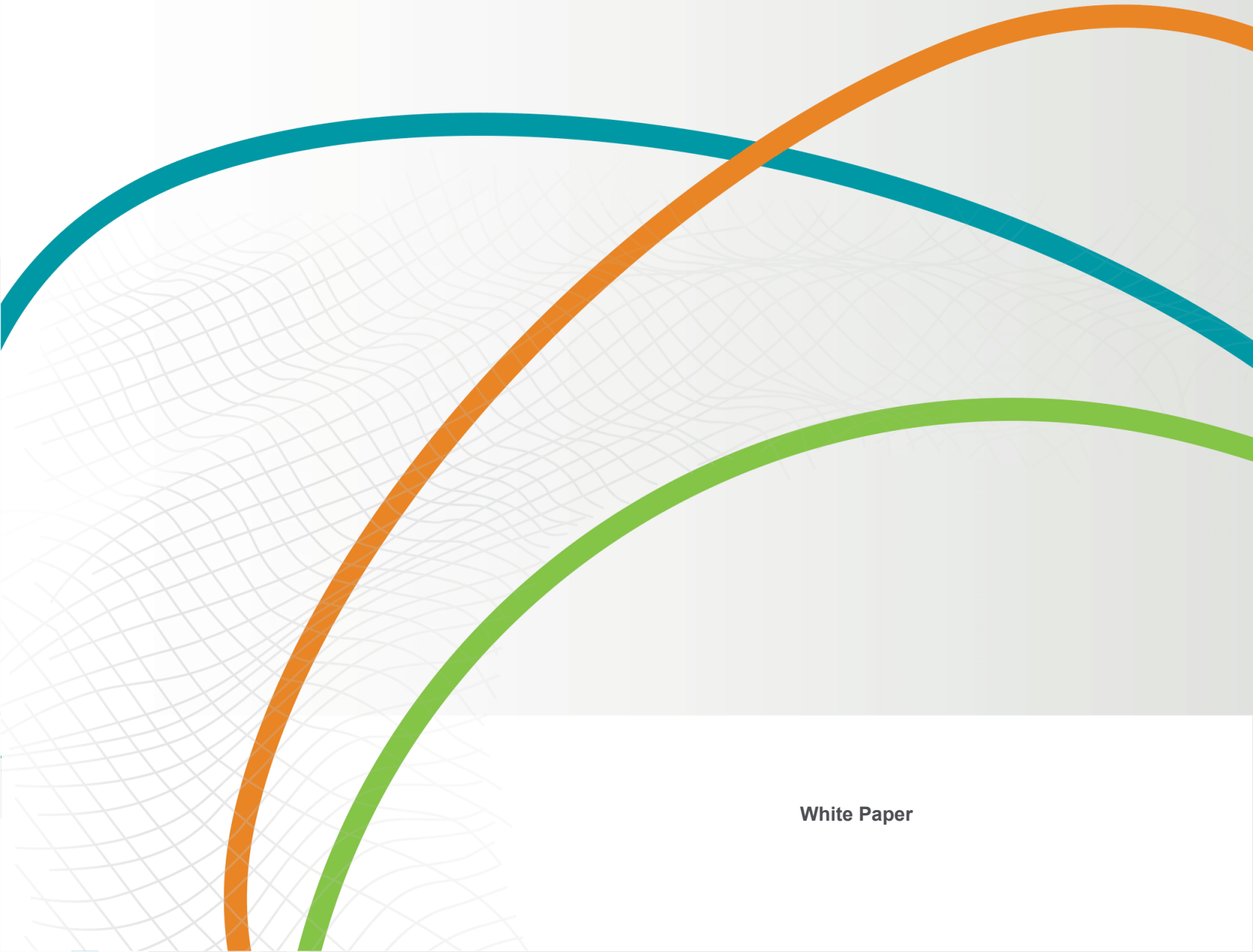
BIG-IP Cloud Edition Solution Guide

Introduction F5® BIG-IP® Cloud Edition™ was built to help network operations teams and applications teams collaborate more effectively in the rapid delivery of secure, appropriately supported applications. BIG-IP Cloud Edition...



White Paper

Introduction

F5® BIG-IP® Cloud Edition™ was built to help network operations teams and applications teams collaborate more effectively in the rapid delivery of secure, appropriately supported applications. BIG-IP Cloud Edition simplifies and centralizes core device and app services management functions like setup, licensing, upgrades, analytics, and scaling. Operations teams can easily define a self-service catalog of application services that developers can then access, on demand, via a dashboard or API call. These services are defined, updated, and deployed for each application in contrast to the traditional, consolidated model in which a single Application Delivery Controller (ADC) supports multiple applications.

As well as bringing a new level of architectural flexibility to enterprise-class application delivery and security services, BIG-IP Cloud Edition also has several how-to-buy options. Designed to provide financial flexibility to match service flexibility, BIG-IP Cloud Edition is available with subscription, utility, and enterprise license options, as well as a traditional perpetual purchase option.

The BIG-IP Cloud Edition Architecture

BIG-IP Cloud Edition has been specially designed and tested to enable organizations to build an application services delivery solution that offers self-service deployment and scaling—allowing application teams to provide enterprise-grade availability and security for their applications. This approach empowers application owners to better collaborate with NetOps, DevOps, and SecOps within an agile framework to significantly improve the performance, availability, and security of all applications.

BIG-IP Cloud Edition is made up of two infrastructure components: 1) specially licensed BIG-IP Per-App virtual editions (VEs), each dedicated to a single application and 2) F5 BIG-IQ® Centralized Management which provides management, visibility, and licensing services across all instances—no matter where they are located. The auto-scaling solution works in Amazon Web Services (AWS) or VMware vCenter-based private clouds.

Logical Components

BIG-IP Cloud Edition is built on several key logical components:

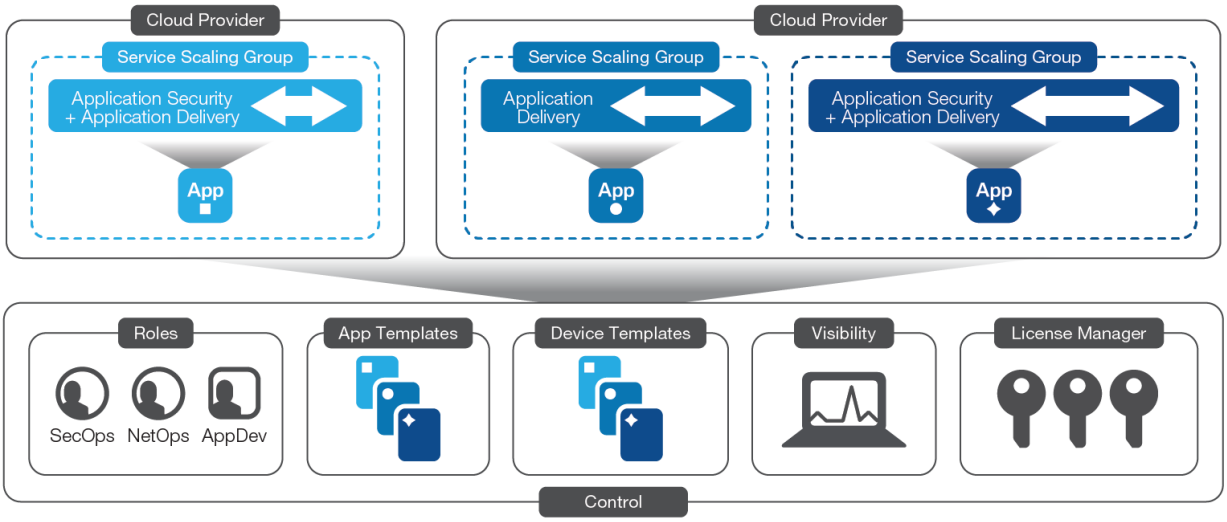


Figure 1: F5 BIG-IP Cloud Edition logical design

Cloud providers

BIG-IP Cloud Edition supports deployment and auto scale of BIG-IP instances on the following cloud platforms:

- Amazon Web Services (AWS)
- VMware vCenter-based private clouds

Support for additional private and public cloud platforms is planned for future releases.

Service catalog

BIG-IP Cloud Edition allows developers to access an on-demand, self-service catalog of application services and to choose an application services deployment template.

Application templates

Application templates define the application delivery and security services that will be deployed for an application, including all BIG-IP objects such as virtual servers, profiles, monitors, SSL certificates, security policies, etc. In addition, the application templates define monitoring and alerting for that application. These templates are typically defined by an application services expert (such as the network and/or security administrator), who configures smart defaults and exposes a limited set of configuration options to the application owner. This simplification removes reliance on network and security operations—and eliminates the need for deep network domain expertise—all while ensuring consistent usage of approved templates and policies in the development and deployment of applications. This results in faster app deployments, as application owners use a simple-to-view dashboard or a single API call to deploy and manage their applications. In addition, BIG-IQ Centralized Management 6.0 comes with a set of predefined templates for common application configurations. Application templates can be delivered by BIG-IP high-availability pairs in non-auto-scaling configurations or by service scaling groups.

Service scaling groups

In addition to using application templates, application teams can take advantage of auto-scaling capabilities by creating a service scaling group. When application services are deployed from an application template and a service scaling group is selected as the target, BIG-IP Cloud Edition manages the availability and elastic scaling of resources to deliver the services, plus manages the lifecycle and upgrade process for the BIG-IP devices delivering those services. Service scaling groups have policy definitions of the minimum and maximum numbers of devices in a group, and the triggers to be used to scale resources.

It is also possible to use application templates from the service catalog to deploy services onto traditional F5 ScaleN® clusters (but without the scaling and lifecycle management benefits).

Device templates

Device templates define all infrastructure-level characteristics (time zone, DNS, hostname, accounts, NTP, licensing, networking, etc.) that are required to deploy a BIG-IP device. Organizations can use device templates are used to create service scaling groups by deploying several new devices using these templates. Device templates are also the primary method for interacting with the BIG-IP devices; if a change to a device in BIG-IP Cloud Edition is required (due to a version upgrade, for example), then the device template is changed and the changes are pushed out to the service scaling group. Device templates contain all of the information required to instantiate a BIG-IP virtual edition, including licensing, provisioning, networking, and other basic device needs.

Device management is different in BIG-IP Cloud Edition.

In most cases, the devices providing application delivery and security services in BIG-IP Cloud Edition are immutable; changes are not made directly to the device configurations but rather to the device template. Then, BIG-IP Cloud Edition rolls these changes out into a service scaling group by deploying new devices, switching traffic to them, and then removing the old devices. This process—sometimes called “BIG-IP Per-App VE and nuke”—is fundamentally different from how a traditional multi-tenant BIG-IP deployment is managed.

Benefits of per-app services:

- Isolate workloads
- Right-size virtual environments
- Move away from traditional in-place upgrades
- Automatically provision and configure new instances

License management

With BIG-IP Cloud Edition, granting, upgrading, and revoking licenses for virtual instances is handled automatically by the license management system in BIG-IQ. This automatic system allows licenses to be pooled and deployed when and where they are needed. When a device is no longer needed, its license is returned to the pool for use by another instance. Although the licensing modes, capabilities, and throughput might vary between deployments, BIG-IQ handles the licensing seamlessly, so there is no need for cumbersome manual license activation when deploying new BIG-IP instances.

Application visibility

To provide triggers for scaling events and deep insight into application and infrastructure performance, BIG-IQ collects and visualizes application-level analytics that are useful for security and network administrators, as well as for application owners. This visibility helps app owners self-diagnose application performance problems to determine whether their application or the network is the source of delays.

Roles

BIG-IP Cloud Edition is designed to drive logical separation of roles. Application owners get self-service application deployment into service scaling groups managed by infrastructure owners. The templates and security policies that the application owners use can be managed by the NetOps and SecOps teams. Some templates might be available to some application owners and not others, and the per-application statistics and dashboards can be restricted to the application owners. Through fine-grained role management, BIG-IP Cloud Edition empowers application teams to support their apps, while enabling operations teams to maintain control over the network.

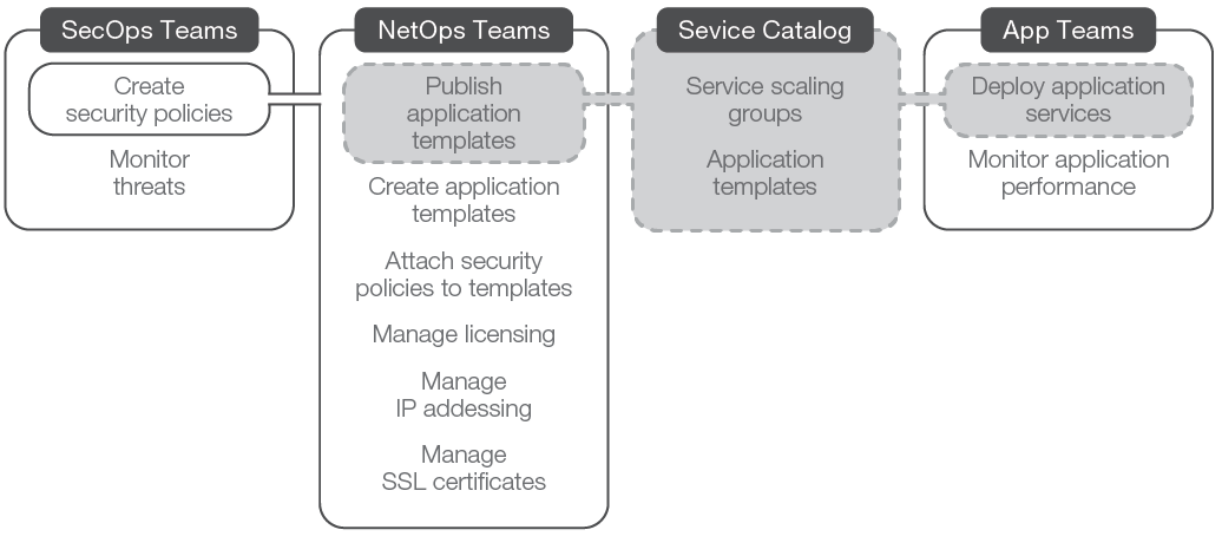


Figure 2: BIG-IP Cloud Edition roles and tasks: The security operations team can build a library of security policies to cover most common application deployments. These policies are then attached to the application templates by the network operations teams, who —in addition to creating non-security-focused application services policies—are also responsible for building the service scaling groups and performing general device and license management tasks. The application teams consume these services by selecting application templates to deploy application services for their apps, and then selecting a service scaling group onto which to deploy these services.

Infrastructure Components

BIG-IP Cloud Edition is composed of several different infrastructure components that work together to deliver the solution.

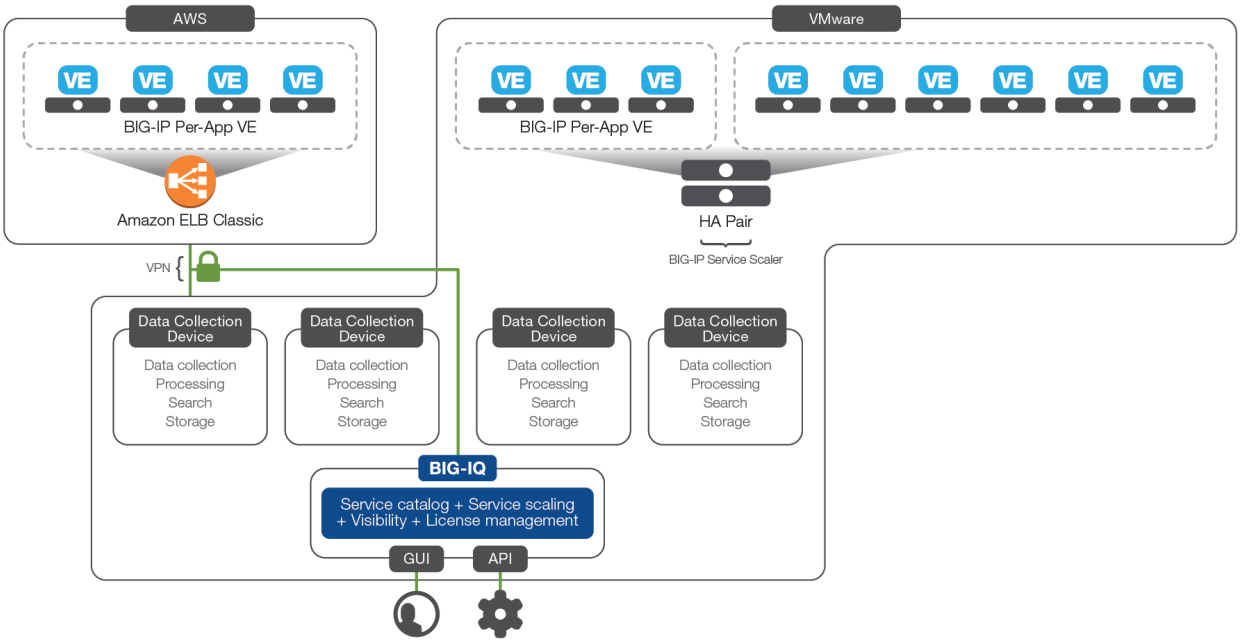


Figure 3: BIG-IP Cloud Edition physical design

BIG-IP Per-App VE

Using BIG-IP Per-App VE outside of BIG-IP Cloud Edition. BIG-IP Per-App VEs can be purchased outside of BIG-IP Cloud Edition. Available as a bundle of licenses, and with a free BIG-IQ license manager component, BIG-IP Per-App

A BIG-IP Per-App VE is a specially licensed BIG-IP instance that has been designed to provide dedicated services for a single application. The full features of BIG-IP software are enabled, but it is right-sized for use as a dedicated device.

Each BIG-IP Per-App VE comes with:

- Single virtual IP address
- Three virtual servers (a combination of a virtual address and a listening port)
- 25 Mbps or 200 Mbps throughput

There are two software module options available in BIG-IP Per-App VEs:

BIG-IP Local Traffic Manager

F5 BIG-IP Local Traffic Manager™ (LTM) software delivers industry-leading application traffic management, including advanced load balancing, rate shaping, content routing, SSL management, and complete control of the application layer traffic in both directions.

F5 Advanced WAF

F5 Advanced WAF offers all the features of a traditional web application firewall (WAF) plus enhanced protection in the form of layer 7 DDoS mitigation, advanced bot detection, and API security management. Advanced WAF comes with a set of BIG-IP LTM traffic management features to effectively manage traffic to downstream application servers. The deployment of Advanced WAF policies are managed as part of the application template component.

Virtual machine requirements

BIG-IP Per-App VEs benefit from the platform streamlining of image and disk sizes that has occurred in recent releases of BIG-IP. In traditional BIG-IP deployments, BIG-IP software versions were done “in place” by downloading a new software image onto a running device then following an upgrade procedure. With BIG-IP Cloud Edition, the devices providing application delivery and security services are, for the most part, immutable so changes are not made directly to the device configurations, but rather are deployed using the device and application templates. Then the old versions are retired in a rolling upgrade. Additional storage space for multiple versions of the BIG-IP software is not, therefore, required and the disk image size can be shrunk.

- In VMware deployments, BIG-IP Per-App VEs are available in non-upgradable images with reduced storage footprints. For details on virtual machine specifications in VMware see the Virtual Edition Setup Guide for ESXi.
- For production use on AWS, F5 recommends M3 or M4 image types, with a minimum of two virtual cores and 4 GB of memory for BIG-IP LTM deployments and 8 GB for Advanced WAF.

Scaling and managing BIG-IP Per-App instances in a service scaling group

VMware–BIG-IP service scalers

In VMware, per-app traffic to BIG-IP Per-App VEs is scaled via a specialized BIG-IP cluster using MAC address forwarding, which preserves the client source and destination IP addresses. This can be important for some of the layer 7 functionality offered by the BIG-IP Per-App VEs, and also ensures accurate data collection for the visibility services that BIG-IQ offers.

BIG-IP service scalers perform basic load balancing across BIG-IP Per-App VEs and have no license limit on throughput (however, virtual hardware resources will obviously limit maximum throughput). Optionally, the service scaler can be enabled with firewall capabilities offering network ACLs and layer 4 DoS mitigation capabilities. The service scalers cannot perform SSL or layer 7 functions at this time.

BIG-IP service scalers require the following virtual machine specifications:

	Minimum	Maximum
vCPU	2[1]	4
Memory	4 GB	16 GB[2]
Disk Space	40 GB[3]	82 GB
Network Interface Cards	4	10

BIG-IP service scalers can belong to more than one service scaling group and can be shared across multiple applications (while BIG-IP Per-App VEs are—as the name suggests—dedicated to a single application).

Setting up and configuring service scalers in a service scaling group is covered in BIG-IQ Centralized Management: Local Traffic & Network Implementations.

[1] Four vCPUs required for additional firewall functionality.

[2] This can be higher as there is not a fixed limit.

[3] 82 GB for firewall features.

AWS ELB Classic

In AWS, services are scaled using Elastic Load Balancing (ELB) Classic instances. ELB Classic provides basic L4 load balancing and availability across BIG-IP Per-App VEs, and a logical instance of ELB is dedicated to a single service scaling group. Each application therefore requires a dedicated ELB configuration. The AWS service manages the scaling of ELB instances to meet demands.

Setting up AWS ELB instances in a service scaling group is covered in BIG-IQ Centralized Management: Managing Applications in an Auto-Scaled AWS Cloud.

BIG-IQ

BIG-IQ can manage more than BIG-IP Per-App VEs.

BIG-IQ can discover and manage BIG-IP instances of all supported software versions—no matter what the platform or location. The platform can perform device management, visualize statistics, and deploy templated application service configurations onto physical, virtual, and cloud-deployed BIG-IP instances. BIG-IQ can even offer autoscaling for supported, traditional (not per-app) BIG-IP VEs on supported platforms (currently AWS and VMware).

BIG-IQ provides centralized management for all components that make up BIG-IP Cloud Edition. All activities and reporting are managed via BIG-IQ and administrative access to BIG-IP Per-App VEs is not required.

BIG-IQ:

- Creates new service scaling groups
  - Within the service scaling group, device templates are referenced to manage the life cycle of BIG-IP Per-App VEs. Device templates include all of the information needed to spin up a BIG-IP Per-App VE and requires no human intervention.
- Provides deep analytics at the application level so that application owners can troubleshoot their own issues.
- Provides device-level performance and capacity metrics for trouble shooting and planning.
- Offers role-based access allowing application owners to deploy F5 L4–7 services for an application via predefined application templates from the service catalog in a self-service manner.

F5 recommends the following virtual hardware for BIG-IQ in a BIG-IP Cloud Edition deployment.

	Minimum	Maximum
vCPU	4	8
Memory	4 GB	16 GB
Disk Space	95 GB	500 GB
Network Interface Cards	2	10

Installing and configuring BIG-IQ is covered in the Planning and Implementing an F5 BIG-IQ Centralized Management Deployment Guide.

BIG-IQ communication with virtual infrastructure management

BIG-IQ is capable of starting, licensing, provisioning, and configuring BIG-IP Per-App VEs on demand, as part of a service scaling group or in a scale-out environment. This requires authenticated access into the virtual infrastructure environment.

In VMware

In VMware, the following is required: credentials to access vCenter, a vCenter hostname, SSL certificate for secure communication, and other information about the ESX environment such as hosts/clusters, datastores, (distributed) virtual switches (vSwitches), and resource pools.

In AWS

In AWS, the following is required: Identity and Access Management (IAM) user access key and associated secret to make API calls and ELBs to provide tier-one traffic distribution. Follow AWS best practices to create and manage the keys.

The IAM user should have the administrator access policy attached and have permission to create auto-scaling groups, Amazon Simple Storage Service (S3) buckets, instances, and IAM instance profiles. For details on permissions and overall AWS configuration, see https://aws.amazon.com/documentation

BIG-IQ high availability and backup

Since BIG-IP Cloud Edition essentially routes all control plane activities through the BIG-IQ management layer—BIG-IQ handles real-time monitoring and scale-in/out events and manages license assignment and revocation—it becomes a critical part of the delivery system and therefore is typically deployed in a highly available, redundant configuration.

Planning should, therefore, include an active-standby BIG-IQ pair, with the appropriate license for the number of BIG-IP instances under management.

Configuring BIG-IQ for high availability is covered in the Planning and Implementing an F5 BIG-IQ Centralized Management Deployment Guide.

BIG-IQ Data Collection Devices

Data Collection Devices in BIG-IQ are responsible for collecting, storing, and processing traffic and performance data from the BIG-IP Per-App VEs. After BIG-IP Per-App VEs send performance and traffic telemetry to Data Collection Devices to process and store, BIG-IQ queries the Data Collection Devices to provide visibility and reporting. Data Collection Devices are arranged into clusters that work together and replicate stored data for redundancy purposes.

F5 recommends the following virtual hardware for Data Collection Devices used in BIG-IP Cloud Edition:

vCPU	8
Memory	32
Disk Space	500 GB
Network Interface Cards	2

A note on disk subsystems: BIG-IQ Data Collection Devices store, process, and analyze data collected from BIG-IP Per-App VEs to produce reports and dashboards for the BIG-IQ system. This is a disk I/O intensive workload, so the underlying storage should be sized for both capacity and performance. For large deployments of BIG-IP Per-App VEs or extensive logging and analysis, high-performance storage subsystems should be deployed. Capture search and indexing operations will generate both random and sequential I/O often with high concurrency of tasks.

For additional information see the BIG-IQ Centralized Management Data Collection Devices Sizing Guide.

Networking and Connectivity

VPN for Data Collection Devices with AWS deployments

When new BIG-IP Per-App VEs are created, they are given the self-IP address of the Data Collection Devices they should connect back to. This is a fixed setting (as of BIG-IQ 6.0). Connections in both directions are required between the Data Collection Devices and the BIG-IP Per-App VE. In many environments—but especially when BIG-IP Per-App VEs are on AWS and BIG-IQ and Data Collection Devices are on the customer premises—VPN connectivity will be required to successfully route traffic in both directions, since the Data Collection Devices will generally have an RFC 1918 non-routable IP address. BIG-IP Cloud Edition requires unique IP address ranges across Amazon Virtual Private Cloud (Amazon VPC), meaning that they cannot have overlapping address spaces in Amazon VPC.

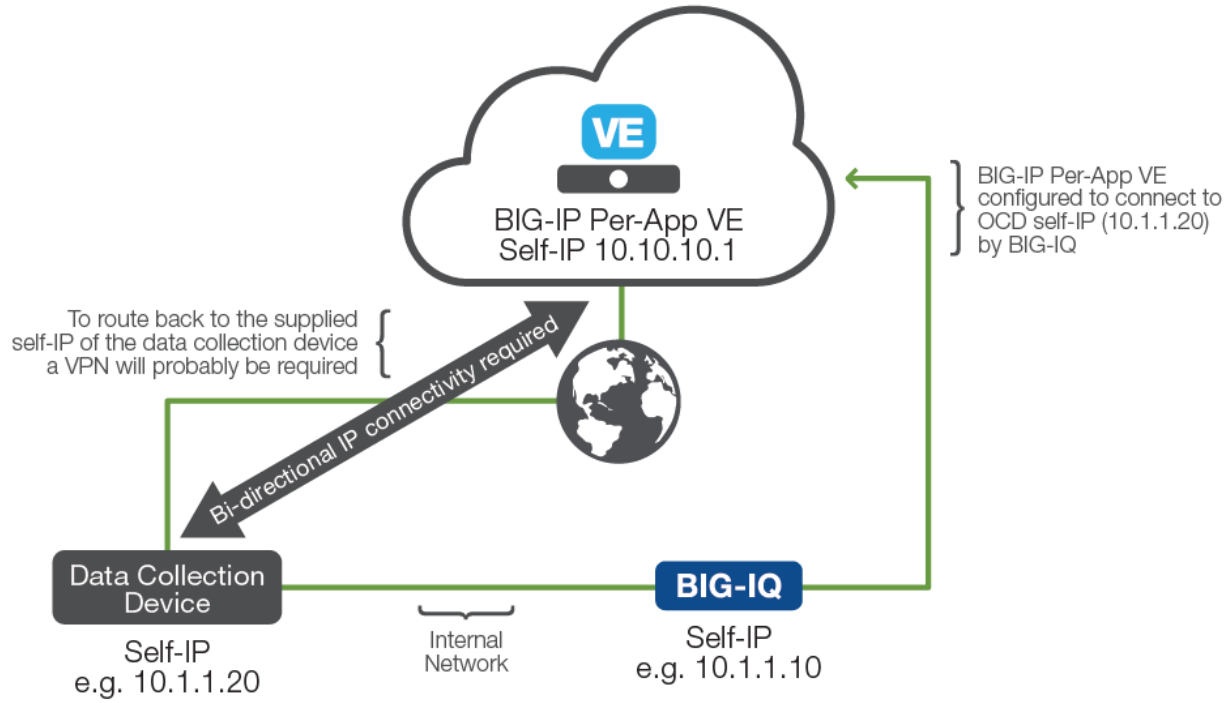


Figure 4: VPN requirements for AWS deployment

There are several different ways to set up a VPN or other private connection to AWS, including services like AWS Direct Connect, or multi-cloud connectivity services like Equinix Cloud Exchange. It is also possible to establish an IPSEC tunnel from on-premises BIG-IP devices to AWS VPN gateways.

Connectivity: Ports and protocols

For port and protocol connectivity details between BIG-IP Cloud Edition components, please see the BIG-IQ 6.0 documentation.

In addition, BIG-IQ needs access to AWS API endpoints for the chosen region on port 443 or to the vCenter server on port 443.

Authentication and Security

BIG-IQ user authentication

BIG-IQ offers both built-in user account management and integration with common external protocols such as TACACS, RADIUS, and LDAP.

Sizing and Capacity Planning

How many BIG-IP Per-App VEs will be needed?

There are two critical limits on BIG-IP Per-App VE instances:

- Objects
- Throughput

Unlike a more traditional deployment, BIG-IP Per-App VEs are generally deployed in an all-active configuration with the tier-one traffic management device taking care of high availability and scaling. Generally, this means more real throughput per provisioned VE instance than in a more hardware-centric, active-standby high-availability (HA) pair where it is necessary to maintain spare capacity for failover, even in an active-active configuration. BIG-IP Per-App VEs are available in 25 Mbps and 200 Mbps throughput licenses and are designed to scale out using service scaling groups.

The first step is determining a base estimate of the required throughput for each application for which traffic management is planned. Next, decide whether the 25 Mbps or 200 Mbps license is appropriate. For larger throughput requirements per application, the 200 Mbps license is appropriate due to having fewer overall devices. For smaller or more fine-grained requirements, the 25 Mbps license is more appropriate. If maximizing use of resources is important, then the 200 Mbps license will make more efficient use of underlying hardware for a particular throughput.

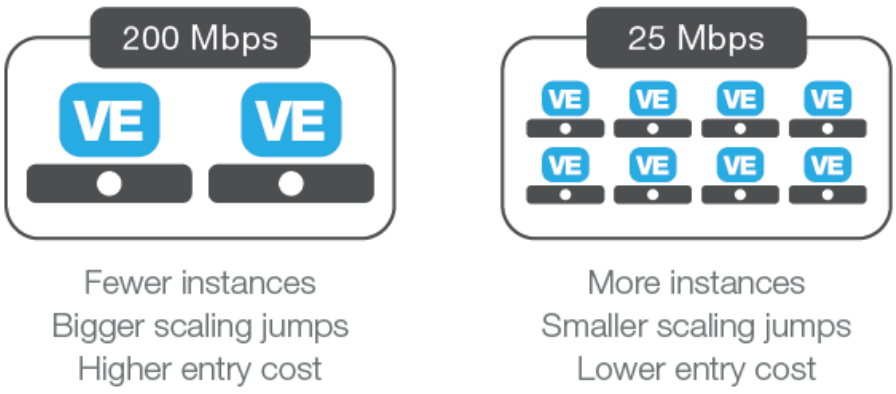


Figure 5: Scaling and sizing BIG-IP Per-App VEs

It is possible to mix and match license types within an environment, but a specific application will only be serviced by one license type.

For each application, determine:

- Total throughput required
- Likely growth
- Volatility

When thinking of volatility, there are a few variables to consider. First, the triggers for scaling events are based on the throughput, CPU thresholds, and/or memory of the busiest device in a service scaling group, taken over a five-minute average. Because a new BIG-IP Per-App VE instance will take a brief amount of time to become active after startup it is recommended to provision the base requirement with capacity for approximately 20 minutes of maximum expected growth. That way, services can flex with demand while still having some capacity to handle expected spikes.

Although sizing can be complex, with BIG-IP Cloud Edition each application's instances can flex on demand, so aiming for perfection—or building in excessive spare capacity—is not necessary.

Sizing BIG-IP service scaler instances

Service scaling capability is implemented differently between environments.

In AWS, scaling in and out is handled by the simple-to-use AWS ELB Classic load balancer.

In VMware, the service scaling, layer 4 DDoS, and firewall functions are provided by special BIG-IP VEs. These VEs are configured to simply distribute traffic to the BIG-IP Per-App VEs and also to provide network layer access control and DDoS mitigation.

The auto-scale instances are designed to be high throughput, low complexity, and shareable between multiple applications.

API and integrations

BIG-IQ offers a REST API, enabling the deployment of application services from the service catalog programmatically. For details on the REST API, please see the BIG-IQ documentation.

Conclusion

BIG-IP Cloud Edition delivers the power, security, and flexibility of F5 application services in new ways. These include a new per-app platform that can scale up and down on demand, and self-service capabilities. Security teams can create application security and DDoS mitigation policies, and network teams can then attach these to application templates and add them to the service catalog for specific users. Application teams can choose from a pre-defined services catalog and deploy services into a service scaling group that will flex to meet their app requirements.

The end result is a highly flexible, scalable solution that delivers enterprise-grade F5 application services—with the flexibility of dedicated instances—accompanied by a measurable reduction in operational overhead. This approach empowers the right teams to do the right work: now application owners can better collaborate with network, development, and security operations teams within an agile framework to significantly improve the performance, availability, and security of all applications.

F5 Networks, Inc.
401 Elliott Avenue West, Seattle, WA 98119
888-882-4447 f5.com

Americas
info@f5.com

Asia-Pacific
apacinfo@f5.com

Europe/Middle-East/Africa
emeainfo@f5.com

Japan
f5j-info@f5.com