先进计算产业发展联盟
Advanced Computing Industry Alliance

分布式异构智能算力的管理和调度技术研究报告

研究单位：中国移动研究院、浪潮电子信息产业股份有限公司、新华三技术有限公司

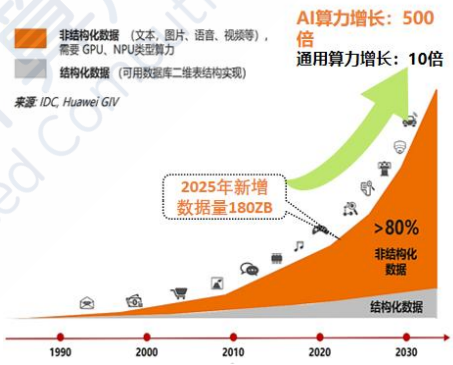
完成日期： 2023 年 12 月

目　录

一、研究背景

随着我国数字经济规模总量的不断攀升，实体经济、数字经济和信息服务的深度融合正加速产业数字化和数字产业化变革。算力作为承载信息数据的重要基础设施，已成为全社会数字化转型的重要基石。根据中国信息通信研究院最新发布的《中国算力发展指数白皮书（2023 年）》显示，至 2023 年我国智能算力规模达到 178.5EFlops，增速为 72%，在我国算力占比达 59%，成为算力快速增长的驱动力；据 IDC 等机构预测，至 2025 年，新增数据量 180ZB，其中 80%的增长来自于文本、图片、语音、视频等非结构化的数据。随着人工智能、元宇宙、高性能计算等领域的发展，激发了更多智能数据处理的需求和场景，对新型智能算力的需求激增。

来源：中国信息通信研究院、IDC、Gartner、TOP500、HPC TOP100



本研究围绕典型智能计算应用对异构算力的协同及调度需求，研究泛在异构算力参与训练或推理过程的协同需求、调度需求，研究泛在异构算力参与训练或推理过程的协同需求，包括异构算力类型、规模要求、性能要求、网络要求、数据传输要求等，分析异构算力协同

的应用场景等特点，考虑同数据中心、跨数据中心、跨云边端多级、池化和非池化异构算力并存等各种场景下，算力协同的需求及可行性。研究分析异构算力资源分类整合、池化重构和智能分配等技术方案。

研究分布式异构算力资源管理技术方案，包括管理跨数据中心、边缘及端侧的 GPU、FPGA 等异构算力设备，已虚拟化或池化的异构硬件，研究对异构算力资源进行标识和监控的方案，对算力进行细力度切分供给的技术方案，研究对计算任务进行异构算力匹配和调度的技术方案。包括如何匹配差异化的计算任务到相应的异构算力节点，如何支持异构算力资源高效和细粒度分配，基于应用场景的负载差异性，建立面向多样化异构算力资源和上层多场景需求的多元异构算力统一调度架构，统一资源实时感知，抽象资源响应和应用调度。研究分布式 AI 框架支持分布式异构算力的管理和调度技术方案。

二、异构算力的发展和应用场景需求

（一）异构算力的发展情况

异构算力通常是指 CPU、GPU、FPGA、ASIC 等多种不同的算力处理体系，能够满足不同场景中的应用需求，实现计算效力最大化。异构算力通常以 AI 芯片的形态被集成在计算机中，AI 芯片是 AI 算力的核心基础设施之一。近年来，面向特定领域体系结构的定制化芯片也不断涌现，已成为 AI 算力发展的主流趋势。

目前异构算力主要有以下类型：

- GPU：

英伟达 GPU 的发展可以追溯到 1999 年，当时英伟达发布了第一代 GPU 架构 GeForce 256，标志着 GPU 时代的开始。随后，英伟达的 GPU 架构不断升级，从 TNT、Rage 到 Geforce 256，再到 Tesla、Fermi、Kepler、Maxwell 等。随着 GPU 技术的不断发展，英伟达的 GPU 架构也不断升级，以适应日益增长的计算需求，GPU 架构也不断推动着图形渲染、人工智能和高性能计算等领域的发展。

近年来，英伟达还发布了多款强大的 GPU 芯片，如 Turing、Ampere 等，这些芯片都具有高性能的计算能力，为各种应用提供了强大的计算支持。2022 年 3 月，英伟达推出了 HGX H100，拥有最高可达 18432 个 FP32（单精度）和 9216 个 FP64（双精度）的 CUDA 核心，辅以 576 个第四代 Tensor 核心。2023 年 11 月，英伟达再次升级其 GPU 产品线，发布了 HGX H200。这款新的 AI 计算平台在原有 H100 的基础上进行了全面升级，主要升级包括提供 141GB 的下一代 HBM3e 内存，这使得 H200 成为了英伟达目前最强的人工智能芯片。

- APU：

APU（Accelerated Processing Unit）中文名字叫加速处理器，AMD 将中央处理器和独显核心做在一个晶片上，它同时具有高性能处理器和最新独立显卡的处理性能，支持 DX11 游戏和最新应用的“加速运算”，大幅提升了电脑运行效率。

从 2010 年以来，AMD 相继推出 GCN 架构、RDNA 架构、RDNA2 架构、RDNA3 架构、CDNA 架构和 CDNA2 架构。最新一代面向高性能计算和人工智能 CDNA2 架构于架构采用增强型 Matrix Core 技

术，支持更广泛的数据类型和应用，针对高性能计算工作负载带来全速率双精度和全新 FP64 矩阵运算。基于 CDNA2 架构的 AMD Instinct MI250X GPU FP64 双精度运算算力最高可达 95.7TFLOPs。

- TPU：

TPU 是由 Google 推出的人工智能芯片 Tensor Processing Unit。之后又陆续推出了 TPUv4 等若干代 TPU 和 TPU Edge。TPU 是计算神经网络专用芯片，是 google 为了为优化自身的 TensorFlow 机器学习框架而打造。

- FPGA：

FPGA 作为一种灵活可编程的硬件平台，具备较高的计算性能和可定制性，能够提供对 AI 算法的加速和优化；在 AI 应用中，可以用于实现神经网络加速器、高性能计算单元等，为计算密集型的 AI 任务提供高性能和低延迟的计算能力。

例如，英特尔 Stratix 10 NX FPGA 就是专门为 AI 设计的，具有 AI 张量块，包含密集的低精度乘法器阵列，针对矩阵和向量乘法进行了调整，可执行 INT4、INT8、Block FP12 或 Block FP16 操作。此外，这些张量块可以级联在一起，支持大型矩阵。

- ASIC：

与更通用的芯片（如 CPU 和 GPU）相比，ASIC 芯片的定制化提供了更高的效率。ASIC 的兴起引起了 NVIDIA、AMD 和英特尔等科技巨头的关注。行业可能会采用混合技术来推动创新和进步。例如，NVIDIA 一直在开发自己的 AI 专用芯片，称为 Tensor Cores。随着亚

马逊、微软和百度等科技巨头探索定制 ASIC，这项新技术显然将在 AI 处理中发挥重要作用。ASIC 领域还持续在可扩展性、可负担性和实施方面开展攻关。

- DPU：

DPU 服务于云计算，主要作用是提升数据中心等算力基础设施的效率，减少能耗浪费，进而降低成本。随着数据中心建设、网络带宽和数据量急剧增长，由于 CPU 性能增长速度放缓，为了寻求效率更高的计算芯片，DPU 由此产生。例如，英伟达将 Mellanox 的 ConnectX 系列高速网卡技术与自己的已有技术相结合，于 2020 年正式推出了两款 DPU 产品 BlueField-2 DPU 和 BlueField-2X DPU 。

（二）异构算力的主要应用场景

异构计算利用不同类型处理器的独特优势，例如 GPU 的并行计算能力和 FPGA 的定制化硬件设计能力，从而提高计算性能和功率效率。它在许多领域都有广泛的应用，如人工智能领域的深度神经网络训练，科学计算领域的模拟和数据处理，物理仿真和计算机视觉等。此外，异构计算还可应用于移动设备和嵌入式系统等领域，在这些领域中，功率和性能都是非常重要的因素。异构计算可以让这些设备更加智能化，同时提高它们的性能和功率效率。总结来看，异构算力的主要应用场景包括：

- 机器学习和深度学习：异构计算可以利用 AI 算力的并行处理能力，加速机器学习和深度学习的训练和推理过程。例如，使用 GPU 进行大规模的矩阵运算，可以大幅提高训练速度和模型准确率。

- 高性能计算（HPC）等科学计算场景：在科学研究、工程仿真等领域，需要处理的数据量巨大，传统的 CPU 计算已经无法满足需求。异构计算可以利用 CPU 和 GPU 联合的方式，实现更高的计算性能和效率。
- 图形处理渲染和游戏开发：异构计算可以利用 AI 算力的并行处理能力，实现图像的实时渲染和处理。例如，在游戏开发中，利用 GPU 卡加速可以实现更加真实的光影效果和更高的帧率。
- 物联网（IoT）：物联网设备数量庞大，需要进行大量的数据处理和管理。通过异构计算，可以实现物联网设备的智能化管理和数据处理，提高物联网应用的效率和可靠性。异构计算可以利用 CPU+GPU 或者 CPU+FPGA+GPU 等异构算力联合的方式，实现更高的计算性能和效率。
- 区块链：区块链技术需要保证交易的安全性和可靠性，同时需要处理大量的交易数据。异构计算可以利用 FPGA 进行加密计算，提高区块链的运算速度和安全性。

除了上述典型的应用场景外，不同行业对异构智能算力的整体需求也呈现差异化分布的特点。

智能算力

来源：中国信息通信研究院、IDC

据信通院与 IDC 的最新统计，由于互联网行业对数据处理和模型训练的需求不断提升，是智能算力需求最大的行业，占智能算力 53%的份额；服务行业由于快速从传统模式向新兴模式发展，算力份额占比位列第二；政府、电信、制造、金融、教育等行业分列第三到八位。

三、分布式异构算力管理和调度的关键技术能力

异构算力多元泛在，对算力的管理平台提出了新的挑战。异构算力管理平台实现多种异构算力的管理和调度，并为智算应用提供应用层的推理和训练技术栈的支持，主要实现以下主要核心能力：

- 动态资源管理：管理 CPU、GPU、FPGA 等异构算力的注册和接入，算力拓扑信息，算力实时状态信息，实现对算力资源的虚拟化和池化的资源重构，提供细粒度的资管管理和隔离；
- 资源调度编排：实现异构算力节点的灵活调度，实现任务与节点资源的灵活编排，多以容器技术基于 Kubernetes 定制化研发实现对任务和资源的灵活编排调度，为上层功能模块提供资源能力；
- 异构算力适配：提供适配异构算力的从底层驱动到应用层框架整

体技术栈的适配支持，以保证应用在不同算力节点上能弹性迁移调度，例如支持不同异构硬件的算子库、编译器、开发工具等；

- 支撑智算的平台能力：基于底层异构算力提供智算应用的数据处理、AI 训练推理框架、模型服务等功能支持。

分布式异构算力的管理和调度是分布式异构算力平台的核心功能，其包括的关键技术主要包括：

（一）异构算力的虚拟化和池化

异构算力虚拟化和池化是指在计算环境中利用不同类型的计算资源（例如 CPU、GPU、FPGA 等）进行虚拟化和资源的池化管理。对于异构资源的虚拟化、池化等资源重构技术方案，将整合硬件资源，形成同类资源池，提高计算资源的利用率和灵活性，从而更好地满足不同应用的需求。

异构算力虚拟化指的是将不同类型的计算资源进行虚拟化，使其能够被多个应用程序或用户共享和管理。这种虚拟化技术可以提高计算资源的利用率和灵活性，比如将 GPU 资源虚拟化供应用程序使用，以满足不同应用对算力的需求。

而池化则指的是将异构计算资源汇聚到一个统一的资源池中，通过统一的管理和调度，按需分配给不同的应用程序或用户。这种池化的方式能够提高整体的资源利用率，降低资源浪费，同时也能够更灵活地满足不同应用对算力的需求。

目前典型的 GPU 虚拟化的技术实现方案包括 MIG 和 vGPU。

MIG（Multi-Instance GPU）作为 Ampere 以及之后的 Hopper 架

构推出的新特性，解决了像 Ampere、Hopper 这种大 GPU 在集群服务应用时一类需求 GPU 切分与虚拟化。MIG 分割的每个 GPU 实例都有完整的独立的内存系统 L2 缓存、内存控制器、DRAM 地址总线等，这样的切分方式也同时以利于容错和吞吐率以及延迟的预测。MIG 的基本方法就是能完成资源的分块+组合，即对物理卡上能用的物理资源进行切分，包括系统通道、控制总线、算力单元（TPC）、全局显存、L2 缓存、数据总线等；然后将分块后的资源重新组合，让每个切分后的子 GPU 能够做到数据保护、故障隔离独立、服务稳定。MIG 可以动态创建和销毁，但是对于没有被分配的 GPU 是无法被使用的。MIG 的资源创建存在两次划分过程，先划分 GI 资源，再划分 CI 资源，这样通过排列组合，增加了配置的多样性。但是这些组合并不是随意的，必须遵循一定的规则，按照 MIG 设定的（profile）进行配置。

基于 vGPU 的虚拟化方案最初由 Nvdia 推出，vGPU 技术允许用户按照规范对 GPU 的计算资源进行切分，就是将一块 GPU 卡的计算能力进行切片，分成多个逻辑上虚拟的 GPU，以 vGPU 为单位分配 GPU 的计算能力，并将单块 GPU 卡分配给多台虚拟机使用，其本质上是通过硬件支持和驱动软件配置的方案，将部分 GPU 暴露给用户。同时为了丰富 GPU 虚拟化的能力，vGPU 也可以支持多种不同的调度机制，使不同的容器可以安全的共享一张物理 GPU，提高 GPU 的利用，例如支持 Round-Robin 调度算法，Equal Share Scheduling 算法，Fixed Share Scheduling 机制等。

智能算力池化的目标是利用软件定义技术，对通过高速无损网络互连互通的 CPU、GPU、AI 芯片等算力资源进行池化整合，实现资源的集中调度、按需分配，使能资源可被充分利用，降低碎片概率，提高总体有效算力。

池化技术下，资源分配方式发生了根本性的变革，软件介入了资源的算力供给，为开启更敏捷的资源管理模式，比如动态伸缩、资源超分等奠定了技术基础，为持续优化智算资源利用率创造了无限可能。池化技术主要通过以下两种实现了软件定义的资源分配：

一是 API 劫持技术：API 劫持技术是目前比较普遍的、针对智能算力的池化技术，它通过劫持对 Runtime API（如 CUDA API）调用实现资源调度。当 AI 应用访问池化运行时的 API 时，则被池化运行时转递至池化服务代理执行，池化服务代理则具备敏捷化的资源管理功能，比如按 1%算力、1MB 缓存的精度细粒度分配资源，实现跨节点远程调用资源等。API 劫持技术的关键在于池化运行时仿真 GPU/AI 芯片的原生运行时，由于 GPU/AI 芯片种类、型号繁多，其原生运行时又相对活跃、升级频繁，仿真工作较为复杂，开发量、维护难度较大。

二是应用程序监视器技术：这是一种完全与 GPU/AI 芯片无关的设备虚拟化和远程处理方法，允许在没有显式软件支持的情况下启用新的硬件体系结构。该项技术通过应用程序监视器工作，该监视器与 Hypervisor 管理虚拟机的方式类似，分为前端、后端，前端监视指定应用程序的活动，拦截至后端处理，后端可以按应用程序申请的数

量分配资源，或将应用程序拆分到多台机器上运行，在保持代码、数据和执行环境一致性的前提下使用这些机器上的智算资源，从而实现资源的细粒度管理、远程调用等资源敏捷化管理功能。

（二）分布式异构算力的调度能力

分布式异构算力的调度将实现底层算力资源与上层应用的匹配，通过节点的动态调度，异构算力节点间的协同，实现分布式异构算力资源使能上层智算应用。

对于跨异构计算节点支撑统一智算应用的调度，依然面临很多技术上的挑战。对于非同质节点的调度，还存在技术上的壁垒问题。由于不同 GPU 等异构硬件在支撑智算应用时，依赖不同的技术栈，包括底层的 CUDA、编译器、前端 AI 框架等，例如运行在英伟达的 GPU 上的应用并不能调度到国产化的 GPU 上无缝运行，也更无法将一个运行在 GPU 上的程序不经过适配改动直接运行在 FPGA 上，技术栈的竖井问题导致一个智算应用目前仍然很难在不同的异构算力节点上无缝迁移，或者同步运行，往往需要对应用本身进行适配和改造才能具备在不同异构算力节点上进行任务调度的前提。产业界也在一致开展跨架构迁移的探索，中国移动提出的算力原生相关技术，能够支撑模型推理在跨异构节点的统一编译，实现不同异构节点的技术栈的拉通，为应用在跨异构节点之间的调度提供了一定的技术基础。

异构算力资源的调度不仅需要考虑异构算力本身的特性，还需要考虑算力资源实时的状态、与算力任务的匹配等。由于当前智算算力集群和资源管理绝大多数以容器和 K8s 的管理体系为主，在异构算力