特色专题

智算中心高性能网络技术研究进展

任杰,刘畅,韩博文,文晨阳,徐博华,曹畅

摘要 随着ChatGPT引领的大模型与AI产业的爆发式发展,大规模分布式计算成为大模型训练常用模式,对应智算算力需求激增。旨在形成智算中心高性能网络技术体系,推动智算中心高性能网络技术持续发展。针对智算中心高性能网络内关键技术进行技术研究,首先,针对大规模智算业务承载场景,分析了智算中心提供高性能网络在传输协议层面、组网层面、管控运维层面的核心需求。随后依据所述需求,详细研究了智算中心高性能网络不同网络层的演进需求及智算中心高性能网络组网、面向智算中心网络的新型负载均衡协议与拥塞控制协议、新型网络管控及运维等领域的关键技术,对不同场景需求提供技术指导。其次,从网络协议发展与全光网络2个层面展开,分析了智算中心网络的未来导向与发展趋势。若要建立完善智算中心高性能网络技术体系,智算网络自身需提供足够的网络性能,如提供近似无丢包的网络环境、足够的互联能力并解决分布式存储场景下的存储性能瓶颈等;同时智算中心高性能网络的发展需要规范组网方案、高性能的新型负载均衡与拥塞控制协议、新型智慧化管控运维技术等方面关键技术的融合协同,提高运营效率;智算中心高性能网络需提供全局范围内设备与资源感知、分配、调度、运维的网络,并提供高性能无损传输能力。

关键词 智算网络;组网;拥塞控制;负载均衡;管控运维

随着人工智能技术的普及与大模型的爆发,智算中心正在加速组网布局建设中。智算中心通过使用异构算力为人工智能应用提供所需基础设施,实现从底层算力供给、通智超资源一体化调度到顶层应用使能的全栈能力。

当前中国智算中心产业发展规模与市场规模潜力巨大。据文献[1]估测,2027年中国智算算力规模或将达到1117.4 EFLOPS,5年复合增长率达33.9%。同时,随着大模型在边缘侧及端侧的算力需求激增,2023年中国智算市场规模达5097亿元,且于2028年预计达到3.4万亿元,实现6.7倍的爆发性增长[2]。

国家发布多项政策推动智算产业的发展。2021年3月,文献[3]中系统布局新型基础设施建设,并明确了加快国家枢纽节点、大数据中心集群、超级计算中心的建设要求。2023年10月,文献[4]中提出2025年算力规模超过300 EFLOPS、智能算力占比达到35%的要求。智算算力需求激增、训练规模不断增长、网络技术发展演进、市场规模不断扩大、政策扶持支持共同推动智算中心及其高性能网络发展。

智算中心的内部互联性能以及远程直接存储器访问(remote direct memory access, RDMA)传输技术下的传输能力是智算中心提供高性能网络的关键,亦成为制约业务性能重要因素。通过分析智算中心高性能网络需求,引出包括RDMA技术、基于融合以太网的远程直接内存访问(RDMA over Converged Ethernet, RoCE)技术、拥塞控制与负载均衡协议在内的多种及智算中心高性能网络技术,分析技术利弊及应用前景,为搭建新型智算中心高性能网络提供依据。

中国联合网络通信有限公司研究院,北京 100176

收稿日期:2024-08-21;修回日期:2024-12-31

作者简介:任杰,工程师,研究方向为新型数据中心网络协议,电子信箱:renj30@chinaunicom.cn

引用格式:任杰,刘畅,韩博文,等. 智算中心高性能网络技术研究进展[J]. 科技导报,2025,43(9):62-75;doi:10.3981/j.issn.1000-7857.2024.08.01038

1 智算中心高性能网络需求分析

1.1 对 RDMA 技术的需求

直接存储器访问(direct memory access, DMA)是一种计算机总线架构的能力,能够让数据从附加设备(如磁盘)直接发送到内存上,数据搬运过程无须中央处理器(central processing unit, CPU)的参与,从而大幅度降低 CPU 拷贝的开销。RDMA 是将 DMA 能力拉远到多台计算机之间,允许主机之间内存直接访问。

传统 TCP/IP 网络技术通过操作系统内核频繁的数据拷贝和中断操作来传输数据,而 RDMA 技术则是通过绕过内核并将网络堆栈卸载到网卡实现 CPU 开销接近零的高吞吐和超低延迟。RDMA 不仅改进了性能,还减少了每个服务器上网络堆栈处理使用的 CPU 核数量。

对于 RDMA 的承载方式,主流的方案有 Infiniband(无限带宽,IB)[5]、RoCE[6−7]、iWARP(Internet Wide Area RDMA Protocol,基于 TCP/IP 协议栈的 RDMA 技术)[8]3 种,如表 1 所示。

表 1 新型智算网络主流的承载方案对比

技术	网络协议	实现	网卡	网络设备	开放性	优缺点
Infiniband	IB	将IB卸载到HCA上	仅Mellanox的IB卡	仅Mellanox的IB交换机	IB专网	性能与稳定性均较好,但专网建设成本高、厂家单一、维护难度大且演进较慢
RoCE	UDP/IP	将RoCEv2卸载到NIC网卡上	支持RoCEv2的NIC网卡厂家较多:Mellanox、Broadcom、Qlogic、华为等	支持无损以太交换机	开放标准和网络架构	性能与IB相当,但要求以太网提供类似IB Link Layer的无损低延迟通路,对网络性能的要求较为严格
iWARP	TCP/IP	将TCP协议卸载到网卡上	仅Intel的iWARP卡	通用交换机	仅Intel	由于TCP在高性能通信方面的弊端,其稳定性和性能损耗较大,尚未被广泛采用

1.2 RDMA 技术对网络性能提高要求

早期组网中 RDMA 主要采用的是 IB 专网承载方案。由于 IB 设备对于丢包具有极强的保障能力,因而 IB 专网中近似于无丢包,使得 RDMA 的 IB 传输层采用较为简单的 Go−back−N 重传方案,在判断出丢包后将后续报文全部重传。

在扩展为 RoCEv2(RDMA over Converged Ethernet version 2,基于融合以太网的 RDMA 协议第 2 版)协议后,图形处理单元(graphics processing unit, GPU)组网从 IB 专网变为通用性更强的 IP 网,IP 设备无法在机理层面提供如 IB 设备的近似无丢包环境。一旦网络中出现丢包,RoCEv2 上层所采用过于简单的 Go−back−N 重传机制会在丢包环境中引发传输效率的严重受损。

文献[9]指出,在高性能计算场景中,0.1% 的丢包率会导致 RDMA 的吞吐量降至 70%,而 1% 的丢包率则会将 RDMA 的吞吐量归零。北京—宁夏 2900 km 实测表明,同机房短距与超过 1000 km 长距场景中,虽然等带宽直通场景下 RDMA 具有较好的传输效率,而当涉及带宽变动时,即使网络本身不存在丢包也不存在背景流,RDMA 的传输效率也极低。而且并非简单的传输效率下降,而是可能出现被重传空占带宽资源的情况,网络中 1 层端口速率接近满速,但 4 层有效传输速率仅为 20%~30%,甚至部分情况下可能接近 0。该现象产生原因即是由于 RDMA 在带宽变动引发的丢包后,频繁地触发 Go−back−N 带来大量的无效重传,而大量重传又持续将链路维持在极高的利用率水平无法缓解。

1.3 智算中心对高性能组网的需求

传统数据中心的组网技术偏向于对外提供服务的数据交互,因此,对外的南北向流量占主要部分,而东西向流量较少。传统数据中心采用的组网方式一般为“核心—汇聚—接入”3 层的 fat−tree 组网,每层所用交换机的上联和下联带宽存在收敛比,网络带宽

有限,网络时延偏大。

智算中心网络以东西向流量与跨智算中心流量为主,尤其在 AI 大模型的训练过程中,网络中仅存在东西向的 AI 服务器间或 AI 服务器和存储服务器间的数据同步流量,无须北向的流量出口,因此,特定场景下亦可被看作高性能但封闭的网络环境。此外,智算中心同样需要部分通算能力、带内和带外的管理能力,以及后续的跨数据中心(data center, DC)高速互联能力等。因此,为满足智算与通算的差异化流量需求,智算中心多需采用可兼容适配多应用场景的组网架构,实现根据不同的业务场景下按需配置不同的带宽大小和网络协议。

除业务需求外,文献[10]指出,智算中心组网还需考虑设备因素,尤其是 GPU 或 AI 计算加速卡等算力设备。在国产化需求日益重要的前提下,如何使用纯国产算力卡或混合算力卡实现高性能组网方案是智算中心组网领域需深入研究的重点。

1.4 智算中心网络流量特征引起的协议层面需求

智算中心在网络侧呈现的流量模型与传统数据中心差异较大。传统通用数据中心中,网络流量典型模型基于数量庞大且突发性强的小流,分布较为均衡。智算中心网络可根据业务特点细分为多类型专用网络,网络流量特征的差异性较大,如表 2 所示。

表 2 智算中心网络的差异性流量特征

网络类型	类型	网络流量特征
通用计算网络 存储前端网络	传统数据中心	由多个小业务流汇聚而成的平稳数据流构成
高性能存储后端网络	智算中心	以数量较多的高带宽大流为主,对时延高度敏感;由于缓存系统的存在,以写多读少的特征为主
AI/ML大模型训练网络		由一定数量具有同步效应的大数据流组成的周期性峰值流量,流间具有明显的同步效应,对时延和抖动要求较高

高性能计算、大模型等新兴应用的涌现使得单 DC 内与跨 DC 间场景下的数据流通量剧增。一方面,需研究负载均衡与灵活调度算法,在宏观上避免平均流量的挤压与冲突,通过生成细粒度的流量切分模型实现流量均衡并适配智算中心内多样性路径,同时进行哈希算法的优化以降低由于哈希冲突导致的 FCT(流完成时间)过高风险;另一方面,需研究端网协同的主动拥塞控制机制,通过端侧与网侧的协同配合,更加精确地感知和通告拥塞状况,在微观上避免瞬时流量的超载与拥塞,并适配智算业务的传输性能要求。通过宏观与微观相结合的综合化手段,解决现有 RDMA 网卡拥塞控制算法在大规模高速智算集群中收敛慢/吞吐低/时延高的瓶颈,实现大规模 RDMA 流量的无损高效承载。

1.5 智算中心的智慧化运维需求

智算中心聚合了海量异构算力、面向复杂业务场景、具备差异化的流量特征,在精细化硬件资源管理、端网一体化管理、设备状态的可视化与监控、灵活的业务部署与快速市场响应、高效的故障管理与业务恢复、多租户需求等诸多层面对管控运维能力提出了需求。

2 智算中心高性能网络关键技术

2.1 组网与拓扑

2.1.1 关键组网技术

1) 分区组网技术

为支撑不同场景的需求,智算中心网络提供分区组网技术,每个区域根据对应场景流量模型选用合适的组网方案。分区组网方案通常由训练区、存储区、推理区、带内带外管理区和汇聚区构成,使用差异化带宽链路进行互联。存储区和训练区含有基于 Spine-Leaf 2 层 CLOS 网络架构(CLOS architecture)的后端高性能无损网络,并与其他区域实现物理隔离,仅接受管理区的配置下发和状态上报。分区组网方案架构如图 1 所示。

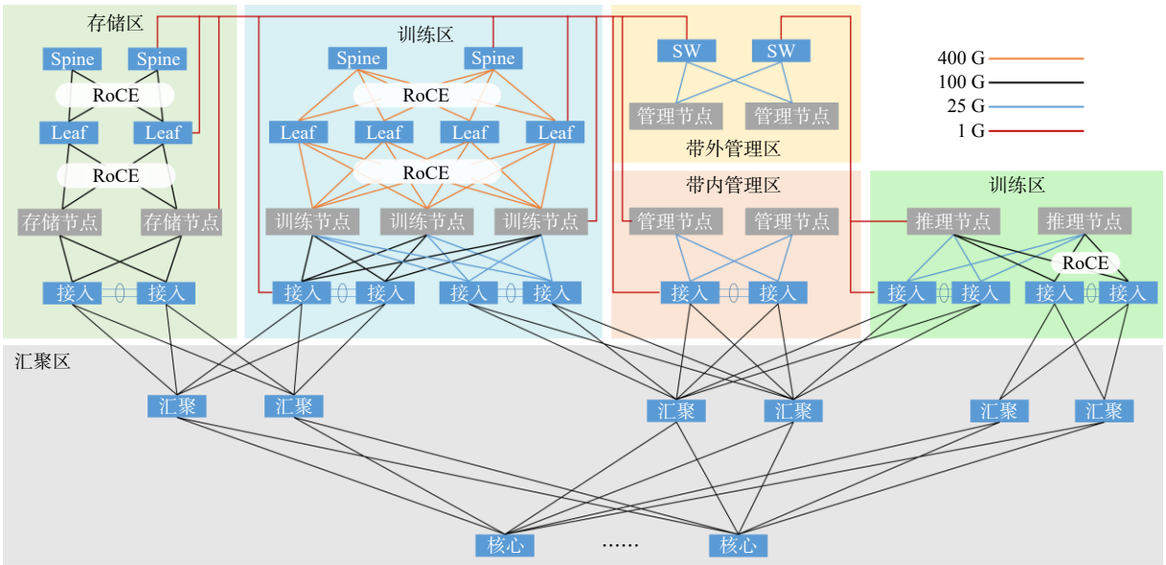

图 1　智算中心分区组网方案架构

分区组网的典型场景包括高性能计算、高性能存储等。

（1）高性能计算（high performace computing, HPC）场景。高性能计算场景要求计算节点间的链路满足大带宽、低延迟、高并发要求，同时需计算节点与存储节点互联。高性能计算场景多采用 RoCEv2 网络方案，2 层 CLOS 拓扑架构（Spine−Leaf）要求网卡和交换机支持 RDMA 能力，计算节点间互联设备采用 400 G/200 G 的端口形态，计算节点与存储节点间的互联设备采用 100 G/25 G 的端口形态。智算网络高性能计算场景架构如图 2 所示。

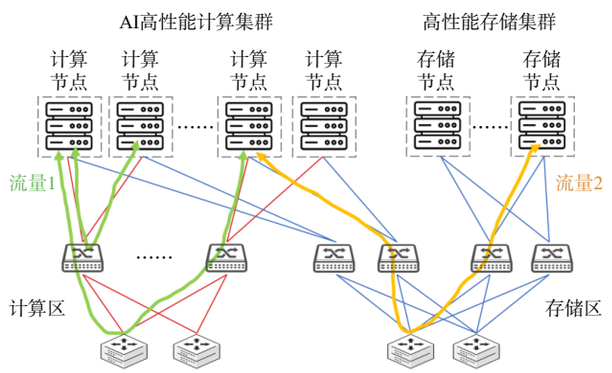

图 2　智算网络高性能计算场景架构

高性能计算场景内流量可分为，流量 1：HPC 计算节点间并行计算流量。在 AI 模型训练中，并行计算节点间相互传递计算结果，对高并发流量实现链路负载分担，对 Incast 流量需通过拥塞控制算法保证网络的无损和可靠性。可将所有 HPC 流量所经过的网络区域划分为计算区。流量 2：计算节点访问存储节点流量。在模型训练中，主要是模型备份和样本导入的交互场景，计算节点从存储中读取数据进行分发训练，并将计算结果放置存储节点，需网络提供无损传输的能力。该流量跨越计算区与存储区。

（2）高性能存储场景。高性能存储场景分为存储与计算节点间的互联及存储节点之间的组网互联 2 类。智算网络内的高性能存储场景同样要求网络提供无损传输能力，但对网络带宽、时延要求相较 HPC 高性能计算场景稍低。此场景仍采用 RoCEv2 方案，且以 100 G/25 G 端口形态为主。智算网络高性能存储场景架构如图 3 所示。

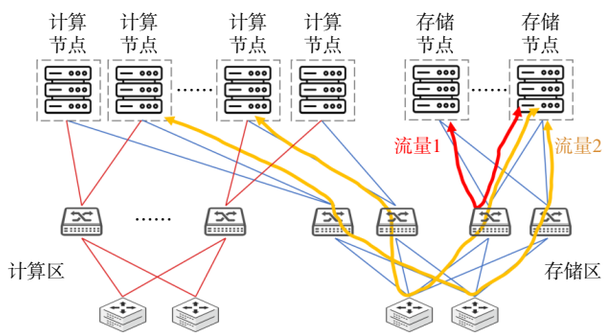

图 3　智算网络高性能存储场景架构

高性能存储场景的流量可分为，流量 1：存储后端网络互访流量。流量报文从源存储节点进入 TOR/Leaf 交换机，依次经过 EOR/Spine、对端 TOR/Leaf，到达

目的存储节点;整个过程在存储后端网络进行,与前端网络物理隔离。可将所有高性能存储流量所经过的网络区域划分为存储区。流量 2:存储节点与计算节点之间的流量。与高性能计算场景所述流量 2 类似,但由于存储后端互访流量需与其他网络流量隔离以形成独立的存储区,因此,还可将该类型流量通过汇聚区或称为业务区承载,可参见智算中心整体组网架构。

2)单轨组网与多轨组网

2023 年,在 MIT 和 META 联合发表的研究[11]中提出“rail-only”的定制化组网方式。在当前智算中心的组网拓扑中,也借鉴该思路形成多轨和单轨 2 种组网模式。2 种组网模式对 GPU 能力、服务器内部设计、网络设备均有不同要求。智算中心建设中多轨和单轨组网方案的选择基于 GPU/AI 加速卡支持度。

(1)多轨组网。多轨组网方式下的网络拓扑如图 4 所示,集群中服务器编号为 N 的网卡需连接 Group 中拥有相同编号 N 的 leaf 交换机。Group 下 leaf 交换机的数量与 1 台 AI 服务器的网卡数量一致,而 Group 中服务器的数量取决于交换机的转发能力。Leaf 层交换机的上下行网络带宽收敛比为 1∶1,并向上与 Spine 层交换机进行全互联。

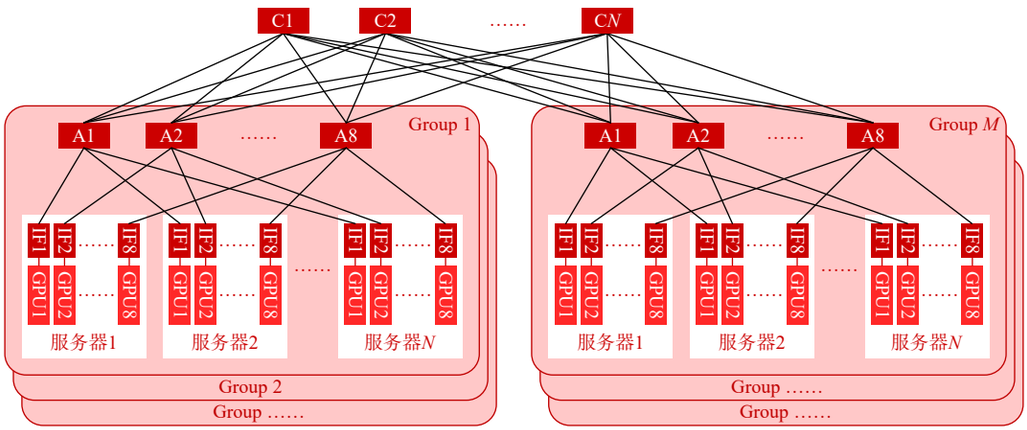

图 4 多轨组网方式下的网络拓扑

同一个 Group 中跨服务器只能进行同号网卡之间的通信,即同轨道通信。异号 GPU/AI 加速卡之间的通信需由上层任意 Spine 交换机中转实现,即跨轨道通信。跨轨道通信也可通过利用服务器机内互联链路将数据传输给同号 GPU/AI 加速卡后再进行同轨道通信实现,如图 5 所示。

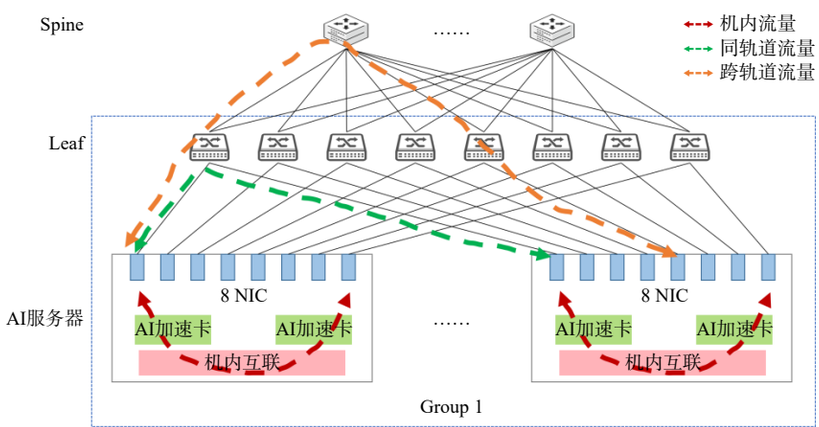

图 5 多轨组网模式下的流量流转模式

多轨组网模式下,流水线串行和数据并行大部分是同号 GPU/AI 加速卡之间的同轨道流量,可在同 1 台 Leaf 下完成大部分任务,仅有少部分需要跨 Spine 基于跨轨道流量实现。因此,多轨组网中的流量呈现

层次化分布,流量分配及负载效果较好。

此外,多轨组网方案可扩展至 3 层结构,如图 6 所示。3 层组网架构最多可支持万卡级别的组网规模,也是当前大型核心智算中心的主流组网方案。

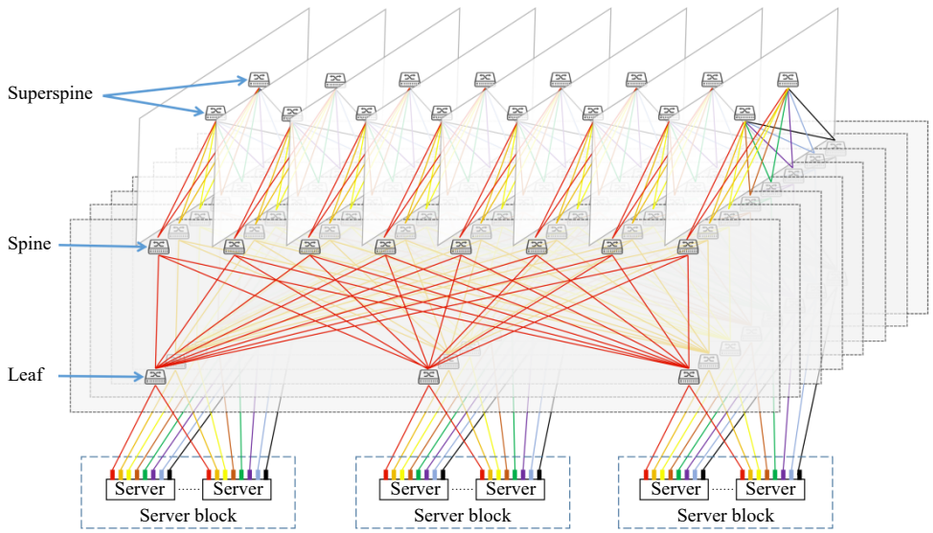

图 6 3 层结构的多轨组网方案

(2) 单轨组网。单轨组网架构的示意图如图 7 所示。单轨组网方式内 1 个 Group 内仅有 1 台 Leaf 交换机,且该 Leaf 交换机与 Group 内所有服务器的所有网卡互联,不分网卡编号。Group 内可容纳的服务器数量取决于 Leaf 交换机的转发能力。Leaf 交换机向上与 Spine 交换机全互联。

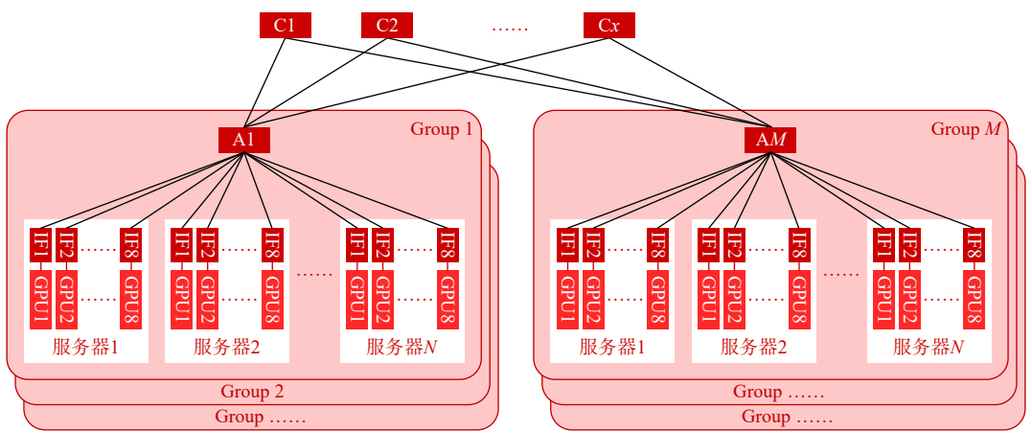

图 7 单轨组网架构的示意

单轨组网模式下的流量流转模式如图 8 所示。Group 内跨服务器通信的流量由 Leaf 交换机中转并通过 Spine 交换机转发。在大规模 AI 训练过程中,若机内互联网络能力不足,张量并行流量除了通过机内互联亦需要通过 Leaf 交换机拓展同 Group 资源实现。单轨组网的数据同步时间长于多轨组网。

2.1.2 智算中心组网拓扑

常用的智算中心组网拓扑包括 Fat-tree[12]、Overlay[13]、Spine-leaf [14]、BiGraph[15]、Dragonfly[9]、Torus[16]等,兼顾智算中心网络对可拓展性、可用性、稳定性、高效性与安全性的需求。

Fat-tree 拓扑最早于 2008 年提出[11],隶属 CLOS 网络架构,利用大量性能较差但成本较低的交换机,构建出大规模无拥塞且无带宽收敛的智算中心网络,其搭建简单,构建成本较低被业内普遍认可。然而,Fat-Free 架构的拓展能力较差,且无法很好地支持 One-to-All 及 All-to-All 模式,对于新型智算中心兼容性较差。

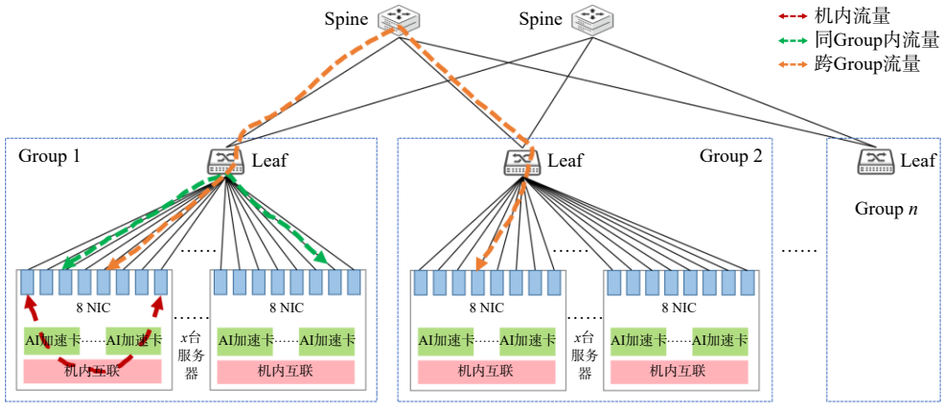

图 8 单轨组网模式下的流量流转模式

Overlay 拓扑将应用虚拟化技术的逻辑网络叠加至实体网络上,将业务部署至虚拟网络上承载,可以为智算中心带来“网业分离”的能力。其主要分为网络 Overlay、主机 Overlay 和混合 Overlay 3 种模式,可以为业务提供灵活的迁移与拓展能力,兼顾了可拓展性与可用性的优势,使业务摆脱了物理层网络限制。但由于虚拟化技术的叠加,性能成为 Overlay 拓扑在智算中心网络中的瓶颈。

Spine–leaf 拓扑是目前应用最广泛的智算中心网络拓扑。其隶属于 CLOS 网络架构,已被广泛证明可为智算中心提供高带宽、低延迟、无阻塞的端到端网络拓扑。其采用脊–叶 2 层架构,具有扁平化的结构,使其无论是 spine 交换机还是 leaf 交换机都具有优异的可拓展性,同时其还具有低收敛比、兼容边缘云业务、易于管理等优势。但智算中心网络的实际部署中,如何实现高效的拥塞控制与负载均衡策略,是需要进一步解决的问题。

BiGraph 将网络分为上下 2 层,每部分的交换机都可以接入服务器,类似于 Spine 和 Leaf 合二为一的组件,并且在上下 2 层交换机间采用 CLOS 架构互联,具有丰富的物理链路资源,这一结构也便于进行集合通信库的针对性优化。其具有网络可靠性强、网络效率更高、交换机端口更多、增强的冗余和容错能力、易于管理和扩展等特性,尤其适合中小规模的集群。

Dragonfly 基于路由器、组和系统的分层设计,每台路由器连接至多个终端,并拥有本地通道和全局通道,分别用于与组内其他路由器和组外路由器的连接,显著扩大了路由器的端口数量,提高了网络的连接性和灵活性。同时,其通过“直连拓扑”与“封装局部性”实现了显著的时延降低和更高的通信带宽与更低的延迟,采用多种路由策略与算法实现了高效负载均衡与拥塞控制。

Torus 行、列均闭环连接的环形网络拓扑拥有极强可扩展性及高带宽、低延迟的通信能力。其可以通过将二维闭环拓展至三维闭环的方式增加新网络节点,从而快速、大幅提升网络的容量与带宽,并降低通信延迟。但 Torus 存在复杂度高、可靠性不足的局限性,常用于超算中心及片上系统,在智算中心尤其是无损网络环境的设计中优势有限。

除上述网络拓扑外,张雅芝还总结了当前行业内其他数据中心拓扑方案,包含以交换机为核心的 Monsoon、Portland、ElasticTree、FBFLY、Jellyfish、F10、Aspen Tree 等,以及以服务器为核心的 DCell、FiConn、BCube、MDCube、BCN、SWCube、BCCC 等[17]。

2.2 主动拥塞控制协议

鉴于智算/超算所需的 RDMA 技术对网络性能提出了极高的要求,因此,需要主动拥塞控制技术在流量源端适当抑制发送速率,尽可能避免在网络中产生拥塞与挤占,从而让网络能够保持在高性能环境下运行。主流拥塞控制技术对比如图 9 所示。

现有的拥塞控制技术,主要包括 DCQCN[18]、TIMELY[19]、HPCC[20]等典型拥塞控制技术。主要的设计思路有:基于 2 层的 PFC 技术[17]、基于纯 4 层路线与基于 3/4 层端网协同路线等。

运营商、通信设备商等智算网络实践多是倾向于使用 RoCEv2,采用 DCQCN+ECN+PFC 的拥塞控制方案,其中 PFC 作为 2 层兜底保障手段,而 DCQCN+ECN 作为主要速率控制手段,并在此基础上针对

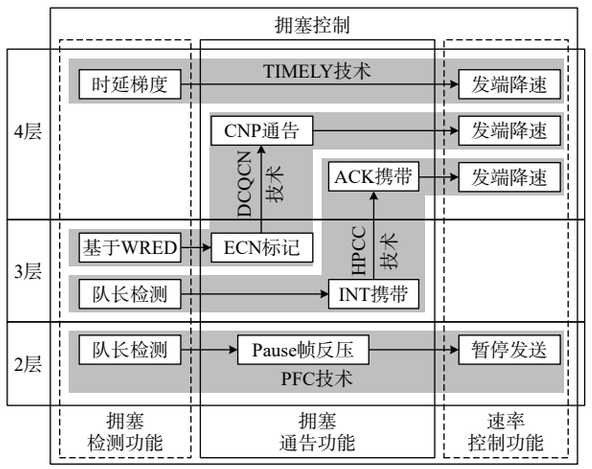

图 9 主流拥塞控制技术对比

DCQCN 做部分的差异性优化。

1）PFC 技术。PFC(priority based flow control)[21]是由 IEEE802.1Qbb 定义的一种基于优先级的流量控制协议,主要用于解决流量拥塞引发的丢包问题。

PFC 技术基于优先级的特性体现在其可针对 8 个虚拟通道分别进行流控(通常对应 8 个 QoS 优先级);流量控制的特性体现在其借助传统以太网的 Pause 帧(IEEE 802.3 Annex 31B)作为反压信号,可要求上游设备按指定时间暂停发送数据。合理地为 PFC 设定门限参数,能够在队列实际产生丢包前及时暂停上游数据发送而有效避免网络丢包。PFC 技术对于零丢包的保障效果在实践中已经得到了充分的验证,并在实际智算网络中有着广泛应用。

PFC 技术也存在较多局限性,主要体现在:流控的粒度较粗极易产生降速误伤、逐跳反压会在多个位置同时出现拥塞时触发互反压导致所有设备无法发送的死锁状态(PFC dead lock)。

在智算网络的实践中,通常是将 PFC 技术作为网络零丢包的兜底保障手段,而在上层附加各种其他拥塞控制技术,尽可能避免或降低 PFC 触发概率。

2）DCQCN 技术。DCQCN(data center quantized congestion notification)[18]称为数据中心量化拥塞通知,是基于端网协同的拥塞控制算法,在 SIGCOMM'15 中由微软等单位提出。其建立在量化拥塞通知(quantized congestion notification, QCN)[22]与数据中心 TCP 协议(data center TCP, DCTCP)[23]的基础之上,专门为 RoCEv2 所打造,其协议功能主要是实现在端侧网卡(NIC)中,但也需要中间网侧设备支持显式拥塞通告(explicit congestion notification, ECN)功能。

DCQCN 由 3 个部分组成:交换机的角色是 CP(congestion point),接收端的角色是 NP(notification point),发送端的角色是 RP(reaction point)。交换机是拥塞的实际发生位置,用于标记拥塞;接收端 NP 感知、反馈拥塞并把控 CNP 间隔;发送端是实际针对拥塞做出调速反应的位置,控制实际发送速率。

DCQCN 支持运行在无损的 3 层数据中心网络中,可有效降低 CPU 开销,且能够在无拥塞情况下超快速启动。DCQCN 设计初衷是在拥塞初期借助 ECN 提前降低传输速率,从而尽可能降低甚至避免 PFC 的触发概率,进而更加平和地应对拥塞问题并保持稳定性。

DCQCN 技术也存在较大局限性,主要体现为收敛较慢、在已发生排队后才能触发后续动作而引起时延的波动、难以高效支撑长距场景、涉及大量未给推荐值且相互关联的参数项导致实际部署中面临烦琐的调参以优化性能的问题。

3）TIMELY 技术。TIMELY[19]是一种端侧的拥塞控制算法,在 SIGCOMM'15 中由谷歌等公司提出。其基于往返时延(round-trip time, RTT)的变化趋势进行速率调节,功能实现在端侧网卡(NIC)中,几乎不需要中间网侧设备做修改。

TIMELY 算法中将时延梯度(即 RTT 的一阶导数)作为核心指标,可反映从发送端到接收端再返回发送端的往返时间 RTT 的变化趋势。通过时延梯度能够更精准地预估拥塞程度变化趋势,从而更快速地做出调速动作。TIMELY 算法架构包含 RTT 测量、速率计算、速率控制 3 个核心模块。相较于 DCQCN/DCTCP 等基于队列阈值的技术,TIMELY 基于时延梯度的思路能够有效预测拥塞变化趋势,能够在拥塞实际发生前做出降速动作。但是,文献[24]中指出,TIMELY 技术也存在难以高效支撑长距场景、收敛较慢的问题。

4）HPCC 技术。HPCC(high precision congestion control)[20]是基于端网协同的高精度拥塞控制算法,于 2019 年由阿里巴巴公司等提出。其借助带内网络遥测技术(inband network telemetry, INT)随流携带网络状态信息,辅助传输协议进行拥塞判别。HPCC 协议功能主要体现在端侧网卡(NIC)中,且要求中间网

侧设备支持 INT 功能。

HPCC 利用 INT 技术实现状态信息收集。通过在发送端得到网络状态数据后通过各段链路的容量与时延综合估算其可承受的数据量,并利用窗口机制对在途数据量(inflight)进行限制,进而形成控制发送速率的效果。由于 HPCC 算法可通过链路的物理带宽与时延预估在途数据量限制,因此,文献[20]宣称其可提供接近 0 的网络排队,以保证低时延。

但是,HPCC 技术存在需要定制化网络设备以支持 INT 技术、数据报文变长而影响传输效率并对网络中最大传输单元(maximum transmission unit, MTU)的设置提出要求、难以高效支撑长距场景、收敛时间较长等问题。

2.3 高效负载均衡协议

负载均衡协议通过合理分配网络内流量并利用智算中心中的多样性路径,避免智算网络中出现单点过载问题,提升智算中心内/间传输的系统性能与传输稳定性。

流量切分是负载均衡协议实施的重要手段。根据流量切分粒度的差异,可将负载均衡协议的流量切分粒度分为 flowlet 级、流级及子流级、包级。同时,还可将负载均衡协议细分为针对 TCP 与针对 RDMA 的 2 类负载均衡方案,以满足单 DC 内与跨 DC 场景下长距离无损传输场景的负载均衡需求。

表 3 总结了智算中心主流负载均衡方案及其特性。

表 3 智算中心主流负载均衡方案及其特性对比

方案	流量切分粒度	方案类型	是否需要修改网卡硬件	是否涉及重排序	是否需要TCP的流特性	是否多路径	是否需要SDN控制器
CONGA	flowlet	针对TCP	否	否	是	否	否
Presto	包粒度	针对TCP	是	是	否	是	否
LetFlow	Flowlet	针对TCP	否	否	是	否	否
Drill	包粒度	针对TCP	否	是	否	否	否
PLB	流/子流粒度	针对TCP	是	是	是	否	否
MP-RDMA	包粒度	针对RDMA	是	是	否	是	否
Maestro	子流粒度	针对RDMA	否	否	否	是	是
ConWeave	流/子流粒度	针对RDMA	否	是	否	否	否

1)ECMP 协议。ECMP(equal-cost multi-path routing)[25]是基于本地转发的逐跳的负载均衡协议,主要用于解决智算网络中的流量的负载均衡问题。

ECMP 协议基于本地转发的特性体现在其在每个交换机本地基于数据包的五元组(源 IP 地址、目的 IP 地址、源端口号、目的端口号、协议号)哈希来为每一个数据包进行路径选择,而不需要下行链路的任何反馈信息;ECMP 协议的逐流的特性体现在其根据五元组进行数据包转发,而同一条流的所有数据包的五元组均相同,因此,同一条流的数据包均会走同一条路径,从而保证了数据包的有序到达。已有实践表明,ECMP 协议在流量大小均匀且流的数量很多的情况下可以很好地均衡负载。

ECMP 协议也存在较多的局限性,主要体现在:不同流量可能会哈希到相同的路径上从而产生哈希冲突导致负载分布不均衡,且在大小流分布不均匀的场景下更加严重;ECMP 对路径的拥塞无感知能力,对流量的转发仅仅基于五元组哈希,无法提前感知下行路径的拥塞情况,从而无法提前避免拥塞的发生。在 RDMA 的传输中,ECMP 是标准化的负载均衡协议。在智算网络的实践中,通常会使用其他负载均衡技术来规避 ECMP 的局限性所带来的性能损耗。

2)CONGA 协议。CONGA[26]是分布式拥塞感知负载平衡机制。其基于本地交换机计算的结构路径上的实时拥塞及来自远端交换机的反馈实现负载均衡。

远端交换机反馈度量存储在每个目标 Leaf 交换机每条上行链路的 Congestion–To–Leaf 表中,并传递路径上所有链路的最大拥塞。源 Leaf 发送数据包时设置 LBTag 字段(发送数据包的上行链路的端口号),并将 CE 字段(沿数据包路径的拥塞程度)设置为零;

数据包通过 Fabric 路由到目标 Leaf,由折扣预估器 DRE(discounting rate estimator)获取链路的拥塞程度并更新 CE 字段;目的地 Leaf 将拥塞信息存储在 Congestion-From-Leaf 表中(基于每个源 leaf、每个 LBTag),当数据包以相反方向发送时,在 FB_LBTag 和 FB_metric 字段中插入 Congestion-From-Leaf 表中源 Leaf 的拥塞信息。

CONGA 利用 Overlay 层使用的 VXLAN 封装格式携带 LBTag、CE、FB_LBTag 和 FB_Metric,Overlay 层也为 CONGA 的 Leaf-to-leaf 拥塞反馈机制提供了理想管道。用于测量链路负载的 DRE 算法相较于广泛使用的 EWMA(exponential weighted moving average),只需 1 个寄存器即可实现(而 EWMA 需要 2 个寄存器),且对流量突发的反应更快。实际应用中,CONGA 的分布式架构解决了集中式方案对于链路中拥塞响应缓慢的问题,且能够有效地平衡负载并无缝处理不对称性,无须修改端侧的 TCP 堆栈。

然而,其针对拥塞的反馈并非完全及时,存在负载均衡效率较低的风险。需要在交换机中使用自定义的 ASIC 实现,增加了实际应用成本。同时由于 RDMA 传输中很难检测到 flowlet,CONGA 协议不适用 RDMA 场景。

3)MP-RDMA 协议。多路径 RDMA(multi-path RDMA,MP-RDMA)[27]算法采用 ACK-clocking 机制,是一种感知全局拥塞的包级负载均衡方案,通过设计机制、算法有效地解决了在 RDMA 多路径传输中网卡硬件内存有限的难题,使 MP-RDMA 适合在硬件中实现。MP-RDMA 对所有路径使用 1 个拥塞窗口,且使用基于 ECN 的拥塞控制算法,可根据拥塞程度比例调整窗口大小。

在实际操作中,可以发现由于 RDMA 网络中丢包率低,其通过 ACK 估计拥塞程度的方法是较为可靠的。而对于多路径传输导致频繁的乱序现象,MP-RDMA 通过乱序感知的路径选择算法动态分配速率较快且延迟相似的路径,有效控制乱序程度并提升网络利用率。MP-RDMA 将乱序数据包直接放入应用程序内存,减少 PCIe 带宽占用和延迟,并在略大于最大时延差的时间后进行内存同步,确保高概率的顺序到达,同时引入重传机制保证正确性,防止内存数据由于内存缓冲不足导致被无序更新的问题。MP-RDMA 还兼容 UDP 和 RoCEv2 包头中的大部分字段,引入了 MSN、iPSN、AACK 等字段拓展其包头。

MP-RDMA 是多路径、感知全局拥塞的高级别的拥塞解决方案,可以大幅提升网络利用率。但其功能实现依托于定制化 NIC 且与量产商用 RNIC 不兼容,因而限制了其大规模部署应用。

4)ConWeave 协议。ConWeave[28]是子流级别的 RDMA 网内负载均衡协议,利用可编程交换机实现对网内的拥塞感知,并通过重路由实现负载均衡。ConWeave 做出一次决策的周期不多于 1 个 RTT,并利用可编程交换机的队列暂停/恢复功能实现网内的重排序,解决网内无序数据包的影响。

ConWeave 的 2 个组件分别运行于源 ToR 交换机与目的地 ToR 交换机中,且 ToR 交换机间通过数据中心网络连接。ConWeave 使用源路技术为每条流进行选路。源 ToR 交换机上的组件可进行延迟监测以识别要避免的拥塞路径,当检测到拥塞时则立刻分配一个新路径,以确保可以“安全”地进行重路由而不会引发终端主机乱序到达。目的地 ToR 交换机上的组件则提供了数据包重排序功能,以解决由于重路由引起的乱序问题。

ConWeave 只会在现有路径拥塞、存在至少 1 个可行的非拥塞路径、重路由引起的乱序数据包已被目的地 ToR 接收的情况下会进行重路由,以确保重路由时存在非拥塞的路径可选且进行重路由时最多有 2 条路径上有数据包在传输。ConWeave 的负载均衡效果较好,平均 FCT 和 99 百分位的 FCT 可分别实现高达 42.3% 和 66.8% 的性能提升[28],但其对拥塞路径的筛选大量依托于重路由,而未考虑到其对速率的实际可能影响。同时,每条拥塞流都占用一个队列的机制会导致多流场景下可编程交换机有限的队列资源耗尽。

2.4 管控运维

1)超大规模集群管理。智算中心超大规模算力的实现需要基于万级数量级高性能 GPU 卡或 AI 计算加速卡的协同工作。因此,面对超大规模的硬件集群管理场景,需运维管控提供总体协调能力、故障恢复能力和系统稳定性保障能力。同时,还需提供对包括计算设备、网络设备、存储设备等硬件资源的精细化管理,以确保资源的高效利用和优化调度。

2）自动化端网部署及测试验收。智算中心网络的硬件环境极其复杂，涉及多种设备以及拥塞控制算法、RDMA 无损等复杂特性的配置和管理，涵盖 GPU、网卡和网络交换机等设备。设备的多样性和复杂性要求智算中心提供硬件选择、配置到优化的全方位管控运维能力。

在网络与端侧的基础配置完成后，为了保证交付质量，还需要管控系统提供自动化验收能力，包括基础环境的校验、通信库性能测试等。

3）精细化数据采集。RDMA 的流量一般呈现较强的突发性，传统简单网络管理协议 SNMP（simple network management protocol）的采样精度已经无法呈现网络的关键带宽业务指标。这意味着智算中心需提供细粒度的监控和调度的管控运维能力，以确保计算任务的高效执行和资源的合理分配。同时 RDMA 流量采集需要从端口级细化到队列级别。

除传统交换机设备相关信息的采集外，智算中心管控运维系统还需采集光模块、服务器和 GPU 等算力设备相关信息及拥塞关键指标等信息。

4）超可视化监控。超可视化监控可更清晰体现智算中心设备状态。包括集群网络可视、节点内部可视化、租户及作业路径可视化等。

5）自动化运维。智算中心网络应该具备自动化运维能力，包括自动巡检、故障快速定位定界、可用节点/慢节点查找、故障快速响应等能力。

6）智能化分析调优。智算中心网络通过智能化分析调优能力实现性能优化。可体现在如下方向。（1）拥塞控制算法参数优化：通过智能化的参数优化能力实现自动化、最优化调参，提高工作效率，同时提高训练效率。（2）融合路径优化：通过持续时间的长期监控，智能归纳网络的拥塞特点，并通过新型负载均衡协议降低网络出现拥塞的概率，从而提升智算网络总体效率。（3）高效的任务调度：实时计算当前提交的大模型任务所需最佳训练资源，并按当前的亲和性最佳资源进行最佳并行切分，减少大模型训练时间。（4）预测性维护：利用机器学习和人工智能技术对性能指标、系统日志及状态等大数据进行深入智能分析与学习优化，从而预测可能出现的问题并采取预防措施。

3 智算中心网络发展趋势

3.1 智算中心网络协议发展

拥塞控制协议方面，可分为纯端侧和端网协同的 2 种设计方向。纯端侧方向利用 RTT 等不涉及网侧设备的指标，判断拥塞情况并仅在端侧做出反应，优势在于可兼容几乎所有网络及组网方案，并简化部署复杂度，但由于无法感知网络设备处理时延及 RTT 的统计周期等情况，拥塞控制精确度有限，且存在收敛慢、反应慢的问题。端网协同方向均需要网侧交换机的功能支持，其中的 DCQCN 是智算中心实际建设中的最常用方案，而学术界更多推崇将基于 2 层或 3 层扩展字段的遥测技术用于拥塞的精确感知，例如，阿里的 HPCC、谷歌 CSIG、华为 CAQM 等。由于各家方案对交换机的功能要求不同，相互也并不兼容，且需要设备商定制化开发，落地场景一般仅限于自建智算中心。

负载均衡协议方面，包喷洒技术以及动态全局负载均衡方案是热门的发展方向。主流的动态全局负载均衡方案放眼网络全局，将“持续但不连续”的传输任务动态调度、卸载至多链路同时进行，与传统流级别负载均衡方案只能使用单一链路传输包含相同五元组流的模式相比，既解决了传统哈希将大象流只聚集在某几个特定链路中传输导致的链路拥塞风险及大量空余链路导致传输效率较低的问题，又解决了包级负载均衡需要耗费大量资源进行重排序的风险。而包喷洒技术则作为动态全局负载均衡方案的增补，可以很好地解决面对“持续且连续”的大象流时逐流的动态全局负载均衡方案仍旧失效的问题。目前主流包喷洒技术主要有新华三技术有限公司的 SprayLink 等，动态全局负载均衡方案有博通公司的 GLB、新华三的 DLB 等。

智算中心网络的协议未来发展目前存在 2 大问题。一是走传统网络的统一标准化道路，还是搭建不同的建设方案并采用不同的协议方案是尚未得到业内共识的问题，跨 DC 数据交互场景下的网络兼容性目前也没有良好解决方案。二是逐包负载均衡会导致很多拥塞控制方案失效。包喷洒是否是负载均衡的发展方向决定了拥塞控制是否需要根据该方向进行优化与创新。

3.2 基于OCS(optical circuit switching)等全光交换机方案的新型智算中心网络

智算中心为提供足够的算力,在资源的规模和灵活性上都有更高要求。传统的物理层连接方式虽然搭建简单且已经成熟,但由于其链路连接较为固定而缺乏可拓展性,使其成为智算中心资源拓展以及资源灵活调度的瓶颈。同时,在智算与超算业务场景下,TB级海量数据高效无损传输一直是业界难题。通过引入高效的、可重构的全光网络以及光交换机可以很好地解决上述瓶颈。

在前期中国联通公司的业内首次3000 km长距RDMA流量传输现网验证中,基于中国联通覆盖全国的骨干全光可重构光分插复用器(reconfigurable optical add-drop multiplexer, ROADM)网络和169骨干互联网,面向上海智算业务训练数据导入宁夏中卫智算训练集群的典型“东数西算”场景,利用光传送网(optical transport network, OTN)无损流控和端网协同拥塞控制技术,将端口带宽利用率从20%提升到90%以上,通过全光一跳直达,实现了高效的入云入算。同时,针对传统的数据搬运耗费时间长,效率低、安全性也比较差的痛点问题,中国联通基于自身全光网络优势,提出通过运营商网络打造跨数据中心数据传输的“数据高铁”的思路。

然而,目前基于光底座的智算中心网络还面临以下问题。(1)现有组网架构难以支撑高复杂度的800 GE光链路。同时800GE光网络的指数级复杂度调度会带来带宽资源调度不灵活的问题。(2)利用全光网络及OCS全光交换机的方案存在成本过高及并行传输链路由于硬件限制而受限(千流级别)的问题,如何面对万卡或更高集群亟须解决。

4 结论

根据智算中心高性能网络现状、行业现状与技术发展情况,从智算中心高性能网络的突出需求与满足需求所必需关键技术角度,研究了智算网络发展趋势及核心能力。

1)智算网络自身需提供足够的网络性能。如智算中心需为RDMA网络(RoCEv2)搭建近似无丢包的网络环境;为智算中心提供足够的互联能力,以解决分布式存储场景下的存储性能瓶颈等。

2)智算中心高性能网络的发展需要规范组网方案、高性能的新型负载均衡与拥塞控制协议、新型智慧化管控运维技术等方面关键技术的融合协同,通过提供专用化分区网络应对不同的需求场景,解决智算中心网络性能瓶颈,提高运营效率。

3)智算中心高性能网络需端网协同配合。需提供全局范围内设备与资源感知、分配、调度、运维的网络,并提供高性能无损传输能力,在此基础上汇聚和共享算力、数据、应用资源,实现智算业务的高效可保障供给。

智算中心高性能网络是智算中心智算业务能力的基座。智算网络的超高性能的构建,为推动智算或超算业务的蓬勃发展、打造极致算力服务体验提供了体系支撑。

参考文献(References)

[1] International Data Corporation. 2023—2024中国人工智能计算力发展评估报告[R]. 北京: IDC, 2023.

[2] 王祺, 李冬露. 2023年中国人工智能产业研究报告[R]. 上海: 艾瑞咨询研究院, 2024.

[3] 中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要[EB/OL]. (2021−03−12) [2024−08−06]. https://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[4] 工业和信息化部. 算力基础设施高质量发展行动计划[EB/OL]. (2023−10−08) [2024−08−06]. https://www.gov.cn/zhengce/zhengceku/202310/P020231009520949915888.pdf.

[5] Infiniband Trade Association. Infiniband architecture volume 1, general specifications, release 1.4[EB/OL]. [2024−08−06]. http://47.92.214.21:8888/rdma/IB%20Specification%20Vol%201-Release-1.4-2020-04-07_ib_spec_vol1.pdf.

[6] Infiniband Trade Association. Infiniband architecture specification release 1.2. 1 annex A16: RoCE[EB/OL]. [2024−08−10]. https://www.afs.enea.it/asantoro/V1r1_2_1.Release_12062007.pdf.

[7] Infiniband Trade Association. Infiniband architecture specification release 1.2. 1 annex A17: RoCEv2[EB/OL]. [2024−08−15].https://websearch.excite.co.jp/?q=InfiniBand+Architecture+Specification+Release+1.2.1+Annex+A17%3A+RoCEv2&page=1.

[8] Internet Engineering Task Force. The architecture of direct data placement (DDP) and Remote direct memory access

(RDMA) on Internet protocols[EB/OL]. [2024−08−15]. https://datatracker.ietf.org/doc/html/rfc4296.

[9] Kim J, Dally W J, Scott S, et al. Technology−driven, highly−scalable dragonfly topology[J]. ACM SIGARCH Computer Architecture News, 2008, 36(3): 77−88.

[10] Agam S. Nvidia shipped 3.76 million data–center GPUs in 2023, according to study[EB/OL]. (2024−06−10) [2024−08−06]. https://www.hpcwire.com/2024/06/10/nvidia-shipped-3-76-million-data-center-gpus-in-2023-according-to-study/.

[11] Wang W Y, Ghobadi M, Shakeri K, et al. Rail−only: A low−cost high−performance network for training LLMs with trillion parameters[C]//Proceedings of IEEE Symposium on High-Performance Interconnects (HOTI). Albuquerque: IEEE, 2024.

[12] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63−74.

[13] Cisco. Data center overlay technologies[R]. USA: Cisco, 2013.

[14] Cisco. Cisco ACI multi-tier architecture white paper[R]. USA: Cisco, 2024.

[15] Dong J B, Cao Z, Zhang T, et al. EFLOPS: Algorithm and system co−design for a high performance distributed training platform[C]//Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA). San Diego: IEEE, 2020: 610−622.

[16] Natalie E J, Tushar K, Li S, et al. On−chip networks[M]. Williston, USA: Morgan & Claypool, 2017.

[17] 张雅芝. 新型数据中心网络拓扑结构及性质的研究[D]. 济南: 齐鲁工业大学, 2024.

[18] Zhu Y B, Eran H, Firestone D, et al. Congestion control for large−scale RDMA deployments[C]//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. New York: ACM, 2015: 523−536.

[19] Mittal R, Lam V T, Dukkipati N, et al. TIMELY[C]//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. New York: ACM, 2015: 537−550.

[20] Li Y L, Miao R, Liu H H, et al. HPCC[C]//Proceedings of the ACM Special Interest Group on Data Communication. New York: ACM, 2019: 44−58.

[21] IEEE. 802.1Qbb. Priority-based flow control[EB/OL]. [2024−08−15]. https://1.ieee802.org/dcb/802-1qbb/.

[22] Alizadeh M, Atikoglu B, Kabbani A, et al. Data center transport mechanisms: Congestion control theory and IEEE standardization[C]//Proceedings of 46th Annual Allerton Conference on Communication, Control, and Computing. Monticello: IEEE, 2008: 1270−1277.

[23] Alizadeh M, Greenberg A, Maltz D A, et al. Data center TCP (DCTCP)[C]//Proceedings of the ACM SIGCOMM 2010 conference. New York: ACM, 2010.

[24] Zhu Y B, Ghobadi M, Misra V, et al. ECN or delay[C]//Proceedings of the 12th International on Conference on Emerging Networking Experiments and Technologies. New York: ACM, 2016: 313−327.

[25] Rhamdani F, Suwastika N A, Nugroho M A. Equal-cost multipath routing in data center network based on software defined network[C]//Proceedings of 6th International Conference on Information and Communication Technology (ICoICT). Bandung: IEEE, 2018: 222−226.

[26] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA[C]//Proceedings of the 2014 ACM conference on SIGCOMM. New York: ACM, 2014: 503−514.

[27] Lu Y W, Chen G, Li B J, et al. Multi-path transport for RDMA in datacenters[C]//Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation. New York: ACM, 2018: 357−371.

[28] Song C H, Khooi X Z, Joshi R, et al. Network load balancing with in−network reordering support for RDMA[C]//Proceedings of the ACM SIGCOMM 2023 Conference. New York: ACM, 2023: 816−831.

Research progress on technologies of high-performance network in artificial intelligence data center

REN Jie, LIU Chang, HAN Bowen, WEN Chenyang, XU Bohua, CAO Chang

Research Institute of China United Network Communications Co., Ltd., Beijing 100176, China

Abstract Amidst the rapid expansion of large-scale models and the AI sector, epitomized by advancements like ChatGPT, there is a burgeoning demand for intelligent computing power to facilitate extensive distributed computing applications. Traditional data centers are facing challenges to accommodate the performance requisites of these scenarios. Currently, the Chinese government has already issued numerous policies to expedite the development of artificial intelligence data centers, purveying clear policy guidance and accelerated planning and construction blueprints. A high-performance network is pivotal within AIDC, serving as the backbone for computational tasks and enabling inter-data center connectivity and efficient data transmission.This paper aims to establish a robust technical framework to propel the continuous development of high-performance network by primarily investigating key technologies for high-performance networks in artificial intelligence data center (AIDC). Core requirements in transport protocols, networking, and operation administration and maintenance (OAM) for large-scale AI tasks are studied. Based on these demands, this paper further investigate the evolving demands on different layers of Artificial Intelligence Network and delves into core technologies, e.g. network architecture, congestion control policy, load balance policy, operation administration and maintenance. Subsequently, from the two major perspectives of network protocol development and all-optical networks, this paper analyzes the future developmental trends of AIDC networks. To establish a robust high-performance network framework within AIDC, this paper concludes that sufficient network performance, such as a near-lossless network environment, adequate interconnectivity, and solutions to storage performance bottlenecks in distributed storage scenarios, must be purveyed effectively. Furthermore, the development of high-performance networks in AIDC necessitates the integration and synergy of key technologies, including standardized networking schemes, innovative load balancing and congestion control protocols, and advanced OAM mechanisms, to enhance operational efficiency. High-performance AIDC networks must also offer comprehensive and universal device and resource awareness, allocation, scheduling, and OAM across the entire network, providing high-performance lossless transmission capabilities.

Keywords artificial intelligence network; networking; congestion control; load balance; operation administration and maintenance

（责任编辑　王微）