分散システム特論レポート

Protocol Independent Multicast-Sparse Mode

情報工学専攻 修士1年 荒木研究室
2000年1月31日
大森 幹之

1　はじめに

IPマルチキャストは、IP上で複数のグループメンバに対してデータグラムを効率的に配送するために提案された。マルチキャストを用いると、既存の配送方法であるユニキャストに比べ、ネットワークの負荷や資源利用を軽減することができる。マルチキャストは、イーサネットのような共有型のローカルネットワーク上では、問題なく動作している。しかし、ローカルのサブネットから外のネットワークへマルチキャストでデータグラムを配送する場合には、マルチキャスト経路制御プロトコルを用いて、マルチキャストの配送木を構築する必要がある。

ここでは、マルチキャスト経路制御プロトコルの一つであるPIM-SM(Protocol Independent Multicast-Sparse Mode)の紹介をおこなう。

2　マルチキャスト

2.1　マルチキャストとユニキャスト

図1は、ホストα からIPのデータグラムの送信を行なう場合のデータの流れを示している。図1において、左側のユニキャスト通信の場合では、同じ2つのデータが網線部分のリンク上に流れているのに対して、右側のマルチキャスト通信の場合では、1つのデータがルータAにおいて複製されて効率的に流れている。

このように、同一のデータを複数のホストに配送を行なう場合には、ユニキャストよりもマルチキャストを用いた方が、使用する帯域を減少させることができ、効率的なデータの配送を行なうことができる。

2.2　用語

1. グループ

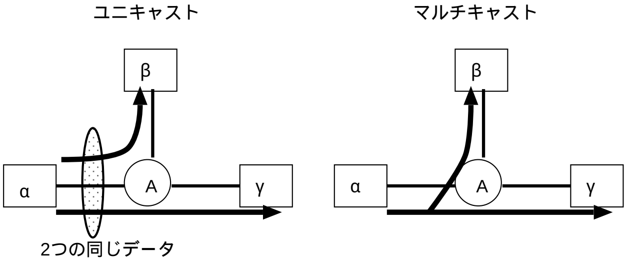

図1: ユニキャストとマルチキャスト

   IPでは、マルチキャストアドレスと呼ばれる特別なアドレスによってグループを識別する。IPv4では、クラスDアドレスが、IPv6では、上位1バイトがff(16進数)で始まるアドレスが、マルチキャストアドレスである。

2. メンバ

   IPでは、マルチキャストパケットを受信するホストを指す。グループに参加(join)することによってマルチキャストパケットの受信を行なう。途中からグループに参加(join)したり、途中で離脱(leave)することもできる。

3. 送信ホスト

   マルチキャストパケットをあるグループに送信するホスト。送信ホストは、必ずしもグループのメンバである必要はない。

2.3　マルチキャストの経路制御の必要性

図1のようなネットワーク構成では、IGMPのみでマルチキャスト通信が可能となるが、インターネットのような広域分散環境では、IGMPのみでは不可能である。

例えば、図2のようなネットワーク構成では、送信ホストに直接接続されているルータR_Aが、受信ホストのH_Bの存在がわからない。そのため、ルータR_AがマルチキャストパケットをH_Bに転送する

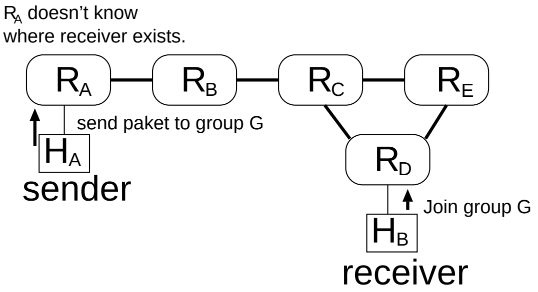

図 2: 経路制御の必要性

ことができない。

このことから、マルチキャスト通信を行なうためには、マルチキャストの経路制御が必要となることがわかる。

3 PIM-SM(Protocol Independent Multicast-Sparse Mode)

PIM-SM(Protocol Independent Multicast-Sparse Mode) とは、マルチキャストの経路制御を行なうプロトコルである。ユニキャストの経路制御プロトコルに依存しないことから、このような名称となった。

3.1 概要

PIM-SM では、マルチキャストパケットは、基本的に RP(Randezvous Point) を根とする共有木 (RPT:RP-Tree) を通じて各メンバに送信される。

図 3 に、PIM-SM によるマルチキャストパケットの配送の様子を示す。

マルチキャストパケットは、以下のようにして、送信ホストから受信ホストまで配送される。

1. 送信ホスト H_A がマルチキャストパケットを送信し、送信ホスト H_A に直接接続されているルータ R_A がマルチキャストパケットを受信する。
2. ルータ R_A は、マルチキャストパケットをカプセル化し、RP までユニキャストで送信する。
3. RP は、カプセル化されたパケットからマルチキャストパケットを取り出し、RP と受信ホスト H_B の間の経路上にあり、RP に隣接しているルータ R_C に転送する。

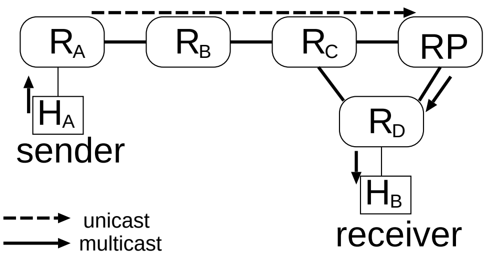

図 3: PIM-SM によるマルチキャストパケットの配送

3.2 ルータの保持している情報

ルータは、マルチキャストパケット配送するために送信ホストの IP アドレスやグループを表す IP アドレスなどの情報を保持しておかなければならない。このため、ルータは、主に次の 3 つエントリを必要に応じて作成し、保持している。

- (*,G) エントリ

  G は、グループを表し、*は、すべての送信ホストを表している。受信ホストがあるグループへ参加したとき、RP と受信ホストの間の経路上にあるルータ (RP も含む) が作成する。

  このエントリには、グループの IP アドレス、マルチキャストパケットの送出インターフェース、受信インターフェースなどの情報が含まれている。

- (S,G) エントリ

  主に、送信ホストに直接接続されているルータが保持しているエントリ。SPT(Shortest Path Tree) と呼ばれる配送木によってマルチキャストパケットが配送される場合には、送信ホストと受信ホストの間の経路上にあるルータも保持しておく必要がある。

  (*,G) エントリと同様に、グループの IP アドレス、マルチキャストパケットの送出インターフェース、受信インターフェースなどの情報が含まれており、それに加えて送信ホストの IP アドレスも登録されている。

- (*,*,RP) エントリ

  他のマルチキャスト経路制御プロトコルとの相互運用のためのエントリ。他のマルチキャスト経路制御プロトコルとの境界ルータは、PMBR(PIM Multicast Border Router) と呼ばれる。

3.3 マルチキャストパケットの送受信

3.3.1 ホストのグループへの参加

ホストが特定のグループに参加する場合は、IGMP (Internet Group Management Protocol) と PIM-SM の JOIN メッセージを用いることによって、ホストから RP までの経路の確立を行なう。

このとき次のような処理が行なわれる (図 4)。なお、図 4 の各ルータリンク上の数字は、インターフェースの識別番号であり、oif(outgoing interface)、iif(incomming interface) は、それぞれ、マルチキャストパケットの送出インターフェース、受信インターフェースを指す。

1. 受信ホスト H_B が IGMP Membership Query Message を送信し、グループ G への参加を表明する。
2. H_B に直接接続されたルータ R_D は、マッチするグループ G のエントリ (G,*) がないかを調べ、なければ、RP へ向かって JOIN メッセージを送信する。
3. RP は、(G,*) エントリがなかった場合には作成し、JOIN メッセージがきたインターフェースを oif として、(G,*) に追加する。

このようにして、受信ホストから RP までの経路が確立される。

3.3.2 ホストによるグループへのパケットの送信

送信ホストに直接接続されているルータが、RP までマルチキャストパケットを転送することによってグループに参加しているメンバ全員に配送される。RP からは、先に述べた RPT によって配送される。

このとき次のような処理が行なわれる (図 5)。なお、図 5 の破線と実線は、それぞれ、マルチキャスト、ユニキャストのパケットの流れを示す。

1. ホスト H_A がグループ G に対して、マルチキャストパケットを送信する。
2. ルータ R_A は、ホスト H_A からパケットを受けとると、(S,G) エントリ存在するかどうかを検証する。もし、(S,G) エントリが存在しない場合には、(S,G) エントリを作成し、
3. ルータ R_A は、マルチキャストパケットを Register メッセージにカプセル化し、RP へ送信する。
4. RP は、Register メッセージからマルチキャストパケットを取り出し、(*,G) エントリを見て、パケットの転送を行なう。

もし、グループ G のメンバが存在していない場合には、RP は、送信ホスト H_A に直接接続されているルータ R_A に Register Stop メッセージをユニキャストで送信し、無駄なトラフィックが生じることを防いでいる。Register Stop メッセージとは、Register メッセージの送信を抑制するものである。

また、パケットのカプセル化によるオーバヘッドを減らすことも考慮されている。このとき次のような処理が行なわれる (図 6)。

1. RP は、到着したパケットをカウントしていきあるしきい値以上のトラフィックが生じていることを検出すると、(S,G) エントリを作成する。
2. RP は、パケットの送信ホストまでのユニキャストの経路にそって、JOIN メッセージを送信する。
3. ルータ R_B は、(S,G) エントリを作成し、ルータ R_A へ JOIN メッセージを送信する。
4. R_A は、JOIN メッセージが送られてきたインターフェースを (S,G) エントリの oif に追加し、そのインターフェースへのパケットの転送を行なう。
5. カプセル化されていないパケットを受信した RP は、ルータ R_A に対し、Register Stop メッセージを送信し、ルータ R_A が Register メッセージを送信することを抑制する。

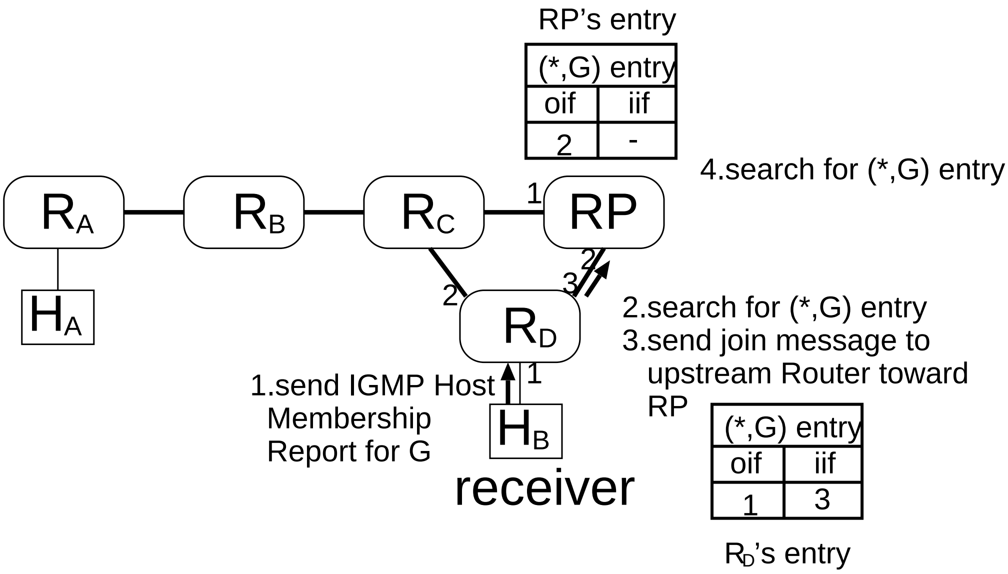

図 4: ホストのグループへの参加

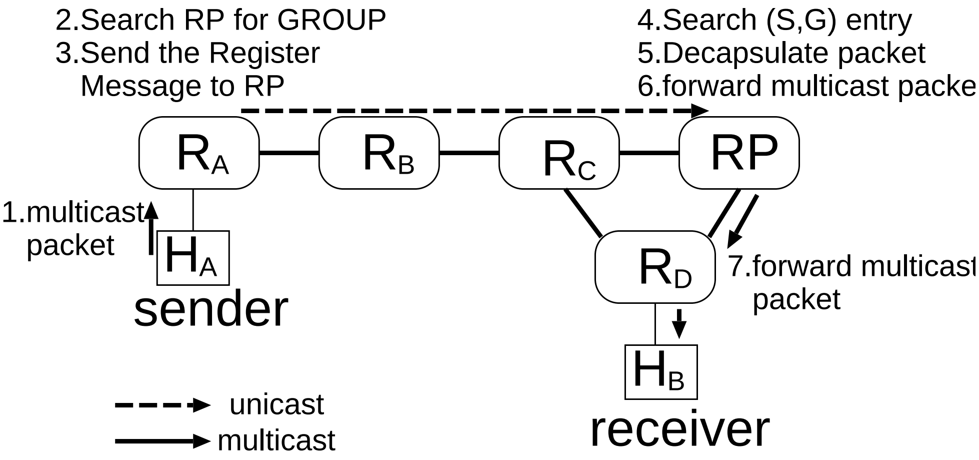

図 5: ホストによるグループへのパケットの送信

3.3.3 SPT(Shortest Path Tree) による効率なパケットの配送

PIM-SM は、通常では RPT でマルチキャストパケットの配送を行なう。しかし、あるしきい値以上のトラフィックが生じた場合には、より効率的な配送を行なうために、SPT(Shortest Path Tree) の形成を行なう。

このとき次のような処理が行なわれる (図 7)。

1. 受信ホストに直接接続されているルータ R_D が、受信したマルチキャストパケットの数を数える。
2. DR で計測されたマルチキャストパケットのトラフィックが、あるしきい値よりも大きい場合には、RPF を行ない、SPT までの経路を計算する。
3. もし、SPT による入力インターフェースが RPT のものと同じ場合には、SPT フラグが立っている JOIN メッセージを上流のルータへ送信する。

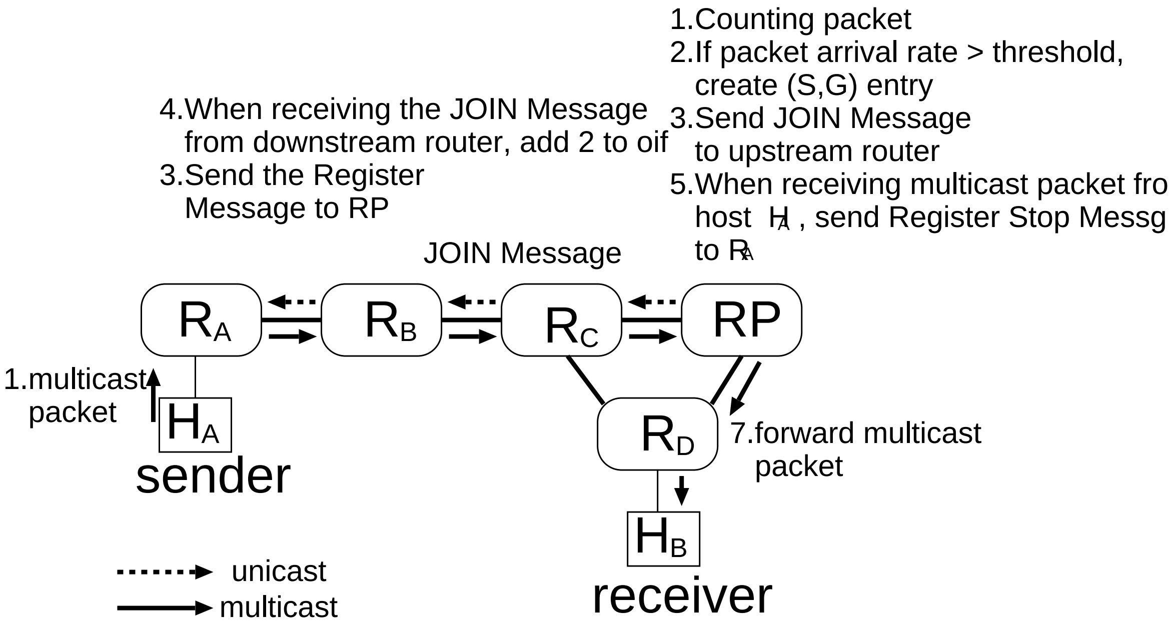

図 6: カプセル化を行なわない送信

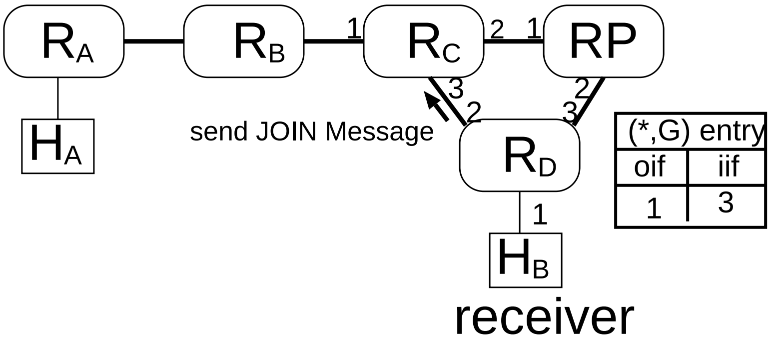

図 7: SPT への移行

もし、異なっている場合には、SPT にそったインターフェースに JOIN メッセージを送信する。その後 SPT からマルチキャストパケットを受信した場合には、RPT にそったインターフェースに PRUNE メッセージを送信する。

3.4 RP の決定

BSR(Bootstrap Router) と呼ばれるルータが RP の決定を行なう。もし、同じグループに対する RP 候補 (Cand-RP: Candidate RP) が複数いた場合には、その中から最も適する思われる RP 候補を RP として選出する。

なお、RP は、グループごとに決めることができ、1 つのグループに対しては、RP は、1 つしか存在しない。

RP の決定は、以下のように行なわれる。

1. 各 RP 候補による BSR への Cand-RP メッセージの送信
   このメッセージは、BSR(BootStrap Router) へユニキャストで送信される。
2. RP 候補が受け持っているグループに対する送信ホストがいた場合、送信ホストに一番近い RP 候補が RP として、BSR によって選出される。

3.5　*RP* の広告

BSR は、あるグループに対する *RP* がどのルータかをすべてのルータに広告するために、Bootstrap メッセージを送信する。Bootstrap メッセージには、BSR が選出した *RP* の IP アドレスとその *RP* が持っている

3.6　Assert メッセージ

Assert メッセージは、同一セグメントに複数のルータが存在する場合に、どのルータがそのセグメントへのマルチキャストパケットを転送するのかを決定するのに用いられる。

あるルータが同一セグメント上に存在している他のルータを検出すると、その時に存在している (S,G)、(*,G)、(*,*,RP) の各エントリ毎に経路の計算を行なう。((S,G) エントリの場合には、送信元ホスト S までの metric、(*,G)、(*,*,RP) エントリの場合には、*RP* までの metric を計算する)。この計算は、ユニキャストの経路表に基づき行なわれる。もし、経路が存在しない場合には、metric は、無限大となる。

4　実装

実装としては、KAME Project で開発されている pim6sd を紹介する。なお、pim6sd は、IPv6 上で動作する。

4.1　ルータの各エントリ

pim6sd に付属しているツール pim6stat を用いると、マルチキャストルーティングテーブルや各エントリの内容などを知ることができる。付録に添付した pim6stat の実行結果は、

4.2　実装とプロトコルの差異

4.2.1　*RP* の決定

RP の決定は、RFC2362 で規定されているものとは異なっており、動的ではなくルータの設定によって行なわれている。

4.2.2　SPT から RPT への移行

RFC2362 では、SPT から再び RPT へ移行することについては言及されていない。pim6sd では、SPT でのパケットの到着レートがあるしきい値よりも小さくなった場合には、再び RPT へ移行するように実装されている。

5　おわりに

PIM-SM では、あるグループに対して *RP* は 1 つしか存在しないため、送信ホストが複数いる場合には、*RP* を決定するのが非常に難しいものとなり、スケーラビリティの問題にもつながる。

また、*RP* と送信ホストとの間の経路に受信ホストが存在した場合には、無駄な帯域を消費してしまうという問題もある。

これらの問題点の解決策も提案されているが [2][3]、PIM-SM 自体が比較的新しいものであるために、PIM-SM がマルチキャストの経路制御プロトコルとして普及するには、もう少し時間が必要である。

参考文献

[1] D. Estrin, D. Frinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. liu, P. Sharma, L. Wei, "Protocol Independent Multicast-Sparse Mode (PIM-SM): Protocol Specification", RFC 2362, June 1998

[2] Dorian Kim, Verio, David Meyer, Cisco Systems, Henry Kilmer, Dino Farinacci, Procket Networks, "Anycast RP mechanism using PIM and MSDP", Internet Draft, Decemter 1999

[3] Deborah Estrin, USC/ISI, Dino Farinacci, Cisco Systems, "Bi-Directional Shared Trees in PIM-SM", Internet Draft, May 17, 1999

付録 pim6stat の出力

```
Multicast Interface Table
 Mif   PhyIF Local-Address/Prefixlen                     Scope Flags
  0     fxp0 3ffe:501:2c24:2100:290:27ff:fe08:a77e/64    0     DR NO-NBR
             fe80::290:27ff:fe08:a77e/64                 1
             Timers: PIM hello = 0:30, MLD query = 1:00
  1      de0 3ffe:501:2c24:2102:240:5ff:fea4:1692/64     0     DR NO-NBR
             fe80::240:5ff:fea4:1692/64                  2
             Timers: PIM hello = 0:30, MLD query = 1:00
  2     fxp1 3ffe:501:2c24:2101:290:27ff:fe1f:6bb/64     0     DR NO-NBR
             fe80::290:27ff:fe1f:6bb/64                  3
             Timers: PIM hello = 0:30, MLD query = 1:00
  3   dummy0 fe80::290:27ff:fe08:a77e/64                 9     DR NO-NBR
             3ffe:501:2c24:2100::1/56                    0
             Timers: PIM hello = 0:30, MLD query = 1:00
  4      lo0 ::1/128                                     0     DISABLED
             fe80::1/64                                  11
             Timers: PIM hello = 0:00, MLD query = 0:00
  5     pvc0 fe80::220:eaff:fe00:2a8e/64                 12    DR PIM
             Timers: PIM hello = 0:10, MLD query = 1:00
  6   regist 3ffe:501:2c24:2100:290:27ff:fe08:a77e/64    0     REGISTER
             fe80::290:27ff:fe08:a77e/64                 1
             Timers: PIM hello = 0:00, MLD query = 0:00

PIM Neighbor List
 Mif  PhyIF Address                                  Timer
 5     pvc0 fe80::220:eaff:fe00:1432                 75
```

```
Multicast Routing Table
 Source          Group           RP-addr         Flags
---------------------------(*,G)----------------------------
 IN6ADDR_ANY     ff1e::1394      3ffe:501:0:801:290:27ff:fe61:b0dd WC RP
Joined   oifs: .......
Pruned   oifs: .......
Leaves   oifs: ..l....
Asserted oifs: .......
Outgoing oifs: ..o....
Incoming     : .....I.
Upstream nbr: fe80::220:eaff:fe00:1432

TIMERS: Entry=0 JP=40 RS=0 Assert=0
  MIF   0   1   2   3   4   5   6   7   8   9
    0   0   0   0   0   0   0   0
---------------------------(S,G)----------------------------
 3ffe:501:2c24:2100:200:e2ff:fe12:3cb7 ff1e::1394      3ffe:501:0:801:290:27ff:fe61:b0dd SPT CACHE SG
Joined   oifs: ......j
Pruned   oifs: .......
Leaves   oifs: .......
Asserted oifs: .......
Outgoing oifs: ..o...o
Incoming     : I......
Upstream nbr: NONE

TIMERS: Entry=210 JP=0 RS=0 Assert=0
  MIF   0   1   2   3   4   5   6   7   8   9
    0   0   0   0   0   0   0   0
--------------------------(*,*,RP)--------------------------
Number of Groups: 1
Number of Cache MIRRORs: 1

---------------------------RP-Set----------------------------
Current BSR address: 3ffe:501:0:801:290:27ff:fe61:b0dd Prio: 10 Timeout: 110
RP-address                               IN  Group prefix     Prio Hold Age
3ffe:501:2400:0:2c0:4fff:fe68:4188       5   ff1e::4321/128   0    150  100
3ffe:501:0:1c01:200:f8ff:fe03:d9c0       5   ff05::/16        10   150  100
3ffe:501:0:801:290:27ff:fe61:b0dd        5   ff1e::/16        0    150  100
```