云计算架构下的深层次动态资源调配

王志钢 汪小林 罗英伟

（北京大学信息科学与技术学院 北京 100871）

摘 要 云计算是当前学术界和工业界都十分关注的热点，被广泛应用于针对海量数据和用户的大规模计算。云计算的特点要求计算机系统能够提供可伸缩的计算能力，而虚拟化技术正是其中的关键层次，在资源管理、服务器整合、提高资源利用率等方面发挥了巨大的作用。通过虚拟化技术，可以实现一个多层次的资源调度机制，以保证高资源利用率和系统性能：首先面向虚拟机的应用特征建立资源预测模型，然后依据预测结果建立资源分配策略，最终通过虚拟机间的资源动态优化技术，实现在同一物理主机或不同物理主机上虚拟机间动态的资源优化使用。这里，不仅要以物理机的宏观资源利用率作为管理依据，更需要关注虚拟机上应用程序在运行过程中的资源需求变化特征，从而为云计算提供一整套的虚拟化资源优化技术及使用方案，从静态部署、动态预测、单机资源动态调配、多机资源动态均衡调度、在线迁移等多个层次为云计算提供全面、有机的支撑。

关键词 虚拟机；云计算；数据中心；资源管理；性能

Dynamic Resource Balancing in Cloud Computing Framework

WANG Zhi-gang WANG Xiao-lin LUO Ying-wei

(*Department of Computer Science and Technology, Peking University, Beijing* 100871, *China*)

Abstract Cloud computing brings the potential to deliver flexibility, consolidation, and high resource utilization to data centers, and virtualization is the key layer. High resource utilization as well as high performance promised by virtualization largely depends on an effective and efficient resource management scheme. There is a great demand on building a multi-layered resource management mechanism via virtualization: using the characters of the virtual machine to build a predictive resource model and then export the allocation strategy, finally the resource can be allocated dynamically on demand. We should focus not only on the resource utilization rate on the physical machines, but also the change of resource demand while running for the application on the virtual machines. The final target is to get a suit of virtual resource management technology, which could be widely used in the static deployment, dynamic prediction, and resource management between virtual machines or physical machines, to support the cloud computing powerfully.

Keywords virtual machine; cloud computing; data center; resource management; performance

1 引 言

随着软硬件技术的高速发展，高性能、低开销、易维护性的计算促进了现代数据中心的快速发展，如何在保证程序性能的前提下，有效地维护分布式的计算资源的高效使用成为了一个富有挑战性的问题。由于虚拟化技术[1]在服务器整合、硬件资源分配、节能等方面具有灵活易操作的特点，近年来 在数据中心以及云计算架构中得到了越来越广泛的应用。

目前，虚拟化技术和云计算已经高度融合，业界已经有了一系列可以应用的产品。这类产品和相关的

基金项目：国家自然科学基金项目（61170055，61272158，61232008）、高等学校博士学科点专项科研基金项目（20110001110101）、国家高技术研究发展计划（863计划）项目（2012AA010905）。

作者简介：王志钢，博士研究生，研究方向为系统虚拟化、云计算，E-mail: cellfires@gmail.com；汪小林，博士，副教授，研究方向为系统虚拟化、云计算和高效能计算；罗英伟，博士，教授，研究方向为高效能计算。

研究，大都是在虚拟机这个宏观层面，以虚拟机为管理对象，通过虚拟在线机迁移技术[2]，实现虚拟机在物理机上的部署、调度和管理，这已经为云计算服务商在快速部署、保证服务质量、提高资源利用率等方面提供了很大的帮助。但是，未来云计算的海量规模将给虚拟化技术带来了更大的挑战。一方面，虚拟化技术不可避免地带来了整个虚拟计算环境的性能降低，因此我们需要更高效的资源虚拟化方法来应对云计算的需求；另一方面，仅仅通过虚拟机级别的迁移技术来实现资源管理是不够的，还应该针对云计算的需求，从更细微的层次提供更精确和有效的资源管理方法。

因此，需要在以下多个方面进行深入的研究：首先，分析虚拟环境下应用程序的特征，研究应用程序行为分析和实时监控方法，建立资源需求的预测模型，通过预测模型为资源的调度提供依据；其次，根据应用程序的特征及其资源需求预测结果，一方面提供面向应用的、高效的资源虚拟化方法；另一方面，从系统虚拟资源管理着手，解决云计算服务的虚拟机动态部署与调度过程中多层次的动态资源管理问题，保证云计算服务质量，提高资源利用效率。

2 相关工作

系统级虚拟化技术源于上世纪60年代，其核心思想是：在一台物理主机上虚拟出多个虚拟计算机（Virtual Machine，VM），每个虚拟计算机可以看作一个物理主机的复本，各虚拟计算机相互隔离，其上能同时运行独立的操作系统，这些客户操作系统（Guest OS）通过虚拟机管理器（Virtual Machine Monitor，VMM）访问实际的物理资源。

目前，大多数虚拟机资源管理方法都是通过虚拟机迁移技术来完成的。虚拟机迁移是指将一台主机（源机器）上运行的虚拟机迁移到另一台主机（目的主机）上运行。为了达到这个目标，需要将虚拟机的运行状态从源机器传输到目的主机，然后在目的主机上恢复虚拟机的运行。在线迁移是指在整个迁移过程中，虚拟机的暂停时间非常短，虚拟机上运行的服务始终能响应用户的请求。目前大部分虚拟机迁移的研究都只关注源主机和目的主机共享磁盘存储的情况，在这样的情形下，只需要迁移虚拟机的内存和CPU状态。Xen[3]和VMware两大虚拟机管理器都实现了共享存储的虚拟机在线迁移[4]，分别称为Xen live migration[5]和VMware Vmotion。

气球技术（Ballooning）是虚拟机系统中所特有的虚拟存储技术[6]，可以从其它虚拟机窃取一些未使用的机器内存页面，给急需内存的虚拟机使用，进而实现动态调整内存。为了实现内存的窃取，VMM需要在Guest OS的内核中安装一个用于窃取内存的模块，称作“Balloon Driver”。但气球技术只是同一物理机上不同虚拟机之间进行内存调配的机制，如何实施具体的调整还需进一步的研究。我们的前期工作之一MEB[7]就是利用了“气球技术”来动态均衡运行在单个物理主机上的各个VM之间的物理内存。它实现了一个基于VMM的内存预测器，先预测出VM的WSS，然后重新通过气球驱动的接口为各个VM重新分配可用内存。MEB只是初步实现了本地内存资源的调配。当所有VM的内存需求超过物理主机上的可用物理内存时，VM只能通过页面交换来应对内存压力。

Overdriver[8]采用了远程缓存和虚拟机迁移两种技术进行虚拟化环境下的资源动态调控，根据负载时间的长短选择调控方法，汲取了以上两种方式的优点。但是Overdriver存在着两个明显的不足：(1)每台物理主机上预留出一部分内存空间作为远程缓存或迁移使用，而未考虑对本机的VM所产生的影响，造成了内存的浪费；(2)VM的内存分配仍然是传统的静态分配方法，不能在本地主机内部进行有效的调节。

3 虚拟机资源动态调配

虚拟化技术的本质是多个VM复用一组物理资源，由VMM实现对底层资源的分时或分割使用。分时复用是指多个虚拟机使用同一个完整的物理设备，分割复用指虚拟机管理器将实际的物理资源分成若干部分，每个虚拟机及虚拟机管理器能访问其中的一个或几个部分。现有的虚拟化系统，一般用分割复用的方式为不同的VM分配内存和外存等资源，用分时复用的方式为VM分配CPU和各种I/O设备。而这种分配，通常是静态方式进行，即每个VM在启动前就已经分配好了各种资源，这在资源的分配和使用上都是不利的。

在图1所示的应用场景中：要在物理内存为4G的物理机上同时启动两台内存需求为3G的虚拟机，静态分配的方式显然就不能满足需求。

事实上，虚拟机的行为具有很强的可变性，其内存需求往往是动态变化的。在图2所示的应用场景中，虚拟机在运行过程中并不总需要那么大的内存。

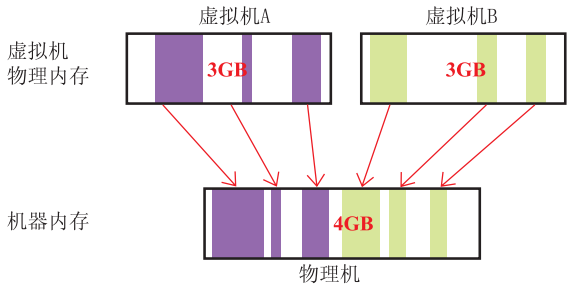

图1　多虚拟机共享内存模型

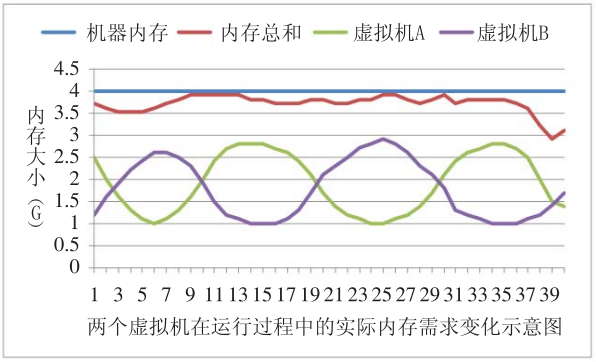

图2　虚拟机内存动态变化

显然，最合适的方法是让虚拟化环境下的资源的管理者VMM为多台虚拟机进行按需、动态调控内存。图2所示为两台虚拟机内存动态调配的理想结果：当一台虚拟机需求增加时，VMM可调配其它虚拟机的空闲内存满足其需求。可见，在资源需求动态变化时，静态分配不利于资源的优化利用，而资源动态调配则可以有效提高资源的利用率。

传统的管理方式对系统资源调度的粒度较大，策略也相对粗糙，而虚拟机的总体性能依赖于对其各方面计算特征需求的均衡满足。因此要实现对系统资源的实时分配和按需分配，就必须对虚拟机的计算特征进行监控和预测。我们应该采用静态分析预测和动态分析相结合的思路，全面研究虚拟机的计算特征，解决虚拟机内部应用层面的资源调度，推动云计算环境中对虚拟资源的管理。

在基于虚拟化技术的云计算服务中，通过为虚拟机分配恰当的资源，来平衡用户服务质量保证和提供商资源开销这对矛盾，是最有效的措施。只有准确地预测虚拟机对资源的需求，才能为它预留出所需的资源，保证虚拟机能即需即得；同时也才能保证当从虚拟机回收部分资源时，不会影响到虚拟机的服务质量。在云计算框架下，我们可以面向虚拟机的应用特征建立预测模型，依据预测结果建立资源分配策略，通过虚拟机间资源动态优化使用技术，实现在同一物理主机内或不同物理主机上虚拟机间动态的资源优化使用。

3.1　资源预测模型

无论是在单机环境下还是在多机环境下，资源协调都是建立在单虚拟机及其上应用程序的资源预测上，问题的核心是单虚拟机的资源需求与其性能、能耗等的关系。因此单虚拟机的资源预测模型是进行资源调控的基础，单一物理主机的资源预测是对其上各虚拟机的汇总；而多机协调则是基于各物理主机的需求进行分布式调配或集中调度。

为了能获得应用系统在虚拟机上运行的特征，我们构想采用静态预测与动态预测相结合、处理器性能监控与动态资源监控相结合、单机资源监控管理与多机资源协同管理相结合等方式，建立一个统一的虚拟机资源预测与管理模型。

首先，对应用系统通过编译和Profiling等方式，静态地预测应用系统对Cache和内存的需求及需求变化的规律，并通过应用程序、Guest OS及VMM等层间的通讯通道把指示性信息传递到虚拟机管理器，支持虚拟机管理器根据指示动态调优系统（包括设置Cache分区大小、调整分配的内存数量以及CPU与IO资源等）。

其次，通过动态预测来弥补静态预测在适应系统变化能力上的不足，对系统运行时资源需求变化进行跟踪监测，并依据历史进行预测，使虚拟机管理器能更全面的调整资源的分配与调度。

在虚拟机执行过程中，如果要进行内存等资源的全面监控，需要占用较大的内存资源和大量的处理器资源。我们可以利用处理器性能监控计数器，首先对系统中的关键指标进行采样，在采样指标达到或超过阈值时，再对特定性质的资源进行针对性的监控，既使监控高效，又保证可对系统资源使用行为进行准确细致地监控分析。

3.2　基于应用程序特征的虚拟机部署

每类应用程序都有其计算的特点，当这些应用程序部署在一个虚拟机上时，就反映为虚拟机的计算特征。虚拟机的计算特征不仅包括其对CPU的使用、对内存的需求、对I/O带宽的占用等方面，还包括系统交互性、内存访问规律等深层次特征。

对于计算特征相同或相似的虚拟机，当把它们部署在同一台物理主机上时，就会存在对相同计算资源的竞争使用，从而影响总体性能。在云计算中服务中，如果能尽量把计算特征相容的虚拟机部署在一

起，则会提高物理主机计算资源的利用率，提高虚拟机的总体性能。

在建立较完备的虚拟机计算特征抽象的基础上，通过静态编译和基准测试等方法，实现对面向应用的特定配置的虚拟机的各方面计算特征的度量是资源预测的进一步目标。当具有不同计算特征的虚拟机在部署于同一物理主机上时，对虚拟机的性能会产生交互影响，特别是当物理主机配置变化时，也会影响该物理机上部署的虚拟机的性能。我们可以总结出具有指导意义的虚拟机部署考查因素表，列出计算特征不相容或相容性差的情况，从而避免把这些虚拟机部署在同一物理主机上；也应总结出计算特征相容性非常好的情况，此时可以把这些虚拟机部署在一起，增强整体性能。

除了CPU利用率、内存使用量、I/O带宽等常规的计算特征外，我们还关注了系统交互性、内存访问规律等深层次的计算特征及其相互间的影响。特别的，对于内存访问规律来说，我们不仅需要检测内存工作集和内存可回收空间，还需要度量和预测Cache的大小对虚拟机性能的影响，以及不同Cache访问模式的虚拟机间相互的性能影响。

3.3 单机环境下虚拟机间的动态资源调整

获得各个虚拟机实时的资源需求之后，就可以按照既定资源调度策略在虚拟机之间进行实时的动态调度，实现系统性能的最大化。

基于应用特征的虚拟机部署可以在很大程度上降低运行在同一物理主机上的多个虚拟机间的相互影响。但是，纯静态的预测并不能反映虚拟机运行状态的动态变化，在虚拟机运行的不同时刻，其计算特征仍然会有所变化。虚拟机所需的实际内存大小就是一个典型的经常变化的计算特征。

通过在虚拟机管理器中增加对虚拟机运行时的计算特征的动态变化的监测，可以采用相应的机制调整虚拟机间的资源分配。例如在内存的预测上，我们可以用内存页面访问的LRU直方图预测虚拟机在一段时间内可以释放多大的内存而不显著影响其性能（如图3所示），通过对交换分区上I/O操作频度的监控可以预测虚拟机的内存不足状况，两者的结合则使得我们可以通过Ballooning机制在虚拟机间按需的调整实际内存分配，使得实际分配给两个虚拟机的物理内存总和小于两者所需的最大内存总和[7]。

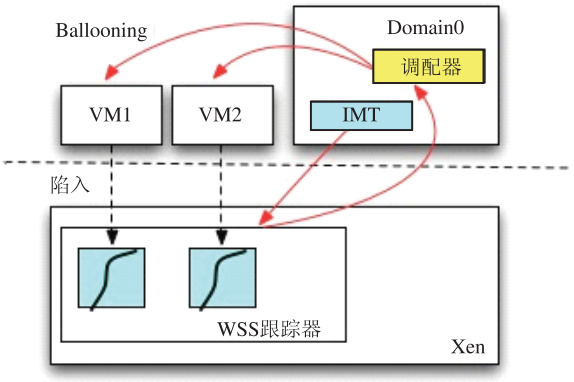

图 3 单机内存资源调配

除了对内存资源的使用情况可以进行监测并做调整外，我们还可以对虚拟机的Cache使用进行监控和调整。

CPU对Cache的使用和主机物理内存地址存在相关性。通过限制虚拟机页面的主机物理地址空间，可使相应的内存页面只使用部分Cache，从而达到Cache分区的目的。Cache分区可以是静态的，这要求在给虚拟机分配页面时，就限定页面的地址空间。但是虚拟机对Cache的影响并非总能预知，因而需要在虚拟机执行过程中，通过页面迁移，把客户物理页面从一个主机页面复制到另一个主机页面，改变页面与Cache的对应关系，以实现动态的Cache划分。页面迁移的代价一般较大，不仅需要复制整个页面，并且需刷新对应的Cache，因而并非迁移页面就一定能带来总体性能的提升。一种好的方式是仅对部分热点页面进行迁移，这样既可减少迁移的总开销，也可增加迁移页面与Cache的相关度[10]。

在多核计算中，CPU调度包括每个虚拟机所获得的核的个数及各个核的时间片的分配。现有的操作系统对进程级CPU调度已有了成熟的算法，这些可直接用到VMM中。我们更要关注不同CPU调度算法的特点，针对云计算环境中，因虚拟机迁移导致同一物理主机上的虚拟机组合特征动态变化的特点，针对不同的组合特征采用针对性的CPU调度算法，使得虚拟机总体性能和个体性能得到平衡。同时，我们还可以研究对单机CPU的利用率的监控，为多机协调提供依据。

对于I/O密集型的虚拟机而言，提高I/O性能对虚拟机性能的提高至关重要。虚拟I/O设备对I/O性能有较大的影响，目前主要的I/O虚拟化优化方向是虚拟机直接设备访问，但是直接设备访问会削弱虚拟机的易管理特性。因此，如果能够采用虚实结合的I/O设

备虚拟化模型，可以很好地改进虚拟机I/O的综合性能：在虚拟机需要高的I/O性能时，通过从虚拟I/O设备切换到物理I/O设备，实现直接I/O设备访问，达到高效I/O的目的；而在虚拟机较空闲，需要通过虚拟机迁移实现能耗管理时，则从物理I/O设备再切换回虚拟I/O设备，从而能够方便的迁移虚拟机。

3.4　多机环境下虚拟机间的动态资源调整

单机资源需求主要表现在CPU计算能力、内存及I/O几个方面。在云计算中，可利用分布式计算、集中管理、P2P等技术，各个虚拟机管理器之间可通过互相通讯来了解彼此的资源需求。当一个物理机的计算能力或网络带宽不够时，将虚拟机在线迁移到有充分资源的物理机是一个可行方案。我们也可以考虑利用虚拟机克隆技术，将资源需求高的虚拟机一分为二，每个虚拟机分配原先的一部分任务，以减少单一虚拟机的资源需求，增加其适配到原物理机或迁移到新宿主机的机会。

当一个物理机的资源不够时，我们可以优先在本地进行资源调度，本地资源不足时，如果过载时间相对较短，我们会通过网络来使用远程机器上的资源（如远程内存[9]）以缓解资源压力；如果过载时间相对较长，我们就把虚拟机整体迁移到资源充裕的机器上。

图4显示了全局内存动态调控的原理。在实现调控策略时，我们采取的原则是“先本地，后远程”，即尽可能地利用本地物理主机的内存和气球技术满足虚拟机的内存需求，当本地物理内存超载时，再借助于数据中心内其他的物理主机来缓解内存压力。在多机环境中，远程内存交换和迁移是调控内存分配，均衡负载的常用技术。二者各有千秋，远程内存交换的

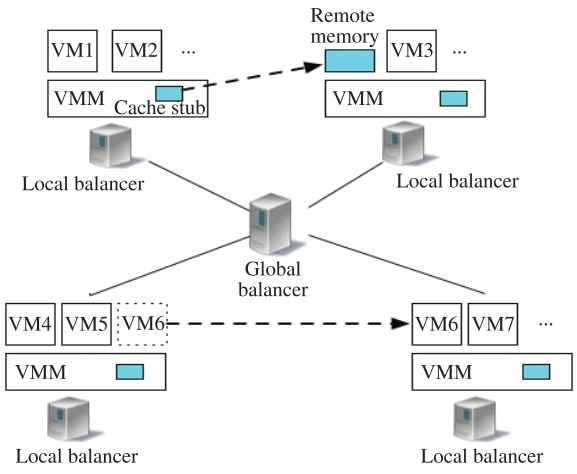

图4　多机资源调度模型

代价较小，因而适用于短期的内存超载；相反虚拟机迁移的代价较大，因此适用于持续的内存超载，但合理地诱发迁移确是一个难题。

3.5　合理的虚拟机迁移诱发机制

虚拟机迁移可以在大粒度上实现多机环境下的动态资源调整，但是迁移本身时间较长（一般要在分钟级别），开销较大（需要占用大量带宽、并降低虚拟机性能），因此，只有在预期虚拟机计算特征明显且将较持久的改变为另一种状态时，进行迁移才是最优的选择。

我们需要针对云计算中同一数据中心内虚拟机的资源需求以及整个环境的节能需求，设计并实现一套合理诱发虚拟机迁移的机制：一方面，当一台物理主机内的资源调配无法满足某个虚拟机的资源需要时，则在环境内选择另一台适当的物理主机，将该虚拟机或其他虚拟机迁移过去；另一方面，在不影响虚拟机资源需求时，可以在环境内将运行在更多物理主机上的虚拟机合并到少数物理主机上，从而可以关闭部分物理主机，达到节能的目的，在性能和节能之间找到一个平衡点。

3.6　支持多个数据中心之间的虚拟机迁移

在共享网络存储的局域网环境内，虚拟机的在线迁移只需拷贝内存状态到目标主机。但是在云计算平台下，多个企业数据中心互联，每个数据中心有单独的存储设备，这就意味着考虑虚拟机迁移时必须要同时迁移磁盘的状态，也即虚拟机的全系统迁移。影响虚拟机的在线迁移的因素包括客户机内存大小、应用程序的内存工作集、回写脏页率、网络带宽等。尤其在跨数据中心的环境下，有限的网络带宽所造成的延迟将会加大虚拟机的停机时间。针对上述问题，我们需要一个支持全系统迁移同时尽量减少虚拟机停机时间的机制，来支持多个数据中心之间虚拟机的无缝迁移。

4　总　结

当前，如何针对云计算中的虚拟化支撑技术展开研究，从虚拟机深层次的资源管理着手，解决云计算服务的虚拟机动态部署与调度过程中多层面的动态资源管理问题，保证云计算服务质量，提高资源利用效率，是一个值得深入研究的课题。这需要我们研究的虚拟机资源管理涉及虚拟机资源预测、虚拟机部署、单机资源动态调配、多机资源动态协调、在线迁移等

多个层面。

在以虚拟化技术为基础的云计算服务中，大量虚拟机运行在众多的物理主机上，此时，需要对虚拟机资源进行全面有效的动态综合管理，才能既满足服务质量，又提高资源利用效率，但这很难通过单一的机制和策略来解决。以内存资源的使用为例，对于两个运行在相同物理主机上的虚拟机，在它们都分配到了足够的物理内存资源时，对于一些应用来说，通过调整低级Cache的共享与划分策略，会显著地改善两个虚拟机的总体性能。在一个虚拟机内存紧张，而另一个虚拟机有空闲内存时，可以通过诸如Ballooning等机制把空闲的内存页面回收并重新分配给内存紧张的虚拟机，缓解内存紧张造成的性能降低。而当一台物理主机上的其他虚拟机都无法释放空闲页面给内存紧张的虚拟机时，一种可行的方法是选择一个虚拟机在线迁移到另外一台能够满足资源需求的物理主机上去，释放出来的内存资源则可用来缓解其他虚拟机的内存紧张。

因此，未来的研究主要应考虑如何把多个层面的监控分析和预测的机制有机地融合在一起，动态识别影响性能的关键因素并预测虚拟机对资源的需求，同时设计各种资源的优化分配和使用技术，进而选取有效的整合方案对资源进行动态再分配及整合，以保证既提升虚拟机的性能，又提高资源利用效率。

参 考 文 献

[1] Goldberg R P. Survey of virtual machine research [J]. IEEE Computer, 1974, 7: 34-45.

[2] Wood T, Shenoy P, Ramakrishnan K K, et al. A platform for optimized WAN migration of virtual machines [C] // Proceedings of the ACM Special Interest Group on Programming Languages international conference on Virtual execution environments. New York, USA, 2011.

[3] Barham P, Dragovic B, Fraser K, et al. Xen and the art of virtualization [J]. Special Interest Group on Operating Systems, Operating Systems Review, 2003, 37(5):164-177.

[4] Clark C, Fraser K, Hand S, et al. Live migration of virtual machines [C] // Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation. Berkeley, USA, 2005: 273-286.

[5] Nelson M, Lim B, Hutchins G. Fast Transparent migration for virtual machines [C] // Usenix Annual Technical Conference, 2005.

[6] Carl A. Waldspurger. Memory resource management in vmware esx server [J]. Special Interest Group on Operating Systems, Operating Systems Review, 2002, 36(SI): 181-194.

[7] Zhao W M, Wang Z L, Luo Y W. Dynamic memory balancing for virtual machines [C] // Special Interest Group on Operating Systems, Operating Systems Review, 2009.

[8] Williams D, Weatherspoon H, Jamjoom H. Overdriver: handling memory overload in an oversubscribed cloud [C] // Proceedings of the Special Interest Group on Programming Languages International Conference on Virtual Execution Environments, New York, USA, 2011.

[9] Comer D, Griffioen J. A new design for distributed systems: the remote memory model [C] // In Usenix Summer Conference, 1990: 127-135.

[10] Wang X L, Wen X, Li Y C, et al. Dynamic cache partitioning based on hot page migration [J]. Frontiers of Computer Science, 2012,6(4): 363-372.