情報爆発時代のグリッド環境に対応した MPI集団通信アルゴリズムの最適化

千葉 立寛* 遠藤 敏夫† 松岡 聡‡

東京工業大学 *, †, ‡ 国立情報学研究所 ‡

1 はじめに

近年，大規模な科学技術計算の実行環境としてグリッド環境の利用が高まってきた．これらのアプリケーションの多くが並列計算ライブラリである MPI を用いて記述されており，これまでにも様々なグリッド向け MPI 実装 [1, 2] が提案されている．MPI アプリケーションの実行性能を大きく左右する要因の一つとして MPI 集団通信が挙げられる．

LAN に比べて高遅延・低バンド幅な WAN で構成されるグリッド環境において，その遅延やサイト間を結ぶバンド幅の影響を十分に考慮した通信・トポロジ最適化を行わない場合，MPI アプリケーションの性能が著しく低下する．

本稿では集団通信の中で Scatter/Gather 通信に対して，グリッドを構成する WAN のバンド幅の十分な利用を可能とするマルチレーン集団通信アルゴリズムを提案する．

2 関連研究と問題点

これまでの集団通信アルゴリズムの最適化に関する研究の多くでは，WAN の通信遅延やバンド幅の影響を最小にするツリーが提案されてきた．MPICH-G2[1] では，サイト間を代表ノード同士がまとめて通信しあうことで，ボトルネックリンクに対する性能低下を極力小さくしている．しかしながら，近年，高バンド幅を実現可能な WAN も登場しており，これらのアルゴリズムの前提としていた，"LAN の性能よりも WAN の性能が低い"，という仮定が成立しなくなってきている．

GridMPI[2] では，高バンド幅な WAN を備えるグリッド環境に対して，複数のリンクで WAN 間通信を行う Bcast 通信と AlltoAll 通信アルゴリズムを提案している．また論文 [3] では，パイプライン通信によって高速な転送が可能になる Bcast 通信用のマルチレーンツリーアルゴリズムを提案している．

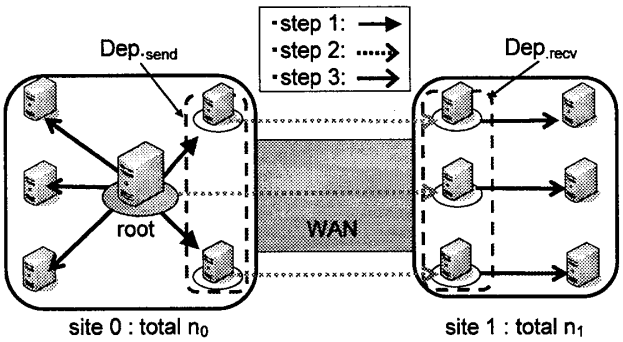

図 1: 2 サイト環境での提案アルゴリズムによる scatter アルゴリズム

今後ますます WAN のバンド幅性能が向上し，扱うメッセージサイズも増大することが予測されるため，WAN のバンド幅を十分に利用可能な通信アルゴリズムを考えることが重要である．

3 提案手法

3.1 集団通信アルゴリズム

step.1(転送準備フェーズ)，step.2(サイト間転送)，step.3(サイト内転送) の 3 ステップで scatter 通信を実現する提案アルゴリズムの説明を行う．なお，gather 通信に関しては通信の向きが逆方向であるだけなので，ここでは scatter 通信のみを扱う．記号の意味と通信の流れ，トポロジ構造を図 1 に示す．

step.1 *root* は dep_{send} に対してそれらが転送を担当するメッセージを *LAN* 内で転送する．

step.2 dep_{send} は対応する dep_{recv} に対してサイト間転送を実行する．

step.3 dep_{recv} は担当する子ノードに対して転送を行う．また *root* も dep_{send} 以外の子ノードへの転送を行う．

"Optimization for MPI Collective Operations on Grid Utilizing Multilane Transfer"

*Tatsuhiro Chiba, Tokyo Institute of Technology

†Toshio Endo, Tokyo Institute of Technology

‡Satoshi Matsuoka, Tokyo Institute of Technology/NII

3.2 通信コストモデル

提案アルゴリズムによるscatter通信を実現するため，WAN間転送を行う dep_{send} と dep_{recv} の組を決定する必要がある．WAN間転送を行う組数 P の取りうる範囲は，$1 \leq P \leq max(n_0, n_1)$ である．下記のようなコストモデルによって P を変化させたときのアルゴリズムの転送コストを見積もり，コストが最小となる dep_{send} と dep_{recv} の組を決定する．サイト間遅延を L ms，サイト間を P 組で同時にWAN間通信するときの1リンクあたりのバンド幅を $b_L(P)$，LAN内のバンド幅を B，メッセージサイズを M とすると，通信コスト $T_L(P)$ は，

$$T_L(P) = L + \frac{x_L(P) \cdot M}{b_L(P)} + \frac{y_L(P) \cdot M}{B} + \alpha$$
$$x_L(P) = P_{max}/P + P_{max}\%P$$
$$x_L(P) + y_L(P) = n_0 + n_1 - 1$$
$$P_{max} = max(n_0, n_1)$$

として表すことができる．$x_L(P)$ と $y_L(P)$ は転送するメッセージの個数，α は通信以外のオーバーヘッドを表す．また，一般的なアルゴリズムの代表ノードが1組ずつでWAN間通信を行う場合のコストは $T_L(1)$ と同値とする．

3.3 トポロジ構築

アルゴリズムとコストモデルから図1のようなマルチレーンツリートポロジを以下の手順で構築する．(1). 遅延 L ms に対するリンク1本辺りのバンド幅 $b_L(P)$ をあらかじめ計測する．(2).$T_L(P)$ を計算し最小となる P_{opt} を求める．(3).P_{opt} 台の代表ノード群をそれぞれのサイトに割り当て，各ノードの転送順をstep.1 2 3となるようスケジューリングする．

4 評価

合計32ノード，WANの遅延8msの2サイトグリッド環境をGNET-1を用いてエミュレートし，各アルゴリズムでの集団通信の性能を測定し比較した(図2)．250KB付近で段階的に性能が低下しているが，概ね線形に性能が推移し，常にマルチレーンのほうが1.5倍程度高速に通信が行えていることが確認できた．

また，図3からシングルレーンだと5Mbps程度の性能しか利用できていないが、マルチレーン化することで，最大2倍程度WANのバンド幅を利用できていることが確認できた．しかしながら，マルチレーンに対する予測実行時間はほぼ正確に出来ている一方，シングルレーンに対しては1.5倍多く見積もってしまっている．

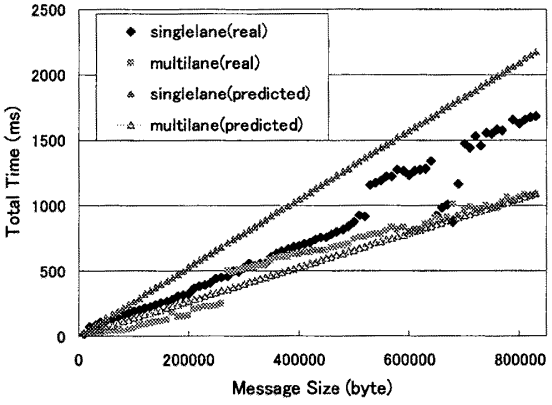

図2: 2サイト32ノードでのscatterアルゴリズムの比較 (遅延：8ms)

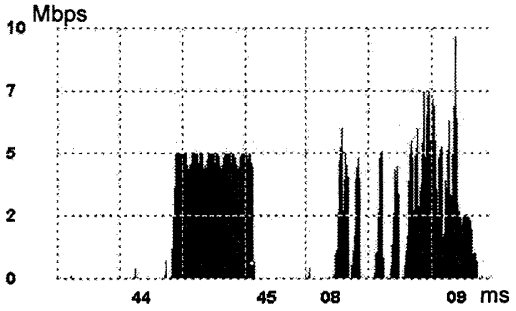

図3: WANのスループット (左：シングルレーン，右：マルチレーン)

5 おわりに

本稿では，WANを効率的に利用可能なマルチレーンツリー集団通信アルゴリズムを提案し，エミュレートしたグリッド環境でその有効性を確認した．今後の課題として，モデルによる見積もりコストを精度を高めることや，より大規模な実験環境でのマルチレーンツリーアルゴリズムの評価を考えている．

参考文献

[1] T. Karonis et al. *MPICH-G2: A grid-enabled implementatoin of the message passing interface.* JPDC 2003.

[2] M. Matsuda et al. *Efficient mpi collective operations for clusters in long-and-fast networks.* IEEE Cluster 2006.

[3] T. Chiba et al. *High-Performance MPI Broadcast Algorithm for Grid Environments Utilizing Multi-lane NICs.* IEEE CCGrid 2007.