SOLUTIONS GUIDE

F5 Application, Network, and Security Solutions for Amazon Web Services

aws PARTNER

Table of Contents

Table of Contents

WHY F5?

Based on data collected by various technology research companies: organizations are using two or more clouds in combination with on-premises systems to run production applications. The diversity of these environments creates multiple challenges including discrepancies in capabilities, consistency in visibility, and bifurcation of deployment processes. With the increasingly diverse array of application architectures and deployment locations combined with the need for consistent operations visibility and control, F5 is uniquely positioned to deliver value to your organization across your application, cloud, and data center estate, while accelerating the value you can derive from AWS.

Overview of F5 Offers

F5 has several offers that can be used in Amazon Web Services (AWS). There will be overlap in where and how you can use these offers in conjunction with AWS services. As we explore the different use cases, we will highlight the technical options for you. Outside of the technical options customers need to consider their own operational readiness for a technology—perhaps they have deep knowledge and experience with F5® BIG-IP® or F5 NGINX—or the delivery model (centralized via F5 Distributed Cloud vs device/instance centric with BIG-IP or NGINX). As you evaluate the use cases and need to refine which is the best option for you an F5 solutions engineer or an engineer from one of F5's partners can assist you.

Product	Commercial Models	Strengths
F5® BIG-IP®	Perpetual, Subscription, AWS Marketplace (hourly or annual)	Broadest support for application protocols, greatest flexibility for traffic controls, load balancing, security services, and access control
F5® NGINX®	Perpetual, Subscription, AWS Marketplace (hourly or annual)	Kubernetes deployments, HTTP focus, Web Application Firewall (WAF), and L7 Denial of Service (DoS)
F5® Distributed Cloud Mesh	AWS Marketplace Subscription	Multi-cloud networking, Edge proxy, Web App, and API Protection (WAAP)
F5® Distributed Cloud App Stack	AWS Marketplace Subscription	Mesh and Distributed Application Management

F5 CUSTOMERS STARTING AWS JOURNEY—SAMPLE QUESTIONS

How many virtual servers do we need to deploy?
The goal here is to understand if the current plan is for a vertically scaled, highly dense deployment. Different customers have different expectations and knowledge about the AWS environment and efforts may be required to find the right architecture vs. the existing architecture. AWS instances have limits in the number of IP addresses and network interfaces that they can have; these limits are lower than customers are used to in on-premises deployments.

Do all of these virtual servers need public IPs?
Following on the number of virtual servers we have, the number of virtual servers that need to have public IP addresses is material since it impacts the architecture. If we need many virtual IPs (VIPs), and they fit in the performance dimensions, we can look at using routed topologies to alien IP addresses if they do not need public IP addresses.

Do the applications and IT polices allow traffic to be translated between BIG-IP and the application?
If we can NAT the traffic between BIG-IP and the application servers, we have more topology and scaling options. If address translation is not allowed, it becomes a limiting factor on how the systems can be deployed into solely active-standby configurations. Consequently, you will have to bifurcate environments at scale limits since the total network and SSL capacity is much lower in AWS compared to F5 hardware systems.

How to Use This Guide

This guide provides high-level guidance on the design patterns and examples of AWS services that can be combined to build an end-to-end architecture for your application. To properly use the guide a reader needs to digest the following:

- Which software modules can be deployed in a basic pattern set of active-standby, active-active, and Amazon Auto Scaling.
- How the deployment patterns of active-standby, active-active, and Auto Scaling can help them meet the high availability (HA) and resiliency requirements of their application and organization.
- How AWS networking behavior impacts architecture options: Amazon Virtual Private Cloud (VPC) peering, AWS Transit Gateway (TGW), the placement of Amazon virtual gateways (VPN and TGW), and Internet ingress.
- Performance and scale dimensions.

Organizations can combine the different F5 offers, patterns, and capabilities to address the application services requirements they have in AWS and align them to areas that they have technical competency.

- Each organization and application may have different needs. This document provides guidance on how F5 offers can be deployed which then need to be matched to an organization's needs to operate an application.

Here is an example of questions asked of F5 customers as we begin the architecture journey in AWS.

What scale and performance metrics are known?
The scale metrics can drive the topology on if it needs to support more capacity than a single instance. Please see the section on migration for more information.

How do you connect to your Amazon VPCs from your corporate location?
How data centers are connected to VPCs is material. AWS Direct Connect and VPNs that are on TransitGateway lead to the greatest flexibility. DirectConnect or VPN to a VPC creates the least flexibility in how we can architect applications.

How do you interconnect your Amazon VPCs?
Understanding the interconnections and topologies allows different options if we need to create high density deployments or address internal applications that are not DNS enabled but we need to create availability zone (AZ) fault domain resilience.

It is highly recommended that you ground your knowledge of F5 and AWS networking with this 5-part article series. While portions of the capabilities have changed the foundations will be critical for other concepts in this document.

F5 BIG-IP Traffic Modules and the Services Presented in AWS

Module	Services	Topologies	Notes
Local Traffic Manager™ (LTM)	Traffic Management, Load Balancing, iRules, Basic DoS	Active-Standby, Active-Active and Auto Scaling	Can be integrated with GWLB
Advanced Firewall Manager™ (AFM)	L4 Firewall, IPS Services, L4 DoS, IP Intelligence, Protocol Compliance	Active-Standby, Active-Active, Auto Scaling	Can be integrated with GWLB
Advanced WAF®	Web Application Firewall, OWASP, Bot Protection, L7 DoS, HTTP Compliance, IP Intelligence, Threat Campaigns	Active-Standby, Active-Active, Auto Scaling	Can be integrated with GWLB
DNS	DNS, DNS Firewall DNSSEC, GSLB	Active-Standby, Active-Active	
Access Policy Manager® (APM)	Access, Authentication, SSL VPN	Active-Standby (single AZ), Active-Active	
SSL Orchestrator® (SSLO)	SSL Inspection, Policy-Based Security Service Chains	Active-Active, Active-Standby (Single AZ)	Can be integrated with GWLB

Application Deployment Compute Services

When deploying applications in AWS customers have an array of choices to leverage. Each one of these choices may not be deterministic and applications can consist of components that run on all these options. F5 BIG-IP, NGINX, and Distributed Cloud supports applications that are running on any of the following:

Amazon Elastic Compute Cloud (Amazon EC2): EC2 is the AWS IaaS compute service allowing customers to run virtual machines. F5 software is deployed on Amazon EC2 instances and can leverage applications running on EC2. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

Amazon Elastic Kubernetes Service (Amazon EKS): EKS is a managed Kubernetes (K8S) service from Amazon. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

Region: AWS Regions are the large data center campuses that many think of for public cloud. Currently F5 offers can be deployed in all regions globally, including AWS China and Government.

Local Zones: AWS Local Zones present as a special availability zone in a given region but are placed in close geoproximity to metro areas. F5 supports integration with the listed AWS compute services running in an AWS Local Zones with BIG-IP, NGINX, and Distributed Cloud. Local Zones can be considered a network edge application deployment.

Wavelength Zone: AWS Wavelength Zones are like Local Zones but are focused in attaching to carrier networks. F5 supports integration between the listed AWS compute services with BIG-IP, NGINX, and Distributed Cloud. Wavelength can be considered a carrier network edge application deployment.

Outposts: AWS Outposts are a method to deploy AWS compute services in hybrid and private edge use cases. F5 supports running BIG-IP, NGINX, and Distributed Cloud on AWS Outposts. Outposts can be considered an on-premises, hybrid, or edge deployment.

Amazon EKS Anywhere (EKS-A): EKS-A is a managed Kubernetes solution that can run in non-AWS environments. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

Amazon Elastic Container Service (Amazon ECS): AWS ECS allows customers to run containerized applications on AWS. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

Amazon Elastic Container Service (ECS) Anywhere (ECS-A): ECS-A is a managed container service that runs outside of an AWS data center. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

AWS Fargate: AWS Fargate is a managed container service where the container hosts are external to the customer's VPC. F5 supports integration with Amazon EKS with BIG-IP, NGINX, and Distributed Cloud.

K8S/Docker on AWS: For customers that are running their own K8S or container runtimes on top of Amazon EC2, F5 can integrate with these technologies the same manner as your on-premises environment.

F5 Distributed Cloud App Stack: Managed Kubernetes cluster.

AWS Deployment Locations for F5 BIG-IP, NGINX, and Distributed Cloud

AWS has developed an array of offers for where you can deploy their compute services. F5 is agnostic on which compute services you are using in AWS and you can leverage F5 offers to either integrate or interoperate with those services bring greater security and traffic management than would be achievable with native services alone. Depending on the intersection of the compute location and the F5 offer you may need to customize the standard template that you are using.

Resiliency and High Availability in the Public Cloud

Prior to diving into the different ways one can accomplish resiliency and high availability in the public cloud it is important that we define terms and how they impact availability. The choices in these patterns affect operations and are material based on both the technical requirements and organization comfort with the differences between them.

Highly Available: A deployment is highly available when there is one instance in one region or availability zone that can process the traffic in an active-standby manner. In a highly available architecture, users commonly deploy active-standby instances and integrate them with the cloud APIs using F5 Cloud Failover Extension. The benefit of this pattern is that it mimics what has been deployed in data centers for some time and is easy to understand. A drawback of this pattern is that traffic processing is commonly limited to a single instance.

An example of a good reason to use active-standby is that you need all traffic to be processed by one instance because you do not want to SNAT (secure network address translation) client traffic to the backend application. If the instance fails over, the secondary instance moves the necessary routes and Elastic IP (EIPs) and customer-owned IPs to it. Configurations can be managed by direct user interaction with a command line interface (CLI), traffic management user interface (TMUI) or application programming interface (API).

Resilient: A deployment is fault tolerant when there are *'n'* number of active instances to process traffic spread across availability zones and regions. In a resilient architecture, traffic is distributed across the instances with the use of Global Server Load Balancing (GSLB), AWS Global Accelerator, Amazon CloudFront, F5 Regional Edges (RE), or other service. A failure of an instance does not require any changes or API integrations. Auto Scaling deployments of BIG-IP fit into this model.

An example of a fault tolerant deployment is when you have two BIG-IP, NGINX, or Distributed Cloud instances that allow traffic in for the application and traffic is SNAT-ed to backend systems. You can add or remove instances and it does not impact backend traffic. Configurations need to be managed via API or templates.

Combining high availability and resilient patterns: It is possible to combine the two patterns, but the decision to do so needs to ensure that it accounts for costs of doing so and whether the combined pattern is justified.

In the charts below we will explore different fault scenarios in how to mitigate them. The cost listed in the TCO is only relative to the same chart.

Availability Concern—Device Failure: When we are talking about a device failure in the cloud, we could mean a failure of software (hypervisor, virtual machine [VM], or application), or a failure of an underlying server in the cloud impacting users. To address this concern, we need to deploy a resilient or highly available topology.

Deployment Type	Server Failure	AZ Failure	VPC Failure	Region Failure	Cloud Failure	TCO
Standalone	No	No	No	No	No	$
Intra-AZ HA	Yes	No	No	No	No	$$
Intra-AZ HA	Yes	Yes	No	No	No	$$

Availability Concern—AZ failure: To protect against an AZ failure, it is required that the customer deploy either active-active standalone, single active-standby inter-AZ. While the active-active standalone only impacts half of the users should an AZ fail, all users would be impacted at a VPC or Region level.

Deployment Type	Server Failure	AZ Failure	VPC Failure	Region Failure	Cloud Failure	TCO
Active-Active Standalone 1 VPC, Multi-AZ	Yes	No	No	No	No	$
Intra-AZ HA (single)	Yes	Yes	No	No	No	$$

Availability Concern—Region failure: At this point the concern is that a region failure will take the system down, and this is a concern they should have from a business continuity and disaster recovery (BCDR) perspective. Let's look at some design patterns and the different paths to mitigation.

Deployment Type	Server Failure	AZ Failure	VPC Failure	Region Failure	Cloud Failure	TCO
Active-Active or Active-Standby (Multi-AZ, Multi-VPC, Multi-Region)	Yes	Yes	Yes	Yes	No	$
Active-Active Inter-AZ HA (Multi-AZ, Multi- VPC, Multi-Region)	Yes	Yes	Yes	Yes	No	$$
Active-Active Intra-AZ HA (Multi-AZ, Multi-VPC, Multi-Region)	Yes	Yes	Yes	Yes	No	$$$

In all of the above scenarios we have to use an external mechanism such as DNS to distribute the traffic and the significantly increasing costs of using a high availability (HA) solution are diminishing in returns to the organization.

Traffic distribution for resilient patterns

As you can see above, resilient patterns require traffic to be distributed at a layer above a single BIG-IP, NGINX, and Distributed Cloud deployment. The different choices of distribution have different impacts to the traffic, supported protocols, and scope of distribution.

Availability Concern—Cloud failure: The complete failure of AWS seems like an extremely unlikely event. This pattern does get discussed from time to time and to mitigate the concern, not only does an organization need a robust automation solution, they need a robust data replication solution while leveraging cloud services to the least extent possible. It is included here to complete the picture of what the pattern and solution would look like. If your organization has this concern, it will be a mix of the different models above repeated in multiple cloud environments.

Deployment Type	Server Failure	AZ Failure	VPC Failure	Region Failure	Cloud Failure	TCO
Active-Active or Active-Standby (Multi-Cloud, Multi-AZ, Multi-VPC, Multi-Region)	Yes	Yes	Yes	Yes	No	$
Active-Active Inter-AZ HA (Multi-AZ, Multi- VPC, Multi-Region)	Yes	Yes	Yes	Yes	Yes	$$
Active-Active Intra-AZ HA (Multi-AZ, Multi-VPC, Multi-Region)	Yes	Yes	Yes	Yes	Yes	$$$

Traffic distribution for resilient patterns

As you can see above, resilient patterns require traffic to be distributed at a layer above a single BIG-IP, NGINX, and Distributed Cloud deployment. The different choices of distribution have different impacts to the traffic, supported protocols, and scope of distribution.

Mechanism	Network Protocols and Behavior	Scope	Notes
Global Server Load Balancing	All network protocols—does not require SNAT	Global, multiple IP addresses, environment agnostic	Examples include F5 DNS, and AWS Route 53. Supports Auto Scaling with F5 DNS
AWS Global Accelerators	All network protocols—does not require SNAT if traffic is on Eth0 of BIG-IP	Global, single IP address, AWS specific	Supports Auto Scaling with AWS Lambda script—BIG-IP does not need to be exposed to the Internet
AWS CloudFront	HTTP/S only—requires SNAT	Global, multiple IP addresses, environment agnostic	Supports Auto Scaling with AWS Elastic Load Balancer (ELB)—can also be a third-party content delivery network (CDN)
AWS Network Load Balancer	TCP/UDP protocols—does not require SNAT if traffic is on Eth0	Regional, multiple IP addresses, AWS specific	
F5 Regional Edges	HTTP, TCP, UDP – requires SNAT	Global	

F5 Deployment Patterns–BIG-IP

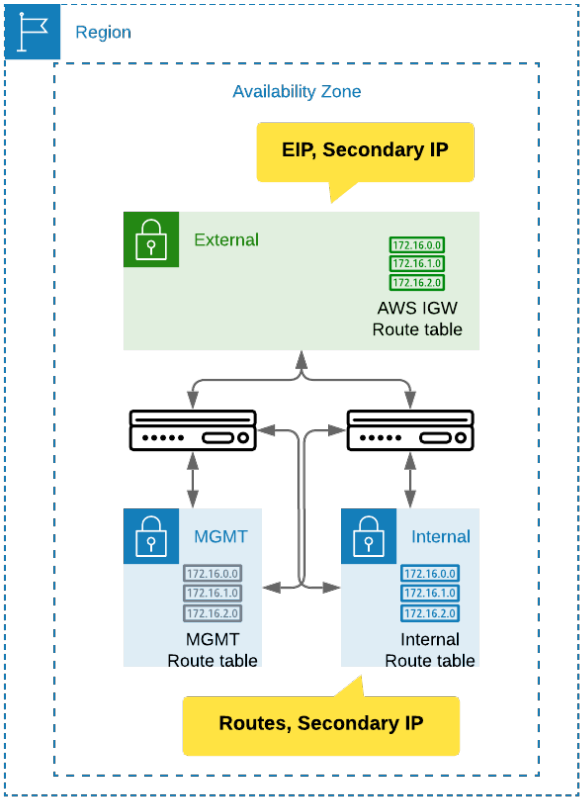
Applications and organizations vary in their architecture needs when it comes to designing application resilience. Even within a single organization there may be different needs for different applications, or stages of an application deployment. BIG-IP supports multiple patterns that can be used to meet the needs of your organization.

Active-standby

In this pattern you leverage an active and a standby instance for each traffic insertion point or scale block. Only one instance in each deployment processes traffic at a time. These deployments may be inter or intra AZ. The Active/standby pattern leverages F5 BIG-IP Cloud Failover Extension combined with IAM rules and AWS environmental tags to manage the mapping of Elastic IPs (EIP), customer owned IP addresses, Secondary IPs addresses, and routes to the active instance in the cluster.

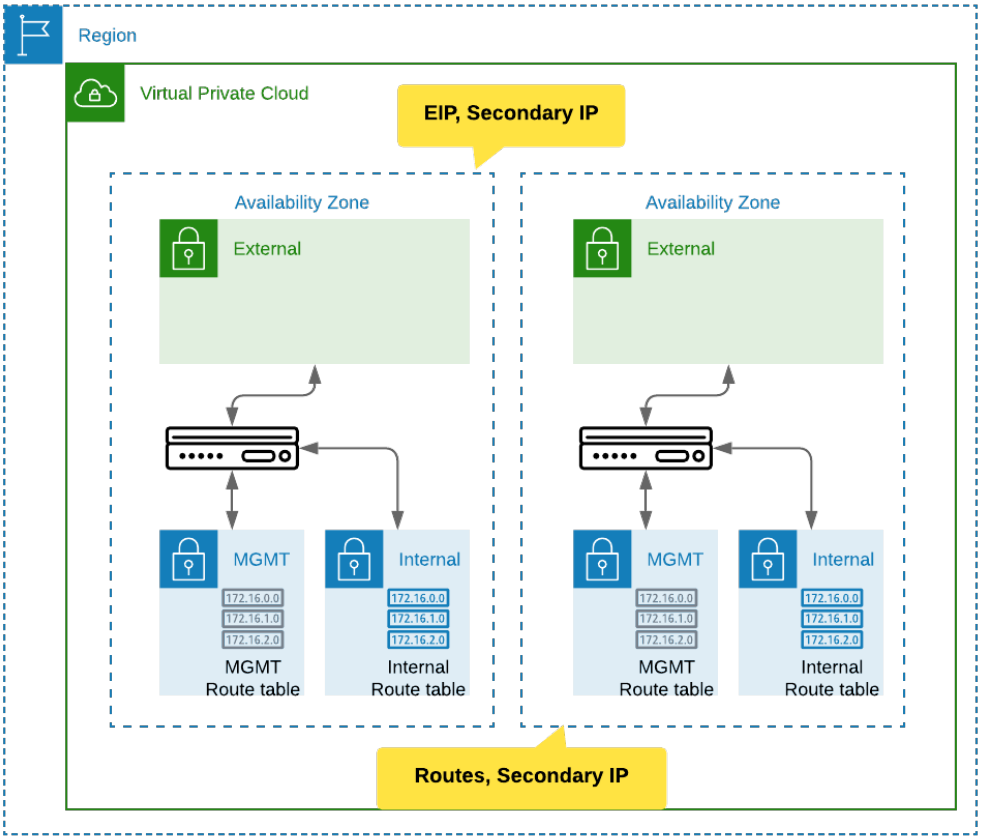
Active-standby single AZ

This creates high availability in a single AZ, but does not protect you from the failure of the AZ or a failure of a subcomponent in an AZ—the VMs could be on the same server, on servers in the same rack, or in racks in the same subsection of the facility.



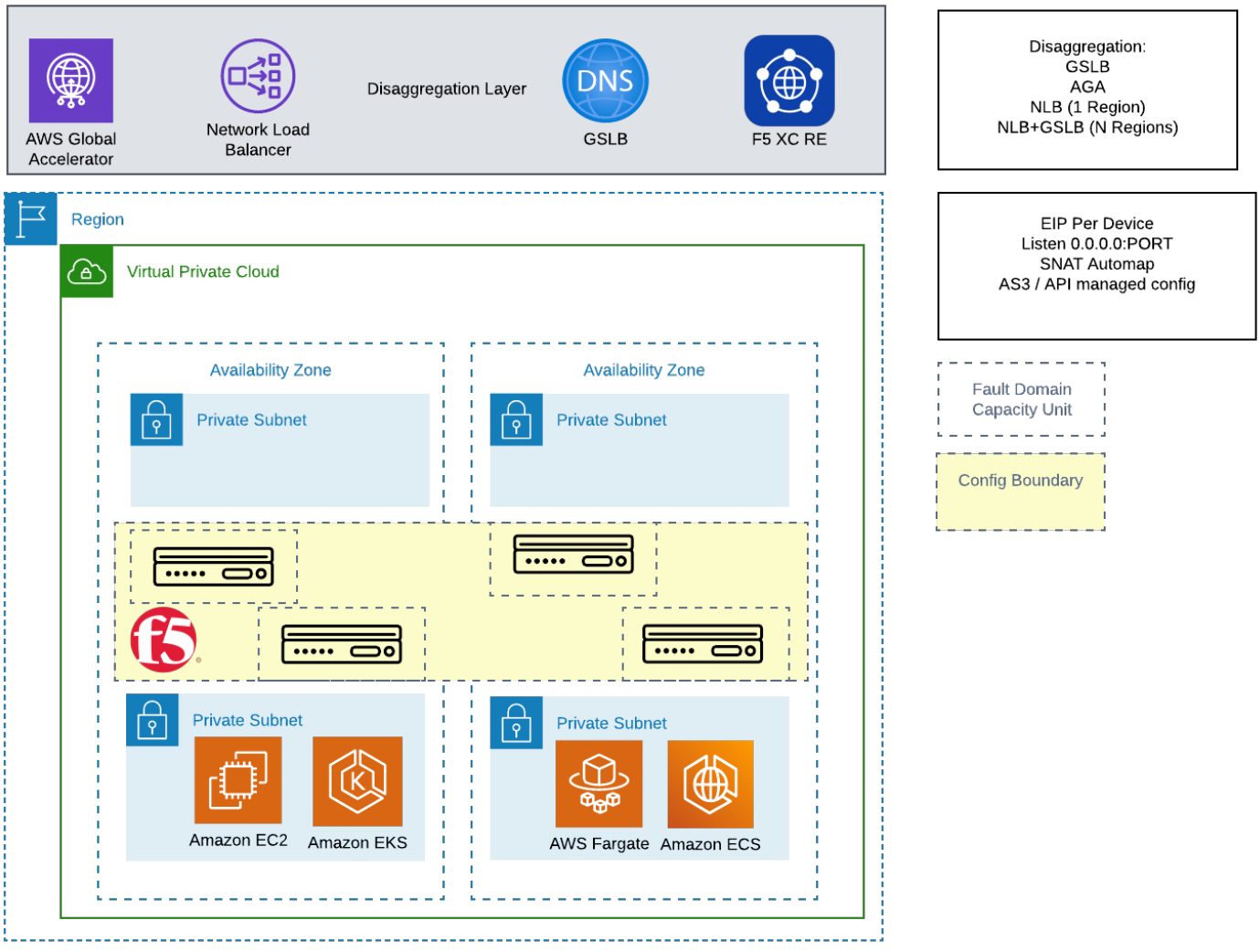
Active-standby multi-AZ

This creates high availability across AZs and protecting you from an AZ failure and the subset of AZ failures, but it does not protect you from a region failure.



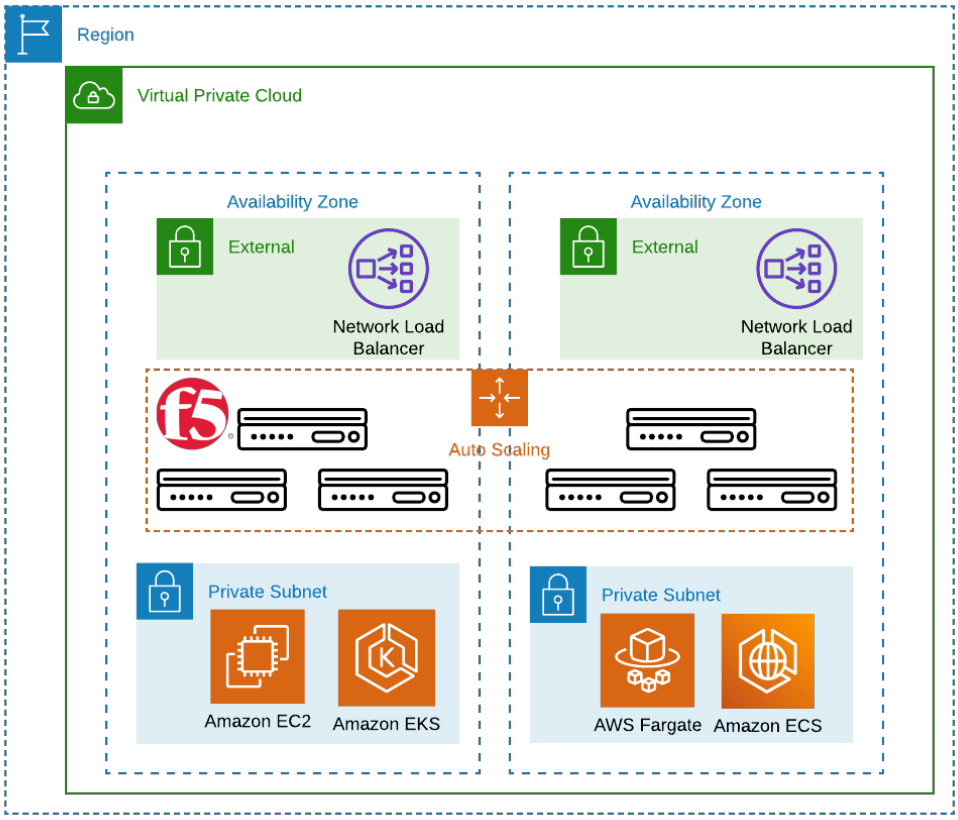
Active-active

Active-active deployments have ‘n’ number of instances active at any given time. Resilience is provided by a client disaggregation service such as F5 Distributed Cloud Regional Edge (RE), GSLB, Amazon Elastic Load Balancer (ELB), or AWS Global Accelerator. Active-active deployments may have VIPs that are listening on 0.0.0.0/0:443 or :0/443 and will process traffic for all traffic that maps the TCP/UDP port that arrives on it regardless of the destination IP address. Active-active deployments are well aligned with applications that have massive scale and applications that are tolerant of SNAT to backend servers. To address operational complexity, active-active users should be leveraging F5 APIs to manage the configurations such as F5 BIG-IP Application Services 3 Extension (AS3) or F5® iControl®.



Auto scaling (Elastic active-active)

In an autoscaling deployment, BIG-IP systems are deployed in an automated fashion using AWS Auto Scaling, F5 declarative API interfaces, and stack updates to manage the configuration over time. Client traffic disaggregation can be provided by AWS Application Load Balancer, Network Load Balancer, F5 Distributed Cloud RE, Global Accelerator, Gateway Load balancer (GWLB) or GSLB. In the diagram below we see the high-level architecture of a regional deployment (NLB) used to distribute traffic of an Auto Scaling group of BIG-IP instances providing L7 Security and docs protection for the application. Users of the Auto Scaling pattern need to leverage the same APIs as the active-active pattern.

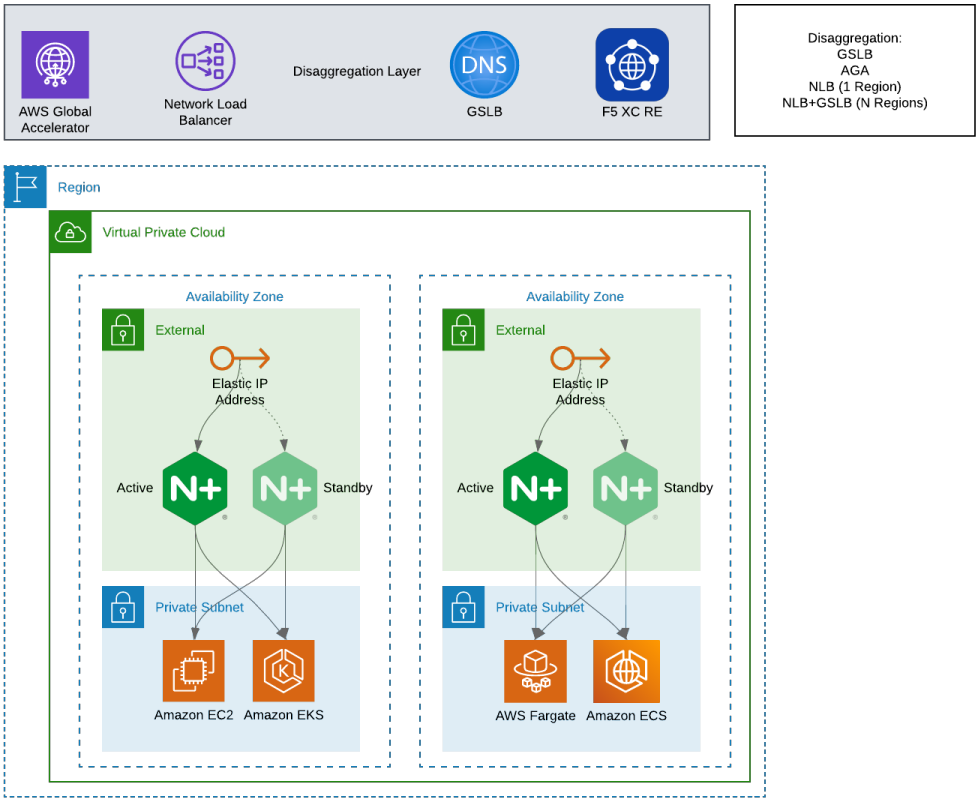


F5 Deployment Patterns–NGINX

Active-standby

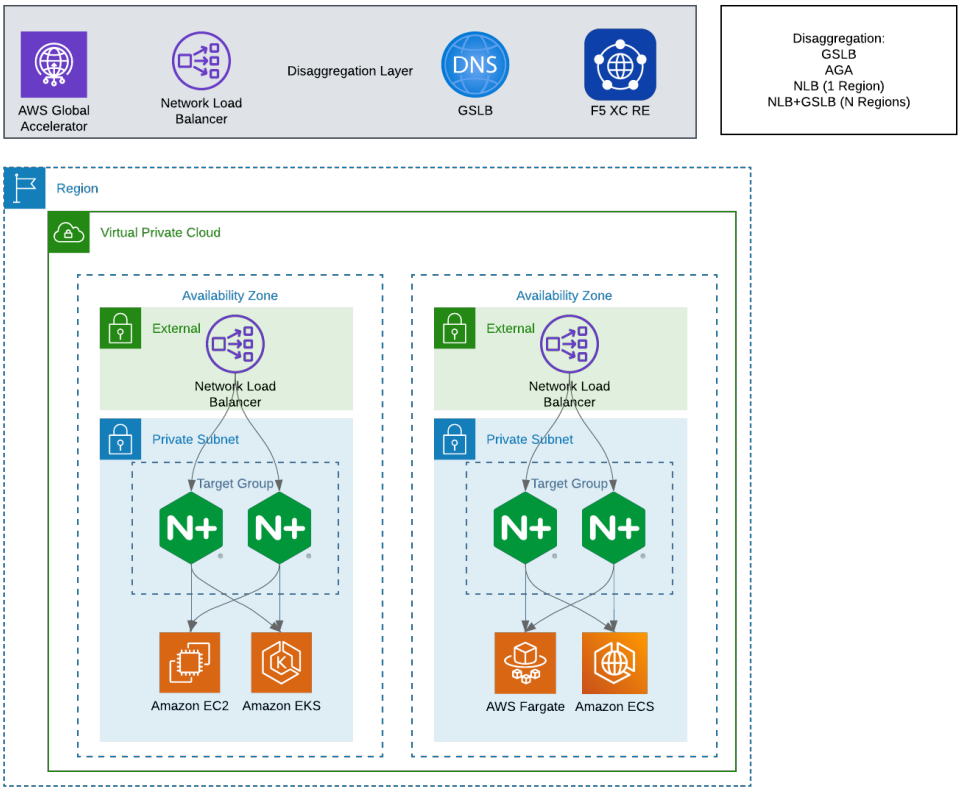
Active-standby deployments have one instance active at any given time. It combines the keepalived based solution with the AWS EIP address feature. This method addresses the requirement for a single IP address: as long as the primary F5® NGINX Plus® instance is operating correctly, it has the EIP address. If the primary fails, the backup instance becomes the primary and reassociates the EIP address with itself. NGINX provides the scripts invoked by keepalived, but note these scripts are not covered under the NGINX Plus support contract. Solution details can be found at Active-Passive HA for NGINX Plus on AWS Using Elastic IP Addresses.

NOTE: Since keepalived requires L2 connectivity, this solution can only be used within in a single AZ. If cross-AZ or cross-region resilience is required, active-standby pairs can be deployed in multiple regions with a DNS-based global-availability solution such as F5 Distributed Cloud RE or GSLB to distribute traffic across regions.



Active-active

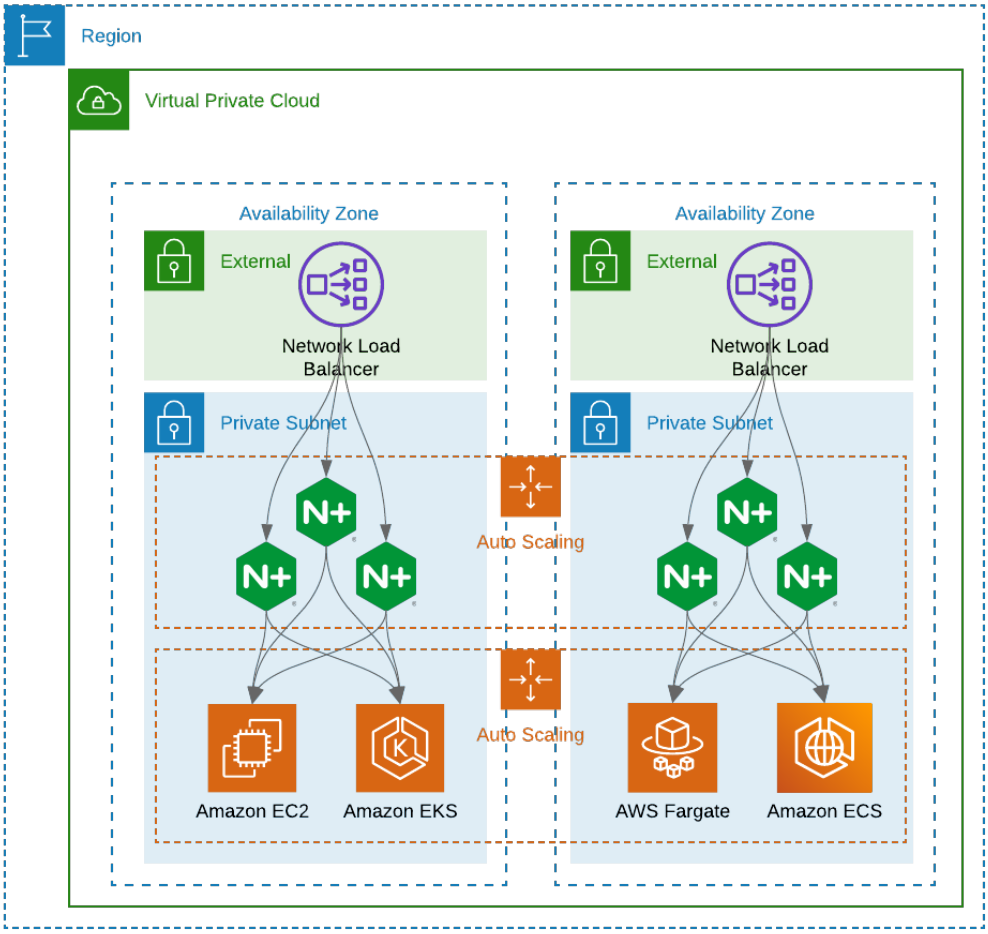
Active-active deployments have ‘n’ number of instances active at any given time. The solution combines the AWS Network Load Balancer (NLB) for fast and efficient handling of Layer 4 traffic with NGINX Plus for advanced, Layer 7 features, such as load balancing, caching, and content-based routing. Resilience is provided by a client disaggregation service such as F5 Distributed Cloud RE, GSLB, or AWS Global Accelerator. [Active-Active HA for NGINX Plus on AWS Using AWS Network Load Balancer](#).



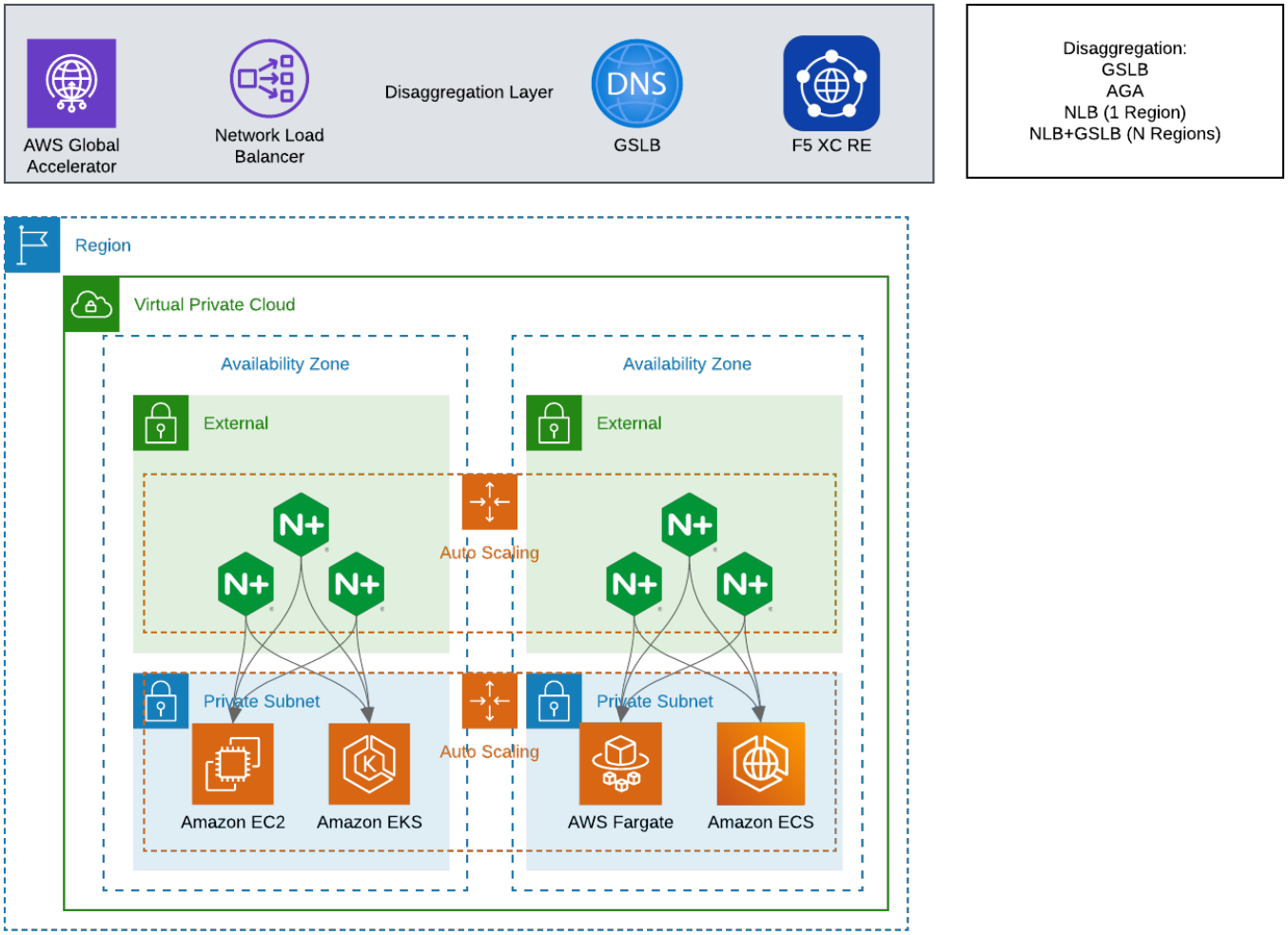
Auto Scaling (Elastic active-active)

In an autoscaling deployment of Amazon EC2 instances for NGINX are deployed in an automated fashion using AWS Auto Scaling—specifically using Auto Scaling groups provided by AWS ELB family or AWS Global Accelerator.

Application tiers can use Auto Scaling groups, and NGINX will dynamically send traffic to these group members as they are created. Elastic Load Balancing (ELB) can be used to proxy traffic to these upstream groups, or the community-supported **nginx-asg-sync** package can be used to discover members of the upstream Auto Scaling group. Details on these options can be found at Load Balancing AWS Auto Scaling Groups with NGINX Plus.



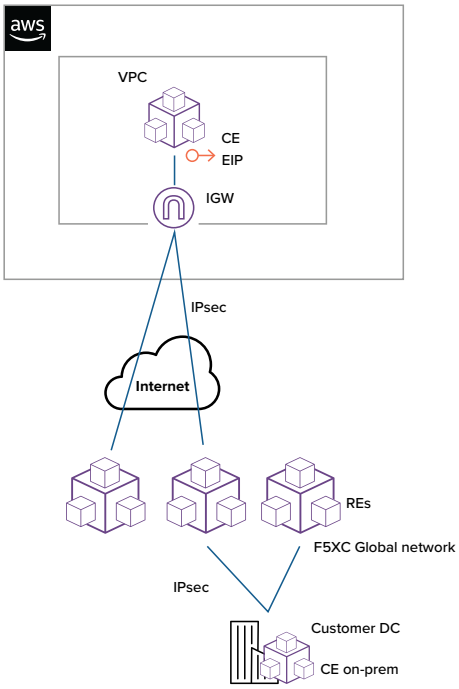
As an alternative, NGINX Plus can be deployed in an external subnet with Public IPs assigned to each instance in the Auto Scaling group. In this kind of deployment, GSLB can be used to disaggregate the client traffic to the instances without the need for ELB. Note: GLSB should be configured to subscribe to the list of public IP addresses associated with each instance in the Auto Scaling group, emit low-TTL responses, and use health checks to validate the health of each instance.



F5 Deployment Patterns–Distributed Cloud

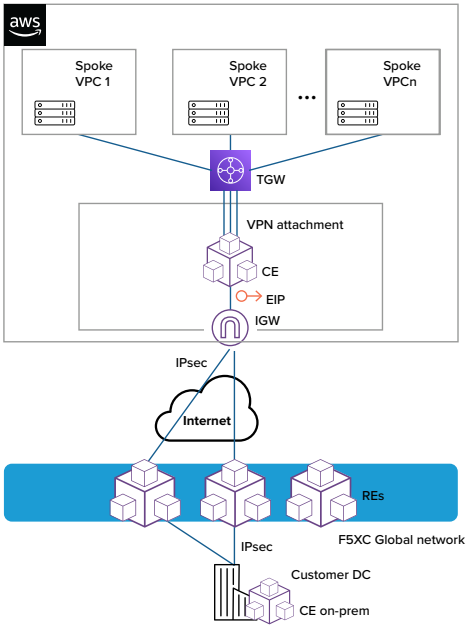
AWS “sites” (one or more customer edges) are created through the F5 Distributed Cloud Console using cloud credentials with limited IAM roles. All sites support single or multiple node configurations. Single Availability Zone or 3 Availability Zone deployments are available. Within AZs, worker nodes can be added and scaled through the F5® Distributed Cloud Console.

Amazon VPC Site Type



Distributed Cloud Console is used to deploy a customer edge site in a new or existing VPC. For a new VPC deployment, network objects (route tables, security groups, etc.) are created to allow egress through an Internet Gateway and ingress through an EIP.

AWS Transit Gateway Site Type



Distributed Cloud Console is used to deploy a customer edge site in a new or existing VPC similar to the VPC site topology along with an AWS Transit Gateway. For Transit Gateway sites created through the Distributed Cloud Console, spoke VPC attachments can be managed through the Console.

F5 Combinations with AWS Services

When deploying applications into AWS customers will adopt an array of services to meet specific use cases. F5 compliments these solutions to meet a complete application traffic management and security solution.

Elastic Load Balancer (ELB): F5 leverages ELB services to create the disaggregation layer for active-active and active-active Auto Scaling solutions. In addition to using ELB as a disaggregation solution, F5 can also use ELB as pool members.

Gateway Load Balancer (GWLB): Integrating F5 with GWLB allows organizations to build single function (WAF, AFM) and multi-function (SSLO + WAF, AFM, third-party NGFW, open source security software) insertion into security inter subnet, Internet ingress, and inter-VPC topologies.

CloudFront: F5 products integrate or interoperate with Amazon CloudFront in multiple ways. F5's connector for bot defense can leverage AWS Lambda at edge to insert bot security into the traffic pattern at the CDN layer. CloudFront can leverage F5 systems running inside and outside of AWS as the origin server(s) for content.

Transit Gateway: AWS Transit Gateway makes it easier to control traffic into, out of and between VPCs and your hybrid network. F5 allows you to use TGW for Border Gateway Protocol (BGP) peering, creating the ability to perform RHI failover, equal cost multi-path (ECMP), or drive VIP density by using Alien IP addressing to exceed the secondary IP limits of a given Amazon EC2 instance.

EKS, ECS, EC2: F5 offers run on Amazon EC2 instances and support applications running on Amazon EC2, EKS, ECS, EKS Anywhere and running in different AWS facilities such as regions, local zones, wavelength zones, and outpost. Additionally, F5 can leverage other AWS services as pool members such as API Gateway network interfaces.

CloudWatch: F5 supports sending metrics to AWS CloudWatch. These metrics are used for Auto Scaling BIG-IP patterns used in our Cloud Formation templates and by F5 Telemetry Streaming if you desire to send log data to Amazon CloudWatch.

F5 Service Discovery on AWS

Depending on the application or the operational model of your organization you may need to use service discovery.

Service discovery mechanisms for BIG-IP

	EC2	EKS	ECS	Elastic Network Interfaces
Discovery Type				
DNS A Record	Yes	Yes	Yes	Yes
DNS SRV Record	No	No	Yes	No
Instance Tag	Yes	No	No	No
Interface TAG	Yes		Yes	Yes
External Event Driven	Customer customized	CIS	Customer customized	Customer customized

Tag-Based Service discovery

With tag-based service discovery F5 BIG-IP Application Services 3 (BIG-IP AS3) interrogates the AWS API for the tags applied to instances or network interfaces and applies a filter based on preference for using public or private IP addresses compute resources you intend to target. The BIG-IP instance requires proper IAM permissions to describe instance and network objects in AWS. For more information please see F5 CloudDocs.

DNS-based service discovery

DNS-based service discovery is used from generate DNS servers and services or from a service registry such as AWS Cloud Map or the AWS Route 53 resolver for EKS.

Elastic Network services can be any service you would like the BIG-IP to direct traffic to that presents an ENI into the environment. Examples of this include service link and internal API gateways to name a few.

Container ingress-based service discovery

When using BIG-IP to route traffic into EKS, it is recommended to use F5® BIG-IP® Container Ingress Services. With this solution, you can route external traffic directly to Services deployed in Kubernetes by their names. No integration with service registries external to Kubernetes is required.

NGINX Plus-based service discovery

NGINX Plus supports service discovery of upstreams by specifying a DNS resolver in the NGINX configuration. This enables you to register services using products such as Route 53, and NGINX Plus will consume these services by use of its resolver directive.

NGINX Kubernetes ingress-based service discovery

When using F5® NGINX® Ingress Controller, Kubernetes services can be located if they are of types Services, Endpoints and Pods, and used as upstreams. Additionally, services of type ExternalName can be used as upstreams
if a proper resolver is configured via ConfigMap.

	NGINX Plus
Discovery Type	
DNS A Record	Yes
DNS SRV Record	Yes
Instance Tag	No
Interface TAG	No
External Event Driven	Customer Customized

F5 Traffic Management into Different ECS Networking Modes

Host: In this mode the container listens on the host network interface to a statically mapped port. The EC2 Auto Scaling group

Name	Description	Discovery Mechanism	Discovery without Custom Code?
Amazon VPC Network	The task is assigned a unique ENI at runtime and is directly routable. The container may or may not also have an EIP assigned. SG applies to the ENI of the task.	DNS A or SRV Records	Yes Fully qualified domain name
Bridge	Containers use the docker bridge and are not routable. The containers use dynamic port mappings, SG applies to the ENI of the Host. Uses the Linux bridge so offers lower network performance than host.	DNS SRV records, AWS API	No Note: A workaround is the user can preconfigure the task with a port mapping on the container. This limits scalability—for example, a HOST is limited to only a given task.
Host	Uses the host network stack. Normally leverage dynamic port mappings, SG applies to the ENI of the Host.	DNS SRV records, AWS API	No Note: A workaround is the user can preconfigure the task with a port mapping on the container. This limits scalability—for example, a HOST is limited to only a given task.
None	Containers are not reachable.	Not Applicable	Not Applicable

NGINX as an external proxy into ECS

For use cases where NGINX is deployed external to the ECS cluster—not running as an EC2 task—you use any of the NGINX deployment patterns to route traffic from external clients into the ECS cluster.

NGINX as a container in ECS

When deploying NGINX as a container inside the ECS cluster is desired, it is possible as a standalone container deployment—
as in this example—or as a sidecar running alongside each application container.

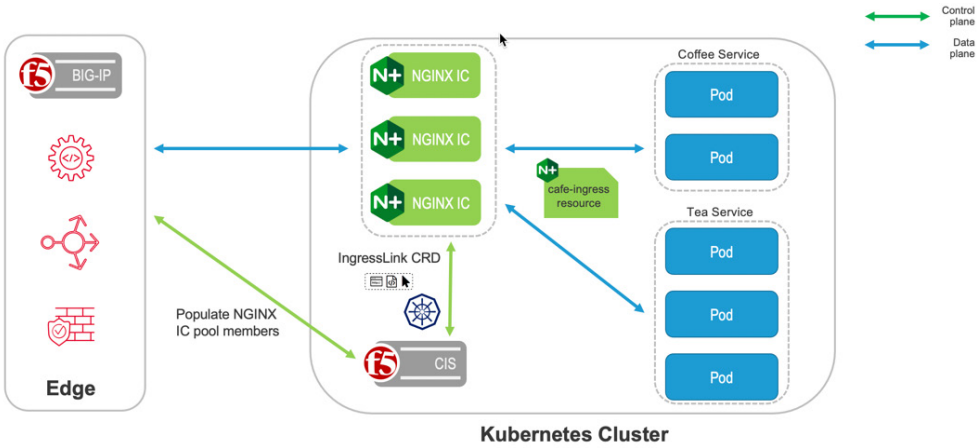
F5 Security and Traffic Management for EKS and EKS-A

F5 BIG-IP has two deployment patterns with EKS and EKS-A. Please see DevCentral for examples of these solutions.

BIG-IP with NGINX

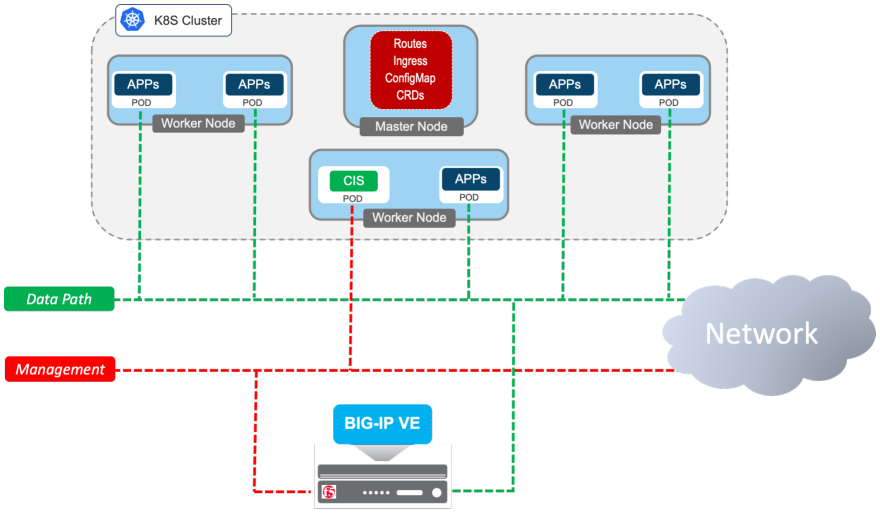
When we look at EKS and EKS Anywhere F5 provides two options. BIG-IP Container Ingress Services (CIS) and F5 BIG-IP with NGINX and F5 IngressLink. Both are able to direct traffic to pods in the cluster; when organizations are looking to separate the function of networking outside of the cluster form networking inside of the cluster, then F5 IngressLink is the correct solution.

If we apply this to an organization where they are using EKS, EKS-A, or their own K8S solution on AWS we now have a consistent pattern in deployment across all models. For deployment of BIG-IP Container Ingress Service and F5 IngressLink.



BIG-IP Container Ingress Services

If your organization does not want to have the additional layer of traffic control and steer the traffic directly from BIG-IP you would have a topology such as:

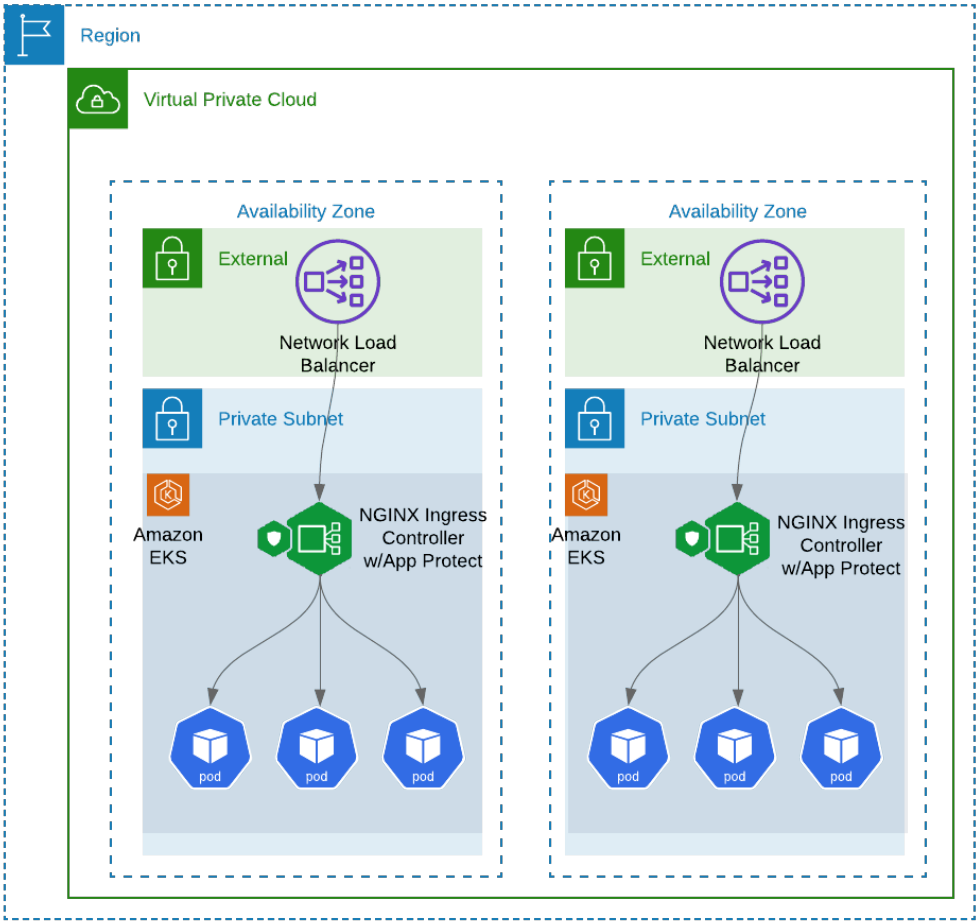


NGINX Ingress Controller

NGINX can be used as an Ingress Controller in EKS environments. Introducing NGINX Ingress Controller to process north-south traffic provides a number of benefits:

- Granular role-based access control for application teams to specify their own routing rules and security policies
- Insertion points for F5® NGINX® App Protect WAF and F5® NGINX® App Protect DoS solutions
- Authorization services for applications and APIs using OIDC, OAuth2, and SAML
- JSON Web Token (JWT) validation for APIs
- Rate limiting
- OpenAPI (formerly Swagger) Specification enforcement for APIs using NGINX App Protect WAF
- Dynamic allow or denylisting by IP, classless inter-domain routing (CIDR), or geolocation database information

NGINX Ingress Controller can be installed using Helm or standalone manifests from the NGINX private container registry or consumed via the AWS Marketplace and installed from the Amazon Elastic Container Registry (ECR)registry. Please see DevCentral for an example of this solution.

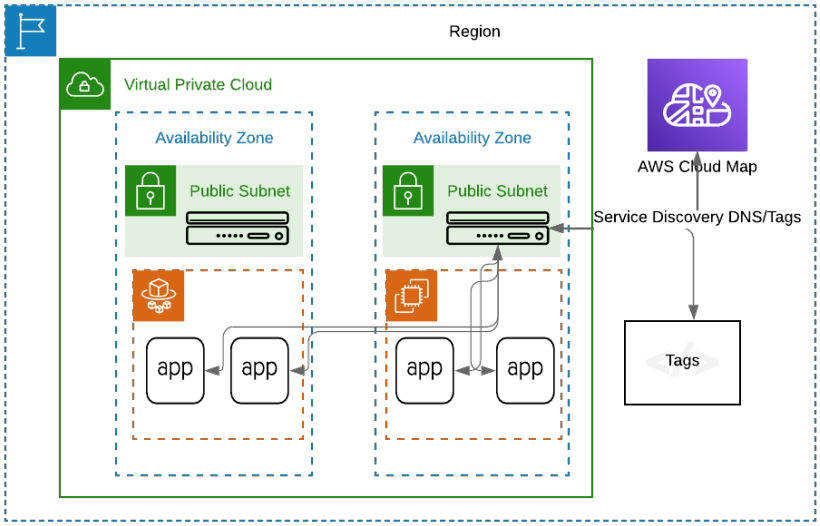


F5 Security and Traffic Management for ECS

APPLICABLE PATTERNS: Active-Standby

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

When applications are deployed on EC2 instances it is like when they are deployed into a data center. The EC2 instances become pool members on BIG-IP instances running in AWS, your data center or at a Cloud Interconnect location. For more details, please see the F5 GitHub.



NGINX Security and Traffic Management for ECS

When using NGINX Plus to securely deliver services residing in ECS, one or more proxy instances can receive client traffic via DNS-based disaggregators and EIPs, or receive traffic from an ELB. Once the traffic is proxied by NGINX, upstreams in ECS can be selected with the help of DNS resolvers combined with the Route 53 Service registry. Service discovery can also be accomplished by configuring NGINX to resolve upstreams from AWS CloudMap A or SRV records. Please see DevCentral for more information on this solution.

When NGINX Plus is established in the path, NGINX App Protect products such as WAF and anti-DoS can be introduced right in NGINX without adding additional hops combined with minimal latency overhead.

Additionally, NGINX Plus can provide authorization services for traditional applications and APIs by securing them using OIDC, OAuth2, or SAML.

NGINX Plus can also perform dynamic denylisting by IP, CIDR, or geolocation database information if the source IP is preserved at ingress.

Finally, rate limiting can be enforced on a per-client basis to prevent abuse and ensure availability of your traditional applications and APIs.

NGINX Security and Traffic Management for Amazon ECS-A

NGINX can also be used in an ECS Anywhere deployment, with the caveat that any services that rely on calls to the AWS APIs will require network connectivity to them, and the proper AWS Identity and Access Management (IAM) roles need to be set up. See the FAQ and Considerations for more information.

BIG-IP Security and Traffic Management for Amazon ECS-A

APPLICABLE PATTERNS: Active-Standby

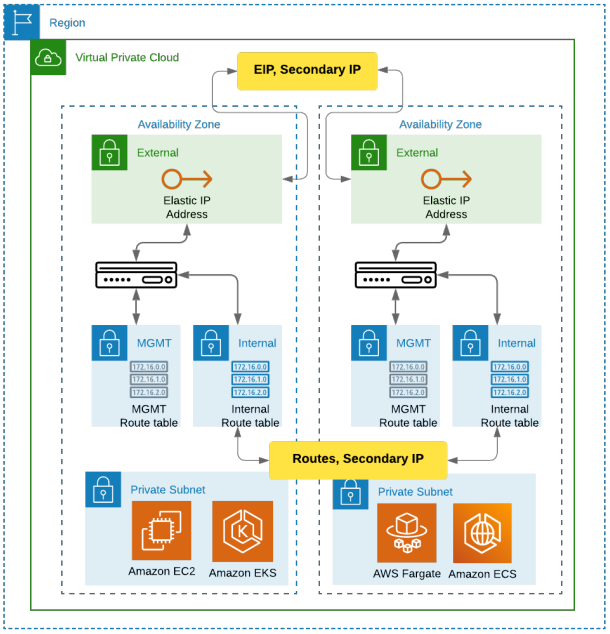
APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

BIG-IP can support integration and service discovery with ECS Anywhere using a community supported pattern. Please see DevCentral for more information.

F5 Security and Traffic Management for Amazon EC2 Applications

APPLICABLE PATTERNS: Active-Standby, Active-Active, and Auto Scaling

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)



When applications are deployed on EC2 instances it is like when they are deployed into a data center. The EC2 instances become pool members on BIG-IP instances running in AWS, your data center or at a Cloud Interconnect location. For more details, please see the F5 GitHub.

F5 Security for ELB and Direct to EC2 Deployed Applications

Applications deployed on Amazon EC2 may be static in capacity size or they may be deployed in Auto Scaling groups. Either way F5 offers solutions to provide traffic management, security, and visibility in the application.

Proxied: In proxied mode a user leverages BIG-IP, NGINX, or Distributed Cloud to provide the application access point and traffic is both proxied and load balanced across an array of application endpoints. The EIP is mapped to one or more BIG-IP, NGINX, or Distributed Cloud instances.

Transparent: Insertion of F5® BIG-IP® Advanced WAF®, F5®BIG-IP®Advanced Firewall Manager (AFM) between the deployed application endpoints such as ELBs and Instances where the EIP is mapped to the ELB or the instances.

F5 Transparent Security Use Cases of East/West, External/Internal and Internal/External Traffic Flows

APPLICABLE PATTERNS: Active-Standby or Active-Active

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO

IT environments are increasingly complex. When you evaluate all the network flows that you have in your estate you have many different patterns and security needs. On top of the diversity of traffic flows you also have diversity in security requirements. To provide security in an environment one needs to be able to accomplish the following:

Determine the security use case profile: Is this security use case solved by a single service, such as Web Application Firewalls, or do we need multiple security services? Do we need to include multiple vendors or OSS security offers?

Discern the flow characteristics. Is this flow east-west, external to internal, or internal to external. For example, for an internal to external flow we may need to forge a certificate to perform Layer 7 traffic inspection. Simply put, we need to define what traffic to intercept and apply a default security policy to.

Apply advanced logic to further refine if we should apply different security policies than the default policy.

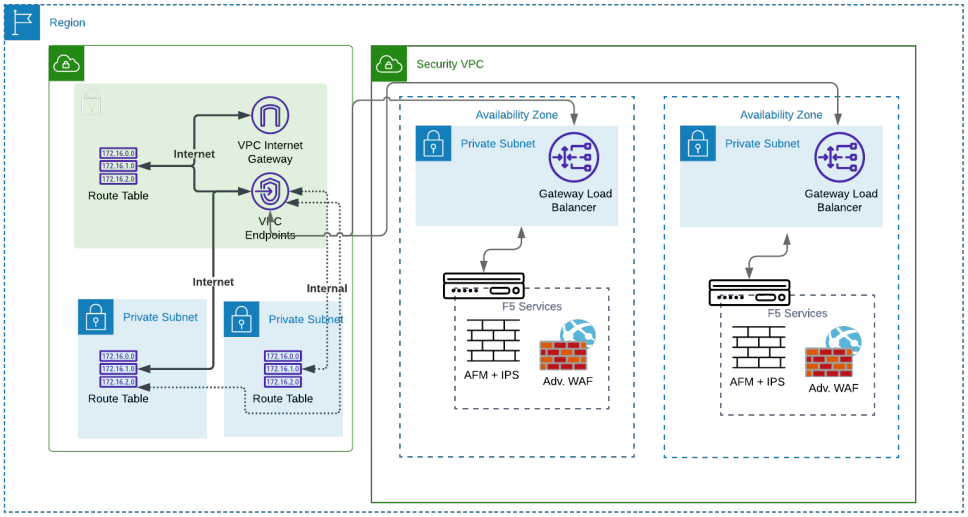
Creating of services and service chains to applied to the security policy.

If we look at the traffic pattern; we will use a Gateway Load Balancer to load balance traffic across one or more BIG-IP instances that are in a target group, and Gateway Load balancer endpoints will be inserted into AWS consumer VPC routing tables to steer the traffic into the Gateway Load Balancer endpoints, over to the security VPC GWLB.

Single System Security Function–BIG-IP Advanced WAF and BIG-IP Advanced Firewall Manager

In this model we are working with F5 security services. Traffic that is inspected is intercepted in the route table by the Gateway Load Balancer endpoint in the consumer VPC and is directed towards an array of F5 instances in the Gateway Load Balancer target group. BIG-IP is listening on the GENEVE tunnel for the traffic with a wild card virtual server, applies the security logic and traffic is sent back on the GENEVE tunnel. Please see DevCentral for an example of this topology.

Security solutions inserted: Advanced WAF, Intrusion Prevention, IP intelligence, Threat Campaigns, L4 DOS, L7 DOS, Bot Defense.

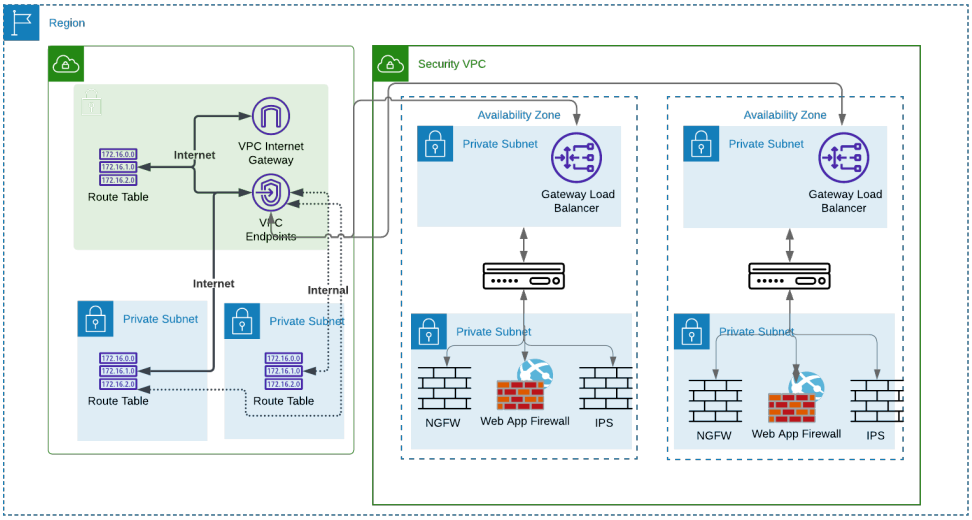


Multiple security functions–F5 and third-party products

In this model we are working with F5 security services and third-party services. Traffic that is inspected is intercepted in the route table by the Gateway Load Balancer endpoint in the consumer VPC and is directed towards an array of F5 instances running SSL Orchestrator in the Gateway Load Balancer target group. BIG-IP is listening on the GENEVE tunnel for the traffic with a wild card virtual server, applies the interception rules and security polices steering traffic to different appliances in a security service chain. These services can be from F5 or from a third-party service provider. For an example, please see this article on DevCentral.

Security solutions inserted: Advanced WAF, Intrusion Prevention, IP intelligence, Threat Campaigns, L4 DoS protection, L7 DoS protection, Bot Defense.

Examples of third-party software: Suricata, Arkime



This setup requires that we integrate with the AWS Gateway Load Balancer GENEVE Protocol.

F5 Solutions for Internal Applications that cannot use DNS

APPLICABLE PATTERNS: Active-Standby or Active-Active

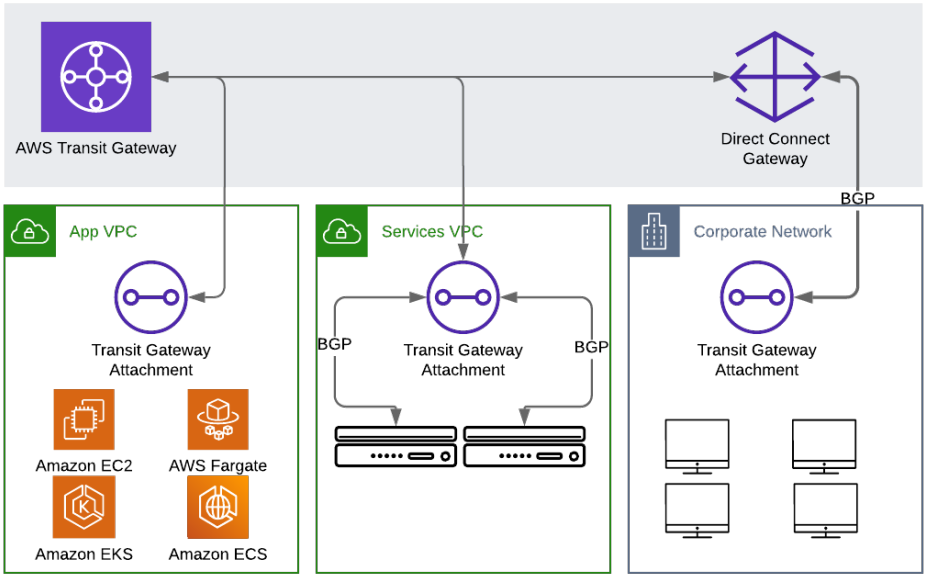
APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO

Many organizations have an internal application that is not integrated into DNS. Perhaps this is due to firewall rules that are in place, or it could be due to a middleware that does not honor DNS TTLS and once it caches an IP address, it will not perform another lookup until the service restarts. For public applications you can abstract the AZ fault domain and network boundary with an EIP construct or with AWS Global Accelerator. When it comes to internal applications many organizations find this more complex. F5 uses a concept we refer to as “Alien IP” addresses. In the simplest form, an Alien IP address is a route targeting an elastic network interface that exists outside of the VPC address space that can “float” from one AZ to the next. For this pattern we use BIG-IP instances and have three deployment models.

1	2	3
Static route targets controlled by BIG-IP Cloud Failover Extension	BGP RHI with TGW GRE Attachments	VGW VPN attachments—environments that have VPN/TGW terminate on a VPC

RHI over TGW Connect Attachments

AWS Transit Gateway (TGW) allows BIG-IP to BGP peer with it via a Transit Gateway Attachment. Once BGP is established routes are exchanged between BIG-IP and TGW populating the TGW route table. Please see DevCentral for an example of this topology.

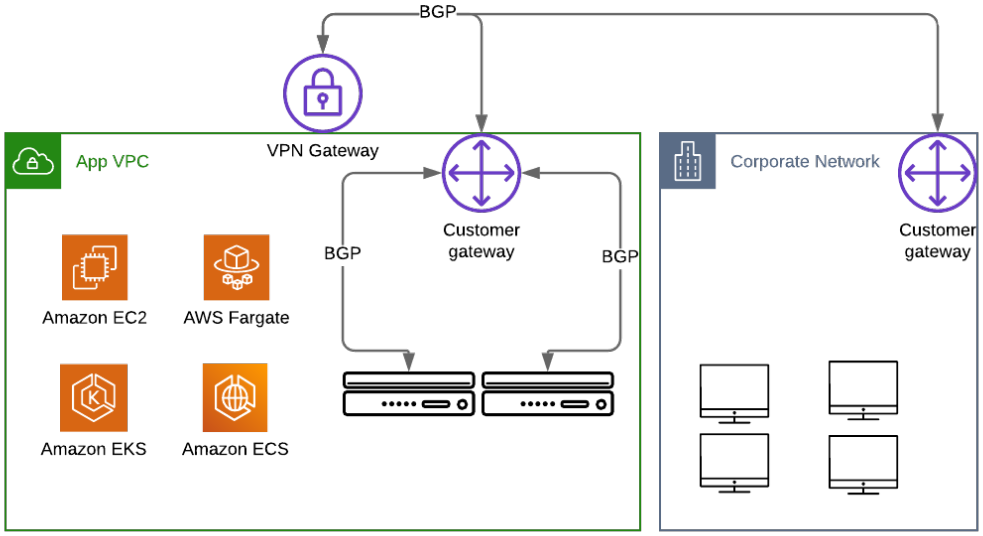


RHI via VGW

In concept this topology is very similar to the Transit Gateway model. There are a few key differences.

1. BIG-IP instances are deployed into one more Availability Zones.
2. The EIPs of the BIG-IP external interfaces are used to generate customer gateway configurations in AWS VPN.
3. Traffic from the VPC will towards the alien range will go to the VGW and then to the active BIG-IP over the VPN tunnel.

This pattern should only be used in scenarios where Transit Gateway cannot be used.

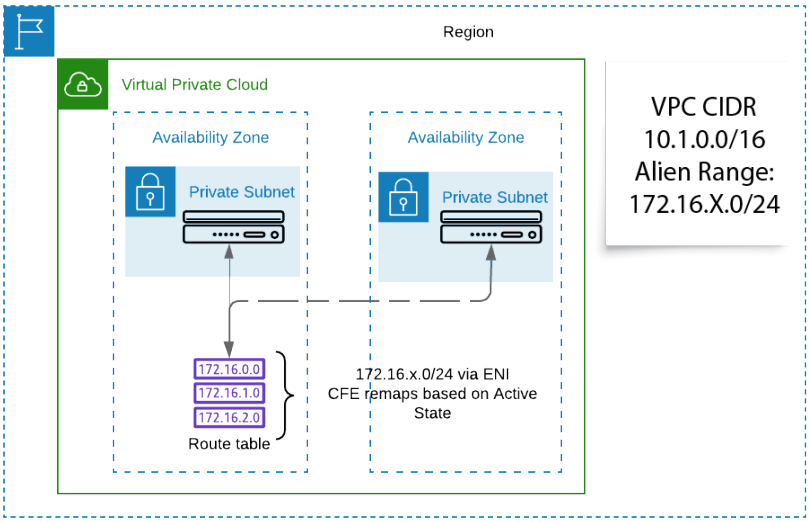


Routing and Alien IP Addresses for Virtual Server Density

APPLICABLE PATTERNS: Active-Standby

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

When organizations need to host a large number of VIPs in AWS, but do not need a public IP address we can driver virtual server density beyond what an EC2 instance normally supports with a concept called “Alien IP address.” An alien IP address is an IP address or address range that exists in a VPC route table and is routed to an instance. This IP address is not from the VPC CIDR block (thus Alien). For more information please refer to F5 DevCentral.



Interconnection considerations for routed applications

If your application cannot use DNS special attention needs to be paid to both your connection to VPCs and your connections between VPCs to enable traffic to flow.

Nontransitive Configurations: Several configurations in AWS do not support transitive properties. VPC peering and TGW/VGW attachments to VPCs. In these scenarios traffic that does not match the VPC CIDR ranges will be dropped.

Transitive Configurations: Transitive configurations support routing traffic that would be in the alien IP address range.

VPC to VPC interconnect

Security VPC with VPC Peering	Security VPC with TransitGateway	Security VPC with VPN Interconnect
Benefits • Easy and quick to setup • Simple routing • High redundancy • High bandwidth	Benefits • Easy to setup • Flexible routing without SNAT • High redundancy • High bandwidth • Easy to manage at scale	Benefits • Flexible routing without SNAT • Easy insertion of security inspection between VPCs
Drawbacks • Only supports traffic from VPC assigned CIDR ranges • Cannot insert security inspection between VPCs	Drawbacks • Routing is more complex (VPC route tables and TransitGateway Route tables)	Drawbacks • Low bandwidth • Complex vendor specific dependent failover • Complex to manage at scale—all are point-to-point

Client (Sends SYN)	Transit Gateway	VPC Peering	VPN (Between VPCs)	Solution Overview/ Considerations
Internet/DC to service in a single VPC with a public subnet or private	N/A	N/A	N/A	• Traffic traverses IGW, or Virtual Gateway—does not need to cross more than VPC boundary • VPC acts as designed—a stub network(s) • Traffic ingresses by designed manner from DC (Direct Connect, VPN)
Internet/DC to a service in a VPC with clients in other VPCs (for example, pool members in another VPC)—no SNAT	Yes	No	Yes	• Transit Gateway or VPNs allow the traffic to bypass the VPC Peering filter of only VPC assigned CIDRs can pass • VPN solutions will be constrained—no ECMP (only a single route) and bandwidth (about 1.2 Gb/S per tunnel— in general only one tunnel)

Client (Sends SYN)	Transit Gateway	VPC Peering	VPN (Between VPCs)	Solution Overview/ Considerations
Internet/DC to a service in a VPC with Clients in other VPCs (IE pool members in another VPC)—SNAT	Yes (but not needed)	Yes	Yes (but not needed)	• Since the interconnection between the VPCs is seeing traffic from VPC assigned CIDRs any will work • VPN solutions will be constrained—no ECMP (only a single route) and bandwidth (about 1.2 Gb/S per tunnel—in general only one tunnel)
Inside of VPC to Service in same VPC	N/A	N/A	N/A	• All traffic constrained to a single VPC—interconnect is not required
Inside of one VPC to a Service VPC—service is in destination VPC CIDR	Yes (but not needed)	Yes	Yes (but not needed)	• Since the interconnection between the VPCs is seeing traffic from VPC assigned CIDRs any will work • VPN solutions will be constrained—no ECMP (only a single route) and bandwidth (about 1.2 Gb/S per tunnel—in general only one tunnel)
Inside of one VPC to a Service VPC—service is outside (alien) of VPC CDIR range	Yes	No	Yes	• Since the interconnection between the VPCs is seeing traffic from outside a VPC assigned CIDR range a VPC Peering cannot be used • VPN solutions will be constrained—no ECMP (only a single route) and bandwidth (about 1.2 Gb/S per tunnel—in general only one tunnel)
Inside of a single VPC to an Internet service	N/A	N/A	N/A	• Traffic is from a VPC assigned CIDR as long as EIP/NAT/Route Table constructs are inline traffic will flow
Inside a VPC to an Internet service—routing out though a security or inspection VPC	Yes	No	Yes	• Since the interconnection between the VPCs is seeing traffic from outside a VPC assigned CIDR range a VPC—peering cannot be used • VPN solutions will be constrained—no ECMP (only a single route) and bandwidth (about 1.2 Gb/S per tunnel—in general only one tunnel)

Data center to VPC interconnect

Connectivity Method	Routing Protocol Support	Bandwidth Limits	End Point IP addressing (Public/Private/Both)
Internet	N/A	You link into AWS. Instance outbound from AWS	Public
VPN—VPC	Static, BGP, (OSPF or third-party)	IP SEC limits (about 1.2 Gb/S per tunnel)	Private
VPN with Transit Gateway	Static, BGP, (OSPF or third-party)	IP SEC limits (about 1.2 Gb/S per tunnel)	Private
Direct Connect—VPC	Static, BGP	Direct Connect Limits (Supports Bonding), Individual Instances limited to 5 Gb/S	Both (Configurable)
Direct Connect—Gateway	Static, BGP	Direct Connect Limits (Supports Bonding), Individual Instances limited to 5 Gb/S	Both (Configurable)
Direct Connect Gateway— Transit Gateway	Static, BGP	Direct Connect Limits (Supports Bonding), Individual Instances limited to 5 Gb/S	Verbal Confirmation from AWS Architect team
Transit Gateway Attach	BGP	Direct Connect Limits (Supports Bonding), Individual Instances limited to 5 Gb/S	Verbal Confirmation from AWS Architect team

Data center to VPC interconnect

Connectivity Method	Support for Alien Address Space (IE routing non VPC CIDRs into and out of AWS)	Multi-VPC Support for 1 Connection (IE 1 link to Complex AWS topologies)	Multi-Region Support
Internet	No	Yes	Yes
VPN—VPC	Yes, you must setup an additional IPSEC tunnel form the BIG-IP in the VPC to the VGW connected to the VPC. Traffic traverses the tunnel	No	No
VPN with Transit Gateway	Yes	Yes	No Note: If TGW is extended will impact
Direct Connect—VPC	No	No	No
Direct Connect—Gateway	No	Yes	Yes
Direct Connect Gateway— Transit Gateway	Yes. Verbal from AWS. Do not have lab to verify	Yes	Yes
Transit Gateway Attach	Yes	Yes	Yes

F5 Security and DoS Solutions for AWS Global Accelerator

APPLICABLE PATTERNS: Active-Standby, Active-Active, and Auto Scaling

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

AWS Global Accelerator provides an optimized network path with static IP addresses for your applications. Global Accelerator allows you to have region and AZ resiliency patterns across your applications and provides L3/L4 DoS/DDoS security for your applications but does not address L7 security and DoS concerns. Organizations need to address the following challenges:

Do I have a L7 security solution?

Does my security only address known threats based on signatures?

Will my security solution detect bots?

What about malicious behavior?

Do I have the ability to mitigate by creating dynamic signatures?

Can I create stepped mitigation such as CAPTCHA, rate limit, and block?

Can my security solution learn my application?

Does my security solution evaluate threats on multiple vectors such as signature, behavior, protocol compliance?

How do I prevent credential stuffing attacks?

How do I mitigate AI, automated CAPTCHA solving, and human attacker farms?

How do I prevent fraud?

How do I deal with the ever-changing landscape of IP's that have attack reputations?

What can I do about emerging threat campaigns leveraging current attack trends?

Does my security solution help me score traffic for false positive and false negatives?

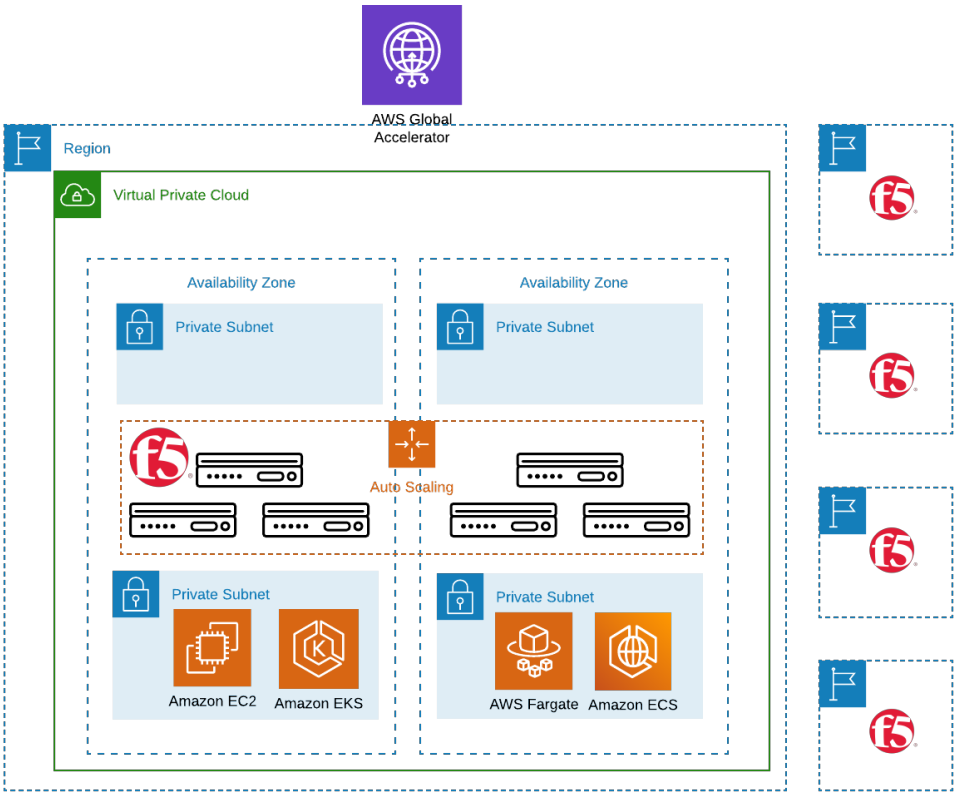
Does my security solution support web, mobile, and API traffic?

Does my security solution support as code management and deployment?

The topology consists of an AWS Global Accelerator with BIG-IP or NGINX running EC2 endpoint targets in one or more AWS regions and intern the BIG-IP and NGINX systems target any of the application compute types listed. By default, AWS Global Accelerator will direct traffic to the region closest to the user and there are resources to serve the request. This will provide you with a single global IP and with both region and AZ diversity capabilities.

This pattern relies on using an AWS provided script to update the endpoint group of Global Accelerator. You can find the AWS blog here and a copy of the script here. Please see DevCentral for an example of this pattern.

To deploy this with NGINX, the Auto Scaling deployment pattern provides the necessary foundation to support Global Accelerator. More information at NGINX Auto Scaling (Elastic active-active) deployment pattern.

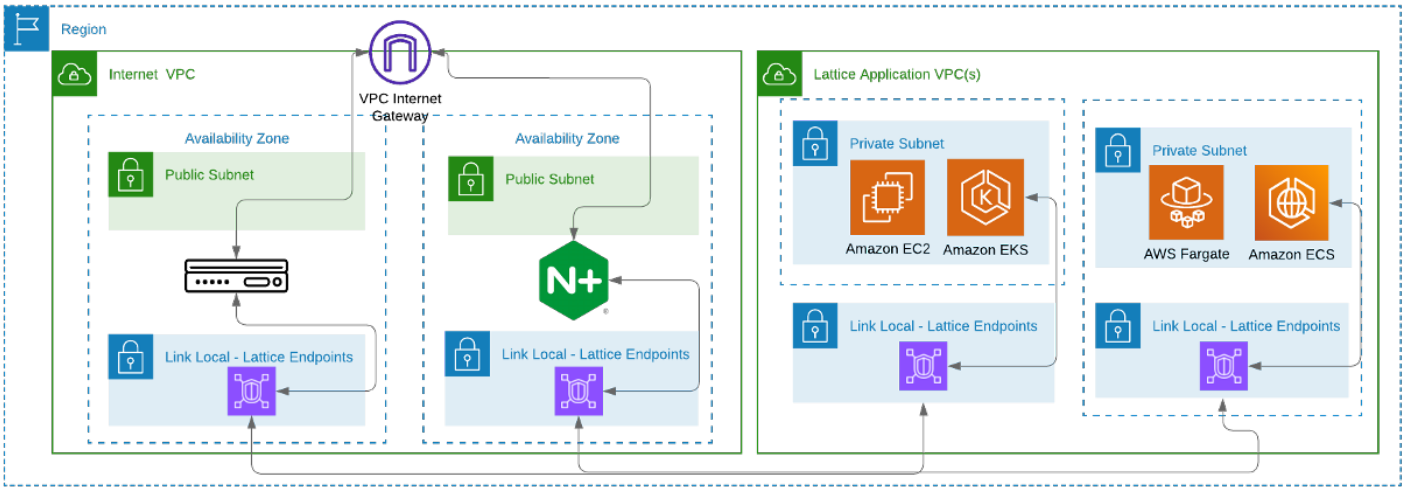


F5 Solutions for VPC Lattice

APPLICABLE PATTERNS: Active-Standby, Active-Active, and Auto Scaling

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

Amazon VPC Lattice is a multi-account application strategy that builds a "lattice" across VPCs and accounts and presents those application endpoints in the link local address space of 169.x.x.x only reachable from in AWS. If you have applications that are running in a VPC Lattice and need connections from systems that are external to the lattice or external to AWS, you will need an ingress proxy. F5 can address this need with both BIG-IP and NGINX. Please see DevCentral for an example of this solution.



F5 Solutions for Bot and Fraud Protections on AWS

APPLICABLE PATTERNS: Active-Standby, Active-Active

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active Distributed Cloud only)

Organizations require bot and fraud protection for AWS base applications. F5 offers a standard (shared service) and advanced solution to address these concerns.

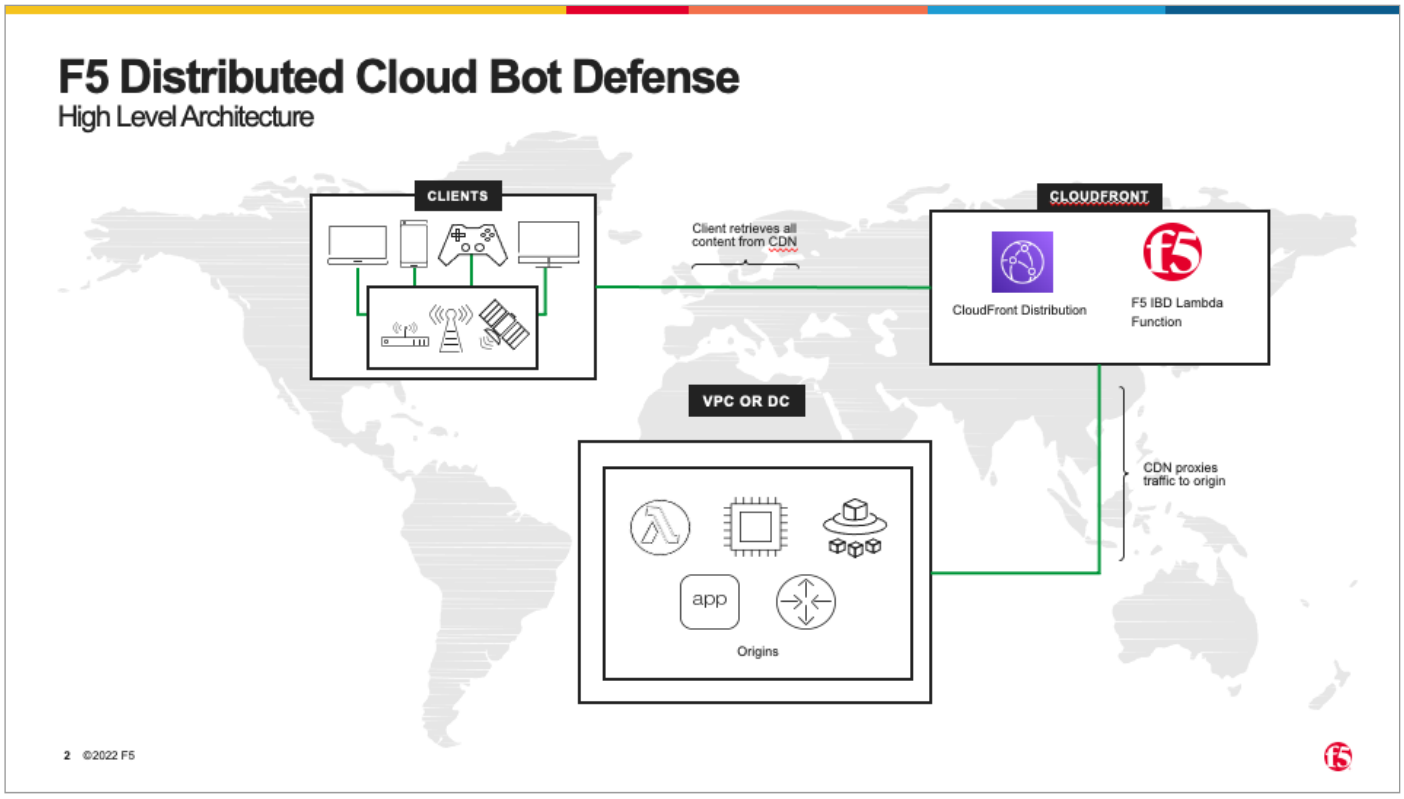
In our bot defense patterns users need to have an account setup on F5 Distributed Cloud and have subscribed to the service. Once that step is complete the different insertion points.

F5 bot mitigation insertion points

Depending on your application architecture and which F5 components you have deployed you can select where to insert the bot defense inspection.

AWS CloudFront Connector

Users who are leveraging AWS CloudFront can use the F5® Distributed Cloud Bot Defense Connector to insert bot defenses via Lambda@Edge. For applications that can leverage a CDN this creates a beneficial insertion point as the bot threats are being addressed near the edge. Please see DevCentral for an example of this pattern.



BIG-IP

Users can leverage BIG-IP integrations to insert bot defense. This pattern is compatible with active-standby and active-active. Traffic is attracted to the BIG-IP by mapping an EIP, ELB, Global Accelerator, or CloudFront to BIG-IP as a target. BIG-IP in turn sends traffic that passes the security policy to any of the compute types listed.

NGINX

Similar to BIG-IP, the insertion points for bot defense can also be used in NGINX when used at L7 in any of the deployment patterns.

Centralized Shared Deployments in Public Cloud

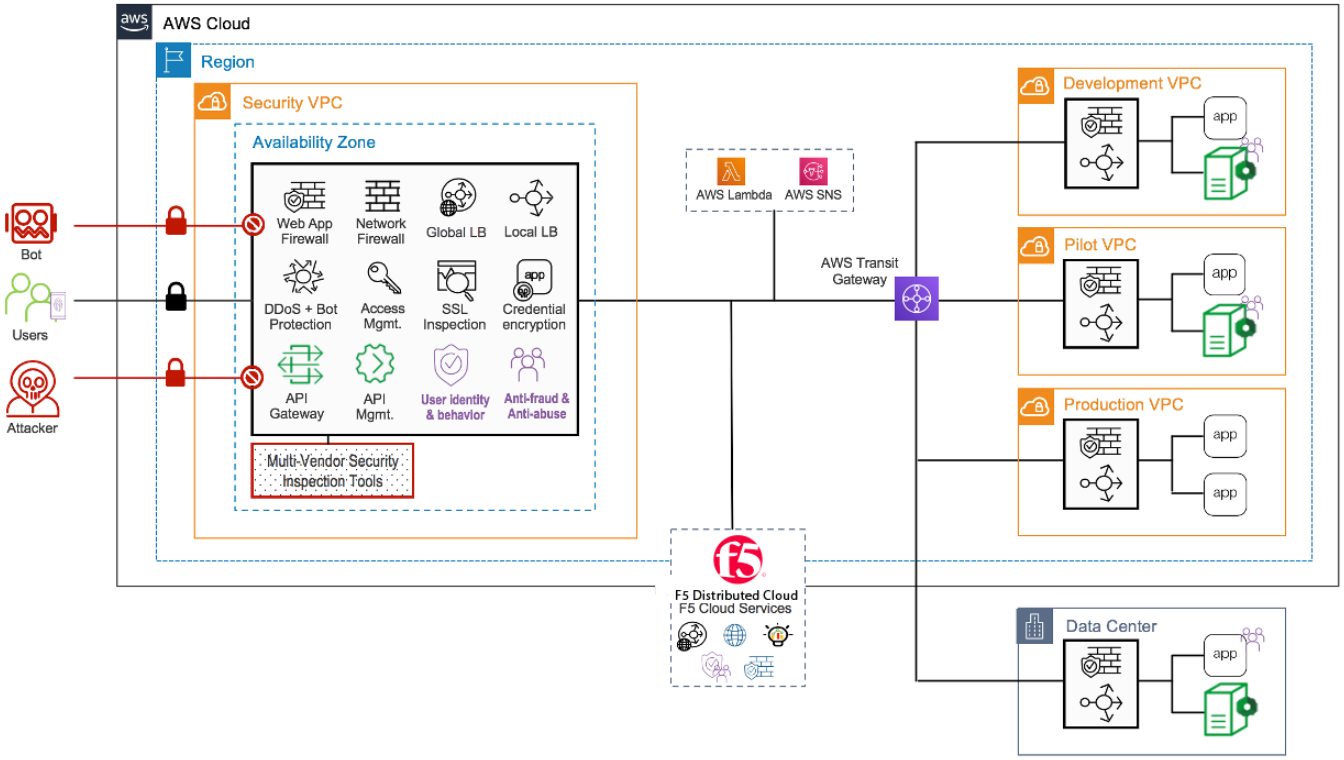
APPLICABLE PATTERNS: Active-Standby, Active-Active

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

Organizations are often interested in running their BIG-IP services in a centralized model in the public cloud. This is possible, with nuances of the architecture depending on the performance dimensions. A centralized deployment pattern can be at the edge, similar to a DMZ structure, or at the core of the VPC architecture.

F5 in cloud DMZ architectures

A cloud DMZ architecture operates in a similar manner to a data center DMZ pattern.



Shared VPC

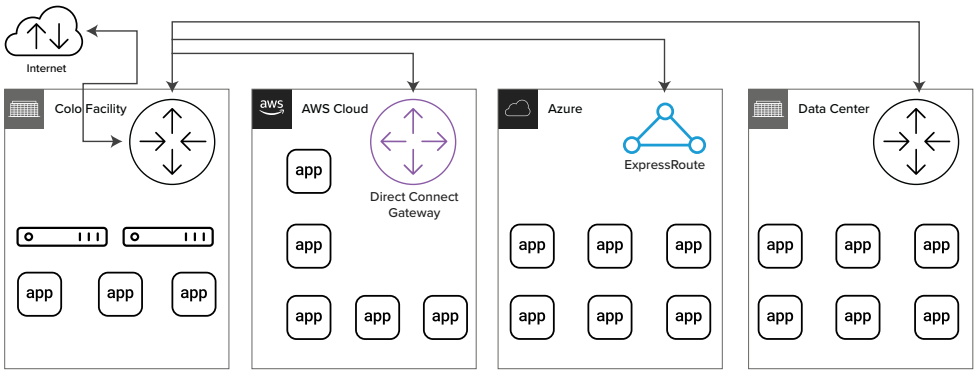
Organizations that leverage a shared VPC model expose subnets to different accounts. For example, there could be two application accounts that each have access to 2 unique subnets per AZ (four total) and the BIG-IP would leverage its own subnets using the workloads from the application accounts as pool members. Please refer to F5 Security for ELB and Direct to EC2 Deployed Applications.

Cloud Interconnect–BIG-IP

APPLICABLE PATTERNS: Active-Standby, Active-Active

APPLICABLE MODULES: LTM, AFM, Advanced WAF, AFM, SSLO, and APM (Active-Standby or Active-Active only)

Many customers are running in multiple clouds/hybrid situations and are already leveraging network point of presence colocation services. In a scenario where you have network point of presence colocation it is possible to place vertical scale system use cases external to AWS while leveraging network connection into AWS and other services such as service discovery to locate resources. This pattern allows users to leverage F5 hardware and be in proximity to multiple cloud providers at the same.



Migration and Scaling Considerations

Organizations that are migrating F5 from on-premises to the public cloud may be faced with a descaling situation. Descaling means that they have either large configurations or large traffic volumes that need to be ported from accelerated hardware systems to virtualized software systems. A review of the systems and understanding the scaling dimensions are required prior to finalizing your cloud architecture.

Configuration objects that need to be evaluated to migrate to AWS:

Number of Virtual Servers with Unique Ips	Number of VIPs that must be mapped to Public IPs in AWS	Non-AWS bound traffic (Internet, VPN or Direct Connect)
Total Network Traffic	Licensed and Enabled modules	SSL TPS
Number of VLANS/ Network Topology		

Once the sizing exercise is complete than an architecture exercise may need to happen. Depending on how you connect to Amazon VPCs from outside of AWS and how you connect between VPCs will create the topology, which intern creates the final sizing. Please see DevCentral for mor information about this topic.

Encrypted Disk Images

Customers that require the use of encrypted disk images in AWS have two options based on how they are consuming and using F5 software. All of these steps are achievable via the CLI and API. Screen shots are used to reinforce the concepts. Please see DevCentral for more information about this solution.

Support for AWS CloudHSM

For customers that need to support FIPS 104-2 L2. BIG-IP supports the use of AWS CloudHSM. The process of installing the CloudHSM client can be found in this article. Please see DevCentral for more information on this integration.

Support for AWS CloudHSM–NGINX

On Linux, the NGINX integrates with OpenSSL to support HTTPS. The AWS CloudHSM dynamic engine for OpenSSL provides an interface that enables NGINX to use the HSMs in your cluster for cryptographic offloading and key storage. The OpenSSL engine is the bridge that connects the NGINX to your AWS CloudHSM cluster. For details, see the prerequisites and walkthrough in using AWS CloudHSM with NGINX.

iRules and Public Cloud

Organizations rely heavily on F5 iRules, and the same application logic you have on-premises can be ported to BIG-IP software running in AWS.

TLS Ciphers and Public Cloud

Instances running in AWS no longer have access to FPGA offload for SSL. Organizations that are concerned with SSL capacity and performance should consider using elliptic curve ciphers. RSA certificate/key pairs can still be used if necessary. Please follow this link for details on custom cipher filters. For information on how to generate certificates and certificate signing requests on BIG-IP please follow this link. For performance testing results please work with your F5 solutions engineer.

F5 Validated Instance Types in AWS

F5 supports an array of instance types in AWS. We can broadly look at these as compute optimized (C 3/4/5/6) and memory optimized (M/3/4/5/6) combined with network optimized (c5n compared to a c5). The most accurate instance selection and sizing will come from engaging with your F5 solutions engineer, but the following rules of thumbs can be used for guidance.

1. If using LTM C types are sufficient, they support 2:1 GB RAM:CPU core and are offered in different sizes, with different network characteristics.
2. If using Advanced WAF, AFM, APM, SSLO, or others M types would be recommended as they support 4:1 GB RAM:CPU cores

In all scenarios network optimized instances (*'n'* designation) should be considered as these systems have superior network performance compared to their similarly named peers.

Network Bandwidth—AWS lists multiple ways in which bandwidth may be restricted. Please familiarize yourself with them as it could be material to your deployment and traffic patterns.

Cloning Instances and Creating Organization Specific AMIs

F5 has regular releases of our software in AWS, but we do not release all builds as an Amazon Machine Image (AMI). For customers that have specific version builds and want to create a standard template they can support cloning of images no matter how they are licensed.

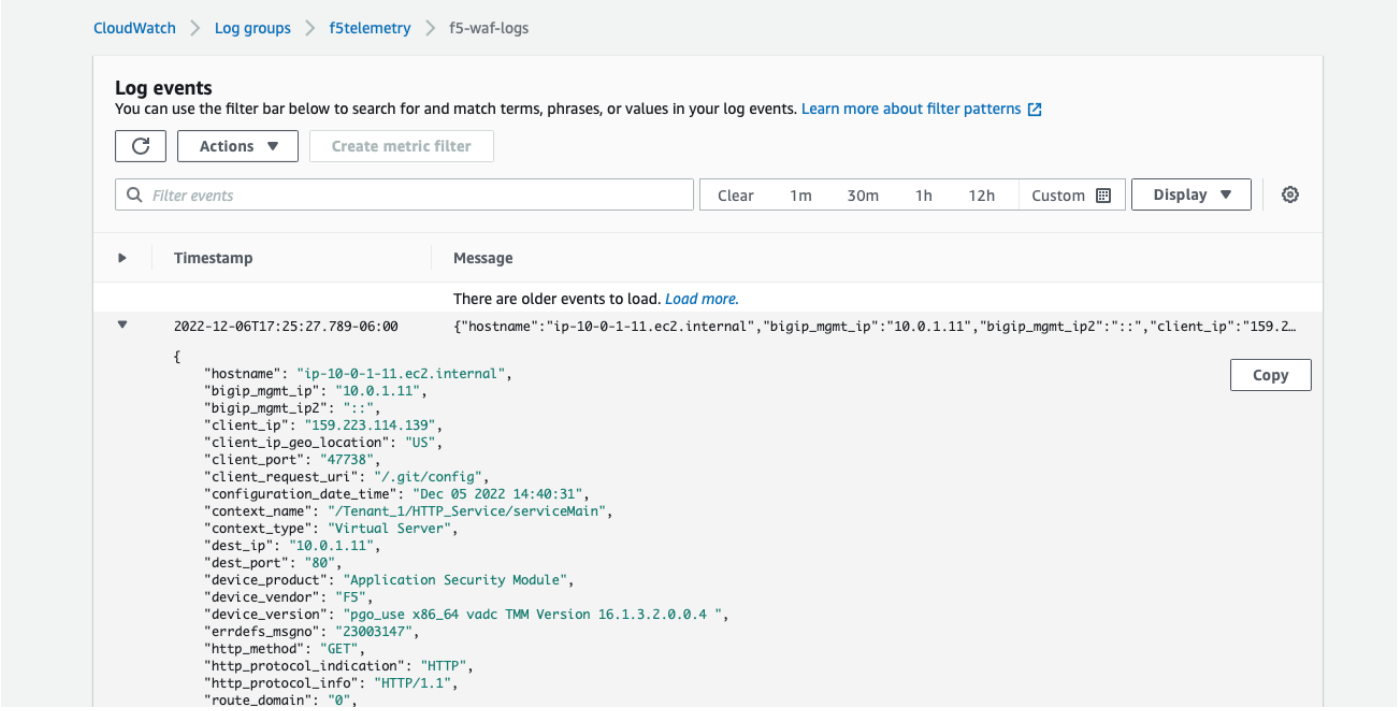
Customers that are using pay-as-you-go based instances will need to prepare an instance with the proper code versions and then run the steps to remove unique attributes from BIG-IP as listed in this knowledgebase article: https://my.f5.com/manage/s/article/K44134742

Once the instance is shut down follow the standard steps as documented by AWS.

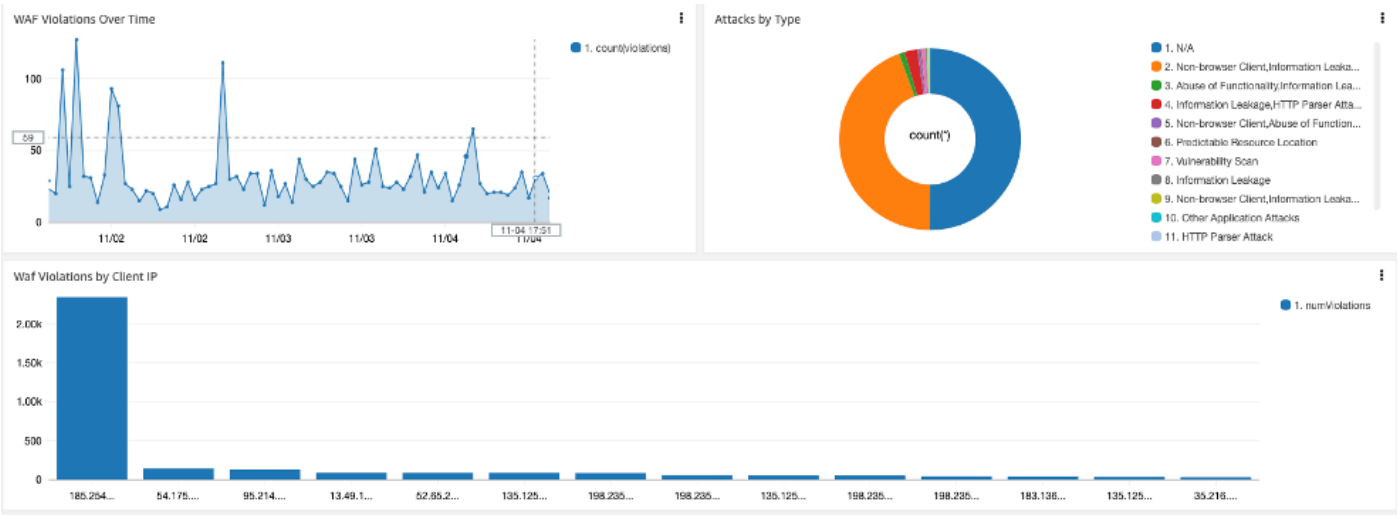
For customers that leverage perpetual or F5® BIG-IQ® Centralized Management licenses systems they can generate their own images with the F5 image generator and upload them to AWS: https://my.f5.com/manage/s/article/K18908626

F5 BIG-IP Logging and Visibility in AWS

F5 customers are diverse in where and how they want to consume logs. For customers that want to leverage AWS CloudWatch we can use F5 BIG-IP Telemetry Streaming (BIG-IP TS) to send logs to CloudWatch logs or CloudWatch metrics. When we consider an environment that is “all” in AWS this makes sense allowing the construction of dashboards that also display the relevant adjacent systems and services in the AWS console.



In the image below we have constructed a dashboard from BIG-IP AFM data allowing use to see the volume of violations over time, the types of violations and the volume of attacks from different source IP addresses.



Example Log Data

CloudWatch > Log groups > f5telemetry > f5-waf-logs

Log events

You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

Actions | Create metric filter

Filter events | Clear | 1m | 30m | 1h | 12h | Custom | Display

Timestamp	Message
	There are older events to load. Load more.
2022-12-06T17:25:27.789-06:00	{"hostname":"ip-10-0-1-11.ec2.internal","bigip_mgmt_ip":"10.0.1.11","bigip_mgmt_ip2":"::","client_ip":"159.2…

Copy


```
{
    "hostname": "ip-10-0-1-11.ec2.internal",
    "bigip_mgmt_ip": "10.0.1.11",
    "bigip_mgmt_ip2": "::",
    "client_ip": "159.223.114.139",
    "client_ip_geo_location": "US",
    "client_port": "47738",
    "client_request_uri": "/.git/config",
    "configuration_date_time": "Dec 05 2022 14:40:31",
    "context_name": "/Tenant_1/HTTP_Service/serviceMain",
    "context_type": "Virtual Server",
    "dest_ip": "10.0.1.11",
    "dest_port": "80",
    "device_product": "Application Security Module",
    "device_vendor": "F5",
    "device_version": "pgo_use x86_64 vadc TMM Version 16.1.3.2.0.0.4 ",
    "errdefs_msgno": "23003147",
    "http_method": "GET",
    "http_protocol_indication": "HTTP",
    "http_protocol_info": "HTTP/1.1",
    "route_domain": "0",
```

Automation and F5 APIs

https://clouddocs.f5.com/cloud/public/v1/shared/cloudinit.html

https://clouddocs.f5.com/products/extensions/f5-appsvcs-extension/latest/

AWS Cloud Integrations APIs and Logging

https://clouddocs.f5.com/products/extensions/f5-cloud-failover/latest/

https://clouddocs.f5.com/products/extensions/f5-telemetry-streaming/latest/

F5 to AWS Service

This chart allows you to navigate questions about AWS services and align to different F5 offers based on different characteristics of the offer such as form factor (VM, software, SaaS) and commercial model.