One of the core Amazon® Web Services (AWS®) values is the ability to scale your applications dynamically as traffic demands dictate. AWS Elastic Load Balancing (AWS ELB) and AWS Auto Scaling allow you to add new resources, such as a web server, to your AWS VPC based on observed, increased traffic.

**Auto Scaling the VM-Series on AWS**

In order to ensure that security scales in a manner that is commensurate with the increased workloads, an external and internal load balancer are used to create a load balancer “sandwich.” By using two load balancers in this manner, the workload Auto Scaling Group and the security Auto Scaling Group can scale independently of each other based on resources demands.

Through integration with AWS services, like Auto Scaling and AWS ELB, you can now build a next-generation security infrastructure that will dynamically, yet independently, scale to protect your AWS workloads as their traffic patterns fluctuate. This architecture will allow you to reduce costs by utilizing the firewall capacities intelligently and efficiently based on user-defined metrics. By scaling the security separately from the application workloads, this solution allows each firewall to be configured identically and managed centrally, resulting in lower administrative costs.

**Image 1:** Elastic Load Balancing “sandwich” allows two tiers to scale independently
Native Integration Enables Scalable Security

Auto Scaling the VM-Series on AWS utilizes native AWS services and VM-Series features to deliver a next-generation firewall solution that can scale dynamically as your traffic patterns dictate. The AWS services and VM-Series features used include:

- AWS CloudFormation Template is used to deploy the entire solution from a single AWS CloudFormation template. This creates a simple-to-deploy, all-inclusive package for the solution.
- AWS Lambda is used for several predefined services: add network interfaces (ENIs) on newly deployed VM-Series instances, monitor VM-Series traffic metrics, and communicate with Amazon CloudWatch.
- AWS S3 is used to store the VM-Series bootstrap configuration and the Lambda services. S3 storage can also be used to store other types of files, such as CloudFormation templates, used for additional automation.
- AWS CloudWatch monitors your AWS workloads, collecting relevant statistics that can be used in conjunction with the VM-Series metrics to initiate the deployment or removal of a VM-Series Auto Scaling Group.
- Bootstrapping (VM-Series) allows you to create a fully configured VM-Series firewall instance. Each bootstrapped firewall can include firewall configuration, security policies, and inclusion in a Panorama™ network security management device group.
- PAN-OS® (VM-Series) API pulls user-defined metrics from the VM-Series firewall and uses Lambda to send them to CloudWatch.

How Auto Scaling the VM-Series on AWS Works

The AWS CloudFormation Template deploys an initial VM-Series firewall using a bootstrapped image stored in AWS S3. The AWS CloudFormation Template can also attach the VM-Series firewall to Panorama if it has been deployed.

When predefined web server traffic metrics are observed by CloudWatch, a scale-out event will deploy added web servers. An AWS Lambda function collects and sends VM-Series traffic metrics to CloudWatch, triggering a VM-Series scale-out event using the bootstrap VM-Series firewall image.

Image 3: Web server traffic demands initiate an AWS Auto Scaling event for the VM-Series

Summary

The VM-Series Auto Scaling on AWS solution utilizes native AWS and VM-Series services to dramatically reduce the "friction" commonly associated with deploying and configuring a next-generation firewall on AWS. As your AWS workload traffic increases, demand for security increases as well. Integration with AWS Auto Scaling and Elastic Load Balancing will allow your VM-Series next-generation firewalls to scale automatically, yet independently of your workloads, ensuring continual protection from cyberattacks.