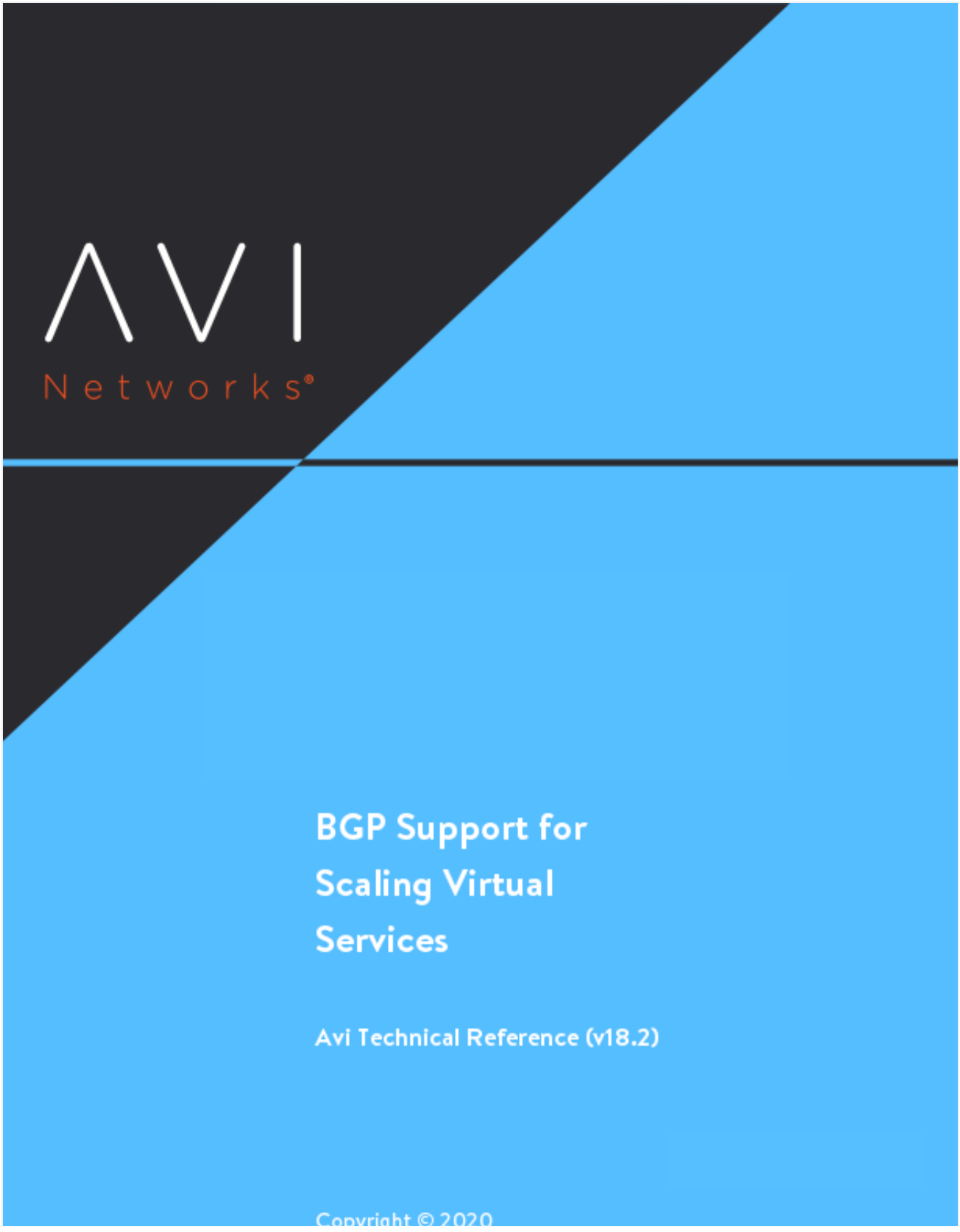


BGP Support for Scaling Virtual Services

Avi Technical Reference (v18.2)

BGP Support for Scaling Virtual Services

view online

Overview

One of the ways Avi Vantage adds load balancing capacity for a virtual service is to place the virtual service on additional Service Engines (SEs). For example, capacity can be added for a virtual service when needed by scaling out the virtual service on to additional SEs within the SE group, then removing (scaling in) the additional SEs when no longer needed. In this case, the primary SE for the virtual service coordinates distribution of the virtual service traffic among the other SEs, while also continuing to process some of the virtual service's traffic.

An alternative method for scaling a virtual service is to use a Border Gateway Protocol (BGP) feature, route health injection (RHI), along with a layer 3 routing feature, equal cost multi-path (ECMP). Using Route Health Injection (RHI) with ECMP for virtual service scaling avoids the managerial overhead placed upon the primary SE to coordinate the scaled out traffic among the SEs.

BGP is supported in legacy (active/standby) as well as elastic (active/active and N+M) high availability modes.

If a virtual service is marked down by its health monitor or for any other reason, the Avi SE withdraws the route advertisement to its virtual IP (VIP) and restores the same only when the virtual service is marked up again.

Notes on Limits

Service Engine Count

By default, Avi Vantage supports a maximum of four SEs per virtual service, and this can be increased to a maximum of 64 SEs. Each SE uses RHI to advertise a /32 host route to the virtual service's VIP address, and is able to accept the traffic. The upstream router uses ECMP to select a path to one of the SEs.

The limit on SE count is imposed by the ECMP support on the upstream router. If the router supports up to 64 equal cost routes, then a virtual service enabled for RHI can be supported on up to 64 SEs. Likewise, if the router supports a lesser number of paths, then the virtual service count enabled for RHI will be lower.

Subnets and Peers (added in 18.1.5)

Avi Vantage supports 4 distinct subnets with any number of peers in those 4 subnets. Consequently, a VIP can be advertised on more than 4 peers as long as those peers belong to 4 or less subnets. To illustrate: * A VIP can be advertised to 8 peers, all belonging to single subnet. * A VIP can be advertised to 4 pairs of peers (once again, 8 peers), with each pair belonging to separate subnet.

Supported Ecosystem

BGP-based scaling is supported on the following. * VMware * Linux server (bare-metal) cloud * OpenShift and Kubernetes Note: Peering with OpenStack routers is not supported. However, peering with an external router is possible.

BGP-based Scaling

Avi Vantage supports use of the following routing features to dynamically perform virtual service load balancing and scaling.

- Route health injection (RHI): RHI allows traffic to reach a VIP that is not in the same subnet as its SE. The Avi Service Engine (SE) where a virtual service is located advertises a host route to the VIP for that virtual service, with the SE's IP address as the next-hop router address. Based on this update, the BGP peer connected to the Avi SE updates its route

table to use the Avi SE as the next hop for reaching the VIP. The peer BGP router also advertises itself to its upstream BGP peers as a next hop for reaching the VIP.

- Equal cost multi-path (ECMP): Higher bandwidth for the VIP is provided by load sharing its traffic across multiple physical links to the SE(s). If an Avi SE has multiple links to the BGP peer, the Avi SE advertises the VIP host route on each of those links. The BGP peer router sees multiple next-hop paths to the virtual service's VIP, and uses ECMP to balance traffic across the paths. If the virtual service is scaled out to multiple Avi SEs, each SE advertises the VIP, on each of its links to the peer BGP router.

When a virtual service enabled for BGP is placed on its Avi SE, that SE establishes a BGP peer session with each of its next-hop BGP peer routers. The Avi SE then performs RHI for the virtual service's VIP, by advertising a host route (/32 network mask) to the VIP. The Avi SE sends the advertisement as a BGP route update to each of its BGP peers. When a BGP peer receives this update from the Avi SE, the peer updates its own route table with a route to the VIP that uses the SE as the next hop. Typically, the BGP peer also advertises the VIP route to its other BGP peers.

The BGP peer IP addresses, as well as the local Autonomous System (AS) number and a few other settings, are specified in a BGP profile on the Avi Controller. RHI support is disabled (default) or enabled within the individual virtual service's configuration. If an Avi SE has more than one link to the same BGP peer, this also enables ECMP support for the VIP. The Avi SE advertises a separate host route to the VIP on each of the Avi SE interfaces with the BGP peer.

If the Avi SE fails, the BGP peers withdraw the routes that were advertised to them by the Avi SE.

BGP Profile Modifications

In Avi Vantage BGP peer changes are handled gracefully as explained below.

- If a new peer is added to the BGP profile, the virtual service IP is advertised to the new BGP peer router without needing to disable/enable the virtual service.
- If a BGP peer is deleted from the BGP profile, any virtual service IPs that had been advertised to the BGP peer will be withdrawn.
- When a BGP peer IP is updated, it is handled as an add/delete of the BGP peer.

BGP Upstream Router Configuration

The BGP control plane can hog the CPU on the router in case of scale setups. Changes to CoPP policy are needed to have more BGP packets on the router, or this can lead to BGP packets getting dropped on the router when churn happens.

Note: The ECMP route group or ECMP next-hop group on the router could exhaust if the unique SE BGP next-hops advertised for different set of VS VIPs. When such exhaustion happens, the routers could fall back to a single SE next-hop causing traffic issues.

Example:

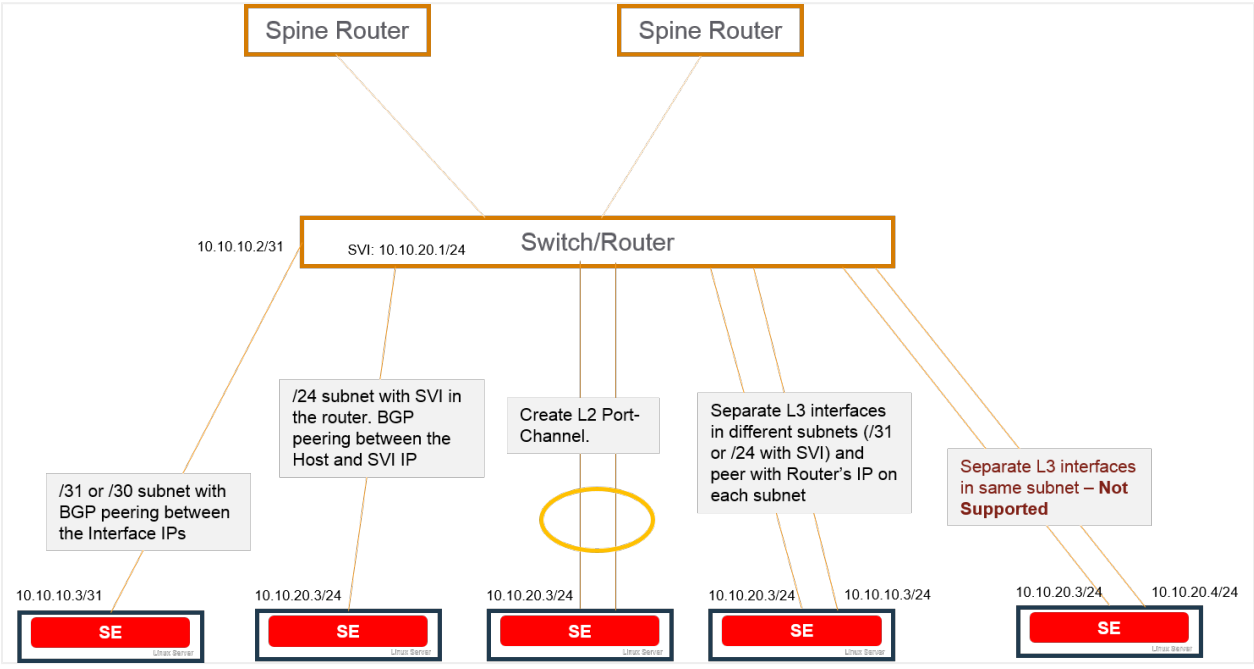
The following is the sample config on a Dell S4048 switch for adding 5k network entries and 20k paths:

```
w1g27-avi-s4048-1#show ip protocol-queue-mapping
 Protocol   Src-Port   Dst-Port   TcpFlag  Queue   EgPort     Rate (kbps)
 --------   --------   --------   -------  -----   ------     -----------
TCP (BGP)     any/179    179/any    _        Q9      _          10000
UDP (DHCP)    67/68      68/67      _        Q10     _          _
UDP (DHCP-R)  67         67         _        Q10     _          _
TCP (FTP)     any        21         _        Q6      _          _
ICMP          any        any        _        Q6      _          _
```

```
IGMP          any        any        _        Q11     _          _
TCP (MSDP)    any/639    639/any    _        Q11     _          _
UDP (NTP)     any        123        _        Q6      _          _
OSPF          any        any        _        Q9      _          _
PIM           any        any        _        Q11     _          _
UDP (RIP)     any        520        _        Q9      _          _
TCP (SSH)     any        22         _        Q6      _          _
TCP (TELNET)  any        23         _        Q6      _          _
VRRP          any        any        _        Q10     _          _
MCAST         any        any        _        Q2      _          _
w1g27-avi-s4048-1#show cpu-queue rate cp
 Service-Queue         Rate (PPS)      Burst (Packets)
 --------------        -----------      ----------
Q0                        600             512
Q1                        1000            50
Q2                        300             50
Q3                        1300            50
Q4                        2000            50
Q5                        400             50
Q6                        400             50
Q7                        400             50
Q8                        600             50
Q9                        30000           40000
Q10                       600             50
Q11                       300             50
```

SE-Router Link Types Supported with BGP

The following figure shows the types of links that are supported between Avi Vantage and BGP peer routers.



BGP is supported over the following types of links between the BGP peer and the Avi SEs.

- Host route (/30 or /31 mask length) to the VIP, with the Avi SE as the next hop.
- Network route (/24 mask length) subnet with Switched Virtual Interface (SVI) configured in the router.
- Layer 2 port channel (separate physical links configured as a single logical link on the next-hop switch or router).
- Multiple layer 3 interfaces, in separate subnets (/31 or /24 with SVI). Separate BGP peer session is set up between each Avi SE layer 3 interface and the BGP peer.

Each SE can have multiple BGP peers. For example, an SE with interfaces in separate layer 3 subnets can have a peer session with a different BGP peer on each interface. Connection between the Avi SE and the BGP peer on separate Layer 3 interfaces that are in the same subnet and same VLAN is not supported. Using multiple links to the BGP peer provides higher throughput for the VIP. The virtual service also can be scaled out for higher throughput. In either case, a separate host route to the VIP is advertised over each link to the BGP peer, with the Avi SE as the next hop address.

Note: Starting with release 18.1.2, this feature is supported for IPv6 in Avi Vantage.

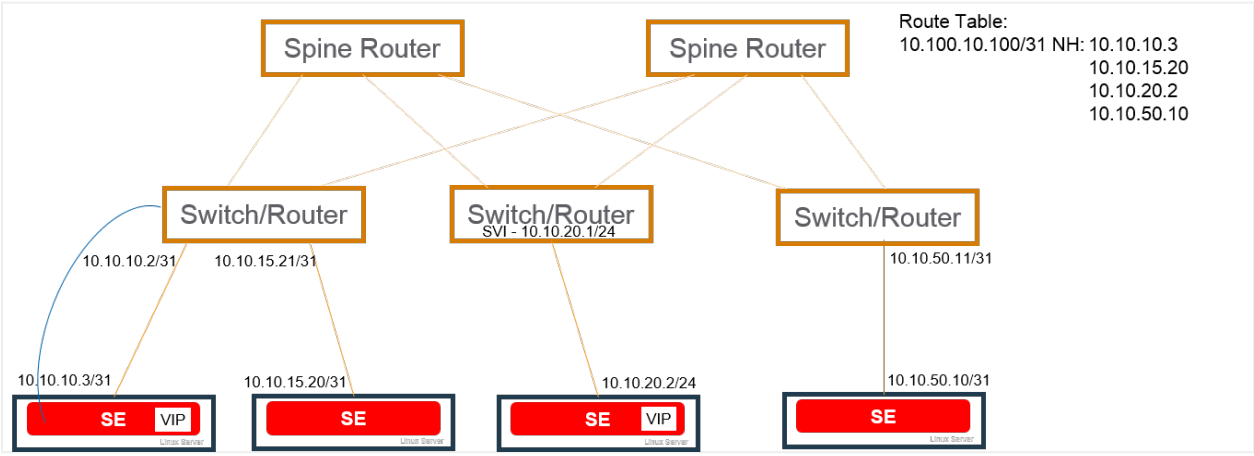
Bidirectional Forwarding Detection (BFD)

BFD is supported for fast detection of failed links. BFD enables networking peers on each end of a link to quickly detect and recover from a link failure. Typically, BFD detects and repairs a broken link faster than by waiting for BGP to detect the down link.

For example, if an Avi SE fails, BFD on the BGP peer router can quickly detect and correct the link failure.

Scaling

Scaling out/in of virtual services is supported. In this example, a virtual service placed on the Avi SE on the 10.10.10.x network is scaled out to 3 additional Avi SEs.



Flow Resiliency During Scale Out/In

A flow is a 5-tuple: src-IP, src-port, dst-IP, dst-port, and protocol. Routers do a hash of the 5-tuple to pick which equal cost path to use. When an SE scale out occurs, the router is given yet another path to use, and its hashing algorithm may make different choices, thus disrupting existing flows. To gracefully cope with this BGP-based scale-out issue, Avi Vantage supports resilient flow handling using IP-in-IP (IPIP) tunneling. The following sequence shows how this is done.

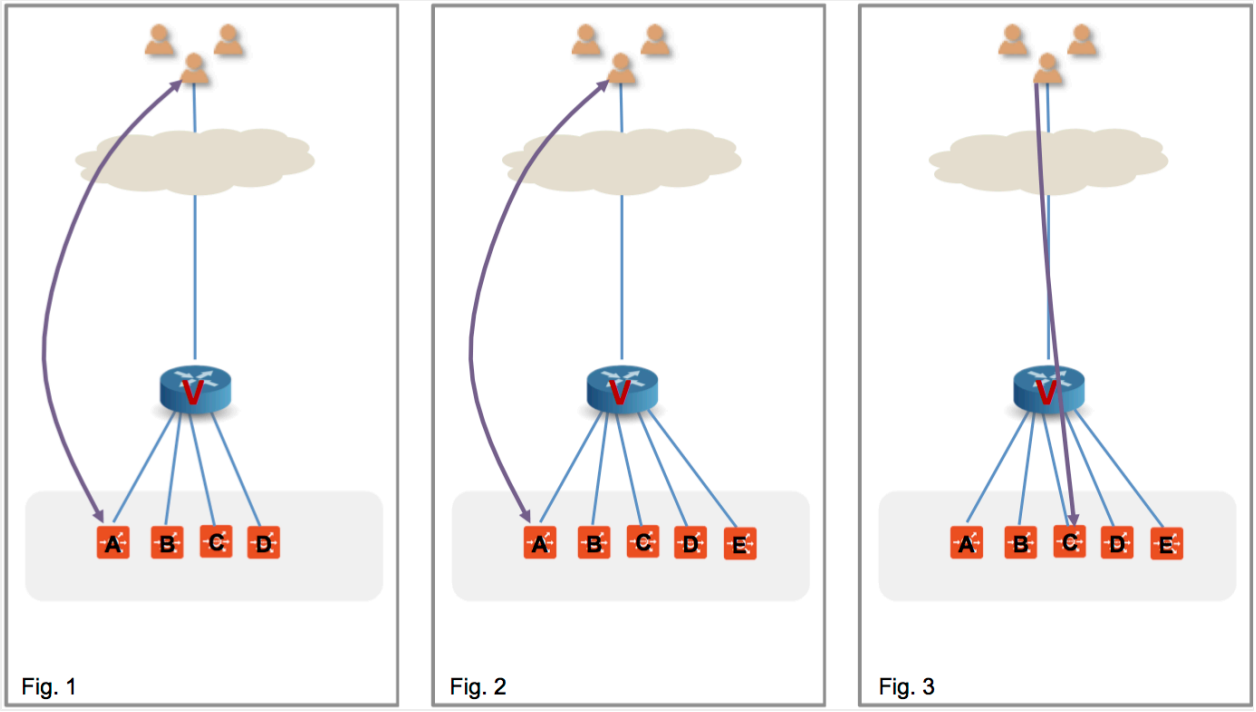
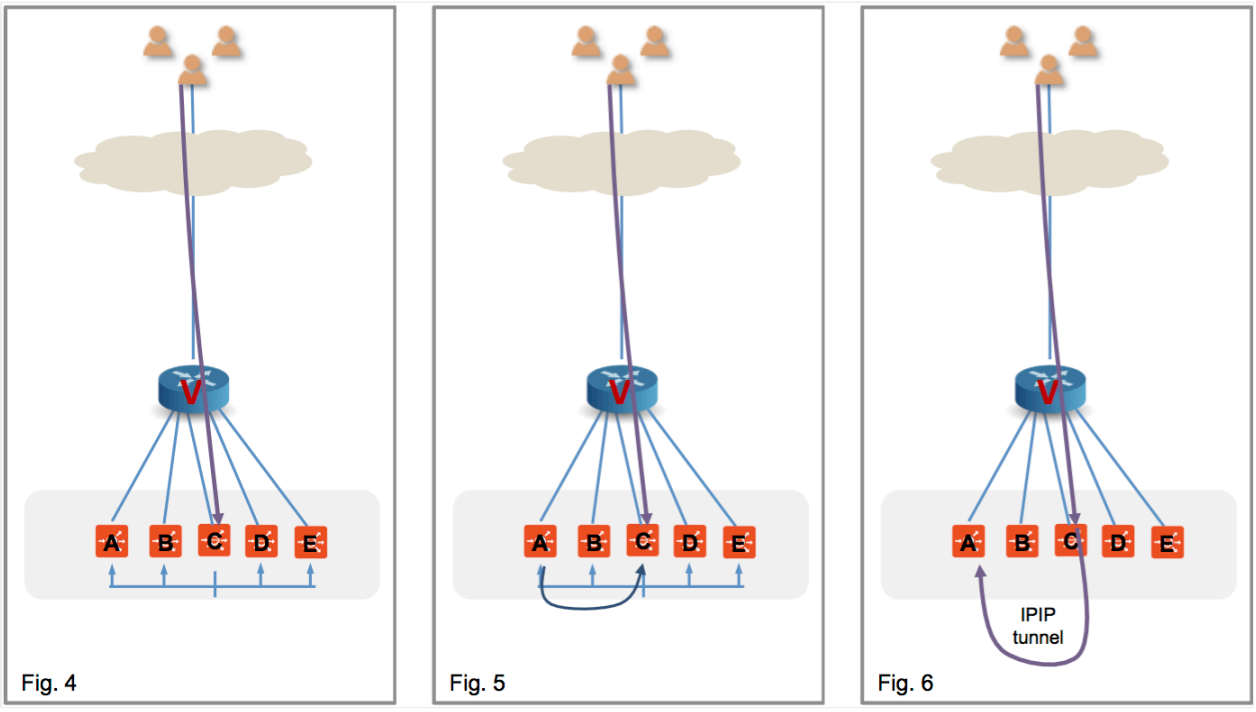


Figure 1 shows the virtual service placed on four SEs, with a flow ongoing between a client and SE-A. In figure 2, there is a scale out to SE-E. This changes the hash on the router. Existing flows get rehashed to other SEs. In this particular example, suppose it is SE-C.



In the Avi Vantage implementation SE-C sends a flow probe to all other SEs (figure 4). Figure 5 shows SE-A responding to claim ownership of the depicted flow. In figure 6, SE-C uses IPIP tunneling to send all packets of this flow to SE-A.

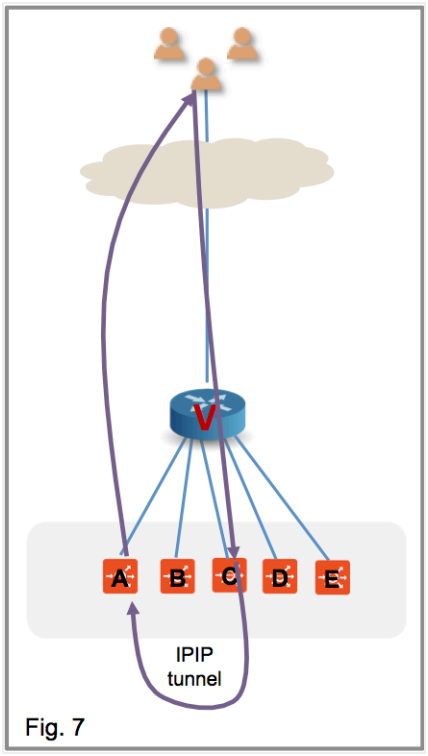


Fig. 7

In figure 7 SE-A continues to process the flow and sends its response directly to the client.

Message Digest5 (MD5) Authentication

BGP supports authentication mechanism using the Message Digest 5 (MD5) algorithm. When authentication is enabled, any TCP segment belonging to BGP exchanged between the peers, is verified and accepted only if authentication is successful. For authentication to be successful, both the peers must be configured with the same password. If authentication fails, BGP peer session will not be established. BGP authentication can be very useful because it makes it difficult for any malicious user to disrupt network routing tables.

Enabling MD5 Authentication for BGP

To enable MD5 authentication, specify `md5_secret` in the respective BGP peer configuration. MD5 support is extended to OpenShift cloud where the Service Engine runs as docker container but peers with other routers masquerading as host.

Mesos Support

BGP is supported for north-south interfaces in Mesos deployments. The SE container that is handling the virtual service will establish a BGP peer session with the BGP router configured in the BGP peering profile for the cloud. The SE then injects a /64 route (host route) to the VIP, by advertising the /64 to the BGP peer.

The following requirements apply to the BGP peer router.

- The BGP peer must allow the SE's IP interfaces and subnets in its BGP neighbor configuration. The SE will initiate the peer connection with the BGP router.

- For eBGP, the peer router will see the TTL value decremented for the BGP session. This could prevent the session from coming up. This issue can be prevented from occuring by setting the eBGP multi-hop time-to-live (TTL). For example, on Juniper routers, the eBGP multi-hop TTL must be set to 64.

To enable MD5 authentication, specify `md5_secret` in the respective BGP peer configuration. MD5 support is extended to OpenShift cloud where the Service Engine runs as docker container but peers with other routers masquerading as host.

Enabling BGP Features in Avi Vantage

Configuration of BGP features in Avi Vantage is accomplished by configuring a BGP profile, and by enabling RHI in the virtual service's configuration.

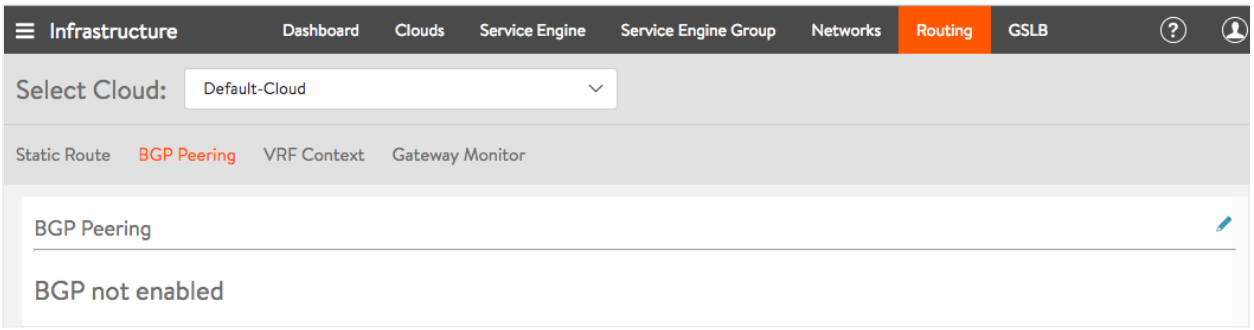
- Configure a BGP profile. The BGP profile specifies the local Autonomous System (AS) ID that the Avi SE and each of the peer BGP routers is in, and the IP address of each peer BGP router.
- Enable the Advertise VIP using BGP option on the Advanced tab of the virtual service's configuration. This option advertises a host route to the VIP address, with the Avi SE as the next hop.

Note: When BGP is configured on global VRF on LSC in-band, BGP configuration is applied on SE only when a virtual service is configured on the SE. Till then peering between SE and peer router will not happen.

Using Avi UI

To configure a BGP profile via the web interface,

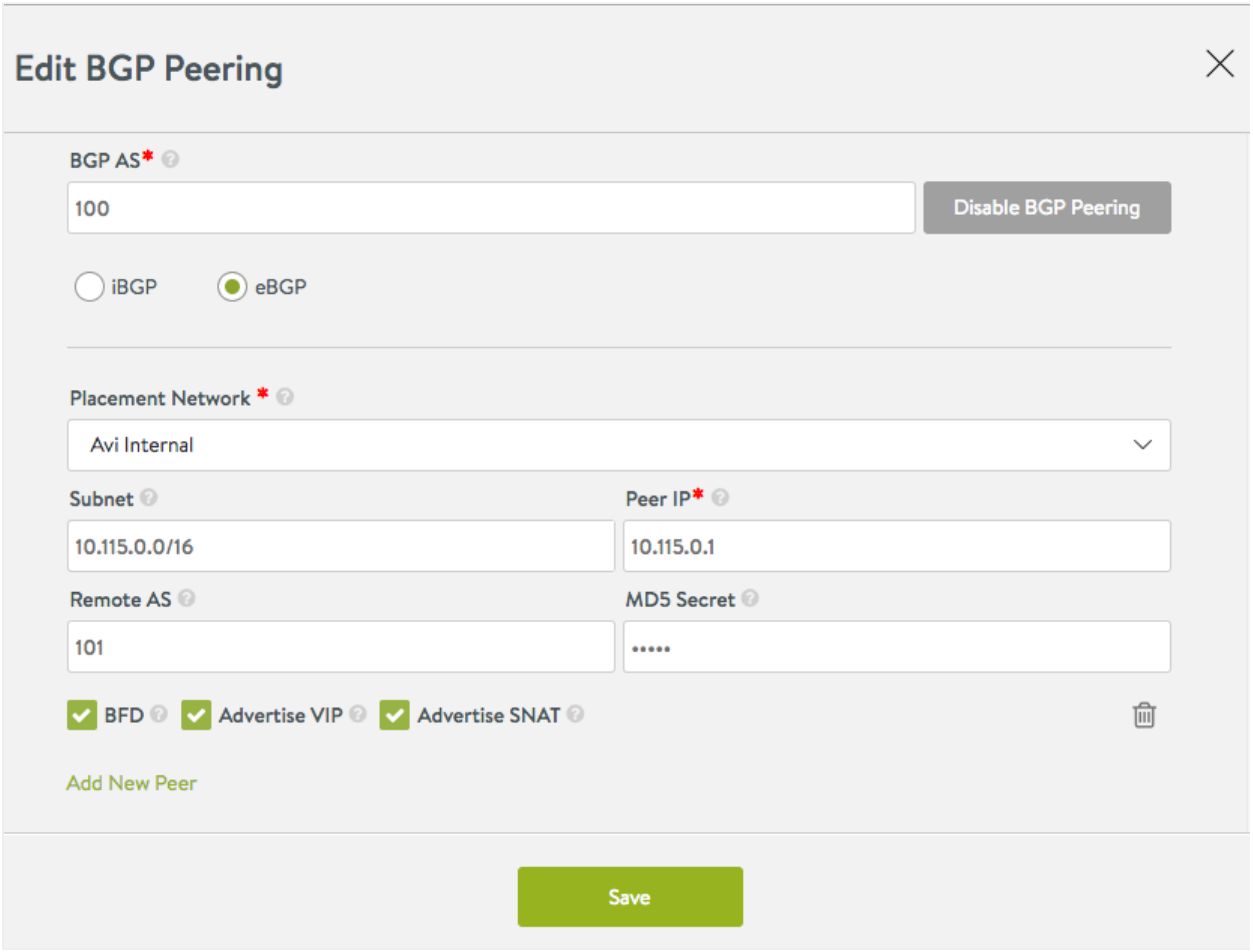
1. Navigate to Infrastructure > Routing.
2. Select the Cloud.



3. Click on the BGP Peering tab, and then click on the Edit icon to reveal more fields.
4. Enter the following information.
    - Local Autonomous System ID: a value between 1 and 4294967295
    - BGP type: iBGP or eBGP
5. Click Add New Peer to reveal a set of fields appropriate to iBGP or eBGP.
   Note: Remote AS is an additional field in eBGP. BGP peering (as eBGP) is explained below.

- SE placement network
- Subnet providing reachability for peer
- Peer BGP router's IP address
- Remote AS, a value between 1 and 4294967295, is applicable only to eBGP
- Peer Autonomous System MD5 digest secret key
- BFD option (on by default, enables very fast link failure detection via BFD, only async mode is supported)
- Advertise VIP (to this peer, on by default)
- Advertise SNAT (to this peer, on by default)

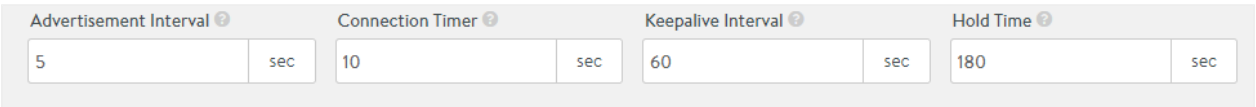
6. Click on Save. The BGP Peering screen appears as below.



Enabling BGP Timers

Starting with Avi Vantage release 18.2.5, BGP timers ? *Advertisement Interval*, *Connection Timer*, *Keepalive Interval*, and *Hold Time* can be configured using Avi UI as well.

Navigate to Infrastructure > Routing, and select BGP Peering. Enter the desired values for the timers as shown below.



CLI

The following commands configure the BGP profile. The BGP profile is included under Avi Vantage's virtual routing and forwarding (VRF) settings. BGP configuration is tenant-specific and the profile. Accordingly, sub-options appear in a suitable tenant `vrfcontext`.

```
: > configure vrfcontext management
Multiple objects found for this query.
       [0]: vrfcontext-52d6cf4f-55fa-4f32-b774-9ed53f736902#management in tenant admin, Cloud AWS-Cloud
```



```
        [1]: vrfcontext-9ff610a4-98fa-4798-8ad9-498174fef333#management in tenant admin, Cloud Default-Cloud
Select one: 1
Updating an existing object. Currently, the object is:
+---------------+-------------------------------------------------+
| Field         | Value                                           |
+---------------+-------------------------------------------------+
| uuid          | vrfcontext-9ff610a4-98fa-4798-8ad9-498174fef333 |
| name          | management                                      |
| system_default | True                                           |
| tenant_ref    | admin                                           |
| cloud_ref     | Default-Cloud                                   |
+---------------+-------------------------------------------------+
: vrfcontext > bgp_profile
: vrfcontext:bgp_profile > local_as 100
: vrfcontext:bgp_profile > ibgp
: vrfcontext:bgp_profile > peers peer_ip 10.115.0.1 subnet 10.115.0.0/16 md5_secret abcd
: vrfcontext:bgp_profile:peers > save
: vrfcontext:bgp_profile > save
: vrfcontext > save
: >
```

This profile enables iBGP with peer BGP router 10.115.0.1/16 in local AS 100. The BGP connection is secured using MD5 with shared secret "abcd."

The following commands enable RHI for a virtual service (vs-1):

```
: > configure virtualservice vs-1
: virtualservice > enable_rhi
: virtualservice > save
: >
```

The following commands enable RHI for a source-NAT'ed floating IP address for a virtual service (vs-1):

```
: > configure virtualservice vs-1
: virtualservice > enable_rhi_snat
: virtualservice > save
: >
```

The following command can be used to view the virtual service's configuration.

```
: > show virtualservice
```

Two configuration knobs have been added to configure per-peer ?advertisement-interval? and ?connect? timer in Quagga BGP:

advertisement_interval: Minimum time between advertisement runs, default = 5 seconds

connect_timer: Time due for connect timer, default = 10 seconds

Usage is illustrated in this CLI sequence:

```
[admin:controller]:> configure vrfcontext management
Multiple objects found for this query.
    [0]: vrfcontext-52d6cf4f-55fa-4f32-b774-9ed53f736902#management in tenant admin, Cloud AWS-Cloud
    [1]: vrfcontext-9ff610a4-98fa-4798-8ad9-498174fef333#management in tenant admin, Cloud Default-Cloud
Select one: 1
Updating an existing object. Currently, the object is:
+---------------+------------------------------------------------+
| Field         | Value                                          |
+---------------+------------------------------------------------+
| uuid          | vrfcontext-9ff610a4-98fa-4798-8ad9-498174fef333 |
| name          | management                                     |
| system_default | True                                          |
| tenant_ref    | admin                                          |
| cloud_ref     | Default-Cloud                                  |
+---------------+------------------------------------------------+
[admin:controller]: vrfcontext> bgp_profile
[admin:controller]: vrfcontext:bgp_profile> peers
New object being created
[admin:controller]: vrfcontext:bgp_profile:peers> advertisement_interval 10
Overwriting the previously entered value for advertisement_interval
[admin:controller]: vrfcontext:bgp_profile:peers> connect_timer 20
Overwriting the previously entered value for connect_timer
[admin:controller]: vrfcontext:bgp_profile:peers> save
[admin:controller]: vrfcontext:bgp_profile> save
[admin:controller]: vrfcontext> save
```

Configuration knobs have been added to configure the keepalive interval and hold timer on a global and per-peer basis:

```
[admin:controller]: > configure vrfcontext global
[admin: controller]: vrfcontext> bgp_profile
```

Overwriting the previously entered value for `keepalive_interval`:

```
[admin: controller]: vrfcontext:bgp_profile> keepalive_interval 30
```

Overwriting the previously entered value for `hold_time`:

```
[admin: controller]: vrfcontext:bgp_profile> hold_time 90
[admin: controller]: vrfcontext:bgp_profile> save
[admin:controller]: vrfcontext> save
[admin:controller]:>
```


The above commands configure the keepalive/hold timers on global basis, but those values can be overridden for a given peer using following per-peer commands. Both the global and per peer knobs have default values of 60 seconds for the keepalive timer and 180 seconds for the hold timer.

```
[admin:controller]: > configure vrfcontext global
[admin: controller]: vrfcontext> bgp_profile
[admin: controller]: vrfcontext:bgp_profile> peers index 1
```

Overwriting the previously entered value for `keepalive_interval`:

```
[admin: controller]: vrfcontext:bgp_profile:peers> keepalive_interval 10
```

Overwriting the previously entered value for `hold_time`:

```
[admin: controller]: vrfcontext:bgp_profile:peers> hold_time 30
[admin:controller]: vrfcontext:bgp_profile:peers> save
[admin:controller]: vrfcontext:bgp_profile> save
[admin:controller]: vrfcontext> save
```

Example

The following is an example of router configuration when the BGP peer is FRR:

You need to find the interface information of the SE which is peering with the router.

```
[admin-ctlr1]: > show serviceengine 10.79.170.52 interface summary | grep ip_addr
|     ip_addr          | fe80:1::250:56ff:fe91:1bed   |
|     ip_addr          | 10.64.59.48                  |
|     ip_addr          | fe80:2::250:56ff:fe91:b2     |
|     ip_addr          | 10.115.10.45                 |
```

Here 10.115.10.45 matches the subnet in the peer configuration in `vrfcontext->bgp_profile` object.

In the FRR router, the CLI is as follows:

```
# vtysh
Hello, this is FRRouting (version 7.2.1).
Copyright 1996-2005 Kunihiro Ishiguro, et al.

frr1# configure t
frr1(config)# router bgp 100
frr1(config-router)# neighbor 10.115.10.45 remote-as 100
frr1(config-router)# neighbor 10.115.10.45 password abcd
frr1(config-router)# end
frr1#
```

You need to perform this for all the SEs that will be peering.

Enable Gratuitous ARP

Starting with Avi Vantage version 18.2.3, you can enable gratuitous ARP for the virtual service allocated via BGP. This feature is enabled at the Service Engine group level as shown below:

```
[admin:controller]: > configure serviceenginegroup se_group_test
[admin:controller]: serviceenginegroup> enable_gratarp_permanent
```

Additional Recommended Reading

- Multihop BGP is supported in Avi Vantage. Click here to know more.
- Configuring BGP Graceful Restart.