从协同视角论云资源调度技术：综述

王玉钊[1]，于俊清[1]，喻之斌[2]
[1]华中科技大学计算机科学与技术学院，中国武汉市，430074
[2]中国科学院深圳先进技术研究院异构智能计算体系结构与系统研究中心，中国深圳市，518055

摘要：当前公有云中的资源竞争管控仍然是一个悬而未决的问题。新型应用框架（如深度学习和微服务）和专用硬件（如GPU和TPU）的开发与部署给资源管理系统的设计带来新的挑战。现有的解决方案往往为保证应用性能而牺牲集群效率，如资源超额分配导致的低利用率。由于涉及到了软件栈中的不同模块，突破该困境并非易事。尽管如此，产学界为寻找高效的性能隔离和资源调度进行了大量的研究。本文从协同的角度对相关工作进行了全面概述，并揭示其中的技术发展趋势。简言之，本文涉及如下四个主题：不同层次上（包括微体系结构、系统和虚拟层）的资源隔离机制，包括GPU多任务处理；机器层和集群层的资源调度技术，包括面向深度学习应用的GPU调度技术；自适应资源管理技术，包括微服务相关的最新研究；最后探讨了未来的研究方向。希望本文能帮助相关研究人员了解公有云中资源管理技术的概貌，并更好地把握其发展趋势。

关键词：协同；同宿；异构计算；微服务；资源调度技术
https://doi.org/10.1631/FITEE.2100298