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

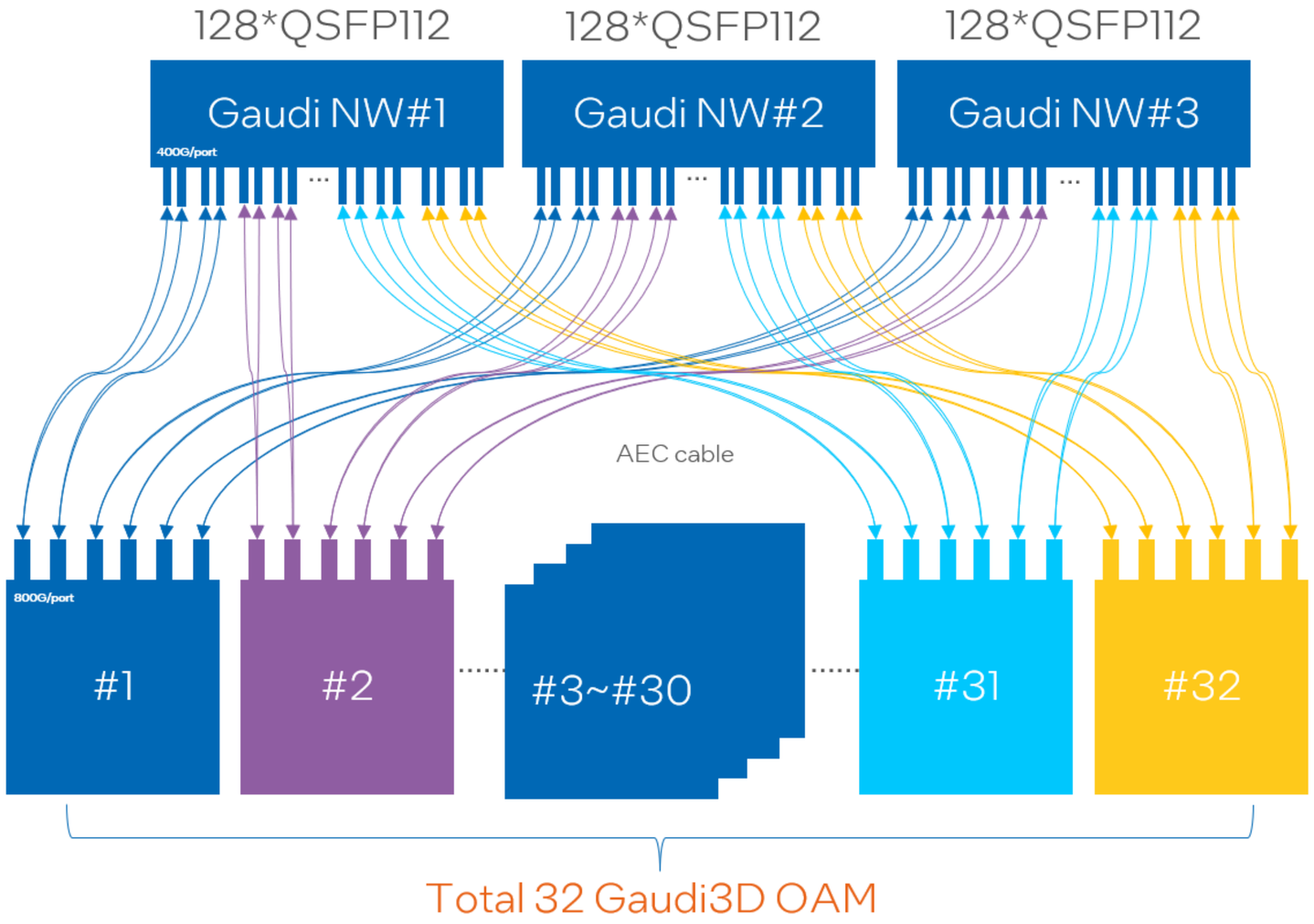
Xiaohu Xu@China Mobile
Zongying He@Broadcom
Nan Wang @Intel
Hua Wang@Moore Threads
Tianyou Zhou@Resnics Technology
Yongtao Yang@Centec
Yinben Xia@Tencent
Weifeng Zhang@Tecent
Peilong Wang@Baidu
Yan Zhuang@Huawei
Fajie Yang@Cloudnine
Xiaojun Wang@Ruijie Networks
Chao Li@Metanet

IETF123, Madrid

Backgrounds

- Scale-up networks are essential for both AI training and AI inference.
- These networks typically consist of multiple network planes, with GPUs being multi-homed.
- The characteristics of collective-communication traffic include: low entropy, elephant flows, and burstiness.
- Static ECMP load-balancing across multiple network planes, performed by GPUs, can lead to a high probability of collisions.
- Adaptive routing has been widely recognized by the community including the UEC and the UAL as an effective approach for improving load-balancing in multi-plane networks.

Illustration



- Generic Scale-up Network Topology
  - Each GPU is equipped with multiple physical ports, which can be further divided into multiple sub-ports.
  - One or more port of a given GPU may be attached into a single network plane, which usually contains a single switch in the single-stage CLOS scale-up network.

Solution Overview

- Adaptive routing relies on real-time awareness of the path bandwidth and/or even congestion status of each ECMP route to perform per-packet, or per-flow WECMP load-balancing.
- Fully Adaptive Routing Ethernet (FARE) using BGP (draft-xu-idr-fare) is a standard-based adaptive routing mechanism, which is applicable to both 3-stage and 5-stage CLOS scale-out networks.
- It seems straightforward to extend the FARE-BGP protocol from switches to GPUs further in scale-up networks.
- The traffic between GPUs is balanced across all available network planes according to the path bandwidth values associated with those network planes.

Illustration (con't)

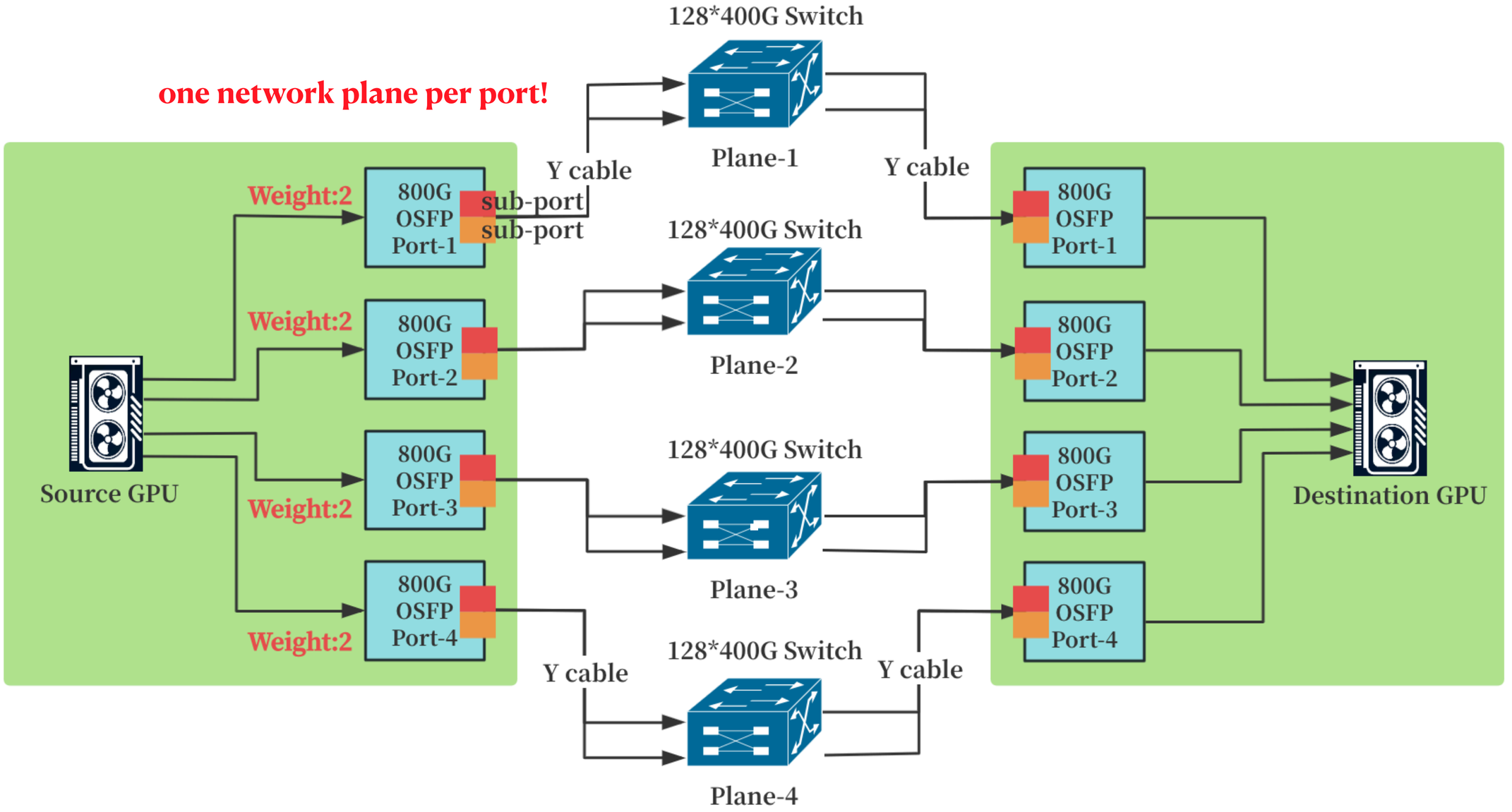


Illustration (con't)

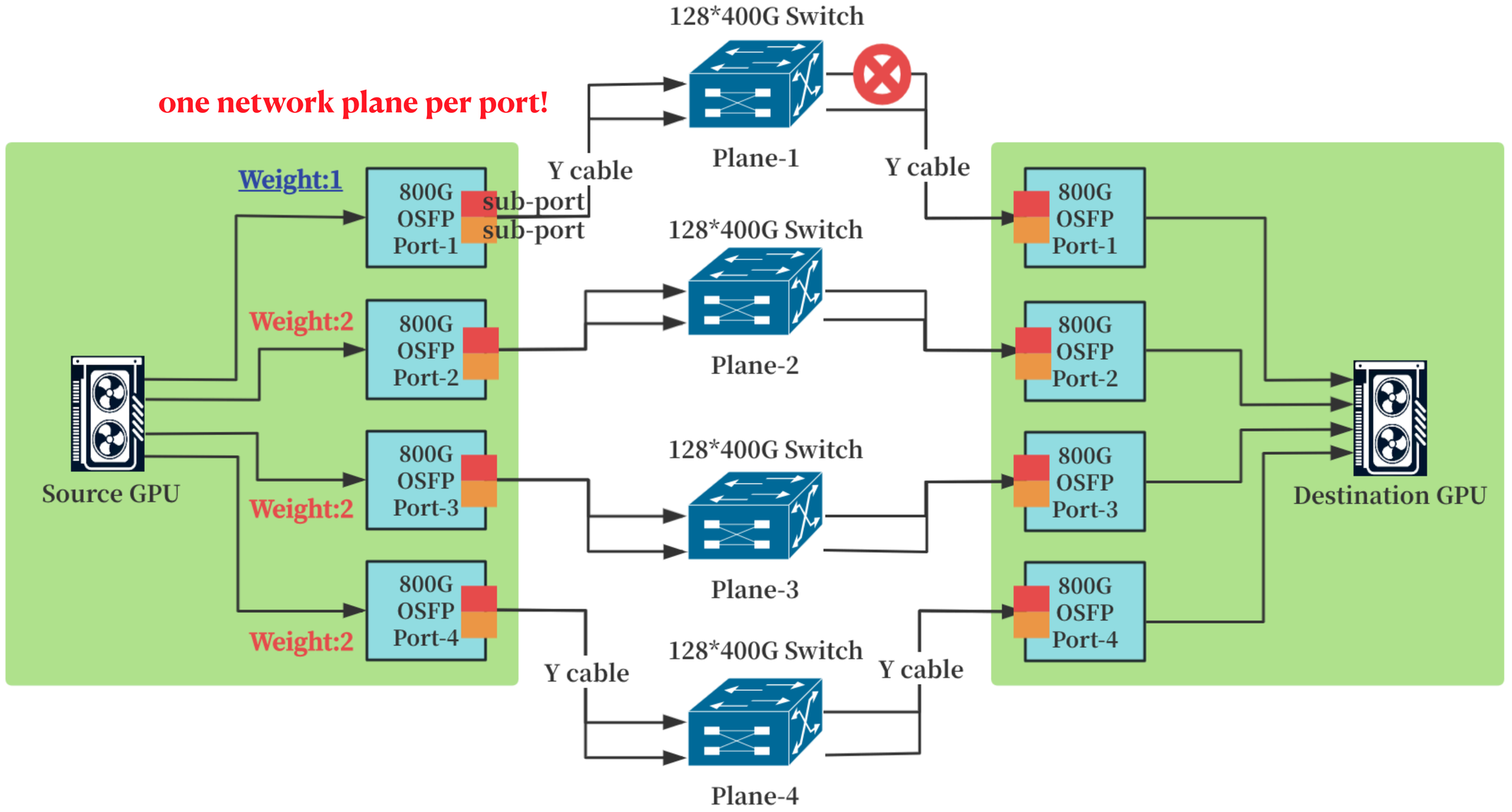
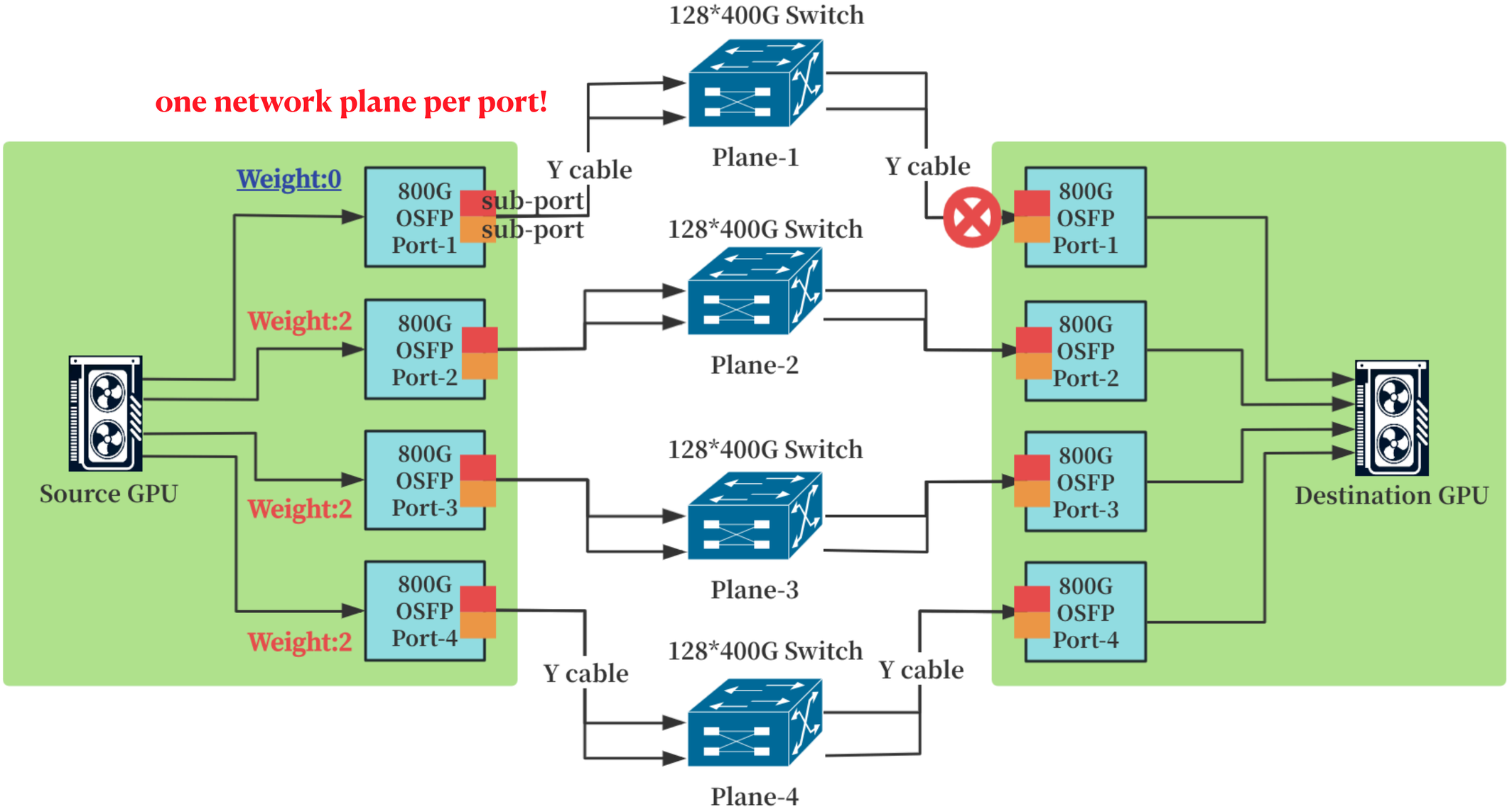


Illustration (con't)



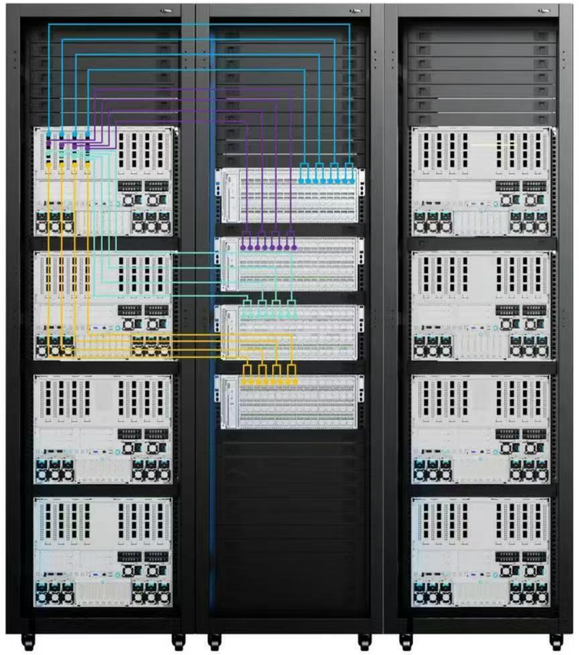
Two WECMP Options

- Per-flow weighted load-balancing
  - 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 to assure ordered packet delivery.
- Per-packet weighted load-balancing
  - 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.

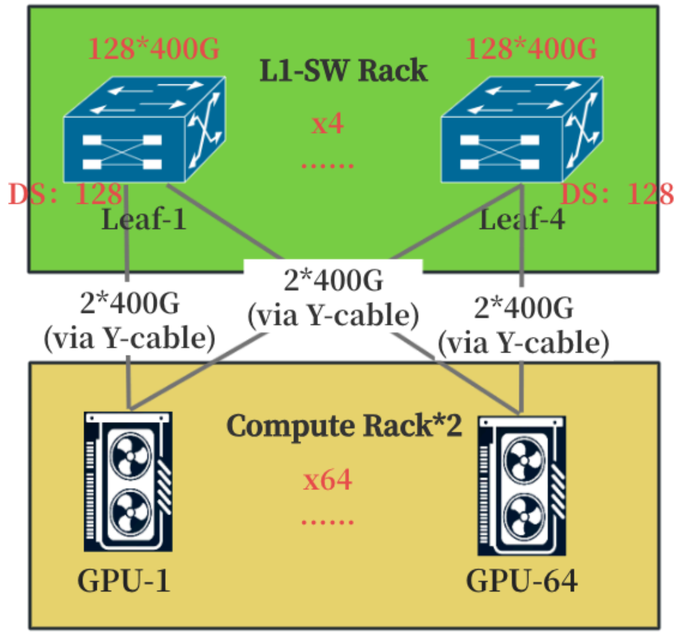
Implementation Status

- A PoC built on the **Disaggregated SuperPod,** which contains a **single-stage scale-up network,** is currently underway.

64-GPU Disaggregated SuperPoD
(Picture of the actual object)



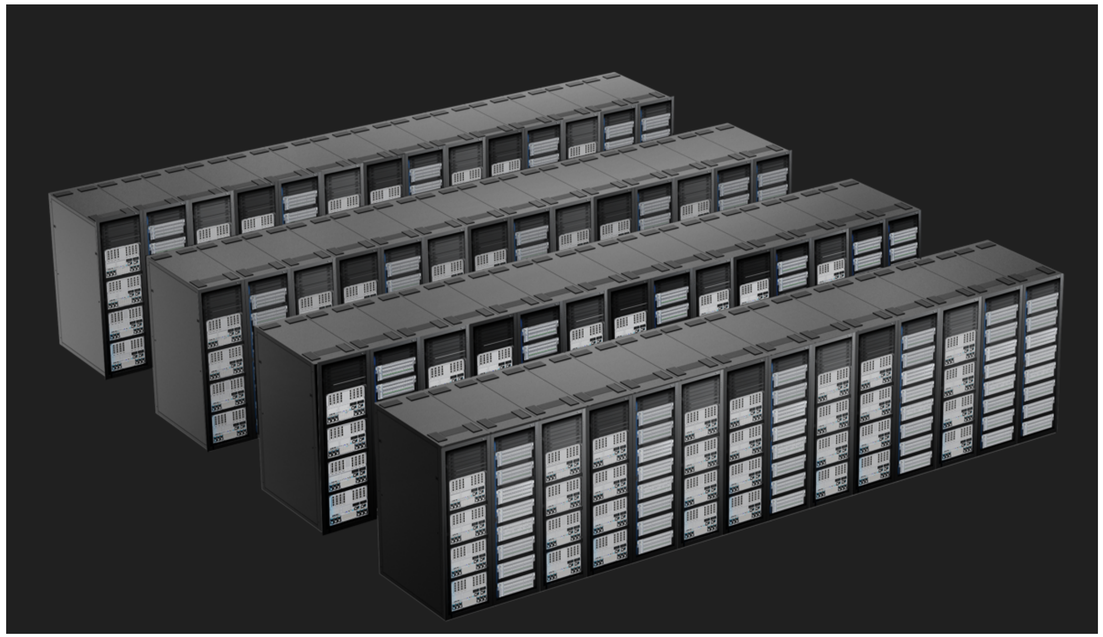
64-GPU Disaggregated SuperPoD
(Conceptual graph)



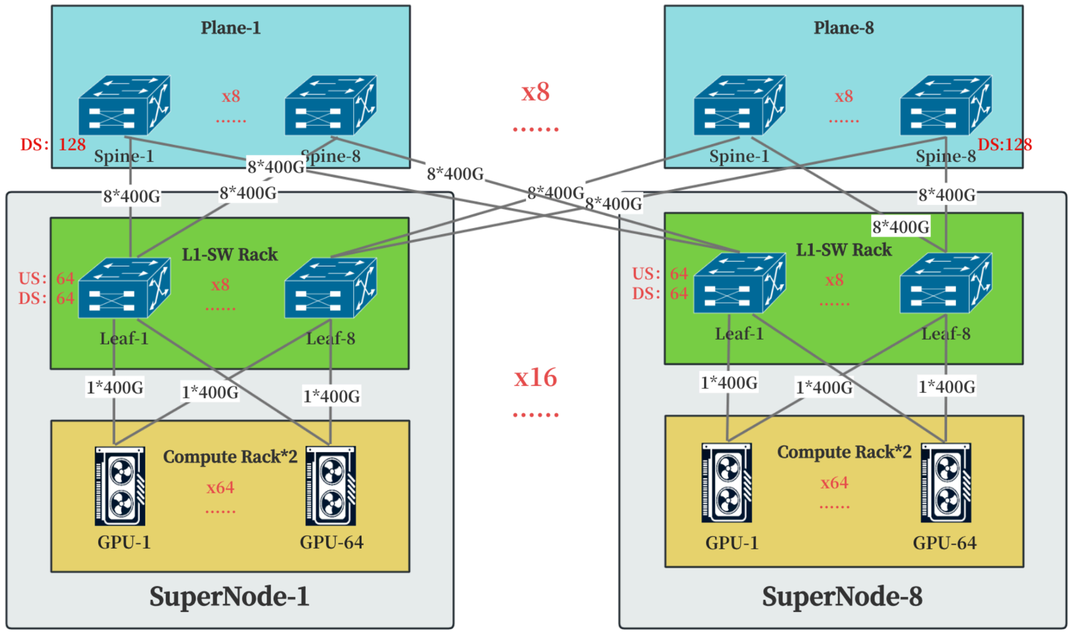
64-GPU Disaggregated SuperPoD
(Topology)

Next Steps

- The POC would be extended further to the **Disaggregated SuperPoD Ultra**, which contains a **3-stage CLOS scale-up network,** in the near future.
- Any suggestions or comments?



1024-GPU Disaggregated SuperPoD Ultra（Conceptual graph）



1024-GPU Disaggregated SuperPoD Ultra（Topology）