



中国科学院软件研究所

杰出青年人才发展专项计划年度进展报告书

(2011 年度)

姓名	张云泉	课题名称	并行算法与并行软件
资助金额	160.00 万元	课题起止时间	2009. 9-2013. 9
资助类别	应用基础	所在部门	并行软件与计算科学实验室

研究工作主要进展和阶段性成果（含论文、研究生培养等）

2011 年，在所基础青年人才发展专项计划的支持下，开展了如下研究工作：

1. 主持的国家 863 子课题，开展地球外核热流动的大规模可扩展并行数值模拟软件平台的研究，在国产千万亿次平台天河一号 A 上实现了超过数万处理器核的可扩展性：从 3072 至 24576 核，并行效率达到 87%，且成功尝试了 120 亿网格的运算规模，超过课题验收指标一个数量级，受到 863 重点项目专家组的好评。

2. 在核高基国家重大专项（软件类）的资助下，作为课题负责人，带领团队研制成功针对龙芯 3 多核 CPU 的新一代国产高性能扩展数学库 CLeXML V1.0 版 和针对 X86 平台的国产高性能扩展数学库 CASIML V1.0 版。

两个软件包都包含 BLAS、LAPACK、FFT、直接解法器和迭代解法等 5 个模块，采用了自适应性能优化、地址交错和多核并行化等技术进行性能优化。已经申请计算机软件著作权登记：X86 CPU 多核高性能数学库软件（登记号：2011SR092896）并与 Intel 公司开发的 MKL 10.1 版做了对比测试。在 X86 平台，CASIML 库 5 个模块单线程和 8 线程相对于 Intel 公司开发的 MKL 10.1 版本的平均加速比达到了 1.08 和 1.46。CLeXML 是我国自主开发的一款面向国产处理器的多核版高性能数学库，拥有自主知识产权、掌握源代码，且与 Intel 公司开发的 MKL 数学库性能相当。该数学库的开发，打破国外高性能计算机厂商在高性能数学库方面的长期垄断，填补国内无国产高性能数学库的空白，也大大降低了我国国产 CPU 的普及和应用难度。

3. CPU 上并行算法优化研究进展

- 1) 基于 RAM(h)层次存储计算模型的理论分析结果，创新性的提出 CRSD 新稀疏矩阵存储格式。该工作的论文已被 Euro-Par 2011 录取。

SpMV 是求解 PDE 的一个核心算法，根据应用矩阵非零元的对角线式的分布特点创新性地提出了 CRSD 矩阵存储格式。该存储格式针对同一应用对角线格式在不同问题规模下，对角线分布具有相似性进行优化。通过对角线格式表示对角线分布。并通过自适应优化方法基于具体运行平台选择最优的 SpMV 实现。测试结果表明，所提出的新存储格式针对不同问题规模的矩阵，在不同线程下均优于 MKL 中基于 CSR 的 SpMV 实现，测试的性能与加速比结果如下图 6 所示。可以看出 CRSD 在单线程下，相对经典存储格式 CSR 运行效率，加速比平均为 3.05 倍，最高可以到达 4.38 倍；相对于 MKL 的 DIA 存储格式，对于不适合 DIA 存储格式的矩阵，CRSD 同样

有很好的加速效果，如图中编号为 4 和 5 的矩阵，加速比可以达到 90 倍以上，除此之外，其余最高加速比为 2.02，平均加速比为 1.74。在多核情况下，CRSD 的加速效果在相同处理器下，是 CSR 存储格式的 2.16 倍，最高可以达到 4.61 倍。

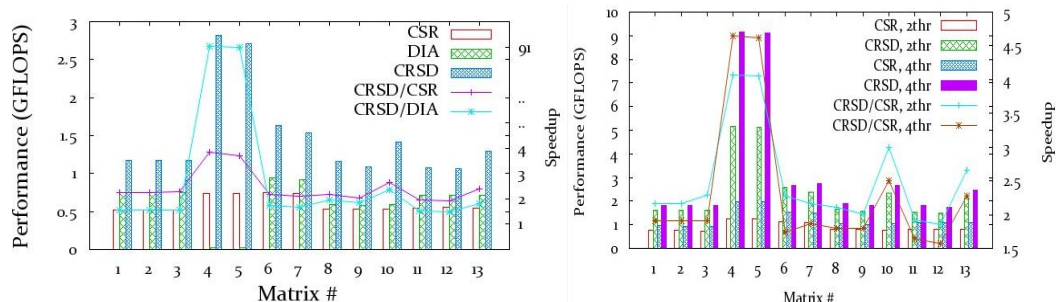


图 1 不同矩阵存储格式的性能比较

- 2) 龙芯 3 上 BLAS 的性能优化工作，相关工作在 HPC China2011 上进行了交流，并被推荐到《软件学报》发表。所研制的 OpenBLAS 软件包已经成为国际开源项目，<https://github.com/xianyi/OpenBLAS>，下载量达到 1676 次，有若干国际知名公司开始资助该项目，逐渐取代国际知名的 GotoBLAS 软件包的位置。

BLAS(Basic Linear Algebra Subprograms)是一个基本线性代数核心子程序集,主要包括向量和矩阵的基本操作。OpenBLAS 是我们在 GotoBLAS 2-1.13 BSD 版基础上发展起来的 BLAS 库开源项目，当前主要工作是为龙芯 3A CPU 提供高性能的 BLAS 库实现。该项目针对龙芯 3A CPU 的结构特点，通过分块选择，手工优化汇编，应用 128bit 访存指令，指令重排等技术对 BLAS 三级函数 GEMM（矩阵乘法）函数进行实现与优化。

```

gsd.QC1(R8,F5,F4,2)
MADD t1,t1,a0,b0
MADD t2,t2,a1,b0

gsd.QC1(R8,F13,F12,2)
MADD t12,t12,a0,b1
MADD t22,t22,a1,b1

```

龙芯3 128bit访存指令

浮点乘加计算指令

图 2 GEMM 核心汇编代码片断

在双精度和双精度复数情况下，OpenBLAS 优势明显，平均性能超过 ATLAS 39.6% 和 27.7%，相对于 GotoBLAS 平均性能提升分别为 109.5%和 107.3%。在单精度和单精度复数下，OpenBLAS 相对于 ATLAS 的优势并不明显，分别只高出 5.0%和 2.0%；相对于 GotoBLAS，性能分别提升 52.0% 和 51.1%。因此，OpenBLAS 在单精度和单精度 GEMM 还具有明显的改进空间，我们推测是在 OpenBLAS 中没有使用单精度浮点向量（ps）指令，造成性能稍低。

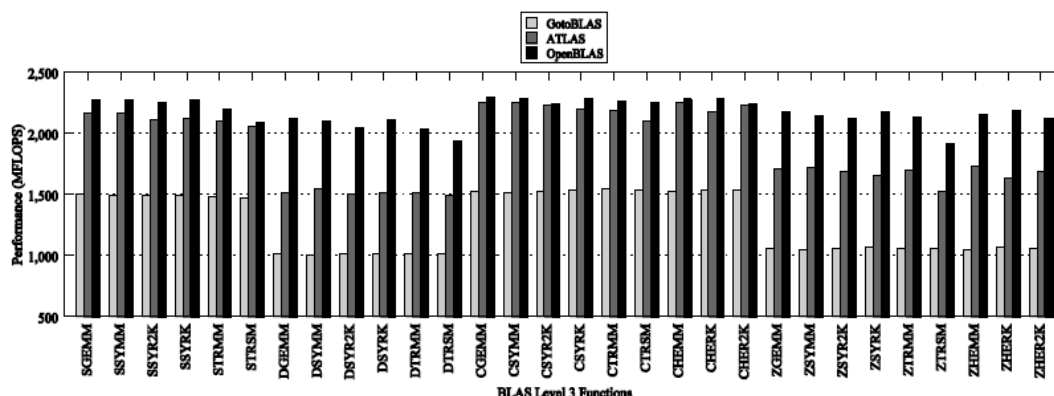


图 3 单线程 BLAS 3 级函数平均性能对比图

在多核情况中，各线程所需计算的子矩阵的总空间约为 3.6M，由于龙芯 L2cache 大小为 4M，采用四路组相连随机替换策略会导致各线程间对共享的 L2 Cache 的争。因此，我们一方面减小多线程下子矩阵的大小来减小对 L2 Cache 的消耗，另一方面，将各个线程使用的数据起址进行交错的排列布局，可尽量避免在 L2 Cache 中各个线程间的相互冲突和替换。

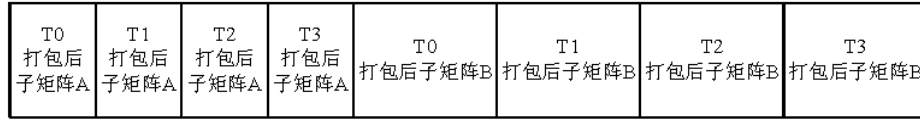


图 4 多线程子矩阵 A 与 B 交错布局示意图

OpenBLAS BLAS 3 级函数的 4 线程并行加速比达到 3.47。4 线程 BLAS 3 级函数平均性能高于 GotoBLAS 和 ATLAS 69%和 34%。。

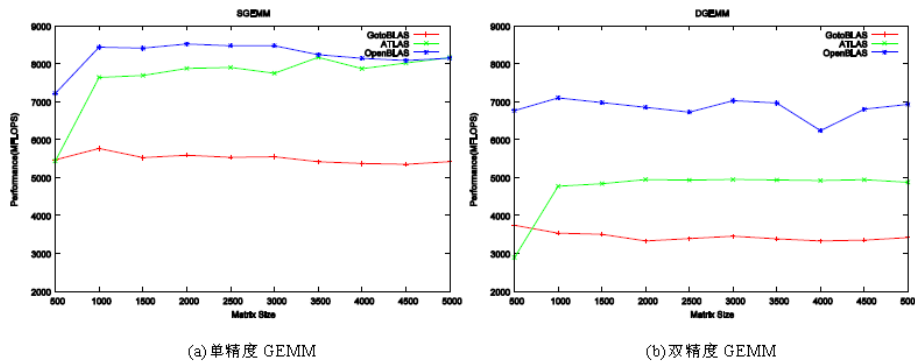
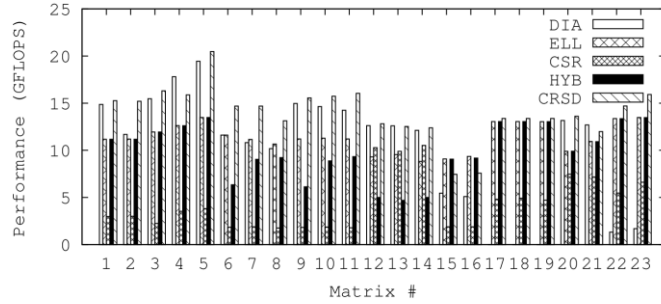


图 5 多线程性能对比效果图

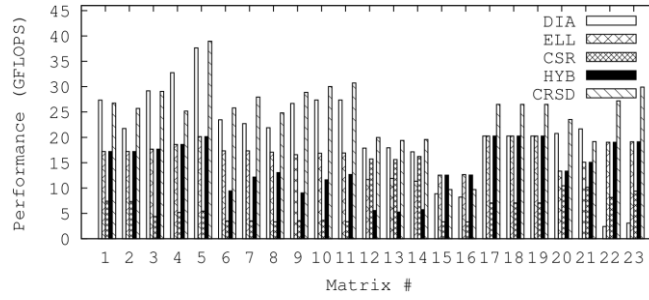
4. GPU 算法优化研究进展

- 1) 基于 CRSD 格式的 SpMV 算法在 GPU 上的性能优化方法和技术研究，相关文章已在 IEEE ICPP 2011 上发表。

在将基于 CRSD 存储结构的 SpMV 算法实现移植到 GPU 上的过程中，根据 GPU 硬件结构特点开展了如下优化技术研究：a)有效利用局部存储器。CRSD 存储结构能够将 SpMV 处理相邻对角线时的公共 x 元素存放到局部存储器中，减少对延迟较大的全局存储器的访问次数。b)将离散点的存储方式由 CPU 上的 CSR 存储格式修改为 ELL 存储格式。由于 GPU 处理 CSR 存储格式的跳转操作时效率很低。c)CRSD 存储格式对矩阵进行分割时，分段大小是 wavefront 的整数倍，每个 work-group 处理一个对角线格式，避免 thread divergence 以及在访问全局内存时能够进行访问合并的优化方法。d)利用代码生成器，在运行时根据矩阵的对角线格式自动生成可执行代码，并采用全局展开等优化策略。与经典 GPU SpMV 实现(Bell and Garland, SC2009) 的四种矩阵存储格式 (DIA、ELL、CSR 及 HYB 四种存储格) 的单双精度进行对比，实验结果如图 6 所示。相对经典实现四种存储格式的最优实现，CRSD 在处理双精度浮点的运行效率加速比为 1.52X，单精度为 1.94。



a) 双精度



b) 单精度

图 6 CRSD 在 GPU 上的性能加速

- 2) GPU 上国产自适应 FFT 软件包 MPFFT 的研制,相关工作发表在 IEEE ICPADS 2011 上。

快速傅里叶变换(Fast Fourier Transform, FFT)是离散傅里叶变换(Discrete Fourier Transform, DFT)的快速算法,具有广泛的应用,常被用于数字滤波、求解偏微分方程和信号分解等。事实上,FFT 已被评选为二十世纪科学和工程界最具影响力的十大算法之一。我们针对目前主流的 GPU 提出了一个新的 FFT 自适应框架,且基于该框架实现了自主知识产权的 FFT 库 MPFFT,如下图所示:

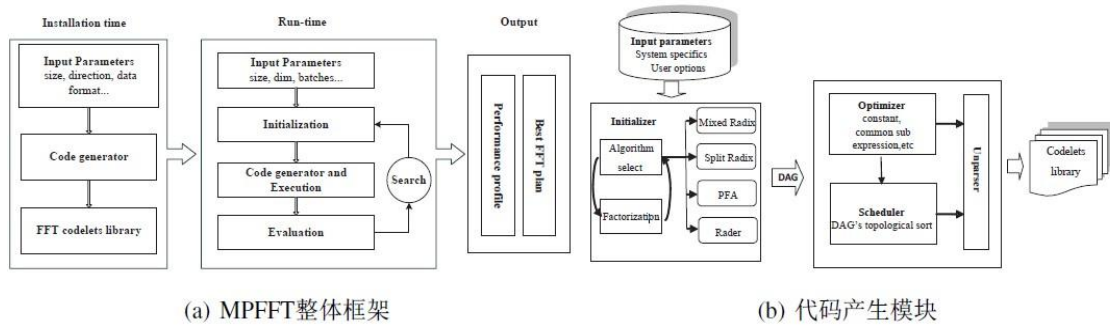


图 7 MPFFT 的整理框架和代码生成模块示意图

首先在程序安装时,只需要借助一个简单的脚本文件,代码生成模块生成可以被 GPU kernel 调用的小因子 FFT 集合(codelets Library); 在程序运行时,获取输入问题的相关参数(DFT 的维度,长度以及数量 batches),并根据这些参数来初始化对性能影响较大的相关硬件参数譬如寄存器大小、LDS 容量和其 bank 数等。其次,运行时模块考虑这些参数和 FFT 算法的特点来对输入问题规模按照不同策略分解,从而生成多个 plan,然后在 Search 模块的监督下进行执行和性能评估。最终,Search 模块输出所有分解策略对应 plan 的性能数据,同时得到最佳 DFT 分解策略,并且生成的 profile 数据对其后问题规模的分解和构造能够提供向导。

我们在不同的 GPU 上评估了 MPFFT 库的性能,为了便于对比,CPU 版本我们

选取 FFTW3.2.2 作为性能基准，而 AMD GPU 和 NVIDIA GPU 上的 FFT 分别选取 clAmdFft 1.4 和 CUFFT 4.0 作为对比。其中一、二、三维的测试结果依次如下图所示：

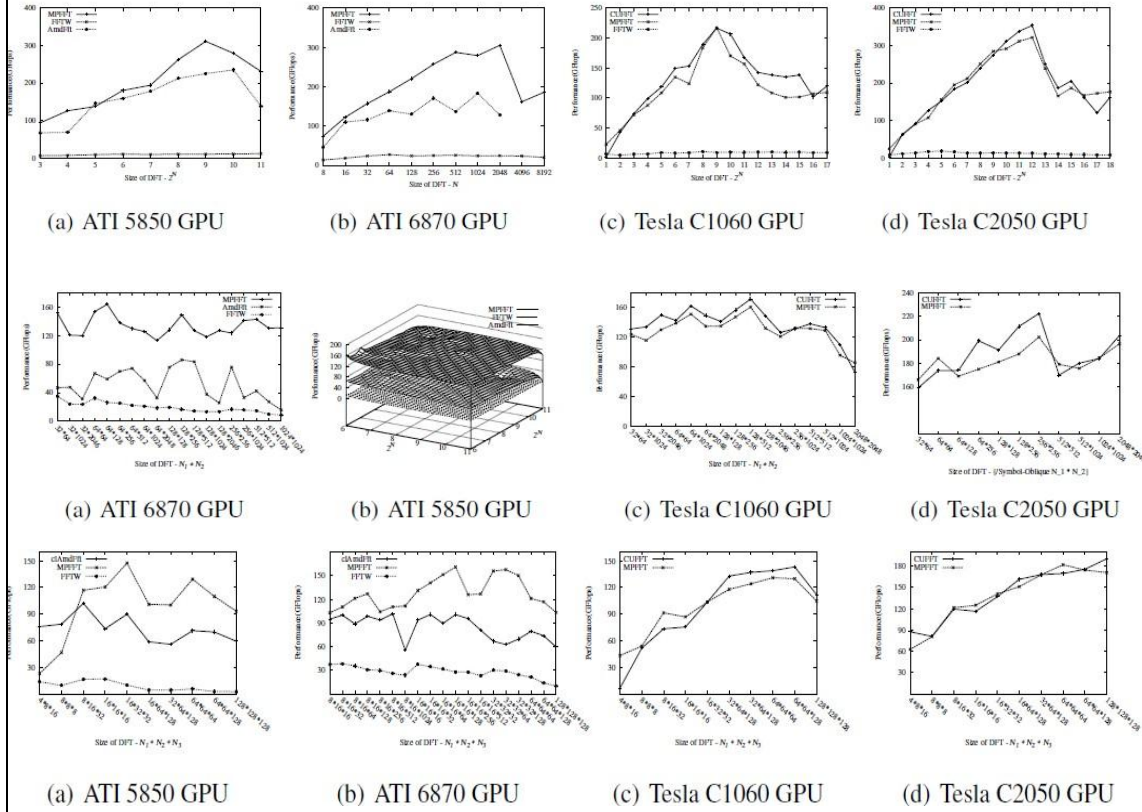


图 8 MPFFT 在不同测试平台上的性能比较

实验结果表明：一维 MPFFT 的性能相对于 clAmdFft 库，取得的加速比为 1.5x - 4x，接近于 CUFFT，是四线程 FFTW 的 6x -18x。二维 FFT 在 ATI 5850 上相对于 clAmdFft 取得的加速比为 1.5x -36.81x，平均加速比为 2.6x。此外，在 ATI 6870 上相对 clAmdFft 平均加速比 2.4x，最大 8.33x，是八线程 FFTW 的 6.52x，在 NVIDIA GPU 上性能均达到了 CUFFT 的 90%。而三维 FFT 的性能，在 ATI 5850 和 ATI 6870 上相对 clAmdFft 的平均加速比分别为 1.81x 和 1.37x，同时，在 NVIDIA GPU 上性能也达到了 CUFFT 性能的 90%。此外，MPFFT 支持输入规模为非 2 的幂，相对 Tesla C2050 和 Tesla C1060 以及四线程 FFTW 的加速比分别为 1.5 x- 28x、20.01x、31.87x，由于 clAmdFft 对非 2 的幂支持不太完善，只支持 $2i \times 3j \times 5k$ ，因此在性能曲线图中没有列出。

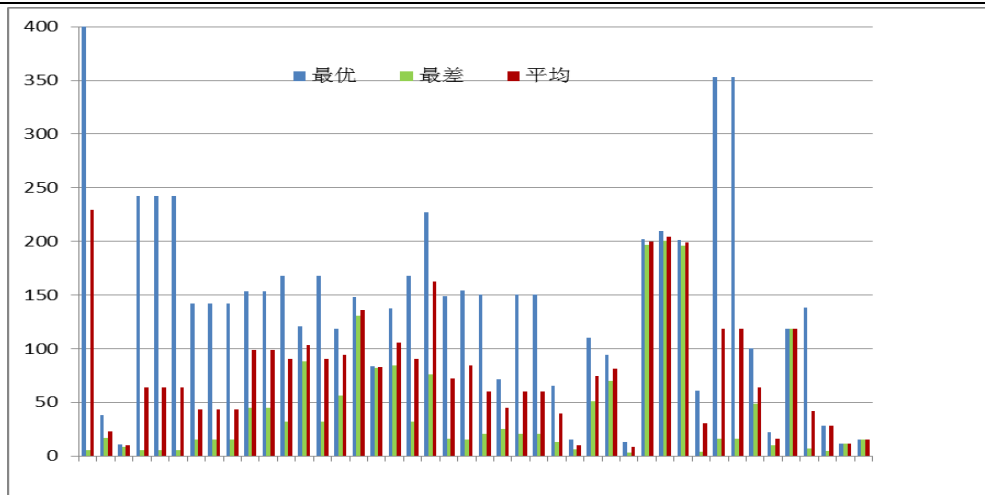


图 10 相对于 CPU 的加速比（横轴表示函数共 55 个）

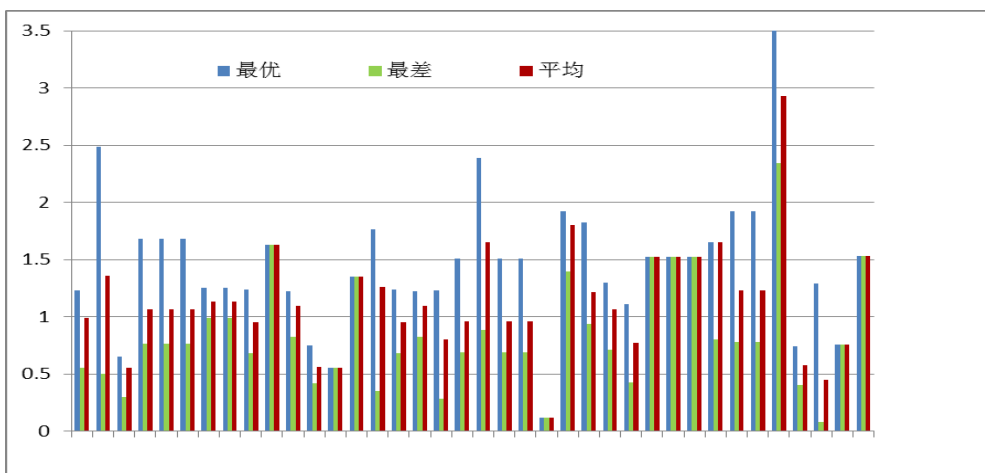


图 11 相对于 CUDA 版本的加速比（横轴表示函数共 55 个）

5. 并行计算模型研究进展

1) 提出一个新的基于横向局部性的多核计算模型 MRAM(h)

RAM(h)模型扩展 RAM 中单层存储层次为多层，但是其存储复杂度只描述了存储层次间的数据重用，并没有考虑多核共享缓存所带来的线程间共享存储的数据共享，我们在多核上对 RAM(h)进行横向扩展，同时考虑数据重用和数据共享。RAM(h)和 MRAM(h)模型的主要目的都是在实现算法时要充分考虑数据所在存储层次，尽量考虑数据重用，以提高算法的性能，MRAM(h)还要求考虑并行计算任务之间的数据共享率，尽量使得共享数据较多的任务分配的更近，也就是在更高的存储层次共享数据，此外，还要保证共享数据的访问时间相离的较近。相关工作发表在 HPC China2011 上。

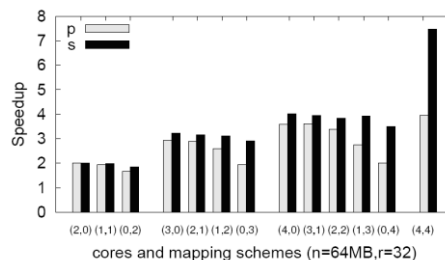
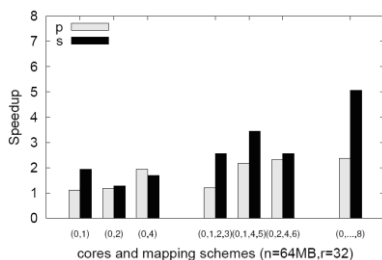


图 12. 平台 1 上共享 LLC 横向局部性

图 13. 平台 2 上共享 LLC 横向局部性

2) 提出一个新的对随机图遍历的局部性预测模型。

大规模图算法最近几年是并行计算中一个研究热点，美国发布了类似 TOP500 Linpack 的 Graph500 测试包和排行榜，对图算法的优化和分析也成为主流会议的热点。我们从图算法中提炼出一种典型模式，包括节点生成序列以及相应节点距离概念，结合基于边分布取随机图连接度因子，设计了一个基于概率的局部性模型。对利用 RMAT 生成的随机图，利用 METIS 获取连接度因子，带入模型进行预测，与实测结果对比如下图所示。相关工作与美国罗切斯特大学丁晨教授等联合撰写“Modeling the Locality in Graph Traversals”英文文章一篇，待发表。

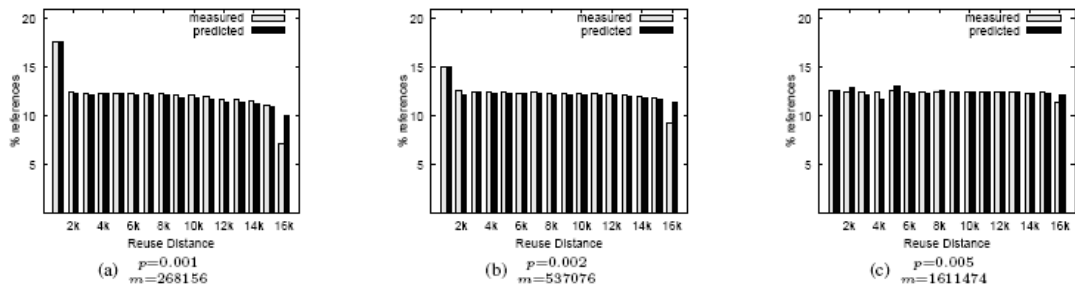


图 14. 局部性预测和实测结果对比

6. 集合通信算法研究进展

开展了利用节点内的通信延迟差异优化 MPI 集合通信的性能得研究。我们设计了一个节点内 NUMA-aware 的集合通信性能分析模型，分别在 Intel 和 AMD 两种多路多核平台上评估了不同进程映射对 MPI 集合通信的影响，依据该分析模型，为各集合通信算法设计了较优的通信模式。试验表明新通信模式可明显提高 MPI 集合通信的性能。如短消息的 MPI_Bcast 在 Intel 和 AMD 平台上可分别提高 76.5% 和 14%。我们通过穷举分析所有进程-核映射模式的方法证实理论分析结果，中短消息 Bcast 性能至少可以提升 31%。相关研究成果与美国阿岗实验室 Pavan Balaji 博士联合撰写文章一篇，并投稿至 IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing(CCGrid2012)。

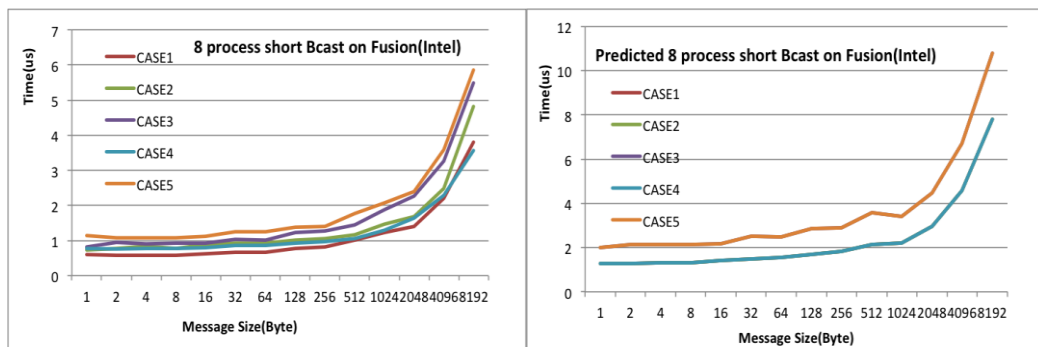


图 16 Intel 平台不同进程映射下 MPI_Bcast()性能 实测（左图）模拟（右图）

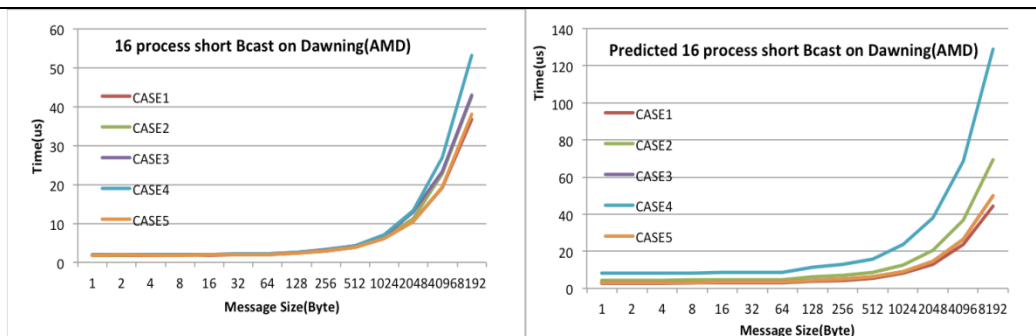


图 17 AMD 平台不同进程映射下 MPI_Bcast()性能实测（左图）模拟（右图）

7. 总结

综上所述，2011 年指导并毕业博士研究生一名,硕士研究生三名,一名硕士生转为博士生。目前正指导硕士研究生十二名，博士研究生六名，联合指导博士研究生两名（中科大，中国海洋大学各一名）。相关工作发表在 EuroPar 2011，IEEE ICPP 2011，IEEE ICPADS 2011 等著名国际会议上，在《软件学报》和《计算机科学》等国内期刊发表文章 6 篇，国内会议发表文章 8 篇，获得软件著作权登记 2 项。

大会特邀报告:

- Zhang Yunquan, ISC 2011, Hamburg, Germany, June 19-23, 2011, Invited Speaker.
- Zhang Yunquan, GPU Technology Conference Asia For 2011, China National Convention Center, Beijing, China, Dec. 14-15, 2011. Invited Speaker.
- 张云泉, 2011 年 7 月 25-27 日, 第七届中国 CAE 工程分析技术年会, 昆明, 云南 (大会特邀报告)
- 张云泉, 2011 年 10 月 26-10 月 29 日, 2011 全国高性能计算学术年会(济南, 大会特邀报告)
- 张云泉, 2011 年 12 月 21-12 月 23 日, 2011 第七届全国高性能算法软件研究开发研讨会 (北京, 大会特邀报告)。
- 张云泉, 2011 年 11 月 6-11 月 9 日, 第一届中国科学院超级计算应用大会(SCA2011), 青岛, 山东.大会特邀报告
- 张云泉, 2011 年 12 月 15-16 日, 第二届中国科研信息化研讨会 (北京, 大会特邀报告)。

2011 年完成和申请课题情况:

- “十一五” 863 计划 “高效能计算机及网络服务环境” 重大项目 “高性能计算与网络应用” 课题, 《天体大规模并行数值计算软件平台的研制》(No.2006AA01A125), 2007.1-2010.12, 课题组副组长, 子课题组组长 ; 顺利通过验收。
- 核高基重大专项 (软件类)《支持国产 CPU 的编译系统及工具链》子课题《龙芯 CPU 多核并行国产高性能数学库研究开发》(2009ZX01036-001-002-3), 2009.1-2011.12, 课题负责人, 正在进行验收工作。
- 国家 863“高效能计算机及网络服务环境”重大专项课题, 《曙光 6000 千万亿次高效能计算机系统研制》子课题 《面向数万个以上处理器的新型基础算法研究》(No. 2009AA01A129), 2009.1-2010.12, 子课题组组长 ; 顺利通过验收。
- 2011 年与湖南大学联合申请到国家自然科学基金重点基金一项, 四年;
- 与美国 AMD 公司联合成立 APU 软件联合实验室的项目通过二期评估, 第二批经费已拨付;

- 与美国阿岗国家实验室成立并行处理联合实验室一个;

发表文章列表:

1. Sun Xiangzheng, Zhang Yunquan, Wang Ting, Long Guoping, Zhang Xianyi, and Li Yan. Crsd: Application Specific Auto-tuning Of Spmv For Diagonal Sparse Matrices. In: Euro-Par 2011 Parallel Processing. 2011. 316-327.
2. Xiangzheng Sun, Yunquan Zhang, Ting Wang, Xianyi Zhang, Liang Yuan, Li Rao: Optimizing SpMV for Diagonal Sparse Matrices on GPU. ICPP 2011: 492-501
3. Yan Li, Yunquan Zhang, Haipeng Jia, Guoping Long and Ke Wang, Automatic FFT Performance Tuning on OpenCL GPUs, IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS 2011). Dec 7, 2011 - Dec 9, 2011, Tainan, Taiwan, China.
4. 费辉 张云泉 王可,许亚武. 基于 GPU 的分子动力学模拟并行化及实现. 计算机科学, 2011, (09): 275-279
5. 解庆春, 张云泉, 王可, 李焱, 许亚武. SIMD 技术与向量数学库研究. 计算机科学, 2011, (07): 298-302
6. 李焱 张云泉 王可,赵美超. 异构平台上基于 OpenCL 的 FFT 实现与优化. 计算机科学, 2011, (08)
7. 颜深根, 张云泉,龙国平,李焱. 基于 OpenCL 的归约算法优化研究. 软件学报, 2011 年增刊
8. 李超,张云泉,郑昌文,胡晓慧, Intensity model with blur effect on GPUs applied to large-scale star simulators,软件学报, 2011 年增刊
9. 张先轶,王茜,张云泉, OpenBLAS:龙芯 3A CPU 的高性能 BLAS 库,软件学报, 2011 年增刊
10. 费辉, 张云泉, 王靖. 基于 GPU 的非标记定量软件 QuantWiz 并行化实现. HPCChina2011.
11. 袁良,张云泉. 基于横向局部性的多核计算模型 . HPC China 2011
12. 张樱,张云泉,龙国平. 基于 OpenCL 的图像模糊化算法优化研究. HPC China 2011
13. 吕渐春, 张云泉, 王婷, 肖玄基, PLASMA 自适应调优与性能优化的设计与实现, HPC China 2011
14. 贾海鹏, 张云泉, 龙国平, 徐建良, 李焱, 基于 OpenCL 的拉普拉斯图像增强算法优化研究, HPC China 2011
15. 颜深根, 张云泉,龙国平,李焱. 基于 OpenCL 的归约算法优化研究. HPC China 2011 (大会优秀论文提名)
16. 李超,张云泉,郑昌文,胡晓慧, Intensity model with blur effect on GPUs applied to large-scale star simulators, HPC China 2011 (大会优秀论文提名)
17. 张先轶,王茜,张云泉, OpenBLAS:龙芯 3A CPU 的高性能 BLAS 库, HPC China 2011 (大会优秀论文)

软件著作权:

- A. 计算机软件著作权登记, 行星流体力学大规模数值模拟并行软件 V1.0, 登记号: 2011SR008456。
- B. 计算机软件著作权登记, X86 CPU 多核高性能数学库软件 CASIML V1.0 版, 登记号:

2011SR092896。

学术活动和兼职：

- 2011 年全国高性能计算学术年会程序委员会副主席
- 第七届中国 CAE 工程分析技术年会专家委员会副主席
- 第一届中国科学院超级计算应用大会专家委员会委员
- 核心期刊《计算机科学》编委
- 中国科学院《科研信息化技术与应用》编委
- 上海超算《高性能计算机发展与应用》编委
- 中国计算机学会《计算机学会通讯》编委
- 中科院超级计算中心用户委员会委员
- 中国计算机学会理事
- 中国软件行业协会常务理事
- 中国计算机学会高性能计算专委会秘书长
- 中国软件行业协会数学软件分会秘书长
- SC2011 SoTP, IPDPS 2012, CCGrid 2012, e-Science 2.0, SC2012 程序委员会委员
- ISC 2012 Steering Committee Member
- 中国传媒大学和华南理工大学兼职教授和博士生导师

下一年度工作计划，包括国内外合作与交流计划

1. 完成CLeXML数学库在龙芯3B平台上的测试，包括正确性、性能和健壮性测试；该项目于明年6月底结题，需要根据工信部的验收细则，按步完成各项工作，确保顺利结题。
2. 继续研究和完善新的基于micro-benchmark的GPU性能指导模型，通过具体应用实例，验证模型的正确性和完整性，根据实验测试结果，找出模型存在的缺点和不足，并加以改进和完善。同时关注具有不规则负载特征的应用在GPU上的实现和优化，并验证将其集成到性能指导模型的可行性。
3. 在OpenCL框架下对更加复杂的通用算法如：Scan、BFS、Sort等进行研究与优化，并争取将现已实现的算法能充分利用OpenCL的跨平台特性，实现不仅是代码的跨平台移植，同时也实现性能的跨平台移植。
4. 调研Data-Parallel算法，以Scan为研究基点，探讨多GPU和多核之间混合并行编程，并初步构建其自适应编程框架。同时调研新的优化策略进行改进以及混合编程模式。此外，调研out-of-card FFT算法，初步建立GPU+CPU计算框架，完善已有的自适应FFT框架。
5. 继续完善基于横向局部性的并行计算模型，从程序角度，利用PIN等工具，进行深入分析，与理论模型比较，检验模型正确性和易用性。
6. 继续完善图论模型，利用模型对更多算法进行局部性预测工作。
7. 拟派出一名博士生到阿贡国家实验室继续合作交流。相关成果撰写高水平学术文章，并投稿著名国际会议。

年度经费使用情况概要

2011 本项目到账 45 万元。支出劳务费约 15 万元，用于学生及临时聘用人员劳务费支出；设备费约 5 万元，用于 PC 机、耗材购买；国际交流与合作费约为 3 万元，用于参加国际会议出访及接待来访交流外宾；其他业务费约为 12 万元，用于差旅费、会议费、版面费等；房屋水电费约 7.5 万元，用于项目人员使用的房屋水电费用；直接管理人员工资约为 2.5 万元。

存在的问题、建议及其他需要说明的情况

备注：1.杰出青年人才入选者应实事求是地填写此报告，禁止弄虚作假；
2.此报告作为专项计划跟踪、管理的主要依据，每年报送人力资源处；
3. 此报告将在所网站对外发布。