Fully Adaptive Routing Ethernet using BGP draft-xu-idr-fare-05

Xiaohu Xu@China Mobile
Shraddha Hegde@HPE
Keyur Patel@Arrcus
Zongying He@Broadcom
Junjie Wang@Centec
Hongyi Huang@Huawei

Qingliang Zhang@H3C
Hang Wu@Ruijie Networks
Yadong Liu@Tencent
Yinben Xia@Tencent
Peilong Wang@Baidu
Tiezheng Li@IEIT SYSTEMS
Roman Glebov@Yandex

IETF126, Vienna

Changes Since -04

- Add a new co-author: Roman Glebov@Yandex.
- Add clarifications for the following two scenarios:
  - Multiple BGP sessions between a pair of switches.
  - Packet forwarding over tunnels.

Multiple BGP Sessions Per Switch Pair

- The document mainly assumes one BGP session per switch pair (e.g., leaf–spine pair), with peering addresses sourced from trunk interfaces or loopback interfaces.
  - BGP neighbor discovery mechanism as defined in [I-D.xu-idr-neighbor-autodiscovery] is recommended, especially when loopback addresses are used for BGP session establishment.
- When multiple BGP sessions exist between a switch pair (one session per parallel physical link), path-bandwidth calculations **MUST operate at node granularity, not session granularity**.
  - Path bandwidth = min {Aggregate bandwidth of all parallel physical links toward the advertising node, Path-bandwidth value advertised in the route updates}
  - Two options to allocate the path bandwidth weight across ECMP routes:
    - Distribute the total path bandwidth evenly across all ECMP paths leading to the advertising node.
    - Prune a subset of ECMP paths toward the advertising node such that the aggregate link bandwidth of the remaining ECMP paths matches the computed path bandwidth.

Packet Forwarding over Tunnels

- The node originating a prefix (such as a leaf node) attaches the path bandwidth extended community carrying a maximum bandwidth value to the route.
- The advertised IP prefix can be bound to either a directly attached physical interface or a loopback interface.
- When tunneling is used for packet forwarding, loopback addresses typically serve as tunnel endpoints for service routes (including VPN routes and public routes for attached hosts):
  - WECMP load-balancing logic is applied to tunneled packets.

Implementation Status

- The FARE protocol has been implemented by at least **2 cloud providers including China Mobile Cloud, Baidu Cloud and 6 network vendors, including Ruijie, IEIT System, H3C, Accton, EdgeCore and Cloudnine.**
- It has also been successfully adapted to more than 6 mainstream switch chips, including but not limited to:
  - Broadcom Tomahawk 5 (51.2T) & Tomahawk6 (102.4T);
  - Nvidia Spectrum-4 (51.2T);
  - Huawei HiSilicon SD58XX(51.2T);
  - Centec Arctic (25.6T&12.8T);
  - Clounix CLX86 (25.6T)

Next Step

- The document is ready for WG adoption call.