虚拟化数据中心资源调度研究※

宋维佳[1]，马皓[1]，肖臻[2]，张晓军[1]，张蓓[1]
（1 北京大学计算中心，2 北京大学信息科学技术学院，北京 100871）

摘要：作为云计算的重要支撑技术，虚拟化提供了热迁移、负载转移等技术丰富了云计算资源调度手段。利用这些技术，资源调度要解决如何将计算资源合理分配给服务，一方面保证在负载动态变化的情况下服务质量不受影响，另一方面减少数据中心的能源消耗。本文对虚拟资源调度技术进行了分类，讨论了云计算资源调度的最新研究成果的特点和发展趋势。

关键词：云计算；资源调度；虚拟化；热迁移
中图分类号：TP302.7　**文献标志码：**A

A Survey of Resource Scheduling for Virtualized Data Center

Weijia Song[1], Hao Ma[1], Zhen Xiao[2], Xiaojun Zhang[1], Bei Zhang[1]
(1 Computer Center, 2 Department of Computer Science, Peking University, Beijing 100871, P.R.China)

Abstract: Virtualization technology enriches the tools including live-migration for resource scheduling. Resource scheduler adjusts resource allocation to keep quality of service against load fluctuation as well as save energy. Taxonomy of resource scheduling for virtualized data center are explained in this work. The latest technologies are introduced and compared. We also indicate the trend of its development.

Keywords: cloud computing, resource scheduling, virtualization, live migration

1. 前言

在校园网中，云计算能提高已有网络信息服务的管理维护效率、节约运行成本，并支持如 IaaS 等新服务，有广阔的应用前景。为保证性能，云计算数据中心往往按照最高负载需求部署服务器，因此不少服务器在大多数的时间里都处于空闲状态，造成浪费。调查表明，数据中心服务器利用率一般在 5-15%[1]。而另一方面，数据中心的耗电惊人：根据粗略推算 Amazon EC2[2]能耗占总成本 15%以上。仅 2006 年，美国数据中心总能耗已占全美总耗电 1.5%。

除用节能方法新建或改造数据中心外，由虚拟化发展的服务器整合技术提供了解决上述问题的一种方法。例如，将高峰时段互相错开的不同应用封装在各自的虚拟机中，部署到同一台物理服务器以提高资源利用率。虚拟化数据中心资源调度利用热迁移、负载转移等调度手段挖掘系统剩余性能，保证性能稳定并减少能源浪费。

本文介绍虚拟化数据中心资源调度的研究现状和最新进展，对其进行分析比较和归纳总结。本文结构如下：第二节介绍虚拟化数据中心运行概况和资源调度机制，并对调度技术进行分类；第三节介绍、分析各类资源调度技术的典型研究成果；第四节对各种资源调度技术进行对比；第五节小结。

2. 虚拟化数据中心与资源调度机制

※ **基金项目：**国家发改委基金资助项目（CNGI2008-108），国家自然科学基金资助项目（61170056）
通讯联系人：宋维佳（1982-），男，四川人，博士研究生；E-mail: songweijia@pku.edu.cn

在虚拟化数据中心环境下，应用被封装在一个或多个虚拟机中。例如典型的互联网应用由 Web 服务器前端、中间应用逻辑和后端数据库三个层次构成；这些层次可能各自封装在独立的虚拟机中。有的应用为了处理高负载，复制了多个副本，每个副本各自封装在一个或一组虚拟机中。图 1 为虚拟化数据中心的四台物理机和四个应用的运行情况，第 n 个应用的第 m 台虚拟机记为“应用 n.m”。

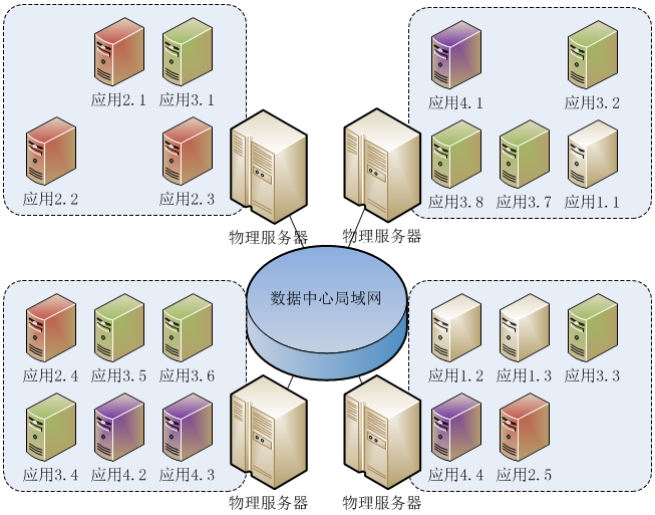

图 1 虚拟化数据中心示意图

虚拟化资源的调度存在于两个层次：局部调度涉及如何在虚拟机之间合理共享物理机的 CPU、内存和 IO 资源，由虚拟机管理器负责；全局调度解决如何优化组合虚拟机、在物理机间平衡负载，这相当于把资源合理地分配给应用。本文主要讨论后者。

资源调度技术由策略和机制两部分组成。全局资源调度主要有三种机制。最便于实施的是静态服务器整合：根据应用对资源需求的特点，结合物理服务器的性能，寻找固定的搭配组合方式将应用汇集到物理服务器上运行。由于负载的波动性，有的物理服务器可能出现过于繁忙的情况，这时需要人工干预解决。

另一种是利用虚拟机热迁移[3]技术，动态调整虚拟机在物理服务器上的优化组合。在负载升高时，把繁忙服务器上的虚拟机迁移到较空闲的服务器上；在负载下降时，把虚拟机从低负载的服务器上集中起来，空闲的服务器则可进入待机的状态节能。

通过调整用户请求在应用副本上的派发来控制虚拟机的负载强度，这种机制称为“负载重定向”。该调度机制利用负载重定向辅以应用的启、停和复制，使负载在物理服务器上合理地分配。

根据调度机制的不同，虚拟化数据中心资源调度技术可相应地分为静态服务器整合、基于虚拟机热迁移的调度和基于负载重定向的调度三种类型。

3. 虚拟化数据中心资源调度研究

3.1. 静态服务器整合

静态服务器整合的第一步是用虚拟机的历史负载和服务品质协议（Service Level Agreement，SLA）确定虚拟机的资源配额；第二步用虚拟机的资源配额和物理服务器的容量来计算服务器整合方案，即每个虚拟机应该在哪些物理服务器上运行。

静态服务器整合的好处是实施简单、稳定可靠。但是，寻找优化的组合方案很困难。首先是应用的负载变化难以被精确把握，虚拟机的资源配额只能是一个近似估计；即使确定了虚拟机的资源配额，在此基础上找到最优方案也是个 NP 完全问题[4]，系统规模较大时获得精确解的时间代价太大。实用方案中，虚拟机资源配额一般被表述为两部分：一部分是保证该虚拟机运行的资源分配下限，另一部分是资源分配超过下限时，性能与资源分配的关系函数（一般是多元线性函数）。寻找优化搭配组合也一般采取近似算法。

把虚拟机看成物品，物理服务器看成箱子；虚拟机各维度资源的配额就是物品的大小而物理服务器各维度资源的性能/容量就是箱子的大小，静态整合问题就转化为矢量装箱问题（vector bin-packing problem）[5]，即可用现有矢量装箱算法进行求解[6,7,8, 9]。

也可用遗传算法模型[10]来描述该问题：把解看成是个体，个体的染色体是一个长度为 N 的序列，其中 N 为虚拟机的个数；序列元素取值为 1 到 M 的整数，M 为物理服务器的个数；第 i 个元素的取值为 h，解释为“第 i 个虚拟机运行在物理服务器 h 上”。 组合的杂交过程通过交换两个染色体的基因片段进行，变异过程则通过随机交换两个物理服务器上的虚拟机完成[4]。若某个染色体对应的解是合理的---即所有服务器都能满足其上的虚拟机资源配额下限，则定义其适应度为系统中所有应用中的最低性能（由性能-资源

配额曲线计算得出）；否则，适应度取值为[-M,-1]中的一个整数，其绝对值为不能满足虚拟机运行条件的物理服务器的个数。也即适应度越高的方案能均衡地顾及所有应用，而适应度低的方案顾此失彼。

实验结果表明，在实用算法中，矢量装箱算法的成功率、最优近似度和计算时间是最优的[4]；在众多矢量装箱算法中，[9]最接近精确解。在时间允许的范围内，基因算法比最简单装箱算法的成功率和优化度更好，但是其优化度不及矢量装箱算法。

3.2. 基于虚拟机热迁移的调度

热迁移[3]将虚拟机透明地迁移到另一台物理机上而不影响其运行。该特性方便于系统资源的动态调整：当某个物理服务器繁忙时，将其上部分虚拟机迁移到别处以保证应用性能；一些物理服务器趋于空闲时，又可以把虚拟机聚集起来，释放空闲的物理机以节约能源。但是热迁移受限于集中式存储，在处理 Map Reduce 等 IO 密集型应用时，存储带宽易成为瓶颈；在成本敏感的场景下，集中式存储较高的价格也是不利因素。

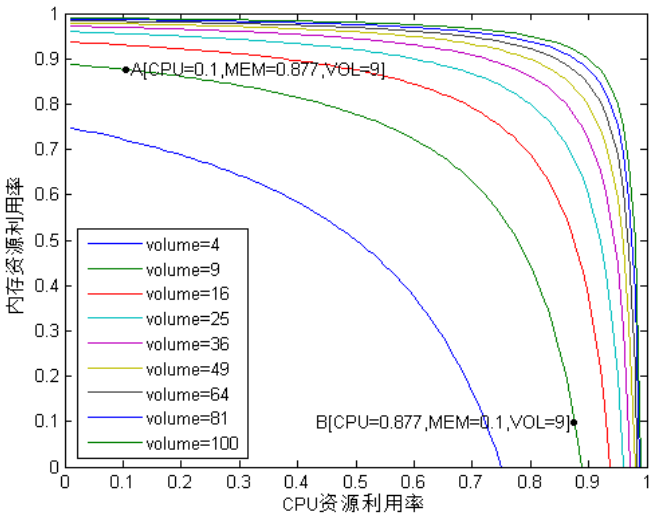

图 2 volume 与资源利用率关系

文献[11]设计了基于热迁移技术的负载均衡调度器 Sandpiper，它由一个集中控制器和各个物理机上的监控器组成。 监控器定期将虚拟机的 CPU、内存和网络 IO 使用统计数据发送给集中控制器（若有可能，从虚拟机内部看到的内存、网络 IO 甚至应用的性能统计也发送给集中控制器；这种方法涉及到虚拟机的内部修改，因而称为灰盒策略）。集中控制器根据各物理机和虚拟机的统计数据判断哪里发生了资源短缺，然后用启发式算法计算迁移调节方案，派发给监控器实施。

Sandpiper 考虑了三种资源类型：CPU、内存和网络，并用一个取值为三类资源空闲率乘积之倒数的 volume 指标来表示物理机或虚拟机的繁忙程度。调度算法从热点（发生资源短缺的物理机）上挑选单位内存承载的 volume 最高的虚拟机，为之寻找当前系统中 volume 最小的物理机作为迁移目标；直到所有热点被消除。引入 volume 指标的好处是将多种资源统一处理，方便调度算法决策，但是它不能表达是哪个资源紧张。如图 2 所示，物理机的 volume 为 9 但是它可能处于 A 点和 B 点两个资源利用率迥然不同的状态。这可能导致从 CPU 过载的物理机上迁走一个大内存而 CPU 并不繁忙的虚拟机。另外，文献[11]仅考虑了负载均衡而没有绿色计算，这两者实际上是可以结合起来[12]。

VirtualPower[13]的目标则是绿色计算。它在虚拟机管理器层为虚拟机提供了软件 ACPI 和 DVFS 的接口，支持虚拟机内部的降频和 ACPI 休眠操作。因各虚拟机降频和 ACPI 操作不同，由物理机的 hypervisor 最终决策通过 CPU 调度、物理降频还是真正的 ACPI 休眠来节能省电。从该研究的结果来看利用硬件 ACPI 和 DVFS 最多只能节省 8%的电能，而结合动态服务器整合腾出空闲物理主机节能的方式能够节省约 34%的电能。文献[13]是第一个用虚拟机迁移技术进行绿色计算的工作，但总的说来，该文献所提的算法[13]还不能算是一个真正的调度算法，它没有考虑如何提高系统性能的问题。

3.3. 基于负载重定向的调度

将负载（如 HTTP 请求）通过应用层交换机重定向到应用某个副本的方法，可以导向负载以控制资源在应用之间的分配。这种模式在非虚拟化环境下就已经广泛使用，其相关研究成果兼具借鉴和实用价值。

文献[14]让每个应用提供自己的估价函数来描述其期待的性能，如：为每分钟 1000 个以内的 HTTP 请求付 1 分/请求，而为每分钟 1000 到 2000 之间付给 0.5 分/请求；而运行物理服务器是有开销的，譬如 0.1 元/分钟；调度算法根据每个应用不同的出价和运行他们的成本来计算资源配比方案以最大化收益。其优点

是可以描述动态的复杂的 SLA，而且同时考虑了负载均衡和绿色计算。文献[14]假设数据中心的所有应用可以在一台物理服务器上同时运行，这在早期的数据中心是可行的；但随着应用的日益复杂和庞大，该假设不再成立。

针对应用庞大复杂的现状，文献[5]建立了一种新的调度模型。该工作只关心 CPU 和内存资源：CPU 资源随着负载的变化而变化，是“负载相关资源”；而应用占用内存的大小几乎恒定不随负载而变化，是“负载无关资源”。把应用的负载看作流，把 CPU 的资源看成边的容量，可把应用在服务器上的布局（placement）转化为一张流图，并用最大流算法求出其可承受的最大负载。如图 3 所示，服务器 A 运行着应用 w、x，服务器 B 运行着应用 x、y、z，服务器 C 运行着应用 z。算法[15]定期地检查流图，确定应用负载是否被消化；否则，尝试根据当前负载和布局重新运行最大流算法看是否能在不改变布局的情况下通过调整负载来满足应用需求；如果尝试失败，再考虑用调整应用布局的方法来解决。

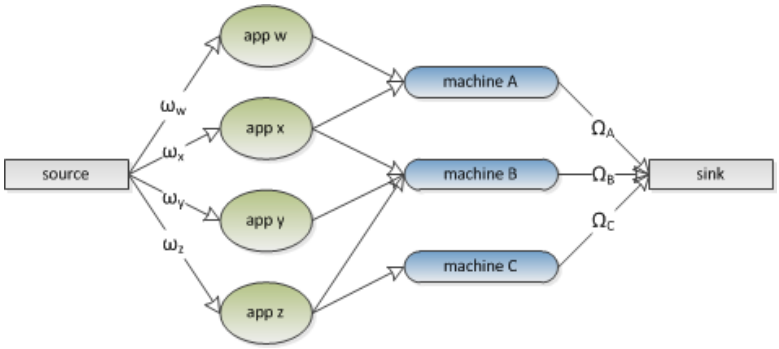

图 3 用流图表达应用在物理服务器上的布局

虽然针对当今应用复杂庞大，提出了合理的资源调度算法，但是[5]和[15]缺乏对节能的考虑。并且，与其他大多数基于负载调度的应用一样，只能处理应用中无状态的层次，不能解决后端存储的性能瓶颈。

3.4. 对比分析

从适用范围、调度开销、是否支持绿色计算、是否支持集中式存储、是否要求虚拟机无状态，以及对 SLA 的支持等角度对三种资源调度技术进行比较，如表 1 所示。

表 1 三种资源调度技术的对比汇总

	静态服务器整合	基于虚拟机热迁移的调度	基于负载重定向的调度
适用范围	负载变化有规律，起伏平缓的应用	所有应用	基于请求-响应机制的 Internet 应用
调度开销	当前静态整合不再合理时，需要管理员手工进行调整	虚拟机迁移引起的网络开销，与虚拟机内存镜像大小成正比	虚拟机复制时的网络带宽和启动时间。
绿色计算	否	是	是
依赖集中式存储	否	是	否
要求虚拟机无状态	否	否	是
对 SLA 支持	不好，只能预先估计、不能动态调整。	不能支持应用层的 SLA	好，可以支持应用层的 SLA

4. 小结

虚拟化数据中心环境下，资源调度主要解决保证应用的性能和节能省电两个互相制约的问题。根据调度机制的种类本文将资源调度技术分为静态服务器整合、基于虚拟机迁移的调度和基于负载重定向的调度三种类型。本文介绍了三种调度类型的最近研究成果，接着分析了各种调度技术的优缺点。静态服务器整合简单稳定，但适应性不好；基于虚拟机热迁移的调度适用范围广，但是依赖于集中式存储，调度网络开销较大；基于负载重定向的调度支持应用层的 SLA，也不依赖于集中式存储，但是只支持基于请求-响应的应用，而且要求应用无状态。三种资源调度方法并不是非此即彼的，静态整合可以优化系统虚拟机的初始布局；而基于负载重定向的调度机制因为代价较低，可以在迁移物理机之前进行调整；热迁移作为高代价

的调度手段，在前述手段都不能很好解决问题时使用。从目前的发展来看，集中式存储是大趋势，基于虚拟机热迁移的调度技术应用范围将越来越广。另外，针对的 SLA 的资源调度技术更加受到重视。

参考文献

[1] “Server Consolidation”, http://www.vmware.com/solutions/consolidation/

[2] “Amazon EC2 Pricing”, http://aws.amazon.com/ec2/pricing/

[3] Christopher Clark, Keir Fraser, Steven Hand, Jacob Gorm Hansen, Eric Jul, Christian Limpach, Ian Pratt, and Andrew Warfield. “Live migration of virtual machines”. In Proceedings of the 2nd conference on Symposium on Networked Systems Design \& Implementation - Volume 2 (NSDI'05), Vol. 2. USENIX Association, Berkeley, CA, USA, 273-286.

[4] Mark Stillwell, David Schanzenbach, Frederic Vivien, Henri Casanova, “Resource allocation algorithms for virtualized service hosting platforms”, Journal of Parallel and Distributed Computing, Volume 70, Issue 9, September 2010, Pages 962-974, ISSN 0743-7315

[5] A. Karve , T. Kimbrel , G. Pacifici , M. Spreitzer , M. Steinder , M. Sviridenko , A. Tantawi, “Dynamic placement for clustered web applications”, Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland

[6] Chandra Chekuri, and Sanjeev Khanna, “On Multi-dimensional Packing Problems”, SIAM Journal on Computing, 2004

[7] Chandra Chekuri, and Sanjeev Khanna, “On multi-dimensional packing problems”, Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms, 1999

[8] L. T. Kou , G. Markowsky, “Multidimensional bin packing algorithms”, IBM Journal of Research and Development, v.21 n.5, p.443-448, September 1977

[9] William Leinberger , George Karypis , Vipin Kumar, “Multi-Capacity Bin Packing Algorithms with Applications to Job Scheduling under Multiple Constraints”, Proceedings of the 1999 International Conference on Parallel Processing, p.404, September 21-24, 1999

[10] J. H. Holland, “Adaptation in Natural and Artificial Systems”, MIT Press, Cambridge, MA, 1992

[11] Wood, T., Shenoy, P., and Arun. “Black-Box and gray-box strategies for virtual machine migration”. In Proceedings of the ACM Symposium on Networked Systems Design and Implementation (NSDI’07). 229--242.

[12] Gong Chen, Wenbo He, Jie Liu, Suman Nath, Leonidas Rigas, Lin Xiao, and Feng Zhao, “Energy-aware server provisioning and load dispatching for connection-intensive internet services”, Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, 2008

[13] Ripal Nathuji , Karsten Schwan, “VirtualPower: coordinated power management in virtualized enterprise systems”, Proceedings of twenty-first ACM SIGOPS symposium on Operating systems principles, October 14-17, 2007, Stevenson, Washington, USA

[14] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, Amin M. Vahdat, and Ronald P. Doyle, “Managing energy and server resources in hosting centers”, Proceedings of the eighteenth ACM symposium on Operating systems principles 2001

[15] Chunqiang Tang, Malgorzata Steinder, Michael Spreitzer, and Giovanni Pacifici. 2007. “A scalable application placement controller for enterprise data centers”. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 331-340.