长距高性能网络关键技术研究

Research on Key Technologies for Long-Distance High-Performance Networks

王亚晨/Wang Yachen，杨峰/Yang Feng，耿竞一/Geng Jingyi
（腾讯科技（深圳）有限公司，中国 深圳 518038）
(Tencent Technology (Shenzhen) Co., Ltd., Shenzhen 518038, China)

DOI：10.12142/ZTETJ.202603002
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260626.0951.002
网络出版日期：2026-06-26
收稿日期：2026-05-10

摘要：针对AI大模型训练规模突破万卡、需跨园区长距组网的新需求，提出了一套端到端的联合优化解决方案：分级流控技术将任务完成时间（JCT）降低一个数量级；快速拥塞反馈机制将拥塞反馈延迟从毫秒压缩至数十微秒；逐流负载均衡协议消除端口负载不均问题；采用拓扑亲和的ZeRO 1策略替代ZeRO 3，大幅减少跨园区通信次数。实验网测试表明，通过整合上述关键技术，在千亿参数模型训练场景下可将跨园区训练的算力损失从10%~15%降低至5%以内，验证了长距网络支撑超大规模AI训练的可行性。

关键词：长距高性能组网；大模型训练；拥塞控制；负载均衡

Abstract: This paper proposes an end-to-end solution to address the new requirements of AI large-model training at the scale of over ten thousand GPUs and the need for long-distance, cross-campus networking. The hierarchical flow control reduces job completion time (JCT) by an order of magnitude. The fast congestion feedback mechanism compresses congestion feedback latency from milliseconds to tens of microseconds. Per-flow load balancing protocol eliminates port load imbalance. The topology-aware ZeRO 1 strategy instead of ZeRO 3 can greatly reduce the number of cross-campus communications.Testbed experiments show that by integrating these key techniques, in the training of models with hundreds of billions of parameters, the compute efficiency loss of cross-campus training was reduced from 10%-15% to less than 5%, validating the feasibility of long-distance networks in supporting ultra-large-scale AI training.

Keywords: long-distance high-performance networking; large-model training; congestion control; load balancing

引用格式：王亚晨, 杨峰, 耿竞一. 长距高性能网络关键技术研究 [J]. 中兴通讯技术, 2026, 32(3): 3-9. DOI: 10.12142/ZTETJ.202603002
Citation: Wang Y C, Yang F, Geng J Y. Research on key technologies for long-distance high-performance networks [J]. ZTE technology journal, 2026, 32(3): 3-9. DOI: 10.12142/ZTETJ.202603002

1 跨园区训练的需求

2022年ChatGPT的发布掀起了全球生成式人工智能（AIGC）热潮，以大语言模型（LLM）为代表的AIGC演变为新一轮全球竞争焦点。在LLM的演进历程中，模型参数量与算力成为驱动模型能力跃迁的关键因素，三者的关系被描述成“规模定律”，即Scaling Law。以OpenAI的GPT系列为例，从GPT-3到GPT-4，参数量从1 750亿（175B）提升至约1.8万亿（1.8T），增长了约10倍；训练数据量则从0.3万亿Token激增至13万亿Token，增长了约42倍；配套使用的图形处理器（GPU）卡数量也从千卡规模增长到万卡规模。最新一代的GPT-5更是使用了10万张H100 GPU。

随着大模型训练规模的增长，从单个园区获取全量GPU卡面临较大的挑战。这些挑战包括两个方面：

1）电力瓶颈。20万张H100集群总功耗直逼吉瓦级。单一数据中心受限于供电容量、散热成本及土地资源，无法承载超大规模GPU部署[1]。例如，Meta训练Llama 4需要10万卡级集群，被迫将算力分散至多栋楼宇，通过聚合层交换机（ATSW）实现跨楼互联[2]。

2）GPU碎片化问题。由于业务发展对GPU卡有强烈需求，一个新建成的GPU模块一般会交付多个业务部门。这导致一个业务部门的卡分散在多个园区。虽然通过置换卡可以延缓碎片的产生，但是仍存在停卡、沟通成本居高不下，以及最终换无可换等一系列问题，因此亟需配套技术来整合GPU碎片，满足百卡规模的训练需求[3]。

腾讯在2024年搭建了一张长距高性能实验网，以满足业务探索多模态混训的需求。本文基于该实验网的测试数据和为解决现网问题而研发的原型样机，阐述长距高性能网络的挑战、关键技术以及性能评估结果。

2 长距高性能网络的挑战

如图1所示，实验网使用了百公里光纤，400G级大端口、数据中心路由器（DR）连接了两个GPU园区。为了降低园区内产生拥塞的概率，本文规划了16∶1的收敛比（DR–核心交换机（GPULC）互联带宽是DR–DR带宽的16倍），使出园区的流量能够“一路畅通”地抵达DR，将最为重要的拥塞控制问题交给DR处理，最大程度地降低跨园区流量对园区内业务的干扰。

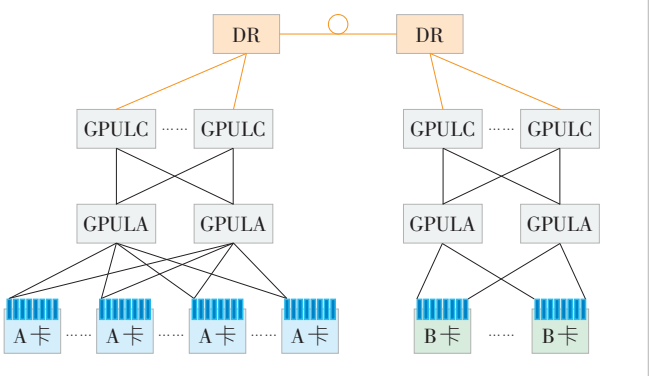
在未优化的长距实验网上进行集合通信基准测试（NCCLTest）和7B–dense/混合专家（MoE）训练测试后，本研究发现几乎所有的性能指标相对于数据中心内部（DCN）基线都出现了明显下降。实验表明，长距离高性能组网存在以下4个方面的挑战：

1）消息传输延迟显著增加导致数据传输欠吞吐

在NCCLTest的SendReceive测试中，本研究发现任务完成时间（JCT）出现了10～100倍的显著增长。受物理距离约束的固有传播延迟大幅提升了端到端往返时延（RTT），导致网络带宽–时延积（BDP）显著增加。在DCN窗口机制下，拥塞窗口（cwnd）规模无法匹配膨胀的BDP需求，形成传输瓶颈。这迫使数据传输长期处于欠吞吐状态，无论是流水线并行（PP）下的激活值传输这类小流，还是数据并行（DP）下的梯度同步这类大流，任务完成时间均出现数量级增长。

2）拥塞反馈不及时导致拥塞丢包和欠吞吐

在perftest中，本研究发现当前实验网存在丢包问题，链路利用率不到理论带宽的80%。这主要由拥塞控制算法数据中心量化拥塞通知（DCQCN）在长距控制器环路下响应失效引发：当传输距离达到百公里时，拥塞状态反馈产生1 ms以上的滞后，导致端侧无法及时响应动态变化。这种延迟一方面会使得交换机缓存溢出丢包，另一方面会使得发射端因过度降速而出现欠吞吐，导致链路的实际吞吐低于有效带宽理论值。



DR：数据中心路由器　GPULC：核心交换机
GPULA：接入交换机

图1　长距高性能实验网架构

3）随机哈希导致端口负载不均和流完成时间增加

根据大数定理，当流数量足够多时，随机等概的出端口调度机制可以实现交换机端口间理想的负载均衡。传统数据中心互联（DCI）流数通常在$O(10^6)$量级，因此DCI交换机通常采用基于五元组的哈希出端口调度方法。LLM训练任务产生的跨园区流（即五元组）数量通常为$O(10^3) \sim O(10^4)$，呈现所谓的低熵特征。这导致基于随机哈希的出端口调度机制出现显著的端口负载不均衡问题。

7B–dense的小规模测试显示，部分端口的负载超出平均值的20%以上，这意味着高性能DCI需要更精细化的负载均衡控制机制。

4）显存优化技术（ZeRO 3）导致大量小消息数据跨园区通信

跨园区的算力损失主要取决于跨园区的通信时间。当带宽为非瓶颈因素时，跨园区的通信时间主要取决于跨园区的通信次数。图2展示了跨园区采用的PP分割和DP分割方案。由于DP跨园区通信次数少，所以通常认为跨园区使用DP分割算力损失更小。然而，7B–dense小规模测试发现，DP分割出现了大量小消息数据跨园区通信，实测算力损失高达10%~15%。因此，跨园区训练需要使用拓扑感知的训练框架。

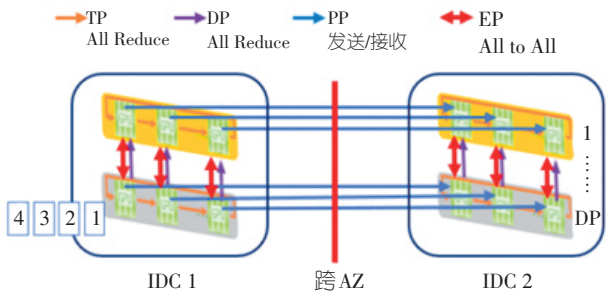
针对上述挑战，本文分别从长距流控技术、快速拥塞反馈、逐流负载均衡和训练框架优化4个方面提出对应的解决方案。

3 长距高性能网络关键技术

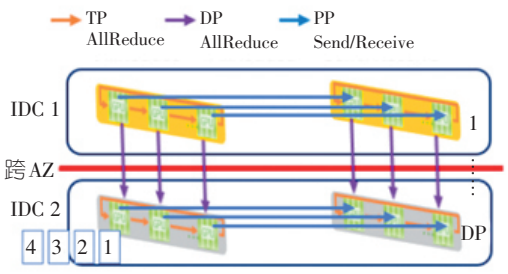
3.1 长距流控技术

DCN的RTT只有10 μs量级，对应的BDP为1 Mbit。因此，DCN通常采用1 Mbit左右的发送窗口。这样既可以充分利用带宽，又可以减少拥塞。而跨DCI场景下，RTT扩大100倍至1 ms量级，对应的BDP增加到100 Mbit。此时如果仍使用1 Mbit的小窗口，发送端会长时间等待接收端的确认而不发送数据，导致较为严重的链路欠吞吐问题。

如图3所示，为了提升长距链路的利用率，本文使用分级的流控方案：对DCN的流，采用小窗口；对跨DCI的流，则使用大窗口，以应对大RTT造成的确认延迟。这一方案需



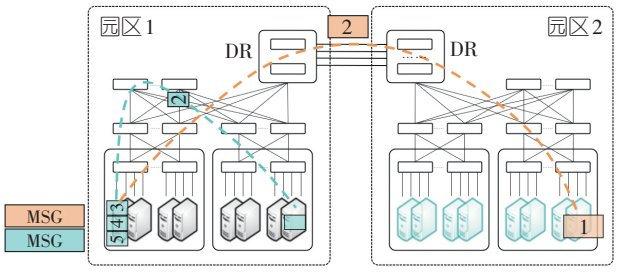
（a）PP分割



（b）DP分割

AZ：可用区　DP：数据并行　EP：专家并行　IDC：跨数据中心　PP：流水线并行　TP：张量并行

图2　跨园区PP分割和跨园区DP分割



DR：数据中心路由器　MSG：消息数据

图3　分级流控技术

要配合大容量缓冲区和快速拥塞控制算法，以便动态管理拥塞风险，确保数据传输的可靠性。对于跨DCI的流，发送窗口的大小接近BDP，即以链路带宽和传播延迟的乘积为基础，最大化长距传输的吞吐效率。这种分级方法基于网络拓扑的物理特性，实现了端到端流控的自适应，有效提升了分布式训练的整体性能和长距链路利用率。

3.2 快速拥塞反馈

本文使用快速拥塞反馈技术解决传统端到端拥塞通知机制在长距场景的反馈延迟问题。该技术将传统显式拥塞通知（ECN）的“交换机标记→接收端回传→发送端响应”三级延迟链路，缩短为交换机直连发送端的单跳拥塞通知包（CNP）信令通道以实现近端反馈。

快速拥塞反馈技术的难点包括：

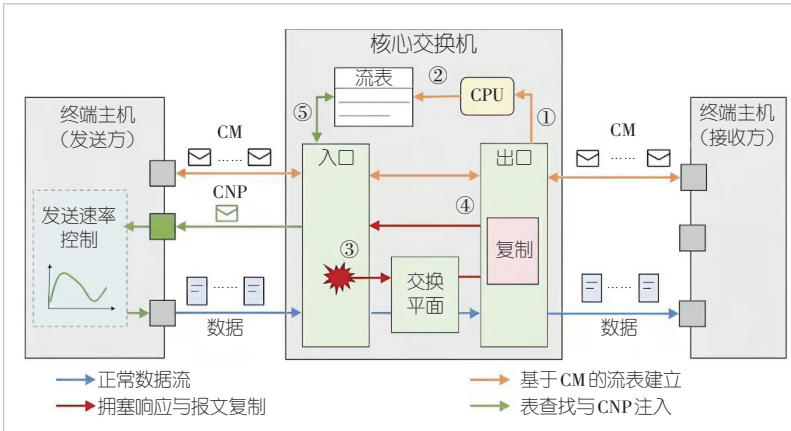
• CNP报文字段队列对编号（QPN）的获取。根据标准[4]规定，通信双方在建链的时候交换QPN信息，之后的通信中使用对方分配的QPN来标识该连接，因此CNP的QPN无法从数据报文解析中得到。

• CNP报文的实时性。有厂商曾尝试使用CPU产生CNP报文，由于软件难以保证实时性，CNP的延迟抖动超过了1 ms，完全无法满足快速拥塞反馈的要求。

为支持快速拥塞反馈，本文对终端和交换机均进行了“技术改造”。

为了让交换机可以获取QPN，本文采用了基于通信管理（CM）来建立通信链路，其整体流程和包含的步骤如图4所示。

步骤1：在远程直接内存访问（RDMA）连接建立过程



CM：通信管理　CNP：拥塞通知包　CPU：中央处理器

图4　快速拥塞反馈流程

中，客户端解析服务器地址并绑定到连接管理编号（CM ID）上，自动获取并管理QPN、PSN和重传上限等关键参数，随后发送请求报文至服务器。服务器接收请求后校验参数，通过应答报文确认连接并回送自身参数。客户端最终发送确认报文，完成链路建立。

步骤2：在RDMA建链控制面处理时，交换机“监听”CM建链请求，应答和确认3类信令报文（①），并将其上传至CPU（②）。

步骤3：CPU基于截获的信令报文建立流表，并使用QPN标识双向数据流路径。

步骤4：当交换机检测到端口缓冲区超过阈值时，将从数据包中提取三元组：发送源IP地址、接收目的IP地址、接收端QPN。

步骤5：获取这些元数据后，交换机CPU查询其内部流表，解析出与接收端QPN对应的发送端QPN。借助这些映射，交换机通过硬件流水线生成CNP报文，设置发送端QPN为目标接收者，从而直接通知发送端QP调整发送速率。

在交换机内部，为了确保拥塞通知的实时性，本文通过硬件处理流程快速生成相关报文。当检测到拥塞时，交换芯片会触发专用流程：首先，将正在经过交换机入口的拥塞报文添加生成通知标记；随后，在出口处理时复制该报文，原始报文继续正常转发，而副本则被重新导入入口处理；最后在入口处，根据内部配置修改副本的特定字段，将其转换为标准的拥塞通知控制帧。这种硬件实现方式有效保证了拥塞反馈的快速性。

3.3 逐流负载均衡

3.3.1 负载均衡优化方法

负载均衡在超大规模路径网络（如Clos架构）内通过集中式或分布式细粒度调度，将流量分散到不同端口上。调度的粒度由大到小可以分为flow级别[5–6]、flowlet级别[7–8]、packet级别[9–10]。以上分布式细粒度调度通常需要硬件、协议支持，如数据报文头修改、RDMA网卡（RNIC）匹配、交换机大缓存区等，难以利用存量机器完成端口级别调度。

此外，近年聚焦大语言模型训练JCT优化的负载均衡备受关注。集合通信库多以JCT为核心目标设计调度，兼顾链路均衡。例如，技术感知集合通信库（TACCL）[11]采用拓扑感知算法最小化GPU通信时间；流量工程集合通信库（TECCL）[12]重建模MCF问题，用混合整数线性规划（MILP）生成调度均衡负载。它们在容量约束下分配数据块，平衡传输效率与求解时间。

3.3.2 逐流出端口调度

针对跨园区训练流量调度不均的问题，本文提出一种轻量级精细负载均衡方案。

本文以最小化JCT为目标，对负载均衡问题进行建模，并提出一种解耦优化算法：在数据中心内部，采用随机哈希结合全局优化路由（GOR）进行局部调整；在DCI瓶颈处采用基于排队模型的轮询算法。该框架整合了哈希引导、混合整数规划（MILP）和排队模型，实现了精确建模与求解。实验表明，该方案可将JCT缩短约20%，并通过完全预计算的路由调度，实现了DCI与DCN协同优化，有效消除了负载失衡问题，并具有显著部署优势。

为实现以上基于轮询的出端口流调度，本文进一步设计了一种端网协同的逐流负载均衡协议。该协议的核心思想是由控制器下发流级别的调度决策，结合用户数据报协议（UDP）字段标记实现流到端口的精准映射，如图5所示。

整体协议的实现流程包含以下4个步骤：

步骤1：发送端策略标记。发送端交换机根据预计算策略，确定流在瓶颈处的分发端口，并利用GOR在UDP源端口字段中嵌入DR出端口标记，无需修改包头结构就能精准保留路径信息。该方案利用UDP头部字段中的源端口srcport

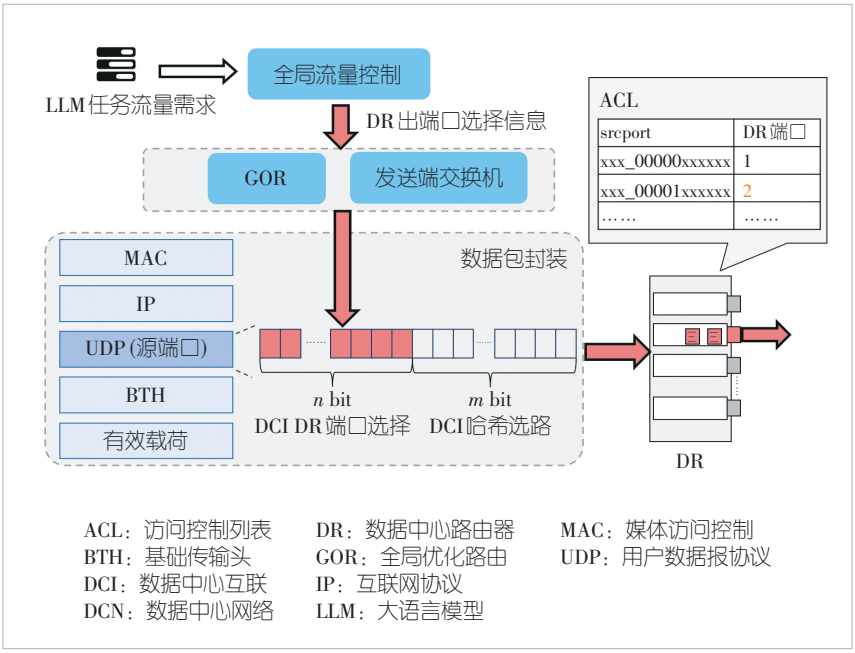


图5 逐流负载均衡流程

字段来承载负载均衡标签。srcport字段共2字节，分为高位n比特和低位m比特，分别用于DCI出端口选取和DCN内部哈希选路。

步骤2：DCN内部哈希选路。中间交换机建立srcport低位到路径的哈希映射，通过低位标记DCN内部的选路和到DR的哈希。实验表明，随机哈希选路冲突会导致可用带宽下降25%。为此，本文基于等价多路径路由（ECMP）哈希函数的线性特性，避免同设备对之间QP的路径冲突。设ECMP哈希函数为$H(p)$（p为流源端口部分字段），对于源端口字段p和$p \oplus \delta$，哈希结果仅取决于δ。只要$H(\delta) \neq H(0)$，即可保证两者采用不同哈希路径。

步骤3：DCI出口轮询分配。DCI出园区交换机用srcport高位作为流编号，通过ACL匹配预计算的轮询结果端口，实现出端口轮询分配。

步骤4：DCN负载监测与冲突调整。该协议实时监测DCN端口负载状态，对哈希冲突路径进行局部调整，并依托集中控制器GOR实时监测故障。收到拥塞信号后，控制器应用ECMP算法去除拥塞流所在的下一跳成员，将其调度至低负载路径，以此消除路径冲突。

3.4 训练框架优化

7B-dense小规模测试结果表明，DP分割方案中存在大量小消息数据跨园区传输。该问题源于框架利用ZeRO 3对DP进行优化[13]。ZeRO 3同时分割优化状态、参数和梯度，其通信流程在前向和反向传播中均需进行All-Gather操作以收集完整参数，最后通过Reduce-Scatter同步梯度。ZeRO 3的本质是以通信换取显存，该技术不适用于跨园区场景。对此，本文采用ZeRO 1方式，仅分割优化状态，以显存为代价，减少跨园区通信次数。

4 性能评估

本研究对长距分布式训练的各项性能指标进行了评估，具体实验网架构可参考图1。

4.1 长距流控的测试结果

在100 Gbit/s、30 km的DCI基础测试中，当发送数据大小为16 MB时，长距窗口优化可以将传输延迟降低至原有的1/7；当发送数据大小为256 MB时，该延迟优化效果提升至20倍。

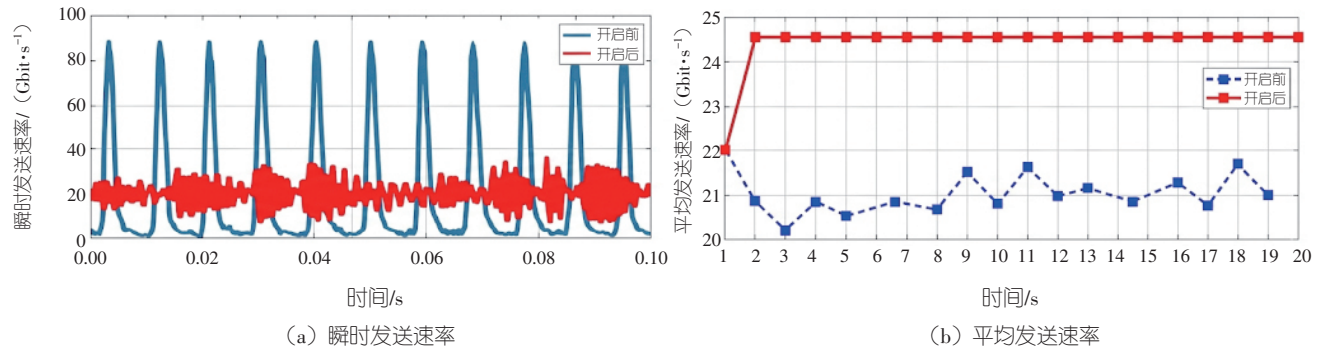
4.2 快速拥塞反馈的测试结果

CNP响应时间的测试结果表明，优化后CNP首包响应时间由1 ms缩短至240 μs，首包响应时间减少76%。当拥塞发生时，CNP频率会迅速增加。发送端触发反压后，速率快速下降，随后维持平稳状态。

图6为网卡平均发送速率和瞬时发送速率随时间变化的测试结果，其中网卡线速为100 Gbit/s，而交换机线速（链路瓶颈）为25 Gbit/s。结果显示，发送端平均吞吐量由20.42 Gbit/s提升至24.51 Gbit/s，链路利用率接近100%；同时开启快速拥塞反馈时，网卡瞬时速率不会在0～100 Gbit/s之间剧烈波动，而是稳定在25 Gbit/s附近，证实了快速拥塞反馈的有效性。

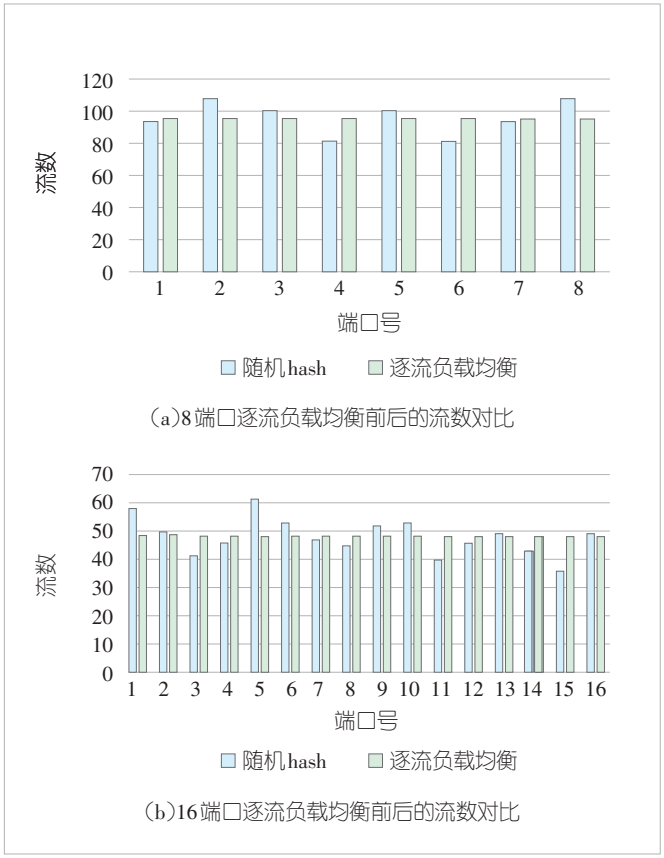
4.3 负载均衡的测试结果

图7为千卡级GPU集群测试中随机哈希和逐流负载均衡在8个出端口上的负载分布情况。随机哈希无法均分流量，负载最高的8号端口和最低的4号端口流数差值超过20%。受木桶效应影响，通信时间将被8号端口拉长。逐流负载均衡将所有的流平均分到8个端口，此时负载不均差异降为0，通信时间缩短。随着端口数量增大，随机哈希和逐流负载均衡在不同端口的流数差值超过40%，逐流负载均衡的优势进一步凸显。



（a）瞬时发送速率　　（b）平均发送速率

图6　开启快速拥塞通知包前后网卡瞬时、平均发送速率时序对比

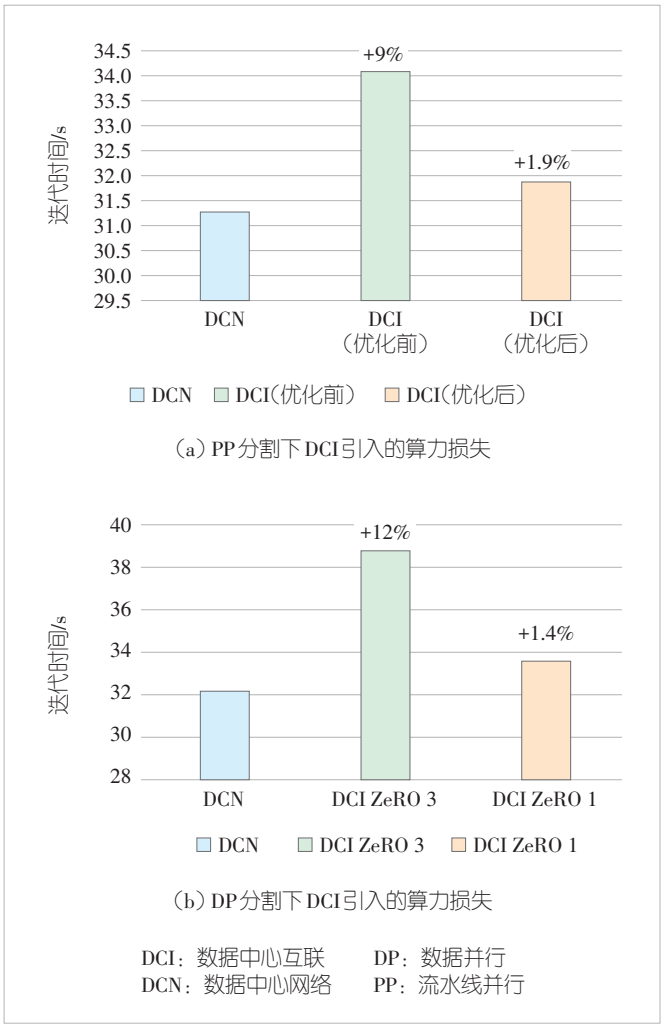


(a)8端口逐流负载均衡前后的流数对比

(b)16端口逐流负载均衡前后的流数对比

图7　随机哈希和逐流负载均衡的端口负载分布情况

信时间也随之增加，混训效率相应降低。在千卡集群下，混训效率从同园区的98%降至95%，整体算力损失为3%。



(a) PP分割下DCI引入的算力损失

(b) DP分割下DCI引入的算力损失

图8　小规模基准测试结果

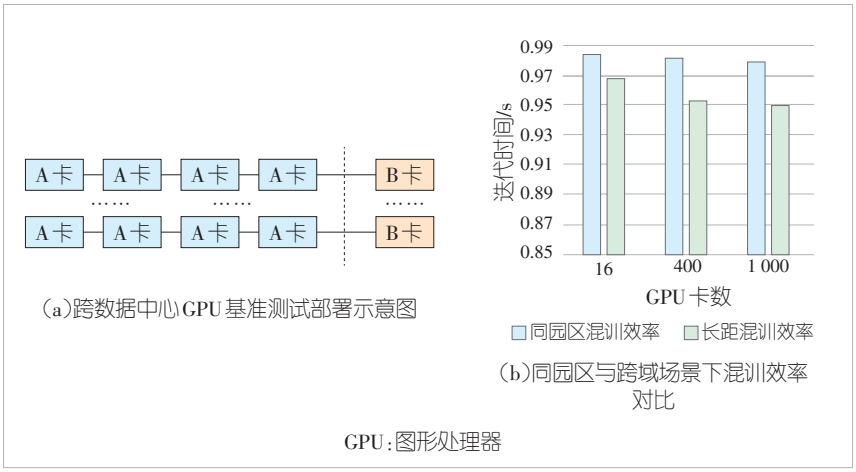
4.4 LLM训练测试结果

1）小规模基准测试

在7B-dense模型的跨园区分布式训练基准测试中，本文评估了PP与DP两种分割策略的性能。测试发现，跨园区网络的高延迟与小流控窗口会带来显著的额外算力开销。在PP分割中，优化流控窗口大小并启用快速拥塞控制，可将跨DCI引入的算力损失从9%降低至2%以内；在DP分割中，将内存优化策略从ZeRO 3切换为ZeRO 1，可避免大量跨园区参数同步通信，将算力损失从12%同样降低至2%以内，如图8所示。结果表明，针对跨园区网络特性进行通信优化，能有效控制性能损耗，使其与理论预期相符。

2）大规模基准测试

7B-dense模型的大规模基准测试使用了百卡级和千卡级两种集群配置，跨园区场景采用PP分割策略。如图9所示，随着卡规模的增加，跨园区通信的数据量线性增加，通



(a)跨数据中心GPU基准测试部署示意图

(b)同园区与跨域场景下混训效率对比

图9　大规模基准测试结果

5 结束语

本文采用长距流控技术、快速拥塞反馈机制和逐流负载均衡算法，并结合拓扑亲和的训练框架，实现了分布式训练端到端的联合优化。实验网测试表明，在千亿参数模型训练场景下，算力损失从10%～15%降低至5%以内，验证了长距网络支撑超大规模AI训练的可行性。

参考文献

[1] NVIDIA. NVIDIA introduces spectrum-XGS ethernet to connect distributed data centers into giga-scale AI super-factories [EB/OL]. (2025-08-22) [2026-04-16]. https://nvidianews.nvidia.com/news/nvidia-introduces-spectrum-xgs-ethernet-to-connect-distributed-data-centers-into-giga-scale-ai-super-factories

[2] Simonite T. Meta's next Llama AI models are training on a GPU cluster 'bigger than anything' else. [EB/OL]. (2024-10-30)[2026-04-16]. https://reportwire.org/metas-next-llama-ai-models-are-training-on-a-gpu-cluster-bigger-than-anything-else/

[3] Weng Q Z, Yang L Y, Yu Y H, et al. Beware of fragmentation: scheduling GPU-sharing workloads with fragmentation gradient descent [EB/OL]. (2023-08-15) [2026-04-16]. https://www.usenix.org/system/files/atc23-weng.pdf

[4] InfiniBand Trade Association. InfiniBand architecture specification Volume 1. Release 1.4 [Z]. 2020

[5] Al-Fares M, Radhakrishnan S, Raghavan B, et al. Hedera: dynamic flow scheduling for data center networks [EB/OL]. (2010-04-28) [2026-04-16]. https://www.usenix.org/legacy/event/nsdi10/tech/full_papers/al-fares.pdf

[6] Zhang Z H, Zheng H Y, Hu J Y, et al. Hashing linearity enables relative path control in data centers [EB/OL]. (2021-07-14)[2026-04-16]. https://www.usenix.org/system/files/atc21-zhang-zhehui.pdf

[7] Alizadeh M, Edsall T, Dharmapurikar S, et al. CONGA: distributed congestion-aware load balancing for datacenters [J]. ACM SIGCOMM computer communication review, 2015, 44(4): 503-514. DOI: 10.1145/2740070.2626316

[8] Vanini E, Pan R, Alizadeh M, et al. Let it flow: resilient asymmetric load balancing with flowlet switching [EB/OL]. [2026-04-16]. https://people.csail.mit.edu/alizadeh/papers/letflow-nsdi17.pdf

[9] Ghorbani S, Yang Z B, Godfrey P B, et al. DRILL: micro load balancing for low-latency data center networks [C]//Proceedings of the Conference of the ACM Special Interest Group on Data Communication. ACM, 2017: 225-238. DOI: 10.1145/3098822.3098839

[10] Li W X, Liu X Z, Zhang Y X et al. Revisiting RDMA reliability for lossy fabrics [EB/OL]. [2026-04-16]. https://www.cse.ust.hk/~kaichen/papers/dcp-sigcomm25.pdf

[11] Shah A, Chidambaram V, Cowan M, et al. TACCL: guiding collective algorithm synthesis using communication sketches [EB/OL]. (2021-11-09)[2026-04-16]. https://arxiv.org/pdf/2111.04867

[12] Liu X T, Arzani B, Kakarla S K R, et al. Rethinking machine learning collective communication as a multi-commodity flow problem [C]//Proceedings of the ACM SIGCOMM 2024 Conference. ACM, 2024: 16-37. DOI: 10.1145/3651890.3672249

[13] Rajbhandari S, Rasley J, Ruwase O, et al. ZeRO: memory optimizations toward training trillion parame ter models [EB/OL]. [2026-04-16]. https://aiichironakano.github.io/cs596/Rajbhandari-ZeRO-SC20.pdf

作 者 简 介

王亚晨，腾讯云副总裁、腾讯网络总经理，现任中国通信学会网络专委会和边缘计算专委会副主任委员、算力网络专委会委员；主要研究方向为智算网络、骨干网络、云网络系统；获省部级奖励12项；发表论文10余篇，拥有中国和国际授权专利30余项，制定国际标准20余项。

杨峰，腾讯高级网络架构师；主要研究方向为高性能DCI网络；拥有中国和国际授权专利20余项；发表论文10余篇。

耿竞一，腾讯DCI网络负责人；主要研究方向为超大规模DCI网络、高性能网络、网络虚拟化。发表论文10余篇，拥有中国和国际授权专利20余项，制定国际标准10余项。