Autoscaling in Public Clouds

Avi Technical Reference (v18.1)

Autoscaling in Public Clouds

view online

This article explains the various autoscaling capabilities provided by Avi Vantage and their integration with public cloud ecosystems, such as Amazon Web Services (AWS) and Microsoft Azure.

Overview

Avi Vantage is an elastic fabric architecture. Various resources, such as the Service Engines and application servers, can be scaled up and down on demand, based on load and capacity requirements.

For the public cloud ecosystems which can provide elastic autoscaling capabilities for workloads, Avi Vantage uses these capabilities and even manages their behaviour based on the metrics collected by Avi Vantage.

Avi Vantage provides the following scaling functionality:

- Scaling the virtual service to more (or fewer) Service Engines, so that traffic can be serviced by more (or fewer) load-balancing instances as the Avi Service Engines reach (underutilize) capacity.
- Scaling the application server pool to more (or fewer) application instances, so that traffic can be serviced by a right-sized back-end pool.

Both types of scaling can be performed automatically via pre-set Avi Vantage policies, based on load and capacity measurements done via Avi Vantage.

Ecosystem Integration

Avi Vantage supports the above-mentioned autoscaling features in all ecosystems. This article discusses integration considerations related to the below public clouds:

- Amazon Web Services
- Microsoft Azure

Virtual Service Scaling

Each Service Engine has a maximum capacity for processing traffic (typically measured in terms of traffic throughput or SSL transactions per second). The SE capacity is a function of various parameters, such as SE VM size (number of vCPUs, or memory), type of traffic, and the ecosystem in which the Service Engine is functioning.

In the default configuration, a virtual service is placed on a single Service Engine. However, if the Service Engine is not sufficient to handle traffic for the virtual service, the virtual service can be scaled out to additional SEs. In this case, more than one SE handles traffic for the virtual service.

Scaling out or scaling in of virtual services can be performed manually or automatically.

In the case of automated scaling of VS placements, one of the following Service Engine parameters can be used to configure thresholds beyond which a virtual service should be scaled out to a new SE, or scaled back into fewer SEs:

- CPU utilization of the Service Engine
- Bandwidth, in Mbps, being served by the Service Engine
- Connections per second (CPS) being served by the Service Engine
- Packets per second (PPS)

For more information on virtual service scaling, refer to Virtual Service Scaling.

Application Server Scaling

Along with the virtual service load balancing, it is important to ensure enough capacity is available at the application instance tier to handle traffic loads.

As public cloud infrastructure is charged based on usage or uptime, it is important to have enough capacity based on usage, along with the ability to scale resources on-demand.

Public clouds provide autoscaling features. The templates for autoscaling servers can be used to spawn virtual machines and configure them. The scale out or scale in can either be done manually or based on certain load conditions.

The relevant features are: * Amazon Web Services: Amazon EC2 Auto Scaling * Microsoft Azure: Virtual Machine Scale Set

Avi Vantage Integration with Public Cloud

There are two variations of Avi Vantage support for autoscaling groups which are as follows: * Autoscale decision managed by public cloud * Autoscale decision managed by Avi Vantage

Autoscale Decision Managed by Public Cloud

In this method of autoscaling, the appropriate autoscaling group is added to the server pool on an Avi Controller. The Avi Controller tracks the autoscaling group. As VM instances are added or removed from the group, Avi Vantage adds or removes the VM from its pool member list.

In this manner, Avi Vantage distributes traffic requests to the requisite VM instances.

The scaling in or scaling out of the pool is controlled based on policies associated with the autoscale group, and Avi Controller does not influence this operation.

Autoscale Decision Managed by Avi Vantage

In this method of autoscaling, Avi Vantage takes over the decision to scale the VM instances.

In this method also, the public cloud autoscale group is added to Avi server pool.

Also, an autoscale policy is created on the Avi Controller and is associated with the pool.

This autoscale policy contains parameters and thresholds for triggering the scale-out and scale-in event, based on a wide range of metrics and alerts that Avi Vantage supports.

When the threshold is crossed, the Avi Controller communicates with the public cloud to initiate a scale-out or a scale-in operation and also manages the pool membership.

A key advantage of this method is the ability to use a much richer set of metrics for performing scaling decisions, as compared to the metrics available with the public cloud.

Configuring Autoscale Integration with Avi Vantage

For more details on configuring autoscale groups with public clouds, refer to the following articles :

- Amazon Web Services (autoscaling managed by public cloud): Avi Integration with AWS Auto Scaling Groups

- Amazon Web Services (autoscale managed by Avi Vantage): Configuration and Metrics Collection on Avi Vantage for AWS Server Autoscaling

- Microsoft Azure: Virtual Machine Scale Set integration with Avi Vantage