

# 広域並列分散システムのための信頼できるマルチキャスト

下國 治、古賀久志、陣崎 明

新情報処理開発機構 並列分散システム富士通研究室

〒211-8588 川崎市中原区上小田中 4-1-1

*E-mail: {osamus, koga, zinzin}@flab.fujitsu.co.jp*

数千台の計算機がインターネット経由で接続された広域並列分散システムにおいて、信頼性のあるマルチキャスト (Reliable Multicast) 通信を利用して一対多通信機構を実現することを検討する。パケットロス時にパケット配布元から再送するリカバリー時間を短縮するため、配送経路途中にあるルータから再送することを考えた。経路のすべてのルータに再送可能な機構を持たせるのはコストがかかるので、再送の効果が大きな箇所に優先的に再送可能なルータを配置し、再送の効果が大きな箇所とはどのような特性を持つのか、実際のネットワークを基にした通信モデル上で検討した。検討の結果、バックボーンネットワークとローカルネットワークの境界にあるルータにだけ再送させれば十分な効果が得られるとの結論を得た。

## A Reliable Multicast for Wide-Area Parallel and Distributed Systems

Osamu Shimokuni, Hisashi Koga, Akira Jinzaki

Parallel and Distributed Systems Fujitsu Laboratory, RWCP

1-1, Kamikodanaka 4-Chome, Nakahara-ku, Kawasaki, 211-8588 Japan

*E-mail: {osamus, koga, zinzin}@flab.fujitsu.co.jp*

In this paper, we discuss multicast implementations on wide area parallel and distributed systems which connect large number of computers through the Internet by reliable multicast communication. Since packet retransmission from the original sender lead to large recovery time, we investigate an approach in which the retransmission is made by intermediate routers on the multicast routes. It would need noticeable cost, if all routers had retransmission ability. Therefore, it is important to place those routers only on effective location. We study characteristics of the effective location on the multicast routings. We examine the above approach on our communication model which reflects the characteristics of the actual networks. And we obtain the result that it is enough to locate routers supporting packet retransmission only on the boundaries between the backbone network and the local area networks.

## 1. はじめに

近年、ATM をはじめとした広域ネットワーク媒体の転送性能向上に従い、世界中に分散した PC やワークステーションをインターネット経由で接続し、並列計算環境を構築する試みが盛んである。

並列計算においてデータを1つのノードから複数の他のノードに送るマルチキャストは出現頻度の高い重要な機能である。例えば、MPI では collective communication 命令としてライブラリレベルでサポートしており、こうしたライブラリレベルでのマルチキャストを実際の通信にどうマッピングするかは性能に影響を与える大きな要因となる。

広域網においては、単純に考えると IP マルチキャストにマッピングすることになるが、現状の規格ではパケット到達の信頼性がなく並列計算に適さない。広域並列計算に関する既存研究では、信頼性確保のため受信者毎に TCP コネクション（つまりユニキャスト）を1本ずつ張ってマルチキャストを実現する [1]。しかし、この方式では受信者の数に比例した通信オーバーヘッドがかかるためスケーラビリティに欠ける。実際、Albatross Project [5]ではローカルに構成された PC クラスタを4箇所相互接続するに留まっている。

我々は、次世代インターネットにおいて、数百から数千サイトを接続した大規模な広域並列計算環境を構築することを目的に研究し、前に述べたスケーラビリティの欠如を解決するため、近年提案された通信保証を行う IP マルチキャスト通信（リライアブルマルチキャスト）の利用を検討している。すでに並列計算に適したリライアブルマルチキャストが持つべき性質を調べ、プロトコル Comet RM の提案を行った [3][4]。Comet RM は配送経路途中に存在するルータを簡単な構成にするというポリシーで設計され、パケットロス時の再送を必ず配布元が行う。この結果、

- ルータが再送用バッファを持たなくてよく、またタイマーなど通信状態の管理もしなくてよい

という利点がある一方で

- パケットロスに対する回復時間が大きい

という弱点も有していた。

本稿では、逆に途中のルータがパケットを溜め、パケットロス時に再送元となるモデルを考慮する。このモデルではネットワークを構成するルータ群のどれにパケットを保持させるかで様々なバリエーションが考えられる。我々は、元々の Comet RM も含めいくつかのバリエーションをモデル化して検討を行い

- パケットがすべての受信者に確実に受信されるのに必要な時間（伝達遅延）
- ルータで必要となる再送用バッファ量

を比較して、その結果からルータが再送用バッファを持つ方式の有効性について考察した。

以下に本稿の構成を示す。第2節では、まず我々が想定するネットワークモデルを明確にし、次にリライアブルマルチキャストプロトコル Comet RM を振り返る。第3節でルータが再送元となるプロトコルモデルを説明する。第4節で各モデルに対して検討を行い、その比較結果を元に考察する。

## 2. 準備

本節ではまず想定するネットワークモデルについて説明し、次に Comet RM プロトコル [3] を説明する。

### 2.1. ネットワークモデル

本稿では物理的に離れた学内 LAN あるいは企業内 LAN に存在するクラスター計算機や単独の計算機間を JGN (Japan Gigabit Network) [6] のような高速バックボーンネットワークで結ぶ疎なマルチキャストネットワークを仮定する。以下にその特性を示す。

- 基本的に高帯域である。帯域幅は学内、企業内 LAN では数 Gbps、バックボーンは数百 Mbps である。
- エラー率については M. Yajnik [2]らが行った MBone (Multicast Backbone) パケットロス率測定結果に従う。バックボーンではパケットロスが極めて少なく (0.2%)、むしろ学内 LAN で数%のパケットロスが発生する。
- リンク遅延は隣接バックボーンルータ間で数十 ms であり、学内、企業内 LAN では数 ms である。

マルチキャスト経路はバックボーンでほぼ固定され、変更は起きないと仮定し、経路変更に関しては以下の議論では取り扱わない。

## 2.2. Comet RM プロトコルの概要

Comet RM は以下のような特徴を持つ。

- ACK パケットによる到達確認を行う

到達確認は TCP と同じく ACK パケットによって行う。ACK パケットの中身は受信側ノードのバッファサイズ (window サイズ) 及び、どこまでパケットデータ列を受け取ったかを示すシーケンス番号からなる。送信側ノードは TCP のようなスライディングウィンドウを持つ。

- パケットの再送は送信元ノードからのみ行う

配送経路途中のルータはパケットを送出したら、すぐバッファを解放できる。言い換えると、ルータは再送バッファを用意しない。

- ルータで ACK パケットの統合及びパケット再送の抑制を行う

ルータは到着する ACK パケットを監視して、自分の各子ノードがどこまでパケットを受信したかを示すシーケンス番号を記憶する。そして、この記憶された値とデータパケット、ACK パケットに付加されたシーケンス番号を比較して、ACK パケットの統合および、パケット再送の抑制を行う。

ルータは、マルチキャストセッション毎に自分の子ノードがどこまで受け取っているかを記憶するだけで、パケットフォワーディング時にテーブルを検索してシーケンス番号比較をする程度なので非常に単純な構成になる。再送バッファも持たない。

## 3. ルータによる再送方式

Comet RM では配布元からのみパケットを再送するので、パケットロス時の (I) タイムアウトによる ACK パケット不到達検出、(II) その後のパケット再送、にそれぞれ端点間の遅延 (広域網では数百 ms) を要する。この方式は平均伝達遅延を増大させ、遅延が問題視される広域並列計算にとっては短所となる。また、配布元ノードも RTT と通信帯域の積に比例する大容

量な再送バッファを用意せねばならない。

この問題点への対応としてルータから再送を行うモデルを検討する。この方式の長所は、不到達検出およびパケット再送に必要な時間が短縮されることである。欠点としては、ルータでタイムアウトを検出するためのタイマー管理、再送バッファを用意しなければならず、ハードウェア量が増大することが挙げられる。

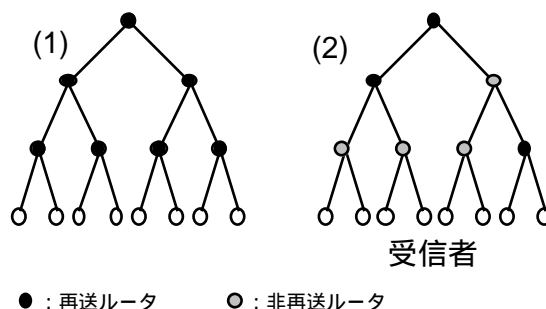


図1 ルータ再送方式

ルータによる再送方式において、選択しうるバリエーションは大きく以下の2種類に分けられる (図1)。

(1) すべてのルータが再送を行う方式

(2) 配送経路の中に再送ルータと非再送ルータが存在する方式

以下では、パケット再送を行うルータを再送ルータ、そうでないものをパケット非再送ルータと呼ぶ。

純粋に伝達遅延を小さくするだけなら(1)の方式が最良である。しかし、上述のようにルータのコストが増大することを考慮すると、(2)の方式でルータを適切に配置することにより、ネットワーク全体としてのハードウェア量増加を抑制しつつ、できるだけ(1)に匹敵する平均伝達遅延を達成することが望ましい。最適な配置はリンクのパケットロス率やリンク遅延などに依存すると考えられる。

以下の章ではネットワークのどこに再送ルータを置くべきか、またその時に達成される伝達遅延は(1)と比較してどの程度悪化するかについて検討し、Comet RM、(1)、(2)方式の有効性について考察を行う。

## 4. 検討

この節では、前出の3つのモデルそれぞれの優劣を検証するため、伝達遅延に関する検討を行う。まず、パ

ケットエラー率、リンク遅延といったパラメータが伝達遅延に与える影響の基本的性質を見るため、送信元から1つの受信者までの非常に単純な経路（リンク数3）に着目して検討する。次に、その結果を利用して、2.1節で想定したネットワークモデルに近いエラー率、リンク遅延パラメータ、経路の長さ（リンク数8）で検討をし、同時にネットワーク全体に必要なバッファ量を見積もる。その後、枝分岐の影響について述べる。

### 検討モデル

実際のインターネットは様々な特性をもったネットワークの集合体であるが、本論文の検討では、それぞれのネットワークを媒体の遅延  $T_i$ 、エラー率  $E_i$  というパラメータで特徴付けられるとし、これらの値を基に、伝達遅延の期待値を求めた。ルータの処理遅延はネットワークの伝達遅延と比較すると無視できる程小さいので、考慮しない。

検討 1-3 では議論を明確にするために、図 2 のように、ある受信者への単純な経路を仮定し、検討を行った。

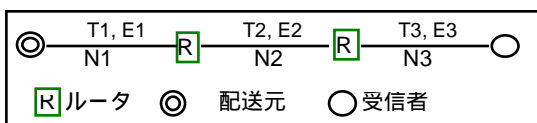


図 2 検討 1-3 で用いる単純な経路

図 2 中の 2 つのネットワーク接続点に再送ルータを置くか否かで以下の表のように 4 通りのバリエーションがあり、それぞれについて遅延の期待値を求めた。

	再送ルータ	備考
A	全く置かない	Comet RM に相当
B	点 に置く	(2) の一つの例
C	点 に置く	(2) の一つの例
D	点 、 共に置く	(1) に相当

### 4.1. 均一なネットワーク

まず、均一なネットワークでの比較を行う。それぞれのネットワークは総て同じ特性を持ち、一般的なバックボーンネットワークの数値より、エラー率、伝達遅延を以下の表のように定めた（検討 1）。

	N1	N2	N3
T ( msec )	3	3	3
E ( % )	すべて同一値 (0.1--10)		

この結果より、ルータでバッファリングすることの効果があることが確認された。

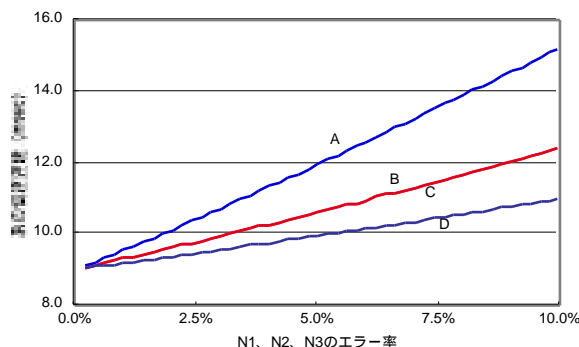


図 3 検討 1

### 4.2. エラー率が高いネットワーク

経路にエラー率の高いネットワークが含まれている状況を考える。まず、N3 がエラー率の高いネットワークとし、N3 のエラー率を変化させて系全体の伝達遅延の変化を観察した（検討 2.1）。

	N1	N2	N3
T ( msec )	3	3	3
E ( % )	0.1	0.1	0.1 -- 10

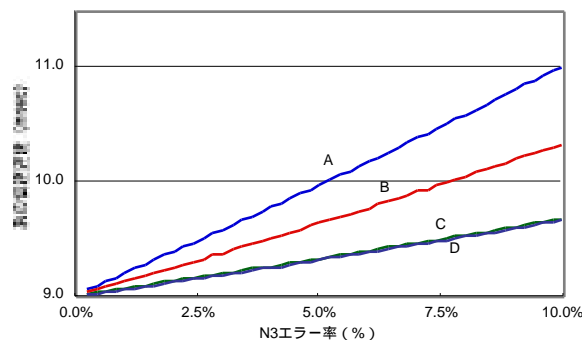


図 4 検討 2.1

検討 2.1 では、B と C ではエラー率の高いネットワークの直前に再送ルータを置く C が伝達遅延が短く、エラー率が変化しても D と殆ど同じ値をとることが確認できる。

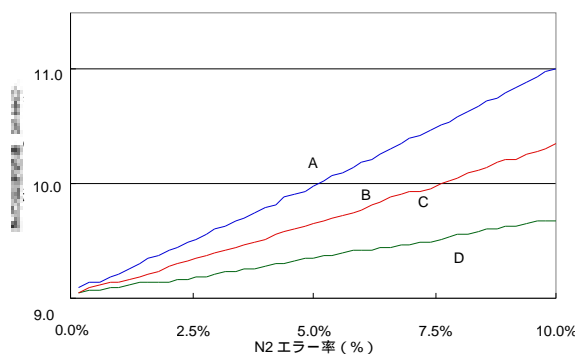


図 5 検討 2.2

次に N1、N3 を固定し、N2 がエラー率が高いネットワークになるとして同様の検討を行った（検討 2.2）。検討モデルの対称性から B、C 共に同一の値をとり、A と D の中間の値をとることが分かった。

これらから、以下の結果を得た。

- エラー率の高いネットワークがある場合、そのネットワーク前後に再送ルータを置くと伝達遅延を小さくできる
- エラー率が低いネットワークが連続している場合、再送ルータで中継する効果は低い
- エラー率が高いネットワークがある場合、そのネットワークを含む再送区間の合計の遅延が大きくなれば、全体の遅延の増大に寄与する

### 4.3. リンク遅延が大きいネットワーク

次に遅延が大きなネットワークが含まれている状況を考える。検討 2 と同様に再送ルータの位置を変化させ、その違いを調べた。今回も以下のように N3 の媒体の遅延時間を変化させ、全体の伝達遅延時間の期待値を計算した（検討 3）。

	N1	N2	N3
T ( msec )	3	3	3 -- 40
E ( % )	0.1	0.1	0.1

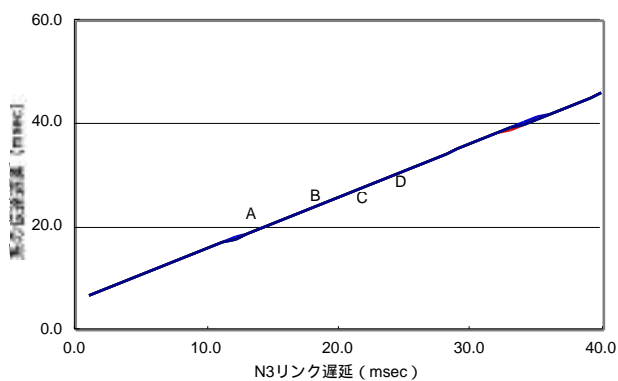


図 6 検討 3

N3 の遅延が 40msec になっても系の伝達遅延は A と D で 0.2%の差しか出なかった。

また、エラー率が 10%の際には、B と C を比較して、検討 3 と同様に C が伝達遅延が小さく、C は D と同様の伝達遅延であることが分かった。

これらから、以下の結果を得た。

- エラー率の低いネットワークの場合、リンク遅延

が遅延の主要因となり、再送ルータを用いる効果は小さい

### 4.4. より実際的なネットワーク

検討モデルを 2.1 節で我々が最初に仮定したネットワークモデルに近付けて、実際的なネットワークでの振舞いについて考察する。

MBone を参考にしてリンク遅延、エラー率 [2]を仮定し、配布元、受信者間に 8 つのネットワークを介する場合について検討を行った。端のネットワークにはローカルネットワーク、中間のネットワークはバックボーンの特性を与える。

	N1	N2-N7	N8
T ( msec )	3	30	3
E ( % )	5.0	0.2	5.0

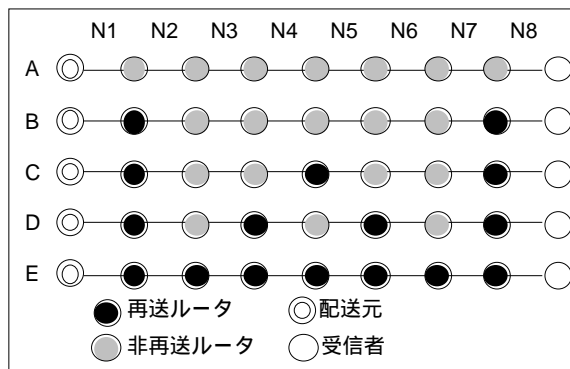


図 7 検討 4 のネットワーク

この条件のもとで、バックボーンに置く再送ルータを図 7 のように、各段、1 段おき、2 段おきと変化した場合の性質を検討した。

また、このモデルでバッファ量についても考察する。ここで仮にこのネットワークが 2 分木状に構成されているとする（受信者は  $2^7=128$  台）。あるルータにおける再送のためのバッファの大きさは、次の再送ルータもしくは受信者までの RTT を T (sec) とし、帯域を B (bps)とすると、BT で計算できる。

N1, N8 のネットワーク帯域を 1Gbps、その他の帯域を 622Mbps と仮定する。系全体の通信帯域は経路の帯域の最小値よりは大きくなるので、 $B = 622$  Mbps と仮定する。

検討の結果、この条件の下で、各段で再送した場合には、全く再送を行わない場合の約 80%の時間で転送

が行なわれることが分かった。また、エラー率が小さいバックボーンでは、再送ルータの配置密度を増してもあまり効果がないことが分かった。

	伝達遅延		バッファ量 (MByte)	
	期待値 (msec)	A に対する比率	系全体	1 台あたり最大量
A	231	1	28.9	28.9
B	191	0.826	58.3	28.0
C	189	0.817	156.2	14.0
D	188	0.813	226.3	9.33
E	187	0.810	324.2	4.67

#### 4.5. 分岐点

今までの検討では一経路における伝達遅延を扱ってきたが、マルチキャストでは当然、ルータで複数のリンクにパケットを送る。図 9 のように のルータから n (n = 1, 2, ..., 10) 本のリンクがある場合の伝達遅延を検討した。

	N1	N2	N3
T (msec)	3	3	3
E (%)	0.1	0.1	0.1

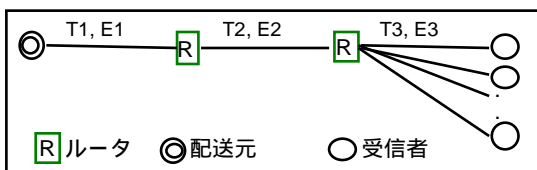


図 9 検討 5 のネットワーク

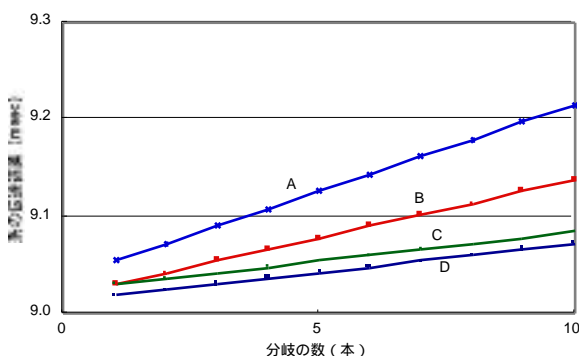


図 10 検討 5

分岐点では接続しているリンクが増えるに従ってエラー率が高くなる。よって、分岐が多い点に再送ルータを置くと伝達遅延を減少できる事が分る。

#### 5. まとめ

リライアブルマルチキャストを使って広域並列計算環境を構築する際に問題となる平均伝達遅延を低くする

ため、ルータがパケット再送を行なう方式を検討した。特に (1) すべてのルータが再送を行なう方式、(2) 再送ルータと非再送ルータが存在する方式、の 2 方式の比較を行なった。

実ネットワークに近いパケットエラー率、リンク遅延のモデル上で検討した結果、以下のことが判明した。

- (1)、(2) とともに、Comet RM より平均伝達遅延を 20%程度削減する。(2) に関してはバックボーンの両端に再送ルータを配置すれば十分で、それ以上再送ルータ配置比率を挙げても性能向上は見られない
- システム全体の再送用バッファ必要量については再送ルータ配置密度をあげると急激に増大する

これから、(2) でバックボーンとローカルネットワークの境界にあるルータに再送ルータを配置する方式が最良方式であると言える。

今後はこの検討結果を元に、実際のネットワークで実験を行い、再送ルータの効果を検証する予定である。

#### 参考文献

- [1] T. Keilmann. et al.: MAGPIE: MPI's Collective Communication Operations for Clustered Wide Area Systems, Proc. of Symposium on Principles and Practice of Parallel Programming (PPoPP'99), May 1999.
- [2] M. Yajnik, J. Kurose, and D. Towsley.: Packet Loss Correlation in the Mbone Multicast Network, Proc. of Global Internet Conference, Nov. 1996.
- [3] 古賀他: インターネットでの並列分散処理の実装検討, SWoPP'99, CPSY99-68, 1999 年 8 月.
- [4] 下國他: インターネットでの並列分散処理の方式検討, SWoPP'99, CPSY99-67, 1999 年 8 月.
- [5] <http://www.cs.vu.nl/albatross/>
- [6] <http://www.shiba.tao.go.jp/JGN/>