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　3 月 31 日現在

機関番号：82112
研究種目：基盤研究（C）
研究期間：2009　～　2011
課題番号：21510217
研究課題名（和文）超ハイスループットシークエンスデータ解析を目指した高速配列解析環境の構築
研究課題名（英文）Design and Implementation of high-performance sequence analysis environment using GPGPU
研究代表者
末次　克行（SUETSUGU YOSHITAKA）
独立行政法人農業生物資源研究所・昆虫ゲノム研究ユニット・主任研究員
研究者番号：80533471

研究成果の概要（和文）：
本課題では、膨大な配列データの効率的な処理を目的とし、GPU を科学技術計算に用いる General Purpose computation on Graphic Processing Unit (GPGPU)を使ったソフトウェア開発を行った。解析の計算量およびデータサイズから最適だと思われる実装を施すことで高いパフォーマンスを得ることが出来た。いくつかの配列解析アルゴリズムを実装した結果、SSE やマルチコア CPU を使った実装と比較して２桁以上の高速化を達成した。

研究成果の概要（英文）：
Since the emergence of high-throughput sequencing platform, the cost for both experimental and time were much reduced for DNA sequencing. This makes easy to produce enormous amount of DNA data. As the result, analyzing huge sequence database has become a next target to find an e□cient solution. To overcome this problem, we applied General Purpose computation on Graphic Processing Unit (GPGPU) for sequence analysis. We implemented several algorithms including Smith-Waterman algorithm based on GPGPU. The GPU-based programs showed higher performance compared to CPU-based ones.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,500,000	450,000	1,950,000
2010 年度	500,000	150,000	650,000
2011 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,700,000	810,000	3,510,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・ゲノム情報科学
キーワード：情報工学、配列解析、並列処理、ゲノム、GPGPU

１．研究開始当初の背景

次世代 DNA シークエンサーの登場により塩基配列解読の低コスト化が進み、塩基配列情報の蓄積量は急激な伸びを示している。これら莫大な量の塩基配列情報を超高速処理するための新しい情報処理技術が求められている。バイオインフォマティクス分野では、PC クラスタをはじめとする並列計算機はスーパーコンピュータや専用ハードウェアに

比べ低コストで高い処理能力を得られることから広く用いられている。しかし、PC クラスタはトータルとしての発熱量が多く、空調が整った専用のサーバ・ルーム内に設置する必要があり、省エネルギー、省スペース性の面において問題がある。また、 PC クラスタ全体の性能向上が必要となった場合、構成要素である PC をすべて入れ替える必要があり、性能向上を達成するために必要となる人的、金銭的コストが高い。さらに、PC クラスタの運用は計算機に関する高度の知識を必要とし、誰にでも容易に取り扱えるものではない。また、構成する部品点数も多いことから維持・管理に要するコストが高い。等の問題点がある。

一方で、PC 上で描画を担当する GPU の性能向上はめざましく、GPGPU(General Purpose computation on GPU)と呼ばれる科学技術計算を高速な GPU 上で行う技術が脚光を浴びており、バイオイメージング、分子動力学計算、信号処理といった分野において GPU を科学計算分野へ利用しようとする動きが急速に広がりつつある。バイオインフォマティクス分野においては、超ハイスループットシークエンスデータを見据えた GPU の本格的な利用はほとんど行われていないのが現状であった。

２．研究の目的

本課題では、GPGPU による低コスト且つ超高速塩基配列解析環境の構築を目指した。配列のペアワイズアラインメント等、ゲノム研究グループにおいて使用頻度の高い配列処理を実装して、配列解析統合環境を提供することを目標とした。

３．研究の方法

(1) まず、配列解析アルゴリズムの理解を進め、並列化による高速化の可能性について十分に検討する。配列解析の並列化には粗粒度並列化と細粒度並列化の 2 通りのアプローチが考えられる。粗粒度並列処理は並列化が容易である半面、パフォーマンス向上に限界がある。一方、細粒度並列処理は一つのタスクを細かく分割し、各プロセッサが分割された小問題を担当する方式でより高いパフォーマンスが期待される。細粒度並列処理は高度な並列アルゴリズムが必要となるが、高速配列解析処理を GPU 上で達成するため、細粒度並列化アルゴリズムを極力開発する。アルゴリズムの検討を十分に行った後に順次 GPGPU による実装を進める。ソフトウェアの開発には NVIDIA 社が提供する C 言語による統合開発環境である CUDA により行う。

(2) 実装された GPU 版プログラムを従来の CPU 版と比較することにより性能の評価を行う。性能評価は公的データベース上のベンチマーク用塩基配列データを用いて行い、性能比較の結果 GPU 版プログラムの性能が十分でない場合にはアルゴリズムの修正を含めた検討を行う。

(3) 実装されたプログラムを WWW 上で公開・配布する。

４．研究成果

(1) 相同性検索アルゴリズムの 1 つである Smith-Waterman algorithm (SWA) は BLAST や FASTA などのヒューリスティックなアルゴリズムと比べ，検出感度が高いが，膨大な処理時間を要するという問題があるため、GPGPU による並列化実装を行った(図 1)。細粒度並列処理を適用した SWA に対し，複数の分割されたスコア行列計算を同時に行う粗粒度並列処理を適用したハイブリッド並列化実装を行った。この実装をベースに、以下の改良を加えた。① 計算に寄与しない未利用スレッド率を極力下げるような実装とした ② スレッド間の同期の仕方を工夫して高速化を実現した。③ メモリアクセスの効率化、即ち、レジスタと同等に高速に動作するシェアードメモリを極力活用することで高速化を図った。さらに、SW algorithm の場合，それぞれの DP は独立しているため，特に GPU 間での同期の必要はない。そこで、複数の GPU を協調動作させるようにアルゴリズムの改良を行った。具体的にはデータベース全体に対し，GPU 一基毎にクエリ 1 配列を入力することで並列化を図った。また、クエリの長さにより各 GPU での実行時間が異なるため，クエリの長さにより昇順にソートし，順々に各 GPU にクエリとして入力することで、GPU 間での実行時間に差が極力でないように工夫した。

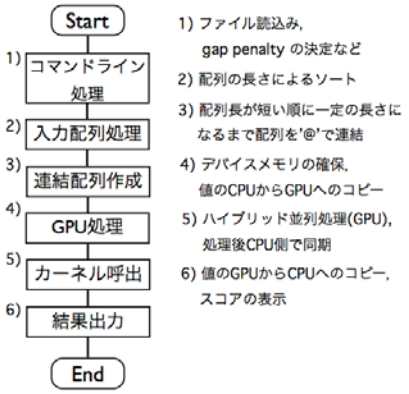

図 1： SW アルゴリズムの GPGPU 実装

図 2 はハイブリッド並列化のみの実装（青)、スレッド間同期およびスレッド利用率を改善した実装(緑)、メモリアクセスの効率化まで行った実装(橙)での、5,000 クエリ配列の処理時間を比較している（CPU は Intel Core i7 2.8 GHz, GPU は NVIDA Tesla C1060 を使用）。アルゴリズムをチューニングすることにより、大きく性能向上を果たすことができた。

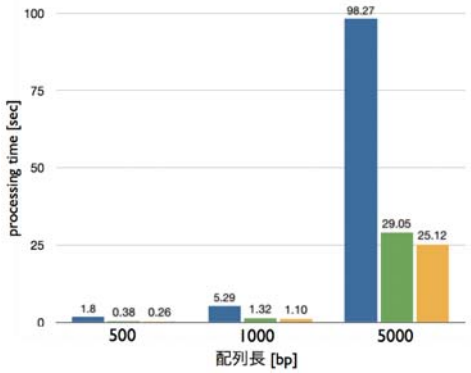

図 2: 実装の違いによる速度比較

CUDASW++, SSEARCH といった既存ソフトウェアとの速度比較による性能評価も行った(GPU: NVIDIA GeForce GTX 480, CPU: Intel Corei7 2.8GHz)。同じく GPGPU 実装である CUDASW++とはアミノ酸比較ではほぼ同等の性能が得られている。一方、塩基配列比較では SSEARCH(ギャップペナルティ: 12/2 または 4/3 を使用)と速度を比較した。ギャップペナルティが 12/2 の SSEARCH と比較して平均 2.9 倍の高速化、ギャップペナルティ 4/3 の SSEARCH とは平均 7 倍の高速化を達成することができた。

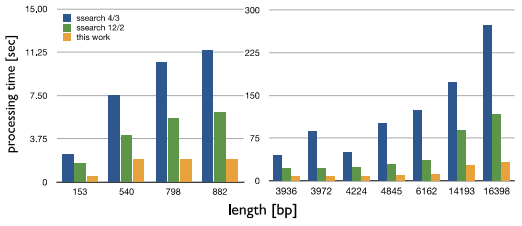

図 3: 既存ソフトウェアとの速度比較

図 4 に、複数の GPU を同時に起動できる実装を行った場合の性能評価結果をまとめた。評価には、CPU として Xeon E5502 1.87 GHz 2 core を搭載するラックマウント型 1U サーバに NVIDIA Tesla C1060(GPU) を 4 台接続したシステムを使った。図から、GPU

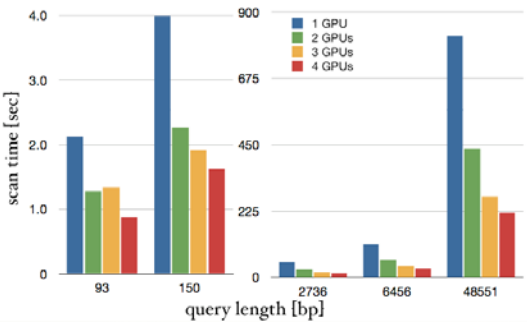

図 4: Multi-GPUs 実装の場合の性能評価結果

数が増加するにしたがい、概ね線形に処理時間が減少しており、良好な結果が得られている。これは、各 GPU 上の SWA が独立であり、他の GPU 上の SWA の処理結果に依存しないことが主な要因であると考えられる。

また、GPU ベンダに依存しない開発環境である OpenCL を使った SWA 実装も行い、CUDA による実装のプログラムとの速度比較を行った結果、配列長に依存せず、CUDA 実装版のほうが OpenCL 版に比較しておよそ 1.5 倍高速であった。OpenCL 使用によりマルチベンダ対応となるメリットに対して、速度が大きく低下するデメリットが大きいため、現状では CUDA による実装が望ましいと考えられる。

(2) その他の実装

その他、complementary sequence(相補鎖), translate nucleotide to amino acid (翻訳), sixpack (ORF 解析), nussinov(動的計画法による RNA の 2 次構造の予測)など、EMBOSS (European Molecular Biology Open Software Suite)に含まれているプログラムについて GPGPU 実装を行った(図 5)。実装したこれらのプログラム解析アルゴリズムについて得られた高速化の程度は、計算量のオーダー(time complexity)に依存する傾向が見られた。即ち、オーダーが高いほど高速化度合が高い傾向を示し、計算量が $O(n)$ の解析では高速化度合が最大でも 2 桁を超えず, GPGPU による並列化の効率が低かった。この原因としては、並列化し得る計算量に対し, メモリ転送など GPGPU を用いたことが原因によるオーバヘッドの割合が大きいことが考えられる。
一方で, O(n) を超える計算量を持つ解析に対しては GPGPU を用いることによるオーバヘッドよりも並列化したことによる計算効率が高いといえる。$O(n)$ のアルゴリズムについて GPGPU により高速化を試みる場合には粒度が粗い実装を行う必要があると考えられる。

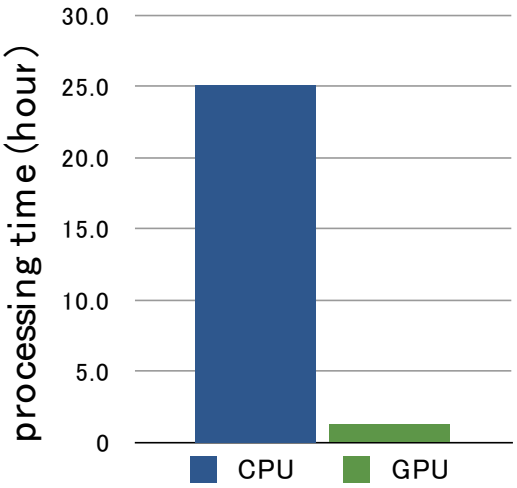

図 5: RNA の二次構造予測アルゴリズム (nussinov)の速度比較

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　0 件）

〔学会発表〕（計 2 件）

①Kazuki Hirata, Nobuyuki Ohta, Noriko Hiroi, Akira Funahashi, Yoshitaka Suetsugu、Design and Implementation of high-performance sequence analysis environment using GPGPU、第 32 回日本分子生物学会年会、2009 年 12 月 11 日、横浜市

②平田一樹，太田信行，上樂明也，広井賀子，舟橋啓，末次克行、MULTI-GPUs による高速配列解析環境の構築、第 33 回日本分子生物学会年会・第 83 回日本生化学会大会　合同大会、2010 年 12 月 9 日、神戸市

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
末次　克行（SUETSUGU YOSHITAKA）
独立行政法人　農業生物資源研究所・昆虫ゲノム研究ユニット・主任研究員
研究者番号：80533471

(2)研究分担者
舟橋　啓（FUNAHASHI AKIRA）
慶應義塾大学・理工学部・生命情報学科・准教授
研究者番号：70324548

(3)連携研究者
（　　）

研究者番号：