受付 ID	15a-45
分野	HPCS

密結合演算加速機構アーキテクチャに向けた アプリケーションの開発と性能評価

塙　敏博
東京大学情報基盤センター

１．研究目的

GPU に代表される演算加速装置は、その高い演算性能とメモリバンド幅、電力当たり性能のため HPC 用途のクラスタに搭載され広く用いられている。しかし、クラスタ上の演算加速装置間の通信では，これまでホストメモリを介した転送が必要であり，特に小データの転送ではレイテンシがボトルネックとなる。そこで，レイテンシとバンド幅の改善を目指した独自開発の演算加速装置向け専用相互結合機構 TCA(Tightly Coupled Accelerators)の開発を行っている。

本研究では，マルチノード・マルチ GPU を用いたアプリケーションとして，QCD や宇宙物理のアプリケーション，数値計算などを対象に， TCA に向けた変更を継続して行う。TCA はノードを超えた GPU 間通信のレイテンシを大きく改善することが可能であり，アプリケーションの強スケーリングにおける性能改善に期待されている。さらに，TCA と InfiniBand からなる複合ネットワークにおける通信の最適化について検討を継続する。TCA によるミニクラスタはノード数が制限されるため，それを超える規模のアプリケーションでは，GPU 間の通信には，従来と同様 InfiniBand を経由した MPI 通信を使って記述する。そこで TCA と MPI の組み合わせ手法についても検討する。

２．研究成果の内容

今年度は、まず TCA における基本通信性能の評価を行った。ノードをまたぐ GPU 間での TCA による通信性能は 2.0μs であり、既存の MPI 実装と比較して 2 倍以上高速である。また、姫野ベンチマークの通信部分に TCA を用いた場合、小サイズの問題においては 60%以上の性能向上が見られた。また TCA において集団通信関数を実装し、CG 法や FFT などのアプリケーションに適用して性能評価を行った。その結果、小さいデータサイズにおいては低レイテンシの効果が高く、MPI に比べて高い性能を得ることができた。

また、「メニーコアおよび演算加速機構を持つクラスタシステム向け並列プログラ

ミング言語の開発と評価」プロジェクトと共同で、TCA サブクラスタを超える通信を実現するため、TCA と InfiniBand によるハイブリッド通信について検討し評価を行った。隣接ノード間の袖領域通信を、TCA と MPI+IB の特性に合わせて最適な割り当てを考慮することで、性能の向上が得られた。記述の複雑さを解消するために Xcalable ACC のフレームワークを用いることで、指示文のみの追加で実現できた。

一方で、GMPI と呼ぶ GPU セルフ MPI の提案および実装を行った。クラスタでは GPU 間での通信が必須であるが、一般的に GPU 上のデータであってもホスト CPU 上で MPI によって通信処理を行う必要がある。そのため，通信が発生する毎に GPU 上の CUDA カーネルからホストに一旦制御を戻す必要があり，カーネル関数の起動や同期に伴うオーバーヘッドが生じる。特に並列処理における通信粒度が細かいほど，カーネル関数の起動回数も増え，オーバーヘッドも増加する。それだけでなく，プログラミングコストが高く，CPU 向け MPI プログラムを GPU 並列化する場合にソースコードが煩雑になりやすいといった生産性の低下も問題となる。これらの問題を解決するため，GPU カーネル内から MPI 通信の起動を可能とする並列通信システム “GMPI”を提案・開発した。これまでに GMPI の実装と，Ping-Pong 通信および姫野ベンチマークの性能評価を行い、コード量の削減および並列処理効率の向上が可能になった。

３．学際共同利用として実施した意義

本プロジェクトでは TCA における通信機能の実現および性能評価を目的としていたため、主として HA-PACS/TCA を用いた。TCA を搭載した GPU クラスタとしては唯一の環境であり、研究の遂行には学際共同利用プログラムが必要不可欠であった。本研究の成果は，TCA アーキテクチャ，ならびに HA-PACS/TCA における効率的な通信の実現につながっており、他の HA-PACS/TCA 利用者に対して、TCA のみならず、MPI における最適なパラメータなどフィードバックされている。

４．今後の展望

本プロジェクトにおいて基本性能に加えて実アプリケーションにおいても TCA の有用性が確認されている。今後のエクサスケールシステムに向けて、演算加速機構とそれを支える通信機構の要素技術開発に大きく貢献すると考えられる。

５．成果発表

(1) 学術論文

A) 藤田 典久、藤井 久史、塙 敏博、児玉 祐悦、朴 泰祐、藏増 嘉伸、Mike Clark:「GPU 向け QCD ライブラリ QUDA への TCA アーキテクチャの適用」、情報処理学会論文誌(コンピューティングシステム)、Vol. 8, No.2, pp. 25–35, 2015 年 6 月

B) 松本 和也, 塙 敏博, 児玉 祐悦, 藤井 久史, 朴 泰祐:「密結合並列演算加速機構 TCA による GPU 間直接通信における Collective 通信の実装と性能評価」、情報処理学会論文誌コンピューティングシステム(ACS), Vol. 8, No. 4, pp. 36−49, 2015 年 11 月

C) 小田嶋 哲哉, 朴 泰祐, 塙 敏博, 児玉 祐悦, 村井 均, 中尾 昌広, 田渕 晶大, 佐藤 三久,「アクセラレータ向け並列言語 XcalableACC における TCA/InfiniBand ハイブリッド通信」, 情報処理学会論文誌コンピューティングシステム(ACS), Vol. 8, No.4, pp. 61−77, 2015 年 11 月

(2) 学会発表

A) 松本 和也, 他:「密結合並列演算加速機構 TCA を用いた GPU 間直接通信による Collective 通信の実装と性能評価」, ハイパフォーマンスコンピューティングと計算科学シンポジウム(HPCS2015)論文集, pp. 120--128, 2015 年 5 月.

B) Kazuya Matsumoto, et al.: “Implementation of CG Method on GPU Cluster with proprietary Interconnect TCA for GPU Direct Communication,” International Workshop on Accelerators and Hybrid Exascale Systems (AsHES2015), pp. 647-655, May 2015.

C) Toshihiro Hanawa, et al.: “Improving Strong-Scaling on GPU Cluster Based on Tightly Coupled Accelerators Architecture,” IEEE Cluster 2015 (short paper), pp. 88--91, Sep. 2015.

D) Tetsuya Odajima, et al.: “Hybrid Communication with TCA and InfiniBand on A Parallel Programming Language XcalableACC for GPU Clusters,” Workshop on Heterogeneous and Unconventional Cluster Architectures and Applications (HUCAA) 2015, pp. 627--634, Sep. 2015.

E) Toshihiro Hanawa, et al.: “Evaluation of FFT for GPU Cluster Using Tightly Coupled Accelerators Architecture, Workshop on Heterogeneous and Unconventional Cluster Architectures and Applications (HUCAA) 2015, pp. 635--641, Sep. 2015.

(3) その他

A) 佐藤 賢太, 藤田 典久, 塙 敏博, 朴 泰祐: 「密結合演算加速機構 TCA における Verbs 実装による MPI 環境の実現」, 情報処理学会研究報告, 2015-HPC-150(42), pp. 1-9, 2015 年 8 月

B) 松本 和也, 塙 敏博, 藤田 典久, 桑原 悠太, 朴 泰祐: 「密結合並列演算加速機構 TCA による並列 GPU コードの性能予測モデル」, 情報処理学会研究報告, 2015-HPC-150(35), pp. 1-8, 2015 年 8 月

C) 桑原 悠太, 塙 敏博, 朴 泰祐: 「GMPI : GPU クラスタにおける GPU セルフ MPI の提案」, 情報処理学会研究報告, 2015-HPC-151(12), pp. 1--8, 沖縄産業

支援センター, 2015 年 9 月

D) 佐藤 賢太, 藤田 典久, 塙 敏博, 松本 和也, 朴 泰祐, Khaled Ibrahim:「密結合並列演算加速機構 TCA による GPU 対応 GASNet の実装」, 情報処理学会研究報告, 2016-HPC-153(28), pp. 1--10, 道後温泉, 2016 年 2 月

使用計算機	使用計算機に○	配分リソース※
HA-PACS	○	32
HA-PACS/TCA	○	128
COMA	○	72
※配分リソースについては 32node 換算時間をご記入ください。		