ICS 33.040.40

CCS I631

T/NIDA

全 球 固 定 网 络 创 新 联 盟

T/NIDA 001-2024

智算数据中心网络建设技术要求

Technology Requirements for Network Construction of Intelligent Computing Data Center

2024-12-25 发布 2024-12-25 施行

全球固定网络创新联盟（NIDA）发布

目　次

前　言

本文件按照 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利权和著作权。本文件的发布机构不承担识别专利和著作权的责任。全球固定网络创新联盟不对标准涉及专利的真实性、有效性和范围持有任何立场；不涉足评估专利对标准的相关性或必要性；不参与解决有关标准中所涉及专利的使用许可纠纷等。

本文件由全球固定网络创新联盟技术委员会提出并归口。

本文件起草单位：中国信息通信研究院、中国工商银行数据中心、平安科技(深圳)有限公司、科大讯飞股份有限公司、中国电信研究院、中兴通讯股份有限公司、珠海星云智联科技有限公司、苏州盛科通信股份有限公司、深圳花儿数据技术有限公司、华为技术有限公司

本文件主要起草人：郭亮、王少鹏、余学山、杨飘飘、蒙祖瑞、鲍中帅、陈映、邵会勇、马国强、王俊杰、郝斌、张力、潘洋

智算数据中心网络建设技术要求

1 范围

本文件规定了全球固定网络创新联盟中智算数据中心网络建设技术要求，包括智算数据中心的参数面网络、样本面网络、业务面网络及管理面网络建设场景。

本文件适用于智算数据中心网络建设，主要应用于指导智算数据中心网络规划、设计和验收。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

3 术语和定义

3.1 智算数据中心集群算力指标定义

3.1.1

集群总算力 Cluster Total Computing Power

本文中使用“集群总算力”泛指智算集群所有计算节点的计算能力总和。智算集群是由多个互联的计算节点组成的系统，这些节点可以协同工作，共同解决计算问题。总算力是衡量集群处理大规模计算任务能力的重要指标。

3.1.2

有效算力率 Effective Computational Power Rate

本文中使用“有效算力率”泛指智算集群中衡量数据中心或计算平台实际计算能力利用效率的指标。它通常指的是实际可用于计算任务的算力与总可用算力的比值。

3.1.3

算力可用率 Computing Power Availability Rate

本文中使用“算力可用率”泛指在训练任务过程中，计算资源实际可用于计算任务的时间与总时间的比率。它反映了计算资源的可用性，是衡量数据中心或计算平台资源利用率的一个重要指标。

3.1.4

集合通信效率 Collective Communication Efficiency

本文中使用“集合通信效率”泛指在并行计算环境中，多个计算节点（或进程）共同参与的通信操作的效率。这些通信操作包括Broadcast、Gather、All-Gather、Scatter、Reduce、All-Reduce和All-to-All等。集合通信效率直接影响到并行计算任务的性能。

3.2 智算数据中心网络术语定义

3.2.1

网络通信效率 Network Communication Efficiency

本文中使用“网络通信效率”泛指智算数据中心网络环境中数据传输的效能，通常用来衡量数据在网络中传输时的速率、准确性和资源利用情况，包含吞吐率、丢包率、时延、带宽利用率等多个方面。

3.2.2

吞吐量 Throughput

网络在单位时间内成功传输的数据量，通常以每秒比特数（bps）来衡量。

3.2.3

时延 Latency

本文中时延指针对不同计算任务网络需要保障的端到端时延。

3.2.4

丢包率 Packet Loss Ratio

未发送成功报文个数占总报文个数的比例。

3.2.5

带宽利用率 Bandwidth Utilization

网络带宽被有效使用的比例，高带宽利用率意味着网络资源得到了充分的利用。

3.2.6

Dragonfly网络拓扑 Dragonfly Network Topology

是一种专为高性能计算设计的网络拓扑，其具有分层设计、全互联、优化路由、高可扩展性的特点，设计目标是在保持低延迟的同时，实现大规模网络的高吞吐量和负载均衡，这使得它成为智算网络架构中一个重要的技术选择。

4 缩略语

下列缩略语适用于本文件。

IP：互联网协议（Internet Protocol）

CLOS：CLOS网络拓扑结构（CLOS）

Scaling law：扩展定律

HPC：高性能计算（High-Performance Computing）

AI：人工智能（Artificial Intelligence）

PFC：基于优先级的流量控制（Priority-based Flow Control）

ECN：显式拥塞通告（Explicit Congestion Notification）

RDMA：远程直接存储器访问（Remote Direct Memory Access）

5 智算数据中心网络建设概述

5.1 AI 业务发展趋势

随着ChatGPT引爆大模型热潮，行业进入了生成式人工智能时代，各界对人工智能加快社会各领域数字化转型及智能化发展，促进全社会生产效率提升，抱有极高的期望。算力既是智能时代的“引擎”，也是智算时代最宝贵的资源。例如金融行业，已经全面推进AI大模型场景，加快新金融战略布局，将大模型广泛应用于智能客服、智能营销、智能运营、智能风控和智慧综合办公等多场景，大幅提升客服、运营效率，提升风控精准度，缩短研发周期，增强用户使用体验。

生成式人工智能训练的第一性原则是扩展定律（Scaling law），即大模型的智能水平与模型参数、数据样本和算力三个因素成正比。

— 算法：迈入万亿参数大模型时代，开启通用人工智能的大门

过去6年里，AI大语言模型参数量从Transformer的6500万，增长到GPT4的1.8万亿，模型规模增长超2万倍，如图1所示。

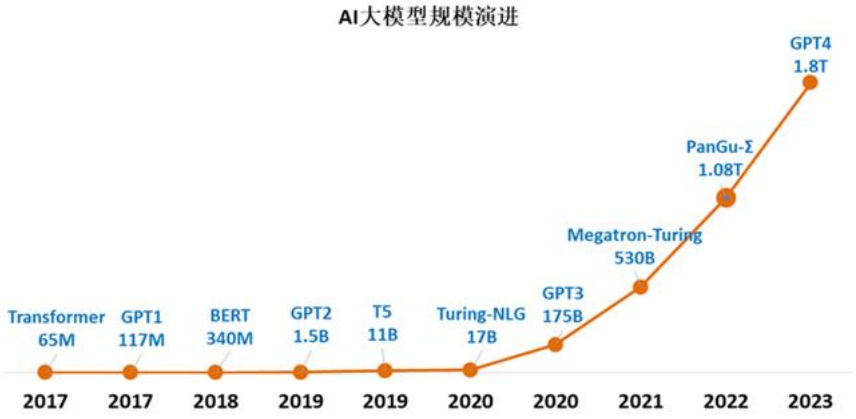

图 1 大语言模型参数规模演进

根据AI大模型的Scaling law，增大模型的参数规模、训练数据集，投入更多的算力，就能持续提升大模型性能。正是扩展定律和涌现能力，驱动着AI大模型规模的持续增大。随着GPT4、盘古等万亿模型的发布，也正式标志着，AI大模型进入了万亿模型时代。目前文本、音频、图像等单模态大语言模型已经相对成熟，大模型正加速朝着多模态模型的方向发展。从Vision Transformer的提出，再到GPT4的图文处理能力，多模态模型取得了明显的进步。

— 算力：单卡算力2-3年翻倍，算力集群规模从千卡走向万卡

AI模型参数量的持续增大带来算力需求的指数级增长，2012年至2019年AI训练算力平均每100天翻倍。而GPU的单卡算力则需要2-3年增长一倍，由此可见，单卡算力的发展速度远远落后于模型发展的算力需求。

近年来，随着各行各业都投入到AI大模型的研发中来，AI智算的算力规模增长迅猛。据《中国算力发展报告（2024）》披露，到2023年底中国的智能算力规模达到246EFLOPS；据IDC（International Data Corporation）预测，预计到2026年，智能算力规模将进入ZFLOPS级别，达到1,271.4EFLOPS。为了满足高速增长的算力需求，AI大模型厂商都在加速建设大规模的GPU卡集群。

— 数据：数据需求持续增长，对高质量数据需求迫切

随着AI模型规模的持续增长，对数据集质量也提出了更高的要求，数据不仅要多，而且要质量高。研究表明，在低质量数据集的预训练，比如噪声数据、有毒数据、重复数据，会损坏模型的性能。Meta的研究表明，更高质量的数据，比如高质量人工标注数据，可以弥补模型规模的差距。麻省理工大学等研究机构的研究显示，高质量的语言数据将在2026年耗尽，低质量的语言数据将在2030~2050年间枯竭。AI大模型马上就将面临训练样本不足的挑战，人类需加强高质量的数据处理、标注，建立完善的数据收集和评估体系，以更高质量的数据推动AI大模型性能的进一步提升。

可见，随着模型参数量向万亿、十万亿增长，单卡算力增长无法满足模型发展的算力需求，模型训练使用的算力卡数量也向千卡、万卡、十万卡发展。为了实现数据中心内算卡的互联，并高效利用算力资源，智算数据中心的网络需要具备超大规模组网、无损高吞吐，以及智能容错能力，进而实现算力效率的极致释放。

5.2 智算数据中心网络建设挑战

随着AI模型参数的规模越来越大，从千亿增长到万亿、十万亿级，客户将面临数亿级美元的投入，以及长达数周乃至数月的训练周期，训练难度和成本与日俱增。集群总算力取决于单芯片算力、集群规模、有效算力率以及算力可用率。而网络作为计算集群重要的组成部分，在提升集群总算力方面也面临着巨大的挑战。

- 挑战一：算力需求倍增，催生超大规模算力集群:大模型的智能涌现能力需要在模型的参数规模足够大，训练数据量足够多，并且能够不断迭代的情况下出现。根据OpenAI的测算，自2012年以来，全球头部AI模型训练参数规模每年增长30%，算力需求每3-4个月翻一番，每年增长幅度高达10倍。随着模型参数量向万亿、十万亿增长，模型训练使用的算力卡数量也向万卡、十万卡发展。同时AI大模型并行计算模式，也带来了更大的通信量。以GTP-3为例，在每轮迭代中，如果使用数据并行方式，通信量可达到9.5GB/iter；如果使用流水线并行方式，通信量可达到13.5GB/iter；而使用张量并行方式，通信量可达到567GB/iter。智算数据中心的算力需求倍增，需要更大规模网络支撑，这对数据中心网络建设的网络容量、组网架构、建设成本、维护成本都提出了新的挑战。
- 挑战二：AI大模型训练成本高，智算网络要求提升通信效率:AI大模型训练需要海量算力的支撑，需要由大量的服务器作为节点，通过高速网络组成集群，服务器之间互联互通，相互协作完成任务，大规模、长时间的 GPU 集群训练任务，仅仅是单次计算迭代内梯度同步需要的通信量就达到了百 GB 量级，此外还有各种并行模式、加速框架引入的通信需求。基于典型模型建模发现，类似GPT3的千亿参数模型，通信的端到端耗时占比达到20%，针对某个万亿参数MoE（Mixture of Experts）模型建模发现，通信的端到端耗时占比则急剧上升到约50%。由此可见，集群规模越大，通信量和复杂度也会越大。要想通过集群发挥出更强的算力，计算节点需协同工作并共享计算结果，需要对服务器之间的通信、拓扑、模型并行、流水并行等底层问题进行整体优化。这其中大带宽、高吞吐、无损的高性能网络至关重要。
- 挑战三：AI大模型训练周期长，智算网络要求提升可靠性:算力可用率是衡量计算资源使用效率的重要指标之一，它直接关系到计算任务的执行效率和计算平台的运营成本。随着超大规模智算网络建设，网络中的故障点倍增，故障更频繁，故障有可能会触发训练任务重调度，影响集群的可用率，增加了大模型训练的成本。这对增强智算网络可靠性部署，提升算前故障检测准确度，增强智算网络故障检测、定界和可视化能力也提出了挑战。

因此智算数据中心建网需要从网络规模、通信效率、可靠性、可维护性等方面来保障和提升集群算力规模、有效算力率和算力可用率。

6 智算数据中心网络架构

6.1 逻辑架构

人工智能计算中心提供训练算力和推理算力。人工智能计算中心多机柜联网形成AI集群。按照服务类型及安全等级，把整个网络分成不同的业务区块：接入区、管理区、业务区。各区块间通过核心交换机连接在一起，不同类型的流量，可根据数据中心的实际情况进行隔离和保护。

通用人工智能计算中心总体逻辑网络拓扑如图2所示，包括如下不同的业务区块：

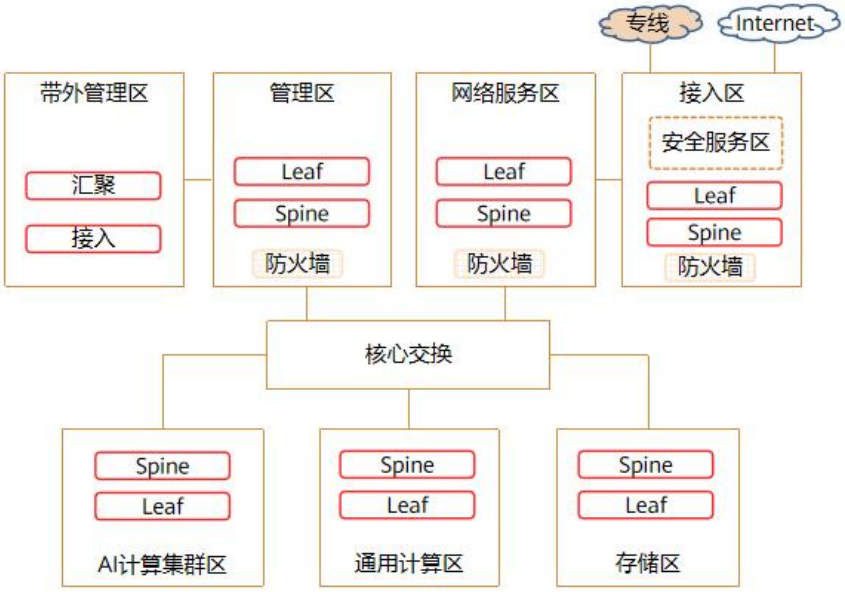

图 2 人工智能组网逻辑架构

- 接入区：Internet和专线网络接入，部署数据中心的外网接入设备。
- 安全服务区：提供DDoS、入侵检测等安全防护能力。
- 网络服务区：提供网络基础服务，例如vRouter、vLB、vFW等。
- 管理区：部署平台的服务管理系统及运维管理支持组件，用于维护管理数据中心的AI服务器、通用服务器、存储设备和网络交换机设备等。根据客户部署实践，部分运维管理组件也可以部署在带外管理区。
- 带外管理区：主要连接网络区设备管理口以及服务器BMC口，为物理设备提供带外管理网络。该网络除物理设备管理流量外不承载其他业务流量。
- AI计算集群区：AI服务器宜集成NPU,CPU,DPU，实现一体架构的AI计算节点，其中DPU要求支持带有RDMA网络卸载加速的能力，为AI计算集群之间的集合通信提供高性能无损AI计算集群网络，实现AI高性能计算。
- 通用计算区：提供AI训练相关的通用计算资源，例如部署深度学习平台等软件。
- 存储区：高速大带宽互联的存储系统，AI场景下主要用于训练数据和训练模型的存储，存储节点可通过支持存储协议加速和加解密能力的DPU网卡进行赋能，以满足AI训练场景中对于训练样本与模型的高性能安全传输。

6.2 物理架构

为应对挑战，智算数据中心网络架构进行了优化，如图 3 所示，划分为参数面、样本面、业务面，及管理面四个网络平面：

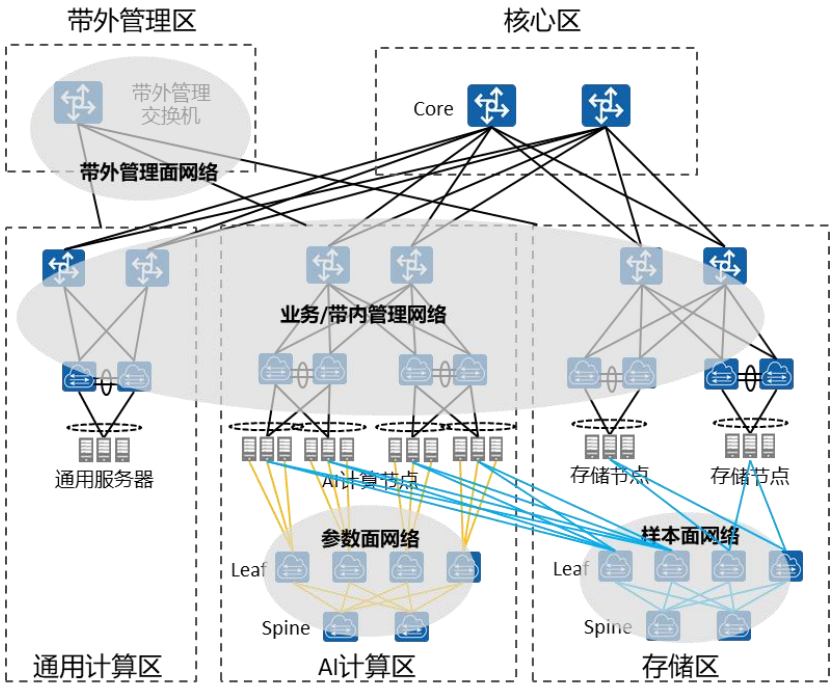

图 3 人工智能组网物理架构

- 参数面网络：承担模型训练过程中AI计算节点之间参数交换的流量，是影响智算集群算力的关键网络要素，要求部署1:1收敛比的高带宽、大规模智能无损以太网络，网络架构可选CLOS二层、CLOS三层组网架构，在考虑支撑大规模组网、减少网络层次以节省设备开销、避免流量绕行减少路由复杂度等因素影响时，也可选择DragonFly+、Group-wise DragonFly+等二层扁平化组网架构。参数面网络要求达成网络规模、网络高吞吐、高可靠性、智能运维等的关键技术能力要求。
- 样本面网络：承担模型训练过程中AI计算节点访问存储系统的流量，如样本数据的读取、Checkpoint的读写等，随着模型扩展，样本面网络要求支持大带宽、低时延、智能无损以太网络RoCE。样本面网络通常采用两层CLOS组网，接入交换机宜采用1:1无收敛组网，并根据业务要求选择合适的算存比。样本面网络同样要求达成可靠性的关键技术能力要求。
- 业务面网络：承担系统业务调度与带内管理流量，采用多层CLOS组网，通常部署为TCP/IP有损网络，对可靠性有要求。
- 带外管理面网络：承担集群设备（包括服务器、交换机、防火墙等）的带外管理流量，接口速率以千兆为主，通常采用接入-汇聚层次化组网，可以采用较高的收敛比，并对可靠性有要求。

7 智算数据中心网络建设关键技术能力要求

7.1 网络规模

网络规模随着集群总算力要求的增长而同步扩大。在Scaling law原则下，随着模型参数量从万亿、十万亿的增长，模型训练使用的算力卡也从万卡到十万卡发展，智算网络规模也相应的要求支持万卡、十万卡级高速无损互联。而随着ChatGPT的引爆，未来400GE网络的服务器成为主流，对应的AI训练网络速率同样需要400GE。因此大规模的智算网络互联，一方面要提升网络设备带宽容量，另一方面要增强组网方案，以支持万卡、十万卡级别互联，并同时确保通信效率和可靠性不下降。

针对智算网络规模定义如下关键能力要求：

a) 网络设备支持200G/400G以及更大的通信带宽：200G/400G适配当前不同AI训练速率的服务器要求，提供对应的通信服务。未来随着智算需求的增长，要能支持提供更大带宽设备能力；
b) 两级CLOS组网要能扩展支持万卡规模集群互联。当前主流AI大模型参数规模达到万亿级别，需要万卡规模集群才能满足训练诉求。而两级CLOS扁平化组网架构是当前最主流的智算网络架构，要求能支持万卡规模集群互联；
c) 智算网络组网架构支持集群规模从万卡级演进到十万卡级。支持通过组网方案优化来提升十万卡规模集群互联能力，同时要求不降低网络通信效率及可用性。

7.2 网络通信效率

集群总算力与有效算力率紧耦合。有效算力率是一个衡量数据中心或计算平台实际计算能力利用效率的指标。它通常指的是实际用于计算任务的算力与总可用算力的比值。在智算网络中，有效算力受集合通信效率和网络通信效率的影响，而且模型参数增大，网络通信的端到端耗时占比将急剧上升。高效率的AI大模型训练对智算网络的核心诉求是高吞吐。网络高吞吐涵盖AI大模型训练的各类场景，包括单任务和多任务大模型训练场景，多任务在交换机下共部署场景，以及故障重调度导致的碎片等常见业务场景，要求各类场景下网络的有效吞吐均大于90%。

针对智算网络高吞吐定义如下关键能力要求：

a) 支持网络级负载均衡：AI集群训练场景，流量周期性循环进行、单流带宽大、流数量少，训练性能受限于最慢的流。在人工智能训练场景中,传统的基于五元组(源IP、目的IP、源端口、目的端口和协议)的静态负载均衡方式存在明显不足。由于AI训练任务通常具有流量周期性强、单流带宽大但流数量少的特点,导致使用五元组进行哈希算法时,容易出现哈希冲突,进而引发流量在链路间分配不均，出现网络吞吐下降，从而导致集群业务性能不高。网络拥塞存在本地冲突和全局冲突，本地冲突可以通过分布式决策算法解决，但全局冲突需要基于智算集群全局视角进行集中资源分配。支持网络级负载均衡，用以解决本地和全局的流量负载不均衡导致的流量拥塞，提升网络有效吞吐；
b) 支持流控PFC（Priority-based Flow Control，基于优先级的流量控制）：PFC用于解决多类型流量共享同一物理链路时的流量管理和拥塞控制，通过优先级的机制来管理流量，保障计算流量的低时延无损传输，支持死锁检测、死锁恢复与死锁控制，具备应对PFC死锁故障的能力，防止PFC死锁扩散，降低因PFC控制业务而导致网络瘫痪的风险，有效提升网络吞吐；
c) 拥塞控制ECN（Explicit Congestion Notification，显式拥塞通告）：ECN是一种端网协同的拥塞控制机制，网络检测拥塞，并通过数据包标识拥塞通知发送端进行降速解除拥塞，提升网络吞吐。

7.3 高可靠性

集群算力的另一个重要因素是算力可用率，指在训练任务过程中，计算资源实际可用于计算任务的时间与总时间的比率。它反映了计算资源的可用性，是衡量数据中心或计算平台资源利用率的一个重要指标。而网络的可靠性影响算力可用率。随着算力网络规模的扩张，构建网络的设备、光模块数量急剧增加，网络故障域也相应变大，一个网络节点故障将影响数十个计算节点的连通性。要达成高可靠性，一方面要优化组网架构，如每台服务器的参数面网口接在同一台接入交换机下以减少接入交换机整机故障时所影响的服务器数量，同时尽可能不跨spine转发以避免影响转发性能。另一方面，由于依赖人工分析排查将耗时数小时，网络的可观测、可预防、可快速恢复就变得至关重要。

针对智算网络高可靠性定义如下关键能力要求：

a) 支持设备升级不断训：大模型训练的周期长，对集群的可靠性要求高，但设备升级的情况难以避免，要支持设备升级；

b) 支持光通道级故障不断训：在多光通道情况下，对光通道进行抗损检测，单光通道级故障不影响其他通道通信，光接口仍可用，避免训练流量中断；
c) 支持光链路脏污和松动检测：AI智算场景下光模块闪断频发，多数由光链路脏污或松动导致。支持光链路脏污和松动检测，可以实现训前健康检查，以及训中端口闪断问题快速定位。

7.4 可视运维

智算网络作为集群共享资源，性能波动会导致计算资源利用率受到影响，但是当前仍缺乏网络性能统一观测和风险预警的手段。 随着智算网络故障域扩大，一个网络节点故障将影响数十个计算节点的连通性，依赖人工分析排查将耗时数小时，效率非常低。可视化运维可有效提升智算网络运维效率。

针对智算网络的可视运维定义如下关键能力要求：

a) 支持Telemetry数据采集机制：支持通过Telemetry实时采集网络设备运行状态数据，提升数据采集精度；
b) 网络丢包检测：智算场景需要网络进行无损转发，少量丢包会导致集合通信大幅下降并拖慢整个集群训练速度；因此智算网络需要具备NPU/GPU卡间通信流量1‰以上的丢包率检测精度，同时具备丢包位置识别能力，以便在发生丢包时及时进行干预避免单点丢包拖慢整个智算集群；
c) 网络丢包原因定位：智算网络出现丢包时要根据丢弃原因快速闭环以便恢复训练业务，因此智算网络需要能直接查询、定位因端口拥塞、ACL丢弃、路由查表失败等原因导致的丢包计数；
d) RDMA通信性能监控：AI训练过程中卡间通信通常为毫秒级的突发流量，通过端口带宽利用率无法有效监控RDMA流级(IP对)通信性能，因此智算网络需要能够监控RDMA通信的流完成时间、流有效吞吐等指标；
e) 网络拥塞监控：网络中存在拥塞时会影响集合通信性能，但是端口/队列的PFC报文计数受xon/xoff水线、流量模型等影响无法量化评估网络拥塞程度，因此智算网络需要具备端口/队列反压时长/反压Pause监控和统计能力，量化拥塞程度；
f) 任务视角网络监控及故障分析：AI智算场景网络规模大、一个训练任务最多调用几千至几万个NPU/GPU卡，任务出现异常时网络运维团队无法将训练任务和网络IP地址、网络设备/端口进行关联，因此智算网络需要支持自动关联到计算侧训练任务调度到的训练卡IP地址信息进行网络监控及故障自动化分析排障功能；
g) 北向能力集成：AI智算集群包含计算、存储、网络等多个领域的软硬件系统，因此智算数据中心网络运维系统需要支持将上述关键能力的监控指标及故障分析数据通过北向接口提供，以便被AI智算集群统一运维系统集成、提供跨域协同的指标监控及故障定界、定位能力。

7.5 快速部署

由于智算项目的高资本投资的属性，具备极致交付效率，减少交付时间，缩短业务上线时间就是保护投资。整体方案需要支持自动化部署，能力包含网络规划工具，设备配置自动化生成和下发，连线拓扑正确性检查（万卡数十万卡，线缆量非常庞大，人员施工易错，拓扑校验可大幅降低工程排查时间），算网参数自动化校验，业务上线前验收自动化等能力。

针对智算网络的快速部署定义如下关键能力要求：

a) 网络规划工具自动化：网络规划功能，一方面要满足网络侧部署参数的设定，另一方面网络和计算之间的耦合参数需要实现统一规划，如算网对接的接口参数等，以保证准确性；
b) 设备配置自动化生成和下发：网络规划到设备配置自动化生成，减少人工干预，提高正确率，同时支持自动完成配置下发，提升部署效率；

c) 自动化连线拓扑正确性检查：设备完成自动化部署后，在万卡到十万卡大规模组网场景下，线缆量庞大，依赖人工排查，工作量大且容易出错，提供自动化连线拓扑校验可大幅降低工程排查时间，并提升准确度；

d) 算网参数自动化校验：在万卡到十万卡大规模组网场景下，网络与计算的对接接口数量激增，配置错误点也响应的增加。自动化校验网络和计算对接参数，如光模块抗损参数、PFC无损参数等，可以快速排除系统间对接失败问题。

7.6 安全

大规模智算集群投资巨大，因此模型、样本、训练数据安全尤其重要，一旦产生数据安全问题，企业将产生巨大的损失。智算网络对内涉及多平面互联，对外涉及与服务器端侧互联，需要提供对应的安全能力，保障安全通信。

智算网络针对安全定义如下关键能力要求：

a) 支持租户隔离：智算网络是多用户共享的网络，租户间需要隔离，避免数据泄露；

b) 支持TLS安全加密通道：智算集群的四个典型平面间互联，要支持安全通道连接；同时支持与计算服务器端侧进行训练任务同步等对接时的安全加密通道。

7.7 高效开放

智算网络规模的快速扩张，对算力、网络、存储在组网规模、负载均衡、拥塞控制的诉求不断增强，联合技术创新节奏加快。从保护既有投资和生态构建角度，要求智算数据中心网络建设具备更开放的网络方案。以太网技术在超大规模组网（含100G、200G、400G、800G及1.6T）标准上领先与其他技术，同时在负载均衡、流控和拥塞控制方面更具有优势。同时针对智算数据中心运营一体化的诉求，网络运维能力应具备北向开放能力，实现与其他系统集成。

智算网络针对高效开放定义如下关键能力要求：

a) 支持以太网络技术：基于参数面、样本面的智算网络建设优选以太网络协议方案，以实现融合承载，保持可持续演进性；

b) 支持智算网络运维北向接口开放：可以与计算端侧、存储侧等运维通过北向接口集成，增强智算数据中心端到端的一体化运维能力。