広域ネットワークにおける大規模データ転送手法の検討

吉野 純平 † 阿部 洋丈 ‡ § 加藤 和彦 † §

1 はじめに

広域ネットワークを介して接続された多数の受信者に対してデータを一斉に転送する手法が多く研究されている．これらの研究成果は，ストリーミング配信や，ミラーサーバや CDN の構築のためのデータ転送等の性能向上に貢献することができる．

本研究の目的は，短時間でデータ配布が可能となる転送路の構築である．我々は，配布ツリーの形の工夫で短時間での転送を実現しようとする従来研究 [1] と同じ方針でこれを研究している．この研究との違いは複数転送元のデータ配布を目標としている点である．

データ配布の際にノードが他のノードへデータを少しづつパイプライン式に転送する方式においては，分岐が少ないツリー構造がデータ配布を短時間で完了させることができるのではないかと仮説を立てた．転送路が分岐していると転送するための帯域を分割することとなり，受信ノードの受信スループットが低下するためである．また，分岐が少なすぎると経由するノードが多くなり，遅延が蓄積されてしまうことでデータ転送時間が延びてしまうのではないかと考えたためである．

分岐の少ないツリー構造には転送先にネットワーク性能の差があるとパイプラインが詰まるという問題点があることが予測される．しかし，従来研究 [2] の複数の転送路を用意する手法を利用することで問題を解決できると考えられる．従来研究はボトルネックの影響を小さくするものであり，分岐が少ないツリー構造が有効性が高いと考えた．

我々は分岐の少ないツリー構造を構築するために，リング構造に注目した．リングは分岐の全くない形状である．この形状に分岐を意図的に作り出すことで分岐を抑えたツリー構造が構築できるためである．

そこで我々はリングにショートカットリンクを追加した構造から最小全域木を作成して分岐を抑えたツリー構造を作成する方式を考案した．最小全域木を作成する際の重みとしてデータの先頭部分の転送時間を利用した．

この転送路がデータを短時間で転送することに有効であるかを検証するため，実際にシステムを作成しデータを転送して測定した．測定の結果，同程度のネットワーク性能を仮定した場合，単純なツリーの形状を利用した場合に比べて短時間でデータを転送できることが確認できた．本稿では，リング構造が他の単純なグラフから作成した転送路に比べ短時間での転送に適していることを確認し，リング構造においてどのようなショートカットリンクの追加が効果的であるかを検討する．

† 筑波大学 システム情報工学研究科
‡ 豊橋技術科学大学 情報工学系
§ 科学技術振興機構 戦略的創造研究推進事業

2 関連研究

Overcast[1] は単一転送元のデータ転送手法である．10k バイトのデータ転送時間を利用してツリー構造の最適化を行っている．Narada[3] はメッシュ型の構造から DVMRP[4] を利用して最小全域木を作成することで配布ツリーを構築している．

SplitStream[2] や Bullet[5] は，ボトルネックの影響を小さくする方式を考案した．これらの方式と対立するものではなく，我々の方式にも同様の手法が適用可能である可能性がある．

3 方式の検討

提案する方式は，転送路レイヤに分岐の少ないツリー構造を作ることで短時間に転送が可能になるのではないかという着想によるものである．本節では，以下の 3 つの層に基づいて分岐の少ない転送路を構築する方法を述べる．

転送路レイヤ データを実際に転送する通信路
転送路作成レイヤ 転送路を作成するための通信路
物理ネットワークレイヤ 物理的な通信路

3.1 転送路作成レイヤ

分岐の少ない転送路レイヤを作成するためにはいくつか方式が考えられる．

方式 1 多量の接続を持つグラフから各リンクの性能を予測して分岐の少ない最小全域木を作成する
方式 2 最初から分岐の少ないグラフから部分的な性能予測で最小全域木を作成する

方式 1 の多数の接続を持つグラフを考えると課題が多いことがわかる．すべての接続について性能を予測するための測定は，大量の資源を利用することが想定される．また，測定ができ最適解を求めたとしてもすべてのノードにそれを伝えなくてはならない．

方式 2 で得られる分岐の少ないグラフから作成される最小全域木は，最適な転送路でない可能性が高いことが問題である．転送路の作成においては，最適化できる個所はグラフ中の分岐のある地点だけである．最適化を転送路作成レイヤにおける分岐点の周辺だけで行うだけで良いため分散環境に適している．これらの理由により方式 2 の転送路作成レイヤの形状を検討した結果，リングにショートカットリンクを追加する方式を考案した．

3.2 転送路レイヤ作成のための方式

転送路作成レイヤのグラフは，ショートカットリンクの影響で任意の 2 つのノードを選んだ際に複数の経路が存在する．本方式では，複数の経路がある場合はどちらかを選

択して使う方式を採用した．選択する手法は，転送データの先頭部分の転送完了時刻という転送性能を利用した．パラメータで与えた先頭部分のサイズまで先にデータを転送してきた経路を転送元として扱い，その他の経路からの転送を拒否する．この方式によって全域最小木を作成した．

4 実験

実験は提案手法を Java で実装し，IBM のブレード型計算機 HS20(Xeon3.6Ghz × 2, Memory2Gbyte, NIC1Gbps) を利用しの計算機 36 台を IBM ブレードセンターのネットワークスイッチと DELL PowerConnect5324 で接続した環境で行った．ソフトウエアは，Linux カーネル 2.6.23 と Java1.6 を利用した．

4.1 リング構造の有効性

リング，トーラス，2 分岐ツリー，完全グラフから作成した転送路と比較した結果を図 1 に示す．ツリーの測定では，ツリー構造の頂点からの転送結果を示す．ネットワークの性能に差がない今回の実験環境においてリングは他の方式より短時間でデータを転送できていることがわかる．

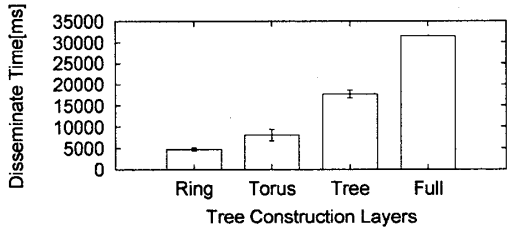


図 1 リングと他の転送路作成レイヤの比較

4.2 ショートカットリンクによる性能の変化

同程度の性能を持ったネットワーク環境においては，リングを利用した転送路が短時間でデータを転送できることが確認された．我々はさらに性能を向上させる方式として，ショートカットリンクの追加による方式を考え計測した．ショートカットリンクの接続は，すべてのノードの中から 2 つのノード選ぶ組み合わせは $p = {}_{node}C_2$ だけ考えられる．また，そのショートカットリンクの中から任意の本数を追加することを考慮すると，すべての組み合わせは 2^p とおり考えることができる．ノード数が 5 のとき 1024 通りとなるため現実的ではない．

今回の実験では，ショートカットリンクを張るノードを限定することとショートカットリンクを張った結果できるグラフが同形であるものを排除することで対象を絞った．具体的には，36 台の計算機に 0 から 35 の番号をリングにおける並び順に割り振り，その中から 0,9,18,27 というノードにおいてのみショートカットリンクを追加するという限定を行った．この結果として得られた 18 パターンについて実験を行った．

測定は 0,4,8,...,32 の 9 地点から転送した．ショートカットリンクの効果を測定した結果の平均と標準偏差を図 2 に示す．横軸がパターンの番号を示し，縦軸が配布にかかった時間を示した．特徴的な結果であったパターン 2 とパターン 7 とパターン 9 のショートカットリンクの貼り方を図 3 に示す．図 3 は，各パターンにおいてリングに追加

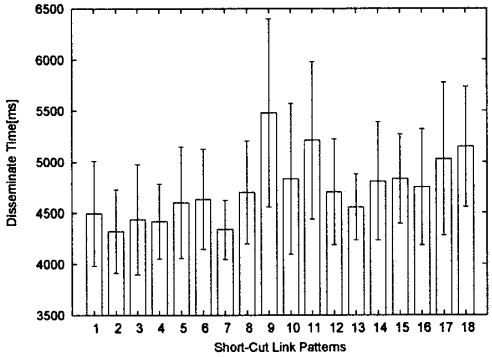
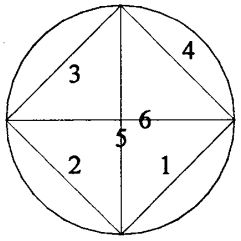


図 2 ショートカットリンク数による性能の変化



パターン	リンクの番号
パターン 1	なし
パターン 2	6
パターン 7	1,3
パターン 9	1,4,5

図 3 ショートカットリンクの張り方

されるショートカットリンクを示したものである．ショートカットの張り方によって転送性能に差が生じることが確認できた．

5 まとめ

本稿では，リング構造にショートカットリンクを追加することで分岐の少ないデータ転送路を作成する方式の性能を測定した．リング構造はネットワーク性能が同程度の場合において短時間にデータを転送できることが確認できた．今後の課題として，ネットワーク性能が異なる状況での測定や関連研究の手法を利用した性能向上が挙げられる．

参考文献

[1] Jannotti, J., Gifford, D. K., Johnson, K. L., Kaashoek, M. F. and James W. O'Toole, J.: Overcast: reliable multicasting with on overlay network, *OSDI'00* (2000).

[2] Castro, M., Druschel, P., Kermarrec, A.-M., Nandi, A., Rowstron, A. and Singh, A.: SplitStream: high-bandwidth multicast in cooperative environments, *SOSP '03* (2003).

[3] hua Chu, Y., Rao, S. G. and Zhang, H.: A case for end system multicast, *SIGMETRICS '00* (2000).

[4] Waitzman, D., Partridge, C. and Deering, S. E.: Distance Vector Multicast Routing Protocol (1988).

[5] Kostić, D., Rodriguez, A., Albrecht, J. and Vahdat, A.: Bullet: high bandwidth data dissemination using an overlay mesh, *SIGOPS Oper. Syst. Rev.*, Vol. 37, No. 5 (2003).