4 Things You Need in a Cloud Computing Infrastructure

Lori MacVittie, 2008-10-07

Cloud computing is, at its core, about delivering applications or services in an on-demand environment. Cloud computing providers will need to support hundreds of thousands of users and applications/services and ensure that they are fast, secure, and available. In order to accomplish this goal, they'll need to build a dynamic, intelligent infrastructure with four core properties in mind: transparency, scalability, monitoring/management, and security.

Transparency

One of the premises of Cloud Computing is that services are delivered transparently regardless of the physical implementation within the "cloud". Transparency is one of the foundational concepts of cloud computing, in that the actual implementation of services in the "cloud" are obscured from the user. This is actually another version of virtualization, where multiple resources appear to the user as a single resource.

It is unlikely that a single server or resource will always be enough to satisfy demand for a given provisioned resource, which means transparent load-balancing and application delivery will be required to enable the transparent horizontal scaling of applications on-demand. The application delivery solution used to provide transparent load-balancing services will need to be automated and integrated into the provisioning workflow process such that resources can be provisioned on-demand at any time.

Related Articles from around the Web

What cloud computing really means
How Cloud & Utility Computing Are Different
The dangers of cloud computing
Guide To Cloud Computing

For example, when a service is provisioned to a user or organization, it may need only a single server (real or virtual) to handle demand. But as more users access that service it may require the addition of more servers (real or virtual). Transparency allows those additional servers to be added to the provisioned service without interrupting the service or requiring reconfiguration of the application delivery solution. If the application delivery solution is integrated via a management API with the provisioning workflow system, then transparency is also achieved through the automated provisioning and de-provisioning of resources.

Scalability

Obviously cloud computing service providers are going to need to scale up and build out "mega data centers". Scalability is easy enough if you've deployed the proper application delivery solution, but what about scaling the application delivery solution? That's often *not* so easy and it usually isn't a transparent process; there's configuration work and, in many cases, re-architecting of the network. The potential to interrupt services is huge, and assuming that cloud computing service providers are servicing hundreds of thousands of customers, unacceptable.

The application delivery solution is going to need to not only provide the ability to transparently scale the service infrastructure, but itself, as well. That's a tall order, and something very rarely seen in an application delivery solution.

Making things even more difficult will be the need to scale *on-demand* in real-time in order to make the most efficient use of application infrastructure resources. Many postulate that this will require a virtualized infrastructure such that resources can be provisioned and de-provisioned quickly, easily and, one hopes, automatically. The "control node" often depicted in high-level diagrams of the "cloud computing mega data center" will need to provide on-demand dynamic application scalability. This means integration with the virtualization solution and the ability to be orchestrated into a workflow or process that manages provisioning.

Intelligent Monitoring

In order to achieve the on-demand scalability and transparency required of a mega data center in the cloud, the control node, i.e. application delivery solution, will need to have intelligent monitoring capabilities. It will need to understand when a particular server is overwhelmed and when network conditions are adversely affecting application performance. It needs to ***know*** the applications and services being served from the cloud and understand when behavior is outside accepted norms. While this functionality can certainly be implemented externally in a massive management monitoring system, if the control node sees clients, the network, and the state of the applications it is in the best position to understand the real-time conditions and performance of all involved parties without requiring the heavy lifting of correlation that would be required by an external monitoring system.

But more than just knowing ***when*** an application or service is in trouble, the application delivery mechanism should be able to take action based on that information. If an application is responding slowly and is detected by the monitoring mechanism, then the delivery solution should adjust application requests accordingly. If the number of concurrent users accessing a service is reaching capacity, then the application delivery solution should be able to not only detect that through intelligent monitoring but participate in the provisioning of another instance of the service in order to ensure service to all clients.

Security

Cloud computing is somewhat risky in that if the security of the cloud is compromised potentially all services and associated data within the cloud are at risk. That means that the mega data center must be architected with security in mind, and it must be considered a priority for every application, service, and network infrastructure solution that is deployed.

The application delivery solution, as the "control node" in the mega data center, is necessarily one of the first entry points into the cloud data center and must itself be secure.

It should also provide full application security - from layer 2 to layer 7 - in order to thwart potential attacks at the edge. Network security, protocol security, transport layer security, and application security should be prime candidates for implementation at the edge of the cloud, in the control node. While there certainly will be, and should be, additional security measures deployed within the data center, stopping as many potential threats as possible at the edge of the cloud will alleviate much of the risk to the internal service infrastructure.

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