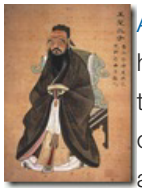
CloudFucius Ponders: High-Availability in the Cloud

Peter Silva, 2010-05-05



According to Gartner, "*By 2012, 20 percent of businesses will own no IT assets*." 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. Cloud computing 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. No matter which cloud service — IaaS, PaaS, or SaaS (or combination thereof) — a customer or service provider chooses, 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.

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, enabling business units and IT departments to focus on the services and workflows that best serve the business. 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, 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.

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 applications, 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. An Application Delivery Controller (ADC) 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 a Virtual ADC to your development infrastructure, you can build, test and deploy your code with ADC enhancements from the start.

With an ADC, 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 being able to bring up or take down application instances seamlessly — as needed and without IT intervention — helps to prevent unnecessary costs if you've contracted a "pay as you go" cloud model. An ADC 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 an ADC 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.

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. Additional cloud deployment scenarios like disaster recovery or seasonal web traffic surges might require a global server load balancer added to the architecture. A Global ADC uses application awareness, geolocation, and network condition information to route requests to the cloud infrastructure that will respond best and using the geolocation of users based on IP address, you can route the user to the closest cloud or data center. In extreme situations, such as a data center outage, a Global ADC will already know if a user's primary location is unavailable and it will automatically route the user to the responding location.

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.

And one from Confucius: *The man of virtue makes the difficulty to be overcome his first business, and success only a subsequent consideration.*

ps

The CloudFucius Series: Intro, 1, 2, 3

F5 Networks, Inc. | 401 Elliot Avenue West, Seattle, WA 98119 | 888-882-4447 | 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