 DOI: https://doi.org/10.12349/iser.v7i5.10535

Research on Resource Scheduling Optimization Strategy in Cloud Computing Environment

Zhe Wu Jian Xu Mingzi Tao

Xingcheng City Tax Bureau, State Taxation Administration China, Xingcheng, Liaoning, 125100, China

Abstract

Against the backdrop of the continuous evolution of cloud computing environments, computing resources exhibit highly virtualized and dynamic characteristics, and the efficiency of resource scheduling directly affects overall system performance and service quality.Focusing on resource scheduling optimization, a systematic analysis is conducted on key influencing factors in the scheduling process from the perspectives of resource heterogeneity, task diversity, and load fluctuations.By integrating the optimization concepts of heuristic and meta-heuristic algorithms, the mechanisms of multi-objective collaborative scheduling models in improving resource utilization, reducing task completion time, and enhancing system stability are explored.Meanwhile, from the perspectives of resource awareness, load prediction, and elastic scaling, the implementation approaches of scheduling strategies are further examined.Through the construction of a multi-dimensional performance evaluation system, the applicability and optimization potential of scheduling strategies in complex application scenarios are analyzed, providing technical support for improving the level of resource scheduling in cloud computing environments.

Keywords

cloud computing; resource scheduling; scheduling optimization; load balancing; multi-objective optimization

云计算环境下资源调度优化策略研究

吴哲 徐健 陶洺梓

国家税务总局兴城市税务局，中国·辽宁 兴城 125100

摘要

在云计算环境持续演进的背景下，计算资源呈现高度虚拟化与动态化特征，资源调度效率直接影响系统整体性能与服务质量。围绕资源调度优化问题，从资源异构性、任务多样性及负载波动等特征出发，对调度过程中的关键影响因素进行系统分析。结合启发式算法与元启发式算法的优化思路，探讨多目标协同调度模型在提升资源利用率、降低任务完成时间及增强系统稳定性方面的作用机制。同时，从资源感知、负载预测及弹性伸缩等技术路径角度，对调度策略的实现方式进行梳理。通过构建多维性能评价体系，进一步分析调度优化策略在复杂应用场景中的适用性与优化空间，为提升云计算资源调度水平提供技术参考。

关键词

云计算；资源调度；调度优化；负载均衡；多目标优化

1 引言

随着信息技术体系不断向分布式与服务化方向演进，云计算平台逐步成为支撑大规模数据处理与应用部署的重要基础设施。计算、存储与网络资源在虚拟化环境中实现高度集成，不同类型任务在同一资源池中并发运行，使得资源分配过程呈现出显著的复杂性与动态性。资源调度不仅需要兼顾资源利用效率，还需满足响应时间、服务质量以及系统稳定性等多维需求。在实际运行过程中，资源异构结构与业务负载波动相互叠加，传统静态调度方式难以适应复杂环境下的运行要求。与此同时，任务规模持续扩大，调度策略面临更高的实时性与精细化管理需求。围绕资源调度优化展开系统分析，有助于提升云平台整体运行效率，并推动资源管理模式向智能化与协同化方向发展。

【作者简介】吴哲（1974–），女，满族，中国辽宁兴城人，本科，中级，从事云计算环境下资源调度优化策略研究。

2 云计算环境下资源调度优化的理论基础与关键特征

2.1 云计算资源调度的基本内涵与体系结构

云计算资源调度是指在虚拟化资源池中，根据任务需求与系统状态，对计算、存储及网络资源进行动态分配与协调管理的过程，其核心在于实现资源利用效率与服务质量之间的平衡。体系结构通常由资源感知层、调度决策层与执行控制层构成，资源感知层通过监测CPU利用率、内存占用率及网络带宽等指标，实现对资源状态的实时采集，监测周期可控制在1 s至5 s范围内以保证数据时效性。调度决策层基于任务特征与资源状态构建调度模型，采用队列调度、优先级调度等机制完成资源分配策略生成。执行控制层通过虚拟机调度、容器编排及任务迁移等手段落实调度指令，保证调度策略在分布式环境中的高效执行。三层结构之间通过数据反馈形成闭环，使调度过程具备动态调整能力。

2.2 多维资源属性对调度机制的影响特征

云计算环境中的资源呈现多维属性特征，包括计算能力、存储容量、网络延迟及能耗水平等指标，这些属性之间存在复杂耦合关系，对调度机制产生直接影响。计算资源以CPU主频与核心数为主要衡量指标，不同节点之间性能差异可达到20%至50%，导致任务执行时间存在显著波动。存储资源受I/O吞吐能力制约，当磁盘读写速率低于150 MB/s时，数据密集型任务的处理效率明显下降。网络资源方面，带宽利用率与时延对分布式任务执行具有关键影响，在跨节点通信场景中，网络延迟每增加10 ms，任务整体响应时间可能提升5%以上。能耗属性逐渐成为调度决策的重要约束因素，在高密度数据中心中，单节点功耗通常在300 W至600 W之间，调度策略需在性能与能耗之间进行权衡。多维属性的综合作用，使调度机制需要具备多指标协同处理能力[1]。

3 云计算环境下资源调度存在的主要问题与瓶颈

3.1 资源异构性对调度效率的制约

云计算平台由多类型硬件与虚拟资源构成，不同节点在计算性能、存储能力及网络条件方面存在明显差异，这种异构性使资源调度难以实现统一优化。部分节点CPU性能差异可达到30%以上，导致相同任务在不同节点上的执行时间出现明显偏差，从而降低整体调度效率。存储设备类型差异同样影响调度结果，固态硬盘与机械硬盘在I/O性能上的差距可达到3倍以上，使数据密集型任务在资源分配过程中面临性能不均衡问题。网络拓扑结构的不一致也会导致通信效率波动，在跨机架环境中，数据传输延迟明显高于同机架节点。异构资源环境下，传统基于平均分配的调度策略难以实现最优效果，资源利用率往往下降至50%至60%，制约系统整体性能提升。

3.2 动态负载变化引发的调度不稳定性

云计算环境中的任务负载具有显著波动特征，业务请求在短时间内可能出现快速增长或下降，负载变化幅度可达到40%以上，给资源调度带来较大挑战。在高峰时段，资源需求迅速增加，若调度策略响应不及时，容易造成部分节点过载，CPU利用率超过90%，系统响应时间显著上升。低负载阶段则可能出现资源闲置现象，利用率降至40%以下，导致资源浪费。负载变化还会引发频繁的任务迁移与资源调整，迁移过程中的数据复制与状态同步会增加系统开销，迁移时延一般在100 ms至500 ms之间，影响整体调度稳定性。动态环境下，调度策略若缺乏预测能力与快速响应机制，将难以保持系统运行的平衡状态。

3.3 多任务竞争下资源分配冲突问题

在多租户与高并发场景中，大量任务同时竞争有限资源，容易引发资源分配冲突，影响系统运行效率。任务在请求CPU、内存及网络资源时存在优先级差异，若调度机制缺乏合理区分，关键任务可能因资源不足而延迟执行。资源竞争还会导致队列拥塞现象，任务等待时间显著增加，在高负载条件下，平均等待时间可增长至原有水平的2倍以上。不同任务之间的资源需求模式存在差异，计算密集型与数据密集型任务在同一节点运行时，容易产生资源争用问题，降低整体执行效率。资源分配冲突还可能引发服务质量下降，任务失败率上升，当资源分配不合理时，失败率可超过5%。多任务竞争环境下，调度策略需要具备精细化管理能力，以缓解资源冲突带来的性能损耗[2]。

4 云计算环境下资源调度优化的核心方法体系

4.1 基于启发式算法的调度优化方法

基于启发式算法的调度优化方法通过构建规则驱动的快速决策机制，在有限时间内获得近似最优解，适用于实时性要求较高的云计算环境。典型方法包括最短作业优先调度、最小完成时间调度及最大最小公平分配策略等，通过对任务执行时间、资源占用率及等待时间进行排序，实现资源分配的快速匹配。在实际应用中，可通过定义任务执行代价函数$C=\alpha T+\beta R$，其中T表示执行时间，R表示资源占用量，α与β为权重系数，用于控制调度偏好。调度周期通常控制在100 ms以内，以满足高并发任务处理需求。在规模为1000个任务、200个虚拟机节点的仿真环境中，采用最小完成时间算法可将平均任务执行时间降低约18%，资源利用率提升至75%以上。结合优先级队列机制，对高优先级任务设置响应时间阈值小于50 ms，可有效保障关键业务处理效率。启发式方法具有计算复杂度低、实现简单的特点，在动态调度场景中能够快速响应资源变化。

4.2 基于元启发式算法的全局优化策略

元启发式算法通过模拟自然过程或群体行为实现全局搜索能力，在复杂资源调度问题中具有较强的优化性能。常用方法包括遗传算法、粒子群优化算法及蚁群算法等，通过编码任务与资源映射关系，构建适应度函数以衡量调度方案优劣。以遗传算法为例，染色体长度与任务数量一

致，通过交叉率0.8与变异率0.05控制种群演化过程，在迭代次数为200代时，系统可获得较稳定的最优解。在包含500个任务与100个节点的调度实验中，粒子群算法可将任务总完成时间降低约22%，负载均衡度提升至0.85以上。蚁群算法通过信息素更新机制实现路径优化，当信息素挥发系数设定为0.4时，可有效避免局部最优问题。元启发式方法计算复杂度较高，单次调度计算时间通常在1 s至3 s之间，适用于离线优化或周期性调度场景，通过与局部优化策略结合，可提升整体调度性能[3]。

4.3 面向多目标的协同调度模型构建

多目标协同调度模型通过构建综合优化函数，实现资源利用率、任务完成时间与能耗指标之间的协调优化。模型通常采用加权求和或帕累托最优方法进行求解，在加权模型中，目标函数可表示为F= ω 1U+ ω 2T+ ω 3E，其中U为资源利用率，T为任务完成时间，E为能耗水平，权重参数根据业务需求动态调整。在数据中心规模为300台服务器的环境中，通过优化调度模型可将平均CPU利用率提升至80%，任务平均完成时间降低约20%，单位任务能耗下降约15%。模型约束条件包括节点容量限制、任务依赖关系及带宽限制，在网络带宽为10 Gbps条件下，通过合理分配通信任务，可将数据传输延迟控制在20 ms以内。结合多目标优化算法，如NSGA−II，可在200代迭代过程中获得多组帕累托最优解，实现不同调度目标之间的平衡。协同模型通过多指标融合，有效提升复杂环境下调度决策的科学性。

5 云计算环境下资源调度优化的关键技术路径

5.1 资源感知与负载预测技术集成

资源感知与负载预测技术通过对系统运行状态进行实时采集与趋势分析，为调度决策提供数据支撑。资源感知模块基于监控系统对CPU利用率、内存占用率及网络流量进行采集，采样间隔一般设置为1 s至3 s，数据通过时间序列方式存储与处理。负载预测采用长短期记忆网络或自回归模型进行建模，在历史数据长度为1000条的情况下，预测误差可控制在5%以内。在实际调度中，通过预测未来5 min负载变化趋势，可提前调整资源分配策略，使系统负载波动幅度降低约30%。结合资源感知与预测结果，调度系统可动态调整虚拟机分配数量，当CPU利用率超过80%时自动扩展资源节点，低于50%时进行资源回收。技术集成实现了从被动响应向主动调度的转变，提高系统运行稳定性[4]。

5.2 弹性伸缩机制与动态调度策略融合

弹性伸缩机制通过动态调整资源规模，实现资源供给与需求之间的匹配，在云计算环境中具有重要作用。水平扩展通过增加虚拟机或容器实例数量实现负载分担，扩展时间通常在30 s至60 s之间，垂直扩展通过调整单节点资源配置，如CPU核心数从4核提升至8核，实现性能提升。在高并发场景中，当请求量增长至原有水平的1.5倍时，通过水平扩展可将系统响应时间控制在200 ms以内。动态调度策略结合实时监测数据，对任务进行重新分配与迁移，任务迁移时间一般控制在200 ms以内，迁移过程对系统性能影响较小。通过弹性伸缩与动态调度融合，可将资源利用率稳定在70%至85%之间，同时减少资源浪费约20%。该技术路径能够有效提升系统应对突发负载的能力。

5.3 任务优先级与服务质量保障机制设计

任务优先级与服务质量保障机制通过对任务进行分级管理，实现关键业务优先处理与整体服务质量提升。优先级划分通常基于任务类型、响应时间要求及资源需求，将任务分为高、中、低三个等级，高优先级任务响应时间要求控制在100 ms以内。调度过程中采用加权公平队列算法，对不同优先级任务分配资源权重，例如高优先级任务权重设定为0.6，中优先级为0.3，低优先级为0.1，从而保障关键任务的执行效率。服务质量指标包括响应时间、吞吐量及任务成功率，在调度优化后，系统吞吐量可提升约25%，任务成功率提高至98%以上。通过引入SLA约束机制，对违反服务协议的任务进行优先调度或资源补偿，可将违约率控制在2%以下。该机制在多租户环境中有效提升资源分配公平性与服务稳定性[5]。

6 结语

综上所述，云计算环境下资源调度优化是提升系统性能与资源利用效率的关键环节。通过构建多目标调度模型，融合启发式与元启发式算法，并结合资源感知、负载预测及弹性伸缩等技术手段，可有效缓解资源异构与负载波动带来的不利影响。调度策略在保障服务质量的同时，实现资源配置的动态优化与精细化管理，为云平台稳定运行提供有力支撑。

参考文献

[1] 毛锋晨.云计算环境下资源动态分配算法的优化策略[J].信息记录材料,2025,26(02):178−180.

[2] 张桂兰.云计算环境下的资源调度与优化算法研究[J].信息系统工程,2023,(12):64−67.

[3] 石学鹏.云计算环境下的虚拟机迁移算法优化与资源调度研究[J].电子元器件与信息技术,2023,7(08):105−109.

[4] 刘獟.云计算环境下虚拟机动态调度优化策略研究[D].导师：肖志娇.深圳大学,2023.

[5] 赵丽花.云计算环境下基于公平与效率的多资源调度优化研究[D].导师：杜明辉.华南理工大学,2021.