F5 Public Cloud Design Patterns

Designing F5 security, optimization, and availability services into Infrastructure as a Service scenarios
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>4</td>
</tr>
<tr>
<td>Public Cloud: IaaS</td>
<td>5</td>
</tr>
<tr>
<td>Key Infrastructure Components in IaaS Systems</td>
<td>5</td>
</tr>
<tr>
<td>F5 Ready Cloud Providers</td>
<td>8</td>
</tr>
<tr>
<td>F5 Application Delivery Services</td>
<td>8</td>
</tr>
<tr>
<td>The BIG-IP Platform</td>
<td>10</td>
</tr>
<tr>
<td>How the BIG-IP Platform Works</td>
<td>11</td>
</tr>
<tr>
<td>Designing BIG-IP Services into Public Cloud IaaS</td>
<td>12</td>
</tr>
<tr>
<td>System Architectures</td>
<td></td>
</tr>
<tr>
<td>Management networks</td>
<td>12</td>
</tr>
<tr>
<td>Application traffic networks</td>
<td>13</td>
</tr>
<tr>
<td>Routed deployment</td>
<td>13</td>
</tr>
<tr>
<td>One-armed deployment</td>
<td>14</td>
</tr>
<tr>
<td>Using SNAT</td>
<td>14</td>
</tr>
<tr>
<td>Multi-tenant or per-application instances</td>
<td>15</td>
</tr>
<tr>
<td>Local availability options</td>
<td>17</td>
</tr>
<tr>
<td>Global availability</td>
<td>21</td>
</tr>
<tr>
<td>Scaling</td>
<td>23</td>
</tr>
<tr>
<td>Secure cloud connectivity</td>
<td>27</td>
</tr>
<tr>
<td>Security Considerations</td>
<td>29</td>
</tr>
<tr>
<td>Deployment Automation and Templates</td>
<td>29</td>
</tr>
<tr>
<td>REST API</td>
<td>29</td>
</tr>
<tr>
<td>Application templates</td>
<td>29</td>
</tr>
<tr>
<td>Integration with cloud and third-party tools</td>
<td>30</td>
</tr>
<tr>
<td>Conclusion</td>
<td>30</td>
</tr>
<tr>
<td>Resources</td>
<td>30</td>
</tr>
</tbody>
</table>
Introduction

By designing F5 application delivery services into public cloud Infrastructure as a Service (IaaS) offerings, organizations can achieve both the economic and scaling advantages of IaaS clouds and the security, high availability, and optimization services of the F5® BIG-IP® platform. A handful of key use cases embrace most situations for combining F5 application delivery services with IaaS. Within those use cases, a number of design patterns provide options for delivering services. These design patterns are worth considering as organizations plan for an architecture that will best meet their needs.

These design patterns are deliberately cloud provider agnostic. Although F5 application delivery services are available in many public cloud offerings, the overall architecture designs will not differ significantly for different providers, although implementation details will. Implementation details for specific IaaS offerings will be described in a forthcoming series of F5 “Recommended Practices” documents for specific providers.
Public Cloud: IaaS

Public IaaS offerings such as Amazon Web Services or Microsoft Azure combine on-demand infrastructure, (almost) limitless capacity, and a range of computing, storage, application, and networking services. By combing these services with application software, customers can build agile, flexible IT infrastructures to deliver a wide range of applications. The flexibility of utility billing, combined with the ability to scale up and down on demand, means you can deliver these services at a considerable CapEx savings, while the availability of automation and templating tools saves operational time and money.

Despite the considerable growth in IaaS services, both in terms of revenue and breadth of services, organizations still need to ensure:

- Security.
- Availability.
- Performance.

F5 application delivery services offer a suite of tools designed to mitigate these risks and to ensure that applications are fast, secure, and available, no matter where they run. These tools help transform the service-centric offering of the public cloud into the application-centric model that powers modern data center design.

Key Infrastructure Components in IaaS Systems

Most public IaaS platforms share some common functional components, even if the names and implementation details differ. Understanding these components and their high-level roles will make it easier to make informed architectural choices.

Virtual networks

Cloud offerings usually include a private virtual network within which customers can create multiple subnets for further segregating traffic. This allows the use of RFC1918 private IP address spaces as opposed to addresses that are publicly routable. Such traffic segregation can make it considerably easier to migrate an existing application infrastructure, since you may be able to use the existing IP addressing schema. The private networks allow you to build an infrastructure analogous to the DMZ/internal network design that is common in data centers.

Virtual subnets

Within virtual networks, administrators can usually configure virtual subnets. Virtual subnets can be marked as private or public and permit control of access to the Internet and other subnets.

In many cloud infrastructures, the virtual network will not expose the underlying layer 2 functionality. Any mechanism relying on features like ARP or cloning traffic to a particular MAC address will need significant alteration in a cloud environment.
Public IP mapping

Virtual networks and virtual subnets contain only RFC1918 non-routable subnet addresses. Cloud infrastructures provide tools to map public IP addresses to internal private addresses. Sometimes you can do this simply by mapping a public IP address to an internal interface. In other systems both the public IP and port can be mapped. For example, a public IP might be mapped to different servers for ports 443 and 25.

Internet gateways

Internet gateways enable Internet access from within the virtual network subnets. Sometimes this Internet gateway is an explicit infrastructure component, and you will need attach it to a private network. Other times, the gateway is an implicit part of the cloud infrastructure, and is merely represented by a “next hop” in the routing table.

Routing tables

Default routing tables are usually created when creating virtual networks and subnets, but you can often override the default with your own user-defined routing table. This lets you route traffic to specific subnet through devices—such as an Application Delivery Controller—to provide security, acceleration, or other application delivery services.

Network security

Different clouds implement network security features in a variety ways, but all seek to control what traffic can enter the virtual networks, traverse the virtual subnets, or reach individual hosts. These rules are generally stateless, so that matching outbound rules (or a wildcard rule) must be in place to allow outbound traffic. Advanced features such as rules based on source location or reputation are not commonly available in IaaS systems.

VPN gateways

Virtual private network (VPN) gateways provide an IPsec endpoint within a virtual network for connecting an on-premises VPN. This creates an encrypted private network transmitted over the public Internet. Traffic to the private network runs encrypted but over a standard Internet connection. It is important to make sure that the service-level agreement (SLA) and management of the connection match the requirements of the application or business.

Private connections

An alternative to VPN for encrypted connections is a true private network. Cloud providers collaborate with service providers to deploy private links onto customer premises. These private links connect into the cloud provider’s infrastructure, usually through an intermediary or peering point. These high bandwidth links often offer better SLAs and latency than public Internet links, but at greater cost.

Load balancers

A load balancer is a key service within cloud infrastructures. Load balancers in clouds are used to ensure high availability for applications, as well as scaling. In most cloud environments, load balancers can be linked into the cloud provisioning system to enable automatic scaling of servers in load-balanced pools.
Current infrastructure-native load balancers are basic, offering a limited range of protocol support, monitor types, and load distribution algorithms, as well as limited persistence (session affinity) techniques. Most offer SSL termination features and, usually, re-encryption.

In some cases, these load balancers offer sufficient load balancing and availability services. If more advanced traffic management or a wider range of protocol support is required, BIG-IP® Local Traffic Manager™ (LTM) can provide a massive range of functionality, programmability, and application fluency.

Templates and automation

Because business agility is often the key driver for cloud adoption, cloud vendors have invested heavily in creating templates and automation systems that help to streamline the deployment of a single instance or a whole application environment. Cloud systems offer web-based dashboards as well as API access and command-line interface (CLI) tools to manage the virtual infrastructure. Using these and common automation software, you can create a range of tools, from simple monitoring scripts to automated application stack deployments.

A range of consumption models

Cloud vendors offer utility pay-as-you-go billing for services. This model is great for volatile workloads where the ability to flex the required computing power delivers huge savings over paying for peak use 100 percent of the time. For more static, longer-term workloads, it is often possible to pay a yearly fee at a reduced cost. Third-party solutions such BIG-IP products usually offer bring-your-own-license (BYOL) and utility billing options.

For many organizations, a mix and match approach will bring the best return on investment (ROI). Yearly commitments and BYOL billing work well for core components, and base workload and utility billing can be used to provide peak capacity or for short-lived requirements.

Local availability

Local availability services differ by cloud provider. Many limit local availability services to ensuring that multiple servers of a particular architectural layer (such as a tier of web servers) are separated into different racks or devices within the provider’s data center. Amazon Web Services offers a greater degree of separation with Availability Zones—physically separate data centers that are near enough for high bandwidth, low latency links, enabling them to be considered part of the same virtual network.

Global availability

Cloud providers generally locate their data centers in widely separated regions (such as Europe, North America, and Asia). These locations, more analogous to traditional separate data centers, can be viewed as separate cloud installations, rather than an extended virtual cloud network. As a result, they offer greater separation for disaster recovery and business continuity purposes, as well as the ability to position applications or services close to their users.
F5 Ready Cloud Providers

F5 has worked closely with a number of key cloud providers to offer an F5 certification program called F5 Ready. F5 Ready cloud service partners have cloud infrastructures verified by F5 to run BIG-IP images. Most F5 Ready providers offer a range of flexible licensing, including BYOL, on-demand utility billing, or volume licensing subscriptions.

See F5 Ready at f5.com for a current list of F5 Ready partners and their capabilities.

F5 Application Delivery Services

F5 application delivery services are application centric. This means that configuration of those services is performed on a per-application basis, ensuring that each application has the exact combination of services it needs to optimize performance, availability, and security.

Some key (but not exhaustive) examples include:

<table>
<thead>
<tr>
<th>Security</th>
<th>Advanced Network Firewall Services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- Controlling traffic by more criteria than just IP:port:protocol, e.g., geographical location or endpoint reputation</td>
</tr>
<tr>
<td></td>
<td>- HTTP protocol validation</td>
</tr>
<tr>
<td></td>
<td>- Day and time schedules</td>
</tr>
<tr>
<td>Web Application Firewall Services</td>
<td>- Comprehensive tools to identify web application threats and block malicious traffic</td>
</tr>
<tr>
<td></td>
<td>- Outbound data loss prevention (DLP) services</td>
</tr>
<tr>
<td>Access and Identity Services</td>
<td>- Advanced authentication services, e.g., two-factor tokens, captcha, or geographical restrictions</td>
</tr>
<tr>
<td></td>
<td>- Client certificate checking, endpoint inspection</td>
</tr>
<tr>
<td></td>
<td>- SAML service provider (SP) and identity provider (IdP) services</td>
</tr>
<tr>
<td>Denial-of-Service (DoS) Mitigation</td>
<td>- Proactive bot defense</td>
</tr>
<tr>
<td></td>
<td>- Layer 7 DoS detection and mitigation</td>
</tr>
<tr>
<td>SSL and Encryption</td>
<td>- SSL decryption, traffic inspection, and re-encryption</td>
</tr>
<tr>
<td>Availability</td>
<td>Application-Level Monitoring</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>· Advanced application health checks (using a multiple-step monitor)</td>
</tr>
<tr>
<td></td>
<td>· Multi-level health checks (e.g., checking that both the database and application are available)</td>
</tr>
<tr>
<td></td>
<td>· Non-HTTP health checks (e.g., SIP, Microsoft Windows SQL server, FTP)</td>
</tr>
<tr>
<td></td>
<td>· Advanced algorithms to better distribute traffic to the servers functioning best</td>
</tr>
<tr>
<td>Global Availability</td>
<td>· Application availability across a heterogeneous mix of different cloud providers or data centers</td>
</tr>
<tr>
<td></td>
<td>· Integration with BIG-IP advanced monitors</td>
</tr>
<tr>
<td></td>
<td>· DNSSec support</td>
</tr>
<tr>
<td>Performance</td>
<td>Network and Transport Optimization</td>
</tr>
<tr>
<td></td>
<td>· A configurable TCP stack that can be optimized to deliver across WAN and cellular networks</td>
</tr>
<tr>
<td></td>
<td>· An HTTP/2 gateway that brings the advantages of additional compression and request multiplexing without altering the back-end infrastructure</td>
</tr>
<tr>
<td>Application and Data Optimization</td>
<td>· Selective image optimization for on-the-fly optimization depending on detected network or client characteristics</td>
</tr>
<tr>
<td></td>
<td>· WAN acceleration over SSL encrypted tunnels with adaptive compression and TCP optimization</td>
</tr>
<tr>
<td>Flexibility</td>
<td>Data Path Programmability</td>
</tr>
<tr>
<td></td>
<td>· Complete programmatic control of application traffic</td>
</tr>
<tr>
<td></td>
<td>· The ability to read, write, and inspect all aspects of application data</td>
</tr>
<tr>
<td></td>
<td>· Event-driven and comprehensive language</td>
</tr>
<tr>
<td>Control Plane Programmability</td>
<td>· The ability to modify configuration in response to events such as changes to server load, application behavior, or infrastructure</td>
</tr>
<tr>
<td></td>
<td>· Fully autonomous or external API-driven triggers</td>
</tr>
</tbody>
</table>
The BIG-IP Platform

The BIG-IP platform, which is available in physical, hypervisor, and cloud editions, works with BIG-IP software modules to provide F5 application delivery services as noted below. A BIG-IP platform can run one or more software modules to suit the needs of the application, and the platform can be deployed as a stand-alone unit or in highly available clusters. See the discussion of local availability below for more information about high availability and scalability designs.

<table>
<thead>
<tr>
<th>Function</th>
<th>F5 Software Module</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security</td>
<td></td>
</tr>
<tr>
<td>Network layer security</td>
<td>BIG-IP® Advanced Firewall Manager™ (AFM)</td>
</tr>
<tr>
<td>Application layer security</td>
<td>BIG-IP® Application Security Manager™ (ASM)</td>
</tr>
<tr>
<td>Identity and access</td>
<td>BIG-IP® Access Policy Manager® (APM)</td>
</tr>
<tr>
<td>Availability</td>
<td></td>
</tr>
<tr>
<td>Application availability and traffic optimization</td>
<td>BIG-IP Local Traffic Manager (LTM)</td>
</tr>
<tr>
<td>Global availability and DNS</td>
<td>BIG-IP® DNS (formerly BIG-IP® Global Traffic Manager™)</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
</tr>
<tr>
<td>Application and network optimization</td>
<td>BIG-IP® Application Acceleration Manager™ (AAM)</td>
</tr>
</tbody>
</table>
How the BIG-IP Platform Works

The BIG-IP platform is a very high performance, stateful, bidirectional zero copy proxy. Understanding this basic architecture principle can help illuminate how the BIG-IP platform delivers services, which in turn helps clarify architectural choices.

Clients connect to the BIG-IP device or instance, which connects to back-end servers (or in some cases, such as with BIG-IP DNS, handles the application traffic and responds back to the client directly). This creates a TCP “air gap,” with complete TCP session regeneration between the client and the server. Within this logical gap, the BIG-IP platform provides application delivery services. As application traffic transits the platform, it can be inspected, altered, and controlled, so the BIG-IP platform gives complete control of both inbound and outbound application traffic.
Designing BIG-IP Services into Public Cloud IaaS System Architectures

Once you’ve established an understanding of common cloud elements, cloud providers, and F5 application delivery services, you’re ready to delve into deployment design patterns and key options, starting with basic cloud housekeeping and management.

Management networks

Where cloud infrastructures allow, F5 recommends separating management and application traffic onto separate networks and interfaces. This enables better use of the cloud provider’s network security features to protect the management interface of the BIG-IP virtual device. In some cases, however, the cloud infrastructure design will make separating that traffic impossible. In this case, the best alternative is to create dedicated management IP addresses and endpoints on the BIG-IP platform. These endpoints may have to run on different ports from standard (for example, 3443 instead of 443) to keep commonly known ports available for application traffic such as an HTTPS website. Guidance for this configuration is available from F5.

![Diagram 1: Management and application traffic on separate networks and interfaces (recommended)](image1)

![Diagram 2: Dedicated management IP addresses and endpoints when separate traffic is impossible](image2)
Application traffic networks

BIG-IP devices can be connected to application traffic networks in one of two ways: a routed deployment or a one-armed deployment.

Routed deployment

In routed deployment, client traffic connects to the BIG-IP platform on one virtual subnet. The relevant application delivery services are applied, and then the traffic exits the platform on another virtual subnet. This configuration is common in both physical and virtual BIG-IP deployments. Since the BIG-IP platform is stateful, it is necessary to ensure that return application traffic from the servers also passes back via the BIG-IP device. This can be done using routing changes or Source Network Address Translation (SNAT).

Figure 4: Routed deployment
One-armed deployment

In some architectures or cloud infrastructures, the BIG-IP platform can only be connected to a single virtual subnet, in which case, a “one-armed” deployment is required. This deployment method—also known as “on a stick” or “same segment” deployment—is also used to manage application traffic to servers connected to the same virtual subnet. Organizations using this scenario should also use SNAT. (Although the flexibility of the F5 platform enables a variety of configurations, SNAT is the best choice in almost every case of one-armed deployment.)

Using SNAT

SNAT is a very common deployment option in which all inbound traffic reaching the back-end servers is modified to use the source IP address of the BIG-IP system. This ensures that the back-end server will always route traffic back to the BIG-IP device or virtual instance—even if the server has an alternate route to the real client IP address. However, this does have the disadvantage of obscuring the client source IP address from the back-end servers. This can usually be addressed with a layer 7 application value injection such as the HTTP X-FORWARDED-FOR header of the SIP VIA header.

In scenarios where traffic is expected to exceed 254,000 concurrent connections, and where the cloud infrastructure allows the binding of multiple IP addresses to a single interface, a pool of source IP addresses can be configured to eliminate the risk of ephemeral port exhaustion.
Multi-tenant or per-application instances

In traditional data center designs, it has been common practice to install high capacity network devices such as firewalls or F5 Application Delivery Controllers and to consolidate multiple services onto one highly available configuration (most commonly an active/passive failover pair).

In cloud environments, this design still has significant merit, as it offers a simple, familiar deployment model that is straightforward to manage and license. New services can be provisioned easily though automation via the REST API without the additional cost and complexity of deploying additional virtual machines.

Figure 6: Multiple application services consolidated to one BIG-IP instance in a multi-tenant design
Challenges arise when the organization has chosen to make extensive use of multiple virtual networks to separate applications into their own network environments. Where the cloud infrastructure permits it, one architecture design option is to create a virtual network dedicated to all common services and use this as a gateway into virtual networks for each application.

Figure 7: A virtual network providing common services

An alternative is to build per-application deployments with a dedicated BIG-IP instance serving each application. This allows for easy portability and provides a failure boundary against misconfiguration.

Figure 8: Dedicated BIG-IP services for each application environment

This approach requires a greater focus on management and licensing choices, and it is occasionally less efficient in terms of utilization, so flexible licensing through utility billing, license pooling, or volume licensing subscriptions is usually a requirement for large deployments.
Local availability options

For any application infrastructure stack, it is critical to understand the availability requirements of your BIG-IP platform and know how to meet them. Organizations have various options for ensuring high availability (HA) locally (within a region), and each has benefits and drawbacks.

No local HA

Uses such as dev/test or quality assurance (QA) may have no need for local high availability services for the BIG-IP platform. Similarly, if multiple regions host an application and DNS services are used to manage client distribution or failover, then you might opt for no local high availability. Organizations without local HA need a process for recovering or restarting the BIG-IP image and for performing software updates in the event of a failure. Organizations using license pooling or utility billing could reasonably deploy new upgraded instances of the BIG-IP platform rather than upgrading in place. In that case, the cloud provider’s public IP address mapping service can be used to perform the switch from the old BIG-IP platform to the upgraded one.

Figure 9: Deployment without local HA
Local HA using BIG-IP platform failover

As a long-established data center networking component, the BIG-IP platform has a feature-rich and highly configurable system for determining device availability and triggering a failover of virtual servers from one system to another. In many cloud environments, the BIG-IP platform can be used much like a physical device. The same rich inter-device monitoring can detect problems with an active device (including failures in connectivity as well as instance failure). What may change is how traffic is directed to the newly active device.

BIG-IP platform failover works at layer 2 by default. A newly active device will transmit a gratuitous ARP or deploy MAC masquerading to ensure traffic destined for a particular IP address gets to the new active device. In many cloud infrastructures, this default technique will not work. Instead, BIG-IP platforms can interact via API with the cloud IP address mapping infrastructure to remap the IP addresses to the interfaces of the newly active system.

BIG-IP devices synchronize configuration changes between members of a cluster, ensuring that the configuration of the devices is the same and the system will function as intended in the event of a failover.

This local HA system provides the most advanced failover detection and configuration options, but also requires the most configuration because the devices require a trust relationship between them. Identity tokens may be required to reside on the device so it can interact with the cloud infrastructure. Thus this method will not be available in some cloud environments.

Figure 10: Local HA provided by the BIG-IP platform
Local availability using BIG-IP DNS services

BIG-IP DNS is a powerful and flexible DNS system that provides global server load balancing (GSLB) as well as a high performance DNS and DNS firewall solution. In this use case, BIG-IP DNS can manage traffic to multiple standalone BIG-IP devices while monitoring their availability and performance. The monitoring ensures that only active devices are resolved for a particular service name. This same DNS system ensures availability both locally and across multiple regions. BIG-IP DNS can be enabled on separate BIG-IP platforms and run concurrently with other application delivery services.

Figure 11: Local HA using BIG-IP DNS

While this is an easy way to provide local and global availability, the reliance on the BIG-IP DNS system can be a drawback, since failover times will be determined by the Time to Live (TTL) setting, which is 30 seconds by default. When faster failovers are required, the organization will need an alternative solution.

Finally, the DNS services themselves must be made highly available (via multiple name server entries or IP anycast). Organizations do not necessarily need multiple BIG-IP devices running BIG-IP DNS per cloud, but a minimum of two devices is needed in the overall design. Since those BIG-IP devices are stand-alone devices, rather than a cluster, they also require methods to ensure configuration synchronization across the devices.

---

1In certain cloud architectures, BIG-IP sync-only device groups can potentially be used to achieve synchronization.
Local availability using the IaaS native load balancer

While the native load balancers provided in cloud systems are primitive, in some cases they will provide adequate failover between BIG-IP platforms delivering more advanced application delivery services such as web application firewall or access and identity services. You can use the same native load balancers to scale devices to deliver more throughput.

Figure 12: Using a cloud infrastructure load balancer for simple HA

As with the BIG-IP DNS solution for high availability, standalone BIG-IP devices will need configuration synchronization.

Outbound traffic

None of these failover mechanisms will work well for dealing with explicit outbound traffic, for example, where a BIG-IP device is the default outbound route for traffic from a VLAN. This is another excellent reason to adopt SNAT as the default mode for cloud implementations with BIG-IP products.
Global availability

Global availability designs solve traffic management issues when applications will be running in:

• Multiple geographical regions.
• The clouds of multiple cloud providers.
• A mix of private and public clouds.

Depending on the use case, traffic may be directed only to the active location or split across multiple (active/active) locations.

As with local availability, BIG-IP DNS provides a powerful, highly configurable, and high performance solution for GSLB. With F5 services, nearly any load distribution algorithm can be created to direct traffic to a particular location based on availability, latency, geographical location, or load.

Figure 13: Using BIG-IP DNS for global availability between regions or cloud providers
BIG-IP DNS can run on BIG-IP devices or virtual editions. The devices can be hosted anywhere with Internet connectivity and do not need to be collocated with the rest of the infrastructure, although it often makes sense for them to be so. The platforms hosting BIG-IP DNS need connectivity to the resolving DNS servers (LDNS) used by clients and to the BIG-IP platforms handling the traffic (or other traffic-handling components, such as infrastructure provided load balancers or server instances).

Detailed descriptions of the required ports, monitoring process, and other recommendations are available in the Best Practices guide called “Application Availability Between Hybrid Data Centers” on f5.com.
Scaling

A key feature of cloud infrastructures is the ability to scale resources up and down to meet demand while only paying for the resources actually consumed. While not all workloads require this elasticity, the ability to auto-scale resources up and down is a de facto requirement of many cloud application architectures.

Resource scaling considerations include the scaling of both the back-end server resources and that of the BIG-IP platform.

Figure 15: Scaling the BIG-IP platform
Figure 16: Scaling back-end resources

**Auto-scaling back-end resources**

Scaling back-end resources enables an application to meet changing demands in workloads due to fluctuating user demands or scheduled activities such as end-of-day batch processing or simulation runs.

Auto-scaling back-end resources is a three-part process:

1. Monitoring the thresholds of servers in the pool.
2. Communicating with the cloud infrastructure to create or destroy resources.
3. Adding or removing resources in the workload pool.
Figure 17: Three-step auto-scaling of back-end servers

Each of these steps requires API-driven interaction with the cloud infrastructure management system. While each cloud management system uses different API calls, the basic process is the same. Further documentation for specific cloud environments will become available from F5.
Scaling the BIG-IP platform
Scaling the BIG-IP platform involves similar requirements as for scaling back-end servers, but with an additional need for a method to distribute traffic to the newly created BIG-IP platform resources. Two options exist:

• Scaling using BIG-IP DNS
  Just as BIG-IP DNS services can direct traffic for local or global high availability, they can be used to distribute traffic across multiple BIG-IP instances. The same benefits and caveats apply as for the high availability solution.

• Scaling using the IaaS native load balancer
  Scaling using the cloud’s native load balancer is similar to using the load balancer for high availability services, except that the load balancer has to select where to send traffic, choosing from among a number of BIG-IP instances. Cases where an application requires persistence or “sticky sessions,” require careful configuration of the session persistence settings on the load balancer. These cases may require SSL decryption and layer 7 traffic processing on the load balancer before that traffic is sent to the BIG-IP instance for more advanced application delivery services such as web application firewall protection. As a result, the functionality limitations of the load balancer may be exceeded, especially if the application is being migrated from an existing F5-managed deployment.

Auto-scaling the BIG-IP platform
Auto-scaling the BIG-IP platform requires similar integration into the cloud management system as when scaling back-end servers, with the additional traffic distribution requirements discussed above. In general, to auto-scale a BIG-IP platform, the cloud management system must be able to monitor BIG-IP instances and deploy new template instances when thresholds are met.

The implementation details of auto-scaling will be addressed in recommended practices for each cloud provider.
Secure cloud connectivity

Transmitting data from the data center to the cloud or between clouds—whether for backup or as part of production application traffic—usually requires steps to preserve data security. Two methods are available:

- Private circuit connection into the cloud provider's infrastructure.
- Encrypting the data over a public Internet connection using a VPN.

Private circuit connections, which are provided as a partnership between service providers and the cloud infrastructure vendor, offer high bandwidth and low latency, along with a range of SLAs to suit your needs.

In many cases, however, private circuit connections are prohibitively expensive, and the lead time for establishing physical connectivity to a customer site may not meet the needs of the business. Private circuits also are not capable of connecting the infrastructures of two disparate cloud vendors.

Encrypting the data over a public link via a VPN may often meet business needs for bandwidth and security at a reduced cost. SLAs for local Internet connections can be as good as those for private circuits. However, they do not guarantee end-to-end connectivity.

Connecting BIG-IP devices using SSL

The BIG-IP platform includes a technology called F5® iSession®. An iSession is an SSL encrypted, accelerated connection between BIG-IP instances. These connections map the application traffic connections 1:1 but are SSL encrypted and transmitted on port 443. The iSession tunnels support only TCP connections.

An iSession is simple to configure and works between any two BIG-IP devices, physical or cloud (with software version dependencies, of course). iSession tunnels can be enabled with adaptive compression; the system then will select the best compression technique to achieve maximum throughput on the link. Data deduplication is also an option, although it can place significant loads on the platform and the I/O subsystems, so interested customers should work with F5 to size the solutions correctly.

![Figure 18: Data encryption over VPN using SSL connections](image-url)
Connecting BIG-IP devices using IPsec

The BIG-IP platform also enables device connectivity with IPsec tunnels, which are useful when you need to manage non-TCP traffic or prefer not to use SSL. iSession data can also be encapsulated in IPsec tunnels.

IPsec tunneling replaces the 1:1 application connection to tunnel connection with a single tunnel containing encapsulated traffic.

Figure 19: Data encryption over VPN using IPsec

Connecting the BIG-IP platform to the cloud VPN gateway using IPsec

As discussed in the section above on common infrastructure components of IaaS systems, many cloud infrastructures provide a VPN gateway. The BIG-IP platform supports IPsec connections to some third-party VPN devices, enabling connections from a BIG-IP device to a native VPN endpoint.

As a result, it’s possible to use an IPsec connection from a BIG-IP platform within one cloud provider’s virtual network to the VPN gateway of another to interconnect clouds from two different vendors.

Figure 20: Using IPsec to connect to a cloud VPN gateway
Security Considerations

The BIG-IP platform is a secure, stateful, default-deny system. This means it will only process or pass traffic that you explicitly configure. It does not act as an open router unless explicitly configured to do so.

Traffic is only accepted and passed to back-end servers when administrators create “traffic-passing” objects that explicitly define the source, destination port, and protocol. In addition, return traffic from servers must be part of an established traffic flow.

Adding services such as a firewall, web application firewall, or application access control provides additional layers of security to application traffic. These security services should be used in tandem with network security features provided by the cloud.

Organizations need to pay particular attention to securing management access. F5, Solution 13092, “Overview of Securing Access to the BIG-IP System,” provides an excellent overview of BIG-IP management security recommendations, including the required ports for management and inter-device communications. Wherever possible, access to management ports should be whitelisted to a set of known addresses. This can be achieved using both the cloud platform security controls and BIG-IP platform configuration.

Deployment Automation and Templates

Automating and orchestrating deployments can be critical to realizing value from a cloud deployment. A number of key components enable automating the BIG-IP platform deployment:

REST API

The REST API provides programmatic access to configuration of the BIG-IP platform. You can control nearly every configuration object through API calls. F5® DevCentral™ contains all the relevant REST API documentation. Writing scripts and integrations via the REST API allows you to control configuration from a wide range of tools and enables F5 to build integration with cloud management platforms.

Application templates

Configuring the optimum combination of application delivery services for a particular application (such as security, performance, or monitoring) can require a large number of configuration objects. While these are generally all accessible via the API, they can place an unnecessary burden of complexity on the systems used to automate the BIG-IP platform.

Application templates provide a way to mask this complexity from upstream systems and provide uniform, best practice deployments every time. To configure application delivery for a template application, all the automation system needs to do is supply the name of the template and the site-specific values (such as IP addressing). The new application delivery service is created using the standard configuration in the template and the values passed by the orchestration system. (Certain prerequisites, such as SSL certificates, may need to be pre-loaded to make this possible, but can also be API driven).
Integration with cloud and third-party tools

Cloud services provide automation and orchestration tools to enable the creation of infrastructure components and application architectures. The BIG-IP platform API enables easy integration with many of these systems to enable the programmatic creation and configuration of BIG-IP platforms and application delivery services. Individual cloud infrastructure guides contain further documentation for each cloud provider.

In addition to cloud-provided automation services, organizations can take advantage of many other automation solutions, such as:

- Ansible.
- Puppet.
- Chef.

F5 is actively working with a number of these solution providers to test, document, and validate their automation solutions. Many of the results will be published on the F5 GitHub account.

In many orchestration scenarios, the F5® BIG-IQ® management and orchestration platform can reduce both the cost and complexity of managing the automation. The BIG-IQ platform offers element management, licensing, and cloud integration facilities in a modular suite of software components.

Conclusion

F5 application delivery services provide the security, availability, and advanced traffic management required by applications in the public cloud. By adding security, access control, and optimization services, organizations can ensure their cloud-based applications remain highly available and uncompromised. These public cloud design patterns and the F5 recommended practices for your specific cloud environment can help you build the configuration that best suits your organization and its business needs.

Resources

Find additional information on these websites, or contact sales@f5.com.

- F5 cloud computing solutions
- Cloud resources on F5 DevCentral
- F5 resources on GitHub