文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

异构算力网络中绿色能源感知的算力调度策略

向星宇　周金和
北京信息科技大学信息与通信工程学院
DOI:10.12238/acair.v3i3.15551

[摘　要] 在全球算力需求爆发式增长与人工智能产业高速发展的时代背景下,算力基础设施的能源消耗与碳排放问题已成为制约数字经济绿色转型的关键瓶颈。其中异构算力网络中的资源调度与绿色能源供给的时空错配问题比较突出。基于此,本文聚焦算力网络的绿色化发展需求,提出一种绿色能源感知的算力调度策略。其主要是通过感知任务时延敏感性、计算偏好特征与节点绿能储备状态,构建差异化调度策略,实现异构算力资源与可再生能源的时空协同匹配,为破解算力基础设施高能耗难题提供了兼具理论创新与实践价值的解决方案。
[关键词] 算力调度；异构计算；绿色能源感知
中图分类号：G623.58　**文献标识码**：A

Green Energy-Aware Computing Resource Scheduling Strategy in Heterogeneous Computing Networks

Xingyu Xiang　Jinhe Zhou
School of Information and Communication Engineering, Beijing Information Science and Technology University

[Abstract] Against the backdrop of exponential growth in global computing demands and rapid development of the artificial intelligence industry, the energy consumption and carbon emissions of computing infrastructure have emerged as critical bottlenecks constraining the green transformation of the digital economy. However, the heterogeneous computing power networks face prominent spatio-temporal mismatches between resource scheduling and green energy supply. Addressing the green development requirements of computing power networks, this paper proposes a green energy-aware computing scheduling strategy. By comprehensively perceiving task latency sensitivity, computational preference characteristics, and node-level green energy reserves, the strategy constructs differentiated scheduling mechanisms to achieve spatio-temporal coordination between heterogeneous computing resources and renewable energy. This approach provides a theoretically innovative and practically valuable solution to address the high energy consumption challenges of computing infrastructure.
[Key words] computing scheduling; heterogeneous computing; green energy-aware

引言

全球算力需求呈指数级增长驱动背景下,以GPT系列为代表的大语言模型参数规模与训练算力需求呈现超线性膨胀,2022年全球算力规模达到906 EFLOPS且年均增速超47%,预计2030年将突破20 ZFLOPS[1]。同步发展的中国人工智能产业已进入高速发展期,2023年首批通过备案的8家机构大模型产品正式商用,支撑产业升级的数字基建核心——数据中心规模正以30%年增速扩张。然而,算力基础设施的能耗代价日益严峻：2030年中国数据中心能耗预计达三峡电站5年发电总量,碳排放将突破3.1亿吨,承担着数字经济碳中和的关键攻坚任务[2]。

在此背景下,提升算力绿色化水平具有双重战略价值：一方面,算力作为数字经济新质生产力,每1元投入可撬动3-4元经济产出[3],但当前数据中心平均利用率仅55%的粗放运营模式造成巨大资源浪费；另一方面,通过时空维度优化算力调度策略,实现算力负荷与可再生能源的时空协同,已被证实是兼具经济效益与环境效益的核心路径[4]。本文聚焦算力网络中的绿色调度机制,旨在突破算网资源全局优化难题,通过提升绿色算力利用效率实现“提质增效、降本节碳”双重目标,既为破解AI算力瓶颈提供基础设施支撑,也为数字经济可持续发展提供理论范式与实践路径。

1 算力网络概述

算力网络目的是在网络、存储和计算等多维度资源之间实

现统一管理与资源调配,从而实现算力资源的全局调度。算力网络的整体架构如图1所示,其核心分为四层：基础设施层、资源池化层、任务调度层和网络运营层。

基础设施层中,计算资源由本地计算中心、边缘MEC服务器以及远程云计算中心组成,而通信系统则由光传输网和路由转发节点组成。在算力网络的覆盖范围内,通过通信网络连接的各点计算资源在资源池化层进行统一度量,涵盖CPU资源、GPU资源、存储资源和带宽资源等关键指标。这些指标的度量为资源的高效分配和利用提供了基础。

资源池化层中,每个计算节点的状态信息会通过北向接口实时上报至上层SDN控制器。这些状态信息包括计算节点的电价信息、网络可用性以及当前计算集群的负载情况。SDN控制器通过整合这些信息,为全局资源管理提供数据支持,从而实现资源的动态优化和任务的高效调度。

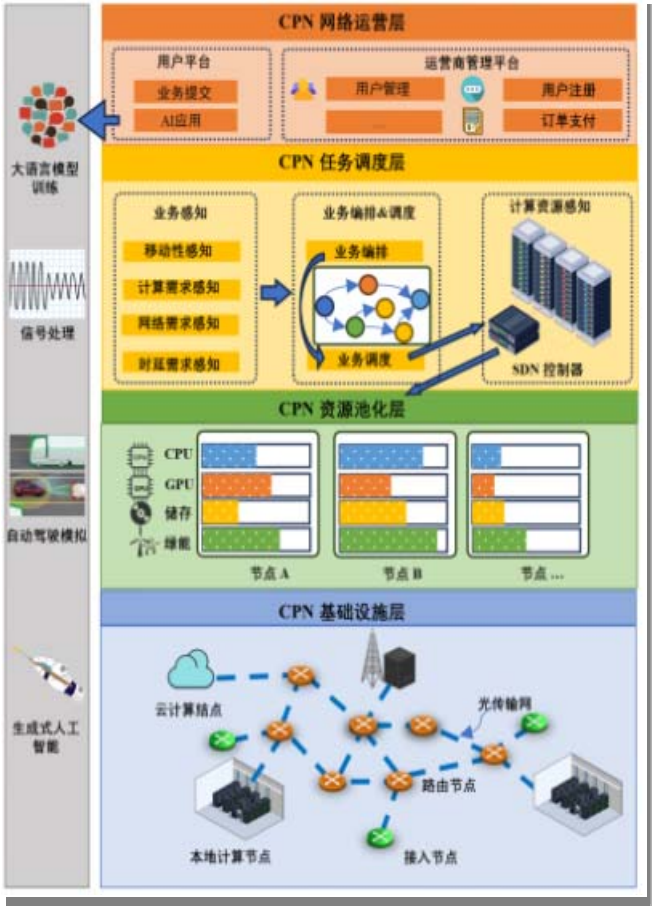

图1 算力网络的整体架构

任务调度层中,当用户平台提交一笔业务时,任务调度层会首先对该业务进行多维感知,包括移动性感知、计算需求感知、网络需求感知以及时延需求感知。其中,时延需求感知特指该业务从用户提交时刻到任务计算完成并返回结果时刻的最大可容忍响应时延。针对可分布式计算任务中多个任务资源竞争的场景,业务编排单元会对当前业务队列进行智能编排,协调不同计算任务和数据流的执行顺序,并将编排结果下发至业务调度单元。业务调度单元从SDN控制器获取算网全局状态信息,结合业务编排结果和实时状态信息,通过预定义策略生成业务部署方案,并通过南向接口对接算力网络基础设施,执行网络配置激活、计算节点容器配置等操作,从而实现算网资源的高效利用与任务调度的优化。

网络运营层通过用户平台向用户提供AI应用及配套的算力服务,算力运营商根据用户对于算力服务的质量要求,将用户的计算任务通过南向接口提交给任务调度层。

2 算力网络中的绿色能源感知调度策略

算力调度是算力网络优化的核心问题,其主要目标是通过合理的资源分配和任务调度策略,优化系统的整体性能,包括任务完成时间、算力资源利用率以及能耗等关键指标。算力调度的核心在于将用户提交的计算任务高效地分配到合适的计算节点上,以满足任务的时延要求、计算需求及能耗约束[5]。

算力网络中算力调度面临多重挑战。首先,算力网络中的计算节点通常具有异构的计算资源(如CPU、GPU、TPU等),如何高效匹配任务的计算需求与节点的计算能力是一个关键问题。其次,绿色能源(如太阳能、风能)的供给具有波动性和不确定性,如何在调度中充分利用绿色能源并降低其消耗,是任务调度的重要目标。此外,任务调度还需在时延和能耗之间进行权衡：时延敏感型任务需要优先考虑任务完成时间,而时延非敏感型任务则可以更多地关注能耗优化。任务调度的优化目标通常是多目标的,主要包括以下几个方面：

(1)时延保障：确保任务在规定时延要求内完成,满足用户的服务质量需求。

(2)能耗优化：通过合理调度,减少系统的总能耗,尤其是棕色能源的消耗,以实现可持续发展。

(3)资源利用率提升：通过合理的任务分配,提高计算资源的利用率,避免资源浪费,从而最大化系统性能。

图2展示了一个典型的算力调度过程。用户通过网络将计算需求提交给算力运营商,调度器根据任务特性和网络时延选择合适的计算节点分配算力资源。计算任务按照算力网络调度器给出的调度方案,传输至相应的节点,执行计算,在计算完成时将计算结果返回给用户。

在任务调度过程中需要考虑如何解决几个关键问题：

(1)任务与节点的匹配：任务的计算偏好(如串行计算密集型、并行计算密集型)需与节点的计算能力匹配,以确保任务的高效执行。

(2)绿色能源的利用：调度算法根据绿色能源供给情况,优先选择绿色能源丰富的节点进行任务分配,减少棕色能源的消耗。

(3)时延与能耗的权衡：调度算法需在时延和能耗之间权衡,确保时延敏感型任务能够按时完成,同时尽量减少系统的总能耗。

文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

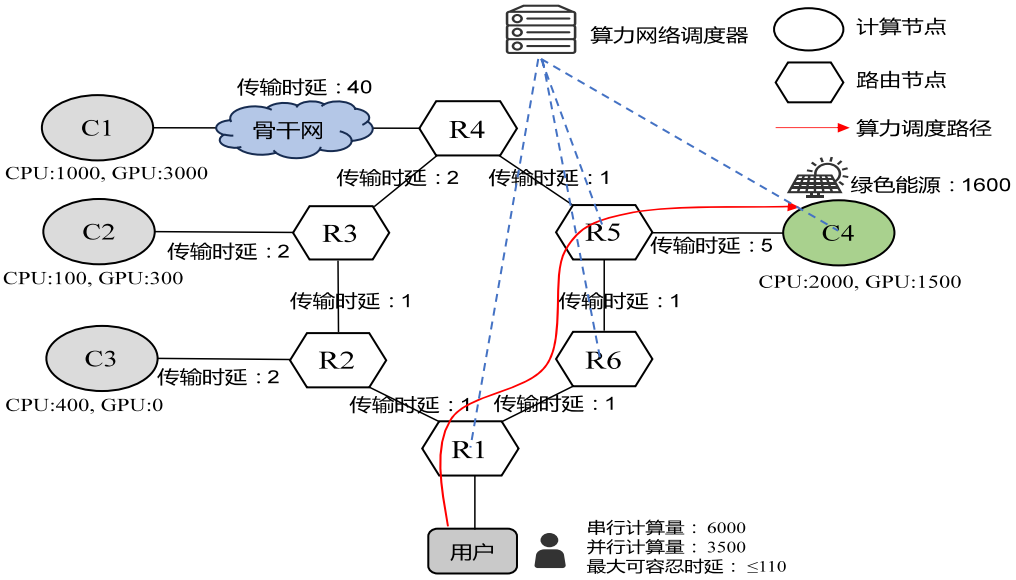

图 2 绿色能源感知的算力调度

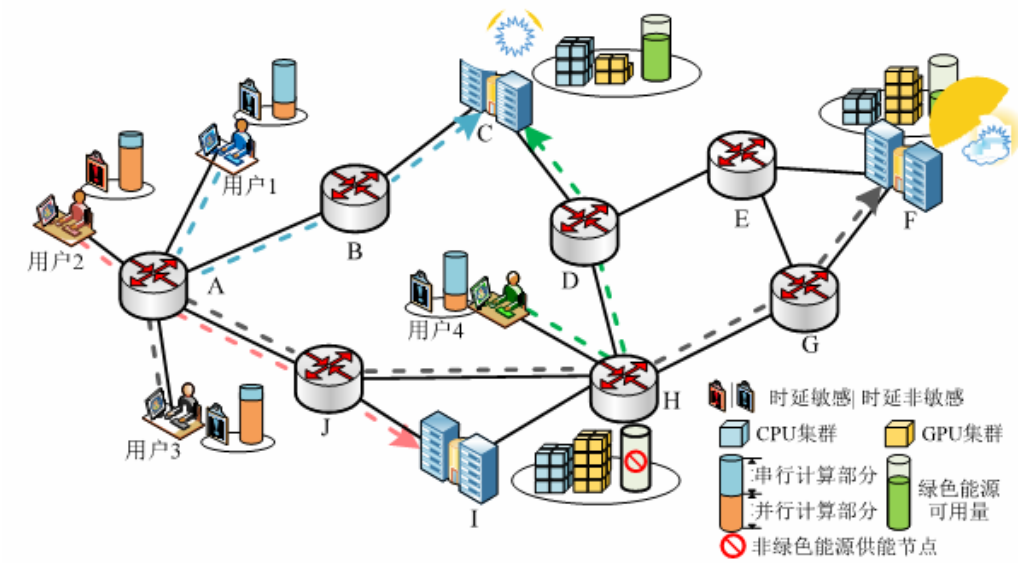

图 3 绿色能源感知的算力调度策略

算力调度在算力网络中具有重要的研究价值，尤其是在绿色能源感知的背景下，如何通过合理的调度策略优化计算任务的能耗和时延，是未来研究的重点方向。针对上述的三个关键问题，本文提出一种绿色能源感知的算力调度(Green Energy-Aware Computing Resource Scheduling Strategy，GECRS)策略，该策略根据计算任务的时延要求和计算偏好，结合算力网络中计算节点的绿能储量信息，平衡计算任务在执行时的能耗与时延。

如图3所示，本文提供一个简短的示例来解释GECRS策略的工作原理。为了便于对比，假设示例中各节点之间的带宽速率和传输损耗相同。算力网络中包含三个计算节点(C、F、H)，每个计算节点由于地理和气候因素的差异，所能利用的可再生能源容量也存在较大差异。此外，各节点的计算资源容量也有所不同。示例中，节点I的CPU/GPU计算单元数量均高于C和F，但由于环境限制，I地区的电力来源主要依赖于棕色能源；而C地区有丰富的太阳能资源，节点C部署的CPU计算单元数量是GPU计算单元数量的两倍。图3还给出了四个计算任务示例，这些任务可以从时延敏感性和计算偏好两个维度进行大致分类：

(1)时延敏感-并行计算密集型(用户2)；(2)时延非敏感-

串行计算密集型(用户1、用户4)；(3)时延非敏感-并行计算密集型(用户3)。

当用户1提交计算任务时,传统的最短路径优先(Shortest Path First, SPF)方法提供了两条路径：路径1：A-B-C,路径2：A-J-I。然而,SPF策略无法根据任务的计算偏好选择合适的计算节点,这可能导致算力网络整体资源利用率的下降。与此同时,由于SPF不具备绿色能源感知能力,无法有效利用算力网络中的可再生能源。而GECRS策略首先会考虑计算任务的时延敏感性,并在保证任务质量的前提下,遵循优先利用绿色能源的原则进行任务调度。对于用户1来说,其计算任务属于时延非敏感-串行计算密集型任务,因此可以优先调度到具备绿色能源供能条件且CPU资源丰富的节点。在示例中,节点I没有绿色能源,而节点F是GPU丰富型节点,与用户1的计算偏好不匹配,因此GECRS策略选择的最佳调度路径为：A-B-C。

与用户1计算类型相似的用户4,位于算力网络中距离节点I最近,且到达节点C和节点F的传输时延和能耗相同。根据GECRS策略的思想,最符合用户4需求的节点仍为节点C,最佳路径为:H-D-C。

用户2属于时延敏感-并行计算密集型任务,GECRS优先选用低时延的调度策略。尽管节点C和I的传输时延相等且都小于到节点F的时延,但节点I拥有最多的计算资源,能够提供最低的时延,因此,用户2的最佳调度方案为：A-J-I。

用户3与用户1在算力网络中拥有相同的起始位置,并且均为时延非敏感型任务。然而,用户3的计算任务是并行计算密集型的,因此节点C相比节点F更能匹配用户3的计算偏好。基于此,用户3的最佳调度策略为：A-J-H-G-F。值得强调的是,尽管与路径A-B-C相比,用户3的调度会产生更多的传输时延和能耗,但这种策略能够避免因计算类型不匹配导致的整体资源利用率下降,GECRS策略能够在保证服务质量的前提下,为算力运营商降低服务成本。

3 结语

综上所述,随着未来算力需求的增速加快,算力基础设施的能耗问题愈发突出。在异构算力环境中如何有效利用绿色能源降低计算能耗,已成为绿色发展的核心目标之一。本文介绍了算力网络架构,并提出一种绿色能源感知的算力调度策略。该策略能够同时考虑任务计算偏好、算力网络中的异构资源与绿能储量,根据任务的时延敏感性差异性制定调度方案,以此实现计算需求与绿色算力供给的时空匹配。

[参考文献]

[1]中国信息通信研究院.中国算力指数发展白皮书[R].北京:中国信息通信研究院,2023.

[2]王永真,唐豪,魏一鸣.中国数据中心综合能耗及其灵活性预测[J].北京理工大学学报(社会科学版),2025,27(2):12-18.

[3]罗茂林.大模型应用推升算力需求GPU租赁业务“钱”景广阔[N].上海证券报,2023-09-23(005).

[4]可思为,董萍,马铭宇,等.考虑风光荷时空互补的多能源绿色数据中心多目标配置方法[J].电力系统保护与控制,2024,52(22):22-33.

[5]刘子腾,向佳霓,李进鑫,等.促进新能源消纳的数据中心负荷优化调度方法研究[J].现代建筑电气,2024,15(10):1-8.

作者简介：

向星宇(1998--),男,贵州贵阳人,硕士研究生,研究方向：绿色算力调度。