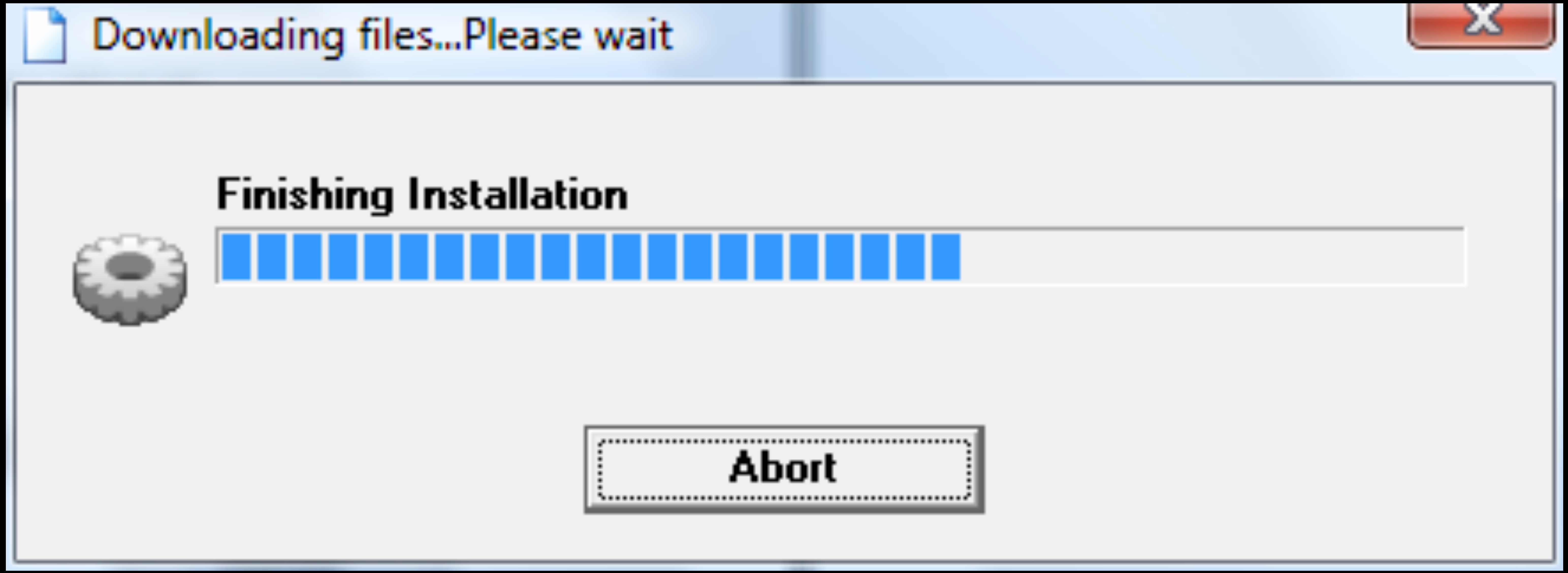
To Push, or Not to Push

Mark Nottingham, fastly

@mnot

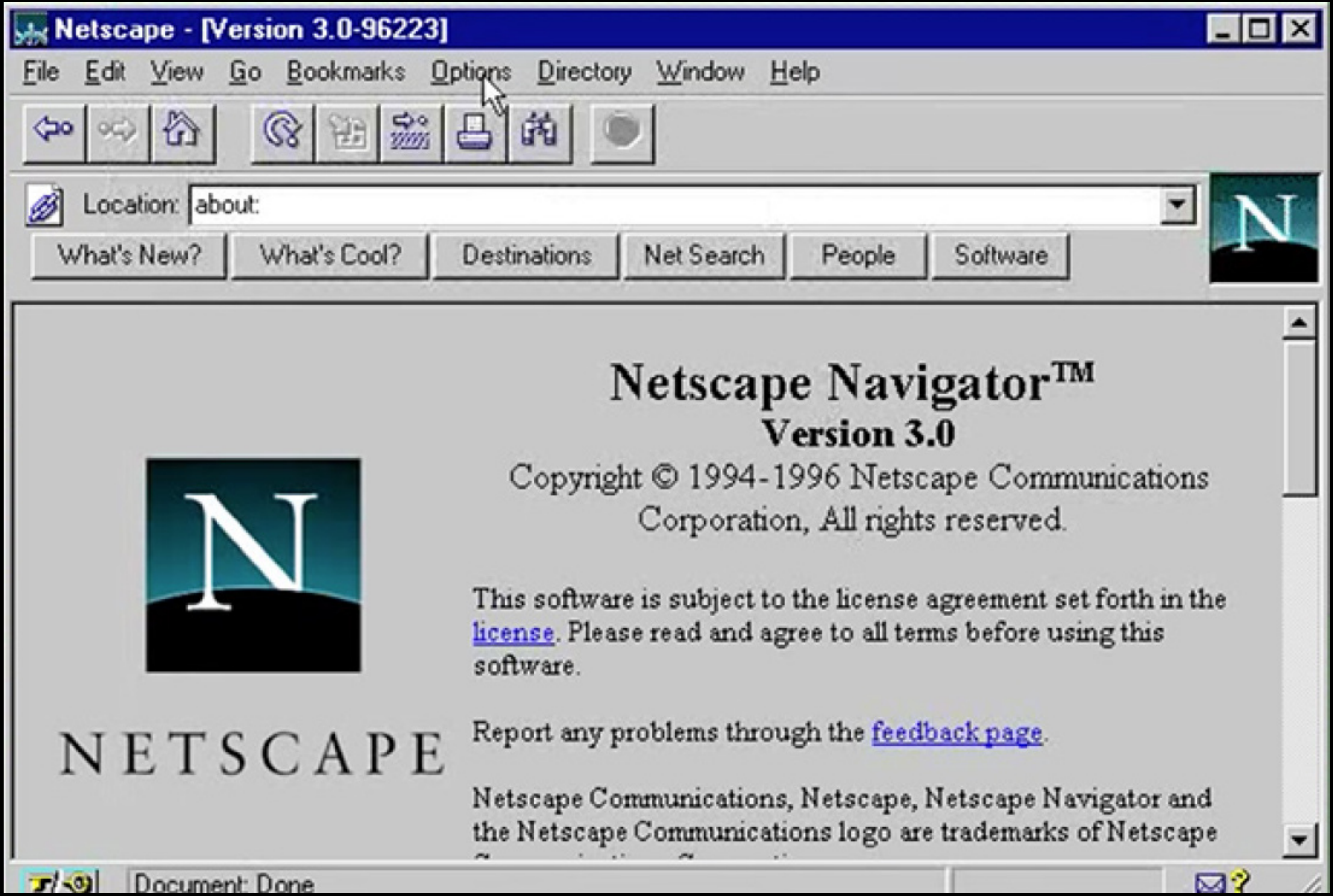
In the beginning,
We downloaded software.



Downloading files...Please wait

Finishing Installation

Then came the Web



Netscape Navigator™

Version 3.0

Copyright © 1994-1996 Netscape Communications Corporation, All rights reserved.

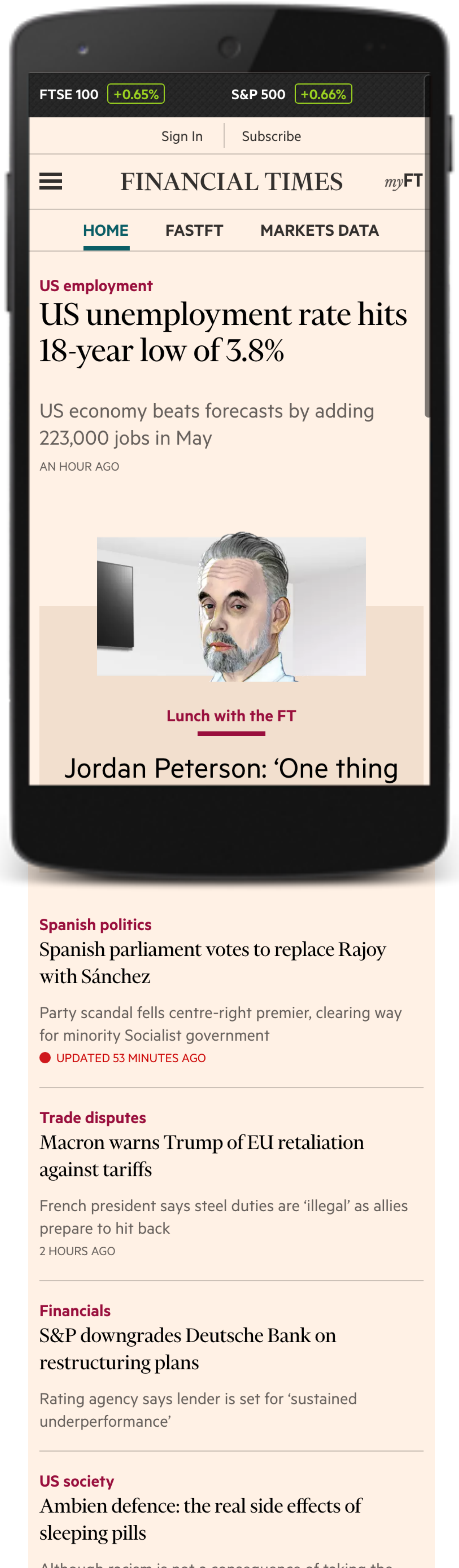
This software is subject to the license agreement set forth in the license. Please read and agree to all terms before using this software.

Report any problems through the feedback page.

Netscape Communications, Netscape, Netscape Navigator and the Netscape Communications logo are trademarks of Netscape

```
<html>
  <head>
    <link rel="stylesheet" href="/sty
    <script src="/script.js"></script
  </head>
  <body>
    <h1>Hello</h1>
    <img src="hello.jpg"/>
    <img src="other.gif"/>
```

Resources can be prioritised.



US employment

US unemployment rate hits 18-year low of 3.8%

US economy beats forecasts by adding 223,000 jobs in May

AN HOUR AGO

Lunch with the FT

Jordan Peterson: 'One thing

Spanish politics

Spanish parliament votes to replace Rajoy with Sánchez

Party scandal fells centre-right premier, clearing way for minority Socialist government

● UPDATED 53 MINUTES AGO

Trade disputes

Macron warns Trump of EU retaliation against tariffs

French president says steel duties are 'illegal' as allies prepare to hit back

2 HOURS AGO

Financials

S&P downgrades Deutsche Bank on restructuring plans

Rating agency says lender is set for 'sustained underperformance'

US society

Ambien defence: the real side effects of sleeping pills

Although racism is not a consequence of taking the

Resources are often shared between pages.

Why Web Directions Summit?

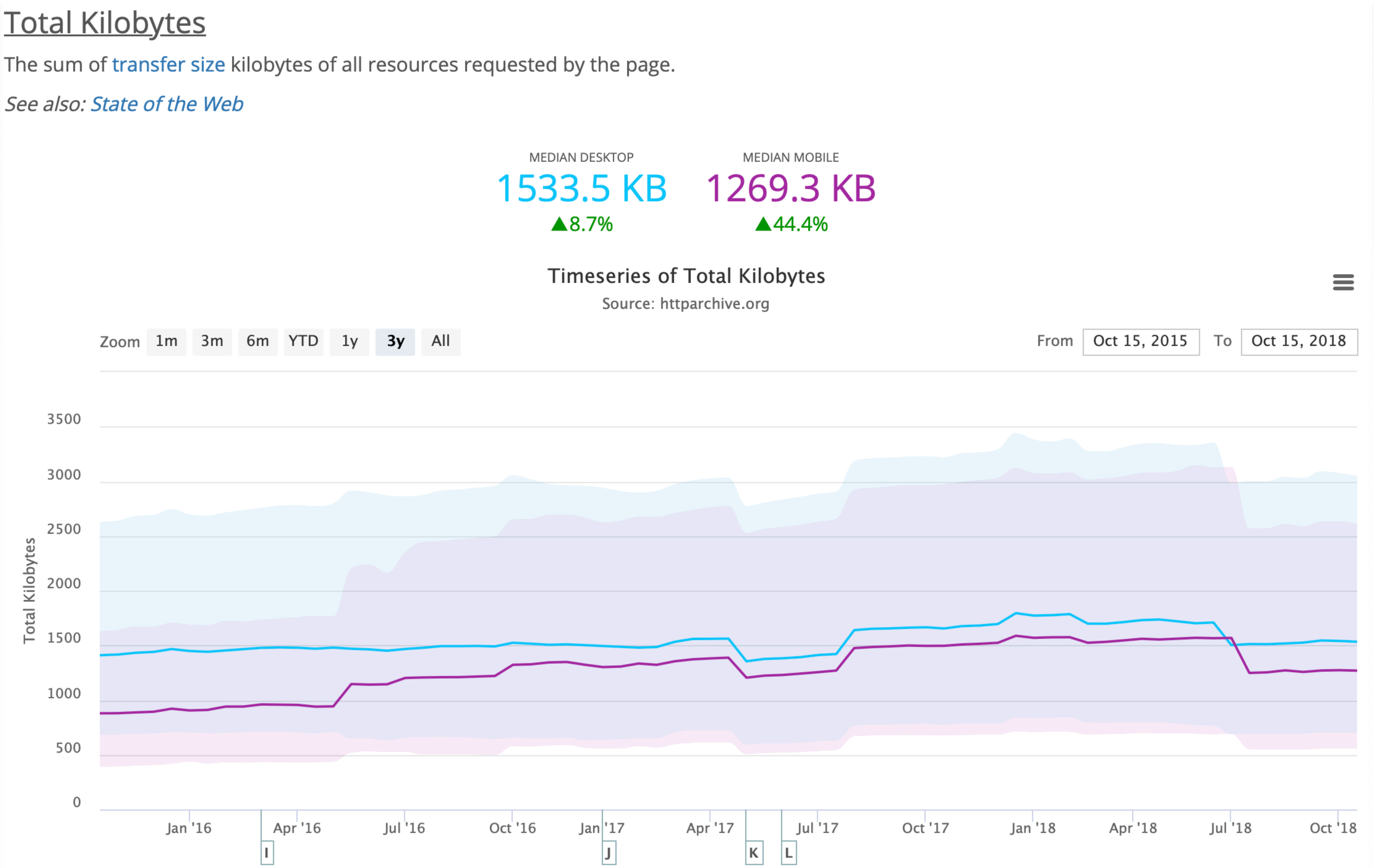
Our field is constantly changing, where last year's cutting edge is this year's commonplace, and today's best practice is tomorrow's old hat. For well over a decade, we've tracked practices, patterns and technologies to keep our audience up to date.

Web Directions Summit brings together the whole team, with two curated tracks, one focused on development and engineering, one focused on design. For this, we've brought together the finest minds at the intersection of technology and design, in an atmosphere unlike any other.

Who's it for?

The Design Team

UX, IxD, visual, Web, Front End and CX experts, Art Directors, Creative Directors,



Total Kilobytes

The sum of transfer size kilobytes of all resources requested by the page.

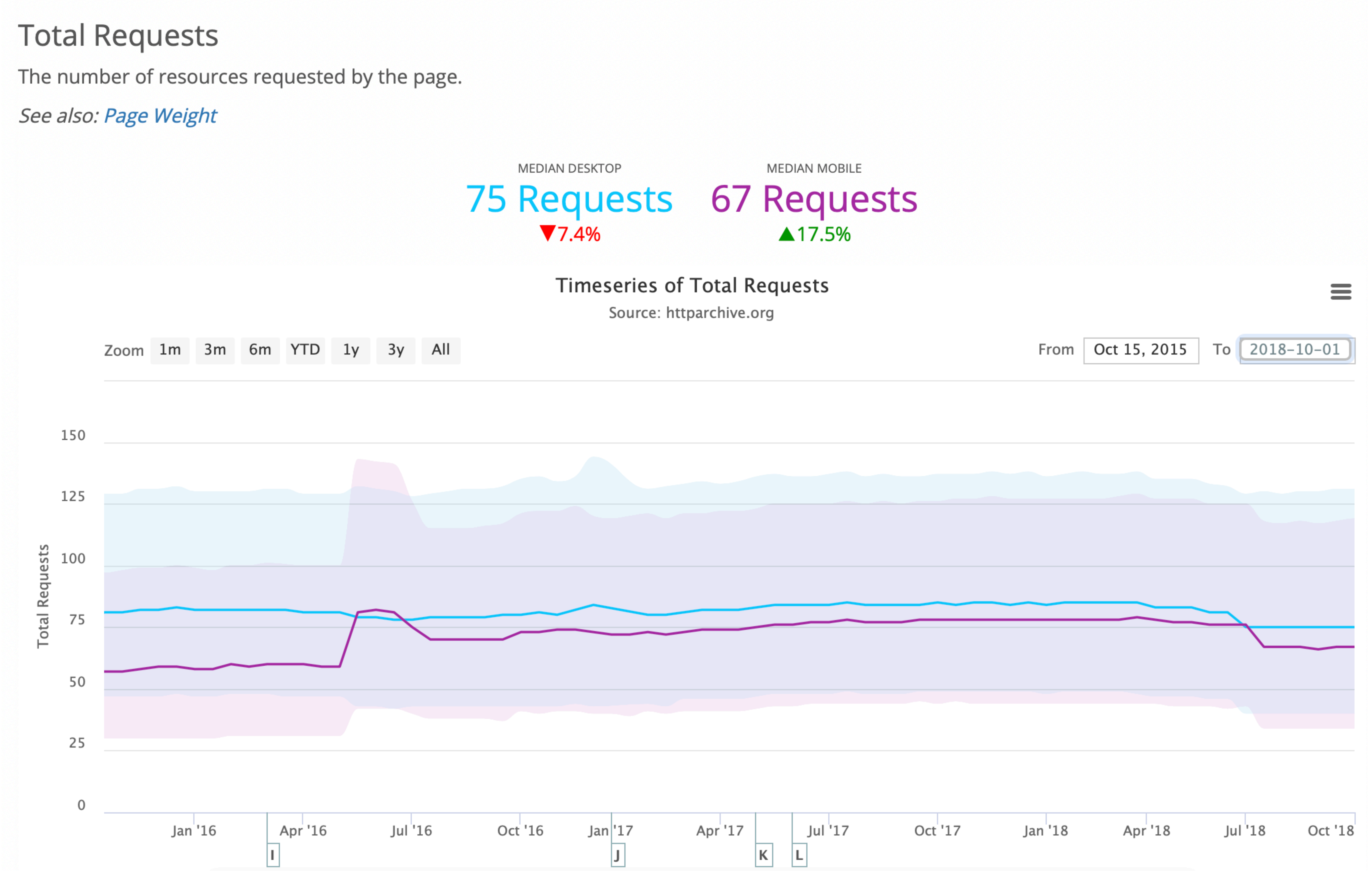
See also: State of the Web

MEDIAN DESKTOP
1533.5 KB
▲8.7%

MEDIAN MOBILE
1269.3 KB
▲44.4%

Timeseries of Total Kilobytes

Source: httparchive.org



Total Requests

The number of resources requested by the page.

See also: Page Weight

MEDIAN DESKTOP

75 Requests

▼7.4%

MEDIAN MOBILE

67 Requests

▲17.5%

Timeseries of Total Requests

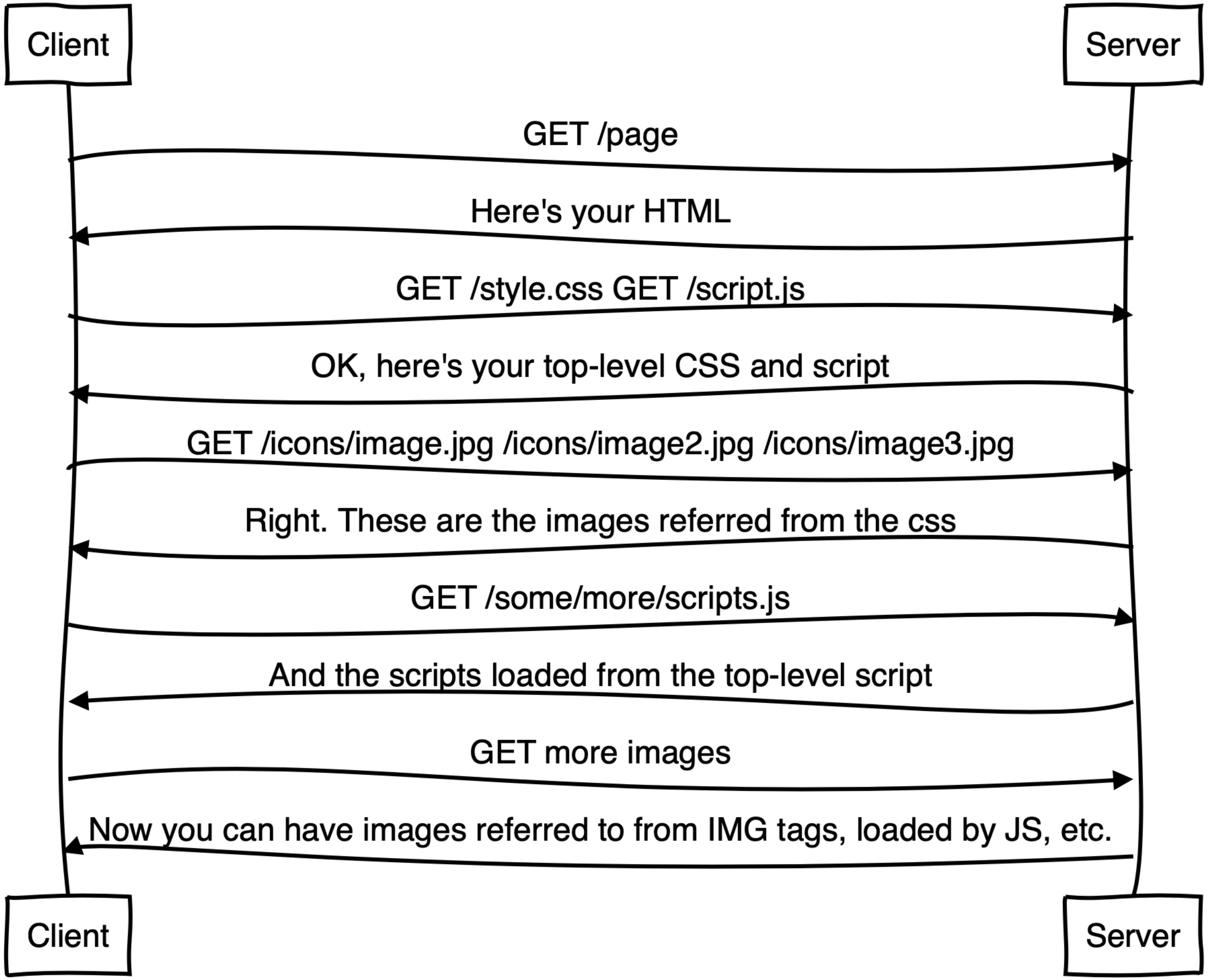
Source: httparchive.org

HTTP/1 requests are expensive, because only one can be active on a connection at a time, and lots of competing TCP connections is bad for performance.

HTTP/2 fixes that with multiplexing.

But, there’s another problem.

We make requests to find out what requests to make.



Client

Server

GET /page

Here's your HTML

GET /style.css GET /script.js

OK, here's your top-level CSS and script

GET /icons/image.jpg /icons/image2.jpg /icons/image3.jpg

Right. These are the images referred from the css

GET /some/more/scripts.js

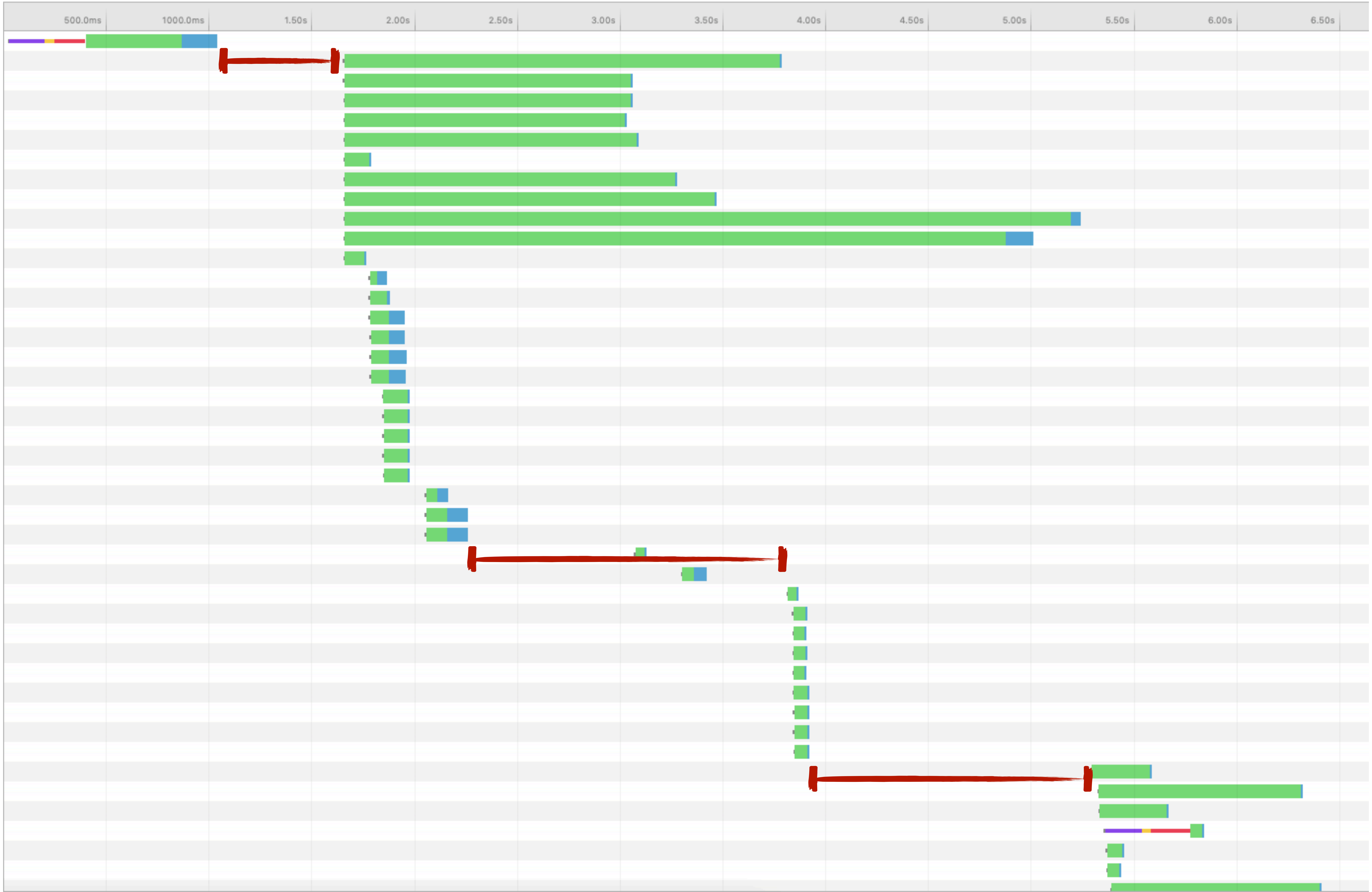
And the scripts loaded from the top-level script

GET more images

Now you can have images referred to from IMG tags, loaded by JS, etc.

Client

Server


500.0ms
1000.0ms
1.50s
2.00s
2.50s
3.00s
3.50s
4.00s
4.50s
5.00s
5.50s
6.00s
6.50s

“Chatty”

CSS Spriting

data: URL Inlining

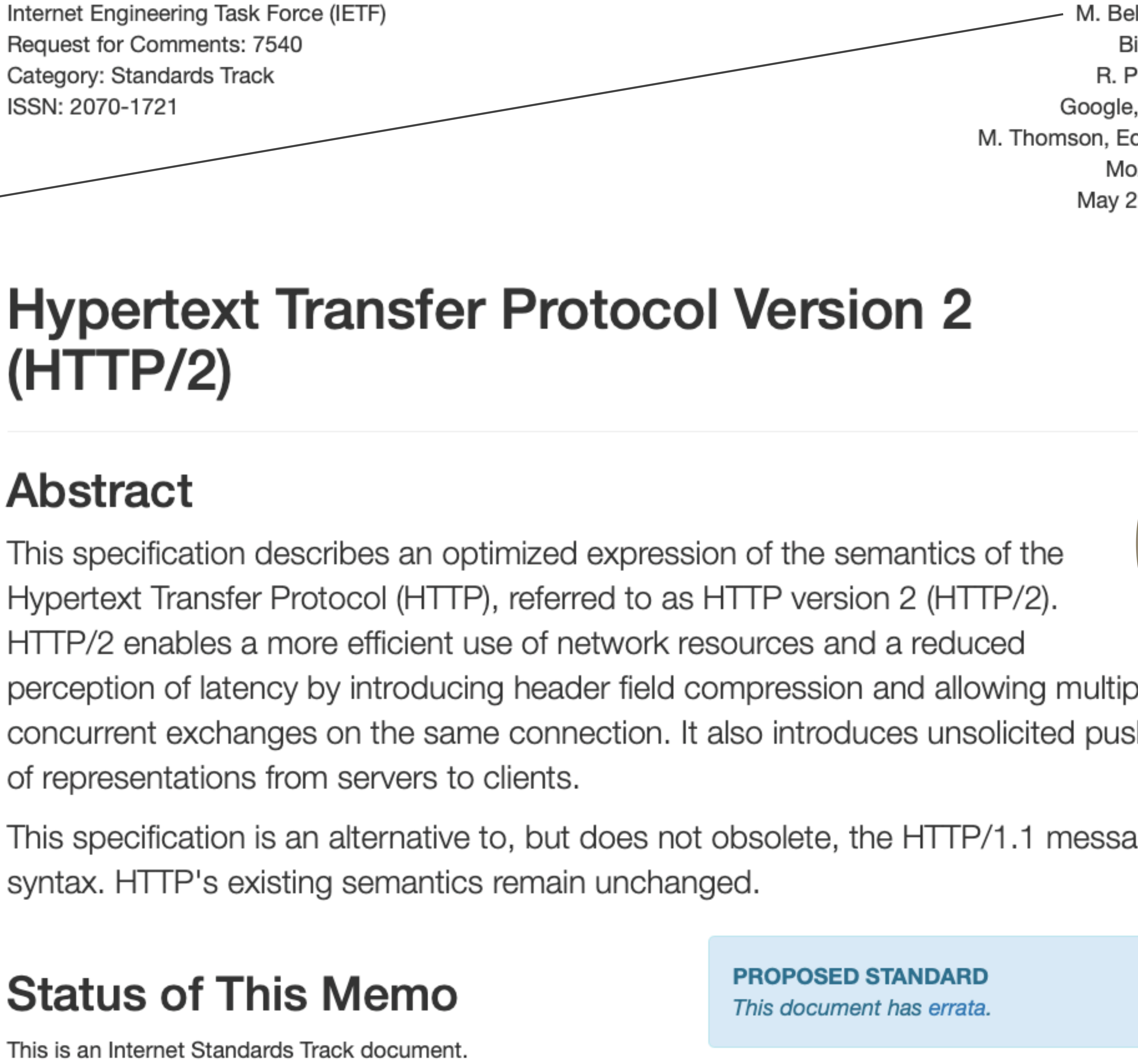
JS and CSS Concatenation

macromedia®
FLASH™
ENABLED

macromedia®
SHOCKWAVE®
ENABLED

We need a way to
avoid the request gap
for “deep” resources

Enter: Server Push



Internet Engineering Task Force (IETF)
Request for Comments: 7540
Category: Standards Track
ISSN: 2070-1721

M. Belshe
BitGo
R. Peon
Google, Inc
M. Thomson, Editor
Mozilla
May 2015

Hypertext Transfer Protocol Version 2 (HTTP/2)

Abstract

This specification describes an optimized expression of the semantics of the Hypertext Transfer Protocol (HTTP), referred to as HTTP version 2 (HTTP/2). HTTP/2 enables a more efficient use of network resources and a reduced perception of latency by introducing header field compression and allowing multiple concurrent exchanges on the same connection. It also introduces unsolicited push of representations from servers to clients.

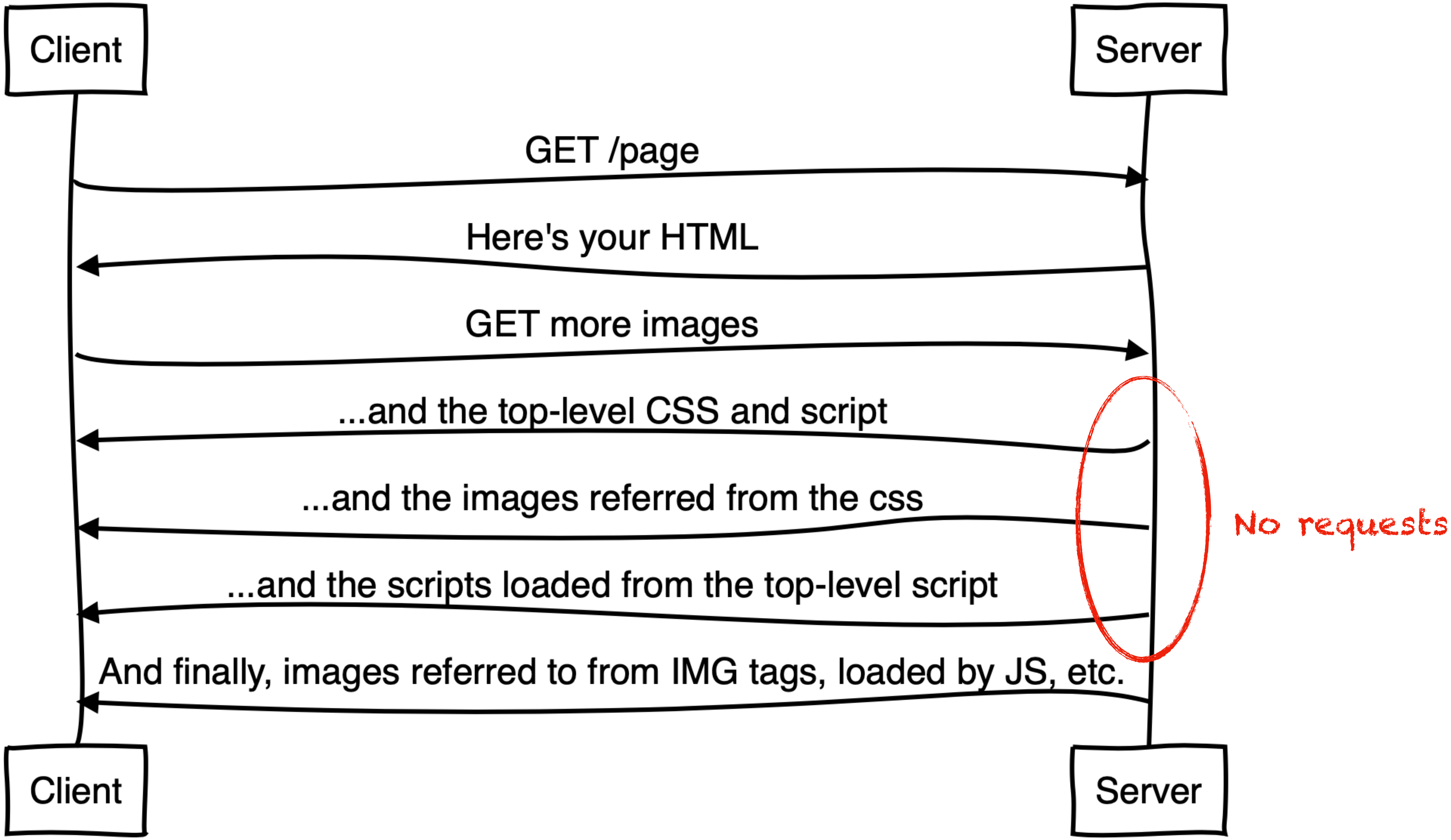
This specification is an alternative to, but does not obsolete, the HTTP/1.1 message syntax. HTTP's existing semantics remain unchanged.

PROPOSED STANDARD
This document has errata.

Status of This Memo

This is an Internet Standards Track document.

“Here’s a request I think you’re about to make, and its response.”



Client

Server

GET /page

Here's your HTML

GET more images

...and the top-level CSS and script

...and the images referred from the css

...and the scripts loaded from the top-level script

And finally, images referred to from IMG tags, loaded by JS, etc.

No requests

Client

Server

PUSH_PROMISE

Synthetic request

Has to be cacheable

Has to be associated with a previous request

Hop-by-hop

tus	Protocol	Initiator	Size	Co…	Waterfall
	http/1.1	Other	241 B	2261	
)	h2	www.mnot.net/	3.0 KB	2273	
)	h2	Push / Other	3.2 KB	2273	
)	h2	Push / Other	38.9 KB	2273	
)	h2	Push / Other	43.8 KB	2273	
)	h2	(index)	61.2 KB	2273	
)	http/1.1	(index)	118 KB	2309	
)	h2	Other	703 B	2333	

ory
Application
Security
Audits
Disable cache
Font
Doc
WS
M
2500 ms
3
Size
Co..
241 B
226
3.0 KB
227
3.2 KB
227
38.9 KB
227
43.8 KB
2273

Queued at 914.62 ms

Started at 920.24 ms

Server Push	TIME
Receiving Push	66.92 ms
Resource Scheduling	TIME
Queueing	5.62 ms
Request/Response	TIME
Reading Push	52.27 ms
Explanation	**129.71 ms**

What if the client
doesn’t want it?

SETTINGS_ENABLE_PUSH

RST_STREAM

Cache Digest*

How does the server know what the client needs now?

Server Push is not Magical.

Maximum usefulness of Push

$$S_{mp} = \min(BW_i \times RTT, IW) - S_{mr}$$

S_{mp} = Maximum size of pushed resources

BW_i = Initial throughput

RTT = Round Trip Time

S_{mr} = Size of main resource

IW = Initial connection window

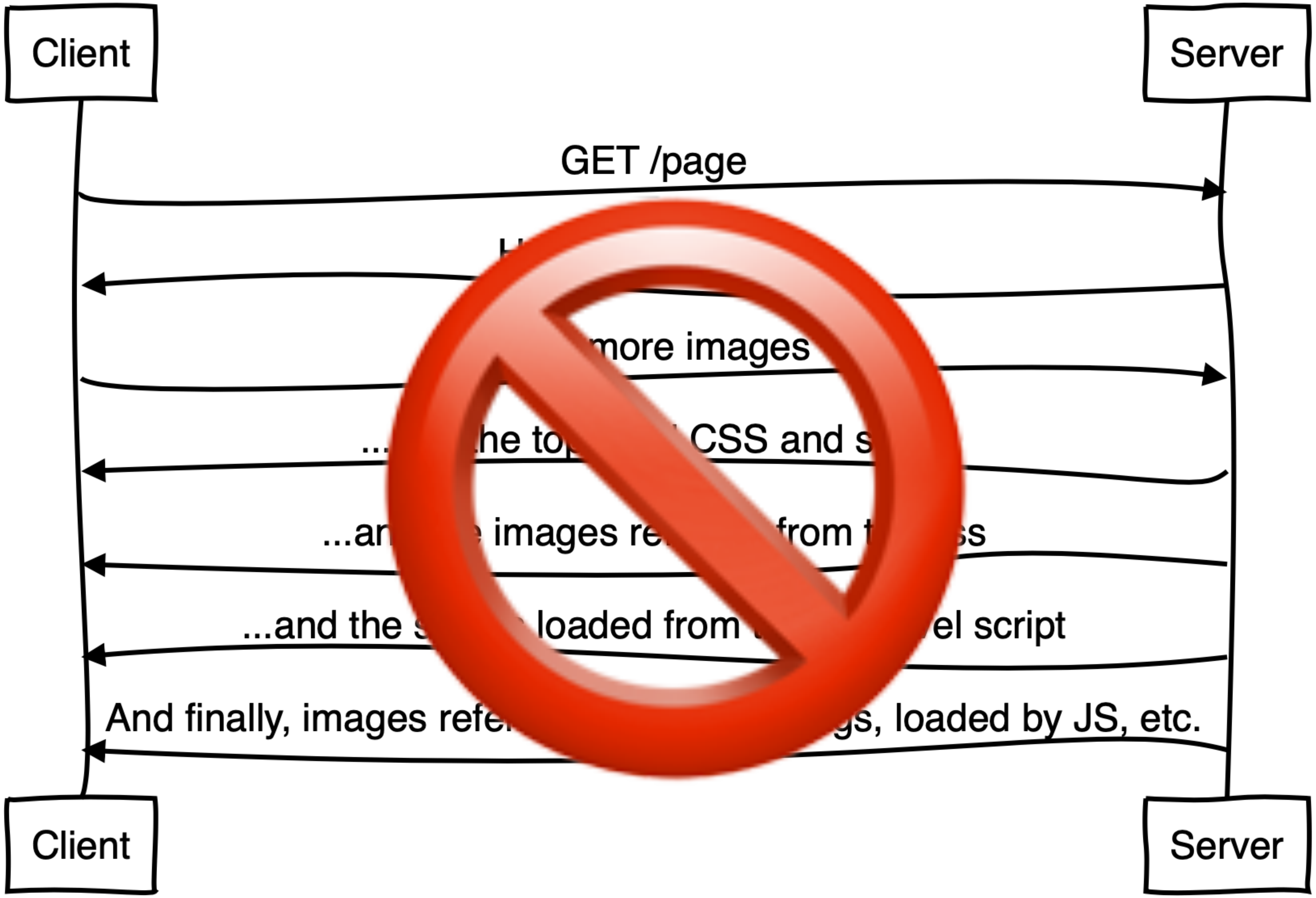
From “Chrome’s View on Push”, IETF102

Some Examples

Country	Mean Min RTT (ms)[1]	Mean Connection Speed (Mb/s)[2]	Max 1RT Data (kb)
South Korea	38	28.6	135.85
US	50	18.7	116.87
India	188	4.9	115.15

- Despite different network conditions, max 1RT data is similar
- But…. Initial CWND caps this
- IW10(rfc6928) equates to ~14600 bytes

From "Chrome's View on Push", IETF102



Client

Server

GET /page

H

more images

... he to CSS and s

...an images re from t ss

...and the s loaded from el script

And finally, images refe gs, loaded by JS, etc.

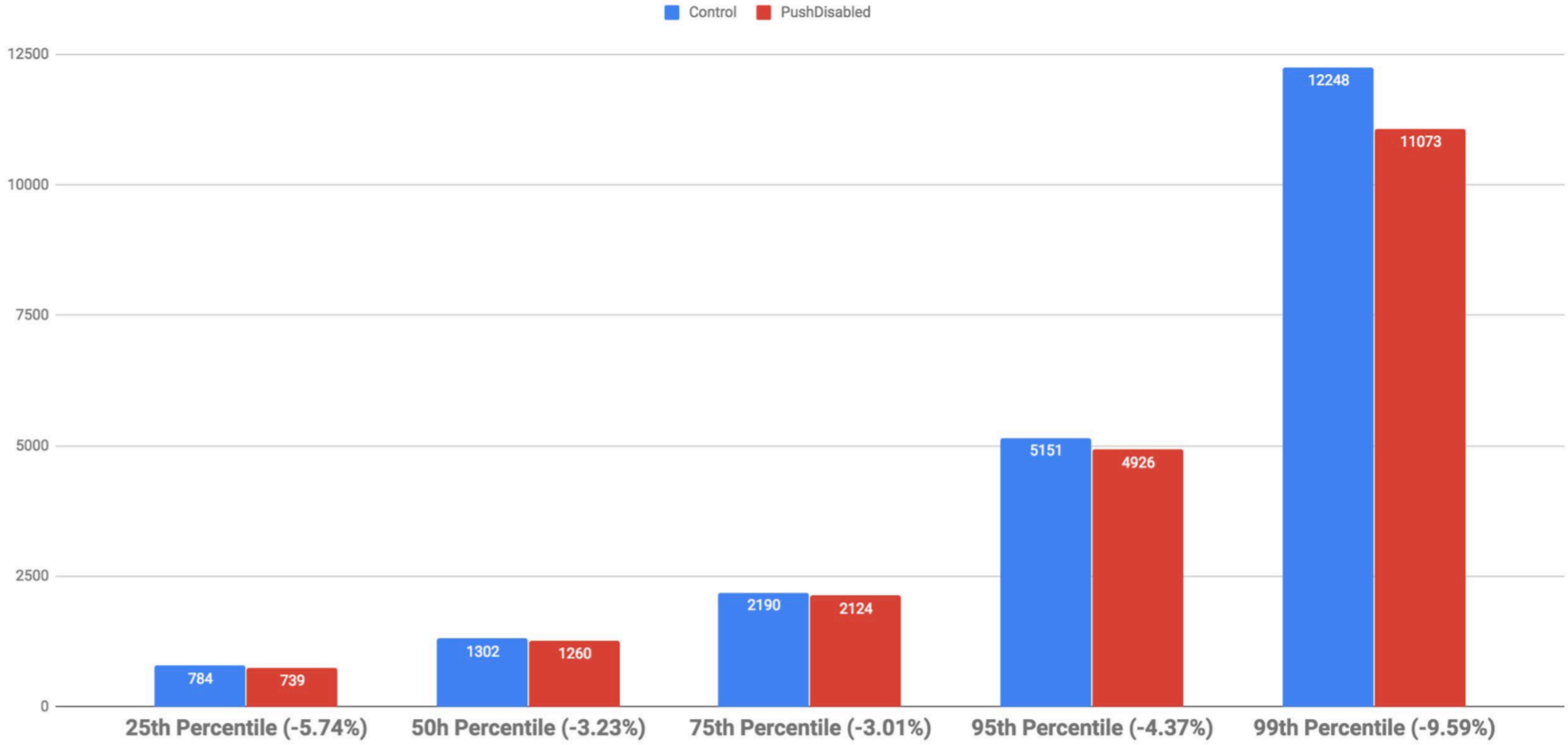
Client

Server

“Push is generally a one round trip optimization.”

– Patrick McManus

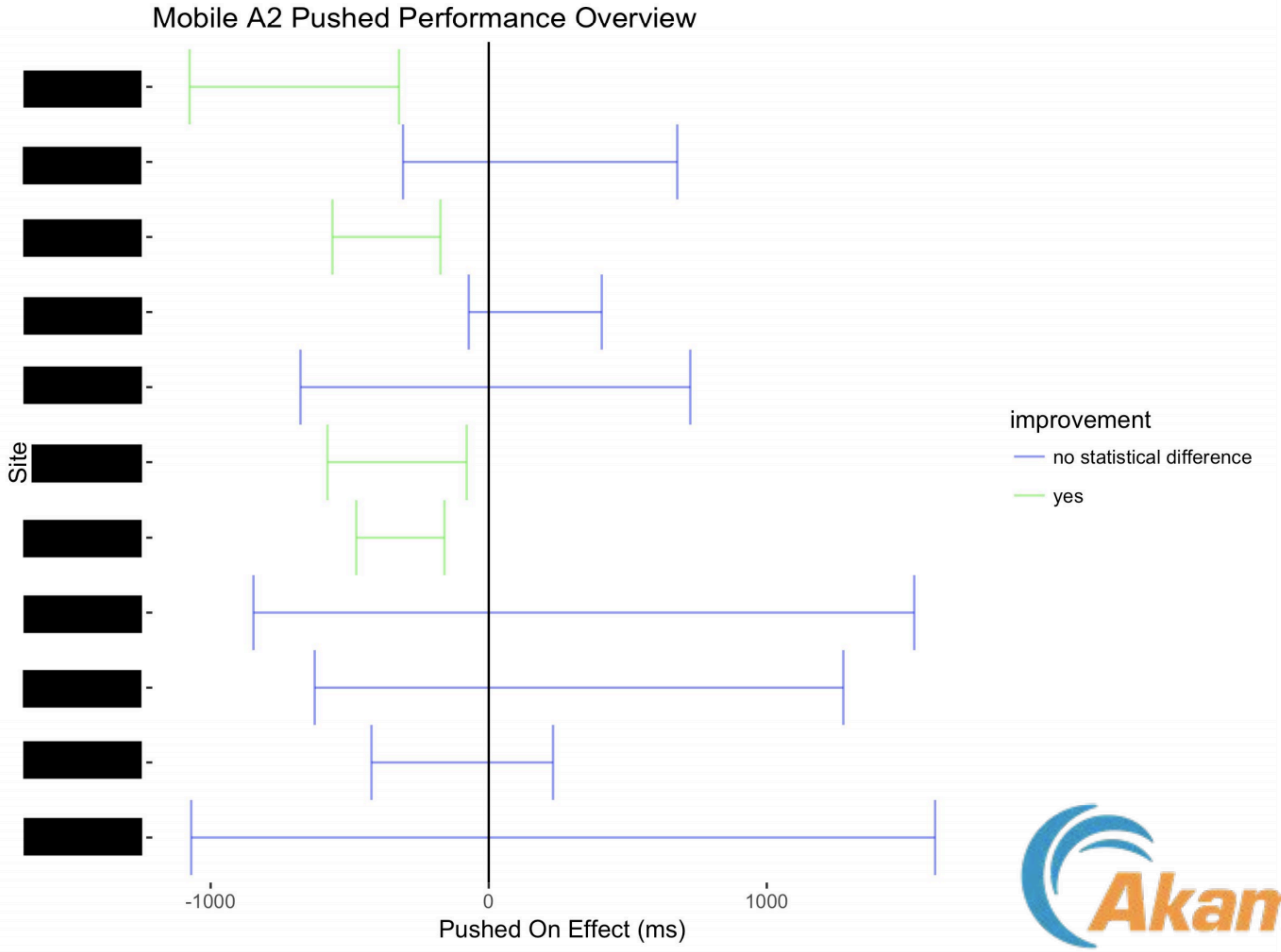
A/B Experiment, Filtered by Domains that Push[1]



[1] From http archive: https://bigquery.cloud.google.com/savedquery/1058239268713:ec65e4a42dbd486fb091718584d73efd

From “Chrome’s View on Push”, IETF102

Confidence Interval : Mean	Site
(-1075.72,-322.71) : -699.21	[redacted]
(-308.19,677.97) : 184.89	[redacted]
(-561.96,-173.59) : -367.78	[redacted]
(-71.86,406.39) : 167.26	[redacted]
(-676.99,724.75) : 23.88	[redacted]
(-579.9,-79.3) : -329.6	[redacted]
(-476.4,-159.31) : -317.85	[redacted]
(-846.19,1530.16) : 341.99	[redacted]
(-625.86,1275.13) : 324.64	[redacted]
(-421.81,231.24) : -95.29	[redacted]
(-1069.9,1605.25) : 267.68	[redacted]



Mobile A2 Pushed Performance Overview

From “H2 Server Push measured over 11 days”, IETF102

Jake Archibald wrote…

HTTP/2 push is tougher than I thought

Posted 30 May 2017

"HTTP/2 push will solve that" is something I've heard a lot when it comes to page load performance problems, but I didn't know much about it, so I decided to dig in.

HTTP/2 push is more complicated and low-level than I initially thought, but what really caught me off-guard is how inconsistent it is between browsers – I'd assumed it was a done deal & totally ready for production.

This isn't an "HTTP/2 push is a douchebag" hatchet job – I think HTTP/2 push is really powerful and will improve over time, but I no longer think it's a silver bullet from a golden gun.

Map of fetching

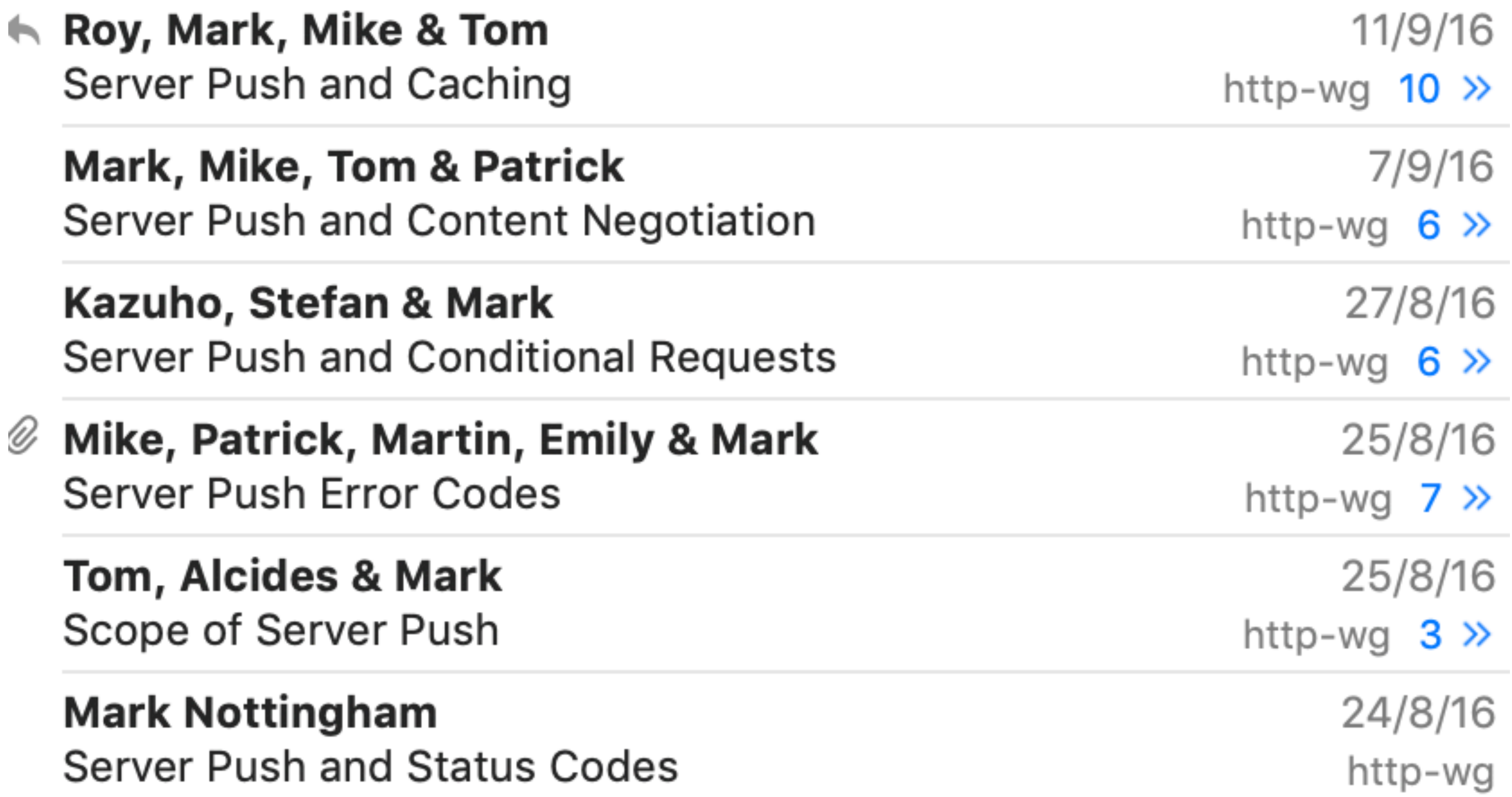
Between your page and the destination server there's a series of caches & things that can intercept the request:

Hello, I'm Jake and that is my face. I'm a developer advocate for Google Chrome.

Elsewhere

Contact

Feel free to throw me an email, unless you're a recruiter, in which case destroy every email-capable device you own to prevent this possibility.



Roy, Mark, Mike & Tom 11/9/16
Server Push and Caching http-wg 10 »

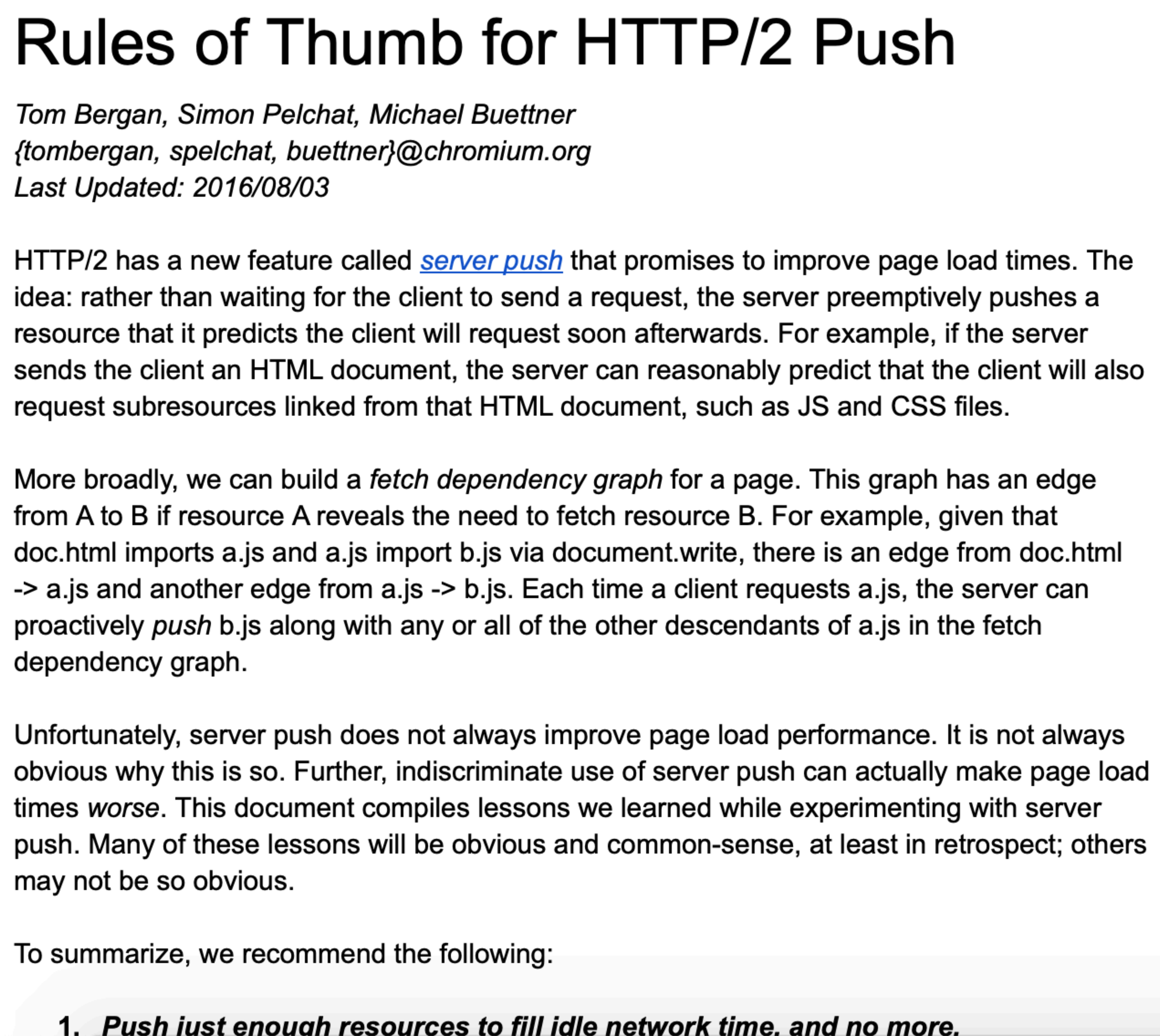
Mark, Mike, Tom & Patrick 7/9/16
Server Push and Content Negotiation http-wg 6 »

Kazuho, Stefan & Mark 27/8/16
Server Push and Conditional Requests http-wg 6 »

Mike, Patrick, Martin, Emily & Mark 25/8/16
Server Push Error Codes http-wg 7 »

Tom, Alcides & Mark 25/8/16
Scope of Server Push http-wg 3 »

Mark Nottingham 24/8/16
Server Push and Status Codes http-wg



Rules of Thumb for HTTP/2 Push

Tom Bergan, Simon Pelchat, Michael Buettner
{tombergan, spelchat, buettner}@chromium.org
Last Updated: 2016/08/03

HTTP/2 has a new feature called *server push* that promises to improve page load times. The idea: rather than waiting for the client to send a request, the server preemptively pushes a resource that it predicts the client will request soon afterwards. For example, if the server sends the client an HTML document, the server can reasonably predict that the client will also request subresources linked from that HTML document, such as JS and CSS files.

More broadly, we can build a *fetch dependency graph* for a page. This graph has an edge from A to B if resource A reveals the need to fetch resource B. For example, given that doc.html imports a.js and a.js import b.js via document.write, there is an edge from doc.html -> a.js and another edge from a.js -> b.js. Each time a client requests a.js, the server can proactively *push* b.js along with any or all of the other descendants of a.js in the fetch dependency graph.

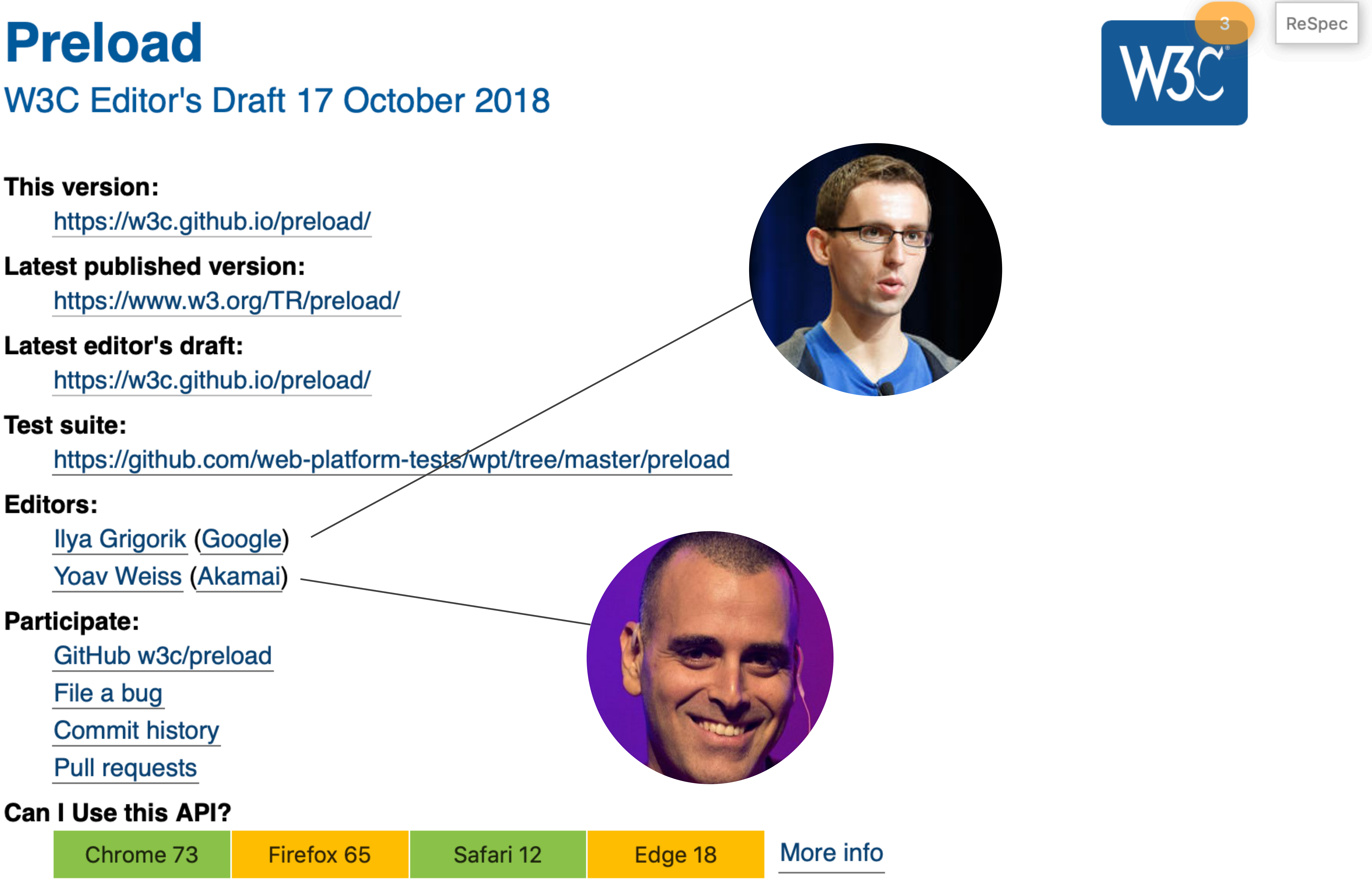
Unfortunately, server push does not always improve page load performance. It is not always obvious why this is so. Further, indiscriminate use of server push can actually make page load times *worse*. This document compiles lessons we learned while experimenting with server push. Many of these lessons will be obvious and common-sense, at least in retrospect; others may not be so obvious.

To summarize, we recommend the following:

1. ***Push just enough resources to fill idle network time, and no more.***

✅ Can push “deep” resources
✅ Can be sent as soon as the HTML request is received
⛔ Server may not know what’s best to push when
⛔ Pushed responses can compete with more important browser requests
⛔ Supported by many browsers, but lots of gotchas

In the meantime...



Preload

W3C Editor's Draft 17 October 2018

This version:
https://w3c.github.io/preload/

Latest published version:
https://www.w3.org/TR/preload/

Latest editor's draft:
https://w3c.github.io/preload/

Test suite:
https://github.com/web-platform-tests/wpt/tree/master/preload

Editors:
Ilya Grigorik (Google)
Yoav Weiss (Akamai)

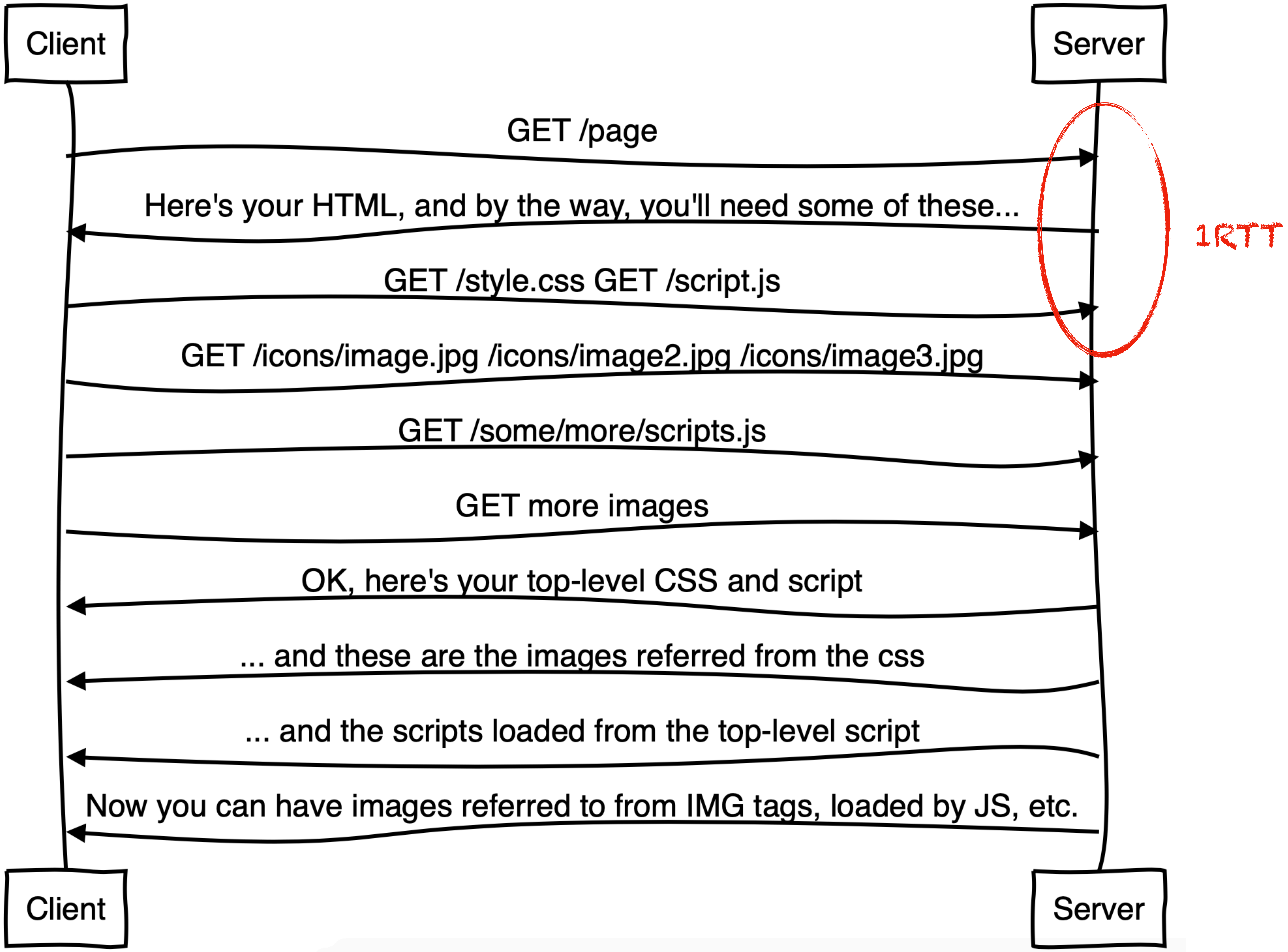
Participate:
GitHub w3c/preload
File a bug
Commit history
Pull requests

Can I Use this API?

Chrome 73	Firefox 65	Safari 12	Edge 18	More info

```
<html>
  <head>
    <link rel="stylesheet" href="/sty
    <link rel="preload" href="/other-
    <script src="/script.js"></script
  </head>
  <body>
    <h1>Hello</h1>
    <img src="hello.jpg"/>
    <img src="other.gif"/>
```

“You probably will need these.”



Client

Server

GET /page

Here's your HTML, and by the way, you'll need some of these...

GET /style.css GET /script.js

GET /icons/image.jpg /icons/image2.jpg /icons/image3.jpg

GET /some/more/scripts.js

GET more images

OK, here's your top-level CSS and script

... and these are the images referred from the css

... and the scripts loaded from the top-level script

Now you can have images referred to from IMG tags, loaded by JS, etc.

Client

Server

1RTT

“Shopify’s switch to preloading fonts saw a 50% (1.2 second) improvement in time-to-text-paint. This removed their flash-of-invisible text completely.”

– Shopify

✅ Can request “deep” resources

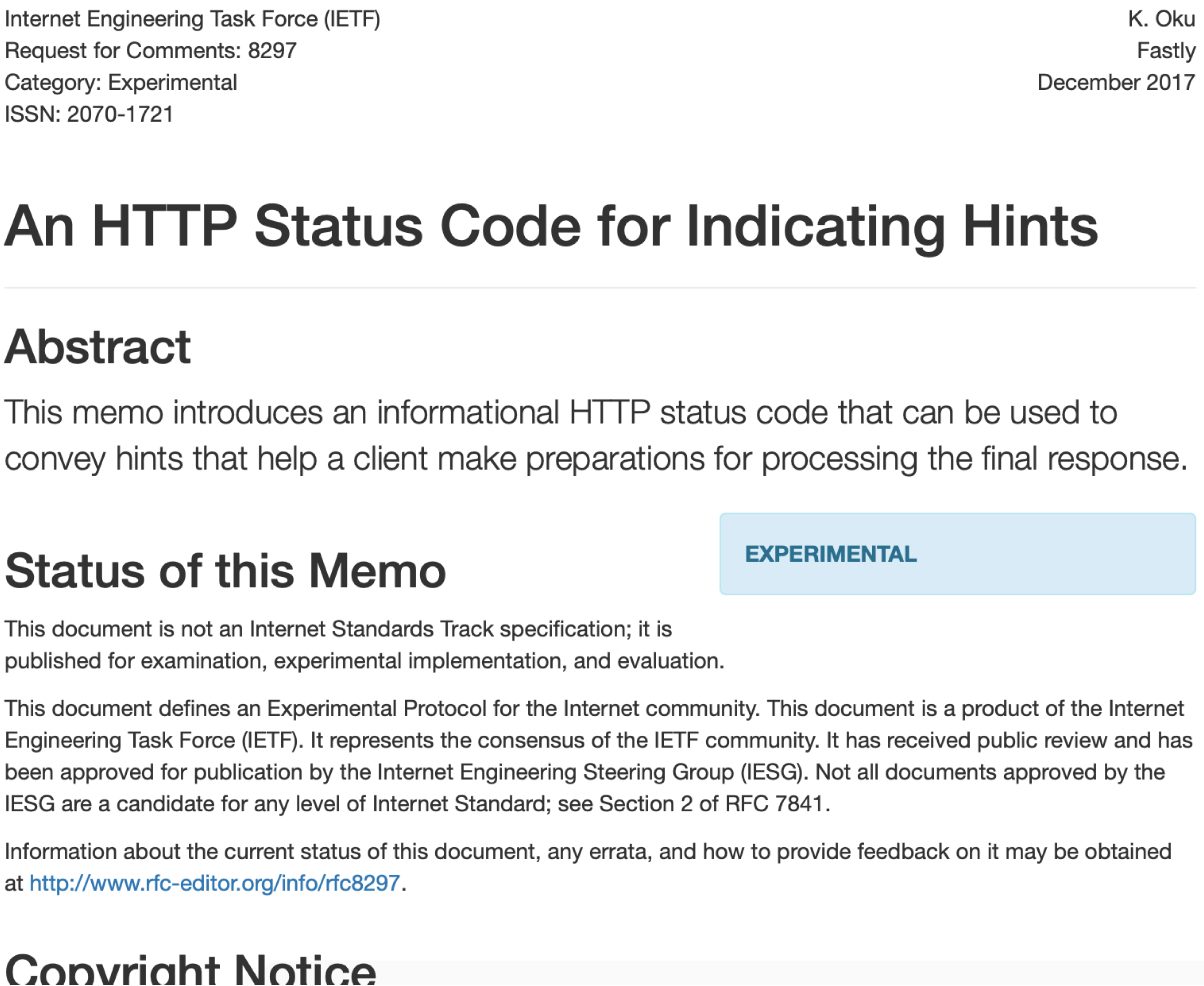
✅ Browser decides priority, whether to fetch

⛔ ... but only after HTML response starts

What about server think time?

✅ Server can Push during think

⛔ Preload relies on HTTP or HTML headers



Internet Engineering Task Force (IETF)
Request for Comments: 8297
Category: Experimental
ISSN: 2070-1721

K. Oku
Fastly
December 2017

An HTTP Status Code for Indicating Hints

Abstract

This memo introduces an informational HTTP status code that can be used to convey hints that help a client make preparations for processing the final response.

Status of this Memo

EXPERIMENTAL

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

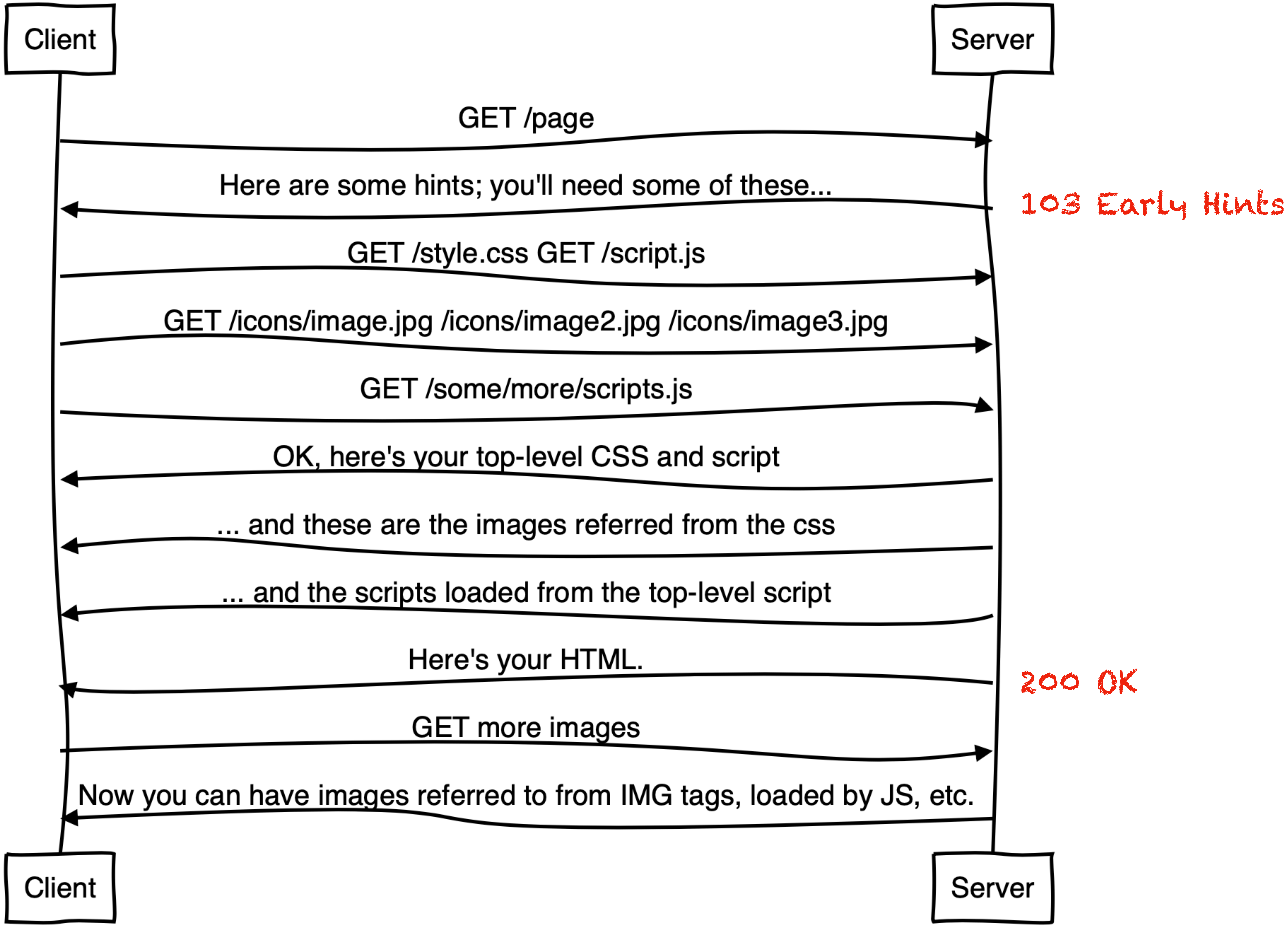
This document defines an Experimental Protocol for the Internet community. This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc8297.

Copyright Notice

RFC 8297

non-final response



Client

Server

GET /page

Here are some hints; you'll need some of these...

103 Early Hints

GET /style.css GET /script.js

GET /icons/image.jpg /icons/image2.jpg /icons/image3.jpg

GET /some/more/scripts.js

OK, here's your top-level CSS and script

... and these are the images referred from the css

... and the scripts loaded from the top-level script

Here's your HTML.

200 OK

GET more images

Now you can have images referred to from IMG tags, loaded by JS, etc.

Client

Server

“It’s **difficult**.”

– The Browsers

✅ Can request “deep” resources

✅ Browser decides priority, whether to fetch

✅ ... as soon as request is received

⛔ ... but still requires 1RT for hint + request

⛔ Not yet supported in browsers

“Push is generally a one round trip optimization.”

– Patrick McManus

If we destroyed push, would anyone really notice?

Currently only 0.04% of sessions

Seems to be a footgun

Better things to work on:

- Connection Pooling
- Prioritization
- DoH
- QUIC
- Alt svc
- ????

From "Chrome's View on Push", IETF102

Use preload for “deep” resources

In many cases, Server Push isn't necessary

If you use push, use it:

- ⇢ to fill 1RT after HTML, no more
- ⇢ to fill server think time (but keep an eye on 103)

All of this is still evolving

Collect metrics!