審 査 の 結 果 の 要 旨

氏　名　　羅　成

本論文は「Research on Performance Optimization Methods based on Performance Analytical Modeling and Communication Latency Hiding in GPU（GPU の性能分析モデリングと通信レイテンシの隠蔽に基づく性能の最適化方法に関する研究）」と題し，6 章から成る．簡明かつ精度の高い数学的モデルに基づいた GPU（Graphic Processing Unit）処理の最適化を目指した．まず最適化に必要な処理時間予測のために，従来より高精度な予測モデルを提案した．さらに，CPU と GPU との間の通信レイテンシを，GPU 内部の計算との並列処理により削減する TPSM（Task Partition and Scheduling Method）手法を提案した．

従来，GPU 処理時間予測においては，予測のための数学的モデルやシミュレータが用いられていたが，精度の低いモデルに基づいていたため，予測精度も低いものとなっていた．また最適化においては，処理の分割に基づく手法，および通信処理と計算処理の並列化が従来より研究されているが，GPU に特徴的な CPU との双方向通信が考慮されておらず，GPU の最適化としては不十分であった．

そこで本研究においては，GPU 処理を計算タスクとメモリアクセスタスクに分け，それぞれについて個別の性能モデル（CPD と MPD）を確立し，それらを統合した，GPU 処理全体の予測モデルを用いて処理時間予測を行う手法を提案している．本手法は従来のものより正確な予測を可能としている．次に最適化については，アプリケーションの種類に応じて異なる方針によりデータ送受信処理と計算処理のそれぞれを分割し，各部分毎に並列実行することにより，全体の処理時間を軽減する手法を提案している．さらにシングル GPU，対称的マルチ GPU，および非対称的マルチ GPU のそれぞれの場合毎の工夫も示している．本手法は並列処理部分を可能な限り多くすることにより，レイテンシの大幅な削減を実現している．

本論文は以下の 6 つの章から構成されている．

第1章では研究の概要を述べている．続いて第2章では，前提となる背景知識を整理している．具体的には，GPU のアーキテクチャと処理の枠組み，Ocelot と呼ばれる GPU アプリケーション処理系，および処理の分割に基づく最適化手法を紹介している．

第3章では，GPU の性能予測モデル，すなわち与えられたプログラムに対する処理時間を予測するためのモデルを提案している．詳細は次の通りである．まず予測モデルを，計算タスクの実行における並列度を算出する CPD モデルと，並列実行可能なメモリアクセス処理数を算出する MPD モデルに分け，それぞれを確立する．CPD，MPD 各モデルのパラメータの推定のために，プログラムを Ocelot により PTX と呼ばれる仮想機械のコードに変換し分析する．CPD，MPD 各モデルにより，計算処理とメモリアクセス処理の各予測時間を算

出した後，それらの関係に応じて場合分けされた計算式により，全体の処理時間を計算する．さらに，各種の GPU とベンチマークを用いて，提案モデルの評価を行ない，総じて高い精度で予測できていることを明らかにしている．

第4章では，GPU 利用アプリケーションの最適化のために，CPU/GPU 間の通信レイテンシを軽減する手法を提案している．現在のアーキテクチャでは，CPU と GPU 間のデータ送受信のそれぞれと GPU の計算処理との計 3 者の並列実行可能性を利用する．提案手法では，アーキテクチャをシングル GPU，対称的マルチ GPU，および非対称的マルチ GPUに分類し，シングル GPU の場合の手法を次のように示している．まずデータ送受信と計算の各処理時間の大小に応じて，アプリケーションを 6 種類に分ける．その上で，各種類ごとの方針でデータ送受信処理と計算処理のそれぞれを分割し，各部分毎に並列実行することにより，全体の処理時間を軽減する．その際にどの分割単位でも処理時間の大小関係を保つようにするための工夫を行なっている．対称的マルチ GPU については，各 GPU へのデータ送受信の時間の割り当て方の導出を，2次元空間上の凸包を利用して行い，非対称的マルチ GPU については，対称的の場合とは異なる方法で最適スケジューリングを求め，またGPU 間の負荷分散の最適化を第3章の性能予測モデルの利用などにより行うことで，シングル GPU の場合を拡張している．さらに各種の GPU とベンチマークを用いて評価を行ない，アプリケーションにより差はあるが，総じて提案手法により高速化できており，マルチ化の効果もあることを明らかにしている．

第5章では，関連研究との比較を行なっている．

第6章では，研究結果をまとめ，今後の課題と展望を述べている．

以上のように本論文は， 簡明かつ精度の高い数学的モデルに基づいた GPU（Graphic Processing Unit）処理の最適化という目標に向けて，着実に技術を進歩させている．特に，GPU 処理を計算タスクとメモリアクセスタスクに分けて性能予測を行うというのは，従来にない新たなアプローチである．また最適化については，アプリケーションの種類に応じた個別の方針によりデータ送受信処理と計算処理のそれぞれを分割し，各部分毎に並列実行することにより，全体の処理時間を軽減するという，やはり従来にない新たなアプローチをとっている．

よって本論文は博士（情報理工学）の学位請求論文として合格と認められる．