ATMワークステーションクラスタにおけるグループ通信MPIライブラリの実現と評価

4G－7

王 欣丹 大澤 範高 弓場 敏嗣*
電気通信大学大学院情報システム学研究科

1 はじめに

EthernetやTCP/IPに代表される従来型のネットワーク技術を用いたクラスタシステムは、狭帯域、転送遅延、プロトコルのオーバヘッドなどの問題がある。一方、ATM[3]は通信帯域が155Mbps～622Mbpsと大きく、ネットワークの転送遅延やエラー率も非常に小さい。さらに、ポイントツーマルチポイントコネクション[1]を提供し、上位のグループ通信の機能に対応することができる。

また、MPI[2]は並列プログラムのためのメッセージ通信用ライブラリインタフェースの標準仕様である。MPIのグループ通信には、一つのプロセスからグループ内の全てのプロセスへ、同一データを伝送するというbroadcast機能と、一つのプロセスからグループ内の各プロセスへ、異なるデータを伝送するというscatter機能がある。

本稿では、ATMネットワークで接続したワークステーションクラスタ上でのMPIのグループ通信ライブラリの実装及び通信性能に関する評価実験について述べる。

2 ATM通信方式モデル

コネクションの確立、解放に要する時間を含む通信時間を定量的に与えるATM通信方式モデルを提案した。これにより、コネクションオーバヘッドを含む通信時間を得ることができ、適切なコネクション管理と通信方法を明らかにできる。

実行手順はコネクション要求の送信、結果の待ち受け及び結果の処理という三つ部分から構成される。モデルと実測に基づくパラメータ値に基づいて、コネクション数とデータサイズによる各通信方式を用いた場合の通信時間を計算する。その結果に基づいて適切な伝送方式を選択する。

3 MPIライブラリの実装

ATMのAAL層上にグループ通信プロトコルを設計して、MPIライブラリのプロトタイプ(MPI over ATM)を実装した。設計したbroadcast機能とscatter機能に対応する1対多通信プロトコルの概要を以下に示す。

1. 受信バッファのオーバフローを回避するために、ノード受信速度、通信バッファサイズ及び伝送データサイズによる最適な送信レートを求める[4]。broadcast通信プロトコルは、グループ中の一番低い受信速度のノードに対応する送信レートを選ぶ。

2. 通信プロトコルは、ATM通信方式モデルを用いた解析に基づいて、より通信時間が短くなるコネクションを利用する方針とする。それにより、broadcast通信プロトコルはポイントツーマルチポイントコネクションでデータを伝送し、双方向通信可能なポイントツーポイントコネクションで制御メッセージのやり取りを行う。また、ポイントツーマルチポイントコネクション確立の時間のために通信時間が長くなる場合には、ポイントツーポイントコネクションでデータ伝送及び制御メッセージのやり取りを行う。scatter通信プロトコルは双方向通信可能のポイントツーポイントコネクションでデータ伝送及び制御メッセージのやり取りを行う。

3. broadcast通信プロトコルは、再送信が必要な場合に、再送信すべきデータに基づいて、再送信時間が短くなるようにポイントツーポイントコネクションもしくはポイントツーマルチポイントコネクションを選択する。

4 実装環境

1. ノード
   SUN-SS5 × 4台
   MicroSparcII 70MHz, 32MB, Solaris2.5
   SUN-SS20 × 2台
   SuperSparc 75MHz, 48MB, Solaris2.5

2. ATMスイッチ
   住友電気工業(株)SUMINET-3700SH
   SONET/SDH, 155.52Mbps, UTP-5, 8ポート

3. ATMネットワークインタフェースカード
   Efficient Networks, Inc. ENI-155s-U5-c
   155.52Mbps,Sbus adapter, 512KB Memory
   Efficient Software Development Kit(SDK)3.3.0

---

* Implementation and Evaluation of a Collective Communication MPI Library on a ATM Workstation Cluster. Xindan WANG, Noritaka OSAWA and Toshitsugu YUBA, Graduate School of Information Systems, the University of Electro-Communications

5 MPIライブラリの評価

IP over ATM、IP over Ethernet(10Mbps)及びMPI over ATM上で、MPIプログラムを実行し、データ伝送の評価実験を行なった。その結果について考察する。

1. ノード数を6に固定した場合のMPI_Bcast関数とMPI_Scatter関数のデータサイズに対する実行時間を図1に示す。

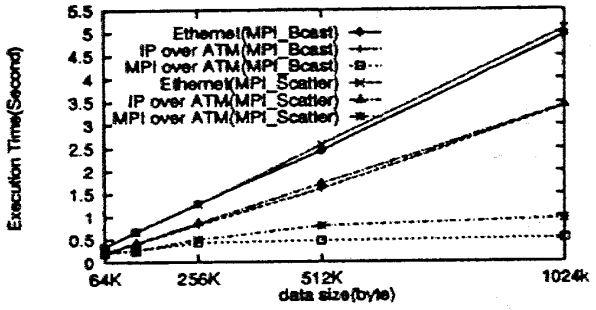

図 1. MPI関数のデータサイズに対する実行時間

IP over ATM と Ethernet 上の MPI_Bcast 関数は、ポイントツーポイント通信で実現するので、MPI_Scatterと同じデータ伝送方法である。その結果、MPI_Bcast関数の実行時間は、MPI_Scatterとほぼ等しくなる。MPI over ATM上のMPI_Bcast関数は、ポイントツーマルチポイントコネクションを利用しているので、MPI_Scatterより実行時間が短くなることがわかる。

2. 伝送データサイズが小さい時(64KB)のMPI_Bcast関数のノード数に対する実行時間を図2に示す。IP over ATMとEthernetよりも、MPI over ATM上での実行時間が長いことがわかる。

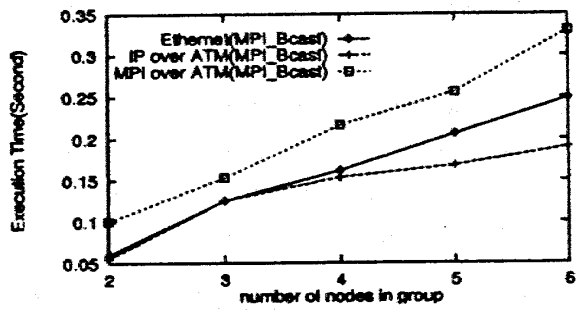

図 2. MPI関数のノード数に対する実行時間(64KB)

MPI over ATM上のMPI関数の実行時間にはコネクションの確立、解放時間を含む。一方、IP over ATMは、各ノードでのIP over ATM環境を初期化する際に、グループ中の他のノードにコネクションを確立するので、MPI関数の実行時間がデータ伝送時間だけである。また、Ethernetでデータを伝送する際にも、他のノードに対するコネクションが必要ないので、MPI関数の実行時間は、データ伝送時間だけである。従って、伝送データサイズが小さい場合に、MPI over ATM上のMPI関数はEthernet及びIP over ATMよりも、実行時間が長くなる。

3. 伝送データサイズが大きい時 (1024KB) のMPI_Bcast関数のノードに対する実行時間を図3に示す。IP over ATMとEthernetより、MPI over ATM上での実行時間が短いことがわかる。

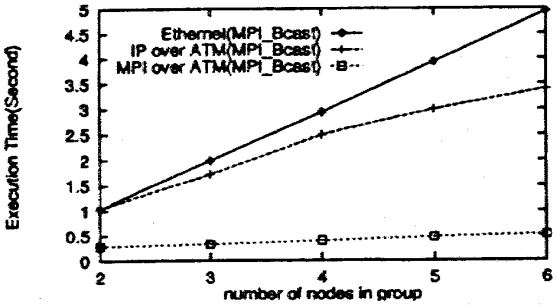

図 3 MPI関数のノード数に対する実行時間(1024KB)

設計した通信プロトコルは、ATMドライバの通信機能を直接に利用するので、IPパケットヘッダやTCPセグメントヘッダを持たない。さらに、スループットを向上するため、データをユーザバッファに直接に送受信するので、システムバッファ間のメモリコピーがない。その結果、伝送データサイズが大きい場合に、MPI over ATM上のMPI関数はEthernet及びIP over ATMよりも、実行時間が短くなる。

6 おわりに

本稿では、ATMワークステーションクラスタにおけるグループ通信用のプロトコルを提案して、MPIライブラリとして実装し、IP over ATM上とEthernet上で動作するMPIライブラリとの比較を行った。データサイズ、ノード数とMPI関数の実行時間の関係を調べ、実装したMPIライブラリの利点、欠点を考察した。今後は、コネクション管理方法の改善及びスキャン、リダクション、バリア同期などのMPI機能の追加を行う予定である。

参考文献

[1] The ATM Forum. *ATM User-Network Interface (UNI) Specification Version 3.1*. Prentice Hall PTR, 1995.

[2] M.Snir, S.W.Otto, S.H.Lederman, D.W.Walker, and J.Dongarra. *MPI The Complete Reference*. The MIT Press, 1996.

[3] マーチン・ドゥプライカー. ATM詳解、新世代通信網構築技術. 株式会社プレンティスホール, Sep. 1996.

[4] 大井拓哉, 大澤範高, 弓場敏嗣. ATMワークステーションクラスタにおける通信スループットの改善. 分散システム運用技術シンポジウム論文集, pp. 13-18. 情報処理学会, Feb. 1997.