Client-side Performance of Web-based Applications: the State of the Art

Jasper van Riet
jaspervanriet@gmail.com
Supervisor: Ivano Malavolta

October 2019

1 Introduction

The performance of web pages is an everlasting focus of both academia and industry. The performance of web pages is vital, with even tiny differences in page performance having the potential to significantly impact a company's profit, and the general usage experience [1]. Furthermore, increases in page performance can lead to a decrease in energy usage [2]. Research into the performance of web-based applications has approached the problem from a multitude of angles, from finding ways to more accurately represent the experience of users, and thus allowing for more accurate optimisation approaches, to off-loading portions of the page-load process to proxy servers. Even when two papers use a similar approach, there is the potential for a significant amount of variance in actual method, goal and subject of experimentation. For example, while many studies focus on the Page Load Time (PLT) metric as reported by browsers, others may take an approach they find more perceptive to the experience of users.

This study reviews the work published on the issue of web page performance, with a particular focus on the client-side. Doing so allows for an overview into topics that are commonly studied, and those which are not. Furthermore, this literature study can provide future research with insight into gaps in the literature, and ways to improve on the currently available research.

2 Approach

2.1 Research approach

This literature study focuses on the following research question, which it set out to answer:

What is the state of the art in methods to measure and optimise for client-side performance of web-based applications?

The approach taken by this literature study to answer this question can be summarised as (1) select papers related to web page performance (2) collect data on each paper (3) analyse the characteristics of each paper. Each of these steps will be clarified in the coming sections.

2.2 Selection process

In order to select papers from the literature, a snowballing selection technique was used [3]. The snowballing technique was used in both a backwards and forwards fashion. These results were augmented with manual queries using Google Scholar. The snowballing technique makes use of a start set of papers, which are then used to discover new papers by iterating through the list of references in each of the papers in the start set, called backwards snowballing, and using the "cited by" method in Google Scholar to find papers citing the papers in the start set, called forwards snowballing. Papers are then selected if they fit the selection criteria. See Table 2 for the inclusion and exclusion criteria. Due to the quick development of the field, a relatively recent year was chosen as the cutoff point.

ID	Authors	Title	Venue	Year
ANRW_2019	Konrad Wolsing, Jan Rüth, Klaus Wehrle, Oliver Hohlfeld	A Performance Perspective on Web Optimized Protocol Stacks: TCP+TLS+HTTP/2 vs. QUIC	Applied Networking	2019
ATC_2016	Jamshed Vesuna, Colin Scott, Michael Buettner, Michael Piatek, Arvind Krishnamurthy, Scott Shenker	Caching Doesn't Improve Mobile Web Performance (Much)	USENIX Annual Technical Conference	2016
CAL_2015	Javier Verdu, Alex Pajuelo	Performance Scalability Analysis of JavaScript Applications with Web Workers	IEEE Computer Architecture Letters	2015
ICSE_2016	Marija Selakovic, Michael Pradel	Performance Issues and Optimizations in JavaScript: An Empirical Study	Proceedings of the 38th International Conference on Software Engineering	2016
NSDI_2018	Ravi Netravali, Vikram Nathan, James Mickens, Hari Balakrishnan	Vesper: Measuring Time-to-Interactivity for Modern Web Pages	Symposium on Networked Systems Design and Implementation	2018
PAM_2019	Theresa Enghardt, Thomas Zinner, and Anja Feldmann	Web Performance Pitfalls	International Conference on Passive and Active Network Measurement	2019
WWW_2016	Javad Nejati, Aruna Balasubramanian	An In-depth study of Mobile Browser Performance	Proceedings of the 25th International Conference on World Wide Web	2016
WWW_2018	Maarten Wijnants, Robin Marx, Peter Quax, Wim Lamotte	HTTP/2 Prioritization and its Impact on Web Performance	Proceedings of the 2018 World Wide Web Conference	2018

Table 1: The starting set of papers

Due to the snowballing process' reliance on the starting set, the chosen papers have the potential to have a large amount of influence on the literature study. The starting set can seen in Table 1. The selection process resulted in 21 papers. See Table 10 in the Appendix for the complete list of papers used in this literature study, including the starting set.

Inclusion criteria	Exclusion criteria
A study that is written in English	A study that is still in the preliminary stage of research
A study that is peer reviewed	
A study that is written in or after 2012	
A study that directly proposes a new way of measuring, improving or analyzing the performance of web-based applications, or indirectly studies the effects of such approaches in either the field of performance or a field that is closely related.	

Table 2: The inclusion and exclusion criteria

2.3 Collecting data and categorising papers

Every paper selected was read. Summaries were produced from all papers, reducing them down to their key points. Doing so enabled more efficient work in the next step: categorisation of papers. Each paper was categorised according to a set of criteria. These criteria enable detailed analysis of the overall set of papers. Some of these criteria applied to all papers, whereas other only applied to a subset of papers. For example, some papers focused purely on page performance metrics, and thus did not propose technical contributions.

3 Results

There are several ways to distinguish the papers that were selected on the subject of client-side web application performance. The first is to divide the selection of papers based on their type: papers proposing a new solution, papers presenting a survey of other studies and papers providing a comparison or measurement of existing technologies. This is not a particularly effective method, since papers tend to cross-over in these categories. A paper presenting new results from a measurement may then address found inefficiencies with a newly proposed solution. Another way to distinguish is to divide the papers between four different types of studies, based on their research focus: page performance metrics, web protocols, web languages and web page performance. While this distinction is not perfect; certain papers will appear in multiple buckets, it allows for closer analysis of a subset of papers, without wading through the intrinsic complexity of the collection of papers in total. Through the following Sections 3.1, 3.2, 3.3 and 3.4, these subsets will be explored. Afterwards, this literature study will zoom out and look at the overall collection of papers, and its characteristics.

3.1 Papers focused on performance metrics

From the complete set of papers, six focus on page performance metrics. Page performance metrics play an important role in the process of improving page performance. The metrics used in research provide a target to aim towards. If research is optimizing towards the wrong goal, then its value is lessened. See Table 3 for a list of papers in this category. From now on, these papers will be referred to by their ID. From this list of papers, one is a survey of other papers, PAM_2019 [4]. PAM_2019 studies papers in the literature and tests them for their reliability and consistency. Its findings show that the PLT is often not documented well in research. In particular, authors tend to not mention when their measurement of PLT is started.

ID	Authors	Title	Venue	Year
CCR_2016	Enrico Bocchi, Luca de Cicco, Dario Rossi	Measuring the Quality of Experience of Web users	ACM SIGCOMM Computer Communication Review	2016
CCR_2017	Qingzhu Gao, Prasenjit dey, Parvez Ahammmad	Perceived Performance of Top Retail Webpages In the Wild	ACM SIGCOMM Computer Communication Review	2017
NSDI_2017	Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian	Improving User Perceived Page Load Times Using Gaze	Symposium on Networked Systems Design and Implementation	2017
NSDI_2018	Ravi Netravali, Vikram Nathan, James Mickens, Hari Balakrishnan	Vesper: Measuring Time-to-Interactivity for Modern Web Pages	Symposium on Networked Systems Design and Implementation	2018
PAM_2019	Theresa Enghardt, Thomas Zinner, and Anja Feldmann	Web Performance Pitfalls	International Conference on Passive and Active Network Measurement	2019
QoMEX_2018	Tobias Hoßfeld, Florian Metzger, Dario Rossi	Speed Index: Relating the Industrial Standard for User Perceived Web Performance to Web QoE	International Conference on Quality of Multimedia Experience	2018

Table 3: Literature on performance metrics

Table 4 provides an overview of the research aim and approach of the remaining five papers that are not surveys. A number of patterns are immediately apparent. Firstly, all papers involved provide a comparison of page performance metrics. Due to the nature of the subject matter, this seems logical. Further, in the two cases where an entirely new metric was proposed, a method to optimise for this new metric was also presented.

ID	Research Aim and Research Approach				
	Research Aim			Research Approach	
	Propose a new metric	Improve a current metric	Provide a way to optimize for a metric	Perform a user study	Compare relationship between metrics
CCR_2016		X			X
CCR_2017		X		X	X
NSDI_2017	X		X	X	X
NSDI_2018	X		X	X	X
QoMEX_2018					X

Table 4: The research aim and approach of the five papers focused on performance metrics, excluding surveys

NSDI_2017 [5] follows this pattern by first defining a metric representing the experience of a user during the web page load, and then optimising for this metric by using gaze tracking to optimise those parts of the page that the user pays attention to in the loading process. NSDI_2018 [6] uses what has been learned from SpeedIndex (SI), a metric that uses an integral formula over the loading progress of a page to give a measure to the progressive nature of the loading process, and then proposes a new metric. Its metric, Ready Index (RI), captures not only the visual progression of the page, but also its interactivity state. In other words, not only should the page be rendered, it should also be accepting of input. This metric is then shown in a user study to more accurately represent the experience of users.

The authors present a way to optimise for RI by designing a custom HTTP/2 scheduler that prioritises resources that contribute to the interactivity of the page.
Further of note is the prevalence of user studies in informing research. Of the three papers that performed a user study, all three went on to provide contributions to the community in the form of either a new metric, or an improvement on currently available metrics. All three user studies result in a new performance metric, with the authors concluding from the user study that PLT is not accurate to the user perception. In the case of CCR_2017 [7], the paper proposes a complementary metric to SI, the Perceptual Speed Index (PSI). The PSI complements the SI in that while the SI provides a measure of how quickly the visible content (above the fold) of the web page loads, the PSI provides a measure of whether this loading is done in a smooth manner, without any noticeable visual jumps or jitter. Meanwhile, the other paper focusing on improving metrics, CCR_2016 [8], does not focus on a conceptual contribution to the current thinking on metrics. Instead, CCR_2016 analyses the computational complexity of SI, and proposes two new metrics which combined will aid the same conclusions as SI by itself. These two metrics, the ByteIndex (BI) and the ObjectIndex (OI) are both computed using metrics available from the browser, such as the percentage of objects received at time t, as opposed to using a video recording of the loading process, as SI does.
QoMEX_2018 [9] does not perform a user study, but also reaches the conclusion that the PLT is not accurate as to the experience of users. It uses both the PLT and SI to predict Mean Opinion Scores (MOS). It finds that the model using the PLT is unable to accurately predict the MOS. By contrast, the SI performs much better. It finds that integral metrics such as the SI, but also the BI, are more accurate in predicting the MOS, and thus are more appropiate as a measure of user experience.

3.2 Papers focused on web protocols

Four papers focus on the web protocols used in the modern-day, to measure their contribution to the page load process. Of particular mention are HTTP/2 and QUIC [10]. See Table 5 for the papers in the set. Each of the four papers seeks to measure the performance of protocols. All 4 papers were published in the two years before this literature study. Collectively, the papers find usage of QUIC to be low. NTMS_2019 [11] reports that QUIC usage is mainly isolated to Google servers, by measuring web pages in an automated fashion for a year. Further, it finds that HTTP/2 delivers 63% of resources, a number that has gone up just 4 percent points in a year. Additionally, WWW_2019 [12] finds, using browsing data from 2 million pageviews, that in cases where QUIC is used, it often has to default back to HTTP/2 or HTTP/1.1 for certain resources, limiting any potential performance improvements QUIC has to offer. Thus, the paper finds, in many cases there is not much difference between QUIC and HTTP/2. ANRW_2019 [13] tests this same assertion by tuning TCP parameters to be more similar to QUIC's defaults, and then comparing whether this makes for a noticable improvement. Its findings show that a tuned TCP outperforms a default TCP by a significant margin, although QUIC still wins out, albeit by a more narrow gap.

ID	Authors	Title	Venue	Year
ANRW_2019	Konrad Wolsing, Jan Rüth, Klaus Wehrle, Oliver Hohlfeld	A Performance Perspective on Web Optimized Protocol Stacks: TCP+TLS+HTTP/2 vs. QUIC	Applied Networking	2019
NTMS_2019	Antoine Saverimoutou, Bertrand Mathieu, Sandrine Vaton	Influence of Internet Protocols and CDN on Web Browsing	International Conference on New Technologies, Mobility and Security	2019
WWW_2018	Maarten Wijnants, Robin Marx, Peter Quax, Wim Lamotte	HTTP/2 Prioritization and its Impact on Web Performance	Proceedings of the 2018 World Wide Web Conference	2018
WWW_2019	Mohammad Rajiullah, Andra Lutu, Ali Safari Khatouni, Mah-rukh Fida, Marco Mellia, Özgü Alay, Anna Brunstrom, Stefan Alfredsson, Vincenzo Mancuso	Web Experience in Mobile Networks: Lessons from Two Million Page Visits	The World Wide Web Conference	2019

Table 5: Literature on web protocols

The final paper of this batch, WWW_2018 [14], investigates how HTTP/2 fares against HTTP/1.1, and whether browsers make optimal use of HTTP/2 prioritisation schemes. It finds that HTTP/2 significantly outperforms HTTP/1.1, although many browsers do not yet make efficient use of its prioritisation capabilities. While browsers tend to use their own custom prioritisation schemes, these do not always outperform naive schemes such as First-Come-First-Serve (FCFS), particularly in the case of smaller web pages.

ID	Authors	Title	Venue	Year
CAL_2015	Javier Verdu, Alex Pajuelo	Performance Scalability Analysis of JavaScript Applications with Web Workers	IEEE Computer Architecture Letters	2015
ICSE_2016	Marija Selakovic, Michael Pradel	Performance Issues and Optimizations in JavaScript: An Empirical Study	Proceedings of the 38th International Conference on Software Engineering	2016
ICSE_2017	Zheng Gao, Christian Bird, Earl T. Barr	To Type or Not to Type: Quantifying Detectable Bugs in JavaScript	Proceedings of the 39th International Conference on Software Engineering	2017

Table 6: Literature on JavaScript

3.3 Papers focused on programming language

Three papers selected focus specifically on JavaScript. These papers can be seen in Table 6. Two of the papers address the problem of detecting and fixing bugs found in code, and both do so by analysing the fixes to bugs done in open-source JavaScript projects. ICSE_2016 [15] does so by analysing performance issues filed in open-source projects, and then analysing how these were fixed, in order to find patterns. From their findings, the authors conclude that many performance problems are similar in nature, for example: 52% of the performance issues found are caused by the inefficient usage of an API. Thus, these issues can also be solved in a similar way to each other. ICSE_2017 [16]

focuses on the improvements that could be made by using a typing system, and find that 15% of the bug fixes they find could have been found earlier in the process via a typing system, such as Microsoft's TypeScript. CAL_2015 [17] takes a different approach than the previous two papers. It studies the scalability of JavaScript's Web Workers. JavaScript Web Workers act as a separate thread, not sharing any state with the main thread, and thus there is a particular interest in how well these scale with an increase in processor cores, or on a particularly busy machine. The paper finds that the optimal number of web workers is highly dependent on a number of factors, including the number of cores, the busyness of the system, and the browser.

ID	Authors	Title	Venue	Year
ATC_2016	Jamshed Vesuna, Colin Scott, Michael Buettner, Michael Piatek, Arvind Krishnamurthy, Scott Shenker	Caching Doesn't Improve Mobile Web Performance (Much)	USENIX Annual Technical Conference	2016
MOBILESoft_2017	Ivano Malavolta, Guiseppe Procaccianti, Paul Noorland, Petar Vukmirović	Assessing the Impact of Service Workers on the Energy Efficiency of Progressive Web Apps	International Conference on Mobile Software Engineering and Systems	2017
NSDI_2013	Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, David Wetherall	Demystifying Page Load Performance with WProf	Symposium on Networked Systems Design and Implementation	2013
NSDI_2015	Victor Agababov, Michael Buettner, Victor Chudnovsky, Mark Cogan, Ben Greenstein, Shane McDaniel, and Michael Piatek, Colin Scott, Matt Welsh, Bolian Yin	Flywheel: Google's Data Compression Proxy for the Mobile Web	Symposium on Networked Systems Design and Implementation	2015
NSDI_2016	Xiao Sophia Wang, Arvind Krishnamurthy, David Wetherall	Speeding up Web Page Loads with Shandian	Symposium on Networked Systems Design and Implementation	2016
NSDI_2017	Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian	Improving User Perceived Page Load Times Using Gaze	Symposium on Networked Systems Design and Implementation	2017
NSDI_2018	Ravi Netravali, Vikram Nathan, James Mickens, Hari Balakrishnan	Vesper: Measuring Time-to-Interactivity for Modern Web Pages	Symposium on Networked Systems Design and Implementation	2018
NSDI_2018_2	Ravi Netravali, James Mickens	Prophecy: Accelerating Mobile Page Loads Using Final-state Write Logs	Symposium on Networked Systems Design and Implementation	2018
WWW_2012	Zhen Wang, Felix Xiaozhu Lin, Lin Zhong, and Mansoor Chishtie	How Far Can Client-Only Solutions Go for Mobile Browser Speed?	International conference on World Wide Web	2012
WWW_2016	Javad Nejati, Aruna Balasubramanian	An In-depth study of Mobile Browser Performance	Proceedings of the 25th International Conference on World Wide Web	2016
WWW_2019	Mohammad Rajiullah, Andra Lutu, Ali Safari Khatouni, Mah-rukh Fida, Marco Mellia, Özgü Alay, Anna Brunstrom, Stefan Alfredsson, Vincenzo Mancuso	Web Experience in Mobile Networks: Lessons from Two Million Page Visits	The World Wide Web Conference	2019

Table 7: Literature on web page performance

3.4 Papers focused on web page performance

Papers focusing on the general field of web page performance consist of a significant subset of the complete selection of papers. For the papers fitting this category, see Table 7. In all, eleven papers fit this description.

While there is a variety of topics to be found in the literature, a number of papers share their research topics within the field of web page performance. This distribution of topics can be found in Table 8. The acronym PWA stands for a Progressive Web Application, a technology advocated by Google utilising the concept of a service worker, a separate JavaScript thread that is capable of handling functionality such as push notifications, or preloading assets. MOBILESoft_2017 [2] asks the question what impact this may have on the energy efficiency. The authors found that the impact of service workers on the energy efficiency of a PWA is not significant, even with different qualities of signal.

NSDI_2013 [18] presents the tool WProf, a tool that allows instrumentation of the browser in order to discover the dependencies that make up the web page load process. Using this tool, the authors find that due to the shared state of the DOM, HTML parsing and JavaScript evaluation can result in a significant amount of blocking. Further of note is its finding that only 20% of objects loads can be found on the critical path, thus fundamentally reducing the effectiveness of caching. WWW_2016 [19] iterates upon this work, and present WProf-M, a mobile version of WProf. It finds that unlike desktop devices, computation is the key bottleneck for mobile browser performance. On desktop devices, network is the bottleneck. It further finds that only 20% of the critical path is shared between mobile and desktop devices.

ATC_2016 [20] provides further proof of the lack of effectiveness of caching on mobile devices, underpinning the conclusion that a lack of objects on the critical path reduces its effectiveness, in addition to slower CPU speeds. WWW_2012 [21] utilizes a dataset of 24 iPhone users over the span of a year to document the same fact, demonstrating the lack of effectiveness of both caching and prefetching. Instead, it proposes predicting subresources on a web page, which it shows can achieve a much higher hit ratio.

ID	Research topic					
	Caching/prefetching	Pre-processing	Dependencies	PWAs	Signal quality	HTTP
ATC_2016	X					
MOBILESoft_2017				X	X	
NSDI_2013	X		X			
NSDI_2015		X				
NSDI_2016		X				
NSDI_2017			X			X
NSDI_2018			X			X
NSDI_2018_2		X				
WWW_2012	X					
WWW_2016			X			
WWW_2019					X	

Table 8: Research topics of papers on web page performance

Another popular approach to improving the performance of, in particular, mobile devices in this subset of papers is the usage of a proxy server to pre-process web pages, thereby reducing the aforementioned computation cost for the mobile device. One of such proposals can be found in NSDI_2016, in which the presented framework Shandian is able to identify the vital parts of the initial state of a web page, and thus only transfer the critical sections. After the initial load, the rest of the state is loaded in. This allows for a 60% reduction in PLT in the median case. This significant reduction is caused by the almost elimination of the performance bottleneck: computation on mobile devices. The approach taken with Shandian is criticised by NSDI_2018_2 [22], which presents a similar framework given the name Prophecy. The authors men-

tion Shandian is reliant on a custom browser, and goes less far than Prophecy in pre-processing. In the case of Prophecy, the entirety of the DOM tree and JavaScript heap state is computed beforehand, and then transmitted to the client using so-called write logs. The client will then reconstruct the state using these logs. Prophecy records a 53% reduction in PLT in the median case.

While NSDI_2015 [23] also describes such a pre-processing server, there are two major differences. First, its goal is not performance, but minimising data usage. Second, the service has been in use inside the Chrome browser on Android and iOS. The pre-processing server attempts to optimise requested web pages by applying data usage optimisations, such as image compression, minification and more. By doing so, it trades off a higher PLT in most cases for, on average, a 58% reduction in data transferred to mobile devices.

Two papers mentioned earlier in Section 3.1, NSDI_2017 and NSDI_2018, both make use of HTTP/2 techniques, Server Push and scheduling of object requests respectively. Both use their respective techniques to implement prioritisation according to certain criteria. In the former's case, this is according to a user's gaze and in the latter case, to optimise for the paper's proposed metric RI.

Finally, in addition to its protocol findings described in Section 3.2, WWW_2019 describes the significance of signal quality on web page performance. The paper finds a median PLT worsening by 36,5% when switching from a wired to a mobile broadband connection. Furthermore, in some cases a handover between technologies such as 3G and 4G can cause up to five times slower page load times.

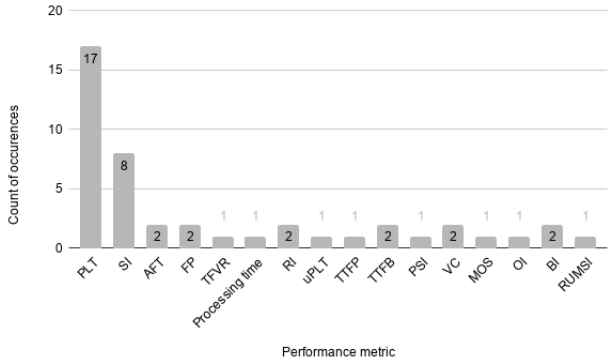

Figure 1: Performance metrics used in all papers

3.5 Overall

The overall picture of the selection of papers is murky, with a great variety in topics and methods used. The entire range of research topics featured in the collection of papers can be found in Table 9. Clear is the common practice of measuring performance as part of the paper, yet also the large diversity that is present within this field.

This amount of diversity in topics leads to the question of how consistent the methods used in each paper are with other papers in the field. For example, what performance metrics are used to measure the page load time? The answer to this question can be found in Figure 1. Both PLT and SI seem be significantly more common than other metrics, with PLT being used in every single paper that performs performance measurements. In six papers, only PLT is used, without any other performance metric. All in all, sixteen different performance metrics are used across the literature featured in this study. Of these sixteen, five originate from papers in this study. Of these five, only 2, BI and RI, were mentioned in other papers, and both in just one other paper.

The architectural scope of papers is another point of interest. In terms of web applications, this scope consists of two parts: network and computation. Resources are fetched by the client and then parsed in order to be rendered. The particular architectural scope of papers can be seen in Figure 2. No mention of the performance aspect of the rendering pipeline in browsers was found, in any of the papers.

Further of interest is the dataset used by these papers. If all papers were to use the exact same set of web pages to optimise page load performance on, that could be problematic. Of the 21 papers, of which 17 perform performance measurements, 13 use Alexa top lists as a resource for web pages. There is a variety in which exact top list is used, although 9 out of 13 use the full or a subset of the top 500 list, with one additional paper leaving the exact list unspecified. Furthermore, two papers also employ a top list from Moz, and the remaining papers use unique datasets, such as the LiveLab dataset [24], a collection of web browsing data collected from iPhone users over the span of a year.

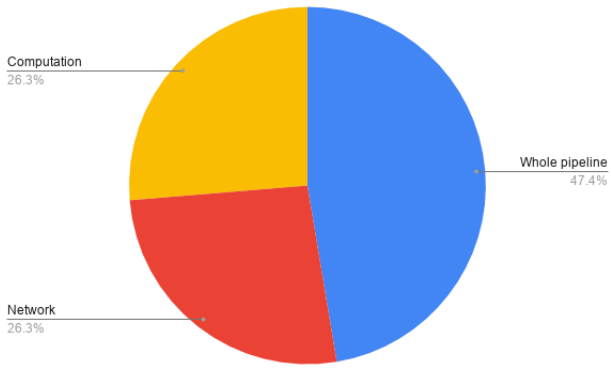

Figure 2: Architectural scope of all papers performing performance measurements

Five papers present performance solutions. Of these five, two papers present a solution that uses pre-processing on a server, both NSDI_2016 [25] and NSDI_2018_2. Two papers are reliant on browser changes, namely NSDI_2016 and WWW_2012, which proposes a solution using speculative loading of subresources for a web domain by assuming that subresources loaded within a domain tend to be shared between pages. The papers shows that this achieves a much

ID	Research topic												
	Caching/prefetching	Pre-processing	Dependencies	PWAs	Signal quality	Browser	Measure performance	Page load metrics	JavaScript	QUIC	HTTP	Web Workers	Improve performance
ANRW_2019					x		x			x	x		
ATC_2016	x						x						
CAL_2015							x		x			x	
CCR_2016							x	x					
CCR_2017							x	x					
ICSE_2016									x				
ICSE_2017									x				
MOBILESoft_2017				x	x				x				
NSDI_2013	x		x				x						
NSDI_2015		x				x	x						
NSDI_2016		x				x	x						x
NSDI_2017			x		x		x	x			x		x
NSDI_2018			x				x	x			x		x
NSDI_2018_2		x					x						x
NTMS_2019							x			x	x		
PAM_2019								x					
QoMEX_2018							x	x					
WWW_2012	x					x	x						x
WWW_2016			x				x						
WWW_2018						x	x				x		
WWW_2019					x		x			x	x		

Table 9: Research topics of all papers

higher hit ratio than either caching or prefetching. The final two papers, NSDI_2017 and NSDI_2018 both attempt to optimise for a certain metric, uPLT and RI respectively, by utilizing HTTP/2 techniques. The former does so via gaze tracking, while the latter optimizes for interactivity. Both analyse dependencies to determine which objects should be prioritised.

For each of these papers, reproducibility is important. Reproducibility allows for reproduction of results, and thus verification of those results. Specific to the field of performance research, this means that papers should report as many of the variables involved in the research as possible. Important variables can be the CPU used in the device, the amount of RAM available, and the specific browser that was used. Thus, it's important to note that in six of the papers featured here, authors failed to report the versions of the browsers that were used in experimentation. Updates for browsers such as Google's Chrome are released as often as every month [26], meaning results that were produced at the start of research may not be reproducible at the end of it. Of the six papers, one paper did not report browsers used at all: MOBILESoft_2017.

3.6 Validity of results

3.6.1 Paper selection

Due to the method used to find papers, the snowball method along with manual queries, the initial selection of papers is key to the resulting selection. Verification of the selection was done via snowballing in a backwards fashion on the paper selection and verifying that the papers found were part of the selection already. Furthermore, the literature featured in this study is limited to the year 2012 and after, however, due to the fast developing pace of this field there is ample reason to do so.

3.6.2 Lack of coverage of industry

Due to the scope of this literature study consisting of academia only, literature outside of academia that is relevant may have been missed. The field of web application performance has many stakeholders, and is thus not only covered in academia. Topics that are observed to not be covered extensively in academia may have the underlying reason that these were covered in the industry instead, and thus would not be covered in this literature study.

3.6.3 Categorising the selection of papers

Categorisation of the papers featured in this literature study was done with care, however, due to the immense diversity of the field, this was often hard to accomplish. Papers may fit many buckets at one time, complicating the process. The initial categories were extended during the process of data collection after certain patterns amongst papers became apparent. In order to perform categorisation with these extended categories, papers were re-read and re-categorised, with some amount of potential for inaccuracies being introduced as a result.

4 Discussion

This study presents a complex picture of the literature available, with a great variety of topics being covered. The measurement of performance, either current technologies or proposed, lies without a doubt at the core of the field. In all three papers that performed user studies featured as part of this literature study, the PLT was found to have a poor correlation with the user perception of the page load process. One may find it surprising then, to see the field's reliance on the PLT metric, with all papers performing performance measurements showing its result. Eight papers make use of the SI, which seems to be a metric more closely aligned with the experience of users, as measured by QoMEX_2018, yet this is still a point of debate, as CCR_2017 found SI to have no significant correlation

with the actual user perception. Thus, for now, the SI is not the definitive answer to a search for a better page performance metric. Problematically, while the field does propose new metrics, with five metrics being proposed by papers featured in this study, none seem to garner significant uptake amongst the community.
While the literature has a relatively well-spread breadth of topics, it should be noted that the number of proposed performance optimisations is a relatively small part of this study. Furthermore, a notable absentee in the list of topics is the time taken up by the rendering pipeline in browsers. Browsers in general show themselves to be a small part of the field, with just 3 papers proposing any changes that would improve the performance of browsers. Further gains in the field are to be made with a greater amount of attention on reproducibility, with six papers showing themselves to be problematic by not providing browser versions used in experiments. Providing better reproducibility would benefit the field in a myriad of ways, and should be a target for research.

5 Conclusion

This study of literature features 21 different papers on the field of web application performance. It does so in order to deliver an overview of the state of the art in this field. In order to establish the state of the art in research on web application performance, papers were carefully selected and categorised. A bisection of the research was created, showcasing the breadth of the field of web application performance. The state of the art seems diverse, and solutions wide-ranging. The number of performance optimisations proposed in this field seems relatively small, with just five papers contributing solutions. Per contrast, seventeen papers provide a performance comparison of technologies, showcasing that to be the primary focus of the research. A greater amount of attention should go towards reproducibility, with six papers containing issues in this department. Future research would be wise to pay attention to the datasets used, and make sure the current literature's reliance on Alexa's Top 500 list does not become more dominant. Finally, research should be aware of the lacking correlation between the PLT and the user experience. Performance metrics more closely aligned to the user perception have been proposed, yet are rarely used.

References

[1] K. Eaton, "How one second could cost amazon $1.6 billion in sales," Jul 2012. [Online]. Available: https://www.fastcompany.com/1825005/how-one-second-could-cost-amazon-16-billion-sales

[2] I. Malavolta, G. Procaccianti, P. Noorland, and P. Vukmirović, "Assessing the impact of service workers on the energy efficiency of progressive web apps," in *Proceedings of the 4th International Conference on Mobile Software Engineering and Systems*. IEEE Press, 2017, pp. 35–45.

[3] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. Citeseer, 2014, p. 38.

[4] T. Enghardt, T. Zinner, and A. Feldmann, "Web performance pitfalls," in *International Conference on Passive and Active Network Measurement*. Springer, 2019, pp. 286–303.

[5] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving user perceived page load times using gaze," in *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, 2017, pp. 545–559.

[6] R. Netravali, V. Nathan, J. Mickens, and H. Balakrishnan, "Vesper: Measuring time-to-interactivity for modern web pages," in *USENIX NSDI*, 2018.

[7] Q. Gao, P. Dey, and P. Ahammad, "Perceived performance of top retail webpages in the wild: Insights from large-scale crowdsourcing of above-the-fold qoe," in *Proceedings of the Workshop on QoE-based Analysis and Management of Data Communication Networks*. ACM, 2017, pp. 13–18.

[8] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of web users," *ACM SIGCOMM Computer Communication Review*, vol. 46, no. 4, pp. 8–13, 2016.

[9] T. Hoβfeld, F. Metzger, and D. Rossi, "Speed index: Relating the industrial standard for user perceived web performance to web qoe," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[10] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang, F. Yang, F. Kouranov, I. Swett, J. Iyengar *et al.*, "The quic transport protocol: Design and internet-scale deployment," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*. ACM, 2017, pp. 183–196.

[11] A. Saverimoutou, B. Mathieu, and S. Vaton, "Influence of internet protocols and cdn on web browsing," 2019.

[12] M. Rajiullah, A. Lutu, A. S. Khatouni, M.-R. Fida, M. Mellia, A. Brunstrom, O. Alay, S. Alfredsson, and V. Mancuso, "Web experience in mobile networks: Lessons from two million page visits," in *The World Wide Web Conference*. ACM, 2019, pp. 1532–1543.

[13] K. Wolsing, J. Rüth, K. Wehrle, and O. Hohlfeld, "A performance perspective on web optimized protocol stacks: Tcp+ tls+ http/2 vs. quic," *arXiv preprint arXiv:1906.07415*, 2019.

[14] M. Wijnants, R. Marx, P. Quax, and W. Lamotte, "Http/2 prioritization and its impact on web performance," in *Proceedings of the 2018 World Wide Web Conference*. International World Wide Web Conferences Steering Committee, 2018, pp. 1755–1764.

[15] M. Selakovic and M. Pradel, "Performance issues and optimizations in javascript: an empirical study," in *Proceedings of the 38th International Conference on Software Engineering*. ACM, 2016, pp. 61–72.

[16] Z. Gao, C. Bird, and E. T. Barr, "To type or not to type: quantifying detectable bugs in javascript," in *Proceedings of the 39th International Conference on Software Engineering*. IEEE Press, 2017, pp. 758–769.

[17] J. Verdú and A. Pajuelo, "Performance scalability analysis of javascript applications with web workers," *IEEE Computer Architecture Letters*, vol. 15, no. 2, pp. 105–108, 2015.

[18] X. S. Wang, A. Balasubramanian, A. Krishnamurthy, and D. Wetherall, "Demystifying page load performance with wprof," in *Presented as part of the 10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, 2013, pp. 473–485.

[19] J. Nejati and A. Balasubramanian, "An in-depth study of mobile browser performance," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 1305–1315.

[20] J. Vesuna, C. Scott, M. Buettner, M. Piatek, A. Krishnamurthy, and S. Shenker, "Caching doesn't improve mobile web performance (much)," in *2016 {USENIX} Annual Technical Conference ({USENIX}{ATC} 16)*, 2016, pp. 159–165.

[21] Z. Wang, F. X. Lin, L. Zhong, and M. Chishtie, "How far can client-only solutions go for mobile browser speed?" in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 31–40.

[22] R. Netravali and J. Mickens, "Prophecy: accelerating mobile page loads using final-state write logs," in *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*, 2018, pp. 249–266.

[23] V. Agababov, M. Buettner, V. Chudnovsky, M. Cogan, B. Greenstein, S. McDaniel, M. Piatek, C. Scott, M. Welsh, and B. Yin, "Flywheel: Google's data compression proxy for the mobile web," in *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, 2015, pp. 367–380.

[24] C. Shepard, A. Rahmati, C. Tossell, L. Zhong, and P. Kortum, "Livelab: measuring wireless networks and smartphone users in the field," *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 3, pp. 15–20, 2011.

[25] X. S. Wang, A. Krishnamurthy, and D. Wetherall, "Speeding up web page loads with shandian," in *13th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 16)*, 2016, pp. 109–122.

[26] [Online]. Available: https://www.chromestatus.com/features/schedule

Appendix

Filename	Authors	Title	Venue	Year
ANRW_2019	Konrad Wolsing, Jan Rüth, Klaus Wehrle, Oliver Hohlfeld	A Performance Perspective on Web Optimized Protocol Stacks: TCP+TLS+HTTP/2 vs. QUIC	Applied Networking Research Workshop	2019
ATC_2016	Jamshed Vesuna, Colin Scott, Michael Buettner, Michael Piatek, Arvind Krishnamurthy, Scott Shenker	Caching Doesn't Improve Mobile Web Performance (Much)	USENIX Annual Technical Conference	2016
CAL_2015	Javier Verdu, Alex Pajuelo	Performance Scalability Analysis of JavaScript Applications with Web Workers	IEEE Computer Architecture Letters	2015
CCR_2016	Enrico Bocchi, Luca de Cicco, Dario Rossi	Measuring the Quality of Experience of Web users	ACM SIGCOMM Computer Communication Review	2016
CCR_2017	Qingzhu Gao, Prasenjit dey, Parvez Ahammmad	Perceived Performance of Top Retail Webpages In the Wild	ACM SIGCOMM Computer Communication Review	2017
ICSE_2016	Marija Selakovic, Michael Pradel	Performance Issues and Optimizations in JavaScript: An Empirical Study	Proceedings of the 38th International Conference on Software Engineering	2016
ICSE_2017	Zheng Gao, Christian Bird, Earl T. Barr	To Type or Not to Type: Quantifying Detectable Bugs in JavaScript	Proceedings of the 39th International Conference on Software Engineering	2017
MOBILESoft_20	Ivano Malavolta, Guiseppe Procaccianti, Paul Noorland, Petar Vukmirović	Assessing the Impact of Service Workers on the Energy Efficiency of Progressive Web Apps	International Conference on Mobile Software Engineering and Systems	2017
NSDI_2013	Xiao Sophia Wang, Aruna Balasubramanian, Arvind Krishnamurthy, David Wetherall	Demystifying Page Load Performance with WProf	Symposium on Networked Systems Design and Implementation	2013
NSDI_2015	Victor Agababov, Michael Buettner, Victor Chudnovsky, Mark Cogan, Ben Greenstein, Shane McDaniel, and Michael Piatek, Colin Scott, Matt Welsh, Bolian Yin	Flywheel: Google's Data Compression Proxy for the Mobile Web	Symposium on Networked Systems Design and Implementation	2015
NSDI_2016	Xiao Sophia Wang, Arvind Krishnamurthy, David Wetherall	Speeding up Web Page Loads with Shandian	Symposium on Networked Systems Design and Implementation	2016
NSDI_2017	Conor Kelton, Jihoon Ryoo, Aruna Balasubramanian	Improving User Perceived Page Load Times Using Gaze	Symposium on Networked Systems Design and Implementation	2017
NSDI_2018	Ravi Netravali, Vikram Nathan, James Mickens, Hari Balakrishnan	Vesper: Measuring Time-to-Interactivity for Modern Web Pages	Symposium on Networked Systems Design and Implementation	2018
NSDI_2018_2	Ravi Netravali, James Mickens	Prophecy: Accelerating Mobile Page Loads Using Final-state Write Logs	Symposium on Networked Systems Design and Implementation	2018
NTMS_2019	Antoine Saverimoutou, Bertrand Mathieu, Sandrine Vaton	Influence of Internet Protocols and CDN on Web Browsing	International Conference on New Technologies, Mobility and Security	2019
PAM_2019	Theresa Enghardt, Thomas Zinner, and Anja Feldmann	Web Performance Pitfalls	International Conference on Passive and Active Network Measurement	2019
QoMEX_2018	Tobias Hoßfeld, Florian Metzger, Dario Rossi	Speed Index: Relating the Industrial Standard for User Perceived Web Performance to Web QoE	International Conference on Quality of Multimedia Experience	2018
WWW_2012	Zhen Wang, Felix Xiaozhu Lin, Lin Zhong, and Mansoor Chishtie	How Far Can Client-Only Solutions Go for Mobile Browser Speed?	International conference on World Wide Web	2012
WWW_2016	Javad Nejati, Aruna Balasubramanian	An In-depth study of Mobile Browser Performance	Proceedings of the 25th International Conference on World Wide Web	2016
WWW_2018	Maarten Wijnants, Robin Marx, Peter Quax, Wim Lamotte	HTTP/2 Prioritization and its Impact on Web Performance	Proceedings of the 2018 World Wide Web Conference	2018
WWW_2019	Mohammad Rajiullah, Andra Lutu, Ali Safari Khatouni, Mah-rukh Fida, Marco Mellia, Özgü Alay, Anna Brunstrom, Stefan Alfredsson, Vincenzo Mancuso	Web Experience in Mobile Networks: Lessons from Two Million Page Visits	The World Wide Web Conference	2019

Table 10: List of papers used