NGINX+

The Fastest, Most Cost-Effective Web Application Platform for the Next-Generation Web

INTRODUCTION

Many users fail to think of Web speed first. They think of design; they think of content; they think of functionality. The speed at which the Web page loads is at best relegated to the IT guy to figure out and at worse is left to chance or a "call me if there's a problem" approach. However, study after study has shown that Web speed is absolutely critical to the Web experience, driving not just increased page views, but conversions, customer satisfaction and revenue.

Potential customers are trying to access your products and services online from their laptops, mobile phones, tablets, and connected devices such as Smart TVs, fitness monitors, and much more. If you're going to convert interested prospects into paying customers your Website, regardless of the amount of media or functionality it boasts, needs to load near-instantaneously.

Don't believe us? Let's look at the numbers.

Fast

↑ **Conversions**
A single second can increase conversions by 7 percent

↑ **Revenue**
When a large comparison-shopping site, dropped latency from 7 seconds to 2, revenue went up 7%-12%

↑ **Customer Satisfaction**
A single second can cause a 16% change in customer satisfaction.

Slow

↓ **Page Views**

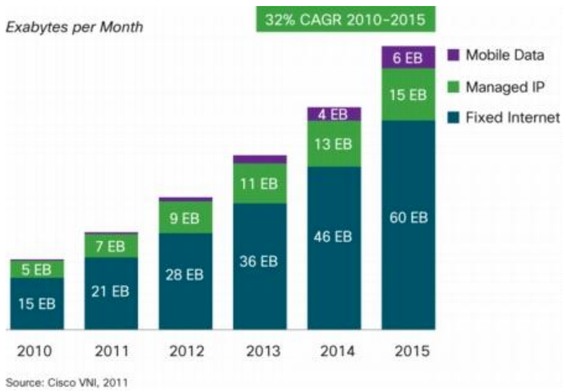
↓ **Abandonment**

↓ **Loyalty**

Speed has become so important to the customer experience on the Web that Google has made it a key factor in determining search rankings. Your website speed is not just impacting the quality of your customer's experience; it could prevent your page from being viewed at all!

UNDERSTANDING THE PROBLEM

The Web has evolved faster than anyone could have imagined in the last 5 years, let alone the last 10 or 20 years.



Source: http://www.fiercecable.com/story/cisco-forecasts-zettabyte-ip-traffic-2015/2011-05-31

Based on a study of global Internet traffic by CISCO systems, Internet traffic has grown from just under 2 Petabytes a month in 1996 to 14,000 times that number to 27,483 Petabytes a month in 2011. It's forecast to exceed 1 Zettabyte in 2016. Along with the growth in traffic has come a growth in the number of devices or what is today commonly referred to as the "Internet of things." Each of these devices, be it the latest smart phone, a five year old desktop, or a Smart TV makes requests to your servers over different levels of bandwidth with varying levels of efficiency and computing power.

All this growth in traffic has come at a time when user expectations have never been higher. The Internet has moved from a nice-to-have technical platform for accessing certain goods and services to a generally used commodity integral to daily life.

Then	Now
Desktop browsers	Desktop, mobile, embedded browsers
Static informational Websites	Dynamic Web applications
Slow, unreliable mobile browsing	Faster, more reliable connections
Internet of people	Internet of things
Thousands of Terabytes a month	Millions of Terabytes a month

THE LIMITATIONS OF TRADITIONAL WEB INFRASTRUCTURE

While the Web has exploded, innovation in the underlying Web server infrastructure has lagged behind. The most common technology in use today to serve web content has changed little since it was first used in the 1990's.

THE C10K PROBLEM

The problem was known as the 'C10K' problem when it was first identified. "How can a web platform handle 10,000 concurrent connections?"

Traditional web platforms use a model where they dedicate a thread or process to each concurrent connection. The problem arises from an imbalance between the complexity of a network connection (lightweight, requires few operating system resources) and an operating system thread or process (heavyweight, requires CPU and memory to manage).

Network connections are created by web clients – many will open several connections to your web application to retrieve the resources they need to render your web page or interact with your application. The non-linear overhead of managing the corresponding threads and processes means that even the most powerful server will struggle with more than several thousand active threads and processes competing for resources.

With 1000's of clients competing to access your website, you'll potentially have 1000's of processes competing for resources on your webserver. Many webservers wisely employ a brake – a maximum limit on the number of processes – so the server is not overloaded, but some users have to wait until others complete. Legacy infrastructure can get overwhelmed and your customers will receive unreliable performance and even service disruptions.

Simply put, the web technologies of the previous generation can't scale to the volume and capacity needed by the new generation of internet clients and services.

SCALING YOUR INFRASTRUCTURE TO MEET TODAY'S DEMANDS

There are multiple ways to address this performance problem. Each has its advantages and its challenges, and it's not uncommon to use two or more of the following techniques:

- **Content Delivery Networks** replicate common content across multiple locations. Use traffic is routed through these locations and CDNs can respond directly for content they have stored.
- **Web Content Optimization** techniques may be used to optimize the structure of your web content to reduce the load on the web infrastructure
- **Usage-based services** such as Amazon Web Services can be a cost-effective way to add more capacity when user traffic demands it
- **More efficient web application platforms** can sustain many more users on the same hardware infrastructure, and absorb significant spikes of traffic without impairing service levels

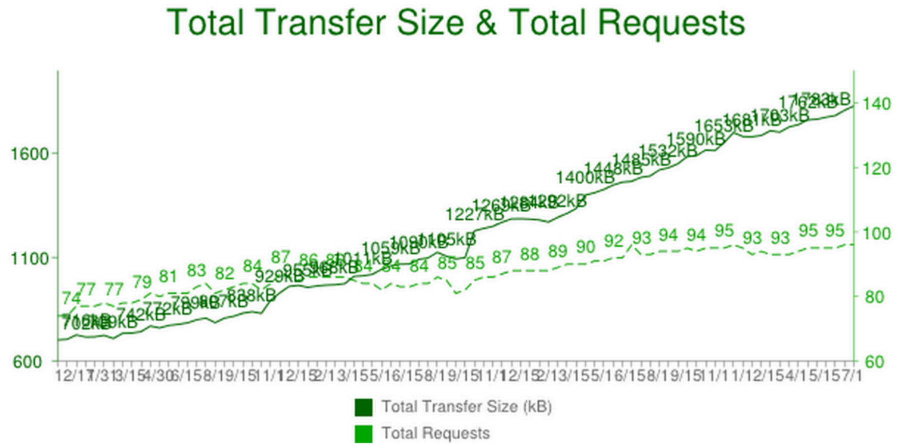
CONTENT DELIVERY NETWORKS

A **CDN (Content Delivery Network) service** uses intermediate webservers operated by the CDN provider to automatically store copies of content on your website – a process known as 'caching'. Client requests to your website are routed through an intermediate CDN server, and if the server has a copy of the resource the client has requested, it responds directly. This has the dual benefits of getting responses to the client more quickly, and taking load of your own web servers.

CDN services come at a cost. Clearly, there's the commercial cost to you of consuming CDN services. There's also the operational cost that using a CDN can impair the speed at which you update your website, and the issue that CDN providers can only reliably cache content that does not change and is shared between multiple users. If your web content is highly personalized ("dynamic"), then a CDN will be less effective than if your web content is very 'static'.

WEB CONTENT OPTIMIZATION

Web pages can be very ‘heavy’ – a large number of resources are required to render a page – and the average weight of pages is increasing year on year:



Source: http://httparchive.org/trends.php

Web Content Optimization seeks to reduce the weight of web pages (number of requests and total transfer size) using techniques such as image merging, compression, browser cache control and JavaScript optimization. It can be effective, but is complex to deploy and advanced optimizations require a lot of additional end-user testing.

USAGE-BASED SERVICES

Usage-based services (Amazon, Rackspace and other cloud providers) provide resources on demand and allow organizations to scale out in advance of large changes in traffic volume. This may be adequate when your website traffic is very seasonal, and you only require additional resources at a few predictable times, but in general it’s not a long-term way to address your capacity problems in a cost effective manner.

None of these approaches address the fundamental problem at home in your datacenter – the web platform cannot scale with the success of your business.

A BETTER WEB INFRASTRUCTURE

NGINX Plus is a different type of application platform for the modern Web. NGINX Plus was developed specifically to solve the speed limitations inherent in traditional web applications and platforms, and its core open-source technology is used by about 40% of the worlds busiest websites, including Facebook, Zynga, Netflix and DropBox. Today, hundreds of millions of websites worldwide use NGINX and NGINX Plus.

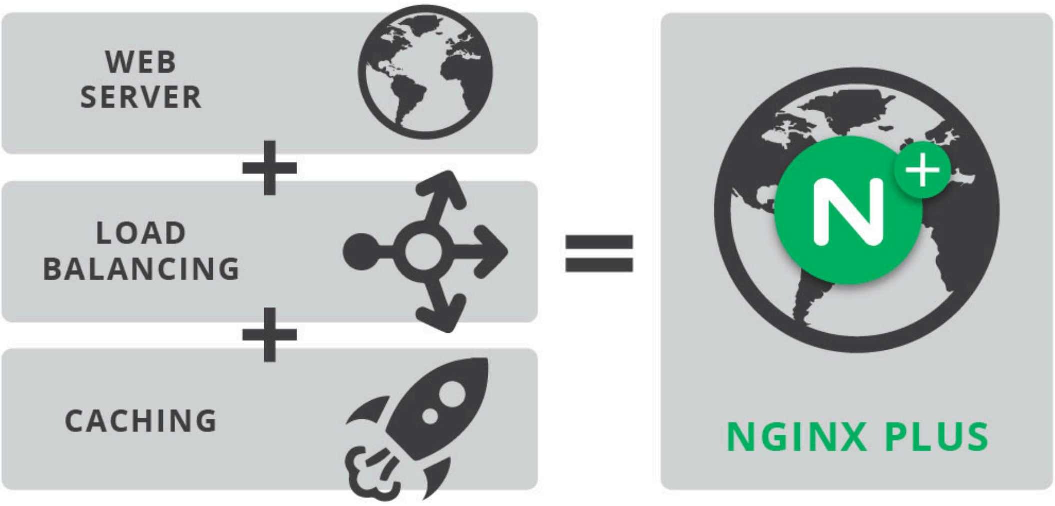
Case Study: WordPress.com

WordPress.com is the cloud version of WordPress, serving more than 33 million sites to over 409 million customers who view over 15.5 billion pages each month. Since April 2008, WordPress.com has experienced about 4.4 times growth in page views. All that growth presents a challenge: delivering high performance and reliable Web performance to every user. WordPress.com sought a solution that could scale to its needs, in a dynamic manner so that reconfiguration and infrastructure changes would not affect user experience. After evaluating many competing options, WordPress.com and hosting partner Automattic chose NGINX:

1. The only solution that could scale to support 10,000's of requests per second of live traffic from a single server
2. The ability to reconfigure and upgrade NGINX instances on-the-fly, without interrupting customer activity.
3. An efficient solution that could perform well on a low hardware footprint

With NGINX, WordPress.com is now able to scale to accommodate its runaway growth without sacrificing the quality of the customer experience even during peak times. Wordpress.com is serving an average of 70,000 requests per second and over 15 gigabytes per second of traffic from 36 NGINX powered Web acceleration and load balancer instances, with plenty of room to grow.

A NEW WEB APPLICATION PLATFORM FOR THE NEXT-GENERATION WEB



NGINX Plus combines the functionality of a web server, load balancer and web accelerating cache to create a platform that can deliver the maximum performance for you web applications.

NGINX Plus is the commercially supported version of NGINX. It extends NGINX into the domain of load balancing and application delivery, adding high-performance Load Balancing, application-aware Health Checks, advanced Content Caching, Streaming Media Delivery, monitoring and on-the-fly reconfiguration.

NGINX Plus uses an event-based architecture that doesn't suffer from the same performance limitations as traditional Web servers. This allows each NGINX Plus process to easily scale to tens of thousands or even hundreds of thousands of connections simultaneously. The result is incredible Web performance for any web application[1].

> "While there are hardware-based solutions that can do [what NGINX Plus does], they are tens of thousands, if not hundreds of thousands of dollars more expensive than NGINX Plus. We found that Nginx was the best and the easiest."
>
> *Gogo's VP of Data Centers and Infrastructure, Vinay Kudithipudi*

[1] Read more: http://www.aosabook.org/en/nginx.html

HOW DOES NGINX PLUS ACCELERATE THE WEB EXPERIENCE?

TRAFFIC OPTIMIZATION:

The NGINX Web Accelerator sits between your servers and your customers, in an architecture known as a 'Reverse Proxy'. It handles all of the user HTTP requests for web content, acting as a buffer to protect your vulnerable web servers and applications.

NGINX Plus applies a range of optimization techniques to deliver these HTTP requests to your web servers in the most efficient fashion possible; techniques such as buffering and offload, HTTP upgrades, optimized use of HTTP keepalive connections and careful management of idle keepalive connections to avoid overloading the limited resources on your webservers.

What would have been a noisy flood of traffic coming from multiple clients each competing for resources is transformed into a consistent and predictable stream of requests coming from NGINX, placing your applications into the environment where they operate the best.

CACHING:

A typical web application has to handle many repeat requests for the same content from different clients.

NGINX Plus can identify content that is 'static' (i.e. changes very infrequently) and automatically store this content in a local 'cache'. Multiple requests for the same content can be answered directly from NGINX Plus' cache, reducing the number of duplicate requests to the upstream servers. You can also optimize NGINX Plus' caching behaviour for more sophisticated use cases, caching dynamic content based on user keys or short expiry times.

Because you have direct control over the caching behavior, you can tune NGINX's cache more effectively than a generic CDN. NGINX's caches are persistent (they are not deleted when the software restarts) and NGINX employs a range of advanced techniques when verifying cached content and requesting updates in order to minimize the load on the origin servers.

EXTEND YOUR SERVER CAPACITY

NGINX can take on a host of additional tasks to allow you to support more traffic without scaling out your hardware capacity. Some of these common tasks include compressing large files for optimal delivery across the network, encrypting and decrypting content for Secure Socket Layer (SSL) communications, performing access control to secure content, and enforcing traffic limits on individual clients. NGINX is designed to process these tasks in a very scalable way, freeing up your Web and application servers and thus extending their capacity to service additional requests.

HOW IS NGINX PLUS DEPLOYED?

While NGINX Plus can act as a primary web server, most companies have already invested in a Web platform and are looking to squeeze more performance and value out of the dollars they've already spent. NGINX Plus is deployed in front of your existing Web infrastructure to load balance traffic, cache and accelerate content delivery, and speed up your web applications through a range of expert HTTP protocol optimizations. It's common for users to report increases of 10x or more in capacity from their infrastructure, with a knock-on effect of lower downtime or degradation in service in the face of traffic spikes.

Over time, organizations see further benefits by moving functionality from their legacy servers onto the NGINX layer – caching, static content delivery, interfacing with web applications using FastCGI and other protocols. This results in simplification and a decoupling of application components (moving from a tiered architecture to a star-like architecture with NGINX Plus as the central routing and application delivery component), with the associated advantages of lower latency, fine-grained scaling and less disruptive upgrades.

CONCLUSION:

The Web has fundamentally changed. Traffic has grown exponentially over the last twenty years while the number and types of devices connecting to the Internet has exploded. Yet many organizations are still running Web servers built in the nineties that simply cannot cost-effectively scale to meet the demands of the modern Web.

NGINX Plus is a modern web application platform for the next-generation Web. Using an event-driven model, and a host of advanced techniques such as traffic optimization, compression, off-loading of static content and media streaming, NGINX Plus delivers blazing speed for your Web and mobile customers. With powerful load balancing and traffic management capabilities you have the freedom to manage and update your infrastructure, while giving your customers reliable and consistent service.

Evaluation and deployment is simple. NGINX Plus does not require a rip and replace of your existing Web and application servers. NGINX Plus sits in front of your current infrastructure and helps you get the most from your existing investment.

> "Since Gogo moved from Apache to NGINX Plus, they have achieved nearly 100% uptime and have experienced no issues with NGINX Plus to date.
>
> "Today – over 90% of revenue passes through NGINX Plus and Gogo's IT organization can focus their attention on innovation and driving the business forward without worrying about downtime and the risk of lost revenue."
>
> *Inflight Internet with NGINX Plus at Gogo: http://nginx.com/products/*

Find out more about the products, support and services offered by Nginx, Inc. at nginx.com