White Paper

Cloud Balancing: The Evolution of Global Server Load Balancing

Cloud balancing moves global server load balancing from traditional routing options based on static data to context-aware distribution across cloud-based services, both internal and external. Consequently, automation reduces errors and IT labor hours while speeding the resource response to changing environmental conditions.

by Lori MacVittie
Senior Technical Marketing Manager

Contents

Introduction

The mysticism of cloud computing has worn off, leaving those required to implement cloud computing directives with that valley-of-despair feeling. When the hype is skimmed from cloud computing—private, public, or hybrid—what is left is a large, virtualized data center with IT control ranging from limited to non-existent. In private cloud deployments, IT maintains a modicum of control, but as with all architectural choices, that control is limited by the systems that comprise the cloud. In a public cloud, not one stitch of cloud infrastructure is within the bounds of organizational control. Hybrid implementations, of course, suffer both of these limitations in different ways.

But what cloud computing represents—the ability to shift loads rapidly across the Internet—is something large multi-national and even large intra-national organizations mastered long before the term "cloud" came along. While pundits like to refer to cloud computing as revolutionary, from the technologists' perspective, it is purely evolutionary. Cloud resources and cloud balancing extend the benefits of global application delivery to the smallest of organizations.

In its most basic form, cloud balancing provides an organization with the ability to distribute application requests across any number of application deployments located in data centers and through cloud-computing providers. Cloud balancing takes a broader view of application delivery and applies specified thresholds and service level agreements (SLAs) to every request. The use of cloud balancing can result in the majority of users being served by application deployments in the cloud providers' environments, even though the local application deployment or internal, private cloud might have more than enough capacity to serve that user.

A variant of cloud balancing called cloud bursting, which sends excess traffic to cloud implementations, is also being implemented across the globe today. Cloud bursting delivers the benefits of cloud providers when usage is high, without the expense when organizational data centers—including internal cloud deployments—can handle the workload.

Multiple Data Center Capabilities Important for Cloud Providers

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

Cloud Balancing

In one vision of the future, the shifting of load is automated to enable organizations to configure clouds and cloud balancing and then turn their attention to other issues, trusting that the infrastructure will perform as designed.

This future is not so far away as it may appear. Consider the completely automated scenario in the diagram below.

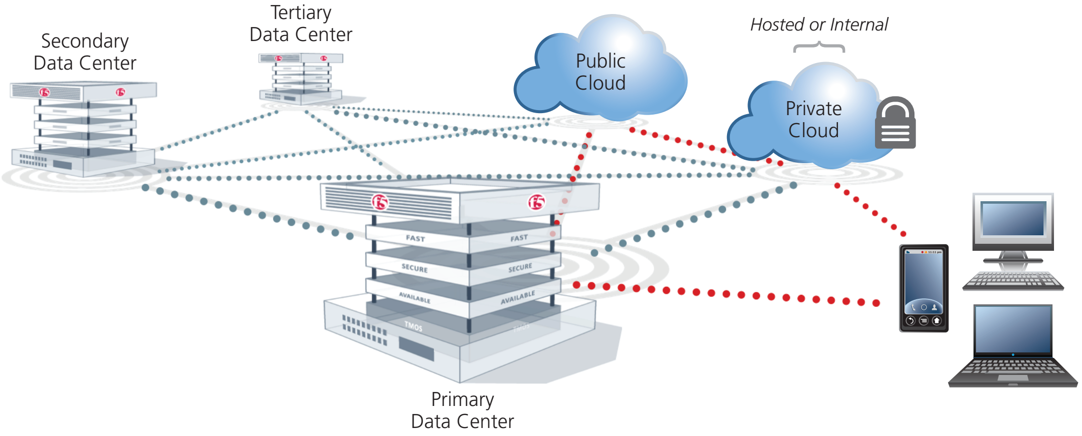

Figure 1: Automated cloud balancing

The global server load balancing (GSLB) and global DNS functionality that has been in place for a very long time is, given the correct architecture, also valid in cloud balancing. The point of both is to present a unified DNS for a variety of locations and determine the best place from which to serve an application when a customer connects.

Consider the scenario of a simple web application that must be available 24x7 and must be served as quickly as possible. Customers enter personally identifiable information (PII) into the application, so data must be safeguarded no matter where it resides.

Configuring GSLB and global DNS to direct traffic to available installations based upon the organization’s criteria and the state of the application permits routing to the geographically closest data center or, if it is down, to an alternate data center, all from the same URL.

Put another way, cloud balancing extends the architectural deployment model used in conjunction with GSLB to the cloud, which increases the choices available for organizations when determining from where a given application should be delivered. What is new in global application delivery is the ability to make application routing decisions based on variables other than those traditionally associated with network layer measurements. Business leaders in the midst of a decision-making process are demanding visibility into metrics, such as the costs associated with responding to a given request, the ability to meet a SLA, and user device and location, among others. Plus, these business leaders seek the capability to balance requests across application instances in various cloud locations based on the value of a transaction or current step within a business process.

Goals

Cloud balancing uses a global application delivery solution to determine, on a per user or customer basis, the best location from which to deliver an application. The decision-making process should include traditional GSLB parameters such as:

- Application response time.
- User location.
- Availability of the application at a given implementation location.
- Time of day.
- Current and total capacity of the data center or cloud computing environment in which the application is deployed.

Additionally, the organization must consider business-focused variables, including:

- Cost to execute the request at a given location.
- Regulatory compliance and legal restrictions.
- Business continuity planning.
- Energy consumption metrics.
- Services required by the user/customer to fulfill the request based on contractual obligations.

It is these business-focused variables, which are admittedly difficult to incorporate, that make cloud balancing an attractive strategy for maximizing the performance of applications while minimizing the costs associated with delivering them. These variables are exacerbated by the inclusion of internal cloud balancing, which, while often more appealing, uses a different set of cost metrics to determine suitability. Those metrics must be translated and comparable to external cloud metrics for true cloud balancing to incorporate an internal cloud.

The key to business continuity planning is in the GSLB and DNS portions of cloud balancing. Just as corporations with multiple data centers eventually moved toward an active/active environment, having active instances of an application in multiple data centers provides for business continuity in the worst of disasters. If data center

A and data center B are both running copies of an application and a natural disaster takes data center A offline, in the worst case there will be a lag while global DNS is moved to data center B. In the best case, global DNS is not in the affected data center, and operations continue practically without interruption. Those people connected to the failed data center when it fails will have to reconnect to data center B, but no other user will notice the failover.

Energy consumption as a cost metric has grown in importance over the years and now must be a consideration in load balancing decisions. Spinning up a copy of an application in a data center might be less cost-effective from a TCO perspective than spinning up a copy in a cloud environment. The adaptability of the cloud allows such decisions to be made, and once a destination for a new copy of an application is determined, GSLB does not care where the application is hosted; it will be included in the rotation of connections regardless.

Likewise, contractual obligations—be they uptime requirements, general information security concerns, or specific data encryption requirements—must be met by an application no matter where it is served from. It's necessary to consider the capabilities of a given cloud provider or internal location as guidelines for where to deploy an application, but after such decisions are made, GSLB and global DNS will send traffic to the instance. The same applies to regulatory compliance issues. The decision-making process is all in where to start a copy of the application. GSLB automates everything else.

The ultimate goal of cloud balancing is to deliver an application to a user or customer as quickly as possible while using the fewest resources at the lowest cost.

Technical Goals of Cloud Balancing

From a purely technical perspective, the goals of cloud balancing are similar to those associated with traditional GSLB: ensure the availability of applications while simultaneously maximizing performance, regardless of the location or device from which users are accessing the application. Whether that access point is within an organization's data center utilizing private cloud resources or via a cloud provider, DNS requests are sent to the most appropriate location.

These technical goals are met through a combination of application and network awareness and collaboration between the global application delivery solution and local load balancing solutions. By coordinating across application deployments in multiple data centers, whether in the cloud or traditionally based, organizations can, through careful monitoring of capacity and performance-related variables, achieve optimal application performance while ensuring availability.

Business Goals of Cloud Balancing

As noted above, the business goals for application and delivery include minimizing costs, ensuring compliance with government and industry regulations, and meeting requirements specific to the line of business. These goals are increasingly difficult to achieve because the decision-making process requires the inclusion of variables that are nontraditional or unavailable for global application delivery solutions. Cloud balancing doesn't just balance applications across cloud implementations, however; it also helps balance business goals, such as cost reduction, with technical goals, such as automatic failover, response time, and availability metrics.

Cost-related variables include the expense of delivering an application based on the core costs associated with a specific deployment. For example, in most cloud computing environments, determining the total expense of delivering an application would require the use of a formula to calculate the costs incurred by the application instance, as well as those of the bandwidth used by the request and response. Because the costs might be highly dependent on the total resources used by the application during a specific period of time (such as monthly or weekly), this formula can very quickly become complex.

Compliance with regulations and contractual obligations, including SLAs, is even more complex. Variables regarding regulations and performance must be clearly defined so the global application delivery solution can incorporate them into the decision-making process. One viable way to take advantage of cloud balancing is utilizing it to achieve compliance by minimizing the investment necessary to deploy and implement specific services, such as application security or acceleration. An organization might choose to offer customers SLAs or services at a premium that includes additional application delivery options, and then subsequently choose to offer these options from a cloud-based environment to minimize the associated costs.

Cloud balancing also offers automation, which not only frees up human resources in IT but reduces errors by eliminating the manual performance of repetitive tasks. Applications can be deployed to the cloud with pre-configured templates for security, resources required, and monitoring. Routing decisions must be made in an automated fashion, but current cloud balancing solutions enable automated consideration of many more criteria. Device type, geographic location, time of day, and username are just a few of the variables that can be used when determining where to send an incoming customer for service.

To adequately meet contractual obligations, the application delivery infrastructure must be able to identify users in the context of request data such as IP address, pre-existing cookies, and credentials for which the obligation must be met. Secondly, the infrastructure must be able to correctly determine from which environment the obligation can best be met. The latter requires integration with the application layer of the infrastructure and the ability to provide metrics based on CPU and RAM utilization, response time, current load, and even financial cost per transaction.

Challenges (and Some Solutions)

There are multiple challenges involved in the implementation of a fully functional cloud balancing strategy. Some of these challenges are a result of the immaturity of current cloud-based offerings, and, as such, they might be automatically addressed as cloud environments continue to mature based on market demand and experience. Other challenges, however, are likely to require standards before they will be sufficiently addressed.

An Evolving Market

One of the first challenges is for organizations to find a cloud computing provider that meets its needs. Transparency in provider services is still in its infancy, and discovering specific service offerings can be time-consuming. Making this process more difficult is the dynamism of the market today. As providers' environments continue to evolve and providers react to the demands of customers and the market, offerings will inevitably change. By the same token, internal cloud computing has picked up momentum as one of the viable options, but a comparison of public cloud providers to a private cloud is not always a simple task, since the costs are in different scales. Purchasing a server to boost internal cloud capacity is a one-time event, for example, but adding capacity in a public cloud involves multiple monthly fees.

Application Portability

The lack of standards across cloud providers in regard to the migration of applications —and the deployment and delivery meta-structure that should accompany migration— makes application portability difficult, if not impossible, in many situations.

Further complicating portability—which ultimately will be a requirement for intercloud and cloud balancing solutions—is the lack of interoperability at the application layer. While virtualization is the primary mechanism through which applications are deployed into almost all cloud computing environments, virtualization can vary from proprietary to commercially available platforms. Proprietary platforms can make it challenging to implement a cloud balancing solution that incorporates local data center deployments. Commercially available platforms can provide easier implementations if the virtualization platforms are homogeneous, but a heterogeneous virtualization environment may prove as challenging as a proprietary platform.

Portability across cloud computing implementations will therefore need to occur at the container layer, with virtualization-agnostic environments that allow for the movement of the entire container across cloud computing boundaries. This portability may be achieved through a combination of APIs and the adoption of a single, virtual data descriptor model such as the Open Virtualization Format (OVF).

There has also been significant progress of late in application virtualization, allowing portability between servers that share a common operating system. This is another piece of the puzzle that will ultimately result in complete automation of the application delivery network. When an existing installation can be copied and moved onto an alternate infrastructure, only the ability to start and stop that infrastructure at will, based on demand, is missing in the equation to enable end-to-end cloud balancing automation.

Integration

For cloud balancing to be most effective, good integration is necessary between the global application delivery and local application delivery solutions. Cloud balancing depends on variables that require visibility into the local environment; thus, the global and local solutions must be able to share that variable information. Adopting a single-vendor strategy to address this challenge is certainly an option, but one that not many organizations are comfortable with—both because of reluctance to rely on one vendor for service and because it weakens their bargaining position at the licensing table. At the same time, there is no guarantee that every cloud will share the same vendors' solutions. Therefore, implementing a cloud balancing strategy requires a dynamic, cross-environment, and vendor-neutral solution. This neutral solution will almost certainly be found in standards-based APIs and Infrastructure 2.0 efforts.

Until a vendor-neutral solution is developed, organizations will need to leverage existing component APIs to achieve the integration of variables not typically associated with network-layer measurements, such as cost per transaction for both internal and external clouds. These variables can be calculated at a regular interval externally and then provided to the application delivery controller via its API to ensure decision-making data is up to date.

Architectural Continuity

Closely related to the challenge of integrating global and local application delivery solutions is architectural continuity. Having a standardized application delivery framework mitigates issues arising from operational differences across solutions and cloud computing environments. These issues include an increase in deployment costs and time while operators and administrators become familiar with different solutions.

While virtual appliances can resolve some of the issues arising from architectural inconsistency, they do not provide a total solution because some cloud computing models are not based on commercial virtualization technology and are proprietary in nature. This makes it difficult for an organization to replicate its architecture across clouds and maintain architectural continuity across cloud computing deployments.

One of the ways to address architectural similarity comes with the introduction of the virtual Application Delivery Controller (vADC). An ADC provides the local load balancing component required to implement a cloud balancing architecture, but there are no guarantees that cloud providers will have available the required load balancing solutions for customers. Deploying a vADC with the application in a cloud computing environment ensures the organization has the means to monitor and manage the health of that cloud-based application deployment. A vADC also provides for the architectural heterogeneity required by the global application delivery controller to include the myriad variables used in cloud balancing to make global application routing decisions.

A vADC can also provide a platform for global load balancing and DNS routing to enable all cloud implementations—internal and external—to behave in unison, as if one single network that offers the desired service based upon the best fit In geographic proximity, capacity costs, and other variables defined by the organization. With vADCs in the cloud architecture and a physical ADC in the primary data center, a coordinated response to changes in networking or application conditions can be implemented through automation. If a vADC suddenly stops responding, the GSLB and global DNS systems on the primary ADC can stop sending requests to that provider and alert operators of a problem.

F5® BIG-IP® Global Traffic Manager™ (GTM) is an ADC with both physical and virtual editions that can provide such a global load balancing platform, delivering continuity between the cloud and the data center, including integration with a local load balancing solution such as BIG-IP® Local Traffic Manager™ (LTM). Utilizing the physical edition BIG-IP GTM in data centers with heavy loads and the virtual edition to support a variety of cloud vendors enables organizations to address the needs of cloud balancing with wide IPs, global DNS, and GSLB.

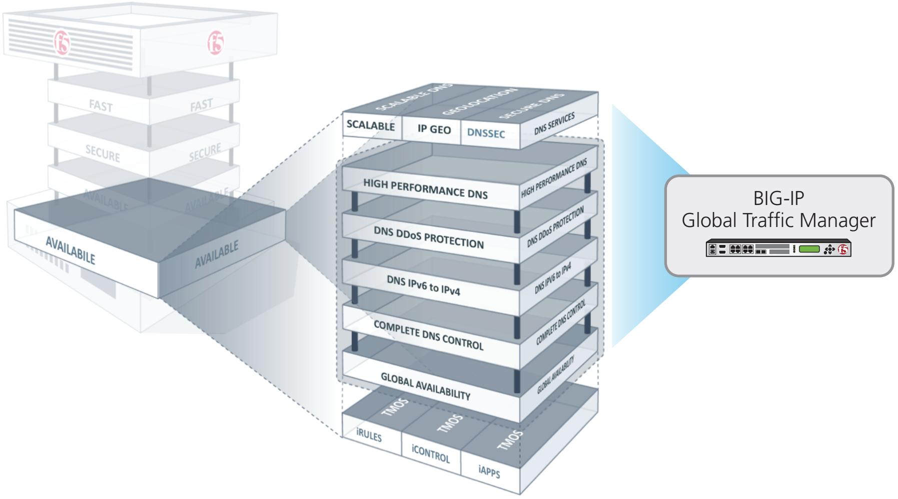

Figure 2: Extending GSLB and global DNS to the cloud

Security and Availability

One cannot control that which one does not have access to. That is a simple principal of life, not just computer science. Introducing a vADC with GSLB into a cloud environment offers the control of an ADC and the convenience of cloud deployment.

Security—from distributed denial of service (DDoS) protection to DNSSEC—requires a level of control that is not offered by most cloud providers today. Utilizing a virtualized GSLB solution in a cloud architecture to work hand-in-hand with a physical GSLB solution in the data center provides DNS DDoS protection and the flexibility to deploy DNSSEC in coordination with the physical, "master" GSLB device. Combining intelligent, geographic sensitive switching with these additional security measures provides peace of mind to IT staff while offering high availability even in the case of a natural disaster.

Conclusion

It is important to evaluate solutions for cloud balancing implementations with an eye toward support for the needs of an actual IT department. The global and local application delivery solution chosen to drive a cloud balancing implementation should be extensible, automated, and flexible, and the vendors involved need to look favorably upon standards. Meeting those criteria is paramount to ensuring the long-term success of a cloud balancing strategy. Combining high availability with security is just as important. When the organization is using a network that's not its own for mission-critical application delivery, stability and security become paramount.

Cloud balancing is still new, but the technology to add value is available today. The ability to distribute connections across the globe based upon an array of inputs such as geographic location, device type, the state of servers in one location or another, and balanced loads is real. There will no doubt be more advances in the future as cloud balancing becomes more mainstream. A solution that is poised to take on new standards and enables use of existing standards, such as IPv6 and DNSSEC, should be the first stop for IT in the quest for agile data centers.

Cloud computing has introduced a cost-effective alternative to building out secondary or even tertiary data centers as a means to improve application performance, assure application availability, and implement a strategic disaster-recovery plan. When they can leverage cloud application deployments in addition to local application deployments, organizations gain a unique opportunity to optimize application delivery from technical and business standpoints.

There are challenges associated with the implementation of such a strategy, some of which might take years to address. But the core capabilities of global and local application delivery solutions today make it possible to build a strong, flexible foundation that will enable organizations to meet current technical and business goals and to extend that foundation to include a more comprehensive cloud balancing strategy in the future.

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