
# Internet Routing! But it’s not just BGP…

AusNOG 2023

Tom Paseka

# Internet Routing

# RIPv2?

# OSPF / ISIS ?

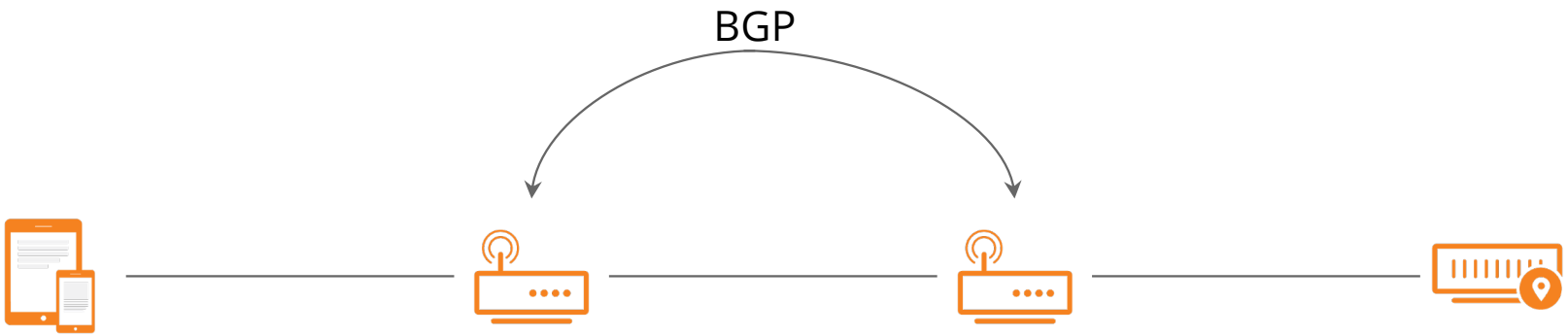
# OK then…

# …BGP?

# Internet Routing

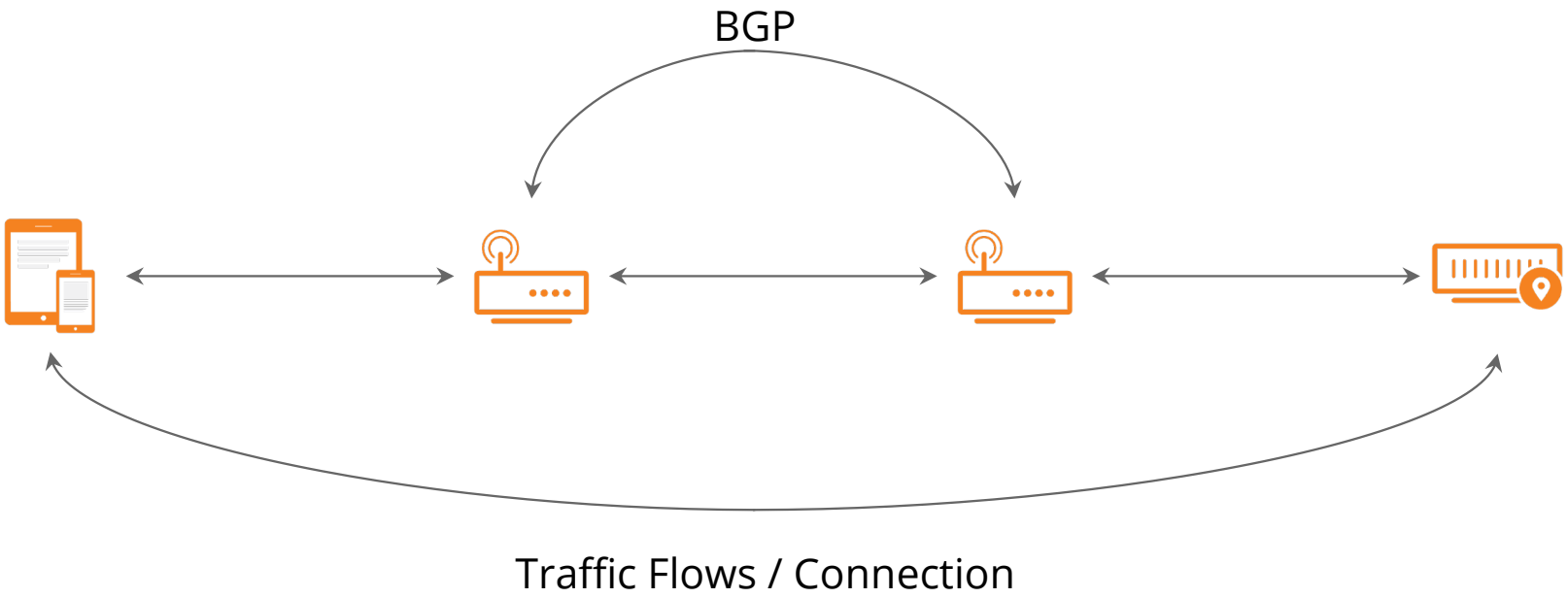
CLOUDFLARE

# Internet Routing




CLOUDFLARE

# Internet Routing




CLOUDFLARE

# Internet Routing

**example.com**
**203.0.113.5**

CLOUDFLARE

# Is DNS a routing platform?

# Most traffic over your network is “Routed” by DNS

Approximately 50% of traffic in your network:

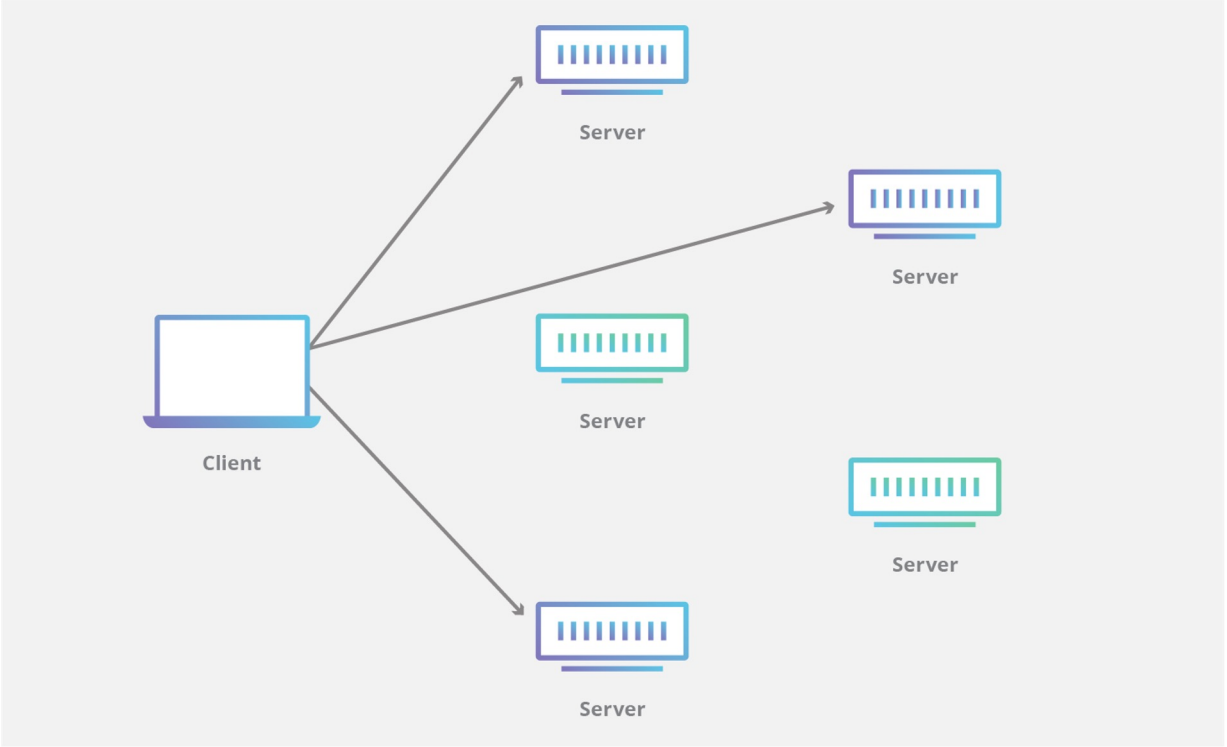
- Not just directed to an IP
- But “routed” to a specific location inside or adjacent to your network.

Others also use application logic for routing, not always visible in DNS.

CLOUDFLARE

# Exceptions

## Anycast




CLOUDFLARE

https://www.cloudflare.com/learning/cdn/glossary/anycast-network/

# How does it work?

# How does it work?

Different classes:

- Network (eg: Akamai)
- Application/Network (eg: Netflix)
- Application (eg: Disney+)

Some Tools used:

- IP Geolocation Feeds
- EDNS-Client-Subnet
- BGP

# Networks

The networks are the simplest to reverse-engineer:
They give an easy path to their customers to use.

Looking first at Akamai

# Networks

```
% dig www.sbs.com.au
###
;; ANSWER SECTION:
www.sbs.com.au.              300      IN       CNAME
        www.sbs.com.au.edgekey.net.
www.sbs.com.au.edgekey.net. 300       IN       CNAME
        e7065.b.akamaiedge.net.
e7065.b.akamaiedge.net.   20       IN       A        96.16.68.225
```

CLOUDFLARE

# Networks

## How did they know to give that IP?

```
% traceroute -I 96.16.68.225
traceroute to 96.16.68.225 (96.16.68.225), 64 hops max, 72 byte packets
 1  104.28.0.0 (104.28.0.0)  9.093 ms  7.520 ms  7.640 ms
 2  172.69.21.1 (172.69.21.1)  8.309 ms  8.164 ms  7.973 ms
 3  172.68.188.21 (172.68.188.21)  9.708 ms  55.153 ms  40.752 ms
 4  172.68.188.87 (172.68.188.87)  9.818 ms  12.150 ms  7.270 ms
 5  ae34.r03.border101.sjc01.fab.netarch.akamai.com (23.203.158.23)  12.367 ms  14.769 ms  7.944 ms
 6  192.168.225.43 (192.168.225.43)  23.001 ms  14.022 ms  11.265 ms
 7  192.168.236.149 (192.168.236.149)  13.908 ms  12.880 ms  10.462 ms
 8  192.168.246.135 (192.168.246.135)  9.742 ms  10.872 ms  8.320 ms
 9  a96-16-68-225.deploy.static.akamaitechnologies.com (96.16.68.225)  9.062 ms  8.902 ms  10.452 ms
```

# Networks

## How did they know to give that IP?

```
% dig whoami.akamai.net +short
172.71.157.103

% curl -s ipinfo.io/172.71.157.103 | grep city
  "city": "San Jose",
```

## Geo IP Feed?

## How do I know its mapped like this?


Let’s test it from far away

# Networks

## Test

```
 1  203.50.77.49 (203.50.77.49)  0.808 ms  0.724 ms  0.493 ms
 2  TenGigE0-0-0-21.lon-dlr20.melbourne.telstra.net (203.50.233.22)  0.744 ms  0.610 ms  0.495 ms
 3  * bundle-ether30.exi-core30.melbourne.telstra.net (203.50.11.246)  0.856 ms  3.486 ms
 4  bundle-ether2.cla-core30.melbourne.telstra.net (203.50.13.124)  1.742 ms  1.986 ms  1.494 ms
 5  bundle-ether3.hay-core30.sydney.telstra.net (203.50.13.132)  11.987 ms  11.858 ms  12.363 ms
 6  bundle-ether2.oxf-gw30.sydney.telstra.net (203.50.6.106)  15.234 ms  15.978 ms  12.114 ms
 7  203.50.13.94 (203.50.13.94)  14.860 ms  13.981 ms  12.988 ms
 8  i-10304.sydo-core03.telstraglobal.net (202.84.222.129)  13.364 ms  12.607 ms
 9  i-10104.sydp-core03.telstraglobal.net (202.84.222.137)  12.858 ms  12.980 ms
10  i-10203.sydp-core04.telstraglobal.net (202.84.222.169)  149.152 ms
11  i-10201.sydp-core04.telstraglobal.net (202.84.222.134)  13.986 ms
12  i-92.eqnx03.telstraglobal.net (202.84.247.17)  149.148 ms  147.652 ms
13  i-20802.eqnx-core02.telstraglobal.net (202.84.141.25)  148.151 ms
14  ae34.r04.border101.sjc01.fab.netarch.akamai.com (23.203.158.25)  148.276 ms
```


https://www.telstra.net/cgi-bin/trace

# Networks

## Test

```
 1  203.50.77.49 (203.50.77.49)  0.876 ms  0.723 ms  0.619 ms
 2  TenGigE0-0-0-21.lon-dlr20.melbourne.telstra.net (203.50.233.22)  0.868 ms  0.734 ms  0.495 ms
 3  bundle-ether30.exi-core30.melbourne.telstra.net (203.50.11.246)  2.992 ms  3.860 ms  4.117 ms
 4  ae10.lon-ice301.melbourne.telstra.net (203.50.61.129)  0.371 ms  0.611 ms  0.497 ms
 5  ae20-20.win-ice301.melbourne.telstra.net (203.50.61.131)  1.492 ms  0.614 ms  0.619 ms
 6  203.46.69.73 (203.46.69.73)  2.118 ms  6.612 ms  2.494 ms
```

https://www.telstra.net/cgi-bin/trace

CLOUDFLARE

# Networks

## Fastly

```
 ;; ANSWER SECTION:
www.theguardian.com.       6702          IN            CNAME         dualstack.guardian.map.fastly.net.
dualstack.guardian.map.fastly.net. 13 IN A            151.101.1.111
dualstack.guardian.map.fastly.net. 13 IN A            151.101.65.111
dualstack.guardian.map.fastly.net. 13 IN A            151.101.129.111
dualstack.guardian.map.fastly.net. 13 IN A            151.101.193.111
```

```
 % ping -c 1 www.theguardian.com
PING dualstack.guardian.map.fastly.net (151.101.193.111): 56 data bytes
64 bytes from 151.101.193.111: icmp_seq=0 ttl=55 time=7.383 ms

--- dualstack.guardian.map.fastly.net ping statistics ---
1 packets transmitted, 1 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 7.383/7.383/7.383/nan ms
```

CLOUDFLARE

# Networks

## Fastly

```
Translating "www.theguardian.com"...domain server (203.131.52.11) [OK]

Type escape sequence to abort.
Tracing the route to dualstack.guardian.map.fastly.net (151.101.29.111)

  1 gi9-9.sglebdist01.nw.aapt.net.au (202.10.15.152) 0 msec 0 msec 0 msec
  2 bu11.sglebcore01.aapt.net.au (202.10.12.7) [MPLS: Label 24389 Exp 1] 4 msec 0 msec 4 msec
  3 gi3-0-3025.nSYDNpe05.aapt.net.au (203.131.60.81) 8 msec 4 msec 0 msec
  4 59-100-201-110.syd.static-ipl.aapt.com.au (59.100.201.110) 0 msec 4 msec 0 msec
  5 ae-14.r00.sydnau04.au.bb.gin.ntt.net (129.250.9.242) [AS 2914] 0 msec
    ae-29.a00.sydnau05.au.bb.gin.ntt.net (129.250.9.218) [AS 2914] 12 msec
    ae-14.r00.sydnau04.au.bb.gin.ntt.net (129.250.9.242) [AS 2914] 4 msec
  6  *
    ae-7.r21.sydnau06.au.bb.gin.ntt.net (129.250.3.176) [AS 2914] 4 msec 0 msec
  7 ae-2.a00.sydnau05.au.bb.gin.ntt.net (129.250.7.49) [AS 2914] 4 msec *  *
  8  *  *  *
```

CLOUDFLARE

```
http://looking-glass.connect.com.au/lg
```

# Networks

## Fastly

```
% traceroute -I 151.101.29.111
traceroute to 151.101.29.111 (151.101.29.111), 64 hops max, 72 byte packets
 1  unifi (192.168.8.1)  4.904 ms  2.759 ms  3.492 ms
 2  192.168.99.254 (192.168.99.254)  4.169 ms  3.483 ms  3.250 ms
 3  45-26-52-1.lightspeed.sntcca.sbcglobal.net (45.26.52.1)  9.593 ms  13.597 ms  2.987 ms
 4  71.148.165.2 (71.148.165.2)  4.684 ms  19.810 ms  2.770 ms
 5  12.242.117.22 (12.242.117.22)  7.300 ms  13.908 ms  12.989 ms
 6  192.205.37.58 (192.205.37.58)  35.666 ms  16.227 ms  10.637 ms
 7  ae-9.r24.snjsca04.us.bb.gin.ntt.net (129.250.2.2)  16.560 ms  8.627 ms  13.175 ms
 8  ae-5.r26.osakjp02.jp.bb.gin.ntt.net (129.250.2.119)  120.932 ms  118.449 ms  118.855 ms
 9  ae-6.r20.sydnau05.au.bb.gin.ntt.net (129.250.4.65)  198.109 ms  197.105 ms  195.493 ms
10  ae-0.a00.sydnau05.au.bb.gin.ntt.net (129.250.2.132)  198.406 ms  195.132 ms  195.695 ms
11  ae-0.fastly.sydnau05.au.bb.gin.ntt.net (202.68.66.118)  193.776 ms  194.585 ms  194.509 ms
12  151.101.29.111 (151.101.29.111)  197.975 ms  194.521 ms  194.080 ms
```

CLOUDFLARE

# Application/Network

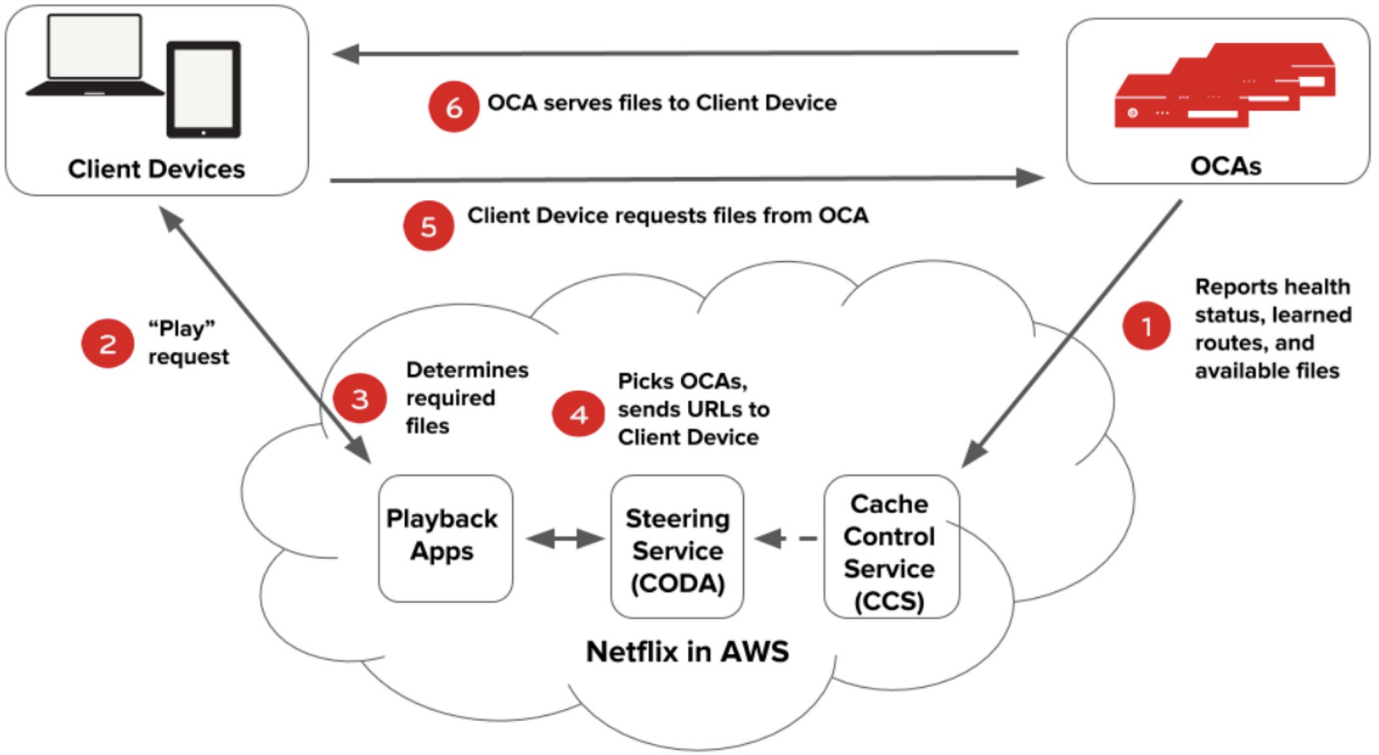
Companies like Netflix, run their own Network, and have routing built into their application!

You can ask DNS, but it won't help you for the major bit flows

```
;; ANSWER SECTION:
www.netflix.com.        90          IN          CNAME       www.dradis.netflix.com.
www.dradis.netflix.com. 30          IN          CNAME       www.us-west-
2.internal.dradis.netflix.com.
www.us-west-2.internal.dradis.netflix.com. 30 IN CNAME apiproxy-website-nlb-prod-2-
e98cb8cf33ff3581.elb.us-west-2.amazonaws.com.
apiproxy-website-nlb-prod-2-e98cb8cf33ff3581.elb.us-west-2.amazonaws.com. 30 IN     A 44.237.234.25
apiproxy-website-nlb-prod-2-e98cb8cf33ff3581.elb.us-west-2.amazonaws.com. 30 IN     A 44.242.60.85
apiproxy-website-nlb-prod-2-e98cb8cf33ff3581.elb.us-west-2.amazonaws.com. 30 IN     A 44.234.232.238
```

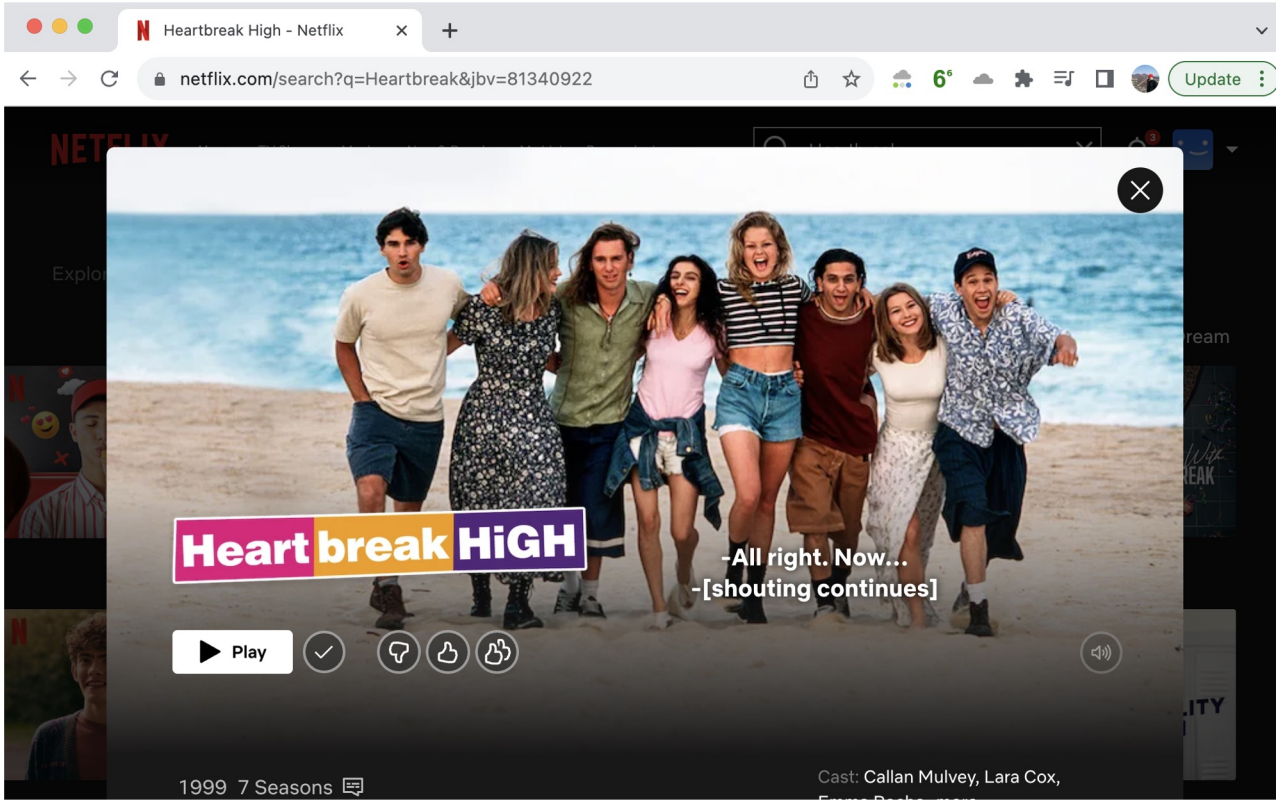
# Application/Network

The following diagram illustrates how the playback process works:



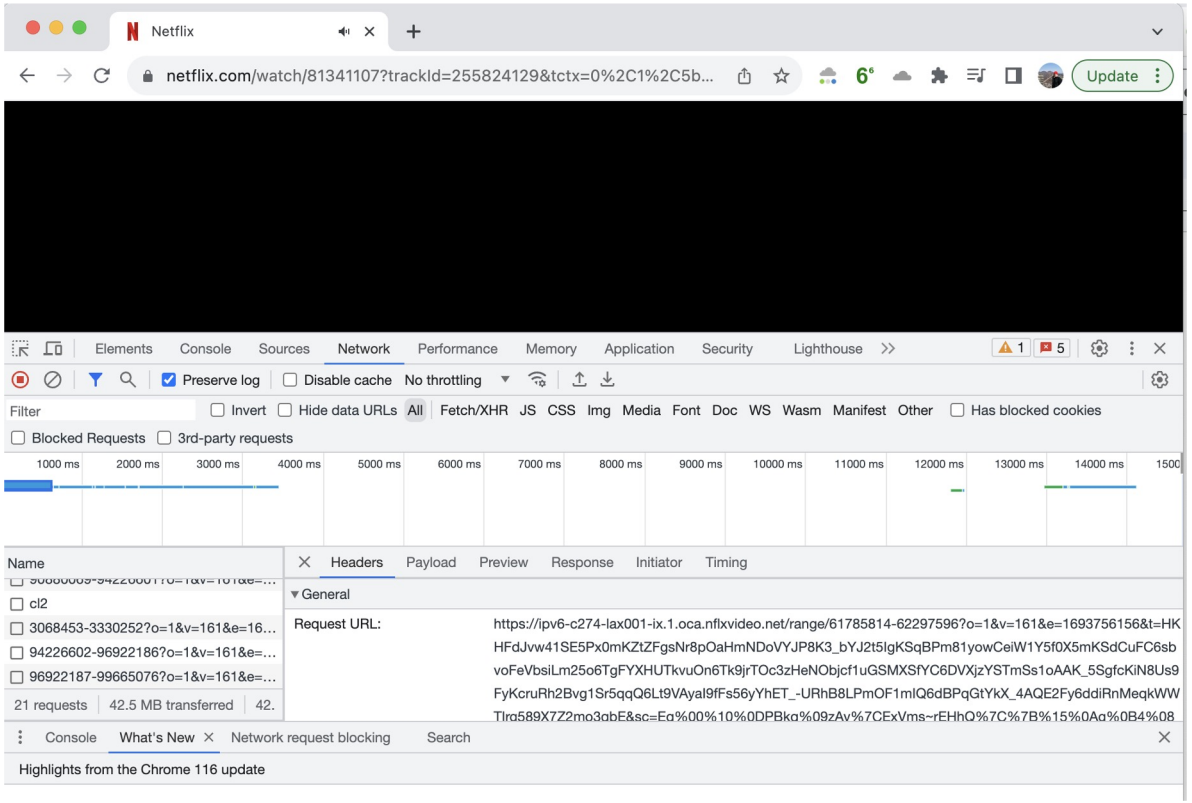

https://openconnect.netflix.com/Open-Connect-Overview.pdf

# Application/Network




CLOUDFLARE

# Application/Network




CLOUDFLARE

# Application/Network

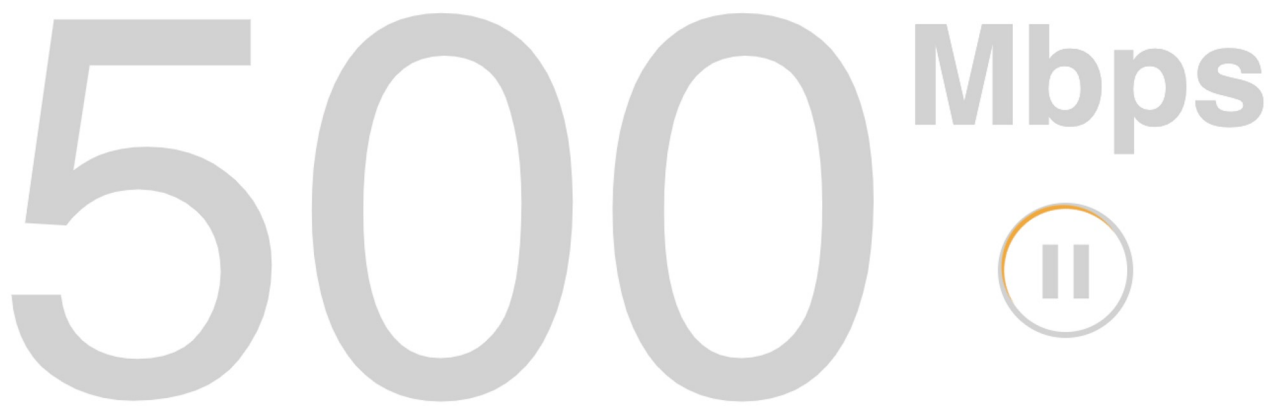
Request URL: https://ipv6-c242-sjc002-ix.1.oca.nflxvideo.net/range/57736523-59868423?o=18
64&t=YVt2uclKe7dfys1E7zRADsqlsYgGlfFpl6lKEpDFTXSlVF-HykdJfVgCsTlJ-gL

So, the application is giving this answer, we can't really see why? But there is a way!

CLOUDFLARE

# Application/Network

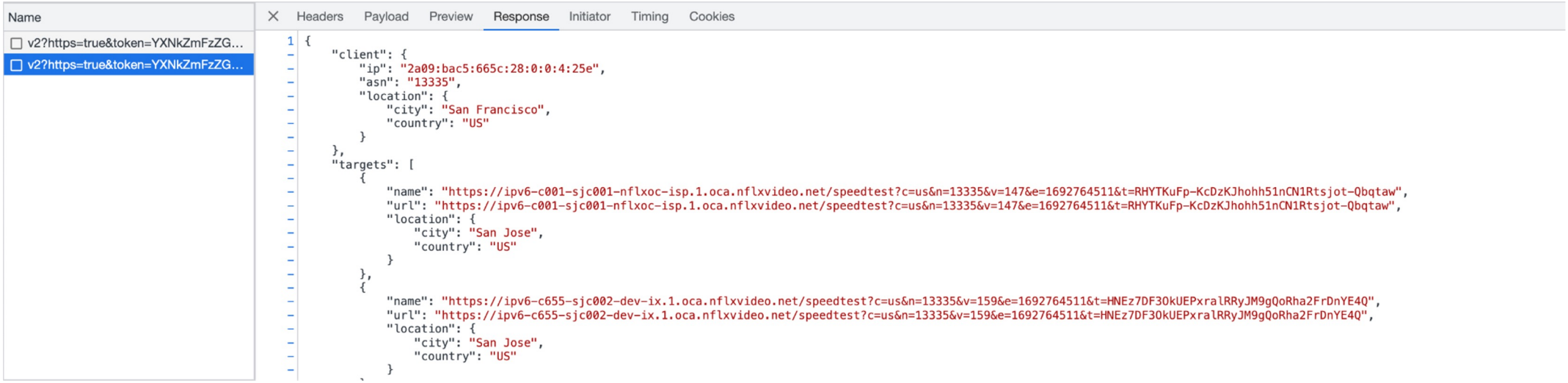
fast.com


CLOUDFLARE

# Application/Network

fast.com exposes these URLs




CLOUDFLARE

# Application/Network

fast.com exposes these URLs

With some reverse engineering, can run this from the command line

# Application/Network

```
% for apiKey in `curl -s https://fast.com/app-a32983.js | awk -F "apiEndpoint,token" '{print $2}' |  awk -F "\"" '{print $2}' |
grep -Ev "^$"`; do curl -s "https://api.fast.com/netflix/speedtest/v2?https=true&token=$apiKey&urlCount=5" | jq . ; done

{
 "client": {
    "ip": "2a09:bac5:665c:28:0:0:4:25e",
    "asn": "13335",
    "location": {
      "city": "San Francisco",
      "country": "US"
    }
  },
  "targets": [
    {
      "name": "https://ipv6-c002-sjc001-nflxoc-
isp.1.oca.nflxvideo.net/speedtest?c=us&n=13335&v=159&e=1692764735&t=uAXwXXWHH5P7TpvxalwwIo4SuC1ayY2e2a9-bg",
      "url": "https://ipv6-c002-sjc001-nflxoc-
isp.1.oca.nflxvideo.net/speedtest?c=us&n=13335&v=159&e=1692764735&t=uAXwXXWHH5P7TpvxalwwIo4SuC1ayY2e2a9-bg",
      "location": {
        "city": "San Jose",
        "country": "US"
      }
    },
```

CLOUDFLARE

# Application/Network

But how?

https://openconnect.zendesk.com/hc/en-us/articles/115001068691-Managing-BGP-sessions

Netflix is using **BGP** to their embedded OCA nodes (if you host one), from their IX sessions (if you IX Peer or have PNI with them), or from your transit networks sessions with them.

# Application

How do the applications work?

Looking at Disney+


CLOUDFLARE

# Application

```
;; ANSWER SECTION:
vod-akc-na-west-1.media.dssott.com. 3458 IN CNAME vod-akc-na-west-1.media.dssott.com.akamaized.net.
vod-akc-na-west-1.media.dssott.com.akamaized.net. 458 IN CNAME a1851.dscw80.akamai.net.
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.41
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.17
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.24
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.35
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.32
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.34
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.27
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.33
a1851.dscw80.akamai.net. 14         IN          A           23.56.3.26
```

# Important Tools

# Tools

GeoIP Feeds:

- Maxmind most ubiquitous.
- Can test your IPs: https://www.maxmind.com/en/geoip-demo

# Tools

GeoIP Feeds:

- Important to publish a GeoIP feed. If you host a Google GGC, they probably already ask you for one :)
- Format in RFC8805:

  https://www.rfc-editor.org/rfc/rfc8805.txt

```
% curl -s https://api.cloudflare.com/local-ip-ranges.csv | grep 104.28.196.203
104.28.196.203/32,AU,AU-NSW,Sydney,
```

# Tools

GeoIP Feeds:

- RFC 9092 - Finding and Using Geofeed Data
- https://github.com/massimocandela/geofeed-finder

- Another great resource: https://thebrotherswisp.com/index.php/geo-and-vpn/

CLOUDFLARE

# Tools

EDNS-Client-Subnet (ECS)

RFC 7871

Google has good documentation:
https://developers.google.com/speed/public-dns/docs/ecs

# Tools

## BGP

# What’s Next

# Next?

If this was useful, I'll create a Github page, where anyone can help edit and provide information as to how this mapping occurs, to help share and spread knowledge!

Questions?

Thank you!