**作业 2**

**deadline: 2023.09.23**

认真阅读论文“PARTIES: QoS-Aware Resource Partitioning for Multiple Interactive Services”，回

答以下问题（论文到网上下载）：

1. 对论文的主要贡献进行简要总结（200 字以内）；

考虑了交互式服务的不同质量需求，并采用了一种自适应的资源分配策略来确保服务的质量，可以在系统资源有限的情况下，最大限度地提高每个服务的用户体验，并提供公平的资源共享，实现多个交互式服务之间的资源分区和调度，从而提高整体系统的性能和用户体验。

资源置换性（fungibility），减少找出一个满足QoS分配所需的时间，对给定的应用与负载，资源上更多的灵活度能满足所有协同调度的应用的QoS，探索可能分配的所用的启发式可以保持相对简单的特性;

提出PARTIES（对多个交互式服务的划分），在线的资源控制器，它能够让多个延迟敏感的应用来在不违反QoS前提下共享一个物理主机,既利用硬件也利用软件隔离机制以保护QoS，并假设没有任何应用的QoS,对比评估了PARTIES和前沿的机制，展示出PARTIES实现了可观的高吞吐量，并在面对波动负载的同时满足QOS，并且它的收益随着混布应用的数目而增加;

PARTIES能在满足应用的QoS前提下大幅提升吞吐率,针对多个延迟敏感（LC）应用的QoS有感资源管理器，它不需要应用的先验知识，通过动态监测，利用系统/硬件层面可划分9个共享资源。PARTIES适用于动态变化的负载，并利用用资源替换性来快速实现收敛。

1. 分析论文所提出方法的主要缺点（500 字以内）；

总搜索空间随着混布的应用的增长而指数级增加，收敛时间实际上增长地更缓慢，PARTIES不会去寻找最优的资源分配，它在所有应用满足QoS时停止搜索，这极大地降低了搜索时间,PARTIES接着依赖Downsize()来进一步缩小当前选择的分配与最优的分配之间的差距；

PARTIES作为运行时（runtime）绑定在核心0上，占它的CPU利用率的15%（监测与资源调整分别占用10%和5%）；

PARTIES花费数秒（当初始划分就有效时）到60多秒（遇到所有6个LC应用混布的最差情况）以收敛到一个没有QoS违规的分配，通常，收敛时间依赖于每个应用的负载，以及混布的应用的数目；

当Memcached与Xapian有相同的负载时，PARTIES的NGINX比Heracles 的高10%-30%，这增加了成本效益；

PARTIES总是以最小延迟松弛度来划分资源，这会导致应用之间的ping-pong；

PARTIES的任何资源调整将花500+ms生效，这会导致资源过度分配；

大部分所研究的应用不涉及磁盘操作，增加他们的内存容量的分配至超过他们各自的数据集规模并不能改善性能；

在控制器参数确定时需注意，决策区间默认是500ms：更短的区间可以更快地侦测出潜在的QoS违规，但过多的调查导致噪声和不稳定的结果；更长的区间提供更好的稳定性，但延迟收敛。

Upsize中延迟松弛度被设为5%：较大的值可找出更多潜在的QoS违规，但是容易出现误报，这有损资源效率；Downsize中则被设为20%：较小的值会导致激进的资源回收，损害性能，但较高的值会降低资源利用率。

依最差案例，触发迁移计时为1分钟。缩短它会有不必要的迁移，延长它会让QoS违规长期存在。

更粗粒度的资源调整可能导致过于激进的资源回收和QoS违规，而更细粒度的调整会延长收敛。

1. 讨论如何改进论文所提出的方法，或者提出新方法解决论文所研究的问题（不限字数）；

文中特征化6个热门的开源LC服务，然后提出资源置换性，Memcached，Xapian，NGINX，Moses，MongoDB，Sphinx，是否应该根据特性分类，主要分为几大类，在大类中选取相关具有代表性的几个交互式服务进行实验，使结果更具说服，避免偶然性，且适当考虑其差异。

文中模型建立主要关注服务质量和资源分配的优化，没有明确考虑能源效率，可以通过引入能源消耗相关的如功耗、能源利用率等指标，建立模型;更有，考虑到网络拓扑结构对服务的影响，也可引入如带宽限制、网络拥塞等因素优化资源分配策略。

如何做到动态资源分配，服务的需求随时间的推移而变化，具体到文中，在控制器参数确定时需注意如何调整参数，决策区间默认是500ms：更短的区间可以更快地侦测出潜在的QoS违规，但过多的调查导致噪声和不稳定的结果；更长的区间提供更好的稳定性，但延迟收敛。Upsize中延迟松弛度被设为5%：较大的值可找出更多潜在的QoS违规，但是容易出现误报，这有损资源效率；Downsize中则被设为20%：较小的值会导致激进的资源回收，损害性能，但较高的值会降低资源利用率。依最差案例，触发迁移计时为1分钟。缩短它会有不必要的迁移，延长它会让QoS违规长期存在。更粗粒度的资源调整可能导致过于激进的资源回收和QoS违规，而更细粒度的调整会延长收敛。

预测机器学习算法的应用，对PARTIES寻找最优的资源分配,因为所有应用满足QoS时停止搜索，这极大地降低了搜索时间，甚至于，充分利用算法，以解决总搜索空间随着混布的应用的增长而指数级增加，以便于缓解消除   PARTIES花费数秒（当初始划分就有效时）到60多秒（遇到所有6个LC应用混布的最差情况）以收敛到一个没有QoS违规的分配，通常，收敛时间依赖于每个应用的负载，以及混布的应用的数目；

提供的思路让我们推广，文中主要考虑有QoS指标，如延迟、带宽等，可引入其它相关的指标，交互式服务用户体验，如流畅度等，CPI/IPC,充分利用其等价性、“资源置换性”。