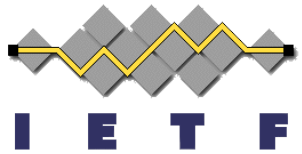
BGP-Based SPF
RTGWG - Jan 2017

Keyur Patel, Arrcus
Acee Lindem, Cisco
Shawn Zandi, Linkedin
Gunter Van de Velde, Nokia
Derek Yeung, Arrcus
Abhay Roy, Cisco
Venu Venugopal, Cisco



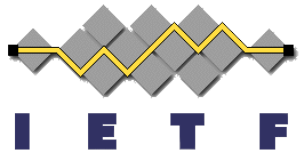
Motivation

- Massively Scalable Data Centers (MSDCs) have implemented simplified layer3 routing
- Operational simplicity has lead MSDCs to converge on BGP as their routing protocol
  - Has support to announce both the underlay routes as well as the overlay routes using separate AFI/SAFIs
- Centralized route control using some controller-based solution for simplified management

Motivation (Cont'd)

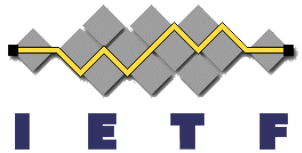
- Route Controller has a similar functionality as a Route Reflector
  - May Reflect Routes
  - Central Database for policy enforcements, management, etc.
- However Route Reflector (not in the forwarding path) assumes a presence of an IGP protocol that help resolve bgp nexthop and its adjacencies for its clients
  - IGP's SPF builds directed graphs that ultimately resolves non connected nexthops
- BGP based MSDCs solves this problem by establishing hop-by-hop (in-band) peering sessions
  - Nexthop is always an address of a directly connected switch
- Proposed solution helps towards deployment of Route Controllers and yet preserve operational simplicity by using BGP for underlay as well as the overlay routing

Advantages of BGP SPF over Traditional BGP Distance Vector



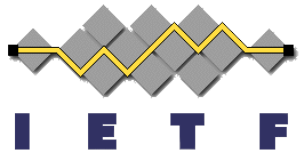
- Nodes have complete view of topology
  - Ideal when BGP is used as an underlay for other BGP address families
- Only network failures (e.g., link) need be advertised vis-à-vis all routes impacted by failure.
  - Faster convergence
  - Better scaling
- SPF lends itself better to optimal path selection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

Advantages of BGP-Based Solution



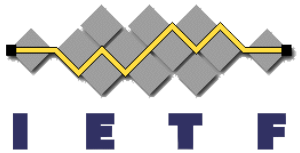
- **Already movement toward BGP as sole MSDC protocol as evidenced by "Use BGP for Routing in Large-Scale Data Centers" work in RTGWG**
- Robust and scalable implementations exist
- Wide Acceptance – minimal learning curve
- Reliable Transport
- Guaranteed In-order Delivery
- Incremental Updates
- Incremental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

Advantages of BGP-Based Controller Solution



- Sparse BGP session topology with session paths for redundancy. Two cases:
  - Control channel to every node.
  - Hierarchy of Route-Reflectors which are directly or loopback connected (with static routes for the latter)
- BGP SPF used for underlay with other BGP AFs used for overlay services
  - Faster convergence of underlay and services in BGP
- Controller also used for end-to-end traffic engineering paths possibly spanning other networks (e.g., core or access).

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.
  - Assured by existing protocol mechanisms

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

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”

Next Steps

- Further discussion
- Collaboration
- Consider Draft adoption