

智算中心组网技术及应用



Intelligent Computing Center Networking Technology and Applications

段威/DUAN Wei^{1,2}, 于浩/YU Hao², 李和松/LI Hesong^{1,2},
张征/ZHANG Zheng^{1,2}, 王宁诚/WANG Ningcheng^{1,2}

(1. 移动网络和移动多媒体技术全国重点实验室, 中国 深圳 518055;

2. 中兴通讯股份有限公司, 中国 深圳 518057)

(1. State Key Laboratory of Mobile Network and Mobile Multimedia Technol-
ogy, Shenzhen 518055, China;

2. ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202502009

网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20250508.1356.002.html>

网络出版日期: 2025-05-08

收稿日期: 2025-03-05

摘要: 智算中心网络架构正面临高性能、低延迟、高扩展性和可靠性的挑战。当前多种主流的智算中心组网拓扑结构各具优缺点。针对智算中心各种组网方案, 边界网关协议 (BGP)、开放式最短路径优先 (OSPF)、中间系统到中间系统 (IS-IS)、胖树路由协议 (RIFT) 等具有不同适用性。中兴通讯在智能计算中心网络架构与部署方案领域取得突破性进展, 创新性地提出了 Scale-Up 与 Scale-Out 融合组网解决方案, 并通过协议优化提升等价多路径路由 (ECMP) 路径选择精度与故障收敛效率。这些研究和实践可为智算中心网络设计提供参考, 推动智算中心技术的发展和应用。

关键词: 网络协议; 组网技术; 智算中心网络协议; 融合组网

Abstract: The network architecture of intelligent computing centers is facing challenges in terms of high performance, low latency, scalability, and reliability. Currently, there are several mainstream networking topologies for intelligent computing centers, each with its own advantages and disadvantages. Meanwhile, for various networking solutions to intelligent computing centers, protocols such as border gateway protocol (BGP), open shortest path first (OSPF), intermediate system to intermediate system (IS-IS), and routing in fat trees (RIFT) have different applicabilities. ZTE Corporation has made breakthrough progress in intelligent computing center network architecture and deployment solutions, innovatively proposing a converged networking solution integrating both Scale-Up and Scale-Out approaches. By employing protocol optimizations, the precision of equal-cost multi-path (ECMP) path selection and the efficiency of fault convergence have been significantly enhanced. These studies and practices can provide references for the network design of intelligent computing centers and promote the development and application of intelligent computing center technologies.

Keywords: network protocol; networking technology; intelligent computing center network protocol; converged networking

引用格式: 段威, 于浩, 李和松, 等. 智算中心组网技术及应用 [J]. 中兴通讯技术, 2025, 31(2): 63-71. DOI: 10.12142/ZTETJ.202502009

Citation: DUAN W, YU H, LI H S, et al. Intelligent computing center networking technology and applications [J]. ZTE technology journal, 2025, 31(2): 63-71. DOI: 10.12142/ZTETJ.202502009

近年来, 大模型技术迅猛发展^[1-5], 模型能力快速增强。大模型技术已经深入到人们生产生活的方方面面, 人类社会正在快速进入人工智能 (AI) 时代。参数规模是影响模型能力的关键要素, 已经从千亿级别向十万亿级别快速增加。

由于未来需要支撑 10 万亿级别参数规模, 图形处理器 (GPU) 的需求急剧增加。根据摩根士丹利的研究报告, 生成式预训练变换器 GPT-5 至少需要 25 000 个 GPU。为了支撑如此大规模 GPU 卡的应用, 智算中心组网技术已经成为业界研究的热点^[6-7]。

1 智算中心网络拓扑和协议

1.1 智算中心主流网络拓扑介绍

智算中心的网络设计需满足高性能、低延迟、高扩展性

和可靠性的需求。不同业务对网络性能 (带宽、延迟、吞吐量等) 的要求不同, 因此需要选择合适的网络设计。例如, 智算网络需要低延迟和高带宽, 而云服务则需要高扩展性和弹性。此外, 数据中心的规模从几十台服务器到数十万台不等, 网络设计需满足不同的扩展需求。目前, 主流的智算中心组网拓扑结构包括 CLOS、Dragonfly、Torus。每种网络拓扑有各自的特点及应用场景, 在标准组网下也衍生了一些变种组网。

1.1.1 CLOS 网络

CLOS 网络采用多级分布式架构, 分为接入层 (Leaf 层)、汇聚层 (Spine 层) 和核心层 (可选), 具有分层结构、无阻塞性、模块化设计和高带宽利用率等特点。CLOS 网络通过多路径设计实现流量无阻塞转发, 适用于大规模通信负

载。其模块化设计易于扩展，只需增加交换机和链路即可扩展网络规模。

1) 两层CLOS架构

如图1所示，假设交换机端口为 n ，则两层CLOS最大规格可支持 $n^2/2$ 个GPU卡互联，所需交换机的数量为 $N+N/2$ 。

以128口400 Gbit/s的交换机、连接带有8个GPU和8个网口的服务器为例，组网按照无收敛比来设计。单个交换机上行64端口连接Spine交换机，下行64端口可连接8台服务器（每台包含8块GPU卡），单个交换机与8台服务器组成一个机架顶交换机（ToR）区域。Spine层有64个交换机，共连接128个ToR区域。其中，Spine1交换机的128个端口连接每个ToR交换机的1号口，Spine2交换机的128个端口连接每个ToR交换机的2号口。依次类推，Spine64交换机的128个端口连接每个ToR交换机的64号口。两层组网组成一个Pod，该Pod拥有8 192个GPU和192台交换机。

2) 三层CLOS架构

三层CLOS架构是在二层架构的基础上增加Core层交换机而形成的。通过Core层交换机，三层CLOS架构能够实现Spine层交换机的互联。假设交换机端口为 n ，则三层CLOS最多可支持 $2 \times (n/2)^3$ 个GPU卡互联，所需交换机的数量为 $N^2+(N/2)^2$ 。

以128口400 Gbit/s的交换机、连接带有8个GPU和8个网口的服务器为例，组网按照无收敛比设计。对于ToR区域，与两层组网相同，单个Leaf交换机下挂8机共64卡GPU。对于Spine层，每个交换机需要有一半端口向上连接第三层，因此剩余64下联口向下连接64个ToR，以便组成一个集群Pod。其中，每个Pod拥有4 096个GPU卡、128台

交换机、4 096个上联口。

第三层Core交换机共有128个下联口连接Pod，每个交换机连接每个Pod中具有相同编号的上联口。例如：Core1交换机连接Pod1~Pod128中Spine1的1口，Core2交换机连接Pod1~Pod128中Spine1的2口，依此类推，Core4096交换机连接Pod1~Pod128中Spine64的64口。因此，Core层共需要4 096台交换机。三层组网可支持 $2 \times (128/2)^3=524 288$ 个GPU，共需要20 480台交换机。

1.1.2 多轨CLOS网络

多轨CLOS网络在传统CLOS的基础上增加了多轨设计，支持更高的并发通信性能和带宽利用率。其特点包括多轨设计、低延迟、高容错和简化布线，适合超大规模网络。多轨设计通过独立的数据路径实现流量的分布式转发，显著提高了带宽利用率和容错能力。典型的组网应用如阿里的高性能网络（HPN）^[8]、字节的超大规模网络Megescale^[9]、百度的自研数据中心等^[10]。

1) 两层多轨CLOS架构

如图2所示，假设交换机端口为 N ，单台服务器GPU数为 M ($M < N$)，则两层CLOS多轨可支持 $2 \times (N/2)^2 \times M$ 个GPU卡互联，所需交换机的数量为 $(N+N/2) \times M$ 。

以128口400 Gbit/s的交换机、连接带有8个GPU和8个网口服务器为例，组网按照无收敛比设计。在多轨CLOS架构的设计中，由于每个服务器的同号GPU与同一个交换机连接，因此对于一个ToR来说，8卡服务器就需要8台交换机（Leaf1~Leaf8），每台交换机64口上联Spine层，64口下联服务器网卡。其中，Leaf1交换机1口接服务器1的GPU1，

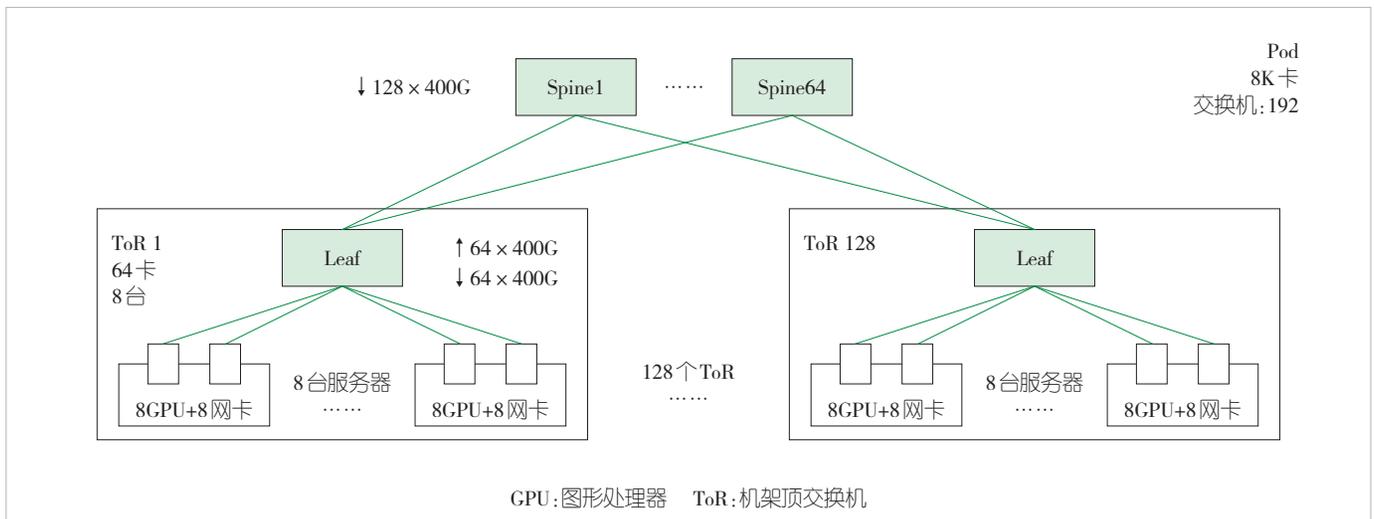


图1 两层CLOS组网架构

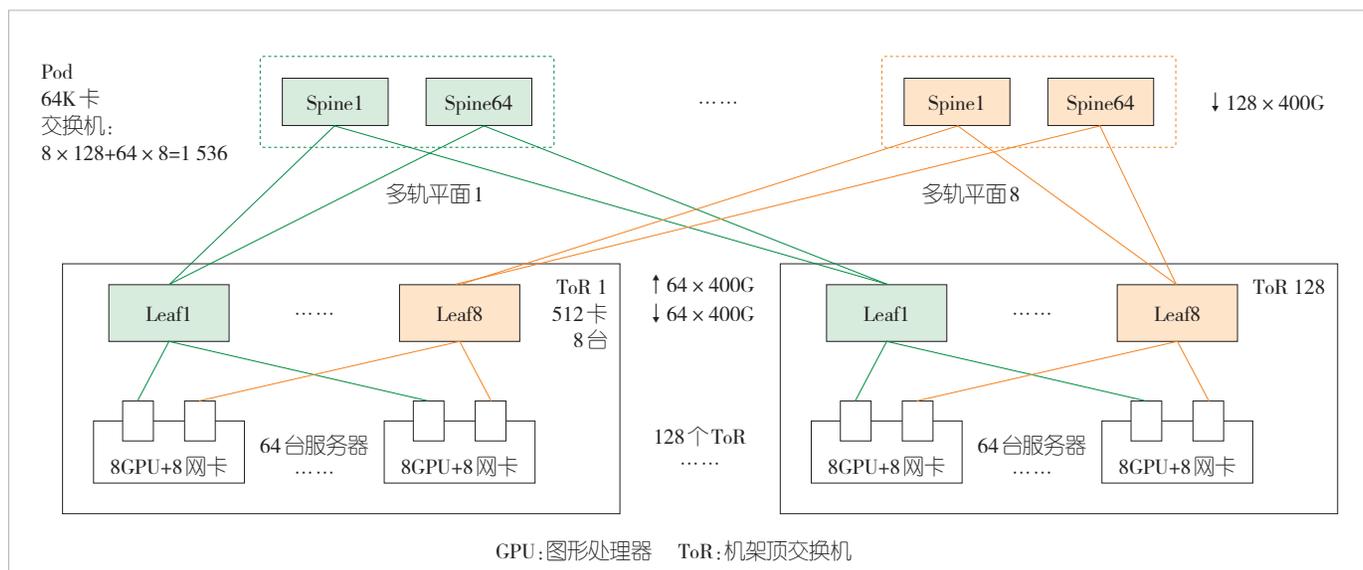


图2 两层多轨CLOS组网架构

2口接服务器2的GPU1, 依此类推, 64口接服务器64的GPU1; Leaf2交换机连接所有服务器的GPU2, Leaf8连接所有服务器的GPU8。因此, 单个ToR共有8台交换机, 64台服务器共有512卡。

网络第二层也是按平面进行划分的, 共有8个平面。对于单个平面而言, 其与传统CLOS网络相同, Leaf层交换机上联64口分别接Spine层64台交换机, Spine层每个交换机128口可以连接128个ToR。两层网络组成一个Pod, 共有 $2 \times (128/2)^2 \times 8 = 65536$ 个GPU卡, 交换机数量为 $(128 + 128/2) \times 8 = 1536$ 台。

2) 三层多轨CLOS架构

三层多轨CLOS架构增加Core层交换机, 使Spine交换机按照轨道划分进行互联。假设交换机端口为 N , 单台服务器GPU数为 M ($M < N$), 则三层多轨CLOS可支持 $2 \times (N/2)^3 \times M$ 个GPU卡互联, 所需交换机数量为 $[N^2 + (N/2)^2] \times M$ 。

以128口400 Gbit/s的交换机、连接8GPU 8网口服务器为例, 组网按照无收敛比设计。三层CLOS多轨ToR与两层相同。Spine层交换机需要有一半端口上接Core层, 下接的ToR区域数量为两层网络的一半, 即64个ToR区域。因此, 单个Pod可连接32768个GPU卡、1024台交换机。

第三层网络与第二层网络类似, 也采用分平面方式进行组网, 共有8个平面, 且每个平面的第三层交换机共有4096台。该层网络的连接方式与普通CLOS网络一样, 相同编号的GPU连接在同一个交换机上。因此, 三层多轨CLOS网络最大可支持 $2 \times (128/2)^3 \times 8 = 4194304$ 个GPU卡, 需要的交换机数量为 $(128^2 + (128/2)^2) \times 8 = 163840$ 。

1.1.3 Dragonfly 网络

Dragonfly是当前应用最广泛的直连拓扑网络架构^[11]。本地通信通过组内交换机完成, 组间通信通过高效的全局链路完成, 减少了传统分层结构中的多跳路径。Dragonfly的拓扑结构分为3层: Switch层、Group层、System层。

Switch层: 包含1个交换机及其相连的 p 个计算节点。

Group层: 包含 a 个Switch层, 这 a 个Switch层的 a 个交换机是全连接(All-to-all)的。换言之, 每个交换机都有 $a-1$ 条链路分别连接到其他的 $a-1$ 台交换机。

System层: 包含 g 个Group层, 这 g 个Group层也具有全连接的特点。

对于单个Switch交换机, 它有 p 个端口连接到计算节点, $a-1$ 个端口连接到Group内其他交换机, h 个端口连接到其他Group的交换机。因此, 我们可以通过计算得到网络的如下属性:

- 每个交换机的端口数为 $k = p + (a-1) + h$;
- Group的数量为 $g = ah + 1$;
- 网络中计算节点数共有 $N = ap(ah + 1)$ 个;
- 如果我们把一个Group内的所有交换机视为一个总交换机, 那么这个交换机的端口数为 $k' = a(p + h)$ 。

1.1.4 Dragonfly+网络

Dragonfly+网络是在Dragonfly网络基础上进一步优化的智算中心网络和数据中心网络架构^[12], 它将Group内的通信结构变为了Fat-tree。相较于传统的Dragonfly网络, Dragonfly+主要在拓扑结构、全局链路分布、路由策略以及可扩展性方面进行了改进。Dragonfly+的组网方式与Dragonfly基

本相同，其基本组网结构如下：

节点分组（Pod）：网络被划分为多个组，每组由若干交换机和计算节点组成，组内可采用CLOS架构。

组内通信：组内交换机和节点通过高带宽连接实现低延迟通信。

组间通信：使用分层全局链路连接各组，链路被分为近邻和远程两种。

Dragonfly+组网架构可以灵活分配组内通信和组间通信的连接数量，以适应不同规模的数据中心网络。在最大拓扑结构下，每对组都由一个全局链接连接，从而形成保留Dragonfly属性的最大拓扑。在中型拓扑结构下，每个骨干路由器都通过一个全局链路连接到其他所有组。在小型拓扑结构下，每个脊椎路由器都通过几个并行链路连接到其他组。Dragonfly+最大拓扑组网如图3所示。

组内网络采用Leaf-Spine两层CLOS架构。其中Leaf与Spine采用全连接方案，每个Leaf交换机都连接到组中的 p 个主机和 s 个Spine交换机，每个Spine交换机都连接到同组的 l 个叶子路由器和其他分组的Spine交换机。为了保证组内网络性能，一般建议： $p=l=s=h$ 。假设交换机的端口数为 k ，此时 $p=h=k/2$ 。单个组内的主机数量为 $N=p \times l=k^2/2$ 。

1.1.5 Torus 网络

Torus 网络采用网格化设计，通常用于计算芯片直连。典型的Torus网络有Google张量处理器TPU V4^[13]。Torus网络的节点排列表现为一维、二维或三维结构^[14]，具体特点如下：

一维Torus：所有节点连接成一个环状拓扑；

二维Torus：节点排列成二维网格，行列两端节点通过额外链路闭合；

三维Torus：节点扩展为三维网格，每个维度的边界闭合，形成环状结构，如图4所示。

1.1.6 网络拓扑对比分析

各类组网架构根据性能需求、延迟要求、扩展性和成本进行选择，以应对不同的数据中心和计算任务需求。表1展示了各种组网架构的优缺点。

1.2 智算中心网络协议设计

如何在多种智算中心组网架构上实现高效路由是智算中心设计面临的关键问题之一。传统路由协议模式分为集中式、分布式和混合式。其中，集中式路由是指引入软件定义网络（SDN）控制器，由控制器下发流表并指导交换机转发的方式。分布式路由是指利用动态路由协议，让所有交换机自主交互路由信息。各个交换机自主计算路由，得出整网的无环路由表以供转发使用。这里所指的动态路由协议包括边界网关协议（BGP）、开放式最短路径优先（OSPF）、中间系统到中间系统（IS-IS）、胖树路由协议（RIFT）、路由信息协议RIP等。混合式路由是指，将SDN控制器与分布式路由协议相结合，利用各自的优势进行部署的模式。

在现有智算中心中，主流趋势是采用分布式路由模式。在一些需要精细化控制的场景中，混合式的路由模式也有一定应用。然而，无论是分布式模式还是混合式模式，动态路由协议都是其中的核心组成部分。路由协议的对比如表2所示。

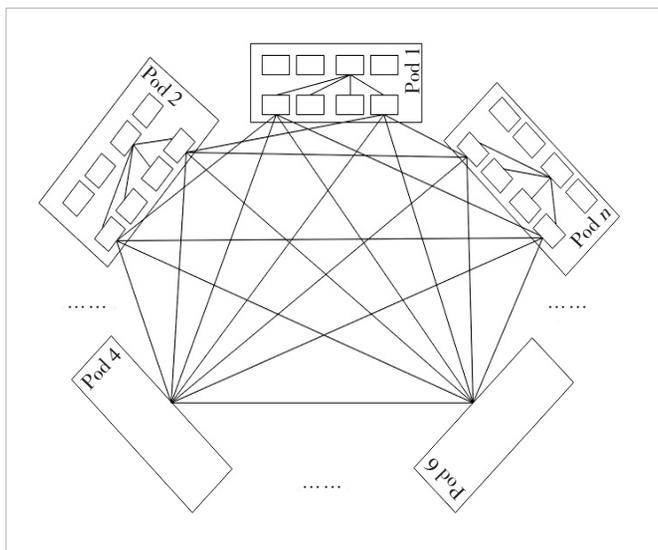


图3 Dragonfly+组网架构

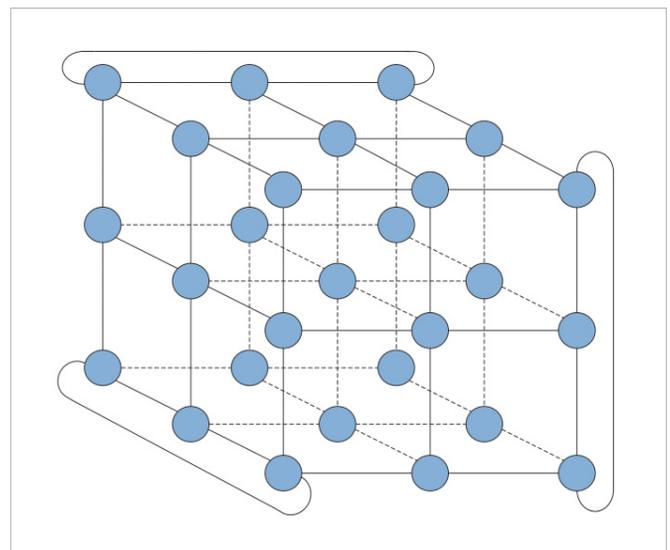


图4 三维Torus组网架构

表1 智算中心网络拓扑对比

网络架构	优点	缺点
CLOS	扩展性强,带宽利用率高,无阻塞性,商用成熟度高	布线复杂,成本和能耗高
Dragonfly	低延迟、高性能,可实现全局通信优化	部署复杂,不适合小规模网络
Dragonfly+	链路利用率更高,灵活性强,超大规模扩展,容错能力强	拓扑复杂,部署成本高,非典型流量模式适应性弱
Torus	低延迟、低成本,拥有路径冗余保护	扩展性差,灵活性低,性能不足
单机多轨	性能更高,容错能力强,布线简化,扩展灵活	布线复杂,成本高,功耗较高

表2 智算中心网络协议对比

内容	BGP	OSPF	IS-IS	RIFT
协议类型	路径矢量协议	链路状态协议	链路状态协议	链路状态+距离矢量
最初设计目标	跨自治系统(AS)互联	企业级内部网络路由	企业级内部网络路由/运营商骨干网	大规模 CLOS/Spine-Leaf 架构
分层能力	基于AS分层	基于Area分层	基于Level分层	基于Level分层
收敛速度	较慢(依赖定时器)	快(泛洪+SPF计算)	快(泛洪+SPF计算)	极快(优化泛洪与计算机制)
扩展性	极高(支持超大规模网络)	中等(Area规划复杂度高)	高(基于Level规划)	极高(专为CLOS架构优化)
策略控制	极强	较弱	较弱	强
支持零接触部署(ZTP)	否	否	否	是
适用架构	大规模平面网络	传统三层架构	CLOS/Fat-tree/运营商架构	CLOS/Fat-tree
典型应用场景	超大规模数据中心	中小型数据中心	运营商/电信级数据中心	新一代数据中心

BGP:边界网关协议 IS-IS:中间系统到中间系统 OSPF:开放式最短路径优先 RIFT:胖树路由协议 SPF:最短路径优先

在数据中心网络设计中,动态路由协议的选择需综合考虑网络规模、拓扑结构、扩展性、收敛速度和运维复杂度等因素。

- 从架构角度考虑:在数据中心选择使用CLOS/Fat-tree架构时,可优先考虑RIFT、IS-IS或者BGP协议。如果是传统的三层架构,仍然可以考虑使用OSPF和IS-IS协议。

- 从规模角度考虑:如果是小型数据中心(节点数量小于500),可考虑使用OSPF协议;如果是中型数据中心(节点数量为500~5000),则可考虑IS-IS、RIFT或者BGP协议;如果是大型数据中心(节点数量在5000以上),则可考虑BGP或者RIFT协议。

- 从大规模运维难度考虑:如果管理能力强,可考虑使用BGP协议;如果非常需要自动运维的能力,则可以考虑RIFT协议。

尽管上述拓扑结构和协议为智算中心组网提供了多样化选择,但在超大规模部署场景中仍面临挑战:CLOS架构的布线成本与能耗随规模增长呈指数级上升,Dragonfly+的复杂拓扑增加了运维难度,传统协议在等价多路径路由(ECMP)路径动态优化方面存在响应滞后问题。此外,当前组网方案普遍采用机内(Scale-Up)与机间(Scale-Out)网络独立部署模式,导致资源利用率低下、运维复杂度激增。针对这些挑战,我们提出面向超大规模智算中心的融合

组网架构与协议创新方案,通过技术整合突破传统架构的局限性。

2 中兴通讯智算中心组网方案和协议创新

当前智算中心商用规模部署的组网方案以CLOS为主。该组网方案凭借高性能、低延迟、强扩展性和高可靠性优势,已成为现代智算中心大规模网络设计的重要选择。CLOS也是当前中兴通讯推荐的主流组网方案。基于该组网方案,我们在网络架构优化、传输效率提升等方面不断进行创新和改进,以更好地满足智算中心日益增长的业务需求。

2.1 Scale-Up和Scale-Out融合组网架构

在现有智算中心网络架构中,Scale-Up网络和Scale-Out网络是两类最重要的后端网络。Scale-Up网络(如NV-Link)与Scale-Out网络(如CLOS)通常采用独立部署方案,存在成本、性能和运维三方面核心矛盾。针对这一问题,我们提出Scale-Up与Scale-Out融合组网架构,通过协议层与拓扑层的协同设计实现资源的统一管理。

Scale-Up网络的功能定位是构建更大的逻辑GPU,旨在解决单个计算节点或者超节点内GPU之间的高速互联。典型的实现方式是英伟达通过NVLink/NVSwitch实现最大576个GPU之间的高速无阻塞互联。与Scale-Out网络相比,Scale-Up网络具有如下3个特征:

1) 超高带宽

Scale-Up 网络对带宽的要求比 Scale-Out 高出一个数量级。以 GB200 为例，Scale-Up 网络内 GPU 点到点通信带宽为双向 1.8 TB/s，对应 Scale-Out 网络带宽为单向 800 Gbit/s。Scale-Up 网络带宽是 Scale-Out 网络带宽的 9 倍。

2) 超低延迟

由于需要支持 Load/Store 同步内存语义，Scale-Up 网络是一种延迟敏感型网络，要求纳秒级的网络延迟。相对而言，Scale-Out 的网络延迟需求通常在微秒级。以 Ualink 为例，它要求用于 Scale-Up 场景的交换芯片延迟控制在 300 ns 以内。这一指标与当前 NVSwitch 基本持平。

3) 有限规模

Scale-Up 网络另外一个重要的特征是组网规模有限。从 AI 基础设施的发展趋势来看，未来几年 Scale-Up 网络的规模将维持在千卡以下。这与 Scale-Out 网络动辄要求万卡甚至百万卡形成鲜明的对比。

总体而言，Scale-Up 网络侧重性能，Scale-Out 网络侧重规模，两者密切协同完成一个高效 AI 集群的构建。在典型的 AI 大模型训练和推理过程中，通常会设计复杂的混合并行策略，根据并行流量的特征差异综合利用 Scale-Up 和 Scale-Out 的网络资源，例如：将对带宽要求极高的张量并行 (TP) 流量和对延迟要求较高的控制流量规划到 Scale-Up 通信域，将对带宽和延迟不太敏感的流水线并行 (PP) 和数据并行 (DP) 流量疏导到 Scale-Out 网络中。随着 AI 模型参数规模和模型架构的不断演进，并行策略变得日益复杂，但这种模型切分策略和网络的最优匹配逻辑却一直存在。

正是由于存在上述明显的差异性，目前 Scale-Out 和 Scale-Up 被建设为两张独立的网络。这两张网络存在多种技术路线，异构特征明显。具体而言，Scale-Out 网络的技术路线分为无限带宽 (IB) 和以太网两个封闭的产业生态。其中，IB 以英伟达为主，而以太网被大部分厂商所支持。Scale-Up 网络的技术实现则更加多样化，总体而言可以分为总线网络化和网络总线化两大类。其中，总线网络化是指，以传统总线型协议 (如 PCIe) 为基础，通过吸收部分以太网的底层技术来满足 Scale-Up 网络的需求，典型的实现方式如 NVLink 和 Ualink；网络总线化是指，以传统网络协议 (如以太网) 为基础，通过引入总线协议的设计来满足 Scale-Up 网络的需求，典型的实现方式如美国的基于以太网的特斯拉传输协议 TTPoE、中国的高通量以太网等。总体而言，当前 Scale-Up 和 Scale-Out 独立组网仍存在如下缺陷：

1) 建设成本高昂：两张独立的网络导致资源利用率不

高，为了达到最优的组合性能，通常只能从一家供应商购买，进而导致整体建设成本高昂。

2) 管理运维困难：运维人员需要同时熟悉两套技术栈，故障定位、故障恢复难度将成倍增加，导致运营成本居高不下。

为了解决上述组网痛点问题，基于对未来智算互联行业趋势的准确把握，中兴通讯创新性地提出了 Scale-Out 与 Scale-Up 融合组网的理念和对应的产品解决方案，总体特征如下：

1) 统一互联协议

针对 Scale-Out 和 Scale-Up 场景的需求，中兴通讯基于以太网技术为基础，构建了开放互联协议 OLink 技术栈。该技术栈在底层兼容以太网的基础上，通过物理层和事务层的创新设计，配合极简的端侧互联网协议 (IP) 设计，使系统既可以满足 Scale-Out 网络超大规模扩展的需求，也可满足 Scale-Up 网络高性能互联的需求。

2) 融合极简组网

由于底层协议都被统一到以太网，在不影响业务性能的前提下，Scale-Up 和 Scale-Out 网络可以融合形成一张物理网络。这样在极致组网场景中可以不再需要独立的数据处理单元 (DPU) 网卡，从而大幅降低整体组网的复杂性和成本。该组网方式通过灵活的服务质量 (QoS) 策略实现不同业务流量之间的性能隔离。

3) 降低组网成本。当 Scale-Up 和 Scale-Out 融合后，原来分布在 GPU 和 DPU 的端侧功能被集中到 GPU 侧。此时硬件构成变得更为简单 (不再需要独立的 DPU 网卡和配套的 PCIe Switch)，管理运维也进一步统一，整体组网成本大幅下降。

图 5 展示了中兴通讯基于 OLink 技术推出的融合组网架构。GPU 集成 OLink 提供的端侧 IP，与支持 OLink 协议的交换芯片结合，实现统一的 Scale-Out 和 Scale-Up 通信。在超节点 (典型如 64 卡) 内的 GPU 通过第一级交换节点实现高带宽通信域，并通过 OLink 协议实现低延迟、高带宽、内存语义以及在网计算等与 Scale-Up 相关的功能。超节点之间通过两级甚至三级交换实现大规模组网，满足 Scale-Out 通信的基本要求。Scale-Up 和 Scale-Out 网络在 GPU 和第一级交换机上物理共享，但在协议上存在差异。

基于以太网技术的 Scale-Up 和 Scale-Out 融合组网不是一蹴而就的，它需要在传统以太网的基础上针对场景需求做大量的技术创新。业界一直都有总线网络化和网络总线化两种探索路径。Ualink 是典型的总线网络化技术路线，而 OLink 则选择了网络总线化这种具有更强演进能力的技术路线，两者的对比如表 3 所示。从场景需求来看，Scale-Up 是一个介于片上网络 (NOC) 总线和网络 Fabric 之间的互联场

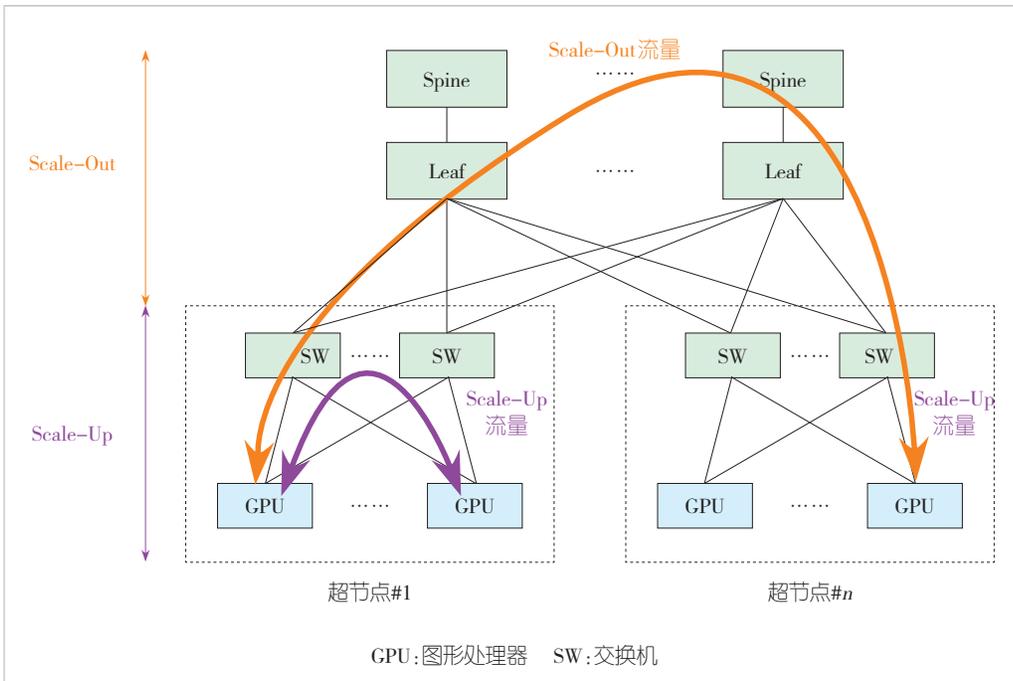


图5 中兴OLink融合组网示意图

景，要求满足TB级超高互联带宽、K级互联规模、原生内存语义、低功耗、纳秒级延迟以及统一内存编址的需求。以太网在互联带宽、互联规模等方面相比于传统总线技术有明显优势，但在其他几个指标维度则存在不足。OLink借助Transaction Packing、芯片PPA（指性能、功耗、面积）优化、低时延前向纠错（FEC）和定制化报文格式等技术措施，充分满足Scale-Up在内存语义、功耗、低延迟以及内存统一编址方面的需求。在技术指标上，OLink与Ualink殊

途同归，但依托以太网快速发展的产业生态，OLink在降低成本、提升兼容性等方面更具优势。总体而言，OLink是中兴通讯长期基于以太网底层技术创新提出的面向未来AI基础设施的互联解决方案。作为Scale-Up和Scale-Out融合组网的技术倡导者和产品践行者，中兴通讯将依托OLink方案，持续深化技术迭代与生态共建，加速AI基础设施网络架构的升级演进。

2.2 基于BGP扩展的ECMP动态优化

智算网络（如AI训练/推理场景）对网络的低时延、高带宽和零丢包特性提出了极高要求。尽管第1章所述的RIFT协议在CLOS网络中实现了快速收敛，但其ECMP路径选择仍然依赖静态负载均衡策略，无法感知链路的实时状态（如拥塞或故障）。在智算中心万卡级集群中，此类静态策略可能导致AI训练任务因单路径故障而中断。为解决这一问题，我们提出基于扩展BGP协议的动态路径选择机制，通过链路状态感知实现ECMP路

表3 OLink和Ualink对比分析

内容	Ualink	OLink	说明
生态现状	头部互联网厂商和非英伟达技术体系支持	中兴通讯主导，与部分图形处理器（GPU）客户达成合作意向	中国也有Ualink，阿里集团在积极推动
场景定位	聚焦1K GPU之间Scale-Up互联场景	考虑Scale-Up和Scale-Out融合场景	融合的代价是需要更多地遵循以太网标准，其中OLink支持更大的组网规模
协议架构	按照典型总线型协议设计，基于时钟同步并行传输	基于以太网协议设计，以异步消息通信为主	两者的协议架构思路不同，OLink能够共享以太网生态红利
交换延迟	< 300 ns	< 500 ns	Ualink时延性能优于OLink，但可通过计算和通信重叠消除性能损耗
链路级可靠性	支持链路级重传(LLR)	支持LLR	当前Ualink 1.0版本LLR待完善
原生内存语义加载(LD)/存储(ST)	原生支持高并发LD/ST语义	通过Packing方式间接支持LD/ST	Ualink的LD/ST性能优于OLink，OLink可针对特定场景优化性能
直接内存访问(DMA)语义	协议未对DMA进行描述，但技术上支持	原生支持DMA语义	Ualink虽然支持DMA性能，但可能会受线缆包长过小的影响
在网计算	1.0版本不支持在网计算，后期版本可支持	支持在网计算	Ualink将在网计算的需求放在1.0版本之后

径的智能负载均衡。

在大规模智算网络场景下，当使用三层CLOS网络架构时，多路径的问题会凸显。然而，传统ECMP技术依赖静态负载均衡策略（如五元组哈希），缺乏对链路实时状态的感知能力，导致流量可能持续转发至拥塞路径，引发AI任务中断。为解决这一问题，我们提出一种基于扩展BGP协议的链路状态感知机制，通过动态通告链路状态信息，实现ECMP路径的智能选择，显著提升网络可靠性和AI任务效率。相较于传统BGP协议仅通告下一跳信息，本方案通过下一跳能力（NHC）扩展TLV（指类型、长度、值格式）字段引入下一跳描述符，使Leaf设备可实时感知Spine-Leaf链路状态（如带宽利用率与故障状态），从而突破第1章所述协议在动态响应能力上的局限。

Spine/Core设备通过在BGP路由通告中引入NHC属性的扩展TLV，通告下一跳和链路描述符，可选携带带宽信息。在链路出现故障或者拥塞（带宽利用率达到一定程度）时，Spine/Core设备将发送控制面或数据面通知消息给Leaf/Spine设备，包括相应的链路描述符以及具体的故障或者拥塞通知。

Leaf/Spine设备在收到Spine/Core设备发过来的协议通告时，建立下一跳、下一跳、链路描述符及链路状态表，以监控下一跳与下一跳之间的链路状态。在收到状态通知消息时，Leaf/Spine设备会更新该表项，并在转发流量时参考该链路状态表决定下一跳设备的选择。为了实现快速的故障恢复，Spine设备在进行故障或者链路带宽预警（预示可能出现拥塞）通告时，通过数据面进行快速通告。需要说明的是，这里的数据面可以基于用户数据报协议（UDP），也可以基于其他类型的IP。Leaf设备可以直接基于链路描述符更新如表1所示的链路状态，以进一步提高处理效率。需要注意的是，当故障链路恢复正常或者拥塞链路不再拥塞时，Spine设备同样需要通告Leaf设备该信息。通告方式可以采用更新属性的方式，以便恢复链路被重新投入使用。

为了避免流量振荡，Leaf设备可以预先向Spine设备发出链路状态订阅消息，即表示自身对某个链路的特别关注。Spine设备在发现该链路出现故障或者拥塞时，直接将报文通知发送给订阅该消息的Leaf设备，实现Leaf设备对该故障的快速反应。这种订阅方式还适用于其他场景，比如：Spine设备在缺省情况下，会在链路带宽占用率达到80%的时候发出带宽告警消息。但某个Leaf设备可能在链路带宽占用率达到50%时就需要得到带宽告警消息。因此，Leaf设备在发送针对链路描述符的订阅信息时，可以携带其要求，比

如要求带宽利用率达到50%时就需要得到通知。Spine设备将记录下这些需求，并在达到这些要求时发送通知消息。

本研究提出的协议创新通过链路状态感知与动态路径选择，解决了智算网络中ECMP路径拥塞与故障导致的AI任务中断问题。实验表明，该机制可降低丢包率90%以上，显著提升大规模智算任务的效率。未来我们将探索与SRv6、Inband Telemetry的深度集成，进一步提升网络智能化水平。

3 结束语

智算中心网络作为大模型训练的核心基础设施，其架构设计与技术演进直接影响AI算力的释放效能。本文从拓扑结构、协议优化与组网融合3个维度系统分析了智算网络的演进路径：在拓扑层面，CLOS架构通过分层扩展能力支撑超大规模GPU集群，Dragonfly+/多轨CLOS等变体技术有效平衡了带宽、时延与部署成本；在协议层面，基于扩展BGP的链路状态感知机制，显著提升了ECMP路径选择精度与故障收敛效率。我们在组网方案中提出的OLink融合架构，通过统一以太网协议栈实现Scale-Up与Scale-Out网络的深度整合，突破了传统异构组网瓶颈。实测数据表明，该架构可大幅提升万卡级集群的通信效率。这些技术创新为智算中心构建高带宽、低时延、强扩展性的网络底座提供了实践范例。未来，随着智算中心规模的进一步扩大和技术的不断演进，中兴通讯将持续探索更加高效、灵活的组网技术，以推动智算中心技术的持续发展，为AI和高性能计算的广泛应用提供坚实的技术支撑。

参考文献

- [1] 田海东, 张明政, 常锐, 等. 大模型训练技术综述 [J]. 中兴通讯技术, 2024, 30(2): 21-28. DOI: 10.12142/ZTETJ.202402004
- [2] 唐宏, 武娟, 徐晓青, 等. 生成式大模型承载网络架构与关键技术探索 [J]. 中兴通讯技术, 2024, 30(2): 50-55. DOI: 10.12142/ZTETJ.202402008
- [3] REN T Q, LI R P, ZHAO M M, et al. Separate source channel coding is still what you need: an LLM-based rethinking [J]. ZTE communications, 2025, 23(1): 30-44. DOI: 10.12142/ZTECOM.202501005
- [4] 裴丹, 张圣林, 孙永谦, 等. 大语言模型时代的智能运维 [J]. 中兴通讯技术, 2024, 30(2): 56-62. DOI: 10.12142/ZTETJ.202402009
- [5] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30(2): 76-88. DOI: 10.12142/ZTETJ.202402012
- [6] 段威, 李和松, 周昆. 智算中心网络技术发展与应用 [J]. 中兴通讯技术, 2024, 30(6): 39-47. DOI: 10.12142/ZTETJ.202406007
- [7] 段晓东, 程伟强, 王瑞雪, 等. 面向新型智算中心的全调度以太网技术 [J]. 中兴通讯技术, 2023, 27(2): 57-63. DOI: 10.12142/ZTETJ.202304011
- [8] QIAN K, XI Y Q, CAO J M, et al. Alibaba HPN: a data center network for large language model training [C]//Proceedings of the ACM SIGCOMM 2024 Conference. ACM, 2024: 691-706. DOI: 10.1145/3651890.3672265

- [9] JIANG Z H, LIN H B, ZHONG Y M, et al. MegaScale: scaling large language model training to more than 10, 000 GPUs [EB/OL]. [2025-03-01]. <https://arxiv.org/abs/2402.15627v1>
- [10] 百度智能云开发者中心. 大规模 AI 高性能网络的设计与实践 [EB/OL]. [2023-06-25]. <https://cloud.baidu.com/article/364290>
- [11] KIM J, DALLY W J, SCOTT S, et al. Technology-driven, highly-scalable dragonfly topology [C]//Proceedings of International Symposium on Computer Architecture. IEEE, 2008: 77-88. DOI: 10.1109/ISCA.2008.19
- [12] SHPINER A, HARAMATY Z, ELIAD S, et al. Dragonfly: low cost topology for scaling datacenters [C]//Proceedings of IEEE 3rd International Workshop on High-Performance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB). IEEE, 2017: 1-8. DOI: 10.1109/HiPINEB.2017.11
- [13] JOUPPI N, KURIAN G, LI S, et al. TPU v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings [C]//Proceedings of the 50th Annual International Symposium on Computer Architecture. ACM, 2023: 1-14. DOI: 10.1145/3579371.3589350
- [14] WANG T, SU Z Y, XIA Y, et al. NovaCube: a low latency torus-based network architecture for data centers [C]//Proceedings of IEEE Global Communications Conference. IEEE, 2014: 2252-2257. DOI: 10.1109/GLOCOM.2014.7037143

作者简介



段威, 中兴通讯股份有限公司资深研发总工, 高级工程师; 主要从事IP网络、智算中心网络关键技术研究; 申请专利30余项。



于浩, 中兴通讯股份有限公司网络系统专家; 主要从事IP网络相关的技术研究和研发工作。



李和松, 中兴通讯股份有限公司技术规划专家; 主要从事智算、光电子以及网络相关的技术研究和规划工作。



张征, 中兴通讯股份有限公司资深标准工程师; 主要从事智算、广域等数通网络相关的技术研究和标准工作。



王宁诚, 中兴通讯股份有限公司算法专家; 主要从事智算网络、智能化相关的技术和算法研究工作。