ICS 33.040.40
CCS I631

T/NIDA

全 球 固 定 网 络 创 新 联 盟

T/NIDA 008-2025

智能计算数据中心网络关键能力测试方法

Test Method of Key Capability of Intelligent Computing Data Center Network

2025-10-26 发布 2025-11-01 施行

全球固定网络创新联盟（NIDA）发布

目录

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云计算与大数据研究所、华为技术有限公司

本文件主要起草人：郭亮、王少鹏、张力、侯延祥、潘洋、顾咸勇

智能计算数据中心网络关键能力测试方法

1 范围

本文件规定了智算网络关键能力的测评项，包括：建网开局能力、配套昇腾和英伟达算力的单训练任务、多训练任务及推理网络性能指标、故障主动预防能力、高可用能力、网络及业务质量深度运维能力、算网协同能力。

本文件适用于当前商用智算网络产品的成熟度评测，也可以供算力中心建设方作为产品选型参考。

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

4 缩略语

下列缩略语适用于本文件。

DC: 数据中心 (Data Center)

DPU: 数据处理芯片 (Data Processing Unit)

ECMP: 等价路由负载分担 (Equal-Cost Multi-Path)

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

FCT: 流完成时间 (Flow Completion Time)

FET: 流有效吞吐 (Flow Efficient Throughput)

FNR: 流重传率 (Flow NAK Rate)

LLD: 详细设计 (Low-level Design)

LLDP: 链路层自动发现协议 (Link Layer Discovery Protocol)

LLM: 大语言模型 (Large Language Model)

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

PRBS: 伪随机码序列 (Pseudo Random Binary Sequence)

RDMA: 远程直接内存访问 (Remote Direct Memory Access)

KPI: 关键性能指标 (Key Performance Indicator)

KQI: 关键质量指标 (Key Quality Indicator)

XCCL: X集合通信库 (X Collective Communication Library)

ZTP: 零接触配置 (Zero Touch Provisioning)

5 测试环境

6 测试框架

智算数据中心网络测试是基于集中控制平台+网络设备的综合能力测试。测试框架如图 1 所示：

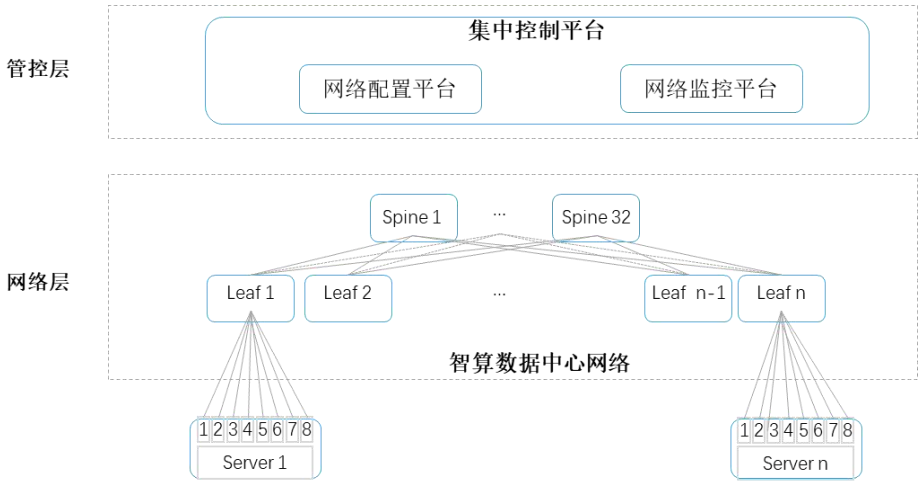


图 1 智算数据中心网络测试架构

集中控制平台包含网络配置、网络监控等平台能力，用于对智算数据中心网络的部署、管理、控制和运维。

智算数据中心网络由交换机组成，提供满足智算需求的网络层能力。

6.1 测试床-昇腾算力

智算数据中心网络针对华为昇腾算力卡的网络成熟度测试采用 1:1 收敛的 CLOS 二层组网架构，由两台 Spine 设备，8 台 Leaf 设备及 32 台服务器共 256 张昇腾算力卡卡组成。其中，两台 Spine 设备下挂 8 台 Leaf 设备，Spine 和 Leaf 之间互联接口速率为 400Gbps；每台 Leaf 设备下挂 4 台包含 8 张昇腾算力卡的算力服务器，服务器接入 Leaf 设备的速率为 200Gbps。基于华为昇腾算力卡的测试组网拓扑如图 2 所示：

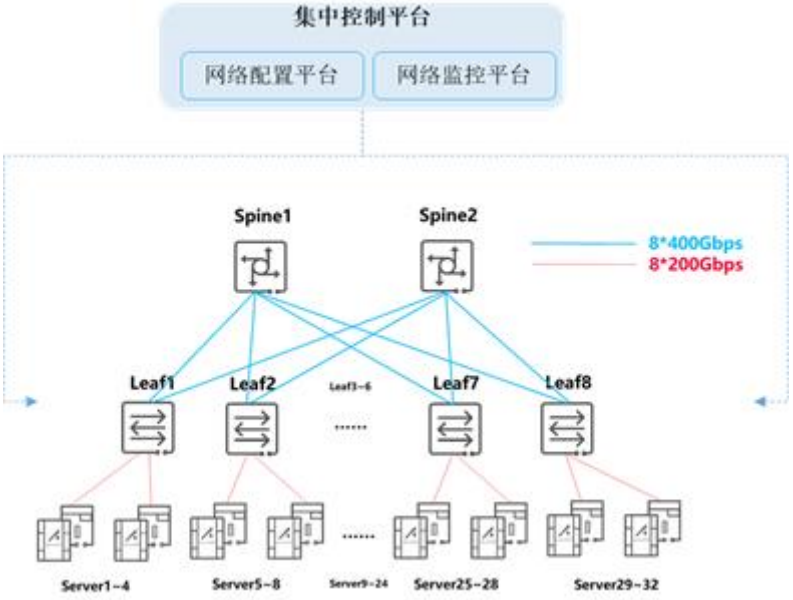


图 2 基于华为昇腾算力卡的智算数据中心网络测试组网

6.2 测试床-英伟达算力

智算数据中心网络针对英伟达算力卡的网络成熟度测试采用 1:1 收敛的 CLOS 二层组网架构，由两台 Spine 设备，8 台 Leaf 设备及 8 台服务器共 64 张英伟达算力卡卡组成。其中，两台 Spine 设备下挂 8 台 Leaf 设备，Spine 和 Leaf 之间互联接口速率为 400Gbps；每台 Leaf 下挂 1 台包含 8 张英伟达算力卡的算力服务器，服务器接入 Leaf 设备的速率为 200Gbps。基于英伟达算力卡的测试组网拓扑如图 3 所示：

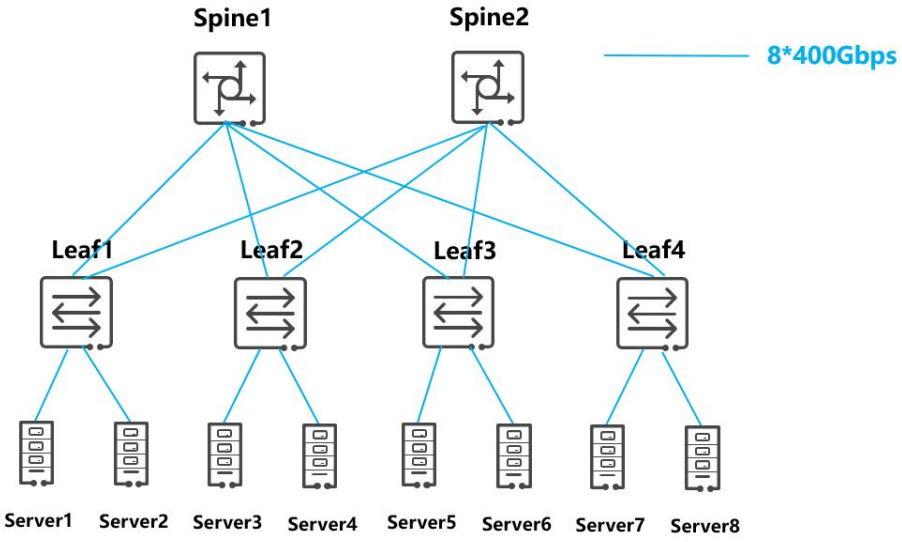


图 3 基于英伟达算力卡的智算数据中心网络测试组网

7 建网开局能力测试

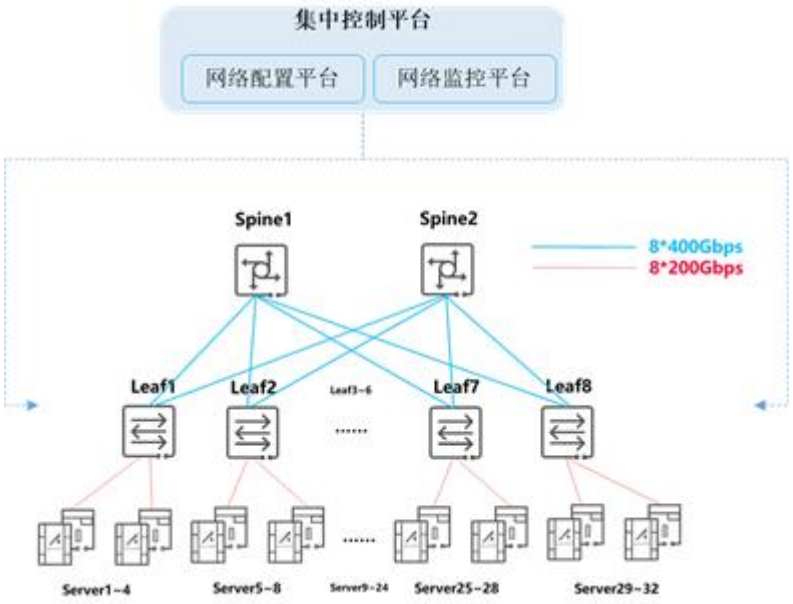
7.1 网络部署自动化

当前智算数据中心网络部署与算侧部署解耦，网络侧通过人工完成 LLD 详细设计后，再由人工基于 LLD 手动生成网络设备配置。这种依赖人工的配置方式低效易错，且容易与计算侧产生配置冲突。随着万卡、十万卡规模的智算集群的出现，网络部署效率低、网络配置错误率高等问题日渐突出，且人工排查难度急剧上升。

训前网络部署自动化指的是基于端网协同的网络零接触配置 ZTP 开局技术。网络配置平台在网络建设阶段支持 LLD 导入，同时支持自动获取基于 LLDP 的网络设备组网拓扑信息。网络配置平台通过对 LLD 和 LLDP 拓扑进行映射，自动生成设备配置并完成自动化部署，以保持端网参数一致无差错。

网络部署自动化可以缩短大规模智算数据中心网络的部署时间，并提升网路部署准确率。

测试编号	
测试目的	验证网络零接触配置 ZTP 能力

组网拓扑	
预置条件	. 网络配置平台已完成部署,已设置 SNMP、NETCONF、SSH 等参数，且设置 DHCP 资源池。 . 网络配置平台按需选择带内或带外管理网络接入方式
测试步骤	1. 操作员将智算数据中心网络中Spine、Leaf设备按需连接带内或带外管理网络（与网络配置平台接入网络方式保持一致），设备上电，检查是否有预期结果1； 2. 操作员在网络配置平台上导入LLD模板，选择要下发配置的设备，检查是否有预期结果2； 3. 操作员确认配置下发，网络配置平台自动执行设备配置下发，检查是否有预期结果3。
预期结果	1. Spine、Leaf设备可被网络配置平台成功纳管，可在网络配置平台上看到设备； 2. LLD模板导入成功，并自动生成设备配置，且配置与选定的设备对应关系正确； 3. 登录设备检查下发的配置，确认设备互联配置、路由协议等配置下发成功。
测试结果	
备注	无

7.2 训前网络全量自检

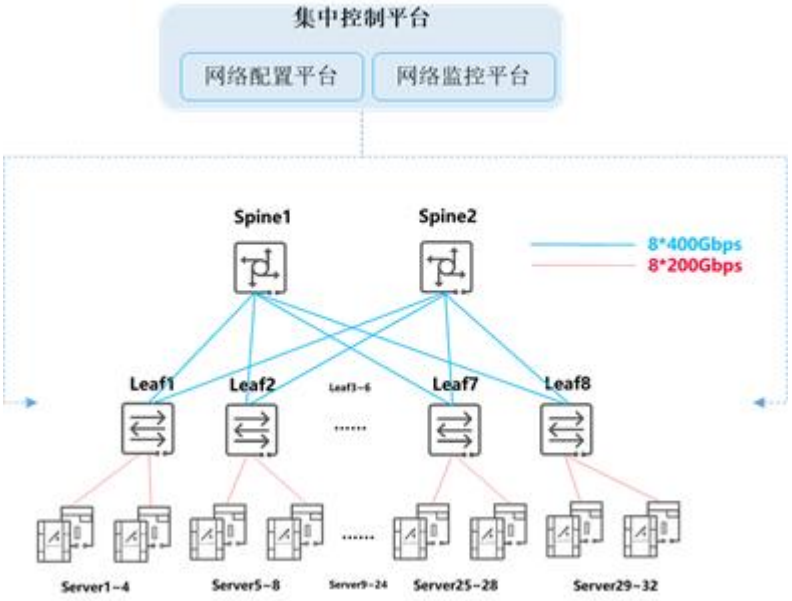
训前网络全量自检能力指的是在模型训练任务开始前对智算数据中心网络的健康度进行评估，包含设备、端口、光模块状态，以及工程连线是否正确、网络端口线速转发能力是否与预期相符等各种能力。本用例设计选取大模型训练更为关注的工程连线、网络端口线速转发能力做重点测试。

物理工程连线错误指物理网络连接与 LLD 规划不相符，在大规模网络中出现的概率很高，且连线错误通过人工排查常常耗时很长。提供自动化的工程连线错误排查手段，并能够给出正确的连线建议，

能有效提高大规模建网的效率。

网络端口线速转发能力，是指在开启大模型训练之前，对网络中各链路光模块的线速转发能力和误码情况进行测试。提供自动化的网络链路光模块测试（如伪随机码序列(Pseudo Random Binary Sequence, PRBS)码流压测）和测试结果分析，能提高检测效率。PRBS 码流压测是一种测试光模块线速转发情况下误码率的技术。

训前网络全量自检，可通过多维度网络状态检测和分析，识别网络隐患，提前进行网络优化，有效支持大模型训练任务的长时间不间断运行。

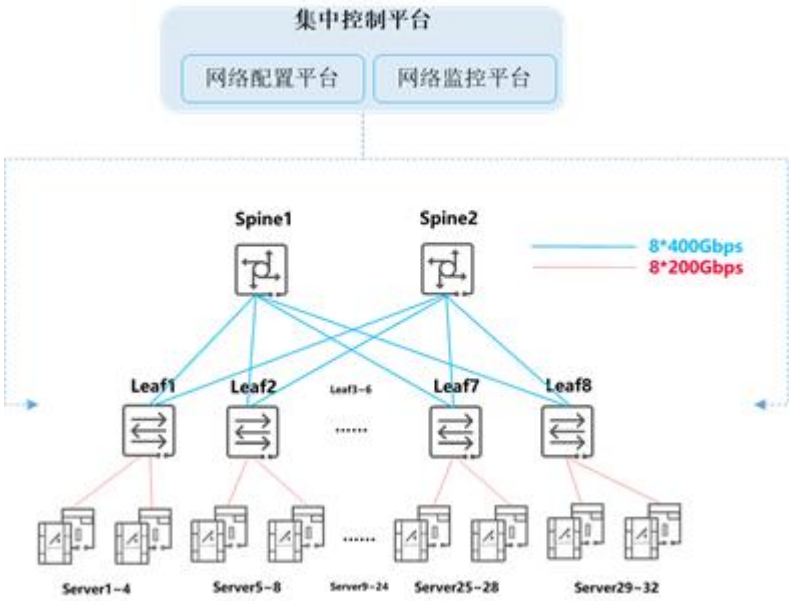
测试编号	
测试目的	验证训前智算数据中心网络对工程连线错误、端口转发带宽和误码率等异常自检能力。
组网拓扑	
预置条件	1. 网络配置平台、网络监控平台功能正常，设备被正常纳管
测试步骤	1. 在连线时构造工程连线错误，Server 到 Leaf 连线错误（如 Server1 连到 Server2 的接入口，Server2 连到 Server3 的接入口，Server3 连到 Server1 的接入口）或 Leaf 到 Spine 连线错误，在网络配置平台检查是否有预期结果 1; 2. 选择 Leaf 和 Spine 间或者 Server 和 Leaf 间的连接批量自动化执行光模块 PRBS 码流的双向压测，检查是否有预期结果 2； 3. 将光模块替换为存在误码故障的光模块，再次执行步骤 2，有预期结果 3。
预期结果	1. 网络监控平台可以检测到 Server1、Server2 和 Server3 的连线错误，或 Leaf 和 Spine 间的连线错误，并给出纠错建议； 2. 网络监控平台可呈现压测结果，并展示光链路质量信息； 3. 网络监控平台可显示被替换为故障光模块的光链路有异常。
测试结果	

备注	

7.3 队列一致性检测

队列一致性检测是指当网络设备开启无损队列功能时，集中控制平台可以针对 Leaf 和 Spine 设备的无损队列配置进行一致性校验，避免出现无损队列在不同设备上配置不一致，导致大模型训练无法端到端执行无损队列保障。

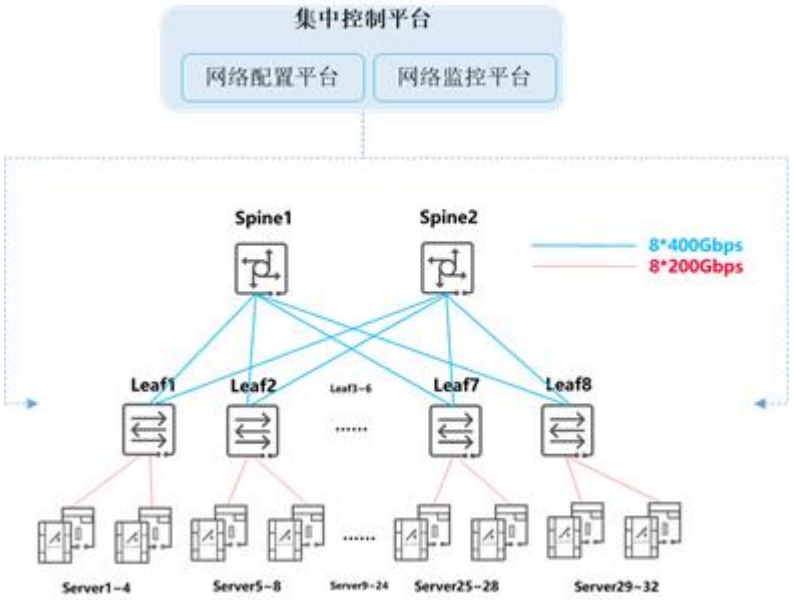
该功能的引入，可自动化执行无损队列不一致校验，提升准确度和时效性。

测试编号	
测试目的	验证无损队列一致性自动化校验能力
组网拓扑	
预置条件	集中控制平台各系统功能正常，设备被正常纳管。
测试步骤	1、配置 Leaf 设备的无损队列为队列 3，Spine 设备的无损队列为队列 4，对网络进行风险评估，有预期结果 1。
预期结果	1、集中控制平台可识别出设备无损队列配置不一致风险，可呈现 Leaf 的无损队列为队列 3，Spine 的无损队列为队列 4。
测试结果	
备注	

7.4 网络配置调优

训前网络配置调优指的是在 AI 大模型训练任务开始前，可以根据智算网络连接拓扑和 AI 训练任务通信特点，自动优化智算数据中心网络的静态配置，以更适配大模型训练性能要求。

网络配置调优可优化缺省网络配置，起到提高大模型训练效率的作用。

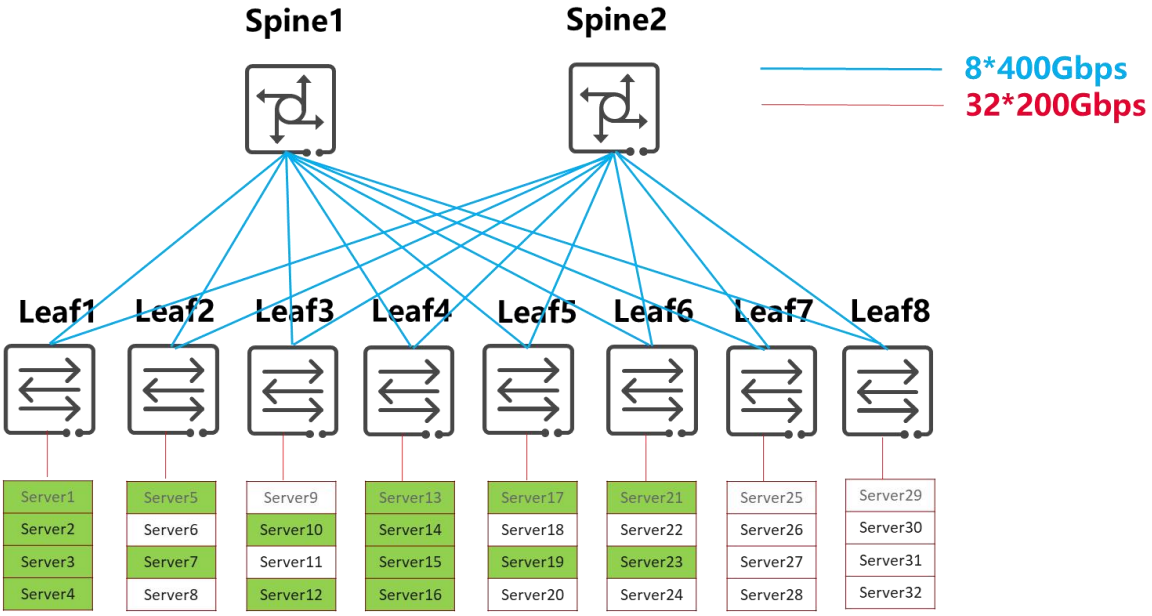
测试编号	
测试目的	测试针对大模型训练任务的智算数据中心网络配置优化能力
组网拓扑	
预置条件	集中控制平台各系统功能正常，设备被正常纳管。 已导入/同步 AI 训练任务。
测试步骤	1. 启动一个all-reduce业务，得到预期结果1； 2. 网络配置平台自动生成优化静态配置，并下发网络设备，检查是否有预期结果2； 3. 重复步骤1，有结果3。
预期结果	1. 记录 all-reduce 业务通信性能 R1； 2. 记录配置下发的方式（如方式 1：网络配置平台可以自动识别网络拓扑并下发网络优化配置；方式 2：网络配置平台支持导入网络设备配置并批量下发网络优化配置），并观察配置下发成功后； 3. 再次记录 all-reduce 业务新的通信性能 R2，R2 相较于 R1 性能有所提升。
测试结果	
备注	

8 昇腾算力训练任务网络性能测试

8.1 单训练任务性能测试

同一 GPU 集群在长期运行后，由于承载多个训练任务，且各训练任务所需服务器数量以及训练任

务的起始和终止时间均不相同，会导致同一个模型训练任务所需的算力服务器呈现分散在相同集群内不同交换机下的非连续性分布状态，即算力服务器资源的碎片化。如下图所示：



单训练任务指的是智算集群内仅运行一个大模型训练任务。但是单训练任务运行在碎片化分布的物理服务器上时，也会因集合通信而竞争网络资源。因此单训练任务性能测试也需要充分考虑算力服务器资源的碎片化场景测试。

单训练任务性能的测试结果体现为训练的平均带宽和平均迭代时长两个指标。训练平均带宽指完成一次训练任务所涉及的各个算力卡的平均带宽，而平均迭代时长则是完成一次训练迭代所需的时间。通常，影响测试指标的因素包括模型参数量、XCCL 集合通信库（包括算子、算法）、网络带宽和时延、网络拥塞和丢包等。其中网络拥塞和丢包严重影响基于 RDMA 协议的应用性能，因此智算网络要求是无损网络，而网络级负载均衡对保障网络无损非常重要。

因此单训练任务性能测试要考虑基于集合通信库测试。不同厂商的集合通信库（X Collective Communication Library, XCCL)会根据硬件能力和系统环境（包括计算、存储、网络等）在算子（如 ALLReduce，ALLtoALL，ReduceScateer，ALLGather）的基础上提供多套常见算法（如 Ring，Tree，以及 HCCL 提供的 H,N,B,P 等算法）。通常在模型训练时要选取不同的算子和算法组合，才能达到最佳的计算性能。用例设计考虑多种算子和算法组合情况下的性能测试。且为有效测试各类算子和算法组合在各种碎片化场景中的能力，取如下 10 种场景进行测试以评估网络性能：

表 1 单训练任务 XCCL 测试资源组合场景

测试场景	资源组合
场景 1	0 1 2 3 4 6 9 11 12 13 14 15 16 18 20 22
场景 2	0 1 2 3 4 6 16 18 12 14 9 11 20 21 22 23
场景 3	0 1 2 3 4 5 6 8 12 13 14 15 16 20 22 23
场景 4	4 5 6 9 0 1 2 3 17 19 21 23 12 13 14 15
场景 5	12 14 5 7 20 21 22 23 1 3 9 11 16 17 18 19

场景 6	3 1 0 2 4 6 8 10 15 12 13 14 7 5 11 9
场景 7	6 4 9 11 13 14 12 15 5 7 10 8 3 1 2 0
场景 8	8 10 9 11 1 7 3 5 0 6 2 4 13 14 15 12
场景 9	12 6 14 4 0 3 1 2 5 13 7 15 8 11 9 10
场景 10	4 6 5 7 2 12 0 14 9 8 11 10 15 1 3 13

注：资源组合中，0代表Server1，1代表Server2，31代表Server32，以此类推。

针对单任务基于 XCCL 的测试，需要考虑字节遍历。字节遍历是为了测试一种算子和算法组合在不同通信量情况下的性能表现。建议每种资源组合场景下，字节遍历取 128MB、256MB、512MB、1GB、2GB、4GB 的情况，每类字节数场景测试 3 轮，最终取所有场景下的测试结果平均值为该算子和算法组合情况下的最终测试结果。具体计算如下：

表 2 单任务独占碎片化环境 XCCL 测试资源组合结果记录

资源组合 / 字节遍历		场景 1	场景 2	场景 3	场景 4	场景 5	场景 6	场景 7	场景 8	场景 9	场景 10
第一轮	128MB										
	256MB										
	512MB										
	1GB										
	2GB										
	4GB										
第二轮	128MB										
	256MB										
	512MB										
	1GB										
	2GB										
	4GB										
第三轮	128MB										
	256MB										
	512MB										
	1GB										
	2GB										

	4GB										

除基于集合通信库的测试外，还应引入大模型进行基于大模型的测试，通过大模型训练结果来端到端检测算网的执行结果，主要体现为大模型训练平均迭代时长。在大模型测试中同样取多种碎片化场景进行测试:

表 3 单训练任务大模型测试资源组合场景

测试场景	资源组合
场景 1	0 14 20 26 1 15 21 27 2 8 22 28 3 9 23 29 4 10 16 30 5 11 17 31 6 12 18 24 7 13 19 25
场景 2	0 12 16 28 1 13 17 29 2 14 18 30 3 15 19 31 4 8 20 24 5 9 21 25 6 10 22 26 7 11 23 27
场景 3	0 15 22 29 1 8 23 30 2 9 16 31 3 10 17 24 4 11 18 25 5 12 19 26 6 13 20 27 7 14 21 28
场景 4	0 13 18 31 1 14 19 24 2 15 20 25 3 8 21 26 4 9 22 27 5 10 23 28 6 11 16 29 7 12 17 30

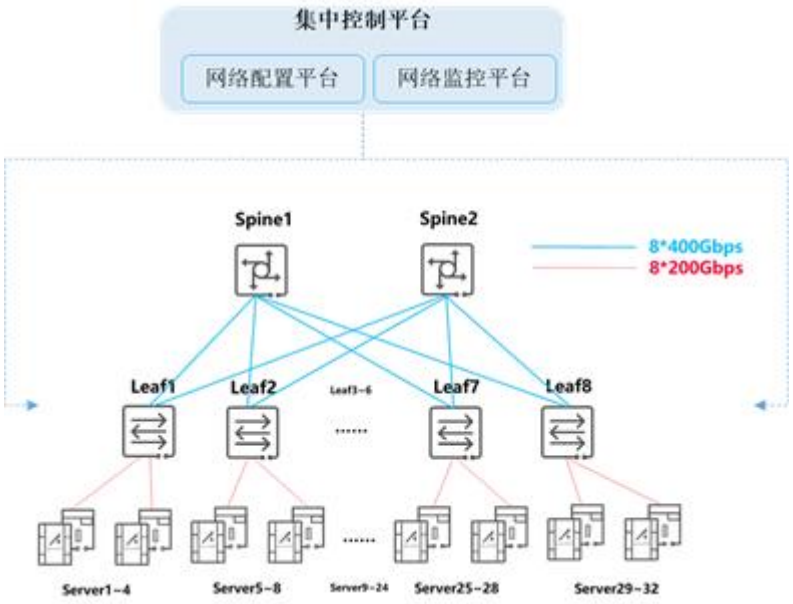
每种大模型分别给定测试结果:

表 4 单任务独占碎片化环境大模型测试资源组合结果记录

模型 / 场景	大模型
场景 1	T1
场景 2	T2
场景 3	T3
场景 4	T4
测试结果	$(\sum_{i=1}^{4} MTi)/4$

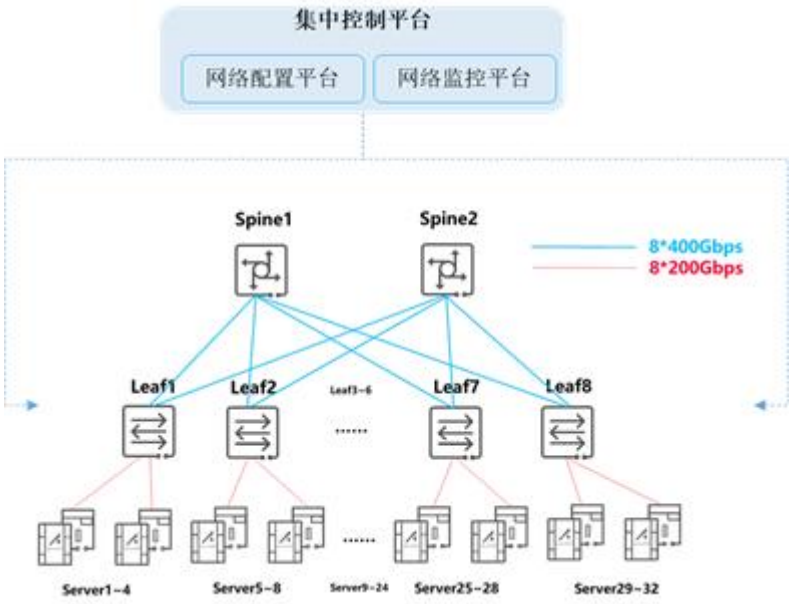
8.1.1 AllReduce-HD 测试

测试编号	
测试目的	测试整个集群为一个 AllReduce 工作负载在 HD 算法操作中的平均带宽。

组网拓扑	
预置条件	. 集中控制平台各系统部件运行正常； . 智算数据中心网络已纳管，且通过自检；
测试步骤	1. 使用HCCL测试跨128卡节点的AllReduce-HD通信模型的性能； 2. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； 3. 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1； 4. 计算测试结果，有预期结果2。
预期结果	1. 收集测试结果，并记录表2； 2. 基于表2计算HCCL AllReduce-HD每个字节遍历场景下的平均带宽（GBps）
测试结果	
备注	

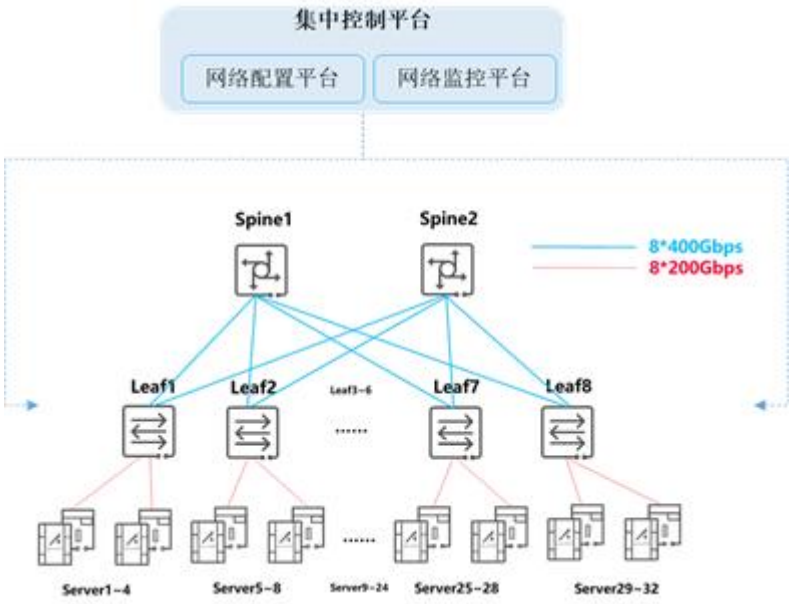
8.1.2 AllReduce-NHR 测试

测试编号	
测试目的	测试整个集群为一个 AllReduce 工作负载在 NHR 算法操作中的平均带宽。

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	. 使用HCCL测试跨128卡节点下AllReduce-NHR通信模型的性能； . 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； . 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1； . 计算测试结果，有预期结果2。
预期结果	1. 收集测试结果，并记录表2； 2. 基于表2计算HCCL AllReduce-NHR每个字节遍历场景下的平均带宽（GBps）
测试结果	
备注	

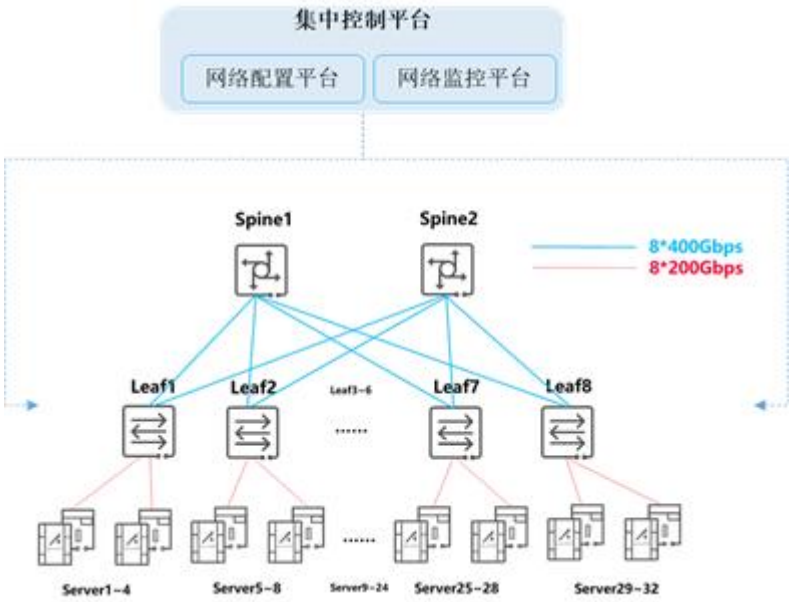
8.1.3 AllReduce-NB 测试

测试编号	
测试目的	测试整个集群为一个 AllReduce 工作负载在 NB 算法操作中的平均带宽。

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用HCCL测试跨128卡节点下AllReduce-NB通信模型的性能； 2. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； 3. 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1； 4. 计算测试结果，有预期结果2。
预期结果	. 收集测试结果，并记录表2； . 基于表2计算HCCL AllReduce-NB每个字节遍历场景下的平均带宽（GBps）
测试结果	
备注	

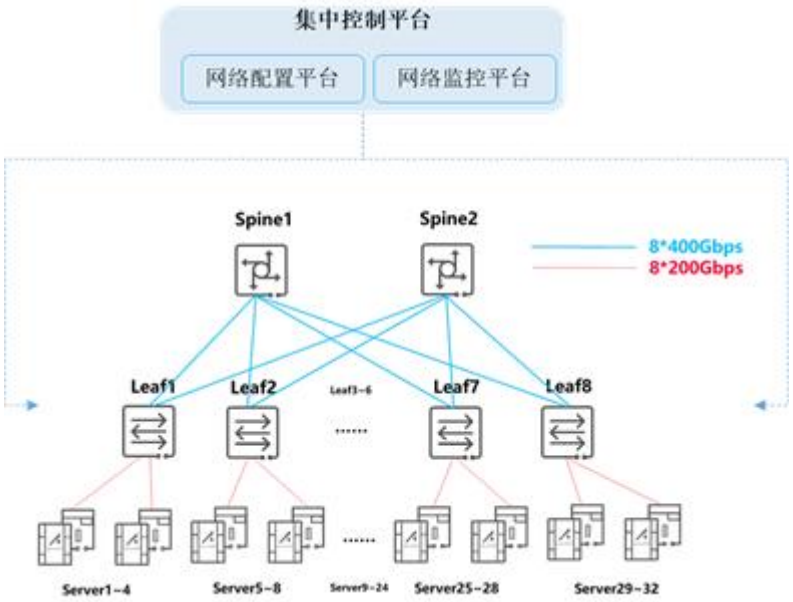
8.1.4 All To All 测试

测试编号	
测试目的	测试整个集群为一个 All to All 工作负载的平均带宽表现。

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用HCCL测试跨128卡节点下All to All 通信模型的性能； 2. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； 3. 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1 4. 计算测试结果，有预期结果2。
预期结果	. 收集测试结果，并记录表2； . 基于表2计算All To All 每个字节遍历场景下的平均带宽（GBps）。
测试结果	
备注	

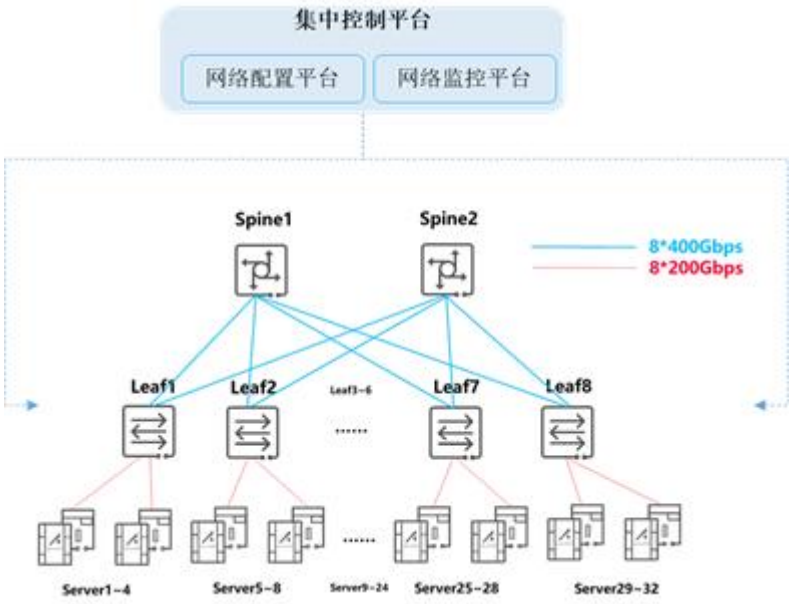
8.1.5 ReduceScatter-HD 测试

测试编号	
测试目的	测试整个集群为一个 ReduceScatter 工作负载在 HD 算法操作中的平均带宽。

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用HCCL测试跨128卡节点下ReduceScatter-HD通信模型的性能； 2. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； 3. 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1； 4. 计算测试结果，有预期结果2。
预期结果	1. 收集测试结果，并记录表2； 2. 基于表2计算ReduceScatter-HD每个字节遍历场景下的平均带宽（GBps）。
测试结果	
8.1.6 备注	**8.1.7**

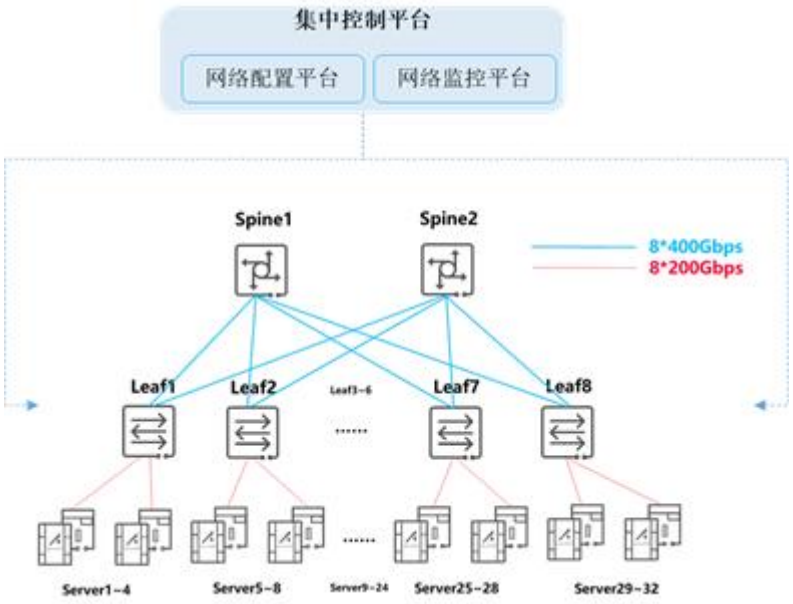
8.1.8 AllGather-HD 测试

测试编号	
测试目的	测试整个集群为一个 AllGather 工作负载在 HD 算法操作中的平均带宽。

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用HCCL测试跨128卡节点下AllGather-HD通信模型的性能； 2. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次； 3. 节点Rank选取参考表1，逐个场景启动字节遍历测试，记录预期结果1； 4. 计算测试结果，有预期结果2。
预期结果	1. 收集测试结果，并记录表2； 2. 基于表2计算AllGather-HD每个字节遍历场景下的平均带宽（GBps）。
测试结果	
备注	

8.1.9 LLM 端到端性能测试

测试编号	
测试目的	验证整个集群网络在特定 AI 模型训练下的性能

组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。 3. 服务器上已完成Mixtral 8*7B、llama2 13B模型的部署。
测试步骤	1. 测试32服务器256卡规模无损网络的大模型训练性能； 2. 选择训练模型为Mixtral 8*7B，节点Rank选取参考表3，逐个场景按照Mixtral 8*7B专家模型训练基线启动测试，记录预期结果1； 3. 切换训练模型为llama2 13B，重复步骤2，逐个场景中按照llama2 13B模型训练基线启动测试，有预期结果2； 4. 计算两种模型测试结果，有预期结果3。
预期结果	1. 在表4中记录Mixtral 8*7B模型训练任务各场景的训练平均迭代时长； 2. 在表4中记录llama2 13B模型训练任务各场景的训练平均迭代时长； 3. 在表4中记录Mixtral 8*7B和llama2 13B的训练平均迭代时长。
测试结果	
备注	1. 资源组合编排中，0 代表 Server1，1 代表 Server2，31 代表 Server32，以此类推。 2. Mixtral 8*7B 专家模型训练基线：https://gitee.com/ascend/MindSpeed-LLM/blob/master/examples/mcore/mixtral/pretrain_mixtral_8x7b_ptd.sh 3. 为了匹配 256 卡的测试床规模，避免训练时间过长，Mixtral 8*7B 专家模型需变更部分默认参数，分别为： TP=8 PP=1 EP=4 CP=4

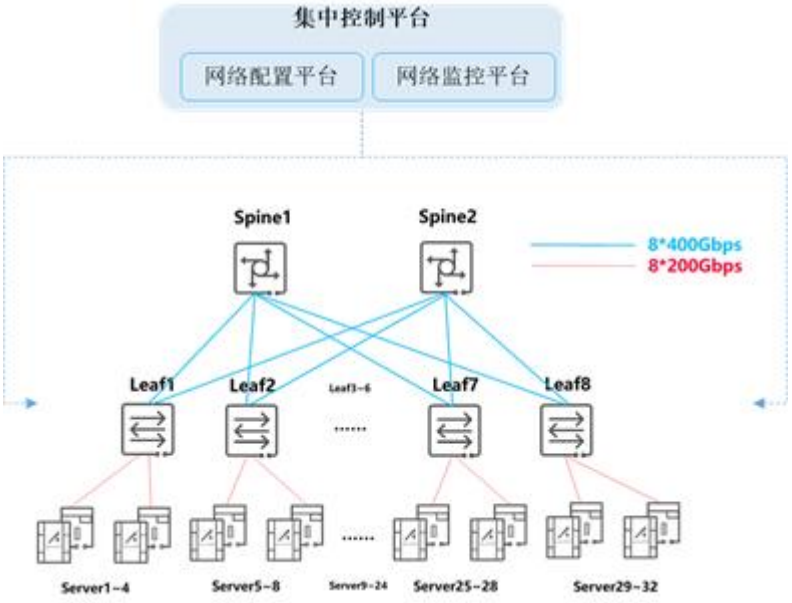
	NUM_LAYERS=16 --global-batch-size 16 \ 4. llama2 13B 模型训练基线： https://gitee.com/ascend/MindSpeed-LLM/blob/1.0.0/examples/mcore/llama2/pretrain_llama2_13b_ptd.sh 5. llama2 13B 训练参数调整： 256 卡 global-batch-size 调整为 256；64 卡 global-batch-size 调整为 128。

8.2 多训练任务性能测试

多训练任务指的是一个智算集群内运行多个大模型训练任务，在此过程中既存在单任务内因集合通信竞争网络资源，又存在多任务之间争抢网络资源情况。针对多个模型训练任务的性能测试，需同时考虑测试环境规模和测试执行效率的基础，选取两组任务和多组（≥3）任务进行设计。

基于多训练任务的测试例选取模型训练最常见的算子算法进行设计。

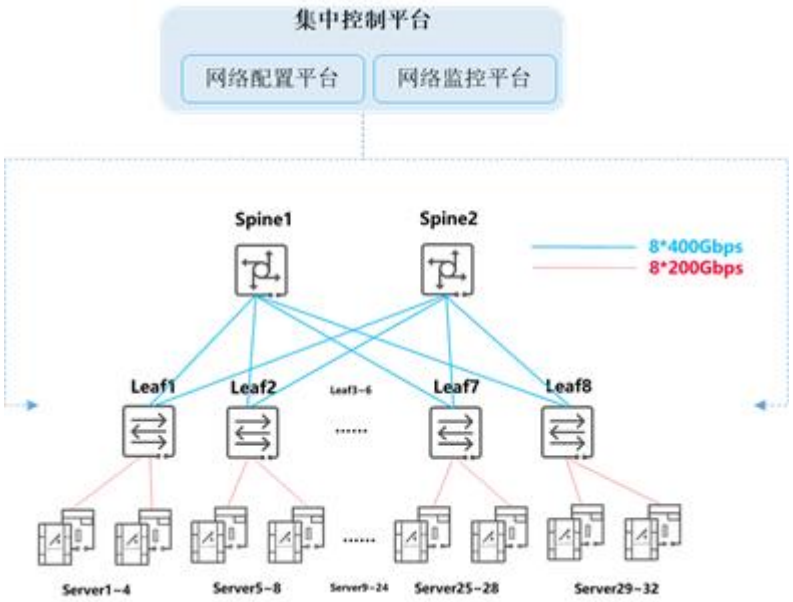
8.2.1 AllReduce 性能隔离测试

测试编号	
测试目的	分别测试两组工作负载和三组工作负载 HCCL AllReduce 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响
组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用 HCCL 同时测试两个工作负载在 AllReduce 通信模型的性能，其中负载 1 占用 16 个计算节点，采用 AllReduce-NHR 算子算法，负载 2 占用 8 个计算节点，采用 AllReduce-HD 算子算法，节点 rank 选取如下： 场景 1 资源组合：负载 1{4 6 7 0 14 12 19 13 18 1 10 11 9 8 5 15}；负载 2{23 22 20 3 17 21 16 2}

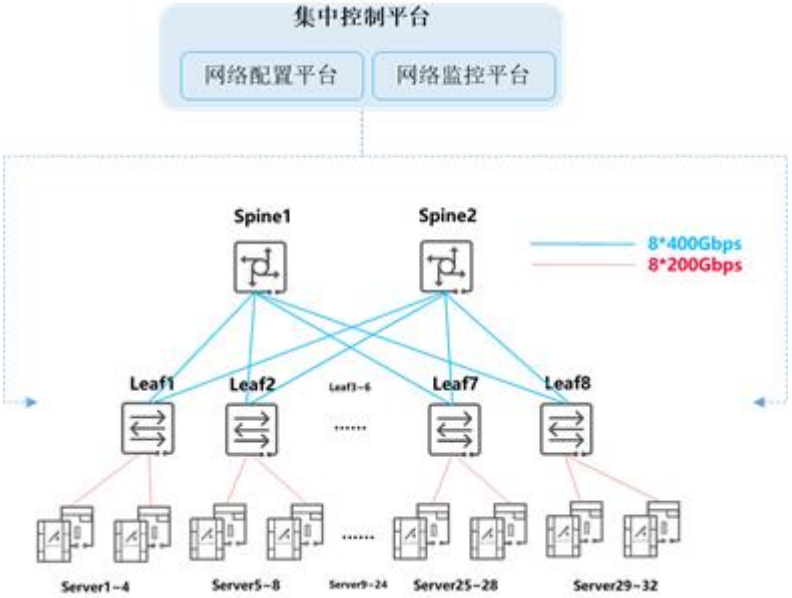
	场景 2 资源组合：负载 1{4 10 19 0 14 9 1 15 12 6 7 8 18 11 5 13}；负载 2{23 2 16 21 3 22 20 17} 场景 3 资源组合：负载 1{1 3 13 19 12 0 2 9 8 14 17 16 15 4 11 18}；负载 2{10 22 5 20 23 7 6 21} 场景 4 资源组合：负载 1{1 15 13 19 12 0 16 18 9 3 17 11 8 4 18 14}；负载 2{10 7 21 20 6 23 5 22} 测试字节为 4GB，分别测试 3 次，记录预期测试结果 1； 2. 使用 HCCL 同时测试三个工作负载的 AllReduce 通信模型的性能，其中负载 1 占用 8 个计算节点，采用 AllReduce-NHR 算子算法，负载 2 占用 8 个计算节点，采用 AllReduce-NB 算子算法，负载 3 占用 8 个计算节点，采用 AllReduce-HD 算子算法，节点 rank 选取如下： 场景 1 资源组合：负载 1{2 11 3 13 0 1 9 12}；负载 2{18 14 15 17 16 8 19 10}；负载 3{22 7 21 5 6 4 20 23} 场景 2 资源组合：负载 1{13 4 7 12 20 22 23 21}；负载 2{15 14 18 5 17 6 16 19}；负载 3{6 0 10 9 1 2 8 11} 测试字节为 4GB，分别测试 3 次，记录预期测试结果 2。
预期结果	1. 收集并分析两组工作负载的 HCCL AllReduce 的平均带宽 (GBps)； 2. 收集并分析三组工作负载的 HCCL AllReduce 的平均带宽 (GBps)。
测试结果	
备注	

8.2.2 All to All 性能隔离测试

测试编号	
测试目的	分别测试两组工作负载和三组工作负载做 XCCL All To All 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响。

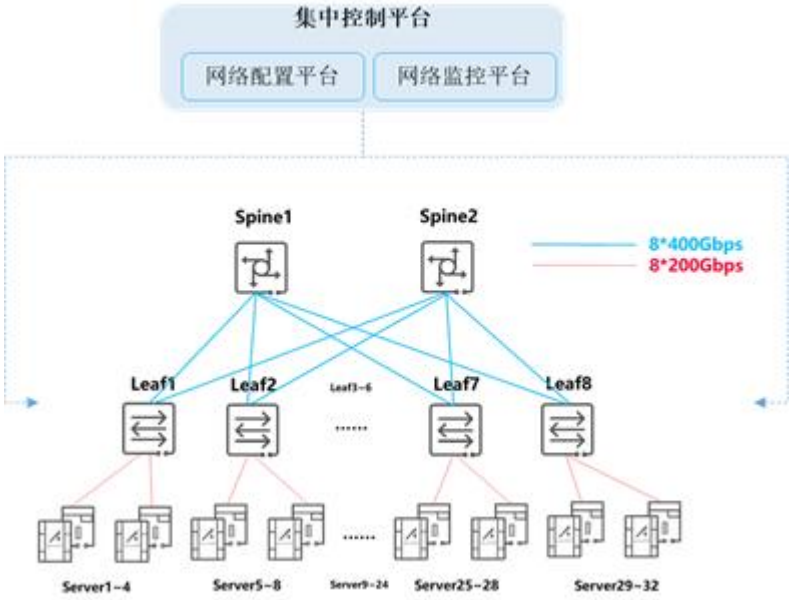
组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	1. 使用 HCCL 同时测试两个工作负载的 All To All 通信模型的性能，其中负载 1 占用 16 个计算节点，采用 All To All-Pairwise 算子算法，负载 2 占用 8 个计算节点，采用 All To All-Pairwise 算子算法。节点 rank 选取如下： 场景 1 资源组合：负载 1{15 13 3 2 17 0 8 16 18 4 19 1 12 10 6 14}；负载 2{7 9 20 5 22 23 11 21} 场景 2 资源组合：负载 1{22 7 14 23 0 8 2 13 15 6 1 20 3 5 21 12}；负载 2{18 9 4 10 17 11 16 19} 测试字节 4GB，分别测试 3 次，记录预期测试结果 1； 2. 使用 HCCL 同时测试三个工作负载的 All To All 通信模型的性能,其中负载 1 占用 8 个计算节点，采用 All To All-Pairwise 算子算法，负载 2 占用 8 个计算节点，采用 All To All-Pairwise 算子算法，负载 3 占用 8 个计算节点，采用 All To All-Pairwise 算子算法，节点 rank 选取如下： 场景资源组合：负载 1{2 13 12 15 14 0 3 1}；负载 2{20 7 22 23 11 10 9 21}；负载 3{18 4 5 6 19 8 17 16} 测试字节 4GB，分别测试 3 次，记录预期测试结果 2。
预期结果	1. 收集并分析两组工作负载的 HCCL All To All 的平均带宽 (GBps)； 2. 收集并分析三组工作负载的 HCCL All To All 的平均带宽 (GBps)。
测试结果	
备注	

8.2.3 AllReduce + All to All 性能隔离测试

测试编号	
测试目的	测试两组工作负载和三组工作负载，做 XCCL AllReduce 和 All To All 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响
组网拓扑	
预置条件	1. 集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。
测试步骤	使用 HCCL 同时测试两个工作负载的 AllReduce + All To All 通信模型的性能，其中负载 1 占用 16 个计算节点，采用 AllReduce-NHR 算子算法，负载 2 占用 8 个计算节点，采用 All To All-Pairwise 算子算法，节点 rank 选取如下： 场景资源组合：负载 1{10 0 15 16 17 1 3 6 4 12 2 7 8 11 5 9}；负载 2{18 19 23 13 14 20 22 21} 测试字节 4GB，分别测试 3 次，记录预期测试结果 1； 使用 HCCL 同时测试三个工作负载的 AllReduce + All To All 通信模型的性能，其中负载 1 占用 8 个计算节点，采用 AllReduce-NHR 算子算法，负载 2 占用 8 个计算节点，采用 AllReduce-NB 算子算法，负载 3 占用 8 个计算节点，采用 All To All-Pairwise 算子算法，节点 rank 选取如下： 场景 1 资源组合：负载 1{15 17 19 13 12 18 14 16}；负载 2{1 6 20 23 0 5 22 21}；负载 3{10 3 9 8 11 4 2 7} 场景 2 资源组合：负载 1{0 2 10 4 7 3 9 1}；负载 2{12 13 18 19 16 17 15 14}；负载 3{8 22 20 6 23 5 11 21} 测试字节 4GB，分别测试 3 次，记录预期测试结果 2。
预期结果	1. 收集并分析两组工作负载的 XCCL AllReduce + All To All 的平均带宽 (GBps)； 2. 收集并分析三组工作负载的 XCCL AllReduce + All To All 的平均带宽 (GBps)。

测试结果	
备注	

8.2.4 LLM 应用隔离测试

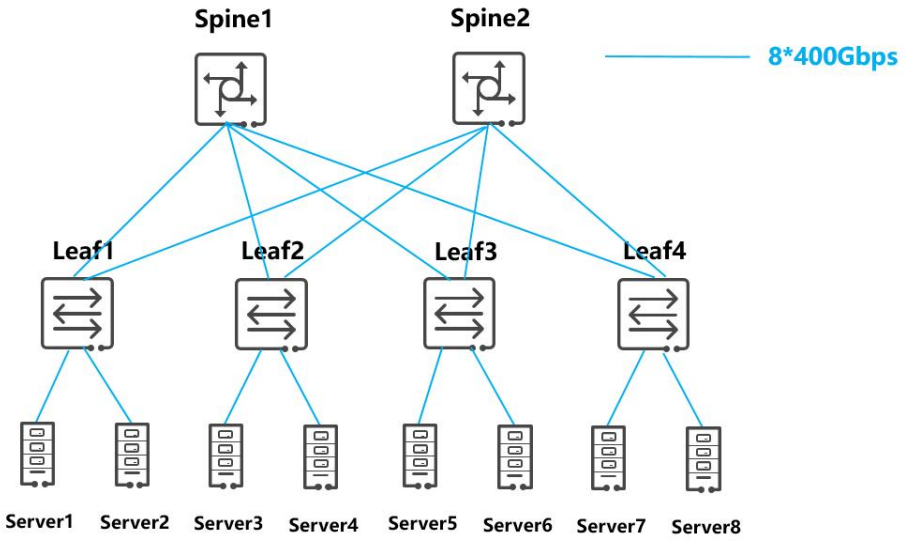
测试编号	
测试目的	测试启动两个 LLM 模型同时训练和启动三个 LLM 模型同时训练的训练性能。
组网拓扑	
预置条件	1. 网集中控制平台各系统部件运行正常。 2. 智算数据中心网络已纳管，且通过自检。 3. 服务器上已完成 Mixtral 8*7B、llama2 13B 模型的部署。
测试步骤	1. 测试 24 服务器 192 卡规模下两个大模型训练任务的性能，其中 128 卡运行 Mixtral 8*7B 大模型的训练任务，另外 64 卡运行 llama2 13B 大模型的训练任务，节点 rank 选取如下： 第1次资源组合：Mixtral 8*7B{0,5,1,9,2,13,3,4,15,6,19,7,10,8,11,17}；llama2 13B{12,20,16,21,14,22,18,23} 第2次资源组合：Mixtral 8*7B{0,4,1,8,2,6,3,9,13,5,17,11,15,7,19,10}；llama2 13B{12,20,16,21,14,22,18,23} 第3次资源组合：Mixtral 8*7B{0,4,1,8,2,6,3,9,13,5,17,11,15,7,19,10}；llama2 13B{20,21,12,18,16,14,22,23} 第4次资源组合：Mixtral 8*7B{13,0,4,8,15,1,5,9,17,2,6,10,19,3,7,11}；llama2 13B{20,21,12,18,16,14,22,23} 第5次资源组合：Mixtral 8*7B{13,0,4,8,15,1,5,9,17,2,6,10,19,3,7,11}；llama2 13B{12,20,16,21,14,22,18,23} 有预期结果1；

	2. 切换训练模型为 llama2 13B，测试三组 llama2 13B 大模型训练的性能，节点 rand 选取如下： 第1次资源组合：llama2 13B{4,0,8,1,6,2,9,3}；llama2 13B{5,16,13,17,7,18,15,19}；llama2 13B{10,20,12,21,11,22,14,23} 第2次资源组合：llama2 13B{0,1,4,9,8,6,2,3}；llama2 13B{16,17,5,15,13,7,18,19}；llama2 13B{20,21,10,14,12,11,22,23} 第3次资源组合：llama2 13B{4,0,8,1,6,2,9,3}；llama2 13B{16,17,5,15,13,7,18,19}；llama2 13B{20,21,10,14,12,11,22,23} 第4次资源组合：llama2 13B{4,0,8,1,6,2,9,3}；llama2 13B{5,16,13,17,7,18,15,19}；llama2 13B{20,21,10,14,12,11,22,23} 第5次资源组合：llama2 13B{0,1,4,9,8,6,2,3}；llama2 13B{5,16,13,17,7,18,15,19}；llama2 13B{10,20,12,21,11,22,14,23} 有预期结果2。
预期结果	1. 记录Mixtral 8*7B + llama2 13B两个大模型训练任务的平均带宽（GBps）； 2. 记录三个llama2 13B大模型训练的平均带宽（GBps）。
测试结果	
备注	1. 资源组合编排中，0 代表 Server1，1 代表 Server2，31 代表 Server32，以此类推。 2. Mixtral 8*7B 专家模型训练基线： https://gitee.com/ascend/MindSpeed-LLM/blob/master/examples/mcore/mixtral/pretrain_mixtral_8x7b_ptd.sh 3. 为了匹配 256 卡的测试床规模，避免训练时间过长，Mixtral 8*7B 专家模型需变更部分默认参数，分别为： TP=8 PP=1 EP=4 CP=4 NUM_LAYERS=16 --global-batch-size 8 \ 4. llama2 13B 模型训练基线： https://gitee.com/ascend/MindSpeed-LLM/blob/1.0.0/examples/mcore/llama2/pretrain_llama2_13b_ptd.sh 5. llama2 13B 训练参数调整： 256 卡 global-batch-size 调整为 256；64 卡 global-batch-size 调整为 128。

9 英伟达算力训练任务网络性能测试

9.1 单训练任务性能测试

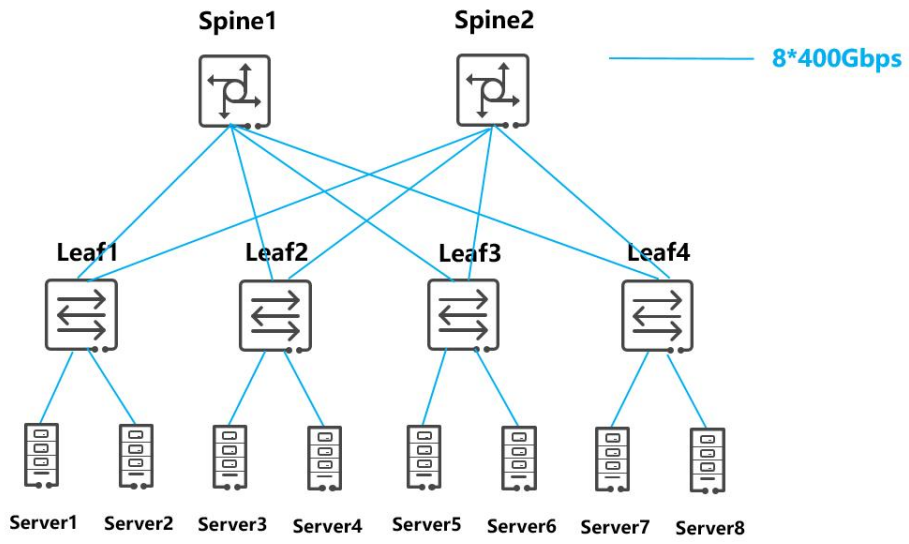
9.1.1 AllReduce-Ring 测试

<table>
<tr><td>测试编号</td><td></td></tr>
<tr><td>测试目的</td><td>测试一个 AllReduce-Ring 工作负载的平均带宽。</td></tr>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 集中控制平台各部件运行正常。
. 智算数据中心网络已纳管，且通过自检。</td></tr>
<tr><td>测试步骤</td><td>. 使用NCCL测试跨64卡节点下AllReduce-Ring通信模型的性能；
. 参考下表的资源组合，字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每组测试三次，记录预期结果1；
<table>
<tr><th>测试场景</th><th>资源组合</th></tr>
<tr><td>场景1</td><td>0 1 2 3 4 5 6 7</td></tr>
<tr><td>场景2</td><td>0 2 4 6 1 3 5 7</td></tr>
<tr><td>场景3</td><td>0 1 2 5 3 4 6 7</td></tr>
<tr><td>场景4</td><td>0 1 2 4 5 3</td></tr>
<tr><td>场景5</td><td>0 2 1 3 5 4</td></tr>
</table>
. 计算测试结果，有预期结果2。</td></tr>
<tr><td>预期结果</td><td>. 收集测试结果，并记录到结果汇总表；
<table>
<tr><th>资源组合
字节遍历</th><th>场景 1</th><th>场景 2</th><th>场景 3</th><th>场景 4</th><th>场景 5</th></tr>
<tr><td>第一</td><td></td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

	次						
	第二次						
	第三次						
	. 基于表计算NCCL AllReduce-Ring每个字节遍历场景下的平均带宽（GBps）。						
测试结果							
备注							

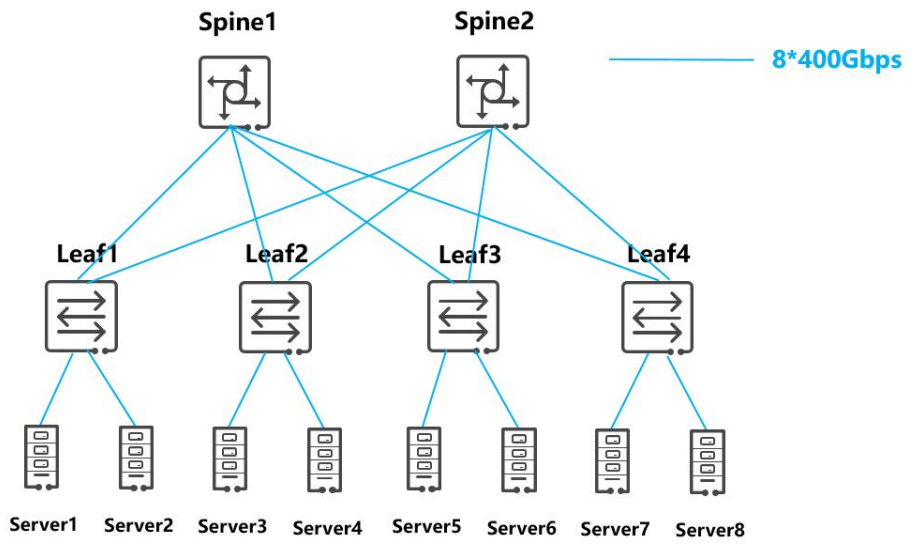
9.1.2 All to All 测试

测试编号	
测试目的	配置一个 All to All ~~-Pairwise~~ 工作负载的平均带宽。

<table>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 集中控制平台各部件运行正常。
. 智算数据中心网络已纳管，且通过自检。</td></tr>
<tr><td>测试步骤</td><td>. 使用NCCL测试跨64卡节点下All to All通信模型的性能；
. 设置字节遍历128MB、256MB、512MB、1GB、2GB、4GB，每字节数测试3次，记录预期结果1；
. 计算测试结果，有预期结果2。</td></tr>
<tr><td>预期结果</td><td>. 收集测试结果，并记录下表；
<table>
<tr><td>资源组合
字节遍历</td><td>第一次</td><td>第二次</td><td>第三次</td><td>平均</td></tr>
<tr><td>128MB</td><td></td><td></td><td></td><td></td></tr>
<tr><td>256MB</td><td></td><td></td><td></td><td></td></tr>
<tr><td>512MB</td><td></td><td></td><td></td><td></td></tr>
<tr><td>1GB</td><td></td><td></td><td></td><td></td></tr>
<tr><td>2GB</td><td></td><td></td><td></td><td></td></tr>
<tr><td>4GB</td><td></td><td></td><td></td><td></td></tr>
</table>
. 基于表2计算All To All每个字节遍历场景下的平均带宽（GBps）；</td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td></td></tr>
</table>

9.1.3 LLM 端到端性能测试

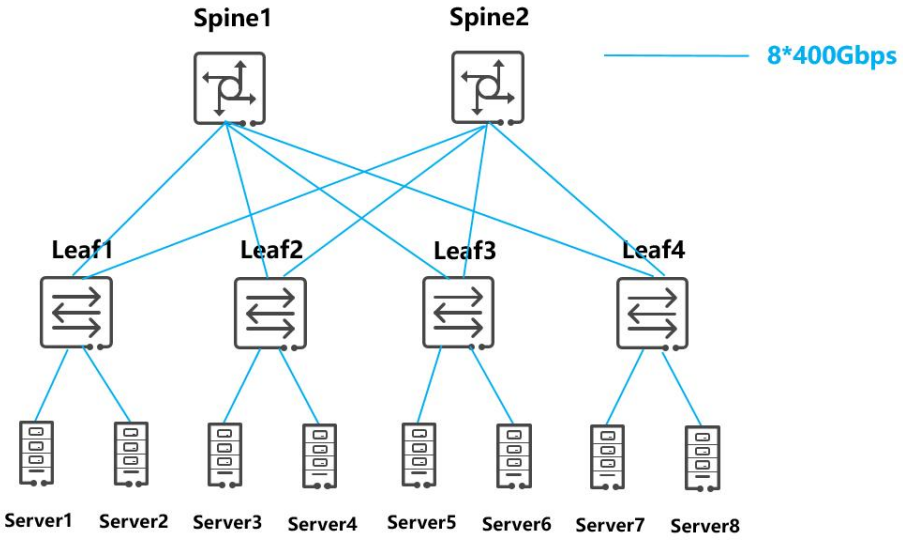
测试编号	

<table>
<tr><td>测试目的</td><td>测试 AI 大模型训练任务的性能</td></tr>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 集中控制平台各部件运行正常。
. 智算数据中心网络已纳管，且通过自检。
. 服务器上已完成llama2 13B模型的部署。</td></tr>
<tr><td>测试步骤</td><td>. 测试64卡规模无损网络大模型的训练性能；
. 训练模型为llama2 13B，节点Rank选取参考下表，逐个场景中按照llama2 13B模型训练基线启动测试，有预期结果1。
<table>
<tr><th>测试场景</th><th>资源组合</th></tr>
<tr><td>场景1</td><td>0 1 2 5 3 4 6 7</td></tr>
<tr><td>场景2</td><td>0 2 4 6 1 3 5 7</td></tr>
<tr><td>场景3</td><td>0 1 2 4 5 3</td></tr>
<tr><td>场景4</td><td>0 2 1 3 5 4</td></tr>
</table></td></tr>
<tr><td>预期结果</td><td>. 在下表中记录llama2 13B模型训练任务的平均带宽和平均迭代时长。
<table>
<tr><th>测试场景</th><th>平均带宽</th><th>平均迭代时长</th></tr>
<tr><td>场景1</td><td></td><td></td></tr>
<tr><td>场景2</td><td></td><td></td></tr>
<tr><td>场景3</td><td></td><td></td></tr>
<tr><td>场景4</td><td></td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td>. 资源组合编排中，0 代表 Server1，1 代表 Server2，7 代表 Server8，以此类推。
. llama2 13B 模型训练基线：</td></tr>
</table>

	https://huggingface.co/meta-llama/Llama-2-13b/tree/main . llama2 13B 训练参数调整： 256 卡 global-batch-size 调整为 256；64 卡 global-batch-size 调整为 128。

9.2 多训练任务性能测试

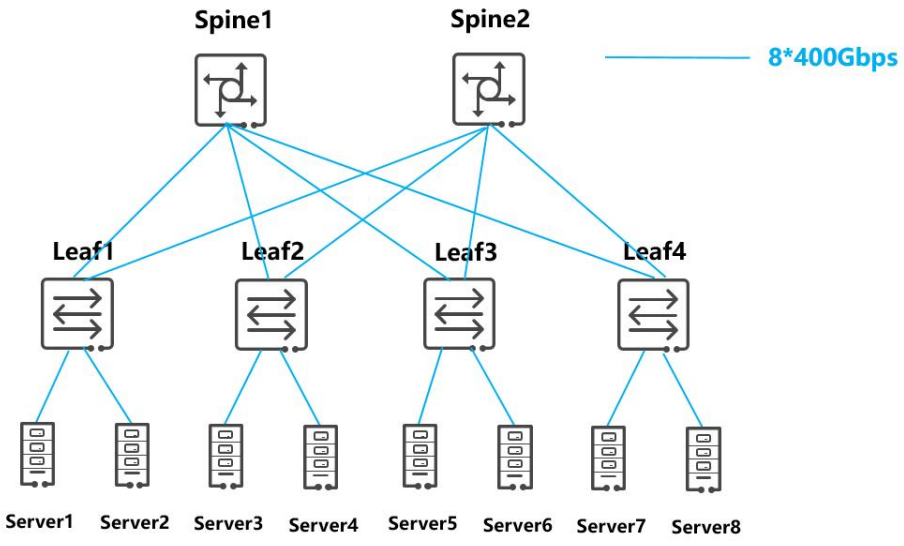
9.2.1 AllReduce-Ring 性能隔离测试

测试编号	
测试目的	分别测试两组工作负载和三组工作负载做 NCCL AllReduce 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响。
组网拓扑	
预置条件	. 集中控制平台各部件运行正常。 . 智算数据中心网络已纳管，且通过自检。
测试步骤	. 使用 NCCL 同时测试两个工作负载的 AllReduce 通信模型的性能，其中负载 1 和负载 2 均使用 4 个节点，均采用测试 AllReduce-ring 算子，节点 rank 选取如下： 场景 1 资源组合： 负载 1{0 1 2 3 }；负载 2{4 5 6 7 } 场景 2 资源组合： 负载 1{0 1 2 4 }；负载 2{3 5 6 7 } 场景 3 资源组合： 负载 1{0 2 4 6 }；负载 2{1 3 5 7 } 测试字节 4GB，分别测试 3 次，记录预期测试结果 1； . 使用 NCCL 同时测试三个工作负载的 AllReduce 通信模型的性能，其中负载 1

<table>
<tr><td></td><td>占用 4 个计算节点，负载 2 和负载 3 占用 2 个计算节点，测试 AllReduce-ring 算子,节点 rank 选取如下：
场景 1 资源组合：
负载 1{0 1 2 3 }；负载 2{4 5 }；负载 3{6 7}
场景 2 资源组合：
负载 1{0 1 2 3 }；负载 2{4 6 }；负载 3{5 7}
场景 3 资源组合：
负载 1{0 2 4 6 }；负载 2{1 3 }；负载 3{5 7}
场景 4 资源组合：
负载 1{0 1 2 4 }；负载 2{3 6 }；负载 3{5 7}
场景 5 资源组合：
负载 1{0 1 2 4 }；负载 2{3 5 }；负载 3{6 7}
测试字节 4GB，分别测试 3 次，记录预期测试结果 2。</td></tr>
<tr><td>预期结果</td><td>. 收集并分析两组工作负载的 NCCL AllReduce 的平均带宽 (GBps)；
<table>
<tr><th colspan="2"></th><th>第一次</th><th>第二次</th><th>第三次</th><th>平均</th></tr>
<tr><td rowspan="2">场景 1</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">场景 2</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">场景 3</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
</table>
. 收集并分析三组工作负载的 NCCL AllReduce 的平均带宽 (GBps)。
<table>
<tr><th colspan="2"></th><th>第一次</th><th>第二次</th><th>第三次</th><th>平均</th></tr>
<tr><td rowspan="3">场景 1</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 3</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">场景 2</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

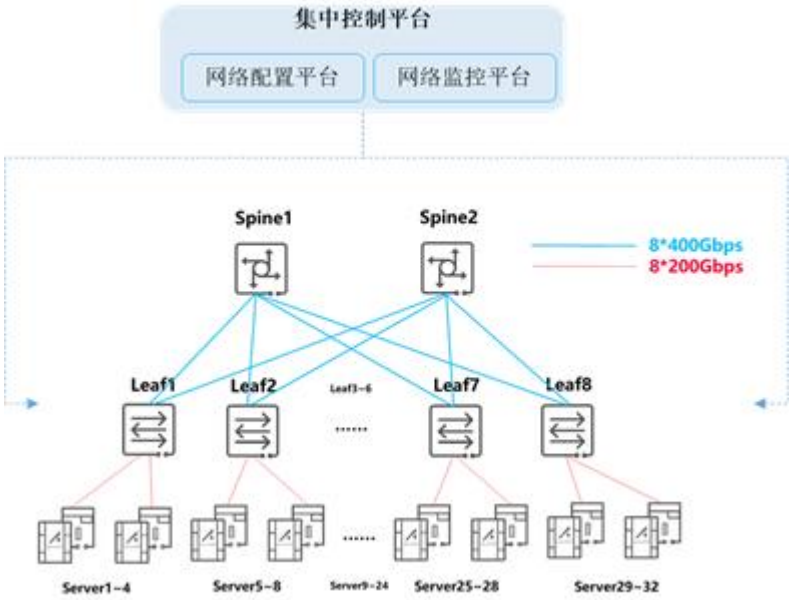
		负载 3				
	场景 3	负载 1				
		负载 2				
		负载 3				
	场景 4	负载 1				
		负载 2				
		负载 3				
	场景 5	负载 1				
		负载 2				
		负载 3				
测试结果						
备注						

9.2.2 All to All 性能隔离测试

测试编号	
测试目的	测试两组工作负载做 NCCL All To All 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响。
组网拓扑	
预置条件	. 集中控制平台各部件运行正常。 . 智算数据中心网络已纳管，且通过自检。
测试步骤	. 使用 NCCL 同时测试两个工作负载的 All To All 通信模型的性能，其中负载 1 和负载 2 各占用 4 个计算节点，测试 All To All，有预期结果 1，节点 rank 选取如

<table>
<tr><td></td><td>下：
场景 1 资源组合：负载 1{0 1 2 3 }；负载 2{4 5 6 7 }
场景 2 资源组合：负载 1{0 2 4 6 }；负载 2{1 3 5 7 }
测试字节 4GB，分别测试 3 次，记录预期测试结果 1。</td></tr>
<tr><td>预期结果</td><td>. 收集并分析两组工作负载的 NCCL All To All 的平均带宽 (GBps)；
<table>
<tr><td colspan="2"></td><td>第一次</td><td>第二次</td><td>第三次</td><td>平均</td></tr>
<tr><td rowspan="2">场景 1</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">场景 2</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td></td></tr>
</table>

9.2.3 AllReduce + All to All 性能隔离测试

<table>
<tr><td>测试编号</td><td></td></tr>
<tr><td>测试目的</td><td>分别测试两组工作负载和三组工作负载同时启动工作做 NCCL AllReduce 和 All To All 集合通信操作的平均带宽。观测该场景下不同组工作负载之间的影响。</td></tr>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 集中控制平台各部件运行正常。
. 智算数据中心网络已纳管，且通过自检。</td></tr>
</table>

<table>
<tr><td>测试步骤</td><td>. 使用 NCCL 同时测试两个工作负载的 AllReduce + All To All 通信模型的性能，其中负载 1 占用 4 个计算节点，采用 AllReduce-ring 算子，负载 2 占用 4 个计算节点，测试 All To All 节点 rank 选取如下：
场景 1 资源组合：负载 1{0 1 2 3 }；负载 2{4 5 6 7 }
场景 2 资源组合：负载 1{0 2 4 6 }；负载 2{1 3 5 7 }
测试字节 4GB，分别测试 3 次，记录预期测试结果 1；
. 使用 NCCL 同时测试三个工作负载的 AllReduce 通信模型的性能，其中负载 1 占用 4 个计算节点，测试 All To All 算子；负载 2 和负载 3 占用 2 个计算节点，测试 AllReduce-ring 算子,节点 rank 选取如下：
场景 1 资源组合：负载 1{0 1 2 3 }；负载 2{4 5 }；负载 3{6 7}
场景 2 资源组合：负载 1{0 1 2 3 }；负载 2{4 6 }；负载 3{5 7}
场景 3 资源组合：负载 1{0 2 4 6 }；负载 2{1 3 }；负载 3{5 7}
测试字节 4GB，分别测试 3 次，记录预期测试结果 2。</td></tr>
<tr><td>预期结果</td><td>. 收集并分析两组工作负载的 NCCL AllReduce + All To All 的平均带宽 (GBps)；
<table>
<tr><th colspan="2"></th><th>第一次</th><th>第二次</th><th>第三次</th><th>平均</th></tr>
<tr><td rowspan="2">场景 1</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="2">场景 2</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
</table>
. 收集并分析三组工作负载的 NCCL AllReduce + All To All 的平均带宽 (GBps)。
<table>
<tr><th colspan="2"></th><th>第一次</th><th>第二次</th><th>第三次</th><th>平均</th></tr>
<tr><td rowspan="3">场景 1</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 3</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">场景 2</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 3</td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="3">场景 3</td><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 3</td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
</table>

测试结果	
备注	

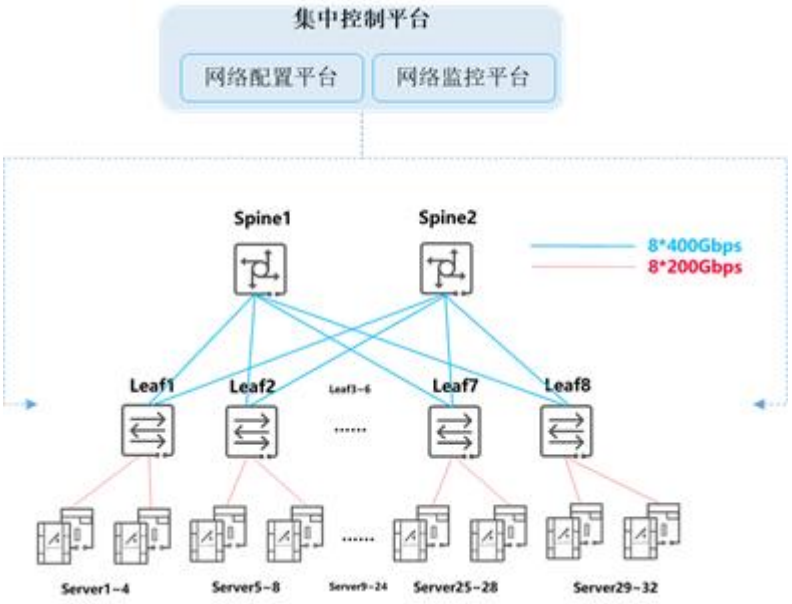
10 昇腾算力推理任务网络性能测试

随着模型参数量和序列长度增长，以及多模态的兴起，促使推理从单机形态走向多机互联，以满足推理对内存和算力的需求。大模型推理包含预填充（Prefill）和解码（Decode）两个阶段，分别负责对输入数据进行初步处理，生成初始的上下文表示，以及基于预填充阶段的结果，逐步生成输出结果。其中，Prefill 是计算密集型应用，Decode 是内存带宽密集型应用，在实际商用部署中推荐采用 PD 分离方式，即将 Prefill 和 Decode 分别部署在不同的推理池，PD 间互不干扰，可以提高算力资源的利用率，并提升并发业务规模。

随着 Deepseek 推理模型的发布，混合专家结构（Mixture of Experts，MOE）成为主流。与传统稠密模型以总线 TP 和大网 DP 为主不同，MoE 稀疏模型模式以大规模随机 EP 通信为主，这带来流量模型的改变。EP 变化快，通信模式不固定，随机通信关系微妙级变化，每一层都不相同。为满足业务对推理时延的需求，智能计算数据中心网络需要满足推理对负载均衡提出的更高要求。

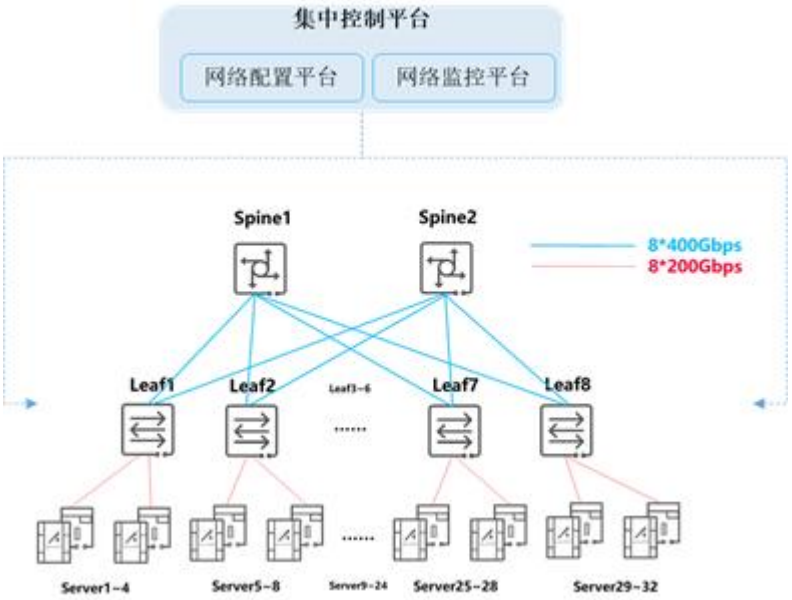
基于昇腾采用 HCCL 方式测试通信库，针对 Prefill 的多任务及训推一体进行吞吐性能测试。

10.1 HCCL All to All 多推理任务性能测试

测试编号	
测试目的	基于 All-to-ALL 算子，分别测试 4 组负载在默认算法和 Pipeline 算法场景下，各工作负载之间的影响。
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已纳管，且通过自检。
测试步骤	使用 HCCL 测试 4 个工作负载在 All To All 算子下的性能。4 个工作负载的节点 rank 选取如下：

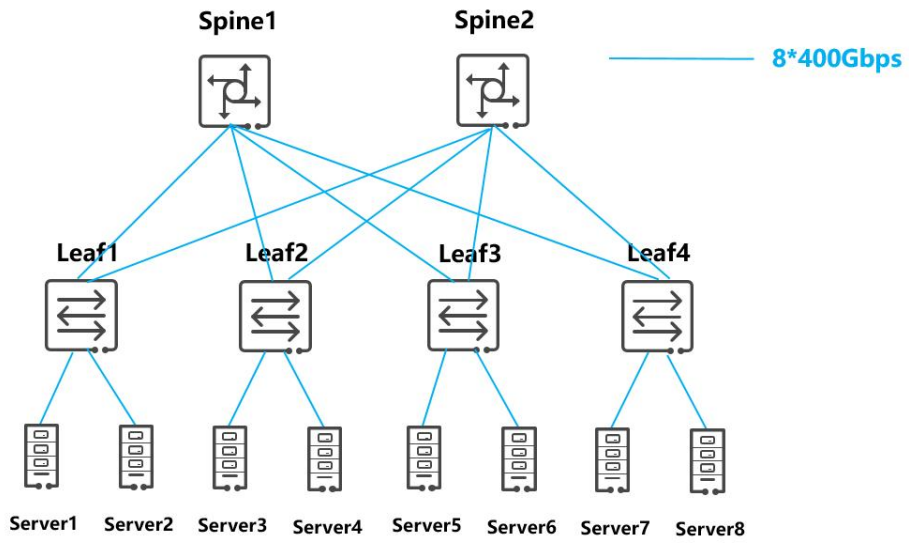
	负载 1{0,4 }；负载 2{1,5}；负载 3{2,6}；负载 2{3,7} 同时启动 4 个工作负载，分别测试 2 GB、4GB、8 GB、12 GB 字节数，每字节数分别测试 3 次，记录预期结果 1； 使用 HCCL 测试 4 个工作负载在 All To All 算子和 Pipeline 算法下的性能。4 个工作负载的节点 rank 选取如下： 负载 1{0,4 }；负载 2{1,5}；负载 3{2,6}；负载 2{3,7} 同时启动 4 个工作负载，分别测试 2 GB、4GB、8 GB、12 GB 字节数，每字节数分别测试 3 次，记录预期结果 2。
预期结果	收集并分析 4 组工作负载各自在 All To All +默认算法的算法带宽 (GBps)； 收集并分析 4 组工作负载各自在 All To All +Pipeline 算法的算法带宽 (GBps)；
测试结果	
备注	

10.2 HCCL All to Allv 多推理任务性能测试

测试编号	
测试目的	基于 All to Allv 算子，分别测试 4 组负载在默认算法和 Pipeline 算法场景下，各工作负载之间的影响。
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已纳管，且通过自检。
测试步骤	使用 HCCL 测试 4 个工作负载在 All to Allv 算子下的性能。4 个工作负载的节点 rank 选取如下： 负载 1{0,4 }；负载 2{1,5}；负载 3{2,6}；负载 2{3,7}

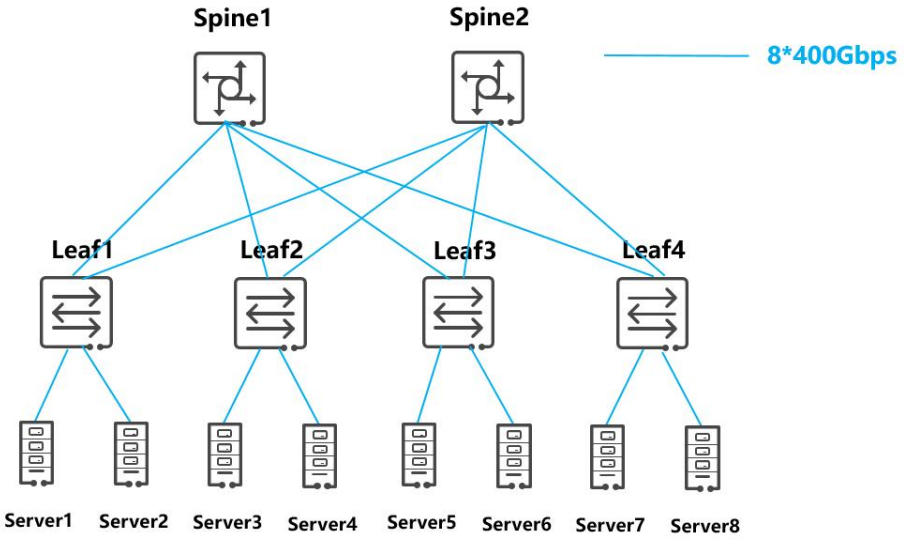
	同时启动 4 个工作负载，分别测试 2 GB、4GB、8 GB、12 GB 字节数，每字节数分别测试 3 次，记录预期结果 1； 使用 HCCL 测试 4 个工作负载在 All to Allv 算子和 Pipeline 算法下的性能。4 个工作负载的节点 rank 选取如下：负载 1{0,4 }；负载 2{1,5}；负载 3{2,6}；负载 2{3,7} 同时启动 4 个工作负载，分别测试 2 GB、4GB、8 GB、12 GB 字节数，每字节数分别测试 3 次，记录预期结果 2。
预期结果	. 收集并分析 4 组工作负载各自在 All to Allv 的算法带宽 (GBps)； . 收集并分析 4 组工作负载各自在 All to All v+Pipeline 算法的算法带宽 (GBps)；
测试结果	
备注	

10.3 基于大模型多推理任务性能测试

测试编号	
测试目的	测试基于大模型同时启动 4 组推理任务情况下的性能影响。
组网拓扑	
预置条件	. 集中控制平台各部件运行正常。 . 智算数据中心网络已纳管，且通过自检。
测试步骤	. 使用Deepseekv3同时启动4组推理负载。4组负载的节点rank选取如下： 负载1{0,8}；负载2{1,10}；负载3{2,12}；负载4{3,14} 推理负载均为8K token输入和1 token输出，测试3次，记录预期结果1。
预期结果	. 收集并分析4组推理负载各自的TTFT。
测试结果	

备注	

10.4 基于大模型训推一体性能测试（3P+训练）

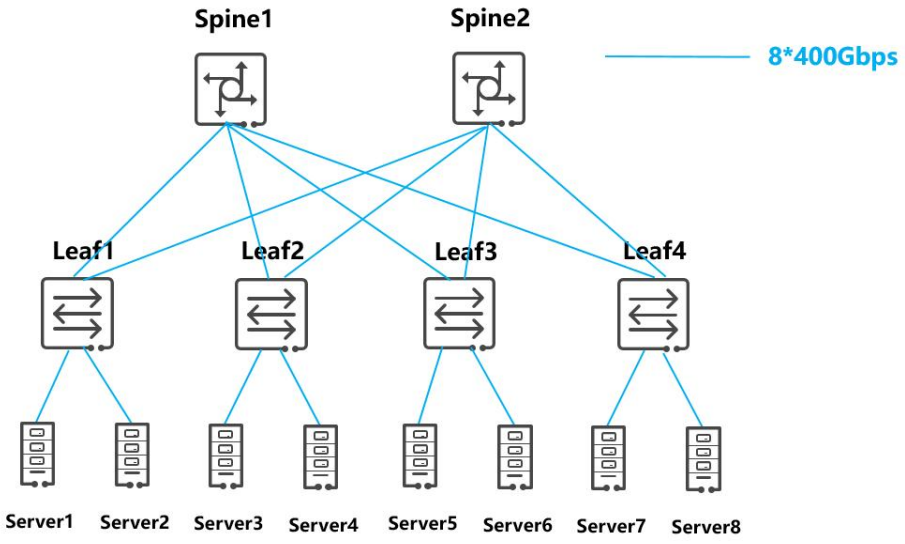
测试编号	
测试目的	基于模型同时启动 3 组推理负载和一组训练负载任务情况下的性能影响。
组网拓扑	
预置条件	. 集中控制平台各部件运行正常。 . 智算数据中心网络已纳管，且通过自检。
测试步骤	. 使用Deepseekv3同时启动3组推理负载，并通过HCCL启动一组训练负载。四组负载节点rank选取如下： 推理：负载1{0,8}；负载2{1,10}；负载3{2,12}； 训练：负载4{3,14} 推理负载均为8K token输入和1 token输出，训练负载使用HCCL 8GB AllReduce Ring，测试3次，记录预期结果1和结果2。
预期结果	. 收集并分析3组推理负载的TTFT； . 收集并分析1组训练负载的算法带宽 (GBps)；
测试结果	
备注	

11 英伟达算力推理任务网络性能测试

针对英伟达算力，DeepEP 通信库专为专家混合（MoE）和专家并行（EP）提供高吞吐量和低延迟的 all-to-all GPU 内核，也称为 MoE dispatch 和 combine。DeepEP 可以直接测试通信性能，包括推理 E2E 性能及 Prefill 和 Decode 不通阶段以及各阶段内 dispatch 和 combine 过程的性能。

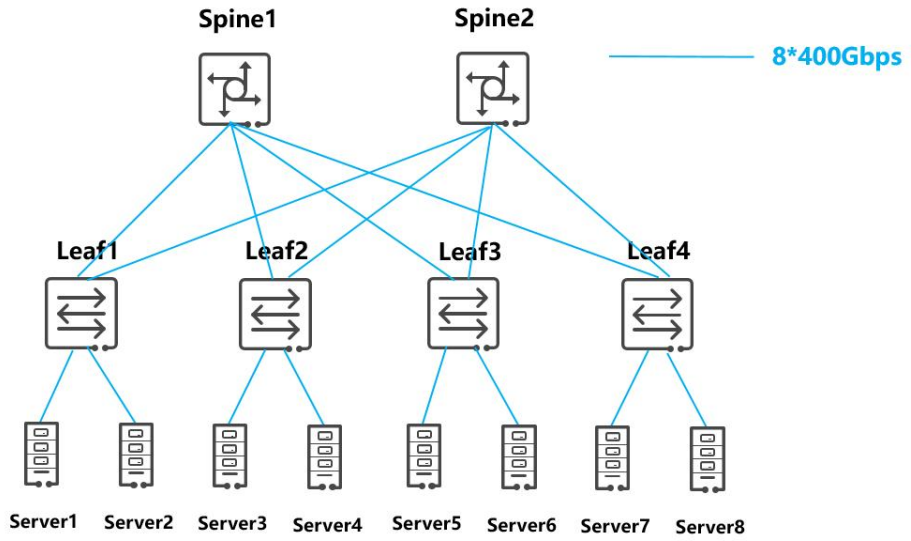
针对英伟达算力的推理性能测试，采用基于腾讯优化后的 DeepEP 通信库，分别针对 Prefill 和 Decode 的单任务、多任务、混合任务及训推一体进行吞吐性能测试。

11.1 单 Prefill 推理任务性能测试

<table>
<tr><td>测试编号</td><td></td></tr>
<tr><td>测试目的</td><td>测试整个集群为一个推理 Prefill 工作负载时的平均带宽。</td></tr>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 智算数据中心网络已纳管，且通过自检。
. 服务器上完成DeepEP的推理模型部署。</td></tr>
<tr><td>测试步骤</td><td>. 使用Prefill执行脚本，模拟执行Prefill阶段业务；
. 测试8服务器（64卡）Prefill阶段业务性能，记录dispatch(FP8)的带宽性能值和Combine性能值，有预期结果1。</td></tr>
<tr><td>预期结果</td><td>. 收集Prefill测试结果，并记录下表；
<table>
<tr><td></td><td>Dispatch （FP8）（RDMA）
(GBps)</td><td>Combine （BF16）（RDMA）
(GBps)</td></tr>
<tr><td>Prefill</td><td></td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td>1.DeepEP 推理模型基线：
https://github.com/deepseek-ai/DeepEP/tree/main
2.推理关键参数如下：
num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8
Prefill 的 batchsize=4096, decode 的 batchsize=128</td></tr>
</table>

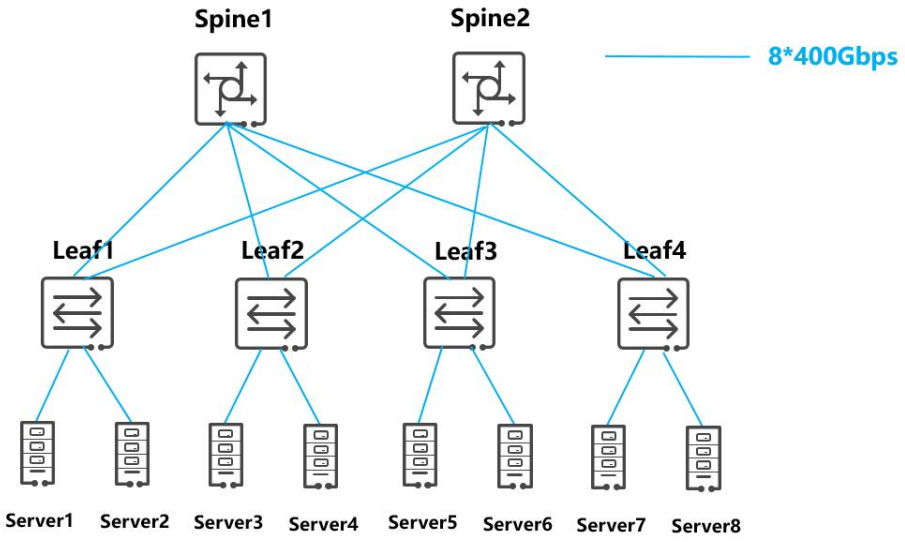
11.2 单 Decode 推理任务性能测试

测试编号	

<table>
<tr><td>测试目的</td><td>测试整个集群为一个推理 Decode 工作负载时的平均带宽。</td></tr>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 智算数据中心网络已纳管，且通过自检。
. 服务器上完成DeepEP的推理模型部署。</td></tr>
<tr><td>测试步骤</td><td>. 使用Decode执行脚本，模拟执行Decode阶段业务；
. 测试8服务器（64卡）Decode阶段业务性能，记录dispatch(FP8)的带宽性能值和Combine性能值，有预期结果1。</td></tr>
<tr><td>预期结果</td><td>. 收集Decode测试结果，并记录下表。
<table>
<tr><td></td><td>Dispatch
(GBps)</td><td>Combine
(GBps)</td></tr>
<tr><td>Decode</td><td></td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td>1.DeepEP 推理模型基线：
https://github.com/deepseek-ai/DeepEP/tree/main
2.推理关键参数如下：
num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8
Prefill 的 batchsize=4096, decode 的 batchsize=128</td></tr>
</table>

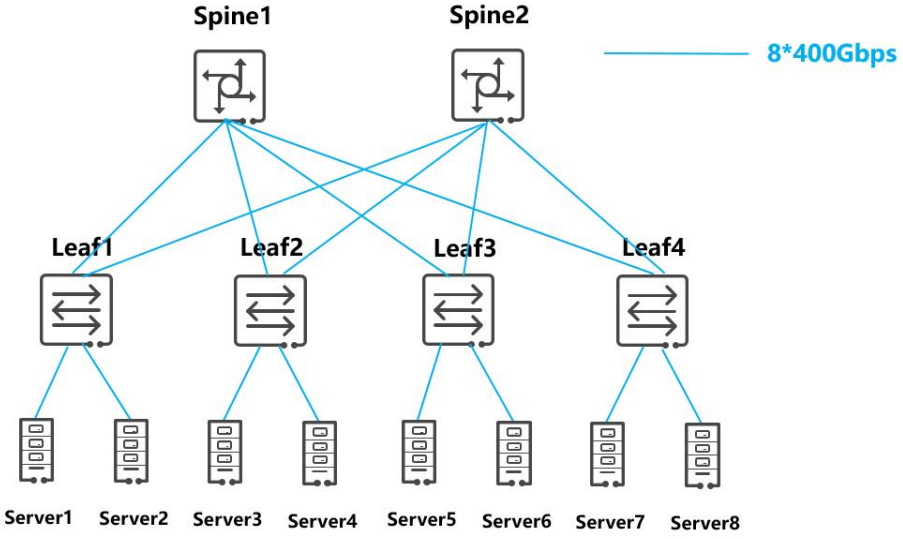
11.3 多 Prefill 推理任务性能测试

测试编号	
测试目的	测试整个集群为多个推理 Prefill 阶段工作负载时的平均带宽。

<table>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 智算数据中心网络已纳管，且通过自检。
. 服务器上完成DeepEP的推理模型部署。</td></tr>
<tr><td>测试步骤</td><td>. 使用Prefill执行脚本，同时测试两个负载，其中两个负载各占用4个计算节点，节点rank选取如下：
场景1：负载1{1 2 3 5}；负载2 {4 6 7 8}
记录dispatch(FP8)的带宽性能值和Combine性能值，记录为预期测试结果1；
. 使用Prefill执行脚本，同时测试四个工作负载的推理的性能，其中四个负载各占用2个计算节点，节点rank选取如下：
场景2：负载1{1 3}；负载2{2 7}；负载3{4 5}；负载4{6 8}
记录dispatch(FP8)的带宽性能值和Combine性能值，记录为预期测试结果2。</td></tr>
<tr><td>预期结果</td><td>. 收集测试结果，将场景1的Prefill的带宽性能值记录至下表；
. 收集测试结果，将场景2的Prefill的带宽性能值记录至下表。
<table>
<tr><td rowspan="2"></td><td colspan="2">场景 1</td><td colspan="2">场景 2</td></tr>
<tr><td>Dispatch (FP8) (RDMA) (GB/s)</td><td>Combine (BF16) (RDMA) (GB/s)</td><td>Dispatch (FP8) (RDMA) (GB/s)</td><td>Combine (BF16) (RDMA) (GB/s)</td></tr>
<tr><td>负载 1</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 2</td><td></td><td></td><td></td><td></td></tr>
<tr><td>负载 3</td><td>--</td><td>--</td><td></td><td></td></tr>
</table></td></tr>
</table>

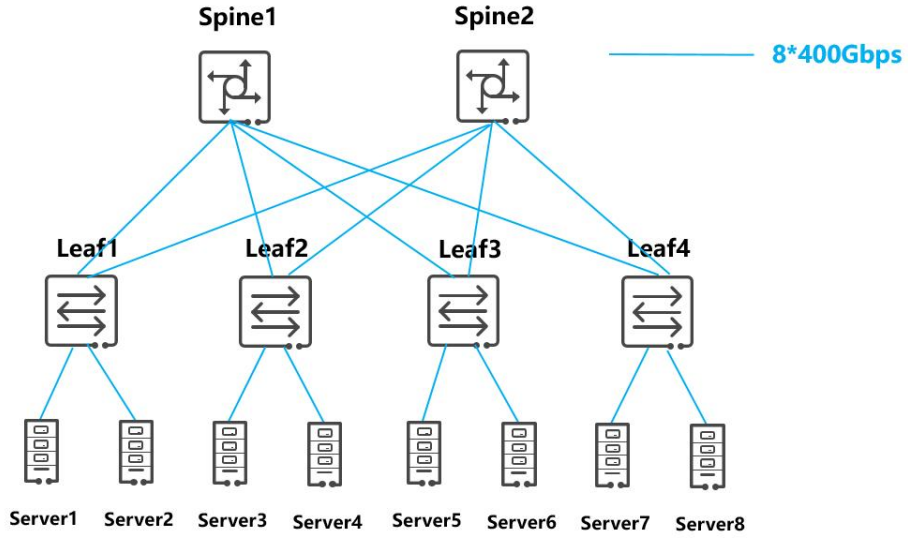
	负载 4	--	--		
测试结果					
备注	1.DeepEP 推理模型基线： https://github.com/deepseek-ai/DeepEP/tree/main 2.推理关键参数如下： num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8 Prefill 的 batchsize=4096, decode 的 batchsize=128				

11.4 多 Decode 推理任务竞争碎片化环境性能测试

测试编号	
测试目的	测试整个集群为多个推理 Decode 阶段时的平均带宽。
组网拓扑	
预置条件	. 采用NVIDIA合轨组网。 . 智算数据中心网络已纳管，且通过自检。 . 服务器上完成DeepEP的推理模型部署。
测试步骤	. 使用Decode执行脚本，同时测试两个负载，两个负载各占用4个计算节点，节点rank选取如下： 场景2：负载1 {1 2 3 5}；负载2 {4 6 7 8} 记录dispatch(FP8)的带宽性能值和Combine性能值，记录为预期测试结果1。
预期结果	. 收集测试结果，将Decode的带宽性能值记录至下表； Dispatch (GB/s)　Combine (GB/s)

	负载 1		
	负载 2		
测试结果			
备注	1.DeepEP 推理模型基线： https://github.com/deepseek-ai/DeepEP/tree/main 2.推理关键参数如下： num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8 Prefill 的 batchsize=4096, decode 的 batchsize=128 3. 4 负载场景下，单个负载分配算卡不足以支撑 Decode 运算，因此不涉及 4 负载的 Decode 测试。		

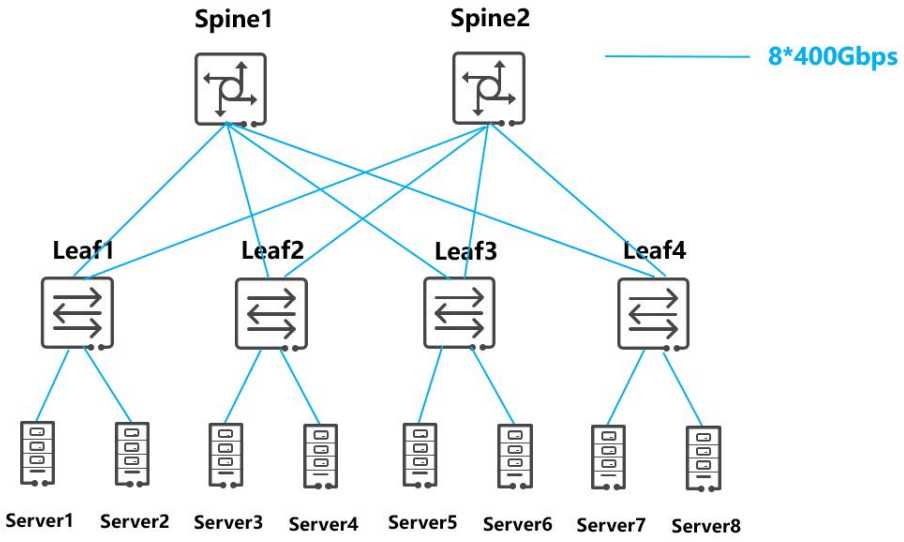
11.5 混合推理任务（Prefill+Decode）共存推理测试

测试编号	
测试目的	测试一个推理 Prefill 负载和一个 Decode 负载在一个集群中同时运行时的平均带宽。
组网拓扑	
预置条件	. 采用NVIDIA合轨组网。 . 智算数据中心网络已纳管，且通过自检。 . 服务器上完成DeepEp的推理模型部署。

<table>
<tr><td>测试步骤</td><td>使用Prefill和Decode执行脚本，分别测试两个工作负载的推理的性能，其中负载1占用5个计算节点，执行Decode脚本，负载2占用2个计算节点，执行Prefill脚本，节点rank选取如下：
场景1：负载1{1 2 3 5 7}；负载2 {4 6}；
记录dispatch(FP8)的带宽性能值和Combine性能值，记录为预期测试结果1。</td></tr>
<tr><td>预期结果</td><td>收集测试结果，将Prefill和Decode的带宽性能值记录至下表。
<table>
<tr><td rowspan="2"></td><td colspan="4">场景 1</td></tr>
<tr><td>Dispatch（FP8）
(RDMA)
(GB/s)</td><td>Combine
(BF16)
(RDMA)
(GB/s)</td><td>Dispatch
(GB/s)</td><td>Combine
(GB/s)</td></tr>
<tr><td>Prefill（负载2）</td><td></td><td></td><td>-</td><td>-</td></tr>
<tr><td>Decode（负载1）</td><td>-</td><td>-</td><td></td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td>1.DeepEP 推理模型基线：
https://github.com/deepseek-ai/DeepEP/tree/main
2.推理关键参数如下：
num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8
Prefill 的 batchsize=4096, decode 的 batchsize=128</td></tr>
</table>

11.6 训推一体性能测试

测试编号	
测试目的	测试整个集群为一个推理 Prefill 或 Decode 工作负载与训练业务（NCCL 模拟）共存，时的平均带宽。

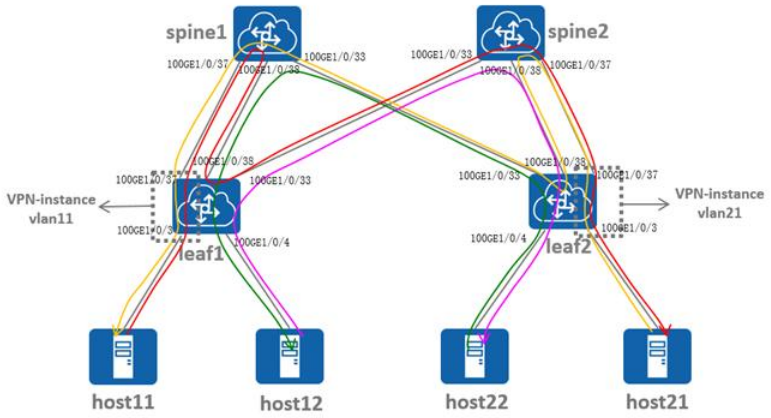
<table>
<tr><td>组网拓扑</td><td> 
</td></tr>
<tr><td>预置条件</td><td>. 采用NVIDIA合轨组网。
. 智算数据中心网络已纳管，且通过自检。
. 服务器上完成DeepEp的推理模型部署。</td></tr>
<tr><td>测试步骤</td><td>. 使用Prefill执行脚本（负载1）和NCCL-Test执行任务（负载2），同时测试两个工作负载的性能，其中负载1占用4个计算节点，负载2占用4个计算节点，使用all-reduce算子ring算法，message_size为4GB。节点Rank选取如下：
场景1：负载1{1 2 3 5}；负载2{4 6 7 8}
记录dispatch(FP8)的带宽性能值和Combine性能值，记录为预期测试结果1。
. 。NCCL-Test的总线带宽（busbw，单位 GB/s）,结果记录为预期测试结果2。</td></tr>
<tr><td>预期结果</td><td>. 收集测试结果，将Prefill的带宽性能值记录至下表；
. 收集测试结果，将NCCL-test的带宽性能值记录至下表。
<table>
<tr><td rowspan="2"></td><td colspan="3">场景 1</td></tr>
<tr><td>Dispatch （FP8）（RDMA）
(GB/s)</td><td>Combine (BF16)
(RDMA)
(GB/s)</td><td>NCCL(GB/s)</td></tr>
<tr><td>负载 1</td><td></td><td></td><td>--</td></tr>
<tr><td>负载 2</td><td>--</td><td>--</td><td></td></tr>
</table></td></tr>
<tr><td>测试结果</td><td></td></tr>
<tr><td>备注</td><td>1.DeepEP 推理模型基线：
https://github.com/deepseek-ai/DeepEP/tree/main
2.推理关键参数如下：
num_tokens=4096, hidden=7168, num_topk_groups=4, num_topk=8</td></tr>
</table>

12 故障主动预防能力测试

12.1 PFC 死锁预防

智算数据中心采用 RoCEv2，此时远程直接存储器访问（Remote Direct Memory Access, RDMA）开始以线速发送，这使得 incast 等常见问题比 TCP/IP 更严重。网络通过 PFC 机制进行流量控制以避免丢包。但 PFC 机制有其缺陷，当设备 PFC 检测到数据包丢失的风险时，会暂停所有该网络节点上游接口，并且该暂停将继续通过树状图向远端传播，即拥塞蔓延。这种拥塞蔓延可能会触发 PFC 死锁和 PFC 风暴，导致即使网络有空闲容量，也会使许多发送方保持沉默(停止发送)。

尽管 PFC 死锁和风暴的可能性非常小，但它们仍然对网络和应用程序构成威胁。PFC 死锁预防能力可用于减少网络出现 PFC 风暴的概率。

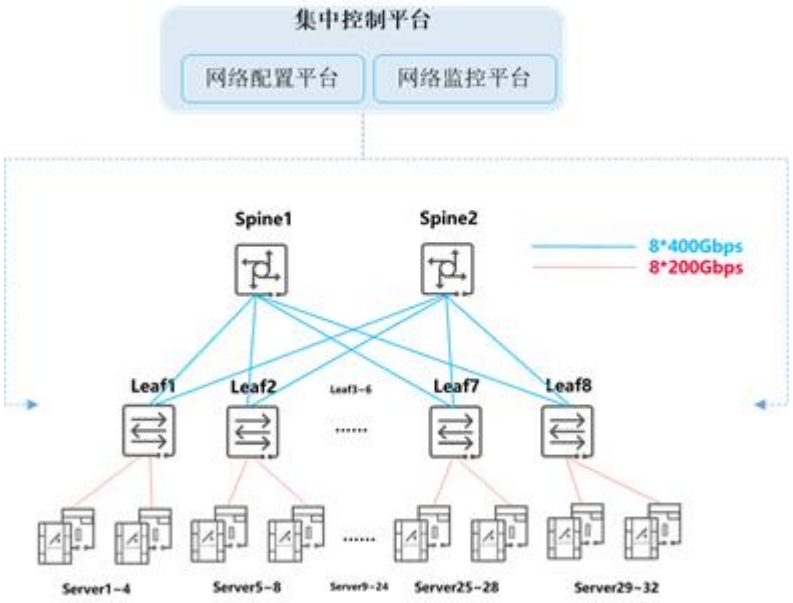
测试编号	
测试目的	验证交换机的 PFC 死锁预防能力
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。 2. 智算数据中心网络已部署，且已被纳管。 3. 服务器网卡配置 PFC 功能，关闭 ECN 4. 智算数据中心网络设备配置 PFC 功能，关闭 ECN
测试步骤	1. 四台主机按照组网拓扑打四条 DSCP 为 26 的流(ROCEv2)，流量走向如下（配置静态路由、策略路由等方式进行控制 ）： 流量 1：host21→leaf2→spine2→leaf2→spine1→leaf1→host11 流量 2： host11→leaf1→spine1→leaf1→spine2→leaf2→host21 流量 3： host22→leaf2→spine1→leaf1→host12 流量 4： host12→leaf1→spine2→leaf2→host22 2. 对 Leaf1 去往 Spine2 的上行流量限速，触发 PFC，构造死锁，有

	预期结果 1； 3. 停止所有流量，有结果 2； 4. 手动去使能 Leaf1 与 Spine2 互联端口的 PFC 功能，解除死锁； 5. 在设备上下发死锁预防配置，按步骤 1 重新打流量，有预期结果 3；
预期结果	1. 流量出现断流现象； 2. 观察 PFC 死锁仍然持续； 3. 重新打流后不再出现断流现象，流量转发正常。
测试结果	
备注	

12.2 光模块单激光器故障检测能力

光模块单激光器故障指的是有多个通道的光模块（如 QSFP-DD-SR8 有 8 个通道），当某个通道的激光器故障无法工作的情况。激光器故障高可用能力指的是多通道光模块单个或多个通道的激光器故障时，光链路可以运行在低速模式保证训练不中断.

光模块激光器故障模式在数据中心中光模块器件故障的占比约为 90%，且故障会造成断训。该功能可有效避免大模型断训。

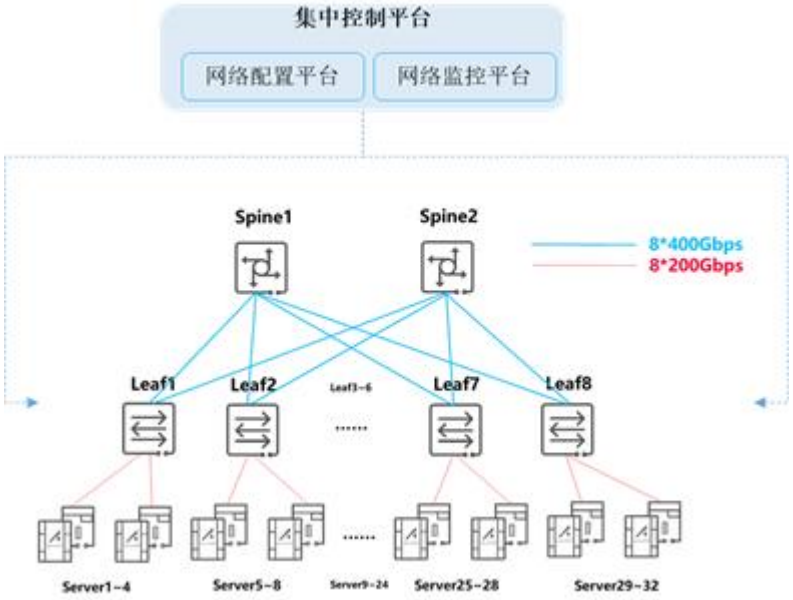
测试编号	
测试目的	验证光模块单激光器故障后网络不中断的能力
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。。 2. 智算数据中心网络已部署，且已被纳管。
测试步骤	1. 操作员选取任意两个交换机以及互联的 P1、P2 两个端口，通过光纤法兰适配器进行连接（光纤法兰适配器将光模块的光

	纤分割成多路，可以通过单独进行插拔来进行单激光器中断测试）。 2. 操作员在 P1、P2 两个端口间持续打满带宽流量； 3. 通过光纤法兰适配器进行单通道插拔，P1 和 P2 间网络接口单激光器出现故障，有预期结果 1； 4. 观察 P1 和 P2 连接带宽的变化，有预期结果 2。
预期结果	1. 触发光模块单通道故障后，网络监控平台有对应光模块的故障告警；但 P1 和 P2 端口不 down； 2. P1 和 P2 端口的带宽减半。
测试结果	
备注	

12.3 光模块脏污检测能力

大规模算力集群中需要使用到的光模块和光纤线缆越多，产生的光纤端面越多。光纤端面如果进灰或脏污容易导致链路闪断，从而影响集群持续训练或导致网络通信效率降低。

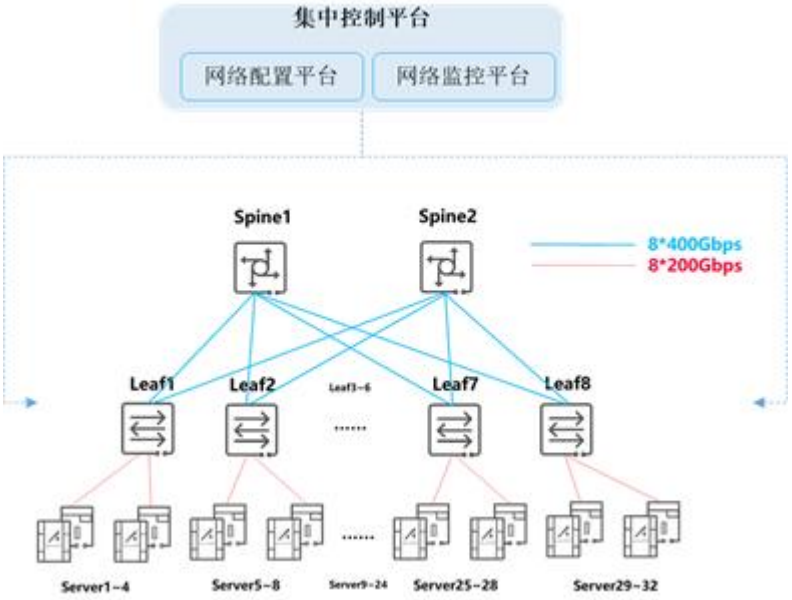
算力集群光模块脏污而导致的故障率很高，且检测手段有限。提供自动化光模块脏污检测可以有效提升网络可靠性。

测试编号	
测试目的	测试智算数据中心网络中光模块脏污的检测能力
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。 2. 智算数据中心网络已部署，且已被纳管。
测试步骤	1. 在 Server 和 Leaf 之间或 Leaf 和 Spine 之间构造光模块脏污故障，有预期结果 1。
预期结果	1. 网络监控平台可以识别光模块脏污异常，并提供修复建议。

测试结果	
备注	

12.4 光模块松动检测能力

光模块松动是导致链路闪断的主要原因之一。在集群经过长时间运行之后，光模块松动发生的概率也会上升。提供光模块松动的自动检测能力，有助于及时发现网络隐患，提升网络可靠性。

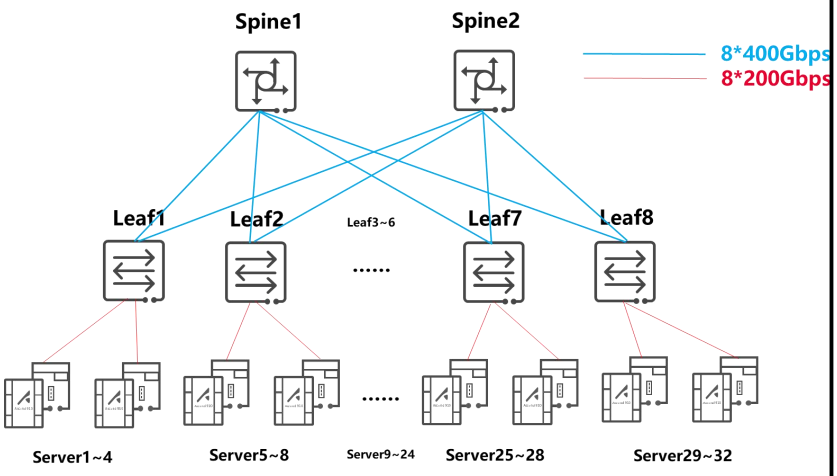
测试编号	
测试目的	测试智算数据中心网络中光模块松动的检测能力
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。 2. 智算数据中心网络已部署，且已被纳管。
测试步骤	1. 在 Server 和 Leaf 之间或 Leaf 和 Spine 之间构造光模块松动故障，有预期结果 1。
预期结果	1. 网络监控平台可以识别光模块松动异常，并提供修复建议。
测试结果	
备注	

13 高可用能力测试

AI 大模型训练周期长，如何降低 MTBF（平均无故障时间），是在使用中面临的最大挑战。提升智算数据中心网络的高可用能力至关重要。

13.1 PFC 风暴自动处理能力

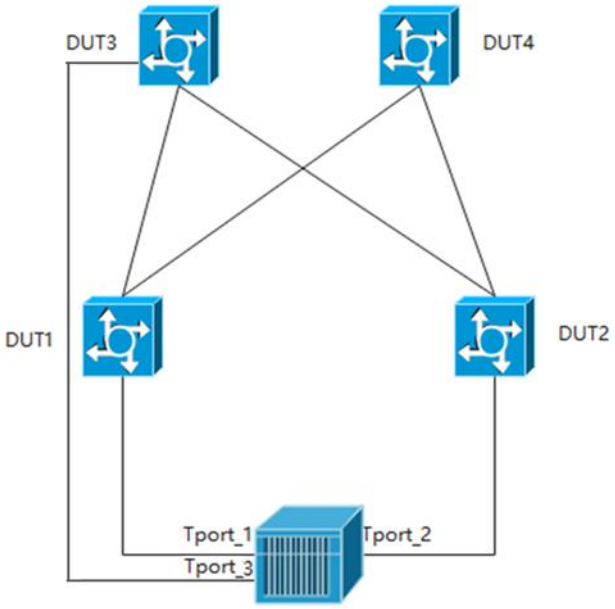
该能力是指设备检测到 PFC 风暴后的自处理能力，要求可以自动解除 PFC 死锁环路。
该技术可以破除 PFC 死锁环路，释放缓存依赖，解决 PFC 风暴类问题，减少训练任务中断。

测试编号	
测试目的	验证交换机 PFC 风暴自处理能力
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已部署，且已被纳管。 服务器网卡配置 PFC 功能，关闭 ECN。 智算数据中心网络设备配置 PFC 功能，关闭 ECN。
测试步骤	1. 启动大模型训练任务； 2. 使用测试仪打 PFC 帧，模拟 spine 和 leaf 之间出现 PFC 风暴场景，观察训练任务以及交换机状态，有预期结果 1； 3. 停止发送触发 PFC 风暴流量，等待 PFC 恢复时间后，观察交换机接口状态，有预期结果 2。
预期结果	1. 训练任务出现中断现象，且交换机上有死锁计数，当死锁频次达到设备阈值时接口的 PFC 自动去使能； 2. 交换机接口 PFC 功能恢复，重新启动训练任务，任务不中断。
测试结果	
备注	

13.2 网络路径故障自动切换

网络路径故障自动切换指的是在交换机检测到发生网络侧端口、链路、设备故障时自动路径切换的能力。该能力提升可大大降低模型训练中断概率。

测试编号	
测试目的	检验网络路径自动切换效率

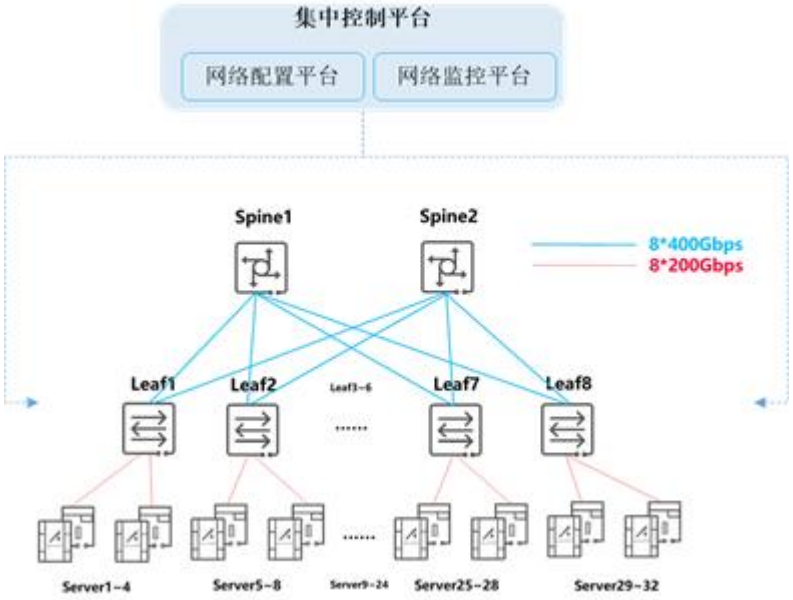
组网拓扑	
预置条件	. 完成上图测试环境的搭建。
测试步骤	1. DUT1、DUT2、DUT3 和 DUT4 之间的互联口配置 IP 地址，并通过 OSPF 协议建立邻居，Tport_1 网关在 DUT1 上，网关地址在 OSPF 中发布，Tport_2 网关在 DUT2 上，网关地址在 OSPF 中发布； 2. 测试仪 Tport_1 向 Tport_2 创建 UDP 流量 S1，UDP 流量的目的端口号变化 1000 次，报文目的 MAC 是 DUT1 设备与测试仪互联接口的 MAC，TTL 是 4，发包速率为端口带宽的 90%，测试仪上观察到流量转发无丢包,被测设备上观察到流量负载均匀，同时有 DUT1-->DUT3-->DUT2 以及 DUT1-->DUT4-->DUT2 的流量路径的流量，测试仪收到的流量的 TTL 为 1，测试仪 Tport_3 向 Tport_1 创建 IP 流量 S2，发包速率为端口带宽的 100%，流量 S2 正常转发无丢包； 3. 在 DUT2 设备上拔出 DUT2 与 DUT3 之间的光模块，有预期结果 1。
预期结果	1. DUT3 设备 S1 流量快速切换至 DUT4 上，DUT3 和 DUT1 之间的连线的端口状态正常，仍然 up，流量 S2 正常转发无丢包，测试仪收到的报文 TTL 为 1，根据测试仪上的丢包计数计算收敛时间（丢包个数/发包速率*1000*2）ms。
测试结果	
备注	1.

13.3 网络故障感知定界

AI 集群的集合通信存在木桶短板效应，最慢的流会拖慢整个通信过程。在发现整个集群异常时，需

要一种高效可视的手段来界定是计算还是网络的问题。

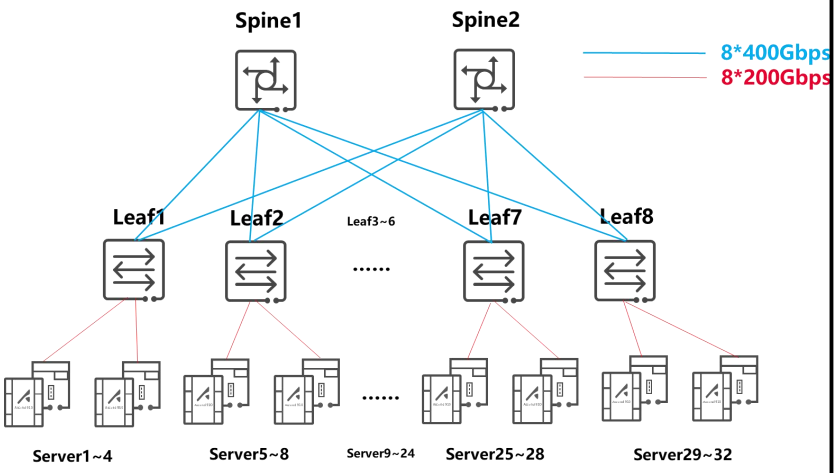
网络故障感知定界能力指的是网络集中控制器对网络设备上报的硬件、软件、转发丢包/错包等运行过程中出现的异常状态进行感知、分析和定界功能。快速进行故障定界可以有效缩短故障恢复时间。

测试编号	
测试目的	验证网络故障感知定界能力
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。 2. 智算数据中心网络已部署，且已被纳管。
测试步骤	1、导入/同步 AI 训练任务； 2、选择 AI 任务，有预期结果 1； 3、在其中一台设备上构造拥塞丢包，有预期结果 2。
预期结果	1、可呈现基于 AI 任务的多层网络拓扑，包括算力层、网络层、AI 任务层，并可以基于拓扑呈现 Issue/风险、指标统计和指标分析趋势图；指标包含丢包的设备个数、丢包率、时延、网络有效吞吐的变化趋势等； 2、支持会话级排障，并可以呈现异常会话详情信息，可呈现逐跳转发路径，并支持基于路径呈现每一跳节点上设备内和设备间入接口和出接口针对该会话的总包数、丢包数、平均时延、最大时延等网络指标，基于指标定界网络丢包和时延异常点
测试结果	
备注	

13.4 网络设备控制模块高可用能力

交换机控制面异常指的是交换机出现看门狗复位、驱动程序故障等异常情况。由于大模型训练中，异常复位问题需要在断训时间范围内修复，达成业务无损效果。

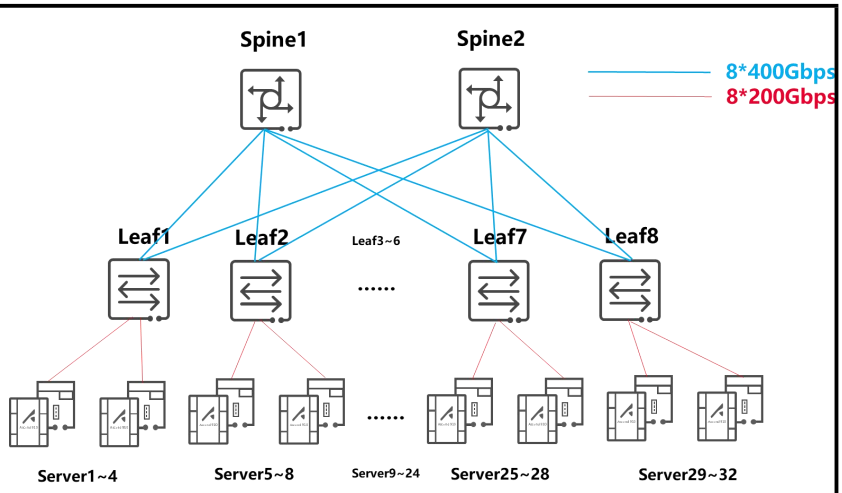
本用例采用快速重启方式进行控制模块故障定义。快速重启是一种 warm reboot 的重启方式，可有效缩短重启引起的流量中断时间。

测试编号	
测试目的	检验交换机控制模块高可用能力
组网拓扑	
预置条件	完成上图测试环境的搭建。
测试步骤	1. 在服务器上通过 XCCL 执行 AllReduce-HD； 2. 在训练任务结束前，选择组网中的一台设备进行快速重启（不涉及配置及版本升级），观察 AllReduce-HD，有预期结果 1。
预期结果	1. 设备升级 AllReduce-HD 任务不断训。
测试结果	
备注	

13.5 交换机升级高可用能力

大模型训练周期长，并要求不断训，通常没有升级和维护窗口。升级高可用能力指设备执行升级操作的时间要小于模型训练断训的时间，避免引起模型训练回退到 checkpoint 点重新训练。

测试编号	
测试目的	检验交换机升级不断训的高可用能力

组网拓扑	
预置条件	完成上图测试环境搭建
测试步骤	1. 在服务器上通过 XCCL 执行 AllReduce-HD； 2. 在训练任务结束前，选择组网中的一台设备加载新版本做升级操作，观察 AllReduce-HD，有预期结果 1。
预期结果	1. 设备升级 AllReduce-HD 任务不断训。
测试结果	
备注	

14 网络质量深度观测能力测试

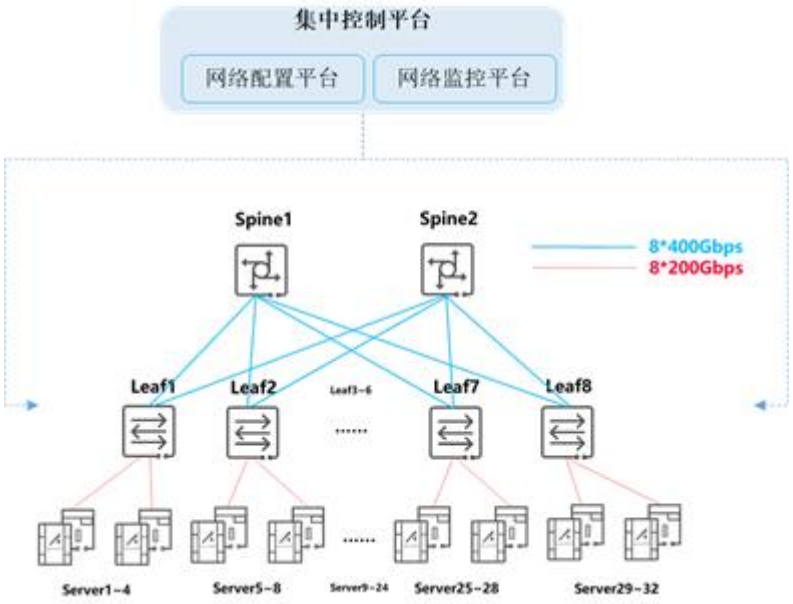
大规模数据中心内网络节点上的等价路由多达数百条，端到端经过多跳时，其路径组合更是几何级增加。某数据流出现丢包或者延时增加时，需要快速定位该数据流的具体路径、及该路径上的故障节点或链路。

14.1 AI 训练任务路况逐跳观测能力

AI 训练任务路况逐跳观测能力指是针对一个特定的 AI 训练任务执行算力卡对之间的路径和网络指标监控能力。该能力由网络侧实现训练任务所在训练算力卡的 IP 地址信息到接入网络的自动关联，并自动发现两算力卡之间的网络路径，对路径状态进行监控。通过该能力，可以在 AI 训练任务异常时针对路径上的网元和链路状态进行分析并快速进行故障定位和排障。

。

测试编号	验证基于 AI 训练任务的算力卡间路况逐跳观测能力
测试目的	

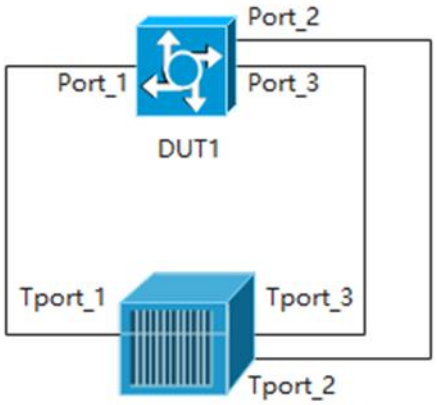
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已部署，且已被纳管。
测试步骤	在端侧启动 AI 训练任务； 在网络侧导入或同步 AI 训练任务，有预期结果 1； 在网络侧选定当前任务所在路径的两个节点，选择不同的接入 leaf 下的多个算力卡对同一个算力卡打流，构造多打一的拥塞丢包场景，经过一段时间后恢复，有预期结果 2； 在网络侧选定当前任务所在路径的两个节点，通过快速插拔光纤模拟链路闪断，有预期结果 3。
预期结果	支持基于 AI 训练任务的可视化能力，可支持展示指定任务最小网络拓扑，并支持指定两卡间的网络路径及路况展示，同时显示该任务的指标详情，如任务 FCT（流完成时间）、FET（流有效吞吐）、FNR（流重传率）的指标趋势； 可以在算力卡间的网络路径上呈现设备的拥塞异常，和拥塞产生的时间，并可对拥塞原因进行分析； 可以在算力卡间的网络路径上呈现闪断异常及闪断产生的时间。
测试结果	
备注	

14.2 丢包根因观测能力

该能力指的是实现设备节点丢包根因分析（如 ACL 丢弃，路由查询失败丢弃）及详细信息展示（如丢包位置信息、丢包的起始和结束时间、丢包的原因、丢包的流量和数量）。

由于 RDMA 对丢包敏感，丢包会引起网络性能的显著下降。该技术可以起到快速定位丢包位置和丢包原因的作用，缩短丢包故障的恢复时间。

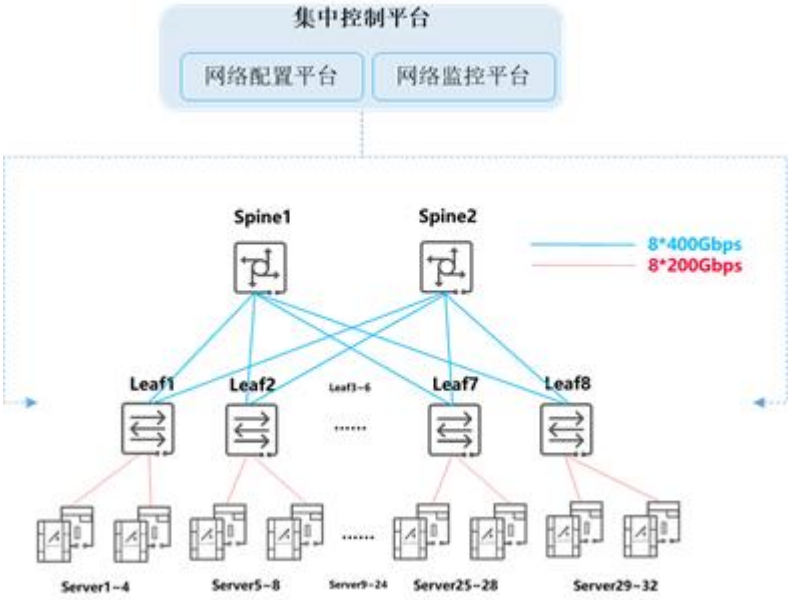
测试编号	

测试目的	验证设备转发丢包根因可视能力
组网拓扑	
预置条件	1. 完成上图测试环境的搭建
测试步骤	1. 将 DUT1 设备上 Port_1~Port_3 接口配置为三层口并配置 IPv4 地址； 2. 设备上开启丢包可视化查询功能，并配置将丢包信息发给 Tport_3； 3. Tport_1 构造源 mac 为 0100-5e00-0011，目的 MAC 为 Port_1 的地址的 UDP 流量，报文长度为 128 字节，流量大小为满带宽，有预期结果 1； 4. Tport_1 向 Tport_2 发送 TTL=0 的 TCP 流量，目的 MAC 为 Port_1 的地址，报文长度为 128 字节，流量大小为满带宽，有预期结果 2；
预期结果	1. Tport_2 收不到流量，DUT1 上通过命令行能查看丢包流表信息，包括丢包流量的五元组信息（源、目的 IP&端口号）、流量入端口号、丢包原因（源 mac 异常丢包）、丢包数、时间戳等信息，并将信息上送到 Tport_3，在 Tport_3 抓包能查看到丢包五元组信息、丢包原因 id、丢包数，其中五元组信息、丢包数与测试仪上的信息一致； 2. Tport_2 收不到流量，DUT1 上能查看丢包流表信息，包括丢包流量的五元组信息（源、目的 IP&端口号）、流量入端口号、丢包原因（TTL 终结丢包）、丢包数、时间戳等信息，并将信息上送到 Tport_3，在 Tport_3 抓包能查看到丢包五元组信息、丢包原因 id、丢包数，其中五元组信息、丢包数与测试仪上的信息一致；
测试结果	
备注	

15 算网协同能力测试

15.1 算网协同网络流量调度能力

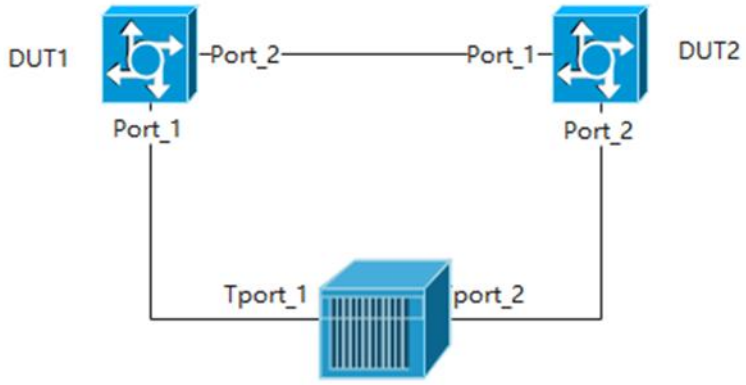
算网协同流量调度指的是网络侧 Leaf 设备接收端侧 GPU/NPU 的输入，且 Leaf 设备可上报控制器，由控制器根据 AI 训练任务的执行计划，对网络流量进行全局负载均衡调度。

测试编号	
测试目的	验证算网协同流量调度能力
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已部署，且已被纳管。
测试步骤	全集群开启 HCCL AllReduce-HD 训练任务，训练任务开始后，观察交换机上的配置，有预期结果 1。 查看控制器上任务信息，有预期结果 2。 HCCL AllReduce-HD 训练任务结束，全集群开启 HCCL AllReduce-NHR 训练任务，训练任务开始后，观察交换机上的配置，有预期结果 3。 查看控制器上任务信息，有预期结果 4。
预期结果	无需人为干预，集中控制平台下发了基于训练任务的负载均衡优化配置到交换机上； 控制器上可查看到任务信息，包含通信算法、运行时长、调度涉及的交换机以及配置下发信息、调度涉及的算力卡 IP 信息； 无需人为干预，控制器刷新了交换机上基于训练任务的负载均衡优化配置； 控制器上可查看到任务信息，包含通信算法、运行时长、调度涉及的交换机以及配置下发信息、调度涉及的算力卡 IP 信息。
测试结果	
备注	

16 网络安全能力测试

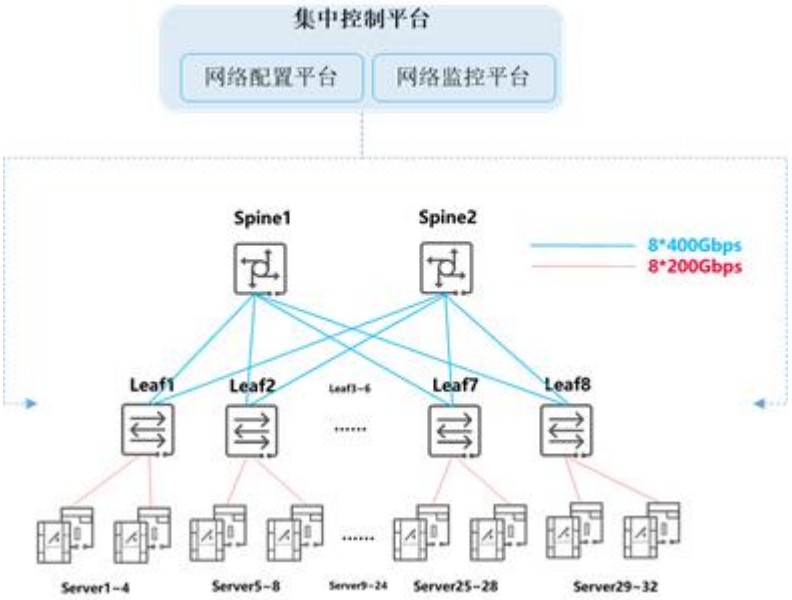
16.1 流量加密能力

该能力指的是在智算数据中心跨楼宇部署时，数据跨楼宇间的光纤存在恶意监听、仿冒、篡改的攻击暴露面,为保证算力数据安全，在交换机采用 MACsec 加密来高效进行数据加密保护。

测试编号	
测试目的	验证交换机支持流量加密能力
组网拓扑	
预置条件	集中控制平台各部件运行正常。 智算数据中心网络已部署，且已被纳管。
测试步骤	在 DUT1、DUT2 上创建 VLAN 10，在 DUT1 和 DUT2 的 Port_1 和 Port_2 上放通 VLAN 10； 在 DUT1、DUT2 上配置 MACsec 功能，配置数据加密和完整性校验，配置国密的 MKA 密钥生成算法、MACsec 数据报文加密算法，有预期结果 1； 测试仪 Tport_1 与 Tport_2 互发流量，发包速率为 20%接口速率，并在设备间抓包查看 MacSec 报文，有预期结果 2。
预期结果	密钥生成算法支持国密算法 SM4，数据报文加密算法支持国密算法 SM4，DUT1、DUT2 间 MACsec 会话协商成功； 测试仪上查看到流量正常转发无丢包，设备上可查看 MACsec 保护的数据报文统计计数，与测试仪的收发包计数一致，与设备间能抓到 MACsec 报文，确认报文已按照 MACsec 加密。
测试结果	
备注	

16.2 访问隔离能力

智算中心网络通常由多租户共享，应支持网络设备上配置租户隔离策略，包括但不限于 VXLAN、ACL 等技术，支持租户间流量隔离。

测试编号	
测试目的	验证交换机支持租户访问隔离
组网拓扑	
预置条件	1. 集中控制平台各部件运行正常。 2. 智算数据中心网络已部署，且已被纳管。
测试步骤	1. 将 Server1、Server2、Server5 规划为同一租户 A，Server3、Server6 为同一租户 B，Server4、Server7 为同一租户 C，在 SSpine 以及 Leaf 上通过 VLAN+ACL 的方式配置多租户隔离策略，有预期结果 1。
预期结果	1. 同一租户内的服务器间可以运行训练任务，不同租户之间的服务器无法互访。
测试结果	
备注	

参考文献

[1] IEEE 802.1Qbb Priority-based Flow Control

[2] RFC 3168 The Addition of Explicit Congestion Notification (ECN) to IP