メッセージ衝突を防止した適応的な集合通信

吉富　翔太†　　田浦　健次朗†

†東京大学大学院 情報理工学系研究科

1　はじめに

集合通信は並列分散処理に欠かせない操作であり，従来より数多くの最適化に関する研究がなされてきた．その多くは計算機間の遅延やバンド幅，通信データ長の値を用いて集合通信をモデル化し，最適な通信経路を決定するものである [1, 2, 5]．ところがこれらの手法の多くは共起する通信同士の影響を厳密に考慮していないので，各ノードが協調せずに通信してしまい通信性能が非常に悪化する．また上記を考慮している手法でも，ノード間でパケットを送受信することだけで通信の協調を図るものなので例えば高遅延な環境では性能が悪化してしまうなど，どのような環境でも最適な手法となるわけではない．

一方我々は過去にネットワークトポロジーの情報を利用して，複数ノード間で同期を取りながら逐次に通信を行うことに加えパイプライン状にノード間でデータを転送させる手法を組み込むことで，コンテンションを抑制し，なおかつ同期に要する総コストも削減させた性能の良い gather のアルゴリズムを提案した [3]．さらに他に比べて広帯域なリンクにおいては一斉に複数ノードが通信を行うようにすることで，特に複数クラスタを利用した環境でスループットの向上を実現させている．

そこで本稿では上記の手法を踏まえ，コンテンションを効果的に抑制することで一般の並列分散環境において柔軟に適応可能な多対多型の集合通信のスケジューリング手法を提案する．

2　関連研究

既存の MPI ライブラリには標準で集合通信の API が用意されている．一例として OpenMPI や MPICH では，集合通信の API 呼び出し時に様々なアルゴリズムに基づき通信ツリーを構築して通信を行う．ところがこれらは各ノードが協調せず一斉に通信するという実装が行われているため，コンテンションにより通信性能が悪化する可能性がある．また，均質なネットワークを対象としたものであり，不均質な環境では通信性能が極端に悪化してしまうという問題もある．一方でグリッド環境に適応した MPI ライブラリの研究もまた数多く行われており，Kielmann らによる MagPIe[1] や松田らによる GridMPI[4] は，LAN のと WAN により別々のアルゴリズムを適用することでクラスタ間の通信について高い性能を得ているものの，コンテンションを抑制することまでは厳密に考慮しておらず，環境によっては通信性能が悪化する．

また個々の通信を一つのジョブと見なしたスケジューリング問題として捉え，集合通信の最適化を図る研究も存在する [2, 5]．遅延・バンド幅・通信データ長からノード間の通信コストを算出し，全体の総通信時間が最小となるように各ノード間の通信をスケジュールする．ところがこれらの手法は同時に同一のリンクを多数のノードが利用するような通信の割当がなされてしまったり，スケジュールした通りに実際に通信するにあたり，ノード間で同期を行う際に要する同期のコストが大きくなってしまうなど，アルゴリズムを実際の環境で実行させたときに期待する性能が得られない可能性がある．

3　提案手法 – 全対全通信のスケジューリング –

3.1　コンテンションの影響とその抑制方法

スイッチに接続された任意の一つのリンクに対し，他の複数のリンクから同時に大量のデータが流入しようとすると，リンクの容量やスイッチの性能によってはパケットロスが生じる．パケットロス発生時に TCP では特定の条件下で一定時間 (RTO) だけ通信を停止させる性質が存在するため，パケットロスが多発すると結果的に集合通信の性能が非常に低下することがある．これが本稿で問題としている通信のコンテンションであり，これを抑制することが多対一・多対多型の通信で高い性能を得るためには必要不可欠である．

本研究では，(1) メッセージを別のノードを介して転送させるかもしくは (2) ノード間で同期取りつつ逐次に通信させるかの 2 通りの手法により，コンテンションを抑制した通信を実現する．これらの 2 手法の直接の目的はネットワーク上のどのリンクについてもその容量を超えるようなデータが一度に流れ込まないようにすることである．一方で両手法ともいくばくかの余分なコストが発生する．メッセージの転送により発生する余分な通信のコストやノード間の同期コストがそれに相当するのであるが，それらはデータ長や遅延・バンド幅などのネットワークのパラメータの関数となるためどちらのコストが大きくなるのかは自明ではなく，どちらかの手法のみを用いるだけでは効率が悪くなることがある．故にネットワークの構造に応じて両者のうちコストが小さくなる方法を随時選択することが高性能な通信を行う上で必要不可欠となる．

3.2　スケジューリングにおける制約条件

全対全通信を例に提案手法のスケジューリング手順の概略を述べる．個々のノード間通信を ノード間遅延 + データ長 ÷ ノード間バンド幅 なるコストを持つジョブと見なすと，全対全通信の性質より全ノード数の 2 乗個のジョブが存在することになる．各ノードが送信と受信を同時に独立にできるとすれば，全対全通信は通信をジョブに見立てたスケジューリング最適化問題と見なせる．以降では，一組のノード間通信をジョブと呼び，ジョブをスケジュールするとは当該ジョブの開始と終了の時刻をそれぞれ遅延やバンド幅をもとに決定することと定義する．さらに主目的であるコンテンションの抑制のため，スケジューリングを行うにあたりいくつかの制約を導入する．

An Adaptive Collective Communication Suppressing Contention
Shota YOSHITOMI†　Kenjiro TAURA†
The University of Tokyo†

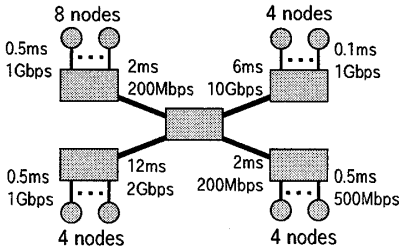

図 1　実験環境

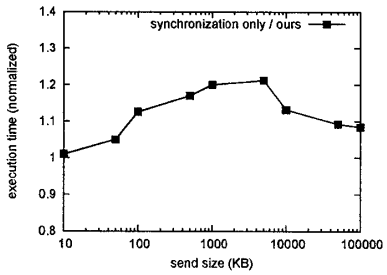

図 2　シミュレーション結果

3.2.1　コンテンションに関する仮定

前述した通り，通信性能を悪化させないためには，リンクの容量を大きく超えるデータを一度に流入させないことが第一である．そこで，任意の一本のリンクにおいて同時刻に同時にデータを流せるのを 1 ノードのみに制限する．ただし (1) データが流入する側のリンクのバンド幅の合計が流出側のリンクのバンド幅よりも小さいか (2) 流入するデータ量の和が流出側のリンクの帯域遅延積以下である場合はリンクの容量に余裕があるため，例外的に複数ノードが同時に同一リンクを使用するように通信を割り当てても良いものとする．

3.2.2　通信時間の下界とスケジューリングの優先度

さて，各リンク上をデータが流れている時間 T_c は通過データ量の総和 ÷ リンクのバンド幅で表せる．この値は言わばリンクの通信コストの下界と呼べるものであり，もしデータがリンクを間断なく通過した場合の各々のリンクの使用時間は T_c となる．系全体のトータルの通信時間を少なく抑えるためには，T_c の値が大きいリンクには常にデータが流れている必要がある．故に T_c の値を各リンクの優先度と定義し，それぞれのジョブにおける通信に使用する経路上のリンクの優先度の高い順に優先的にジョブをスケジューリングしていく．

3.3　スケジューリング手順

基本的に 3.2.2 の優先度の順にジョブをスケジュールする．ただし，ジョブの通信が行われようとする時間中に通信経路上の一箇所以上のリンクで 3.2.1 の制約が満たされていない場合は，そのジョブは制約が満たされるまで割り当て不能となる．その場合の処理方法が 3.1 で挙げた 2 手法となる．

まず簡単な方法がノード間で同期を取ることである．これは，そのジョブにおける通信が使用する経路における制約が満たされるまでジョブの開始時刻をずらすことに相当する．各ノードはデータを受信し終わると同時にデータを全て受信したことを示すパケットを他ノードに broadcast する．制約を満たすために必要十分なノード全てからそのパケットを当該ジョブの通信を行う送信者側のノード受け取った時刻をもってそのジョブの開始時刻として，ジョブをスケジュールする．

もう一つの方法がデータを転送してしまうことである．これは $A \to C$ なるジョブを $A \to B, B \to C$ なる 2 つのジョブに分割することを意味する．もし $A \to B, B \to C$ のどちらも 3.2.1 の制約が満たされているのであれば，データは即座に $A \to B \to C$ と送信できる．制約の満たされていないジョブを制約が満たされている複数のジョブに分割することで，複数ノード間で同期を取る回数を減らすことができる．

どちらの手法が適しているかは 3.1 で述べた通りネットワークの構造に依存する．証明は省略するが，仮にジョブのスケジュールを待機させることで増加する時間 (これを同期コストとする) と，仮に転送を行った場合に各ノードの通信時間の増分および各リンクをデータが通過する時間の増分 (これを転送コストとする) とを計算・比較し，コストが小さくなる手法を選択してそのジョブをスケジュールする．

以上を繰り返し全てのジョブを割り当て終わったら全対全通信のスケジューリングが完了したものとする．

4　実験と評価

コンテンションを抑制させる通信手法という点で，ノード間で同期を行うのみの既存手法に対して，加えてデータ転送も考慮する提案手法がどれほど有効であるかについて検証を行った．具体的には，図 1 のような構造を持つネットワークに対して，提案手法 (同期及び転送を行う) と比較手法 (同期のみを行う) について，通信データ長をそれぞれ変化させた場合の各々の総通信時間についてシミュレーションを行った．その結果を図 2 に示す．この結果より，全対全通信において単に同期を取るだけではなく，データの転送を行うことがある程度有効であることがわかる．

5　おわりに

本論文ではノード間で同期を取りつつ，加えてデータを別ノード間で転送させる手法を組み合わせることによりコンテンションを効果的に抑制できる多対多型集合通信の手法を提案した．さらにシミュレーションを行い，コンテンションを抑えた通信を行うという制約を課した場合でデータの転送を行うことの効果を確認した．今後は提案手法を適用した集合通信を実装し，実際のヘテロな並列分散環境で既存手法との比較を行い，性能を検証していくことを目指す．

謝辞　本研究の一部は文部科学省科学研究費補助金特定領域研究「情報爆発に対応する新 IT 基盤研究プラットフォームの構築」の助成を得て行われた．

参考文献

[1] Thilo Kielmann, Rutger F.H.Hofman, Henri E.Bal, Aske Plaat, and Raoul A.F.Bhoedjang. Magpie: Mpi's collective communication operations for clusterd wide area systems. *PPoPP'99*, 1999.

[2] Yuya Jinno, Masashi Ito, and Akihiro Fujiwara. Efficient scheduling algorithms for total exchange on grid environment. *PDPTA'03, Vol. 1*, pp. 81–87, 2003.

[3] Shota Yoshitomi, Ken Hironaka, and Kenjiro Taura. An adaptive collective communication suppressing contention. *SACSIS 2009*, pp. 71–78, 2009.

[4] Motohiko Matsuda, Yutaka Ishikawa, Tomohiro Kudoh, Yuetsu Kodama, and Ryousei Takano. Efficient collective algorithms for grid environment. *IPSJ SIG Notes Vol.2006*, pp. 257–262, 2006.

[5] Ahmad Faraj and Xin Yuan. Message scheduling for alltoall personalized communication on ethernet switched clusters. *IEEE IPDPS*, 2005.