

ZTE中兴



面向智算场景的 高性能网络白皮书



编制说明

本白皮书在编制过程中得到了多家单位的大力支持，联合编制单位如下（排名不分先后）：

中国联通研究院

中国信息通信研究院

2025 版权所有 中兴通讯股份有限公司 保留所有权利

版权声明：

本文档著作权由中兴通讯股份有限公司享有。文中涉及中兴通讯股份有限公司的专有信息，未经中兴通讯股份有限公司书面许可，任何单位和个人不得使用 and 泄漏该文档以及该文档包含的任何图片、表格、数据及其他信息。

本文档中的信息随着中兴通讯股份有限公司产品和技术的进步将不断更新，中兴通讯股份有限公司不再通

中兴通讯版权所有未经许可不得扩散

目录

1 前言	1
2 术语和缩略语	1
3 高性能网络关键需求和挑战	3
3.1 高性能数据中心网络(HP-DCN)	3
3.1.1 支持超大规模组网是基础	3
3.1.2 超高稳定性是前提	4
3.1.3 极致高性能是核心	5
3.1.4 多维自动化运维体系是必需.....	6
3.1.5 可规模扩展的安全机制是保障	7
3.2 高性能广域网（ HP-WAN ）	8
3.2.1 有效高吞吐量是焦点	8
3.2.2 公平共享高带宽是关键	8
4 高性能网络技术架构	9
4.1 高性能网络技术现状和趋势	9
4.2 中兴高性能网络技术架构.....	10
5 高性能数据中心网络关键技术	12

- 5.1 超大规模组网关键技术 12
 - 5.1.1 大规模组网交换机：硬件基础，容量速率双升 12
 - 5.1.2 大规模组网路由协议：可扩展快速部署，组播能力提供 13
- 5.2 超高稳定性关键技术 19
 - 5.2.1 故障无感恢复：硬件检测，多级保障 19

5.2.2 链路级可靠：轻量级 FEC，链路层重传	21
5.2.3 端网协同的路径控制：端侧传递需求 网络精准控制	22
5.2.4 网络隔离与资源保障：网络拓扑隔离，资源合理分配	23
5.3 极致高性能关键技术	24
5.3.1 层次化负载均衡：整网规划，局部调优，多粒度负载均衡	24
5.3.2 拥塞控制：算法无关，迅捷智能	25
5.3.3 集合通信卸载：统一编排，轻量传输	27
5.4 多维自动化运维关键技术：层次化可观测体系，高精度感知	29
5.5 可规模扩展安全机制关键技术：零信任模型，轻量级加密，安全会话无关	32
5.5.1 零信任安全架构	32
5.5.2 可扩展安全协议	33
6 高性能广域网关键技术	34
6.1 主动拥塞避免	34
6.2 近源端反馈机制	34
6.3 端网协同速率协商	34
6.4 任务式传输及配额调度	34
6.5 基于流的网络监控	35
7 展望	35

8 参考文献	37
--------------	----

1 前言

智算场景的普惠化正带来新一轮网络技术的革新浪潮。

随着生成式人工智能的发展，AI大模型参数量从GPT-3.5的1.75亿，到GPT-4的1.8万亿，预计未来GPT-5将达到十万亿参数规模，迅速膨胀的AI模型需要更大规模的算力集群执行训练。AI大模型以GPU集群分布式训练为基础，根据阿姆达定律，串行占比决定并行效率上限，网络成为影响算力的重要因素。AI训练任务的高精度并行协同特性以及超大集群互联吞吐量对网络性能提出了数量级的提升需求。AI大模型训练的时间往往长达数月，也使得网络的长稳运行变得前所未有的重要。从网络流量模型来看，AI大模型训练流量与通算流量呈现出完全不同的特征，突发的稀疏大流成为网络常态，聚合流量具备波峰波谷效应明显、周期性等特征，也使得很多通用数据中心的网络技术不再适用。HPC同样对网络提出高性能需求，特别是在可扩展性以及分布式资源的高效利用方面，HPC与AI的需求趋同。一般来说HPC对于时延更加敏感，但部分采用并行通信的计算模型，同样也关注长尾时延。

AI和HPC集群规模和服务范围的扩大对广域网传输也提出全新需求，包含数据协同和数据快递两大应用场景。数据协同应用主要面向AI/HPC的分布式协同，例如在跨DC的AI训练过程中的训前模型和数据上载，以及训练期间数据和状态同步过程；数据快递场景包括数据灾备、大规模科学数据传递等。以上都需要广域网具备高性能海量数据传输的能力。

综上，面对大规模AI/HPC的计算、存储和通信需求，不仅数据中心内部的大规模密集数据交换需要高性能网络的支撑，还需要网络能够高效地连接多个数据中心或站点，实现跨地域的AI/HPC业务的高效协同。

本白皮书从面向智算业务的高性能网络需求和技术挑战出发，分析高性能网络技术发展现状和趋势，并探索更适合行业协同发展的高性能网络技术架构和关键技术。

2 术语和缩略语

以下缩略语适用于本白皮书。

缩略语	英文全称	中文含义
AEAD	Authenticated Encryption with Additional Data	带有身份认证的加密算法
AES	Advanced Encryption Standard	高级加密标准

AI	Artificial Intelligence	人工智能
----	-------------------------	------

AIGC	Artificial Intelligence Generated Content	生成式人工智能
ARN	Adaptive Routing Notification	自适应路由通知
BBR	Bottleneck Bandwidth and Round-trip propagation time	瓶颈带宽和往返传播时间
BGP	Border Gateway Protocol	边界网关协议
BIER	Bit Indexed Explicit Replication	位索引显式复制
CCO	Collective Communication Offloading	集合通信卸载
CCOM	Collective Communication Offloading Manager	集合通信卸载管理
CNP	Congestion Notification Packet	拥塞通告报文
CPU	Central Processing Unit	中央处理器
CSIG	Congestion Signaling	拥塞信令
DSF	Distributed Scheduled Fabric	分布式全调度网络
DOM	Digital Optical Monitoring	数字光学监控
ECMP	Equal-Cost MultiPath routing	等价多路径路由
ECN	Explicit Congestion Notification	显式拥塞通知
ENCC	End-Network Cooperation Congestion Control	端网协同的拥塞控制
FEC	Forward Error Correction	前向纠错
GCM	Galois/Counter Mode	伽罗瓦/计数器模式
GPU	Graphics Processing Unit	图形处理器
HPC	High Performance Computing	高性能计算
HPCC++	Enhanced High Precision Congestion Control	增强的高精度拥塞控制
HP-WAN	High Performance Wide Area Network	高性能广域网
IFA	Inband Flow Analyzer	带内流分析器
IGLB	Intelligence Global Load Balance	全局负载均衡
IGP	Internal Gateway Protocol	内部网关协议
INT	Inband Network Telemetry	带内网络遥测
IOAM	In situ Operations, Administration, and Maintenance	带内 OAM

JCT	job completion time	任务完成时间
KDF	Key Derivation Function	密钥派生函数
MOD	Mirror On Drop	丢包镜像

MTU	Maximum Transmission Unit	最大传输单元
NCPC	Network-coordinated Path control	网络协同的路径控制
PIM	Protocol Independent Multicast	协议无关组播
RIFT	Routing In Fat Trees	胖树路由协议
RTT	Round-Trip Time	往返时延
ZTP	Zero Touch Provisioning	零接触配置

3 高性能网络关键需求和挑战

3.1 高性能数据中心网络 (HP-DCN)

3.1.1 支持超大规模组网是基础

在Scaling Law（扩展定律）的驱动下，万卡GPU训练集群已成为AIGC核心玩家的及格线，智算中心正迅速朝着超万卡级别的规模迅速发展，国内云商如阿里巴巴、百度等陆续宣布具备10万卡集群的支持能力，而Grok-3的训练集群已达到20万卡级别，预计会有更多10万卡+智算集群出现。

如此大的组网规模势必引发网络技术的质变，高性能网络架构的主要功能设计以及性能要求，都需要放到支持超大规模网络的框架下重新考量。

支持超大规模组网的主要挑战包括：

1) 交换机接入容量的限制

服务器GPU网卡数量和接口速率在逐渐增加，呈现每两年翻倍的趋势，当前规模商用的GPU服务器网卡接口达到8*400G，支持800G的GPU服务器也已推出，为了满足接入需求，减少设备数量，对单交换机容量提出了越来越高的需求，而单交换芯片的容量提升节奏，明显落后于IO总线的发展，并且存在物理上限。

2) 组网拓扑的限制

为了满足数十万卡乃至更大规模的组网需求，在交换机支持端口数短期无法跃升的情况下，传统的CLOS架构需要采用更多的网络层次，更多的网络层次意味着转发跳数增加，在带来更大时延的同时，更多跳数的路径也增大了故障发生的概率和定位难度，使得网络难以

运维；同时，各层级之间用于互联的端口数量剧增，若采用光纤连接，光模块部分的成本增加也不容忽视。

3) 异构网络的互通挑战

大规模网络的构建可能会涉及多厂商设备，当前大模型训练网络仍处于技术方案耦合度较高、标准不完备的状态，未来设备间互通可能面临挑战。

3.1.2 超高稳定性是前提

AI和HPC均是典型的分布式系统，网络作为分布式系统的连接底座，网络的故障或者性能波动会影响集群计算效率，因此网络自身的稳定性是整个系统稳定运行的基础。此外大规模训练或计算任务可能持续数周或数月时间，因此要求网络需具备长期持续的稳定性。

高性能网络的稳定性可采用如下两方面的指标衡量：

1) 网络可用性：即网络无故障运行的时间，该指标主要与故障间隔时间以及故障恢复时间相关；

2) 性能一致性：即在不同网络负荷状态下，网络性能指标具备稳定性与一致性。

高性能网络的稳定性主要面临如下挑战：

1) 网络设备故障感知能力不足，协议软件层面的故障感知时延在毫秒级，故障恢复效率低。

2) 在大规模网络中，错包和丢包无法避免，链路中如误码产生的随机丢包对于模型训练的影响巨大，但缺乏与高性能组网需求匹配的恢复机制。

由于对带宽的强劲需求，网络需要采用更高的端口速率，但误码率也随之增加，为了纠错而引入的高精度FEC机制又会带来更大的时延。例如，400G及以上速率以太网普遍采用PAM4调制代替NRZ，以在不新增光纤的基础上增加网络带宽，有效提升传输效率；为了应对PAM4带来的比特错误率更高的问题，400G标准中采用了纠错能力更强的RS(544,514) FEC技术，但是这种纠错技术带来了更高的传输时延。基于400G以太网，对比RS(544,514)与轻量级前向纠错RS(272,258)两种FEC技术的时延，前者每包传输的时延约是后者的1.8倍。

针对丢包，现有技术多采用端到端重传机制，对网络吞吐影响较大。已有研究表明，基于RoCEv2的Go-back-N丢包恢复模型，当丢包率达到1%时，RDMA报文吞吐量接近为0。

3) 不同负载情况下，网络性能指标波动，性能表现不一致。通用数据中心网络在轻载时一般都可以达到较高的性能，但AI训练是多任务集群，在多任务并存的情况下，同时满足各任务的高性能需求，对于网络资源的规划和保障提出了更高要求。

3.1.3 极致高性能是核心

为了最大化集群算力利用率，AI大模型训练通常采用并行处理机制，将一个任务分布在多个GPU上。并行训练整体分为三个步骤：处理、通知和同步。在处理阶段，每个GPU完成各自的任务部分；在通知和同步阶段，GPU进行卡间通信并汇总出整个任务的结果。整体的JCT由GPU返回计算结果的速度以及网络的同步速度决定。因此，网络性能是影响集合通信效率的重要因素，主要包括以下指标：

1) 超低时延：网络时延分为静态时延和动态时延，静态时延为交换机收发报文以及报文在线缆上转发的固有时延，而动态时延则包括排队时延以及拥塞、丢包引入的时延。在AI大模型训练中，集合通信的网络时延和业务吞吐性能呈现正相关，决定了训练加速比的上限，因此需要网络尽可能降低时延，目标在亚微秒级。

2) 极低抖动：网络抖动同样是影响JCT的重要因素，集合通信的流程一般可分解为计算节点之间的多次平行的点对点通信，存在木桶效应，每一步计算中，需要所有节点全部完成点对点通信后，才能进行下一步计算。网络抖动会引发点对点通信时延增加，直接影响集合通信的计算效率。

3) 有效高吞吐：大模型训练期间的数据通信体量巨大，可达几百TB级别。机间GPU高速互联，要求网络提供与机内IO总线性能相匹配的单点接入带宽以及网络整体带宽。然而，需要注意的是，对于网络的需求不仅是高带宽，最终目标是业务的有效高吞吐，有效吞吐率需尽可能接近满吞吐。

对于网络极致高性能的需求主要面临如下挑战：

1) 时延方面：动态时延一般比静态时延高出几个量级，是网络低时延的主要矛盾，由拥塞、丢包引起的时延往往在毫秒级，因而需要精度更高的流控/拥塞控制及故障恢复机制。

2) 抖动方面：现有拥塞控制技术如DCQCN等，主要面向丢包和吞吐，对于控制抖动的考虑较少，需要更精准的拥塞及流控机制。另一方面，针对大模型训练产生的大象流，网络负载均衡粒度过粗也会导致网络抖动难以管控。

3) 吞吐方面：高吞吐设计是一个复杂的系统工程，涉及机内外和软硬件的精细协同，单项指标的提升无法有效提高吞吐率，还可能导致此消彼长效应，丢包、时延、吞吐往往会相互影响。

3.1.4 多维自动化运维体系是必需

与通用数据中心网络相比，HP-DCN网络具有节点链路规模大、流量突发性强、集群系统复杂等特性，在网络运维领域，对运维体系提出了以下需求：

1) 广度——全面覆盖每个潜在故障点。AI大模型训练网络中的任何链路问题都可能产生广泛影响，鉴于GPU集群网络中链路数量和等价路径数量庞大，传统的Fullmesh监控方式难以覆盖所有链路。未监控链路上的故障会导致故障发现时间显著增加，定位难度也随之加大。

2) 精度——捕捉突发拥塞。AI大模型训练产生的突发流更为同步、粒度更小。传统的秒级或亚秒级监控无法满足对链路状态实时测量的需求。例如，在一秒内可能经历多次持续时间仅为毫秒至数十毫秒的吞吐量波动。在秒级尺度下观察时，链路似乎处于稳定低负载状态而无明显波动。这种表象与实际情况存在巨大差异，在逻辑上将导致不同的故障发现和定位结果。

3) 深度——节点级别的根因监控实现快速排障和自愈功能。对于承载AI大模型训练流量的ROCE高性能网络而言，一旦检测到异常情况，运营系统必须深入分析以确定根本原因。这包括多个数据源之间的关联分析以及因果关系推理。系统需追踪复杂的关系链并找出最初的触发点，在监控层面直接给出根本原因分析结果，进而加速高性能网络的故障排查和自我修复过程。

尽管通用数据中心网络已采取诸如性能监控、告警系统、电信级OAM等运维策略，并引入随流检测技术以增强运维能力，在面向AI高性能网络运维时仍面临如下挑战：

1) 业务视角受限：当前随流检测技术仅停留在网络域，并未延伸至业务端到端层面，从而限制了对业务质量的全面评估和故障定位能力。

2) 故障诊断深度不足：在网络故障检测方面存在局限性，尤其是在处理静默丢包和拥塞等复杂问题上缺乏准确判断其根本原因的能力。

3) 资源监控缺陷：队列和端口资源监控范围有限且精度不足；对于RoCE网络而言，在满足基础运维需求的同时还需要实现RoCE指标的可视化管理。

4) 流量监控精度低：现有流量监控采样机制精度不高，无法实现对流量的精准监控与全面覆盖；这直接影响了网络性能优化效果以及故障定位准确性。

针对AI大模型训练网络的特定需求与挑战，必须建立一个多维分层可观测运维体系：多维度包括实现监控、调优与排障三大核心维度能力；分层覆盖业务层、网络服务层以及基础

网络层。旨在确保从不同层面全面洞察系统状态，为三大核心维度能力提供关键数据支撑，有效应对各种运维挑战。

3.1.5 可规模扩展的安全机制是保障

在计算集群规模日益扩大及网络智能化发展大背景下，网络安全的重要性愈发凸显。将网络安全融入到业务流程，以实现对用户敏感数据保护，传输数据的机密性及完整性保护，以及提升网络安全的可用性，是实现网络安全运营，应对数据泄露及隐私侵犯等风险的必要手段。

随着网络云化的加速，传统的物理控制管理已无法满足现代云环境中的安全需求。在大规模数据中心中可能面临的网络威胁包括：拒绝服务（DoS）、伪造请求源和修改请求消息数据等中间人攻击。

基于大规模智算中心网络的安全管理场景，传统的IPsec和MACsec协议需要通过集中化管理、动态SA管理、实时监控和分布式架构等方法进行优化，以提高网络安全性能和管理效率。这将有助于确保在大规模部署中实现高效、安全的网络通信。在传统安全协议进行优化或调整的同时，应满足以下需求：

1) 性能增强：通过优化加解密算法的应用，有效降低内存消耗、优化资源使用，并减少对专用硬件的需求，从而降低整体实现成本。

2) 安全功能增强：提供增强的加密和身份验证功能，确保加/解密密钥的安全性和不重复使用，增加数据机密性及完整性保护，提供网络安全可用性。

3) 支持规模扩展：提供安全密钥生成及维护可扩展性支持，降低安全处理对硬件资源的需求，解决安全源依赖性，支持大规模安全密钥状态及维护。

在满足以上需求过程中可能面临如下挑战：

1) 组网规模带来的资源消耗剧增：智算中心计算量呈几何级数增长及网络规模的快速演进，要求新安全技术实现规模支持同时，有效降低对硬件资源依赖；

2) 用户多样化需求复杂：智算中心网络用户需求多样化，包括可靠、定序、重传等不同服务类型。为满足差异化服务的安全传输，需结合报文序列号检测，数据包完整性验证及重传检测等技术，通过安全协议与数据链路层及传输层等技术协同实现对特定安全服务支持；

3) 用户隐私数据泄露及滥用风险增加：随着AI大模型对数据需求不断增加，用户数据面临更高的泄露及滥用风险。需加强对用户数据保护，确保信息交换中用户身份和敏感信息不被泄露，强化数据安全管理和隐私保护，以确保数据的安全性和可靠性。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。
如要下载或阅读全文，请访问：

<https://d.book118.com/108057004060007041>