様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　５ 月 １３ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００１７０

研究課題名（和文）GPUにおけるプログラムの最適化手法の開発

研究課題名（英文）Development of the optimization technique of the program in GPU

研究代表者
本多　弘樹（Honda, Hiroki）

電気通信大学・大学院情報理工学研究科・教授

研究者番号：２０１９９５７４

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究課題では，高性能・低消費電力な並列GPUコンピューティングを実現する最適化手法について取り組み，次の二つの主要な成果が得られた．
・GPUを用いたネットワーク侵入検知システムにおいて性能低下と消費電力増加を引き起こす条件分岐を多く含むパターンマッチング処理の最適化手法の考案
・複数GPU環境において，最適なタスク分割とタスクのGPUへの割当を行う際に必要となる個々のGPUの性能・消費電力の特性を取得する手法の構築

研究成果の学術的意義や社会的意義
コンピュータに必要とされる性能が高くなっていく一方，コンピュータが消費する電力を少なくすることも求められている．GPUを複数搭載したコンピュータはそれを実現する方式として有望なものである．本研究成果はGPUにおけるアプリケーションプログラムの高性能・低消費電力な実行に不可欠な最適化のための重要項目を明らかにした点において意義がある．

研究成果の概要（英文）：In this study on optimization technique to realize the parallel GPU computing that was high efficiency, low consumption electricity, two next main result was provided.
 - Optimization technique of the pattern matching processing including the condition divergence to cause degradation and consumption electricity increase in Network Intrusion Detection System using GPU
 - Construction of the technique to acquire performance, consumption electricity properties of individual GPU in the most suitable task division, allotment in the parallel GPU environment

研究分野：並列処理，高性能コンピューティング

キーワード：GPU　並列処理　高性能コンピューティング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

(1) 科学技術計算のためのハイパフォーマンスコンピューティングシステム（HPC システム）では，計算アクセラレータとしての GPU を複数備えたハイブリッドアーキテクチャのノードを複数結合することによりハードウェアを構成し，GPU コンピューティング（GPGPU）を可能とするものが多い．また，単一ノードで構成される小型 HPC サーバでも複数 GPU を装備し，GPU コンピューティングを可能とするものもある．これは，GPU により電力及び電力量あたりの計算性能を向上させることができるなどの理由による．HPC システムの環境負荷低減が求められる中で，低消費電力な HPC システムの構築をするには今後も GPU コンピューティングが不可欠と考えられる．

(2) 一方，GPU 利用にあたっては，これまで，CUDA をはじめ，OpenCL，OpenACC，といった抽象度の高いプログラミングモデルも整備されてきており，従来の DirectX や OpenCL などのグラフィックス API によるプログラミングより簡便になってきているが，高性能・低消費電力なプログラムの開発はユーザの知見によるところが多く，これらに習熟していないユーザにとってはそのようなプログラムの開発は困難である．GPU を備えた HPC システムで GPU の効率良い利用を達成していないプログラムを実行することは，そのユーザが GPU による実行時間短縮の恩恵を得られないだけでなく，当該システムで消費する電力や電力量を増大させてしまうという問題がある．

(3) このような中で，応募者らは，CPU 向け並列プログラミング経験者が CUDA などを習得せずとも GPU プログラムの開発をすることを可能とし，また既存の CPU 向け並列プログラムの GPU 上での実行を可能とすることを目指して，CPU 向け並列プログラミングモデルである OpenMP で記述されたプログラムを GPU で実行されるプログラムへ変換することに着眼し，その処理系(参考文献[1])の開発に取り組むとともに，これに関連して，GPU プログラムの高性能化・低消費電力化を達成するための個別の手法を開発してきている．その後，類似の研究[2,3]がなされていることから，このアプローチが有用なものであることがわかる．

(4) 一方，OpenMP4.0 でも GPU へのオフロード処理の記述が可能となり，OpenACC2.0 では並列性やデータ管理などにおいて記述の自由度が上がるなど，個々の GPU プログラミング環境が高度化するとともに，GPU プログラミング環境の多様化も進んでおり，この傾向は今後も続くものと考えられる．このような傾向は熟練ユーザにとってはプログラムの高性能化の可能性を広げることとなるので好ましい状況と言える反面，GPU プログラミングに不慣れユーザにとってはかえって複雑感が増し，GPU コンピューティングの恩恵を受ける敷居を高くしてしまう問題がある．

２．研究の目的

本研究課題では，応募者らがこれまでに行ってきた GPU プログラムの最適化の研究成果を発展させ，タスクの最適分割と最適スケジューリング，CPU-GPU 間/GPU-GPU 間データ転送最適化，カーネル関数内最適化などについてプログラミング環境に依存しない汎用的な手法の考案を進めるとともに，省電力化も有用な要素とし，多様化する GPU プログラミング環境の中にこれらの最適化手法を適用するための技法を実現しようと試みるものである．具体的には次の点に重点をおく．

(1)タスク分割・割当の最適化：
マルチ GPU 環境において並列ネステッドループを分割して複数タスクとする際に，実行時間短縮・省電力化の視点から最適なタスク粒度，割当先(CPU，GPU，複数 GPU)を決定するための最適化手法を明らかにする．

(2)データ転送最適化：
GPU を利用するに当たっては，CPU-GPU 間，および，GPU-GPU 間でのデータ転送が処理性能に大きな影響を及ぼすため，転送データ量の削減，転送回数の削減が求められる．本研究では真に必要なデータのみを転送する最適化手法，テータ転送と演算をオーバーラップさせる最適化手法，タスク分割・割当の工夫などにより転送回数を低減するための最適化手法を明らかにする．

(3)カーネル関数内最適化：
カーネル関数の実行時間や消費電力は，使用するブロック数，スレッド数，レジスタ数，データが配置されるメモリの種類の選択，メモリアクセスパターンや実行中の条件分岐の特性などによって大きな影響を受ける．それぞれのプログラムに応じて最適化を行うための手法を明らかにする．

３．研究の方法

研究は 3 年間かけて行う.

(1) 初年度は，最適化手法の項目に関しての検討から着手する．具体的には，最適化項目としてタスク分割最適化，タスク割当最適化，データ転送最適化，カーネル関数内最適化，メモリ最適化を当初の候補とし，これに加えて GPU プログラミングにおけるハンドチューニング技法をサーベイの上，最適化項目としてさらに対象とすべきものが有るか否かを検討し，さらに検討に際しては GPU のネットワーク機器への応用など，視野を広くとることとする.

(2) 二年度は，最適化項目の検討を進める．具体的には，最適化項目としてタスク分割最適化，タスク割当最適化，データ転送最適化，カーネル関数内最適化，メモリ最適化を当初の候補とし，これに加えて GPU プログラムの実システム上での実行時間及び消費電力の測定により最適パラメタの同定を行う．また GPU テストベッドシステムを構築するとともに，必要となるソフトウエアの導入，消費電力測定機構の実装などを行い有効性検証環境の整備を行う.

(3) 最終年度は，前年度までの最適化項目の検討の結果をもとに，最優先とする最適化項目を絞り込むとともに，その項目についてさらに最適化手法を考案する．また，前年度に導入した GPU テストベッドシステムと異なる GPU による GPU テストベッドシステムを構築するとともに，必要となるソフトウエアの導入，消費電力測定機構の実装などを行い有効性検証環境の整備を行う．さらに研究の総括を行う.

４．研究成果

(1) 初年度には，カーネル関数内最適化手法に関しての検討を行い，その結果，GPU コンピューティングでは GPU におけるカーネル関数内の条件分岐においてスレッドによって分岐先が異なる場合に分岐先とは反対の実行時時間分の待ちが生じてしまい，総合的な性能が低下してしまう問題に着目した.
そして，そのような分岐が生じるアプリケーションとしてネットワーク侵入検知システム(IDS)におけるパターンマッチング処理の GPU での実行を具体例として，サーベイを実施した．具体的には，異なる分岐先となってしまった際の現象の調査，パターンマッチングに用いられる Aho-Corasik 手法の GPU での処理の調査を進めるとともに，それによって生じる問題を解決するための最適化手法について考察を進め，その結果， GPU における重要な最適化項目のひとつであることを明らかにし本研究で対象とすべき項目とした.

(2) 二年度目には， タスク分割・割当最適化項目の検討を進め，その結果，複数の GPU を用いる並列 GPU コンピューティングにおいて実行する並列プログラムの消費電力を考慮に入れてタスク分割・割当の最適化を行うためには，個々の GPU の消費電力特性を明らかにする必要がある点に着眼した.
また，GPU の消費電力特性を明らかにするに際して，同一 GPU を複数備えるシステムでのプログラム実行特性の測定を実施した.
具体的には，東京大学の GPU を備えた大規模 HPC システムを用いてプログラムの特性測定のための環境構築を進めるとともに，ローカルな GPU テストベッドシステムを構築した.
その結果， GPU における消費電力を考慮した最適化においては，個々の GPU の消費電力特性を基に最適化を行うことが必要であることを明らかにし本研究で対象とすべき項目とした.
また，Tesla K40 GPU テストベッドシステムを構築し，その上で必要となるソフトウエアの導入などの環境整備を行った.

(3) 最終年度は，前年度に構築した東京大学の GPU を備えた大規模 HPC システムでのプログラムの特性測定のため実行環境において実行時間及び消費電力の測定により最適パラメタの同定とこれを用いた GPU プログラムの最適化方式を検討した.
また，前年度導入した GPU テストベッドシステムとは異なる TITAN V GPU を搭載した GPU テストベッドシステムを構築し，その上で必要となるソフトウエアの導入などの環境整備を行った.
その結果，複数 GPU における消費電力を考慮したタスク分割・割当の最適化においては，個々の GPU の消費電力特性を同定し，そのうえでそれを基に最適化を行うことが必要であることを明らかにした.

＜引用文献＞

[1]大島聡史，平澤将一，本多弘樹：既存の並列化手法を用いた GPGPU プログラミングの提案，情報処理学会研究報告(ARC-175)，pp.7-10(2007).
[2]S.Lee，S.-J.Min，R Eigenmann: OpenMP to GPGPU: A compiler framework for automatic translation and optimization，ACM SIGPLAN symposium on Principles and Practice of

Parallel Programming, pp.101-110(2009).
[3]D.Unat: Mint: An OpenMP to CUDA Translator, GPU Technology Conference (2010).

５．主な発表論文等

〔雑誌論文〕（計０件）

〔学会発表〕（計０件）

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
なし

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし