CLOUDFLARE®

Load Balancing for High Performance & Availability in the Cloud

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

CLOUDFLARE

I. Executive Summary

Every year, enterprises lose millions of dollars to site sluggishness and downtime, most of it in the form of missed business opportunities. Slow or unavailable sites and apps also negatively impact internal productivity and degrade search engine rankings. Latency and availability 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, and even the type of device a visitor is using 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 multicloud and hybrid cloud data environments that most businesses have today.

II. Understanding Load Balancing

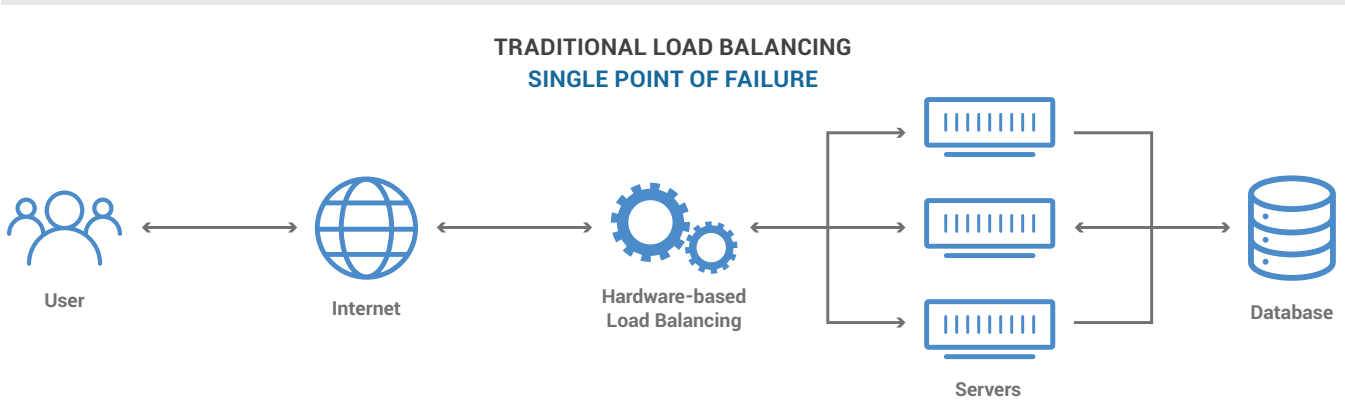
A load balancer is a layer that sits between a network of servers and the internet, managing the flow of information between the servers and end users. The purpose of load balancing is to evenly distribute workloads across multiple servers. This ensures application reliability, efficiency, and responsiveness by ensuring that individual servers do not become overwhelmed during traffic spikes. Load balancing also provides failover in the event of a server crash. Load balancers monitor server health, and if one server goes down, the load balancer simply routes the traffic through healthy servers.

Traditional Load Balancers

Traditional load balancers are hardware devices deployed in on-premise data centers. They are usually deployed in pairs to provide backup if one device fails.

- Hardware-based load balancers have numerous drawbacks.
- They must be purchased upfront, and the cost may be significant.
- They do not scale. To determine how many load balancers to purchase, an enterprise must calculate how much traffic they expect their website or app to generate. If traffic is lighter than expected, the enterprise is stuck with capacity they don't need. If traffic is heavier than expected, end users will experience sluggishness or downtime until new devices are purchased, configured, and installed.
- They run specialized operating systems and can be quite tricky to configure and maintain, adding to their total cost of ownership (TCO).
- They can only be used in data centers. Deploying applications in a cloud requires a virtual appliance, which must be uniquely configured for each cloud or data center it will operate in.

Cloud-based load balancers are less expensive, easier to use, and are more suitable for today's computing environments, where 94% of enterprises have migrated to the cloud, and 84% run multicloud environments.[1] Because they use an elastic network of servers, cloud-based load balancers give enterprises the flexibility and scalability to immediately adjust to both seasonal traffic spikes and long-term changes in workload.

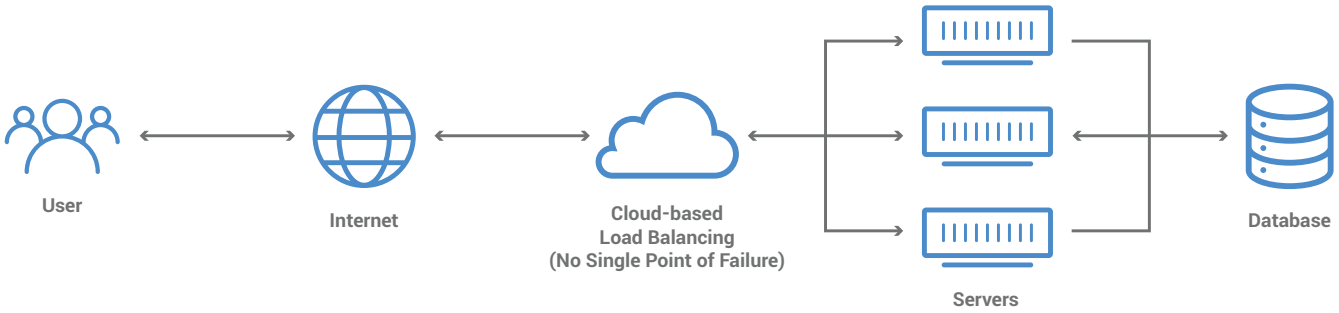


Next Generation, Cloud-based Load Balancers

While all public cloud providers offer load balancers, they aren't platform-agnostic. They are native to the vendor's cloud and can only be used with applications running in that provider's environment. If an enterprise wants to move the application to another cloud provider or run it on-premise, the load balancer won't move with it, forcing the enterprise to reconfigure load balancing each time they want to move an application. The situation is even more complex for the 58% of organizations that have hybrid cloud environments[2] and may be using traditional load balancers on-premise.

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, vendor-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 their own infrastructure, the standalone cloud-based load balancer automatically fails over to healthy providers or servers.

NEXT GENERATION, CLOUD-BASED LOAD BALANCING



III. Challenges solved by Cloud-based Load Balancers

Enterprises looking for a cloud-based load balancer have a number of different options to choose from. Selecting the right solution for your enterprise's needs requires an understanding of the challenges that load balancers solve — latency and downtime.

Costs of latency and downtime

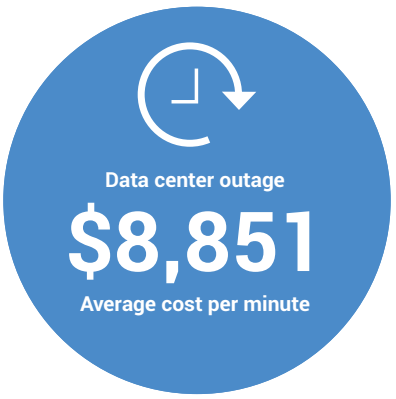
At the dawn of the millennium, the human attention span was 12 seconds long; today, it is only eight seconds.[3] This has profound implications for online businesses of every size and in every industry. Today's digital consumers demand websites, applications, and APIs that load instantaneously and are never offline. Recognizing this, Google uses page speed as a ranking factor for both desktop and mobile search.[4]

Even tiny delays can noticeably impact engagement and conversion rates. Latency becomes noticeable to the average user at 30 milliseconds,[5] and delays as short as 100 to 400 milliseconds have a measurable impact on consumer behavior.[6] Just one additional second of load time can cause conversions to drop by 7%.[7] Latency also damages operational productivity. The average employee wastes one week annually waiting on their company's network to respond.[8]

Left unchecked, latency results in the worst-case scenario of websites and applications becoming unavailable altogether The cost of down-time is rising dramatically. In 2010, the average per-minute cost of a data center outage was USD $5,617; by 2016, this had risen to USD $8,851. Most of these costs stem from end-user productivity, lost revenue, and business disruption.[9]

Waiting On Networks

The average employee wastes one week annually waiting on their company's network to respond.



Causes of latency & downtime

Since many variables impact how quickly web assets load, enterprises face a perennial challenge in striving to achieve low latency and high availability. Among other factors, latency and availability are impacted by:

UNEVENLY DISTRIBUTED SERVER WORKLOADS

Over-utilized servers run more slowly, which can cause websites and applications to run sluggishly or even go down altogether. Distributing workloads uniformly across a network of servers maximizes performance and prevents downtime. Effective load balancing can significantly improve performance; one SaaS company experienced a 2-3 second improvement in page load times after deploying Cloudflare Load Balancing.[10]

GEOGRAPHIC DISTANCE

Global Internet penetration is exploding. In 2019, 57% of the world's population was connected, and over one million people per day were coming online for the first time.[11]

This impacts speed and availability in two ways. More people online means less bandwidth to go around, and the distance between users and servers vexes businesses with global customer bases. It's estimated that every 100 miles of geographic distance between an app or website's resources and an end user adds 0.82 milliseconds of latency.[12]

SITE AND APP COMPLEXITY

Internet content is becoming richer and more sophisticated, making modern websites bulkier than ever. Total page size has steadily climbed since at least 2011.[13] Content-rich apps, such as those used for games, virtual reality, and augmented reality applications, have caught on world-wide; video games are now the world's most popular form of entertainment.[14]

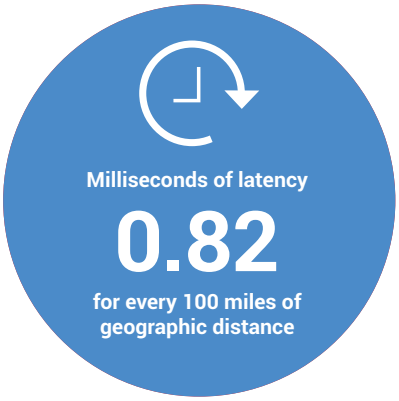
Game sizes were once constrained by the size of physical delivery mediums such as CD-ROM discs. Thanks to wide-spread high-speed Internet access, today's games are limited only by end user bandwidth and hard drive space. The high-resolution videos, 5.1 surround sound audio, and intricate textures used to create modern games have caused file sizes to balloon. In the mid-2000s, Red Orchestra 1 was considered quite large at 2.6GB. Forza Motorsport 7, released in 2017, is a whopping 96.5GB.[15]

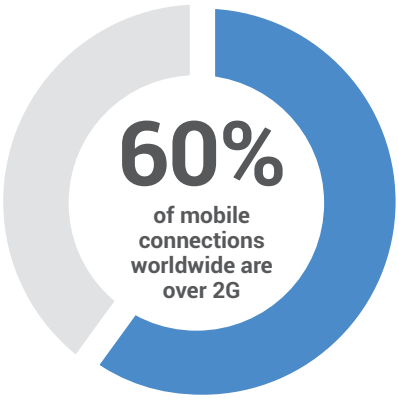
DEVICE TYPE

Optimizing sites and apps for mobile is no longer optional. Nearly 60% of web searches originate on mobile,[16] and about half of mobile users expect apps to respond in two seconds or less.[17] Achieving high availability and low latency on mobile presents a unique set of challenges. Mobile performance is constrained by network connectivity and availability. Despite the widespread availability of 4G and 5G networks in some countries, 60% of mobile connections worldwide are over 2G.[18] In some regions, mobile network providers will throttle bandwidth past a certain amount.[19]

Page Size

Total page size has steadily climbed since at least 2011.





SLOW DNS RESOLUTION

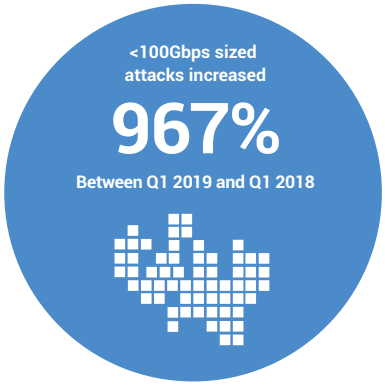
When users access a web asset, their devices must query a DNS resolver that will map the asset's domain name to its IP address, then send the correct IP address back to the device. This is called DNS resolution, and optimizing it is an important part of optimizing performance. Not all DNS providers are optimized for speed – many DNS providers take over 50 milliseconds to resolve each DNS query. The fastest DNS providers will resolve queries in under 20 milliseconds; Cloudflare DNS, for example, resolves queries in under 12 milliseconds on average.[20]

SERVER HEALTH

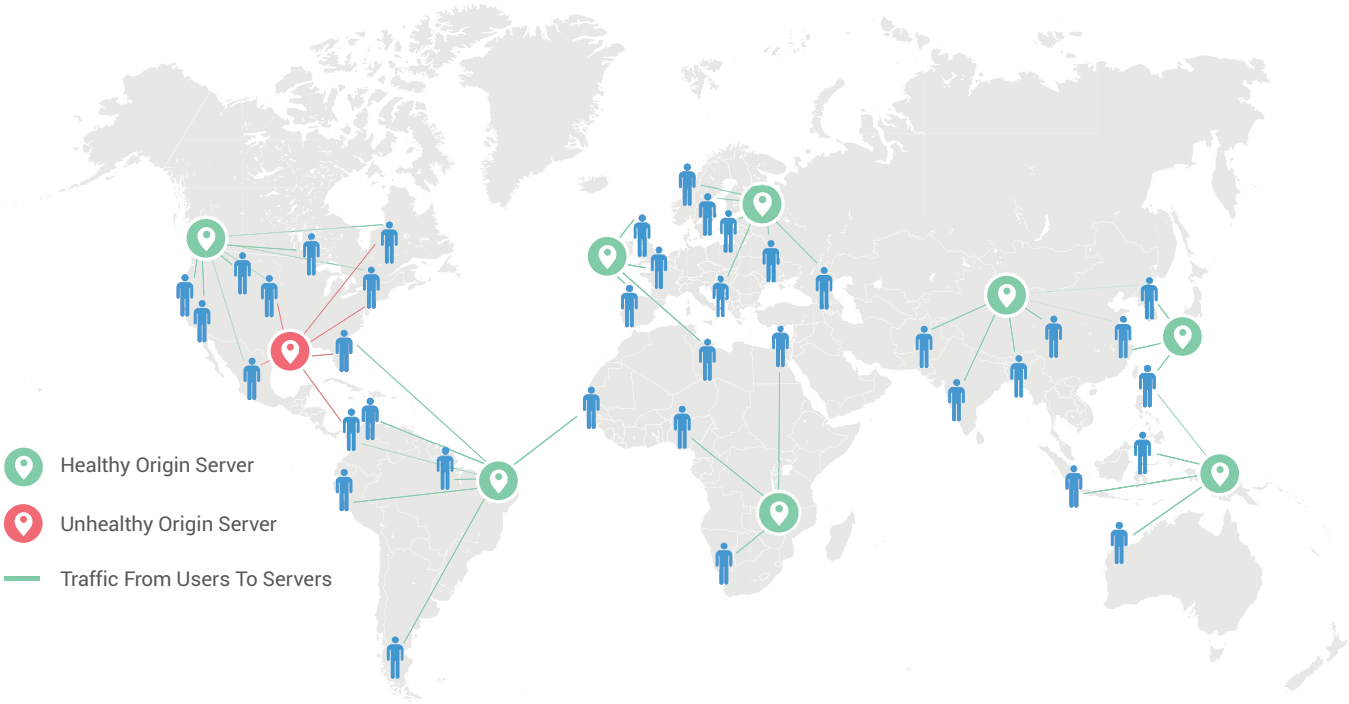
Like all computers, servers can develop problems. Monitoring the health of servers and applications is critical to reducing latency and ensuring application availability in the event of a crash. Load balancing solutions that fail to monitor server health can inadvertently route traffic to a server that is experiencing problems, resulting in users experiencing long delays and outages.

60 seconds

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



DOWNTIME CAUSED BY UNHEALTHY ORIGINS



DISTRIBUTED DENIAL OF SERVICE (DDOS) ATTACKS

DDoS attacks are a major threat to server health, and they are growing in frequency, size, and severity. Attacks sized 100Gbps and higher skyrocketed by 967% between Q1 2019 and Q1 2018, and over three-quarters of attacks targeted more than one vector.[21] Many DDoS attacks utilize "zombie armies" of hijacked IoT devices, as was the case in the Mirai botnet attacks against DNS provider Dyn in 2016.[22]

In addition to causing sluggishness and outages, DDoS attacks are sometimes used as smokescreens for other cyberattacks. Taking down a website or app is a good way to keep a company's IT team distracted during a data breach.

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

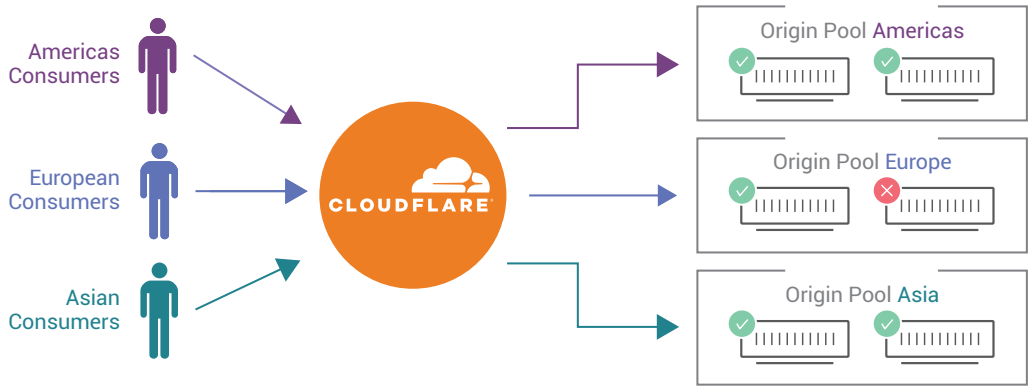
It is imperative to select a load balancing solution that not only meets your business requirements today but that can easily scale to accommodate your future needs. The solution must also be robust enough to accommodate modern traffic volume levels, app complexity, and DDoS attack size. Here's a list of features to look for.

Integration with a global content delivery network (CDN)

Load balancing and CDNs work together to solve latency and availability problems. A CDN caches static content at the network edge so that it can be delivered from a server located nearest the end user. This greatly enhances performance and minimizes bandwidth consumption by reducing the number of requests sent to the origin server.

Global geolocation-based routing

Because geographic distance between the server and the end user plays such an important role in combating latency, a modern load balancer must have the ability to connect visitors to infrastructure that is in the same part of the world. For example, U.K. traffic should be directed to a data center in London, not one in New York.



DDoS resiliency

The smaller the capacity of the global CDN, the more likely it is that a DDoS attack can bring it down. With the size of DDoS attacks growing rapidly, the CDN must have the capacity to withstand even the largest DDoS attack, with room to spare, so that the load balancer can always route traffic to healthy servers even when under stress.

Layers 3 & 4 load balancing functionality

Attackers can directly send volumetric DDoS traffic to the custom TCP and UDP communication protocols used for custom gaming protocols, remote server access (SSH), secure file transfer services (SFTP), and email (SMTP). They can also use these ports to intercept unencrypted data in transit. Defending these ports and protocols without compromising performance requires additional resources. Be sure your load balancer supports protection against layers 3 and 4 DDoS attacks, along with TLS/SSL protection to encrypt customer data.

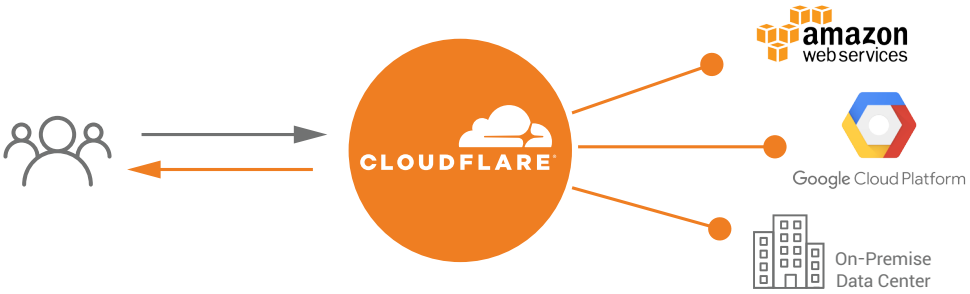
Near real-time failover

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

Multicloud & hybrid cloud support

To avoid vendor lock-in, reduce complexity, and minimize misconfigurations in multicloud 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, but rather it will work in tandem with them so that they function smoothly together.

VENDOR AGNOSTIC LOAD BALANCER



Ease of use

Time spent managing your load balancing solution is time you cannot spend managing your business. 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 as your business needs change.

Analytics

Because load balancing solutions sit between end users and applications, they are in the perfect position to gather actionable business intelligence regarding customer behavior, application performance, security posture, and other operational insights. Make sure your load balancing solution captures these analytics, and that it integrates with your existing analytics provider.

Conclusion

Modern websites and applications will not perform properly or stay online consistently 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 high-performing.

Endnotes

1. Flexera, "Cloud Computing Trends: 2019 State of the Cloud Survey," https://www.flexera.com/blog/cloud/2019/02/cloud-computing-trends-2019-state-of-the-cloud-survey/. Accessed October 10, 2019.
2. Ibid.
3. The Human Attention Span [Infographic], Digital Information World, https://www.digitalinformationworld.com/2018/09/the-human-attention-span-infographic.html. Accessed August 6, 2019.
4. "Using page speed in mobile search ranking," Google Webmaster Central Blog, https://webmasters.googleblog.com/2018/01/using-page-speed-in-mobile-search.html. Accessed August 6, 2019.
5. Rouse, Margaret. "What Is Latency?" TechTarget. https://whatis.techtarget.com/definition/latency. Accessed October 10, 2019.
6. Brutlag, Jake. "Speed Matters," Google AI Blog, https://ai.googleblog.com/2009/06/speed-matters.html. Accessed August 6, 2019.
7. Rodman, Tedd. "Marketing & Web Performance: How Site Speed Impacts Metrics," Yottaa, https://www.yottaa.com/marketing-web-performance-101-how-site-speed-impacts-your-metrics. Accessed August 6, 2019.
8. Hastreiter, Nick. "Why Slow Internet Costs Companies Money," HuffPost, https://www.huffpost.com/entry/why-slow-internet-costs-companies-money_b_5801a4b2e4b0985f6d1570e5. Accessed October 10, 2019.
9. Priceonomics Data Studio. "Quantifying the Staggering Cost of IT Outages," https://priceonomics.com/quantifying-the-staggering-cost-of-it-outages/. Accessed October 10, 2019.
10. "Cloudflare Case Study: Crisp." Cloudflare, https://www.cloudflare.com/case-studies/crisp/. Accessed July 26, 2019.
11. Kemp, Simon. "Digital 2019: Global Internet Use Accelerates," We Are Social, https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates. Accessed October 10, 2019.
12. Sherman, Fraser. "Network Latency Milliseconds Per Mile," Techwalla, https://www.techwalla.com/articles/network-latency-milliseconds-per-mile. Accessed 6 August 2019.
13. Report: State of the Web, HTTP Archive. http://beta.httparchive.org/reports/state-of-the-web#bytesTotal. Accessed August 6, 2019.
14. D'Argenio, Angelo M. "Statistically, Video Games Are Now the Most Popular and Profitable Form of Entertainment." https://www.gamecrate.com/statistically-video-games-are-now-most-popular-and-profitable-form-entertainment/20087. GameCrate. Accessed 27 August 2019.
15. Wilde, Tyler. "How game sizes got so huge, and why they'll get even bigger." https://www.pcgamer.com/how-game-sizes-got-so-huge-and-why-theyll-get-even-bigger/. PCGamer. Accessed 27 August 2019.
16. Sterling, Greg. "The mobile, desktop split may have stabilized at roughly 60% – 40%," Search Engine Land, https://searchengineland.com/mobile-desktop-search-traffic-split-may-have-stabilized-at-roughly-60-40-317091. Accessed August 6, 2019.
17. Dimensional Research. "Failing to Meet Mobile App User Expectations: A Mobile App User Survey," https://techbeacon.com/sites/default/files/gated_asset/mobile-app-user-survey-failing-meet-user-expectations.pdf. Accessed August 6, 2019.
18. Schwarz, Ben. "Beyond the Bubble: Real world performance." Calibre (Medium), https://building.calibreapp.com/beyond-the-bubble-real-world-performance-9c991dcd5342. Accessed October 10, 2019.
19. O'Donoghue, Ruadhán. "You've been throttled, but don't stop browsing!" mobiForge,https://mobiforge.com/news-comment/youve-been-throttled-dont-stop-browsing. Accessed October 10, 2019.
20. "DNS Performance Analytics and Comparison." DNSPerf, https://www.dnsperf.com/. Accessed 23 July 2019.
21. Rayome, Alison DeNisco. "Major DDoS attacks increased 967% this year," TechRepublic, https://www.techrepublic.com/article/major-ddos-attacks-increased-967-this-year/. Accessed August 6, 2019.
22. Dignan, Larry. "Dyn confirms Mirai botnet involved in distributed denial of service attack," ZD Net, https://www.zdnet.com/article/dyn-confirms-mirai-botnet-involved-in-distributed-denial-of-service-attack/. Accessed August 6, 2019.