

Fully Adaptive Routing Ethernet (FARE) in Scale-Up Network (SUN) draft-xu-rtgtw-fare-in-sun-02

Xiaohu Xu@China Mobile

Zongying He@Broadcom

Nan Wang @Intel

Nan Wang@Hygon

Hua Wang@Moore Threads

Jian Guo@Biren Technology

Xiang Li@Enflame Technology

Tianyou Zhou@Resnics Technology

Yongtao Yang@Centec

Yinben Xia@Tencent

Weifeng Zhang@Tencent

Peilong Wang@Baidu

Yan Zhuang@Huawei

Fajie Yang@Cloudnine

Xiaojun Wang@Ruijie Networks

Chao Li@Metanet

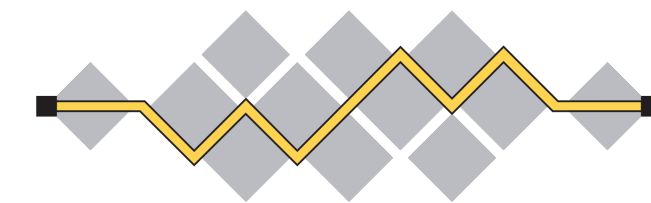
IETF125, Shenzhen

Changes Since -01

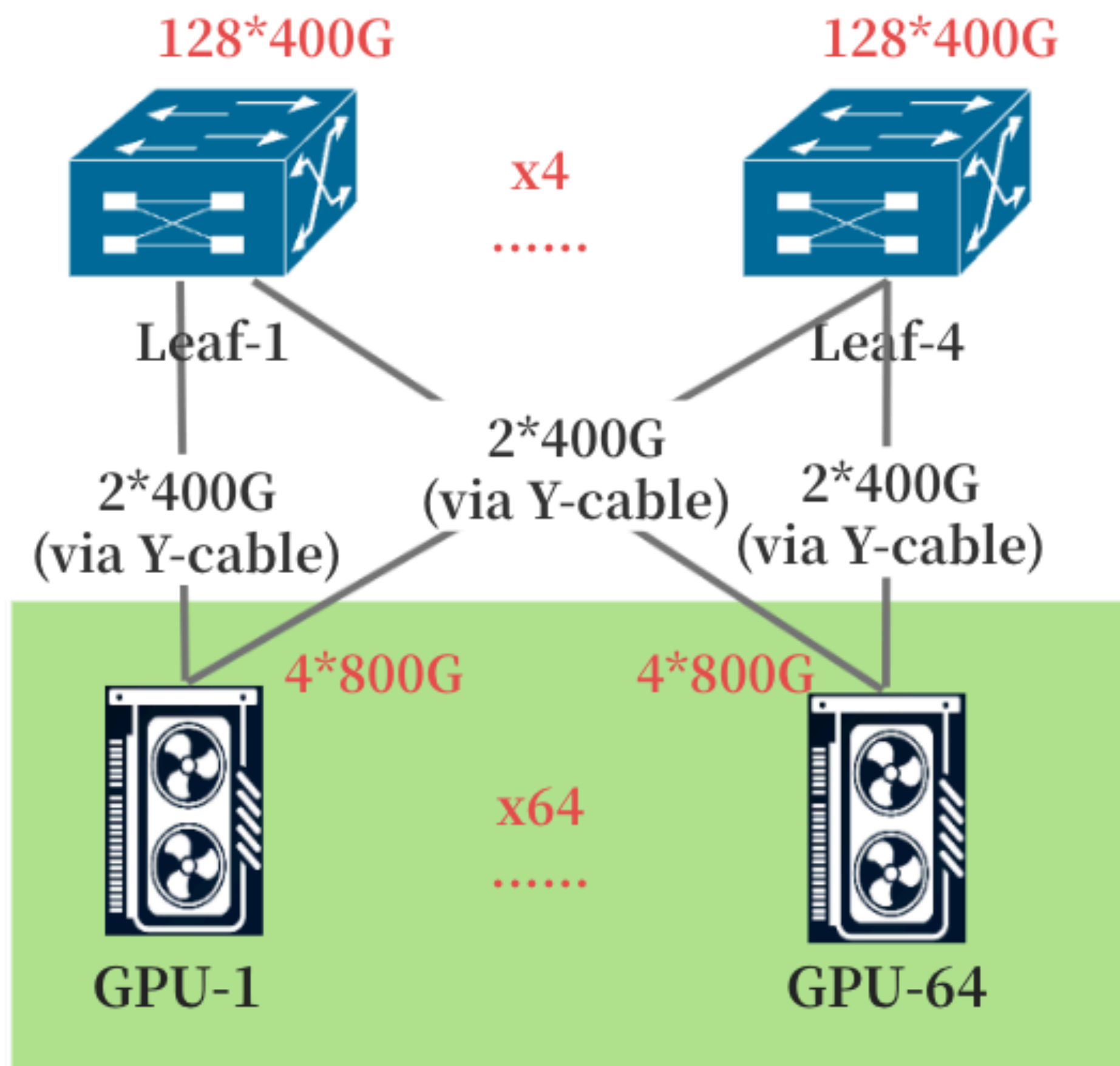
- Add five new co-authors mainly from GPU vendors:
 - Nan Wang@Intel
 - Nan Wang@Hygon
 - Jian Guo@Biren Technology
 - Xiang Li@Enflame Technology
 - Weifeng Zhang@Tencent
- Add considerations on memory semantic operations.

Motivations

- Ethernet-based scale-up networks (e.g., OCP ESUN, Broadcom SUE-T, AMD UALoE, Intel Gaudi, Microsoft Maia, Meta MTIA) have become industry standard. These networks typically employ multiple planes, with GPUs multi-homed to each plane.
- The collective communication traffic in such environments is characterized by low entropy, elephant flows, and burstiness. When GPUs use static ECMP for load balancing across planes, collision probability is high.
- Adaptive routing—using real-time path bandwidth or congestion feedback to perform per-packet or per-flow WECMP—has emerged as an effective solution for improving load balancing in multi-plane networks.
- Fully Adaptive Routing Ethernet (FARE) using BGP (draft-xu-idr-fare) offers a standards-based adaptive routing mechanism originally designed for scale-out networks. Extending FARE-BGP to scale-up networks is straightforward: the protocol can be extended from switches to GPUs, enabling GPUs to act as routers performing adaptive routing.

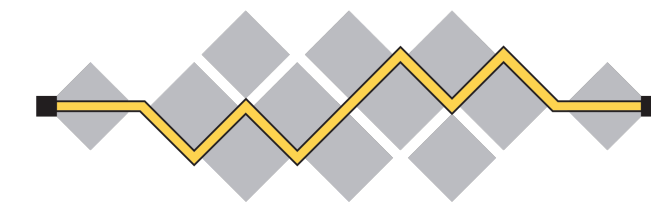


Scale-up Network Topology Example

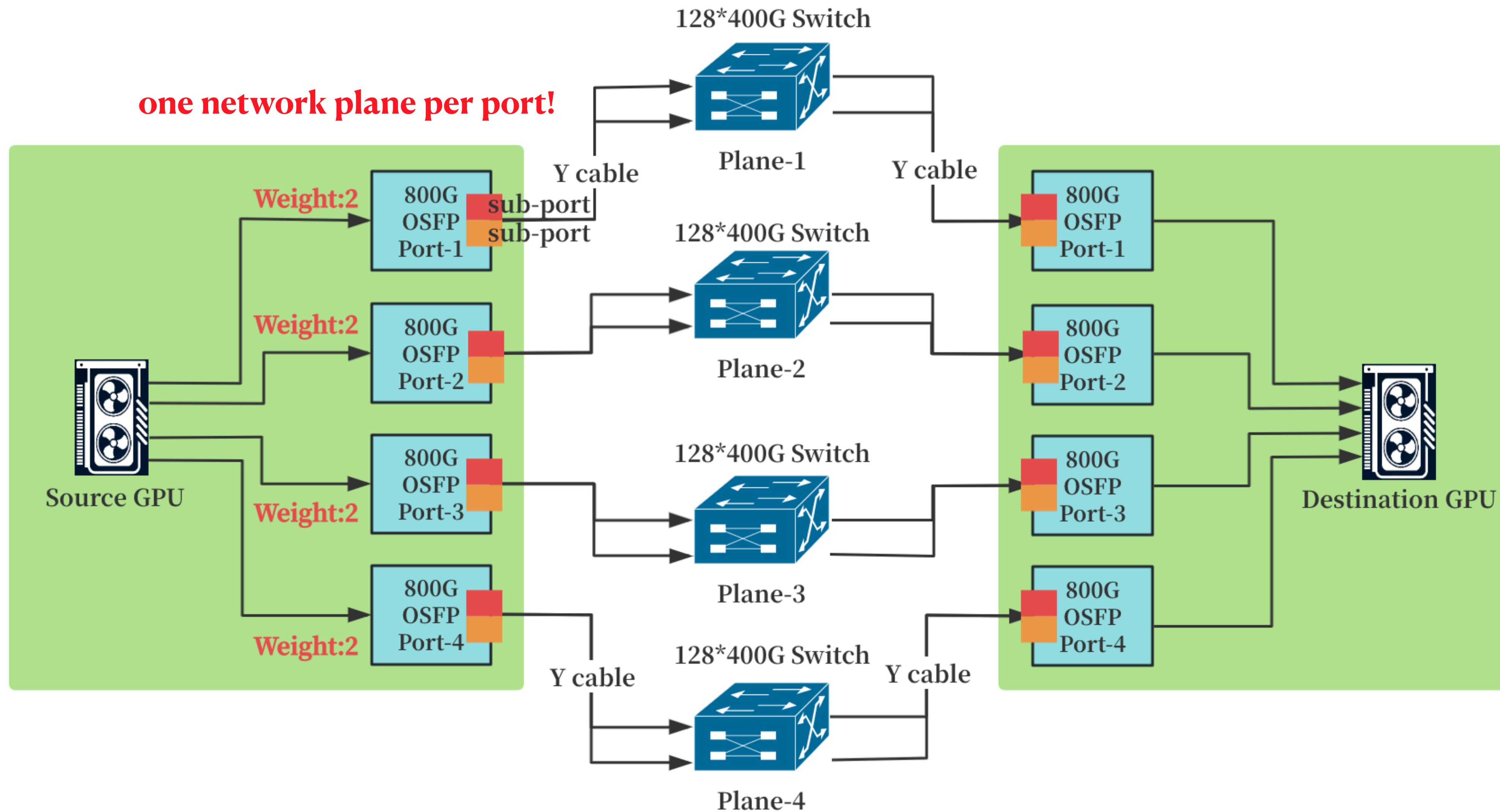


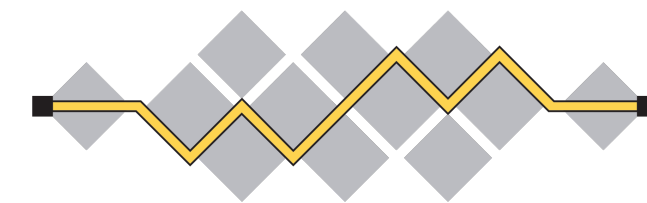
64-GPU SuperPod

- Each GPU is equipped with multiple 800G ports, with at least one sub-port connected to a given scale-up network plane.
- The reasons for using multiple links per plane include, but are not limited to:
 - Facilitating an increase in superpod scale within a single-tier scale-up network.
 - Simplifying further upgrades to a two-tier scale-up network.

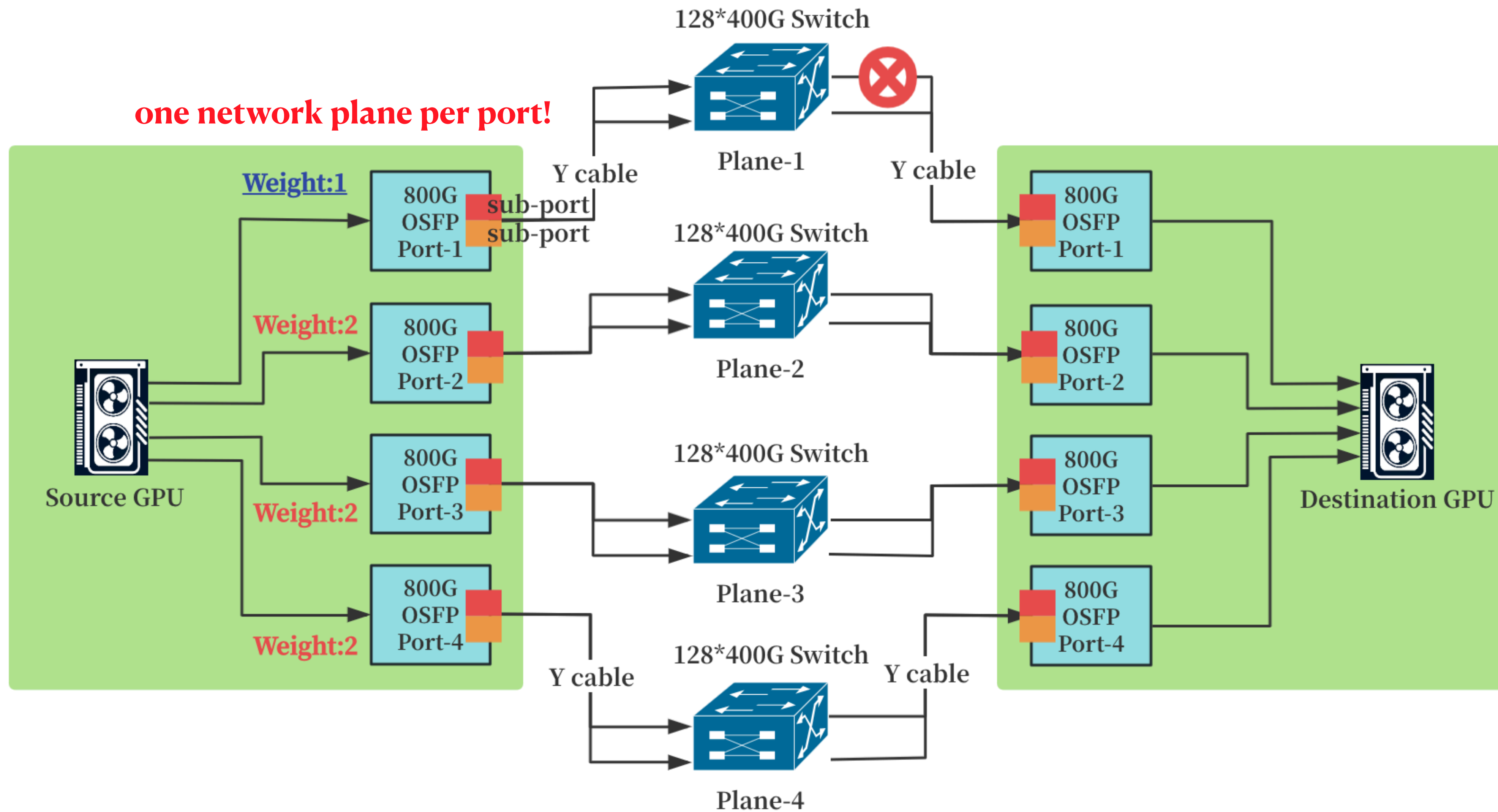


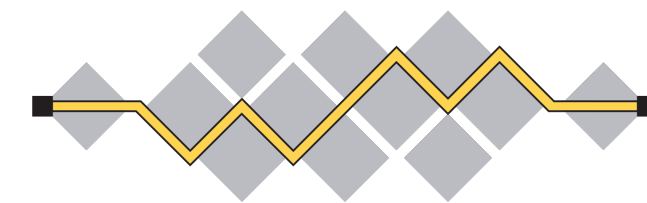
Path-Bandwidth-Aware WECMP (1/3) I E T F[®]



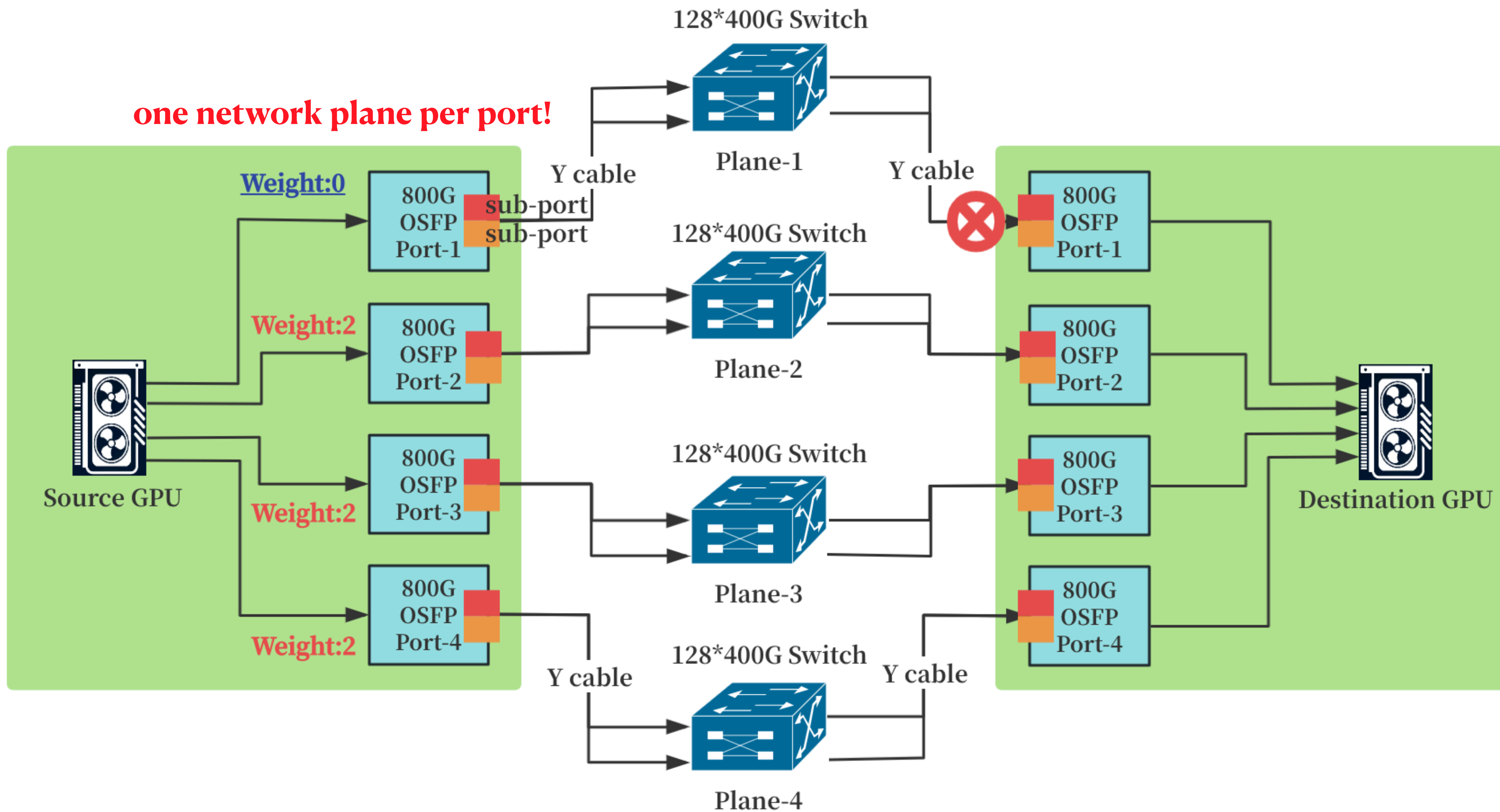


Path-Bandwidth-Aware WECMP (2/3) I E T F[®]



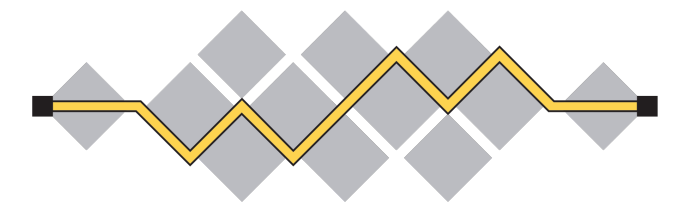


Path-Bandwidth-Aware WECMP (3/3) I E T F[®]



Two WECMP Options

- Per-flow weighted load-balancing @ GPUs
 - It's applicable to the ordered packet delivery mode.
 - At least one RDMA Queue Pair (QP) per sub-port must be established between a given GPU pair.
 - Switches must perform per-flow load-balancing.
- Per-packet weighted load-balancing @ GPUs
 - It's applicable to the disordered packet delivery mode.
 - A single QP between a given GPU pair suffices.
 - Packets are sprayed across all available network planes by the source GPU.
 - Switches should also perform per-packet weighted load balancing.

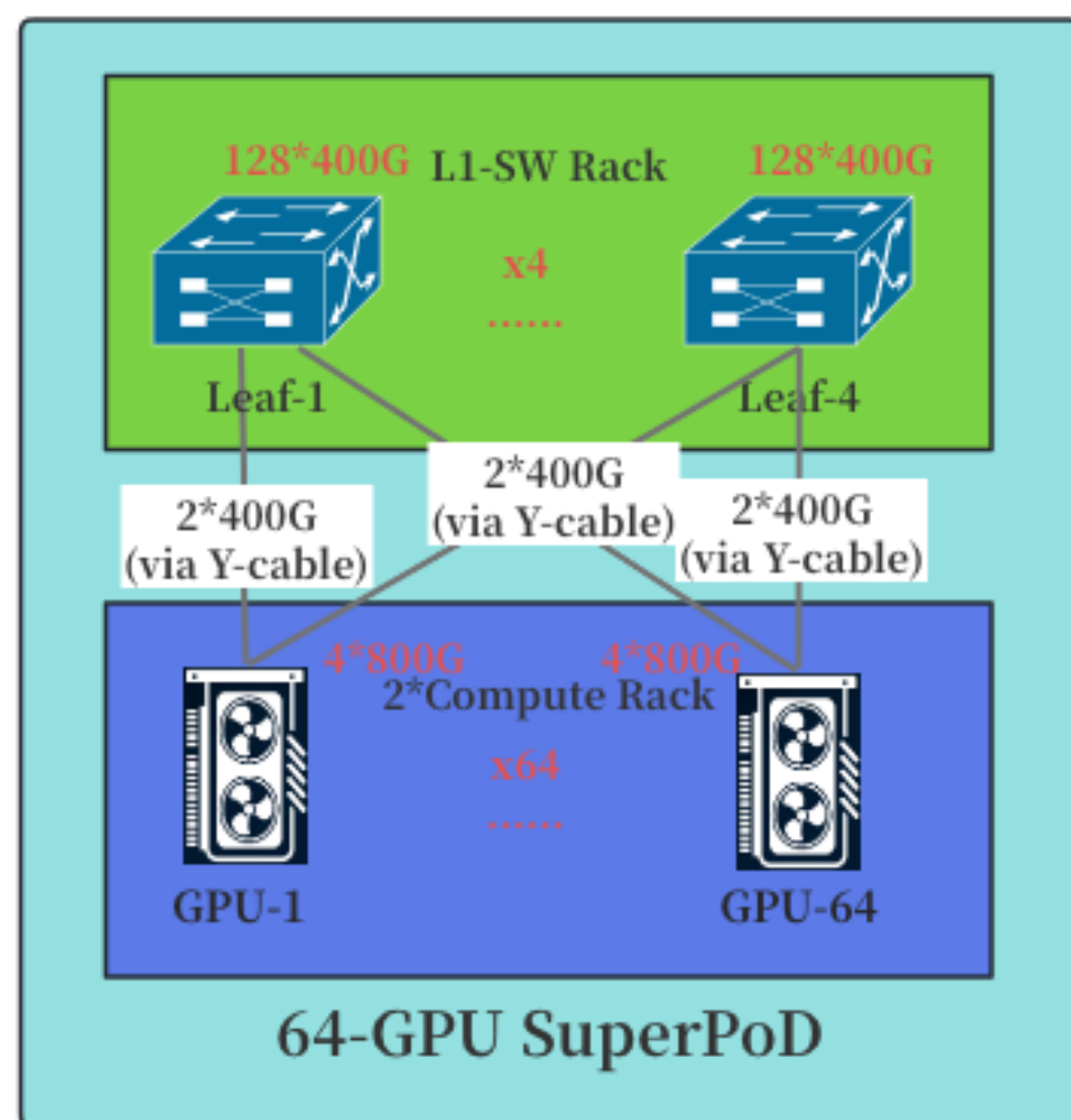


Considerations on Memory Semantic I E T F[®]

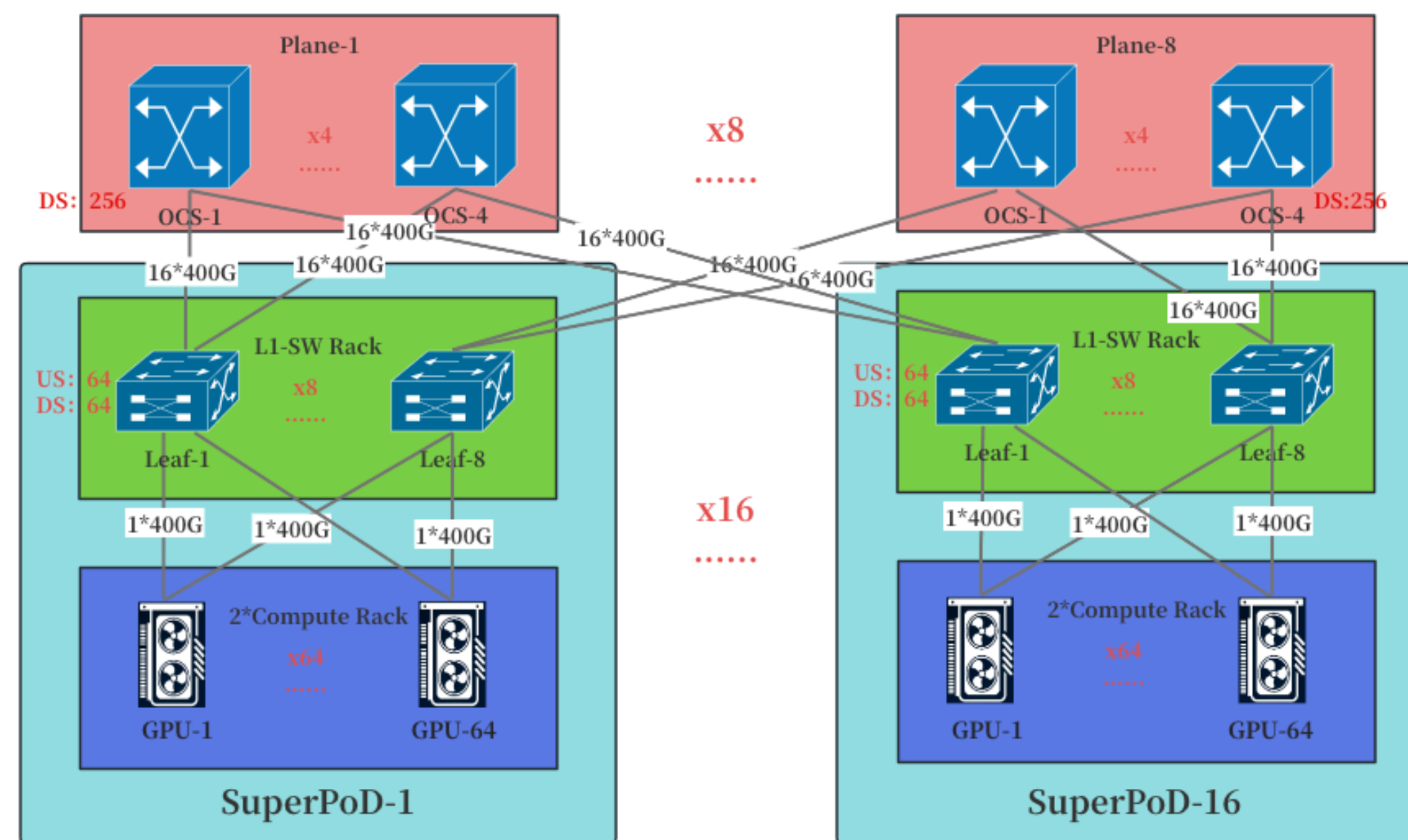
- When implementing memory semantics, network transmission ordering guarantees fall into three categories:
 - **Weak Ordering:** Relies on full packet spraying and GPU Reorder Buffers.
 - **Strong Ordering:** Mandates in-order delivery for the entire transaction stream, simplifying GPU implementation.
 - **Partial Ordering:** Uses in-order delivery for strict operations and out-of-order delivery for others, balancing flexibility and control.
- Flexible selection between per-packet and per-flow weighted load balancing at the GPU can satisfy the above memory semantics operation requirements.

Implementation Status

- A PoC for a 64-GPU disaggregated SuperPod with a single-stage CLOS scale-up network is near completion, while a 1024-GPU cascaded disaggregated SuperPod with an optical-electronic-hybrid scale-up network is under development.



64-GPU Disaggregated SuperPod
(Topology)



1024-GPU Disaggregated SuperPod
(Topology)



Next Step

- Any suggestions or comments?