面向算力互联网的柔性计算调度机制研究

顾炯炯[1]，李佩珊[2]，曹伟朋[3]，蔡智源[1]，徐传飞[3]，闫丹[2]，毛馨纬[2]，明仲[3]
（1.华为云计算技术有限公司，广东 深圳 518129；2.中国信息通信研究院，北京 100191；
3.人工智能与数字经济广东省实验室（深圳），广东 深圳 518107）

摘 要：算力互联网本质是在互联网体系架构上构建统一算力标识符，以算网云调度操作系统和高性能传输协议为基础，形成算力标准化、服务化的大市场和算力相互连接、灵活调用的逻辑网。当前算力互联网面临着算力资源利用率低和成本高等挑战。为解决这一问题，柔性计算应运而生，作为一种数据驱动、AI 使能的新一代算力互联网资源调度范式。柔性计算通过精确的负载画像和服务质量（QoS）驱动的动态资源调度，优化算力资源池的利用率，在保障用户业务 QoS 的前提下显著降低用户的算力使用成本。深入研究了柔性计算的关键技术以及在算力互联网中的应用潜力，重点展示了其在提升算力利用效率和优化资源配置方面的创新优势。

关键词：算力互联网；柔性计算；弹性计算；服务质量
中图分类号：TP302.1
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2025184

Research on flexible computing scheduling mechanism for Internet of computing resource

GU Jiongjiong[1], LI Peishan[2], CAO Weipeng[3], CAI Zhiyuan[1], XU Chuanfei[3], YAN Dan[2], MAO Xinwei[2], MING Zhong[3]
1. Huawei Cloud Computing Technologies Co., Ltd., Shenzhen 518129, China
2. China Academy of Information and Communications Technology, Beijing 100191, China
3. Guangdong Laboratory of Artificial Intelligence and Digital Economy (Shenzhen), Shenzhen 518107, China

Abstract: The Internet of computing resource was architected to establish a unified computing power identifier within the Internet framework. A computing-network-cloud scheduling operating system, integrated with high-performance transmission protocols, was utilized. Consequently, a large-scale, standardized, and service-oriented computing power market was formed, resulting in a logical network that enables the interconnection and flexible invocation of computing power. However, Internet of computing resource faces some important challenges such as low resource utilization and high costs. To solve these challenges, flexible computing has emerged as a new-generation resource scheduling paradigm that is data-driven and AI-enabled for Internet of computing resource. Through precise user profiling and quality of service (QoS)-driven dynamic resource scheduling, flexible computing can improve the resource utilization, significantly reducing the costs while ensuring QoS. The key technologies and applications of flexible computing are explored in Internet of computing resource, with the advantages in improving resource utilization and optimizing resource allocation highlighted.

Keywords: Internet of computing resource, flexible computing, elastic computing, quality of service

收稿日期：2025–08–05；**修回日期：**2025–10–16
基金项目：广东省基础与应用基础研究基金资助项目（No.2025A1515011259）
Foundation Item：Guangdong Basic and Applied Basic Research Foundation (No.2025A1515011259)

0　引言

从互联网的发展历程看，自1950年开始，全球互联网经历了消费互联网、移动互联网、工业互联网等多个典型发展形态，演进至今日，伴随人工智能对算力需求的急剧增长，算力的架构互通、高速互联、弹性调度要求日益提升，产业各界开启构建算力互联网。根据中国信息通信研究院《算力互联网对形成新型生产关系的作用逻辑与实践方式》定义，算力互联网是面向算力应用与调度需求，通过能力增强和系统升级形成的新型基础设施和技术产业体系，其本质是在互联网体系架构上构建统一算力标识符，以算网云调度操作系统和高性能传输协议为基础，增强异构计算、弹性网络等能力，具备智能感知、实时发现、随需获取能力，形成算力标准化、服务化的大市场和算力相互连接、灵活调用的逻辑网[1]。

我国算力基础设施体系庞大。截至2024年年底，我国算力总规模达246 EFLOPS（每秒246百亿亿次浮点运算）[2]，居全球第二。算力资源结构复杂，算力主体较为分散，市场算力服务主体超5 000家[3]。在算力互联网快速发展背景下，涉及大体量终端设备与平台连接，海量数据资源、多种类算力资源与网络基础设施，对算力资源调度管理水平提出了全新挑战和更高要求。然而，当前我国大部分运营主体调度水平不够高，灵活部署及服务能力欠缺，难以针对各类场景合理、精准匹配算力资源。

谷歌、微软、亚马逊等业界知名云厂商在算力资源调度及运行时技术方面也有不少探索和实践：比如作为Kubernetes容器调度引擎前身的Borg[4]是谷歌内部自研的一套算力资源调度引擎，它通过用户指定应用资源模板，并指定资源最大、最小预留空间控制中央处理器（CPU）超分以及进程级资源隔离等，实现了具备一定程度应用差异感知的算力分配效率提升；基于Borg已验证成功的经验，谷歌进一步设计了一套更加精简的集群管理调度系统Omega[5]。相比于Borg单一中心式调度，Omega采用多调度器模式，迭代修改更加方便。微软Hyper-V[6]内置与Windows Server操作系统兼容的企业级虚拟机管理（VMM）运行时技术，并支持与Azure公有云算力调度引擎配合实现跨云租户的算力资源高效池化复用。爆竹（Firecracker）[7]是由亚马逊云开源的轻量级虚拟化运行时技术，主要目的是解决虚拟机启动速度慢、系统开销大，以及容器的跨租户安全隔离保护能力不足的问题，上述所有涉及虚拟机及容器的运行时技术及其资源池调度机制，包括Borg、Omega、Hyper-V以及Firecracker等，均未改变基于静态规则和固定虚机/容器规格进行资源池算力分配管理的公共范式，本质上仍然属于基于分配率而非利用率的算力调度，无法实现依据每个虚拟机/容器实例的动态忙闲度及邻居干扰程度变化进行动态自适应的算力调度。学界麻省理工学院（MIT）面向效用计算终极愿景所提出的流沙（QuickSand）计算架构，目标是引入比当前容器或无服务器计算（Serverless）更细粒度的资源调度单位，以便有效降低算力资源浪费，并为提供用户基于精确计算资源消耗量的付费，然而，考虑其实现需要依赖于将应用计算实例拆分为细粒度、可在主机间无缝迁移的Proclet计算单元，改造代价高昂，并不具备实际的工程可实现性，且在性能开销方面也面临诸多挑战。文献[8]详细分析总结了通算的关键挑战（如资源限制、数据持久化和状态管理、通信效率、性能优化等），虽然预言了未来Serverless将成为云时代默认的计算范式，并终结服务器化计算的时代，但并未给出具体完整的实现路径。

在算力分配与供给方面，业界云厂商多数采用弹性计算的技术与商业模式，即由云服务商预定义一系列固定配比的算力规格，再由云租户根据业务应用的大致资源需求和性能敏感程度选择资源规格和算力所在区域，后续云服务商以租户选定的算力规格作为计费单元。这种基于弹性计算的资源服务模式实现简单，但存在严重缺陷，即租户的业务负载具有高度动态性，对算力资源的使用存在波峰波谷现象，会导致租户为大量占而未用的资源额外付费，同时，云服务商数据中心的资源有效利用率偏低。AWS burstable性能实例和Azure B系列也具备弹性能力来提升资源利用率，但它们没有服务质量（QoS, quality of service）保障。主要是因为AWS burstable性能实例和Azure B系列性能取决于对应的历史积分积累情况，如果积分耗尽，性能会被限制到极低的水平，可能会导致QoS急剧下降。通过实际观察，当前业界云算力的平均CPU有效利用率仍仅达20%[9-10]，远低于80%+的CPU分配率水平。为解决该问题，在对海量云应用的资源使用特征的深入洞察基础上，本文创新性地提出了应用驱

动的新一代算力资源服务调度体系——柔性计算，旨在为用户提供具有实时QoS保障的量体裁衣式算力资源服务。

近年来，随着业界对数据中心资源效率的日渐重视，以性能为中心的传统算力服务设计理念正在向以能效为中心转变。例如，在中国工商银行的云平台成本优化建设规划中，明确提出要建设以“按实际运行需要进行资源分配”为核心思想的云平台，采用一系列柔性资源调配措施提升资源利用率。该思路与本文提出的柔性计算理念相吻合，但该银行所提规划更多针对自身经营的私有云环境进行创新，难以普适到公有云和混合云场景。国际上以AWS为代表的云厂商近年来也推出了具有一定柔性算力供给能力的产品，如AWS 2024年5月推出的C7i-Flex，该产品通过动态超分提升资源运营效率，但仍然缺少对云应用QoS表现的实时度量能力，因此并不能做到具有实时QoS保障的极致资源服务能力。

本文提出的柔性计算可被认为是弹性计算的下一跳，它的核心本质在于将算力资源的分配与供给机制从传统弹性计算的计划经济模式转变为市场经济模式，即根据应用的实际需求进行按需供给，能够有效解决弹性计算计划经济模式下普遍存在的供过于求和供不应求问题。在柔性计算的市场经济模式下，生产、资源分配及产品消费不再仅依据预先指定的计划进行，而是采用一切依托动态变化的客户需求为中心进行驱动。在保障云租户的业务应用QoS满足服务级别协议（SLA, service-level agreement）的前提下，对资源池内的算力进行极致的动态空分和时分复用，从而最大限度提升云服务商算力资源池的有效利用率、资产运营效率及能效比。

柔性计算强调对于应用负载的动态资源需求更为准确和精细化的感知与画像，并基于该画像信息，在算力资源分配过程中，实现多租户、多应用负载所需算力资源的最大化时空复用，最终达成真正动态自适应的算力供需最优化平衡。柔性计算对租户业务应用的资源需求洞察，分为实例级和集群级2个层面。

实例级：在最小可分配的算力实例层面，无论该算力资源实例的颗粒度大小及运行时技术是虚机、容器还是函数，柔性计算强调打破弹性计算CPU–内存固定配比的思维定势，依托应用负载在其画像周期内的历史资源用量采样数据，为实例设置量体裁衣的精细化规格。而算力实例发放后，系统仍将持续监控实例的资源用量，对实例进行动态画像，作为后续主机资源分配策略的输入和依据，从而实现单台服务器范围内多租户、多应用负载算力资源的最大化空分复用。

集群级：在应用负载集群维度的资源需求层面，柔性计算也打破了传统弹性计算模式下算力资源水平弹性伸缩单纯依赖人工设置的固定资源水位线及固定触发条件的约束，全面引入基于应用负载集群资源用量、业务请求等监控统计指标等历史数据及机器学习技术驱动的智能弹性伸缩资源预测模型，从而使柔性计算平台得以依据预测结果，在恰当的时机对合理的算力资源量进行分配与释放，实现大规模服务器集群范围内多租户、多应用负载算力资源的最大化时分复用。

1　基本理念

根据业界开源数据集及华为公有云现网数据分析，当前数据中心算力资源实际利用率偏低的原因主要归结为以下3个方面。

一是算力规格与应用负载需求不匹配。云服务商预定义的算力规格系列和云租户应用负载多样化的算力需求不匹配，虚拟机/容器实例的CPU-内存配比通常为1:2的n次方，比如2核4 GB（2U4G），8核32 GB（8U32G）等，如图1所示，但应用对各维度算力资源的真实需求往往并不符合该固定配比，从而导致了不必要的资源浪费，增加了用户上云成本。

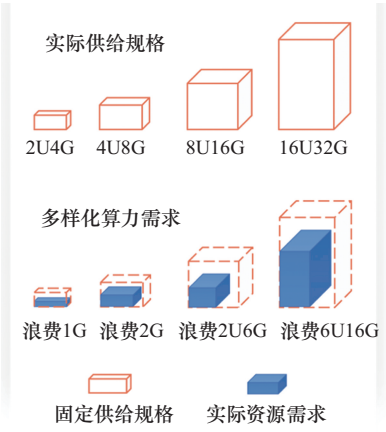
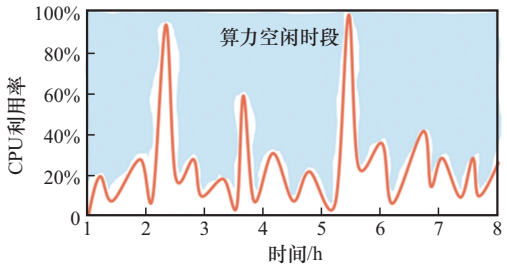


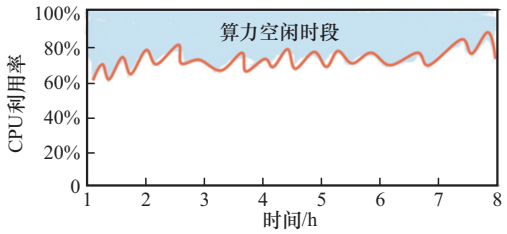
图1　算力规格与应用负载需求不匹配

二是资源调度缺乏对业务负载忙闲特征的感知。如图2所示，主机内分配多个虚拟机/容器往往

采用非超分或固定超分的分配算法，缺乏对算力实例CPU实际用量忙闲度分布特征的感知，导致对于CPU平均利用率低但偶尔冲高的业务应用负载，超分比不足导致算力浪费；而对于CPU平均利用率持续较高的场景，则超分比过高导致性能劣化显著。



(a) 轻载应用的CPU利用率特征



(b) 重载应用的CPU利用率特征

图2　资源调度缺乏对业务负载忙闲特征的感知

三是弹性伸缩依赖固定阈值导致算力分配释放与应用需求无法及时同步。如图3所示，当前业界对于集群级弹性伸缩，缺乏对应用负载随业务量变化所需的资源总量随时间动态变化的特征感知，依赖保守原则确定的资源预留水位线往往导致较大算力浪费；依赖CPU利用率或业务指标动态触发资源申请释放，则可能导致资源就位滞后于业务量的突发增长导致业务请求失败或任务排队等待等QoS劣化问题。

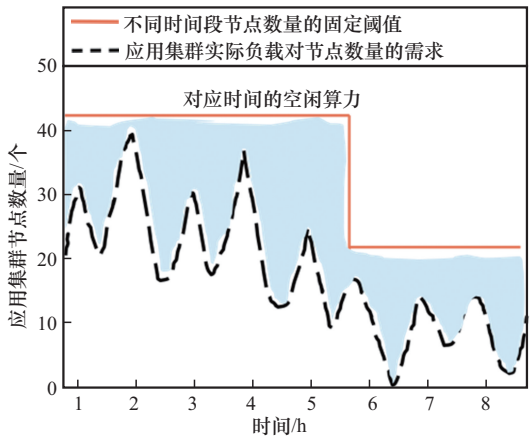


图3　弹性伸缩依赖固定阈值导致算力分配释放与应用需求无法及时同步

柔性计算的主要目的是解决上述三大痛点问题。柔性计算的价值，还在于从根本上解决了Serverless函数计算模式在云算力利用率提升方面存在的一个关键约束：应用服务需满足无状态或轻状态的前置条件，进行事件驱动应用架构（EDA）范式的改造，且必须将应用关键状态信息下沉到云服务商后端数据库、云存储服务，导致有状态应用负载及服务（含数据库、消息/缓存类中间件）的Serverless化不可避免面临数据访问性能瓶颈，以及巨大的适配改造代价。

柔性计算与Kubernetes容器化、函数计算范式等云原生技术生态的差异在于，柔性计算作为下一代普适、泛在、AI原生驱动的Serverless算力范式，可广泛适用于虚拟机、容器及函数等多样化的控制面及数据面运行时技术，对于承载多样化应用负载及云服务的全栈公有云及私有云的大规模算力资源有效利用率提升及租户上云成本优化，具有重要的突破性意义。

柔性计算作为下一代Serverless算力范式的普适性、泛在性，除了体现在上述对虚拟机、容器及函数等多样化控制面及数据面运行时技术的广泛兼容性外，还体现在硬件算力资源层面上对CPU、神经网络处理器（NPU）、图形处理器（GPU）和数据处理器（DPU）等多元异构算力类型，以及通用在线与离线类业务、AI模型训推任务、Agent应用的广泛适用与支持。 在云上多租算力品价比最优化的目标驱动下，云资源利用率的提升与应用的性能保障之间永远存在“鱼与熊掌”的权衡矛盾，考虑到不同云租户对性能和成本有不同的诉求，柔性计算基于业务对QoS保障水平和资源冲突后二次调度优先级的选择，为柔性实例设计了3档优先级。

柔性高优先级：按95-th峰值画像为实例进行CPU、内存、存储、网络I/O资源预留，并通过动态绑核机制保障应用性能的稳定性。此优先级的单位算力价格最高，主要面向性能敏感类业务，承诺业务性能抖动小于5%。

柔性中优先级：按均值画像为实例进行各维度资源预留，除柔性高优先级实例外，主机资源发生冲突时优先保留，资源空闲时允许其资源使用超过预留值。此优先级能接受一定的性能劣化（小于20%），故单位算力价格居中，主要面向高阶服务有状态业务，或者租户侧的云原生应用。

柔性低优先级：按最小规格预留资源，主机资源空闲时允许其资源使用超过预留值，资源发生冲突时优先驱逐。此优先级能接受较大幅度的性能抖动（小于50%），且局部实例中断不影响整体业务连续性。因此单位算力价格最低，主要面向离线批处理业务，以及具备调度框架的业务集群。

3档优先级的设置打破了弹性计算模式下不同算力规格独占物理资源池的制约，实现了跨不同CPU硬件代次、多优先级混部与抢占式调度的资源池化归一化，大幅减少资源分配的碎片化浪费。

通过引入多优先级柔性实例，云租户可以权衡业务性能需求和算力成本开销，找到最优的实例优先级组合；同时，云服务商也可以对不同优先级采取不同的调度算法和策略，通过跨优先级混合部署进一步提升主机的资源利用率，降低运营成本，实现租户成本与云服务商资源利用率的双赢。

2　柔性计算架构设计

为支撑上述理念，本文设计了的柔性计算架构，如图4所示。

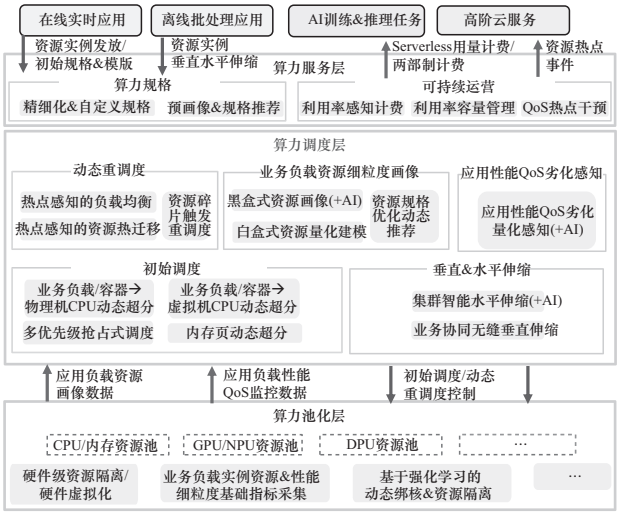


图4　柔性计算架构

算力池化层：包含CPU/内存资源池、GPU/NPU资源池、DPU资源池等多样化算力。将云算力分配与供给模式从以资源为中心转变为以应用需求为中心，云租户的算力实例规格以及整体算力实例数量，采用动态在线预画像数据与AI驱动机制，结合多维资源协同调度算法，在保障租户业务QoS的前提下，最大化降低算力分配过程中的资源过剩和浪费。此外，柔性计算架构下前端逻辑资源与后端物理资源解耦，云租户不感知硬件代次，柔性计算通过性能建模仿真模块评估特定任务SLA下不同硬件代次的需求，以应用性能视角决策特定算力资源需求，从而实现跨硬件代次的算力资源池化共享。同时在硬件和操作系统基础层，提供硬件级虚拟化/资源管理、细粒度业务负载基础指标采集，以及动态绑核等能力。

算力调度层：柔性计算关键技术实现层，需要根据负载结构进行智能分配。通过在线分析应用负载结构，给出最符合历史数据特征的画像，基于画像对负载进行分类，作为调度器输入。调度器以离线基准测试结果作为性能基线，给出同时满足QoS以及资源要求的调度决策，并执行负载资源映射绑定。算力调度层具体包括以下内容。

1) 初始调度：从弹性计算装箱式调度转变为预/动态画像数据+概率卷积驱动的确定性概率调度，实现了共宿物理算力的多个应用负载实例在CPU等资源维度的智能动态超分。

2) 动态重调度：实现了资源热点及资源碎片感知的二次热迁移及业务负载均衡，资源调度从静态开环转变为动态闭环系统。

3) 应用性能QoS劣化感知：通过细粒度的观测指标采集，实现白盒/黑盒式资源量化建模，作为算力资源动态调度及无缝伸缩的关键输入。同时，基于非侵入式的QoS劣化检测模型实时监控实例的性能表现，在触发劣化阈值时驱动二次迁移实现主动型QoS保障。

4) 垂直&水平伸缩：弹性计算模式下，资源规格调整操作会导致业务中断，柔性计算可实现业务零中断的无缝垂直和水平伸缩，极大提升用户体验。

业务负载资源细粒度画像支撑以上4点调度。

算力服务层：柔性计算技术具备普适性，各种部署形态的租户应用都可以利用柔性计算平台能力，在保障性能的同时减少资源浪费，提升成本优势；相比弹性计算，主要差异包括：1)精细化算力规格。面向云用户，通过预画像/在线迭代画像能力，实现与应用负载各维度算力需求精确匹配的推荐能力，支持任意资源配比（例如1U1G），显著优于当前弹性计算模式下Flavor规格配比单一的问题；2)可持续运营。面向云运营管理，容量管理从资源分配率为中心转变为利用率为中心，计量计费

从固定资源规格转变为按实际资源用量。

针对以上范式，本文总结了柔性计算的九大关键特征。

特征 1：精细化资源用量度量（预画像&在线迭代画像），使能量体裁衣式的云主机规格选择和实例发放。为了为租户提供与应用资源需求严格匹配的多维算力规格，需要对应用负载在时间和空间 2 个维度进行资源预画像以及在线迭代画像更新。柔性计算采用秒级指标采集确保不遗漏应用资源需求的波峰，以输出准确的应用画像，并通过精细化的多元算力取值和配比保证应用能选择到最匹配的实例规格。

特征 2：共宿实例资源用量画像驱动的动态超分算法，使能应用性能与资源复用的最佳平衡。对于共宿同一物理节点的多个实例进行细粒度的资源用量统计与上报，基于概率卷积技术实现资源冲突概率可控的动态超分，显著区别于当前业界普遍采用的固定超分比策略。

特征 3：非侵入式 QoS 劣化实时检测模型，使能 QoS 驱动的资源复用和热点快速消除。保障租户的业务应用 QoS 是算力资源服务的基本前提。因此，柔性计算提供了面向公有云、私有化、混合云普适的非侵入式 QoS 劣化检测模型，通过可观测内核指标实时推理实例内部应用的性能表现及资源使用特征，从而为资源复用和热点消除提供决策依据。

特征 4：QoS 感知的主机内算力资源分配与隔离，使能主机本地实时 QoS 动态保障。根据 QoS 劣化检测模型的反馈，在必要时触发动态绑核等操作，有效保障突发性能劣化实例的业务稳定性。

特征 5：QoS 热点感知的二次调度，使能主机热点快速消除。通过 QoS 劣化检测模型的输出结果触发底层热迁移和业务均衡相结合的二次调度，实现主动型 QoS 保障能力。同时引入迁移决策模型，实现热迁移一步到位，有效避免热迁移场景容易出现的乒乓效应。

特征 6：柔性内存，实现内存资源的安全可控超分。基于虚拟机内存用量画像合理分配资源，实现内存安全复用；基于主机内存热点分析及热迁移能力，进一步降低内存复用风险；基于内存缺页触发热迁移，作为最终性能保障措施。

特征 7：AI 与数据驱动的智能水平伸缩及资源池预热。基于资源池历史用量数据训练 AI 时序预测模型，基于该模型的预测输出指导集群资源快速、精准伸缩，达成业务未动、资源先行的目标。

特征 8：统一运行时框架，使能高效资源利用的跨运行时统一资源管理和调度。虚拟机、容器、函数在资源池拉通调度，通过资源使用特征差异进行互补，进一步减少资源碎片化，提升资源利用效率。

特征 9：支持基于计算实例的实际资源用量进行精细化计量和计费。引入 CPU、NPU、GPU 核时及内存、显存 GB 时作为计费量纲。

3　关键技术解析

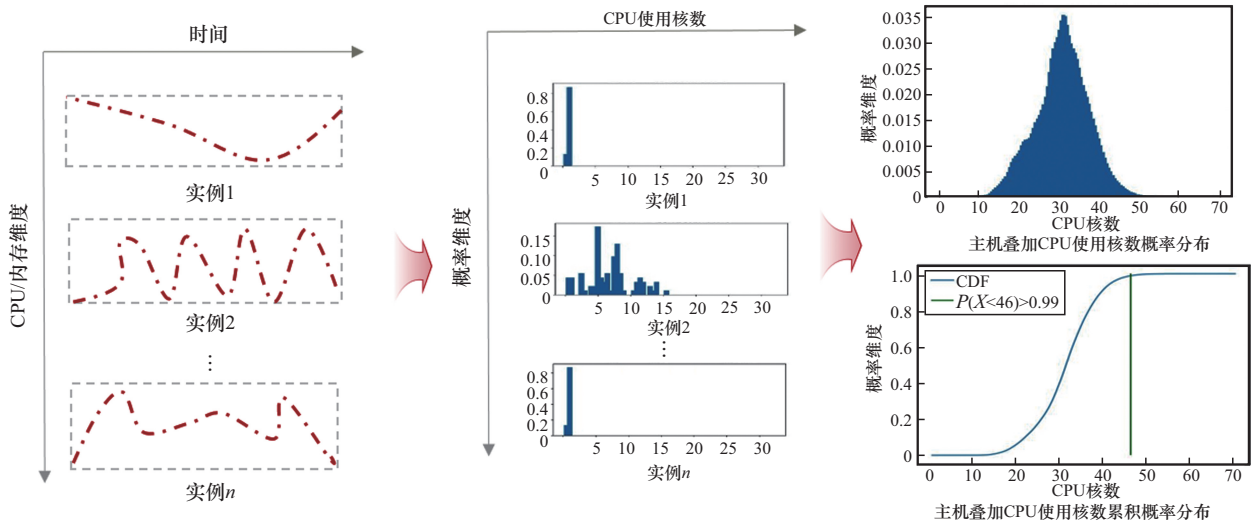
本节对柔性计算架构涉及的三大关键技术进行解析，即：基于概率卷积的主机 CPU 动态超分技术、非侵入式 QoS 劣化检测技术，以及数据驱动的集群资源智能预测技术。

3.1　基于概率卷积的主机 CPU 动态超分技术

基于概率卷积的主机 CPU 动态超分，本质上是在算力资源的分配过程中，确保共宿同一物理机的多个虚拟机/容器实例在同一时刻 CPU 算力资源之和不会超越物理算力 CPU 资源总量，使多租户算力资源的极致复用，大幅提升性价比。

该技术通过对共宿物理机的所有虚拟机实例，以及共宿虚拟机的所有容器实例进行资源用量统计，实现对其资源使用概率分布特征的洞察，并基于同一物理机内各虚拟机/容器实例的概率分布数据进行离散卷积，如图 5 所示，从而对多虚拟机/容器实例叠加算力的概率分布进行量化，进而指导柔性计算调度系统基于可量化的概率阈值，确定共宿虚拟机/容器实例的动态复用超分比，有效避免了当前业界普遍采用的固定超分策略导致的超分不足引发资源浪费与过度超分引发性能劣化等问题。

具体地，首先对节点上实例的 CPU 用量进行采样统计，并据此计算出节点 CPU 用量的概率分布。在图 5 的实验配置中，长租类虚拟机的画像周期缺省为 7 天，而动态弹性虚拟机的画像周期缺省为 1 h，并支持对历史累积的概率分布进行持续迭代更新。如图 6 所示，输入实例的 CPU 用量采样数据，以一定的分段粒度对 CPU 用量落在各个分段的样本数进行统计，得到 CPU 用量的离散概率分布。



(a) 资源利用率时序图　　(b) 资源用量概率分布　　(c) 主机内多实例CPU用量的概率叠加分布

图5　概率卷积叠加原理示例

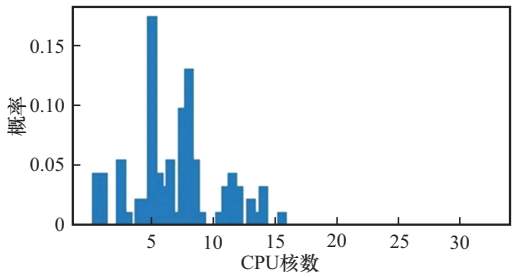


图6　画像周期内实例CPU用量的离散概率分布

离散随机变量的和通过卷积进行计算。

$$Z = X + Y$$
$$P(Z = z) = \sum_{k} P(X = k)\, p(y = z - k) \qquad (1)$$

其中，X、Y、Z为随机变量，$P(.)$表示相关概率分布。基于节点上各个实例的CPU用量概率分布数据，节点自身的CPU用量概率分布可以如下获得。据此本文可以进一步计算节点CPU用量的累积概率分布，并取一定概率阈值下的CPU核数作为节点CPU用量的评估值，具体实现如下（以Python为例）。

```
host_estimate_core = 0
for prob in numpy.cumsum(host_historgram):
    if prob > 0.99
        break
    host_estimate_core += 0.5 # 0.5 为频数统计时分段的粒度
```

基于以上测算，本文实现了资源冲突概率可控的主机级CPU动态超分。

3.2　非侵入式QoS劣化检测技术

上述基于概率卷积的主机CPU动态超分技术虽然在理论上保障了资源冲突概率的最小化，但由于云负载的高度动态性，仍然存在极小概率出现CPU资源热点，导致租户的业务性能劣化。因此基于QoS劣化检测触发的二次调度（可以是业务应用无感的资源热迁移，或业务感知的任务调度）消除算力集群内的资源热点，将业务性能恢复至SLA承诺水平，对于资源调度系统来说依然必不可少。

传统的QoS劣化检测通常采用基于专家经验的启发式规则算法[11-12]，其缺点在于动态超分环境下影响业务应用性能的因素非常复杂，启发式算法难以覆盖多样化的应用负载场景，QoS劣化检测的普适性和准确性偏低[13-14]。为了解决该问题，本文基于Transformer算法架构进行QoS预测。在公有云场景，应用层数据属于租户的私有数据，鉴于隐私保护协议约束，云服务商无法直接获取租户实例内部应用的业务层指标。因此，本文通过可以观测的实例运行时内核指标，利用AI建模的方式，构建起内核指标与实例内部无法直接观测的应用层业务指标的关联关系，从而实现对租户实例业务性能的非侵入式检测。因此，本文挖掘公有云可观测的各类内核指标时序数据（如虚拟机运行时的CPU、内存、存储、网络I/O相关维度的资源用量信息）与应用性能劣化百分比之间的映射关系，得到非侵入式的QoS劣化检测模型。

为使得该模型具备足够的泛化能力，柔性计算

引入了与大语言模型（LLM）类似的自监督预训练+有监督后训练的训练模式：基于云上生产环境可收集的多样化应用负载运行数据进行自监督学习，从海量的多维度资源用量及内核层性能统计指标数据中提取出应用负载的抽象高维空间特征，再向下对接下游的有监督微调训练任务，结合少量基于已知应用负载类型的性能劣化百分比标签数据对模型参数进行微调，从而提升模型的预测准确度。上述训练过程如图7所示。

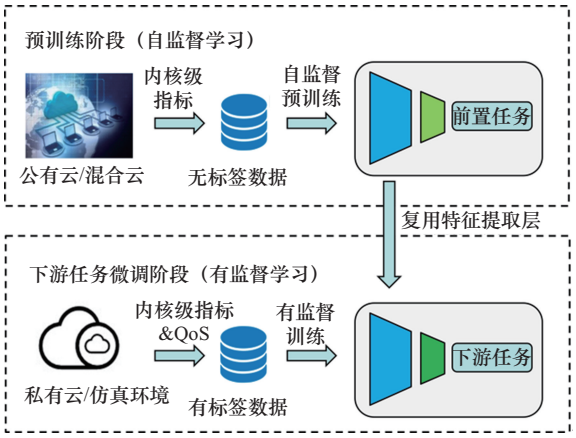


图7　QoS劣化检测模型的训练过程

具体来说，在自监督学习阶段，由于数据量巨大，为了更好地提取其中的有效信息和加速训练，本文选择Transformer算法作为特征提取模型的基础架构。在训练策略上，基于对比学习设计损失函数[15]，假设前提是：同一时间序列中相邻时间窗口内的内核指标，其变化规律是相近的；而不同时间序列，其时间上相邻窗口间的内核指标，其变化规律应该是不同的。基于此策略定义了对比损失函数，如式(2)所示。

$$L=\sum_i -\log\frac{\exp\left(\frac{\mathrm{sim}(h_i,f_j)}{\tau}\right)}{\sum_j \exp\left(\frac{\mathrm{sim}(h_i,f_j)}{\tau}\right)},\mathrm{sim}(h_i,f_j)=\frac{h_i\cdot f_j}{|h_i||f_j|} \tag{2}$$

其中，h_i和f_j是编码器编码后的特征向量，2个向量的下标相同表示其来自同一时间序列，下标不同则来自不同时间序列。τ是一个缩放因子（本文实验中τ=0.1），sim(·)表示求解2个输入向量的余弦相似度。同一时间序列的特征向量相似度越高，并且不同时间序列的特征向量相似度越低，损失函数就越小，符合本文的训练策略。图8展示了对比损失的构造过程。

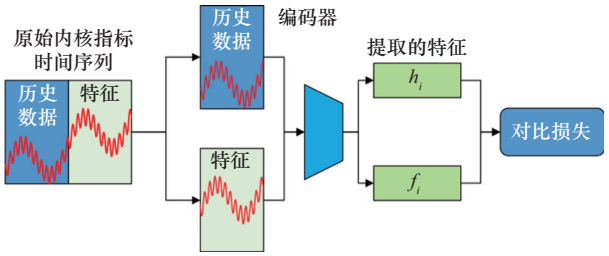


图8　对比损失的构造过程

原始的可观测性指标经过上述特征提取模型转换后，相同的应用类型将会被聚类到一个簇内。如图9所示，本文通过t分布随机邻域嵌入（t-SNE）技术降维可视化Cassandra数据库、HBase数据库等6种应用经过特征提取模型转换后的高维空间特征，可以直观地观察到，相同应用的特征基本聚集在一个簇内。

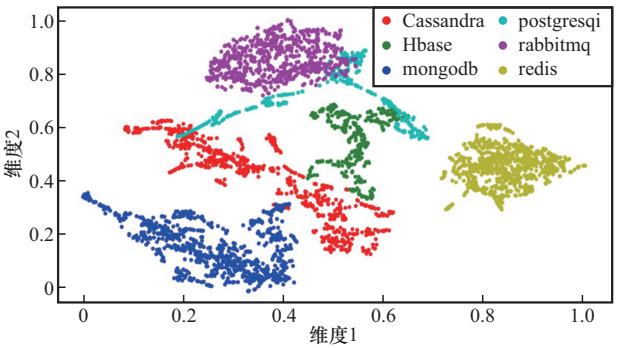


图9　经过特征提取模型转换后的应用特征聚类效果

上述特征抽取模型能够将公有云可观测的几十维指标转换为高维空间线性可分的抽象特征，在此基础上，采用多任务学习框架构建以QoS劣化检测为主任务的学习器，如图10所示。

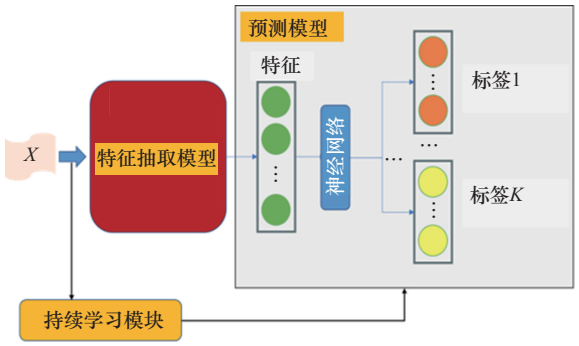


图10　面向QoS劣化检测的多任务学习框架示意

本文通过多个相关任务协同的方式，为主任务QoS劣化检测提供更丰富的辅助信息，以提升其预测准确度。通过perf采集工具采集CPU、内存、网络I/O等使用率指标，每3 s采集一个样本作为QoS模型训练数据，为了使其他研究人员可复现实验，本文把采集的数值型数据集公开[16]。

为测试QoS劣化检测模型效果，以Cassandra应用为例，其在现网的QoS真实劣化与模型预测的结果对比如图11所示。从图11可以看出，模型的预测结果与应用的真实QoS表现基本贴合，绝对误差（MAE）控制在5%以下，能够为实例性能保障提供准确的参考。目前业界云服务商在公有云服务中，通常不对租户承诺业务层QoS保障指标，本文提出以QoS劣化检测模型的输出进行资源动态隔离来保障实例的性能QoS会比当前现行的方案更优。更多实验结果可以参考文献[17]。

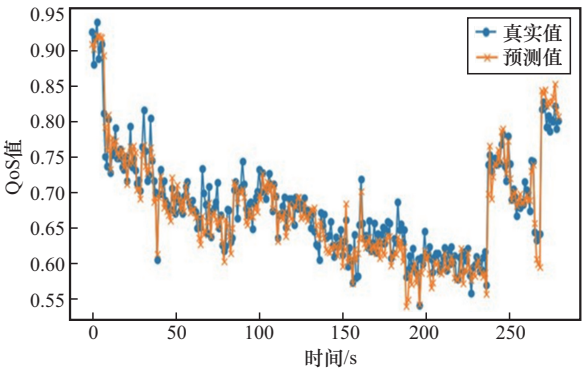


图11 QoS真实劣化与模型预测的结果对比

3.3 数据驱动的集群资源智能预测技术

当前业界普遍采用基于经验值（固定水位线）的方法作为集群资源备量的决策依据，这种方法易导致资源严重浪费且在业务突发时性能无保障。为解决这一挑战，本文在柔性计算范式中创造性地引入了数据驱动+AI使能的智能资源用量预测模型，有效解决了固定水位线配置法存在的局限性。

基于对云上应用负载系统的并发算力资源需求总量/业务并发量随时间变化的历史数据和AI算法训练时序预测模型，并基于该模型与实时测量的资源用量/业务并发量输入进行模型推理，输出与集群动态资源用量曲线匹配的智能动态伸缩资源量的实时预测，从而实现多租户应用负载在公共算力资源池集群范围内的最大化时分复用共享[18]。

具体地，Transformer算法具备的序列建模能力与可扩展性在自然语言、图像处理等领域已被广泛验证，在时序预测领域，Transformer算法也逐渐成为主流架构。iTransformer算法[19]针对多维时间序列的数据特性，创新性地提出对特定维度输入进行转置编码，从而避免不同维度特征之间因为量纲差异导致数值缩放干扰以及破坏特定维度的变化规律信息，在业界多个时序预测任务中取得了具有竞争力的性能表现。本文基于资源池多维度时序数据进行建模，与iTransformer算法的创新架构较为匹配，故选择了iTransformer算法作为相关模型的基础架构，在实践中，仍需结合场景约束对其进行进一步优化。例如，原始iTransformer算法的损失函数并未对预测结果相对真实值出现高估或低估做出差异性惩罚，本文根据场景业务层约束条件，调整了模型对预测结果高估和低估的惩罚力度。在华为云的资源池智能水平伸缩场景中，鉴于业务侧要求任务等待时间要尽可能地短，因此本文加大了对模型的预测低估的惩罚度（10倍于高估惩罚），从而最大程度避免了资源准备不足导致的上层业务等待问题。

因此，本文在iTransformer算法的基础上，结合业务场景的条件约束（如采集维度限制、请求等待率约束等），将资源池多维度的资源时序数据分而治之，每个维度的变量被编码成独立的文本处理中的最小词元（Token），并利用注意力机制和前馈网络分别建模不同变量间相关性和变量的时序相关性，从而获取更好的时序表征，并提升最终时序预测的泛化能力。

图12展示了时序预测模型在华为云数据中心某资源池上的预测效果（直线对应业务负载的实际算力资源需求总量随时间推移的动态曲线，标圆点的线对应该模型的预测值，两条曲线越接近，则资源浪费越少，时分复用效率越高。灰色虚线则是该业务原有的基于经验设置的固定资源水位曲线）。本文实验中采集了3年历史数据作为训练数据，由于训练数据中包含节假日及促销活动日，采集信息包括时间戳信息和多维资源用量信息等。因此，节假日等周期性出现的事件，模型能够捕获特定时间与资源用量的关联关系，故可以较为准确地预测出特殊时段的资源用量值。对于非周期的突发性营销活动，可以通过系统预留的应用程序接口（API）对模型的输出结果进行覆盖，以应对临时突发的资

源需求。从实验结果来看，节假日和促销活动预测效果也表现良好。

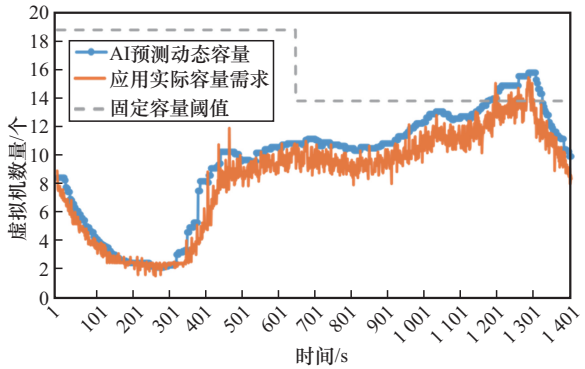


图 12　数据驱动的集群资源智能预测

在实践中，对于一些计划任务导致的资源需求突增，如果历史数据中未出现类似突增场景，模型可能无法提前预测。因此，该模型应和业务调度系统配合，获取用户配置的计划任务并输入到预测服务中，对模型的预测结果进行修正，以覆盖此类突增场景。

3.4　基于柔性内存的实例级资源垂直伸缩能力

目前业界虚拟机/安全容器的内存资源动态伸缩通常基于内存气泡/空闲内存页指示技术，由于采用异步通知和回收机制，存在性能开销大、空闲内存页回收不及时等关键缺陷，导致跨虚拟机/安全容器的内存页资源无法实现实时和高效的动态伸缩与复用。此外，微软和 Intel 等通过计算快速链接技术（CXL）高速互联，形成内存池，从而提升内存资源使用率[20-21]，但它们依然采用异步通知和回收机制，在内存访问频繁场景，内存会频繁释放、回收再复用，性能依然存在挑战。

针对上述挑战，本文创造性提出了柔性内存这一更为可靠高效的内存动态复用机制，引入基于单层内存管理单元（MMU）内存管理的同步式空闲内存页释放与回收机制，替代基于双层 MMU 内存管理的异步空闲内存页释放与回收机制，彻底解决了内存气泡/空闲内存页技术的弊端。

为了测试柔性内存与 Linux 原生内存是否存在差距，针对 1 GB 内存进行复制、缩放、加法、组合（Copy、Scale、Add、Traid）等操作测试。为了保证实验公平性，分别对每种操作测试带宽和时延，并测试 100 次取平均值。如图 13 所示，柔性内存的性能与 Linux 原生内存管理系统基本接近，无论是带宽还是时延性能差异都在 1% 以内。由此可知，柔性内存不会降低原生内存的性能。

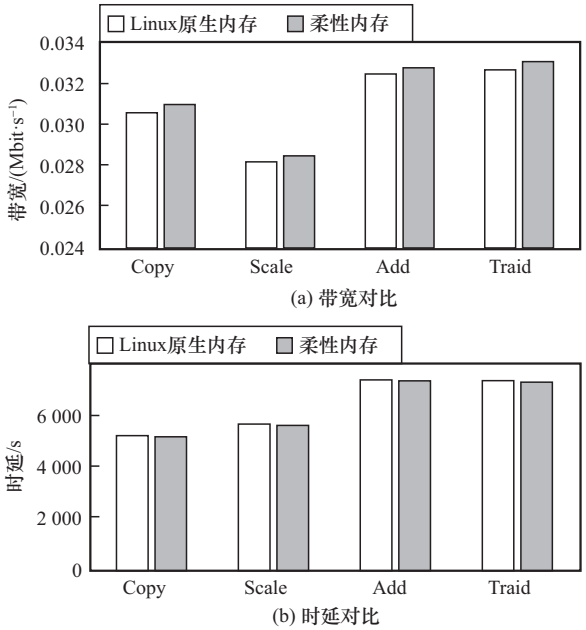


图 13　柔性内存性能效果

更进一步，柔性内存通过重载主机侧内核的内存分配管理流程，基于客户机/主机共享的空闲内存页元数据，在保留虚拟机和安全容器内存页虚拟机监视器（Hypervisor）级安全隔离保障的前提下，通过跨虚拟机和安全容器的同步式空闲内存的释放、回收和再复用，实现了跨虚拟机/安全容器的内存页资源实时、高效、安全的动态伸缩与复用，也使得承载应用负载的虚拟机及安全容器具备了毫秒级的垂直伸缩能力，为数据库、中间件及更多有状态应用负载的无服务器计算化提供了有力支撑。

4　应用研究与案例分析

4.1　公有云的应用上云成本优化

对于公有云服务商而言，能否在相同总体拥有成本（TCO）投入且满足云上应用性能 SLA 需求的前提下，最大限度优化云租户的算力成本支出，对于其市场竞争力及客户满意度具有决定性的意义。

基于柔性计算的华为公有云 Flexus X 云主机实例[22]，通过量体裁衣的精细化调度、基于概率卷积的主机 CPU 动态超分一次调度，基于非侵入式 QoS 劣化检测的二次调度等能力，使中小企业（SMB）用户上云成本平均降低 30%。例如，如果

检测到某共享实例的QoS劣化比例超过了特定阈值，将对该实例进行动态绑核，以保障其性能稳定性。如果通过绑核依然无法使其QoS表现恢复到预期范围内，将触发热迁移操作，将其二次调度到更空闲的主机上。

华为公有云某互联网公司A离线任务系统，引入Flexus X实例+柔性容器，对接柔性集群容量预测服务，相比之前采用的静态预留方案，算力成本节省40%+。如图14所示，离线任务系统只需新增和集群容量预测服务的查询API交互，系统原先的伸缩能力可以复用，仅需投入很小的适配工作量。其中，AOM为应用运维管理，OBS为对象存储服务。

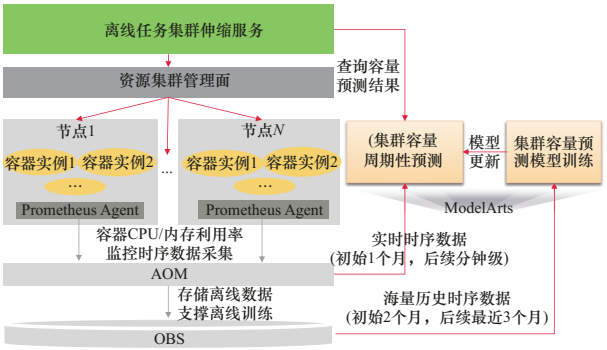


图14　互联网公司A离线任务系统示例

4.2　私有云的资源利用率提升

如何有效提升IT算力的总体利用率，最大限度减少不必要的资源闲置浪费，实现花更少的钱办更多的事，是所有政府和企业自建和运维的私有云所关注的核心焦点与永恒主题。

华为流程IT编译构建云（百万核规模）：编译构建业务是一种典型的Job任务调度执行系统，从业务负载上看属于重载高弹性业务，主要的挑战是维持资源池算力容量和任务量的合理平衡。结合柔性计算的概率卷积和智能弹性伸缩等关键技术优化，能够使编译构建云节省35%的资源。编译构建云与柔性计算技术的集成关系如图15所示。

具体地，本文首先基于任务的历史监控数据对任务的资源需求进行画像，为任务推荐最佳资源规格。其次将概率卷积叠加算法应用到任务实例的调度系统中，以合理评估任务容器叠加后的资源用量，指导调度系统进行合理的资源复用。与基于固定比例复用的调度策略相比，柔性计算将任务运行时长的波动从分钟级缩短至秒级。同时，通过集群容量预测指导调度系统对各个虚拟机资源池进行智能弹性伸缩，为未来任务流量提前预热虚拟机，以及及时释放闲置虚拟机。与基于固定时段的水位预留策略相比，柔性计算达成了35%的构建虚拟机资源节约。

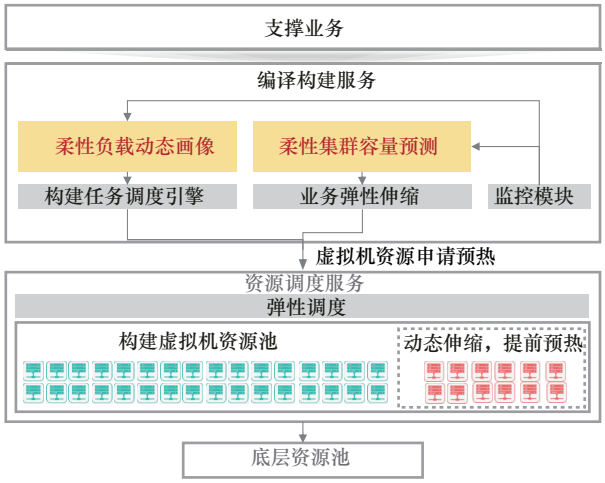


图15　编译构建云与柔性计算技术的集成关系

此外，华为云与某银行私有云团队配合，对该银行私有云的虚拟机资源指标数据进行了收集和分析，以评估引入柔性计算后带来的收益。根据虚拟机的规格和资源用量对这些虚拟机分别采用弹性计算调度和柔性计算调度算法进行模拟调度，其中弹性计算调度取1~3倍3档固定超分比配置，并对比在固定物理机数量的条件下，两者可以发放的虚拟机的数量，如图16所示。从图16可以看到，与该私有云生产环境的3倍固定超分相比，柔性调度可将vCPU分配数量提升约122%（29 320核提升至65 166核），资源利用率可由原来的16.2%大幅提升至34.4%。

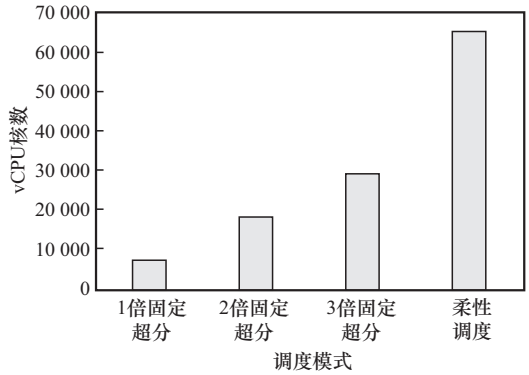


图16　私有云的虚拟机资源指标

5　结束语

针对当前云计算领域普遍采用的弹性计算范式所暴露的资源利用率与应用性能保障间的结构性矛盾，本文突破传统资源分配率依赖型调度模式的局限性，提出面向算力互联网的新型柔性计算架构及智能调度体系。相较于传统弹性计算存在的资源超配冗余与性能波动缺陷，柔性计算通过数据驱动+AI 使能的方式创新性地实现了算力资源的精准适配与弹性伸缩。在华为云数据中心的大量实验和实践表明：在承载同等业务规模下，相比弹性计算，柔性计算可节省算力资源投入 40%+。因此，柔性计算实现了在保障实例应用性能满足 SLA 的前提下，为租户节省了大量算力成本，同时有效提升了数据中心的资源利用率。

下一步工作将通过积累华为云现网海量的日志数据并结合更先进的机器学习算法进一步优化柔性计算架构中预测模块的准确率和泛化性。在算力互联网场景，柔性计算提供了一种普适的算力供给范式，为用户提供极致品价比的算力服务。

参考文献：

[1] 余晓晖 . 算力互联网对形成新型生产关系的作用逻辑与实践方式[J]. 人民论坛·学术前沿, 2024(9): 13-18.
YU X H. The logic and practical means of the arithmetic Internet on new-type production relations[J]. Frontiers, 2024(9): 13-18.

[2] 刘诗萌 . 专访工信部信息通信发展司负责人：2024 年中国算力总规模达 280EFLOPS，有序推进算力网络建设和应用丨两会时间[N]. 华夏时报, 2025-03-11.
LIU S M. Interview with the head of the Information and Communication Development Department of the Ministry of Industry and Information Technology: By 2024, the total scale of China’s computing power will reach 280EFLOPS, and the construction and application of computing power networks will be promoted in an orderly manner|Two Sessions[N]. ChinaTimes, 2025-03-11.

[3] 余晓晖 . 加快推进算力互联互通，构建算力服务统一大市场[N]. 人民邮电报, 2024-03-06.
YU X H. Accelerate the interconnection of computing power and build a unified market for computing power services[N]. People’s Posts and Telecommunications News, 2024-03-06.

[4] ABHISHEK V, LUIS P, MADHUKAR K, et al. Large-scale cluster management at Google with Borg[C]//2015 Proceedings of the 10th ACM European Conference on Computer Systems (EuroSys). New York: ACM Press, 2015: 182-199.

[5] SCHWARZKOPF M, KONWINSKI A, ABD-EL-MALEK M, et al. Omega: flexible, scalable schedulers for large compute clusters[C]//Proceedings of the 8th ACM European Conference on Computer Systems. New York: ACM Press, 2013: 351-364.

[6] ZHANG C, BI J, ZHOU Y, et al. HyperV: a high performance hypervisor for virtualization of the programmable data plane[C]//Proceedings of the 2017 26th International Conference on Computer Communication and Networks (ICCCN). Piscataway: IEEE Press, 2017: 1-9.

[7] ALEXANDRU A, MARC B, ANDREEA F, et al. Firecracker: lightweight virtualization for serverless applications[C]//Proceedings of the 17th Usenix Conference on Networked Systems Design and Implementation (USENIX). New York: ACM Press,, 2020: 54-76.

[8] ERIC J, JOHANN S S, VIKRAM S, et al. Cloud programming simplified: a berkeley view on serverless computing[R]. 2019

[9] SHI J C, FU K H, CHEN Q, et al. Characterizing and orchestrating VM reservation in geo-distributed clouds to improve the resource efficiency[C]//Proceedings of the 13th Symposium on Cloud Computing. New York: ACM Press, 2022: 94-109.

[10] 郭静, 胡存琛, 包云岗 . 面向多应用混部的性能保障方法综述[J]. 计算机研究与发展, 2024, 61(1): 43-65.
GUO J, HU C C, BAO Y G. Survey on guaranteeing the performance of co-located applications[J]. Journal of Computer Research and Development, 2024, 61(1): 43-65.

[11] 张鹏程, 魏芯淼, 金惠颖 . 移动边缘计算下基于联邦学习的动态 QoS 优化[J]. 计算机学报, 2021, 44(12): 2431-2446.
ZHANG P C, WEI X M, JIN H Y. Dynamic QoS optimization method based on federal learning in mobile edge computing[J]. Chinese Journal of Computers, 2021, 44(12): 2431-2446.

[12] YANG H L, BRESLOW A, MARS J, et al. Bubble-flux: precise online QoS management for increased utilization in warehouse scale computers[J]. ACM SIGARCH Computer Architecture News, 2013, 41(3): 607-618.

[13] GUO J, HU C C, BAO Y G. Survey on guaranteeing the performance of co-located applications[J]. Computer Research and Development, 2024, 61(1): 43-65.

[14] CAO W P, TAO X, PAN Y H, et al. QoS perception for cloud databases: necessity, trends, and challenges[C]//Proceedings of the 2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS). Piscataway: IEEE Press, 2024: 1-2.

[15] DELDARI S, SMITH D V, XUE H, et al. Time series change point detection with self-supervised contrastive predictive coding[C]//Proceedings of the Web Conference. New York: ACM Press, 2021: 3124-3135.

[16] 华为云 . QoS 劣化检测模型使用的数据集[R]. 2025.
Huawei Cloud. The dataset used for QoS degradation detection model[R]. 2025.

[17] CAO W P, GU J J, MING Z, et al. Flexible computing: a new framework for improving resource allocation and scheduling in elastic computing[J]. IEEE Transactions on Services Computing, 2025, 18(1): 198-211.

[18] QIU Y Y, CAO W P, XIAO Z J, et al. WGGAL: a practical time series forecasting framework for Dynamic cloud environments[C]//Knowledge Science, Engineering and Management. Berlin: Springer, 2024: 16-27.

[19] LIU Y, HU T, ZHANG H, et al. iTransformer: inverted transformers

are effective for time series forecasting[C]//Proceedings of 12th International Conference on Learning Representation. Vancouver: ICLR, 2024: 210-218.

[20] LI H C, BERGER D S, HSU L, et al. Pond: CXL-based memory pooling systems for cloud platforms[C]//Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. New York: ACM Press, 2023: 574-587.

[21] ZHONG Y H, BERGER S D, CARL A, et al. Managing Memory Tiers with CXL in Virtualized Environments[C]/Proceedings of the 18th USENIX Symposium on Operating Systems Design and Implementation (OSDI).Berkeley: USENIX Association, 2024: 37-56.

[22] 华为. 华为云耀云服务器Flexus X实例[R]. 2025.
Huawei. Huawei cloud Yaoyun server Flexus X instance[R]. 2025.

[作者简介]

顾炯炯（1973– ），男，陕西咸阳人，华为云计算技术有限公司工程师，主要研究方向为云计算架构、柔性计算等。

李佩珊（1993– ），女，河南开封人，中国信息通信研究院助理工程师，主要研究方向为云计算和算力调度等。

曹伟朋（1990– ），男，河南洛阳人，博士，人工智能与数字经济广东省实验室（深圳）副研究员、硕士生导师，主要研究方向为机器学习、AI4Cloud、柔性计算等。

蔡智源（1988– ），男，广东汕尾人，华为云计算技术有限公司工程师，主要研究方向为华为云资源利用效率提升和业务性能QoS保障。

徐传飞（1984– ），男，辽宁沈阳人，博士，人工智能与数字经济广东省实验室（深圳）副研究员，主要研究方向为自然语言处理、大模型、数据挖掘等。

闫丹（1989– ），女，辽宁沈阳人，中国信息通信研究院工程师，主要研究方向为云计算和多元算力等。

毛馨纬（1993– ），男，湖南邵阳人，中国信息通信研究院助理工程师，主要研究方向为存储技术、算力应用等。

明仲（1967– ），男，江西赣州人，博士，人工智能与数字经济广东省实验室（深圳）教授，主要研究方向为云计算、人工智能、柔性计算等。