審 査 の 結 果 の 要 旨

氏　　名　　額田 彰

高速フーリエ変換（FFT）は，コンピュータ科学・計算科学なかでも特異な重要性を持つアルゴリズムである．離散フーリエ変換としての自然な定義では入力サイズ N に対して O(N^2) の計算量がかかる計算を，FFT は O(N log N) の計算量で実現する．この圧倒的な性能向上のため，数値計算や信号処理のみならず，多数のコンピュータ科学のアルゴリズムの基礎となっている．FFT が高速化されれば，多くのアプリケーションが多大な恩恵を受ける．

本論文では，近年高性能計算アーキテクチャとして非常に高い注目を集めている GPU (グラフィックプロセッシングユニット)を用いた FFT の高性能実装技術について論じている．GPU はもともとグラフィックス用の付加プロセッサとして設計されていたが，近年はグラフィックス以外の汎用計算にも使えるようにアーキテクチャおよびプログラミングモデルが改善されている．GPU は多数の演算コアを持ち，さらに各演算コアは大規模な SIMD 型の並列処理を行う．このため GPU は，大規模 FFT のような大規模並列性を持つ計算において CPU をはるかに上回る性能を達成する．

しかし，GPU において実際に高い性能を実現するために必要なのは，並列性だけではない．GPU の演算能力は非常に高いが，FFT は入力サイズ N に対して計算量が O(N log N) しかないため，メモリと演算器の間のデータ転送が所要時間の主要な要因となる．従って，メモリアクセスを最適化することが高性能実装の本質になる．また，複数 GPU を用いる場合は GPU 間でのデータ転送，GPU を複数搭載する GPU クラスタを用いる場合にはノード間のデータ転送が性能を決める重要要因となる．

本論文では，単一 GPU，複数 GPU，GPU クラスタのそれぞれの場合に対して有効な新規手法を開発しており，それぞれ世界最高水準の FFT を実現している．本論文は以下の 8 つの章から成る．

第１章は導入部であり，課題の表明，本研究の貢献を概論している．

第２章は背景説明であり，FFT および GPU についての基本的知識を説明している．

第３章は単一 GPU のための 3 次元高速 FFT のアルゴリズムとして，Bandwidth Intensive 3-D FFT を提案，評価している．通常，メモリアクセスの最適化のためにはブロック化が行われるが，FFT をブロック化した six-step FFT には 3 回の行列転置

演算が必要である．しかし GPU には連続したスレッドが連続したアドレスにアクセスするコアレッシングと呼ばれる条件が成り立たないデータアクセスは格段に遅くなってしまうという性質がある．しかし行列転置はこのような GPU の性質に向かず，行列転置のために FFT は遅くなってしまう．これに対し本論文では，あえてブロック化によるメモリアクセス最適化を用いず，コアレッシングを優先するアルゴリズムを提案している．これによりベンダー提供の CUFFT3D の 3 倍以上という性能を達成した．

第４章では，単一 GPU のための FFT ライブラリの自動チューニングによる最適化が述べられている．基底とその順序の選択，スレッドブロック数の選択，シェアードメモリのパディングについて網羅的検査によるオフライン自動チューニングを行い，2，3，5 のべきを含む FFT に対して常に高性能を達成する．

第５章では，単一 GPU の FFT のためのワープレベル最適化について述べている．GPU は SIMD 型計算機であるが，多数の基底を含む FFT では必要とするスレッド数が異なる．実行のためには必要数の最大値にあわせてスレッドを準備する必要があるが，一部の基底では不要なスレッドが発生するため，それによるオーバーヘッドを削減する手法を論じている．

第６章では，複数 GPU による FFT のためのデータ交換の最適化が論じられている．CPU と GPU は PCI Express により接続されているが，単一ノードに実装された複数 GPU を用いて FFT を行う場合，GPU 間でのデータ通信がオーバーヘッドとなる．本研究では，GPU 間直接通信，および PCI Express レーンの衝突を避けるスケジューリングにより，高い実効性能を実現する手法を提案している．

第７章では，GPU クラスタを用いた FFT のためのデータ交換の最適化手法が論じられている．本研究で用いたプラットフォームは GPU スーパーコンピュータである TSUBAME 2.0 である．TSUBAME 2.0 はノード間を 2 組の QDR x4 インフィニバンドで接続している．本論文では，ノード間データ転送と CPU-GPU 間データ転送を最適なチャンクサイズでパイプライン化するのみならず，2 組のレールの使用方法をオンラインで最適化することにより，高いスケーラビリティを達成している．

第８章は本論文の貢献と議論とをまとめている．

以上のように，本論文は GPU を用いた FFT について，単一 GPU，単一ノード複数 GPU，複数ノード GPU のいずれについても新規性のある手法を開発し，世界最高水準の高性能を達成している．今後プロセッサ性能のさらなる向上により，所要時間の主要部分はデータ移動となる傾向は一層強まると考えられ，メモリ律速な FFT というアルゴリズムにおけるこれらの多大な成果は，今後の高性能計算・スーパーコンピューティング，ひいては計算科学の発展に寄与することが極めて大である．

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。