
facebook
INFRASTRUCTURE

Building a Billion User Load Balancer

Patrick Shuff
Production Engineer, Traffic Team

facebook

We'll be talking about

Serving Dynamic Facebook Requests

Created by Arthur Shlain

L4/L7 Load Balancing

Edge PoP and Reducing Latency

Created by Alexandria Eddings from the Noun Project

Global DNS Load Balancing

Q&A

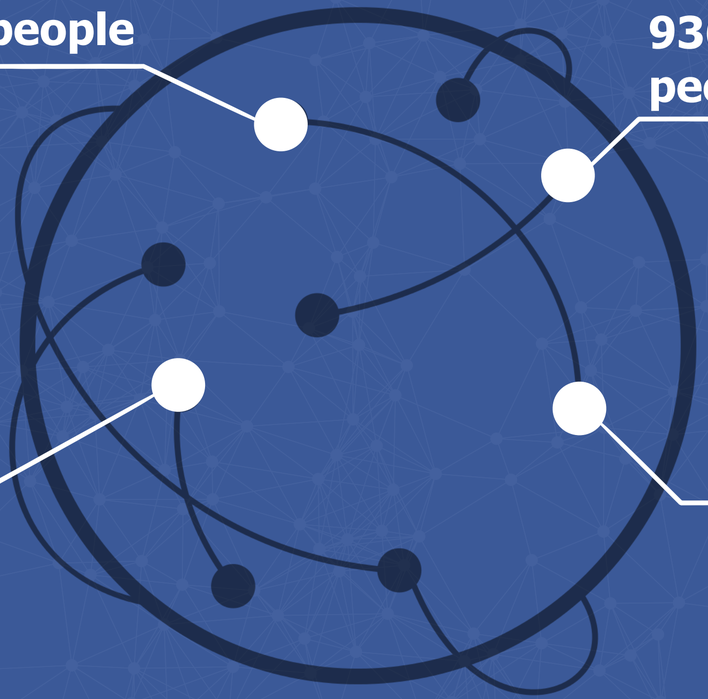
Traffic @ fb

Facebook scale

as of March 2015

1.25 billion mobile monthly active people

936 million daily active people on average



1.44 billion monthly active people

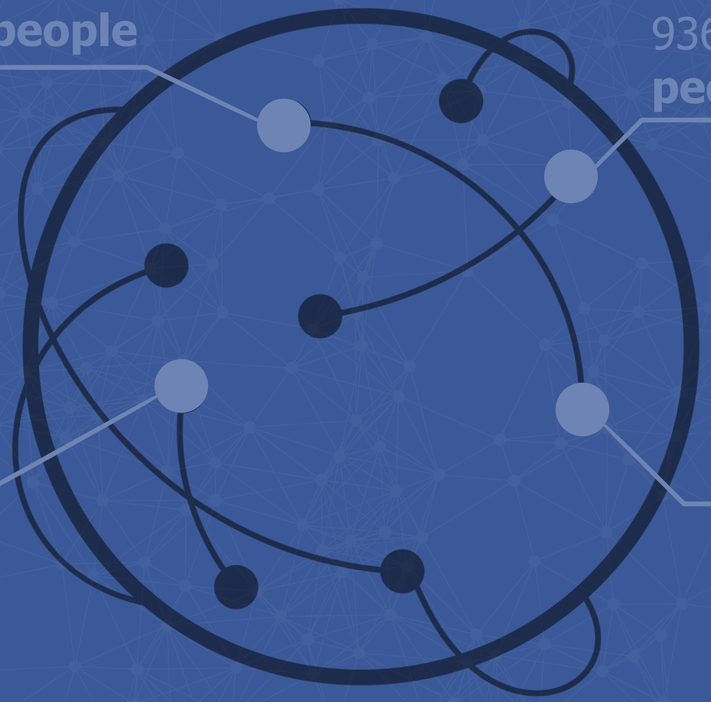
798 million mobile daily active people on average

Facebook scale

as of March 2015

1.25 billion mobile monthly active people

936 **million daily active people** on average



1.44 billion monthly active people

798 million mobile daily active people on average

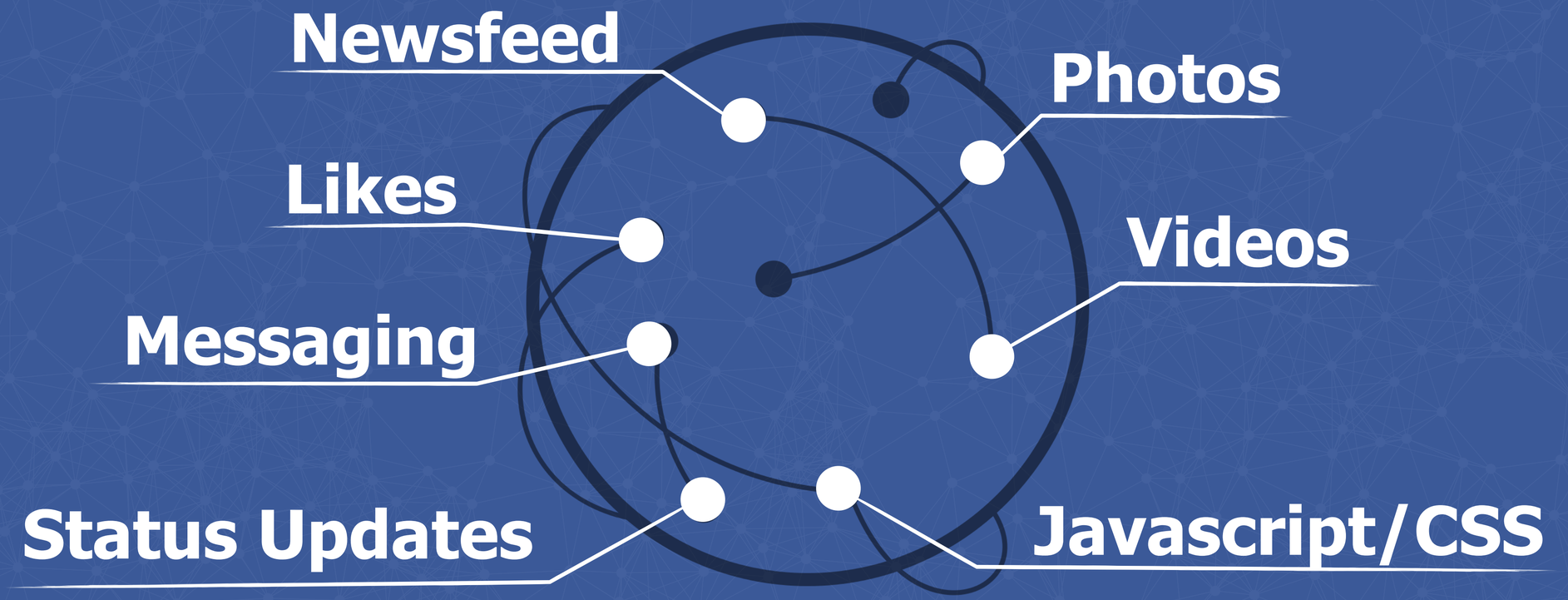
Approximately 82.4% of our daily active users are outside the US and Canada

What is facebook?

(from traffic's perspective)

Dynamic Requests

Static Requests

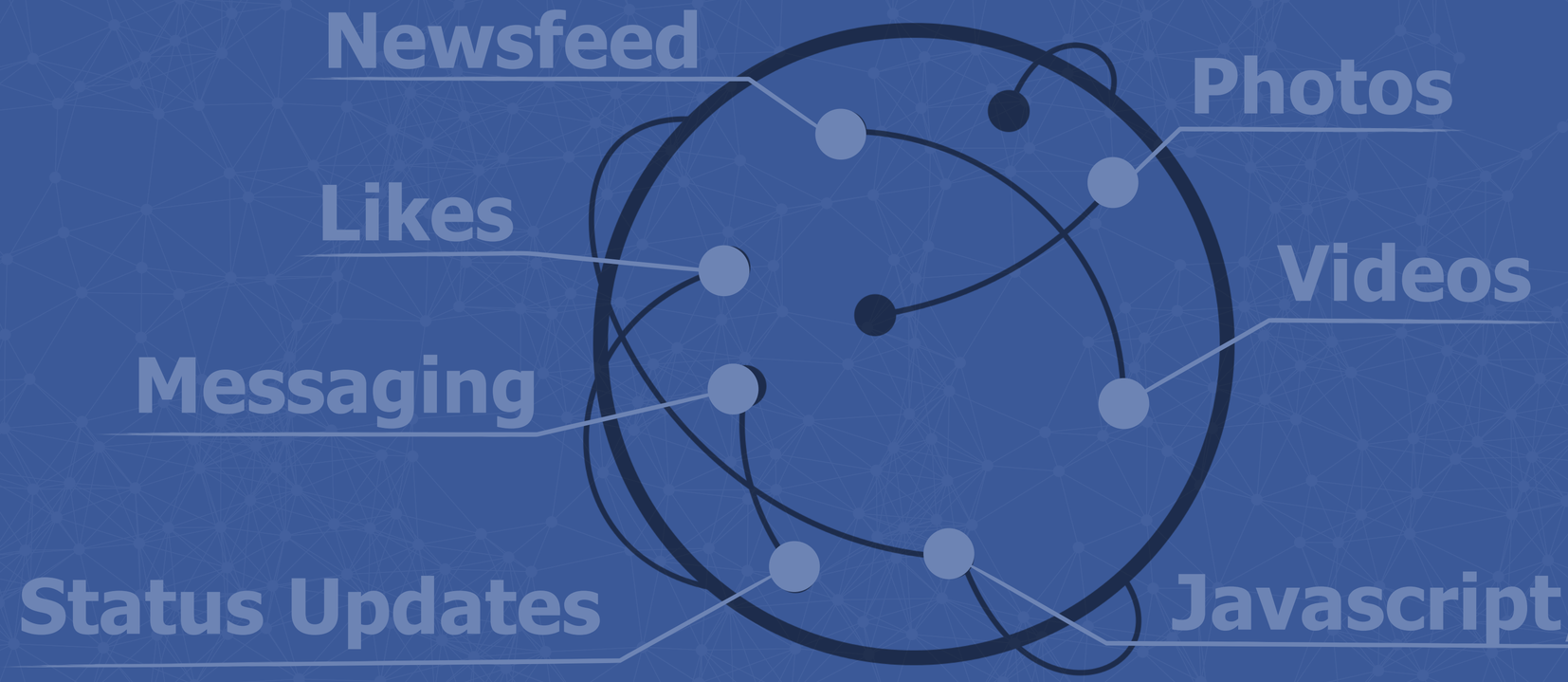


What is facebook?

(from traffic's perspective)

Dynamic Requests

Static Requests



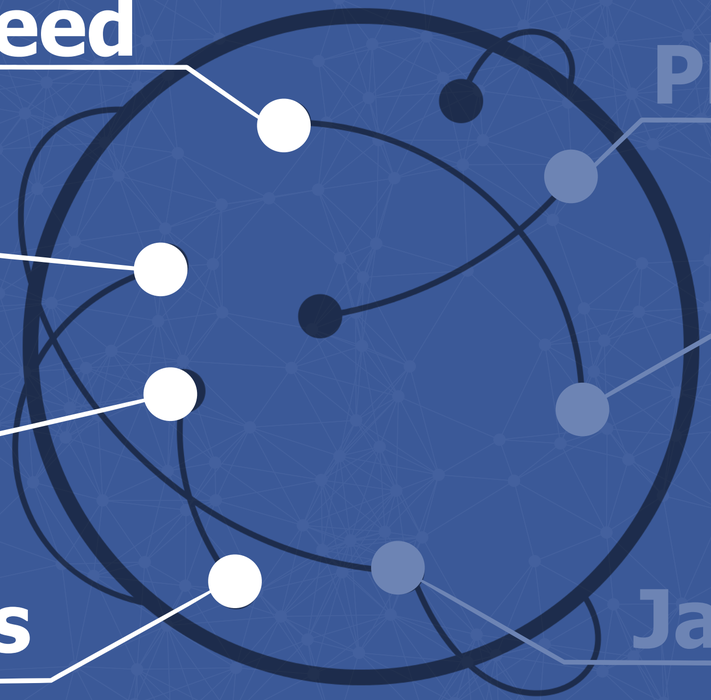
Terabits of egress (outgoing bits per second)

What is facebook?

(from traffic's perspective)

Dynamic Requests

Static Requests



Newsfeed

Likes

Messaging

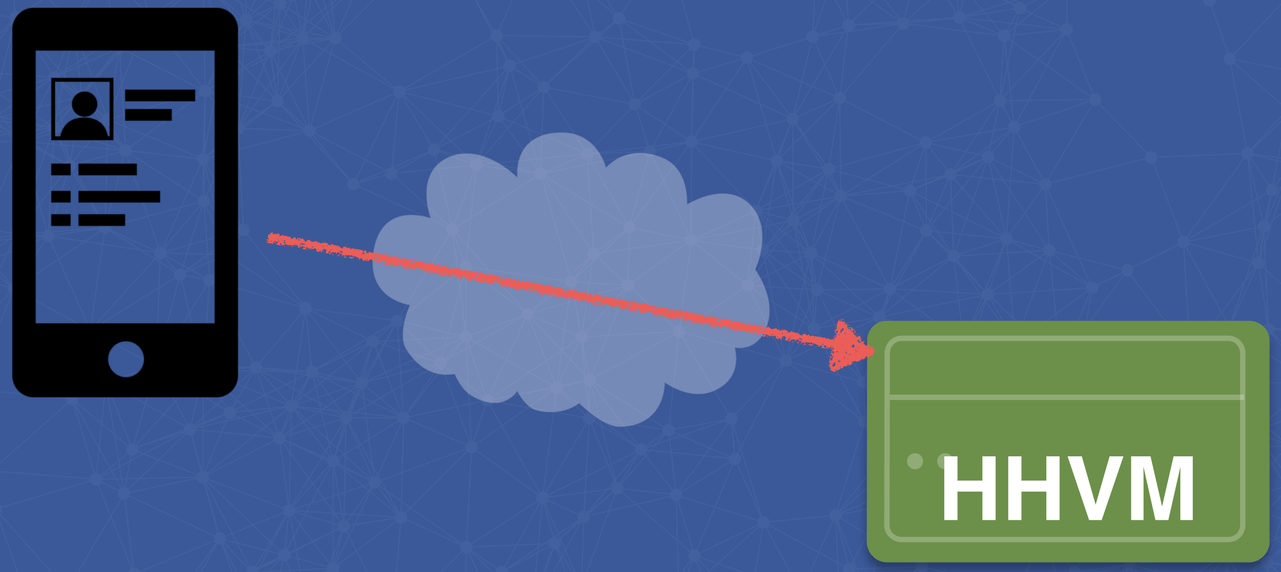
Status Updates

Photos

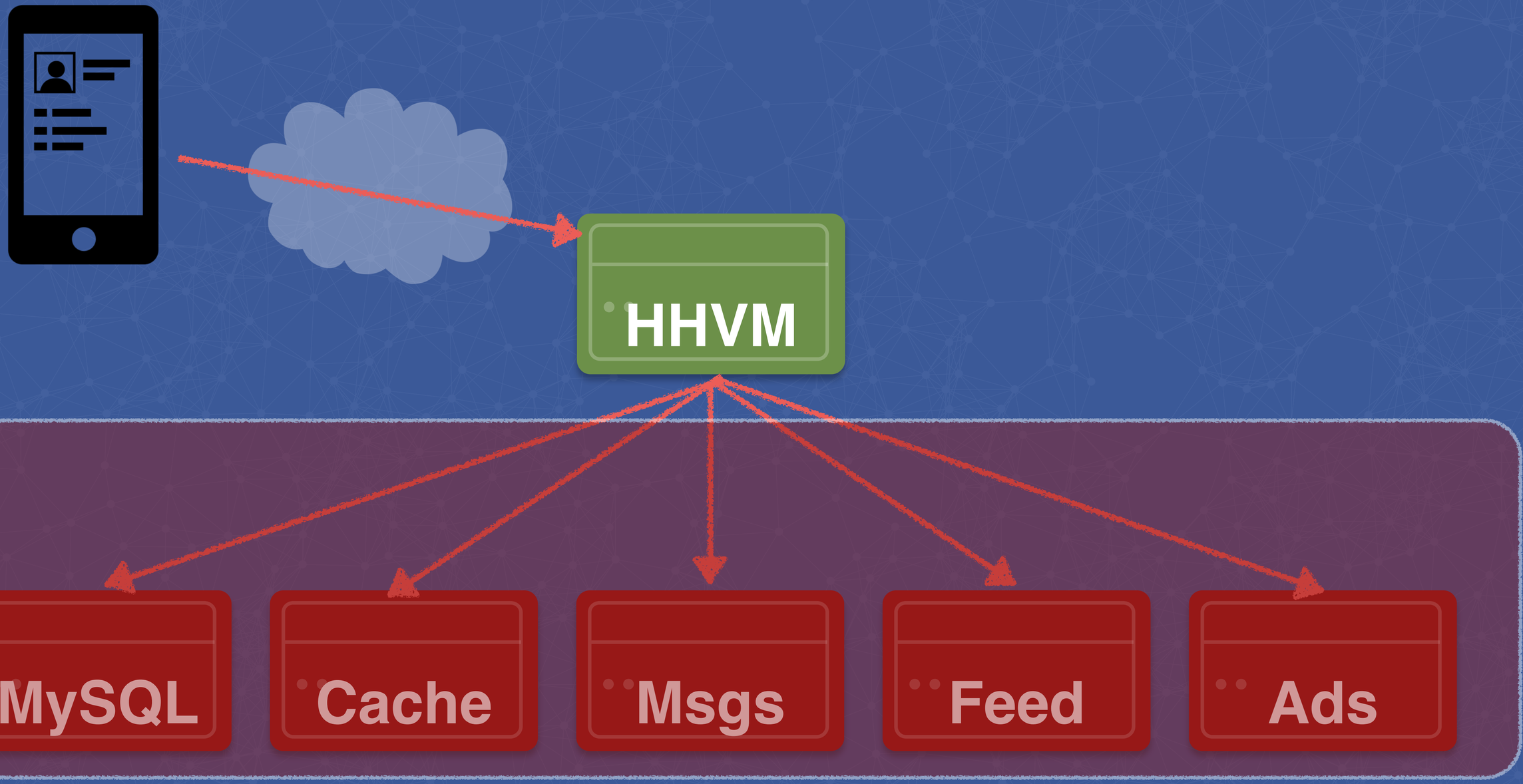
Videos

Javascript/CSS

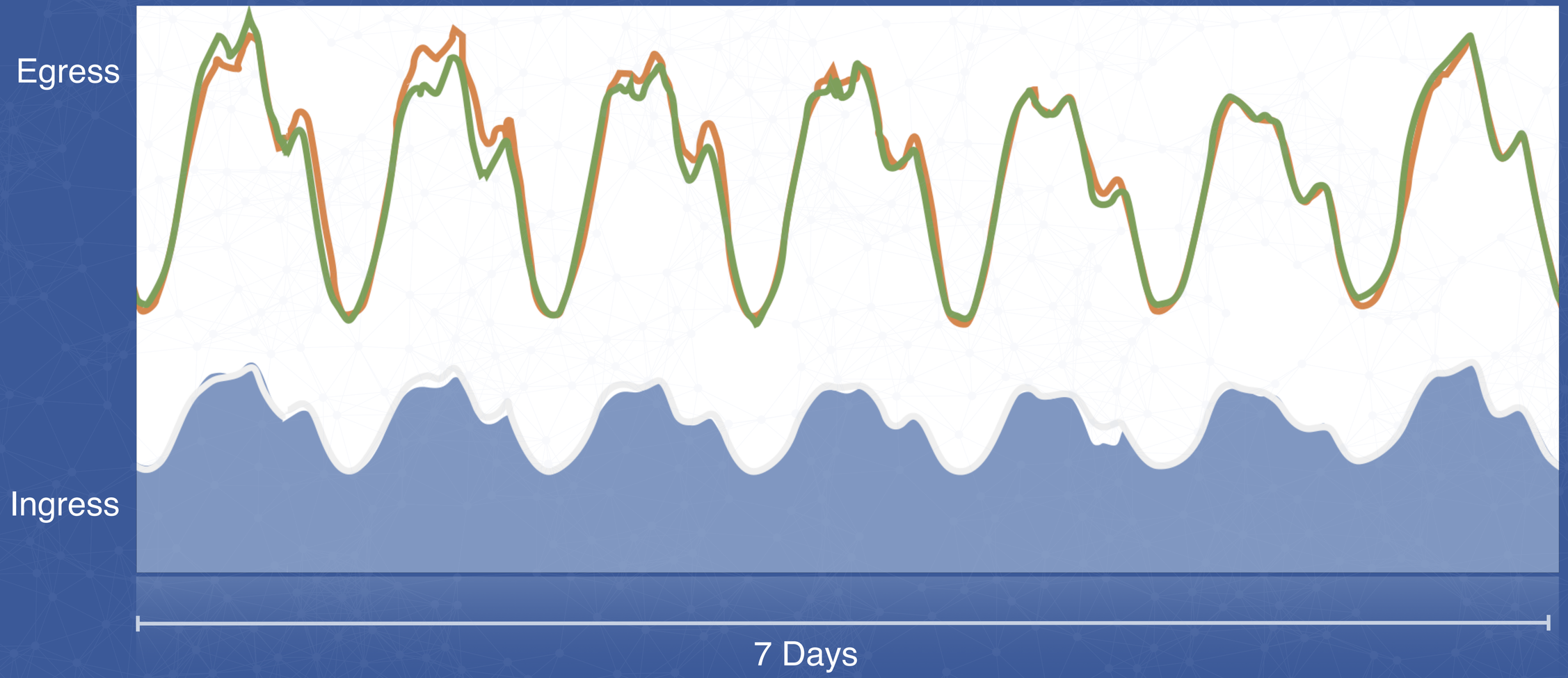
What are we talking about?



What are we not talking about?

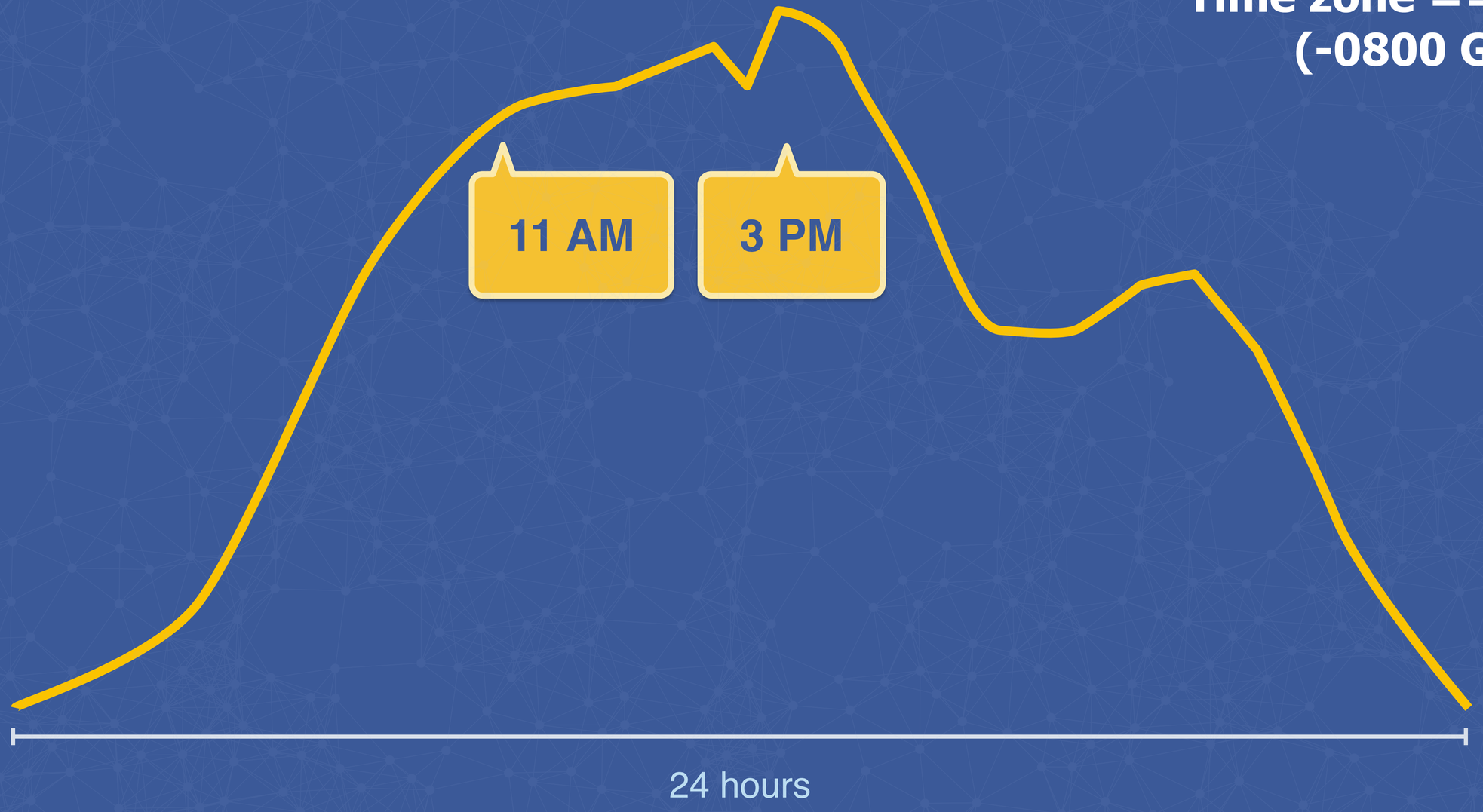


Weekly egress cycle



Diurnal egress Cycle

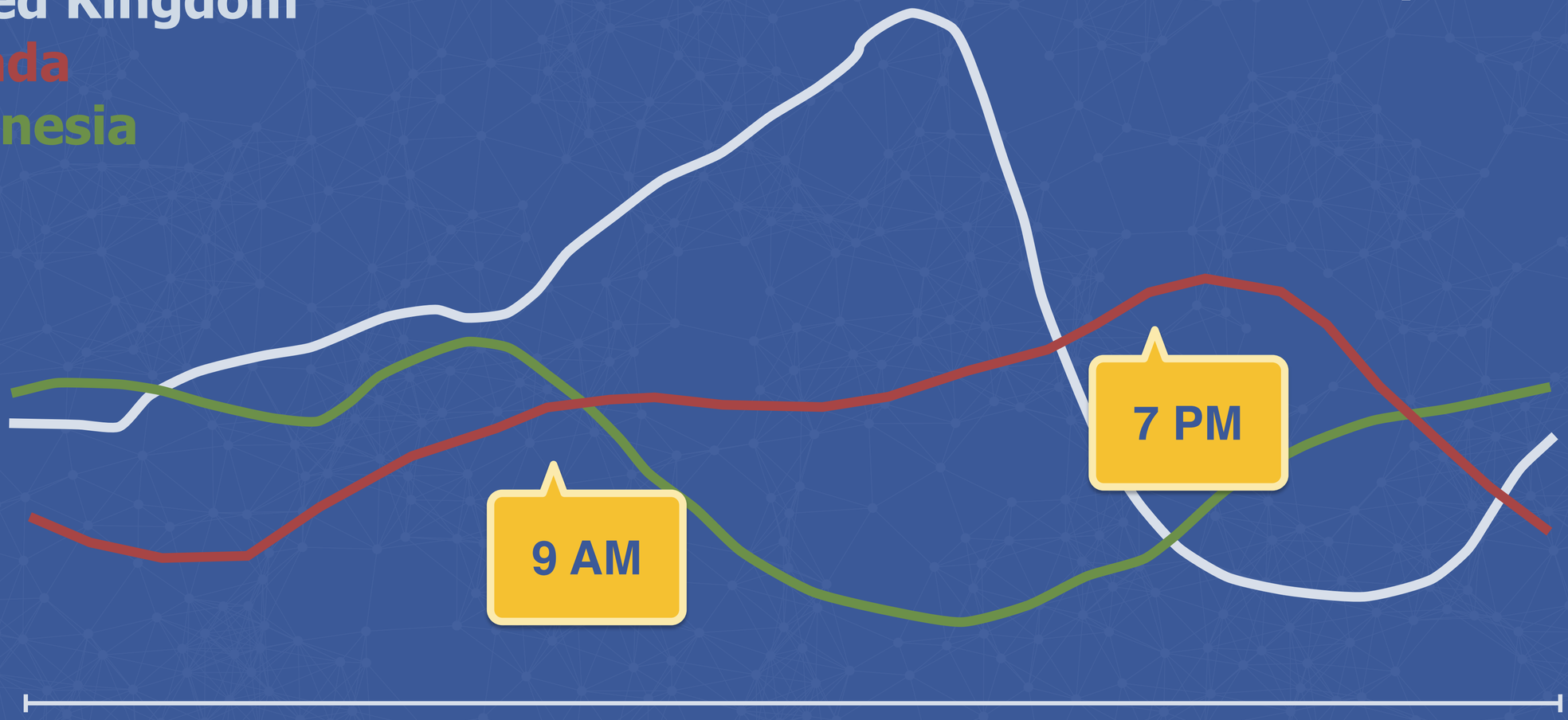
Time zone == Pacific (-0800 GMT)



Sum of timezones

Time zone == Pacific
(-0800 GMT)

United Kingdom
Canada
Indonesia

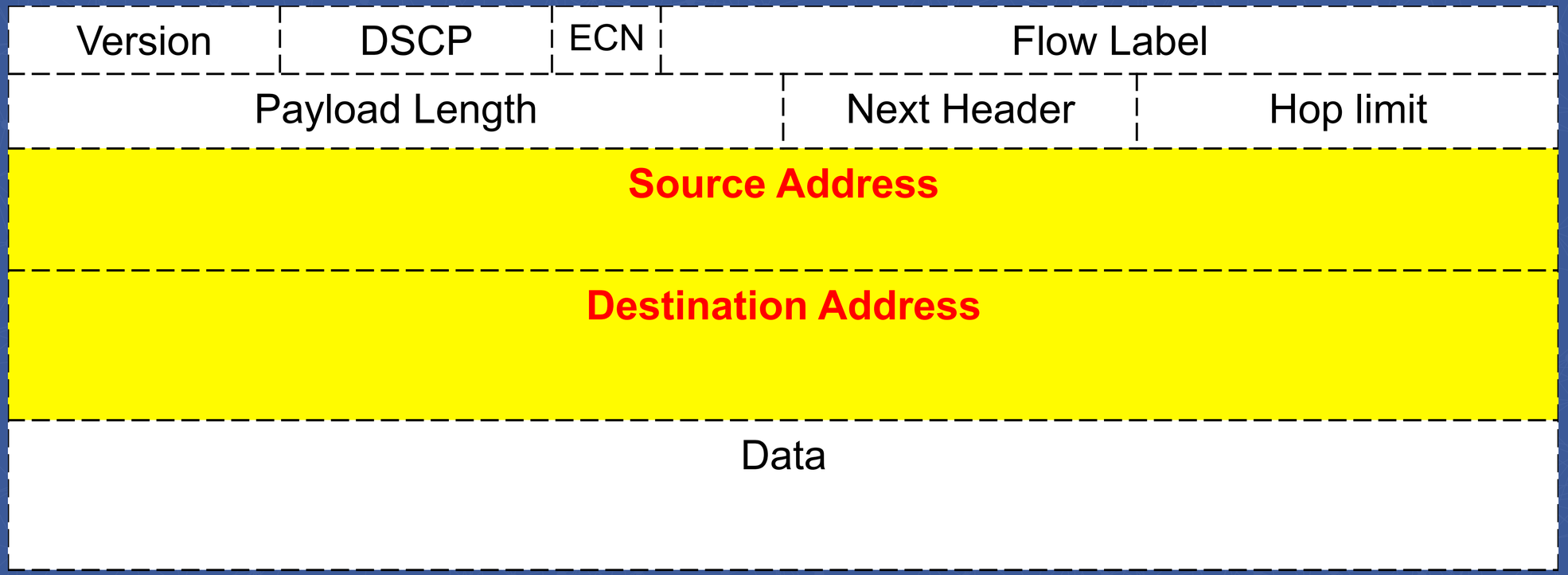


TCP/IP Review

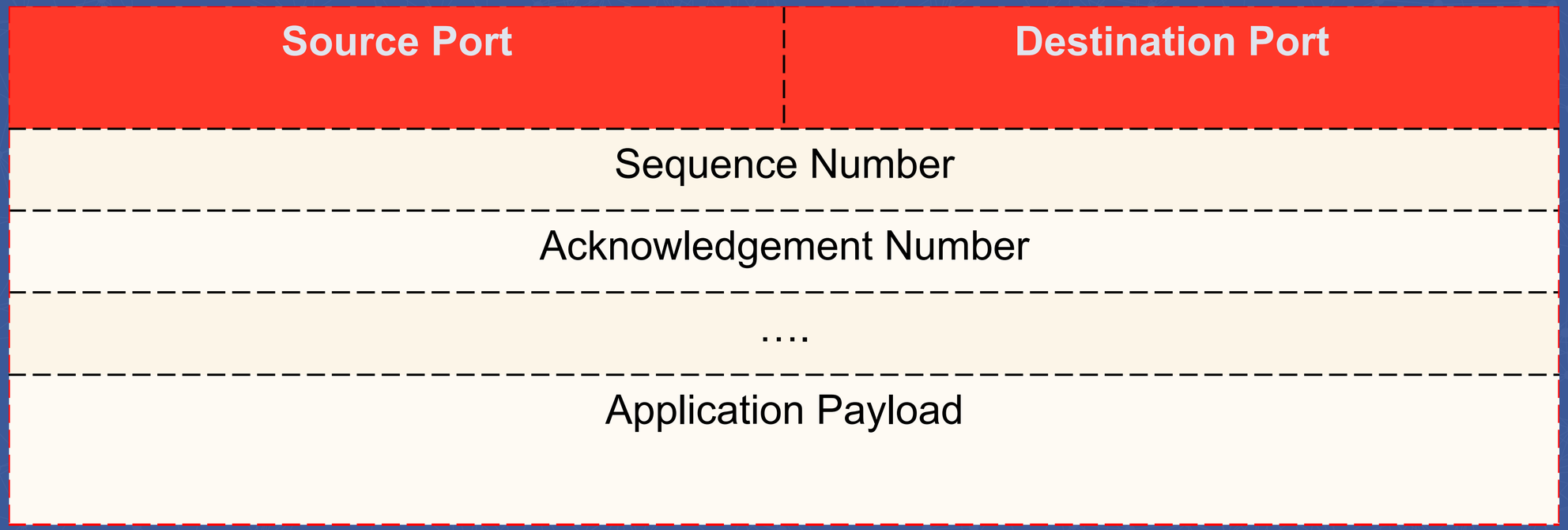
OSI Model

Layer	Purpose	Ex
7: Application	**High-Level API**	**HTTP**, SPDY, MQTT
6: Presentation	Data Translation	ASCII, JPEG
5: Session	Communication Session	RPC
4: Transport	**Transmission**	**TCP**, UDP
3: Network	**Address, Routing, Flow**	**IPv6**, IPv4
2: Data Link	Reliable Physical Comm.	IEEE, 802.2
1: Physical	Raw bit transmission	DSL, USB

IP Header (OSI Layer 3)



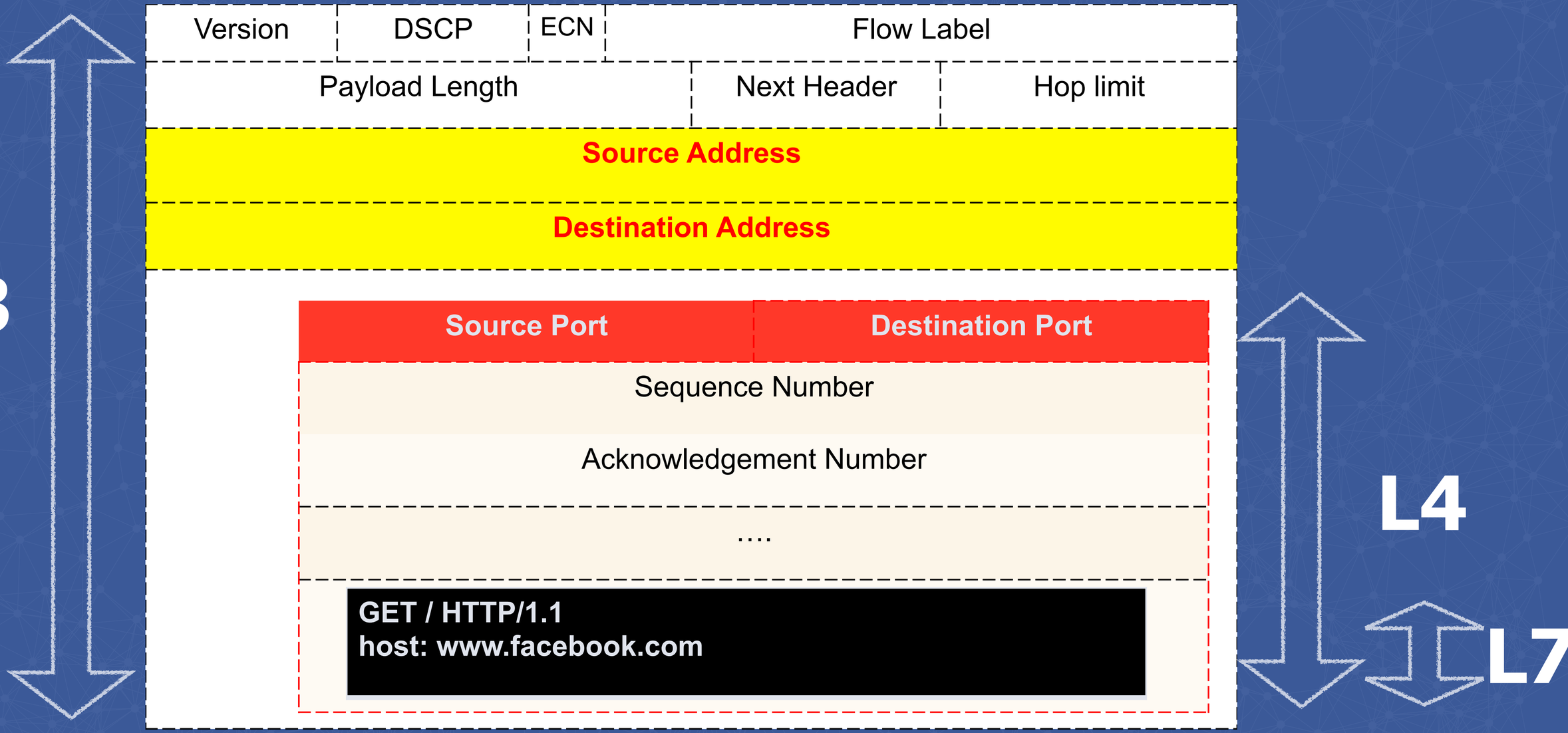
TCP Header (OSI Layer 4)



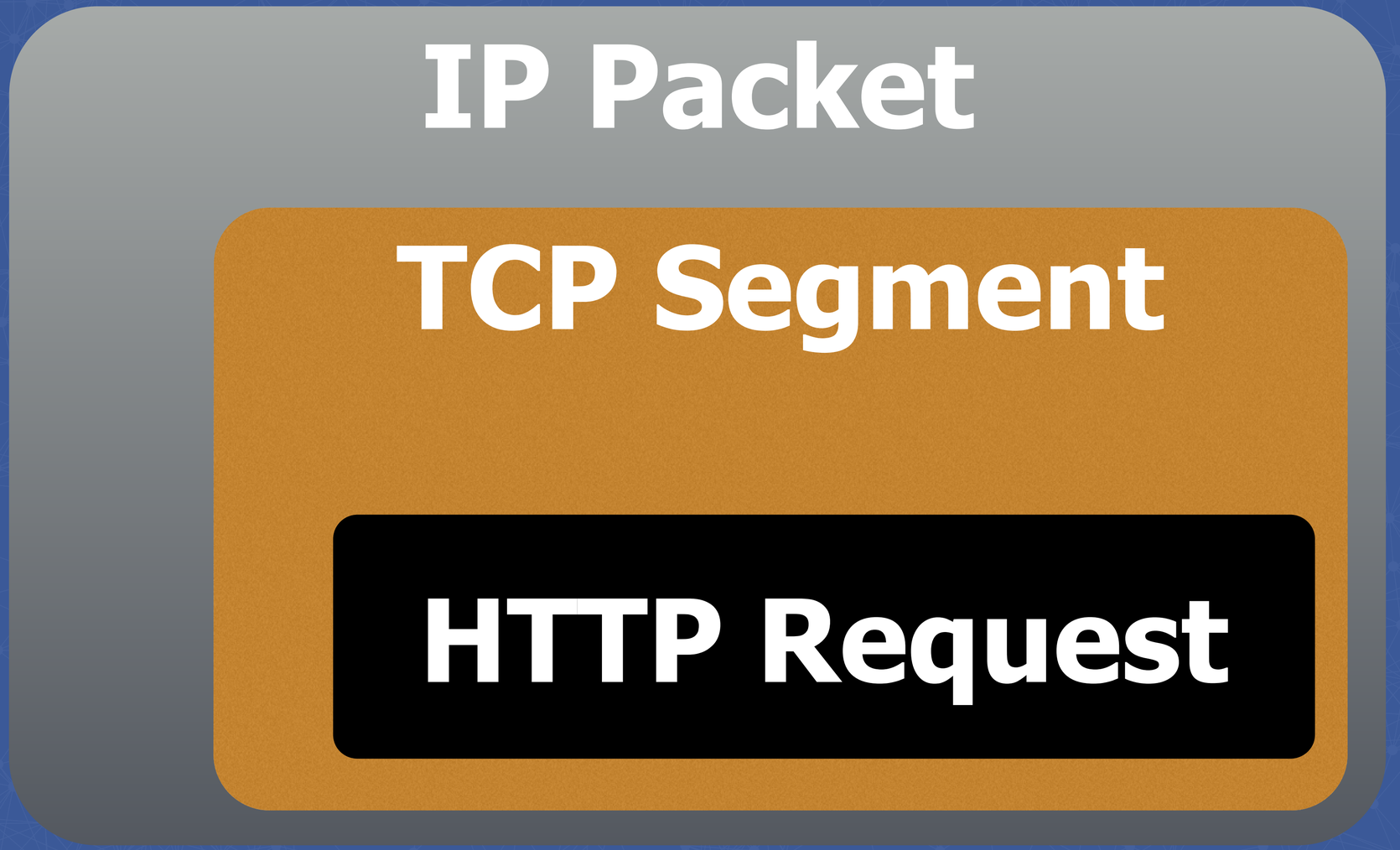
HTTP Request (OSI Layer 7)

```
GET / HTTP/1.1
host: www.facebook.com
```

Putting it all together



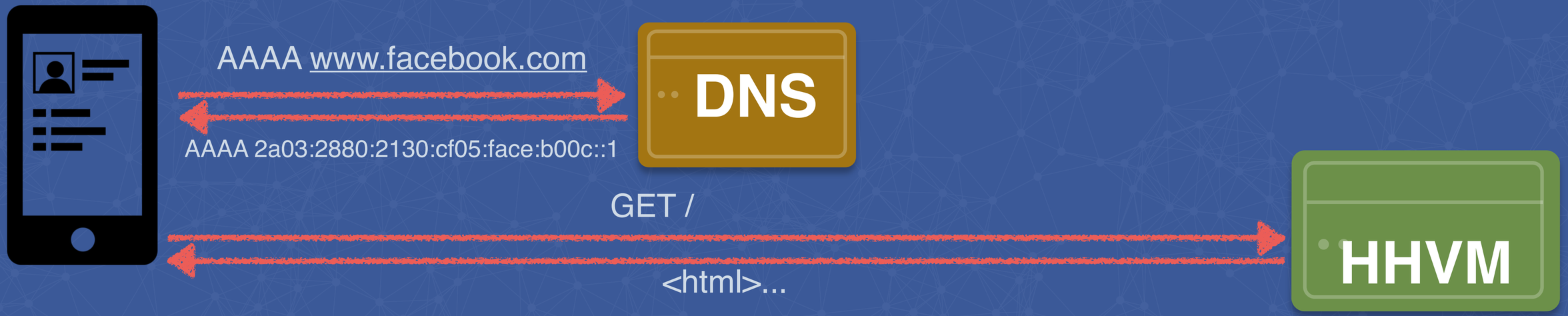
Putting it all together



Serving Dynamic Facebook Requests

FB Request -- one web server

rps = requests per second

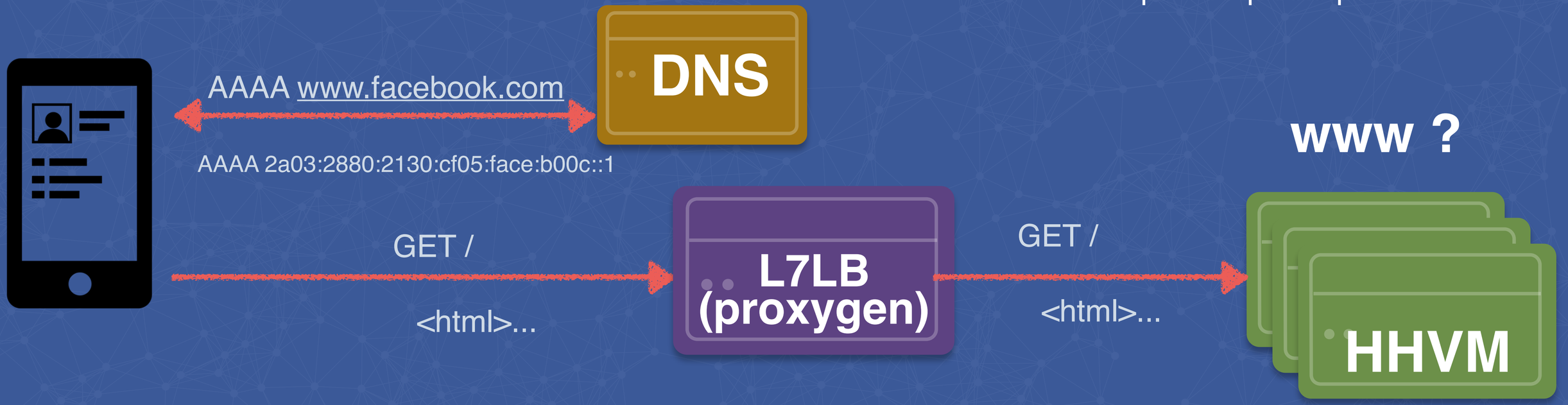


low rps

how do we get more rps?!

Add a load balancer!

rps = requests per second

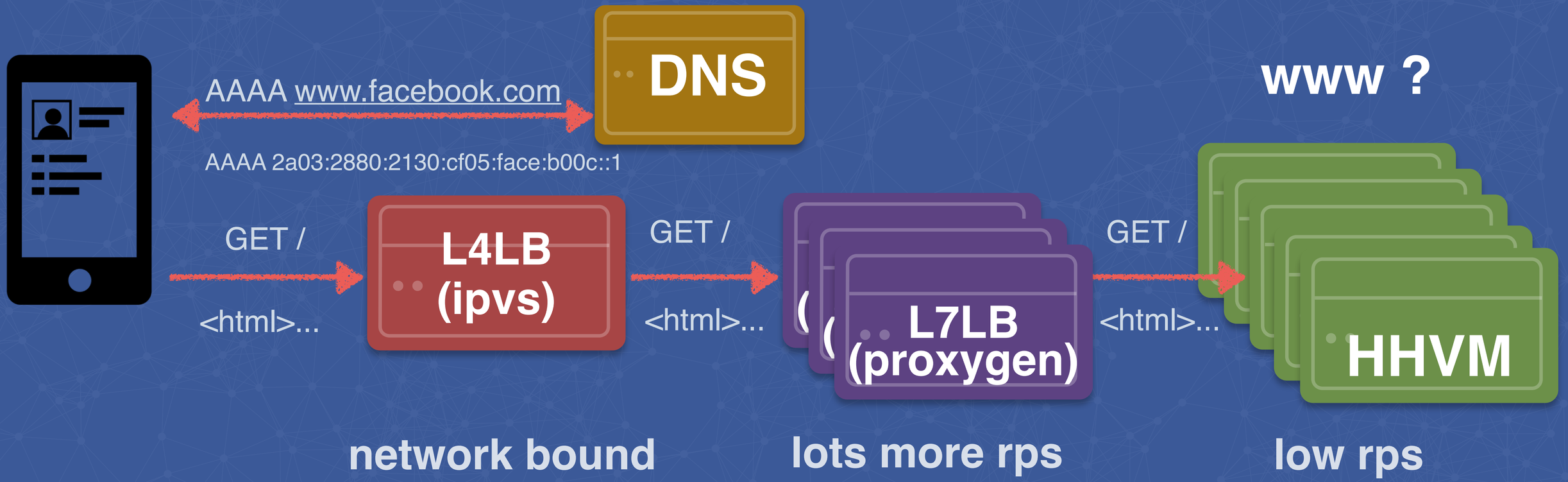


lots more rps

low rps

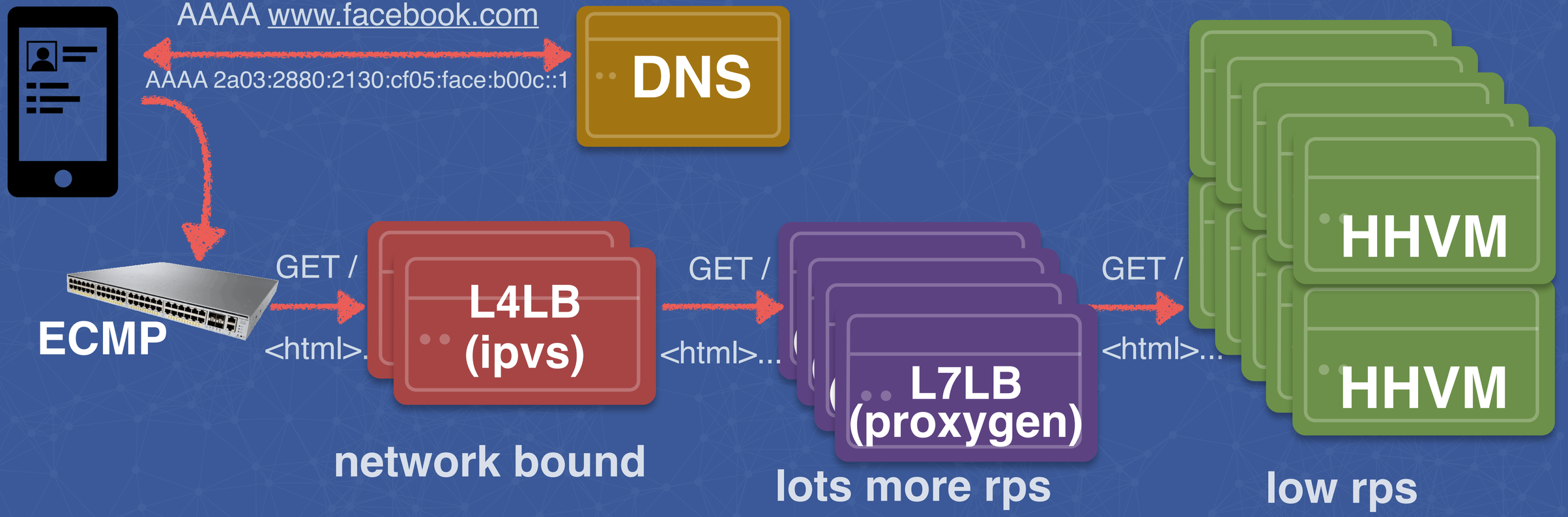
how do we get more rps?!

Add another load balancer!



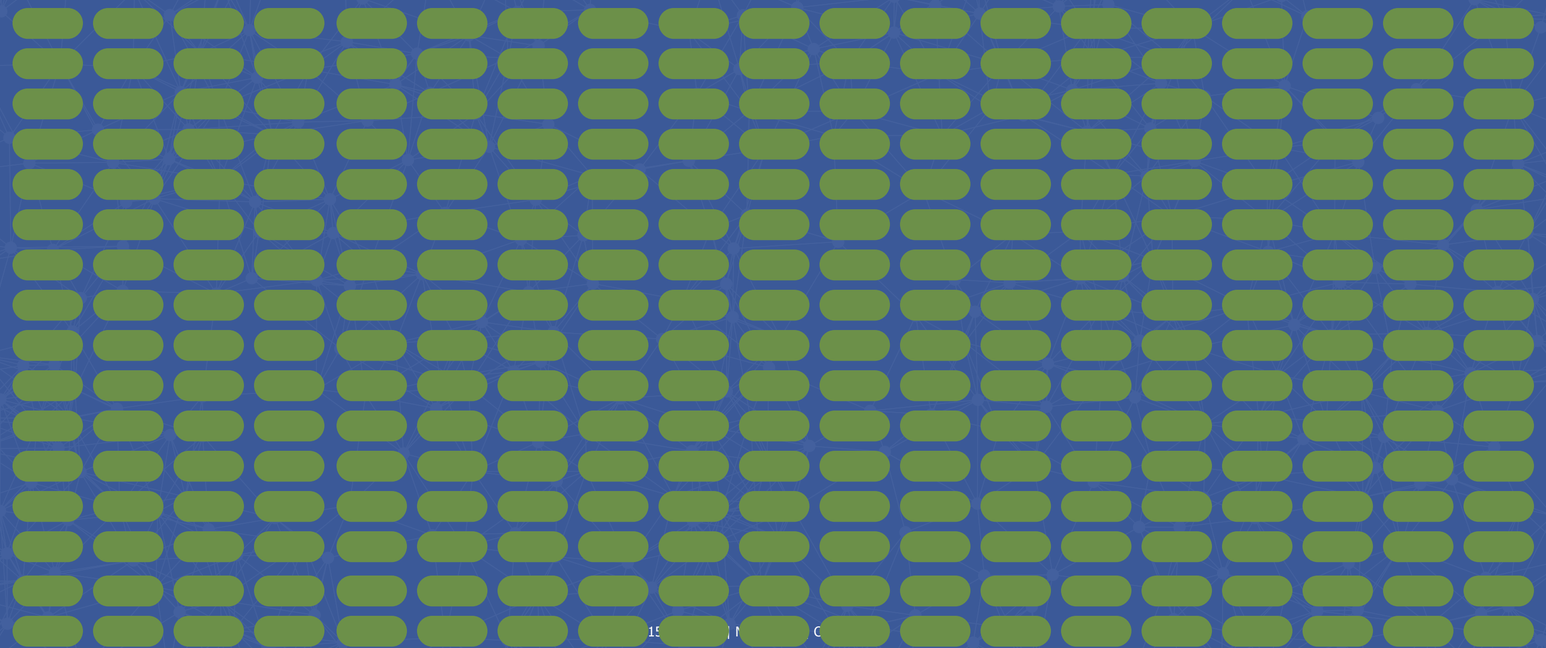
how do we get more rps?!

Add another load balancer!

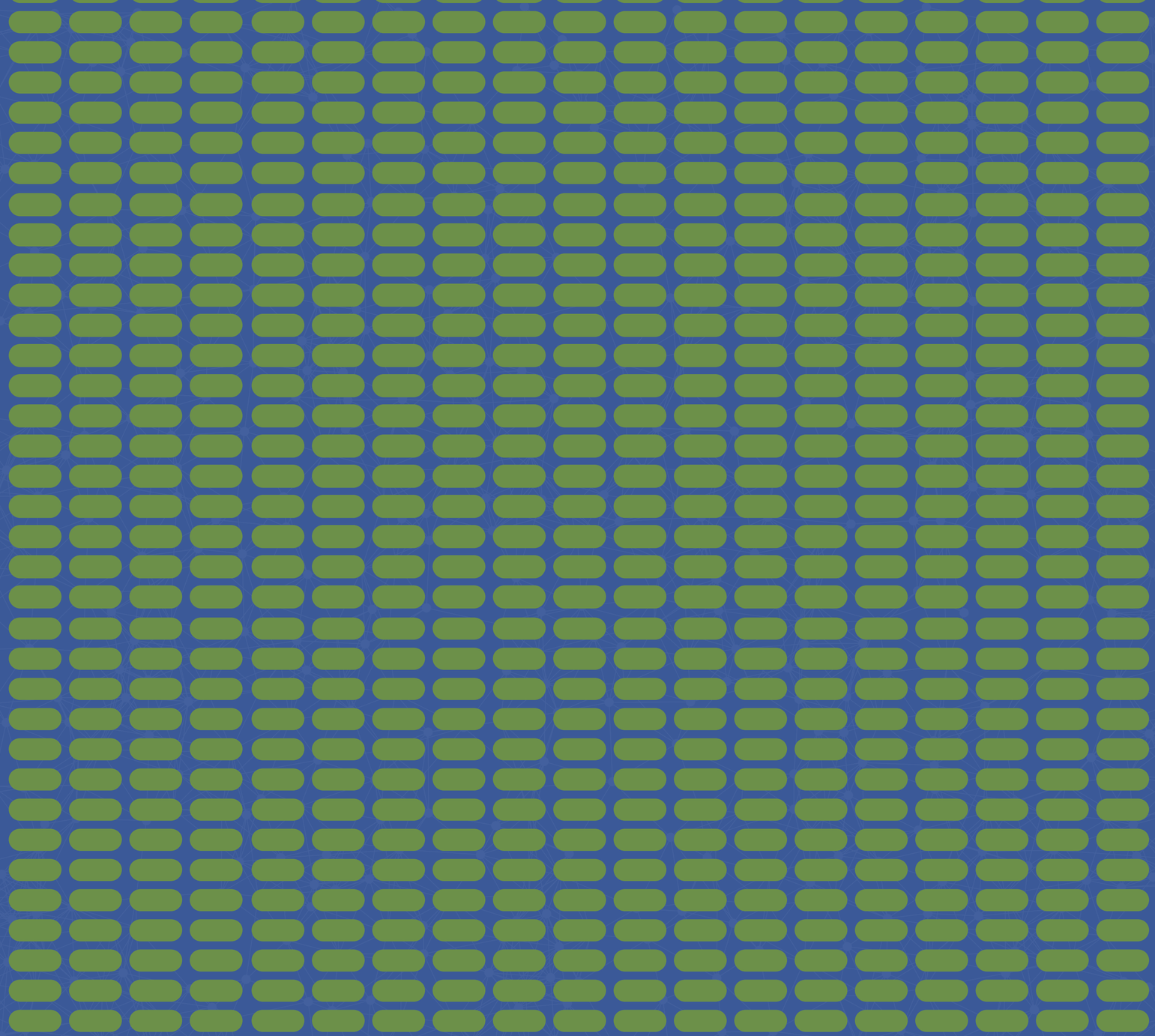


Front end Web Cluster

Thousands



cont.



x 10 or more

More RPS? Add another cluster!

AAAA www.facebook.com

AAAA 2a03:2880:2130:cf05:face:b00c::1

DNS

GET /

<html>...

L4LB (ipvs)

GET /

<html>...

L7LB (proxygen)

GET /

<html>...

HHVM

GET /

<html>...

L4LB (ipvs)

GET /

<html>...

L7LB (proxygen)

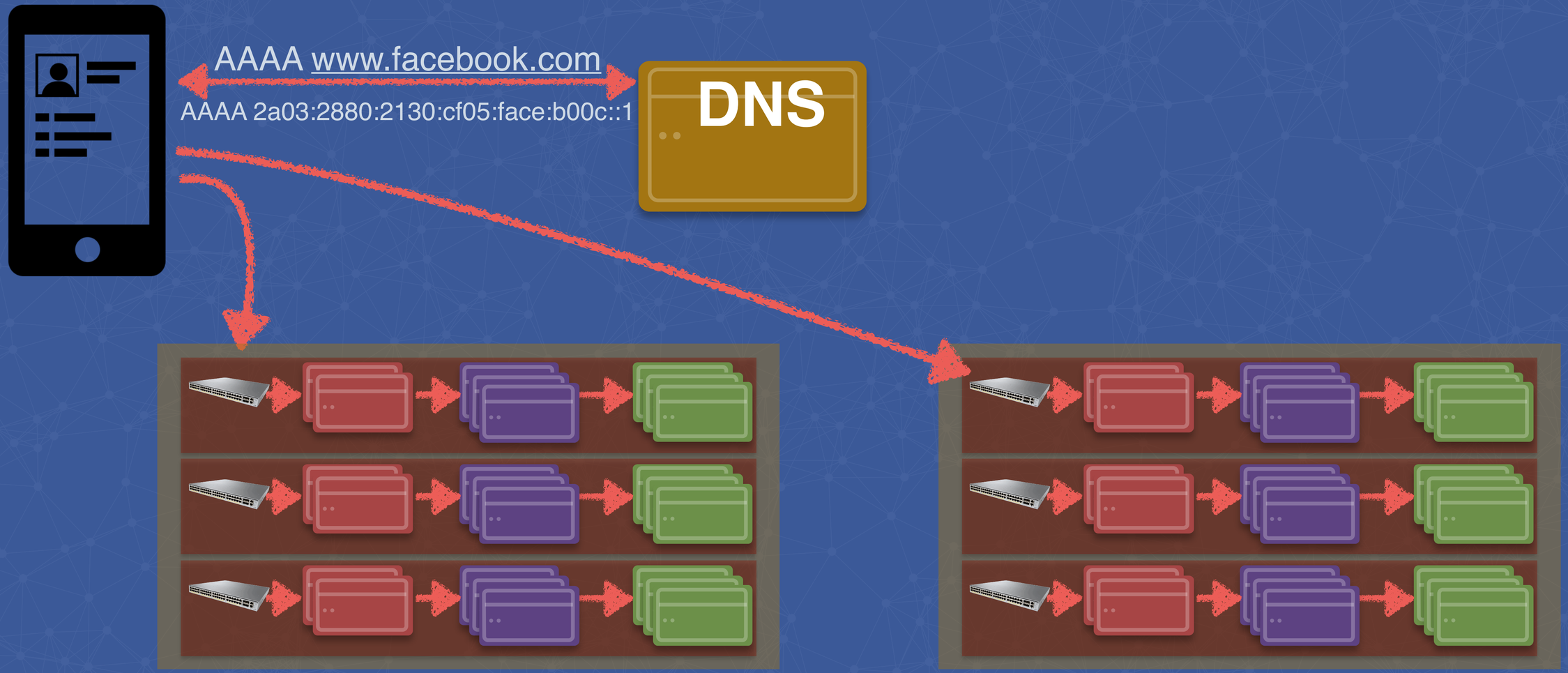
GET /

<html>...

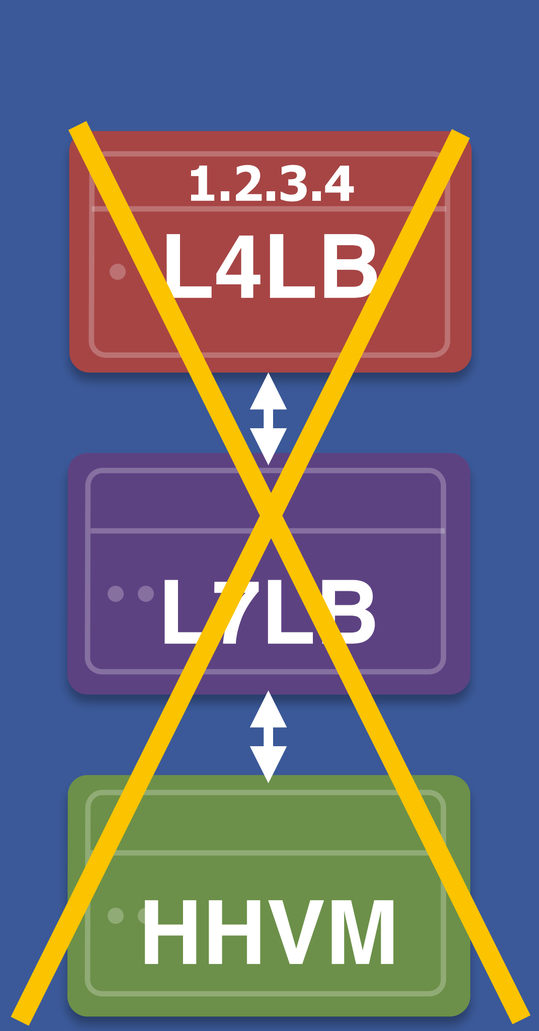
HHVM

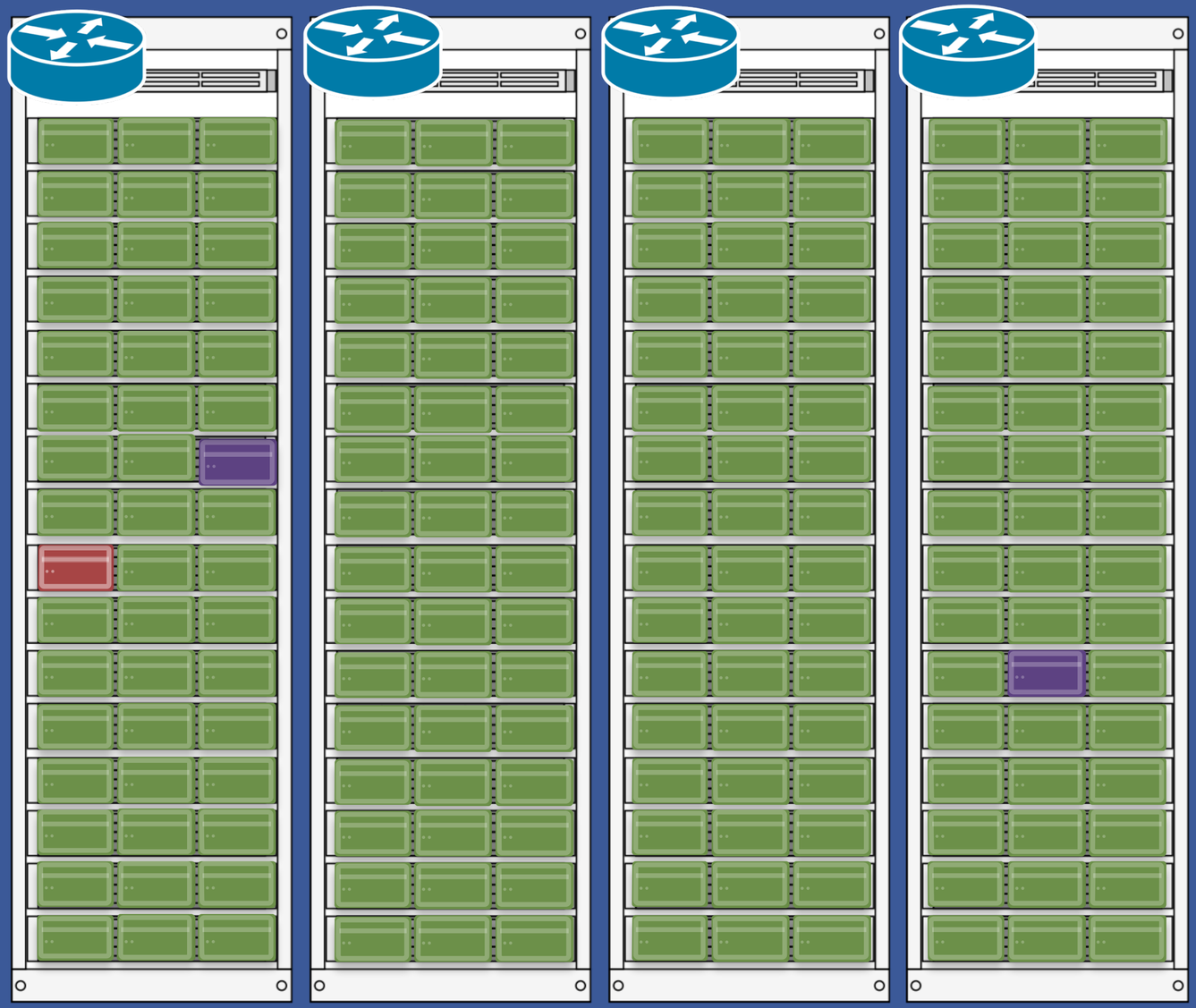
how do we get more rps?!

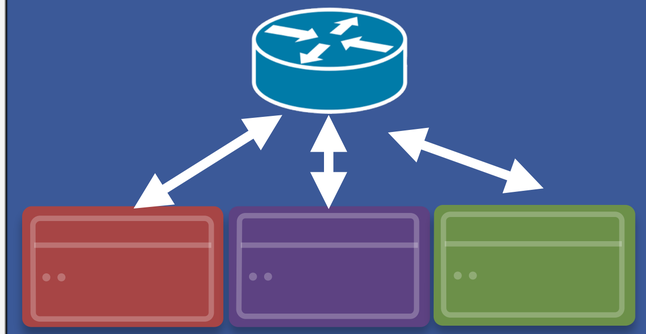
Add another datacenter!



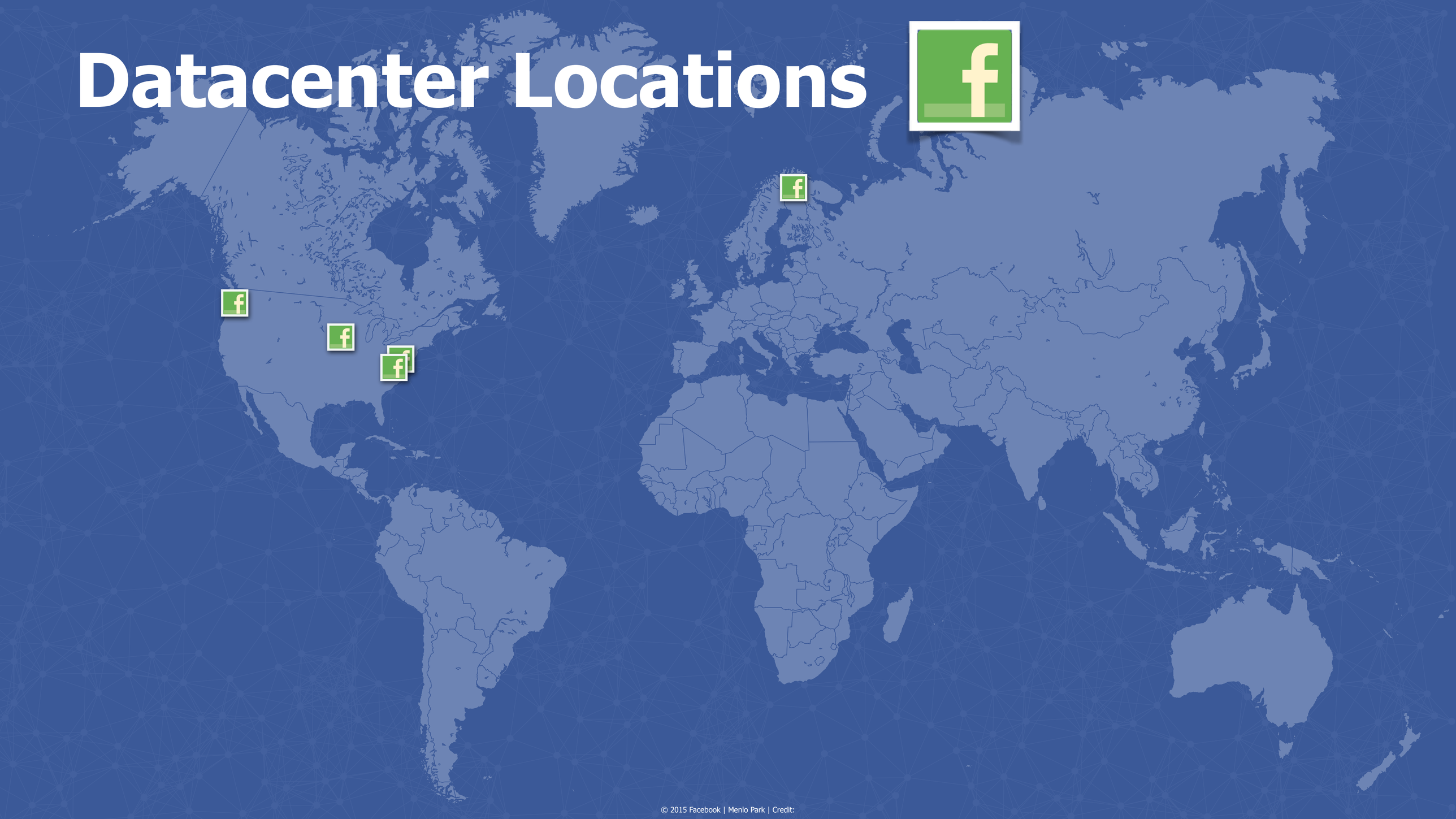
Not really top down







Datacenter Locations



Load Balancing: L4/L7

Let's break it down

www?

A 1.2.3.4

DNS

L4 = IP/port
L7 = HTTP

GET /

<html>...

L4LB
(ipvs)

GET /

<html>...

L7LB
(proxygen)

GET /

<html>...

HHVM

OSI Model: What is L4/L7?

Layer	Purpose	Ex
7: Application	**High-Level API**	**HTTP**, SPDY, MQTT
6: Presentation	Data Translation	ASCII, JPEG
5: Session	Communication Session	RPC
4: Transport	**Transmission**	**TCP**, UDP
3: Network	**Address, Routing, Flow**	**IPv6,** IPv4
2: Data Link	Reliable Physical Comm.	IEEE, 802.2
1: Physical	Raw bit transmission	DSL, USB

L7LB (proxygen)

L4LB (ipvs)

ECMP

L4LB

ECMP Hash

BGP

face:b00c::1

face:b00c::1

L4LB

L4LB

face:b00c::1 (lo)

face:b00c::1 (lo)

face:b00c::1 (lo)

face:b00c::1 (lo)

L7LB

L7LB

L7LB

L7LB

NOTE: L7 (proxygen) Listens to the VIP on loopback (lo) interface. Not eth0.

L4LB

BGP

ECMP Hash

face:b00c::1

L4LB

face:b00c::1

L4LB

State Table + Hash

face:b00c::1 (lo)

L7LB

face:b00c::1 (lo)

L7LB

face:b00c::1 (lo)

L7LB

face:b00c::1 (lo)

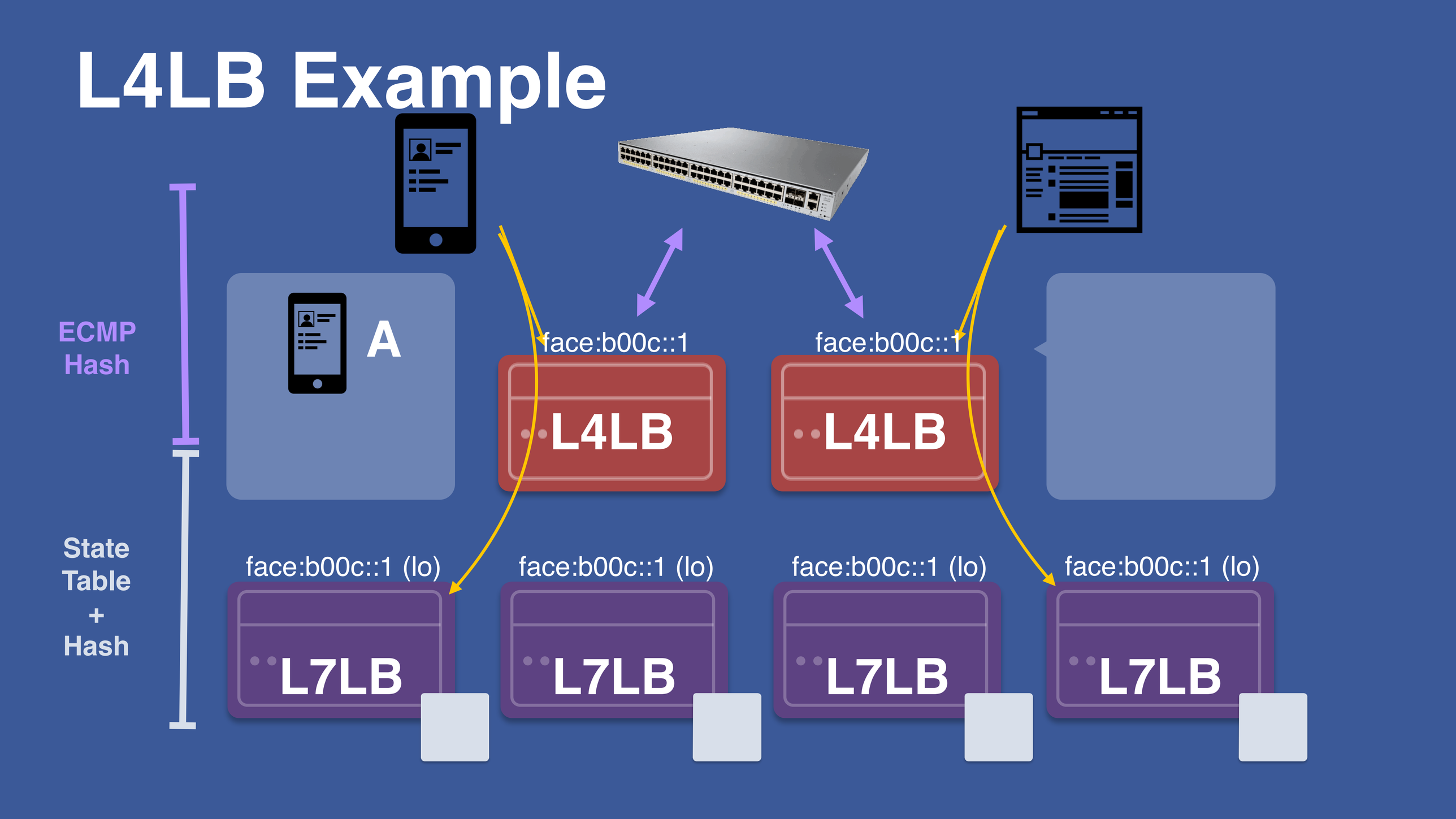
L7LB

Service Discovery

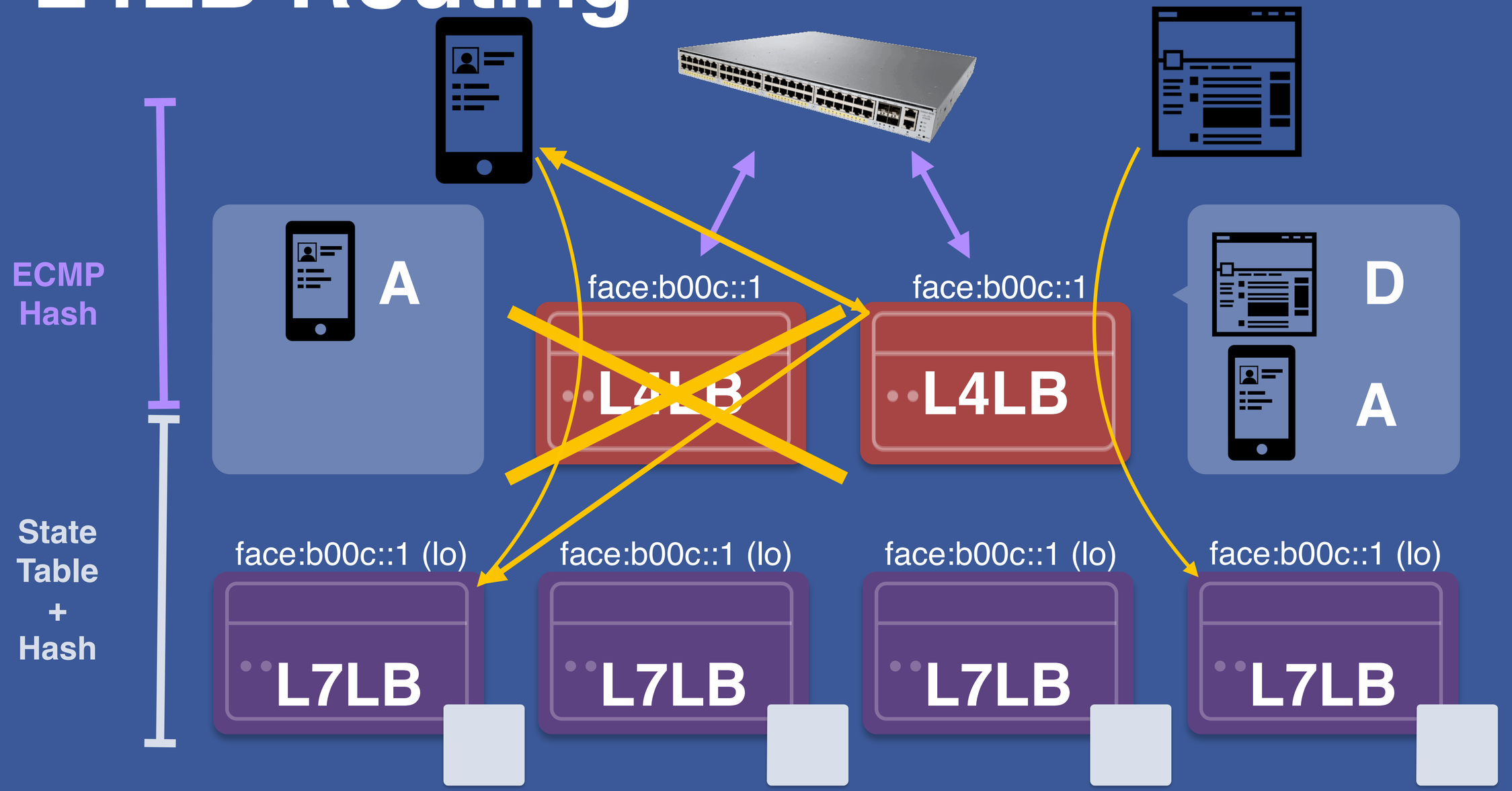
(Zookeeper)

NOTE: L7 (proxygen) Listens to the VIP on loopback (lo) interface. Not eth0.

L4LB Example



L4LB Routing



L4LB Routing

ECMP Hash

State Table + Hash

B

face:b00c::1

L4LB

face:b00c::1

L4LB

face:b00c::1 (lo)

face:b00c::1 (lo)

face:b00c::1 (lo)

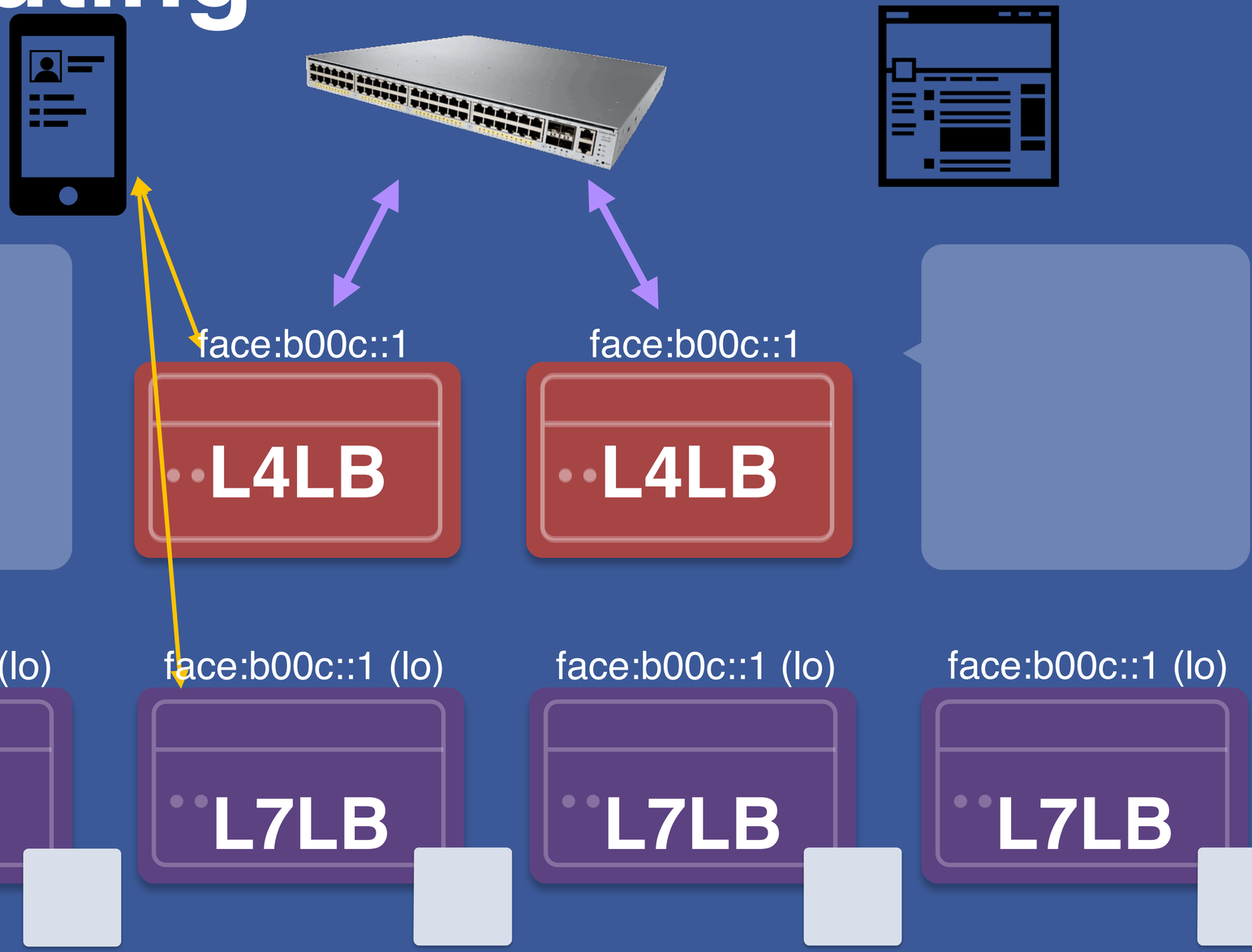
face:b00c::1 (lo)

L7LB

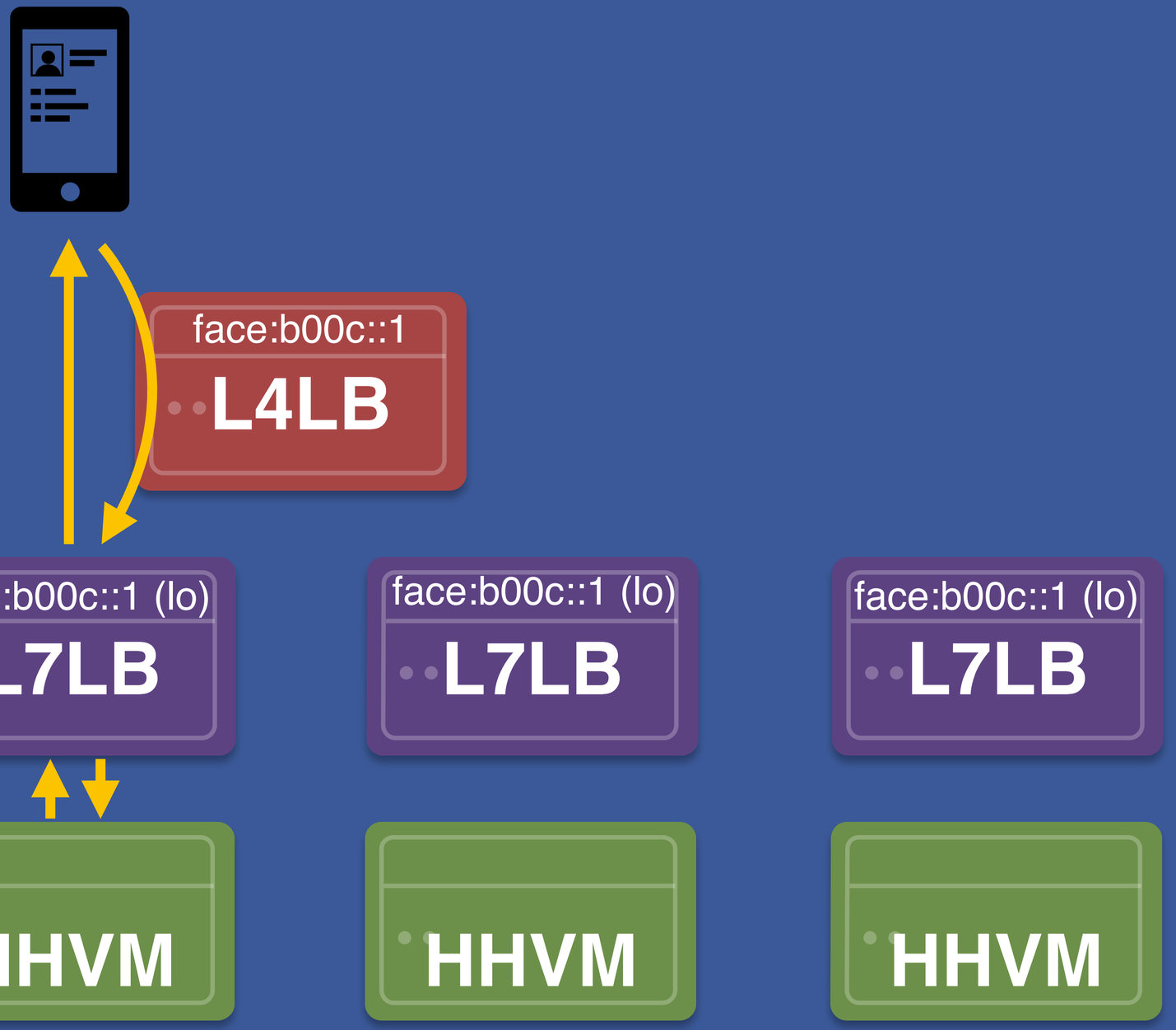
L7LB

L7LB

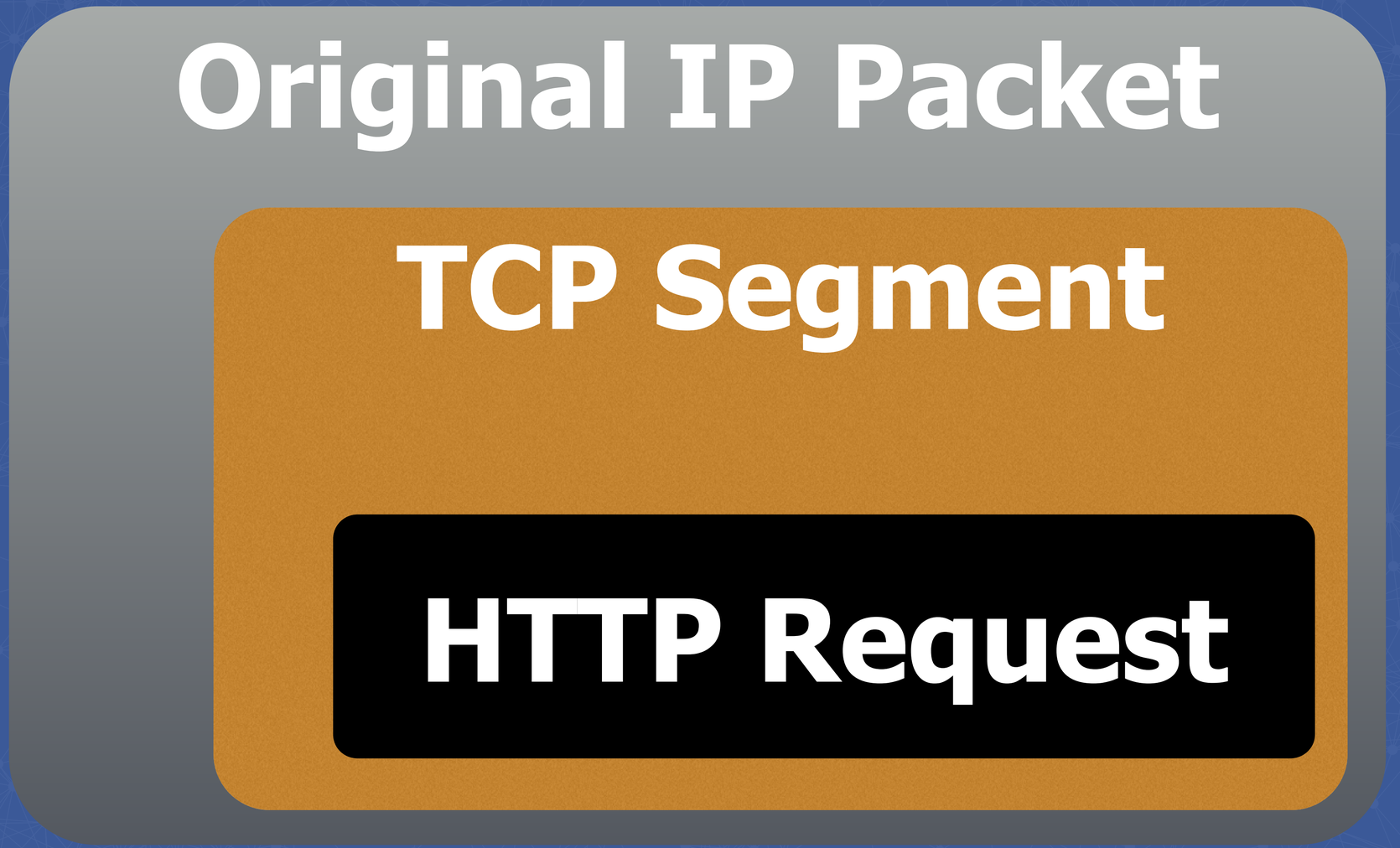
L4LB Routing



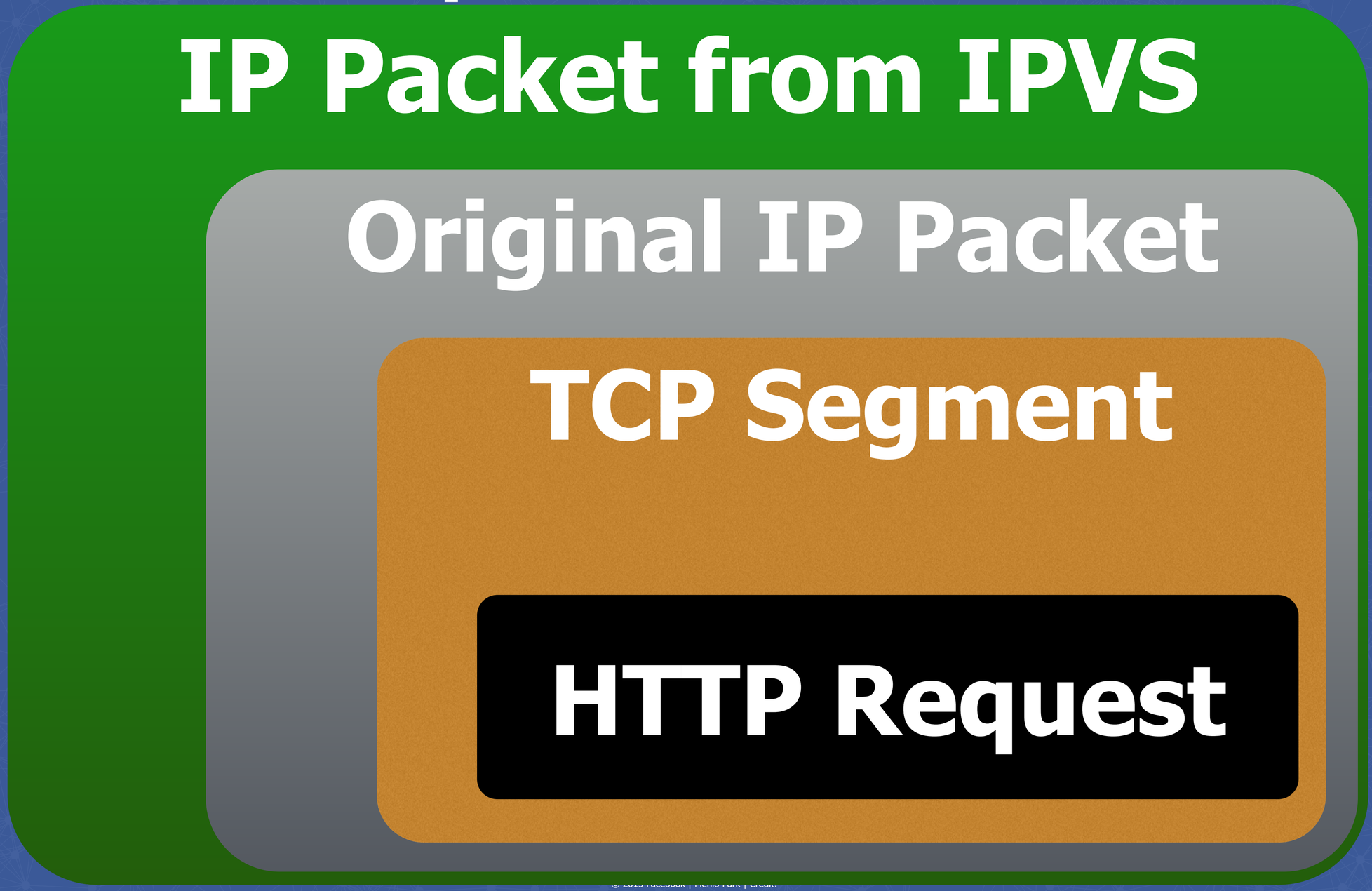
Direct Server Return



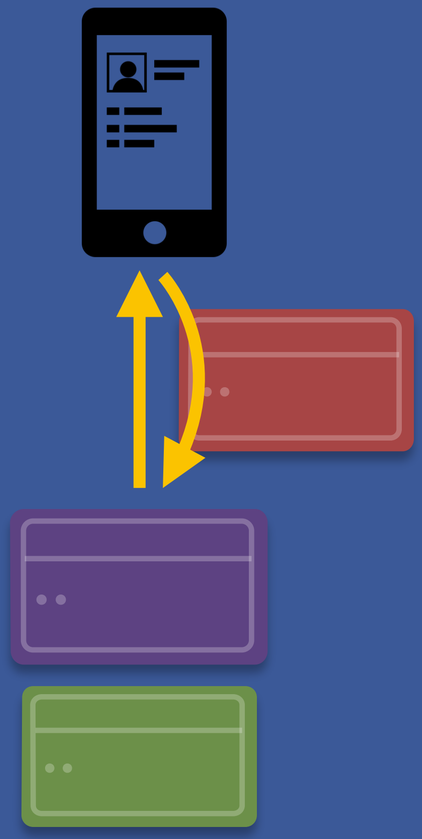
Remember this?

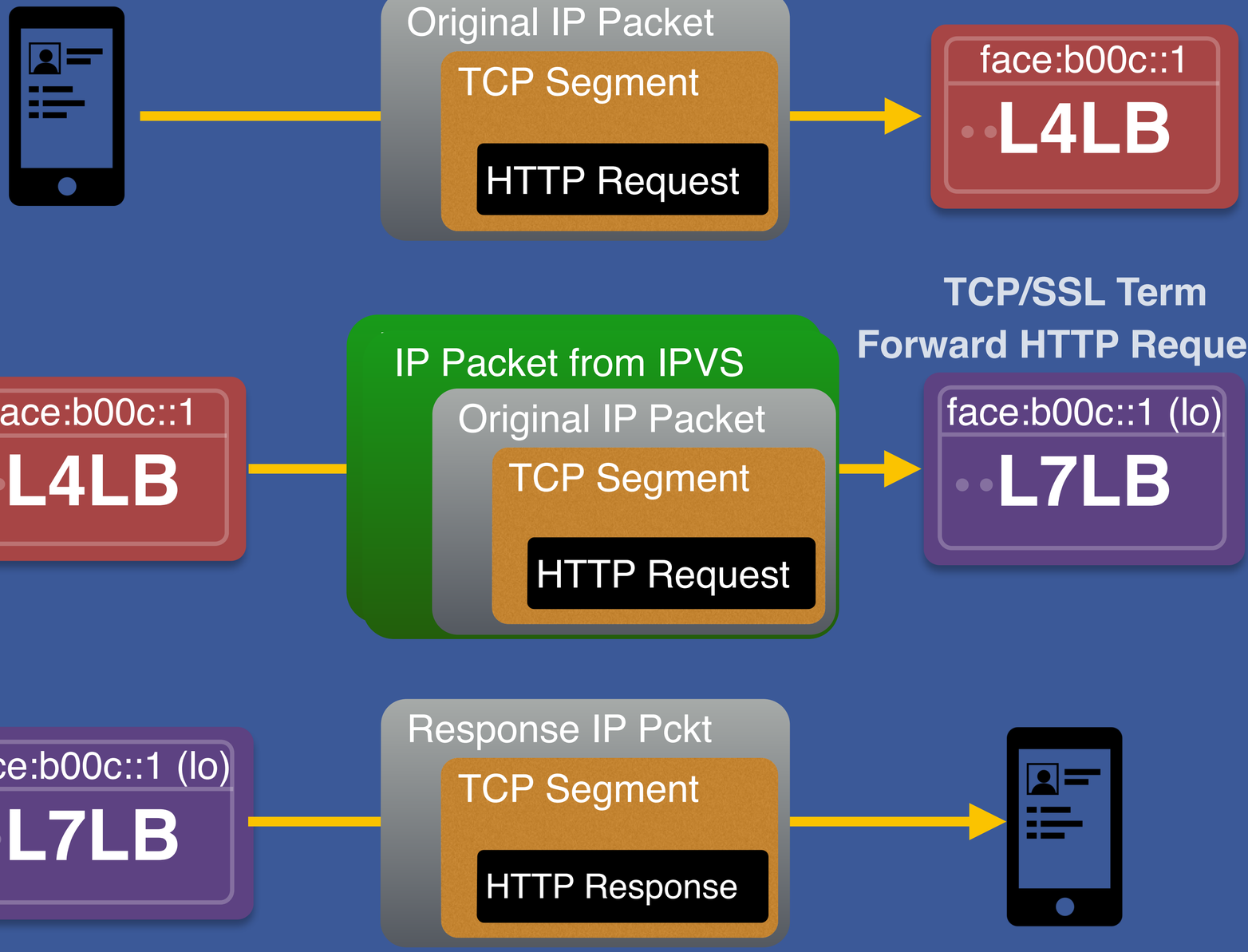


IP in IP encapsulation



Direct Server Return

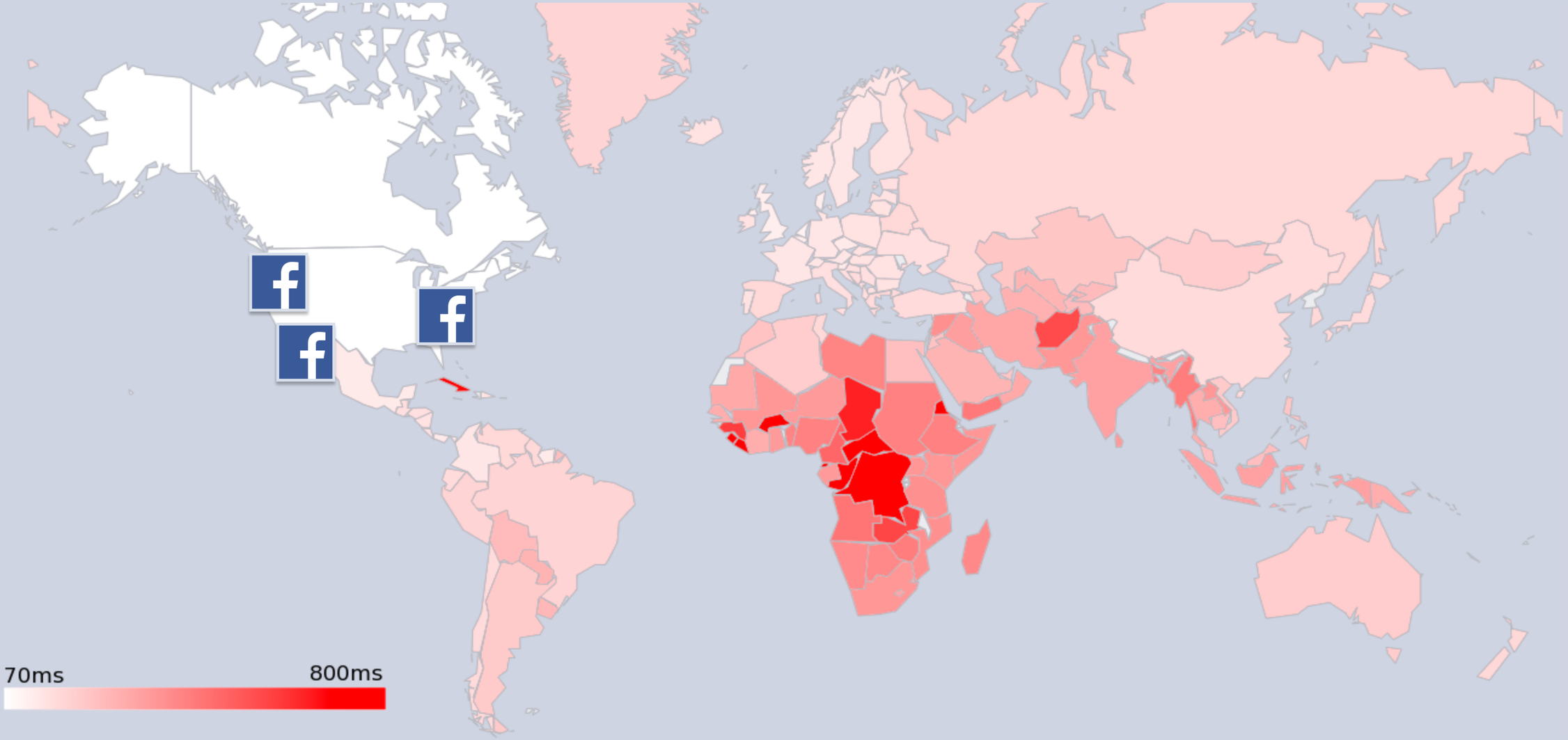




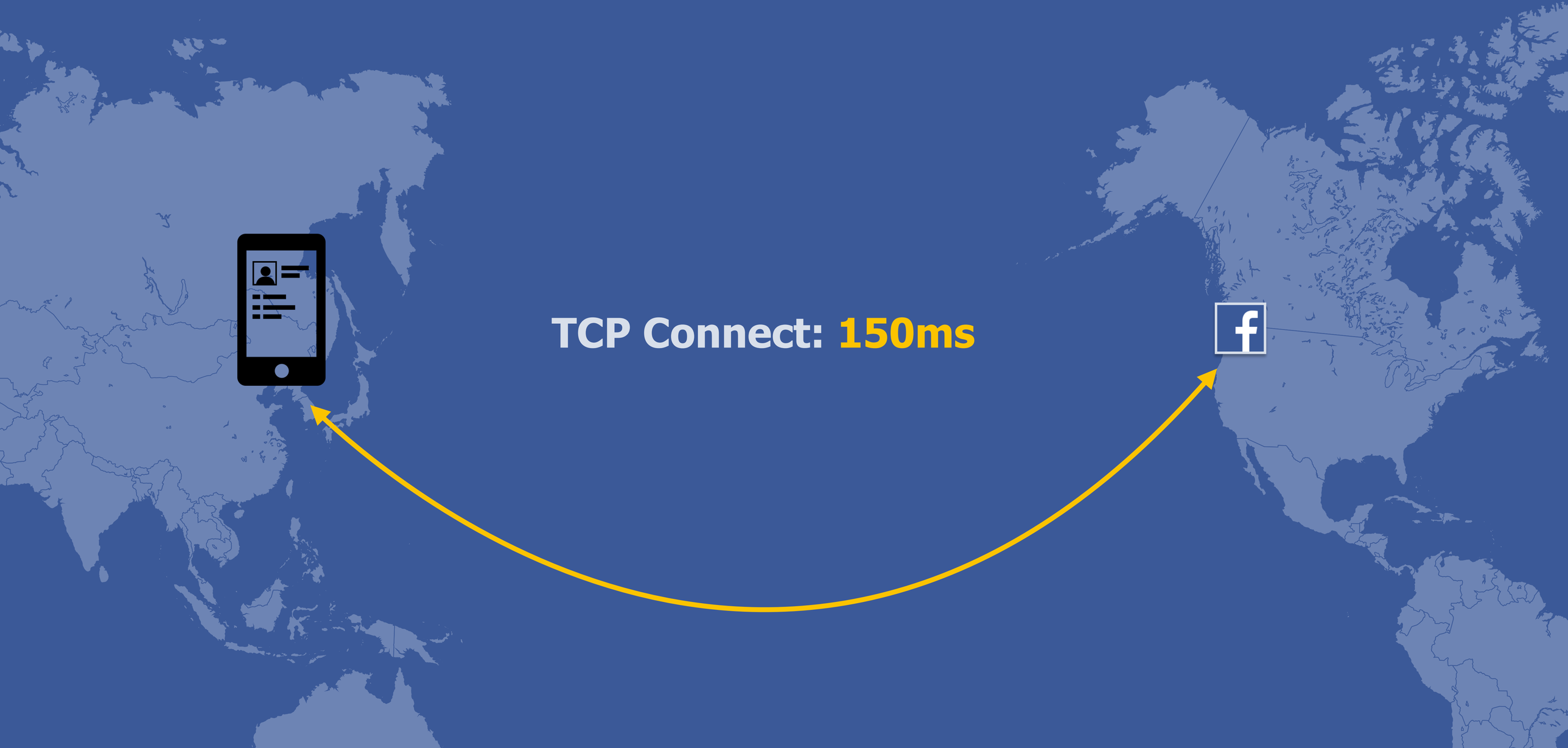
Edge PoP’s & Reducing Latency

International RTT

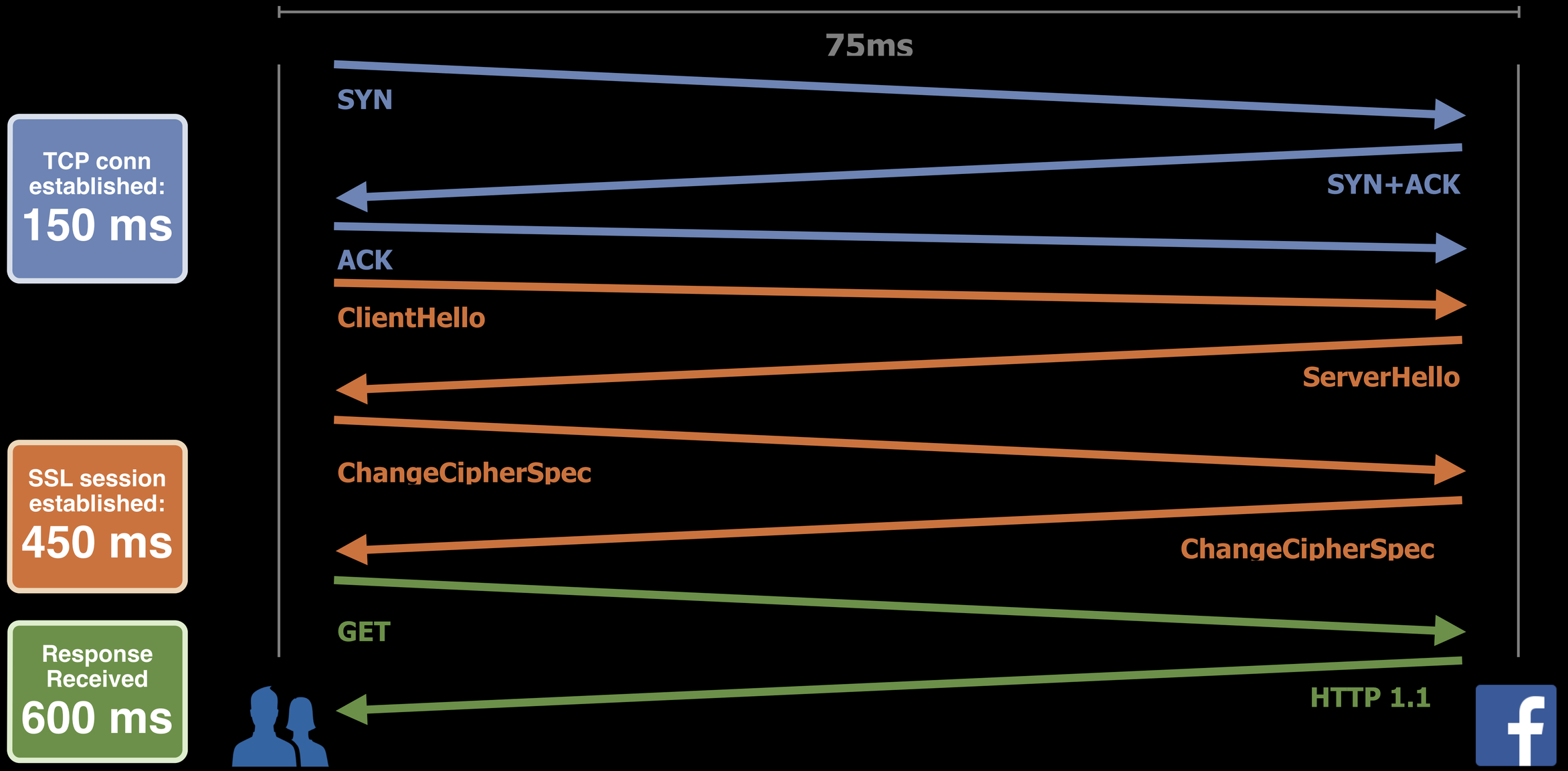
circa 11/2011



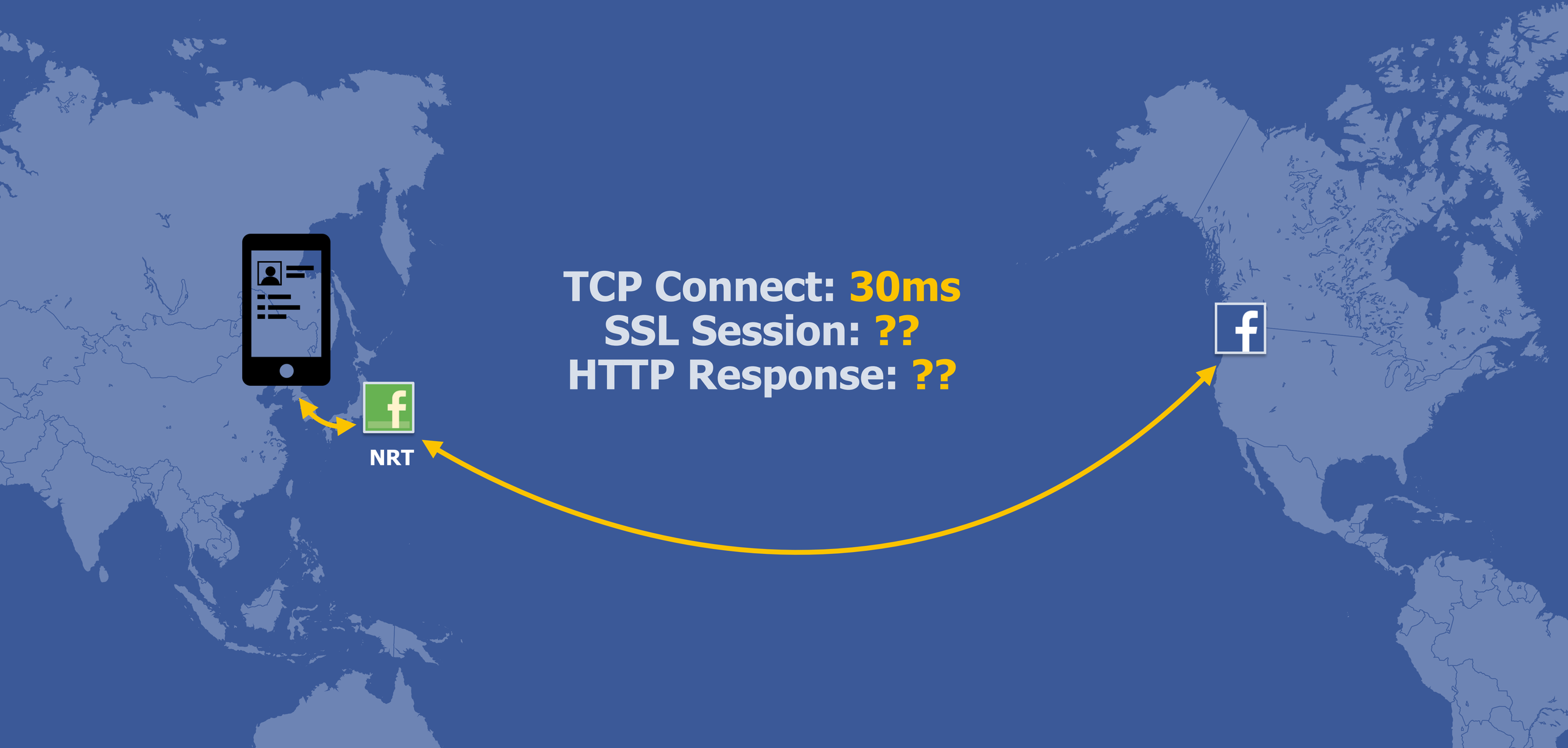
Seoul -> Oregon



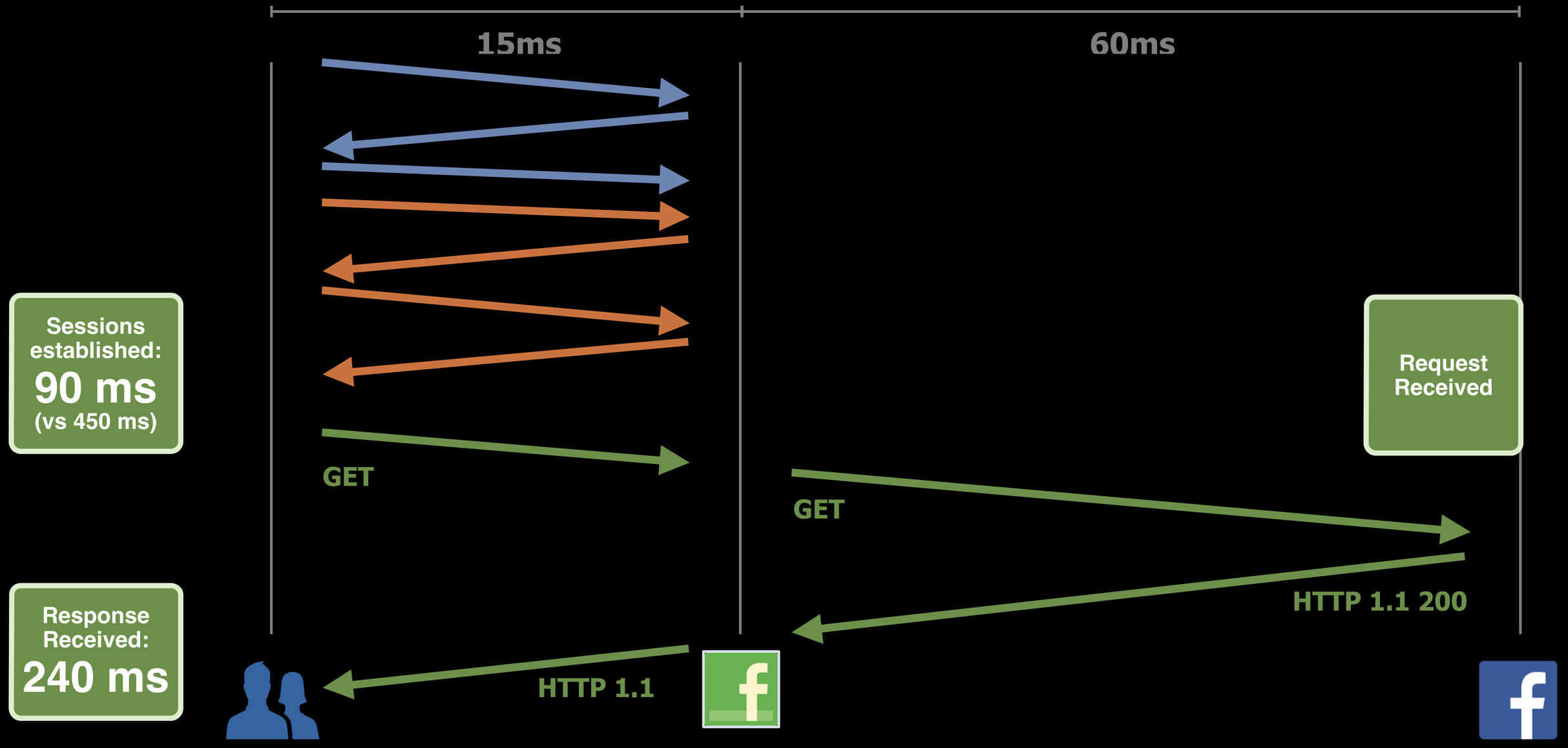
HTTPS Seoul -> Oregon



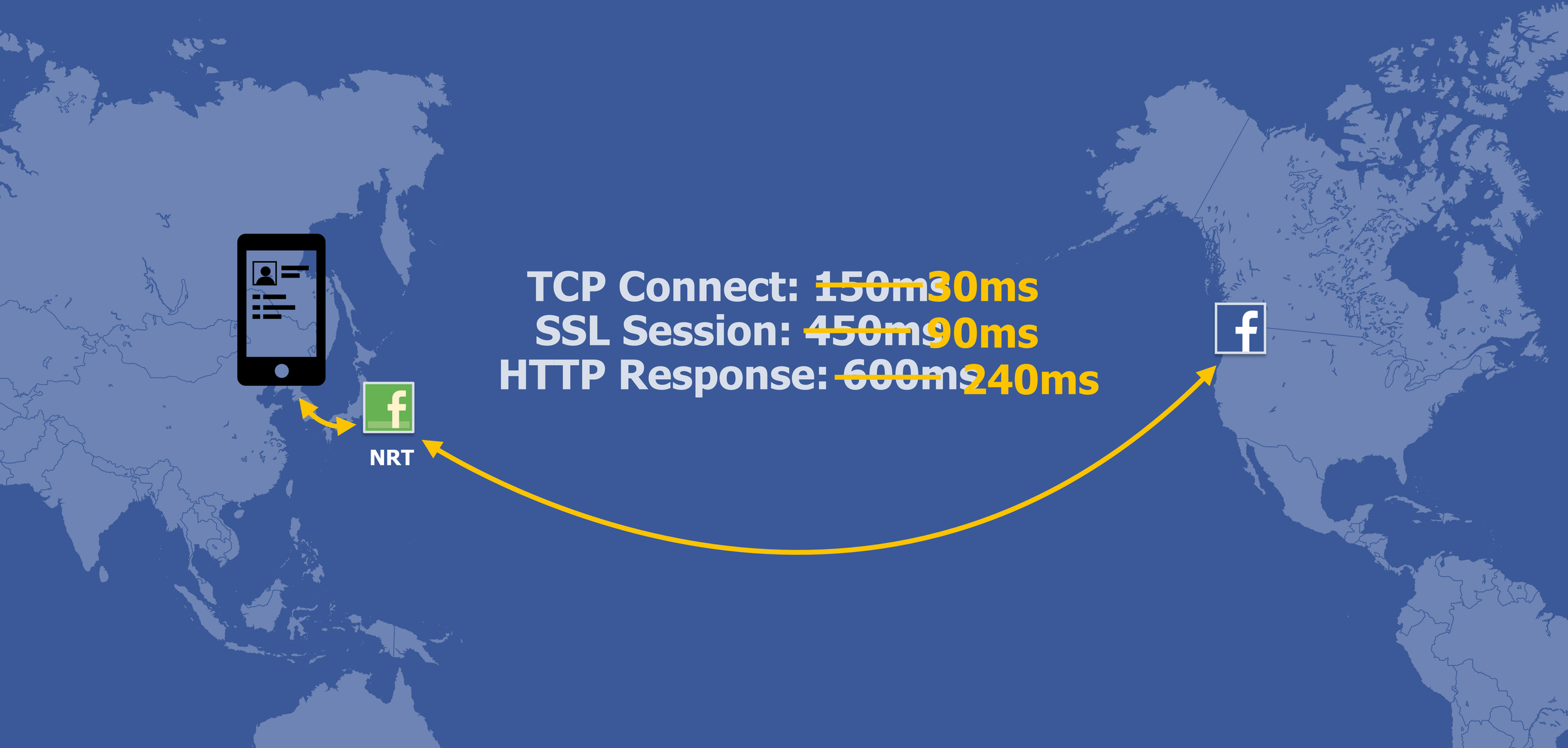
Seoul -> Tokyo -> Oregon



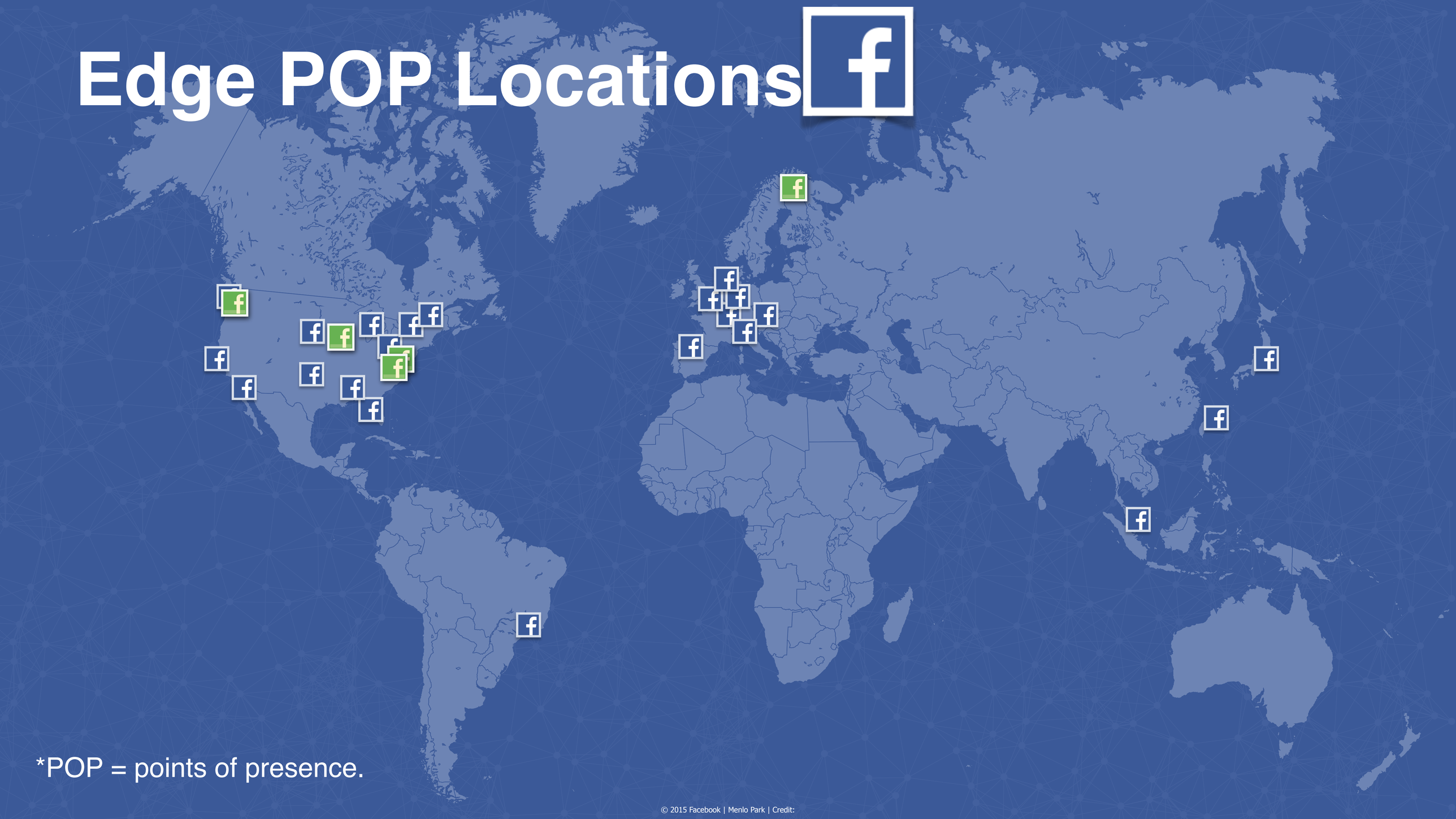
HTTPS Seoul->Tokyo->Oregon



Seoul -> Oregon

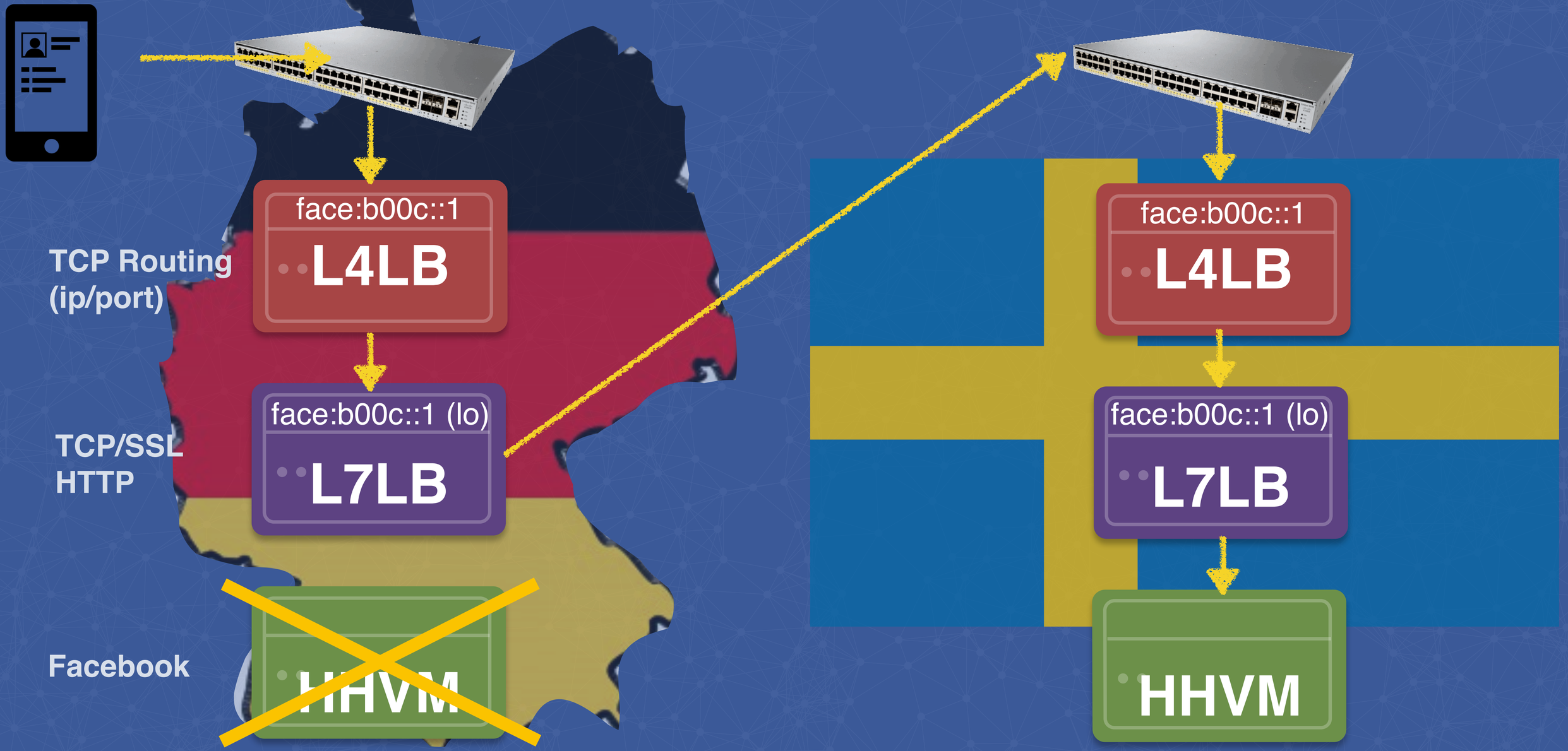


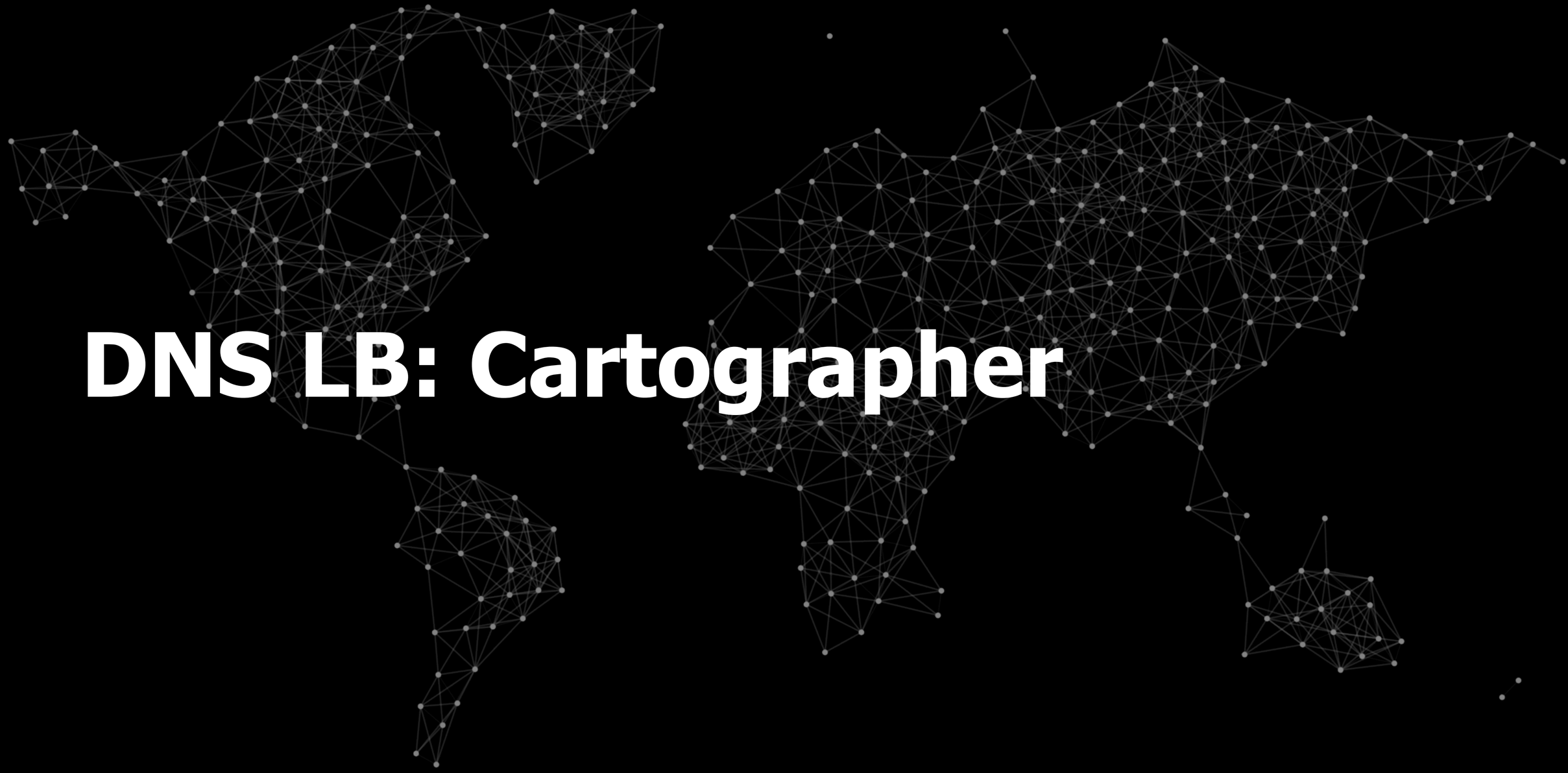
Edge POP Locations



*POP = points of presence.

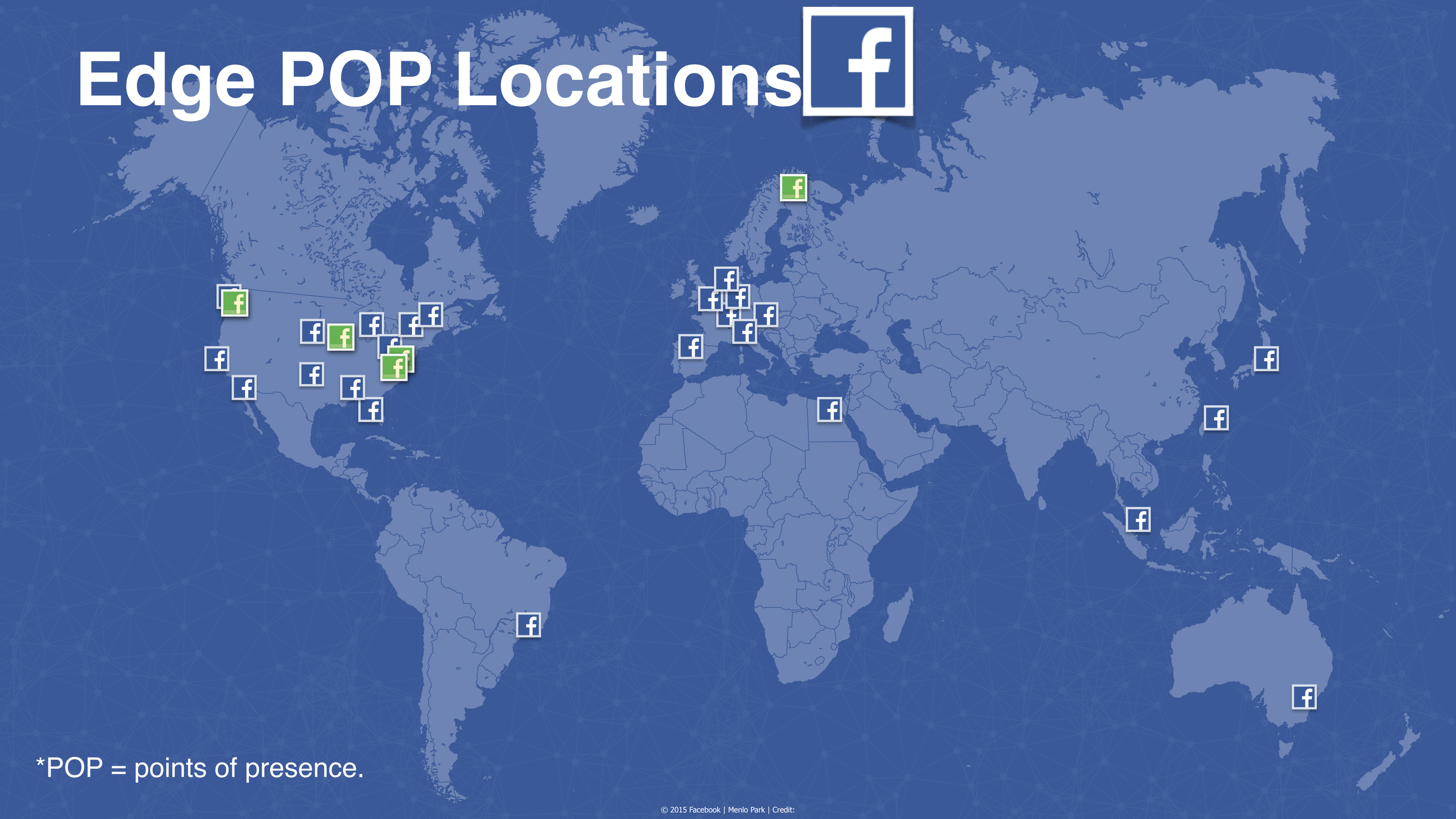
How do the LB's in PoP's work?





DNS LB: Cartographer

Edge POP Locations



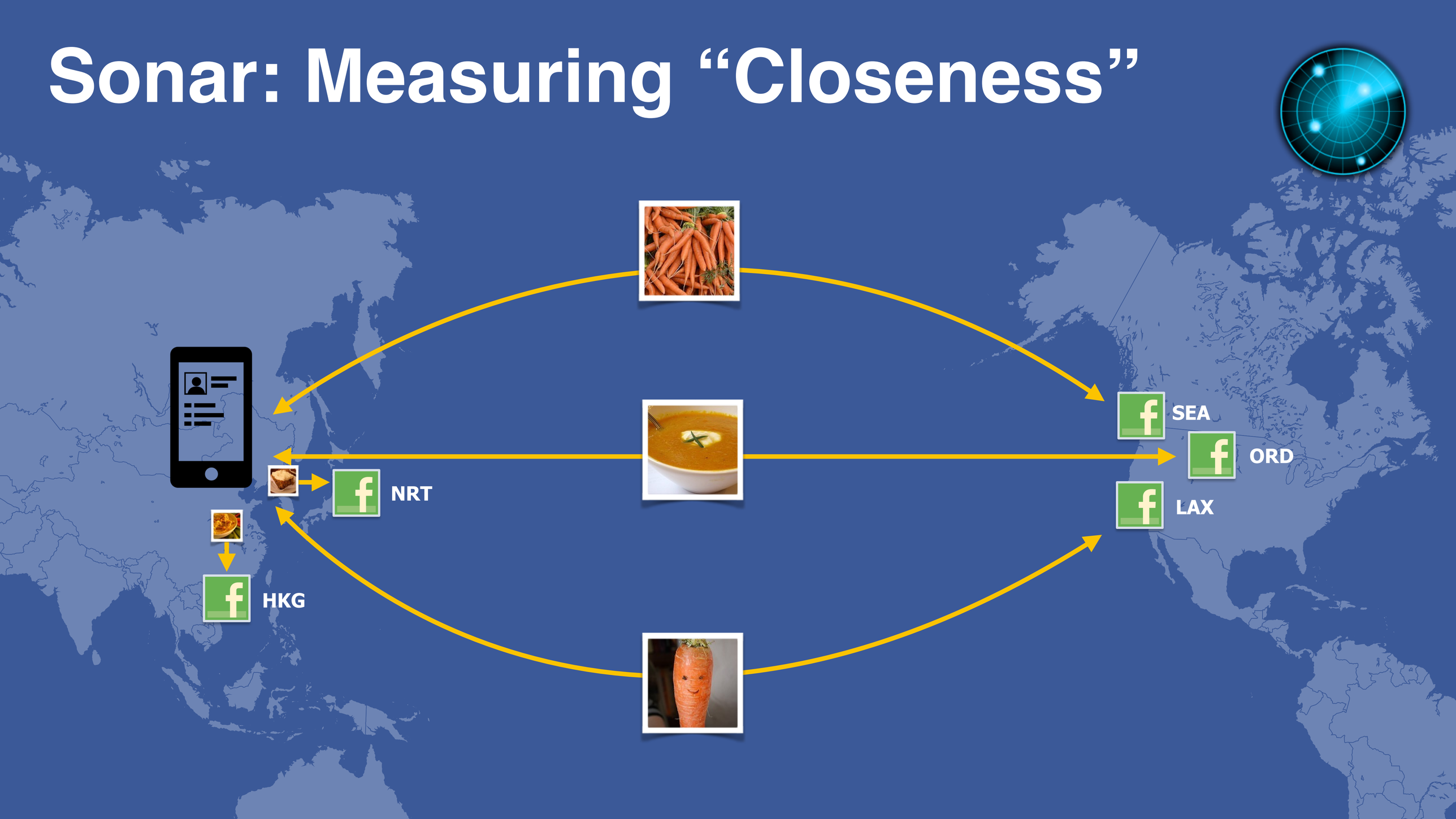
*POP = points of presence.

DNS LB Decision

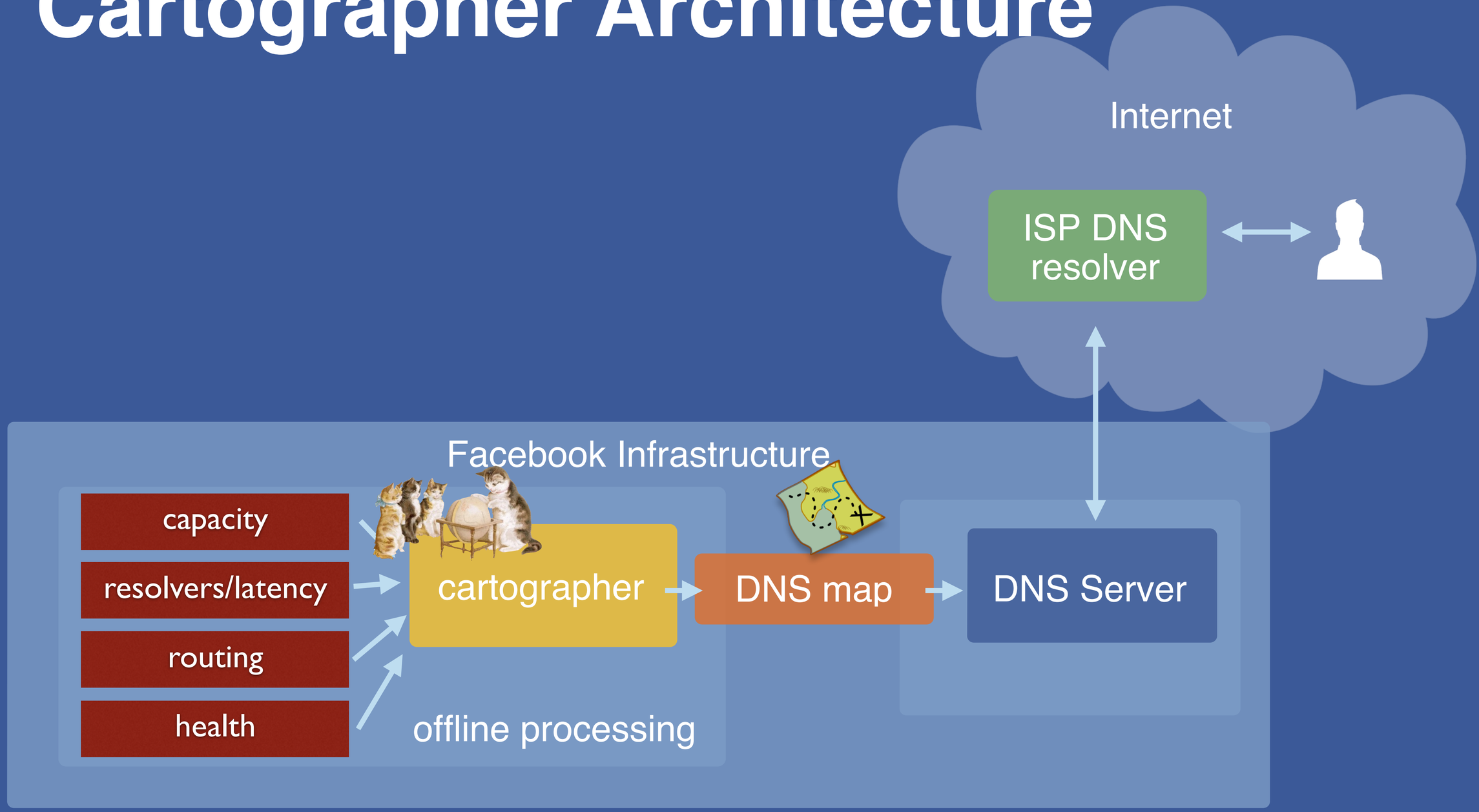
Considerations:

- **Closest Edge** to user**
- **Capacity**
- **Health**
- **Geo data**

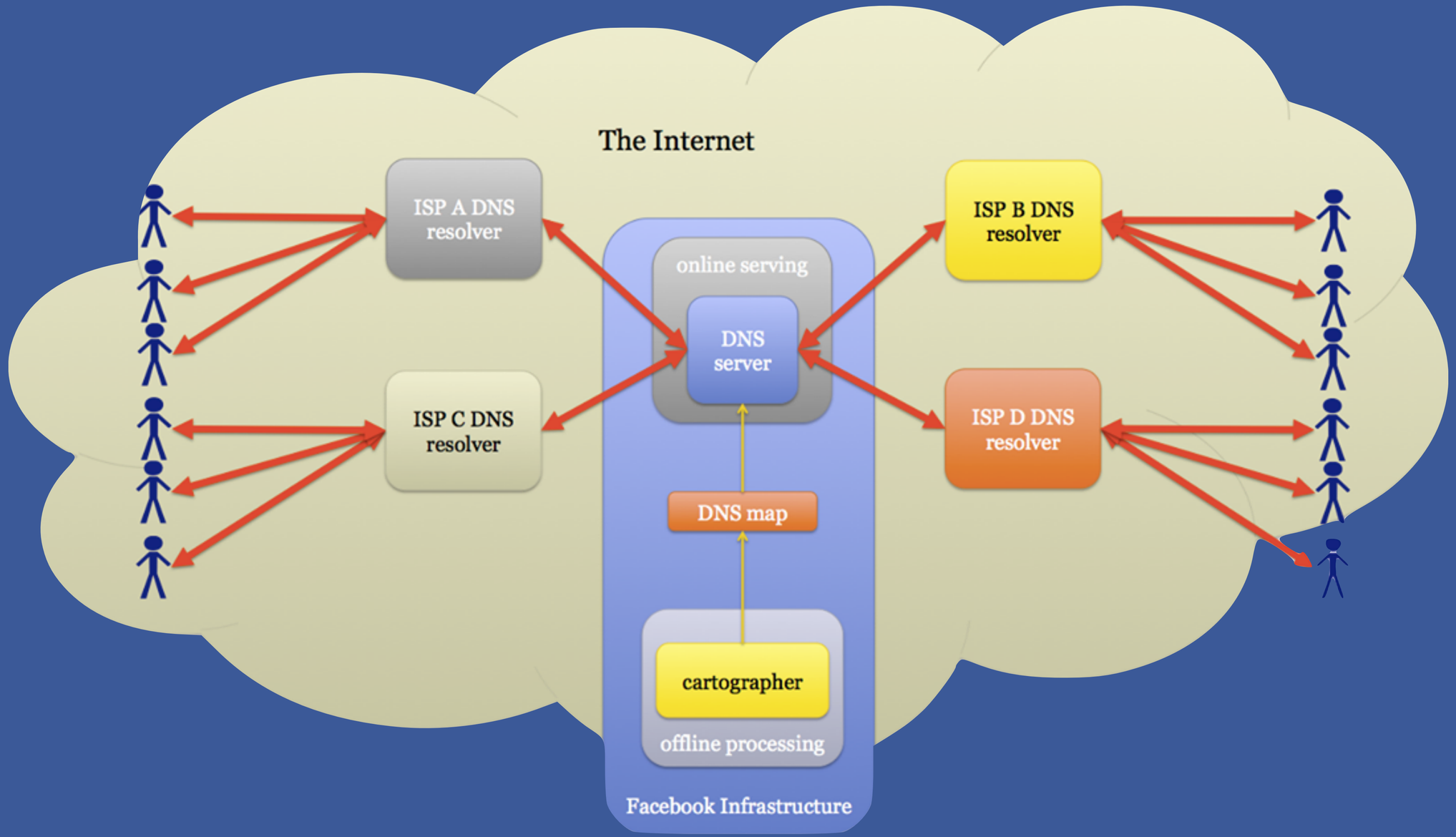
Sonar: Measuring “Closeness”



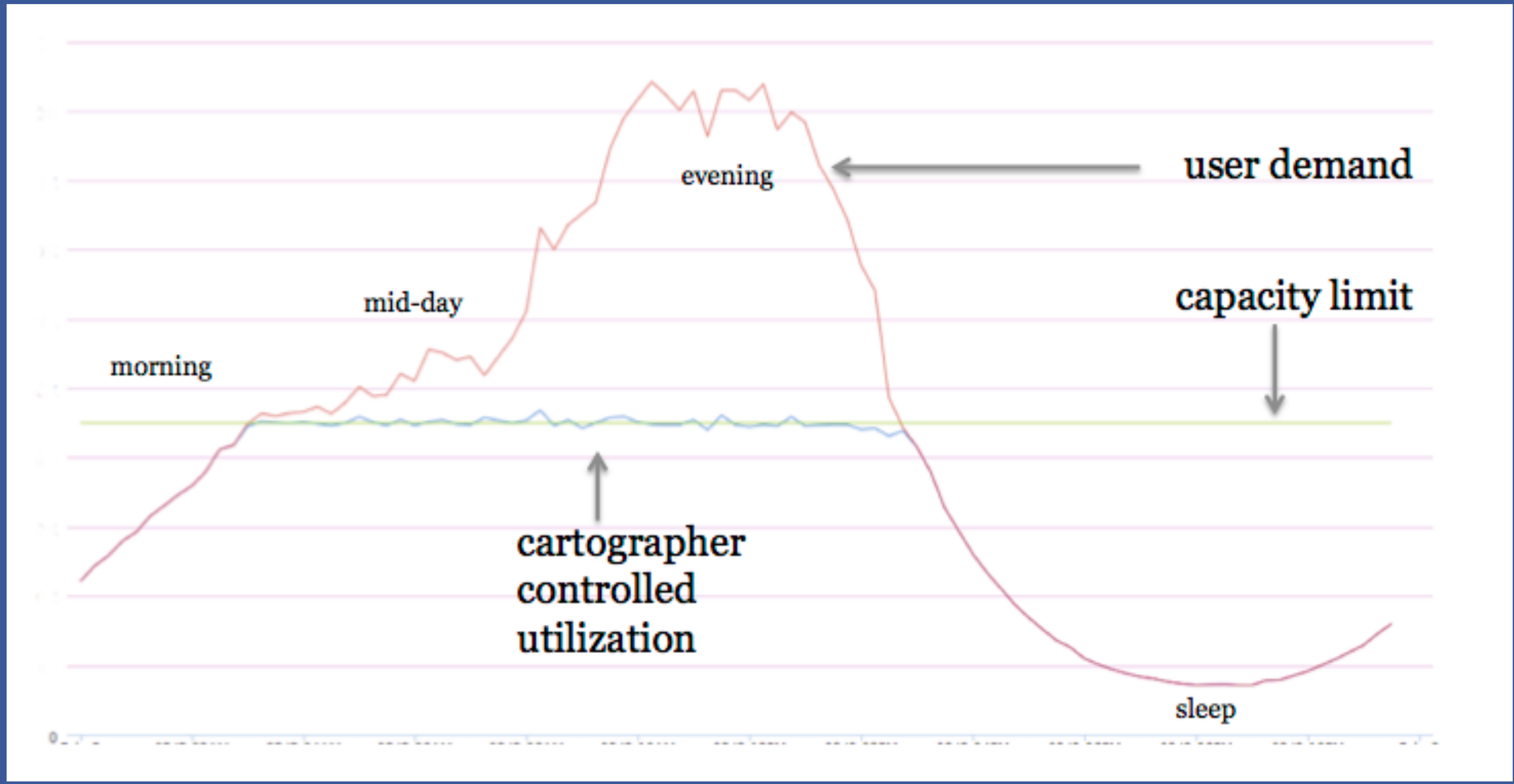
Cartographer Architecture



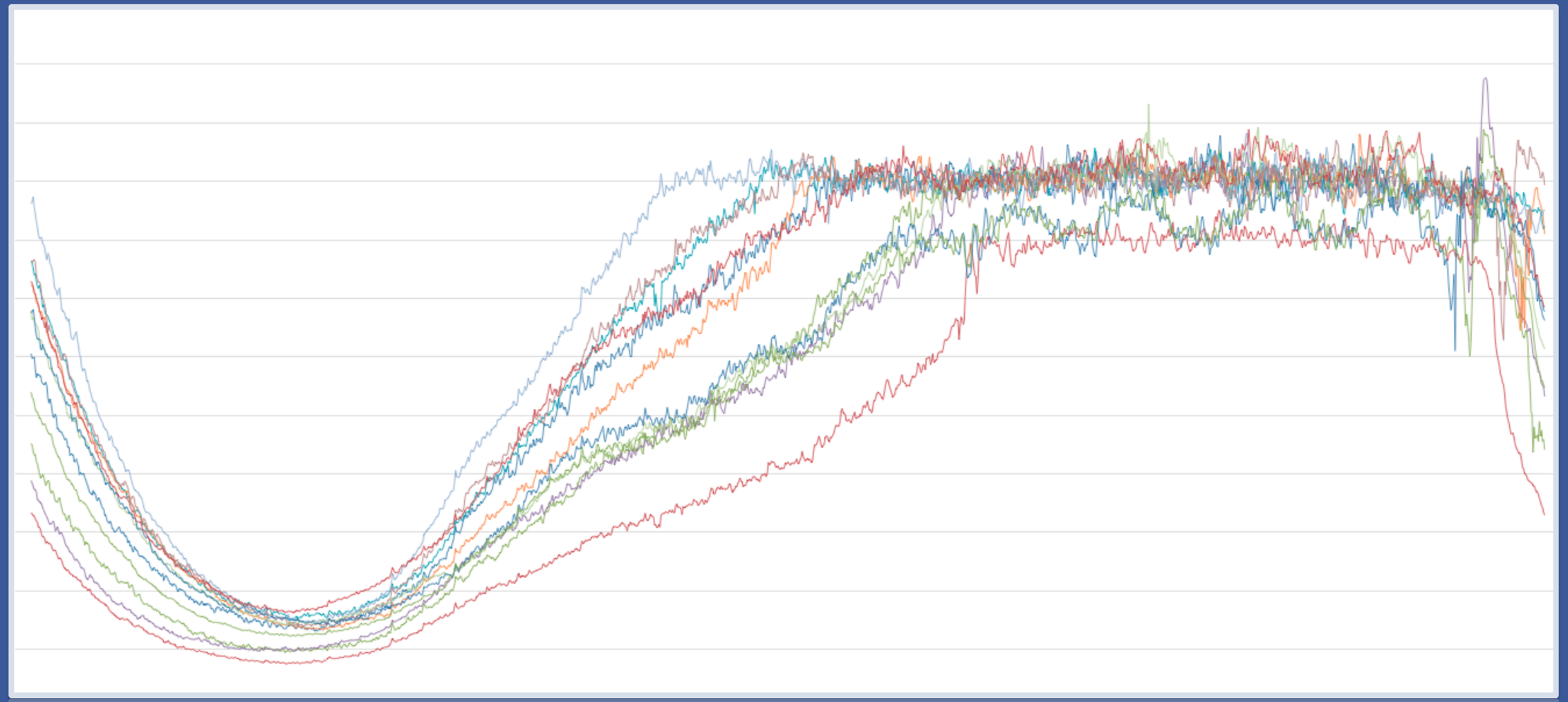
Cartographer Architecture



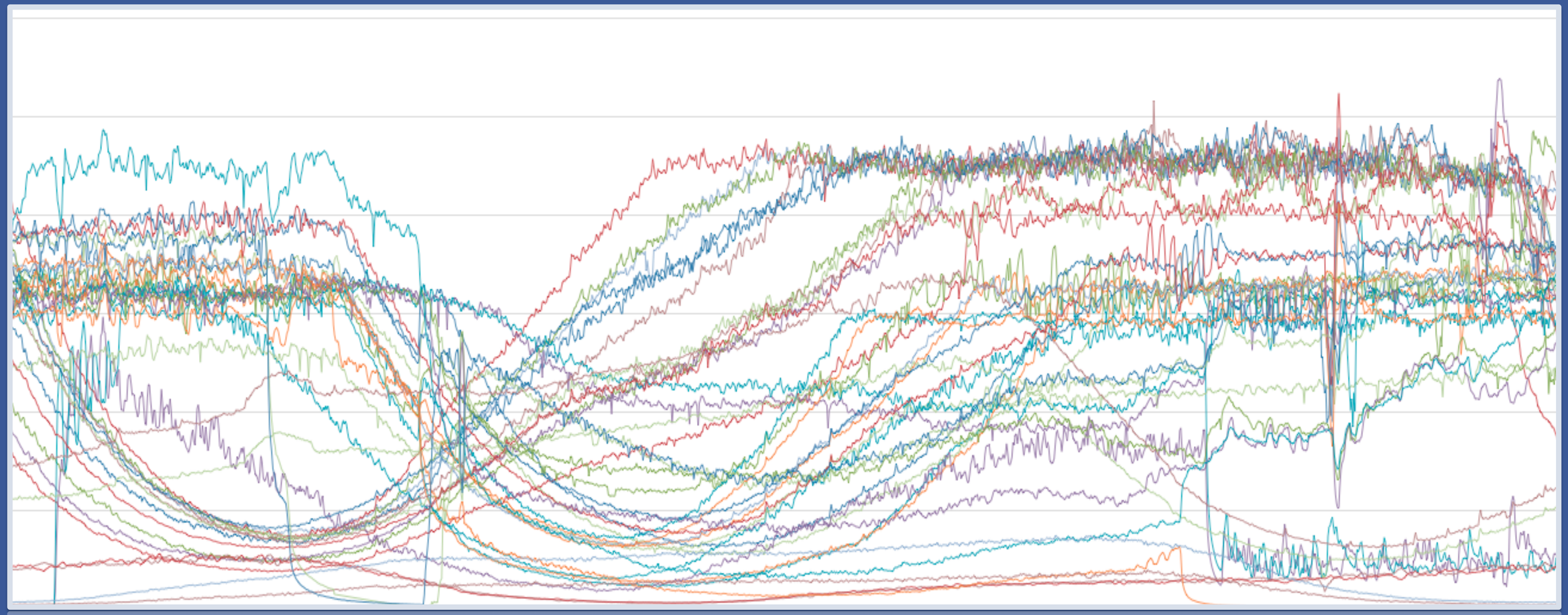
Cartographer in action



Regional Load Shedding

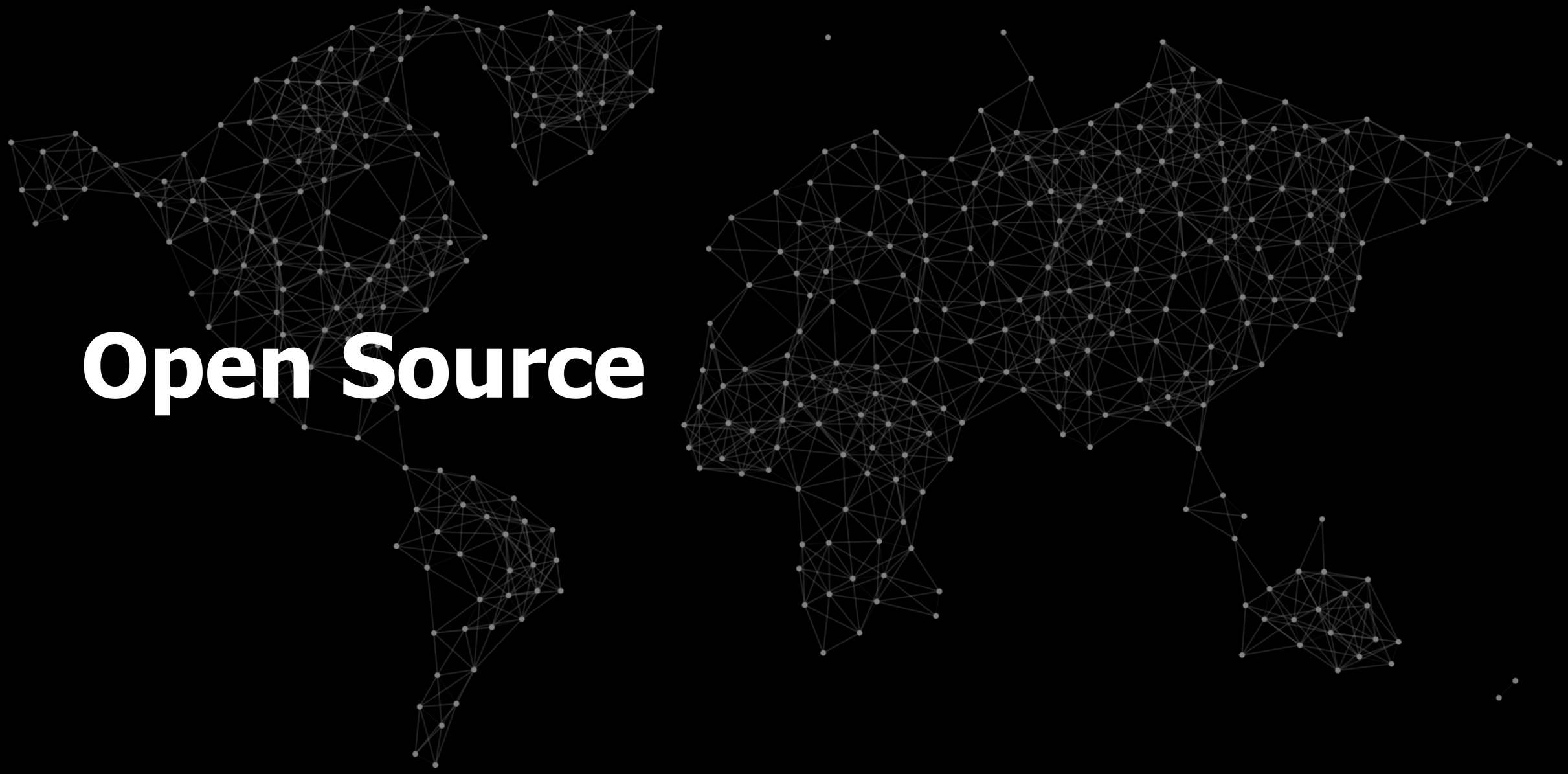


Global Load Shedding



Monitoring

~~Monitoring (Not enough time)~~



Open Source

Open Source Components

- Proxygen Libs

https://github.com/facebook/proxygen

- HHVM

https://hhvm.com

- TinyDNS

https://cr.yp.to/djbdns/tinydns.html

- IPVS (IP Virtual Server)

http://www.linuxvirtualserver.org/software/ipvs.html

- ExaBGP

https://github.com/Exa-Networks/exabgp

- Python

https://python.org

- Zookeeper

https://zookeeper.apache.org

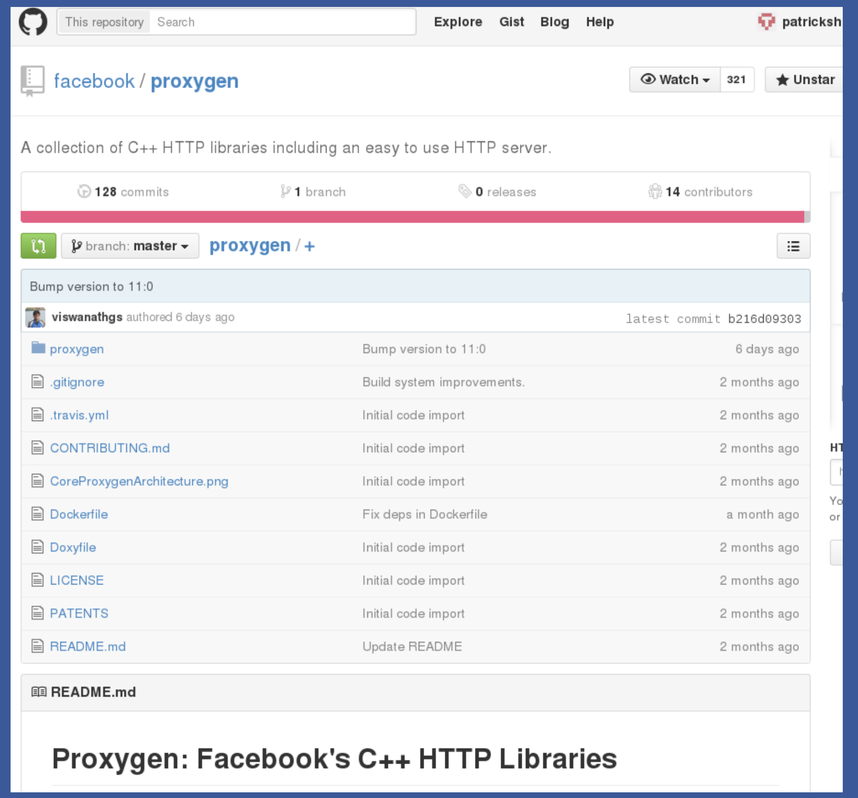


Photo Credits

http://www.flickr.com/photos/27587002@N07/5170590074

http://www.flickr.com/photos/yaketyyakyak/7001664846

http://www.flickr.com/photos/hinnosaar/3778903507

http://www.flickr.com/photos/eamoncurry/8698726494

http://www.flickr.com/photos/43158397@N02/4514113429

http://www.flickr.com/photos/nobusue/6876280595

http://www.flickr.com/photos/29487672@N07/14760573314

http://www.flickr.com/photos/joyosity/3595242078

http://www.flickr.com/photos/kyntharyn74/3262089319

http://www.flickr.com/photos/rexipe/826987087

http://www.flickr.com/photos/lablasco/6815671096

https://thenounproject.com/term/iphone-profile/54906/

https://thenounproject.com/term/browser/59091/


facebook

facebook