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

Xiaohu Xu@China Mobile
Shraddha Hegde@Juniper
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

IETF123, Madrid

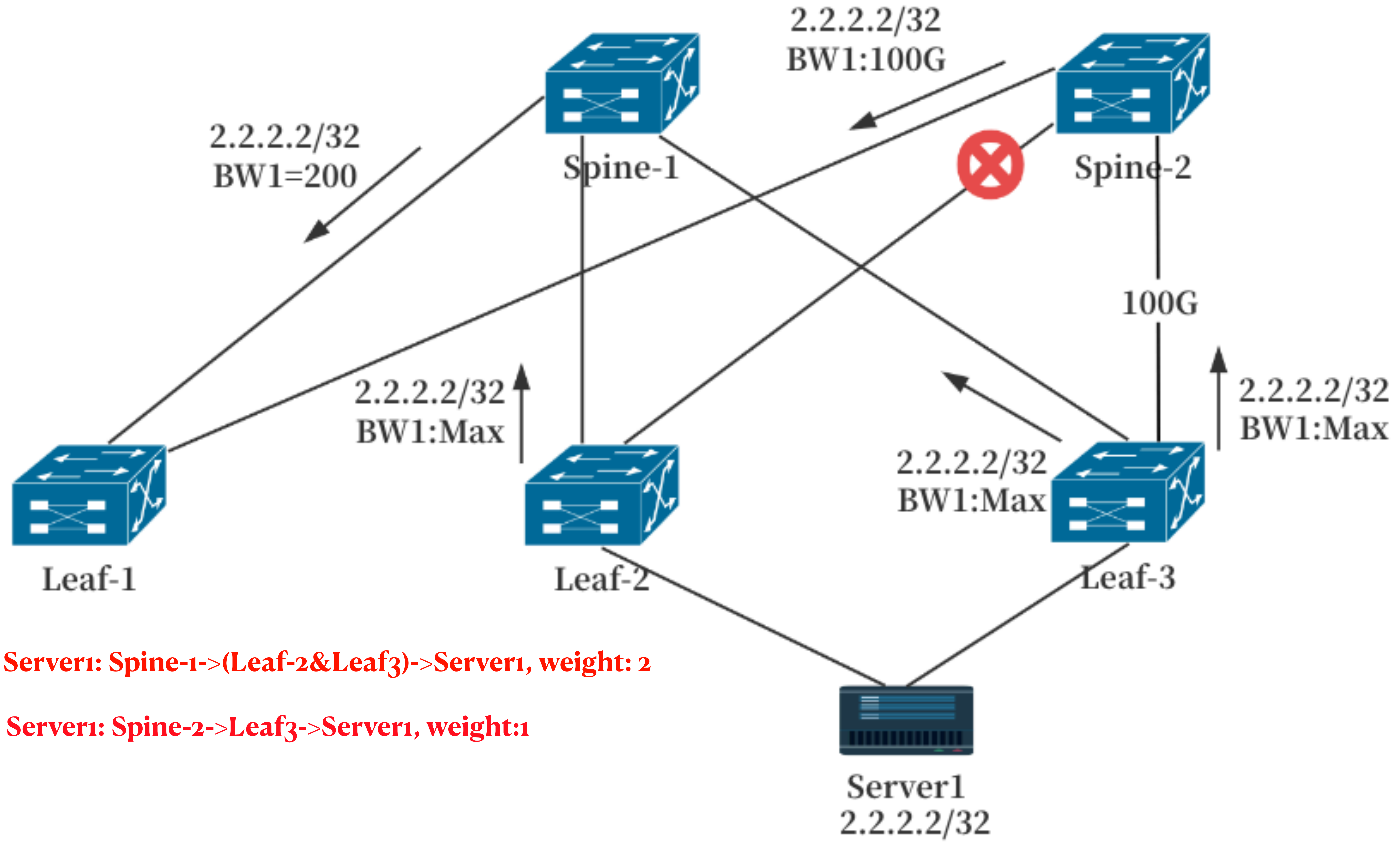
Backgrounds

- 10K-GPU AI training clusters using 3-stage or 5-stage CLOS networks become normal.
- Network throughput is much critical for AI training .
- Characteristics of AI collective-communication traffic： low-entropy， elephant-flows and burstiness.
- Static ECMP load-balancing would result in high probability of collision.
- Adaptive routing has been widely recognized as a valid approach for improving load-balancing effect in AI networks.

Motivations

- Adaptive routing relies on real-time awareness of the path bandwidth of each ECMP route to perform per-packet or per-flowlet WECMP load-balancing.
- BGP is widely used as an underlay routing protocol for large-scale data center networks, making it straightforward to leverage BGP for propagating and calculating the end-to-end path bandwidth information of each ECMP route.
- The link-bandwidth attribute (draft-ietf-idr-link-bandwidth) is not intended and not suitable for this scenario, particularly for 5-stage CLOS networks.
- A BGP-based adaptive routing mechanism called FARE (Fully Adaptive Routing Ethernet) is built upon the newly defined BGP path bandwidth attribute and is applicable to both 3-stage and 5-stage CLOS networks.

Illustration (3-stage CLOS)

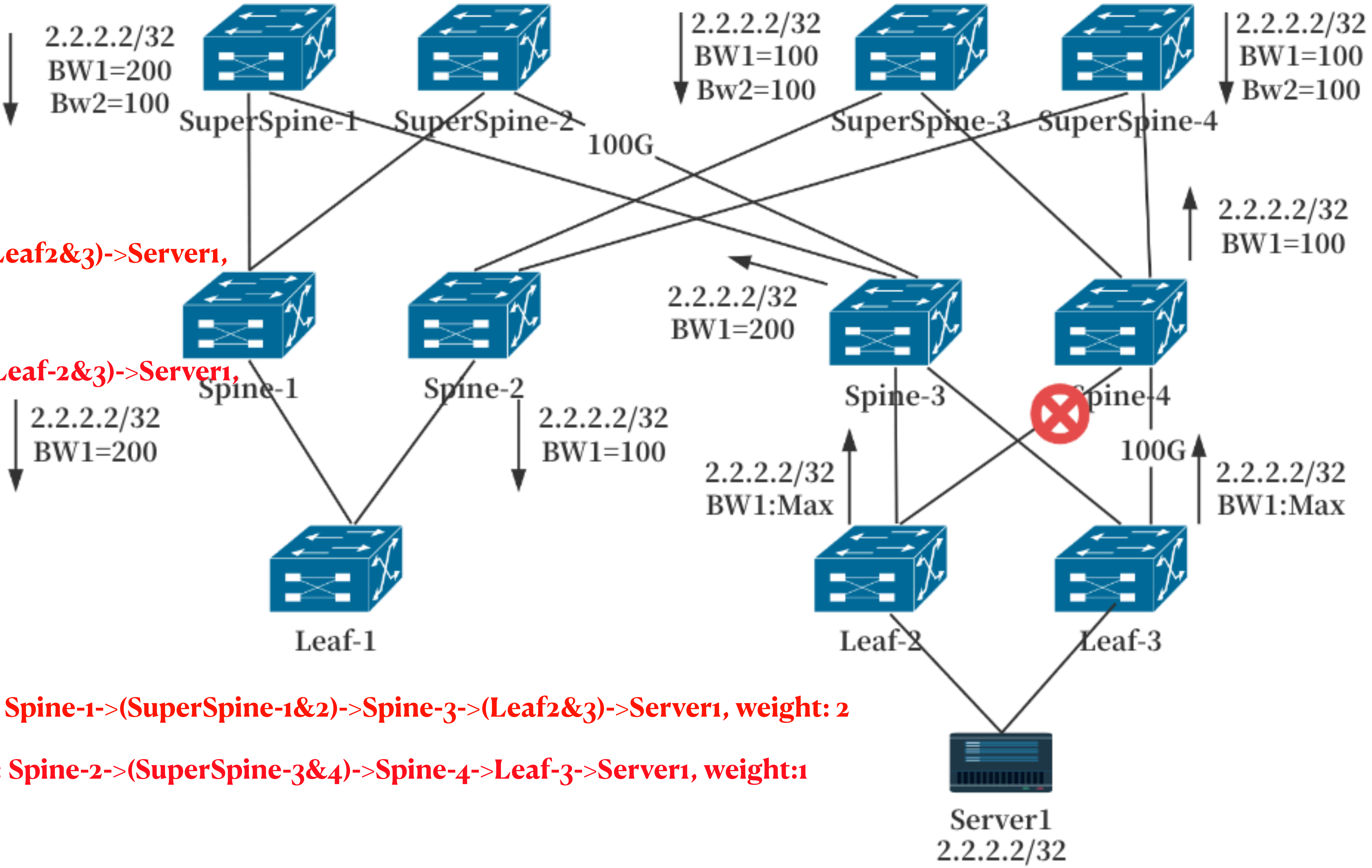


@Leaf-1

ECMP route-1 to Server1: Spine-1->(Leaf-2&Leaf3)->Server1, weight: 2

ECMP route-2 to Server1: Spine-2->Leaf3->Server1, weight:1

Illustration (5-stage CLOS)



@Spine-1

ECMP route-1 to Server1:
SuperSpine-1->Spine-3->(Leaf2&3)->Server1,
weight: 1

ECMP route-2 to Server1:
SuperSpine-2->Spine-3->(Leaf-2&3)->Server1,
weight:1

@Leaf-1

ECMP route-1 to Server1: Spine-1->(SuperSpine-1&2)->Spine-3->(Leaf2&3)->Server1, weight: 2

ECMP route-2 to Server1: Spine-2->(SuperSpine-3&4)->Spine-4->Leaf-3->Server1, weight:1

Implementation Status

- The FARE protocol has been implemented by over two vendors/cloud providers on their Ethernet switches, including China Mobile Cloud, Baidu Cloud, Ruijie, and IEIT System.
- It has also been successfully adapted to more than two mainstream switch chips, such as 1) Broadcom Tomahawk 5 and 2) Nvidia Spectrum-4.
- These 51.2T Ethernet switches with FARE enabled have been successfully deployed in over two 1K-GPU training clusters, one from China Mobile Cloud and another from Baidu Cloud.
- The bandwidth utilization ratio can exceed 95% on average, which is 1.6 times higher than the static ECMP ratio, even when upstream or downstream links are partially broken..

Next Steps

- The document is ready for the WG adoption call.
- The Early IANA Code Point Allocation is needed.
  - The Path Bandwidth attribute.