocking your neighbor's BGP announcements, use the following commands to add the *soft-reconfiguration* command to the BGP neighbor session in order to see what routes they are truely sending.

```
conf t
router bgp <as number>
neighbor x.x.x.x soft-reconfiguration inbound
```

The above commands allow you to see what announcements are actually coming in without removing your route filter and getting flooded with route announcements. To see ALL the routes the neighbor is announcing to you, type the following command:

```
show ip bgp neighbor x.x.x.x received-routes
```

If you don't see the routes after setting up soft-reconfig and checking for received routes, THEY AREN'T BEING ANNOUNCED! The administrator of the neighboring router has some work to do.

Keep in mind that altering a BGP session configuration on a Cisco router with IOS prior to 12.0 will cause the BGP session to reset and cause the connection to drop. Purposely resetting the session does not count as a route flap in the local session between you and your ISP. However, your route may flap on the Internet while work on the BGP session is going on.

*Practical experience has shown that routes tend to propagate to the major route servers after the session has been up and stable for about 15-20 minutes. --InetD*

## 12.3 Network Not Connected

The fifth most common cause of a BGP failure is that the network it is supposed to be advertising is DOWN (often not even connected). This is fairly uncommon because the network that BGP is usually supposed to advertise is the LAN that the Administrator actually manages, which is rarely down. Still, administrators usually refuse to connect the LAN to the router before trying to set up BGP and don't have enough expertise to know how to work around this.

BGP is doing what it was *designed* to do and is not advertising a route for a network it cannot reach. Customers used to make this mistake *all the time*. They wanted to advertise their routes, but were too afraid to connect the Internet router to the local network, thus, the Internet router thought the local network was down and wouldn't advertise the route for the LAN.

## 12.4 Synchronization

Cisco turns on route synchronization by default. This is to assure that the networks BGP is advertising are in fact reachable. However, if the administrator on the receiving end of the Internet connection is using static routes on his LAN, it may be advisable to turn off synchronization at the Internet router, especially if their network is unstable and tends to crash or fail with any significant frequency. By turning off synchronization and redistributing static routes into BGP on the customer side of the Internet connection, the ISP sees the customer as "always up". ISP's actually use this mechanism after aggregating their routes to make sure major peering connections stay active.

# 13 Cisco 'show ip route' Command

The *show ip route* command shows you which BGP routes have made it to the IP routing table. BGP is NOT used to make forwarding decisions on data, it is used to learn the single best path to a destination network. To do this, BGP listens for routes and advertises the best routes it knows. BGP also places the best route it knows for each destination (prefix) into the IP routing table. A Cisco router uses the Best Path Selection Algorithm to select the single best route to a destination and then inserts it into the IP routing table.

To check a router's IP routing table for a given destination (such as Yahoo), simply use the *show ip route* command:

```
route-server>show ip route 216.109.127.30
   Routing entry for 216.109.112.0/20, supernet
 Known via "bgp 65000", distance 20, metric 0
 Tag 7018, type external
 Last update from 12.123.9.241 1d17h ago
 Routing Descriptor Blocks:
```

```
* 12.123.9.241, from 12.123.9.241, 1d17h ago
Route metric is 0, traffic share count is 1
AS Hops 5
Route tag 7018
```

From the above output, the route to Yahoo (216.109.127.30 in this case) is known via an external BGP peer. As this is the ONLY route to the destination, this is the *preferred route.*

## 14 Addendum

# 14.1 The Route Arbiter Project (and the RADB)

The Route Arbiter project was a research grant awarded by the National Science Foundation to the Merit Network and the University of Southern California's Information Sciences Institute in July of 1994. The goal of this project was to "support leading-edge routing, tool development and research for the U.S. Internet". Work on this project was completed in 1998 to help address routing issues occurring at Internet Exchanges. The result of this research was a commercial product called the Route-Arbiter Database (RADB). The RADB contains information about the routing policies of those who have registered for the service and provided the information. The fee for this privledge is $250.

At that time, it was not uncommon for a destination to become unreachable because of the routing policies of a given provider. The point of the RADB was that many ISP's filter the routes they receive from other ISP's and thus 'delete' many routes from the Internet. This causes Internet users be unable to reach certain websites. The RADB became a "routing registry" where the routing policy of a registered participant in the RADB could be placed on file for others to use. IN THEORY this would make troubleshooting Internet routing problems easier by being able to see the route filtering policies of all the ISP's along the path to a location that is unable to reach your website. IN PRACTICE, none of the top backbone ISP's participated in this project, rendering the tool almost useless.

There are a number of software tools developed to allow registered participants and Internet users to query the RADB; however, these tools are no longer supported.

# 14.2 Books on BGP

Architectures (Bassam Halabi)

This is perhaps the definitive work on the subject of BGP and the ideal book if you're just starting out. Nevermind it is over 10 years old, it's STILL the definitive work on the core BGP functionality.

The book contains a clear explanation of how BGP functions, shows exact Cisco IOS configuration examples and guides the user through the complex nature of routing announcements, aggregation, controlling announcements and filtering received routes. The author even makes recommendations for special routing environments used in the backbone of very large networks such as confederations.

Ideal reference for Network Architects, Network Operations personnel and anyone who deals with BGP on a regular basis. An excellent primer for those wishing to learn BGP. It really is *that* good.

*I keep this book in ready reach whenever I am working with BGP. --InetDaemon*

Routing in the Internet

Christian Huitema is an author of a number of Request For Comments documents (RFC's) and is working for INRIA on high speed routing in networks above 1Gbps. This work is more broad than the Halabi book but approaches it from a slightly different perspective.