责任编辑：严粒粒 版式：陈仰东 联系电话：0571-85312335 邮箱：youfenglai2014@sina.com

边缘计算，让智能触手可及

机器人“爱宝”。 潮新闻记者 于诗奇 摄

■ 本报记者 金春华

最近，“养龙虾”（科技圈对部署和使用开源AI智能体框架OpenClaw的戏称）火爆出圈。仅凭Mac mini、手机等普通终端，就能实现原本需要大型数据中心支撑的AI服务，备受网友追捧。

这一切的实现，离不开一项关键技术——边缘计算。

而今，边缘计算走到了技术与产业发展的“C位”。去年年底，工信部等八部门印发《“人工智能+制造”专项行动实施意见》，提出“强化人工智能算力供给，推动边缘计算服务器部署”。日前，《浙江省“十五五”数字经济和数字基础设施规划（征求意见稿）》公布，其中提到“按需建设边缘算力节点”。杭州、宁波、温州、金华等地在推动人工智能创新发展中，也在落地边缘计算相关布局。

什么是边缘计算？它为什么这么强？对我们的生活又带来了怎样的影响？

走向“去中心化”

“边缘计算是一种分布式计算，和我们常说的中心化部署的云计算相对。”浙江大学人工智能专家张微说，边缘计算的“边缘”，可以是一部智能手机、一台工厂传感器网关、一辆自动驾驶汽车或是一个街角的基站，能让大模型在离用户最近的终端上稳定运行。

前不久，特斯拉受监督版全自动驾驶系统（FSD）获准在荷兰公共道路上使用。这是该系统首次在欧盟主要成员国落地。

更让车迷们津津乐道的是，作为车规级边缘计算的标杆性应用，FSD系统的全链路延迟据悉可控制在100毫秒以内。若采用云计算，常规场景延迟普遍达500毫秒，车辆在高速上会多冲出去10多米，直接关系到是否会造成严重车祸。

而这正是边缘计算的厉害之处。它把计算任务、算力能力从云端“下沉”到离数据产生源头、用户需求最近的“边缘侧”，就近完成数据处理与决策，可以缓解云端算力压力、降低传输带宽要求和能耗成本，大大降低延迟时间。

张微打了个比方，“海洋智者”章鱼有超5亿个神经元，其中40%集中在中央大脑，负责统筹决策与高级认知，60%分布在8条触手上，可脱离大脑独立完成抓取、辨别材质等复杂动作。“AI要真正落地千行百业、走进千家万户，就得向章鱼学习，具备这种自主响应、快速决策的能力。”

但AI不像章鱼天生具备这样的能力，而是要经历从“中心化”到“去中心化”的迭代。

2022年，依托万余枚高端GPU（图形处理器）芯片的ChatGPT横空出世，标志着大模型时代全面到来，其核心依托的正是云端中心化算力——把所有数据、计算任务都上传到云数据中心统一处理，以集中调度海量算力。

当时，很多专家以“大力出奇迹”来形容中心化布局对AI的重要性。其后问世的Gemini、Claude、DeepSeek等主流语言大模型，几乎都用此模式。

但随着大模型加速落地推广，中心化模式的弊端日益凸显：延迟高、成本高，高度依赖网络、断网便无法使用，数据隐私风险……

近段时间，相关吐槽就明显多了：“高峰期常要排队，让人抓狂”“问个小问题就要消耗大量token，烧不起”……前不久，话题“千万不要跟AI说谢谢”还冲上热搜榜：“它回你一句‘不客气’，大概要消耗0.0003度电，可以让你的手机多待机约10小时。”

不少业内人士甚至认为，AI靠云计算发展已进入瓶颈期，亟需边缘计算来“救场”。

“中心化与去中心化相互迭代，是一个推动数字世界发展的深层规律。”杭州城市大脑公司总经理申永生说。1998年，边缘计算的技术雏形——内容分发网络（CDN）诞生。CDN通过将静态内容缓存至靠近用户的边缘节点，有效缓解了网络拥堵，大幅提升响应速度。而在此之前，互联网的算力与服务高度集中于中心化服务器，当时的网民调侃万维网（World Wide Web）是“World Wide Wait（全球等待）”。

正如CDN的出现推动了全球互联网规模化爆发，边缘计算也正承担着推动人工智能规模落地的使命。

这在具身智能、智慧交通、智能制造等领域尤为急切。

“50分26秒！”近日，在2026北京亦庄人形机器人半程马拉松赛事中，机器人“闪电”以这一打破人类男子半马纪录的成绩夺冠。

但比起机器人跑得快，业内人士更惊叹于边缘智能发展的迅速。跑步中，脚掌触地瞬间，机器人必须在极短时间内完成“感知—计算—响应”的闭环，哪怕延迟一眨眼时间（约100毫秒），都有可能摔倒。目前，搭载边缘计算的主流具身智能机器人，已能将响应延迟控制在50毫秒以内，甚至更低。但主流云计算的延迟仍在200毫秒左右，是前者的四倍，慢了一大截。

边缘计算还是数据安全的重要屏障。

4月初，字节跳动发布的语音大模型Seeduplex获得网友“AI真正有了人味”的肯定评价。成功背后，“端云协同”的部署架构就很关键。

“类似的人工智能拟人化互动服务，特别是情感陪护机器人等个性化产品，会涉及大量个人信息，最安全的办法就是留在本地设备、交给边缘计算。即使为了优化服务体验必须传到云端，也应先在本地完成数据清洗和脱敏处置。”同济大学电信学院研究员吴迪表示。

在2026中关村论坛年会期间，人们从北京中关村国际创新中心大厅的飞书“龙虾”宣传板旁经过。 新华社发

模型“瘦身”，硬件“壮骨”

4月8日，国内边缘计算重要交流平台——边缘计算社区，发布“2026中国边缘计算企业20强”榜单。入选企业涉及的领域，包括AI芯片全栈自研、可信安全计算，以及工业智能体落地、企业级云边协同等，折射了一个快速发展并加速差异化的新赛道。

专家指出，边缘计算与人工智能相融合，正在催生更多新的技术，掀起“边缘智能革命”。

目前，大模型压缩领域已形成“知识蒸馏”“模型量化”“剪枝”“架构搜索”等核心技术。其中，前两种工程实践尤为成熟，“炙手可热”。记者查询了国家知识产权局专利公布公告系统发现，自去年10月10日至今年4月17日，短短半年多时间，有关“知识蒸馏”的发明授权就有1500余项，而同期有关“智能驾驶”的发明授权不到500项。

简而言之，“知识蒸馏”能让大模型化身为“导师”，即通过特定技术，将训练成熟、参数量庞大的“教师模型”所积累的知识，高效迁移到结构精简、参数量少的“学生模型”中，让小模型也具备接近大模型的性能。“模型量化”则在尽可能保留模型推理精度前提下，为大模型“精准瘦身”，以解决存储、计算与功耗等难题，目前行业基本可以将模型“体积”压缩到原来的四分之一甚至八分之一。

这意味着，原本用售价上万元的专业级显卡才能勉强跑起来的模型，如今在两三千元的智能手机上就能顺利运行。

边缘计算如此厉害，是否会取代云计算？

吴迪给出的答案是否定的：“两者会分层协同，如云端负责训练与复杂推理、边缘侧承担实时推理与数据预处理、终端设备执行轻量感知与交互，进一步支撑AI规模化落地。”吴迪介绍，面对复杂多元的现实需求，业界、学界又探索了云边协同、云边端协同、多接入边缘计算（MEC）、联邦学习（Federated Learning）等更精细的架构、模式。

4月15日，英特尔全球高管一行到访位于杭州上城区西子智慧产业园内的“西子智平方未来小馆”。机器人“爱宝”取杯、制作，为客人端上咖啡，效率已接近人工。这一场景，正是英特尔最新的云边端协同分布式具身智能解决方案在现实中的应用。

据悉，该方案将“爱宝”的计算功能放在园区的边缘服务器中，机器人只管干活，而不像以前那样“一台机器人装一个大脑”。该园区可以用同一套算力同时调度多台机器人，成本更低，效率更高。

模型“瘦身”之际，跑模型的硬件也在不断升级。例如，从“临时凑数”的CPU（中央处理器），到“半路跨界”的GPU，再到“为模型而生”的NPU（神经网络处理器）、TPU（张量处理器）等，关键硬件芯片的专业化程度越来越高。

“边缘计算芯片受功耗、延迟、成本等限制很大，比如CPU的内存，2G、4G是很常见的，但NPU中几十MB就算豪华了，完全是两个数量级，业界形容是‘螺蛳壳里做道场’，要精准设计、做到极致。”多位从业者说。这一新赛道竞争也日益激烈。

今年2月，总部位于宁波的爱芯元智半导体股份有限公司在港交所敲钟上市。这一在港股上市的首家边缘计算AI芯片企业，靠着自研的“爱芯通元”混合精度NPU等架构，推出了多款AI芯片，覆盖视觉智能终端、智能汽车、边缘AI推理三大核心赛道，去年营收达5.62亿元，同比增长18.8%。

“iPhone时刻”或将来临

自动驾驶、工业互联网、具身智能……眼下，边缘智能在多个场景加速落地。有分析人士将今年称为边缘计算“物理化”元年。

全世界已为此期盼了很久。早在2023年，欧盟委员会批准一项12亿欧元的国家援助计划，以支持欧洲先进云计算和边缘计算技术的研究、开发和首次工业部署。北美地区因在全球互联网布局上的传统优势，长期占市场主导地位，其中仅微软、亚马逊等公司就分别在全球运营着约200个、180个边缘站点，苹果、英伟达等公司也在推出边缘AI芯片等硬件。

在国内，工信部等六部门在2023年10月联合印发《算力基础设施高质量发展行动计划》，提出“加速边缘算力协同部署”。今年3月，工信部等九部门联合印发的《推动物联网产业创新发展行动方案（2026—2028年）》更明确要求，“推动人工智能大模型在终端和边缘的轻量化部署”。

“边缘计算的部署，影响到电力、网络等基础设施布局。”浙江移动技术专家陈韩玮介绍，浙江作为数字经济大省，也早早开始了相关部署。2021年，浙江移动就参与了全省11个设区市和49个区县的“边缘云”等网络基础设施建设，将算力下沉到需求所在地。

近日，2026年世界移动通信大会（MWC 2026）在西班牙巴塞罗那举行。多家厂商展出了智能手机、人形机器人、智能物联网等与边缘智能相关的技术、硬件和应用。

业内专家表示，就像iPhone开启人机交互革命一样，边缘智能的“iPhone时刻”或将来临。

这将开启一个庞大的新市场。根据权威市场研究机构Omdia预测，到2027年，全球企业边缘服务市场规模将达到2450亿美元。

在浙江，边缘智能正在加速融入生产、生活——宇树科技的机器人、云深处的机器狗等具身智能，乐奇AI眼镜等智能穿戴设备，公牛集团等的“黑灯工厂”，以“安诊儿”等为代表的智慧医疗，飞行在山区海岛的各类无人机……它们让人们看到了边缘智能落地的无限可能。

“应用需求是技术进步的重要推动力。丰富的场景，将进一步培育市场需求，形成良好的供需互动，从而推动整个边缘智能产业健康成长。”申永生说。

在杭州城市大脑有限公司，记者看到了多款智能玩具样品。“智能玩具是进入边缘智能一个相对成熟的切入口。”申永生解释，00后、10后等群体从小与“天猫精灵”等语音助手相伴，对智能玩具接受度高、需求也更大。“我们去调研时，有不少家长说很想有电影《长江七号》中的‘七仔’那类智能玩偶，陪伴孩子。这一市场空间很大。把它做好了，也可以打开其他更多人机互动的场景。”

边缘智能，正在将种种梦想变为现实。

链接

相关未来应用场景一览

从工业生产到日常生活，从城市运行到数字娱乐，边缘计算与人工智能相结合，正在解锁一个又一个全新应用场景。未来，边缘智能将渗透到生活、工业、医疗等各领域。

智能交通：边缘计算节点会部署在路侧单元、5G基站等载体中，实时采集车辆速度、位置及路况信息，短时间内完成分析决策，动态调整信号灯配时，甚至向车辆推送碰撞预警，让出行更安全高效。

智能驾驶：边缘计算与车联网深度融合，能让无人配送车、无人出租车、无人货运卡车规模化落地，靠本地实时计算规避风险，无需完全依赖云端并支撑智能汽车、电动飞行载具等多类交通工具协同运行，解锁全新出行方式。

智慧工厂：边缘智能就是“全能大管家”，从接订单、排生产计划、产品质检，到设备故障预判、节能降耗、对接物流，全流程本地自动完成，大大降低生产成本，我们买到的商品也会更实惠、品控更稳定。

智慧医疗：远程会诊、手术示教将实现“零卡顿”，基层医院的医学影像可在边缘节点快速处理，让诊断响应时间从几十分钟缩短至几分钟。在偏远地区，边缘智能还能辅助基层医生完成诊断，让优质医疗资源触手可及，同时保障医疗数据的隐私安全。

智慧城市：城市会变成一个“会思考的智慧体”，AI能实时感知城市里的每一个角落，哪里堵车了、哪里路灯坏了、哪里餐馆后厨卫生不达标、哪里老人需要帮助，都能及时发现并解决，让城市管理从“被动应对”转向“主动预判”，让城市生活更方便、更安全、更舒心。

智能家居：你会有一个“全能管家”，能读懂你的生活习惯，不用反复指令，就能提前打理好家务：下班到家前自动开空调、烧热水，早上按你的日程叫醒你、播报天气路况，还能帮你预约挂号、缴水电费、对接物业快递。所有数据都在你家本地处理，不用传到云端，响应更快、隐私更安全。

智慧农业：田间的传感器、无人机将采集土壤湿度、光照、病虫害等数据，边缘节点就近分析处理，精准控制灌溉、施肥设备，实现“按需供给”。它还能实时识别病虫害，提前发出预警，减少农药使用，让农业生产更高效、更绿色。

智能零售：便利店、超市的边缘设备可实时分析客流数据、商品库存，精准推送促销信息，还能通过视觉识别实现无人结算，减少排队等待时间。同时，边缘计算可实现商品库存的实时监控，自动提醒补货，让零售运营更高效，提升消费者购物体验。

智能电网：在智能电网中，边缘计算节点可实时监测光伏、风电等新能源设备的运行状态，优化电力调度，平衡电网负荷，避免能源浪费。它还能实时预警电网故障，快速定位问题并处理，保障电力供应的稳定可靠，助力“双碳”目标实现。

智能娱乐：元宇宙的实时渲染、多用户协同对低延迟要求极高，边缘计算节点可在用户附近完成高清渲染，将交互延迟控制在50毫秒内，支撑虚拟办公、VR演唱会等场景。无论是多用户在虚拟空间协同工作，还是万人同步观看VR演出，都能实现无缝流畅的交互体验。

个人级边缘计算：个人电脑、手机将成为新型边缘节点，具备强大的本地AI推理能力。无论是离线处理高清视频、进行AI绘图，还是实现隐私保护下的语音助手交互，都能快速响应，不再完全依赖云端，让个人计算更高效、更安全。

（本报记者 金春华 整理）

杭州城市大脑有限公司的智能玩具样品。智能玩具是边缘智能的重要落地场景之一。 受访者供图