TEU00311
What is the Internet doing to me? (witidtm)

Stephen Farrell
stephen.farrell@cs.tcd.ie

https://github.com/sftcd/witidtm
https://down.dsg.cs.tcd.ie/witidtm

URLs accessed 20230904 (not all content updated)

The Internet is not the web

- An important point!
- The web is (roughly) the set of computers that speak the HTTP protocol
  - HTTP == HyperText Transfer Protocol (http://example.com)
  - HTTPS == HTTP/Transport Layer Security (https://example.com)
- Email doesn’t use HTTP, but rather (mostly) the Simple Mail Transfer Protocol (SMTP) which is a couple of decades older than HTTP
- Mobile network internals (3G, 4G, 5G...) mostly run over IP using a bunch of protocols you’d prefer to never have to know about
- But lots of our interactions with the Internet are via the web
  - Many phone “apps” are just a simple user interface using an on-device web library

The web has fantastic things! Too many to list really but here's some...

- The Internet archive: Books, music, video, s/w
  - https://archive.org/ Includes the wayback machine
- Wikipedia ('nuff said)
  - https://en.wikipedia.org/wiki/Trojan_Room_coffee_pot
- Project Gutenberg, out of copyright books, in various formats
  - https://www.gutenberg.org/
  - In 2021 : "Italy blocks Gutenberg book publishing website" https://ooni.org/post/2021-italy-blocks-gutenberg-book-publishing-website/

    Still seems partially blocked according to OONI
- I like OpenStreetMap
  - https://www.openstreetmap.org/
- DuckDuckGo is a reasonable search engine
  - https://duckduckgo.com/
- As is Brave's search
  - https://search.brave.com/
- Arxiv.org and domain-specific friends are good
  - https://arxiv.org/
- Google scholar is academically useful (there are others too)
  - https://scholar.google.com/

What web sites do you people like?

Mechanics of the Web
(Plumbing:-)

So you click a link...

- What do you think happens?

So you click a link...

10 minute video that's a bit simplified and out of date, but that might be a good thing:-

https://gizmodo.com/what-actually-happens-when-you-click-on-a-link-1665573786

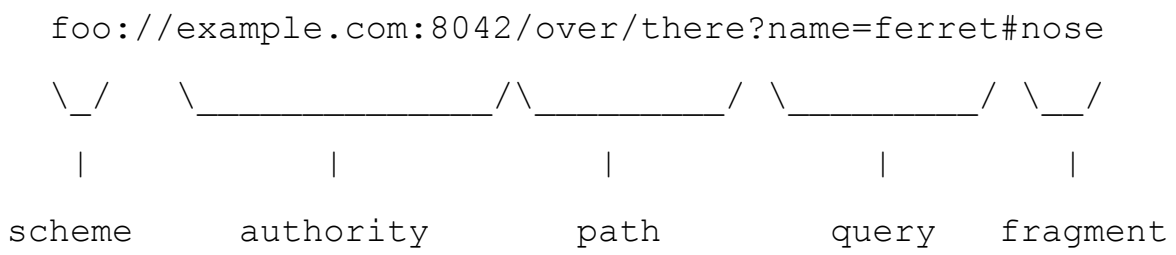
- Link (URL) is part of a rendered HTML page (or typed/bookmarked) and is clicked…
- Domain Name System (DNS) resolution...DNS is a recursive protocol that uses a world-wide piece of infrastructure (the DNS) that depends on a complex name registration system (next slides)
- Transport Connection Established (TCP)… that's the network connection between your browser and a web server instance
- Transport Layer Security (TLS, which used be called SSL) session established… that (usually) authenticates the server and sets up encryption of traffic between browser and web server instance - Not mentioned in video but now happens >80% of the time as HTTPS URLs are used
- HTTP request sent… may include site-specific parameters, e.g. amount of money to transfer to where or which bit of a map you want to see
- HTTP response received… probably contains HTML with links, some of which are automatically fetched (e.g. images), so GOTO step 1 for each of those

Actual browser behaviour is **much** more complex (pre-fetching, caching, QUIC instead of TLS/TCP, Javascript, tab isolation...)

What’s a URL?

- A web “link” is really a **Uniform Resource Locator (URL)**
- Knowing what a URL looks like or *should look like* can save you from being phished!
- URI/URL definitions are in RFC3986
  - https://datatracker.ietf.org/doc/html/rfc3986
  - URI = Uniform Resource Identifier is a kind of generalisation of URL
  - RFC3986 has been sorta “forked” by browser makers (a fine example of a boring controversy;-)
    - https://url.spec.whatwg.org/
- URIs are used as web links but also for many other things, e.g. voice over IP signalling, e.g. SIP URIs can represent phone numbers
  - sip:1-999-123-4567@voip-provider.example.net
  - https://en.wikipedia.org/wiki/SIP_URI_scheme

Parts of a URI/URL



“Real” example:

https://down.dsg.cs.tcd.ie/witidtm/examples/stuff.html#middle

- Mostly, the URL schemes you’ll see will be “https” or “http” but there are many more
- The “authority” part is essentially the DNS name of the host (with an optional port number)
- The “path” can be thought of as a directory/folder name in a web server’s document root
- The “query” part provides a way to parameterise URLs sent to programs
- The “fragment” provides a way to “land” your browser at some place in a web page

Referring to URLs in academic work

- In academic work, e.g. a publication, or assignment, its an excellent idea to add the date on which you accessed the URL as part of the reference
- Because the content that a browser gets at that URL can change anytime
- Example text you might see in a paper/report:

```
Bibtex provides a way to note the date on which a resource was
accessed. [1]


[1] “How to add ‘date accessed’ or ‘date retrieved’ in BibLaTeX?”,
https://tex.stackexchange.com/questions/111630/how-to-add-date-accessed-or-date-retrieved-in-biblatex
, accessed 2023-09-04
```

Domain Name System (DNS) 1/3

- Internet Assigned Numbers Authority (IANA, https://iana.org/) keeps lists of “top level domain names” (TLDS, e.g. .com, .ie), IP address allocations and protocol numbering registrations
  - That’s a fantastic bit of bookkeeping but no more than that, policies are set elsewhere despite what some “Internet Governance” folks might say
  - IANA is homed in ICANN which is one of the policy setting/operations entities
  - The Regional Internet Regisgtries (RIRs) are another, and the IETF controls the protocols that create the space in which the policies operate
- The DNS is structured as a tree with a single root, below which we have .com, .ie, etc. and below those we have example.com, tcd.ie etc. and within tcd.ie we have whatever e.g. TCD might want e.g. down.dsg.cs.tcd.ie
- There are a set of (13 logical, physically: ~1000) DNS root servers on the Internet that serve as the DNS “root zone” and who can tell you set of IP addresses where you can find more authoritative information about .com or .ie. or any of the Top Level Domains (TLDs)
  - Entities that do the bookkeeping for a TLD are called registries

Domain Name System (DNS) 2/3

- There are 1464 TLDs, all recent ones being outrageously expensive, about 200 of those are country-code TLDs (ccTLDs), e.g. for “.ie” or “.de”
  - https://www.iana.org/domains/root/db
- DNS supports internationalised domain names (IDNs) because the world has many languages and writing systems (the Internet started out in ASCII)
  - ουτοπία.δπθ.gr is a (no longer functioning) Greek IDN https://en.wikipedia.org/wiki/Internationalized_domain_name
    - another representation for that is (in Punycode) is xn—kxae4bafwg.xn--pxaix.gr)
  - You may see IDNs listed in punycode as XN--<<stuff>>
  - IDNs create lots of confusion possibilities!
- DNS authoritative servers store the canonical information for some “zone” (e.g. all the hosts in/below tcd.ie) but can also delegate to another server as happens in the case of cs.tcd.ie
- When you want a new name (e.g. jell.ie) you have to go to a registrar who works with the relevant registry (e.g. Tolerant Networks Limited is me-as-a-registrar for .ie) and then pay to rent that for a few years (IEDR, rebranded as https://weare.ie/ is the name of the Irish ccTLD registry)
  - Not all names are available – could be taken or a trade mark – some lawyers love this stuff!

Domain Name System (DNS) 3/3

- You need a server machine with an IP address for a DNS name to be useful
  - You often get from a hosting company or cloudy service provider like AWS or Azure or whomever
- And then that machine's name and IP address need to be published in the DNS of the parent zone
- Then you can e.g. install nginx (a web server) and make your web site and interact with e.g. LetsEncrypt.org to get a public key certificate for your DNS name so TLS to your server will work
  - Then browsers can nicely visit your web site
- Web crawlers and attackers of all sorts will also constantly ping your server, all the time
- At that point you may decide to be an advertiser or not, if you do, you'll probably start to record things about people who visit you and maybe you'll sign up to some advertising platform to make money for you and them
- But you might also decide not to track anyone (what I do)
  - Modulo normal web logs!

IP Addresses

- There are two versions of the Internet Protocol (IP) in use – versions 4 and 6
- They are incompatible at the IP layer but mostly the same from the transport or application layer view
- An **IPv4** address: **134.226.36.81** (that's the address of down.dsg.cs.tcd.ie)
- An **IPv6** address: **2001:770:10:20b:a47a:3ff:fed2:9d22** (forget what that is, but it's in TCD and still responding:-)
- There are "private" address ranges as well as public IPv4 and IPv6 addresses. Often private IPv4 addresses (e.g. 10.0.0.1 or 192.168.1.1) are used within home networks and Network Address Translation (NAT) is used to map those to shared public IP addresses
- NAT is mainly to get around the fact that **all 4 billion IPv4 addresses have been allocated** by IANA already, modulo few corner cases and a grey-market in re-sold addresses
- There are plenty of IPv6 addresses to go around – my home network gets 2^{72} of those from Eir!
- IP addresses sometimes are and sometimes are not personally identifying information (PII) – has been established in some court cases

IP Addressing and Routing

- How does a browser make that TCP connection to a web site after it gets the IP address?
- Regional Internet Registries (RIRs) such as RIPE allocate Autonomous System (AS) numbers and blocks of IP addresses to network operators (e.g. ISPs, enterprises, hosting companies)
- Addresses are further allocated downstream, eventually to e.g. your laptop via DHCP, or (semi-)manually to a server hosting a web site in a data centre
- ASes tell one another about who has what blocks of addresses and how to route packets to one another via the Border Gateway Protocol (BGP) – remember there are tens of thousands of ASes – that's done by **announcing prefixes to peers**, e.g. "134.226/16" is TCD as is "2001:770:10::/48"
- Once your browser has a source IP address and knows the web site (destination) IP address then they can establish the TCP connection, assuming the relevant ISPs have done a good job with BGP, and with managing their routers, which mostly happens
  - The source address is needed to get the answer from the web site

TCP and TLS session

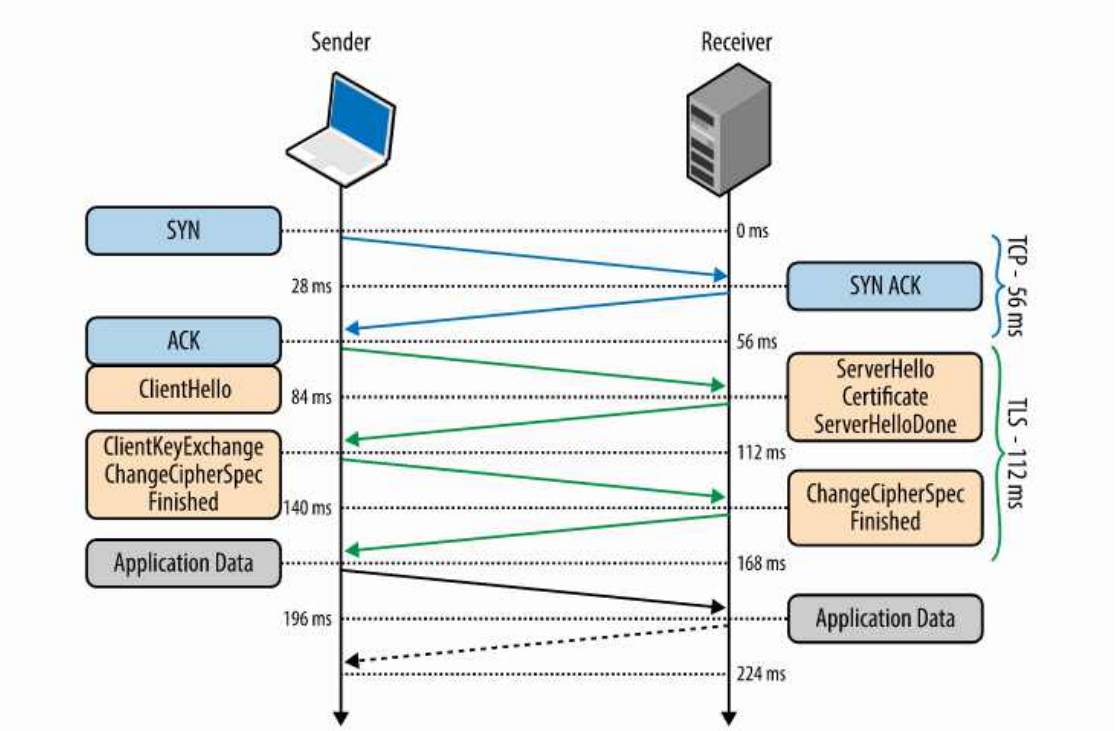


Image is from https://hpbn.co/transport-layer-security-tls/ which also has **lots** of text describing what's going on.

I think I'm ok with using the diagram according to
https://creativecommons.org/faq/#can-i-reuse-an-excerpt-of-a-larger-work-that-is-licensed-with-the-noderivs-restriction

HyperText Transfer Protocol (HTTP)

- HTTP/0.9 was an early version that saw a lot of deployment
- HTTP/1.1 is a text-based protocol and still widely used
- HTTP/2 (aka h2) is semantically the same but binary and with a few other efficiency improvements – go-faster-stripes basically, but with TLS as an almost-MUST
- HTTP/3 (aka h3) standardised in 2022, is HTTP over QUIC (QUIC is a recent alternative that does what TLS over TCP does)
- HTTP/TLS => HTTPS URL scheme

HTTP

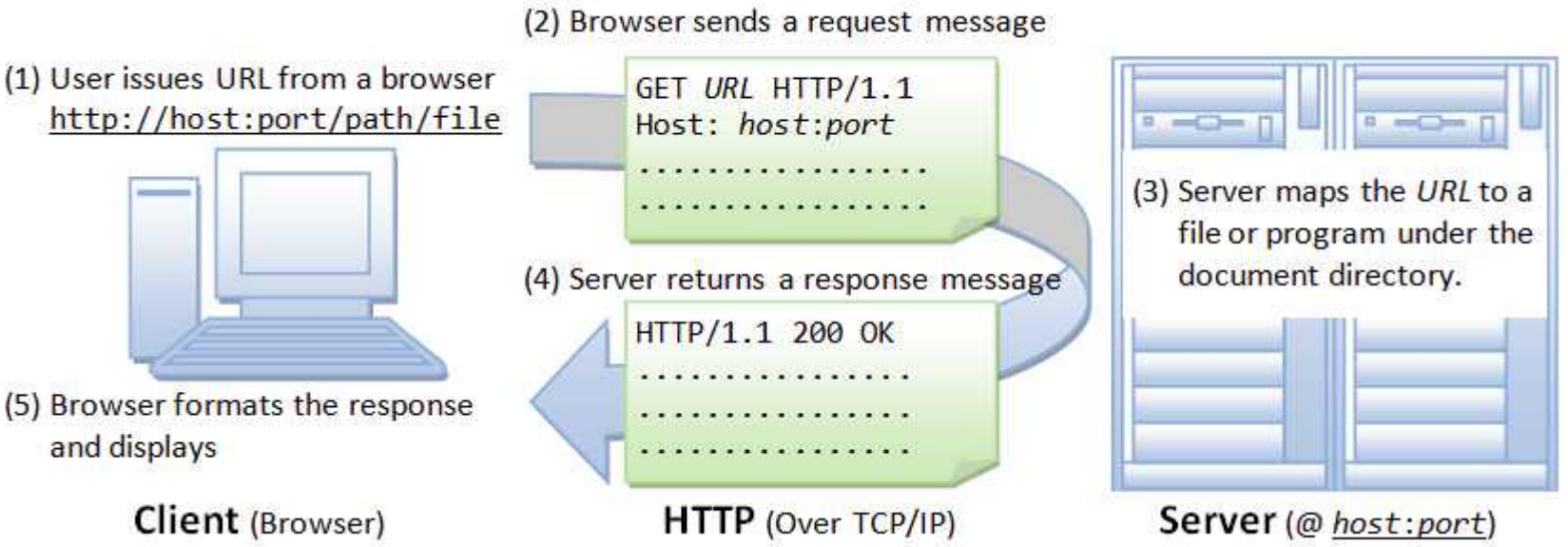


Diagram is from a long, detailed description of HTTP at:
https://personal.ntu.edu.sg/ehchua/programming/webprogramming/HTTP_Basics.html

That used to be at: https://www.ntu.edu.sg/home/ehchua/programming/webprogramming/HTTP_Basics.html
...but the university wanted to disclaim responsibility for “personal” content, or something...there’s some legal text at https://personal.ntu.edu.sg/ **...**sigh – this is an example of **link rot!**

“link rot”

- That’s where some URL is included in some document (HTML or other) but later, when someone clicks that URL, the get an error
- HTTP 404 – file not found most commonly
- Is the existence of “link rot” a good or bad thing?

HyperText Markup Language (HTML)

- Simplistically, the web pages that you download via HTTPS are HTML files
  - There’s also lots of non-HTML content: video, Javascript, images
  - Reality is nowhere near this simple… but it can be!
- HTML describes the structure of the web page
- Browser renders that
- Key concept: hypertext links

HTML for a trivial web page

https://down.dsg.cs.tcd.ie/witidtm/examples/trivial.html

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
    <title>A trivial web page</title>
</head>
<!-- Background white, links blue (unvisited), navy (visited), red (active) -->
<body bgcolor="#FFFFFF" text="#000000" link="#0000FF" vlink="#000080" alink="#FF0000">
    <p>The trivial content is just a link to the
        <a href="https://jell.ie/news/">jell.ie news</a>
    </p>
</body>
</html>
```

A link on our trivial web page

https://down.dsg.cs.tcd.ie/witidtm/examples/trivial.html

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
    "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
    <title>A trivial web page</title>
</head>
<!-- Background white, links blue (unvisited), navy (visited), red (active) -->
<body bgcolor="#FFFFFF" text="#000000" link="#0000FF" vlink="#000080" alink="#FF0000">
    <p>The trivial content is just a link to the
        <a href="https://jell.ie/news/">jell.ie news</a>
    </p>
</body>
</html>
```

TLS Sessions per “front page”

During 2019Q1, I did some tests, loading the “front page” of these sites and recording the network traffic. The numbers below are the number of separate TLS sessions that were created when the initial page HTML is rendered, to display content and (mostly) ADs

site	N	min	max	stdev	avg
ietf.org	22	4	12	2.12	9.29
irishtimes.com	22	74	**158**	22.47	126.36
jell.ie	22	4	12	1.96	6.73
nytimes.com	22	29	**98**	16.07	81.23
rte.ie	19	38	63	6.33	50.32
tcd.ie	22	69	**102**	10.61	92.18
www.ietf.org	14	4	7	0.73	5.07

- N = count of tests done
- Min,max,stdev,avg refer to the number of TLS sessions for each test
- Automation tool used was Selenium on Ubuntu which mostly used FF, but also chrome/opera for some tests
- Browsers/selenium drivers are “out of the box” with no special config, nor plug-ins, extensions etc.

Why did the Irish Times front page have an average of 120+ TLS sessions?

Why did tcd.ie have an average > 92?

Why did the Irish Times front page have an average of 120+ TLS sessions?

Why did tcd.ie have an average > 92?

With a desktop browser... you can see what’s happening: “shift-ctrl-I”
(but we’re getting a bit ahead with that)

Scaling and Content Delivery Networks

Big web sites aren’t trivial

- Caveat: I don’t run a big web site! So I don’t really know this stuff in full detail.
- There’s a web called “high scalability” that publishes “war stories” for people who manage large systems
  - https://highscalability.com/ (re-directs back to plaintext!! - bad practice)
- One article gives a nice perspective from a sys admin point-of-view – probably too much detail for us, but good as a description of how a “middle-sized” web site has developed in the back-end over the last decade
  - https://www.betabrand.com/ - that used to say: “Best. Pants. Ever.” (sigh;-)
  - https://boxunix.com/2018/12/10/from-bare-metal-to-kubernetes/
  - If you do read that, I’d say the take-away is that there’s enough complexity in even a middle-sized merchant web site that many things can go wrong on their end, including some that could affect you (e.g. data leaks, which we’ll look at later)
- Another article talks about how Netflix do stuff...

“Netflix: What Happens When You Press Play?”

- We’ll look at some quotes from this Dec. 2017 article about Netflix (author: Todd Hoff)
  - https://highscalability.com/blog/2017/12/11/netflix-what-happens-when-you-press-play.html
  - 2022 stats from https://www.businessofapps.com/data/netflix-statistics/
  - Slightly newer (2018) info at https://blog.apnic.net/2018/06/20/netflix-content-distribution-through-open-connect/ but mostly same

Some 2017/2018 Netflix statistics (with some updates):

- Netflix had more than 110 million subscribers in 2017 (2021: 209 million, 2022: 220 million)
- Netflix operated in more than 200 countries (2022: 190+ https://help.netflix.com/en/node/14164)
- Netflix had $11.6 billion revenue in 2017 (2021: $29.6 billion, 2022: $31.6)
- Netflix played more than 1 billion hours of video each week. As a comparison, YouTube streams 1 billion hours of video every day while Facebook streams 110 million hours of video every day.
- Netflix played 250 million hours of video on a single day in 2017.
- Netflix accounted for over 37% of peak internet traffic in the United States.
- Netflix planned to spend $7 billion on new content in 2018.

The quotes... (1)

- “Netflix collects a lot of information. Netflix knows what everyone has watched when they watched it and where they were when they watched. Netflix knows which videos members have looked at but decided not to watch. Netflix knows how many times each video has been watched…and a lot more.”
- “When browsing around looking for something to watch on Netflix, have you noticed there’s always an image displayed for each video? That’s called the header image. The header image is meant to intrigue you, to draw you into selecting a video. The idea is the more compelling the header image, the more likely you are to watch a video. And the more videos you watch, the less likely you are to unsubscribe from Netflix.”

The quotes... (2)

- “Everyone used to see the same header image. Here’s how it worked. Members were shown at a random one picture from a group of options, like the pictures in the above Stranger Things collage. Netflix counted every time the video was watched, recording which picture was displayed when the video was selected. For our Stranger Things example, let’s say when the group picture in the center was shown, Stranger Things was watched 1,000 times. For all the other pictures, it was watched only once each. Since the group picture was the best at getting members to watch, Netflix would make it the header image for Stranger Things forever. ”
- “That’s why Netflix now personalizes all the images they show you. Netflix tries to select the artwork highlighting the most relevant aspect of a video to you. How do they do that? Remember, Netflix records and counts everything you do on their site. They know which kind of movies you like best, which actors you like the most, and so on. Let’s say one of your recommendations is the movie Good Will Hunting. Netflix must choose a header image to show you. The goal is to show an image that lets you know about a movie you’ll probably be interested in. Which image should Netflix show you?”

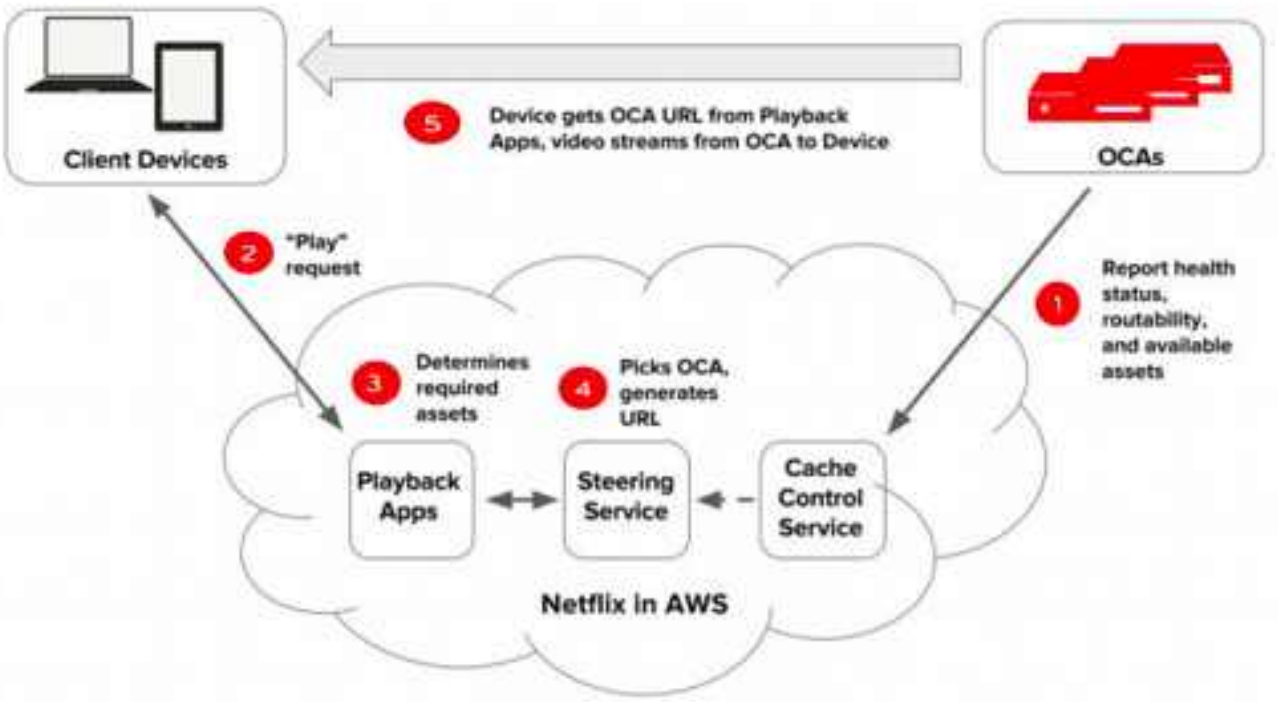
Content Deliery Networks

- Netflix use “hundreds of thousands” of Amazon EC2 instances for their application logic – stuff that happens before video is streaming
  - Anyone know what “Amazon EC2” is? If not, you wanna?
- For video, Netflix built their own Content Delivery Network (CDN) – most web sites use 3rd party CDNs like Akamai, Cloudflare etc.
  - CDNs represent a kind of Internet centralisation that you may not have known about?
- Basic idea is to put large data files near where the customer is so data (video) gets there faster, and to be more reliable when failures happen
- So where are Neflix CDN points-of-presence (PoPs)?

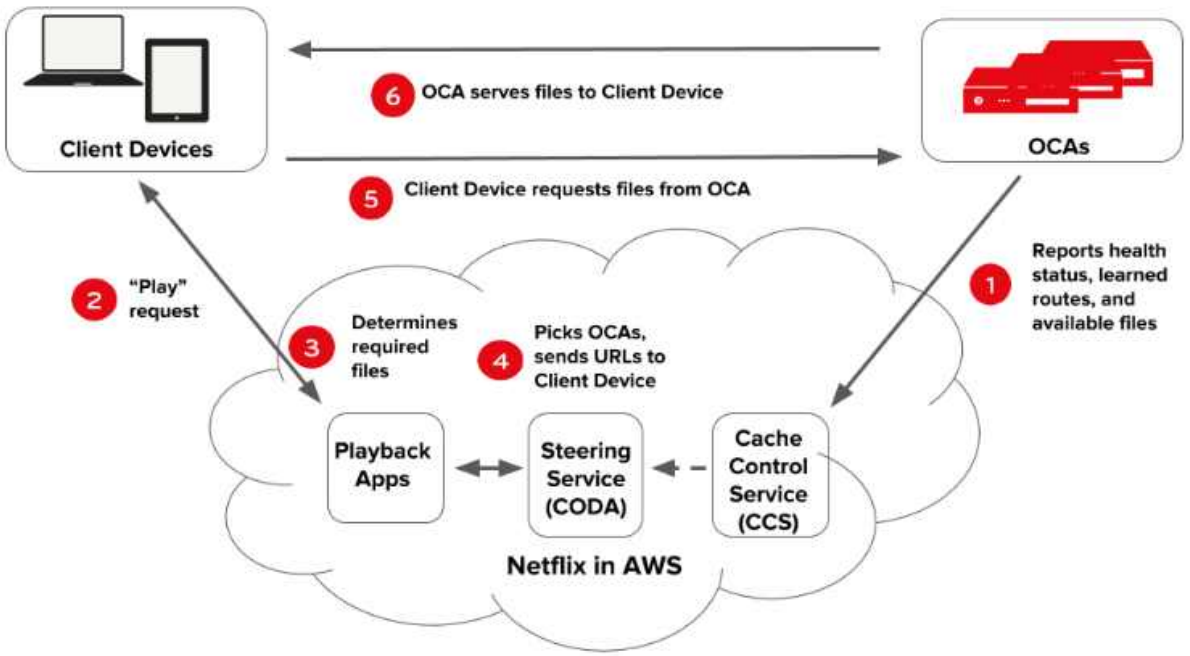
Netflix PoPs (2017)

- PoPs are mostly hosted by ISPs or IXPs
- Because Netflix send lots of data (37% of the 2107 internet traffic in the US) and otherwise the ISPs would have to pay interconnect charges to the ISPs between them and the content
- That’s done with Neflix-designed hardware physically located in the ISP’s network – called an Open Connect Appliacnce (OCA)

So what happens when you press Play?



More up to date pic



I have no idea if the differences in these diagrams are significant:-)

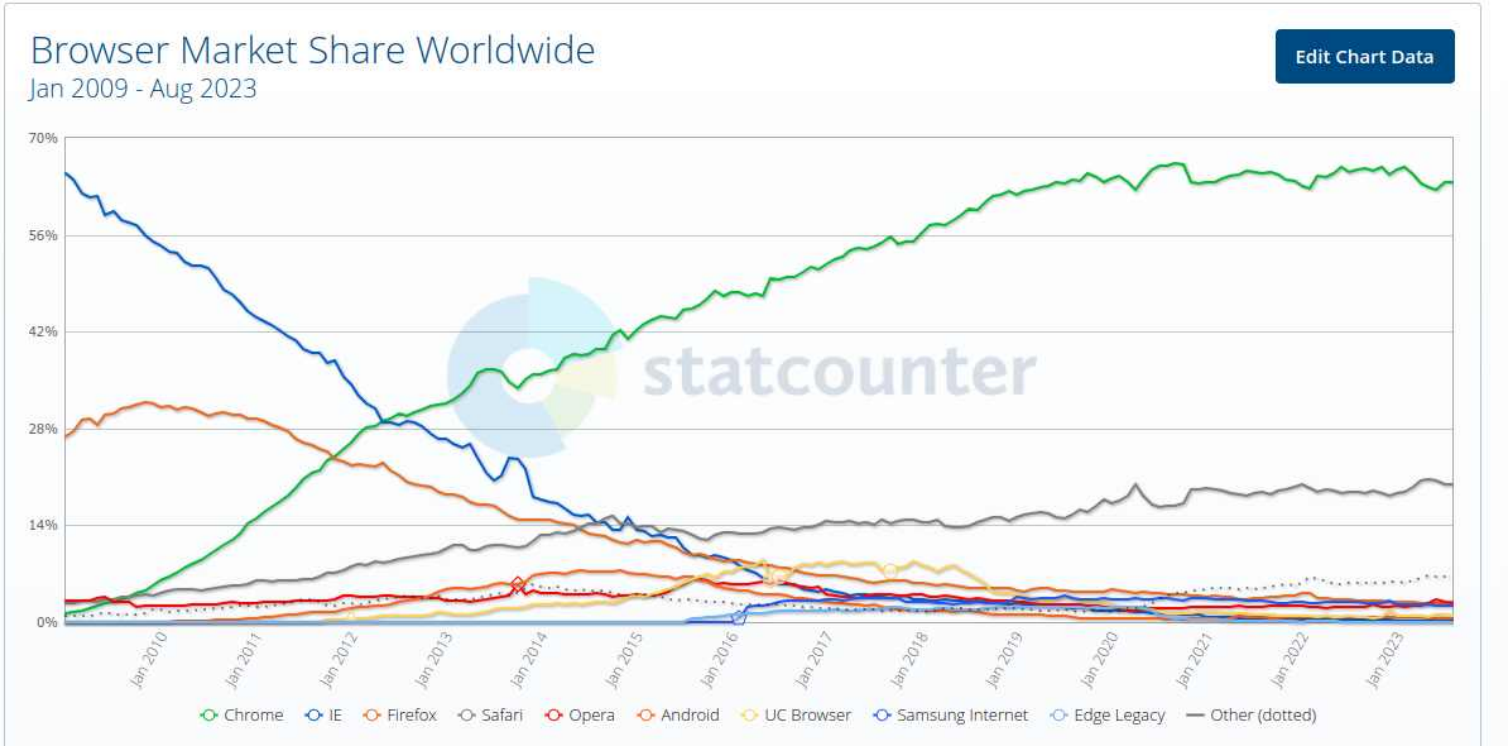
https://openconnect.netflix.com/Open-Connect-Overview.pdf

Browsers & Browser Hygiene

Web Browsers...

- Generic software used as a web client is called a browser, e.g. safari, chrome, firefox
- Some platforms (android, iOS) include a “browser” that can be embedded into other apps – often those apps are just a thin shim layer on top of the embedded browser
- So it’s interesting to consider how that can affect you

Browser worldwide market share



2009-2023

Chrome
Safari
Edge
Firefox: 2.94% !!
Opera
Android
UC browser
...
Brave
Vivaldi

https://gs.statcounter.com/browser-market-share#monthly-200901-202308

Overall browser landscape

- Browser defaults are chosen by browser implementers (Google, Mozilla, Microsoft, Apple, Handset vendors...)
  - Generally they allow Javascript and cookies, do telemetry, try get you to login, keep lots of state ...
- Historically, browser-makers seemed to care most about market share
  - Performance and rendering were their main concerns as they lose market share if they’re slower or sites don’t render (well)
- They started getting significantly better at security a while back (2013+)
- Some browser-makers are starting to get a bit better at privacy, perhaps esp. Brave
- IMO they don’t behave as if they think you should be the one in control

Why browser hygiene matters...

- Developer of popular (300k installs) chrome ad blocking extension hadn’t time to keep maintaining that...
- Someone offered to buy the code and promised to maintain it...
- That someone added malware to the code that stole cookies and session tokens, and maybe more...
  - That “someone” seems to have been a repeat offender
- Result: 300k very unhappy people changing passwords all over and one very very embarassed original maintainer whose name is now mud (for some).
- Happened in 2020: https://github.com/jspenguin2017/Snippets/issues/2

My browser setup (1)

- Default browser: FF “nightly” + NoScript/Ghostery & disallowing cookies, with some white-listed sites, and search via brave search
  - This is the only browser that saves logins, but not for sensitive things (we’ll consider passwords later)
  - Some sites don’t work with the above; mostly: screw ‘em
- Opera for managing home network, Brave for TCD stuff
- Tor Browser: If searching for anything sensitive (e.g. medical info)
- If-need-be: vivaldi or chromium/incognito setup to delete-all on exit (at least I hope so;-)
  - Use that e.g. for airline/hotel bookings, but also Vivaldi as default for mail links
- If-all-else-fails: chrome – for e.g. RTE player very rarely

My browser setup (2)

- ~~On phone: Sailfish OS (not Apple and not Android) - sailfish browser with no JS/no cookies and 2ndary open-kimono browsers if-need-be (Webcat/Web pirate)~~
  - ~~Or a 2nd phone phone (android, yuk!) with Brave or FF~~
- On phone: e/OS de-googled android with Brave set for no cookies
  - Backup: FF nightly, zero'ing on exit

Your browser setup?

- What do you do?

- Recommend you figure out some browser-hygiene you consider ok and follow that
- Requires some self-discipline!
- Be willing to help others do the same!