



F5 White Paper

# Cloud Balancing: The Evolution of Global Server Load Balancing

Cloud balancing evolves global server load balancing from traditional routing options based on static data to context-aware distribution across cloud-based services.

**by Lori MacVittie**

Technical Marketing Manager, Application Services



# Contents

<b>Introduction</b>	<b>3</b>
<hr/>	
<b>Cloud Balancing</b>	<b>3</b>
Goals	4
Technical Goals of Cloud Balancing	4
Business Goals of Cloud Balancing	5
<hr/>	
<b>Challenges (and Some Solutions)</b>	<b>6</b>
An Evolving Market	6
Application Portability	6
Integration	7
Architectural Continuity	7
<hr/>	
<b>Conclusion</b>	<b>8</b>



## Introduction

While cloud balancing sounds like a futuristic concept, it is actually based on proven, well-understood global application delivery technology. Today, large organizations already use global application delivery to enhance application performance, ensure availability, and control the costs associated with application delivery. Cloud balancing provides the means by which organizations with smaller budgets and/or IT staff can also enjoy the myriad benefits of global application delivery.

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 cloud computing providers' environments even though the local application deployment, if there is one, might have more than enough capacity to serve that user.

## Cloud Balancing

Cloud balancing extends the architectural deployment model used in conjunction with global server load balancing to the cloud, which increases the choices available for organizations when determining from where a given application should be delivered. What is being slowly introduced into global application delivery is the ability to make application routing decisions based on variables other than those traditionally associated with network layer measurements. Customers in the midst of a decision-making process are demanding visibility into business metrics, such as the costs associated with responding to a given request, the ability to meet a SLA, and user device and location, among other factors. Plus, customers 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.

Cloud balancing itself does not provide a technical framework for collecting or supplying the variables required for making application routing decisions. That framework will be provided by the standards closely associated with Intercloud or Infrastructure 2.0. What is necessary from a technical implementation standpoint is the ability to integrate the network, application, and business variables into

### 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



the decision-making process. In other words, cloud balancing requires that the global application delivery strategy includes solutions that are contextually aware, made so by the use of Infrastructure 2.0 capabilities.

## Goals

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

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

Additionally, customers must consider business-focused variables, including:

- Cost to execute the request at a given location.
- Total cost to deliver the request to a user/customer.
- Regulatory compliance and/or legal restrictions.
- Services required by the user/customer to fulfill the request based on contractual obligations.

It is these business-focused, and admittedly difficult to incorporate, variables that make cloud balancing an attractive strategy for maximizing the performance of applications while minimizing the costs associated with delivering them.

The ultimate goal of cloud balancing is to deliver an application to a user as quickly as possible with the least amount of resources and for 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 global server load balancing: Ensure the availability of applications while simultaneously maximizing performance regardless of the location or device from which users are accessing the application.



These technical goals are met through a combination of application and network awareness and the 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 customer-specific obligations. These goals are increasingly difficult to achieve because they require the inclusion of variables in the decision-making process that are untraditional or unavailable for global application delivery solutions. However, cloud balancing doesn't just balance applications across cloud implementations; it also helps balance business goals, such as cost reduction, with technical goals, such as 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 requires 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 on a monthly or weekly basis), this formula can very quickly grow to be complex.

Compliance with regulations and contractual obligations, including SLAs, are even more complex. Variables regarding regulations and performance must be clearly defined so that the global application delivery solution can incorporate these variables into the decision-making process. One of the ways in which cloud balancing is a viable option is that it can be leveraged 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 include additional application delivery options, and then subsequently choose to offer these options from a cloud-based environment to minimize the associated costs.

In order to adequately meet such contractual obligations, however, the application delivery infrastructure must first be able to identify users using the context of



the request data such as IP address, pre-existing cookies, and credentials for which the obligation must be met; secondly, it 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 the ability to discover specific service offerings can be a time-consuming process. 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.

### Application Portability

The lack of standards across cloud providers in regard to the migration of applications—and the deployment and delivery metastructure 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, it 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; however, 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 Open Virtualization Format (OVF).

## Integration

In order for cloud balancing to be most effective, good integration between the global application delivery and local application delivery solutions is necessary. Cloud balancing is dependent on variables that might 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 many organizations are not comfortable with this strategy because of the reluctance to rely on one vendor for service and because it weakens bargaining position at the licensing table. At the same time, there is no guarantee that every cloud will share the same vendors' solutions. Therefore, in order to implement a cloud balancing strategy, a dynamic, cross-environment, and vendor-neutral solution must be used. 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, such as cost per transaction, that are not typically associated with network-layer measurements. 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 that of architectural continuity. This similarity alleviates 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 the need for architectural similarity comes with the introduction of a 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 cloud providers will have available the required load balancing solutions for customers. However, deploying a vADC with the application in a cloud computing environment ensures organizations have the means by which to monitor and manage the health of cloud based application deployments. A vADC also provides 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

## Conclusion

It is important to evaluate solutions for cloud balancing implementations with an eye toward support for forthcoming standards. Ensuring the global and local application delivery solution chosen to drive a cloud balancing implementation is extensible and flexible, and that the vendors involved look upon standards favorably, is paramount to ensuring the long-term success of a cloud balancing strategy.

Many of the elements that will add business value through cloud balancing are not yet available today. The challenge of moving forward with cloud balancing in what is an emerging, evolving market has pros and cons. Benefits include the ability to help crystallize the needs for cloud computing providers and encourage the implementation of supporting services. Risks include those normally associated with the early adoption of any technology or architecture: potential obsolescence as vendors and providers move quickly in new and different directions.

Nonetheless, cloud computing has introduced a cost-effective alternative to building out secondary or even tertiary data centers as a means to improve application performance, ensure application availability, and implement a strategic

**White Paper**

Cloud Balancing: The Evolution of Global Server Load Balancing

disaster-recovery plan. The ability to leverage cloud computing based application deployments along with local application deployments gives organizations—including those with limited resources—the unique opportunity to extend their ability to optimize application delivery from technical and business standpoints.

**F5 Networks, Inc.** 401 Elliott Avenue West, Seattle, WA 98119 888-882-4447 [www.f5.com](http://www.f5.com)

---

F5 Networks, Inc.  
Corporate Headquarters  
[info@f5.com](mailto:info@f5.com)

F5 Networks  
Asia-Pacific  
[apacinfo@f5.com](mailto:apacinfo@f5.com)

F5 Networks Ltd.  
Europe/Middle-East/Africa  
[emeainfo@f5.com](mailto:emeainfo@f5.com)

F5 Networks  
Japan K.K.  
[f5j-info@f5.com](mailto:f5j-info@f5.com)

