



F5 White Paper

Global Distributed Service in the Cloud with F5 and VMware

Using F5 BIG-IP Global Traffic Manager to orchestrate and deliver access to services in the cloud.

by Alan Murphy

Technical Marketing Manager



Contents

Introduction	3
<hr/>	
The Traditional Data Center	3
<hr/>	
The Cloud: A New Distributed Data Center Model	4
The IT Challenge: Managing Access to the Cloud	5
<hr/>	
Clouds Inside and Out	6
<hr/>	
Follow That App!	8
Global Application Delivery	8
<hr/>	
Conclusion	9



Introduction

The term “cloud computing” has been a prevalent, forward-thinking component of IT for the past year, fueled by new technologies in the local data center such as virtualization. Like most new terms in technology, cloud computing means different things to different people. On one end of the spectrum, it’s nothing more than a marketing phrase used to make people think about how their data center is changing and evolving to include off-site services, with no meat to back it up. You’ll hear, “Tomorrow your applications and data will be different in the cloud!” but there’s not much talk about how the cloud will change your services. On the other end, the cloud is a completely new computing paradigm with shared yet secure resources—both on- and off-site—providing a dynamic “just-in-time” (JIT) computing model. While the end goal may be to achieve a true JIT dynamic data center in the cloud, the reality of cloud computing today lies somewhere in the middle of the spectrum: using technologies readily available today to better align IT needs with the needs of the business. The goal is to provide an environment where applications are no longer bound to the local data center and can scale across available resources as needed.

The Traditional Data Center

For there to be a new data center model, there must be an old data center model to move away from. This old, or traditional, model is still a tried and true architecture today—albeit one that is beginning to show its age and is in desperate need of redesign—and in reality is where most data centers begin. In this static data center model, client server applications consist of a farm of physical servers sitting behind a firewall and communicating with users through a traditional load balancer. Physical data centers were built to focus on speeds and feeds, with plumbing built to handle network congestion and traffic direction. As new services were added to the data center, physical servers and cables were added and continued to grow as long as there was available real estate.

Historically, enterprises have built out static data centers from a central location, either by building a structure themselves or by leasing rack space from a hosting provider. They were isolated and self-contained, single-function structures. Directing users to a traditional data center was as easy as allocating a domain name to an IP address in that data center. Once a user was directed to that data center via one domain name, they stayed isolated in that same data center.



But that was only for the initial, or primary, data center. A secondary redundant data center was also needed to provide a rollover location for all application services in the data center should the primary become unavailable due to a human, system, or natural failure. In essence, every physical data center required a geographically removed twin, just in case. Every part of the primary data center needed to be replicated at the secondary location and kept in constant synchronization with the primary.

This primary/secondary data center model has worked well for providing a disaster recovery architecture but it still has challenges: capital expenses to build and maintain parity between systems and facilities within each data center, and operating expenses to manage each location and keep each in perfect mirror sync. These requirements continue to drain the business of operating funds. Despite the challenges with this model, it has become the de facto architecture for redundant system and application distribution. When an application is mission-critical, virtually no expense is too high to change this architecture, even if it requires maintaining multiple geographically disparate data centers.

That high cost barrier to entry for mission-critical applications and services is changing, however, and new computing models are attempting to do away with that barrier all together. Enter the cloud.

The Cloud: A New Distributed Data Center Model

From this antiquated, bifurcated data center model came the need to dynamically direct users to different data centers. The decision over where to direct them could be based on data center availability: "Is my primary data center able to serve my shopping cart to users?" If yes, then new and existing user connections would continue to be routed to the primary data center. If/when the answer became no, connections would be routed to the secondary data center. User direction decisions could then be based on additional information such as geographic location or fastest service response time on a user-by-user basis. F5® BIG-IP® Global Traffic Manager™ (GTM) pioneered the science of managing user sessions to unique data centers based on application service availability and location.



For the new dynamic data center, the primary driver is a model where discrete resources are allocated for applications rather than completely isolated systems. These resources are decoupled from physical boundaries (such as servers and data centers). The new model allows us to build data centers around the applications rather than the network. It enables us to break through the physical requirements of primary and secondary single-purpose computing sites. And most importantly, it lets us begin treating all parts of the data center—network, IP, storage, and applications—as fluid services rather than isolated technologies. This is the basis for cloud computing.

Cloud computing, as it's typically defined for remote deployments, enables an IT department to run services off-premise in a hosted data center, decoupling applications from physical servers located in isolated data centers. Unlike a traditional hosted environment, however, where the physical data center model is replicated at an off-site location, off-premise cloud computing enables companies to spin up new application services using technologies such as VMware's vSphere™ virtualization platform. This eliminates the need to physically install and manage servers in an off-site data center.

The IT Challenge: Managing Access to the Cloud

There's no question that the cloud is a new computing model that will change and challenge how IT manages local and remote applications. Tools are beginning to emerge from off-premise cloud vendors to help manage application growth into the cloud via APIs, along with tools for on-premise clouds such as VMware's vCloud™ set of services. The long-term goal is to enable seamless and fluid management of applications as they are provisioned between internal and external clouds.

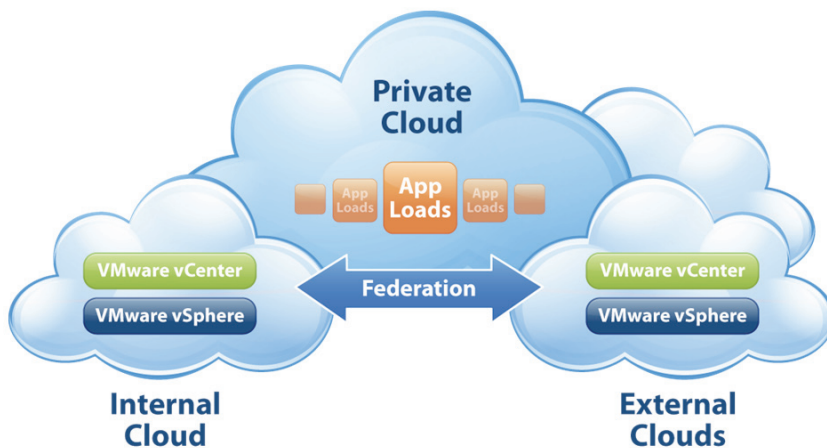
But what about managing how users access applications as those services float around the cloud? Unlike the primary/secondary distributed model, applications in the cloud can be provisioned and de-provisioned in real-time and even be moved from one physical location to another, oftentimes spanning large geographic regions. IP, DNS, and user management become paramount issues in a true cloud model. Working alongside VMware, F5 BIG-IP® Local Traffic Manager™ (LTM) and BIG-IP GTM work in concert to enable users to follow applications as they come up and down around the cloud.



Clouds Inside and Out

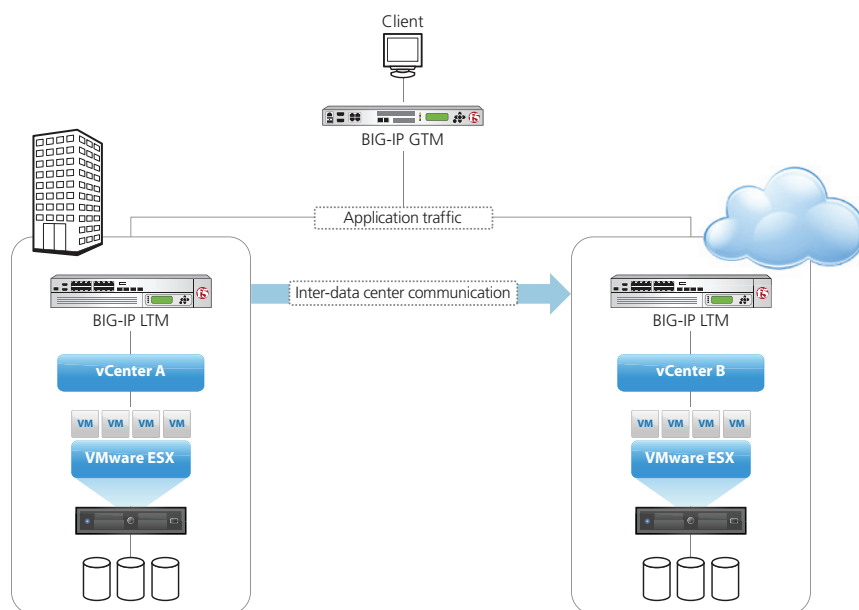
VMware's vCloud initiative is a set of services that couple VMware virtualization platform products with a set of management APIs designed to manage virtual machines and applications as they move in and out of the cloud. vCloud supports the internal deployment model of a private or on-premise cloud and the external deployment model of a public or off-premise cloud, as well as linking the two cloud types together so they can be managed as one solution.

vCloud is targeted towards both enterprise customers and cloud service providers. In an optimal deployment scenario, a customer will use VMware vCloud technologies to build an internal cloud, running application services and virtual machines on VMware's virtualization platform vSphere. These application services can be bundled together and managed as one component using vApp, or they can be made up of very specific virtual application services such as VMware View virtual desktop images, all managed by vCenter and vCloud.



VMware distributed cloud environments

Within this on-premise cloud, the enterprise can then deploy F5 BIG-IP LTM to manage application access and delivery to the apps running on those virtual machines. BIG-IP LTM sits in-line in front of the virtual machine applications and manages all user connections in and out of those virtual machines. In essence, BIG-IP LTM becomes the extended networking arm for applications running on the ESX™ hypervisor; it becomes part of VMware's virtual networking fabric. BIG-IP LTM plugs directly into vCenter so that it can be managed as part of the on-premise virtual infrastructure.



BIG-IP LTM works with VMware vCenter within and between data centers to manage virtual application delivery.

Once an enterprise IT group has deployed an on-premise cloud, they can then begin looking at external providers to host their virtual machines in an off-premise cloud—if there is a need to move services off-site. Building an on-premise cloud doesn't mean it is also necessary to use an off-premise solution. In fact, many IT groups will build an internal cloud to enable an agile data center while keeping their services internal. It is still important, however, to plan for managing internal and external clouds together for future growth.

A traditional off-premise cloud computing model (if it's possible to have a traditional model this early in the cloud lifecycle) requires that customers build their infrastructure and applications "on site" with the cloud provider; new virtual machines with new application installations must be provisioned at each cloud provider site from scratch. While this model works for customers looking to build from the ground up, it's not optimal for customers who already have an on-premise cloud deployment in place and are already using virtual machines throughout their infrastructure.

Enterprise customers that are or will be using vCloud internally for on-premise clouds can look for and choose cloud providers that offer vCloud and support running existing virtual machines, enabling an enterprise to communicate directly with their off-premise cloud and manage it as though it's internal—an extension of their own on-premise deployment. Virtual machines (and vApp bundles) can be transferred as needed between the on- and off-premise clouds, enabling a highly dynamic infrastructure for situations such as cloud bursting.



Follow That App!

One of the challenges of building and deploying such a dynamic application environment with vCloud —and cloud architecture models such as cloud bursting—is managing access to applications and services running with the clouds as they move from place to place. When dealing with the older primary/secondary model of application location, applications are static and isolated to a specific location. Once the application is live in the primary data center, that's where it stays. One of the main advantages of the cloud is the ability to move applications in and out as needed, from one location to the next, breaking that isolation barrier. Where those applications live at any given moment and how users access the application is a critical component that needs to be addressed with cloud deployments.

Global Application Delivery

In both physical and virtual data centers, F5 BIG-IP GTM manages application access as applications move from on- to off-premise and vice versa. BIG-IP GTM follows the application and is always aware of its location, processing state, availability, performance, and secure access requirements. When applications move from one cloud to another, BIG-IP GTM moves user access to those applications to ensure that the apps are always available to users, even as they're moving from one cloud to the next.

A good example of this scenario is cloud bursting. With cloud bursting, an application resides within a local virtual data center (an on-premise cloud) and users are served from this location as long as the application and bandwidth can handle the level of requests. Once a certain threshold is met, such as too many user connections to the application, new user requests will be directed outside the data center to an off-premise cloud; existing connections and users will continue to be served from within the internal cloud. Once the internal application connections fall below the acceptable threshold, connections that were burst into the cloud will be moved back to the internal data center. This elaborate system of managing individual user connections to the application, and making sure that new and existing connections go to different locations appropriate for the app, is handled by BIG-IP GTM. By communicating directly with BIG-IP LTM within each data center, BIG-IP GTM is always aware of how the local applications are responding and if they're meeting SLAs and performing as expected. BIG-IP LTM can employ threshold management through rate shaping and connection throttling, and notify BIG-IP GTM when a connection redirection event needs to occur.

55% of IT organizations reported that the ability to redirect, split, or rate-shape application traffic between multiple data centers is valuable when choosing a cloud provider.

– Source: TechValidate,
TVID: 3D4-C64-27A

White Paper

Global Distributed Service in the Cloud with F5 and VMware

Local application management goes hand in hand with global application management. Since BIG-IP LTM can be managed as part of vCenter and with VMware's AppSpeed™ product, information from the virtual infrastructure can also be relayed directly to BIG-IP GTM. For example, when virtual machines need to come down for offline patching overnight, BIG-IP LTM can bleed those connections from both the local virtual machines as well as the entire virtual data center, moving those connections to another data center via BIG-IP GTM. When used with a cloud-based architecture like VMware vSphere, BIG-IP GTM and BIG-IP LTM enable IT groups to manage external clouds as if they were their own data centers. They can move users between locations, minimize downtime, and use the cloud to create a more fault-tolerant Application Delivery Network.

Conclusion

More than any technology in the past 10 years, cloud computing enables IT agility; it allows the data center to move for business needs rather than technology need. VMware is leading the pack in developing a cloud platform for both internal and external clouds. By working with VMware and by providing a robust Application Delivery Networking infrastructure for virtual machines and platforms, F5 BIG-IP LTM and BIG-IP GTM provide the backbone for application delivery inside and out of the cloud with solutions such as vCloud. BIG-IP products work in concert across the entire application stack—from BIG-IP LTM monitoring application health and managing local user connections through BIG-IP GTM distributing those connections across the globe. Together, they deliver applications running in the cloud on VMware virtual platforms to meet changing business needs.

Applications are becoming mobile packages rather than static systems and machines. As this mobility grows and becomes more commonplace, F5 is there to help manage the growth and guarantee that applications are always secure, fast, and available, enabling organizations to achieve IT agility.

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
info.asia@f5.com

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

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

