**Border Gateway Protocol** (**BGP**) is a standardized [exterior gateway protocol](https://en.wikipedia.org/wiki/Exterior_gateway_protocol) designed to exchange [routing](https://en.wikipedia.org/wiki/Routing) and reachability information among [autonomous systems](https://en.wikipedia.org/wiki/Autonomous_system_(Internet)) (AS) on the [Internet](https://en.wikipedia.org/wiki/Internet).[[1]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-orbit-1) BGP is classified as a [path-vector routing protocol](https://en.wikipedia.org/wiki/Path-vector_routing_protocol),[[2]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-2) and it makes [routing](https://en.wikipedia.org/wiki/Routing) decisions based on paths, network policies, or rule-sets configured by a [network administrator](https://en.wikipedia.org/wiki/Network_administrator).

BGP may be used for routing within an autonomous system. In this application it is referred to as **Interior Border Gateway Protocol**, **Internal BGP**, or **iBGP**. In contrast, the Internet application of the protocol may be referred to as **Exterior Border Gateway Protocol**, **External BGP**, or **eBGP**.

**History**

The Border Gateway Protocol has been in use on the Internet since 1994.[[3]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-3) The current version of BGP is version 4 (BGP4), which was published as [RFC 4271](https://tools.ietf.org/html/rfc4271) in 2006,[[4]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-4) after progressing through 20 drafts from documents based on [RFC 1771](https://tools.ietf.org/html/rfc1771) version 4. [RFC 4271](https://tools.ietf.org/html/rfc4271) corrected errors, clarified ambiguities and updated the specification with common industry practices. The major enhancement was the support for [Classless Inter-Domain Routing](https://en.wikipedia.org/wiki/Classless_Inter-Domain_Routing) (CIDR) and use of [route aggregation](https://en.wikipedia.org/wiki/Route_aggregation) to decrease the size of [routing tables](https://en.wikipedia.org/wiki/Routing_table).

[IPv6](https://en.wikipedia.org/wiki/IPv6) BGP was first defined in [RFC 1883](https://tools.ietf.org/html/rfc1883) in 1995, and it was improved to [RFC 2283](https://tools.ietf.org/html/rfc2283) in 1998. The new RFC allows BGP4 to carry a wide range of [IPv4](https://en.wikipedia.org/wiki/IPv4) and [IPv6](https://en.wikipedia.org/wiki/IPv6) "address families". It is also called the Multiprotocol Extensions which is Multiprotocol BGP (MP-BGP).

**Operation**

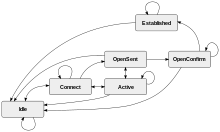
Ideally, even if impractical, all possible links would connect to one another, like spokes in a bicycle tire. BGP, "the most scalable of all routing protocols,"[[5]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-5) provides a workaround and ways to measure how effectively it is configured.

BGP neighbors, called peers, are established by manual configuration among [routers](https://en.wikipedia.org/wiki/Router_(computing)) to create a [TCP](https://en.wikipedia.org/wiki/Transmission_Control_Protocol) session on [port](https://en.wikipedia.org/wiki/Port_(computer_networking)) 179. A BGP speaker sends 19-byte keep-alive messages every 60 seconds[[6]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-6) to maintain the connection.[[7]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-7) Among routing protocols, BGP is unique in using TCP as its transport protocol.

When BGP runs between two peers in the same [autonomous system](https://en.wikipedia.org/wiki/Autonomous_system_(Internet)) (AS), it is referred to as *Internal BGP* (*i-BGP* or *Interior Border Gateway Protocol*). When it runs between different autonomous systems, it is called *External BGP* (*eBGP* or *Exterior Border Gateway Protocol*). Routers on the boundary of one AS exchanging information with another AS are called *border* or *edge routers* or simply *eBGP peers* and are typically connected directly, while *i-BGP peers* can be interconnected through other intermediate routers. Other deployment [topologies](https://en.wikipedia.org/wiki/Network_topology) are also possible, such as running eBGP peering inside a [VPN](https://en.wikipedia.org/wiki/VPN) tunnel, allowing two remote sites to exchange routing information in a secure and isolated manner. The main difference between iBGP and eBGP peering is in the way routes that were received from one peer are propagated to other peers. For instance, new routes learned from an eBGP peer are typically redistributed to all iBGP peers as well as all other eBGP peers (if *transit* mode is enabled on the router). However, if new routes are learned on an iBGP peering, then they are re-advertised only to all eBGP peers. These route-propagation rules effectively require that all iBGP peers inside an AS are interconnected in a full mesh.

How routes are propagated can be controlled in detail via the *route-maps* mechanism. This mechanism consists of a set of rules. Each rule describes, for routes matching some given criteria, what action should be taken. The action could be to drop the route, or it could be to modify some attributes of the route before inserting it in the [routing table](https://en.wikipedia.org/wiki/Routing_table).

**Extensions negotiation**

[](https://en.wikipedia.org/wiki/File:BGP_FSM.svg)

BGP state machine

During the peering handshake, when OPEN messages are exchanged, BGP speakers can negotiate[[8]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-8) optional capabilities of the session, including [multiprotocol extensions](https://en.wikipedia.org/wiki/Multiprotocol_BGP) and various recovery modes. If the multiprotocol extensions to BGP[[9]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-9) are negotiated at the time of creation, the BGP speaker can prefix the Network Layer Reachability Information (NLRI) it advertises with an address family prefix. These families include the IPv4 (default), IPv6, IPv4/IPv6 Virtual Private Networks and multicast BGP. Increasingly, BGP is used as a generalized signaling protocol to carry information about routes that may not be part of the global Internet, such as VPNs.[[10]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-10)

In order to make decisions in its operations with peers, a BGP peer uses a simple [finite state machine](https://en.wikipedia.org/wiki/Finite_state_machine) (FSM) that consists of six states: Idle; Connect; Active; OpenSent; OpenConfirm; and Established. For each peer-to-peer session, a BGP implementation maintains a state variable that tracks which of these six states the session is in. The BGP defines the messages that each peer should exchange in order to change the session from one state to another. The first state is the "Idle" state. In the "Idle" state, BGP initializes all resources, refuses all inbound BGP connection attempts and initiates a TCP connection to the peer. The second state is "Connect". In the "Connect" state, the router waits for the TCP connection to complete and transitions to the "OpenSent" state if successful. If unsuccessful, it starts the ConnectRetry timer and transitions to the "Active" state upon expiration. In the "Active" state, the router resets the ConnectRetry timer to zero and returns to the "Connect" state. In the "OpenSent" state, the router sends an Open message and waits for one in return in order to transition to the "OpenConfirm" state. Keepalive messages are exchanged and, upon successful receipt, the router is placed into the "Established" state. In the "Established" state, the router can send/receive: Keepalive; Update; and Notification messages to/from its peer.

* **Idle State**:
  + Refuse all incoming BGP connections.
  + Start the initialization of event triggers.
  + Initiates a TCP connection with its configured BGP peer.
  + Listens for a TCP connection from its peer.
  + Changes its state to Connect.
  + If an error occurs at any state of the FSM process, the BGP session is terminated immediately and returned to the Idle state. Some of the reasons why a router does not progress from the Idle state are:
    - TCP port 179 is not open.
    - A random TCP port over 1023 is not open.
    - Peer address configured incorrectly on either router.
    - AS number configured incorrectly on either router.
* **Connect State**:
  + Waits for successful TCP negotiation with peer.
  + BGP does not spend much time in this state if the TCP session has been successfully established.
  + Sends Open message to peer and changes state to OpenSent.
  + If an error occurs, BGP moves to the Active state. Some reasons for the error are:
    - TCP port 179 is not open.
    - A random TCP port over 1023 is not open.
    - Peer address configured incorrectly on either router.
    - AS number configured incorrectly on either router.
* **Active State**:
  + If the router was unable to establish a successful TCP session, then it ends up in the Active state.
  + BGP FSM tries to restart another TCP session with the peer and, if successful, then it sends an Open message to the peer.
  + If it is unsuccessful again, the FSM is reset to the Idle state.
  + Repeated failures may result in a router cycling between the Idle and Active states. Some of the reasons for this include:
    - TCP port 179 is not open.
    - A random TCP port over 1023 is not open.
    - BGP configuration error.
    - Network congestion.
    - Flapping network interface.
* **OpenSent State**:
  + BGP FSM listens for an Open message from its peer.
  + Once the message has been received, the router checks the validity of the Open message.
  + If there is an error it is because one of the fields in the Open message does not match between the peers, e.g., BGP version mismatch, the peering router expects a different My AS, etc. The router then sends a Notification message to the peer indicating why the error occurred.
  + If there is no error, a Keepalive message is sent, various timers are set and the state is changed to OpenConfirm.
* **OpenConfirm State**:
  + The peer is listening for a Keepalive message from its peer.
  + If a Keepalive message is received and no timer has expired before reception of the Keepalive, BGP transitions to the Established state.
  + If a timer expires before a Keepalive message is received, or if an error condition occurs, the router transitions back to the Idle state.
* **Established State**:
  + In this state, the peers send Update messages to exchange information about each route being advertised to the BGP peer.
  + If there is any error in the Update message then a Notification message is sent to the peer, and BGP transitions back to the Idle state.

**Router connectivity and learning routes**

In the simplest arrangement, all routers within a single AS and participating in BGP routing must be configured in a full mesh: each router must be configured as peer to every other router. This causes scaling problems, since the number of required connections [grows quadratically](https://en.wikipedia.org/wiki/Quadratic_growth) with the number of routers involved. To alleviate the problem, BGP implements two options: [route reflectors](https://en.wikipedia.org/wiki/Route_reflector) ([RFC 4456](https://tools.ietf.org/html/rfc4456)) and [BGP confederations](https://en.wikipedia.org/wiki/BGP_confederation) ([RFC 5065](https://tools.ietf.org/html/rfc5065)). The following discussion of basic UPDATE processing assumes a full iBGP mesh.

A given BGP router may accept Network Layer Reachability Information (NLRI) UPDATEs from multiple neighbors and advertise NLRI to the same, or a different set, of neighbors. Conceptually, BGP maintains its own "master" routing table, called the Local [Routing Information Base](https://en.wikipedia.org/wiki/Routing_Information_Base) (*Loc-RIB*), separate from the main routing table of the router. For each neighbor, the BGP process maintains a conceptual Adjacent Routing Information Base, Incoming (*Adj-RIB-In*) containing the NLRI received from the neighbor, and a conceptual *Adj-RIB-Out* (Outgoing) for NLRI to be sent to the neighbor.

"Conceptual", in the preceding paragraph, means that the physical storage and structure of these various tables are decided by the implementer of the BGP code. Their structure is not visible to other BGP routers, although they usually can be interrogated with management commands on the local router. It is quite common, for example, to store the two Adj-RIBs and the Loc-RIB together in the same data structure, with additional information attached to the RIB entries. The additional information tells the BGP process such things as whether individual entries belong in the Adj-RIBs for specific neighbors, whether the peer-neighbor route selection process made received policies eligible for the Loc-RIB, and whether Loc-RIB entries are eligible to be submitted to the local router's routing table management process.

By *eligible to be submitted*, BGP will submit the routes that it considers best to the main routing table process. Depending on the implementation of that process, the BGP route is not necessarily selected. For example, a directly connected prefix, learned from the router's own hardware, is usually most preferred. As long as that directly connected route's interface is active, the BGP route to the destination will not be put into the routing table. Once the interface goes down, and there are no more preferred routes, the Loc-RIB route would be installed in the main routing table. Until recently, it was a common mistake to say *BGP carries policies*. BGP actually carried the information with which rules inside BGP-speaking routers could make policy decisions. Some of the information carried that is explicitly intended to be used in policy decisions are communities and multi-exit discriminators (MED).

The BGP standard specifies a number of decision factors, more than the ones that are used by any other common routing process, for selecting NLRI to go into the Loc-RIB. The first decision point for evaluating NLRI is that its next-hop attribute must be reachable (or resolvable). Another way of saying the next-hop must be reachable is that there must be an active route, already in the main routing table of the router, to the prefix in which the next-hop address is reachable.

Next, for each neighbor, the BGP process applies various standard and implementation-dependent criteria to decide which routes conceptually should go into the Adj-RIB-In. The neighbor could send several possible routes to a destination, but the first level of preference is at the neighbor level. Only one route to each destination will be installed in the conceptual Adj-RIB-In. This process will also delete, from the Adj-RIB-In, any routes that are withdrawn by the neighbor.

Whenever a conceptual Adj-RIB-In changes, the main BGP process decides if any of the neighbor's new routes are preferred to routes already in the Loc-RIB. If so, it replaces them. If a given route is withdrawn by a neighbor, and there is no other route to that destination, the route is removed from the Loc-RIB, and no longer sent, by BGP, to the main routing table manager. If the router does not have a route to that destination from any non-BGP source, the withdrawn route will be removed from the main routing table.

After verifying that the next hop is reachable, if the route comes from an internal (i.e. iBGP) peer, the first rule to apply, according to the standard, is to examine the LOCAL\_PREFERENCE attribute. If there are several iBGP routes from the neighbor, the one with the highest LOCAL\_PREFERENCE is selected unless there are several routes with the same LOCAL\_PREFERENCE. In the latter case the route selection process moves to the next tie breaker. While LOCAL\_PREFERENCE is the first rule in the standard, once reachability of the NEXT\_HOP is verified, Cisco and several other vendors first consider a decision factor called WEIGHT which is local to the router (i.e. not transmitted by BGP). The route with the highest WEIGHT is preferred.

The LOCAL\_PREFERENCE, WEIGHT, and other criteria can be manipulated by local configuration and software capabilities. Such manipulation is outside the scope of the standard but is commonly used. For example, the COMMUNITY attribute (see below) is not directly used by the BGP selection process. The BGP neighbor process however can have a rule to set LOCAL\_PREFERENCE or another factor based on a manually programmed rule to set the attribute if the COMMUNITY value matches some pattern matching criterion. If the route was learned from an external peer the per-neighbor BGP process computes a LOCAL\_PREFERENCE value from local policy rules and then compares the LOCAL\_PREFERENCE of all routes from the neighbor.

At the per-neighbor level – ignoring implementation-specific policy modifiers – the order of tie breaking rules is:

1. Prefer the route with the shortest AS\_PATH. An AS\_PATH is the set of AS numbers that must be traversed to reach the advertised destination. AS1-AS2-AS3 is shorter than AS4-AS5-AS6-AS7.
2. Prefer routes with the lowest value of their ORIGIN attribute.
3. Prefer routes with the lowest MULTI\_EXIT\_DISC (multi-exit discriminator or MED) value.

*Before the most recent edition of the BGP standard, if an UPDATE had no MULTI\_EXIT\_DISC value, several implementations created a MED with the highest possible value. The current standard however specifies that missing MEDs are to be treated as the lowest possible value. Since the current rule may cause different behavior than the vendor interpretations, BGP implementations that used the nonstandard default value have a configuration feature that allows the old or standard rule to be selected.*

Once candidate routes are received from neighbors, the Loc-RIB software applies additional tie-breakers to routes to the same destination.

1. If at least one route was learned from an external neighbor (i.e., the route was learned from eBGP), drop all routes learned from iBGP.
2. Prefer the route with the lowest interior cost to the NEXT\_HOP, according to the main routing table. If two neighbors advertised the same route, but one neighbor is reachable via a low-bitrate link and the other by a high-bitrate link, and the interior routing protocol calculates lowest cost based on highest bitrate, the route through the high-bitrate link would be preferred and other routes dropped.

*If there is more than one route still tied at this point, several BGP implementations offer a configurable option to load-share among the routes, accepting all (or all up to some number).*

1. Prefer the route learned from the BGP speaker with the numerically lowest BGP identifier
2. Prefer the route learned from the BGP speaker with the lowest peer IP address

**Communities**

BGP communities are attribute tags that can be applied to incoming or outgoing prefixes to achieve some common goal ([RFC 1997](https://tools.ietf.org/html/rfc1997)). While it is common to say that BGP allows an administrator to set policies on how prefixes are handled by ISPs, this is generally not possible, strictly speaking. For instance, BGP natively has no concept to allow one AS to tell another AS to restrict advertisement of a prefix to only North American peering customers. Instead, an ISP generally publishes a list of well-known or proprietary communities with a description for each one, which essentially becomes an agreement of how prefixes are to be treated. [RFC 1997](https://tools.ietf.org/html/rfc1997) also defines three well-known communities that have global significance; NO\_EXPORT, NO\_ADVERTISE and NO\_EXPORT\_SUBCONFED. [RFC 7611](https://tools.ietf.org/html/rfc7611) defines ACCEPT\_OWN. Examples of common communities include local preference adjustments, geographic or peer type restrictions, DoS avoidance (black holing), and AS prepending options. An ISP might state that any routes received from customers with community XXX:500 will be advertised to all peers (default) while community XXX:501 will restrict advertisement to North America. The customer simply adjusts their configuration to include the correct community or communities for each route, and the ISP is responsible for controlling who the prefix is advertised to. The end user has no technical ability to enforce correct actions being taken by the ISP, though problems in this area are generally rare and accidental.

It is a common tactic for end customers to use BGP communities (usually ASN:70,80,90,100) to control the local preference the ISP assigns to advertised routes instead of using MED (the effect is similar). The community attribute is transitive, but communities applied by the customer very rarely become propagated outside the next-hop AS. Not all ISPs give out their communities to the public, while some other do.[[11]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-11)

The BGP Extended Community Attribute was added in 2006, in order to extend the range of such attributes and to provide a community attribute structuring by means of a type field. The extended format consists of one or two octets for the type field followed by seven or six octets for the respective community attribute content. The definition of this Extended Community Attribute is documented in [RFC 4360](https://tools.ietf.org/html/rfc4360). The IANA administers the registry for BGP Extended Communities Types.[[12]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-12) The Extended Communities Attribute itself is a transitive optional BGP attribute. However, a bit in the type field within the attribute decides whether the encoded extended community is of a transitive or non-transitive nature. The IANA registry therefore provides different number ranges for the attribute types. Due to the extended attribute range, its usage can be manifold. [RFC 4360](https://tools.ietf.org/html/rfc4360) exemplarly[[*check spelling*](https://en.wikipedia.org/wiki/Template:Typo_help_inline)] defines the "Two-Octet AS Specific Extended Community", the "IPv4 Address Specific Extended Community", the "Opaque Extended Community", the "Route Target Community", and the "Route Origin Community". A number of BGP QoS drafts[[13]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-13) also use this Extended Community Attribute structure for inter-domain QoS signalling.

Note: since [RFC 7153](https://tools.ietf.org/html/rfc7153), extended communities are compatible with 32 bits ASNs.

With the introduction of 32 bits AS numbers, some issues were immediately obvious with the community attribute that only defines a 16 bits ASN field, which prevents the matching between this field and the real ASN value. It is the reason why [RFC 8092](https://tools.ietf.org/html/rfc8092) and [RFC 8195](https://tools.ietf.org/html/rfc8195) introduce a [Large Community](http://largebgpcommunities.net/) attribute of 12 bytes, divided in three field of 4 bytes each (AS:function:parameter).

**Multi-exit discriminators**

MEDs, defined in the main BGP standard, were originally intended to show to another neighbor AS the advertising AS's preference as to which of several links are preferred for inbound traffic. Another application of MEDs is to advertise the value, typically based on delay, of multiple AS that have presence at an [IXP](https://en.wikipedia.org/wiki/IXP), that they impose to send traffic to some destination.

**Message header format**

The following is the BGP version 4 message header format:

|  |  |  |  |
| --- | --- | --- | --- |
| **bit offset** | **0–15** | **16–23** | **24–31** |
| **0** | Marker | | |
| **32** |
| **64** |
| **96** |
| **128** | Length | Type |  |

* **Marker**: Included for compatibility, must be set to all ones.
* **Length**: Total length of the message in [octets](https://en.wikipedia.org/wiki/Octet_(computing)), including the header.
* **Type**: Type of BGP message. The following values are defined:
  + Open (1)
  + Update (2)
  + Notification (3)
  + KeepAlive (4)
  + Route-Refresh (5)

**Internal scalability**

An autonomous system with internal BGP (iBGP) must have all of its iBGP peers connect to each other in a [full mesh](https://en.wikipedia.org/wiki/Complete_graph) (where everyone speaks to everyone directly). This full-mesh configuration requires that each router maintain a session to every other router. In large networks, this number of sessions may degrade performance of routers, due to either a lack of memory, or high CPU process requirements.

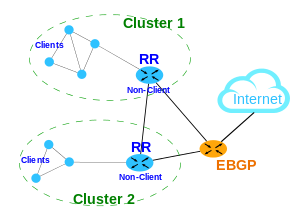
**Route reflectors**

[Route reflectors](https://en.wikipedia.org/wiki/Route_reflector)[[14]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-14) reduce the number of connections required in an AS. A single router (or two for redundancy) can be made a route reflector: other routers in the AS need only be configured as peers to them. A route reflector offers an alternative to the logical full-mesh requirement of [internal border gateway protocol](https://en.wikipedia.org/wiki/Internal_border_gateway_protocol) (IBGP). A RR acts as a focal point[[*clarify*](https://en.wikipedia.org/wiki/Wikipedia:Please_clarify)] for IBGP sessions. The purpose of the RR is concentration. Multiple BGP routers can peer with a central point, the RR – acting as a route reflector server – rather than peer with every other router in a full mesh. All the other IBGP routers become route reflector clients.

This approach, similar to [OSPF](https://en.wikipedia.org/wiki/OSPF)'s DR/BDR feature, provides large networks with added IBGP scalability. In a fully meshed IBGP network of 10 routers, 90 individual CLI statements (spread throughout all routers in the topology) are needed just to define the remote-AS of each peer: this quickly becomes a headache to administer. A RR topology could cut these 90 statements down to 18, offering a viable solution for the larger networks administered by ISPs.

A route reflector is a [single point of failure](https://en.wikipedia.org/wiki/Single_point_of_failure), therefore at least a second route reflector may be configured in order to provide redundancy. As it is an additional peer for the other 10 routers, it comes with the additional statement count to double that minus 2 of the single Route Reflector setup. An additional 11\*2-2=20 statements in this case due to adding the additional Router. Additionally, in a BGP multipath Environment this also can benefit by adding local switching/Routing throughput if the RRs are acting as traditional Routers instead of just a dedicated Route Reflector Server role.

**Rules**

[](https://en.wikipedia.org/wiki/File:RR_BGP.svg)

A typical configuration of BGP Route Reflector deployment, as proposed by Section 6, [RFC 4456](https://tools.ietf.org/html/rfc4456).

RR servers propagate routes inside the AS based on the following rules:

* If a route is received from a non-client peer, reflect to clients only and EBGP peers.
* If a route is received from a client peer, reflect to all non-client peers and also to client peers, except the originator of the route and reflect to EBGP peers.

**Cluster**

RR and its clients form a "Cluster". The "Cluster-ID" is then attached to every route advertised by RR to its client or nonclient peers. Cluster-ID is a cumulative, non-transitive BGP attribute and every RR MUST prepend the local CLUSTER\_ID to the CLUSTER\_LIST in order to avoid routing loops. Route reflectors and confederations both reduce the number of iBGP peers to each router and thus reduce processing overhead. Route reflectors are a pure performance-enhancing technique, while confederations also can be used to implement more fine-grained policy.

**BGP confederation**

Confederations are sets of autonomous systems. In common practice,[[15]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-15) only one of the confederation AS numbers is seen by the Internet as a whole. Confederations are used in very large networks where a large AS can be configured to encompass smaller more manageable internal ASs.

The confederated AS is composed of multiple ASs. Each confederated AS alone has iBGP fully meshed and has connections to other ASs inside the confederation. Even though these ASs have eBGP peers to ASs within the confederation, the ASs exchange routing as if they used iBGP. In this way, the confederation preserves next hop, metric, and local preference information. To the outside world, the confederation appears to be a single AS. With this solution, iBGP transit AS problems can be resolved as iBGP requires a full mesh between all BGP routers: large number of TCP sessions and unnecessary duplication of routing traffic.

Confederations can be used in conjunction with route reflectors. Both confederations and route reflectors can be subject to persistent oscillation unless specific design rules, affecting both BGP and the interior routing protocol, are followed.[[16]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-16)

However, these alternatives can introduce problems of their own, including the following:

* route oscillation
* sub-optimal routing
* increase of BGP convergence time[[17]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-17)

Additionally, route reflectors and BGP confederations were not designed to ease BGP router configuration. Nevertheless, these are common tools for experienced BGP network architects. These tools may be combined, for example, as a hierarchy of route reflectors.

**Stability**

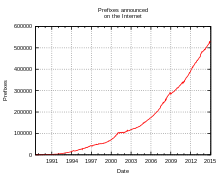
The routing tables managed by a BGP implementation are adjusted continually to reflect actual changes in the network, such as links breaking and being restored or routers going down and coming back up. In the network as a whole it is normal for these changes to happen almost continuously, but for any particular router or link, changes are supposed to be relatively infrequent. If a router is misconfigured or mismanaged then it may get into a rapid cycle between down and up states. This pattern of repeated withdrawal and re-announcement known as [route flapping](https://en.wikipedia.org/wiki/Route_flapping) can cause excessive activity in all the other routers that know about the broken link, as the same route is continually injected and withdrawn from the routing tables. The BGP design is such that delivery of traffic may not function while routes are being updated. On the Internet, a BGP routing change may cause outages for several minutes.

A feature known as *route flap damping* ([RFC 2439](http://www.ietf.org/rfc/rfc2439.txt)) is built into many BGP implementations in an attempt to mitigate the effects of route flapping. Without damping, the excessive activity can cause a heavy processing load on routers, which may in turn delay updates on other routes, and so affect overall routing stability. With damping, a route's flapping is [exponentially decayed](https://en.wikipedia.org/wiki/Exponential_decay). At the first instance when a route becomes unavailable and quickly reappears, damping does not take effect, so as to maintain the normal fail-over times of BGP. At the second occurrence, BGP shuns that prefix for a certain length of time; subsequent occurrences are timed out exponentially. After the abnormalities have ceased and a suitable length of time has passed for the offending route, prefixes can be reinstated and its slate wiped clean. Damping can also mitigate [denial of service](https://en.wikipedia.org/wiki/Denial_of_service) attacks; damping timings are highly customizable.

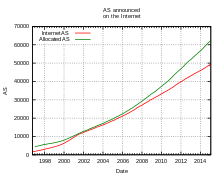
It is also suggested in [RFC 2439](https://tools.ietf.org/html/rfc2439) (under "Design Choices -> Stability Sensitive Suppression of Route Advertisement") that route flap damping is a feature more desirable if implemented to Exterior Border Gateway Protocol Sessions (eBGP sessions or simply called exterior peers) and not on Interior Border Gateway Protocol Sessions (iBGP sessions or simply called internal peers); With this approach when a route flaps inside an autonomous system, it is not propagated to the external ASs – flapping a route to an eBGP will have a chain of flapping for the particular route throughout the backbone. This method also successfully avoids the overhead of route flap damping for iBGP sessions.

However, subsequent research has shown that flap damping can actually lengthen convergence times in some cases, and can cause interruptions in connectivity even when links are not flapping.[[18]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-18)[[19]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-19) Moreover, as backbone links and router processors have become faster, some network architects have suggested that flap damping may not be as important as it used to be, since changes to the routing table can be handled much faster by routers.[[20]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-20) This has led the RIPE Routing Working Group to write that "with the current implementations of BGP flap damping, the application of flap damping in ISP networks is NOT recommended. ... If flap damping is implemented, the ISP operating that network will cause side-effects to their customers and the Internet users of their customers' content and services ... . These side-effects would quite likely be worse than the impact caused by simply not running flap damping at all."[[21]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-21) Improving stability without the problems of flap damping is the subject of current research.[[22]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-22)

**Routing table growth**

[](https://en.wikipedia.org/wiki/File:BGP_Table_growth.svg)

BGP table growth on the Internet

[](https://en.wikipedia.org/wiki/File:Internet_AS.svg)

Number of AS on the Internet vs number of registered AS

One of the largest problems faced by BGP, and indeed the Internet infrastructure as a whole, is the growth of the Internet routing table. If the global routing table grows to the point where some older, less capable routers cannot cope with the memory requirements or the CPU load of maintaining the table, these routers will cease to be effective gateways between the parts of the Internet they connect. In addition, and perhaps even more importantly, larger routing tables take longer to stabilize (see above) after a major connectivity change, leaving network service unreliable, or even unavailable, in the interim.

Until late 2001, the global routing table was [growing exponentially](https://en.wikipedia.org/wiki/Exponential_growth), threatening an eventual widespread breakdown of connectivity. In an attempt to prevent this, ISPs cooperated in keeping the global routing table as small as possible, by using [Classless Inter-Domain Routing](https://en.wikipedia.org/wiki/Classless_Inter-Domain_Routing) (CIDR) and [route aggregation](https://en.wikipedia.org/wiki/Route_aggregation). While this slowed the growth of the routing table to a linear process for several years, with the expanded demand for [multihoming](https://en.wikipedia.org/wiki/Multihoming) by end user networks the growth was once again superlinear by the middle of 2004.

**512k day**

A Y2K-like overflow triggered in 2014 for those models that were not appropriately updated.

While a full IPv4 BGP table as of August 2014 (512k day)[[23]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-23)[[24]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-24) is in excess of 512,000 prefixes,[[25]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-Potaroo_—_BGP_Table_data-25) many older routers have a limit of 512k (512,000–524,288)[[26]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-26)[[27]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-27) routing table entries. On August 12, 2014, outages resulting from full tables hit [eBay](https://en.wikipedia.org/wiki/EBay), [LastPass](https://en.wikipedia.org/wiki/LastPass) and [Microsoft Azure](https://en.wikipedia.org/wiki/Microsoft_Azure) among others.[[28]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-28) A number of Cisco routers commonly in use have [TCAM](https://en.wikipedia.org/wiki/Ternary_Content-Addressable_Memory), a form of high-speed [content-addressable memory](https://en.wikipedia.org/wiki/Content-addressable_memory), for storing BGP advertised routes. On impacted routers, the TCAM is default allocated to 512k entries for IPv4 routes, and 512k entries for IPv6 routes. While the reported number of IPv6 advertised routes was only about 20k, the number of advertised IPv4 routes reached the default limit, causing a [spillover effect](https://en.wikipedia.org/wiki/Spillover_effect) as routers attempted to compensate for the issue by using slow software routing (as opposed to fast hardware routing via TCAM). The main method for dealing with this issue involves operators changing the TCAM allocation to allow more IPv4 entries, by reallocating some of the TCAM reserved for IPv6 routes. This requires a reboot on most routers. The 512k problem was predicted in advance by a number of IT professionals.[[29]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-29)[[30]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-30)[[31]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-31)

The actual allocations which pushed the number of routes above 512k was the announcement of about 15,000 new routes in short order, starting at 07:48 UTC. Almost all of these routes were to [Verizon](https://en.wikipedia.org/wiki/Verizon_Communications) [Autonomous Systems](https://en.wikipedia.org/wiki/Autonomous_System_(Internet)) 701 and 705, created as a result of [deaggregation](https://en.wikipedia.org/wiki/Route_aggregation) of larger blocks, introducing thousands of new [*/24* routes](https://en.wikipedia.org/wiki/Classless_Inter-Domain_Routing), and making the routing table reach 515,000 entries. The new routes appear to have been reaggregated within 5 minutes, but instability across the Internet apparently continued for a number of hours.[[32]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-32) Even if Verizon had not caused the routing table to exceed 512k entries in the short spike, it would have happened soon anyway through natural growth.

Route summarization is often used to improve aggregation of the BGP global routing table, thereby reducing the necessary table size in routers of an AS. Consider AS1 has been allocated the big address space of *172.16.0.0/16*, this would be counted as one route in the table, but due to customer requirement or traffic engineering purposes, AS1 wants to announce smaller, more specific routes of *172.16.0.0/18*, *172.16.64.0/18*, and *172.16.128.0/18*. The prefix *172.16.192.0/18* does not have any hosts so AS1 does not announce a specific route *172.16.192.0/18*. This all counts as AS1 announcing four routes.

AS2 will see the four routes from AS1 (*172.16.0.0/16*, *172.16.0.0/18*, *172.16.64.0/18*, and *172.16.128.0/18*) and it is up to the routing policy of AS2 to decide whether or not to take a copy of the four routes or, as *172.16.0.0/16* overlaps all the other specific routes, to just store the summary, *172.16.0.0/16*.

If AS2 wants to send data to prefix *172.16.192.0/18*, it will be sent to the routers of AS1 on route *172.16.0.0/16*. At AS1's router, it will either be dropped or a destination unreachable [ICMP](https://en.wikipedia.org/wiki/Internet_Control_Message_Protocol) message will be sent back, depending on the configuration of AS1's routers.

If AS1 later decides to drop the route *172.16.0.0/16*, leaving *172.16.0.0/18*, *172.16.64.0/18*, and *172.16.128.0/18*, AS1 will drop the number of routes it announces to three. AS2 will see the three routes, and depending on the routing policy of AS2, it will store a copy of the three routes, or aggregate the prefix's *172.16.0.0/18* and *172.16.64.0/18* to *172.16.0.0/17*, thereby reducing the number of routes AS2 stores to only two: *172.16.0.0/17* and *172.16.128.0/18*.

If AS2 wants to send data to prefix *172.16.192.0/18*, it will be dropped or a destination unreachable ICMP message will be sent back at the routers of AS2 (not AS1 as before), because *172.16.192.0/18* would not be in the routing table.

**AS numbers depletion and 32 bits ASN**

The [RFC 1771](https://tools.ietf.org/html/rfc1771) (*A Border Gateway Protocol 4 (BGP-4)*) planned the coding of AS numbers on 16 bits, for 64510 possible public AS, since ASN 64512 to 65534 were reserved for private use (0 and 65535 being forbidden). In 2011, only 15000 AS numbers were still available, and projections[[33]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-33) were envisioning a complete depletion of available AS numbers in September 2013.

[RFC 6793](https://tools.ietf.org/html/rfc6793) extends AS coding from 16 to 32 bits (keeping the 16 bits AS range 0 to 65535, and its reserved AS numbers), which now allows up to 4 billion available AS. An additional private AS range is also defined in [RFC 6996](https://tools.ietf.org/html/rfc6996) (from 4200000000 to 4294967294, 4294967295 being forbidden by [RFC 7300](https://tools.ietf.org/html/rfc7300)).

To allow the traversal of router groups not able to manage those new ASNs, the new attribute OT AS4\_PATH is used.

32 bits ASN assignments started in 2007.

**Load balancing**

Another factor causing this growth of the routing table is the need for load balancing of multi-homed networks. It is not a trivial task to balance the inbound traffic to a multi-homed network across its multiple inbound paths, due to limitation of the BGP route selection process. For a multi-homed network, if it announces the same network blocks across all of its BGP peers, the result may be that one or several of its inbound links become congested while the other links remain under-utilized, because external networks all picked that set of congested paths as optimal. Like most other routing protocols, BGP does not detect congestion.

To work around this problem, BGP administrators of that multihomed network may divide a large contiguous IP address block into smaller blocks and tweak the route announcement to make different blocks look optimal on different paths, so that external networks will choose a different path to reach different blocks of that multi-homed network. Such cases will increase the number of routes as seen on the global BGP table.

One method growing in popularity to address the load balancing issue is to deploy BGP/LISP ([Locator/Identifier Separation Protocol](https://en.wikipedia.org/wiki/Locator/Identifier_Separation_Protocol)) gateways within an [Internet exchange point](https://en.wikipedia.org/wiki/Internet_exchange_point) to allow ingress traffic engineering across multiple links. This technique does not increase the number of routes seen on the global BGP table.

**Security**

By design, routers running BGP accept advertised routes from other BGP routers by default. This allows for automatic and decentralized routing of traffic across the Internet, but it also leaves the Internet potentially vulnerable to accidental or malicious disruption, known as [BGP hijacking](https://en.wikipedia.org/wiki/BGP_hijacking). Due to the extent to which BGP is embedded in the core systems of the Internet, and the number of different networks operated by many different organizations which collectively make up the Internet, correcting this vulnerability (such as by introducing the use of cryptographic keys to verify the identity of BGP routers) is a technically and economically challenging problem.[[34]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-34)

**Extensions**

An extension to BGP is the use of multipathing – this typically requires identical MED, weight, origin, and AS-path although some implementations provide the ability to relax the AS-path checking to only expect an equal path length rather than the actual AS numbers in the path being expected to match too. This can then be extended further with features like Cisco's dmzlink-bw which enables a ratio of traffic sharing based on bandwidth values configured on individual links.

Multiprotocol Extensions for BGP (MBGP), sometimes referred to as Multiprotocol BGP or Multicast BGP and defined in IETF RFC 4760, is an extension to (BGP) that allows different types of addresses (known as address families) to be distributed in parallel. Whereas standard BGP supports only IPv4 unicast addresses, Multiprotocol BGP supports IPv4 and IPv6 addresses and it supports unicast and multicast variants of each. Multiprotocol BGP allows information about the topology of IP multicast-capable routers to be exchanged separately from the topology of normal IPv4 unicast routers. Thus, it allows a multicast routing topology different from the unicast routing topology. Although MBGP enables the exchange of inter-domain multicast routing information, other protocols such as the Protocol Independent Multicast family are needed to build trees and forward multicast traffic.

Multiprotocol BGP is also widely deployed in case of MPLS L3 VPN, to exchange VPN labels learned for the routes from the customer sites over the MPLS network, in order to distinguish between different customer sites when the traffic from the other customer sites comes to the Provider Edge router (PE router) for routing.

**Uses**

BGP4 is standard for Internet routing and required of most [Internet service providers](https://en.wikipedia.org/wiki/Internet_service_provider) (ISPs) to establish routing between one another. Very large private [IP](https://en.wikipedia.org/wiki/Internet_Protocol) networks use BGP internally. An example is the joining of a number of large [Open Shortest Path First](https://en.wikipedia.org/wiki/Open_Shortest_Path_First) (OSPF) networks, when OSPF by itself does not scale to the size required. Another reason to use BGP is multihoming a network for better redundancy, either to multiple access points of a single ISP or to multiple ISPs.

**Implementations**

Routers, especially small ones intended for [Small Office/Home Office](https://en.wikipedia.org/wiki/Small_office/home_office) (SOHO) use, may not include BGP software. Some SOHO routers simply are not capable of running BGP / using BGP routing tables of any size. Other commercial routers may need a specific software executable image that contains BGP, or a license that enables it. Open source packages that run BGP include [GNU Zebra](https://en.wikipedia.org/wiki/GNU_Zebra), [Quagga](https://en.wikipedia.org/wiki/Quagga_(software)), [OpenBGPD](https://en.wikipedia.org/wiki/OpenBGPD), [BIRD](https://en.wikipedia.org/wiki/Bird_Internet_routing_daemon), [XORP](https://en.wikipedia.org/wiki/XORP), and [Vyatta](https://en.wikipedia.org/wiki/Vyatta). Devices marketed as [Layer 3 switches](https://en.wikipedia.org/wiki/Layer_3_switch) are less likely to support BGP than devices marketed as [routers](https://en.wikipedia.org/wiki/Router_(computing)), but high-end Layer 3 Switches usually can run BGP.

Products marketed as switches may or may not have a size limitation on BGP tables, such as 20,000 routes, far smaller than a full Internet table plus internal routes. These devices, however, may be perfectly reasonable and useful when used for BGP routing of some smaller part of the network, such as a [confederation-AS](https://en.wikipedia.org/w/index.php?title=Confederation-AS&action=edit&redlink=1) representing one of several smaller enterprises that are linked, by a BGP [backbone of backbones](https://en.wikipedia.org/w/index.php?title=Backbone_of_backbones&action=edit&redlink=1), or a small enterprise that announces routes to an ISP but only accepts a [default route](https://en.wikipedia.org/wiki/Default_route) and perhaps a small number of aggregated routes.

A BGP router used only for a network with a single point of entry to the Internet may have a much smaller routing table size (and hence RAM and CPU requirement) than a multihomed network. Even simple multihoming can have modest routing table size. See [RFC 4098](https://tools.ietf.org/html/rfc4098) for vendor-independent performance parameters for single BGP router convergence in the control plane. The actual amount of memory required in a BGP router depends on the amount of BGP information exchanged with other BGP speakers and the way in which the particular router stores BGP information. The router may have to keep more than one copy of a route, so it can manage different policies for route advertising and acceptance to a specific neighboring AS. The term view is often used for these different policy relationships on a running router.

If one router implementation takes more memory per route than another implementation, this may be a legitimate design choice, trading processing speed against memory. A full IPv4 BGP table as of August 2015 is in excess of 590,000 prefixes.[[25]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-Potaroo_—_BGP_Table_data-25) Large ISPs may add another 50% for internal and customer routes. Again depending on implementation, separate tables may be kept for each view of a different peer AS.

Notable free and open source implementations of BGP include:

* [BIRD Internet Routing Daemon](https://en.wikipedia.org/wiki/Bird_Internet_routing_daemon), a GPL routing package for Unix-like systems.
* [FRRouting](https://en.wikipedia.org/wiki/FRRouting), a fork of Quagga for [Unix-like](https://en.wikipedia.org/wiki/Unix-like) systems.
* [GNU Zebra](https://en.wikipedia.org/wiki/GNU_Zebra), a [GPL](https://en.wikipedia.org/wiki/GNU_General_Public_License) routing suite supporting BGP4. (decommissioned)[[35]](https://en.wikipedia.org/wiki/Border_Gateway_Protocol#cite_note-35)
* [OpenBGPD](https://en.wikipedia.org/wiki/OpenBGPD), a [BSD licensed](https://en.wikipedia.org/wiki/BSD_licence) implementation by the [OpenBSD](https://en.wikipedia.org/wiki/OpenBSD) team.
* [Quagga](https://en.wikipedia.org/wiki/Quagga_(software)), a fork of GNU Zebra for [Unix-like](https://en.wikipedia.org/wiki/Unix-like) systems.
* [XORP](https://en.wikipedia.org/wiki/XORP), the eXtensible Open Router Platform, a BSD licensed suite of routing protocols.

Systems for testing BGP conformance, load or stress performance come from vendors such as:

* [Agilent Technologies](https://en.wikipedia.org/wiki/Agilent_Technologies)
* [GNS3](https://en.wikipedia.org/wiki/GNS3) [open source](https://en.wikipedia.org/wiki/Open-source_software) [network simulator](https://en.wikipedia.org/wiki/Network_simulator)
* [Ixia](https://en.wikipedia.org/wiki/Ixia_(company))
* [Spirent Communications](https://en.wikipedia.org/wiki/Spirent_Communications)

**Standards documents**

* [Selective Route Refresh for BGP](https://tools.ietf.org/html/draft-utgikar-serr-00), IETF draft
* [RFC 1772](https://tools.ietf.org/html/rfc1772), Application of the Border Gateway Protocol in the Internet Protocol (BGP-4) using SMIv2
* [RFC 2439](https://tools.ietf.org/html/rfc2439), BGP Route Flap Damping
* [RFC 2918](https://tools.ietf.org/html/rfc2918), Route Refresh Capability for BGP-4
* [RFC 3765](https://tools.ietf.org/html/rfc3765), NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control
* [RFC 4271](https://tools.ietf.org/html/rfc4271), A Border Gateway Protocol 4 (BGP-4)
* [RFC 4272](https://tools.ietf.org/html/rfc4272), BGP Security Vulnerabilities Analysis
* [RFC 4273](https://tools.ietf.org/html/rfc4273), Definitions of Managed Objects for BGP-4
* [RFC 4274](https://tools.ietf.org/html/rfc4274), BGP-4 Protocol Analysis
* [RFC 4275](https://tools.ietf.org/html/rfc4275), BGP-4 MIB Implementation Survey
* [RFC 4276](https://tools.ietf.org/html/rfc4276), BGP-4 Implementation Report
* [RFC 4277](https://tools.ietf.org/html/rfc4277), Experience with the BGP-4 Protocol
* [RFC 4278](https://tools.ietf.org/html/rfc4278), Standards Maturity Variance Regarding the TCP MD5 Signature Option ([RFC 2385](https://tools.ietf.org/html/rfc2385)) and the BGP-4 Specification
* [RFC 4456](https://tools.ietf.org/html/rfc4456), BGP Route Reflection – An Alternative to Full Mesh Internal BGP (iBGP)
* [RFC 4724](https://tools.ietf.org/html/rfc4724), Graceful Restart Mechanism for BGP
* [RFC 4760](https://tools.ietf.org/html/rfc4760), Multiprotocol Extensions for BGP-4
* [RFC 4893](https://tools.ietf.org/html/rfc4893), BGP Support for Four-octet AS Number Space
* [RFC 5065](https://tools.ietf.org/html/rfc5065), Autonomous System Confederations for BGP
* [RFC 5492](https://tools.ietf.org/html/rfc5492), Capabilities Advertisement with BGP-4
* [RFC 5575](https://tools.ietf.org/html/rfc5575), Dissemination of Flow Specification Rules
* [RFC 7752](https://tools.ietf.org/html/rfc7752), North-Bound Distribution of Link-State and Traffic Engineering Information Using BGP
* [RFC 7911](https://tools.ietf.org/html/rfc7911), Advertisement of Multiple Paths in BGP
* [draft-ietf-idr-custom-decision-08](https://tools.ietf.org/html/draft-ietf-idr-custom-decision-08) – BGP Custom Decision Process, Feb 3, 2017
* [RFC 3392](https://tools.ietf.org/html/rfc3392), Obsolete – Capabilities Advertisement with BGP-4
* [RFC 2796](https://tools.ietf.org/html/rfc2796), Obsolete – BGP Route Reflection – An Alternative to Full Mesh iBGP
* [RFC 3065](https://tools.ietf.org/html/rfc3065), Obsolete – Autonomous System Confederations for BGP
* [RFC 1965](https://tools.ietf.org/html/rfc1965), Obsolete – Autonomous System Confederations for BGP
* [RFC 1771](https://tools.ietf.org/html/rfc1771), Obsolete – A Border Gateway Protocol 4 (BGP-4)
* [RFC 1657](https://tools.ietf.org/html/rfc1657), Obsolete – Definitions of Managed Objects for the Fourth Version of the Border Gateway
* [RFC 1655](https://tools.ietf.org/html/rfc1655), Obsolete – Application of the Border Gateway Protocol in the Internet
* [RFC 1654](https://tools.ietf.org/html/rfc1654), Obsolete – A Border Gateway Protocol 4 (BGP-4)
* [RFC 1105](https://tools.ietf.org/html/rfc1105), Obsolete – Border Gateway Protocol (BGP)
* [RFC 2858](https://tools.ietf.org/html/rfc2858), Obsolete – Multiprotocol Extensions for BGP-4

**See also**

* [AS 7007 incident](https://en.wikipedia.org/wiki/AS_7007_incident)
* [Internet Assigned Numbers Authority](https://en.wikipedia.org/wiki/Internet_Assigned_Numbers_Authority)
* [Packet forwarding](https://en.wikipedia.org/wiki/Packet_forwarding)
* [Private IP](https://en.wikipedia.org/wiki/Private_IP)
* [QPPB](https://en.wikipedia.org/wiki/QPPB)
* [Regional Internet registry](https://en.wikipedia.org/wiki/Regional_Internet_registry)
* [Route analytics](https://en.wikipedia.org/wiki/Route_analytics)
* [Route filtering](https://en.wikipedia.org/wiki/Route_filtering)
* [Routing Assets Database](https://en.wikipedia.org/wiki/Routing_Assets_Database)

**References**

 [*"BGP: Border Gateway Protocol Explained"*](https://web.archive.org/web/20130928115120/http:/www.orbit-computer-solutions.com/BGP.php)*. Orbit-Computer Solutions.Com. Archived from* [*the original*](http://www.orbit-computer-solutions.com/BGP.php) *on 2013-09-28. Retrieved 2013-10-08.*

  *Sobrinho, João Luís (2003).* [*"Network Routing with Path Vector Protocols: Theory and Applications"*](https://conferences.sigcomm.org/sigcomm/2003/papers/p49-sobrinho.pdf) *(PDF). Retrieved March 16, 2018.*

  [*"The History of Border Gateway Protocol"*](https://datapath.io/resources/blog/the-history-of-border-gateway-protocol/)*. blog.datapath.io.*

  [*"RFC 4271 - A Border Gateway Protocol 4 (BGP-4)"*](http://tools.ietf.org/html/rfc4271)*. ietf.org.*

  [*"Border Gateway Protocol (BGP)"*](https://www.cisco.com/c/en/us/products/ios-nx-os-software/border-gateway-protocol-bgp/index.html)*.* [*Cisco*](https://en.wikipedia.org/wiki/Cisco)*.com.*

  [*"BGP Keepalive Messages"*](http://www.inetdaemon.com/tutorials/internet/ip/routing/bgp/operation/messages/keepalives.shtml)*. InetDaemon's IT Tutorials.*

  [RFC 4274](https://tools.ietf.org/html/rfc4274)

  [Capabilities Advertisement with BGP-4](http://www.ietf.org/rfc/rfc2842.txt), [RFC 2842](https://tools.ietf.org/html/rfc2842), R. Chandra & J. Scudder, May 2000

  [Multiprotocol Extensions for BGP-4](http://www.ietf.org/rfc/rfc2858.txt), [RFC 2858](https://tools.ietf.org/html/rfc2858), T. Bates *et al.*, June 2000

  [BGP/MPLS VPNs.](http://www.ietf.org/rfc/rfc2547.txt), [RFC 2547](https://tools.ietf.org/html/rfc2547), E. Rosen and Y. Rekhter, April 2004

  [*"BGP Community Guides"*](http://www.onesc.net/communities/)*. Retrieved 13 April 2015.*

  [IANA registry for BGP Extended Communities Types](http://www.iana.org/assignments/bgp-extended-communities), IANA,2008

  [IETF drafts on BGP signalled QoS](http://www.bgp-qos.org/forum/viewforum.php?f=6) [Archived](https://web.archive.org/web/20090223214439/http:/www.bgp-qos.org/forum/viewforum.php?f=6) 2009-02-23 at the [Wayback Machine](https://en.wikipedia.org/wiki/Wayback_Machine), Thomas Knoll,2008

  [BGP Route Reflection: An Alternative to Full Mesh Internal BGP (iBGP)](http://www.ietf.org/rfc/rfc4456.txt), [RFC 4456](https://tools.ietf.org/html/rfc4456), T. Bates *et al.*, April 2006

  [*"Info"*](http://www.ietf.org/rfc/rfc5065.txt)*. www.ietf.org. Retrieved 2019-12-17.*

  [*"Info"*](http://www.ietf.org/rfc/rfc3345.txt)*. www.ietf.org. Retrieved 2019-12-17.*

  [*"Info"*](http://www.ietf.org/rfc/rfc4098.txt)*. www.ietf.org. Retrieved 2019-12-17.*

  [*"Route Flap Damping Exacerbates Internet Routing Convergence"*](http://web.eecs.umich.edu/~zmao/Papers/sig02.pdf) *(PDF). November 1998.*

  *Zhang, Beichuan; Pei Dan; Daniel Massey;* [*Lixia Zhang*](https://en.wikipedia.org/wiki/Lixia_Zhang) *(June 2005).* [*"Timer Interaction in Route Flap Damping"*](http://www.cs.arizona.edu/~bzhang/paper/05-icdcs-dtimer.pdf) *(PDF). IEEE 25th* [*International Conference on Distributed Computing Systems*](https://en.wikipedia.org/wiki/International_Conference_on_Distributed_Computing_Systems)*. Retrieved 2006-09-26. We show that the current damping design leads to the intended behavior only under persistent route flapping. When the number of flaps is small, the global routing dynamics deviates significantly from the expected behavior with a longer convergence delay.*

  [*"BGP Route Flap Damping"*](https://tools.ietf.org/html/rfc2439)*. Tools.ietf.org.*

  *10 May 2006 (2006-05-10).* [*"RIPE Routing Working Group Recommendations On Route-flap Damping"*](http://www.ripe.net/ripe/docs/ripe-378)*. RIPE Network Coordination Centre. Retrieved 2013-12-04.*

  [*"draft-ymbk-rfd-usable-02 - Making Route Flap Damping Usable"*](http://tools.ietf.org/html/draft-ymbk-rfd-usable)*. Tools.ietf.org. Retrieved 2013-12-04.*

  as of the 12th of August 2014, multiple [Internet routers](https://en.wikipedia.org/wiki/Router_(computing)), manufactured by [Cisco](https://en.wikipedia.org/wiki/Cisco_Systems) and other vendors, encountered a default software limit of 512K (512,000 - 524,288) [*"Cisco switch problem"*](http://www.cisco.com/c/en/us/support/docs/switches/catalyst-6500-series-switches/117712-problemsolution-cat6500-00.html#.U-okMKwClYO.twitter)*.*

  [*"Renesys 512k global routes"*](http://www.renesys.com/2014/08/internet-512k-global-routes)*.*

  [*"BGP Reports"*](http://bgp.potaroo.net/index-bgp.html)*. potaroo.net.*

  [*"CAT 6500 and 7600 Series Routers and Switches TCAM Allocation Adjustment Procedures"*](http://www.cisco.com/c/en/us/support/docs/switches/catalyst-6500-series-switches/117712-problemsolution-cat6500-00.html#.U-okMKwClYO.twitter)*. Cisco. 9 March 2015.*

  *Jim Cowie.* [*"Internet Touches Half Million Routes: Outages Possible Next Week"*](http://www.renesys.com/2014/08/internet-512k-global-routes/)*. Dyn Research.*

  *Garside, Juliette; Gibbs, Samuel (14 August 2014).* [*"Internet infrastructure 'needs updating or more blackouts will happen'"*](https://www.theguardian.com/technology/2014/aug/14/internet-infrastructure-needs-updating-more-blackouts-will-happen)*. The Guardian. Retrieved 15 Aug 2014.*

  [*"BOF report"*](https://www.nanog.org/meetings/nanog39/presentations/bof-report.pdf) *(PDF). www.nanog.org. Retrieved 2019-12-17.*

  *Greg Ferro.* [*"TCAM — a Deeper Look and the impact of IPv6"*](http://etherealmind.com/tcam-detail-review/)*. EtherealMind.*

  [*"The IPv4 Depletion site"*](http://www.ipv4depletion.com/?p=672)*. ipv4depletion.com.*

  [*"What caused today's Internet hiccup"*](https://www.bgpmon.net/what-caused-todays-internet-hiccup/)*. bgpmon.net.*

  [*16-bit Autonomus System Report*](http://www.potaroo.net/tools/asn16/), Geoff Huston 2011 (original archived at [https://web.archive.org/web/20110906085724/http://www.potaroo.net/tools/asn16/](https://web.archive.org/web/20110906085724/http:/www.potaroo.net/tools/asn16/))

  *Craig Timberg (2015-05-31).* [*"Quick fix for an early Internet problem lives on a quarter-century later"*](https://www.washingtonpost.com/sf/business/2015/05/31/net-of-insecurity-part-2/)*. The Washington Post. Retrieved 2015-06-01.*

* 1.  [*"GNU Zebra"*](https://www.gnu.org/software/zebra/)*.*

**Further reading**

* [Chapter "Border Gateway Protocol (BGP)"](http://docwiki.cisco.com/wiki/Border_Gateway_Protocol) in the [Cisco](https://en.wikipedia.org/wiki/Cisco_Systems) "IOS Technology Handbook"

**External links**

* [BGP Routing Resources](https://www.bgp4.as) (includes a dedicated section on [BGP & ISP Core Security](http://www.bgp4.as/security))
* [BGP table statistics](http://bgp.potaroo.net/)

## What is BGP?

Border Gateway Protocol (BGP) is the postal service of the Internet. When someone drops a letter into a mailbox, the postal service processes that piece of mail and chooses a fast, efficient route to deliver that letter to its recipient. Similarly, when someone submits data across the Internet, BGP is responsible for looking at all of the available paths that data could travel and picking the best route, which usually means hopping between autonomous systems.

BGP is the protocol that makes the Internet work. It does this by enabling data routing on the Internet. When a user in Singapore loads a website with [origin servers](https://www.cloudflare.com/learning/cdn/glossary/origin-server/) in Argentina, BGP is the protocol that enables that communication to happen quickly and efficiently.

## What is an autonomous system?

The Internet is a network of networks; it’s broken up into hundreds of thousands of smaller networks known as autonomous systems (AS). Each of these networks is essentially a large pool of routers run by a single organization.

If we continue to think of BGP as the postal service of the Internet, AS’s are like individual post office branches. A town may have hundreds of mailboxes, but the mail in those boxes must go through the local postal branch before being routed to another destination. The internal routers within an AS are like mailboxes, they forward their outbound transmissions to the AS, which then uses BGP routing to get these transmissions to their destinations.

The diagram above illustrates a simplified version of BGP. In this version there are only 6 autonomous systems on the Internet. If AS1 needs to route a packet to AS3, it has two different options:

Hopping to AS2 and then to AS3:

AS2 → AS3

Or hopping to AS6, then to AS5, AS4, and finally to AS3:

AS6 → AS5 → AS4 → AS3

In this simplified model, the decision seems simple. The AS2 route requires fewer hops than the AS6 route, and therefore it is the quickest, most efficient route. Now imagine that there are hundreds of thousands of AS’s and that hop count is only one part of a complex route selection algorithm. That’s the reality of BGP routing on the Internet.

The structure of the Internet is constantly changing, with new systems popping up and existing systems becoming unavailable. Because of this, every AS must be kept up to date with information regarding new routes as well as obsolete routes. This is done through peering sessions where each AS connects to neighboring AS’s with a [TCP/IP](https://www.cloudflare.com/learning/ddos/glossary/tcp-ip/) connection for the purpose of sharing routing information. Using this information, each AS is equipped to properly route outbound data transmissions coming from within.

Here’s where part of our analogy falls apart: Unlike post office branches, autonomous systems are not all part of the same organization. As such, they have no reason to be friendly to each other and are often times business competitors! For this reason, BGP routes sometimes take business considerations into account. Autonomous Systems often charge each other to carry traffic across their networks, and the price of access can be factored into which route is ultimately selected.

## Who operates BGP autonomous systems?

Autonomous systems typically belong to ISPs or other large high-tech organizations, such as tech companies, universities, government agencies, and scientific institutions. Each autonomous system wishing to exchange routing information must have a registered autonomous system number (ASN). Internet Assigned Numbers Authority (IANA) assigns ASNs to Regional Internet Registries (RIRs), which then assigns them to ISPs and networks. ASNs are 16 bit numbers between 1 and 65534 and 32 bit numbers between 131072 and 4294967294. As of 2018, there are approximately 64,000 ASNs in-use worldwide. These ASNs are only required for external BGP.

## What’s the difference between external BGP and internal BGP?

Routes are exchanged and traffic is transmitted over the Internet using external BGP or eBGP. Autonomous systems can also use an internal version of BGP to route through their internal networks, which is known as internal BGP, or iBGP for short. It should be noted that using internal BGP is NOT a requirement for using external BGP. Autonomous systems can choose from a number of internal protocols to connect the routers on their internal network.

External BGP is like international shipping; there are certain standards and guidelines that need to be followed when shipping a piece of mail internationally. Once that piece of mail reaches its destination country, it has to go through the destination country’s local mail service to reach its final destination. Each country has its own internal mail service that doesn’t necessarily follow the same guidelines as those of other countries. Similarly, each autonomous system can have its own internal routing protocol for routing data within its own network.

## How BGP can break the Internet

In 2004 a Turkish Internet service provider (ISP) called TTNet accidentally advertised bad BGP routes to its neighbors. These routes claimed that TTNet itself was the best destination for all traffic on the Internet. As these routes spread further and further to more autonomous systems, a massive disruption occurred, creating a 1-day crisis where many people across the world were not able to access some or all of the Internet.

Similarly, in 2008 a Pakistani ISP attempted to use a BGP route to block Pakistani users from visiting YouTube. The ISP then accidentally advertised these routes with its neighboring AS’s and the route quickly spread across the Internet’s BGP network. This route sent users trying to access YouTube to a dead end, which resulted in YouTube being inaccessible for several hours.

These are examples of a practice called BGP hijacking, and it isn’t always accidental. In April of 2018, attackers deliberately created bad BGP routes to redirect traffic that was meant for Amazon’s DNS service. The attackers were able to steal over $100,000 worth of [cryptocurrency](https://www.cloudflare.com/learning/ddos/cryptocurrency-ddos-attacks/) by redirecting this traffic to themselves.

Incidents like these can happen because the route-sharing function of BGP relies on trust, and autonomous systems implicitly trust the routes that are shared with them. While there have been a number of ambitious proposals intended to make BGP more secure, these are hard to implement because they would require every autonomous system to simultaneously update their behavior. Since this would require the coordination of hundreds of thousands of organizations and potentially result in a temporary takedown of the entire Internet, it seems unlikely that any of these major proposals will be put into place anytime soon.

The Border Gateway Protocol (BGP) is the routing protocol of the Internet, used to route traffic across the Internet. For that reason, it's a pretty important protocol, and it can also be the hardest one to understand.

From our [overview of Internet routing](http://www.enterprisenetworkingplanet.com/netsp/article.php/3613781/Networking-101--Understanding-Internet-Routing-and-Peering.htm), you should realize that routing in the Internet is comprised of two parts: the internal fine-grained portions managed by an IGP such as OSPF, and the interconnections of those autonomous systems (AS) via BGP.

**Who needs to understand BGP?**

BGP is relevant to network administrators of large organizations which connect to two or more ISPs, as well as to Internet Service Providers (ISPs) who connect to other network providers. If you are the administrator of a small corporate network, or an end user, then you probably don't need to know about BGP.

**BGP basics**

* The current version of BGP is BGP version 4, based on RFC4271.
* BGP is the path-vector protocol that provides routing information for autonomous systems on the Internet via its AS-Path attribute.
* BGP is a Layer 4 protocol that sits on top of TCP. It is much simpler than OSPF, because it doesn’t have to worry about the things TCP will handle.
* Peers that have been manually configured to exchange routing information will form a TCP connection and begin speaking BGP. There is no discovery in BGP.
* Medium-sized businesses usually get into BGP for the purpose of true multi-homing for their entire network.
* An important aspect of BGP is that the AS-Path itself is an anti-loop mechanism. Routers will not import any routes that contain themselves in the AS-Path.

**Why do you need to understand BGP?**

When BGP is configured incorrectly, it can cause massive availability and security problems, as Google discovered in 2008 when its YouTube service became unreachable to large portions of the Internet. What happened was that, in an effort to ban YouTube in its home country, Pakistan Telecom used BGP to route YouTube's address block into a black hole. But, in what is believed to have been an accident, this routing information somehow got transmitted to Pakistan Telecom's Hong Kong ISP and from there got propagated to the rest of the world. The end result was that most of YouTube's traffic ended up in a black hole in Pakistan.

More sinisterly, 2003 saw a number of BGP hijack attacks, where modified BGP route information allowed unknown attackers to redirect large blocks of traffic so that it travelled via routers in Belarus or Iceland before it was transmitted on to its intended destination.

Clearly, BGP is significant. Here we'll provide a short overview of how BGP works, along with the problems it solves and causes.

**Autonomous systems**

First a little terminology. In the world of BGP, each routing domain is known as an autonomous system, or AS. What BGP does is help choose a path through the Internet, usually by selecting a route that traverses the least number of autonomous systems: the shortest AS path.

You might need BGP, for example, if your corporate network is connected to two large ISPs. To use BGP you would need an AS number, which you can get from the American Registry of Internet Numbers (ARIN).

Once BGP is enabled, your router will pull a list of Internet routes from your BGP neighbors, who in this case will be your two ISPS. It will then scrutinize them to find the routes with the shortest AS paths. These will be put into the router's routing table. (If you only connect to a single ISP then you don't need BGP. That's because there's only one path to the Internet, so there's no need for a routing protocol to select the best path.)

Generally, but not always, routers will choose the shortest path to an AS. BGP only knows about these paths based on updates it receives.

**Route updates**

Unlike Routing Information Protocol (RIP), a distance-vector routing protocol which employs the hop count as a routing metric, BGP does not broadcast its entire routing table. At boot, your peer will hand over its entire table. After that, everything relies on updates received.

Route updates are stored in a Routing Information Base (RIB). A routing table will only store one route per destination, but the RIB usually contains multiple paths to a destination. It is up to the router to decide which routes will make it into the routing table, and therefore which paths will actually be used. In the event that a route is withdrawn, another route to the same place can be taken from the RIB.

The RIB is only used to keep track of routes that could possibly be used. If a route withdrawal is received and it only existed in the RIB, it is silently deleted from the RIB. No update is sent to peers. RIB entries never time out. They continue to exist until it is assumed that the route is no longer valid.

**BGP path attributes**

In many cases, there will be multiple routes to the same destination. BGP therefore uses path attributes to decide how to route traffic to specific networks.

The easiest of these to understand is Shortest AS\_Path. What this means is the path which traverses the least number of AS "wins."

Another important attribute is Multi\_Exit\_Disc (Multi-exit discriminator, or MED). This makes it possible to tell a remote AS that if there are multiple exit points on to your network, a specific exit point is preferred.

The Origin attribute specifies the origin of a routing update. If BGP has multiple routes, then origin is one of the factors in determining the preferred route.

**BGP issues**

To get a true sense of how BGP works, it's important to spend some time talking about the issues that plague the Internet.

First, we have a very big problem with routing table growth. If someone decides to deaggregate a network that used to be a single /16 network, they could potentially start advertising hundreds of new routes. Every router on the Internet will get every new route when this happens. People are constantly pressured to aggregate, or combine multiple routes into a single advertisement. Aggregation isn't always possible, especially if you want to break up a /19 into two geographically separate /20s. Routing tables are approaching 200,000 routes now, and for a time they were appearing to grow exponentially.

Second, there is always a concern that someone will "advertise the Internet." If some large ISP's customer suddenly decides to advertise everything, and the ISP accepts the routes, all of the Internet's traffic will be sent to the small customer's AS. There's a simple solution to this. It's called route filtering. It's quite simple to set up filters so that your routers won't accept routes from customers that you aren't expecting, but many large ISPs will still accept the equivalent of "default" from peers that have no likelihood of being able to provide transit.

Finally, we come to flapping. BGP has a mechanism to "hold down" routes that appear to be flaky. Routes that flap, or come and go, usually aren't reliable enough to send traffic to. If routes flap frequently, the load on all Internet routes will increase due to the processing of updates every time someone disappears and reappears. Dampening will prevent BGP peers from listening to all routing updates from flapping peers. The amount of time one is in hold-down increases exponentially with every flap. It's annoying when you have a faulty link, since it can be more than an hour before you can get to many Internet sites, but it is very necessary.

This quick discussion of BGP should be enough to get you thinking the right way about the protocol but is by no means comprehensive. Spend some time reading the RFCs if you're tasked with operating a BGP router. Your peers will appreciate it.