虚拟化高性能计算集群上 MPI 应用性能分析*

陆华俊, 任开军, 宋君强, 刘少伟

(国防科学技术大学 计算机学院,长沙 410073)

MPI applications performance analysis in HPC virtual clusters *

LU Hua-Jun, REN Kai-Jun, SONG Jun-Qiang, LIU Shao-Wei

(School of Computer, National University of Defense Technology, Changsha 410073, China)

+ Corresponding author: E-mail: luhuajun@nudt.edu.cn

Abstract: Due to the advantages of customizing operating environment, resource utilization and fault tolerance, virtualization are considered to have the potential solving the problem that high-performance computing(HPC) facing to. However, the major problem of applying virtualization to HPC is the additional performance overhead. In order to verify the feasibility of virtualization using in HPC application, we build the virtualization environment in HPC cluster, and carry out a series comparative examination of issues relevant to MPI-style applications. Our results show that, NPB LU and SP application runs longer time in virtual cluster. IS, MG, CG and BT in 16 nodes runs faster by 40%-50% per node than in 32(36)nodes, while FT runs faster by 27%. Excessive communication overhead in applications will reduce the efficiency of the virtual cluster.

Key words: virtualization; high performance computing; Xen; scheduling policy

摘 要: HPC 上搭建虚拟集群可以较好地实现资源复用且可定制，同时保证各个用户需求并不会产生资源冲突。被认为可以解决高性能计算面临的诸多问题的潜质，同时其主要面临的问题是额外的性能开销。为了验证虚拟化在高性能计算中应用的可行性，本文在高性能计算集群中构建了虚拟化系统环境，并针对 MPI 类应用进行了系列对比测试。实验结果显示，NPB 测试中，LU 和 SP 在虚拟集群上运行时间更长，IS、MG、CG、BT 在 16 节点的集群单节点运行速度比 32(36)节点大约快 40%-50%，FT 在 16 节点的集群单节点运行速度比 32 节点大约快 27%，任务具有较大通信需求时，会降低虚拟集群的执行效率。

关键词: 虚拟化;高性能计算;Xen; 调度策略

中图法分类号: TP393 **文献标识码**: A

* Supported by the National Natural Science Foundation of China under Grant No.60903042 (国家自然科学基金)

作者简介: **陆华俊**(1989－),男,广东湛江人,硕士研究生,主要研究领域为虚拟化、云计算和大数据处理;任开军(1975－),男,副研究员,主要研究领域为高性能计算、云计算、科学工作流、Web 服务合成;宋君强(1962－)，研究员，博士生导师，主要研究领域为数值天气预报、大型并行应用软件、CPU/GPU 异构混合计算.刘少伟(1987－),男,博士研究生,主要研究领域为科学工作流,云计算和大数据处理.

1 虚拟化高性能计算集群概述

近年来，高性能计算系统的规模日益扩大，节点规模可达几十万甚至百万、千万，存储容量跃升到 PB、ZB 级别。最新 Top 500[1]中，排名第一的 TH-2 超级计算机，核心包含 2120000 个可运行内核，硬盘容量 12.4PB，内存大小 1.4PB。然而，高性能计算系统在性能提升的同时，也带来一系列管理和使用的问题。大规模科学计算软件部署在不同的 HPC 系统中，通常需要花费大量人力及时间进行优化和移植，而当前在 HPC 上进行开发时，研究人员往往缺乏方便快捷的系统软件调试环境。另外，高性能计算系统虽然峰值性能很高，但实际性能往往只有峰值性能的 0.5%-10%，资源利用率亟待提升。如何在满足大规模的科学工程应用基础上，尽可能地提高系统系能，充分利用硬件资源，将成为当前和今后一段时间的研究热点。

虚拟化因其运行环境可定制，资源复用，容错性高等特点，引起科学计算领域研究者的极大兴趣。当前，高性能计算中的虚拟化技术应用仍处在起步阶段，将系统虚拟化技术引入到高性能计算当中，仍需要解决许多关键技术问题。减少性能开销是研究人员关注的重点，其包括如何构建系统管理程序 Hypervisor、对客户操作系统(Guest OS)进行定制和部署，以及优化网络带宽和控制通信。在 VMM 设计上，Xen、KVM 均可采用半虚拟化方式配合修改客户端操作系统来减少性能损失；在硬件上，Intel 和 AMD 等处理器均对虚拟化进行 VT 支持，提高 CPU，内存等方面的性能；在操作系统上，研究人员根据高性能应用的不同类型进行定制以最大限度提升性能。然而，对于部署虚拟化技术的高性能计算集群和传统集群的差异，目前还没有一个详实的数据进行验证分析。此外，和虚拟化在云计算中应用不同的是，超级计算机多核、CPU/GPU 异构等特点也是虚拟化在应用于 HPC 中需要考虑的问题。

本文通过在天河高性能计算集群子系统上部署虚拟计算平台和远程并行开发环境，针对不同的性能指标进行了系列测试，对测试集和实际应用产生的实验结果进行量化讨论和进一步分析。本文贡献主要表现在以下几个方面：首次在大规模高性能计算集群上搭建虚拟化环境并进行相应测试。和当前其它虚拟集群测试不同，本实验将直接在高性能计算集群上搭建虚拟集群环境，与先前小规模的集群实验相比，更具有说服力和代表性。在测试用例选取上，根据实际高性能应用需求，选取特定标准测试集进行定量分析和讨论。本文通过对比实验，为后续对虚拟化是否能在高性能计算上进一步得到应用提供详实的理论依据和定量的数据分析。

2 相关工作

2.1 虚拟化在HPC集群中应用

目前虚拟化技术应用到高性能计算中已经有一些成功案例。Cray 公司提出了 adaptive supercomputing[2]的设想，其观点就是：系统要适应用户的代码，也就是在同一个高性能计算平台上系统能针对提交的应用确定哪一种处理技术最适应处理该代码，再进行相应的编译，虚拟化技术是该系统整合方案的核心所在。Huang 和 Abali 等人提出了一种虚拟化技术应用到高性能技术的框架[3]，其通过采用 VMM-bypass I/O 旁路技术访问 InfiniBand 高速互联网络，使得虚拟化系统获得了较高的通信性能，该文献还提出了可扩展 VM 映像管理技术以减少 VM 分布和管理的开销。2008 年 IBM 推出的“蓝云（Blue Cloud）”[4]通过架构一个分布的、可全球访问的资源结构，使数据中心可以在类似互联网的环境下运行大规模科学计算。 Amazon 在 2010 年 7 月推出了面向高性能的 Cluster Compute Instances 或 CCIs[5]，美国劳伦斯国家实验室的研究者在测试一个由 7040 个处理器核心构成的集群后得到的结果表明，这一集群在当时的 TOP500 中排名第 231 位，这也标志着云计算从此进入了 TOP500 的行列。2011 年 Amazon 测试了一个由 1064 个 CCIs 构成的集群，获得了 240.09 TeraFLOPS Linpack 性能，从而使得这一集群在 2011 年 11 月 TOP500 排名中位列第 42 位。华中科技大学的金海，钟阿林等人[6]深入调研了 VCPU 调度策略和机制，指出了当前多核处理器同步机制，共享 Cache，不对称多核结构是影响虚拟机在高性能应用面临的主要问题和挑战。

2.2 HPC虚拟集群测试的相关工作

从 08 年开始，研究人员对虚拟化在高性能计算中应用进行了一系列的测试工作，早期的基于科学工作流

的测试结果表明，通过提供合适的资源，应用性能不会得到明显下降[7-10]，但科学工作流并非典型的高性能应用。研究人员开始针对紧耦合 MPI 和 OPENMP 应用[11-16]以及实际应用[11, 15, 17]进行相应的测试。这些测试实验环境和侧重点不同，存在着以下几个共同的特点和不足：（1）测试规模偏小，大部分实验采用的是 NPB 和 LINPACK 进行测试，NPB 采用的是 A，B 类小规模测试，（2）其次，对比实验多采用云计算设施（Amazon EC2 或者 CCI）及本地构建的小集群，构建相同硬件环境进行对比实验，在操作系统和集群管理软件选取上存在着差异，导致实验存在误差；（3）针对虚拟机 I/O 方面测试较少，对虚拟化在 I/O 方面影响缺乏深入的分析和讨论。

清华大学的 Yan Zhai 等人[18]利用 NAS NPB[19]、Intel MPI[20]、IOR[21]、以及三个实际的应用对 Amazon EC2 高性能计算云基础设施进行了全面的测试，并将结果与一个本地使用 InfiniBand 互连的高性能集群进行了对比。测试结果表明，尽管 Amazon 使用 10GB 以太网仍然是制约性能提升的瓶颈，但是并行应用在 CCIs 集群上已经能够达到本地集群的性能水平，例如 NPB 最坏情况下（NPB LU）的性能损失为 50%， 而 EP、MG、SP、与 CG 则已经达到了本地集群相当的性能。然而他们的实验是使用两个完全不同的平台搭建进行的，虽然硬件性能基本一致，但不同的搭建流程，使用不同的管理软件，都可能给对比实验带来误差。

3 系统模型与性能评测指标

虚拟化对于高性能计算的主要缺陷是性能上的额外开销，为了定量分析虚拟化给高性能计算集群带来的影响，我们在部分高性能计算集群节点上重新搭建了虚拟化环境。

3.1 系统模型与框架

系统采用统一的用户认证机制，并搭建了 Eclipse PTP 远程开发环境，用户通过 Eclipse 远程桌面查看节点和作业信息，并通过 PTP-proxy 提交作业到集群管理软件 SLURM 中。为了测试虚拟化的性能差异，同时尽可能地减少环境搭建和实验误差，如图 1，我们将计算节点分为两部分，其中一部分部署了虚拟化环境，每个物理节点部署一个虚拟机，通过 Hypervisor VMM 管理 CPU 资源，虚拟化后的节点，通过配置高速互联网络 InfiniBand 进行网络通信。另外一部分不做改动，仍然为传统的高性能集群。为了减少磁盘访问带来的实验误差，采用了一致的磁盘文件管理系统，所有的虚拟节点和物理节点均可访问同样的文件系统。

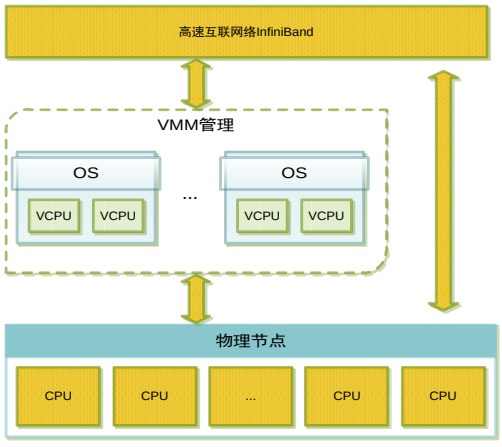

图 1 虚拟集群框架

3.2 性能评测指标

性能测量指标的选取对虚拟高性能计算系统的评测结果有很大影响，虚拟计算集群性能的测试主要考察虚拟机管理器、网络通信、服务器整合中虚拟机性能和共存虚拟机影响下的性能等几个方面。另外，高性能计算系统的使用因用户和作业的不同而产生不同的计算行为和目的，对 CPU、网络带宽等集群资源也有不同的要求。 影响虚拟化在高性能计算系统中应用的最主要的因素就是性能上的额外开销，主要包括即网络方

面的通信开销和 CPU 调度开销。而虚拟集群中单个物理节点上的虚拟机网卡是分时复用模式，当节点内和节点间进行大量的通信时，网络带宽的资源瓶颈将制约整个科学应用的性能和效率。

4 性能对比实验和数据分析

4.1 实验环境建立

实验集群包含 512 个物理节点，每个物理节点上包含 8 核 Intel Xeon E7520 CPU，存储采用 Lustre 文件系统，节点间通过 InfiniteBand 高速互联网络进行网络通信，每个物理节点启动一个虚拟机。虚拟化软件采用 Xen-4.1.2，采用全虚拟化模式。操作系统为 Ubuntu 12.04，编译环境为 Intel Composer_xe_2013.3.163 组件，默认编译优化选项为-O 3 级，为了使得数据区申请空间大于 2G，添加编译选项-mcmodel=medium。测试用程序采用 NAS-NPB 测试集进行 C 类测试。每个测试实验进行一百次，取平均值作为结果。

4.2 NPB结果分析

NPB 版本为 3.3，选用 C 类测试。分别在在 16 个物理节点、16 个虚拟节点、32 个物理节点和 32 个虚拟节点上进行 NPB C 类测试，由于 SP 和 BT 只能在平方数个节点上运行，因此，SP 和 BT 使用 36 节点，其它测试集则使用 32 个节点。

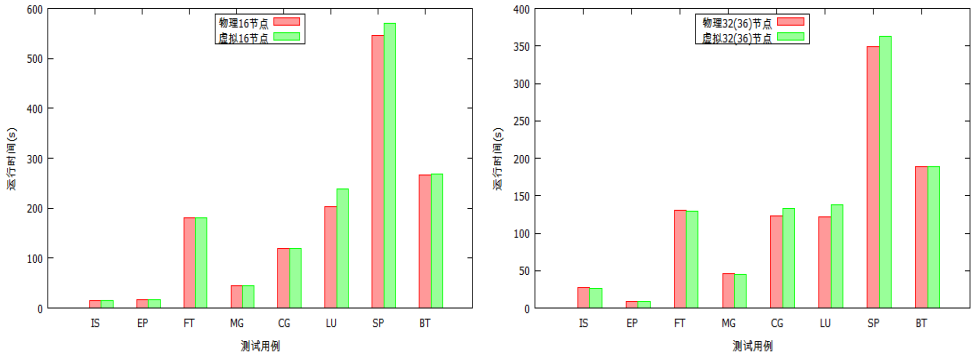

图 2 天河 1-A 集群物理节点 NPB C 类规模测试结果

由图 2 可以看出，16 节点的 LU 和 SP 在虚拟集群上的运行时间稍长，其他测试集运行时间大致相同。而在 32 (36) 节点中，CG、LU 和 SP 在虚拟集群上的运行时间稍长，其他测试集运行时间大致相同。

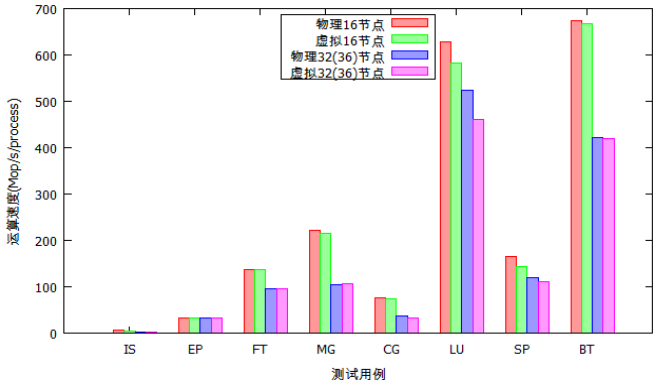

图 3 天河 1-A 集群物理节点 NPB C 类对比测试结果

从上图可以看出，EP 在四个集群中运算速度相差无几；IS、FT，MG，CG、BT 在节点数量相同时，虚拟集群和实际集群单节点运行速度相当；IS、MG、CG、BT 在 16 节点的集群单节点运行速度比 32(36)节点大约快 40%-50%，FT 在 16 节点的集群单节点运行速度比 32 节点大约快 27% ；LU 和 SP 在虚拟集群上的运

行时间稍长，其他测试集运行时间大致相同。

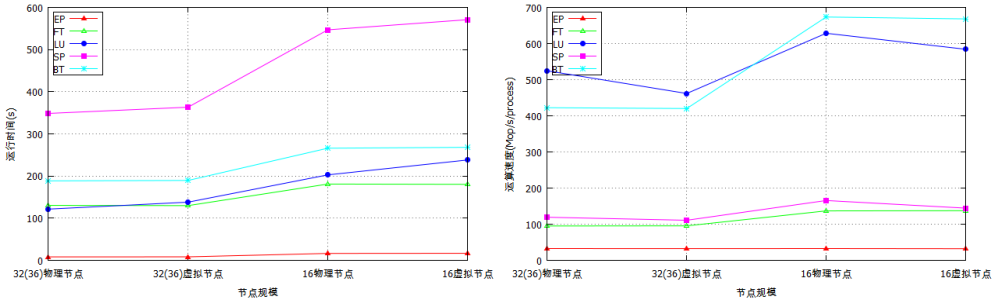

图 4 天河 1-A 集群物理节点 NPB C 类运算时间及速度对比测试结果

从图 4 可以看出，对于 LU，SP 而言，同等节点虚拟集群耗时更长；同时对于 LU、SP，节点数量越少，斜率越大，即额外损耗越大。当节点数量相同时，虚拟节点的运算速度有所降低，而且节点数量越多，这种降低就越明显。

总结来说，LU 和 SP 在虚拟集群上运行时间更长，这是因为 LU 和 SP 有更多的处理器间的通信，所以虚拟集群在处理通信时会有一部分额外的开销，这导致了虚拟集群运行时间更长，而对于其他标准测试集，由于通信较少，更偏重于处理器的计算，所以虚拟集群和实际集群的运行时间相差无几。IS、MG、CG、BT 在 16 节点的集群单节点运行速度比 32(36)节点大约快 40%-50%，FT 在 16 节点的集群单节点运行速度比 32 节点大约快 27%。这是因为 FT 执行了大量的计算，运算速度=作业量/节点数量，当相同的作业量提交到 16 节点的集群时，单个节点的负载更多，会出现部分节点在个别时间段发生过载，所以运算速度并没有因为节点数量减半而翻倍，只是提升了 27%左右，其它测试集则提升了将近 50%。

5 结论

由以上实验可以得出以下结论，集群间过多通信会降低虚拟集群的效率。下一步，我们准备在每一个物理节点上部署 5-7 个虚拟机，使虚拟集群中包括约 70-100 个节点，此时，每个物理机上因为有多个虚拟机，所以会存在网卡复用的现象，当节点间通信增多时，如果虚拟节点位于不同的物理节点上，那么虚拟节点可能会去抢占网卡，那么会造成虚拟集群执行效率更加低下，但如果需要通信的虚拟节点位于同一个物理节点上，那么它们之间的通信是不需要经过网卡的，这将极大的降低虚拟集群的开销。更进一步的研究还包括高性能计算集群上虚拟集群的调度方法研究。

References:

[1] (2013, 7.1). top 500 supercomputer sites. Available: http://www.top500.org/

[2] C. Lazou. Cray's Adaptive Supercomputing - A Paradigm Shift. Available: http://archive.hpcwire.com/hpc/601369.html

[3] W. Huang, J. Liu, B. Abali, and D. K. Panda, "A case for high performance computing with virtual machines," presented at the Proceedings of the 20th annual international conference on Supercomputing, Cairns, Queensland, Australia, 2006.

[4] (2013). IBM Blue Cloud Solution. Available: http://www-31.ibm.com/ibm/cn/cloud/index.shtml

[5] High Performance Computing (HPC). Available: http://aws.amazon.com/ec2/hpc-applications/

[6] Jin Hai, Zhong Alin, Wu Song, and Shi Xuanhua, "Virtual Machine VCPU Scheduling in the Multi-core Environment: Issues and Challenges," Journal of Computer Research and Development, vol. 48, pp. 1216-1224, 2011.

[7] E. Deelman, G. Singh, M. Livny, B. Berriman, and J. Good, "The Cost of Doing Science on the Cloud: The Montage Example," presented at the ACM/IEEE conference on Supercomputing, Austin, Texas, 2008.

[8] C. Hoffa, G. Mehta, T. Freeman, E. Deelman, K. Keahey, B. Berriman, and J. Good, "On the Use of Cloud Computing for Scientific Workflows," presented at the IEEE Fourth International Conference on eScience, 2008.

[9] C. Vecchiola, S. Pandey, and R. Buyya, "High-Performance Cloud Computing: A View of Scientific Applications," presented at the Proceedings of the 2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks, 2009.

[10] L. Jie, M. Humphrey, D. Agarwal, K. Jackson, C. van Ingen, and R. Youngryel, "eScience in the cloud: A MODIS satellite data reprojection and reduction pipeline in the Windows Azure platform," in Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on, 2010, pp. 1-10.

[11] C. Evangelinos and C. N. Hill, "Cloud Computing for parallel Scientific HPC Applications: Feasibility of running Coupled Atmosphere-Ocean Climate Models on Amazon’s EC2," ratio, vol. 2, p. 2.34, 2008.

[12] R. J. Figueiredo, P. A. Dinda, and J. A. B. Fortes, "A case for grid computing on virtual machines," in Distributed Computing Systems, 2003. Proceedings. 23rd International Conference on, 2003, pp. 550-559.

[13] Z. Hill and M. Humphrey, "A quantitative analysis of high performance computing with Amazon's EC2 infrastructure: The death of the local cluster?," in Grid Computing, 2009 10th IEEE/ACM International Conference on, 2009, pp. 26-33.

[14] J. Napper and P. Bientinesi, "Can cloud computing reach the top500?," presented at the Proceedings of the combined workshops on UnConventional high performance computing workshop plus memory access workshop, Ischia, Italy, 2009.

[15] K. R. Jackson, L. Ramakrishnan, K. Muriki, S. Canon, S. Cholia, J. Shalf, H. J. Wasserman, and N. J. Wright, "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud," in Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on, 2010, pp. 159-168.

[16] S. Ostermann, A. Iosup, N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing Cloud Computing." vol. 34, D. R. Avresky, M. Diaz, A. Bode, B. Ciciani, and E. Dekel, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 115-131.

[17] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. H. J. Epema, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing," IEEE Transactions on Parallel and Distributed Systems, vol. 22, pp. 931-945, 2011.

[18] Y. Zhai, M. Liu, J. Zhai, X. Ma, and W. Chen, "Cloud versus in-house cluster: evaluating Amazon cluster compute instances for running MPI applications," presented at the State of the Practice Reports, Seattle, Washington, 2011.

[19] (2013). The NAS Parallel Benchmarks. Available: http://www.nas.nasa.gov/publications/npb.html

[20] (2013). Intel MPI Benchmarks. Available: http://software.intel.com/en-us/articles/intel-mpi-benchmarks/

[21] S. Hongzhang, K. Antypas, and J. Shalf, "Characterizing and predicting the I/O performance of HPC applications using a parameterized synthetic benchmark," in High Performance Computing, Networking, Storage and Analysis, 2008. SC 2008. International Conference for, 2008, pp. 1-12.

附中文参考文献:

[4] (2013). IBM 蓝云解决方案. Available: http://www-31.ibm.com/ibm/cn/cloud/index.shtml

[6] 金海, 钟阿林, 吴松, 石宣化, "多核环境下虚拟机 VCPU 调度研究: 问题与挑战," 计算机研究与发展, vol. 48, pp. 1216-1224, 2011.