THE MOST MISLEADING MEASURE OF RESPONSE TIME

How Optimizely Dramatically Improved Response Times with CDN Balancing
<table>
<thead>
<tr>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 INTRODUCTION</td>
</tr>
<tr>
<td>02 UNDERSTANDING CDN BALANCING</td>
</tr>
<tr>
<td>WHAT IS A CDN?</td>
</tr>
<tr>
<td>RESPONSE TIME IMPROVEMENT</td>
</tr>
<tr>
<td>RELIABILITY IMPROVEMENT</td>
</tr>
<tr>
<td>03 MEASURING WEB PERFORMANCE</td>
</tr>
<tr>
<td>THE PROBLEM WITH AVERAGE</td>
</tr>
<tr>
<td>THE PERCENTILE-BASED APPROACH</td>
</tr>
<tr>
<td>THE 99TH PERCENTILE</td>
</tr>
<tr>
<td>04 IMPACT OF CDN BALANCING FOR OPTIMIZELY</td>
</tr>
<tr>
<td>BREAKOUT BY CONTINENT</td>
</tr>
<tr>
<td>05 CONCLUSION</td>
</tr>
<tr>
<td>ABOUT OPTIMIZELY</td>
</tr>
</tbody>
</table>
Website load times have been an important topic of conversation since the early days of the Internet. However in recent years, response time and server latency have become especially hot topics among web developers and business leaders alike. Web visitors have an ever shrinking tolerance for long load times. According to Compuware’s 2013 Application Performance Trends Report, “a one-second delay in website response time results in 11% fewer page views, a 16% decrease in customer satisfaction and a 7% loss in conversions.”
Bounce rate also increases significantly with each incremental second that a page takes to load. For businesses in industries like e-commerce and media, this could mean a real loss of revenue opportunities. Compuware indicates that “a one second increase in Amazon.com’s page load would annually cost them $1.6 billion in sales.”

This paper provides an overview of how Optimizely has improved the response time of its experiment code by implementing a balanced CDN architecture. This paper explores the various ways businesses can measure response times for the client-side, third party code they host on their websites. It also explores why using the average response time paints an unclear picture of the amount of time visitors actually experience when loading a page. Finally, this paper shows how Optimizely improved the 99th percentile of response times of its variation code by 42% globally and by more than 160% in North America.
02
UNDERSTANDING CDN BALANCING

WHAT IS A CDN?
Due to the ubiquity of the Internet, users can visit websites from anywhere in the world, and they are doing so with increasing frequency as global markets become even more penetrated with web access. While this is generally a good thing for website owners, there is a risk that response times to their page could suffer. The reason is that response times can be impacted both by (a) the number of visitors making a request from a web server at the same time and (b) the physical distance of that visitor from the web server. This is why it’s become increasingly important for website owners...
to use CDNs to host their content. A CDN, or Content Delivery Network, is a network of web servers located around the world. Each of which hosts a version of the assets that make up a website (e.g. images, styles, JavaScript, etc.). This architecture reduces the load on any individual server and decreases the amount of physical distance between visitors and the web server.

At the highest level, a “balanced” CDN architecture is one that leverages two or more CDNs hosting identical content to increase (a) the overall number of physical Points of Presence for the network (PoPs) and (b) the proximity of those PoPs to the end users accessing them around the world.

**RESPONSE TIME IMPROVEMENT**

As an example, imagine “CDN A” has PoPs in the following locations around the world:

Then, assume there is another CDN, “CDN B,” with PoPs in these following locations:

As you can see, CDN A has significant coverage in North America and Europe, but only one server in Asia-Pacific. On the other hand, CDN B has no coverage in Europe but very heavy coverage in North America and Asia-Pacific.

The vast majority of a CDN’s performance depends upon the proximity of a PoP to the end user accessing the content. If a business has a website that receives a significant deal of traffic from all three of these geographies, this business cannot choose only one of CDNs A or B without accepting that one of their markets will suffer a major decline in performance.
In this situation, a balanced CDN is the optimal solution. It allows the business to combine the PoPs from CDN A and CDN B into a larger combined network:

If a visitor from Asia-Pacific is requesting assets from a page, the system will likely switch them to the nearest PoP in CDN B. If a visitor is coming from North America, the system will determine the nearest PoP on either CDN A or CDN B, and funnel them to that one so they experience the fastest response time.

RELIABILITY IMPROVEMENT
At its core, a balanced CDN is also a redundant system with each CDN representing an independent component of the system. As is the case in other areas of engineering, the likelihood of an overall CDN failure decreases each time a redundant CDN is added.

This formula represents how one should calculate the probability of system failure:

$$p = \prod_{i=m}^{n} p_i$$

Where:
- $n =$ number of CDNs
- $p_i =$ probability of CDN $i$ failing
- $p =$ probability of all CDNs failing (i.e. a total system failure)

If we assume that both CDN A and CDN B have a 0.1% probability of failure (i.e. 99.9% uptime), the two systems combined would have a probability of failure equal to 0.01%, or 99.99% uptime.

Depending on what assets are stored on the CDN, a failure can either (a) cause the page to look jumbled or broken, or (b) produce a response error and prevent the page from loading entirely. In either case, a failure is an expensive risk few businesses should want to take. It not only causes site abandonment in the short term, but can also leave a lasting impact on the visitors’ perception of and confidence in the company’s focus on stability and quality.
Anyone who has spent time analyzing website performance has likely looked at the average, or mean, response time for a page to summarize the page’s overall performance. There are many situations where this is an appropriate metric, but when it comes to analyzing website response time, the uneven way results are typically distributed make it a highly misleading metric.
THE PROBLEM WITH AVERAGE

As an example, let us assume web response times were generally distributed in a normal, bell-shaped curve that looks like this:

The change in the average is equal to the change across the entire distribution.

Unfortunately actual response times are determined by a number of factors like visitors’ bandwidth and their proximity to a CDN server. As a result, there are often significant outliers in the distribution. An actual distribution may look something like this:

In this example, average response time decreased by 17% but despite that, some visitors actually experienced an increase in response times. If the team at this fictional company were only to look at average response time, they might go on not noticing that there are still a segment of visitors receiving a degraded experience and likely abandoning at a higher-than-normal rate.

THE PERCENTILE-BASED APPROACH

This is why it’s important to look at the entire distribution of results rather than one specific metric. Percentiles tend to be the most efficient way to describe a distribution, so a good place to start is to pick key percentiles and look for change at those levels.
For example if we look at the 75th percentile using the data above, you can see that there was a 23% improvement in response time.

The 75th percentile provides insight into how the long tail of the distribution curve has changed. It says that 75% of the transactions are as fast or faster than that amount. But it still does not tell us the entire story. Looking at the 90th and 99th percentiles expose the problem with the long tail of visitors:

The new distribution is not better for the vast majority of visitors in the long tail. In fact, it actually increased response times at the 90th percentile of visitors by 17%, and there was no change for visitors at the 99th percentile. This indicates the change was actually not successful in improving the site experience for all visitors, despite the fact that response times actually did decline for the average and median set of visitors.

THE 99TH PERCENTILE

If one only metric does need to be reported, then the 99th percentile may be the most telling because the raw number tells us that 99% of visitors who come to the site experience that speed or better. Further, a change that reduces the 99th percentile indicates there was a change that actually moved the entire distribution of response times closer to zero. The 99th percentile is a critical metric when trying to assess overall impact of a change because it explains that a website’s slowest response times are now faster.

In practice, this will tend to indicate a faster response time across the board for all visitors, but it is still important to look at key percentiles along the distribution curve wherever possible in case speed improvements in the long tail are coming at the expense of visitors in the 50th percentile, for example.
Prior to introducing CDN Balancing, Optimizely’s experiment code was stored in a JavaScript file that sits on Akamai’s CDN, one of the fastest and most reliable CDNs in the world. The code in this file is important because it’s responsible for making client-side changes to a page and serving them to visitors according to the targeting and A/B testing criteria set up by the Optimizely customer who created the experiment.
Akamai's CDN serves more than 20% of web traffic worldwide, so it has been built for scale and reliability. This chart illustrates a distribution of response times for Optimizely’s experiment code over the course of 30 days:

The chart below shows the distribution of response times using the Balanced CDN architecture for the same Optimizely experiment snippet as above over the same period of time:

Due to the importance of response time and the value of CDN balancing illustrated in previous sections, Optimizely introduced CDN Balancing using the second most widely used CDN on the web, EdgeCast. With the global delivery networks of Akamai and EdgeCast combined, the results were significant performance improvements across the board.

Finally, here are the two distributions combined to show the relevant impact:
Overall, CDN Balancing resulted in a dramatic improvement in response times for visitors at every point along the distribution curve. Here are a few key percentile statistics to illustrate just how significantly the long tail of results was reduced with this change:

<table>
<thead>
<tr>
<th>PERCENTILE</th>
<th>% CHANGE*</th>
</tr>
</thead>
<tbody>
<tr>
<td>25th Percentile</td>
<td>-268%</td>
</tr>
<tr>
<td>50th Percentile (Median)</td>
<td>-133%</td>
</tr>
<tr>
<td>75th Percentile</td>
<td>-58%</td>
</tr>
<tr>
<td>85th Percentile</td>
<td>-42%</td>
</tr>
<tr>
<td>95th Percentile</td>
<td>-39%</td>
</tr>
<tr>
<td>99th Percentile</td>
<td>-42%</td>
</tr>
</tbody>
</table>

* LOWER NUMBERS ARE BETTER.

### BREAKOUT BY CONTINENT

The volume of web activity worldwide is increasing rapidly, and the same is true for the traffic coming to sites owned by Optimizely customers. That’s why it is also important to look at the geographic breakdown of web visitors.

The chart below shows the difference in response times at the 99th percentile with and without CDN Balancing across key continents alongside the overall (worldwide) number:

The 99th percentile of response times improved in all continents. The most dramatic improvement was in North America where there was a 161% (more than 1.5x) improvement in response times. This is significant because the vast majority of web traffic still comes from North America, so this change illustrates that having a balanced CDN is very much worthwhile. The following charts illustrate the change in the overall distribution curves for these key markets:
For many website owners, it can be difficult to weigh the value of adding third party code to a page against the relative impact of adding additional milliseconds to the page's response time.

Prior to implementing a balanced CDN architecture, Optimizely’s snippet already responded in a virtual blink of the eye. Now, Optimizely’s CDN Balancing strategy allows experiment code to be served by a combination of the two fastest and most widely used CDN providers in the world.

For customers, this means that Optimizely’s experiment code integrates into their web pages even more seamlessly than before. Delivering such fast response times complements Optimizely’s mission to empower the world’s businesses and organizations to turn data into action.
Optimizely is the world’s leading web optimization platform, providing A/B testing, multivariate testing, and website personalization for individuals and organizations with and without technical expertise. The platform’s ease of use empowers organizations to conceive of and run experiments that help them make better data-driven decisions. With targeting and segmentation using powerful real-time data, Optimizely meets the diverse needs of any business looking to deliver unique web experiences to their visitors.

Email: hello@optimizely.com
Phone: 1-800-252-9480
Facebook: facebook.com/optimizely
Twitter: twitter.com/optimizely

San Francisco Office
631 Howard Street, Suite 100
San Francisco, CA 94105

Amsterdam Office
Egelantiersstraat 143
1015 RA Amsterdam

To test out Optimizely or request a demo, visit OPTIMIZELY.COM/CDN