RIFT: Routing in Fat Trees Motivation, Additional Requirements and Use Cases in User Access Networks

Yan Filyurin

Bloomberg LP

IETF 101, London

Actual Draft Status

- Designed to be informational and help evolve RIFT.
- Submitted, but ultimately waiting for other interested parties.
- Designed to promote RIFT for use in general access networks.
- As well as mixed use DC/User Access setups found in a traditional enterprise campus network.
- A call for certain vendors to consider RIFT in their integrated fabric solutions.
- A call for Open Source projects to start developing RIFT for access networks.
- Includes various Open Source routing software or integrated in projects like ODL or ONOS.

“Enterprise” Motivations for RIFT

- Enterprise is just a marketing term, to attract the right sales people.
- Network environment for end user devices and their services.
- No major distinction between campus or multitenant access network.
- And not to take away from the Data Center use case.
- Even some potential for DCs with multiple security zones.
- So why campus/access?
  - It is a much harder problem and frequently ignored.
  - Core/Distribution/Access = Superspine/Nodes (N Levels)/Leafs.
  - Mobility and security requirements difficult to address.
  - Often addressed by scaling broadcast domains.
  - Which makes it tough to horizontally scale.
  - More capabilities (control plane, forwarding operations) at higher tiers. A lot less at lower tiers.
- Decoupling of reachability from prefix information makes it possible to advertise any information (like if we do scale broadcast domains).
- Operation over unnumbered networks
- K/Vs and Policy Guided Prefixes simplify management and create framework for separate connectivity domains.

Desired RIFT Capabilities: “Network Slicing”

- Network Slicing – Ability to create virtual private routed networks within our RIFT access network.
- Incoming packets associated with a slice at the UNI interface, distinctly identified at NNI and associated with the proper connectivity domain and right UNI at egress.
- Instead of Transport/Service IP VPN model the goal is to follow multi-instance model.
- Plus our Edge/Core model is reversed.
- Start with Auto-discovery:
  - Configure slice components on the edges (leafs, hosts, etc.)
  - Flood K/V TIEs Northbound to auto-configure instance ID. (possibly borrow from RFC8196)
  - If there is “route” Southbound, create network instance.
  - What is installed dependent on protocol and default origination.
- Give lower tier nodes the option of explicitly requesting default origination.
- Default (aggregation) breaks all forms of leaf to leaf tunnels, shim encapsulations and any other network virtualization technology.

RIFT Network Slicing Control Plane Control and Data Planes

- Establish separate adjacencies for each instance.
- Separate Link Information Exchange for each instance.
- If using separate link monitoring protocol, utilize a single protocol session to notify every adjacency.
- Use negotiated UDP ports to establish and maintain.
- Perform standard TIE exchange for each instance or slice.
- Carry optional instance ID as part of Prefix TIEs.
- Separate N-SPF and S-SPF for each instance.
- Install in FIBs based on AD process.
- Use any tunneling/encapsulation technology, as long as point of default origination has decap-route-encap bidirectional capabilities.
- Nodes should be allowed to flood their capabilities to determine nodes acting as route aggregation point. Default aggregators can be explicitly configured.
- Lower level nodes nodes do not populate RIBs and propagate advertisements.
- Leverage PGPs for policy control.

RIFT Network Slicing Control Plane Variations

- No per instance adjacency.
- Single topology, instance IDs part of Prefix TIE
- No per instance Auto Discovery and ID establishment.
- All devices are part of the same consensus group
- Prefix TIEs carry instance ID and optional parameters specifying tunneling, encapsulation, SR path, etc.
- Whether default is originated or all Prefix TIEs are propagated in the Southbound, they are propagated to all leafs.
- Whether a particular leaf hosts a certain instance or not.
- Also use PGP communities in propagating the prefix.

External Routing Information From End Systems

- Leafs and network boundaries will run other routing/information exchange protocols with upstream and downstream systems
- There may be a need to flag those Prefix TIEs, if we need to differentiate between them.
- Such as if we ever have to propagate them Southbound and we must be sure they can not cause a routing loop.
- Some of those may be workload specific
- Same network layer addresses are given to different workloads and move around the network.
- Mobile workloads – the same workload and its address moves around the network.
- Ability to do “purge” Prefix TIE in Southbound direction.
- After nodes are flood it Northbound.
- Consider carrying a special “mobile” flag in Prefix TIEs.
- Flag to keep a route in the RIB and only remove from FIB and reinstall later.
- Becomes a sort of caching system, but now we have to worry about expiration.
- Aggregated prefix TIE when flooding Northbound.

External Connectivity and Superspine Interconnectivity

- Many Data Centers may chose to deploy external connectivity off leaf nodes.
- Should be treated no differently than any other external route.
- Default can be flooded Northbound.
- Some people will deploy a set of special Border Nodes off Superspines.
- Or Superspines will act as border nodes.
- Superspines and Border Nodes can form their own flooding domain.
- Northbound flooding becomes E/W flooding.
- This is NOT a requirement to turn Superspines into a backbone.
- Distinct Fat Trees or RIFT domains must rely on a more traditional backbone protocol to interconnect.

External Connectivity and Superspine Interconnectivity

- Many Data Centers may chose to deploy external connectivity off leaf nodes.
- Should be treated no differently than any other external route.
- Default can be flooded Northbound.
- Some people will deploy a set of special Border Nodes off Superspines.
- Or Superspines will act as border nodes.
- Superspines and Border Nodes can form their own flooding domain.
- Northbound flooding becomes E/W flooding.
- This is NOT a requirement to turn interconnected Superspines into a backbone interconnecting distinct node domains.
- Distinct Fat Trees or RIFT domains must rely on a more traditional backbone protocol to interconnect.

“Daisy-Chained” Leaf Nodes

- Very much access network use case.
- Unlikely to ever happen in a data center.
- Two rightmost and leftmost leaf nodes connect to level 1 nodes.
- Setup in fiber constrained campus environments.
- Reverse of Superspine use case, except in Southbound direction .
- Deviation from original RIFT Spec – North TIEs in E/W direction.
- Leafs still run S-SPF only.
- S-TIEs go E/W in both directions of the daisy chain.
- No need for S-TIEs get reflected back Northbound to insure no disaggregation loop.
- As no S-TIEs can ever be propagated if already learned from leaf.
- Break in a daisy must force relatively quick re-convergence.
- Utilize “purge” S-TIE in both direction of the break to withdraw stale routes.
- Standard N-TIEs force upstream nodes to run N-SPF.
- Purge forces all leafs to rerun S-SPF and reroute from the break.

Security Considerations

- All the typical ones.
- Neighbor discovery addressed by Secure-ND (RFC6494)
- RIFT Migration towards the use of QUIC will make it encrypted.
- Secure adjacency establishment.
- RIFT makes it very easy for leafs to join the network, whether they are DC compute aggregation devices or hosts themselves.
- Consideration for new leafs to be registered and manually authenticated.
- Leaf Prefix TIEs (outside on in-band management) become valid only after they are properly signed.