北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

金融业 AI 大模型智算网络研究报告

北京金融科技产业联盟

2025 年 5 月

版权声明

编制委员会

主任：

聂丽琴

编委会成员：

吴仲阳　张　勇　张志鹏　李建高　成晓强

编写组成员：

陈　鹏　余学山　黄　海　张治铧　王世媛　叶　鑫　张　杰　常　东　刘　佳

编审：

黄本涛　周豫齐

参编单位：

北京金融科技产业联盟秘书处

中国工商银行股份有限公司

中国邮政储蓄银行股份有限公司

华为技术有限公司

摘　要

2023年10月，中国人民银行等六部门联合印发《算力基础设施高质量发展行动计划》，指出“算力是集信息计算力、网络运载力、数据存储力于一体的新型生产力”，针对网络运载力提出“优化算力高效运载质量、强化算力接入网络能力、提升枢纽网络传输效率、探索算力协同调度机制”的重点任务，明确通过“算力+金融”加快算力在金融领域的创新应用，为金融业务发展提供更为精准、高效的算力支持。

AI大模型智算网络技术是算力集群的重要基础底座，是新型算力中的网络运载力，是助力大模型实现跨节点分布式训练，提升大规模训练效率的重要支撑。

本文深入分析 AI 大模型技术在模型能力、结构、算力、效率等方面的技术发展趋势，提出作为底座的智算网络所面临的新问题和新挑战。围绕 AI 大模型智算网络“高性能连接、高效率传输、高可维网络、高安全保障”等关键技术进行研究，提供一套适应金融特征的覆盖数据中心、骨干及分支的 AI 智算网络技术方案。结合行业业务及技术发展方向，将金融业 AI 智算网建设演进划分为打造底座、构建系统、完善生态 3 个阶段，并给出了新技术发展及创新方向，为金融机构开展 AI 大模型智算网络规划及建设提供参考。

关键词：大模型训练、智算网络、负载均衡、流控技术、拥塞管理

目 录

一、研究背景

（一）AI 大模型发展趋势及挑战

随着新一轮科技革命和产业变革加速推进，AI 大模型浪潮席卷全球，成为最具影响力的创新科技，大模型被认为是未来人工智能领域的关键基础设施。AI 大模型正加速定义及形成新服务、新制造、新业态，成为数字时代的新质生产力。

随着技术演进，AI 大模型技术呈现以下显著发展趋势:

一是模型能力持续提升。随着深度学习技术不断发展，AI 大模型的参数规模和计算能力不断增加,使得模型能够处理更加复杂的任务和数据。2022 年发布的自然语言模型 GPT-3，能生成高质量的自然语言文本，能进行翻译、问答、文本生成等任务;在 2024 年诞生的 Sora 模型，不仅演进到视频生成的能力，还具备理解和应用现实世界物理规律的强大能力。AI 大模型逐渐从能说会道过渡到了突破多模态，形成了人机混合、自主理解、规划决策、执行复杂任务的智能体 AI Agent。

二是模型结构持续演进。稠密模型由于其结构简单及易于实现，在早期成为大模型的主流。但是随着 AI 模型规模不断扩大，计算和存储资源的需求不断增加，成为新的挑战。稀疏模型因其支持参数剪枝，在保持模型性能的同时极大降低了计算成本，因而受到更多关注并逐步成为演进方向。

三是模型算力持续增长。从 2016 年到 2024 年，GPU 单卡算力增长了 1000 倍；以英伟达 GPU 为例，其单卡算力增长速度达

到每 6 个月翻一番，超过了摩尔定律。新出现的 Super pod 超节点技术可将多个 GPU 集成在一个较大的服务器中，通过高速总线互联，实现高带宽、低延迟的数据交换和通信，以持续提升单节点算力(例如英伟达 GB200)。另一方面，大模型的参数量从 GPT-1 的 0.1B 增长到 ChatGPT 的 175B，模型所需算力在四年间也从 GPT-1 的 1PFlops 增长到 ChatGPT 的 3000+PFlops，如表 1 所示。

表1 AI大模型算力变化趋势

	GPT-1	GPT-2	GPT-3	ChatGPT
发布时间	2018年6月	2019年2月	2020年5月	2022年12月
参数量（亿）	1.17	15	1750	~1750
数据集（GB）	5	40	45,000	~80,000
训练算力（P）	1P （8张V100）	32P （256张V100）	1250P （1万张V100）	~3120P ~1万张A100
单任务训练时间	数天	数天~数周	数周~数月	数天~数周

四是模型效率持续优化。随着AI大模型的规模和复杂性增加，训练效率面临严峻挑战。业界通过并行通信算法优化、模型算法优化、混合精度训练优化等技术在训练框架层、通信算法层持续提升AI模型训练的效率。随着技术的不断进步，未来必定会有更多高效训练AI模型的方法出现。

AI大模型持续加速演进，其庞大的训练任务需要大量服务器节点通过高速网络互联组成AI算力集群协同完成。但AI算力集群并非通过简单算力堆叠即可实现完美线性扩展，而是取决于节点间网络通信及集群系统资源调度能力。网络系统的性能及可用性

成为AI算力集群的线性度和稳定性的关键，也面临新的挑战：

一是高性能传输挑战。大模型需要大量的数据进行训练和推理，千亿模型单次计算迭代内，梯度同步需要的通信量达百GB量级；MoE稀疏模型下张量并行的卡间互联流量带宽需求达到数百至上千GBps量级。服务器节点间互联网络会承载数据并行和流水线并行流量，千亿参数模型如GPT-3并行训练节点间带宽需求达到13.5GB（108Gbps），如表2所示。万亿模型参数面带宽需求增加到200Gbps至400Gbps。AI智算网络需提供更高的带宽来支持数据快速传输，并且支持算力的横向扩展能力。

表2 千亿稠密模型GPT3千卡PTD训练通信量

并行模式	通信范围	集合通信模式	一轮迭代通信数据量(计算公式)	一轮迭代通信数据量
数据并行	跨主机/跨Pod	AllReduce	$\frac{48(D-1)N_{decoder}h^2}{PTD}$	$9.5GB$
Tensor并行	主机内	AllReduce	$\frac{24(T-1)N_{decoder}Bsh}{PTD}$	$567GB$
Pipeline并行	跨主机同Rail GPU	Send/Recv	$\frac{6Bsh}{D}$	$13.5GB$
Pipeline并行优化	跨主机同Rail GPU	Send/Recv	$\frac{6Bsh}{TD}$	$1.7GB$

注：PDT，P指Pipeline 并行，D指Date数据并行，T指Tensor 并行

参数：模型 GPT3-175B，h=12288，S=2048，$N_{decoder}$=96，B=1536，D=16，T=8，P=8

二是高可用互联挑战。由于AI并行训练通信具备不规整的特征，即单流通信量大，ECMP选路不均衡，极易导致网络出现局部堵点，从而导致训练效率下降。以GPT3-175B大模型千卡并行训练为例，训练期间网络中同时存在的流数目千条以内，ECMP选路方式下，高负载链路利用率：低负载链路利用率达7:1，即流量无法有效hash，高负载链路堵点概率极大。因此对网络负载均衡

调优、无损传输等提出了更高要求。同时大模型的训练和推理也对网络的可靠性提出了更高要求，任何网络中断都可能导致训练失败或推理错误，降低集群算力的效率。

三是高可维网络挑战。大模型单次训练时间在数天-月级。训练期间如果出现网络不稳定的问题，会影响整个训练任务的进度。且大模型训练环境涉及各软硬件组件配合，运维复杂。例如Meta OPT-175B训练，故障定位平均时长约11小时，复杂应用故障定位长达80小时。因此需要一套具备精细化监控、端网一体化的，且可一键故障定界、定位及自愈的技术手段，来提升智算网络易用性。

四是高安全模型保障。在推理和训练的各个阶段，大模型都可能成为网络攻击的对象，因此需要采取额外的安全措施来保护模型不受侵害，保障数据的保密性和完整性，防止数据泄露和滥用。此外，大模型基础设施在端到端供应链的安全性、稳定性和坚韧性也存在巨大挑战，需加强AI大模型与自主可控芯片适配，建设基于自主可控人工智能芯片、训练框架、交互网络的智算中心。

（二）金融领域应用规划

2023年10月，中国人民银行等六部门联合印发《算力基础设施高质量发展行动计划》，明确算力是集信息计算力、网络运载力、数据存储力于一体的新型生产力。并提出“提升算力高效运载能力”的重点任务，要求针对智能计算、超级计算和边缘计算

等场景，开展数据处理器（DPU）、无损网络等技术升级与试点应用，实现算力中心网络高性能传输。并提出“算力+金融”赋能金融行业应用的发展计划。

中央金融工作会议强调要“做好数字金融大文章”，金融业要全面适应数字经济时代的经济社会发展变化，深化数字技术的金融应用，以大模型为重要抓手推进产业创新和解锁新质生产力。北京金融科技产业联盟依托人工智能专业委员会，加快金融业人工智能的发展和落地。2022年10月，发布《人工智能金融应用发展报告》，分析人工智能技术发展与金融创新应用情况，以加快人工智能与金融应用深度融合。2023年8月，发布《金融数据中心人工智能算力建设指引》，给出了算力与网络协同的指导意见。

二、智算网络方案综述

（一）智算网络技术需求

根据业界论文的推论，AI 大模型训练端到端理论时间计算公式如下:

$$\mathrm{E}\_t = \frac{8 \times T \times P}{N \times X}$$

其中 T 为训练数据的 token 数量，P 为模型参数量，N 为 AI 硬件卡数，X 为每块卡的有效算力，N × X 则为集群算力。

在 T 和 P 一定的情况下，提升 AI 集群算力 N × X 是降低整体时间，节省训练成本的关键。而 AI 集群算力能力很大程度上依赖于高性能和高可用的网络。在分布式计算环境中，多个计算节点需要频繁地交换数据和模型参数，这一过程的流畅与否直接关

乎集群计算效率。高性能的网络能够确保数据快速传输，减少节点间的等待时间，从而加速训练或推理过程；高可用的网络使得AI任务并行处理更加稳定高效，从而优化网络通信瓶颈。因此，高性能、高可用，且具备高效运维的网络是AI大模型训练的重要条件。

（二）业界智算网络方案

围绕着智算网络提升AI大模型训练效率，AI芯片提供商、互联网厂商、运营商网络团队及网络设备厂商，分别从不同角度进行技术探索和实现。

一是以英伟达、华为为代表的AI芯片提供商通过网络和计算联合调优，有效避免通信拥塞。英伟达在AI以太互联解决方案中，通过Spectrum交换机和BlueField网卡的协同，完成逐包均衡以缓解流量拥塞。华为提出网络级逐流负载均衡，通过网络控制器的全局视角获取全网拓扑，与端侧配合获得计算任务信息，通过对流量的主动干预、主动调度，从而达到近乎满吞吐的目标。此外，英伟达在超节点组网中引入了超高速互联通信机制。以英伟达为例，如图1所示，节点间在高速InfiniBand/RoCE连接基础上，基于NVLink形成GPU ALL-to-ALL的超高速网络，并在NVLINK网络中引入SHARP协议实现在网计算，将端侧（服务器）计算任务的部分处理操作卸载到互联网络中，由分布式交换机协同端侧应用完成集群的集合通信（Reduce、Multi-Cast等），降低网络流量负载。

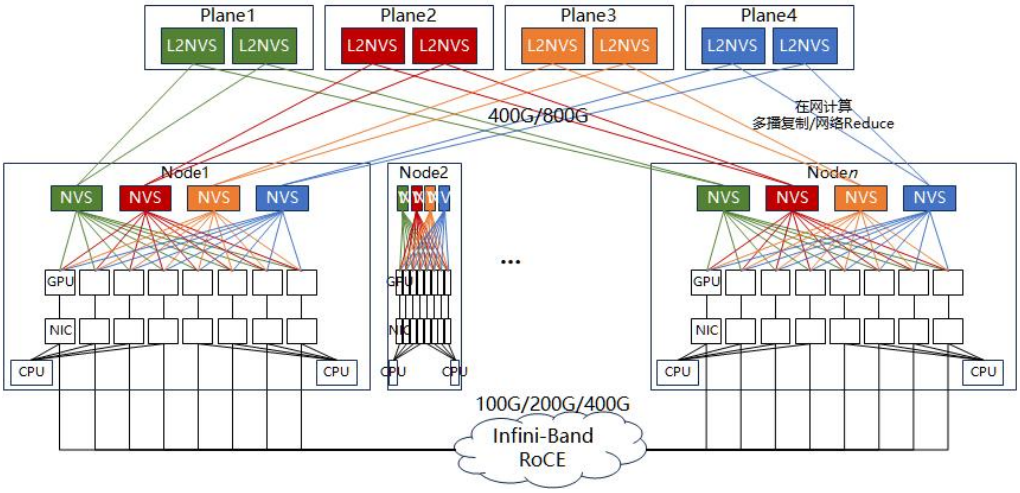

图1 英伟达NVLink超高速网络架构分析[1]

二是以Google为代表的互联网厂商，主要通过端侧技术优化来解决关键负载不均的问题，尽量减少对网络设备的能力依赖。Google提出Timely/Swift，基于端侧精准测量RTT的拥塞控制机制；Google/SRD/UEC通过逐报文对IPv6 Flowlabel/Entropy等字段修改，结合现有网络设备已具备的ECMP技术，对大规模单流进行逐报文的多路径喷洒，以提升网络流量负载。从近期发展看，该技术路线也慢慢从端侧延展到网络侧，如Google提出的CSIG，以及阿里巴巴提出的HPCC，均希望充分利用网络设备的实时测量信息，为端侧调速、选路提供更优参考。

三是以博通、中国移动为代表的网络设备厂商或运营商，主要通过推动网络侧进行方案优化。博通在DDC采用信元为粒度的网络调度方案。与以太网逐流ECMP对比，信元交换网络的负载均

[1] 来源：参考 SHARP 论文：https://ieeexplore.ieee.org/abstract/document/7830486/ Graham, R. L., Bureddy, D., Lui, P., Rosenstock, H., Shainer, G., Bloch, G., ... & Zahavi, E. (2016, November). Scalable hierarchical aggregation protocol (SHArP): A hardware architecture for efficient data reduction. In 2016 First International Workshop on Communication Optimizations in HPC (COMHPC) (pp. 1-10). IEEE.
参考 GTC 2025 官方信息公开数据，NVL 的网络拓扑推测，NV576/NV288 的 4 个子框之间电缆连接，2 级 CLOS 架构，每个 NVLink switch 是 288L@448G

衡粒度更优，但DDC依赖大缓存交换设备以及严格可靠的VoQ调度机制，给网络带来压力。中国移动提出全调度以太网，在网络中通过虚拟的报文容器机制，将流量均衡打散并利用出口设备的重排能力完成流量恢复。从近期发展看，该类技术路线有网络下延至端侧趋势，如博通发展端侧EQDS拥塞控制能力，中国移动推出端网协同负载均衡，即端和网联合参与全局流量调度。

四是以UEC、Google、AWS为代表的产业联盟及公有云厂商，持续推动对端侧及传输层协议进行优化。超以太网联盟（UEC），致力于开发物理层、链路层、传输层和软件层以太网技术以满足规模化人工智能等高性能计算需求。2023年10月，谷歌宣布开放其硬件传输协议Falcon，基于以太网基础实现高带宽、低延时、大规模工作负载的性能和效率提升。AWS推出SRD数据报文协议，即基于Nitro芯片，为实现高性能计算而开发的一种高性能、低延时的网络传输协议，以解决AWS的云性能挑战。整体上各新型网络协议总体思路类似，即在以太网完善的生态和兼容性基础上，为应对大规模高性能、低延时的计算负载诉求，优化乃至重构传输协议，例如多路径和报文散传、支持灵活传递顺序、端到端遥测等。详细对比如表3所示：

表 3 业界主流传输协议对比

对比项		TCP	RoCE	Falcon	UET	SRD
主导方		IETF	BTA	谷歌	UEC 联盟	AWS
负载均衡	包级			√	√	√
	流级	√	√			

对比项		TCP	RoCE	Falcon	UET	SRD
拥塞管理	链路层反压	基于端口反压	PFC：基队列反压		CBFC，基于信用调度	
	传输拥塞控制（检测机制）	丢包	ECN	RTT	EQDS：端侧检测 SMarTTrack:ECN+RTT+BDP+丢包	RTT
	算法	TCP 基础	DCQCN、AI-ECN、零队列拥塞管理	SWIFT	EQDS、SMarTTrack	类 BRR 算法
	实现	端侧	端侧+网络侧	端侧	EQDS：端侧 SMarTTrack：端侧+网络侧	端侧

五是以 OTT 厂商为代表打造可运维网络，减轻运维成本。OTT 厂商通过采用交换机双归方法来缓解光电端口闪断等常见故障问题，探索光模块故障快速定位定界、快速自恢复等全新方法，尝试建立有效的网络性能观测和风险预警机制。整体上，业界对网络运维能力提升对保障算力运营效率的重要性已形成共识，但目前仍然缺乏成熟有效的运维手段，常规的流量采集方案在智算场景下效果不佳。

此外，业界还在尝试创新和研究在AI智算网络中部署CLOS架构外的Dragonfly+、Torus等新型拓扑，以及多轨网络架构来满足特定大模型应用，并逐渐衍生出混合拓扑架构。此类新型拓扑易构造出非对称路径网络及拥塞，对流量均衡机制的优化要求更高，因此仍需进一步研究和验证才能使方案成熟、得到推广。

综上，针对 AI 大模型智算网场景，产业各芯片厂商、互联网公司及运营商和网络厂商，通过大带宽及网络架构优化构建高性能，基于芯片及网络机制优化构建高可用，且探索构建适用于智算场景的最优网络运维。高性能、高可用、高效运维同样也是金融行业构建智算网所必需，同时兼顾金融行业业务连续性、数据隐私保护等特征需求，AI 大模型智算网还需关注可靠性及安全性方面的能力。

三、智算网络整体架构及关键技术

金融机构普遍采用多地多中心、多分支网络互联架构。在 AI 大模型训练初期，集群规模较小，单数据中心即可集中部署训练资源池；后续随着算力规模增长，如万卡集群，可能会涉及同城多数据中心甚至跨城市数据中心的 AI 集群融合承载。另外，边缘数据中心可部署靠近用户的推理任务，以实现业务的快速决策与处理能力，提升客户体验。

金融 AI 智算网络在基础设施之上，以网络运载力支撑 AI 算力充分释放，不仅涉及数据中心网络，还涉及高吞吐的骨干网络和敏捷低时延的分支网络，AI 算力网络如图 2 所示。都需要在性能、可用性、可靠性和安全性多方面保障，以提升算力网络的智能化水平和算力能效。

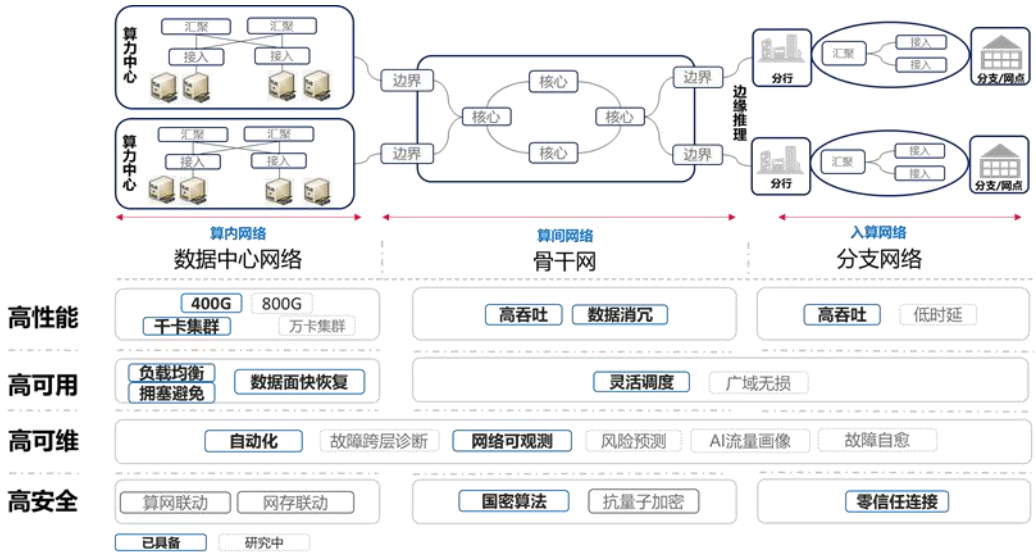

图 2 AI 智算网络架构

（一）高性能网络拓展算力规模

金融 AI 大模型具有更高的泛化能力，提升了智能化程度，也带来了模型参数量增大，数据规模增大，集群算力急剧提升的需求。网络性能 10%的提升，能够撬动整体性能、投入产出和能耗效能数倍的提升，因此高算力集群的构建，依赖高性能互联的网络。

一是使用高带宽网络设备释放算力性能。千亿参数大模型训练过程中通信占比最大达 50%，且模型越大、通信占比越高。以 GPT3.5 为例，当接入带宽提升 16 倍，通信占比从 35%降低至 3.7%，All-Reduce 训练周期缩短 14 倍。由此可见网络带宽是构建高集群算力的基础。当前业界 AI 服务器的单端口带宽已普遍具备 100G/200G 能力，未来网络设备应具备单端口 400G/800G 能力，以满足 AI 集群训练的高性能数据传输。

二是使用 CLOS 架构支撑大集群规模。大规模训练集群场景

网络通常采用 CLOS 组网架构，其优点是全互联组网支持大算力集群，网络带宽上限更高，配合负载均衡技术可使链路达到近满带宽传输数据，同时通用性和扩展性也更好。

三是使用数据消冗提升跨中心传输带宽。AI 大模型智算网范围不仅包含在数据中心内，例如生产中心和智算中心部署在不同数据中心，需要将生产数据以批量或实时方式传输到训练区域，此时会涉及跨骨干网传输，而骨干网租用运营商专线费用高昂。广域网络数据消冗技术，采用路由器设备插板方案，能有效减小跨 DC 的传输数据量，大幅减少专线租用费用。

（二）高可用网络提升算力效率

算力效率的充分发挥依赖高可用网络基础，需构建快速故障恢复能力的高可用网络，减少因网络故障中断、网络拥塞低效等问题带来的算力资源浪费，保障分布式计算任务的稳定进行。

1.高可靠传输网络

相较于传统网络，大模型训练网络对丢包中断等异常情况的容忍度更低，对故障敏感度更高，收敛时间要求更严，有更高的可靠性要求。传统网络依赖控制面协议探测协商，故障中断时可能产生百毫秒左右的短暂中断，但是这百毫秒中断若发生在数据读取或模型更新等关键阶段，系统会丢弃这批数据或在恢复后重新计算，从而浪费计算资源，延长训练时间，因此网络异常的收敛时间越短，对 AI 训练网可用性的提升越大。

使用数据面快速故障恢复技术，实现 AI 算力网故障快速恢

复。以 CLOS 架构远端设备故障场景下为例，如图 3 所示，技术整体实现包括三个步骤:

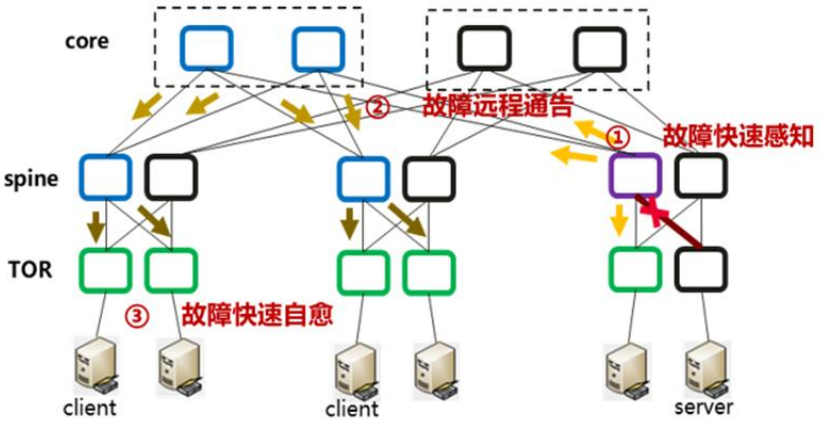

图 3 故障快速恢复技术

①故障快速感知: 转发芯片快速感知链路故障，路由选路联动故障状态，感知故障影响的业务流。

②故障远程通告: 硬件生成故障远程通告报文，携带故障路由/流，通告上游设备，解决本地设备无法保护切换。

③故障快速自愈: 远端设备基于远程故障通告，快切流量转发路径，实现业务自愈。

使用数据面快速故障恢复技术网络收敛性能，相比传统网络百毫秒的故障收敛时长，最快可提升至亚毫秒级，显著减少故障场景对训练任务的影响。

2. 高效率传输网络

在 AI 大模型训练环境中，算力服务器间需频繁通信做模型参数交换，网络传输效率优劣直接影响分布式集群训练效率。因此为了最大限度提升传输效率，AI 大模型智算网按照 1:1 无收敛网络架构设计，实现均衡无损传输，从而使整网利用率达到

100%。而在实际应用中，网络高效利用遇到两个重大难题，如图 4 所示：

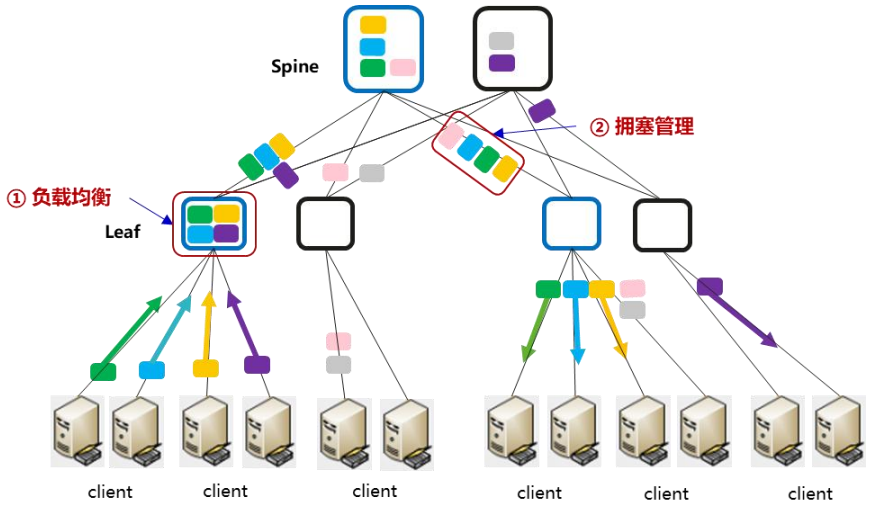

图 4 网络高效利用难题

一是流量负载不均衡。AI 大模型训练是同步模式的集群训练，即一个训练迭代周期取决于处理最慢的流。网络拥塞概率越大，通信时延越大，则 AI 大模型训练周期越长。负载均衡是避免网络拥塞的关键手段，而传统 ECMP 技术无法解决大模型训练场景诉求。有测试数据表明，即使在不产生拥塞情况下，ECMP 流级负载均衡会导致约 10% 的应用流完成时间指标是理想状态下的 1.5 倍以上，应用性能劣化明显。因此大模型业务负载不均，需要更优技术手段来解决。

二是拥塞影响范围大。高性能网络 go-back-N 的丢包重传机制，以及 PFC 队列拥塞反压机制决定了 AI 训练网络拥塞时的影响远比传统网络大。传统 TCP 网络采用丢包选择性重传及滑动窗口机制实现拥塞控制，而高性能 RoCE 网络传输层是基于无连接

UDP 实现，需要依赖上层 go-back-N 重传机制，从丢包处到最新的所有数据包进行传输，重传数据量大。据调研数据显示，当丢包率超过 10^{-5}，RoCE 网络吞吐出现急剧下降。此外，传统拥塞控制采用基于队列的 PFC 反压机制，以保证业务无损，但 PFC 是基于端口进行反压，即使能 PFC 队列的端口流量都会受影响。针对这些问题，业界均在探索有效的拥塞管理技术手段来解决。

针对流量负载不均衡问题，流级负载均衡逐步向包级负载均衡演进，细化颗粒度提升网络吞吐率。流级负载均衡通过转控分离的方式实现，训练前先基于控制面规划好流量路径，训练时根据规划好的路径进行流量转发；控制面实时感知大模型训练业务情况，自动调整、优化流量路径，相比传统 ECMP 流负载均衡技术，网络有效吞吐 40%提升。此外，随着算网协同技术完善，负载均衡技术未来将从流级进一步向包级技术演进。包级负载均衡是端侧（即服务器）将业务流量分割成多个大小相当的小包后发出，以数据包的颗粒度在网络中均衡转发，有望将网络负载提升 90%以上。转发过程如图 5 所示：

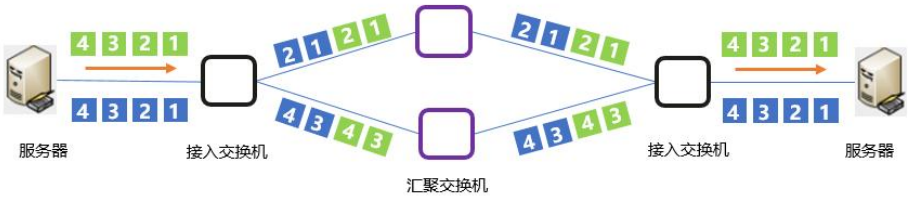

图 5 包级负载分担

值得一提的是，使用包级负载均衡技术，需要解决报文在网络中乱序的问题。当接收方接收到的报文顺序与发送方发送的报

文顺序不一致，会造成业务中断。目前解决报文乱序问题有两种方案，一种是在端侧进行报文排序，此方案对交换机的要求比较低，仅需支持报文分片和流控机制；另外一种是在网络侧进行报文排序，此方案需要交换机支持报文分片和流控，以及支持报文重组。报文重组目前受网络侧实现及应用侧限制，还未规模应用。

针对拥塞影响范围大问题，使用端网协同的拥塞控制技术，精准控制避免拥塞。在大模型智算网训练过程中，当多节点向单节点传输数据时，瞬态拥塞不可避免。针对此问题，需为训练业务流量规划优先等级队列，并使能网络流控 PFC 反压进行拥塞控制。同时，要避免 PFC 反压导致源端网络设备拥塞问题，防止相关队列中后续数据包无法转发，导致业务中断。使用端网协同的拥塞控制技术，可将网络拥塞情况向源端通告，端侧提前降速以避免拥塞发生，此技术关键是控制算法。

一是基于 DCQCN 的静态控制算法：使用 ECN 静态水线（静态配置），当大模型训练流量经过设备队列超过 ECN 水线时，即会触发拥塞通知给源端进行降速，由此进行拥塞避免。该技术是 RoCEv2 网络标准的拥塞控制技术。

二是基于 ECN 的动态控制算法：使用动态算法如 AI ECN 技术自动调整 ECN 阈值和参数，可简化控制算法部署难度。动态 ECN 技术需要作为“拥塞点”的网络设备支持，目前国内主流厂商已普遍满足。

三是基于零队列拥塞控制技术的拥塞控制算法：零队列拥塞控制技术，主动计算网络空闲带宽。各端侧发送数据窗口请求，网络设备根据端口空闲情况分配增量窗口返回给端侧，从而提高无阻塞网络的吞吐，进一步提升大模型训练效率。该技术需要交换机、网卡配合实现，目前还未规模应用。

（三）高可维网络增强算力可用性

使用 AI 大模型训练体系化网络运维架构，可有效支撑 AI 训练任务开展，运维体系如图 6 所示。AI 训练系统规模大，上下游系统庞杂，保持任务长时间不中断对于大模型训练十分重要。大模型智算网络作为算力运转的关键环节，其稳定性提升及训练性能劣化后能及时故障处置、有效自证是重要的业务诉求。因此，网络运维需与 AI 集群层协同，且网络自身亦需具备智能化的能力。

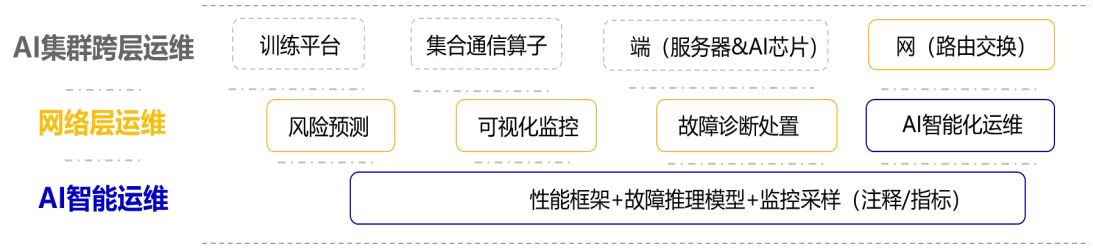

图 6 AI 大模型网络运维体系

一是网络风险预测能力：大模型智算网光纤、光模块使用量大。例如构建万卡集群训练网，需 2.5 万个光模块、1 万条光链路，管理对象繁多、故障风险高。因此网络需具备光模块训前风险预测能力，以提前排除风险隐患，提升训练系统的稳定性。

二是网络可观测能力：大模型训练流量具有突发性，易出现流量采集不准。网络需通过 Telemetry 技术获取细粒度的业务指标，包括流吞吐、丢包、PFC 反压帧、PFC 反压时长、ECN 标记、队列缓存使用率、关键告警等信息，建立有效可视的大模型智算网运维可视系统。

三是网络故障自动修复能力：大模型智算网规模大、故障排查面广，人工处理及业务恢复困难。网络硬件需具备原生自修复能力，例如光模块多 lane 自动降速，芯片故障感知、快切恢复等功能；并通过提前搭建容错服务器及网络设备，进行故障设备快速替换，实现故障快速处置和一键修复能力。

四是故障跨层诊断能力：大模型集群通信类故障主要包括训练任务无法拉起、训练任务异常中断和训练任务性能下降。训练平台、集合通信算子、端（服务器&AI 芯片）、网（路由交换）串行分析排查周期长、定位效率低。因此需构建面向训练任务的一体化故障诊断平台，可跨层获取本层故障定位及有效自证所需数据，实现快速定界定障。

五是 AI 智能运维能力：学术界提出将 AI 技术应用在网络运维新模式，即网络大模型。其核心方向是通过建立分布式网络性能框架，为每个监控对象建立丰富的注释与指标，再结合端到端的检测值，通过模型算法来推断故障组件。该技术为未来构建高精度检测分析，高效率故障闭环，以及风险预测、提前规避的运维系统建立奠定基础。

（四）高安全网络保障算力安全

金融行业智算应用关乎金融数字资产安全、生产系统安全。在大模型建设中，需网络与安全建设并举，以确保智算中心安全合规性。

AI 大模型训练应用场景日益多元化，将面临算力和数据开放，同时也面临核心资产安全风险加大，如表 4 所示：

表 4 AI 大模型智算网络安全挑战

典型场景	关键业务	安全挑战
大模型智算网接入场景	**管理接入：**本地或远程登录智算网(内网)进行开发、训练、测试。 **数据导入：**数据从外部导入，数据外发操作等。	**安全合规：**满足等保、AI 安全国标/行标规范要求。 **数据资产安全：**数据来源可追溯，数据资产不泄露。 **模型资产安全：**模型内容符合价值观要求。
算力开放出租场景	**算力共用：**算力出租方(如集团总部)提供租户级算力、存储资源，多租户(如三方、子公司)在同一平台下训练。 **数据私有：**租户从私有网络把数据导入到算力平台训练。	**算力资产安全：**算力提供者需避免算力被滥用，保障租户利益。 **数据资产安全：**租户私有数据，需保障在共用平台训练的隐私性、不滥用，不窃取。
模型及样本传输场景	**模型导入：**AI 模型训练方(如集团总部，及各租户三方子公司)，需将模型导入训练资源池，以及将训练好的模型推理到边，涉及模型传输。	**数据资产安全：**数据提供者如何控制数据资产不被滥用。 **模型资产安全：**面向模型训练方，是否能追溯模型训练过程，AI 全生命周期的责任追溯。

	样本传输：AI 大模型训练需要大量数据样本(包括集团总部及租户数据)，送到算力平台训练，涉及数据传输。	

通过对如上典型场景及业务流分析，训练场景中模型样本即数据，以及算力均属于高价值核心资产，数据被泄露、算力被盗用或破坏，将是 AI 大模型场景面临的两大关键威胁。因此，构建零信任连接、网存联动防数据泄漏、网算联动防入侵的体系化安全架构，是打造高安全 AI 大模型智算网的关键。

一是基于零信任连接，为模型拥有者及使用者提供安全接入。通过采用零信任终端接入，基于 xSEC 抗量子加密网络连接，结合零信任管理平台和态势感知平台进行威胁识别、分析、阻断，以构建零信任安全连接，保障 AI 大模型平台的用户接入安全。

二是基于网存联动数据标识及加密，防止数据泄漏及窃取。通过存储为敏感数据打标签，联动安全火墙基于标签进行安全策略控制，防止核心数据泄露；以及通过对数据进行租户级加密，保障多租户的训练数据在上传-存储-读取过程端到端防泄漏、防窃取。

三是基于网算联动检测、隔离及阻断，防止安全入侵。通过端侧(计算)内生硬件辅助检测识别恶意软件、未知威胁等，联动网络进行安全隔离、边界防护及阻断、横向扩散阻断等措施；以及通过端侧(计算)内生算力异常检测，结合外网防火墙入口流量检测，实现精准安全检测及联动处置闭环。

四、智算网络发展趋势

（一）主要发展阶段

AI 大模型智算网的建设演进需适配企业战略及业务发展，业务需求驱动技术发展，同时新技术革新反哺业务，有效驱动金融行业加速数字化、智能化转型，整体来看会经历 3 个主要发展阶段。

一是打造 AI 网络底座，支撑行业千亿模型训练，优化智能场景服务，如提升风险投顾精准度，人工客服及个性化推荐满意度等。智算网络需支持 200G/400G 速率连接百卡至千卡规模，并通过负载均衡、流量控制等技术，打造高性能、高可用的智算网络基础能力；同时围绕智算网络，构建简化运管复杂性的自动化运维网络能力。

二是构建 AI 网络系统，支撑行业万亿模型训练及推广推理应用，多元化智能场景服务，如数字柜员与无人银行，基于音/视频/文复杂业务流程一体化自助办理。智算网络需具备 400G/800G 速率实现高性能万卡连接，并基于端网协同，与 AI 计算平台联合调度，构建高速无阻塞高可用智算网络系统；同时以 AI 技术反哺网络运维，构建网络大模型以构建智能化运维网络能力。

三是完善 AI 网络生态，支撑模型能力延展，完善生态化服务场景。支撑企业 AI Agent 系统构建，从自挖掘价值业务场景，到自优化服务对象；从业务需求到网络能力，最终构建一个可自

检、自治、自愈、自闭环的 AI 网络智能体。

AI大模型智算网是金融科技数字化转型的关键技术之一，各金融机构正积极探索、试点及推动建设。目前部分大型金融机构已完成阶段一构建基础能力智算网，并论证试点阶段二能力，其他机构也在阶段一的探索和筹备构建中。

（二）新技术创新方向

从技术研究角度看，AI 大模型智算网技术，未来主要面临 3 个新技术研究及创新方向。

一是重构协议栈，优化网络能力。以 UEC 全栈协议技术为典型代表，重定义网络分层、协议能力以优化网络，聚焦提升带宽利用率，精准控制拥塞，优化反压机制。通过模型分层重构，物理层、链路层、传输层和软件层，并基于每一层围绕大规模、高性能为优化目标，构建端到端全栈增强系统。网络向下与端芯协同，向上与集合通信、AI 应用联合，纵观全产业，各技术流派方向趋同、技术不一，但最终效果如何、能否达到提升 AI 性能，同时此上下协同模式是否会因单领域限制产生木板效应，还需在后续课题继续研究。

二是突破单机卡限制，构建超万卡集群。传统的单机 8 卡配置，可满足中等规模任务训练的需求，但面对未来数万卡乃至十万卡规模训练任务，其算力和扩展性将面临挑战，产业已开始研究在 AI 集群中引入了超节点技术。超节点设计能有效整合和调度集群中的资源，突破单机 8 卡硬件限制，实现 AI 超万卡集群

的构建。该技术的成熟度、是否具备可推广使用能力，兼顾金融业务对构建超万卡 AI 集群的紧迫性和必要性，还待继续考察。

三是构建网络智能体，支撑 AI agent 基础设施生态构建。随着 AI 大模型在行业的推广，以及基于 LLM 驱动的 Agents 自动化逐步落地，金融各机构会拥有越来越多的 AI Agents 处理任务。网络也将具备 AI Agent(网络智能体)能力，支撑 AI Agent 基础设施生态构建。当前金融行业已在金融分析、金融风控、贷后处置三类场景进行 AI agent 研究创新，但各场景涌现的准度不高，网络智能体及 AI Agent 基础设施生态构建，还待继续考察。

五、案例实践

（一）工商银行 AI 大模型算网融合创新实践

随着业界算力规模持续扩大，底层网络如何支撑算力规模、算力效率、持续运行能力提升成为很大的挑战。工商银行围绕算内网络、算间网络和入算网络等 3 个方面积极开展实践。

1.在算内网络方面，工商银行选定 RoCE 作为高性能网络的技术路线并推进建设，在新型集中存储网络和 AI 算力集群等场景落地。一是基于 RoCE 高性能网络实现对 FC 存储交换网络的自主替代，率先建成支持全栈国产化的新型 RoCE-SAN 存储体系并推广应用。二是先后建设多地多中心千亿 AI 算力集群，支撑工商银行智慧金融业务创新发展。在网络技术创新方面，AI 算力集群流量模型不同于传统的联机业务流量，存在流量大流数少的

特点，传统的 ECMP 算法不适用于 AI 训练集群，极易形成局部的堵点影响 AI 整体训练效率。工商银行以技术创新赋能负载均衡能力提升，采用负载均衡优化算法，网络级联端口负载均衡差异从 5%～33%降低到 12%～16%以内，提升 AI 集合通信带宽吞吐约 24%，更好地支撑 AI 算力效率提升。

2.在算间网络方面，工商银行在全行一级骨干网部署 SRv6 网络实现广域网灵活调度基础上，率先完成广域网流量压缩技术生产落地。目前已有数据中心、开发中心、业务研发中心 3 张广域网，在核心系统异地灾备、研发测试等多个场景落地发挥作用。设计全生命周期字典压缩方案提升压缩率，提出压缩池化实现压缩持续在线等，多项技术均为首次探索。压缩流量带宽节约率达 45%，以新技术创新践行“勤俭办行”理念。

3.在入算网络方面，工商银行推广 SD-WAN 技术至多家分行网点，SD-WAN 技术的应用实现了 MSTP、MPLSVPN、4/5G 多类线路的统一管理。工商银行还自研集中运管工具实现对异构厂商控制器的屏蔽，支撑工商银行智慧网点建设。

未来，工商银行将在网络支撑计算实践基础上，进一步探索网络感知计算能力提升。

（二）邮储银行 AI 大模型算力网络创新实践

中国邮政储蓄银行在服务“三农”、小微金融、主动授信、财富管理和金融市场等领域，凭借差异化的竞争优势，为实体经济注入了源源不断的金融“活水”，致力于建设成为客户信赖、

特色鲜明、稳健安全、创新驱动、价值卓越的一流大型零售银行。以数字化转型为中心，邮储银行提出了“SPEEDS”科技战略，即智慧（Smart）、平台（Platform）、体验（Experience）、生态（Ecosystem）、数字化（Digital）和协调（synergism）。其中人工智能 AI 技术的发展和应用处于该战略的核心位置。邮储银行制定人工智能 AI 技术发展的整体蓝图，纵向分成 3 层并行建设，以数字化基础服务建设作为资源层底座，以企业 AI 技术中台建设作为平台层，在应用层不断进行场景挖掘及建设。今年年内达到百亿模型投产上线、并具备千亿模型二次训练能力，以及实现 AI 感知、洞察向创作转型升级。

大模型训练由于模型参数和数据量的规模不断增加，单机很难满足训练业务诉求，需要利用分布式并行计算将成千上万个节点高效调度起来，并通过将训练任务的数据或模型参数分片，部署到多个 NPU（神经处理单元）或其他类型的加速器进行并行计算，同时每次计算结束需要进行交叉参数协同。所以 AI 大模型训练是一个并行+串行的过程，千卡规模的算力基础设施包含了参数面、样本面、业务面、管理面和存储面网络。其中最重要的是参数面网络，因其承载着训练过程中每次迭代的参数同步和交互，这对网络的要求最为严苛，直接决定了 AI 集群的实际算力。其与传统网络的区别显而易见：高带宽、低延迟、高可靠性、扩展性、容错能力、安全性、高效的数据分发与训练调度软件协同。

邮储银行围绕大模型的需求开展了智能无损网络的探索建设，包括：

1.采用 200G ROCE 网络，匹配了昇腾的 HCCL 集合通讯库，规模上具备了万卡的扩展能力。可以在多机多卡之间建立直通无收敛组网，通过集合通讯能力支持大模型的高线性度并行训练。

2.为了实现高效的负载均衡，采用了控制器网络调优算法，实现网络动态路由，计算调度协同保障训练过程实际带宽，避免训练过程局部拥塞，确保大模型训练稳定、快速完成。

3.AI 集群是算网存高度协同的集群，其规模和复杂度增长导致故障发生概率增加，集群故障直接导致训练中断影响训练效率。邮储银行搭建了面向 AI 算力集群的运维系统，实现 AI 网络关键指标实时监控和预检查，故障时能快速定界定位，并辅助断点续训，全面提升 AI 集群的训练效率。

2024 年，中国邮政储蓄银行在人工智能（AI）领域取得了显著成果。推出了多项创新应用，如“星辰平台”“邮储大脑”“看未来”模型、RPA（机器人流程自动化）技术等，展望未来，邮储将继续深化 AI 数据中心网络的建设和应用，紧跟大模型、生成式 AI、通用人工智能技术趋势，向新技术要效益，向新要素要价值，将继续围绕 AI 基础设施建设、运维大模型和网络安全继续开展实践和探索。

术语与缩略词表

英文缩写	英文全称	中文全称
RDMA	Remote Direct Memory Access	远程直接数据存取
AI	Artificial Intelligence	人工智能
VxLAN	Virtual eXtensible Local Area Network	虚拟扩展局域网
HPC	High Performance Computing	高性能计算
DPU	Data Processing Unit	数据处理单元
NVMe	NVM Express	非易失性内存主机控制器接口规范
NVMe-oF	NVMe over Fabric	基于网络的非易失性内存主机控制器接口规范
ECN	Explicit Congestion Notification	明确的拥塞通知
PFC	Priority-Based Flow Control	基于优先级流量控制
CBFC	Credit Based Flow Control	基于优先级流量控制
ZSTD	Zstandard	无损数据压缩算法
ML	Machine Learning	机器学习
ECMP	Equal Cost Multipath	等价多路径
UCMP	Unequal Cost Multipath	非等价多路径
SHARP	Scalable Hierarchical Aggregation and Reduction Protocol	可扩展的分层聚合和归约协议
DDC	Distributed Disaggregated Chassis	分布式机框解耦
SRD	Scalable Reliable Datagram	可扩展的可靠数据报
UEC	Ultra Ethernet Consortium	超级以太联盟
VoQ	Virtual output Queueing	虚拟输出队列
BBR	Bottleneck Bandwidth and Round-trip propagation time	瓶颈带宽和往返传播时间
RUD	Reliable, Unordered Delivery	可靠无序传输技术

参考文献

[1] Susan Zhang, 等OPT: Open Pre-trained Transformer Language Models

[2] Radhika 等TIMELY: RTT-based Congestion Control for the Datacenter

[3] MetaOPT 175B 训练日志

[4] OpenAIGPT3：Language Models are Few-Shot Learners

[5] NVIDIA&Stanford University &Microsoft ResearchEfficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM

[6] Gautam 等Swift: Delay is Simple and Effective for Congestion Control in the Datacenter

[7] Yuliang 等HPCC: high precision congestion control

[8] Vladimir 等An edge-queued datagram service for all datacenter traffic

[9] 中国移动通信研究院全调度以太网技术架构白皮书

[10]中国移动通信研究院面向超万卡集群的新型智算技术白皮书（2024 年）

[11]清华大学等 2024 金融业生成式 AI 应用报告

[12]中国移动研究院 面向 AI 大模型的智算中心网络演进白皮书(2023 年)

[13] NVIDIA Spectrum-X Network Platform Architecture The First Ethernet Network Designed to Accelerate AI Workloads white paper

[14]NVIDIA,Stanford University,Microsoft Research Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM