中兴通讯技术 简讯

ZTE TECHNOLOGIES | 2024年3月/第3期
准印证号：（粤B）L011030048

视点

专题：智能算力

内部资料 免费交流

扫码体验移动阅读

第28卷/第03期
总第426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办

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：4000本
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：2024年3月25日

王卫斌
中兴通讯产品规划首席科学家

打造新型智算，赋能千行百业

过去的一年是大模型蓬勃发展的一年，海内外“百模大战”愈演愈烈。模型规模的提升，带来模型精度、泛化性、能力涌现、内容生成等一系列令人惊艳的能力，大幅提升了人工智能的生产力和应用场景。AI时代已然来临。

大模型需要大算力，智能算力是大模型发展的基石。据IDC预测，中国智能算力规模未来年复合增长率将达到52.3%，远超通用算力，到2025年我国智算规模将达到300EFLOPS，占整体算力的35%。我国智算产业发展将迎来重大机遇。

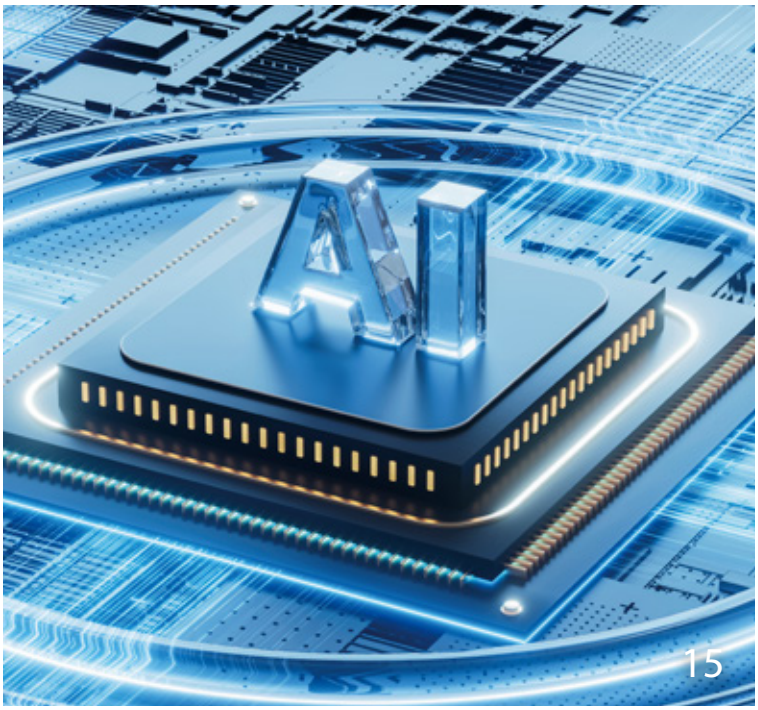
然而机遇面前挑战同样艰巨。当前智算已经成为国家竞争的关键技术方向，我国智算发展起步晚，国产AI芯片多而不强，AI软件栈及框架生态能力弱，AI基础算法领域相对落后，需要业界群策群力推动产业发展。面对大模型训练场景，需要围绕算、存、传以及资源管理平台的有效协同，提升大规模智算集群的算效、能效和可靠性；面对训练推理一体化场景，需要结合具体场景采用通算扩容、升级以及智算新建等多种可行方式，支持多样性算力互补提升资源利用率；面对行业智能化场景，需要软硬集成的一体机，提供私域快速部署、一体化服务。

面对机遇和挑战，中兴通讯推出星云智算解决方案，提供从计算、存储、网络等智算基础设施，AI训推平台，到1+N+X大模型及应用等全栈智算产品，方案兼顾国际先进技术及国产化智算生态，助力多场景智算中心建设，赋能千行百业数智化生产及智能化转型。

坚持“数智经济筑路者”定位，做极致的AI公司，中兴通讯愿与业界伙伴携手，共同迈进智能化新时代！

目次

中兴通讯技术（简讯）2024年第3期



面向大模型，中兴通讯全栈智算解决方案赋能千行百业

人工智能发展至今，已经经历了三次高潮、两次低谷。2022年11月，OpenAI公司发布的ChatGPT及其采用Transformer算法和预训练大模型的生成式AI技术，使得第三次人工智能技术发展达到了前所未有的新高度，并由此迎来AI大模型技术拐点和炒作高峰。

视点

专题：智能算力

媒体转载

中兴通讯荣获GTI Awards 2024多项大奖

2月27日，GTI Awards 2024获奖名单在MWC24巴塞罗那正式公布，中兴通讯斩获3项大奖。其中，中兴通讯“5G-A赋能VR多人对战游戏”和中兴通讯、土耳其移动合作推出的“智能电网：土耳其移动的能源分配网络与5G集成项目”均荣获“移动创新服务应用奖”，中兴通讯动态智能超表面D-RIS 2.0荣获“移动技术创新突破奖”。

移动创新服务应用奖——5G-A赋能VR多人对战游戏

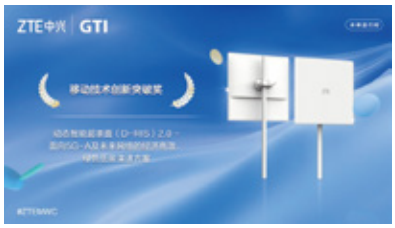
中兴通讯携手中国移动研究院、北京移动重客中心、高通和当红齐天集团，在当红齐天首钢一高炉SoReal科幻乐园项目，联合完成基于5G-A网络XR专属保障方案的多并发大空间VR竞技游戏业务验证。

移动创新服务应用奖——智能电网：土耳其移动的能源分配网络与5G集成项目

中兴通讯与土耳其第一大无线运营商Turkcell携手合作，在土耳其推出了一种创新的智能电网系统，助力电力行业取得了重大突破。

该系统作为Turkcell在电力配电网络中扩展5G应用的一部分，为土耳其第一大城市伊斯坦布尔的两个变电站之间建立低延迟的5G网络通信。项目与ABB、BEDAŞ和NETAŞ公司深度合作，并通过使用GSL Engineering公司提供的3GPP Release 16 Robustel EG5120 5G工业调制解调器，实现在故障发生时快速激活保护电路，最大限度地减少故障影响的区域和持续时间，成为土耳其5G行业垂直应用案例的典范。

移动技术创新突破奖——动态智能超表面D-RIS 2.0



D-RIS是业界首个支持和基站动态协同的智能超表面，此次获奖的D-RIS 2.0是其第二代产品，具备高增益、低功耗、易部署等特性，通过5G-A基站和D-RIS 2.0协同部署，能够扩展5G-A基站30%的覆盖范围，信号增益高达30dB，功耗低至30W。

中兴通讯加入中国移动RedCap商用推广行动倡议

2月28日，在巴塞罗那举办的世界移动通信大会上，中国移动携手GSMA及全球产业伙伴，共同发布RedCap商用推广行动倡议，中兴通讯高级副总裁张万春出席发布会现场，中兴通讯将与产业合作伙伴一起，共同推动RedCap产业加速发展和规模应用。

中兴通讯参与中国电信天翼云全球发布

2月27日，在MWC24巴塞罗那，中国电信云网高峰论坛顺利举办，全球通信业领袖和专家代表云集，共同探讨释放全球互连的数字潜力。论坛上，天翼云全球发布仪式正式启动，标志着中国电信云服务全球化布局与服务进一步延伸。作为中国电信的战略合作伙伴，中兴通讯受邀出席论坛，将全面助力中国电信打造云服务出海新航标。

中国联通与中兴通讯成功举办5G网络创新与应用发布会

2月26日，中国联通携手中兴通讯在世界移动通信大会（MWC24）成功举办“算网为基，智领未来”5G网络创新与应用发布会，联合发布5G确定性工业互联网的技术试验成果及《以用户为中心的无线算网架构及关键技术白皮书》《面向工业智能化的确定性技术白皮书》两大白皮书。

武钢有限、中国联通和中兴通讯荣获GSMA GLOMO “最佳互联经济移动创新奖”

2月28日，在MWC24巴塞罗那，由宝钢股份武汉钢铁有限公司（简称“武钢有限”）、中国联通和中兴通讯联合打造的“5G全连接智慧钢铁工厂”项目荣获GSMA全球移动大奖（GLOMO）——“最佳互联经济移动创新奖”（Best Mobile Innovation for Connected Economy）。

武钢有限“智慧钢铁工厂”项目目前已建成全球钢铁行业最大的5G专网，在园区实现99%的5G覆盖，并基于5G专网部署了6大类25个钢铁应用场景，贯通智慧物流、生产管控、数字设备、能环管控、质量管控、安全管控等全流程，建成一个公司级管控中心和炼铁、炼钢、CSP、热轧四大厂区操控中心，实现一键式炼钢。

5G专网已应用于武钢有限的核心生产领域，实施无人化改造的行车已超过100个。智慧铁水运输实施以来，运输效率不断提升，TPC铁水温降历史性突破100°C，创造铁钢界面极致效率新纪录，每年减少碳排放超过75万吨。

中兴通讯举办“5G-Advanced联合创新及新品发布会”

2月27日，世界移动通信大会期间，中兴通讯举办“5G-Advanced联合创新及新品发布会”，系统展示了中兴通讯面向5G-A时代的全景规划及成果，并发布了5G-A十大新品，为5G-A商用做好全面准备。发布会特邀国内三大运营商及产业合作伙伴共同参与。

中兴通讯高级副总裁张万春先生在开场主旨发言中指出，5G-Advanced技术的推出是未来通信技术发展潜力的一个重要里程碑。

发布会上，中兴通讯发布了面向5G-A时代的十大创新产品，涵盖了业界首创的恒定功放效率UBR产品、可提供超万兆体验的系列AAU产品、拓展5G低空和星连的新品，以及通信与算力融合的系列产品，为5G-A全景筑基。

中国联通、中兴通讯、GlobalData联合发布新通话白皮书

2月26日，在MWC24巴塞罗那，中国联通、中兴通讯联合全球知名咨询公司GlobalData共同发布《Remonetizing Voice with Next-Generation 5G Voice and Video Services》白皮书。白皮书围绕5G新通话的价值生态链、业务体验、网络架构、部署建议、发展趋势等进行了系统性的阐述，为运营商、设备提供商、合作伙伴，以及行业客户评估和建设5G新通话提供了科学的借鉴。

中兴通讯推出高密度全液冷整机柜IceCube

在MWC24世界移动通信大会，中兴通讯推出了高密度全液冷整机柜IceCube，展现了其在技术领域的深厚实力，更是对通过创新推动高效未来的坚定承诺。IceCube机柜，通过零间隙部署，能够容纳高达40台1U服务器，极大提升了单位空间的算力。随着高密度服务器的引入，单机柜的总功耗的将会达到100KW。得益于IceCube高效冷板制冷技术，叠加液冷机柜门，IceCube可大幅降低能耗，实现pPUE值低于1.1。

中兴通讯发布零碳能源网V3.0解决方案

在MWC24世界移动通信大会，中兴通讯发布了零碳能源网V3.0解决方案。这一创新方案通过系列化的产品和技术，实现了绿色发电、高效转电、智能储电、精准用电以及智能运维等环节的创新，从而为通信站点供电全链端到端的节能减排和提效降费做出贡献，助力运营商网络的绿色低碳演进。

中兴通讯推出全球首款基于AI技术的AI 5G FWA

MWC24世界移动通信大会期间，中兴通讯于终端新品发布会推出由AI驱动的全场景智慧生态3.0，进一步丰富“1+2+N”产品队列，并同步发布全新的AI 5G FWA（fixed wireless access），以AI重塑连接体验，为用户创造美好通信生活。

据国际专业咨询公司TSR最新报告显示：2023年中兴FWA&MBB产品市场占有率已连续三年稳居全球第一。

数智兴农　中国移动携手中兴通讯打造5G智慧农业项目蝉联GSMA Foundry大奖

2月28日，在MWC24世界移动通信大会期间，GSMA Foundry颁奖典礼举办，GSMA Foundry旨在释放互联互通的全部力量，快速开发应对行业挑战的现实解决方案，打造数字未来。由中国移动、中国移动吉林公司、中兴通讯联合打造的“Automated Farming”项目（吉林省大安市盐碱地5G智慧农业项目）荣获GSMA Foundry“5G货币化卓越奖”（GSMA Foundry 5G Monetisation Excellence Award）。这是该项目继2023年荣获GSMA Foundry“卓越贡献奖”（GSMA Foundry Excellence Award）后再次被授予荣誉，表明GSMA Foundry对项目在智慧农业领域持续深耕的示范效应和经济效益给予高度认可。

中国移动、中国移动吉林公司携手中兴通讯在吉林首创盐碱地5G“单元化”无人农场理念，通过5G技术打造全视角、全场景、全过程、全成本的国内首个盐碱地5G智慧农业项目；运用5G远程驾驶+5G智慧灌溉技术针对稻田pH和ESP值进行持续检测，实现盐碱地生态修复，同时运用5G无人耕地、5G无人插秧、5G无人机巡田、5G无人收割等技术融合中国移动“AICDE”，实现农业生产“改、耕、种、管、收”全生命周期数智化管理。

项目从2021年至今，5G“单元化”无人农场建设从2个扩展到10个，盐碱地生态修复成果也从7万亩扩大到20万亩，实现直接经济效益超千万。

中国电信携手中兴通讯联合发布Cluster DRS创新成果，赋能低空经济

2月26日，2024年世界移动通信大会期间，中国电信携手中兴通讯联合发布了Cluster DRS（dynamic radio sharing）创新技术和成果。

中国电信联合中兴通讯推出的Cluster DRS解决方案，5G商用网络动态生成以无人机为中心的基站簇，簇内多小区间实现空域波束共享；簇间灵活降低干扰，稳定保障无人机高清视频的实时回传体验。

中兴通讯行业首个5G+XR网媒融合解决方案亮相2024世界移动通信大会

2024年世界移动通信大会期间，中兴通讯推出了行业首个5G+XR的网媒融合解决方案。在展台可以体验到基于该方案赋能的两个领域创新业务，分别是智能制造业领域的5G MR沉浸式协作和数字文旅领域的5G VR大空间沉浸剧场，为大家呈现了行业数智化发展的新动能和新机遇。

中兴通讯推出AiCube训推一体机，加速大模型走深向实

在2024年世界移动通信大会上，中兴通讯推出了AiCube训推一体机，为运营商和行业用户提供Ai-In-One一站式智算解决方案，为企业数字化转型赋智提效。

中兴通讯的AiCube训推一体机具有快速交付、资源按需分配和安全易用的特点。该产品用户无需复杂的部署和配置过程，即可快速投入使用，节省时间和资源。

中兴终端亮相MWC24，“Better for All”全球愿景发布　多款新品推出

2月26日，中兴通讯亮相MWC24世界移动通信大会，发布终端全球品牌愿景“Better for All”，推出由AI驱动的全场景智慧生态3.0，并发布多款创新产品和技术，包括全球首款5G+AI的裸眼3D平板nubia Pad 3D II、全新AI理念的5G FWA、首款小折叠屏手机nubia Flip，以及主打影像、音乐和游戏的多款特色新品。努比亚Z60 Ultra、红魔9Pro系列、领先全球的MBB&FWA系列产品也在展会亮相。

AI技术的加持，为终端融合创新带来了巨大的机遇。中兴通讯推出全场景智慧生态3.0，以多终端智能互联和生态延展为核心，基于中兴星云OS系统，实现各业务接入统一的AI平台，整合大模型、大数据和3D技术，提供全场景联接、多维度感知和多端融合计算，覆盖运动健康、影音娱乐、商务出行、家庭教育和智能驾驶五大场景。

在创新终端领域，中兴通讯带来全球首款5G+AI驱动的裸眼3D平板nubia Pad 3D II，并全球首发Neovision 3D Anytime技术，支持系统级2D到3D实时自由转换，用户打开任意2D内容，都可以一键实时转换，随时随地体验3D视觉效果。

在移动互联领域，2023年中兴通讯5G FWA&MBB产品市占率已连续三年稳居全球第一。此次展会全球首发AI 5G FWA——中兴G5 Ultra，基于底层设备行为和网络应用分析，通过多维度AI网络算法的应用，实现不同应用场景下的网络策略控制，极大提升了带宽利用率和网络性能。此外还发布了第五代5G室外FWA中兴G5F，支持5G Advanced-ready，支持Sub6G和mmW载波聚合及双连接，峰值速率高达10Gbps，为用户提供前所未有的超高速网络体验。在汽车生态领域，中兴RCU产品Y2002集5G、V2X、MEC于一体，融合通信和算力功能，连接感知设备，通过100TOPS的AI处理，对道路上的人、车实现智能识别和提前预警，赋能辅助驾驶和智慧交通。

中兴通讯发布端到端算力基础设施解决方案，加速全行业数智化转型

2024年MWC世界移动通信大会上，中兴通讯发布端到端算力基础设施解决方案，提供算/存/网/IDC完整的全套解决方案，实现全栈软硬件一体化部署，加快业务上云速度。

此次中兴通讯发布的解决方案，在硬件层面，全系列服务器提供高品质异构算力，高性能分布式全闪存储实现海量数据的快速读写，提升大模型训练的速度；无损网络实现0丢包、微秒级时延；全液冷模块化预制IDC，PUE低至1.13。在软件层面，打造AI Booster智算平台，通过自动并行训练最大化GPU利用率，通过可视化开发和自适应参数优化，大幅降低开发门槛。

此外，中兴通讯为不同的应用场景提供了差异化的解决方案。在通算领域，针对互联网云和电信云等通用场景，提供系列化通用服务器，全系支持液冷，为客户提供高性价比和高扩展性的通用算力。针对海量的大视频存储的场景，推出了大存储服务器。对于金融、科学计算等关键应用场景，推出业界领先的4路服务器。此外，针对高密度计算场景，中兴通讯推出IceCube全液冷整机柜方案，推进数据中心的绿色高效发展。在智算领域，中兴通讯提供系列化训练服务器、推理服务器和AiCube训推一体机，满足中心万卡规模训练池、区域通算/智算融合推理池、边缘训推一体等全场景需求。

中国移动携手中兴通讯发布全球首台算力路由器，共创数字产业新生态

2月26日，在2024年巴塞罗那世界移动通信大会（MWC）期间，中国移动副总经理高同庆与中兴通讯执行副总裁、首席技术官王喜瑜等嘉宾在中国移动展台发布全球首台算力路由器。该产品将计算因子引入路由系统，实现算网联合路由，提升算网系统的整体性能和处理容量，降低业务端到端时延，可应用于时延和计算敏感新型业务以及大规模AI推理等业务。

中兴通讯荣获Lightwave三项大奖

近日，全球光网络领域知名媒体Lightwave公布2024年光通信年度创新大奖（Lightwave Innovation Reviews）评选结果，中兴通讯哑资源故障智能管理系统、Tbit全光接入平台ZXA10 C600E、FTTR-B解决方案成功获奖，充分印证了中兴通讯在光通信领域的强大实力，体现了业界对中兴通讯的高度评价和肯定。

中兴通讯OTN网络哑资源故障智能管理系统可自动识别OTN网络中的同缆同路由风险，提前预警光纤劣化，在光缆发生意外中断后，能够在GIS地图上快速定位断点，使光层运维效率提升90%，保障用户业务高可靠性。

中兴通讯Tbit全光接入平台ZXA10 C600E是为满足50G PON的大规模部署而研发的新一代大容量全光接入平台，具备超高带宽、确定性、开放及绿色等特征。

中兴通讯FTTR-B（fiber to the room for business）解决方案通过光纤进行企业内网建设，能为中小微企业任何一个角落提供2Gbps+的Wi-Fi服务，且光纤组网不受带宽限制，可满足未来20年技术演进升级需求。目前中兴通讯FTTR-B方案已经在中国30多个城市成功商用，并在泰国、印尼、奥地利、巴西等二十多个国家建立试商用局，助力提升全球中小微企业数字化转型效率。

Lightwave已连续多年评选光通信各技术领域的年度创新奖，旨在表彰光通信领域的顶尖产品和解决方案。评委团由网络运营商、技术供应商以及行业研究和分析公司的高管、分析师及工程师等组成。

中兴通讯重磅发布家庭AI媒体算力中心　拓展智慧家庭交互新体验

2月26日—29日，在MWC24世界移动大会期间，中兴通讯发布了全新家庭AI媒体算力主机。该产品将传统机顶盒能力与存储、算力、智控相结合，助力运营商打造智慧家庭能力中心。

天津联通携手中兴通讯联合打造BBU风液混合柜方案商用示范

2月，天津联通携手中兴通讯，在天津打造了BBU风液混合柜方案的商用示范，标志着双方绿色节能创新实验室的合作取得了实质性进展。本次BBU风液混合柜方案成功落地，不仅是对双方合作的肯定，也是中国联通在绿色节能创新方面的积极探索，体现了中国联通落实国家“双碳”政策，助力社会可持续发展的责任担当。

中兴通讯携手德国O2探索零碳站点建设

2月，中兴通讯采用太阳能、甲醇制氢燃料电池和智能储能的创新绿色供电解决方案成功为德国O2建设了一座通信站点，在探索零碳站点方面取得了积极的进展。这是德国O2首次在商用中采用这一绿色方案，在边远通信站点展现绿色创新力量。

中兴通讯携手中国移动、高通和当红齐天完成业界首个5G-A多并发大空间XR竞技游戏业务试点

2月，中兴通讯携手中国移动、高通技术公司和当红齐天集团，在当红齐天首钢一高炉SoReal科幻乐园项目，联合完成基于5G-A大容量、低时延及智能化XR业务保障方案的多并发大空间XR竞技游戏业务试点，结果显示，在近千平米的大空间内，12路XR业务同时接入时，画面清晰流畅无卡顿，平均空口时延低于10ms，无线大空间多人XR免背包体验如约而至。

在大空间XR体验中，传统的背包渲染方案面临着包括背包设备重量大、散热差、续航短等多重挑战，而使用WiFi等无线方案通过网络侧渲染可以解决背包方案的问题，但也存在容量受限、信号不稳定和易受干扰等问题，这些问题无法满足多并发用户的体验，阻碍了XR应用的发展。针对这些挑战，寻求一个高效的无线解决方案至关重要。

在此次当红齐天XR多人智能体育竞技端到端验证中，中兴通讯采用基于内生智能的SuperMicell解决方案，提供高速率、低时延的5G-A室内网络覆盖及智能化的XR业务保障，配合采用骁龙®X75 5G调制解调器及射频系统的高通测试终端，以及第一代骁龙XR2平台的VR一体机进行验证，游戏过程中画面清晰度达到业界主流的4K@90fps高画质。多人竞技游戏过程中大量数据交互，通过“免背包”方式，将背包式的本地渲染上移至云渲染，大量的数据承载于5G-A网络上，这要求网络提供稳定的大带宽和低时延能力，并满足数十路XR业务并发的确定性业务保障。本次测试验证了中兴通讯SuperMicell方案，通过联合智能波束管理，完善业务覆盖，满足XR应用的移动性需求；通过5G-A XR业务专属保障方案，以时隙级的智能调度和优化策略，实现XR业务的端到端性能保障，平均空口时延低于10ms。联合验证结果表明，该解决方案实现了无线大空间XR体验，免背包、去线缆，可支持超50路XR用户同时在线，畅游虚拟世界。

土耳其移动Turkcell携手中兴通讯共创800G传输世界纪录

2月，土耳其移动（Turkcell）携手中兴通讯基于800G可插拔相干光模块完成单波800G无电中继传输验证，并创造陆地系统最远2000km的传输距离，为800G在中长距场景下的商用奠定坚实基础。

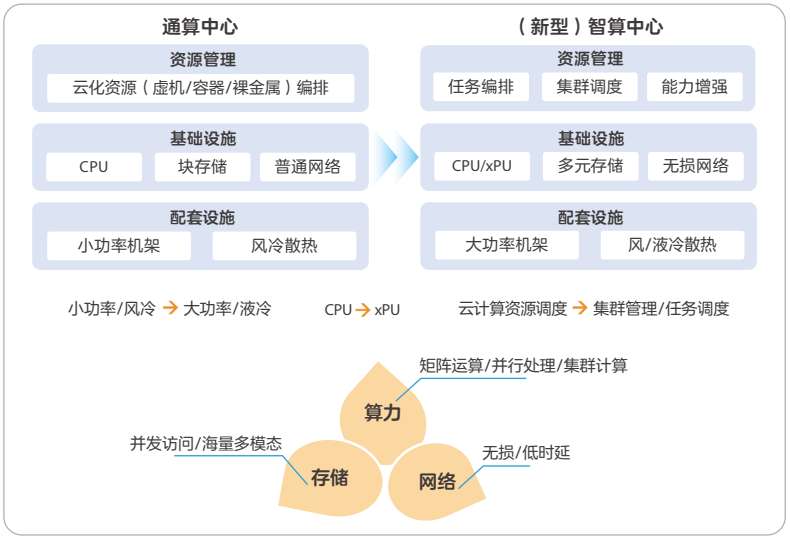
土耳其横跨欧、亚两洲，处于欧亚连接处的重要地理位置，因此长距传输成为其光网络发展的必要条件。作为当地第一大运营商，Turkcell为五个欧洲跨境网关提供网络基础设施，并通过多重保护和冗余方案向欧洲主要城市提供独立通信线路。同时，Turkcell为通往中东、阿拉伯半岛、土耳其共和国和高加索地区的7个东部跨境网关提供网络服务。此外，未来土耳其5G网络的部署以及云计算和数字服务等新兴技术发展也将激发网络流量大幅增长，进而增加运营商对带宽提升的需求。

为解决以上问题，Turkcell携手中兴通讯基于创新800G可插拔方案完成单波800G无电中继传输验证。在该项目中，中兴通讯提供光传输旗舰产品ZXONE 9700，采用业界首个800G插拔相干光模块，实现在G.652光纤中2000km以上的陆地长距离传输。此次验证成果将为Turkcell在现网中引入800G及创新解决方案、确定未来光网络发展方向奠定坚实基础。值得一提的是，与业内标准800G固定模块相比，该可插拔模块功耗可降低68%，符合土耳其可持续发展策略需求，并可助力Turkcell达成发展绿色、低碳光网络的战略目标。

智能算力
发展趋势洞察

朱堃
中兴通讯云计算规划总工

随着ChatGPT横空出世，人工智能（AI）技术在短时间内呈现“涌现”态势，并成为推动社会进步的关键力量。AI技术的广泛应用给我们的生活和工作带来了巨大的改变，而这一切的背后离不开算力基础设施的支持。AI训练任务以及推理应用对算力有着高性能、大规模并行、低时延互联的要求，导致对计算、存储、互联网络有了不同于通用计算的要求，同时对算力聚合的要求也引发了基础设施管理平台的创新（见图1）。



▲ 图1 智算中心创新架构

AI芯片

除了大模型训推有高性能矩阵运算的要求之外，大模型参数量越大对内存容量的需求越大，同时多颗AI芯片间的大量数据交互也带来了对互联总线高带宽、低时延的要求。因此，算力、显存，以及互联总线形成了对AI芯片的三大能力要求。

算力方面，由于人工智能采用基于多层神经网络的机器学习技术，需要对大量数据进行矩阵运算，例如矩阵乘法、卷积、激活函数等。传统CPU以复杂数据流程见长，为此将更多的空间让渡给了控制单元和缓存单元，计算单元只占25%的空间，一般只有几十个算术逻辑单元（ALU），处理这些并行化和向量化运算的效率不高。而处理图像和图形相关运算的GPU计算单元占90%的空间，高达几千的ALU适合对密集数据进行并行处理。在2017年后，主流AI芯片厂家发布专门针对矩阵运算加速的AI GPU（GPGPU，general purpose computing on GPU），为大模型训练提供了更高的计算性能。除硬件之外，GPU厂家通常会提供相应的开发平台（如NVIDIA CUDA），它使得开发者能够直接使用GPU进行编程和优化，充分发挥GPU的计算能力。

显存方面，Transformer类模型参数量按照

平均每两年翻240倍的速度增长，与之相比，AI内存容量仅以每两年翻2倍的速率增长，已经远远不能匹配大模型增长速率。为解决该问题，内存统一寻址的“超级节点”是目前比较可行的方案，如：定制AI服务器，通过高速互联技术组成1个超级节点（包含256颗GPU和256颗CPU），支持GPU和CPU之间的内存统一寻址，内存容量可以提升230倍。此外，AI芯片内采用计算和存储分离的冯·诺依曼架构，芯片60%~90%的能量消耗在数据搬移过程中。按照H800的最大功耗700W的60%来估算，数据搬移消耗了420W。为解决该问题，存算一体技术将内存与计算完全融合，避免数据搬移，大幅提升了能效。

互联总线方面，大模型3D并行拆分后，带来了芯片间数据传输的要求。其中数据传输量最大的张量并行（TP），在传输时间中的占比超90%。有测试数据表明，使用同样数量的服务器训练GPT-3，采用NVLink相比PCIE，一个Micro-batch在相邻GPU之间的传输时间从246.1ms降低到78.7ms，整体训练时间从40.6天降低到22.8天，因此互联总线的带宽成为关键。

智算存储

在大模型开发端到端的多个环节中，都对存储提出了创新需求。具体包括：

- 多元存储：视频、图像、语音等多模态数据集带来块、文件、对象以及大数据等多元存储以及协议互通的要求；
- 海量存储：为保证大模型训练的精准性，数据集通常为参数量的2~3倍，在当前大模型从千亿到万亿飞速发展的时代，存储规模是一个重要的指标；
- 并发高性能：大模型并行训练场景下，多个训练节点需要同时读取数据集。在训练过程中，训练节点需要定时保存检查点（checkpoint）以保障系统的断点续训能力。这些读写操作的高性能能够大大提升大模型训练的效率。

因此，作为智算存储，首先需要提供多元数据存储能力以及块（iSCSI）/文件（NAS）/对象（S3）/大数据（HDFS）多协议互通能力。可通过软硬件综合调优来提升性能，硬件加速手段包括：通过DPU卸载存储接口协议以及去重/压缩/安全等操作，数据按热度自动分级及分区存储；软件调优手段包括分布式缓存、并行文件访问系统/私有客户端等。同时，采用NFS over RDMA以及GPU直接存储（GDS）技术也能够大大降低数据访问的时延。

无损网络

AI大模型训练的并行计算特性带来大量通信开销，使得网络成为制约训练效率的关键因素，无损网络成为刚需，具体表现为零丢包、高吞吐大带宽、稳定低时延以及超大规模组网。

目前的无损网络协议主要分为英伟达的IB与RoCE两大类。IB网络最初为高性能计算（HPC）设计，具备低延迟高带宽、SDN化拓扑管理、拓扑组网丰富以及转发效率高的优势，但存在产业链封闭的问题。RoCE为统一承载网络设计，具备高带宽/高弹性组网，对云化服务支持较好以及生态开放的优势，是国产化的必选之路。但是在网络性能和技术成熟度方面不如IB，需要结合芯片进一步优化时延。

传统网络拥塞和流量控制算法端侧和网侧独立，网络仅提供粗颗粒度的拥塞标记信息，很难确保网络高吞吐满负荷场景下不出现拥塞、丢包以及排队时延。因此需要端网协同实现精准、快速的拥塞控制和流量调度算法，进一步强化网络性能。

在网络拓扑方面，Fat-Tree CLOS和Torus轨道多平面拓扑为当前两种主流形态，从组网设计上解决网络拥塞问题。Fat-Tree CLOS网络基于传统树型网络增强，采用上下行带宽1:1低收敛比，保障任意两个节点间无阻塞路径；Torus轨道多平面网络将不同服务器相同位置GPU连接到同一组交

图2 智算中心层次化部署模式

换机，构成一个轨道平面。同时，服务器不同位置GPU连接到不同交换机，形成多个轨道平面。

资源任务调度平台

和通用算力资源管理平台通过虚拟云化技术将资源分发给多个租户不同，智算场景更强调的是算力聚合，即在AI任务训练中，可能同时运行数百个任务和上千个节点。通过任务调度平台，可以将执行的任务与可用资源进行最佳匹配，从而最小化任务在队列中等待的时间长度，最大化任务并行量，获得最优资源利用率。目前主流的调度系统有Slurm、Kubernetes两种。

Slurm主要应用于HPC场景下的任务调度，已经被世界范围内的超级计算机（包括天河等）和计算机群广泛采用；而Kubernetes作为容器编排平台，用于调度以及自动部署、管理和扩展容器化应用。目前Kubernetes和更广泛的容器生态系统逐渐成熟，正在形成一个通用的计算平台和生态系统。

在AI任务调度场景下，Slurm和Kubernetes面对着不同的挑战：深度学习工作负载的特征与HPC工作负载的特征非常相似，因此可以使用HPC任务调度器Slurm来管理其机器学习集群。但是，Slurm不是围绕容器开发的机器学习生态系统的一部分，因此很难将Kubeflow等AI平台集成到此类环境中。此外，Slurm使用较为复杂，维护难度大；另一方面，Kubernetes更易于使用，并与常见的机器学习框架集成，越来越多的企业和学术机构在他们的大模型训练中使用Kubernetes。但是使用Kubernetes调度GPU资源时会遇到资源闲置时间过长，导致的集群平均利用率低（约为20%）、资源调度只能整卡调度，不能切分或按照卡的类型调度、不能进行任务排队等问题。

部署场景

由于基础大模型预训练、行业大模型精调以及客户场景大模型微调对算力特征及部署位置的要求均不同，结合运营商算力网络DC层次化分布的架构，智算中心部署也呈现枢纽大模型训练中心、省份训推融合资源池、边缘训推一体机三级部署模式（见图2）。

当前，智算已经成为国家竞争的关键技术方向，运营商肩负着提升智算关键软硬件技术创新能力及智算基础设施建设的使命。中兴通讯拥有从IDC、芯片、服务器、存储、数通等基础设施到资源管理平台的全系列产品，结合在电信、政企领域的丰富经验，将助力运营商在智算技术创新及建设中大展宏图。ZTE中兴

杜永生
中兴通讯AIM/无线UME大模型产品
首席架构师

郜艳琴
中兴通讯AIM/无线UME大模型产品
总体规划总工

2024年AI Agent技术洞察：高朋满座，群智涌现

自2023年初ChatGpt3.5推出后，不到3个月时间，随着论文《LLM Powered Autonomous Agents》的发布，AI Agent（智能体）技术立刻引起业界高度关注，继而OpenAI在11月开发者大会推出GPTS。目前，Agent在国内外已经成为大模型的主流产品形式。

相对于大模型这种面向通用思考能力的产品，Agent作为一种角色代理或者说作为专注于特定领域任务的代理，技术上更加容易控制，输出准确率也更高，用户理解和调教起来更加方便。

经过8个月的发展，Agent技术出现一些收敛的征兆，在这个阶段，我们针对Agent定义和基本原理、价值分析、价值驱动的技术发展分析，再结合我们的实验情况，对Agent做一个回顾和洞察，以指导我们后续的研究和工作方向。

当前Agent技术和基本原理

从初期的结构化定义到当前的多模态，Agent的定义经过了Tool、社交型、工作流、多智能体合作、多模态等阶段。当前，我们把基于大模型，拥有学习和记忆能力，可以结合对外部环境的理解和历史记忆，提供对目标任务的思考，按照思考的结果进行技能执行，进而影响环境的虚拟角色，称之为Agent。其中环境（Environment）、感知（Perception）、思考（Think）、技能（Skill），是智能体感知、影响外部环境的行为，我们称为外驱行为。而记忆（Memory）和学习（Learning）则是一个改变自己的过程，我们称为内驱行为。

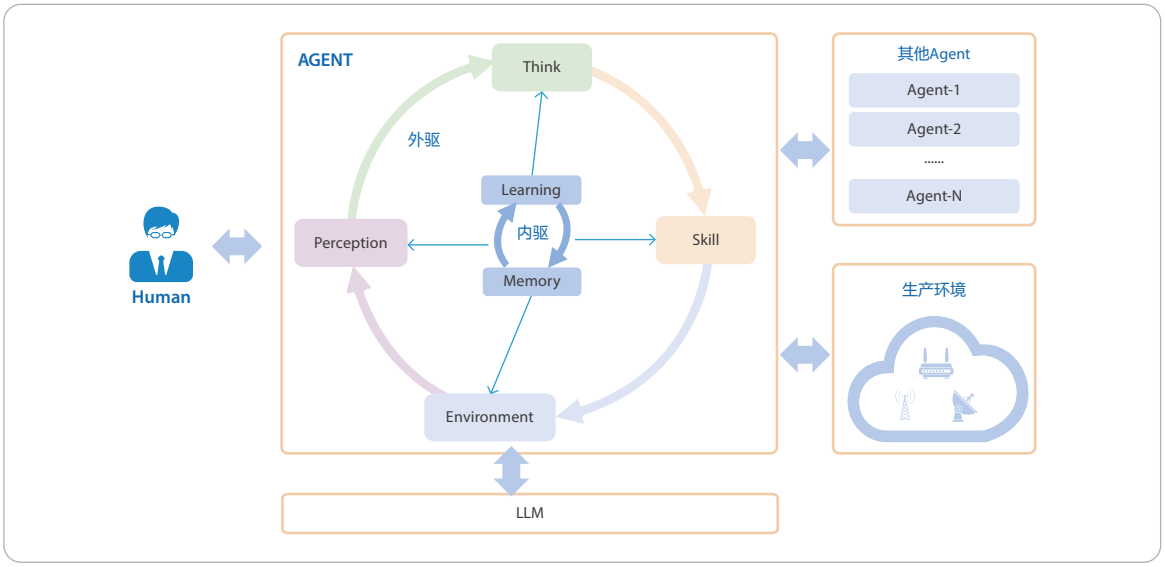


图1 Agent部署和定义 ▶

Agent的典型部署场景如图1所示：Agent主要是衔接人类的意图、LLM（large language model）和生产环境。Agent把人类的意图目标，通过大模型拆解为子目标和相关任务，再把任务通过指令下发到生产环境。实际使用场景包扩通信领域工作流中的保障专家、企业办公领域的Office助理等拟人化场景等。

基于图1，Agent的各个部分介绍说明如下：

- 思考和技能：Agent接收用户的任务目标，通过LLM进行任务思考规划，然后映射子任务到对应技能。技能包括对其他Agent生产任务的间接指令交流，以及对生产环境直接执行的指令。对任务思考分解的方式，可以是整体规划后执行，也可以是迭代边分解边执行。
- 环境和感知：对于任务的分析执行，需要结合当时任务的下上文环境。对于环境的理解，则需要首先通过一种建模方式，把实际的环境信息转换成机器能识别的语言，比如元宇宙是一种物理世界的环境建模方式，通信行业的数字孪生方案则是一种通信网络的环境建模方式。
- 记忆和学习：Agent通过模仿或者强化学习从其他智能体、环境反馈中进行学习，并把学习的成果放入记忆中，下次遇到类似的问题进行借鉴处理。这种随环境学习改变的行为，对于智能体的自我演进有至关重要的作用。

Agent价值分析

首先当前Agent的价值主要由LLM支撑，而LLM本质上是一个条件概率的生成模型，LLM通过提示词的不同，生成不同类型的输出，例如文本生成、任务拆解、逻辑推理、场景理解等。Agent的能力以LLM的输出能力为基础，构建拟人化的角色，服务于生产领域。

其次从业界的主流观点看，Agent体现出来的价值，表现为人+多个虚拟人形成混合专家团队，撬动更大范围的工作，也就是一个人可以干多个人的事情。人做事的方式从人利用工具做事，转变为人驱动多个智能体，多个智能体再使用工具做事。相对工具来说，基于大模型的智能体能提供更加泛化、灵活的判断和决策思考。

中兴通讯实验成果

目前中兴通讯结合自己对于LLM前沿知识的

追踪、价值理解和对通信行业的深刻理解，构建了4种不同类型场景的Agent，包括保障助手、智能问答、故障助手、看网助手。

其中用于重大活动保障场景的保障助手自动化程度较高，是把现实工作流投影到虚拟世界，由虚拟世界的重保专家、助理、排障专家等协作自动完成工作流，并通过总结上报、风险评估等方式和人进行衔接沟通。这是一个复杂的Job Agent类型，以L5- Full Autonomous Network为目标进行设计。

另外3种Agent类型从技术角度看是Task Agent类型：

- 智能问答是结合RAG+Agent技术构建面向ToB的知识库应用；
- 故障助手结合故障知识库和API映射协助运维人员快速排出故障；
- 看网助手：基于大小模型结合，多个Agent通过不同维度进行网络分析后，交给总的网络洞察Agent进行汇总，输出看网结果。

Agent发展趋势和技术拆解

当前学术界的智能体主流分类如下，和中兴通讯的实验结果比较一致：

- Logic Agent：基于对输入语言、多模态的理解再次生成语言和多模态输出的一类Agent；
- Task Agent：面向具体任务，分解计划执行对应操作，过程中没有长期状态记忆的Agent；
- Job Agent：面向较为抽象的工作职责和总体目标，感知环境，记忆过程状态，自生子目标推动工作前进的Agent。

从发展趋势看，自我演进型Agent也非常重要，这类Agent能够自我学习。

下面按照技术层次对主流Agent产品进行拆解，如表1所示。

我们对以上技术进行进一步论文扫描和研究，可以发现：

	Logic Agent	Task Agent	Job Agent	自我演进Agent
环境感知	1. 文字 2. 角色Profile	1. 文字 2. 角色Profile	1. 文字 2. 成熟角色Profile 3. 多模态	1. 文字 2. 成熟角色Profile 3. 多模态 4. 自我执行过程感知
环境理解	上下文	长、短期记忆	1. 长短期记忆 2. 环境建模	1. 长短期记忆 2. 环境建模 3. 有效的抽象记忆
问题思考	思维链技术	思维链技术	智能化提示词	智能化提示词
解决模型	无	1. 服务 2. 小模型 3. SQL	1. 服务 2. 小模型 3. SQL 4. Prompt驱动其他专家	1. 服务 2. 小模型 3. SQL 4. Prompt驱动其他专家 5. 代码生成自动探索
反馈学习	搜索+RAG	搜索+RAG	1. 搜索+RAG 2. 仿制学习	1. 搜索+RAG 2. 仿制学习 3. 增强学习
多智能体协作	无	补充	1. 补充 2. 协商 3. 启发	1. 补充 2. 协商 3. 启发 4. 有效探索的启发
智能体组织模型	无	人控制智能体	智能体控制智能体	自适应组织
国内外产品分类	1. 知识库 2. Voice	1. 自动化 2. 通用助理 3. 开发 4. 软硬件结合类	1. 数字职员 2. 工作流	实验研究为主

◀ 表1 主流Agent技术拆解

未来一年中，普通智能体的数量会快速增加，群智现象可能会先于强大智能体涌现。

- 技术成熟性分析：表中有下划线的相关技术当前论文虽然不少，但在工业环境中还没有成熟解决方案；
- 疑难项技术分析：其中环境模型、自学习技术最难解决。主要是因为其提出时间较长，但在物理生产中没有很好的实际使用方案，另外和大模型关联性不是很强，大模型的进展对这个技术影响小。
- 潜力技术分析：自适应组织、探索、智能提示词、记忆、对话目前看有进一步发展的空间，可能在短期内是拉开Agent水平的关键。
- 发展趋势分析：综合以上分析，Task Agent涉及的非成熟技术相对较少，只有1项；Job Agent涉及到5项非成熟技术，其中包括1个疑难项——环境建模；自我演进Agent涉及关键技术基本全都是疑难项。所以当前Task Job的发展速度可能最快，价值最高。
- 当前产品分析：国内外主要产品集中在Task Agent类型上。

Agent趋势洞察

通过以上分析，我们可以进一步得到如下结论：

- 当前阶段以简单的任务智能体（Task Agent）为主，这种智能体涉及的技术较为成熟，容易复制推广，这和我们的产品实验情况感受一致，这类智能体的数量可能会快速增加。
- 在上述成立的情况下，拉开Agent差距的是记忆、对话等技术。
- 强大的个体智能体由于涉及增强学习和环境模型技术，实现比较困难，这和我们实验时在环境建模等方面投入的成本和最终的效果表现是相符合的。

简单的任务智能体在大模型帮助下，能对其他智能体提供有启发的信息，如果能达到一定的数量，则满足群智涌现的两个必要条件中的一个；其次在大模型的抽象总结能力的帮助下，一个团队的智能体，能把来自不同智能体的多个关联性高的不同信息片段融合，形成信息增加，这样可能满足群智的另外一个必要条件；两个必要条件得到满足后，群体智慧现象可能会开始涌现。

综上，经过学术跟踪和产品的实验探索，以及不同类型Agent的技术分解，我们提出一个Agent的洞察：未来一年中，普通智能体的数量会快速增加，群智现象可能会先于强大智能体涌现。

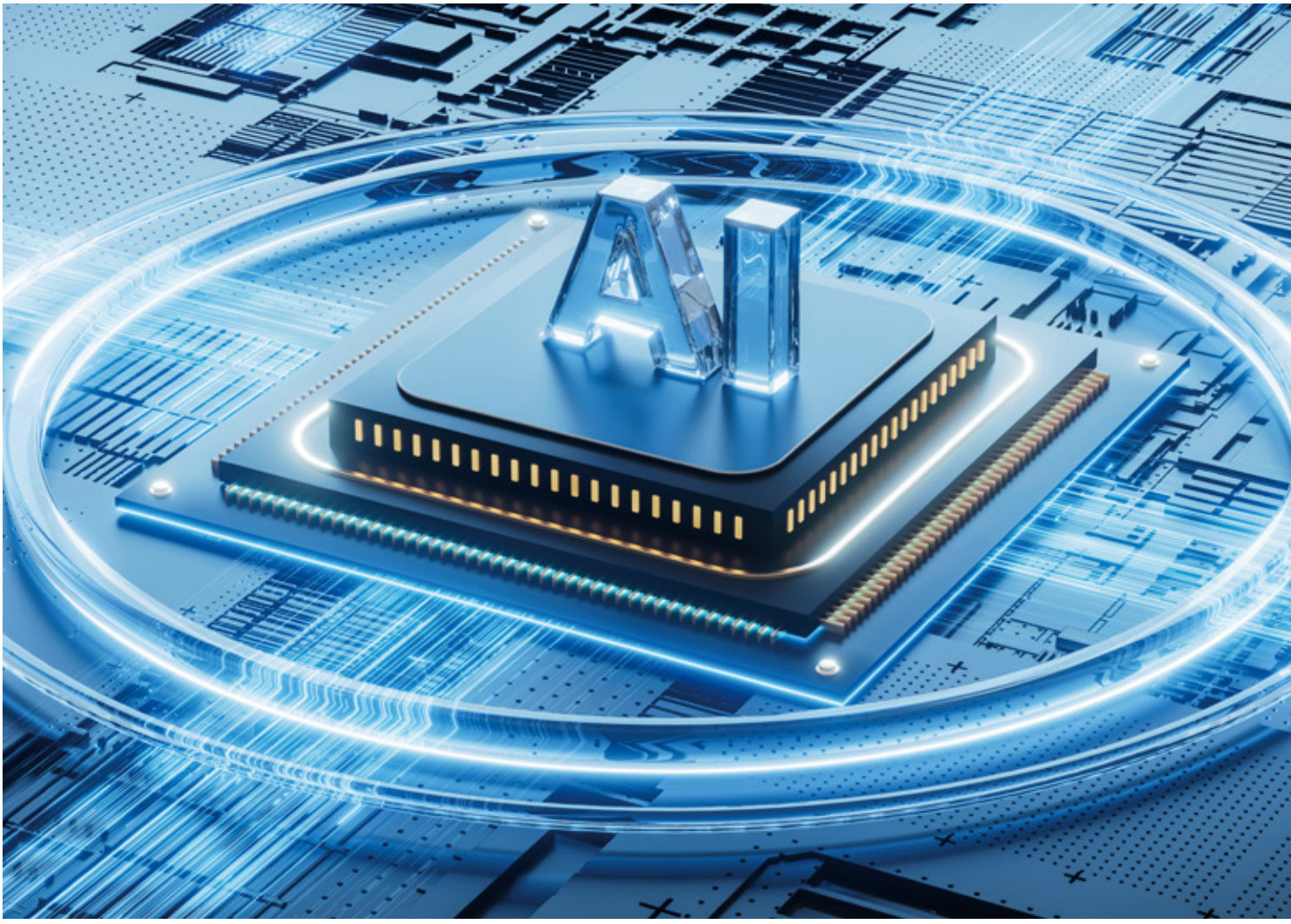
基于这一洞察，我们未来需要进一步考虑在以下方向努力：

- 建设可以快速生成Agent的低学习成本、低技术门槛框架技术；建设多智能体协作，管理群智涌现现象的群智控制中心；
- 追踪Agent演进关键能力，包括环境模型建立、记忆、学习设计，深入挖掘记忆的潜力；
- 建设给企业带来增益的企业数据分析、SOP工作流相关Agent产品。 ZTE中兴

面向大模型，中兴通讯全栈智算解决方案赋能千行百业

中兴通讯　王卫斌，陆光辉

面对生成式AI技术发展的机遇和挑战，中兴通讯坚持“数智经济筑路者”定位，做极致的AI公司，成为大模型应用的企业范例，同时致力于助力千行百业构建端到端的智算基础设施和智能化的企业数字化转型解决方案。



王卫斌
中兴通讯产品规划首席科学家

陆光辉
中兴通讯CCN产品首席架构师

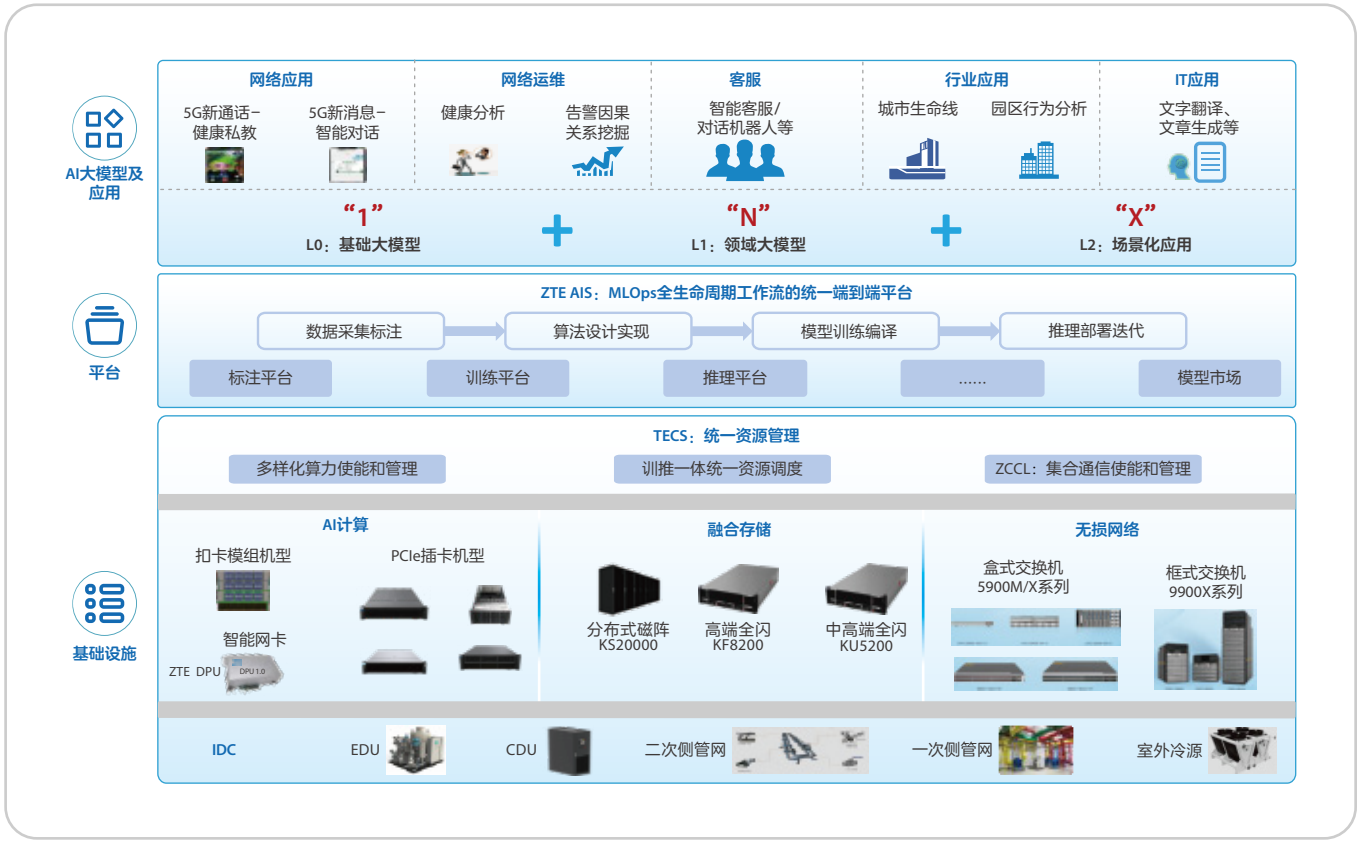
人工智能发展至今，已经经历了三次高潮、两次低谷。2022年11月，OpenAI公司发布的ChatGPT及其采用Transformer算法和预训练大模型的生成式AI技术，使得第三次人工智能技术发展达到了前所未有的新高度，并由此迎来AI大模型技术拐点和炒作高峰。

生成式AI技术，具有生成新内容、模仿人类创造力和创新性的能力，使其在众多领域都能发挥重要作用，从而推动了人工智能领域的繁荣和进步。规模创造奇迹，更大的模型带来更高的智能。随着AI技术不断发展，千行百业将可以利用AI更好地实现运营效率的提升和商业价值的创造，从“数字化”迈向“数智化”。

面对生成式AI技术发展的机遇和挑战，中兴通讯坚持“数智经济筑路者”定位，做极致的AI公司，成为大模型应用的企业范例，同时致力于助力千行百业构建端到端的智算基础设施和智能化的企业数字化转型解决方案。在原有通用算力解决方案基础上，中兴通讯推出星云智算解决方案——Nebula Intelligent Computing Solution，以开放、高效、智能和安全理念为指引，面向训练和推理两类场景，打造智能基础设施、AI平台、大模型及应用三个层次的开放生态全栈智算解决方案，助力运营商智算中心建设，赋能千行百业数智化生产，实现数智化转型（见图1）。

智算基础设施，高效安全

中兴通讯智算基础设施层，包括IDC、AI计算、融合存储、无损网络和资源管理平台，以支撑多样化多层次的智算基础设施建设为目标，从



▲ 图1 中兴通讯星云智算解决方案

▲ 图2 中兴通讯智算一体机解决方案

大模型训练智算中心到训推混合智算中心再到边缘训推一体机，不同层次智算基础设施满足不同场景下的性能、成本和服务的差异化需求。这种多层次的智算基础设施设计，使得解决方案在适应性上更为灵活，为用户提供了更加个性化的选择。

高效为本，大模型单次训练成本高，因此需要高效的智算基础设施。中兴通讯围绕硬件、资源管理和产品方案三个要素，构建高效智算基础设施。

在硬件方面，一方面通过选择高算力、大显存和高速互联的处理器以及高性能并发多元存储来提高系统并行率，从而提升集群有效算力，另一方面自主研发DPU智能网卡，提供超大带宽和超低时延的无损网络，使得整体解决方案具有更高的可靠性和算效性。

在资源管理平台方面，通过资源管理平台向下链接多种异构硬件，满足AI大模型训练和推理的资源高效管理需求。中兴通讯AI资源管理平台产品TECS，为AI训练/推理任务提供JOB调度和智算集群管理，主要包括计算增强（如vGPU技术等）、存储增强（如支持高性能文件存储等）、网络增强（如支持集合通信技术等）和集群管理调度等相关能力。AI资源管理平台产品TECS，是在原有自研通算资源管理产品基础上针对AI大模型训练推理相关需求进行的产品增强，与原有产品功能可分可合，可根据应用场景需求选择融合部署，实现通算和智算统一管理和编排，也可以选择独立部署提供智算资源管理和编排。

在产品方案层面，为了精准解决行业二次训练及实时推理业务场景需求，中兴通讯推出了一站式、开箱即用的训推一体机，如图2所示。一体机集成了计算、存储、网络设备和AI平台软件，支持主流AI框架，为用户降低私域模型的训练和推理成本，降低技术门槛。这意味着用户无需复杂的部署和配置过程，可以快速投入使用，实现了训推资源的灵活分配。

安全为基，人工智能的三个基本要素（算力、算法和数据）中，算力是推动人工智能系统整体发展并快速应用的核心要素和主要驱动力，因此提供安全可靠的算力是关键。在智能算力发展上，中兴通讯致力于构建国内外多渠道供应

链，面向AI大模型训练和推理场景，一方面可提供基于国际主流GPU厂家高性能AI服务器和IB交换机的全套成熟方案；另一方面也联合国内头部GPU厂家进行了大量自研工作，可提供高性能端到端多样化智算方案，包括基于这些头部GPU厂家芯片的高性能AI服务器、盒式和框式RoCE交换机、支持高性能和多元存储（文件、对象、块和大数据等）的分布式存储服务器等。

此外，百千亿级参数规模的大模型训练，由于训练数据大，预训练耗时长，为了保障训练的稳定可靠，避免硬件故障引起训练中断的巨大损失，中兴通讯资源管理平台TECS提供安全可靠的可视化管理平台实现自动监控，并提供断点续训服务，减少训练中断时间，大幅降低训练进程中断的损失。

AI平台，开放解耦

中兴通讯AI平台层以开放、解耦为核心，拥有完备的AI平台产品，AI平台向上提供统一的编程环境及工具链，最大化降低模型开发及迁移成本，助力生态建设。

为帮助开发者和使用者更好地以更具可扩展性的方式开发、训练、评估、实施和更新AI大模型，中兴通讯提供面向大模型的组件化AI平台（AIE，AI Enabler）。作为智算一体机或AI应用的内嵌平台，AIE涵盖数据采集、数据标注、模型训练、模型精调、知识库、编译优化、推理部署等全栈工作流，支持PyTorch等主流AI框架，为客户提供端到端的智算中心解决方案，为AI应用提供模型能力及运行引擎。

AI大模型作为数智化转型的核心技术，直接关系到千行百业在新时代的转型和商业成功，中兴通讯已经做好了充分准备，将与合作伙伴一起拥抱这一重大机遇，让AI普惠千行百业。

大模型及应用，从“通用”到“专用”

针对大模型赋能企业数字化转型，中兴通讯将其发展总结为“1+N+X”的策略，从“通用”到“专用”。

一系列基座大模型

中兴通讯以工程化能力为优势，自主研发星云系列基座大模型，包括NLP大模型和多模态大模型，通过收集大量训练数据，利用无监督或自监督学习方法，从而使其在不同任务和领域中具备优异的理解和表达能力。

N个领域大模型

领域大模型是在基座大模型基础上，通过加入领域Know-How增量预训练等方式，提高专业性能力。在研发领域方面，中兴通讯自2022年开始使用大模型技术全流程助力研发提效，辅助开发人员进行需求分析、产品设计、编程、测试、版本发布以及产品文档编写。目前，中兴通讯研发的编码领域大模型在HumanEval评估的编码类模型能力方面处于第一梯队，编码语言种类多样性和中文编码能力均达到业内领先水平。在电信领域方面，中兴通讯在电信领域拥有大量、高质量的网络运维和业务运营数据，将大量高质量的领域数据以及Know-How知识积累注入到电信领域大模型中，在通信领域的知识超过其他大模型。中兴通讯电信领域大模型支持通信领域的多模态数据，可较好解决覆盖、容量、性能报表、看网讲网等复杂问题。中兴通讯电信领域大模型支持更强的意图引擎，与自智网络高度结合，通过高效的工作流串接，可帮助运营商提升网络运营效率。

X个场景应用

中兴通讯基于领域大模型开发了各种细分应用，如基于计算机视觉（computer vision）大模型，针对水、电、气、热、交通等城市重要基础设施的城市安全风险综合检测预警场景，推出城市生命线解决方案；基于编码大模型，开发出覆盖研发全流程的一站式AI开发助手；基于网络大模型，开发出系列运维工具，如故障运维机器人等，为不同场景提供支持；基于大语言模型开发出短信反诈业务应用。

丰富应用，助力客户数智转型

为助力运营商和合作伙伴构建端到端的智算基础设施和企业数智化转型解决方案，实现数智化转型，中兴通讯推出开放解耦的星云智算方案Nebula Intelligent Computing Solution，提供AI全栈产品，并在智算中心、研发提效、通信领域、反诈治理和城市治理等多个领域得到应用。在通信领域，中兴通讯2023年发布了业界首个基于大模型的“智御”短信反诈治理系统；在行业领域，与业界数百个伙伴展开合作及签署战略合作协议，并在机器视觉、工业生产等领域落地多个项目。AI大模型作为数智化转型的核心技术，直接关系到千行百业在新时代的转型和商业成功，中兴通讯已经做好了充分准备，将与合作伙伴一起拥抱这一重大机遇，让AI普惠千行百业。ZTE中兴

中兴通讯系列化智算服务器方案，助力数字经济蓬勃发展

周赞鑫
中兴通讯服务器及存储产品总工

人工智能（AI）领域正迎来新一轮快速发展，生成式AI对算力的需求迅速增加，这将成为AI计算市场新的增长点和加速器。

中国智算服务器市场

2023年中国智算服务器保持了快速增长。据IDC2023H1数据统计，2023年加速服务器预计发货规模达31.6万台，同比增长11.3%；营收约89.9亿美元，同比增长79.7%；其中GPU加速服务器（智算服务器）占比约90%。IDC预测，2027年加速服务器营收将加速增长，达164亿美元；发货规模将达到69.1万台。

目前，单机配置8或4张GPU加速卡的智算服务器是客户的主流选择，其中Nvidia GPU加速卡依然是市场主流，份额高达90%左右。此外，面向推理方向应用的智算服务器份额约占60%左右。

AI应用对智算服务器的要求

智算服务器相比通用服务器主要有以下特点：

- 高性能CPU：AI训练和推理需要大量的计算资源，需要配备高性能CPU，以满足大数据集的处理需求。
- GPU加速卡：GPU可以提供比CPU更高效的并行计算，从而加速深度学习模型的训练和推理，插卡型GPU加速卡可以满足大部分中小模型训练&推理应用需求，单台服务器支持4~8张GPU加速卡实现并行处理，可提升计算性能和效率。
- 大容量内存：具有足够容量的内存可以加速数据流和算法处理速度。
- 高带宽网络接口：需要高速网络带宽（100Gbps及以上），以便在训练过程中传输大量数据。

AI大模型的兴起对智算服务器提出了更高的要求，特别是大模型训练计算量巨大，单个GPU无法满足训练算力需求，需要使用单机多卡或多机集群实现TP/DP/PP等并行训练。大模型对智算服务器的特殊要求体现在以下几个方面：

- 高性能&大显存GPU：大模型需要大量的并行计算能力，且需要存储大量的参数和梯度信息，因此需要高性能&大显存GPU来进行训练和推理。
- 机内GPU高速互联：单机多卡TP并行对智算服务器的多个GPU之间通信带宽有极高的要求，需要使用支持高速互联通道的扣卡型GPU加速卡，实现机内8卡高速互联，以加速数据传输和模型同步。
- 机间高性能互联网络：采用多机集群时，为了充分发挥GPU集群计算资源的强大算力，机间参数面互联网络需采用高速多轨道流量聚合架构。一方面，要求PCIe5.0插槽以便使用200/400G高性能、低延迟的IB/RoCE网卡；另一方面，要求至少10个以上网卡插槽，管存面至少2个网卡，GPU和参数面网卡按照8:8配比，以实现多台智算服务器间相同位置GPU卡所连参数面网卡都归属于同一交换机，优化通信效率，加速并行传输。

◀图1 中兴通讯智算服务器“3+2+3”方案

- 高速内存&存储：大模型训练过程中需要快速读取和写入数据，需支持DDR5内存和NVMe SSD等高速部件提供更高的数据传输速度和更低的延迟，从而提高训练效率。
- 液冷散热：扣卡型GPU加速卡的超高算力密度导致智算服务器功耗激增，风冷方案限制了智算数据中心的算力密度，且无法满足节能降耗要求，液冷散热是必选方向。

鉴于大模型训练推理对智算服务器提出的特殊要求，需要设计专用的智算服务器以适配扣卡型GPU卡和机内机间高速互联网络，并进行合理的配置和优化，使其不断适应新的挑战和要求。

中兴通讯智算服务器“3+2+3”方案

为应对人工智能的快速规模发展，中兴通讯推出“3+2+3”智算服务器解决方案，全面满足各行各类客户的AI全场景应用需求（见图1）。

基于3大CPU平台

中兴通讯针对3大CPU平台都已推出不同形态的智算服务器，满足客户的多样性CPU选择需求，包括业界主流的国外X86架构CPU平台、国产X86架构CPU平台，以及中兴通讯自研ZFX CPU平台。

支持2种GPU形态

中兴通讯智算服务器支持插卡型GPU加速卡和扣卡型GPU加速卡（支持卡间高速互联），比如SXM扣卡型GPU加速卡（Nvidia）或OCP OAM扣卡型GPU加速卡（壁仞、寒武纪等）。

面向3类应用场景

中兴通讯系列化智算服务器具有多种组合方式，满足大、中、小不同等级的AI模型训练、推理场景。

- 小模型训练&中小模型推理场景：采用通用机架服务器，单服务器配置4张双/单宽全高GPU插卡或6/8张单宽半高GPU插卡，对应中兴通讯R53xx/59xx系列服务器。
- 中小模型训练&大模型推理场景：采用专用插卡型智算服务器，单服务器配置8张（或10张）双宽全高全长GPU卡或16张（或20张）单宽全高全长GPU卡，对应中兴通讯R65xx系列智算服务器。
- 大模型训练场景，采用专用扣卡型智算服务器，单服务器配置8张SXM/OAM GPU卡，为满足多节点集群计算需求，GPU&参数面互联网卡&NVMe SSD支持1:1:1配置，对应中兴通讯R69xx系列智算服务器。

智算服务器市场正在经历一个快速发展的阶段，已成为服务器市场中的高增长领域，且未来几年的复合增速也有望保持在较高水平。中兴通讯推出的系列化智算服务器，为用户提供优质、高效的最强算力解决方案，以坚实的智算基础设施助力数字经济进一步蓬勃发展。ZTE中兴

多样化的AI芯片

高振中
中兴通讯算力及核心网硬件总工

1956年，在美国达特茅斯学院的夏季研讨会上，麦卡锡、明斯基等科学家首次提出AI概念。此后的60余年中，AI发展历经多次沉浮，经历了漫长的探索期。直到2015年，AI的视觉识别精度超越人类，开始在视频领域规模商用。2022年，现象级产品ChatGPT横空出世，推动大模型成为产业应用的主要方向。

AI芯片是AI发展的重要基石，经历了两个主要阶段。2012年以前，AI研究和应用主要基于CPU；2012年，多伦多大学的Alex Krizhevsky首次将GPU用做AI，只用了4颗英伟达Geforce GTX580（同时期的谷歌方案采用16000颗CPU）就在ImageNet竞赛中获得冠军，震撼了学术界，开启了AI芯片多样化的大门。

AI芯片的关键需求

从功能上，AI芯片可分为训练芯片和推理芯片两大类。训练，指的是通过向模型提供大量标注或未标注的数据，并基于优化算法来调整模型的参数，使其能够从数据中学习到相关的模式和规律。推理，是指将已经训练好的模型应用到实际场景中，进行预测、分类或决策。

训练芯片的关键需求是如何提供更高的AI算力，降低模型的训练时间。大模型趋势发生以来，模型的数量、规模，在短短几月内剧增，百亿千亿级别大模型飙升至数十个，万亿参数大模型已正式诞生。训练所需的计算量也随之呈指数增长，且翻倍时间约三四个月，远快于芯片工艺的摩尔定律，导致大模型的训练时间不断被拉长。以OpenAI为例，2022年训练一次1750亿参数的GPT-3模型大概需要1024块A100 GPU运行34天，2023年训练一次1.8万亿参数的GPT-4模型大概需要25000块A100运行约100天。相比GPT-3，GPT-4的训练时间增加近2倍。

推理芯片的需求呈现多样化的趋势，主要由业务场景决定。如线上问答场景，AI芯片的算力需要跟上人类的阅读速度（平均每分钟阅读250个单词，最大1000个单词）。如5GC的新通话场景，需要在AI算力的基础上，叠加语音编解码和图像处理能力。

AI芯片的部署位置

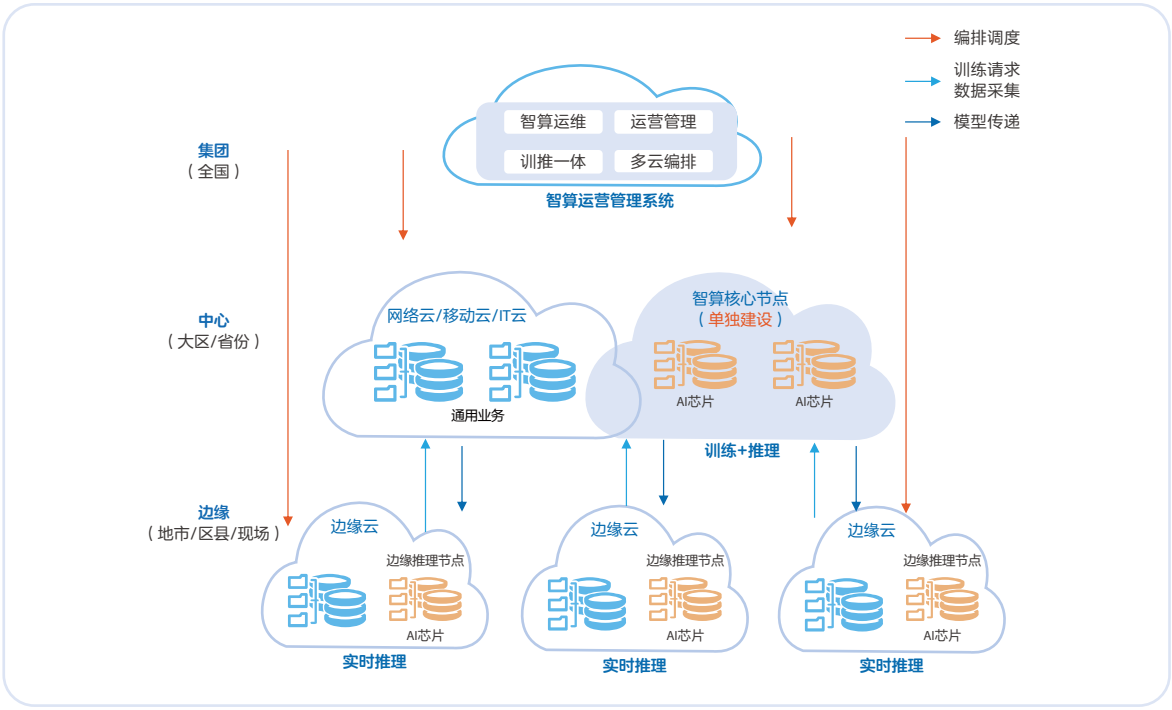
AI芯片主要部署在云侧和端侧。云侧一般指云端数据中心，端侧一般指个人可接触或使用，不需要远程访问的设备内（如手机、PC等）。

云端数据中心，以运营商的智算中心为例，可进一步细分为集团节点、中心节点和边缘节点（见图1）。集团节点用于智算运营管理；中心节点用于训练和非实时推理；边缘节点与边缘云混合建设，用于实时推理。

端侧AI具备安全性、独立性、低时延、高可靠性等特点，能很好地完成各类AI推理任务。目前，多个大模型均已推出“小型化”和“场景化”版本，其轻量化提供了端侧运行的基础。

AI芯片的技术路线

以GPU为代表的通用并行计算架构以及以针



◀图1 运营商的智算中心部署架构

对AI领域加速为代表的专用定制架构，是目前两大主流AI芯片技术路线。

GPU设计初衷是进行图形渲染。图形处理涉及到相当多的重复计算量，因此GPU芯片上排布了数以千计专为同时处理多重任务而设计的图像计算核心，正好和AI运算的数据量规模大、可并行的特点相匹配。

不同于GPU，AI专用芯片是一种针对AI运算的专用处理器，内部以AI专用核为主，相比GPU，减少了视频渲染、高性能计算等功能。AI专用芯片在功耗、体积等方面有一定的优势，但由于是专用定制的设计思路，开发周期较长，在通用性和可编程性方面也弱于GPU，整体处于多而不强的局面。

未来展望

万亿大模型已成为事实，不远的未来很可能出现十万亿的超大模型。随着模型规模的不断增长，无论是GPU还是AI专用芯片，性能和功耗都出现了瓶颈，导致云端数据中心的规模不断增大，从千卡演进到万卡。功耗不断增加，需要引入液冷才能满足散热要求。另一方面，模型在端侧落地也面临着功耗问题，AI手机和AI PC作为两种典型的端侧设备，功耗的增加都会影响消费体验。针对以上问题，下一代AI芯片设计有以下方向：

在计算架构层面，引入存算一体架构，降低功耗。当前的主流GPU和AI专用芯片均采用冯·诺依曼架构，计算和存储分离，芯片60%~90%的能量消耗在数据搬移过程中。存算一体架构将内存与计算完全融合，避免数据搬移，可大幅降低功耗。

在芯片实现层面，采用Chiplet和3D堆叠技术，提升芯片良率和性能。Chiplet将芯片分割成多个具有特定功能的芯粒（如计算芯粒、存储芯粒等），各种芯粒选择最适合的半导体制程进行分别制造，实现最优的良率，再通过高速总线将彼此互联，最终集成封装为一颗芯片。3D堆叠把芯片从二维展开至三维，在不改变原本的封装体积大小的基础上，通过在垂直方向进行芯粒叠放，增加芯片内的芯粒数量，进而提升芯片性能。ZTE中兴

面向AI大模型训练的高性能网络

杨茂彬
中兴通讯Cloud&AI网络规划总工

ChatGPT的火爆，催生了人工智能从判决式到生成式的跨越式发展，百亿千亿参数规模的AI大模型训练如火如荼的展开，对高性能网络提出了迫切需求。AI大模型训练依赖于分布式并行计算，包括数据、流水和张量并行，为了最大化发挥GPU算力，需要将通信时间开销占比控制在5%以内，这就要求AI大模型训练的网络必须是满足零丢包、低时延、高吞吐大带宽以及大规模组网的高性能网络。

当前高性能网络主流解决方案

应用于AI大模型训练场景的两大主流高性能网络技术为IB网络和RoCEv2网络。

IB网络起源于上世纪九十年代，原旨在替代PCI总线技术。然而，它在高性能计算和AI领域的数据中心中意外受到欢迎并得到广泛应用。IB网络通过信用流控机制实现了无丢包传输，并提供QoS服务质量以优化特定流量。尽管IB网络有诸多优点，但由于其配置、维护和扩展的复杂性，以及需要专门的硬件和子网管理器，导致成本较高，并不像以太网那样普及。

RoCEv2网络是基于以太网演进的，它允许通过封装RDMA帧在IP/UDP报文中实现远程直接内存访问。当数据包抵达GPU服务器的RDMA网卡时，数据可被直接传输到GPU内存，绕过CPU以降低时延。另外，通过部署DCQCN等拥塞流控方案，降低RoCEv2网络的拥塞和丢包。RoCEv2网络为统一承载网络设计，满足高带宽、高弹性组网，云化服务化和扩展性支持较好，是国产化高性能网络的必选之路。

当前RoCEv2网络拥塞及流控机制问题分析

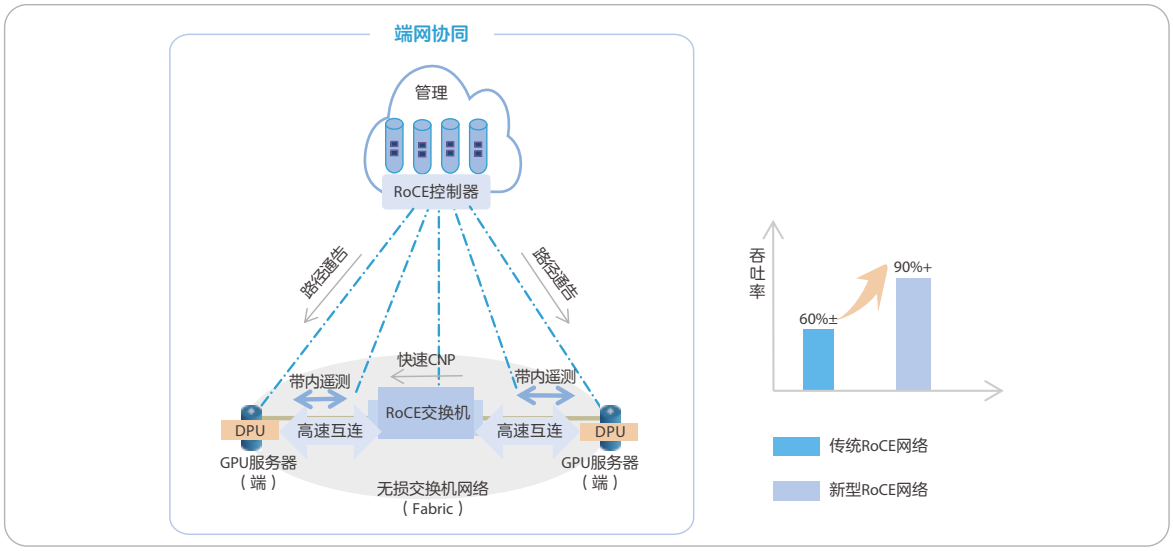
RoCEv2网络中，DCQCN是最常用的拥塞控制算法，它通过交换机的ECN标记来检测并指示网络拥塞。交换机在发现拥塞时，会概率性地在数据包上加上ECN标记，RDMA网卡则根据这些标记来判断网络状况，并通过CNP报文来调整数据传输速率。DCQCN算法公平高效，非常适合高性能计算和AI学习等需要高吞吐、低时延的应用场景。

但DCQCN也存在如下不足，导致网络吞吐率徘徊在50%~60%：

- 拥塞指示不够精确：ECN标记只有1bit，无法细致区分不同程度的拥塞；
- 速率调整反应缓慢，精度不足：仅依赖CNP报文来调整速率，缺乏其他网络信息反馈；
- 没有结合流量特征调优：没有考虑长短流的不同特性，以及调度间隔周期；
- 没有考虑多路径均衡调度：多打一流量分布不均，未能充分利用AI网络多路径带宽资源。

中兴通讯RoCEv2网络端网协同创新方案

传统DCQCN网络因其拥塞标记信息粗略和端侧与网络侧流控机制的相对独立，难以在高吞吐、满负荷的网络环境下避免拥塞、丢包和时延等问题。为提升高性能网络的传输性能，中兴通讯提出了RoCEv2网络端网协同创新解决方案，通过端网协同联动机制实现精准、快速的拥塞控制和流量调度算法，使网络的吞吐率提升到90%以



◀ 图1 中兴通讯RoCEv2网络端网协同创新方案

上（见图1）。该方案在拥塞控制和精准流控两个方向实现端网协同创新。

端网协同新型拥塞控制技术

网络设备通过快速CNP和带内遥测技术及时准确的向端侧提供链路拥塞信息，实现新型拥塞控制技术。

- 快速CNP技术

传统DCQCN网络，当网络设备出现拥塞时，相关链路的数据报文会打上ECN标记，目的端网卡收到ECN标记报文再向源端网卡发送CNP报文，源端接收到CNP报文后进行调速，该过程周期较长，调整速率响应缓慢。我们提出快速CNP解决方案，中间交换机检测到拥塞时，会迅速向源端网卡发送包含详细拥塞信息的CNP报文，源端网卡能更快地利用这些信息精准调整流量，从而迅速缓解网络拥塞。

- 基于带内遥测机制的精准拥塞流控技术

传统DCQCN中的ECN拥塞指示只有1bit，无法精确表达链路拥塞程度，源端也就无法进行精准流量调控。我们提出了基于带内遥测技术携带更多路径负荷信息的解决方案，中间设备在遥测报文中填充可用带宽、队列深度、时间戳、发送字节数等信息，端侧收集齐路径所有网络设备的遥测信息后，根据训练调优后生成的流量调度算法对流量进行实时精准调控，使端到端路径流量达到高吞吐、低时延、无拥塞的最佳状态。

端网协同多路径精准流控技术

网络侧与端网配合，充分利用RoCEv2网络ECMP路径和多种负载均衡技术，提升数据传输效率。

- ECMP路径端网协同通告

AI大模型训练数据中心的RoCEv2网络采用胖树CLOS架构，拥有丰富的ECMP路径。RoCEv2网络控制器掌握全网拓扑，并向端侧同步ECMP路径信息，以优化数据传输，提升网络效能。

- 根据流量特征匹配的负载均衡技术

端侧根据流量特征（如老鼠流、大象流）选择不同的负载均衡技术，通过报文哈希或源端口散列进行选路，并可根据网络负载实时调整策略，以提升数据传输效率。

随着AI大模型参数从千亿迈向万亿，以及AI芯片算力供给受限，万卡规模的智算集群网络成为必然，大规模组网场景下的精细化端网拥塞控制成为业界亟待解决的共同挑战。中兴通讯提出的RoCEv2网络端网协同创新解决方案，旨在改进网络的吞吐率，强化AI大模型训练网络性能，进一步释放AI算力，提升AI大模型训练效率。ZTE中兴

中兴通讯智算AI平台，助力大模型训推工程化

周祥生
中兴通讯AI平台研发经理

孙文卿
中兴通讯AI算法工程师

在数据爆发式增长、算法性能持续提升以及算力产品不断跨越迭代的背景下，我们正处在AI引领产业全方位变革的阶段。此过程中，AI平台扮演着至关重要的角色。AI平台通过集约化管理数据、算力、算法和服务，将作坊式、离散的算法研究转为标准化、自动化的生产流程，避免重复造轮子，让用户聚焦于智能业务中的高价值问题。

AI平台成为企业智能化转型的关键基础设施

AI平台充当着连接计算力与算法的关键桥梁，它不仅将算法开发过程中的共性需求工具化、流程化，而且向用户提供定制化的能力和服务。同时，平台还要具备共享复用、高效训练推理、快速交付、持续迭代的特性。为此，中兴通讯开发了异构算力管理与AI模型训练推理平台——智算AI平台。平台由硬件适配层、引擎层、服务层和能力层组成（见图1）。

- 硬件适配层：成千上万的GPU、CPU提供算力，既支持国际主流显卡，也对国产显卡进行了适配。
- 引擎层：包括机器学习引擎、超参调优引擎、训练引擎、编译引擎以及推理引擎。该层集成了多种高性能训练和推理引擎框架，如Tensorflow、Pytorch、Oneflow、Deepspeed等。
- 服务层：服务层包括数据集管理、数据标注、模型训练、超参调优、模型评测、模型

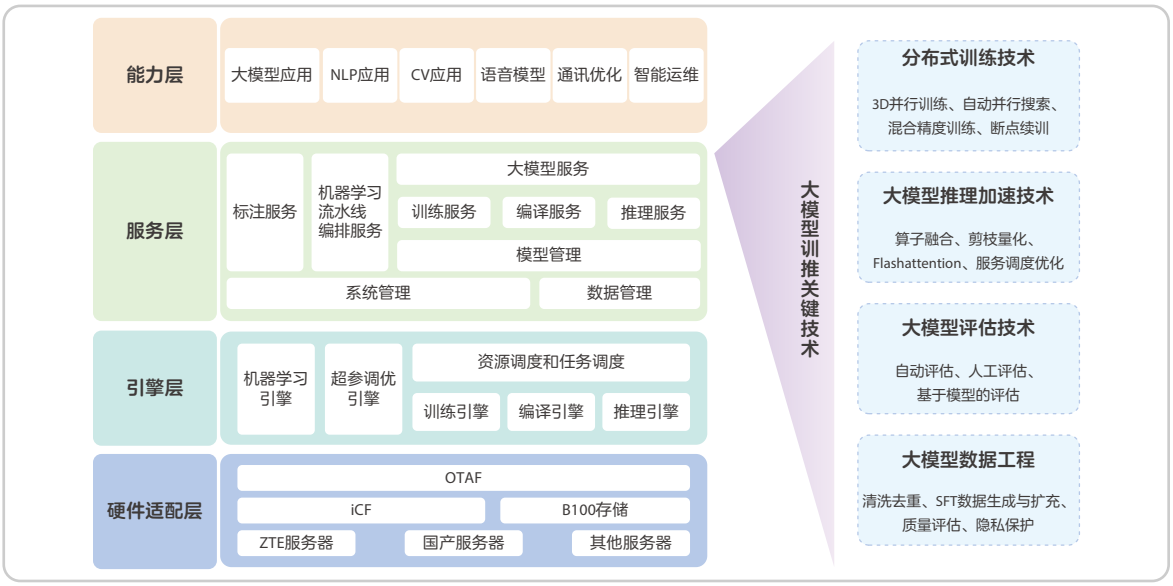


图1 智算AI平台 ▶

编译及模型推理等服务，涵盖AI模型端到端的全业务。

- 能力层：内置多种解决实际问题的算法包、推理包，供直接部署和调用。

从基础算力和调度技术、深度学习框架及引擎，到NLP、视觉、语音、大模型等感知、认知能力，AI平台作为推动企业智能化转型的关键基础设施，不仅整合了计算硬、软件工具，还提供了AI算法的研发接口。通过这种全面的整合，AI平台大大提高了资源的利用效率，加速了AI的落地应用。

从“大炼模型”转向“炼大模型”

当前，在AI落地场景中，许多解决中间任务或特定领域任务的小模型正被通用性更强的大模型所取代，人工智能全面向AGI（artificial general intelligence）转型。伴随而来的，是大模型对完备、稳定且高效的数据存储和清洗方式、训练推理技巧、集群资源的需求日益增长，这给AI平台的建设提出了新的挑战。

大模型的出现，带来了模型结构和训练-推理（训推）范式的统一化。首先，Transformer结构一直是骨干模型基本部件的首选；其次，在训练和推理方法上，以大语言模型为例，OpenAI最初提出的训练方法（包括预训练、指令精调、强化学习精调）和推理方法（如随机采样解码）仍是大模型训推的主流解决方案。

然而，这种结构和应用范式的统一并没有缩小行业平均水平与AI头部公司之间的差距，而是将AI竞争的焦点从算法研发创新转移到了大模型训推工程化的规模和效率的竞争上。这就使得集成大模型训练和推理关键技术成为AI平台建设的首要需求。

大模型训推工程化关键技术

大模型训练和推理过程中的关键技术包括分布式训练技术、大模型推理加速技术、大模型评估技术和大模型数据工程。

- 分布式训练技术：分布式训练能将训练扩展到多个AI硬件上，从而突破单个硬件内存和算力的限制。中兴通讯智算AI平台已集成3D混合并行技术，以及自主研发的自动并行工具，这些工具支持数据并行（DP）、张量并行（TP）、流水线并行（PP）以及激活重计算等大模型训练技术，并能根据集群和模型特点自动调整并行超参。
- 大模型推理加速技术：大模型推理加速技术是降低推理过程中显存消耗和计算延迟的综合技术。智算AI平台从服务调度、显存优化、量化压缩及算子融合等多个方面提高推理效率。在中兴通讯推出的业界首个基于大模型的“智御”短信反诈治理系统中，智算AI平台所提供的推理方案相比于业界通用方案，成功将推理时延降低30%。
- 大模型评估技术：大模型评估方法与传统模型区别很大。为此，智算AI平台一方面提供了全面的客观评估数据集，从多维度评估大模型的性能。其次，平台融合了基于模型的评估机制，评估生成内容的语义准确性和逻辑连贯性。
- 大模型数据工程：高质量训练数据能够缓解大模型幻觉问题，缩短训练周期。智算AI平台提供了Model-in-the-loop的数据标注、SFT数据生成与扩充、数据清洗与去重、质量评估、隐私保护等智能化数据工程流水线。

凭借大模型工程化关键技术的支撑，中兴通讯的智算AI平台在公司内部及与国内运营商客户的合作中已取得初步成效。在公司层面，AI平台支持了电信、编码、CV以及多模态等领域多个大模型的训练。在运营商客户方面，AI平台完成了客户集团31个省份训推集群建设，提供了模型训练、模型管理和推理服务等九大核心功能，成为客户AI开发的重要工具云。ZTE中兴

大模型赋能通信运维智能提效

何伟
中兴通讯MANO产品规划总工

随着行业数字化转型的加速，通信领域运维需求日益复杂，智能运维成为数字化时代保持竞争力的关键因素之一。然而，随着业务的快速发展和技术的不断更新，传统运维方式已难以满足通信设备运维的需求。大模型技术的出现为智能运维领域带来了突破，它能够提供更人性化的人机交互模式，同时能够处理海量格式化数据，提供高精度的分析和预测，为智能运维提供强大的技术赋能。

大模型在智能运维中的应用

大模型技术在通信领域智能运维中得到了广泛的应用拓展，主要包括：

- 运维知识问答：大模型对于通信知识有存储、记忆、理解和运用能力，灵活结合上下文的理解，能够准确检索和提取相关信息，生成问题答案，反馈给提问运维人员；同时大模型能够不断更新和修正自己的知识库，从而保持与最新知识的同步。
- 故障异常检测：利用大模型智能算法和模型，系统对采集到的数据进行处理和分析，可发现与正常状态不符的异常数据或行为；通常涉及特征提取、数据建模和分类、异常判断标准制定等步骤。
- 根因定位：在异常检测的基础上，进一步对异常数据进行深入分析，推断出导致异常的原因和位置，从而确定故障的具体类型和位置；这需要运用各种诊断技术和方法，如故障树分析、专家系统等。
- 故障预测与预防：大模型可以对海量的历史运维数据进行学习，从中发现故障发生的规律和趋势，建立故障发生模型。基于此模型，通过对实时数据的监控和分析，大模型亦可以预测潜在的故障风险，提前发出预警，使运维人员有足够的时间采取预防措施，降低故障率。

相较于传统AIOps，大模型给予智能运维更进一步的能力加持，如交互更简单、知识覆盖更全面、能够实现故障自我学习、模型架构更灵活等，使用门槛更低，并且实现了运维能力的不断泛化。

中兴通讯核心网运维大模型体系架构及关键技术

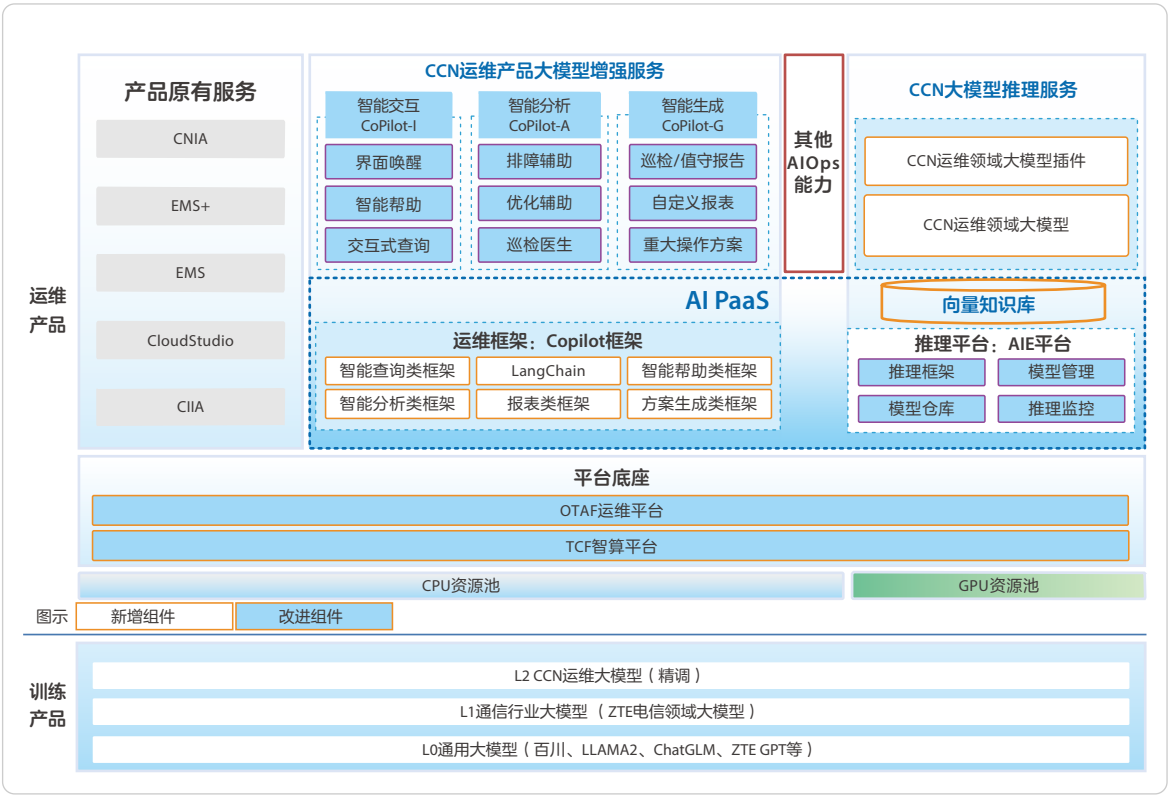
中兴通讯核心网运维大模型基于中兴通讯自研训练的电信领域星云大模型，使用高质量的语料对基座模型进行精调，生成面向核心网及网络云的运维大模型（见图1）。运维大模型应用有三大类能力：

- 智能交互（CoPilot-I）：包括专业知识问答、网络健康度查询、关键指标信息查询等功能；
- 智能分析（CoPilot-A）：包括故障分析辅助、网络优化辅助、巡检报告排查等功能；
- 智能生成（CoPilot-G）：包括巡检报告生成、操作方案生成、网络报表生成等功能。

为满足以上大模型运维能力，在中兴通讯核心网及网络云各运维大模型产品中，分别应用了当前热门的关键技术，包括RAG（retrieval-augmented generation，检索增强生成）、多智能体协同等。

- RAG（检索增强生成）

要完成更复杂和知识密集型的任务，需要构



◀ 图1 面向核心网及网络云的大模型运维产品系统架构图

建一种精度更高、更可靠的系统，并且缓解大模型“幻觉”问题。RAG是一种大模型的关键技术，它通过从数据源中检索信息来辅助大语言模型生成答案。RAG技术可以极大地提升内容的准确性和相关性，有效缓解幻觉问题，提高知识更新的速度，并增强内容生成的可追溯性。RAG已成为当前解决大模型获取外部新知识问题最受欢迎的系统架构。

- 多智能体协同架构

多智能体协同是指多个智能体在共享环境中通过相互通信和协作，实现协同行动以达成共同目标的过程。每个智能体都具备一定的自主性和智能性，能够根据环境信息进行感知、决策和执行。多智能体协同通过相互之间的交互与合作，使整个系统能够从各个智能体的优势和特长中受益，实现更高效、更智能的决策和行动。基于多智能体协同架构，我们可以创建知识专家、故障专家、值守专家、方案专家等独立智能体个体，通过相互协同，共同构建网络的智能运维体系架构。

大模型在智能运维中的挑战与未来发展

尽管大模型在智能运维领域具有广泛的应用前景和优势，但仍然存在一些挑战。例如，如何提高大模型的自适应能力、降低模型的复杂度、解决数据隐私和安全问题等。

未来，随着技术的不断进步和应用场景的不断拓展，大模型在智能运维领域的应用将会更加广泛和深入。例如，随着边缘计算的普及和发展，大模型将逐渐向边缘端迁移，实现更高效、实时的智能运维；同时，大模型将与机器学习、深度学习等技术结合得更加紧密，进一步提高智能运维的效率和精度，大模型将面临更多的数据挑战和机遇。因此，我们需要不断地探索和创新，结合具体场景和需求进行应用和实践。同时，也需要进一步加强相关技术的研究和开发，推动智能运维技术的进步和发展。ZTE中兴

大模型+5G，赋能行业智能化

王朝营
中兴通讯CCN规划架构师

刘西亮
中兴通讯CCN产品规划总工

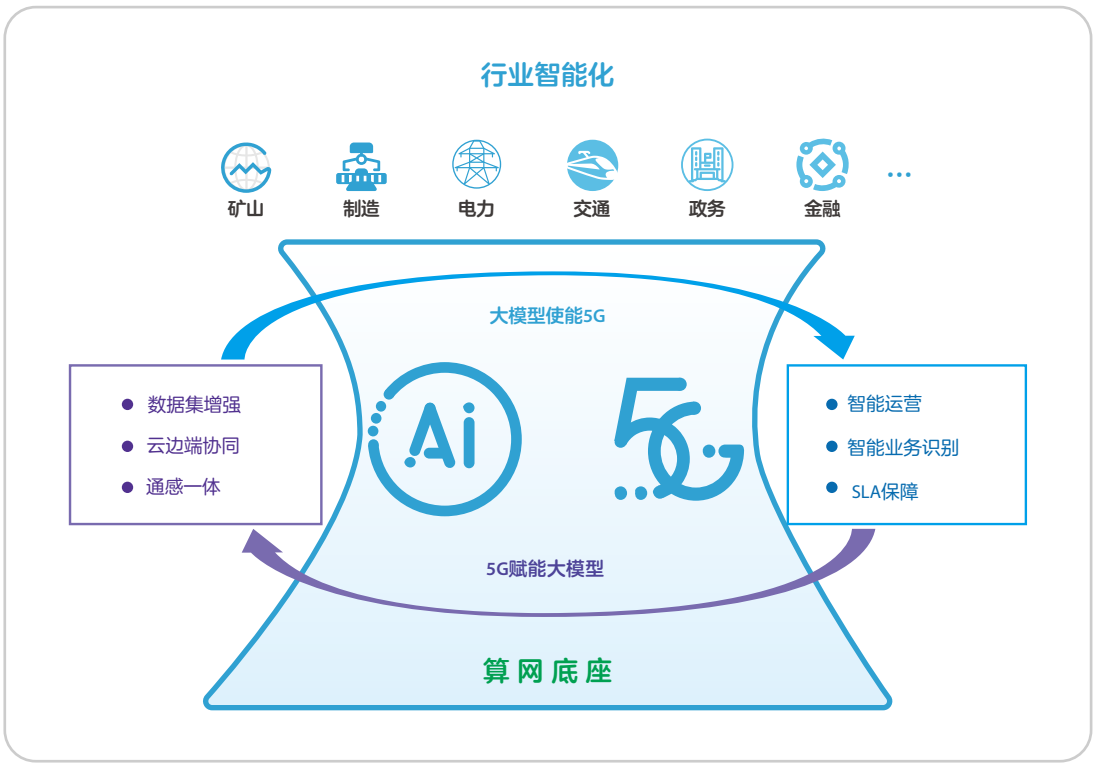
2022年下半年以来，ChatGPT为代表的AI大模型热潮席卷全球，标志着人工智能技术正式进入大模型时代。然而，模型发展和应用在2023年呈现冰火两重天的局面。一方面，模型发展一骑绝尘；另一方面，模型应用发展不温不火，没有跟上大模型的发展速度。大模型作为一个新兴技术，强力推动了基础智能科学的快速发展，但广泛的应用还需要跟行业、其他技术结合。本文探讨大模型结合5G在行业中的应用新范式，探讨“5G+模型”推进行业智能化的发展的前景和演进之路。

国内连续举办六届“绽放杯”5G应用征集大赛，促进数字经济与实体经济深度融合。在此过程中也暴露了较多问题，如5G网络技术门槛高、行业使用者维护及运营困难、网络数据价值无法体现等。模型应用在行业，同样遇到很多困难，优质语料获取、快速行业部署等问题依旧突出。

行业智能化发展需要大模型和5G相辅相成，共同推进：大模型使能5G，推动行业数智化转型；5G赋能大模型，加速模型应用落地（见图1）。

大模型使能5G

5G向行业拓展中遇到的问题和困难，主要包



▲ 图1 大模型和5G在行业中的协作框架

括网络运维复杂、业务SLA不确定性、针对特定业务差异化处理手段欠缺等。在大模型引入5G网络之后，上述问题均可解决。

大模型助力5G网络智能运营

行业用户部署5G网络后，业务开通及日常运维需要专业知识支撑及人员储备，增加了用网成本及投资，将大模型应用在5G运维领域，实现网络智能运营，可大幅降低行业投入。大模型帮助行业用户将业务需求转换为网络规划及配置，实现业务的意图驱动，提升业务开通效率；通过分析大量的网络日志/告警数据，大模型可以识别或预测故障，并提供相应的解决方案；大模型通过深度挖掘用户行为数据，可以预测用户的网络需求和业务偏好，生成相关业务套餐，帮助行业用户更好地运营网络。

大模型实现网络SLA保障

不同行业应用对网络的带宽时延需求不同，比如智能制造领域的视频监控类业务有大带宽需求，生产控制类业务有低时延需求，因此需要5G网络针对不同的用户或业务提供相应的SLA保障。在核心网侧通过大模型，可以对用户的业务质量进行评估及预测并及时生成业务保障策略；在无线侧，大模型通过分析无线信号和传感数据，实现更精确的资源分配和调度，提高网络的效率和质量。

大模型实现智能业务分析

行业应用场景，有数据不出园区的保密性要求。通过引入大模型及AI算法，5G网络对接入的业务流量在网元内实时分析及智能标注，进而进行差异化业务保障，满足行业特定业务需求，并满足数据保密性要求。

5G赋能大模型

5G网络作为数字化时代的信息基础设施，具备高带宽、低时延、海量接入特性，极大拓宽了行业的数据采集能力，丰富了大模型的数据集。同时，随着5G网络广泛使能行业领域，将带动大模型深度参与行业数智化转型。

5G网络增强大模型数据集

在行业智能化的进程中，数据是构建大模型竞争力的核心要素。5G网络支持IP/LAN等多种接入方式，解决了移动性限制，使得行业内的语音、图像、视频等数据可以随时随地接入，极大丰富了大模型的数据集。同时，5G网络通过用户鉴权、传输加密等技术，保证了接入数据的安全性及有效性，大幅度提升大模型的数据集质量。

5G网络拓展大模型应用场景

5G技术是充满活力的创新引擎，正在千行百业大展身手。5G的应用拓展了大模型的应用场景：一方面，通过部署5G专网，能够支持一线生产现场传感器、摄像头等监控设备的异构海量连接，这些新的场景也催生了对行业大模型的需求，通过大模型的多模态分析能力，实现智能精准化故障预警和风险管理，大幅提升生产效率；另一方面，5G网络也促进了大模型的协同能力，当前大模型主要以云边协同为主，随着5G网络的应用，大模型的智能体（Agent）可以延伸到5G智能终端，形成云边端协同体系。

为推进大模型+5G的行业应用，中兴通讯推出了AiCube算网一体机。算网一体机由算力硬件、云平台、AI平台组成。算力硬件能够适配多厂家多型号CPU/GPU；云平台实现资源统一管理，算力按需分配，具备部署5G网络和大模型的能力。AI平台为运营商和企业用户提供数据管理、模型开发、模型训练、模型推理、应用统计等工具，方便用户使用，提升效率。

作为全栈智算解决方案提供商，中兴通讯将与运营商、行业用户紧密合作，持续推动大模型+5G应用，共同构建智算新生态，迎接智算新未来，为数字经济的发展注入新的动力。ZTE中兴

双剑合璧，“智御”反诈大模型护民生

黄小兵
中兴通讯消息产品规划总工

王巍
中兴通讯产品规划专家工程师

据国家反诈中心的公开资料显示，近年来电信网络诈骗已经成为发案最多、上升最快、涉及面最广的犯罪类型。截至2022年底，公安部门共破获电信网络诈骗案件115.6万起，抓获犯罪嫌疑人155.3万名，止付冻结涉案资金9165亿余元。电信网络诈骗态势日益严峻，严重威胁大众人身安全与财产安全。

短信诈骗监控难点与挑战

诈骗短信作为电诈最常见的手段之一，其内容不断变异和升级，以穿透电信运营商的短信监控系统处理策略：

- 通过组合变异、转义字符、谐音、形近等种种手段突破关键词规则；
- 通过汉字、字符、数字的变异组合来表达标准URL和号码，突破现网正则监控策略；
- 通过海量号码池规避流量和关键字门限；
- 通过拨测等方法，一点突破，海量发送。

传统治理方案升级周期长，面临巨大挑战，策略过松拦截效率低，策略过严影响用户正常通信需求。

AI大模型开启新技术革命

2022年11月30日，OpenAI公司发布ChatGPT，其成为有史以来最短时间用户量突破1亿的应用。ChatGPT基于Transformer神经网络架构，在大规模自然语言、序列数据和目标检测等多个深度学习领域取得重大突破，并可通过大量的语料库来训练模型，使得大模型具备泛化知识，能深刻理解语言和对话；此外，可针对性训练解决特定领域的问题，迅速适应新的任务和场景。

对于诈骗短信的精准识别，首先需要能深刻理解自然语言；其次，需要对敏感信息进行分类，并识别内容真正的意图；第三，对于不停变换的诈骗短信内容，需要能对样本学习，完成知识和模型的动态升级。这些正是Transformer架构大模型擅长的技术，基于大模型研发新型短信反诈技术和产品，非常值得快速技术穿刺和尝试。

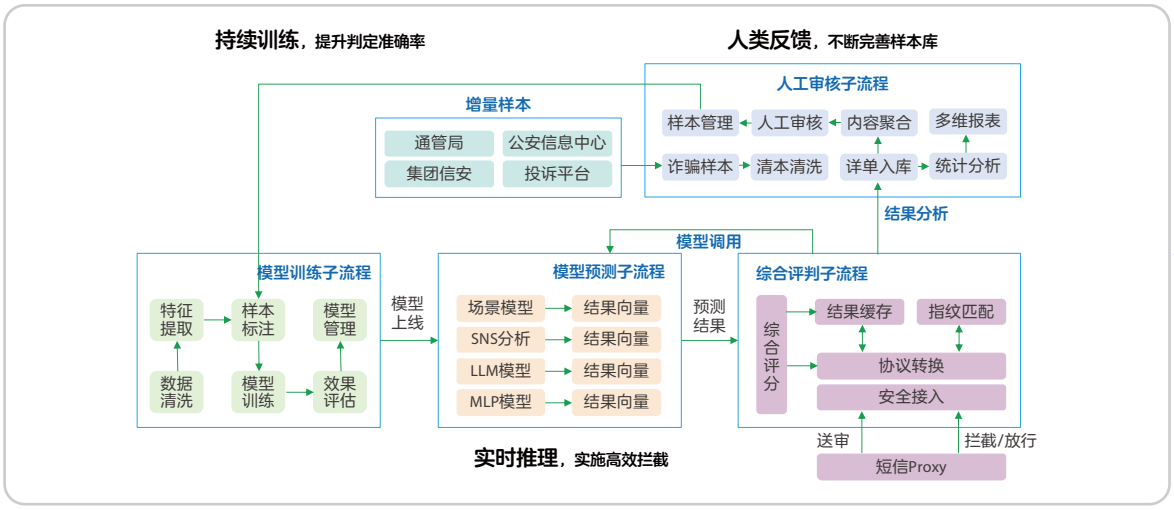
快速技术穿刺，攻关难点

项目初期，我们在AI大模型选用方面面临几大困难：

- 模型不确定：如何选择最合适的大模型，并确保合法合规；
- 语料及训练方案不确定：语料质量、数量、格式、提示词要求不了解，训练、推理方案从零开始；
- GPU和服务器成本高：中期推理性能低，大业务量下GPU数量和成本过高。

我们基于快速穿刺，敢于试错，及时调整方案逐个解决难点。模型选择方面，最初摸索阶段，从小于1亿参数规模，到3.4亿，再到70亿和130亿，尝试了包括国内和国外多种大模型，总计4种参数规模、6种国内外模型及自研模型、20多种组合，进行了大量穿刺比较。

语料和精调方面，获取一手高质量、合规语



◀ 图1 中兴通讯“智御”反诈大模型系统

料，尝试多种精调方案，最终设计出“专用提示词+样本微调”最佳方式，识别准确率和召回率效果均大幅提升。

针对GPU数量和成本过高问题，设计多层架构，将缓存加速作为前置，以小模型与大模型叠加结合的方案来解决，并通过推理加速达到最优。

最终基于模型评估效果和成本两项指标，选取综合最优方案完成选型，并通过法务合规评审。

通信与AI完美结合，双剑合璧

经过不断创新，中兴通讯成功发布业界第一个“智御”反诈大模型系统（见图1）。该系统无需策略配置，开箱即用，自动识别非法短信，大幅降低现场策略运维的复杂度和工作量的同时，提升了非法短信识别的查准率与召回率，实现垃圾/诈骗短信的识、防、管、打一体化治理。

该系统目前已在A、B两个运营商样板局点开展业界首个基于大模型的短信反诈治理试点，达成目标，并快速转商用。

A运营商成果：系统在省公司上线后，诈骗短信拦截率得到显著提升，境外垃圾短信发送量从日均50万~60万条断崖式下跌为日均2万~3万条，预测准确成功率和拦截准确率最高可达99%；与此同时，有效减少了涉诈案件数量，2023年8月，境外涉诈案件环比下降64%；局点上线后得到客户及该省反诈中心的高度认可。

B运营商成果：国内终端发起短信（MO）总量400万条/天，全部进入“智御”系统。日均拦截11万条左右垃圾和诈骗消息，拦截准确率从57.25%提升到93.60%；误拦截比例从42.75%降低到6.4%。

此外，智御反诈技术成果荣获工业和信息化部反诈专班《防范治理电信网络诈骗创新技术遴选应用》一等奖，并全面向全国推广。

未来演进与展望

反诈大模型是通信大模型应用的一个开端，“智御”系列大模型将在服务范围、媒体能力和行业应用等多个方向深入发展、演进和应用。

- 领域拓展和能力开放

实现能力复制和开放能力，进一步深拓反诈治理领域到5G新通信领域、IT领域和内容发布等领域应用。

- 媒体CV大模型

除了短信文本内容反诈，多媒体内容是一种上升很快的电信诈骗形式，为保障5G新通信时代媒体内容可信、安全、可靠，“智御”大模型未来必须支持多媒体内容的高效识别和反诈。

- 行业大模型

5G新通信行业客户有广泛的智能对话、行业知识服务和企业应用需求，“智御”大模型可以通过支持L0/L1/L2大模型，在5G消息平台等新通信网络侧平台集成和升级，来快速满足和实现5G行业通信AI能力需求，服务政企客户。ZTE中兴

那夜，一只藏羚羊路过我的帐篷

摘编自《C114通信网》

有了5G做支撑，以高清视频回传为代表的一系列数字化应用就成为可能。这里的藏羚羊和它们的守护者，再不会因与世隔绝而孤立无援。

睡在可可西里无人区的时候，人会对微小的声响格外敏感。比如此时，一只产仔路上的母藏羚羊正路过我的帐篷。天光里，它的侧影悠然又肃穆，仿佛是沿着自然划定的无形轨迹走去，直到和广远天地融为一体。

正是这样的寂静，让我接起电话时情不自禁压低了音量。电话里传来保护站同事兴奋的声音，藏羚羊们的身影已经实时显示在了观测平台上，或许在餐后散步，又或许在寻觅今夜的入眠地。随即他又感叹，能通过一方屏幕看它们享受宁静的时光，这样真好。

我本就澎湃的心被搅得更没了睡意。

身往“禁区”的理由

现场工作者们的帐篷依次扎在青海可可西里腹地，卓乃湖周边。可可西里是我国面积最大、海拔最高、野生动植物资源最丰富的自然保护区之一，卓乃湖更是被称为“藏羚羊大产房”，是保护藏羚羊的重要观测点和中继点。每年5至7月，数万头雌性藏羚羊都会经长途迁徙来到卓乃湖周边集中产仔，完成周而复始的生命循环。而我们尽管分别来自青海移动、中国铁塔、中兴通讯，却都是为了同一群生灵跋涉千里，带着数字技术前来加入“高原精灵”守护者队伍的。

今天是5月的最后一天，卓乃湖5G信号正式开通的日子。于我那感觉不太像交付了一项工作，更像成就了一个近乎神圣的使命。思及此，我索性披上外套，缓步走出帐篷远望。因着暮色的勾勒，这块土地显得格外静谧祥和。

实际上，夏天的可可西里迷人却极度危险，正所谓“野生动物的天堂”也是“人类的禁区”。临行前，当地有经验的前辈再三告诫我们，除却高寒、大风、缺氧、冻土这些极端状况外，这里还是一片泥泞不堪的沼泽地，“你们要来，就得做好各种意义上的准备。”很快，我们就在前往卓乃湖的路上有了亲身体会。风雪肆虐、路况艰险，解放车和应急车数次陷入泥潭，我们不得不用绞盘拉，拿铁锹挖。这对于因高原反应相继头疼、胸闷乃至呕吐的同事们而言，无疑是极漫长、艰难的旅程。仅仅是进入到预先规划好的位置，我们就耗费了一整天时间。

此外，这里苦寒难耐，风声猎猎，随时可能有野生动物出没。因此我们不仅要对湖区的覆盖

和容量做好精细预案，还要经得起大风、严寒、脆弱生态环境的重重考验。从前期规划设计到现场部署开通，整个过程前后经历了近两个月的时间。来到可可西里后，我们几乎每天都要经历两三场雪，有时风力达到8级，只能利用清晨没风的一小段时间抓紧调试。

许多次，我看到他们因持续的头疼、乏力在睡袋中辗转反侧，难以入眠，看到大家交替搬运设备时止不住微颤的手，也看到有人靠在施工区域旁将就一碗没泡开的面。5月31号上午，我们完成了微波全线调测和5G基站开通。终于，视频电话于茫茫无人区成功接通的那一刻，我们从相互对望的神情里知道，谁都没忘记来到可可西里的初心：

有了5G做支撑，以高清视频回传为代表的一系列数字化应用就成为可能。这里的藏羚羊和它们的守护者，再不会因与世隔绝而孤立无援。

这就是支撑我们远赴海拔4800米高原的原因。

“戍守”的故事

在可可西里，许多幼时以为自己会做牧羊人的孩子，都成为了“藏羚羊大产房”的守卫者。有人在雨季被困卡点，无法与救援队联络，白天的雨、夜晚的雪都成为最后的口粮；有人因常年巡山吃不上热饭而落下胃疾，去哪儿都得随身携带丹参滴丸；有人为保护藏羚羊幼崽驱赶过怒吼的棕熊；还有人指着自己的胸口说：“这里没有一块儿是健康的，但我放不下可可西里。”

我想他们不是不怕，只是选择了“与世隔绝”的坚守。正如一位保护站老队员在日记中写的：“再往前就没有信号了，踏入无人区意味着

故事的开头是：夜色淡去，天光渐明，碎石与小花被露水打湿，产仔路上的藏羚羊结伴而行，从山坳边缘向湖边奔腾而去。保护站的工作人员照常起早，用手机拍下救助站里小羊羔的照片，实时传送给千里之外的动物专家寻求建议。经幡随风微动，一只藏野驴慢悠悠地踱步经过，不远处，5G基站的轮廓正在消弭的晨雾中逐渐清晰。

没有3G、4G、Wi-Fi等等所有的信号，意味着失联的状态，不论泥泞沼泽、雨雪风霜，只有兄弟们同舟共济才能回归现代社会，记着15岁那年冬天第一次和前辈们一起深入无人区出来，看见青藏公路上的车灯简直有种重生的感觉……”

是的，那时的卓乃湖没有信号，踏入无人区就意味着失联的状态，所以守护自然生灵的长路只能用双脚、用孤独，甚至用生命丈量。对他们而言，巡山前轻轻对亲人说出的那句“走了”，是平淡的牵挂，也是生死未卜的告别。十余日里一通来之不易的卫星电话，对着无人区的夜空与星星哼出的摇篮曲，已是联结守护者们与远方家人的唯一纽带。

所以我们要来加入他们，我对自己说。夜色渐浓，可是天很快就会亮起来。正如跋山涉水的“戍守”，也会被注入“数守”的力量。

“数守”的意义

现在，藏羚羊的身影能通过视频实时回传到观测平台上，一旦发现猛兽袭击、异常天气等紧急状况，工作人员可以及时联络，以便实施救助或请求专业增援。通信设备有了信号，守护者们也不会陷入孤立无援的困境。他们会从以往生死不知的“失联”状态走向“自由连接”，在生命安全得到保障的同时，能和身在都市的人一样，在闲暇时看看视频、线上聊天、网上购物……不仅如此，他们还能去到以往难以巡检到的“盲区”，当地生物保护工作的广泛性和灵活度因此得到提升。父母、爱人和孩子也不必再苦苦等候那通卫星电话，而是可以尽情享受视频通话的时光，分享千里之外的趣闻轶事，一解思念之苦。

这只是卓乃湖环境保护、生态监测与科学考察模式走向高效化、智慧化的起点。不久的将来，5G远程视频巡检、无人机监测、土壤水质观测分析等技术也有望相继覆盖至此，可可西里的生态保护与管理还有无限可能等待发掘。

我想，到了那时，五湖四海的人们都有机会一睹可可西里的面容，听到卓乃湖的风雪，望见这片神秘土地上新生小藏羚羊的湿润的眼睛。

也许有一天，这些广袤天地间的生灵会替我们讲一个故事。故事的开头是：夜色淡去，天光渐明，碎石与小花被露水打湿，产仔路上的藏羚羊结伴而行，从山坳边缘向湖边奔腾而去。保护站的工作人员照常起早，用手机拍下救助站里小羊羔的照片，实时传送给千里之外的动物专家寻求建议。经幡随风微动，一只藏野驴慢悠悠地踱步经过，不远处，5G基站的轮廓正在消弭的晨雾中逐渐清晰。

它是我们一同守护这群生灵、这湾湖畔、这块土地的见证，是我们对可可西里自然保护区生态文明建设的答卷，也是我们的责任与荣幸所在。ZTE中兴

ZTE中兴

5G领衔 别出新彩

中兴云电脑双风系列

纤薄至简 | 缤纷配色 | 大美无界

ZTE中兴

让沟通与信任无处不在