The Computer Network behind the Social Network

James Hongyi Zeng

Engineering Manager, Network Infra

APNet 2019, Beijing, China

Facebook Family

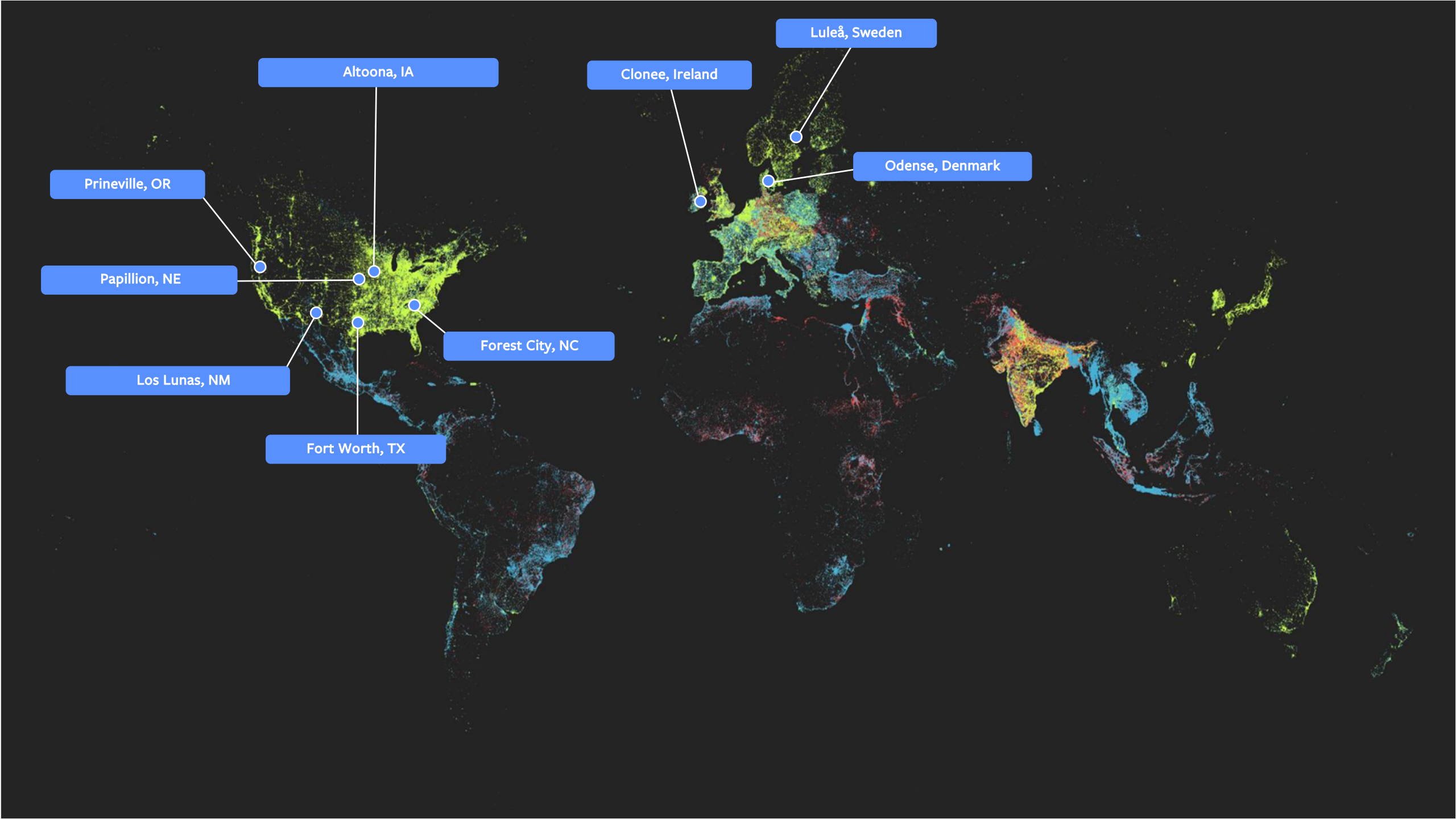
2.7B people every month

2.1B people every day

(Q2, 2019)

About Me

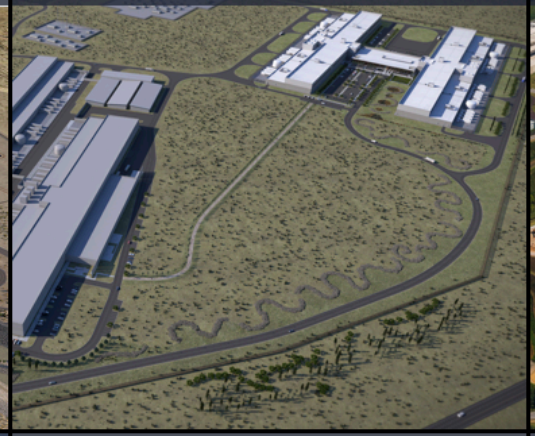
- Joined Facebook networking in 2014
- Supporting Routing and UI team
- https://research.fb.com/category/systems-and-networking/


Luleå, Sweden
Altoona, IA
Clonee, Ireland
Odense, Denmark
Prineville, OR
Papillion, NE
Forest City, NC
Los Lunas, NM
Fort Worth, TX

PAPILLION, NE

LOS LUNAS, NM

PRINEVILLE, OR



FOREST CITY, NC

FORT WORTH, TX

NEWTON COUNTY, GA

NEW ALBANY, OH

ODENSE, DENMARK

LULEÅ, SWEDEN

ALTOONA, IA

CLONEE, IRELAND

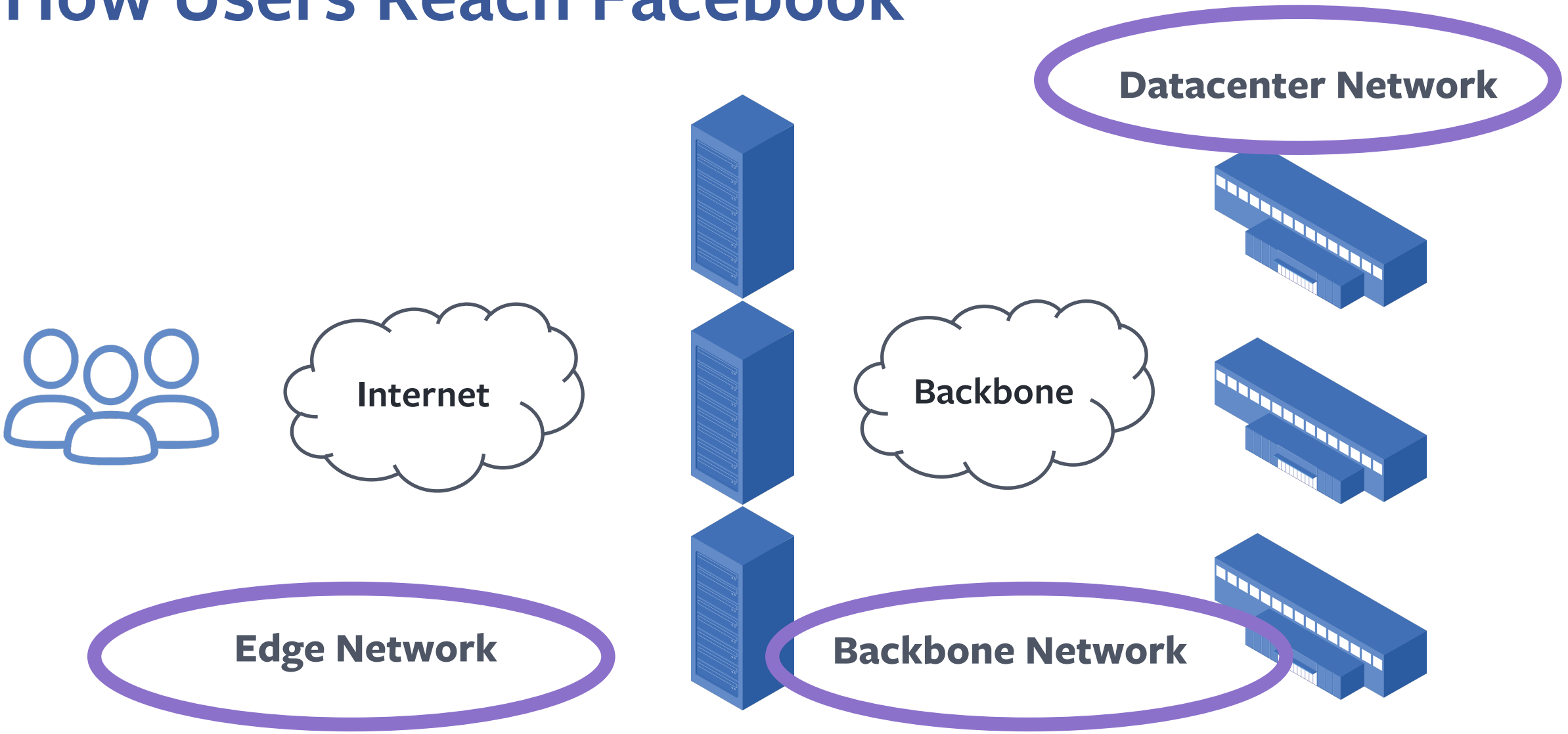
HENRICO, VA

EAGLE MOUNTAIN, UT

HUNTSVILLE, AL

SINGAPORE

How Users Reach Facebook



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

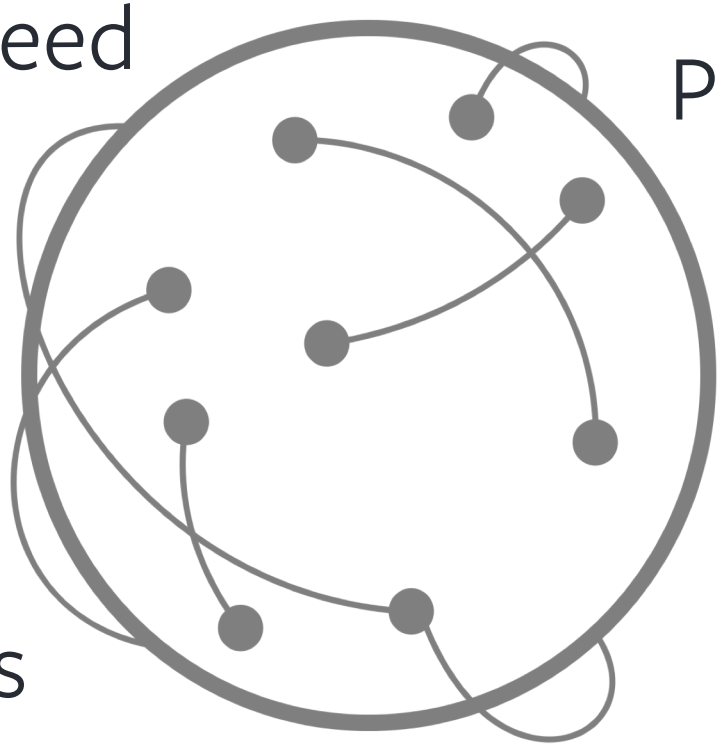
Dynamic Requests
(not Cachable)

News Feed

Likes

Messaging

Status Updates



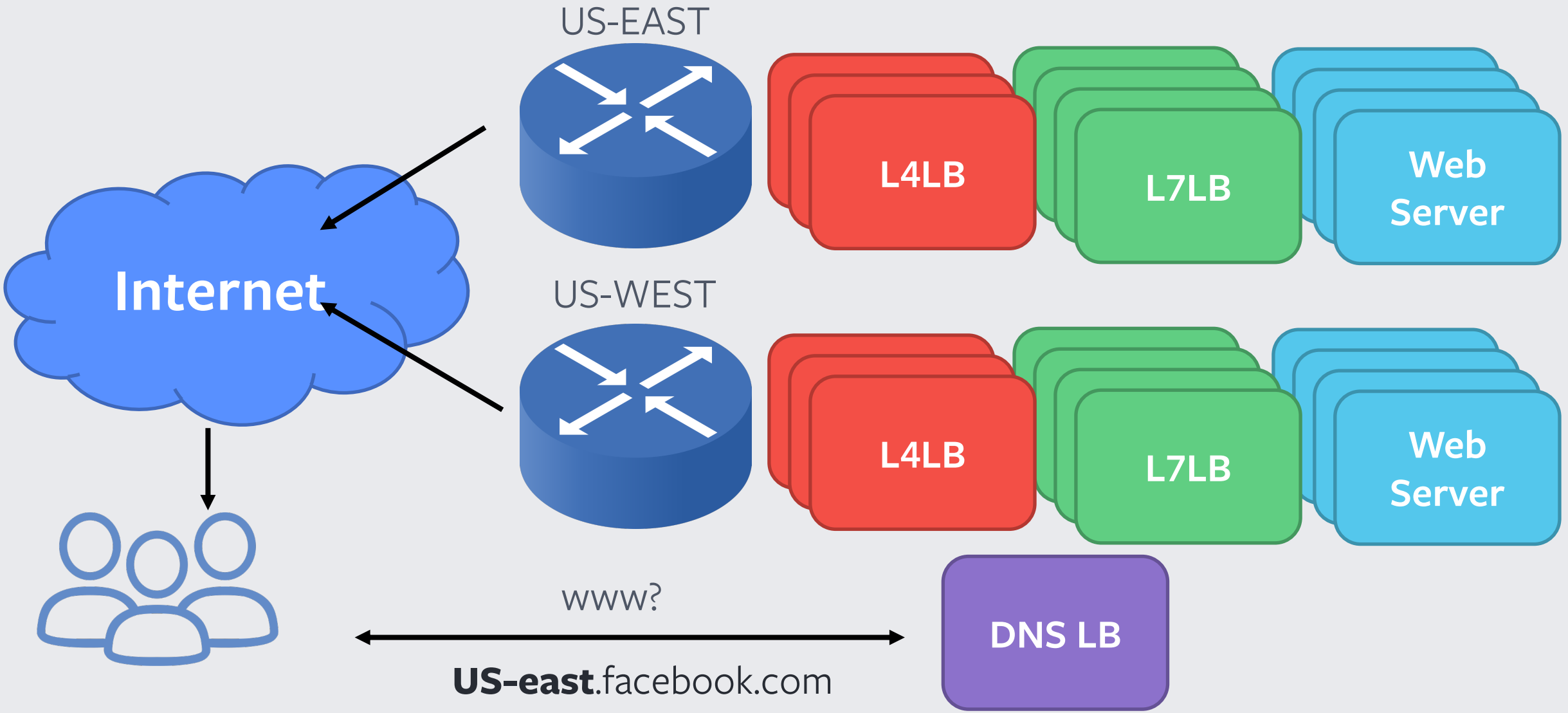
Static Requests
(Cachable)

Photos

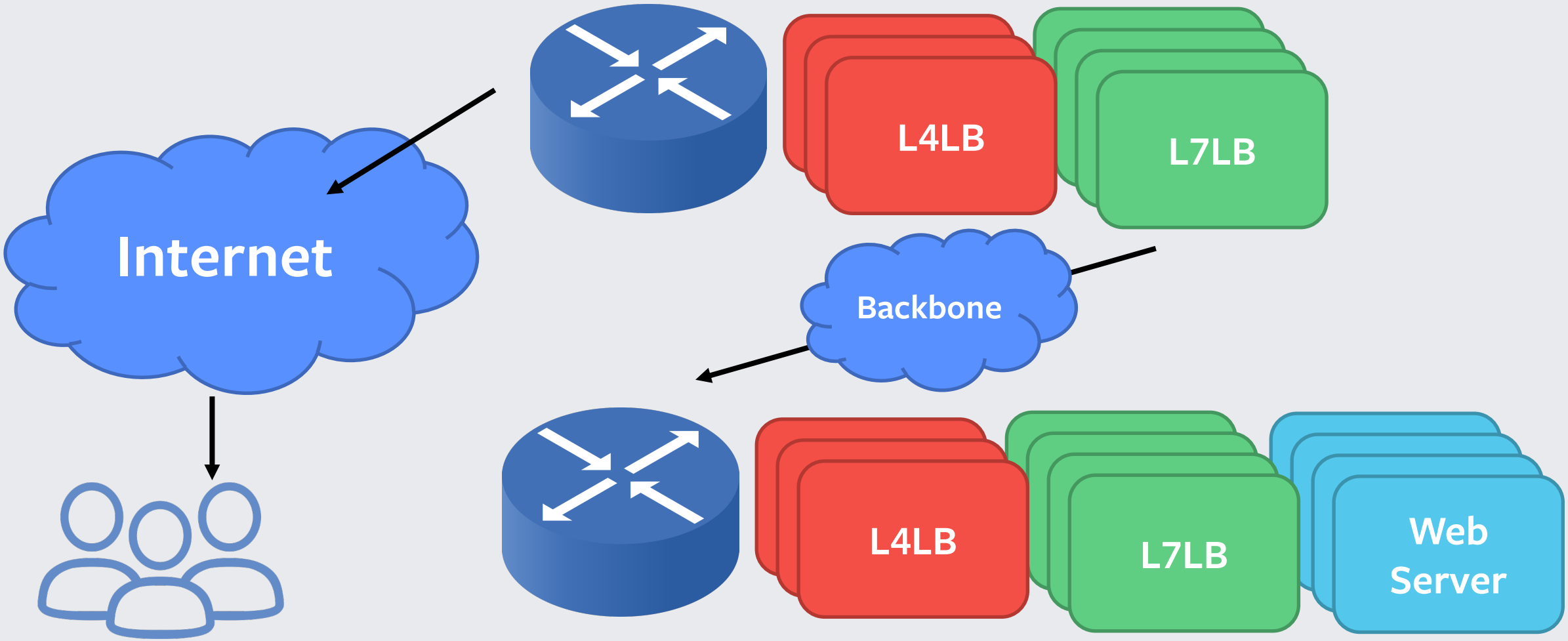
Videos

JavaScript

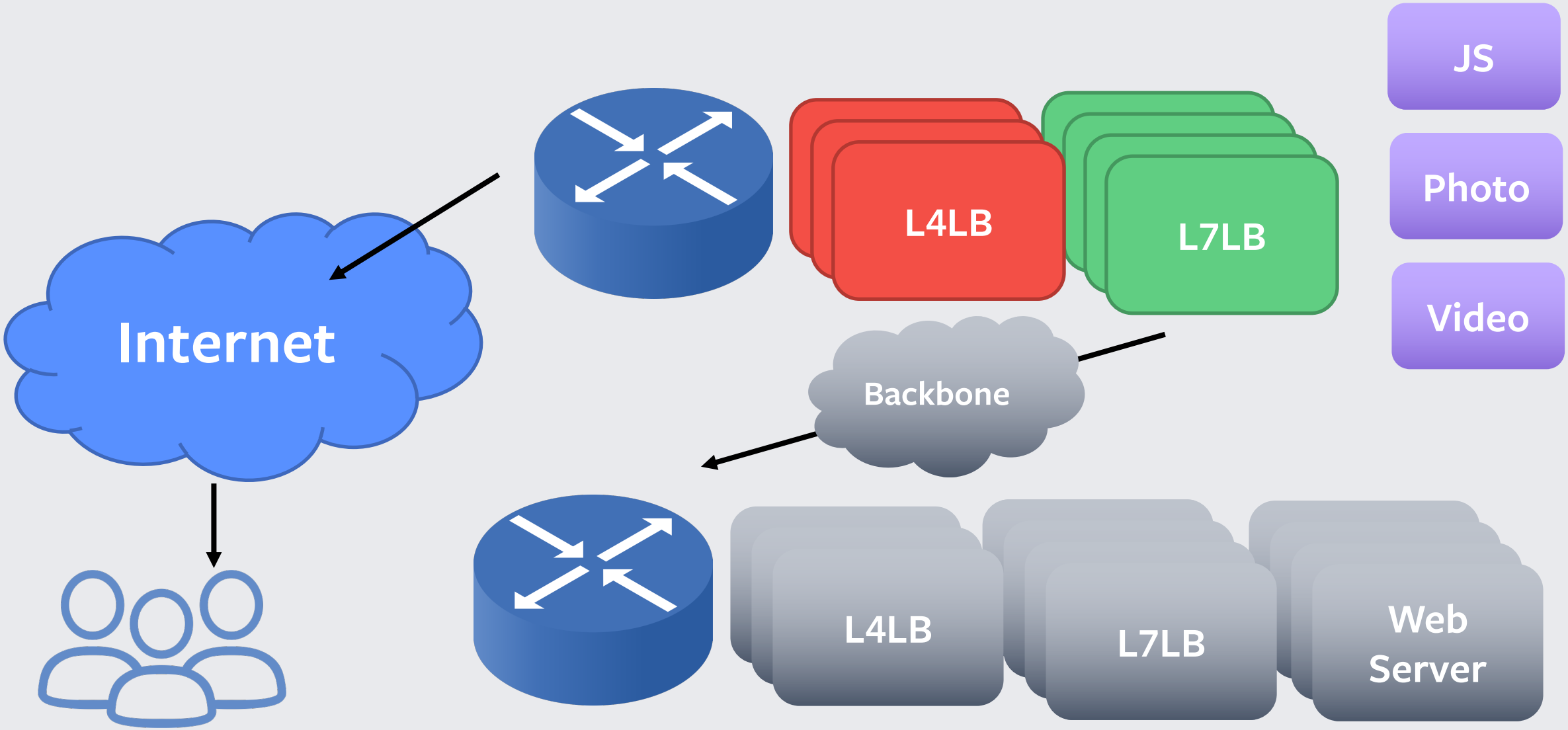
DNS Based Load Balancing



POP + DC



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

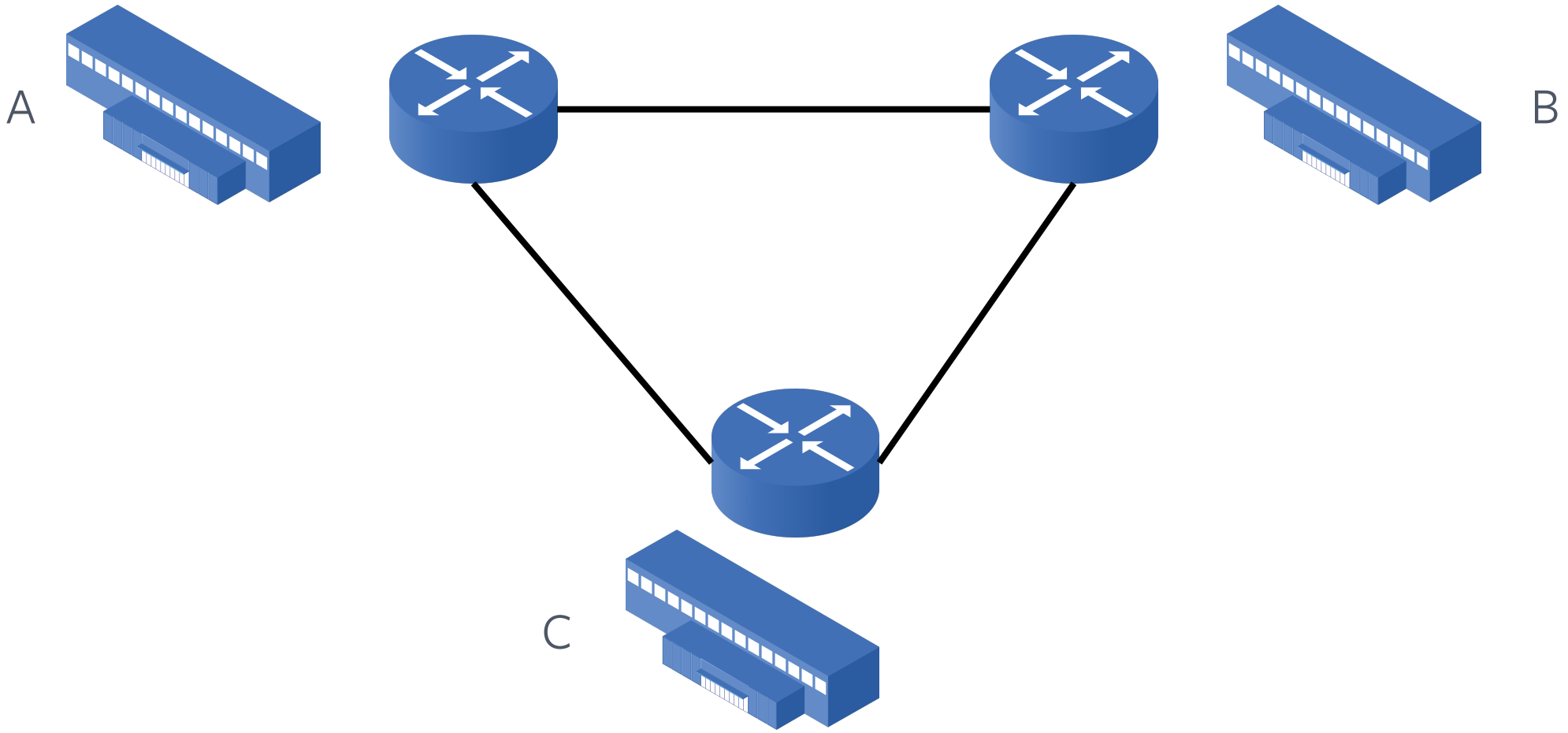
Agenda

- Edge Network
- **Backbone Network**
- Datacenter Network

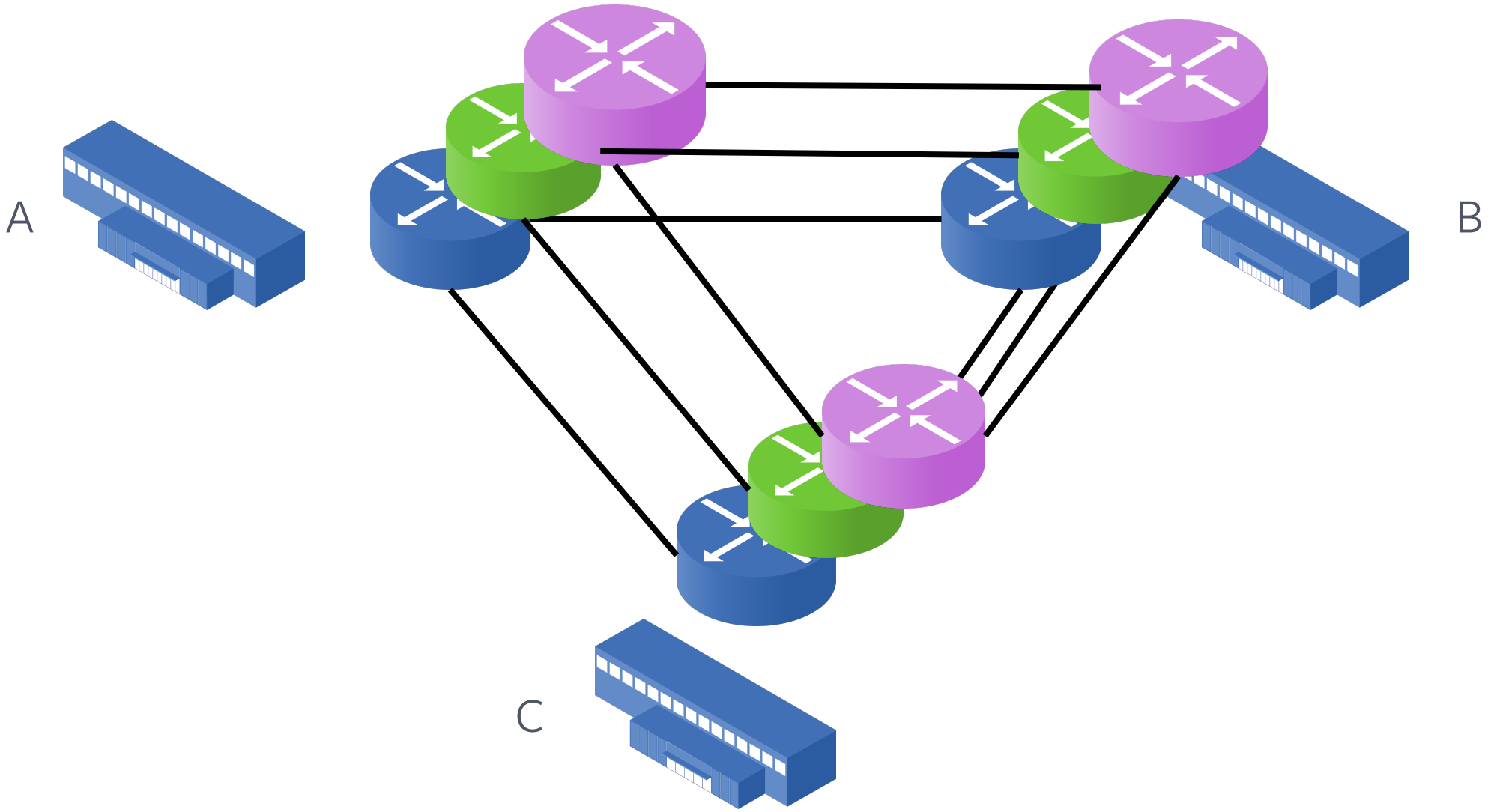
Backbones at Facebook

- Classic Backbone (CBB)
  - Connects POP and DCs
  - RSVP-TE, Vendor software solution
- Express Backbone (EBB)
  - Connects DC and DC
  - Centralized control

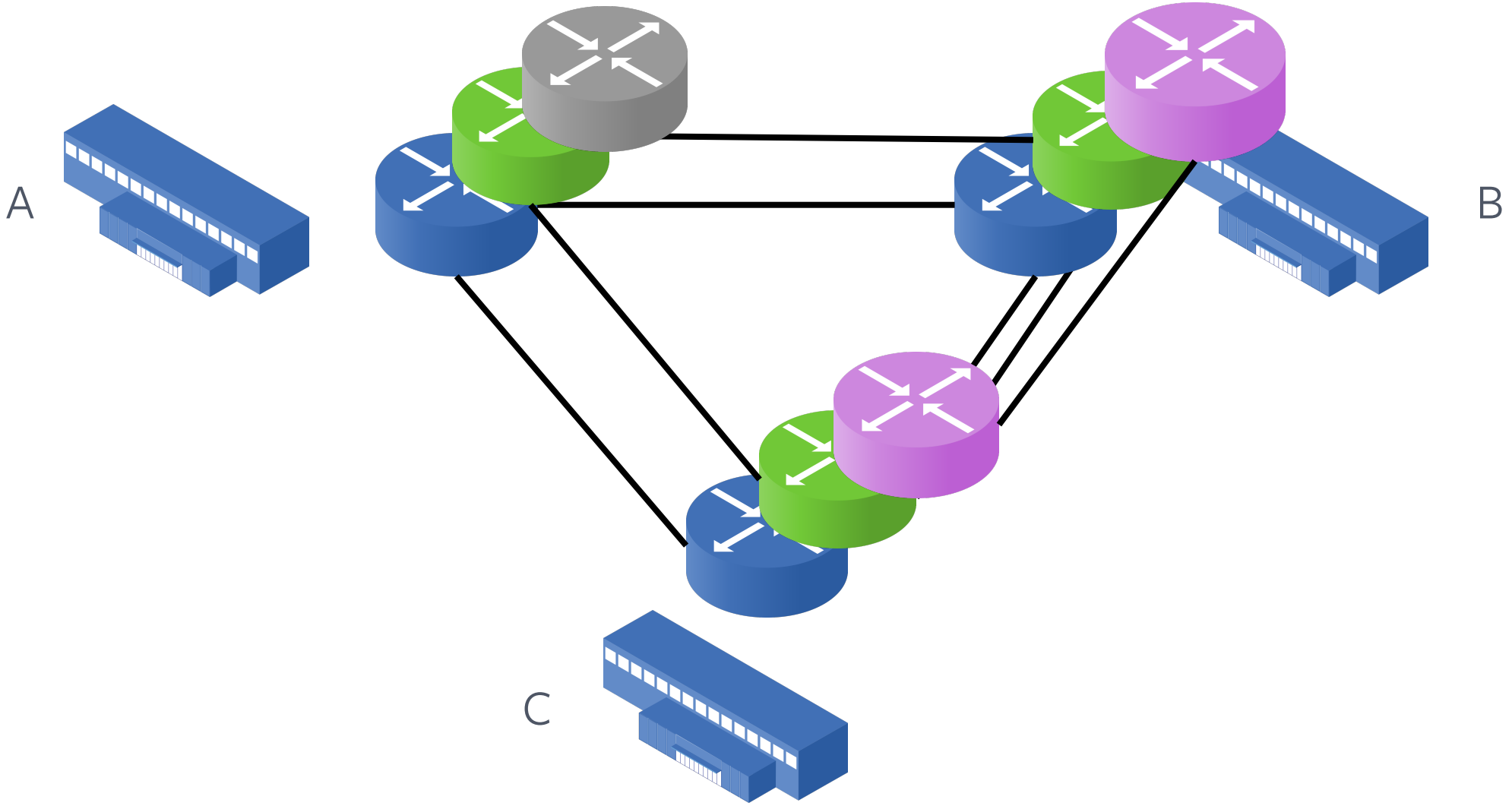
Three Datacenters



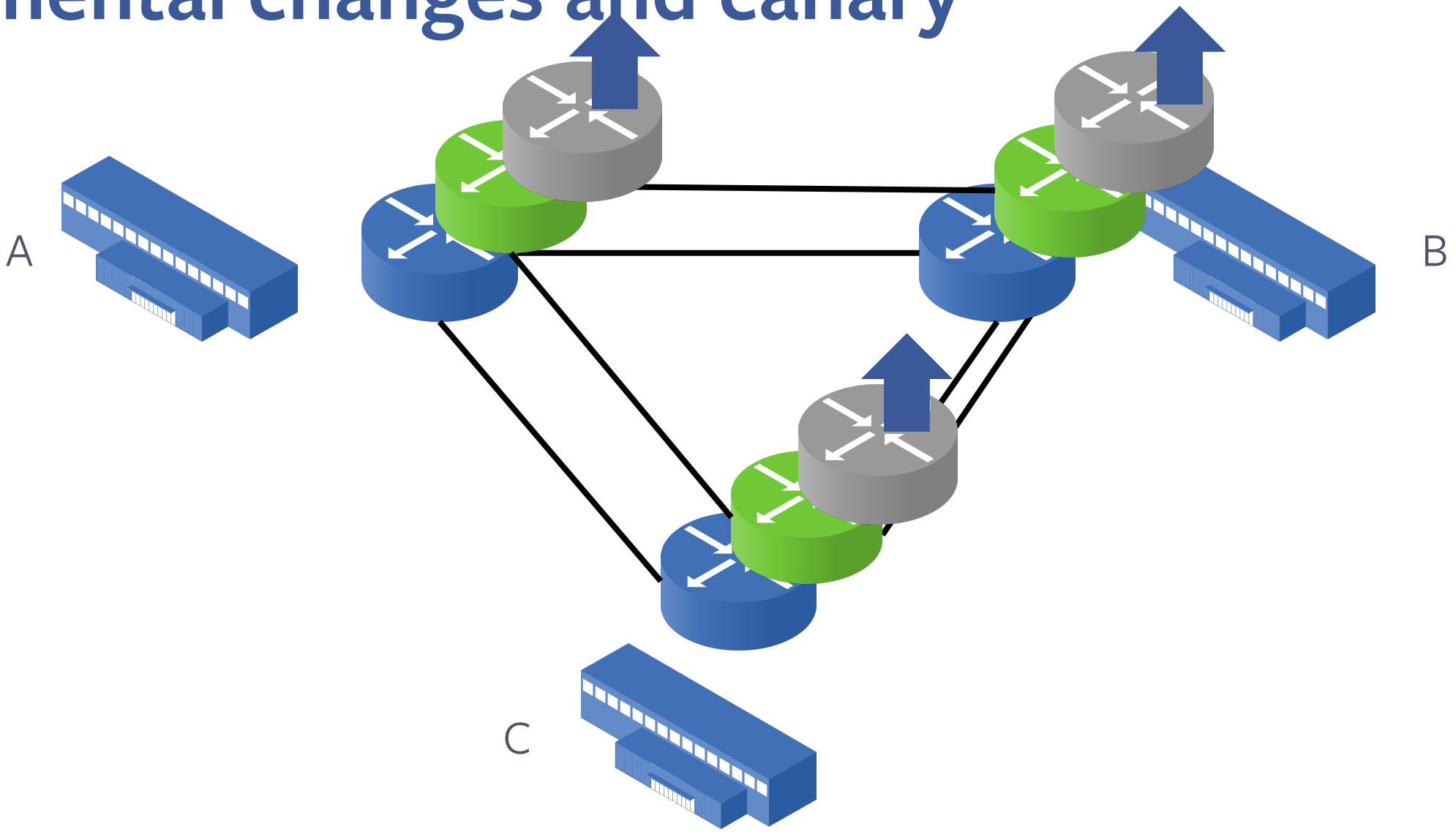
Add Planes



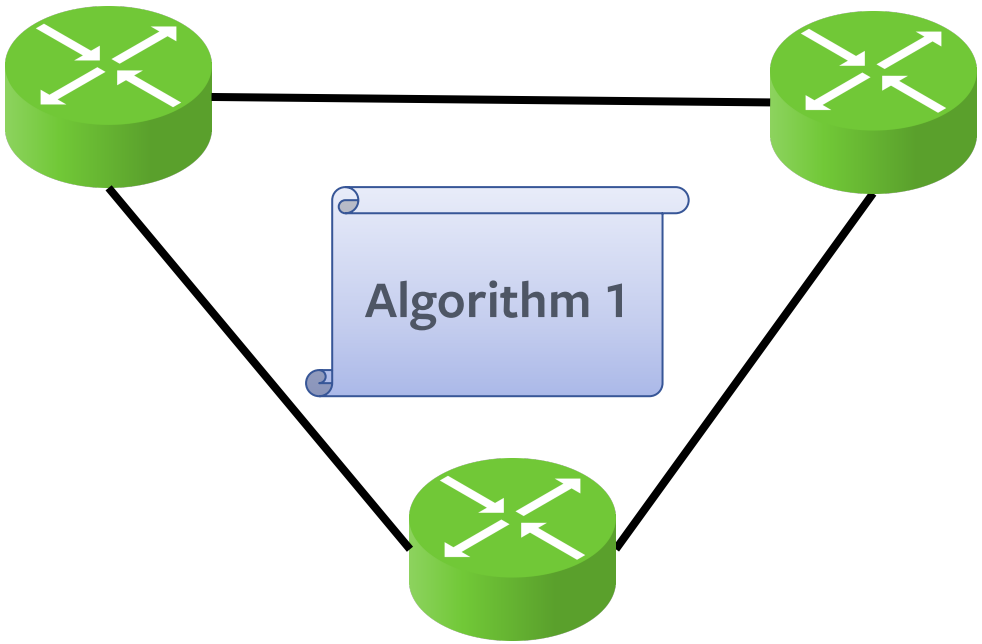
N-way Active-active Redundancy

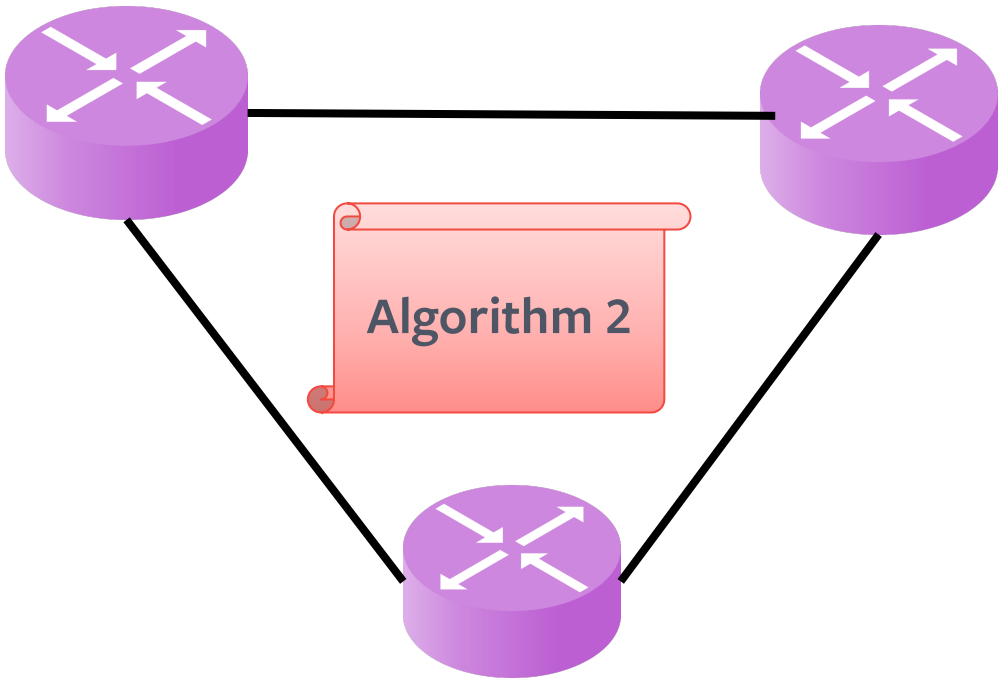


Incremental changes and canary



A/B Testing

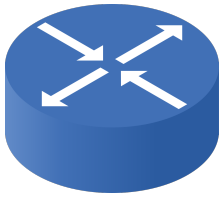




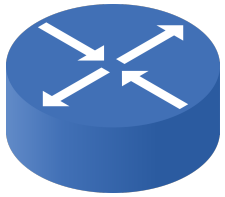
Open/R

- 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

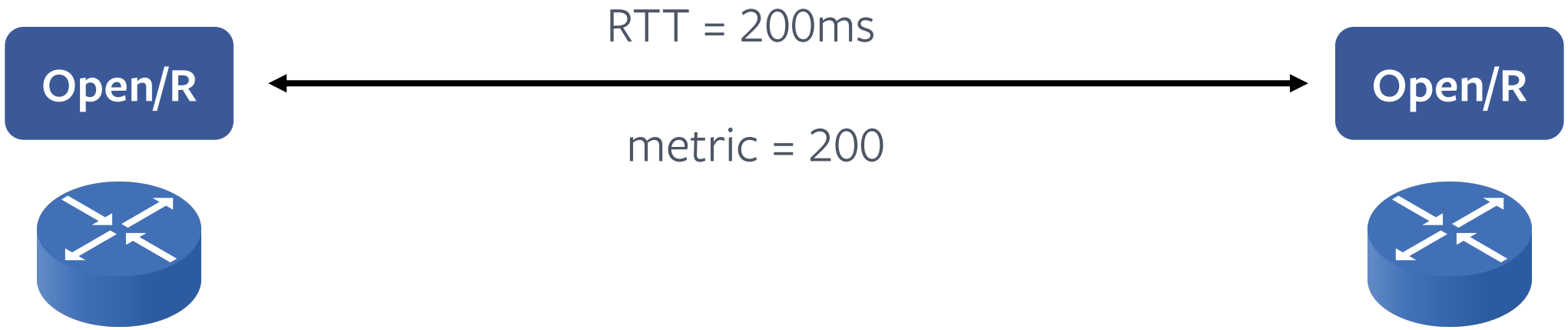
Typical IGP metric configuration



Type	Link Metric
Trans Atlantic	100
Trans Pacific	150
US-West to US-East	50



Open/R: Calculate link metric with RTT



Backbone Network Summary

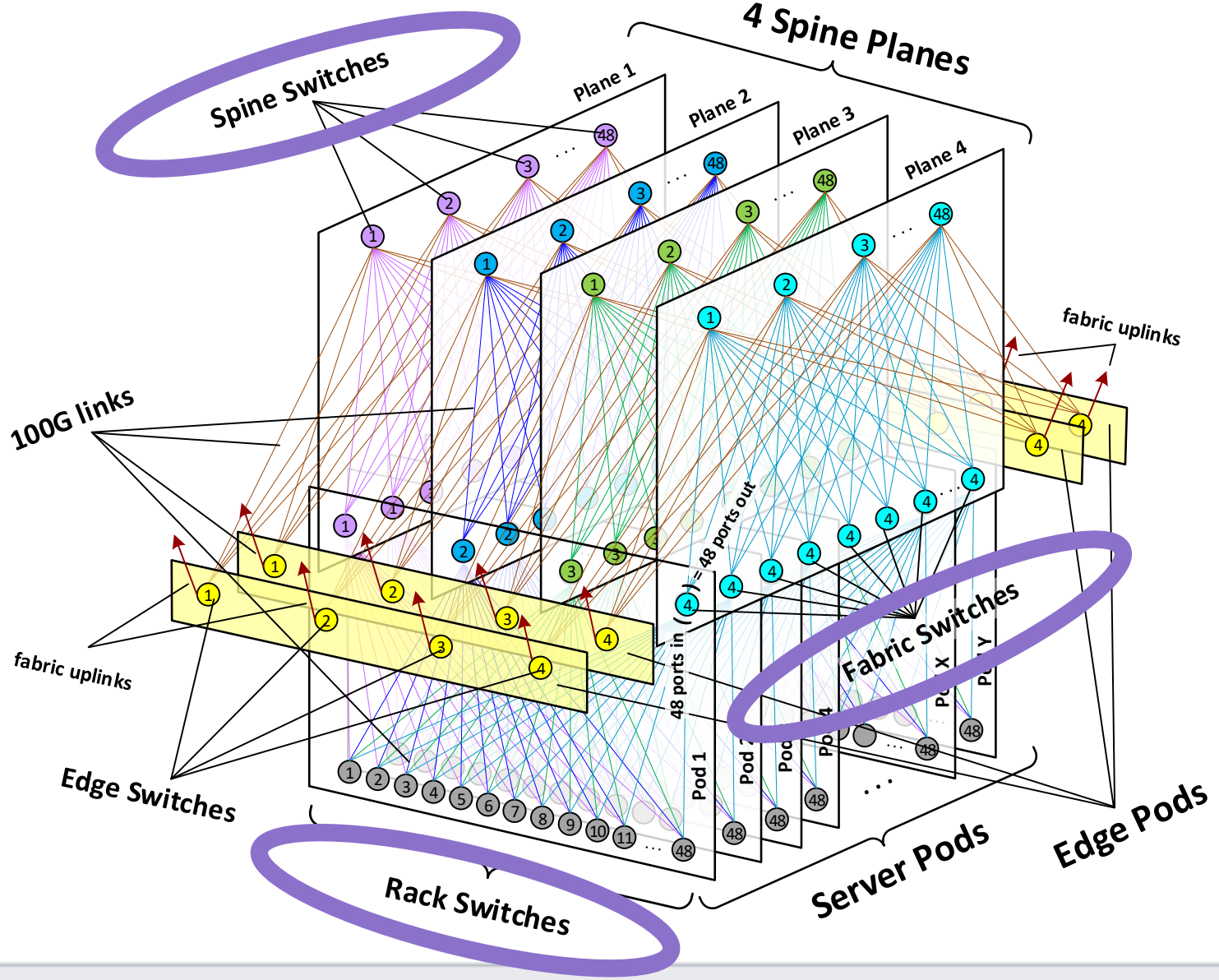
- Two backbones
  - CBB: Connects POPs and DCs
  - EBB: Inter-DC backbone
- Plane architecture
  - Reliability, maintenance, experiment
- Software
  - Centralized control
  - Innovative distributed routing protocols to minimize configuration


CATERPILLAR

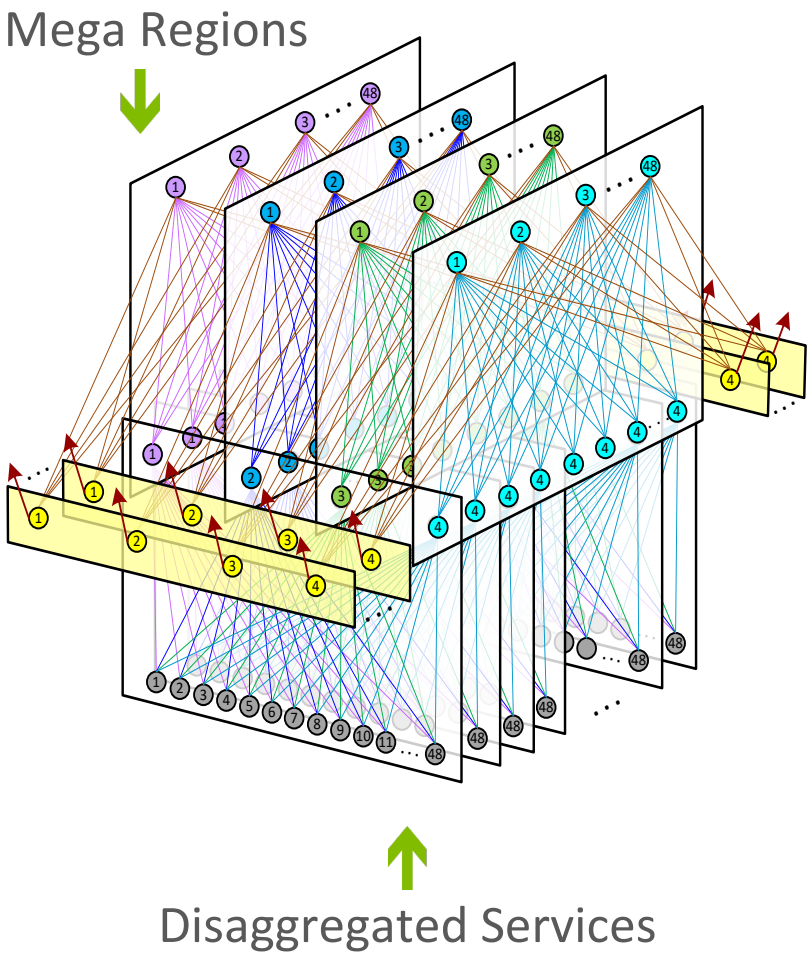
Agenda

- Edge Network
- Backbone Network
- **Datacenter Network**

Classic Facebook Fabric



Growing Pressure

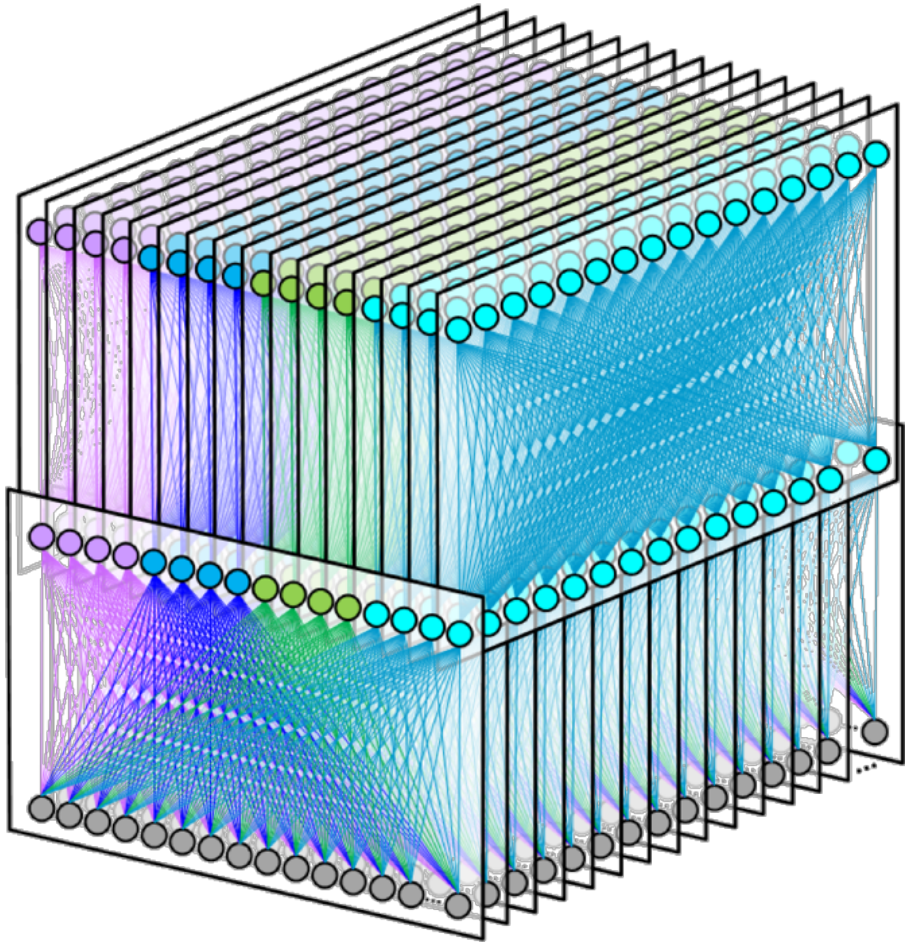


Expanding **Mega Regions** (5-6 buildings) = accelerated fabric-to-fabric East-West demand

Compute-Storage and AI disaggregation requires **Terabit** capacity **per Rack**

Both require larger fabric Spine capacity (by **2-4x**) ...

F16 – Facebook's new topology



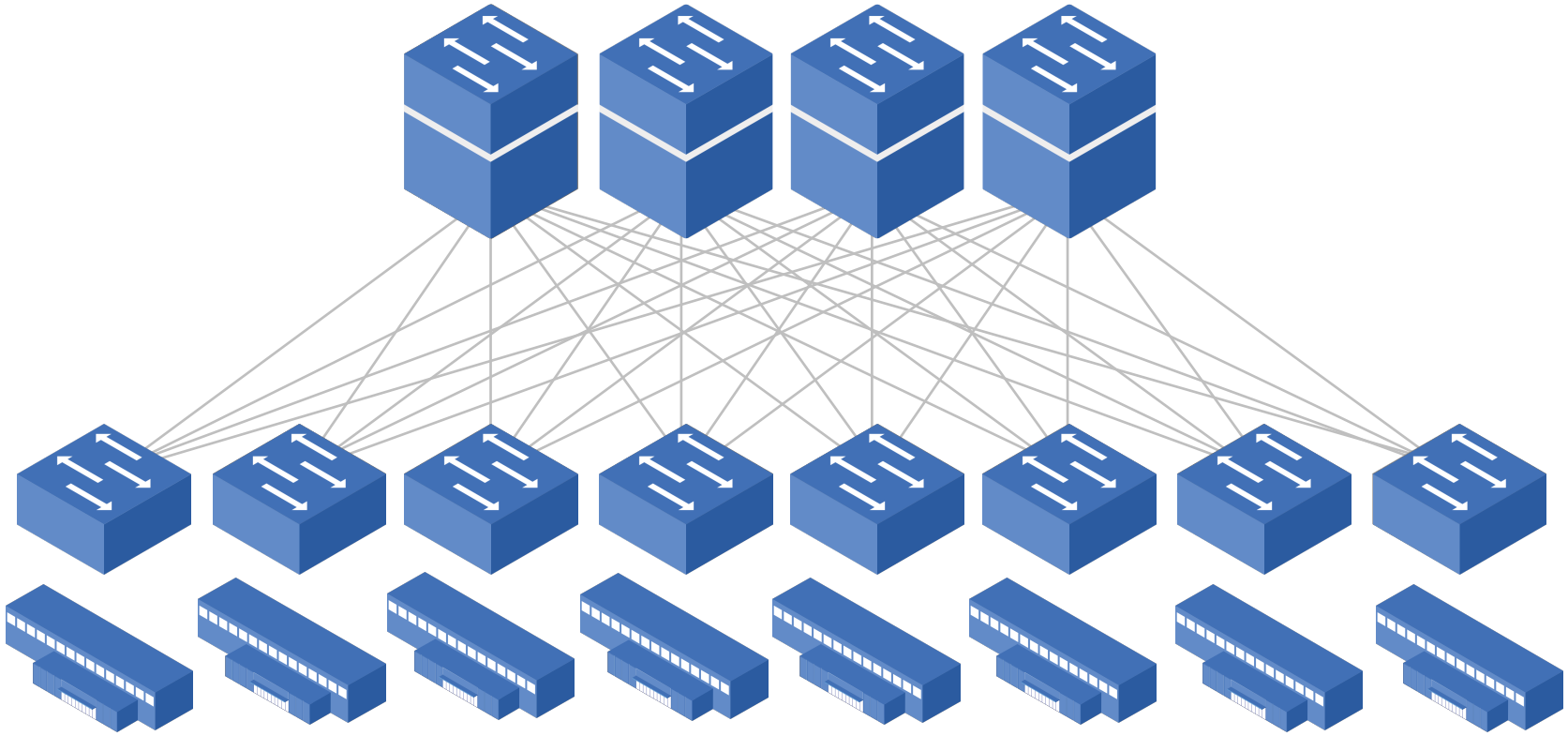
16-plane architecture

6-16x spine capacity on day 1

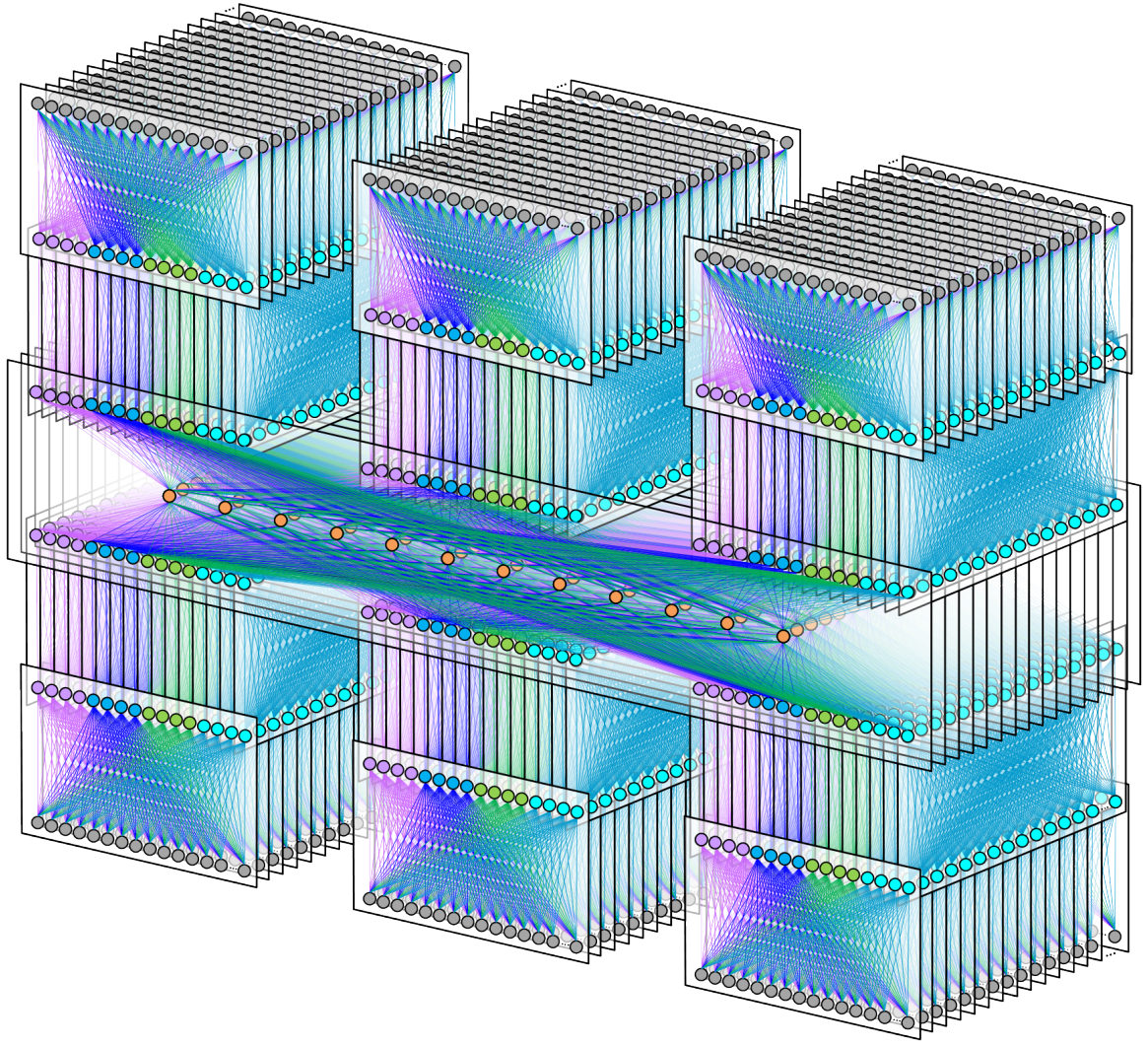
1.6T raw capacity per rack

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

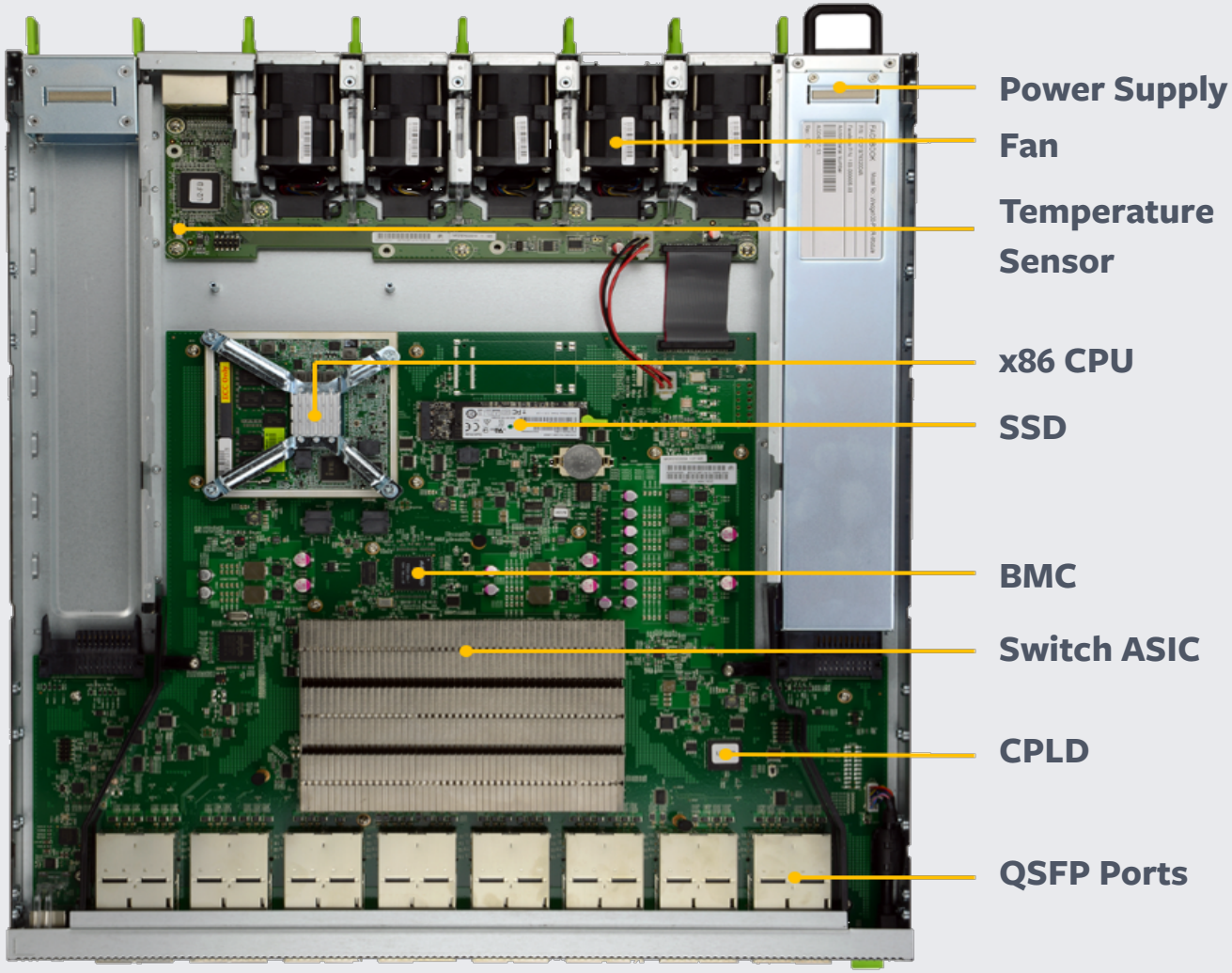
Fabric Aggregator

- Disaggregated design for scale
- Built upon smaller commodity switches

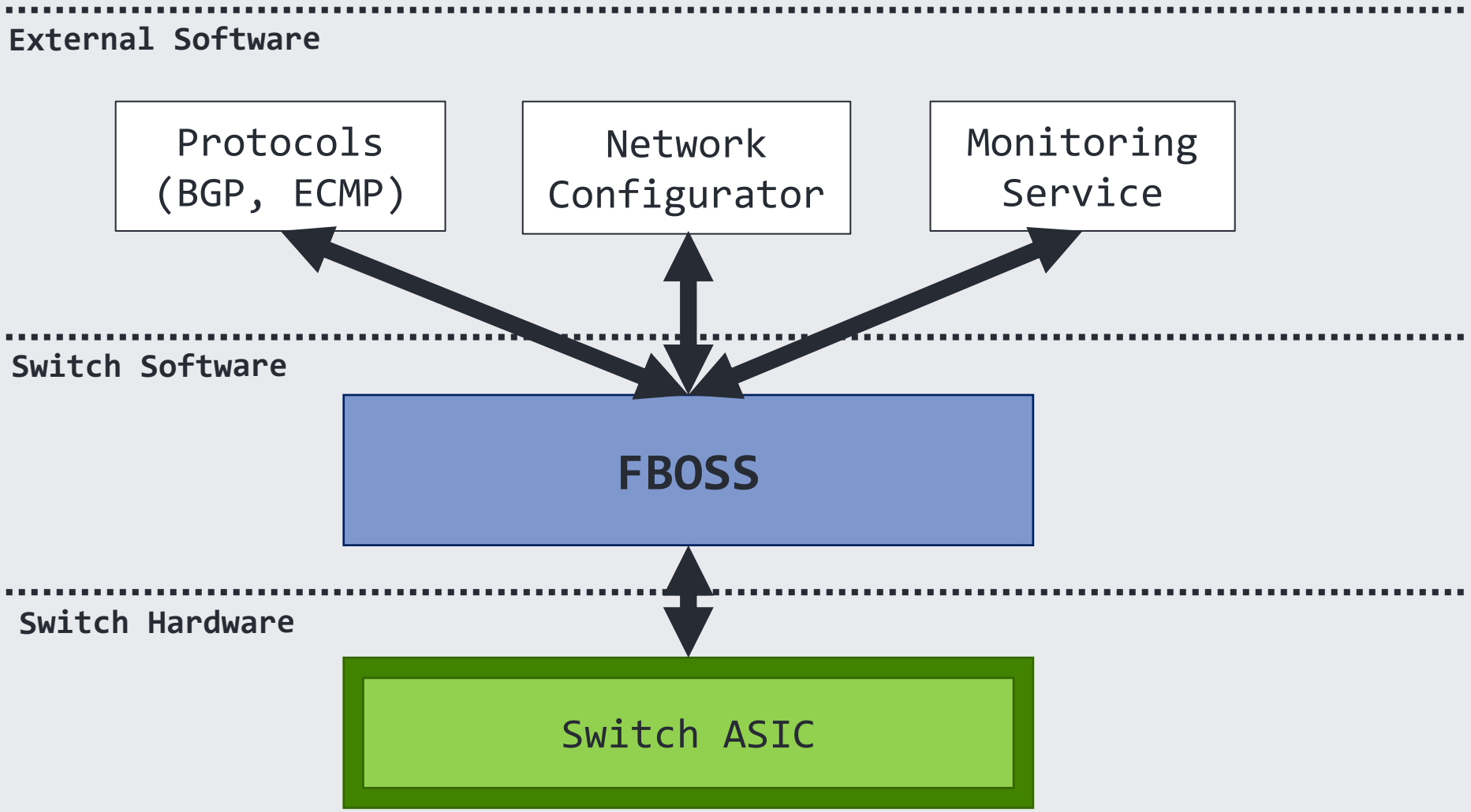
White Box Switch

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



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

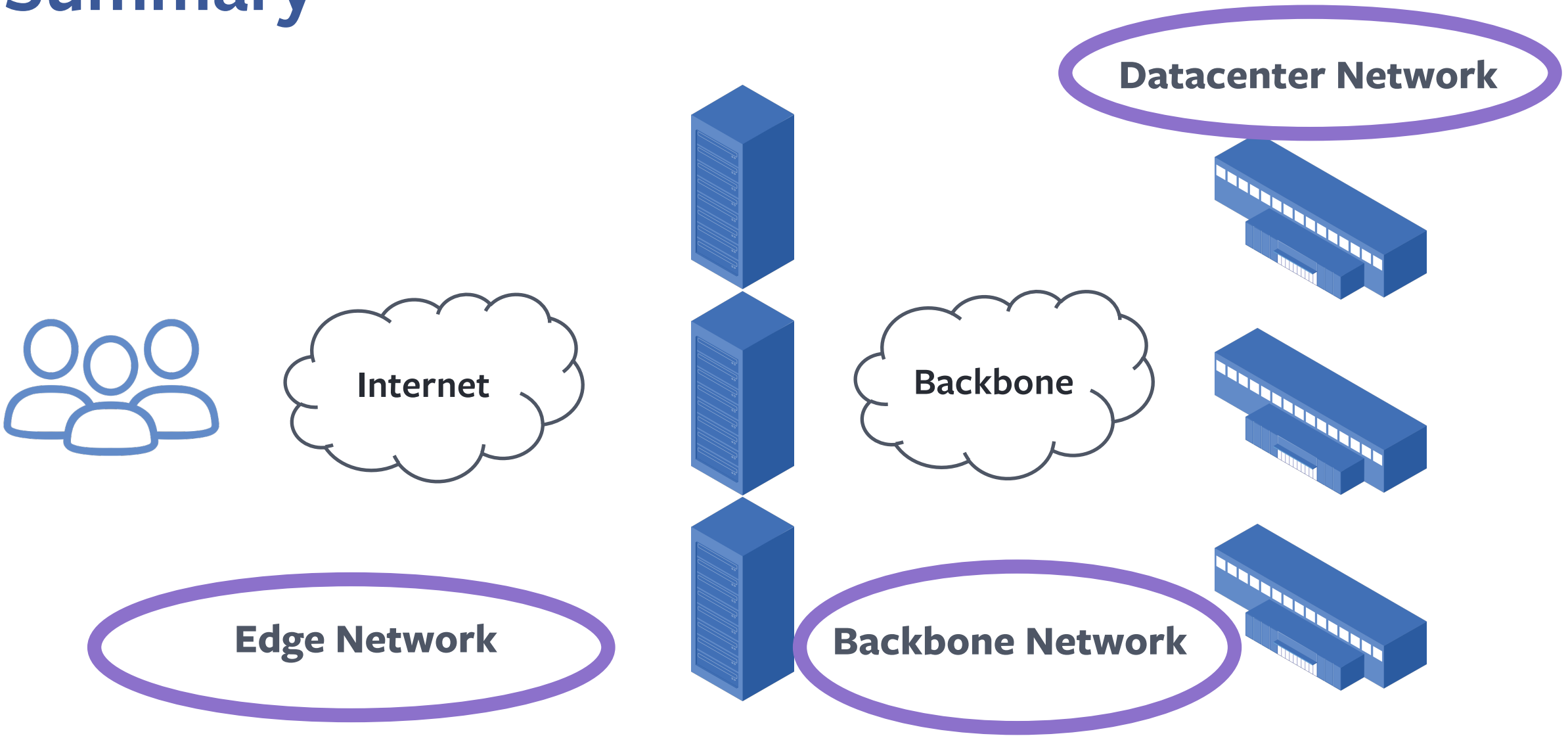
3 Stage Deployment via *fbossdeploy*

- **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

Summary


Datacenter Network
Internet
Backbone
Edge Network
Backbone Network

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

facebook