**Web Performance Optimization**

Focus

Classic Thinking

* Performance has to with the server.
* Adding more CPU, memory and more bandwidth.
* More database performance solutions (optimize database queries).

Current

* Most of the page load time happens on the frontend side (84% based on Alexa data).
* The backend side Is usually the first to load and usually takes the least time.
* Optimizing the backend is not simple and is expensive (more servers, more CPU, buy load balancers). 30 % optimization of the backend would mean 5 % increase.
* Optimizing the frontend is simple and cheap and we can test and measure the different techniques. 30% improvement would impact 25% on the web performance. This would mean that the focus of optimization would mean focusing on network optimization.

## Performance golden rule

**“80 % of end-user response time on average is spent on the fronted leaving 20% on the backend”.**

WPO Thinking

* Frontend decisions makes the difference.
* Simpler and easier to measure solutions.
* Focus is on frontend tuning.
* For enterprise WPO backend optimization is applicable as well.

### Performance Goals

Load Times Vs Perception

* Human perception is more important than the actual load time being fast.
* Human milliseconds for an immediate feedback is 100 ms ,1 seconds is the limit of user flow of thoughts and limit to keep the users’ attention on the task is 10 seconds.

**RAIL Performance Goals**

This was created by google.

**R**- Response (This must appear within 100ms)

1. Animation (This must take only 10ms) e.g. lists

**I**- Idle (This should not take more than 50ms) Tasks in code that are executed without the users’ reaction

**L**-Load (Page load should take only one second) When users are typing the URLs

**Mobile Web**

Users opening the websites from their phones

**Problems**

* Not everyone is using Safari or Chrome
* Not everyone is on 4G
* Even on 4G the latency is high and one can be downgraded
* VPN latency is high as well

**Android Data**

**Websites visits**

* 62% visits from the browser
* 34% visits from the Facebook
* 4% others

100% use the browsers 66% of the users are using chrome

Raises the need to test the performance on different browsers.

Bandwidth(Speed) might not matter on the performance of a website

* Latency refers to the time it takes a packet to reach its destination (4G latency is 300ms)

### Performance Metrics

1. **Waterfall Charts**

Diagram to see how the resources are being downloaded and parsed by the engine in a timeline that lets you see the sequence and dependencies between resources. Depends on the browsers and is created by the browser developer tools.

X axis has the time in milliseconds

Y axis has the resources by order (rows are the resources)

**First view:** Empty Cache

**Repeat view:** Full Cache for the page – Less resources are needed to load the page

**Flexible view:** Some resources cached such as login page

**Elements**

1. Vertical line represents the milestones – e.g. page load
2. Row colors
   1. Yellow rows – means the resource was wreck
   2. Red rows – means there was an error
3. Resource Timing Bar
   1. DNS Lookup – converts the human friendly domain into an IP address
   2. TCP Connection/Handshakes
   3. Time to First Byte (TTFB) – Time the server takes to send the first byte(Backend)
   4. Download

**Waterfall Chart in Action**

**Tool -** <https://www.webpagetest.org/>

Analysis Example - [https://www.webpagetest.org/result/201106\_DiZW\_259d601e3c3043fcca336c5bf48de4b4/1/details/#waterfall\_view\_step1](file:///C:\Users\Spectre\AppData\Roaming\Microsoft\Word\auto.wazinsure.com)

1. **The HAR format**

This refers to the HTTP Archive Format. The format that can be used by a web browser to export detailed performance data about web pages as it loads. One can save in a file the waterfall chart.

Har files can be created on the developer tools of different browsers.

**Har File Example**



**Tool** - <https://httparchive.org/>

**Tool to analyze HAR file –** <https://toolbox.googleapps.com/>

- <http://www.softwareishard.com/har/viewer/>

- <http://onlinecurl.com/har-diff/new>

### Performance Milestones

Meaningful moments in the timeline and it depends on the tools used. Some tools render them in the waterfall chart as vertical lines in the timelines.

* TTFB (Time to First Byte) – When the server answers with the first byte of response and is useful to know when the server is taking too much time.
* First Render/First Render – User sees something other than a blank space or the previous page after CSS has loaded
* First Meaningful Paint – User can now see and read text
* First Interactive – User starts interacting with the app (Custom metric and defined in the codebase)
* Consistently Interactive
* ATF (After Fold Render) Render – Content seen on the screen without scrolling
* Page Load – Initial resources have been loaded

### Performance importance Indicator

This refers to selling the performance of webpage. This can be done using the **Filmstrip view**. Show the frames of the page every 100ms. Then generate a video with the same. Can be found on the google chrome developer tools.

**Speed Index -** How much blank content the user has seen during the loading process. Render as much as possible as soon as possible. The goal is to have a speed index of 1500

**Perpetual Speed Index** – Indicates how quickly visible parts are populated

## Measuring Performance