未来网络发展大会
FUTURE NETWORK
DEVELOPMENT CONFERENCE
NANJING · CHINA / 中国 · 南京

未来网络技术发展系列白皮书（2025）

分布式算力感知与调度技术白皮书

第九届未来网络发展大会组委会

2025年8月

版权声明

编委会

专家指导组成员

刘韵洁 中国工程院院士、江苏省未来网络创新研究院荣誉院长、首席科学家

郭宇辉 中国铁塔通信技术研究院院长、中国通信企业协会低空经济专委会秘书长

黄　韬 北京邮电大学教授

麻文军 中国铁塔行业拓展部总经理、铁塔智联技术有限公司执行董事兼总经理

窦　笠 中国铁塔科技创新部总经理

吴晓梅 中国铁塔通信技术研究院副院长

何　杰 中国铁塔行业拓展部副总经理、铁塔智联技术有限公司副总经理

编制组成员

中国铁塔股份有限公司

闫亚旗、董玉池、潘三明、聂昌、贾平胜、徐佳祥、安颖、王东、汪涛

北京邮电大学

谢人超、唐琴琴、杨煜天、马霄鹏、汪硕

江苏省未来网络创新研究院

魏亮、方辉、孙玉刚、尹鹏、林枭、韩风、占昊天、王磊

前　言

随着算力网络的飞速发展，算力资源呈现出泛在化、异构化、分布化的显著趋势。如何高效感知、协同调度这些广泛分布且动态变化的算力资源，以支撑日益复杂的智能应用需求，已成为推动产业数字化转型和智能化升级的关键挑战与核心技术方向。

本白皮书首先详细阐述了分布式算力感知与调度的背景、需求、体系架构以及关键技术，同时介绍了该技术在远程医疗、智慧城市、大模型分布式训推以及云游戏等领域的典型应用场景，并探讨了当前技术落地、基础设施建设与改造以及标准化建设面临的挑战和发展建议。

目前，工业界和学术界对分布式算力感知与调度技术的研究尚处于起步阶段，并仍处于快速发展之中，新的架构、算法和应用模式不断涌现，本白皮书作为阶段性研究成果，还存在需要不断完善的地方，真诚地企盼读者批评指正。

目　　录

一、分布式算力感知与调度背景

本白皮书创新提出分布式算力感知与调度模型与架构。分布式算力是一种新型的计算模式，在实时感知多类型、多数量计算设备资源状况的基础上，借助统一的度量范式对资源量进行对比与评估，再结合任务的计算强度、时延要求和数据依赖等特征，以及网络带宽和能量预算等约束，运用自适应的智能调度算法将大规模的计算任务分散到不同的计算节点上，从而实现高效的数据处理和分析。本白皮书阐述了分布式算力感知与调度的背景、体系结构、关键技术、应用场景、发展建议，旨在为有兴趣了解分布式算力感知与调度相关概念和技术的研究人员提供介绍与指导。具体而言，本章从分布式算力感知与调度概念和特征入手，进一步分析分布式算力感知与调度的研究意义和各种需求，并提出分布式算力感知与调度的发展目标。

1.1 分布式算力感知与调度概念和特征

随着全球数字化浪潮的到来，5G、AI、大数据等新技术不断涌现，算力已成为驱动社会进步的核心生产力。随着人工智能、物联网、元宇宙等技术的爆炸式发展，传统的集中式算力计算模式面对如此庞大且多样化需求，已经难以有效应对。分布式算力感知与调度技术应运而生，成为应对海量、泛在、实时计算需求的关键基础设施。这一理念旨在构建一个能够动态感知全网算力资源，并根据任务需求进行智

能化、自动化、最优化调度的新型信息基础设施，降低计算延迟与成本，支撑新型智能化应用的落地。

分布式算力是相对于传统集中式算力（如单一超级数据中心）而言的算力部署与利用模式，其核心是将一个大的计算任务分解成若干个小任务，然后把这些小任务分配给地理、网络层级或逻辑上相互独立的多个节点。这些计算节点可涵盖数据中心、边缘设备（如基站、物联网网关）、终端设备甚至个人闲置设备等，通过网络连接形成协同体系，实现算力资源的分布式协同与高效利用。分布式算力并非单一形态，边缘算力是前者重要组成部分，是分布式思想的一种具体体现。边缘算力强调“地理近端性”，即计算能力的部署靠近数据源，以满足低延迟和高实时性的需求；而分布式算力更关注“全局最优性”，侧重任务的分解与协同，以处理大规模和复杂的计算任务可能调度至边缘、核心云或两者协同，例如“云-边-端”分层推理。

分布式算力感知与调度的核心在于“感知”与“调度”两个相互依存、紧密结合的环节。“感知”是基础和前提，它指的是系统具备全面、实时、精准地获取分布式网络中各个计算节点资源状态的能力。具体而言，感知过程涵盖多维度：首先，系统需自动发现并注册新计算节点，识别其 CPU、GPU、FPGA 等类型，以及内存、存储容量等基础属性。其次，通过轻量探针或节点遥测数据，实时监控 CPU/GPU 利用率、存储占用率、网络带宽与延迟、负载水平、功耗及环境温度等核心指标。更关键的是，感知需深入能力评估，如量化计算单元的理论峰值性能（如 FLOPS）及对特定负载的实际加速效能。网络感知需

精确测量任务提交点、计算节点间的拓扑关系、带宽、延迟、丢包率及抖动，以保障低延迟应用。此外，还需感知资源使用的经济成本、能源成本及数据主权、SLA 等策略性约束。这些信息经清洗、融合与抽象后，将形成支撑智能决策的多维度量化算力资源模型。

“调度”则是基于“感知”结果所采取的行动，是整个系统的“大脑”和中枢。它根据感知到的全网算力资源分布图景和实时状态，在复杂约束条件下，通过智能高效的算法，将计算任务合理地分配到最合适的节点上执行，从而实现全局最优的资源利用率、最低的运营成本和最佳的用户体验。调度决策是一个高度复杂的优化问题，其目标函数通常是多维度的，需要在性能目标、经济目标和系统目标之间寻求最佳平衡点。分布式调度策略多种多样，从传统的基于静态规则的调度，如轮询、随机分配等，到更为复杂的动态调度策略，如基于负载均衡的调度、基于服务质量（QoS，Quality of Service）的调度、基于经济效益的调度等。现代的算力调度系统越来越倾向于采用人工智能和机器学习技术，通过对历史数据的学习和对未来负载的预测，实现预测性、主动性的智能调度，甚至能够做到“算力路由”，即像网络路由一样，为计算任务规划出一条从数据源到最优计算节点、再到结果返回的最佳路径。

分布式算力感知与调度具有如下几个显著的特征：

- 异构性：算力节点的硬件类型、操作系统、网络协议存在显著差异，系统必须能够充分识别并利用这种异构性，将不同类型的计算任务精准匹配到最适合的硬件上执行，从而实现

整体计算效能的最大化。感知系统需通过统一的“算力单位”实现异构资源的归一化描述；调度系统则需针对不同类型任务设计适应性的分配策略。

- 动态性：分布式环境本质上充满不确定性。资源可能随时加入、离开、发生故障或性能波动；网络状况瞬息万变；任务需求和依赖关系也可能动态调整。因此，算力感知必须是实时的，调度决策也必须是动态调整的。系统需要具备快速响应变化的能力，在某个节点出现故障时，能够迅速将其上的任务迁移至其他健康节点，保证业务的连续性；在检测到网络拥塞时，能够智能地选择其他通信路径，避免性能瓶颈。这种动态适应能力是保障系统稳定性和可靠性的关键。
- 跨域协同与互操作性：理想的分布式算力池往往跨越不同管理域（多个公有云、私有云、边缘站点、终端设备）。实现高效的感知与调度，必须解决跨域资源发现、认证授权、状态信息交换、任务协同执行等挑战。这依赖于开放的 API 标准、通用的资源描述语言、安全的跨域通信机制以及可能的跨域调度协调器或联邦学习机制。
- 能耗与可持续性感知：随着“双碳”目标的推进，算力调度的绿色属性愈发重要。感知需纳入能耗与碳足迹的实时监测；调度决策则需将能耗和碳排放作为重要优化目标或约束条件，例如优先将任务调度到使用可再生能源的数据中心或能效比更高的节点，或利用电价谷值进行计算，实现“绿色调度”。

分布式算力感知与调度是现代计算范式的核心支柱。它通过构建全域资源认知神经网和智能调度决策中枢，实现了对泛在、异构、动态算力资源的有效整合与按需供给。其核心在于全局化资源视图、多目标动态优化、高度环境适应、跨域无缝协同、智能学习进化以及对可持续性的深度关切。随着算力网络（CPN，Computing Power Network）概念的兴起和“东数西算”等国家级工程的推进，分布式算力感知与调度技术将持续演进，其智能化、自动化、绿色化水平将不断提升，为构建高效、敏捷、普惠、可持续的下一代数字基础设施提供核心动能，赋能千行百业的数字化转型与智能化升级。

1.2 分布式算力感知与调度研究意义

在数字化浪潮席卷全球的今天，算力已不再是单纯的技术指标，而是驱动社会经济形态深刻变革、与热力、电力并驾齐驱的关键生产力，是支撑数字经济高质量发展的战略基石。中国信通院指出，随着新一代通信规模建设和边缘计算应用的持续部署，越来越多的应用运行和数据生产处理在边端侧开展，这对于传统算力基础设施的部署、调度提出了新要求，分布式算力通过不同范围部署不同规模算力，为政企数智化转型各场景随需获取算力提供新思路。在此背景下，分布式算力感知与调度技术作为构建下一代算力基础设施的核心神经系统，其研究意义已远超单纯的技术优化范畴，上升至关乎国家数字竞争力、产业革命性变革以及社会可持续发展的战略高度。

开展分布式算力感知与调度的研究，是服务于国家发展战略、保

障数字主权的迫切需要。当前，算力已成为大国博弈的新焦点，构建自主可控、高效协同的算力体系是提升国家核心竞争力的关键。国家发改委等部门推动的“东数西算”工程，核心目标并非简单的“数据西迁”，而是构建全国一体化的算力网络体系。推动该技术发展，能将地理上广域分布、架构上高度异构、权责上分属多域的海量算力资源，通过智能化感知与调度，整合成逻辑统一、弹性敏捷、安全可控的国家级“算力资源池”。这不仅从根本上解决我国东西部算力供需不平衡的结构性矛盾，更能通过统一调度形成规模效应，为国家重大科研项目、经济社会发展提供澎湃且经济的算力支撑，从而在全球数字竞争中掌握战略主动权。

同时，这也是激活数据要素价值、推动产业智能化转型、催生新质生产力的核心技术引擎。算力作为数字经济的“发动机”，其渗透力决定产业升级的深度和广度。在前沿科学探索领域，如基因测序、新药研发、宇宙模拟等，分布式调度能汇聚全球顶级计算资源，为复杂科学问题求解提供前所未有的算力规模与效率。对于人工智能产业，尤其是大模型的训练与推理，异构算力调度可将计算任务精细化拆解，精准匹配到最高效的处理单元，最大化计算效率，加速 AI 在各行业的研发与应用。在工业互联网、智慧城市、自动驾驶等实体经济领域，实时感知能力与低延迟调度决策是支撑高级应用落地的关键。而开放共享的算力服务平台，能降低中小企业获取先进算力的门槛，激发全社会创新活力，为新产业、新业态、新模式提供沃土。

深入研究分布式算力感知与调度，核心价值在于推动构建集智能

高效、安全可信、绿色低碳于一体的下一代算力基础设施新范式，三者互为表里，不可分割。以智能为核心，利用人工智能算法赋能调度决策，通过全面、精准、实时感知全网状态，做出全局最优资源匹配决策，最大化资源利用率。以安全为基石，系统能实时感知网络攻击、节点故障等异常，智能进行任务迁移等操作保障业务连续性，同时确保敏感数据在可信域内流转，构筑数据安全屏障。以绿色为目标，将“绿色低碳”作为核心优化目标，与智能、安全深度融合。智能调度系统把能耗与碳足迹作为核心调度因子，感知各数据中心实时情况，智能分配计算任务，实现“算力调度”与“能源调度”协同，降低运营成本和碳排放，落实“东数西算”绿色发展理念。

综上所述，分布式算力感知与调度的研究具有深远的战略意义和广泛的应用价值，它上承国家发展战略，下接各行业数字化需求，内含技术创新驱动，外显绿色发展理念。随着“东数西算”工程的深入推进、算力网络概念的兴起以及人工智能应用的持续爆发，该领域的研究将不断深化，向着更高程度的智能化、更精细的感知、更强大的跨域协同能力、更强的安全可信保障以及更优的绿色效能演进。它不仅是解决当前算力供需矛盾的有效途径，更是塑造未来数字社会形态、驱动经济高质量发展、提升国家综合实力的关键所在。

1.3 需求分析

当下通信与算力的多样化以及算力资源分布式特性与多类型应用场景的深度耦合极大地推动了分布式算力感知与调度的产生，本节

将从国家战略需求、产业发展需求、技术演进需求、用户需求和功能需求五个方面进行分析。

1.3.1 国家战略需求分析

分布式算力感知与调度系统的发展深度融入国家战略布局，是支撑新型基础设施建设、数字经济转型及算力资源优化配置的关键技术支撑。从国家战略层面看，其建设需求源于多个维度的政策导向与实际发展需求。国家“十四五”规划将算力基础设施纳入新型基础设施重点领域，明确提出构建“全国一体化算力网络”，要求推动“中心-边缘”协同计算体系落地，《关于进一步深化电信基础设施共建共享的实施意见》中要求“促进基础设施智能化升级”，而分布式算力感知与调度系统通过整合边缘节点算力资源，能够弥补集中式数据中心在地理覆盖和低时延响应上的短板，形成“云边协同”的算力供给体系。例如，中国铁塔拥有的 210 万站址资源，通过统一的感知与调度机制，可转化为支撑“双千兆”网络、工业互联网、车联网等新型基础设施的底层算力支撑。

在数字经济与产业转型领域，国家《工业互联网创新发展行动计划》中强调“推动边缘计算与工业场景融合”，要求边缘节点具备实时数据处理、算力动态调度能力。分布式算力感知与调度系统恰好满足这一需求，其能够支撑工业互联网中设备互联的低时延算力需求，如智能制造中的实时控制；满足车联网中的路侧算力协同，如自动驾驶数据的本地处理；还能支撑智慧城市中的泛在感知计算，如视频监

控的实时分析，为产业数字化转型提供坚实的算力调度保障。

“东数西算”工程提出“优化算力资源空间布局”，要求边缘算力节点与枢纽节点协同联动，分布式算力感知与调度系统通过对全国边缘算力的统一度量和动态调度，可实现算力资源的跨区域优化配置：在东部经济发达地区，通过边缘节点分担中心算力压力，降低网络拥塞；在中西部地区，通过算力调度激活存量资源，支撑区域数字经济发展，直接响应《全国一体化大数据中心协同创新体系算力枢纽实施方案》中“提升算力资源利用效率”的目标。

此外，国家高度重视关键技术自主可控。分布式算力感知与调度系统的研发部署，可推动边缘计算领域的技术标准化，如算力度量、调度策略的统一规范；促进国产化软硬件适配，如基于鲲鹏、昇腾芯片的边缘服务器应用；并整合运营商、设备商、行业用户等多方资源，培育自主可控的边缘算力产业生态，打破国外技术垄断。在应急与公共服务领域，国家要求算力资源具备“全域覆盖、快速响应”能力，该系统依托广泛分布的边缘节点，可在自然灾害、重大活动保障等场景下，快速调度就近算力资源，支撑应急通信、视频会商、数据汇聚等服务，响应《国家应急通信保障预案》中“构建分布式应急算力支撑体系”的战略需求。

1.3.2 产业发展需求分析

随着数字经济的深度渗透，各行业对算力的需求呈现出“泛在化、低时延、差异化”的特征，分布式算力感知与调度系统成为支撑产业

升级的核心技术纽带。从产业实践来看，其需求主要体现在边缘算力资源的高效利用、跨行业应用的适配支撑以及产业生态的协同构建三个层面。

在算力资源利用层面，当前边缘算力基础设施呈现“规模庞大但分散异构”的特点。以中国铁塔为例，其 210 万站址资源和超 100 万机房资源广泛分布于全国，但由于资源零散、管理分散、网络接入方式多样（如 4G/5G、企业宽带、园区 NAT 网络等），导致算力资源利用率不足、调度效率低下。产业界迫切需要通过统一的感知与调度系统，实现异构资源的抽象建模与池化管理，例如通过标准化算力度量体系（涵盖 CPU、GPU、内存、网络等指标），将分散的边缘节点转化为可统一调度的“虚拟算力池”，提升资源利用效率。同时，边缘算力节点的“弱网、单通”等网络特性，也要求系统具备离线自治、断点续传等能力，以适应产业现场的复杂网络环境。

在跨行业应用支撑层面，不同行业对边缘算力的需求差异显著，推动调度系统向“场景化、定制化”方向发展。例如，工业互联网场景中，智能制造需要边缘节点提供毫秒级实时算力支撑，用于设备状态监测与实时控制；车联网场景中，路侧边缘节点需协同车辆终端，实现低时延的数据处理与协同决策，保障自动驾驶安全；智慧城市场景中，边缘算力需支撑视频监控、环境监测等泛在感知应用，要求系统具备高并发处理能力。此外，AI 训练推理、泛在数据采集等新兴场景，进一步要求调度系统能够根据业务需求动态匹配算力资源，例如为 AI 推理任务优先调度 GPU 资源，为数据采集任务优化网络带宽

分配。这些需求推动产业界从“通用算力调度”向“场景化算力服务”转型，而分布式感知与调度系统正是实现这一转型的核心载体。

在产业生态构建层面，边缘算力的商业化运营需要打通“供给-需求-交易“全链条，这依赖于开放、协同的调度体系。当前，边缘算力的供给方包括电信运营商、铁塔公司、第三方数据中心等，需求方涵盖政府、企业、社会公众等，各方亟需通过标准化的调度接口实现资源互通与业务协同。例如，铁塔边缘算力资源可通过调度系统接入公共算力交易平台，向企业提供按需付费的算力服务；同时，系统需支持第三方调度算法的灵活接入，满足不同行业的定制化需求。这种开放生态的构建，不仅能降低算力交易成本，还能促进边缘算力在电子政务、企业数字化、新兴业态等领域的规模化应用，推动产业从“硬件堆砌”向“服务增值”升级。

综上，分布式算力感知与调度系统的发展，既是解决当前边缘算力资源分散、利用低效等问题的技术手段，也是支撑各行业数字化转型、构建边缘算力产业生态的必然需求，其产业价值已成为推动数字经济高质量发展的重要引擎。

1.3.3 技术演进需求分析

在数字经济蓬勃发展以及各行业数字化转型持续深化的当下，分布式算力感知与调度系统的技术演进已是大势所趋，旨在契合日益繁杂的业务需求，从容应对激烈的市场竞争。

从硬件基础层面来看，算力基础设施朝着异构化与分布式方向加

速迈进。不同类型的计算芯片，诸如 CPU、GPU、FPGA 以及新兴的 ASIC 芯片等，在性能、功耗以及适用场景等方面呈现出显著差异，共同构建起复杂的异构计算环境。与此同时，计算节点的分布范围不断拓展，从传统的数据中心延伸至各类边缘计算节点，如基站、工厂、园区等。这一发展态势要求分布式算力感知与调度系统必须能够精准识别各类硬件资源的详细信息，包括处理器的型号、核心数、频率、缓存大小，内存的容量、类型、速度，存储设备的容量、读写速度、接口类型，以及 GPU 的型号、显存大小等。只有做到这些，系统才能够依据业务的具体需求，将任务合理且高效地分配至最适宜的硬件资源上，从而充分挖掘和发挥各类硬件的性能优势。举例来说，在处理大规模数据的并行计算任务时，GPU 能够凭借其强大的并行处理能力大幅提升运算速度；而在执行复杂逻辑运算与顺序指令时，CPU 则更具优势。因此，系统需要准确把握不同硬件的特性，实现任务的最优分配。

在软件与算法领域，相关技术同样处于快速迭代升级的进程中。随着深度学习模型规模与复杂度的与日俱增，模型训练与推理对于算力的需求呈现出爆发式增长。为有效缓解硬件算力的紧张压力，一系列模型优化技术应运而生，如模型压缩、量化、剪枝等。这些技术通过减少模型参数数量、降低数据精度等手段，在不明显影响模型性能的前提下，显著降低了计算量与存储需求。例如，借助模型压缩技术，部分深度学习模型的计算量能够大幅降低。与此同时，调度算法也在持续优化创新。传统基于规则的调度算法在面对复杂多变的网络环境、动态波动的业务负载以及多样化的硬件资源状态时，显得力不从心。

基于强化学习、机器学习等人工智能技术的智能调度算法顺势崛起，这些先进算法能够实时采集和分析网络延迟、节点负载、业务优先级等多维度数据，通过不断尝试不同的调度策略，并根据实际执行效果进行动态优化，最终探寻出最优的任务调度方案，进而大幅提升资源利用率与任务执行效率。

从应用场景的视角出发，不同行业对分布式算力感知与调度系统提出了丰富多样且标准日益严苛的要求。在工业互联网领域，为切实达成生产过程的实时控制与精细化优化，对算力的实时性、可靠性以及精准性提出了极高要求。生产线上的设备运行数据需要在毫秒级的极短时间内完成处理与深入分析，以便及时、精准地调整生产参数，确保产品质量的稳定以及生产效率的提升。在智能安防领域，随着视频监控分辨率的持续提高以及多模态感知技术的广泛应用，智能安防系统需要同时高效处理来自高清摄像头、红外传感器、声纹识别设备等多种设备的海量数据流，并实现实时的行为分析、异常事件的快速检测与精准目标追踪。这无疑要求系统具备强大的并行计算能力以及极低延迟的数据传输能力，以此保障安防应用的高效、稳定运行。在医疗影像领域，为实现医学影像的快速处理与精准诊断，需要分布式算力感知与调度系统能够有力支持大规模数据的快速传输与高效运算，同时严格满足医疗数据的隐私安全要求。例如，通过巧妙运用边缘计算与云计算的协同技术，将部分数据处理任务前置至边缘节点，既有效减少了数据传输延迟，又切实保障了数据安全。

分布式算力感知与调度系统的技术演进，需构建智能、安全、资

源高效协同的体系，以适配数字经济多元需求。智能调度作为核心引擎，依托机器学习，深挖计算节点的性能参数、负载趋势、业务适配性等多维数据，动态构建节点画像。同时，安全防护贯穿调度全流程，从节点接入时的身份认证，到数据传输加密、调度策略防篡改，构建多层次防护网。同时，结合智能调度与安全机制，让算力资源在安全流转中，高效支撑业务运行，实现智能调度精准匹配、安全防护全程护航、资源利用极致优化的协同发展，为数字经济筑牢坚实算力底座，也在技术迭代中响应绿色发展需求。

综上所述，分布式算力感知与调度系统正处于技术快速演进的关键转折点，面临着从硬件基础、软件算法到应用场景适配，乃至绿色低碳发展等多方面的严峻挑战与难得机遇。只有紧密追踪技术发展前沿趋势，持续不断地开展技术创新与优化升级工作，才能全方位满足国家战略、产业发展以及社会民生等多层面对于分布式算力的急切需求，为数字经济的高质量、可持续发展筑牢坚实的技术根基。

1.3.4 用户需求分析

分布式算力感知与调度系统的用户群体涵盖各级政府机构、全国性/区域性企业、社会公众及行业合作伙伴，其需求聚焦于算力资源的“可感知、可调度、可信赖”，并随业务场景的多样化呈现显著差异。

从政府与公共服务领域来看，用户核心需求集中在算力资源的统筹管理与安全可控。例如，电子政务场景中，各级政府需要通过系统

实现跨区域边缘算力的统一调度，支撑政务数据本地化处理（如身份证核验、社保信息查询），要求满足低时延、高可靠及数据隐私保护需求。此外，应急指挥、智慧城市等场景需系统具备快速响应能力，在突发事件中可动态扩容边缘算力，保障视频会商、实时监控等服务的连续运行。

企业用户的需求则围绕业务效率与成本优化展开。工业企业通过系统调度厂区边缘节点算力，支撑设备实时监控与工艺优化，要求算力调度响应时间极低，以满足智能制造的实时性要求；互联网企业（如短视频、直播平台）需利用边缘算力降低内容分发时延，要求系统支持动态调整算力节点分布，匹配用户访问热点的迁移。同时，企业普遍关注算力资源的可视化管理，需系统提供资源使用率、成本消耗等数据报表，辅助决策。

社会公众作为终端用户，其需求体现在算力服务的“无感可用”。例如，车联网场景中，驾驶员通过车载终端获取实时路况分析，依赖边缘算力的低时延响应，而系统需自动调度路侧节点算力，保障服务连续性；智能家居场景则要求系统按需分配算力，支撑语音识别、安防监控等轻量级业务，同时控制终端能耗。

合作伙伴（如第三方算力提供商、应用开发商）的需求聚焦于系统的开放性与兼容性。第三方算力提供商需通过标准化接口接入系统，实现资源互通与收益分成；应用开发商则要求系统支持多类型应用部署（如容器化、虚拟机化），并提供灵活的调度策略接口，适配不同算法对算力的差异化需求（如 AI 推理需 GPU 资源优先调度）。

此外，所有用户群体均对系统提出共性需求：一是弱网环境适配，在 4G/5G 信号不稳定区域（如偏远地区、地下停车场）仍能保障算力服务可用；二是安全防护，需具备数据传输加密、节点身份认证等能力，防止算力资源被非法占用或数据泄露；三是低成本运维，通过自动化部署、远程监控功能降低人工干预，尤其适合边缘节点分散的场景。

综上，用户需求呈现“分层化、场景化、个性化”特征，分布式算力感知与调度系统需通过模块化设计、灵活的策略配置及开放接口，满足不同用户在功能、性能、安全等维度的多样化要求。

1.3.5 功能需求分析

分布式算力感知与调度系统的功能需求围绕算力资源的全面感知、精准调度、高效协同及可靠运维展开，旨在解决当前算力资源分布不均、利用率低、协同困难等问题，满足不同行业对算力的多样化需求，实现“一点接入、即取即用”的算力服务目标。

算力感知功能：系统需具备对各类异构算力资源（CPU、GPU、FPGA、ASIC 等）的实时感知能力，包括硬件配置（核心数、主频、显存大小等）、负载状态（利用率、任务队列长度）、能耗指标等。通过标准化接口（如 Telemetry 协议）及轻量化采集代理，实现资源信息的秒级采集与上报，为调度决策提供数据基础。同时，系统应支持对网络资源（带宽、时延、丢包率）的动态监测，通过带内网络遥测（INT）、主动探测（IPP/IFIT）与被动分析（sFlow/IPFIX）等技术，构建“资

源-网络”协同视图，保障任务执行的网络质量。

算力路由功能：在大规模分布式算力资源的寻址过程中，为避免传统网络路由机制对计算节点实时负载、任务处理能力等关键算力状态参数考量的忽视，而形成与算力资源状态割裂的寻址模式，需构建算力与网络深度融合的新型路由体系，构建分布式算力路由这一创新的网络-计算协同调度范式，通过在传统 IP 路由架构中融入“服务标识”、“算力资源状态”和“算网多因子选路算法”三大核心要素，实现网络路径与算力资源的联合优化调度框架，以及网络寻址方式的根本性变革，在保障网络稳定性的同时实现了算网资源的协同优化。

算力调度功能：基于感知数据，系统需实现智能、灵活的算力调度。一方面，支持多维度调度策略，如计算优先、网络优先、成本优先等，以满足不同业务对算力、网络的差异化需求。例如，AI 推理任务可优先调度 GPU 资源，实时性业务（如自动驾驶、云游戏）则侧重网络时延优化。另一方面，调度算法应具备自适应能力，根据资源动态变化及业务负载波动，动态调整调度策略，提升资源利用率与任务执行效率。此外，系统需支持任务的跨节点、跨区域调度，实现“东数西算”“东数西渲“等跨域协同，通过算力路由协议将任务精准匹配至最优算力节点。

资源管理功能：对分布式算力资源进行统一管理，涵盖资源注册、注销、状态监控、故障诊断等全生命周期管理。通过资源虚拟化与池化技术，将分散的物理资源整合为逻辑资源池，实现资源的灵活分配与弹性扩展。例如，利用 GPU 虚拟化技术（MIG、vGPU）将单块 GPU

切分为多个虚拟实例，供不同任务共享使用；通过智能算力池化，对 CPU、GPU 等资源进行统一调度，降低资源碎片化，提升资源整体利用率。

业务适配功能：系统需具备良好的业务适配能力，支持多样化应用的快速部署与运行。通过容器化（Docker、Kubernetes）、虚拟机（VM）等技术，实现应用的隔离与高效运行。同时，提供丰富的 API 接口与开发工具，方便第三方应用接入与定制化开发，满足不同行业（工业、医疗、金融等）对算力服务的个性化需求。例如，工业互联网应用可通过 API 获取实时算力资源状态，动态调整生产任务；医疗影像处理应用可利用开发工具优化算法，适配系统算力特性。

安全可信功能：鉴于算力资源的重要性与敏感性，系统需构建全方位安全防护体系。在数据安全方面，支持数据传输加密（SSL/TLS）、存储加密（AES），防止数据泄露；在身份认证与访问控制方面，采用多因子认证、RBAC 权限模型，确保只有授权用户可访问与调度算力资源；在安全审计方面，对所有操作进行日志记录与审计，实现操作可追溯；此外，通过区块链技术保障算力交易的可信任性与透明度，防止算力资源被非法占用或滥用。

综上所述，分布式算力感知与调度系统的功能需求紧密围绕算力资源的全生命周期管理，通过技术创新与功能优化，为数字经济发展提供坚实的算力支撑，推动算力资源的高效利用与广泛普及。

1.4 分布式算力感知与调度发展目标

在数字经济蓬勃发展、数据量呈指数级增长的当下，分布式算力感知与调度技术的重要性愈发凸显，其发展目标涵盖了体系构建、技术突破、场景适配以及生态营造等多个关键维度，致力于打造一个高效、智能、安全且开放的分布式算力服务网络，如图 1-1 所示，从算力度量、调度引擎、跨域协同、安全机制、效能优化五方面推进，最终集成算力服务网络。

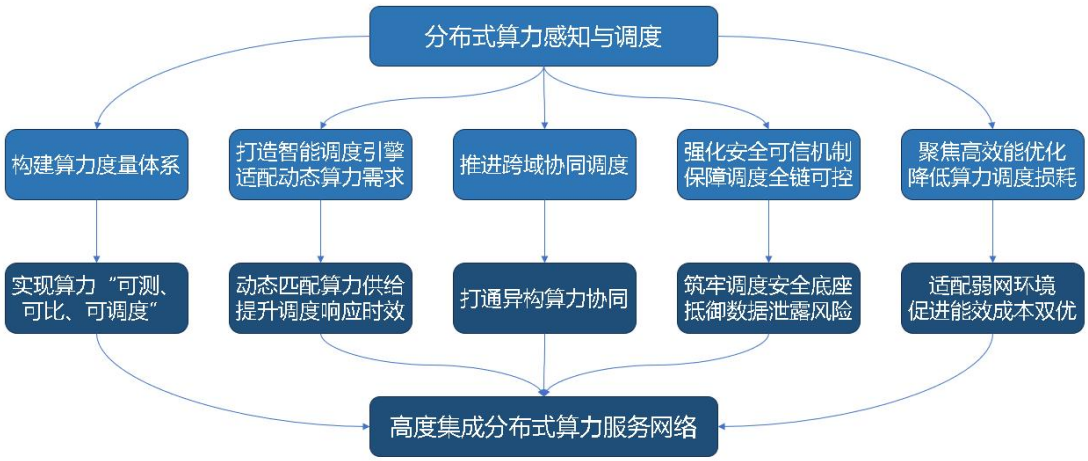

图 1-1 分布式算力感知与调度发展目标图

构建统一、标准的算力度量与管理体系是首要目标。当下，算力资源呈现出显著的异构性，CPU、GPU、NPU 等多元算力单元在性能、应用场景等方面各有千秋。这就迫切需要建立一套全面且精准的资源建模、性能建模以及服务能力建模体系。在资源建模中，不仅要对各类硬件的基础参数，如 CPU 的核心数、主频，GPU 的显存容量、带宽等进行细致梳理，还要考虑硬件的架构特性与兼容性，实现算力“可

测、可比、可调度”。性能建模则需综合考量算力在不同负载、不同应用场景下的实际表现，例如在复杂图形渲染时 GPU 的帧率稳定性，在大规模数据运算中 CPU 的计算精度与速度。而服务能力建模要涵盖从算力的交付效率到运维保障能力等多方面因素，确保对算力资源实现全方位、多层次的量化描述与精准评估，让不同类型、处于不同场景下的算力资源，都能基于这套体系具备“可测、可比、可调度”的基础条件，为后续的高效管理与合理调配奠定坚实根基。

与此同时，为了契合大规模分布式节点的复杂特性，必须构建起跨层级的协同管理架构。从集团级的统筹规划，到省、市级的协调执行，再到区县级的具体落实，形成“集团-省-市-区县”这样一套严密且灵活的分级分域感知与调度体系，推进跨域协同调度，打通异构算力协同。该体系既要确保各层级对本地边缘站址资源实现实时监控，掌握诸如资源的实时负载、运行状态等关键信息，又要能够依据全局资源的动态变化进行动态调配。当某一区域因突发业务需求导致算力紧张时，上级层级可迅速协调周边区域的闲置算力资源进行支援，实现资源的高效利用与协同优化，既保证各域在一定程度上的自主性，以应对本地的特殊情况，又能从整体上保障资源调配的科学性与合理性。

在技术攻坚层面，分布式算力感知与调度的发展目标聚焦于突破异构网络与复杂环境下的重重瓶颈。边缘节点的网络接入状况极为复杂，涵盖了互联网专线、企业宽带、4G/5G 移动通信网络以及园区 NAT 网络等多种类型。不同网络在带宽、时延、稳定性等方面差异巨大，

这给算力信息的及时准确传递与调度指令的有效下达带来了极大挑战。因此，研发自适应的感知与通信机制迫在眉睫。通过对各类网络协议进行深入研究与优化，构建能够在不同网络间无缝切换、智能适配的通信体系，解决弱网环境下双向访问难题，保障算力信息的实时通告与调度指令的高效传达，确保即便在网络状况不佳、波动频繁的情况下，分布式算力系统依然能够稳定运行，维持服务的连续性与可靠性并且，要构建起智能调度引擎，这一引擎需融合网络延迟、算力位置、资源负载等多因子算法。

在网络延迟方面，精确测算数据在不同链路、不同节点间传输所需的时间，结合实时网络拥塞状况，动态调整数据传输路径；考虑算力位置时，充分权衡物理距离与网络拓扑结构，优先选择距离近且网络连接质量优的算力节点，降低传输损耗；而资源负载的监控与分析，则能让调度引擎知晓各算力节点当前的工作饱和度，避免将任务过度集中于高负载节点，实现“路径+节点”的联合优化。如此一来，业务请求便能精准匹配到最合适的算力节点，极大提升资源利用率，显著缩短业务响应时间，为用户提供更为流畅、高效的服务体验。

在场景适配与服务能力提升方面，分布式算力感知与调度旨在实现对多元业务的深度、精准支撑。以政企领域为例，电子政务涉及大量数据的安全处理与高效流转，企业业务则因行业特性、业务规模的不同，在算力需求上呈现出多样化特点。车联网场景中，自动驾驶对实时性要求极高，车辆行驶过程中的决策需在极短时间内完成，这就要求分布式算力系统能够提供低时延的算力支持，端到端时延需严格

控制在极短范围内，保障行车安全；工业互联网领域，生产过程的连续性与稳定性至关重要，设备控制、实时数据分析等业务不容有丝毫差错，对算力的可靠性提出了严苛要求，任何算力故障都可能导致生产线停滞，造成巨大损失。而在 AI 训练推理场景中，面对海量数据与复杂算法，需要适配异构算力资源，充分发挥 CPU、GPU、NPU 等不同芯片的优势，加速模型训练与推理过程。通过制定灵活多变的调度策略，结合资源的动态扩缩容机制，无论业务需求如何波动，都能确保各类业务获得稳定、充足的算力支撑。与此同时，积极推动算力资源向公共服务属性拓展，搭建算力交易平台，完善交易规则与流程，让算力如同水电一般，用户可根据自身实际需求，便捷地获取相应算力资源，真正实现算力的按需使用与灵活交易。

分布式算力感知与调度的长远发展目标还包括构建一个开放、安全的生态体系。在开放性方面，通过标准化接口设计，打造一个兼容第三方调度算法与插件的平台，吸引云计算厂商、行业解决方案提供商、科研机构等产业链各方积极参与。不同主体可基于自身优势，开发各具特色的调度算法与应用插件，丰富分布式算力系统的功能与应用场景，形成一个充满活力、互利共赢的算力资源共建共享格局。在安全保障上，建立起全流程的安全防护机制。从节点接入阶段的严格身份认证，确保只有合法、可信的节点能够进入分布式算力网络；到数据传输过程中的加密处理，运用先进的加密算法，保障数据在传输过程中的保密性、完整性与可用性，防止数据被窃取或篡改；再到调度日志的全程追溯，借助区块链等技术，详细记录每一次调度操作的

相关信息，一旦出现问题，可快速溯源，查明原因。通过这样全方位、多层次的安全保障措施，确保算力调度的安全性与合规性，满足诸如金融、医疗等对数据安全极为敏感行业的严苛要求，为分布式算力网络的大规模、高可靠性应用筑牢安全防线。

通过在这些方面持续发力，分布式算力感知与调度技术将逐步实现从基础能力构建到深度场景应用，从单一技术突破到生态体系完善的全面跨越，最终构建起一个强大、高效、智能且安全的分布式算力服务网络，成为推动数字经济发展、支撑社会数字化转型的核心基础设施。

二、分布式算力感知与调度架构

为应对算力资源日益呈现分布化、异构化的发展趋势，亟需构建面向多源异构算力的高效协同与智能调度能力。为此，提出基于“分层分域”设计理念的分布式算力感知与调度系统架构。该架构由基础设施层、网关管理层、算力管控层、级联控制层、安全保障层和应用服务层六大功能层构成，分别承担资源接入、状态感知、统一管控、跨域协同、安全隔离与业务支撑等关键任务。各层协同联动，支撑算力资源的泛在接入、智能编排与可信运行，构建统一、智能、可扩展的分布式算力底座，全面赋能多行业、多场景的数字化和智能化转型。如图 2-1 所示，本章将对该分布式算力感知与调度系统架构进行详细设计与分析。

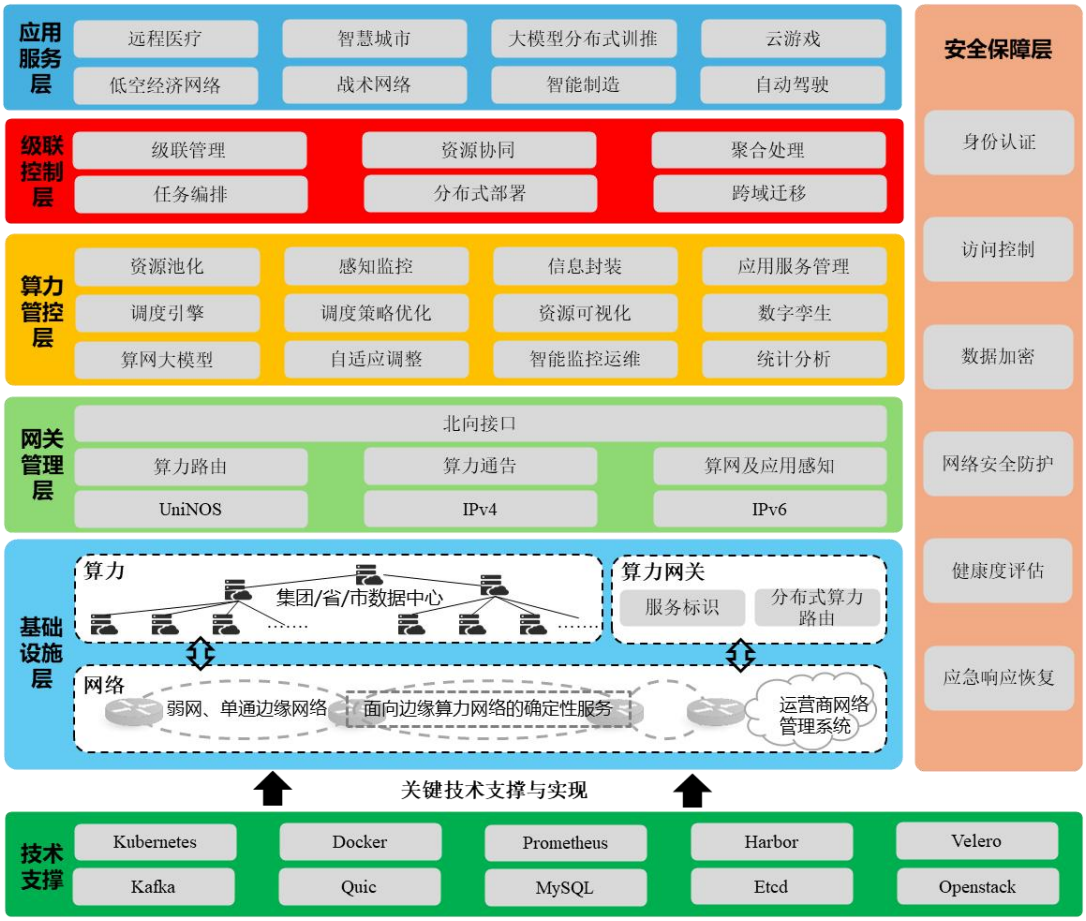

图 2-1 分布式算力感知与调度架构图

（1）基础设施层

基础设施层作为分布式算力感知与调度体系的资源支撑底座，面向多源异构算力与多样化网络环境，提供统一接入、抽象管理与弹性供给能力。该层涵盖从边缘到核心的数据中心资源，包括适配弱网环境的边缘节点、具备单向通信能力的安全隔离网络、广覆盖的运营商通信网络、集团、省、市各级数据中心与算力网关，支撑资源的分级部署、异构接入与高弹性调度，构建覆盖广泛、能力多样的算力资源基础。

（2）网关管理层

网关管理层负责实现对多类型接入网络的统一接入控制、通信协议解析与运行状态采集，该层支持多协议栈与南向接口的灵活适配，涵盖 IPv4、IPv6、UniNOS 等网络协议，具备算力路由、算力通告、算网及应用状态感知等关键功能。通过对接入链路与设备运行状态的精准感知，实现异构算网资源的高可靠接入，为上层算力管控与服务编排提供实时、可信的网络态势支撑。

（3）算力管控层

算力管控层是分布式算力系统的核心控制中枢，聚焦异构算力资源进行统一池化、智能调度与全局编排，实现资源“从分散部署到统一编排”的能力跃迁。该层涵盖资源池化、感知监控、信息封装、应用服务管理、调度引擎、调度策略优化、资源可视化、数字孪生建模、算网大模型支持、自适应资源调整、智能监控运维、统计分析等功能。通过构建跨平台、跨设备的协同机制，打通资源边界，提升资源的弹性供给与高效利用能力，实现算力按需调度与动态匹配，从而为上层多样化应用提供稳定、高性能、可持续的算力支撑。

（4）级联控制层

级联控制层作为跨域协同与系统全局优化的关键枢纽，负责实现多集群、多算域间、多层级的资源协同与联动调度。包括级联管理、资源协同、聚合处理、任务编排、分布式部署、跨域迁移功能，可将上层调度指令下发至各域控制单元，进行各域算力池状态实时协商与协调分配与聚合处理，完成在不同地域、不同算力域内灵活编排与布署，实现运行中任务和数据的动态迁移与无缝切换。该层通过全局视

角与局部自治的有机结合，有效提升系统的调度效率、负载均衡能力与故障自愈性能。

（5）安全保障层

安全保障层构建全链路、纵深式的安全防护体系，覆盖从终端接入、数据传输到任务执行的各关键环节。包括份认证、访问控制、数据加密、网络安全保护、健康度评估、应急响应恢复等功能，支持多策略联动的风险识别与处置能力。有效提升在复杂动态环境下的系统稳定性、业务连续性与数据可信性。

（6）应用服务层

应用服务层聚焦典型算力应用场景，面向不同行业、不同形态的任务调度需求，提供灵活、高效的分布式计算服务能力。支持包括远程医疗、智慧城市、大模型分布式训练、云游戏、低空经济网络、战术网络、智能制造、自动驾驶等典型场景的分布式算力感知与调度，实现任务在多源异构环境中最优部署，全面提升任务执行效率、资源利用率与用户体验，助力多行业智能化转型。

三、分布式算力感知与调度关键技术

分布式算力感知与调度作为智能互联网基础能力体系中的核心支撑技术，面向大规模、异构化、多域协同的算力资源环境，致力于实现算力资源的全面感知、智能决策与高效调度。该技术面向算力泛在部署、需求多元涌现的发展趋势，突破传统集中式资源调度的局限，构建了具有层次性、自治性与协同性的资源感知与调度架构，可对分布在云、边、端不同层级、不同地域的算力节点进行精细化建模、动态化评估与灵活化编排，有效提升资源使用效率与业务响应能力。本章以典型业务场景为切入点，展示分布式算力感知与调度技术在赋能数字社会和智能产业中的关键价值。随着技术体系的持续演进和应用需求的不断深化，分布式算力感知与调度还将面临更多挑战与机遇，亟需产业、学界与研究机构协同推进，不断丰富关键技术体系，拓展应用广度与深度，持续释放算力价值。

3.1 分布式算力资源度量模型

为支撑分布式算力感知与调度系统的智能化与高效化运行，亟需构建统一的算力资源度量模型，面向分布式、异构、多域的复杂环境，对多类型算力资源实现标准化建模、特征提取与精准量化。该模型由分布式算力资源标识体系与多维资源度量指标体系两大核心模块组成，旨在实现算力资源的统一识别、动态管理，为多场景算力调度提

供基础支撑。

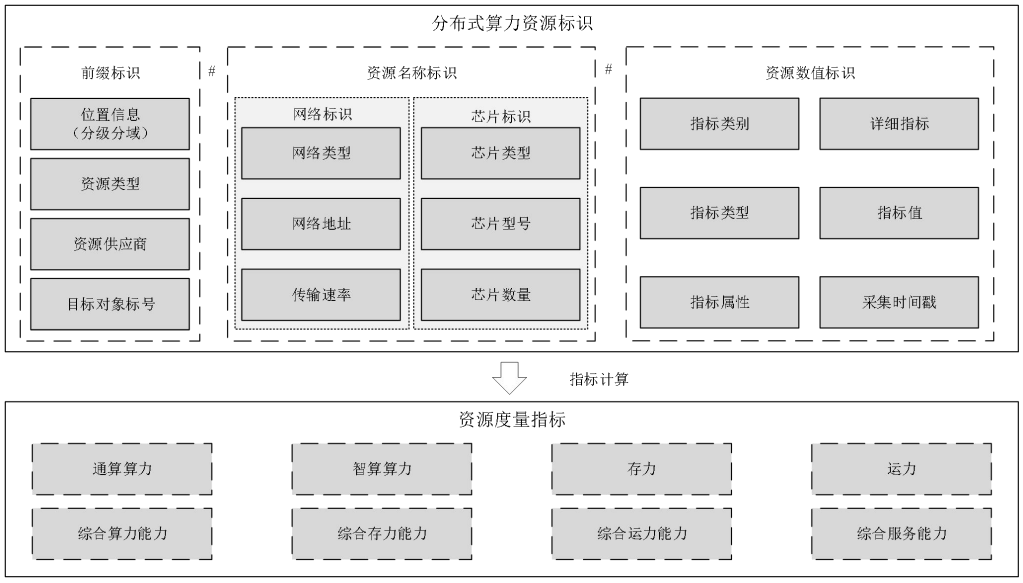

图 3-1 分布式算力资源度量架构图

针对边云协同、异构部署和多域融合等典型应用场景，算力资源需具备唯一性、可溯源性与可组合性标识能力。为此，设计统一的资源标识编码体系，构建由“前缀标识—资源名称标识—资源数值标识”组成的三级结构化模型，采用嵌套编码方式实现资源的多维抽象与唯一命名，支持跨域协同与资源调度。具体包括。

前缀标识：用于描述资源的唯一基本信息，涵盖位置信息、资源类型、供应商和目标对象标号等要素，明确了资源的来源、属性和用途；

资源名称标识：用于描述网络及芯片信息，包括网络类型、网络地址、传输速率、芯片类型、芯片型号和芯片数量；

资源数值标识：用于描述具体的度量指标信息，包括指标类别、指标类型、指标属性、详细指标、具体指标值，以及指标采集的时间

戳。

该标识体系与任务调度引擎深度集成，实现对资源的实时感知、快速匹配与统一管理，增强调度系统的智能响应能力。

在此基础上，为支撑面向任务的精确资源调度，需构建覆盖计算、存储、网络及能耗等多维度的统一资源度量指标体系。该体系通过结构化与标准化建模，系统定义各类异构资源的量化维度与评估方法，确保资源能力具备可比性、可预测性与可优化性，为分布式环境下的智能调度提供基础支撑。

在具体指标构建上，体系涵盖通用算力（基于 CPU 的处理能力）、智能算力（基于 GPU 的处理能力）、存力（服务器整体存储容量）、运力（站点公网带宽总和）等基础维度，并进一步提出综合能力指标，包括。

综合算力能力：融合通用算力（GFLOPS）、内存容量（GB）、智能算力（TFLOPS）及显存容量（GB），采用向量模方式标准化处理，反映节点整体计算能力；

综合存力能力：整合存储总容量、磁盘读/写 IOPS 等关键指标，表征数据存储性能与承载能力；

综合运力能力：结合带宽（Mbps）与时延（ms）构建通信性能度量，反映网络传输能力；

综合服务能力：以算力、存力与运力综合指标为基础，形成反映平台整体服务保障能力的统一度量体系。

通过统一的分布式资源度量模型构建，可有效提升异构、多域、

多类型资源的可观测性与可控性，为实现资源高效利用、任务智能调度和多场景适配提供核心能力支撑，助力泛在算力网络的持续演进与产业化应用拓展。

3.2 分级分域算力资源感知技术

分级分域算力资源感知技术是支撑算力网络高效运行的基础能力，是实现大规模异构、分布式、动态算力资源精准掌控与智能调度的前提条件。面对当前算网架构日益复杂、资源形态日益多元的演进趋势，传统的集中式、静态化感知机制在数据更新效率、系统扩展能力、感知精度等方面逐渐暴露瓶颈，难以满足多源异构资源的协同管理与实时调度需求。构建具有分层架构、域间协同、自适应更新能力的感知机制，已成为分布式算力调度体系亟需解决的核心问题。

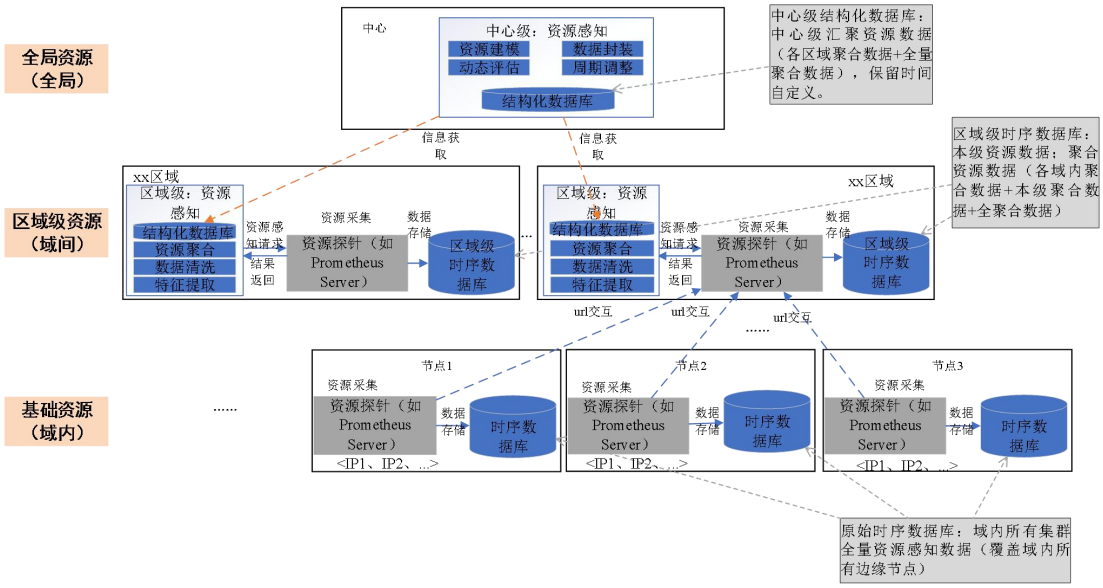

图 3-2 分级分域算力资源感知模型

该技术采用“域内自治、域间协同、全局融合”的三级感知架构，通过在边缘节点、区域集群与全局控制域之间建立分级感知通道，提升感知系统的可扩展性与实时性。一方面，在边缘域内部署轻量级资源探针，针对本地 CPU、GPU、内存、存储、I/O 等关键算力指标进行快速采集与预处理，并融合容器运行态、操作系统状态等多源信息，实现本地资源的快速建模与上报。另一方面，在区域级集群层引入聚合感知模块，支持多源数据的清洗、归一化与特征提取，对来自不同边缘节点的数据进行汇总分析，结合地理分布、负载状态、应用特征等，形成区域级资源态势视图，为局部调度与负载均衡提供策略依据。

在全局控制域，依托统一算力资源建模与标准化数据封装协议，实现不同区域感知数据的融合处理与时序对齐。通过引入动态权重机制与感知质量评估模型，实现对异常数据的剔除、对感知盲区的补偿；并结合自适应感知周期调整策略，根据资源负载波动、任务密度变化及调度优先级动态调节感知频率与深度，实现算力感知效率与系统负载之间的动态平衡，兼顾系统轻量性与感知精度。

整体上，分级分域算力感知技术具备高扩展性、高实时性与高鲁棒性，为异构算力资源的统一建模、精准度量与智能编排提供可靠的底层支撑，显著增强了跨域调度、弹性部署等算网能力，有效支撑算力网络在多场景、多区域、多维度下的广泛应用与持续演进。

3.3 分级分域算力调度技术

在分布式异构算力网络日益复杂的背景下，算力资源呈现出跨地

域、跨运营主体、异构架构与多样部署等多维度差异，资源状态动态变化频繁、服务需求类型日益多元，传统集中式调度方式已难以满足其对实时感知、高效响应与灵活协同的综合性要求，面临感知粒度粗、调度路径长、中心瓶颈明显等突出挑战。为应对算力网络规模持续扩展和服务形态加速演进的趋势，亟需构建具备高可扩展性、低延迟响应能力与全局—局部协同能力的分级分域算力调度技术体系。

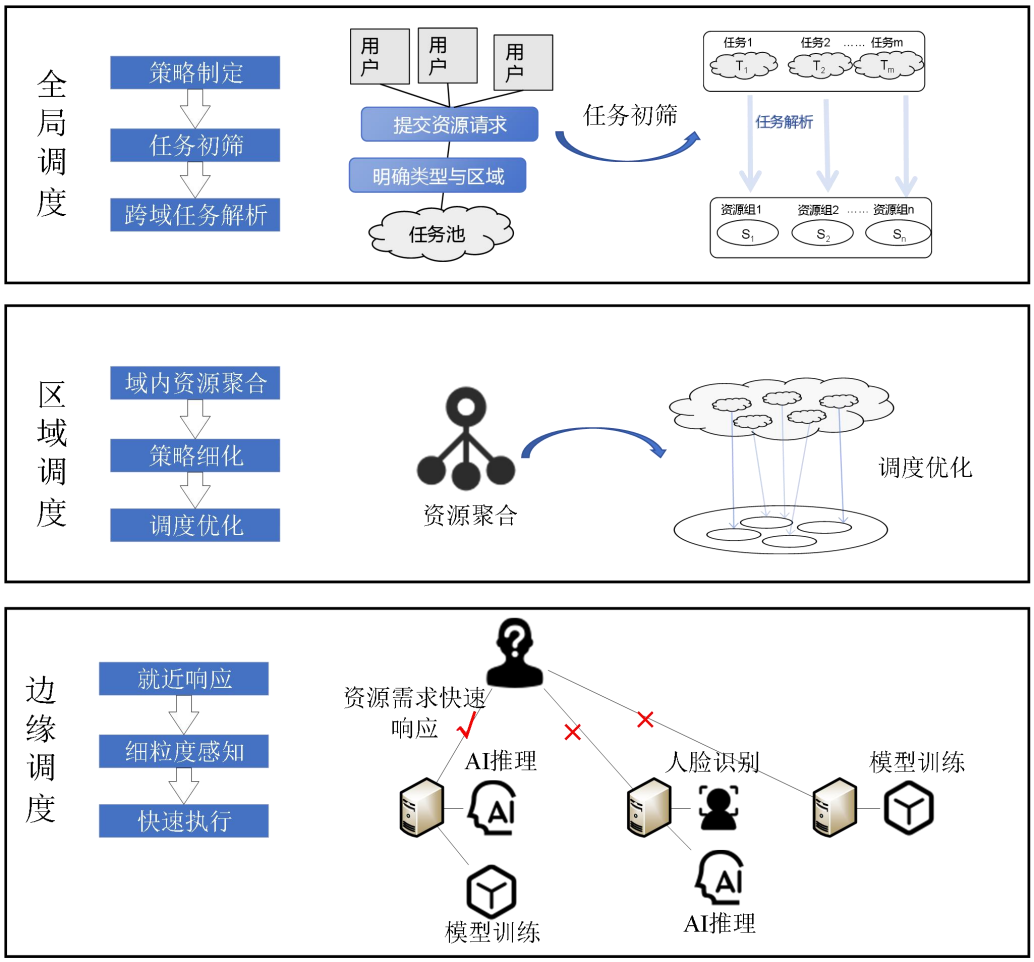

图 3-3 分级分域算力调度示意图

分级分域调度技术立足多层次算力系统架构，通过划分“全局—区域—边缘”三级调度域，形成一体化、自适应的调度控制体系，满足不同层级算力节点对资源管理与任务响应的差异化需求。其中，全

局调度中心聚焦于策略制定、任务初步筛选与跨域任务解析，承担整体资源视角下的调度决策与任务分派；区域调度中心面向本域内资源聚合、策略细化与调度优化，实现区域内的资源高效利用与跨域间负载均衡；边缘调度单元部署于算力节点侧，具备就近响应、细粒度感知与快速调度执行能力，能够在毫秒级内对本地突发请求作出决策，有效提升系统的服务敏捷性与抗扰鲁棒性。

该技术体系的核心在于调度粒度与路径的双重优化：一方面，构建多层级资源状态感知机制与任务上下文理解能力，支持感知粒度随业务 QoS 要求与网络状态灵活调整，实现自适应调度策略动态收敛；另一方面，通过分级调度路径分散中心调度压力，实现由上至下逐级下发与自治优化的协同路径，有效避免高频交互对中心节点的性能冲击，提升整体调度系统的可用性、实时性与稳定性。

此外，为进一步提升调度智能性与泛化能力，分级分域调度体系需融合资源异构性建模、任务优先级动态评估、多策略协同编排与调度意图表达等关键技术，构建统一建模、规则驱动与策略自适应相结合的调度框架。在策略层面，引入面向多业务类型与多网络状态的多维度决策模型，实现资源利用率、任务响应效率与服务质量保障之间的有机统一。

综上所述，分级分域算力调度技术面向分布式异构算力网络的演进趋势，既提升了算力资源调度的精准性、响应的敏捷性与体系的可扩展性，也为构建算力网络智能调度体系、支撑算网深度融合奠定了关键基础，是推动算力服务泛在可得、高效可控的核心技术路径之一。

3.4 分布式算力路由技术

在大规模分布式算力资源的寻址过程中，若采用传统网络路由机制，其决策依据通常局限于路由算法、可达性、路径跳数、传输时延、带宽利用率等网络层指标，而忽视了对计算节点实时负载、任务处理能力等关键算力状态参数的考量。这种与算力资源状态割裂的寻址模式，可能导致用户请求被导向网络指标优良但算力资源过载、服务响应迟缓的算力节点，进而引发分布式系统的性能塌陷效应，造成底层算力资源无法通过智能调度实现全局能效优化。

针对传统网络路由机制存在的算网割裂问题，亟需构建算力与网络深度融合的新型路由体系。在此背景下，分布式算力路由（Distributed Computing Force Routing, DCFR）作为一种创新的网络-计算协同调度范式应运而生。该技术通过在传统 IP 路由架构中融入"服务标识"、"算力资源状态"和"算网多因子选路算法"三大核心要素，实现了网络路径与算力资源的联合优化调度框架。

分布式算力路由技术通过构建基于服务标识的分布式多实例服务寻址体系，实现了网络寻址方式的根本性变革。针对同类算力服务会广泛分布于不同物理位置的云化资源池的服务部署特性，算力路由面向用户服务层提出了“服务标识“的抽象概念，实现了对同质化服务的抽象表征，并在技术实现层面动态构建维护了服务标识到一系列同质化候选算力服务实例的映射关系。在实际寻址时，用户基于服务标识发起寻址，算力路由会基于本地维护的候选算力服务实例集，根

据一定算法，选择出最符合用户需求的算力服务实例节点，实现众多候选实例到最优服务实例节点的选择。该机制的提出使得用户对算力服务的请求是位置无关的、主机无关的，用户对算力服务的请求仅表达意图，不关心服务的部署位置信息。这是算力路由跟传统基于主机位置的 IP 路由最本质的区别 ，也标志着网络寻址从"位置寻址"向"服务寻址"的代际演进。

分布式算力路由技术相较于传统路由的另一个核心变化在于实现了算力资源状态的动态感知机制，通过持续感知获得算力节点的实时负载情况、服务响应速度、资源可用量等关键参数，为每个算力服务实例构建算力资源状态动态画像，并与网络层的时延、抖动、带宽等网络性能指标相结合，形成包含网络-计算资源状态的多维状态矩阵。这种将算力度量指标深度融入进路由体系的做法，突破了传统 IP 路由仅关注网络层指标的局限，实现了网络层对网算资源联合状态的全面感知，为智能调度决策提供了多维数据基础。分布式算力路由技术在算力资源状态感知实现层面，当前形成两条演进思路：其一是依托集中式算力感知平台，构建全局算力资源状态视图，通过标准化接口将算力资源状态同步至网络转发节点；其二是基于 BGP 协议扩展的分布式感知机制，通过在 BGP 更新报文中嵌入算力资源状态信息，实现算力资源状态基于网络协议的分布式扩散与学习；这两种技术路径分别对应集中式感知与分布式感知的不同设计，在可扩展性、状态一致性和维护成本等方面有较大差异。通过算力感知平台，有利于实现数据压缩与通告优化，应用分级分域管理体系，有助于支撑跨地域

大规模算力资源状态同步；基于协议扩展的分布式通告机制主要适配小规模算力场景，通过拓扑感知的差异化同步策略，有效防止海量算力状态信息扩散对网络稳定性的冲击。

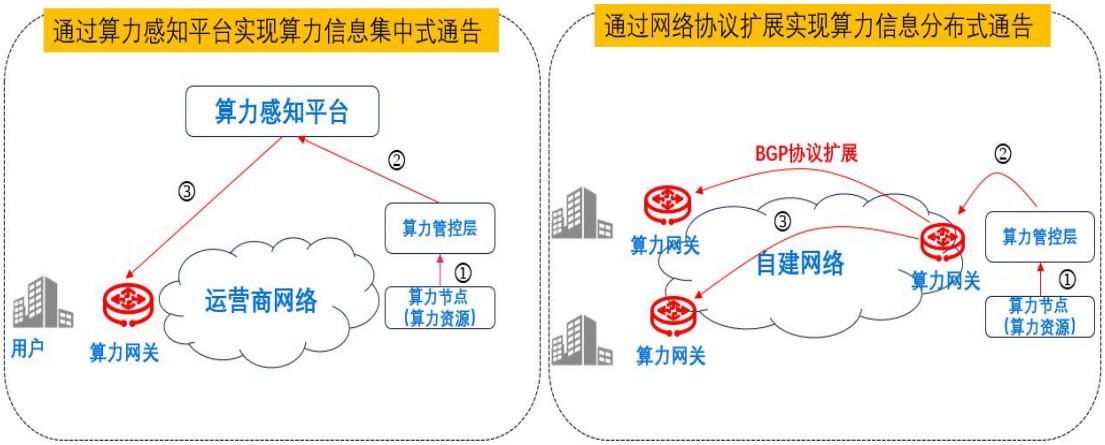

图 3-4 算力路由对算力资源状态感知的两种不同技术路线

分布式算力路由技术在寻址算法上采用了算网多因子选路算法，将传统的基于网络状态的单因子选路算法，升级为基于算力-网络状态多因子选路算法，实现网络平面的算网联合决策模型。算力路由需要构建维护到达该算力服务实例的网络状态和算力资源状态的多维状态矩阵。网络状态维度一般涉及网络丢包率、时延、时延抖动、带宽等信息，算力资源状态维度一般涉及服务器/虚机/容器等的 CPU/GPU 资源类型/负载、内存/硬盘可用容量等信息。算网多因子选路算法因为涉及到对众多信息的筛选选择，可采用分级筛选策略，首先对于用户请求的服务标识对应的所有候选服务实例集合，基于网络状态矩阵筛选满足时延、带宽等 SLA 约束的候选实例子集一，然后再基于候选实例子集一，通过算力状态矩阵过滤出具备充足计算资源和良好资源状态的实例子集二，最后基于实例子集二运用多因子加权评

分机制对各候选实例进行综合评估，选取最优算力服务实例作为路由决策。这种分层递进的决策架构有效降低了多维空间搜索复杂度，在保障网络稳定性的同时实现了算网资源的协同优化。

在坚持平滑兼容现有网络协议架构的原则下，算力路由必须与传统 IP 路由体系保持高度兼容。但算力路由与 IP 主机路由在运行机制方面存在显著差异，对实际部署工作构成了诸多挑战。从设备实现的角度来看，若要构建更为灵活高效的算力路由转发流程，则需要深入探索可编程芯片技术的应用潜力。目前，该技术已在标准制定、设备研发及试点应用等关键环节取得阶段性成果，未来，仍需集中力量攻克高动态环境适配、标准化统一等核心技术难题，以此夯实 “算力无处不在” 智能化基础设施建设的基础。

3.5 分布式算力自智技术

分布式算力自智技术，作为下一代计算基础设施的关键演进方向，其核心在于通过深度整合自动化与智能化机制，实现基于单域自治与跨域协同的算力资源管理与服务供给模式。在理论层面，单域自治强调系统在局部范围内的自主决策与闭环控制能力，涉及资源动态分配、故障预测与自愈、性能闭环优化等关键技术环节，旨在提升基础设施层面的智能化水平与运营效率。而跨域协同则侧重于构建分布式系统间的智能交互与协同机制，通过信息共享、策略联合优化等方式，打破传统计算资源孤岛，实现全局资源的最优调度与业务流程的端到端一体化整合。

分布式算力自智技术不仅是提升现有算力资源利用效率的手段，更是推动计算范式从被动响应向主动智能、从孤立运行向协同共生的根本性变革。首先，基础设施智能化不仅意味着运维成本的降低和系统可靠性的增强，更代表着计算平台本身具备了适应复杂环境和负载变化的能力。其次，业务流程一体化通过打通数据与算力在异构环境下的流转壁垒，显著提升跨系统、跨地域的业务协同效率与响应速度。再者，服务场景定制化能力使得算力服务能够根据特定应用场景（如大规模科学计算、实时工业控制、个性化推荐系统等）的差异化需求，提供高度适配的资源组合与服务模式，从而最大化性能与用户体验。

从服务供给的角度看，分布式算力自智技术致力于构建一种新型的服务化算力供给范式，其目标是为日益多样化的应用负载提供泛在可达、按需高效、实时响应、弹性灵活且安全可控的算力服务。这种范式不仅极大地丰富了算力资源的利用形态，也为新兴应用的发展提供了坚实的底层支撑。尤为关键的是，该技术内在地蕴含了使能网络基础设施实现高级别自治运行与持续演进的潜力。通过内嵌的自感知、自决策、自执行与自优化能力，网络系统能够模拟生物体的适应性，实现对自身状态的实时监控、异常行为的智能诊断与高效修复，并依据应用需求与运行经验进行自适应的架构调整与功能演进，从而形成一种可持续发展的、高度智能化的算网融合新生态。

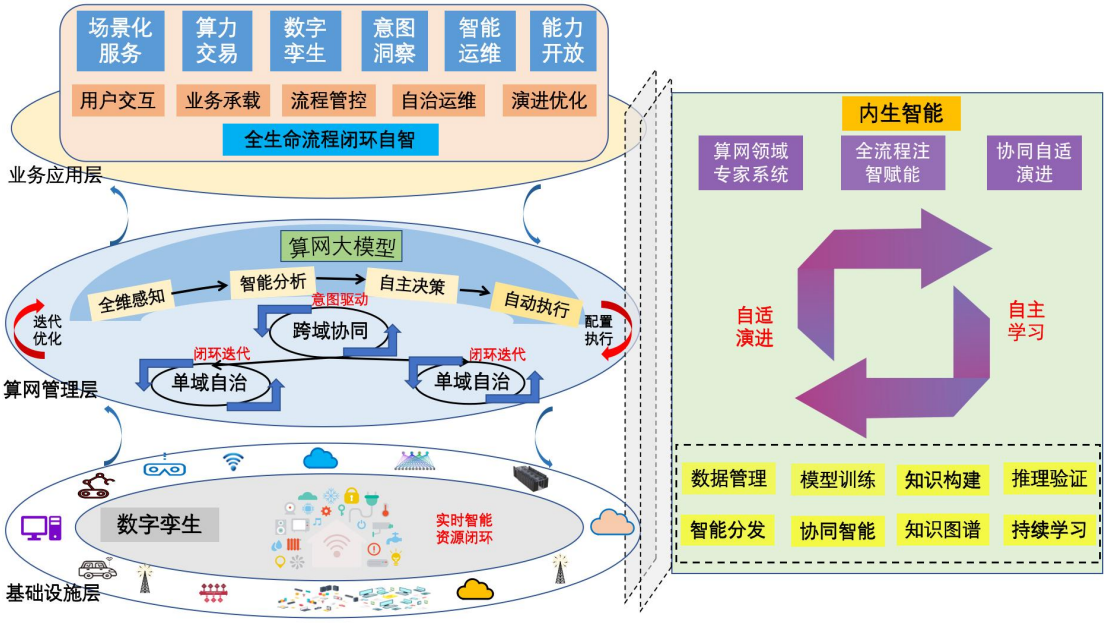

图 3-5 自智算力网络参考架构示意图

分布式系统本身固有的异构性、动态性、规模可变性以及跨域交互带来的不确定性，构成了分布式算力自智技术需要应对的基础挑战。因此，首要任务在于如何设计具备高度鲁棒性与适应性的单域自治机制，使其能在局部信息不完备或环境突变的情况下，依然保持稳定的资源管理与服务提供能力。在跨域协同中，需要建立高效、可信且低开销的通信与决策框架，以实现全局最优或次优的资源调度与任务协同，同时有效处理域间冲突与利益博弈。此外，自智系统如何实现从被动响应到主动预测的转变，即具备基于历史数据与实时监测进行故障预测、性能衰退预警乃至潜在安全威胁识别的能力。这些问题的解决将为后续的理论建模、算法设计与系统实现奠定基础。

3.6 分布式算力安全保障技术

随着分布式计算架构逐步成为国家、行业与企业级算力基础设施

的主流形态，其开放性、异构性和跨域性也带来了全新的安全挑战。分布式算力体系的建设带来了资源利用效率的显著提升，但也暴露出跨域主体身份难以统一、算力资源接入过程不可信、算力服务过程中风险不可控等一系列安全问题，传统静态、安全边界明确的防护模型已难以适应现代算力体系的安全需求。因此，亟需构建内生安全、动态可控、泛在协同、可验证可监管的分布式算力安全保障体系，为未来可信算力网络提供坚实的安全支撑。

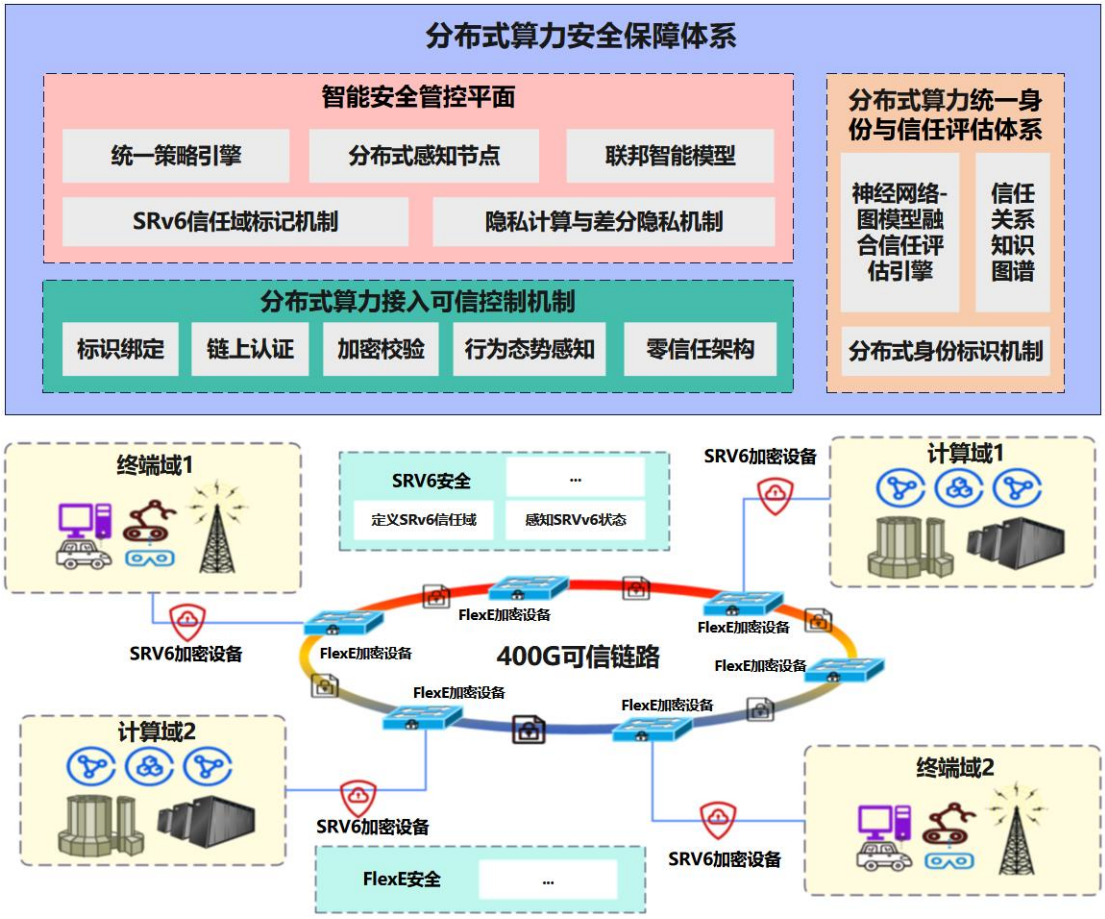

图 3-6 分布式算力保障技术架构示意图

面对分布式算力主体身份多元、权限粒度细化、行为隐蔽性增强等复杂态势，亟需建立面向分布式算力资源全生命周期的统一身份与

信任评估体系。由于算力提供方与使用方往往位于不同物理域和管理域，传统的局域信任模型难以延伸至广域协作场景。针对这一问题，以分布式身份标识（DID）机制为基础，为每个算力节点、资源主体、任务请求分配唯一标识，确保身份的唯一性、不可伪造性、不可抵赖性与上链可追溯性。通过多维数据特征（如身份、权限、历史行为、任务贡献度等）训练构建神经网络-图模型融合的信任评估引擎，动态输出主体现阶段的可信等级，并与算力调度平台联动调整访问权限和行为策略，实现基于信任的差异化资源调配。同时，构建信任关系知识图谱，形成跨域、跨平台、跨时间的动态安全关系映射，支撑算力安全决策中的风险推理、异常追踪与行为溯源，成为分布式可信基础设施的重要元数据支柱。

在多源异构算力节点不断接入算力网络的环境下，节点可信性成为系统安全保障的关键门槛。需建立基于泛在边界的动态可信接入控制机制，实现对全域算力资源的准入管理与状态感知。该机制可采用统一接入认证架构，融合标识绑定、链上认证、加密校验、行为态势感知等能力，支持对边缘节点、云节点、专用终端等算力载体进行分层分域的接入校验和实时状态同步。同时，结合零信任架构（ZTA）理念，对每次资源访问请求施加临时授权、行为分析与动态策略下发，实现无边界化条件下的“最小权限”使用原则，提升整个算力生态系统的抗破坏能力。此外，设计基于可信交易与行为审计平台的算力终端安全管理机制，通过分布式账本记录调度行为与服务调用路径，结合智能合约触发异常分析与告警通知，保障算力服务全过程中的行为

合规性与责任可溯源性。

分布式算力体系的安全风险不再是局部突发事件，而是以动态扩散性、高度联动性和隐蔽变异性为特征，需建立面向全域、可预测、可响应的风险管控机制。针对该问题，可构建智能安全控制平面（Security Control Plane），以统一策略引擎 + 分布式感知节点 + 联邦智能模型协同运行的方式，实现算力资源访问、使用、调度过程中的全域风险识别与动态应对。引入 SRv6 信任域标记机制，对算力流路径进行可信标签绑定与状态监测，基于状态感知判断调度链路是否存在异常行为（如资源漂移、频率异常、权限超越等）；同时，结合隐私计算与差分隐私机制，实现数据与行为的脱敏建模，保障风控策略训练过程中的数据合法合规。风险事件一旦被识别，系统可根据预定义响应模型进行自动阻断、隔离调度、权限收缩或联邦通报。所有行为响应与风险处置结果将在审计平台中登记，实现可持续监管能力与决策闭环。

四、分布式算力感知与调度应用场景

分布式算力感知与调度通过实时采集和评估多源异构计算节点的性能状态，结合网络带宽与业务需求动态优化任务分布，不仅极大地丰富和拓展了算力资源的供给方式，也提升了算力与网络协同的灵活性与利用效率。在保障高吞吐、低时延和高可用性的同时，调度策略还能根据峰谷负载自动弹性扩缩，降低资源闲置和运营成本，从而

为各种新兴业务场景提供可按需伸缩的计算服务能力。该技术聚焦于产业数字化升级与智能化转型，面向大模型训练、云游戏、智慧医疗、低空经济网等多领域，提供新型高效的解决方案，赋能各行各业。本章将以若干典型应用场景为例，阐述分布式算力感知与调度的核心价值与实践成效；同时也指出，该领域尚处于不断演进阶段，未来还将涌现更多创新应用，有待产学研各界持续探索与协同推进。

4.1 远程医疗

在远程医疗场景中，分布式算力感知与调度技术可以打破地理与机构壁垒。它整合从一线城市三甲医院高性能影像服务器，到县级医疗中心边缘计算节点，乃至偏远乡镇卫生院轻量级终端及云端算力集群，构建“广域医疗算力网”，让优质诊断能力和 AI 辅助分析能力高效、安全、普惠地服务患者，为医疗资源公平可及提供技术支撑。

实现这一目标，需先精准感知网络内异构、分散的算力资源，形成全局视图。系统周期性汇聚各级节点实时状态，这不仅包括院内 PACS 服务器、GPU 工作站的 CPU/GPU 利用率、显存内存余量、I/O 带宽等传统指标，也涵盖了部署在社区卫生服务中心的轻量级 ARM 架构边缘盒子乃至 5G 救护车上移动终端的网络往返时延等关键信息。更进一步，感知系统还会对节点计算架构、支持的 AI 算子集与驱动版本深度画像，建立“有效医学影像算力单元”等统一度量范式，为智能调度提供决策基准。

基于此，智能调度器可根据远程医疗任务需求规划最优计算路径。

基层医院的远程诊断请求被抽象为包含数据预处理、分布式 AI 推理等步骤的工作流，调度器评估子任务计算强度、时延阈值等，结合资源视图与网络负载模型映射执行路径。如急诊影像分析这类超低时延任务，会下沉到就近边缘节点并通过网络切片保障传输；常规体检图像筛查等则优先上传至云端大规模集群处理，以获规模效益。

分布式算力感知与调度还天然支持隐私保护的协同推理与训练。在联合推理模式下，患者原始影像块仅在本地进行 DICOM 去标识和压缩编码，随后通过加密隧道传输特征张量至云推理服务器；对联邦学习而言，GPU 工作站根据本地标签数据训练梯度，调度器按“通信带宽×梯度稀疏度”优化同步顺序，确保大规模医院联盟在不共享原始影像的前提下快速收敛。区块链-可信执行环境用于记录任务调度日志、模型版本与推理结果，保障诊断过程的全链路可追溯与不可篡改。

在实际应用中，这一体系让放射科医生在高峰时段也能在 2–3 秒内获得 AI 辅助肺结节检出结果；同一时间，偏远乡镇卫生院通过 5G 专网将疑难病例的 MRI 切片拆分上传，云端完成分区推理与拼接后把定量分析报告回传，仅耗时数十秒。大规模资源池的协同让设备利用率提升 30-50%，而调度算法对时延违约率的持续压缩，使危急重症的影像诊断平均响应时间缩短至传统单点部署的 1/5。最终，分布式算力感知与调度为智慧医疗影像诊断提供了可扩展、高并发且时延可控的数字底座，显著提升诊断效率、准确率与区域医疗公平性，并为未来多模态融合诊疗与实时远程手术导航等创新应用奠定了坚实的算力基础。

4.2 智慧城市

在智慧城市的宏大构想中，城市被视作一个庞大复杂的生命体，其高效、安全、可持续运行依赖强大的“中枢神经系统”。分布式算力感知与调度技术便是构建这一系统的核心，它将遍布城市的感知设备、边缘节点与云端数据中心相连，形成协同整体，实现对城市脉搏的实时感知、资源的智能调度和事件的精准响应，推动城市管理从被动割裂迈向主动一体。

实现城市智慧化治理，首要的是对其运行状态进行全面深入的感知。这是个多层次、异构化的算力与数据感知体系。城市末梢的百万级物联网设备，如高清摄像头、激光雷达、空气质量监测器等，是感知基础。分布式算力感知系统不仅采集海量数据，还实时掌握智能灯杆、路边单元等边缘计算节点的“健康状况”与“能力图谱”，包括 CPU/GPU 负载、网络连接质量、硬件架构等。为实现跨平台资源公允调度，系统建立面向城市场景的统一度量范式，如“有效城市计算单元”，量化不同节点执行特定任务的真实效能，形成全局统一、动态更新的“城市算力地图”。

基于这张“算力地图”，作为“城市大脑”核心决策引擎的智能调度器，能依据不同应用场景需求，进行毫秒级任务分发与路径规划。城市治理因场景多样，调度策略需有极强的情境自适应能力。比如交通摄像头捕捉到主干道车辆碰撞事故，事件被定为最高优先级，调度器会立即将实时视频流分析任务下沉到最近的路边边缘计算单元，实

现低延迟的事故识别与定性，将分析结果而非原始视频推送至交管和急救中心，同时协同调整周边交通信号灯配时，为救援车辆开辟“绿色生命通道”，整个过程数秒内自动完成，实现感知、决策与控制的无缝联动。

对于非紧急但计算密集型任务，调度器策略不同。像城市规划部门分析全年交通流量数据以模拟新地铁线路影响，这是对时延不敏感但计算量大的批处理任务。调度器会安排在夜间或周末等城市计算资源负载低的时段，将海量历史数据传输至云数据中心，利用云端算力进行大规模并行计算和仿真。这种分时分域的调度策略，大幅提升算力资源利用率，降低城市运营计算成本，实现资源效益最大化。

在智慧城市运行中，数据安全与市民隐私是红线。分布式算力感知与调度遵循“计算贴近数据”原则，为可信城市治理体系提供天然优势。大量敏感原始数据如人脸影像、车辆轨迹等，在边缘侧本地化处理，AI 模型部署在摄像头或边缘服务器上，仅将脱敏后的结构化分析结果上传至中心平台，原始视频数据分析后即刻销毁或本地按规存储，最大限度保护市民隐私。此外，结合区块链等技术，能为城市关键基础设施的控制指令提供不可篡改的执行记录，确保操作可追溯、可问责，提升城市治理的透明度与公信力。

最终，分布式算力感知与调度技术将城市从功能孤岛融合成能自我感知、调节、优化的智能有机体。它缩短交通拥堵时长，加快公共安全事件响应速度，通过精细化管理降低城市能源消耗，是提升城市治理现代化水平的关键技术，更是打造安全、高效、绿色、宜居未来

城市不可或缺的数字底座。

4.3 大模型分布式训推

由于机器学习与人工智能的迅猛发展，超大规模语言模型已跃升为科技竞逐的新高地。然而，要训练和上线诸如 GPT-4 之类的模型，必须依托巨大算力与高速网络协同配合，这对计算与通信基础设施都提出了极为严苛的要求。训练过程中需吞吐 TB-级乃至 PB-级的文本语料，并执行庞大的矩阵运算，对 GPU/TPU 的浮点性能、显存容量以及存储带宽形成高压。大规模生成式模型的训练与推理对算力提出了“高并行度、长持续、高带宽”三重要求：训练阶段需要数百到上千颗 GPU 进行同步梯度聚合，推理阶段则要在全球多地为 API 调用提供快速响应，同时保证模型权重版本一致。因此，分布式算力感知与调度能够在大模型分布式训推场景中得到广泛应用。

分布式算力感知与调度首先利用资源探针实时采集各数据中心与超算机房的 GPU 型号、显存余量、NVLink / InfiniBand 带宽和节点健康状态，并用统一的有效算力单元来衡量不同加速卡在主流 Transformer 模型上的 token-per-second 吞吐。调度器根据这些度量，将并行化策略（数据并行、张量并行、流水并行或 MoE 路由）与硬件特征匹配，把通信密集的训练算子尽可能聚集在同一低时延互联域内，把带宽压力较小的校验、存储、蒸馏等任务分散到网络次优但算力富余的节点。到了推理阶段，会实时计算各区域请求量、权重缓存命中率与 GPU 温度，自动把模型副本热迁移到用户最近的边缘集群，

并在负载回落时回收冗余实例。应用该机制后，单步训练时间可缩短两成以上，推理 QPS 提升三至四成，同时跨集群 GPU 利用率从 50%提高到 80%左右，显著压缩模型迭代周期与空闲算力成本。

为了提升系统鲁棒性与资源利用率，分布式算力感知与调度技术还能够集成“预测驱动的弹性扩缩容”功能，通过对历史训练作业和推理流量的时序分析，提前预估未来算力波峰，并在多云环境中按需预启动抢占式实例或竞价实例，结合 SLA 优先级对不同任务进行分层调度。此外，控制平面与节点代理配合，实现了秒级故障转移，当探针监测到某个交换机队列异常或 GPU 性能掉点，能够迅速将任务切换至健康机房或边缘节点，最大限度减少训练中断和推理超时。

通过这一整套实时资源探针、统一算力量化和智能调度策略的协同工作，大模型训练能够在异构硬件和多云环境中实现高效协同，通信瓶颈得到显著缓解，训练作业的吞吐率和并行效率大幅提升；在推理环节，模型副本可根据请求分布和节点健康状况灵活下沉与回收，为全球用户提供低时延一致性响应。预测驱动的弹性扩缩容与秒级故障切换机制进一步增强了系统的鲁棒性和资源利用率，使得超大规模语言模型的迭代周期缩短、运行成本降低，并在面对突发负载或硬件故障时依然能够保持平滑、稳定的服务表现。因此，分布式算力感知与调度能够成为支撑下一代生成式 AI 平台快速演进和大规模部署的关键支柱。

4.4 云游戏

在云游戏场景中，玩家对操作反馈的及时性和画面的流畅度提出了极高的期许，任何一帧渲染或一次编码的延迟都可能让操控体验大打折扣。玩家分布在不同城市甚至国家，网络质量随时可能出现抖动、丢包或带宽骤降，而日常时段与峰值时段的并发量波动又往往在数分钟内剧增，这就要求算力和网络资源能够像流动的液体一样随需而动。

为此，分布式算力感知与调度在各级渲染服务器、硬件编码器和网络接入节点中部署轻量化探针，持续采集 GPU 利用率、编码队列深度、网络往返时延和丢包率等指标。不同类型和代际的硬件性能通过统一的“渲染单元”进行量化，使得边缘节点、城域机房与云端计算资源能够在同一基准下横向比较。这样，当一位玩家发起连接时，调度逻辑便会根据其地理位置、所玩游戏类型和当前网络状况，将整条渲染—编码链路精确地分配给最近且负载最轻的节点，以保证每一次画面生成都在毫秒级内完成。在流量突发激增时，系统会迅速唤醒预留的边缘实例或启动云端竞价实例，在后台无缝迁移活跃会话，玩家几乎感受不到资源切换带来的抖动；流量回落之后，又能自动释放多余算力，避免资源闲置和成本浪费。

为了进一步提升网络抖动时的体验平滑性，平台引入了预测驱动的内容预加载与帧缓存技术。通过对玩家历史会话数据、网络波动模式和时段流量特征的深度分析，系统能够在玩家即将进入高带宽消耗场景（如大型团战、场景切换）前，提前在边缘节点或客户端缓存关

键渲染帧与差分数据。即便随后网络出现短时中断或延迟飙升，播放器也能凭借本地缓存继续输出流畅画面，待链路恢复后再快速补齐缺失帧和增量信息，从而有效削减了因网络突变带来的卡顿感。

在多租户并发运行的环境中，不同业务的资源隔离与优先级管理同样关键。分布式调度会根据各服务的协议等级划分渲染单元配额，当某项服务承压过大时，系统会重新调整资源分布，优先保证关键会话的流畅性，防止因突发流量引起的相互干扰。同时，运维团队可通过可视化仪表盘实时监控全球各区域的负载热力图与 QoE 指标，一旦监测到潜在瓶颈或性能波动，系统会自动触发策略建议，如在邻近区域预配更多渲染资源、调整网络路由优先级或优化压缩参数，确保平台长期稳定运行。

此外，为了兼顾成本与可持续发展，系统将能耗模型纳入调度决策：在非高峰时段或对画质要求不高的场景下，优先利用能效更高的硬件编码器和共享算力资源；在高价值关键会话中，则自动启用性能最强但相对耗电更高的 GPU 实例，并在会话结束后迅速回收。通过对功耗、性能与用户付费等级之间的动态平衡，运营方既能最大程度降低能源开支，也能确保用户在关键时刻获得最佳体验。

展望未来，随着虚拟现实、增强现实和下一代无线技术（5G/6G、Wi-Fi 7）的普及，云游戏的网络和算力需求将更加苛刻。分布式算力感知与调度将进一步扩展至玩家终端侧，在多接入网环境中实现链路聚合与动态切换，实时选择最优路径或并行传输，以进一步压缩端到端时延。同时，AI 驱动的网络风险检测模块将能在极短时间内发

现并规避恶意抖动或突发攻击，保障平台的安全与稳定。通过这一系列精准感知与弹性调度策略，云游戏平台将在更为复杂多变的网络环境中持续为玩家提供无缝、沉浸的互动娱乐体验，同时实现高效节能与稳健运营。

4.5 低空经济网络

在低空经济网络场景中，各类载体如大型无人机、旋翼机、气球和轻型飞艇等在低空空域与地面基站共同编织出一个灵活的三维通道，用于承载物流配送、环境监测、应急通信、临时零售与数据采集等多元化业务。随着这些平台在城郊乃至乡村上空的持续巡航，任务对计算和通信的需求随时在变化：物流环节需要在飞行器上实时完成包裹条码识别与路径优化；环境监测需对多源传感器数据进行边缘聚合、清洗和初步分析；在群众聚集或演出活动现场，需要快速部署移动支付和库存查询服务以支撑临时商铺；而在应急救援或大型活动中，低空平台又要承担视频转发和通信中继的重任。这种动态生态对算力的并发性、链路的可靠性和能源的可持续性都提出了高要求，同时不同业务之间的负载高峰常常出现错峰重叠，给资源调度带来极大挑战。

分布式算力感知与调度通过在每一个飞行平台和地面节点中安装轻量化探针，持续采集包括 CPU/GPU 利用率、硬件编码器队列长度、网络往返时延、丢包率及平台剩余电量在内的多维指标，并将异构设备的性能映射为统一的“空中算力单元”。在此基础上，调度器根据当前业务类型的时延敏感度、数据量大小与处理复杂度，将紧急

的条码识别和支付验证任务优先分配给网络最稳定、计算负载最轻的临近无人机或地面服务车；对容许短暂延迟的环境数据清洗与批量分析，则集中调度到飞行器群中计算资源富余的节点或后端边缘机房。每当检测到流量骤增或某个平台电量临界，系统会自动唤醒预置的备用载体、启动竞价式算力实例，并在后台平滑迁移正在运行的子任务，从而保证业务不中断。任务完成后，调度器还会迅速回收已用算力，避免能源与资源的浪费。

通过这种面向多业务、多载体的精细化感知与弹性调度，低空经济网络在实践中取得了显著效果。包裹分拣与跟踪的响应延迟大幅缩减，复杂环境监测的初步分析结果能够更快送达指挥中心，临时零售点的支付和查询服务也始终保持高可用，而在突发演练或救援场景中，通信中继能力得以随需扩展，且在保障任务执行的同时，平台总体能耗与运营成本均得到有效控制。未来，随着更多轻量化 AI 算子和高效能算力模块的加入，这一底层架构还将进一步提升对超低时延和高并发业务的支撑能力，为低空经济的发展奠定坚实的技术基础。

4.6 战术网络

战术网络是现代军事通信系统的核心组成部分，其主要功能是为战场上的作战单元提供实时、可靠的通信支持和信息共享。随着军事技术不断发展，战术网络的复杂性和对算力的需求也在不断增加，分布式算力感知与调度技术为战术网络的优化与发展开辟了新路径。

战术网络需要处理大量的实时数据，以此为根据做出战术决策。

例如，在战场态势感知中，需要实时处理来自多个传感器的数据，以生成准确的战场态势图；在目标识别中，需要快速处理图像和视频数据，以识别潜在威胁；在通信加密中，需要实时加密和解密大量数据，以确保通信的安全性。这些任务的高效执行依赖于强大的算力支持。分布式算力感知与调度系统将网络从一个单纯的数据传输管道，转变为一个分布式的、可协同工作的计算平台，使战术网络能实时感知网络各节点的算力状态，将任务快速分配至最合适的节点，满足战术应用对实时性的严苛要求。

在分布式算力架构可以提升战术网络的可靠性与抗毁性。战场环境复杂多变，网络节点随时可能因敌方攻击或自然因素受损。分布式算力系统中，即使部分节点失效，其他节点仍可继续承担计算任务，保障系统的整体运行。算力感知与调度系统持续监控承载服务节点的状况，一旦某个指挥节点因敌方火力打击或强电磁干扰而离线，调度系统则会依据预设策略，迅速将服务和计算“迁移”到网络中其他节点上，并重新建立服务连接。这将改变传统指挥体系“中心即是弱点”的困境，在部分网络被摧毁的情况下，指挥能力可以延续、作战体系也能保持核心功能的运作。

分布式算力调度是加速战术决策、赋能自主协同作战的核心。装备了分布式计算能力的战术网络，可以实现“边缘决策”，不必等待后方指挥链层层下达指令，前沿的传感器节点发现目标后，可立即触发局域的“决策任务”。算力调度系统将计算结果直接分发给网络内最合适的攻击单元，实现“发现即摧毁”的快速闭环。更进一步，通

过在战术边缘部署强化学习等 AI 模型，整个作战编组可以进行自主协同进化。例如，一个无人机蜂群在执行任务时，可以利用分布式算力，根据实时战场环境和战损情况，集体重新计算和优化队形、分工和攻击策略，而无需依赖任何中心节点的微操控制。这种由数据和算力驱动的自主协同，将极大提升作战单元的智能化水平和任务的成功率。

综上所述，分布式算力感知与调度技术并非简单地将计算资源分散化，而是通过赋予战术网络以智能的“感知”和自主的“调度”能力，从根本上重构了信息时代战场的作战模式，推动战术网络从脆弱的通信链路演变为一个坚韧、智能、高效的分布式作战中枢，为赢得未来高科技战争奠定坚实的算力基石。

4.7 智能制造

在第四次工业革命的浪潮推动下，分布式算力感知与调度的应用具有重大意义，正深刻变革着智能制造领域的生产模式与效率。智能制造的本质，是将制造系统从一个由物理设备和人力构成的集合体，转变为一个由数据驱动、模型定义、软件控制的智能物理系统（CPS，Cyber-Physical System）。其中海量数据的产生、传输、处理与分析，对算力提出了空前巨大、异构且时延敏感的要求。传统集中式算力显现出响应迟缓、成本高昂、资源利用率不高等问题，分布式算力感知与调度则可将部分数据处理任务分流至靠近设备的边缘计算节点，实现数据就近快速处理，减少传输延迟。

分布式算力感知与调度在智能制造领域的应用，是对传统生产范式的一次深刻重构，它将算力作为一种可灵活调配的核心生产要素，深度融入到制造的全生命周期中。

在智能制造中，产品设计与研发阶段是算力需求最为集中的环节之一。分布式算力调度系统能够实时感知整个算力网络中可用的算力资源池，无论是企业私有云中的计算集群，还是远在“东数西算”节点上的国家超算中心，系统都能根据仿真任务的规模、优先级和预算，自动选择并调度最合适的算力资源，从而将以往需要数周乃至数月的仿真周期缩短至几天甚至几小时。此外，分布式算力调度还可以支持多学科优化设计，例如同时进行结构优化和热力学分析，提高产品的整体性能。

在生产制造阶段，分布式算力感知与调度技术可以用于优化生产过程。制造企业生产任务多样，需考虑设备状态、订单优先级、物料供应等多因素。分布式算力感知与调度系统能实时获取各生产环节信息，利用本地与云端算力动态调整生产计划和资源分配。例如，在智能工厂中，机器人和自动化设备需要实时接收任务指令并进行协同工作。此外，通过分布式算力调度，还可以实现生产过程中的故障预测和预防性维护，减少设备停机时间，提高生产效率。

产品质量检测与控制需要对产品进行高精度的图像识别和数据分析，分布式算力感知与调度技术可以支持大规模的图像处理和数据分析任务，提升检测精度与效率。例如，在边缘节点利用轻量化图像识别模型实时对图像进行初步筛选，识别明显缺陷，复杂缺陷图像再

上传至云端进行深度分析；实时监测各个计算节点的负载情况，并将图像处理任务动态分配到空闲的节点上，从而加快检测速度。此外，分布式算力调度还可以支持多模态数据融合，例如将图像数据和传感器数据结合起来进行综合分析，提高质量检测的准确性。

供应链管理涉及到多个环节的协同工作，分布式算力感知与调度可以推动跨企业、跨区域协同制造。在未来的智能制造生态中，订单、设计、生产、物流等环节将在不同的企业主体之间动态共享与协同。这需要一个强大的“产业大脑”来进行全局的资源优化。平台可以调度闲置的算力资源，实现跨域算力与制造能力的协同调度，打破企业间的信息壁垒和资源孤岛，使得整个产业链能够像一个紧密耦合的虚拟工厂一样运作，极大地提升区域制造产业集群的整体竞争力和市场响应速度。此外，分布式算力调度还可以支持供应链中的风险预测和应急响应，提高供应链的稳定性和可靠性。

4.8 自动驾驶

自动驾驶技术正朝着 L4/L5 级别的高度自动化迈进，这使车辆需要实时处理海量的环境感知数据、进行复杂的决策规划与控制计算，对算力的需求是海量、瞬时且不容出错的。一辆高级别自动驾驶汽车每小时产生的数据量可达 TB 级别，其内部的计算平台需要在毫秒级的时间内完成从数据融合、目标识别到轨迹预测、行为决策等一系列复杂运算。单纯依靠车载计算单元的算力，会面临功耗、散热、成本以及算力天花板的巨大挑战；而完全依赖远端中心云的计算模式，其

固有的网络延迟对于实时的驾驶决策是不可接受的。因此，构建一个“车-路-云”一体化的协同计算体系，并引入分布式算力感知与调度的先进理念，对车载、路侧、云端的异构算力进行统一管理和智能分配，已成为突破单车智能瓶颈、实现安全、高效、可扩展自动驾驶的关键路径。

分布式算力感知与调度在自动驾驶领域的应用，其精髓在于将车辆从一个算力孤岛，转变为一个能够与外部环境进行计算资源实时交互的智慧体。这里的“感知”具有双重维度：一是车载操作系统对自身计算资源的“内省感知”，需持续监控其高性能计算平台的负载率、内存占用、芯片温度和功耗等状态；二是车辆通过车对外部可用算力资源的“环境感知”，包括感知路侧单元（RSU，Road Side Unit）的算力负载、网络连接质量以至中心云数据中心的宏观资源状况。只有建立在这种内外兼修、实时动态的全局算力资源图谱之上，智能调度才成为可能。

“调度”则是基于感知结果所执行的核心动作，即智能化的计算任务卸载（Computational Offloading）。车载的智能调度器，如同一个运筹帷幄的“算力总管”，它根据不同驾驶任务的特性决定该任务是在本地执行，还是卸载到多接入边缘计算（MEC，Multi-access Edge Computing），抑或是提交给中心云。、

分布式算力感知与调度为自动驾驶的算力管理提供了全新的范式，构建了一个可感知、可调度、可协同的“车-路-云”分布式计算架构。通过将安全关键任务锁定在本地、将复杂感知任务协同于边缘、

将海量训练任务汇聚于云端，该技术使得车辆能够在确保绝对安全的前提下，突破自身物理算力上限，获得近乎无限的“云端外脑”支持。

五、分布式算力感知与调度行业发展建议

5.1 发展面临的挑战

技术挑战：分布式算力感知与调度技术面临的核心技术挑战在于如何高效协同异构、动态的算力资源。边缘节点与算力中心的算力、存储和网络资源呈现高度动态性，传统静态感知机制难以实时捕获资源状态变化，而频繁探测又会带来额外开销，分级分域感知技术需要在精度与效率间取得平衡。异构计算单元的性能差异显著，现有资源度量模型缺乏统一的跨平台量化标准，影响调度决策准确性。同时网络环境的不稳定性导致边缘节点间通信质量波动，跨域协同还面临管理策略差异带来的标准化难题。算力自智技术受限于数据稀疏性，AI模型的训练效果和决策可解释性面临挑战。这些问题的解决需要突破轻量级感知、智能调度算法和隐私计算等关键技术，构建自适应、高可靠的分布式算力调度体系。

基础设施挑战：网络传输的协同能力亟待加强，边缘接入网与核心骨干网间的带宽落差导致跨级调度指令的端到端时延保障存在波动，多运营商网络边界策略差异更使得跨域算力的路由稳定性面临挑战，因此需要构建更精细的互联协商机制以平滑传输路径。广域节点时钟同步精度亟待提升以抑制微秒级偏差；同时需深度适配异构协议栈、突破转换层微延迟瓶颈并优化轻量化终端通信开销。物理层与协议层的协同演进是释放跨域服务确定性的关键。

标准挑战：当前，主流云服务商、电信运营商边缘平台及工业设备厂商普遍采用私有化的资源描述框架与异构接口规范，在算力性能表征、拓扑关系建模、实时负载度量等关键维度缺乏统一语义定义。这种标准缺失导致跨管理域的资源发现与调度需通过复杂的定制化中间件实现，显著推高系统集成成本与生态协同门槛，阻碍产业级算力网络的集约化演进。

经济挑战：在分布式算力感知与调度技术蓬勃发展的当下，行业在经济层面遭遇诸多挑战。资源成本方面，地域与厂商差异导致算力资源成本结构复杂多样。不同地区算力资源在价格、性能及可用性上参差不齐，低价算力性能欠佳，高价算力性能卓越。这种异构性极大增加了统一调度的难度，要求调度系统在决策时必须精准权衡性能与成本，力求实现资源的最优配置与高效利用；在算力交易市场机制层面，当前尚未形成统一成熟的规则体系。跨主体、跨地域的算力资源流通效率有待提升，相关交易模式仍处于实践探索的初期阶段，市场基础设施与协作机制需进一步健全完善。

5.2 发展阶段划分

目前分布式算力感知与调度的发展仍处在重要的建设阶段，无论是学术界、产业界还是研究领域都在持续推进理论创新与工程实践。分布式算力感知与调度平台建设涉及三方协同，不同领域之间需要打通技术壁垒，进行标准的升级互通与人才的交流协作，方案设计与技术落地都存在很多挑战。因此分布式算力感知与调度技术发展可分为

以下四个阶段。

5.2.1 起步阶段

分布式算力感知与调度行业的起步阶段，技术突破是核心驱动力，这一时期主要以技术验证与初步探索为特征。受限于当时网络带宽、数据传输效率以及分布式系统管理技术的不足，行业参与者聚焦于解决基础架构的可行性问题，例如如何实现跨地域、跨组织的算力资源识别、连接与简单协同。早期尝试多基于单机集群的扩展思维，通过定制化协议和中间件技术，验证分布式算力调度的基本逻辑，如任务分解、负载均衡和结果聚合等关键环节。这一阶段的典型场景集中在科研机构和头部科技企业的内部实验环境中，用于处理高性能计算、大数据分析等特定领域的需求。

5.2.2 整合阶段

整合阶段标志着行业从技术探索向规模应用的关键跃迁。伴随5G、软件定义网络等基础设施技术成熟，算力资源实现跨域全局化聚合，形成覆盖多数据中心与云平台的协同体系。企业通过构建统一资源池，推动离散算力向可度量、可流通的服务形态转化，智能调度系统依托自适应算法实现精准动态供给。行业实践表明，超大规模云服务商已建立体系化调度框架，显著提升资源集约效能；混合云架构通过能力下沉构建全域协同的算力供给网络。开放标准体系持续深化——硬件层依托开放计算推进异构环境兼容，软件层基于云原生规范

统一编排范式，为产业集约化发展奠定基础。

5.2.3 智能化阶段

智能化阶段是行业从“量变”到“质变”的关键跃迁。人工智能、机器学习技术的深度融合，使算力感知与调度具备自主决策能力。感知层面，系统通过实时监测节点温度、功耗、负载等参数，结合历史数据预测故障风险，实现预防性维护；调度层面，强化学习算法能够根据任务优先级、资源成本、网络延迟等多维度因素，动态优化分配策略，使算力利用率得到很大突破。例如，谷歌通过 Borg 系统将任务调度时间从分钟级缩短至毫秒级，年节省算力成本超 10 亿美元；华为云 AI 调度器在 AI 训练场景中，通过智能拓扑感知将数据搬运时间减少 40%。这一阶段，行业应用场景从互联网向制造、医疗、金融等传统领域渗透，成为数字化转型的核心基础设施。

5.2.4 生态化阶段

生态化阶段标志着行业竞争范式从单体创新向全产业链协同的系统性跃迁。随着算力基础设施逐步成为支撑经济社会发展的公共基础资源，产业生态链加速完善——上游异构计算架构持续演进，推动算力资源弹性供给能力升级；中游绿色集约化设施构建稳健算力底座；下游开放平台通过标准化接口赋能垂直领域创新应用；终端用户基于服务化模式实现普惠接入。行业实践表明，领先云平台已形成繁荣的应用开发生态，国家级算力枢纽工程有效促进跨区域资源协同。当前，

产业边界持续融合重构，算力与数据、算法深度耦合，共同构筑数字经济发展的核心要素基座。这一阶段，行业边界逐渐模糊，跨界融合成为主流，算力与数据、算法共同构成数字经济的新生产要素。

5.3 发展对策建议

技术创新与研发：需重点突破动态感知与智能调度技术。研发轻量级分级分域感知算法，结合边缘计算与数字孪生技术，实现低开销、高精度的资源状态实时捕获；构建跨平台、多维度的资源统一度量模型，通过标准化算力、存储、网络等关键性能指标，提升调度决策的精准性；发展基于强化学习与联邦学习的智能调度算法，优化多目标（时延、成本、能耗等）动态权衡能力，避免局部最优问题；推动算力自智技术演进，利用边缘侧增量学习与小样本训练提升 AI 模型的适应性，同时增强决策可解释性以满足关键领域合规需求；加强隐私计算与安全协同技术研发，确保跨域数据交互的可信性与安全性。

基础设施建设与改造：需强化网络传输与协议协同能力。优化边缘接入网与核心骨干网的代际协同，通过 SDN/NFV 技术实现带宽资源的动态调配，降低跨级调度时延；推动多运营商网络互联协商机制建设，统一跨域路由策略，提升算力流传输稳定性；加强广域节点时钟同步技术研发，满足分布式协同计算的高精度时序需求；促进工业 TSN 与云原生 IPv6 等异构协议栈的深度适配，提升跨域数据传输的确定性；研发高效协议转换中间件，降低边缘设备接入算力网络的通信开销，提升全域资源感知敏捷性。

标准制定与完善：为推动分布式算力感知与调度行业 标准发展，需多管齐下促进标准统一与市场机制完善。一方面，由行业协会联合产业各方力量，加快制定涵盖分布式算力资源描述、接口协议及调度规则的全面行业标准，明确算力性能表征、拓扑关系建模等关键维度的语义定义，打破主流云服务商、电信运营商边缘平台及工业设备厂商私有化标准的壁垒，推动其开放私有化 API，降低生态协同成本，实现跨管理域资源的高效发现与调度。

经济协同机制：为推动分布式算力感知与调度行业的经济协同发展，需重点突破资源价值评估与市场机制建设的关键环节。建议采取以下措施：建立全域算力动态价值评估体系，通过标准化模型量化性能、时延、成本等多维参数，生成实时资源价值图谱，支撑调度系统的多目标优化决策；构建弹性分层定价机制，基于服务质量承诺与实时负载状态动态调节资源溢价，形成价格性能联动的市场调节能力；健全算力交易基础设施，由产业联盟主导制定资源描述规范、跨域服务等级协议（SLA）及结算规则，降低协同摩擦成本；打造可信交易执行平台，确保多主体协作的可验证性。通过技术标准与经济机制的深度融合，系统性释放分布式算力资源的协同效能。

六、总结与展望

分布式算力感知与调度技术作为支撑国家算力网络战略落地的核心引擎，致力于整合边缘节点、区域级算力中心等异构资源，构建“全域协同、智能弹性”的算力服务体系。本白皮书系统剖析了该技术在推动国家数字化转型、产业智能化升级及应对技术演进趋势中的迫切需求，明确提出“构建全域智能算力网络，实现异构资源统一度量、跨域协同调度与安全开放服务，赋能数字化转型”的核心发展目标。在关键技术层面，创新性地设计了分级分域协同的算力网络技术体系，其核心涵盖分层感知、统一度量、智能调度、算网路由、自治优化与全域安全六大要素，旨在实现对异构算力资源的高效管控与按需服务，驱动算力基础设施向泛在化、智能化方向持续演进。最终，通过架构革新与标准体系建设，本技术聚焦支撑产业数字化升级与智能化转型，面向远程医疗、智慧城市、大模型训练、云游戏等多元化应用领域，提供新型高效解决方案，全面赋能各行各业发展。

分布式算力感知与调度技术正迈向智能化与生态化融合的新阶段。AI 驱动的动态优化算法与异构资源适配技术突破效能边界，开放接口协议与可信生态构建降低使用门槛，推动跨行业协同规模化落地。随着量子计算、光子计算等前沿技术融入，行业加速向"泛在算力"演进，最终实现算力资源无感接入与智能流动，为全社会数字化、智能化转型提供底层支撑。

附录 A：术语与缩略语

CPN	Computing Power Network，算力网络
QoS	Quality of Service，服务质量
CPU	Central Processing Unit，中央处理器
AI	Artificial Intelligence，人工智能
IPV6	Internet Protocol Version 6，互联网协议第 6 版
SLA	Service Level Agreement，服务水平协议
SDN	Software Defined Network，软件定义网络
NFV	Network Function Virtualization，网络功能虚拟化

参考文献

[1]工业和信息化部等十四部门.《关于进一步深化电信基础设施共建共享 促进“双千兆”网络高质量发展的实施意见》[EB/OL].(2023-05-25)

[2]工业和信息化部.(2021).《工业互联网创新发展行动计划(2021—2023 年)》[EB/OL].

[3]国家发展改革委,中央网络安全和信息化委员会办公室,工业和信息化部,国家能源局.《全国一体化大数据中心协同创新体系算力枢纽实施方案》:发改高技〔2021〕709 号[Z].2021-05-24. https://www.cac.gov.cn/2021-05/26/c_1623610318323289.htm

[4]国务院办公厅. 国家应急通信保障预案: 国办函〔2021〕112 号[Z].2021-12-29.

[5]中国信息通信研究院.绿色算力白皮书[R].2023

[6]生态环境部环境规划院.中国区域电网二氧化碳排放因子研究[R].2023

[7]P. Mokshith and P. K. Pullela, "Cloud Gaming: Revolutionizing the Video Gaming Industry," 2023 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Mysuru, India, 2023, pp. 165-169.

[8]C. Zhu et al., "Intelligent Management and Computing for Trustworthy Services Under 6G-Empowered Cyber-Physical-Social System," in IEEE Network, vol. 39, no. 2, pp. 124-133, March 2025.

[9]S. Fu, W. Zhang and Z. Jiang, "A network-level connected autonomous driving evaluation platform implementing C-V2X technology," in China Communications, vol. 18, no. 6, pp. 77-88, June 2021.