边缘算力蓝皮书

(2024)

边缘赋智 应用无垠

前　言

当前全球数字化浪潮蓬勃兴起，边缘算力通过就近提供计算、网络、智能等关键能力，加速赋能经济转型升级，已逐步成为计算体系的新方向、信息领域的新业态、产业转型的新平台，是垂直行业实现数字化、智能化的重要基础设施。

我国高度重视边缘算力技术创新、产业发展和应用探索。2023 年，工信部等六部委联合发布《算力基础设施高质量发展行动计划》 ，明确提出促进边缘算力协同部署，加快边缘算力建设，支撑工业制造、金融交易、智能电网、云游戏等低时延业务应用，推动“云边端”算力泛在分布、协同发展。

经过全社会的共同努力，我国边缘算力技术体系初步构建，典型应用实践不断涌现，正在从单点、局部应用向多点全面应用演进。目前，边缘算力已经成为国内外经济社会各领域数字化转型和智能化升级的切入点和优先方向，受到了产学研用高度关注。

本蓝皮书分析了边缘算力整体发展态势，具体阐述了边缘算力概念及特征，梳理了边缘算力技术体系，总结了边缘算力典型应用场景，并提出边缘算力未来展望，希望为推进边缘算力技术产业、基础设施建设及应用发展提供参考。

本蓝皮书在编写过程中得到了工业互联网产业联盟、边缘计算产业联盟、中国通信标准化协会边缘计算产业发展及技术标准推进委员会等大力支持。

边缘算力正处于快速发展阶段，本蓝皮书仍有诸多不足，恳请各界批评指正。

参编单位：

中国铁塔股份有限公司

中国信息通信研究院

中国联合网络通信集团有限公司

中兴通讯股份有限公司

中国移动通信有限公司

中国电信股份有限公司

北京研华兴业电子科技有限公司

北京交通大学

东南大学

中国科学院计算技术研究所

联通智网科技股份有限公司

上海天数智芯半导体有限公司

北京城建智控科技股份有限公司

上海道客网络科技有限公司

广州市政务服务和数据管理局

暨南大学

工业互联网产业联盟公众号

编写组成员（排名不分先后）：

郭宇辉、闫亚旗、潘三明、魏华、张民贵、冉沛、董玉池、张阔、刘文睿、张文龙、王哲、王涵、朱瑾瑜、刘洋、韩玲、张文博、周光涛、辛亮、徐成杰、许恒斌、尤鸿、邱述洪、梁文昭、束裕、赵孝武、王旭辉、魏彬、黄震宁、马程然、彭涛、陈菲雨、袁方正、仲其伟、曹亚平、姜丽丽、孙颖、李子龙、马萍、张维庭、廖培希、李博睿、王帅、张宗帅、刘诗瑶、彭伟、程伟、姚涛、张辉、张利宽、任超、张红兵、侯玲玉、杨红军、刘国栋、龙赛琴、王泽平

目录

1 边缘算力整体发展态势

1.1 算力发展持续加速

1.1.1 算力是数字经济发展的重要支撑

近年来，数字经济发展速度之快、辐射范围之广、影响程度之深前所未有，正在成为重组全球要素资源、重塑全球经济结构、改变全球竞争格局的关键力量。党的十八大以来，党中央高度重视发展数字经济，将其上升为国家战略，数字经济的新引擎作用不断凸显。2022 年 1 月，习近平总书记发表题为《不断做强做优做大我国数字经济》的署名文章，进一步深刻阐明发展数字经济意义重大，是把握新一轮科技革命和产业变革新机遇的战略选择[1]。

算力作为数字经济时代的核心生产力，为经济增长提供了智能升级、融合创新的新动力，一方面，算力正加速向政务、工业、交通、医疗等各行业各领域渗透，成为传统产业智能化改造和数字化转型的重要支点。另一方面，以 AIGC（Artificial Intelligence Generated Content）为代表的人工智能应用、大模型训练等新需求、新业务的崛起,推动算力规模快速增长、计算技术多元创新、产业格局加速重构。算力作为数字经济核心产业的重要底座支撑，对上下游软硬件产业的拉动作用日渐凸显，2022 年全国电子信息制造业实现营业收入 15.4 万亿元，同比增长 5.5%。软件业收入跃上十万亿元台阶，达 10.81 万亿元，同比增长 11.2%，保持较快增长[2]。

1.1.2 边缘算力价值凸显

根据 Machina Research 研究报告显示，2025 年全球物联网连接数将增长至 270 亿个，联网设备的指数式增长造成网络传输能力及中心云处理能力捉襟见肘。同时，接入网络的终端每年产生数据达 847 ZB，增量数据呈现分散性、碎片化的特点，超过 50%的数据需要在网络边缘侧分析、处理与存储[3]。而以大型数据中心等为主的集中式算力虽然有强大的数据处理能力，但是在面对海量数据以及有限网络带宽带来的挑战时，无法实现全面计算覆盖。此外，随着 AR/VR、云游戏、智能控制等新业务蓬勃发展，产生了大量低时延响应需求，也需要边缘算力完成数据的就近处理和分析，以满足用户的实时性要求。因此，边缘算力已成为支撑 IT、CT、OT 创新演进的关键基础设施。

（a）IT 领域：当前，计算架构的革新和硬件制程的进步推动算力部署模式

从集中式向互联协同的范式转变。多异构芯片集成、Chiplet 技术以及新型封装技术等为边缘算力设备提供更强大的计算能力和灵活的定制化选项。此外，数据隐私和安全性受到各方高度关注，边缘算力可以实现数据本地处理分析，以减少数据传输风险，提高安全性。

（b）CT 领域：新型网络架构将采用服务化设计，实时感知用户需求，支持资源可按需调用，为不同垂直行业提供快速响应和灵活部署。而实时感知能力需要边缘算力支撑响应，在 5G 等网络架构中明确定义了边缘算力作为重要组成部分，以实现网络智能化演进。

（c）OT 领域：依托算力总量的持续增长和算力类型的不断丰富，工业企业不断加快数字化转型步伐，其中，边缘算力已成为工业数字化转型的关键基础设施之一。基于边缘算力部署的工业应用种类繁多且贯穿于各生产环节，目前已在设计模拟、订单排产、生产制造、运营维护、安环管理、质量追溯、物流库存等重要场景形成一批典型应用。例如，工业生产质检系统基于部署在产线附近的边缘算力节点，实时采集产品图像数据，并利用深度学习模型进行缺陷识别，能够解决传统机器视觉方案中数据传输延迟高、云端处理算力不足的问题，实现毫秒级缺陷检测和实时反馈控制，大幅提升产品良率，降低人工质检成本。

1.2 边缘算力政策逐步完善

我国高度重视边缘算力发展，出台了一系列政策措施（参见表 1.1）[4-6]，推动边缘算力技术创新、应用推广和产业发展。同时，各级地方政府积极响应，支持边缘算力在工业互联网、车联网等垂直领域的应用试点。

表 1.1 我国边缘算力相关政策

序号	发布时间	部门/省	政策名称	具体内容
1	2021.7	工信部	《新型数据中心发展三年行动计划（2021-2023 年）》	积极构建城市内的边缘算力供给体系，支撑边缘数据的计算、存储和转发，满足极低时延的新型业务应用需求
2	2021.11	工信部	《“十四五”信息通信行业发展规划》	到 2025 年实现数据与算力设施服务能力显著增强的目标。形成数网协同、数云协同、云边协同、绿色智能的多层次算力设施体系。

3	2022.1	国务院	《关于印发“十四五”数字经济发展规划的通知》	加快实施“东数西算”工程，推进云网协同发展，提升数据中心跨网络、跨地域数据交互能力，加强面向特定场景的边缘计算能力，强化算力统筹和智能调度
4	2023.2	国务院	《数字中国建设整体布局规划》	推进数字技术与经济、政治、文化、社会、生态文明建设“五位一体”深度融合，强调系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局
5	2023.10	工信部、网信办、教育部、卫健委、中国人民银行、国资委	《算力基础设施高质量发展行动计划》	结合算力基础设施产业现状和发展趋势，明确提出促进边缘算力协同部署，加快边缘算力建设，支撑工业制造、金融交易、智能电网、云游戏等低时延业务应用，推动“云边端”算力泛在分布、协同发展
6	2024.5	网信办、市场监管总局、工信部	《信息化标准建设行动计划（2024-2027年）》	建设“算、存、运”一体化算力基础设施标准体系，面向融合共生的技术发展趋势，推进云计算、边缘计算、高性能计算等异构算力中心的共性标准研究。开展算力接入、调度、服务等相关标准研制。开展云网协同标准研制，促进云间互联互通。
7	2022.11	北京	《2023年北京数字经济促进条例》	强化算力统筹、智能调度和多样化供给，提升面向特定场景的边缘计算能力，促进数据、算力、算法和开发平台一体化的生态融合发展
8	2023.1	江西	《江西省未来产业发展中长期规划（2023-2035年）》	加快提升技术创新能力，加强边缘计算、人工智能等新兴技术领域研究。统筹建设协同集约的算力基础设施。科学布局一批算力中心、边缘计算节点、区块链节点等，建设智能高效的融合基础设施，支撑智能发展的行业赋能能力

9	2023.4	河南	《2023 年河南省大数据产业发展工作方案》	打造一批新型数据中心和边缘数据中心，探索形成一批大数据产业技术规范和标准，灵活部署边缘计算中心，建设边缘计算城市节点，构建城市内边缘算力供给体系
10	2023.9	上海	关于印发《上海市进一步推进新型基础设施建设行动方案（2023-2026 年）》的通知	加快建成支撑人工智能大模型和区块链创新应用的高性能算力和高质量数据基础设施。建成多元供给、云边协同、随需调度、高效绿色的城市高性能算力网络体系。

1.3 边缘算力标准体系初步构建

目前，边缘算力标准化工作主要围绕四个方面开展：一是边缘算力关键设备及组件标准；二是边缘算力平台标准；三是基于边缘算力构建的算力网络、分布式云等相关标准；四是边缘算力应用标准。

中国通信标准化协会 CCSA TC13 针对边缘算力关键设备及组件开展标准研制工作，完成边缘边缘网关、边缘控制器等技术标准近 20 项[7]，构建了以边缘算力设备功能架构、参考架构、技术要求和测试方法为核心的边缘算力关键设备及组件标准体系，有效推动了边缘算力底层基础设施的标准化进程。

中国通信标准化协会 CCSA TC5 及 TC13 围绕边缘算力平台开展标准研制工作，完成移动通信网络、工业互联网等领域边缘云等关键技术标准的制定，覆盖资源管理、设备管理、应用管理、运维管理等功能规范，为边缘算力平台的研发方向提供技术指导。

中国通信标准化协会 CCSA TC1 TC3、TC7 及 TC8 围绕基于边缘算力构建的算力网络和分布式云开展标准研制工作，完成算力溯源、算力并网、算力度量、算力可信、算力安全、分布式云安全、分布式云运维管理等技术标准制定，为构建安全、可信、高效的边缘算力网络和分布式云生态体系奠定基础。

中国通信标准化协会 CCSA TC1 及 TC13 针对边缘算力应用开展标准研制工作，已涵盖公共通信、智能计算中心及工业数据中心等多个领域，为各应用的算力能力要求提供重要参考。

1.4 边缘算力已成为产业各方布局重点方向

随着 5G 规模化应用，5G+边缘算力产业进入快速发展期，基础设施运营商充

分利用自身网络资源，将算力和网络接入端扩展成边缘算力节点，形成边缘算网一体化融合的独特优势。中国移动构建广泛边缘算力节点，并聚焦边缘应用生态构建，基于边缘算网融合能力提供泛在算力一体服务；中国电信自主研发边缘云，赋能低延时、大带宽及数据安全应用需求；中国联通融合 5G、通信、大数据能力，构建 5G 边缘算力体系，面向工业、视频、车联网等领域推出定制化边缘服务；中国铁塔依托遍布全国的站址机房资源优势，积极推进“通信塔”变“数字塔”、“通信机房”变“数据机房”，利用铁塔站址按需网格化部署边缘算力，从基础设施共享、边缘智算服务、算力共享服务等三个层面助力边缘算力建设与发展，推动实现算力的泛在化、随需化、普惠化；阿里云扩展中心云至边缘，推出 OpenYurt 开源框架及边缘节点服务，实现云原生应用边缘部署。

基础设施服务商持续加大边缘算力产业布局。浪潮、新华三均推出边缘超融合一体机，实现边缘算力敏捷部署、数据就近接入、本地智能处理，同时集成 AI 智能、物联网平台、数据平台等多项能力；中兴、华为积极探索边缘算力技术创新，提供基于边缘算力的行业解决方案，已在电力、制造、农业、医疗等行业广泛应用。

工业企业基于行业经验积累，将工业软件部署至边缘算力节点，提高生产运营效率。三一重工、海尔、商飞等在边缘算力节点实现远程 I/O 毫秒级实时控制及逻辑控制器集中虚拟化部署，动态调配网络、控制、算力资源，降低设备固定资产投入，成功打造边缘算力应用于制造业的“样板间”。西门子、罗克韦尔等工业自动化企业推出基于边缘算力的强化学习服务，通过强化学习智能算法创建大型设备预测性维护模型，实现工业生产高效运维管理。

2 边缘算力概念及关键特征

2.1 边缘算力概念

传统上，算力的部署主要依赖于数据中心等基础设施，这些设施通常位于特定的地理位置，如数据中心园区或企业自建的机房内。然而，随着云计算、边缘计算等技术和业务的快速发展，算力部署方式正在发生深刻变革。集中式算力通过虚拟化技术将计算资源、存储资源和网络资源封装成一个独立的虚拟环境，为用户提供按需使用、弹性扩展的计算服务；边缘算力则将计算资源部署在更接近数据源和用户的地方，以减少数据传输延迟和提高响应速度。

边缘算力是指在终端、本地或离用户较近的位置部署的计算能力。边缘算力利用多种通信网络技术（如 5G、WiFi、光纤通信等）连接分布式算力节点，通过虚拟化和多层次的算力资源协同，实现对资源的灵活、按需和实时调度，提高边缘算力及网络的利用率，增强业务的服务质量和安全性，降低服务延迟和提高数据处理效率，提升用户体验。

边缘算力不仅仅是新设施、新架构以及新模式，而是三维一体的新生态，通过在网络边缘侧汇聚网络、计算、存储、应用、智能等五类资源及能力，提高服务性能（“提速”）、开放控制能力（“敏捷”），提升用户体验，从而激发类似于移动互联网生态的新模式和新应用。

边缘算力的“边缘”可以从多个角度去理解。**一是地理位置的边缘**，指的是数据源和用户设备所在的网络边缘位置。这些位置可能包括智能手机、智能家居设备、工业传感器、自动驾驶汽车等物联网（IoT）设备的所在地。这些设备生成大量数据，并在其所在的网络边缘进行处理，以减少数据传输到中央云端的延迟和带宽消耗。**二是计算资源的边缘**，指的是在用户侧部署的计算设备和资源。这些设备包括边缘服务器、网关、路由器等，它们具有一定的计算能力和存储能力，可以在本地处理数据并做出决策。这种分布式计算模式使得数据处理更加高效和灵活。**三是数据流的边缘**，指的是数据在网络中传输的路径上的关键节点。这些节点可以对数据流进行预处理、分析和决策，以减少需要传输到中央云端的数据量，通过在网络边缘进行数据过滤和压缩，可以显著提高数据传输的效率和安全性。

边缘算力在靠近数据源或用户的地方提供计算、存储等基础设施，可以满足业务的低时延需求，有效缓解网络带宽压力。单个算力节点的算力资源有限，但可以借助运营商现有网络基础，将分布式算力节点网联起来，打造边缘算力集群，将计算任务调度至最优的边缘算力节点进行处理，促使算力部署从中央走向边缘，进而促进基础网络与计算深度融合。

因此，边缘算力部署的位置可以从端到云端的任何网络位置。比如终端、边缘控制器、边缘网关、边缘云、基站、核心网用户面、有线的汇聚层（例如城域网）或接入层（例如 FTTH）等。

2.2 边缘算力特征

相比传统集中式算力部署，边缘算力具有如下特征：

资源异构：边缘算力类型多样，包括 CPU（ARM 或 X86）、GPU、NPU、FPGA 等。因此，边缘算力需要通过资源抽象对算力进行统一组织和管理，并封装成用户所需要的能力，融合不同协议的各类设备进行通信交互，并提供自适应的数据格式与内容适配。

网络异构：边缘侧网络制式多元，同时面临不同运营商的网络平台问题，因此边缘算力网络需要通过技术创新打造高可靠、低时延、高速率的网络接入和算网协同调度能力，加快用户接入的进程，使得用户可随时、随地、随需地通过无所不在的网络接入无处不在的算力。

泛在分布：边缘算力的显著特征在于用户和计算节点在地理上的泛在分布，由此产生的用户数据也呈现出地理上的分散性。算力管控调度等技术实现分布式算力节点协同联动提供统一的分析处理能力。

低时延：由于数据处理在数据产生的源头附近进行，因此可以显著降低数据传输的延迟，提高应用的实时性，这对于需要快速响应的应用场景尤为重要，如自动驾驶、实时游戏等。

低成本：边缘算力可以减少对云端计算资源的依赖，降低数据传输的网络带宽成本。同时，由于边缘设备通常具有较低的功耗和成本，因此可以进一步降低整体运营成本。

高隐私：边缘算力可以实现数据本地处理，减少数据在传输过程中的泄露风险，从而增强数据的隐私保护。

3　边缘算力关键技术

边缘算力的技术体系架构如图 3.1 所示，主要包括：边缘算力基础设施、边缘算力网络、边缘智能、边缘算力安全等四方面。其中，边缘算力基础设施聚焦于计算、存储、网络等物理硬件资源及其虚拟化，边缘算力网络关注分布式算力资源的感知、度量、并网、调度、管控等，边缘智能涉及系统部署、数据处理、模型优化、边缘训练、边缘推理等关键问题，边缘算力安全则贯穿始终提供对从基础设施到上层服务的全面安全保障[8]。

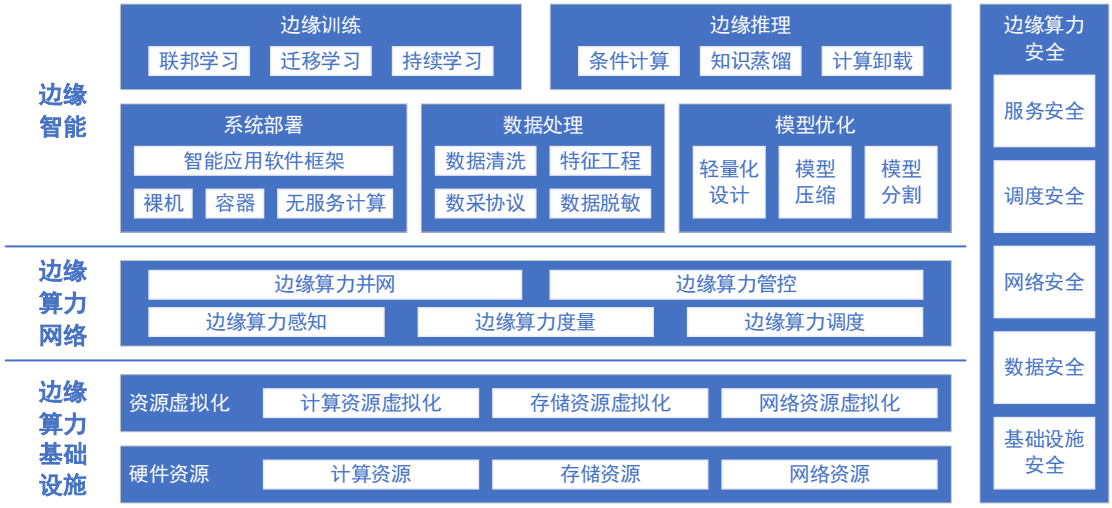

图 3.1 边缘算力技术体系架构图

3.1 边缘算力基础设施

边缘算力基础设施由硬件资源和资源虚拟化组成，前者提供边缘算力所需的计算、存储、网络等基础硬件资源，后者则通过虚拟化技术将各类异构的基础硬件资源抽象为逻辑资源，便于统一管理、调度和使用。

3.1.1 硬件资源层

边缘算力硬件资源层主要涵盖计算、存储、网络等多种基础设施资源，共同构筑了边缘算力的物理承载底座，其性能与效率直接决定着边缘算力的整体服务能力。

3.1.1.1 计算资源

计算资源主要是指 CPU、GPU、ASIC、FPGA、DSP 等各类处理器芯片及其组合所构建的加速卡。其中，X86、ARM 和 RISC-V 等 CPU 芯片主要面向通用计算，GPU 主要面向图形和 AI 训练推理，而 ASIC、FPGA、DSP 等芯片则专注于定制化/半定制化计算和数字信号处理等功能。计算资源旨在满足边缘设备对于实时性、可靠性和计算能力的需求，能够在接近数据源的位置进行本地化计算处理，有效减少数据传输延迟，并降低对网络带宽的依赖。边缘算力资源支持异构架构，能够高效处理包括人工智能推理、图像处理、信号处理等在内的复杂任务。

3.1.1.2 存储资源

存储资源主要负责在边缘节点附近保存并缓存数据，主要包括边缘算力设备

中的 RAM、HDD、SSD 及各类 RAID 阵列等。边缘存储将数据从远距离的云服务器端迁移到离数据更近的边缘存储设备端，可以提供实时可靠的数据存储和访问。边缘存储作为一种基于边缘算力的新型分布式存储架构，将数据分散存储在邻近的边缘存储设备或边缘数据中心，可大幅度缩短数据产生、计算、存储之间的物理距离，提供高速低延迟的边缘数据访问和智能处理能力。边缘存储需具备高性能、高稳定、高可靠等特点，以便满足与边缘算力、网络协同及数据中心存储的协同，从而实现数据的互联互通与共享。

3.1.1.3 网络资源

网络资源在边缘算力基础设施中扮演着至关重要的角色，是确保数据高速传输与高效处理的核心通道。边缘算力的网络资源复杂多样，包括各类以太网、光纤、无线等多种设备，共同构建可靠边缘网络基础设施。

3.1.2 资源虚拟化层

边缘资源虚拟化将物理硬件资源抽象为逻辑资源，使得多个虚拟机或容器可以共享相同的硬件资源。虚拟化技术已在云计算中得到广泛应用，在边缘算力中同样具有重要价值。与云计算不同的是，边缘算力需要处理来自于各种异构设备的多种分布式资源，例如不同厂商的服务器、路由器、网关、传感器甚至是用户终端设备。通过边缘算力资源虚拟化，这些异构资源可以被抽象为统一的资源池。虚拟化层可以屏蔽底层硬件的差异，使得上层应用无需关心具体的硬件类型和配置，只需与虚拟资源接口进行交互。

3.1.2.1 计算资源虚拟化

边缘算力的计算资源虚拟化是指通过虚拟化技术将边缘设备上的物理计算资源抽象为多个虚拟资源，提高硬件利用率，增强系统灵活性，并满足多样化的应用需求。

边缘算力虚拟化通常依赖虚拟机监控程序或容器技术来实现。虚拟机监控程序可以将边缘设备的 CPU 和内存等资源分割为多个虚拟机，使得每个虚拟机能够运行不同的操作系统和应用程序。而容器技术则进一步简化了虚拟化的开销，允许多个容器在同一操作系统内核上运行，具有更轻量、更高效的特点。Kubernetes 等容器编排工具在边缘算力中得到广泛应用，用于自动化地管理和调度这些容器化应用。

边缘环境中计算资源虚拟化的价值不仅体现在资源整合和提高利用率上，更重要的是支持多租户环境下的应用隔离和弹性扩展。通过虚拟化技术，边缘设备可以在运行多个应用的同时，确保各应用之间的资源隔离，防止相互干扰。同时，虚拟化还能够根据实时需求动态调整资源分配，使得边缘算力能够灵活应对突发负载和复杂应用场景，进一步提升了边缘算力平台的服务能力和响应速度。

3.1.2.2 存储资源虚拟化

边缘算力的存储虚拟化是将边缘设备上物理存储资源抽象为虚拟存储资源的技术。边缘算力设备通常具备不同类型和容量的存储介质，如固态硬盘(SSD)、闪存、甚至是低成本的机械硬盘。通过存储虚拟化，这些异构存储资源可以被整合为一个统一的虚拟存储池，以满足不同应用和服务的数据存储需求，同时简化数据管理和提升存储资源的利用效率。

在边缘算力环境中，存储虚拟化不仅有助于克服存储资源分散的问题，还能够提高数据的可用性和可靠性。通过存储虚拟化技术，边缘设备可以支持数据的分布式存储、自动化备份和跨节点的冗余存储。这种方式确保了即使某些边缘节点发生故障，数据依然可以从其他节点快速恢复，从而减少数据丢失的风险。

此外，存储虚拟化在边缘算力中的另一个重要应用是数据的分层存储。根据数据的访问频率和重要性，可以将数据智能地存储在不同的存储介质上。例如，频繁访问的数据可以存储在高速的 SSD 上，而较少访问的冷数据则可以转移到更大容量但访问速度较慢的硬盘上。这种分层存储机制不仅优化了存储资源的使用，还提高了数据访问的效率。

存储虚拟化还为边缘算力中的数据共享和协作提供了技术保障。通过虚拟化的存储资源，多个边缘节点能够更方便地访问和共享数据，支持边缘算力环境中的协同计算和实时数据处理。同时，存储虚拟化还可以结合数据加密和访问控制机制，确保数据在边缘设备之间传输和存储过程中的安全性。

3.1.2.3 网络资源虚拟化

在边缘算力环境中，网络虚拟化技术扮演着至关重要的角色。边缘节点通常部署于地理位置分散的环境，节点间需要借助网络进行通信和数据交互。然而，传统的网络架构难以满足边缘算力环境下网络拓扑和流量负载的动态变化需求。

网络虚拟化技术支持边缘算力平台对网络资源进行灵活配置的能力。通过软

件定义网络(SDN)和网络功能虚拟化(NFV)等技术，边缘算力平台可以实现虚拟网络的按需创建、动态调整和高效销毁，从而确保应用和服务能够在异构的网络环境下保持高性能运行。

网络虚拟化还支持边缘算力中的多租户隔离。在边缘算力场景中，多个应用或服务可能需要共享同一个物理网络资源。通过网络虚拟化技术，平台可以为不同的租户创建独立的虚拟网络，确保各租户之间的网络流量相互隔离，从而提高网络安全性。此外，虚拟网络的配置和管理也更加灵活，可以根据应用的需求动态调整网络带宽、延迟和可靠性等参数。

在边缘算力的网络虚拟化中，NFV 技术的应用也十分关键。传统的网络功能（如防火墙、负载均衡、路由等）通常依赖于专用硬件设备，而 NFV 通过将这些功能以软件的形式虚拟化，可以在通用硬件上灵活部署，从而减少对专用设备的依赖，提高网络服务的灵活性和可扩展性。在边缘算力中，NFV 可以帮助快速部署和更新网络功能，支持复杂的网络环境和多样化的应用需求。

3.2 边缘算力网络

随着边缘算力的快速发展，边缘节点数量和层级不断增加，为构建灵活高效的算力网络提供了基础。边缘算力网络通过对分布式算力资源进行感知、度量、并网、调度、管控和交易等操作，整合泛在分布的边缘算力资源，实现一体化的接入和管理，从而满足不断增长的算力资源分配需求，成为边缘算力技术的重要研究方向。

3.2.1 边缘算力感知

边缘算力感知是针对具体场景下的边缘算力需求的感知和边缘算力资源的感知，通过面向边缘算力、SLA 需求的感知、分析技术，构建边缘算力度量及感知体系。边缘算力感知的核心在于对异构算力资源进行一体化接入与管理，不仅需要提供快速响应的能力，还需要实现资源利用的动态优化。这对分布式计算节点的管理和调度提出了较高的要求。通过实时感知算力资源的状态、负载和需求，边缘算力感知技术可以动态调整资源分配策略，确保计算任务的高效执行和资源的合理利用。

边缘算力池是实现边缘算力感知的关键机制之一，将计算资源集中于资源池中，使用户可以通过网络以便利的、按需申请的方式获取算力资源。这种集中式

管理不仅实现了算力资源的可视化，还为资源的感知、调度和编排提供了基础。算力资源池可以显著提高资源利用率，通过统一管理降低运维成本，并提高算力资源的调度效率。通过设立算力资源池可实现：（1）资源整合与可视化，算力资源池集中管理异构算力资源，包括其位置分布、计算特性和动态占用等情况，使得资源的可视化更加直观，有助于对资源状态的实时监控和分析。（2）高效调度与动态优化，通过实时感知和分析资源需求，算力资源池能够动态调整资源的分配策略，不仅平衡了算力资源的利用率，还降低了网络拥堵和系统延迟，提高了计算任务的响应速度和系统的整体性能。（3）降低运维成本，集中式资源池可以减少对分散资源的维护和监控成本，实现对整个资源池的高效管理，可以降低运营成本，并提升系统的可靠性和稳定性。

当前，边缘算力感知技术正在向智能化、自动化和全局优化的方向发展。未来，随着人工智能和机器学习技术的引入，边缘算力感知系统将能够更精确地预测资源需求和负载变化，实现更加智能的资源调度和动态优化。此外，边缘算力感知技术将与 5G 等网络技术进一步融合，推动端到端的资源管理和优化，满足更加复杂和多样化的计算需求。

3.2.2 边缘算力度量

在边缘算力的发展过程中，边缘算力度量成为了一个关键指标，用以评估边缘算力系统的性能、效率以及资源利用率，其核心在于量化边缘算力资源以及多样化 SLA 需求，建立统一的标准化的度量指标体系，以实现高效的算力利用和任务处理。通过准确的算力度量，可以更好地优化边缘算力资源的分配，提高系统的整体效能。因此，算力度量技术是实现边缘算力网络的重要基础。类比云计算计量方法，边缘算力度量可以分为计算、存储和网络分别进行度量，以更好的满足业务对算力资源的需求。

边缘算力的计算能力度量通常采用运算能力作为关键指标，常用单位包括 OPS（Operations Per Second）和 FLOPS（Floating-point Operations Per Second）。OPS 泛指处理器每秒所能执行的操作次数，而 FLOPS 特指每秒可执行的浮点运算次数，更侧重于反映处理器的数值计算能力。这里的处理器涵盖范围广泛，不仅包括传统的 CPU，还包括 GPU、FPGA、ASIC、DPU、TPU 等各类专用芯片。不同类型的处理器拥有各自独特的性能指标和应用场景，实际应用中往往需

要采用多维度的度量体系来全面、准确地描述其性能。例如，除了运算能力之外，还需要考虑内存带宽、功耗、延迟等因素。

边缘算力的存储能力决定了其可以存储和处理的数据量。存储能力包括持久存储（如 HDD、SSD）和临时存储（如 RAM）的容量和读写速度。对于数据密集型应用，如视频监控、数据缓存、日志分析等，充足的存储能力是保证系统稳定运行的基础。存储资源可从磁盘/内存存储容量、IO 读写效能、吞吐率等维度进行度量。

边缘算力的网络能力度量除了网络带宽、时延之外，还包括可用私网个数、可用公网 IP 地址数等。网络延迟是指数据在边缘节点之间传输所需的时间，通常以毫秒（ms）为单位衡量。低网络延迟是边缘算力的一大优势，因为它能够显著减少数据从源头到处理节点的传输时间，提高系统的实时性和响应速度。

随着边缘基础设施的发展，电信运营商提供算力和网络的综合服务，其核心能力在于对行业应用场景中算力及网络时延的确定性保障。在这一过程中，电信运营商通过对算力资源与网络资源提供统一的度量标准，对行业应用场景中相关算网资源进行量化描述，然后根据业务 SLA 对算网资源进行转译建模，进而支持将业务按需映射到网络、计算及存储资源，为业务的动态调度和路由寻址提供可靠支撑，实现多样性算力资源调度与管理，大幅提高算网应用中各个网元间的协同工作效率。目前，边缘由于其异构性算力资源的度量还缺乏统一的标准和衡量方式，业界研究机构、产业联盟、标准组织等尚未形成统一结论。基于 SLA 通过统一的量化描述将异构算力资源与多样化的业务需求综合考虑，进而实现对算力资源的度量，或可成为边缘算力度量技术发展的方向之一。随着边缘算力与网络资源的深度融合，边缘算网统一度量将有效推动算力资源和网络资源的有机融合与协同优化，从而实现形成标准化归一化的度量体系。

3.2.3 边缘算力调度

边缘算力调度技术对分布式计算资源（如 CPU、GPU、FPGA 等）进行动态管理和调度，根据实时需求、资源状态和网络条件将工作负载智能分配到边缘算力节点，以优化性能、减少延迟和提升资源利用效率。

边缘算力调度技术主要包括以下方面：（1）动态资源分配，能够实时监测各个边缘节点的资源状况，如 CPU、内存、存储和网络带宽，动态调整任务分配。

（2）负载均衡，通过算法自动将任务分散到多个边缘节点，防止个别节点过载，从而提升系统整体性能。（3）延迟优化，根据用户位置和网络条件，智能选择最合适的边缘节点来执行请求，减少数据传输的延迟。（4）智能分析与预测，利用机器学习技术分析历史数据，预测未来的资源需求变化，提前做好相应的资源准备。（5）故障恢复与容错，在边缘节点发生故障时，系统能够迅速将任务迁移到其他健康节点，确保服务的连续性。

未来，边缘算力调度将更多地依赖于智能化技术，如可编程网络技术和智能感知网络技术，以提升调度效率和灵活性。同时，面向算力大规模落地的趋势，边缘算力调度通过原生 AI 算力工具让不同种类的芯片大规模并行，同时发挥最大效率，并让算力使用者无需关注不同芯片生态，做到随取随用。

3.2.4 边缘算力管控

边缘算力管控技术是指对边缘算力资源进行有效管理，实现计算、存储、网络等资源的协同与优化，以满足多种应用场景的需求。

边缘算力管控根据任务类型、优先级、资源要求等因素，采用合适的调度算法（如最长任务优先、最短作业优先等）对算力任务进行合理调度和分配，使得资源得以充分利用，并确保任务的高效执行。同时，边缘算力管控通过深度集成云端和边缘节点，实现跨层次的服务，优化资源利用和任务调度，提升整体系统的效率和响应速度，并通过对历史数据和未来趋势的分析，进行算力需求的预测，动态调整边缘缓存的内容和位置，以提高数据访问速度和用户体验。

3.2.5 边缘算力并网

边缘算力并网是将边缘算力的分布式计算能力与网络资源深度融合，形成一种新型的信息服务模式，以满足算网融合的需求。算力并网通过网络将大量闲散的资源连接起来并进行统一管理和调度，同时实现多级资源节点的协同调度与应用的灵活部署。

边缘算力并网作为算力网络提供服务的重要方式之一，以解决现网资源调度需求为目标，面向典型业务场景实现多方算力对接互联与协同共享、算力资源一体化调度，构建动态共享的新型基础设施合作模式。

边缘算力具有分布式特性，需要通过网络调度能力实现算力资源的调度、共享等，因此，算力并网需要具备较强的感知能力，不仅能够感知不同边缘应用的

算力需求，还要能实时感知算力互联网络的连接状况，为边缘应用提供差异化高可靠的算力服务。

3.3 边缘智能

边缘智能通过在边缘节点应用人工智能算法进行部分训练和推理决策。移动终端等设备通过将深度学习模型的推理或训练任务卸载到临近的边缘算力节点，以完成终端设备的本地计算与边缘服务器强计算能力的协同互补，进而降低终端设备自身资源消耗和任务推理的时延或模型训练的能耗，保证良好的用户体验。同时，将人工智能模型部署在边缘设备上，可以为用户提供更加实时的智能应用服务。此外，依托远端的云计算服务，根据设备类型和场景需求，可以进行近端边缘设备的大规模安全配置、部署和管理以及服务资源的智能分配，从而让智能能力在云端和边缘之间按需流动。

边缘智能是边缘算力的重要使能技术，涉及到系统部署、数据处理、模型优化、边缘训练、边缘推理等关键问题。

3.3.1 系统部署

边缘算力设备异构性强、计算资源受限，软件部署框架对在不同环境中实现和运行边缘智能起着至关重要的保障作用。边缘智能系统部署不仅包括了主流的裸机、容器、无服务器计算（Serverless）等隔离部署方式，也涵盖了各种智能应用部署软件框架。这些环境和技术的选择对系统的性能、可靠性和可扩展性有着直接影响。

裸机部署是直接在物理服务器上运行应用程序的部署方式，避免了虚拟化带来的性能开销，从而提供了较高的计算性能和快速响应能力。该方式使得应用能够充分利用处理器、内存和存储等硬件资源，提供优异的计算性能和响应速度。但其缺点在于其灵活性较低，硬件资源必须事先分配和配置，可能导致资源闲置同时扩展和维护通常需要物理访问服务器，复杂且成本高，大规模部署可能需要更多的时间和专业人员。因此在实际系统中，通常采用**容器化方式**进行快速部署，能够提供轻量级、可移植和可扩展的方式来打包、部署和运行应用程序及其依赖项。**无服务器计算**提供了比容器更加便捷的边缘智能模型部署方式，开发者可以专注于代码的功能开发而无需关心底层硬件的管理，平台自动管理应用的扩展需求，根据实际使用来动态分配资源，不仅提高了资源的利用率，还可能降低运营

成本。无服务器计算极大简化了部署和维护的过程，加速了开发和应用的迭代周期，尤其适合于处理边缘端间歇性或不连续的工作负载，例如基于事件的触发器、轻量级微服务等场景，能够灵活响应业务需求的变化。除上述基础部署框架外，机器学习领域还有如 TVM、PyTorch、TensorFlow、ONNX 和 PaddlePaddle 等一系列的智能应用框架，支持从模型训练到部署的全过程，使开发者能够针对不同硬件环境实现高效的运行，有效提升应用的开发与执行的效率、可扩展性和最终的性能。

3.3.2 数据处理

边缘智能以分布式方法实现人工智能训练和推理，对分布式的数据处理存在较高要求，技术上需要对数采协议、数据脱敏、数据清洗、特征工程等进行深入研究。

3.3.2.1 数采协议

在边缘智能场景中，随着网络自动化、通感和网络分析等 5G 应用的兴起，人工智能和机器学习的使用愈加频繁，带来了高频率和海量的数据交互，具有巨大的通信开销。面对高频次、大流量的数据交互需求，需要对现有的服务化接口增强优化，以便在用户面、核心网或无线侧的会话级或设备级的数据交互场景中更高效的汇集所有数据。例如，采用 UDP 协议替代 HTTP+TCP 协议，可降低数倍 CPU 开销。这些增强型协议未来也可用于模型交互和模型分发等应用场景，在多节点间完成更低消耗、更高效率的数据交换和任务协调，以确保数据可以在最短时间内被处理和响应，并保障数据传输的可靠性和可扩展性。3GPP R19 已开展智能数采协议的相关研究。

3.3.2.2 数据脱敏

数据脱敏是一种通过数据变形等方式处理敏感数据，从而降低敏感数据暴露风险的数据处理技术。在边缘智能场景中，大量数据在边缘侧产生并部分传输至中心服务器进行处理。这些数据中可能包含个人隐私信息或企业敏感数据，需进行数据脱敏处理，有效防止隐私数据在采集、传输、使用等环节中暴露。边缘数据脱敏需在边缘数据传输到云服务器之前，在保留数据原本特性的基础上，根据不同的数据脱敏规则和算法，对特定敏感信息进行模糊化、无效化、替换、乱序、加密或掩码处理，并且保证脱敏后的数据依旧存在可用性，满足后续 AI 分析、

机器学习、关联分析等应用场景的使用需求。

3.3.2.3 数据清洗

数据清洗是对数据进行去重、填补缺失值、处理异常值和转换格式等操作，以去除或修复数据集中存在的错误、不一致、不完整和冗余等现象，以确保数据质量和准确性的技术。在边缘智能场景中，边缘侧收集的大量数据往往具有噪声或异常值，直接分析未经清洗处理的数据会影响效率和准确性。在实际应用中，一些工具和技术如 Pandas、NumPy 和 OpenRefine 可用于数据清洗，消除错误数据和干扰噪声，从而提高智能分析和建模精度。例如，在工业互联网场景中，传感器数据易受到工业现场恶劣环境的干扰或设备故障影响，而数据清洗可过滤掉无效数据，确保后续数据分析的可靠性和高效性。

3.3.2.4 特征工程

特征工程是数据科学和机器学习中的关键步骤之一，是从原始数据中提取、选择、转换和创建对预测模型有用的特征，以提高模型性能和效果的技术。首先，特征提取是特征工程的重要部分之一，指从原始数据中抽取有效的信息。例如，在图像处理任务中，边缘设备采集的原始图像数据包含成千上万的像素值。通过特征提取，可将像素值转化为诸如边缘、纹理或形状等更具代表性的特征，以便后续模型训练或推理时可以更高效地理解图像内容。其次，特征选择也是特征工程的关键步骤。边缘侧数据集通常具备大量数据特征，但并非所有特征都与分析任务存在直接关联。特征选择技术可以筛选出对模型的训练和推理最具贡献的典型特征，减少冗余和相关性低的特征，防止模型过拟合。在实际应用中，常见的特征选择方法包括过滤法（如方差阈值）、包裹法（如递归特征消除）以及嵌入法（如 L1 正则化）等。特征变换是特征工程的另一核心环节，需对现有边缘数据特征进行数学或统计变换，以提升特征的表现力。例如，可通过数据的标准化和归一化调整特征的范围或分布，从而使模型更快收敛。在边缘侧进行特征工程，从原始数据中提取、转换和选择有意义的特征，使模型更好地理解数据的内在模式，从而可以显著提升模型的准确性、鲁棒性和泛化能力。

3.3.3 模型优化

在边缘智能场景中，模型的优化处理直接影响到应用的性能和效率。目前有主流优化技术包括轻量化设计、模型压缩、模型分割等，旨在解决边缘设备的资

源限制问题，通过智能的算法和策略，优化数据处理过程，降低延迟，节省内存资源和能耗。

3.3.3.1 轻量化设计

模型轻量化设计是指通过一系列技术和方法，对原始的大型 AI 模型进行优化和压缩，减少深度神经网络的参数和计算量，以减少其所需的计算资源、存储空间和运行时间，同时保持或提升模型的预测精度和泛化能力，特别是在边缘设备、移动设备或资源受限的环境中。如 MobileNet、ShuffleNet 和 EfficientNet 等更小、更高效的网络架构，通过引入深度可分离卷积、组卷积和通道分组等技术，大幅减少计算量和参数数量。在此基础上衍生出了神经架构搜索（NAS）等自动模型轻量化方法，通过优化搜索和评估过程，自动生成高性能的神经网络架构，从而提高模型的性能并减少人工设计的精力，使其更易于在实际应用中推广和使用。

3.3.3.2 模型压缩

模型压缩指的是通过减少模型的参数量来减小模型的体积和计算量，从而使得复杂的深度学习模型能够在资源受限的边缘设备上运行，这对于计算资源相对受限的边缘设备来说非常重要。按照压缩过程对网络结构的破坏程度，一般将模型压缩技术分为“前端压缩”和“后端压缩”两部分。前端压缩，是指在不改变原网络结构的压缩技术，主要包括知识蒸馏、轻量级网络（紧凑的模型结构设计）以及滤波器层面的剪枝（结构化剪枝）等；后端压缩，是指包括低秩近似、未加限制的剪枝（非结构化剪枝/稀疏）、参数量化以及二值网络等，目标在于尽可能减少模型大小，会对原始网络结构造成极大程度的改造。

3.3.3.3 模型分割

模型分割通常指的是将一个大型或复杂的机器学习模型（尤其是深度神经网络）拆分成多个部分，这些部分可以在不同的硬件或设备上独立运行。模型分割的主要目的是优化计算资源的使用，优化目标主要在于时延最小化、内存最小化和能耗最小化，以改善性能表现、减少延迟、降低运行成本。在许多实时应用和移动设备中，由于资源限制，模型分割变得尤为重要。在多个设备或处理单元之间智能地分配计算任务，可以最大化硬件的使用效率，尤其是在异构计算环境中，不同的硬件可能对不同类型的计算任务有不同的优化和加速，从而使得任务可以

更好的完成。根据分割的维度，可分为将不同神经网络层级切分的流水线并行、将一层分配到不同设备上的数据并行等策略。在松耦合的边缘侧，也即通信开销较大的节点间，通常使用流水线并行策略，降低层内数据的传输开销；在紧耦合的边缘数据中心，可考虑采用数据并行与流水线并行混合的策略，更好的利用计算资源。同时，模型分割可以提高模型的伸缩性，通过和其余设备共享计算的结果，一些计算资源受限的设备上也可以完成复杂的计算。在决策延迟方面，模型分割可以帮助将关键的决策计算前置到本地执行，从而减少了响应时间。模型分割的实现通常涉及多个步骤，包括模型的结构分析、计算瓶颈的识别、计算任务的划分以及跨多个平台的同步与通信优化。

3.3.4 边缘训练

传统云端训练需要稳定且高速的网络连接来上传和下载数据及模型，对网络带宽和时延都提出了很高的要求。边缘训练是一种模型分布式训练方法，主要在本地处理数据，数据传输量更少、距离更短，可以有效减少对网络带宽、时延等方面的苛刻要求。然而，由于边缘侧数据来源复杂、资源异构受限，且在训练参与方更多，导致边缘训练机制方法更加复杂。传统云端训练的很多大模型训练技术技巧（如 3D 混合并行训练、激活重计算等）和自动并行工具（如数据并行、张量并行、流水线并行等）仍需进一步研究适配到边侧训练。目前，学术界针对边缘训练过程中多边缘节点协作训练与隐私保护、边缘智能模型的知识迁移、边缘模型的持续优化等三方面主要问题，分别提出了联邦学习、迁移学习与持续学习三类技术分支进行解决。

3.3.4.1 联邦学习

联邦学习（Federated Learning）是一种分布式机器学习方法，允许模型在多个设备上进行训练，而无需将数据集中到中央服务器。这种方法通过在本地设备上进行训练，并仅共享模型参数或梯度，保护了用户的隐私和数据安全，为实现大规模分布式机器学习提供了新的可能。联邦学习适用于边缘设备多、数据分散且敏感的场景，在如移动设备上的个性化服务、医疗数据共享、物联网设备管理等领域都有较大的应用潜力。现有研究中，技术创新主要集中在提高通信效率、保护数据隐私、实现个性化和自适应的训练。

3.3.4.2 迁移学习

迁移学习（Transfer Learning）是将从一个任务（称为源任务或源域）学到的知识应用到另一个不同但相关的任务（称为目标任务或目标域）上的机器学习技术。其核心思想是利用已有的知识来解决新的问题，可以减少对大量标记数据的依赖，提高学习效率，帮助模型在新的任务上获得更好的泛化能力。迁移学习在计算机视觉、自然语言处理、语音识别等多个领域都有广泛应用，特别是在那些标记数据稀缺或获取成本高昂的领域。通过迁移学习，可以有效地利用有限的资源来训练出性能较好的模型，但也在源域和目标域具有较大差异时，存在负迁移、领域适应、模型选择等问题。在边缘算力中，由于资源限制（如计算能力、存储空间和带宽）和数据的动态性，传统的集中式学习方法往往无法满足实时处理和高效利用数据的需求。因此，面向边缘训练的迁移学习成为了提高边缘设备上模型性能和降低延迟的有效手段。面向边缘训练的迁移学习关键技术有以下几类：（1）分布式模型训练，在多个计算节点上并行训练模型的方法，以提高训练效率和模型性能，由于资源受限和数据分布的非独立同分布性，该方法能够有效利用边缘设备的计算能力和存储资源，同时减少数据传输的需求。（2）去中心化模型训练，无需中心服务器参与的模型训练方法，各边缘节点通过交换参数或梯度信息来共同更新模型，可以提高系统的鲁棒性和安全性，避免单点故障。（3）智能任务负载迁移，根据边缘设备的资源状态和任务需求，动态调整任务的执行位置，以优化资源使用和降低延迟，涉及到基于深度强化学习等复杂决策算法。（4）对抗知识迁移，将复杂模型中的知识迁移到简化模型的技术，以提高后者的鲁棒性，通常涉及从复杂模型到简化模型的知识蒸馏过程。（5）计算迁移策略，在不同的计算节点（如云端和边缘设备）之间动态地分配和重新分配计算任务和资源，通过将计算密集型的任务迁移到具有更多资源的中心节点，或者将简单或不那么密集的任务保留在边缘设备上执行。

3.3.4.3 持续学习

持续学习（Continuous Learning）是一种允许机器学习模型在持续接收新数据的同时不断改进和调整其性能的机器学习范式。它能够不断地学习和适应而不会遗忘之前学到的知识，使得模型在面对不断变化的数据或任务时，能够保持高效的学习能力。持续学习中模型可以处理未标记数据，不断扩展其知识库。已有的学术研究借助记忆回放、参数隔离等方式实现高效的持续学习。在边缘智能

场景中，由于计算设备的算力往往受限，因此通常将全量大模型蒸馏为小模型使用。小模型仅能在部分场景中达到最优，若遇到“数据漂移”情况，也即模型无法适应场景的变化，就需要使用持续学习来进行更新。其面临挑战包括以下几个方面：（1） 过拟合，模型可能会对新数据过度适应，导致性能下降。（2） 资源限制，某些情况下模型可能无法实时处理新数据，特别是在资源受限的环境中。（3） 数据分布变化，随着时间的推移，数据的分布可能会发生变化，这可能需要模型进行重大调整。

3.3.5 边缘推理

分布式推理是边缘智能的一项关键技术，通过在多个边缘设备上分布部署和执行推理任务，实现高效的数据处理和智能决策。例如，在智能监控中，可以将视频分析任务分布到多个摄像头上进行本地化处理，通过分布式推理引擎实现实时的目标检测和行为分析；在智能制造中，可以将生产线上的多个传感器数据分布处理，使用智能边缘网关进行数据聚合和边缘推理加速，实现设备状态的实时监测和故障预测。工程实现方面，通常利用分布式计算框架实现多节点的协同工作完成分布式推理，边缘设备间通过高效的通信协议进行数据交换和任务协调。分布式推理不仅提高了系统的响应速度，还增强了系统的鲁棒性和可靠性，适用于各类需要实时决策的应用场景。目前边缘推理的关键技术主要包括条件计算、知识蒸馏、计算卸载等。

3.3.5.1 条件计算

条件计算（Conditional Computation）通过在推理过程中仅激活模型的部分计算路径来提高效率和减少资源消耗。其核心思想是根据输入数据的特性或模型当前的状态，动态地决定哪些计算路径（如网络层或神经元）应该被激活。这样可以避免对不相关的计算路径进行不必要的计算，从而节省计算资源并加快推理速度，同时也增强模型灵活性，使模型能够更好地适应多样化的输入数据。

条件计算特别适用于输入数据具有高度可变性的场景，如自然语言处理(NLP)中的文本分类或机器翻译。在图像识别任务中，条件计算可以帮助模型专注于图像中的特定区域，而不是整个图像。

3.3.5.2 知识蒸馏

知识蒸馏通过训练模仿较大模型的输出，将知识从一个大型训练模型（或模

型集合）转移到一个较小的模型中进行部署。知识蒸馏的核心思想是通过引导轻量化小模型模仿性能更强大的大模型的行为，将大模型的知识迁移至小模型。根据小模型所模仿的不同行为类型，知识蒸馏可分为基于对元的知识蒸馏、基于特征的知识蒸馏、基于关系的知识蒸馏，以及面向大模型涌现能力的知识蒸馏四类。

3.3.5.3 计算卸载

计算卸载是指资源受限的设备将资源密集型的计算任务部分或全部迁移到资源丰富的附近设备上，以解决移动设备在资源存储、计算性能和能效方面的不足。计算卸载技术不仅减少了核心网络的压力，还减少了传输导致的延迟。该技术主要集中在两个问题上：卸载决策和资源分配。其中，卸载决策涉及如何卸载计算任务、卸载多少以及为移动设备卸载什么；资源分配则研究在哪里进行资源的卸载。卸载决策的执行分为三个步骤，首先是代码解析器确定可以卸载的内容，具体的卸载内容取决于应用类型和代码数据分区；接着，系统解析器负责监控各种参数，如可用带宽、待卸载数据的大小或执行本地应用程序的能源消耗，最决策引擎决定是否卸载。卸载决策结果分为三种情况：本地执行、完全卸载和部分卸载，具体决策结果由完成计算任务时的能量消耗和延迟决定。根据卸载决策的优化目标，计算卸载可以分为三种类型：减少延迟、减少能量消耗，以及平衡能量消耗和延迟。

3.4 边缘算力安全

边缘算力节点分布广、环境复杂、数量庞大等特点，很多应用在设计之初未能完备的考虑安全风险，包括算力用户数据的安全和算力提供者基础设施的安全等，传统的安全防护手段已经不能完全适用泛在计算的安全防护需求，一旦被攻击控制，可能会带来较大的安全风险，影响用户的数据，阻碍行业的发展。目前边缘算力还处于发展初期，将安全与业务同步规划建设，才能保证边缘算力的健康发展。

3.4.1 基础设施安全

随着业务应用的推广，边缘算力面临边缘节点被攻击、边缘节点作为跳板向算力网络发动横向或纵向攻击、边缘节点部署环境不安全等风险。针对上述风险，需重点从网络服务安全、边缘算力平台安全、能力开放安全、数据安全等维度进行安全防护。在网络服务安全方面，需实现网安全和用户面网络功能安全。在边

缘算力平台安全方面,重点考虑系统安全、边缘服务授权、应用切换过程中的服务认证和授权、用户接入安全等。在能力开放安全方面,应对 API 进行安全管理、发布和开放。在数据安全方面，应提供轻量级数据加密、数据安全存储、敏感数据处理和敏感数据监测等关键技术能力，保障数据全生命周期安全,保障边缘算力所处理数据的完整性、保密性和可用性。

3.4.2 数据安全

边缘算力中数据分散到多方算力节点进行计算,使数据面临隐私泄露和结果篡改的风险。因此，需要采用数据安全防护、数据流转安全以及计算安全等多种技术实现对数据的可用不可见和实时感知。边缘算力中的数据安全防护是对数据全生命周期的防护,覆盖数据采集、数据传输、数据存储数据处理、数据共享、数据销毁等各阶段,同时对各阶段的数据安全风险进行集中监测与预警处置。数据流转过程中需对数据做好标识,对数据流转节点、数据操作、数据流向等信息进行记录,要构建跨系统的统一的数据流转标识和预授权能力,实现数据出网可管控、数据流转可感知,在算力网络中引入数据标识和流转监测技术，及时发现数据流转安全威胁,做到对数据流转的安全可控。在多方算力节点参与计算的场景下，为保障数据安全引人隐私计算技术。隐私计算主要包括密码学、可信执行环境、信息混淆脱敏、分布式计算等技术路线,分别适合不同的应用和场景。

3.4.3 网络安全

目前视频内容已成为移动网络流量主要来源,其传输处理消耗大量边缘算力资源，对基础设施提出了很大的挑战。在大型网络中，往往存在恶意设备的潜在渗透，恶意设备通过向网络中注入虚假信息或篡改数据等恶意内容，破坏系统可靠性。此外，内部攻击者为避免被发现，可能采取相互保护措施，需要对网络中的恶意设备进行有效识别。构建基于主动检测的信任评估方法，消除对主观经验或正确数据包转发的依赖，使信任证据更加准确。设备将被检测设备提供的可验证内容的校验和转发给基站，并对可疑节点进行有针对性的主动检测，足够的交互可以使设备之间的信任链更加准确和牢固,从而减弱恶意设备碰撞攻击的影响,实现算力信息的安全通信。

3.4.4 调度安全

边缘算力设备通过交换机路由器等联网，构建成了云边端的服务架构，弹性

地向客户提供各种服务。为保障边缘算力能够高效调度、顺畅运转，一方面要做好设备和通信链路的主从备份，另外还要做好调度策略的应急预案，保证在部分设备失效的情况下，仍然保证边缘算力可以有效调度资源，保障计算任务的高效完成。

3.4.5 服务安全

边缘算力系统中，用户频繁地通过网络使用相关的算力服务和资源，因此需要保障在用户和算力服务资源提供商之间的算力交易安全可信。通过采用区块链技术，将算力服务交易和结算的逻辑规则部署在区块链的智能合约层，可实现高效的算力服务交易和结算。另外，由于边缘算力资源的泛在特性，计算资源不仅可以由边缘算力节点来提供，也可由家庭网关设备、智能移动终端等设备提供，通过区块链的分布式可信机制，将提供计算资源的各异构节点或设备统一纳入分布式泛在计算资源的感知和管理中，构建基于区块链的泛在计算资源可信管理机制，以保证计算资源的可信接入和可信服务，进而保证各计算资源提供方的利益以及用户的业务安全。

4 边缘算力典型应用场景

4.1 工业互联网

4.1.1 工业互联网场景概述

工业互联网是新一代信息通信技术与工业经济深度融合的演进方向，实现了工业数据的全面感知、互联互通、汇聚分析和优化应用，为制造业转型升级和新业态发展提供了强大动力。

传统工业生产过程中存在着信息孤岛化、资源利用率低、产品质量难以保证、生产模式僵化、安全生产管理薄弱等诸多挑战。而工业互联网的出现，将连接、数据和智能的力量注入到工业生产的各个环节，为构建现代化产业体系助力，推动经济高质量发展。

4.1.2 工业互联网对边缘算力的需求

传统工业企业依赖云计算中心集中式地处理数据，存在高延迟、带宽压力大、数据安全风险、应用灵活性不足以及对网络连接依赖性强等问题，难以满足实时性、个性化需求，且易受网络中断影响，尤其在偏远地区应用困难。

边缘算力的出现为解决上述问题提供了新的思路，通过将计算资源下沉至网

络边缘，缩短数据处理距离，降低延迟，提升实时处理能力，满足了工业场景实时控制和故障预警等需求。一方面，边缘算力可进行本地数据处理和过滤，降低网络带宽压力和运营成本，并提升数据安全性；另一方面，边缘算力可灵活部署和扩展，满足个性化需求，加速功能部署和系统升级，提高生产效率。

总体而言，工业互联网对边缘算力的要求可以体现为：

1）边缘和现场装置设备的高频实时通信访问；

2）边缘和云端的同步，包括部署等生命周期管理以及合理粒度周期的数据通信访问；

3）借助虚拟化和安全隔离，边缘算力的计算、存储和网络等资源，能满足工业互联网应用的有效运用；

4）边缘算力提供服务的框架，能便捷的提供典型工业互联网应用的服务。

4.1.3 工业互联网边缘算力典型应用

（1）应用背景

生产管控是生产过程中的必备环节，早期主要靠人工操作。在工业互联网应用阶段，往往为了将生产资源、经验等固化，并共享为更多批量化工位使用，生产管控正演进为云边协同方式。由于不同产品需要的生产周期和生产项差异很大，为避免生产中断，以往是工控机本地部署代理程序，这样本地的生产执行和生产任务平台可以解耦，无需等待云端平台的确认即可顺序、连续地执行生产任务。

（2）应用方案

云边协同架构通过在边缘部署代理程序，实现云端模型按需下发至本地执行，满足不同产品的定制化需求。如图 4.1 所示的架构，优势在于利用边缘算力，提高本地响应速度，并能够满足企业多基地、多车间及外协工厂的管理需求，实现统一生产管控模式。

通过集中管控的 5G 网络和边缘算力，可以高效构建和优化云边协同架构。该架构通过集成现场端算力，替代传统工控机，并利用边缘算力节点进行算力分发和管理，实现边缘算力节点与现场端之间的算力协同。这种架构能够在工厂生产一线实现快速部署，具备即插即用的便捷性。

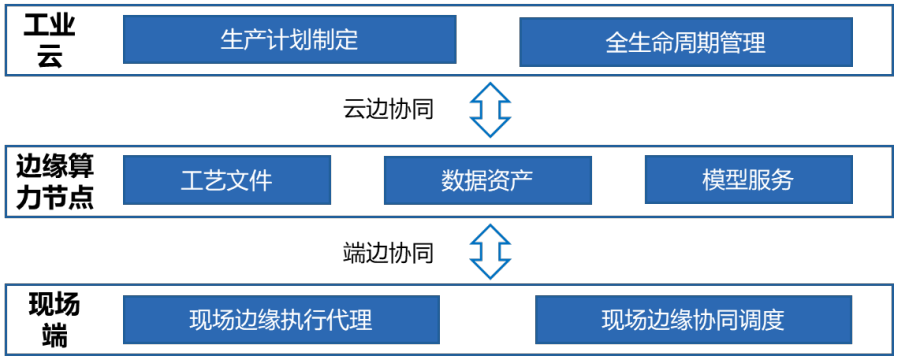

图 4.1 基于边缘算力的工业生产系统架构

考虑到工业设备在不同工作位置可能需要执行不同的工种或任务，边缘算力服务应具备对现场端应用进行动态编排和部署的能力。同时，企业云化应用会随着生产不同产品而迭代演进，更新程序。为了避免频繁调试和升级边端应用，建议采用 PaaS+FaaS 解决方案，解决典型工业边缘算力场景的问题。

在现场端侧，基于容器部署的云边协同应用框架，用户只需在云端开发脚本或进程，即可通过框架将应用派发到边缘网关执行。云端集中管理应用开发、测试和版本存储，实现应用共享。边缘侧设备可自动调度并执行不同任务，只需在框架中编排不同脚本进程即可完成业务行为。

5G 与云边协同框架的结合，能够有效提升工业场景下的效率和灵活性。通过云端任务调度、本地高速执行、动态部署、容器化应用等技术，实现高效的生产流程，并满足不断变化的生产需求。

（3）应用成效

通过将计算资源下沉至工业场景的边缘，利用实时响应、需求定制化、统一管控、稳定可靠、动态部署和应用共享等优势，有效提升了生产效率和灵活性。未来，边缘算力架构将在工业场景中扮演更加关键的角色，促进工业互联网创新发展。

4.2 智慧社区

4.2.1 智慧社区场景概述

智能社区是指通过利用各种智能技术和方式，整合社区现有的各类服务资源，为社区群众提供政务、商务、娱乐、教育、医护及生活互助等多种便捷服务的模式，实现社区管理精细化、服务智能化、居民生活便捷化，最终提升社区治理能

力。

4.2.2 智慧社区对边缘算力的需求

边缘算力作为连接感知终端与智慧应用的桥梁，通过边缘网关汇聚、清洗社区终端采集的视频、图片、结构化数据，并进行边缘智能分析与决策，最终将数据推送至统一管理平台，实现社区智慧化应用。目前，智慧社区在提升整体处理数据能力上对边缘算力需求凸显：

1）低延迟响应：智能安防、家居控制等应用需毫秒级响应，将计算能力推向数据源，减少数据传输延迟，保证实时性和有效性。

2）隐私安全：边缘算力在本地处理社区数据，降低敏感信息传输风险，保护居民隐私和数据安全。

3）网络带宽优化：边缘算力进行数据初步处理和筛选，仅上传关键数据，降低网络传输压力，节省带宽资源和运营成本。

4）高可靠性保障：边缘算力在网络不稳定时提供本地处理和决策能力，保障关键功能不受影响，提升系统可靠性和稳定性。

4.2.3 智慧社区边缘算力典型应用

（1）应用背景

传统社区面临着终端设备布局分散、数量不足、智能化水平低、过度依赖人工等问题,缺乏有效技术手段支持日常管理,导致在应对突发事件时响应速度慢，处理效率低，增加居民安全风险，降低社区整体防范效能。

智慧社区建设通过部署边缘终端设备（摄像头、传感器、监测设备等），实现对社区全方位覆盖和实时监控，并利用边缘算力节点进行数据智能分析。边缘算力可实现对采集数据的实时处理和解析，精准识别安全异常场景（人员聚集、非法入侵、火灾等），及时发出告警通知，确保安全隐患得到有效处置。

（2）应用方案

基于上述背景，社区内部建设边缘算力节点，对感知终端数据进行就近汇聚和治理，数据资源开放共享，满足社区智慧化应用。

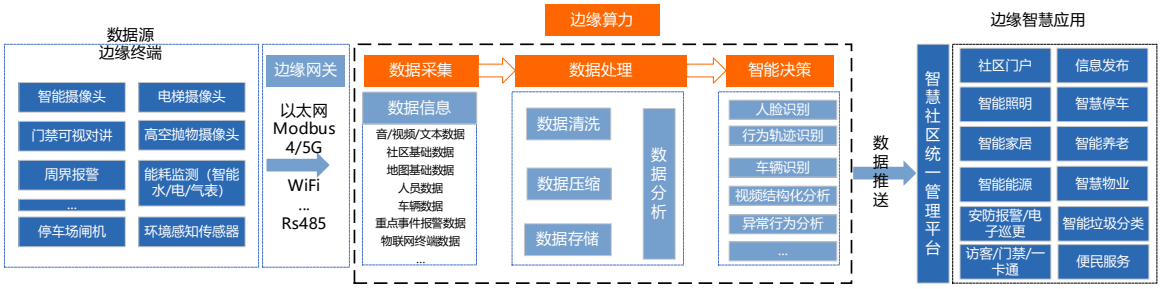

图 4.2 基于边缘算力的智慧社区系统架构

通过在社区内部搭建边缘算力机房，实现数据本地处理与分析，提升社区安全防控能力及快速响应能力。

1）本地化智能分析：利用边缘算力平台对智能摄像头、门禁、环境感知、能耗监测等设备采集的数据进行实时处理，实现人/车/物识别、异常行为分析、重大安防/消防事故监测。

2）感知终端智能化：终端嵌入边缘算力模块，实现数据就近处理与分析，如实时视频分析、人脸识别、行为模式识别等，并可根据异常情况快速触发警报。

3）模块化边缘设备部署：采用标准化接口与协议，实现设备的统一管理、维护、扩容与升级，并优化机房空间利用。

4）数据安全与备份：边缘算力将关键数据本地存储，并与云端平台实现无缝对接，保障数据安全、可访问性及远程备份与共享。

（3）应用成效

边缘算力在智慧社区应用中，通过本地数据处理，提升了数据采集效率和安全性，并增强了应急响应能力。具体表现为：

1）数据采集更实时准确：边缘算力节点对源数据进行本地清洗和预处理，减少无效数据传输，减轻云端压力。

2）数据管理更规范：边缘网关汇聚数据，进行筛选、清洗和压缩，结合智能分析模型识别异常事件，提高处理效率并增强数据隐私安全。

3）应急处置更及时：边缘侧数据分析可智能识别安全异常，及时告警，并进行常态化监测，实现远程集中监控和应急响应。

4.3 智慧能源

4.3.1 智慧能源场景概述

智慧能源是指利用信息通信技术、物联网、人工智能等新兴技术,对能源系统进行全面感知、互联互通和智能化管理的新型能源体系。旨在提高能源利用效率,优化能源结构,实现能源生产、传输、存储、消费等全过程的智能化和协同化。

智慧能源主要涵盖能源的生产、储运、消费和管理四个主要环节。如图 4.3 所示，在能源生产方面,包括新能源(如太阳能、风能)和传统能源(如石油、天然气)在内的多种能源形式,对其进行智慧开采生产。能源储运环节涉及煤炭运输、油气管道等传统方式和智能电网等现代化设施,还涉及集中式储能和分布式储能。能源消费则聚焦于智慧城市、园区，新能源汽车等新型用能模式。最后,能源管理串联其他环节,通过综合能源服务、需求侧响应和能效检测等手段,实现能源、信息和价值的高效流动。

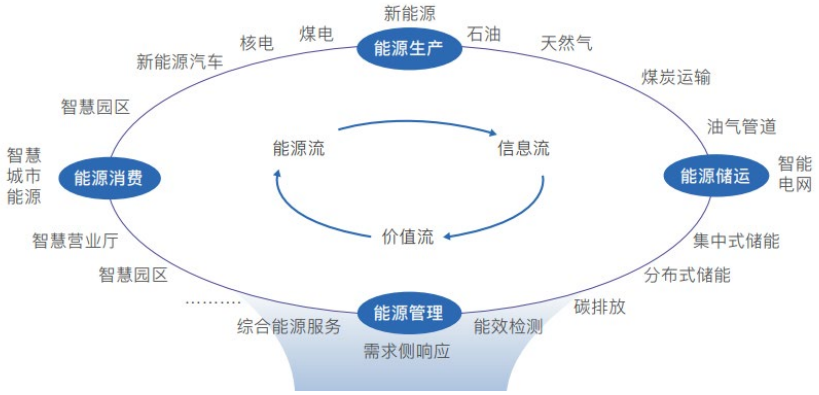

图 4.3 智慧能源主要环节

4.3.2 智慧能源对边缘算力的需求

智慧能源迈向全环节智能化和协同化的进程中,对边缘算力的需求日益迫切。边缘算力作为一种将计算、网络和存储能力下沉至更接近数据源的技术，可以为智慧能源场景提供实时性、可靠性和灵活性，满足其特定需求。

1）实时计算：智慧能源场景对实时性要求较高，例如电网故障处理等，需要边缘侧具备毫秒级甚至微秒级的响应速度。

2）灵活组网：智慧能源场景部署环境复杂，需要边缘侧能够适应不同的网络环境，支持多种通信协议，例如 5G、Wi-Fi、ZigBee 等。

3）自主决策：边缘侧需要具备一定的自主决策能力，能够根据实时数据和预设策略进行自动控制，例如微电网的自主运行等。

4)海量数据存储:智慧能源场景会产生海量的传感器数据、设备运行数据、

用户用能数据等，需要边缘侧具备大容量、高可靠的数据存储能力。

边缘算力是赋能智慧能源的关键，能够有效解决其在数据处理、实时响应、智能化管理等方面的需求，推动能源系统向着更加高效、可靠、智能的方向发展。

4.3.3 智慧能源边缘算力典型应用

（1）应用背景

数字化变电站是电力系统智能化升级的关键，利用前沿技术实现全面的感知、互联和智能化控制。边缘算力作为核心技术，在站内感知设备与上层网络间建立桥梁，通过就近处理提升数据处理效率、实时性和可靠性，降低传输延迟和能耗。边缘算力在数字化变电站的应用解决了传统集中式数据处理模式的延迟、带宽和可靠性问题，为海量设备状态数据的实时监测和分析提供支持，保障电网稳定运行和供电质量。

（2）应用方案

通过在变电站部署边缘算力节点，构建起一个高速、泛在、智能、安全的电力算力体系。边缘节点能够实时处理和分析设备状态数据，如变压器油温、断路器状态、保护装置信号等，利用本地智能模型进行故障诊断，并在必要时通过 5G 等通信技术将数据上传至云端或控制中心。

1）数据采集与处理。变电站内部署的传感器和智能设备采集电压、电流、功率质量等数据。边缘算力节点对接收到的数据进行清洗、特征提取和异常检测，为智能决策提供支持。

2） 智能决策与控制。边缘算力节点根据分析结果，自动触发相应的控制策略，如负荷调整、故障隔离、设备维护预警等，实现智能化的电网运行管理。

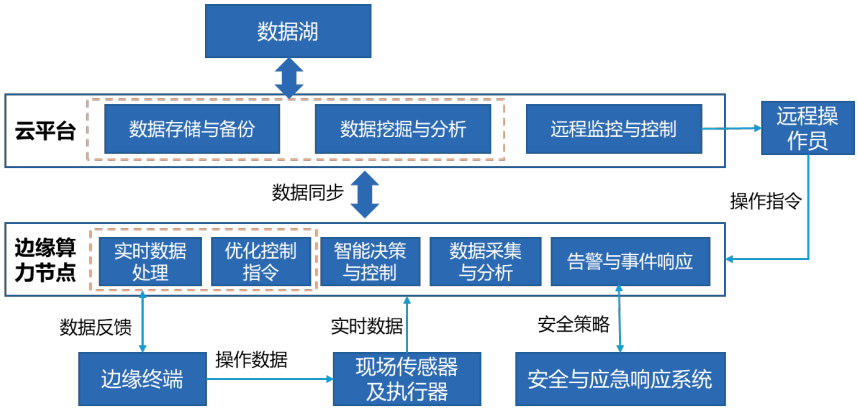

图 4.4 数字化变电站典型应用架构图

如图 4.4 所示，该架构通过边缘算力节点连接终端设备，并利用现场传感器与执行器实现实时数据采集和控制。云平台负责数据存储、分析和远程监控，边缘算力节点负责数据处理和控制指令执行，终端设备负责现场数据采集，现场传感器与执行器负责物理设备的交互。此外，该架构还包含远程操作员、安全与应急响应系统和数据湖，分别用于远程操控、安全保障和数据分析。

（3）应用成效

边缘算力的应用，通过实时数据处理和智能化分析，显著提升了数字化变电站的监控能力、故障诊断速度和应急响应效率，进而增强了电网的可靠性和安全性，降低了运维成本。边缘算力在源头进行数据处理，提高了监控精度和实时性，并支持快速识别故障模式，缩短故障诊断时间；同时，通过自动化控制策略，实现了对突发事件的快速响应，有效降低事故影响，为电力系统的数字化转型提供了有力支撑。

4.4 云游戏

4.4.1 云游戏场景概述

云游戏作为一种基于云计算的交互式在线视频流，有效解决了传统游戏受限于终端硬件和场景的瓶颈，同时弥补了移动游戏在操作体验和内容丰富度方面的不足，实现了便捷性和沉浸式体验的融合。近年来，随着算力基础设施的不断完善，特别是高性能计算、分布式渲染和网络加速技术的进步，为云游戏产业的蓬勃发展提供了强大的驱动力，推动了整个生态系统的快速演进。

4.4.2 云游戏对边缘算力的需求

云游戏作为一种游戏服务模式，其核心在于将原本在本地终端进行的游戏渲染、逻辑运算等高负载任务卸载至边缘算力节点，用户终端仅需负责接收串流的音视频数据并进行简单的输入操作。这种架构对边缘算力节点的计算能力、图形处理能力以及网络传输性能提出了极高的要求，以确保流畅的游戏体验和低延迟的交互响应。

1）低延迟计算：边缘算力节点需具备低延迟计算能力，将复杂的图像渲染、物理运算和 AI 处理任务分布到接近用户的节点上，以减少传输延迟并提升游戏体验。

2）动态资源调度：边缘算力节点需具备动态资源调度能力，根据玩家数量

和游戏场景的变化，毫秒级别地完成算力资源分配，确保流畅的游戏运行。

3）高带宽低延迟网络：云游戏高度依赖高带宽、低延迟的网络环境，特别是在高清画面的实时传输方面，边缘网络通过缩短数据传输距离，减少跨数据中心的延迟。

4.4.3 云游戏边缘算力典型应用

（1）应用背景

近年来，游戏产业蓬勃发展，玩家对游戏体验的要求日益提升，云游戏凭借其便捷性和跨平台特性成为业界焦点。然而，受限于网络传输带来的高延迟、高带宽成本以及中心化架构的并发瓶颈，玩家的游戏体验和云游戏产业的发展都面临着巨大挑战。高延迟导致的画面卡顿、操作滞后严重影响玩家的沉浸感，高带宽需求则推高了云游戏运营成本，而中心化架构难以应对玩家数量激增带来的并发压力，这些问题都亟待有效的解决方案来推动云游戏产业的快速发展。

（2）应用方案

如图 4.5 所示，基于边缘算力的云游戏系统架构核心思路是将云游戏的部分计算、渲染和网络传输功能从中心云迁移至网络边缘，利用边缘算力节点的低时延、高带宽、分布式部署和丰富的云服务优势，打造更加流畅、高效、便捷的云游戏体验。

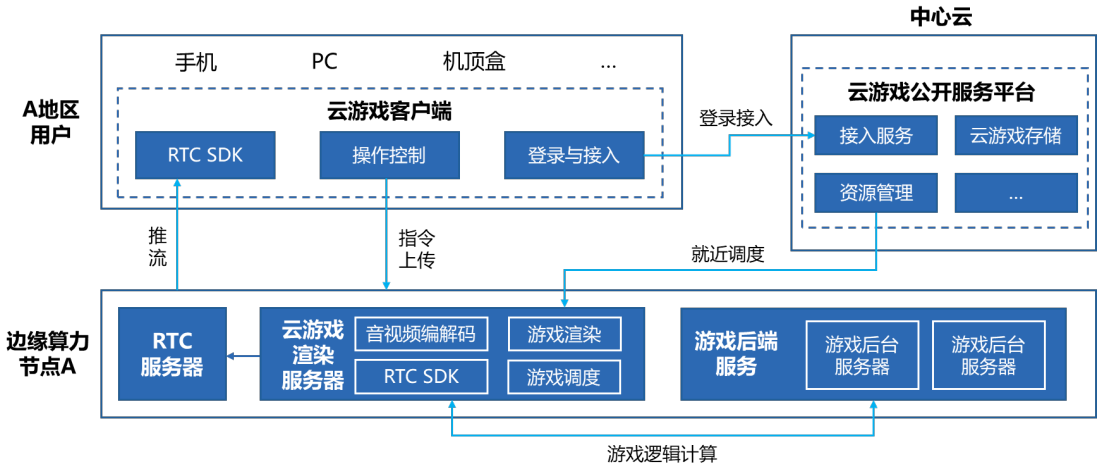

图 4.5 基于边缘算力的云游戏系统架构

该架构主要由本地客户端、云游戏公共服务、云游戏后端服务、云游戏边缘算力节点等几部分组成：

云游戏客户端：云游戏客户端作为用户与云游戏平台交互的入口，集成解码、

用户管理、操作控制等核心功能，将用户的键盘、鼠标等操控指令发送至云游戏实例，同时接收平台回传的音视频流，进行解码并最终呈现给用户，保障流畅的游戏体验。

云游戏公共服务平台：云游戏公共服务平台部署在中心云，提供涵盖游戏生命周期管理的一系列服务。平台基于用户地域、网络状况和游戏算力需求等因素，从资源池中为用户动态分配最优边缘算力节点，确保用户获得流畅的游戏体验。

游戏后端服务：游戏后端服务部署在边缘算力节点，承担着处理云游戏渲染服务器逻辑运算的核心职责，负责接收渲染服务器的输入数据，执行游戏逻辑计算，并将计算结果返回给渲染服务器，最终呈现给玩家。

边缘算力节点:通过构建分布式算力节点实例资源池,为云游戏提供低延迟、高性能的运行环境。云游戏平台根据地域、网络状况及游戏算力需求等因素，智能调度包括 X86+GPU、ARM 等多种规格的边缘节点实例，以适配不同游戏的算力要求。游戏应用部署于这些实例之上，负责指令解析、逻辑运算、渲染、编解码等核心处理流程，并将处理后的游戏画面通过实时音视频传输技术（如 RTC）推流至用户终端，保障流畅的游戏体验。

（3）应用成效

边缘算力为云游戏场景带来了显著的优化和发展机遇。通过将计算和数据处理能力部署至网络边缘，边缘算力有效降低了游戏延迟和卡顿，提升了玩家的游戏流畅度和沉浸感。此外，利用边缘算力节点的带宽资源优势，云游戏平台能够大幅降低运营成本，提升盈利能力。边缘算力的分布式部署特性也增强了系统并发处理能力，满足更多玩家同时在线需求，推动云游戏产业的快速发展。

4.5 轨道交通

4.5.1 轨道交通场景概述

轨道交通是城市公共交通的重要组成部分，包括城内地铁、轻轨、有轨电车与城际铁路、高铁等多种形式。随着城市化进程的加快，轨道交通在缓解交通拥堵、改善出行条件等方面发挥着越来越重要的作用，但与此同时也面临着运营效率、安全可靠性等方面的挑战。为了应对这些挑战，轨道交通引入先进的信息通信技术、人工智能等手段，对运输环境进行全面感知、互联互通和智能化管理以提高轨道交通运营效率和服务质量，为乘客提供更加便捷、舒适的出行体验。

4.5.2 轨道交通对边缘算力的需求

智慧轨道交通利用新兴技术实现全面感知、互联互通和智能化管理，提升运营效率、降低能耗，并为用户提供安全可靠的服务体验。边缘算力作为层级架构核心，连接感知层与网络层，实现近源数据处理，降低时延和能耗，提升安全性和可靠性，简化平台层映射和同步，为智慧轨道交通提供以下关键能力：

1）实时控制响应：为列车控制和信号系统提供低延迟数据处理和决策能力。

2）多源异构数据处理：就地采集、清洗、存储和分析车辆状态、设备工况、视频监控等数据，优化传输效率。

3）高安全性保障：综合分析轨道沿线的环境数据和视频监控，通过和巡检人员、指挥中心的联动快速处置安全隐患，保障轨道交通运营安全。

4)高可靠性保障:通过分布式部署和冗余设计提升系统鲁棒性和恢复能力，确保业务连续性。

5）层级架构支撑：简化数字孪生映射和同步，为深度强化学习提供高质量决策依据。

边缘算力赋能智慧轨道交通，解决实时响应、数据处理、安全性、可靠性和架构支撑等需求，推动系统向更可靠、智能方向发展。

4.5.3 轨道交通边缘算力典型应用

（1）应用背景

传统列车运行状态监测系统依赖集中式数据处理，面临时延、带宽和可靠性挑战。为满足快速故障诊断等需求，亟需在列车中或轨道沿线部署边缘算力节点，对温度、制动压力、轨道环境等信息进行实时分析，并利用预训练模型在毫秒级时间尺度内识别潜在安全隐患，触发预警和应急处置，提升列车运行安全裕度。同时，边缘算力节点对数据进行压缩、特征提取和加密，降低车地通信负荷，提高数据传输安全性和效率。边缘算力将有效解决传统系统瓶颈，实现轨道交通运行状态的实时监测和智能化管理。

（2）应用方案

轨道交通架构如图 4.6 所示，采用分层部署策略，在列车、车站和指挥中心分别部署边缘算力节点，并与云端协同，构建高速、泛在、智能、安全的算力体系。该架构旨在通过边缘算力的实时性与云计算的强大算力相结合，为轨道交通

行业数字化发展提供新动能。

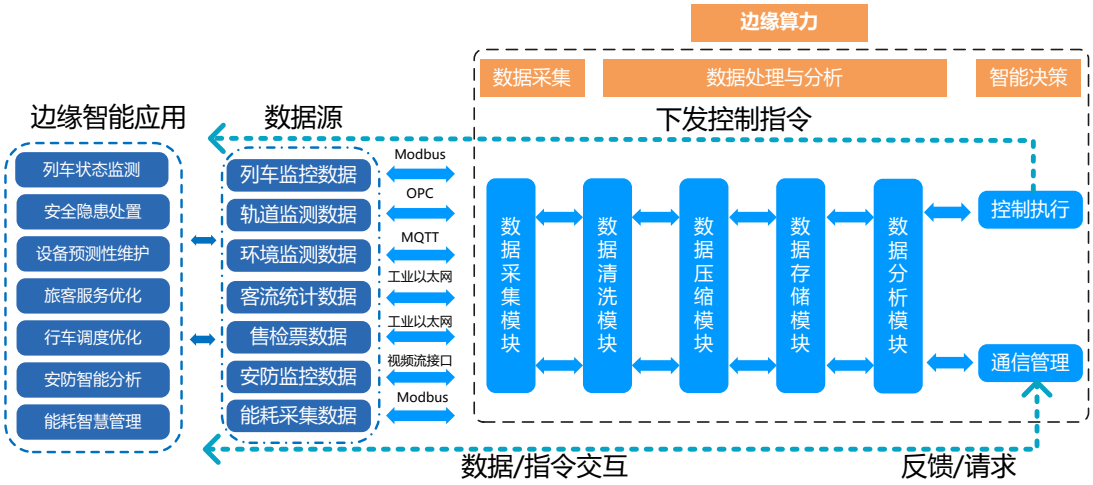

图 4.6 基于边缘算力的轨道交通智能运维系统架构

在列车车载设备中部署边缘算力模块，对车载监测数据进行实时处理，包括数据清洗、特征提取和异常检测，并将预处理后的数据通过 5G 回传至云端。当 5G 网络中断时，边缘节点可调用本地智能模型，对列车状态进行初步判断，保障列车安全运行。

在轨道沿线部署传感器或视频监控设备，通过在轨道沿线重点区域部署的缘算力设备对轨道交通沿线的环境进行分析，实现对轨道交通闯入人员和物体的预警、线路地质灾害预警、违规堆放物的识别，保障轨道交通的安全运营。

在车站和车辆段部署边缘算力服务器，对闸机、售检票机等设备的运行状态数据和视频图像数据进行本地处理，实现设备故障预警、客流统计分析、人脸识别等功能，提高站场智能化水平。

在指挥中心边缘部署边缘算力平台，整合各专业系统数据，构建轨道交通全局数字孪生模型，依托强大的实时计算能力，对列车运行、设备状态、客流分布和轨道线路状态等进行全局实时监控和优化调度，提高轨道交通的安全性、效率和智能化水平。同时，边缘算力平台可作为应急指挥系统，在突发事件发生时提供本地计算和通信能力，保障应急处置决策的时效性。

基于边缘算力的轨道交通智能运维系统将主要实现以下核心功能：

数据采集：轨道交通数据采集涵盖列车、轨道、车站等场景，通过传感器和监控设备获取数据。例如，列车状态监测系统采集车辆关键部件振动、温度等数

据；轨道监测设备实时检测轨道几何状态；售检票系统记录旅客进出站信息；环境监测设备感知站台温湿度和空气质量；能耗设备采集牵引供电系统电压电流数据。

数据处理与分析：边缘算力平台支持多种通信协议接入数据，并提供设备接入管理和元数据管理功能。数据经过预处理，包括格式转换、时间同步、数据清洗等，以解决数据异构性和提升数据质量。平台提供机器学习算法库，支持Python、R、C++等多种开发语言，可实现实时数据分析。应用场景包括：列车振动数据分析诊断部件健康状态、轨道几何状态数据预测变形发展、轨道交通封闭区域闯入物的检测与识别、售检票数据分析优化运力调度等。

智能决策：边缘算力节点基于数据分析结果，驱动各场景应用的智能化闭环控制。例如，根据设备健康诊断结果，节点可向设备控制器下发预警或维护指令，实现设备的预防性维护；根据轨道状态监测与分析，实现轨道异常事件及时处置；根据旅客流量预测结果，节点可优化进出站导流策略；根据环境质量分析结果，节点可调节站台的通风、空调系统参数，提升旅客候车体验。

（3）应用成效

轨道交通边缘算力应用通过在列车、轨道、车站等位置部署传感器、监控设备和边缘算力节点，实时采集并分析振动、位移、电流、图像等信息，实现异常状态的早期预警和诊断，从而有效提升轨道交通系统运行安全性和可靠性，并显著降低运维成本，避免因设备故障导致的行车事故和服务中断。该场景下边缘算力的部署与应用开创了轨交行业智能化升级的新模式，引领行业数字化发展，提升了城市交通运输效率。

4.6 车联网

4.6.1 车联网场景概述

车联网（Internet of Vehicles，IoV）[9]被认为是物联网体系中最有产业潜力、市场需求最明确的领域之一。通过借助新一代信息移动通信技术，实现车内、车与车、车与路、车与人、车与服务平台的全方位网络连接，提升汽车智能化水平和自动驾驶能力，构建汽车和交通服务新业态，从而提高交通效率，改善汽车驾乘感受，为用户提供智能、舒适、安全、节能、高效的综合服务。

4.6.2 车联网对边缘算力的需求

随着智能网联汽车渗透率的不断提高，由此产生的车载信息娱乐、传感器、车路协同、地图等数据传输、处理、存储的需求极大地增加的网络负荷，并对网络时延、带宽、可靠性提出了更高的需求，将边缘算力应用于车联网之后，可以将业务部署在边缘节点，减少数据传输路由长度，从而降低端到端通信时延；还可以作为本地服务托管环境，提供强大的计算、存储资源，支持视频流的实时分析与处理、违章预警、危险驾驶处理等，及部署本地更具地理和区域特色、更高吞吐量的车联网服务。

为了支撑海量的车联网应用，需要边缘算力具备以下的能力：

1)低延迟：应当支持低延迟的端到端数据交互，以满足实时路况信息传递和车辆控制的需求。

2)高带宽：应当支持高带宽的高清视频监控数据传输，以实现对道路状况的精准监控和分析。

3)边缘智能：应当支持轻量化 AI 算法及 AI 模型在边缘侧的部署、推理及训练，以实现实时路况判断、交通流量预测、自动驾驶辅助等功能。

4)异构加速：应当支持异构平台和视频卡和网络加速卡的部署，以提升数据处理效率和智能分析能力。

4.6.3 车联网边缘算力典型应用

（1）应用背景

城市停车难、取车难问题日渐突出，其根源在于停车资源供需失衡、车位利用率低下以及传统停车方式效率低等因素。为解决这一难题，自主代客泊车（Automated Valet Parking，AVP）应运而生。AVP 是一种基于 L4 级自动驾驶技术的解决方案，能够在特定区域内，例如停车场，实现车辆的自主驾驶和泊车，无需人工干预。随着自动驾驶技术的快速发展，AVP 被业界普遍认为是最具商业化落地潜力的自动驾驶应用场景之一，有望有效缓解城市停车压力。

（2）应用方案

基于边缘算力的自主泊车方案，以 5G 通信网络为基础，融合人工智能、北斗定位等新技术，实现人、车、场、云的协同规划、协同感知、协同控制以及协同定位，构建自主泊车场景的商业闭环。车辆可循迹安全行驶，结合场侧融合感知结果，并提供全方位的智慧泊车运营服务，包括行标 V2X 预警、场端/车端运

营监控、车位预约、车辆引导、一键泊车/召车、车辆的多视角实时孪生监控、数据统计和业务分析等服务，可大幅提高停车场运营效率和用户出行效率。

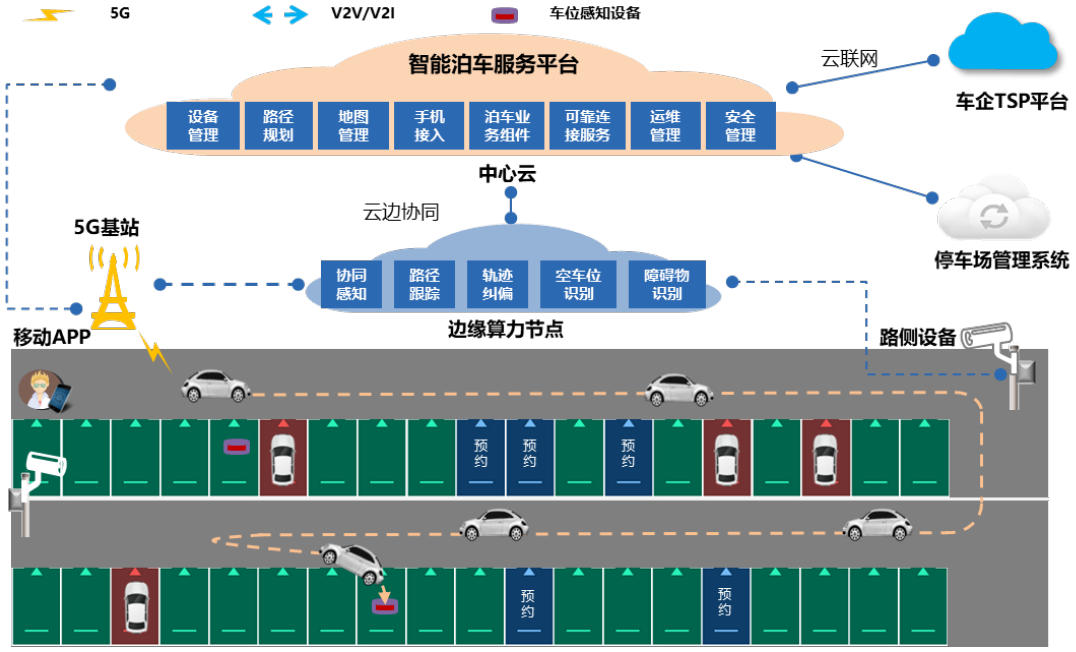

图 4.7 基于边缘算力的自主泊车系统架构

协同感知能力：通过部署在现场端的传感器和部署在边缘算力节点的感知算法进行障碍物识别，并结合车辆自身周边的障碍物感知，进行多源融合感知，使车辆具备盲区感知能力，提升综合感知能力，解决停车场盲区多、遮挡多的问题。

协同规划能力：通过部署在边缘算力节点的全局路径规划算法，在车辆开始泊车时，动态生成全局路径规划结果，自动分配目标停车位，通过 5G 网络，实时将线路发送给车辆，进行全局导航。在车辆端，车辆根据云端下发的车道级导航，结合周围的实时障碍物感知结果，生成局部路径规划，进行局部导航行驶，通过车端和云端的协同规划，解决单车智能无法实现的最优车位选择问题，提升了提升停车场车位整体使用效率，避免停车场拥堵。

协同控制能力：边缘算力节点可以根据现场端收集到的点云数据，判断感知物运动趋势，并结合车辆局部路径规划结果，判断碰撞风险，第一时间对碰撞风险的车辆下达刹车指令，另一方面，车端在实时接收边缘算力节点下发的控制指令的同时，也可以根据自身的感知、预测与规划结果，生成控制指令对车辆进行控制，在提升车辆感知准召率的同时，有效应对行人鬼探头、路口冲出等场景，提升行车安全性，预计可以有效减少 80%以上碰撞风险。

（3）应用成效

基于边缘算力的自主泊车系统，通过实现车辆自动泊车和取车，有效解决了传统泊车方式中用户操作繁琐、体验不佳等问题，并大幅降低了车辆调度和人工运维成本，优化停车资源配置效率。该系统与智慧泊车小程序深度融合，构建了完整的停车闭环体验流程，为用户提供便捷、高效的智能化停车服务。

4.7 未来产业

随着 5G 网络、云服务、AI 等技术的蓬勃发展，边缘算力有望在更多领域发挥作用，如元宇宙、脑机接口等领域。边缘算力通过在靠近数据源的本地设备上处理和分析数据，可显著减少传输延迟，提高系统响应实时性，降低云端算力负担，提供更流畅的用户体验。

以元宇宙为例，边缘算力是推动元宇宙实现低成本、即时响应、灵活扩展以及高隐私性的关键力量。元宇宙是一个与现实世界映射与交互的虚拟世界与协作空间，现有虚拟环境的生成大多源于特定区域，如虚拟街区、虚拟教室，或特定领域，如工业领域、能源领域等。随着扩展现实（XR）、360 度视频等前沿技术的迅猛发展，具有高保真度的沉浸式元宇宙体系正逐步趋于完善。然而，当前的部署模式面临多重挑战：在云端，多媒体内容从全球范围内的云服务器实现流式传输，但难以满足严格的数据隐私保护要求，且传播过程及云基础设施自身的延迟问题显著，成本高昂。而在端侧，用户的头戴式显示器等终端设备，可依托本地计算和存储资源直接渲染和显示虚拟现实场景，但其有限的资源限制了应用的流畅运行，在端侧设备上的独立安装方式也制约了应用扩展的灵活性[10]。

为解决上述问题，边缘算力的部署策略应运而生。将边缘算力部署在距离用户较近的区域，作为连接云与端的桥梁。这种部署方式允许端侧设备将部分处理任务卸载至边缘，减少远程通信开销，从而更有效地处理用户在元宇宙体验中生成的交互数据。同时，云端也可将部分轻量化的深度学习模型任务迁移至边缘执行，提供本地推理服务。例如，边缘节点上可部署不同参数规模、算力需求的轻量化模型，调度不同压缩率数据输入，或是根据用户差异化的精度和时延需求，规划数据版本，分配到各节点的模型上进行推理，带来更大的推理服务优化空间，满足低成本、低延迟、高扩展性和高隐私性的需求，为元宇宙的未来发展铺设更加坚实的技术基础。

边缘算力在元宇宙领域有丰富的场景应用。例如，在元宇宙游戏中，玩家需即时的反馈交互以及流畅的视频画面，通过在边缘侧进行图形渲染和物理计算，可显著降低网络延迟，确保游戏的流畅性和沉浸感。在虚拟旅游场景中，用户可以虚拟浏览世界各地的名声古迹，而边缘算力可支持生成高清 3D 场景，使用户仿佛身临其境。例如，“张家界星球”项目运用了 XR 技术和边缘算力，实现了对张家界自然景观的精准复刻和沉浸式游览。在虚拟街区中，孪生城市的构建需要处理海量的实时数据，包括交通流量、环境监测、城市管理等，边缘算力可以就近处理大量数据，实现快速响应和智能决策，为虚拟城市的呈现提供有力支持。

此外，脑机接口也是边缘算力的重要应用方向之一。脑机接口通过采集用户的脑部活动信息，配合外骨骼及其他可穿戴设备，可为残障人士及神经系统疾病患者提供辅助治疗方案。当前，大多脑电分析设备商在其云端上部署数据分析平台，提供高精度大规模脑电图记录、分析和解码服务。但是，脑机接口需实时处理大量神经信号，这些脑电信号需要高带宽和低延迟的计算能力进行有效解析和响应。因此，边缘算力未来有望成为脑机接口技术的算力分担节点，减少脑电数据传输延迟，快速响应，降低云端数据处理工作量。英国某公司推出的脑电图环，患者可以自行通过平板电脑或手机查看脑电闪烁图像，监测患者脑电变化，再通过云平台和人工智能技术诊断脑状态。

未来，边缘算力将成为支撑下一代互联网应用的基石。例如，多模态、多样化数字内容的生成，大量数据样本的深度学习，用户特征的精准提取等，均离不开边缘算力的应用部署，并为用户提供更加个性化、安全、高效的服务，助力各行各业实现数字化智能化转型，迎接未来时代更多的挑战和机遇。

5　边缘算力未来展望

随着大型语言模型、工业大模型等智能化应用热度不断高涨，边缘算力作为产业智能化发展的数字化底座将迎来战略机遇期，但由于边缘算力属于跨领域融合概念，参与主体众多，在实际部署中存在“三难”问题：一是标准统筹难，云原生、边缘算力、垂直行业边缘算力等发展路径的相关标准组织均从各自领域对边缘算力进行了标准化工作，导致边缘算力标准化工作缺乏统一布局，标准内容上存在一定的冲突和重复，这也成为边缘算力基础设施规模化部署的障碍。二是产业集约化难，目前各个垂直行业在边缘算力领域中独自探索，产业链上下游联

系不够紧密，产业呈现碎片化发展。三是规模部署难，目前产业各方正在积极推进边缘算力基础设施规模化应用部署，但成熟且可复制的建设模式尚未形成，需要进一步探索。在商业模式尚不清晰的前提下，运营商、云计算服务商以及工业企业等核心参与者将难以应对边缘算力基础设施建设运营的资金投入。

为推动边缘算力部署应用，建议产业各方从以下五方面协力共同推进：

（1）加速标准规范体系从建立走向健全

以数字化转型需求为牵引，强化边缘算力与 5G、人工智能等数字技术协同攻关力度，聚焦边云协同、边缘算力网络及边缘智能为代表的边缘算力关键技术方向的基础理论研究攻关，完善边缘算力标准体系，制定实施核心设备、互通接口、测试规范、应用指南等急用先行标准，为边缘算力技术的应用转化夯实基础。

（2）加强边缘算力基础设施的统筹规划

推动边缘算力基础设施与城市发展、建筑物建设的同步规划、同步设计、同步建设，加强电力、网络等基础设施配套建设，并作为重要基础设施纳入国土空间规划。研究制定边缘算力基础设施在机房结构、配套设施、网络、安全等统一建设标准，规范化发展。

（3）建立边缘算力供需对接机制

搭建边缘算力资源匹配对接和交易渠道，引导相关服务商实现资源对接和互联互通，建立统一边缘算力调度平台；规范算力交易和监管机制，为企业提供低时延、低成本的园区、楼宇等边缘算力资源供给，提升边缘算力综合服务水平。

（4）积极推动边缘算力共享试点部署

结合新型基础设施建设规划及发展需求，以共建共享模式统筹规划边缘算力基础设施建设，鼓励边缘算力基础设施与变电站、基站、通信机房等基础设施开放共享，健全跨行业规划协调机制，建立共建共享智慧平台，盘活存量边缘资源价值，提升资源利用效率与效益，助力边缘算力规模化部署发展。

（5）持续促进边缘算力创新应用

依托我国行业门类齐全优势，引导产业链积极开展边缘算力应用试点，构建“边缘算力+N”应用体系，支持各行业龙头企业按需灵活部署边缘算力基础设施，聚焦重点领域共性应用场景，形成一批可复制、可推广的应用模板，“以用促建”，加快边缘算力应用推广落地。

参考文献

[1] 中国移动通信集团有限公司.算力网络技术白皮书[R],2022

[2] 中国信息通信研究院.中国算力发展指数白皮书[R],2023

[3] 边缘计算产业联盟&工业互联网产业联盟.Edge Native 技术白皮书 2.0[R], 2023

[4] IDC&浪潮信息&清华大学全球产业研究院.2021-2022 全球计算力指数评估报告[R],2022

[5] 中国信息通信研究院.中国综合算力评价白皮书[R],2023

[6] 中国信息通信研究院.中国算力中心服务商分析报告[R],2024

[7] 王哲.边缘计算发展现状与趋势展望[J].中国信息通信研究院,2021

[8] YD/T 4255-2023,算力网络 总体技术要求[S].北京：中国通信标准化协会,2023

[9] 中国信息通信研究院&华为技术有限公司&电信科学技术研究院.车联网白皮书[R],2017

[10] 中国信息通信研究院&虚拟现实与元宇宙产业联盟.元宇宙白皮书[R],2023