中国移动 研究院
China Mobile CMRI

面向 AI 大模型的智算中心网络演进白皮书

（2023 年）

中国移动通信研究院

目 录

前 言

人工智能是数字经济的核心驱动力，AI 大模型是人工智能的新引擎。AI 大模型指通过在海量数据上进行预训练，能够适应多种下游任务的模型，具有强大的泛化能力、自监督学习功能和精度突破性能。其已经在自然语言处理、计算机视觉、气象预报等多个领域取得了令人瞩目的成果。大模型的发展是大势所趋，未来将会助推数字经济，为智能化升级带来新范式。

近年来，随着 ChatGPT 等生成式人工智能（AIGC）的突飞猛进，全球范围内的经济价值预计将达到数万亿美元。尤其在中国市场，生成式 AI 的应用规模有望在 2025 年突破 2000 亿元。这一巨大的潜力不仅吸引着业内领军企业竞相推出万亿、10 万亿参数量级别的大模型，而且对底层 GPU 支撑规模提出了更高的要求，达到了万卡级别。 然而，如何满足如此庞大规模的训练任务，对网络的规模、性能、可靠性和稳定性等方面提出了前所未有的挑战。以 GPT3.5 为例，其训练过程依赖于微软专门建设的 AI 超算系统，由 1 万个 V100 GPU 组成的高性能网络集群，总计算力消耗约为 3640 PF-days。在这种情况下，寻求提供极致高性能网络已成为人工智能领域的重要研究方向之一。

本白皮书将从 AI 业务发展的历程出发，深入研究大模型对网络能力的需求，分析当前网络与业务需求的差距，并探索网络技术发展趋势以弥补这一差距。我们希望，通过本白皮书的研究和分析，为未来面向 AI 大模型的智能计算中心网络发展提供有益的参考和启示。

本白皮书由中国移动研究院牵头编制，联合编制单位：华为技术有限公司、锐捷网络股份有限公司、思博伦通信科技（北京有限公司）、中兴通信股份有限公司、上海云脉芯联科技有限公司、星云智联科技有限公司、中科驭数(北京)科技有限公司、博通公司、是德科技（中国）有限公司、北京大禹智芯科技有限公司

1. AI 业务发展趋势

1.1. 人工智能技术发展趋势

人工智能（AI）是一种使计算机和机器能够表现出智能和类似人类思维的能力的技术和方法论。它通常包括学习与推理、语言和语音识别、视觉感知、自动化控制等多个领域。自从 20 世纪 50 年代，人工智能的研究开始以来，AI 已经走了一个漫长的历程，经历了许多发展与进步，也经历了漫长的寒冬。

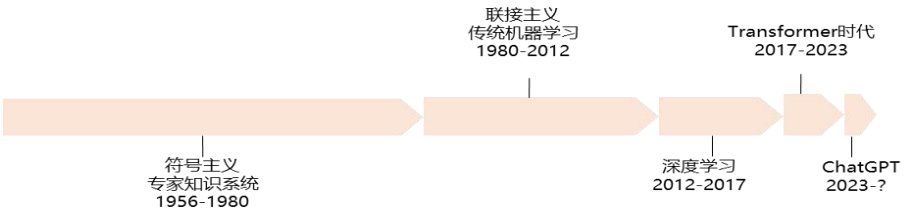

图 1-1 人工智能发展时间轴

符号主义与专家系统（1956 年-1980 年），AI 领域的创始人之一约翰·麦卡锡在 1956 年提出了“人工智能”这一术语后，AI 的符号推理阶段就正式开始了。符号推理阶段的主要发展是建立起了人工智能的推理基础。在这个阶段，人们将人类智能中的逻辑进行了形式化，创造了一种称为“推理形式”的数学表示方法。

联接主义与机器学习（1980 年-2012 年），AI 开始了一些深度学习和神经网络方面的探索，这两种技术是使用机器学习算法进行自动学习和推理的两种方式。1986 年，Rumelhart 和 McClelland 的研究进一步加强了神经网络代表的连接主义观点，这标志着 AI 进入了“连接主义时代”。

深度学习（2012 年-2017 年），20 世纪 50 年代，人们开始尝试模拟人脑的神经网络，以解决一些计算机视觉和语音识别问题。后来的时间，神经网络由于计算复杂度和可解释性等问题，经历了长时间的寒冬。直到 2012 年，Hinton 等人提出了深度学习中一种新的神经网络结构--卷积神经网络，并在 ImageNet 图像识别竞赛中获得了显著的成果。卷积神经网络具有重大的意义，推动了计算机视觉和深度学习的发展，并开拓了探索神经网络的新领域。2016 年基于深度学习的 AlphaGo 战胜围棋世界冠军，再次点燃人们对深度学习探索的热情。

Transformer 模型预训练（2017 年-2022 年），2017 年谷歌发布论文《Attention Is All You Need》，Transformer 模型引入了一种新的机制——注意力机制（Attention），用于学习不同位置的词汇之间的关联关系，从而更好地表征大型语言文本中的语义和词法关系。在

Transformer 中，可以使用多头注意力机制来学习输入序列中不同的信息，并根据这些信息进行分类、生成或其他任务。

Transformer 架构由多个堆叠的自注意力层和前馈神经网络层组成，这种设计使得它在构造大型深度神经网络时具有巨大优势。自注意力机制解决了长序列输入的信息传递问题，允许不同位置的单词或符号与其他单词或符号之间产生交互，从而更好地捕捉序列之间的依赖关系。这意味着 Transformer 可以处理极长的文本序列，而不会产生梯度消失或爆炸问题。同时，Transformer 架构具有并行计算的能力，可以同时处理输入序列的不同部分。这样可以更快地训练和推理大型深度神经网络，尤其是在使用分布式计算和 GPU 并行计算的情况下。由于它的特殊结构和设计，Transformer 架构适合构造大型神经网络，由此开启了深度学习大模型时代。大模型也被称为基础模型（Foundation Model），其通常定义为：参数规模较大（亿级）并使用 Transformer 结构，在大规模无标注语料进行自监督训练后，可以赋能一系列下游任务的模型。

BERT 和 GPT 是两种最知名的基于 Transformers 的自然语言处理模型。虽然都是基于 Transformers，但 GPT 只使用了 Transformer 的解码器部分，而 BERT 使用了双向 Transformer 的编码器部分；GPT 是从左到右建模文本，确保下一个预测是来自上下文的正确，而 BERT 是双向建模文本，不仅考虑上下文，还考虑了文本的未来信息。由于建模方式的不同使得 GPT 更适用于自然语言生成任务，如文本摘要、对话生成等，而 BERT 更适用于下游任务，如自然语言理解、文本分类、问答系统等。

2018 年 10 月，Google 团队发布了 BERT 模型。2019 年 7 月，华盛顿大学研究团队在 BERT 模型上进行了改进，提出了 RoBERTa 模型。RoBERTa 采用了更大的训练数据集和更长的训练时间，并且修改了模型的 Mask 机制，取得了更好的效果。2020 年 2 月，Google 团队提出了 ALBERT 模型，这是 BERT 模型的一个轻量级变体。ALBERT 通过参数共享和跨层参数连接的方式减少了模型大小，同时在性能上与 BERT 相当甚至略有提升。BERT 模型经过不断的改进和迭代，逐渐成为现代自然语言处理领域中的基础和标准之一。

2018 年，OpenAI 团队发布了 GPT-1，它使用了 Transformer 架构，采用了无监督学习的方法进行训练，其目标是预测下一个词语。该模型使用了 8 个 Transformer 编码器层和 12 个 Transformer 解码器层。它被证明在自然语言生成任务中比较有效。2019 年，GPT-2 被提出，相比 GPT-1，GPT-2 具有更多的参数和更高的预测能力。它使用了 48 个 Transformer 编码器层和 12 个 Transformer 解码器层，参数数量达到了 1.5 亿个。2020 年，OpenAI 团队提出了 GPT-3，它是目前最大的语言模型之一，拥有 1750 亿个参数。相比于 GPT-2，在生成

文本的质量、多样性和准确性等方面都有明显提升。GPT-3 采用的是自回归的生成方式，通过预测下一个词来生成文本。

ChatGPT 是 OpenAI 公司于 2022 年 11 月底上线的一款具有跨时代意义的大规模智能语言模型，它使用了 OpenAI 实验室开发的 GPT-3.5 和 GPT-4 系列大型语言模型，并采用了监督学习和强化学习等技术进行微调。具体而言，为了让预训练的语言模型更加智能和准确，可以在少量已标注的数据上进行调优。这种方法会使用已标注的数据训练一个有监督的策略，用于生成从给定的提示列表所需的输出。标注者们会对 SFT 模型输出结果进行打分，这样便可以创建一个由输出数据组成的打分（排序）数据集。然后，会在该数据集上进行训练一个打分模型。最后，使用近端策略优化进一步通过打分模型对语言模型进行调整，这种方法旨在提高 ChatGPT 生成输出时的准确性和自然度。ChatGPT 展现的语言能力令人印象深刻，是具有跨时代意义的语言模型。

由于大模型巨大的参数量，需要分布式计算、GPU/CPU 等异构技术及多种并行模式等方式进行训练与推理。而智算中心网络用于连接 CPU、GPU、内存等池化异构算力资源，贯穿数据计算、存储全流程，网络性能的提升对提升算力水平具有关键意义。

1.2. 人工智能业务发展趋势

中国的人工智能研究和发展方面已经取得了许多重大成就，包括在自动驾驶、机器人、语音识别和自然语言处理等领域。AI 能力已渗透多行业多环节，其中对话式 AI 产品已在部分行业进入规模化落地阶段，优化人机交互形式、流程与赋能方案，为企业 “降本增效”。ChatGPT 的出现将助力对话式 AI 进一步对产业赋能。

云服务提供商提供的三种基础服务模式为 SaaS、PaaS 及 IaaS。伴随着人工智能的发展，涌现出了人工智能即服务（AIaaS）和模型即服务 (MaaS)。

人工智能即服务，已经成为了中国 IT 行业的一个关键词。AIaaS 为企业和机构提供了一种创新的商业模式，使得他们能够更加便捷地获得人工智能服务，同时也促进了人工智能技术的进步与发展。可以提供 AIaaS 的企业数量也在持续增加，包括如华为云、百度云、阿里云和腾讯云等等。这些企业在 AI 领域投入巨大的资金和人力，为各行业提供了高品质的人工智能服务。AIaaS 的应用领域也不断拓展。AI 让传统的企业也能够通过数字化的方式创新发展。例如，金融和零售业领域的使用人工智能技术，使得他们能够更加准确的进行风险管理和发现消费者需求等。随着 AIaaS 市场的规模不断扩张,越来越多的企业也纷纷开辟了

自己的人工智能服务领域。迫切需要 AI 赋能的领域包括教育、医疗、智慧城市和智能制造等。随着互联网和人工智能技术的融合，这些领域都会产生诸多的新的商业模式，从而带动整个行业的发展。

模型即服务（MaaS）是基于云端提供预先训练好的机器学习模型，无需自己从头构建和维护模型。换句话说，MaaS 是为那些需要支持应用程序或工作流程的开发人员、数据科学家和企业提供预先构建好的模型的方式。MaaS 平台方通过大量数据 L0 层的基础大模型，再结合行业数据训练 L1 层的行业模型，用户通过 API 或模型压缩的方式获得 L2 层的垂直领域模型。

MaaS 提供商通常提供经过大量数据集训练和优化的模型，以支持特定的用例，例如图像识别、自然语言处理、预测分析和欺诈检测，这些模型用户可以通过多方式(API、在线部署)进行使用并获得推理结果。

这种方法有多个好处，包括减少开发时间和成本，以及降低那些可能没有构建自己的模型所需资源或专业知识的组织的门槛。此外，MaaS 提供商通常通过提供按需付费的定价模型，使扩展变得更加容易。一些 MaaS 提供商包括 Amazon SageMaker、Microsoft Azure Machine Learning、百度文心大模型和华为云 ModelArts 等。

1.3. 人工智能政策发展趋势

中国一直非常关注人工智能，并将其列为国家发展计划的优先领域之一。在过去一段时间，中国多部门分别发布了多项人工智能的规划性和政策性文件，目标加速人工智能在我国的发展脚步。

2017 年国家工信部颁布了《促进新一代人工智能产业发展三年行动计划（2018-2020 年）》的未来三年规划性文件，文件指出需要将人工智能和制造业深度赋能作为发展基础，将全新的人工智能技术的大规模落地作为发展重心，推动智慧工厂的发展，使我国未来制造业具备竞争力。

2017 年 7 月国务院全新公布了政策性文件《新一代人工智能发展规划》，文件中明确了构建以人工智能为主研究方向的创新机构、会逐步提升人工智能的投入、大力培养人工智能全才等指导性方案，目标加快人工智能在我国的前进脚步。

2018 年 9 月，国家科技部推出了 “新一代人工智能开放创新平台”的新一批名单，在名单上的企业被称为“人工智能国家队”，并且数量已经增至 15 家。

2020 年 8 月，国家五大部门联合发布了新一代人工智能的建设指引文件《国家新一代人工智能标准体系建设指南》，文件的目标是指定人工智能的标准，未来需要将重点工作投入在数据层面、算法层面、系统层面等，并优先将既有的成果应用于制造业、智慧交通、智慧金融、智慧安防等重点民生行业，并构建人工智能统一的评价平台。

2023 年 2 月 24 日，国家科技部官员陈家昌发表讲话，介绍国家科技部已经将人工智能视为中国的战略性新兴技术，作为经济发展的催化剂，国家各部门后续将给予人工智能发展更多政策和资金上的支持。

在刚刚 2023 年的两会报告中，ChatGPT（大模型）的人工智能词汇多次被提及，并且提出了深入产业领域的核心建议和提案，重点关注数据安全和提升产业质量。

2. AI 大模型对智算中心网络的需求

从 Transformer 问世至 2023 年 ChatGPT 爆火，人们逐渐意识到随着模型参数规模增加，模型的效果越来越好，且两者之间符合 Scaling law 规律，且当模型的参数规模超过数百亿后，AI 大模型的语言理解能力、逻辑推理能力以及问题分析能力迅速提升。同时，随着模型参数规模与性能提升后，AI 大模型训练对于网络的需求相比于传统模型也随之产生变化。

为满足大规模训练集群高效的分布式计算，AI 大模型训练流程中通常会包含数据并行、流水线并行及张量并行等多种并行计算模式，不同并行模式下均需要多个计算设备间进行集合通信操作。另外，训练过程中通常采用同步模式，需多机多卡间完成集合通信操作后才可进行训练的下一轮迭代或计算。因此，在 AI 大模型的大规模训练集群中，如何设计高效的集群组网方案，满足低时延、高吞吐的机间通信，从而降低多机多卡间数据同步的通信耗时，提升 GPU 有效计算时间占比（GPU 计算时间/整体训练时间），对于 AI 分布式训练集群的效率提升至关重要。以下将从规模、带宽、时延、稳定性及网络部署角度分析 AI 大模型对于网络的需求。

2.1. 超大规模组网需求

AI 应用计算量呈几何级数增长，算法模型向巨量化发展，人工智能模型参数在过去十年增长了十万倍，当前 AI 超大模型的参数目前已经达到了千亿~万亿的级别。训练这样的模型，毫无疑问需要超高算力。此外，超大模型对于显存的需求页很高。以 1T 参数模型为

例，使用 16bit 精度存储，首先需要消耗 2TB 的存储空间。除此之外，在训练过程中，前向计算产生的激活值、反向计算产生的梯度、参数更新需要的优化器状态等中间变量均需要存储，且中间变量在单次迭代中也会不断增加。一个使用 Adam 优化器的训练过程，峰值会产生 7 倍于模型参数量的中间变量。如此高的显存消耗，意味着需要几十上百个 GPU 才能完整存储一个模型的训练过程。

可是，仅仅有了大量 GPU，仍然无法训练出有效的大模型。合适的并行方式才是提升训练效率的关键。目前超大模型主要有三种并行方式：数据并行、流水线并行、张量并行。

在千亿~万亿级别的大模型训练时，以上三种并行都会存在。训练超大模型需要数千 GPU 组成的集群。表面上看，这和云数据中心当前已经达到数万服务器的互联规模相比，还处于下风。但实际上，几千节点的 GPU 互联，比数万服务器的互联更具有挑战，因为网络能力和计算能力需要高度匹配。云数据中心使用 CPU 计算，网络需求一般在 10Gbps~100Gbps，并且使用传统 TCP 传输层协议。但 AI 超大模型训练使用 GPU 训练，算力比 CPU 高好几个数量级，互联网络需求在 100Gbps~400Gbps，此外使用了 RDMA 协议来减少传输时延，提升网络吞吐。

具体来说，数千 GPU 的高性能组网，在网络规模上有以下问题需要考虑

- 大规模 RDMA 网络遇到的问题，例如链路头阻、PFC 死锁风暴
- 网络性能优化，包括更高效的拥塞控制、负载均衡技术
- 网卡连接性能问题，单主机受到硬件性能限制，如何构建数千 RDMA 的 QP 连接
- 网络拓扑选择，是传统 Fat Tree 结构更好，还是可以参考高性能计算的 Torus，Dragonfly 等组网

2.2. 超高带宽需求

在 AI 大模型训练场景下，机内与机外的集合通信操作将产生大量的通信数据量。从机内 GPU 通信角度看，以千亿参数规模的 AI 模型为例，模型并行产生的 AllReduce 集合通信数据量将达到百 GB 级别，因此机内 GPU 间的通信带宽及方式对于流完成时间十分重要。服务器内 GPU 应支持高速互联协议，且其进一步避免了 GPU 通信过程中依靠 CPU 内存缓存数据的多次拷贝操作。从机间 GPU 通信角度看，流水线并行、数据并行及张量并行模式需要不同的通信操作，部分集合通信数据将达到百 GB 级别，且复杂的集合通信模式将在同一时刻产生多对一与一对多的通信。因此机间 GPU 的高速互联对于网络的单端口带宽、节

点间的可用链路数量及网络总带宽提出了高要求。另外，GPU 与网卡间通常通过 PCIe 总线互联，PCIe 总线的通信带宽决定网卡单端口带宽能否完全发挥。以 PCIe3.0 总线（16lane 对应单向 16GB/秒带宽）为例，当机间通信配备 200Gbps 的单端口带宽时，机间的网络性能将无法完全被使用。

2.3. 超低时延及抖动需求

在数据通信传输过程中产生的网络时延由静态时延和动态时延两个部分构成。静态时延包含数据串行时延、设备转发时延和光电传输时延，静态时延由转发芯片的能力和传输的距离决定，当网络拓扑与通信数据量确定时，此部分时延通常为固定值，而真正对网络性能影响比较大的是动态时延。动态时延包含了交换机内部排队时延和丢包重传时延，通常由网络拥塞和丢包引起。

以 1750 亿参数规模的 GPT-3 模型训练为例，从理论估算模型分析，当动态时延从 10us 提升至 1000us 时，GPU 有效计算时间占比将降低接近 10%，当网络丢包率为千分之一时，GPU 有效计算时间占比将下降 13%，当网络丢包率达到 1%时，GPU 有效计算时间占比将低于 5%。如何降低计算通信时延、提升网络吞吐是 AI 大模型智算中心能够充分释放算力的核心问题。

除时延外，网络变化因素引入的时延抖动也对训练效率产生影响。训练过程中计算节点的集合通信过程一般可以拆解成多个节点间并行执行 P2P 通信，例如 N 个节点间 Ring AllReduce 集合通信包含 2*(N-1)次的数据通信子流程，每个子流程中所有节点均完成 P2P 通信（并行执行）才可结束这个子流程。当网络出现波动时，某两个节点间的 P2P 的流完成时间（FCT）将明显变长。因网络抖动引入的 P2P 通信时间变化可理解为木桶效率的最弱一环，将会导致其所属的子流程的完成时间也随之变长。因此，网络抖动导致集合通信的效率变低，从而影响到 AI 大模型的训练效率。

2.4. 超高稳定性需求

Transformer 诞生以后，开启了大模型快速演进的序章。过去 5 年时间，模型从 61M，增长到 540B，翻了近 1 万倍！集群算力决定了 AI 模型训练速度的快慢，单块 V100 训练 GTP-3 需要 335 年，10000 张 V100 的集群，集群系统完美线性扩展需要 12 天左右时间。

网络系统的可用性是作为基础来决定整个集群的计算稳定性。一方面，网络故障域大，

集群中一个网络节点的故障可能会影响数十个甚至更多的计算节点的连通性，降低系统算力的完整性；另一方面，网络性能波动影响大，网络作为集群共享资源相较于单个计算节点不容易被隔离，性能波动会导致所有计算资源的利用率都受影响。因此在 AI 大模型训练任务周期中，维持网络的稳定高效是极其重要的目标，对网络运维带来了新的挑战。

在训练任务期间一旦发生故障，可能需要容错替换或者弹性扩缩容的方式来处理故障节点。一旦参与计算的节点位置发生了变化，导致当前的通信模式或许就不是最优的，需要通过作业重新排布和调度，以此来提升整体训练的效率。另外，一些网络故障（例如静默丢包）的发生是不可被预期的，一旦发生不仅会导致集合通信效率降低，同时还会引发通信库超时，造成训练业务长时间卡死，很大程度上影响训练效率。因此需要通过获取细粒度的业务流吞吐、丢包等信息，可避障自愈的耗时控制在秒级别内。

2.5. 网络自动化部署需求

智能无损网络的构建往往基于 RDMA 协议及拥塞控制机制，但与之相伴随的是一系列复杂多样化的配置。其中任一个参数配置错误都可能会影响到业务的性能，还有可能会引出些许不符合预期的问题。据统计，超过 90%的高性能网络故障是由配置错误导致的问题，出现这一问题的主要原因是网卡配置参数多，其中参数量取决于架构版本、业务类型和网卡类型。由于 AI 大模型训练中集群规模大，进一步增大配置的复杂度。因此，高效或自动化部署配置能够有效的提升大模型集群系统的可靠性和效率。自动化部署配置需要能够做到多台并行部署配置的能力，自动选择拥塞控制机制相关参数以及根据网卡类型和业务类型选择相关配置。

同样的，在复杂的架构和配置条件下，在业务运行过程中可快速准确地故障定位，能够有效保障整体业务效率。自动化的故障检测一方面可以快速定界问题，精准推送问题至管理人员，另一方面可以减少问题定位成本，快速定位问题根因并给出解决方案。

3. 当前网络能力与业务需求的差异点

根据前面的分析可知，AI 大模型对网络的需求主要体现在规模、带宽、稳定性、时延/抖动以及自动化能力 5 个方面。从当前数据中心网络的实际能力来看，完全匹配 AI 大模型的需求在技术上仍然有一定的差距。

3.1. 规模差距分析

AI 大模型分布式机器学习场景的集群规模通常在 10K 级别以上，且要求在规模组网环境下实现稳定的高传输性能，相比之下，当前数据中心网络存在以下的明显不足：

(1) 网络性能需求制约着组网规模的增长

单纯从 AI 集群规模来看，10K+节点规模相对于采用数据中心多级 CLOS 组网架构完全可以胜任。但多级 CLOS 架构下避免拥塞并维持稳定的时延、抖动以及吞吐性能保障却是当前数据中心网络能力所不具备的。由于 AI 网络特有的流量模型（低熵、高带宽利用率、少数大象流、同步效应等），传统数据中心所采用的负载均衡技术（通常使用 ECMP 或者 LAG 等）以及微突发应对策略（通常采用较低的带宽利用率预留 Headroom）在该场景中的能力不足会导致 AI 业务性能受损，从而制约着 AI 集群的规模。。

(2) 网卡资源不足限制了集群规模的增长

RDMA 技术可以大幅提升通信节点之间的数据访问性能并降低 CPU 的负荷，在 AI/HPC 集群中有着广泛的应用,是面向应用开发者高性能通信库的底层支撑技术。而原生 RDMA 协议中通常采用可靠面向连接的传输方式，RDMA 网卡需要为每一个连接维护大量的协议状态，进而消耗掉大量的片上缓存。综合来看需要占用网卡缓存资源的信息主要包括：

- QP Context 上下文信息：用于缓存 QP 对应上下文信息，经验值每个 QP 需要缓存 200B 以上的内容
- 内存地址翻译表（MTT）：内部逻辑地址与主机内存物理地址的映射表
- 内存保护表（MPT）：用于本地和远端 RDMA 访问时做鉴权功能
- 拥塞控制/流控状态：每一个拥塞控制/流控组都会对应维护一组拥塞控制/流控的状态信息以及对应的限速或窗口数据，通常这些数据会随着部署规模的增加而需要更多的缓存空间，也是影响大规模 QP 部署的主要因素

由于在芯片设计时有限面积对应的 RAM 空间终究也是有限的，通常分配到如上缓存类别中，整体规模都不会太大，进而网卡的资源限制了网卡可以支持的 QP 对数量，考虑到大模型训练的集群规模，如何减少 QP 需求以及优化 QP 可支持数量是当前迫切需要解决的问题。

(3) 拥塞控制算法能力不足是限制集群规模的重要因素

根据 AI 大模型训练的组网规模需求，网络中的通信节点可达数千卡规模，且训练过程中包含多种并行模式，通信数据模型呈现多点互相通信与“大象流”的特性。而当前网络的

交换容量与缓存空间有限，易产生网络拥塞和丢包问题。当前 RoCEv2 网络中最常用的拥塞控制算法为 DCQCN 算法，该算法在在 10K+节点级的 AI 大模型网络中存在明显的性能不足问题，主要包含以下 3 点：

- 流控调参复杂度高：主流的拥塞控制算法都基于启发式算法，涉及众多的算法参数的配置和调优。不同参数的组合对特定物理网络中业务的性能影响较大。调参的复杂性在 AI 大模型网络中显得尤为突出，进而成为制约网络规模的重要因素。以典型的 DCQCN 算法为例，实际生产系统中算法参数的调整涉及 Alpha 因子更新、降速阶段、升速阶段以及拥塞通知等 15+算法参数的设置。此外网络设备侧的参数含 ECN/PFC 水线、QoS 策略等可变参数，流控调参工作的复杂性自是不言而喻。实践表明，即便在小规模 ROCE 网络中，流控调参工作往往需要专业人士持续投入数周的时间, 其高昂的精调成本和经验在 AI 大模型网络中显然不具备可复制性。
- PFC 协议有缺陷：当前几乎所有的拥塞控制算法均将 PFC 作为拥塞控制失效场景下的最后一道屏障，然而，由于 PFC 协议本身的局限性，导致依赖 PFC 协议的网络规模受限。首先，在高度冗余的网络拓扑中（如多级 CLOS 网络），传统的 PFC 协议容易出现死锁问题，可导致网络性能急速归零，而通过 Watchdog 等技术手段预防死锁也会导致协议配置的复杂化。其次，由于 PFC 协议仅支持接口队列级流控，这种粗颗粒度的流控机制极易引发头端阻塞和流间公平性问题，目前尚没有一种拥塞控制算法能完美地解决这些问题。在 AI 大模型网络中，高吞吐和低时延抖动需求的叠加要求网络最大限度避免 PFC 以及报文排队现象的发生，这对当前拥塞控制算法的能力提出了更高的要求；
- 水线调节不灵活：为了配合端侧拥塞控制算法的实施，网络设备涉及到 ECN、PFC 等协议的水线配置和灵活调整。这些水线的合理设置对于网络的整体性能影响极大，其具体的取值与业务流量模型、网络设备架构、网络拓扑、网络规模等信息息息相关。传统小规模网络中基于人工的配置方式显然不满足 AI 大模型网络规模化建设和运维的需求，需要一定的自动化水线调节甚至 AI 智能水线能力的建设和积累。

3.2. 有效带宽差距分析

在带宽需求方面，一方面 AI 大模型对网络的互联带宽有明确的要求，另一方面需要在高互联带宽的前提下保持 AI 应用通信的吞吐性能。这些需求虽然在传统数据中心中也有体

现，但在面向 AI 业务的网络中仍然呈现出不同的特征，具体分析如下：

(1) 负载均衡能力不足带来的挑战

在传统数据中心网络中，数量较多的小流使得传统基于流的负载均衡技术虽然不感知网络的实际状态，却仍然可以达到较好的负载均衡和拥塞避免的效果。而 AI 场景流量特征的巨大差异导致传统负载均衡技术失效，其本质原因是基于流的负载均衡技术并不能感知上下游网络实际的利用率和拥塞状态，引发链路极化进而导致频繁的拥塞、丢包以及时延抖动指标的劣化。有测试数据表明，在不产生拥塞的情况下，ECMP 流级负载均衡导致约有 10%的应用流完成时间指标是理想状态下的 1.5 倍以上，最坏的情况下甚至达到 2.5 倍，应用性能劣化明显。因此在面向 AI 的网络中，需要网络基于实时状态信息支持更细颗粒度的负载均衡能力。

(2) RDMA 拥塞控制算法的挑战

分布式高性能应用的特征是多对一通信的 Incast 流量模型，对于以太网的设备，Incast 流量易造成设备内部队列缓存的瞬时突发拥塞甚至丢包，带来应用时延的增加和吞吐的下降，从而损害分布式应用的性能。解决网络拥塞丢包实际上是要防止过多的数据注入到网络中造成拥塞，使设备缓存或链路容量不会过载。

DCQCN 目前是 RDMA 网络应用最广泛的拥塞控制算法，也是典型的被动拥塞控制算法。其发送端根据接收到的 ECN 标记报文，利用 AIMD 机制调整发送速率。由于 1 个比特的 ECN 信号只能定性不可定量地表示拥塞，端测需要探测式调整发送速率，导致收敛速度慢，引起网络吞吐性能下降。

(3) 超高互联带宽的挑战

AI 服务器当前采用的普遍是 PCIE4.0,、5.0，目前正在向 6.0 的规格发展。相比 PCIe 4.0 相比，PCIe 5.0 速率提升 1 倍，带宽最大支持 x16，可支撑更高性能的业务。AI 集群当前普遍采用单卡 100GE/200GE 的高性能网卡，高端网卡已经达到 400G 接口，对于网络接入层的盒式交换机，其也需要在接入端配套为 100G/200G 甚至更高速率的 400G 交换机，汇聚端需要 800G 交换机，这对交换机设备容量的需求提出了挑战。

3.3. 稳定性差距分析

当 AI 集群规模达到一定量级后，如何保障集群系统的稳定性，是除了性能外必须面对的另一个挑战。网络的稳定性一方面决定了整个集群的计算稳定性，另一方面其引发的影响

具有放大效应，根本原因在于：

- 网络故障域大：相比单点 GPU 故障只影响集群算力的千分之几，网络故障会影响数十个甚至更多 GPU 的连通性，只有网络稳定才能维持系统算力的完整性。
- 网络性能波动影响大：相比单个低性能 GPU 或服务器容易被隔离，网络作为集群共享资源，性能波动会导致所有计算资源的利用率都受影响。

对比当前数据中心在稳定性方面的能力，在如下几个方面仍然略显不足：

(1) 故障收敛时间过长导致业务性能受损

在 AI 大模型场景下，网络故障收敛时间越长，算力损失越大，性能敏感业务体验也越差。然而可靠性再高的网络仍然不可避免出现链路级以及节点级的故障，网络规模越大，出现故障的概率越大。在大规模网络环境中，网络节点和链路数量激增的同时也带来了故障事件的增加（典型云数据中心交换机的硬件故障率通常在 0.15 左右）。当链路故障发生时，传统收敛技术依赖控制面的动态路由协议的信息交互和重新选路，收敛时间较长，通常达到秒级甚至十秒级，即便采用数据面故障快速检测恢复技术（如 BFD 检测，主备路径切换），其故障收敛性能仍然在几十毫秒以上，其收敛时长均远大于 AI 高性能网络的 RTT 时延。网络故障发展成为性能损伤事件基本是必然且不可接受的。如何提升网络在故障场景中的收敛性能是当前网络亟待解决的问题之一。

(2) 缺乏高效的端网协同机制导致算侧无法快速响应网络故障

当前数据中心网络故障通常依靠网络本身的收敛能力或者运维手段实现故障隔离和恢复，对于丢包、时延不敏感的业务流量而言已经足够。在传统的无损网络中，ECN/PFC 等粗颗粒度端网协同机制也可以有效实现拥塞避免。然而，在 AI 高性能网络中，业务对丢包、时延以及抖动性能都异常敏感，如果网络侧故障不能快速准确地传递到端侧（智能网卡/DPU）并进行精准合理的源端行为控制（包括速率调节和路径控制等），拥塞导致的丢包、时延抖动以及吞吐性能下降则是必然的结果。由此可见，支持高效的端网协同机制是 AI 网络稳定性的重要一环，也是当前网络的主要能力短板。

3.4. 时延、抖动差距分析

AI 大模型应用对端到端通信时延和抖动性能提出了较高的要求，通常要求平均时延需要控制在数 us，长尾时延控制在 10us 及以下。对比当前的网络能力，存在如下差距：

(1) 网络拥塞导致的动态时延是实现低时延通信的主要障碍

典型数据中心交换机的硬件转发时延（静态时延）通常在 500ns-10us 之间，在 AI 业务节点端到端通信时延（通常都在几十甚至上百 ms）中的占比较小，而由拥塞导致的排队时延（动态时延）可以达到几十 ms 甚至亚秒级，是导致时延指标不达预期的主要原因。由前面的分析可知，当前主流的拥塞控制算法在 AI 高性能网络中均无法避免局部拥塞的问题，需要更精准、及时的拥塞控制机制实现 AI 业务低时延的基本需求。

(2) 集合通信的流同步效应导致抖动成为影响应用性能的关键因素

AI 场景中常用的集合通信具有明显的流同步效应，这种同步效应要求网络不仅要做到低时延，且时延抖动要尽可能降到最低。由于木桶效应，集合通信会放大长尾时延对应用性能的影响，因而抖动的控制相比时延的平均值而言更具挑战性。即便无拥塞丢包，不合理的负载均衡、随机的排队时延依然会让抖动指标劣化，进而导致应用性能的下降。相关测试数据表明，在 AI 场景中，对比传统基于流的负载均衡技术，逐包负载均衡带来时延抖动下降的同时，应用 JCT 指标可以获得高达 40%的性能增益。由此可见，对时延抖动的有效控制是 AI 高性能网络的重要需求，需要合理的技术手段来弥补当前网络抖动控制能力的不足。

(3) 机内和机间网络缺乏协同导致整体通信性能受限

当前机内节点间通信通常以 PCIE、NVLink、UPI、CXL 等高速互联总线技术为主，机间通信则由网卡和网络设备组成高性能网络。机内互联总线具有带宽高性能好的优势，但总体扩展能力有限且容易出现局部性能瓶颈。机间通信虽然性能方面略逊一筹，但扩展性好。当前机内网络和机间网络缺少灵活的协同机制，容易出现局部热点导致端到端通信性能受限，需要通过合理的软件、机内、机间网络的协同设计实现硬件资源的高效利用。

3.5. 自动化能力差距分析

SDN 已经诞生近 10 年时间，相关的自动化技术也相对成熟。但传统的 SDN 自动化主要是建立在通用计算网络之上，通过网络设备部署 VXLAN 特性，将业务平面与物理网络状态解耦。网络控制器在自动化部署、变更时只需要编排业务网络，映射到基础物理网络就是建立 IP 可达的隧道，自动化管理能力简单、高效。

在 AI 大模型训练场景下，当大规模 AI 网络或者对安全隔离有独特的需求时，网络建设可以引入 VXLAN 特性，传统网络控制器具备自动化编排能力。但多数情况下 AI 参数面网络是一个封闭的专用网络。基于训练效率考虑，一种典型的网络架构是 Underlay 直接承载 AI 训练任务，不再划分 Overlay 平面。同时为了充分利用设备转发能力，设备组间不再

配置 M-lag，GPU 使用单归方式接入网络。

最后，由于 AI 训练场景下，网流动辄 100G，200G 乃至 400G，传统的智能流分析技术已经无法解决 AI 训练场景下的可视化问题。隐患识别和故障预测、闭环一定程度上依赖可视化技术，因此需要新的技术解决相关问题。

4. 面对差异网络应对举措

智算中心网络作为连接 CPU、xPU、内存、存储等资源重要基础设施，贯穿数据计算、存储全流程，算力水平作为三者综合衡量指标，网络性能成为提升智算中心算力的关键要素，智算中心网络向超大规模、超高带宽，超高稳定性、超低时延、自动化等方向发展。

4.1. 大规模组网关键技术

为了支持更大规模的组网，首先需要组网设备本身硬件能力的支持，其次需要研究与 AI 大模型协同的新型拓扑优化时延和成本方案。同时在组建大规模网络过程需要强大的拥塞控制机制来解决大规模网络的拥塞问题，以便在大规模网络中有高性能网络指标。

4.1.1 网络设备硬件本身改进

网络设备应从提升自身能力出发，联合端网协同机制，为应对 AI 大模型对智算中心网络超大规模需求的挑战，在以下两方面提出改进措施：

(1) RDMA 智能网卡针对大规模 QP 部署措施优化

基于 RC 的通讯是为每一对需要通讯的 QP 建立、维护一组连接，此方式导致连接数的规模巨大，进而限制了组网规模。为减少对 QP 连接数的需求，提出以下四种优化措施。：

- 每连接多路径的能力优化。基于多个五元组的会话进行数据包的传输时，每连接多路径可将连接上的数据可以分担到多个不同的五元组。这样一方面可以提升网络的可靠性，如在数据中心 fat-tree 组网存在充分的等价路径前提下，任意一个单点故障仅影响部分路径的转发，不会导致整个连接中断。另一方面，网络均衡性将提高，使得网络的利用率得到改善，从而提高 RoCE 传输的性能。AWS 已经将多路径技术应用到其自研的协议 SRD 中，其在流量收敛性能上得到了显著的优化。
- 从 RC 模式往连接数依赖更小的模式演进。目前基于 RC 的通讯是为每一对需要通讯的

QP 建立、维护一组连接，导致了连接数的规模巨大，进而限制了组网规模。针对该问题有两种解决方案，方案一是不再提供更粗粒度的传输服务，即协议栈不提供面向连接的保序传输可靠传输能力，硬件协议栈仅负责可靠报文传递，保序等复杂的服务由驱动软件完成；方案二是优化连接的层次拆分，构建连接池，实现连接的动态共享。AWS 的 SRD 及 Mellanox 的 DC 技术分别为这两种方案的代表。

- 从 go back N 往选择性重传演进。go back N 重传是一种简单的重传方式，所以在早期芯片资源受限的情况下硬件卸载的协议栈选择实现此方式来实现重传，加上有 PFC 加持，一般来说丢包概率非常低（在 PFC 参数配置合理的情况下，一般只会在出现链路错包，链路故障的情况下才会发生丢包），芯片实现 go back N 重传不失为一种合理的选择。但随着 RoCE 组网规模不断增加，引发对 PFC 风暴整网流量骤停的担忧，同时半导体工艺的提升在帮助网卡硬件芯片中能实现更为复杂的协议，RoCE 的重传方式将会逐渐从 go back N 的全量重传演进到选择性重传。
- 可编程能力优化。目前行业内的探索方向包括可编程拥塞控制算法、可编程 DMA 能力等，主要目的是根据实际应用中业务模型实现更有针对性的拥塞控制算法，以及根据 DMA 技术的方式可以及时更新 DMA 的实现机制，能够保障在更先进的 DMA 机制或者拥塞控制机制被提出的时候，RDMA 智能网卡能够及时通过可编程能力更新对应的能力，进而提高部署规模。

(2) 芯片容量是智算中心网络规模部署的重心

25.6Tbps 容量芯片也早已大规模部署在国内外互联网或云计算数据中心。25.6Tbps 容量芯片常见的数据中心交换机形态为 200G 或者 400G。25.6Tbps 容量芯片的成熟稳定部署，使得 200G/400G 光模块放量速度加快，生态拉齐，价格已经平坦化。同时，51.2Tbps 容量的芯片已经量产并即将规模性部署，如使用 51.2Tbps 芯片，则可加倍设备 400G 接口的密度，在 16K 和 32K 典型配置下，减少设备数量，并提供未来更大规模的可能性。

越大带宽的容量，可实现 GPU 大规模模型的网络承载，并具有未来可扩展性，增强网络基础设施的先进性和寿命，投资回报率极高。更大的带宽，意味着单芯片网络设备更高的端口密度，更高的端口速率，减少网络设备数量，节省成本和功耗，当前即能实现两级 CLOS 架构 384 台交换机即可支持 32K 个 CPU 的部署。

(3) 测试仪表需具备模拟 AI 大模型业务能力

测试仪模拟大模型的业务分为两种场景，其中：

- 使用无状态流量测试仪表，在指定测试端口数量后（模拟服务器的多对多通信），提供

模拟常用高性能计算通信库的典型流量模型（比如根据 NCCL 的 broadcast，reduce，all-reduce 等典型操作）的能力。从 M:N 通信场景，流量大小，持续时间，突发设置，大流小流设置等角度，进行针对这些典型通信操作进行模拟。可以精准测试报文时延，抖动，丢包等指标。

- 使用有状态的 RoCE 测试仪表（完整实现 RDMA 协议状态），同样在指定测试端口数量后，通过定义 job 来模拟典型的通信操作，并通过多对多的通信模式进一步模拟大模型的流量。每个 job 由基本的 RDMA 操作（比如 ib read/write 等，包括 qp 数量以及消息长度大小）和“等待”，“循环”等通用动作组成。这样当定义好 job 的构成后，可以精准测试网络中的带宽占用情况，报文时延，job 的完成时延等信息。

4.1.2 新型拓扑

当前智算中心网络通常采用 CLOS 网络架构，主要关注通用性，无法满足超大规模超算场景下低时延和低成本诉求，业界针对该问题开展了多样的架构研究和新拓扑的设计。

直连拓扑在超大规模组网场景下，因为网络直径短，具备低成本、端到端通信跳数少的特点。64 口盒式交换机 Dragonfly 最大组网规模 27w 节点，4 倍于 3 级 CLOS 全盒组网。以构建 10 万个节点超大规模集群为例，传统的 CLOS 架构需要部署 4 级 CLOS 组网，端到端通信最大需要跨 7 跳交换机。使用 Dragonfly 直连拓扑组网，端到端交换机转发跳数最少减少至 3 跳，交换机台数下降 40%。同时，通过自适应路由技术实时感知网络流量负载，动态进行路由决策，充分利用网络链路带宽，提升网络整体吞吐和性能。

4.1.3 端网协同的流控改进

当前主流拥塞控制算法的优化思路仍然在端侧实现，需要至少 1 个 RTT 的响应时长，同时针对网络中存在的多拥塞点问题，仍然需要多个周期才能收敛。因此需要一种新型的端网配合的拥塞控制算法，越来越多的无损网络设计者意识到，网络遥测信息对拥塞控制算法的重要性，网络遥测可以获得精确的链路负载信息、时延信息、丢包信息、甚至缓存状态，配合网卡和拥塞控制控制算法，可以达到精确控制流量、快速收敛、充分利用空闲带宽，最终避免拥塞提高带宽利用率的效果，保障大规模分布式 AI 任务的高效完成。

4.2. 超高带宽关键技术

为了支持更大规模的组网，首先需要组网设备本身硬件能力的支持，其次在组建大规模网络过程需要强大的拥塞控制机制来保证大规模网络的拥塞问题，以便在大规模网络中有高性能网络指标。

4.2.1 网络-应用协同设计释放算力

网络带宽的增长主要依赖网卡/交换机转发芯片的发展，遵循10G->25G->100G->200G->400G->800G 的路线。近几年随着摩尔定律的逐步失效，芯片演进越来越慢，带宽提升难度也越来越大。因此，除了芯片本身的提升，可预见将来将通过网络-应用协同的方式，尽可能释放已有网络的带宽和性能。

随着聚合算力的规模不断增长、计算复杂度的增加，集合通信中数据交互的次数也会有明显的增长，随之网络通信效率对 AI 应用完成时间的制约作用也越来越明显。以目前较流行的集合通信操作 MPI ring all-reduce 为例，需要 2(N-1)次的数据交互才能完成，其中 N 为参与的节点数量。深度学习同样需要调用 AllReduce 操作进行梯度聚合，且每个节点的传输数据量是深度学习模型尺寸的 2(N-1)/N 倍。当 N 值较大时，传输量接近原始模型尺寸的 2 倍，相当于额外增添了网络带宽的负担。

近年来，随着可编程交换机的兴起和部署，可通过在网计算压缩数据流量，实现计算传输效率的提升，该方式成为一个有效提升分布式系统的方法。在集合通信原语中，Reduce 和 AllReduce 含有计算的语义，因此可以使用在网计算进行加速，减少数据交互次数和入网数据量。

组播是分布式计算系统中最常使用的通信模式之一。由于数据被重复发送，应用层组播任务完成时间大于数据量与通信带宽之比。交换机可完成组播报文的复制分发，以网络层组播替代应用层组播，避免相同数据的重复发送，实现组播任务完成时间逼近理论最优值（即数据量与带宽之比）的效果，相比于应用层组播任务完成时间减少约 50%。

4.2.2 链路负载均衡优化技术

现有基于流的负载分担技术为：网络设备接收到一条流进行转发时，此流经过 hash 计算确定一个转发路径，若不发生网络路径的变化，此流所有的报文都将持续在确定的路

径上转发。

在 AI/ML 的应用中，GPU 或其他类型的 AI/ML 计算单元之间有着非常简单的通讯关系（流的数量非常少），且由于他们有着极高的计算能力，导致一对通讯单元间的数据吞吐极高（单个流很大，所需的网络带宽极大）。这就导致在这样的应用中存在极端的负载分担不均衡，而且这种不均衡一旦引发网络丢包，就会对整体 AI/ML 的任务完成时间带来显著的负面影响。因此业界越来越重视 Spine 和 Leaf 节点之间链路的负载均衡算法优化方案，以实现流量更加均衡的哈希在多条等价路径中。

在链路负载均衡的优化算法中，已经成熟部署的案例有动态负载平衡 (DLB)。DLB 是一种质量感知负载分配的方案，它根据本地交换机的端口质量为数据包选择下一跳。且 DLB 支持 flowlet 颗粒度的调度，和基于流的负载均衡完美兼容。

近期新兴的感知路由（Cognitive routing）已经普遍被行业认为是负载均衡算法的最佳实践之一。基于感知路由的负载均衡技术实际上是一个基于全局信息的负载均衡算法。全局负载平衡通过使用在下游交换机感知到的路径质量或队列深度，来调制本地交换机的路径选择，并支持 DLB 方式动态平衡流量负载。迭代路由的 GLB 功能允许上游交换机避开下游拥塞热点并选择更好的端到端路径。GLB 保留了 DLB 的所有优质属性，例如当链路出现故障时受影响流的自动快速故障转移及非等价路径的能力。

同时，另一个路径也开始逐渐萌芽和发展——基于信元交换实现均衡负载分担。信元交换机制下，接收端设备接收到报文后，会将报文拆分成若干信元。信元会基于目的端发送的调度信令选择空闲的链路进行转发；到的目的后，信元被重新拼装成报文发出设备。在这样的机制下，不同于包转发（一个固定的流仅能利用单个路径），两个交换机之间的所有链路都可以利用，而且完全是动态的、基于微观负载实时调整的均衡利用。实际上信元交换本身并不是一项崭新的技术。在目前广泛应用的框式设备中，线卡芯片与网板芯片之间的流量交换普遍都采用了信元交换的技术，以实现机框内无阻塞交换。现在业界已经开始尝试将此技术应用到网络中，比如博通发布的 DDC 网络架构--在整个网络设备之间采用信元交换。DDC 网络架构证实了此项技术确实可以有效解决链路负载均衡的难题。将此项技术进一步扩展，应用到整个网络上，会是 AI/ML 等专有网络未来解决负载均衡问题的方向之一。

4.2.3 低功耗的 400G/800G 互联方案

随着 Serdes 技术推动数据中心进入 400G，800G 的时代，端口功耗成为了业界普遍关注的热点。低功耗的 400G/800G 互联解决方案相继推出，引起业界广泛关注，也被普遍认为是 AI 和机器学习等智算数据中心的关键技术。

(1) CPO 旨在解决下一代带宽和功率挑战

随着对网络和计算结构带宽的持续加速，需要在系统和芯片架构方面进行创新，以减缓摩尔定律的放缓。与此同时，铜互连正迅速达到其带宽距离极限。硅光子学对于维持快速数据增长和高带宽应用至关重要。共封装光学 (CPO) 是把交换机芯片 ASIC 和光 / 电引擎（光收发器）共同封装在同一基板上，光引擎尽量靠近 ASIC，以最大程度地减少高速电通道损耗和阻抗不连续性，从而可以使用速度更快、功耗更低的片外 I / O 驱动器。

通过使用 CPO 不仅可以实现联网，还可以实现 GPU 到 GPU 的互连、资源池和内存的分解。其可以满足 AI/ML 训练集群的需求，且具备高带宽和基数连接、最低的每比特成本，以及最低的电源使用效率。

(2) 线性直驱可插拔模块亦可降低功耗

在 400G、800G 时代，除了可插拔光模块和 CPO 解决方案外，在今年 3 月 OFC，Linear Direct Drive（直接驱动，也称线性驱动）可插拔 400G/800G 光模块成为了研究热点。该光模块方案最大的优势在于光模块可以省掉 DSP 芯片，极大程度降低在模块层面的信号处理的功耗和延迟。

服务于 AI 和机器学习等应用的 GPU 服务器在提供出色算力的基础上，服务器功耗也会相应的增加。400G/800G 的高速互联使得光模块以及网络设备的功耗也会相应的增长。无论 CPO 还是线性直驱可插拔模块可能都是未来智算中心的互联解决方案，通过从互连中移除所有可能的有源组件来提供最低的系统级功率。

4.3. 超高稳定性关键技术

AI 大模型下的智算中心网络作为业务流量的调度中枢，其稳定性决定着整个 AI 集群的运行效率。因此，除关注网络正常运行状态下的性能指标外，如何隔离故障域、提升故障事件的感知和恢复能力也是智算中心网络当前要解决的关键问题。

4.3.1 基于硬件的快速感知能力

AI 大模型网络通常要求网络实现亚 ms 级的故障恢复时间。故障快速感知作为故障恢复的前提，其感知性能通常在数十 us 级以下。当前大部分基于报文探测保活机制的感知技术仅能保障 50ms 级的故障感知性能。因此通过设备硬件提供更高性能的故障感知能力成为了研究重点。具体而言，硬件转发芯片可以充分利用接口物理层的统计信息（如收发光、FEC 错包统计等）提供快速的故障感知及预测的功能，实现为上层系统提供亚 ms 级故障感知基础能力的支持。

4.3.2 基于硬件的快速收敛能力

为了解决故障收敛慢的问题，一个可行的优化思路是数据面硬件卸载典型场景的故障收敛全过程，即完全由数据面感知、传递、处理故障。通过这种方式，有望将故障收敛性能提升至亚毫秒级。该技术基于转发芯片的硬件可编程能力构建，从传统的基于控制面协议软件的收敛方式演进到基于数据面硬件极速感知故障的收敛方式，并且基于数据面硬件实现远程通告和快速换路。该技术可达到亚毫秒级（<1ms）的收敛速度，将对业务性能的影响降至最低。基于硬件的故障快速收敛为高性能数据库、存储以及超算等关键应用提供了极致的高可靠性保证和稳定性体验。

4.3.3 层次化的网络故障自愈能力

在以性能为导向的大规模网络中，面向各种网络故障场景下的自愈能力是保障业务可靠性的关键。网络故障自愈能力需要在链路级、设备级以及网络级开展层次化方案的制定。其主要宗旨是最大限度降低业务性能的影响，核心技术在于提升各类网络故障事件响应的实时性。具体而言，在链路层面，通过充分挖掘网络多路径的资源价值，在最合适的节点以最快的速度实现流量转发路径的切换保护；在设备层面，通过利用节点级保护技术，实现流量的快速重路由；在网络层面，借助自动化和智能化的手段对常见的网络级故障开展根因分析和问题关联，通过快速响应预案的积累形成网络自动止血的能力，确保网络故障恢复指标在可预期的范围内。

4.4. 超低时延关键技术

为了满足 AI 大模型对超低时延的需求，智算中心网络需要从集合通讯与网络拓扑协同、硬件卸载加速技术以及静态时延优化等方面进行优化和创新。

4.4.1 集合通讯算法和网络拓扑协同

集合通信允许一组进程以定义明确、协调一致的方式交换消息和共享数据，是分布式 AI 训练系统实现数据并行、模型并行以及混合并行的的核心。如 NVIDIA 公司开源的 NCCL 可在 PCIe, NVLink，Ethernet 以及 Infiniband 网络上实现较高带宽、低延迟的 GPU 通信。

集合通信的性能和网络拓扑密切相关。NCCL 能够针对拓扑特征和 GPU 特征进行定制优化，具有比传统集合通信库 MPI 更高的性能。比如 PXN 方法将不同服务器上位于相同位置的网卡，都归属于同一 ToR switch；不同位置的网卡，归属于不同的 ToR switch。该方式下，不同 host 上相同位置的 GPU 仍然走机间网络通信，一跳可达；不同 host 上不同位置的 GPU，则先通过机内网络转发到对应位置的 GPU 代理上，然后通过该 GPU 代理走机间网络来完成通信。该方法可以有效地减少跨 host 集合通信过程的网络跳数，从而提升整网性能。

4.4.2 DPU 硬件卸载

在当前 GPU 的算力能力下，100Gbps 或更大的数据量才能够充分发挥单个 GPU 的算力。在这样的发展趋势下，基于 RDMA 协议的 GPUDirect RDMA 技术，在 DPU 与 GPU 通信的过程中可绕过主机内存，直接实现对 GPU 内存的读写能力。而且，DPU 上全硬件实现的 RDMA 能够支持单流百 G 以上的数据收发能力，进而实现了 GPU 算力聚合且最大化提升了 GPU 集群算力。GPUDirect RDMA 技术已经成为当前算力资源总线级互联高性能网络的主流技术。

另一个 GPU Direct 技术是 GPU Direct Storage，简称 GDS。GDS 是为了解决 GPU 从 Storage 获取数据的延时和效率问题，可实现 GPU 到 Storage 的直接访问。在 GPU 使用 GDS 访问远端存储时，通过网卡实现 NVMe-oF 的卸载和加速就变得异常重要。NVMe-oF 的实现是在标准的 NVMe 操作上进行了网络部分的封装。NVMe 实现的各种队列操作，包括 MQ，SQ，CQ 等，可以清晰的通过硬件逻辑实现。同时，根据 NVMe-oF 所支持的网络协议，进一步判断哪种协议适合硬件卸载，或者其使用硬件卸载方式付出的代价最小。

利用网卡对 RDMA 的支持，再叠加 NVMe 的实现逻辑，网卡可以完整实现 NVMe over RDMA 的硬件卸载，进而为 GDS 提供 NVMe-oF 卸载及加速方案。

4.4.3 静态转发时延优化

静态转发时延主要是由转发芯片引入的。转发芯片主要有 PHY/MAC 模块、包处理（PP）模块和缓存管理（BM）模块组成，可针对不同模块分别进行时延优化处理。

- PHY/MAC 模块：为了支持更广泛的场景应用，在保证接口可靠性的同时追求更低的时延，新的接口形态和编码算法有待进一步探索。
- 包处理（PP）模块：为了降低包处理模块的时延，可通过简化业务部署的方式，如关闭报文转发路径上不需要的子模块、关闭下行 ACL 功能（设备上未部署下行 ACL 时）及不建议部署 VxLAN 业务等方式。同时，包处理模块内存在较多的查表（MAC 表/FIB 表）过程，主要表项因为容量较大普遍采用算法查找，其查表深度也会影响转发时延。为了追求更低的时延，需要探索更好的并行查表设计及高效的查表算法。
- 缓存管理（BM）模块：为进一步降低缓存管理静态时延，需要优化芯片内缓存布局和总线设计。随着应用流量模型的变化和链路利用率的提升，影响时延的主要因素不再是静态时延，而是拥塞带来的动态时延。动态时延的控制依赖精细的缓存管理，包括各种拥塞通知门限和反压门限的适应和调整，以及端网协同等相关技术。

4.5 自动化关键技术

面对 AI 大模型场景下网络的特殊性，AI 网络需要实现多维度自动化能力的支持，包含以下四方面。

(1) 端到端部署自动化能力是 AI 集群扩展性的前提

AI 大模型网络典型的特征是规模较大，且必须支撑业务集群的按需扩容。然而，网络中涉及拥塞控制算法、RDMA 无损等复杂特性的配置，且配置工作涵盖网卡和网络交换机。面对 AI 网络特殊性和复杂性，通过充分识别并分析 AI 场景网络特征及变更特点，从而设计符合 AI 场景的网络模型，支撑自动化能力，尽力实现“即插即用式开局”。

(2) 测试验收自动化能力是 AI 集群高品质交付的基础

在网络与端侧的部署工作完成后，需要结合场景针对配置一致性、可靠性、业务性能等开展一系列自动化测试和验收的活动。通过自动化测试建立验收基准，而非依靠人工经

验，是确保 AI 集群高品质交付的基础。

(3) 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，利用转发芯片的原生能力，对网络的状态、数据进行高性能可视化监控。例如通过呈现网络的拥塞状态、负载不均状态等，为自动化调度调优提供数据支持，可实现端到端可视化、自动化运维等，实现故障的快速定位和一键修复的能力。

(4) 变更自动化是网络能力自演进的基本保障

在 AI 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。

5. 总结和展望

随着 ChatGPT、Copilot、文心一言等大模型应用的横空出世，AI 大模型下的智算中心网络也将带来全新的升级。

本白皮书从 AI 大模型发展情况、AI 大模型下智算中心网络的需求、当前技术与需求的差距及技术演进四个方面，开展了相关研究，以期抛砖引玉，更盼得到更多同行的参与和讨论。中国移动也希望按照高价值优先、先易后难的原则，逐步推动 AI 大模型下的智算中心网络关键技术的成熟与落地。我们期盼与众多合作伙伴一起，汇聚行业力量，共同打造大规模、高带宽、高性能、低时延以及智能化的 AI 大模型智算中心网络。

术语定义

词语	解释
数据并行（Data Parallelism）	通过将训练样本集拆分成多个mini-batch，在多GPU上训练。每个GPU根据自己mini-batch得到模型梯度，然后多个GPU将各自得到的梯度进行平均，再进行参数更新，开始下一轮迭代训练
流水线并行（Pipeline Parallelism）	将大模型按照层为单位，切分到多个设备。模型层之间有依赖关系，负责第k层的GPU需要在负责第k-1层的GPU完成计算后，传递相关参数。为了避免这种依赖关系导致的GPU等待，mini-batch进一步被分割成micro-batch，然后多个micro-batch按照流水线的模式依次计算。这样，后一个GPU处理第n个micro-batch时，前一个GPU可以开始计算第n+1个micro-batch，实现流水线计算，提升系统效率
张量并行（Tensor Parallelism）	将大模型每一层进一步切分，从而减少存储一层带来的显存压力。张量并行可以理解成矩阵乘法拆分成分块乘法，在多个GPU分块完成任务后，再进行AllReduce/AllGather集合通信模式进行结果汇总。张量并行对通信时延和带宽要求都极高

缩略词表

英文缩写	英文全称	中文全称
ACL	Access Control List	访问控制列表
AI	Artificial Intelligence	人工智能
AIGC	AI Generated Content	人工智能生成内容
AIMD	additive-increase/multiplicative-decrease	线性增速乘性降速
Adam	Adaptive Moment Estimation	自适应矩估计
API	Application Programming Interface	应用程序接口
AIaaS	AI as a Service	人工智能即服务
ASIC	Application Specific Integrated Circuit	专用集成电路
BERT	Bidirectional Encoder Representations from Transformers	一个预训练的语言表征模型
BFD	Bidirectional Forwarding Detection	双向转发检测
BM	Buffer Memory	缓存管理
CPU	Central Processing Unit	中央处理器
CPO	Co-packaged Optics	共封装光学
CQ	Completion Queue	完成队列
CXL	Compute Express Link	开放性互联协议
DCQCN	Data Center Quantized Congestion Notification	拥塞控制算法
DDC	Distributed Disaggregated Chassis	分布式分散式机箱
DPU	Data Processing Unit	数据处理器
DMA	Direct Memory Access	直接内存访问

DLB	Dynamic Load Balance	动态负载平衡
ECMP	Equal Cost Multipath	等价多路径
ECN	Explicit Congestion Notification	明确的拥塞通知
FCT	Flow Completed Time	流完成时间
FIB	Forwarding Information Base	转发表
FEC	Forward Error Correction	前向纠错
GB	Gigabyte	吉字节
GLB	Global Load Balance	全局负载平衡
GPU	Graphics Processing Unit	图形处理器
GPT	Generative Pre-trained Transformer	生成预训练变压器
IaaS	Infrastructure as a Service	基础设施即服务
IP	Internet Protocol Address	互联网协议地址
JCT	job completion time	任务完成时间
LAG	Link Aggregation Group	链路汇聚组
MaaS	Model as a Service	模型即服务
MPI	Message Passing Interface	传统集合通信库
MQ，SQ，CQ 队列操作	Message Queue	消息队列
MAC	Media Access Control Address	媒体存取控制位址
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
NCCL	NVIDIA Collective Communication Library	NVIDIA 聚合通信库
PF-day	Petaflop/s-day	一天进行约 10 的 20 次方运算
P2P	point to point	点对点
PaaS	Platform as a Service	平台即服务
PFC	Priority-Based Flow Control	基于优先级流量控制
PCIe	Peripheral Component Interconnect Express	高速串行计算机扩展总线标准
PP	packet processor	包处理
PHY	Physical	端口物理层
QoS	Quality of Service	服务质量
QP	Queue Pair	队列对
RAM	Random Access Memory	随机存取存储器
RC	reliable connection	可靠连接
RDMA	Remote Direct Memory Access	远程直接数据存取
RoCE	RDMA over Converged Ethernet)	远程内存直接访问协议
RTT	Round-Trip Time	往返时延
SFT	supervised fine-tuning	生成模型 GPT 的有监督精调
SaaS	Software as a Service	软件即服务
SDN	Software-defined Networking	软件定义网络
SQ	Submission Queue	提交队列

SRD	Scalable Reliable Datagram	可扩展的可靠性数据报
ToR	Top of Rack	接入交换机
TCP	Transmission Control Protocol	传输控制协议
TB	Terabyte	太字节
UPI	Quick Path Interconnect	快速通道互联
VxLAN	Virtual eXtensible Local Area Network	虚拟扩展局域网