软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2011,22(Suppl.(2)):157−162

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

数万核上复杂应用程序的性能测试与分析*

高兴誉[1,2+], 曹小林[1,2], 赵伟波[1,2], 张爱清[1,2], 莫则尧[1,2]

[1](北京应用物理与计算数学研究所 高性能计算中心,北京 100094)

[2](北京应用物理与计算数学研究所 计算物理国防科技重点实验室,北京 100094)

Performance Study of Complex Application Programs on Tens of Thousands Cores

GAO Xing-Yu[1+], CAO Xiao-Lin[1,2], ZHAO Wei-Bo[1,2], ZHANG Ai-Qing[1,2], MO Ze-Yao[1,2]

[1](High Performance Computing Center, Institute of Applied Physics and Computational Mathematics, Beijing 100094, China)

[2](Key Laboratory of Science and Technology for National Defense, Institute of Applied Physics and Computational Mathematics, Beijing 100094, China)

+ Corresponding author: E-mail: gao_xingyu@iapcm.ac.cn

Gao XY, Cao XL, Zhao WB, Zhang AQ, Mo ZY. Performance study of complex application programs on tens of thousands cores. ***Journal of Software*, 2011,22(Suppl.(2)):157−162.** http://www.jos.org.cn/1000-9825/11036.htm

Abstract: Oriented to the large-scale computation on tens of thousands cores, the parallel software infrastructure named JASMIN has released a new version which has improved the enabling techniques and numerical algorithms. With downward compatible programming interfaces, the new version can enhance the scalability of the programs free of application users' effort. To investigate the scalability of the programs based on JASMIN, we test and analyze the performance of five complex application programs on tens of thousands cores of TH-1A supercomputer. These programs were developed for the high-performance computation arising from inertial confinement fusion, material science as well as the high-power microwave. It is shown that four programs achieve a parallel efficiency of over 60% on 42,000 cores and three ones achieve a parallel efficiency of over 45% on 84,000 cores.

Key words: TH-1A supercomputer; tens of thousands cores; complex application; JASMIN (J Adaptive Structured Meshes applications INfrastructure)

摘 要: 面向数万核大规模计算,JASMIN 框架在使能技术和数值算法上进行了发展和完善,推出了新版本.新版 JASMIN 框架保持编程接口兼容,无需用户修改程序,可直接提升已有程序的并行可扩展能力.为考察应用程序在 JASMIN 框架支撑下的并行可扩展能力,在天河一号 A 超级计算机的数万核上测试和分析了 5 个复杂应用程序的并行性能.这些程序是激光聚变、材料科学、高功率微波研究中最具典型代表性的高性能计算需求.结果表明,4 个应用程序可扩展到 42000 核,并行效率均在 60%以上;3 个应用程序可进一步扩展到 84000 核,并行效率均在 45%以上.

关键词: 天河一号 A;数万核;复杂应用;并行自适应结构网格应用支撑软件框架 JASMIN

* 基金项目: 国家自然科学基金(61033009);国家重点基础研究发展计划(973)(2011CB309702);国家高技术研究发展计划(863)(2010AA012303)

收稿时间: 2011-07-15; 定稿时间: 2011-12-02

高性能计算已深入我国若干重大应用领域的数值模拟中,成为科技创新和提升核心竞争力的重要研究手段[1].在武器物理、激光聚变、材料科学等领域,高性能数值模拟不可或缺[2−4].为形成高置信度的数值模拟能力,物理建模必须科学与精细,计算量随之成千上万倍的增加.为满足需求,我国高性能计算机的峰值性能正以“十年千倍”的速度提升,已部署多台百万亿次和千万亿次计算机[5].2010 年 11 月,国产千万亿次计算机“天河一号A”(以下简称 TH-1A)荣登国际高性能计算机 TOP 500 排名表榜首[5].

随着实际应用的日趋复杂和高性能计算机的迅猛发展,数值模拟应用软件高效使用好计算机越来越难,迫切需要解决计算效率低和研制周期长的两大瓶颈[6,7].并行自适应结构网格应用支撑软件框架 JASMIN (J parallel adaptive structured meshes applications infrastructure)面向高性能科学与工程计算中广泛适用的结构网格数值模拟,致力于解决这两个难题.基于 JASMIN 框架提供的编程接口,用户可以在不熟悉并行程序设计、自适应计算和高性能算法的前提下,在个人计算机上通过串行编程实现问题相关的物理模型、离散格式、数值算法等,多人协同研制适应高性能计算机体系结构的并行应用程序.

为适应数万处理器核(以下简称核)上的大规模计算,2010 年底推出的 JASMIN 框架 2.0 版[8]发展和完善了 3 方面关键技术.第一,与千万亿次计算机五层复杂体系结构“结点-CPU-核-部件-流水线”匹配的五层粒度数据结构“区-块-对象-数组-向量”.其中,子区域(区)被进一步划分为若干逻辑矩形“块”,自适应结构网格并行应用的数据结构以“块”为基本单位.第二,屏蔽可扩展到数万处理器核的 MPI/OpenMP 混合并行编程技术、非规则并行数据通信算法与动态负载平衡方法.这些关键的使能技术均基于“块”:混合并行是“区”的 MPI 并行结合“块”的 OpenMP 并行;非规则通信采用一类新的快速算法构建“块”间通信关系,能支撑数万核并行计算;以“块”为单位的负载调整粒度可灵活地适合应用需求.第三,可扩展到数千上万处理器核的数值并行算法及解法器,包括并行代数多重网格算法、辐射输运通量扫描算法、多层快速多极子算法、快速 Fourier 变换算法等.

新版 JASMIN 框架保持编程接口兼容,无需用户修改程序,可直接提升已有程序的并行可扩展能力.为此,本文以国产千万亿次计算机 TH-1A 为平台,通过测试和分析应用程序在数万核上的并行性能,实际考察复杂应用在 JASMIN 框架支撑下的并行可扩展能力.

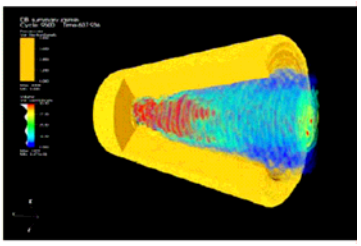
1 基于 JASMIN 框架的复杂应用程序

JASMIN 框架已经成功应用于流体力学、辐射流体力学、弹塑性流体力学、辐射和中子输运、分子动力学、位错动力学、粒子模拟、计算电磁学、以及多物理过程的耦合计算,其中大部分程序能高效使用数百至数千核[9].为考察 JASMIN 框架对数万核上复杂应用的支撑能力,本文采用激光聚变[10]、材料科学、高功率微波研究中 5 个典型的数值模拟应用作为测试程序.见表 1,它们具有多物理、多尺度、高温高压、复杂三维构型等复杂性特征,计算规模可达“十亿网格、百亿粒子”,属于最具典型代表性的高性能计算需求.

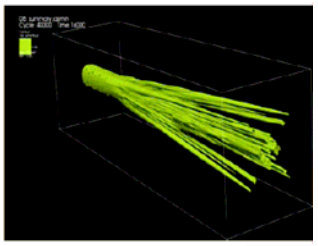
Table 1 Summary of five selected complex applications
表 1 5 个复杂应用程序简介

程序名称	应用邻域	物理模型	计算规模
LARED-P	激光聚变	激光等离子体相互作用粒子模拟,研究强激光与锥结构靶相互作用中光束的聚焦特征和高能电子能谱分布	200 亿粒子,7.68 亿网格,1 万时间步
LAP3D	激光聚变	模拟三维激光光束的传播和成丝过程,研究不同相干手段对成丝的抑制作用	21 亿网格,4 万时间步
LARED-S	激光聚变	ICF 点火靶丸内爆阻滞阶段流体力学不稳定性多尺度模拟	1.6 亿粒子,77 万时间步
MD3D	材料科学	聚变材料短程分子动力学模拟,研究孔洞塌缩的微观机理	2.56 亿粒子,3 万时间步
FDTD3D	高功率微波	模拟计算机机箱对电磁波耦合、散射的全过程	6 亿网格,25 万时间步

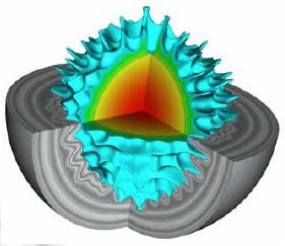
在 JASMIN 框架的混合并行、断点续算、并行 I/O 等功能的支撑下,这些应用已经在作者单位的国产百万亿次计算机上实现了上万核的长时间模拟.如图 1 所示,得到的可视化分析结果与现有的理论认识完全一致.



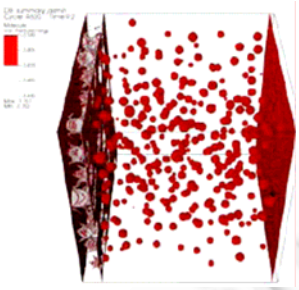
(a) LARED-P 模拟强激光与锥结构靶相互作用



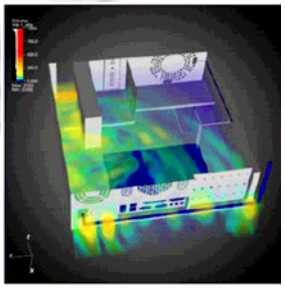
(b) LAP3D 模拟激光光束的聚焦与成丝



(c) LARED-S 模拟 ICF 点火靶丸内爆阻滞阶段流体力学不稳定性



(d) MD3D 模拟聚变材料在冲击波下的孔洞塌缩



(e) FDTD3D 模拟机箱与电磁波耦合、散射的时域场

图 1 复杂应用模拟结果

2 TH-1A 数万核上的性能测试与分析

我们以表 1 中 5 个典型应用为测试模型并取定问题规模,见表 2.在并行可扩展性测试中,通过增加处理器核数测得并行效率,考察 JASMIN 框架支撑数万核上复杂应用的能力.在此基础上,我们结合应用程序和计算机体系结构的特点分析测试结果,为进一步优化提供参考.

Table 2 Summary of five benchmarks for parallel scalability tests

表 2 5 个并行可扩展性测试模型

程序名称	应用邻域	测试模型规模
LARED-P	激光聚变	766 亿粒子,7.68 亿网格,10 个时间步
LAP3D	激光聚变	3 亿网格,10 个时间步
LARED-S	激光聚变	3 亿网格,100 个时间步
MD3D	材料科学	5.12 亿粒子,10 个时间步
FDTD3D	高功率微波	16.8 亿网格,20 个时间步

2.1 线程并行可扩展性测试

大量实验表明,万个以上 MPI 进程的稳定运行时间短(平均在 2 小时以下),难以满足实际应用需求,需要考虑 MPI 进程和 OpenMP 线程的混合并行.因此,一个至关重要的问题是,应用程序必须在 OpenMP 线程级实现可扩展.在 JASMIN 框架的支撑下,应用程序可以自动获得线程可扩展能力.

TH-1A 的计算结点集成了 2 个 6 核 CPU,每个结点拥有 12 个物理线程.据此,我们设计如下的线程可扩展性测试方案:问题规模固定,固定 MPI 进程数为 7 000,每个进程启动的线程数从 1、2、4、6,…,12,最大并行规模达到 84 000 核.

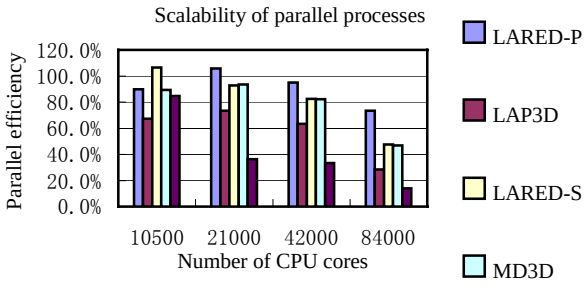


图 2 线程并行效率柱状图

图 2 展示了相对 7000 核的并行效率,5 个应用程序的 6 线程并行效率均在 38%以上,3 个程序的 12 线程并行效率在 20%以上,其中 LARED-P 的并行效率为 37.4%.图 2 只是初步的测试结果,JASMIN 框架还未优化线程并行:线程并行部分的比例可进一步提高;线程间负载可进一步平衡.尽管如此,这样的结果充分表明 JASMIN 框架提供的线程并行功能是高效的,适合 TH-1A 计算机多核体系结构.

2.2 进程扩展性测试

在进程可扩展性测试中,问题规模固定,每个进程启动 6 个线程,进程数从 875、1 750、3 500 增长到 7 000.出于对应用程序运行稳定性的考虑,最大并行规模使用 7 000 个进程,每个进程启动 12 个线程.图 3 展示了相对 7 000 核的并行效率.在 TH-1A 上,4 个应用程序可以扩展到 42 000 核,并行效率均在 60%以上,3 个应用程序通过线程并行进一步扩展到 84 000 核,并行效率均在 45%以上.

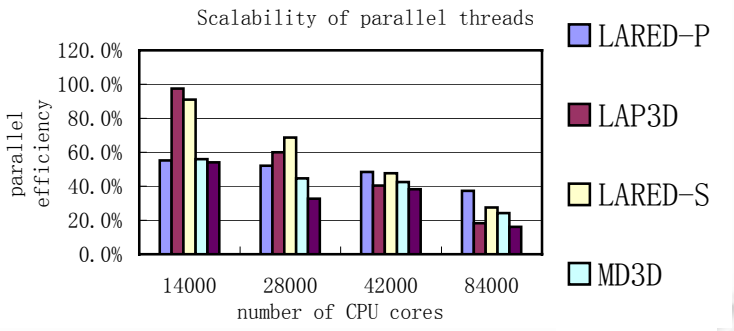


图 3 进程并行效率柱状图

注意到,处理器核数增加一倍,结点数相应地增长一倍,84 000 核的并行任务分布在 7 000 个结点上.此时,对于计算通信比较小的应用程序 LAP3D 和 FDTD3D,结点间通信带宽是并行可扩展的主要瓶颈,从而导致它们在 84 000 核上的扩展性不理想.具体讲,LAP3D 在每个时间步实施并行 3 维 FFT,其中 all-to-all 的通信带宽与结点数平方成反比;FDTD3D 数据量远大于计算量,扩展时通信带宽效应逐渐显现.此外,在 FDTD3D 程序的进程可扩展性测试中,21 000 核的运行时间出现增长,这很可能是当时存在个别结点运行较慢所致.

2.3 计算机体系结构对并行可扩展性的影响

高性能优化需要从计算机体系结构特征来理解测试结果.为此,我们分别在 TH-1A 和国产某百万亿次计算机上测试 LARED-P 的并行性能.这两台计算机不仅计算结点相同,而且体系结构也非常相似,均采用高交换密度的片间互连网络和光电混合的两级胖树结构.主要不同在于这台百万亿次机采用更多的通信光缆,从而显著提高了折半带宽.

在测试中,计算规模固定,每个进程启动 6 个线程,进程数从 750、1 500、3 000,…,6 000.从表 3 列出的测试结果不难看出,4 500 核的并行执行时间接近;扩展超过 9 000 核时,LARED-P 在这两台计算机上的并行性能表

现出明显差异.注意到一个机柜包含 512 个计算结点、每个结点集成 2 个 6 核 CPU,那么从互连网拓扑结构的折半带宽不难理解这样的测试结果:当扩展超过 6 144 核时,柜间结点的通信带宽差异是引起应用程序在两台计算机上并行性能差异的主要因素.

Table 3 Comparison test of LARED-P on between TH-1A and some domestic MPP computer

表 3 LARED-P 在 TH-1A 和国产某 MPP 计算机上的比较测试

任务数	线程数	处理器核数	TH-1A 上的运行时间(秒)	TH-1A 上的并行效率 (%)	国产某 MPP 机上的运行时间(秒)	国产某 MPP 机上的并行效率 (%)
750	6	4 500	138.85	100.0	132.11	100.0
1 500	6	9 000	109.25	63.5	73.44	89.9
3 000	6	18 000	81.85	42.4	46.02	71.8
6 000	6	36 000	43.62	39.8	23.51	70.2

3 总 结

复杂应用和复杂计算机体系结构对应用程序研制提出了巨大挑战.集成共性研制支撑软件框架、基于框架研发并行应用软件是应对这种挑战的新途径.在 TH-1A 数万核上的测试表明,无需用户修改程序,JASMIN 框架在大规模并行计算关键技术上的突破直接推动了已有复杂应用程序的并行性能从数千核提升到数万核.

测试和分析发现,网络间通信带宽可能成为大规模并行程序的性能瓶颈,需结合应用程序和计算机体系结构的特点具体分析.作为一个不断发展和完善的支撑软件框架,JASMIN 将进一步优化多线程并行,从而更好地平衡网络通信带宽与访存带宽.

致谢 感谢北京应用物理与计算数学研究所 JASMIN 框架开发团队和所有应用用户.

References:

[1] Benioff MR, Lazowska ED. Computational science: Ensuring America's competitiveness. Technical Report, Reports of President's Information Technology Advisory Committee (PITAC), Arlington VA, 2005.

[2] LLNL computation directorate annual report 2007. Technical Report, UCRL-TR-402150, Lawrence Livermore National Laboratory, 2007.

[3] Dimitri FK. Advanced simulation & computing: The next ten years. NNSA Defense Programs, 2009.

[4] Bader DA, Wrote; Du ZH, Trans. Petascale Computing: Algorithms and applications. Beijing: Tsinghua University Press, 2008.

[5] Dongarra J, ed. The international top 500 list for computer. http://www.top500.org

[6] Post DE, Votta LG. Computational science demands a new paradigm. Physics Today, 2005,58(10):35−41.

[7] Sarkar V, Harrod W, Snavelg AZ. Software challenges in extreme scale systems. Journal of Physics: Conf. Series, 2009, 180(1):012045. [doi:10.1088/1742-6596/180/1/012045]

[8] Mo ZY, Zhang AQ, eds. User's guide for JASMIN (2.0 version). T09-JMJL-01, Beijing: Institute of Applied Physics and Computational Physics, 2011 (in Chinese).

[9] Mo ZY, *et. al*. JASMIN: A parallel software infrastructure for scientific computing. Front. Comput. Sci. China, 2010,4(4):480−488.

[10] Pei WB. The construction of simulation algorithms for laser fusion. Comm. Comput. Phys., 2007,2(2):255−270.

附中文参考文献:

[4] Bader DA,著;都志辉,译.面向千万亿次计算的算法和应用.北京:清华大学出版社,2008.

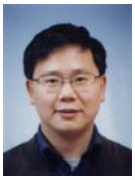
[8] 莫则尧,张爱清,主编.并行自适应结构网格应用支撑软件框架(JASMIN 2.0 版)用户指南.T09-JMJL-01,北京:北京应用物理与计算数学研究所,2011.

高兴誉(1981－),男,江苏靖江人,博士,助理研究员,主要研究领域为特征值并行算法,第一原理材料计算.

曹小林(1974－),男,博士,博士生导师,主要研究领域为动态负载平衡,大规模粒子模拟.

赵伟波(1982－),男,博士,主要研究领域为非结构网格并行算法.

张爱清(1976－),女,博士,副研究员,主要研究领域为并行支撑软件框架.



莫则尧(1971－),男,博士,研究员,主要研究领域为大规模并行快速算法,并行支撑软件框架.