CLOUDFLARE®

Cloud Load Balancing: Achieving Application Availability, Performance and Security

Executive Summary

The business impact of high latency and downtime for many businesses can run into the hundreds of thousands or even millions of dollars very quickly. Primary causes for these business challenges include server capacity constraints, application failures, and inefficient routes. Companies have traditionally deployed physical load balancers in their data centers to thwart latency and unavailability issues. But the dynamics of this solution approach is changing, shifting from application delivery control in data center environments to the edge of the network. To thwart high latency and unavailability, organizations are turning to load balancing. With this in mind, the white paper concludes with a checklist of nine criteria that an organization needs to consider when evaluating load balancing solutions.

The Business Impact of High Latency and Downtime

In today's digital age, customers, partners, and employees expect response times in a matter of a second or even milliseconds and have no tolerance for downtime. Slow online experiences result in lost revenues, dissatisfied customers, and diminished productivity, impacting companies of all shapes and sizes—B2C to B2B and small businesses to large enterprises across all industry segments.

High Latency

Speed matters, and it is measured in milliseconds. A study by Google found that 400 millisecond page-load times (.4 second) result in users conducting fewer web searches. It also uncovered that a 250-millisecond difference (.25 second) between your site and that of a competitor is enough to prompt customers to turn to the competitor's site instead.[1] So, what's the ideal page-response time? The same Google research reveals that the visual sensory memory processor in the human brain works in bursts of 100 milliseconds (.1 second).

Research shows that abandonment rates increase as page response times go up. Forty percent of visitors expect pages to load in two seconds or less, with a one-second delay resulting in a seven percent reduction in conversions for e-commerce sites. What does this mean in terms of revenue? For an e-commerce site generating $100,000 in sales per day, this translates into $2.5 million in lost revenue annually.[2]

One second counts

A one-second latency delay translates into a seven percent reduction in conversions. For an e-commerce business generating $100,000 in sales each day, this adds up to $2.5 million in lost revenue annually.

As mobile commerce continues to grow rapidly, already exceeding desktop transactions, the problem of latency for many companies grows in complexity. Mobile phones often have slower connections and low-powered CPUs that increase demands for faster response times. And the problem is going to continue growing in scope: even though mobile commerce in the U.S. is at 35 percent, it is growing rapidly at an annual rate of 17 percent and is expected to comprise more than half of online transactions very soon. Further, this does not account for the fact that consumers rely heavily on their smartphones when shopping in brick-n-mortar stores—using them to compare prices, check product reviews, and more.[3]

Latency not only aversely impacts e-commerce, but it reduces operational productivity [4] and efficiencies in manufacturing environments, financials services institutions, healthcare, and education. Research shows that latency above 30 milliseconds increases user annoyance and affects productivity. According to a report by SanDisk, the average employee wastes one week annually waiting on their company's network to respond. This equates to thousands of dollars in lost productivity for each worker. When multiplied across an entire workforce, this quickly tallies into the hundreds of thousands of dollars or even millions for some enterprises. [5]

Not available: unplanned downtime

In addition to latency, companies must deal with unavailability of services. In its annual survey of downtime, ITIC finds that 98 percent of organizations indicate that one hour of downtime equates to more than $100,000. Eighty-one percent report that one hour costs their businesses over $300,000. A frightening one-third say one hour translates into $1 to $5 million[9]. In a separate report, IDC pinpoints the average total cost of downtime for the Fortune 1000 to be between $1.25 and $2.5 billion per year.[10] And the cost will continue to burgeon; it rose between 25 and 30 percent since 2008.[11]

The Cause of Latency and Disruption

There are several causes of latency and downtime. These include inefficient routing, server capacity limits, and unpredictable network congestion.

Global Load Balancing

As companies expand globally, their customers, partners, and employees become farther from servers, thereby increasing latency. Without intelligent load balancing that accounts for user location (viz., geographical location-based steering), user sessions are not processed on servers closest to them. This often adds latency that degrades user experiences. Further, these same global models often fail to assess server health, and application sessions get sent to servers that are unhealthy and unable to provide high performance.

Server scalability

As Events such as traffic spikes or distributed denial-of-service (DDoS) attacks can overwhelm servers, causing latency and unavailability. The need for rapid failover and intelligent routing to other servers is critical. Seconds count for many applications such as online trading platforms, and it often can take a

Industry Examples: Linking High Latency to Conversion Rates

Amazon increased revenue by one percent for every 100-millisecond reduction in site latency[6] Yahoo increased traffic by nine percent for every 400-millisecond reduction in site latency[7] Walmart saw a sharp decline in conversion rates when site load time increased from one to four seconds[8]

Downtime repercussions

One hour of downtime equates to more than $100,000 in lost revenue and diminished productivity for the vast majority of businesses.

Downtime and the e-Commerce business

The average uptime for a business is 97.8 percent. This equates to eight years of downtime per year. For an e-commerce business with $100,000 in daily sales, this adds up to $800,000 in lost revenue per year.

DNS change at least 60 seconds to propagate. Instead, load balancing needs to reroute proxied in virtual real time, thus ensuring there is no impact on latency or disruption of services.

Misconfigured Servers and Applications

Monitoring the health of servers and applications is critical when it comes to latency and availability. When a server is unhealthy, users suffer through long application response delays and application outages. Load balancing that fails to monitor the health of servers in its network can inadvertently route traffic to a server that is unhealthy.

Cloud Platform Lock-In

When it comes to the cloud and load balancing, organizations require a solution that delivers flexible options not only across regions, but also across cloud providers. Thus, for organizations that rely on one cloud provider, they do not have the option to failover servers to other cloud providers in the event the servers for that cloud provider become unhealthy. High latency and unavailability occur when organizations are unable to direct traffic to healthy servers.

Internet Congestion

The amount of Internet traffic is exploding globally. Over half of the world's population now uses the Internet. Last year alone, the number of Internet users grew 10 percent—equating to 354 million new users.[12] The growth in traffic is much broader than new users. Social media users shot up 21 percent—or 482 million—in the same timeframe. Mobile traffic also continues to add to network congestion (grew five percent last year), and video consumption is exploding—with 80 percent of Internet traffic expected to be video by 2019.[13]

Rapid adoption of cloud-based apps and storage is another factor behind the explosion in Internet traffic. Businesses of all shapes and sizes are opting to transition away from monolithic, on-premises solutions to the cloud. Ninety-three percent of enterprises utilize cloud services in some form, and 79 percent of all workloads run in the cloud today.[14] All of this increases the amount of traffic and data traversing the Internet.

Concurrently, Internet content is becoming richer and more sophisticated, and thus the amount of network bandwidth they consume is increasing. These content-rich apps such as gaming, virtual reality, and augmented reality applications consist of more extensive code and employ bandwidth-intensive video and images.

60 seconds

It can often take up to 60 seconds for a DNS change to take effect.

Size of DDoS Attacks Spikes

The size of DDoS attacks quadrupled last year. The largest in 2015 was 500 Gbit(s). Last year, two separate attacks reached two terabytes.[15]

10 Million

The number of DDoS attacks this year is expected to exceed 10 million.[16]

Internet Traffic is Exploding

Over half of the world's population uses the Internet. Last year alone, the number of users shot up by 10 percent.

Video Becomes a Traffic Hog

Over 80 percent of Internet traffic in 2019 is expected to be video consumption. In addition to growing utilization of video in business as well as continued consumer adoption, growth in virtual reality and augmented reality applications will add to this upward spiral.

A Checklist for Evaluating Load Balancing Solutions

Organizations can follow a checklist when evaluating load balancing solution options. The following eight areas can be the difference between a subpar solution that does not scale to meet your business requirements or the size and volume of the security threats that exist today and one that does.

- **Cloud-Based:** Traditionally enterprise companies have deployed physical load balancers in their on-prem datacenters to address latency and availability challenges. While these solutions are well established and feature rich, the industry is rapidly shifting to public cloud and SaaS based solutions. Along with this shift comes a change in requirements such as significantly higher and global scalability, ease of use and price. Evaluate SaaS-based load balancers located at the network edge.

- **Global Content Delivery Network:** Load Balancing in the cloud can take advantage of a Global Content Delivery Network (CDN). The CDN caches static content eliminating the need for the Load Balancer to route this type of traffic to the origin server. This not only reduces the load on the origin server, it also reduces bandwidth costs and it accelerates performance since requests are served from a physically close edge server. A CDN with a large network capacity helps to absorb spikes in network traffic as well, which may be caused by seasonal holiday shoppers or corporate events and activities. Spikes in network traffic can create high latency and even interruptions in service causing lost e-commerce revenue, poor employee productivity, and disrupted supply chains and communications, among others. Look for a Load Balancer based on a content delivery network. The CDN should consist of a large number of globally distributed datacenters / network interchanges and should provide a large network capacity.

- **Global geolocation-based routing:** Geographic distance between the server and the visitor causes latency issues with websites, web apps or APIs. This is a concern for companies, which are expanding into new regions or which are operating on a global scale. Especially for dynamic content, which needs to be requested directly from the origin server, such as pricing or availability information, latency can have a significant business impact. Look for load balancers to provide the ability to connect visitors to infrastructure that is in the same part of the world e.g. send European customers to the London datacenter, Australian customers to the Sydney datacenter, and anywhere in-between

- **DDoS Relisient Service:** With the size of DDoS attacks growing, the global network needs to have the capacity to withstand even the largest DDoS attack, to ensure that the load balancer can always route traffic to healthy servers even when under stress. Carefully consider the size of today's DDoS attacks and the capacity of the global network. The smaller the capacity of the network, the more likely a DDoS attack can bring it down.

- **Near Real-Time Failover:** Seconds literally count on today's digital age. For e-commerce, B2B and B2C, customers will abandon their sessions and purchase from elsewhere. In the case of internal- or partner-facing applications, application delays or disruptions drain productivity and inhibit operations. Cloud based load balancers frequently rely on public DNS, which are plagued by slow propagation of changes, delaying the failover. Organizations need to look for a load balancer, based on DNS with fast time-to-lives (TTLs) to expire, to ensure that failover can occur in a matter of seconds.

- **Multi-Cloud Support:** Organizations increasingly rely on multiple cloud platforms for redundancy to avoid outages and to reduce vendor lock in. Look for Load Balancers with the ability to route traffic to healthy servers in different clouds or even on-premise environments.

- **Deployment and Flexibility:** No business of any size wants to spend significant time deploying and managing a load balancing solution. Here, look for a solution based in the cloud that can be configured and setup in minutes, that requires minimal management. Configuration should be easy and there should be support for a graphical UI and powerful API's. And as your business requirements change, whether you need additional network bandwidth or expand and open offices in a new location, the load balancing solution should provide you with the flexibility to easily and quickly adapt to those changes.

Sign up for Cloudflare today.
1 888 99 FLARE
enterprise@cloudflare.com
www.cloudflare.com

References

[1] Steve Lohr, "For Impatient Web Users, an Eye Blink Is Just Too Long to Wait," The New York Times, February 29, 2012.

[2] "How Loading Time Affects Your Bottom Line," Kissmetrics.com, accessed July 28, 2017.

[3] "Don't Get Left Behind: Improving Your E-Commerce Site Performance and Security for the Mobile Consumer," Cloudflare, 2016.

[4] "Latency," WhatIs.com, accessed on July 17, 2017.

[5] Mark Tyson, "Users Lose a Full Working Week Every Year Due to Slow Computers," Hexus.net, October 8, 2013.

[6] Nati Shalom, "Amazon Found Every 100ms of Latency Cost Them 1% in Sales," GigaSpaces.com, accessed June 15, 2017.

[7] Jennifer Bland, "How I Decreased My Website's Page Load Speed by More Than 50%," Ratracegrad.com, February 15, 2014.

[8] "Real User Monitoring," SlideShare, September 7, 2013.

[9] "Cost of Hourly Downtime Soars: 81% of Enterprises Say It Exceeds $300K on Average," ITIC, August 2, 2016.

[10] Alan Shimel, "The Real Cost of Downtime," DevOps.com, February 11, 2015.

[11] "Cost of Hourly Downtime Soars."

[12] "Internet Disruption Study: Strategic Market Research Report," Spiceworks, January 2017.

[13] Carla Marshall, "By 2019, 80% of the World's Internet Traffic Will Be Video," Tubular Insights, June 11, 2015.

[14] "Building Trust in a Cloudy Sky: The State of Cloud Adoption and Security," McAfee, January 2017.

[15] "Technology, Media, and Telecommunications Predictions," Deloitte Global, 2017.

[16] Ibid.

CLOUDFLARE®

1 888 99 FLARE | enterprise@cloudflare.com | www.cloudflare.com