修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工学研究科　情報・通信工学専攻　博士前期課程		
氏　　名	伊藤　一輝	学籍番号	1531009
論 文 題 目	複数 GPU を用いた DS-CUDA による P2P 機能使用時の性能評価及び最適化		

要　旨

GPU は数値流体シミュレーションやディープラーニングの分野に応用する GPGPU として多岐に渡る発展を見せ広く普及してきた． GPU は豊富な計算資源を所有しており大規模な並列演算が可能で年々性能が飛躍的に向上している． しかし，大規模なシミュレーションや数値流体計算問題をアプリケーションプログラムとして実行するには単体の GPU ではメモリなどの計算資源が不足する． 通常は並列コンピューティングで用いられる MPI と GPGPU で用いられる CUDA を利用することで GPU 間あるいはノード間での通信を行い不足する計算資源を補う． また当研究室で扱っている DS-CUDA はネットワークに接続されたサーバ上の GPU を仮想化するミドルウェアで，クライアント側でソフトを書き換えることなくリモートの GPU の計算資源を用いた GPGPU によって同様の問題を解消することが可能である． しかし，大規模なデータに対し高並列化のプログラムを実装するとレイテンシが大きくなり通信速度が向上しないという問題が新たに発生する． 対策としては DS-CUDA API の dscudaMemcopies()を利用することで，サーバ上の GPU 間の通信を Peer to Peer(P2P)で並列に処理することで高速化が可能になっている．

そこで本研究では，3 次元 Euler 方程式から Rayleigh-Taylor 不安定性の成長シミュレーションを解く数値流体計算用のコードを複数 GPU を用いて最適化を行った． さらに DS-CUDA に搭載されている P2P 機能を用いてノード間の通信をサーバ側だけで行う通信の最適化を行った． アプリケーションプログラムは最大 8 つの GPU を用いて Native 時，DS-CUDA を利用した InfiniBand ネットワーク使用時，DS-CUDA を利用した P2P 機能使用時における性能評価を行った． 予備実験では，P2P 機能の通信速度を測定し転送データ量やパラメータ数，メモリアクセスの手法を変えることでどのくらいの転送速度が出るか測定した． その結果，連続領域で 4 台のノード間における P2P 機能を使用した場合に最大で約 5.08 倍の通信時間が高速化されており，本研究で利用する数値流体計算用のコードを想定した検証では，8 台のノード間において最大で約 1.95 倍の通信時間高速化が見込まれることを示した．

実際のアプリケーションコードでは 8GPU における P2P 機能使用時に 256^3 グリッドサイズにおいて InfiniBand ネットワーク使用時と比較して約 2.56 倍高速化された． また，通信時間については P2P 機能使用時 512^3 グリッドサイズにおいて同様の比較を行い約 72.51%の通信時間を削減した． これらにより，DSCUDA API の P2P 機能はサーバ側におけるノード間 GPU 通信において転送データ量を大きくすること，高並列化することが通信時間削減に有効であることを示した．