基于KVM虚拟化技术的Hadoop架构

夏上云[1]，王旻超[1]，张惠然[1]，戴东波[1]，谢 江[1,2]，李 青[1,2]，张 武[1,2]

(1 上海大学 计算机工程与科学学院，上海 200072；2 上海大学 高性能计算中心，上海 200072)

摘 要：本文提出并实现了一种Hadoop与虚拟化技术相结合的模型.该模型将多核计算机虚拟成多节点集群，最大限度地利用计算资源.在实验测试的基础上，通过分析任务在Hadoop环境中的并行机制，弄清计算任务在节点中与核中的分配方式，以达到提高并行效率、降低硬件开销的目的，最后通过对典型应用问题进行计算，从而对本文所提出模型的正确性和有效性进行了验证.

关键词：大规模数据；Hadoop；虚拟化；并行计算

中图分类号：TP301.6 **文献标识码**：A **文章编号**：1000－7180(2013)03－0063－04

Hadoop Architecture Based－on Virtual Technology of KVM

XIA Shang-yun[1], WANG Min-chao[1], ZHANG Hui-ran[1],
DAI Dong-bo[1], XIE Jiang[1,2], LI Qing[1,2], ZHANG Wu[1,2]

(1 School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China;
2 High Performance Computing Center, Shanghai University, Shanghai 200072, China)

Abstract: We propose a model which combines the Hadoop with the virtualization technology. In the model, each processing core of multi－cores computer is virtualized as a computing node, then the use of computing resources is maximized. Moreover, parallel working mechanism of the Hadoop is discussed. Experiments are fulfilled by typical applications finally.

Key words: large－scale data; Hadoop; KVM; parallel computing

1 引言

近年来，大规模数据的重要性越来越被人们所重视，因为其普遍存在于现代科学技术发展的各个领域，并且伴随着数据量的迅猛增长，使得计算对象的规模越来越大.由于单个计算机的存储空间及运算能力有限而无法满足大规模数据处理的需求，因此基于多台服务器的并行计算就是一个可靠的选择.作为一项起支撑作用的技术，它可以满足实际工作中涉及的大规模计算的需求[1－2].为了提高大规模数据处理的精度和减少运算时间，可以借助并行计算技术来寻求一种更快速、更容易、更廉价的方式用以获取可利用的数据以及存储数据的方法.

通常认为，高性能计算机的内存结构可以分为共享存储的内存结构和分布式存储的内存结构两大类[3].现在以分布式存储为代表的层次并行体系结构的高性能计算机发展迅速.在过去的几年间，MPI和OpenMP的编程模型已经成为并行计算的主流选择[4－6].由于使用MPI编程的消息传递模型时，细粒度的并行会引发大的通信量，造成了动态负载平衡相对困难[7].而且，OpenMP只能在共享存储的机器上运行，因此对于大规模密集型数据而言，采用分布式存储的并行集群将会使计算效率得到有效的提高.当前Hadoop架构成为高性能计算机发展的主流，它采用分布式存储系统和并行执行机制，将大规模数据分布在集群的各节点中.目前多核计算机

收稿日期：2012－06－29；**修回日期**：2012－07－22

基金项目：国家教育部博士点基金项目(20113108120022)；上海市科学技术委员会重点项目(11510500300)；上海市教委创新基金(11YZ03)；上海市重点学科建设基金项目(J50103)

被广泛应用,使用多核计算机组成 Hadoop 集群时,分发给各个数据节点的多个任务会产生资源竞争,像 CPU、内存、输入输出带宽等,会使得暂时不用的资源处于等待状态,这样能够引起一些资源的浪费和响应时间的延长,资源开销相应有所增长,同时对系统性能产生消极的影响.

针对上述问题我们提出了一种基于 Hadoop 结合虚拟化技术的模型,此模型利用 Kernel－based Virtual Machine(KVM)虚拟技术按照每台计算机 CPU 的个数将核模拟成计算机,即把一台多核计算机模拟成多台计算机,把一定的计算资源合理分配,允许一个平台同时运行多个操作系统,这样做的好处是可以提高计算效率. 原因在于,在该模式下不同的应用程序工作时处于相互独立的状态,即在独立的空间中运行,从而避免了彼此受到影响.

本文首先讨论了 Hadoop 结合虚拟化技术的框架机制,接着给出了三台双核计算机集群虚拟成六台计算机集群的方法;然后通过实验比较虚拟前与虚拟后的计算性能;最后分析实验并给出结论.

2 Hadoop 结合虚拟机运行机制

Hadoop 由许多元素构成,作为一个软件框架,其最底部是 Hadoop Distributed File System (HDFS),它存储 Hadoop 集群中所有存储节点上的文件[8－9],能够处理大量数据. HDFS 的上一层是 MapReduce 引擎,可对数据进行分布式处理,该引擎由 JobTrackers 和 TaskTrackers 组成[10]. 最为常见的 Hadoop 模型是以计算机为节点将多台计算机部署为一个多节点集群,将大数据集分割成小数据集并分发给各个从节点[11].

为了充分利用硬件资源,本文把 KVM 虚拟化技术引入 Hadoop 架构中. KVM 是一个开源的系统虚拟化模块,需要硬件的支持,是基于硬件的完全虚拟化[12－13]. KVM 由两部分组成,一部分是 KVM-Driver[14],已经成为 Linux 内核的一个模块,负责虚拟机的创建,虚拟机内存的分配、虚拟 CPU 寄存器的读写以及 CPU 的运行等;另一部分是稍微修改过的 Qemu[15],用于模拟 PC 硬件的用户空间组件、提供 I/O 设备模型以及访问外设的途径. CPU 的虚拟化技术可以单 CPU 模拟多 CPU 并行,本文利用 KVM 虚拟化技术的这一特性将 Hadoop 集群的双核计算机中的 CPU 模拟成一台计算机,从而使得硬件资源被充分利用并在一定的硬件设备上扩大集群规模. 图 1 给出了 Hadoop 结合虚拟化技术的简要模型. 此模型是基于 Hadoop 思想的扩充.

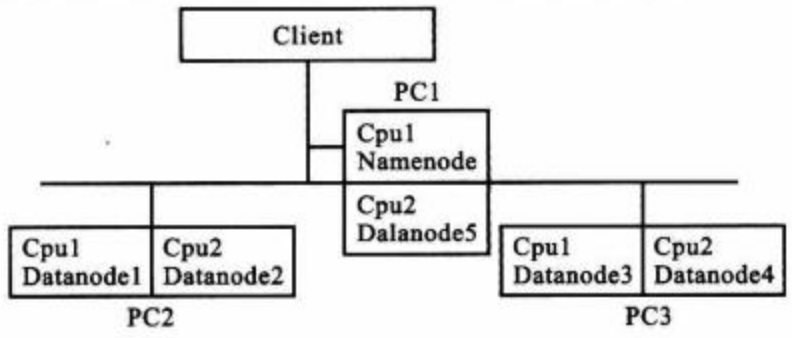

图 1 虚拟后的简要模型

3 性能测试

3.1 实验环境

根据我们使用的数据,三台双核计算机的资源已经满足计算量. 所以本次实验使用三台相同配置的双核计算机作为实验设备构建了一个三节点集群,并使用两种实验环境作比较来说明问题:实验环境一,以其中一个节点作为主节点,其他两个作为从节点,不使用 KVM 虚拟化模块技术. 实验环境二,在三台计算机集群基础之上使用 KVM 虚拟化模块技术把六核虚拟为六个节点,其中一个节点作为主节点,其他五个节点作为从节点. 两实验环境中的三台计算机使用相同配置的 PC,软硬件配置具体如表 1 所示.

表 1 软硬件配置

硬件配置	处理器:Pentium(R) Dual－Core CPU E5800 @3.20GHz 内存:6GB
软件配置	Ubuntu 10.04;JDK 版本:1.6; Hadoop 版本:0.20.2

3.2 实验内容

所做实验分为两组. 第一组求证在 Hadoop 模型中计算机任务在集群中的分配机制. 使用 Hadoop 自带的蒙特卡洛求 PI 的程序作为测试程序,使用实验环境一. 设置计算任务为 10 个,计算量在 10 的 8 次方量级,计算任务运行三遍,记录在系统管理器中每个核的使用情况.

第二组实验是将三台双核计算机组成的集群利用虚拟技术变为六个节点组成的集群,即每个核虚拟为一个节点. 这么做可以在很大程度上提高计算机资源的利用率. 本组实验设置 Hadoop 的块大小为 512KB,冗余备份设置为 3,使用的程序是字符统计程序(wordcount.jar)和矩阵相乘算法程序(MatrixMulti.jar).

4 结果分析

程序中设置每个节点一次分 2 个任务. 第一组

中的第一次实验主节点把 10 个任务分别分配给 3、4、5、6、7 这五个 CPU，第二次实验主节点把任务分配给 2、3、4、6、7 这五个 CPU，第三次实验主节点把任务分配给 2、3、6、7、8 这五个 CPU. 其结果如图 2 所示. 结果表明 Hadoop 的任务分配机制是随机的、不固定的，每个节点得到的任务量是由程序设定的.

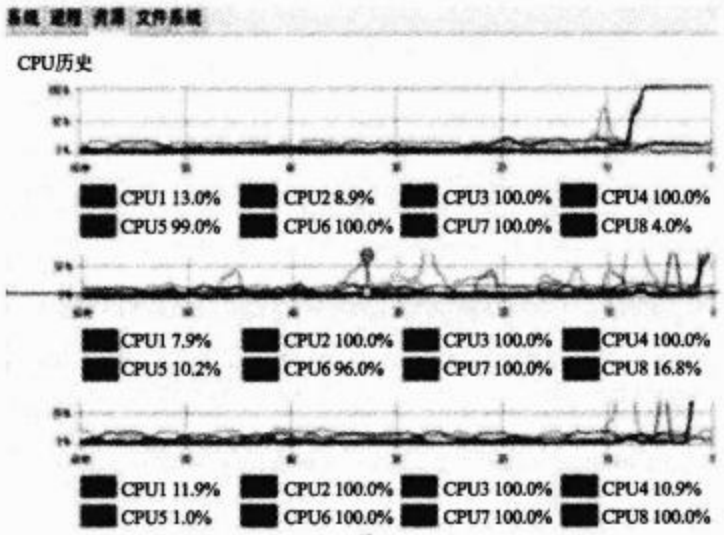

图 2　第一组实验的 CPU 占用情况

第二组使用字符统计程序和矩阵相乘算法程序在上述两种实验环境中运行. 对于字符统计程序我们使用了 4 组数据，其数据大小分别为 1GB、2GB、4GB 和 6GB. 实验结果如图 3 所示.

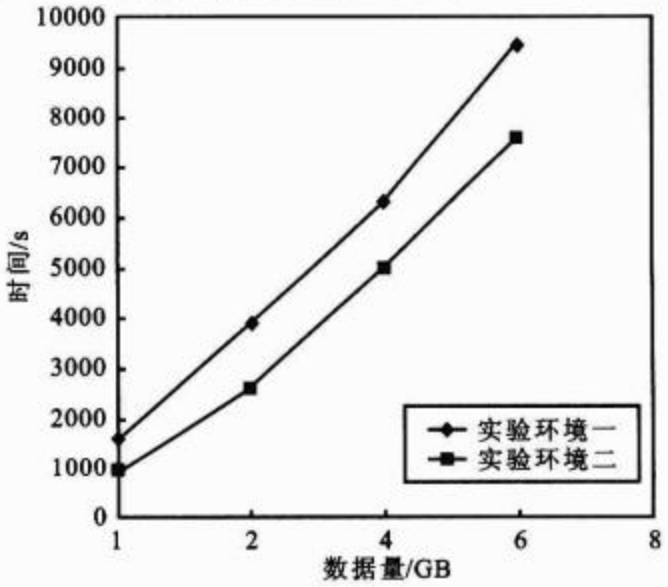

图 3　字符统计程序在两种环境下运行的性能对比

在上述两个实验环境中使用相同内容相同大小的数据，得出的结果表明实验环境二的性能明显好于实验环境一的性能.

对于矩阵相乘算法程序我们使用了七组数据，分别为 3×3、4×4、5×5、6×6、7×7、8×8 和 9×9 矩阵. 实验结果如图 4 所示.

同样的，在上述两个实验环境中使用相同的矩阵，从图 4 可以看出虚拟的六节点集群所用时间少于三节点集群而且计算时间在直线上下波动. 结果表明虚拟集群的性能优于没有虚拟的集群而且虚拟的集群相对于没有虚拟的集群而言计算性能呈线性增长.

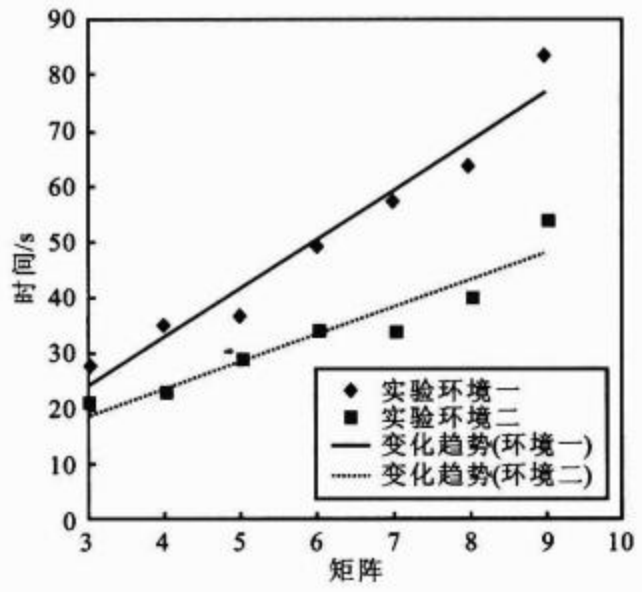

图 4　矩阵相乘算法程序在两个实验环境下的性能对比

实验过程中，三节点集群的 CPU 利用率大约在 60%～70%左右，由于分配给各节点的任务相互竞争资源，一些暂时不用的资源处于等待状态而闲置不用造成资源浪费，从而 CPU 达不到很高的利用率. 三节点集群经过虚拟化之后形成六节点集群资源得到合理分配，减少了资源空闲浪费. CPU 利用率几乎是 100%. 另一方面 Blocksize 变量对不同大小的数据文件计算性能也有影响. 当数据少而 Blocksize 设置超过总数据的大小时，使得较多的 CPU 产生空闲，不利于计算. 当数据多时，Blocksize 设置超过总数据或超出节点计算能力时，也不利于计算. 只有为处理不同数据文件大小的程序设置合适的 Blocksize 变量才能够提升计算性能[16]. 对高并行性密集型数据而言，充分利用计算资源是很有必要的. 虚拟化技术可以把资源细分，减少由于作业竞争资源造成的浪费[17]. Hadoop 采用的是分布式数据文件系统，每个数据节点按就近原则处理数据，减少通信消耗. 通信并行性程序则要考虑虚拟化带来的通信消耗. 虚拟的集群跟未虚拟的集群调度任务的策略是不一样的，因为在实验环境一中每隔 2s 为各个节点分配一轮任务，而实验环境二则是连续为各个节点分配任务.

5　结束语

本文针对大规模数据计算消耗时间长的问题，使用了虚拟化技术结合 Hadoop 模型进行了优化. 并且通过实验进行了验证. 其结果表明，虚拟化技术与 Hadoop 相结合的模型性能更高，更能充分利用硬件资源. 从而为降低硬件设备的成本提供了理论依据. 我们将会做进一步的工作，通过增大数据量把 Hadoop 结合虚拟化模型扩展到四核、八核计算机上.

参考文献：

[1] Taylor R C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics [C]//BMC Bioinformatics. Washington: BMC Bioinformatics, 2010.

[2] Gunarathne T, Wu T L, Qiu J, et al. Cloud computing paradigms for pleasingly parallel biomedical applications[C]// HPDC '10 Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. Chicago: ACM, 2010.

[3] Nakajima Kengo. Parallel multigrid solvers using OpenMP/MPI hybrid programming models on multi－Core/Multi－socket clusters[C] // VECPAR 2010. Berkeley, CA, United States, 2011: 185－199.

[4] Dean Jeffrey, Ghemawat Sanjay. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008(51):107－113.

[5] 熊盛武,王鲁,杨婕. 构建工性能集群计算机系统的关键技术[J]. 微计算机信息, 2006, 22(1/3): 86－88.

[6] 陈靖,张云泉,张林波,等. 一种新的 MPI Allgather 算法及其在万亿次集群系统上的实现与性能分析[J]. 计算机学报, 2006, 29(5): 808－814.

[7] 郭本俊,王鹏,陈高云,等. 基于 MPI 的云计算模型[J]. 计算机工程, 2009, 35(24): 84－85.

[8] Borthakur Dhruba. The hadoop distributed file system: architecture and design[EB/OL]. [2008－05－29]. document on Hadoop Wiki.

[9] 周轶男,王宇. Hadoop 文件系统性能分析[J]. 电子技术研发, 2011, 38(5): 15－16.

[10] Hu Weisong, Tian Chao, Liu Xiaowei, et al. Mutiple－job optimization in MapReduce for heterogeneous workloads[C]// Sixth International Conference on Semantics. China, Beijing: Knowledge and Grids, 2010.

[11] 刘鹏. 实战 Hadoo[M]. 北京:电子工业出版社, 2011.

[12] Fraser Keir, Hand Steven, Neugebauer Rolf, et al. Safe hardware access with the xen virtual machine monitor[C]// Proceedings of the OASIS ASPLOS Workshop. UK, 2004.

[13] Borden T, Hennessy J, Rymarczyk J. Multiple operating systems on one processor complex[J]. IBM Systems Journal, 1989, 28(1): 104－123.

[14] Deshane T, Shepherd Z, Matthews NJ, et al. Quantitative comparison of Xen and KVM[C]// Xen Summit. Boston, USA, 2008.

[15] Bartholomew D. Qemu: a multi host, multitarget mulator[J]. Linux Journal, 2006(145):3.

[16] Tom White. Hadoop: 权威指南[M]. 北京:清华大学出版社, 2011.

[17] Qin An, Tu Dandan, Shu Chengchun, et al. XConveryer: guarantee hadoop through－put via lightweight OS－level virtualization [C]//Eighth International Conference on Grid and Cooperatie Computing. China: Lanzhou, 2009.

作者简介：

夏上云 女,(1990－),硕士研究生. 研究方向为生物信息学.

王旻超 男,(1988－),硕士研究生. 研究方向为生物信息学

张惠然 男,(1981－),博士,讲师. 研究方向为生物信息学

戴东波 男,(1977－),博士,讲师. 研究方向为生物信息学

谢 江 女,(1971－),博士,高级工程师,研究生导师. 研究方向为生物信息学

李 青 男,(1962－),博士,教授,博士生导师. 研究方向为高性能计算与应用

张 武 男,(1957－),博士,教授,博士生导师. 研究方向为高性能计算与应用、生物信息学