BGP-Based SPF
IETF 101, London

Keyur Patel, Arrcus
Acee Lindem, Cisco
Shawn Zandi, Linkedin
Wim Henderickx, Nokia
Derek Yeung, Arrcus
Abhay Roy, Cisco
Venu Venugopal, Cisco

Motivation

- Massively Scalable Data Centers (MSDCs) have implemented simplified layer3 routing
- Centralized route control using some controller-based solution for simplified management
- Operational simplicity has lead MSDCs to converge on BGP as their routing protocol

Motivation (Cont'd)

- Route Controller has a similar functionality as a Route Reflector
  - May Reflect Routes
  - Central Database for policy enforcements, management, etc.
- However Route Reflector (not in the forwarding path) assumes a presence of IGP that help resolve nexthop and its adjacencies for its clients
- BGP based MSDCs solve this problem by establishing hop-by-hop (in-band) peering sessions
- Proposed solution helps towards deployment of Route Controllers and yet preserve operational simplicity by using BGP
  - Route Controllers may or may not be in the forwarding path

Advantages of BGP SPF over Traditional BGP Distance Vector

- Nodes have complete view of topology
  - Ideal when BGP is used as an underlay for other BGP address families
- Only network failures (e.g., link) need be advertised vis-à-vis all routes impacted by failure.
  - Faster convergence
  - Better scaling
- SPF lends itself better to optimal path selection in Route-Reflector (RR) and controller topologies.

Advantages of BGP-Based Solution

- **Already movement toward BGP as sole MSDC protocol as evidenced by "Use BGP for Routing in Large-Scale Data Centers" work in RTGWG**
- Robust and scalable implementations exist
- Wide Acceptance – minimal learning curve
- Reliable Transport
- Guaranteed In-order Delivery
- Incremental Updates
- Incremental Updates upon session restart
- No Flooding and selective filtering
- Lends itself to multiple peering models including Route-Reflectors and controllers.

BGP based Link-State Routing

- Defined a new SAFI
  - NLRI format is exactly same as BGP LS Address Family to carry link state information
- BGP MP Capability and BGP-LS Node attribute to assure compatibility
- Multiple Peering Models
- BGP runs Dijkstra instead of Best Path Decision process

BGP Best-Path

- Next-Hop and Path Attribute basically along for the ride for BGP Link-State Address Family anyway
  - Need to be announced based on RFC 4271 error handling
- Decision Process Phases 1 and 2 replaced by SPF algorithm
- Decision Process Phase 3 may be short-circuited since NLRI is unique per BGP speaker.
- Need to assure the most recent version of NLRI is always used and re-advertised.
  - Augmented with sequence numbers

BGP SPF

- Starting with greatly simplified SPF with P2P only links in single area (i.e., SPT)
- Will scale very well to many use cases.
- Could support computation of LFAs, Segment Routing SIDs, and other IGP features.
  - BGP-LS format includes necessary Link-State
- Link-State AF is dual-stack AF since both IPv4 and IPv6 addresses/prefixes advertised
  - BGP-LS format also supports VPNs but SPF behavior not defined.
  - Work needed to define interaction with existing unicast AFs.
    - Matter of local implementation policy

BGP SPF Convergence Mechanisms

- Local NLRI changes always advertised before any attendant best-path changes
  - Implementation recommendation
- BGP session outage normally results in withdrawal of all NLRI
  - Investigate ways to dampen this behavior
  - Variation on graceful restart where is marked stale before withdrawn
  - Jury out as to whether we want to change BGP to be more like IGPs – May not be required.

BGP SPF AFI/SAFI Interaction

- BGP SPF used for IPv4/IPv6 unicast underlay
- Legacy BGP also used for IPv4/IPv6 unicast
- No implicit route leaking between legacy BGP and BGP SPF
  - Normal implementation policy applies
- Recommend that default route preference (aka, Admin Distance) for BGP SPF is preferred over legacy BGP

Peering Model

- BGP sessions with Route-Reflector or controller hierarchy.
  - Link discovery/liveliness detection outside of BGP.
- RR hierarchy can be less than fully connected but must provide redundancy
  - Must not be dependent on SPF for connectivity
- Controller could learn the expected topology through some other means and inject it.
  - SPF Computation is distributed though.
  - Similar to “Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google’s Datacenter Network”
- BGP SPF applicability draft covers in detail

Next Steps

- Further discussion
- Consider Draft adoption