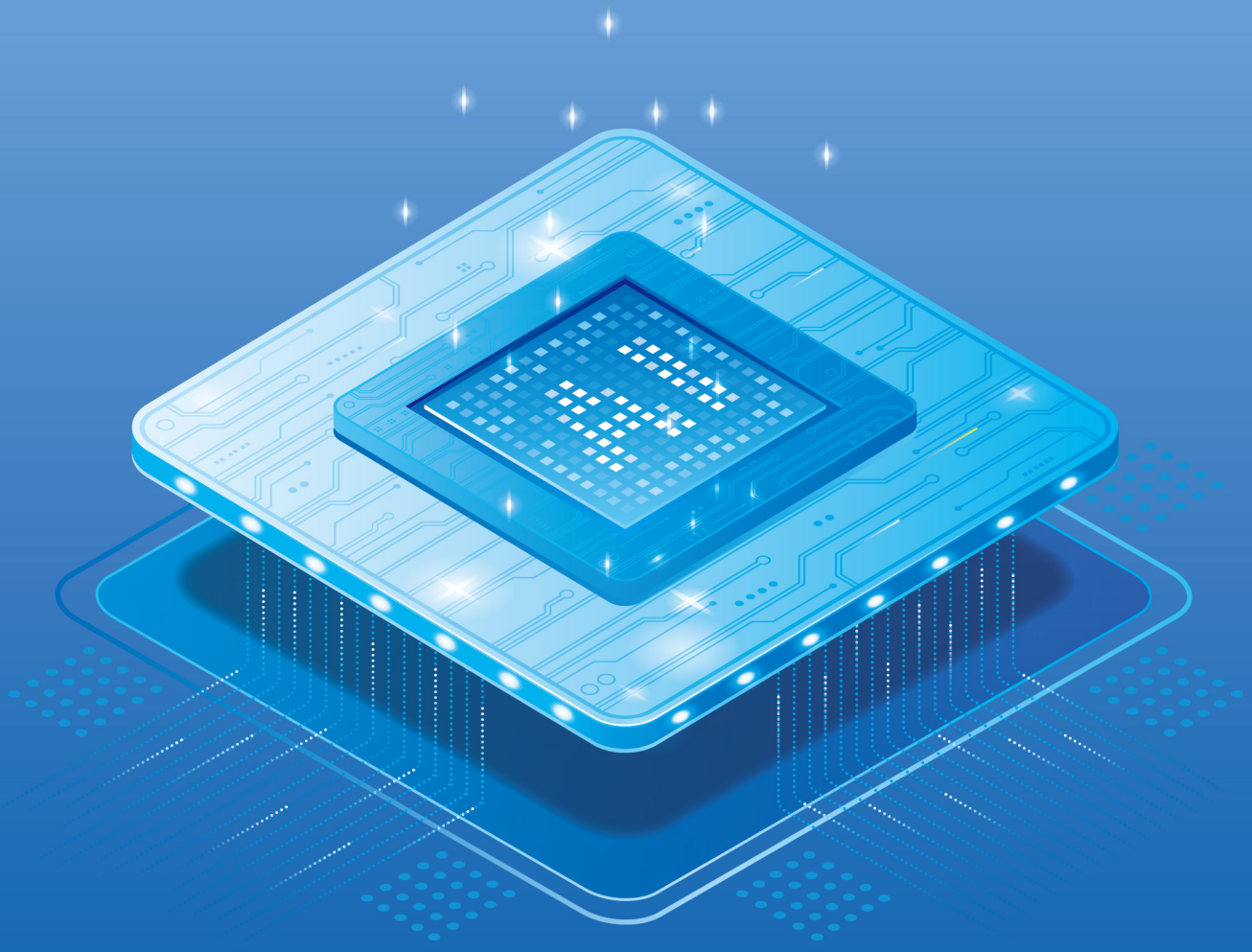
准印证号：（粤B）L011030048

中兴通讯技术 简讯

ZTE TECHNOLOGIES | 第29卷 第04期 · 2025年04月

视点



专题：智算网络

内部资料 免费交流

扫码体验移动阅读

1996年创办 总第439期
2025年04月 第29卷 第04期

中兴通讯技术（简讯）
ZHONG XING TONG XUN JI SHU（JIAN XUN）
中兴通讯股份有限公司主管

主办单位：中兴通讯技术杂志社
编辑：《中兴通讯技术（简讯）》编辑部
发行范围：国内业务相关单位
印数：5000本
出版频次：按月
地址：深圳市科技南路55号
邮编：518057
发行部电话：0551-65533356
网址：http://www.zte.com.cn

设计：深圳市奥尔美广告有限公司
印刷：深圳市旺盈彩盒纸品有限公司
印刷日期：2025年4月25日

李新双
中兴通讯承载网产品副总经理

中兴星云智算网络，领航智算新发展

2024年3月5日，国务院总理李强在第十四届人大会议的《政府工作报告》中提出，要加快发展新质生产力，深化大数据、人工智能等研发应用，开展“人工智能 +”行动。这是国家首次提出“人工智能+”理念并将其上升到战略行动层面。2024年12月，DeepSeek发布，对传统Transformer模型进行了一系列技术革新与优化，显著降低了AI应用门槛，快速拉动了推理算力需求。人工智能产业发展势头迅猛，据IDC最新预测，2025年中国人工智能市场规模将达到4721亿元，年复合增长率将超过30%。

为助力智算业务发展，中兴通讯创新推出“星云智算网络解决方案”。该方案具备超强组网、超高性能、自主可控、敏捷创新等特点，为客户提供完善的智算网络支持。

超强组网方面，中兴通讯星云智算网络提供全盒式、单层多轨、框盒混合三大组网架构，一套方案即可覆盖从“千卡”到“十万卡”乃至“百万卡”的全场景应用需求。

超高性能方面，中兴通讯采用业界领先的CELL/VOQ交换架构，设备负载均衡和拥塞控制能力全面领先Packet交换架构。同时，中兴通讯在国内率先实现112Gbps Serdes产品化，其单机支持576个QSFP112 400G端口，且可平滑演进支持800G。此外，创新提出“智能网卡+智算交换机”的端网协同、以网强算方案，将网络带宽利用率从60%提升至99%。

自主可控方面，中兴通讯智算交换机的核心器件均实现自主研发，整机器件国产化率可达100%；自主研发的全套NOS网络操作系统，已成为国内工业级嵌入式实时操作系统领域领军者。

敏捷创新方面，中兴通讯采用业界领先的MASA全可编程架构设计，在无需进行产品硬件变更的情况下，能快速支持全新的协议与功能。在智算网络领域，中兴通讯不断创新，已先后支持中国移动GSE全调度以太网、谷歌CSIG拥塞信号等多项前沿新型网络方案与技术。

展望未来，中兴通讯将秉持开放合作、持续创新的理念，携手全球智算领域合作伙伴，顺应算力时代浪潮，共同构建新型智算网络，为国家新质生产力的快速发展贡献力量。

目次

中兴通讯技术（简讯）2025年第04期

中兴星云智算网络，新一代高阶智算网络解决方案

自2021年ChatGPT 3引发AI智算热潮以来，短短3年时间里，模型参数量便从千亿规模跃升至十万亿级别，平均每年保持10倍的惊人增长速度。随着AI大模型参数量的迅猛增长，所需的GPU集群规模也在同步扩大。从万卡集群迈向十万卡集群的过程中，业界对网络产生了新的需求。

视点

专题：智算网络

成功故事

中兴通讯智能家用摄像头荣膺IDEA大奖

在IDEA工业设计大奖评选中，中兴通讯智能家用摄像头以其新颖独特的造型荣获FINALIST奖。面对在设计、功能上都高度同质化的家用摄像头市场，中兴通讯智慧家庭以独特的审美设计和功能创新带来了别开生面的产品。

以小兴看看SC50为例，其首创的物理遮蔽时钟显示功能与多彩配色，不仅考虑到了摄像头在家庭环境中所承载的隐私信任和人文关怀，同时兼顾现代家庭的家居颜值需求。

中铝集团与中兴通讯达成战略合作

3月，中国铝业集团有限公司（以下简称中铝集团）与中兴通讯股份有限公司（以下简称“中兴通讯”）在北京举行战略合作签约仪式。双方将依托各自的资源与技术优势，在数字化领域、产业上下游生态建设、前瞻性课题研究与人才培养、产品互推广及联合品牌宣传等多个维度展开深度合作，共同为有色金属行业的智能化与数字化发展注入新动力。

中铝集团党组副书记、总经理王石磊，党组成员、副总经理魏成文，中兴通讯总裁徐子阳、中兴通讯副总裁张继军等出席本次签约仪式。

王石磊在签约仪式上指出，中铝集团作为中央管理的国有重要骨干企业以及国有资本投资公司试点企业，肩负着保障国家战略性矿产资源与先进材料供应的重任，同时也是行业创新与绿色发展的引领者。目前，中铝集团正大力推进信息化建设与传统产业的迭代升级，已在智能方案、智能工厂等方面取得了阶段性成果。中铝集团深厚的行业底蕴以及积极创新求变的决心，为双方的合作奠定了坚实基础。

徐子阳介绍了中兴通讯在芯片、5G、AI等领域的技术优势与业务布局。他表示，中兴通讯凭借长期高强度研发投入、稳定的供应链体系及优质的服务能力，在行业内树立了卓越口碑。未来，中兴通讯将聚焦网络侧与算力业务，持续为行业发展注入新动能。

中兴通讯携手河北移动发布肿瘤AI诊断应用，分钟级病理分析，准确率突破95%

2025中国移动云智算大会上，中兴通讯联合河北移动及三甲医院，展示了一款专为医疗行业打造的AI赋能产品——肿瘤细胞诊断应用。该应用基于中兴通讯AiCube智算一体机，将AI大模型深度融入医疗诊断分析流程，可在分钟级生成分析结果与诊断报告，准确率突破95%，显著优化诊断效能，为医疗智能化的发展提供有力支撑。

中兴通讯ECSS全球合规在线筛查服务上线

面对日趋复杂和不确定的国际经贸环境，以及不断升级的全球合规压力，中国企业亟需更经济、更高效、更智能的合规管理工具。中兴通讯深刻理解企业合规的痛点与难点，依托自身深厚的合规实践经验和领先的数智技术，推出ECSS“全球合规在线筛查服务”，旨在助力中国企业以更低的成本、更高的效率，应对日益严峻的合规挑战，在全球舞台上行稳致远。

中兴通讯与万华化学举行ECSS项目启动会

为积极顺应时代发展趋势，提升企业合规管理智能化水平，3月，中兴通讯与万华化学集团股份有限公司在烟台磁山万华研发中心举行ECSS（企业合规服务系统）实施项目启动会，正式开启数智化合规管理体系建设合作。这一合作不仅标志着两大企业在合规数字化道路上迈出了重要步伐，更为行业的合规管理变革树立了崭新的标杆。

中兴通讯获2025 Lightwave+BTR全球光通信多项创新大奖

3月，全球光网络领域知名媒体Lightwave公布了2025年光通信年度创新大奖（Lightwave+BTR Innovation Reviews）评选结果。中兴通讯凭借其在光通信领域的卓越创新，一举斩获多项大奖，包括最强C+L全频一体化光模块、空芯光纤传输系统突破和创新、可插拔800G OTN解决方案、基于意图的OTN业务自动开通、三模对称50G PON Combo方案以及Wi-Fi 7 D-WLAN Mesh方案，彰显了中兴通讯在光网络领域的卓越技术实力和领导地位。

最强C+L全频一体化光模块

中兴通讯业界首款C+L全频一体化光模块，频谱调整范围为业界2倍，系统容量提升25%且备件种类减半。此外，该模块采用可插拔设计，与固定光模块相比，尺寸减小60%，每Gbit功耗降低68%，大幅节省机房空间并降低能耗，助力运营商构建更高效、更灵活、更绿色的光网络。

空芯光纤传输系统突破和创新

中兴通讯在光传输技术上取得重大突破，使用空芯光纤成功实现超高速传输。不仅单波长传输速度达到1.2Tbps，还在20km光纤上实现了超100Tbps的容量，成功连接智能计算中心。这项新技术的亮点包括：更智能的算法提升传输效率，高功率放大器和高级调制技术让传输更快更远，支持超宽带传输，以及比传统光纤低30%的延迟。这一创新为未来数据中心和算力网络的高效、快速传输铺平了道路。

可插拔800G OTN解决方案

中兴通讯业界首个可插拔800G OTN解决方案，具有四大显著优势：大容量、高性能、极致灵活、全场景覆盖。该方案已在运营商现网成功部署，实现最大单纤容量64T，支持2000km无电中继传输，线路侧端口密度翻倍，功耗降低68%。该方案可广泛应用于骨干、城域和DCI等多种场景，即插即用，部署快捷、运维轻松，助力客户有效节省网络CapEx和OpEx。

基于意图的OTN业务自动开通

随着OTN网络应用场景的不断扩展和客户个性化需求的日益增长，OTN业务的配置难度大幅增加，基于意图的业务自动开通正是针对这一挑战的有效解决方案。它通过意图输入、意图编排、意图发放等环节的自动化功能，实现最少参数输入、最少人工干预、最佳资源匹配的业务快速发放，使维护人员工作量减少20%，网络资源利用率提高10%，并将业务开通时间缩短至秒级，从而显著提高了运营商的市场竞争力。

三模对称50G PON Combo方案

中兴通讯三模对称50G PON Combo方案，实现GPON、10G PON、50G PON三代PON技术平滑升级，ODN无需改动，ONU按需部署，接入带宽上下行均可达50Gbps。此外，该方案通过支持硬管道传输机制实现精准带宽、时延和抖动，灵活满足更多业务需求，助力构建极速万兆基础设施。

Wi-Fi 7 D-WLAN Mesh方案

中兴通讯Wi-Fi 7 D-WLAN Mesh方案，采用Co-EDCA技术、中心化控制和VBSS漫游三大核心技术，并结合AI数据识别，优化高优先级业务的传输和用户漫游体验。通过动态调整空口参数、智能调度Mesh网络各节点，确保高优先级数据以低时延、低抖动的方式传输，在漫游过程中实现零丢包，达到无损漫游体验。该方案显著提升了用户体验，助力运营商增强竞争力。

Lightwave+BTR主编Sean Buckley向中兴通讯表示祝贺。他强调了这一认可的重要性，特别指出中兴通讯通过其突破性的解决方案、技术和项目为行业所做的创新贡献。

智算网络演进趋势

陈志伟
中兴通讯承载网规划总工

从2022年11月ChatGPT推出后，大模型的发展就进入加速阶段，大模型网络参数每3年增长1000倍，集群规模每2年扩大至原来的4倍。2023年，千卡GPU训练池开始部署，2024年，万卡集群完成部署，xAI的十万卡集群也已启动。从万卡集群迈向十万卡集群的过程中，业界对网络产生了新的需求。英伟达作为AI基础设施的标杆企业，从横向扩展（Scale-out）和纵向扩展（Scale-up）两个维度对更大规模的集群进行拓展。在Scale-out方面，英伟达主要推行IB（InfiniBand）技术，其次是RoCE（RDMA over converged ethernet）；在Scale-up上，则采用其私有的NVLink技术，而其他多数厂家选择了开放解耦的发展路径，其中互联网企业在这方面的探索更为领先。

DeepSeek和Grok 3代表了大模型发展的两个侧重方向。

2025年1月，DeepSeek火爆出圈。它凭借低成本训练取得了显著成果，仅使用2000张H800卡，经过15天的训练，就达到了OpenAI、LLaMA等模型采用万卡规模训练所获得的效果。这体现了以低成本投入、依靠算法优化来降低成本的发展逻辑。

2025年2月，xAI发布了Grok 3。该模型使用了10万张H100卡，在数学、科学、编程三大评测中均排名第一，彰显了通过堆积大算力，不断提升AI性能和拓展新能力的发展逻辑。

我们认为，上述两种发展路径都将继续推进，“扩展定律”（Scaling Law）依然有效。

Scale-out：RoCE部分替代IB，GSE/UEC胜过IB指日可待

对于Scale-out而言，其核心诉求为实现大规模互联，尽可能提升带宽利用率，并减少网络阻塞。当前，借助大容量的盒式交换机与框式交换机，叠加多轨道优化技术，能够搭建起千卡、万卡乃至十万卡规模的网络。同时，头部互联网企业基于标准的RoCE，通过端网协同的算法优化，以及更有效的拥塞控制，进而提高带宽利用率。腾讯、阿里、谷歌、AWS等企业都采用了类似思路，在具体实施过程中存在细节差异。互联网

从目前GSE/UEC的进展来看，标准在超大规模组网、缩短作业完成时间（JCT时间）、优化负载分担算法以及端到端协同等方面，均取得了良好的技术创新与显著进展，超越IB指日可待。

基于RoCE的优化方案虽具有优势，但存在两个缺点：其一，该方案是建立在端到端自研基础上的私有实现，仅能供自身使用，难以广泛部署；其二，其主要创新点集中在软件层面，硬件部分依旧采用当前标准的商用器件，存在诸多制约因素。

鉴于此，业内成立了联盟，共同构建下一代智算网络的新标准。其基本逻辑包含三点：一，基于以太网，物理层（phy层）和介质访问控制层（mac层）均采用以太网标准；二，从物理层到传输层都要进行优化；三，实现端到端支持，即从网卡到网络设备都需满足要求。下一代智算网络标准以GSE（国内）和UEC（海外）为代表。预计2025—2026年，会有支持GSE、UEC的产品及解决方案推出。

从目前GSE/UEC的进展来看，标准在超大规模组网、缩短作业完成时间（JCT时间）、优化负载分担算法以及端到端协同等方面，均取得了良好的技术创新与显著进展，超越IB指日可待。

Scale-up：内存语义和消息语义并行发展

对于Scale-up而言，随着张量并行（TP）扩展，从8卡扩展到16卡，再到英伟达NVL72的72卡，高带宽域（HBD）可扩展支持至NVL576，多GPU的HBD域互联显得极为重要。其核心需求在于具备高带宽、低时延以及在网计算能力，以便更好地支持TP并行。英伟达作为行业引领者，其最新的NVLink5.0已支持224G serdes。

Scale-up的标准需求主要源于两大驱动力：其一，Scale-up的多卡扩展趋势已然形成；其二，NVLink属于私有技术。因此，2024年UAL联盟迅速成立，国内也相继出现ETH-X、OISA、Alink等组织。与Scale-out一开始就在技术逻辑上统一采用以太网不同，Scale-up始终存在两条技术路线：一条基于总线型，在逻辑上需在PCIe基础上进行扩展，以支持内存load/store操作，并实现内存一致性，代表性厂商是AMD的Infinity；另一条基于网络型，在逻辑上要对RoCE进行裁剪与扩展，代表性厂商是英特尔。

最新进展是，UAL的第一个版本已决定采用以太网作为物理层和链路层，放弃了PCIe，“Infinity over Eth”取代了最初的“Infinity over PCIe”，当然仍会支持内存语义。之所以选择这一方案，主要是因为以太网能够提供比PCIe更宽的带宽，且以太网发展更为迅猛。与此同时，国内ODCC也立项了ETH-X项目，单机柜可容纳64

卡。可将其视为中国版本的NVL72，但该项目采用以太网进行机内互联，计算板（compute tray）和交换板（switch tray）能够解耦，可兼容多家GPU和交换机，从而为客户提供更多选择。

从当前情况来看，Scale-up方面的进展相较于Scale-out稍显滞后。不过，在整体需求定义、架构选择以及芯片路径等方面均取得了快速进展。期待OISA/UAL能够推出如同英伟达交换机一样高品质的方案，为非英伟达GPU的多卡应用奠定网络基础。

拉远、CPO等新技术在大规模智算网络中加速演进，重要性日益凸显

在打造超大规模智算集群时，拉远技术可能成为关键考量因素。这主要源于机房功耗问题，以NVL72设备为例，单台功耗达120kW，万卡集群至少需20MW，而100万卡集群则高达2GW，这远远超出了普通园区的供电能力。然而，大模型计算所采用的众多算法对时延极为敏感，因此针对拉远技术的优化势在必行。目前，研究重点聚焦于300km以内的网络方案，这要求在网卡侧对RDMA协议进行拉远算法的拥塞控制（CC）升级，同时交换机要与网卡实现拉远协同，并具备缓存正在传输（in flight）流量的能力。若传输距离超过100km，就需引入OTN设备，使用新型空心光纤则有助于降低时延。

供电限制引发了对拉远技术的需求，不过换个思路，也可通过降低功耗来提升单个集群的规模上限。降低功耗可从优化光模块、采用液冷技术等几方面着手。

交换芯片从640G升级到51.2T，容量提升80倍，芯片功耗仅提升8倍，而光模块功耗却飙升了26倍，降低网络功耗，光模块是关键。针对光模块优化，主要有LPO、LRO、CPO几种新技术。LPO（linear-drive pluggable optics，线性驱动可插拔光模块）技术通过去除光模块中的数字信号处理器（DSP），大约能降低1/3的功耗，但存在设备对接难题。目前部分互联网企业已开始尝试应用。LRO（linear receive optics，线性接收光模块）技术则是一种折衷方案，仅去除发送方向的DSP，对兼容性要求较低，相应地，功耗降低幅度也较小。CPO（co-packaged optics，光电共封装）技术较为激进，虽能大幅降低功耗，却给部署和运维带来极大挑战。因此，我们认为可插拔式光模块（LPO）是当前优选方案，CPO可作为最后的考量选项。

当单芯片功耗达到1300W，热流密度高达140W/cm^{2}时，就需引入液冷系统。在各类液冷方案中，冷板式液冷相较于浸没式液冷，因部署和维护更为便捷，成为推荐方案。

总结：AI网络未来可期

中兴通讯认为，和传统HPC超算相比，AI大模型市场规模庞大，全行业均有需求，更多供应商参与、采用解耦的解决方案将是大势所趋。为此，中兴通讯开展了诸多工作：

在Scale-out方面，中兴通讯深度参与GSE核心技术研发。同时，中兴通讯积极投身UEC规范研讨，确保下一代产品能够全面支持GSE和UEC协议。

在Scale-up方面，中兴通讯加入OISA、UAL以及ODCC ETH-X等组织，深度参与超节点的设计工作。

在智算拉远领域，中兴通讯致力于打造端到端开放解耦方案，通过自研网卡、交换机、路由器以及OTN，实现端到端的优化。

在降低功耗方面，中兴通讯积极探索液冷解决方案以及CPO/LPO解决方案，力求在设备层和芯片层实现优化。

中兴通讯坚信，AI大模型给国内网络从业者带来了发展机遇，我们将坚持走开放解耦的道路，与各方协同合作、共同推进，为构建更优质的AI智算网络而不懈努力。ZTE中兴

中兴星云智算网络，新一代高阶智算网络解决方案

中兴通讯 周昆

周昆
中兴通讯数据中心网络架构师

自2021年ChatGPT 3引发AI智算热潮以来，短短3年时间里，模型参数量便从千亿规模跃升至十万亿级别，平均每年保持10倍的惊人增长速度。随着AI大模型参数量的迅猛增长，所需的GPU集群规模也在同步扩大。目前，国际上已经开展了超万卡甚至十万卡规模的工程建设，例如OpenAI拥有2.5万张GPU卡，特斯拉更是配备了20万张GPU卡。国内由于单卡算力性能存在一定限制，为了追赶国际先进的算力水平，就需要部署更多数量的GPU卡来提升整体算力性能。这使得大规模智算组网能力在国内智算业务建设中的重要性愈发突出。

大规模智算组网的挑战

大规模智算组网作为释放强大算力的基础支撑，其重要性不言而喻。随着ChatGPT、DeepSeek应用的爆发式增长，对智算能力的需求呈指数级攀升。迈向大规模智算组网的征程并非坦途，在超大规模、极致性能以及安全可控等维度，正面临着诸多严峻且亟待攻克的挑战。

超大规模

在Scaling Law（扩展定律）的驱动下，当下万卡GPU训练集群仅仅是AIGC核心参与者的入门标准。随着特斯拉率先宣告十万卡集群投入使用，国内云服务提供商如阿里巴巴、百度等也相继宣称具备支持十万卡集群的能力。可以预见，未来将会涌现出更多十万卡甚至百万卡规模的智算集群。

如此庞大的组网规模必然促使网络技术产生质的飞跃。高性能网络架构的主要功能设计以及性能要求，都需要置于支持超大规模网络的框架内重新审视与考量。

超大规模组网面临诸多主要挑战：

- 单机接入的容量限制

服务器GPU网卡的数量和接口速率呈持续增长态势，基本每两年翻一倍。当前，大规模商用的GPU服务器网卡接口已达8×400G，支持800G的GPU服务器也已面市。为了满足日益增长的接入需求，同时减少设备数量，对单交换机容量的要求越来越高。然而单交换芯片的容量提升速度远落后于IO总线的发展速度，还受到物理层面的限制。

- 组网架构的规模限制

为满足数十万卡乃至更大规模的组网需求，鉴于单台盒式交换机支持的端口数量在短期内难以实现大幅提升，传统的盒式交换机组网架构不得不增加网络层次。但网络层次的增多意味着数据转发跳数增加，这不仅导致路径规划更为复杂，还加大了故障发生的概率以及故障定位的难度，使得网络运维工作更困难。

极致性能

为最大程度提升集群算力利用率，AI大模型训练一般会采用并行处理机制，将单个任务分配到多个GPU上同步运算。并行训练总体可分为处理、通知、同步3个步骤：在处理阶段，每个GPU分别执行任务的一部分；到通知和同步阶段，GPU之间进行卡间通信，汇总得出整个任务的最终结果。整体的作业完成时间（job completion time，JCT）取决于GPU返回计算结果的速度以及网络同步的速度。所以，网络性能对集合通信效率有着重要影响。要实现网络的极致高性能，面临以下挑战：

- 时延方面：动态时延通常比静态时延高出几个数量级，是造成网络低时延问题的主要因素。由拥塞、丢包引发的时延往往达到毫秒级别，因此需要更精确的流量控制/拥塞控制以及故障恢复机制。
- 抖动方面：现有的拥塞控制技术，如DCQCN等，主要针对丢包和吞吐量进行优化，在控制网络抖动方面考虑不足，需要更为精准的拥塞及流控机制。此外，大模型训练产生的流量以大象流为主，若网络负载均衡的粒度太粗，也会使网络抖动难以有效管控。
- 吞吐方面：高吞吐设计是一项复杂的系统工

程，涉及到机内外以及软硬件之间的精细协同。单独提升某一项指标，不仅无法有效提高吞吐率，还可能产生此消彼长的负面效应，因为丢包、时延和吞吐这几个因素往往相互影响。

安全可控

在国内各行各业智算业务迅速发展的大背景下，网络产品和技术的安全可控变得愈发重要。将网络安全融入业务流程，提升安全可控的网络产品和技术的可用性，是实现智算业务安全运营的必然要求。满足网络安全可控的需求，面临以下挑战：

- 网络产品器件的供应链安全

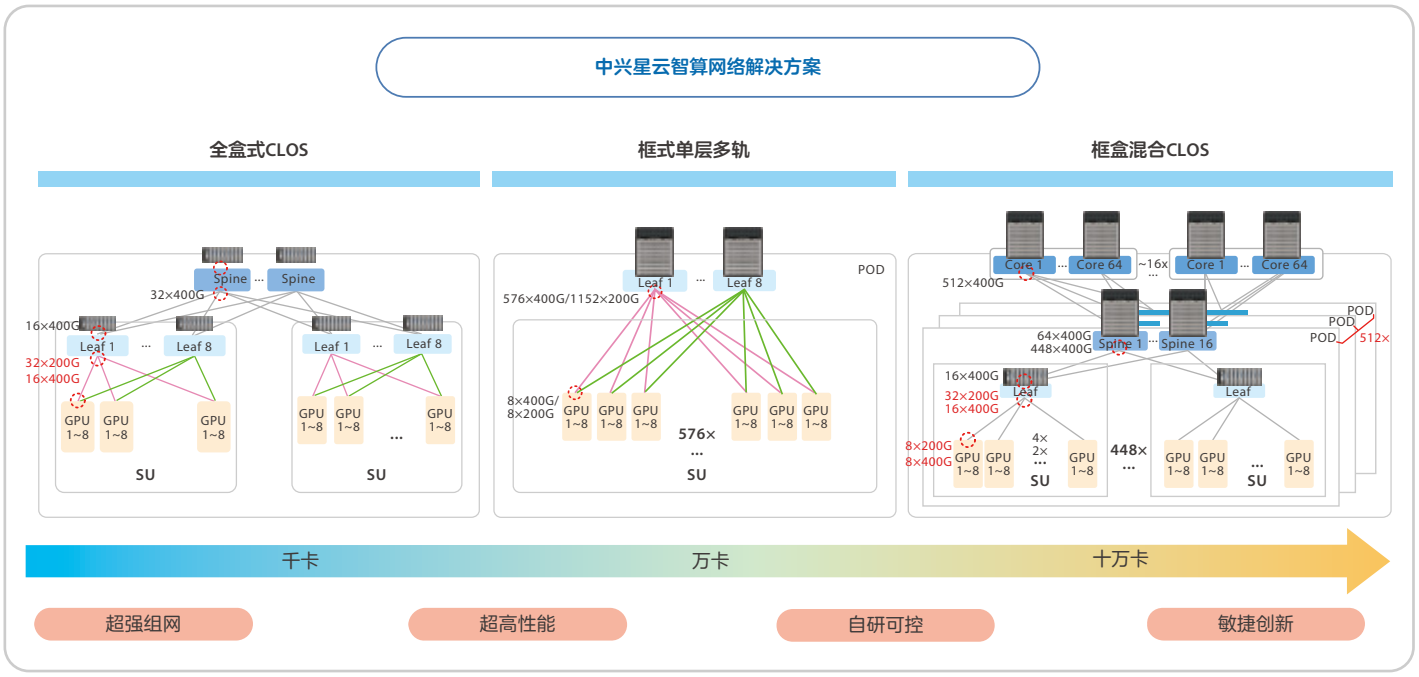
智算网络方案与产品涵盖CPU、转发、交换以及内存、连接器、电源、I2C控制器等众多电子器件，只要任何一类器件被独家垄断，就会对产业生态的健康与稳定造成制约。因此，加强核心器件的自主研发，提高外围器件的国产化率，构建稳固且多元化的供应链体系刻不容缓。

- 网络软件系统的知识产权

一方面，网络软件研发过程存在代码复用现象，有侵权风险，还可能引入安全隐患，比如隐藏的后门程序等，一旦被恶意利用，可能导致网络系统被远程操控。另一方面，开源软件在网络领域应用广泛，虽然带来了诸多便利，但部分开源协议较为复杂，如果企业在使用时未能严格遵守，很容易引发知识产权纠纷。如何在合法合规的前提下，强化自身知识产权保护并突破相关障碍，是网络软件发展面临的关键课题。

中兴星云智算网络解决方案

为契合各大运营商以及政企客户在智算业务发展与网络建设方面的需求，中兴通讯创新推出中兴星云智算网络解决方案。该方案依托全自研的ZXR10 5960和ZXR10 9900X智算网络交换机产品家族，为客户打造出具备超强组网能力、超高性能表现、自主可控、敏捷创新优势的新一代高阶智算网络解决方案（见图1）。



▲ 图1 中兴星云智算网络解决方案

超强组网

智算的大模型训练对算力有着极高要求。目前，大模型训练高度依赖大规模GPU集群互联，需要构建大规模网络来支撑GPU并行计算，从而充分释放GPU集群的最大算力。

针对智算网络的参数面、存储面、业务面、管理面，中兴星云智算网络精心设计了超大规模组网方案。这一方案能够全面覆盖从千卡、万卡到十万卡，乃至未来百万卡的全场景智算业务应用。

基于新一代全自研的ZXR10 5960X和ZXR10 9900X智算网络交换机，中兴通讯为客户提供全盒式、单层多轨、框盒混合三大组网架构，客户可根据自身需求灵活选用。其中，框盒混合组网采用多POD（point of delivery）矩阵架构设计，单个POD最大可支持7000张400Gbps的GPU卡集群组网。该架构可依据不同阶段的智算建设规模需求，支持以POD为单位进行灵活弹性扩展，实现分阶段持续建设交付。整个网络最大可支持512个POD，满足未来百万卡的超大规模组网需求。

超高性能

传统RoCE网络中，拥塞和流控算法在端侧与网侧相互独立，网络仅能提供粗粒度的拥塞标记信息。在网络高负荷场景下，这种机制很难保证不出现拥塞丢包以及排队时延的问题。

针对传统RoCE网络在拥塞控制精度方面存在的性能限制，中兴通讯创新性地设计出ENCC（end to network congestion control）这一高精度端网协同拥塞控制技术。通过运用ENCC技术，方案实现了以网络优势助力计算性能提升，端到端网络带宽利用率从60%大幅提升至99%。

在大规模智算组网环境中，存在着大量等价链路，为了充分发挥每张GPU的算力，需要网络充分挖掘每条链路的转发性能，避免出现链路负载不均衡的现象。传统基于Flow五元组哈希的负载分担策略，极易引发哈希极化以及负载不均衡的问题，无法有效保障网络负载分担效率；基于Packet逐包哈希的策略，则会导致报文乱序问题，使其难以在商业场景中得到实际部署应用。

基于上述情况，中兴通讯提出了层次化负载分担架构设计方案。该方案通过不同层级负载均衡机制的协同配合，弥补单一负载均衡方案存在的缺陷，以更好地实现全网流量均衡，并达到高吞吐的目标。

首先，从全局视角出发，设计了智能全局负载分担（intelligent global load balance，iGLB）方案。网络控制器通过API接口，接收算侧调度平台传递过来的流量特征信息，然后依据网络负载状态，对全局路径进行统一规划，以此确保全网每条链路的负载分担效率达到最优水平。

其次，针对模型训练过程中可能出现的局部网络故障或拥塞情况，还配备了自适应路由ARN（adaptive routing notification）技术。交换机可根据本地出口的可用状态和拥塞情况，动态选择出口。在全局规划的基础上，该技术主要针对因网络突发事件导致的瞬时流量不均衡问题，及时对路径进行局部调整。若本地不存在其他满足条件的可用路径，交换机会自动通过数据面报文向上游节点发送通知，促使其进行路径切换，从而实现远端路径的快速调整 。

自主可控

中兴通讯于1996年设立IC设计部，正式启动设备核心器件自研工作，至今已有20余年的持续技术沉淀，中兴通讯已成功量产各类器件100余种，累计发货量突破12亿颗。

在软件方面，中兴通讯早在2002年就着手自研嵌入式OS内核，拥有超20年的操作系统研发及商用经验，具备完全自主知识产权。其嵌入式实

在百度文心、阿里通义、字节云雀、移动九天、电信星辰等国内头部大模型智算业务蓬勃发展的带动下，国内智算中心正快速向超万卡级别的规模迈进，并且部分企业已在规划十万卡规模的智算方案。

时操作系统年发货量超1500万套，累计发货量达2亿套，服务范围覆盖全球130多个国家和地区。

中兴通讯星云智算网络选用的ZXR10 5960X和ZXR10 9900X智算网络交换机，核心器件均采为中兴微电子自主研制。这些自研器件能够完全替代商业器件方案，从根本上化解了核心器件的供应链风险。其中，ZXR10 9900X采用业界领先的CELL/VOQ交换架构，在设备交换性能上全面超越传统Packet交换架构。此外，中兴通讯还是国内首个完成112Gbps Serdes产品化的厂商，ZXR10 9900X单机最大可支持576个400G QSFP112端口，并且能够平滑升级，以支持576个800G QSFP112-DD端口。同时，针对智算网络交换机的其他外围器件，中兴通讯也全面实现了国产化替代，目前整机的器件国产化率可达到100%。

敏捷创新

中兴通讯星云智算网络依托自研的MASA全可编程技术优势，全面支持P4编程语言。基于自研的PPC+RTC融合架构，数据报文由网络接口接收后，会先经过Pre-Parse报文解析。对于简单业务，可在PPC的多级微引擎流水中进行处理，处理操作包括分类、查表、过滤、修改等；RTC则是完全可编程的微引擎，作为PPC的补充处理部分，它支持在无需替换和修改产品硬件的前提下，仅通过软件编程就能快速适配全新的协议标准和网络特性，这使得产品具备丰富且灵活的无损网络调优能力，能够持续创新智算相关的新特性。目前，中兴通讯星云智算网络已通过敏捷创新，成功支持了中国移动GSE全调度以太网等全新网络方案。

在百度文心、阿里通义、字节云雀、移动九天、电信星辰等国内头部大模型智算业务蓬勃发展的带动下，国内智算中心正快速向超万卡级别的规模迈进，并且部分企业已在规划十万卡规模的智算方案。可以预见，在不久的将来，将会有更多十万卡规模的智算集群相继出现。中兴通讯期待与众多合作伙伴携手共进，共同推动智算网络技术的发展，拓展其应用场景，实现由人工智能驱动的新质生产力的提升，促进社会的共同繁荣。ZTE中兴

智算网络Scale-out的演进思路：GSE和UEC

王恒
中兴通讯交换机产品规划经理

在AI人工智能、高速计算等业务迅猛发展的当下，算力需求呈爆炸式增长态势。算力集群正从千卡规模，逐步向万卡、十万卡乃至百万卡规模演进。其中，智算网络作为连接各算力单元的外部纽带，对于能否充分释放整体最大算力、发挥更优算力能效，起着至关重要的作用。

然而，由于所承载业务发生变化，由传统数据中心网络发展而来的智算数据中心Scale-out网络，面临诸多新挑战。

- 智算网络的流量具有低熵、大象流、同步效应等特征，致使传统负载均衡策略难以发挥作用，网络拥塞冲突频发。这大幅降低了网络的有效吞吐能力，进而影响算力效率。
- AI大模型运算依赖多GPU并行计算，多GPU之间存在海量通信，对时延要求极高。一旦发生拥塞，动态时延大幅增加，GPU就不得不等待数据传输完成后才能开展下一步计算，导致其无法满负荷运行，算力无法充分释放。
- RDMA协议采用“Go-back-N”机制，这要求网络必须确保零丢包。一旦出现丢包，进行重传操作，网络的有效吞吐将急剧下降，严重影响算力效率。

因此，如何解决因网络因素导致的AI算力浪费问题，如何让网络与算力更紧密适配，避免网络成为大模型算力的瓶颈，成为智算网络亟待攻克的关键目标。目前，业内各方针对上述挑战开展了诸多努力与探索，催生出各种技术与理念。

结合智算流量的业务特征，各厂家针对RoCE网络开展了大量的增补与优化工作，期望实现“网”与“算”的良好适配。由于不同厂家的主打产品和技术切入点存在差异，该优化路线又细分为网侧优化、端网协同优化等方向。以中兴通讯为例，其iGLB（Intelligent Global Load Balance，智能全局负载均衡）、ZRLB（ZTE Rail Load Balance，端口级负载均衡）等技术，是从网侧视角出发解决负载均衡问题。而谷歌提出的CSIG、阿里提出的HPCC等技术，则采用端网协同方式，端侧借助网侧随路采集的拥塞状态信息进行流量调控，从而有效解决网络拥塞问题。

然而，受现有网络底层逻辑的限制，针对Incast流量拥塞、少流哈希等问题的处理，仅在特定场景下有效，且方法复杂，存在一定局限性。为此，国内外厂家携手合作，尝试对网络进行底层改造，从而衍生出网络架构重构路线。该路线旨在打造全新的网络体系，通过采用新的转发机制，从底层逻辑层面化解当前无损以太网面临的困境，让Scale-out网络在AI人工智能、高性能计算等场景中，能够达成无阻塞、高带宽、超低时延的目标。这一路线的主要代表方案有GSE（Global Scheduling Ethernet）、UEC（Ultra

Ethernet Consortium）等。

全调度以太网GSE

2023年5月，中国移动主导提出GSE概念，并联合业界的设备商、芯片商、云服务商、科研院所等机构，旨在针对智算场景，革新以太网基础转发机制，全面提升网络性能，共同打造一个新型的高速无损、安全可靠、开放兼容的网络架构技术体系。

在GSE技术体系中，考虑到智算产业现状与技术发展节奏，其发展分为GSE1.0和GSE2.0两个阶段。

GSE1.0技术，基于现有产品能力，提出了一系列优化RoCE网络性能的技术，如端口级负载均衡、算网协同负载均衡、端网协同的拥塞感知等，这些技术属于对现有网络的改进范畴。

GSE2.0技术，则提出了一系列重新定义下一代网络转发机制与上层协议的技术，比如基于容器的多路径喷洒、基于DGSQ（dynamic global scheduling queue，动态全局调度队列）的拥塞控制等，这些技术属于网络重构范畴。

通常所说的GSE，更多是指用于规范下一代网络的GSE2.0技术。在GSE2.0协议标准中，对GSE技术进行了全面阐述，并定义了GSE网络的基本架构，如图1所示：

- 控制层：包含GSOS（global scheduling operating system，全调度操作系统）和NOS（network operating system，网络侧操作系统），主要负责实现全局信息编址、映射等配置与管理功能。
- 网络层：包含源GSP（global scheduling processor，全调度网络处理节点）、目的GSP、GSF（global scheduling fabric，全调度交换网络）三类设备，这些设备对GSE报文进行识别、封装、解封装以及转发等操作，从而实现全网的流量调度和链路间的负载均衡。
- 计算层：包含高性能计算卡和网卡，作为全调度以太网的服务层。

GSE2.0协议标准提出的核心技术包括报文容器（PKTC）、DGSQ、控制面设计、健壮性设计、网络可视化设计等。

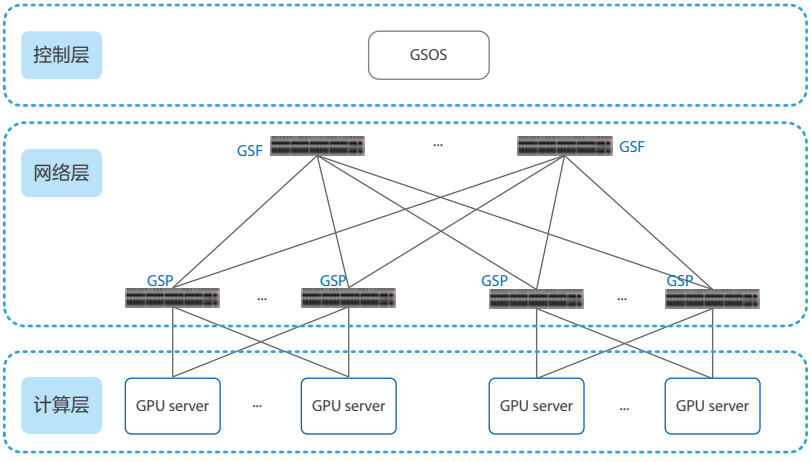
PKTC技术运用主动均衡的理念，构建等长的虚拟容器，并将其作为网络转发的最小单元。如此一来，原有的以太网报文基于流的哈希转发方式转变为基于逻辑容器的多路径喷洒方式，有效解决了因智算流量具有低熵或大象流等特征而引发的负载不均衡问题。

DGSQ拥塞控制技术通过端到端的授权机制，对发往目的端口的总流量进行控制。只有在源GSP收到授权后，才能够发送报文，并且发送的报文数量不能超过授权数量，这从根本上避免了因Incast因素导致的拥塞情况。当源GSP达到设定的水线后，会通知发送源降低发送速率，从而实现流量速率的动态平衡。

控制面设计从全局和分布式两个维度对控制面进行定义，以实现对网络的配置与控制管理。

健壮性设计涵盖故障检测、故障通告、故障收敛等方面。同时提出了更有效的实现方式，比如通过扩展O码来减少故障通告所产生的带宽开销等。

网络可视化设计包括丢包检测、时延检测、根因检测、报文统计等功能，有助于提升网络的



▲ 图1 GSE网络基本架构

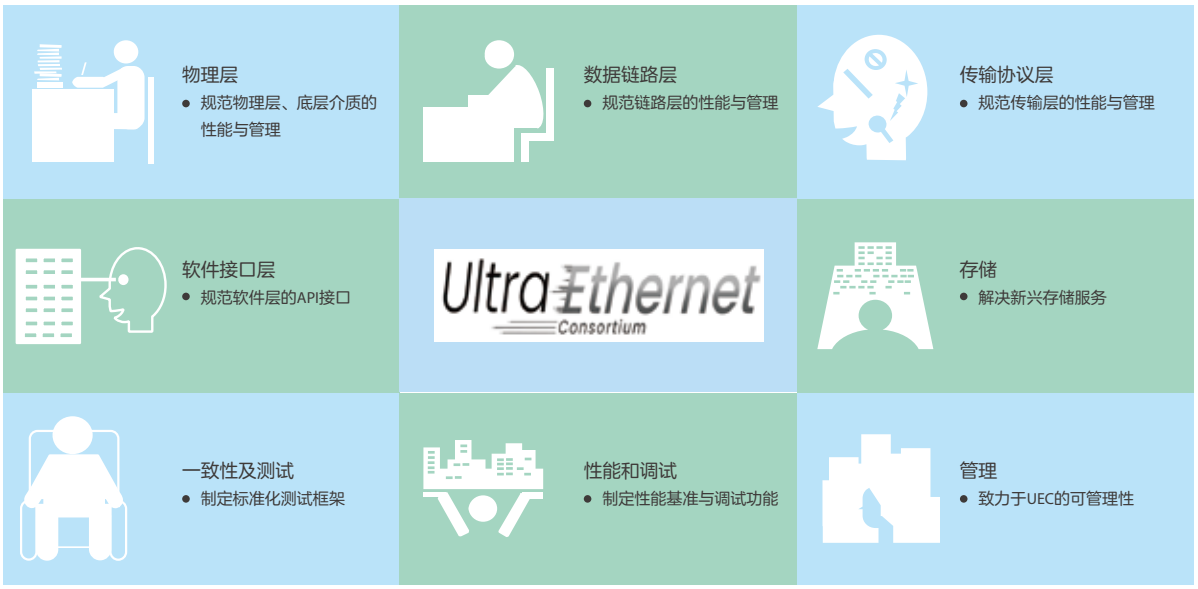


图2 UEC工作组介绍 ▶

运维能力。

超以太网联盟UEC

UEC于2023年7月成立，由Linux基金会主办，汇聚了半导体、设备和云厂商等多方企业，开展全行业合作，旨在构建新一代基于以太网的完整通信栈架构，让超以太网能够更好地满足AI和HPC（high performance computing，高性能计算）对网络的需求。

UEC最初设立了物理层、链路层、传输层、软件层4个工作组，后续又相继成立了存储、管理、兼容性&测试、性能&调试工作组（见图2）。这些工作组从不同层面入手，对网络进行改进或重新定义，并提出了一系列关键技术：

- 多路径和数据包喷洒：借助数据包喷洒技术实施路径的精细化负载均衡操作，能够让每条数据流同时利用所有可通往目的地的路径，进而实现网络链路的均衡使用。
- 灵活的排序：对于那些仅关注消息最后部分何时抵达目的地的传输服务场景而言，在将数据包传送给应用程序之前，无需进行重新排序操作。适当放宽对逐包排序的严格要求，能够有效减少数据传输的时延，实现快速且批量的数据传输。
- 优化拥塞控制：为了有效规避Incast拥塞问题的产生，提出了基于接收端的拥塞控制策略。具体来说，由接收端向发送端分配信用值（Credit），发送端依据接收端所分配的信用值以及接收端的实际处理能力，相应地调整数据发送速率。
- 端到端遥测：着重增强拥塞通知的能力，向终端节点提供更为详尽的拥塞位置、拥塞原因等相关信息。同时，提出有效的拥塞缓解算法，使终端节点能够快速感知拥塞事件并做出反应。

当前，智算网络Scale-out尚处于探索的初级阶段，技术路线呈现多样化态势是必然现象。随着实践的持续深入，无论是重点聚焦网侧优化或端网协同优化的优化路线，还是以GSE、UEC为代表的重构路线，乃至其他新兴技术与新生态，其目标都高度一致，即让网络更好地适配大规模算力，为构建新一代智算网络筑牢坚实基础。ZTE中兴

智算网络Scale-up技术路线：总线型还是网络型

AI技术发展至今，大模型呈现出超百万亿参数、长序列、多模态、推理/测试时计算（test-time scaling）以及物理AI几大明显的发展趋势，可以预见的是，AI对集群算力的需求仍将保持高速增长的态势，智算集群发展到十万卡甚至百万卡规模已成为行业发展的必然需求。在超十万卡规模的智算集群中，由Scale-up网络构成的高带宽域（即业界所说的超节点域）将扮演着重要的角色。以NVIDIA NVL72为例，相比上一代单机8卡服务器，在同等的32k集群规模下，GPT-MOE-1.8T模型推理性能提升30倍，训练效率提升4倍。由于相比Scale-out网络这种明显的业务加速收益，Scale-up网络从2024年开始成为业界研究的焦点。国外以AMD为代表的GPU厂商牵头成立了UAlink技术联盟，而国内短时间出现腾讯ETH-X、

李和松
中兴通讯智算网络资深架构师

中国移动OISA以及中兴通讯提出的OLink（即Open Link）等多种技术方案，并出现了“总线型”和“网络型”两种技术路线之争。这两种技术路线到底是水火不容还是殊途同归，是当前行业关注的焦点问题。

第一性原理看Scale-up网络：核心需求是什么

Scale-up网络作为智算集群所引入的一种新型网络类型，在进行技术路线选择时，首先要明确其核心的技术需求是什么。以当前业界Scale-up网络的行业标杆NVIDIA为例，其NVLink技术目前已经发展到了第5代，最初是为了解决PCIe带宽不足问题而设计的。当GPU之间的通信带宽需求达到300GB/s，为了实现8卡之间的高速全互联，出现了第一代NVSwitch芯片。从NVIDIA官方透露的未来3年芯片规划来看，Scale-up网络呈现出如下发展趋势：NVLink接口带宽逐代稳步提升（NVLink5的1.8TB/s到NVlink 6的3.6TB/s），Scale-up互联规模渐进式提升（NVL576到NVL1k），随之而来的是其配套的交换芯片NVSwitch的容量同步提升。基于这些基本的技术信息，再综合应用场景的基本特点，可以总结出Scale-up的一些基本特点。

总体而言，Scale-up需要的既不是传统的总线技术，也不是传统的网络技术，Scale-up是具有其特定诉求的新型互联应用场景。任何将总线技术（如PCIe）或网络技术（如以太网、Infiniband等）直接照搬应用在Scale-up场景的做法，都将引入某个维度的性能代价。

Scale-up网络本身是“总线”和“网络”技术在特定场景融合应用的产物，其综合了两者的特点，并融入了自身的核心诉求（见图1）。

一方面，Scale-up继承了总线技术的一部分用户侧需求，如支持Load/Store内存语义、要求控制芯片PPA（performance power area，能效比）代价等，但引入了超高带宽和相对宽松的时延需求。以最新PCIe 6.0×16为例，其带宽只有256GB/s，这与NVLink动辄TB级的通信带宽存在明显的代差。从Scale-up本身的应用场景出发，由于存在计算–通信掩盖等优化技术，Scale-up网络相对而言具有更高的时延容忍度，然而，由于Load/Store语义的同步通信特征，延迟又不能过于宽松。

另一方面，Scale-up网络也继承了网络技术的扩展性需求。由于技术、成本等多方面因素的约束，Scale-up互联规模从最初的8卡发展到如今的百卡，未来有可能扩展到千卡，这种规模需求介于总线互联规模和网络互联规模之间，因此必须借鉴网络技术高可扩展性的设计经验。

此外，由于Scale-up应用场景的特点，为了给上层应用提供一个高效、友好的通信环境，需要支持DMA（direct memory access）语义和内存统一编址，其中DMA语义可有效提升数据批量传输的性能，内存统一编址能为上层应用提供更友好的编程模型。

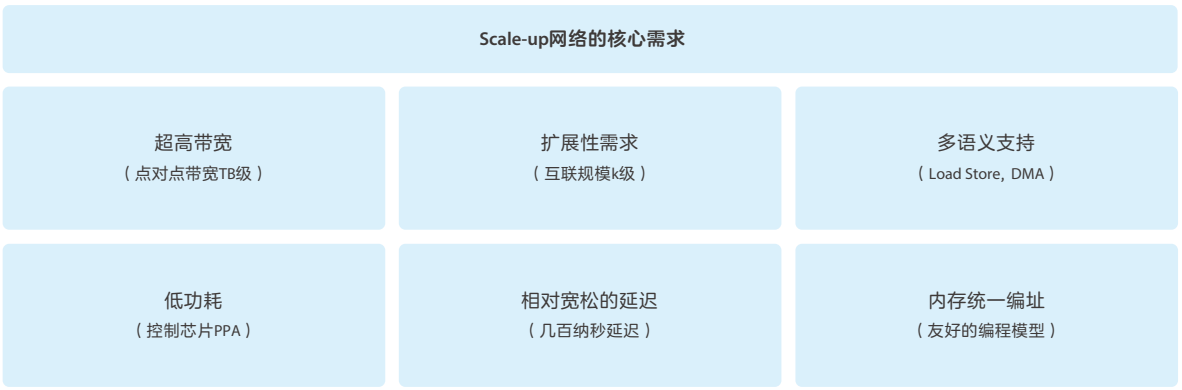
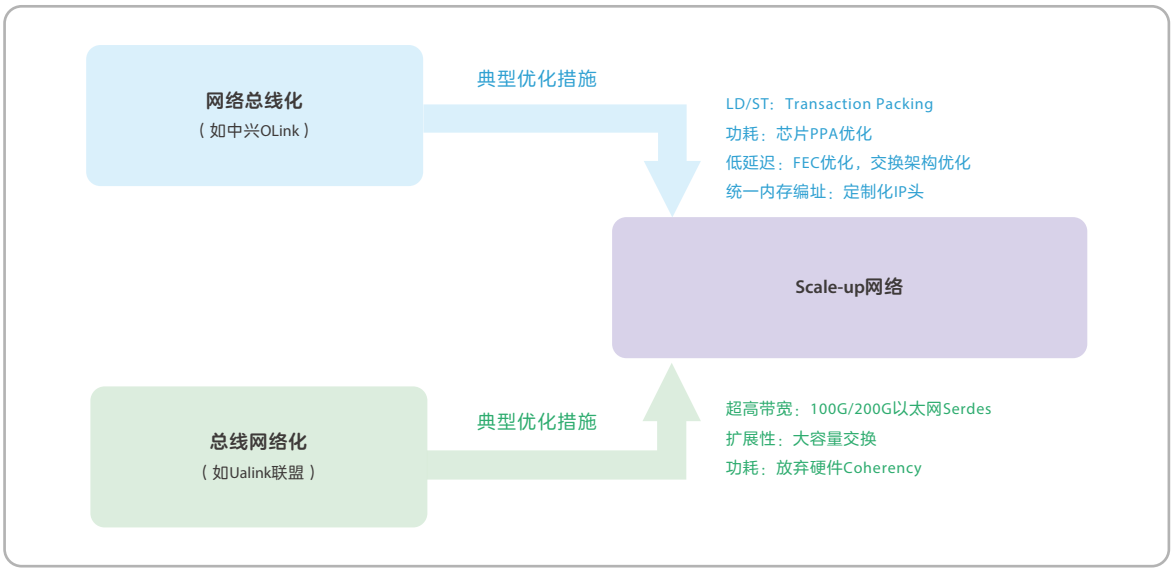


图1 Scale-up网络的核心需求



◀ 图2 Scale-up网络的两种技术路线

综上所述，Scale-up的核心需求是TB级超高带宽、k级扩展性、多语义、低功耗、百纳秒延迟以及内存统一编址等。

Scale-up网络技术路线：网络总线化还是总线网络化

鉴于Scale-up网络对未来AI基础设施的重要性，当前业界存在多种技术路线，大致可以分为“网络总线化”和“总线网络化”两种思路（见图2）。

“网络总线化”技术路线

所谓的“网络总线化”技术路线，其主流思路是在传统以太网协议以及交换技术的基础上，针对Scale-up网络的需求开展协议优化以及交换芯片架构创新，从而满足未来一段时间内Scale-up超节点组网的需求。这种技术思路常见于以太网解决方案提供商，其抓住Scale-up高带宽的首要诉求，通过成熟开放、快速发展的以太网生态解决智算“生态封闭”的问题，通过拥抱以太网高速Serdes成熟产业生态解决高带宽互联需求。然而，该技术路线通常需要采用一定的技术手段满足Scale-up网络对延迟性能、低功耗、多语义、内存统一编址等需求，这些工作存在大量的创新空间。中兴OLink是该技术路线的践行者，通过低延迟FEC、LD/ST Packing、内存统一编址、在网计算等一系列创新功能，可以实现Scale-up网络的整体诉求。

“总线化网络化”技术路线

所谓的“总线网络化”技术路线，其主流思路是以传统的总线技术为基础，摒弃一些高代价低收益的总线需求，再引入网络技术元素，满足Scale-up高带宽和扩展性需求。这种思路常见于GPU厂商或以总线技术擅长的厂商，如NVIDIA NVLink和AMD Infinity Fabric。AMD牵头成立的Ualink技术联盟是该技术路线的践行者，其协议主体更多借鉴了PCIe和CXL的设计思想，一方面结合Scale-up的特点摒弃了硬件Cache Coherency的需求，此外，通过在物理层引入以太网Serdes能力，解决了传统总线技术带宽能力不足的问题。

两种技术路线的共性和不同

从技术的角度看，无论是“总线网络化”还是“网络总线化”，其本质都是围绕Scale-up网络的核心需求开展的不同设计，最终都是一个与传统总线和传统网络都不同的特殊网络，过度

中兴OLink协议采用“网络总线化”的架构思路，提供完善的端侧IP和交换芯片的产品解决方案，并基于公共以太网技术底座实现Scale-out和Scale-up融合组网，大幅降低了网络建设和运维成本。

强调兼容“总线”或兼容“网络”都会带来一些需求或者性能的损失，因此两者并不存在绝对的优劣。

从产业推进的角度看，两种Scale-up技术路线均存在落地上的困难，均需要GPU和网络基础设施进行深度联合设计，且在此过程中，GPU通常处于相对强势的地位。两者的差异在于，“总线网络化”的技术路线更容易被GPU厂商所接受和理解，但通常需求各异难于统一，这加剧了产业落地的难度。而当前数量众多的GPU厂商已经切换以太网接口，这种情况下“网络总线化”反而更容易得到落地。

中兴OLink协议采用“网络总线化”的架构思路，提供完善的端侧IP和交换芯片的产品解决方案，并基于公共以太网技术底座实现Scale-out和Scale-up融合组网，大幅降低了网络建设和运维成本。

Scale-up网络发展趋势和产业建议

从当前AI基础设施发展趋势来看，Scale-up网络将扮演越来越重要的角色，其呈现出如下发展趋势：

- Scale-up网络规模在技术上不断突破，在应用中有序提升

未来几年，随着高速互联技术、芯片工艺、系统工程等技术不断取得突破，Scale-up的理论互联规模将从当前的百卡发展到千卡甚至数千卡的规模，但由于部署成本、边际收益、功耗密度、RAS（reliability，availability，and serviceability）等方面因素的限制，Scale-up实际落地的超节点规模将有序提升，且长期稳定在百卡左右。

- 光互联深入到芯片级，Scale-up网络架构将迎来重构契机

随着电互联逐渐触达香农定律的极限，光互联技术将深入到芯片级，以CPO/Optical IO为代表的技术成为AI芯片的重要发展方向。通过光互联可以极大提升带宽密度并降低互联功耗，服务器、交换机的产品形态和互联方式均发生重大变化，Scale-up网络将迎来架构重构。

- Scale-up和Scale-out网络在协同中融合

随着Scale-up网络和Scale-out网络技术的进一步发展，基础设施层面的融合需求将推动产品技术的进一步靠拢，当Scale-up网络和Scale-out网络在底层技术逐渐从相似走向统一时，Scale-up和Scale-out网络将逐渐融合。

当前Scale-up网络的重要性和技术路线的多样性之间的矛盾显得尤为突出，很多所谓的技术路线之争更多存在于具体的技术方案上，在要解决的核心问题上其实没有根本性的矛盾。在当前智算基础设施产业事实性垄断明显的情况下，中兴通讯呼吁产业界能更多凝聚共识，以“小步快跑”的模式推动Scale-up网络的繁荣发展。ZTE中兴

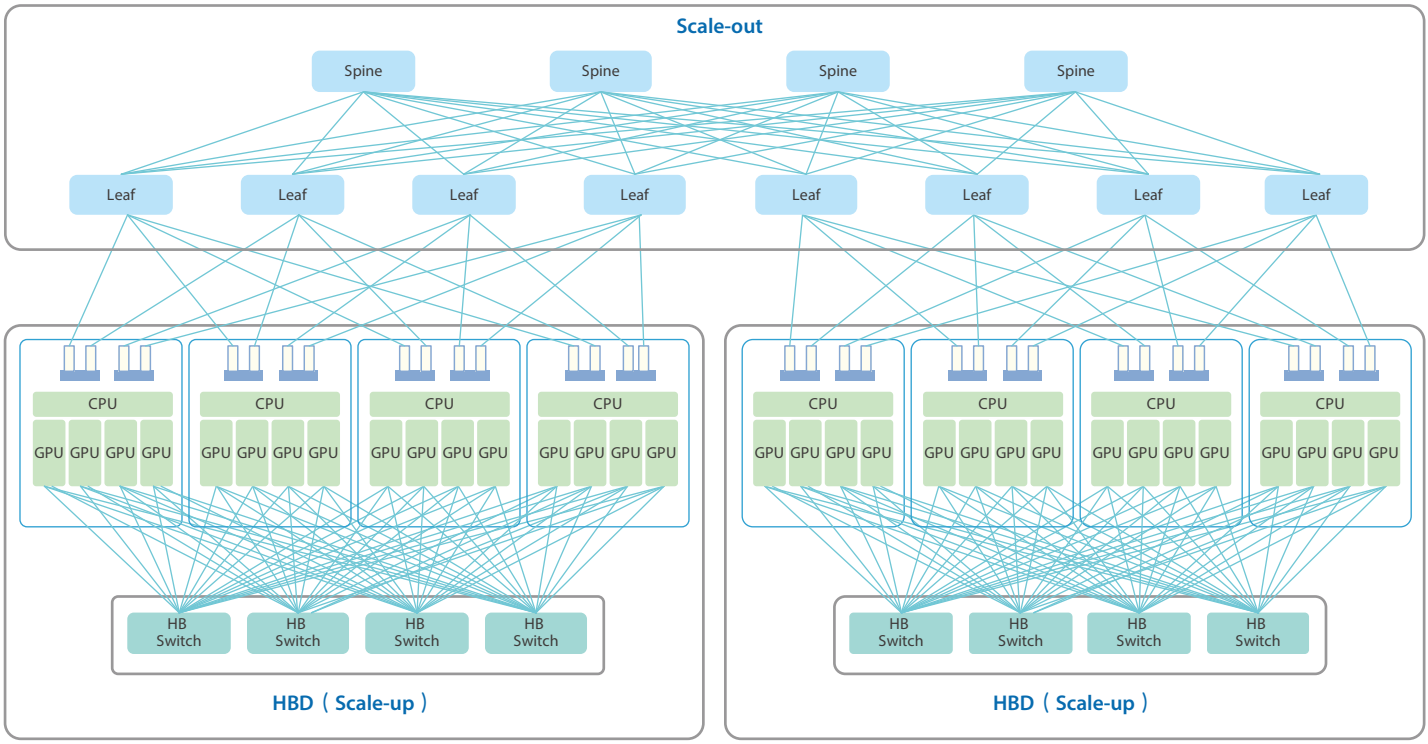
超节点技术：NVL72和ETH-X

随着人工智能技术的飞速发展，特别是AI大模型参数规模的快速增长，对计算资源的需求呈现出爆炸性增长，需要极高的算力来处理和训练，同时模型的注意力机制和前馈网络都需要大量的内存资源。最理想的方式就是开发一个超级大的GPU，具备超级大的计算能力和内存资源，由这个超级GPU完成所有大模型数据的处理。但现实上是不可能的，业界发展出超节点技术来应对这一问题。目前，在超节点技术领域，英伟达推出了基于NVLink的NVL72方案，凭借其私有协议的优势，实现了高性能的GPU互联；与此同时，ODCC（开放数据中心委员会）则基于以太网RoCE技术提出了ETH-X方案，以开放标准为基础，为行业提供更具兼容性和灵活性的选择。本文重点探讨这两种超节点解决方案的特点与应用场景，深入分析他们在高性能计算领域的价值与潜力。

潘文斌
中兴通讯数据中心网络架构师

Scale-up和Scale-out网络

为了应对大模型参数规模的快速增长，可以把大模型分解为两大类，分别处理（见图1）。一类是需要在高频度进行数据交互的，例如张量并行，把这些并行处理放置到GPU之间，通过超高带宽、超低时延互联的网络进行处理，形成一个超节点，压缩超节点内部GPU之间的通信开销



▲图1 智算网络的Scale-up和Scale-out联接

成本，这个网络就是Scale-up网络。Scale-up网络是一个追求极致性能的互联网络，支持Load/Store内存语义。另一类是将数据分解为相对独立的并行任务，如流水线并行和数据并行，这个网络就是Scale-out网络。Scale-out网络利用现有的Infiniband或RoCE网络，支持消息语义。

Scale-out网络通过网卡提供对外接口，并借助高性能、高密度的交换机组网实现节点间的互联扩展。当前，常见的组网方式包括框盒组网和盒盒组网，这两种组网方式为超节点在Scale-out方向上的扩展提供了灵活且高效的连接能力。

Scale-up网络则聚焦于超节点内部的深度互联，由GPU内部I/O与HB Switch相结合，形成all-to-all的全互联拓扑结构。在Scale-up连接的技术路线上，业界目前存在两种主要方向：基于私有协议的方案和基于标准开放协议的方案。这些技术路线旨在实现超节点内部GPU之间的高速互联，从而满足复杂计算任务对性能的极致追求。

相较于超节点之间的Scale-out网络，超节点内部的Scale-up网络具备显著的优势：更高的带宽、更低的通信时延，以及更大的缓存一致性内存空间。这些特性使得Scale-up网络能够更好地支持超节点内部密集型计算任务的需求，进一步提升整体计算效率。

英伟达NVL72

NVL72是英伟达推出的机柜级超节点，整个系统由18个Compute Tray和9个Switch Tray构成（见图2）。1个Compute Tray包含2个GB200超级

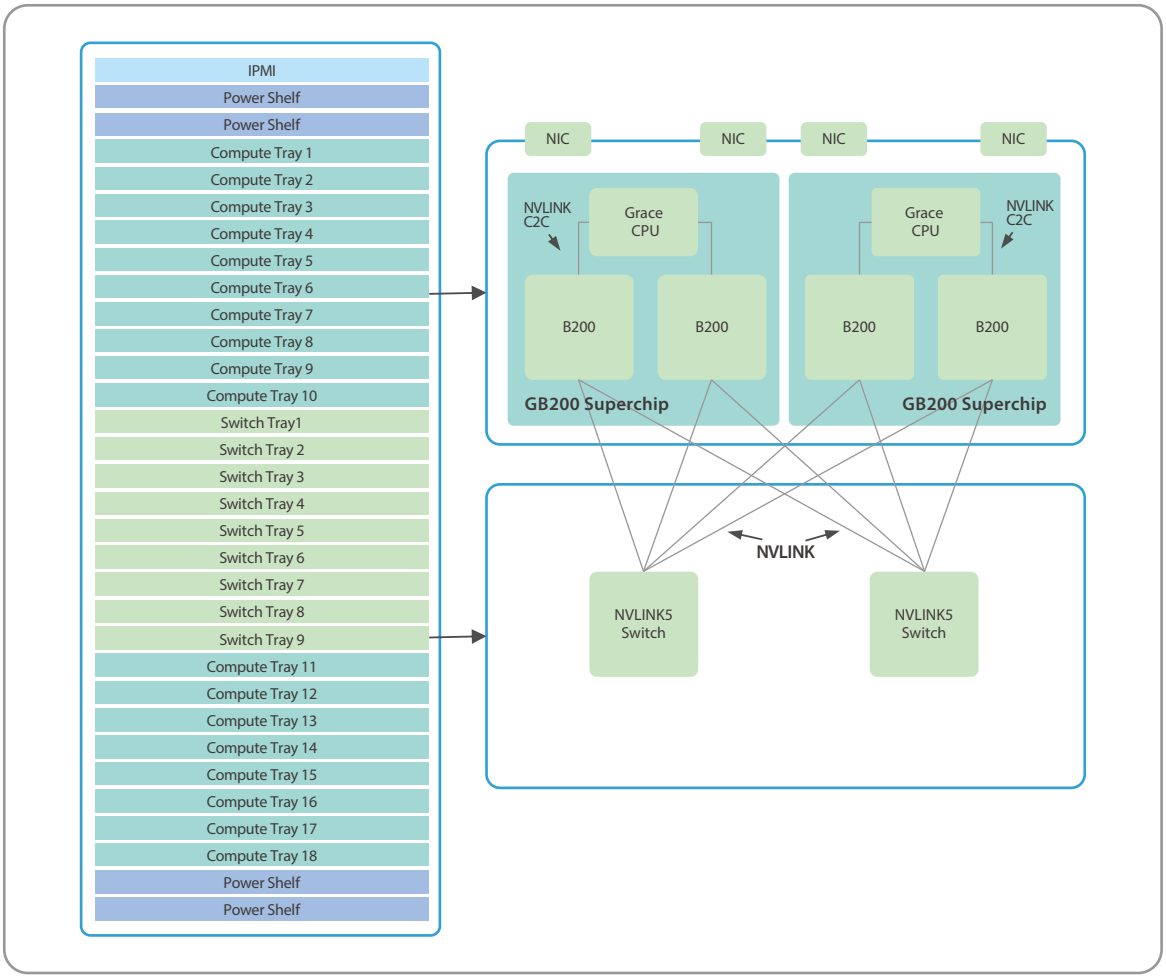


图2 NVL72整机柜示意图

芯片（Superchip），每个GB200超级芯片有2个Blackwell系列的B200 GPU，整个机柜共72个Blackwell GPU。同时每个Compute Tray提供4个网络接口卡（NIC）用于Scale-out方向的扩展。1个Switch Tray包含2颗NVLINK Switch芯片，整个机柜提供18个NVLink Switch芯片。整机柜后部通过线缆将Compute Tray和Switch Tray进行互联。

B200采用最新一代的NVLink 5连接方案，对外可提供1.8TB/s（NVIDIA采用双向计算，即单向7.2Tb/s）的NVLink连接，单个Compute Tray提供7.2TB/s（单向28.8Tb/s）带宽，NVL72整机柜的Compute Tray提供129.6TB/S的NVLink带宽。NVLink5 Switch对外可提供7.2TB/s（单向28.8Tb/s）的NVLink连接，单个Switch Tray提供14.4TB/s（单向57.6Tb/s）带宽，NVL72整机柜的Switch Tray提供129.6TB/s的NVLink带宽。这样超节点整机柜Compute Tray的GPU和Switch Tray的交换芯片之间就可以实现全连接。

B200和NVLink5采用200G的serdes，为实现B200的单向7.2Tb/s的带宽，需要72个差分对，NVL72超节点整机柜就需要5184个差分对。Compute Tray和Switch Tray通过机柜后面的线缆连接，每根线缆包含1个差分对，NVL72超节点整机柜需要5184根线缆。

NVL72通过NVLink连接将72个GPU组成一个超大Fabric网络，这个网络解决了GPU之间的高速通信带宽和效率问题，同时通过NVLink，所有GPU都可以任意访问其他GPU的内存空间。另外，英伟达还设计了NVLink C2C，B200和Grace CPU之间采用NVLink C2C连接，创建了一个NVLink可寻址的内存地址空间，B200和Grace CPU之间的内存可以互相访问。通过NVLink和NVLink C2C，每个B200 GPU可以访问超节点其他所有超级芯片的内存，包括B200和Grace CPU。每颗B200提供192GB的HBM3e内存，每颗Grace CPU提供480GB的LPDDR5X内存。这样每个GB200超级芯片提供384GB HBM内存和480GB LPDDR5X内存，NVL72整机柜支持13.5TB的HBM和17TB的LPDDR5X内存容量。

GB200超级芯片的功耗为2700W，每个Compute Tray的功耗约为6.3kW，每个Switch Tray功耗超过800W，NVL72整机柜的功耗预计达到120kW，采用冷板液冷进行散热。

考虑到实际机房提供120kW机柜能力的难度，英伟达还支持规格减半的NVL36。有两种方案：

- Switch Tray结构不变，Compute Tray同样也是有2个GB200超级芯片，包含4个B200和2个Grace CPU，但尺寸改为2U，整个NVL36超节点的Compute Tray数量减半、GPU数量减半，Switch Tray可以有一半的带宽（28.8Tb/s）用于对外连接扩展；
- Switch Tray结构不变，Compute Tray尺寸不变保持为1U，但GB200超级芯片包含的B200数量减为1个，整个NVL36超节点的GPU数量减半，Switch Tray可以有一半的带宽（28.8Tb/s）用于对外连接扩展。

方案2可以提供更大的LPDDR内存空间，但价格会更贵。两个NVL36超节点之间通过ACC线缆互联，同样可以提供72卡的计算能力。通过L2 NVLink Switch进行16个NVL36超节点互联，可以完成Scale-up方向NVL576的扩展，提供576卡的计算能力。

我们来看NVL72怎么满足Scale-up网络的特性的。

- 高带宽

NVL72的每个B200 GPU提供7.2Tbps的Scale-up连接带宽，同时通过PCIe对外提供400Gbps的Scale-out连接带宽，Scale-up带宽是Scale-out带宽的18倍。

- 低时延

英伟达官方没有提供NVLink Switch的转发时延具体数据，但以低时延作为一个卖点，同时从设计上充分考虑低时延的架构。Switch Tray和Compute Tray之间采用的是电缆连接，这样可以节省因CDR或DSP引入的将近100ns的时延，同时也降低了成本。

- 大内存空间

NVL72利用NVLink和NVLink C2C，所有GPU都可以访问整个超节点其他GPU的HBM内存和Grace CPU的DDR内存，NVL72整机柜支持13.5TB的HBM和17TB的LPDDR5X内存容量。

ODCC ETH-X

由中国信通院、腾讯在ODCC牵头发起的ETH-X项目可以支持单个超节点64卡的计算能力，和英伟达的私有NVLink方案不同，ETH-X采用更为开放的RoCE方案。

整个系统有16个Compute Tray和8个Switch Tray。每个Compute Tray包含4张GPU和1个X86 CPU，CPU和GPU之间通过PCIe Switch对接。整个机柜共64张GPU。同时每个Compute Tray提供4个NIC用于Scale-out方向的扩展。每个Switch Tray包含1颗支持RoCE的高性能51.2Tbps以太网交换芯片，整个机柜提供8个Switch芯片。GPU和Switch芯片支持100G serdes。当前主流的GPU互联带宽为3.2Tbps，ETH-X整机柜GPU互联带宽为204.8Tbps。8个Switch Tray支持409.6Tbps的带宽，一半用于超节点柜内连接GPU，另一半的带宽用于背靠背连接旁边机柜的超节点或者通过L2 HB Switch做更大的HBD域Scale-up扩展。对于Intel Gaudi3 GPU，可以提供4.8Tbps的带宽，因此超节点机柜需要12个Switch Tray。同时，ETH-X也支持Switch Tray没有外部Scale-up扩展口的方案，所有serdes连接都用于柜内互联，这时候只需要4个2U高的Switch Tray（Gaudi3为6个）。

▼表1 NVL72和ETH-X关键指标对比

	GPU数	Serdes	GPU-CPU互联协议	GPU-Switch互联协议	GPU单芯片互联带宽	Switch单芯片带宽	差分对数量
NVL72	72	224Gbps	NVLink C2C	NVLink	7.2Tbps	28.8Tbps	5184
ETH-X	64	112Gbps	PCIe	RoCE	3.2Tbps/4.8Tbps	51.2Tbps	4096/6144

ETH-X对Scale-up网络特性的支持情况：

- 高带宽

ETH-X的每个GPU提供3.2Tbps（或4.8Tbps）的Scale-up连接带宽，同时通过PCIe对外提供400Gbps的Scale-out连接带宽，Scale-up带宽是Scale-out带宽的8~12倍。

- 低时延

ETH-X没有限定Switch Tray的芯片型号，可以采用Broadcom的Tomahawk5，也可以采用Marvell的Teralynx10，甚至还可以采用国产化的25.6T芯片2片进行设计。总体来说，Scale-up方向的Switch时延控制在纳秒级是大家的一个共识。同时ETH-X也借鉴了NVIDIA NVL72的经验，Switch Tray和Compute Tray之间采用的是更低成本和更低时延的电缆连接。

- 大内存空间

NVIDIA NVL72通过GPU-Switch-GPU的NVILink实现统一内存地址空间，通过GPU-CPU的NVLink C2C实现缓存一致性。而ETH-X的GPU-Switch-GPU之间为RoCE连接、GPU-CPU之间为PCIe连接，需要进一步的开发互通协议，向应用层提供支持Direct Copy、Direct Access以及UVA统一编址等内存语义，实现GPU之间的访存协议。

总结和展望

NVL72和ETH-X超节点都可以提供高带宽、低时延、大内存空间的Scale-up网络扩展。NVL72方案采用NVLink和NVLink C2C连接，超节点内的GPU之间的内存都可以互访。ETH-X采用开放的以太网解决方案，优点是生态开放，可以推广为ODCC组织的一个标准，不过由于没有NVLink这种类总线的协议，ETH-X后续还需要进行内存语义支持的开发。两种超节点的关键指标对比如表1所示。

NVL72凭借其先发优势，在国外OTT大厂中获得了较多的订单，展现出强大的市场竞争力。然而，它也存在一定的局限性，其基于私有协议的生态体系相对封闭，可能在一定程度上限制了更广泛的行业协作与创新。

ETH-X作为开放标准，在进度上稍落后于NVL72，这主要是由于公开标准的制定过程需要投入大量时间和精力。这一过程中不仅涉及复杂的技术讨论，还需在标准成员间进行多方面的协调与博弈，包括技术细节、商业利益以及战略方向等非技术因素。尽管如此，开放标准的特性为ETH-X带来了广阔的潜在应用空间和行业包容性。

独行快，众行远，NVL72和ETH-X作为当前超节点技术的两大代表，各自展现了独特的魅力。在未来的发展中，我们相信这两种技术将在各自的生态系统中绽放异彩，共同为超节点技术的发展书写精彩篇章。ZTE中兴

高性能以太网
助力智算中心长距互联

李连华
中兴通讯交换机产品规划经理

从2023年开始ChatGPT爆火，在大模型训练场景下，随着参数规模从亿级提升到万亿级别，算力需求也出现爆发式增长。据统计，2012—2022年模型算力需求每年约增长4倍，而2023年后模型算力需求预计会以每年10倍的速度增长。

算力需求的指数级增长对AI基础设施带来极大挑战，当GPU算力增长速度低于算力需求增长时，组建超大规模GPU智算中心就成为必然趋势。由于机房空间、电力、机房散热等问题限制，智算中心建设在达到一定规模后存在单点物理中心规模受限的问题。因此在构建万卡甚至十万卡集群时搭建智算拉远分布式智算中心成为一种必然选择，智算中心长距互联需求应运而生。

训练拉远和存储拉远

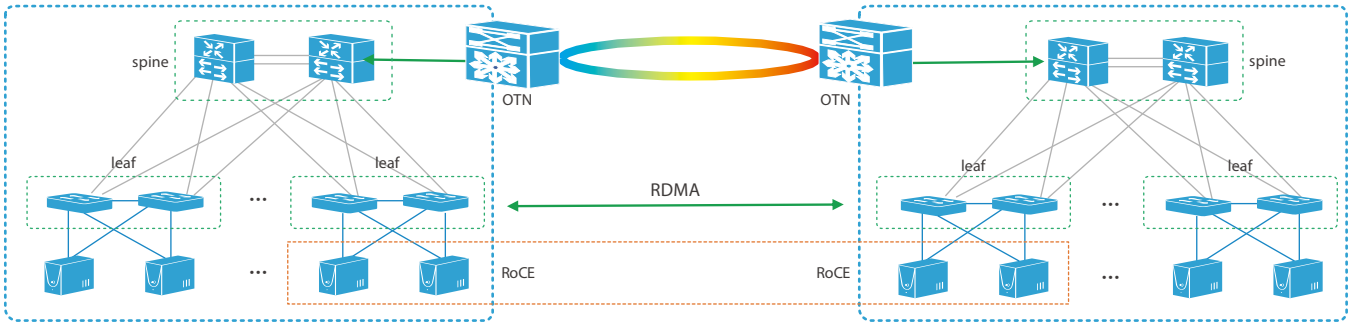
目前，智算中心长距互联主要分为两类场景：训练拉远场景和存储拉远场景。

- 训练拉远场景

国内智算中心现状为，单点智算中心规模普遍偏小，存在零散、欠规划、算力闲置等问题，智算中心拉远互联可以将多个已经建成的智算中心的算力进行整合。且未来长时间内，国内GPU算力能力要低于国外最先进制程，相同算力国内需要更多数量GPU，通过智算互联进行分布式训练，可以弥补单智算中心算力不足的问题。在训练拉远组网时，一般以OTN设备作为长距链路底座，直连不同智算中心，如图1所示。

- 存储拉远场景

在智算中心算力出租或智算中心长距互联后，算力训练处理过程中会存在部分数据样本安全性要求较高的数据不便外迁，以及智算中心互联数据样本跨中心调度的需求。存储拉远场景可以将计算集群和存储集群无损互联，满足数据本地化需求，保障数据安全高效。存储拉远组网时



▲图1 智算中心训练拉远组网

依据用户实际需求，可灵活选用组网方案，保障网络无损传输数据即可。

无损以太网保障长距互联无丢包

为充分释放GPU集群算力，智算中心网络需要搭建高吞吐、高可靠的RoCE（RDMA over converged ethernet）网络，充分释放AI生产力。随着GPU算力增长，网络带宽从以前的100Gbps到现在的200/400Gbps，以及未来的800Gbps，都是为了实现端口高吞吐。中兴通讯承载交换机与OTN产品依托自主芯片在高性能以太网路线上持续演进，为未来800G Fabric网络提供平滑升级方案。

根据智算中心长距互联网络特点，中兴通讯承载数据中心交换机及OTN产品线联合推出ZXR10 9900X加光传输跨智算中心高速互联无损以太网解决方案，完美匹配AI算力互联需求特点，保障智算中心网络高吞吐、高可用。中兴通讯承载产品针对智算互联场景进行了多项技术创新，包括ZRLB负载均衡技术、长距PFC流量控制技术、智能主动拥塞告知（IPCN）等技术。

ZRLB负载均衡技术

无收敛的AI智算网络中，常遇到多入口多出口的复杂场景。此类场景中，我们希望来自多个入口相同目的地址的流量均匀地分发到多个出口。然而在实际环境中，往往会出现多个入口的流量汇聚到某一个出口，而其他出口上没有流量的情况，导致负载分担不均衡和网络拥塞。传统Hash算法基于五元组信息进行逐流Hash，在AI智算网络中出现Ring/HD等点到点流数少、单流带宽大的情况时极易出现Hash同质导致链路负载极化从而造成部分链路重载、部分链路轻载，产生丢包、影响整体网络吞吐率。

中兴通讯推出ZRLB（ZTE Rail Load Balance）负载均衡技术，通过对端口进行分组和智能编排，基于设备连接关系配置接口组形成入口和出口之间的一对一映射关系，发往同一台或同一类型设备的流量在接口组成员间基于接口ID进行Hash，将流量精准地负载分担到不同的出口以提升网络吞吐率。

在智算中心长距互联场景，ZRLB功能通过实现数据中心内部负载均衡，从而一定程度解决长距链路拥塞，合理利用宝贵OTN带宽的同时避免将数据中心内的不均衡问题传导到数据中心外。

长距PFC流量控制

PFC（priority-based flow control）通过逐跳提供基于优先级的流量控制实现整网链路的无丢包功能，是构建无损以太网的必选手段之一。数据中心间网络通常具有几十公里的链路长度，带来的固定往返时延往往是毫秒级别的，这使得普通PFC和ECN（explicit congestion notification）机制在该场景下会产生巨大的响应延时，从而导致对拥塞的控制不够及时，无法应对网络流量的变化。特别是，如果采用普通浅缓存网络设备，在PFC触发到生效，长距链路上产生的飞行流量很可能撑爆缓存区空间造成丢包，同时从取消PFC到上游流量再次到达，也会因缓存流量不足而造成吞吐损失。因此数据中心间网络实现无损数据传输就对网络设备提出更高的性能要求。

以网络设备400GE接口100km光纤链路为例，100km单向会引入0.5ms固定传输时延，网络设备产生拥塞通过PFC通知对端设备停流的周期内需要预留1ms的飞行报文缓存空间以防止链路丢包，当接口带宽为400GE时，所需缓存大小约为47.68MB，所需缓存空间已超过部分ASIC芯片的整机芯片最大值。业内对该场景的应对方式一般为采用高性能芯片、外置HBM（high bandwidth memory）或增加智算网关等方式，以满足数据中心间的无损数据传输。

中兴通讯数据中心交换机ZXR10 9900X系列设备通过芯片大缓存解决方案，支持长距离智算互联场景，为跨智算中心训练业务提供技术保障。

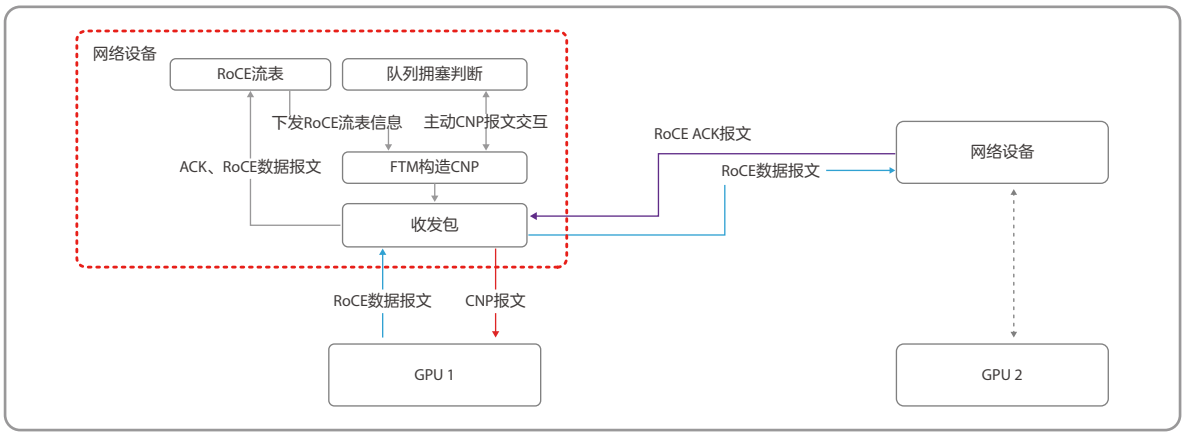


图2 IPCN工作原理

IPCN解决DC内拥塞

ECN通过网卡进行流速控制从而调整网络拥塞，这强依赖于网卡的实时响应。若将两台网卡距离拉远至2个长距离数据中心，则传统的数据中心内的DCQCN算法就不再适用，无法对网络标记的ECN报文进行快速响应避免拥塞加剧的问题。

IPCN（intelligent proactive congestion notification）功能支持在网络设备上智能识别拥塞状态，主动发送CNP报文，准确控制服务器发送RoCEv2报文的速率，既可以确保拥塞时的及时降速，又可以避免拥塞已经缓解时的过度降速，最终确保数据中心互联这种长距场景中RoCEv2业务的低时延和高吞吐。IPCN的工作原理图2所示，网络设备上启用IPCN功能的接口会对过路的RoCE报文进行分析并建立流表。接口根据队列的拥塞状态向发送端服务器主动发送CNP报文，服务器收到CNP后降低数据报文的发送速率达到缓解网络拥塞的目的。

中兴通讯数据中心交换机ZXR10 9900X、ZXR10 5960X系列设备采用自研芯片，通过自主可控硬件支撑IPCN功能，为跨智算中心训练业务提供全网络技术保障。

智算拉远网络技术展望

针对智算拉远无损技术，RoCE网络需更多地参与到流量控制和拥塞控制机制优化中。未来智算拉远网络中，下面几个技术方向是可能的发展趋势：

- AI ECMP：基于流量基线信息，综合流带宽统计、ECMP路径带宽以及流的生命周期，选择最优路径，下发到设备转发面。在智算拉远时设备可以响应流量动态变化，及时保证数据中心出口的ECMP路径负载均衡。
- AI长距主动性流控：AI算法在识别不同流量模型后结合本地缓存流量数据与历史信息，通过预测性的估计来预测未来网络流量的拥塞状态，从而对网络流量进行提前控制，使数据传输达到最优。
- 精细化流控：数据中心内智算流量和长距智算流量的响应时间和流量特性是有所区分的，精细化流控技术可以准确区分数据中心内流量和长距流量，从而精准控制长距链路数据报文，达到整网流量的性能最优。
- 全维度负载均衡：业内已有整网全局负载均衡技术，如iGLB（intelligent global load balance）等，当前该类技术仅局限在数据中心内部使用。在智算拉远环境中，通过调整算法的相关参数，可以做到长距跨数据中心的整网负载均衡，将长距链路的拥塞情况提前规划到最低。

未来中兴通讯会紧跟时代步伐不断创新方案，与合作伙伴一起探索智算中心建设发展路径，在实践中不断取得突破，助力AI应用全面落地。ZTE中兴

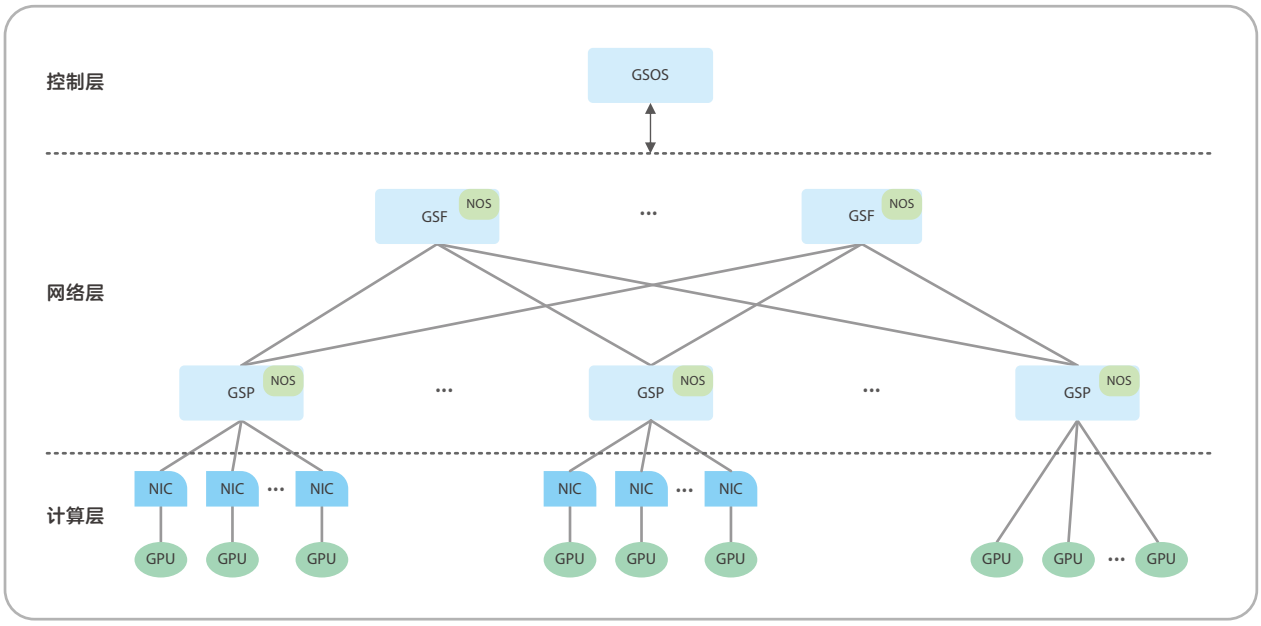
GSE关键技术及产品实现

近年来，AI技术在自动驾驶、语音识别、自然语言处理等领域取得了显著成就，推动了算力经济的高速发展。智算中心作为AI技术的核心支撑平台，对网络的性能要求极高。然而，传统以太网在面临大规模、高速度的数据传输时，存在网络拥塞、丢包和延迟增大等问题，难以满足智算中心对高性能网络的需求。因此，业界迫切需要一种新型网络技术来突破传统以太网的性能瓶颈。

GSE（Global Scheduling Ethernet）技术是由中国移动联合产业合作伙伴共同提出的一种以太网技术架构，旨在突破智算中心网络性能瓶颈，打造无阻塞、高带宽及超低时延的新型智算中心网络。GSE技术通过革新以太网基础转发机制，对以太网协议栈和配套通信产品产业进行革新，实现了对大规模AI训练和推理需求的全面支持。

GSE技术原理与关键技术

GSE技术架构主要包括控制层、网络层和计算层三个层级（见图1）。控制层包含全局集中式全调度操作系统（global scheduling operating system，GSOS）以及设备端分布式节点操作系统（node operating system，NOS）。GSOS提供整网管控的集中式网络操作系统能力，实现基于全局信息编址、日常运维管理等功能；设备端NOS可实现动态负载均衡、动态全局调度队列（dynamic global scheduling queue，DGSQ）调度等分布式网络管控功能。网络层主要通过网络设备的分工协作，构建出具有全网流量有序调度、各链路间负载均衡、网络异常精细反压等技术融合的交换网络，由GSF（global scheduling fabric）和GSP（global scheduling processor）



▲图1 GSE整体架构

吕勇
中兴通讯有线芯片架构师

两类设备组成。计算层则包含高性能计算卡（如GPU或CPU）及网卡，通过优化交换网络能力提升计算集群训练性能。

GSE关键技术包括基于报文容器的转发及负载分担机制、基于DGSQ的全局调度技术、GSOS的集中管理及分布式控制等。

基于报文容器的转发及负载分担机制

智算中心网络通常采用胖树（fat-tree）架构，任意出入端口之间存在多条等价转发路径。然而传统以太网逐流负载分担方式在以低熵大象流为主的智算中心网络中，会因为Hash极化导致链路利用率不均，从而引起网络拥塞。为了解决这个问题，GSE技术提出了一种基于报文容器（packet container，PKTC）的转发及负载分担机制。该机制根据最终设备或设备出端口，将数据包逻辑分组，并组装成长度较长的“定长”容器进行转发。属于同一个报文容器的数据包被标记为相同的容器标识，沿着相同路径转发，以保证同属于一个报文容器的数据包保序传输。

在负载均衡调度时，报文容器被作为转发单位。由于报文容器是逻辑组装，无需额外的硬件开销来对数据包进行组装和还原，不增加转发的延时，而报文容器为等长的，可以保证在多条等价转发路径上的流量负载是相等的，可在网络上实现最理想的负载均衡效果。

基于DGSQ的全局调度技术

DGSQ是GSE技术的核心组件之一，它通过接收侧授权和动态队列实现了对整网流量全局调度和整网所有设备缓存的全局管理。在GSE网络中，每条流的首个报文在进入网络后都会在GSP为该流动态分配队列，并构建虚拟的报文容器，报文在出队前需向接收侧的目的GSP请求授权，目的GSP根据该流对应的出接口的负载情况，给予授权响应，实现将出接口的带宽在所有源之间进行动态分配。

通过DGSQ动态队列和接收侧授权技术的组合，GSE网络实现了精细的反压机制，当网络出现拥塞时，DGSQ会及时通知源端降低发送速率，避免网络拥塞的进一步恶化。这种精细的反压机制可保证全网中前往任何一个端口的流量既不会超过该端口的负载能力，也不会超出中间任一网络节点的转发能力，同时可将造成拥塞的流量分布式地缓存在源侧，实现全网所有设备缓存的全局管理。

在以Alltoall/AllReduce集合通信为主的智算中心网络中会大量产生Incast流量，GSE的机制对于持续的Incast可通过精细反压减少全网反压的产生，对于突发的Incast通过利用全网多个设备的缓存进行吸收而保持高吞吐，使采用GSE技术的智算中心网络能够提供稳定的吞吐率和网络延迟。

GSOS的集中管理及分布式控制

GSOS是GSE技术的核心控制平面，它提供整网管控的集中式网络操作系统能力。GSOS通过收集全网资源使用情况、设备状态等信息，实现对网络的全面监控和管理。同时，GSOS还支持基于全局信息的流量调度、负载均衡、故障恢复等功能，确保网络的稳定性和可靠性。

设备端NOS则实现了分布式网络管控功能，它负责本设备容器的负载均衡管理、DGSQ调度等属于设备自身的网络功能。通过设备分布式管控能力，NOS可以提升整网可靠性，降低网络故障的影响范围。

GSE技术的产品实现

GSE技术的产品实现主要为在传统以太网交换机产品的基础上，增加基于报文容器的转发及负载分担机制、基于DGSQ的全局调度机制和快速故障感知和故障收敛机制，从工作模式上可分为GSP设备和GSF设备，与GSOS/NOS协同实现高可靠的高性能网络。

GSP设备实现

GSP设备是GSE网络中的边缘处理节点，它

负责对计算层数据包的接收、处理和转发。GSP设备需要支持基于报文容器的转发及负载分担机制，能够按照DGSQ的调度结果进行流量调度。同时，GSP设备还需要支持精细的反压机制，能够根据网络拥塞情况及时通知计算层调整发送速率。

与传统以太网交换机相比，GSP设备在TM（traffic management）上有较大的变化，主要体现在队列的管理和队列调度上。传统以太网交换机的TM队列是静态绑定到接口上，队列的调度是基于队列绑定的本地接口的状态。GSP设备的TM队列由流触发动态生成，队列的调度需要基于队列中流在网络中最终的出接口（可能位于远端）的状态，该状态是通过授权请求和授权响应反馈到源端GSP。

GSF设备实现

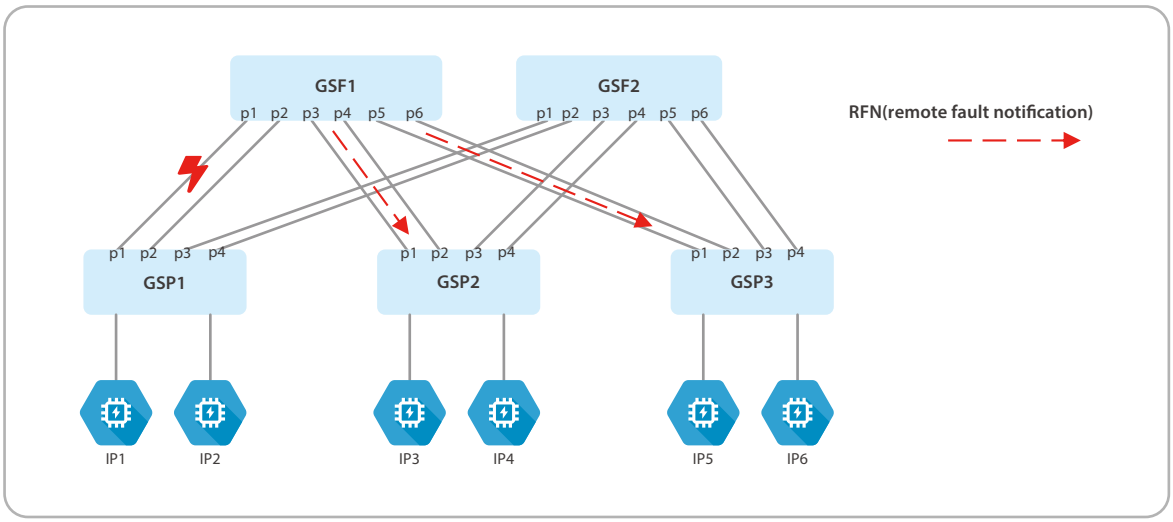
GSF设备是GSE网络的核心交换节点，负责构建融合全网流量有序调度和链路负载均衡的交换网络。它需支持高速数据交换和基于报文容器的负载均衡。与传统以太网交换机不同，GSF设备的转发引擎从无状态改为有状态转发。报文路由不仅由目的地址决定，还需参考所属虚拟容器中前序报文的选路结果，确保同一容器内的报文选择相同路径转发。采用有状态转发时，报文转发需增加读取和更新状态的操作。大容量交换设备需采用多个转发核实现高性能，而有状态转发要求在多个转发核上保持状态同步。GSF设备通过设计多端口RAM来实现状态的更新和同步。

故障感知与故障收敛

传统以太网交换机故障感知是事后感知，需检测到丢包后才能发现故障，故障感知时间通常为毫秒级，无法满足智算中心需求。GSE设备利用以太网物理层的FEC机制，统计误码数量，在丢包前预测链路质量，将故障感知从事后变为事前，使得故障对网络的影响最小化。如图2所示，感知故障后，GSF设备利用其在fat-tree拓扑中的根位置，主动向所有GSP设备发送故障通告，故障通告消息借助以太网O码扩展实现，不占用业务带宽，快速通告故障，快速完成故障收敛。

GSE技术的发展前景

随着AI大模型参数量呈指数级增长，智算中心对网络性能的需求已从“高性能”向“超性能”跃迁。传统以太网在应对千卡级GPU集群训练任务时，往往因网络时延抖动、链路利用率不足60%等问题，导致算力效率损失超过30%。GSE技术通过颠覆性架构设计，为解决这一全球性难



◀ 图2 GSE故障通告

GSE技术的应用边界正从智算中心向更广域延伸，作为超大规模算力网，在国家“东数西算”工程中，GSE的“逻辑长容器”技术可有效解决跨区域数据中心间带宽波动问题。

题提供了中国方案，其后续发展将从技术深化、生态扩展、产业融合、政策驱动几个维度展开。

技术深化：从理论突破到场景适配

GSE核心技术的持续迭代将围绕两大方向：智能调度算法升级和全场景覆盖能力。

- 智能调度算法升级：当前DGSQ机制已实现微秒级动态授权，未来拟引入AI驱动的流量预测模型，通过分析历史通信模式（如AllReduce/AllGather操作特征），预判突发流量并动态调整容器调度策略，目标是将授权响应速度提升至百纳秒级。
- 全场景覆盖能力：针对多模态AI训练中混合流量（视频、文本、参数并行）特性，研发差异化容器构建策略。例如，对时延敏感的小包（如梯度同步）采用“小容器+高优先级”，对大包（如模型参数）采用“巨型容器+带宽保障”，实现网络QoS精细化控制。

生态扩展：构建自主可控技术体系

GSE正加速形成从标准到产品的全栈生态。在标准体系构建方面，中国移动联合华为、中兴通讯等企业发布《智算中心全调度以太网技术白皮书》，推动CCSA（中国通信标准化协会）制定GSE行业标准，并积极向ITU-T国际标准演进。智算集群实践方面，中国移动采用GSE技术改造的GPU集群测试显示，性能相比原生RoCE网络平均提升52.6%，相比优化后的RoCE网络，平均提升26.6%。BERT-Large训练中网络零丢包，单任务节约成本超百万美元。

产业融合：赋能新型数字基础设施

GSE技术的应用边界正从智算中心向更广域延伸，作为超大规模算力网，在国家“东数西算”工程中，GSE的“逻辑长容器”技术可有效解决跨区域数据中心间带宽波动问题；作为6G算力网络，GSE动态队列管理机制和其微秒级故障感知能力，可支持空天地一体化网络中星间链路的快速自愈。

政策驱动：数字中国战略下的发展机遇

我国《“十四五”数字经济发展规划》明确要求数据中心PUE<1.3，而GSE技术通过全局流量调度可使网络能耗降低。通过产业合作伙伴的共同推动，有望将GSE列入新基建重点项目技术目录，预计到2027年将带动超千亿规模的智算网络改造市场。

可以预见，随着GSE专用设备量产、GSOS智能管控系统与Kubernetes等编排平台深度融合，GSE技术有望成为智算网络的“操作系统”，在实现网络传输零损耗、算力释放100%的终极目标中发挥核心作用。这一技术路径不仅关乎企业算力成本竞争，更是国家在AI 2.0时代掌握算力基础设施主导权的战略基石。ZTE中兴

中兴通讯助力快手

打造数据中心无损互联网络

作为全球知名的短视频社交平台，快手拥有海量用户数据和复杂业务场景。为更好地应对业务增长压力，提升自身算力水平，满足短视频、直播等核心业务以及人工智能、大数据等新兴业务的发展需求，快手积极布局人工智能和大数据算力中心建设。

数据中心网络作为算力互联的“神经系统”，如何实现服务器、存储设备以及应用系统之间快速、稳定、高效的数据传输，是支撑各类数字化业务正常运转的关键所在。快手与中兴通讯深入合作，坚持数据中心交换机产品软硬件自主创新，共同助推国产智算网络技术能力不断升级。

自主可控，生态开放

快手坚持“软件自主可控，硬件生态开放”的技术战略，与中兴通讯携手打造满足全场景业务发展的新一代数据中心交换机产品。

软件方面，快手积极拥抱开源社区，以SONiC为基础构建了开放架构的网络操作系统平台KNOS（Kuaishou Network Operating System）。依托KNOS，快手与中兴通讯携手研发了一系列数据中心场景关键技术特性，如远程直接内存访问（RDMA）、非等值负载分担（UCMP）、双向转发检测（BFD）/链路时延（Link-Delay）、ISIS协议、长距光模块（ZR）等，同时配套全业务场景统一的网络管理平台KNP（Kuaishou Network Platform），实现了端到端高性能、智能化网络自动规建维优。

硬件方面，中兴通讯ZXR10 5960X/M数据中心交换机有2T/8T/12.8T/51.2T多个产品机型，可覆盖快手通算、智算、存储、管理等多个大规模数据处理和复杂业务场景。交换机系列产品基于中兴通讯自主研发的硬件平台，产品性能优异。以51.2T盒式交换机为例，4RU紧凑高度即可搭载128个400G QSFP112端口，其中心交换、接口单

韩云霞
中兴通讯数通产品规划总监

元、主控单元等均采用模块化设计，模块间采用高速SLIMSAS总线互联，可靠性极高；创新式的两层PCB板设计，不仅节省了1块高速PCB板材和装配，双层固定扣板结构更保障了112Gbps速率信号的稳定传输；主控单元的CPU模组设计采用OCM标准，支持BMC（baseboard management controller）进行外设管理，盒体前面板可拆卸，支持多元化交换芯片和接口板，可适配不同端口形态的机型。

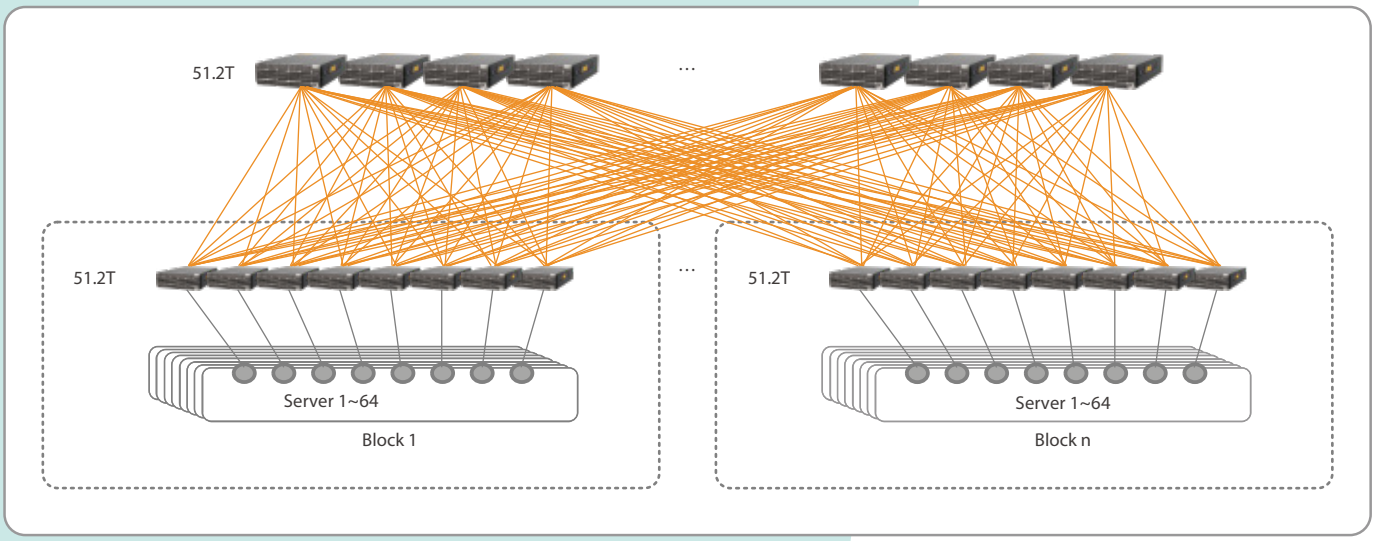
这些软硬件创新技术全面覆盖了快手DCN数据中心网络、HPN高性能智算网络、DCI城域网、CDN等主要网络场景，极大地提升了网络运营效率与稳定性保障能力，助力快手数据中心网络向智能化、高效化迈进。

全栈智算，超宽无损

基于自主可控、开放解耦的软硬件产品平台，快手构建了面向AI智算的全新一代数据中心网络架构。方案深度整合RoCEv2端到端无损以太技术，以中兴通讯51.2T交换机作为核心交换节点，成功构建了万卡级大模型集群网络（见图1），率先在国内实现400G RoCEv2高性能网络的部署应用。

在该万卡智算集群，快手携手中兴通讯进行了一系列领先无损网络技术创新：

- RDMA：利用KNOS中的RDMA相关特性，与统一网络管控平台KNP紧密协作，构建了一套集功能支撑、精细化网元监控、带内遥测可视化、自动化管控调度、流量调优等功能于一体的端到端高性能网络解决方案，全方位提升RDMA网络的带宽吞吐和时延，将网络性能发挥到极致，全方位提升系统性能。
- UCMP：创新采用了UCMP（unequal cost multiple path）协议及动态负载分担功能，根据实时可用带宽比例智能调整流量分配，有效缓解链路故障下的拥塞丢包问题，增强了网络的稳定性和灵活性，使自研交换机能在更多复杂网络场景中成功部署与应用。
- 无损热升级：基于KNOS系统通过无损热补



▲ 图1 快手万卡级大模型智算集群

> 基于自主可控、开放解耦的软硬件产品平台，快手构建了面向AI智算的全新一代数据中心网络架构。

丁设计技术，实现了对交换机软件的快速、无损升级，覆盖了全场景运维需求，确保在业务无感知的情况下完成软件修正与功能增强。

- 网络丢包检测（MOD）：实时捕捉并分析硬件层面各类常见丢包事件，精准记录丢包原因及被丢弃报文的关键特征，随后将这些宝贵信息传输至采集器。极大缩减故障排查时间，为数据中心网络的稳定运行提供了坚不可摧的保障
- 网络可视化/带内遥测（INT）：集成先进的带内遥测（INT）技术，交换机网元在数据包流转间巧妙嵌入核心运行数据，实现状态与数据的同步传递。沿途设备接力标注，最终汇聚至监控分析中心，通过深度数据挖掘与拓扑融合，为运维人员呈现报文全路径视图与端到端时延细节，助力网络性能优化决策更加精准。

上述方案，不仅与业界传统IB网络方案在性能上并驾齐驱，更实现了成本的大幅削减，实现了基于以太网的全栈增强无损网络方案规模商用部署落地。

全域覆盖，智引未来

除了大模型智算数据中心，中兴通讯开放解耦的全系交换机产品规模部署全场景数据中心网络。基于全自研框盒式交换机构建的数据中心网络，两层CLOS即可轻松驾驭数十万台服务器的接入需求，其容量之巨，较上一代产品实现了质的飞跃，同时前瞻性地兼容了100GE/200GE/400GE服务器的接入，确保了技术投资的长期价值。

历经3年的稳健发展，快手与中兴通讯的深度合作在新技术新产品的研发与落地方面取得了令人瞩目的成就，依托自主研发与快速迭代能力，持续推动交换机向更大带宽、更高容量的极限挑战，打造了行业开放生态合作的全新范式。未来，快手将继续加强与中兴通讯的联合研发，持续创新INT/SDN、端网融合、在网计算等前沿技术，满足新一代AI/大模型算力对数据中心网络的极致带宽与超强无损需求，为数字经济时代的快速发展提供更强大的网络支撑。ZTE中兴

中兴通讯助力南京电信智算中心基础设施项目建设，打造AI智能新时代

崔健
中兴通讯有线产品规划总监

生成式人工智能发展迅速，催生出大量新场景、新模式、新生态，引发算力需求爆发式增长。建设智算资源池可以为人工智能大模型训练、推理等工作提供强大的基础设施支撑，满足AI训练、推理需求。

南京吉山云计算中心，作为全国一体化算力网络长三角国家枢纽节点的核心，紧密对接国家“十四五”规划，旨在通过大数据、AI等前沿科技提供算力服务，打造高性能算力服务平台，同时整合本土生态产业链企业，共同培育“算力+生态”应用环境。南京电信携手中兴通讯共同打造了集约高效、安全可靠的国产化千卡智算资源池，为千行百业提供强大的算力支持，赋能数字经济高质量发展。

集约高效，安全可靠

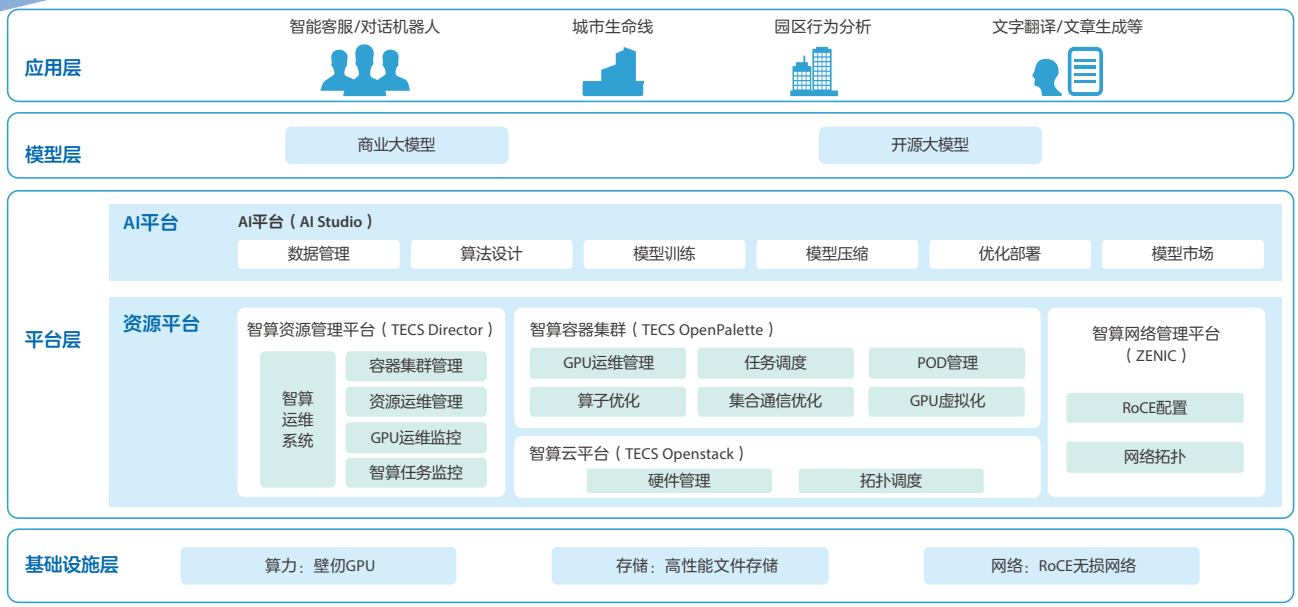
资源池采用“集约高效、共享开放、安全可靠、按需服务”的理念设计，中兴通讯量身打造了业界领先的AI基础设施及平台软件，包括中兴通讯智算服务器、国产化OAM卡、中兴通讯自研RDMA交换机、自研AI平台（AIS）和资源管理平台（TECS），全面覆盖智算训练和推理全流程。同时，为该资源池定义了数据处理和分析流程，包括数据的收集、清洗、转换、分析和可视化等步骤，可根据企业的具体需求和业务场景来进行定制。资源池整体架构如图1所示。

开放生态，优算提效

面向智算生态不开放，存在厂家绑定、资源整合难、运营成本高、生态建设难等问题，中兴通讯千卡智算资源池基于模型解耦、训推解耦、软硬件解耦等方式，积极探索统一生态建设途径，打造开放、解耦的生态。

大模型的训练周期长，训练中断是影响训练效率的核心问题。提供长时稳定的训练环境是大模型训得快的必要保障。中兴通讯智算资源池实现集群资源可管可视，故障能够快速定位、隔离、修复，提供高效的训练中间文件缓存以及读取、断点续训机制，缩短模型训练中断时间。

此外，传统大模型推理还面临算力成本高的问题。推理场景、客户需求侧重都有所不同，如果为所有业务不加区别地提供算力和服务，是对资源的浪费，也难以获得成本竞争优势。构建多样化的推理算力，为不同的业务需求选择恰到好处的推理算力服务是实现性价比最优的必要手段。智算资源池在算法层通过模型优化、模型压缩等多种手段降低模型规模和对算力的需求，基于业务量的潮汐效应，做到算力按需部署、弹性



▲ 图1 南京电信算力资源池建设目标架构

伸缩，发挥算力的最大价值，降低成本。

以网强算，自主创新

在智算中心，数据密集型任务（如大规模AI模型训练）需要在短时间内传输海量数据，当多个计算任务同时进行，尤其是在数据并行或模型并行的计算场景中，网络需要处理大量的并发数据传输。但网络设备的吞吐量有限，容易出现拥塞，使得数据传输效率降低。在智算中心内部不同计算节点之间需要频繁通信协作以及计算任务对存储设备的数据访问均需要快速响应。除此之外，随着智算中心计算资源的不断扩充，对网络的扩展性提出了更高要求。传统网络架构在添加新的计算节点或存储设备时，可能需要重新配置网络拓扑结构、更换网络设备，这会导致网络扩展的周期长、成本高。

南京电信本次建设的国产化千卡智算资源池采用中兴通讯ZXR10 9900X/5960M/5960X系列交换机。作为中兴通讯面向智算数据中心网络的旗舰级产品，凭借其创新的硬件架构设计、高性能可编程产品与智能化软件生态系统，实现了对智算和通算全场景的无缝覆盖，并在RoCE（基于RDMA的以太网）无损网络领域树立了行业标杆。

该系列交换机融合了全局均衡调优iGLB技术与端网协同ENCC算法，实现了高达98%的全网吞吐效率与微秒级快速流量拥塞调控，为大规模模型训练提供了零丢包、极致吞吐、超低时延的无损网络环境。同时，其强大的灵活扩展性支持万卡级智能计算集群的超大规模组网，满足未来数据中心对算力与性能的极致追求，通过网络“无损”实现AI算力“无损”。

在算力产业蓬勃发展的当下，南京电信与中兴通讯强强联合，成功搭建千卡国产化GPU资源池，意义深远。这一举措不仅有力推动了国产化技术的广泛应用，显著提升了自主可控能力，更有效促进了产业上下游的协同共进，形成强大示范效应，为地区算力产业生态的构建注入了强劲动力。ZTE中兴

河南移动携手中兴通讯升级IP承载网，打造绿色安全算力支撑网络

刘义
河南移动规划技术部
数据通信资深专家

徐晓蕾
河南移动规划技术部
基础网规划主管

李劭阳
中兴通讯河南分公司
产品方案经理

算网融合时代，业务驱动网络成为业界共识，运营商需要更灵活智能的网络技术来满足多样化的算力业务需求。满足算网融合对网络提出的敏捷开通、灵活随选、安全可靠、智能可视等能力需求的SRv6/G-SRv6、网络切片、随流检测、ARN（应用响应网络）等“IPV6+”技术应运而生。

河南移动IP专用承载网主要承载三云（移动网、网络云、IT云）算力平台的云管、云间及部分入云流量，保障全省4G/5G、光纤宽带、政企等各类用户的通信、计算、存储等业务需求，是河南移动重要的算力枢纽网络。因部分节点在网设备老旧，网络无法整体升级支持“IPV6+”新技术，制约了河南移动向算网一体、智慧内生算力网络技术架构的演进。

在此背景下，河南移动携手中兴通讯对现网部分节点进行替换升级改造，共同打造安全可靠、绿色节能、技术领先的自有业务算力调度的枢纽承载网，为河南移动自有高服务质量和高安全可靠性要求的电信业务提供安全可靠的承载网络，赋能河南数字经济高质量发展。

聚焦算网融合需求，产品前瞻选型

在网络升级改造过程中，河南移动充分考虑网络当前及未来演进的需求，引入的中兴通讯M6000-S系列产品具备大容量、高性能、大容量、灵活可扩展、新技术演进全面支持等技术优势，全面满足算网融合时代的IP承载网络的业务需求。

- 全面支持新技术演进，支撑算力网络战略落地

产品硬件采用中兴通讯自研灵活编程的转发芯片，基于灵活的编程能力更容易实现SRv6/G-SRv6、网络切片、随流检测、ARN等“IPV6+”新技术的演进部署。软件系统采用模块化设计，业务功能模块智能动态加载，卸载轻松自如，支持灵活按需扩展新业务，可支撑算力网络新需求的简易落地。

- 大容量、高性能，支撑业务未来发展

采用业界先进的分布式并行处理和无阻塞交换架构，以及业界领先工艺的自主研发芯片，具备长期演进能力，可支撑系统容量的平滑扩展。

- 灵活可扩展，有效降低网络建设成本

设备板卡采用灵活子母卡结构，通用母卡可混插不同业务类型的接口子卡，使端口组网更加灵活，可按实际需求为网络建设提供低成本、个性化的解决方案，降低整体网络建设成本。

坚持自主创新生态，保障网络安全可靠

河南移动和中兴通讯积极贯彻国家构建安全可控信息技术体系的相关要求，项目选用的设备为全面符合最新安全可控要求的M6000-S系列设备，该系列设备的核心元器件为中兴通讯自研，非关键器件采用安全可控供应链生态。

通过本项目的实施，实现了对现网存量设备的替换，确保了河南移动业务支撑系统算力互联的信息安全。

> 河南移动携手中兴通讯共同打造安全可靠、绿色节能、技术领先的自有业务**算力调度**的枢纽承载网。

探索绿色节能，全面契合双碳战略

项目建设中河南移动积极响应国家提出的碳达峰碳中和行动计划，为满足绿色节能的相关要求，与中兴通讯协同探索最新的设备节能和网络节能技术在项目中的部署落地。该项目采用的中兴通讯M6000-S系列产品支持多级节能技术，大幅降低网络设备能耗。

- 芯片多合一高集成设计，功耗更低

产品核心芯片采用中兴通讯自研多合一高集成先进工艺，将传统的多个芯片采用先进工艺集成在1个芯片中，单个芯片功耗相较于传统多芯片功耗降低50%以上。

- 智能多级风扇调速技术，降低风扇能耗

采用智能多级风扇调速技术，当路由器工作负载较低、设备温度不高时，风扇降低转速，减少能耗；在路由器高负载运行、产生较多热量时，风扇提高转速加强散热，保障设备正常运行。

- 灵活下电未使用的板卡，按需节能

当业务板卡端口未使用时，支持对板卡灵活下电，按需降低系统的整体功耗。

- 交换链路动态节能

当业务板卡不在槽位时，对应的交换链路可自动关闭，从而降低整机能耗。

作为河南省领先的通信及信息服务企业，河南移动锚定中国移动集团“世界一流信息服务科技创新公司”定位，坚持“安全可靠、绿色节能、技术领先”的既定方向，持续加大投入，巩固提升河南网络枢纽地位。后续河南移动将继续携手中兴通讯等产业合作伙伴发力新基建、人工智能、数据要素、算力等新质生产力，为河南省千行百业数智化升级赋能，为省内数字经济高质量发展做出贡献。ZTE中兴

ZTE中兴

让沟通与信任无处不在