高性能智算网络架构探讨与实践

High-Performance AI Computing Network Architecture: Exploration and Practice

黄光平/Huang Guangping，肖敏/Xiao Min，张征/Zhang Zheng
（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202603008
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260629.1134.002
网络出版日期：2026-06-26
收稿日期：2026-03-25

摘要：智算网络对超高吞吐、超低时延、超高可用性的极致性能需求，是传统网络架构的核心挑战和瓶颈。论述了一种新型高性能智算网络架构，其关键特征包括端网精细化协作、精准流控、无状态组播、轻量级在网计算等。基于智算网络的关键性能堵点，提出了一种新智算网络架构，其主要特征是端网双域协同、网侧定量反馈以及无状态组播，概述并分析了原型测试及主要结论数据。

关键词：智算网络；端网协同；无状态组播；精准流控；轻量级在网计算

Abstract: The extreme performance requirements of artificial intelligence (AI) computing networks, namely ultra-high throughput, ultra-low latency, and ultra-high availability, pose critical challenges and bottlenecks for traditional network architectures. A novel high-performance AI network architecture is presented, characterized by fine-grained host-network coordination, precise flow control, stateless multicast, and lightweight in-network computing. To address the key performance bottlenecks of AI networks, a new architectural design is proposed, featuring dual-domain host-network collaboration, quantitative network-side feedback, and stateless multicast. The corresponding prototype tests and primary performance data are outlined and analyzed.

Keywords: artificial intelligence network; host-network coordination; stateless multicast; high accuracy flow control; lightweight in-network computing

引用格式：黄光平, 肖敏, 张征. 高性能智算网络架构探讨与实践 [J]. 中兴通讯技术, 2026, 32(3): 40-44. DOI: 10.12142/ZTETJ.202603008
Citation: Huang G P, Xiao M, Zhang Z. High-performance AI computing network architecture: exploration and practice [J]. ZTE technology journal, 2026, 32(3): 40-44. DOI: 10.12142/ZTETJ.202603008

相对传统网络业务，高频且超大体量的数据传输是智算网络最典型的差异化业务特征。同时，智算网络对丢包和拥塞的容忍度极低，这对传统数据中心网络及广域网均构成重大挑战。传统网络在架构设计上本质是容损、容拥塞的，其设计目标仅为“尽可能减少”丢包与拥塞，因此无法满足智算网络的极致性能需求。

就丢包问题而言，在多业务共存的流量转发模式下，网络侧通过缓存与队列机制降低拥塞导致的丢包，但受限于缓存资源容量及多流间的资源竞争。在端侧，现有方案通过网络拥塞节点为业务报文添加拥塞标签并反馈至端侧，端侧据此进行业务流调速，以减轻网络拥塞压力。然而，这种依赖定性参数的端网粗放式协调机制，无法充分调度网络可用资源，难以从根本上解决拥塞问题。

在混合专家模型训练及推理场景中，高频且体量巨大的全互连（All-to-All）通信模式下，点对多点（P2MP）与多点对点（MP2P）流量不仅消耗宝贵的网络节点带宽资源，还会对端侧输入/输出（I/O）造成严重的数据交互瓶颈。然而，这种资源消耗与性能瓶颈并非不可避免。

基于此，本文提出一种全新的智算网络架构视角。针对拥塞问题，将解决范畴从单一网络域拓展至端网精细化协同，将当前网络对端侧的黑盒模式转变为白盒模式，即端网协同由定性参数升级为定量参数，实现业务需求与网络资源的精细化动态匹配，从而从根本上提前规避拥塞，而非被动感知拥塞后再作通告。在智算模型训练与推理的P2MP与MP2P场景中，分别引入无状态组播机制与轻量级在网计算机制，将关键性能瓶颈点（如端侧I/O）上的N条相同数据流，通过组播复制与在网聚合计算归并为1条数据流，从而彻底解决性能瓶颈问题。

1 精细化端网协同

面向大容量数据在广域网中的高性能传输，端网精细化

协同架构引入了端网协商速率机制。该机制向端侧主机动态分配并授权发送流量的配额，以主动防止拥塞；同时，基于配额进行资源调度、准入控制和流量控制，从而满足最优传送时间目标。具体而言，通过端网速率协商，应用端能够以更精细的方式高效、及时地调整发送速率，避免被动依赖端侧传输协议的调速机制，从而有效提升广域传输吞吐量；通过基于配额的资源预留，网络增强了对流量的主动调节与资源调度能力，保障所有任务的传输需求与资源供应；通过流量调度与准入控制，实现多流之间网络带宽资源的合理分配与动态调度，在满足高吞吐要求的同时避免慢流长尾现象，并实现集合通信多流传输的高精度同步。此外，该架构还采用快速拥塞感知及通告机制。当网络拥塞发生时，系统可在近源端进行快速反馈，从而迅速、准确地对流量速率进行通告与调整，及时缓解网络拥塞。

1.1 端网速率协商机制

突发大容量数据流量的传输可能引发网络瞬时拥塞、丢包与排队延迟。在传统拥塞控制机制中，端侧对网络带宽资源状态缺乏感知，导致调速出现锯齿效应，并使吞吐量下降。高性能广域网需引入端网协商速率机制，实现速率协同与拥塞避免。

精细化端网协同机制下包含3种速率协商策略[1]：

1）最优速率传输。网络通过资源调度机制为大容量数据提供服务质量（QoS）保障，实现最优速率传输，端侧主机可按协商的最优速率或最优速率范围进行传输。

2）最小速率传输。网络为大容量数据提供最小资源预留保障，实现最小速率传输，端侧主机可按不低于协商值的速率进行传输。

3）最大速率传输。网络为大容量数据提供资源预留的上限，实现最大速率传输，端侧主机可按不高于协商值的速率进行传输。

1.2 基于配额的资源预留

在智算场景下，任务式数据传输是主流模式，其主要特征为流量的周期性与可预期性。基于配额（quota）的调度是一种资源管理策略。该策略可扩展至成熟协议，例如资源预留协议–流量工程（RSVP–TE），使其携带配额信息。配额可定义为一定时间内的可用资源（包括带宽、队列、缓存等），涵盖动态资源与静态资源两类。静态资源如确定性链路，未申请资源预留的资源仅用于准入判断。网络可根据数据传输任务的需求分配和授权配额，并基于配额实施资源调度，从而实现主动拥塞避免，保障基于配额及对应速率的高效转发。网络配额保障策略需根据当前网络资源状态进行选择，为此网络需先收集资源信息，再依据具体任务需求选取合适的资源保障策略[2]。

1.3 流量调度及准入

网络侧在接收流量后，依据协商速率进行流量调度与策略执行，具体包括：对流量进行分类，按业务类型划分优先级，并为关键流量提升服务质量（QoS）等级；对流量进行整形，例如聚合“老鼠流”或分片“大象流”等。网络边缘的流量策略执行能够规范数据流，消除拥塞并最小化流量传输完成时间。基于协商速率实施速率调控时，可依托特定信令协议（如RSVP–TE）进行基于速率的动态资源调度与准入控制。其中，通过最小速率对应的最小带宽配额，保障单流传输完成时间；通过最大速率对应的最大带宽上限，避免多流竞争引发拥塞丢包[3]。网络节点应依据协商的QoS等级与速率执行准入与流量控制。该节点通过结合准入控制与拥塞控制，可在高效利用网络容量的同时，实现高吞吐量与低时延。

2 精准流控

在传统数据中心网络中，基于优先级的流量控制（PFC）是主流解决方案。拥塞节点通过专用的二层报文，向上游节点发送拥塞反压消息，并逐跳回传，直至到达数据发送源端。然而，随着跨数据中心智算业务的兴起，这类面向无损传输的流控方案需延伸至广域网，这对现有广域网提出了两大挑战：1）PFC逐节点模式要求广域网全网升级，部署代价高昂；2）广域网需承载多业务共存，必须极力避免特定业务（如智算模型训练与推理）的拥塞对其他业务流量造成严重影响。

相对于PFC基于队列的粗粒度流控模式，为适配广域网多业务、多租户共存共享网络资源的现实场景，本文提出一种面向租户或切片的精细化流控方案，即基于互联网协议第6版段路由（SRv6）的精准流控方案。如图1所示，当拥塞发生时，系统触发向指定源端的反压信号传输。该反压路径可沿SRv6路径向源端通告本地拥塞信息，包括队列状态、缓存深度等。与当前数据中心网络PFC机制不同，SRv6可由路径源端或控制器指定转发路径，因此内在地支持插花式组网模式——即无需所有节点均具备PFC报文处理能力，仅需在选定节点（如数据中心边缘节点、长距广域网中继节点）支持PFC段标识（SID）处理能力。PFC报文的目标节点则可基于SID精准指向特定租户的子接口或切片。目的节点收到PFC报文后，可根据具体拥塞状态，选择停止或调低

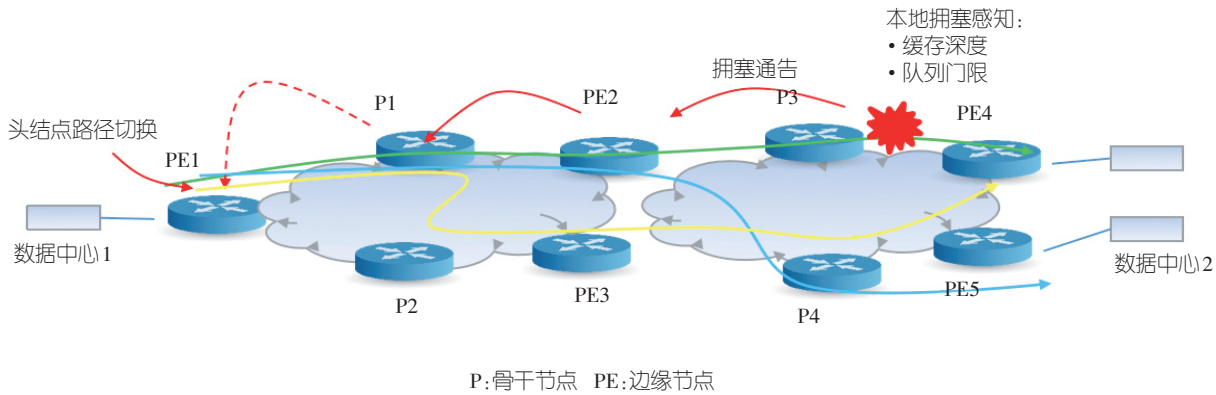


图1　基于SRv6的精准流控方案示意图

发送速率，或切换至备用路径。

拥塞控制的另一个成熟机制是显式拥塞通告（ECN）。在智算场景的极致性能需求下，广域网端到端的往返时延（RTT）成为关键瓶颈。为实现快速拥塞通告与反馈，网络拥塞节点可向就近的代理网关节点发送快速拥塞通告报文，再由该代理网关节点将报文转发至发送端设备。为此，系统需为每个端侧设备指定一个用于快速拥塞通告的代理网关节点，该节点应能够识别并解析端侧设备所支持的拥塞通告报文格式。代理网络节点通过内部网关协议（IGP）或边界网关协议（BGP）向外通告自身的拥塞通告代理能力及其所代理的端侧设备的IP前缀。网络中的其他设备收到该通告后，记录代理节点与端侧设备之间的映射关系表。当网络中发生拥塞时，检测到拥塞的网络节点根据拥塞报文的源IP地址查找对应的代理网络节点，并扩展互联网控制报文协议（ICMP）或用户数据报协议（UDP），向该代理节点发送快速拥塞通告报文，最终由代理节点将拥塞通告报文发送至源端设备[3-4]。

3 无状态智算组播

极致的传输性能要求使得智算网络同样需要极致的转发效率。在混合专家模型（MoE）下的All-to-All传输场景中，组播技术能够在I/O性能和带宽资源利用上带来显著收益。然而，传统组播技术如协议无关组播（PIM）通常需要在网络节点维护较重的状态信息，因此难以满足智算场景下组播传输的需求。基于位索引的显式复制组播（BIER）因其内生无状态且业务与网络分离的特性，具有明显优势，并能更好地满足动态性需求。

图2展示了MoE组播技术框架。在应用层与网络层的BIER之间，设置了一个网络适配层。该层负责将LLM的组播请求与BIER组播转发机制进行衔接，从而使BIER技术能够应用于智算MoE组播场景。

3.1 MoE专家组播实现

MoE通过将超大网络拆分为若干“小专家”，每次输入仅激活少数相关专家，从而在不增加计算量的前提下大幅提升模型参数量。MoE模型的核心流程分为两个阶段：首先，通过路由机制（Gating函数）为每个令牌（token）分配指定数量的专家；其次，在分发（Dispatch）阶段将令牌发送至对应专家，待选定专家完成计算后，再通过汇总（Combine）阶段对计算结果进行整合[5]。

在Gating过程中，由于专家选择过程极短（通常在毫秒级内），且选择具有高动态性，为实现此场景下的高效组播，BIER是最优的基础方案。

在应用层执行Gating操作并选定需要投递的专家后，将目的专家集合通知组播适配层，由适配层将其封装至网络层BIER报文头中。在BIER中，组播目的地址通过报文头中的位串（BitString）字段标识，BitString中的每一位均可对应一个目的节点。根据转发芯片的实现，BitString可支持64～4 096个目的地址。在MoE组播专家投递过程中，选定专家

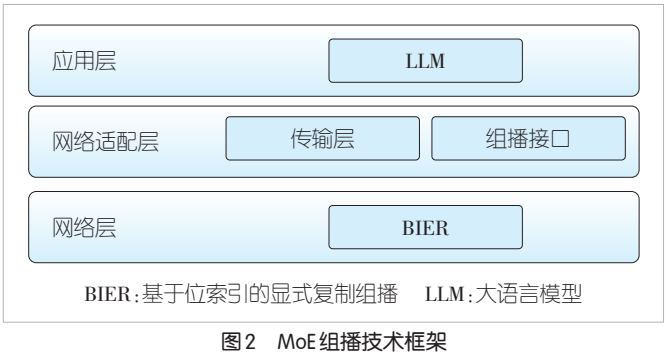


图2　MoE组播技术框架

群作为BIER组播转发的终点，数据经BIER封装后，Bit-String中的每一位表示一个目的专家。这些专家可位于同一节点内，也可分布在跨越Leaf或Spine设备的不同节点上。

在实际AI任务开始之前，需先对专家进行编号。例如，对于包含256个专家的大语言模型（如DeepSeek-V3），可将专家编号为1～256。每个专家IP地址与编号的对应关系，可通过两种方式建立：一是在节点内部交换机及Leaf交换机上进行静态配置；二是通过路由协议，如BGP、开放最短路径优先（OSPF）、中间系统到中间系统（IS-IS）或胖树路由协议（RIFT）进行自动通告。随后，节点内部交换机、Leaf及Spine交换机根据这些信息计算并生成BIER转发表项。至此，每个专家均可根据转发表项，查找到其对应的下一跳交换机或出接口。

每次Gating完成专家选择后，被选中的专家编号将直接映射至BIER报文的BitString字段，对应的位被设置为1。例如，DeepSeek-V3模型中若选中9个专家（含8个路由专家与1个共享专家），其编号分别为1、5、6、12、15、21、22、32、50，则BIER报文中相应位（第1、5、6、12、15、21、22、32、50位）均被设置为1。封装完成后的BIER报文，经节点内部交换机或Leaf/Spine交换机转发，最终到达目的专家。整个过程无需为选中的专家预先建立组播转发树，而是在Gating决策完成后直接封装并发送组播报文，从而极大提升了组播效率。

3.2 传输层协作

在智算MoE场景中，可能涉及大量内存操作，特别是远程直接内存访问（RDMA）的单边操作（如READ/WRITE）。这些操作对应于特定图形处理器（GPU）上不同的内存地址。在使用BIER技术时，存在两种不同的实现方式：

方式1：基于映射的转换方案

如果每个GPU为该AI任务分配的内存地址不同，则BIER数据传输后需写入各GPU对应的内存地址。当应用程序存在实现限制，无法直接交由网络适配层封装BIER报文时，也可采用此方式进行转换。

当网络适配层收到应用发来的报文时（该报文可能为基于RDMA协议的单播报文），为实现与BIER技术的结合与转换，传输层协议（如RDMA）可约定某些AI任务与操作对应的组播地址。此地址可以是传统组播IP地址，但这并非回退使用传统组播技术，而是指I/O接口、节点内部交换机或Leaf交换机在收到该报文后，可执行BIER报文的封装与发送。当数据报文丢失时，传输层可迅速重新发送丢失的报文，重传仍然采用BIER封装，仅需将未收到数据的专家对应的目的位填入BIER头即可。

BIER的封装转换可发生在GPU的I/O接口、相连的内部交换机或Leaf交换机上。这些节点仅需预先保存组播IP地址与BIER目的位之间的映射关系，收到报文后即可进行封装转换。报文经BIER组播转发到达目的后，同样依据映射关系，定位并写入各个目的GPU及其对应的内存地址。

方式2：统一内存地址方案

在开始AI任务之前，由参与该任务的所有GPU划分出统一的内存地址以供组播操作。例如，在训练或计算时，所有专家所在的GPU上均预先划分出一块统一的内存地址，用于执行READ或WRITE组播操作。此方案省略了映射环节，从而提高了BIER组播应用的效率。

3.3 多BIER域适配不同应用

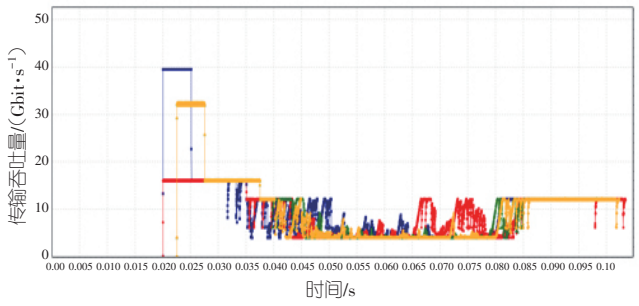
BIER技术支持根据不同应用场景划分独立的子域进行适配。例如，在MoE专家投递场景中，可为MoE专家划分一个BIER子域；在Agent场景中，可为Agent划分另一个BIER子域；对于跨数据中心（DC）的组播数据发送，则可将相关设备划分至第3个BIER子域。不同子域中的目的编号可以重复使用。BIER技术本身支持多达256个子域，能够充分满足各类场景的差异化需求[6]。

4 轻量级在网聚合

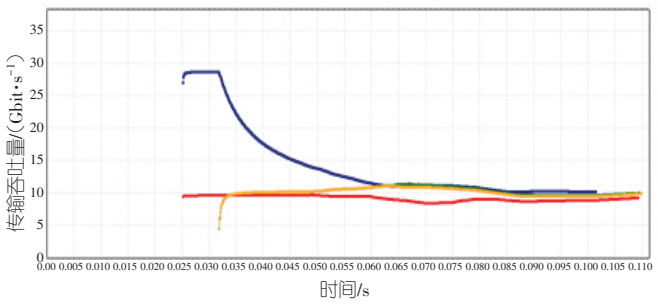
与内容分发网络（CDN）中的单向组播不同，智算网络All-to-All场景下的组播接收端往往需要向发送端返回确认消息，即组播逆向流程。同样，这种MP2P的确认消息回传也会给端侧I/O系统带来极大的性能压力，并造成带宽资源浪费。若在边缘网络节点对这些结构简单的确认消息进行轻量级聚合处理（例如引入定时限内的目的节点位图机制），则仅需将聚合后的消息发回组播发送端，从而有效避免大量确认消息经由I/O系统逐条返回端侧。

5 高性能智算网络原型测试数据及结论分析

高性能智算网络仿真系统由若干端侧服务器与7台跨区域路由器构成，路由器采用星型与网状混合组网拓扑。该仿真系统引入了端网精细化速率协商、近源端快速通告及精准流控机制。如图3所示，左侧为未引入高性能智算网络机制时的发送速率收敛曲线，锯齿效应和长收敛周期较为明显；右侧为引入本文所述部分方案后的对比曲线，锯齿效应基本消除，收敛周期显著缩短。仿真数据表明，精细化端网速率协同机制及面向广域网的精准流控机制，能够显著改善智算



（a）普通WAN流量速率收敛示意图



（b）高性能智算WAN流量速率收敛示意图

图3　高性能智算网络原型仿真测试速率收敛对比

网络场景下超大数据流量的广域网传输吞吐量，并缩短任务完成时间，从而为跨数据中心模型训练与推理的综合效率提供有力支撑。

6 结束语

高性能智算网络是支撑人工智能应用的核心网络底座，其在跨数据中心场景下的性能瓶颈将严重制约智算网络的规模化部署与拓展。本文提出了全新的架构视角，并针对关键性能瓶颈给出了端到端的解决方案：将端网协同从定性参数层面升级至定量参数层面，将拥塞处理从被动通告转变为主动规避；同时，针对广域网多业务共存的现实特征，提出了基于SRv6的精准流控与快速通告机制；在MP2P和P2MP的智算模型训练与推理场景中，分别引入无状态组播与轻量级在网聚合机制，有效缓解了网络带宽资源浪费及端侧I/O性能瓶颈问题。基于该架构的原型系统测试结果表明，端网精细化协同机制显著提升了数据传输吞吐量与整体传输效率。

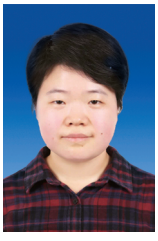
参考文献

[1] Xiong Q, Huang D, Yao K, et al. Framework for high performance wide area network (HP-WAN) draft-xhy-hpwan-framework-03 [EB/OL]. (2026-04-23)[2026-04-25]. https://datatracker.ietf.org/doc/draft-xhy-hpwan-framework/
[2] Xiong Q, Zhu X Y, Lin C W. Signaling solution for HP-WAN draft-xiong-hpwan-signaling-solution-01 [EB/OL]. (2026-03-02) [2026-04-05]. https://datatracker. ietf. org/doc/draft-xiong-hpwan-signaling-solution/
[3] Min X, Zhang K, Hu Z H. Fast congestion notification packet (CNP) with Proxydraft-xiao-rtgwg-proxy-congestion-notification-03 [EB/OL]. (2026-03-15)[2026-04-05]. https://datatracker.ietf.org/doc/draft-xiao-rtgwg-proxy-congestion-notification/
[4] Liu Y, Yao J, Lin C W, et al. Congestion control based on SRv6 pathdraft-liu-rtgwg-srv6-cc-01 [EB/OL]. (2026-02-27) [2026-04-15]. https://datatracker.ietf.org/doc/draft-liu-rtgwg-srv6-cc/
[5] Zhang H, Xu X H, Zhang Z, et al. Optimized use of BIER in AIML data centers draft-zzhang-bier-optimized-use-in-aidc-01 [EB/OL]. (2026-03-16)[2026-04-25]. https://datatracker.ietf.org/doc/draft-zzhang-bier-optimized-use-in-aidc/
[6] Zhang Z, Duan W, Xu X H, et al. Multicast use case in LLM MoE draft-zhang-rtgwg-llmmoe-multicast-02 [EB/OL]. (2026-04-13) [2026-04-26]. https://datatracker. ietf. org/doc/draft-zhang-rtgwg-llmmoe-multicast/

作　者　简　介

黄光平，中兴通讯股份有限公司有线标准总监；长期从事IP网络、算力网络、确定性网络以及智算网络等领域的技术研究和标准制定工作；发表论文10余篇，申请发明专利50余项，提交国际标准提案100余项。

肖敏，中兴通讯股份有限公司技术预研资深专家；长期从事数据通信领域的技术研究和标准预研工作；申请发明专利70余项，参与制定标准22项。



张征，中兴通讯股份有限公司技术预研资深专家；长期从事智算网络、广域网络领域的技术研究和标准制定工作；申请发明专利40余项，参与多项标准的制定。