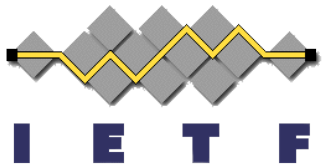
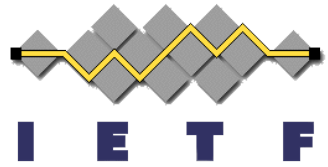
BGP-Based SPF
IETF 96, Berlin

Keyur Patel, Cisco
Acee Lindem, Cisco
Derek Yeung, Cisco
Abhay Roy, Cisco
Venu Venugopal, Cisco

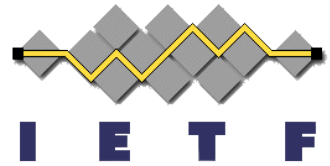


Data Center Routing Routing Problem Space



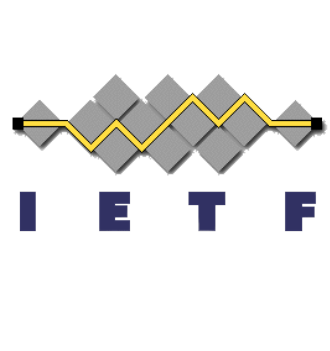
- Routing scaling for Massively Scalable Data Centers (MSDCs)
- Support both traditional and controller-based solutions
- Many MSDCs migrating towards layer 3 only solutions for simplified management

Advantages of BGP-Based Solution



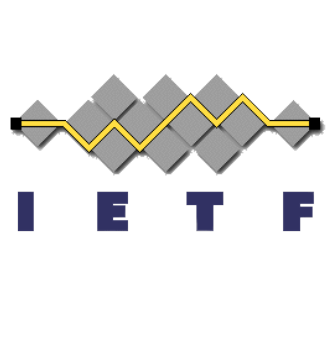
- Robust and scalable implementations exist
- Wide Acceptance – minimal learning curve
- Already movement toward BGP as sole MSDC protocol as evidenced by “Use BGP for Routing in Large-Scale Data Centers” work in RTGWG
- Reliable Transport
- Guaranteed In-order Delivery
- Incremental Updates
- Incremental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

Advantages of BGP SPF over Traditional BGP Distance Vector



- Nodes have complete view of topology
- Only network failures (e.g., link) need be advertised vis-à-vis all routes impacted by failure.
  - Faster covergence
  - Better scaling
- SPF lends itself better to optimal path selection in Route-Reflector (RR) and controller topologies.

BGP based Link-State Routing



- Build on BGP LS Address Family to carry link state information
- BGP Capability and BGP-LS Node attributed to assure compatibility
- Multiple Peering Models
- BGP runs Dijkstra instead of Best Path

BGP-LS Usage

- Uses existing BGP-LS Encodings from RFC 7752 and subsequent drafts.
  - Completely incremental updates
  - Considered new BGP-SPF Protocol-ID but this would limit usage to SPF. It would allow implicit policy to prevent advertisement to peers that don't support BGP SPF.
- Node NLRI
  - Carries a new BGP-LS node attribute to indicate BGP SPF capability and specify the algorithm
- Link NLRI
  - Identifies links with advertisement dependent on peering model
  - Dual stack advertisement of IPv6/IPv4 addresses
  - Unnumbered interfaces use local/remote identifier (so far no BGP extension for remote identifier discovery)
- Prefix NLRI – Advertise connected prefixes or even prefixes imported from other AFs or protocols

BGP Capability

- New capability indicating support for the BGP-LS SPF computation
- Requires support for MP Address-Family Link-State
- Used to determine whether to include link in topology.

BGP Best-Path

- Next-Path and Path Attribute basically along for the ride for BGP Link-State Address Family anyway
  - Need to be validated 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.
  - More work required here.

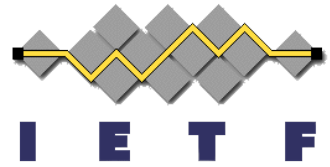
BGP SPF

- 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 includes necessary Link-State
- Link-State AF is dual-stack AF since both IPv4 and IPv6 addresses/prefixes advertised
  - BGP-LS also supports VPNs but SPF behavior not defined (at least not yet).
  - Work needed to define interaction with existing unicast AFs.

Peering Models

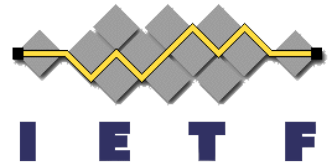
- BGP run on all inter-fabric connections
  - Similar to RTGWG BGP Data Center solution
  - Assures all nodes in SPT support SPF capability
  - May want to wait for End-of-RIB (RFC 4724) to advertise link to avoid black holes.
- BGP run between all inter-fabric BGP speakers
  - Single session when multiple links – can be multi-hop.
  - Links/Liveliness discovered outside BGP.

Peering Models (Continued)



- BGP sessions with Route-Reflector or controller hierarchy.
  - Link discovery/liveliness detection outside of BGP.
- Less control over avoiding links until new node has complete topology.
- 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



- More refinement of mechanisms
  - Changes to Best-Path selection for Link-State Address Family
  - Interactions with other Address Families
- Further discussion
- Collaboration