

算力互联网体系架构与AI驱动的资源调度关键技术研究

宋美芳 张紫明 韩冰心

河南省信息咨询设计研究有限公司 河南 郑州 450008

摘要：在数字经济蓬勃发展的当下，算力需求呈现多元化与规模化趋势。本文聚焦算力互联网体系架构与AI驱动的资源调度关键技术，阐述算力互联网基于分布式计算等的基础原理、核心特征及技术支撑。剖析分层体系架构设计与运行机制，探讨AI在资源调度中的应用原理、需求预测、动态分配算法及调度优化技术。研究算力互联网与AI调度的适配方式、协同运行流程及技术衔接，为构建高效、智能的算力互联网，满足多元算力需求提供理论支撑与技术参考。

关键词：算力互联网；体系架构；AI驱动；资源调度；协同机制

引言：数字经济的快速发展使算力成为关键生产要素，其需求规模不断扩大且呈现多元化特征。传统算力供给模式难以满足复杂多变的算力需求，算力互联网应运而生。算力互联网通过整合分散算力资源，实现跨域协同与高效流转。而AI技术凭借强大的数据分析与智能决策能力，为资源调度带来新的解决方案。研究算力互联网体系架构与AI驱动的资源调度关键技术，对于提升算力利用效率、推动数字经济发展具有重要意义，是当前信息技术领域的重要课题。

1 算力互联网基础理论

1.1 算力互联网的基本原理

算力互联网的基本原理植根于分布式计算、网络虚拟化与资源池化技术，核心是实现算力资源的跨域协同、高效流转与按需供给。依托高速互联网络，将地理上分散的算力节点有机联结，打破算力资源的物理壁垒与地域限制，构建一体化算力资源池^[1]。通过标准化的接口协议与数据传输规范，实现不同类型、不同规格算力资源的互联互通，让算力如同水电等公共资源般可灵活调度、便捷取用。通过资源抽象、调度协同与传输优化，实现算力资源的动态分配与高效利用，支撑各类算力密集型应用的规模化部署，适配数字经济发展对算力的多元化需求，这一原理源于分布式系统理论与网络协同技术的深度融合。

1.2 算力互联网的核心特征

算力互联网具备鲜明的技术与应用特征，这些特征依托其架构设计与运行机制形成，区别于传统算力供给模式。分布式布局体现为算力节点广泛分布于不同地域，通过网络实现协同联动，避免单一节点故障对整体算力供给造成影响。异构性体现在接入资源的多样性，涵盖通用算力、智能算力、超算算力等多种类型，可适

配不同应用场景的差异化需求。弹性伸缩能力通过资源池化与动态调度技术实现，能够根据应用负载变化实时调整算力供给规模，避免资源闲置与供给不足。高效协同性体现在算力、网络、存储等各类资源的深度联动，通过层间协同机制实现资源的最优配置，提升整体算力利用效率。

1.3 算力互联网的技术支撑基础

算力互联网的稳定运行与高效应用，依赖多领域技术的协同支撑，各技术模块相互配合、形成合力。网络传输技术是基础支撑，高速光网络、软件定义网络等技术为算力跨域传输提供高带宽、低时延的传输通道，保障算力数据的快速流转。算力虚拟化技术通过对物理算力资源的抽象与封装，实现算力资源的池化管理，为资源调度提供灵活支撑。调度管理技术通过智能算法实现算力资源的动态分配与负载均衡，优化算力利用效率。安全防护技术涵盖数据加密、访问控制、漏洞检测等，保障算力传输与存储过程的安全可靠，边缘计算技术将部分算力下沉至网络边缘，降低时延、提升本地应用响应速度，为算力互联网的广泛应用提供重要支撑。

2 算力互联网体系架构设计

2.1 体系架构的分层设计

算力互联网体系架构的分层设计遵循模块化、可扩展、协同化原则，基于网络分层思想与算力资源调度需求，构建层次清晰、功能独立且协同联动的架构体系。分层设计核心是将整体功能拆解为不同层级，每个层级承担特定职责，通过层级间的协同配合实现算力资源的全流程管理与高效调度^[2]。架构分层需兼顾算力供给、网络传输、调度管理与应用服务的全链路需求，通常划分为基础设施层、网络传输层、算力调度层与应用服务层，各层级边界清晰、接口标准化，既保障单一层级的

功能完整性,又实现层级间的无缝衔接,为架构的稳定运行与灵活扩展提供支撑,该设计思路源于网络分层架构理论与分布式资源管理技术的实践延伸。

2.2 各层级功能定位

算力互联网体系架构各层级依据自身定位承担差异化功能,形成协同高效的功能体系。基础设施层作为架构的底层支撑,整合各类算力资源与存储资源,通过资源池化技术实现物理资源的抽象封装,为上层提供标准化的算力供给。网络传输层负责算力数据与控制信号的跨域传输,依托高速互联技术构建低时延、高带宽、高可靠的传输通道,保障算力资源跨域流转的高效性与稳定性。算力调度层是架构的核心枢纽,承担算力资源感知、调度决策与动态调整功能,实现算力资源的最优分配与负载均衡。应用服务层面向各类算力需求场景,提供标准化的算力服务接口,将抽象的算力资源转化为可直接调用的应用服务,适配不同领域的算力需求。

2.3 层级间接口设计

层级间接口设计是保障算力互联网体系架构协同运行的关键,需遵循标准化、规范化、可兼容原则,实现不同层级间的数据交互与指令传输。接口设计需基于统一的通信协议与数据格式,确保层级间信息传输的准确性与高效性,避免因接口不兼容导致的协同障碍。基础设施层与网络传输层的接口,主要实现算力资源状态信息与传输需求的交互,支撑算力资源的接入与传输调度。网络传输层与算力调度层的接口,负责传输算力调度指令与资源状态反馈,保障调度决策的精准落地。算力调度层与应用服务层的接口,提供标准化的算力服务调用方式,实现应用需求与算力资源的精准匹配,接口设计需兼顾灵活性与安全性,适配架构扩展与不同应用场景的差异化需求。

2.4 体系架构的运行机制

算力互联网体系架构的运行机制围绕算力资源的全生命周期管理构建,实现各层级、各模块的协同联动与高效运转。资源感知机制实时采集各层级的资源状态、负载情况与需求信息,为调度决策提供数据支撑,依托传感技术与数据采集协议实现信息的实时获取与同步更新。调度协同机制基于资源感知数据,通过智能调度算法实现算力资源的动态分配,协调各层级功能衔接,优化资源利用效率。容错备份机制针对架构运行中的故障场景,通过冗余设计与故障转移技术,避免单一节点或层级故障影响整体架构运行,保障算力供给的连续性。动态适配机制能够根据应用需求变化与资源状态调整,实时优化层级间的协同模式与资源分配策略,确保架构

始终适配算力需求的动态变化,支撑算力互联网的稳定高效运行。

3 AI驱动的资源调度关键技术

3.1 AI在资源调度中的应用原理

AI在算力资源调度中的应用,依托数据挖掘、机器学习等核心技术,打破传统调度模式的局限,实现调度过程的智能化与高效化^[3]。应用原理围绕算力资源全生命周期的状态感知与动态调控展开,通过多维度采集算力网络中各类资源的运行数据、业务需求数据,经过系统的数据清洗、特征提取与归一化处理,构建贴合实际调度需求的智能模型。借助模型对资源供需关系进行深度剖析,实现资源运行状态的精准判断与调度决策的自主生成,替代传统人工调度的经验驱动模式,提升调度决策的合理性与时效性,精准适配算力互联网动态变化的资源运行场景,为后续各项调度技术的落地提供基础支撑。

3.2 算力资源需求预测技术

算力资源需求预测技术是AI驱动资源调度的前提,通过AI算法对历史需求数据、实时运行数据进行分析,实现对未来算力需求的精准预判。该技术依托时序预测算法,整合不同应用场景的需求特征、历史负载数据、环境因素等多维度信息,挖掘需求变化的内在规律,构建需求预测模型。模型通过对数据的清洗、特征提取与模式识别,剔除干扰因素对预测结果的影响,提升预测精度。预测过程中,算法可根据需求场景的动态变化,自适应调整模型参数,适配不同类型应用的需求波动特点,提前预判算力需求的规模、时效与类型,为资源调度决策提供科学依据,避免资源供给与需求脱节,该技术基于时序分析与机器学习算法的交叉应用构建。

3.3 动态资源分配与调度算法

动态资源分配与调度算法是AI驱动资源调度的核心载体,通过AI算法实现算力资源的动态分配与高效调度,适配算力需求与资源状态的实时变化。算法基于资源感知数据与需求预测结果,结合算力资源的特性与约束条件,构建动态调度模型,实现资源在不同节点、不同应用之间的合理分配。调度算法兼顾资源利用效率与任务运行质量,通过负载均衡策略,避免部分节点过载、部分节点闲置的情况,优化整体资源利用率。同时,算法具备快速响应能力,可实时捕捉资源状态与需求变化,动态调整资源分配方案,解决算力资源异构性、需求随机性带来的调度难题,常见算法包括强化学习调度算法、深度学习自适应调度算法等,均基于智能算法与资源管理技术的协同优化。

3.4 AI驱动的调度优化技术

AI驱动的调度优化技术围绕资源调度的效率、精度与稳定性展开,通过AI算法对调度过程进行全流程优化,弥补传统调度技术的不足。该技术整合资源调度各环节的数据,通过多目标优化算法,平衡资源利用效率、任务响应速度、能耗成本等多方面需求,生成最优调度方案。优化过程中,AI算法可对调度策略进行持续迭代优化,结合调度过程中的反馈数据,修正调度偏差,提升调度效果。同时,通过智能容错优化,识别调度过程中的潜在故障与风险,提前采取应对措施,保障调度过程的稳定性。此外,该技术可结合边缘计算、网络虚拟化等技术,优化跨域调度的传输效率与资源适配性,实现调度全流程的智能化优化,支撑算力互联网资源调度的高效、稳定运行。

4 算力互联网与 AI 调度的协同机制

4.1 AI调度与体系架构的适配方式

AI调度与算力互联网体系架构的适配,是实现两者高效协同的基础前提,需依托算力互联网各层级功能特性,构建适配性运行模式^[4]。适配设计需贴合算力互联网分层架构的功能定位,使AI调度模块能够深度融入各层级运行,实现调度指令与架构运行的精准呼应。需结合算力互联网的资源调度需求,优化AI调度的功能模块,使AI调度能够精准匹配架构的资源传输、算力分配逻辑,避免调度指令与架构运行脱节。通过优化接口设计,打通AI调度与体系架构各层级的信息传输通道,确保调度信息能够快速、准确传递,实现AI调度与体系架构的无缝适配,为后续协同运行提供保障。

4.2 协同运行的关键流程

算力互联网与AI调度的协同运行围绕资源全生命周期管理,构建闭环式关键流程,各环节衔接流畅、高效联动。流程始于资源与需求的双重感知,AI调度实时采集算力互联网各层级的资源状态数据,同步解析应用侧的算力需求,构建全面的资源与需求视图。基于感知数据开展协同决策,AI调度结合架构运行状态,生成适配的资源调度方案,同步反馈至架构各层级。方案下发后进入资源分配与任务执行环节,架构各层级根据调度指令完成资源绑定与任务启动,AI调度持续跟踪运行状态。运行过程中开展动态协同调整,AI调度捕捉资源负载与需求变化,实时优化调度方案,架构各层级同步响

应调整指令,确保协同运行的稳定性。任务完成后进入数据反馈环节,将运行数据回传至AI调度模块,为后续协同优化提供数据支撑,形成完整协同闭环。

4.3 协同过程中的技术衔接

技术衔接是保障算力互联网与AI调度协同运行的核心支撑,需突破各环节技术壁垒,实现技术层面的无缝衔接。信息传输层面,采用高效的信息交互技术,确保AI调度指令与算力互联网资源状态信息的快速传递,减少传输延迟,保障调度指令的实时性。资源调度层面,整合AI调度算法与算力互联网资源管理技术,使调度算法能够精准适配架构的资源分配逻辑,提升资源调度的合理性^[5]。数据交互层面,建立统一的数据交互标准,规范AI调度与体系架构之间的数据格式,避免数据不兼容导致的衔接不畅。通过技术衔接的优化,实现AI调度与算力互联网体系架构的深度协同,充分发挥两者的技术优势。

结束语

算力互联网体系架构与AI驱动的资源调度关键技术研究,实现了算力资源的高效整合与智能调配。通过构建分层体系架构,明确了各层级功能与接口设计,保障了架构的稳定运行。AI技术的应用,使资源调度更加精准、智能,提升了算力利用效率与任务响应速度。两者的协同运行,解决了算力资源异构、需求多变等难题。该研究成果为算力互联网的建设与应用提供了有力支持,有助于推动数字经济向更高水平发展。

参考文献

- [1]张印科,任语铮,肖壹夫,等.AI驱动的通感算一体化智能体通信网络架构与关键技术研究[J].移动通信,2025,49(12):20-30,48.
- [2]顾炯炯,李佩珊,曹伟朋,等.面向算力互联网的柔性计算调度机制研究[J].通信学报,2025,46(10):326-338.
- [3]鄢智勇,陈浩,丁立戈,等.面向算力互联网的高可用智算互联互通平台体系架构[J].通信学报,2025,46(9):65-80.
- [4]央广网.我国启动算力互联网试验网建设[J].文化数字化,2025(5):93.
- [5]许丽丽,杨哲建,张文博,等.面向工业互联网的算力网络技术研究及应用[J].邮电设计技术,2024(11):1-6.