BGP

In This Chapter

This chapter provides information about the Border Gateway Protocol (BGP) and its implementation in SR-OS.

Topics in this chapter include:

- BGP Overview on page 620
- BGP Sessions on page 621
- BGP Design Concepts on page 629
- BGP Messages on page 634
- BGP Path Attributes on page 639
- BGP Routing Information Base (RIB) on page 658
- BGP Applications on page 680
- BGP Configuration Process Overview on page 695
- Configuration Notes on page 696

BGP Overview

Border Gateway Protocol (BGP) is an inter-Autonomous System routing protocol. An Autonomous System (AS) is a set of routers managed and controlled by a common technical administration. BGP-speaking routers establish BGP sessions with other BGP-speaking routers and use these sessions to exchange BGP routes. A BGP route provides information about a network path that can reach an IP prefix or other type of destination. The path information in a BGP route includes the list of ASes that must be traversed to reach the route source; this allows inter-AS routing loops to be detected and avoided. Other path attributes that may be associated with a BGP route include the Local Preference, Origin, Next-Hop, Multi-Exit Discriminator (MED) and Communities. These path attributes can be used to implement complex routing policies.

The primary use of BGP was originally Internet IPv4 routing but multi-protocol extensions to BGP have greatly expanded its applicability. Now BGP is used for many purposes, including:

- Internet IPv6 routing
- Inter-domain multicast support
- L3 VPN signaling (unicast and multicast)
- L2 VPN signaling (BGP auto-discovery for LDP-VPLS, BGP-VPLS, BGP-VPWS, multi-segment pseudowire routing, EVPN)
- Setup of inter-AS MPLS LSPs
- Distribution of flow specification rules (filters/ACLs)

The next sections provide information about BGP sessions, BGP network design, BGP messages and BGP path attributes.

BGP Sessions

A BGP session is a TCP connection formed between two BGP routers over which BGP messages are exchanged. There are three types of BGP sessions: internal BGP (IBGP), external BGP (EBGP), and confederation external BGP (confed-EBGP).

An IBGP session is formed when the two BGP routers belong to the same Autonomous System. Routes received from an IBGP peer are not advertised to other IBGP peers unless the router is a route reflector. The two routers that form an IBGP session are usually not directly connected. Figure 25 shows an example of two Autonomous Systems that use BGP to exchange routes. In this example the router ALA-A forms IBGP sessions with ALA-B and ALA-C.

An EBGP session is formed when the two BGP routers belong to different Autonomous Systems. Routes received from an EBGP peer can be advertised to any other peer. The two routers that form an EBGP session are often directly connected but multi-hop EBGP sessions are also possible. When a route is advertised to an EBGP peer the Autonomous System number(s) of the advertising router are added to the AS Path attribute. In the example of Figure 25 the router ALA-A forms an EBGP session with ALA-D.

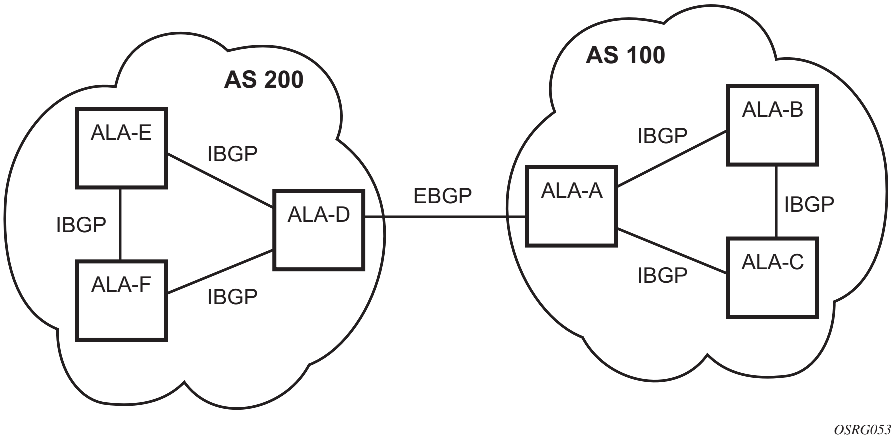

Figure 25: BGP Sessions

A confederation EBGP session is formed when the two BGP routers belong to different member AS of the same confederation. More details about BGP confederations are provided in the section titled BGP Confederations on page 632.

In SR-OS a BGP session is configured using the **neighbor** command. This command accepts either an IPv4 or IPv6 address, which allows the session transport to be IPv4 or IPv6. By default 7x50 is the **active** side of TCP connections to remote neighbors, meaning that as soon as a session leaves the *Idle* state 7x50 attempts to setup an outgoing TCP connection to the remote neighbor in addition to listening on TCP port 179 for an incoming connection from the peer. If required, a BGP session can be configured for **passive** mode so that the 7x50 router only listens for an incoming connection and does not attempt to setup the outgoing connection. The source IP address used to setup the TCP connection to the peer can be configured explicitly using the **local-address** command. If a **local-address** is not configured then the source IP address is determined as follows:

- If the neighbor's IP address belongs to a local subnet the source IP address is this router's IP address on that subnet
- If the neighbor's IP address does not belong to a local subnet the source IP address is this router's system IP address

BGP Session States

A BGP session is in one of the following states at any given moment in time:

- ***Idle***. This is the state of a BGP session when it is administratively disabled (with a **shutdown** command). In this state no incoming TCP connection is accepted from the peer. When the session is administratively enabled it transitions out of the *Idle* state immediately. When the session is restarted automatically it may not leave the *Idle* state immediately if **damp-peer-oscillations** is cnfigured. **damp-peer-oscillations** holds a session in the *Idle* state for exponentially increasing amounts of time if the session is unstable and resets frequently.
- ***Connect***. This is the state of a BGP session when the router, acting in active mode, is attempting to establish an outbound TCP connection with the remote peer.
- ***Active***. This is the state of a BGP session when the router is listening for an inbound TCP connection attempt from the remote peer.
- ***OpenSent***. This is the state of a BGP session when the router has sent an OPEN message to its peer in reaction to successful setup of the TCP connection and is waiting for an OPEN message from the peer.
- ***OpenConfirm***. This is the state of a BGP session after the router has received an acceptable OPEN message from the peer and sent a KEEPALIVE message in response and is waiting for a KEEPALIVE message from the peer. TCP connection collision procedures may be performed at this stage. Refer to RFC 4271 for more details.
- ***Established***. This is the state of a BGP session after the router has received a KEEPALIVE message from the peer. In this state BGP can advertise and withdraw routes by sending UPDATE messages to its peer.

Detecting BGP Session Failures

If a router suspects that its peer at the other end of an established session has experienced a complete failure of both its control and data planes the router should divert traffic away from the failed peer as quickly as possible in order to minimize traffic loss. There are various mechanisms that the router can use to detect such failures, including:

- BGP session hold timer expiry. See the section titled Keepalive Message on page 636 for more details about this mechanism.
- Peer tracking
- BFD
- Fast external failover

When any one or these mechanisms is triggered the session immediately returns to the ***Idle*** state and a new session is attempted. Peer tracking, BFD and fast external failover are described in more detail in the following sections.

Peer Tracking

When peer tracking is enabled on a session the neighbor IP address is tracked in the routing table; if a failure occurs and there is no longer any IP route matching the neighbor address or else if the longest prefix match (LPM) route is rejected by the configurable **peer-tracking-policy** then after a 1 second delay the session is taken down. By default peer-tracking is disabled on all sessions. The default peer-tracking policy allows any type of route to match the neighbor IP address except aggregate routes and LDP shortcut routes.

Peer tracking was introduced when BFD was not yet supported for peer failure detection. Now that BFD is available peer-tracking has less value and is used less often.

NOTE: Peer tracking should be used with caution. Peer tracking can tear a session down even if the loss of connectivity turns out to be short-lived — for example while the IGP protocol is re-converging. Next-hop tracking, which is always enabled, handles such temporary connectivity issues much more effectively.

Bidirectional Forwarding Detection (BFD)

SR-OS also supports the option to setup an async-mode BFD session to a BGP neighbor so that failure of the BFD session can trigger immediate teardown of the BGP session. When BFD is enabled on a BGP session a 1-hop or multi-hop BFD session is setup to the neighbor IP address and the BFD parameters come from the BFD configuration of the interface associated with the **local-address**; for multi-hop sessions this is typically the system interface. With a 10 ms transmit-interval and a multiplier of 3 BFD can detect a peer failure in a period of time as short of 30 ms.

Fast External Failover

Fast external failover applies only to single-hop EBGP sessions. When fast external failover is enabled on a single-hop EBGP session and the interface associated with the session goes down the BGP session is immediately taken down as well, even if other mechanisms such as the hold-timer have not yet indicated a failure.

High Availability BGP Sessions

A BGP session reset can be very disruptive – each router participating in the failed session must delete the routes it received from its peer, recalculate new best paths, update forwarding tables (depending on the types of routes), and send route withdrawals and advertisements to other peers. It makes sense then that session resets should be avoided as much as possible and when a session reset cannot be avoided the disruption to the network should be minimized. To support these objectives the BGP implementation in SR-OS supports two key features:

- BGP high availability (HA)
- BGP graceful restart (GR)

BGP HA refers to the capability of a 7x50 router with redundant CPMs to keep established BGP sessions up whenever a planned or unplanned CPM switchover occurs. A planned CPM switchover can occur during In-Service Software Upgrade (ISSU). An unplanned CPM switchover can occur if there is an unexpected failure of the primary CPM.

BGP HA is always enabled on 7x50 routers with redundant CPMs; it cannot be disabled. BGP HA keeps the standby CPM in-sync with the primary CPM, with respect to BGP and associated TCP state, so that the standby CPM is ready to take over for the primary CPM at any time. Note that the primary CPM is responsible for building and sending the BGP messages to peers but the standby CPM reliably receives a copy of all outgoing UPDATE messages so that it has a synchronized view of the RIB-OUT.

BGP Graceful Restart

Some BGP routers do not have redundant control plane processor modules or else do not support BGP HA with the same quality or coverage as 7x50 routers. When dealing with such routers or certain error conditions BGP graceful restart is a good option for minimizing the network disruption caused by a control plane reset. BGP graceful restart assumes that the router restarting its BGP sessions has the ability/architecture to continue packet forwarding throughout the control plane reset. If this is the case then the peers of the restarting router act as helpers and “hide” the control plane reset from the rest of the network so that forwarding can continue uninterrupted. Forwarding based on stale routes and hiding the “staleness” from other routers is considered acceptable because the duration of the control plane outage is expected to be relatively short (a few minutes). In order for BGP graceful restart to be used on a session both routers must advertise the BGP graceful restart capability during the OPEN message exchange; see the section titled BGP Advertisement on page 635 for more details.

On 7x50 routers BGP graceful restart is enabled on one or more BGP sessions by configuring the **graceful-restart** command in the global, group or neighbor context. The command causes the GR capability to be advertised and enables helper mode support for IPv4 (AFI1, SAFI1), IPv6 (AFI 2, SAFI1), VPN-IPv4 and VPN-IPv6 routes. Note that the GR capability advertised by a 7x50 router does not list the supported AFI/SAFI unless **enable-notification** is configured.

On a 7x50 router helper mode is activated when one of the following events affects an ***Established*** session:

- TCP socket error
- New inbound TCP connection from the peer
- Hold timer expiry
- Peer unreachable
- BFD down
- Sent NOTIFICATION message (only if **enable-notification** is configured under **graceful-restart,** and the peer set the 'N' bit in its GR capability, and the NOTIFICATION is not a *Cease* with subcode *Hard Reset*)
- Received NOTIFICATION message (only if **enable-notification** is configured under **graceful-restart,** and the peer set the 'N' bit in its GR capability, and the NOTIFICATION is not a *Cease* with subcode *Hard Reset*)

As soon as the failure is detected the helping 7x50 router marks the received IPv4, IPv6, VPN-IPv4 and VPN-IPv6 routes from the peer as 'stale' and starts a restart timer. (As noted above the 'stale' state is not factored into the BGP decision process and not made visible to other routers in the network.) The restart timer derives its initial value from the Restart Time carried in the peer's last GR capability. (The default Restart Time advertised by 7x50 routers is 300 seconds but this can be changed using the **restart-time** command.) When the restart timer expires helping stops if the session has not yet re-established. If the session is re-established before the restart timer expires and the new GR capability from the restarting router indicates that forwarding state was preserved then helping continues and the peers exchange routes per the normal procedure. When each router has advertised all its routes for a particular address family it sends an **End-of-RIB** marker (EOR) for the address family. The EOR is a minimal UPDATE message with no reachable or unreachable NLRI for the AFI/SAFI. When the helping router receives an EOR it deletes all remaining stale routes of the AFI/SAFI that were not refreshed in the most recent set of UPDATE messages; there is an upper limit on the amount of time that routes can remain stale (before being deleted if they were not refreshed) and this is configurable using the **stale-routes-time**.

NOTE: 7x50 routers always abort the GR helper process, regardless of the failure trigger, if there is a second reset before GR has successfully completed.

BGP Session Security

TCP MD5 Authentication

The operation of a network can be compromised if an unauthorized system is able to form or hijack a BGP session and inject control packets by falsely representing itself as a valid neighbor. This risk can be mitigated by enabling TCP MD5 authentication on one or more of the sessions. When TCP MD5 authentication is enabled on a session every TCP segment exchanged with the peer includes a TCP option (19) containing a 16-byte MD5 digest of the segment (more specifically the TCP/IP pseudo-header, TCP header and TCP data). The MD5 digest is generated and validated using an authentication key that must be known to both sides. If the received digest value is different from the locally computed one then the TCP segment is dropped, thereby protecting the router from spoofed TCP segments.

TTL Security Mechanism

The TTL security mechanism relies on a simple concept to protect BGP infrastructure from spoofed IP packets. It recognizes the fact that the vast majority of EBGP sessions are established between directly-connected routers and therefore the IP TTL values in packets belonging to these sessions should have predictable values. If an incoming packet does not have the expected IP TTL value it is possible that it is coming from an unauthorized and potentially harmful source.

On 7x50 routers TTL security is enabled using the **ttl-security** command. This command requires a minimum TTL value to be specified. When TTL security is enabled on a BGP session the IP TTL values in packets that are supposedly coming from the peer are compared (in hardware) to the configured minimum value and if there is a discrepancy the packet is discarded and a log is generated. TTL security is used most often on single-hop EBGP sessions but it can be used on multi-hop EBGP and IBGP sessions as well.

NOTE: When a 7x50 router sends IP packets to an IBGP peer they are originated with an IP TTL value of 64. When a 7x50 router sends IP packets to an EBGP peer they are originated with an IP TTL value of 1, except if **multihop** is configured, and in that case the TTL value is taken from the **multihop** command.

BGP Groups

In SR-OS every neighbor (and hence BGP session) is configured under a **group**. A group is a CLI construct that saves configuration effort when multiple peers have a similar configuration; in this situation the common configuration commands can be configured once at the group level and need not be repeated for every neighbor. A single BGP instance can support many groups and each group can support many peers. Most SR-OS commands that are available at the **neighbor** level are also available at the **group** level.

BGP Design Concepts

BGP assumes that all routers within an Autonomous System can reach destinations external to the Autonomous System using efficient, loop-free intra-AS forwarding paths. This generally requires that all the routers within the AS have a consistent view of the best path to every external destination. This is especially true when each BGP router in the AS makes its own forwarding decisions based on its own BGP routing table. The basic BGP specification does not store any intra-AS path information in the AS Path attribute so basic BGP has no way to detect routing loops within an AS that arise from inconsistent best path selections.

There are 3 solutions for dealing the issues outlined above.

- Create a full-mesh of IBGP sessions within the AS as shown in Figure 26. This ensures routing consistency but does not scale well because the number of sessions increases exponentially with the number of BGP routers in the AS.
- Use BGP route reflectors in the AS. Route reflection is described in the section titled Route Reflection on page 630. BGP route reflectors allow for routing consistency with only a partial mesh of IBGP sessions within the AS.

Create a confederation of autonomous systems. BGP confederations are described in the section titled BGP Confederations on page 632.

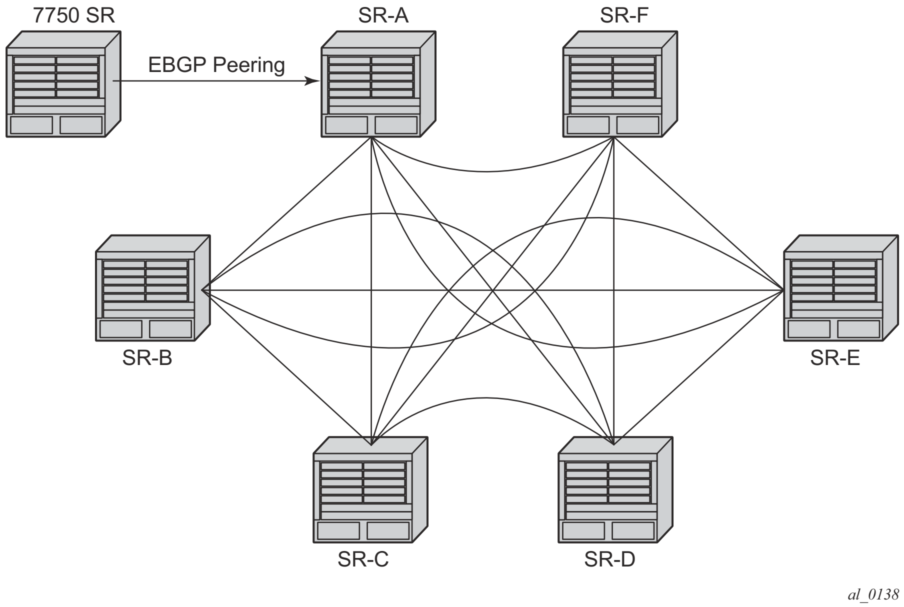

Figure 26: Fully Meshed BGP Configuration

Route Reflection

In a standard BGP configuration a BGP route learned from one IBGP peer is not re-advertised to another IBGP peer. This rule exists because of the assumption of a full IBGP mesh within the AS. As discussed in the previous section a full IBGP mesh imposes certain scaling challenges. BGP route reflection eliminates the need for a full IBGP mesh by allowing routers configured as *route reflectors* to re-advertise routes from one IBGP peer to another IBGP peer.

A route reflector provides route reflection service to IBGP peers called *clients*. Other IBGP peers of the RR are called *non-clients*. An RR and its *client* peers form a *cluster*. A large AS can be sub-divided into multiple clusters, each identified by a unique 32-bit *cluster ID*. Each cluster contains at least one route reflector which is responsible for redistributing routes to its clients. The *clients* within a cluster do not need to maintain a full IBGP mesh between each other; they only require IBGP sessions to the route reflector(s) in their cluster. (If the clients within a cluster are fully meshed consider using the **disable-client-reflect** functionality.) The *non-clients* in an AS must be fully meshed with each other.

Figure 27 depicts the same network as Figure 26 but with route reflectors deployed to eliminate the IBGP mesh between SR-B, SR-C, and SR-D. SR-A, configured as the route reflector, is responsible for reflection routes to its clients SR-B, SR-C, and SR-D. SR-E and SR-F are non-clients of the route reflector. As a result, a full mesh of IBGP sessions must be maintained between SR-A, SR-E and SR-F.

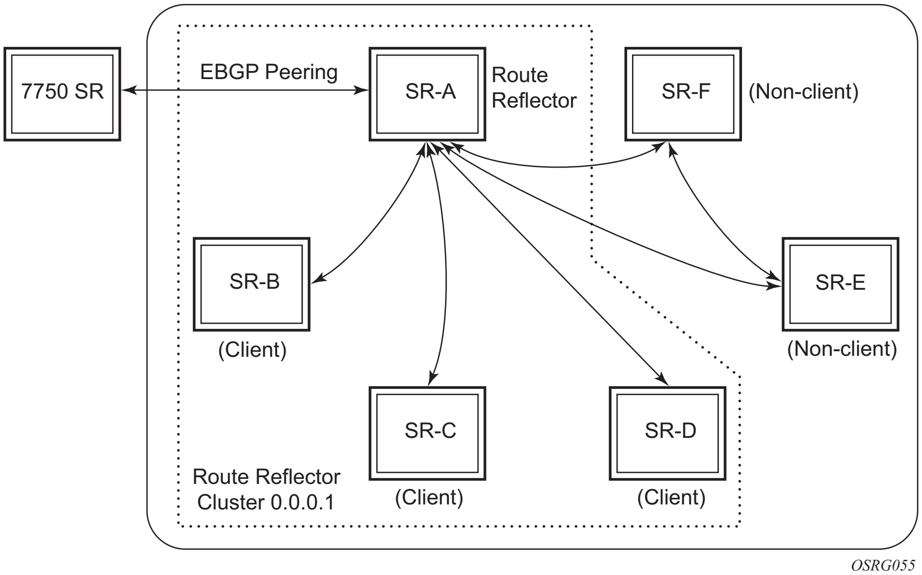

Figure 27: BGP Configuration with Route Reflectors

A 7x50 router becomes a route reflector whenever it has one or more client IBGP sessions. A client IBGP session is created with the **cluster** command, which also indicates the cluster ID of the client. Typical practice is to use the router ID as the cluster ID, but this is not necessary.

Basic route reflection operation on a 7x50 router (without Add-Path configured) can be summarized as follows:

- If the best and valid path for an NLRI is learned from a *client* and **disable-client-reflect** is NOT configured then advertise that route to all *clients*, *non-clients* and EBGP peers (as allowed by policy). If the client that advertised the best and valid path is a neighbor to which the **split-horizon** command (at the **bgp**, **group** or **neighbor** level) applies then the route is not advertised back to the sending client. In the route that is reflected to *clients* and *non-clients*:
  - → The route reflector adds an ORIGINATOR_ID attribute if it did not already exist; the ORIGINATOR_ID indicates the BGP identifier (router ID) of the *client* that originated the route.
  - → The route reflector prepends the cluster ID of the *client* that advertised the route and then the cluster ID of the *client* receiving the route (if applicable) to the CLUSTER_LIST attribute, creating the attribute if it did not previously exist.
- If the best and valid path for an NLRI is learned from a *client* and **disable-client-reflect** is configured then advertise that route to all *clients* in other clusters, *non-clients* and EBGP

peers (as allowed by policy). In the route that is reflected to *clients* in other clusters and *non-clients*:

- → The route reflector adds an ORIGINATOR_ID attribute if it did not already exist; the ORIGINATOR_ID indicates the BGP identifier (router ID) of the *client* that originated the route.
- → The route reflector prepends the cluster ID of the *client* that advertised the route and then the cluster ID of the *client* receiving the route (if applicable) to the CLUSTER_LIST attribute, creating the attribute if it did not previously exist.

- If the best and valid path for an NLRI is learned from a *non-client* then advertise that route to all *clients* and EBGP peers (as allowed by policy). In the route that is reflected to *clients*:
  - → The route reflector adds an ORIGINATOR_ID attribute if it did not already exist; the ORIGINATOR_ID indicates the BGP identifier (router ID) of the *non-client* that originated the route.
  - → The route reflector prepends the cluster ID of the *client* receiving the route to the CLUSTER_LIST attribute, creating the attribute if it did not previously exist.
- If the best and valid path for an NLRI is learned from an EBGP peer then advertise that route to all *clients*, *non-clients* and other EBGP peers (as allowed by policy). The ORIGINATOR_ID and CLIUSTER_LIST attributes are not added to the route.
- If the best and valid path for an NLRI is locally originated (by the RR) — i.e. it was learned through means other than BGP — then advertise that route to all *clients*, *non-clients* and EBGP peers (as allowed by policy). The ORIGINATOR_ID and CLUSTER_LIST attributes are not added to the route.

The ORIGINATOR_ID and CLUSTER_LIST attributes allow BGP to detect the looping of a route within the AS. If any router receives a BGP route with an ORIGINATOR_ID attribute containing its own BGP identifier the route is considered *invalid*. In addition if a route reflector receives a BGP route with a CLUSTER_LIST attribute containing a locally configured cluster ID the route is considered *invalid*. Invalid routes are not installed in the route table and not advertised to other BGP peers.

BGP Confederations

BGP confederations are another alternative for avoiding a full mesh of BGP sessions inside an Autonomous System. A BGP confederation is a group of Autonomous Systems managed by a single technical administration that appear as a single AS to BGP routers outside the confederation; the single externally visible AS is called the confederation ID. Each AS in the group is called a *member AS* and the ASN of each member AS is visible only within the confederation. For this reason member ASNs are often private ASNs.

Within a confederation EBGP-type sessions can be setup between BGP routers in different member AS. These confederation-EBGP sessions avoid the need for a full mesh between routers in different member ASes. Within each member AS the BGP routers must be fully-meshed with IBGP sessions or route reflectors must be used to ensure routing consistency.

In SR-OS a confederation EBGP session is formed when the ASN of the peer is different from the local ASN and the peer ASN appears as a member AS in the **confederation** command. The confederation command specifies the confederation ID and up to 15 member AS that are part of the confederation.

When a route is advertised to a confederation-EBGP peer the advertising router prepends its local ASN, which is its member ASN, to a confederation-specific sub-element in the AS_PATH that is created if it does not already exist. The extensions to the AS_PATH are used for loop detection but they do not influence best path selection (i.e. they do not increase the AS Path length used in the BGP decision process). The MED, NEXT_HOP and LOCAL_PREF attributes in the received route are propagated unchanged by default. Note that ORIGINATOR_ID and CLUSTER_LIST attributes are not included in routes to confed-EBGP peers.

When a route is advertised to an EBGP peer outside the confederation the advertising router removes all member AS elements from the AS_PATH and prepends its confederation ID rather than its local/member ASN.

BGP Messages

BGP protocol operation relies on the exchange of BGP messages between peers. 7x50 and most other routers support the following 5 message types: Open, Update, Notification, Keepalive and Route Refresh. Details about each one are described in the following sections.

The minimum length of a BGP message is 19 bytes and the maximum length is 4096 bytes. BGP messages appear as a stream of bytes to the underlying TCP transport layer so there is no direct association between a BGP message and a TCP segment. One TCP segment can carry parts of one or more BGP messages. The maximum size of a BGP TCP segment sent by a 7x50 router is 1024 bytes (assuming a 40 byte TCP/IP header) if path MTU discovery is not enabled for the BGP session and the interfaces have default **tcp-mss** configurations. When path MTU discovery is enabled (with the **path-mtu-discovery** command) the maximum TCP segment size is discovered from received ICMP messages.

Open Message

After a TCP connection is established between two BGP routers the first message sent by each one is an Open message. If the received Open message is acceptable a Keepalive message confirming the Open is sent back. (See the section titled BGP Session States on page 622 for more details.) An Open message contains the following information:

- Version — The current BGP version number is 4.
- Autonomous System number — The 2-byte AS of the sending router. If the sending router has an ASN greater than 65535 this field has the special value 23456 (AS_TRANS). On a 7x50 router the ASN in the Open message is based on the confederation ID (if the peer is external to the confederation), the global AS (configured using the **autonomous-system** command) or a session-level override of the global AS called the local AS (configured using the **local-as** command). More details about the use of local-AS are described in the section titled Using Local AS for ASN Migration on page 642. More details about 4-byte AS numbers are described in the section titled 4-Octet Autonomous System Numbers on page 643.
- Hold Time — The proposed maximum time BGP will wait between successive messages (Keepalive and/or Update) from its peer before closing the connection. The actual hold time is the minimum of the configured **hold-time** for the session and the hold-time in the peer's Open message. If this minimum is below a configured threshold (**min** hold-time), the connection attempt is rejected. Note that changes to the configured **hold-time** trigger a session reset.
- BGP Identifier — The router ID of the BGP speaker. In Open messages sent by 7x50, the BGP Identifier comes from the **router-id** configured under **bgp**; if that is not configured then the **router-id** configured under **config>router** (or **config>service>vprn**) is used and if that too is not configured then the system interface IPv4 address is used. Note that a

change of the router ID in the **config>router>bgp** context causes all BGP sessions to be reset immediately while other changes resulting in a new BGP identifier only take effect after BGP is shutdown and re-enabled.

- Optional Parameters — A list of optional parameters, each encoded as a TLV. The only optional parameter that has been defined is the optional parameter. The optional parameter supports the process of BGP advertisement, which is described in the next section. When a BGP router receives an Open message with an unsupported optional parameter type it terminates the session. A 7x50 router always sends a optional parameter in its Open message unless **disable-capability-negotiation** is configured.

Changing the Autonomous System Number

If the AS number is changed at the router level (**config**>**router**) the new AS number is not used until the BGP instance is restarted either by administratively disabling and enabling the BGP instance or by rebooting the system with the new configuration.

On the other hand if the AS number is changed in the BGP configuration (**config**>**router**>**bgp**) the effects are as follows:

- A change of the local-AS at the global level causes the BGP instance to restart with the new local AS number.
- A change of the local-AS at the **group** level causes BGP to re-establish sessions with all peers in the group using the new local AS number.
- A change of the local-AS at the **neighbor** level causes BGP to re-establish the session with the new local AS number.

Changing a Confederation Number

Changing the a confederation value on an active BGP instance will not restart the protocol. The change will take affect when the BGP protocol is (re) initialized.

BGP Advertisement

BGP advertisement allows a BGP router to indicate to a peer, using the optional parameter, the features that it supports so that they can coordinate and use only the features that both support. Each capability in the optional parameter is TLV-encoded with a unique type code. SROS supports the following capability codes:

- Multi-protocol BGP (code 1)
- Route refresh (code 2)

- Outbound route filtering (code 3)
- Graceful restart (code 64)
- 4-octet AS number (code 65)
- Add-path (code 69)

Update Message

Update messages are used to advertise and withdraw routes. An Update message provides the following information:

- Withdrawn routes length — The length of the withdrawn routes field that is described next (may be 0).
- Withdrawn routes — IPv4 prefixes that are no longer considered reachable by the advertising router.
- Total path attribute length — The length of the path attributes field that is discussed next (may be 0).
- Path attributes — The path attributes presented in variable length TLV format. The path attributes apply to all the NLRI in the UPDATE message.
- Network layer reachability information (NLRI) — IPv4 prefixes that are considered reachable by the advertising router.

For fast routing convergence a 7x50 router packs as many NLRI into a single Update message as possible. This requires identifying all the routes that share the same path attribute values.

Keepalive Message

After a session is established each router sends periodic Keepalive messages to its peer to test that the peer is still alive and reachable. If no Keepalive or Update message is received from the peer for the negotiated hold-time duration the session is terminated. The period between one Keepalive message and the next is 1/3 of the negotiated hold-time duration or the value configured with the **keepalive** command, whichever is less. If the active hold-time or keepalive interval is zero Keepalive messages are not sent. On 7x50 routers the default hold-time is 90 seconds and the default keepalive interval is 30 seconds.

Many times a peer (reachability) failure is detected through faster mechanisms than hold-timer expiry, as explained in the section titled Detecting BGP Session Failures on page 622.

Notification Message

When a non-recoverable error related to a particular session occurs a Notification message is sent to the peer and the session is terminated (or restarted if graceful restart is enabled for this scenario; see the section titled BGP Graceful Restart on page 625 for more details). The Notification message provides the following information:

- Error code — Indicates the type of error: message header error, Open message error, Update message error, Hold timer expired, Finite State Machine error, or Cease.
- Error subcode — Provides more specific information about the error. The meaning of the subcode is specific to the error code.

UPDATE Message Error Handling

The approach to handling Update message errors has evolved in the past couple of years. The original BGP protocol specification called for all UPDATE message errors to be handled the same way — send a NOTIFICATION to the peer and immediately close the BGP session. This error handling approach was motivated by the goal to ensure protocol “correctness” above all else. But it ignored several important points:

- Not all UPDATE message errors truly have the same severity. If the NLRI cannot be extracted and parsed from an UPDATE message then this is indeed a “critical” error. But other errors such as incorrect attribute flag settings, missing mandatory path attributes, incorrect next-hop length/format, etc. can be considered “non-critical” and handled differently.
- Session resets are extremely costly in terms of their impact on the stability and performance of the network. For many types of UPDATE message errors a session reset does not solve the problem because the root cause remains (e.g. software error, hardware error or misconfiguration). If a session reset is absolutely necessary then the operator should have some control over the timing.
- Some degree of protocol “incorrectness” is tolerable for a short period of time as long as the network operator is fully aware of the issue. In this context “incorrectness” typically means a BGP RIB inconsistency between routers in the same AS. Such inconsistency has become less and less of an issue over time as edge-to-edge tunneling of IP traffic (e.g. BGP shortcuts, IP VPN) has reduced the number of deployments where IP traffic is forwarded hop-by-hop.

In recognition of these points and the general trend towards more flexibility in BGP error handling SR-OS supports a BGP configuration option called **update-fault-tolerance** that allows the operator to decide whether the router should apply new or legacy error handling procedures to UPDATE message errors. If **update-fault-tolerance** is configured then non-critical errors as described above are handled using the “treat-as-withdraw” or “attribute-discard” approaches to error handling; these approaches do not cause a session reset. If **update-fault-tolerance** is not configured then legacy procedures continue to apply and all errors (critical and non-critical) trigger a session a reset.

Route Refresh Message

A BGP router can send a Route Refresh message to its peer only if both have advertised the route refresh capability (code 2). The Route Refresh message is a request for the peer to re-send all or some of its routes associated with a particular pair of AFI/SAFI values. AFI/SAFI values are the same ones used in the MP-BGP capability (see the section titled Multi-Protocol BGP Attributes on page 655).

A 7x50 router only sends Route Refresh messages for AFI/SAFI associated with VPN routes that carry Route Target extended communities - i.e. VPN-IPv4, VPN-IPv6, L2-VPN, MVPN-IPv4 and MVPN-IPv6 routes. By default routes of these types are discarded if, at the time they are received, there is no VPN that imports any of the route targets they carry. If at a later time a VPN is added or reconfigured (in terms of the route targets that it imports) a Route Refresh message is sent to all relevant peers so that previously discarded routes can be relearned. Note that Route Refresh messages are not sent for VPN-IPv4 and VPN-IPv6 routes if **mp-bgp-keep** is configured; in this situation received VPN-IP routes are kept in the RIB-IN regardless of whether or not they match a VRF import policy.

BGP Path Attributes

Path attributes are fundamental to BGP. A BGP route for a particular NLRI is distinguished from other BGP routes for the same NLRI by its set of path attributes. Each path attribute describes some property of the path and is encoded as a TLV in the Path Attributes field of the Update message. The type field of the TLV identifies the path attribute and the value field carries data specific to the attribute type. There are 4 different categories of path attributes:

- **Well-known mandatory.** These attributes must be recognized by all BGP routers and must be present in every Update message that advertises reachable NLRI towards a certain type of neighbor (EBGP or IBGP).
- **Well-known discretionary**. These attributes must be recognized by all BGP routers but are not required in every Update message.
- **Optional transitive**. These attributes are allowed to be unrecognized by some BGP routers. If a BGP router does not recognize one of these attributes it accepts it, passes it on to other BGP peers, and sets the Partial bit to 1 in the attribute flags byte.
- **Optional non-transitive**. These attributes are allowed to be unrecognized by some BGP routers. If a BGP router does not recognize one of these attributes it is quietly ignored and not passed on to other BGP peers.

SR-OS supports the following path attributes, which are described in detail in upcoming sections:

- ORIGIN (well-known mandatory)
- AS_PATH (well-known mandatory)
- NEXT_HOP (well-known, required only in Update messages with IPv4 prefixes in the NLRI field)
- MED (optional non-transitive)
- LOCAL_PREF (well-known, required only in Update messages sent to IBGP peers)
- ATOMIC_AGGR (well-known discretionary)
- AGGREGATOR (optional transitive)
- COMMUNITY (optional transitive)
- ORIGINATOR_ID (optional non-transitive)
- CLUSTER_LIST (optional non-transitive)
- MP_REACH_NLRI (optional non-transitive)
- MP_UNREACH_NLRI (optional non-transitive)
- EXT_COMMUNITY (optional transitive)
- AS4_PATH (optional transitive)
- AS4_AGGREGATOR (optional transitive)

- CONNECTOR (optional transitive)
- PMSI_TUNNEL (optional transitive)
- AIGP (optional non-transitive)

Origin

The ORIGIN path attribute indicates the origin of the path information. There are 3 supported values:

- IGP (0)
- EGP (1)
- Incomplete (2)

When a 7x50 router originates a VPN-IP prefix (from a non-BGP route) it sets the value of the Origin attribute to IGP. When a 7x50 originates an BGP route for an IP prefix by exporting a non-BGP route from the routing table it sets the value of the Origin attribute to Incomplete. Route policies (BGP import and export) can be used to change the Origin value.

AS Path

The AS_PATH attribute provides the list of Autonomous Systems through which the routing information has passed. The AS_PATH attribute is composed of segments. There can be up to 4 different types of segments in an AS_PATH attribute: AS_SET, AS_SEQUENCE, AS_CONFED_SET and AS_CONFED_SEQUENCE. The AS_SET and AS_CONFED_SET segment types result from route aggregation. AS_CONFED_SEQUENCE contains an ordered list of member AS through which the route has passed inside a confederation. AS_SEQUENCE contains an ordered list of AS (including confederation IDs) through which the route has passed on its way to the local AS/confederation.

The AS numbers in the AS_PATH attribute are all 2-byte values or all 4-byte values (if the 4-octet ASN capability was announced by both peers).

A BGP router always prepends its AS number to the AS_PATH attribute when advertising a route to an EBGP peer. The specific details for a 7x50 router are described below.

- When a route is advertised to an EBGP peer and the advertising router is not part of a confederation:
  - → The global AS (configured using the **autonomous-system** command) is prepended to the AS_PATH if **local-as** is not configured
  - → The local AS followed by the global AS are prepended to the AS_PATH if **local-as** is configured.
  - → Only the local AS is prepended to the AS_PATH if **local-as no-prepend-global-as** is configured
  - → Private AS numbers (64512 - 65534 inclusive) are removed from the AS_PATH if **remove-private** is configured.
- When a route is advertised to an EBGP peer outside a confederation:
  - → The confederation ID is prepended to the AS_PATH if **local-as** is not configured
  - → The local AS followed by the confederation ID are prepended to the AS_PATH if **local-as** is configured. (Note that the **no-prepend-global-as option** has no effect in this scenario.)
  - → Member AS numbers are removed from the AS_PATH as described in the section titled BGP Confederations on page 632.
  - → Private AS numbers (64512 - 65534 inclusive) are removed from the AS_PATH if **remove-private** is configured.
- When a route is advertised to a confederation-EBGP peer:
  - → If the route came from an EBGP peer and **local-as** was configured on this session (*without* the **private** option) this local AS number is prepended to the AS_PATH in a regular AS_SEQUENCE segment
  - → The global AS (configured using the **autonomous-system** command) is prepended, as a member AS, to the AS_PATH if **local-as** is not configured
  - → The local AS followed by the global AS are prepended, as member AS, to the AS_PATH if **local-as** is configured
  - → Only the local AS is prepended, as a member AS, to the AS_PATH if **local-as no-prepend-global-as** is configured
  - → Private AS numbers (64512 - 65534 inclusive) are removed from the AS_PATH if **remove-private** is configured (except for the local AS added as a member AS).

- When a route is advertised to an IBGP peer:
  - → No information is added to the AS_PATH if the route is locally originated or if it came from an IBGP peer.
  - → The local AS number is prepended to the AS_PATH if the route came from an EBGP peer and **local-as** is configured *without* the **private** option.
  - → The local AS number is prepended, as a member AS, to the AS_PATH if the route came from a confederation-EBGP peer and **local-as** is configured *without* the **private** option.
  - → Private AS numbers (64512 - 65534 inclusive) are removed from the AS_PATH if **remove-private** is configured.

BGP import policies can be used to prepend an AS number multiple times to the AS_PATH, whether the route is received from an IBGP, EBGP or confederation EBGP peer. The AS path prepend action is also supported in BGP export policies applied to these types of peers, regardless of whether the route is locally originated or not. Note that AS path prepending in export policies occurs before the global and/or local ASes (if applicable) are added to the AS_PATH.

When a BGP router receives a route containing one of its own Autonomous System numbers (local or global or confederation ID) in the AS_PATH the route is normally considered *invalid* for reason of an AS path loop. However SR-OS provides a **loop-detect** command that allows this check to be bypassed. If it known that advertising certain routes to an EBGP peer will result in an AS path loop condition and yet there is no loop (assured by other mechanisms, such as the Site of Origin (SOO) extended community) then **as-override** can be configured on the advertising router instead of disabling loop detection on the receiving router. The **as-override** command replaces all occurrences of the peer AS in the AS_PATH with the advertising router's local AS.

AS Override

The AS Override feature can be used in VPRN scenarios where a customer is running BGP as the PE-CE protocol and some or all of the CE locations are in the same Autonomous System (AS). With normal BGP, two sites in the same AS would not be able to reach each other directly since there is an apparent loop in the AS Path.

When as-**override** is configured on a PE-CE EBGP session the PE rewrites the customer ASN in the AS Path with the VPRN AS number as the route is advertised to the CE.

Using Local AS for ASN Migration

The description in the previous section does fully explain the reasons for using **local-as**. This BGP feature facilitates the process of changing the ASN of all the routers in a network from one number to another. This may be necessary if one network operator merges with or acquires another network operator and the two BGP networks must be consolidated into one Autonomous System.

For example suppose the operator of the ASN 64500 network merges with the operator of the ASN 64501 network and the new merged entity decides to renumber ASN 64501 routers as ASN 64500 routers so that they the entire network can be managed as one Autonomous System. The migration can be carried out using the following sequence of steps:

1. Change the global AS of the route reflectors that used to be part of ASN 64501 to the new value 64500.
2. Change the global AS of the RR clients that used to be part of ASN 64501 to the new value 64500.
3. Configure **local-as 64501 private no-prepend-global-as** on every EBGP session of each RR client migrated in step 2.

This migration procedure has several advantages. First, customers, settlement-free peers and transit providers of the previous ASN 64501 network still perceive that they are peering with ASN 64501 and can delay switching to ASN 64500 until the time is convenient for them. Second, the AS path lengths of the routes exchanged with the EBGP peers are unchanged from before so that best path selections are preserved.

4-Octet Autonomous System Numbers

When BGP was developed it was assumed that 16-bit (2-octet) ASNs would be sufficient for global Internet routing. In theory a 16-bit ASN allows for 65536 unique autonomous systems but some of the values are reserved (0 and 64000-65535). Of the assignable space less than 10% remains available. When a new AS number is needed it is now simpler to obtain a 4-octet AS number. 4-octet AS numbers have been available since 2006. A 32-bit (4-octet) ASN allows for 4,294,967,296 unique values (some of which are again, reserved).

When 4-octet AS numbers became available it was recognized that not all routers would immediately support the ability to parse 4-octet AS numbers in BGP messages so two optional transitive attributes called AS4_PATH and AS4_AGGREGATOR were introduced to allow a gradual migration.

A BGP router that supports 4-octet AS numbers advertises this capability in its OPEN message; the capability information includes the AS number of the sending BGP router, encoded using 4 bytes (recall the ASN field in the OPEN message is limited to 2 bytes). By default OPEN messages sent by 7x50 routers always include the 4-octet ASN capability but this can changed using the **disable-4byte-asn** command.

If a BGP router and its peer have both announced the 4-octet ASN capability then the AS numbers in the AS_PATH and AGGREGATOR attributes are always encoded as 4-byte values in the

UPDATE messages they send to each other. These UPDATE messages should not contain the AS4_PATH and AS4_AGGREGATOR path attributes.

If one of the routers involved in a session announces the 4-octet ASN capability and the other one does not then the AS numbers in the AS_PATH and AGGREGATOR attributes are encoded as 2-byte values in the UPDATE messages they send to each other.

When a 7x50 router advertises a route to a peer that did not announce the 4-octet ASN capability:

- If there are any AS numbers in the AS_PATH attribute that cannot be represented using 2 bytes (because they have a value greater than 65535) they are substituted with the special value 23456 (AS_TRANS) and an AS4_PATH attribute is added to the route if it is not already present. The AS4_PATH attribute has the same encoding as the AS_PATH attribute that would be sent to a 4-octet ASN capable router (i.e. each AS number is encoded using 4 octets) but it does not carry segments of type AS_CONFED_SEQUENCE or AS_CONFED_SET.
- If the AS number in the AGGREGATOR attribute cannot be represented using 2 bytes (because its value is greater than 65535) it is substituted with the special value 23456 and as AS4_AGGREGATOR attribute is added to the route if it is not already present. The AS4_AGGREGATOR is the same as the AGGREGATOR attribute that would be sent to a 4-octet ASN capable router (i.e. the AS number is encoded using 4 octets).

When a 7x50 router receives a route with an AS4_PATH attribute it attempts to reconstruct the full AS path from the AS4_PATH and AS_PATH attributes, regardless of whether **disable-4byte-asn** is configured or not. The reconstructed path is the AS path displayed in BGP show commands. If the length of the received AS4_PATH is N and the length of the received AS_PATH is N+t then the reconstructed AS path contains the t leading elements of the AS_PATH followed by all the elements in the AS4_PATH.

Next-Hop

The NEXT_HOP attribute indicates the IPv4 address of the BGP router that is the next-hop to reach the IPv4 prefixes in the NLRI field. If the Update message is advertising routes other than IPv4 unicast routes the next-hop of these routes is encoded in the MP_REACH_NLRI attribute and the NEXT_HOP attribute is not included in the Update message; see the section titled Multi-Protocol BGP Attributes on page 655 for more details.

In IPv4 and IPv6 routes advertised by a 7x50 router the BGP next-hop address is set as follows:

- When a route is advertised to an EBGP peer the BGP next-hop is always changed to the local-address used with the EBGP peer and this behavior cannot be overridden, even with a BGP export policy. (See the section titled BGP Sessions on page 621 for an explanation of how the local-address is determined.) The one exception to this rule occurs when the third-party-nexthop command is applied:

  - → When a route is received from one EBGP peer and is advertised to another EBGP that is in the same IP subnet and has been configured with the third-party-nexthop command (at the BGP instance, group or neighbor level), the BGP next-hop in the advertised route remains unchanged.

- When a route is advertised to an IBGP or confederation-EBGP peer and the route is not locally originated the advertising router does not modify the next-hop by default, however:
  - → If the **next-hop-self** command is applied to a confederation-EBGP peer this changes the next-hop to the local-address used with that peer.
  - → If the **next-hop-self** command is applied to an IBGP peer this changes the next-hop to the local-address used with that peer, but only if the route came from a confed-EBGP or EBGP peer.
  - → A BGP export policy applied to an IBGP or confederation-EBGP session can change the next-hop to any IPv4 address, regardless of the route source (IBGP, EBGP, confed-EBGP).
- When a route is locally-originated and advertised to an IBGP or confederation-EBGP peer the BGP next-hop is by default copied from the next-hop of the route that was exported into BGP, with certain exceptions (e.g. black-hole next-hop).

In VPN-IPv4 routes advertised by a 7x50 router the BGP next-hop address is set as follows:

- When a route is advertised to an EBGP peer the BGP next-hop is changed to the local-address used with the EBGP peer if **enable-inter-as-vpn** is configured; otherwise there is no change to the next-hop.
- When a route is received from an EBGP peer and advertised to an IBGP or confederation-EBGP peer the BGP next-hop is changed to the local-address used with the IBGP or confederation-EBGP peer if **enable-inter-as-vpn** is configured. If **enable-inter-as-vpn** is not configured the next-hop may be changed with the **next-hop-self** command but this is not recommended because it can result in a change of the next-hop without a change in the VPN label.
- When a route is reflected from one IBGP peer to another IBGP peer the RR does not modify the next-hop by default, however if the **next-hop-self** command is applied to the IBGP peer receiving the route and **enable-rr-vpn-forwarding** is configured then this combination of commands changes the next-hop to the local-address used with the peer.

In Label-IPv4 routes advertised by a 7x50 router the BGP next-hop address is set as follows:

- When a route is advertised to an EBGP peer the BGP next-hop is always changed to the local-address used with the EBGP peer and this behavior cannot be overridden.

- When a route is received from an EBGP peer and advertised to an IBGP or confederation-EBGP peer next-hop-self is applied automatically (i.e. the next-hop is modified to the local-address used with the peer), however:
  - → A BGP export policy applied to the IBGP or confederation-EBGP session can change the next-hop to any IPv4 address
  - → If the **next-hop-unchanged label-ipv4** command is applied to the receiving IBGP or confederation-EBGP peer this overrides the automatic next-hop-self and causes no modification to the BGP next-hop
  - → *At the current time SR-OS does not support next-hop-self for label-IPv4 routes advertised to a confed-EBGP peer.*
- When a route is received from an IBGP peer and reflected to another IBGP peer the next-hop is not modified by default, however:
  - → If the **next-hop-self** command is applied to the receiving IBGP peer this changes the next-hop to the local-address used with that peer.
  - → A BGP export policy applied to the IBGP session can change the next-hop to any IPv4 address.

In 6PE routes advertised by a 7x50 router the BGP next-hop address is set as follows:

- When a 6PE route is locally-originated and advertised to any BGP peer the BGP next-hop is an IPv4-mapped IPv6 address allocated from the ::FFFF/96 range. The bottom 32 bits of the IPv6 address is the IPv4 local-address used with the peer.
  - → *At the current time SR-OS does not support sending and receiving 6PE routes with EBGP peers.*
- When a route is received from an IBGP peer and reflected to another IBGP peer the next-hop is not modified by default, however:
  - → A BGP export policy applied to the IBGP session can apply **next-hop-self** or change the next-hop to any IPv4-mapped IPv6 address. Note that the **next-hop-self** command at the group/neighbor configuration level has no effect in this case.
- When a route is advertised to a confederation-EBGP peer the next-hop is not modified by default, however:
  - → If the **next-hop-self** command is applied to the session this changes the next-hop to the IPv4-mapped IPv6 address corresponding to the IPv4 local-address used with the peer.
  - → A BGP export policy applied to the IBGP session can change the next-hop to any IPv4-mapped IPv6 address.

Next-Hop IPv4 Address Family over IPv6

For IBGP sessions, next-hop information is taken from the system interface. If the system interface does not have an IPv4 address configured, no next-hop will be populated without a routing policy applied to the BGP session, and BGP NLRI messages is not sent for the IPv4 address family. The use of an export policy allows the operator to configure next-hop information explicitly.

For EBGP sessions, the next-hop information must be taken from an export routing policy that explicitly sets the next-hop based on operator configuration. If the export policy is not set, the BGP NLRI messages are not sent for the IPv4 address family due to no `next_hop`.

Next-Hop VPN-IPv4 Address Family over IPv6

For IBGP and EBGP sessions, next-hop information is specified as the system IP address.

Next-Hop VPN-IPv6 Address Family over IPv6

For IBGP sessions, the next-hop information is specified as the system IP address encoded as an IPv4-mapped-IPv6 address.

For EBGP sessions, the next-hop information is specified as the system IP address encoded as an IPv4-mapped-IPv6 address, by the way of an export policy configured by the user.

Next-Hop Resolution

For a BGP router to use a BGP route for forwarding it must know how to reach the BGP next-hop of the route. The process of determining the local interface or tunnel that should be used to reach the BGP next-hop is called next-hop resolution. The BGP next-hop resolution process depends on the type of route (the AFI/SAFI) and various configuration settings. The SR-OS details are explained below:

- Next-hop resolution is always done for IPv4 routes. Note the following:
  - → BGP routes are eligible to resolve a BGP next-hop only if the **use-bgp-routes** command is configured.
  - → If there are multiple eligible routes that match the BGP next-hop the longest prefix match (LPM) route is selected.
  - → If the LPM route is rejected by the user-configured **next-hop-resolution policy** or if there are no eligible matching routes the BGP next-hop is unresolved and all the routes with that next-hop are considered *invalid* and not advertised to peers.
  - → If the LPM route (accepted by the policy) is a BGP route then the BGP next-hop of that route is looked up and this time other BGP routes are not eligible to be resolving routes, whether or not **use-bgp-routes** is configured. In other words 7x50 routers support BGP routes resolving BGP routes with one level of recursion.
  - → BGP shortcuts are discussed further in the section titled Next-hop Resolution Using Tunnels on page 680.
- Next-hop resolution is always done for IPv6 routes. The 7x50 router looks for an eligible IPv6 route that matches the BGP next-hop address in the route table. Note the following:
  - → BGP routes are eligible to resolve a BGP next-hop only if the **use-bgp-routes** command is configured.
  - → If there are multiple eligible routes that match the BGP next-hop the longest prefix match (LPM) route is selected.
  - → If the LPM route is rejected by the user-configured **next-hop-resolution policy** or if there are no eligible matching routes the BGP next-hop is unresolved and all the routes with that next-hop are considered *invalid* and not advertised to peers.
  - → If the LPM route (accepted by the policy) is a BGP route then the BGP next-hop of that route is looked up and this time other BGP routes are not eligible to be resolving routes, whether or not **use-bgp-routes** is configured. In other words 7x50 routers support BGP routes resolving BGP routes with one level of recursion.
- SR-OS attempts to resolve the next-hop of a VPN-IPv4 or VPN-IPv6 route only if it is imported into one or more VPRNs or if it is advertised with a new BGP next-hop. If the

next-hop is part of a local subnet the next-hop is automatically resolved by the direct route. If the next-hop is remote (more than one hop away):

  - → The 7x50 router looks for a tunnel in the tunnel-table with a destination that matches the BGP next-hop address. If the route is imported into VPRNs the tunnel types eligible to resolve the BGP next-hop are controlled by the **auto-bind-tunnel** configurations of the VPRNs. If the route is advertised with a new BGP next-hop the eligible tunnel types are controlled by the **transport-tunnel** command.
  - → If there is no matching tunnel-table entry then the BGP next-hop is unresolved and the VPN-IP route is effectively *invalid* despite displaying as *valid* and *best*. A VPN-IP route that is *invalid* due to an unresolved next-hop can be advertised to any type of peer, but only if the next-hop is not changed.

- SR-OS always attempts to resolve the next-hop of a label-IPv4 route. If the next-hop is part of a local subnet the next-hop is automatically resolved by the direct route. If the next-hop is remote (more than one hop away):
  - → The 7x50 router looks for a tunnel in the tunnel-table with a destination that matches the BGP next-hop address and a type allowed by the **transport-tunnel** command. If there are multiple matches the tunnel with the lowest preference is used (RSVP is preferred over LDP).
  - → If there is no matching and eligible entry in the tunnel table but there is a /32 static route with a black-hole next-hop that matches the BGP next-hop address this static route automatically resolves the BGP next-hop.
  - → If there is no matching and eligible entry in the tunnel table and no /32 static black-hole route then the BGP next-hop is unresolved and the label-IPv4 route is considered *invalid*. However note that a label-IPv4 route that is *invalid* due to an unresolved next-hop can still be reflected to an IBGP peer, whether or not **next-hop-self** is applied to the route.
- SR-OS always attempts to resolve the next-hop of a 6PE route.
  - → The 7x50 router looks for an LDP tunnel in the tunnel-table with a destination that matches the IPv4 address contained in the IPv4-mapped IPv6 BGP next-hop address.
  - → If there is no matching LDP entry in the tunnel table but there is a /128 static route with a black-hole next-hop that matches the IPv4-mapped IPv6 BGP next-hop address this static route automatically resolves the BGP next-hop.
  - → If there is no matching LDP entry in the tunnel table and no /128 static black-hole route then the BGP next-hop is unresolved and the 6PE route is considered *invalid*. However note that a 6PE route that is *invalid* due to an unresolved next-hop can still be reflected to an IBGP peer, whether or not **next-hop-self** is applied to the route.
- SR-OS does not check for next-hop reachability in Flow-spec and RTC routes.

Next-Hop Tracking

In SR OS next-hop resolution is not a one-time event. If the IP route or tunnel that was used to resolve a BGP next-hop is withdrawn due to a failure or configuration change an attempt is made to re-resolve the BGP next-hop using the next-best route or tunnel. If there are no more eligible routes or tunnels to resolve the BGP next-hop then the BGP next-hop becomes unresolved. The continual process of monitoring and reacting to resolving route/tunnel changes is called next-hop tracking. In SR-OS next-hop tracking is completely event driven as opposed to timer driven; this provides the best possible convergence performance.

Next-Hop Indirection

SR OS supports next-hop indirection for most types of BGP routes. Next-hop indirection means BGP next-hops are logically separated from resolved next-hops in the forwarding plane (IOMs). This separation allows routes that share the same BGP next-hop(s) to be grouped so that when there is a change to the way a BGP next-hop is resolved only one forwarding plane update is needed, as opposed to one update for every route in the group. The convergence time after the next-hop resolution change is uniform and not linear with the number of prefixes; in other words the next-hop indirection is a technology that supports *prefix independent convergence* (PIC). SR-OS uses next-hop indirection whenever possible; there is no option to disable the functionality.

Using Multiple Address Families over IPv6 BGP Sessions

To ease transition to IPv6 and the deployment of IPv6 into service provider environments, SR-OS permits the transport of the following address families over an IPv6 transported BGP session (a BGP session where both neighbors are configured and transported over IPv6):

- IPv4
- VPN-IPv4
- IPv6
- VPN-IPv6

As the IPv4, VPN-IPv4 and VPN-IPv6 address families require an IPv4 NEXT_HOP address to be present in the BGP NLRI messaging, the following approaches are taken in SR-OS:

- For iBGP sessions, SR-OS will use the configured System Interface IPv4 address as the NEXT_HOP address; unless specifically overwritten by a routing export policy.
- For eBGP sessions, SR-OS requires the use of a routing export policy to set the NEXT_HOP to an appropriate address, such as the IPv4 address configured on the interface between eBGP neighbors.

MED

The Multi-Exit Discriminator (MED) attribute is an optional attribute that can be added to routes advertised to an EBGP peer to influence the flow of inbound traffic to the AS. The MED attribute carries a 32-bit metric value. A lower metric is better than a higher metric when MED is compared by the BGP decision process. Unless the **always-compare-med** command is configured MED is compared only if the routes come from the same neighbor AS. By default if a route is received without a MED attribute it is evaluated by the BGP decision process as though it had a MED containing the value 0, but this can be changed so that a missing MED attribute is handled the same as a MED with the maximum value. SR-OS always removes the received MED attribute when advertising the route to an EBGP peer.

Deterministic MED

Deterministic MED is an optional enhancement to the BGP decision process that causes BGP to groups paths that are equal up to the MED comparison step based on the neighbor AS. BGP compares the best path from each group to arrive at the overall best path. This change to the BGP decision process makes best path selection completely deterministic in all cases. Without **deterministic-med**, the overall best path selection is sometimes dependent on the order of route arrival because of the rule that MED cannot be compared in routes from different neighbor AS.

Local Preference

The LOCAL_PREF attribute is a well-known attribute that should be included in every route advertised to an IBGP or confederation-EBGP peer. It is used to influence the flow of outbound traffic from the AS. The local preference is a 32-bit value and higher values are more preferred by the BGP decision process. The LOCAL_PREF attribute is not included in routes advertised to EBGP peers. (If the attribute is received from an EBGP peer it is ignored.)

In SR-OS the default local preference is 100 but this can be changed with the **local-preference** command or using route policies. When a LOCAL_PREF attribute needs to be added to a route because it does not have one (e.g. because it was received from an EBGP peer) the value is the configured or default **local-preference** unless overridden by policy.

Route Aggregation Path Attributes

An aggregate route is a configured IP route that is activated and installed in the routing table when it has at least one *contributing* route. A route R contributes to an aggregate route S1 if:

- The prefix length of R is greater than the prefix length of S1
- The prefix bits of R match the prefix bits of S1 up to the prefix length of S1
- There is no other aggregate route S2 with a longer prefix length than S1 that meets the previous two conditions
- R is actively used for forwarding and is not an aggregate route

When an aggregate route is activated by a 7x50 router it is not installed in the forwarding table by default. In general though it is advisable to specify the **black-hole** next-hop option for an aggregate route so that when it is activated it is installed in the forwarding table with a black-hole next-hop; this avoids the possibility of creating a routing loop. SR-OS also supports the option to program an aggregate route into the forwarding table with an **indirect** next-hop; in this case packets matching the aggregate route but not a more-specific contributing route are forwarded towards the indirect next-hop rather than discarded.

An active aggregate route can be advertised to a BGP peer (by exporting it into BGP) and this can avoid the need to advertise the more-specific contributing routes to the peer, reducing the number of routes in the peer AS and improving overall scalability. When a 7x50 router advertises an aggregate route to a BGP peer the attributes in the route are set as follows:

- The ATOMIC_AGGREGATE attribute is included in the route if at least one contributing route has the ATOMIC_AGGREGATE attribute or the aggregate route was formed without the **as-set** option and at least one contributing route has a non-empty AS_PATH. The ATOMIC_AGGREGATE attribute indicates that some of the AS numbers present in the AS paths of the contributing routes are missing from the advertised AS_PATH.
- The AGGREGATOR attribute is added to the route. This attribute encodes, by default, the global AS number (or confederation ID) and router ID (BGP identifier) of the router that formed the aggregate, but these values can be changed on a per aggregate route basis using the **aggregator** command option. The AS number in the AGGREGATOR attribute is either 2 bytes or 4 bytes (if the 4-octet ASN capability was announced by both peers). The router ID in the aggregate routes advertised to a particular set of peers can be set to 0.0.0.0 using the **aggregator-id-zero** command.
- The BGP next-hop is set to the local-address used with the peer receiving the route regardless of the BGP next-hops of the contributing routes.
- The ORIGIN attribute is based on the ORIGIN attributes of the contributing routes as described in RFC 4271.

- The information in the AS_PATH attribute depends on the **as-set** option of the aggregate route.
  - → If the **as-set** option is not specified the AS_PATH of the aggregate route starts as an empty AS path and has elements added per the description in the section titled AS Path on page 640.
  - → If the **as-set** option is specified and all the contributing routes have the same AS_PATH then the AS_PATH of the aggregate route starts with that common AS_PATH and has elements added per the description in the section titled AS Path on page 640.
  - → If the **as-set** option is specified and some of the contributing routes have different AS paths the AS_PATH of the aggregate route starts with an AS_SET and/or an AS_CONFED_SET and then adds elements per the description in the section titled AS Path on page 640.
- The COMMUNITY attribute contains all the communities from all the contributing routes.
- No MED attribute is included by default. Note that SR-OS does not require all the contributing routes to have the same MED value.

Community and Extended Community Attributes

A BGP route can be associated with one or more standard communities and one or more extended communities. All the standard communities are carried in a single COMMUNITIES attribute and all the extended communities currently supported by SR-OS are carried in a single EXTENDED_COMMUNITIES attribute.

Each standard community is 4 bytes; the first 2 bytes encode the AS number of the administrative entity that assigned the value in the last 2 bytes. In SR-OS a standard community member is input as *AS*:*value* to reflect this format. There are several well-known standard communities that 7x50 and most other BGP routers recognize:

- NO_EXPORT: When a route carries this community is must not be advertised outside a confederation boundary (i.e. to EBGP peers).
- NO_ADVERTISE: When a route carries this community it must not be advertised to any other BGP peer.
- NO_EXPORT_SUBCONFED: When a route carries this community it must not be advertised outside a member AS boundary (i.e. to confed-EBGP peers or EBGP peers).

Standard communities can be added to or removed from BGP routes using BGP import and export policies. When a BGP route is locally originated by exporting a static or aggregate route into BGP, and the static or aggregate route has an associated community, this community is automatically added to the BGP route. (Note that this may affect the advertisement of the locally originated route if one of the well-known communities is associated with the static or aggregate route.)

If it is necessary to remove all the standard communities from all routes advertised to a BGP peer SR-OS supports the **disable-communities standard** command.

Extended communities provide more flexibility than standard communities. Each extended community is 8 bytes. The first 1 or 2 bytes identifies the type/sub-type and the remaining 6 or 7 bytes is a value. As of release 12.0R1 SR-OS supports the following types of extended communities:

- Transitive 2-octet AS-specific
  - → Route target (type 0x0002)
  - → Route origin (type 0x0003)
  - → OSPF domain ID (type 0x0005)
  - → Source AS (type 0x0009)
  - → L2VPN identifier (type 0x000A)
- Transitive IPv4-address-specific
  - → Route target (type 0x0102)
  - → Route origin (type 0x0103)
  - → OSPF domain ID (type 0x0105)
  - → L2VPN identifier (type 0x010A)
  - → VRF route import (type 0x010B)
- Transitive 4-octet AS-specific
  - → Route target (type 0x0202)
  - → Route origin (type 0x0203)
  - → OSPF domain ID (type 0x0205)
  - → Source AS (type 0x0209)
- Transitive opaque
  - → OSPF route type (type 0x0306)
- Transitive experimental
  - → OSPF domain ID (type 0x8005)
  - → Flow-spec traffic rate (type 0x8006)
  - → Flow-spec traffic action (type 0x8007)
  - → Flow-spec redirect (type 0x8008)
  - → Layer 2 info (type 0x800A)

- EVPN
  → MAC mobility (type 0x0600)

Route target and route origin extended communities can be added to or removed from BGP routes using BGP import and export policies. Other types of extended communities are added automatically to the relevant types of routes.

If it is necessary to remove all the extended communities from all routes advertised to a BGP peer SR-OS supports the **disable-communities extended** command.

Route Reflection Attributes

The ORIGINATOR_ID and CLUSTER_LIST are optional non-transitive attributes that play a role in route reflection, as described in the section titled Route Reflection on page 630.

Multi-Protocol BGP Attributes

As discussed in the BGP chapter overview the uses of BGP have increased well beyond Internet IPv4 routing due to its support for multi-protocol extensions, or more simply MP-BGP. MP-BGP allows BGP peers to exchange routes for NLRI other than IPv4 prefixes - for example IPv6 prefixes, Layer 3 VPN routes, Layer 2 VPN routes, flow-spec rules, etc. A BGP router that supports MP-BGP indicates the types of routes it wants to exchange with a peer by including the corresponding AFI (Address Family Identifier) and SAFI (Subsequent Address Family Identifier) values in the MP-BGP capability of its OPEN message. The two peers forming a session do not need indicate support for the same address families; as long as there is one AFI/SAFI in common the session will establish and routes associated with all the common AFI/SAFI can be exchanged between the peers.

The list of AFI/SAFI advertised in the MP-BGP capability of a 7x50 router is controlled primarily by the **family** command. The AFI/SAFI supported by SR-OS as of Release 12.0R1 and the method of configuring the AFI/SAFI support is summarized in Table 11.

Table 11: Multi-Protocol BGP support in SR-OS

Name	AFI	SAFI	Configuration Commands
IPv4 unicast	1	1	**family ipv4**
IPv4 multicast	1	2	**family mcast-ipv4**
IPv4 labeled unicast	1	4	**family ipv4** **advertise-label ipv4**
NG-MVPN IPv4	1	5	**family mvpn-ipv4**
MDT-SAFI	1	66	**family mdt-safi**
VPN-IPv4	1	128	**family vpn-ipv4**
VPN-IPv4 multicast	1	129	**family mcast-vpn-ipv4**
RT constrain	1	132	**family route-target**
IPv4 flow-spec	1	133	**family flow-ipv4**
IPv6 unicast	2	1	**family ipv6**
IPv6 multicast	2	2	**family mcast-ipv6**
6PE	2	4	**family ipv6** **advertise-label ipv6**
NG-MVPN IPv6	2	5	**family mvpn-ipv6**
VPN-IPv6	2	128	**family vpn-ipv6**
IPv6 flow-spec	2	133	**family flow-ipv6**
Multi-segment PW	25	6	**family ms-pw**
L2 VPN	25	65	**family l2-vpn**
EVPN	25	70	**family evpn**

To advertise reachable routes of a particular AFI/SAFI a BGP router includes a single MP_REACH_NLRI attribute in the UPDATE message. The MP_REACH_NLRI attribute encodes the AFI, the SAFI, the BGP next-hop and all the reachable NLRI. To withdraw routes of a particular AFI/SAFI a BGP router includes a single MP_UNREACH_NLRI attribute in the UPDATE message. The MP_UNREACH_NLRI attribute encodes the AFI, the SAFI and all the withdrawn NLRI. Note that while it is valid to advertise and withdraw IPv4 unicast routes using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes SR-OS always uses the IPv4 fields of the UPDATE message to convey reachable and unreachable IPv4 unicast routes.

4-Octet AS Attributes

The AS4_PATH and AS4_AGGREGATOR path attributes are optional transitive attributes that support the gradual migration of routers that can understand and parse 4-octet ASN numbers. The use of these attributes is discussed in the section titled 4-Octet Autonomous System Numbers on page 643.

AIGP Metric

The accumulated IGP (AIGP) metric is an optional non-transitive attribute that can be attached to selected routes (using route policies) to influence the BGP decision process to prefer BGP paths with a lower end-to-end IGP cost, even when the compared paths span more than one AS or IGP instance. AIGP is different from MED in several important ways:

- AIGP is not intended to be transitive between completely distinct autonomous systems (only across internal AS boundaries)
- AIGP is always compared in paths that have the attribute, regardless of whether or not they come from different neighbor AS
- AIGP is more important than MED in the BGP decision process (see the section titled BGP Decision Process on page 659)
- AIGP is automatically incremented every time there is a BGP next-hop change so that it can track the end-to-end IGP cost. All arithmetic operations on MED attributes must be done manually (for example, using route policies).

In the 7x50 implementation AIGP is supported only in the base router BGP instance and only for the following types of routes: IPv4, label-IPv4, IPv6 and 6PE. The AIGP attribute is only sent to peers configured with the **aigp** command. If the attribute is received from a peer that is not configured for **aigp** or if the attribute is received in a non-supported route type the attribute is discarded and not propagated to other peers (but it is still displayed in BGP show commands).

When a 7x50 router receives a route with an AIGP attribute and it re-advertises the route to an AIGP-enabled peer without any change to the BGP next-hop the AIGP metric value is unchanged by the advertisement (RIB-OUT) process. But if the route is re-advertised with a new BGP next-hop the AIGP metric value is automatically incremented by the route table (or tunnel table) cost to reach the received BGP next-hop and/or by a statically configured value (using route policies).

BGP Routing Information Base (RIB)

The entire set of BGP routes learned and advertised by a BGP router make up its BGP Routing Information Base (RIB). Conceptually the BGP RIB can be divided into 3 parts:

- RIB-IN
- LOC-RIB
- RIB-OUT

The RIB-IN (or Adj-RIBs-In as defined in RFC 4271) holds the BGP routes that were received from peers and that the router decided to keep (store in its memory).

The LOC-RIB contains modified versions of the BGP routes in the RIB-IN. The path attributes of a RIB-IN route can be modified using BGP import policies. All of the LOC-RIB routes for the same NLRI are compared in a procedure called the BGP decision process that results in the selection of the best path for each NLRI. The best paths in the LOC-RIB are the ones that are actually ‘usable’ by the local router for forwarding, filtering, auto-discovery, etc.

The RIB-OUT (or Adj-RIBs-Out as defined in RFC 4271) holds the BGP routes that were advertised to peers. Normally a BGP route is not advertised to a peer (in the RIB-OUT) unless it is ‘used’ locally but there are exceptions. BGP export policies modify the path attributes of a LOC-RIB route to create the path attributes of the RIB-OUT route. A particular LOC-RIB route can be advertised with different path attribute values to different peers so there can exist a 1:N relationship between LOC-RIB and RIB-OUT routes.

The following sections describe many important 7x50 BGP features in the context of the RIB architecture outlined above.

RIB-IN Features

SR-OS implements the following features related to RIB-IN processing:

- UPDATE message fault tolerance. This is described in the section titled UPDATE Message Error Handling on page 637.
- BGP import policies

BGP Import Policies

The **import** command is used to apply one or more policies (up to 15) to a neighbor, group or to the entire BGP context. The **import** command that is most-specific to a peer is the one that is applied. An **import** policy command applied at the **neighbor** level takes precedence over the same command applied at the **group** or global level. An **import** policy command applied at the **group** level takes precedence over the same command specified on the global level. The **import** policies applied at different levels are not cumulative. The policies listed in an **import** command are evaluated in the order in which they are specified.

NOTE: The **import** command can reference a policy before it has been created (as a **policy-statement**).

When an IP route is rejected by an import policy it is still maintained in the RIB-IN so that a policy change can be made later on without requiring the peer to re-send all its RIB-OUT routes. This is sometimes called soft reconfiguration inbound and requires no special configuration in SR-OS.

When a VPN route is rejected by an import policy or not imported by any services it is deleted from the RIB-IN. For VPN-IPv4 and VPN-IPv6 routes this behavior can be changed by configuring the **mp-bgp-keep** command; this option maintains rejected VPN-IP routes in the RIB-IN so that a Route Refresh message does not have to be issued when there is an import policy change.

LOC-RIB Features

SR-OS implements the following features related to LOC-RIB processing.

- BGP decision process
- BGP route installation in the route table
- BGP route installation in the tunnel table
- BGP fast reroute
- QoS Policy Propagation via BGP (QPPB)
- Policy accounting
- Route flap damping (RFD)

These features are discussed in the following sections.

BGP Decision Process

When a BGP router has multiple routes in its LOC-RIB for the same NLRI its BGP decision process is responsible for deciding which one is the best. The best path can be used by the local router (e.g. for its own forwarding) and advertised to other BGP peers.

On 7x50 routers the BGP decision process orders *valid* LOC-RIB routes based on the following sequence of comparisons (if there multiple routes tied at step N then proceed to step N+1):

1. Select the route with the best (numerically lowest) route preference.
2. Select the route with the highest Local Preference (LOCAL_PREF).
3. From all routes with an AIGP metric (if any) select the route with the lowest sum of:
   → a.AIGP metric value stored with the LOC-RIB copy of the route.
   → b.The route table (or tunnel table) cost between the calculating router and the BGP NEXT_HOP in the received route.
4. Select the route with the shortest AS Path. Note that AS numbers in AS_CONFED_SEQ and AS_CONFED_SET elements do not count towards the AS path length. Skip this step if **as-path-ignore** is configured for the address family.
5. Select the route with the lowest Origin (IGP=0, EGP=1, Incomplete=2).
6. Select the route with the lowest MED. Only compare MED in routes from the same neighbor AS by default. A missing MED attribute is considered equivalent to a MED value of 0 by default. Defaults can be changed with the **always-compare-med** command.
7. Prefer routes learned from EBGP peers over routes learned from IBGP and confed-EBGP peers.
8. Select the route with the lowest route or tunnel table cost to the NEXT_HOP. If i**gnore-nh-metric** is configured skip this step.
9. Select the route with lowest next-hop type (resolved in route-table = 0, resolved in tunnel-table = 1). If **ignore-nh-metric** is configured skip this step.
10. Select the route received by the peer with the lowest Router ID; this comes from the ORIGINATOR_ID attribute (if present) or else the BGP identifier of the peer (received in its OPEN message). If **ignore-router-id** is configured and two EBGP routes are being compared keep the current best path and skip steps 11 and 12.
11. Select the route with the shortest CLUSTER_LIST length.
12. Select the route received from the peer with the lowest IP address.

Always Compare MED

By default, the MED path attribute is used in the decision process only if the routes being compared come from the same neighbor AS; if one of the paths lacks a MED attribute it is considered equal to a route with a MED of 0. These default rules can be modified using the **always-compare-med** command.

The **always-compare-med** command without the **strict-as** keyword allows MED to be compared in paths from different neighbor autonomous systems; in this case, if neither **zero** or **infinity** is part of the command, **zero** is inferred, meaning that a route without a MED attribute is handled as though it had a MED with value 0. When the **strict-as** keyword is present MED is only compared between paths from the same neighbor AS and in this case **zero** or **infinity** is mandatory and tells BGP how to interpret paths without a MED attribute.

Table 12 shows how the MED comparison of two paths is influenced by different forms of the **always-compare-med** command.

Table 12: MED Comparison with always-compare-med

Command	MED comparison step in decision process
no always-compare-med always-compare-med strict-as zero	Only compare the MED of two paths if they come from the same neighbor AS. If one path is missing a MED attribute treat it the same as MED=0.
always-compare-med always-compare-med zero	Always compare the MED of two paths, even if they come from different neighbor AS. If one path is missing a MED attribute treat it the same as MED=0.
always-compare-med infinity	Always compare the MED of two paths, even if they come from different neighbor AS. If one path is missing a MED attribute treat it the same as MED=infinity.
always-compare-med strict-as infinity	Only compare the MED of two paths if they come from the same neighbor AS. If one path is missing a MED attribute treat it the same as MED=infinity.

Ignore Next-Hop Metric

The **ignore-nh-metric** command allows the step comparing the distance to the BGP next-hop to be skipped. When this command is present in the **config**>**service**>**vprn** context it applies to the comparison of two imported BGP-VPN routes. When this command is present in the **config**>**router**>**bgp** context it applies to the comparison of any two BGP routes received by that instance. And when this command is present in the **config**>**service**>**vprn**>**bgp** context it applies to the comparison of two BGP routes learned from VPRN BGP peers (that is, CE peers). In all cases, this option is useful when there are multiple paths for a prefix that are equally preferred up to (but not including) the IGP cost comparison step of the BGP decision process and the network administrator wants all of them to be used for forwarding (*BGP-Multipath*).

BGP Route Installation in the Route Table

If the best BGP path for an IPv4 or IPv6 prefix is the most preferred route to the destination it is installed in the IP route table unless **disable-route-table-install** is configured. The best BGP path is the most preferred route if has the numerically lowest route preference among all routes, of all protocols, to the destination. The default preference value for BGP routes is 170 but this can be changed using the **preference** command in the BGP or policy configuration.

NOTE: Consider configuring the **disable-route-table-install** command on control-plane route reflectors that are not involved in packet forwarding (i.e. that do not modify the BGP NEXT_HOP); this improves the performance and scalability of such route reflectors.

If the best path can be installed in the route table and there are other BGP paths (LOC-RIB routes) for the same IPv4 or IPv6 prefix that are nearly as good as the best path the additional paths can also be installed in the route table. This is called *BGP-Multipath* and it must be explicitly enabled using the **multipath** command. The **multipath** command specifies the maximum number of BGP paths (up to 32), including the overall best path, that BGP can install in the route table for any particular IPv4 or IPv6 prefix; in this scenario each BGP path is effectively one ECMP next-hop of the IP route and traffic matching the IP route is load-shared across the ECMP next-hops based on a per-packet hash calculation.

By default the hashing is not *sticky*, meaning that when one or more of the equal-cost BGP next-hops fail all traffic flows matching the route are potentially moved to new BGP next-hops. If required, a BGP route can be marked (using the **sticky-ecmp** action in route policies) for sticky ECMP behavior so that BGP next-hop failures are handled by moving only the affected traffic flows to the remaining next-hops as evenly as possible.

NOTE: In order for BGP to install a route with *N* ECMP next-hops in the route-table the associated routing instance must have the **ecmp** command in its configuration and the max number of ECMP next-hops specified as part of that command must have a value greater than or equal to *N*.

In SR-OS a BGP path to an IPv4 or IPv6 prefix is a candidate for installation as an ECMP next-hop (subject to the path limits of the **multipath** and **ecmp** commands) only if it meets both of the following criteria:

1. It is the overall best BGP path or else it is tied with the overall best path up to and including step 9 of the decision process as summarized in the section titled BGP Decision Process on page 659.
2. Compared to other paths with the same BGP NEXT_HOP it is the best path (based on evaluation of all steps of the BGP decision process).

NOTE: VPRN routing instances support a special mode of BGP multipath called *EIBGP-Multipath*. In *EIBGP-Multipath* BGP routes learned from CE devices that are typically EBGP peers are combined with imported VPN-IP routes that typically come from IBGP peers to form an IP ECMP route. When *EIBGP-Multipath* is enabled a route is a candidate for installation as an ECMP next-hop if it is the overall best route or else it is tied with the overall best route up to and including the MED step of the BGP decision process.

SR-OS also supports a feature called *IBGP-Multipath*. In some topologies a BGP next-hop is resolved by an IP route (for example a static, OSPF or IS-IS route) that itself has multiple ECMP next-hops. When **ibgp-multipath** is not configured only one of these ECMP next-hops is programmed as a next-hop of the BGP route in the IOM. But when **ibgp-multipath** is configured the IOM attempts to use all of the ECMP next-hops of the resolving route in forwarding.

Although the name of the **ibgp-multipath** command implies that it is specific to IBGP-learned routes this is not the case; it applies to routes learned from any multi-hop BGP session including routes learned from multi-hop EBGP peers.

It is important to note that *BGP-Multipath* and *IBGP-Multipath* are not mutually exclusive and work together. *BGP-Multipath* enables ECMP load-sharing across different BGP next-hops (corresponding to different LOC-RIB routes) and *IBGP-Multipath* enables ECMP load-sharing across different IP next-hops of IP routes that resolve the BGP next-hops.

The final point about *IBGP-Multipath* is that it does not control load-sharing of traffic towards a BGP next-hop that is resolved by a tunnel, such as the case when dealing with BGP shortcuts or labeled routes (VPN-IP, label-IPv4, 6PE). When a BGP next-hop is resolved by a tunnel that supports ECMP the load-sharing of traffic across the ECMP next-hops of the tunnel is automatic.

NOTE: At the current time SR-OS does not support direct resolution of a BGP next-hop to multiple RSVP-TE tunnels. However a BGP next-hop can be resolved by multiple LDP ECMP next-hops that each correspond to a separate LDP-over-RSVP tunnel. It is also possible for a BGP next-hop to be resolved by an IGP shortcut route that has multiple RSVP-TE tunnels as its ECMP next-hops.

Weighted ECMP for BGP Routes

In some cases, the ECMP BGP next-hops of an IP route correspond to paths with very different bandwidths and it makes sense for the ECMP load-balancing algorithm to distribute traffic across the BGP next-hops in proportion to their relative bandwidths. The bandwidth associated with a path can be signaled to other BGP routers by including a Link Bandwidth Extended Community in the BGP route. The Link Bandwidth Extended Community is optional and non-transitive and encodes an autonomous system (AS) number and a bandwidth.

In SR OS, a Link Bandwidth Extended Community can be added to an IPv4, IPv6, VPN-IPv4 or VPN-IPv6 route using either route policies or the **ebgp-link-bandwidth** command. The **ebgp-link-bandwidth** command is supported in BGP group and neighbor configuration contexts and automatically adds (on import) a Link Bandwidth Extended Community to received routes from single-hop (directly connected) EBGP peers. Note that when a route is advertised to an EBGP peer, the Link Bandwidth Extended Community, if present, is always removed. The Link Bandwidth Extended Community associated with a BGP route can be displayed using the **show router bgp routes** commands; for the bandwidth value, the system automatically converts the binary value in the extended community to a decimal number in units of Mbps (1000000 bit/s).

7x50 routers automatically performs weighted ECMP for an IP BGP route when the ingress card is FP2 or better and all the ECMP BGP next-hops of the route include a Link Bandwidth Extended Community. The relative weight of traffic sent to each BGP next-hop is visible in the output of the **show router route-table extensive** and **show router fib extensive** commands.

Weighted ECMP across the BGP next-hops of an IP BGP route is supported in combination with ECMP at the level of the route or tunnel that resolves one or more of the ECMP BGP next-hops. This ECMP at the resolving level can also be weighted ECMP when the following conditions all apply:

- The BGP next-hop is resolved by an IP route (OSPF, IS-IS or static) with MPLS LSP ECMP next-hops
- **ibgp-multipath** is configured under BGP
- **config router weighted-ecmp** is configured (requires chassis mode D)

BGP Route Installation in the Tunnel Table

If the best BGP path for a /32 IPv4 prefix is a label-IPv4 route (AFI 1, SAFI 4), and if it has the numerically lowest **preference** value among all routes (regardless of protocol) for the /32 IPv4 prefix, and if **disable-route-table-install** is *not* configured, the label-IPv4 route is automatically added, as a *BGP tunnel* entry, to the tunnel table. In SR-OS the tunnel-table is used to resolve a BGP next-hop to a tunnel when required by the configuration or the type of route (see the section titled Next-Hop Resolution on page 648 for many of these details). BGP tunnels play a key role in the following solutions:

- Inter-AS IP VPN model C
- Inter-AS L2 VPN model C
- Carrier Supporting Carrier (CSC)
- Intra-AS seamless MPLS

BGP tunnels have a preference of 10 in the tunnel table, compared to 9 for LDP tunnels and 7 for RSVP tunnels, so if the router configuration allows all types of tunnels to resolve a BGP next-hop an RSVP LSP is preferred over an LDP tunnel and an LDP tunnel is preferred over a BGP tunnel.

If **multipath** and **ecmp** are configured appropriately a BGP tunnel can be installed in the tunnel table with multiple ECMP next-hops, each one corresponding to a path through a different BGP next-hop; the multipath selection process outlined in the previous section (BGP Route Installation in the Route Table on page 661) also applies to this case.

For BGP tunnels there is no support for the equivalent of *IBGP-Multipath*. That is, if a BGP next-hop of the label-IPv4 route in the tunnel table is resolved by an LDP tunnel with multiple ECMP next-hops load-sharing is not supported across the LDP ECMP next-hops; only the first next-hop carries traffic towards the BGP next-hop.

BGP Fast Reroute

BGP fast reroute is a feature that brings together indirection techniques in the forwarding plane and pre-computation of BGP backup paths in the control plane to support fast reroute of BGP traffic around unreachable/failed BGP next-hops. BGP fast reroute is supported with IPv4, labeled-IPv4, IPv6, 6PE, VPN-IPv4 and VPN-IPv6 routes. The scenarios supported by the base router BGP context are outlined in Table 13.

Note that BGP fast reroute information specific to IP VPNs is described in the BGP Fast Reroute in a VPRN section of the 7x50 SR OS Services Guide.

Table 13: BGP Fast Reroute Scenarios (Base Context)

Ingress Packet	Primary Route	Backup Route	Prefix Independent Convergence
IPv4	IPv4 route with next-hop A resolved by an IPv4 route or an LDP or RSVP shortcut tunnel	IPv4 route with next-hop B resolved by an IPv4 route or an LDP or RSVP shortcut tunnel	Yes
IPv6	IPv6 route with next-hop A resolved by an IPv6 route OR 6PE route with next-hop A resolved by an LDP tunnel	IPv6 route with next-hop B resolved by an IPv6 route OR 6PE route with next-hop B resolved by an LDP tunnel	Yes, but if the 6PE routes are label-per-prefix the ingress card must be IOM3 or better for PIC
IPv4	Lbl-IPv4 route with next-hop A resolved by an LDP or RSVP tunnel	Lbl-IPv4 route with next-hop B resolved by an LDP or RSVP tunnel	Yes, if ingress card is IOM3 or better
IPv4	Lbl-IPv4 route with next-hop A resolved to an interface	Lbl-IPv4 route with next-hop B resolved to an interface	Yes, if ingress card is IOM3 or better
MPLS or Service	Lbl-IPv4 route with next-hop A resolved by an LDP or RSVP tunnel	Lbl-IPv4 route with next-hop B resolved by an LDP or RSVP tunnel	Yes
MPLS or Service	Lbl-IPv4 route with next-hop A resolved to an interface	Lbl-IPv4 route with next-hop B resolved to an interface	Yes

Calculating Backup Paths

In SR-OS BGP fast reroute is optional and must be enabled using either the **backup-path** or **install-backup-path** command.

The **backup-path** command is used in the base router context to control fast reroute on a per-routing instance and per-family (IPv4 and IPv6) basis. The command supports options to enable fast reroute for IPv4 prefixes only, for IPv6 prefixes only, or for all IPv4 and IPv6 prefixes.

The **install-backup-path** command is used to designate a specific set of IPv4 or IPv6 prefixes that are eligible for BGP fast reroute protection. The command enables a BGP import policy to restrict the set of routes that are programmed with a backup path.

When BGP fast reroute is enabled the control plane attempts to find an eligible backup path for every received IPv4 and/or IPv6 prefix, depending on configuration. In general the backup path is the single best path remaining after the primary ECMP paths and any paths with the same BGP next-hops as these paths have been removed. However the following points should be noted:

- A backup path is not calculated for a prefix if the best path is a labeled-IPv4 route and it has been programmed with multiple ECMP next-hops through different BGP next-hops.
- For labeled-IPv4 prefixes that are re-advertised with a new BGP next-hop the programmed backup path is the same for all prefixes that have the same best path and received label, even if the calculated backup path is different for some of the prefixes.

Failure Detection and Switchover to the Backup Path

When BGP fast reroute is enabled the IOM reroutes traffic onto a backup path based on input from BGP. When BGP decides that a primary path is no longer usable it notifies the IOM and affected traffic is immediately switched to the backup path.

The following events trigger failure notifications to the IOM and reroute of traffic to backup paths:

- Peer IP address unreachable and peer-tracking is enabled
- BFD session associated with BGP peer goes down
- BGP session terminated with peer (for example, send/receive NOTIFICATION)
- There is no longer any route (allowed by the next-hop resolution policy, if configured) that can resolve the BGP next-hop address
- The LDP tunnel that resolves the next-hop goes down. This could happen because there is no longer any IP route that can resolve the FEC, or the LDP session goes down, or the LDP peer withdraws its label mapping.
- The RSVP tunnel that resolves the next-hop goes down. This could happen because a ResvTear message is received, or the RESV state times out, or the outgoing interface fails and is not protected by FRR or a secondary path.
- The BGP tunnel that resolves the next-hop goes down. This could happen because the BGP label-IPv4 route is withdrawn by the peer or else becomes invalid due to an unresolved next-hop.

QoS Policy Propagation via BGP (QPPB)

QPPB is a feature that allows different QoS values (forwarding class and optionally priority) to be associated with different IPv4 and IPv6 BGP LOC-RIB routes based on BGP import policy processing. This is done so that when traffic arrives on a QPPB-enabled IP interface and its source or destination IP address matches a BGP route with QoS information the packet is handled according to the QoS of the matching route. SR-OS supports QPPB on the following types of interfaces:

- Base router network interfaces
- IES and VPRN SAP interfaces
- IES and VPRN spoke-SDP interfaces
- IES and VPRN subscriber interfaces

QPPB is enabled on an interface using the **qos-route-lookup** command. There are separate commands for IPv4 and IPv6 so that QPPB can be enabled in one mode (source or destination or none) for IPv4 packets arriving on the interface and a different mode (source or destination or none) for IPv6 packets arriving on the interface.

NOTE: Source-based QPPB is not supported on subscriber interfaces.

Different LOC-RIB routes for the same IP prefix may be associated with different QPPB information. If these LOC-RIB routes are combined in support of ECMP or BGP fast reroute then the QPPB information becomes next-hop specific. This means that in destination QPPB mode the QoS assigned to a packet depends on the BGP next-hop that is selected for that particular packet by the ECMP hash or fast reroute algorithm. In source QPPB mode the QoS assigned to a packet comes from the first BGP next-hop of the IP route matching the source address.

BGP Policy Accounting

Policy accounting is a feature that allows different *accounting classes* to be associated with IPv4 and IPv6 BGP LOC-RIB routes based on BGP import policy processing. This is done so that per-accounting-class traffic statistics can be collected on policy accounting-enabled interfaces of the router. Policy accounting interfaces are only supported on IOM3 or better cards. The following types of interfaces are supported:

- Base router network interfaces
- IES and VPRN SAP interfaces
- IES and VPRN spoke-SDP interfaces
- IES and VPRN subscriber interfaces

Policy accounting is enabled on an interface using the **policy-accounting** command. The name of a policy accounting template must be specified. Each policy accounting template contains a list of *source classes* and *destination classes*. 7x50 routers support up to 255 different source classes and up to 255 different destination classes. Each source class is identified by an index number (1-255) and each destination class is identified by an index number (1-255). The policy accounting template tells the IOM what accounting classes to collect stats for on a policy accounting interface. SR-OS supports up to 1024 different templates, depending on the chassis type.

NOTE: Policy accounting templates containing one or more source class identifiers cannot be applied to subscriber interfaces.

Through policy mechanisms a LOC-RIB route for an IP prefix can have a source class index (1-25), a destination class index (1-255) or both. When an ingress packet on a policy-accounting enabled interface [I1] is forwarded by the IOM and its destination address matches a BGP route with a destination class index [D], and [D] is listed in the relevant policy accounting template, packets-forwarded and IP-bytes-forwarded counters for [D] on interface [I1] are incremented accordingly. Similarly, when an ingress packet on a policy-accounting enabled interface [I2] is forwarded by the IOM and its source address matches a BGP route with a source class index [S], and [S] is listed in the relevant policy accounting template, the packets-forwarded and IP-bytes-forwarded counters for [S] on interface [I2] are incremented accordingly.

It is possible that different LOC-RIB routes for the same IP prefix are associated with different accounting class information. If these LOC-RIB routes are combined in support of ECMP or BGP fast reroute then the destination-class of a packet depends on the BGP next-hop that is selected for that particular packet by the ECMP hash or fast reroute algorithm. If the source address of a packet matches a route with multiple BGP next-hops its source-class is derived from the first BGP next-hop of the matching route.

Route Flap Damping (RFD)

Route flap damping is a mechanism supported by 7x50 and other BGP routers that was designed to help improve the stability of Internet routing by mitigating the impact of route flaps. Route flaps describe a situation where a router alternately advertises a route as reachable and then unreachable or as reachable through one path and then another path in rapid succession. Route flaps can result from hardware errors, software errors, configuration errors, unreliable links, etc. However not all perceived route flaps represent a true problem; when a best path is withdrawn the next-best path may not be immediately known and may trigger a number of intermediate best path selections (and corresponding advertisements) before it is found. These intermediate best path selections may travel at different speeds through different routers due to the effect of the min-route-advertisement interval (MRAI) and other factors. RFD does not handle this type of situation particularly well and for this and other reasons many Internet service providers do not use RFD.

In SR-OS route flap damping is configurable; by default it is disabled. It can be enabled on EBGP and confed-EBGP sessions by including the **damping** command in their group or neighbor configuration. The **damping** command has no effect on IBGP sessions. When a route of any type (any AFI/SAFI) is received on a non-IBGP session that has **damping** enabled:

- If the route changes from reachable to unreachable due to a withdrawal by the peer then damping history is created for the route (if it does not already exist) and in that history the Figure of Merit (FOM), an accumulated penalty value, is incremented by 1024.
- If a reachable route is updated by the peer with new path attribute values then the FOM is incremented by 1024.
- In SR-OS the FOM has a hard upper limit of 21540 (not configurable).
- The FOM value is decayed exponentially as described in RFC 2439. The **half-life** of the decay is 15 minutes by default, however a BGP import policy can be used to apply a non-default damping profile to the route, and the **half-life** in the non-default damping profile can have any value between 1 and 45 minutes.

- The FOM value at the last time of update can be displayed using the **show router bgp damping detail** command. Note that the time of last update can be up to 640 seconds ago; SR-OS does not calculate the current FOM every time the show command is entered.
- When the FOM reaches the suppress limit, which is 3000 by default but can be changed to any value between 1 and 20000 in a non-default damping profile, the route is suppressed, meaning it is not used locally and not advertised to peers. The route remains suppressed until either the FOM exponentially decays to a value less than or equal to the **reuse** threshold or the **max-suppress** time is reached. By default the **reuse** threshold is 750 and the **max-suppress** time is 60 minutes, but these can be changed in a non-default damping profile: **reuse** can have a value between 1 and 20000 and **max-suppress** can have a value between 1 and 720 minutes.

RIB-OUT Features

SR-OS implements the following features related to RIB-OUT processing.

- BGP export policies
- Outbound route filtering (ORF)
- RT constrained route distribution
- Configurable min-route-advertisement (MRAI)
- Advertise-inactive
- Best-external
- Add-path
- Split-horizon

These features are discussed in the following sections.

BGP Export Policies

The **export** command is used to apply one or more policies (up to 15) to a neighbor, group or to the entire BGP context. The **export** command that is most-specific to a peer is the one that is applied. An **export** policy command applied at the **neighbor** level takes precedence over the same command applied at the **group** or global level. An **export** policy command applied at the **group** level takes precedence over the same command specified on the global level. The **export** policies applied at different levels are not cumulative. The policies listed in an **export** command are evaluated in the order in which they are specified.

NOTE: The **export** command can reference a policy before it has been created (as a **policy-statement**).

The most common uses for BGP export policies are as follows:

- To locally originate a BGP route by exporting (or redistributing) a non-BGP route that is installed in the route table and actively used for forwarding. The non-BGP route is most frequently a direct, static or aggregate route (exporting IGP routes into BGP is generally not recommended).
- To block the advertisement of certain BGP routes towards specific BGP peers. The routes may be blocked on the basis of IP prefix, communities, etc.

- To modify the attributes of BGP routes advertised to specific BGP peers. The following path attribute modifications are possible using BGP export policies:
  - → Change the ORIGIN value
  - → Add a sequence of AS numbers to the start of the AS_PATH. Note that when a route is advertised to an EBGP peer the addition of the local-AS/global-AS numbers to the AS_PATH is always the final step (done after export policy).
  - → Replace the AS_PATH with a new AS_PATH. Note that when a route is advertised to an EBGP peer the addition of the local-AS/global-AS numbers to the AS_PATH is always the final step (done after export policy).
  - → Prepend an AS number multiple times to the start of the AS_PATH. Note that when a route is advertised to an EBGP peer the addition of the local-AS/global-AS numbers to the AS_PATH is always the final step (done after export policy). Also note that the add/replace action on the AS_PATH supersedes the prepend action if both are specified in the same policy entry.
  - → Change the NEXT_HOP to a specific IP address. Note that when a route is advertised to an EBGP peer the next-hop cannot be changed from the local-address.
  - → Change the NEXT_HOP to the local-address used with the peer (next-hop-self).
  - → Add a value to the MED. If the MED attribute does not exist it is added.
  - → Subtract a value from the MED. If the MED attribute does not exist it is added with a value of 0. If the result of the subtraction is a negative number the MED metric is set to 0.
  - → Set the MED to a particular value.
  - → Set the MED to the cost of the IP route (or tunnel) used to resolve the BGP next-hop.
  - → Set LOCAL_PREF to a particular value when advertising to an IBGP peer.
  - → Add, remove and/or replace standard communities
  - → Add, remove and/or replace extended communities
  - → Add a static value to the AIGP metric when advertising the route to an AIGP-enabled peer with a modified BGP next-hop. The static value is incremental to the automatic adjustment of the LOC-RIB AIGP metric to reflect the distance between the local router and the received BGP next-hop.
  - → Increment the AIGP metric by a fixed amount when advertising the route to an AIGP-enabled peer with a modified BGP next-hop. The static value is a substitute for the dynamic value of the distance between the local router and the received BGP next-hop.

Outbound Route Filtering (ORF)

Outbound route filtering (ORF) is a mechanism that allows one router, the ORF-sending router to signal to a peer, the ORF-receiving router, a set of route filtering rules (ORF entries) that the ORF-receiving router should apply to its route advertisements towards the ORF-sending router. The ORF entries are encoded in Route Refresh messages.

The use of ORF on a session must be negotiated —i.e. both routers must advertise the ORF capability in their Open messages. The ORF capability describes the address families that support ORF, and for each address family, the ORF types that are supported and the ability to send/receive each type. 7x50 routers support ORF type 3, which is ORF based on Extended Communities. It is supported for only the following address families:

- VPN-IPv4
- VPN-IPv6
- MVPN-IPv4
- MVPN-IPv6

In SR-OS the send/receive capability for ORF type 3 is configurable (with the **send-orf** and **accept-orf** commands) but the setting applies to all supported address families.

The SR-OS support for ORF type 3 allows a PE router that imports VPN routes with a particular set of Route Target Extended Communities to indicate to a peer (for example a route reflector) that it only wants to receive VPN routes that contain one or more of these Extended Communities. When the PE router wants to inform its peer about a new RT Extended Community it sends a Route Refresh message to the peer containing an ORF type 3 entry instructing the peer to *add* a *permit* entry for the 8-byte extended community value. When the PE router wants to inform its peer about a RT Extended Community that is no longer needed it sends a Route Refresh message to the peer containing an ORF type 3 entry instructing the peer to *remove* the *permit* entry for the 8-byte extended community value.

In SR-OS the type-3 ORF entries that are sent to a peer can be generated dynamically (if no Route Target Extended Communities are specified with the **send-orf** command) or else specified statically. Dynamically generated ORF entries are based on the route targets that are imported by all locally-configured VPRNs.

A router that has installed ORF entries received from a peer can still apply BGP export policies to the session. If the evaluation of a BGP export policy results in a reject action for a VPN route that matches a permit ORF entry the route is not advertised — i.e. the export policy has the final word.

NOTE: The 7x50 implementation of ORF filtering is very efficient. It takes less time to filter a large number of VPN routes with ORF than it does to reject non-matching VPN routes using a conventional BGP export policy.

Despite the advantages of ORF compared to manually configured BGP export policies a better technology, when it comes to dynamic filtering based on Route Target Extended Communities, is RT Constraint. RT Constraint is discussed further in the next section.

RT Constrained Route Distribution

RT constrained route distribution, or RT-constrain for short, is a mechanism that allows a router to advertise to certain peers a special type of MP-BGP route called an RTC route; the associated AFI is 1 and the SAFI is 132. The NLRI of an RTC route encodes an Origin AS and a Route Target Extended Community with prefix-type encoding (i.e. there is a prefix-length and “host” bits after the prefix-length are set to zero). A peer receiving RTC routes does not advertise VPN routes to the RTC-sending router unless they contain a Route Target Extended Community that matches one of the received RTC routes. As with any other type of BGP route RTC routes are propagated loop-free throughout and between Autonomous Systems. If there are multiple RTC routes for the same NLRI the BGP decision process selects one as the best path. The propagation of the best path installs RIB-OUT filter rules as it is travels from one router to the next and this process creates an optimal VPN route distribution tree rooted at the source of the RTC route.

NOTE: RT-constrain and Extended Community-based ORF are similar to the extent that they both allow a router to signal to a peer the Route Target Extended Communities they want to receive in VPN routes from that peer. But RT-constrain has distinct advantages over Extended Community-based ORF: it is more widely supported, it is simpler to configure, and its distribution scope is not limited to a direct peer.

In SR-OS the capability to exchange RTC routes is advertised when the **route-target** keyword is added to the relevant **family** command. RT-constrain is supported on EBGP and IBGP sessions of the base router instance. On any particular session either ORF or RT-constrain may be used but not both; if RT-constrain is configured the ORF capability is not announced to the peer.

When RT-constrain has been negotiated with one or more peers SR-OS automatically originates and advertises to these peers one /96 RTC route (the origin AS and Route Target Extended Community are fully specified) for every route target imported by a locally-configured VPRN or BGP-based L2 VPN; this includes MVPN-specific route targets.

SR-OS also supports a group/neighbor level **default-route-target** command that causes the 7x50 router to generate and send a 0:0:0/0 default RTC route to one or more peers. Sending the default RTC route to a peer conveys a request to receive all VPN routes from that peer. The **default-route-target** command is typically configured on sessions that a route reflector has with its PE clients. Note that a received default RTC route is never propagated to other routers.

The advertisement of RTC routes by a route reflector follows special rules that are described in RFC 4684. These rules are needed to ensure that RTC routes for the same NLRI that are originated by different PE routers in the same Autonomous System are properly distributed within the AS.

When a BGP session comes up, and RT-constrain is enabled on the session (both peers advertised the MP-BGP capability), the 7x50 router delays sending any VPN-IPv4 and VPN-IPv6 routes until either the session has been up for 60 seconds or the End-of-RIB marker is received for the RT-constrain address family. When the VPN-IPv4 and VPN-IPv6 routes are sent they are filtered to include only those with a Route Target Extended Community that matches an RTC route from the peer. VPN-IP routes matching an RTC route originated in the local AS are advertised to any IBGP peer that advertises a valid path for the RTC NLRI — i.e. route distribution is not constrained to only the IBGP peer advertising the best path. On the other hand VPN-IP routes matching an RTC route originated outside the local AS are only advertised to the EBGP or IBGP peer that advertises the best path.

NOTE: SR-OS does not support an equivalent of *BGP-Multipath* for RT-Constrain routes. There is no way to distribute VPN routes across more than one 'almost' equal set of inter-AS paths.

On 7x50 routers received RTC routes have no effect on the advertisement on MVPN-IPv4, MVPN-IPv6 and L2-VPN routes.

Min Route Advertisement Interval (MRAI)

According to the BGP standard (RFC 4271) a BGP router should not send updated reachability information for an NLRI to a BGP peer until a certain period of time, called the *Min Route Advertisement Interval*, has elapsed since the last update. The RFC suggests the MRAI should be configurable per peer but does not propose a specific algorithm and therefore MRAI implementation details vary from one router operating system to another.

In SR-OS the MRAI is configurable, on a per-session basis, using the **min-route-advertisment** command. The **min-route-advertisement** command can be configured with any value between 1 and 255 seconds and the setting applies to all address families. The default value is 30 seconds, regardless of the session type (EBGP or IBGP). When all RIB-OUT routes have been sent to a peer the MRAI timer associated with that session is started and when it expires the RIB-OUT changes that have accumulated while the timer was running trigger the sending of a new set of UPDATE messages to the peer.

It may be important to send UPDATE messages that advertise new NLRI reachability information more frequently for some address families than others. SR-OS offers a **rapid-update** command that overrides the peer-level **min-route-advertisement** time and applies the minimum setting to routes belonging to specific address families; routes of other address families continue to be advertised according to the session-level MRAI setting. The address families that can be configured with **rapid-update** support are:

- L2-VPN
- MVPN-IPv4

- MVPN-IPv6
- MDT-SAFI
- EVPN

In many cases the default MRAI is appropriate for all address families (or at least those not included in the above list) when it applies to UPDATE messages that advertise reachable NLRI but it is less than ideal for UPDATE messages that advertise unreachable NLRI (route withdrawals). Fast re-convergence after some types of failures requires route withdrawals to propagate to other routers are quickly as possible so that they can calculate and start using new best paths and this is impeded by the effect of the MRAI timer at each router hop. SR-OS provides a solution for this problem by supporting a configuration command called **rapid-withdrawal**. When **rapid-withdrawal** is configured UPDATE messages containing withdrawn NLRI are sent immediately to a peer — without waiting for the MRAI timer to expire. UPDATE messages containing reachable NLRI continue to wait for the MRAI timer to expire, and this timer remains governed by the **min-route-advertisement** time or the **rapid-update** command, if it applies. When **rapid-withdrawal** is enabled it applies to all address families.

Advertise-Inactive

Standard BGP rules do not allow a BGP route to be advertised to peers unless it is the best path and it is 'used' locally. An IPv4 or IPv6 BGP route is considered 'used' if it is the *active* route to the destination in the route table. If there a multiple routes from different protocols for the same IP destination the BGP route is 'used' only if it has the numerically lowest route preference among all these routes; for further details refer to the section titled BGP Route Installation in the Route Table on page 661.

In some cases it may be useful to advertise the best BGP path to peers despite the fact that is *inactive* —i.e. because there are one or more lower-preference non-BGP routes to the same destination and one of these other routes is the *active* route. One way SR-OS supports this flexibility is using the **advertise-inactive** command; other methods include *Best-External* and *Add-Paths*.

As a global BGP configuration option the **advertise-inactive** command applies to all IPv4 and IPv6 routes and all sessions that advertise these routes. When the command is configured and the best BGP path is inactive it is automatically advertised to every peer unless rejected by a BGP export policy.

Best-External

Best-External is a BGP enhancement that allows a BGP speaker to advertise to its IBGP peers its best "external" route for a prefix/NLRI when its best overall route for the prefix/NLRI is an "internal" route. This is not possible in a normal BGP configuration because the base BGP specification prevents a BGP speaker from advertising a non-best route for a destination.

In certain topologies *Best-External* can improve convergence times, reduce route oscillation and allow better loadsharing. This is achieved because routers internal to the AS have knowledge of more exit paths from the AS. Enabling *Add-Paths* on border routers of the AS can achieve a similar result but *Add-Paths* introduces NLRI format changes that must be supported by BGP peers of the border router and therefore has more interoperability constraints than *Best-External* (which requires no messaging changes).

Best-External is supported in the base router BGP context. (A related feature is also supported in VPRNs; consult the Services Guide for more details.) It is configured using the **advertise-external** command, which provides IPv4 and IPv6 as options. *Best-External* for IPv4 applies to both regular IPv4 unicast routes as well as labeled-IPv4 (SAFI4) routes. Similarly, *Best-External* for IPv6 applies to both regular IPv6 unicast routes as well as 6PE (SAFI4) routes.

The advertisement rules when **advertise-external** is enabled can be summarized as follows:

- If a router has **advertise-external** enabled and its best overall route is a route from an IBGP peer then this best route is advertised to EBGP and confed-EBGP peers, and the "best external" route is advertised to IBGP peers. The "best external" route is the one found by running the BGP path selection algorithm on all LOC-RIB paths except for those learned from the IBGP peers.

NOTE: A 7x50 route reflector with **advertise-external** enabled does not include IBGP routes learned from other clusters in its definition of 'external'.

- If a router has **advertise-external** enabled and its best overall route is a route from an EBGP peer then this best route is advertised to EBGP, confed-EBGP, and IBGP peers.
- If a router has **advertise-external** enabled and its best overall route is a route from a confed-EBGP peer in member AS X then this best route is advertised to EBGP, IBGP peers and confed-EBGP peers in all member AS except X and the "best external" route is advertised to confed-EBGP peers in member AS X. In this case the "best external" route is the one found by running the BGP path selection algorithm on all RIB-IN paths except for those learned from member AS X.

NOTE: If the best-external route is not the best overall route it is not installed in the forwarding table and in some cases this can lead to a short-duration traffic loop after failure of the overall best path.

Add-Paths

Add-Paths is a BGP enhancement that allows a BGP router to advertise multiple distinct paths for the same prefix/NLRI. This provides a number of potential benefits, including reduced routing churn, faster convergence, and better loadsharing.

In order for a router to receive multiple paths per NLRI from a peer, for a particular address family, the peer must announce the BGP capability to send multiple paths for the address family and the local router must announce the BGP capability to receive multiple paths for the address family. When the Add-Path capability has been negotiated this way all advertisements and withdrawals of NLRI by the peer must include a path identifier. The path identifier has no significance to the receiving router. If the combination of NLRI and path identifier in an advertisement from a peer is unique (does not match an existing route in the RIB-IN from that peer) then the route is added to the RIB-IN. If the combination of NLRI and path identifier in a received advertisement is the same as an existing route in the RIB-IN from the peer then the new route replaces the existing one. If the combination of NLRI and path identifier in a received withdrawal matches an existing route in the RIB-IN from the peer then that route is removed from the RIB-IN.

An UPDATE message carrying an IPv4 NLRI with a path identifier is shown in Figure 28.

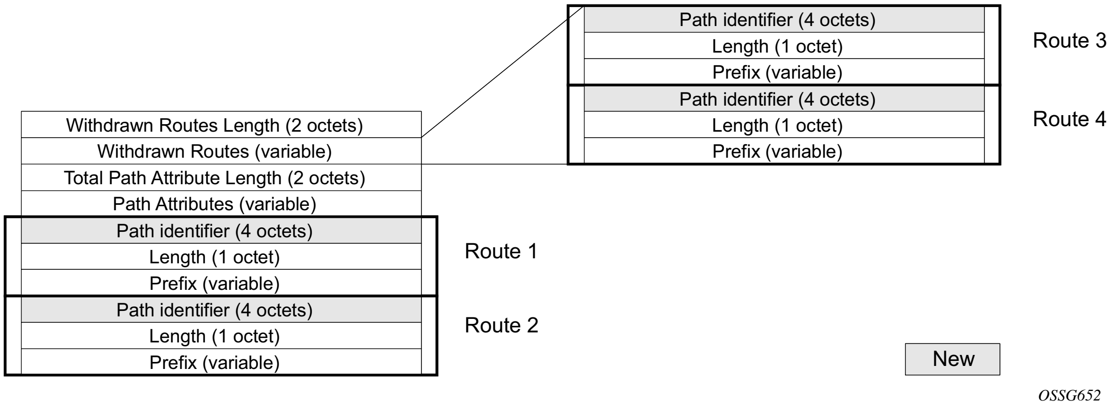
Path identifier (4 octets)
Length (1 octet)
Prefix (variable)
Route 3
Path identifier (4 octets)
Length (1 octet)
Prefix (variable)
Route 4
Withdrawn Routes Length (2 octets)
Withdrawn Routes (variable)
Total Path Attribute Length (2 octets)
Path Attributes (variable)
Path identifier (4 octets)
Length (1 octet)
Prefix (variable)
Route 1
Path identifier (4 octets)
Length (1 octet)
Prefix (variable)
Route 2
New
OSSG652

Figure 28: BGP Update Message with Path Identifier for IPv4 NLRI

Add-Paths is only supported by the base router BGP instance and the EBGP and IBGP sessions it forms with other *Add-Paths* capable peers. The ability to send and receive multiple paths per prefix is configurable per family, with the supported options being:

- IPv4 (including labeled IPv4 routes)
- VPN-IPv4
- IPv6 (including labeled IPv6 routes)
- VPN-IPv6

Path Selection with Add-Paths

The LOC-RIB may have multiple paths for a prefix. The path selection mode refers to the algorithm used to decide which of these paths to advertise to an Add-Paths peer. SR-OS supports the Add-N path selection algorithm described in *draft-ietf-idr-add-paths-guidelines*. The Add-N algorithm selects, as candidates for advertisement, the N best paths with unique BGP next-hops. In the SROS implementation, the default value of N is configurable, per address-family, at the BGP instance, group and neighbor levels, however, this default value can be overridden, for specific prefixes, using route policies. The maximum number of paths to advertise for a prefix to an Add-Paths neighbor is the value N assigned by a BGP import policy to the best path for P, otherwise it defaults to the neighbor, group or instance level configuration of N for the address family to which P belongs.

Add-Paths allows non-best paths to be advertised to a peer, but it still complies with basic BGP advertisement rules such as the IBGP split horizon rule: a route learned from an IBGP neighbor cannot be re-advertised to another IBGP neighbor unless the router is configured as a route reflector.

Split-Horizon

Split-horizon refers to the action taken by a router to avoid advertising a route back to the peer from which it was received. By default SR-OS applies split-horizon behavior only to routes received from IBGP non-client peers. This split-horizon functionality, which can never be disabled, prevents a route learned from a non-client IBGP peer to be advertised to the sending peer or any other non-client peer.

To apply split-horizon behavior to routes learned from RR clients, confed-EBGP peers or (non-confed) EBGP peers the **split-horizon** command must be configured in the appropriate contexts; it is supported at the global BGP, **group** and **neighbor** levels. When **split-horizon** is enabled on these types of sessions it only prevents the advertisement of a route back to its originating peer; for example SR-OS does not prevent the advertisement of a route learned from one EBGP peer back to different EBGP peer in the same neighbor AS.

BGP Applications

SR OS implements the following BGP applications:

- Next-hop Resolution Using Tunnels on page 680
- BGP Flow-Spec on page 685

Next-hop Resolution Using Tunnels

BGP Routes

The user enables the resolution of IPv4 prefixes using tunnels to BGP next-hops in TTM with the following command:

```
configure>router>bgp>next-hop-resolution
            shortcut-tunnel
                  [no] family {ipv4}
                        resolution {any|disabled|filter}
                        resolution-filter
                              [no] bgp
                              [no] ldp
                              [no] rsvp
                              [no] sr-isis
                              [no] sr-ospf
                        [no] disallow-igp
                        exit
                  exit
            exit
```

The **shortcut-tunnel** and **family** nodes are contexts to configure the binding of BGP unlabelled routes to tunnels.

The default resolution of a BGP unlabelled route is performed in RTM. The user must configure the **resolution** option to enable resolution to tunnels in TTM. If the **resolution** option is explicitly set to **disabled**, the binding to tunnel is removed and resolution resumes in RTM to IP next-hops.

If **resolution** is set to **any**, any supported tunnel type in BGP shortcut context will be selected following TTM preference. If one or more explicit tunnel types are specified using the **resolution-filter** option, then only these tunnel types will be selected again following the TTM preference.

The following tunnel types are supported in a BGP shortcut context and in order of preference: RSVP, LDP, Segment Routing (SR), and BGP.

- The **rsvp** value instructs BGP to search for the best metric RSVP LSP to the address of the BGP next-hop. This address can correspond to the system interface or to another loopback used by the BGP instance on the remote node. The LSP metric is provided by MPLS in the tunnel table. In the case of multiple RSVP LSPs with the same lowest metric, BGP selects the LSP with the lowest tunnel-id.
- The **ldp** value instructs BGP to search for an LDP LSP with a FEC prefix corresponding to the address of the BGP next-hop.
- The **bgp** value instructs BGP to search for a BGP LSP with a RFC 107 label route prefix matching the address of the BGP next-hop.
- When the **sr-isis** or **sr-ospf** value is enabled, an SR tunnel to the BGP next-hop is selected in the TTM from the lowest preference ISIS or OSPF instance and if many instances have the same lowest preference from the lowest numbered IS-IS or OSPF instance

The user must set **resolution** to **filter** to activate the list of tunnel-types configured under **resolution-filter**.

If **disallow-igp** is enabled, the BGP route will not be activated using IP next-hops in RTM if no tunnel next-hops are found in TTM.

BGP Labeled Routes

The user enables the resolution of RFC 3107 BGP label route prefixes using tunnels to BGP next-hops in TTM with the following command:

```
configure>router> bgp>next-hop-resolution>
            label-route-transport-tunnel
                  [no] family {ipv4, vpn}
                        resolution {any|disabled|filter}
                        resolution-filter
                              [no] ldp
                              [no] rsvp
                              [no] sr-isis
                              [no] sr-ospf
                        exit
                  exit
            exit
```

Note that the **label-route-transport-tunnel** and **family** CLI nodes are contexts used to configure the binding of IPv4 or IPv6 BGP labeled routes to tunnels.

The **label-route-transport-tunnel** command provides a separate control for the different families of RFC 3107 BGP label routes: core IPv4 routes and inter-AS option B vpn-ipv4 and vpn-ipv6 routes at ASBR.

By default, core IPv4 routes and inter-AS option B VPN label routes resolve to LDP without the user needing to enter this command. IPv6 BGP labeled routes routes currently resolve only to IPv4 LDP tunnel with the 6PE feature and do not require this command.

If the **resolution** option is explicitly set to **disabled**, the default binding to LDP tunnel resumes. If **resolution** is set to **any**, any supported tunnel type in BGP label route context will be selected following TTM preference.

The following tunnel types are supported in a BGP label route context and in order of preference: RSVP, LDP, segment routing OSPF, and segment routing IS-IS.

- The **rsvp** value instructs BGP to search for the best metric RSVP LSP to the address of the BGP next-hop. This address can correspond to the system interface or to another loopback used by the BGP instance on the remote node. The LSP metric is provided by MPLS in the tunnel table. In the case of multiple RSVP LSPs with the same lowest metric, BGP selects the LSP with the lowest tunnel-id.
- The **ldp** value instructs BGP to search for an LDP LSP with a FEC prefix corresponding to the address of the BGP next-hop.
- When the **sr-isis** or **sr-ospf** value is enabled, an SR tunnel to the BGP next-hop is selected in the TTM from the lowest preference ISIS or OSPF instance. If many instances have the same lowest preference from the lowest numbered IS-IS or OSPF instance

If one or more explicit tunnel types are specified using the **resolution-filter** option, then only these tunnel types will be selected again following the TTM preference.

The user must set **resolution** to **filter** to activate the list of tunnel-types configured under **resolution-filter**.

VPN-IPv4 and VPN-IPv6 Routes

The user enables the resolution of vpn-ipv4 and vpn

-ipv6 prefixes using tunnels to MP-BGP peers with the following command:

```
configure>service>vprn>
            auto-bind-tunnel
                  resolution {any|disabled|filter}
                  resolution-filter
                  [no] gre
                  [no] ldp
                  [no] rsvp
                  [no] sr-isis
                  [no] sr-ospf
                  exit
            exit
```

The same auto-bind command is supported with BGP EVPN service:

```
configure>service>vpls>bgp-evpn>mpls>
     auto-bind-tunnel
          resolution {any|disabled|filter}
          resolution-filter
                [no] bgp
                [no] ldp
                [no] rsvp
                sr-isis
                no sr-isis
                sr-ospf
                no sr-ospf
         exit
     exit
```

The **auto-bind-tunnel** node is simply a context to configure the binding of VPRN or BGP EVPN routes to tunnels. The user must configure the **resolution** option to enable auto-bind resolution to tunnels in TTM. If the **resolution** option is explicitly set to disabled, the auto-binding to tunnel is removed.

If **resolution** is set to **any**, any supported tunnel type in VPRN or BGP EVPN context will be selected following TTM preference. If one or more explicit tunnel types are specified using the **resolution-filter** option, then only these tunnel types will be selected again following the TTM preference.

The following tunnel types are supported in a VPRN or BGP EVPN context in order of preference: RSVP, LDP, Segment Routing (SR), and GRE.

- The **rsvp** value instructs BGP to search for the best metric RSVP LSP to the address of the BGP next-hop. This address can correspond to the system interface or to another loopback used by the BGP instance on the remote node. The LSP metric is provided by MPLS in the tunnel table. In the case of multiple RSVP LSPs with the same lowest metric, BGP selects the LSP with the lowest tunnel-id.
- The **ldp** value instructs BGP to search for an LDP LSP with a FEC prefix corresponding to the address of the BGP next-hop.
- When the **sr-isis** or **sr-ospf** value is enabled, a SR tunnel to the BGP next-hop is selected in the TTM from the lowest preference ISIS or OSPF instance. If many instances have the same lowest preference from the lowest numbered IS-IS or OSPF instance.
- The **gre** value instructs BGP to use a GRE encapsulated tunnel to the address of the BGP next-hop.
- The BGP tunnel type is not explicitly configured in VPRN resolution and is thus implicit. It is always preferred over any other tunnel type enabled in the auto-bind-tunnel context. However, the BGP tunnel type is configurable as a new tunnel type for BGP EVPN prefixes. If the user does not enable the BGP tunnel type, inter-area or inter-as prefixes will not be resolved.

The user must set **resolution** to **filter** to activate the list of tunnel-types configured under **resolution-filter**.

When an explicit SDP to a BGP next-hop is configured in a VPRN or BGP EVPN service (**configure>service>vprn>spoke-sdp**), it overrides the **auto-bind-tunnel** selection for that BGP next-hop only. There is no support for reverting automatically to the **auto-bind-tunnel** selection if the explicit SDP goes down. The user must delete the explicit spoke-SDP in the VPRN or BGP EVPN service context to resume using the **auto-bind-tunnel** selection for the BGP next-hop.

BGP Flow-Spec

Flow-spec is a standardized method for using BGP to distribute traffic flow specifications (flow routes) throughout a network. A flow route carries a description of a flow in terms of packet header fields such as source IP address, destination IP address, or TCP/UDP port number and indicates (through a community attribute) an action to take on packets matching the flow. The primary application for Flow-spec is DDoS mitigation.

Flow-spec is supported for both IPv4 and IPv6. To exchange IPv4 Flow-spec routes with a BGP peer the **flow-ipv4** keyword must be part of the **family** command that applies to the session and to exchange IPv6 Flow-spec routes with a BGP peer **flow-ipv6** must be present in the **family** configuration.

The NLRI of an IPv4 flow route can contain one or more of the subcomponents shown in Table 14.

Table 14: Subcomponents of IPv4 Flow Route NLRI

Subcomponent Name [Type]	Value Encoding	SROS Support
Destination IPv4 Prefix [1]	Prefix length, prefix	Yes
Source IPv4 Prefix [2]	Prefix length, prefix	Yes
IP Protocol [3]	One or more (operator, value) pairs	Partial. No support for multiple values other than “TCP or UDP”.
Port [4]	One or more (operator, value) pairs	No
Destination Port [5]	One or more (operator, value) pairs	Partial. No support for multiple ranges.
Source Port [6]	One or more (operator, value) pairs	Partial. No support for multiple ranges.
ICMP Type [7]	One or more (operator, value) pairs	Partial. Only a single value is supported.
ICMP Code [8]	One or more (operator, value) pairs	Partial. Only a single value is supported.
TCP Flags [9]	One or more (operator, bitmask) pairs	Partial. Only SYN and ACK flags can be matched.
Packet Length [10]	One or more (operator, value) pairs	No
DSCP [11]	One or more (operator, value) pairs	Partial. Only a single value is supported.

Table 14: Subcomponents of IPv4 Flow Route NLRI (Continued)

Subcomponent Name [Type]	Value Encoding	SROS Support
Fragment [12]	One or more (operator, bitmask) pairs	Partial. No support for matching DF bit, first-fragment or last-fragment.

The NLRI of an IPv6 flow route can contain one or more of the subcomponents shown in Table 15.

Table 15: Subcomponents of IPv6 Flow Route NLRI

Subcomponent Name [Type]	Value Encoding	SROS Support
Destination IPv6 Prefix [1]	Prefix length, prefix offset, prefix	Partial. No support for prefix offset.
Source IPv6 Prefix [2]	Prefix length, prefix offset, prefix	Partial. No support for prefix offset.
Next Header [3]	One or more (operator, value) pairs	Partial. Only a single value supported.
Port [4]	One or more (operator, value) pairs	No
Destination Port [5]	One or more (operator, value) pairs	Partial. No support for multiple ranges.
Source Port [6]	One or more (operator, value) pairs	Partial. No support for multiple ranges.
ICMP Type [7]	One or more (operator, value) pairs	Partial. Only a single value is supported.
ICMP Code [8]	One or more (operator, value) pairs	Partial. Only a single value is supported.
TCP Flags [9]	One or more (operator, bitmask) pairs	Partial. Only SYN and ACK flags can be matched.
Packet Length [10]	One or more (operator, value) pairs	No
Traffic Class [11]	One or more (operator, value) pairs	Partial. Only a single value is supported.
Flow Label[13]	One or more (operator, value) pairs	No

Validating Received Flow Routes

Table 16 summarizes the actions that may be associated with an IPv4 or IPv6 flow route and how each type of action is encoded.

Table 16: IPv4 Flowspec Actions

Action	Encoding	SROS Support
Rate Limit	Extended community type 0x8006	Partial. Only rate=0 is supported.
Sample/Log	Extended community type 0x8007. S-bit	Yes
Next Entry	Extended community type 0x8007. T-bit	No
Redirect to VRF	Extended community type 0x8008.	Yes
Mark Traffic Class	Extended community type 0x8009.	No

IPv4 and IPv6 flow routes received from a BGP peer must be validated before they can be installed as filter entries. A flow route is considered invalid if:

1. The flow route is received from an EBGP peer and the left most AS number in the AS_PATH attribute does not equal the peer's AS number (from the **group**/**neighbor** configuration).
2. The **flowspec-validate** command is enabled, the flow route has a destination prefix subcomponent D, and the flow route was received from a peer that did not advertise the best route to D and all more-specific prefixes.

After received flow routes are validated they are processed by the relevant import policies, if applicable.

NOTE: A flow route never matches a prefix entry in a prefix-list, even if the destination IPv4 (or IPv6) prefix subcomponent or the source IPv4 (or IPv6) prefix subcomponent of the NLRI is a match.

Using Flow Routes to Create Dynamic Filter Entries

When the base router BGP instance receives an IPv4 or IPv6 flow route and that route is valid and best the system attempts to construct an IPv4 or IPv6 filter entry from the NLRI contents and the action(s) encoded in the UPDATE message. If successful, the filter entry is added to the system-created 'fSpec-0' IPv4 or 'fSpec-0' IPv6 filter policy. The 'fSpec-0' IPv4 filter policy is applied to the following:

- Ingress IPv4 traffic on a network interface, if its configuration includes the **flowspec** command.
- Ingress IPv4 traffic on an IES SAP interface, if its configuration includes the **flowspec** command.
- Ingress IPv4 traffic on an IES spoke SDP interface, if its configuration includes the **flowspec** command.

Similarly the ‘fSpec-0’ IPv6 filter policy is applied to the following:

- Ingress IPv6 traffic on a network interface, if its configuration includes the **flowspec-ipv6** command.
- Ingress IPv6 traffic on an IES SAP interface, if its configuration includes the **flowspec-ipv6** command.
- Ingress IPv6 traffic on an IES spoke SDP interface, if its configuration includes the **flowspec-ipv6** command.

A user-defined filter policy can be applied to a base router interface that has flow-spec enabled. When an interface has both a user-defined filter policy and the system-created ‘fSpec-0’ filter policy, the filter rules are installed in the following order:

1. User-defined filter entries
2. Flow-spec entries (in order, determined by comparison of the NLRI described in RFC 5575).
3. User-defined filter default action.

Configuration of TTL Propagation for BGP Label Routes

This feature allows the separate configuration of TTL propagation for in transit and CPM generated IP packets at the ingress LER within a BGP label route context.

TTL Propagation for RFC 3107 Label Route at Ingress LER

For IPv4 and IPv6 packets forwarded using a RFC 3107 label route in the global routing instance, including 6PE, the following command specified with the **all** value enables TTL propagation from the IP header into all labels in the transport label stack:

- config router ttl-propagate label-route-local [none | all]
- config router ttl-propagate label-route-transit [none | all]

The **none** value reverts to the default mode which disables TTL propagation from the IP header to the labels in the transport label stack.

These commands do not have a no version.

Note that the TTL of the IP packet is always propagated into the RFC 3107 label itself. The commands only control the propagation into the transport labels, for example, the labels of the RSVP or LDP LSP which the BGP label route resolves to and which are pushed on top of the BGP label.

Note that if the BGP peer advertised the implicit-null label value for the BGP label route, the TTL propagation will not follow the configuration described, but will follow the configuration to which the BGP label route resolves:

- RSVP LSP shortcut:
  → `configure router mpls shortcut-transit-ttl-propagate`
  → `configure router mpls shortcut-local-ttl-propagate`
- LDP LSP shortcut:
  → `configure router ldp shortcut-transit-ttl-propagate`
  → `configure router ldp shortcut-local-ttl-propagate`

This feature does not impact packets forwarded over BGP shortcuts. The ingress LER operates in uniform mode by default and can be changed into pipe mode using the configuration of TTL propagation for RSVP or LDP LSP shortcut.

TTL Propagation for RFC 3107 Label Routes at LSR

This feature configures the TTL propagation for transit packets at a router acting as an LSR for a BGP label route.

When an LSR swaps the BGP label for a IPv4 prefix packet, thus acting as a ABR, ASBR, or data-path Route-Reflector (RR) in the base routing instance, or swaps the BGP label for a vpn-IPv4 or vpn-IPv6 prefix packet, thus acting as an inter-AS Option B VPRN ASBR or VPRN data path Route-Reflector (RR), the all value of the following command enables TTL propagation of the decremented TTL of the swapped BGP label into all LDP or RSVP transport labels.

- `config router ttl-propagate lsr-label-route [none | all]`

Note that when an LSR swaps a label or stitches a label, it always writes the decremented TTL value into the outgoing swapped or stitched label. What the above CLI controls is whether this decremented TTL value is also propagated to the transport label stack pushed on top of the swapped or stitched label.

The **none** value reverts to the default mode which disables TTL propagation. Note this changes the existing default behavior which propagates the TTL to the transport label stack. When a customer upgrades, the new default becomes in effect. The above commands do not have a no version.

The following describes the behavior of LSR TTL propagation in a number of other use cases and indicates if the above CLI command applies or not:

1. When an LSR stitches an LDP label to a BGP label, the decremented TTL of the stitched label is propagated or not to the LDP or RSVP transport labels as per the above configuration.
2. When an LSR stitches a BGP label to an LDP label, the decremented TTL of the stitched label is automatically propagated into the RSVP label if the outgoing LDP LSP is tunneled over RSVP. This behavior is not controlled by the above CLI.
3. When a LSR pops a BGP label and forwards the packet using an IGP route (IGP route to destination of prefix wins over the BGP label route), it pushes an LDP label on the packet and the TTL behavior is like described in (2) when stitching from a BGP label to an LDP label.
4. Carrier Supporting Carrier (CsC) VPRN. The ingress CsC PE swaps the incoming eBGP label into a VPN-IPv4 label. The reverse operation is performed by the egress CsC PE. In both cases, the decremented TTL of the swapped label is propagated or not to the LDP or RSVP transport labels as per the above configuration.
5. SR OS does not support ASBR or data path RR functionality for labeled IPv6 routes in the global routing instance (6PE). As such the CLI command above has no impact on prefix packets forwarded in this context.

BGP Prefix Origin Validation

BGP prefix origin validation is a solution developed by the IETF SIDR working group for reducing the vulnerability of BGP networks to prefix mis-announcements and certain man-in-the-middle attacks. BGP has traditionally relied on a trust model where it is assumed that when a peer AS originates a route it has the right to announce the associated prefix. BGP prefix origin validation takes extra steps to ensure that the origin AS of a route is valid for the advertised prefix.

7x50 routers support BGP prefix origin validation for IPv4 and IPv6 routes received by the base router BGP instance from selected peers. When prefix origin validation is enabled on a session using the **enable-origin-validation** command every received IPv4 and/or IPv6 route received from the peer is checked to determine whether the origin AS is valid for the received prefix. The origin AS is generally the right most AS in the AS_PATH attribute and indicates the autonomous system that originated the route.

For purposes of determining the origin validation state of received BGP routes, the router maintains an Origin Validation database consisting of static and dynamic entries. Each entry is called a VRP (Validated ROA Payload) and associates a prefix (range) with an origin AS.

Static VRP entries are configured using the **static-entry** command available in the **config>router>origin-validation** context of the base router. In SR-OS, a static entry can express that a specific prefix and origin AS combination is either valid or invalid.

Dynamic VRP entries are learned from PRKI local cache servers and express valid origin AS and prefix combinations. The router communicates with RPKI local cache servers using the RPKI-RTR protocol. SR-OS supports the RPKI-RTR protocol over TCP/IPv4 or TCP/IPv6 transport; at the current time, TCP-MD5 and other forms of session security are not supported. A 7x50 router can setup an RPKI-RTR session using the base routing table or the management router.

An RPKI local cache server is one element of the larger RPKI system. The RPKI is a distributed database containing cryptographic objects relating to Internet Number resources. Local cache servers are deployed in the service provider network and retrieve digitally signed Route Origin Authorization (ROA) objects from Global RPKI servers. The local cache servers cryptographically validate the ROAs before passing the information along to the routers.

The algorithm used to determine the origin validation states of routes received over a session with **enable-origin-validation** configured uses the following definitions:

- A route is **matched** by a VRP entry if the prefix bits in the route match the prefix bits in the VRP entry (up to its min prefix length), AND the route prefix length is greater than or equal to the VRP entry min prefix length, AND the route prefix length is less than or equal to the VRP entry max prefix length, AND the origin AS of the route matches the origin AS of the VRP entry.
- A route is **covered** by a VRP entry if the prefix bits in the route match the prefix bits in the VRP entry (up to its min prefix length), AND the route prefix length is greater than or equal to the VRP entry min prefix length, AND the VRP entry type is static-valid or dynamic.

Using the above definitions, the origin validation state of a route is based on the following rules.

1. If a route is matched by at least one VRP entry, and the most specific of these matching entries includes a static-invalid entry then the origin validation state is Invalid (2).
2. If a route is matched by at least one VRP entry, and the most specific of these matching entries does not include a static-invalid entry then the origin validation state is Valid (0).
3. If a route is not matched by any VRP entry, but it is covered by at least one VRP entry then the origin validation state is Invalid (2).
4. If a route is not covered by any VRP entry then the origin validation state is Not-Found (1).

Consider the following example. Suppose the Origin Validation database has the following entries:

10.1.0.0/16-32, origin AS=5, dynamic

10.1.1.0/24-32, origin AS=4, dynamic

10.0.0.0/8-32, origin AS=5, static invalid

10.1.1.0/24-32, origin AS=4, static invalid

In this case, the origin validation state of the following routes are as indicated:

10.1.0.0/16 with AS_PATH {…5}: Valid

10.1.1.0/24 with AS_PATH {…4}: Invalid

10.2.0.0/16 with AS_PATH {…5}: Invalid

10.2.0.0/16 with AS_PATH {…6}: Not-Found

The origin validation state of a route can affect its ranking in the BGP decision process. When **origin-invalid-unusable** is configured, all routes that have an origin validation state of 'Invalid' are considered unusable by the best path selection algorithm, that is, they cannot be used for forwarding and cannot be advertised to peers.

If **origin-invalid-unusable** is not configured then routes with an origin validation state of 'Invalid' are compared to other 'usable' routes for the same prefix according to the BGP decision process.

When **compare-origin-validation-state** is configured a new step is added to the BGP decision process after removal of invalid routes and before the comparison of Local Preference. The new step compares the origin validation state, so that a route with a 'Valid' state is preferred over a route with a 'Not-Found' state, and a route with a 'Not-Found' state is preferred over a route with an 'Invalid' state assuming that these routes are considered 'usable'. The new step is skipped if the **compare-origin-validation-state** command is not configured.

Route policies can be used to attach an Origin Validation State extended community to a route received from an EBGP peer in order to convey its origin validation state to IBGP peers and save them the effort of repeating the Origin Validation database lookup. To add an Origin Validation State extended community encoding the 'Valid' result, the route policy should add a community list that contains a member in the format **ext:4300:0**. To add an Origin Validation State extended community encoding the 'Not-Found' result, the route policy should add a community list that contains a member in the format **ext:4300:1**. To add an Origin Validation State extended community encoding the 'Invalid' result, the route policy should add a community list that contains a member in the format **ext:4300:2**.

BGP Route Leaking

It is possible to leak a copy of a BGP route (including all its path attributes) from one routing instance to another in the same 7x50 system. This BGP route leaking capability applies to IPv4 and IPv6 routes (without labels). Leaking is supported from the GRT to a VPRN, from one VPRN to another VPRN and from a VPRN to the GRT. Any valid BGP route for an IPv4 or IPv6 prefix can be leaked. A BGP route does not have to be the best path or used for forwarding in the source instance in order to be leaked, but it does have to be valid (that is, the next-hop must be resolved, the AS PATH must not exhibit a loop etc.).

An IPv4 or IPv6 BGP route becomes a candidate for leaking to another instance when it is specially marked by a BGP import policy. This special marking is achieved by accepting the route with a bgp-leak action in the route policy. Routes that are candidates for leaking to other instances show a leakable flag in the output of various show router bgp commands. In order to copy a leakable BGP route received in a source instance S into the BGP RIB of a target instance T the target instance must be configured with a leak-import policy that matches and accepts the leakable route. There are separate leak-import policies for IPv4 and IPv6 routes and multiple (up to 15) leak-import policies can be chained together for more complex use cases. The leak-import policies are configured under the rib-management CLI node.

NOTE: Using a leak-import policy to change the BGP attributes of leaked route (compared to the original source copy) is NOT supported. The only attribute that can be changed is the RTM preference.

In the target instance leaked BGP routes are compared to other (leaked and non-leaked) BGP routes for the same prefix based on the complete BGP decision process, but note that leaked routes do not have information about the router ID and peer IP address of the original peer and use all-zero values for these properties.

The BGP next-hop of a leaked BGP route is always resolved in the original (source) routing instance. There is no need to leak resolving routes and tunnels into the target instance. If there is no resolving route/tunnel in the source instance then the unresolved route is not leaked. If the cost to reach the BGP next-hop in the source instance is N then this is next-hop cost used by the BGP decision process in both the source and target instances.

If a target instance has BGP multipath and ECMP enabled and some of the equal-cost best paths for a prefix are leaked routes they can be used along with non-leaked best paths as ECMP next-hops of the route.

If the original (source) routing instance has IBGP multipath and ECMP enabled and the route or tunnel that resolves the BGP next-hop of a leakable route has multiple ECMP next-hops then traffic matching the leaked route in the target instance is load-shared across the ECMP next-hops the same way as traffic matching the original route in the source instance. Note that in this case the ECMP and IBGP-multipath configurations of the target instance are effectively ignored.

When BGP fast reroute is enabled in a target instance T (for a particular IP prefix) BGP attempts to find a qualifying backup path considering both leaked and non-leaked BGP routes. The backup path criteria are unchanged by this feature – i.e. the backup path is the best path remaining after the primary paths and all paths with the same BGP next-hops as the primary paths have been removed.

A leaked BGP route can be advertised to direct BGP neighbors of the target routing instance. The BGP next-hop of a leaked route is automatically be reset to self whenever it is advertised to a peer of the target instance. Normal route advertisement rules apply, meaning that by default the leaked route is advertised if and only if (in the target instance) it is the overall best path and it is used as the active route to the destination and it is not blocked by the IBGP-to-IBGP split-horizon rule.

A BGP route leaked into a VPRN can be exported from the VPRN as a VPN-IPv4/v6 route if it matches the VRF export policy. Normal VPN export rules apply, meaning that by default the leaked route is exported if and only if (in the VPRN) it is the overall best path and it is used as the active route to the destination. Note that a leaked route cannot be exported as a VPN-IP route and then re-imported into another local VPRN.

BGP Configuration Process Overview

Figure 29 displays the process to provision basic BGP parameters.

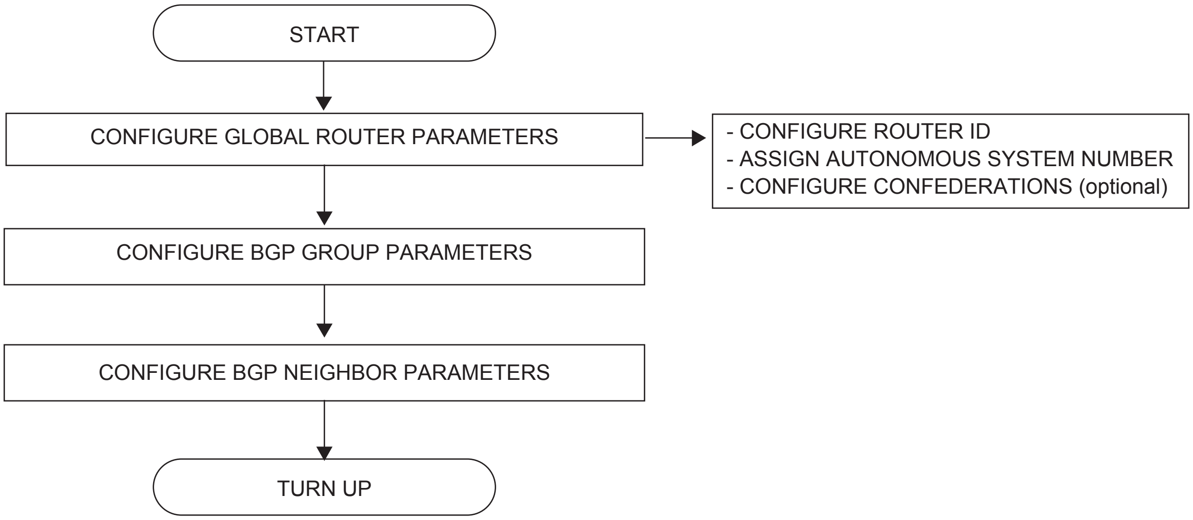
START
CONFIGURE GLOBAL ROUTER PARAMETERS
- CONFIGURE ROUTER ID
- ASSIGN AUTONOMOUS SYSTEM NUMBER
- CONFIGURE CONFEDERATIONS (optional)
CONFIGURE BGP GROUP PARAMETERS
CONFIGURE BGP NEIGHBOR PARAMETERS
TURN UP

Figure 29: BGP Configuration and Implementation Flow

Configuration Notes

This section describes BGP configuration caveats.

General

- Before BGP can be configured, the router ID and autonomous system should be configured.
- BGP must be added to the router configuration. There are no default BGP instances on a router.

BGP Defaults

The following list summarizes the BGP configuration defaults:

- By default, the router is not assigned to an AS.
- A BGP instance is created in the administratively enabled state.
- A BGP group is created in the administratively enabled state.
- A BGP neighbor is created in the administratively enabled state.
- No BGP router ID is specified. If no BGP router ID is specified, BGP uses the router system interface address.
- The router BGP timer defaults are generally the values recommended in IETF drafts and RFCs (see BGP MIB Notes on page 697)
- If no *import* route policy statements are specified, then all BGP routes are accepted.
- If no *export* route policy statements specified, then all best and used BGP routes are advertised and non-BGP routes are not advertised.

BGP MIB Notes

The router implementation of the RFC 1657 MIB variables listed in Table 17 differs from the IETF MIB specification.

Table 17: SR OS and IETF MIB Variations

MIB Variable	**Description**	**RFC 1657 Allowed Values**	SR OS **Allowed Values**
bgpPeerMinRouteAdvertisementInterval	Time interval in seconds for the MinRouteAdvertisementInterval timer. The suggested value for this timer is 30.	1 — 65535	[a]1 — 255

[a] A value of 0 is supported when the rapid-update command is applied to an address family that supports it.

If SNMP is used to set a value of X to the MIB variable in Table 18, there are three possible results:

Table 18: MIB Variable with SNMP

Condition	**Result**
X is within IETF MIB values and X is within SR OS values	SNMP set operation does not return an error MIB variable set to X
X is within IETF MIB values and X is outside SR OS values	SNMP set operation does not return an error MIB variable set to “nearest” SR OS supported value (e.g., SR OS range is 2 - 255 and X = 65535, MIB variable will be set to 255) Log message generated
X is outside IETF MIB values and X is outside SR OS values	SNMP set operation returns an error

When the value set using SNMP is within the IETF allowed values and outside the SR OS values as specified in Table 17 and Table 18, a log message is generated.
The log messages that display are similar to the following log messages:

Sample Log Message for setting bgpPeerMinRouteAdvertisementInterval to 256

```
535 2006/11/12 19:40:53 [Snmpd] BGP-4-bgpVariableRangeViolation: Trying 
to set bgpPeerMinRouteAdvInt to 256 - valid range is [2-255] - setting to 
255
```

Sample Log Message for setting bgpPeerMinRouteAdvertisementInterval to 1

```
566 2006/11/12 19:44:41 [Snmpd] BGP-4-bgpVariableRangeViolation: Trying 
to set bgpPeerMinRouteAdvInt to 1 - valid range is [2-255] - setting to 2
```