The Computer Network behind the Social Network

James Hongyi Zeng
Engineering Manager, Network Infra
APNet 2019, Beijing, China
Facebook Family

2.7B people every month
2.1B people every day

(Q2, 2019)
About Me

- Joined Facebook networking in 2014
- Supporting Routing and UI team
- [https://research.fb.com/category/systems-and-networking/](https://research.fb.com/category/systems-and-networking/)
How Users Reach Facebook

- Internet
- Backbone
- Datacenter Network

- Edge Network
- Backbone Network
Agenda

- Edge Network
- Backbone Network
- Datacenter Network
Agenda

• Edge Network
• Backbone Network
• Datacenter Network
Edge Network

- Goal: Delivers the traffic to ISP and ultimately to users
- Majority of users are on mobile
- Majority of users are on IPv6
  - IPv6 penetration rate is at 56% in the United States
  - https://www.facebook.com/ipv6/
Facebook’s Traffic

**Dynamic Requests**
(Not Cachable)
- News Feed
- Likes
- Messaging
- Status Updates

**Static Requests**
(Cachable)
- Photos
- Videos
- JavaScript

Dynamic Requests:
- Java
- JavaScript

Static Requests:
- Cachable

Status Updates:
- (not Cachable)
DNS Based Load Balancing

Internet

US-EAST
L4LB
L7LB
Web Server

US-WEST
L4LB
L7LB
Web Server

www?

US-east.facebook.com

DNS LB
How about static content?
Edge Network Summary

• Software Hierarchy to scale
  • DNS Load Balancer (to Datacenter/POP)
  • Router + Anycast BGP, Layer 3 Load balancer (to Layer 4 Load Balancer)
  • Layer 4 Load Balancer (to Layer 7 Load Balancer)
  • Layer 7 Load Balancer (to Web Server)
• POP + DC to scale
  • Reduce RTT for initial setup
  • Cache content closer to users
Agenda

• Edge Network
• **Backbone Network**
• Datacenter Network
Backbones at Facebook

- Classic Backbone (CBB)
  - Connects POP and DCs
  - RSVP-TE, Vendor software solution
- Express Backbone (EBB)
  - Connects DC and DC
  - Centralized control
Three Datacenters
Add Planes
N-way Active-active Redundancy
Incremental changes and canary
A/B Testing

Algorithm 1

Algorithm 2
Open/R

• Routing Protocol supports EBB
  • Establish basic reachability among routers (OSPF, IS-IS)
• Extensible (e.g., key-value store)
• In-house software
• Run as agent on EBB routers
• EBB is first production network where Open/R is the sole IGP
Typical IGP metric configuration

<table>
<thead>
<tr>
<th>Type</th>
<th>Link Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans Atlantic</td>
<td>100</td>
</tr>
<tr>
<td>Trans Pacific</td>
<td>150</td>
</tr>
<tr>
<td>US-West to US-East</td>
<td>50</td>
</tr>
</tbody>
</table>
Open/R: Calculate link metric with RTT

RTT = 200ms

metric = 200
Backbone Network Summary

- Two backbones
  - CBB: Connects POPs and DCs
  - EBB: Inter-DC backbone
- Plane architecture
  - Reliability, maintenance, experiment
- Software
  - Centralized control
  - Innovative distributed routing protocols to minimize configuration
Agenda

• Edge Network
• Backbone Network
• **Datacenter Network**
Classic Facebook Fabric

Spine Switches

4 Spine Planes

100G links

fabric uplinks

Edge Switches

Rack Switches

Server Pods

Edge Pods

Pod 1

Pod 2

Pod 3

Pod X

Pod Y

48 ports in Pod 1
48 ports in Pod 2
48 ports in Pod 3
48 ports in Pod X
48 ports in Pod Y
Growing Pressure

Expanding Mega Regions (5-6 buildings) = accelerated fabric-to-fabric East-West demand

Compute-Storage and AI disaggregation requires Terabit capacity per Rack

Both require larger fabric Spine capacity (by 2-4x) ...
F16 – Facebook’s new topology

- 16-plane architecture
- 6-16x spine capacity on day 1
- 1.6T raw capacity per rack
- Fewer chips* = better power & space
Mega Region
Mega Region

F16

Fabric Aggregator
Minipack – 128 x 100G Switch

- Single 12.8T ASIC
- Modular design
- Mature optics
- Lower power/smaller size
Fabric Aggregator

- Disaggregated design for scale
- Built upon smaller commodity switches
White Box Switch

Customizable switch hardware and software

- **Customized** hardware
- **Pick** the **minimal** software needed for the specific network
- **Powerful CPU** to run more complex software
FBOSS Overview

External Software

- Protocols (BGP, ECMP)
- Network Configurator
- Monitoring Service

Switch Software

FBOSS

Switch Hardware

Switch ASIC
FBOSS Design Principles

• **Switch-as-a-Server**
  • Continuous integration and staged deployment
  • Integrate closely with existing software services
  • Open-source software

• **Deploy-Early-and-Iterate**
  • Focus on developing and deploying minimal set of features
  • Quickly iterate with smaller “diffs”
FBOSS Testing and Deployment

3 Stage Deployment via *fbossdeploy*

- **Continuous Canary**
  - Deploy all commits continuously to 1~2 switches for each type

- **Daily Canary**
  - Deploy all of single day’s commits to 10~20 switches for each type

- **Staged Deployment**
  - Final stage to push all the commits to all the switches in the DC
  - Performed once every two weeks for reliability
Datacenter Network Summary

- Datacenters are huge
  - Internally: Clos topology
  - Intra-region connectivity is challenging too
- In-house Hardware and Software
  - Minipack, Fabric Aggregator
  - FBOSS
Extended Reading

• Inside the Social Network’s (Datacenter) Network, SIGCOMM 2015
• Robotron: Top-down Network Management at Facebook Scale, SIGCOMM 2016
• Engineering Egress with Edge Fabric: Steering Oceans of Content to the World, SIGCOMM 2017
• FBOSS: Building Switch Software at Scale, SIGCOMM 2018