平成 30 年度　利用報告書

課題名(タイトル):

通信・ファイル I/O に関する研究開発

利用者氏名:

〇石川　裕*、高木　将通*、堀　敦史*、Gerofi Balazs*、森江　善之*、畑中　正行*、小倉　崇浩*、Martsinkevich Tatiana*、亀山　豊久*

理研における所属研究室名:

***計算科学研究センター　フラッグシップ２０２０プロジェクト　システムソフトウェア開発チーム**

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

本課題では、FX100 が有する Tofu2 ネットワークインターフェイスを用いてポスト「京」向けに開発している通信・ファイル I/O 機能のプロトタイプ実装および性能検証を実施することを目的とする。我々は、ポスト「京」向けの OS、MPI ライブラリ等のシステムソフトウェアの研究開発を行なっている。このようなシステムソフトウェアの開発は、ターゲット計算機のハードウェア開発と並行して進めている。ポスト「京」で採用されるネットワークインターフェイス TofuD は、FX100 の Tofu2 ネットワークの後継版であり互換性を有する。FX100 を用いることによりポスト「京」向けの通信に関係するソフトウェア実装実験が可能となる。

２．　具体的な利用内容、計算方法

今回、ポスト「京」に向けて MPI 通信ライブラリで規格化が進んでいる永続型集団通信の設計、評価を行った。

Tofu2 および TofuD にも適用可能な Open Fabrics Interfaces (OFI)が提供する fi_trigger と呼ばれるネットワークオフロードインターフェイスを用いて broadcast, allgather などを例に永続型集団通信アルゴリズムをオフロード可能であることを示した。

３．　結果

昨年度(平成29年度)開発した Trinaryx3 アルゴリズムの broadcast 永続型集団通信の実装を、OFI のfi_triggerインターフェイスで書き直した。

barrier、allgather, alltoallなどの永続型集団通信はbutterflyアルゴリズムで実装できる。Tofu2のオフロード機能を用いて実装可能かどうかの確認を行った。まず、Tofu2 インターコネクト・ハードウェアのオフロード機能をモデル化し、butterfly アルゴリズムの通信の各段で通信相手が変わってゆく場合に、通信ハードウェア資源が O(1)になるような手法を設計した。さらにこの手法に基づいて、butterfly アルゴリズムの barrier、allgather (recursive doubling) および alltoall でプロトタイプ実装を行った。

これらの実装により、演算を伴うリダクションを除く永続型集団通信をOFI のfi_triggerインターフェイスを用いてオフロード実装が可能であることを確認した。なお、Tofu2にはバリア同期時にリダクション同様の演算を行う機能が提供されており、リダクション型集団通信はこの機能で実現できる。

４．　まとめ

昨年度のツリー型のアルゴリズムの実装であった Trinaryx3 アルゴリズム に加え butterfly アルゴリズムを用いた集団通信も O(1)の通信ハードウェア資源消費で実装可能であることを確認し、省資源型のオフロード可能な永続型集団通信関数群を提供できる見込みがついた。

５．　今後の計画・展望

今回実装していない永続型集団通信についても同様に設計、評価を行う。これらの設計、評価を元にポスト「京」向けの永続型集団通信の適切な実装を行う予定である。

平成 30 年度 利用研究成果リスト

【口頭発表】

森江　善之, 畑中　正行, 高木　将通, 堀　敦史, 石川　裕, “Open Fabrics Interfaces オフロード API を用いた MPI 永続型集団通信の設計と評価”, 情報処理学会研究報告, 2018(HPC-168), 2018 年 3 月, 石川県加賀市