CLOUDFLARE

Optimize web performance and reliability with load balancing best practices

Executive Summary

Every year, enterprises lose millions of dollars to website sluggishness and downtime — most of it in the form of lost revenue. Slow or unavailable sites and apps also negatively impact internal productivity and degrade search engine rankings. These performance and reliability problems can be caused by numerous factors, including:

- Overworked or unhealthy servers

- Geographic distance between end users and servers

- Slow DNS resolution times

- Distributed denial of service (DDoS) attacks

- The type of device a visitor uses to access the Internet.

Load balancers mitigate latency and availability problems by uniformly dispersing web traffic across a network of servers, ensuring that no single server becomes overwhelmed and that web assets will still be available even if one server fails. Traditionally, companies deployed physical load balancers in data centers, but as computing moves into the cloud, enterprises are gravitating towards more flexible, less costly, and easier-to-use cloud-based load balancing solutions.

However, not all cloud-based load balancing solutions are created equal. A robust solution will integrate with a global content delivery network (CDN) and offer features such as global geolocation-based routing, DDoS resiliency, layers 3 and 4 load balancing functionality, analytics capabilities, and near real-time failover. It will also seamlessly integrate into the multi-cloud and hybrid cloud data environments that most businesses have today.
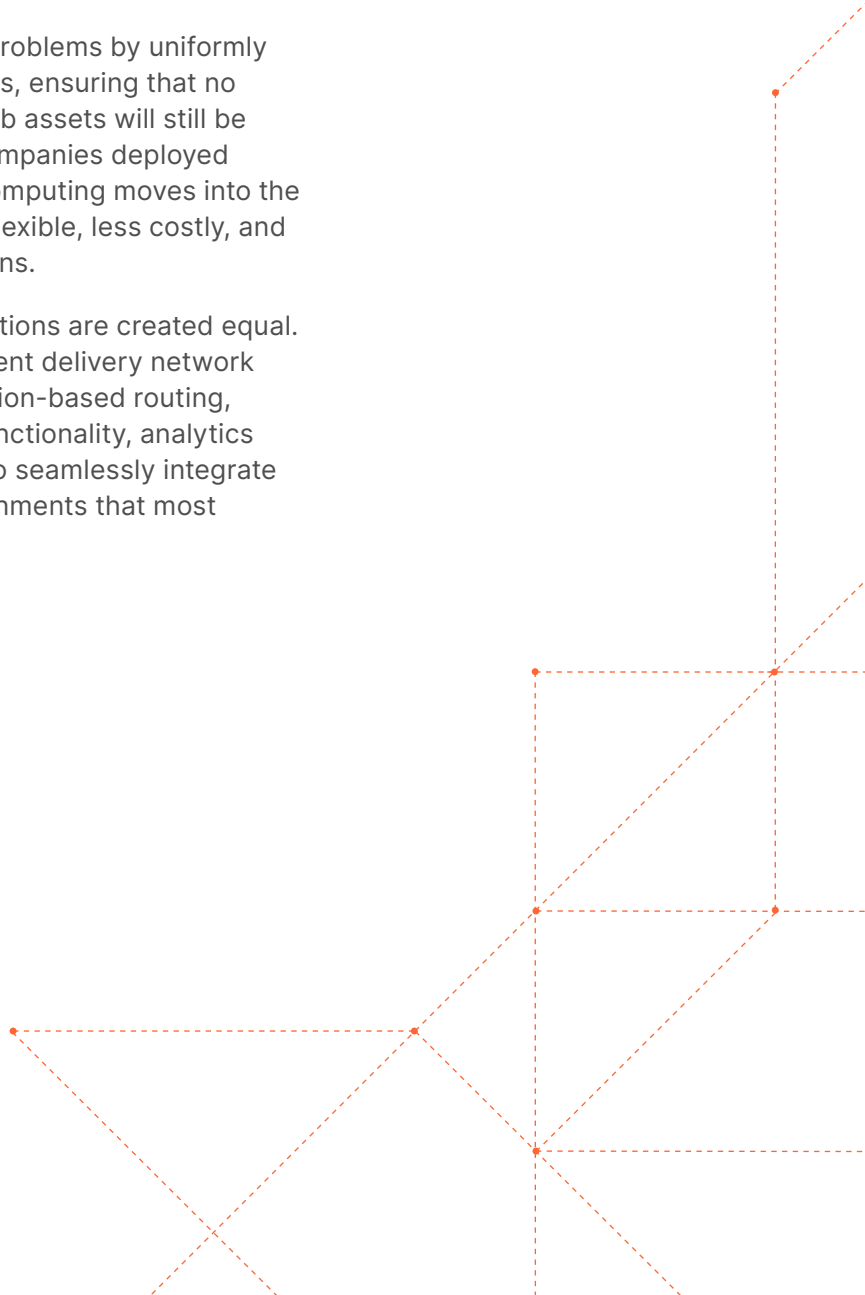
Causes of latency and downtime

Latency and downtime have significant negative impacts on the business. Companies can experience latency and downtime due to various causes

Unevenly distribute server workloads

Overutilized servers run more slowly as requests compete for limited resources. An overburdened server can reduce website and application performance or render them completely unavailable.

Effective load balancing distributes workloads uniformly across a network of servers, which can significantly improve performance. For example, one SaaS company's customers were having issues with latency across different regions globally. However, after deploying Cloudflare Load Balancing, they experienced an immediate improvement in latency and saw a 2-3 seconds improvement in page load times.[1]

Geographic distance

Global Internet penetration is exploding. **In January 2023, 64.4% of the world's population was connected, and over one hundred million people connected for the first time in 2022.**[2]

The globalization of the Internet has multiple impacts on network performance. As the number of active users grows, the available bandwidth per user decreases, causing delays.

In the past few years, users have also become more distributed with the expansion of remote work.What was previously considered east-west traffic within corporate environments is now moving north-south as it traverses the internet to reach remote users. This transition places additional load on the infrastructure and extends the round trip distance traffic travels from the user to the servers and back, adding latency.[3]

Site and application complexity

The Internet has undergone multiple stages of evolution, and each iteration adds more complexity to websites and applications. Modern websites are bulkier than ever, with total page size steadily climbing since 2011.[4]

Video conferencing, online games, and similar online services also add to the size and complexity of websites and applications. These applications consume significant bandwidth and are latency-sensitive, placing additional load and pressure to deliver on corporate networks and infrastructure.

Device type

Over 60% of web traffic is from mobile devices,[5] and about half of mobile users expect apps to respond in two seconds or less.[6] Designing and optimizing websites and applications for mobile devices is a necessity.

The emergence of 5G mobile networks does not guarantee high-speed, unconstrained network access for mobile users. Customer conversion rates depend on the ability to rapidly deliver content on mobile devices.

Slow DNS resolution

DNS resolvers translate domain names to IP addresses, providing computers with the information necessary to route a request for a web asset. DNS resolution is a vital first step to accessing online resources, and optimizing it is vital to maximizing performance.

Not all DNS resolvers are optimized for speed, and many DNS providers take 20-120 milliseconds to resolve each DNS query.[7] The fastest DNS providers will resolve queries in under 20 milliseconds; Cloudflare DNS, for example, resolves queries in 8.92 milliseconds on average.[8]

While these numbers may seem insignificant, it's important to consider that rendering a single page may require multiple HTTP and DNS requests. For example, **the average web page involves 71 HTTP requests on desktops and 66 requests on mobile.[9]** While some of these requests may be for the same domain, each unique DNS request adds latency.

Server health

Servers can fail for a variety of different reasons. If a server crashes, then the application(s) and web pages hosted on it may become unavailable to users.

Users are also consuming more video content, and, in today's world, if something goes viral, a large amount of traffic can render your services unresponsive. Adding load balancing solutions and redundancy into IT infrastructure is essential to protect against legitimate traffic taking down applications just like a DDoS attack would.

Load balancing solutions should monitor server health to maintain application availability. Otherwise, traffic may inadvertently be routed to a server that is experiencing problems, resulting in long delays or outages for users.

Cyber attacks

Distributed denial-of-service (DDoS) attacks pose a significant threat to the health and availability of online services. DDoS attacks flood web servers with spam requests, drowning out legitimate traffic and potentially acting as a smokescreen to conceal other attacks.

The growing number of insecure Internet of Things (IoT) devices — a common target for DDoS botnet malware — has contributed to a rise in DDoS attacks. **In Q4 2022, DDoS attacks increased 79% YoY.[10]**
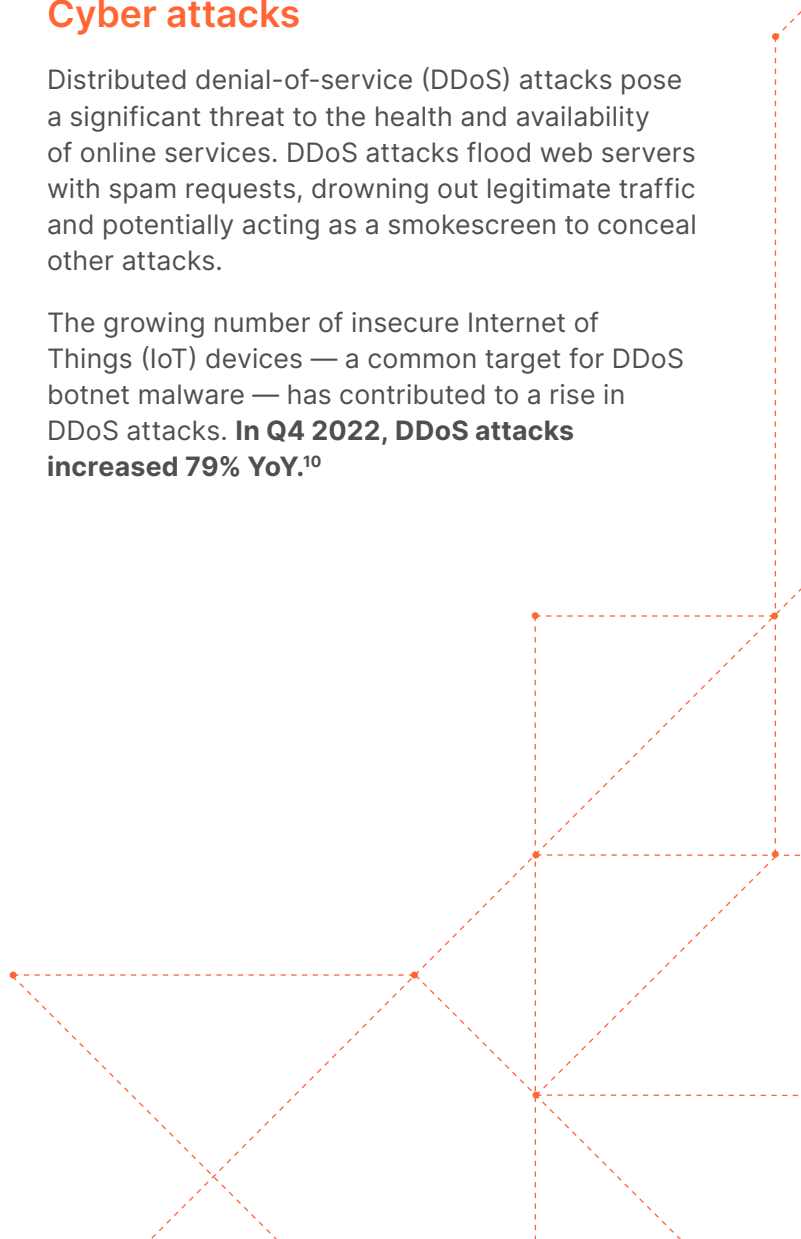
Cost of latency and downtime

Network latency and site load times have a significant impact on the customer experience and conversion rates. In fact, delays as short as 100 milliseconds have a measurable impact on consumer behavior.

Latency can have various negative impacts on the business. Common costs of latency and downtime include:

- **Revenue Loss:** Companies increasingly connect and provide services to their customers via their websites. Downtime and latency can result in missed sales opportunities when customers cannot reach an organization's website or abandon their cart due to slow page load times.

- **Customer churn:** Slow page load times equate to lost sales. **A page with a 1 second load time has a 3x higher conversion rate than one with a 5 second load time.[11]**

- **Lost productivity:** Latency and downtime for internal applications also impact employees' productivity. For example, **the average U.S. employee spends about 1 second waiting for an app per minute of usage.[12]** This equates to losing over 4 days of work per year.

- **Brand visibility:** Google uses page speed as a ranking factor for both desktop and mobile search.[13] Pages with slow load speed can harm brand visibility.

- **Legal and regulatory compliance:** Providers of online services are likely bound by service level agreements (SLAs) that include availability and uptime. Downtime and latency can result in penalties and the potential for legal action.

Downtime is expensive for the business. While the average cost of downtime is about $9,000 per minute,[14] this varies based on industry and the size of the business. For example, Facebook lost an estimated $90 million in a 14-hour outage for a cost of over $107,000 per minute.[15]

Understanding Load Balancing

Latency and downtime carry significant costs to an organization. A load balancer is a service that sits between a network of origin servers and the Internet and can help to mitigate these costs by evenly distributing across multiple servers. This ensures application reliability, efficiency, and responsiveness by ensuring that individual servers do not become overwhelmed by traffic spikes.

Why do we need load balancers?

When an end user visits a web page, an origin server receives and responds to this request. This involves processing the request, collecting the desired content, and sending it to be rendered in the user's browser.

The number of requests that a single origin server can handle depends on the physical infrastructure and the code complexity. However, the number of requests that a website receives can outpace even the best hardware and most performant web application. If this happens, then requests may need to wait in a queue — increasing latency — or are dropped entirely.

A load balancer prevents individual servers from falling victim to these issues. Load balancers sit between the end user and a cluster of origin servers and uniformly balance the load across the server pool. By reducing the load on each server, a load balancer improves website performance and resiliency.
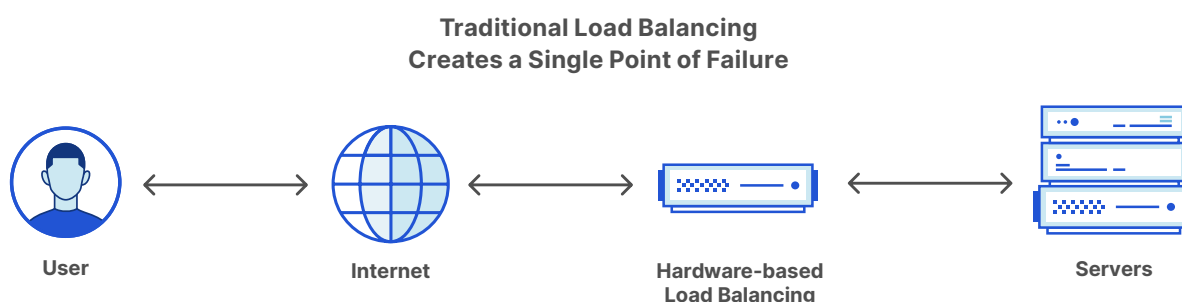
Legacy load balancers

Traditionally, load balancers were deployed in on-premises data centers. Often, these were implemented using dedicated hardware, but virtualized options were also available. To ensure resiliency, these devices were commonly deployed in pairs so that the backup system could take over if the primary one failed.

These legacy hardware-based load balancers had significant limitations.The challenges that they created include the following:

- **Up-front costs:** Load balancer appliances must be purchased and installed before use. This can be expensive, and all costs are incurred upfront.

- **Scalability:** Hardware-based solutions have a set maximum capacity, and the load balancer may become a bottleneck if the organization experiences exceptional surges in traffic. As an organization's bandwidth needs grow, existing solutions must be augmented or replaced with new hardware.

- **Geographic limitations:** Load balancer appliances can only be deployed in data centers where companies can install physical hardware. As a result, they can only manage traffic to on-prem applications, not cloud-based ones.

- **Skill gaps:** An in-house load balancer must largely be configured and operated by in-house personnel. Companies may struggle to attract and retain employees with the necessary skill sets.

- **Lack of flexibility:** Hardware load balancers are appliances connected to an organization's physical network infrastructure. This makes it difficult for companies to adapt to changing requirements.

**Traditional Load Balancing
Creates a Single Point of Failure**

User   Internet   Hardware-based   Servers
                                Load Balancing

Next Generation, Cloud-based Load Balancers

The vast majority of companies are rapidly moving to the cloud. 87% of organizations have multi-cloud infrastructure, and 72% have hybrid cloud environments that incorporate both public and private clouds.[16] A growing percentage of corporate apps can no longer sit behind hardware load balancers.

A robust standalone cloud-based load balancer can be used in conjunction with traditional hardware-based devices in hybrid environments, as well as with load balancers native to public clouds. A standalone load balancer is a neutral, cloud-agnostic layer that sits atop an enterprise's hardware-based and public cloud-native load balancers. The enterprise selects a primary provider to direct all traffic to. When the load balancer detects a failure, it automatically routes traffic to backup providers or regions. If the enterprise experiences outages or intermittent network connectivity in a public cloud or its own infrastructure, the standalone cloud-based load balancer automatically fails over to healthy providers or servers.

Virtualized load balancers can be deployed in the cloud to manage traffic to these applications. These cloud-based load balancers provide various benefits, including the following:

- **Virtually Unlimited Scalability:** Cloud load balancers have the advantages of cloud flexibility and scalability. Additional capacity can be quickly spun up as needed to manage surges in traffic to corporate web applications.

- **Cost Savings with Usage-based billing:** Cloud load balancers are commonly available under service-based models. Companies only pay for the capacity that they use rather than purchasing oversized appliances.

- **Greater Geographic reach:** Cloud load balancers should ideally run on a network with a global presence, putting them within close reach of applications living anywhere.
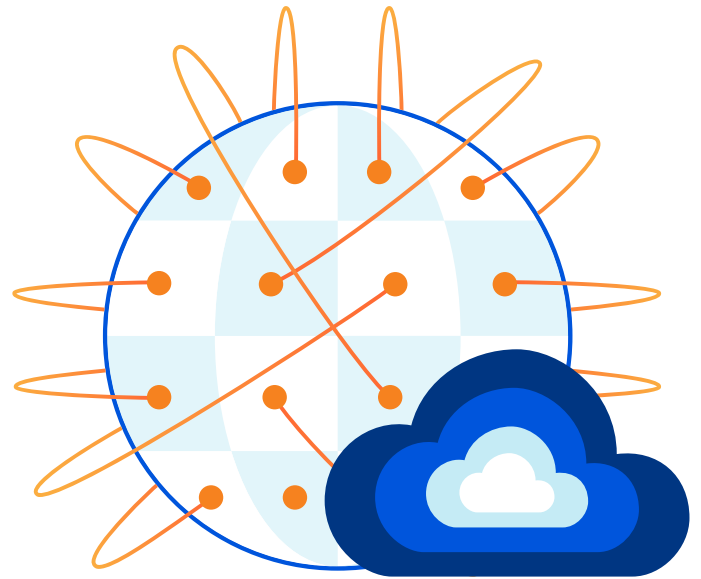
- **Ease of Configuration and management:** If a load balancer is offered as a service, the service provider performs much of the configuration and management. This reduces the overhead for the organization and its need for specialized personnel.

- **Flexibility**: A standalone cloud load balancer can easily be reconfigured or moved to support applications operating in a new environment. This enables companies to rapidly adapt to change and avoids vendor lock-in.

- **Resiliency:** Cloud-based load balancers can take advantage of the built-in resiliency and availability guarantees of the cloud. This reduces the risk that an outage will take the applications behind the load balancer offline.

- **Consolidation of Features:** With a cloud based solution, after onboarding with Load Balancing, it's easy to add additional modules such as Web Application Firewall (WAF), Bot Management, etc. as needed without any additional effort. With hardware solutions, upgrades commonly require either replacing the whole hardware device or adding physical modules or blades. These modifications can force companies to schedule maintenance downtime, which can leave customers without protection and negatively impact businesses.

What to look for when evaluating cloud-based load balancing solutions

Cloud-based load balancing solutions can help an organization drive down latency and downtime and decrease their impacts on the business. When evaluating load balancing solutions, look for the following features.

## Integration with a global content delivery network (CDN)

Load balancers and CDNs are both solutions designed to reduce latency and improve availability. A CDN caches static content at the network edge, reducing the distance that requests and responses need to travel. Additionally, by serving content from distributed CDN servers, the load at the origin server is reduced.

Integrating load balancing with CDNs optimizes content delivery. The load balancer distributes requests across CDN clusters and origin servers to optimize performance and minimize bandwidth consumption.

Global geolocation-based routing

The geographic distance between the server and the end user has a dramatic impact on the latency of the request and response. A load balancer should route traffic to the nearest available infrastructure, minimizing the distance that it needs to travel. For example, U.K. traffic should be directed to a data center in London, not one in New York.

The load balancer should also offer optimized, fast DNS lookups. For example, DNS queries should be directed to the nearest, healthy DNS server to minimize the latency incurred by DNS lookups.
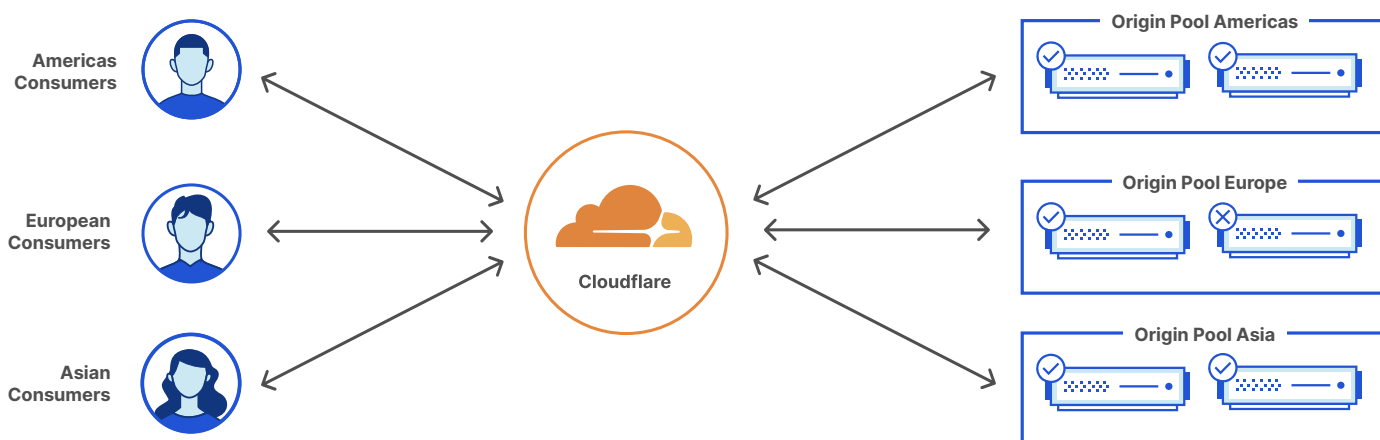
Unification of application delivery and security

Load balancer and CDN networks must be designed to address various security concerns. For example, DDoS attacks pose a significant threat to server health and availability. As a result, CDN networks should be scaled and secured to withstand even the largest DDoS attack.

Another major concern for load balancers and CDNs is compliance with privacy and security standards. For example, load balancers should support the use of TLS/SSL to encrypt customer data and authenticate web traffic.

Layers 3 & 4 load balancing functionality

DDoS attacks can operate at multiple layers of the OSI model. Volumetric DDoS attacks flood a web server with large volumes of traffic sent to the ports that implement various services. For example, DDoS attacks can target SMTP ports to disrupt email or the custom ports used to implement custom gaming protocols and other online services. A load balancer should have protection against Layer 3/4 DDoS attacks and sufficient capacity to maintain normal service during these attacks.

Near real-time failover

Cloud-based load balancers frequently rely on public DNS, which is plagued by slow change propagation, delaying failovers in the event of problems. A load balancer should use a DNS resolver with short time-to-live (TTLs), ensuring that failover can occur in a matter of seconds.

Multi-cloud and hybrid cloud support

Most companies have multi-cloud environments or hybrid cloud environments. To avoid vendor lock-in, reduce complexity, and minimize misconfigurations in multi-cloud and hybrid environments, make sure the load balancing solution is a neutral layer that can work both on-premise and in any public cloud.

A vendor-agnostic load balancer won't replace cloud vendors' native load balancers or traditional hardware appliances. However, it can work in tandem with them so that multi-cloud infrastructure functions smoothly.

Automation and DevOps support

Load balancers are designed to distribute requests across a cluster of servers. With the emergence of agile and DevOps processes and cloud computing, corporate applications infrastructure may be constantly changing, making this a moving target.

Relying on human operators to define and implement configuration changes poses significant risks to availability and performance. Load balancers should integrate automation and DevOps support to ensure that changes can be made rapidly at scale as corporate IT infrastructure evolves.

Ease of use

Configuration and management of load balancing solutions can be a time-consuming and resource-intensive task for skilled personnel. A good cloud-based load balancer can be configured and set up in minutes and should require minimal management. There should be support for a graphical UI and powerful APIs, and the solution should be easily reconfigurable to support evolving business needs.

Detailed Analytics

Load balancers' location between end users and applications is for collecting actionable business intelligence. Load balancers have visibility into customer behavior, application performance, security posture, and other operational insights. A load balancing solution should capture these analytics and integrate with your existing analytics provider.
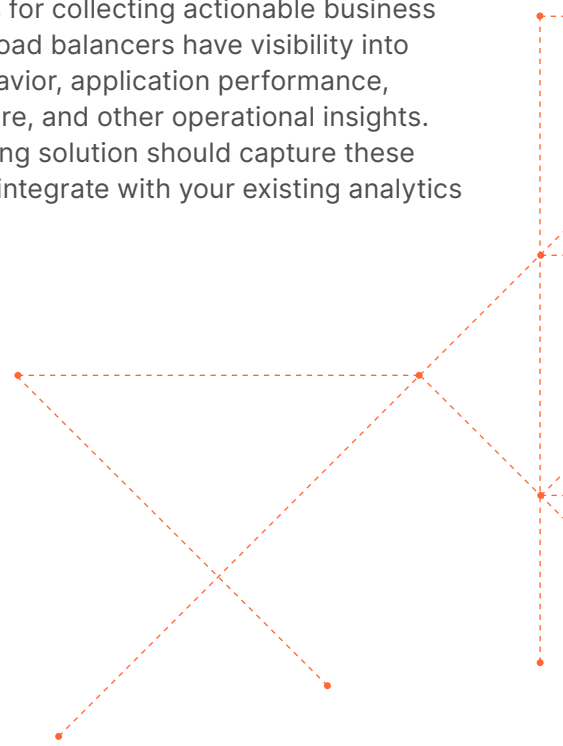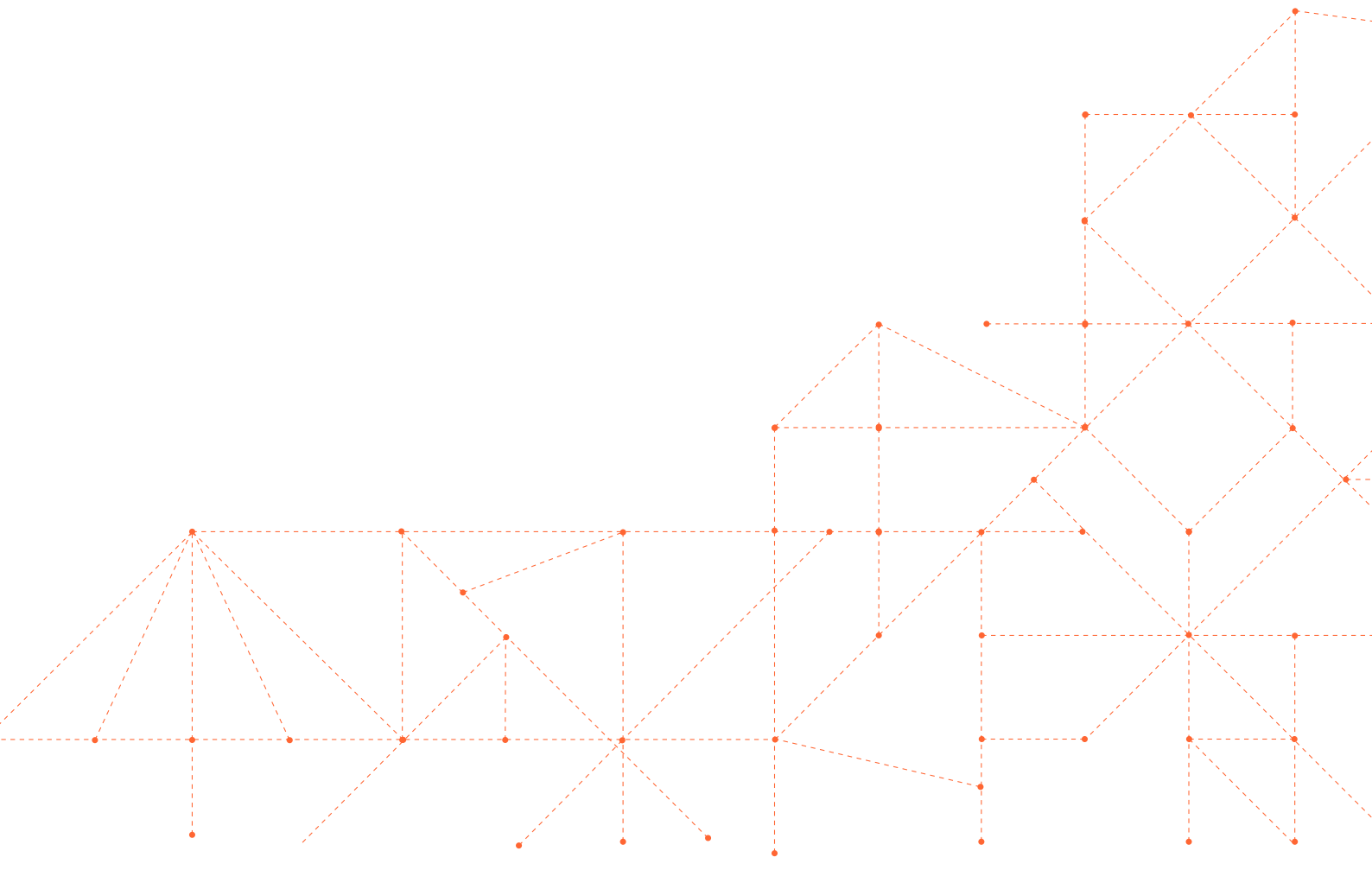
Conclusion

Modern websites and applications will not perform properly or remain consistently online without the use of a load balancer. A robust cloud-based load balancer is a much better choice than a traditional hardware-based solution.

In addition to being less expensive, easier to use, and scalable, a standalone cloud-based load balancer augments both traditional hardware-based load balancers as well as proprietary solutions offered by public cloud providers, ensuring that web assets always remain available and performant.

Cloudflare's global network and high-performance CDN help organizations maximize availability and minimize latency.
Learn more about Cloudflare Load Balancing.

References

1. https://www.cloudflare.com/case-studies/crisp/

2. https://datareportal.com/global-digital-overview

3. https://www.techwalla.com/articles/network-latency-milliseconds-per-mile

4. https://httparchive.org/reports/state-of-the-web#bytesTotal

5. https://gs.statcounter.com/platform-market-share/desktop-mobile/worldwide/#yearly-2011-2022

6. https://techbeacon.com/sites/default/files/gated_asset/mobile-app-user-survey-failing-meet-user-expectations.pdf

7. https://sematext.com/glossary/dns-lookup-time/

8. https://www.dnsperf.com/

9. https://httparchive.org/reports/state-of-the-web#reqTotal

10. https://blog.cloudflare.com/ddos-threat-report-2022-q4/

11. https://www.portent.com/blog/analytics/research-site-speed-hurting-everyones-revenue.htm

12. https://www.apmdigest.com/the-impact-of-app-performance-on-productivity

13. https://developers.google.com/search/blog/2018/01/using-page-speed-in-mobile-search

14. https://www.vertiv.com/globalassets/documents/reports/2016-cost-of-data-center-outages-11-11_51190_1.pdf

15. https://www.ccn.com/facebooks-blackout-90-million-lost-revenue/
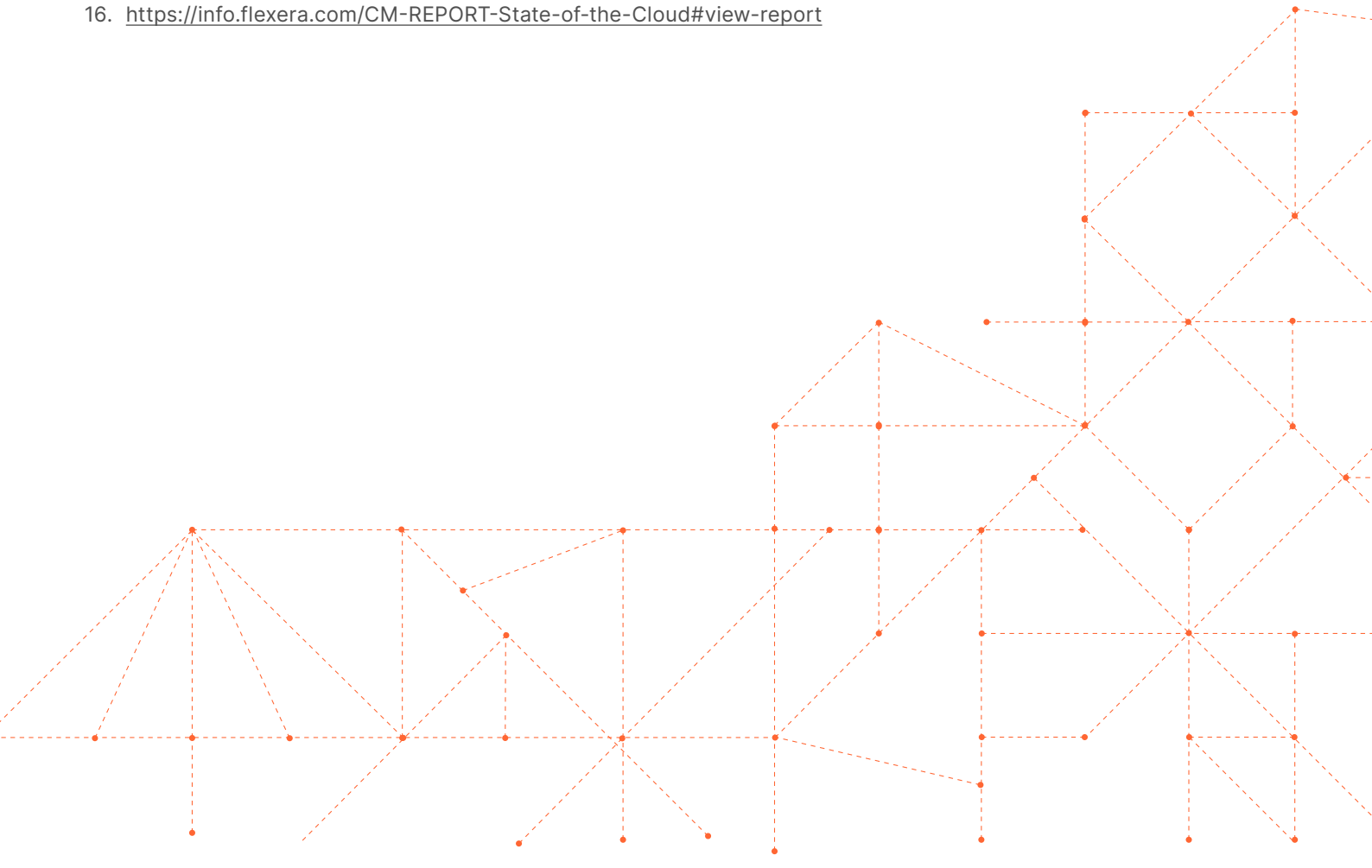
16. https://info.flexera.com/CM-REPORT-State-of-the-Cloud#view-report

CLOUDFLARE

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

REV:BDES/4505.2023APR27