

“天河一号”大规模并行应用程序测试

朱小谦^{1,2} 孟祥飞¹ 曹晓东¹ 冯景华¹

(国家超级计算天津中心 天津 300457)¹ (国防科技大学计算机学院 长沙 410073)²

摘 要 目前安装在国家超级计算天津中心的“天河一号”是我国首台千万亿次超级计算机,在 2010 年 11 月世界超级计算机 Top500 排名中,位列世界第一。“天河一号”采用了 CPU 与 GPU 相结合的异构融合计算体系结构,自主设计实现了高速互连通信系统,在多个高性能计算应用领域具有应用适应性强、系统运行稳定可靠、性能可扩展性好等特点,为科学研究和应用提供了重要的高性能计算平台。采用石油地震数据处理、飞行器流场模拟、生物分子动力学模拟、磁约束聚变数值模拟、湍流数值模拟、晶体硅分子动力学模拟、全球大气浅水波全隐式数值模拟、地球外核热流动数值模拟等应用的典型算例对“天河一号”进行了大规模并行程序测试,结果表明,“天河一号”在上述应用领域具有良好的可扩展性和并行效率。

关键词 天河一号,高性能计算,应用,并行效率

中图法分类号 TP399 **文献标识码** A

Performance Test of Massively Parallel Program on TH-1A

ZHU Xiao-qian^{1,2} MENG Xiang-fei¹ JIAN Xiao-dong¹ FENG Jing-hua¹

(National Supercomputer Center in Tianjin, Tianjin 300457, China)¹

(College of Computer, National University of Defense Technology, Changsha 410073, China)²

Abstract TH-1A is the china's first petaflops supercomputer and is now installed in national supercomputer center in Tianjin, which is ranked No. 1 on Top500 list released in Nov. 2010. TH-1A system adopts a hybrid architecture of heterogeneous integration of CPU+GPU and self-intellectual high-speed interconnection system, and shows distinguished system usability, performance stability and application scalability in quite a few high performance computing areas, provides an important platform for science research and technology innovation. In this paper, several large-scale application tests on TH-1A were introduced, such as the oil seismic data processing, aircraft flow field simulation, biomolecular dynamics simulation, magnetic confinement fusion numerical simulation, turbulent flow simulation, crystal silicon molecular dynamics simulation, fully implicit simulation of global atmospheric shallow wave, heat flow simulation of earth outer core. These results show that TH-1A has good parallel efficiency and scalability in practical applications.

Keywords TH-1A, High performance computing, Application, Parallel efficiency

1 引言

当今世界科技发展日新月异,以信息技术和计算机技术为基础的高性能计算快速发展,成为继实验科学和理论研究后的第三大科学研究方法,并迅速成为商业计算和信息化服务发展的核心动力,推动社会生产力不断发展。高性能计算机的研制和应用是一个国家的核心竞争力,是一个国家综合国力的体现。

超级计算是在高性能计算机上,采用并行计算技术,针对大型科学与工程计算问题,实现多个(通常数千个以上)计算处理器协同并行工作,共同求解大型计算问题。超级计算主要面向计算密集型和通信密集型应用,通常具有较高的计算精度和时效性要求。

高性能计算应用是充分发挥超级计算机的计算性能,为科研和生产服务的关键。从 2010 年 11 月“天河一号”正式发

布以来,中石油东方地球物理公司(BGP)、中石化物探技术研究院、中科院力学所、中科院上海药物研究所、北大聚变中心、北大工学院与信息学院、中科院过程所、中科院软件所等应用单位与国家超级计算天津中心合作,在“天河一号”上完成了多个典型应用的大规模并行程序测试,取得了令人鼓舞的性能测试结果。目前,“天河一号”在国家超级计算天津中心为 150 多个用户提供计算服务,具有良好的推广应用前景。

2 系统概况

“天河一号”(TH-1A)是我国首台千万亿次超级计算机。它采用 CPU 和 GPU 相结合的异构融合计算体系结构,峰值速度 4700TFlops,持续速度 2566TFlops(LINPACK 实测值),内存总容量 262TB,存储总容量 2PB。

全系统包含 7168 个计算结点和 1024 个服务结点。其中,每个计算结点包含 2 个英特尔至强 X5670 CPU(2.93GHz、

6核)和1个英伟达 M2050GPU(1.15GHz,14核/448个CUDA核);每个服务结点包含2个飞腾-1000CPU(1.0GHz,8核)。全系统共有3552个微处理器,其中CPU核102400个,GPU核100352个。

“天河一号”采用自主设计的高阶路由芯片NRC和高速网络接口芯片NIC,实现光电混合的二层胖树结构高阶路由网络互连。链路双向带宽160Gbps,延迟1.57us,单背板交换密度61.44Tbps。输入、输出系统采用Lustre全局分布共享并行I/O结构,总容量2PB。监控诊断系统采用分布式集中管理结构,实现系统实时安全监测、控制和调试诊断。编译系统支持C、C++、Fortran77/90/95、Java语言,支持OpenMP、MPI并行编程,支持异构协同编程框架,高效发挥CPU和GPU的协同计算能力。

3 应用领域与测试

3.1 石油勘探数据处理

目前,我国已逐步进入能源消费的高增长期,能源需求大幅增加,油气供需矛盾日趋突出,能源安全面临巨大挑战,迫切需要利用高性能计算机系统,研究突破地球物理勘探关键技术,为我国油气资源勘探开发提供物探装备和技术保障。

中石油东方地球物理公司(BGP)与国家超级计算天津中心合作,在“天河一号”计算机上对其自主研发的GeoEast-lightning单(双)程波叠前深度偏移软件系统进行了实际应用测试。表1是采用GeoEast-lightning单(双)程波叠前深度偏移软件系统处理中亚某国勘探区域和我国东北某勘探区域的地震资料数据的计算参数和实际计算性能^[1]。

表1 GeoEast-lightning 软件的应用测试

处理区域	面积 (平方公里)	深度 (公里)	数据规模 (炮)	计算结点	系统 利用率	计算时间 (小时)
中亚某国	1050	5	70000	7100	99%	16
中国东北 某区域	680	9	80000	7000	99%	40

中石化物探技术研究院采用具有自主知识产权的软件iCluster地震叠前偏移成像系统中的炮域单程波动方程叠前深度偏移、kirchhoff积分法叠前时间体偏移和高斯束叠前深度偏移3个偏移计算模块对“天河一号”进行性能测试,测试结果如表2所列^[2]。

表2 Kirchhoff 时间偏移及炮域深度偏移的应用测试

偏移算法	数据量 (GB)	面积 (平方公里)	炮数 (CDP数)	结点数 (进程数)*	计算时间 (小时)
Kirchhoff 叠前时间偏移	486	186	464400	1024 (10240)	14.4
炮域波方程 叠前深度偏移(高密度)	1300	31	4096	410 (4097)	1.3

*天河每结点最大进程数是12,实际启用进程数由应用需求决定

经过与应用单位在其它计算机系统上的测试结果对比表明,“天河一号”超级计算机系统计算能力强,互连通信速度快,运行稳定可靠。经过分析,各偏移模块能准确成像,完全满足石油勘探地震数据处理领域的计算需求。

3.2 飞行器流场模拟仿真

自主开发的飞行器流场模拟仿真软件CCFD由中航第一飞机设计研究院、中国科学院力学研究所、中国航空计算技术研究所和中科院网络信息中心4家单位共同开发,其主要用于面向航空航天领域飞行器设计流场模拟仿真,以提高飞行

器气动设计效率。

CCFD软件在“天河一号”上取得了良好的并行计算加速比,见图1。该测试算例为可压缩平板绕流数值模拟,来流马赫数等于8,来流温度等于169.44K(壁面温度等于1.9倍来流温度),雷诺数等于2000000。

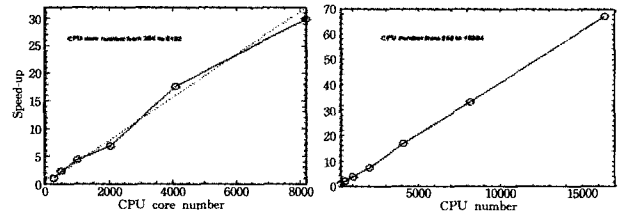


图1 CCFD加速比测试

除测试并行效率外,还进行了RAE2822绕流模拟和强冷壁平板绕流模拟等测试,均采用了8192CPU核,其中RAE2822超临界翼型是NASA NPARC Alliance Verification and Validation Archive标准算例,有可靠实验数据、高质量网格及他人计算结果(RANS),但目前尚无DNS或LES的计算结果,本应用测试尝试了这方面的工作。高马赫数高雷诺数平板绕流是当前国际上的计算难题之一,目前报道较多的结果是马赫数小于6,雷诺数约为2百万或更低的结果。本次测试对马赫数等于8的平板绕流做了直接数值模拟并取得了良好的测试结果,这充分表明“天河一号”超级计算机在航空航天领域的应用价值。

3.3 生物分子动力学模拟

近几十年来,分子动力学(MD)模拟的发展为研究蛋白质动力学行为、蛋白质动态相互作用和生物网络分析提供了新的途径,大大扩展了蛋白质科学和其他生命科学研究领域的广度和深度。目前,MD模拟与各种实验方法结合,已经成为研究生命科学问题的一个重要手段。

中科院上海药物研究所(简称上海药物所)与国家超级计算天津中心合作,在“天河一号”上采用Gromacs、NAMD等测试软件,针对不同大小模拟体系开展了一系列MD模拟测试,评估“天河一号”在蛋白质构象变化研究和蛋白质折叠研究领域的应用价值。

上海药物所采用Gromacs软件对中小规模的模拟体系(约两万个原子)进行测试。截断半径为12Å,采用512个核,计算模拟速度大约为330纳秒/天,是测试人员目前见到的最好计算性能结果。大规模的模拟体系(约四十万个原子),截断半径为12Å,采用2048个核,Gromacs计算模拟速度为85纳秒/天,也是非常不错的性能测试结果。

上海药物所采用NAMD软件对包括蛋白质、磷脂和水,共967172个原子数的体系进行测试。截断半径为12Å,Particle Mesh Ewald(PME)全静电势,时间步长设置为2fs,温度为310K,能量和压力输出的频率为0.1ps,模拟时长为500皮秒。各次测试采用相同的初始文件和不同的CPU核数运行,并行计算加速比见图2^[4]。

国家超级计算天津中心选取美国伊利诺伊州大学香槟分校贝克曼研究所理论生物物理小组公布的卫星烟草花叶病毒算例(STMV,此体系共有1,066,628个的原子)^[5],在“天河一号”上采用NAMD软件进行测试,测试算例使用条件截断半径为12Å,Particle Mesh Ewald(PME)全静电势,时间步长1飞秒,模拟时长为5纳秒。对于百万原子规模的生物体

系,采用 6144 个核计算,模拟速度大约为 20 纳秒/天。

如图 3 所示,1024CPU 核时并行效率为 87.20%,2048 CPU核时并行效率为 80.27%,该测试结果与中科院网络中心深腾 7000 的测试结果比较^[6]表明,“天河一号”具有更好的并行效率。

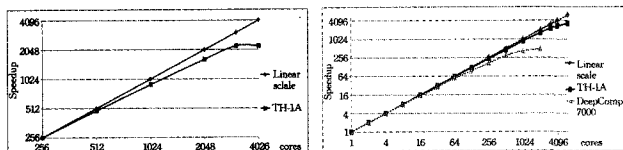


图 2 药物所搭建体系在 TH-1A 上的测试结果

图 3 STMV 在 TH-1A 上的测试结果

3.4 磁约束聚变数值模拟

磁约束聚变大规模数值模拟软件“回旋环等离子体代码”(GTC),是磁约束核聚变领域的重要应用软件,主要用于研究磁约束溶解等离子体湍流传输问题,是托卡马克溶解实验设备的自然配置结果^[7]。北京大学、中国科学技术大学、浙江大学、核工业西南物理研究院、中科院等离子体物理研究所等国内承担磁约束聚变模拟研究的单位共同参与 GTC 合作开发,用于研究聚变实验的物理过程。

GTC 是能够充分利用超级计算机整机性能的少数计算机模拟程序之一,因此被选为美国能源部采购超级计算机及评估超级计算机性能的基准程序,是目前为止能够充分发挥“天河一号”整体性能的唯一磁约束聚变模拟代码。如图 4 所示,GTC 在“天河一号”上采用的最大规模达到 4096 个结点(共 49152 核),计算速度和网络通信性能与美国能源部橡树岭国家实验室的超级计算机“美洲豹”(Jaguar)进行了比较测试,性能提升接近 3 倍,在 3000 核以内计算规模保持了更好的线性加速比,最高每秒模拟的粒子步数达到了 10 的 10 次方^[8]。

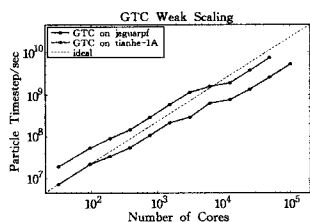


图 4 GTC 在 TH-1A 上的测试结果

3.5 湍流数值模拟

湍流是自然界和生产生活中普遍存在的物理现象,由于其强非线性、对外界条件的敏感性、不规则性以及不可预测性,使得对湍流的研究既复杂又具有重要意义。湍流研究与国防建设和国民经济中的能源、化工、航空、造船、环境保护等领域密切相关,掌握湍流结构的运动规律并进行有效控制和合理运用,对基础研究和实际工程应用将产生重大的推动作用。

北京大学湍流与复杂体系国家实验室和信息学院与国家超级计算天津中心合作,在“天河一号”上进行大规模湍流数值模拟,最大计算规模达到全系统 7168 个 CPU+GPU 结点,进行 14336 立方数据规模的湍流数值模拟,其流动参数已经与自然界中的实际湍流相当。这一模拟计算结果能够直接检验目前湍流理论和经验在实际流动中的合理性和适用性^[9]。

在模拟测试过程中,北京大学湍流与复杂体系国家实验室基于“天河一号”开发了 GPU 加速的 FFT 求解器,可以达

到计算能力为 30TFlops 的直接湍流数值模拟。该 FFT 求解器的性能优于其它函数库的 FFT 求解器,见图 5、图 6。

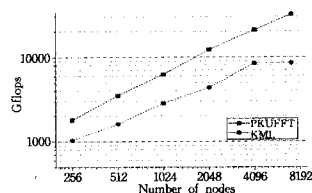


图 5 FFT 与 Intel MKL 扩展性比较

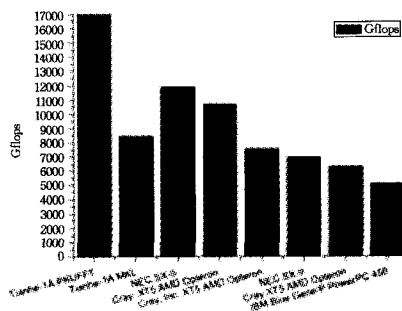


图 6 不同机器的 FFT 求解计算性能测试

3.6 晶体硅分子动力学模拟

中科院过程工程研究所(简称过程所)与国家超级计算天津中心合作,采用“天河一号”全系统 7168 个 GPU 进行晶体硅的分子动力学模拟(MD)模拟,模拟太阳能面板、半导体行业中常用的晶体硅的分子结构。

该模拟计算取得了单精度为 1.87PFlops 的峰值性能。模拟体系含有 1101 亿个原子,而为了模拟 0.116 纳秒物理时间内的系统演化,整套模拟系统计算了大约 50 万的时间步长,每一步计算时间为 25 微妙,总共计算了大概 3 个小时。通过反复验证,确认模拟结果与真实材料完全相符。在此之前,同类模拟计算的最高纪录是 490 亿个原子、每秒 369 万亿次的浮点运算能力^[10]。

在分子模拟测试中,过程采用了自主开发的多体键序作用势(BOP)程序,图 7 描述了不同结点规模的 BOP 程序的实测性能。图 8 描述了不同结点的耦合计算性能(CPU+GPU)。采用“天河一号”全系统 7168 个 GPU 时,浮点计算性能达到了 1.87PFlops,取得了良好的单机效率和并行加速比。

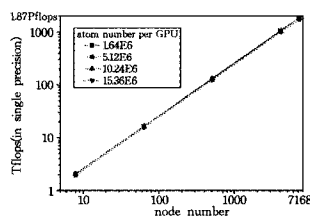


图 7 不同结点的 BOP 算法的性能

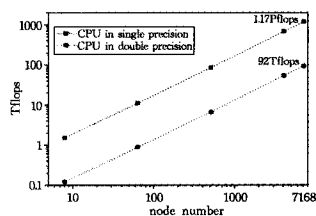


图 8 不同结点的耦合计算性能 (CPU+GPU)

3.7 全球大气浅水波全隐式数值模拟

全球气候数值模拟耦合了大气、海洋、陆地、海冰等 4 个模式,其中大气模式计算占主要计算时间。浅水波方程属于双曲守恒型偏微分方程组,是最基本的大气动力方程,研究其可扩展算法十分重要。目前国际上浅水波方程的数值模拟计算规模仅能扩展至数千 CPU 核左右。

(下转第 282 页)

基序查找算法[J]. 小型微型计算机系统, 2007, 28(11): 1963-1967

- [4] Wang C, Parthasarathy S. Parallel Algorithms for Mining Frequent Structural Motifs in Scientific Data[C]//Proc. of 18th annual international conference on Supercomputing, Malo, France, ACM ICS04, 2004: 31-40
- [5] Maddimsetty R P, Buhler J. Accelerator Design for Protein Sequence HMM Search[C]//Proc. of 20th Annual International Conference on Supercomputing, ACM ICS06, Cairns, Queensland, Australia, 2006: 288-296
- [6] Sandve G K, Nedland M. Accelerating Motif Discovery: Motif Matching on Parallel Hardware[C]//Proc. of WABI 2006. LNBI 4175, Zurich, Switzerland, 2006: 197-206

- [7] 沈一飞, 陈国良, 张强峰. 基于纳米计算结构上的生物序列模式发现算法[J]. 小型微型计算机系统, 2007, 28(4): 635-639
- [8] 李锦, 钟诚, 卿海军. 基于分群策略的机群系统上 Motif 发现并行算法[J]. 计算机工程与科学, 2008, 30(11): 126-128
- [9] 钟诚, 李显宁. 异构机群系统上带返回信息的可分负载多轮调度算法[J]. 计算机研究与发展, 2008, 45(S1): 99-104
- [10] 钟诚, 范大娟. 异构机群系统上基于多轮分配方式的近似串匹配并行算法[J]. 计算机研究与发展, 2008, 45(S1): 105-112
- [11] Wang Jian-xin, Yang De. UPNT: Uniform Projection and Neighborhood Thresholding method for motif discovery [J]. Bioinformatics Research and Applications, 2008, 4(1): 96-106
- [12] Buhler J, Tompa M. Finding Motifs using random projections [J]. J Comput Biol, 2002, 9(2): 225-242

(上接第 267 页)

中科院软件所与国家超级计算天津中心合作, 在“天河一号”上测试其自主开发的全隐式数值模拟程序, 通过使用尽量多的 CPU 核(最多至 82944 核)来进行强/弱可扩展性测试, 并尝试了数百亿网格未知数的问题规模。该测试采用 Williamson 标准测试集中的 Isolated mountain 算例。在强可扩展性测试中, 计算网格数目不变, 随着 CPU 核数增多, 求解时间相应减少, 图 9 给出了强可扩展性的测试结果, 程序从 4608 核扩展至 82944 核, 并行效率约为 60%。在弱可扩展性测试中, 固定每个 CPU 上的计算网格规模以及时间步长; 如果采用传统显式算法, 受稳定性条件的限制, 若 CPU 核数增加, 则总网格规模相应增加, 由于时间步长无法固定而造成计算步数增多, 导致计算时间随之增长; 但采用全隐式方法后, 随着 CPU 核数由 36 增至 57600(1600 倍), 计算步数无需增长, 但由于算法复杂度的关系, 计算时间缓慢增长(约 6.3 倍), 见图 10^[11]。

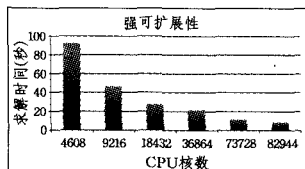


图 9 强可扩展测试结果

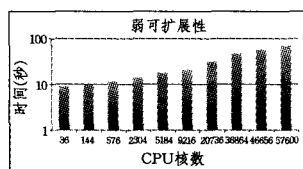


图 10 弱可扩展测试结果

3.8 地球外核热流动数值模拟

地球外核普遍被认为是由高热的液态金属构成的, 地磁发电机理论认为地球磁场由地球外核的热对流运动而产生。研究地球外核的热流动, 有助于加深对地球以及磁场形成的了解, 并有助于地外行星的探索。地球外核的热流动过程, 可以采用快速旋转球壳内的不可压 Navier-Stokes 方程来描述, 由于其非线性强、物理参数极端, 数值模拟有极大难度。

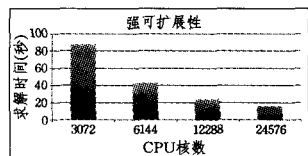


图 11 强可扩展测试结果

中科院软件所与国家超级计算天津中心合作, 在“天河一号”上对其自主开发的地球外核热流动数值模拟程序进行强可扩展性测试, 见图 11。测试采用 Christensen 测试集中的 0 号算例, 拟定常状态时的模拟结果。程序从 3072 核扩展至 24576 核, 并行效率达到 87%^[12]。

结束语 本文通过对石油地震数据处理、飞行器流场模拟、生物分子动力学模拟、磁约束聚变数值模拟、湍流数值模拟、晶体硅分子动力学模拟、全球大气浅水波全隐式数值模拟、地球外核热流动数值模拟等应用进行了测试。测试结果表明, “天河一号”系统运行稳定, 可用性高、可扩展性强, 网络性能优于目前商用主流互连网络系统, 全系统整体性能具有一定优势。

下一步工作在于, 开展通用大规模科学计算软件的移植与优化, 充分发掘“天河一号”性能, 支持重点高性能计算应用领域; 基于“天河一号”CPU 与 GPU 异构协同并行架构, 面向重要科学计算应用领域, 与合作者共同开展 GPU 应用程序开发和 CPU+GPU 异构并行算法设计研究, 探索 GPU 以及 CPU+GPU 协同并行在高性能计算应用领域的发展趋势。

参考文献

- [1] BGP INC. China National Petroleum Corporation, Tianhe-1A" GeoEast-lighting Application Report
- [2] SINOPEC Geophysical Research Institute. "Tianhe-1A" iCluster Application Report
- [3] Institute of Mechanics, Chinese Academy of Sciences. "Tianhe-1A" Aircraft Flow Field Simulation CCFD software Application-Report
- [4] Shanghai Institute of Materia Medica, Chinese Academy of Sciences. "Tianhe-1A" Molecular Dynamic Simulation Application Report
- [5] <http://www.ks.uiuc.edu/Research/STMV/>
- [6] Liu Q, Jin Z, He P, et al. Performance Analysis of Large-Scale NAMD on DeepComp 7000. e-Science, 2010, 11: 83-89
- [7] Ethier S, Tang W M, Lin Z. Gyrokinetic particle-cell simulations of plasma microturbulence on advanced computing platforms [J]. J. Phys. : Conf. Series, 2005, 16
- [8] GTC Home Page[EB/OL]. <http://phoenix.ps.uci.edu/GTC/>
- [9] College of Engineering PEKING UNIVERSITY. "Tianhe-1A" Numerical Simulation of Turbulent Flow Application Report
- [10] Ge W, Hou C, Xu J, et al. Trans-scale Simulation of Silicon Deposition Process on Tianhe-1A. Submitted to Supercomputing, Seattle, USA, Nov. 2011
- [11] Institute of Software, Chinese Academy of Sciences "Tianhe-1A" Fully Implicit Simulation of Global Atmospheric Shallow Wave Application Report
- [12] Institute of Software, Chinese Academy of Sciences "Tianhe-1A" Heat Flow Simulation of Earth Outer Core Application Report