BGP Routing for Large Scale DCs

draft-rtgwg-bgp-routing-large-dc-00

Petr Lapukhov
Jon Mitchell
Ariff Premji

Purpose of Draft

- Document Working Design for Large Scale DC Routing using EBGP
  - Guidance to operators and implementers on design decisions and BGP behavior expectations
  - Stable reference for related work

Overview of Draft Layout

Section 1 – Introduction and Overview of draft

Section 2 – network design requirements

Section 3 – relationship of this design to other physical DC designs

Section 4 – relationship of this design to other logical DC designs

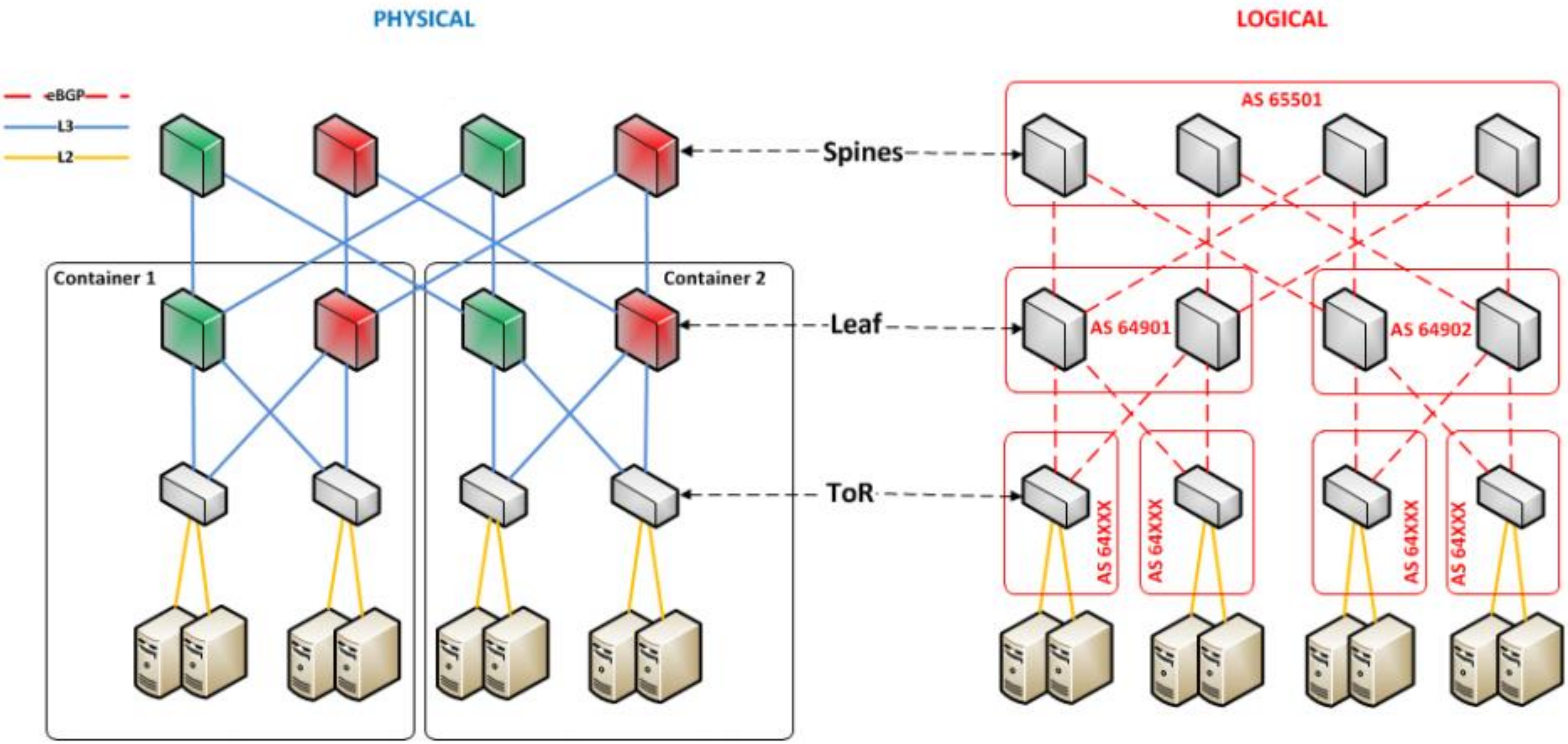
Section 5 – specifics of an EBGP only DC design

Section 6 – discusses Equal Cost Multipath (ECMP) within the design

Section 7 – describes routing convergence properties of the design

Section 8 – reviews optional attributes of the design

Simplified Physical/Logical Diagram



Draft Diagram / Terminology

```
                             Tier-1  → Spine
                            +-----+
                            |     |
                         +--|     |--+
                         |  +-----+  |
                 Tier-2  |           |   Tier-2  → Leaf
                +-----+  |  +-----+  |  +-----+
  +-------------| DEV |--+--|     |--+--|     |-------------+
  |       +-----|  C  |--+  |     |  +--|     |-----+       |
  |       |     +-----+     +-----+     +-----+     |       |
  |       |                                         |       |
  |       |     +-----+     +-----+     +-----+     |       |
  | +-----+-----| DEV |--+  |     |  +--|     |-----+-----+ |
  | |     | +---|  D  |--+--|     |--+--|     |---+ |     | |
  | |     | |   +-----+  |  +-----+  |  +-----+   | |     | |
  | |     | |            |           |            | |     | |
+-----+ +-----+          |  +-----+  |          +-----+ +-----+
| DEV | | DEV |          +--|     |--+          |     | |     |
|  A  | |  B  | Tier-3      |     |      Tier-3 |     | |     |  → TOR
+-----+ +-----+             +-----+             +-----+ +-----+
  | |     | |                                     | |     | |
  O O     O O                                     O O     O O
    Servers                                         Servers
```

Why this design?

(Section 2.5)

REQ1: Select a topology that can be scaled "horizontally" by adding more links and network devices of the same type without requiring upgrades to the network elements themselves.

REQ2: Define a narrow set of software features/protocols supported by a multitude of networking equipment vendors.

REQ3: Choose a routing protocol that has a simple implementation in terms of programming code complexity and ease of operational support.

REQ4: Minimize the failure domain of equipment or protocol issues as much as possible.

REQ5: Allow for traffic engineering, preferably via explicit control of the routing prefix next-hop using built-in protocol mechanics.

BGP Implementation Requirements

(section 5/6)

1. BGP multi-path relax – load-balancing over multiple eBGP paths from different ASNs
2. remove-private-as – useful for deploying multiple fabrics that need to communicate via non-default routing or if routes from fabric need to reach Internet directly
3. allow-as-in or 4 byte ASN support – to scale design given limitation of original number of 2 byte Private ASN’s
4. fast ebgp-fallover – feature to allow P2P eBGP session teardown w/o waiting for hold-time when corresponding connected link fails

Status of Design

- More servers in this design than legacy designs at large scale content provider
  - 100K's of servers
  - Many 10's of DC's
  - Design is proven
- Other DC operators have adopted design, some using this document as a reference
- At least one equipment supplier points to document as reference on how to build BGP based DC design

Main observations in deployments

- Automation is key for deployment of a single ASN per device design
- Most issues not related to BGP rather vendor specific RIB->FIB or Tier3 host connectivity bugs
- Care should be given before any aggregation or route information reduction techniques (such as originating default from Tier1/2 for instance or aggregating at Tier1 towards WAN) are deployed as described in the draft (section 5.4/5) to prevent black holing traffic in certain failure scenarios
  - consider impact of link, node, and physical path failures (multiple links in a fiber tray or optical systems if present to connect parts of topology)
- Modeling should be done in advance based on number of subnets required at Tier3 to understand maximal size based on equipment FIB limitations
  - consider not carrying P2P address to reduce FIB consumption for maximum sizing
  - may not be issue with latest generation of commodity chipsets
- Reduce impact of single link/node failures by fanning out horizontally number of devices Tier3 connects to

Status of Draft

- draft-lapukhov-bgp-routing-large-dc-01 presented @IETF84 IDR/GROW
  - Good feedback and interest received from participants
  - Comments incorporated
- -07 Differences
  - Added details around route aggregation options and convergence properties
  - Large restructuring in -05 for readability
    - Features that expand on / add value to design separated into section from base “spec”
    - Separated more clearly types of alternative designs
  - All comments from recent list exchange incorporated
- -07 became rtgwg-00

Summary

Intent is Informational RFC to provide long term stable reference

- i2rs potential use case draft
- segmented routing potential use case draft
- BGP SDN draft
- Future work

Authors Document is stable

- no large revisions since -05
- all technical comments received on various lists have been integrated

Ready for Last Call?