第 35 卷 第 7 期 计 算 机 工 程 2009 年 4 月
Vol.35 No.7 Computer Engineering April 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)07—0242—03 文献标识码：A 中图分类号：N945.15

HPCC 在 IBM 刀片机群上的诊断测试与结果分析

王宣强[1]，王向前[1]，张云泉[1,2]
(1. 中科院软件研究所并行计算实验室，北京 100080；2. 中国科学院计算机科学国家重点实验室，北京 100080)

摘 要：在 IBM JS21 BladeCenter 上进行 2 次 HPCC 测试，介绍 HPCC 的结果分析方法，并采用分层模型 AHPCC 对 HPCC 的测试结果进行分析。其目的是通过在高性能机群上执行 HPCC 测试说明 HPCC 测试对机群系统的评价和诊断能力。实验发现，在之前的 HPL 测试结果一直不理想并且无法更进一步发现和解决问题的情况下，采用 HPCC 测试能够较好地评价系统和诊断系统问题。通过分层模型的评价，能够得到更多关于目标系统的性能参数和发现可能的性能瓶颈，为系统设计及构建积累有价值的经验。

关键词：HPCC 测试；分层模型；系统评价

HPCC Diagnostic Test and Results Analysis on IBM BladeCenter Cluster

WANG Xuan-qiang[1], WANG Xiang-qian[1], ZHANG Yun-quan[1,2]
(1. Lab of Parallel Computing, Institute of Software, Chinese Academy of Sciences, Beijing 100080;
2. State Key Lab of Computer Science, Chinese Academy of Sciences, Beijing 100080)

【Abstract】This paper describes a comparative performance test on IBM JS21 BladeCenter using High Performance Computing Challenge (HPCC) tests. It applies the hierarchic analytic model AHPCC to analyze the test results of HPCC. The previous HPL tests get poor results and few reasons can be found. However, through AHPCC model, HPCC can obtain more detailed benchmark information and find performance bottlenecks of the system.

【Key words】High Performance Computing Challenge(HPCC) test; hierarchic model; system evaluation

1 概述

HPL(High Performance Linpack)是国际上最流行的用于测试高性能计算机系统浮点性能的基准测试程序。然而，学术界和工业界已经意识到 HPL 的一些不足。在高性能计算机系统越来越复杂化的今天，单单以 CPU 浮点操作能力无法反映计算机系统可能存在的性能瓶颈，特别是高性能计算机系统集群化。多核化之后，网络带宽和网络延迟、内存存取和存储共享及分级机制，都有可能成为制约系统性能的因素，也有可能影响测评结果。

HPCC(High Performance Computing Challenge)测试基准程序是高性能计算机评测的另一个可选途径。HPCC 已经成为 Top500 评测的一个有效补充，有关方面也在积极地关注将 HPCC 作为计算机评测新标准的可行性。

HPCC 包含几个广泛使用的内核测试程序：(1)HPL，测试系统的浮点操作性能；(2)DGEMM，通过矩阵乘测试系统浮点操作能力；(3)STREAM，测试内存带宽；(4)PTRANS，测试多处理器内存之间传送大规模数组的能力；(5)Random Access，测试随机更新内存的能力；(6)FFT 使用双精度一维 DFT 测试浮点操作能力；(7)通信和延时测试。

HPCC 将它们以统一的参数和环境在同一个系统中运行，使用户能够对系统性能和测试结果有一个全面的了解。

HPCC 包含了 HPL 测试并充分弥补了 HPL 的一些缺点，可用它来对计算机系统进行全面的评价或诊断。但 HPCC 测试也存在一些问题，没有一个统一的结果评价标准，评价体系的构建不方便是其复杂性的最大表现。

由于 HPCC 包含了 HPL，本文参考了很多 HPL 测试方面的经验。文献[1]研究了在 IBM e326, e336 大规模机群上的 HPL 测试，总结了测试过程中重要参数的选择规律；文献[2-3]则引入层次分析法来评价 HPCC 测试结果，使得 HPCC 测试更具可操作性。本文则通过 HPCC 在 IBM 千亿次刀片机群上的诊断测试，研究高性能计算机的性能测试和结果分析方法，对实验系统进行评价和诊断，发现和解决了该系统在通信方面的问题，以一次成功的诊断实践说明 HPCC 测试的结果评价方法和测试价值。

2 HPCC测试的结果评价方法

HPCC 测试了高性能计算机系统的各个方面性能，包括 CPU 速度、内存存取速度、网络带宽、网络延迟等方面，因而如何统一地对系统进行评价就显得非常重要。

分层评价模型 AHPCC[2]是基于 HPCC 和层次分析法提出的高性能计算机系统评价模型。使用层次分析法对对象系统建立递阶层次结构模型，根据权重构造各层次中元素对上层元素的判断向量，该层上所有的判断向量构成该层对上层元素的判断矩阵，最后进行层次单排序和层次总排序及一致性检验，得到测试结果集对目标的总排序权重向量，作为系统总的评价值。当给定的系统用 HPCC 测试集测出各项基本性能指标后，使用分层评价模型，可以依据应用需求对各性能指标进行半定量分析，得出各待选系统的相对指标。

基金项目：国家自然科学基金资助项目(60303020, 60533020)；国家“863”计划基金资助项目(2006AA01A102, 2006AA01A125)；中国软件行业协会数学软件分会和北京邮电大学网络与交换技术国家重点实验室开放课题基金资助项目(2005-05)

作者简介：王宣强(1983－)，男，硕士，主研方向：高性能计算，自适应优化；王向前，硕士；张云泉，研究员、博士生导师

收稿日期：2008-10-14 **E-mail：**shmimy-w@163.com

在图 1[2]中，第 1 层是整体系统，它可以通过第 2 层的 CPU、内存带宽及容量、网络带宽和延迟等方面来进行评价，而第3层是从HPCC测试中抽取的有代表性的主要测试子集，它们从不同的方面来反映第 2 层各项指标。根据分层评价模型，从第 3 层的测试结果得到对第 3 层指标的归一化判断矩阵 $\boldsymbol{P}_{4,3}$；再通过目标系统应用对 CPU 速度、内存容量、网络带宽等要求设定第 3 层元素对第 2 层元素的归一化判断矩阵 $\boldsymbol{P}_{3,2}$；根据目标系统应用对整个系统的要求，设定第 2 层元素对整体系统的判断矩阵 $\boldsymbol{P}_{2,1}$；最后就可以得到 HPCC 测试结果对各层元素的评价以及系统的整体评价(归一化权重向量 $\boldsymbol{W}_4=\boldsymbol{P}_{4,3}\boldsymbol{P}_{3,2}\boldsymbol{P}_{2,1}$)。

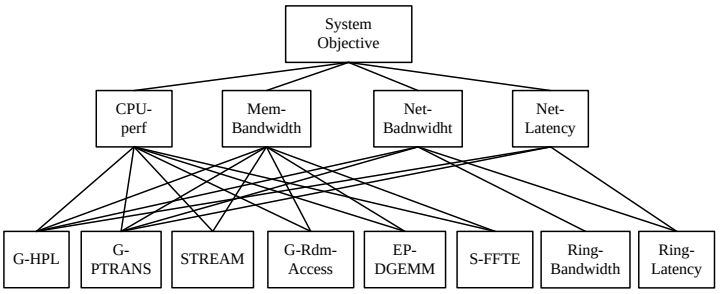

图 1　HPCC 分层评价模型层次图

在实际中，可以通过比较一定规模的不同系统，采用分层评价模型相对地评价它们。本文采用下面的方法来进行 HPCC 结果评价：(1)根据不同的应用需求制定不同的权值标准，用户可以根据实际情况修改该标准。(2)采集几个相近系统配置的测试结果，与本系统测试结果进行横向的相对比较。一般从 HPCC 官方网站获得。(3)在各个层次上可以进行比较，在较低层次上，可以看到本系统可能存在的瓶颈，在较高层次上，有一个统一的分数作为系统性能的衡量标准。

3　HPCC测试实验结果及分析

3.1　实验平台

在千亿次 IBM JS21 刀片机群上进行 HPCC 测试。该机群配置如下：(1)8 个 JS21 刀片节点(2 路 IBM PowerPC 970 2.5 GHz 双核处理器)。(2)每个处理器内核配备 1 MB 二级高速缓存，每个节点配备 4 GB 内存。(3)1 倍速 InfiniBand 网络互联，同时全部刀片以千兆网络互连。

该系统理论峰值 320 Gflop，之前的 HPL 测试在 RHEL4 上用 gcc 3.4.5 编译，数学库采用 GoToBlas 1.14，MPI 采用 MPICH2.1.0.5。得到的测试值仅为 172.4 Gflop，效率为 53.87%，结果并不理想。

3.2　HPCC测试结果及评价

按照第 2 节介绍的方法，采集与本文系统配置相近的几个系统的 HPCC 测试结果，和本文在实验平台上的测试结果放在一起进行对比，见表 1 和表 2。以高性能科学计算作为目标来比较和评价这几个系统。根据第 2 节中介绍的分层评价模型，给出这些系统在分层评价模型中各层间的判断矩阵。

表 1　几个使用 HPCC 测试的机群系统

系统编号	制造商	处理器	峰值/Tflop	网络
1	HP	32	0.384	Infiniband 4x
2	Fujitjsu	64	0.409	Ethernet
3	Tyan	72	0.288	Ethernet
4	Team HPC	56	0.224	Ethernet
5	PathScale	32	0.166	InfiniPath
6	SGI	32	0.205	N/A
7	Sun	64	0.282	Ethernet
8	Cray Inc.	64	0.282	RapidArray
9	Dalco	64	0.307	QsNetII
10	IBM	32	0.320	Infiniband 1x

表 2　HPCC 测试结果

HPL /TFop	*PTRANS* /(GB·s^{-1})	*STREAM* /(GB·s^{-1})	*Rdm access* /(GUP·s^{-1})	*DGEMM* /GFlop	*S-FFT* /GFlop	*Random Ring BW*/(GB·s^{-1})	*Random Ring Lat*/μs
0.270	5.14	1.54	0.015	9.28	1.51	0.300	8.87
0.183	0.39	1.23	0.012	4.11	0.58	0.008	108.00
0.167	0.44	1.53	0.006	3.50	0.22	0.006	72.60
0.160	0.53	1.41	0.008	3.58	0.43	0.008	119.00
0.126	6.71	4.20	0.012	4.77	0.72	0.265	1.31
0.147	6.05	2.00	0.006	5.99	0.74	1.519	3.26
0.218	1.56	2.51	0.008	3.97	0.40	0.041	43.00
0.223	10.5	2.66	0.011	4.03	0.58	0.227	1.63
0.257	9.26	3.51	0.010	4.33	0.49	0.174	4.89
0.172	0.62	1.22	0.004	8.12	0.53	0.011	77.40

判断矩阵 $\boldsymbol{P}_{4,3}$：通过将表 2 中每一列归一化而得到。其中，Ring Latency(表 2 中表示为 Ring Lat)取倒数，再进行归一化。

$$\boldsymbol{P}_{4,3}=\begin{pmatrix}0.140&0.124&0.070&0.163&0.179&0.244&0.117&0.055\\0.950&0.009&0.056&0.132&0.080&0.094&0.003&0.004\\0.087&0.011&0.070&0.066&0.068&0.036&0.003&0.007\\0.083&0.013&0.065&0.084&0.069&0.069&0.003&0.004\\0.065&0.163&0.192&0.137&0.092&0.116&0.104&0.368\\0.077&0.146&0.092&0.062&0.116&0.119&0.593&0.148\\0.113&0.038&0.115&0.087&0.077&0.064&0.016&0.011\\0.117&0.256&0.122&0.119&0.078&0.093&0.089&0.297\\0.134&0.224&0.161&0.109&0.084&0.079&0.068&0.099\\0.089&0.015&0.056&0.040&0.157&0.086&0.004&0.006\end{pmatrix}$$

判断矩阵 $\boldsymbol{P}_{3,2}$：分层模型中第 3 层的 8 个指标对第 2 层指标的判断矩阵。

$$\boldsymbol{P}_{3,2}=\begin{pmatrix}0.5&0.15&0.2&0.25\\0.1&0.1&0.1&0.25\\0.05&0.3&0&0\\0.05&0.25&0&0\\0.2&0.1&0&0\\0.1&0.1&0&0\\0&0&0.5&0\\0&0&0.2&0.5\end{pmatrix}$$

判断矩阵 $\boldsymbol{P}_{2,1}$：分层模型中第 2 层指标对整个系统的判断矩阵。

$$\boldsymbol{P}_{2,1}=\begin{pmatrix}0.6&0.2&0.1&0.1\end{pmatrix}^{\mathrm{T}}$$

所以，表 1 中的 10 个系统对于总目标的归一化权重向量 $\boldsymbol{W}_4=\boldsymbol{P}_{4,3}\boldsymbol{P}_{3,2}\boldsymbol{P}_{2,1}$ 为

$$\boldsymbol{W}_4=(0.14\ \ 0.07\ \ 0.06\ \ 0.06\ \ 0.13\ \ 0.13\ \ 0.08\ \ 0.14\ \ 0.13\ \ 0.07)^{\mathrm{T}}$$

也可以得到表 1 中 10 个系统在模型第 2 层上 CPU、内存等方面的表现。权重向量 $\boldsymbol{W}_3=\boldsymbol{P}_{4,3}\boldsymbol{P}_{3,2}$ 为

$$\boldsymbol{W}_3=\begin{pmatrix}0.1546&0.1377&0.1100&0.0935\\0.0831&0.0823&0.0223&0.0283\\0.0684&0.0621&0.0210&0.0277\\0.0711&0.0680&0.0202&0.0260\\0.0955&0.1389&0.1548&0.2411\\0.0956&0.0926&0.3564&0.1299\\0.0921&0.0910&0.0366&0.0433\\0.1207&0.1265&0.1526&0.2417\\0.1274&0.1343&0.1028&0.1389\\0.0909&0.0661&0.0227&0.0292\end{pmatrix}$$

3.3　结果分析

将这 10 个比较系统的测试结果进行作图比较。图 2 表示了这 10 个系统的 HPL 测试结果和 HPCC 测试的归一化权重向量 $\boldsymbol{W}_4$ 以及 $\boldsymbol{W}_3$ 的各个向量。在图 2 中，HPL 测试结果采用实际值，使用右边的纵轴；而 HPCC 的权重向量是相对比较值，使用左边的纵轴。可以看到：

(1)曲线 HPL 和 HPCC W4 的形状并不一致，特别是在系统 4~系统 7 之间进行比较的时候出现了较大的反差：在曲线 HPL 中，系统 4、系统 7 的 HPL 测试结果高于系统 5、系统 6，而在通过 HPCC 分层模型评估出来的结果曲线 HPCC W4 中，系统 5、系统 6 总体评价高于系统 4、系统 7。

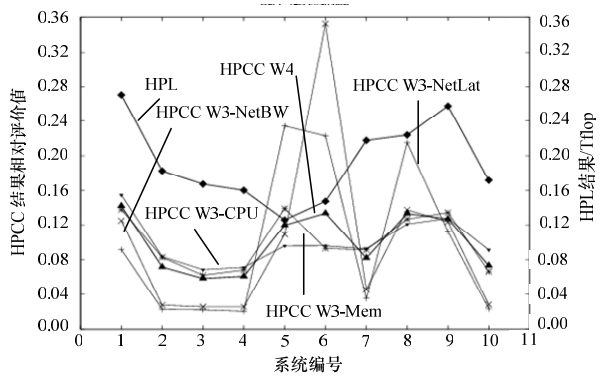

图 2　HPCC 分层评价模型结果比较

(2)HPCC W3 的几条曲线反映了系统的对应方面的性能，不同的系统在不同方面的表现往往差别较大，比如系统 2、系统 3、系统 4、系统 7 以及本文的实验机群，在 HPL 测试中表现都不是最差，但是它们的网络带宽和网络延迟方面的性能都在所有系统中处于较差的位置，因此它们在曲线 HPCC W4 中的位置都低于在 HPCC W3-CPU 中的位置，这说明系统的通信情况严重影响它们的整体性能；相反，系统 5、系统 6 有非常高的网络带宽和较小的网络延迟，因此在 HPCC 整体评价中得分超过了上述的几个系统。通过比较各个系统，能够发现本文目标系统的优缺点。

3.4　第 2 次HPCC测试结果

根据第 3.3 节的分析，本文的实验机群在网络通信方面表现不佳，系统的 CPU 实测性能都只能达到系统理论峰值的 54%。问题在于本文的系统一开始没有安装对应 InfiniBand 网卡通信专有的 MPI，因此实际上通信使用的是系统中的以太网。改用了 InfiniBand 的最新驱动 OFED1.2，MPI 采用 MVAPICH2，重做 HPL 测试，得到的测试值为 192.9 Gflop，效率为 60.29%，单独的网络带宽和延迟测试结果显示，在本文的实验平台上使用以太网通信的带宽和延迟分别是 117.6 MB/s 和 54.59 μs，而通过 InifinBand 网卡通信的带宽和延迟则分别是 241.3 MB/s 和 6.15 μs。但在仅仅做 HPL 测试的时候，根本无从判断系统的性能瓶颈在何处。重做 HPCC 测试，并根据前述的结果分析方法，得到各系统的评价值 $\boldsymbol{W}_4$ 为

$$\boldsymbol{W}_4 = (0.14\ \ 0.07\ \ 0.06\ \ 0.06\ \ 0.12\ \ 0.12\ \ 0.08\ \ 0.13\ \ 0.12\ \ 0.09)^{\mathrm{T}}$$

同样作图以说明，如图 3 所示，相对于图 2，加入了图 2 中各系统的评价值(W4-before)和本次测试得到的评价值(W4-after)进行对比。

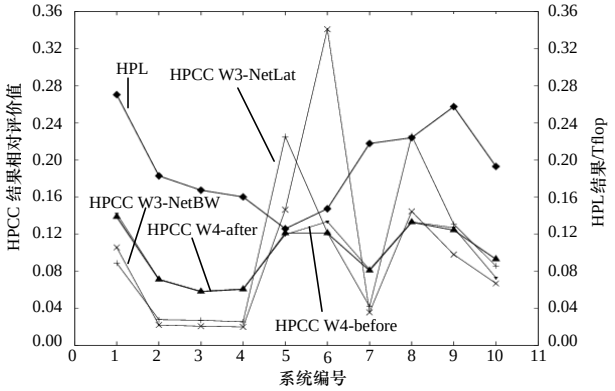

图 3　HPCC 分层评价模型结果比较

从图 3 中可以看到，本文实验系统的网络带宽和延迟的评价值已不像图 2 中那样严重拉低系统总评价值，系统总评价值提高了 27.3%。

4　结束语

本文在 IBM 刀片机群上进行了 HPCC 的诊断测试。通过前后 2 次测试，根据分层模型分析测试结果，找出了系统通信能力方面的问题并解决了问题。通过本文的测试实践和结果评价过程，可以看到，HPCC 测试能够更充分更精细地刻画系统的性能，帮助找到系统可能存在的性能瓶颈。分层评价模型能够有效地分析 HPCC 测试结果。完善和发展 HPCC 的结果评价方法，是下一步的工作。

参考文献

[1] Pase D M. Linpack HPL Performance on IBM eServer 326 and xSeries 336 Servers[EB/OL]. (2005-07-29). ftp://ftp.software.ibm.com/eserver/benchmarks/wp_Linpack_072905.pdf.

[2] 刘川意, 汪东升. 基于HPCC 和层次分析法的高性能计算系统评价模型[J]. 软件学报, 2007, 18(4): 1039-1046.

[3] 王宣强, 王向前, 张云泉. HPL 与 HPCC 在 IBM 刀片机群上的对比测试[C]//CNCC 中国计算机大会. 中国, 苏州: [出版者不详], 2007.

编辑　任吉慧

（上接第 241 页）

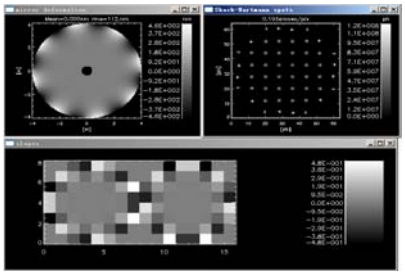

图 7　“变形镜”项目在 SciSimu 上的仿真结果界面

7　结束语

经过大量仿真测试证明本设计原理是正确的，开发工作使 SciSimu 获得了自适应光学组件建模与仿真的能力，也增强了 SciSimu 的可扩展性。

参考文献

[1] Carbillet M, Verinaud C, Guarracino M, et al. CAOS: A Numerical Simulation Tool for Astronomical Adaptive Optics(and Beyond)[C]//Proc. of Advancements in Adaptive Optics. Glasgow, Scotland, UK: [s. n], 2004.

[2] 闫殿武. IDL 可视化工具入门与提高[M]. 北京: 机械工业出版社, 2003.

[3] Campbell S L, Chancelier J P, Nikoukhah R. Modeling and Simulation in Scilab/Scicos[M]. [S. l.]: Springer, 2006.

[4] Carbillet M, Verinaud C, Femenia B, et al. Modelling Astronomical Adaptive Optics-I[J]. Monthly Notices of the Royal Astronomical Society, 2005, 356(4): 1263-1275.

编辑　陆燕菲