智算中心Scale-Out网络的演进及GSE的实践

Evolution of Scale-Out Network in Intelligent Computing Centers and Practice of GSE

程伟强/CHENG Weiqiang[1,2]，李新双/LI Xinshuang[3]，白艳/BAI Yan[2]，吕勇/LYU Yong[3]
（1. 东南大学，中国 南京 211189；
2. 中国移动通信有限公司研究院，中国 北京 100053；
3. 中兴通讯股份有限公司，中国 深圳 518057）
(1. Southeast University, Nanjing 211189, China;
2. The Research Institution of China Mobile, Beijing 100053, China;
3. ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202502003
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250425.1536.003.html
网络出版日期：2025-04-27
收稿日期：2025-03-05

摘要：探讨了人工智能（AI）大模型时代智算中心网络面临的技术挑战，重点分析了传统互联网协议（IP）网络在负载均衡和突发流量处理方面的局限性，并对比了基于以太网融合远程直接内存访问（RoCE）的优化与网络架构重构两种技术路线。研究聚焦中国自主研发的全调度以太网（GSE）技术，详细阐述了其核心技术：基于报文容器（PKTC）的负载均衡机制和动态全调度队列（DGSQ）端到端拥塞控制技术。这些技术有效解决了智算网络中的流量极化和拥塞丢包问题。同时，系统分析了GSE网络设备在接口设计、转发引擎和队列管理等关键环节的创新架构，论证了GSE技术在构建高带宽、低时延、无阻塞新型网络方面的技术优势，为智算中心网络演进提供了重要参考。

关键词：AI大模型；智算中心；Scale-Out；GSE；RoCE；负载均衡；拥塞避免

Abstract: The technical challenges faced by intelligent computing center networks in the era of large-scale artificial intelligence (AI) models are discussed, focusing on analyzing the limitations of traditional Internet Protocol (IP) networks in load balancing and burst traffic handling. Two technical approaches are compared: optimization based on remote direct memory access over converged Ethernet (RoCE) and network architecture reconstruction. The research centers on China's independently developed global scheduling Ethernet (GSE) technology, detailing its core innovations: the packet container (PKTC)-based load balancing mechanism and the dynamic global scheduling queue (DGSQ) end-to-end congestion control technology, which effectively addresses traffic polarization and congestion packet loss in intelligent computing networks. Additionally, it systematically analyzes the innovative architecture of GSE network equipment in key areas such as interface design, forwarding engines, and queue management, demonstrating the technical advantages of this approach in building high-bandwidth, low-latency, and non-blocking next-generation networks. The findings provide important insights for the evolution of intelligent computing center networks.

Keywords: AI large-scale model; intelligent computing center; Scale-Out; GSE; RoCE; load balance; congestion avoidance

引用格式：程伟强, 李新双, 白艳, 等. 智算中心Scale-Out网络的演进及GSE的实践 [J]. 中兴通讯技术, 2025, 31(2): 14-20. DOI: 10.12142/ZTETJ.202502003

Citation: CHENG W Q, LI X S, BAI Y, et al. Evolution of Scale-Out network in intelligent computing centers and practice of GSE [J]. ZTE technology journal, 2025, 31(2): 14-20. DOI: 10.12142/ZTETJ.202502003

1 AI大模型对网络的挑战

以生成式预训练变换器3.0（GPT 3.0）为代表的大模型展现惊人的能力后，人工智能（AI）呈现向海量参数大模型方向发展的技术趋势。随着算力的提升与数据资源的不断扩充，大模型的参数量级将继续扩大。目前，已有多个大模型的参数规模超过了万亿级别。

当前，单独的计算芯片和存储芯片已无法满足AI大模型对参数量和计算量的需求，形成了制约AI技术发展的“算力墙”和“存储墙”两大瓶颈。为此，业界普遍采用通过多计算节点构建高性能集群的方案，以整合分布式计算能

基金项目：国家重点研发计划项目（2024YFB2906600）

力和存储资源，从而突破算力和存储的双重限制。这种基于集群的分布式架构已成为应对上述挑战的主流解决方案[1]。

智算集群中节点间网络的通信效率直接影响集群的整体吞吐量和性能。AI大模型训练业务的网络流量具有以下特征：流数量少（低熵）、单流带宽高（大象流）、同步突发（Incast）等，这对传统基于以太网的IP网络架构提出了两大挑战[2-3]：

挑战1：传统基于流的等价多路径路由（ECMP）负载均衡技术在流数量较少时存在局限性，会导致交换网络中出现流量极化现象，从而造成链路负载不均。具体表现为部分链路拥塞而其他链路利用率不足，这会降低整体网络吞吐量，如图1所示。

挑战2：在集群节点通信过程中，当源端在不了解目的端接收能力的情况下持续发送数据，会形成分布式训练中典型的多对一通信模式。这种模式产生的大量Incast流量将导致网络设备队列缓存出现瞬时突发，进而引发拥塞甚至丢包问题，最终造成应用时延增加和吞吐量下降，如图2所示。

因此，如何构建适配大模型算力的高性能网络，突破现有Scale-Out架构的瓶颈，已成为智算时代亟待解决的关键技术难题。

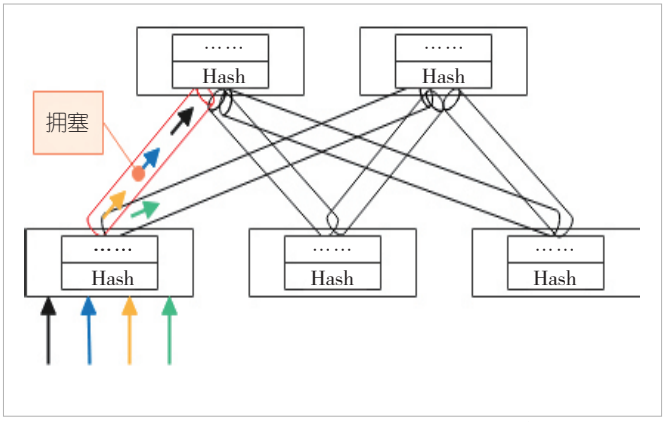


图1　等价多路径路由选路不均造成网络拥塞

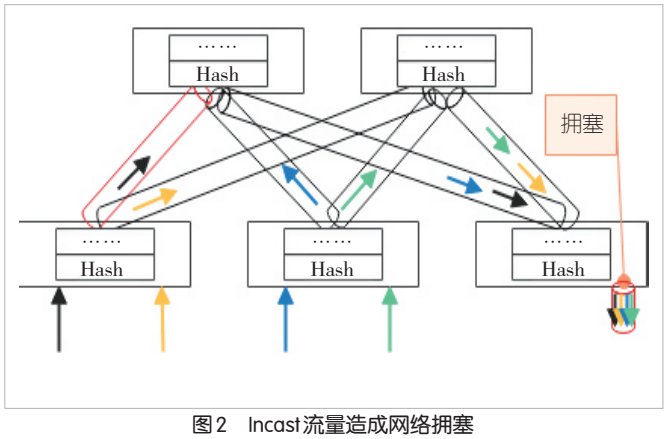


图2　Incast流量造成网络拥塞

2 Scale-Out网络的优化探索

为应对当前网络挑战，人们在探索过程中提出了多种优化思路。根据底层转发优化处理方式的不同，这些方案可归纳为两条发展路线：优化路线与重构路线，具体如图3所示[4-5]。

1）基于RoCE的优化

在传统基于融合以太网的远程直接内存访问（RoCE）技术的基础上进行创新优化，通过引入新技术可以改进智能计算Scale-Out网络架构。该方案保持网络底层转发逻辑不变，使现有网络更好地适配算力流量特征，有效降低Scale-Out网络对计算性能的制约影响。

方式1：纯网络侧的优化。纯网络侧的优化是设备厂商倡导的技术方案。该方案是基于网络设备技术升级的无损网络优化方案，主要包括：（1）通过快速显式拥塞通知（ECN）功能降低队列深度带来的时延影响；（2）采用AI ECN智能调优技术简化复杂的水线参数配置；（3）利用智能全局负载均衡提升少流场景的均衡性能。此类创新技术方案持续涌现，推动着网络性能的不断提升。

方式2：端网协同的优化。端网协同优化是互联网厂商倡导的技术方案，其核心在于通过终端侧的性能优化及网络状态感知，动态协调计算能力与网络资源，实现系统整体效能的提升。例如：阿里μFab方案采用智能网络调度机制，通过μFab-E网卡主动发送探测报文（probe），由μFab-C交换机动态反馈路径带宽和时延信息，基于这些网络状态数据实现网卡级智能限速及动态路径选择；阿里高精度拥塞控制算法（HPCC）与谷歌CSIG（一种用于网络拥塞控制的协议）方案采用端网协同机制，通过网侧随路采集拥塞状态信息，实现端侧流控参数的精细化调优；腾讯的星脉方案，基于多轨异构亲和部署策略，结合自研的拓扑感知集合通信库（TCCL），显著提升网络通信性能，已在多个场景成功落地。

2）网络架构重构

RoCE的优化建立在现有网络基础之上，这种方式仅能缓解算力与网络之间的冲突，并不能从根本上解决问题，因此称不上是最优解决方案。为彻底摆脱网络困境，业界各方希望构建全新的网络及底层转发机制，突破无损以太网的性能瓶颈，实现无阻塞、高带宽、超低时延，以契合AI与高性能计算对新型网络的需求[6]。

在重构路线方面，有两大主流技术方案在业内具备广泛的影响力：

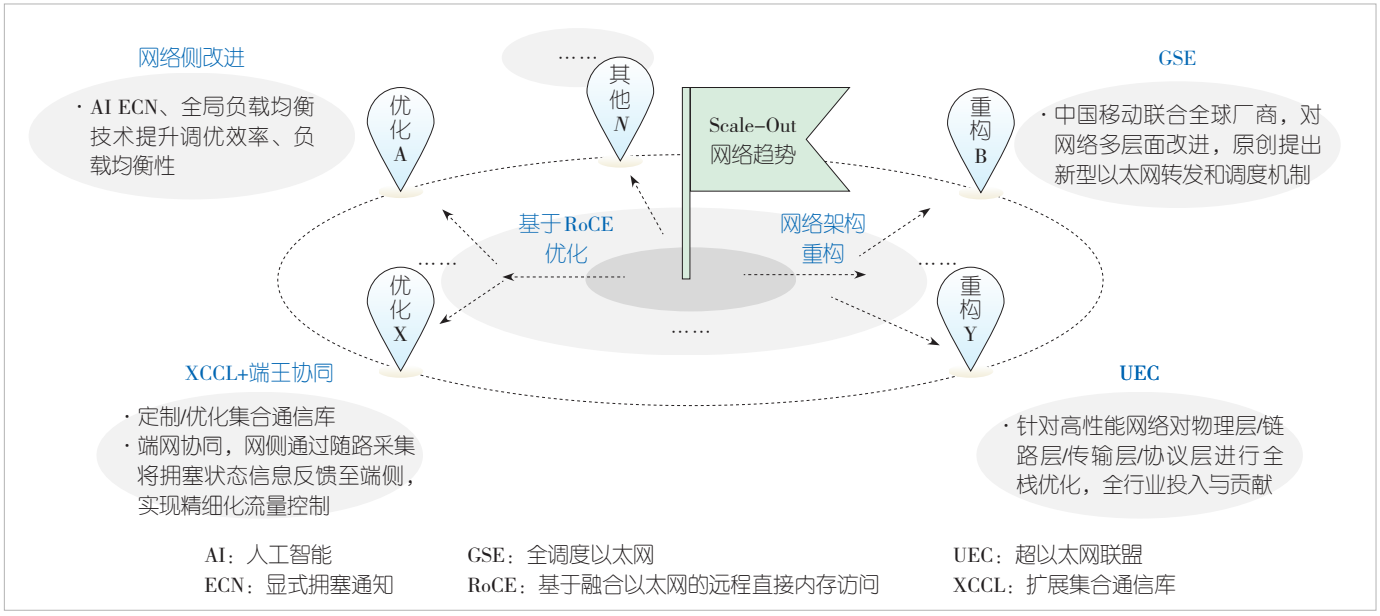


图3　Scale-Out网络路线与实践

（1）由全球多家颇具影响力的企业主导成立的超以太网联盟（UEC）。该联盟专注于高性能网络，对全通信栈开展深入优化工作，积极整合全行业资源，全力投入并贡献力量。其核心目标在于有效解决大规模数据传输过程中存在的尾部延迟等棘手问题，进而达成最佳的算力性能表现与网络利用率。

（2）由中国移动发挥牵头作用，联合多家中国厂商共同提出的全调度以太网（GSE）。此方案原创提出新型以太网转发和调度机制，将工作重点置于解决网络拥塞以及负载均衡等关键问题上。

在打造新一代网络的架构演进中，GSE技术已进入实践阶段。自2023年5月白皮书首次提出GSE概念以来，通过行业各界的通力合作，该技术在标准制定、设备研发等关键领域均取得显著突破。

3 GSE网络的实践

3.1 GSE关键技术

为应对智算网络流量特性所引发的网络拥塞难题，GSE引入两项关键技术：基于包容器的负载均衡技术以及端到端拥塞避免技术。

3.1.1 基于包容器的负载均衡技术

在负载均衡策略的抉择上，GSE采用喷洒技术，并提出以等长包容器作为喷洒单位的方案。这一举措不仅确保了负载分担的均匀性，显著降低了数据传输中的乱序程度，减小了后续保序操作所需付出的代价。

由于以太网支持变长包长传输，若仅采用简单的逐包喷洒方式，当包长可变时，极易导致负载分担不均衡。为化解因变长包引发的喷洒分担不均难题，业内运用切包与拼包两项技术加以应对。

切包技术（如图4所示）是指把数据包切割成等长信元后进行喷洒操作，在目的端再对信元进行重组以恢复数据包并实现转发。不过，该技术既要执行信元切分，又要进行组包，实现过程颇为复杂，并且每个信元都需要额外添加信元头，这无疑会造成较大的带宽开销。

拼包技术的运作机制是，将多个数据包组合成等大小的聚合帧后进行喷洒传输，在目的端对聚合帧解帧，还原出原始数据包并予以转发，具体过程如图5所示。相较于切包技

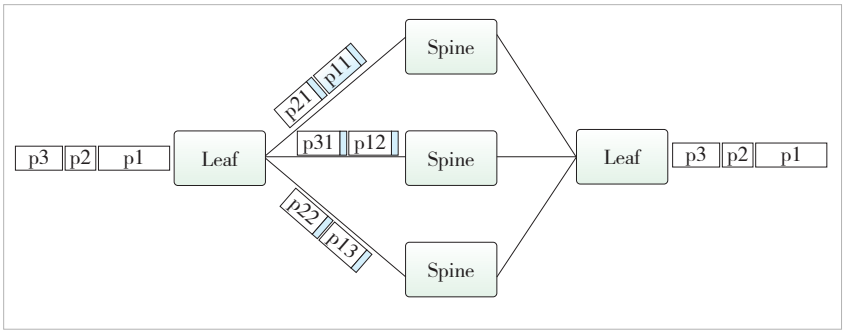


图4　切包喷洒示意图

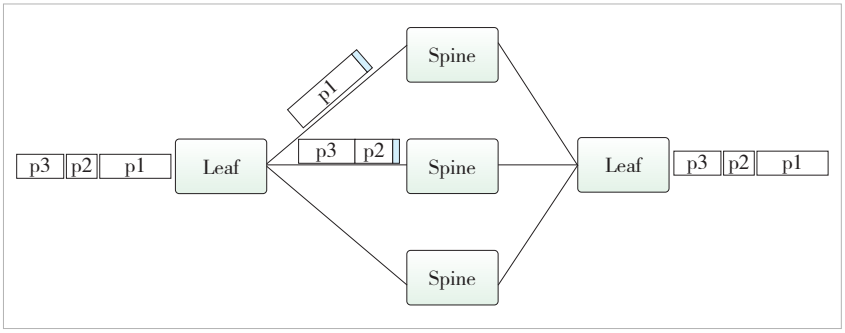


图5　拼包喷洒示意图

术，拼包技术在实现上更为简易，多个数据包只需共用一个聚合帧头，这样可以显著降低带宽开销。然而，拼包过程会引入不确定的等待时长，这就导致转发时延抖动情况难以确定。在对同步性要求较高的智算场景中，这种不确定性会对整体性能产生负面影响。

GSE融合了逐包喷洒低时延以及拼包交换高均衡性的优势，引入了基于报文容器（PKTC）的转发与负载均衡机制，具体如图6所示。该机制会把发往相同目的地的数据包，整合组装成“定长”的虚拟容器来进行转发操作。在这一过程中，同一容器内的数据包均被标记相同标识，确保它们沿着相同路径转发，以此实现保序传输。在进行负载均衡调度时，该机制以报文容器作为分担单位。值得注意的是，这里的报文容器属于逻辑概念，并非是实体，因而在对数据包进行组装与还原的过程中，无需额外的硬件投入。每个数据包仅需添加报文容器标识即可。与切包喷洒技术中每个信元都要添加信元头的做法相比，该方法极大地降低了带宽损耗，达到更好的效果。

我们假设网络中有N条等价路径，流量的包长在[64 B、1 500 B]之间随机分布，需要分担的总流量大小为F_{total}字节，路径的平均负载为L_{avg}，则$L_{\text{avg}}=\frac{F_{\text{total}}}{N}$。在传统逐包喷洒中，路径$i$上的负载$L_i$与其承载的包长总和成正比，即$L_i=\sum_{k=1}^{M_i}P_k$。其中$M_i$为路径$i$上的包数，$P_k$为随机变量表示包长。总体方差$\text{VAR}(L)=\frac{\sum_{i=1}^{N}(L_i-L_{\text{avg}})^2}{N}$，由于$P_k$的变异性，$L$的方差较大，导致负载分布不均。采用PKTC技术后，每个容器被设计为等长C，路径i上的负载为$L_i=C\times K_i$，其中，K_i为该路径上的容器数。由于容器可均匀装载，则$K_i\approx\frac{F_{\text{total}}}{N\times C}$，那么$L_i\approx\frac{F_{\text{total}}}{N}$。总体方差$\text{VAR}(L)\approx\frac{\sum_{i=1}^{N}(\frac{F_{\text{total}}}{N}-\frac{F_{\text{total}}}{N})^2}{N}$。此时，负载方差显著减小，趋近于零。

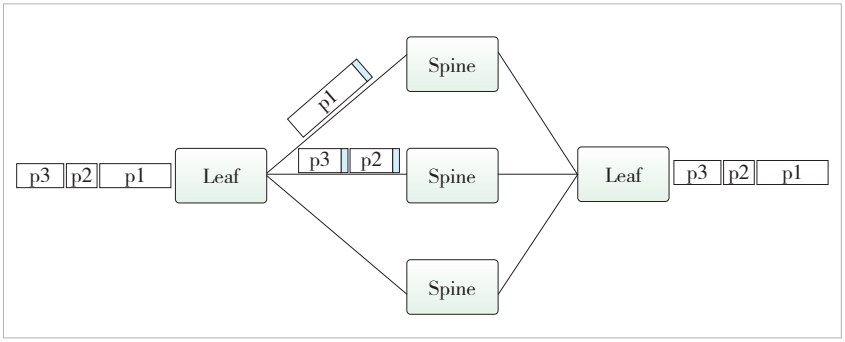


图6　容器喷洒示意图

PKTC技术确保同一容器内的数据包沿相同路径传输，有效减少数据乱序情况与时延抖动问题。假设网络每条路径的时延D_i服从正态分布$N(D,\sigma^2)$，其中，D为平均路径时延，σ为网络抖动的标准差，两路径时延差$D_1-D_2\sim N(0,2\sigma^2)$（两独立正态分布之差），标准差为$\sqrt{2}\cdot\sigma$。在传统的逐包喷洒机制里，每发送完一个报文，随即更换传输路径。一旦两条路径之间的时延差大于报文的发送间隔，便会导致乱序现象的产生。尤其是在数据包连续发送的高负载场景下，报文的发送间隔几乎趋近于0，乱序的条件可以简化为$D_1-D_2>0$：$\text{P}(D_1-D_2>0)=\text{P}(Z>0)$，其中$Z=\frac{D_1-D_2}{\sqrt{2\sigma^2}}\sim N(1,0)$，对于正态分布$N(0,2\sigma^2)$，均值为0，$\text{P}(D_1-D_2>0)=0.5$。采用PKTC技术后，同一容器内的数据包到达顺序始终保持一致，从根本上杜绝了容器内部的乱序情况，乱序概率降为0。不过，在不同容器之间，前一个容器的最后一个报文与下一个容器的第一个报文之间，存在出现乱序的可能性，而这一乱序概率与传统逐包喷洒技术的乱序概率相同。假设容器平均容纳m个包，则两个连续的容器一共$2m$个包，乱序的概率为$\frac{0.5}{2m}$，相比传统逐包喷洒显著下降。借助定长容器的设计，GSE从理论层面达成了负载在多路径间的均衡分配。这一成果有效提升了网络吞吐量，优化了资源利用率，为智算网络的高效稳定运行筑牢根基。

3.1.2 端到端拥塞避免技术

为有效应对Incast流量拥塞问题，GSE采

取了发送方预请求机制。发送数据前，发送方需先向接收端请求发送权限，接收端则依据自身接收能力，向发送方授予相应信用（即授权）。这一机制确保发送方的数据发送量不会超出接收端的接收能力，以此实现网络拥塞的有效避免。

假设接收端的处理能力为R，发送方为$S_1,S_2,\cdots,S_n$的发送速率为$r_1,r_2,\cdots,r_n$。在传统网络中，若$\sum_{i=1}^{n} r_i > R$，就会发生拥塞。采用GSE技术后，接收端根据自身能力授予总信用$C \leqslant R$，并分配给各发送方c_i，满足$\sum_{i=1}^{n} c_i \leqslant C \leqslant R$。因此，发送速率被限制为$\sum_{i=1}^{n} r_i \leqslant R$，这在理论上避免了拥塞的发生。

GSE基于图形处理器（GPU）间实际流量状况，动态构建虚队列，以此减少网络设备所需的队列数量资源。虚队列的调度依据接收端所授予的权限来执行。如此一来，来自不同源设备、发往同一目的端口的多个虚队列，能够依据目的端口的发送能力，进行统一的发送调度，进而达成整个网络的全局调度效果。GSE将这一技术命名为动态全调度队列（DGSQ）。借助DGSQ技术，GSE网络得以实现端到端的拥塞避免，具体原理如图7所示。

当网络出现Incast拥塞时，该方案通过将拥塞流量分布式缓存在参与传输的多个源端设备上，实现了全网缓存资源的协同利用。相较于传统方案仅能在目的端设备缓解Incast突发流量，这种分布式缓存机制使得网络整体缓存效率得到数量级提升，从而更有效地吸收突发流量。采用DGSQ技术后，当源设备本地缓存达到预设阈值时，系统会通过本地PFC机制直接通知源GPU降速。这一方案避免了传统方法中PFC信号需从目的端经Leaf-Spine网络反向传输的问题，从而有效防止了Fabric网络中可能引发的PFC风暴及其导致的网络性能骤降。

3.2 GSE性能测试

GSE基于包容器的负载均衡和端到端拥塞避免两种关键技术，实现了多路径间的均匀负载分担，减少了乱序和时延抖动，提升了网络吞吐量和时延稳定性。同时，从源头避免了拥塞，提升了缓存效率和拥塞响应速度。理论分析表明，该架构能显著提升网络吞吐量与时延稳定性，尤其适用于智算网络中高带宽、高突发及高并发的流量场景。

我们在实验室搭建了32张GPU卡的测试环境（每卡配备200G端口），在相同模型和组网拓扑下对比测试了GSE与RoCE网络的性能。如图8所示，测试网络分别由GSE原型设备和RoCE交换机构建。针对LLama2-13B大模型的测试表明（如表1所示）：在单任务、多任务及Leaf上行链路故障3种场景中，GSE网络性能均显著优于RoCE，平均提升达47.7%。

3.3 GSE设备实现

以太网交换机的核心功能是数据报文转发。经过多年发展，主流厂商的转发架构已趋于标准化，通常包括接口模

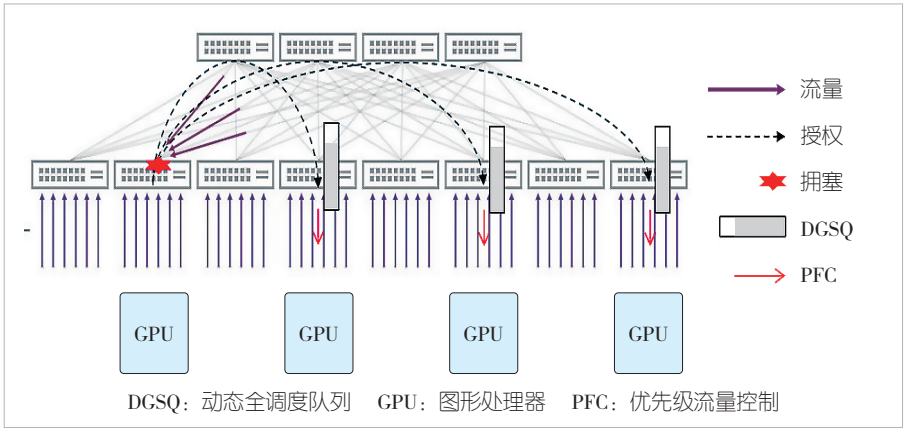


图7　端到端拥塞避免

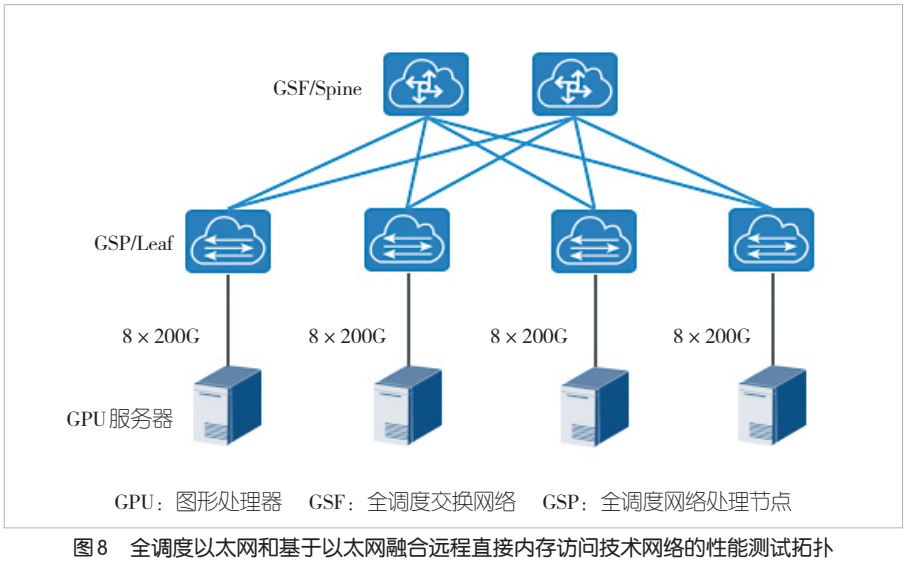


图8　全调度以太网和基于以太网融合远程直接内存访问技术网络的性能测试拓扑

表1　LLama2-13B大模型测试中RoCE网络与GSE网络测试性能数据对比

	RoCE网络性能/（samples·s^{-1}）	GSE网络性能/（samples·s^{-1}）	提升比例/%
单任务	2.66	3.91	47
链路故障(断1条)	2.43	3.84	58
链路故障(断4条)	2.43	3.53	46
多任务	2.45	3.43	40

GSE：全调度以太网　RoCE：远程直接内存访问

块、转发引擎、流量管理（TM）单元及辅助功能模块（如内嵌中央处理器、操作维护管理引擎等）[7-8]。

GSE在传统以太网交换机的架构之上，融入基于容器的负载均衡技术以及端到端拥塞避免技术。在接口、转发引擎以及TM子系统中增添特定功能特性，可以实现GSE设备的构建，具体实现架构如图9所示。

在设备的接口子系统中，我们增设链路质量监控与故障通告机制。该机制对接口上的前向纠错（FEC）误码情况展开持续监测，以此实现对链路质量的精准预测，确保在链路实际发生故障前，便能敏锐感知潜在问题。一旦监测到异常，立即以物理层编码的形式主动发出故障通告，让全网的GSE设备能够迅速察觉本地及远端故障，实时掌握端到端的可用带宽信息。随后，这些关键信息会及时反馈至授权和容器选路模块，将因链路故障导致的业务中断时长有效控制在微秒级，大幅提升整个网络的健壮性。

为增强设备的功能与适用性，在其转发引擎里新增对GSE头的解析、封装及转发能力。此新增功能的转发操作将严格遵循GSE规范要求，全面支持GSE开放生态。

在TM子系统中，我们将队列分配机制从静态分配升级为动态分配，并引入授权管理机制。通过实现队列和缓存资源的池化管理，系统能够根据实际流量需求动态分配资源，仅为活跃流分配队列和缓存。这种优化方案显著减少了大规模组网所需的队列数量，同时大幅提升了整网队列和缓存资源的利用效率。

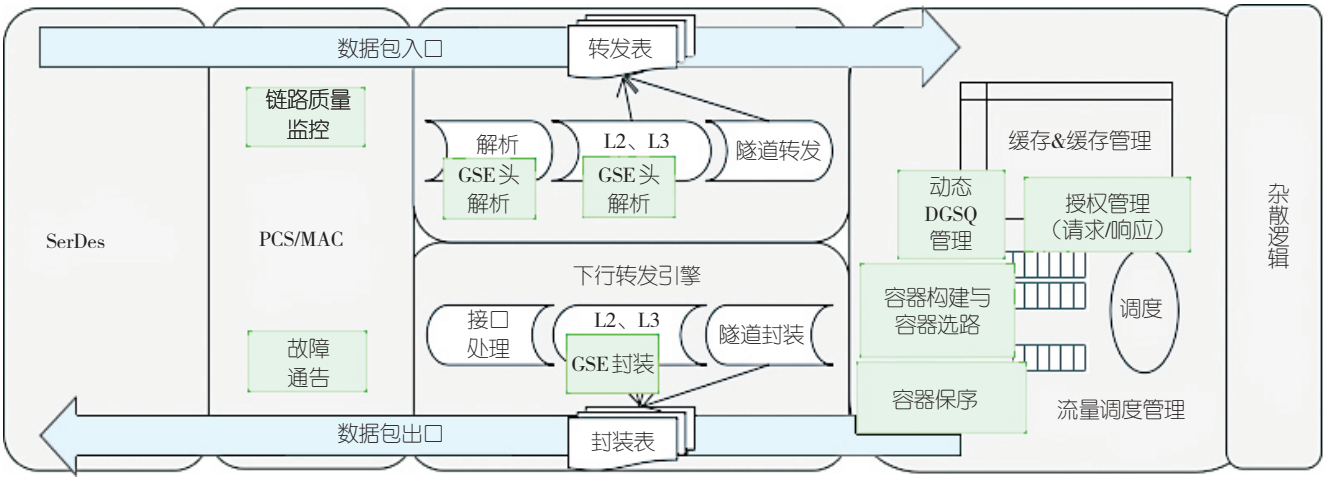
在TM子系统中增加容器构建与以容器为单位的负载均衡选路功能。系统通过系统通过累计算出队报文的包长，除以配置的容器长度，得到商值作为报文所属容器的ID，并以该容器ID为索引选择转发路径。

在TM子系统中增加容器保序功能，针对从远端接收到的、目的为本地直连GPU的流量，按目的地址进行容器级保序处理。该功能可消除因容器间负载均衡引入的报文乱序，避免将乱序传递给目的端而触发不必要的重传，从而提升系统整体性能[9]。

4 结束语

随着AI时代的到来，行业对适配大规模算力的网络需求日益迫切。当前人们对Scale-Out网络的探索不断深入，各类实践呈现百家争鸣之势，技术收敛与共识形成尚需时日。

其中，面向下一代网络重构的重要技术体系GSE，针对智算网络特有的低熵、大象流、同步突发等流量特征，我们创新性地提出几大核心技术：基于GSE封装头的转发机制、



DGSQ：动态全局调度队列　GSE：全调度以太网　MAC：媒体访问控制　PCS：物理编码子层

图9　GSE设备新增特性

基于PKTC的负载均衡、基于DGSQ的授权与调度等。这些技术有效解决了当前智算Scale-Out网络面临的诸多痛点：传统基于流的等价多路径负载均衡效果不佳，大规模Incast流量易引发拥塞甚至丢包等问题。当前，GSE相关硬件研发、协议创新及工程实践已进入快速发展阶段，多项核心技术通过原型机验证取得突破性进展，展现出卓越的性能表现和产业化潜力，正成为智能计算Scale-Out网络架构最具发展前景的技术路线之一。

参考文献

[1] 中国移动通信研究院. GSE 2.0网络侧优化技术标准 [R]. 2024

[2] 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]. [2025-03-02].https://arxiv.org/abs/2402.15627

[3] QIAN K, XI Y Q, CAO J M, et al. Alibaba HPN: a data center network for large language model training[C]//Proc of ACM SIGCOMM. ACM, 2024: 691-706

[4] 中国移动通信研究院. 全调度以太网技术架构白皮书 [R].2023

[5] 中国移动通信研究院. 面向AI大模型的智算中心网络演进白皮书 [R]. 2023

[6] WANG S, GAO K H, QIAN K, et al. Predictable vFabric on informative data plane [EB/OL]. [2025-03-02]. https://dblp. org/rec/conf/sigcomm/0028GQ0MLZZSGZF22.html

[7] 段晓东，程伟强，王瑞雪，等. 面向新型智能计算中心的全调度以太网 技术 [J]. 中兴通讯技术，2023，29(4)：57-63. DOI：10.12142/ZTETJ.202304011

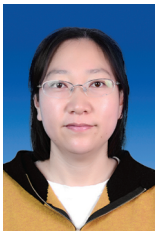
[8] 段晓东，陆璐，孙滔，等. 广域抗损高吞吐URDMA技术 [J]. 中兴通讯技术，中兴通讯技术，2024, 30(6): 23-30. DOI: 10.12142/ZTETJ.202406005

[9] 崔佳怡, 谢人超, 唐琴琴. 基于生成式人工智能的算力网络自智优化研究综述 [J]. 中兴通讯技术, 2024, 30(6): 54-62. DOI: 10.12142/ZTETJ.202406009

作 者 简 介

程伟强，东南大学网络空间安全学院在读博士，中国移动通信有限公司研究院基础网络技术研究所副所长，教授级高工；主要从事下一代互联网、数据中心网络、传输网等方面的技术研究和标准推动工作；参与IETF、ITU-T等10余项国际标准的制定。

李新双，中兴通讯股份有限公司承载网产品副总经理；在数据通信产品等相关领域拥有20余年工作经验。



白艳，中国移动通信有限公司研究院基础网络研究所项目经理；主要从事数据中心网络技术与方案研究工作。

吕勇，中兴通讯股份有限公司有线系统架构师；主要从事数据通信网络设备架构等的研究工作。