BGP Routing for Large Scale Data Centers

draft-lapukhov-bgp-routing-large-dc
Agenda

Design Requirements
Network Design
Why BGP over IGP
Feature Standardization?
Design Requirements
Online Service DC Specifics

Server Perspective

100’s thousands of servers
10G NICs

Distributed Applications

Aware of the network
Explicit parallelism
Example: Web Index computation

“Network as a computer” concept
Online Services DC Specifics (cont.)

Two types of traffic flows

Query
- Latency Sensitive
- Partition/Aggregate

Background
- East/West
- Compute & Synchronize
Design Requirements

**REQ1:** Build upon a topology providing horizontal bandwidth scalability

**REQ2:** Minimize feature/protocol set

**REQ3:** Select simplest most common protocols

**REQ4:** Protocol must support traffic engineering via 3rd party next-hop
Network Design
Topology choice: Clos

Multiple definitions exist...

Has $N$ stages ($N=3,5,7,..$)
  Folded on diagram

Full bisection bandwidth if $M \geq N$

Natural link load-balancing
  ECMP Based – implements Valiant Load Balancing
Scaling Clos Topology

Think multiple parallel Clos topologies
Lower port density on switches

Horizontal capacity scaling at every layer above ToR
Routing Design for Parallel Clos

BGP all the way down to the ToR (eBGP)
Separate BGP ASN per ToR
BGP Specific: Features

Requires “BGP AS_PATH Multipath Relax”

We rely on ECMP for routing
Needed for Anycast prefixes

We use 16-bit Private BGP ASN’s ONLY

Simplifies path hiding at WAN edge (remove private AS)
Simplifies route-filtering at WAN edge (single regexp)

But we only have 1022 Private ASN’s...
BGP Specifics: Allow AS In

Reuse Private ASNs on the ToRs

Use of *Allow AS in* on ToR eBGP peerings

Effectively, ToR numbering is local to the container

Requires vendor support...
Feature Standardization
Features that would benefit standardizing

There isn’t that many requirements...

- ECMP programming
- AS_PATH Multipath Relax
- Allow AS In
- Fast eBGP Fall-over
- Remove Private AS
- Unequal-cost load-balancing
- 32-bit Private ASNs
Questions?