Dive deep on AWS edge networking infrastructure

Lincoln Dale
Senior Principal Engineer
AWS – AS16509

Fredrik Korsbäck
Senior Infrastructure Business Developer
AWS – AS16509

aws

Agenda

AWS Global Infrastructure

Our journey to reinventing our network infrastructure

- our hardware, software and how we put systems together

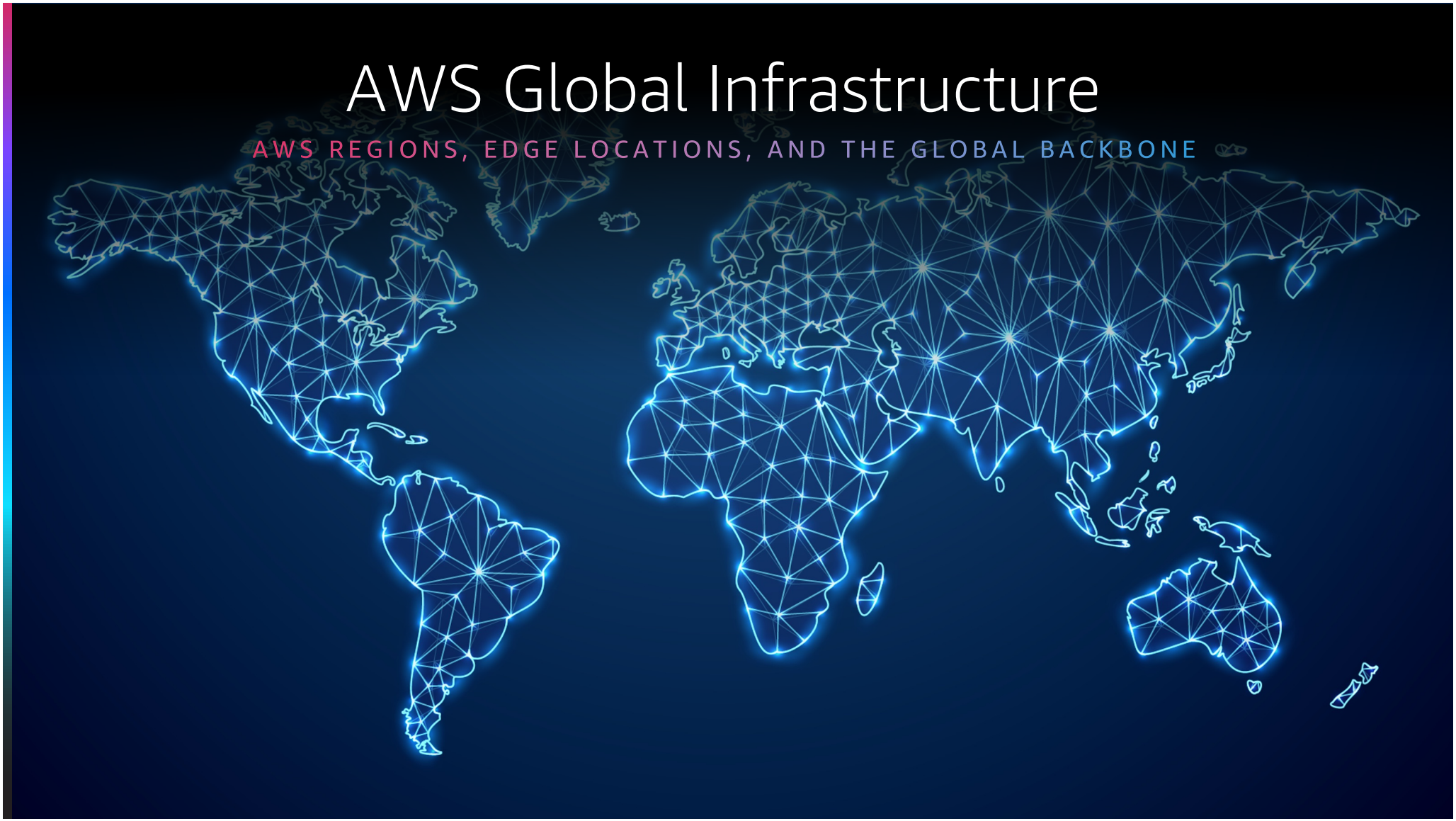
Network architecture and software, tools and controllers

How we build and automate our network, and how it’s going

AWS Global Infrastructure

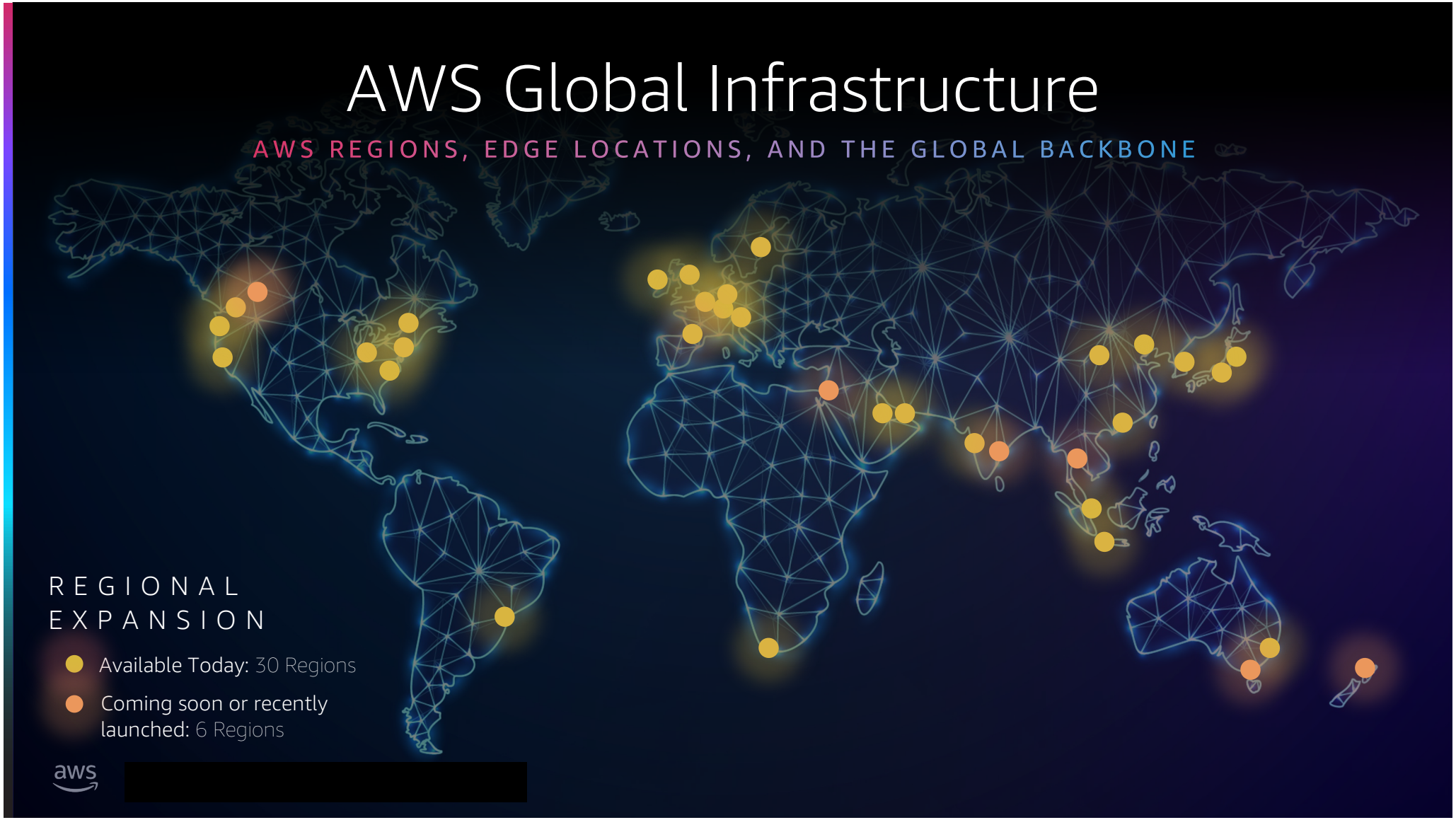
AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



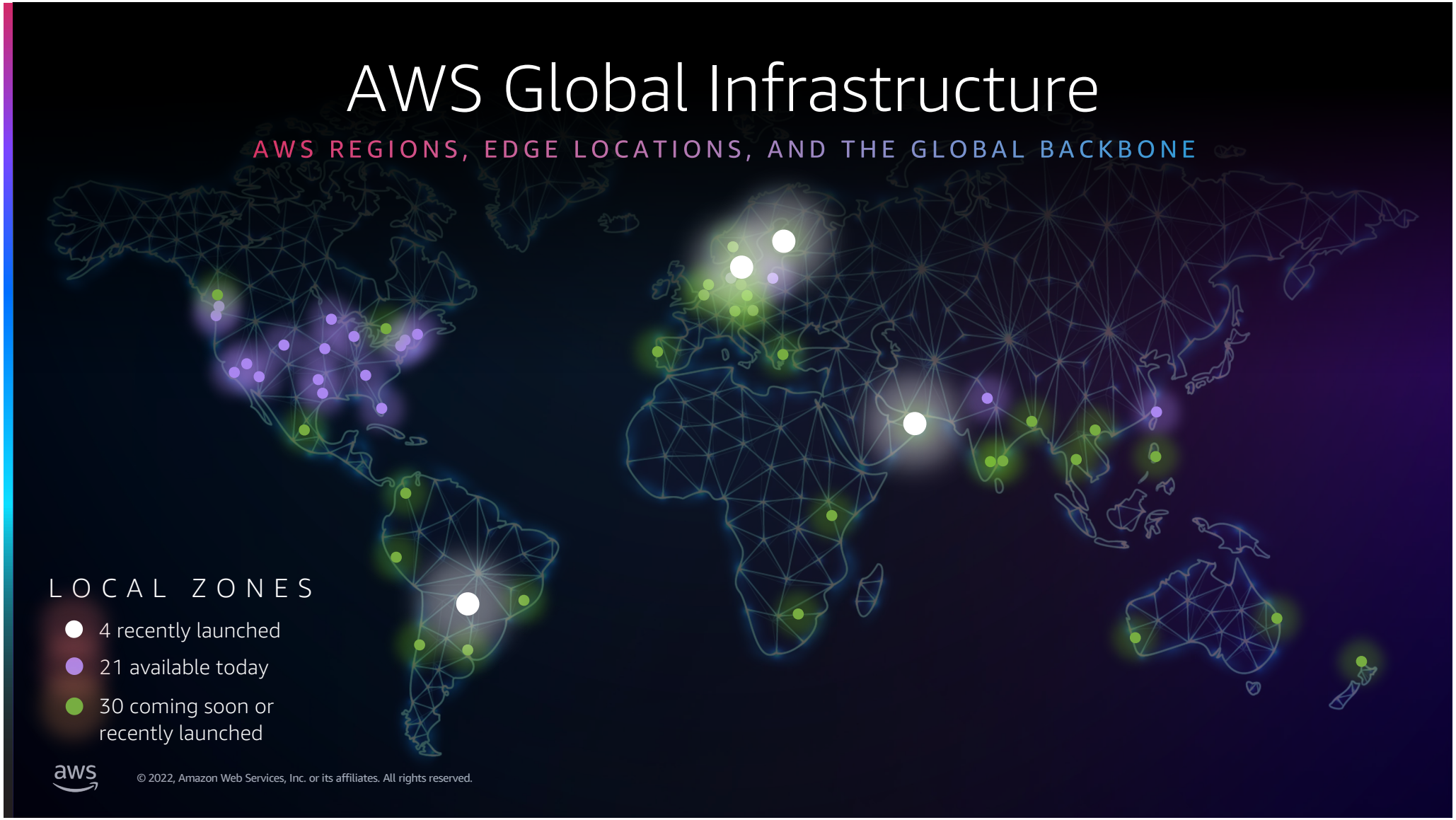
AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



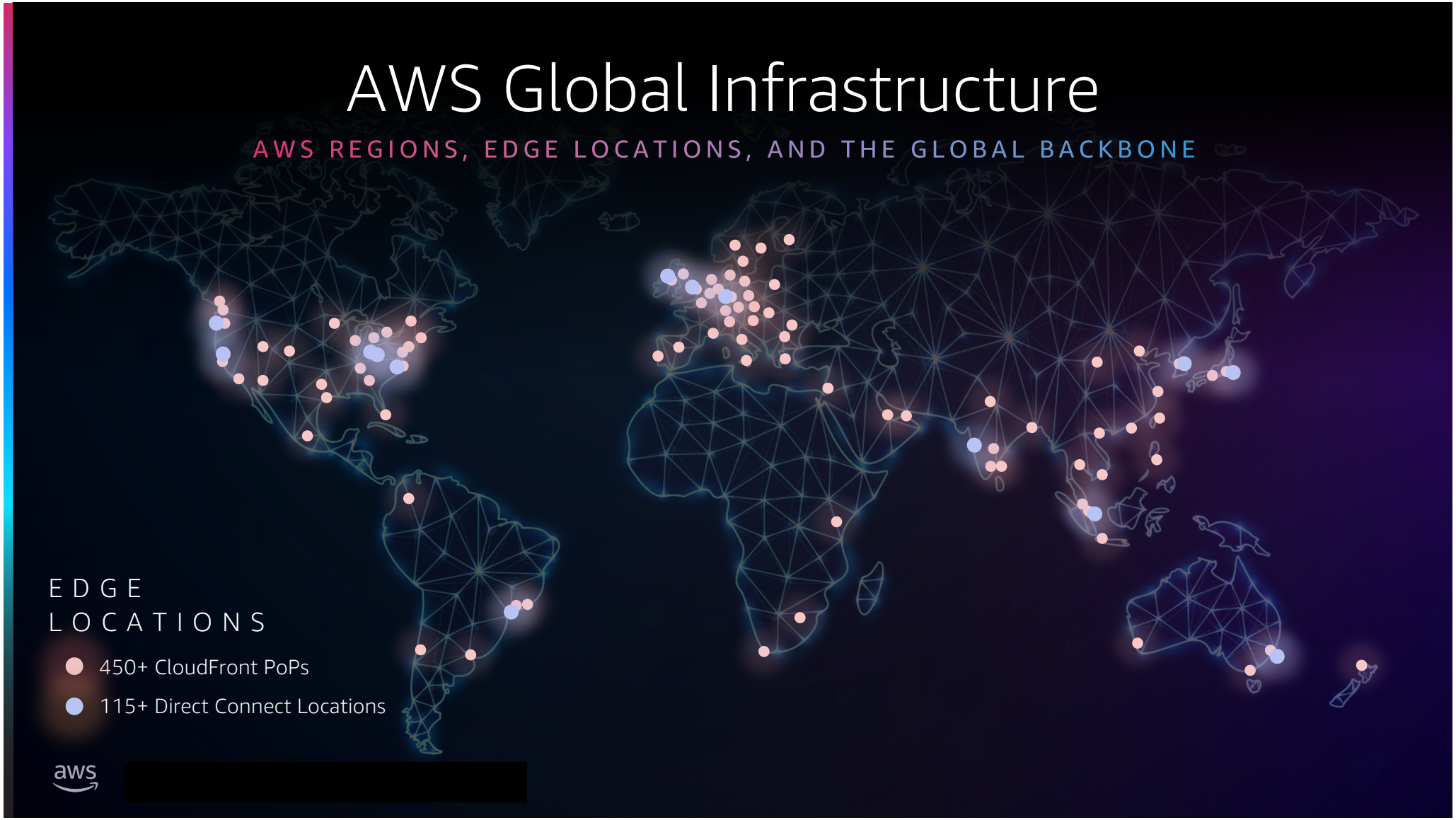
AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



AWS Global Infrastructure

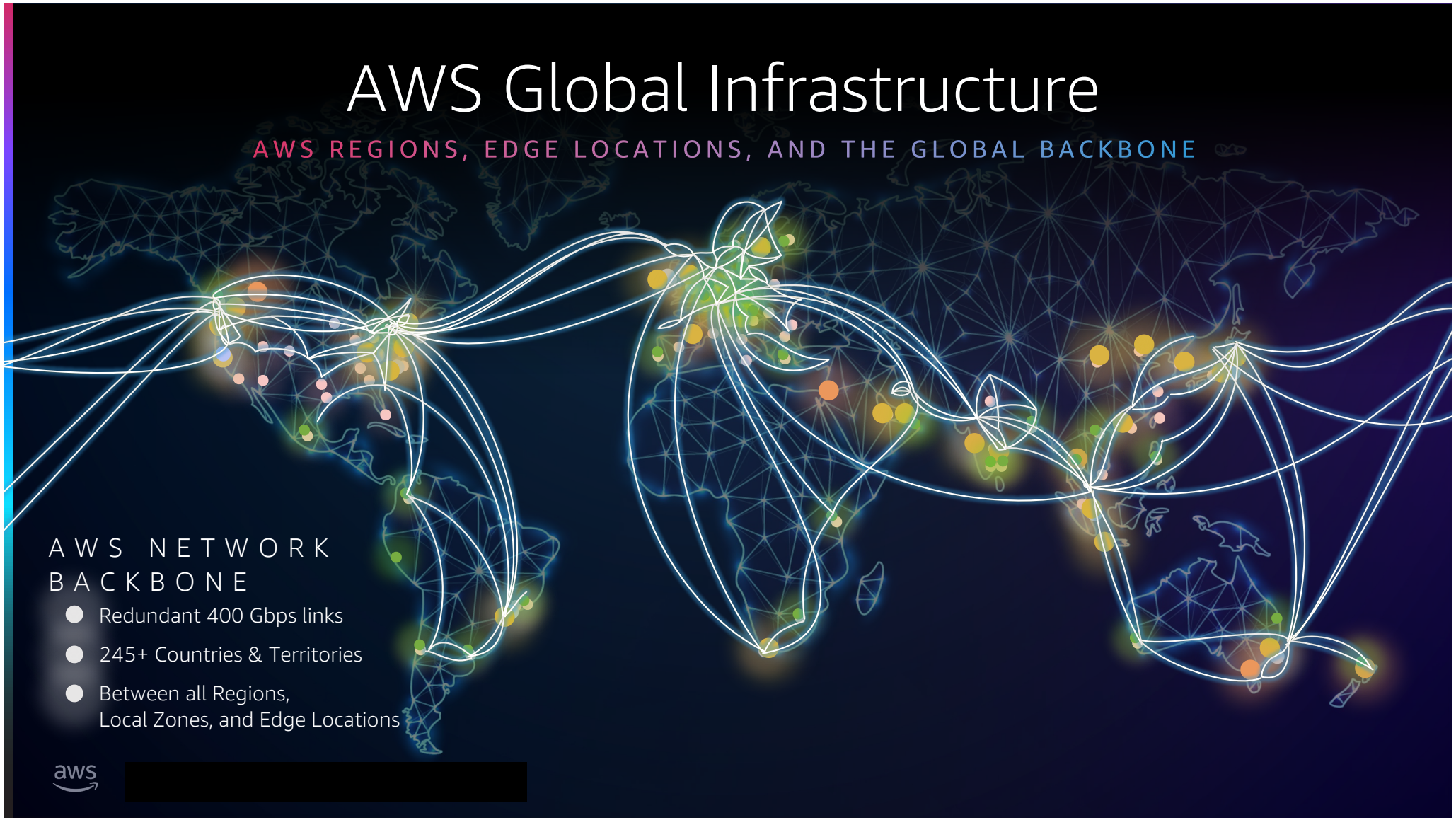
AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE



AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE

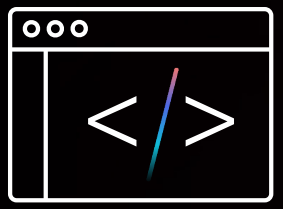
aws


AWS Global Infrastructure
AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE
AWS NETWORK BACKBONE
Redundant 400 Gbps links
245+ Countries & Territories
Between all Regions, Local Zones, and Edge Locations
aws


NFL
prime video
THURSDAY
NIGHT
FOOTBALL

Reinventing our network infrastructure

aws

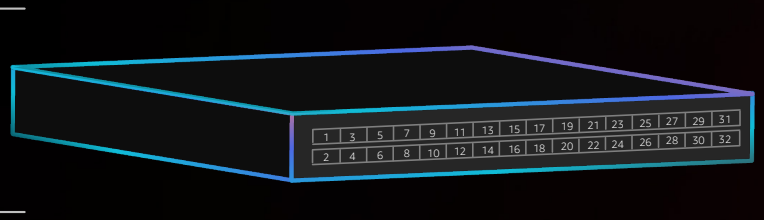


CUSTOM SOFTWARE

CUSTOM HARDWARE

- Simplicity Scales
- Focus on the benefits
- Freedom to examine trade-offs

aws



12.8

TERABITS PER SECOND

DEVICE: 1 x Switch

HEIGHT: 1 x Rack Unit (RU)

PORTS: 32 x 400G

aws

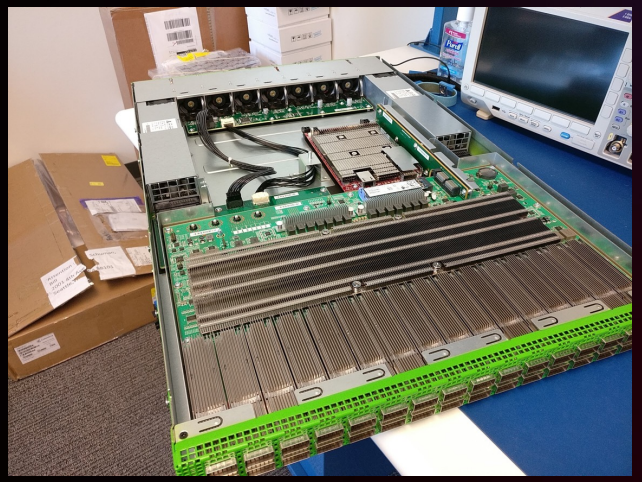
12.8

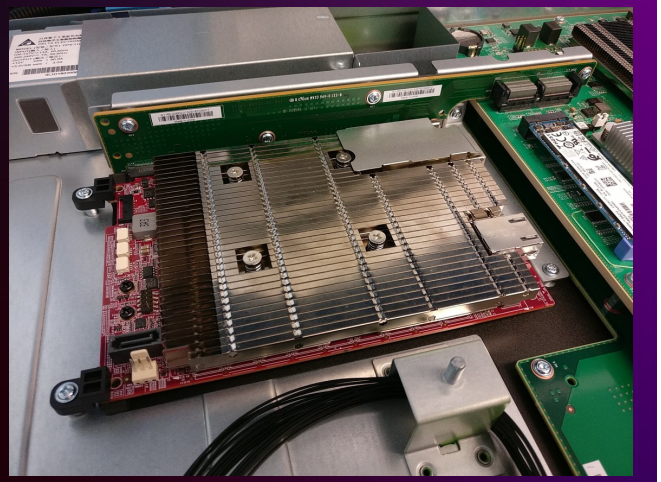
TERABITS PER SECOND

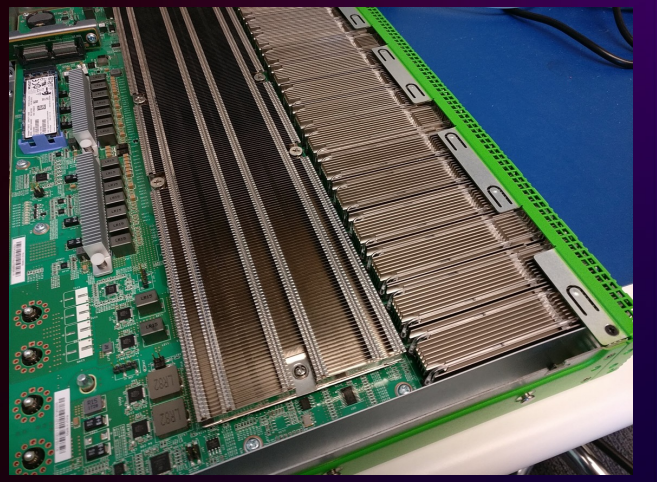
DEVICE: 1 x Switch

HEIGHT: 1 x Rack Unit (RU)

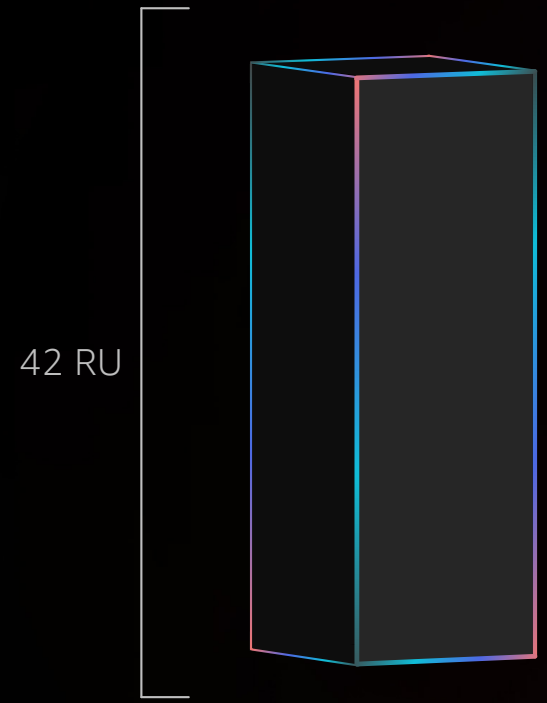
PORTS: 32 x 400G







aws



100

TERABITS PER SECOND

DEVICE: 1 rack (32 x switches)

HEIGHT: 42 x Rack Unit (RU)

PORTS: 32 x 400G (12.8 Tbps)

aws

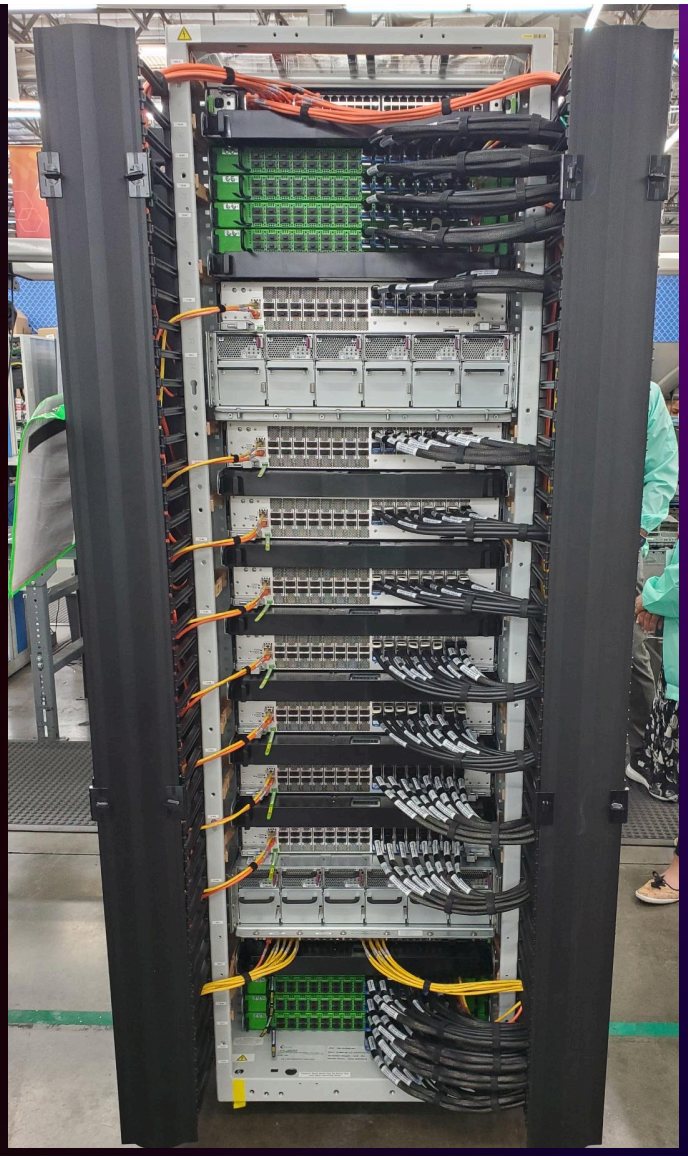
100

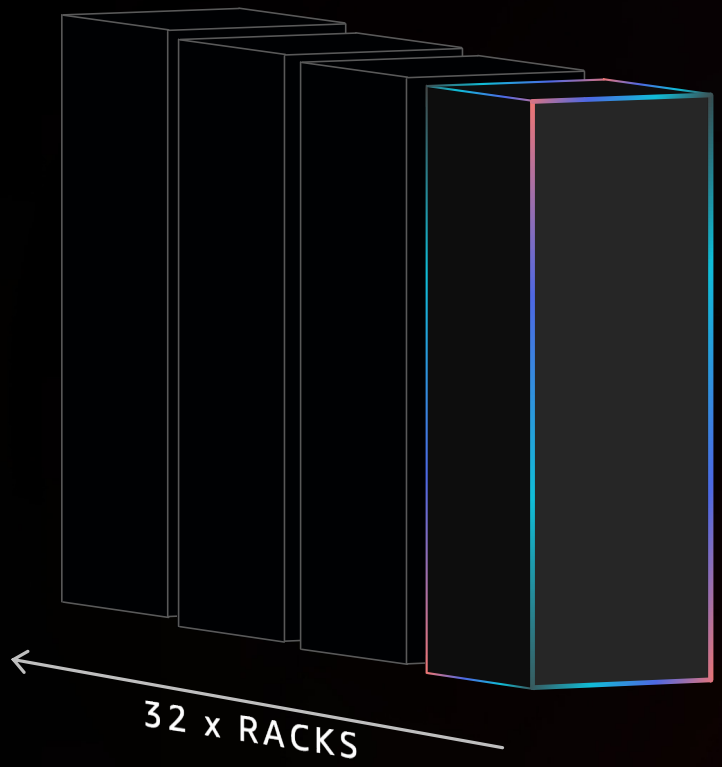
TERABITS PER SECOND

DEVICE: 1 rack (32 x switches)

HEIGHT: 42 x Rack Unit (RU)

PORTS: 32 x 400G (12.8 Tbps)





3,200

TERABITS PER SECOND

DEVICE: 32 racks (32 x switches)

HEIGHT: 42 x Rack Unit (RU)

THROUGHPUT/RACK: 100 Tbps

aws

How we do it – In rack

Direct-attach copper (DAC) cabling

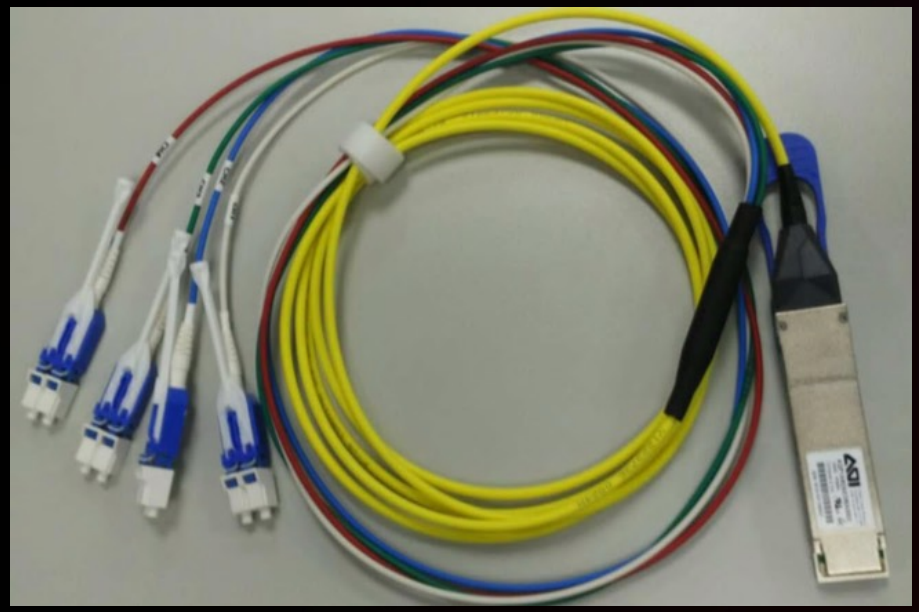
- 100G 6.7mm OD at 2.5m
- 400G 11mm OD at 2.5m

Our Biggest enemy? Cable diameter.

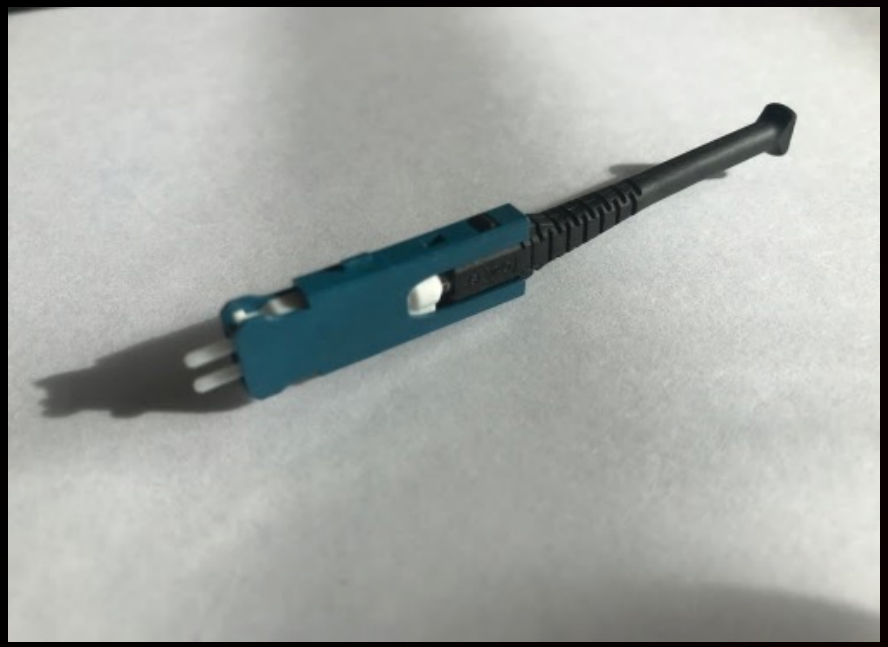
Active DAC with retimers to reduce cable area


POWER
ON/OFF
RESET
LUG STRIP
LUG TAG
400G

How we do it – Short reach



How we do it – SN connector



aws

Network Architecture and Software

Create

Config generation

Deployment coordination

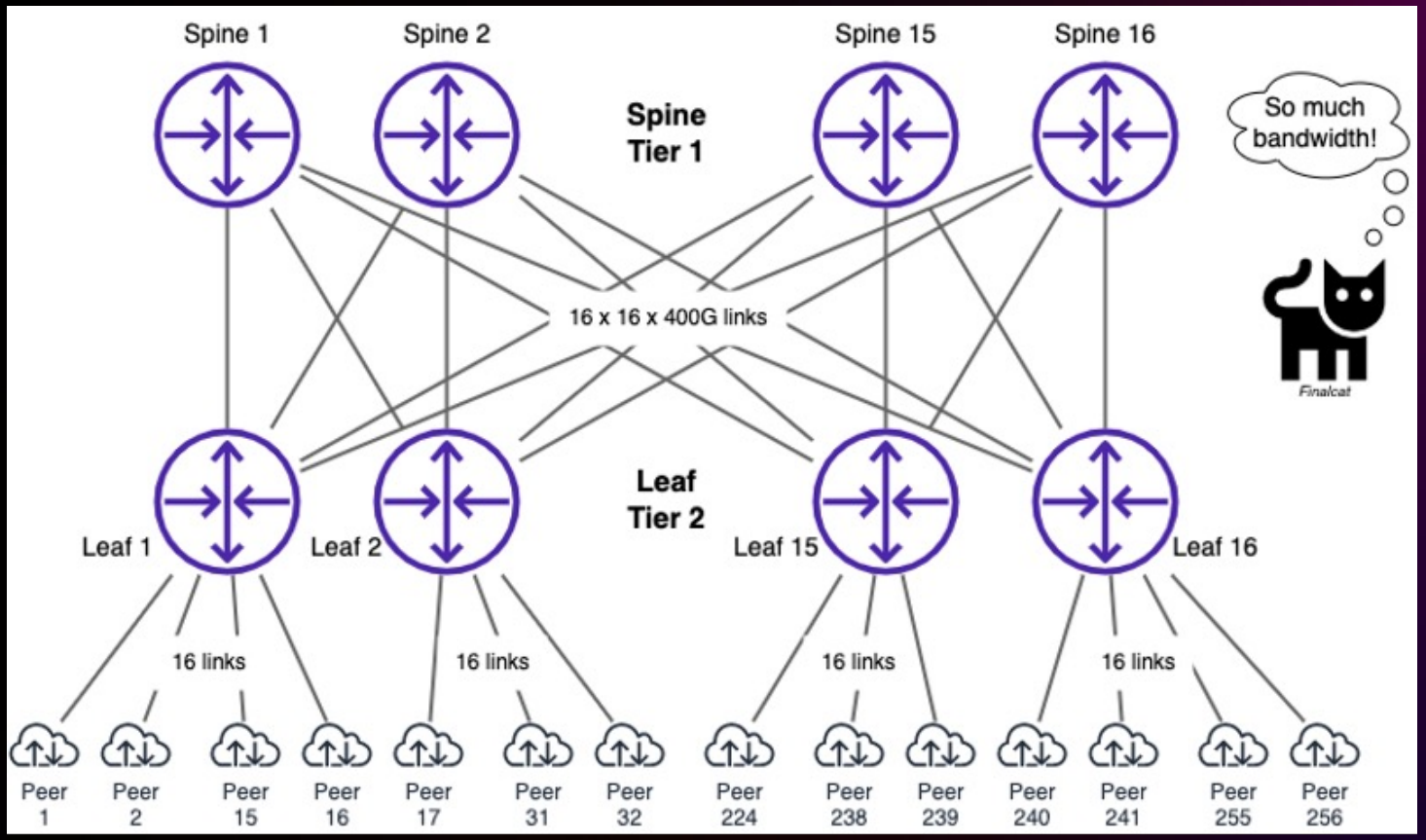
Active telemetry

Auto-remediation

NOC-less

2 tier Clos

NON OVERSUBSCRIBED ANY PORT TO ANY PORT



Metal boxes and a lot of cables

Small number of rack variations

Rack and cable switches for burn-in

Collect inventory and compare with bill of materials

Reprogram with AWS controlled binaries

How we do it

MEDIUM HAUL

Data center interconnect (DCI)

OIF 400G ZR

400G – ZR+ to 400km,
Bright ZR over 1000km

Integrated routing, DWDM, encryption

aws

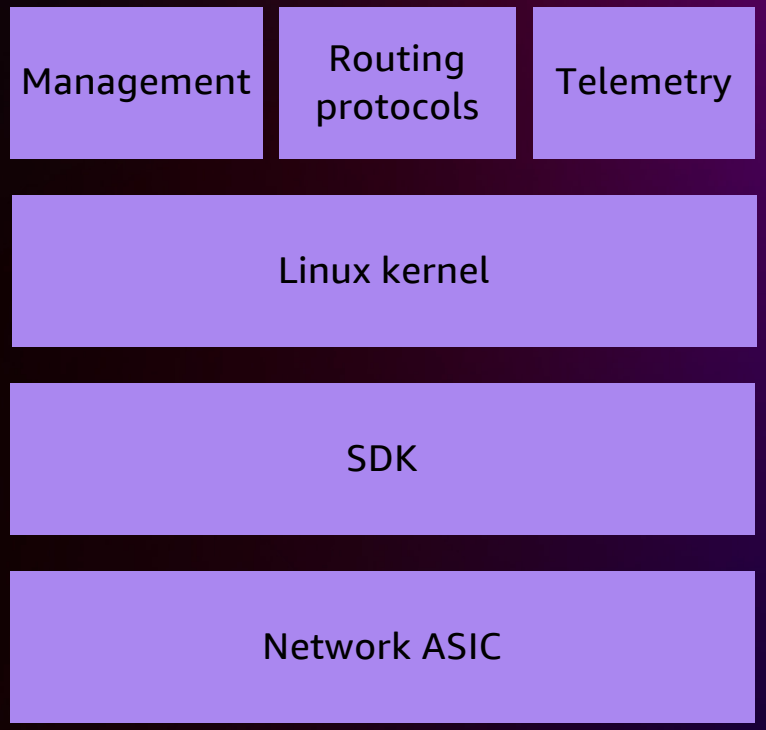
Create

NETWORK OPERATING SYSTEM

Linux-based

Multi-sourced manufacturing

Multi-ASIC



aws

Create

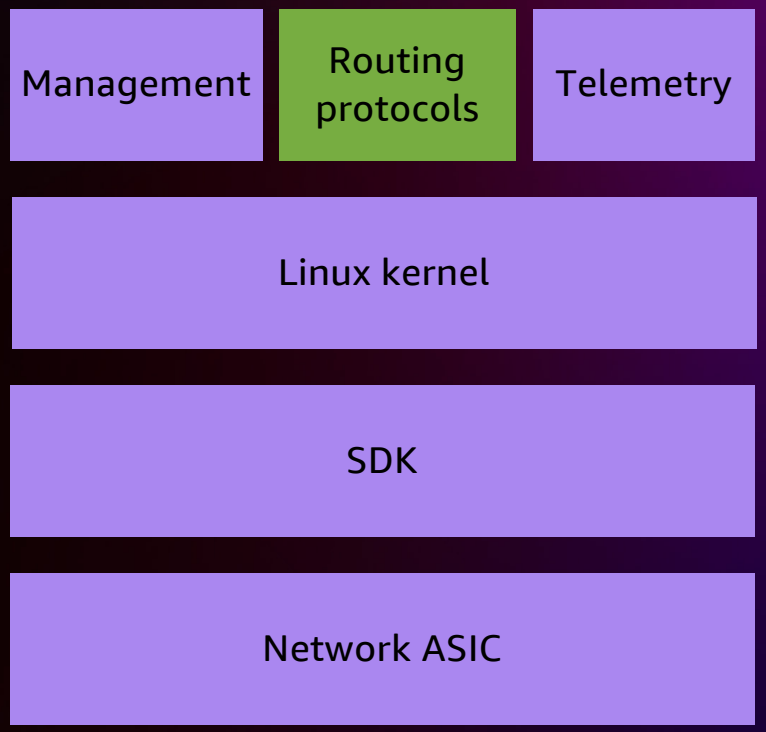
NETWORK OPERATING SYSTEM

Linux-based

Multi-sourced manufacturing

Multi-ASIC

OSPF/BGP ++



aws

Traceroute from outside AWS

```
% traceroute www.amazon.com
...
 5  * * *
 6  * * *
 7  52.93.33.77 (52.93.33.77)  1.984 ms 1.789 ms 1.983 ms
 8  52.93.33.130 (52.93.33.130)  2.316 ms 2.362 ms  2.891 ms
 9  150.222.72.105 (150.222.72.105)  3.682 ms  3.044 ms  3.002 ms
10  * * *
11  * * *
12  * * *
13  * * *
14  * * *
15  server-65-8-32-17.mel50.r.cloudfront.net (65.8.32.17)  3.650 ms  4.866 ms  3.033 ms
```

Traceroute from inside AWS

Take a look at this **traceroute** from **#AWS** EC2 instance towards internet through NAT GW. Check out those **Class E** addresses 👀

```
traceroute to 8.8.8.8 (8.8.8.8), 30 hops max, 60 byte packets
 1  * * *
 2  * * *
 3  * * *
 4  10.117.52.85 (10.117.52.85)  4.268 ms  4.256 ms  4.244 ms
 5  100.64.95.255 (100.64.95.255)  11.343 ms  11.331 ms 100.64.95.253 (100.64.95.253)  6.606 ms
 6  240.1.240.32 (240.1.240.32)  5.200 ms  3.643 ms 240.1.236.32 (240.1.236.32)  3.122 ms
 7  100.66.13.156 (100.66.13.156)  9.308 ms 240.1.236.62 (240.1.236.62)  3.074 ms 240.1.236.57 (240.1.236.57)  2.630 ms
 8  240.1.236.24 (240.1.236.24)  2.552 ms 100.66.14.142 (100.66.14.142)  16.584 ms 240.1.236.29 (240.1.236.29)  2.831 ms
 9  108.166.244.14 (108.166.244.14)  2.776 ms 241.0.12.137 (241.0.12.137)  3.514 ms 108.166.244.15 (108.166.244.15)  3.057 ms
10  108.166.244.18 (108.166.244.18)  2.914 ms 108.166.248.61 (108.166.248.61)  3.565 ms 108.166.244.27 (108.166.244.27)  2.801 ms
11  242.0.78.241 (242.0.78.241)  3.146 ms 108.166.248.50 (108.166.248.50)  3.721 ms 242.0.78.249 (242.0.78.249)  2.798 ms
12  242.0.90.89 (242.0.90.89)  3.555 ms 242.0.91.65 (242.0.91.65)  3.354 ms 15.230.134.185 (15.230.134.185)  3.989 ms
13  15.230.39.40 (15.230.39.40)  4.537 ms 15.230.134.84 (15.230.134.84)  4.083 ms 52.95.2.86 (52.95.2.86)  4.151 ms
14  15.230.140.117 (15.230.140.117)  4.460 ms 15.230.39.234 (15.230.39.234)  4.574 ms 15.230.140.159 (15.230.140.159)  4.195 ms
15  52.93.239.36 (52.93.239.36)  7.733 ms 52.95.3.39 (52.95.3.39)  6.571 ms 100.91.177.167 (100.91.177.167)  14.866 ms
16  100.100.6.57 (100.100.6.57)  14.555 ms 100.91.177.1 (100.91.177.1)  14.291 ms 100.91.177.27 (100.91.177.27)  15.858 ms
17  100.100.77.70 (100.100.77.70)  14.679 ms 100.100.92.72 (100.100.92.72)  14.630 ms 100.100.76.134 (100.100.76.134)  14.319 ms
18  100.100.69.163 (100.100.69.163)  14.312 ms 100.100.64.165 (100.100.64.165)  45.745 ms 100.100.86.99 (100.100.86.99)  14.297 ms
19  100.100.2.32 (100.100.2.32)  14.361 ms 100.100.88.227 (100.100.88.227)  14.704 ms 100.100.4.24 (100.100.4.24)  15.748 ms
20  99.83.113.93 (99.83.113.93)  15.617 ms 100.100.34.94 (100.100.34.94)  14.689 ms 99.82.181.25 (99.82.181.25)  14.959 ms
21  * 99.83.113.93 (99.83.113.93)  16.883 ms 108.170.246.33 (108.170.246.33)  16.294 ms
22  dns.google (8.8.8.8)  15.325 ms * *
```

aws

Disaggregated control plane

COMBINATION OF ON-DEVICE AND OFF-DEVICE

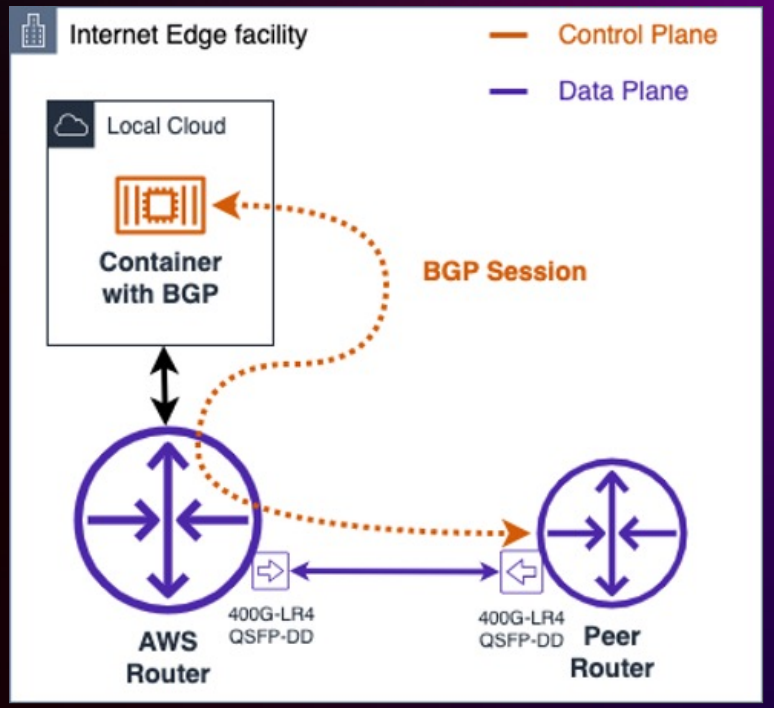
On-device handles local things like LACP, ARP/ND and all aspects of physical connectivity

BGP speaker runs elsewhere

Faster convergence and higher scale than would otherwise be possible

Enables us to iterate/evolve each part separately

Peer doesn't see anything different, TTL1 or TTL255 BGP still works the same way



The curious case of flaky IPv6 NS

LINUX MCAST_RESOLICIT (NON-DEFAULT) REQUIRED FOR NON-LINK-LOCAL IPV6 NS

```
% ip -ts monitor neigh dev bond1
[2023-01-13T02:58:15.544747] 2620:107:4008:xxx::2 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:58:15.649269] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:58:15.650764] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE
[2023-01-13T02:58:45.852977] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 PROBE
[2023-01-13T02:58:45.854469] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 REACHABLE
[2023-01-13T02:58:46.112645] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:58:46.114825] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE
[2023-01-13T02:58:52.413809] 2620:107:4008:xxx::2 dev bond1  router FAILED
[2023-01-13T02:59:07.779235] 2620:107:4008:xxx::2 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE
[2023-01-13T02:59:16.305279] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 PROBE
[2023-01-13T02:59:16.306371] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 REACHABLE
[2023-01-13T02:59:16.473164] 2620:107:4008:xxx::2 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:59:16.570665] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:59:16.574393] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE
[2023-01-13T02:59:46.767019] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 PROBE
[2023-01-13T02:59:46.770263] 99.83.1xx.xx dev bond1 lladdr d4:6a:35:35:4c:92 REACHABLE
[2023-01-13T02:59:47.025611] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router PROBE
[2023-01-13T02:59:47.026513] fe80::d66a:35ff:fe25:4c92 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE
[2023-01-13T02:59:53.341824] 2620:107:4008:xxx::2 dev bond1  router FAILED
[2023-01-13T03:00:07.779211] 2620:107:4008:xxx::2 dev bond1 lladdr d4:6a:35:35:4c:92 router REACHABLE

% tcpdump -i bond1 -n -p --direction=out 'icmp6'
04:15:57.597793 IP6 fe80::a2d0:dcff:fefc:8ed6 > 2620:107:4008:xxx::2: ICMP6, neigh solicitation, who has 2620:107:4008:xxx::2
04:16:02.717802 IP6 fe80::a2d0:dcff:fefc:8ed6 > 2620:107:4008:xxx::2: ICMP6, neigh solicitation, who has 2620:107:4008:xxx::2
04:16:07.837808 IP6 fe80::a2d0:dcff:fefc:8ed6 > 2620:107:4008:xxx::2: ICMP6, neigh solicitation, who has 2620:107:4008:xxx::2
04:16:10.407026 IP6 fe80::a2d0:dcff:fefc:8ed6 > fe80::d66a:35ff:fe35:4c92: ICMP6, neighbor advertisement, tgt is fe80::a2d0:d
04:16:12.957792 IP6 fe80::a2d0:dcff:fefc:8ed6 > ff02::1:ff00:2: ICMP6, neighbor solicitation, who has 2620:107:4008:xxx::2
```

36 IPv6 NS sent, none answered

As soon as we age out the entry..

..it then answers

..rinse/repeat..

aws

AWS Global Infrastructure

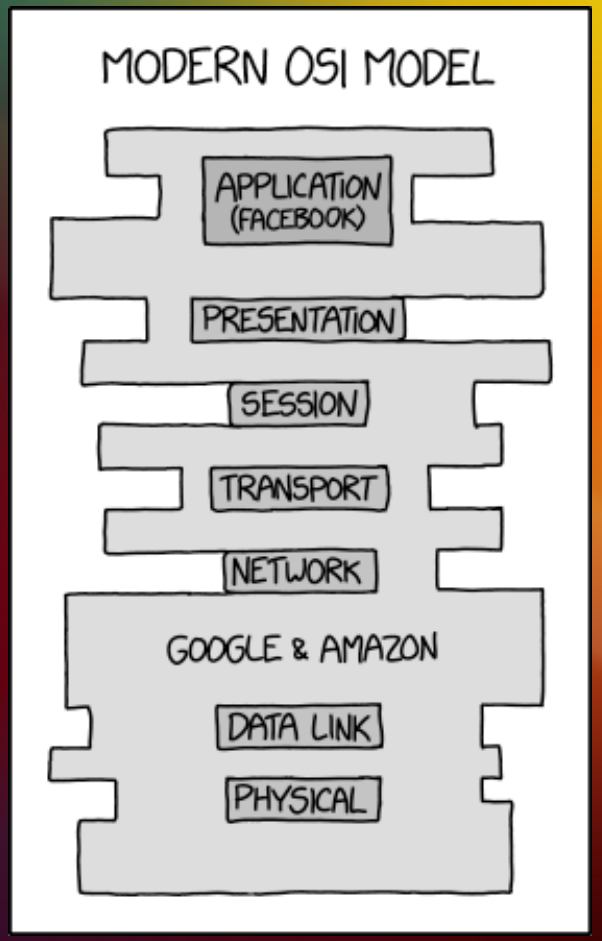
AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE

aws

Thank You

Lincoln Dale
Senior Principal Engineer
AWS – AS16509

Fredrik Korsbäck
Senior Infrastructure Business Developer
AWS – AS16509



Source: https://xkcd.com/2105, Randall Munroe

aws