开放数据中心委员会
Open Data Center Committee

Research Progress of Intelligent Computing Networks in China

guoliang1@caict.ac.cn

ODCC Introduction

开放数据中心委员会
Open Data Center Committee

Open Data Center Committee (ODCC) was founded by Alibaba, Baidu, Tencent, China Telecom, China Mobile, CAICT. It now has over 100 members.

https://www.odcc.org.cn/introduction-en.html

Working Groups in ODCC

开放数据中心委员会
Open Data Center Committee

There are 6 WGs and several ad-hoc task groups.

Server WG

Wangfeng
China Telecom

Infrastructure WG

Li Daicheng
Baidu

Network WG

He Zekun
Tecent

Edge Computing WG

Chen Wei
Tecent

Test WG

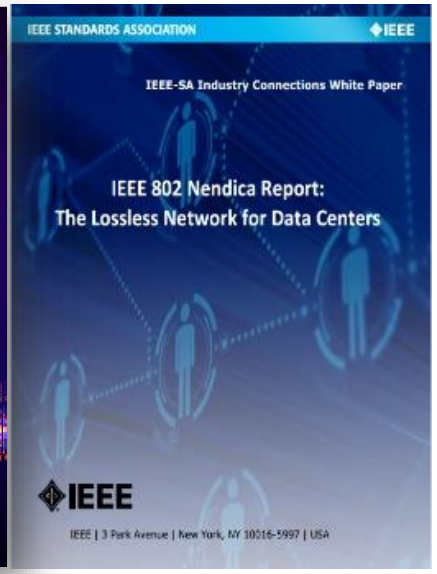
Guo Liang
CAICT

O&M WG

Chao Huaipo
Alibaba

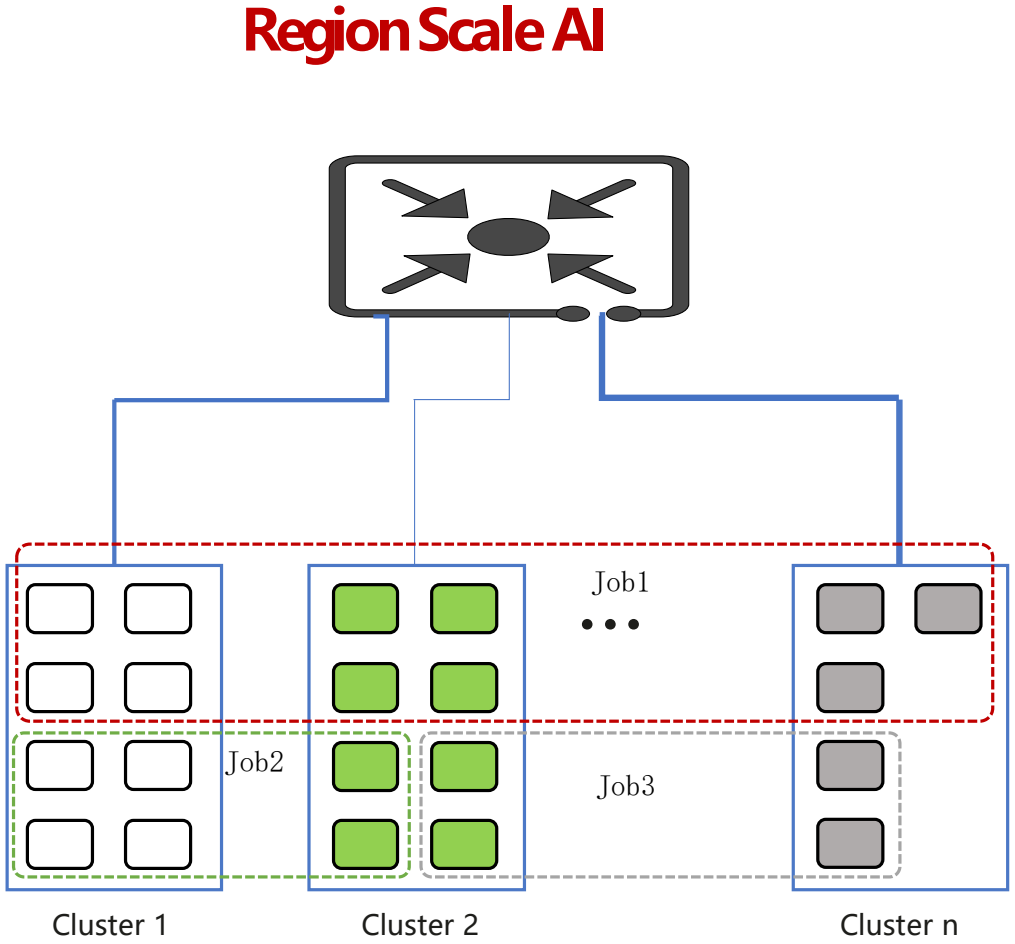
Achievements in network related work

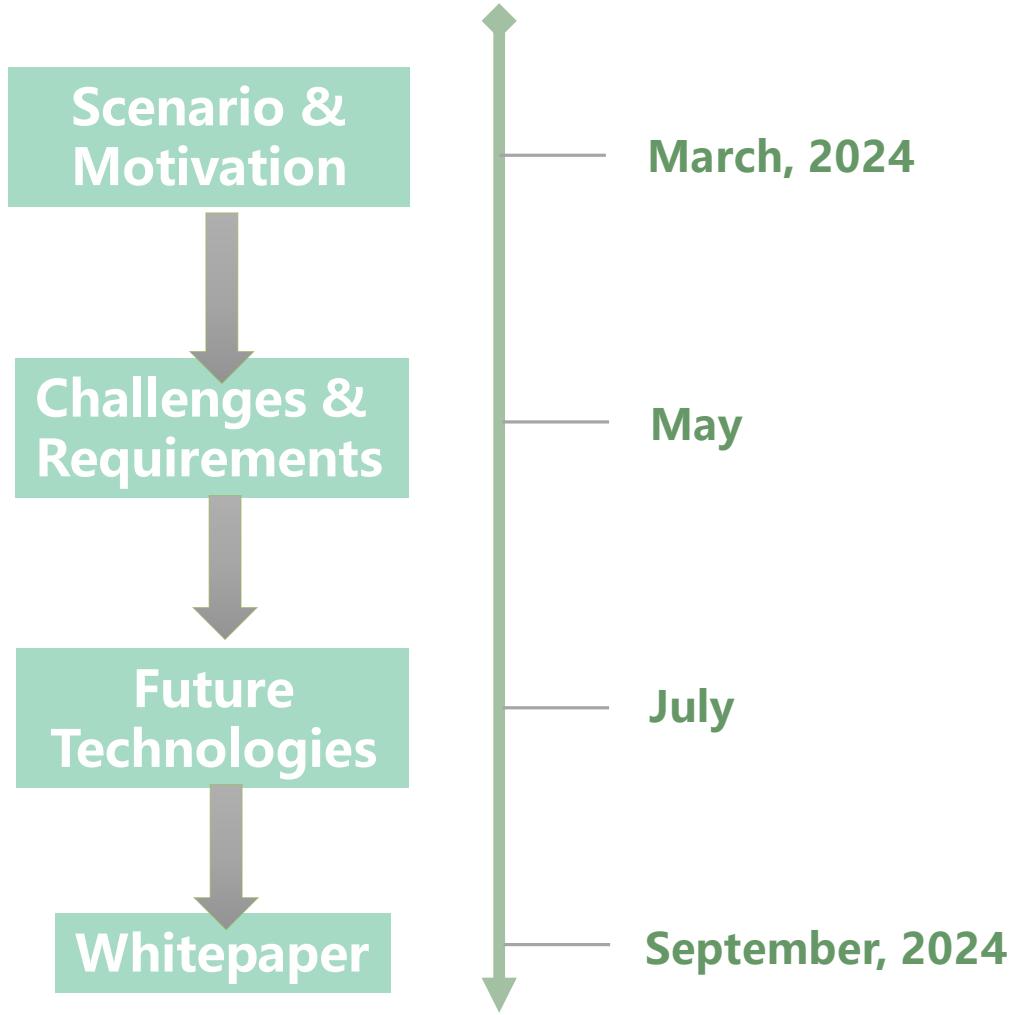
开放数据中心委员会
Open Data Center Committee



Region Scale AI Project

开放数据中心委员会
Open Data Center Committee

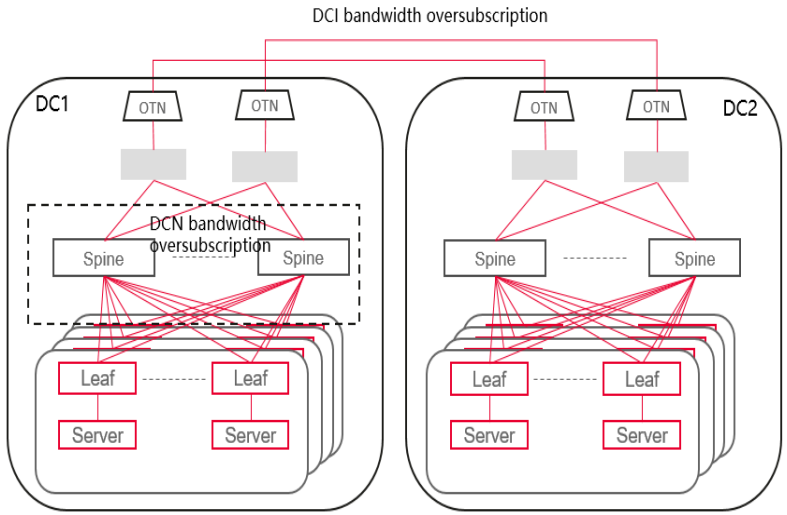




Region Scale AI: Challenges & Requirements

开放数据中心委员会
Open Data Center Committee

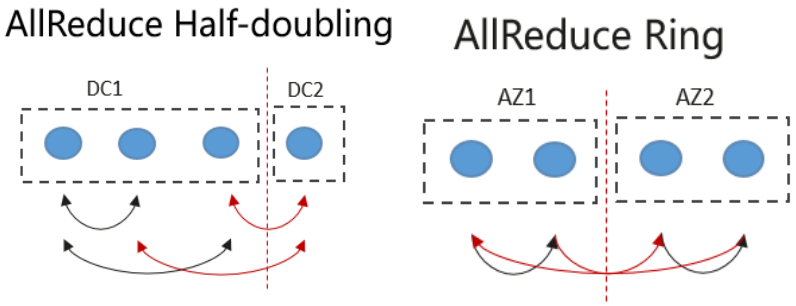
Challenge 1：
Bandwidth oversubscription in AI training network



Requirement：

Minimize traffic amount and communication times on oversubscription networks by optimizing collective communication

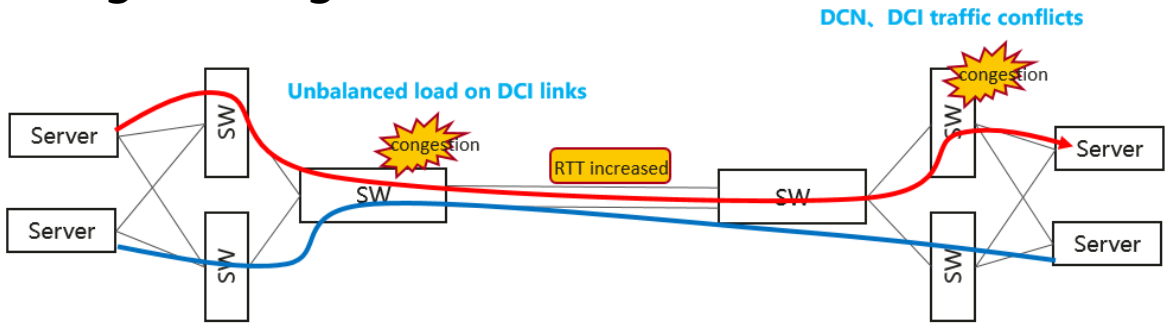
Challenge 2：Asymmetric transmission in collective communication



Requirement：

Model partitioning and collective communication operations are resource-aware(bandwidth, computing power, memory)

Challenge 3：Long distance transmission across DCI links



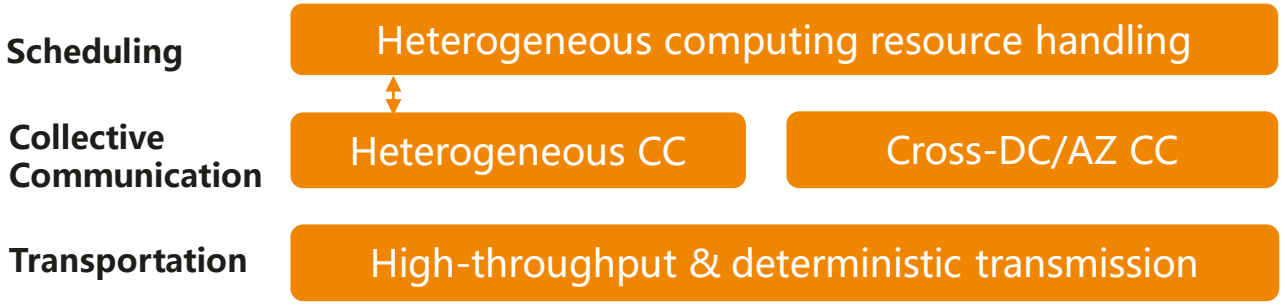
Requirement：

- Load balancing for DCI links
- DCN and DCI traffic classification
- Fast feedback and accurate congestion control mechanisms

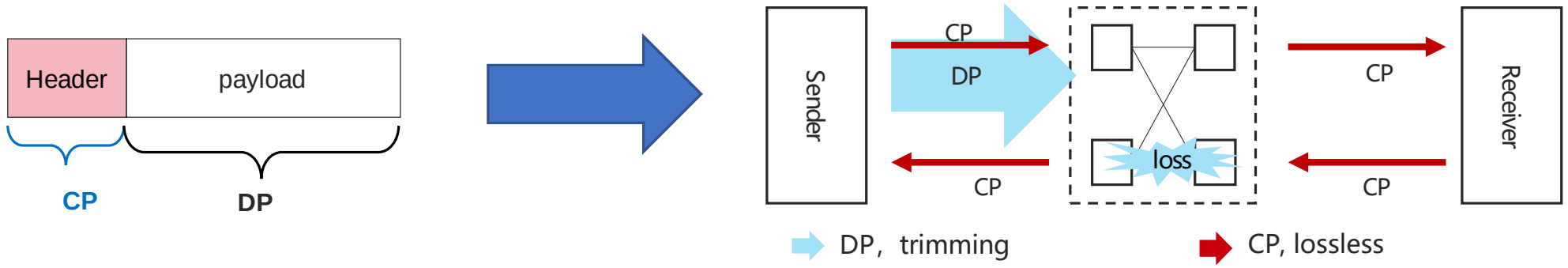
Region Scale AI: Future Technologies

开放数据中心委员会
Open Data Center Committee

To meet the requirements of Region Scale AI, systematic innovations are needed in computing resource scheduling, collective communication, and transmission protocols.



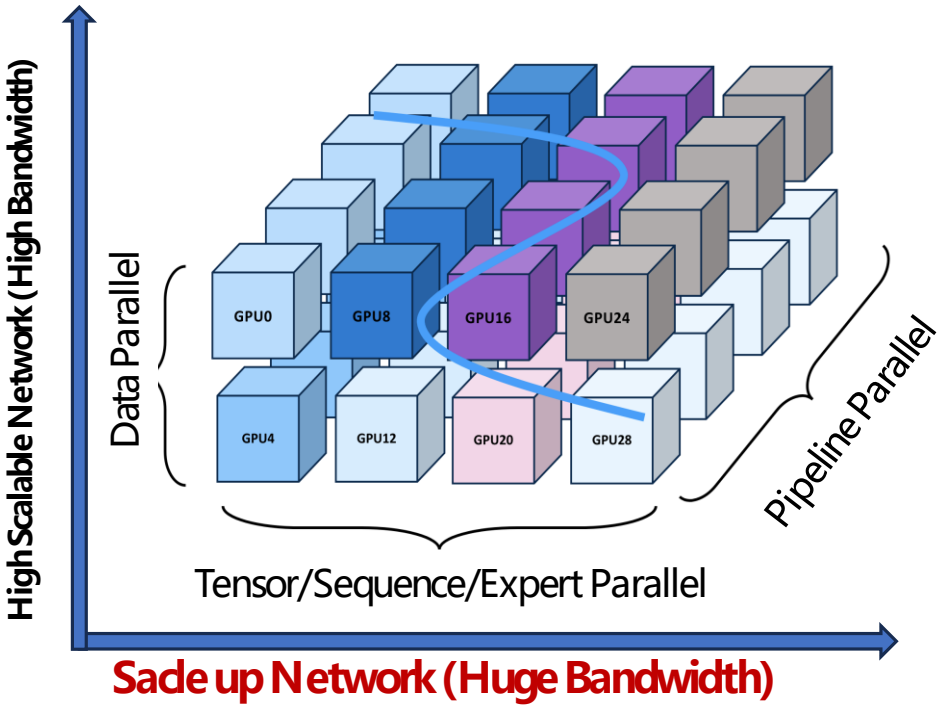
- Resource aware model partitioning and scheduling
- Cross-AZ/cross-DC collective communication operations for DP processing
- Long-distance transmission with high-throughput and determinism
  - Lossless vs. lossy -> Control Plane lossless and Data Plane lossy



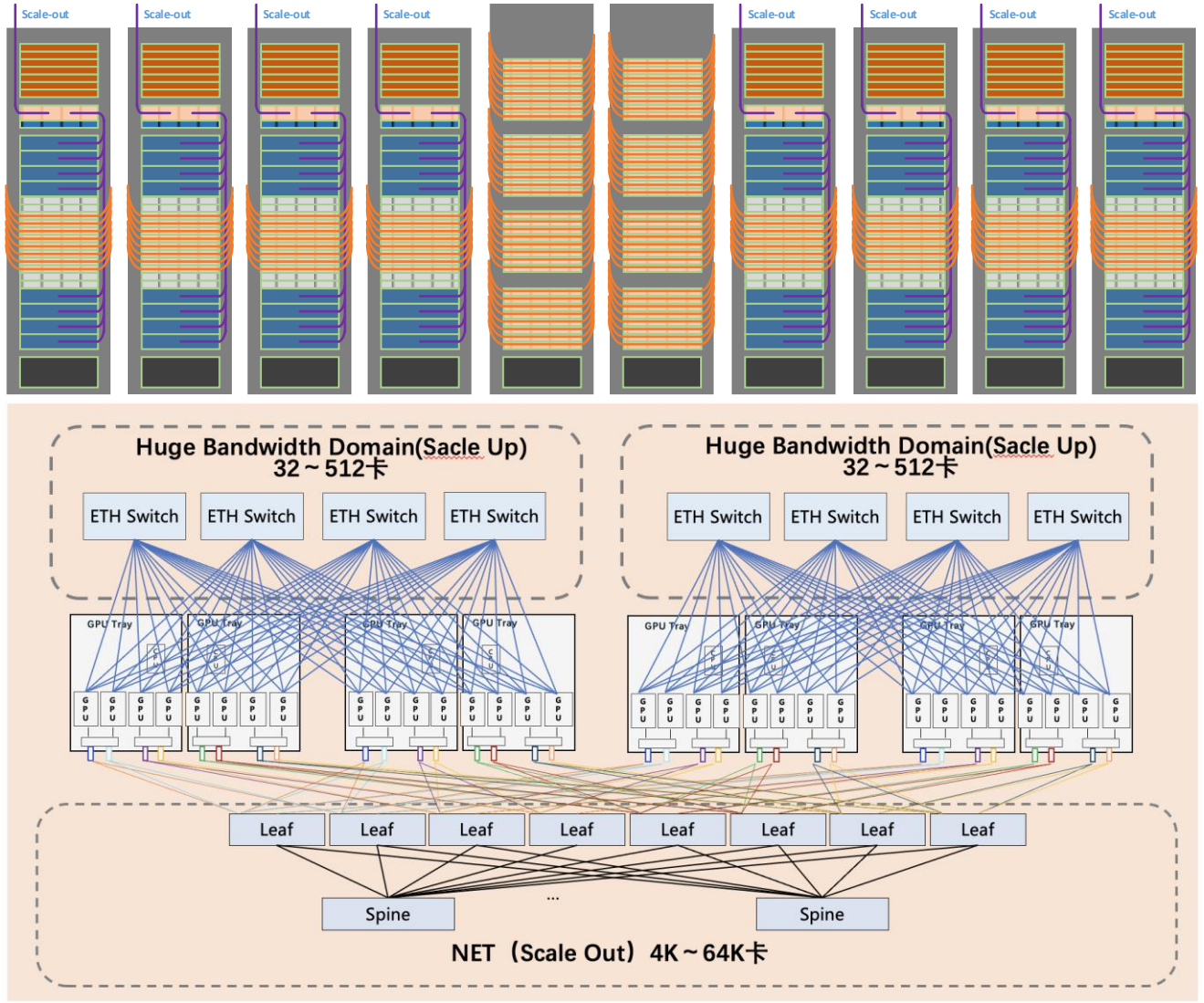
ETH-X Super POD

开放数据中心委员会
Open Data Center Committee

ETH: 以太网



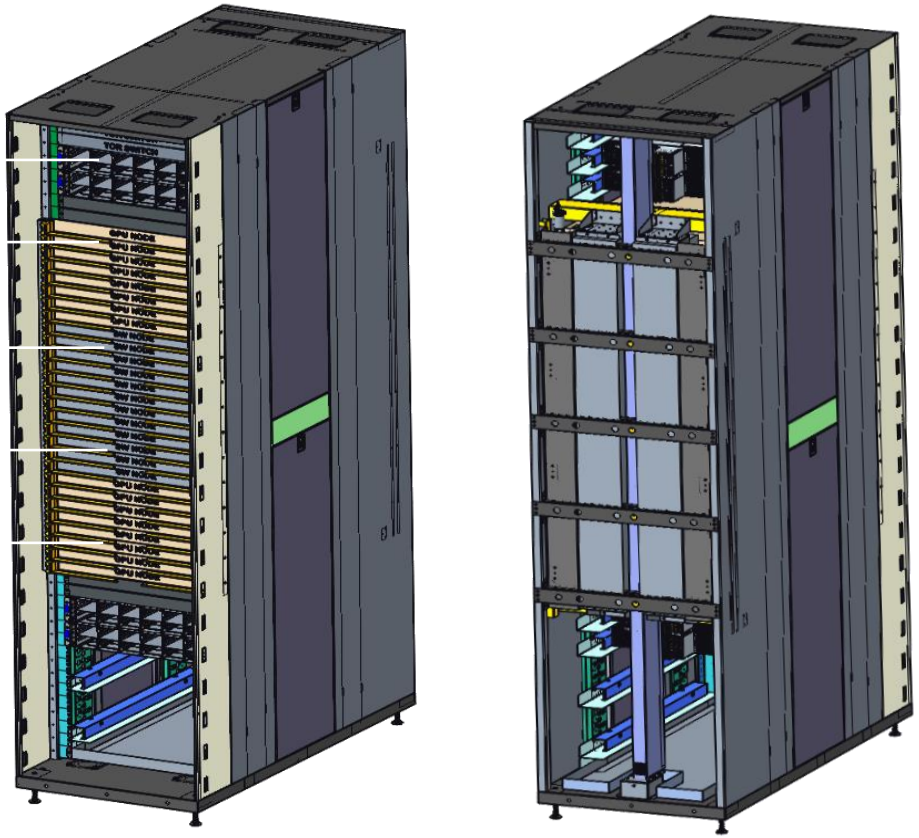
并行策略	数据量(一次iteration)	计算与通信是否可隐藏
DP	GB	大部分可隐藏
PP	GB	大部分可隐藏
TP	TB	大部分不可隐藏
SP	TB	大部分不可隐藏
EP	TB	大部分不可隐藏



建模仿真数据：GPT-110B-16MoE，ETH-128超节点 vs 16台一机8卡，推理TTFT下降 57%，单卡吞吐量提升了59%。

ETH-X

开放数据中心委员会
Open Data Center Committee



ETH-X技术规范名称	进展
rack design specification	Finished
computing node design specification	Finished
switching node design specification	Finished
interconnection design specification	Finished
integration testing specification	Designing
operation and maintenance fault technical specification	Designing
business testing specification	Designing
Scale Up interconnection protocol specification	Designing
Scale Up on network computing specification	Designing

Alink System ： Native Designed for AI Scale Up

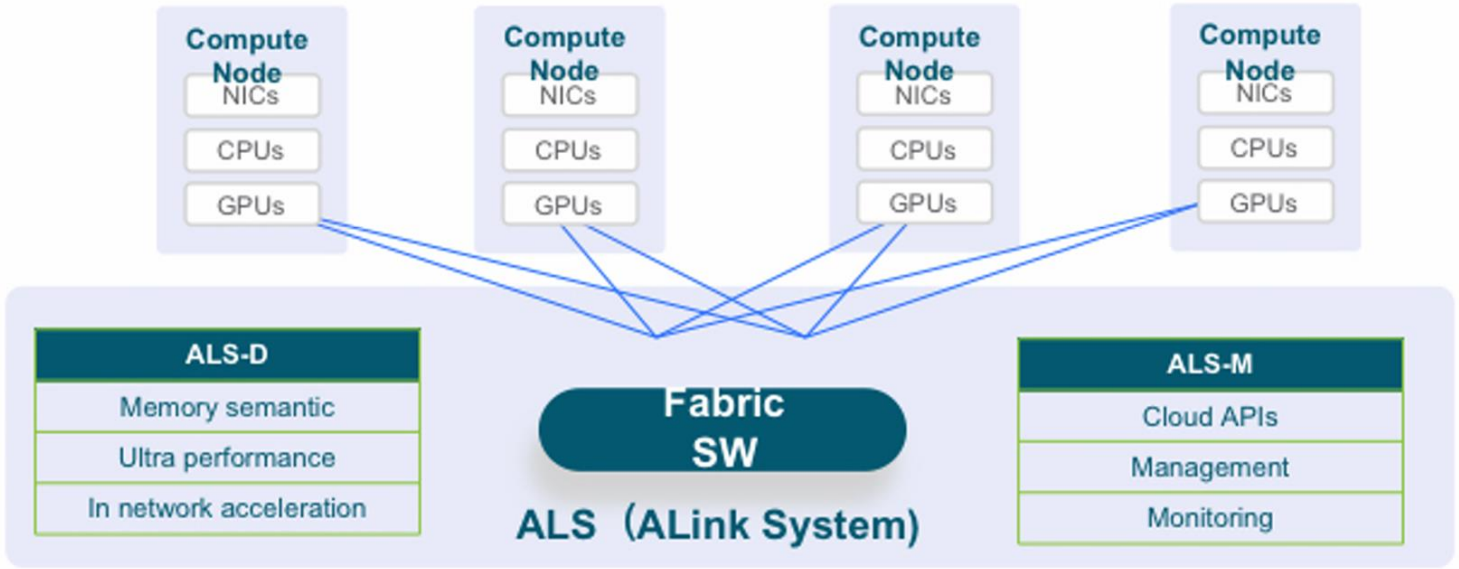
ALS （Alink System）

ALS-D： Data Plane

- UALink （Ultra Accelerator Link ） as Link protocol
- 100+TB/s interconnect bandwidth
- 10+TB level memory fabric
- Single/multiple layer switch topology

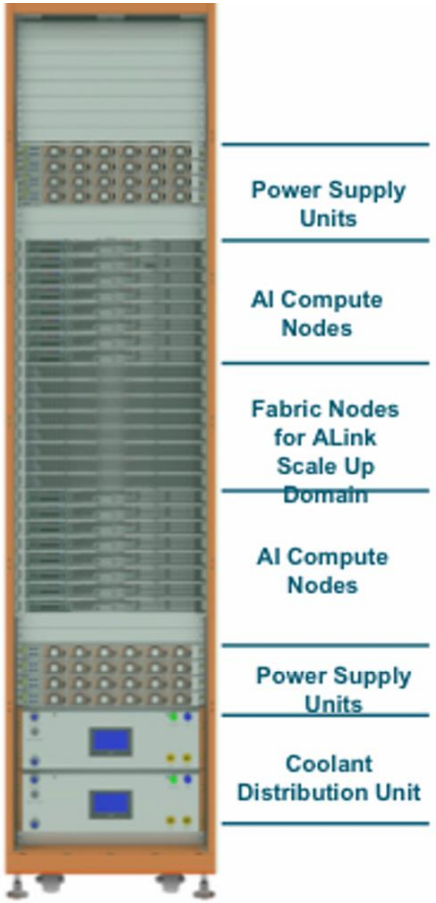
ALS-M： Control Plane

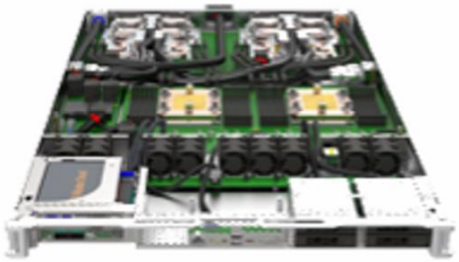
- Standard APIs for Cloud
- Manage different vendor's fabric device
- High flexibility

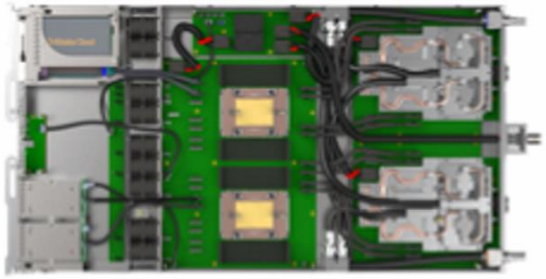


AI infra 2.0 for Alibaba Cloud

开放数据中心委员会
Open Data Center Committee







1 High Performance

Up to 80 AI Accelerators per Rack
200kW per Rack, 2kW per AI Chip
800Gbps Scale Out BW per CIPU

2 High Energy Efficiency

Liquid Cooling
Dynamic CDU, 30% energy saving
HVDC, 98% power efficiency

3 ALink System (Scale-Up Domain)

64/80 accelerators via L1 Fabric
3K+ accelerators via L2 Fabric
PB level memory sharing

Invitation

开放数据中心委员会
Open Data Center Committee

ISCT' 24 in Shanghai, China

THANKS