



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

(2012 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）

2012 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. 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。

2. GPU 上并行算法优化研究进展

- 1) OpenCV 于 1999 年由 Intel 公司建立，2006 年以后由著名的机器人研究实验室和技术孵化器 Willow Garage 提供支持。它是一个基于 BSD 许可证授权发行的跨平台计算机视觉库,包括图像处理和计算机视觉方面的很多通用算法的实现[3]。OpenCV 有 C、C++和 Python 三个版本，支持 Windows、Linux、Android 和 Mac 系统，源代码中包含大概 500 个函数以及一系列使用这些函数实现的应用。从 2.2 版本开始 OpenCV 库中增添了 gpu 部分，该部分基于 CUDA 实现了对部分函数的 GPU 加速。从 2010 年 12 月 OpenCV2.2 版本发布至今，GPU(CUDA)部分不断有新的函数加入，目前为止大概有 90 个函数。

我们的工作是使用 OpenCL 将 OpenCV(Open Source Computer Vision Library) 库进行加速。目前为止我们已经优化超过 60 个基础函数，两个典型应用 Face detection 和 FFT，相对于 CUDA 版本 OpenCV，我们开发的 OpenCV 版本（OpenCVCL）最高达到了 6.25 倍，平均 1.2 倍 Kernel 加速比。相较于 CPU 版本 OpenCV，最高达到了 410 倍，平均 74 倍 Kernel 加速比。

我们总体上将 OpenCV 库中的函数划分为三类：1) 基本的矩阵运算函数，如

四则运算、逻辑运算、最大值最小值、矩阵求和等；2) 基本图像处理函数，如图像滤波、图像缩放、图像分析等；3) 复杂的图像处理算法，如运动分析与对象跟踪、模式识别等。OpenCV 中基本图像处理操作包括去噪、边缘检测、角点检测、采样与插值、色彩变换、形态学处理、直方图和图像金字塔结构等。目前已经实现的函数中第一类 34 个，第二类 26 个，第三类 2 个。其中基本图像处理函数，我们也将其分为三类：数据无关、数据共享和数据相关。表 1 展示了 17 个典型图像处理函数的分类。

类别	函数名	功能
数据共享	blur	图像模糊
	Sobel	使用 Sobel 算子计算图像差分
	Scharr	使用 Scharr 算子进行图像差分
	GaussianBlur	高斯卷积
	Laplacian	拉普拉斯变换
	erode	腐蚀
	dilate	膨胀
	morphologyEX	基于 erode 和 dilate 的高级形态变换
	cornerHarris	Harris 角点检测
	cornerMinEigenVal	计算梯度矩阵的最小特征值进行角点检测
数据无关	Threshold	对图像元素进行固定阈值操作
	copyMakeBorder	复制图像并制作边界
	warpAffine	对图像做仿射变换
	warpPerspective	对图像进行透视变换
	Resize	图像大小变换
数据相关	integral	计算积分图像
	equalizeHist	灰度图像直方图均衡化

表 1 典型图像处理函数分类（17 个）

数据无关函数 Threshold，Resize 优化：Threshold 函数对单通道图像应用固定阈值操作，将源图像中(x, y)位置的像素点与固定阈值进行比较，根据不同的阈值类型，目标图像(x, y)位置的像素点分别取相应的值。函数的典型应用是对灰度图像进行阈值操作得到二值图像，或者是去掉噪声，例如过滤很小或很大像素值的图像点。优化主要通过访存对齐和向量化写回来进行，下图展示了优化结果。

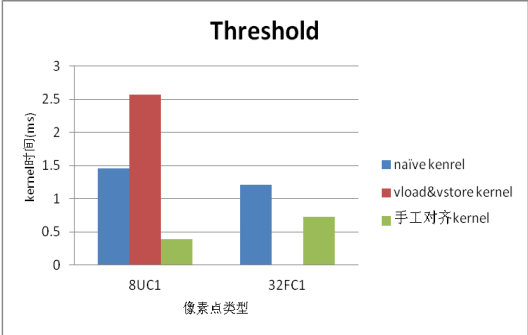
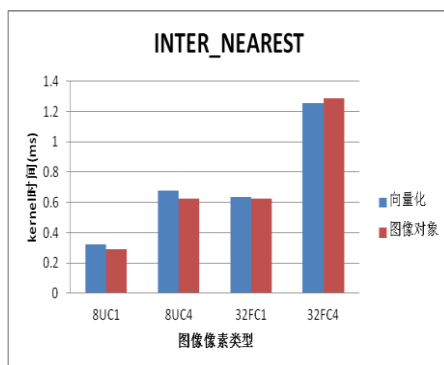
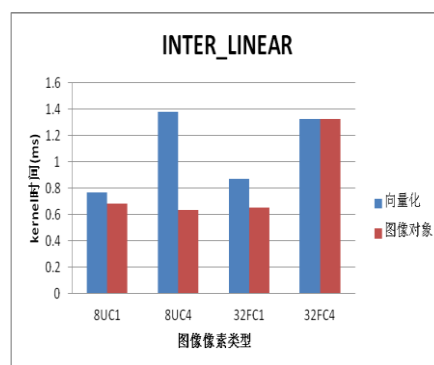


图 3 Threshold 优化结果

Resize 函数对图像进行大小变化，即进行图像缩放操作。缩放操作常用的插值方法有三种：最近邻插值(INTER_NEAREST)、双线性插值(INTER_LINEAR)和立方插值(INTER_CUBIC)。主要通过向量化插值和图像对象插值两种方法进行优化。图 4 和图 5 展示了优化结果。

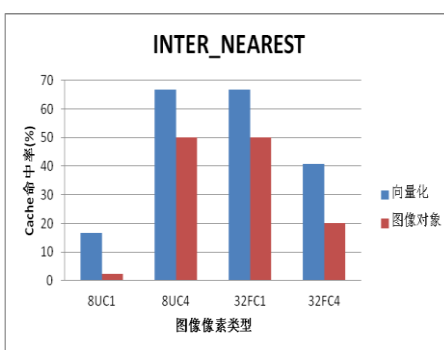


(a)最近邻插值两种方法性能对比

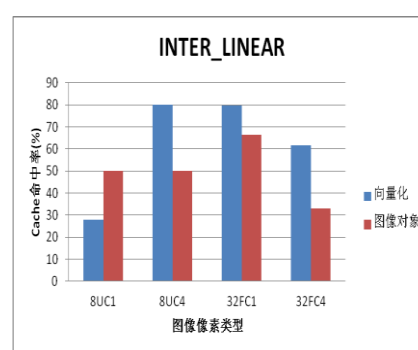


(b)双线性插值两种方法性能对比

图4 向量化和图像对象两种方法性能比较



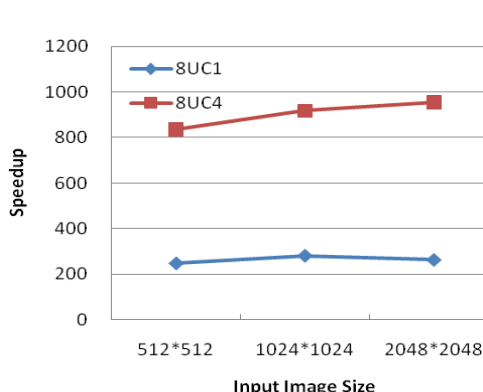
(a)最近邻插值两种方法Cache命中率对比



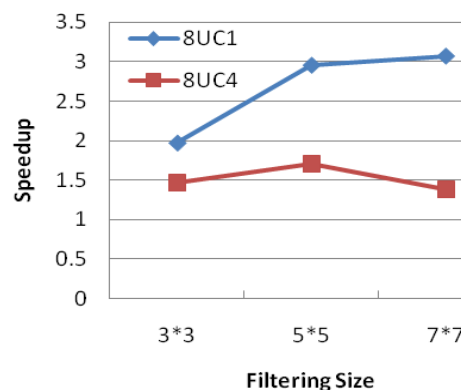
(b)双线性插值两种方法Cache命中率对比

图5 向量化和图像对象两种方法Cache命中率比较

数据共享函数优化：Blur 算法是图像滤波中最基本的一种,其作用是对图像进行模糊处理，算法的基本原理是把图像中某点的值用该点周围一块区域的平均值来替代。主要优化方法为利用 local memory 实现数据重用。



(a)相对 CPU 版 blur 的加速比



(b) 相对 NVIDIA NPP 版 blur 函数的加速比

图 6 Blur 函数加速比

数据相关函数 Integral 优化：Integral 函数用于计算图像积分，我们主要采取了三个方法对该函数进行了优化：（1）对列实施前缀和操作，按行写回存储。（2）一个线

程组计算一整列。(3) 一个线程组同时计算两列。通过这些方法的优化，我们取得了较好的加速比，相对于 CPU 版本函数最多获得了 11.6 倍加速，相对于 CUDA 函数，最佳获得了 1.5 倍加速。

- 2) 基于 GPU 的 Scan 算法的研究。提出了 GPU 上的邻接同步，基于邻接同步，创新性的设计了一种 GPU 上的无全局同步的大规模 Scan 算法，相较于现有算法，对于规模为 N 的数组，我们将 Scan 算法的全局内存读写从 $3N$ 降到了 $2N$ 。Scan（也称 Prefix sum）算法是一种简单但非常常用的基础并行算法，是并行压缩，并行基数排序，并行快速排序，并行内存分配，并行 SpMV，并行 BFS 等其他基础算法中的重要组成部分。现有的基于 GPU 的大规模 Scan 算法，均是分三阶段进行，各个阶段之间需要进行全局同步。我们设计了一种基于 GPU 的新型 Stream-Scan 算法，首先对需要 Scan 的数组进行分段，然后启动固定数量的 Block（Work-Group）依次对每一段进行 Scan，各 Block 内部分三阶段进行，第一阶段 Reduce，第二阶段等待前一个 Block 的 Reduce 结果，第三阶段 Propagate。该算法无需全局同步，仅启动一次 Kernel，不同 Block 之间 Reduce 和 Scan 可以同时进行，能充分利用计算单元。同时各段数据在同一个 Block 中处理的时候，均存储在 Shared Memory（Local Memory）和寄存器中，无需重复从全局内存读取，将全局内存读写从 $3N$ 降到了 $2N$ 。我们将算法（OpenCL 实现）在 NVIDIA 和 AMD GPU 上分别进行了测试，在 NVIDIA Tesla C2050 上，在打开 ECC 的情况下，相较于开源库 Cudpp（CUDA 实现）最高获得了 1.8 倍加速比，相较于开源库 B40C（CUDA 实现）最高获得了 1.5 倍加速比，相较于 Matrix Scan（ICS'08）（OpenCL 实现）最高获得了 1.5 倍加速比。在 AMD HD 5850 上，相较于 Matrix Scan 最高获得了 1.4 倍 kernel 加速比。相关成果已经被国家顶级会议 PPOPP 2013 接收。

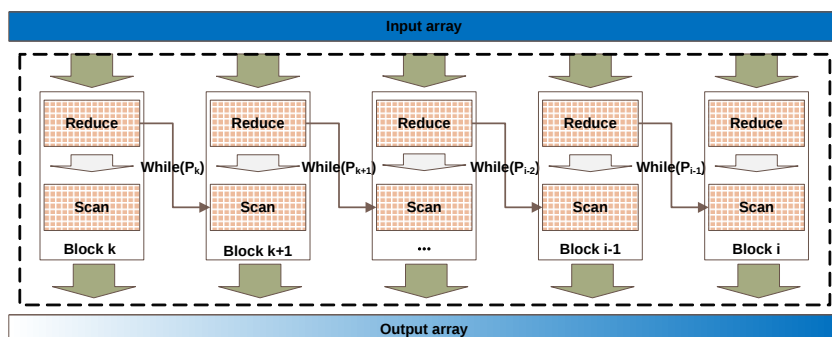


图 7 程序框架图

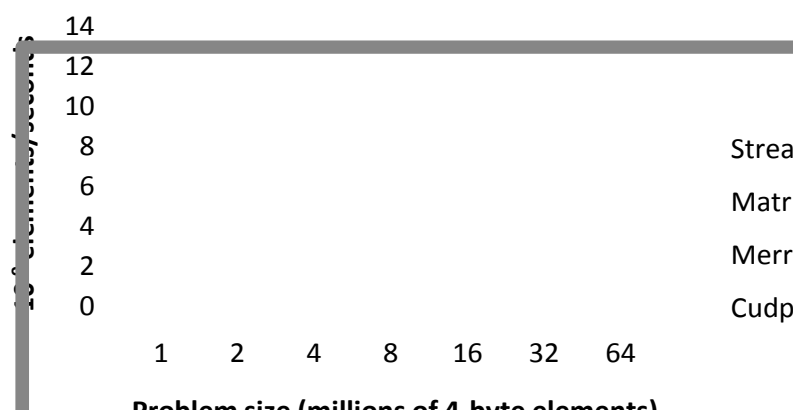


图 8 程序在 NVIDIA Tesla C2050 上吞吐率（打开 ECC）

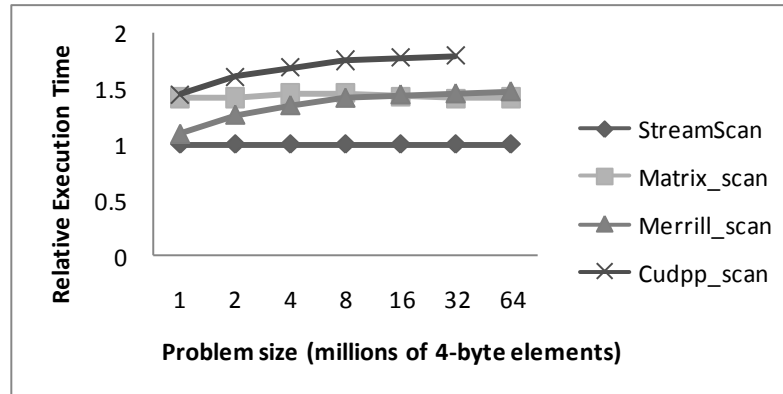


图 9 程序在 NVIDIA Tesla C2050 上加速比（打开 ECC）

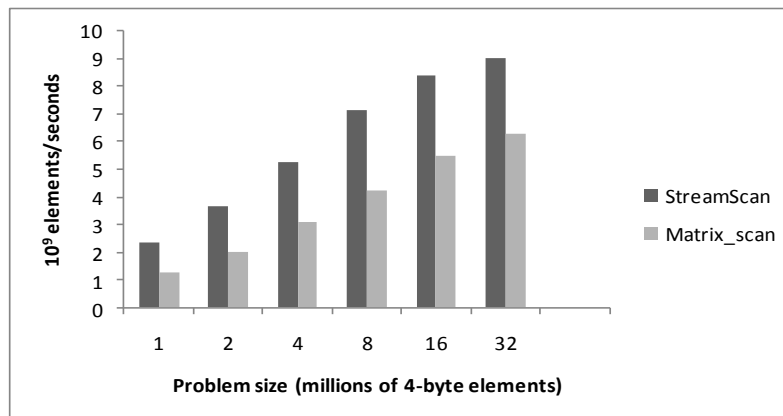


图 10 程序在 AMD HD 5850 上吞吐率

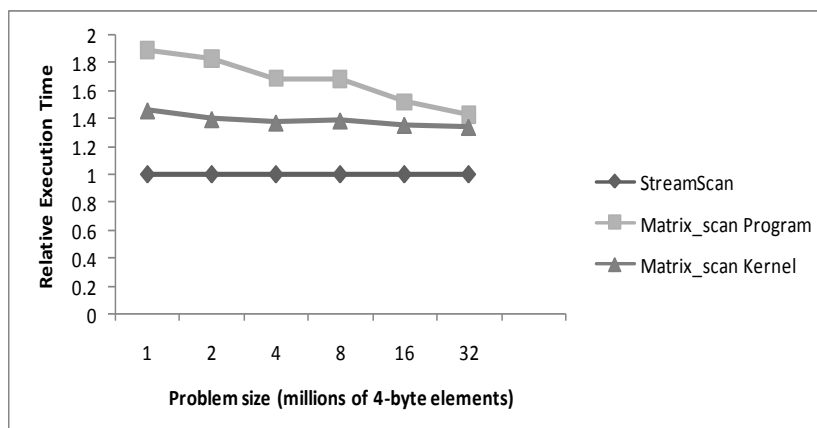


图 11 程序在 AMD HD 5850 上加速比

3) GPU 性能优化指导模型—GPURoofline。相关工作已发表在 EURO-PAR 2012 和 ICA3PP2012 国际会议上。

提出并构建完成了一个与特定硬件平台相关、可视化的 GPU 程序性能优化指导模型—GPURoofline。该模型针对因 GPU 底层硬件架构的复杂性和差异性而导致的 GPU 程序“实现难，优化更难”的特点，为 GPU 程序的性能优化提供可视化指导，以减轻应用程序开发人员的工作负担。GPURoofline 模型通过定义 GPU 程序优化的主要途径：片外访存带宽的利用率、片上计算资源的利用率和数据本地化，以及针对访存和计算特点的性能优化链的构建，使其可以提供性能信息来确定 GPU 程序在特定硬件平台上的性能瓶颈以及应选择的优化策略和方法。最后通过引入 Roofline 模型和计算密度的概念，实现了性能优化链与数据本地化的有效结合以及可视化，从而使应用开发人员特别是对 GPU 底层架构不熟悉的应用开发人员，可在其直观指导下更加容易的实现高性能 GPU 程序。GPURoofline 模型被应用于图形图像核心处理函数的 GPU 移植优化工作中，并在其指导下，不仅使程序的优化工作负担大大减轻，并且都取得了非常好的加速比，有效验证了该模型的可用性和真确性。图 1 为针对 AMD HD5850 GPU 和 NVIDIA C2050 GPU 的 GPURoofline 模型，GPU 程序性能主要受三方面的限制：片外访存带宽、硬件平台的计算峰值以及计算密度。图 2 为在 GPURoofline 模型的指导下，图像拉普拉斯变换和图像积分图算法在 GPU 上的性能提升。

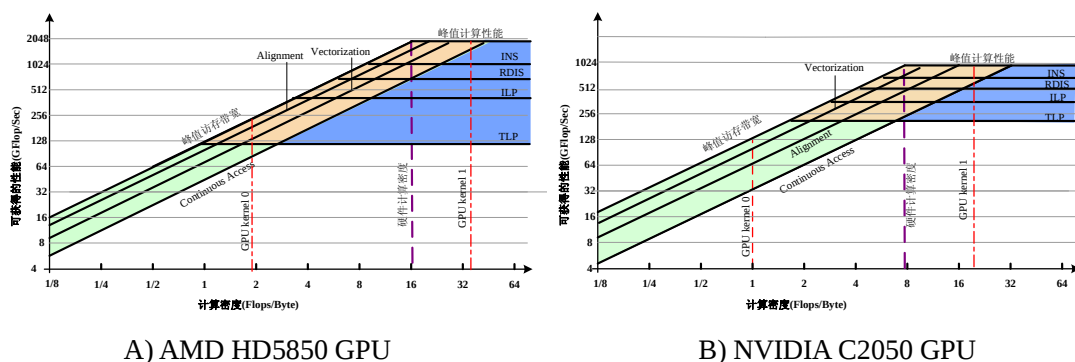


图 1 GPURoofline 模型

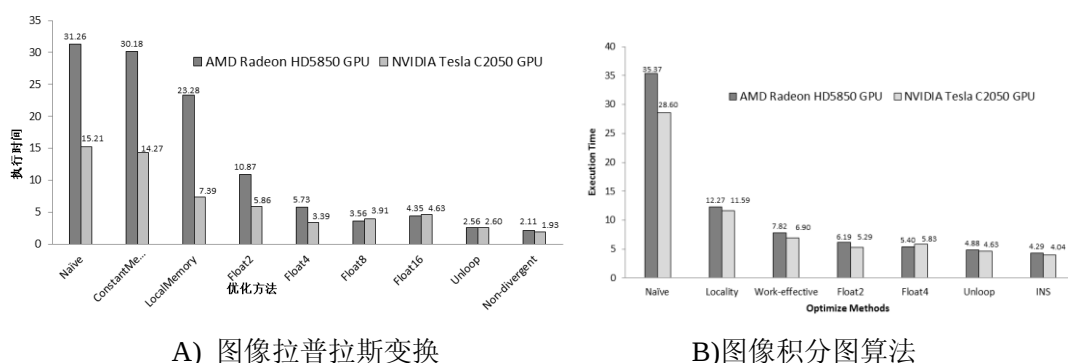


图 2 部分图形图像核心处理函数的优化性能提升

4) 非规则应用在 GPU 上的实现和优化，相关工作已经发表在 IEEE HPCC2012 国际会议上。

引入了非规则应用在 GPU 上实现和优化的五种策略和方法。非规则应用是指该算法的多线程实现会导致线程间存在着严重的负载不平衡或者数据依赖。非规应用的这种特性，对于大规模、细粒度并行的 GPU 硬件架构来说是极不友好的。如果仅仅使用传统的优化方法，非规则应用很难在 GPU 上取得满意的性能加速比。而随着计算

能力和可编程性的不断增强，GPU 不仅仅应用于传统的科学计算领域，在图形图像处理、游戏、三维可视化等非传统领域也得到了越来越广泛的应用。而这些非传统领域的大多数算法具有非规则的特性，这个特性严重制约了这些算法在 GPU 上的性能。针对这种情况，我们引入了非规则应用在 GPU 上实现和优化的五大关键技术：粗粒度并行、Uberkernel、Persistent Kernel、本地队列和全局队列；并通过性能特征参数的定义和抽取，以及性能参数传递机制的建立和完善，完成了针对不同硬件平台的可调优 GPU kernel 的初步实现。最后，完成了具有鲜明非规则特性的 Viola-Jones 人脸检测算法的 GPU 实现和优化，并在两个 GPU 平台上取得了较高的加速比，验证了在不同 GPU 平台间实现性能移植的可行性。图 3 显示了 Viola-Jones 人脸检测算法在两个 GPU 平台上相对于其 CPU 版本（OpenCV2.3）的性能加速比。

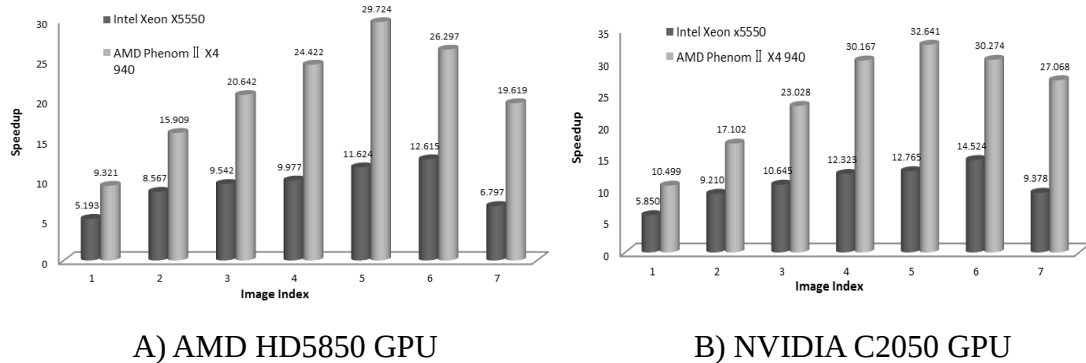


图3 Viola-jones 人脸检测算在不同 GPU 平台上的性能提升

- 5) GPU 上国产自适应 FFT 软件包 MPFFT 的研制，相关工作发表在 Journal of Computer Science and Technology 第 28 卷第 1 期，以及计算机研究与发展第 50 卷。

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。

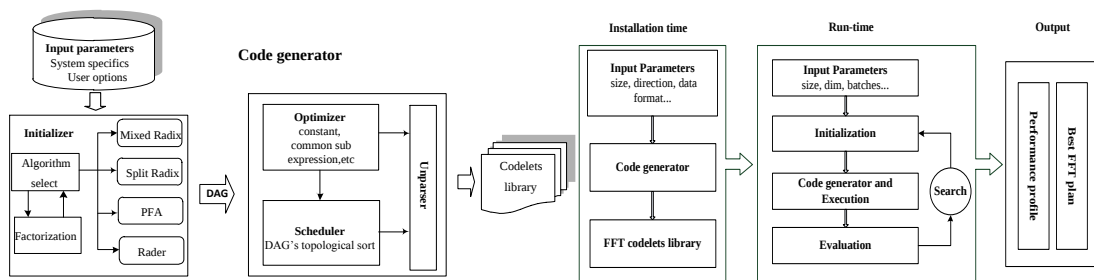


图 7 MPFFT 的整理框架和代码生成模块示意图

自适应性能优化 FFT 的整体框架如图 7 所示，主要涉及到 3 个概念：优化空间，代码产生，搜索。优化空间包含了影响程序性能的参数；搜索部分提供不同的参数组合策略；代码产生模块通过对优化空间可选的不同参数进行组合，从而产生多份不同配置的 kernel 代码。最终通过 benchmark 遴选出最佳的 kernel 实现。由于代码产生和搜索模块主要是在主机端进行，故而本节主要介绍运行时优化空间参数的选择，以及静态调用的小因子代码库(codelet)。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：

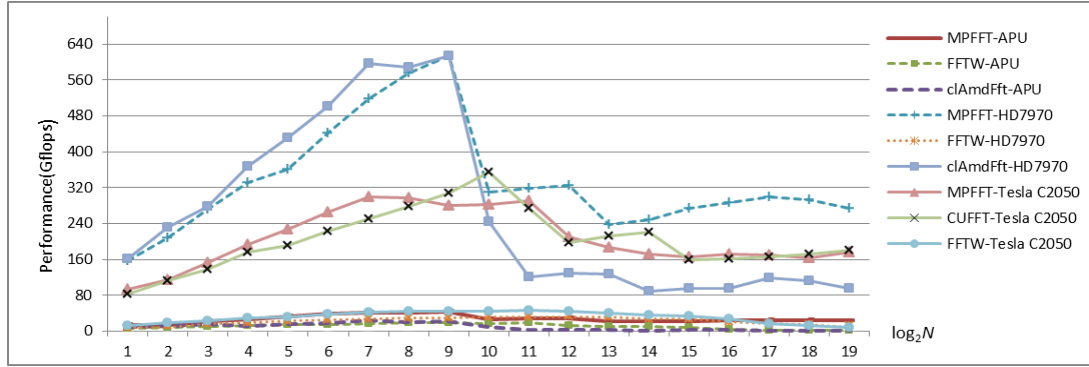


图8 一维FFT的性能测试结果

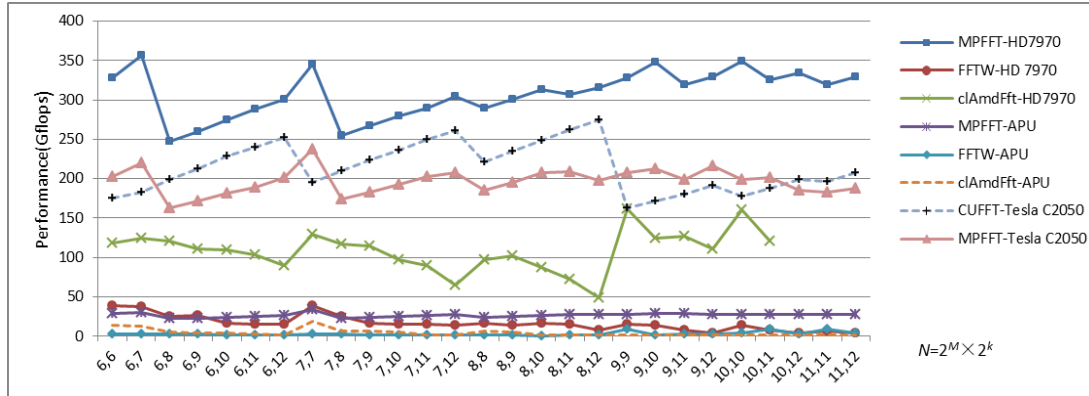


图9 二维FFT的性能测试结果

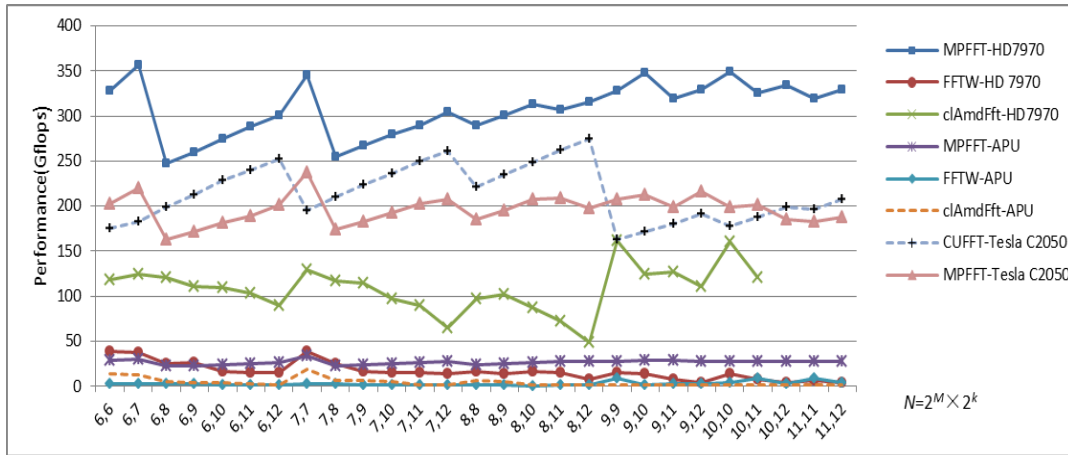


图 10 三维 FFT 的性能测试结果

实验结果表明：一维 MPFFT 的性能，在 APU 上相对于 clAmdFft 库，取得的平均加速比为 3.54，是 4 线程 FFTW 的 7.49 倍。在 AMD HD7970 上，当 FFT 规模 N 为 2048 时取得了最佳性能，值为 614.11Gflops。对于多维 FFT 的计算，MPFFT 采用的是 FFT 行列算法，取得的实验结果如图 9 和图 10 所示。对于二维和三维 FFT，MPFFT 在 APU 上相对于 clAmdFft 取得的平均加速比分别为 15.2 和 4.47，在 AMD HD7970 上相对 clAmdFft 的平均加速比分别为 3.01 和 1.69，最大为 6.48 和 7.27，是 8 线程 FFTW 的 29.19 倍和 10.93 倍。此外，MPFFT 在 NVIDIA GPU 上性能均达到了 CUFFT 的 93%。

3. 并行计算模型研究进展

1) 提出一个新的基于局部性的多核计算模型

针对多核处理器越来越复杂的存储层次，以及面向众多具有加载计算模式的并行编程模型，数据结构设计以及算法优化技术，本文提出了一个基于算法局部性函数的性能分析模型。利用该模型可以根据算法局部性特征，分析在具有不同深度存储层次架构上的性能，除了在纵向角度分析，该模型还可以分析算法横向局部性，以及私有高速缓存和共享高速缓存两者之间的优缺点。此外，我们还从内存数据访问角度扩展了 Amdahl 定律，得到了不同体系模型上的加速比，与一些已有加速比模型不同，我们的模型主要从局部性，即数据移动角度来分析并行算法的加速比，弥补了以往只关注计算能力加速比的不足。相关工作发表在 CLUSTER2012 上。

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，ICPP2012 收录。

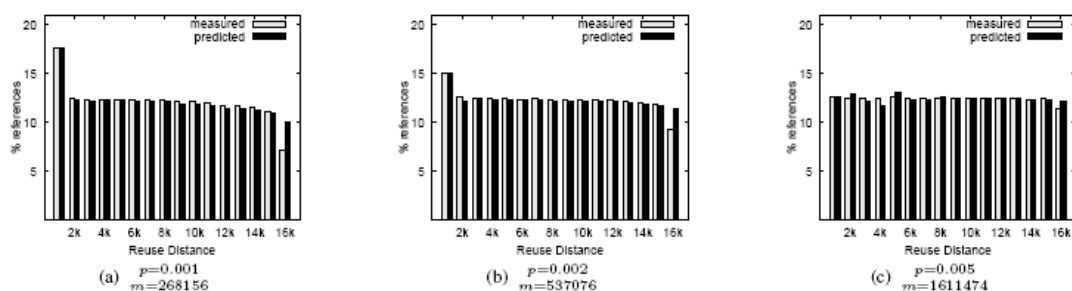


图 14. 局部性预测和实测结果对比

4. 总结

综上所述，2012 年指导并毕业博士研究生一名,联合指导毕业博士生一名,硕士研究生三名,三名硕士生转为博士生。目前正指导硕士研究生七名，博士研究生九名，联合指导博士研究生两名（中科大，中国海洋大学),硕士 1 名。申请软件著作权九项，申请专利一项。

论文发表情况

- 已发表译著 1 本，文章 19 篇，大会特邀报告 6 次，录用待发表文章 4 篇
- 国际顶级会议 PPOPP 2013 录用文章一篇！
- 国际期刊录用/发表文章两篇，JCST 2013（SCI）， Tsinghua Science and Technology（EI）
- 国际著名会议发表文章六篇 EuroPar 2012，Cluster 2012，IEEE ICPADS 2012， IEEE ICPP 2012， IEEE HPCC 2012， ICA3PP 2012
- 国内一级刊物录用两篇（计算机研究与发展）；
- 国内核心期刊发表四篇；
- 国内会议发表论文六篇。（一篇获大会优秀论文提名奖）

大会特邀报告:

- Invited Keynote Speaker, SIAM PP12, Feb.15-17, Georgia, USA

- Invited Speaker, 2nd China US Software Workshop, Mar. 5-7, San Diego, CA, USA.
- Invited Speaker, ATIP/A*CRC Workshop on Accelerator Technologies for High-Performance Computing, May 7-10, Singapore.
- 26th IEEE International Parallel & Distributed Processing Symposium (IPDPS2012), May 21-25, 2012, Shanghai, China (Workshop Co-Chair)
- 18th International European Conference on Parallel and Distributed Computing Euro-Par 2012, Rhodes Island, Greece, August 27-31, 2012. (发表文章)
- 2012 年 7 月 26-27 日, 第八届中国 CAE 工程分析技术年会, 成都, 四川 (大会特邀报告)
- 2012 年 8 月 17-8 月 18 日, 第二届中国科学院超级计算应用大会 (SCA2012), 腾冲, 云南. 大会特邀报告
- 2012 年 10 月 29-10 月 31 日, 2012 全国高性能计算学术年会 (张家界, 大会特邀报告)
- 2012 年 12 月 21-12 月 23 日, 2012 第八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)。

在研课题

- 国家自然科学基金重点基金“大规模异构并行系统的高效能调度理论与方法”, 进展良好;
- 中科院软件所杰出青年人才专项计划, 2009.9-2013.9, 课题负责人, 中期检查;
- 国防科工委项目“基于光线追踪的高精度成像并行数值计算软件”一项, 进展良好;
- 中科院空间中心横向课题一项, 48 万, 课题负责人, 进展良好。
- 美国 AMD 公司联合成立 APU 软件联合实验室, 课题负责人, 进展良好。

学术活动和兼职

- 国家自然科学基金委第十四届信息科学部专家评审组成员
- 中国科技大学、中国传媒大学和华南理工大学兼职教授/博导
- 山东省科学院计算机网络省重点实验室学术委员会委员
- 北京科技大学材料领域知识工程北京市重点实验室学术委员会委员
- HPC China 2012 程序委员会副主席, CAE 2012 专家委员会副主席, SCA 2012 专家委员会委员, 中科院超级计算中心用户委员会委员
- 核心期刊《计算机科学》和《计算机工程与科学》编委
- 中科院《科研信息化技术与应用》编委, CCF《计算机学会通讯》编委, 上海超算《高性能计算机发展与应用》编委
- 中国软件行业协会常务理事/数学软件分会秘书长
- 中国计算机学会理事/高性能计算专委会秘书长
- IPDPS 2012, CCGrid 2012, e-Science 2.0, Cluster 2012, IEEE ICPADS 2012, CGO 2013 程序委员会委员
- ISC Steering Committee, IEEE HPCC 2013 PC Co-Chair

2012 年完成和申请课题情况:

- 江南计算所 国防军工课题, 2009.1-2012.6, 顺利结题, 课题负责人
- 核高基重大专项课题子课题, 2009.1-2012.12, 正在验收, 课题负责人
- “十一五” 863 计划“高效能计算机及网格服务环境”重大项目“高性能计算与网

格应用”课题,《曙光 6000 千万亿次高效能计算机系统研制》(No.2006AA01A129), 2010.1-2013.6,顺利结题,子课题组组长;

- 中国科学院知识创新工程重大项目《面向生命科学前沿的高性能计算研究》课题《高性能蛋白质定量计算软件平台构建与开发》,2008.1-2012.4,顺利结题,课题负责人

发表文章列表:

1. 作者: Benedict R. Gaster Lee Howes David R. Kaeli Perhaad Mistry Dana Schaa,译者: 张云泉 张先轶 龙国平 姚继锋,定价: 76.00 元,出版日期: 2012 年 6 月,出版社: 清华大学出版社,分类: 并行计算/计算机体系结构
2. Yan Li, Yunquan Zhang, MPFFT: An Auto-tuning FFT Library for OpenCL GPUs, JCST, Springer-Verlag;co-published by Science Press, accepted.(SCI), 2013.
3. Shengen Yan, Guoping Long, Yunquan Zhang, StreamScan Fast Scan Algorithms for GPUs without Global Barrier Synchronization, PPOPP 2013, accepted.
4. Xianyi Zhang, Qian Wang, Yunquan Zhang, Model-driven Level 3 BLAS Performance Optimization on Loongson 3A Processor, ICPADS 2012, Singapore.
5. Liang Yuan, Yunquan Zhang: A Locality-based Performance Model for Load-and-Compute Style Computation. CLUSTER 2012: 566-571
6. Haipeng Jia, Yunquan Zhang, Guoping Long, Jianliang Xu, Shengen Yan, Yan Li: GPU Roofline: A Model for Guiding Performance Optimizations on GPUs. Euro-Par 2012: 920-932
7. Haipeng Jia, Yunquan Zhang, Weiyan Wang, Jianliang Xu: Accelerating Viola-Jones Face Detection Algorithm on GPUs. HPCC 2012: 396-403
8. Haipeng Jia, Yunquan Zhang, Guoping Long, Shengen Yan: An Insightful Program Performance Tuning Chain for GPU Computing. ICA3PP (1) 2012: 502-516
9. Liang Yuan, Chen Ding, Daniel tefankovic, Yunquan Zhang: Modeling the Locality in Graph Traversals. ICPP 2012: 138-147
10. Chao Li, Yunquan Zhang, Changwen Zheng, Xiaohui Hu: Implementing High-performance Intensity Model with Blur Effect on GPUs for Large-scale Star Image Simulation. IPDPS Workshops 2012: 1879-1888
11. Shengen Yan, Yunquan Zhang, Guoping Long, Summed-area table algorithm optimization based on the OpenCL, Proceedings of the ATIP/A*CRC Workshop on Accelerator Technologies for High-Performance Computing: Does Asia Lead the Way? pp.798 - 800, ISBN: 978-1-4503-1644-6.
12. Wang, Weiyan; Zhang, Yunquan; Yan, Shengen; Zhang, Ying; Jia, Haipeng, Parallelization and performance optimization on face detection algorithm with OpenCL: A case study, Tsinghua Science and Technology, Vol.17, No.3, pp.287-295, 2012.
13. 李焱, 张云泉等, MPFFT: 异构平台上性能自适应 FFT 框架, 计算机研究与发展, 已录用, 2013
14. 孙相征, 张云泉等, 针对应用对角线矩阵特点的 SpMV 自适应性能优化, 计算机研究与发展, 已录用, 2013
15. 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. 计算机科学, 2012, (05):
16. 张樱 张云泉 龙国平. 基于 OpenCL 的图像模糊化算法优化研究. 计算机科学, 2012, (03): 260-

17. 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. 计算机科学, 2012, (04): 282-287
18. 费辉 张云泉 王靖. 基于 GPU 的非标记定量软件 QuantWiz 并行化实现. 计算机科学, 2012, (06):
19. 吴长茂, 张云泉等, 基于大型场景的高精度成像并行光线追踪算法, HPC China 2012.
20. 庞旭等, 基于 OpenCL 的均值平移算法在多个众核平台的性能优化研究, HPC China 2012.
21. 王靖等, 基于 MPI 和 CUDA 的蛋白质定量软件的设计和分析, HPC China 2012.
22. 蒋丽媛等, 基于 OpenCL 的连续数据无关访存密集型函数并行与优化研究, HPC China 2012.
23. 张龙等, 基于 OpenCL 的 addWeighted 和 magnitudeSqr 算法性能优化研究, HPC China 2012.
24. 肖玄基等, 异构平台数学库 MAGMA 性能测试与分析, HPC China 2012.
25. Shengen Yan, Guoping Long, Yunquan Zhang. StreamScan: Fast Scan Algorithms for GPUs without Global Barrier Synchronization. PPOPP 2013 (已接收)
26. Shengen Yan, Yunquan Zhang, Guoping Long. Summed-Area Table algorithm optimization based on OpenCL. ATIP 2012.

下一年度工作计划, 包括国内外合作与交流计划

1. 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。
2. 进一步完善 GPU Roofline 模型，提高相对性能优化链的精确性，并尽可能提升模型对性能提升的量化处理。并研究将该模型作为 GPU 程序自适应优化框架底层基础的可行性。
3. 结合任务偷取（Task stealing）或者任务贡献（task donation）等任务调度方法改进和完成全局队列的任务策略。从根本上消除 work-group 间负载不均衡的影响。并通过更多非规则应用在 GPU 上的研究和实现，总结出一套提高此类算法在 GPU 上实现和优化的完整通用方案。
4. 可调优 GPU kernel 框架的完美实现。全面提取影响 GPU 程序的性能因素，完成性能调优空间和参数传递机制的建立，最终完成可在不同 GPU 平台上实现性能无缝移植的可调优 GPU kernel 框架。
5. 牢牢把握当前异构计算的趋势和研究状态，提高对新硬件以及新编程模型和方法的敏感性。特别是 Fusion 架构对算法优化的影响。相对现在的架构，在 Fusion 架构中，CPU 和 GPU 共享内存，不需要 PCI，从 ARM 和 Intel 目前实现的情况来看，都有一致的存储层次，这样二者间可以做到更加紧耦合，可大大降低同步开销。要关注在这种技术背景下，算法优化侧重点的改变。
6. 百万亿次高效能计算平台高可扩展基础并行算法研究
 - 1) 进一步优化 OpenCV 中的其他函数。同时开发一些基于 OpenCV 的典型应用。
 - 2) 进一步研究基于 Scan 的基础并行算法，如 Sort, SpMV 等，争取开发一个 OpenCL 版本的基础并行算法开源库。
7. 研究线性递推算法的自动代码生成。

年度经费使用情况概要

2012 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。