Dive deep on AWS edge networking infrastructure

Lincoln Dale
Senior Principal Engineer
AWS – AS16509

Fredrik Korsbäck
Senior Infrastructure Business Developer
AWS – AS16509
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 Regions, Edge Locations, and the Global Backbone

Regional Expansion

- Available Today: 30 Regions
- Coming soon or recently launched: 6 Regions
AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE

LOCAL ZONES

- 4 recently launched
- 21 available today
- 30 coming soon or recently launched

© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.
AWS Global Infrastructure

AWS Regions, Edge Locations, and the Global Backbone

Edge Locations
- 450+ CloudFront PoPs
- 115+ Direct Connect Locations
AWS Global Infrastructure

AWS REGIONS, EDGE LOCATIONS, AND THE GLOBAL BACKBONE
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
Reinventing our network infrastructure
CUSTOM SOFTWARE

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

CUSTOM HARDWARE
12.8 Tbps

TERABITS PER SECOND

DEVICE: 1 x Switch

HEIGHT: 1 x Rack Unit (RU)

PORTS: 32 x 400G
12.8 TERABITS PER SECOND

DEVICE: 1 x Switch
HEIGHT: 1 x Rack Unit (RU)
PORTS: 32 x 400G
100 TERABITS PER SECOND

DEVICE: 1 rack (32 x switches)
HEIGHT: 42 x Rack Unit (RU)
PORTS: 32 x 400G (12.8 Tbps)
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
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
How we do it – Short reach
How we do it – SN connector
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
Create

**NETWORK OPERATING SYSTEM**

- Linux-based
- Multi-sourced manufacturing
- Multi-ASIC

- Management
- Routing protocols
- Telemetry
- Linux kernel
- SDK
- Network ASIC
Create

**NETWORK OPERATING SYSTEM**

- Linux-based
- Multi-sourced manufacturing
- Multi-ASIC
- OSPF/BGP ++

**Diagram:**
- Management
- Routing protocols
- Telemetry
- Linux kernel
- SDK
- Network ASIC
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 108.64.95.253 (100.64.95.253) 6.686 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)  3.874 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) 2.031 ms
9  108.166.244.14 (108.166.244.14)  2.776 ms 240.1.12.137 (240.1.12.137)  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) 2.001 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 242.0.91.15 (242.0.91.15)  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)  4.883 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  52.93.299.36 (52.93.299.36)  7.733 ms 52.95.3.39 (52.95.3.39)  6.571 ms 100.91.177.167 (100.91.177.167) 14.886 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 99.82.101.25 (99.82.101.25) 14.959 ms
21  99.83.113.93 (99.83.113.93) 16.883 ms 100.170.246.33 (100.170.246.33) 16.294 ms
```

Disaggregated control plane

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

% tcpdump -i bond1 -n -p --direction=out 'icmp6'
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

As soon as we age out the entry..
..it then answers
..rinse/repeat..

36 IPv6 NS sent, none answered
AWS Global Infrastructure

AWS Regions, Edge Locations, and the Global Backbone
Thank You

Lincoln Dale
Senior Principal Engineer
AWS – AS16509

Fredrik Korsbäck
Senior Infrastructure Business Developer
AWS – AS16509

Source: [https://xkcd.com/2105](https://xkcd.com/2105), Randall Munroe