



F5 White Paper

Availability and the Cloud

Cloud computing offers IT another tool to deliver applications. While enticing, challenges still exist in making sure the application is always available. F5's flexible, unified solutions ensure high availability for cloud deployments.

by Peter Silva

Technical Marketing Manager



Contents

Introduction	3
<hr/>	
Cloud Forecast: Partly to Mostly Sunny	4
<hr/>	
Delivering Applications in the Cloud	4
<hr/>	
Achieving Availability in the Cloud	5
<hr/>	
Conclusion	8



Introduction

“By 2012, 20 percent of businesses will own no IT assets.”¹ While the need for hardware will not disappear completely, hardware ownership is going through a transition: Virtualization, total cost of ownership (TCO) benefits, an openness to allow users run their personal machines on corporate networks, and the advent of cloud computing are all driving the movement to reduce hardware assets. While many IT departments haven’t yet fully embraced all the potential of cloud computing, there is a growing understanding that cloud computing can offer cost savings, including a reduction in capital expenses. Cloud computing also offers the ability to deliver critical business applications, systems, and services around the world with a high degree of availability, which enables a more productive workforce.

There are three primary deployment models for the cloud: public (resources provisioned and available over the Internet), private (internal provision of resources through intranet and virtualization), and hybrid (a combination of public and private models). Within these deployment models, different delivery services provide infrastructure, platform, and software delivery. Although there is also some confusion (and disagreement) about the parameters of these services—particularly given the ongoing evolution of the cloud—most of the industry recognizes three service delivery models:

Infrastructure as a Service (IaaS): IaaS delivers computing infrastructure as a service. Instead of purchasing hardware and other infrastructure components, customers use some form of virtualization to access outsourced resources. Because consumption is on an on-demand basis, costs directly reflect the amount of use.

Platform as a Service (PaaS): PaaS delivers computing and development platforms (for example, Microsoft .NET, Java EE, Google applications) as a service, giving users the ability to deploy and develop applications without significant hardware and software expense or management time. Since the deployment platform is very specific, like .NET, there might be limitations of the types of applications that might be supported. For instance, Google App Engine only supports applications written using Python while Heroku supports Ruby on Rails application development.

Software as a Service (SaaS): By delivering applications as a service, SaaS offers customers pre-packaged/pre-built applications through a standard web browser. With SaaS, customers can avoid the installation and management of software on their own computers and further benefit from centralized, automatic software updates as well as lower costs. Customers don’t need to dedicate valuable resources to software deployment or management.

Cloud computing is a style of computing in which dynamically scalable and often virtualized resources are provided as a service. Users need not have knowledge of, expertise in, or control over the technology infrastructure in the “cloud” that supports them. Furthermore, cloud computing employs a model for enabling available, convenient, and on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

F5 Networks Cloud Survey,
August 2009



No matter which cloud service—IaaS, PaaS, or SaaS (or combination thereof)—customers or service providers choose, the availability of that service to users is paramount, especially if service level agreements (SLAs) are part of the contract. Even with a huge cost savings, there is no benefit for either the user or business if an application or infrastructure component is unavailable or slow.

Cloud Forecast: Partly to Mostly Sunny

Over the last year or so, cloud computing has quickly grown from a little-understood delivery model to a valuable resource for IT departments. IT departments don't need a cloud expert on staff since many of the offerings and services are either pre-built or are similar to what has traditionally been deployed in house. In fact, the primary difference is that the offerings and services are not located in house, but outside the organization in one or more data centers either off-site or in the cloud.

As hype about the cloud has turned into the opportunity for cost savings, operational efficiency, and IT agility, organizations are discussing, testing, and deploying some form of cloud computing. Many IT departments initially moved to the cloud with non-critical applications and, after experiencing positive results and watching cloud computing quickly mature, are starting to move their business-critical applications. No matter what the deployment model, much of the initial capital outlay for hardware, software, bandwidth, licenses, and more is reduced, enabling business units and IT departments to focus on the services and workflows that best serve the business.

Delivering Applications in the Cloud

Like business, the cloud is dynamic in nature; as such, cloud computing integration and support needs to be flexible. F5® solutions, in general, focus on the task of application delivery. Since the driver for any cloud deployment, regardless of model or location, is to deliver applications in the most efficient, agile, and secure way possible, all F5 solutions can fit within the cloud infrastructure and enhance application delivery. The dynamic control plane of cloud architecture requires the capability to intercept, interpret, and instruct where the data must go and must have the necessary infrastructure, at strategic points of control, to enable quick, intelligent decisions and ensure consistent availability.



F5 products and solutions provide the scalability, extensibility, adaptability, manageability, security, and real-time performance required of a dynamic control plane. Since the cloud exists just about anywhere, controlled, contextual delivery of applications becomes paramount in the realm of availability. F5 products such as BIG-IP® Local Traffic Manager™ (LTM), BIG-IP® Edge Gateway™ and BIG-IP® Global Traffic Manager™ (GTM) help customers and service providers alike build a cloud computing environment that meets specific needs. F5 products provide intelligent, strategic points of control through proxies, policies, and services in a unique, modularized delivery infrastructure that is capable of handling the high-volume of traffic associated with cloud computing. Additionally, because F5 solutions can be deployed on a wide range of hardware platforms, they help ensure availability.

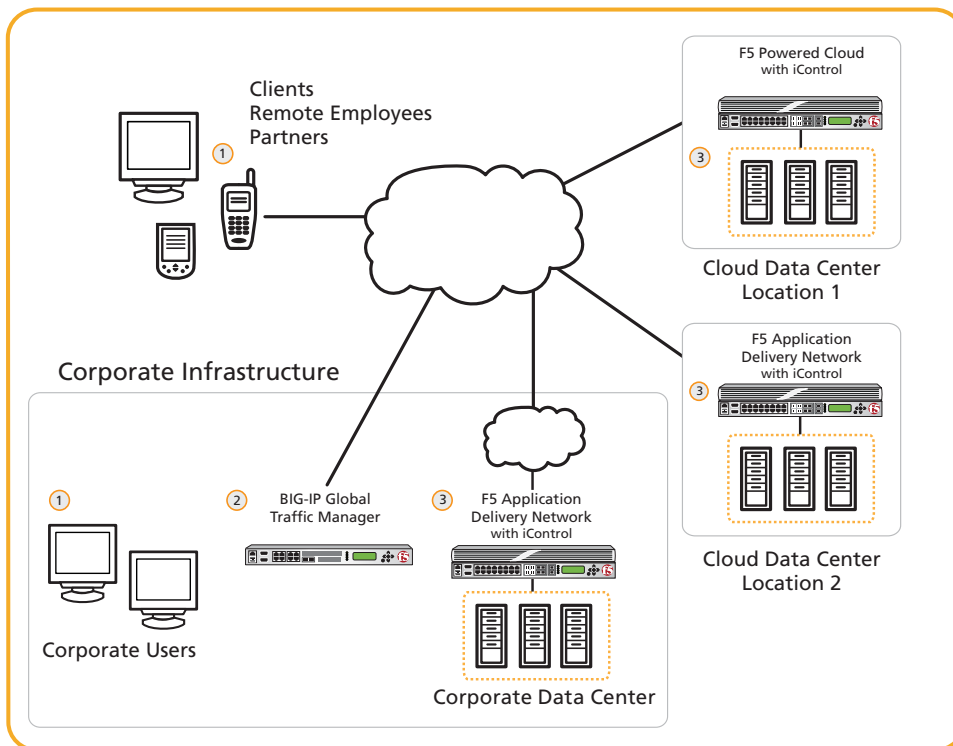
Achieving Availability in the Cloud

The on-demand, elastic, scalable, and customizable nature of the cloud must be considered when deploying cloud architectures. Many different customers might be accessing the same back-end application(s), but each customer has the expectation that only their application will be properly delivered to users. Making sure that multiple instances of the same application are delivered in a scalable manner requires both load balancing and some form of server virtualization. As an Application Delivery Controller (ADC), BIG-IP LTM represents the pinnacle of load balancing. BIG-IP LTM can virtualize back-end systems and can integrate deeply with the network and application servers to ensure the highest availability of a requested resource. Each request is inspected using any number of metrics and then routed to the best available server. Knowing how an ADC can enhance your application delivery architecture is essential prior to deployment. Many applications have stellar performance during the testing phase, only to fall apart when they are live. By adding the BIG-IP® Local Traffic Manager™ (LTM) Virtual Edition (VE) to your development infrastructure, you can build, test and deploy your code with ADC enhancements from the start. By providing a programmable, adaptable platform BIG-IP LTM enables the scalability necessary in the cloud and the availability necessary for the particular application delivered from the cloud.

Since application availability is paramount in the cloud, BIG-IP LTM, built on the TMOS® modular traffic management platform, enables the addition of features and functionality without disrupting the production architecture. Load balancing is just the foundation of what can be accomplished. In application delivery architectures,



additional elements such as caching, compression, rate shaping, authentication, and other customizable functionality, can be combined to provide a rich, agile, secure and highly available cloud infrastructure. Scalability is also important in the cloud and BIG-IP LTM can bring up or take down application instances seamlessly—as needed and without IT intervention—helping to prevent unnecessary costs if you’ve contracted a “pay as you go” cloud model. BIG-IP LTM can also isolate management and configuration functions to control cloud infrastructure access and keep network traffic separate to ensure segregation of customer environments and the security of the information. The ability of BIG-IP LTM to recognize network and application conditions contextually in real-time, as well as its ability to determine the best resource to deliver the request, ensures the availability of applications delivered from the cloud.



- 1 Users (local or remote) make a request to access web resources.
- 2 BIG-IP GTM makes a determination based on capacity, performance, location (and other user-specified parameters) which site—local or cloud—will best service the request. If the cloud is an F5 powered cloud, the parameters from which the administrator can choose will be more extensive.
Administrative domains isolate configuration and management for fine-grained control over access to cloud computing infrastructure.
- 3 The selected data center/cloud answers the request.



Availability is crucial; however, unless applications in the cloud are delivered without delay, especially when traveling over latency-sensitive connections, users will be frustrated waiting for “available” resources. Using SSL technology, F5 BIG-IP® Edge Gateway™ is a high performing controller, offering secure, accelerated access no matter where in the world a user is located. With the BIG-IP® WebAccelerator™ product module and the BIG-IP® WAN Optimization Module,™ web applications are delivered with LAN like speed, and file transfers are smooth even on high latency lines. Using cache, adaptive compression, and data deduplication, access to applications and needed resources is so fast that users could just as well be connected directly to a LAN. If other BIG-IP devices are deployed throughout the overall infrastructure, administrators can create secure optimized tunnels between BIG-IP devices, creating an instant private backbone. When a user initiates a secure connection from their client to the BIG-IP, and there is a secure tunnel between BIG-IP devices, you get double encryption with ultimate optimization and high availability all with BIG-IP devices.

Additional cloud deployment scenarios like disaster recovery or seasonal web traffic surges might require a global server load balancer added to the architecture. BIG-IP GTM, which is built on the same TMOS architecture as BIG-IP LTM, uses application awareness, geolocation, and network condition information to route requests to the cloud infrastructure that will respond best. BIG-IP GTM can also determine the geolocation of users based on IP address and route them to the closest cloud or data center, all without user interaction. In extreme situations, such as a data center outage, BIG-IP GTM will already know if a user’s primary location is unavailable and it will automatically route the user to the responding location. BIG-IP GTM provides global application availability whether you choose IaaS, PaaS, or SaaS.

Conclusion

Cloud computing, while still evolving in all its iterations, can offer IT a powerful alternative for efficient application, infrastructure, and platform delivery. As businesses continue to embrace the cloud as an advantageous application delivery option, the basics are still the same: scalability, flexibility, and availability to enable a more agile infrastructure, faster time-to-market, a more productive workforce, and a lower TCO along with happier users. Even though cloud computing in all its variants is still evolving, F5 provides a set of flexible, unified solutions to address cloud delivery needs now and in the future to help ensure the cloud computing environment is always fast, secure, and highly available.

¹ Press Release: [Gartner Highlights Key Predictions for IT Organizations and Users in 2010 and Beyond](#)

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

F5 Networks, Inc.
Corporate Headquarters
info@f5.com

F5 Networks
Asia-Pacific
apacinfo@f5.com

F5 Networks Ltd.
Europe/Middle-East/Africa
emeainfo@f5.com

F5 Networks
Japan K.K.
f5j-info@f5.com

