

曙光 5000A 天体大规模数值模拟软件性能测试

王婷¹, 孙相征^{1,2,3}, 张云泉^{1,2}, 杨超¹, 李力刚⁴, 刘芳芳¹, 管文华¹, 唐雨新^{1,2,3}, 姚继峰⁵

(1. 中国科学院软件研究所并行计算实验室, 100190, 北京; 2. 中国科学院软件研究所计算机科学
国家重点实验室, 100190, 北京; 3. 中国科学院研究生院, 100190, 北京; 4. 中国科学院
上海天文台, 200030, 上海; 5. 上海超级计算中心, 201203, 上海)

摘要: 在国产百万亿次超级计算机曙光 5000A 上进行了天体大规模数值模拟软件的性能和可扩展测试实验, 详细介绍了软件中的测试程序以及测试环境和过程, 并对测试结果进行了分析. 对于 $80 \times 80 \times 50$ 的网格规模, 采用每节点 4 进程测试了 16~128 个处理器核, 每节点 8 进程、16 进程分别测试了 16~512 个处理器核, 相对加速比最终分别达到 5.33、10.48 和 12.57, 并行效率分别达到 66.66%、32.58% 和 32.29%. 对于 $160 \times 160 \times 100$ 的网格规模, 测试了每节点 16 进程的 64~8 192 个核的性能, 最大相对加速比为 12.46, 并行效率为 9.73%. 测试结果表明, 曙光 5000A 具有良好的性能, 测试结果对软件下一步的优化研究具有重要的指导意义.

关键词: 曙光 5000A; 数值模拟; 性能测试; 可扩展

中图分类号: TP393 **文献标志码:** A **文章编号:** 0253-987X(2009)10-0071-05

Performance Evaluation of the Simulation Software on Dawning 5000A for Large Scale Celestial Bodies

WANG Ting¹, SUN Xiangzheng^{1,2,3}, ZHANG Yunquan^{1,2}, YANG Chao¹, LI Ligang⁴
LIU Fangfang¹, GUAN Wenhua¹, TANG Yuxin^{1,2,3}, YAO Jifeng⁵

(1. Laboratory of Parallel Computing, ISCAS, Beijing 100190, China; 2. State Key Laboratory of Computer Science, ISCAS, Beijing 100190, China; 3. Graduate University of Chinese Academy of Sciences, Beijing 100190, China; 4. Shanghai Astronomical Observatory, CAS, Shanghai 200030, China; 5. Shanghai Supercomputer Center, CAS, Shanghai 201203, China)

Abstract: The performance and scalability evaluation of the large scale simulation software on planetary fluid dynamics are investigated by the experiments on the domestic 100 TFlops super-computer, Dawning 5000A. The details of the software testing procedures, the environment and the process of the testing are introduced. For the mesh size of $80 \times 80 \times 50$, the performance testing on 16 to 128 processor cores with mapping 4 processes on each node and the performance testing on 16 to 512 processor cores with mapping 16 or 8 processes on each node are performed, respectively. The relative speedups reach 5.33, 10.48 and 12.57, respectively, and the corresponding parallel efficiencies of these 3 testing cases are 66.66%, 32.58% and 32.29%, respectively. For the mesh size of $160 \times 160 \times 100$, the performance testing on 64 to 8 192 processor cores with mapping 16 processes on each node gives the results that the relative speedup is 12.46, and its parallel efficiency is 9.73%. The experimental results reveal the good performance of Dawning 5000A and show important clues of the software optimization.

Keywords: Dawning 5000A; numerical simulation; performance evaluation; scalability

在国产百万亿次超级计算机曙光 5000A 上对天体大规模数值模拟软件中的旋转行星球壳内的流体动力学程序进行了性能和可扩展性测试. 实验结果体现了曙光 5000A 良好的性能,同时也对下一步天体大规模数值模拟软件的优化研究有重要的指导意义.

曙光 5000A 高性能计算机是国家“863 计划”高性能计算机及其核心软件重大专项支持的研究项目,是面向网格的高性能计算机,可以为网格提供计算服务,同时也是面向信息服务的超级服务器,可以提供多目标的系统服务. 曙光 5000A 系统峰值运行速度达到 2.33×10^{14} 次/s 浮点操作, Linpack 运行速度超过 1.8×10^{14} 次/s 浮点操作,是目前国内速度最快的商用高性能计算机系统. 2008 年 11 月,曙光 5000A 在国际超级计算机排名 TOP500 中列第 10 位,在中国高性能计算机性能排名 TOP100 中名列第 1 位^[1-2].

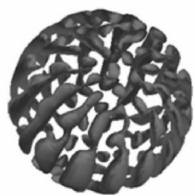
在天文学的行星研究中,行星流体和磁流体动力学数值模拟是近年来国际关注的重点之一. 行星流体动力学中的时空尺度跨度大、非线性耦合、球形几何形状且快速旋转以及极端的动力学参数等特点,使得必须进行大规模的并行求解^[3]. 本文对天体大规模数值模拟软件中描述行星快速旋转时流体运动过程的并程序进行了测试,该程序模拟的不同参数下旋转球壳的温度等值面图如图 1 所示^[4-5].



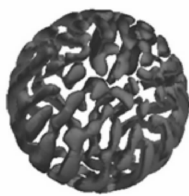
(a)慢速旋转



(b)快速旋转低 Rayleigh 数



(c)快速旋转中 Rayleigh 数



(d)快速旋转高 Rayleigh 数

图 1 旋转球壳内计算结果的温度等值面图

1 天体大规模数值模拟软件简介

天体大规模数值模拟软件包中的测试程序是基于有限差分法求解如下的球壳内行星流体动力学方程组^[6-7]

$$Ek \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) + 2\mathbf{k} \times \mathbf{u} = -\nabla p +$$

$$Ra \frac{\mathbf{r}}{r_o} \Theta + Ek \nabla^2 \mathbf{u} \quad (1)$$

$$\frac{\partial \Theta}{\partial t} + \mathbf{u} \cdot \nabla \Theta = \left(\frac{r_i r_o}{r^3} \right) \mathbf{r} \cdot \mathbf{u} + \frac{1}{Pr} \nabla^2 \Theta \quad (2)$$

$$\nabla \cdot \mathbf{u} = 0 \quad (3)$$

式中: r, θ, ϕ 分别为球坐标中的半径、余纬度和经度; $\mathbf{r}$ 为位置矢量, 可以写为 (r, θ, ϕ) ; $\mathbf{u}(\mathbf{r}, t)$ 、 $\Theta(\mathbf{r}, t)$ 、 $p(\mathbf{r}, t)$ 分别为待求的速度矢量、温度和压强, 其中 $\mathbf{u}$ 的 3 个分量为 (u_r, u_θ, u_ϕ) ; t 为时间; $\mathbf{k}$ 为行星自转轴方向单位矢量; r_i, r_o 分别为球壳的内、外半径; Ek, Ra, Pr 分别为 Ekman 数、Rayleigh 数和 Prandtl 数.

被测程序在半径方向是采用非均匀网格划分的, 采用有限差分方法对 Boussinesq 近似下不可压缩流体的归一化方程式(1)~式(3)进行离散, 时间差分格式采用具有二阶精度的 Crank-Nicolson 格式, 并结合近似因式分解方法^[3,8]分离压强 p .

被测程序模拟了旋转行星内部球壳中流体的热对流情况. 程序中先求解速度方程, 再求解压力方程(共 5 个方程, 2 个大方程组), 完成一个时间步迭代后, 继续返回速度方程求解下一个时间步. 程序的具体流程为: ①读入数据进行初始化后, 形成速度、温度方程矩阵和求解压强的泊松方程矩阵; ②开始时间循环迭代, 先计算速度、温度方程的右边项, 用迭代法求解出速度、温度后, 再计算泊松方程的右边项, 算出压强之后进行速度校正, 同时输出能量、速度和温度数据; ③进行下一个时间循环迭代, 直至规定的时间循环步数. 本文采取 $80 \times 80 \times 50$ 和 $160 \times 160 \times 100$ 两种网格规模, 未知量的个数大约是网格规模的 5 倍, 分别约为 1.58×10^6 和 1.272×10^7 个. 由于实验时间较紧促, 对于这两种网格规模分别测试了 10 个及 8 个时间步, 验证了程序的可行性.

2 测试环境和测试过程

曙光 5000A 的计算节点采用 4 路 4 核 AMD Barcelona (主频 1.9GHz) 处理器, 每个节点的内存为 64 GB, 本次实验中使用的计算节点可以满足 8 192 (512×16) 个处理器核的程序测试. 计算节点用 InfiniBand 高速网络进行互联, 使用 SuSe Linux 操作系统. 采用的编译器版本为 GCC-4.1.2 和 Gfortran-4.1.2, MPI 通信库为 mvapich-1.1.0.

实验调用 Aztec 库^[9]生成了程序的可执行文件. 对于 $80 \times 80 \times 50$ 的网格规模, 采用每节点 16 进

程、8 进程和 4 进程分别测试了 16、32、64、128、512 个核的情形,以下简称测试 A。对于 $160\times160\times100$ 的网格规模,测试了每节点 16 进程的 64、128、256、512、1 024、2 048、4 096、8 192 个核的情形,以下简称测试 B。采用 Mvapich 软件包作为 MPI 通信库, gcc 作为编译器,并对 Mvapich 软件包进行了 InfiniBand 高速互连网络特殊优化。

3 实验结果

3.1 测试 A 的实验结果

当网格规模为 $80\times80\times50$ 时,图 2 中给出了曙光 5000A 天体数值模拟中每节点 16、8 和 4 进程的运行时间随处理器核数的变化情况。图 3~图 5 分

别给出了每节点 16、8 或 4 进程的运行时间和相对加速比随处理器核数的变化情况,其中的相对加速比均是以 16 个核的运行时间记为 1 做的比较。

3.2 测试 B 的实验结果

图 6 为网格规模为 $160\times160\times100$ 时曙光 5000A 天体数值模拟中运行时间或相对加速比随处理器核数的变化情况,其中的相对加速比是取以 64 个核的运行时间记为 1 做的比较。

4 实验结果分析

从图 2 中可以看出,测试 A 中对于 16 到 512 个处理器核的并行模拟中,运行时间随核数的增加呈下降趋势。在处理器核数较少时,相同处理器核数

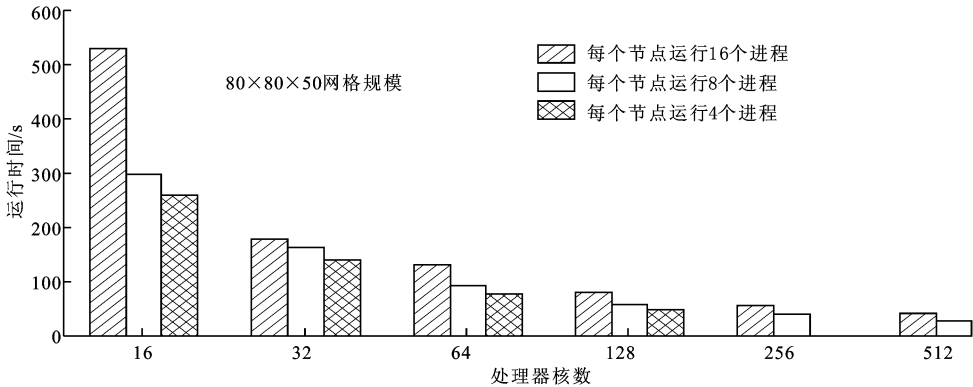


图 2 测试 A 中曙光 5000A 天体数值模拟运行时间随处理器核数的变化情况

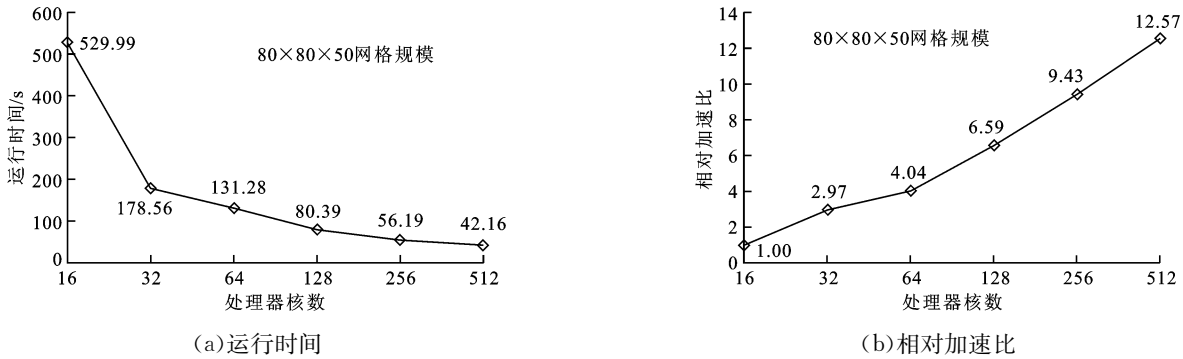


图 3 测试 A 中每节点 16 进程时运行时间和相对加速比随处理器核数的变化情况

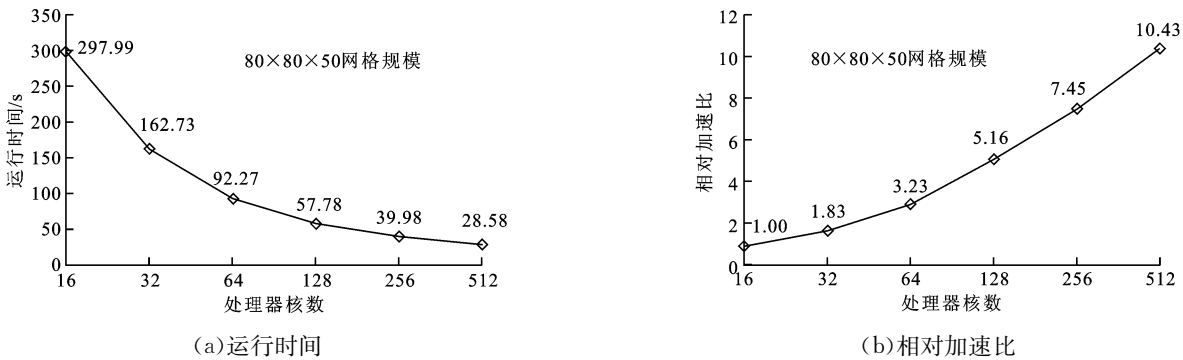
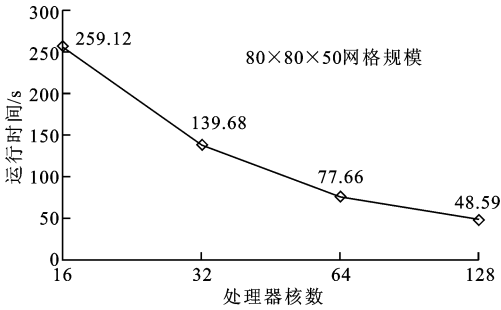
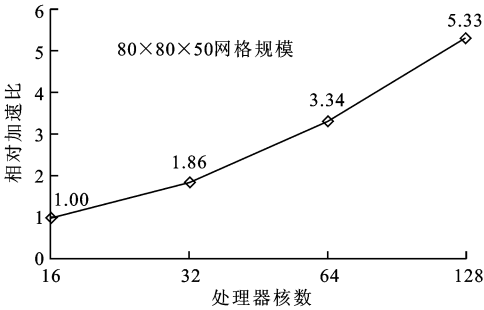


图 4 测试 A 中每节点 8 进程时运行时间和相对加速比随处理器核数的变化情况

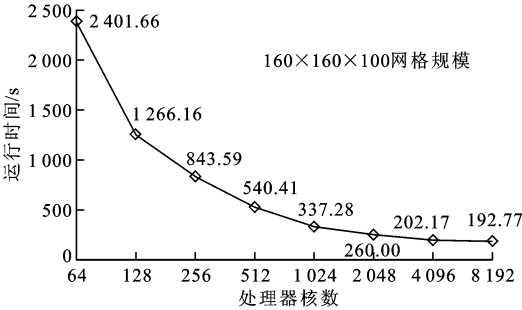


(a)运行时间

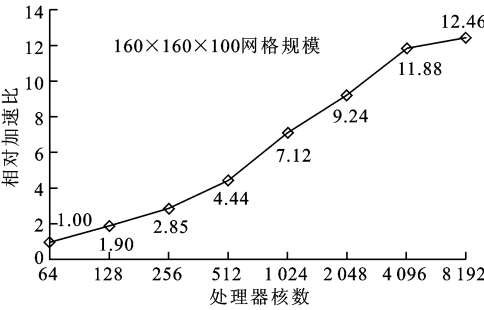


(b)相对加速比

图 5 测试 A 中每节点 4 进程时运行时间和相对加速比随处理器核数的变化情况



(a)运行时间



(b)相对加速比

图 6 测试 B 中运行时间和相对加速比随处理器核数的变化情况

下每节点不同进程时的运行时间变化较大,如在 16 个核时,每节点 8 进程和 4 进程的运行时间比 16 进程的减少了 232 s 和 270.87 s,且运行时间分别是每节点 16 进程的 56.23%和 48.89%。随着处理器核数的增加,运行时间逐渐减少,使得这种差距变得不再那么显著,如在 512 个核时,每节点 8 进程的运行时间比 16 进程减少了 13.58 s。

在每节点 16 进程和 8 进程的 16~512 个处理器核的运行过程中,相对加速比随处理器核数的增加而增长,在 512 个处理器核时的相对加速比分别达到了 16 个核时的 12.57 倍和 10.43 倍,并行效率分别达到 32.29%和 32.58%。每节点 4 进程的相对加速比最终达到 5.33 倍,并行效率则高达 66.66%。这是因为此情况只测试了 16~128 个核,考虑到每节点 16 进程和 8 进程时 128 个处理器核的相对加速比分别为 6.59 和 5.16,也可以和前面的 5.33 倍相对应。对于 128 个核采用每节点 4 进程比 256 个核采用每节点 16 进程的运行时间还要少的问题,原因可能是由于每节点的进程过多,导致存储访问冲突增加^[10-11]。这也提示我们,在大规模并行测试中,在测试节点足够多的情况下,尽量不要把节点上的处理器核全部用满。

在测试 B 的 64~8 192 个处理器核的测试过程

中,随着处理器核数的增加,程序的运行时间逐渐下降,这显示出曙光 5000A 提供了天体大规模数值模拟程序所需的良好网络通信环境。8 192 个核的运行时间是 64 个核运行时间的 10.22%。在 64 到 512 个处理器核的测试过程中,程序随着处理器核规模的上升,所需要的运行时间随之减少,512 核的相对加速比为 64 核的 4.44 倍,并行效率为 55.55%。在 512 到 4 096 个处理器核的测试过程中,相对加速比的增长速度进一步加大,4 096 核的相对加速比增长为 64 核的 11.88 倍,并行效率为 18.56%。最后到 8 192 核时,程序的计算时间仍然在减少,比 64 核时减少了 2 208.89 s,相对加速比最终达到了 64 核的 12.46 倍,由此可以看出曙光 5000A 上可以较顺利地运行此天体大规模模拟应用软件,但是此时的并行效率降为 9.73%,也观察到相对加速比有放缓增长的趋势。这是由于在本次大规模并行性能测试中,保持测试的问题规模固定不变,从而导致随着网格的划分不断细化,每核分配到的计算时间快速减少,而核与核之间的通信开销却快速增加,更短的消息长度和更多的通信次数使得通信的开销增大,计算和通信开销比快速下降^[10-12]。这使得在下一步的研究工作中,需要考虑对并行算法进行进一步改进和优化。本次实验所暴露出的性能瓶颈对于天体

大规模数值模拟软件的研发具有重要的指导意义。

5 小结及未来工作

尽管此次测试时间仓促,我们还是在曙光 5000A 上比较顺利地完成了此次大规模测试。在多路多核处理器、高速互联网络等体系结构设计方面,曙光 5000A 都展现了其对大规模并行测试的有力支撑。天体数值模拟程序在整个测试中的运行时间随处理器核数的增加而不断减少,直到 8 192 个处理器核都始终保持了相对加速比的提高。此次还成功地在曙光 5000A 上测试了 16 384 个核的运行时间,但由于问题规模固定不变,运行时间并未继续减少。下一步准备增大问题规模,争取在 16 384 个核继续得加速比的提高。

如何更进一步地提高天体数值模拟程序在更大的问题规模和更大机器规模^[13]上的性能和可扩展性,需要根据本次测试发现的问题和瓶颈,对并行算法及其实现进行更加深入细致的分析和优化,这将是我们的下一步的工作重点。

致谢 衷心地感谢中国科学院计算技术研究所张佩珩研究员级高级工程师提供宝贵的实验机会以及霍志刚博士和张潘勇博士在实验中的热心帮助,同时感谢本实验室张先轶、余元等对实验工作的支持和帮助。

参考文献:

- [1] 张云泉,孙家昶,袁国兴,等. 2008 年中国高性能计算机 TOP100 排行榜分析与展望 [J]. 科研信息化技术与应用,2008,1(3): 71-78.
ZHANG Yunquan, SUN Jiachang, YUAN Guoxing, et al. Analysis of 2008 China HPC TOP100 rank list and perspectives [J]. e-Science Technology and Application, 2008,1(3): 71-78.
- [2] 中国软件行业协会数学软件分会,国家 863 高性能计算机评测中心,中国计算机学会高性能计算专业委员会. 2008 年中国高性能计算机性能 TOP100 排行榜 [EB/OL]. (2009-03-13) [2009-03-28]. <http://www.samss.org.cn/sites/shuxue/ndhyC.jsp?contentId=2473512102846>.
- [3] CHANA K H, LI Ligang, LIAO Xinhao. Modelling the core convection using finite element and finite difference methods [J]. Phys Earth Planet Interiors, 2006, 157(2): 124-138.
- [4] LI Ligang, LIAO Xinhao, ZHANG Keke. Linear and nonlinear instabilities in rotating cylindrical Rayleigh-Benard convection[J/OL]. Physical Review: E, 2008, 78(5): 12[2009-02-20]. <http://link.aps.org/doi/10.1103/PhysRevE.78.056303>.

- [5] LI Ligang, LIAO Xinhao, ZHANG Keke. Counter-traveling waves in rotating Rayleigh-Benard convection [J/OL]. Physical Review: E, 2008, 77(2): 4[2009-02-18]. <http://link.aps.org/doi/10.1103/PhysRevE.77.027301>.
- [6] 李力刚. 球壳内行星流体动力学方程组的有限差分法 [R]//国家 863 高效能计算机及网格服务环境项目技术报告. 北京:中国科学院,2008: 10-20.
- [7] 杨超. 球壳对流的有限差分程序注记 [R]//国家 863 高效能计算机及网格服务环境项目技术报告. 北京:中国科学院,2008: 30-40.
- [8] DUKOWICZ J K, DVINSKY A S. Approximate factorization as a high order splitting for the implicit incompressible flow equations [J]. Journal of Computational Physics, 1992, 102(2): 336-347.
- [9] TUMINARO R S, HEROUX M, HUTCHINSON S A, et al. Official AZTEC users guide: version 2.1 [DB/OL]. [2009-02-20]. <http://www.CS.sundia.gov/CRF/pspapers/Aztec-og-2.1.ps>
- [10] 张云泉,孙家昶,唐志敏,等. 数值计算程序的存储复杂性分析 [J]. 计算机学报,2000,23(4):363-373.
ZHANG Yunquan, SUN Jiachang, TANG Zhimin, et al. Memory complexity analysis on numerical programs [J]. Chinese Journal of Computers, 2000, 23(4):363-373.
- [11] 张云泉. 面向高性能数值计算的并行计算模型 DRAM(h) [J]. 计算机学报, 2003, 26(12): 1660-1670.
ZHANG Yunquan. DRAM(h): a parallel computation model for high performance numerical computing [J]. Chinese Journal of Computers, 2003, 26(12): 1660-1670.
- [12] CHEN Jing, ZHANG Yunquan, ZHANG Linbo, et al. Performance evaluation of allgather algorithms on terascale Linux cluster with fast ethernet [C]//Proceedings of Eighth International Conference on High-Performance Computing in Asia-Pacific Region. Los Alamitos, CA, USA: IEEE Computer Society, 2005: 437-442.
- [13] 薛正华,董小社,伍卫国,等. 自适应大规模服务器集群监控系统的构建 [J]. 西安交通大学学报, 2008, 42(4): 399-403.
XUE Zhenghua, DONG Xiaoshe, WU Weiguo, et al. Design and implementation of an adaptive monitoring system for large scale server clusters [J]. Journal of Xi'an Jiaotong University, 2008, 42(4): 399-403.