高能物理计算环境中KVM虚拟机的性能优化与应用

黄秋兰 李 莎 程耀东 陈 刚
(中国科学院高能物理研究所计算中心 北京 100049)

摘 要 高能物理是典型的高性能计算的应用,对CPU计算能力要求很高,并且CPU利用率的高低直接影响高能物理的计算效率。虚拟化技术在实现资源共享和资源高利用率方面表现出很大的优势。基于KVM(Kernel-based Virtual Machine)虚拟机进行性能测试和性能优化。首先对KVM虚拟机的处理器、磁盘IO和网络IO等参数进行测试,给出虚拟机和物理机的性能差异和定量分析,然后从KVM虚拟机架构上分析影响KVM性能的各种因素,从硬件级、内核级对影响性能的因素包括扩展页表EPT(Extented Page Table)和CPU的亲和性(CPU affinity)展开研究,以对KVM进行性能优化。优化结果表明,KVM的CPU性能的损失率可以降低至3%左右。最后,给出了高能物理计算的虚拟集群,结果显示虚拟机群的计算性能能够满足高能物理计算的需求。

关键词 高性能计算,KVM,CPU亲和性,扩展页表
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.015

Performance Optimization and Application of KVM in HEP Computing Environment

HUANG Qiu-lan LI Sha CHENG Yao-dong CHEN Gang
(Computing Center, Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China)

Abstract High Energy physics computing is a high-performance computing application, which highly requires computing power, and CPU utilization directly affects the computational efficiency of high-energy physics. Virtualization technology has demonstrated a great advantage in achieving high utilization of resources and resource sharing. This paper gave performance testing and optimization of KVM. Firstly, we showed testing results and quantitative analysis of performance between KVM and physical machines in aspect of CPU, disk IO and network IO. Then, we demonstrated various factors in hardware level and kernel level that affect the performance of KVM through the architecture of KVM virtual machine, including EPT (Extended Page Tables) and CPU affinity which are optimized for KVM. Experiment results show the penalty of CPU performance for KVM can be decreased to about 3%. Finally, virtual cluster of high energy computing was shown and the performance of virtual cluster can meet the needs of high energy physics computing.

Keywords HPC, KVM, CPU affinity, Extended page table

1 引言

我国高能物理领域的重大科学工程包括北京谱仪(BESIII)[8]、大亚湾反应堆中微子实验[9]、硬X射线调制望远镜(HXMT)、中国散裂中子源等。这些大规模高能物理实验的发展及新发现,离不开对海量数据的处理与分析。因此,如何高效率、高精度地分析海量数据是高能物理计算环境中面临的一个巨大挑战。高能物理计算是典型的高性能计算的应用,运行时需要大量的CPU计算资源。如果系统的CPU资源利用率不高,计算效率则大大下降。当前,高能物理计算环境主要通过Torque[1]、Condor[2]、LSF[3]等资源管理和作业调度系统基于系统负载状态和作业信息将作业调度到物理机器上运行。但是,这种资源管理是静态的,难以满足突发、批处理、CPU密集型、数据密集型等不同类型的作业对于不同的物理资源(内存、CPU、IO、网络、磁盘空间等)的需求。通常分配给某些作业队列的资源处于空闲的状态,而需要资源的作业却因为得不到资源无法被执行,导致资源难以充分利用。因此,十分有必要将高能物理计算环境移植到虚拟化平台上。使用KVM虚拟机等虚拟化技术[10]降低应用与基础设施的耦合程度,灵活调度各种类型的作业,实现不同应用需求对资源进行高度共享的目标,从而充分利用资源,提高资源利用率。

KVM(Kernel-based Virtual Machine)[4]开始是由Qumranet公司开发的基于x86硬件虚拟化的全虚拟化解决方案,是一款基于GPL授权方式的开源虚拟机软件。从Linux内核2.6.20版本开始,它以模块的形式成为内核的一部分。KVM是基于Intel VT技术和AMD SVM技术的硬件辅助虚拟化方案,并结合QEMU模拟器实现设备的虚拟化。

到稿日期:2013-12-26 返修日期:2014-03-15 本文受国家自然科学基金(11305192,11205179)资助。
黄秋兰(1982-),女,硕士,助理研究员,主要研究领域为云计算与海量数据存储,E-mail:huangql@ihep.ac.cn(通信作者);李 莎(1988-),女,博士生,主要研究领域为虚拟化技术;程耀东(1977-),男,博士,副研究员,主要研究领域为海量存储、网格计算与云计算;陈 刚(1961-),男,研究员,博士生导师,主要研究领域为高能物理计算。

KVM具有良好的性能，但是在IO密集型的应用下，其开销和性能与物理机上相比有很大的差距，KVM还需要进行进一步的改进和优化；另一方面，随着计算技术的快速发展和多核处理器[5]的流行，如何有效地利用多核环境来提高虚拟机的性能也存在一些挑战。在多核环境下，各核之间共享资源，如网络带宽、内存等资源。系统可以同时运行多个服务，比如一个核上运行Web服务，另一个核上运行数据库服务。如果把所有的服务都调度到一个核上，并不能有效利用系统资源。如果把不同的服务调度到不同的核上运行，各个核之间就会发生资源争用的情况。

本文重点是在高能物理计算环境下如何优化KVM的性能，及KVM在高能物理计算中的应用。首先对KVM虚拟机的处理器、磁盘IO和网络IO等参数进行测试，给出虚拟机与物理机的性能差异和定量分析，然后从KVM虚拟机架构上分析影响KVM性能的各种因素，从硬件级、内核级对影响性能的因素（包括扩展页表EPT[11]（Extented Page Table）、CPU的亲和性（CPU affinity）[6]）展开研究，以对KVM进行性能优化，优化结果显示，KVM的CPU性能损失率可以降低至3%左右。最后给出KVM虚拟机在高能物理计算中的应用实例。

2 KVM虚拟机性能测试

在虚拟化的应用中，虚拟机的性能好坏是实际应用的一个很重要的问题。评价虚拟机的性能主要包括CPU计算能力、磁盘IO和网络IO。对KVM虚拟机的性能测试，主要是通过各种基准测试程序（benchmark），比较KVM虚拟机和物理机的性能差异，并分析KVM的性能特征。

2.1 CPU性能测试

测试环境：Intel(R) Xeon(R) CPU X5650(2.67GHz)，8核，24GB内存，系统版本：SLC release 5.5 (Boron) 2.6.18-194.11.3.el5.cve20103081，64bit，KVM-83。

测试程序：基准测试程序HEPSPEC06[7]，测试值越大，表示CPU性能越好。HEPSPEC06为高能物理界开发，设计初衷是为了满足高能物理实验的需求，现在已经广泛应用到其他很多机构。

测试方法：在物理机上运行8个虚拟机，每个虚拟机分配1个CPU、2GB内存，使用基准测试程序同时在8个虚拟机上执行，得出每个CPU的计算能力的评估值。

图1给出了虚拟机上的CPU的评价值，横坐标为CPU计算能力的评价值，纵坐标为分布频率即密度，Entries表示CPU的评价值个数，Mean表示平均值。

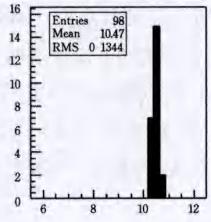

图1 KVM虚拟机上的CPU计算能力的测试值

从图可知，测试中共产生98个CPU的评价值，平均大小为10.47。物理机上运行基准测试程序后，单个CPU的评价值为11.77，可知KVM虚拟机在默认情况下，CPU计算能力损失率约为10%。

2.2 磁盘IO性能测试

测试环境：Intel(R) Xeon(R) CPU X5650(2.67GHz)，8核，24GB内存，系统版本：SLC release 5.5 (Boron) 2.6.18-194.11.3.el5.cve20103081，64bit，KVM-83。

测试程序：IOZONE -Mce -I -+r -r 256k -s 8g -f /usr/vice/cache/iozone_$i.dat$$ -i0 -i1 -i2。

测试方法：在物理机上运行8个虚拟机，每个虚拟机分配1个CPU、2GB内存，使用IOZONE同时在8个虚拟机上执行，得出每个虚拟机上的磁盘IO测试值，在物理机上同时运行8个IOZONE的进程，得出磁盘IO的测试值。

图2给出了写磁盘的性能测试结果。

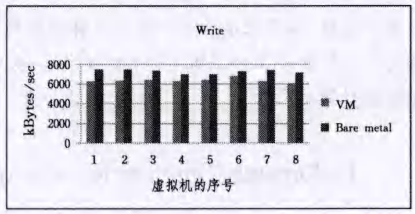

图2 KVM和物理机写磁盘的IO性能比较

图3给出了读磁盘的性能测试结果。

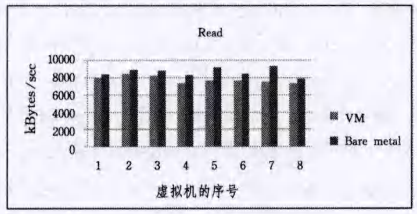

图3 KVM和物理机读磁盘的IO性能比较

从图2和图3可知，KVM虚拟机在磁盘IO上的性能损失比较大，损失率约为12%左右。

2.3 网络IO性能测试

测试环境：Intel(R) Xeon(R) CPU X5650(2.67GHz)，8核，24GB内存，系统版本：SLC release 5.5 (Boron) 2.6.18-194.11.3.el5.cve20103081，64bit，KVM-83。

测试程序：IPERF “-p 11522 -w 458742 -t 60”，TCP窗口值为256kB，测试时间持续60秒。

测试方法：在物理机上运行8个虚拟机，每个虚拟机分配1个CPU、2GB内存，在客户机上使用IPERF同时与8个虚拟机上建立连接，得出每个虚拟机的网络IO测试值，同时，在客户机上用IPERF启动8个并行的线程连接物理机，得出物理机的网络IO的测试值。图4给出了8个虚拟机和物理机上8个并行线程测试网络IO性能的结果。

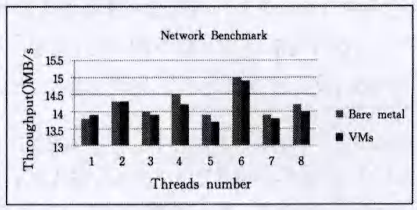

图4 虚拟机与物理机的网络IO性能比较

由图 4 可知，KVM 虚拟机的网络 IO 与物理机上的相差不多，损失率约为 3%左右。

3 性能优化与结果

基于第 2 节的测试结果，发现 KVM 虚拟机的 CPU 和磁盘 IO 性能的损失率在 10%左右，网络 IO 表现较好，损失率在 3%左右。在测试的过程中发现，现有的 KVM 虚拟机在处理器虚拟化、IO 虚拟化方面存在一些不适应虚拟化环境的因素，影响虚拟机的性能。

由 KVM 的虚拟化原理可知，虚拟机的处理器虚拟化利用 VT-x 技术的支持，KVM 中的每个虚拟机可具有多个虚拟处理器 VCPU，每个 VCPU 对应一个 Qemu 线程，VCPU 的创建、初始化、运行以及退出处理都在 Qemu 线程上下文中进行，需要 Kernel、User 和 Guest 3 种模式相互配合，实现比较复杂。而 IO 设备的虚拟化使用软件模拟的方式来实现，通过捕获虚拟机的任何 IO 请求，交给 Qemu 实现相应的 IO 操作，操作结果由 KVM 返回虚拟机中的硬件驱动并进行处理，虚拟机完成一次完整的 IO 操作，这种方法实现简单，但是由于 IO 处理流程中涉及多个环境，切换较多，其 IO 性能不是很理想。虽然 KVM 较新的版本中已经将一些关键设备的虚拟化进行了优化，但是主要的设备如磁盘和网卡虚拟化的性能开销仍旧较大。

下面将从 CPU 亲和性、扩展页表等影响 KVM 性能的因素进行改进和优化。

3.1 CPU 亲和性(Affinity)

在 KVM 模型中，每一个虚拟机都是由 Linux 调度程序管理的标准进程。CPU 亲和性，即是 CPU 的绑定设置，是指将进程绑定到特定的一个或多个 CPU 上来执行，而不允许调度到其他的 CPU 上。在多核环境下，Linux 内核对进程的调度算法也是遵守进程对处理器亲和性设置的。设置进程的处理器亲和性带来的好处是可以减少进程在多个 CPU 之间交换运行带来的缓存命中失效(Cache missing)，从该进程运行的角度来看，如果能够使单个进程始终运行在同一个或多个 CPU 上，减少进程在处理器间频繁迁移，会使 Cache 的命中率得到提高，从而提高进程的性能。

文中基于 CPU 的亲和性，采用进程绑定的方法来降低 KVM 进程在处理器间频繁调度，从而优化性能，具体要求：

(1)用户在创建虚拟机时，可以指定其在某个核上运行；

(2)虚拟机运行在指定的核上时，整个运行期间不会发生迁移。

如图 5 所示，在 KVM 虚拟机运行的过程中，通过 virsh 工具将虚拟机处理器与物理 CPU 绑定，例如某个虚拟机处理器 VCPU0 先和物理 CPU0 绑定，并在某个时刻解除绑定关系，下一个时刻可能会重新绑定到物理 CPU1 上。在任意给定时刻，VMCS 与物理 CPU 是一对一的绑定关系，即一个 VCPU 只能与一个物理 CPU 绑定。

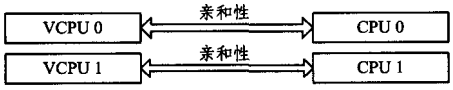

图 5 虚拟处理器与物理处理器的对应关系

3.2 扩展页表 EPT(Extented Page Table)

EPT 是 Intel 在 VT-x [9] 技术基础上增加的一种硬件辅助内存虚拟化技术。支持此类技术的处理器有两种工作模式：根模式和非根模式。VMM 工作在根模式，客户机工作在非根模式。EPT 仅在非根模式下有效，借助一套 EPT 页表结构将给定的客户物理地址转换成为机器物理地址，这种转换由硬件完成，所需的 EPT 页表结构由 VMM 创建、维护和更新。

3.3 优化后的实验结果

实验中，将客户机操作系统中的 VCPU 与物理 CPU 进行绑定，与未绑定的 CPU 性能和磁盘性能进行比较；将启用物理机的扩展页表选项与未启用扩展页表的 CPU 性能和磁盘性能进行比较。

3.3.1 CPU 计算性能

测试环境、测试程序与测试方法同第 2.1 节。

图 6 给出了优化前与优化后的 CPU 计算能力的实验结果。

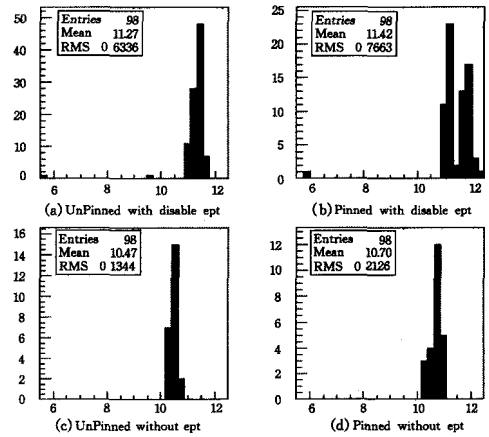

图 6 优化前与优化后的 CPU 计算能力比较

图 6 中，图(a)表示 VCPU 与物理 CPU 未绑定并且关闭扩展页表选项，图(b)表明 VCPU 与物理 CPU 绑定并且关闭扩展页表选项，图(c)表示 VCPU 与物理 CPU 未绑定并且与未启用扩展页表，图(d)表示 VCPU 与物理 CPU 绑定并且未启用扩展页表。

测试中共产生 98 个 CPU 的评价值，优化前 CPU 计算能力的平均大小为 10.47，物理机上运行基准测试程序的单个 CPU 的评价值为 11.77，可知 KVM 虚拟机在默认情况下，CPU 计算能力损失率约 10%。

从图 6(c)、(d)可知，客户机操作系统中的 VCPU 与物理 CPU 进行绑定后，CPU 计算能力的平均值提高到 10.7，并且从值的分布上看，由优化前的区间[10.2,10.9)往右偏移为[10.2,11.1)，说明在 VCPU 与 CPU 进行绑定后，KVM 虚拟机的计算能力提高了约 3%。

从图 6(b)看出，在将 VCPU 与物理 CPU 绑定并且关闭扩展页表选项的情况下，CPU 计算能力的测试值大部分都落在区间[10.9,12.5)，KVM 虚拟机的计算能力得到进一步提高。

3.3.2 磁盘 IO 性能

测试环境、测试程序与测试方法同第 2.2 节。

图 7 给出了优化前后磁盘的写性能与物理机的比较，图 8 给出了优化前后磁盘的读性能与物理机的比较。

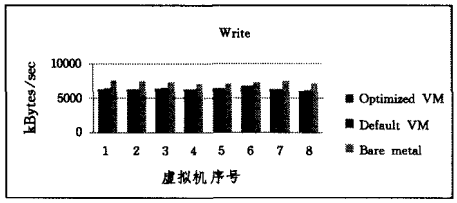

图 7　优化前后的磁盘写性能与物理机的比较

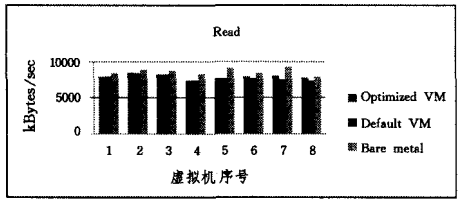

图 8　优化前后的磁盘读性能与物理机的比较

图 7 和图 8 显示的结果表明，绑定 VCPU 与物理 CPU、关闭扩展页表选项，对磁盘 IO 性能的影响不大。

实验证明，在客户机操作系统的 VCPU 与物理 CPU 绑定、关闭扩展页表选项时，KVM 虚拟机的 CPU 性能最好，损失率约为 3%，CPU 计算能力与优化前相比，提高了 6%～8%，而磁盘 IO 性能的提升不大，这与 KVM 虚拟机的 IO 虚拟化有关，需要进一步从 KVM 客户机操作系统源码分析与优化。KVM 虚拟机的 CPU 性能还比较乐观，尤其在处理 CPU 密集型的应用方面，在资源共享和资源利用率提高的前提下，3%的损失率可以不予考虑。因此，可以得出结论：优化后的 KVM 虚拟机比较适合 CPU 密集型和网络 IO 密集型的应用。

4　KVM 虚拟机在高能物理中的应用

高能物理计算需要巨大的 CPU 计算能力，随着集群规模的不断扩大，操作系统与应用软件的不断升级，CPU 等硬件性能的持续提升，传统的集群系统面临着资源利用率不高、应用迁移复杂、多应用支持困难等问题，而虚拟集群技术是解决这些问题的有效手段。虚拟集群基于虚拟机技术，在物理机器上虚拟出不同的操作系统以适应不同种类和不同版本的应用软件，同时在用户看来，这仍然是传统集群的使用方法，用户也不必改变以前的使用习惯。目前，高能物理研究所的虚拟集群核心技术是采用 KVM 虚拟机屏蔽底层基础设施的异构性，动态调整集群的资源分配，大大提高了系统资源的利用率和作业运行效率；同时对外仍然是传统的作业提交方式，用户感觉不到虚拟集群系统与原有系统的不同。虚拟集群的结构图如图 9 所示。

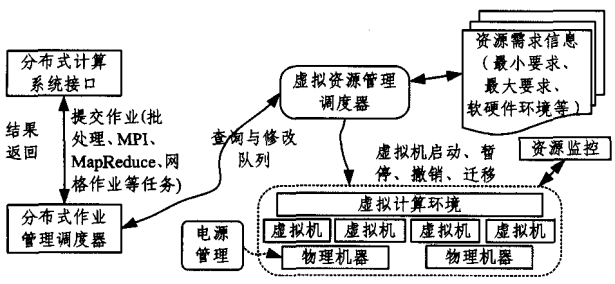

图 9　虚拟集群结构图

高能物理的实际作业运行在虚拟集群上，即用户将物理作业提交到虚拟机上进行计算。高能物理计算分为蒙特卡罗模拟计算、重建计算、分析计算 3 种，主要表现为 CPU 密集型和 IO 密集型，CPU 和 IO 成为数据处理的瓶颈。蒙特卡罗模拟计算主要消耗 CPU，对 IO 要求不高；重建计算从海量存储系统中读取原始数据，然后把重建计算后的数据写入海量存储数据；分析计算需要读取重建后的物理数据，做分析和处理，结果写回用户目录，对 CPU 和网络 IO 的要求都很高。从物理作业的实验结果可知，高能物理模拟计算在 KVM 虚拟机上的性能损失约为 2.9%，高能物理的分析作业在 KVM 虚拟机上的性能损失约为 3%。这一实验结果正好验证了上述优化后的 KVM 虚拟机的性能。在资源高度共享和高利用率的前提下，3%左右的性能损失丝毫不会影响用户体验。

结束语　虚拟化技术在数据中心和云计算领域正发挥着越来越重要的作用，但是虚拟机在 CPU、IO 性能方面还存在一些问题，需要进一步优化。本文通过各种基准工具对 KVM 的 CPU 处理器、磁盘 IO 和网络 IO 进行测试，对虚拟机和物理机的性能差异有了初步的认识，接着从 KVM 虚拟机的虚拟化机制分析影响虚拟机性能的各种因素，从硬件级、内核级对影响性能的因素展开研究，包括扩展页表 EPT(Extented Page Table)和 CPU 的亲和性(CPU affinity)对 KVM 进行性能优化。实验结果表明，KVM 虚拟机的 CPU 和磁盘 IO 性能的损失率在 10%左右，网络 IO 表现较好，损失率在 3%左右。经过优化后，KVM 的 CPU 性能的损失率降低至 3%左右。因此，可以得出结论：优化后的 KVM 虚拟机比较适合 CPU 密集型和网络 IO 密集型的应用，并且 KVM 虚拟机在高能物理环境中的虚拟集群中表现出很好的性能，能够满足高能物理计算的需求。

参 考 文 献

[1] Zhang Yang, Chen Wen-bo, Li Lian, et al. Analysis and Implementation of TORQUE: A High-Performance Cluster Job Management System[J]. Computer Engineering & Science, 2007, 10:42

[2] Frey J, Tannenbaum T, Livny M, et al. Condor-G: A Computation Management Agent for Multi-Institutional Grids[J]. Cluster Computing, 2002, 5(3): 237-246

[3] Etsion Y, Tsafrir D. A Short Survey of Commercial Cluster Batch Schedulers [J]. School of Computer Science and Engineering, The Hebrew University of Jerusalem, 2005(13): 1-4

[4] Kamay K Y, Laor D, Lublin U. Kvm the Linux virtual machine monitor[C]// Proceedings of the Linux. 2007: 225-230

[5] Gepner P, Kowalik M F. Multi-core Processors: New Way to Achieve High System Performance[C]// Proceeding of the International Symposium on Parallel Computing in Electrical Engineering. 2006: 9-13

[6] Guo Zhao-liang, Hao Qin-fen. Optimization of KVM Network Based on CPU Affinity on Multi-cores[J]. International Conference of Information Technology, Computer Engineering and Management Sciences, 2011, 4: 347-351

[7] HEPSPEC06 benchmark [OL]. http://w3.hepix.org/benchmar-ks/doku.php/

[8] 陈和生，张闯，李卫国. 北京正负电子对撞机重大改造工程[J]. 工程研究-跨学科视野中的工程，2009，1(3)：275-281

[9] 郝慧峰. 大亚湾反应堆中微子试验中基于 VME 的 RPC 电子学研制[D]. 合肥：中国科技大学，2012

[10] 张朝鹏. 基于 Intel-VT 处理器的虚拟机内存虚拟化的实现和优化[J]. 应用科技，2009，123：143-144

[11] 李勇，郭玉东，王晓睿，等. 基于 EPT 的内存虚拟化研究与实现[J]. 计算机工程与设计，2010，31(18)：4101-4104