

Bachelor-Thesis

Web Performance für den mobilen Endanwender

Zusammenfassung

1. April 2015

Columbus Interactive
Eywiesenstraße 6
88212 Ravensburg

Fakultät Elektrotechnik und Informatik
Studiengang Angewandte Informatik
Hochschule Ravensburg-Weingarten
Doggenriedstraße, 88250 Weingarten

Autor:

Andreas Lorer
contact@andreaslorer.de
Wilhelmstraße 4
88250 Weingarten



Bachelor-Thesis

Web Performance für den mobilen
Endanwender

Inhaltsverzeichnis

1 Einleitung	3
1.1 Motivation	3
1.2 Zielsetzung	4
1.3 Eigene Leistung	4
1.4 Ist-Zustand	5
2 Begriffe	5
2.1 Latenz	5
2.2 Round Trip Time (RTT)	5
2.3 Http/1.1	5
2.4 TCP Three Way Handshake	5
2.5 TCP Slow Start	5
2.6 Content Delivery Network (CDN)	7
2.7 Above The Fold	8
2.8 Perceived Performance	9
3 Die 1000 ms Barriere	10
3.1 Touch Event	10
3.2 Netzwerke	11
3.2.1 Mobilfunknetz	11
3.3 Der HTTP-Request	12
3.4 Das Herunterladen einer 40 Kilobyte Datei	13
3.5 Zusammengefasst	14
3.6 Kritischer Rendering-Pfad	16
3.6.1 Rendering	16
3.6.2 Rendering-Pfad	17
3.6.3 Critical render path	18
3.6.4 Zusammengefasst	18
3.7 Analyse des Wasserfalls	19
4 Entwicklung	22
4.1 Tools	22
4.1.1 Google Chrome Developer Tool	22
4.1.2 Google Pagespeed Insight	22
4.1.3 Google Closure Compiler	22
4.1.4 Webpagetest	23
4.1.5 Pingdom	25
4.1.6 Speedcurve	25
4.1.7 Google Spreadsheet	25
4.1.8 Feed the Bot	25
4.1.9 What Does My Site Cost?	25
4.1.10 Critical Path CSS Generator	26
4.1.11 Http Archive	27
4.1.12 Perf Tooling Today	27
4.1.13 Twitter	27
4.2 Ausgangspunkt	28
4.3 Best Practices	29
4.3.1 Render Blocking Javascript	29

4.3.2	Render Blocking CSS	32
4.3.3	Inline CSS	32
4.3.4	Ressourcen reduzieren	32
4.3.5	CSS-Bereitstellung optimieren	33
4.3.6	Antwortzeit des Servers reduzieren	33
4.3.7	Browser-Caching nutzen	33
4.3.8	Komprimierung aktivieren	36
4.3.9	„Keep-alive“ ermöglichen	38
4.3.10	HTML5 Link Prefetching	38
4.4	Bilder optimieren	39
4.4.1	Progressive Image Rendering	39
4.4.2	Image Spriting	40
4.4.3	Bild Komprimierung	40
4.4.4	Responsive Images	41
4.4.5	Adaptive Images	42
4.4.6	Verzögertes Laden von Bildern	43
4.5	Zusammengefasst	43
5	Workflow	43
5.1	Nodejs	43
5.2	Node Package Manager	43
5.2.1	Dependency Management	44
5.2.2	Bower	44
5.3	Gulp Task Manager	45
5.4	Yeoman	47
5.4.1	Eigene Generatoren erstellen	48
5.5	Zusammengefasst	49
6	Ergebnis	51
6.1	Wie wurde getestet?	51
6.2	Datenauswertung	51
7	Der Weg zur Performance	55
7.1	Hürden	55
7.1.1	Projekt Manager	56
7.1.2	Aesthetic heavy designers	57
7.2	Performance Budget	58
7.2.1	Budget Metriken	59
7.2.2	Wie schnell, ist schnell genug?	59
7.2.3	Arbeiten mit einem Performance Budget	60
8	Ausblick	62
8.1	Http/2.0	62
8.1.1	Http/1.1 Optimierungen in Http/2	63
9	Fazit	64
10	Anhang	65
10.1	Webpagetest Teststandorte	65
10.2	Argumentations Sammlung	66

1 Einleitung

1.1 Motivation

Niemand mag es zu warten. Sei es auf Bus, Bahn oder an der Kasse im Supermarkt. Wir warten auch nicht gerne im Internet auf das Buffern eines Videos, beim Besuch einer Webseite oder beim Shoppen via APP. Wie oft wurde aus dem „nur mal eben diesen Begriff nachschlagen“ ein endloses Starren auf den weißen Bildschirm? Zu oft! Jeder kennt das.

Larry Page, CEO und Mitgründer von Google, sagt:

„As a product manager you should know that speed is the number one feature.“ (Holzle 2010)

Niemand mag es zu warten, auch nicht auf eine Webanwendung. Die Studie „The Psychology of Web Performance“ kam schon bereits im Jahr 2008 auf folgende Ergebnisse:

„Slow web pages lower perceived credibility and quality. Keep your page load times below tolerable attention thresholds, and users will experience less frustration, lower blood pressure, deeper flow states, higher conversion rates, and lower bailout rates. Faster websites are actually perceived to be more interesting and attractive.“ (WebSiteOptimization.com 2008)

Das Hauptvermarktungsargument für den Chrome Browser war damals: er sei schneller als die Konkurrenz. Tatsächlich ist für Google Geschwindigkeit alles. Deshalb hat Google im Jahr 2010 angekündigt, dass Geschwindigkeit in die Berechnung des Google Page Rankings mit einfließt.

„Faster sites create happy users and we've seen in our internal studies that when a site responds slowly, visitors spend less time there. [...] Recent data shows that improving site speed also reduces operating costs. Like us, our users place a lot of value in speed — that's why we've decided to take site speed into account in our search rankings“ (Google 2010)

Aktuell (2015) geht Google sogar noch einen Schritt weiter und informiert tausende Webmaster per E-Mail über die schlechte Usability ihrer Websites für mobile Besucher und warnt ausdrücklich vor dementsprechend „angepassten Rankings“. (t3n 2015) Im Hinblick auf die Zukunft wird der Marktanteil an mobilen Internetnutzern noch weiter wachsen und die Optimierung der Ladezeiten gewinnt dadurch noch mehr an Bedeutung. Zwischen 2011 und 2014 stieg die Anzahl der Smartphonenuutzer von 18% auf 50% an. Dies ist ein Wachstum von 32% innerhalb von nur 3 Jahren. (TNS Infratest 2014)

Die Antwort auf diesen Trend läutete eine Ära ein, die wir heute unter dem Namen **Responsive Webdesign** kennen. „Responsive“ muss aber sehr viel mehr bedeuten, als nur eine angepasste Darstellung für eine bestimmte Art von Gerät. „Two out of three mobile shoppers expect pages to load in 4 seconds or less.“ (Radware 2013). Der Anwender erwartet also auf dem Smartphone ähnliche oder gleiche Ladezeiten wie er auch von der Nutzung eines Desktop-Pc's gewohnt ist. Diese Erwartungen werden von dem Großteil der im Internet besuchbaren Seiten nicht erfüllt. Der Inhalt einer Seite muss darum so aufbereitet werden, dass dieser auch auf Geräten mit langsamer Internetverbindung, hoher Latenz und einem begrenzten Datentarif, in einer für den Anwender annehmbaren Geschwindigkeit, angezeigt werden kann.

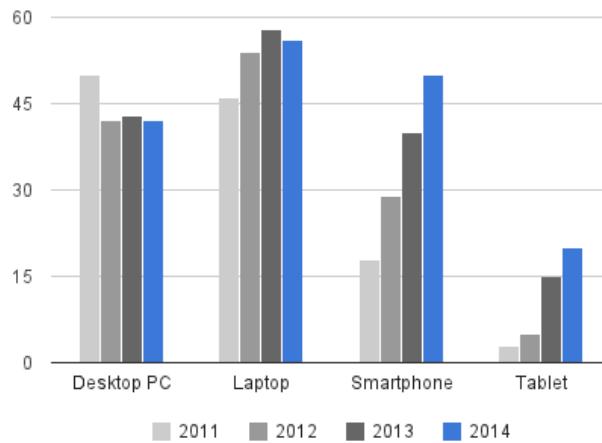


Abbildung 1: Gerätenutzung in der Gesamtbevölkerung (2011 – 2014)(TNS Infratest 2014)

1.2 Zielsetzung

Um gängige Methoden und Techniken der Ladezeit-Optimierung anzuwenden wird das Projekt anhand der Website <http://andreaslorer.de> durchgeführt. Das Ziel ist es, die Ladezeit der Website auf dem Smartphone, wie auch auf dem Desktop von 10 Sekunden auf unter 1 Sekunden zu verringern. Mit Ladezeit ist dabei nicht die Zeit gemeint, die benötigt wird um die Website komplett zu laden, sondern die Zeit bis ein erste visuelle Rückmeldung für den Anwender zu sehen ist. Diese vom Anwender wahrgenommene Rückmeldung nennt man auch „Perceived Performance“ und bedeutet, dass die Ladezeit als schneller empfunden wird, als es eigentlich laut Messwerten der Fall ist. Näheres dazu wird in Punkt 2.8 beschrieben.

1.3 Eigene Leistung

Meine Leistung besteht darin, einen Leitfaden zu erstellen, der einen Gesamtüberblick ermöglicht. Die Arbeit soll es dem Leser ermöglichen Fehler in der Struktur von Webanwendungen zu finden, die für die Geschwindigkeit hinderlich sind. Es soll herausgefunden werden, was die „Best Practices“ sind um die Ladezeit zu minimieren, wie ein moderner „Workflow“ aussehen kann, damit eine Webanwendung schon bei seiner Entstehung schnell ladet und im Projektverlauf schnell bleibt. Des weiteren soll erklärt werden, was für Herausforderungen es zu meistern gilt um eine schnelle Webanwendung zu erreichen, welche Tools es gibt und welche Vor- oder Nachteile diese mit sich bringen.

Diese Arbeit befasst sich nicht, mit der Geschwindigkeit von Datenbanken, SQL-Abfragen oder sonstigen Problemen die durch solch einen Engpass ein schnelles Laden der Seite verhindern könnten.

1.4 Ist-Zustand

Die Webseite ist auf einem **shared Hosting**¹ aufgesetzt und antwortet auf ein Ping Kommando in rund 13ms. Dadurch, dass es keine Möglichkeit gibt **root Rechte**² auf einem shared hosting zu bekommen, können so manche serverseitige Einstellungen nicht durchgeführt werden. Diese werden dann zwar Aufgezeigt, aber kommen für dieses Projekt nicht zum Einsatz.

Die Website hat als Ausgangsbasis einen gängigen Aufbau. Sie besteht aus einer Bilder Gallerie basierend auf PHP und dem Bootstrap Framework.

2 Begriffe

2.1 Latenz

Latenz bezeichnet die Verzögerung, bis ein Paket von Sender A zu Empfänger B gelangt ist.

2.2 Round Trip Time (RTT)

„Round Trip Time“ wird im Deutschen Paketumlaufzeit genannt. Es bezeichnet die Zeit die ein Datenpaket braucht um in einem Netzwerk von Sender A zu Empfänger B und wieder zurück zu gelangen. Bei einer Latenz von 100ms würde die RTT folglich 200ms betragen (Annahme: Hin- und Rückweg haben die selbe Zeit).

2.3 Http/1.1

Der für diese Arbeit wichtige Aspekt der HTTP/1.1 Spezifikation ist die Limitierung von Verbindungen pro Domain Name. Dabei weicht die Limitierung zwischen den Browsern ab und reicht von 6 (Google Chrome) bis 13 (Internet Explorer 11) parallelen TCP Verbindungen. Eine Übersichtsliste ist hier zu finden <http://www.browserscope.org/?category=network>

2.4 TCP Three Way Handshake

TCP ist das meistgenutzte Verbindungsprotokoll im Internet. Auf diesem Protokoll wird der HTTP Request aufgebaut, der die eigentlichen Daten enthält. Bevor Daten zwischen Server und Browser ausgetauscht werden können, muss eine Verbindung aufgebaut werden. Abbildung 2 beschreibt den Prozess des Verbindungsauftausbs.³

2.5 TCP Slow Start

Ein Round Trip kann nicht beliebig viele Bytes transportieren sondern ist durch die sogenannte „Congestion Window Size“⁴ limitiert. Der Überbegriff für dieses Verhalten nennt sich „Slow Start“

- Congestion Control: Nach dem eine neue Verbindung per TCP aufgebaut wurde, können weder Server noch Client wissen, wie schnell die Verfügbare Bandbreite ist, mit der Daten

¹Bei shared Hosting werden mehrere Websites von verschiedenen Website-Betreibern von dem gleichen Webserver gehostet. Bei Shared Hosting teilen sich in der Regel Hunderte andere Websites einen Server (ItWissen.info 2015)

²Standardmäßig existiert unter Linux immer ein Konto für den Benutzer „root“ mit der User-ID 0. Dies ist ein Systemaccount mit vollem Zugriff auf das gesamte System, und damit auch auf alle Dateien und Einstellungen aller Benutzer (wiki.ubuntuusers 2014)

³Für ein tieferes Verständnis empfiehlt sich dieser Artikel: High Performance Browser Networking - Chapter 2: Building Blocks of TCP

⁴engl. congestion: Stauung, Überlastung, Anhäufung

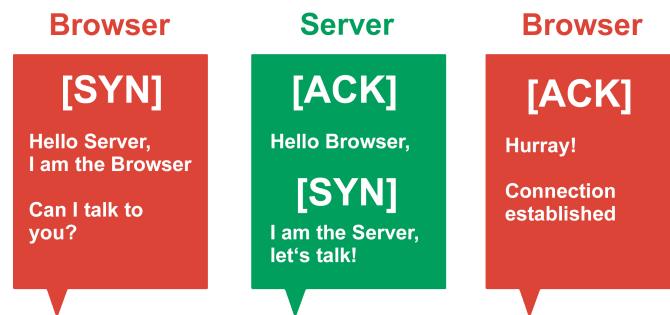


Abbildung 2: Three-Way-Handshake zum Aufbau einer TCP Verbindung zwischen Browser und Server (Eigene Abbildung nach: (Stefanov 2011))

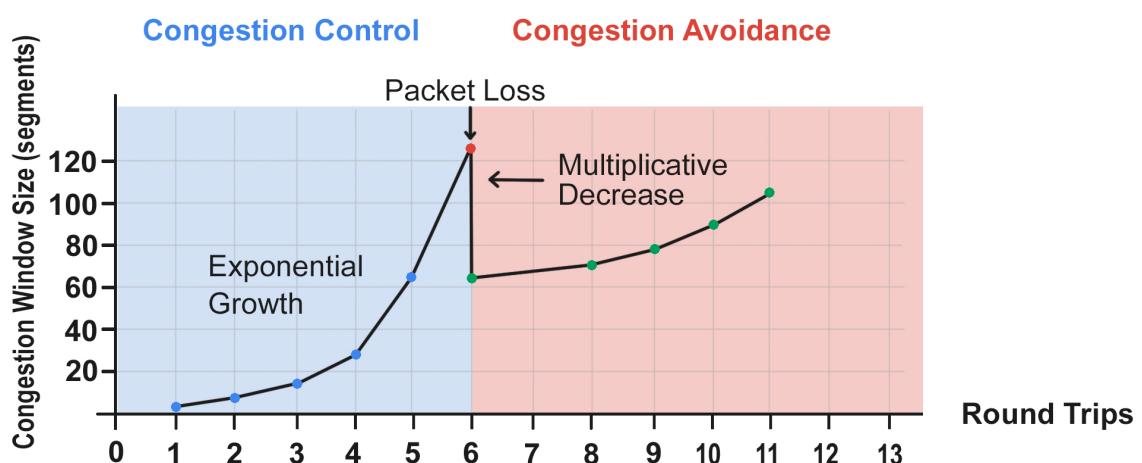


Abbildung 3: Congestion Control und Congestion Avoidance (Eigene Abbildung nach (Grigorik 2013c))

ausgetauscht werden können. Um das Netzwerk, vor einem Datenstau zu schützen, wird mit einem sehr niedrigen Wert begonnen, der dann ansteigt bis das Limit erreicht ist. Dieses Verhalten nennt sich auch „Congestion Control“ und verhindert das Aufstauen von Daten.

- Congestion Window Size: Diese Größe bestimmt, wieviel Bytes der pro Segmente geschickt werden darf, bis diese vom Empfänger per ACK (acknowledgement) bestätigt werden müssen. Die Größe der Segmente ist Standardmäßig 1460 bytes und die Rate bis zum ACK ist im April 2013 von 4 auf 10 Segmente erhöht worden.(Grigorik 2013c). In der Grafik wird davon ausgegangen, dass der erste Round Trip 4 Segmente senden darf. Die Datenrate Wächst exponentiell an, damit möglichst schnell die volle Bandbreite nutzbar ist.
- Congestion Avoidance bedeutet, dass sich die Datenrate wieder um ein Vielfaches verringert, falls es zu einem Paketverlust kommt. Da es besonders bei WLAN oder Mobilfunknetzen des öfteren zu Packetverlusten kommen kann ist dieser Aspekt besonders hervorzuheben, denn er verzögert das erreichen der maximal möglichen Datenrate.

Slow Start bedeutet also aus sicht der Performance, dass bei einer neuen TCP Verbindung nicht die maximale Bandbreite zu Verfügung steht. Bei größeren Dateien wird zwar durch das exponentielle Wachstum das Maximum schnell erreicht, gerade aber bei kleineren Dateien mit wenigen Kilobyte ist dies oft nicht der Fall.

2.6 Content Delivery Network (CDN)

Ein Content Delivery Network (CDN), oder auch Content Distribution Network genannt, ist ein Netz lokal verteilter und über das Internet verbundener Server, mit dem Inhalte ausgeliefert werden. CDN-Knoten sind auf viele Orte verteilt um Anfragen (Requests) von End-Nutzern nach Inhalten (Content) möglichst ökonomisch zu bedienen. Große CDNs unterhalten tausende Knoten mit zehntausenden Servern.(wikipedia 2015a)



Abbildung 4: Schematische Darstellung eines CDN (Eigene Abbildung nach (Ritz 2014))

2.7 Above The Fold

Damit ist der auf einem Bildschirm sichtbare Bereich vor dem Scrollen gemeint. Diesem Bereich wird eine besondere Wichtigkeit zugesprochen.



Abbildung 5: Darstellung des sichtbaren Bereichs vor dem Scrollen

„In an analysis of 57,453 eyetracking fixations, we found that there was a dramatic drop-off in user attention at the position of the page fold. Elements above the fold were seen more than elements below the fold: the 100 pixels just above the fold were viewed 102% more than the 100 pixels just below the fold.“ (Schade 2015)

Wichtige Informationen oder Navigationselemente sind meistens dort zu finden. Eine Webseite die nach dem Paradigma des Responsive-Webdesign aufgebaut ist kann dabei 3 oder mehrere Ansichten haben die alle einen unterschiedlichen „above the fold“ bereich haben. Eine Anwendung kann aber auch unterschiedliche Seiten haben, auf dem der Anwender beim Aufrufen der Seite landen kann. Zum Beispiel wenn dieser an- oder abgemeldet ist. Paradebeispiel dafür sind Facebook oder Twitter.

2.8 Perceived Performance

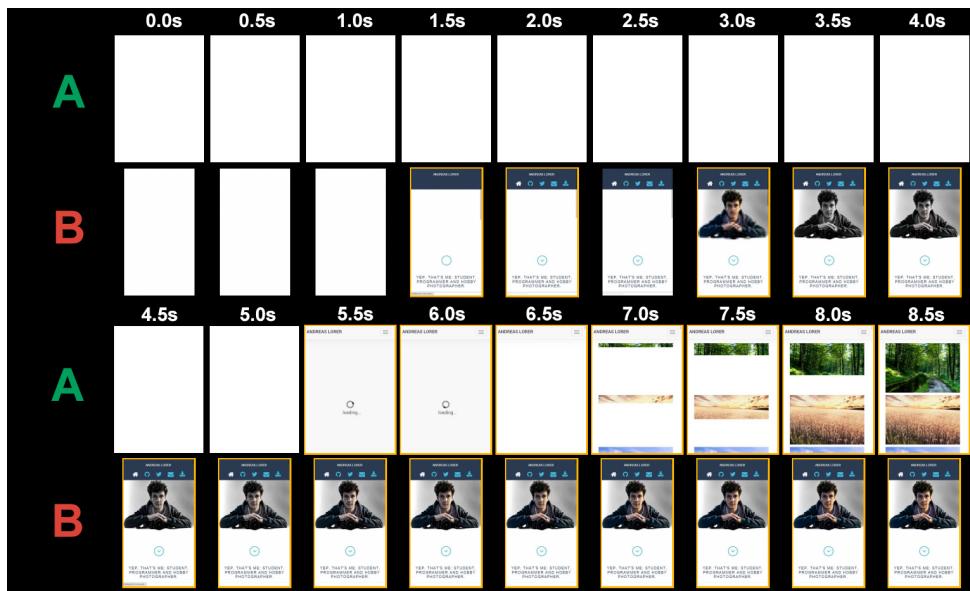


Abbildung 6: Zwei Seiten im Vergleich (Eigene Abbildung via webpagetest.org)

Abbildung 6 zeigt die Seiten A und B, mit nahezu identischer Ladezeit. Der Unterschied besteht darin, dass Seite B bereits nach 1.5 Sekunden eine erste visuelle Rückmeldung für den Anwender zu sehen ist, währendgegen Seite A erst nach 5.5 Sekunden dem Anwender zeigt, dass sie überhaupt ladet. „Perceived Performance“ steht also für die Zeit bis ein erste visuelle Rückmeldung für den Anwender zu sehen ist und bedeutet, dass die Ladezeit als schneller empfunden wird, als es eigentlich laut Messwerten der Fall ist. Warum diese „Perceived Performance“ für eine Webanwendung so wichtig ist zeigen mehrere Studien, deren Daten in folgender Infographik aufbereitet sind.

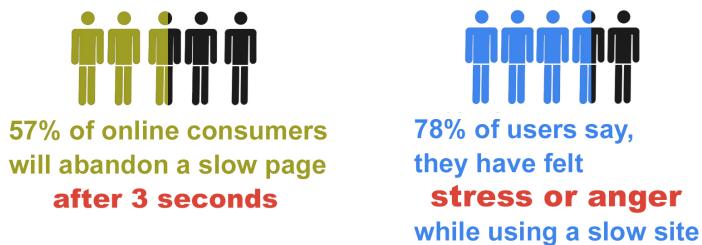


Abbildung 7: Einfluss und Effekt einer langsamen Seite auf den Anwender (Eigene Abbildung nach Daten von: (Radware 2014, p. 8))

Bereits kleine Verbesser- oder Verschlechterungen der Ladezeit können einen großen Einfluss auf den Anwender haben. Yahoo hat herausgefunden, dass wenn eine Seite um nur 400 Millisekunden schneller ist, sich der Traffic um 9% erhöhte.(Stefanov 2008) 57% der Online Konsumenten haben eine Seite, die länger als 3 Sekunden ladet wieder verlassen. 78% der Anwender empfinden sogar Zorn oder Stress wenn eine Seite nicht Ladet oder dies nicht ersichtlich ist.

3 Die 1000 ms Barriere

Das Ziel dieser Arbeit, die 1000 Millisekunden Barriere zu durchbrechen, wurde nicht durch einen Zufall gewählt. Der Anwender nimmt die Geschwindigkeit einer Seite subjektiv wahr. Sie wird in der folgenden Grafik interpretiert:

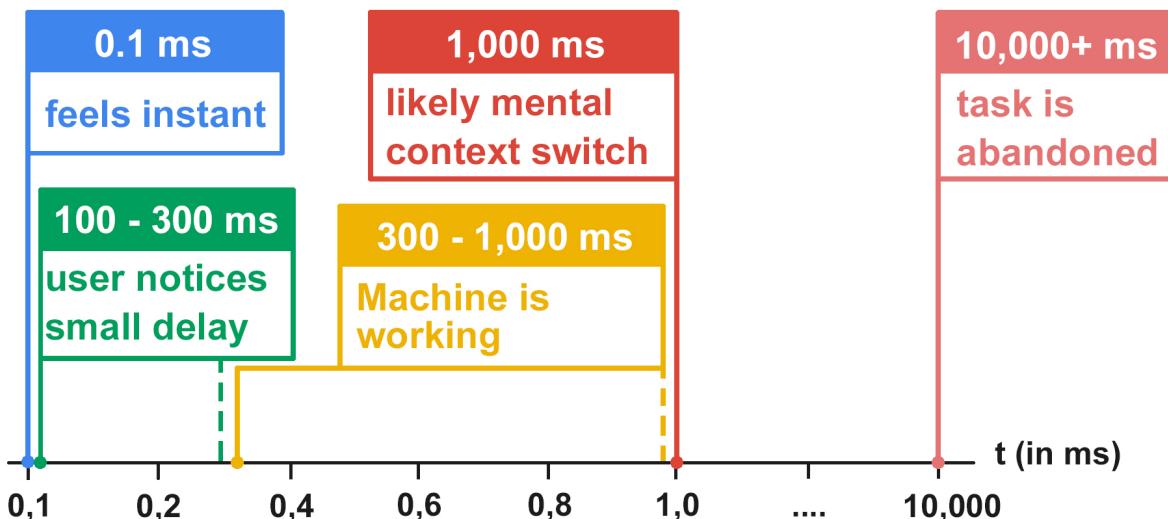


Abbildung 8: Zeit und Wahrnehmung durch den Anwender (Eigene Abbildung nach Daten von: (Grigorik 2013d))

Wie zu sehen ist bleiben gerade einmal eine Sekunde, bevor das Gehirn uns sagt, man solle doch einer anderen Aufgabe nachgehen bis der Ladevorgang abgeschlossen ist. Der Anwender verlangt visuelle Rückmeldung um „am Ball zu bleiben“, dies wurde bereits in Punkt 2.8 „Perceived Performance“ angesprochen. Auf vielen Webseiten sieht man deshalb, Ladebalken oder sogenannte **Spinner**, die dem Anwender sagen, dass der Ladevorgang in Gang, aber noch nicht abgeschlossen ist.

Um das Ziel von einer Sekunde Ladezeit bis zum ersten Render zu erreichen, ist es nötig zu verstehen, womit die meiste Zeit beim Aufrufen einer Webanwendung verbracht wird. Bevor eine Seite mittels Smartphone vom Browser dargestellt wird läuft eine ganze Reihe von Prozessen ab.

3.1 Touch Event

Der Aufruf einer Seite über das Smartphone erfolgt über ein Touch Event auf einen Link, Button oder die Seite wird per URL aufgerufen. Hierbei können je nach Gerät zwischen 50 (iPhone 5) und 123 Millisekunden (Moto X - Android) zwischen der Berührung des Touch Screen und dem Registrieren des Events vergehen.(Takahashi 2013) Der Browser wartet allerdings nochmals bis zu 300 Millisekunden, denn er muss abwarten ob vielleicht noch ein zweiter Finger aufgelegt wird (Multitouch), oder ob der Anwender Scrollen oder Zoomen möchte.(Google 2011)

Dieses Verhalten lässt sich bei vielen Browsern per **Meta Tag** abstellen:

```
1 <meta name="viewport" content="user-scalable=no">
```

Dies setzt natürlich voraus, dass die Webanwendung kein Zoomen benötigt um sie zu bedienen! Gerade bei älteren Webseiten trifft das oftmals nicht zu, da sie keine für das Smartphone

angepasste Ansicht haben (responsive view). Eine vollständige Liste mit Meta Tags für die verschiedenen Browser ist der Fußnote zu entnehmen.⁵

3.2 Netzwerke

Warum gerade das nutzen des Internets per Smartphone so langsam sein kann (und oftmals ist) liegt zu einem Großteil am Netzwerk. Eine Studie untersuchte die Top eine Millionen Webseiten des Internets auf ihre Ladezeiten. Dabei wurde eine Verbindung von 5 Mbit/s und 28ms RTT benutzt. Eine RTT von 28 ms ist sehr schnell, vergleicht man sie zum Beispiel mit der Latenz des 3G Netzes (Abbildung 10). Diese Studie kam zu dem Ergebnis, dass fast 70% der Ladezeit nur durch warten auf das Netzwerk verbracht wird:

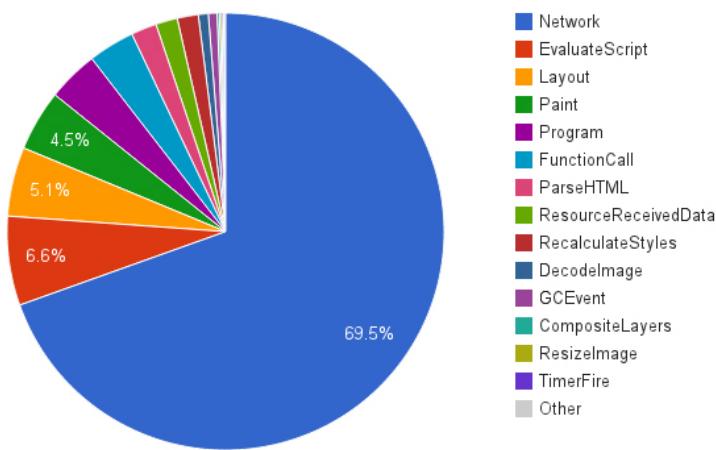


Abbildung 9: Untersuchung der top 1 Millionen Alexa Seiten (Abbildung von: (Tonyg 2013))

Es macht also durchaus Sinn, sich diesen Bereich näher anzusehen, um zu verstehen worauf Einfluss genommen werden kann und wo nicht.

3.2.1 Mobilfunknetz

Es gibt unterschiedliche Mobilfunktsstandards mit denen Anwender Anbindung an das Internet erlangen. Aber selbst wenn einem Anwender 4G vom Mobilfunkanbieter versprochen wird, so ist die Netzaabdeckung mit 4G noch nicht vollständig deckend.

*„Bereits Mitte 2013 konnten laut dem Breitbandatlas der Bundesregierung 70 Prozent der deutschen Haushalte über LTE verfügen. E-Plus startete mit LTE im März 2014.“
(Bundesnetzagentur 2014)*

Das bedeutet, dass der 4G Anwender auf ein niedrigeres Netz wie zum Beispiel 3G ausweichen muss. Die verschiedenen Netzwerke unterscheiden sich entscheidend in ihrer Datenrate und vor allem in der Latenz. Die Tabelle in Abbildung 10 gibt eine Übersicht:

Unser Smartphone ist nicht ständig mit dem „wireless service provider“ verbunden. Ist eine erste Verbindung nötig, so muss das Smartphone dem Sendeturm mitteilen, dass es Kommunizieren möchte. Der Anbieter muss die Anfrage Authentifizieren, die Verbindung herstellen und dann die Anfrage in das Internet weiter leiten. Die Zeit bis eine Authentifizierung erfolgt ist, kann je nach Anbieter und Mobilfunkstandard zwischen <100ms (LTE) und 2,5 Sekunden (3G) liegen

⁵ Suppressing 300ms delay for touchscreen interactions: <http://tinyurl.com/psj5nxz>

Gernation	Data rate	Latency
2G	100 - 400 Kbit/s	300 - 1000 ms
3G	0,5 - 5 Mbit/s	100 - 500 ms
4G	1-50 Mbit/s	< 100 ms

Abbildung 10: Datenrate und Latenz (Eigene Abbildung nach (Grigorik 2013e))

(Grigorik 2013a)! Bereits hier ist zu sehen, dass es „worst case“ Szenarien gibt, durch die es nicht möglich sein kann, dass eine Webanwendung in unter einer Sekunde eine Rückmeldung gibt. Gerade Mobilfunknetze unterliegen Stoßzeiten, die Funksignale von Smartphones können sich gegenseitig stören oder das Signal kann in gewissen Gegenden stärker oder schwächer sein.

3.3 Der HTTP-Request

Nachdem uns unser Mobilfunkanbieter mit dem Internet verbunden hat, kann die eigentliche Anfrage an den Server gestellt werden.

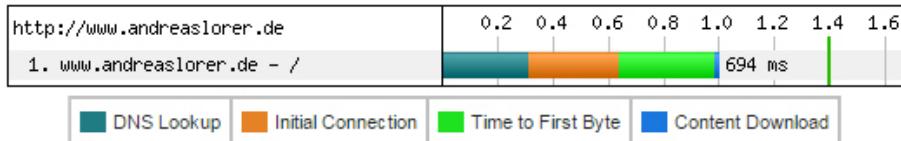


Abbildung 11: Anfrage der HTML-Datei von Irland mittels 3G Netz (Abbildung nach <http://webpagetest.org>)

- DNS Lookup: Um eine Verbindung mit dem Server herzustellen benötigt das HTTP Protokoll die IP Adresse des Ziels. Das heißt der DNS Server wird für den Namen „<http://andreaslorer.de>“ die zu diesem Namen zugehörige IP Adresse zurückgegeben.
- Initial Connection bezeichnet die Zeit die vergeht, bis eine neue Verbindung zum Server hergestellt wurde damit eine Kommunikation zwischen Browser und Server stattfinden kann. Hierbei findet der sogenannte TCP „Three-Way-Handshake“ statt, der dafür einen Round Trip benötigt.
- TTFB: Ist die Abkürzung für „Time to first byte“. Dieser Begriff beschreibt die Zeit die vergeht, bis das erste Byte vom Server beim Browser ankommt. Der Server muss den Request erst zusammenstellen bevor er ihn versenden kann. Dafür werden unter umständen Daten aus der Datenbank abgefragt oder es müssen berechnungen stattfinden. Diese Faktoren beeinflussen die TTFB und kann optimiert werden (schnellerer Server, bessere Datenbankanbindung, Caching).
- Content Download: Die Zeit die benötigt wird bis die Datei vom Server heruntergeladen wurde.
- Nachdem das HTML Dokument heruntergeladen wurde, muss es vom Browser noch gelesen und interpretiert werden. Diese Zeit taucht im Diagramm nicht auf.

3.4 Das Herunterladen einer 40 Kilobyte Datei

Abbildung 12 zeigt Schematisch wie eine 40kb Datei mittels einer neuen TCP Verbindung heruntergeladen wird. Sie soll verdeutlichen, wie vor allem die RTT eine entscheidende Rolle spielt und warum die Latenz die Geschwindigkeit einer Seite viel höher beeinflusst als die Bandbreite.

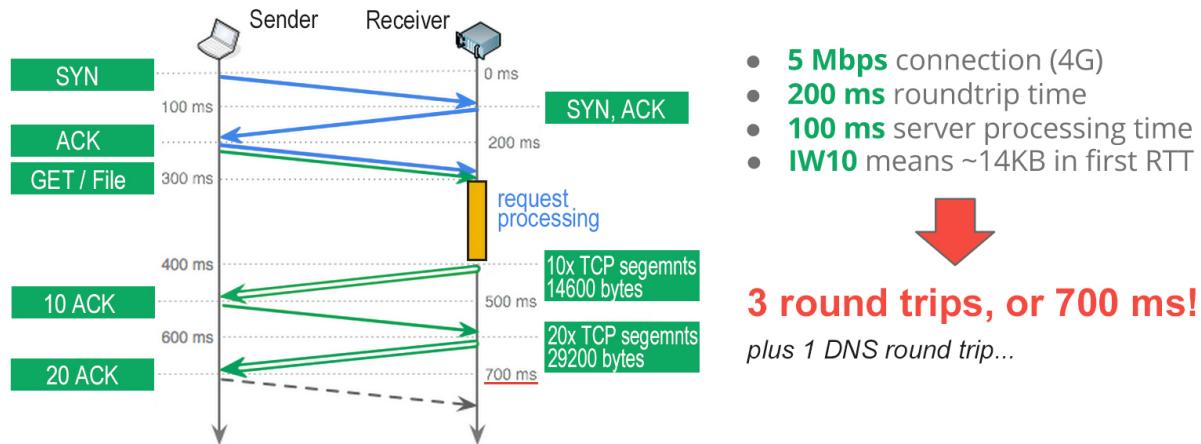


Abbildung 12: Herunterladen von 40kb mittels TCP (Abbildung nach (Grigorik 2013b))

Zuerst erfolgt der DNS Lookup, dann muss TCP per **three way handshake** eine Verbindung aufbauen. Dies kostet bereits 2 round trips, was in diesem Beispiel 400 ms entspricht. Durch den TCP slow start (siehe Punkt: 2.5) steht bei einer neuen TCP Verbindung nicht die volle Bandbreite zur Verfügung. Deshalb kann die volle Datenmenge nicht auf einmal, sondern nur durch zusätzliche round trips heruntergeladen werden. Wenn die Performance einer Webanwendung verbessert werden soll, macht es also Sinn in round trips zu denken. Wieviel round trips sind nötig, bis ich dem Browser Informationen übermittelt habe, so dass dieser etwas anzeigen kann? Idealerweise genau einer.

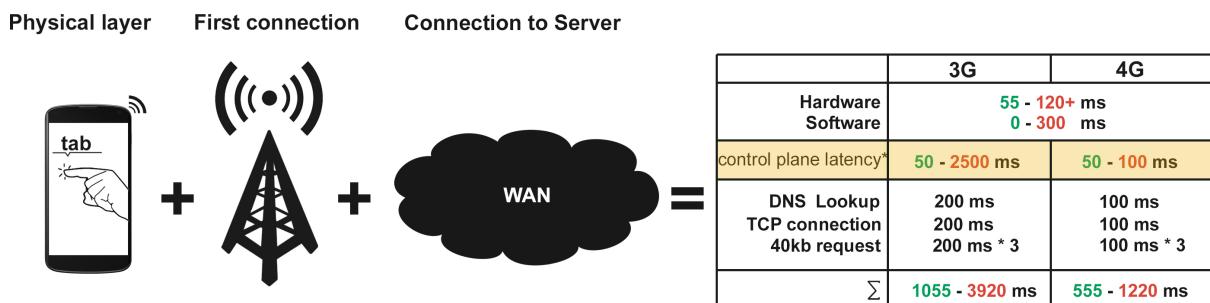


Abbildung 13: *control plane latency: Wenn noch keinerlei Verbindung zu einem Sendeturm aufgebaut wurde, entstehen einmalige Authentifizierungskosten (Eigene Abbildung nach Daten von: (Takahashi 2013)(Grigorik 2013a, p. 7, 12))

Wie in Abbildung 13 zu sehen ist bleibt von dem 1000 Millisekunden Budget nicht mehr viel übrig, wenn alleine die Netzwerkzeiten abgezogen werden. Für Nutzer mit 3G Netz ist es laut dieser These selbst im „best case“ Szenario nicht möglich, die 1000 ms Marke zu durchbrechen. Vor allem wenn man bedenkt, dass das 3G Netz eine Latenz von 200 bis 500 ms haben kann und

hier schon der bestmögliche Wert von 200 ms verwendet wurde.

Für Nutzer des 4G schaut es besser aus und es bleiben 445 ms übrig.

Das bedeutet es müssen in den ersten Kilobytes, soviel nützliche Informationen vorhanden sein, damit der Browser bereits anfangen kann mit dem rendering zu beginnen, obwohl noch nicht alle Daten heruntergeladen sind. Noch spitzer formuliert: Der Browser sollte mit den ersten 14 KB (das ist die Menge an Daten die der erste round trip transportieren kann, siehe Abbildung 12) bereits den **above the fold** Bereich rendern können. Um das zu ermöglichen ist es nötig, den Kritischen Rendering-Pfad zu optimieren.

3.5 Zusammengefasst

Dieses Kapitel hat aufgezeigt, dass die Bandbreite eine nur untergeordnete Rolle spielt, wenn von schnellen Webanwendungen die Rede ist. Die Ladezeit wird dominiert von der Latenz und der damit verbundenen round trip time. Diese entscheidet maßgeblich, wie schnell oder wie langsam der Ladevorgang ist. Folgendes ergibt sich auf der Ebene des Netzwerks:

- Die Ladezeit wird für mobile Anwender durch die Latenz bestimmt. 4G kann hier bereits Zeiten von <100 ms liefern, was das erreichen der 1000ms Barriere enorm erleichtert.
- Bei Anwendern die mittels 3G Netzwerk im Internet surfen besteht wenig Möglichkeiten eine Seite in unter einer Sekunde zu übermitteln, denn die Zeit ab dem Touch Event und im Netzwerk kann bereits schon 900 ms betragen. Diese These deckt sich auch mit den Werten, die im Laufe des Projektes gesammelt wurden. Eine Auswertung davon findet unter Punkt 6 statt.
- Näheres Platzieren der Bits: Durch die Benutzung eines CDN's lassen sich Bits und Bytes näher am Endanwender platzieren was die Netzwerkzeiten verringert.
- Sage das Nutzerverhalten voraus: Wenn der Anwender in einer Einkaufs-App 3 Schritte zum vollenden des Kaufvorgangs benötigt, dann lässt sich bei Schritt 1 bereits vorhersagen was er für weitere Ressourcen im nächsten Schritt benötigt. Diese Ressourcen könnten bereits geladen werden. Nachteil: Wenn der Nutzer nie zu Schritt 2 oder 3 kommt, wurde das Internetvolumen des Anwenders (für die der Nutzer eventuell zahlt) umsonst belastet.
- Wahl eines guten Hostings: Die RTT als auch die **Server Response Time** sind je nach Anbieter unterschiedlich. Ein gutes Hosting kann hier unter Umständen bereits eine enorme Verbesserung bedeuten. Zur Not sollte gewechselt werden!

Wie in der Grafik zu sehen ist, sank die Response Zeit meines Hosting Providers von durchschnittlichen fast 400 Millisekunden auf 183 ms. Dies kann mehrere Gründe haben, so kann sich das Routing zum Server geändert haben, der Server kann ein Update erhalten haben oder die Maschine kann gewechselt worden sein. Was letzten Endes dazu geführt hat kann nicht genau gesagt werden.

- Senden von weniger Daten: Das schnellste Bit ist das, dass nicht gesendet wird. Das zusammenfügen und Verkleinern von Javascript und CSS Dateien verringert die Dateigröße. Zudem lassen sich die Daten per GZIP zwischen Browser und Server komprimieren. Aus Abschnitt 2.3 wissen wir bereits, dass die Anzahl von parallelen TCP Verbindungen limitiert ist und Abschnitt 2.5 den Einfluss des TCP slow start gezeigt. Deshalb macht es Sinn, möglichst wenige Dateien auszuliefern, denn jede Datei benötigt einen extra TCP Verbindungsaufbau und jede neue TCP Verbindung unterliegt dem TCP slow start. Wie dies in der Praxis umgesetzt wird soll unter Punkt ?? konkretisiert werden.



Abbildung 14: Verringerung der response time für <http://andreaslorer.de> (Abbildung nach pingdom.com)

- Stelle nützliche bytes zu Verfügung: Wie in Abbildung 12 zu sehen ist, erfolgt mit dem ersten round trip eine Datenübertragung von ca. 14 KB. Optimal ist es, wenn bereits mit dieser ersten Antwort genügend Informationen vorliegen um etwas auf den Bildschirm des Anwenders zu rendern. Das setzt voraus, dass die HTML Datei nicht größer ist als 14 KB (nach Kompression).
- Vermeiden von Weiterleitungen: Abbildung 15 zeigt den Seitenaufruf von hasbro.com. Wie zu sehen ist, gibt es einen HTTP 301 (Wert in Klammer) Response zurück:

301 - Moved Permanently: „Die angeforderte Ressource steht ab sofort unter der im „Location“-Header-Feld angegebenen Adresse bereit (auch Redirect genannt). Die alte Adresse ist nicht länger gültig.“ (wikipedia 2014)

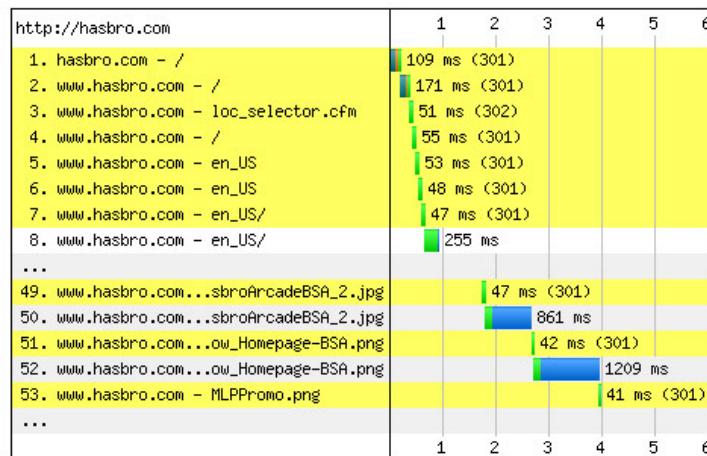


Abbildung 15: Testlauf von hasbro.com: „Dulles, VA USA - Thinkpad T430 - Chrome - Cable“ via webpagetest.org)

Wenn ein Anwender `hasbro.com` eingibt erfolgt der DNS Lookup und die TCP Verbindung wird aufgebaut. Der Browser erfährt dann, dass die Ressource unter einer anderen Adresse zur Verfügung steht. Danach erfolgt die DNS Auflösung für `www.hasbro.com` bei dem der gleiche Vorgang vonstatten geht. Es erfolgt die Weiterleitung auf die jeweilige Ländersprache von `www.hasbro.com`, wieder mit dem selben Vorgang. Nach rund 900ms konnte die erste Anfrage an die richtige Zieladresse aufgegeben werden! Aber auch Bilder werden auf dieser Seite Weitergeleitet wie in den Anfragen 49, 51 und 53 zu sehen ist. Ruft man sich die RTT von einem 3G Netz ins Gedächtnis dürfte klar werden, was Weiterleitungen für den Smartphonenuutzer bedeuten können.



Abbildung 16: Testlauf mittels Smartphone von hasbro.com: „Dulles VA USA - Modell: MOTOG. 3G shaped 1.6Mbps / 300ms RTT“ Detaillierter Test unter: http://www.webpagetest.org/result/150310_AH_HVD/1/details/

Abbildung 16 zeigt den Aufruf von hasbro.com mittels Smartphone mit 3G Netz. Katastrophale 5.5 Sekunden dauert alleine der Verbindungsauaufbau zum HTML Dokument der Seite!

Die Weiterleitung von `hasbro.com` auf `www.hasbro.com` ist für die Suchmaschinenoptimierung⁶ allerdings sinnvoll, denn sonst würde unter zwei Namen der gleiche Inhalt zu finden sein. Dies ist aus Sicht von Google „Duplicated Content“ und kann zu einer Abstrafung im Ranking führen.⁷

3.6 Kritischer Rendering-Pfad

Auf Englisch „critical render path“ genannt, ist der wohl wichtigste Begriff, wenn es um schnelle Ladezeiten geht. Durch die Optimierung des Rendering-Pfads kann die benötigte Zeit für das erste Rendern der Seite erheblich verkürzt werden. Das Verständnis des Rendering-Pfads ist zudem eine wesentliche Voraussetzung für die Erstellung von schnellen Webanwendungen und soll in diesem Abschnitt ausführlich erklärt werden. Dabei wird der Begriff in seine Teile zerlegt: Kritischer, Rendering und Rendering-Pfad.

3.6.1 Rendering

Rendering: Der Browser liest das HTML Dokument und übersetzt es. Diesen Vorgang nennt man auch Parsen. „Das Parsen der HTML- und CSS-Ressourcen und Ausführen von JavaScript be-

⁶engl. SEO - search engine optimization

⁷Mehr zu diesem Thema gibt es unter <http://www.sem-deutschland.de/seo-tipps/duplicate-content-definition/>

anspricht Zeit und Clientressourcen. Je nach Geschwindigkeit des Mobilgeräts und Komplexität der Seite kann dieser Prozess Hunderte von Millisekunden in Anspruch nehmen.“ (Google 2014b) Bei der Optimierung von Webanwendungen ist besonderst das Auftreten des ersten Zeichnens interessant (engl. „First Paint Event“). Je früher das sogenannte **First Paint Event** auftritt umso höher ist die **Perceived Performance** der Webanwendung.

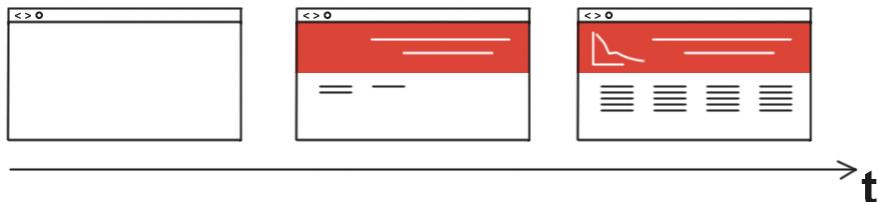


Abbildung 17: Der Render Prozess (Eigene Abbildung)

3.6.2 Rendering-Pfad

Der **Rendering-Pfad** setzt sich aus den für die Anwendung nötigen Ressourcen zusammen. Webanwendungen bestehen schließlich nicht nur aus einer HTML Datei, sondern aus mehreren Javascript und CSS Dateien.



Abbildung 18: Ressourcen die für das Rendern von nötigen sind (Eigene Abbildung)

Gegeben ist das folgende Beispiel einer simplen HTML Datei.

```

1  <!DOCTYPE html>
2  <meta charset="utf-8">
3  <title>Web Performance fuer den mobilen Endanwender</title>
4
5  <link href="assets/styles.css" rel="stylesheet" />
6  <script src="assets/script.js"></script>
7
8  <p> Hello world! </p>
```

Listing 1: Beispiel Code

Nachdem diese Datei heruntergeladen wurde, beginnt der Browser sie von oben nach unten zu Parsen. Dabei stößt er in Zeile 5 auf einen **Link-Tag** der ihn anweist diese Datei herunterzuladen.

In Zeile 6 findet der Browser einen **Script-Tag**. Auch diese Datei muss heruntergeladen, interpretiert und ausgeführt werden, denn jede Javascript Datei kann den DOM-Baum⁸ oder das CSS manipulieren. So können per Javascript sowohl Elemente dem DOM-Baum hinzugefügt, als auch weggenommen werden oder Elemente können eine Änderung ihrer CSS Attribute erhalten. Dieser Umstand verbietet es dem Browser mit dem Rendering zu beginnen, da bis zur Ausführung der Javascript Dateien noch Manipulationen erfolgen können. Bevor also das „Hello world“ in Zeile 8 angezeigt werden kann, ist das rendern blockiert. Dieses Verhalten nennt sich auch „Render Blocking“ und wird sowohl von Javascript als auch von CSS Dateien ausgelöst. Folglich spricht man hierbei auch von „Render Blocking Javascript“ und „Render Blocking CSS“. Erst wenn diese Blockierenden Ressourcen geladen und interpretiert wurden, kann der Browser mit dem Rendern beginnen.

3.6.3 Critical render path

Critical render path sind genau die Javascript und CSS Dateien, die für den für das Rendern des **above the fold** (Punkt: 2.7) von nötig sind. Um dies umzusetzen ist es nötig, die Ressourcen in zwei Teile zu zerlegen: Für das Rendering **absolut** notwendig und nicht notwendig. Alle Dateien die nicht notwendig für das erste Rendern sind sollten so lange mit dem Laden verzögert werden, bis die Anwendung geladen ist. Wie genau so eine Umsetzung aussieht, wird in Punkt: ?? ausführlich gezeigt.

3.6.4 Zusammengefasst

Folgende Pattern lassen sich, bedingt durch den **Kritischen Rendering-Pfad**, für die Erstellung von Webanwendungen ableiten.

- CSS Dateien möglichst weit oben im „<head>“ Bereich platzieren und Javascript vor dem schließen des „</body> tags“. Da Javascript und CSS so lange Blockieren, bis sie heruntergeladen wurden, kann mit dem Parsen des gesamten Dokumentes nicht fortgefahrene werden. (Bart 2014)
Das Platzieren von Javascript am Ende des Dokuments hat allerdings den Nachteil, dass sich der Zeitpunkt des Herunterladens verzögert. Deshalb ist es ratsam genau die Javascript Dateien in den „<head>“ zu verlagern, die **Kritisch** für das Rendern des **above the fold** Bereichs sind und den rest der Scripte vor den „</body> tag“.
- Zusammenfügen von Dateien: Je weniger einzelne Dateien umso weniger wird das Rendern der Seite blockiert.
- Aufteilen von Ressourcen in 2 Gruppen: Für das Rendering kritisch und unkritisch. Das gilt sowohl für CSS als auch für Javascript Dateien. Unkritische Dateien werden solange verzögert, bis der above the fold der Seite geladen wurde.
- Inlining von CSS im HTML Dokument: Durch das Einbetten von CSS direkt in das HTML Dokument wird das CSS bereits mit der ersten Server Antwort mitgesendet. Dadurch muss der Browser die Datei nicht anfordern und heruntergeladen, sondern kann gleich mit dem Parsen beginnen.
- Die Herausforderung besteht darin, in der eigenen Webanwendung die für das Rendern kritischen Ressourcen zu erkennen und aufzuteilen, ohne die Funktionalität der Anwendung in Mitleidenschaft zu ziehen. Dinge die fast immer verzögert geladen werden können sind zum

⁸Der DOM-Baum: <http://wiki.selfhtml.org/wiki/JavaScript/Objekte/DOM>

Beispiel Social Media Buttons (Facebook, Google+, Twitter ect.), Widgets oder Tracking Codes wie Google Analytics.

3.7 Analyse des Wasserfalls

Für ein besseres Verständnis des Kritischen Rendering-Pfads soll ein praktisches Beispiel einer nicht optimierten Seite helfen. In Abbildung 19 ist ein Ausschnitt eines Wasserfallmodells dargestellt.

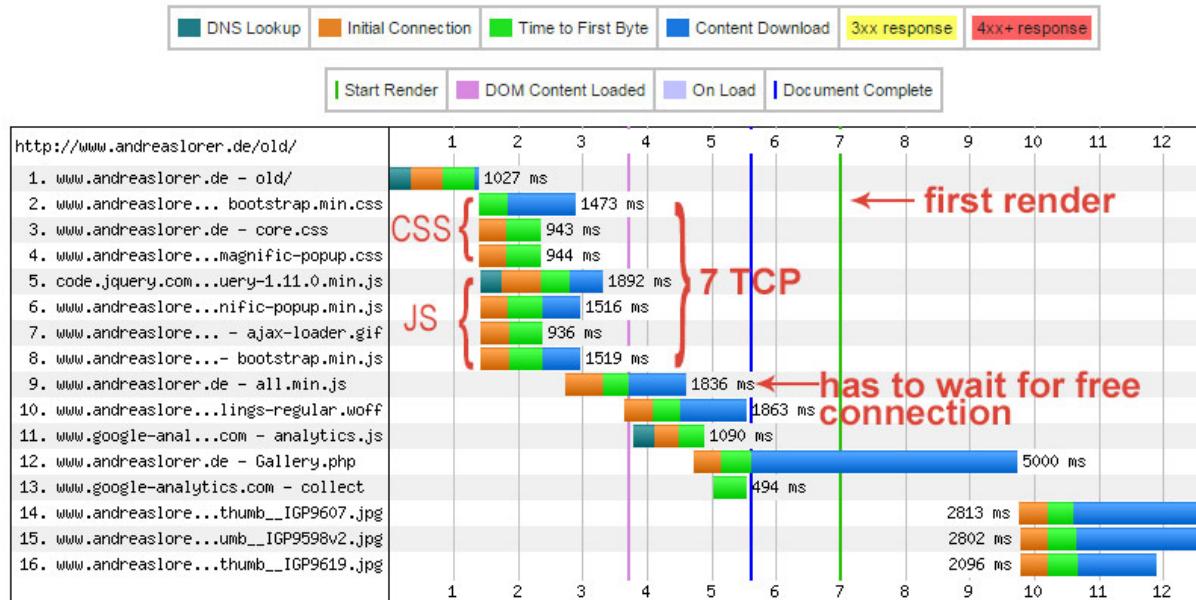


Abbildung 19: Testlauf von: „Dulles VA USA - Modell: MOTO.G. 3G shaped 1.6Mbps / 300ms RTT“ Ganzer Test: http://www.webpagetest.org/result/150308_A1_2W4/8/details/

Die Abbildung zeigt, dass typische Verhalten des Browsers: Es werden zuerst CSS, Javascript und anschließend die Bilder heruntergeladen. Hierbei fällt auf, dass er nicht mit allen Dateien gleichzeitig beginnen kann, sondern wie in Punkt 2.3: HTTP/1.1 nur 6 TCP Verbindungen pro Host Name aufbauen darf.⁹ Das bedeutet, das alle weiteren Ressourcen auf eine frei werdende TCP Verbindung warten müssen. Je weniger einzelne Dateien die Webseite benötigt, umso weniger bilden sich Warteschlangen für eine frei werdende TCP Verbindung (der Wasserfall wird flacher).

Auch in diesem Diagramm ist die Latenz als dominierender Faktor zu sehen. Es fällt auf, dass es nur einen relativ langen blauen Balken gibt (Anfrage #12 gallery.php) bei dem für längere Zeit etwas heruntergeladen wird. Das herunterladen der meisten Inhalte dauert überwiegend nur so lange, wie die Zeit die nötig war um eine Verbindung herzustellen.

Die senkrechte blaue Linie bedeutet: **Document Complete** und das heißt für den Browser, dass alle für das Rendern nötigen Ressourcen fertig heruntergeladen wurden und nun vorhanden sind. Dadurch kann er bei Sekunde 7 (senkrechte grüne Linie) mit dem Rendern beginnen. Das Ziel des Kritischen Rendering-Pfads ist es, diese senkrechte grüne Linie möglichst weit nach links zu schieben, also die Zeit bis zum ersten Rendern zu minimieren. Zum Vergleich nun das Wasserfallmodell einer Optimierte Seite:

⁹Es sind deshalb 7 Verbindungen, da die Datei: „code.jquery“ von einer Google Domäne kommt und deshalb als neuer Host Name zählt.

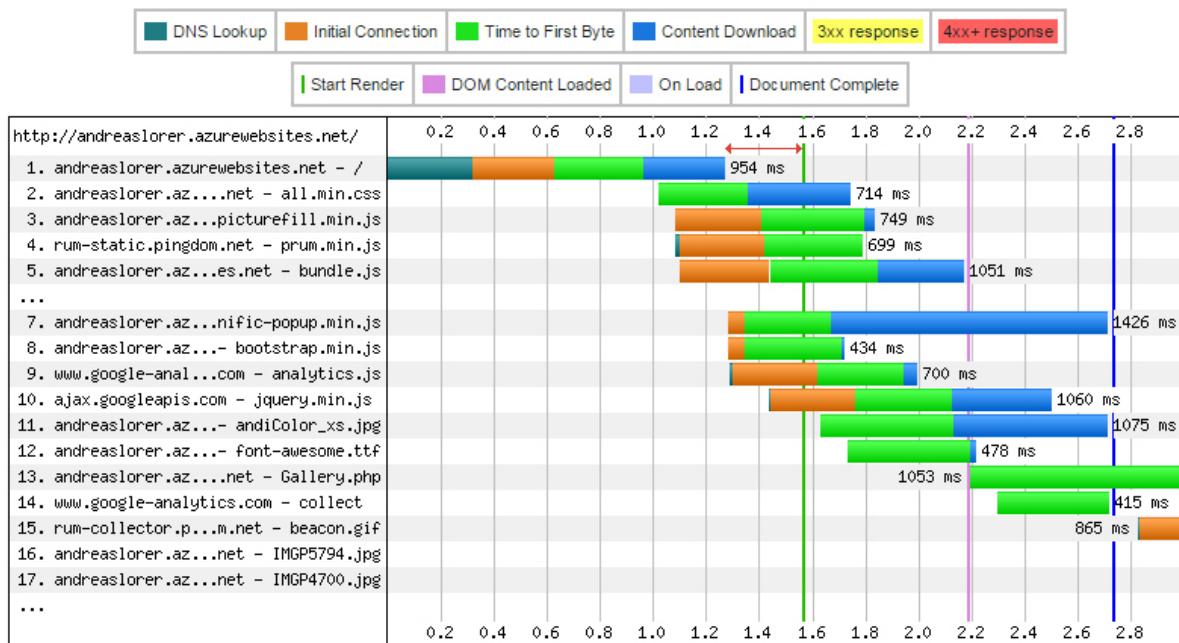


Abbildung 20: Selbe Testbedingungen wie bei Abbildung 19. Ganzer Test: http://www.webpagetest.org/result/150308_5V_JSD/6/details/

Wie zu sehen ist, fällt die senkrechte grüne Linie bereits viel früher bei rund 1.6 Sekunden. Das CSS und Javascript wird zu diesem Zeitpunkt noch heruntergeladen. Daraus lässt sich schlussfolgern, dass in dieser optimierten Version kein **render blocking Javascript / CSS** mehr vorhanden ist. Dadurch ist der Browser nicht blockiert und kann bereits früh mit dem Rendern der Seite beginnen. Dieses Diagramm lässt sich noch viel weiter interpretieren und belegt die Hauptaussage dieses Kapitels „Brechen der 1000 ms Barriere“:

Request Nummer 1 zeigt genau dass, was im Kapitel „3.2 Netzwerke“ beschrieben wurde.

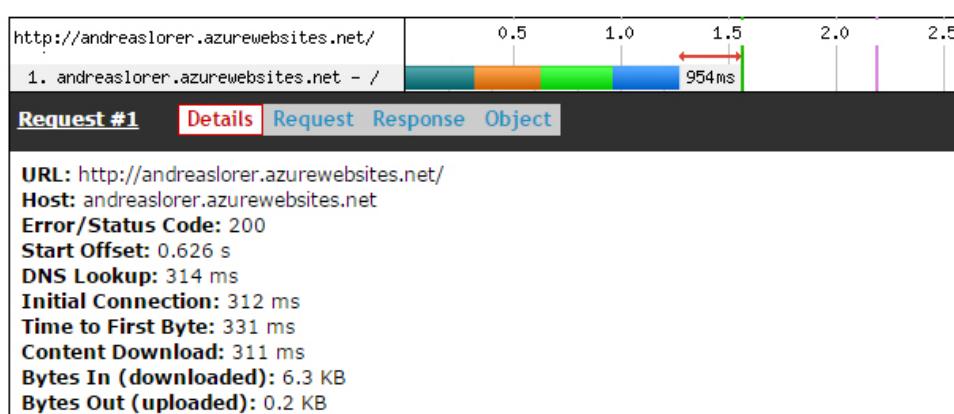


Abbildung 21: Request Nr. 1 im Detail (Abbildung nach webpagetest.org)

Mit gerundeten Werten ergibt sich:

- DNS Lookup: 1 RTT = 300 ms

- Initial Connection (TCP - 3 Way Handshake): 1 RTT = 300 ms
- TTFB: Server Processing Time¹⁰: = 300 ms
- Content Download¹¹: 1 RTT = 300 ms
- Zeit fürs Parsen, Ausführen und Rendern: ca. 500 ms ¹²:

Obwohl hier nur eine 6.3kb Datei heruntergeladen wurde und keine 40kb so wie in Beispiel 13, ist es ohne eine geringere Latenz (4G) nicht möglich unter die 1000 ms Barriere zu kommen. Leider stellt der Serviceanbieter von [Webpagetest.org](#) keine Testumgebung mit 4G zu Verfügung. Ein Beweis für das erreichen eines „first render“ mittels Smartphone in unter 1000 ms Sekunde steht folglich aus. Mittels Kabelverbindung sind Werte um die 300 ms zu erreichen, dies wird per Datenauswertung in Kapitel 6 gezeigt.

¹⁰die Antwort muss erst vom Server generiert werden

¹¹Das HTML Dokument beträgt 6.3 KB (siehe in Abbildung 21 Eintrag: Bytes In (downloaded)), TCP kann mit dem ersten round trip rund 14 KB transportieren. Das HTML Dokument kann also in einem round trip geliefert werden.

¹²Dies ist in der Abbildung mittels rotem Pfeil (<- ->) markiert

4 Entwicklung

Dieses Kapitel soll den Entwicklungsprozess konkretisieren und den Optimierungsprozess einer Webanwendung aufzeigen. Es soll erläutern, welche Fragen sich stellten und welche Antworten darauf gefunden wurden. Wie bestimmte Probleme gelöst wurden. Welche Tools und Hilfsmittel zur Verwendung kamen. Dies soll ein Bewusstsein dafür schaffen, was möglich ist und wie eine technische Umsetzung aussehen kann.

4.1 Tools

Dies ist eine Auflistung an Tools und nützlichen Seiten, die entweder im Projekt verwendet, oder die für Wertvoll befunden wurden und deshalb hier ihren Platz finden, damit jeder für sich entscheiden kann, ob der Einsatz davon sinnvoll sein könnte.

4.1.1 Google Chrome Developer Tool

Dieses Tool ist über die Taste F12 im Chrome Browser zu finden. Nützliche Features sind:

- **Device Emulation**¹³: Damit lassen sich verschiedene Devices wie Smartphones, Ipad oder verschiedene Desktopauflösungen simulieren. Auch das Touchverhalten wird Simuliert.
- In der Device Emulation lässt sich auch die Netzwerkgeschwindigkeit simulieren. Dies ist allerdings nur eine Simulation und kann unter wahren Bedingungen stark abweichen.
- Netzwerk: Hier lässt sich das Wasserfallmodell nachvollziehen. Auch lässt sich hier das Caching des Browsers abschalten, während das Developer Tool geöffnet ist.
- Audits: Unter diesem Reiter bekommt man erste Informationen, welche Verbesserungen es für diese Seite aus dem Gesichtspunkt der Performance ergeben. So wird zum Beispiel aufgezeigt, wie viele CSS Selektoren auf dieser Seite gar keine Verwendung finden (gerade bei CSS-Frameworks wie Bootstrap kann es sein, dass rund 90% der Selektoren keine Verwendung haben)

4.1.2 Google Pagespeed Insight

Pagespeed Insight ist ein Analysetool für Webanwendungen. Per URL Eingabe wird die Anwendung aufgerufen und gegen die „Best Practices“ von Google getestet:¹⁴. Dabei wird ein Rating von 1 (schlecht) bis 100 (gut) vergeben. Mobile und Desktop Version werden voneinander unabhängig bewertet. Findet das Tool Verstöße gegen die **best practices**, so gibt es Hilfestellungen wie zum Beispiel weiterführende Links oder Hinweise zur Behebung des Problems. Für die Verbesserung der Performance ist dieses Tool eines der besten Anlaufziele, um einen Überblick zu bekommen wo sich die Probleme befinden. Pagespeed Insight gibt es auch als Plugin für das Google Chrome Developer Tool.¹⁵

4.1.3 Google Closure Compiler

Ein simples Tool von Google¹⁶, mit der Aufgabe Javascript zu verkleinern. Dieser Vorgang nennt sich auch „minify“ und ist auch für HTML und CSS möglich. Ein Beispiel:

¹³Bei geöffnetem Tool (F12): strg + shift + M oder klick auf das Smartphone Symbol

¹⁴<http://tinyurl.com/nvxksks>

¹⁵Plugin - Pagespeed Insight: <http://tinyurl.com/mv8fcx8>

¹⁶<http://closure-compiler.appspot.com/>

Listing 2: Input

```

1  /**
2   * urlEncodes an object to send it via post
3   * @param {Object} object Object with key value pairs
4   * @return {String} string in format key=value&foo=bar
5   */
6  var urlEncode = function (object) {
7      var encodedString = '';
8      for (var prop in object) {
9          if (object.hasOwnProperty(prop)) {
10             if (encodedString.length > 0) {
11                 encodedString += '&';
12             }
13             encodedString += encodeURIComponent(prop + '=' + object[prop]);
14         }
15     }
16     return encodedString;
17 };

```

Wird zu:

Listing 3: Output

```

1  var urlEncode=function(c){var a="",b;for(b in c)c.hasOwnProperty(b)&&
2  (0<a.length&&(a+="&"),a+=encodeURIComponent(b+"="+c[b]));return a};

```

Wie zu sehen ist, werden nicht nur alle Kommentare, Leerzeichen und Zeilenumbrüche entfernt, sondern auch Variablennamen werden auf 1 Zeichen reduziert um weitere bytes zu sparen. Die Funktionalität bleibt dabei gewährleistet. Dieser Vorgang ist auch unter dem Namen „uglify“ bekannt.

4.1.4 Webpagetest

[Webpagetest.org](http://webpagetest.org) ist das wohl umfangreichste und beste Website-Analysetool das im Internet zu finden ist. Es ist ein kostenloser Service der hauptsächlich von Patrick Meenan entwickelt wurde. Das Tool ist leicht zu bedienen aber schwer zu beherrschen („easy to use, hard to master“) und es gibt zahllose Einstellungen und undokumentierte Funktionen auf die man nur in Vorträgen oder Foren stößt. Es gibt auch ein Buch, dass sich nur mit diesem Tool beschäftigt, beim Verlag O'Reilly.¹⁷ Die Features für Webpagetest sind vielseitig:

- Es lassen sich Webanwendungen mittels eines in der realität existierenden Geräts testen. So kann vom Standort Dulles VA ein MOTOG zum Testen einer Seite verwendet werden. Dieses Gerät ruft dann auch wirklich die eingegebene URL auf und die darunterliegende Schicht misst die Zeit. Abbildung 22 zeigt den Teststandort Dulles VA.¹⁸
- Webpagetest hat die wohl genaueste Erfassung von Netzwerkzeiten und spiegelt damit realitätsgetreu die Ladezeiten einer Seite wieder.
- Webpagetest liefert eine enormes Spektrum an Daten und Diagrammen, was ausführliche Analysen zulässt.
- Speed Index: Dies ist eine von diesem Tool eigene Maßeinheit zum bestimmen der **Perceived Performance** einer Seite. „*The Speed Index metric was added to WebPagetest in April, 2012 and measures how quickly the page contents are visually populated (where lower numbers*

¹⁷ Buch - Using WebPagetest: <http://shop.oreilly.com/product/0636920033592.do>

¹⁸ Einen detaillierten einblick vom Gründervater und Entwickler Patrick Meenan gibt es hier: <http://tinyurl.com/o4b3rxh>

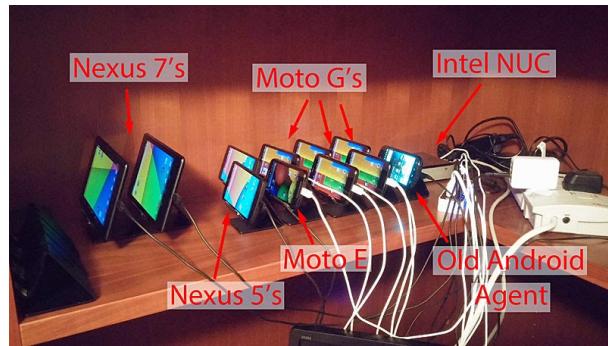


Abbildung 22: Webpagetest Android device farm (Abbildung von (Meenan 2014))

are better). It is particularly useful for comparing experiences of pages against each other (before/after optimizing, my site vs competitor, etc) and should be used in combination with the other metrics (load time, start render, etc) to better understand a site's performance.“(webpagetest.org 2015)

- Man kann Tests direkt miteinander vergleichen. Das ist möglich, indem diese URL eingegeben wird: www.webpagetest.org/video/compare.php?tests= und nach dem „=“ Zeichen die Test ID eingibt, beispielsweise „150310_8E_GRH“. Mit einem Komma getrennt wird eine 2. oder 3. ID angefügt. Die Tests werden dann in einer Vergleichsansicht dargestellt.
- Filmstrip Ansicht: Damit lässt sich visuell erkennen, wann welches Element gerendert wird.
- Video erstellung: Aus der Filmstrip Ansicht lässt sich ein Video erstellen. Das ist vor allem interessant, wenn mit der Vergleichsmethode mehrere Tests geladen sind. Der Ladevorgang der Testläufe wird dann in einem Video Parallel abgespielt. Vor allem für Präsentationen oder vorher / nachher Vergleiche ist dies nützlich.
- Test History: Durch eine Registrierung auf der Seite wird ein eigenes Testprofil angelegt in dem alle Test-ID's gespeichert werden.
- Testen von verschiedenen Standpunkten: Webpagetest ermöglicht es die eigene Seite von ganz verschiedenen Geographischen Standpunkten aus aufzurufen. Dadurch lässt sich ein Eindruck gewinnen, wie schnell die Seite aus dem Ausland aufrufbar ist und wie stark die Abweichung sein kann.
- API: Webpagetest hat eine offene API (Schnittstelle) durch die das Tool von außerhalb erreichbar ist. So lässt sich ein Test beispielsweise in Google-Spreadsheets aufrufen und das Ergebnis direkt in eine Tabelle schreiben. Mehr dazu in Punkt: ???. Diese Schnittstelle Limitiert allerdings die Anzahl an Tests pro Tag auf 200. Für mehr muss man sich eine eigene Private Instanz erstellen.
- Private Instanz: Da webpagetest Open Source ist, gibt es die Möglichkeit eine eigene Private Instanz aufzusetzen. Dies kann sowohl per Amazon Cloud oder auf einem eigenen Server geschehen. Damit lassen sich dann soviele Tests ausführen, wie die Leistungs des Servers bietet.

4.1.5 Pingdom

<http://tools.pingdom.com/fpt/> ist eine Alternative zu Webpagetest. Auch damit lässt sich eine URL nach Performanceproblemen analysieren. Die Ergebnisse sind nicht so genau wie mit Webpagetest und auch ein Testen mit Smartphones fehlt. Bei einer kostenlosen Anmeldung erhält man allerdings ein System zur Überwachung der eigenen Webanwendung. Bei Ausfall oder zu hoher Last kann eine SMS versendet werden um den Admin auf diesen Umstand hinzuweisen. Durch einbetten eines Scripts auf der eigenen Seite lässt sich die Response zeit aufzeichnen (siehe Abbildung 14). Dieses tracking nennt man auch „real user monitoring“ und ist zum Beispiel auch durch Google Analytics in solch einer Form abrufbar.

4.1.6 Speedcurve

Ist ein kommerzielles Tool basierend auf Webpagetest. Es liefert einen „life monitoring“ Service mit dem sich Webanwendungen vergleichen lassen. So kann man zum Beispiel die eigene Webanwendung dauerhaft und über einen längeren Zeitraum mit denen der Konkurrenz vergleichen.



Abbildung 23: Speedcurve Life Monitoring (Abbildung von <http://speedcurve.com/>)

4.1.7 Google Spreadsheet

Ist im Grunde wie Microsofts Excel. In Tabellen können Werte eingetragen und Berechnungen ausgeführt werden. Der große Vorteil an Google Spreadhseet besteht in der Möglichkeit, dass es einen Skript Editor gibt, mit dem sich kleine Programme schreiben lassen. So sind zum Beispiel API Abfragen möglich, dessen Ergebnis dann direkt in die Tabelle geschrieben werden kann.

4.1.8 Feed the Bot

<http://www.feedthebot.com/pagespeed/> bietet umfassende Artikel zu SEO und web performance. Wenn man sich mit dem Thema web performance beschäftigen möchte, ist dies eine erstklassige Anlaufstelle.

4.1.9 What Does My Site Cost?

„Was kostet es eigentlich meinene Seitenbesucher, wenn sein Datenvolumen für diesen Monat aufgebraucht ist und er pro verbrauchtes MB zur Kasse gebeten wird?“ Diese Frage versucht diese Webanwendung zu klären und visuell darzustellen.

<http://whatdoessitecost.com/> benutzt die webpagetest Schnittstelle um eine eingegebene URL zu Analyisieren und berechnet aus den billigsten Anbieteren pro Land einen Preis für den Aufruf der Seite mittels Smartphone:

*„Prices were collected from the operator with the largest marketshare in the country and the for the least expensive plan with a (minimum) data allowance of 500 MB over (a minimum of) 30 days. Prices include taxes. Because these numbers are based on the least expensive plan, they are **best case** scenarios.“*(whatdoessitecost.com 2015)

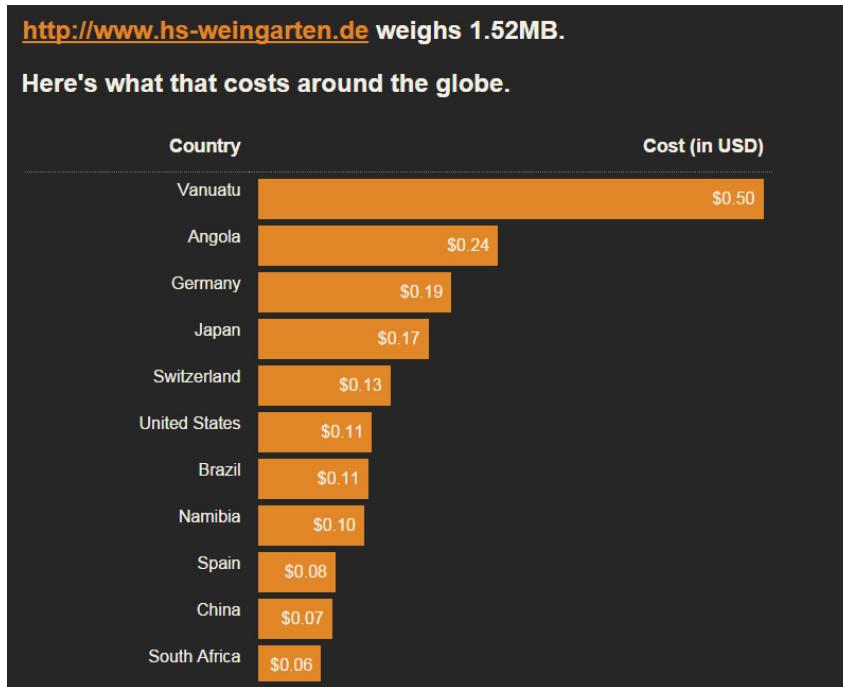


Abbildung 24: Find out how much it costs for someone to use your site on mobile networks around the world.(whatdoessitecost.com 2015)

In Deutschland kostet also der Seitenaufruf von hs-weingarten.de rund 20 Cent. Dieses Tool stellt auf sehr schöne Art und Weise dar, dass schlechte web performance nicht nur den Anwender verärgert, sondern zusätzlich zum Ärger auch noch bares Geld kosten kann.

4.1.10 Critical Path CSS Generator

Im Kapitel „Brechen der 1000 ms Barriere“³ wurde gesagt, man solle das CSS des above the fold direkt in das HTML als inline CSS schreiben. <http://jonassebastianohlsson.com/criticalpathcssgenerator/> erstellt aus einer gegebener URL und dem dazugehörigen CSS genau den CSS-Code, der für den above the fold Bereich nötig ist. Das Ergebnis lässt sich dann bequem in das eigene HTML einfügen.

Dieser Generator funktioniert allerdings nur dann gut, wenn sowohl die Smartphone, als auch Desktop Darstellung identisches CSS haben. Bootstrap zum Beispiel manipuliert die Navigation auf der Smartphone Ansicht per Javascript und fügt dabei Elemente ein. Diese Elemente kennt dieser Generator natürlich nicht und kann sie folglich auch nicht beachten.

4.1.11 Http Archive

<http://httparchive.org/> ist ein Archiv der populärsten Seiten des Internets und bietet eine Vielzahl an statistischen Auswertungen, Trends und Daten.

„[HTTP Archive] is a permanent repository of web performance information such as size of pages, failed requests, and technologies utilized. This performance information allows us to see trends in how the Web is built and provides a common data set from which to conduct web performance research.“ (httpArchive 2015)

4.1.12 Perf Tooling Today

<http://perf-tooling.today/> ist wohl die Umfassendste Sammlung an web performance Tools und Material im Internet. Es hat eine Liste von 105 Tools, 51 Artikel, 27 Videos und 14 Slidedecks (Stand: 12.03.15).

4.1.13 Twitter

Twitter bietet die Möglichkeit am Puls der Zeit zu sein und unter dem Hashtag #webperf und #perfatters erhält man neuste Erkenntnisse, Tools oder Links, die sonst unentdeckt bleiben.

4.2 Ausgangspunkt

Im folgenden Abschnitt soll der Prozess beschrieben werden, um von einer langsamen Webanwendung zu einer schnellen zu gelangen. Von Beginn an war es wichtig, den Verbesserungsablauf zu Dokumentieren und in konkrete Daten zu fassen. Wie bereits unter Punkt 4.1.7 beschrieben, bietet Google Spreadsheets die Möglichkeit Skripte zu schreiben und die Ergebnisse direkt in eine Tabelle auszugeben. Diesen Umstand hat sich **Andy Davies** zu nutzen gemacht und ein Programm¹⁹ geschrieben (MIT License), dass es ermöglicht Webpagetest innerhalb einer Google Tabelle²⁰ aufzurufen. Damit wurde während der Entwicklungsphase täglich tests aufgezeichnet.²¹ Die Auswertung dieser Daten erfolgt in Punkt 6.

Da nur in Dulles VA eine Testinstanz mit richtigen Smartphones zur Verfügung steht, wurde mittels der **Microsoft Azure Cloud** die selbe Seite auch in den USA gehostet, um die Latenz zwischen USA und Europa zu eliminieren. Dadurch lässt sich exakter bestimmen, wie schnell ein Smartphone mit 3G Netz die Seite aufrufen kann. Leider steht keine Testinstanz mit 4G Netz zur Verfügung.

Als Ausgangspunkt dient die Seite <http://andreaslorer.de/old/>. Zu Beginn des Optimierungsprozesses gab es folgenden Ausgangspunkt (Daten via Developer Tool & webpagetest):

Desktop: ²²

- 42 requests: 30 Images, 5 JS, 3 CSS, 4 other
- 1000 kb Seitengröße
- Speed Index: **3584**
- Start Render: **1399** ms
- Load Time: 1926 ms
- TTFB: 690 ms

Mobile: ²³

- 17 requests: 4 Images, 5 JS, 3 CSS, 4 other
- 363 kb Seitengröße
- Speed Index: **10642**
- Start Render **6968** ms
- Load Time: 5587 ms
- TTFB: 1292 ms

¹⁹ WebPageTest Bulk Tester via GitHub: <https://github.com/andydavies/WPT-Bulk-Tester>

²⁰ Das Google Dokument ist hier zu finden: <http://tinyurl.com/nv4pge5>

²¹ Die gesamten Daten sind hier zu finden: <http://tinyurl.com/l5usz79>

²² Webpagetest: http://www.webpagetest.org/result/150312_Z1_18QD/

²³ Webpagetest: http://www.webpagetest.org/result/150308_A1_2W4/

Diese Werte sind nicht gut und für dieses Projekt wurden eine Start Render Zeit von weniger als einer Sekunde und ein Speed Index von unter 1000, für sowohl Mobile- als auch Desktopgeräte, angestrebt.

Der erste Schritt war es, die Seitengröße zu verringern. Aus diesem Grund wurde das Framework gewechselt und die Seite neu Aufgebaut. Bootstrap ist zwar ein sehr populäres Framework, hat aber gerade für kleine Seiten sehr viele Komponenten, die keine Verwendung finden (oft auch als Overhead bezeichnet). Bootstrap lässt sich zwar per „Customize“ Funktion so zusammenstellen, dass nur die Komponenten zur Verfügung gestellt werden, die für das eigene Projekt von Nöten sind, es ist aber dennoch ein Framework mit relativ großem Volumen (30 bis 90 kb). Die Alternativen zu Bootstrap sind vielzählig. Die Entscheidung für dieses Projekt fiel auf <http://purecss.io/>. Dieses Framework von Yahoo ist Komprimiert gerade einmal **4 kb** groß, vollkommen responsive und kommt mit den wichtigsten Komponenten wie einer Navigations Bar, Buttons, Tabellen, Menüs und Form Elementen. Je nach gewählten Komponenten, benötigt es kein Javascript und kein JQuery. Dadurch lassen sich weitere Kilobytes als auch Requests einsparen.

Da Bootstrap seine eigene Icon-Font liefert, musste hier eine Alternative gefunden werden. „Font Awesome“²⁴ bietet dabei eine der umfangreichsten Icon Sammlungen im Web an und ist unter der [Open Font License](#) komplett frei benutzbar (auch kommerziell). Font Awesome ist mit seinen 519 Icons allerdings nicht gerade ein Leichtgewicht und kann bis zu 100 kb groß sein. Da auf der Seite <http://andreaslorer.de> weniger als 20 Icons zum Einsatz kommen ist der Überschuss folglich enorm. Deshalb gibt es eine Webanwendung namens <http://fontello.com>. Damit lassen sich aus einer Vielzahl an Icons genau die wählen, die für die eigene Seite benötigt werden. Auch das Wählen aus verschiedenen Icon-Sammlungen ist möglich. Heruntergeladen wird anschließend eine ZIP-Datei. Das Resultat: Die neue Version der Seite benötigt nur noch 5.6 kb für die Icons. Verglichen mit: Bootstrap 43 kb, Font-Awesome 97 kb.

Als nächstes wird die Webanwendung mittels [Pagespeed Insight4.1.2](#) Analyisiert. Das Ergebnis liefert Anhaltspunkte, was für schnellere Ladezeit alles umgesetzt werden sollte. Im folgenden soll erläutert werden, was es alles an Verbesserungen gibt und wie eine mögliche Umsetzung in der Praxis aussieht.

4.3 Best Practices

4.3.1 Render Blocking Javascript

Bereits unter Punkt 3.6.2 ist das Blockierende Verhalten von Javascript und CSS angesprochen worden. Grundvoraussetzung für diesen Punkt ist, dass das Javascript der Webanwendung in ihre, für das Rendern kritische und für das Rendern unkritische Teile zerlegt wurde.

Der Browser stellt bereits von Haus aus zwei Attribute bereit, mit denen sich Skripte asynchron herunterladen lassen. Diese Attribute heißen „async“ und „defer“ und werden von jedem Browsertyp unterstützt. (caniuse.com 2015) Sie erlauben es, dass der Browser nicht auf das Herunterladen der Dateien warten muss, sondern mit dem Parsen des Dokuments fortfahren darf. Async wird direkt nach dem herunterladen ausgeführt und dafür muss das Parsen pausiert werden. Defer hingegen unterscheidet sich von async in zwei Punkten: 1. Das Skripte wird nach Ende des Parsens ausgeführt. 2. Mit defer verzögert geladene Skripte werden in genau der Reihenfolge ausgeführt, wie die Reihenfolge der Skripte im HTML Dokuments vorliegen.

Diese Methode erlaubt es, Skripte parallel herunterzuladen, ohne dass der Renderprozess warten muss. Was damit nicht erreicht werden kann ist, dass der Download so lange verzögert wird, bis

²⁴Font-Awesome: <http://fontawesome.github.io/Font-Awesome/>

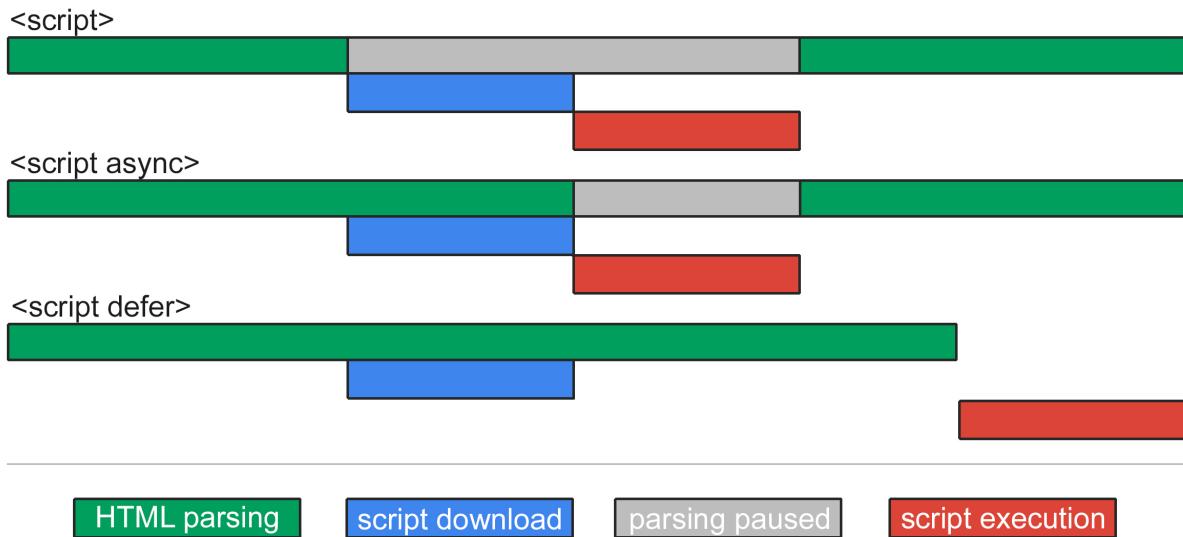


Abbildung 25: Script-Tags mit verschiedenen Attributen (Abbildung nach (growingwiththeweb.com 2014))

die für die Seiten primär wichtigen Ressourcen zuerst heruntergeladen wurden.

Mit Hilfe des Javascripts in Listing 4, kann die Datei „defer.js“ komplett mit dem Laden verzögert werden, bis der Ladeprozess der Seite abgeschlossen ist.

```

1  // the function to asynchronous load js
2  function loadJS( src, cb ){
3      "use strict";
4      var ref = window.document.getElementsByTagName( "script" )[ 0 ];
5      var script = window.document.createElement( "script" );
6      script.src = src;
7      script.async = true;
8      ref.parentNode.insertBefore( script, ref );
9      if (cb && typeof(cb) === "function") {
10          script.onload = cb;
11      }
12      return script;
13  }
14
15 // the function call to load your script
16 loadJS( "path/to/script.js" );

```

Listing 4: Javascript nach (Group 2015)

Dieses Skript hat einen Nachteil: Es kann nicht mehrere Skripte laden, die voneinander abhängen und deren Reihenfolge wichtig ist, um die Funktionalität zu gewährleisten.

„loadJS does nothing to manage execution order of requested scripts, so we do not advise using it to load multiple javascript files that depend on one another. It simply fetches a script and executes it as soon as possible. You can certainly use loadJS to load multiple scripts as long as those scripts are designed to execute independently of any other scripts being loaded by loadJS.“ (Group 2015)

So gibt es Skripte (Skript A) die von anderen Frameworks wie zum Beispiel JQuery (Skript B) abhängen. Das bedeutet, wenn Skript A schneller heruntergeladen und ausgeführt wird als Skript B, A bereits Funktionen von B aufruft, die noch nicht zur Verfügung stehen. Daraus resultiert ein Fehlschlagen des Skripts und somit können Teile der Webanwendung nicht mehr wie beabsichtigt funktionieren.

Darum gibt es Skripte die genau diese Funktionalität bereitstellen können. Skript A und B werden gleichzeitig heruntergeladen, A wird aber erst genau dann ausgeführt, wenn B zur Verfügung steht.

<http://headjs.com/> kann das erreichen. Durch das Herunterladen und Einfügen in das HTML Dokument, kann per Funktionsaufruf die Abhängigkeit festgelegt werden:

```

1 // Load up some script A and then script B
2 head.load("jquery.js", function() {
3     // Call a function when done
4     console.log("Done loading jquery");
5     head.load('defer.js')
6 });

```

Listing 5: Headjs dependency loading (Listing nach <http://headjs.com/>)

Headjs hat den Nachteil, dass es auch noch andere Funktionalitäten außer dem Laden von Javascript ermöglicht. Dies wird aber nicht benötigt und deshalb ist Headjs mit 2.1 kb doch zu groß, um es **Inline** in das HTML Dokument zu schreiben. Eine bessere Alternative ist jQ1.²⁵ Die Verwendung ist sehr simpel:

```

1 <script type="text/javascript">
2     // first include jQ1 inline (missing here) then call these functions
3     jQ1.loadjQ('jquery.js');
4     jQ1.loadjQdep('defer.js');
5 </script>

```

Listing 6: jQ1 asynchronous jQuery-Loader

Dieses Skript sagt im Grunde: Lade beide Dateien gleichzeitig herunter, beachte aber die Abhängigkeit (loadjQdep steht für: load dependency) von defer.js gegenüber JQuery. Ist defer.js früher heruntergeladen als JQuery, so wird gewartet und anschließend werden die Skripte in der richtigen Reihenfolge ausgeführt.

Für die Webseite <http://andreaslorer.de> wurde sowohl defer, als auch das Skript jQ1 verwendet. „<script defer src='critical.js'></script>“ wird dabei in den „<head>“ Bereich des HTML Dokuments platziert, damit es möglichst früh erkannt wird und der Download bereits beginnen kann und das Skript aus Listing 6 befindet sich vor dem „</body>“-Tag.

²⁵jQ1 an asynchronous jQuery Loader: <http://www.yterium.net/jQ1-an-asynchronous-jQuery-Loader>

4.3.2 Render Blocking CSS

Wie Javascript blockiert auch CSS das Rendern der Seite. In Listing ?? ist ein Skript der Filament Group zu sehen. Dieses Skript ermöglicht es CSS verzögert zu laden.

```

1   <script>
2     // minified script after:
3     // https://github.com/filamentgroup/loadCSS/blob/master/loadCSS.js
4     // [c]2014 @scottjehl, Filament Group, Inc.
5     // Licensed MIT
6     function loadCSS(e,a,g,h){function f(){for(var a,c=0;c<d.length;c++)d[c].href&&
7       -1<d[c].href.indexOf(e)&&(a!=0);a?b.media=g||"all":setTimeout(f)}
8     var b=window.document.createElement("link");a=a||
9     window.document.getElementsByName("script")[0];
10    var d=window.document.styleSheets;b.rel="stylesheet";b.href=e;b.media="only x";
11    b.onload=h||function(){a.parentNode.insertBefore(b,a);f();return b
12  };
13
14  loadCSS( "path/to/css" );
15 </script>
16
17  <!-- fallback if javascript is disabled in browser -->
18  <noscript><link href="path/to/css"></noscript>
```

Listing 7: load a CSS file asynchronously

Mehr als dieses Skript ist nicht notwendig.

4.3.3 Inline CSS

Kleinere Mengen an CSS lassen sich direkt **Inline** in das HTML Dokument einfügen. Dadurch sind diese gleich mit dem ersten Request bereits im Dokument enthalten und müssen nicht erst angefordert und heruntergeladen werden.

4.3.4 Ressourcen reduzieren

Das schnellste Byte ist das, dass nicht gesendet wird und der schnellste Request ist der, der nicht gestellt wird. Deshalb gibt es drei Maßnahmen die für eine Webanwendung umgesetzt werden sollte:

- „Minify“: Kommentare, Leerzeichen oder Zeilenumbrüche sind für die Funktionalität nicht notwendig. Bei der Verkleinerung des HTML- CSS oder Javascript Codes werden diese entfernt und dadurch die Dateigröße verkleinert. Wie unter Punkt 4.1.2 angesprochen, gibt es Pagespeed Insight auch als Chrome-Erweiterung. Damit ist es möglich, eine reduzierte Version des HTML Dokuments zu erzeugen. Für Javascript ist der Closure Compiler (4.1.3) das richtige Werkzeug. CSS lässt sich per <http://cssminifier.com/> verkleinern.
- „Uglify“: Dabei werden Variablennamen, auf nur wenige Zeichen reduziert. Aber auch Code wird teilweise umgeschrieben, wenn für einen verwendeten Ausdruck beispielsweise eine kürzere Schreibweise existiert.

```

1      // input:
2      if(foo){
3        bar();
4      }
5      else{
6        boo();
7      }
```

```

8
9      // output:
10     foo?bar():boo();

```

Listing 8: Beispiel: Uglify eines Ausdrucks

Input und Output sind identisch in ihrem Ausdruck, die Zeichenanzahl wurde aber von 27 auf 16 verringert.

- „Concatenate“: Damit ist das Zusammenfügen von einzelnen Dateien zu einer Einzigen gemeint. Dadurch lassen sich zusätzliche Requests einsparen. Dies hat den Vorteil, dass sowohl der TCP Slow start nur einmal eintritt als auch nur eine TCP Verbindung aufgebaut werden muss. Zusätzlich werden weniger TCP Verbindungen belegt, denn dafür gibt es, wie bereits erwähnt wurde (Punkt 2.3), ein Limit.

4.3.5 CSS-Bereitstellung optimieren

Wenn externe Ressourcen klein sind, können diese direkt in das HTML Dokument **Inline** platziert werden. Dabei sollte darauf geachtet werden, dass das HTML Dokument Komprimiert nicht die 14 kb Marke überschreitet. Dadurch kann es im ersten round trip geliefert werden. CSS Dateien die groß sind sollten per **Link-Tag** eingebunden werden und mittels Skript Verzögert geladen werden. Das CSS für den above the fold Bereich sollte **Inline** im „**<head>**“ Bereich der Seite stehen.

4.3.6 Antwortzeit des Servers reduzieren

Die Zeit zur Antwort des Servers lässt sich zum Beispiel mit Webpagtest herausfinden. Ein Server sollte auf eine Response Zeit von unter 200 ms kommen. „*Es gibt Dutzende potenzielle Faktoren, die die Antwortzeit Ihres Servers beeinträchtigen können: eine langsame Anwendungslogik, langsame Datenbankabfragen, langsames Routing, Frameworks, Bibliotheken, CPU-Ressourcenmangel oder Speicherplatzmangel. Berücksichtigen Sie zur Verkürzung der Antwortzeit Ihres Servers alle diese Faktoren.*“ (Google 2015). Bereits unter Punkt: 5.5 wurde es nahegelegt ein gutes Hosting zu wählen. Besser Sie wechseln ihr Hosting als ihre Kunden den Service.

4.3.7 Browser-Caching nutzen

Fehlendes Browser-Caching (das lokale Speichern von Daten) wird von Pagspeed Insight bemängelt, wenn der Server bei seiner Antwort keinen expliziten **Caching-Header** versendet. Durch das Speichern von statischen Ressourcen wie Javascript, Stylesheets und Bildern kann Zeit eingespart werden, wenn der Besucher die Webanwendung ein weiteres mal aufruft. Generell sollten alle statischen Ressourchen außer das HTML Dokument selbst, gecached werden.

Um auf dem Server (Apache) das Caching von statischen Ressourcen zu ermöglichen, ist ein Eintrag in die **htaccess** Datei des Servers nötig. Folgender Eintrag sollte dort platziert werden:

```

1      ## EXPIRES CACHING ##
2      <IfModule mod_expires.c>
3          ExpiresActive On
4          ExpiresByType image/jpg "access 1 year"
5          ExpiresByType image/jpeg "access 1 year"
6          ExpiresByType image/gif "access 1 year"
7          ExpiresByType image/png "access 1 year"
8          ExpiresByType text/css "access 1 year"
9          ExpiresByType text/woff "access 1 year"

```

```

10      ExpiresByType application/pdf "access 1 year"
11      ExpiresByType text/x-javascript "access 1 year"
12      ExpiresByType application/x-shockwave-flash "access 1 year"
13      ExpiresByType image/x-icon "access 1 year"
14      ExpiresDefault "access 1 month"
15      </IfModule>
16      \#\# EXPIRES CACHING \#\#

```

Listing 9: Aktivieren von Browser Caching in Apache (Listing nach: (Sexton 2015b))

Listing 9 hat 2 Aufgaben. Erstens: Es setzt die Ablaufzeit für alle statischen Ressourcen auf 1 Jahr und erfüllt damit den von Google empfohlenen Wert. Längere Speicherzeiten sind dagegen nicht Empfohlen, da dies gegen die RFC-Richtlinien verstößen würde (Google 2014a). Zweitens: Es wird mit dem HTTP Request ein Header mit gesendet. Dieser ermöglicht es dem Browser seine lokal gespeicherten Ressourcen zu managen. Er besteht aus folgenden Teilen und es ist jeweils nur **eine** der Optionen nötig.

- Last-Modified: date
- ETag: ID

Diese beiden Header ermöglichen es dem Browser zu überprüfen, ob sich die gecachten Ressourcen geändert haben oder noch identisch sind. Last-Modified ist dabei das Datum der letzten Änderung und der ETag-Header ist ein automatisch generierter Wert, der die Datei eindeutig Identifiziert.

Beim erneuten Laden einer Seite werden diese Header zurück an den Server gesendet und verglichen. Wenn die Datei auf dem Server geändert wurde stimmen die Werte nicht überein und der Server schickt eine entsprechende Antwort zurück.

- Cache-Control: max-age=value
- Expires: date

Mit diesen Headern ist es möglich Serveranfragen komplett zu vermeiden. „Sämtliche vom Browser ausgegebenen HTTP-Anfragen werden zuerst an den Browsecache weitergeleitet, um zu überprüfen, ob eine gültige Antwort im Cachespeicher vorliegt, die der Anfrage entspricht. Liegt eine Übereinstimmung vor, wird die Antwort aus dem Cache ausgelesen, wodurch sowohl die Netzwerklatenz als auch die durch die Übertragung anfallenden Datenkosten umgangen werden.“ (Grigorik 2014b) Das bedeutet, dass die Latenz bei Smartphones für gecachte Dateien komplett negiert werden kann. Gültige Ressourcen werden erst gar nicht angefragt, sondern gleich aus dem Cache geladen. Ungültige oder abgelaufene Ressourcen werden dagegen vom Server geholt. Ohne diesen Header, muss der Browser für jede in seinem Cache befindliche Ressource, den Server anfragen. Dafür sind jedes mal ein round trip nötig.



Abbildung 26: Gecachte Ressourcen müssen nicht mehr abgefragt werden

Das rot unterstrichene zeigt, dass 59 ms im Netzwerk verbraucht wurde (Kabel Verbindung). Der Server antwortete mit: „Not-Modified“. Grün zeigt, dass keinerlei Kommunikation mit dem

Server nötig ist sondern die Datei, in diesem Fall ein Bild, direkt aus dem Cache geholt wird. Fazit: Ohne Cache-Control lässt sich durch die Browser eigene Caching Funktionalität das erneute Herunterladen der Datei vermeiden. Mit Cache-Control kann sowohl das Herunterladen als auch der gesamte Verbindungsauftakt zum Server vermieden werden.

Was aber wenn sich zum Beispiel eine CSS-Datei geändert hat? Dann würden nun Besucher mit leerem Cache eine andere Darstellung erhalten, wie Besucher mit der gecachten Version. Dafür gibt es mehrere Lösungsansätze.

1. Die HTML Datei sollte nicht gecached werden da sonst Änderungen nicht mehr den Anwender erreichen können.
2. Eine für die Datei angemessene max-age: Dateien die sich oft ändern dürfen auch entsprechend niedrige max-age Werte haben. Dadurch wird die Datei Zeitnah für alle Anwender neu Angefordert.
3. Ressourcen können mit einer ID versehen werden: `styles.css` wird in `styles.v1.0.1.css` umbenannt.
4. Alternativ zur ID ist auch in `Fingerprint` möglich. Dabei wird eine Zahl aus der Datei generiert. Ändert sich die Datei so ändert sich auch der Fingerprint. Dieser Fingerprint wird auch wiederrum dem Dateiname angefügt. Das kann so aussehen: `styles.82s0dfa.css`.²⁶

Browser-Caching ist eine mächtige Funktionalität die sich jeder zu nutzen machen sollte. Sie ist zudem ganz einfach mit nur einem Eintrag in die `htaccess`-Datei realisierbar. Allerdings hat eine Studie von Yahoo ergeben, dass 40-60% der Besucher beim Seitenaufruf einen leeren Cache haben und rund 20% aller aufgerufenen Seiten wurden mit einem leeren Cache aufgerufen.

„[...] I don't know about you, but these results came to us as a big surprise. It says that even if your assets are optimized for maximum caching, there are a significant number of users that will always have an empty cache.“ (Yahoo 2007)

Folglich macht es Sinn, die Geschwindigkeit der Seite für die sogenannten „first users“ zu optimieren und nicht von einem gecachten Zustand der Seite auszugehen.

`Feed-The-Bot` stellt ein Tool²⁷ zu Verfügung, mit dessen Hilfe sich überprüfen lässt, ob die eigene Webseite „Browser-Caching“ richtig einsetzt. Abbildung Nummer 27 zeigt Links eine Seite mit aktiviertem Browser-Caching und Rechts eine Seite ohne.

²⁶Dieses Verfahren lässt sich auch automatisieren, ich verweise auf folgenden Artikel: <https://adactio.com/journal/8504>

²⁷Tool: <http://www.feedthebot.com/tools/if-modified/>



Abbildung 27: Beispiel: Eine Webseite mit und ohne „Cache Control“. (Eigene Abbildung nach feedthebot.com)

4.3.8 Komprimierung aktivieren

Nachdem durch das in Punkt 4.3.4 beschriebene Verfahren die Ressourcen soweit wie möglich verkleinert wurden, können diese vor dem Versenden komprimiert werden. Dies nennt man auch Datenkomprimierung und stellt ein ganz eigenes Forschungsgebiet dar. Deshalb soll nur der Grundgedanke erklärt werden.

Gegeben sei folgende Textnachricht:

```

1 # Lorem ipsum dolor sit amet, consectetur adipisicing elit. Debitis temporibus
   incidunt id.
2 # Cumque molestiae est praesentium magnam, fugit ipsa.
3 format: text/plain
4 date: 21.03.15
5 AAZZBBBBEEEMMM EEETTTAAAA

```

Diese Nachricht ist 200 Zeichen lang und kann durch einfache Regeln verkürzt werden. Zuerst werden die Kommentare, die mit # gekennzeichnet sind entfernt, denn sie sind für die Bedeutung der Nachricht nicht relevant. Das Datum könnte in eine ID konvertiert werden: 210315. Die Nutzdaten der Nachricht wird nach Wiederholungen durchsucht. Daraus ergibt sich dann: 3A2Z4B3E3M 3E3T4A (Grigorik 2014a). Die Neue Nachricht:

```

1 format: text/plain
2 date: 210315
3 3A2Z4B3E3M 3E3T4A

```

Die Zeichenanzahl wurde von 200 Zeichen auf 47 Zeichen reduziert. Das ist eine Reduktion von 76,5%!

Das gängigste Komprimierungsprogramm im Web ist GZIP und wird von allen modernen Browsern unterstützt. Es arbeitet nach dem „Deflate“-Algorithmus und Komprimiert Daten verlustlos.²⁸ Je länger eine Textdatei ist, umso drastischer kann sich die Komprimierung auswirken. GZIP ist meistens Standardmäßig aktiviert. Falls nicht kann GZIP mittels `htaccess` Eintrag

²⁸Eine ausführlichere Beschreibung über den GZIP Algorithmus und der text basierter Dokumentkomprimierung ist hier zu finden: <http://www.infinitepartitions.com/art001.html>. Für ein tiefere Verständnis bietet sich dieses Video von Google an: <http://tinyurl.com/mfxt5zt>

von Listing 10 in Apache aktiviert werden.²⁹ Einträge in die `htaccess`-Datei sollten nur dann eingesetzt werden, wenn man zum Beispiel ein Shared Hosting verwendet und dadurch keinen direkten Zugang zur Serverkonfiguration hat. Grund dafür ist, dass die Konfiguration von Apache mittels `htaccess` den Server verlangsamt.(Apache.org 2015)

```

1      ## gzip Compression if available
2      <ifModule mod_gzip.c>
3          mod_gzip_on Yes
4          mod_gzip_dechunk Yes
5          mod_gzip_item_include file \.(html?|txt|css|js|php|pl)$
6          mod_gzip_item_include handler ^cgi-script$ 
7          mod_gzip_item_include mime ^text/*
8          mod_gzip_item_include mime ^application/x-javascript.*
9          mod_gzip_item_exclude mime ^image/*
10         mod_gzip_item_exclude rspheader ^Content-Encoding:.*gzip.* 
11     </ifModule>
```

Listing 10: GZIP htaccess Eintrag

Mittels <http://www.feedthebot.com/tools/gzip/> lässt sich überprüfen, ob der eigene Server GZIP erlaubt.

Gzip compression test



Abbildung 28: Testen von GZIP anhand von andreaslorer.de (Eigene Abbildung nach: feedTheBot.com)

Diesem Tool sollte aber nicht einfach nur blind vertraut werden. Es gibt Ressourcen die müssen explizit als Komprimierbar gekennzeichnet werden. So sollte mittels dem „Chrome Developer Tool“ (oder vergleichbare Tools wie z.B. Firebug in Firefox) nachgesehen werden, ob alle Ressourcen auch tatsächlich mittels GZIP komprimiert werden. Abbildung 29 zeigt, dass die Datei: „Gallery.php“ trotz aktiviertem GZIP keine Komprimierung verwendete.

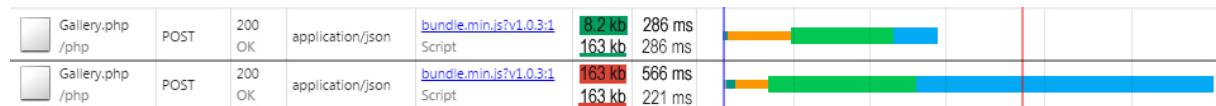


Abbildung 29: Die Serverantwort im JSON-Format ohne und mit GZIP

Der unterstrichene Wert ist dabei die Datengröße vor GZIP und der farblich hinterlegte Wert nach GZIP. Wie zu sehen ist konnte die zu übertragende Datengröße um fast das 20-Fache verringert werden! Entsprechend drastisch sank auch die Zeit für das herunterladen der Datei. Wie konnte das erreicht werden? Um GZIP für json-Dateiformate zu ermöglichen ist nur dieser Eintrag in der PHP-Datei (hier in „Gallery.php“) nötig:

²⁹Für andere Server wie zum Beispiel Nginx oder Node.js gibt es auf GitHub fertige Konfigurationen <https://github.com/h5bp/server-configs>

```

1     header('Content-Type: text/javascript');
2     header('Accept-Encoding: gzip');
3     ob_start('ob_gzhandler');
```

Manchmal sind es sehr kleine Dinge die eine überaus große Wirkung haben können und die oftmals übersehen werden.

4.3.9 „Keep-alive“ ermöglichen

Eine weitere Einstellung die in Apache vorgenommen werden kann ist „keep-live“ und hat folgende Bedeutung:

*„Webpages are often a collection of many files and if a new connection (the brief communication) has to be opened for each and everyone of those files it could take significantly longer to display that webpage.
More officially the definition of HTTP keep-alive would be something like: a method to allow the same tcp connection for HTTP conversation instead of opening new one with each new request.“(Sexton 2015a)*

```

1     ## keep-alive
2     <ifModule mod_headers.c>
3         Header set Connection keep alive
4     </ifModule>
```

Listing 11: htaccess Eintrag nach (Sexton 2015a)

Dieser Eintrag in der htaccess Datei fügt den `keep alive header` bei Serveranfragen hinzu.³⁰ Mittels Webpagetest kann überprüft werden, ob keep-alive auf dem Server aktiviert ist.

4.3.10 HTML5 Link Prefetching

Mit HTML5 gibt es ein neues Link-Attribut namens „prefetch“. Damit lassen sich Ressourcen herunterladen, die auf der aktuellen Seite noch nicht benötigt werden. Zum Beispiel in einem Bestellprozess lässt sich bereits bei Schritt 1 sagen, was auf der Seite von Schritt 2 benötigt wird. Dadurch lässt sich die Latenz und die Zeit zum herunterladen in den Hintergrund verlagern, ohne das der Anwender davon etwas mitbekommt. Natürlich ist dies mit bedacht einzusetzen und es macht wenig Sinn große Mengen an Daten zu laden, die der Anwender vielleicht gar nie verwendet. `Prefetch` ist ganz einfach über den `Link-Tag` zu verwenden:

```

1     <!-- load a single ressource -->
2     <link rel="prefetch" href="http://your-ressource-comes-here.jpg" />
3     <!-- or a full page -->
4     <link rel="prefetch" href="http://your-site/your-sub-site" />
```

Listing 12: Prefetching via Link-Tag

³⁰Bei shared Hostings kann es trotz dieser Einstellung oft nicht erreicht werden, keep-alive zu aktivieren.

4.4 Bilder optimieren

„Ein Bild sagt mehr als Tausend Worte.“ Diesen Spruch gibt es nicht umsonst und auch im modernen Web haben immer mehr und immer größere Bilder Einzug gehalten. Sie werden eingesetzt um eine Botschaft zu übermitteln, Emotionen zu erzeugen oder als **Eyeatcher**. In Abbildung 30 ist die durchschnittliche Seitengröße der Top 1000 Seiten³¹ des Webs abgebildet. So beträgt die durchschnittliche Seitengröße (rechts) zum Zeitpunkt März 2015 rund 1843 kb. Davon sind 65% (1112 kb) Bildmaterial. Das linke Diagramm zeigt einen Aufwärtstrend der Seitengröße in Kilobyte über die Zeitspanne von einem Jahr. Dabei repräsentiert der Graph in Lila die schlechtesten 90% der Internetseiten, Grün die 10% der besten Seiten und Gelb stellt den Median (Mittelwert) dar. Wie zu sehen ist werden die schlechtesten Seiten weiterhin schlechter, indem sie weiter an Kilobytes zulegen, auch der Median und die besten 10% sind über das Jahr hinweg leicht angestiegen. Mit dem optimieren von Bildern können sehr viele Kilobytes gespart werden.

„Most sites fail to leverage best practices for optimizing images.

Despite the fact that images represent one of the single greatest performance challenges (and opportunities), 34% of pages failed to properly implement image compression, and 76% failed to take advantage of progressive image rendering.“ (Radware 2014, p. 4)

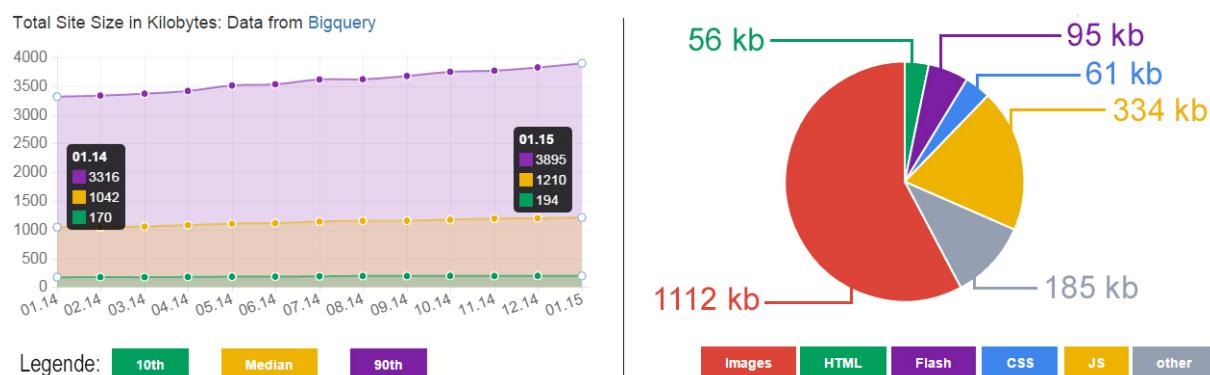


Abbildung 30: Seitenanalyse der populärsten Seiten des Webs (Eigene Abbildung - Daten via bigqueri.es: Gesamte Auswertung <http://tinyurl.com/o8vawxd>)

4.4.1 Progressive Image Rendering

Eigentlich eine sehr alte Technik ist das „Progressive Image Rendering“. Es kam längere Zeit außer Mode Bilder als Progressive zu speichern. Der Trend hat sich aber wieder geändert und Progressive Images gilt als „Best Practice“ im Web. Sowohl Webpagetest als auch Google schlagen vor, Bilder Progressive an den Anwender auszuliefern. Testet man eine Seite via Webpagetest.org, so bekommt man unter dem Reiter „Compress Images“ beispielsweise solch eine Auswertung:

Use Progressive JPEGs: 100/100
124.8 KB of a possible 124.8 KB (100%) were from progressive JPEG images

Abbildung 31 zeigt einmal den normalen Ladevorgang eines Bildes (unten) gegenüber dem „Progressive Rendering“ (oben).

³¹Top 1000 Seiten nach dem Ranking von Alexa: (Alexa gehört zu Amazon und liefert umfassende Analysen über das Web <http://www.alexa.com/>)



Abbildung 31: Progressives Bilder laden

Nicht nur sind Progressive Images fast immer ein paar Kilobyte kleiner, sie erhöhen auch die **Perceived Performance** für den Anwender. Wo bei normalen Bildern lange Zeit eine weiße Lücke klafft, ist bei Progressive Images bereits sehr viel früher das volle Bild zu sehen. Das PNG Äquivalent für Progressive-Jpeg's ist „Interlaced“. Bilder sollten zum Beispiel in Photoshop über den Reiter: Datei -> für Web speichern... entweder Progressive oder als Interlaced gespeichert werden.

4.4.2 Image Spriting

„Image Spriting“ bezeichnet das Kombinieren von vielen kleinen Bilddateien zu einer einzigen großen Datei. Mittels CSS lässt sich aus dem Bild dann das gewünschte Icon auswählen.



Abbildung 32: Image Sprite von verschiedenen Icons

Der Vorteil dieser Methode ist, dass nur ein Request benötigt wird um alle Icons, die für die Seite benötigt werden, herunterzuladen. Der Nachteil ist, dass durch die größere Datei der Download länger dauern kann und sich so die Zeit für das erste Rendern verringert. Auch bei einer einzigen Änderung, muss die ganze Datei (die im besten Fall bereits gecached wurde) neu heruntergeladen werden.

4.4.3 Bild Komprimierung

Bei der Komprimierung von Bildern unterscheidet man zwischen zwei Arten: „Lossless“ und „Lossy“.

- Lossless bezeichnet die Komprimierung eines Bildes, ohne dabei die Qualität des Bildes zu verändern. So werden bei Lossless zum Beispiel die Metadaten eines Bildes entfernt, die von einer Kamera bei der Aufnahme hinzugefügt werden.

- Lossy ist die reduzierung der Bildgröße auf kosten von Qualität. Für die meisten Bilder ist diese Reduzierung aber durchaus Legetim und oftmals sind die Einbußen mit bloßem Auge nicht zu erkennen. Lossy Komprimierung kann aber durchaus bis zu rund 40% der Bildgröße reduzieren.

Google hat ein eigenes Bildformat namens „Webp“ entwickelt und soll bis zu 30% der Bildgröße bei gleichbleibender Qualität einsparen. Allerdings ist dieses Format nur im Chrome Browser und Opera (der auf Chrome basiert) unterstützt.(caniuse.com 2015)

Eine bessere Alternative ist moz-jpeg. Das ist ein von Mozilla entwickelter JPEG Bild Encoder um die Bildgröße ohne Qualitätseinbußen zu verringern. Dabei bleibt das .jpg Format erhalten und ist somit überall einsetzbar.

„For the short term Mozilla has developed MozJPEG — a modernized JPEG encoder that offers better compression while remaining fully standard-compliant, so it's compatible with all browsers, operating systems and native apps, and you can use it today without waiting for the whole world to upgrade“ (Lesiński 2014)

Allerdings ist die Verwendung nicht ganz so einfach wie in diesem Zitat dargestellt und es ist das eigenständige Compilieren von C-Code³² nötig, um dies auf dem eigenen Betriebssystem zu verwenden.

Glücklicherweise es gibt eine Webanwendung³³ mit der ganz einfach per „drag and drop“ Bilder mittels diesem Encoder komprimiert werden können. Je nach Bild lassen sich so mehrere Hundert Kilobyte einsparen (abhängig von der Qualitätseinstellung und Größe des Bildes).

Natürlich bietet auch Photoshop oder IrfanView (Kostenlos) umfangreiche Möglichkeiten die Größe von Bildern zu verringern. Moz-jpeg schafft dies allerdings mit einem qualitativ besseren Ergebnis bei zudem kleinerer Bildgröße. Eine Verwendung sollte auf jedenfall in betracht gezogen werden.

4.4.4 Responsive Images

„Responsive Images“ ist eine auf das Gerät des Endanwenders angepasste Auslieferung von Bildern. Auf vielen Webanwendungen sieht man folgendes:



Abbildung 33: Downsampling auf GitHub (Eigene Abbildung via Chrome Developer Tool)

Das Bild wurde nicht nur viel zu Groß gewählt, es wird auch auf jedem Gerät, egal ob Tablet, Smartphone oder Desktop die selbe Anzahl an Bytes über die Leitung gesendet. Diese Methode

³²Das Repository ist hier zu finden: <https://github.com/mozilla/mozjpeg> und eine Anleitung gibt es hier: <http://calendar.perfplanet.com/2014/mozjpeg-3-0/>

³³Webanwendung zur Verwendung von moz jpeg: <https://imageoptim.com/mozjpeg>

nennt man Downsampling. Das bedeutet, dass nicht das Bild selbst verkleinert wird, sondern der Browser das Bild herunterlädt und dann auf die entsprechende Größe skaliert. Das kostet nicht nur Rechenleistung sondern vor allem sehr viel unnötige Bandbreite und sollte unter allen Umständen vermieden werden.

Seit HTML 5 gibt es ein neues HTML Attribut namens `srcset`. Dieses Attribut ist leider noch nicht in allen Browsern implementiert, hat aber immerhin bereits eine Globale Abdeckung von rund 50% (caniuse.com 2015). Für dieses Problem gibt es allerdings Abhilfe. Es gibt ein Polyfill³⁴ namens „Picturefill“³⁵, dass genau diese Funktionalität für alle Browser zur Verfügung stellen kann. Nötig ist dazu nur das herunterladen und Einbinden des Skripts in das HTML Dokument. Dadurch wird folgendes Listing möglich:

```

1   <picture id="hero-image">
2     <source srcset="someImg_lg.jpg" media="(min-width: 1367px)">
3     <source srcset="someImg_md.jpg" media="(min-width: 768px)">
4     <source srcset="someImg_xs.jpg" media="(min-width: 300px)">
5     <img srcset="fallback_md.jpg" alt="Some alt text">
6   </picture>
```

Listing 13: Srcset in Verwendung

Das bedeutet, dass Smartphones mit einer Breite von $> 300\text{px} < 768\text{px}$ das Bild „someImg_xs“ bekommen. Tablets mit $768\text{px} < 1367\text{px}$ bekommen das mittlere Bild und alles was über 1367px ist bekommen die größte Version zu Verfügung gestellt. Falls der Anwender Javascript im Browser deaktiviert, ist es möglich ein `fallback`-Bild zu setzen (Listing: Zeile 8). Man kann sogar einen Schritt weiter gehen und auch das Bildformat auf entsprechend dem Aufrufenden Gerät anpassen:

```

1   <picture id="hero-image">
2     <source srcset="someImg_lg.webp" type="image/webp" media="(min-width: 1367px)">
3     <source srcset="someImg_lg.jpg" media="(min-width: 1367px)">
4     <source srcset="someImg_md.webp" type="image/webp" media="(min-width: 768px)">
5     <source srcset="someImg_md.jpg" media="(min-width: 768px)">
6     <source srcset="someImg_xs.webp" type="image/webp" media="(min-width: 300px)">
7     <source srcset="someImg_xs.jpg" media="(min-width: 300px)">
8     <img srcset="fallback_md.jpg" alt="Some alt text">
9   </picture>
```

Listing 14: Srcset mit webp

Wie zu sehen ist wurden 3 Zeilen eingefügt, die sich lediglich im Typ unterscheiden. Das bedeutet, dass Anwender mit Chrome Browser das Chrome eigene Bildformat .webp bekommen und alle anderen das normale .jpg Format. Voraussetzung ist natürlich, all diese Bilder in ihren verschiedenen Auflösungen und Formaten anzulegen und bereit zu stellen, was einen Mehraufwand bedeutet. Picturefill ermöglicht es, Downsampling zu vermeiden und einen auf das Gerät angepasste Version des Bildes auszuliefern. Dadurch sinkt die Anzahl von Bytes die Übertragen werden müssen enorm.

4.4.5 Adaptive Images

Adaptive Images ist ein PHP-Skript³⁶, das mit Hilfe einer htaccess Datei Bilder auf dem Server automatisch auf die verschiedenen Gerätegrößen zuschneidet. Ruft ein Anwender die Seite auf,

³⁴ „Ein Polyfill [...] ist ein - meist in Javascript geschriebener - Code-Baustein, der in älteren Browsern eine neuere, von diesen nicht unterstützte Funktion mittels eines Workarounds nachrüsten soll. Beispielsweise sind Features von HTML5 in älteren Browsern nicht verfügbar. Webseiten können diese Funktionen trotzdem verwenden, wenn sie ein entsprechendes Polyfill mitliefern.“ (Wikipedia 2015)

³⁵ Das offizielle Picturefill Projekt ist hier zu finden: <http://scottjehl.github.io/picturefill/>

³⁶ Mehr über dieses Skript ist hier zu finden: <http://adaptive-images.com/>

prüft das Skript die Größe des Bildschirms und liefer anschließend das für ihn passende Bild aus.

4.4.6 Verzögertes Laden von Bildern

4.5 Zusammengefasst

5 Workflow

Im folgenden Abschnitt soll erläutert werden wie ein moderner Workflow aussehen kann. Dabei sollen viele Aufgaben, die von vielen noch per Hand erfolgen, automatisiert werden.

5.1 Nodejs

Nodejs - ist eine auf Chromes Javascript runtime aufbauende Plattform. Während Server vor allem in PHP oder anderen Sprachen programmiert werden ist dies durch Nodejs auch mit Javascript möglich. Nodejs liefert einen eigenen Paket Manager namens „npm“.

5.2 Node Package Manager

Auch als „npm“ abgekürzt, erlaubt das installieren von Paketen mittels der Kommandozeile. Ein Beispiel:

Das Paket „tmi - too many images“ analysiert eine gegebene URL nach ihrer totalen Bildgröße und vergleicht es mit der durchschnittlichen Größe des Webs. Es lässt sich mittels npm-Befehl über die Kommandozeile installieren: Danach lässt es sich ganz einfach per Befehl aufrufen:

```
C:\Users\Andi\Documents\workspace\bachelorthesis (master)
λ npm install -g tmi

C:\Users\Andi\Documents\workspace\bachelorthesis (master)
λ tmi andreaslorer.de
Your image weight
36.15 kB
Median mobile site image weight
17 kB
Median desktop site image weight
47 kB
On Mobile
-78 kB compared to sites in the 25th percentile
-403 kB compared to sites in the 50th percentile
-958 kB compared to sites in the 75th percentile
On Desktop
-233 kB compared to sites in the 25th percentile
-817 kB compared to sites in the 50th percentile
-1.67 MB compared to sites in the 75th percentile
Thanks for keeping the web fast <3
```

Es wäre aber auch möglich das unter Punkt 4.1.2 vorgestellte Google Pagespeed Insight mittels „npm install -g psi“ zu installieren und per „psi someURL.de“ aufzurufen. Es gibt bereits mehr als 134,082 Pakete (Stand 21.03.2015) die verschiedenste Aufgaben erledigen können. Das Spektrum an Paketen ist sehr groß und reicht von sinnlosen Aufgaben (wie dem zufälligen ge-

nerieren von Katzennamen³⁷⁾ bis hin zu hoch komplexen Programmen wie „sitespeed.io“³⁸, dass eine rekursive Performance-Analyse einer ganzen Webanwendung ermöglicht.

5.2.1 Dependency Management

Der klassische Weg wie externe Abhängigkeiten des Projekts geregelt werden sieht wie folgt aus:

- Zu Beginn des Projekts werden Bibliotheken und Frameworks heruntergeladen.
- Es folgt das Entpacken und das Verschieben in das richtige Projektverzeichnis.
- Erscheint beispielsweise eine neue Version des Frameworks beginnt dieser Prozess von neuem.

Dependency Manager (wie z.B. npm oder Bower^{5.2.2}) schaffen hier Abhilfe. Durch Ausführen des Befehls: „npm init“ lässt sich eine neue Abhängigkeitsstruktur für ein Projekt anlegen. Dabei wird eine Datei mit dem Namen: „package.json“ erzeugt. In dieser Datei werden nun sowohl die Beschreibung, die Version als auch alle Abhängigkeiten gespeichert. Will man nun beispielsweise das Programm „Gulp“ ?? installieren, erfolgt dies ganz einfach über das Kommando: „npm install -save gulp“. Die **-save** Option bedeutet dabei, dass folgender Eintrag in die package.json Datei erfolgen soll:

```
1 {
2   "dependencies": {
3     "gulp": "^3.8.11"
4   }
5 }
```

^{3.8.11} bedeutet dabei, dass für dieses Projekt mindestens eine Gulp Version größer als 3.8.11 vorliegen muss. Der große Vorteil besteht nicht nur darin, dass weder ein Seitenaufruf erfolgt, noch die Dateien entpackt und verschoben werden müssen. Zudem lässt sich über den Befehl „npm update“ alle Pakete auf die neueste Version bringen. Die package.json Datei lässt sich zudem in die Versionskontrolle einfügen. Der von npm angelegte Ordner: „node_modules“ sollte unbedingt per **.gitignore** von der Versionierung ausgeschlossen werden! Läßt ein Teammitglied das Repository herunter, so muss nur noch „npm install“ aufgerufen werden und alle Abhängigkeiten, mit den für dieses Projekt verwendeten Versionsnummern werden heruntergeladen und installiert. Damit ist jedes Teammitglied auf dem selben Stand und verwenden die selbe Version. Neue Abhängigkeiten lassen sich so auch ganz einfach an alle Mitglieder verteilen.

5.2.2 Bower

Wie npm so ist auch Bower ein Paket Manager. Um genau zu sein basiert Bower auf npm und lässt sich mittels „npm install -save bower“ für das Projekt installieren. Der Vorteil von Bower besteht darin, dass über eine **.bowerrc** Datei angeben werden kann, in welchem Ordner die Pakete installiert werden sollen. Dies ist bei npm nicht möglich.

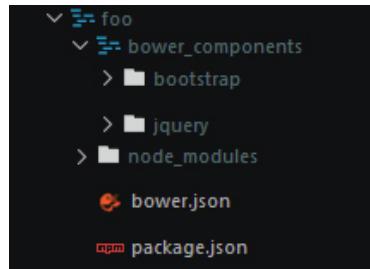
Es lohnt sich Frameworks und Libraries wie z.B. **Bootstrap** oder **Underscore** mittels Bower zu installieren und andere Programme wie zum Beispiel Gulp oder Nodejs Module per npm.

Bower lässt sich, ähnlich wie npm, über das Kommando: „bower init“ Initialisieren. Danach lassen sich per Befehl „bower install -save bootstrap“ das Bootstrap Framework installieren.

Wie zu sehen ist, wurde nicht nur Bootstrap von Bower installiert, sondern auch JQuery. Das liegt daran, dass Bootstrap als interne Abhängigkeit wiederum auch Abhängigkeiten besitzt,

³⁷Cat-names: <https://www.npmjs.com/package/cat-names>

³⁸Sitespeed.io: <http://www.sitespeed.io/>



die bei der Installation von Bootstrap gleich mit installiert werden. Entfernt man Bootstrap nun wieder, so würde auch JQuery entfernt werden. Würde Bootstrap in einer neuen Version erscheinen, die von einer höheren JQuery Version abhängt, so würde Bower automatisch auch JQuery auf die nötige Version aktualisieren.

5.3 Gulp Task Manager

Warum und für was braucht es überhaupt einen Task Manager?

*If you aren't using productivity tools or task automation, you are working **too hard**.
[...] Automation isn't about being lazy. It's about being **efficient**.* (Osmani 2014b, p. 18,78)

Ein Task Manager übernimmt immer wiederkehrende Arbeiten. Dazu zählt zum Beispiel die Aufgabe „minify, uglify und concatenating“ wie in Punkte: 4.3.4 bereits beschrieben. Aber auch das Übersetzen von „Sass“³⁹, oder das optimieren und verkleinern von Bildern lassen sich als Task beschreiben und automatisieren. Die zwei bekanntesten Task Manager heißen **Gulp** und **Grunt**. Hier soll das Arbeiten mittels Gulp beschrieben werden, denn er ist sehr viel einfacher zu benutzen als Grunt.

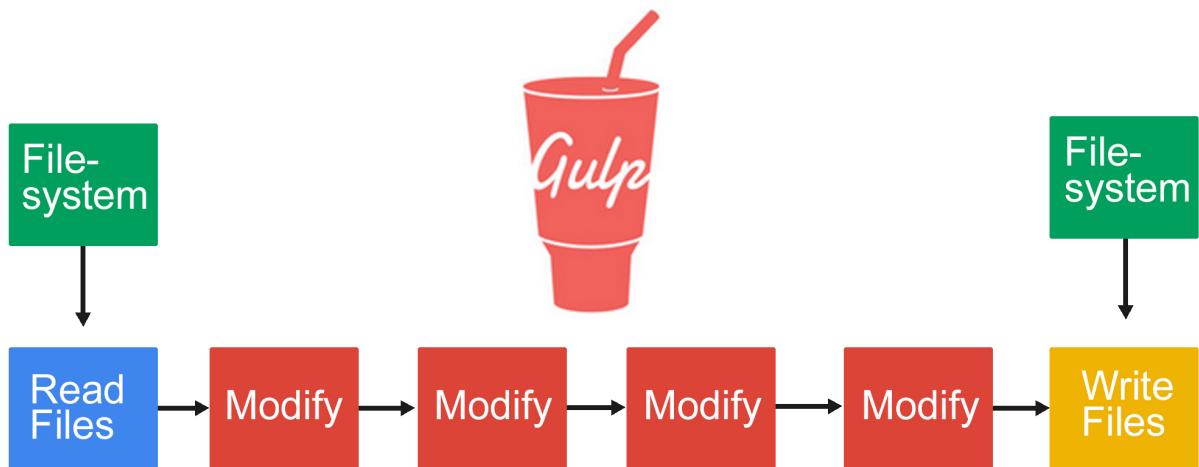


Abbildung 34: Gulp arbeitet nach dem „file stream“ Prinzip (Eigene Abbildung nach (Osmani 2014b, p. 85))

Gulp besteht aus nur 4 Kommandos:

³⁹Sass is the most mature, stable, and powerful professional grade CSS extension language in the world. <http://sass-lang.com>

1. gulp.task: Definiert einen Namen für einen Gulp Task. Dadurch wird dieser Task über die Kommandozeile benutzbar. Für Anwender die wenig mit der Kommandozeile arbeiten wollen gibt es auch eine Google Chrome Browsererweiterung namens Gulp-devtools. Damit lassen sich die verschiedenen Tasks ganz einfach über eine Benutzeroberfläche verwenden.
2. gulp.src: Greift auf eine oder mehrere Dateien zu. Auch Ordner können angegeben werden.
3. gulp.dest: Schreibt die Datei in den angegebenen Ort.
4. pipe: Mittels „pipe“ lassen sich die Dateien (auch „file streams“ genannt) modifizieren. Dabei lässt sich pipe beliebig oft aufrufen. Die Aufgabe: „minify, uglify und concatenating“ könnte dabei so erfolgen:

```

1 // require the gulp modules needed:
2 var gulp = require('gulp');
3 var uglify = require('gulp-uglify');
4 var concat = require('gulp-concat');
5
6 // javascript task: concat, minify, uglify all javascript files in folder:
7 gulp.task('javascript', function () {
8     gulp.src(['site/assets/libs/*.js', 'site/assets/js/*.js'])
9         .pipe(concat('bundle.min.js'))
10        .pipe(uglify())
11        .pipe(gulp.dest('dist/assets/js/'));
12 });

```

Dieser Task mit dem Namen „javascript“ holt nun alle Javascript Dateien aus dem Ordner „libs“ und „js“, fügt diese zu einer einzigen Datei zusammen und benennt sie „bundle.min.js“. Danach erfolgt das „uglify“ (beinhaltet das „minify“). Zum Schluss wird die Datei in den Ordner „dist/assets/js“ geschrieben. Die original Dateien werden dabei nicht modifiziert. Es sind auch so genannte „watch“ Tasks möglich, die bei einer Dateiänderung ausgeführt werden können. So kann der Browser immer dann neu geladen werden, wenn sich eine CSS Datei geändert hat. Dadurch wird ein manuelles neu Laden, um die Änderungen zu betrachten, hinfällig.

Grunt hat gegenüber Gulp den Vorteil, dass es älter ist als sein Konkurrent. Dadurch gibt es viele Pakete, die nur mittels Grunt zur Verfügung stehen. Beispielsweise das Paket „grunt-responsive-images“. Damit lassen sich aus einem gegebenen Bild, automatisch verschiedene Bildgrößen heraurechnen und abspeichern. Dies ist besonders hilfreich, wenn man „responsive images“ wie in Punkt: 4.4.4 beschrieben, verwenden möchte. Einen Blick auf Grunt kann sich also druchaus lohnen.

5.4 Yeoman

Ein Projekt wird oftmals Manuell oder mittels eines Programms, wie zum Beispiel PhpStorm oder Visual Studio erstellt. Es wird ein passender Projekttyp gewählt und das Programm erzeugt dann eine fertige Ordnerstruktur mit den wichtigsten Dateien.

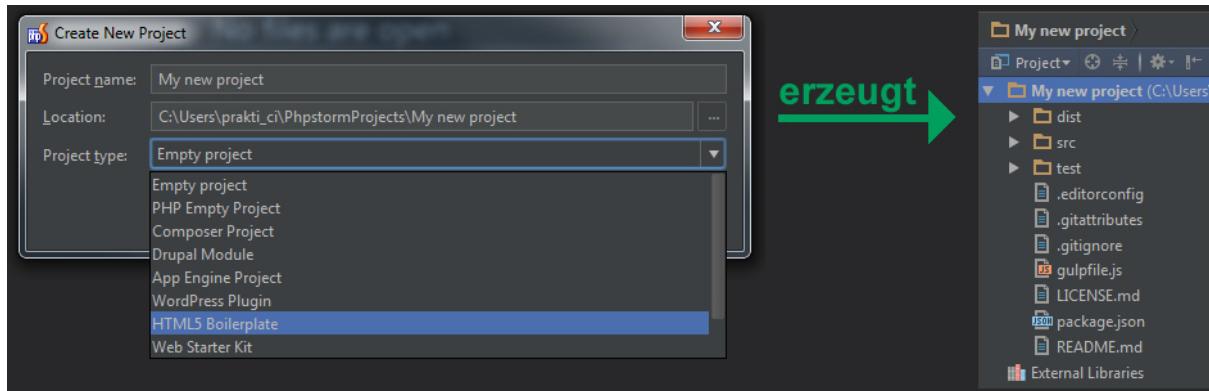


Abbildung 35: Projekt erstellen via PhpStorm (Eigene Abbildung)

<http://yeoman.io/> ist ein Tool für das schnelle Erstellen eines Grundgerüsts für ein Projekt.

„Yeoman helps you to kickstart new projects, prescribing best practices and tools to help you stay productive.

To do so, we provide a generator ecosystem. A generator is basically a plugin that can be run with the ‘yo’ command to scaffold complete projects or useful parts.

Through our official Generators, we promote the „Yeoman workflow“. [...] We take care of providing everything needed to get started without any of the normal headaches associated with a manual setup.“ (<http://yeoman.io>)

Es stellt sogenannte „Generators“ zur Verfügung, die genau diese Funktionalität der Projekterstellung abbildet. Dabei wird das Ziel verfolgt, „Best Practices“ für das jeweilige Projekt umzusetzen. Der Vorteil von Yeoman besteht darin, dass es eine enorme Vielzahl an Generatoren zur Verfügung stellt. So gibt es Generatoren für die Erstellung von Web-Apps, Chrome Browser Erweiterungen, Wordpress Blogs, Firefox OS Apps und noch 1538⁴⁰ weitere (Stand 23.03.15). Es steht als npm Paket zur Verfügung und lässt sich mittels „npm install -g yo“ mit Hilfe der Kommandozeile installieren. Via „npm install -g generator-someName“ lassen sich dann beliebige Generatoren dazu installieren. Yeoman bringt dabei die zuvor vorgestellten Tools wie Gulp / Grunt und Bower zusammen. Durch das Ausführen wird das gesamte Projekt angelegt:

- Es werden alle npm Pakete (Bower, Gulp usw.) die nötig sind automatisch installiert.
- Die Ordnerstruktur mit einer CSS und Javascript Datei wird angelegt und im HTML Dokument eingebunden.
- Nach der Installation stehen die wichtigsten Gulp Tasks zur Verfügung, ohne dass man sie selber schreiben muss.
- Durch Gulp besteht die Möglichkeit einen Webserver mittels Kommando „gulp serve“ zu starten.

⁴⁰Eine volle Liste an verfügbaren Generatoren ist hier zu finden: <http://yeoman.io/generators/>

- Es bestehen zusätzliche Funktionalitäten, wie das automatische Aktualisieren des Browsers nach einer Änderung, zur Verfügung.
- Mittels Eingabeaufforderung lässt sich auswählen, welche Module (Sass, Bootstrap, Modernizr) enthalten sein sollen und welche nicht.

Das Ausführen von:

```
C:\Users\prakti_ci\Documents\myProject
λ yo gulp-webapp
[---]
[---(o)---]
[---U---]
[---A---]
[---~---]
[---.---]
[---|---]
[---o---]
[---Y---]

'Allo 'allo! Out of the
box I include HTML5
Boilerplate, jQuery, and
a gulpfile.js to build
your app.

? What more would you like?
>(•) Sass
(•) Bootstrap
( ) Modernizr
```

Abbildung 36: Starten eines neuen Projekts mittels generator-gulp-webapp (Eigene Abbildung)

würde innerhalb von wenigen Minuten ein Projekt anlegen, dass bereits fertig mit Bootstrap, JQuery und einem HTML Dokument konfiguriert ist. Mittels „gulp serve“ kann das direkt im Browser betrachtet werden:

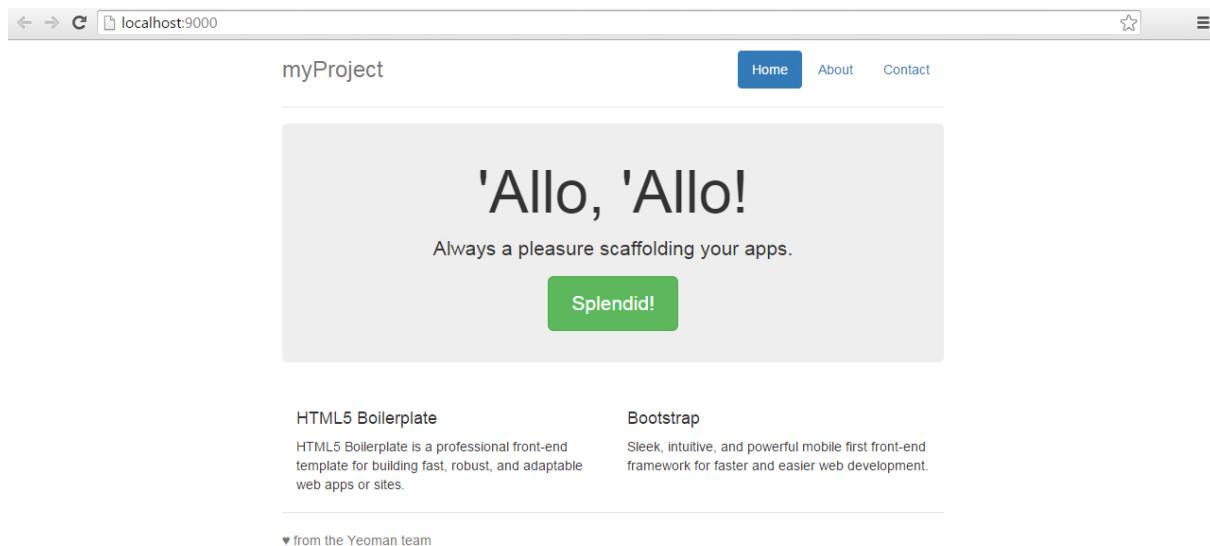


Abbildung 37: Ergebnis des Kommandos: „yo gulp-webapp“ (Eigene Abbildung)

5.4.1 Eigene Generatoren erstellen

Yeoman bietet jedem die Möglichkeit einen eigenen Generator zu erstellen. Auch dafür gibt es wiederum einen Generator, den „generator-generator“. Damit lässt sich eine persönliche Projektstruktur erstellen, die den eigenen Projekten gerecht wird. Dadurch lassen sich beispielsweise Skripte zum verzögerten Laden von Javascript, damit das Rendern nicht blockiert wird, von

Beginn an in das eigene Projekt integrieren. Eigene Generatoren können ganz einfach Veröffentlicht werden. Es muss lediglich ein Benutzer mittels „npm adduser“ angelegt werden und dann kann mittels „npm publish“ der eigene Generator als **npm** Paket zur Verfügung gestellt werden. Auch Updates lassen sich mittels „npm publish“ binnen Sekunden einspielen. Mein persönlicher Generator ist zum Beispiel unter: <https://www.npmjs.com/package/generator-4dev> zu finden und kann von jedem über „npm install -g generator-4dev“ installiert und anschließend mittels „yo 4dev“ aufgerufen werden.

Für ein Unternehmen könnte dies zum Beispiel bedeuten, mit einem eigenen Generator eine einheitliche Projektstruktur einzuführen, der die „best practices“ des Unternehmens abbildet. Alternativ könnte auch ein Konsens gefunden werden, welcher der bereits existierenden Generatoren für die jeweils eigenen Projekte einsetzbar wäre.

Eine ausführliche Anleitung und Dokumentation zur Erstellung von Generatoren ist der offiziellen Seite: <http://yeoman.io/authoring/index.html> zu entnehmen.

5.5 Zusammengefasst

Gulp, Bower und Yeoman arbeiten perfekt zusammen und ermöglichen einen modernen Workflow mit den momentan gültigen „best practices“. Mittels **npm** und **Bower** lassen sich Pakete ganz einfach installieren, deinstallieren und updaten, während gleichzeitig eine einheitliche Entwicklungsumgebung bereitgestellt wird. Außerdem bieten viele Frameworks die Möglichkeit, mittels Bower nur einzelne Module zu installieren. Dadurch ist es möglich nur das Nötigste in die Seite einzubinden.

Gulp ermöglicht eine umfassende Automatisierung von Aufgaben, die gerade für die web performance sehr wichtig sind. Folgende Projektstruktur erweist sich als sinnvoll:

```

1   .
2   |_ site/
3   | |_ bower_components/
4   | |_ index.html
5   | |_ assets/
6   |   |_ images/...
7   |   |_ js/
8   |   | |_ script_A.js
9   |   | |_ script_B.js
10  |   | |_ script_C.js
11  |
12  |   |_ css/
13  |   | |_ style_A.css
14  |   | |_ style_B.css
15  |   | |_ style_C.css
16  |   |_ ...
17  |
18 |_ dist/
19 | |_ bower_components/
20 | |_ index.html
21 | |_ assets/
22 | |_ images/...
23 | |_ js/
24 |   | |_ script_all.js
25 |   |
26 |   |_ css/
27 |   | |_ style_all.css
28 |   |_ ...
29 |

```

```
30      | _ node_modules/
31      | | _ Gulp
32      | | _ Bower
33      | | _ ..
34      |
35      | _ gulpfile.js
36      | _ package.json
37      | _ bower.json
38      | _ .gitignore
39      | _ ...
```

Listing 15: Projektstruktur

Der Ordner „./site“ ist dabei der Ordner, in dem die Entwicklung stattfindet. Der Ordner „./dist“ beinhaltet die für die Veröffentlichung angepasste Version von optimierten Bildern, verkleinertes und zusammengefügtes CSS und Javascript Dateien und eine Kopie der restlichen, für die Webanwendung wichtigen Elemente. Diese werden automatisch mittels Gulp erzeugt (meistens durch einne sogenannten „build Task“) und lassen sich selbst bei kleinsten Änderungen an der Seite mit nur einem Kommando neu erstellen. Durch diese Arbeitsweise ist es möglich, dass das Projekt von beginn an einen für die Veröffentlichung optimalen Stand hat. Gulp bietet für fast alle Aufgaben eine Automatisierung und es lohnt sich ein Blick in die zahllosen Pakete⁴¹ die von einer immer weiter wachsenden Community bereitgestellt werden.

Durch beachten des Kritischen Rendering-Pfads (3.6), das einsetzen der unter Punkt 4.1 vorgestellten Tools und den in dieser Arbeit vorgestellten Optimierungsmaßnahmen lässt sich eine für den Endanwender akzeptable Ladezeit erreichen.

⁴¹Gulp Plugin Suchmaschiene: <http://gulpjs.com/plugins/>

6 Ergebnis

Bereits zu Beginn des Projekts war es wichtig es messbar zu machen. Dafür wurde eine Testumgebung aufgebaut, mit deren Hilfe die Seite nach ihrer Geschwindigkeit getestet werden kann. Ganz entscheidend war dabei die Webpagetest API in Verbindung mit Google Spreadsheets. Damit lassen sich regelmäßig automatisierte Tests durchführen und die Daten werden nach erfolgreichem Test automatisch in einer Spreadsheet Tabelle gespeichert. Die über den Zeitraum der Arbeit hinweg gesammelten Daten sind hier finden: <http://tinyurl.com/l5usz79>. Diese Daten wurden anschließend mittels <http://Chartjs.org> in Diagrammen aufbereitet und alle Diagramme sind auch Online abrufbar unter: <http://bithugger.github.io/bachelorthesis/>

6.1 Wie wurde getestet?

Damit mittels Google Spreadsheets die Webpagetest API verwendet werden kann, ist es nötig einen sogenannten API-Key anzufordern. Ein solcher Key ist kostenlos unter der Adresse: <http://www.webpagetest.org/getkey.php> zu erhalten und bietet die Möglichkeit täglich 200 Seitenaufrufe zu tätigen. Als Seitenaufruf zählt sowohl die „first view“ als auch „repeat view“. Die Tests sind 30 Tage abrufbar und gespeichert.

Für das Testen der Seite kann aus einer Vielzahl an Teststandorten gewählt werden. Damit lässt sich nachvollziehen wie beispielsweise die Ladezeiten aus der USA oder Asien sind. Je nach Zielgruppe sollten Tests von verschiedenen Standorten in betracht gezogen werden.⁴² Für dieses Projekt wurden ausschließlich Teststandorte aus der USA und Europa gewählt.

Als Testparameter wurde eine Anzahl von 9 Tests pro Testlauf gewählt. Dabei wurde sowohl die „first view“ als auch die „repeat view“ aufgezeichnet. Von den 9 Testläufen wurde der Median als Ergebnis des Testlaufs verwendet. Über den Zeitraum der Arbeit wurde die Seite 1089 Tests unterzogen.

Das größte Problem bestand in der möglichst genauen Messzeiterfassung für die Ladezeiten mittels Smartphone. Webpagetest stellt nur einen Smartphones Teststandort in Dulles USA zu Verfügung. Da die Latenz zwischen dem Hosting in Deutschland und dem Seitenaufruf in der USA sehr viel größer ist, als wenn dieser direkt aus Deutschland erfolgt, wurde nach einer Lösung gesucht diese Messungen exakter zu gestalten. Die Lösung dafür ist, ein zweites Hosting mit der selben Seite in den USA zu erstellen. Dafür wurde die „Microsoft Azure Cloud“ verwendet. Eine kostenlose Testversion ermöglicht es, ganz einfach auf verschiedensten Kontinenten eine Webseite zu hosten. Die Seite wurde für die Tests Online geschaltet und nach den Testläufen wieder Offline genommen.

6.2 Datenauswertung

Folgende Daten wurden bei jedem Testlauf erfasst: Speed Index, TTFB (ms), Render start (ms), Visually complete (ms), Dom Content loaded (ms), Site fully loaded (ms), Requests, Bytes in Document.

Das Diagramm in Abbildung 38, zeigt eine Übersicht der verschiedenen Messwerte über den Zeitraum des Projekts. Die Y-Achse bildet die Zeit in Millisekunden ab und die X-Achse ist das Datum der Messung. Innerhalb von 36 Tagen haben sich alle Werte signifikant verbessert.

⁴²Eine volle Liste der zur Verfügung stehenden Teststandorte ist im Anhang unter Punkt: 10.1 zu finden.

Overview: First View

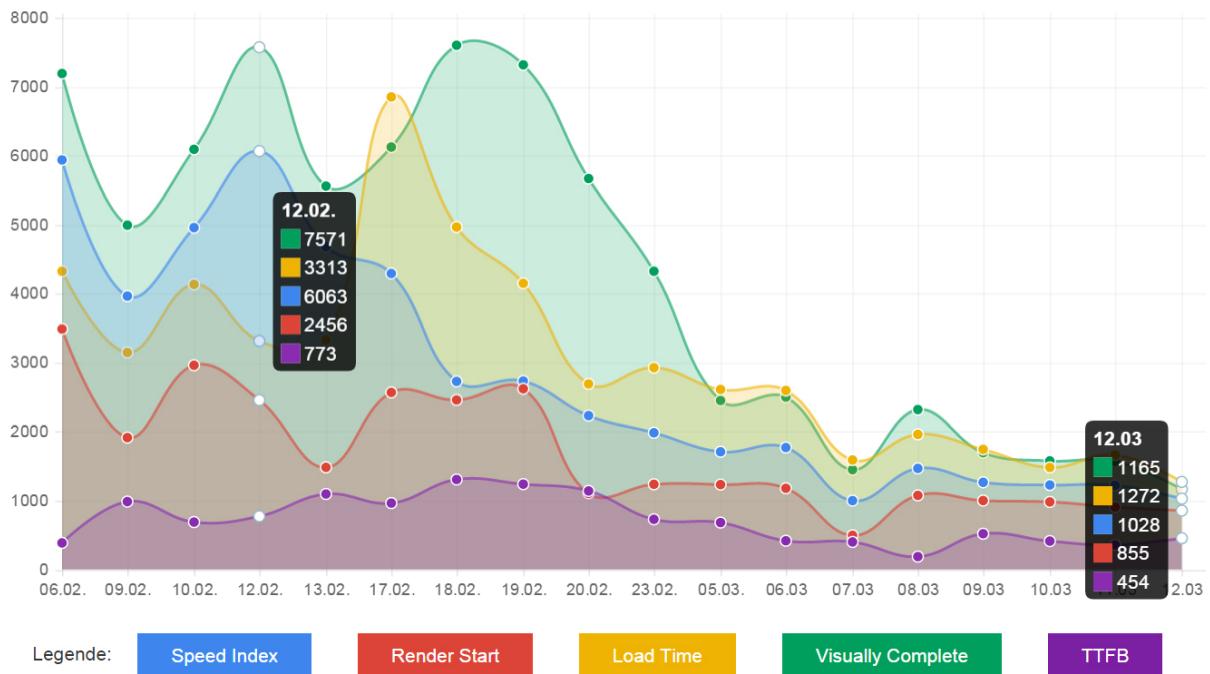


Abbildung 38: Datenauswertung - Überblick

- Speed Index: Ein Speed Index von < 1000 Punkten wird als „schnell genug“ angesehen. Der Median des Speed Indexes konnte von 6063 Zählern auf 1028 Zähler verringert werden.
- Render Start: Während dieser Wert bereits bei Projektbeginn mit 2,4 Sekunden für das erste Rendern recht gut war, konnte auch hier fast das dreifache der Zeit (287%) eingespart werden. Dieser Wert wird durch dieses Diagramm nicht drastisch genug ausgedrückt. So konnte die alte Version der Seite auf dem Smartphone durchaus einen **Render start** von ganzen 10 Sekunden haben. Das dieser Wert im Diagramm vergleichbar gering ausfällt liegt daran, dass auch viele Tests mittels Desktop PC und Kabelverbindung eingeflossen sind. Die jetzige Renderzeit auf dem Smartphone beträgt ungefähr 1,4 Sekunden.
- Load Time bestimmt, wie lange ein Anwender warten muss, bis eine Interaktion mit der Seite möglich ist. Dieser Wert konnte um volle 2 Sekunden verringert werden. Dies wurde vor allem durch eine Verringerung der Seitengröße erreicht.
- Visually Complete: Der höchste Messwert mit rund 7,6 Sekunden konnte auf 1,2 Sekunden verringert werden. Der Hauptgrund dafür ist die Priorisierung des Inhalts „above the fold“. Bei der alten Version erfolgte keine Priorisierung, welche Bilder zuerst und welche zuletzt geladen werden sollen. Dadurch konnte es sein, dass Bilder die sehr weit unten platziert waren, zuerst geladen wurden was die „Visually Complete“ Zeit erhöhte. In der neuen Version wird alles, was unterhalb des „above the fold“ ist verzögert geladen.
- TTFB: Dieser Wert ist grundsätzlich nicht beeinflussbar und sollte durch die Wahl des richtigen Hosting Anbieters so niedrig wie möglich gehalten werden.

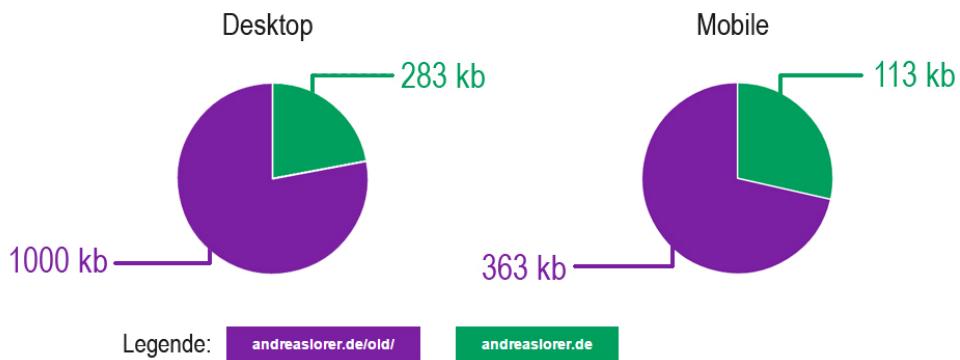


Abbildung 39: Seitengröße in Kilobyte

Die Seitengröße konnte in der Mobilen und Desktop Variante um rund 3/4 reduziert werden. Dadurch sinkt für den Anwender nicht nur die Ladezeit, sondern auch sein Datenvolumen wird weniger in Anspruch genommen. Laut <http://whatdoesmysitecost.com/site/andreaslorer.de> kostet ein Seitenaufruf zwischen 0,01\$ und 0,04\$. Die reduzierung wurde durch mehrere Dinge erreicht:

- Die „best practices“ wurden umgesetzt und der Server, wie oben beschrieben, entsprechend konfiguriert.
- Bilder „below the fold“ werden erst dann geladen, wenn der Anwender dort hin scrollt.
- Bilder wurden umfassend Komprimiert und **Progressive** abgespeichert
- Das Framework wurde von Bootstrap auf **Pure.css** gewechselt.
- Verwendung von „Responsive Images“.
- Es werden mittels **Fontello** nur die Icons geladen, die auch in der Seite eine Verwendung finden.
- JQuery wurde als Abhängigkeit für die Bildergallery entfernt. Dadurch kann das herunterladen von JQuery soweit verzögert werden, bis alle wichtigen Teile der Seite geladen wurden.

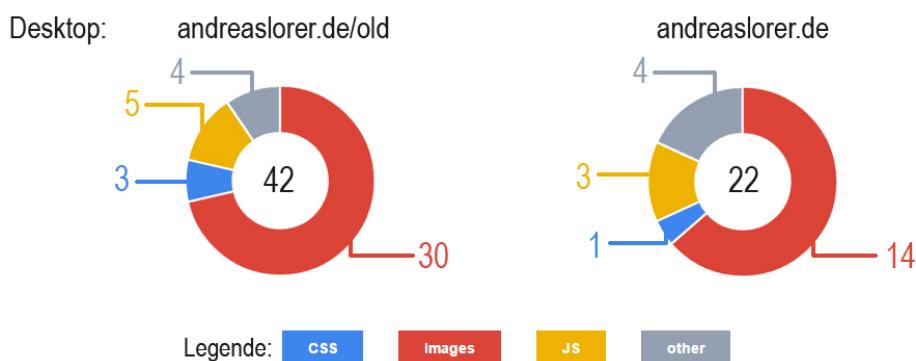


Abbildung 40: Anzahl an Requests via Desktop

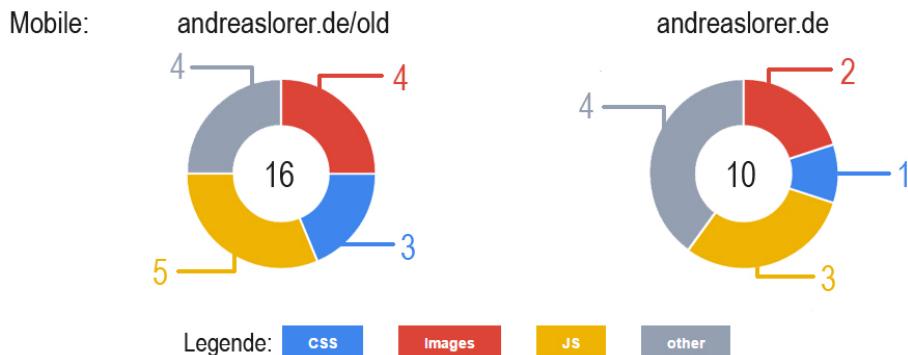


Abbildung 41: Anzahl an Requests via Mobile

- Scripts und sonstige Ressourcen wurden verkleinert und zusammengefasst. Die Anzahl an Requests wurde in der Desktop Variante um die Hälfte verringert.
- In der Mobilen Variante konnten weitere Requests eingespart werden und es wird nur noch der Hintergrund und das erste Bild der Bildergallery geladen.

Beim Seitenaufruf mittels Desktop-PC und Kabelverbindung können „first Render“ Zeiten von unter 189 ms und ein Speed Index von 468 erreicht werden.⁴³ Auf Smartphones mit 3G Netz und 300 ms RTT siehen die Werte etwas schlechter aus und so kann je nach Tag oder Uhrzeit das Ergebnis entsprechend anders ausfallen. Hier sind Werte von 1,3 Sekunden bis zum „first Render“ und einem Speed Index von 1948 möglich.⁴⁴ Ein visueller Vergleich zwischen der alten und neuen Seite ist in Form eines Videos unter der URL: <http://tinyurl.com/o8xoy7m> zu sehen.

⁴³Test Desktop: http://www.webpagetest.org/result/150324_EW_105M/5/details/

⁴⁴Test Mobile: http://www.webpagetest.org/result/150308_5V_JSD/4/details/

7 Der Weg zur Performance

Damit Webanwendungen überhaupt eine gute Performance bieten können, gibt es eine fundamentale Bedingung: Sowohl das ganze Team als auch das ganze Unternehmen muss hinter dem Gedanken „Speed is feature number one“ (Holzle 2010) stehen. Diese Voraussetzung zu erfüllen und alle in ein „Boot“ zu bekommen kann bereits sehr schwierig sein, ist aber für den Erfolg einer schnellen Webanwendung unabdingbar.

„Performance more often comes down to a cultural challenge, rather than simply a technical one.“ (Kovalcin 2015, p. 13)

„If other designers and developers who shape the site aren't educated on performance, how can they make the best decisions about user experience? How can they weigh the balance between aesthetics and page speed? If they aren't empowered to make improvements, any performance champions will simply be playing cleanup after other people's work. Spending your time cleaning up other people's work (especially when it's preventable) is a one-way ticket to burnout.“ (Hogan 2014b)

Damit dies gelingt beschäftigt sich dieser Abschnitt mit der Frage, wie der Weg zur Web Performance aussehen kann und welche Hürden es zu meistern gibt.

7.1 Hürden

Wenn man an das Thema Web Performance denkt, so mag man in erster Linie denken, dass dies ausschließlich für die Entwickler von Belangen ist.



Abbildung 42: (Eigene Abbildung nach (Kovalcin 2015, p. 10))

Genau das Gegenteil ist allerdings der Fall. Wenn eine Webanwendung klassischerweise in die Entwicklungsphase geht, sind bereits die meisten Weichen gestellt. Das Design ist ausgearbeitet und kann mehr oder weniger Mächtig ausfallen. Das Budget für das Projekt wurde vereinbart und der Umfang wurde Vertraglich festgelegt. Das Projektmanagement muss darauf achten, dass die Wünsche und Erwartungen des Kunden erfüllt werden und die Entwicklung sich auf diesen Aspekt konzentriert.



Abbildung 43: (Eigene Abbildung nach (Kovalcin 2015, p. 11))

Deshalb beginnt das Thema Web Performance schon ganz am Anfang. Der Kunde muss mit in das Boot geholt werden und es muss verdeutlicht werden, warum sich schnelle Ladezeiten lohnen, was für negative Auswirkungen langsame und was für positive Auswirkungen schnelle

Ladezeiten auf den Endanwender haben. Der Kunde sieht oftmals nicht den Mehrwert an einer schnellen Webanwendung und möchte für etwas, dass er nicht sehen kann auch kein Geld investieren. Argumente wie:

„We surveyed 3000 users about 17 key product drivers. They rated speed 2nd most important only after easy to find content.“ (Patrick Hamann 2014, p. 8)

„User load time expectations are 2 seconds or less and decreasing“ (Bixby 2013)

oder die in dieser Arbeit genannten Argumente können bei der Überzeugungsarbeit helfen.⁴⁵ Aktuelle Reports oder Zahlen und Fakten die den direkten Zusammenhang zwischen Ladezeit und dem daraus resultierenden Profit verdeutlichen sind, entsprechend aufbereitet, sehr hilfreich. Der Kunde könnte auch Argumentieren, dass er gar nicht so viele Mobile Anwender hat und es sich deshalb nicht lohnt. Dies war damals die selbe Argumentationsweise, die gegen „Responsive Webdesign“ sprach. „*Amongst the top 10,000 websites, almost 8% of websites went responsive within the span of a single year – an incredible growth!*“ (Guypo.com 2014). Diese Argumentation hielt nicht lange und heute spricht jede Firma von Responsive, will eine Responsive Website oder hat zumindest darüber nachgedacht. (Guypo.com 2014) Web Performance ist ein Zukunftstrend, der mit der steigenden Anzahl an Smartphone Nutzer einher geht und damit früher oder später für jeden relevant ist. Ein guter „sales pitch“ könnte sein:

„We'll provide you with a fast, responsive, immersive online experience.“ (Kovalcin 2015, p. 32)

Dem Kunden muss Performance als visuell greifbares Erlebnis präsentiert werden. Dabei sind nicht nur Diagramme sondern auch Tools wie Webpagetest sehr nützlich. Damit können Vergleichsvideos erstellt werden. Denkbar wäre zum Beispiel ein direkter Vergleich mit den Konkurrenzseiten zu zeigen. Auch ein Vorher- / Nachervergleich eines erfolgreichen Projekts könnte dem Kunden präsentiert werden. Zum Launch kann dann ein Vergleich mit pre- und post-Performance aufgezeigt werden.

7.1.1 Projekt Manager

Für das Team der Projekt Manager ergeben sich laut Katie Kovalcin folgende Hürde: (Kovalcin 2015, p. 43)

- Understand the importance
- Advocate with clients
- Help maintain performance budget.

Zuerst müssen die Projekt Manager verstehen, warum Web Performance wichtig für das Projekt ist um den Kunden entsprechend führen und beraten zu können. Des weiteren muss jemand dafür Sorge tragen, dass das Performance Budget (dazu später mehr) eingehalten wird. Neue Anforderungen und features müssen dementsprechend bewertet und mit dem Kunden diskutiert werden.

⁴⁵Eine Sammlung mit Argumenten ist im Anhang unter Punkt 10.2 zu finden.

7.1.2 Aesthetic heavy designers

„Aesthetic heavy designers“ können oft nicht abschätzen, was ihre Entscheidungen für einen Einfluss auf die Webanwendung haben. Die Voraussetzung dafür ist zu wissen, wie das Web funktioniert. Warum Webanwendungen langsam sind und was dazu führt. Erst dann lassen sich Entscheidungen treffen, die sowohl den Endanwender und sein Online Erlebnis, als auch den ästhetischen Anspruch zufriedenstellen. Wikipedia beschreibt den Begriff Design so:

„Insbesondere umfasst Design auch die Auseinandersetzung des Designers mit der Funktion eines Objekts sowie mit dessen Interaktion mit einem Benutzer“
(wikipedia 2015b)

Design ist also mehr als nur die visuelle Aufbereitung von Informationen, es muss in erster Linie dem Anwender dienen. Webseiten die zu groß sind, zerstören den Ansatz von Web Performance bereits schon zu Beginn. Wie in Punkt 2.8 gesagt, verlassen über 50% der Nutzer eine Seite nach einer Verzögerung von nur 3 Sekunden. Die emotionsvollsten Bilder und das prächtigste Design kann dem Besucher keine Botschaft vermitteln, wenn sie niemals zu sehen sein werden.

„When you want to be fast, you have to give up the things slowing you down.“ (Osmani 2014a, p. 2)

Performance darf nicht als schlimmster Feind, sondern muss als bester Freund betrachtet werden. Dabei findet ein Balanceakt statt. Manchmal werden Entscheidungen zugunsten der Performance, ein anderes mal für die Ästhetik getroffen. Der Schlüssel ist es, alle verfügbaren Informationen zu nutzen, um die richtigen Entscheidungen für sich und die Webanwendung zu treffen. (Hogan 2014a, p. 126)

Der Designer muss sich Fragen stellen wie: „Welchen Mehrwert hat der Nutzer durch dieses große Bild auf der Startseite“, braucht es 3 Schriftarten um einen gewissen **Look & Feel** zu vermitteln oder reicht eine? Gibt es eine alternative Schriftart die fast identisch aussehen, aber viel weniger Bytes benötigt?

Tabelle 1: Beispiel: Abwägung - Performance oder Ästhetik
(Tabelle nach (Hogan 2014a, p. 126))

Question	Aesthetische Consideration	Performance Consideration
Can I put a large hero image at the top of every page?	Eye-catching represents the brand	This could be a really large file, we want to minimize page weight
Should I use three Font-Weights plus a text weight?	Lots of flexibility in typography	We want to reduce page weight and requests
Do I need a carousel on the landing page?	Showcases lots of different content	This adds a lot of page weight and additional requests. The user might never see the 2nd image.
How can I demonstrate the product functionalities?	Could use a GIF or embed a video	Videos and GIF's can be very heavy.

Die Antworten können je nach Projekt oder Designer unterschiedlich ausfallen.

Tabelle 2: Beispiel: Entscheidungen (Tabelle nach (Hogan 2014a, p. 127))

Question	Decision
Can I put a large hero image at the top of very page?	Yes, we use responsive images for the different screen sizes and compress them to reduce page weight! We might lose image quality.
Should I use three Font-Weights plus a text weight?	We need the 3 Font-Weights, they are used by the brand. No choice there
Do I need a carousel on the landing page?	No the extra images are not increasing the users experience.
How can I demonstrate the product functionalities?	We will use CSS-Animations instead of a GIF. This will cost some of the developers time.

Für viele Entscheidungsschritte macht es oft Sinn, dass Entwickler und Designer kollaborieren. Dabei können bereits zu einem sehr frühen Zeitpunkt **Bottlenecks** erkannt, alterantiven vorgeschlagen und diskutiert werden. Dafür muss es klare Regeln geben. So haben Phrasen wie: „Das ist zu schwierig, gefällt mir nicht, dumme Idee, da gibt es keine Alternative, das Bild muss genau so aussehen“ generell nichts verloren. Das Wort „Performance“ darf nicht als **Trumpfkarte** für einen Entwickler gegenüber dem Designer gesehen werden. Viel mehr Sinn macht es Alternativen zu erarbeiten, Prioritäten zu diskutieren oder Lösungen aufzuzeigen. Ebenso kann es hilfreich sein, früh mit einem Mockup oder Prototyp auf Code-Basis anzufangen und gemeinsam an diesem auszuprobieren.

Design



Development



Abbildung 44: Projekt Zeitlinie

Auch hier gilt wieder den Balanceakt zwischen Ästhetik und Performance zu finden und nicht gegeneinander, sondern miteinander zu arbeiten. Um dem ganzen einen Rahmen zu geben, in dem sich sowohl Designer, Entwickler als auch der Kunde bewegen darf, wird von vielen Performance Führsprechern ein sogenanntes „Performance Budget“ verwendet.

7.2 Performance Budget

Ein Performance Budget bedeutet, dass ein Wert gesetzt wird, der von der Seite nicht überschritten werden darf. Dieser Wert kann ganz simpel sein, wie die Ladezeit der Seite oder aber eine komplexere Metrik aus verschiedenen Werten.

„The important point is to look at every decision, right through the design/build process, as something that has consequence. Having a pre-defined ‘budget’ is a clear, tangible way to frame decisions about what can and can’t be included, and at a suitably early stage in the project. It can also potentially provide some justification to the client about why certain things have been omitted (or rather, swapped out for something else).“

Die Einführung eines Performance Budgets bietet den Vorteil, einen festgelegten Rahmen über die Zeitspanne des Projekts zu haben. Es ist ein Referenzpunkt der mit darüber entscheidet, welche und welche Komponenten nicht in die Seite mit einfließen können oder dürfen. Es funktioniert wie Spielgeld, dass auf das Projekt aufgeteilt wird. Ist das Spielgeld leer, so gibt es diese 3 Regeln:

1. Optimiere eine existierende Funktion der Seite
2. Entferne eine existierende Funktion von der Seite
3. Füg es nicht hinzu

Das Motto heißt dabei: „you can't spend, what you don't have!“

7.2.1 Budget Metriken

Eine sehr einfache Budget Metrik wähle es, den Speed Index als Budget zu setzen. Jede Seite muss dann unter diesem festgelegten Wert bleiben. Oftmals ist es aber besser verschiedene Werte als Budget zu setzen. Eine Kombination aus: Anzahl Requests, start Render, maximale Seitengröße, maximales Gewicht der Bilder und zusätzlich der Speed Index können eine sinnvolle und verständliche Basis bilden. Es können aber je nach Projekt auch anwendungsbezogene Metriken verwendet werden. So benutzt Twitter zum Beispiel die Metrik „Time to first Tweet“.

„The most important metric we used was “time to first Tweet”. This is a measurement we took from a sample of users, of the amount of time it takes from navigation (clicking the link) to viewing the first Tweet on each page’s timeline. The metric gives us a good idea of how snappy the site feels.“ (Twitter 2012)

7.2.2 Wie schnell, ist schnell genug?

Wie wird das Performance Budget gesetzt? Dazu ist es erforderlich zu wissen, wie schnell eigentlich schnell genug ist? Schnelligkeit ist ein relativer Begriff.

*„In the high-speed world of automated financial trading, milliseconds matter. So much so, in fact, that a saving of just **6 milliseconds** in transmission time is all that is required to justify the laying of the first transatlantic communications cable for 10 years at a cost of more than \$300m between London and the New York Wall Street.“ (Williams 2011, vgl.)*

Aber auch Google stellt fest, dass bereits der Einfluss von nur **100 - 400 ms** einen messbaren Einfluss auf die Anzahl an Suchanfragen (-0,2% bis -0,6%) pro Anwender hat.(Google 2009) Während 0,4% sich nach wenig anhört, so ist bei jährlich 2,2 Billionen Suchanfragen (2013) wohl klar, was das für Google an Mehreinnahmen durch Werbung und Klicks bedeuten kann.(Statista.com 2014)

Studien haben ergeben, dass Menschen den Unterschied zwischen 2 Zeitspannen erst dann erkennen, wenn die Differenz 20% überschreitet.(Steve Seow 2009) Damit der Anwender eine Verbesserung gegenüber der Konkurrenz bemerkt, benötigt es in Folge dessen mindestens einen 20 prozentigen Unterschied. Anhand der Konkurrenzanalyse lässt sich ein Ergebnis bestimmen, mit dessen Hilfe das Performance Budget festgelegt werden kann. Bei einem Relaunch kann auch die alte Seite als Referenzpunkt dienen.

Zu beachten bleibt, die Daten richtig zu interpretieren. Eine Seite die eine 7 Sekunden Ladezeit hat muss nicht unbedingt langsamer sein (Stichwort: Perceived Performance) als eine Seite

mit 5 Sekunden. Hier lohnt sich vor allem die Beachtung des Speed Indexes, der die Seite nach dem „visuellen Fortschritt über Zeit“ bewertet. Die 20%-Methode liefert den wohl minimalsten „schnell-genug“ Wert der erreicht werden sollte, um sich von der Konkurrenz abzusetzen. Paul Irish, Mitarbeiter des Google Chrome Teams, sieht das ganze drastischer und gibt folgendes Statement:

„My answer to how fast is fast enough? A Speed Index of under 1000.

And for professionals that get there, they should shoot for delivering the critical-path view (above the fold) in the first 14kb of the page.“ (Irish 2014)

Für viele (vor allem im Bereich des E-Commerce) kann es sich rechnen, nicht nur schneller als die Konkurrenz zu sein.

7.2.3 Arbeiten mit einem Performance Budget

Im folgenden soll Beispielhaft verdeutlicht werden, wie das Arbeiten mittels Performance Budget aussehen kann. Zu Projektbeginn wird das Performance Budget festgelegt.



Abbildung 45: Projekt Zeitline (Eigene Abbildung nach (Kovalcin 2015, p. 58)

Vor Projektbeginn macht Sinn, denn hat die Seite erst einmal 3 Slider, 1 Karousel plus entsprechendes Plugin und zudem noch ein vollflächiges Hintergrundbild, so hat man keine Chance mehr die Seite in einen geordneten Rahmen zu bringen, ohne sie komplett neu zu überarbeiten. Ein Beispiel Budget kann aus folgenden Metriken bestehen:

- Maximale Seitengröße 600 kb
- Start Render 1000 ms
- Speed Index 1000

Die Seitengröße lässt sich nun aufteilen:

Tabelle 3: Seitengröße mit einem Budget von 600 kb

Content	Size
CSS	50 kb
Fonts	50 kb
Scripts	100 kb
Images	400 kb

Mit diesem Budget ist festgelegt, wieviel Spielraum jede Komponente hat. Will man 500 kb den Bildern zu Verfügung stellen, so muss in den anderen Bereichen eingespart werden. Dieses Beispiel verdeutlicht den Ansatz: „You can't spend what you don't have“.

Nachdem das Budget festgelegt wurde ist sowohl das Design, die Entwicklung als auch das Projektmanagement damit beauftragt, das Budget beizubehalten.

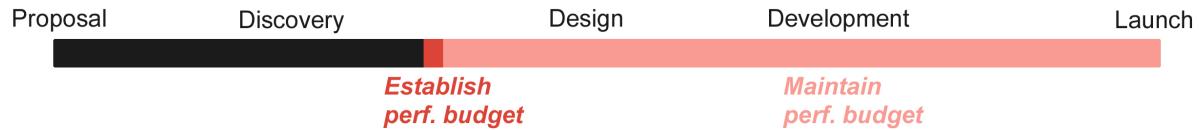


Abbildung 46: Projekt Zeitline (Eigene Abbildung nach (Kovalcin 2015, p. 59))

Dafür muss dem Kunden vermittelt werden, um was es sich bei diesem Budget handelt und warum es verwendet wird: (Aufzählung nach (Kovalcin 2015, p. 72))

- Was ist ein Performance Budget?
- Was ist das Performance Budget für dieses Projekt und wie kam es zustande?
- Was sind die Bedingungen für dieses Budget?
- Warum wird es verwendet?
- Wie werden neue Funktionen hinzugefügt? (Optimieren, entfernen, nicht hinzufügen)
- Wieviel Budget hat jede Seite- / Unterseite

Für Designer (als auch Entwickler) ergibt sich ein Rahmen. Dies muss in keiner Weise schlecht sein und Dan Mall, Art Director und Designer sagt dazu folgendes:

„I believe designers do their best work within constraints, and knowing those constraints before starting a design can be incredibly enabling. What I wouldn't give to know that I could use up to 10 images and 4 webfonts before starting a design! What a day that would be! Here's how to make that possible. [We define a Performance Budget]“(Mall 2014)

Wenn diese Weichen für das Projekt gestellt wurden, dann kann das Entwickler-Team die Anwendung im Sinne der Performance umsetzen.

Mittels Grunt (leider gibt es noch keine gute Gulp Alternative) lässt sich ein npm Paket namens „Grunt perfbudget“ installieren. Dieses ermöglicht es für die Seite einen Budget Task festzulegen. Dieser Task kann nun in den Build-Prozess der Seite integriert werden. Wird das Budget überschritten, so würde der Build-Task fehlschlagen.

```
Running "perfbudget:default" (perfbudget) task
>> -----
>> Test for http://cfarman.com    FAILED
>> -----
>> render: 2092 [FAIL]. Budget is 1000
>> SpeedIndex: 5426 [FAIL]. Budget is 1000
>> Summary: http://www.webpagetest.org/result/141107_HW_19HS/
Warning: Task "perfbudget:default" failed. Use --force to continue.

Aborted due to warnings.
```

Abbildung 47: Fehlschlagender Grunt Budget Task (Abbildung von (Farman 2014))

Das setzen eines Performance Budget und das Arbeiten damit, ist eine Herausforderung. Das Budget unterscheidet sich je nach Projekt und es kann schwierig sein die richtigen Metriken festzulegen. Dieser Abschnitt hat auch gezeigt, dass Web Performance nicht abhängig von einem einzigen Entwickler ist, sondern kann nur gelingen, wenn alle an einem Strang ziehen.

8 Ausblick

Sowohl für die Endanwendner als auch für die Entwickler und Designer, hat die Zukunft vor allem positive Änderungen für sich parat. Das Http/1.1 Protokoll wird durch Http/2.0 abgelöst. Der LTE-Netzausbau schreitet weiter voran, deckt dabei eine immer größere Fläche ab und bringt den Nutzern damit eine niedrigere Latenz und eine höhere Bandbreite. Auf der diesjährigen CeBIT wurde der neue Mobilfunkstandard 5G von Vodafone präsentiert. Dieser steht noch in der Entwicklung, wird aber für das Jahr 2020 für den kommerziellen Einsatz vorhergesagt. 5G soll dabei Latenzen zwischen 1 bis 10 Millisekunden liefern und eine 100 mal höhere Bandbreite von bis zu 10.000 MBit/s zur Verfügung stellen. (lte-anbieter.info 2015) Ob diese Geschwindigkeiten auch wirklich innerhalb von nur 5 Jahren erreicht werden können, oder ob sie in erster Linie nur der PR dienen, bleibt dabei abzuwarten.

8.1 Http/2.0

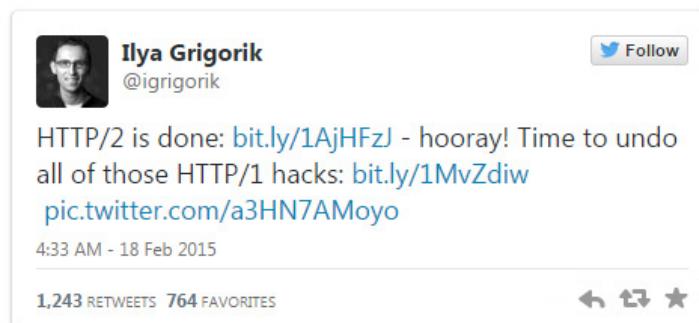


Abbildung 48: Http 2.0 wurde veröffentlicht

Dieses Jahr ist es soweit und das Http/1.1 Protokoll aus dem Jahre 1999 wird durch Http/2 abgelöst. Natürlich verschwindet das alte Protokoll nicht von heute auf morgen. Http/1.1 wird noch viele Jahre erhalten bleiben, vielleicht auch nie gänzlich verschwinden. Dennoch sagt Daniel Stenberg, Mitglieder der Http/2 Work Group, voraus, dass bis Ende 2015 der Anteil an Http/2 traffic mehr als 10% beträgt. (Stenberg 2015) Die großen Browser Chrome und Firefox haben in ihrer aktuellen Version Http/2 bereits aktiviert und auch der Internet Explorer soll mit Windows 10 (in der Technical Preview bereits aktiv) Http/2 unterstützen. (Microsoft 2014) Schlusslicht bildet Apple mit seinem Safari Browser, von denen bis dato (01.04.2015) noch keine Stellungnahme zu Http/2 verlautet wurde. Die zwei weitverbreitesten Server Apache und Nginx sind dabei, Http/2 zur Verfügung zu stellen. Apache hat bereits ein Http/2 Modul in der Alpha Phase. Nginx sagt, dass sie bis Ende 2015 Http/2 unterstützen werden. Google (damit auch Youtube und andere Produkte die zu Google gehören) und Twitter haben bereits Http/2 seit einigen Monaten aktiviert.

● 200 GET / twitter.com	Angefragte Adresse: https://twitter.com/
▲ 304 GET  1f389.png abs.twimg.com	Anfragemethode: GET
▲ 304 GET  1NBmhRZy_bigger.jpeg pbs.twimg.com	Status-Code: ● 200 OK
▲ 304 GET  2f2b89177e469d18f473835dc13... pbs.twimg.com	Bearbeiten und erneut senden Kopfzeilen (unformatiert) Version: HTTP/2.0

Http/2 bietet folgende Performance Verbesserungen:

- Parallel Multiplexing anstatt der Verwendung von Parallelen Verbindungen
- Verwendet nur eine einzige TCP Verbindung
- Server-Push

8.1.1 Http/1.1 Optimierungen in Http/2

Viele Optimierungs Pattern für Http/1.1 stellen sich für das Http/2 Protokoll als Anti-Pattern heraus. Http/2 bringt den Vorteil, dass es genau eine TCP Verbindung benötigt. Damit wird **Domain Sharding** ein Performance Anti-Pattern für HTTP 2.0. Auch die Verwendung von **Image Sprites**, das Zusammenfügen von CSS und Javascript zu einer Datei ist ein Anti-Pattern wenn Http/2 verwendet wird. Eine Vielzahl von kleinen einzelnen Dateien werden zu keinem Performance Problem mehr durch Http/2.(Grigorik 2013f)

„Server Push“ ist eine Mechanik, die in Http/1.1 durch das Inlinen von CSS in das HTML Dokument bereits simuliert wird. De facto ist dies ein Workaround für eine Funktionalität die nun durch Http/2 bereitgestellt wird. Durch Inlinen wird bei einer Anfrage gleich das CSS / Javascript als Antwort mitgesendet, dass der Browser zur Darstellung des „above the fold“ Inhaltes braucht.

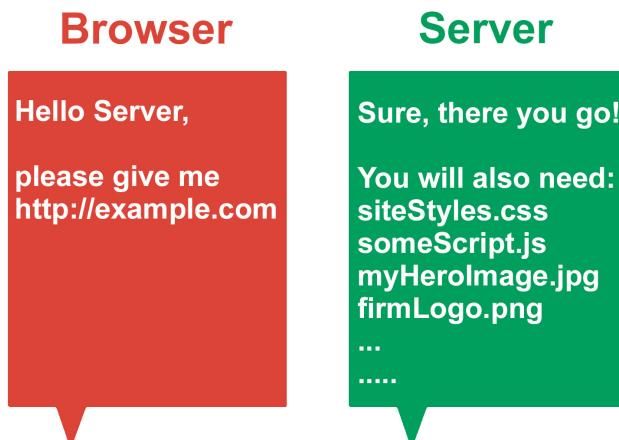


Abbildung 49: Veranschaulichung der Server Push Mechanik (Eigene Abbildung)

Durch **Server Push** wird genau dies möglich, ohne es Inline in das Dokument zu schreiben. Dadurch sind die Ressourcen vom Browser im Cache speicherbar und können auch in den Unterseiten verwendet werden. Bereits bevor der Anwender die Seite besucht, weißt man welche Ressourcen ihm zur Verfügung gestellt werden müssen, um die Seite anzuzeigen. Diese Ressourcen lassen sich somit gleich auf die erste Anfrage als Antwort mitsenden und es wird damit das Explizite Anfragen und Antworten eingespart.⁴⁶

Durch Http/2 ergeben sich viele Vereinfachungen, aber auch neue Herausforderungen. Der Umstieg von Http/1.1 auf Http/2 wird nicht über Nacht geschehen. Für viele Seitenbetreiber wird es deshalb nötig sein, sowohl auf das Alte, wie auch das neue Protokoll zu optimieren und an den jeweiligen Anwender die jeweils richtige Version auszuliefern. Dafür gibt es mehrere Ansätze und eine ausführliche Beschreibung dazu gibt es hier: <http://tinyurl.com/phnq5d9>.⁴⁷

⁴⁶Dieses Video stellt eindrücklich die Auswirkung der Server Push Mechanik vor (Länge 5 Minuten): https://youtu.be/4Ai_rrhM8gA?t=29

⁴⁷Eine detaillierte Erklärung zu Http/2.0 ist dem PDF: http2 explained - Daniel Stenberg: <http://daniel.haxx.se/http2/http2-v1.11.pdf> zu entnehmen.

9 Fazit

Is the web getting faster?

10 Anhang

10.1 Webpagetest Teststandorte

Name	Standort
Eu-West, Chrome, Cable	ec2-eu-west-1:Chrome
Eu-West, Chrome, 3G	ec2-eu-west-1:Chrome.3G
Eu-Central, Firefox	ec2-eu-central-1:Firefox
Dulles, MotoG, Chrome	Dulles_MotoG:Motorola G - Chrome
Us-East, Chrome, 3G	ec2-us-east-1:Chrome.3G
Eu-Central, Chrome	ec2-eu-central-1:Chrome.3G
Eu-Central, IE_11	ec2-eu-central-1:IE 11
Us-East, IE11	ec2-us-east-1:IE 11
Us-East, Firefox	ec2-us-east-1:Firefox
Us-West, Chrome	ec2-us-west-1:Chrome
Us-West, IE11	ec2-us-west-1:IE 11
Us-West, Firefox	ec2-us-west-1:Firefox
Us-West-2, Chrome	ec2-us-west-2:Chrome
Us-West-2, IE_11	ec2-us-west-2:IE 11
Us-West-2, Firefox	ec2-us-west-2:Firefox
Ap-Northeast, Chrome	ec2-ap-northeast-1:Chrome
Ap-Northeast, IE_11	ec2-ap-northeast-1:IE 11
Ap-Northeast, Firefox	ec2-ap-northeast-1:Firefox
Ap-Southeast-1, Chrome	ec2-ap-southeast-1:Chrome
Ap-Southeast-1, IE_11	ec2-ap-southeast-1:IE 11
Ap-Southeast-1, Firefox	ec2-ap-southeast-1:Firefox
Ap-Southeast-2, Chrome	ec2-ap-southeast-2:Chrome
Ap-Southeast-2, IE_11	ec2-ap-southeast-2:IE 11
Ap-Southeast-2, Firefox	ec2-ap-southeast-2:Firefox
SA-East, Chrome	ec2-sa-east-1:Chrome
SA-East, IE_11	ec2-sa-east-1:IE 11
SA-East, Firefox	ec2-sa-east-1:Firefox

10.2 Argumentations Sammlung

Die folgende Sammlung an Argumenten ist der Seite: <http://blog.apakau.com/tag/web-performance/page/2/> entnommen. Dort befinden sich auch alle Links zu den Quellen.

“In 2014, the median top 100 e-commerce page takes 6.2 seconds to render its primary content, 10.7 seconds to fully load; i.e. 27% longer to begin rendering than it did in 2013”

“When faced with a negative mobile shopping experience, 43% of consumers will go to a competitor’s site next”

“Mobile users get frustrated after 1 second of delay. 500 ms delay increases user frustration by 26% and lowers engagement by 8%”

“47% of web users expect a page load of 2 seconds or less”

“57% of online users will abandon a website that takes more than 3 seconds to load”

“Shoppers remember online wait times as being 35% longer than they actually are”

“51% of online shoppers in the US say that site slowness is the top reason they’d abandon a purchase”

“64% of smartphone users expect pages to load in less than 4 seconds”

“32% of online consumers will start abandoning slow sites between 1 and 5 seconds”

“39% of users say speed is more important than functionality for most websites”

“18% of shoppers will abandon their cart if pages are too slow”

“37% of consumers said they would not return to a slow site, and 27% would likely jump to a competitor’s site”

“After experiencing a slow site, 14% of shoppers will begin shopping at another site, and 23% will stop shopping or walk away from their computer”

“80% of users will not return to a site after a disappointing experience. Of these, 37.5% will go on to tell others about their experience”

“64% of shoppers who are unhappy with their site visit will go elsewhere to shop next time”

“52% of online shoppers claim that quick page loads are important for their loyalty to a site”

“73% of mobile users said they’ve encountered a site that was too slow to load”

“Users who experience a 2-second site slowdown make almost 2% fewer queries, click 3.75% less often, and report being significantly less satisfied with their overall experience”

“A site that loads in 3 seconds experiences 22% fewer page views, 50% higher bounce rate and 22% lower conversion rate than a site that loads in 1 second”

“A site that loads in 5 seconds experiences 35% fewer page views, 105% higher bounce rate, and 38% lower conversion rate than a site that loads in 1 second”

“A site that loads in 10 seconds experiences 46% fewer page views, 135% higher bounce rate, and 42% lower conversion rate than a site that loads in 1 second”

“Issues with application performance are affecting overall business revenues by up to 9%”

“A 1-second delay in page response time decreases pages views by 11%, customer satisfaction by 16% and conversion by 7%”

“When pages are slow, business metrics suffer more now than they did a few years ago. For example, a page that took 6 seconds to load in 2010 suffered a -40% conversion hit vs. -50% hit in 2013”

“If Amazon increased page load time by 100 ms they would lose 1% of sales”

“Shopzilla achieved a 7–12% increase in conversion rate and a 25% increase in page views after a 5-second improvement in page load time”

“Shopzilla was able to support the same volume with 50% (402 to 200 nodes) less nodes, cutting server costs in half after a 5-second improvement in page load time”

“Facebook pages that are 500 ms slower result in a 3% drop-off in traffic, and 6% drop-off for 1 second”

“Search engines like Google recommend improving loading time when a site is slower than 95% of others”

“Yahoo saw a 5 – 9% drop in full-page traffic after increasing page load times by 400 ms”

“If Google increased page load by 500 ms, they get 25% fewer searches”

“AOL realized 160% increase in average number of page views from users in the top 10th percentile (fastest experience), compared to the bottom 10% (slowest experience)”

“Bing test results for a 2 second delay in page load time: Linear impact with increasing delay Time-to-click changed by double the delay User satisfaction dropped by 3.8% Revenue per user dropped by 4.3% Number of clicks dropped by 4.4%”

Literatur

- [Apa15] Apache.org. *When (not) to use .htaccess files.* <http://tinyurl.com/m6v5rut> [Aufgerufen am 17.03.2015]. 2015 (siehe S. 37).
- [Bar14] Bart. *Where is the best place to put <script> tags in HTML markup?* <http://stackoverflow.com/questions/436411/where-is-the-best-place-to-put-script-tags-in-html-markup> [Aufgerufen am 08.03.2015]. 2014 (siehe S. 18).
- [Bix13] Joshua Bixby. *Top ecommerce sites are 22 percent slower than they were last year.* <http://www.webperformancetoday.com/2013/03/27/top-ecommerce-sites-are-slower-than-they-were-last-year/> [Aufgerufen am 27.03.2015]. 2013 (siehe S. 56).
- [Bun14] Bundesnetzagentur. *Jahresbericht 2014.* <http://tinyurl.com/18r7flv> [Aufgerufen am 19.03.2015]. Seite 78f. 2014 (siehe S. 11).
- [can15] caniuse.com. *Latest supported data.* <http://caniuse.com/> [Aufgerufen am 13.03.2015]. 2015 (siehe S. 29, 41, 42).
- [Far14] Catherine Farman. *Automate Performance Testing with Grunt.js.* <http://www.sitepoint.com/automate-performance-testing-grunt-js/> [Aufgerufen am 01.04.2015]. 2014 (siehe S. 61).
- [Goo09] Google. *Speed Matters.* <http://googleresearch.blogspot.de/2009/06/speed-matters.html> [Aufgerufen am 30.03.2015]. 2009 (siehe S. 59).
- [Goo10] Google. *Using site speed in web search ranking.* Website. 2010 (siehe S. 3).
- [Goo11] Google. *Creating Fast Buttons for Mobile Web Applications.* https://developers.google.com/mobile/articles/fast_buttons [Aufgerufen am 03.03.2015]. 2011 (siehe S. 10).
- [Goo14a] Google. *Browser-Caching nutzen.* <https://developers.google.com/speed/docs/insights/LeverageBrowserCaching> [Aufgerufen am 16.03.2015]. 2014 (siehe S. 34).
- [Goo14b] Google. *Mobile Analyse in PageSpeed Insights.* <https://developers.google.com/speed/docs/insights/mobile> [Aufgerufen am 09.03.2014]. 2014 (siehe S. 17).
- [Goo15] Google. *Antwortzeit des Servers verbessern.* <https://developers.google.com/speed/docs/insights/Server> [Aufgerufen am 13.03.2015]. 2015 (siehe S. 33).
- [Gri13a] Ilya Grigorik. *Breaking the 1000ms Mobile Barriere.* https://docs.google.com/presentation/d/1wAxB5DPN-rcelwbG06lCOus_S1rP24LMqA8m1eXEDRo/present?slide=id.g11c1373c5_3_0 [Aufgerufen am 04.03.2015]. Slides. 2013 (siehe S. 12, 13).
- [Gri13b] Ilya Grigorik. *Breaking the 1000ms Mobile Barriere.* https://docs.google.com/presentation/d/1wAxB5DPN-rcelwbG06lCOus_S1rP24LMqA8m1eXEDRo/present?slide=id.g11c1373c5_5_35 [Aufgerufen am 04.03.2015]. Slides. 2013 (siehe S. 13).
- [Gri13c] Ilya Grigorik. *High Performance Browser Networking.* <http://tinyurl.com/p5dds9p> [Aufgerufen am 02.03.2015]. Chapter 2 Slow-Start. 2013 (siehe S. 6, 7).
- [Gri13d] Ilya Grigorik. *High Performance Browser Networking.* <http://tinyurl.com/lz8t3mh> [Aufgerufen am 02.03.2015]. Chapter 10 Speed, Performance, and Human Perception. 2013 (siehe S. 10).
- [Gri13e] Ilya Grigorik. *High Performance Browser Networking.* <http://tinyurl.com/nojaxxa> [Aufgerufen am 02.03.2015]. Chapter 7 Table 7.1. 2013 (siehe S. 12).

- [Gri13f] Ilya Grigorik. *Removing 1.x Optimizations*. <http://chimera.labs.oreilly.com/books/1230000000545/ch13.html> [Aufgerufen am 01.04.2015]. 2013 (siehe S. 63).
- [Gri14a] Ilya Grigorik. *Codierung und Übertragungsgröße textbasierter Ressourcen optimieren*. <https://developers.google.com/web/fundamentals/performance/optimizing-content-efficiency/optimize-encoding-and-transfer?hl=de> [Aufgerufen am 21.03.2015]. 2014 (siehe S. 36).
- [Gri14b] Ilya Grigorik. *HTTP-Caching*. <https://developers.google.com/web/fundamentals/performance/optimizing-content-efficiency/http-caching?hl=de> [Aufgerufen am 16.03.2015]. 2014 (siehe S. 34).
- [gro14] growingwiththeweb.com. *async vs defer attributes*. <http://www.growingwiththeweb.com/2014/02/async-vs-defer-attributes.html> [Aufgerufen am 13.03.2015]. 2014 (siehe S. 30).
- [Gro15] Filament Group. *Defer Loading Javascript*. <https://github.com/filamentgroup/loadJS> [Aufgerufen am 13.03.2015]. GitHub. 2015 (siehe S. 30).
- [Guy14] Guypo.com. *Responsive Web Design Adoption, 2014*. <http://www.guypo.com/rwd-2014/> [Aufgerufen am 27.03.2015]. 2014 (siehe S. 56).
- [Hog14a] Lara Hogan. *Designing for Performance*. O'Reilly, 2014. ISBN: 1-4919-0251-5 (siehe S. 57, 58).
- [Hog14b] Lara Hogan. *Performance Cops and Janitors*. <http://davidwalsh.name/performance-cops-janitors> [Aufgerufen am 27.03.2015]. 2014 (siehe S. 55).
- [Hol10] Urs Holzle. *Velocity 2010: Urs Holzle*. <http://tinyurl.com/px7m64m> [Aufgerufen am 27.02.2015]. Video from Velocity Conference. 2010 (siehe S. 3, 55).
- [htt15] httpArchive. *Interesting stats*. <http://httparchive.org/interesting.php> [Aufgerufen am 04.03.2015]. 2015 (siehe S. 27).
- [Iri14] Paul Irish. *Fast-Enough*. <http://tinyurl.com/pdjmbp3> [Aufgerufen am 30.03.2015]. 2014 (siehe S. 60).
- [ItW15] ItWissen.info. *Shared Hosting*. <http://www.itwissen.info/definition/lexikon/Shared-Hosting-shared-hosting.html> [Aufgerufen am 26.02.2015]. 2015 (siehe S. 5).
- [Kov15] Katie Kovalcin. *The Path to Performance*. <https://speakerdeck.com/katiekovalcin/the-path-to-performance> [Aufgerufen am 27.03.2015]. 2015 (siehe S. 55, 56, 60, 61).
- [Les14] Kornel Lesiński. *MozJPEG 3.0*. <http://calendar.perfplanet.com/2014/mozjpeg-3-0/> [Aufgerufen am 12.03.2015]. 2014 (siehe S. 41).
- [lte15] lte-anbieter.info. *Cebit 2015: Vodafone präsentiert 5G made in germany*. <http://www.lte-anbieter.info/lte-news/cebit-2015-vodafone-praesentiert-5g-made-in-germany> [Aufgerufen am 01.04.2015]. 2015 (siehe S. 62).
- [Mal14] Dan Mall. *HOW TO MAKE A PERFORMANCE BUDGET*. <http://danielmall.com/articles/how-to-make-a-performance-budget/> [Aufgerufen am 30.03.2015]. 2014 (siehe S. 61).
- [Mee14] Patrick Meenan. *Titel*. <http://tinyurl.com/o4b3rxh> [Aufgerufen am 11.03.2015]. blog. 2014 (siehe S. 24).
- [Mic14] Microsoft. *Http/2: The Long-Awaited Sequel*. <http://blogs.msdn.com/b/ie/archive/2014/10/08/http-2-the-long-awaited-sequel.aspx> [Aufgerufen am 01.04.2015]. 2014 (siehe S. 62).

- [Osm14a] Addy Osmani. *CSS Performance Tooling*. <https://speakerdeck.com/addyosmani/css-performance-tooling> [Aufgerufen am 27.03.2015]. 2014 (siehe S. 57).
- [Osm14b] Addy Osmani. *Front-end Tooling Workflows*. <https://speakerdeck.com/addyosmani/front-end-tooling-workflows> [Aufgerufen am 21.03.2015]. 2014 (siehe S. 45).
- [Pat14] The Guardian Patrick Hamann. *Breaking news at 1000ms - Front-Trends 2014*. <https://speakerdeck.com/patrickhamann/breaking-news-at-1000ms-front-trends-2014> [Aufgerufen am 27.03.2015]. 2014 (siehe S. 56).
- [Rad13] Radware. *Radware Mobile Infographic*. Website. http://blog.radware.com/wp-content/uploads/2013/11/Radware_SOTU_Fall_2013_Mobile_Infographic_Final1.jpg [Aufgerufen am 15.01.2015]. 2013 (siehe S. 3).
- [Rad14] Radware. *State of the union - Ecommerce Page Speed and Web Performance*. <http://www.radware.com/assets/0/314/6442478110/c810eee1-e86f-438a-b82f-3ad002bf1c75.pdf> [Aufgerufen am 19.03.2015]. 2014 (siehe S. 9, 39).
- [Rit14] Michael Ritz. *Weltkarte in Schwarz*. <http://www.landkartenindex.de/kostenlos/?p=31> [Aufgerufen am 25.02.2015]. 2014 (siehe S. 7).
- [Sch15] Amy Schade. *The Fold Manifesto: Why the Page Fold Still Matters*. Techn. Ber. <http://www.nngroup.com/articles/page-fold-manifesto/> [Aufgerufen am 26.02.2015]. Nielsen Norman Group, 2015 (siehe S. 8).
- [Sex15a] Patrick Sexton. *Enable Keep Alive*. <http://www.feedthebot.com/pagespeed/keepalive.html> [Aufgerufen am 17.03.2015]. 2015 (siehe S. 38).
- [Sex15b] Patrick Sexton. *Leverage Browser Caching*. <http://www.feedthebot.com/pagespeed/leverage-browser-caching.html> [Aufgerufen am 16.03.2015]. 2015 (siehe S. 34).
- [Sta14] Statista.com. *Anzahl der Suchanfragen bei Google weltweit in den Jahren 2000 bis 2013 (in Milliarden)*. <http://de.statista.com/statistik/daten/studie/71769/umfrage/anzahl-der-google-suchanfragen-pro-jahr/> [Aufgerufen am 30.03.2015]. 2014 (siehe S. 59).
- [Ste08] Stoyan Stefanov. *Exceptional Website Performance with YSlow 2.0*. <http://de.slideshare.net/stoyan/yslow-20-presentation> [Aufgerufen am 27.02.2015]. Slide Nummer 4. 2008 (siehe S. 9).
- [Ste09] Ph.D. Steve Seow. *User Interface Timing Cheatsheet*. Techn. Ber. <http://tinyurl.com/nbl3cu> [Aufgerufen am 30.03.2015]. Microsoft, 2009 (siehe S. 59).
- [Ste11] Stoyan Stefanov. *Book of Speed*. <http://www.bookofspeed.com/chapter3.html> [Aufgerufen am 02.03.2015]. siehe Abbildung 3.4 - The-three-way handshake. 2011 (siehe S. 6).
- [Ste15] Daniel Stenberg. *The state and rate of http2 adoption*. <http://daniel.haxx.se/blog/2015/03/31/the-state-and-rate-of-http2-adoption/> [Aufgerufen am 01.04.2015]. 2015 (siehe S. 62).
- [t3n15] t3n. *Ist deine Website „mobile-friendly“? – Google steigert Druck auf Webmaster*. <http://t3n.de/news/google-mobile-friendly-589402/> [Aufgerufen am 25.02.2015]. 2015 (siehe S. 3).
- [Tak13] Dean Takahashi. *Apple's iPhone 5 touchscreen is 2.5 times faster than Android devices*. <http://venturebeat.com/2013/09/19/apples-iphone-5-touchscreen-is-2-5-times-faster-than-android-devices/> [Aufgerufen am 03.03.2015]. Grafik. 2013 (siehe S. 10, 13).

- [TNS14] Google TNS Infratest BVDW. *Global Connected Consumer Study*. Website. <http://www.netzproduzenten.de/wp-content/uploads/2014/08/global-connected-consumer-studie-deutschland.pdf> [Aufgerufen am 14.12.2014]. 2014 (siehe S. 3, 4).
- [Ton13] Rmistry und Tonyg. *Loading measurement: alexa top million netsim*. <https://docs.google.com/document/d/1cpLSSYpqI4SprkJcVxbS7af6avKM0qc-imxvkexmCZs/edit> [Aufgerufen am 04.03.2015]. 2013 (siehe S. 11).
- [Twi12] Twitter. *Improving performance on twitter.com*. <https://blog.twitter.com/2012/improving-performance-on-twittercom> [Aufgerufen am 30.03.2015]. 2012 (siehe S. 59).
- [Web08] WebSiteOptimization.com. *The Psychology of Web Performance*. <http://www.websiteoptimization.com/speed/tweak/psychology-web-performance/> [Aufgerufen am 25.02.2015]. 2008 (siehe S. 3).
- [web15] webpagetest.org. *Speed Index*. <https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index> [Aufgerufen am 11.03.2015]. 2015 (siehe S. 24).
- [wha15] whatdoesmysitecost.com. *What Does My Site Cost?* <http://whatdoesmysitecost.com/site/www.hs-weingarten.de> [Aufgerufen am 12.03.2015]. 2015 (siehe S. 26).
- [wik14a] wikipedia. *HTTP-Statuscode*. <http://de.wikipedia.org/wiki/HTTP-Statuscode> [Aufgerufen am 04.03.2015]. 2014 (siehe S. 15).
- [wik14b] wiki.ubuntuusers. *Der Benutzer root*. <http://wiki.ubuntuusers.de/sudo> [Aufgerufen am 26.02.2015]. 2014 (siehe S. 5).
- [Wik15] Wikipedia. *Polyfill*. <http://de.wikipedia.org/wiki/Polyfill> [Aufgerufen am 19.09.2015]. 2015 (siehe S. 42).
- [wik15a] wikipedia. *Content Delivery Network*. http://de.wikipedia.org/wiki/Content_Delivery_Network [Aufgerufen am 04.03.2015]. 2015 (siehe S. 7).
- [wik15b] wikipedia. *Design*. <http://de.wikipedia.org/wiki/Design> [Aufgerufen am 27.03.2015]. 2015 (siehe S. 57).
- [Wil11] Christopher Williams. *The 300m dollar cable that will save traders milliseconds*. <http://www.telegraph.co.uk/technology/news/8753784/The-300m-cable-that-will-save-traders-milliseconds.html> [Aufgerufen am 30.03.2015]. 2011 (siehe S. 59).
- [Yah07] Yahoo. *Performance Research, Part 2: Browser Cache Usage - Exposed!* <http://yuiblog.com/blog/2007/01/04/performance-research-part-2/> [Aufgerufen am 16.03.2015]. 2007 (siehe S. 35).