Cloud NGFW for AWS

Disaster Recovery Guide

Table of Contents

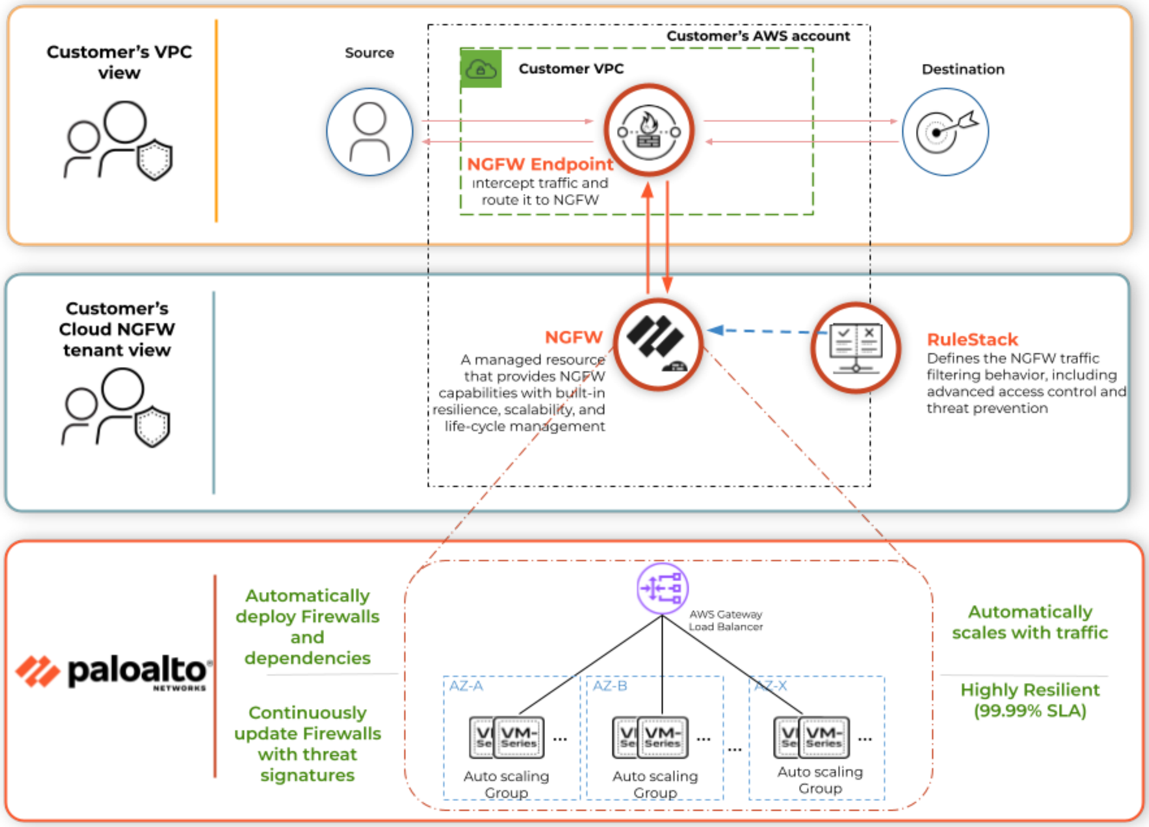
Introduction

Cloud NGFW for AWS is a regional service similar to other AWS services. This service is delivered on the AWS platform to protect your AWS Virtual network (VPC) traffic in an AWS region. A **Cloud NGFW resource** (or simply NGFW) provides next-generation firewall capabilities for your VPC. This resource has built-in resiliency, scalability, and life-cycle management. An NGFW spans multiple AWS availability zones. In AWS networking parlance, an NGFW is a VPC endpoint service. To use an NGFW resource, you create a dedicated subnet in your VPC for each desired AWS availability zone, then create NGFW endpoints on the subnets and update the VPC route tables to send the traffic through these **Cloud NGFW endpoints**. In AWS networking parlance, Cloud NGFW endpoints are Gateway Load balancer endpoints.

Under the hood, each NGFW is a Gateway Load Balancer-based VPC endpoint service managed by Palo Alto Networks, with built-in resiliency, zone-affinity, scalability, and life-cycle management. Each Cloud NGFW resource includes a dedicated set of Autoscaling Groups of Palo Alto Networks VM-Series virtual firewall (EC2) nodes. Cloud NGFW associates an ASG for each Availability Zone (AZ) the NGFW resource spans across. These Autoscaling groups (ASGs) are configured as targets to the Gateway Load Balancer (GWLB).

Cloud NGFW resource scales with your VPC traffic. The Autoscaling group provisioned for each AWS availability zone (within the Cloud NGFW resource) scales out independently and includes more VM-Series instances to inspect higher traffic volume. As the traffic volume decreases within an AWS availability zone, the corresponding Auto scaling group scales in independently.

A **Cloud NGFW rulestack** defines Cloud NGFW resource's advanced access control (App-ID, Advanced URL Filtering) and threat prevention behavior. A rulestack includes a set of security rules, associated objects, and security profiles. To use a rulestack, you associate the rulestack with one or more NGFW resources.



Customer's VPC view
Source
Customer's AWS account
Customer VPC
Destination
NGFW Endpoint
intercept traffic and route it to NGFW
Customer's Cloud NGFW tenant view
NGFW
A managed resource that provides NGFW capabilities with built-in resilience, scalability, and life-cycle management
RuleStack
Defines the NGFW traffic filtering behavior, including advanced access control and threat prevention
paloalto NETWORKS
Automatically deploy Firewalls and dependencies
Continuously update Firewalls with threat signatures
AWS Gateway Load Balancer
AZ-A
AZ-B
AZ-X
VM-Series
Auto scaling Group
Auto scaling Group
Auto scaling Group
Automatically scales with traffic
Highly Resilient (99.99% SLA)

Disaster Recovery

Disaster recovery is the process by which an organization anticipates and addresses technology-related disasters. IT systems in any company can go down unexpectedly due to unforeseen circumstances, such as power outages, natural events, or security issues. Disaster recovery includes a company's procedures and policies to recover quickly from such events.

A disaster event can take down your workload. Therefore, your objective for disaster recovery should be bringing your workload back up or avoiding downtime altogether.

You may deploy application workloads across multiple AWS availability zones and regions for global availability or for reducing Recovery Point Objectives (RPO) and Recovery Time Objectives (RTO) as part of a disaster recovery (DR) plan.   AWS recommends the following methods

1. Multi-AZ Strategy
2. Multi-Region Strategy

1. Multi-AZ strategy

Every AWS Region consists of multiple Availability Zones (AZs). Each AZ consists of one or more data centers in a separate and distinct geographic location. This AWS region architecture significantly reduces the risk of a single event impacting more than one availability zones.

You may deploy your application workloads across multiple availability zones for disaster recovery based on AWS recommendations. In this case, We have identified the following disaster events that can affect your Cloud NGFW resources. Cloud NGFW offers resiliency for these disaster events.

A. VM-Series instance (or AWS EC2) failures
B. AWS Availability zone failures

A. VM Series Instance (or AWS EC2) Failure

Cloud NGFW resource offers built-in **resiliency within an availability zone in an AWS region** by having a minimum of two VM-Series instances running simultaneously in a dedicated Auto scaling group (ASG) for high availability. Cloud NGFWs use the ASGs running VM-Series to enable resiliency for VM-Series (or EC2) instance failures. The fine-grained health check configurations enable the AWS Gateway Load Balancer to detect faults in the VM-Series instances and immediately bring up a new VM-Series instance. Since the recovery heuristic is built into the product and does not require any action from your end, Palo Alto Networks will not notify you about this event.

B. AWS Availability zone failures

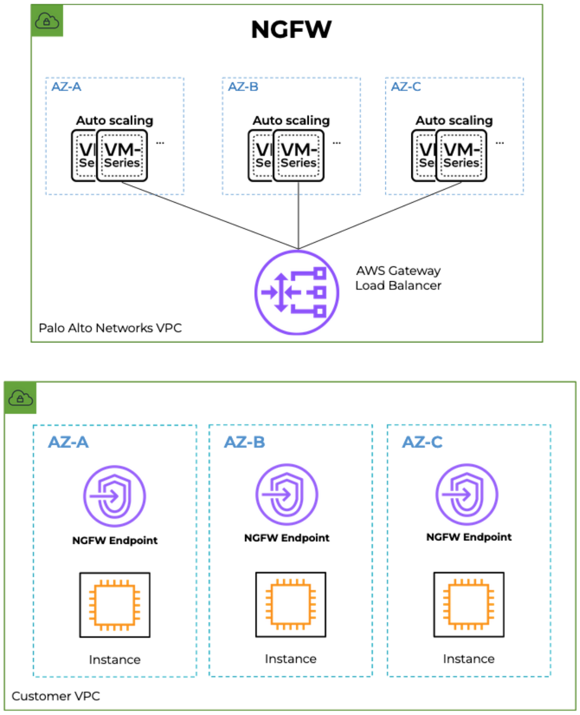
Cloud NGFW resource offers built-in **resiliency across AWS availability zones in an AWS region** by having distinct Autoscaling groups (ASGs) for each AWS availability zone that it spans across.

In a rare event of a complete Availability Zone failure, the blast radius within the Cloud NGFW resource is limited to the Auto-scaling group and VM-Series instances provisioned for that specific availability zone.

The Cloud NGFW resource remains intact and protects traffic in other AWS availability zones using the VM-Series in those zones.

Suppose the entire AWS availability zone is down. In that case, all your application workloads and VPC endpoints in that availability zone will also be down, and Cloud NGFW will receive no traffic in that availability zone.

When the AWS availability zone is back up, the Cloud NGFW resource automatically detects the change and immediately brings up the Autoscaling group and the new VM-Series instances. Since the recovery heuristic is built into the product and does not require any action from your end, Palo Alto Networks will not notify you about this event.



2. Multi-Region strategy

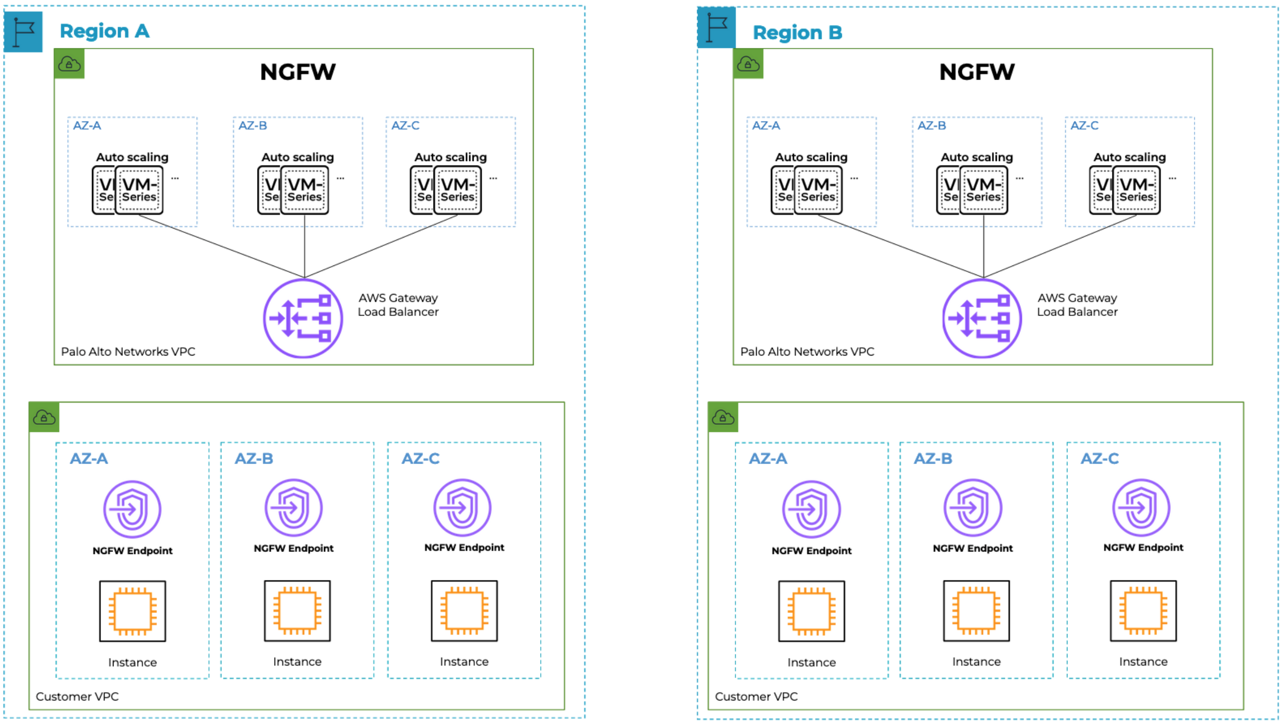
Based on AWS recommendations, you may deploy application workloads across multiple AWS regions for global availability and disaster recovery (DR) using different methods.

A. Multi-Site Active/Active
B. Warm Standby

In both these cases, you can protect the VPC traffic by deploying Cloud NGFW resource(s) in each AWS region you have the applications. You will also synchronize the rules deployed on these two resources using your preferred automation tools (such as CFT or Terraform).

In a rare event of a complete AWS regional failure, all autoscaling groups and VM-Series instances powering your Cloud NGFW resource will be down.

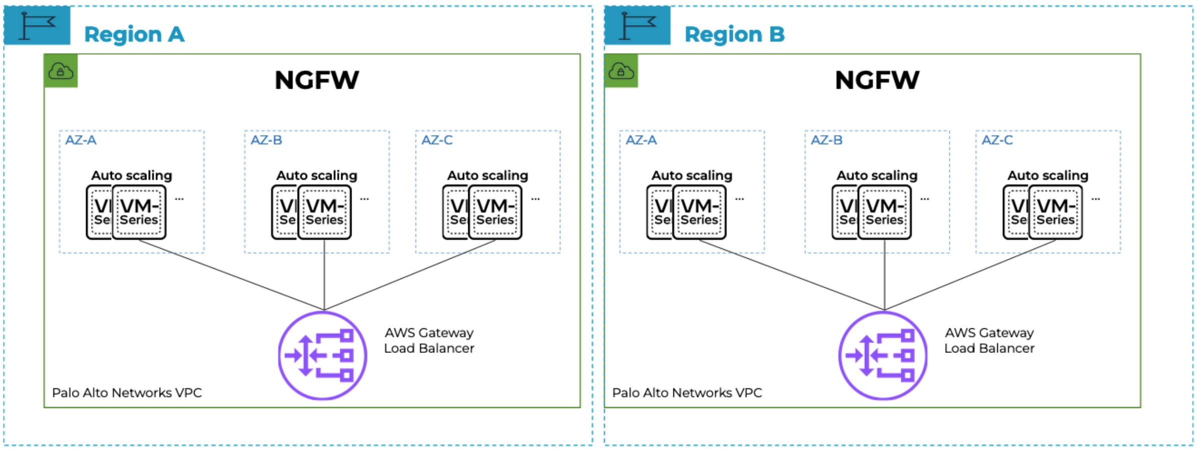
If the entire AWS region is down, all your application workloads and VPC endpoints will also be down, along with your Cloud NGFW resources. There is no traffic in the region to secure during this outage. When the AWS region is back up, Cloud NGFW resource automatically detects the change. It immediately brings up the new VM-Series instances.

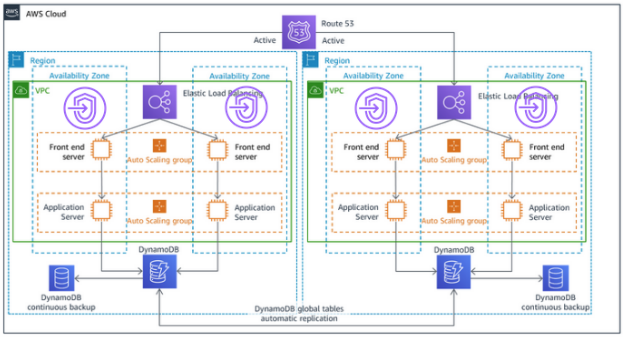


A. Multi-site active/active

With multi-site active/active, your workloads in two or more AWS regions are actively servicing requests. Failover consists of re-routing requests away from a region that cannot serve them. Here, data is replicated across AWS regions and is actively used to serve requests in those Regions.

In this architecture, you will have Cloud NGFW resources deployed in both regions to secure the regional traffic.

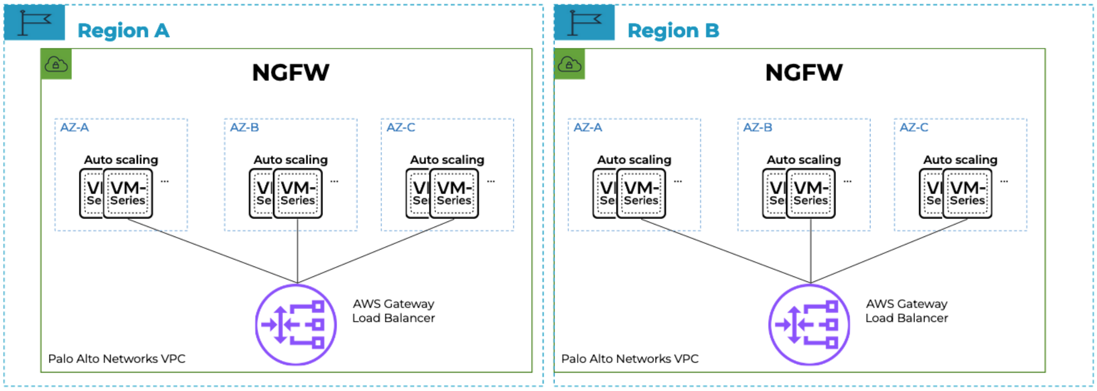


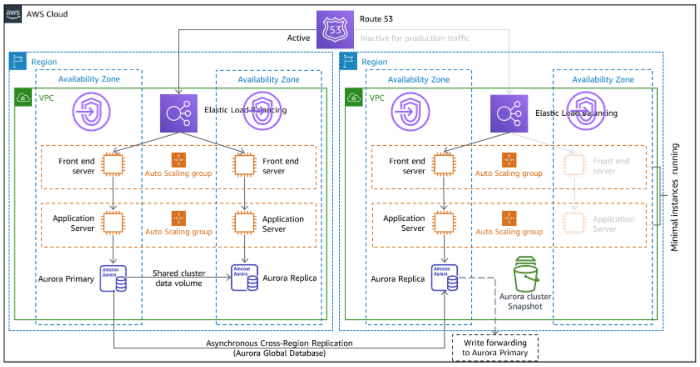


B. Warm standby

The warm standby strategy maintains data across two regions by periodic backups. A warm standby maintains a minimum deployment that can handle requests, but at a reduced capacity—it cannot handle production-level traffic.

To support this architecture, you will deploy Cloud NGFW resources in both regions to secure the regional traffic




AWS Cloud
Route 53
Active
Inactive for production traffic
Region
Availability Zone
VPC
Elastic Load Balancing
Front end server
Auto Scaling group
Application Server
Aurora Primary
Shared cluster data volume
Aurora Replica
Aurora cluster Snapshot
Minimal instances running
Asynchronous Cross-Region Replication (Aurora Global Database)
Write forwarding to Aurora Primary

Conclusion

Disaster events threaten your application workload availability. Cloud NGFW is flexible to work with the disaster recovery architectures you choose for your business needs.