【1】

氏名	池田孝利（いけだたかとし）
博士の専攻分野の名称	博士（情報科学）
学位記番号	第　24919　号
学位授与年月日	平成23年9月20日
学位授与の要件	学位規則第4条第1項該当 情報科学研究科コンピュータサイエンス専攻
学位論文名	**畳み込み演算を応用したアプリケーションのＧＰＵによる高速化に関する研究**
論文審査委員	（主査） 教授　萩原　兼一 （副査） 工学研究科教授　八木　哲也　　教授　増澤　利光 准教授　伊藤　文彦

論文内容の要旨

近年のプログラム可能なGPU（Graphics Processing Unit）は，高い計算性能，メモリバンド幅を持ち，CPUを凌駕する早さで性能が向上している．そのため，有用な計算資源として注目されている．

高速な描画処理の実行を目的としてきたGPUを，3次元グラフィクス用途に限定せず，幅広い用途に応用する取り組み（GPGPU：General-Purpose computation on GPUs）が行われるようになり，多くの成果が見られるようになっている．さらに先進的なアーキテクチャを持つGPUでは，グラフィクスの概念とは独立して，単一命令マルチスレッド並列プロセッサとしてプログラミングが可能である．加えて専用のプログラミング環境も整備されてきているため，多くの分野の問題で，GPUを用いた高速処理の取り組みが一層盛んになっている．

一方，広く応用されている演算に畳み込み演算がある．畳み込み演算は計算負荷が高いものの高い並列性をもつため，GPUを用いた並列化により高い高速化効果が期待できる．本研究では，畳み込み演算を応用したアプリケーションのGPUによる高速化として，（1）2次元の畳み込み演算を応用したアプリケーションとしてガウスフィルタ，（2）3次元の畳み込み演算を応用したアプリケーションとしてネオコグニトロンのGPUによる高速化を行った．

一方，広く応用されている演算に畳み込み演算がある．畳み込み演算は計算負荷が高いものの高い並列性をもつため，GPU を用いた並列化により高い高速化効果が期待できる．本研究では，畳み込み演算を応用したアプリケーションのGPUによる高速化として，（1）2次元の畳み込み演算を応用したアプリケーションであるガウスフィルタ，（2）3次元の畳み込み演算を応用したアプリケーションであるネオコグニトロンのGPU による高速化を行った．

現在，GPUには，3Dグラフィクスの高速処理に特化した伝統的アーキテクチャ，および，単一命令マルチスレッド並列プロセッサとしてプログラミングが可能な先進的アーキテクチャの2種類がある．後者のGPUは，並列プログラムの高速化に有用な共有メモリをもつ．また，前者のGPUとしても使うことができる．幅広く適用可能な手法を目指した（1）では，前者のGPUを用い，より高速な実装を目指した（2）では，共有メモリを使用するため，後者のGPUと専用のプログラミング環境を用いた．

本論文ではまず（1）について述べる．プログラム可能なGPUは，3次元グラフィクスを高速にレンダリング処理するため，それぞれプログラム可能な2種類のプロセッサ，頂点シェーダー（VS：Vertex Shader），および画素シェーダー（PS：Pixel Shader）から構成されている．VSとPSはパイプライン構成であるにも関わらず，多くのGPGPUプログラムはPSのみを多用している点に着目した．このような実装では，PSの計算負荷が高くなり，パイプライン実行における性能ボトルネックになり得る．研究（1）の目的は，この性能ボトルネックを解消し，GPGPUプログラムの性能を向上することにある．その実現のために，与えられたVSとPSのアセンブリ言語プログラムに対して，PSプログラムにおける命令の一部をVSプログラムへ移動する手法を提案する．提案手法は，CPUとGPU間の入出力仕様を変えることなく命令を移動するための条件を明らかにし，VSおよびPSにおける計算資源の制限を考慮して命令を移動する．提案手法を2次元畳み込み演算のアプリケーションであるガウスフィルタプログラムに適用した結果，GPUにおける計算時間をおよそ19ミリ秒から11ミリ秒に短縮でき，その性能を向上できた．

次に，（2）について述べる．単一命令マルチスレッド並列プロセッサとしてもプログラミング可能なGPUを用い，手書き文字認識を行うニューラルネットモデルであるネオコグニトロンを高速化した．ネオコグニトロンは変形や位置ずれに強く，高い認識性能を持つが，計算コストが高く，高速化が求められている．GPUを用いた並列計算により高速化が期待できる反面，その計算には3次元畳み込み演算が含まれており，多くのメモリアクセスを要するといった問題がある．研究（2）では，ネオコグニトロンの処理時間の95%以上を占めるS層応答計算をGPUを用いて高速化し，性能を検証した．GPUではコストが特に高いグローバルメモリへの書き込みを最小化し，共有メモリの利用効率を高めた高速化では，CPU実装に対して約8～20倍高速化した．

論文審査の結果の要旨

本研究は、最近、並列処理の分野で注目されているGPU（Graphics Processing Unit）を、本来の用途である 3次元描画処理ではなく、汎用処理に適用する取り組み（GPGPU：General-Purpose computation on GPUs）に関するものであり、以下の(1)および(2)の内容から成る。

GPGPUが研究されだした当初のGPUアーキテクチャは、3次元の対象物を高速にレンダリング処理するため、2種類のプロセッサ（頂点シェーダー（VS：Vertex Shader）および画素シェーダー（PS：Pixel Shader））がパイプライン接続された構成である。VSおよびPSはそれぞれプログラム可能ではあるが、特に汎用処理を目指したものではなかったので、GPGPUは描画処理の副作用として汎用問題を計算する形態であった。

現在の主流アーキテクチャは、VSとPSを統一したUS(Unified Shader)で構成されており、描画機能を用いずに汎用問題を数百個のUS上でスレッド実行する形態で並列処理できるようになっている。また、メモリ構成は大容量であるがアクセスが低速なグローバルメモリGMと、小容量であるが高速な共有メモリSMの２階層となっているので、効率的なプログラムとするためには、この２階層のメモリを効果的に使い分けることが重要である。

研究(1)は前者のアーキテクチャのもとでの研究であり、研究(2)は後者のもとでの研究である。

（1）前者のGPUアーキテクチャを用いたGPGPUのプログラムでは、一般にPSのみを多用して実装されている。したがって、PSの計算負荷が高くなり、パイプライン実行における性能ボトルネックになり得る。この性能ボトルネックを解消するために、与えられたVSとPSのアセンブリ言語プログラムに対して、PSプログラムにおける命令の一部をVSプログラムへ移動する手法を提案する。CPUとGPU間の入出力仕様を変えることなく命令を移動するための条件を明らかにし、VSおよびPSにおける計算資源の制限を考慮して命令を移動する。この提案手法は一般的なものであるが、2次元畳み込み演算のアプリケーションであるガウスフィルタプログラムに適用した結果、GPUにおける計算時間をおよそ19ミリ秒から11ミリ秒に短縮でき、その性能を向上できた。

（2）後者のGPUアーキテクチャを用いて、手書き文字を認識するニューラルネットモデルであるネオコグニトロンを高速化した。ネオコグニトロンは変形や位置ずれに強く、認識性能が高いが、計算量が多く、高速化が求められている。この計算には多くの3次元畳み込み演算が含まれているので、多くのメモリアクセスを必要とする。ネオコグニトロンの処理時間の95%以上を占めるS層応答計算に関して、低速なGMへの書き込みを最小化し、高速なSMの利用効率を高めることにより、CPU実装に対して約8～20倍高速化した。

これらの研究成果は、並列処理分野において有用なものであり、汎用性をもつものである。よって、博士（情報科学）の学位論文として価値のあるものと認める。