FX700 によるデータ通信プロトコル HpFP の高速化

村田 健史 [1),3)], 柿澤 康範 [2)], 川鍋 友宏 [3)], 深沢 圭一郎 [1)], 高木 文博 [2)], 水原 隆道 [2)]

1) 京都大学 学術情報メディアセンター
2) 株式会社クレアリンクテクノロジー
3) 情報通信研究機構 総合テストベッド研究開発推進センター

murata4stars@gmail.com

High-speed data transfer via HpFP (High-performance and Flexible Protocol) using FX700 supercomputer

Ken T. Murata [1),3)], Yasunori Kakizawa [2)], Tomohiro Kawanabe [3)], Keiichiro Fukazawa [1)], Ayahiro Takaki [2)], Takamichi Mizuhara [2)]

1) Academic Center for Computing and Media Studies, Kyoto Univ.
2) CLEALINK TECHNOLOGY Co., Ltd.
3) ICT Testbed Research and Development Promotion Center, National Institute of Information and Communications Technology

概要

京都大学学術情報メディアセンターの富士通スパコン FX700 を用いたデータ通信プロトコル HpFP（High-performance and Flexible Protocol）の高速化を試みた。HpFP は筆者らのグループが開発した TCP 互換プロトコルであり、これを用いた通信環境計測ツールやファイル転送ツールなども実装および販売を行っている。本実験では HpFP を多重化して通信速度を計測するツール hperf を用いた。これにより FX700 の 2 ノードを直結して実験を行ったところ、30Gbps 程度を達成した。また、仮想的に遅延およびパケットロスを与えて計測を行ったところ、24 並列で 22Gbps となった。これらのボトルネックは、もともとデータ通信高速化での利用を想定していないスーパーコンピュータの NIC（ネットワークカード）やメモリサイズ（バッファサイズ）にあると予想している。

1 はじめに

近年のネットワーク広帯域化に伴って、10Gbps さらには 100Gpbs を超えるような高速データ通信アプリケーションの必要性が高まっている。これらを背景に、SUPERCOMPUTING ASIA[1]では毎年 Data Mover Challenge を開催し、各国の参加者が複数国を接続する国際バックボーンネットワーク回線上での超高速データ伝送技術を競っている。

データ通信高速化の手法の一つとして、トランスポート層のデータ通信プロトコルの技術開発が有効である。これらの手法は主として TCP 改良型と独自プロトコル開発に分けられる。独自プロトコルは UDP をベースとして独自アルゴリズムの通信を実現する方法であり、信頼性を担保する TCP 互換型と信頼性の一部を犠牲にして高速化を行う特殊型に分かれる。インターネット上でのデータ伝送は前者となるが、独自ネットワーク上での映像伝送などは後者で行われることもある。

これまでにも様々な TCP または UDP をベースとしたデータ通信プロトコルが開発されているが、ネットワークにおいて一定値以上のパケット損失や遅延が生じる場合に伝送効率が大きく劣化するため、期待するような高速データ伝送が提供できないという課題があった。筆者らはパケット損失や遅延が生じる長距離広帯域伝送ネットワーク環境においても高い伝送効率を実現するトラスポート層プロトコル HpFP（High-performance and Flexible Protocol）を開発した[2]。

詳細は別論文に譲るが、HpFP をベースとして開発したデータファイル転送アプリケーション HCP tools により DMC2020 および DMC2021 において高速データ通信実験を実施した。その結果 30Gbps から 50Gbps を超えるファイル転送を実現したが、バンド幅いっぱいの 100Gbps または 200Gbps でのファイル転送には至らなかった。室内実験環境において汎用 Linux サーバを用いて類

似のネットワークを仮想的に構築し検証を行ったところ、ボトルネックはメモリアクセス速度にあることが分かった。

そこで本稿では、メモリバンド幅が一般的な Linux サーバと比較して高速であるスーパーコンピュータによる HpFP の高速化を試みる。具体的には京都大学のスパコンである FX700（富士通社）上で HpFP 通信のマルチスレッド通信を行い、IPoverIB 環境上で 100Gbps クラスの HCP によるノード間通信実験を実施した結果を報告する。第 2 節では HpFP について紹介し、第 3 節では利用するスーパーコンピュータの紹介、第 4 節で実験結果について報告する。第 5 節で本稿をまとめる。

2 HpFP プロトコル

2.1 HpFP プロトコルとアプリケーション

HpFP は、経路上でパケットロスが発生する LFN（Long-fat network）においても高信頼性通信を実現するために、TCP ではなく UDP をベースに設計したトランスポート層の独自プロトコルである。衛星通信環境[3]、Web アプリケーション[4]、国際回線でのデータ共有[5]などでの高速通信実績がある。

筆者らのグループでは HpFP をベースとしたネットワーク環境計測ツール hperf[6]をフリーソフトウェアとして公開している[7]。クレアリンク社では HpFP をベースとしたデータファイル伝送・同期ツールである HCP tools を開発、公開（販売）している。HpFP のバージョンは現在 ver.2 であり、たとえば Super Aggressive Mode として、回線帯域が予めわかっており該当回線を占有できる場合にターゲットスループットを設定し、設定した帯域の限界まで活用することが可能なモードなどがある。

次節の DMC 以外にも、JHPCN（学際大規模情報基盤共同利用・共同研究拠点）プロジェクトにおいて、HCP のファイル同期機能により、拠点大学間（たとえば NICT、京都大学、九州大学など）の 1PB 級のデータファイル伝送実験などの実績がある[8] [9]。

2.2 Data Mover Challenge

情報通信研究機構、京都大学およびクレアリンクテクノロジー社は、Team Musashino として、2020 年、2021 年に開催された Data Mover Challenge に 2 年連続で参加した。2021 年は参加 7 チーム中 5 チームが入賞し、同チームは”Most Innovative and Best IPv6 Performance”賞を受賞、2022 年 3 月 2 日にシンガポールで開催された SUPERCOMPUTINGASIA 2022 において表彰を受けた。NICT とクレアリンクテクノロジー社が開発した HpFP プロトコルおよびそれによるファイル転送ツール（HCP）の先進性（他のプロトコルとの親和性の高い独自開発の高信頼プロトコルと IPv6 環境での高スループット）が評価されたものである。同チームは 2020 年度に開催された DMC20 においても”Experimental Excellence Award”（HpFP を用いたデータ伝送ツールの多機能性が評価された）を受賞しており[10]、2 年連続の受賞となった。

3 FX700 による高速化性能検証

3.1 目的

前述の通り、これまでに 100Gbps の NIC を搭載した 2 台のサーバ間での室内通信実験環境において、並列化した HpFP プロトコルを用いたデータ通信実験を実施したところ、通信ボトルネックはメモリアクセス速度にあることが分かった。これは、HpFP が UDP をベースに実装されており、TCP のようなカーネルサポートや NIC による高速化が困難であるためである。特に、パケットロス環境においては再送制御のためのデータ処理量が膨大となり、これがメモリアクセスを圧迫している。

本実験で用いる FX700 の 1 ノードあたりのメモリバンド幅は 1024GB/sec で、上記室内実験で利用している一般的なサーバの 10 倍以上である。したがって、100Gbps までの通信であれば、メモリアクセス性能はボトルネックにならないと予想した。

2.2 FX700 システム構成

本実験では、京都大学学術情報メディアセンターの Fujitsu FX700 を使用した。システム構成を図 1 に示す。

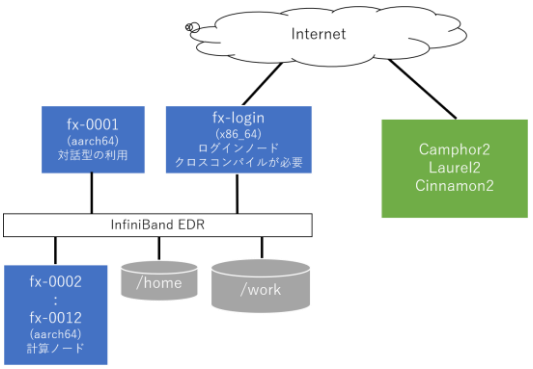

図 1: 京都大学 FX700 のシステム構成図[11]

実験では、fx-0009 と fx-0010 の計算ノードを占有し、2 つのノード間で通信を行った。Fujitsu FX700 の計算ノードのスペックは以下の通り。

- CPU: A64FX 1.8GHz 48 コア (Armv8.2-A SVE)
- アーキテクチャー: 1 CPU/ノード
- メモリ容量: 32 GiB（HBM2, 4 スタック）
- メモリバンド幅: 1,024 GB/s
- インターコネクト: InfiniBand EDR 100Gbps
- 内蔵ストレージ: M.2 SSD Type 2280 スロット（NVMe）
- OS: Red Hat Enterprise Linux 8

3.3 環境設定

OS のネットワークパラメータとして、UDP の性能向上のため、以下のバッファサイズの設定を各計算ノードに行った。

net.core.wmem_max=104857600
net.core.rmem_max=104857600

3.4 メモリアクセス性能計測

本研究ではまず、C 標準関数 memcpy を用いたベンチマークプログラムを作成し、並列数とブロックサイズを変更して、read/write の速度を計測した。上記の通り、FX700 ではメモリバンド幅が 1024GB/sec である。memcpy では read と write が 1 回ずつ発生するので、FX700 での最大理論値は 512GB/sec（4096Gbps）となる。詳細は別論文で述べるが、最大性能は理論性能の約 57%となった。また、並列数を変更して計測したところ通信性能が並列数にほぼ比例したため、メモリコピー処理の CPU 負荷がボトルネックになっていると思われる。

さらにメモリアクセス速度向上のために、FX700 用に開発された専用メモリ性能計測ツールを用いた計測を行った。同ツールでは libnuma により CPU コアとメモリコントローラを制御することでメモリアクセス速度向上を実現している。本ツールによると、1 並列当たりの最大性能が約 2.5 倍、トータルメモリアクセス性能も約 1.2 倍向上した。

3.5 通信性能計測

本節では、TCP および HpFP による通信性能検証を行った。TCP は前節の libnuma ありの場合となしの場合で比較した。HpFP については libnuma 対応のためのプログラム書き換えが膨大となるため、本研究では libnuma なしの場合のみ計測した。計測には iperf や hperf[6]を用いた。計測結果の詳細は別論文で述べるが、ここでは結果の概要を報告する。

まず、NIC を標準状態の Infiniband モードに設定した。MTU をデフォルトの 2044 バイトとした場合、並列度により約 5〜10 倍程度、TCP が HpFP よりも高速であった。また MTU を 65520 バイトに拡張しても TCP および HpFP は高速化しなかった（低速化した）。特に TCP でスピードダウンが顕著であった。これは Infiniband の MTU はこの環境では 2044 が上限で、それ以上の設定は仮想的にパケットの分割と結合をしているためだと考えられる。

次に NIC を Ethernet モードに設定し、MTU を 9000 バイトとした。コンパイラは gcc を用いた。その結果、Infiniband よりも性能は向上したが、TCP は約 70Gbps、HpFP は約 30Gbps となり目標である 100Gbps には届かなかった。なお、fcc でコンパイルしても性能に大きな変化は見られなかった。高速化が達成できなかった要因としては、NIC 性能限界であると予想している。

4 LFN モデルでのパケットロス耐性検証

前節において、HpFP（hperf）は FX700 の 2 ノードを直結した場合に、期待する 100Gbps のスループットを達成できなかった。本節では、同環境において仮想的に遅延とパケットロスを与えることで、HpFP の LFN 上でのパケットロス耐性が汎用 Linux サーバと比較して向上するかについて調査した。

調査結果の詳細は別論文で議論するが、たとえば 24 並列（バッファサイズを 512MB に設定）では 21.1Gbps という結果となり、必ずしも期待するような高速データ通信は実現しなかった。CPU 使用率は送受信側共に合計 500%程度で全体の 1/6 程度であり、負荷が 100%になっているコアもなかった。したがって、高速化を阻害する要因はバッファサイズの不足ではないかと考えられる。

5 おわりに

本研究では、京都大学学術情報メディアセンターの富士通スパコン FX700 を用いたデータ通

信プロトコル HpFP（High-performance and Flexible Protocol）の高速化を試みた。HpFP は筆者らのグループが開発した TCP 互換プロトコルであり、これを用いた通信環境計測ツールやファイル転送ツールなども実装および販売を行っている。本実験では HpFP を多重化して通信速度を計測するツール hperf を用いた。これにより FX700 の 2 ノードを直結して実験を行ったところ、30Gbps 程度を達成した。また、仮想的に遅延およびパケットロスを与えて計測を行ったところ、24 並列で 22Gbps となった。これらのボトルネックは、もともとデータ通信高速化での利用を想定していないスーパーコンピュータの NIC（ネットワークカード）やメモリサイズ（バッファサイズ）にあると予想している。

謝辞

本実験は、京都大学学術情報メディアセンターのスーパーコンピュータ共同研究制度 FX700 小ノード実行枠を利用して行いました。また、本実験を行うにあたり、富士通株式会社・インフラストラクチャシステム事業本部・基盤ソフトウェア事業部・甲斐俊彦様のご協力をいただきました。また、実験のご支援をいただいた京都大学情報部情報基盤課・當山達也様に感謝します。

参考文献

[1] SUPERCOMPUTING ASIA：https://www.sc-asia.org/about/

[2] 村田健史、水原隆道、長屋嘉明、村田健史，長屋嘉明，水原隆道，IoT/M2M 時代に向けた高性能遠隔制御のための通信プロトコル -新しい社会システムデザインに向けた基盤通信技術の創出 -，電波技術協会報 FORN，2016 年 9 月号，No.312，pp.6-9，2016.

[3] K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, N. Katayama, K. Muranaga, T. Mizuhara, A. Takaki and E. Kimura, "An Application of Novel Communications Protocol to High Throughput Satellites," in The 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON2016), Vancouver, Canada, Oct. 13-15, pp. 1-7, 2016, doi: 10.1109/IEMCON.2016.7746274.

[4] K. T. Murata, P. Pavarangkoon, K. Inoue, T. Mizuhara, Y. Kagebayashi, K. Yamamoto, K. Muranaga, E. Kimura and Y. Nagaya, "Development of High-performance and Flexible Protocol Handler for International Web Accesses," in The 21st IEEE International Conferences on High Performance Computing and Communications (HPCC-2019), Zhangjiajie, China, Aug. 10-12, pp. 1958-1963, 2019, doi: 10.1109/HPCC/SmartCity/DSS.2019.00270.

[5] P. Pavarangkoon, K. T. Murata, K. Yamamoto, K. Muranaga, A. Higuchi, T. Mizuhara, Y. Kagebayashi, C. Charnsripinyo, N. Nupairoj, T. Ikeda, J. Tanaka and K. Fukazawa, "Development of international mirroring system for real-time web of meteorological satellite data," Earth Science Informatics, vol. 13, no. 4, pp. 1461-1476, 2020, doi: 10.1007/s12145-020-00488-z.

[6] K. T. Murata, P. Pavarangkoon, K. Yamamoto, Y. Nagaya, T. Mizuhara, A. Takaki, K. Muranaga, E. Kimura, T. Ikeda, K. Ikeda and J. Tanaka, "A quality measurement tool for high-speed data transfer in long fat networks," in 2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), Split, Croatia, Sep. 22-24, pp. 1-5, 2016, doi: 10.1109/SOFTCOM.2016.7772111.

[7] https://hpfp.nict.go.jp

[8] P. Pavarangkoon, K. T. Murata, K. Yamamoto, K. Muranaga, T. Mizuhara, K. Fukazawa, R. Egawa, T. Katagiri, M. Ogino and T. Nanri, "Performance Improvement of High-Speed File Transfer over JHPCN," in The 5th IEEE International Conference on Cloud and Big Data Computing (CBDCom 2019), Fukuoka, Japan, Aug. 5-8, pp. 1086-1089, 2019, doi: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00195.

[9] K. T. Murata, P. Pavarangkoon, K. Yamamoto, K. Muranaga, T. Mizuhara, K. Fukazawa, R. Egawa, T. Katagiri, M. Ogino and T. Nanri, "High-Speed File Transfer of Real Datasets over JHPCN," IEICE Technical Report, vol. 118, no. 466, Okinawa, Japan, Mar. 4-5, pp. 175-179, 2019.

[10] P. Pavarangkoon, K. T. Murata, K. Yamamoto, N. Fujita, H. Ohkawa, H. Mikai, Y. Ikehata, K. Muranaga, T. Mizuhara, A. Takaki and Y. Kakizawa, "Performance Evaluation of High-Performance and Flexible Protocol on Data Mover Challenge," in 2020-5th International Conference on Information Technology (InCIT), ChonBuri, Thailand, Oct. 21-22, pp. 265-269, 2020, doi: 10.1109/InCIT50588.2020.9310956.

[11] 京都大学学術情報メディアセンターFX700 システム図（https://web.kudpc.kyoto-u.ac.jp/manual/ja/fx700 より転載）