 DOI: https://doi.org/10.12349/iser.v6i3.5186

About the construction of heterogeneous computing power scheduling platform and computing power network

Fang Wang[1,2] **Xin Ma**[1,2]

1. China Communications Construction Group Co., Ltd., Beijing, 100073, China
2. Telecom Group Co., Ltd. Henan Branch, Zhengzhou, Henan, 450000, China

Abstract

As a new type of computing resource scheduling mode, the heterogeneous computing force scheduling platform can effectively integrate a variety of different types of computing resources and realize efficient computing task allocation and scheduling. This paper introduces the basic concepts and architectural characteristics of the heterogeneous computing force scheduling platform, analyzes the key technologies and challenges faced, and discusses the construction and optimization methods of the heterogeneous computing force scheduling network. This paper analyzes the cases of application scenarios, summarizes the relevant implementation experience, prospects the development trend of heterogeneous computing power scheduling platform in the future, and provides theoretical support and practical guidance for the further optimization and application of heterogeneous computing power platform.

Keywords

heterogeneous computing power; scheduling platform; computing power network; resource optimization; application case

关于异构算力调度平台及算力网络建设

王芳 [1,2] 马欣 [1,2]

1. 中国通信建设集团有限公司，中国 · 北京 100073
2. 中国电信集团有限公司河南分公司，中国 · 河南 郑州 450000

摘　要

异构算力调度平台作为新型的计算资源调度方式，能有效整合多种不同类型的计算资源且实现高效的计算任务分配与调度。本文介绍了异构算力调度平台的基本概念及其架构特点，分析了异构算力调度的关键技术与面临的挑战，探讨了异构算力网络的构建与优化方法。文章对应用场景的案例进行分析，总结了相关的实施经验，对未来异构算力调度平台的发展趋势进行了展望，为异构算力平台的进一步优化与应用提供了理论支持与实践指导。

关键词

异构算力；调度平台；算力网络；资源优化；应用案例

1 引言

随着信息技术的快速发展，计算需求逐渐呈现出多样化和复杂化的特点，传统的计算平台已难以适应高效处理这些需求。异构算力调度平台应运而生，成为解决这一问题的重要技术手段，异构算力平台借助集成多种类型的计算资源（如 CPU、GPU、FPGA 等），能根据任务需求灵活调度资源，实现资源的高效利用与优化分配，这种平台不仅在提升计算效率方面具有重要意义，还可以在成本控制和系统稳定性等方面带来优势。本文旨在深入研究异构算力调度平台及其算力网络的构建与优化，探讨如何有效整合和调度异构计算资源，以应对现代计算领域中的复杂问题。

【作者简介】王芳（1978–），女，中国河南郑州人，硕士，一级建造师，从事人工智能在信息通信基础设施建设上的技术和应用研究。

2 异构算力调度平台概述

2.1 异构算力平台的定义与特点

异构算力平台是一个集成了多种计算单元的系统，结合不同类型的计算资源（如 CPU、GPU、FPGA、TPU 等）以实现对计算任务的高效处理，异构算力平台的核心优势在于其资源的多样性与灵活性，上述计算资源在性能、处理能力、处理方式以及适应的任务类型等方面都有显著差异。每种计算资源都有其特定的优势，如 GPU 在处理大规模并行计算任务时表现优异，特别适用于深度学习训练和科学计算，FPGA 则能在硬件级别进行数据流处理，适合低延迟、高吞吐量的应用场景，TPU 作为专为机器学习设计的硬件加速器，在神经网络的训练和推理中具备超高效率。平台能根据实时的计算负载和资源状况，自动选择最合适的计算单

元来执行任务，任务调度不仅考虑计算资源的处理能力，还会综合考虑任务的实时性要求、计算复杂度和数据依赖性等因素，在保证计算性能的同时提高资源的利用效率。这种灵活的资源调度能力使得异构算力平台能应对不同类型的计算任务，特别是在处理海量数据、复杂计算和多任务并行时展现出强大的处理能力。

2.2 异构算力调度平台的架构与关键技术

异构算力调度平台的架构通常由计算资源管理、任务调度、数据管理、网络通信和用户接口等多个模块组成，资源管理模块负责监控和管理各类计算资源的状态，保证资源的合理配置与使用，任务调度模块根据任务的计算需求和资源状况，动态选择最合适的资源进行计算，调度策略可能有负载均衡、优先级调度和资源预留等。数据管理模块则负责高效的数据存储与访问，保证任务执行过程中数据的快速传输与同步，网络通信模块提供高效、低延迟的通信手段，保障各类计算单元之间的数据交换顺畅无阻。用户接口模块则提供操作人员与平台交互的接口以便于任务的提交与监控，关键技术有计算资源的动态调度算法、任务的分解与并行化技术、数据流的优化传输技术以及异构计算环境下的容错机制等，上述技术的综合应用。使得异构算力调度平台能在处理复杂计算任务时实现高效、稳定的运行。

2.3 异构算力调度的挑战与解决方案

尽管异构算力调度平台具有一定的性能优势，但在实际应用中仍面临一系列挑战，由于各类计算资源之间的架构差异，如何高效地分配计算任务并利用不同计算单元的性能优势是一大难题。研究人员提出了基于任务特征的调度策略，利用任务的计算复杂度和数据依赖关系动态选择最合适的计算资源，以优化资源的利用率和计算效率，异构平台的资源管理和调度需要在多种计算资源之间进行精确协调，避免出现资源冲突或过载的情况 [1]。解决这一问题的办法是依靠设计高效的负载均衡算法，使得各类资源的使用达到均衡状态，进一步避免出现资源瓶颈或资源浪费的现象，异构算力平台还面临数据传输的挑战。由于计算资源分布广泛且互联复杂，如何高效地传输数据并减少传输延迟成为平台设计中的关键问题，针对这一问题采用分布式存储和高效的数据压缩技术，结合低延迟的网络协议可以有效减少数据传输的瓶颈，并在一定程度上提升系统的整体性能。

3 异构算力网络的构建与优化

3.1 异构算力网络的基本构建框架

异构算力网络的构建框架主要由计算节点、网络拓扑、资源调度层和数据传输层组成，计算节点分为不同类型的硬件资源（如 CPU、GPU、FPGA、TPU 等），以上资源在网络中以分布式的方式部署，根据计算任务的需要动态连接。网络拓扑设计决定了计算资源之间的数据传输路径和延迟性能，为了提升网络的吞吐量和稳定性，通常采用高带宽与低延迟的交换结构，如数据中心内的交换机或者高效的光纤通道连接。资源调度层负责监控和管理网络中各计算节点的负载状况，动态调整任务分配以实现负载均衡和资源的高效利用。数据传输层保证计算任务之间的通信高效可靠，特别是在多种计算资源协同工作时，如何保证数据在网络中快速传递，避免传输瓶颈是该层设计的重点，异构算力网络的构建框架需要协调计算资源、网络拓扑与任务调度，以保障高效且稳定的运行。

3.2 计算资源的动态分配与优化策略

计算资源的动态分配是异构算力网络优化的核心问题，由于计算任务的种类、规模和计算需求差异较大，如何根据任务特性灵活调度资源成为关键。资源分配策略应基于任务的计算强度、数据依赖性及实时性要求进行调度，图像处理类任务可以优先调度 GPU 资源，而密集型数据计算任务则可能更适合使用 CPU 或 FPGA。基于任务的负载预测和性能评估，采用预测性调度方法可有效避免任务拥堵和资源闲置，动态负载均衡策略也是重要的优化手段，借助对当前计算资源的实时监控，调整任务的分配以防止某些节点资源过载，而其他节点资源闲置。为了进一步优化资源利用，可以结合任务的优先级、资源的剩余能力和网络带宽进行智能调度，采用多级调度模型将任务优先级、资源状态和性能需求有机结合，以实现整体系统效能的最优化。

3.3 网络带宽与数据传输的优化问题

网络带宽与数据传输效率在异构算力网络中，是决定整体系统性能的瓶颈之一，由于不同计算节点之间需要频繁交换大量数据，特别是在并行计算环境下，数据传输延迟和带宽限制可能成为影响计算效率的关键因素 [2]。优化网络带宽和数据传输可以从多个方面进行，采用高带宽的网络架构和高效的交换机，可以提升数据传输速率以减少数据瓶颈，借助数据压缩技术减少传输数据量，结合高效的数据编码与解码算法，可以有效降低带宽消耗。数据流的优化传输技术通过动态调整数据传输路径，选择最优传输路径和合适的网络路由算法，以减少传输时间和延迟。网络拥塞控制与流量管理也是优化网络带宽的关键环节，依靠实时监控网络状态并结合流量控制协议，能有效避免网络拥堵、带宽浪费等问题，进一步提高数据传输的稳定性和可靠性。

4 异构算力调度平台的应用实践与案例分析

4.1 典型应用场景与实施策略

异构算力调度平台的应用场景涉及多个领域，特别是在大数据分析、人工智能训练、图像处理、金融计算等领域中具有明显优势，异构算力平台在大数据分析场景下，依靠结合 CPU 和 GPU 等资源，可以高效处理海量数据的并行计算任务，有效提升数据处理能力。GPU 和 TPU 在人工智能训练中，高并行处理单元可以大幅加速深度学习模型的训练过程，降低训练时间。FPGA 和 GPU 在图像处理和视频

编解码等应用中，提供了硬件加速支持使得数据处理能力大幅提高，实施策略上需要对应用场景进行细致分析，评估计算任务的特性、数据依赖性与实时性需求，以此为依据进行资源的合理配置与调度。平台架构需具备高可扩展性和灵活性，可以根据不同的任务需求动态调度计算资源，异构算力平台的实施还需要考虑到网络带宽和数据传输的优化，保证各类计算单元能够高效协同工作，避免因带宽瓶颈导致的性能损失。

4.2 成功案例分析与经验总结

异构算力调度平台在某全球领先的人工智能公司，被成功应用于深度学习模型训练的加速，该公司面临着训练大型神经网络模型时，传统 CPU 平台的计算能力远不能满足需求，引入 GPU 和 TPU 等高并行计算资源，并结合自研的调度算法，成功地实现了多任务的动态调度与负载均衡，有效提高了模型训练的效率，具体数据如表 1 所示：

表 1：传统 CPU 平台与异构算力平台的提升比例

项目	传统 CPU 平台	异构算力平台	提升比例
模型训练时间（小时）	120	30	4 倍
训练成本（美元）	5000	2000	60% 节省
每秒处理样本数	50	200	4 倍
系统资源利用率	65%	90%	38% 提升

从表 1 中可以看出，采用异构算力调度平台后，模型训练时间大幅缩短，训练成本大幅下降，且系统资源利用率有效提升，这表明借助合理的异构资源调度，不仅可以提高计算效率，还能在成本和资源管理方面实现优化。

4.3 持续优化与性能评估方法

异构算力平台的优化不仅限于初期实施阶段，还需要在运行过程中持续进行性能评估与优化，持续优化具体有算法优化、资源调度优化以及数据传输优化等方面。调度算法需要根据系统的实际运行状况进行调整，采用基于任务特性的自适应调度方法，优化负载均衡和任务分配。系统资源的动态调整需要实时监控计算资源的负载情况，依靠引入智能预测模型提前预判资源需求，避免出现资源过度负载或闲置的情况，数据传输优化方面可以借助流量控制和压缩技术，减少传输过程中的瓶颈并且提升数据交换效率 [3]。性能评估方法包含对系统运行中的各项关键指标进行监测，如任务执行时间与资源利用率以及带宽利用率和延迟等，结合实时数据对平台的整体性能进行评估。常用的评估方法具体有对比分析法与回归分析法和 A/B 测试法等，借助以上评估手段可以及时发现平台在不同任务场景中的瓶颈和优化点，进一步为后续的优化提供数据支持与理论依据。

5 结语

本研究围绕异构算力调度平台的构建与优化进行了深入探讨，依靠对异构算力平台的定义与特点和架构的分析，明确了其在高性能计算和人工智能等领域的优势与应用潜力。针对计算资源的动态分配与网络带宽优化，提出了一系列优化策略如智能调度与负载均衡技术，以及基于数据流优化的传输方案，有效提升了系统的计算效率和资源利用率。文章对案例进行分析展示了异构算力调度平台在人工智能领域的成功应用，进一步验证了其高效性和可行性，随着应用需求的不断扩展，异构算力调度平台仍然面临许多挑战，包含复杂的任务调度、数据传输瓶颈和系统的可扩展性问题。平台的优化方向将集中在提高计算资源的自适应调度能力、网络带宽的智能管理以及跨平台协同工作能力的提升，量子计算以及边缘计算等新兴技术的不断发展，使得异构算力平台的应用场景将更加广泛，平台架构也将不断演进以应对更加复杂和多样的计算需求，异构算力调度平台将在未来的技术演进中发挥关键作用，推动各行业计算能力的跨越式发展。

参考文献

[1] 余竞航,赵一辰,王凌,等.基于虚拟网络嵌入的异构算力网络资源管理[J].电信科学,2024,40(10):86-99.

[2] 崔羽飞,毋涛,刘畅,等.面向天地一体场景的异构算力调度与流量卸载应用研究[J].信息通信技术,2024,18(05):48-53+70.

[3] 郭亮,赵精华,赵继壮.异构AI算力操作平台的架构设计与优化策略[J].信息通信技术与政策,2022,(03):7-12.