Data-center networking challenges

Terminology

Words we use often

- Rack
- Pod
- **Cluster** (logical: hundreds of racks)
- Fabric (hundreds of pods)

Aggregation layer

- Pod
- Fabric (DC)
- Region (multiple DC)
- Region attaches to same datacenter routers

Speeds and Feeds

Speeds (2015-2016)

- “Trunks”: 40G/100G
- Servers: 10G/25G/50G

Demand growth

- Storage
- Compute
- Cache
- Data Non-Locality

Over-subscription

- Rack level: fixed
- Pod level: flexible
- Cross-DC (another agg layer)
- DC <-> WAN (highest)

Utilization imbalance

- Hotspots at rack level
- Hotspots at pod level
- Small amount of hot-spots
- Average utilization low
- High cross-DC utilization

Incast/micro-bursts

- Almost none
- Except some cases
- Most often in pod switches

Routing

Main goals

- Stability
- Programmability
- IPv6 predominant
- Enable future opportunities

Right now: BGP

- Aggregation at pod layer
- Tricks to avoid black-holing
- Policies to control prefix propagation
- BGP to servers for VIP injection

BGP: features

- V4/V6 sessions
- Drain/undrain tooling
- Very low packet loss on reconvergence
- Graceful restart for FBOSS

What we want

- Maintain stability
- Keeping things simple
- Innovate fast if needed

Open/R

- Link-State
- Data-bus
- Easy to extend
- Fast convergence

FBOSS

Specifics

- BRCM silicon
- Single-chip devices (ToR) - Wedge
- Multi-chip devices (pod switches) - Backpack

Routing

- BGP
- Open/R
- Static
- Large domains with multi-chip boxes

Goals

- Topology flexibility
- Operational simplicity
- Bulk operations