多DAG共享资源公平调度策略研究

异构平台下多DAG调度相关算法研究

Abstract

1 绪论

* 1. 研究背景
     1. 云计算
     2. 异构计算系统
     3. 工作流概念简述
  2. 研究内容及贡献
  3. 论文组织结构

1. DAG工作流的分类及相关算法
   1. DAG工作流的分类
   2. 现有的DAG调度算法（HSIP, MMHS等）
   3. DAG工作流调度优化模型
2. 带通信开销的工作流优化模型
   1. 调度模型
   2. DAG工作流相关概念
   3. 本章小结？
3. 兼顾费用与公平的带通信开销的多DAG调度策略
   1. 后向求异原则
   2. 优化算法阐述
   3. 实验设计
      1. 相关算法简介(??)
4. 异构平台上的动态优先级多工作流公平调度算法
   1. 问题描述(urgency的必要性)
   2. 算法描述与分析
   3. 实验设计与分析

1. 云计算

云计算是一种基于互联网的计算模型，它将计算处理资源及相关数据按需提供给其他设备。云计算[1]的出现与发展使得用户可以随时随地动态地以最小的管理成本获取相关计算资源，如计算机网络、存储、服务等。数据中心及其相关软件硬件称之为云(cloud)。目前国内很多公司也相继推出了云服务，如七牛云存储、百度云网盘等。按照服务的不同，广义上可以把云计算分为软件即服务(Software as a Service, SaaS)、平台即服务(Platform as a Service, PaaS)、基础设施即服务(Infrastructure as a Service)。图1.1展示了云服务提供商与云服务消费者的关系。

现如今，通过互联网获取信息、知识已变得稀松平常，而且无需了解其底层原理。而云计算更是该模式的一种延伸[2]。云计算已经成为一种流行的资源提供与获取方式。云服务提供商根据其服务质量(Quality of Service, QoS)向消费者收费；消费者如企业为了寻求降低或在一定程度上避免成本，通过各种各样可能的形式使用云服务。现有的云服务使用方式灵活多样，如可以使用云服务(如百度云)进行数据存储，也可以将自己的任务通过互联网提交到远程服务商进行处理并等待结果的返回。两种方式共同的优点是，不再需要大量的大型存储设备。

云服务商面临的用户群体庞大[3-4]，因而如何为用户提供高效得服务就是云计算中一个重要的问题。而其中一个核心问题就是如何调度云服务用户提交的任务。

云服务提供商在大型服务器上部署多种服务[5] ，根据各个服务的服务质量（quality of service, QoS）属性指定收费标准。用户向服务商提交的工作流由这些服务完成，每个工作流中包含多个任务，每个任务可以由若干个服务完成。

云服务提供商为了实现用户的按需请求服务，必然要显著地冗余配置其基础设施并相应地为低利用率的资源支付高昂的运行成本，或者是按比例拒绝一部分用户的任务请求。前者会为云服务提供商带来额外的不合理的成本，而后者无疑会降低用户体验并且某种意义上讲这种服务也就不再是按需请求的了[6]。而从云服务消费者角度来看，并非所有用户都要求完全意义上的按需云服务，甚至许多用户提交的任务在设计之初就设计为允许中断的可分割的任务。一些系统如SETI@HOME[7]被设计为可以利用空闲的资源，同时也可以被强占。这样设计的好处是，可以在服务中断时可以从故障中恢复。因而，云服务商采取怎样的策略或是算法处理用户提交的任务就变得尤为重要。通常，云服务提供商为了高效地调度用户的任务，会根据具体的调度目标采取不同的调度策略，如最低执行开销（花费）、最低执行时间、负载均衡、公平调度等。一个合理而高效的调度策略应该综合考虑任务的完成时间、费用、可靠性。作业调度或者说任务调度一直是云计算的一个核心问题[8]。

2.异构平台

异构平台由具有不同计算能力的电脑资源组成[9]。由于异构平台计算资源间的异构性，目前提高其计算性能的方法仍旧主要是提高计算资源的数目及其计算能力。比如想异构平台系统添加而随着异构平台内部互相联通的计算资源数量逐渐增加，对合理、高效的调度策略的需求也日益提高。

1. 工作流

工作流最早被商业领域广泛使用。工作流管理系统（Workflow Management Systems, WFMS）最早被引入用来支持一般意义上的需要持久运行的商业过程[10]。