様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　4 月　　1 日現在

機関番号：１２６０８
研究種目：若手研究（Ａ）
研究期間：2010～2011
課題番号：22680002
研究課題名（和文）　ＧＰＵによるＦＦＴ計算の自動チューニング手法の研究

研究課題名（英文）　Ａｕｔｏ－ｔｕｎｉｎｇ　FFT　ｕｓｉｎｇ　ＧＰＵ

研究代表者
額田　彰（NUKADA AKIRA）
東京工業大学・学術国際情報センター・産学官連携研究員
研究者番号：40545688

研究成果の概要（和文）：NVIDIA 社製の CUDA 対応 GPU 向けの自動チューニング FFT ライブラリである NukadaFFT ライブラリを開発した．その性能は多くの場合に NVIDIA 社の CUFFT ライブラリを上回る．また複数 GPU 版についても複数 GPU を搭載するシングルノードと複数ノード版を実装し，さらなる速度向上を達成した．

研究成果の概要（英文）：We have developed the NukadaFFT library, which is an auto-tuning FFT library for NVIDIA CUDA GPUs. It outperforms NVIDIA's CUFFT library in many cases. We also implemented a multi-GPU version for both single-node with multi-GPUs and multi-node, and achieved further speed-up.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	5,000,000	1,500,000	6,500,000
２０１１年度	4,000,000	1,200,000	5,200,000
年度			
年度			
年度			
総　計	9,000,000	2,700,000	11,700,000

研究分野：総合領域
科研費の分科・細目：情報学・ソフトウェア
キーワード：ソフトウェア工学

１．研究開始当初の背景

GPU の利用による各種数値計算の高速化は今日既に十分に普及してきたといえる．従来のアクセラレータは多数の演算コアを搭載したもので演算量の極めて多い計算において高速化を実現するものであったが，GPU はメモリバンド幅においても優れたアクセラレータであるためより広範囲な計算に対応することが可能である．高速フーリエ変換(FFT)は信号処理，画像解析，マルチメディア処理のような小規模なものからスーパーコンピュータを利用するような大規模なシミュレーションまで幅広く用いられる重要な計算であるが，演算性能だけでなくメモリアクセス性能の要求も高いため従来のアクセラレータでは対応できなかった．

２．研究の目的

高速な FFT ライブラリは常に多くのアプリケーション開発者等から必要とされている．特に GPU 製品はおよそ年に１回新しいアーキテクチャ世代の製品が発売されるという状況であるため、新アーキテクチャに即対応することが可能である自動チューニング機能を搭載するライブラリを開発する．

３．研究の方法

GPUによるFFT計算の高速化は単純ではない．GPUのメモリアクセスは汎用CPUのメモリやベクトルプロセッサのそれとも異なる性質をもち，理論ピーク性能に近いメモリアクセス性能を得るためにはメモリアクセスの局所性や各スレッド間でアクセス箇所を調整することが不可欠であり，またキャッシュメモリの容量が限られるGPUではキャッシュを搭載するCPU用のアルゴリズムとベクトルプロセッサ向けのアルゴリズムのどちらでもない新たな計算手法を開発する必要がある．

また FFT の演算量を O(N logN)に削減するアルゴリズムはデータサイズに依存するため，ライブラリ化する場合にはデータサイズ毎に自動最適化を行う仕組みが不可欠であった．

また現在主流である CUDA は NVIDIA 社の GPU にのみ対応する．より広範なアーキテクチャに対応する OpenCL での実装も検討する．

４．研究成果

GPU における自動最適化のパラメータは大きく分けて以下の３点に絞られる．

(1) 入力サイズから基底への因数分解方法，及び各基底の計算順序

これはCPUでも考慮しなければならないパラメータであるが，GPU の場合には演算数だけでなくスレッド数やレジスタ数，ロード・ストア命令数などにも影響するため最適なパラメータはGPUアーキテクチャに大きく依存する．

(2) スレッド数

GPU は多数のスレッドが交代で命令を実行する仕組みになっており，メモリアクセス時にはデータが実際に利用されるまでは他のスレッドが命令実行を行うことでメモリアクセスレイテンシを隠ぺいすることが可能である．FFT の計算ではメモリアクセス効率が重要で，最大のデータ転送スループットを得られるように最適なスレッド数を選択する必要がある．

(3) シェアードメモリアクセスパターン

１組のFFTの計算をスレッドブロックが担当する場合，スレッド間で頻繁にデータの交換が必要である．CUDAではスレッド間のデータ交換にはオンチップにある高速にアクセス可能なシェアードメモリを用いる．シェアードメモリは16個や32個などのバンクに分かれており，別々のバンクは同時にアクセス可能である．FFT におけるスレッド間のデータ交換パターンに対してバンクコンフリクトが起こらないように最適化するべく自動的にパディング挿入パターンを決定する．

これらの３つのパラメータを基本的には網羅的探索として全ての組み合わせで試行し，性能が最高となるパラメータを得る．

ライブラリとして実装する場合に，全ての入力サイズ及び各パラメータに対応するコードを予め生成しておくことは現実的ではない．そのため与えられた入力サイズに対して全パラメータに対応するコードを生成し，試行して最適なものを選択する方式を採用する他にない．全パラメータの試行にはかなり時間を要する．試行には①カーネル生成，②コンパイル，③ロード，④デバイスへのデータ転送，⑤カーネル実行，⑥ホストへのデータ転送，⑦結果のチェックの７つのタスクで構成される．特にホストCPU 側で行われる②，④，⑥，⑦のタスクが時間がかかる．中でも②が一番時間がかかるため，C for CUDA 言語ではなくよりローレベルな CUDA PTX 言語でカーネルをメモリ上に生成してコンパイルすることで大幅な時間短縮を実現した．

OpenCL 環境では CUDA とほぼ同じ機能が提供されており，特にカーネル部分のコードは似ている．唯一の差異はFFTで利用する三角関数のテーブルの扱いで，CUDAではテクスチャメモリを利用していたが OpenCL ではコンスタントメモリを使用した．AMD製RADEON HD 7970を用いてCUDA+NVIDIA製GPUを超える性能を実現することができた．一方，性能のポータビリティは OpenCL では確保されていない．NVIDIA製GPUではCUDA版と比べるとOpenCL 版の性能は現時点ではかなり劣る．今後ドライバの成熟を待てばこの差は縮まることが期待される．

さらに複数GPUへの対応も進めている．特にGPU のデバイスメモリ容量は限られるため，実アプリケーションでは容量を確保するために複数GPUを利用するケースも少なくない．GPU 間の転送は PCI-Express インターフェイスや InfiniBand などのノード間インターコネクトの性能に大きく依存し，実行時間の大部分を GPU 間の all-to-all 通信が占めることになる．CUDAではversion 4.0からこのGPU 間通信を強化しており，PCI-Express ネットワークを介した GPU 間の直接通信や，InfiniBand HCAとの協調動作などがサポートされた．これらの機能を活用することで，通信が全く必要ないシングルＧＰＵ実行時と比べても 4GPU 搭載システムで約 2 倍，64 ノード 64GPU のクラスタで最大 13 倍の性能向上を実現した．通信関連の各種自動最適化は今後の課題である．

シングル GPU 用の FFT ライブラリは NukadaFFT ライブラリという名前で既に一般公開しており，現時点（平成２４年３月３１日）でのダウンロード数は５７６である．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計３件）

① 遠藤 敏夫，額田 彰，松岡 聡．スーパーコンピュータ TSUBAME 2.0 における Linpack 性能1 ペタフロップス超の達成．情報処理学会論文誌コンピューティングシステム，査読有，Vol. 4, No.4 (ACS 35), pp.169-179, 2011 年 10 月．
② 額田彰．CUDA による高速フーリエ変換，応用数理，査読有，第 20 巻，第 2 号，pp. 37-43，応用数理学会，2010 年 6 月．
③ 遠藤敏夫，額田彰，松岡聡．異種アクセラレータを持つ TSUBAME スーパーコンピュータの Linpack 評価，応用数理，査読有，第 20 巻，第 2 号，pp.29-36，応用数理学会 2010 年 6 月．

〔学会発表〕（計 18 件）

① Akira Nukada, Yutaka Maruyama, Satoshi Matsuoka. “High Performance 3-D FFT using multiple CUDA GPUs”, In Proceedings of the Fifth Workshop on General Purpose Processing using Graphics Processing Units (GPGPU-5) in conjunction with ACM ASPLOS XVII, London, UK, ACM Press, March 3rd, 2012.
② 遠藤 敏夫,松岡 聡,額田 彰,長坂 真路,四津 匡康，“グリーンスパコン TSUBAME2.0 における電力危機対応運用，情報処理学会研究報告，Vol. 2011-ARC-197/HPC-132，札幌，2011 年 11 月 28 日．
③ Takashi Shimokawabe, Takayuki Aoki, Tomohiro Takaki, Akinori Yamanaka, Akira Nukada, Toshio Endo, Naoya Maruyama, and Satoshi Matsuoka, “Peta-scale Phase-Field Simulation for Dendritic Solidification on the TSUBAME 2.0 Supercomputer”, In Proc. of 2011 ACM/IEEE International Conference for High Performance, Networking, Storage, and Analysis (SC’11), Seattle, ACM Press, Nov. 15th, 2011. (Technical Paper and Gordon Bell Award finalist.)
④ Shuntaro Yamazaki, Akira Nukada, Masaaki Mochimaru, “Hamming Color Code for Dense and Robust One-shot 3D Scanning”, In Proc. of the 2011 British Machine Vision Conference, Dundee, Scotland, Springer, Aug. 30th, 2011.
⑤ Akira Nukada, “Fast Fourier Transform for AMD GPUs”, AMD Fusion Developer Summit 2011, Bellevue, WA. Jun. 15th, 2011.
⑥ 遠藤 敏夫，額田 彰，松岡 聡．スーパーコンピュータ TSUBAME 2.0 における Linpack 性能1 ペタフロップス超の達成．先進的計算基盤システムシンポジウム (SACSIS2011)，秋葉原，2011 年 5 月 27 日．
⑦ Akira Nukada, Hiroyuki Takizawa, and Satoshi Matsuoka. “NVCR: A Transparent Checkpoint-Restart Library for NVIDIA CUDA”, In Proc. of 20th Heterogeneity in Computing Workshop (HCW 2011), in conjunction with IPDPS 2011, Anchorage, AK, USA, May 16th, 2011.
⑧ Leonardo Bautista Gomez, Akira Nukada, Naoya Maruyama, Franck Cappello and Satoshi Matsuoka. Low-overhead diskless checkpoint for hybrid computing systems, In Proc. of International Conference on High Performance Computing (HiPC 2010), Goa, India, Dec.20th, 2010.
⑨ Ali Cevehir, Cevdet Aykanat, Ata Turk, B. Barla Cambazoglu, Akira Nukada and Satoshi Matsuoka. Efficient PageRank on GPU Clusters, 情報処理学会研究報告，Vol. 2010-HPC-128, No.21, 札幌，2010 年 12 月 17 日．
⑩ 遠藤 敏夫，額田 彰，松岡 聡．ヘテロ型スーパーコンピュータ TSUBAME 2.0 の Linpack による性能評価，情報処理学会研究報告，Vol. 2010-HPC-128, No.11, 札幌，2010 年 12 月 16 日．
⑪ 長坂仁，丸山直也，額田彰，遠藤敏夫，松岡聡．ＧＰＵにおけるモデルに基づいた電力効率の最適化，情報処理学会研究報告，Vol. 2010-HPC-128, No. 2, 札幌，2010 年 12 月 16 日．
⑫ Takashi Shimokawabe, Takayuki Aoki, Chiashi Muroi, Junichi Ishida, Kohei Kawano, Toshio Endo, Akira Nukada, Naoya Maruyama and Satoshi Matsuoka, An 80-Fold Speedup, 15.0 TFlops, Full GPU Acceleration of Non-Hydrostatic Weather Model ASUCA Production Code, In Proc. of the 2010 ACM/IEEE conference on Supercomputing (SC’10), New Orleans, IEEE Press, Nov. 17th, 2010.

⑬ Akira Nukada and Satoshi Matsuoka. "NukadaFFT: An Auto-Tuning FFT Library for CUDA GPUs". NVIDIA GPU Technology Conference 2010, Research Summit Poster, San Jose, September 22nd, 2010.

⑭ Hitoshi Nagasaka, Naoya Maruyama, Akira Nukada, Toshio Endo and Satoshi Matsuoka. Statistical Power Modeling of GPU Kernels Using Performance Counters, Proceedings of the First International Green Computing Conference (IGCC'10), Chicago, Aug. 17th, 2010.

⑮ Ali Cevahir, Akira Nukada, and Satoshi Matsuoka. High Performance Conjugate Gradient Solver on Multi-GPU Clusters Using Hypergraph Partitioning, Computer Science - Research and Development, Vol. 25, No. 1-2, Springer, (Proceedings of the 2010 International Supercomputing Conference (ISC'10), Hamburg, Germany, May 31st, 2010).

⑯ Akira Nukada and Satoshi Matsuoka, "Fast Fourier Transform using CUDA GPUs", ETHZ - Tokyo Tech Workshop: Computing with GPUs, Cells, and Multicores, Zurich, Switzerland, May 11th, 2010.

⑰ Naoya Maruyama, Akira Nukada, and Satoshi Matsuoka. A High-Performance Fault-Tolerant Software Framework for Memory on Commodity GPUs, In Proceedings of 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2010), Atlanta, April 20th, 2010.

⑱ Toshio Endo, Akira Nukada, Satoshi Matsuoka, and Naoya Maruyama. Linpack Evaluation on a Supercomputer with Heterogeneous Accelerators, In Proceedings of 24th IEEE International Parallel & Distributed Processing Symposium (IPDPS 2010), Atlanta, April 21st, 2010.

〔図書〕(計１件)

① Tamito Kajiyama, Akira Nukada, Reiji Suda, Hidehiko Hasegawa, and Akira Nishida. Toward Automatic Performance Tuning for Numerical Simulations in the SILC Matrix Computation Framework, Chapter 11 of "Software Automatic Tuning : From Concepts to the State-of-the-Art Results", pp. 175-192, Springer, Sep. 2010.

〔その他〕
ホームページ等

下記 URL にて本研究の成果の一部である NukadaFFT ライブラリソフトウェアを公開.
http://matsu-www.is.titech.ac.jp/~nukada/nufft/

６．研究組織

(1)研究代表者

額田　彰 (NUKADA AKIRA)
東京工業大学・学術国際情報センター・産学官連携研究員
研究者番号：40545688

(2)研究分担者

(　　)

研究者番号：

(3)連携研究者

(　　)

研究者番号：