

HPC 向け変動精度計算基盤および複数 GPU 間ストリーム処理機構

の開発

Development of Computing Platform based on Transprecision and Inter-GPU Streaming Mechanism for HPC

埜 敏博

東京大学

1. 研究目的

FPGA は、AI を始めとした特定用途のアクセラレータとして活用され始めている。OpenCL などの高位合成が可能な開発環境の普及に伴い、限られた電力制約の下 HPC アプリケーションに対してカスタマイズ可能なアクセラレータとして FPGA への期待が高まっている。そこで我々は、1. 変動精度による HPC アプリケーションの精度評価、および性能・電力評価、2. 複数ノード GPU 間のデータ転送に伴うストリーム処理機構の実現、を目指して研究を行っている。

本研究では、宇宙物理のアプリケーション、数値計算などを対象に、FPGA へのオフローディングの実現と変動精度の適用を目指している。さらに、筑波大と理研 R-CCS で開発されている、OpenCL で記述可能な FPGA 間通信フレームワーク CIRCUS (Communication Integrated Reconfigurable CompUting System)の導入に加えて、ホスト上の GPU との連携についても検討する。

2. 研究成果の内容

本年度は、宇宙物理アプリケーションとして、N 体シミュレーションを対象に、可変精度を考慮に入れ FPGA の特性を活かした演算オフローディングについて検討した。同一の N 体シミュレーションの中でも、粒子数が多いケースでは、並列性が高く、GPU で実行することにより高い性能を得られるが、粒子数が少ないケースでは、並列性は少ないものの、ステップ数が多くなるため、相対的に実行時間が大きくなることがわかっている。このようなケースについては、FPGA を用いてパイプライン処理を行うことで高い性能が期待できる。

一方、HBM2 搭載 FPGA である Intel Stratix10 MX を用いて、実現に向けた OpenCL カーネルの記述方法やメモリバンド幅性能について確認した。HBM2 を使うための BSP (Board Support Package)として、32 の独立なメモリチャネルを扱う構造になっており、これらを統合して、カーネルに応じたメモリ容量やバンド幅を得られるようにする必要があり、それを実現するモジュールの開発が必要であることがわか

った。さらに、倍精度、単精度、半精度と精度を変えて行列積計算を実行したところ、単精度に適したハードウェアを持っているため、単精度が最も使用ハードウェア量は少なかったが、半精度にすることで使用リソースが減少し動作周波数が向上したため単精度に比べて 20%性能の向上が見られた。倍精度についても、0.53 倍と、想定よりやや高い性能であった。

3. 学際共同利用が果たした役割と意義

本プロジェクトでは、GPU と FPGA との協調計算を目指しており、GPU と FPGA の複合クラスタとしては筑波大 CCS の Cygnus 以外に存在しない。また、筑波大、理研 R-CCS と協力して研究を進める上での共通プラットフォームとしても Cygnus が最適な環境である。そのため、本研究の遂行には学際共同利用が必要不可欠である。

4. 今後の展望

これまでは Cygnus 以外の環境を主に使って基本的な確認を実施した。来年度から Cygnus を使い、実アプリケーションのカーネルのオフロードに向けて実装を行なっていく予定である。

5. 成果発表

(1) 学術論文 なし

(2) 学会発表

埜 敏博, 三木 洋平, 宇宙物理アプリケーションのための FPGA 演算オフローディングの検討, 情報処理学会ハイパフォーマンスコンピューティング研究会, 2019 年 12 月

(3) その他 なし

使用計算機	使用計算機 に○	配分リソース*	
		当初配分	追加配分
Cygnus	○	5000	
Oakforest-PACS			

※配分リソースについてはノード時間積をご記入ください。