UB-mesh:

An New Interconnection Technology for Large AI SuperNode

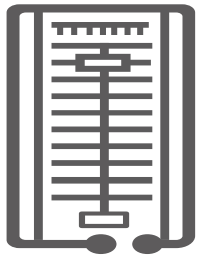
Heng Liao, PhD

Chief Scientist, Hisilicon

2025/8

HUAWEI

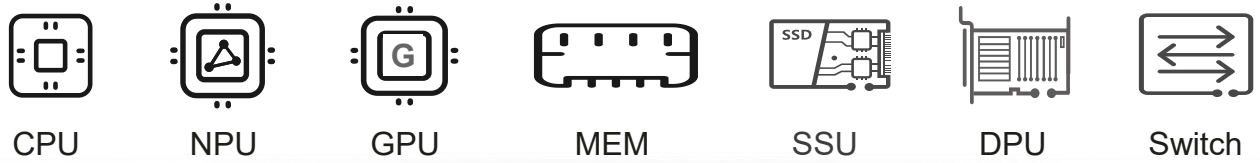
SuperNode is Becoming the Norm for GigaWatt AI Data Center



- SuperNode tightly connects a large number of devices to form a single large computing system
  - Chip number: 8 →1Million
  - Bandwidth/chip: 100Gbps → 10Tbps
  - Latency/hop: us → 150ns
  - Scale: Single Rack → Data Center
  - Schema: Asynchronous DMA → Synchronous Load/Store
  - Connecting all type of devices: GPU only → CPU+GPU+Pool Memory+SSD+NIC+Switches

HUAWEI

Unified Bus
Unifying Common Bus and Network Protocols



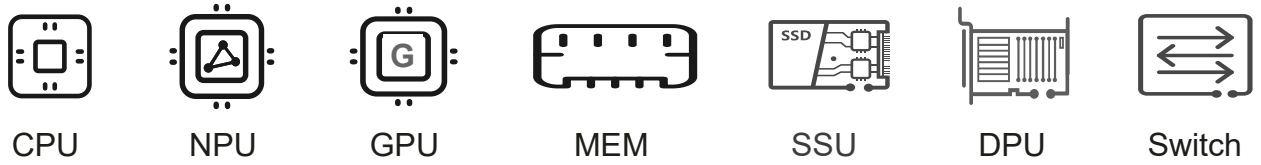
Unified Bus

UPI + PCIe + CXL + RoCE + Nvlink + NoF + FCoE + TCP/IP + gRPC …

- Why
  - Lower latency: any port can connect and forward to any type of devices without protocol conversion overhead
  - Higher bandwidth: all high speed SERDES can utilized for any purpose
  - Simplified Schema: Load/Store+RDMA+uRPC
  - Compatibility: UBoE can run natively over Ethernet Network

HUAWEI

Challenges for Extending Local Bus to Data Center Scale

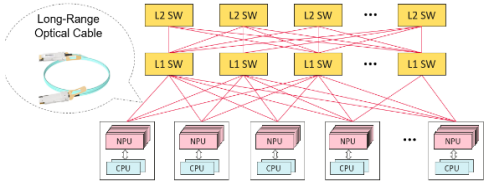


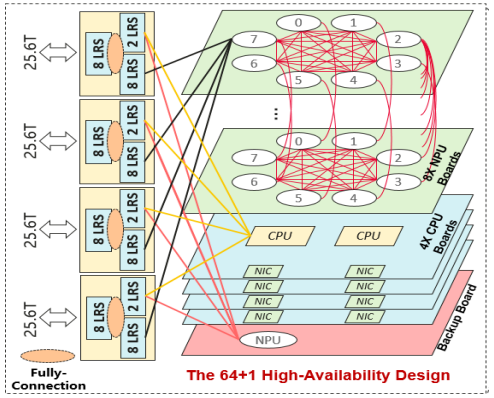
Unified Bus

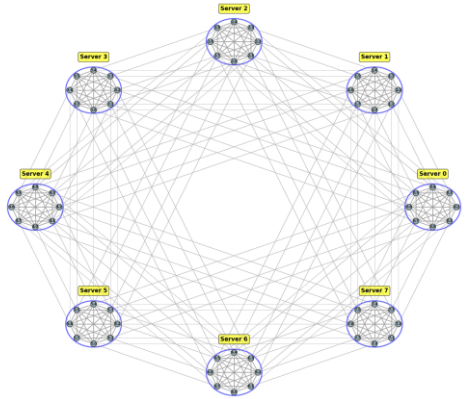
- More nodes at high bandwidth
- Longer reach physical links = optical
- Higher BER of optical links = more sophisticated error recovery scheme (e.g. Link Level Retry + …)
- Load/Store Schema: SoC recovery scheme to prevent processor pipeline deadlock
- Resiliency over node Failure

100x Node Bandwidth without 100x Cost

Requires New Hybrid Physical Topology





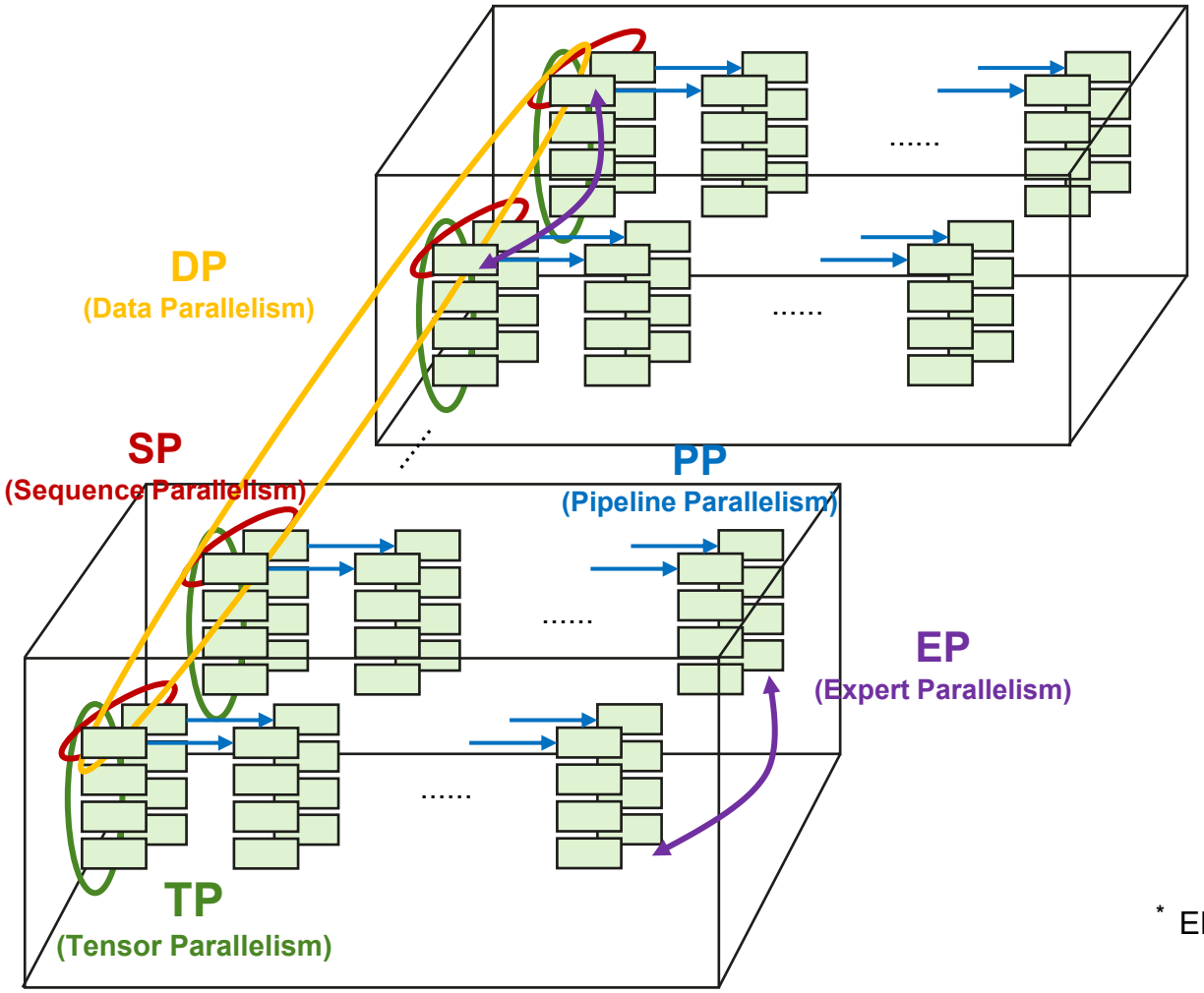


- CLOS = Versatile +Trustworthy, suitable for top level network (1M) at lower bandwidth
- nD mesh = High local bandwidth + diminished remote bandwidth suitable for Rack(~64) large Pod (128~8192)
- nD sparse mesh = Low cost, High bandwidth (16~128) suitable for smaller local deployment

HUAWEI

Key Observation: LLM Training Has Pairwise Hierarchical Traffic Patterns

Five-dimensional parallelism in LLM training



Hierarchical traffic volume and bandwidth requirements

	Traffic per Round	Rounds of Communication	Total Traffic
PP	96 MB	120	**11.25 GB**
DP	1277.86 MB	32	**39.93 GB**
EP (Group-Wise)	5.625 MB	7680	**42.19 GB**
TP	180 MB	7680	**1350 GB**
SP	762 MB	11520	**8572.5 GB**

Low Medium High

Example:

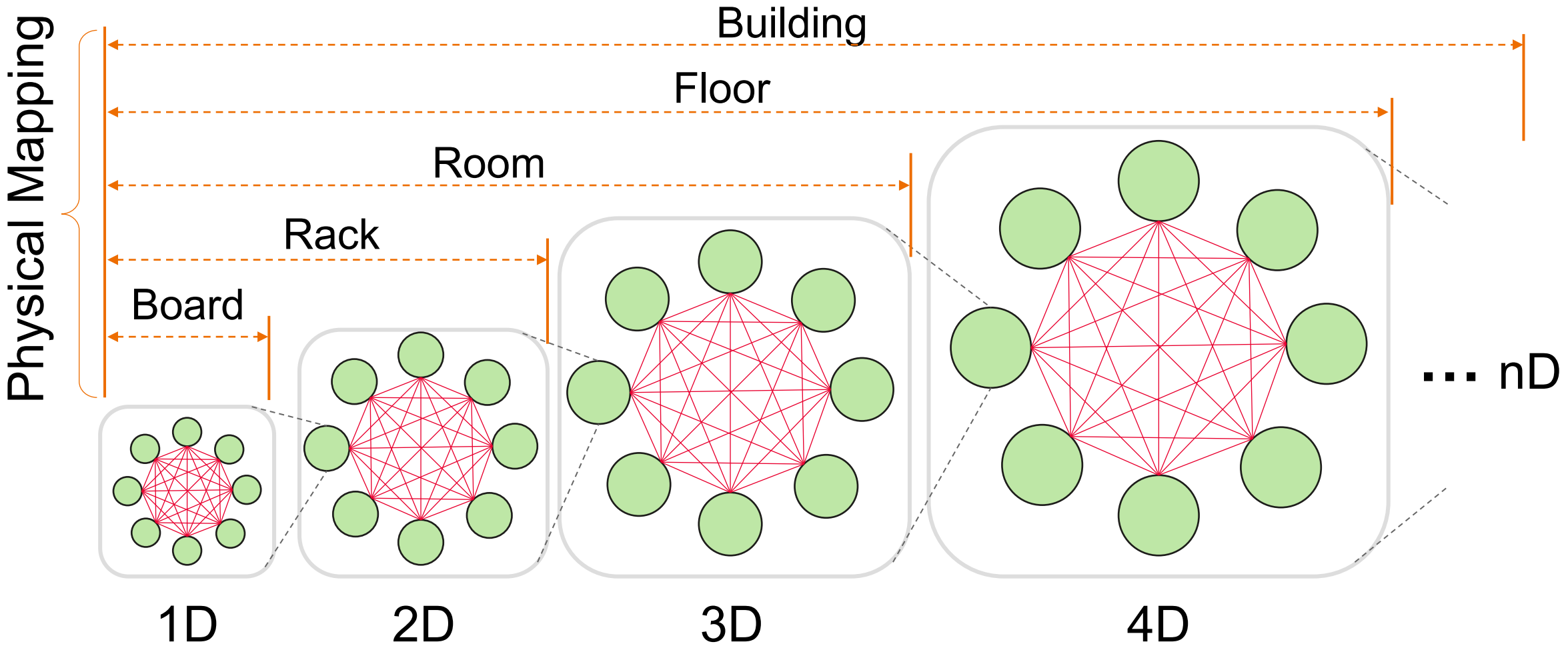
Model size ≈ 2T, Sequence length =1M,

TP=16, SP=128, EP=16, PP=4, DP=8

* EP traffic can be significantly reduced through group-wise All2All.

HUAWEI

UB-Mesh: Hierarchically Localized nD-FullMesh Network Topology



Short-range direct interconnect

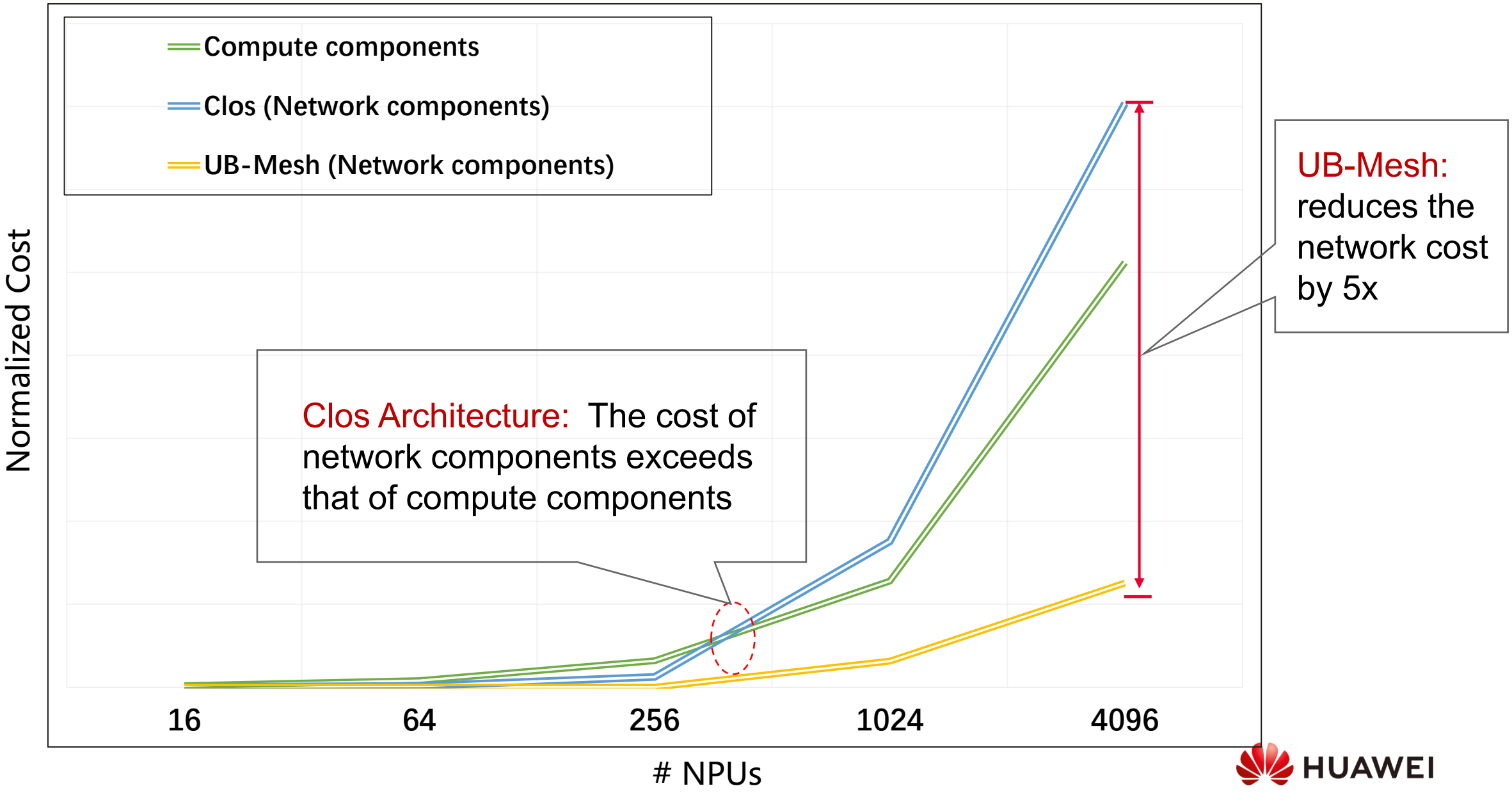
- maximize the usage of electrical cables

Per-tier oversubscribed bandwidth

- matches the dense-to-sparse traffic patterns

HUAWEI

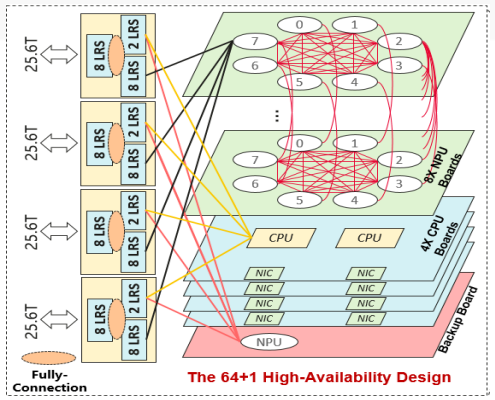
System Cost Comparison Between Clos and UB-Mesh



HUAWEI

8K Node Real Life Example – CLOS + 2D-Mesh

Rack: 64+1

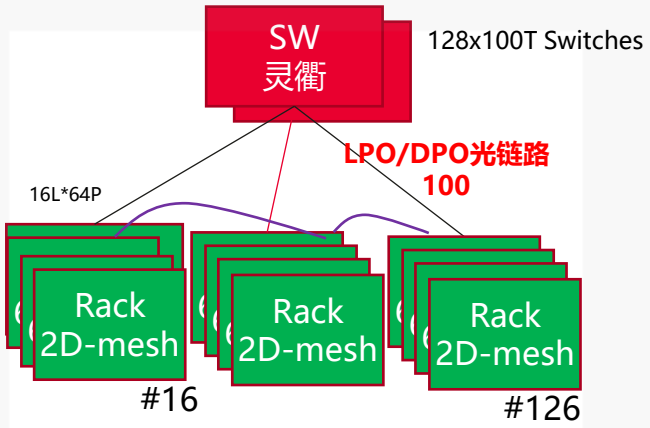


2D-Mesh: 64(NPU)+1(backup)+16(CPU)

Increase Rack Number

Linear Cost Scaling:
256 → 512 → 1024 → 2048 → 4096 → 8K

SuperNode: 8192 NPU (128 Racks)



1-layer CLOS (LPO optical)

HUAWEI

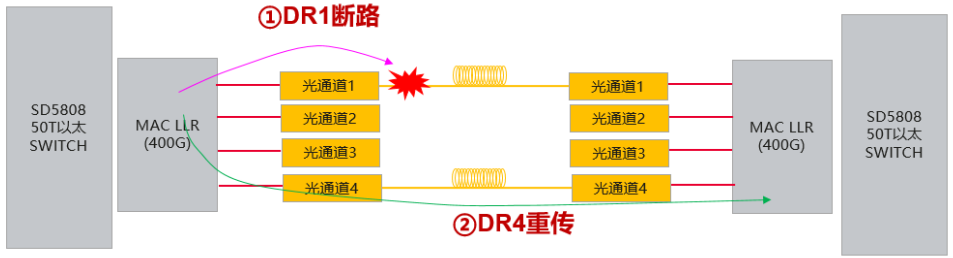
Resilient Optical Links

- Optical Link Flickering > 100ms can cause an SoC failure in Load/Store schema
- 3.9 flickerings/Day in 512P SuperNode

UB LLR + Optical module can improve 100x

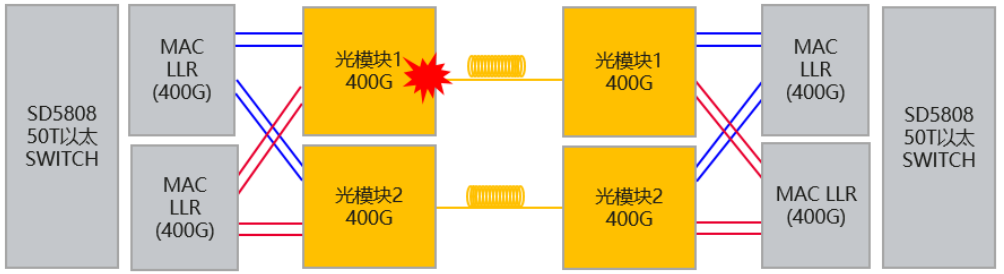
Within OM 4 lane backup = 0 packet loss

LLR over alternative optical link with same OM



Dual OM crossover = Single OM failure 0 packet loss

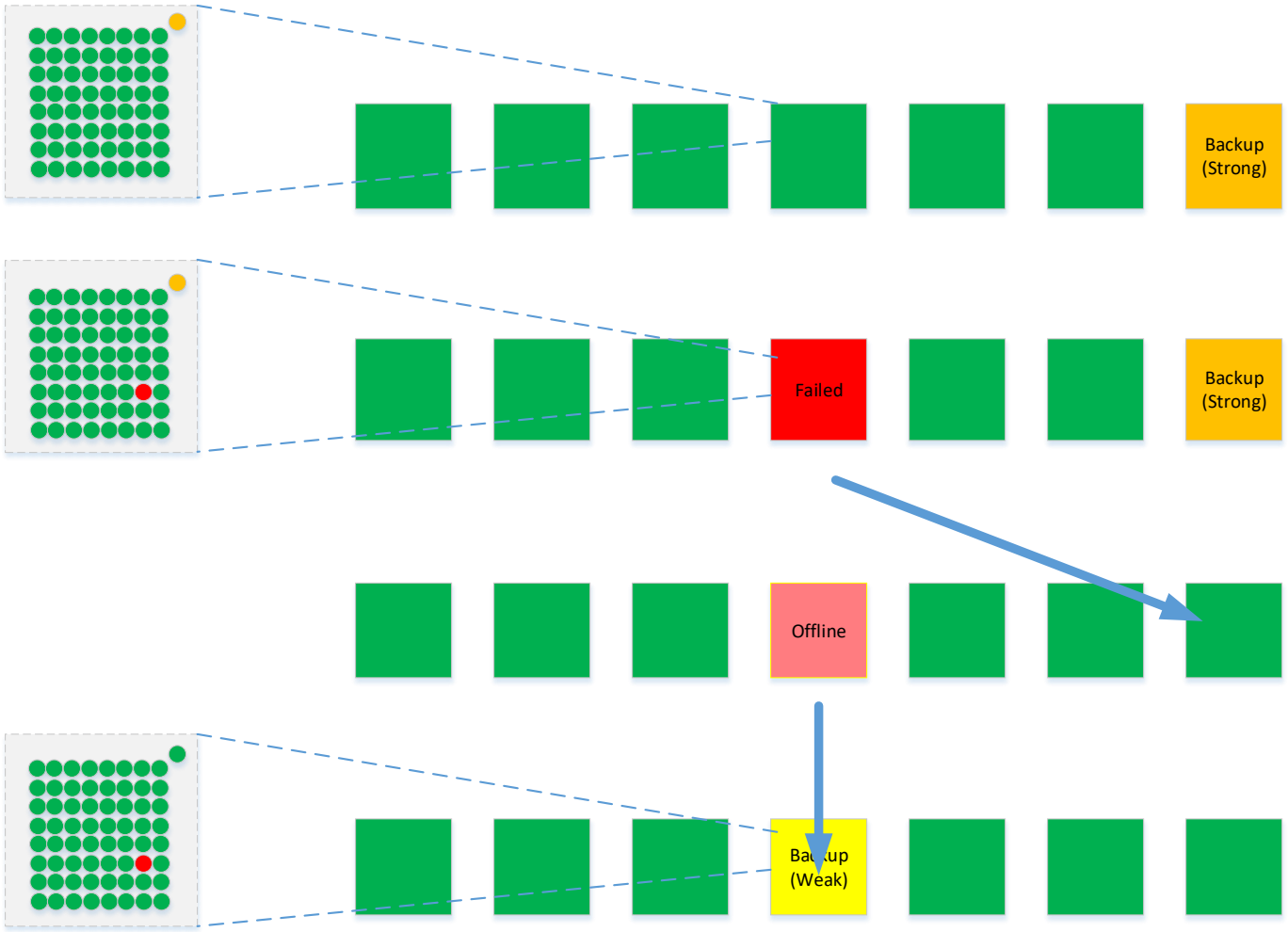
LLR over alternative OM



HUAWEI

Hierarchical System Resiliency: 100x MTBF

Service Time: 1hour → 1 Month



1) N Online + 1 Backup

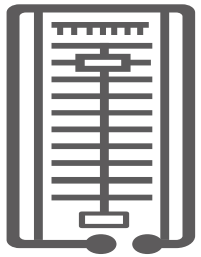
2) Failed Node migrate to Backup

3) Failed Node taken offline

4) Resume Failed Rack as Backup using Backup NPU within the Rack

HUAWEI

Summary



- Unified Protocol → UB
- Novel and pragmatic topology
- Improved protocol and OM design
- System Level Fault Tolerance
- Super Node size scaling: 64 – 8192 – 32768 – 1M?

SuperChips vs SuperNode → SuperChip + SuperNode

For more details: https://arxiv.org/abs/2503.20377

HUAWEI

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI