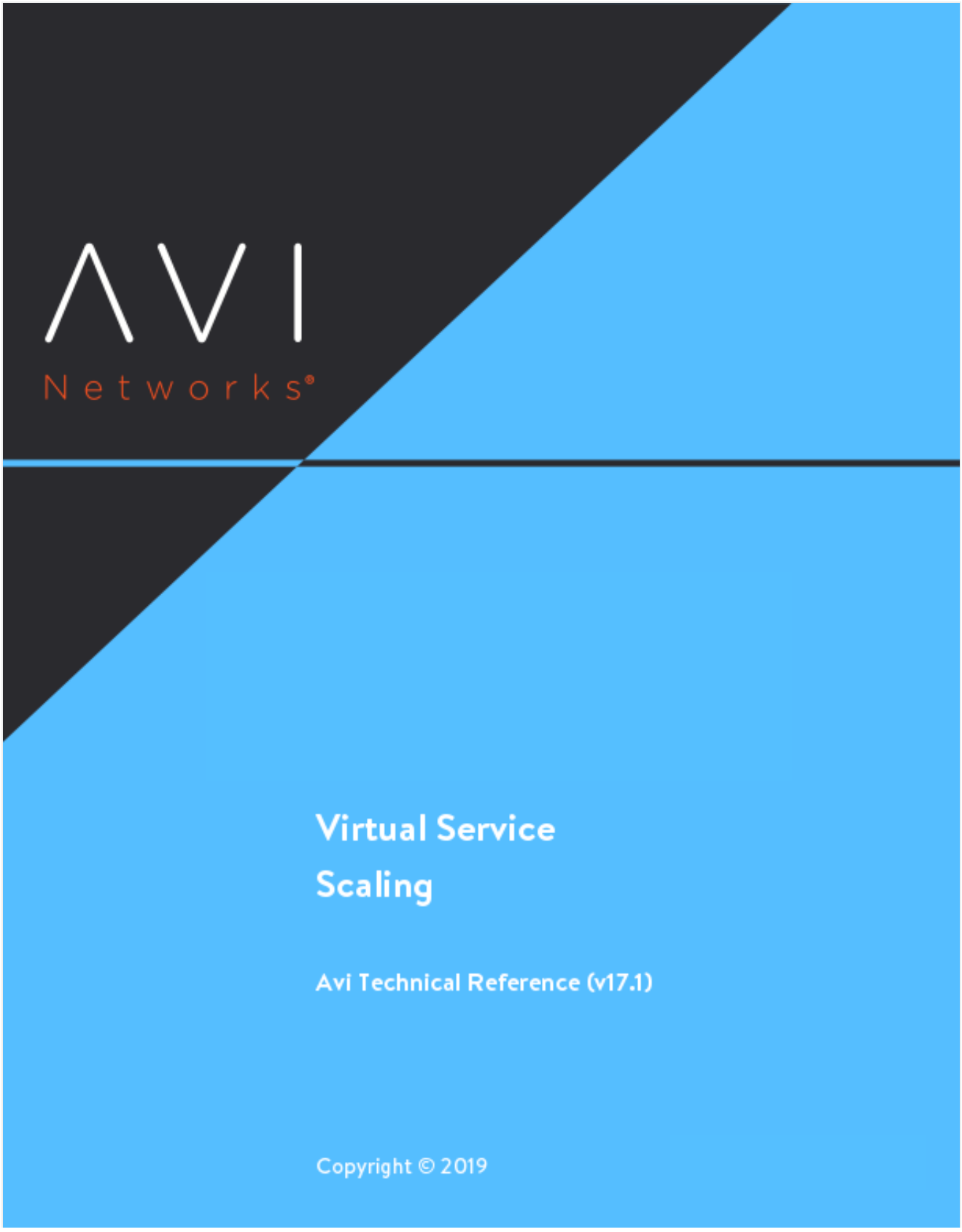


Virtual Service Scaling

Avi Technical Reference (v17.1)

Virtual Service Scaling

view online

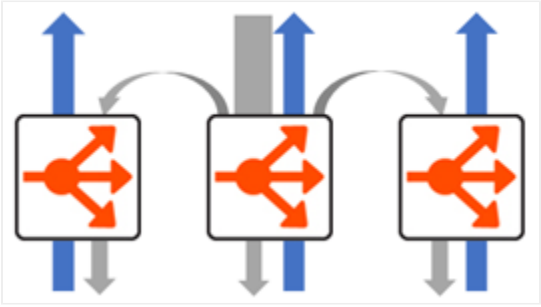
This article covers the following virtual service optimization topics:

- Scaling out a virtual service to an additional Avi Service Engine (SE)
- Scaling in a virtual service back to fewer SEs
- Migrating a virtual service from one SE to another SE

Avi Vantage supports scaling virtual services, which distributes the virtual service workload across multiple SEs to provide increased capacity on demand, thus extending the throughput capacity of the virtual service and increasing the level of high availability.

- Scaling out a virtual service distributes that virtual service to an additional SE. By default, Avi Vantage supports a maximum of four SEs per virtual service when native load balancing of SEs is in play. In BGP environments the maximum can be increased to 32.
- Scaling in a virtual service reduces the number of SEs over which its load is distributed. A virtual service will always require a minimum of one SE.

Operational Notes



Native load balancing of SEs. The primary SE (in the middle) is flanked by two secondary SEs.

This section provides additional information for specific infrastructures.

How Scaling Operates in VMware / OpenStack with Nuage Deployments

For VMware deployments and OpenStack deployments with Nuage, the scaled out traffic behaves as follows:

- The virtual service IP is GARPed by the primary SE. All inbound traffic from clients will arrive to this SE.
- The primary SE may handle a percentage of traffic, which is handled normally.
- Excess traffic is forwarded at layer 2 to the MAC address of additional secondary Service Engine(s).
- The scaled-out traffic to the secondary SEs is processed as normal. The SEs will change the source IP address of the connection to their own IP address within the server network.
- The servers will respond back to the source IP address of the traffic, which may the primary or one of the secondary SEs.
- Secondary SEs forward their response traffic directly back to the origin client, bypassing the primary SE.

How Scaling Operates in OpenStack with Neutron Deployments

For OpenStack deployments with native Neutron, server response traffic sent to the secondary SEs will be forwarded back to and through the primary SE before returning to the origin client.

Avi Vantage will issue an Alert if the average CPU utilization of an SE exceeds the designated limit during a five-minute polling period. Alerts for additional thresholds can be configured for a virtual service. The process of scaling in or scaling out must be initiated by an administrator. The CPU Threshold field of the SE Group > High Availability tab defines the minimum and maximum CPU percentages.

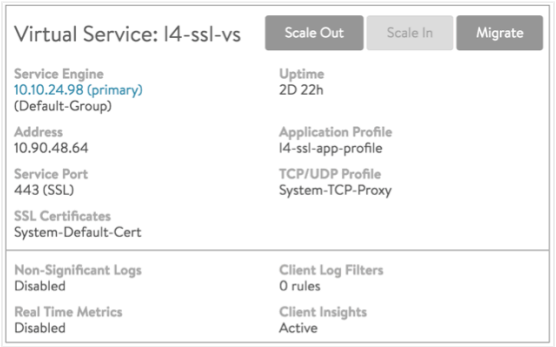
Scaling Process

The process used to scale out will depend on the level of access, write access or read/no Access, that Avi Vantage has to the hypervisor orchestrator:

- If Avi Vantage is in Write Access mode with write privileges to the virtualization orchestrator, then Avi Vantage is able to automatically create additional Service Engines when required to share the load. If the Controller runs into an issue when creating a new Service Engine, it will wait a few minutes and then try again on a different host. With native load balancing of SEs in play, the original Service Engine (primary SE) owns and ARPs for the virtual service IP address to process as much traffic as it can. Some percentage of traffic arriving to it will be forwarded via layer 2 to the additional (secondary) Service Engines. When traffic decreases, the virtual service automatically scales in back to the original, primary Service Engine.
- If Avi Vantage is in Read Access or No Access mode, an administrator must manually create and configure new Service Engines in the virtualization orchestrator. The virtual service can only be scaled out once the Service Engine is both properly configured for the network and connected to the Avi Vantage Controller. >

  Note: Existing Service Engines with spare capacity and appropriate network settings may be used for the scale out; otherwise, scaling out may require either modifying existing Service Engines or creating new Service Engines.

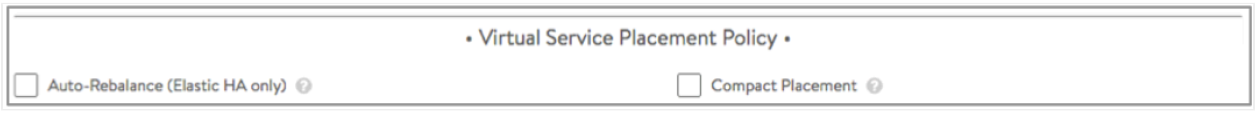
Manual Scaling of Virtual Services



Virtual services inherit from their SE group the values for the minimum and maximum number of SEs on which they can be instantiated. [Note: A virtual service's maximum instantiation count may be well below the maximum number of SEs in its group.] Between the VS min/max values, the user can manually scale the virtual service out or in from the UI, CLI, or REST API. Also, current VS instantiations on SEs can be migrated to other SEs with the same SE group. The mouse-over popup at right shows how these three actions can be accomplished from within the UI.

> *Note: For information related to the SE group settings* `min_scaleout_per_vs` *and* `max_scaleout_per_vs`, *refer to* Impact of Changes to Min/Max Scaleout per Virtual Service.

Automatic Scaling of Virtual Services



Virtual services likewise inherit from their SE group the value set for automatic rebalancing of VS instantiations. [Note: Auto-rebalancing applies only if elastic HA has been selected for the SE group]. As shown above, this setting can be checked in the Virtual Service Placement Policy section of the SE group editor. With auto-rebalance in play, and based upon the CPU utilizations of SEs with the group, Avi Vantage will migrate virtual services and -- if need be -- scale out/in the number of SEs deployed. As a result of an auto-rebalance operation, one or more virtual services in the group may be migrated to alternative SEs and/or their instantiation count adjusted to best serve the current client load.

Scaling Out

To manually scale a virtual service out when Avi Vantage is operating in Write Access mode:

1. Open the Virtual Service Details page for the virtual service that you want to scale.
2. Hover the cursor over the name of the virtual service to open the Virtual Service Quick Info popup.
3. Click the Scale Out button to scale the Virtual Service out to an additional Service Engine per click, up to a maximum of four Service Engines.
4. If available, Avi Vantage will attempt to use an existing Service Engine. If none is available or matches reachability criteria, it may create a new SE.
5. In some environments, Avi Vantage may prompt for additional information in order to create a new Service Engine, such as additional IP addresses.

The prompt "Currently scaling out" displays the progress while the operation is taking place.

> *Note: If a virtual service scales out across multiple Service Engines, then each Service Engine will independently perform server health monitoring to the pool's servers. Note: Scaling out does not interrupt existing client connections.*

Scaling out a virtual service may take anywhere from a few seconds to a few minutes. The scale out timing depends whether an additional Service Engine exists or if a new one needs to be created, as well as network and disk speeds if creating a new SE.

Scaling In

To manually scale in a virtual service in when Avi Vantage is operating in Write Access mode:

1. Open the Virtual Service Details page for the virtual service that you want to scale.
2. Hover the cursor over the name of the virtual service to open the Virtual Service Quick Info popup.
3. Click the Scale In button to open the Scale In popup window.
4. Select Service Engine to scale in. In other words, which SE should be removed from supporting this Virtual Service.
5. Scale the virtual service in by one Service Engine per SE selection, down to a minimum of one Service Engine.

The prompt "Currently scaling in" displays the progress while the operation is taking place.

> *Note: When Scaling In, existing connections are given thirty seconds to complete. Remaining connections to the SE are closed and must restart.*

Migrating

The Migrate option allows graceful migration from one Service Engine to another. During this process, the primary SE will scale out to the new SE and begin sending it new connections. After thirty seconds, the old SE will be deprovisioned from supporting the virtual service. > Note: Existing connections to the migration's source SE will be given thirty seconds to complete prior to the SE being deprovisioned for the virtual service. Remaining connections to the SE are closed and must restart.

How different Scaling Methods works

ARP tables are maintained for scaled out virtual service configuration, which is relevant for VIP scale-out scenarios only, i.e., a single VIP across multiple Service Engines

In L2 scale-out mode, the primary always responds to the ARP for the VIP. It then sends out a part of the traffic to the secondary SEs. The return traffic can go directly from the secondary SEs (DSR mode) or via the primary SE (Tunnel mode) In case of Tunnel mode, the MAC-VIP mapping is always unique. The VIP is always mapped to the primary SE'

In the DSR mode, the return traffic will use VIP as the source IP and the secondary SE's MAC as the source MAC. The ?ARP Inspection? must be disabled in the network, i.e., the network layer should not inspect/block/learn the MAC of the VIP from these packets. Otherwise MAC-IP mapping will flap. This is a case with a few environments, such as OpenStack, Cisco ACI, etc and tunnel mode is required in these environments.

In the L3 scale-out with BGP, this is not applicable since the ARP is done for the next-hop, which is the upstream router, which in turn does the ECMP to individual SEs. The return traffic uses respective SE's MAC as source MAC and VIP as source IP. The router handles this as expected.