智算中心高性能网络流量调度技术研究及实践

Research and Practice of High-performance Network Traffic Scheduling Technology in Intelligent Computing Center

韩博文[1]，徐博华[1]，曹　畅[1]，刘千仞[2]（1. 中国联通研究院，北京 100048；2. 中国联合网络通信集团有限公司，北京 100033）
Han Bowen[1]，Xu Bohua[1]，Cao Chang[1]，Liu Qianren[2]（1. China Unicom Research Institute，Beijing 100048，China；2. China United Network Communications Group Co.，Ltd.，Beijing 100033，China）

摘　要：
AI大模型训练、高性能存储等业务应用场景提出了海量规模的计算需求，与传统数据中心业务相比，在流量模型和网络需求方面有着显著区别，驱使传统的数据中心网络向智算中心和无损网络转型。从智算中心和无损网络的发展背景入手，分析了当前智算中心网络存在的问题，探索了智算中心网络流量调度的关键技术，并进行了流量调度平台的研发实践，为智算中心网络发展和应用提供思路。

关键词：
智算中心；无损网络；大模型训练
doi：10.12045/j.issn.1007-3043.2024.04.003
文章编号：1007-3043（2024）04-0012-08
中图分类号：TP393
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Scenarios such as AI large-scale model training and high-performance storage have proposed massive computing demands. Compared with traditional data center services，there are significant differences in traffic models and network requirements，driving traditional data center networks to transform towards intelligent computing centers and lossless network. From the development of intelligent computing centers and lossless network，it analyzes the problems existing in the current intelligent computing center，explores the key technologies of network traffic scheduling in the intelligent computing centers，and conducts research and practice on the traffic scheduling platform，which provides ideas for the development and application of intelligent computing centers.

Keywords：
Intelligent computing centers；Lossless network；AI large-scale model training

引用格式：韩博文，徐博华，曹畅，等. 智算中心高性能网络流量调度技术研究及实践[J]. 邮电设计技术，2024（4）：12-19.

1 概述

1.1 智算中心发展现状

智算中心与传统数据中心在建设目的、应用领域和主要特征方面存在显著差异。在建设目的方面，传统数据中心主要致力于提供IT资源支持，包括计算、存储和网络等基础设施服务。其设计核心在于支持企业日常运营、业务处理和信息系统托管，确保数据安全存储及业务连续性。相较之下，智算中心不仅提供基础数据存储和处理能力，更专注于智能计算领域，特别是针对人工智能算法的训练、推理和大数据分析等复杂计算任务。智算中心的目标是推动各行各业的数字化和智能化转型，促进产业AI化和AI产业化。在应用领域方面，传统数据中心广泛应用于各类企业和组织的信息系统，如企业资源规划（ERP）、客户关系管理（CRM）、数据库服务和内部办公系统等，覆盖范围广泛，但主要集中在传统行业。而智算中心主要服务于人工智能、机器学习、深度学习和自然语言处理等领域，满足新兴智能化应用的需求。在主要特征方面，传统数据中心主要提供基础设施即服务（IaaS）、软件即服务（SaaS）和平台即服务（PaaS），包含大量标准化的服务器、存储设备和网络设施，但通常

收稿日期：2024-02-26

不针对高性能或异构计算进行优化。其运维和管理相对独立且为静态的，资源扩展性受限，灵活性和自动化程度因技术和管理水平而异。相反，智算中心主要提供任务式服务（TaaS），强调计算性能和能效比的高度优化，以GPU为核心的智能算力，具备大规模存储和高速数据处理能力。智算中心在结构上更倾向于模块化和弹性扩展，同时需具备自动化和智能化的运维能力。

1.2 典型业务场景

1.2.1 AI模型训练

人工智能是数字经济高质量发展的引擎，也是新一轮科技革命和产业变革的重要驱动力量。大模型技术因其良好的通用性与泛化性，其溢出效应正在加速推进新一轮的科技革命和社会产业的变革。目前，越来越多的科技巨头竞相推出千亿参数的大模型，OpenAI的GPT-3模型参数已经达到了1 750亿。截至2023年，国内共发布了超过200个大模型，其中自然语言处理成为大模型研发的核心方向。

在AI模型训练过程中，需要通过数据集对模型进行训练，并根据损失函数的反向梯度计算方法对模型参数进行调整，训练过程会进行多轮的数据迭代。大规模的参数对算力和内存都提出了更高的要求，通常采用分布式训练技术对数据和模型进行切分，采用多机多卡方式，由大量GPU服务器组成算力集群协同计算，减少训练时长。

另一方面，大集群不等于大算力，分布式训练系统的整体算力并不是简单的随着节点的增加而线性增长，而是存在加速比，且加速比小于1。因为GPU集群规模的扩大还会引发额外的通信开销，并行应用中的通信开销可能会抵消更多CPU或GPU的收益，计算节点之间的同步时间，直接影响GPU集群的效率，分布式训练系统已经开始从计算约束转化为网络通信约束（见图1）。

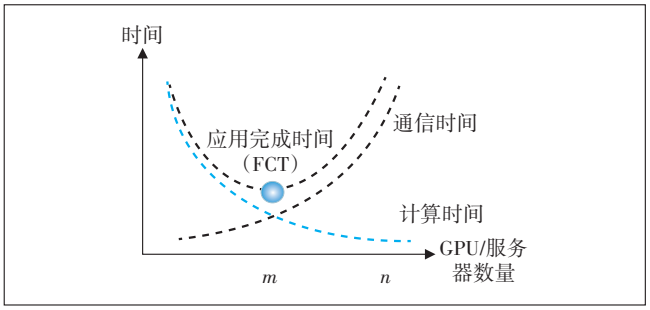

图1　集群规模与FCT关系示意

同时大模型训练也要求网络能够提供更高的稳定性与可靠性。因此，AI模型训练对无损网络的需求非常强烈。

1.2.2 高性能存储

随着大规模计算需求的兴起，高性能存储系统的需求显著增加，这主要体现在以下几个方面。

a）性能需求。大量计算任务对存储性能的需求主要集中在数据加载的初期阶段和最终数据保存阶段。如果存储性能跟不上，将导致高成本的计算资源空闲等待。因此，降低这些计算资源的等待时间是存储系统的重要目标，这要求存储系统具备高吞吐能力。影响分布式存储性能的主要因素就是存储介质的时延和节点间的网络通信时延。SSD的性能增强和NVMe存储接口协议的出现，极大提升了存储系统内部的存储吞吐性能。在存储介质的时延已大幅降低的情况下，网络通信时延占比已从原来的小于5%变化到65%左右，影响分布式存储系统吞吐性能的提升（见图2）。

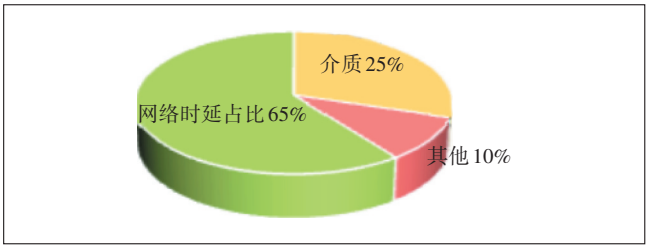

图2　分布式存储SSD场景下网络时延占比

b）文件大小。AI HPC场景对存储系统提出了特殊的挑战，尤其是处理海量小文件的能力。海量小文件对存储系统的扩展性和元数据性能提出了较高的要求。

c）接口需求。POSIX文件接口目前仍是最主流的存储接口。分层文件系统（HCFS）作为POSIX的一个子集，能够简化HCFS SDK的开发过程，满足POSIX标准的要求。特别是在HPC领域，除了需要完全兼容POSIX外，还需要适配MPI-I/O框架，以支持更为复杂的数据处理需求。

d）数据持久化与临时存储。对于存储系统来说，保证关键数据的持久化是基本要求，应确保数据不会丢失。

综上，高性能存储系统需要在性能、文件处理能力、接口兼容性以及数据持久化和临时存储策略等方面进行优化，以满足多样化和复杂化的场景需求。

2 智算中心网络需求

2.1 高性能网络需求分析

智算中心对高性能网络的需求可以从组网规模、带宽、丢包时延抖动和稳定性几方面进行分析。

a）从组网规模来看，传统数据中心网络节点数量通常小于1 000。随着数据并行和模型并行技术的不断完善和提升，智算中心高性能网络的分布式训练可使用千卡或万卡规模的GPU来缩短整体训练时长。

b）从带宽需求来看，传统数据中心接入带宽通常为10G/25G，汇聚核心带宽为40G/100G。而高性能网络中多为1∶1无收敛组网，接入侧智算服务器可以满配8张GPU卡，一般会给每个GPU关联一个网络端口，单端口需求正在从100G向200G、400G演进。

c）从丢包、时延、抖动需求来看，在分布式场景下，单次的计算时间包含了单卡的计算时间叠加卡间通信时间。因此，降低卡间通信时间，是分布式训练中提升加速比的关键，需要重点考虑和设计。传统数据中心网络时延通常为亚毫秒级，拥塞情况下为秒级。而高性能网络中，RDMA要求网络达到零丢包和微秒级时延。

d）从稳定性角度来看，传统数据中心检测时间通常为50 ms，收敛时间达到秒级或分钟级，而高性能网络要求亚毫秒级的网络稳定性。

2.2 智算中心管控需求

在典型的AI大规模网络中，网络部署、测试验收、运维以及变更全流程的复杂程度呈现指数级增长，传统网络依靠人工完成，耗时耗力且容易出错，智算中心网络需要实现多维度自动化能力的升级。

智算中心网络管控系统具备自动化端到端部署、自动化测试和验收、自动化运维的能力。

a）自动化端到端部署。高性能网络中涉及拥塞控制算法、RDMA无损等复杂特性的配置，其多样化的配置也往往被网络运营人员诟病。同时配置工作涵盖网卡和网络交换机，端到端部署自动化能力不仅是提升部署效率的关键，也是系统扩展性的重要前提。

b）自动化测试和验收。部署完成后，需要结合智算场景针对配置、可靠性和业务性能等方面自动化开展一系列测试和验收活动。

c）自动化运维。运维自动化是确保网络性能和可靠性的关键，对于一些突发的网络故障或者性能事件，需要通过端到端可视化、自动化运维实现故障的快速定位和一键修复的能力。自动化运维包括精细化采集、超可视化监控和自动化变更。

（a）精细化采集。从采集周期来看，RDMA的流量一般呈现较强的突发性，SNMP 30 s的采样精度已无法呈现网络的关键带宽业务指标。从采集粒度来看，RDMA流量需要从端口级细化到队列级别。从采集内容来看，除了传统网络需要的性能指标，智算中心的管控系统需要采集PFC和ECN等拥塞关键指标。

（b）超可视化监控。智算中心网络的管控系统需要具备超可视化监控。一方面需要具备集群网络的可视化，包括交换机、服务器和GPU卡之间的组网关系、秒级流量的监控、关键指标统计（丢包、时延、吞吐、PFC、ECN等）监测等，另一方面需要具备节点内部可视化，包括GPU利用率、PCIE和NVLink的带宽监测等。

（c）自动化变更。变更自动化是网络能力自演进的基本保障，在AI高性能网络中，业务需求的变化、新技术的引入、网络故障的修复、网络配置的优化等都会引发网络配置的频繁变更，变更自动化能力是确保过程安全的基本手段，也是网络能力自优化、自演进的基本要求。

2.3 现有拥塞控制方案的问题

因为RoCEv2缺乏完善的丢包保护机制，对于网络丢包异常敏感，因此拥塞控制算法的优化对智算中心的网络性能至关重要。

2.3.1 DCQCN

由于Mellanox网卡的限制，目前最广泛应用的拥塞控制协议是DCQCN（Data Center Quantized Congestion Notification）。它基于QCN（Quantized Congestion Notification）和DCTCP（Data Center TCP），结合了ECN（Explicit Congestion Notification）和PFC（Priority Flow Control），支持端到端无损以太网。ECN有助于克服仅靠PFC实现无损以太网的局限性。DCQCN的原理为：通过ECN在拥塞开始时降低传输速率来控制流量，以尽量减少触发PFC的次数，避免流量完全停止。它主要包括发送端（Reaction Point，RP）、交换机（Congestion Point，CP）和接收端（Notification Point，NP）。

CP算法中，如果出口队列长度超过设定的阈值，到达的数据包将被ECN标记。这一过程通过所有现有交换机所支持的RED（Random Early Detection）功能实现。拥塞标记的概率是队列长度的函数，具体关系如下：采用2个阈值定义了标记的概率，即当队列长度

低于下限阈值时，不会对ECN位进行标记。队列长度超过上限阈值时，所有经过该队列的数据包都将被ECN标记。而当队列长度位于这2个阈值之间时，数据包被ECN标记的概率将随队列长度的增加而线性增加。

NP算法中，到达NP的ECN标记数据包表明网络出现拥塞，NP将这一拥塞信息反馈给发送方。RoCEv2标准中定义了专门用于此目的的显式拥塞通知包（CNP）。NP算法规定了CNP的生成方式及其生成的时机。

对每个数据流，该算法按照下述状态机的指示操作：若一个被ECN标记的数据包到达接收方，且在最近N μs内未对该数据流发送过CNP，则立即发送一个CNP。此后，如果在相同的时间窗口内，该流的任何数据包再次被ECN标记，则网卡每隔N μs最多为该流生成一个CNP数据包。

RP算法，即DCQCN（Data Center Quantized Congestion Notification）的速率控制算法，通过特定的图表进行描述（见图3）。简单来说，发送方依据内部定时器和发送字节计数器持续增加发送速率。一旦接收到CNP包，便会降低发送速率。此外，该算法还维护了一个名为α的参数，该参数反映了网络中的拥塞程度，并用于计算降速。α是一个动态变化的平均值，计算基于CNP到达的时间间隔的比例（如果在同一时间间隔内收到多个CNP，其效果等同于仅收到一个CNP）。每个时间间隔结束时，α参数根据下式更新：new_α=g×old_α+(1−g)×CNP_arrived，其中g是一个介于0和1之间的常量，CNP_arrived是一个比特位字段，用以指示在上一个时间间隔内是否有CNP到达。

当在上一个时间间隙内到达CNP（如果同一个时间间隙内到达多个CNP，仅第1个CNP产生指示）时，队列对（QP）的速率会根据下式进行减少：new_rate=old_rate×(1−2α)，同时，会重置那些用于增加速率的参数。其加速逻辑与QCN（Quantized Congestion Notification）非常相似，分为快速恢复、积极增加（保持探测）、超积极增加（保持探测）3个连续阶段。每个阶段的转换是通过该阶段内累积的加速事件数量来定义的。当某一阶段内的加速事件数量超过预设的阈值时，将转入下一阶段。降速事件会重置所有与加速相关的计数器，并返回到快速恢复阶段。此外，一旦加速后，在降速前，当前速率会被保存在名为target_rate的参数中。从上次加速后，经过预定义的时间间隙或发送字节数之后，如果没有发生降速事件，则会触发加速事件。在快速恢复阶段，面对每个加速事件，速度会增加到target_rate与当前速率之间距离的一半〔即对数接近，current_rate=(current_rate+target_rate)/2〕。这允许在快速恢复阶段快速回到拥塞发生前的速率，并在接近拥塞速率时更加谨慎地增加速度。在接下来的2个阶段中，一旦发生加速事件，速率将按照固定值增加，以便在带宽释放时获得更高吞吐量（见图4）。

由于DCQCN具有超过16个可调节的参数，为了更加适应不同的网络拓扑和流量环境，其参数的调整显得格外重要，不同参数下的网络吞吐会有50%以上的差异。

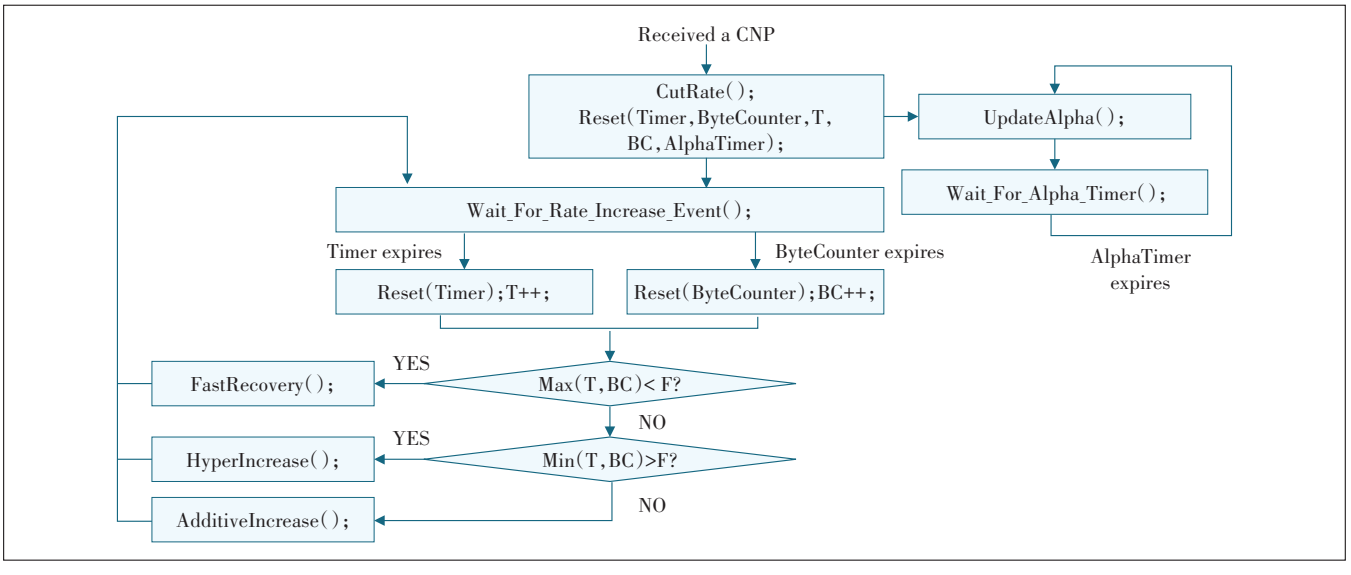

图3　RP算法流程示意

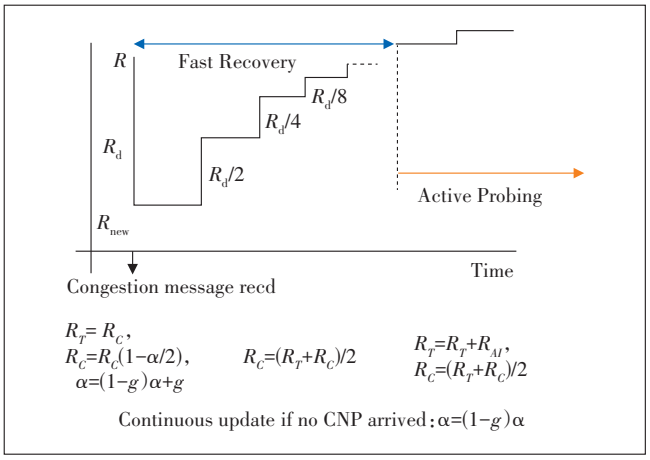

图4　DCQCN降速升速过程

同时DCQCN的参数众多，无论是在交换机上还是接收端/发送端上都需要根据网络流量状况进行调整才能协同发挥出网络最大的带宽吞吐。因此结合管控平台可以更方便实现对端网两侧参数的联合调优来提高网络性能。

2.3.2　AI-ECN

现有的无损网络方案，如某厂家提出的AI-ECN（Artificial Intelligence Explicit Congestion Notification），能根据现网流量模型，智能地调整无损队列的ECN门限，可以保障零丢包下的低时延和高吞吐，以使无损业务达到最优性能。

AI-ECN使用嵌入式AI进行智能计算，嵌入式AI是一个内置在设备中的AI功能通用框架系统，可以为AI-ECN提供模型管理、数据获取和预处理功能，支持向AI-ECN发送推理结果（见图5）。

网络设备内的转发组件会对当前流量的特征进行采集，比如队列缓存占用率、带宽吞吐、当前的ECN门限配置等，然后通过Telemetry技术将网络流量实时状态信息推送给AI-ECN组件。

AI-ECN功能启用后，将自动订阅嵌入式AI系统的服务。依据嵌入式AI系统，AI-ECN组件收到推送的流量状态信息后，将智能地对当前的流量模型进行判断，识别当前的网络流量场景是否是已知场景。

如果该流量模型是嵌入式AI系统内已训练的模型，则判断当前网络流量场景为已知场景，AI-ECN组件将根据嵌入式AI系统推理的最优结果，计算出与当前网络状态匹配的ECN门限配置，这种模式称为模型推理模式，由于其采用NN（NeuralNetwork）算法，因此也称为NN模式。

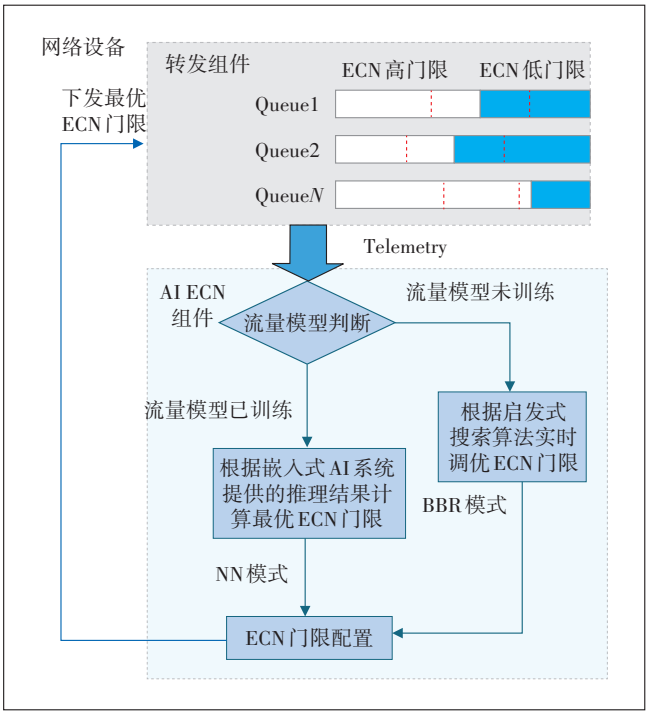

图5　无损队列的AI-ECN功能实现原理

如果该流量模型是嵌入式AI系统内未训练的模型，则判断当前网络流量为未知场景，AI-ECN组件将结合启发式搜索算法，基于现网状态，在保障高带宽、低时延的前提下，对当前的ECN门限不断进行实时修正，最终计算出最优的ECN门限配置，这种模式称为启发式推理模式，由于其采用BBR（Bottleneck Bandwidth and RTT）算法，因此也称为BBR模式。

最后，AI-ECN组件将最优ECN门限下发到设备中，调整无损队列的ECN门限。

对于获得的新的流量状态，设备将重复进行上述操作，从而保障无损业务的最佳性能，尽管它具有一些潜在的优势，但也存在如下局限性。

a）训练数据依赖。AI-ECN的性能高度依赖于训练数据的质量和数量，要建立准确的AI模型，需要大量的实际网络数据。如果训练数据不足或不代表实际网络环境，AI-ECN的性能可能会受到限制。

b）部署复杂性。将AI-ECN部署到实际网络中可能会存在一定的复杂性，需要在网络设备和路由器上集成AI模型，这可能需要硬件支持和软件定制，增加了部署的难度和成本。

c）实时性要求。AI-ECN需要实时监测和响应网络流量的变化，以调整拥塞控制策略。这对于要求低延迟和高性能的应用来说是一个挑战，因为AI模型的

推理和决策需要时间。

d）数据隐私和安全。在实际网络中使用AI-ECN可能涉及到敏感数据的处理，如网络流量信息，确保这些数据的隐私和安全是一个重要的考虑因素。

e）复杂性增加。AI-ECN引入了更多的复杂性和抽象层次到网络拥塞控制中，这可能使网络管理和故障排除更加复杂，因为管理员需要理解和维护AI模型的行为。

f）鲁棒性问题。AI模型可能对未知或异常情况表现出鲁棒性不足的问题，导致意外的网络问题。网络环境中的变化和异常可能会影响AI-ECN的性能。

3 智算中心高性能网络管控平台研发实践

3.1 流量管理平台架构设计

通过对现有拥塞控制方案的分析可以看出，智算中心需要具备管控平台，实现对网络流量以及拥塞控制信息的精细化采集与监控，实现拥塞控制算法的参数调优。对于突发问题的响应和故障排查也需要全面自动化的平台。

为实现智算中心无损网络流量管理，智算中心管控平台可以采用异构分布式计算架构，提供数据驱动的全方位、一体化的智算中心无损网络的运营、维护和监控。总体架构设计如图6所示。

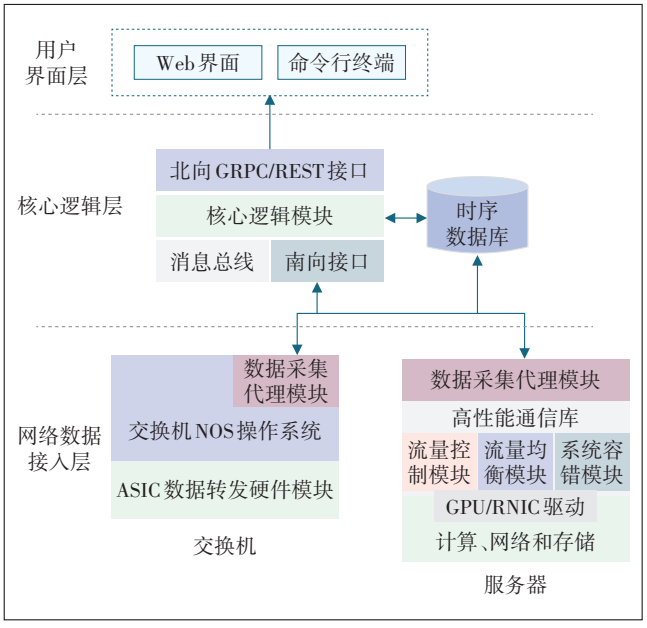

图6　智算中心无损网络流量管理平台架构

智算中心无损网络管控平台的架构从上至下包括用户界面层、核心逻辑层和网络数据接入层。

核心逻辑层南向连接交换机和服务器的数据接入层，北向提供统一的标准服务和接口，支撑命令行（CLI）和Web界面2种用户交互界面。命令行（CLI）界面和Web界面的功能类似，提供了所有的查询和配置命令，可运行定制的自动化脚本。

管理平台需要具备以下核心能力。

a）网络拓扑发现。自动发现网络的设备，包括交换机、服务器以及设备之间的拓扑连接，生成网络拓扑的视图。

b）网络基础设施的集中式管理。智算中心无损网络管控平台全覆盖所有网络设备的详细信息和运行状态信息，一站式的查询服务实现了网络基础设施的集中式管理。

c）自动化配置。平台提供内嵌式的设备配置模版，自动适配设备的功能，生成的配置可一键式地部署到所有设备上，配置校验更加确保了设备的正确性，减轻了配置设备的人工负荷。

d）超可视化监控。自动采集网络设备的运行状态数据以及流量的动态遥测数据，展示高可视化的网络运行视图和流量动态视图，全方位监控网络的运行状态，并有效预测网络流量的趋势，为网络的决策提供数据支持。

e）错误告警。及时给出网络不同级别的错误信息，及时进行故障恢复。

f）主机端侧融合网络功能。主机服务器集成高性能通信库，内嵌智能流量控制机制，实现流量超无损和流量超均衡。另外，在存储侧支持基于ROCE的高性能存储。

3.2 基于端网协同的解决方案

基于端网协同的解决方案主要包括以下2个方面。

a）通过近实时地对端侧（算力服务器）、网侧（spine，leaf交换机）的相关流和拥塞信息报文进行统计（PFC、ECN、CNP）来监测网络的拥塞状态。

b）根据持续的长期监控，智能归纳网络的拥塞特点，调整DCQCN等参数，并对leaf交换机上hash流的实现进行配置或路径指定，从而降低网络出现拥塞的概率，提升AI训练过程的数据同步效率和训练整体效率。

管控平台实现拥塞控制和负载均衡的优化流程如图7所示。

a）高频采集相关转发路径上端侧及网络侧的

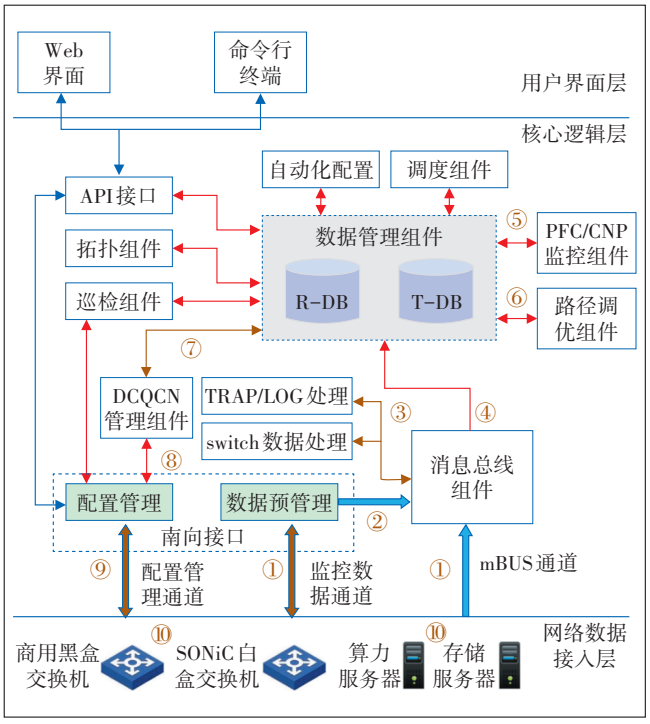

图7　拥塞控制和负载均衡的优化流程

PFC、ECN、CNP的报文统计计数，更新到数据库中。

b）管理平台通过PFC/CNP监控组件定期扫描数据库中各个交换机、算力服务器的PFC、ECN、CNP标记的报文统计数据，根据数据所属的端口信息获取链路连接信息。根据管理员配置的条件策略将产生拥塞的描述信息写入时序型数据库，并将连接关系写入关系型数据库中。

c）路径调优组件定期扫描PFC/CNP监控组件生成的数据信息，并根据生成的结果数据判断拥塞状态，然后根据设置的策略生成需要调整的算力服务器上的DCQCN参数，以及交换机上的流量路径及hash修改规则。然后调用自动化配置组件的接口，生成相关的配置命令，更新到数据库中。

d）DCQCN组件从数据库中读取相关的配置命令信息及相关的配置节点信息，组合为配置命令序列。

e）DCQCN将组合后的配置命令序列依次下发到南向接口的配置管理组件。对单台设备进行串行的配置下发处理，对多台设备根据策略进行并行的配置下发处理，提升整体效率。

f）配置管理组件将相关配置通过已建立的配置管理通道下发到相应设备上。

g）设备将执行结果向上回复给配置管理组件，配置管理组件再反馈给DCQCN管理组件。DCQCN根据配置结果确定后续处理（如成功就继续后续的配置下发，如失败则根据策略进行告警等处理）。

3.3 DCQCN配置调优实践

针对大模型场景的流量情况和网络拓扑状态，不同的DCQCN参数对训练的TFLOPS影响能达到50%，因此调试出适合当前高性能网络环境的一套参数对提升网络利用率十分有效。结合英伟达DCQCN配置参数及实际调优经验，本文给出了DCQCN参数的默认值、调参范围及概念（见表1）。

表1　DCQCN参数默认值、调参范围及概念

参数	默认值	范围	描述
K_{max}	3 200	（1 000，4 000）	参数K_{max}和K_{min}分别表示队列开始触发拥塞通知的阈值和最大队列长度。P_{max}表示交换机标记ECN的最大概率，反映了网络对拥塞的灵敏度和响应速度
K_{min}	800	（10，1 000）	
P_{max}	0.2	（0，1）	
R_{AI}（rpg_ai_rate）/（Mbit/s）	50	（0，line rate）	积极增加阶段的速率增加值
R_{HAI}（rpg_hai_rate）/（Mbit/s）	100	（0，line rate）	超积极增加阶段的速率增加值
R_{MIN}（rpg_min_rate）/（Mbit/s）	100	（0，line rate）	此参数定义了QP队列的最小速率限制
a_{RI}（dce_tcp_rtt）	1	（0，100）	此参数定义α更新的间隔。如果在此期间收到CNP，则α递增；否则，它会递减
g（dce_tcp_g）	4	（0，1 023）	g是介于0到1之间的常数，会影响a 增速公式：$\alpha=\left(\frac{g}{2^{10}}\right)\alpha+\left(2^{10}-g\right)$ 降速公式：$\alpha=\left(\frac{g}{2^{10}}\right)\alpha$
R_{DI}（rate_reduce_monitor_period）	4	（0，100）	降速的时间间隔
RP_{TIMER}（rpg_time_reset）	900	（0，1 500）	增速事件的时间间隔
F（rpg_threshold）	4	（0，50）	快速恢复阶段的增加次数

在DCQCN算法中仅仅考虑端侧或者几个重要的参数往往不够全面，需要不同参数配合作用，要实现更好的结果往往需要端网同步调参，通过模拟退火算法能够快速得到最优参数。

为了对DCQCN的参数进行调优验证，搭建如下实践环境：采用4台服务器共32GPU卡，每台服务器共2个网卡，链路速度均为100 Gbit/s，组网拓扑示意如图8所示。

采用nccl-test对调参前后进行对比测试，对All Reduce（ar）、Broadcast（bc）、All Gather（ag）、All to All

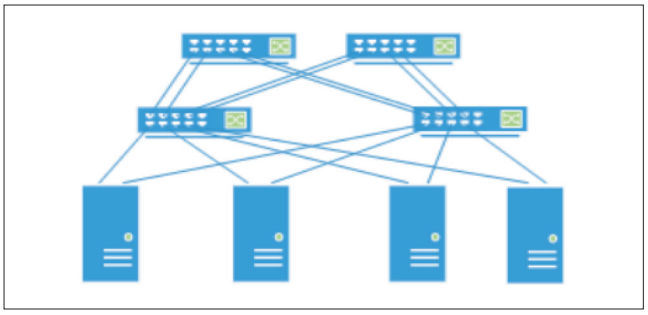

图8　组网拓扑示意

(a2a)4种集合通信操作各测试100组，同步消息大小为1 GB。实验结果统计如图9所示。其中横坐标是100组的分布比例，纵坐标是nccl-test输出的吞吐量，单位为GB/s。

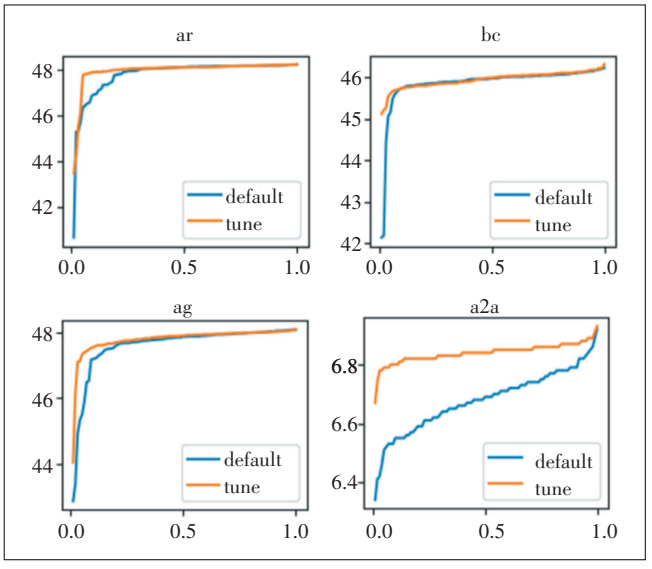

图9　AI-DCQCN调优结果

调参后，对于不同集合通信操作均有稳定性的提升，其中All to All的吞吐量提升最为明显。这是因为通过精细调整DCQCN参数（如降低拥塞通知的响应阈值和优化速率调整策略），可以更有效地管理数据中心的网络流量和拥塞。在All to All操作中，这种优化特别重要，因为数据包在多个节点间的大量交换容易引起网络瓶颈。优化后的参数设置可能有助于更平滑地处理这些数据包，减少拥塞和提高整体数据传输效率。因此在流量情况简单的高性能网络中，AI-DCQCN会有极大的性能提升。

4　结束语

无损网络能够提供低延迟、高吞吐量及高质量的网络服务能力，但也需要结合网络层、管控层和应用层各项技术，并面向各类不同智算应用进行优化，最终最大限度地利用计算和存储网络，从而提高智能计算中心的网络性能。

参考文献：

[1] 中国移动. 新一代智算中心网络技术白皮书[R/OL]. [2023-12-25]. http://221.179.172.81/images/20221213/95531670888719213.pdf.

[2] 中国科学技术信息研究所，AITISA，鹏城实验室. 人工智能计算中心发展白皮书2.0[R/OL]. [2023-12-25]. http://finance.people.com.cn/n1/2021/0926/c1004-32237080.html.

[3] 王少鹏，郑常奎，芦帅，等. 数据中心无损网络关键技术研究[J]. 信息通信技术与政策，2021(10)：68-74.

[4] 百度智能云. 智算中心网络架构白皮书[R/OL]. [2024-01-25]. https://zhuanlan.zhihu.com/p/654153304.

[5] DONG J B，CAO Z，ZHANG T，et al. EFLOPS：algorithm and system co-design for a high performance distributed training platform[C]// 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA). New York：IEEE，2020：610-622.

[6] 华为. CloudFabric解决方案Dragonfly直连拓扑技术白皮书[R/OL]. [2023-12-25]. https://e.huawei.com/cn/material/enterprise/1660f766d5e94870acb46c3b5db35113.

[7] WANG W Y，GHOBADI M，SHAKERI K，et al. How to build low-cost networks for large language models (withoutsacrificingperformance)? [EB/OL]. [2023-12-25]. https://arxiv.org/pdf/2307.12169v3.pdf.

[8] 新华三技术有限公司. 智能无损网络技术白皮书[R/OL]. [2023-12-25]. https://www.h3c.com/cn/Service/Document_Software/Document_Center/Home/Public/00-Public/Learn_Technologies/White_Paper/H3C-4103/?CHID=794344.

[9] 陈乐. 智能无损网络(HPC场景)[R/OL]. [2023-12-25]. https://support.huawei.com/enterprise/zh/doc/EDOC1100212165.

[10] 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.

[11] SHARADA Y. 盘点GPUFabric典型拓扑结构及拥塞控制技术[EB/OL]. [2023-12-25]. https://mp.weixin.qq.com/s/wBWnXemGMGzjkDbifFthXQ.

[12] 中国信通院，华为. 数据中心超融合以太技术白皮书[R/OL]. [2023-12-25]. https://e.huawei.com/cn/news/ebg/2022/data-center-hyperconverged-ethernet-technical.

作者简介：
韩博文，工程师，硕士，主要从事IP新技术和管控运维方向的研究工作；徐博华，高级工程师，硕士，主要从事IP新技术研究和新型网络设备研发工作；曹畅，正高级工程师，博士后，主要从事算力网络、下一代互联网等方向的研究工作；刘千仞，毕业于北京邮电大学，高级工程师，硕士，主要从事云计算、数据通信等规划相关工作。