

## BOSS 在 KVM 平台中的性能研究与优化

马震太<sup>1,2</sup>, 张晓梅<sup>1</sup>

(1. 中国科学院高能物理研究所 计算中心, 北京 100049; 2. 中国科学院大学, 北京 100049)

**摘 要:** 针对 BESIII 离线软件系统 (BOSS) 在内核虚拟机平台上的性能损耗, 结合 BOSS 作业特点给出相关优化方法。对引起性能损耗的各种因素进行研究, 并根据不同客户机规模的测试情况, 定量分析测试结果, 进而确定性能损耗。对 BOSS 作业在物理机和客户机上的性能进行测试, 结果表明, 优化后模拟作业性能损耗降低至 1.1% ~ 1.6%, 重建作业与分析作业性能分别提高 2.6% ~ 4.5%, 7% ~ 18.7%。

**关键词:** BESIII 离线软件系统; 内核虚拟机; 中央处理单元迁移; 透明大页; 磁盘预分配; 多客户机

**中文引用格式:** 马震太, 张晓梅. BOSS 在 KVM 平台中的性能研究与优化[J]. 计算机工程, 2017, 43(7): 70-74.

**英文引用格式:** Ma Zhentai, Zhang Xiaomei. Performance Research and Optimization of BOSS on KVM Platform[J]. Computer Engineering, 2017, 43(7): 70-74.

## Performance Research and Optimization of BOSS on KVM Platform

MA Zhentai<sup>1,2</sup>, ZHANG Xiaomei<sup>1</sup>

(1. Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

**[Abstract]** Aiming at the performance loss of BESIII Offline Software System (BOSS) on Kernel Virtual Machine (KVM) platform, this paper gives related optimization suggestions combining BOSS operation characteristics, and studies the factors which lead to performance loss through quantitative analysis and test results, and determines performance loss. It explores the performance loss of multiple clients. BOSS job performance test on client machina and physics machina, results show that the performance loss optimize simulation job is decreased to 1.1% ~ 1.6%, the performance of reconstruction job is increased by 2.6% ~ 4.5%, the performance of analysis job is increased by 7% ~ 18.7%.

**[Key words]** BESIII Offline Software System (BOSS); Kernel Virtual Machine (KVM); Central Processing Unit (CPU) migration; Transparent Huge Page (THP); disk preallocation; multi-client

**DOI:** 10.3969/j.issn.1000-3428.2017.07.012

### 0 概述

虚拟化即资源的抽象化, 包括单一物理资源的多个逻辑表示和多个物理资源的单一逻辑表示。虚拟化技术通过整合分散的物理资源, 提供统一的管理和服务, 大幅提高了资源利用率; 资源整合不仅减少了物理设备的投入, 还降低了硬件成本和管理成本, 实现资源共享。虚拟化的封装性降低了资源使用的复杂度, 隔离性为用户提供安全、高效的应用环境, 使得用户部署系统更加快捷, 系统维护和资源管理更加简便。

目前常见的企业级虚拟化产品有: VMware, HyperV, Xen, KVM。由于 KVM 是开源项目, 因此

Linux 的新技术可应用到 KVM 上。KVM 的解决方案绝大部分免费, 且拥有数量巨大的技术支持<sup>[1]</sup>。鉴于以上特点, KVM 成为虚拟化平台的首要选择。对 KVM 进行调研发现, KVM 客户机运算能力与物理机差距在 10% ~ 20% 之间<sup>[2]</sup>, 且磁盘性能表现稍差<sup>[3-4]</sup>, 故而应用程序在 KVM 平台上运行时仍存在一定的性能损耗。

北京谱仪 (BESIII) 是北京正负电子对撞机上的大型通用谱仪<sup>[5]</sup>。BESIII 离线软件系统 (BESIII Offline Software System, BOSS) 是基于 Gaudi 框架, 根据 BESIII 探测器和实验的实际需要开发的离线数据处理框架<sup>[6]</sup>。已知 BOSS 系统每年需要处理的数据多达 120 TB, 随着实验的深入, 数据量将不断地积

**基金项目:** 国家自然科学基金 (11375221)。

**作者简介:** 马震太 (1988—), 男, 硕士研究生, 主研方向为离线软件系统、云计算; 张晓梅, 副研究员。

**收稿日期:** 2016-07-07    **修回日期:** 2016-08-31    **E-mail:** mazt@ihep.ac.cn

累增加<sup>[7]</sup>。为提高资源利用率,降低复杂性<sup>[8]</sup>和减少成本,将 BOSS 作业迁移到虚拟化平台,对虚拟化资源进行管理显得十分必要,此时 BOSS 作业在 KVM 平台的性能损耗问题则成为关注的焦点。本文针对 BOSS 在 KVM 平台上的性能损耗问题,结合 BOSS 作业自身的特点展开研究,以期降低 BOSS 作业的性能损耗。

## 1 性能损耗

本节对 BOSS 进行介绍,阐述实验环境,研究 BOSS 作业的特点以及性能表现。

### 1.1 BOSS 简介

BOSS 是根据 BESIII 探测器以及实验的实际需要开发的离线数据处理框架。事例是 BOSS 作业的基本数据单元,对于一组事例数据所进行的数据处理过程,称为一个作业<sup>[9]</sup>。作业是 BOSS 软件的基本执行单元,包括系统初始化、事例处理循环和作业结束。事例处理循环次数取决于事例集大小,每次循环可分为读取、处理、存储 3 个阶段<sup>[10]</sup>。本文主要研究模拟作业、重建作业和分析作业。模拟作业用于研究高能物理实验中各种随机物理过程,以及物理量的统计分布和相关性质。重建作业将原始数据所记录的数字信号还原为与粒子相关的物理量,形成重建数据。分析作业对重建数据进行计算和统计以进行物理分析<sup>[11]</sup>。

### 1.2 实验环境

硬件环境: CPU Intel (R) Xeon (R) CPU E5-2630L v2@ 2.40 GHz, 2 个处理器, 每个 6 核; 内存 48 099 MB; 磁盘容量 825 GB。

软件环境: 操作系统 Scientific Linux 6.5; libvirt 版本 0.10.2; 文件系统 ext4; 客户机镜像格式 qcow2。

测试作业: 模拟作业运行 10 000 个  $J/\psi \rightarrow \rho\pi$  事例; 重建作业运行 10 000 个  $J/\psi \rightarrow \rho\pi$  事例; 分析作业输入文件包括 552 530 个  $J/\psi$  一般衰变事例。

实验方法: 分别在物理机和客户机上进行 3 轮测试, 取平均值, 由两者的测试结果确定性能损耗。

### 1.3 基准工具的性能损耗

因为 HEPSPEC06 工具能准确衡量高能物理环境中 CPU 的性能损耗, 所以本文采用此工具进行测试, 测试结果表明, KVM 客户机相对于物理机性能下降 9.7%。

本文选用基准测试工具 Lmbench<sup>[12]</sup> 测试内存的性能损耗, 发现 KVM 客户机访存性能相对于物理机访存性能下降 4.6%。

本文选用基准工具 iozone 对磁盘性能进行测试<sup>[13]</sup>, 测试文件大小依次设为 2 GB, 4 GB, 6 GB,

8 GB, 10 GB, 因 BOSS 作业的记录大小接近 1 MB, 故 iozone 测试记录大小采用 1 MB, 对测试结果求平均值。由实验数据可知, 相对于物理机, KVM 客户机磁盘写性能下降 15.5%, 磁盘读性能下降 9.3%。

### 1.4 BOSS 作业的性能损耗

依次对 1, 2, 4, 6, 8, 10, 12 个 BOSS 作业在物理机和 KVM 客户机上的性能进行测试, 对于不同的作业类型, BOSS 软件有着不同的性能表现。实验数据表明, 模拟作业性能损耗 3% ~ 4.2%, 重建作业性能损耗 4.3% ~ 7.7%, 分析作业性能损耗 10.7% ~ 33.4%。

用 perf 工具进行监测, 得到 BOSS 作业的 CPU 时间百分比如表 1 所示。

表 1 BOSS 作业的 CPU 时间百分比 %

作业类型	CPU 时间百分比
模拟作业	99.4
重建作业	97.8
分析作业	94.2

由表 1 可知, BOSS 作业的运行时间主要消耗在 CPU 计算上。

对单个模拟作业和重建作业的运行状况进行监测, 资源消耗数据如表 2 所示。从表 2 可以看出, 在模拟作业和重建作业运行初期, cache 消耗较大。

表 2 BOSS 作业的资源消耗 MB

作业类型	实际占用物理内存	Cache 消耗	Buffer 消耗
模拟作业	621	1 235	49
重建作业	762	1 667	45

## 2 性能调优

由表 1 结果可知, BOSS 在 KVM 平台的性能并不理想, 这是因为开发商为了最大程度地保证客户机的兼容性, 一些对客户机性能影响显著的因素并没有得到合理配置。本文对影响 KVM 客户机性能的各个因素展开研究, 以期实现 BOSS 作业在 KVM 平台上的性能优化。

### 2.1 CPU 特性

现代处理器拥有各种增强计算性能的 CPU 特性, 如单指令多数据扩展指令集 SSE、对浮点运算性能提升非常明显的 AVX 指令集等<sup>[14]</sup>, 这些 CPU 特性通过减少作业运行所需的指令数目和降低指令平均时钟周期数来提高计算性能。如图 1 所示, 在创建客户机时, 内核查询物理 CPU 支持特性集合并通知 QEMU, QEMU 默认配置一个兼容性较强的 CPU 特性子集, 然后将该子集返回给内核, 内核根据该子集模拟出虚拟 CPU, 提供给客户机使用。默认子集虽然兼容性较强, 但却会造成一部分的性能损耗。

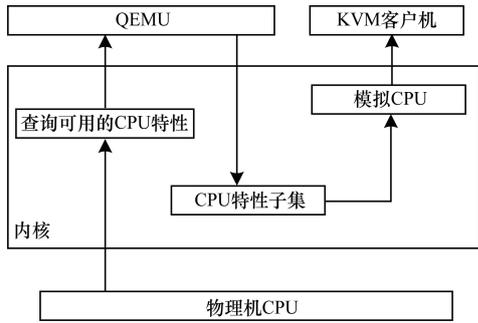


图1 CPU特性的原理

为使客户机 CPU 运算性能最大化,将物理 CPU 支持的所有 CPU 特性传递给客户机。选择 astart 寻路算法进行测试,在客户机继承物理机 CPU 特性后,算法性能提高了 1.9%。

### 2.2 非统一内存访问架构

非统一内存访问架构 (Non Uniform Memory Access Architecture, NUMA)<sup>[15]</sup> 采用分布式存储器模式,处理器可以访问全部内存。然而,每个核及内存在跨互联时都有一个小的性能代偿。由于性能代偿的存在,随着处理器和内存间距离的增大,处理器访问内存的速度逐渐下降<sup>[16]</sup>。为提高客户机的访存性能,应将客户机绑定到某个节点,并从该节点分配内存,避免远程内存访问带来的性能损耗。

启动 2 台客户机,用 virsh 工具进行 NUMA 调优<sup>[17]</sup>,分别将一台客户机绑定到节点 0,另一台绑定到节点 1,使得客户机与 NUMA 节点一一对应,原理如图 2 所示。

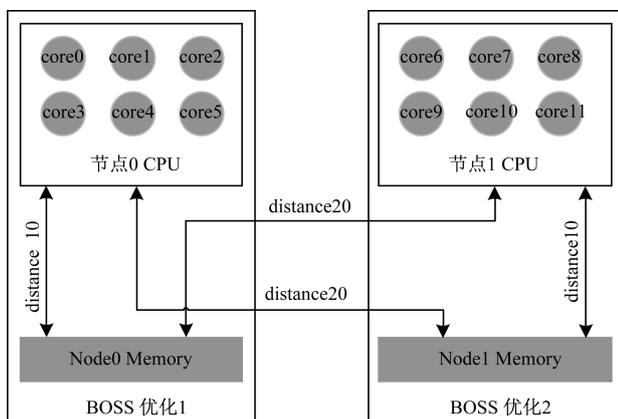


图2 客户机 NUMA 调优原理

在对客户机进行 NUMA 调优前后,对单个 BOSS 作业运行过程中的 CPU 迁移事件进行监测,统计 CPU 的迁移次数,调优前后 CPU 迁移次数对比如图 3 所示。

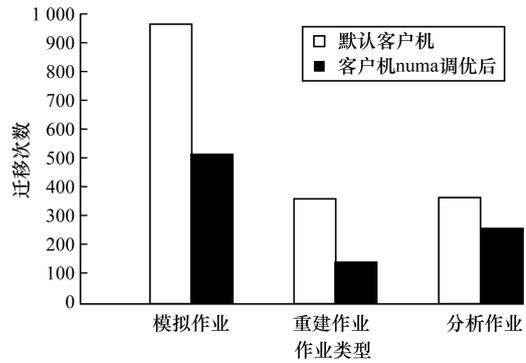


图3 BOSS 作业调优前后 CPU 迁移次数对比

由图 3 可知,NUMA 调优后,BOSS 作业的 CPU 迁移次数大幅减少,进而降低了缓存命中的失效次数,减少了不必要的性能损耗。

### 2.3 透明大页

当前系统为 64 位,采用四级页表管理虚实映射,如图 4 所示,每个页表项占据 8 Byte,若作业需要 2 MB 的内存,则要经历 512 次快表未命中事件和 512 次缺页中断;访问 2 MB 内存,需访存 2 048 次。当采用 2 MB 大页作为分页单位时,页表影射简化为三级页表,如图 5 所示,此时只需经历 1 次快表未命中事件和一次缺页中断,就可为 2 MB 的内存空间建立虚实映射;而访存只需 3 次即可<sup>[18]</sup>。可见,大页的使用显著提高了系统的性能<sup>[19]</sup>,但大页需要预留,预留未用会造成内存的浪费;透明大页 (Transparent Huge Pages, THP) 既有大页的优点,又避免了上述缺点<sup>[20]</sup>。

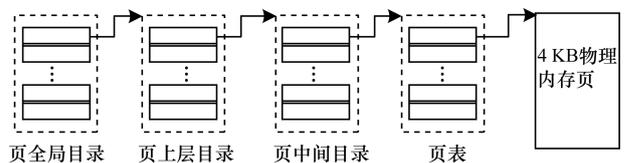


图4 四级页表映射

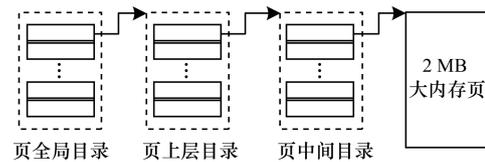


图5 三级页表映射

### 2.4 相关因素

KSM 提高了内存资源的利用率<sup>[21]</sup>,但会为 CPU 带来额外的负担<sup>[22]</sup>。EPT 技术直接在硬件上支持内存地址转换,提升了内存虚拟化的性能<sup>[23]</sup>。因分析作业磁盘 I/O 吞吐率相对较高,故提高 I/O 性能很有必要。KVM 客户机磁盘的 full 预分配模

式在分配磁盘时不仅建立元数据, 而且在每个字节填充 0, 这在任何情况下都保证了磁盘空间, 性能最好; cache 采用回写模式时, 页缓存和磁盘写缓存均开启, 客户机 I/O 性能较好。

### 2.5 调优效果

将物理机、客户机优化前、客户机优化后 3 种作业的实验结果进行对比, 对比数据如图 6 ~ 图 8 所示。

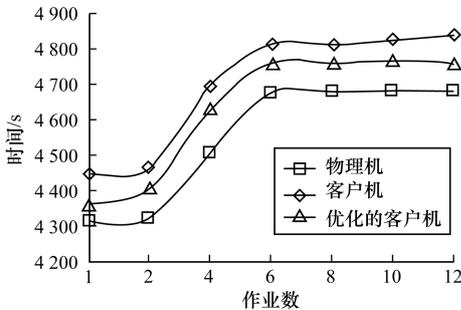


图 6 模拟作业优化前后结果对比

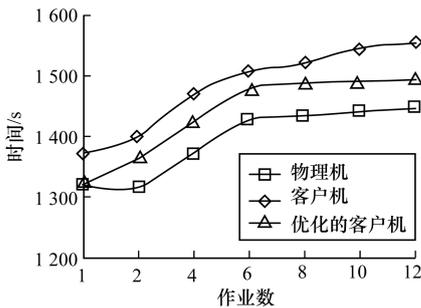


图 7 重建作业优化前后结果对比

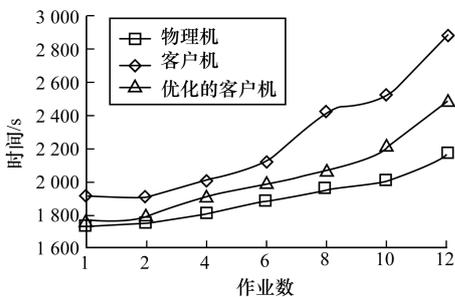


图 8 分析作业优化前后结果对比

分析实验数据可知, 优化以后, 模拟作业性能损耗降低至 1.1% ~ 1.6%, 重建作业性能提高 2.6% ~ 4.5%, 分析作业性能提高 7% ~ 18.7%。

### 3 多客户机的性能损耗

上节本文成功降低了 BOSS 作业在 KVM 平台的性能损耗。然而, 在实际应用环境中仍存在客户机规格多种配置的问题。本节就这一问题展开实验, 首先在物理机上运行 12 个作业, 取得实验结果;

然后依次按以下 6 种客户机规格进行实验: 1) 启动 1 个客户机, 分配 12 个 CPU 核心和 46 600 MB 内存; 2) 启动 2 个客户机, 每个客户机分配 6 个核心和 23 300 MB 内存; 3) 启动 3 个客户机, 每个客户机分配 4 个核心和 15 534 MB 内存; 4) 启动 4 个客户机, 每个客户机分配 3 个核心和 11 650 MB 内存; 5) 启动 6 个客户机, 每个客户机分配 2 个核心和 7 767 MB 内存; 6) 启动 12 个客户机时, 每个客户机分配 1 个核心和 3 884 MB 内存。

对模拟作业进行测试, 得到多客户机的性能分布曲线如图 9 所示, 横轴表示客户机的数量; 纵轴表示 KVM 客户机上模拟作业相对物理机的性能损耗百分比。

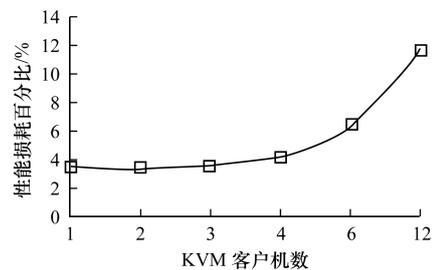


图 9 模拟作业多客户机性能损耗分布

由图 9 可知, 当 12 个 KVM 客户机并行时, 性能出现大幅下滑。用 nmon 对 2 种情况进行监测, 得到两者初始化阶段 CPU 分布如图 10、图 11 所示。

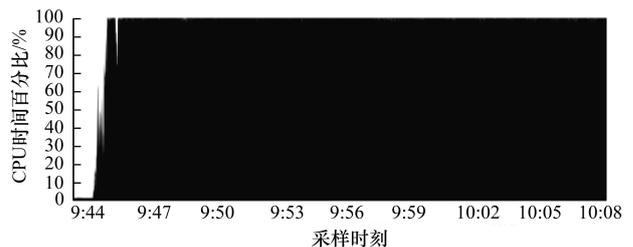


图 10 单个客户机模拟作业初始化阶段

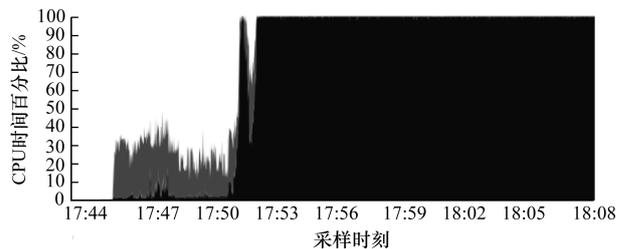


图 11 12 个客户机并行模拟作业初始化阶段

本文结合分析实验数据, 单个客户机运行时作业耗时 4 841.6 s, 12 个客户机并行时作业耗时 5 243.7 s, 作业运行时间延长 402.1 s。1 个客户机运行时初始化等待时间为 35 s, 而 12 个客户机并行

时初始化等待时间为434 s,2种情况的初始化等待时间相差399 s。由以上数据可知,2种情况的运行时间差别主要体现在初始化阶段。经研究发现,在初始化阶段,模拟作业需要加载库文件、几何数据、实验数据,导致作业运行前期 cache 消耗较大,同时伴随着客户机数量的增加,磁盘逐渐饱和,由此引起的时间延迟逐渐加大,性能逐渐下滑,进而导致12个客户机并行时,性能大幅下滑。

与模拟作业类似,重建作业和分析作业也出现了性能下滑现象,原因相同。若要避免这个问题,则采用磁盘阵列不失为一个有效的方法。

#### 4 结束语

本文针对单机环境中 BOSS 作业在 KVM 客户机上的性能损耗给出了相关优化建议,优化后 BOSS 作业的运行性能得到明显提高。在多客户机并行运行时,初始化阶段出现磁盘饱和,进而导致性能下滑的现象。然而,若要大规模应用,则需要应用到 OpenStack,OpenNebula 等虚拟化管理平台。同时,在实际应用环境中,经常需要通过网络文件系统来读取输入文件,本文并没有考虑网络通信所带来的性能损耗。下一步将在虚拟化管理平台上,对大规模 KVM 客户机上 BOSS 作业的性能进行实时监控,分析收集到的性能数据,结合作业调度系统建立模型,以期得到更理想的性能调优方案。

#### 参考文献

- [1] 肖力,汪爱伟,杨俊俊,等. 深度实践 KVM [M]. 北京:机械工业出版社,2015.
- [2] 高清华. 基于 Intel VT 技术的虚拟化系统性能测试研究 [D]. 杭州:浙江大学,2008.
- [3] 张新玲,张东,曹玲玲,等. 云计算虚拟化平台性能研究 [J]. 软件导刊,2013,12(11):1-3.
- [4] 孙琳程. 客户机 KVM 与 XEN 的性能分析 [J]. 电脑知识与技术,2013,9(10):2364-2366.
- [5] 王贻芳. 北京谱仪的设计与研制 [M]. 上海:上海科学技术出版社,2011.
- [6] Arnault C. CMT: A Software Configuration Management Tool [C]//Proceeding of CHEP'00. Padova, Italy: [s. n.], 2000:236-245.
- [7] Barrand G, Belyaev I, Binko P, et al. GAUDI—A Software Architecture and Framework for Building HEP Data Processing Applications [J]. Computer Physics Communications, 2001, 140(1/2):45-55.
- [8] Hoshina K, Fujii K, Nitoh O. Development of Geant4 Solid for Stereo Minijet Cells in a Cylindrical Drift Chamber [J]. Computer Physics Communications, 2003, 153(3):373-391.
- [9] Zou Jiaheng, Li Weidong, Huang Xingtao. SNiPER: An Offline Software Framework for Non-collider Physics Experiments [J]. Journal of Physics, 2015, 664(72):1-5.
- [10] 张晓梅. BESIII 离线数据处理软件框架的研究和开发 [D]. 北京:中国科学院研究生院,2005.
- [11] Li Weidong, Liu Huaimin, Deng Ziyang. The Offline Software for the BESIII Experiment [C]//Proceedings of CHEP'06. Mumbai, India: [s. n.], 2006:215-222.
- [12] 李春艳. 小规模高性能虚拟集群关键技术研究 [D]. 昆明:云南大学,2014.
- [13] 吕庆翰. 面向虚拟化的综合性能测评方法研究 [D]. 广州:华南理工大学,2015.
- [14] 胡伟武,陈云霁,肖俊华,等. 计算机体系结构 [M]. 北京:清华大学出版社,2011.
- [15] 任永杰,单海涛. KVM 虚拟化技术实战与原理解析 [M]. 北京:机械工业出版社,2014.
- [16] Manchanda N, Anand K. Non-uniform Memory Access (NUMA) [D]. New York, USA: New York University, 2012.
- [17] Rajput V, Kumar S, Patle V K. Performance Analysis of UMA and NUMA Models [J]. International Journal of Computer Science Engineering and Technology, 2012, 10(2):1457-1458.
- [18] 宏伟唐. Linux 大页使用与实现简介 [EB/OL]. (2009-09-10). <http://www.ibm.com/developerworks/cn/linux/1-cn-hugetlb/index.html>.
- [19] Jain S. Design and Implementation of Huge-page Aware Memory-overcommitment Solution Using Ballooning and Sharing [D]. Bombay, India: Indian Institute of Technology, 2013.
- [20] Arcangeli A. Transparent Hugepage Support [C]//Proceeding of KVM'10. Boston, USA: Red Hat Inc., 2010:126-135.
- [21] Jones M T. Anatomy of Linux Kernel Shared Memory [EB/OL]. (2010-04-07). <http://public.dhe.ibm.com/software/dw/linux/1-kernel-shared-memory/1-kernel-shared-memory-pdf.pdf>.
- [22] Arcangeli A, Eidus I, Wright C. Increasing Memory Density by Using KSM [C]//Proceedings of Linux Symposium. Montreal, Canada: [s. n.], 2009:159-164.
- [23] Bhatia N. Performance Evaluation of Intel EPT Hardware Assist [EB/OL]. (2009-03-30). [https://www.vmware.com/pdf/Perf\\_ESX\\_Intel-EPT-eval.pdf](https://www.vmware.com/pdf/Perf_ESX_Intel-EPT-eval.pdf).

编辑 索书志