2003－DPS－113 (15)
2003／6／20

オーバレイネットワークを利用した 少人数グループマルチキャストの設計と実装

生野 徳彦 [†1] 寺岡 文男 [†1]

[抄録] これまで、マルチキャストグループに属する小人数のメンバを把握した Small Group Multicast のプロトコルとして、Xcast などいくつかの手法は提案されてきたが、いずれも通信基盤のルータに改良しなければならない、通常の IP マルチキャストと比較して配送経路が冗長になる、などの理由により普及には至らなかった。
そこで本研究では、物理ネットワーク上のエンドホストによって仮想的なオーバレイネットワークを構築し、そのネットワーク上で Small Group Multicast を実現する Overlay ATcast を提案する。Overlay ATcast は物理ネットワーク上のルータに変更を加えることなくマルチキャストを実現でき、仮想ネットワーク上にルータを設置するため Application-layer Multicast と比べて冗長な経路が発生しにくい、という利点を持つ。

Design and Implementation of Small Group Multicast with Overlay Network

Naruhiko IKUNO [†1] Fumio TERAOKA [†1]

[Abstract]Until now, although some techniques had been proposed for Small Group Multicast protocol such as Xcast, which understand the group member each other ,they have not spread for the reasons of a delivery course becoming redundant as compared with the usual IP multicast, or having to put a hand into a router of communication infrastructure.
Therefore, in my research, a virtual overlay network is built by the end host on a physical network, and Overlay ATcast which realizes ATcast on the virtual overlay network is proposed. Overlay ATcast can realize ATcast without adding change to the router on a physical network infrastructure, and has the advantage of seldom generating a redundant course compared with Application-layer Multicast.

1 はじめに

1.1 背景

マルチキャストは、複数のノードが同時にデータを送受信する際に効率良く配送するためのプロトコルである。しかし、これまではルータがマルチキャストに対応しなければならないという通信基盤面での問題などから、あまり普及することがなかった。しかし一方で、ここ数年でインターネットは ADSL の導入などの影響により急速に世の中に普及してきた。またこれと共に、これからは携帯端末がより社会に普及していくものと思われる。このような環境では手軽に情報交換ができるようになることから、マルチキャストを用いることによって、より効率的なデータの送受信が可能になる様々なアプリケーションに対するニーズが、これまでよりも大きくなると考えられる。

またマルチキャストの中でも PIM-SM[1] のような、不特定多数の受信者を想定したマルチキャストとは異なり、マルチキャストグループに属する小人数のメンバを把握した Small Group Multicast を利用するユーザが増加することが想像され、そのために様々な特色を持った Small Group Multicast に対応する技術が必要になってくると思われる。これまで Small Group Multicast を実現するものとして、Xcast となどがあったが、これらのプロトコルはいずれも既存のマルチキャストと同様にルータが各々のプロトコルに対応する必要があるという問題点がある。

この問題点を解決する手法として、マルチキャストのグ

ループに属するエンドホスト同士で配送木を構築し、データを送受信する Application-Layer Multicast と呼ばれる方式を利用した Narada が提案されているが、これは既存のマルチキャストと比較すると、配送経路が冗長になりがちでかつ、経路がより複雑になり、メンバホストに負担がかかるなどの問題点がある。このように、これまで提案されてきたプロトコルはそれぞれ機能的な欠点を抱えているために普及するに至らなかった。

一方、これまで IP マルチキャストについて述べられてきた別の問題点として、データ転送に UDP を用いているために信頼性を保証していないという点がある。

以上のことから、ルータを用いたマルチキャストの利点を生かしつつ、通信基盤となるルータに手を加えなくてはならないという問題を解決し、信頼性を保証する Small Group Multicast の新たなプロトコルを考慮する必要性があると考えられる。

2 関連研究

2.1 Small Group Multicast

不特定多数の受信者を想定した、グループの規模が大きい一般的なマルチキャストに対して、Small Group Multicast は多くとも 10 人程度のグループを想定としたマルチキャストである。そのため、通常のマルチキャストは、受信者が欲しいデータを選んで受信する放送型であるので、送信者はグループのメンバを知ることなく、またメンバの動的な変化に影響されずに送信することができるが、Small Group Multicast ではメンバを把握した通信を行うため、このような放送型の通信を実現することはできない。

Small Group Multicast の例として Xcast[2] というものがある。Xcast ではグループアドレスを用いず、受信者のアドレスをリストにしてヘッダに書き込むことによって、マルチキャストグループに属するメンバにデータを送信する。Xcast の問題点としては、ヘッダを処理する際に Xcast 特有の処理を要するために通常の生活基盤上でのルータでは対応できないこと、そしてグループに属する全てのメンバのアドレスリストをヘッダに書き込むために、パケットヘッダのオーバヘッドが多くなってしまうことなどが挙げられる。

2.2 Application-Layer Multicast

ルータがマルチキャストに対応していない場合でも、マルチキャストを実現するために Application-Layer Multicast が提案された。Application-Layer Multicast は、通常のマルチキャストでは IP 層で実装していたメンバーシップ管理やパケットの複製を含むマルチキャスト機能をメンバとしてマルチキャストグループに join しているエンドシステムで実装する。このような Application-Layer Multicast の例としては Narada[3] のようなプロトコルが提案されている。

しかし Application-Layer Multicast には、物理層リンクでのパケットの冗長な配送経路が発生するという問題や、IP マルチキャストと比較するとエンドホスト間での遅延が大きくなるという問題がある。

3 設計

3.1 3 階層の抽象化

1.1 章でも述べたとおり既存の Small Group Multicast を含めた IP マルチキャストの主な問題点は、マルチキャストの経路制御やパケット複製を実現するために、既存のルータに手を加える必要があるということである。これまで、この問題を解決する手法としては、Application Layer Multicast が提案されている。その中でトポロジー的に散在した小人数のメンバを対象としたマルチキャストのプロトコルとして、Narada が既に提案されている。ところが、Application Layer Multicast はマルチキャストグループのメンバ同士で配送木を構築するので、一般的な IP マルチキャストと比較すると、データの転送経路が冗長になりやすい。さらに、メンバの加入や離脱によってその都度、配送木を再構築するのでプロトコルとしては脆弱となってしまうという欠点がある。したがって本研究では、Application-Layer overlay network を構築し、その上で Small Group Multicast を行う手法を提案する。

具体的な手法としては物理ネットワーク上ではエンドホストであり、トポロジー的に外部のネットワークと近い位置にあるなどの、要所に設置されている端末同士でオーバレイネットワークを構築し、それらの端末上にマルチキャストルータのデーモンを実行させることによってルータの役割を担わせる。マルチキャストグループに join したいメンバホストはオーバレイネットワーク上にあるマルチキャストルータに join request を送信する。このようにして Small Group Multicast を実現する。この様子を図 1 に示した。

このような手法を用いることによって、通信基盤上のネットワークルータに手を加えることなく Small Group Multicast を実現でき、また stress や stretch も Application Layer Multicast と比較すると通常の IP マルチキャストに近い、理想的なシステムを実現することが可能になる。

Overlay ATcast は、オーバレイネットワーク上のマルチキャストルータや、メンバホストに対してオーバレイネットワーク独自のアドレス機構である virtual address を用いる。この virtual address は network を特定するための prefix フィールドと network 内での PC を特定するための host フィールドから構成されており、トポロジー情報を反映させることが可能である。ネットワークを識別するための prefix フィールドと同一の network 内での PC を特定するための host フィールドにそれぞれ 32bit の充分なスペースを設けることにより、スケーラビリティを保証することができると考えられる。

しかしこのオーバレイネットワークは実際のネットワーク上に構築される仮想ネットワークであるので、コネクションを確立する際には各々の Overlay ATcast ルータとなるホストのインタフェースに IP アドレスを付けられて

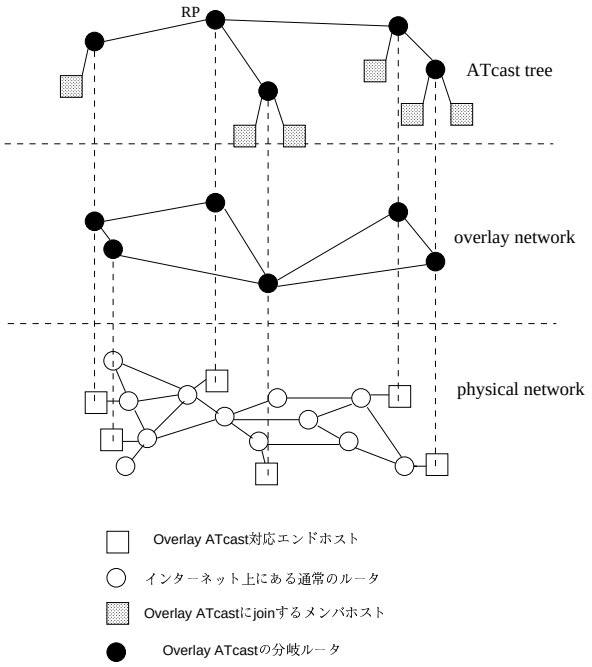

図 1: オーバレイネットワーク上で実現する Small Group Multicast

いる必要がある。これにより、必然的にネットワークインタフェースについての、IP address と virtual address との対応づけをするためのシステムが必要となる。そこで本研究では Overlay ATcast ルータのルーティングデーモンに virtual address と IP アドレスの対を持つテーブルを設け、そのテーブルを用いてコネクションを確立することによってオーバレイネットワークを構築する手法を用いた。

また、これまでのマルチキャストプロトコルはデータ転送に UDP を用いていたのに対して、Overlay ATcast 上のマルチキャストルータ間では TCP を用いたコネクションを確立することによって、ルータ間でのデータグラムの転送において、信頼性を保証することが出来る。ルータ間のコネクションはオーバレイネットワークが構築された後は、このオーバレイネットワーク上で実行されるアプリケーションがすべて終了するまで維持される。一方、メンバホストとルータ間のデータグラムの転送にも TCP コネクションを用いて行うが、このコネクションはマルチキャストデータの送受信やマルチキャストグループへの加入またはグループからの離脱に関するメッセージのやりとりの際にのみ確立され、それら以外の場合ではコネクションは切断される。

3.2 配送木構築

配送木を作成する際には、情報を保持するルータをなるべく減らすために、ネットワークを抽象化し、分岐点にあたるルータのみが自分の下流にあるルータの情報を保持することによってルータのメモリを節約する。このように配送木を抽象化するという特徴から、このプロトコルは ATcast(Abstruct Tree multicast) と名付けられている。ただし、以下のような前提条件を必要とする。

- ATcast を利用するエンドホストはグループアドレスと RP(Rendezvous Point) アドレスをあらかじめ知っている
- すべてのルータが ATcast に対応している

ATcast はマルチキャストグループアドレスを用いて Small Group Multicast を行い、共有木を配送木として構築するプロトコルである。グループメンバの把握は、共有木の根に RP を置き、RP がメンバの認証を行うことによって実現される。

3.3 データ送信

Overlay ATcast のグループに属しているメンバはデータを送信する際、一度 RP にデータを送信し RP から配送木を用いてグループメンバ全体に対してデータが送信される。

3.4 前提条件

Overlay ATcast を設計・実装する上で前提条件となることは以下の通りである。

- 管理権限を持った管理者がオーバレイネットワークに属する全てのホストのトポロジー情報を認識している
- Overlay ATcast に属する全てのホストはグループアドレスと RP アドレス (virtual address)、及びオーバレイネットワーク上でルータの役割を担うホスト (virtual router) は自分も含めた隣接する virtual router の virtual address 及び IP address と、この virtual router の下に属するメンバホストの virtual address、一方グループに属するメンバホストは自分が属する virtual router の virtual address を知っていなくてはならない

virtual address は管理権限を持った管理者によって付けられるので、トポロジー情報を考慮した virtual address を付けるためには、管理者がオーバレイネットワークに属する全てのホストのトポロジー情報を認識していることは必要である。

また ATcast を用いているので、メンバがマルチキャストグループに join するためには、グループアドレスと RP アドレスを知っていなければならない。また、データの転送には通常のルーティングプロトコルではなく、独自のルーティングを行っているために、virtual router であるホストは隣接しているルータの virtual address を知っている必要がある。

4 実装

オーバレイネットワークを構築するホストは、ホスト間で経路制御を行うためにルーティングデーモンを実装し

ている。ATcast のデーモンはそれらのルーティングデーモンの上に実装している。そして、デーモン間でプロセス通信を行うことによって、経路制御を行いながら ATcast を実現する。図 2 にその様子を示す。

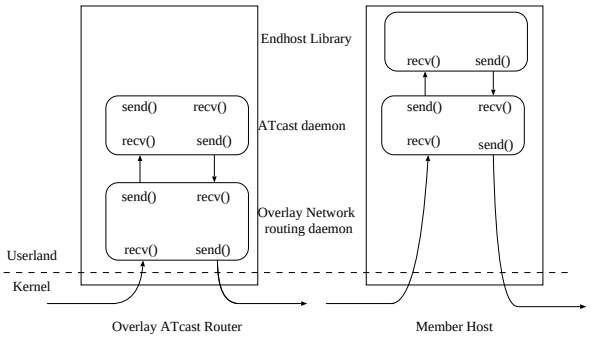

図 2: 同一ホスト内でのデーモン間通信

各ルータ上で動いているデーモン同士がコントロールメッセージをやりとりすることによって ATcast の配送木が形成される。一方ルーティングデーモンでは virtual address と IP address をマッピングすることによって実際のネットワーク上を ATcast のメッセージが正しく転送されるために制御を行う。これによってオーバレネットワーク上のデーモンから見ると、メッセージのフィールド内に virtual address のみを用いて経路制御できるのが可能であるようになる。

4.1 ルータ用デーモン

Overlay ATcast ルータではユーザとのインタフェースの役割を果たす Endhost Library は必要ないが、メンバ情報の管理や配送木を構築するためのメッセージの処理を行う必要がある。またデータ転送の際に、近隣ルータと TCP コネクションを確立することによってオーバレイネットワーク上で 1 ホップずつデータグラムを転送する処理を行うために、メッセージの宛先が自ルータでないものを処理する必要もある。そのようなメッセージの例として RP がメンバホストに対して、マルチキャストグループへの加入を認証する際に送信する AUTH_REPLY メッセージを用いて、Overlay ATcast ルータでのメッセージフロー及び ATcast デーモンの状態遷移の様子を図 3 に示した。なおこの図で JOIN_ACK と共に示されている点線部の直線は、ATcast デーモンの状態が JOIN_WAIT でない時 ATcast デーモンはエラーを出力し、メッセージを破棄してしまうことを表しているものである。図 3 に示したメッセージのうち、四角で囲んだ AUTH_REPLY は宛先フィールドが自ルータでないものを示している。メッセージには、宛先となる virtual address を保持しているフィールドがあるので、その宛先の virtual address を引数としてルータ用のルーティングデーモンが next hop となるルータの virtual address を返り値とする。このように宛先アドレスを記入するフィールドをメッセージ内に設け、そのフィールドをルーティングデーモンが参照して適切な処理を行うことによって、ATcast デーモンにとって必要ないメッセージを受信してしまうことを防ぎ、冗長な処理時間を省く。

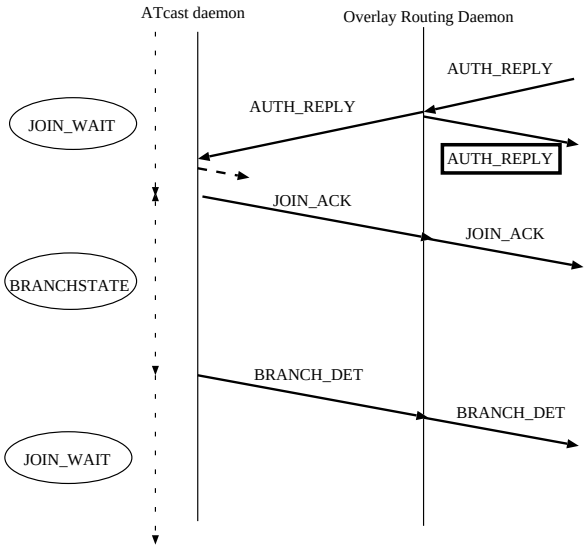

図 3: AUTH_REPLY メッセージの流れと ATcast デーモンの状態遷移

4.2 メンバホスト用デーモン

オーバレイネットワークを構築する Overlay ATcast ルータとは異なり、Overlay ATcast のグループメンバのホストはオーバレイネットワークの構築に関与しないため、ルーティングデーモンは必要ない。しかし ATcast デーモンを Overlay ATcast のプロトコルに対応できるように変更しなくてはならない。オーバレイネットワーク上の ATcast ルータのルーティングデーモンとコネクションを確立するために基になる ATcast のエンドホスト上に実装されていた ATcast のデーモンを変更した。メンバホストではルーティングを行う必要はないのでルーティングデーモンは必要ない。

5 評価

5.1 実装環境下での評価

Overlay ATcast において、関連研究で紹介した ATcast と比較した際に、オーバヘッドとなるのは二つの要因が考えられる。まず一つ目は、ルーティングデーモンでの処理である。そして二つ目は、物理ネットワーク上のルータとオーバレイネットワーク上のルータとのトポロジーの差によって生じる、冗長な配送経路による遅延である。そこで、本研究では、各々の種類のオーバヘッドを以下のように定義した。

- O_{Br} :Overlay ATcast 上の分岐ルータでの処理時間。純粋にルーティングデーモンでの処理時間と、ルーティングデーモンから ATcast デーモンへデータパケットが渡される時間との総和。
- O_R :ルーティングデーモンでの処理された後、ATcast デーモンへ転送されることなく、次のルータまたはメンバホストへ転送される間の処理時間。ATcast では、通常のルーティングの処理のみであるため、ルータのカーネル空間のみの処理となる。
- O_T :物理ネットワーク上のルータとオーバレイネットワーク上のルータとのトポロジーの差によって生じる冗長な配送遅延。

以上のオーバヘッドの測定値を図 1 に示す。 O_{Br} の処理

表 1: O_{Br},O_R,O_T の測定結果

	msec
O_{Br}	1.86
O_R	1.17
O_T	0.72

のオーバヘッドが大きくなる原因としては、ルーティングデーモンでの処理時間分、余計に時間がかかるということだけでなく、オーバレイルータや ATcast 機能をデーモンプロセスとして実装したため、パケット送受信のたびにコンテクストスイッチが発生することが理由になっていると考えられる。一方 O_R で Overlay ATcast のオーバヘッドが劇的に増大する原因としては、ATcast ではカーネルのみの処理であるのに対して、Overlay ATcast ではユーザ空間内のルーティングデーモンであるデーモンプロセス部で処理が行われることが最大の理由であると考えられる。

5.2 任意なトポロジーでの評価

以上の評価結果を用いて、任意のトポロジーについての Overlay ATcast の ATcast に対するオーバヘッドについて考えてみると、次節のようになる。

5.2.1 メンバの Join による配送木再構築の時のオーバヘッド

まず最初に、メンバが Join する際の ATcast に対する Overlay ATcast のオーバヘッドについて考えてみる。関連研究でも示した通り、メンバの Join による配送木の再構築は以下の手順を踏む。

1. join したいメンバが RP 方向の上流ルータへ join message を送信
2. join message を受信したルータは、(RP , G(グループアドレス)) のエントリを持っていない、もしくは状態なしの中間ルータの場合は、そのまま join message を RP 方向へ転送 RP、または (RP , G) のエントリを持っている分岐状態の中間ルータの場合、ルータが waiting 状態でなければ join message の送信者 (メンバ) を下流ルータのリストに追加し、join-waiting 状態になる
3. waiting 状態になった分岐ルータは authentication check を RP へ送信
4. authentication check を受信した RP はメンバリストをチェックした後、authentication reply を送信、一定時間 authentication reply が来なかった分岐ルータは authentication check を RP へ再送
5. authentication reply を受信した分岐ルータはメンバに対して join-ack を送信、下流ルータに next hop が等しいルータがある場合は branch detection を送信
6. next hop が同じ場合はそのまま branch detection を転送、next hop が異なる場合は {G, RP, 下流ルータ } のエントリを作成し、branch detection の送信者に対して branch notification を送信、分岐ルータを受信した場合上流へ branch notification を送信し、更に下流へ branch detection を転送
7. branch notification を受信したルータは (G , RP) のエントリを更新し、join-waiting 状態を解除。さらに、branch notification 中の下流ルータを削除、branch notification の送信者を下流ルータに追加、branch notification を受信しなかったルータは branch detection を再送

このような上で挙げた手順において、各々の段階におけるオーバヘッドを示し、その総和が全体でのオーバヘッドとなる。また RP と Join するメンバとの間に分岐ルータが既に存在する場合としない場合とでは処理が異なるので、オーバヘッドも異なってくる。これを踏まえて各々の段階でのオーバヘッドを示していく。ただし、文中の N_B, N_R,x はそれぞれ、メンバホストと RP との間にある分岐ルータの数, 分岐していないルータの数, メンバホストから RP 方向に向かって最もメンバホストに近い分岐ルータ (以降これを最隣接分岐ルータと呼ぶ) までの、分岐していないルータの数を示している。

RP までの上流方向の Overlay ATcast 分岐ルータがない場合
全体でのオーバヘッドは
$5.58N_R+1.17(N_R+2)+2.28N_R = 9.03N_R+2.34$(msec)
新たな分岐ルータが発生しない場合のオーバヘッドは
$2N_RO_R+(N_R+1)O_{Br}+3N_RO_T = 6.36N_R+1.86$(msec)
となる。
以上により、分岐状態でない Overlay ATcast ルータ数とオーバヘッドは比例関係にあり、分岐状態でない Overlay ATcast ルータが一つ余分にあるだけでも、オーバヘッドが劇的に増大してしまうことがわかる。

RP までの上流方向の Overlay ATcast 分岐ルータがある場合
全体でのオーバヘッドは
$5.03x + 3.82N_R + 1.91N_B + 3.1$(msec)
新たな分岐ルータが発生しない場合は
$3.14x + 3.82N_R + 1.91N_B + 1.24$(msec)
となる。

以上により、オーバヘッドに最も大きな影響を与えるのは、分岐状態でない Overlay ATcast ルータであり、そのなかでも特に、メンバホストと最隣接分岐ルータとの間にある分岐状態でない Overlay ATcast ルータであることがわかる。

5.2.2 メンバの Leave による配送木再構築の時のオーバヘッド

全体でのオーバヘッドは
$3.72x + 2.34x + 0.72(2x + 2) = 7.5x + 1.44$(msec)
または
$2.34N_R + 0.72(2N_R + 4) + 7.44 = 3.78N_R + 10.32$(msec)
となる。
したがって、分岐状態でない Overlay ATcast ルータ数とオーバヘッドは比例関係にあり、分岐状態でない Overlay ATcast ルータが一つ余分にあるだけでも、オーバヘッドが劇的に増大してしまうことがわかる。

上流方向に分岐ルータがない場合は、branch removal を送信する必要がないので、処理としては RP と leave message や leave ack をやりとりするのみとなる。したがってオーバヘッドは
$2O_{Br} + 2N_R O_R + (2N_R + 2)O_T = 3.78N_R + 5.16$(msec)
これは結果的に、上流方向の Overlay ATcast 分岐ルータがある場合での、最隣接分岐ルータが branch detection を送信しない場合の一例としてまとめられる。

5.2.3 データ送受信時

データを送受信する際の、ATcast に対する Overlay ATcast のオーバヘッドは以下のようになる。ただし N_{BS},N_{RS},N_{BJ},N_{RJ} はそれぞれ、データを送信するメンバと RP との間にある分岐ルータの数 , 分岐していないルータの数 , 及びデータを受信するメンバと RP との間にある分岐ルータの数, 分岐していないルータの数を表している。
$1.86(N_{BJ} + 1) + 1.17(N_{BS} + N_{RS} + N_{RJ}) + 0.72(N_{BS} + N_{RS} + N_{BJ} + N_{RJ})$(msec)
となる。

任意のトポロジーでの、ATcast に対する join や leave、またデータ配送の際に生じるオーバヘッドと、オーバレイネットワーク上の分岐ルータでないルータの数、また join や leave を行うメンバホストにとっての最隣接分岐ルータとの間にある、オーバレイネットワーク上の分岐ルータでないルータの数との間にはそれぞれ比例関係が成り立っている。したがって、オーバレイネットワークルータの数を出来るだけ制限することによって ATcast に対する Overlay ATcast のオーバヘッドも減少させることが出来る。しかし、逆にオーバレイネットワークルータの数があまりに少ないと、メンバホストとルータとのトポロジーが離れてしまうため、メンバホストが操作を行う際のメッセージのやりとりに時間がかかってしまうという問題もある。このようなことから、オーバレイネットワークルータをどのように適切に配置するかが今後の課題となる。

6 まとめ

本研究ではマルチキャストグループのメンバを把握した、Small Group Multicast のために提案されているプロトコルである ATcast をオーバレイネットワーク上で実現し、信頼性を付加するデータ配送を行うことを可能にする Overlay ATcast を提案し、実装した。

Overlay ATcast は ATcast と比較して、以下のような特徴を持つ。

- 利点
  - 物理ネットワーク上の既存のルータに手を加えることなく Small Group Multicast を実現できる
  - データ転送において信頼性を保証する
- 欠点
  - ルータでのメッセージやデータパケットの処理などのオーバヘッドが大幅に増加してしまう

これまでマルチキャストプロトコルが普及しなかった重要な要因の一つとして、物理ネットワーク上のルータに手を加える必要なあったことを考えると、上で挙げたような利点は非常に有意義であるといえるであろう。しかしその一方で、オーバレイネットワーク上でのルーティング機能や ATcast 機能をプロセスとして実装したために、パケットを送受信する度にコンテキストスイッチが発生してしまうことにより、オーバヘッドが増大してしまうことがわかった。したがって、これらの機能をプロセスとして実装するのではなく、カーネル内に実装することによってオーバヘッドを減少させることが今後の課題と考えられる。

†1 慶應義塾大学大学院理工学研究科

†1 Graduate School of Science and Technology, Keio University.

参考文献

[1] D.Farinacci and A.Helmy and D.Thaler and S.Deering and M.Handley and V.Jacobson and C.Liu and P.Sharma and L.Wei, “Protocol Independent Multicast-Sparse Mode(PIM-SM):Protocol Specification”, RFC 2362, IETF, Jun. 1998.

[2] R.Boivie and N.Feldman and Y.Imai and W.Livens and D.Ooms and O.Paridaens, “Explicit Multicast(Xcast) Basic Specification”, Internet Draft, IETF, April. 2002.

[3] Y.-H. Chu and S.G. Rao and H.Zhang, “A case for End System Multicast”,In Proceedings of ACM SIGMETRICS, Jun, 2000.