工程与应用

基于虚拟化的GPU异构资源池平台架构设计、关键技术及应用研究

张万才，张楠，杨文清，王涛，张文强
（国电南瑞科技股份有限公司，江苏 南京 211100）

摘　要：人工智能算力资源面临价格高昂、市场断供等现状问题，传统的单卡单用模式导致资源利用率和使用效率低下，现有的技术研究手段难以支撑多元异构图形处理单元（graphics processing unit，GPU）资源的高效管理和调度。基于此，提出一种基于虚拟化的GPU异构资源池平台，首先对平台总体架构、逻辑架构和功能架构进行了规划设计；其次，对关键技术进行研究，提出了虚拟化异构GPU资源池框架和基于时间切片+负载均衡的调度模型；最后，基于所提方法，提出了多业务单卡叠加、交叉拉远、跨机整合、混合部署和时分复用等多种创新应用模式。所提方法为企业级AI应用提供了可兼容多个GPU不同厂商、支持远程访问、可灵活切分和聚合、可弹性调度的GPU算力资源。经测算分析，同等开发和训练量下，GPU卡数量可节省60%、运行效率可提升4倍。

关键词：GPU异构资源池；算力平台；虚拟化；时间切片；负载均衡
中图分类号：TP391
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024216

Architecture design, key technologies, and application research of GPU heterogeneous resource pool platform based on virtualization

ZHANG Wancai, ZHANG Nan, YANG Wenqing, WANG Tao, ZHANG Wenqiang
Nari Technology Development Co., Ltd., Nanjing 211100, China

Abstract: The current challenges facing the field of artificial intelligence include high prices and market supply disruptions. The traditional single-card, single-use model results in low resource utilization and efficiency. Furthermore, existing technological research methods make it difficult to support the efficient management and scheduling of diverse heterogeneous GPU resources. Based on this, a virtualization-based GPU heterogeneous resource pool platform was proposed. Firstly, the overall architecture, logical architecture, and functional architecture of the platform were planned and designed. Secondly, key technologies were studied, and a virtualization heterogeneous GPU resource pool framework and a scheduling model based on time slicing + load balancing were proposed. Finally, based on the methods described, various innovative application models were proposed, including multiservice single-card stacking,

收稿日期：2024−07−05；修回日期：2024−09−15
基金项目：国家电网公司科技项目（No.524608210272）
Foundation Item: State Grid Corporation Technology Project (No.524608210272)

cross-pull, cross-machine integration, hybrid deployment, and time division multiplexing. The research method proposed provides enterprise-level AI applications with GPU computing resources that are compatible with multiple GPU manufacturers, support remote access, flexible partitioning and aggregation, and flexible scheduling. Following the completion of calculations and an in-depth analysis, it has been demonstrated that a reduction of up to 60% in the number of GPU cards can be achieved while simultaneously enhancing operational efficiency by a factor of four.
Key words: GPU heterogeneous resource pool, computing power platform, virtualization, time slicing, load balancing

0　引言

近些年，人工智能作为数字化的基础能力，在能源电力领域得到快速的应用和发展。国网公司在2021年发布的顶层设计中，提出了人工智能基础平台和组件的年度投资预算达1.35亿元。南网公司也在积极开展人工智能顶层规划设计，并结合电网生产业务开展应用融合深化研究，旨在构建能源数字产业生态。国家能源集团、国家电投等公司也纷纷展开了人工智能整体规划。随着各行业对人工智能应用的日益推广，据评估，仅2023年，全国人工智能市场总体规模预期超过1 500亿元[1]。

面对快速的发展趋势和庞大的市场需求，图形处理单元（graphic processing unit，GPU）作为人工智能重要且基础的算力资源，却由于产品价格暴涨等因素，出现智能化建设无法持续、有效开展的问题[2]。与此同时，为了支撑人工智能行业应用的广泛落地及企业业务的顺利开展，当前GPU技术及其应用模式面临着以下问题亟待解决。

（1）算力资源昂贵且利用率低，目前已知GPU算力卡成本可占服务器成本的80%以上，然而，传统的一卡一应用的使用模式导致实际利用率并不高。如何在有限的算力资源条件下有效提升其利用率，已成为目前研究的主要方向[3]。

（2）近年来，国产芯片自给率不断提升，但同时也面临与国外英伟达（NVIDIA）和超威半导体公司（AMD）等高端GPU进行集成的挑战。在多芯异构模式下，如何构建多元算力应用体系，实现不同计算架构的智能算力和通用算力协同，以低成本、高质效地支撑人工智能应用的需求，也是一大挑战[4]。

针对上述背景和现状，围绕异构算力资源的高效利用，相关领域已开展了诸多研究。例如，文献[5]基于小颗粒度资源度量与异构调度技术，从宏观角度提出了构建“东数西训”算力网络的建设思路，但未能就GPU资源如何池化、异构算力如何调度开展深入研究。文献[6]提出了采用负载均衡等技术，设计了基于Kubeflow的异构算力调度框架，提高了并行任务的执行效率，但未涉及多元多芯异构算力资源的调度管理。综上所述，当前急需一种能兼容多元异构GUP算力资源并具备高效调度能力的框架或模型，以应对现有的形势和挑战。

基于此，本文提出了基于虚拟化的GPU异构资源池平台，一是对所述平台的主要架构进行设计，实现算力资源弹性伸缩、动态调用和释放等功能，支撑人工智能应用及业务开展；二是对其关键技术展开研究，提出了基于虚拟化技术的异构GPU池化框架以及基于时间切片+负载均衡的虚拟图形处理单元（virtual GPU，vGPU）调度模型，实现了异构GPU资源池构建和调度管理；三是对比常规模式，创新性提出了混合部署、时分复用、单卡叠加、交叉拉远等应用方法，并分析了资源利用率、使用效率等相关的成效价值。本文所述平台，实现了对主流厂商异构GPU算力资源的高效纳管和调度，有效提升了资源利用率，缓解了当前形势下算力资源供需紧张等问题，为企业人工智能业务开展提供了支撑。

1 研究背景及思路

1.1 研究背景

人工智能技术的快速发展对人工智能算力提出了规模化建设的需求。一方面，人工智能算法愈发复杂，模型规模不断提升，图片、语音、视频等非结构化数据爆炸式增长；另一方面，人工智能与5G、物联网等行业领域结合落地，对算力资源的需求呈现指数级增长。随着智能业务的发展，以国家电网公司为例，人工智能覆盖了其智能安监、规划、运行、调度等业务领域。因此，建设异构算力平台已成为支撑和引领数智化转型的重要基础设施和关键发展路径[7]。

算力是人工智能发展的重要因素之一，大规模深度学习模型的开发与训练带来的高昂成本，导致大量业务或部门难以持续获得这种算力研发条件，成为制约企业人工智能技术发展和应用的沉重负担[8]。同时，传统的算力资源通常按一卡一应用的模式进行使用，导致使用效率低下。以2021年Facebook对其机器学习负载的分析结果为例，大量的GPU算力被闲置浪费，真正的利用率不足30%[9]。因此，建设企业统一的算力平台，打破部门隔阂，消除算力孤岛，提高算力资源的共享能力和利用率，为全场景人工智能业务提供普惠的算力资源成为当务之急[10]。

算力类芯片产品主要包括中央处理器（central processing unit，CPU）、现场可编程门阵列（field programmable gate array，FPGA）和GPU，参考《中国算力白皮书（2022年）》，并结合笔者调研统计，算力资源头部企业中，英特尔占据了全球71%的FPGA、84%的CPU市场份额；英伟达占据了全球95.7%的GPU市场份额。近些年，寒武纪、昆仑芯、中科海光等国产芯片的自给率不断提升[11]。预计到2025年，智能芯片的国产化自给率将达到70%以上。当前的国际形势不仅加速了国产芯片的发展步伐，也给企业的数据中心尤其是不同厂商品牌的多元算力资源优化配置、不同计算架构的智能算力与通用算力协同发展提出了新的要求和挑战[12]。

1.2 研究思路

GPU作为重要的AI算力资源，其使用方式与CPU存在较大不同。CPU时刻在运行，而GPU作为附加在计算机中的设备，只有在需要时才被运行，因此，GPU宜采用动态分配、按需调用的原则进行调度和管理[13]。同时，为实现对异构资源的兼容纳管，本文提出采用虚拟化技术，通过增加软件层将GPU异构资源虚拟化，按照一定颗粒度定义，并将虚拟化后的vGPU按照设定的颗粒度按需分配给AI应用，实现了与物理GPU进行解耦，从而优化资源分配与管理，简化异构资源的使用难度，提高了物效与人效[14]。软件定义异构AI算力资源池示意图如图1所示。

本文所述方法具备如下优势。

（1）利用率高。支持将GPU切片为任意大小的vGPU，从而允许多AI负载并行运行，提高物理GPU利用率，GPU综合利用率提升3～10倍。

（2）弹性扩展。对数据中心所有异构GPU服务器进行池化管理，并利用远程直接内存访问（remote direct memory access，RDMA）技术，包括“无线宽带”（infiniband，IB）或基于以太网的远程直接内存访问协议（RDMA over converged ethernet，RoCE），以及TCP/IP网络进行分布式部署和连接调研，可轻松实现资源池的弹性、横向扩展。

（3）全局管理。为运维人员提供了全局性的资源调度管理策略，增强了资源池性能监控能力[15]。

2 异构算力资源池平台架构设计

2.1 总体架构

本文所述异构算力资源池平台的核心设计目

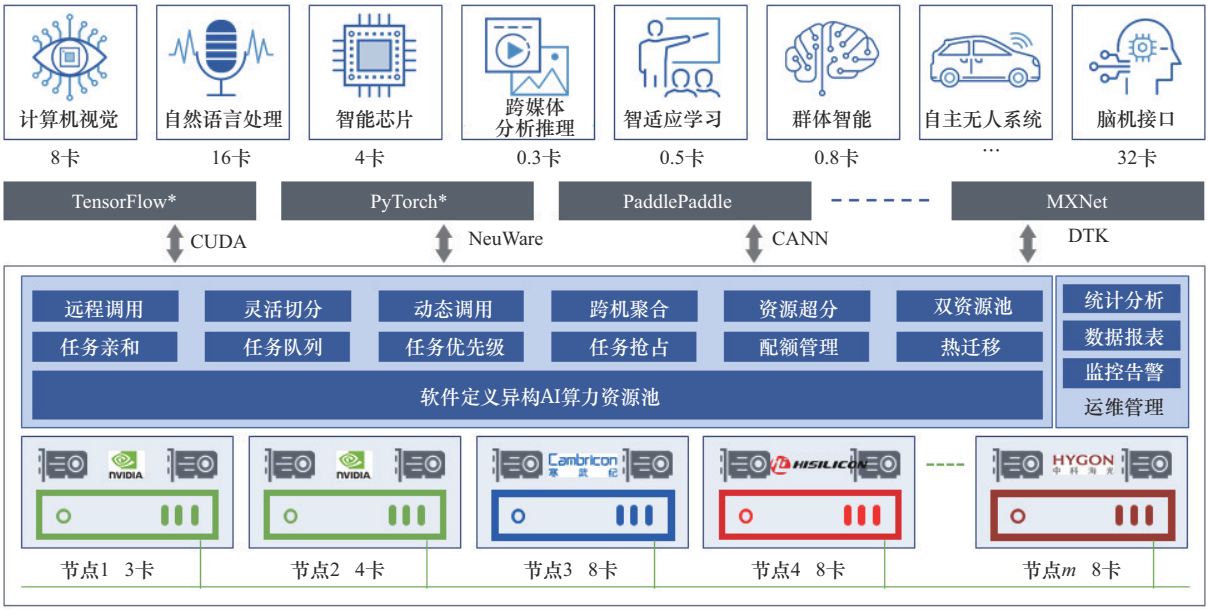

图1　软件定义异构AI算力资源池示意图

标：一是对英伟达、寒武纪等国内外多厂商品牌或计算架构的兼容及池化，实现“一池多芯”；二是对公/私有云、虚拟机和裸金属的纳管支持，实现“一池多云”[16]。异构算力资源池总体架构如图2所示，分为API层、池化调度层和GPU池化引擎层，具体说明如下。

（1）GPU池化引擎。该层包括算力切分、显存切分、资源聚合、动态挂载/调整/释放、远程调用等组件，采用软件定义技术，实现了对池化后的算力资源切分、挂载、聚合、释放等全生命周期流程管理。

（2）池化调度层。该层包括本地优先节点/设备紧凑、节点/设备均衡、任务排队、负载均衡等调度算法，按任务优先级实现对资源池化后的资源调度管理。

（3）API层。该层包括用户API、服务器API、GUI API、资源池API，为底层引擎和上层应用提供标准化调用接口。

通过软件定义技术构建的资源池，兼容并支持容器节点、虚拟化节点和裸金属，面向数据中心运维人员，提供了全局资源管理的技术条件[17]。

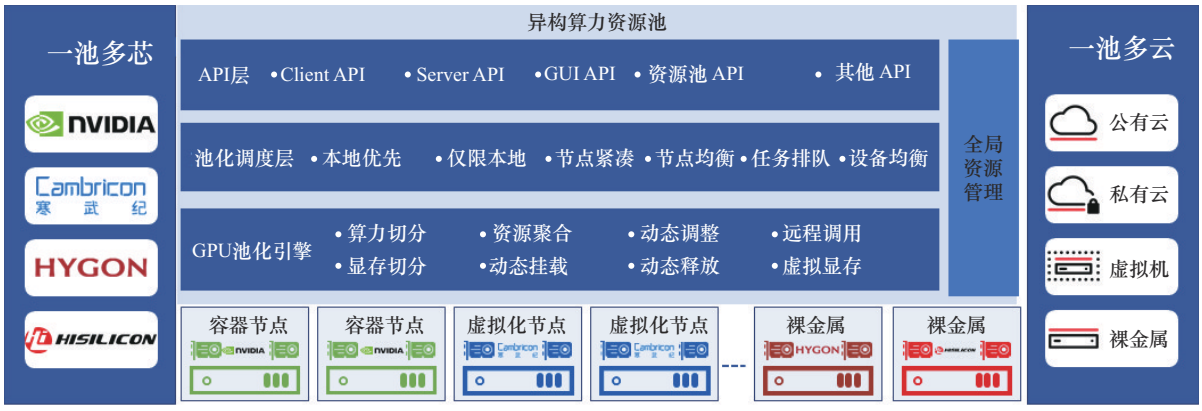

图2　异构算力资源池总体架构

2.2 逻辑架构设计

本文所述异构算力资源池平台的逻辑架构如图3所示，包含了GPU控制器、GPU服务器、客户端运行时组件和图形用户界面（graphical user interface，GUI）等功能组件。各功能组件可根据用户环境需求部署在单服务器上，或者分布式部署在多服务器、虚拟机或容器等云环境中。在分布式环境中，各功能组件可通过多类型的网络建立连接，从而将数据中心的GPU资源管理起来，形成一种可以被全局共享的计算资源，为AI应用提供可远程访问的、可灵活切分的、可聚合的弹性GPU算力[18]。各功能组件的具体说明如下。

（1）GPU控制器组件。通过本地直连或网络连接其他所有功能组件，并在组件中部署调度算法，实现资源统一调度和管理。

（2）GPU服务器组件。该组件部署在GPU服务器节点上，负责物理GPU资源的发现和管理，并通过API和运行时库，将GPU计算能力提供给各业务应用。

（3）客户端运行时组件。该组件通过封装Nvidia CUDA、AMD HIP等主流GPU厂商的编程环境和运行环境，提供应用自动调用，实现异构GPU的兼容管理。

（4）GUI组件。该组件为运维人员提供友好易用的GUI界面，方便管理员对整体资源池进行全面管理。

为支持大部分流行的AI框架，如TensorFlow、PyTorch、MXNet和PaddlePaddle等，本文所述的异构算力资源池，通过封装或模拟CUDA标准接口，为各种AI应用提供与Nvidia CUDA SDK接口功能一致的运行环境，从而使得AI应用透明无感地运行在GPU资源池之上，并通过分布式部署各功能组件，提供分布式的CUDA运行环境[19]。

2.3 功能架构设计

为实现对异构GPU计算资源的池化管理，支持AI视觉、自然语音处理等企业AI应用，对功能架构进行设计，异构算力资源池功能架构如图4所示，包括资源池化、资源调度和运维管理，具体设计说明如下。

（1）资源池化。通过软件定义的方式，将异构资源进行池化以支撑AI应用是平台基础能力。资源池化功能集包含了算力切分、显存切分、远程调用、交叉拉远、资源聚合等功能组件。

（2）资源调度。采用虚拟化、容器化及远程调用等技术，以实现池化后的资源高效率、可弹性化地调度，具体包括本地调用、本地优先、任

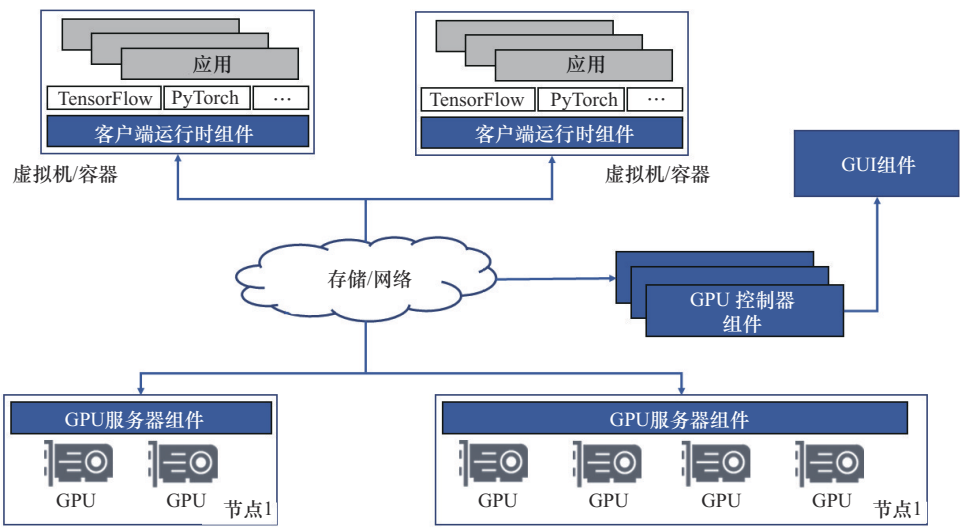

图3 异构算力资源池平台的逻辑架构

图4　异构算力资源池功能架构

务排队、任务抢占、资源回收、指定GPU节点和型号等功能组件。

（3）运维管理。将异构资源池化，有助于对资源分配及使用等情况进行全景监测，按照可配置的资源分配策略实现集约化管理，具体包括节点管理、热迁移、灰度升级、GUP热插拔、高可用等功能组件。

3　关键技术

3.1　异构GPU资源池构建

虚拟化技术是构建异构GPU资源池，对英伟达、寒武纪等国内外多厂商品牌或计算架构进行兼容及池化，实现“一池多芯”，进而实现异构算力资源统一纳管的关键技术。常规的GPU虚拟化技术有API远程调用、GPU直通和完全虚拟化等，GPU虚拟化常规技术优劣势分析见表1[20]。

由于常规技术存在GPU切分或超分、异构资源支撑能力不足、远程调用效率不高等问题，本文提出了一种基于虚拟化技术的异构GPU资源池化框架，对GPU各厂商的服务器进行整合，对不同异构GPU厂商驱动和软件库进行集成，将各计算节点的异构GPU资源进行池化管理，通过封装并调用CUDA API、HIP API等主流厂商接口和运行时库，实现了异构GPU算力切分和调用，进而实现了GPU资源共享。本文所述GPU资源池框架设计示意图如图5所示，主要包含GPU客户端、GPU服务器端和GPU控制器端3个主要组件，具体说明如下。

（1）GPU客户端主要负责按照AI训练和应用需要发起创建、调用和释放vGPU资源请求，为实现不同厂商的异构GPU调用支撑，以VIDIA和AMD为例，客户端端对其原生的CUDA API和HIP API接口进行了封装。

（2）GPU控制器端主要通过GPU监控和GPU调度模块，对资源池中的所有vGPU资源进行管理，GPU监控模块负责监控GPU利用率、网络负载及其状态，GPU调度模块配置了负载均衡算法，对vGPU资源进行最优调度。

（3）GPU服务器端部署在资源池节点GPU服务器上，集成了各主流厂商的硬件驱动程序和运行时库，如CUDA lib和HIP lib，使用不同进程执行响应GPU客户端的vGPU资源调用API请求，满足其资源的调用服务，并在资源使用完后进行回收和释放。

表1　GPU虚拟化常规技术优劣势分析

常规技术	优势	劣势
GPU直通	性能和兼容性好	无法资源切分
完全虚拟化	支持GPU资源切分，性能损失少	GPU实现能力和方式不一
API远程调用	资源使用灵活	性能存在瓶颈

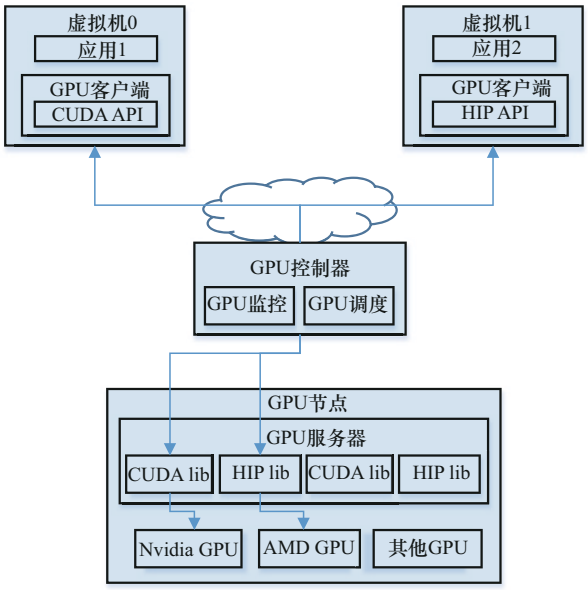

图5　GPU资源池框架设计示意图

异构GPU资源池构建及调用管理流程如图6所示，主要步骤如下。

（1）设备注册。当有新的计算节点或GPU物理资源接入或开机、移除或关机时，物理设备向GPU服务器发送设备注册消息，消息中附带设备的显存和算力大小、设备型号、协议栈等参数信息。

（2）定时轮询设备状态。GPU控制器中的监控模块会定期轮询各计算节点的设备状态，包括资源利用率、使用时长、是否处于正常状态等，并监控网络负载。

（3）发送设备及其参数信息，更新资源链表。GPU服务器收到设备状态轮询指令后，将各GPU设备信息及参数发送给GPU控制器，并构建资源链表。

（4）发送vGPU创建请求。GPU客户端按照应用需求向GPU控制器发送vGPU创建请求消息，包含vGPU数量、显存和算力大小、设备型号、协议栈以及调用模式等设备参数。vGPU的创建，同时也是对一块物理GPU卡按照指定的参数进行切分的过程，算力切分的最小颗粒度原则上为原物理GPU算力的1%，显存切分的最小颗粒度为1 MB。

（5）选址最优匹配资源。GPU控制器收到请求后，在所管理的GPU资源池中，根据调度模块

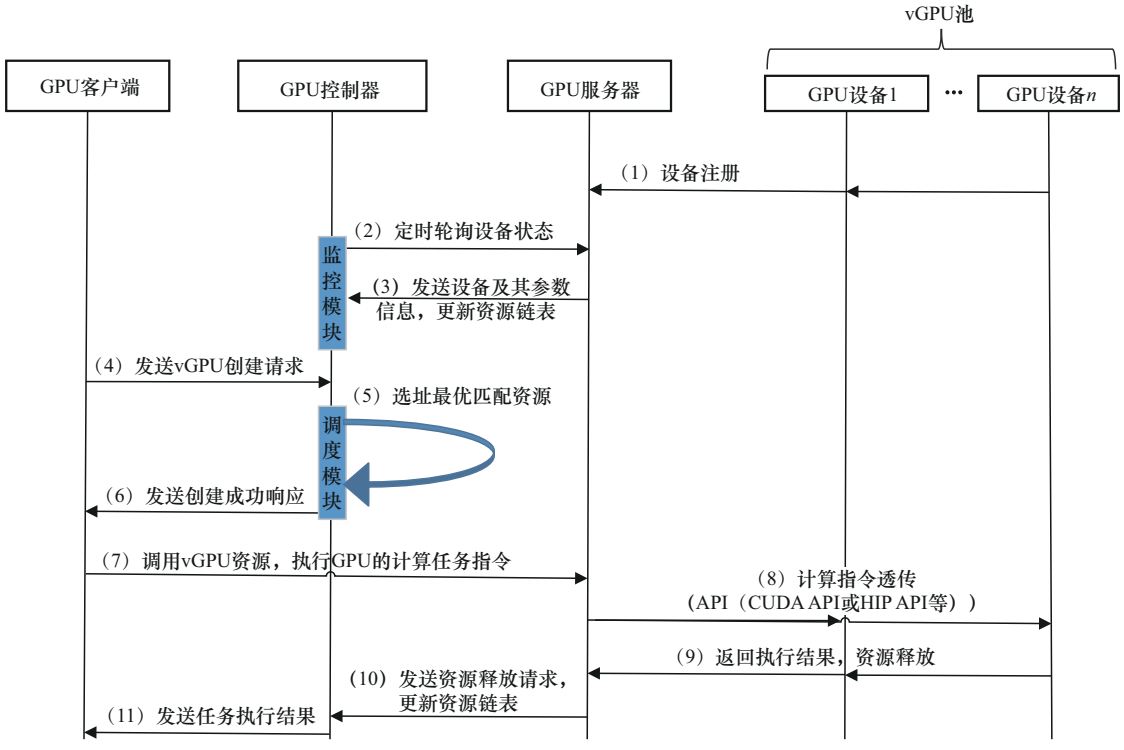

图6　异构GPU资源池构建及调用管理流程

选择满足参数的最优匹配计算节点，并分配vGPU给GPU客户端。资源池中的各计算节点由各本地或远端的服务器组成，每个服务器内建议配置同厂商的GPU卡，不同的服务器可根据实际情况或需求配置不同主流厂商的GPU设备。

（6）发送创建成功响应。GPU控制器完成vGPU资源创建，并返回GPU客户端资源创建成功响应，响应消息包括可调用的GPU资源所在计算节点IP地址及vGPU设备ID。

（7）调用vGPU资源，执行GPU的计算任务指令。vGPU资源创建成功后，GPU客户端可通过封装的API，如CUDA API或HIP API，直接调用GPU服务器集成的GPU设备驱动和运行时库，如CUDA lib或HIP lib，并通过不同进程执行GPU客户端的调用请求。

（8）计算指令透传。GPU服务器通过API（CUDA API或HIP API等），透传计算指令，并直接调用设备底层接口，执行物理GPU的计算指令。

（9）返回执行结果，资源释放。计算节点完成物理GPU计算任务后，向GPU服务器返回执行结果，并将已创建的vGPU资源进行释放。

（10）发送资源释放请求，更新资源链表。GPU服务器向GPU控制器发送资源释放请求，并更新资源链表各资源及计算节点状态。

（11）发送任务执行结果。GPU控制器向GPU客户端发送任务执行结果，本次AI任务执行结束。

3.2　异构GPU资源池调度

为提高资源利用率，本文提出了基于时间切片+负载均衡的vGPU调度模型。时间切片通常指将GPU的计算时间进行划分，按照时间片段分配给不同的用户，实现多用户共享同一个物理GPU的技术。通过时间切片调度，多个用户可以同时访问同一块GPU，并获得相对独立的虚拟图形处理单元。这种虚拟化方法旨在提升资源利用效率，降低物理设备成本，并保持用户间的良好体验。

基于时间切片的vGPU调度策略需要均衡地分配GPU资源，并确保用户任务能够得到及时响应，通常应遵循以下原则：（1）公平分配原则。为每个用户分配相等的时间片，确保公平性和公正性。（2）优先级调度原则。根据用户任务类型和优先级，对时间片进行动态分配。对于需要低时延的任务（如游戏），可以分配更多的时间片。（3）任务切换开销最小化原则。减少任务切换的开销，提高系统吞吐量。

本文提出了资源和负载均衡的调度优化策略，即通过实时监测GPU的负载和资源利用率，设置vGPU权重，并根据权重值动态调整时间片的分配，从而实现资源的最优利用。当GPU的负载较低时，可降低vGPU权重值，并将时间片分配给其他用户，确保按照负载变化情况为vGPU分配合理的时间片资源，从而提高资源利用效率。资源调度流程设如图7所示。

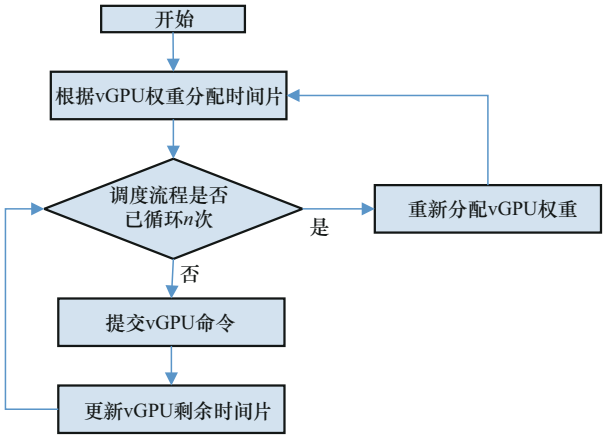

图7　资源调度流程

算法说明如下。

（1）时间片计算。当GPU控制器收到vGPU创建请求时，需要从当前计算节点及GPU资源链表中，根据vGPU创建参数完成资源匹配，计算待创建的$vGPU_i$所需要分配的时间片T_s，计算式如下。

$$T_s=16\ \text{ms}\times W_i/(W_1+W_2+W_3+\cdots+W_n) \quad (1)$$

其中，$W_1,W_2,\cdots,W_n$分别为$vGPU_1,vGPU_2,\cdots,vGPU_n$的权值，通常图片卡顿的时间为16 ms，故将

其设为时间片的最大刻度值，即给$vGPU_i$分配的时间片值需要小于16 ms。

当$vGPU_i$被执行调用时，记录该时刻为T_1，即视为其获得调度权；当$vGPU_i$完成计算任务被释放时，记录该时刻为T_2，即视为其被回收调度权。剩余时间片计算式如下。

$$T=T_s-(T_2-T_1) \tag{2}$$

（2）重新分配权重。由于计算任务、网络及硬件资源状态存在非线性或瞬态变化，vGPU负载执行计算任务存在波动性，本文提出在vGPU资源池完成n次调用/释放次数后，由调度器记录每次vGPU的工作时长以及剩余时间片，统计n次实际耗费时间情况，从而计算出总剩余的时间片T_{rem}。当次数达到n次时，收到资源调用请求的$vGPU_i$需要对其新权重进行重新分配，计算式为：

$$W_{new}=W_i-W_{sub} \tag{3}$$

其中，W_{sub}为根据总剩余时间所计算的可回收权重，计算式为：

$$W_{sub}=T_{rem}/(T_s\times n) \tag{4}$$

当完成本次权值再分配后，需要将其平均分配给剩余vGPU，在下一个n轮调用时，根据其负载情况按上述过程进行时间片动态分配。

上文所述权重重分配算法如算法1所示。

算法1 权重重分配算法

输入 vGPU; //虚拟GPU设备

n=10; //执行轮次

输出 无

//参数设置

Wsub=0；

Num=0； //无剩余时间片的vGPU数量

//分配vGPU时间片

For (vGPUεvGPUs) {

trem = vGPU.T_{rem};

ts=vGPU.Ts; //分配时间片

if (trem) //若时间片存在剩余

Wsub += trem/(ts*n); //可回收的权重

else //无剩余时间片

Num++;

}

//剩余权重再分配

For (vGPUεvGPUs) {

If (!vGPU.Trem)

Vgpu.W += Wsub/Num; //剩余权重平均分配给负载大的vGPU

}

需要说明的是，vGPU.是同一物理GPU分配的系列虚拟vGPU设备，并配置了剩余时间片及权重等参数信息。

4 应用模式及成效

本文基于所述方法在应用模式上进行了相应创新，在资源利用率、使用效率等成效价值方面得以提升。

（1）动态调用、释放及弹性伸缩应用，如图8所示，常规模式下，应用与物理GPU卡静态绑定。经统计分析，平均每GPU卡峰值利用率21.8%、平均利用率10%，采用本文所述方法构建GPU资源池，可实现对不同服务器厂商的物理GPU卡统一纳管，打破了GPU资源孤岛，并根据不同时间或业务量设计不同策略，可实现按需动态配置资源比例，实现了资源动态调用、释放和弹性伸缩，提高了GPU资源利用率，同等开发和训练量条件下，卡数量可节省60%、运行效率可提升4倍。

（2）混合部署和时分复用应用，如图9所示，常规模式下，训练、开发任务各自独占整张GPU卡，每张卡一次只能执行一个任务，通过GPU资源虚拟化，可支持训练和开发任务混合部署到单张物理GPU卡，并同时处理多个任务，单卡业务处理数量翻倍。同时，训练任务可利用开发任务“闲时”进行，不需要独占一整张卡资源，节省了GPU物理卡的配置数量，以某中型企业为例进行实际测算，数据中心年平均节省GPU数量40%以上。

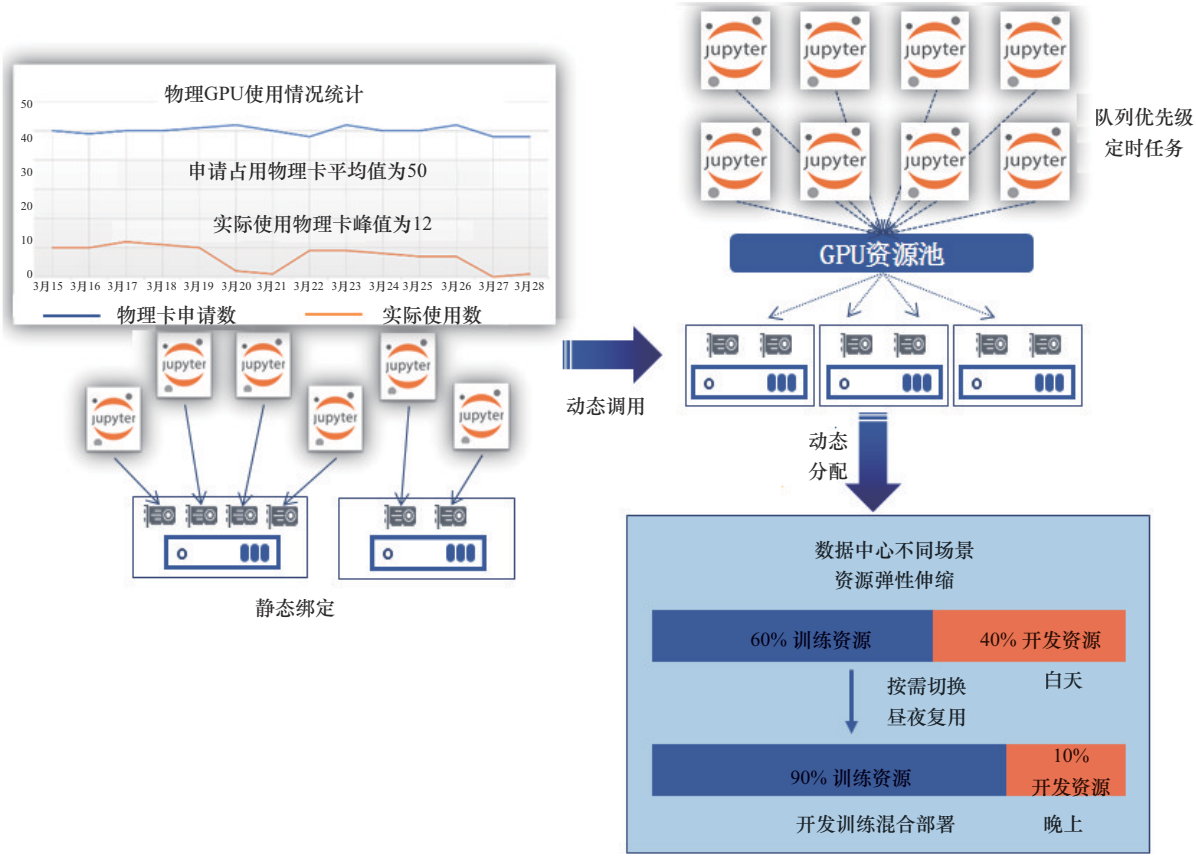

图8　动态调用、释放及弹性伸缩应用

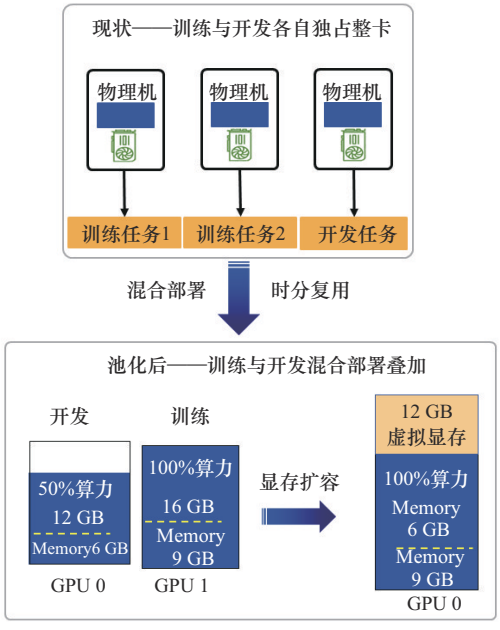

图9　混合部署和时分复用应用

（3）化整为零、多业务单卡叠加应用，如图10所示，常规模式下，为避免业务干扰和争抢，不同推理任务在各自物理GPU卡上独立运行，导致资源利用率低下。采用本文所述方法，可实现物理GPU资源“化整为零”切分为多vGPU虚拟资源，支持多模型并行运行在同一张GPU卡上。同时，通过对各任务进程进行封装和隔离，保证了业务安全性，减少了各业务间的干扰。按最低细粒度的算力1%、显存1 MB进行资源切分，GPU整体利用率提升3倍以上。

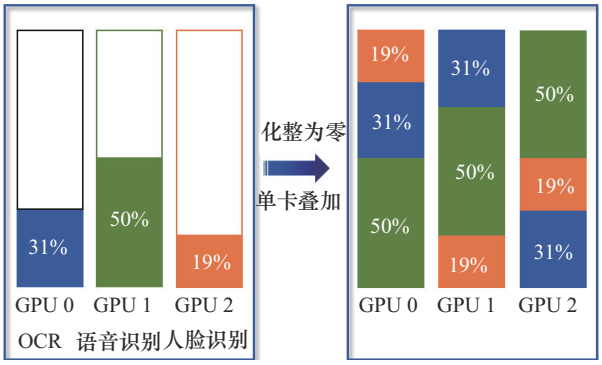

图10　化整为零、多业务单卡叠加应用

5 实验与仿真

5.1 测试配置

本文进行了与性能相关测试，以分析所述方法构建的vGPU资源池算力利用率和运行效率，与常规物理GPU卡绑定模式的差异情况。

（1）测试配置。配置硬件服务器3台，型号为DELL PowerEdge R720。其中一台用于安装本文所述资源池平台软件，另外两台用于配置GPU计算节点。每台GPU计算节点服务器各配置同型号GPU卡两张：NVIDIA英伟达Tesla T4、摩尔线程MTT S80。服务器及GPU硬件配置情况见表2。

（2）软件配置。KVM环境为libvirt 1.3.0 qemu 2.5.0；NVIDIA显卡驱动为418.43；CUDA为9.0；cuDNN为7.4.2。

（3）测试用例配置。TensorFlow v1.12、官方Benchmark、使用Synthetic数据集，测试过程中选用主流机器学习模型：VGG16、ResNet50、ResNet152、InceptionV3。

5.2 测试方法

为评价GPU算力资源池处理性能，本文采用imgs/sec作为实验评测指标，imgs/sec作为衡量图像处理性能参数，指的是每秒处理的图像数量，常用于评估算法或系统处理图像的速度。该参数反映了系统在单位时间内能够处理的图像数量，是衡量图像处理效率的重要指标。例如，在深度学习框架的性能测试中，imgs/sec被用来比较不同框架或不同配置下的图像处理速度。此外，在数据加载和处理的性能测试中，imgs/sec也用来衡量数据加载和处理的速度，这对于实时图像处理系统或需要快速响应的应用尤为重要。例如，在搜索结果中提到的Cifar10数据集的性能测试中，通过调整数据处理进程数和采用不同的数据加载格式（如MindRecord格式），可以观察到处理速度的变化，从而评估不同方法对图像处理性能的影响。这些测试展示了如何通过优化数据处理和加载方式来提高imgs/sec，进而提升整体系统的效率。

实验过程中，加载测试集中的VGG16、ResNet50、ResNet152、InceptionV3这4类机器学习模型进行推理训练模拟，并分别记录常规物理GPU卡绑定模式和GPU资源池模式下imgs/sec指标的变化。这4类模型的加载过程进一步细分为单独加载和混合加载两种方式。

5.3 测试结果

实验测试结果如图11所示。经验证分析，单独加载下，GPU资源池模式较常规物理GPU卡绑定模式imgs/sec平均提升15%；混合加载下，资源池调度算法效果发挥最大，imgs/sec平均提升达到40%。

实验结果表明，本文所述GPU资源池针对机器学习的图像处理性能和效率具备较好的提升作用。

表2 服务器及GPU硬件配置情况

配置项目		参数
服务器配置	CPU架构	2×(Intel)Xeon E5 2620 Six-Core
	CPU频率	12×2.0 GHz
	内存大小	20 GB DDR3
	磁盘大小	2×1 TB
	网络接口卡	Intel quadport Gigabit network adapter
GPU配置	NVIDIA英伟达Tesla T4	显存频率16 000 MHz，FP32浮点性能8.1TFLOPS，INT4浮点性能最高260TFLOP，16 GB GDDR6显存，带宽320 GB/s
	摩尔线程MTT S80	显存频率1.8 GHz，单线程精度达到14.4TFLOPS，16 GB GDDR6显存，位宽256 bit，带宽448 GB/s

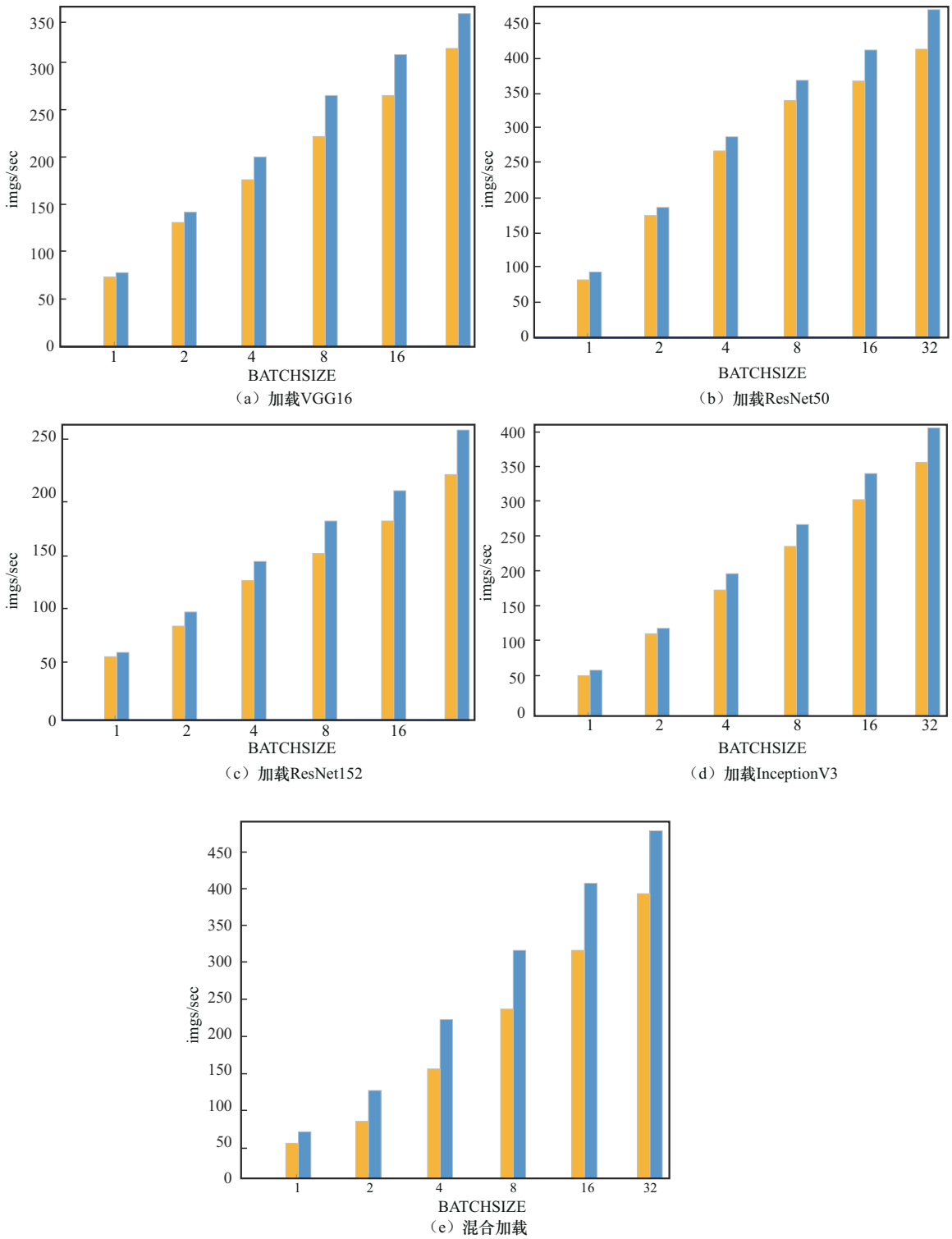

图11　实验测试结果

6　结束语

随着人工智能技术和应用的快速发展，如何在有限且成本高昂的算力资源环境下，提高算力资源利用率和使用效率，成为一个重要的研究课题。现有的技术手段难以满足多元异构GPU资源

的高效调度和科学管理的业务需要。本文提出了一种基于虚拟化的GPU异构资源池平台，一是通过对平台总体架构、逻辑架构和功能架构进行设计，实现了“一池多芯”和“一池多云”；二是对所述GPU资源池关键技术进行研究，提出基于虚拟化技术的异构GPU资源池化框架，实现了异构资源池化纳管，设计的基于时间切片+负载均衡的调度模型，实现了不同厂商异构GPU池化后的资源高效调度，为AI应用提供可远程访问的、可灵活切分的、可聚合的弹性GPU算力。此外，本文基于上述研究成果，对现有应用模式进行创新，引入了多业务单卡叠加、交叉拉远、跨机整合、混合部署和时分复用等应用模式，显著提升了GPU资源利用率和使用效率。

展望未来，研究将进一步拓展至FPGA、专用集成电路（application specific integrated circuit，ASIC）和CPU等其他通用或智能计算资源领域，旨在为企业提供更多更优质的算力资源解决方案。

参考文献：

[1] 于非, 何玉林, 贺颖. 人工智能与数字经济专题序言: 人工智能赋能数字经济，共创未来无限可能[J]. 深圳大学学报(理工版), 2023, 40(3): 253-257.
YU F, HE Y L, HE Y. Editorial of special issue on artificial intelligence and digital economy[J]. Journal of Shenzhen University (Science and Engineering), 2023, 40(3): 253-257.

[2] 傅懋钟, 胡海洋, 李忠金. 面向GPU集群的动态资源调度方法[J]. 计算机研究与发展, 2023, 60(6): 1308-1321.
FU M Z, HU H Y, LI Z J. Dynamic resource scheduling method for GPU cluster[J]. Journal of Computer Research and Development, 2023, 60(6): 1308-1321.

[3] 陈铉, 阚博文, 刘广一. GPU技术的最新进展及其在电力系统中的应用前景探讨[J]. 电力信息与通信技术, 2018, 16(3): 16-25.
CHEN X, KAN B W, LIU G Y. The latest development of GPU and its prospective application in power system[J]. Electric Power Information and Communication Technology, 2018, 16(3): 16-25.

[4] 陈杏仪, 柯清建. 异构算力的应用与展望[J]. 长江信息通信, 2023, 36(11): 226-228.
CHEN X Y, KE Q J. Application and prospect of heterogeneous computing power[J]. Changjiang Information & Communications, 2023, 36(11): 226-228.

[5] 段晓东. 中国移动算力网络推动“东数西算”工程向纵深发展[J]. 通信世界, 2022(5): 20-21.
DUAN X D. China Mobile Computing Network promotes the deep development of the “East Counting and West Computing” project[J]. Communications World, 2022(5): 20-21.

[6] 孙毅, 王会梅, 鲜明, 等. Kubeflow异构算力调度策略研究[J]. 计算机工程, 2024, 50(2): 25-32.
SUN Y, WANG H M, XIAN M, et al. Research on heterogeneous computing scheduling strategy for kubeflow[J]. Computer Engineering, 2024, 50(2): 25-32.

[7] 许俊东, 李兆滨, 宋德华, 等. 算力网络应用平台研究与设计[J]. 信息技术与信息化, 2024(2): 27-30.
XU J D, LI Z B, SONG D H, et al. Research and design of computing network application platform[J]. Information Technology and Informatization, 2024(2): 27-30.

[8] 王月, 柯芊. 智能计算中心: 人工智能时代的算力基石[J]. 中国电信业, 2021(S1): 11-15.
WANG Y, KE Q. Intelligent computing center: the cornerstone of computing power in the era of artificial intelligence[J]. China Telecommunications Trade, 2021(S1): 11-15.

[9] 史庭祥, 张剑波, 曹越, 等. 一种基于超大规模云资源池的算力供给新模式及其关键技术[J]. 移动通信, 2023, 47(1): 83-89.
SHI T X, ZHANG J B, CAO Y, et al. A new model of computing power supply based on super-large-scale cloud resource pool and its key technologies[J]. Mobile Communications, 2023, 47(1): 83-89.

[10] 邢文娟, 雷波, 赵倩颖. 算力基础设施发展现状与趋势展望[J]. 电信科学, 2022, 38(6): 51-61.
XING W J, LEI B, ZHAO Q Y. Development status and trend prospect of computing power infrastructure[J]. Telecommunications Science, 2022, 38(6): 51-61.

[11] 高鹏, 李玮, 唐利莉, 等. 新基建风潮下的全国数据中心算力布局规划研究[J]. 电信工程技术与标准化, 2021, 34(8): 2-5.
GAO P, LI W, TANG L L, et al. Research on the layout planning of national data center computing power under the new trend of infrastructure construction[J]. Telecom Engineering Technics and Standardization, 2021, 34(8): 2-5.

[12] 沈健, 孙道军. 算力网高质量发展探究: 现实挑战、制度基础和技术支撑[J]. 人工智能, 2024, 11(2): 20-31.
SHEN J, SUN D J. Research on the high-quality development of computing network: realistic challenges, institutional basis and technical support[J]. AI-View, 2024, 11(2): 20-31.

[13] 崔雪冰, 张延红, 李国徽. 基于通用计算的GPU-CPU协作计算模式研究[J]. 微电子学与计算机, 2009, 26(8): 30-33.

CUI X B, ZHANG Y H, LI G H. On the basis of general-purpose computing GPU-CPU cooperative computing method research[J]. Microelectronics & Computer, 2009, 26(8): 30-33.

[14] 杨经纬, 马凯, 龙翔. 面向集群环境的虚拟化GPU计算平台[J]. 北京航空航天大学学报, 2016, 42(11): 2340-2348.
YANG J W, MA K, LONG X. Virtualized GPU computing platform in clustered system environment[J]. Journal of Beijing University of Aeronautics and Astronautics, 2016, 42(11): 2340-2348.

[15] 王浩, 王浩枫. 面向CPUs-GPUs系统的OpenCL任务调度框架[J]. 计算机工程与设计, 2022, 43(7): 1955-1963.
WANG H, WANG H F. Scheduling framework for OpenCL programs on CPUs-GPUs heterogeneous platforms[J]. Computer Engineering and Design, 2022, 43(7): 1955-1963.

[16] 崔嘉. 试析虚拟计算环境中资源池的资源聚合机制的研究[J]. 自动化技术与应用, 2017, 36(6): 35-37, 41.
CUI J. Research on resource pool mechanism of resource pool in virtual computing environment[J]. Techniques of Automation and Applications, 2017, 36(6): 35-37, 41.

[17] 查乾. 基于GPU虚拟化的资源优化调度[D]. 武汉: 武汉理工大学, 2022.
ZHA Q. Optimal scheduling of resources based on GPU virtualization[D]. Wuhan: Wuhan University of Technology, 2022.

[18] 朱紫钰, 汤小春, 赵全. 面向CPU-GPU集群的分布式机器学习资源调度框架研究[J]. 西北工业大学学报, 2021, 39(3): 529-538.
ZHU Z Y, TANG X C, ZHAO Q. A unified schedule policy of distributed machine learning framework for CPU-GPU cluster[J]. Journal of Northwestern Polytechnical University, 2021, 39(3): 529-538.

[19] GÓMEZ-LUNA J, GONZÁLEZ-LINARES J M, BENAVIDES J I, et al. Performance models for asynchronous data transfers on consumer Graphics Processing Units[J]. Journal of Parallel and Distributed Computing, 2012, 72(9): 1117-1126.

[20] 吴再龙, 王利明, 徐震, 等. GPU虚拟化技术及其安全问题综述[J]. 信息安全学报, 2022, 7(2): 30-58.
WU Z L, WANG L M, XU Z, et al. GPU virtualization technology and security issues: a survey[J]. Journal of Cyber Security, 2022, 7(2): 30-58.

[作者简介]

张万才（1983– ），男，博士，国电南瑞科技股份有限公司高级工程师，主要研究方向为电力信息通信。

张楠（1982– ），女，国电南瑞科技股份有限公司高级工程师，主要研究方向为人工智能平台、知识图谱、NLP语义分析、RPA机器人在电力系统中的应用和创新。

杨文清（1974– ），男，博士，国电南瑞科技股份有限公司高级工程师，主要研究方向为电网大数据分析、电力系统数据应用及数据价值挖掘。

王涛（1996– ），男，国电南瑞科技股份有限公司工程师，主要研究方向为人工智能、计算机视觉、数字图像处理、目标检测等。

张文强（1993– ），男，国电南瑞科技股份有限公司工程师，主要研究方向为电力信息化系统建设、人工智能、知识图谱。