GSLB Moves to the Cloud

NS1.

WHITE PAPER 2017.4.13

Executive Summary

Server load balancing continues to be a core element of IT infrastructure, even as applications move from traditional data center architectures to the Cloud. There is always a need to intelligently distribute workloads across multiple servers, whether those servers are real or virtual, permanent or ephemeral.

However, there remains a chronic gap in the ability to reliably distribute workloads across multiple Clouds, multiple data centers and hybrid infrastructures. The result is maldistributed workloads and degraded application performance that could be avoided if workloads were better managed at a global level. In short, there is a need for better Global Server Load Balancing, or GSLB.

This white paper reviews the current state of available GSLB solutions, explaining how they work and their limitations. It then reviews the core capabilities an enterprise class GSLB solution should have for effectively addressing the limitations of current GSLB methods and solutions. Lastly, the paper describes how the Cloud based, managed GSLB capabilities from NS1 prevent workload maldistribution from occurring and deliver better overall application performance.

LOAD BALANCING AND THE ADVENT OF CLOUD COMPUTING

Load balancers (also referred to as application delivery controllers or ADCs) are widely deployed in data centers. Their function is to distribute workloads to back end servers so as to ensure optimum use of aggregate server capacity and better application performance. Intelligent distribution of workloads amongst data centers is called Global Traffic Management (GTM), or sometimes Global Server Load Balancing (GSLB). GSLB solutions use DNS to direct incoming requests to the data center best able to respond.

Providers in the traditional load balancer space include vendors such as F5, Citrix, Radware and Kemp Technologies. Their hardware ADCs have been the go-to solutions for infrastructure and operations teams for quite some time. Recently, software based solutions such as HAProxy, Nginx and Amazon ELB have emerged as enterprises have moved applications to the Cloud.

In light of this changing landscape, Gartner recommends that Infrastructure & Operations (I&O) teams "Augment your ADC portfolio with simple load-balancing and other Layer 4 through Layer 7 technologies that can be easily managed at scale, whether deployed on- or off-premises, recognizing you may have to introduce new suppliers." (Skorupa, 2016) This has implications for enterprises that rely on vendor specific GSLB solutions as they typically only work with the vendor's own ADCs.

There are two basic approaches to multi-datacenter, multi-cloud GSLB. One is to use a traditional managed DNS provider for basic traffic management. This has the advantage of being very easy to implement, low cost, requiring no capital outlay and reliable. Unfortunately it offers only minimal traffic management capabilities such as round robin DNS and Geo routing. These approaches do not prevent maldistribution of workloads because they use fixed, static rules rather than basing traffic routing on the actual, real time workloads and capacity at each data center. For example, Geo routing can only ensure users (and their workloads) get sent to the geographically closest data center. It cannot account for uneven distribution of users geographically, localized demand spikes, or server outages within a data center.

To address these limitations, many ADC vendors offer their own purpose built DNS appliances that have a tighter integration with their load balancers. These can make traffic management decisions based on actual utilization levels at each data center by receiving real time load and capacity information from the local load balancers.

However, this benefit comes with trade-offs that are unpalatable for many enterprises, including:

1. The performance of a DNS hosted at a single data center is not adequate to meet the needs of a global user base, but the cost and complexity of deploying a globally ubiquitous DNS is prohibitive for most enterprises
2. Attacks on DNS (DDoS) are widespread and difficult to mitigate. It becomes a single point of failure for the enterprise's internet facing services. The need to deploy and defend the DNS becomes an added operational and cost burden
3. DNS is a mission critical service that requires specialized skills to operate correctly with 100% availability. Most enterprises are not equipped to do this

4. Finally, these are typically high performance network appliances with high capex. And because they must be widely deployed, redundantly configured and defended from attack, the solution overall results in both high capex and high opex

As a result, it is no surprise that most enterprises that have deployed data center load balancers are not using the GSLB functions available from their load balance vendor. Those that have deployed GSLB functions are open to replacing them with a better solution. A superior approach is a Cloud based, managed GSLB solution that uses real time telemetry from load balancers to make intelligent traffic management decisions.

GLOBAL SERVER LOAD BALANCING AS A SERVICE

Global Server Load Balancing (GSLB) is best delivered as a Cloud based, managed service. The core attributes and advantages of such an approach are as follows:

1. **Open interface for ingesting real time telemetry.**
   Most companies currently using Cloud architecture have a hybrid architecture (RightScale, 2017), containing some private datacenter servers and some Cloud based. Because enterprises deploying hybrid infrastructures often use a mix of ADC types (both commercial and open source), the GSLB solution needs an open interface for collecting real time data from disparate ADC types
2. **Cloud based, as a Service**
   As discussed above, basic managed DNS does not offer very good traffic management, but is very attractive from the perspective of being globally available, performant and well managed. A Cloud based GSLB solution needs to retain those attributes while at the same time offering the true, real time GSLB capabilities available from the proprietary ADC vendors
3. **Reduce Capex and Opex**
   Cloud based GSLB as a service by definition reduces capex, as there is no need to purchase hardware or software appliances. Running one's own authoritative DNS includes the need to deploy globally for high performance, must be designed for redundancy, protected from attack, maintained and staffed 24x7. Thus a managed GSLB solution is likely to have lower capex and opex
4. **Proactive**
   An effective GSLB solution needs to do more than direct workloads away points of presence that are overloaded. It should prevent overload conditions from happening in the first place. This requires the capability to detect the onset of overload conditions and shift traffic appropriately, whether those conditions are due to demand spikes, loss of capacity, or both in combination

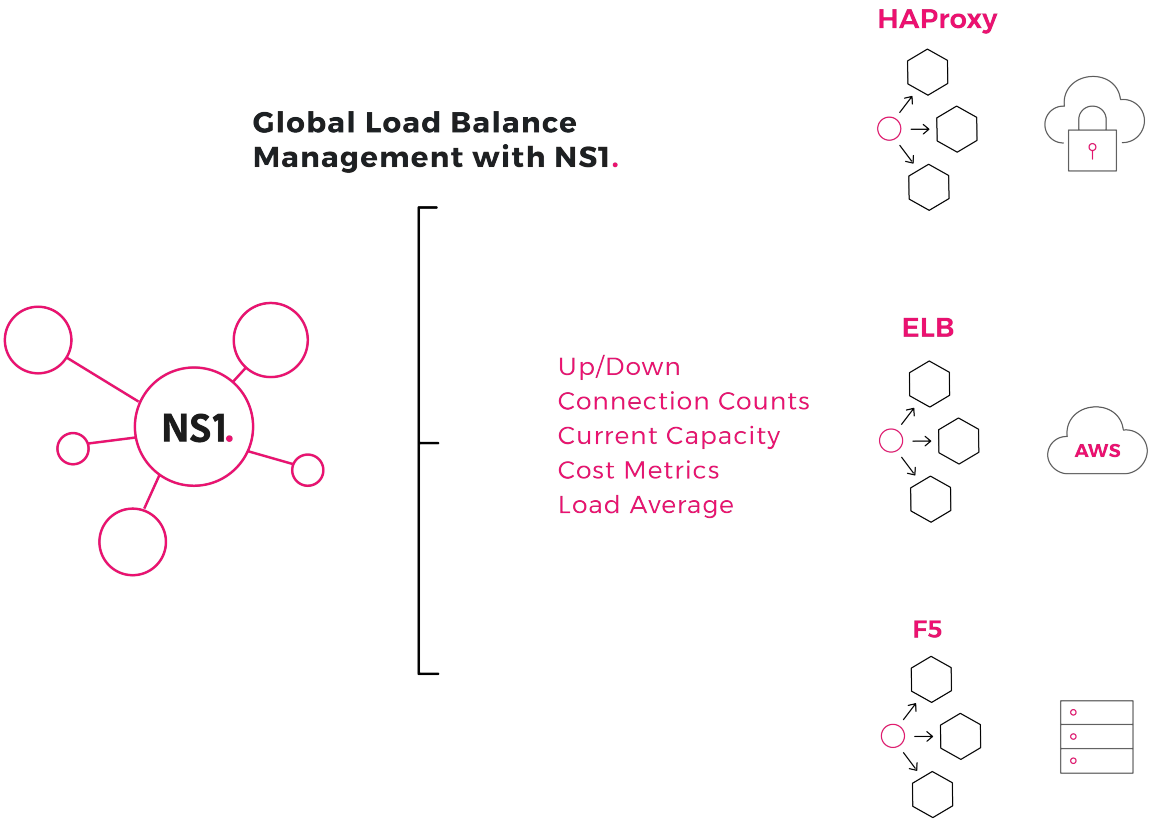
GLOBAL SERVER LOAD BALANCING FROM NS1

NS1 has developed a managed DNS based GSLB solution that addresses the requirements described above. The solution is based on three core capabilities that are unique to NS1:

The first key capability is the ability to attach meta data to DNS records. What this means is every DNS answer can have information associated with it that indicates both static and dynamic attributes that can be used to make intelligent traffic management decisions. Examples include (but are not limited to):

1. The network (ASN) the host is connected to
2. Bandwidth availability to the host
3. Whether the host is available (up/down state)
4. How busy the host is (connections, server load)
5. How costly it is to use this host

The second key capability is NS1 provides an open API for ingesting meta data. In the context of Global Server Load Balancing, the API enables load balancers to send current load conditions (e.g. number of active connections) and current capacity (e.g. number of servers or VMS in the backend) to their corresponding record in the system. The net result is the NS1 platform maintains a wealth of real time information regarding load and capacity at each Cloud based and traditional data center, regardless of the type of load balancers used. That information can be used in combination with other meta data such as geo location, network latency and cost to make optimal traffic management decisions.



The third key capability is NS1's exclusive Filter Chain™ technology. This enables the NS1 system to act on the meta data associated with the DNS records. It allows enterprises to build global traffic management routing algorithms that proactively prevent overload conditions from occurring, and steer user traffic to the best performing locations. In the example shown below, the Filter Chain performs the following traffic management logic:

1. Remove any non-responsive ADC from consideration. NS1's built in monitoring system can test the availability of ADCs and other hosts as frequently as every 5 seconds and update status accordingly

2. Order the available ADCs according to geographical proximity to the end user making the DNS request. NS1 tracks the location of end users by inspecting their source IP address
3. Assess current load and capacity of the ADC that is geographically closest to the user. If the load at that ADC exceeds a threshold apply a weighting factor to steer users to the next closest ADC. Thus in aggregate a proportion of incoming users will be shifted to the alternate ADC. If the load continues to grow the proportion of users sent to the alternate will become greater and greater. At the high water mark all users are sent to an alternate, until the load at the primary is alleviated

CONCLUSION

The combination of a globally performant, reliable managed DNS service offering advanced traffic management capabilities that were previously available only with proprietary ADC solutions is available now. This offers new opportunities for enterprises to prevent maldistribution of application workloads and deliver a better, more consistent end user experience.

Resources

Elias Khnaser. Gartner. Evaluation Criteria for Cloud Infrastructure as a Service. Published May 18, 2016. Retrieved March 22, 2017 from https://www.gartner.com/document/3322017?ref=solrAll&refval=182136052&qid=0cc5f69d865e0f020a8614c8d098eea8

Andrew Lerner, Joe Skorupa, Danilo Ciscato. Gartner. Magic Quadrant for Application Delivery Controllers. Published 29 August 2016. Retrieved March 16, 2017 from https://www.gartner.com/document/3426420?ref=solrAll&refval=181866471&qid=aeaf3c230f0bbbcc0bb60112cf3f1440

RightScale State of the Cloud Report 2017 Retrieved February 20, 2017 from http://www.rightscale.com

Joe Skorupa, Andres Lerner. Gartner. How I&O Teams Can Survive the Return of the Zombie Load Balancers. Published: 27 June 2016. Retrieved March 16, 2017 from https://www.gartner.com/document/code/308377?ref=imq_grbody

NS1.

+1.855.GET.NSONE (6766) NS1.COM @NSONEINC