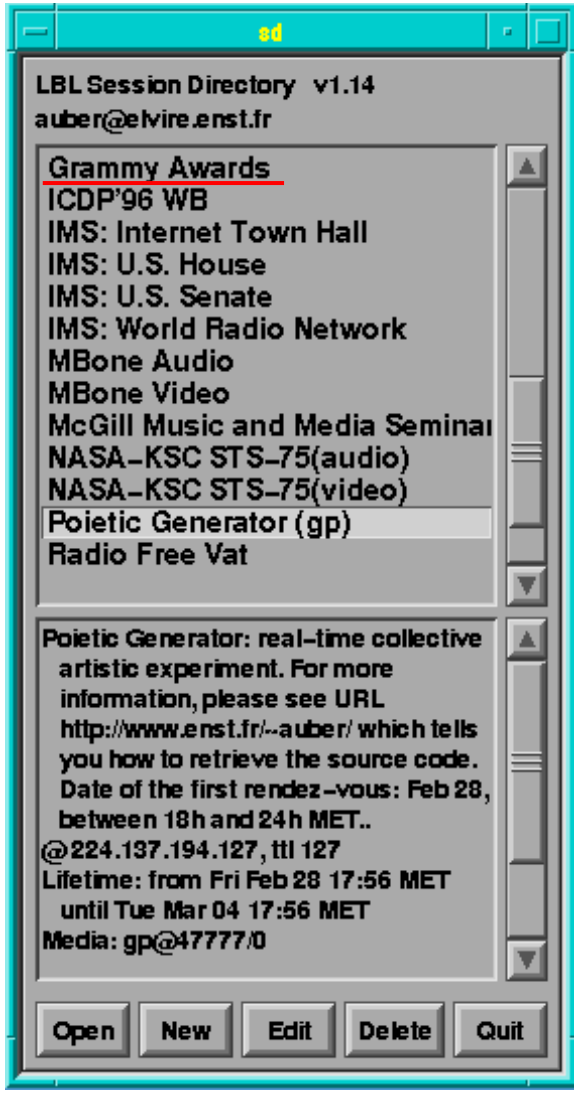
PIM-SM勉強会

2024/6/28

石黒邦宏

Multicastそれは人類の見果てぬ夢

1994年11月18日のMboneでのRolling Stonesのコンサート
当時 WIDE -> SRA の接続でSunのWorkstationで視聴していた



当時のMulticast放送のGroup管理ツールsdのキャプチャ
CableTVや衛星放送を目指していた

しかし。。。オンデマンドの利便性の前に破れ去る

マルチキャストの歴史は人類の夢を追い求めた歴史でもある。。。。

IP Multicastの歴史 – 全てはスティーブ・ディアリングから始まった

スティーブ・ディアリング

4の言語版

ページ　ノート　　閲覧　編集　履歴表示　ツール

出典: フリー百科事典『ウィキペディア（Wikipedia）』

スティーブン・ディアリング(Stephen Deering, 1951年 -)は、カナダ出身の計算機科学者である。シスコシステムズの元研究員で、Internet Protocol(IP)のアーキテクチャ拡張の開発と標準化に取り組んでいる。

スティーブ・ディアリング
Steve Deering

研究分野	計算機科学
研究機関	シスコシステムズ ゼロックス
出身校	ブリティッシュコロンビア大学 スタンフォード大学

プロジェクト:人物伝

テンプレートを表示

経歴 [編集]

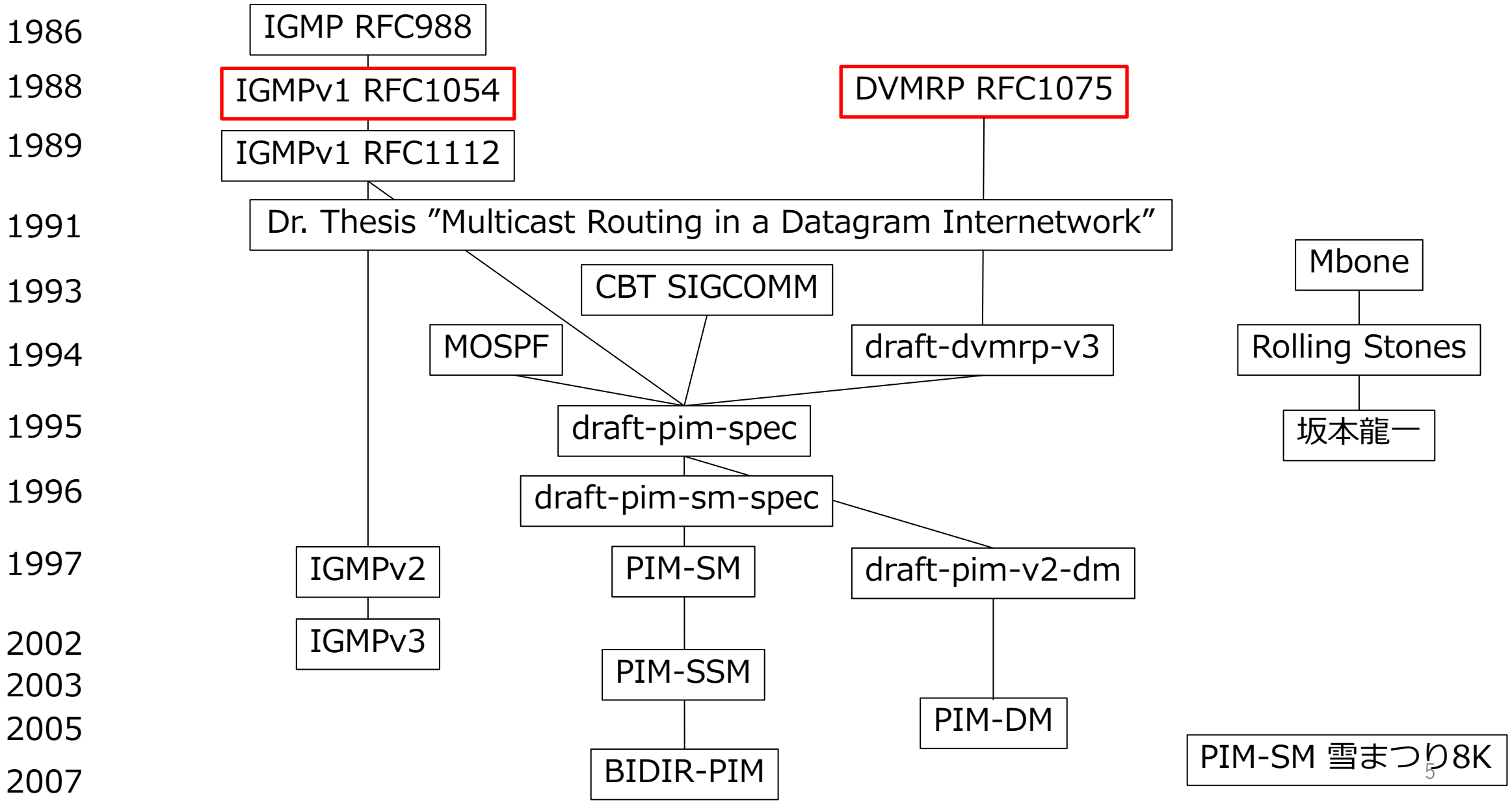
バンクーバー島のShawnigan Lake Schoolの高校を卒業後、ブリティッシュコロンビア大学で1973年に学士(B.Sc.)、1892年に修士(M.Sc.)の学位を取得し、1991年にスタンフォード大学でPh.D.を取得した[1]。その後、ゼロックスのパロアルト研究所に6年間勤務し、マルチキャストルーティング、モバイルインターネットワーキング、スケーラブルアドレッシング、インターネットでのマルチメディアアプリケーションのサポートなど、先進的なインターネット技術に関する研究に従事していた。1996年にシスコシステムズに入社した。

彼はインターネットアーキテクチャ委員会(IAB)の元会員であり、Internet Engineering Task Force(IETF)の多くのワーキンググループの元議長である。IPマルチキャストの発明者であり、新しいバージョンのInternet ProtocolであるIPv6の設計を主導している。

2010年、IPマルチキャストとIPv6に関する業績に対してIEEEインターネット賞が授与された[2]。1994年、Internet Talk Radioにより"Geek of the Year"に選ばれた[1]。

https://ja.wikipedia.org/wiki/スティーブ・ディアリング (最終更新 2021年7月8日 (木) 02:26 UTC)

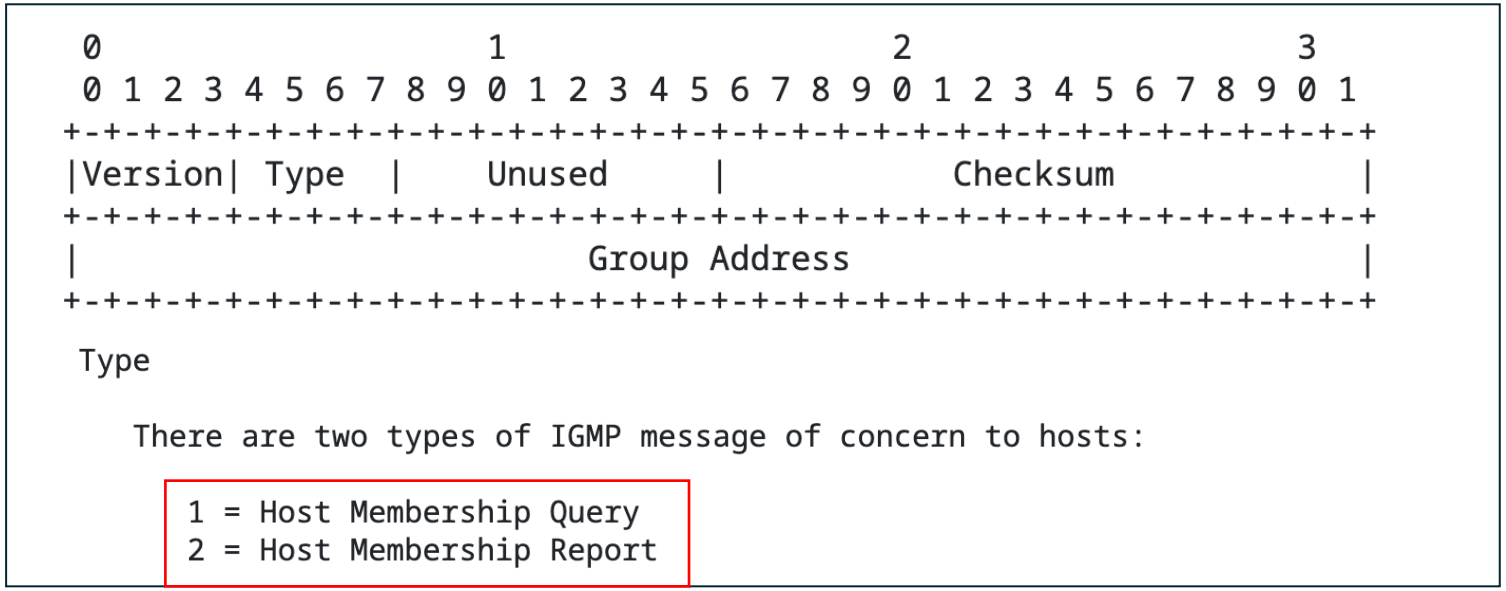
IP Multicastの歴史



1986年 IGMPv1とDVMRP どちらもスティーブ・ディアリング著

IGMPのIP Protocol番号は栄えある２番　インターネット黎明期のまだ牧歌的な時代

0x01	1	ICMP	Internet Control Message Protocol	RFC 792
0x02	2	IGMP	Internet Group Management Protocol	RFC 1112



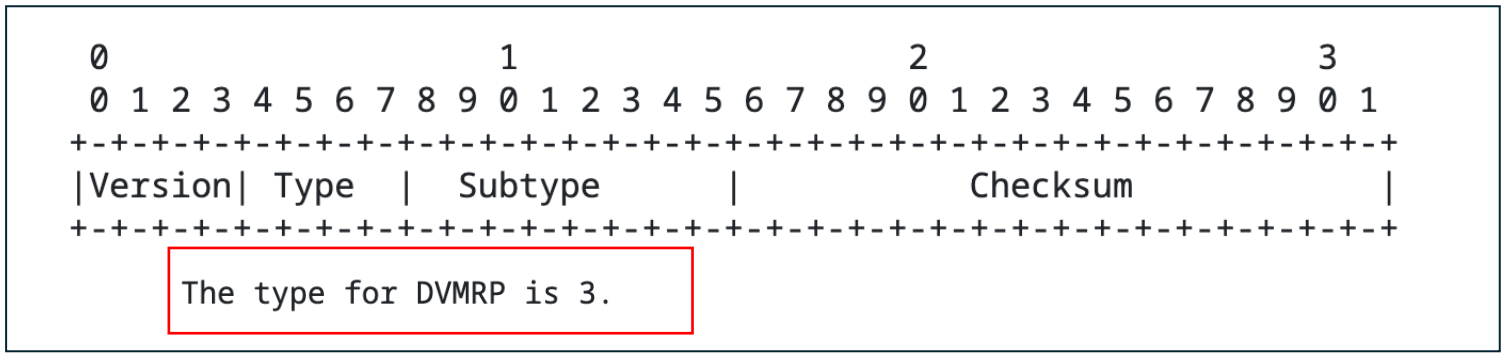
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |    Unused     |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Group Address                         |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Type

There are two types of IGMP message of concern to hosts:

1 = Host Membership Query
2 = Host Membership Report

RFC1054でIGMPv1が策定され
QueryとReportにそれぞれType 1, 2
が割り当てされる



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Type  |  Subtype      |           Checksum            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

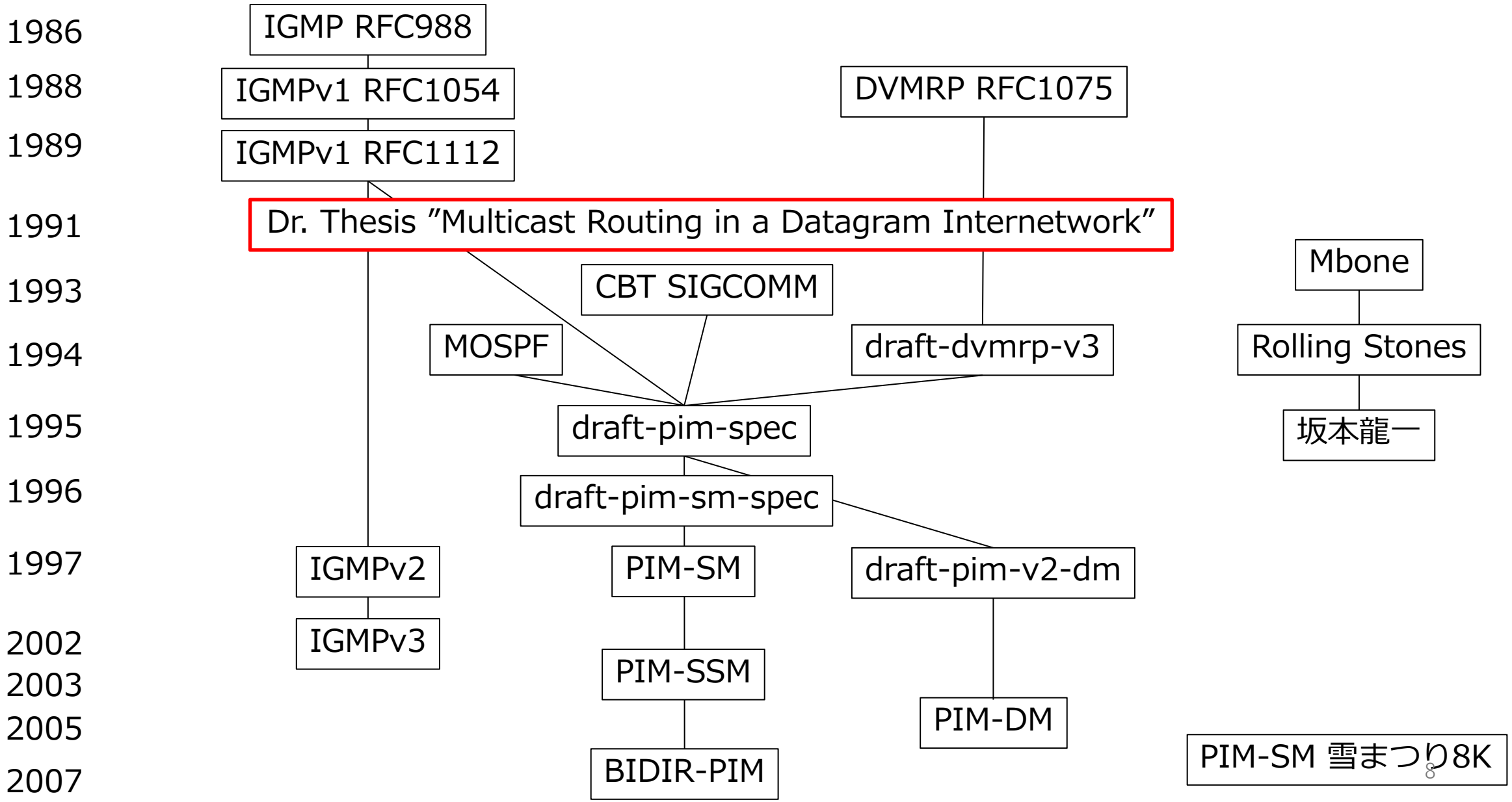
The type for DVMRP is 3.

RFC1075でDVMRPが策定され、
Type 3が割り当てられた
SubtypeでDVMRPのメッセージ
タイプが判別できるようになった

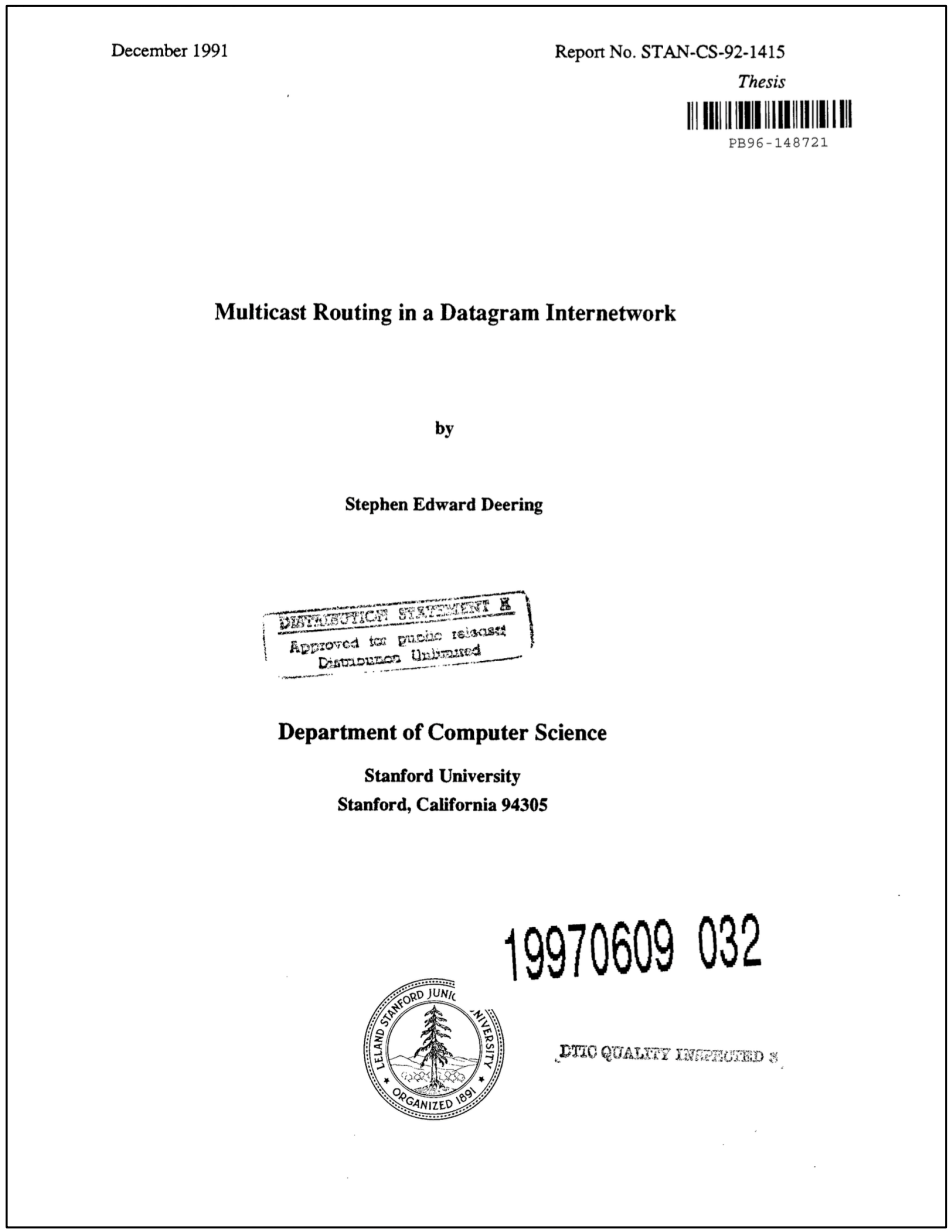
世界初のマルチキャストルーティングプロトコル - DVMRP

- アルゴリズムはのちにDense Modeと呼ばれるもの
- Flood & Pruneとも呼ばれる
- とりあえずDVMRPをしゃべる全ルーターにマルチキャスト経路を公告(Flood)
- そのあとリスナーがいない経路を刈り取る仕組み(Prune)
- 基本的な仕組みはRIPと同じ
- routedをもとにmroutedが実装される
- しかしループが怖い。。。。
- RIPが最大16 Hopなので、とりあえず倍の32 Hopにしとくか！という適当さ

IP Multicastの歴史



1991年 - スティーブ・ディアリングの博士論文



December 1991

Report No. STAN-CS-92-1415

Thesis

PB96-148721

Multicast Routing in a Datagram Internetwork

by

Stephen Edward Deering

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

Department of Computer Science

Stanford University
Stanford, California 94305

19970609 032

LELAND STANFORD JUNIOR UNIVERSITY ORGANIZED 1891

DTIC QUALITY INSPECTED 3

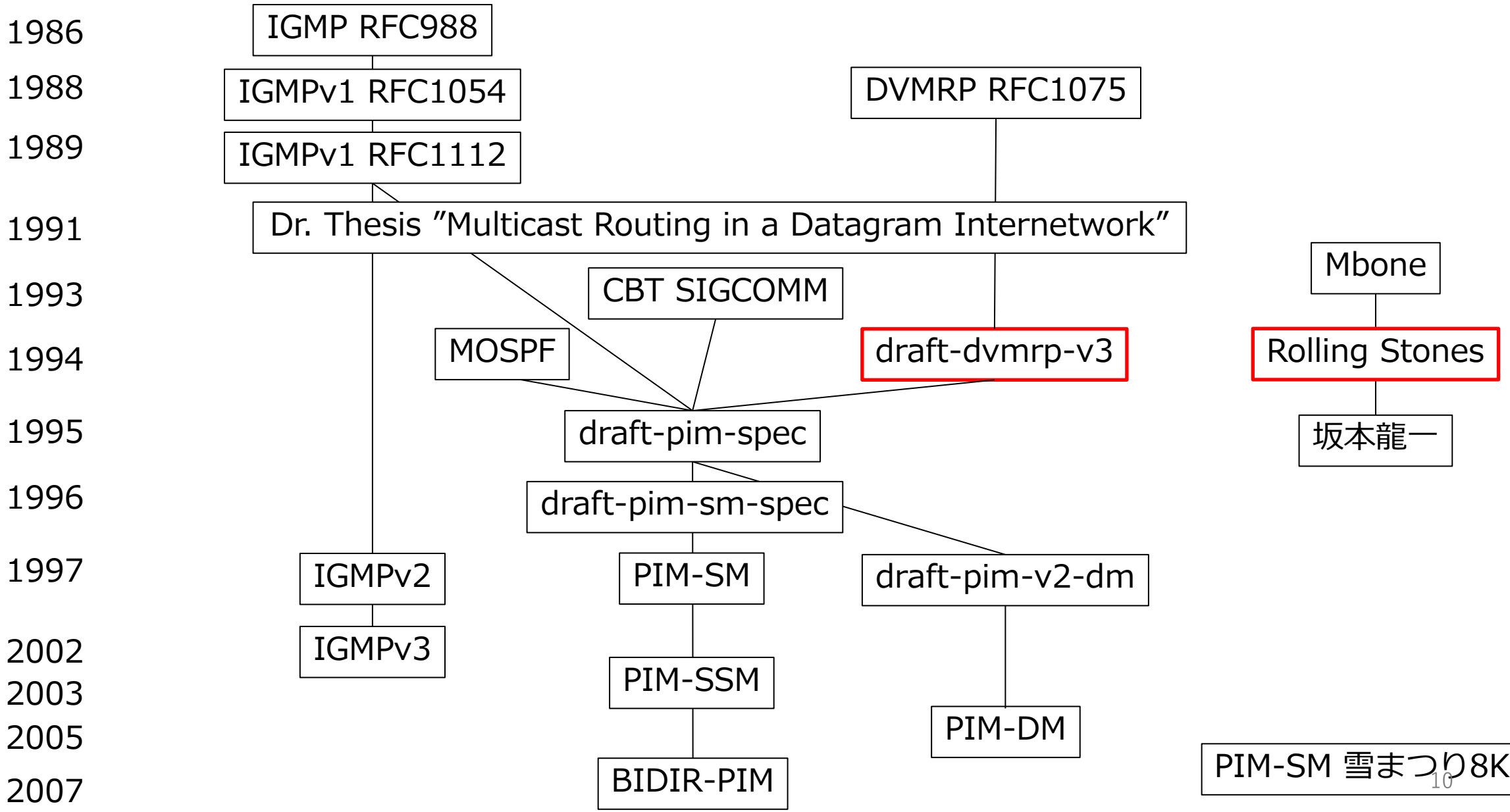
3.1 Host Groups

Under the host group model, the set of destinations of a multicast packet is called a *host group*, and it is identified by a single *group address* or *multicast address*. To accomplish a multicast, a sender simply places a group address, rather than an individual (unicast) address, in the destination address field of a packet.

As pointed out in Section 1.1, the use of group addresses allows a multicast service to be used not only for efficient multi-destination delivery, but also for *logical addressing*, that is, for reaching entities whose individual host addresses are either unknown (to the sender) or changeable—a sending host need know only a group address to reach all hosts belonging to that group.

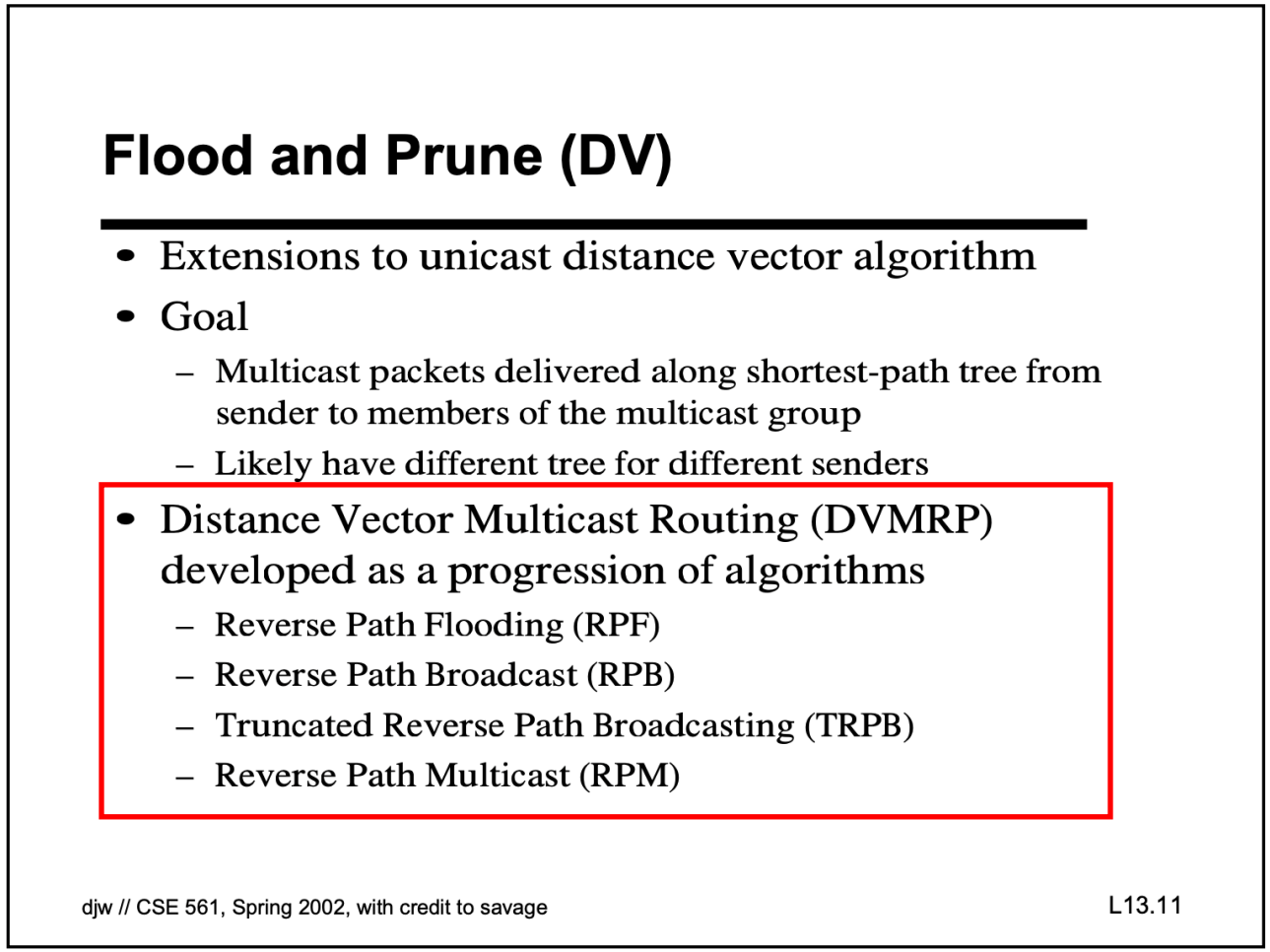
- スタンフォード大学における博士論文(スティーブ40歳)
- 同じマルチキャストアドレスに属するHostにだけパケットが配布されるHost Group Modelを発明した
- 今聞くとあたりまえすぎるが、当時はそういった概念から生み出す必要があった
- ルーティングについても触れており1991年当時のマルチキャスト技術の最前線を概観できる

IP Multicastの歴史



1994年 – Rolling StonesのライブとDVMRPの変遷

- 1992年からMboneと呼ばれる世界的なマルチキャストの実験ネットワークが稼働しはじめる
- 1994年の11月にはRolling Stoneのライブが配信される
- そこで使われていたのはDVMRPだが、既にRFC1075とは似ても似つかぬ姿になっていた。。。



Flood and Prune (DV)

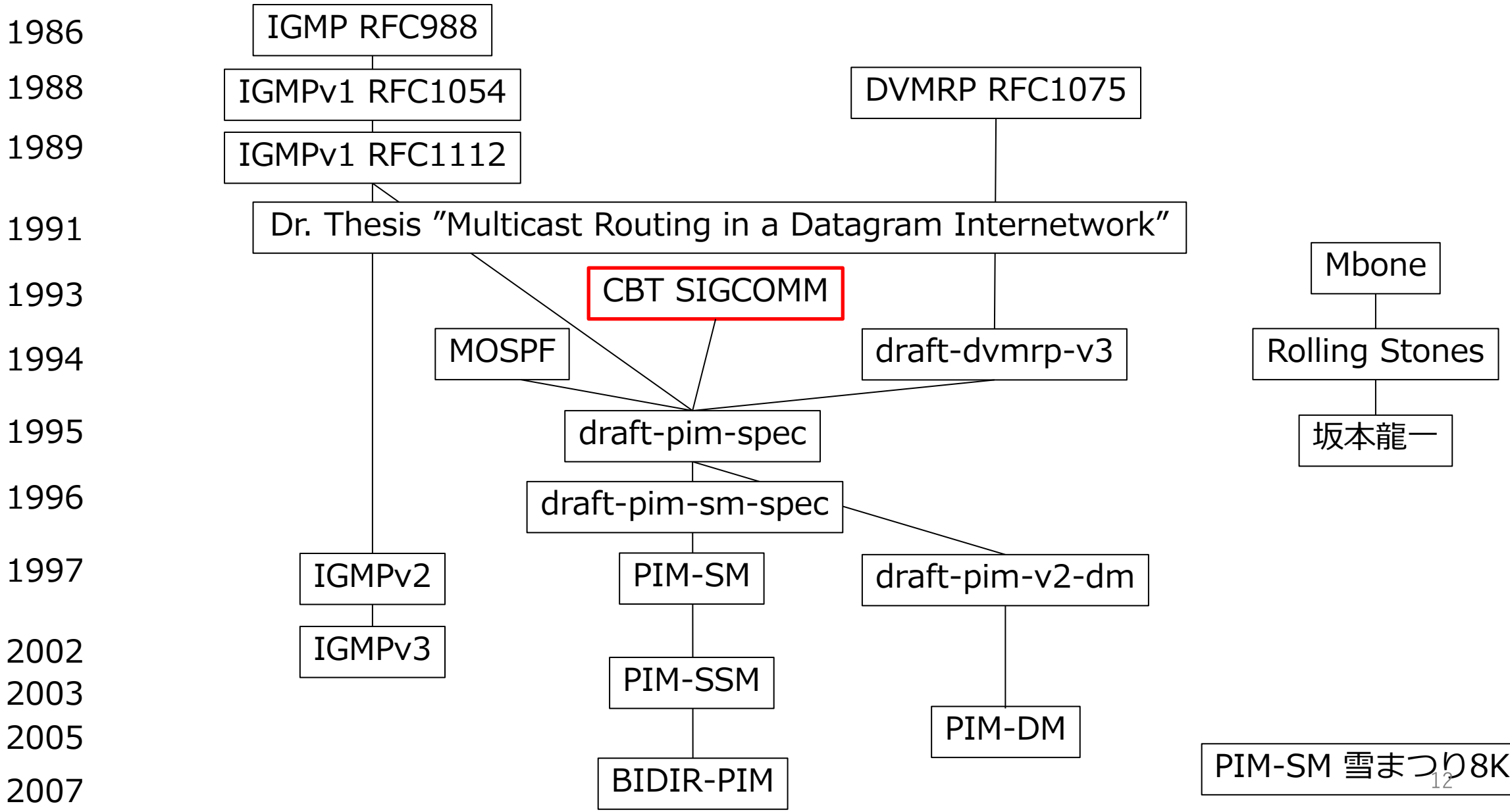
- Extensions to unicast distance vector algorithm
- Goal
  - Multicast packets delivered along shortest-path tree from sender to members of the multicast group
  - Likely have different tree for different senders
- Distance Vector Multicast Routing (DVMRP) developed as a progression of algorithms
  - Reverse Path Flooding (RPF)
  - Reverse Path Broadcast (RPB)
  - Truncated Reverse Path Broadcasting (TRPB)
  - Reverse Path Multicast (RPM)

djw // CSE 561, Spring 2002, with credit to savage

L13.11

- DVMRPのアルゴリズムは時代によって改良されており、最終的に４つのアルゴリズムが試された
- IGMP TypeもRFC1075の3ではなく、13を使うようになっていた
- こうしたことは文書になっておらずmroutedのソースを読むしか理解する手段がなかった
- さすがにそれはいかがなものかということでinternet-draftが書かれるもののPIMの広まりによってDVMRPは使われなくなり、結局RFCにならずじまいだった
- Reverse Pathの考え方はReverse Path ForwardingとしてPIMに生かされた🪦

IP Multicastの歴史



1993年 Rolling Stonesライブの前年に画期的な論文CBTが発表される

Core Based Trees (CBT)

An Architecture for Scalable Inter-Domain Multicast Routing

Tony Ballardie*(University College London)
e-mail: A.Ballardie@cs.ucl.ac.uk
Paul Francis†(Bellcore, N.J., U.S.A.)
e-mail: francis@thumper.bellcore.com
Jon Crowcroft (University College London)
e-mail: J.Crowcroft@cs.ucl.ac.uk

Abstract

One of the central problems in one-to-many wide-area communications is forming the delivery tree - the collection of nodes and links that a multicast packet traverses. Significant problems remain to be solved in the area of multicast tree formation, the problem of scaling being paramount among these.

In this paper we show how the current IP multicast architecture scales poorly (by scale poorly, we mean consume too much memory, bandwidth, or too many processing resources), and subsequently present a multicast protocol based on a new scalable architecture that is low-cost, relatively simple, and efficient. We also show how this architecture is decoupled from (though dependent on) unicast routing, and is therefore easy to install in an internet that comprises multiple heterogeneous unicast routing algorithms.

1 Introduction

Multicast group communcation is an increasingly important capability in many of today's data networks. Most LANs and more recent wide-area network technologies such as SMDS [12] and ATM [7] specify multicast as part of their service, but perhaps the most apparent and widespread growth in multicast applications is being experienced in the IP Internet. We can see evidence of this growth in the MBONE, the set of routers and networks with multicast capapility.

In order to cater to a very large number of internetwork-wide multicast applications, examples of which include audio and video conferencing [15], replicated database updating and querying, software update distribution, stock market information services, and more recently, resource discovery [11], it is important that the multicast routing protocol used be first and foremost scalable with respect to a network of very large size, and low-cost in terms of computational overhead and storage requirements - properties lacking in current IP multicasting techniques. The protocol should also be designed to operate "invisibly" across domain boundaries, i.e. independent of the underlying unicast routing algorithm, so that it can evolve independently.

This paper describes a new multicast routing architecture which is applicable to any datagram network whose switches have multicast forwarding capability. We will present a multicast routing protocol (CBT) for IP networks based on this new architecture that not only satisfies the above criteria, but is also relatively simple in design.

In the following section we discuss the existing multicast architecture. Section 3 describes the current IP multicast environment and goes on to briefly describe two IP multicast routing protocols. Section 4 presents a comprehensive critique of the existing architecture showing how it is inherently non-scalable and bound to particular underlying unicast routing algorithms. This leads us to the new architecture in section 5 followed by a description of a protocol built on this new architecture in section 6. Sections 7 and 8 offer some thoughts on future work and an overall summary, respectively.

2 Existing Multicast Architecture

The existing multicast architecture is not restricted to IP networks, but is being accepted as the solution to multicasting in many different kinds of networks and environments.

For each multicast group, the current architecture builds a shortest-path source-based delivery tree be-

*Principal author
†Previously published under the name Paul Tsuchiya

SIGCOMM'93 - Ithaca, N.Y., USA /9/93
© 1993 ACM 0-89791-619-0/93/0009/0085...$1.50

- UCL ユニバーシティ・カレッジ・ロンドンのチーム
- NATで有名なPaul Francisが著者の一人
- PIMの共有ツリーの直接的なベースになった
- IGMPのGroup Membershipを活用しつつCBT専用のプロトコルを策定するというのはPIMv2と同じ

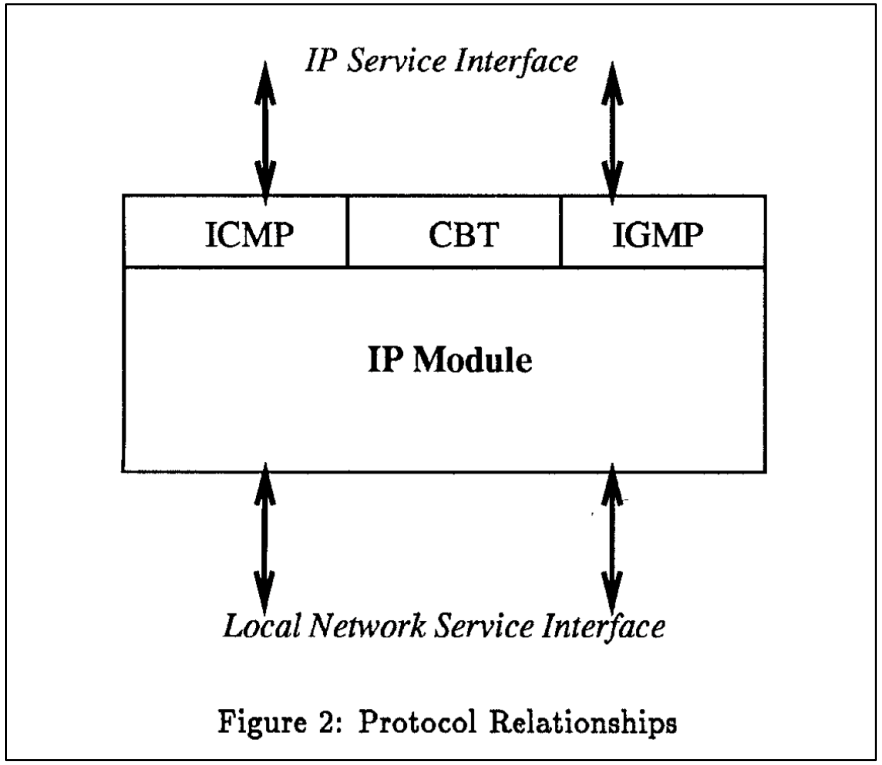
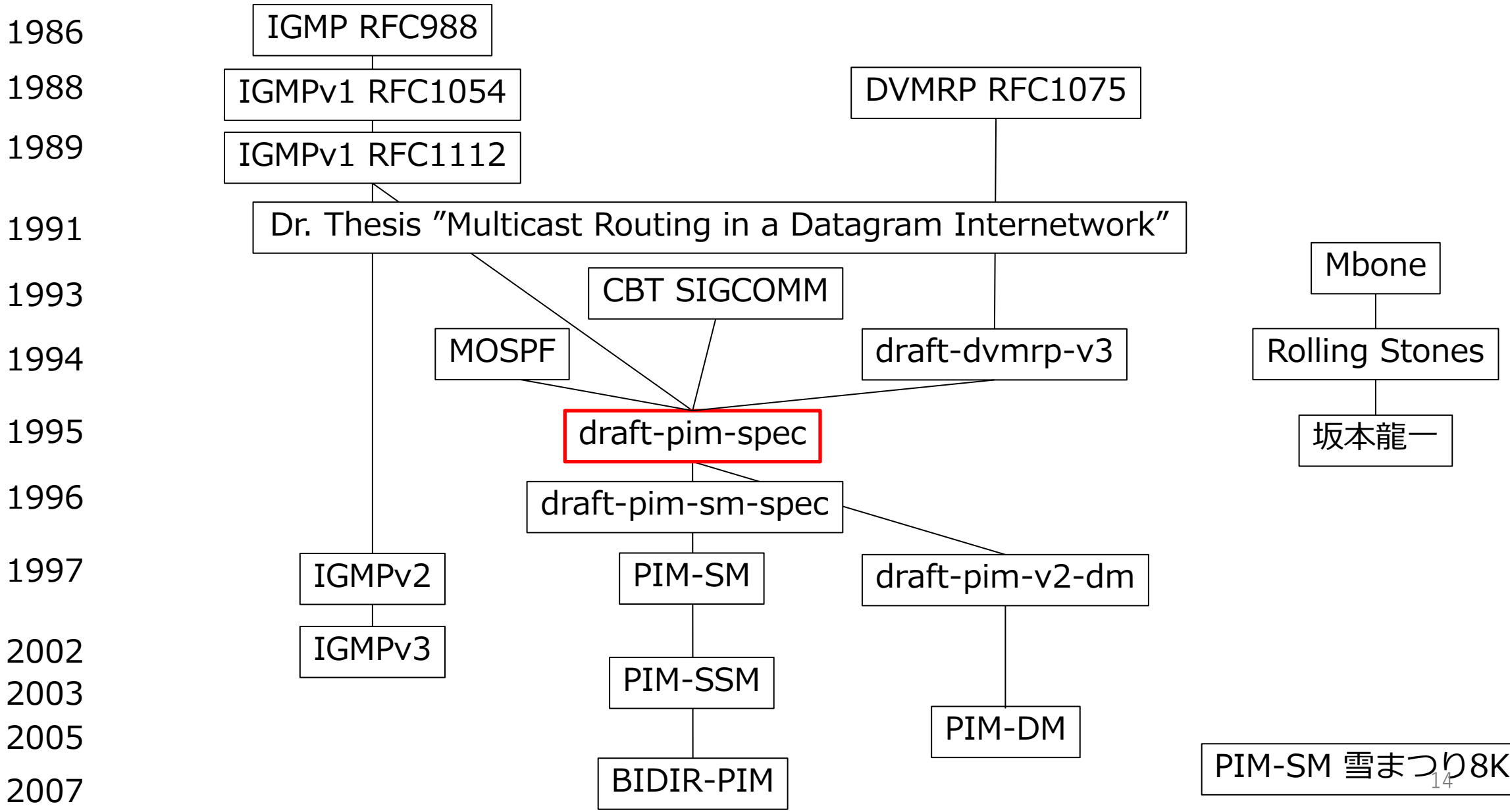


Figure 2: Protocol Relationships

IP Multicastの歴史



1995年 ついにPIMが爆誕！著者は当時のスーパースター達

Protocol Independent Multicast (PIM): Protocol Specification

マルチキャスト産みの親

Stephen Deering

Xerox PARC
3333 Coyoty Hill Road
Palo Alto, CA 94304
deering@parc.xerox.com

Jon Postelの愛弟子

Deborah Estrin

Computer Science Department/ISI
University of Southern California
Los Angeles, CA 90089
estrin@usc.edu

産業界代表Cisco

Dino Farinacci

Cisco Systems Inc.
170 West Tasman Drive,
San Jose, CA 95134
dino@cisco.com

Van Jacobson

Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, CA 94720
van@ee.lbl.gov

TCP輻輳制御
TCP Header圧縮
TCP Slow Start
tcpdump/tracerouteの作者
業績多すぎ

Jon Postelのチーム

Ching-gung Liu

Computer Science Department
University of Southern California
Los Angeles, CA 90089
charley@catarina.usc.edu

Jon Postelのチーム

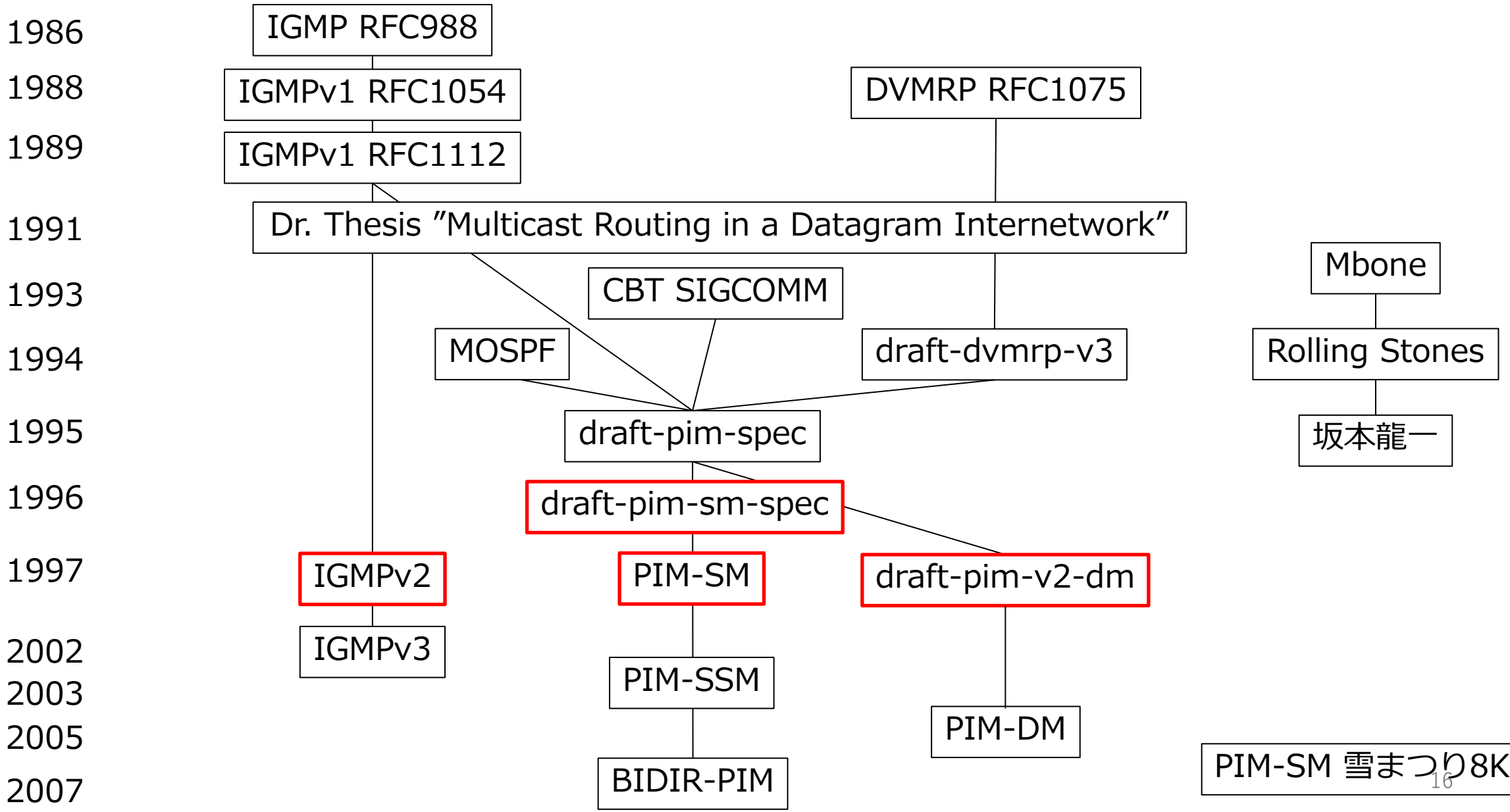
Liming Wei

Computer Science Department
University of Southern California
Los Angeles, CA 90089
lwei@catarina.usc.edu

draft-ietf-idmr-pim-spec-01.ps

January 11, 1995

IP Multicastの歴史



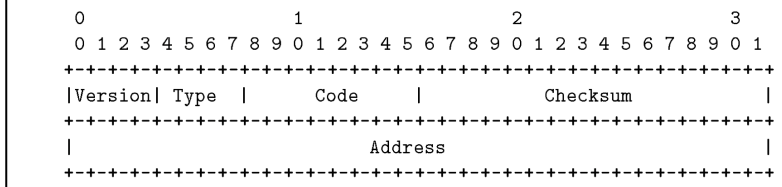
1995 -> 1996 PIMv1からPIMv2へそしてPIM-SMとPIM-DMの文書を分離

draft-ietf-idmr-pim-spec-01
Jan 11 1995

4 Packet Formats

RFC-1112, see [3], specifies two types of IGMP packets for hosts and routers to convey multicast group membership and reachability information. An IGMP Host-Query packet is transmitted periodically by routers to ask hosts to report which multicast groups they are members of. An IGMP Host- Report packet is transmitted by hosts in response to received queries advertising group membership.

This section introduces new types of IGMP packets that are used by PIM routers. The fixed header packet format is:



Version This memo specifies version 1 of IGMP.

Type There are nine types of IGMP messages:

1 = Host Membership Query
2 = Host Membership Report
3 = Router DVMRP Messages
4 = Router PIM Messages
5 = Cisco Trace Messages
6 = New Host Membership Report
7 = Host Membership Leave
14= Mtrace Response
15= Mtrace Request

- PIM-SMもPIM-DMも一つの文書で定義
- PIMパケットはまだIGMPv1の拡張
- PIMのVersionは1
- (*,G)から(S,G)への切り替えやRPFによるLoop検知の仕組みは今のPIM-SMと同じ

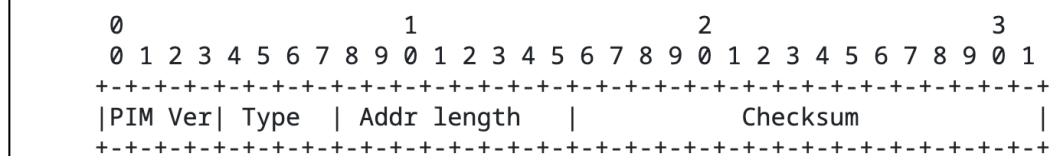
draft-ietf-idmr-pim-sm-spec-02
Jun 6 1996

4 Packet Formats

This section describes the details of the packet formats for PIM control messages.

All PIM control messages have protocol number 103.

Basically, PIM messages are either unicast (e.g. Registers and Register-Stop), or multicast hop-by-hop to `ALL-PIM-ROUTERS' group `224.0.0.13' (e.g. Join/Prune, Asserts, etc.).



PIM Ver
PIM Version number is 2.

- PIM-SMのみの文書に変更
- PIM-DMは翌年internet-draftが書かれるまで放置プレイ
- PIM専用のIP Protocol番号がアサイン
- IGMPはIGMPv2へ
- PIMのVersionは2へ

PIM-SMのアイデア

1. RPランデブーポイント起点の「共有ツリー」(*,G)をまず作成し
2. そのあと「送信元ツリー」 (S,G)を連結
3. さらに(S,G,rpt)の最適化を行う

(*, G)および(S,G)のままでも問題なくMulticast Routingが行えるが、その後(S, G,rpt)への切り替えにより最適化が行えるというのがPIM-SMの一番のアイデア

Reverse Path ForwardingによるHop-by-Hopのマルチキャストパスの作成とループ防止のメカニズム

セッションレスにも関わらず極力アンダーレイの変更に即座に対応できるようにさまざまな工夫がされている、が、あまりに複雑なためこれまでPIM-SMのスペックは数多くの改訂がされている

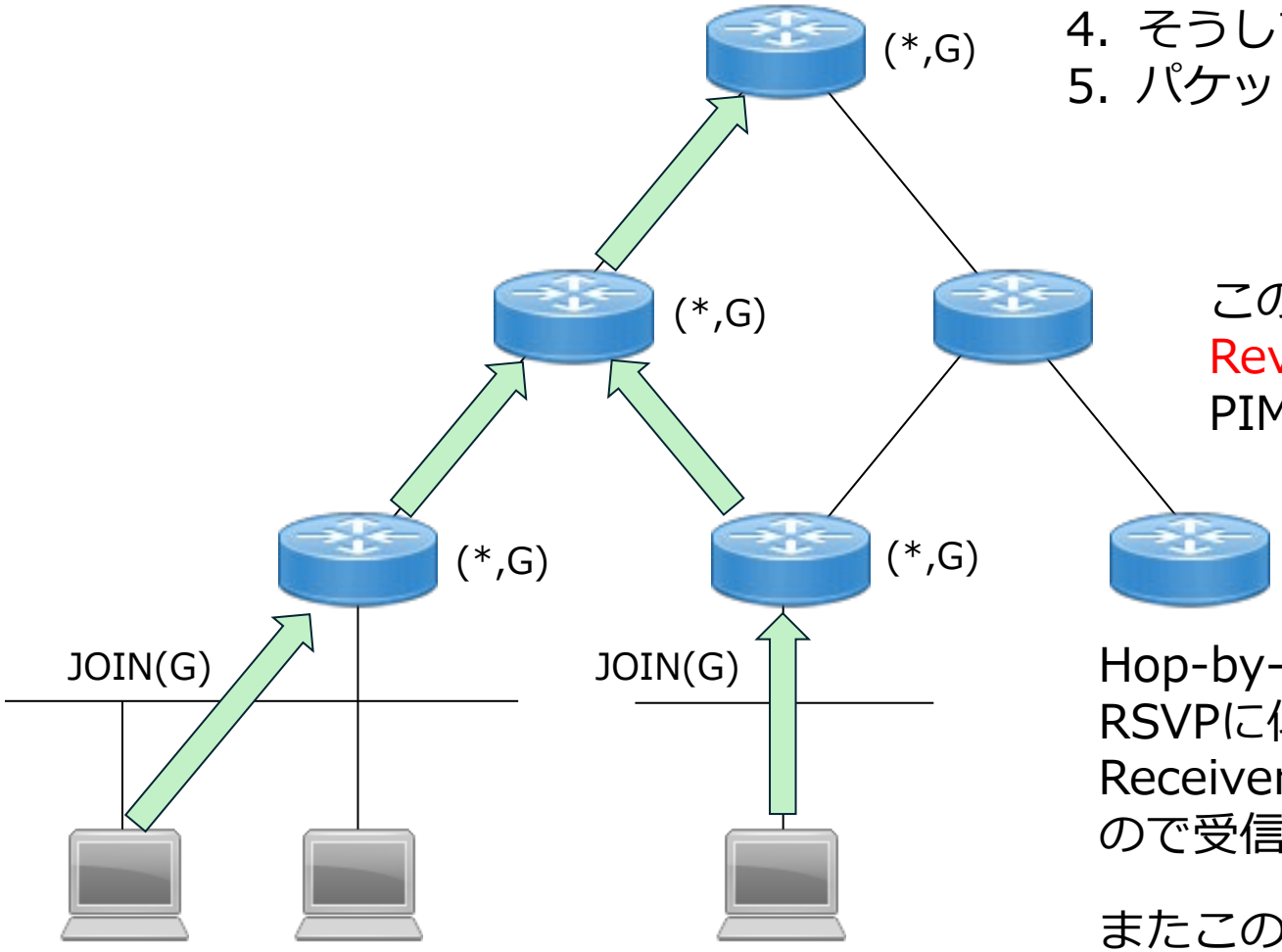
しかしエントリーの更新は基本的にタイマーによるExpirationに依存しているのでDebugがしにくい

2024年に考える - なぜPIM-SMだけが生き残ったのか？

- CBTからSparse Modeを、DVMRPからDense Modeを受け継ぎ悪魔合体できた
- まだ人類がMulticastに夢を持っていた時期だったのでIGMPv2のデプロイがうまく行った
- 中途半端にIP電話の保留音とかのユースケースができてしまい引っ込みがつかなくなった
- なんだかんだいってAny Source Multicastが便利だった
- 仕様は当時のスーパースター勢揃いで、この連中なら凄いものができるのではないかといった期待があった

PIM-SMの仕組み (*,G)

1. まずHostがMulticast GroupにJoinしないと始まらない
2. Joinを受け取ったルータはRPに向けて(*,G)を送る
3. 経由するルータもそれぞれ(*,G)を作成する
4. そうして共有ツリー(*,G)が作成される
5. パケットはこの経路を逆向きにたどって運ばれる



この逆向きにたどってパケットを転送する方式を
Reverse Path Forwardingと呼んでおり、
PIM-SMの特徴的なアルゴリズムになっている

Hop-by-HopのExplicit Pathをたどるという意味だと
RSVPに似ているように思えるが、PIM-SMの場合
Receiverへの共有ツリーを作りたいという動機がある
ので受信者からのReverse Pathを用いてTreeを構成

またこの際のOutgoing IF/Incoming IFを使用して
Loop Preventionを行う

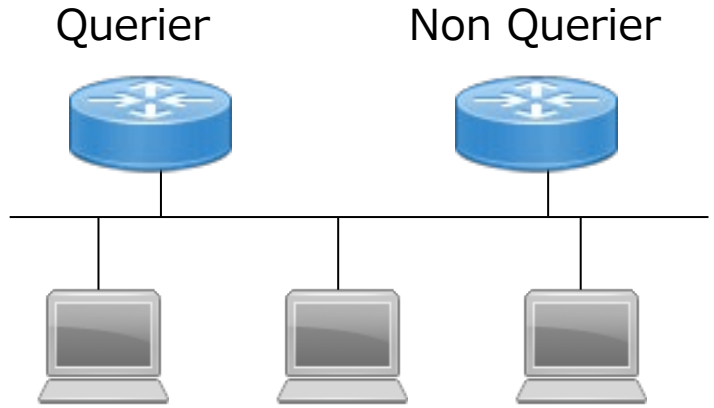
さて、ここで問題です

一般的なPIM-SMの設定の場合
(*,G)の時点でHost, Router, RP合わせて
何種類のタイマーが存在するでしょう？

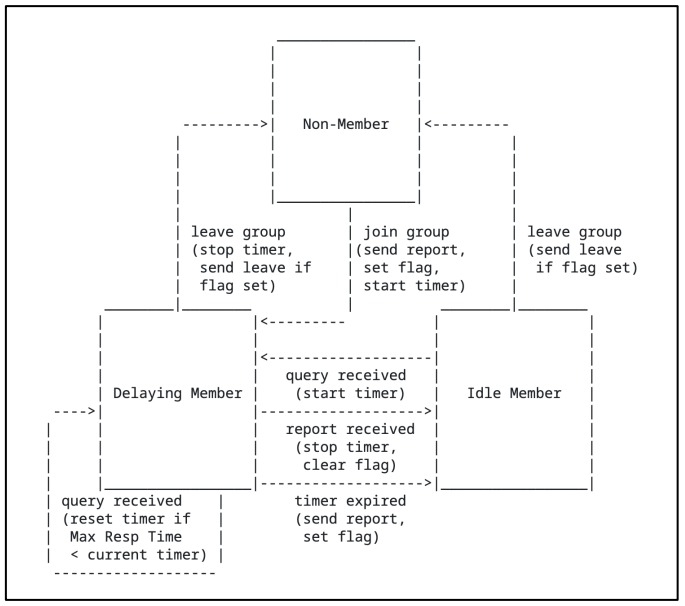
答え：１５種類

IGMPv2関連のタイマー (Leaveを除いて5種類)

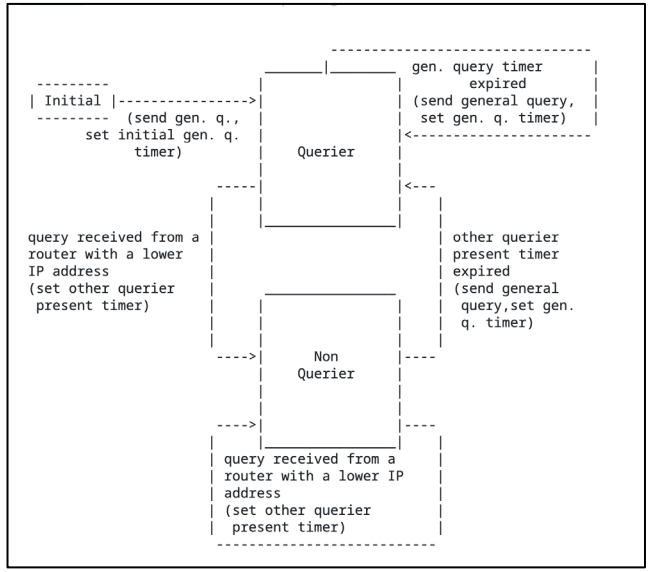
- Host
  - unsolicited report interval (IGMPv2:10sec, IGMPv3:1sec)
  - query response timer
- Router (Querier)
  - query timer
  - group membership retransmit timer
- Router (Non Querier)
  - other querier present timer
  - group membership retransmit timer(Querierと同じ)



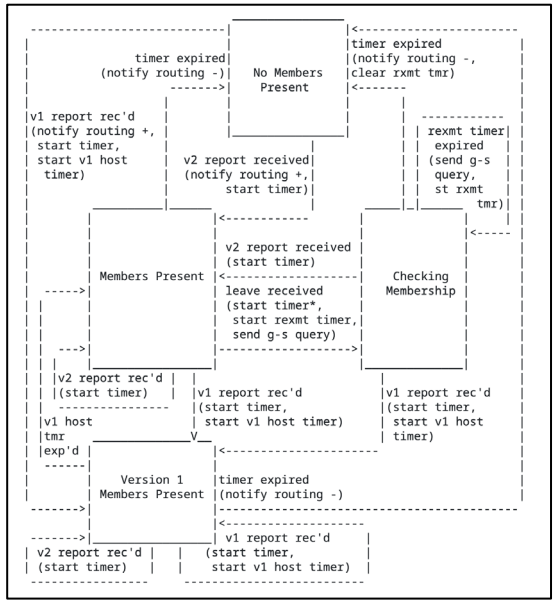
RFC2236 Host



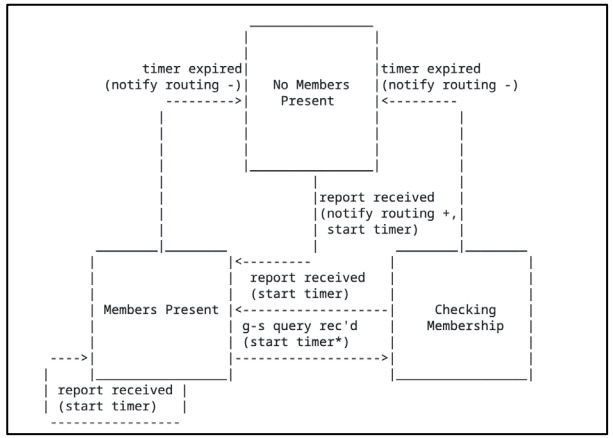
Querier <-> Non Querier



Querier



Non Querier



PIM関連の(*,G)タイマー 6種類

RFC7761

```
Global Timers

Per interface (I):

     Hello Timer: HT(I)

     Per neighbor (N):

          Neighbor Liveness Timer: NLT(N,I)

     Per Group (G):

          (*,G) Join Expiry Timer: ET(*,G,I)

          (*,G) Prune-Pending Timer: PPT(*,G,I)

          (*,G) Assert Timer: AT(*,G,I)

          Per Source (S):

               (S,G) Join Expiry Timer: ET(S,G,I)

               (S,G) Prune-Pending Timer: PPT(S,G,I)

               (S,G) Assert Timer: AT(S,G,I)

               (S,G,rpt) Prune Expiry Timer: ET(S,G,rpt,I)

               (S,G,rpt) Prune-Pending Timer: PPT(S,G,rpt,I)

Per Group (G):

     (*,G) Upstream Join Timer: JT(*,G)

     Per Source (S):

          (S,G) Upstream Join Timer: JT(S,G)

          (S,G) Keepalive Timer: KAT(S,G)

          (S,G,rpt) Upstream Override Timer: OT(S,G,rpt)

At the DRs or relevant Assert Winners only:

     Per Source,Group pair (S,G):

          Register-Stop Timer: RST(S,G)
```

- DR/Liveness関連 2種類
- Downstream Per Interface (*,G) 3種類
- Upstream (*,G) 1種類

Figure 2: Downstream Per-Interface (*,G) State Machine

Prev State	Event: Receive Join(*,G)	Event: Receive Prune(*,G)	Event: Prune-Pending Timer Expires	Event: Expiry Timer Expires
NoInfo (NI)	-> J state start Expiry Timer	-> NI state	-	-
Join (J)	-> J state restart Expiry Timer	-> PP state start Prune-Pending Timer	-	-> NI state
Prune-Pending (PP)	-> J state restart Expiry Timer	-> PP state	-> NI state Send Prune-Echo(*,G)	-> NI state

Figure 5: Upstream (*,G) State Machine

Prev State	Event: JoinDesired(*,G) ->True	Event: JoinDesired(*,G) ->False
NotJoined (NJ)	-> J state Send Join(*,G); set Join Timer to t_periodic	-
Joined (J)	-	-> NJ state Send Prune(*,G); cancel Join Timer

BSR関連のタイマー 4種類

Bootstrap Routerは自動的にRPを見つける仕組み

RFC5059

5. Timers and Timer Values

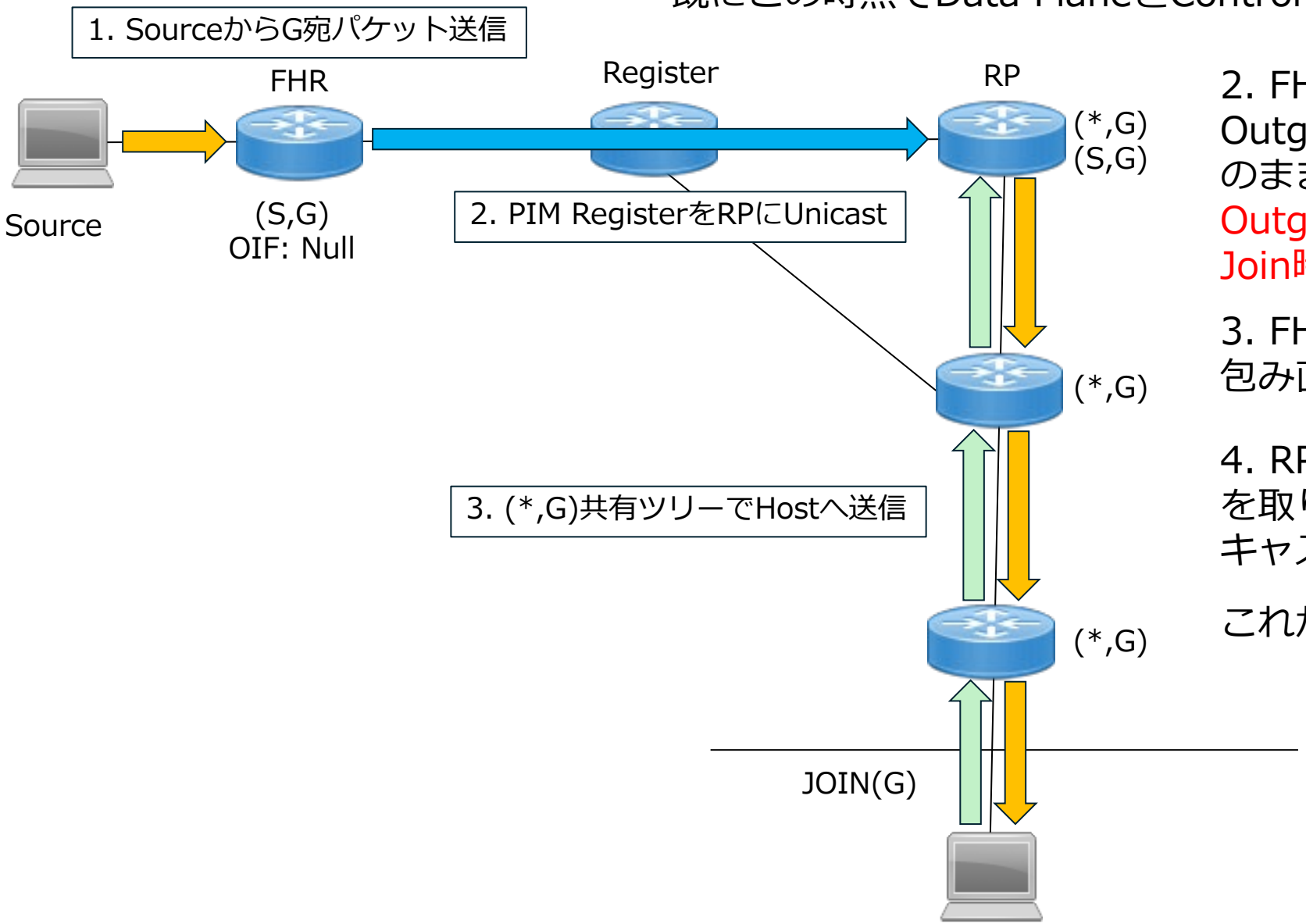
Timer Name: Bootstrap Timer (BST(Z))

Value Name	Value	Explanation
BS_Period	Default: 60 seconds	Periodic interval with which BSMs are normally originated
BS_Timeout	Default: 130 seconds	Interval after which a BSR is timed out if no BSM is received from that BSR
BS_Min_Interval	Default: 10 seconds	Minimum interval with which BSMs may be originated
BS_Rand_Override	see below	Randomized interval used to reduce control message overhead during BSR election

- IGMP, PIM, BSRでそれぞれ独立のStateがある
- すべてコネクションレスでタイマーがある
- なにかおかしい時にはそれぞれ確認
- イベントのデバッグログが重要
- 特にタイムアウト

PIM-SMの仕組み (S,G)

1. Sourceからのマルチキャストデータパケットを First Hop Routerが補足
– 既にこの時点でData PlaneとControl Planeの分離の違反



2. FHRは(S,G)エントリーを作るものの Outgoing Interfaceがわからないので空のまま
Outgoing InterfaceはReverse Pathの Join時に分かる

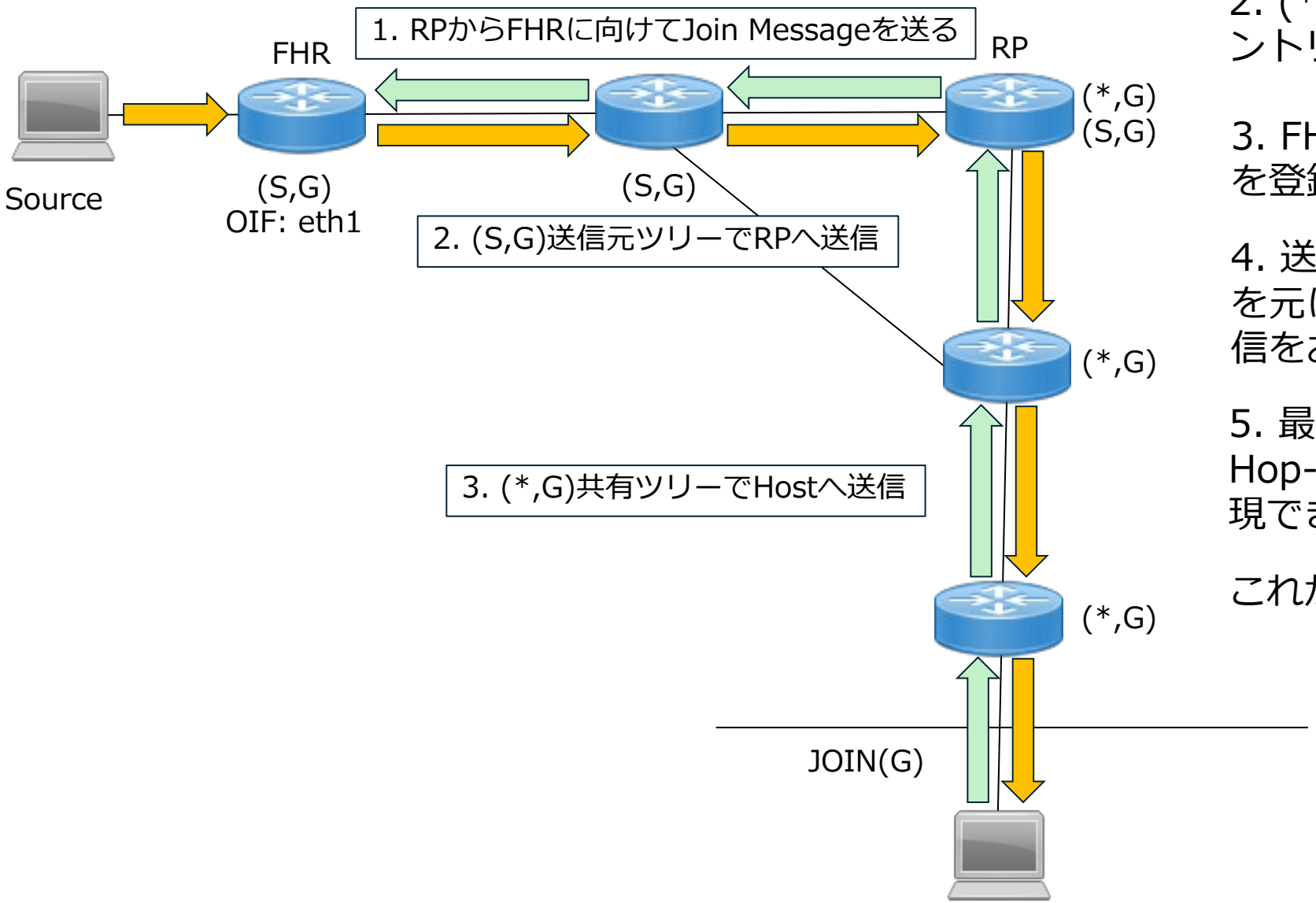
3. FHRはG宛パケットをPIM Registerに包み直してRPへユニキャストで送る

4. RPでPIM Registerから元のパケットを取り出し(*,G)をHop-by-Hopのマルチキャストで Hostに送信

これが第一形態

PIM-SMの仕組み (S,G)

1. Register Messageを受け取ったRPはFHRにむけてJoin Messageを送る



2. (*,G)と同様経由するルータで(S,G)エントリを作成する

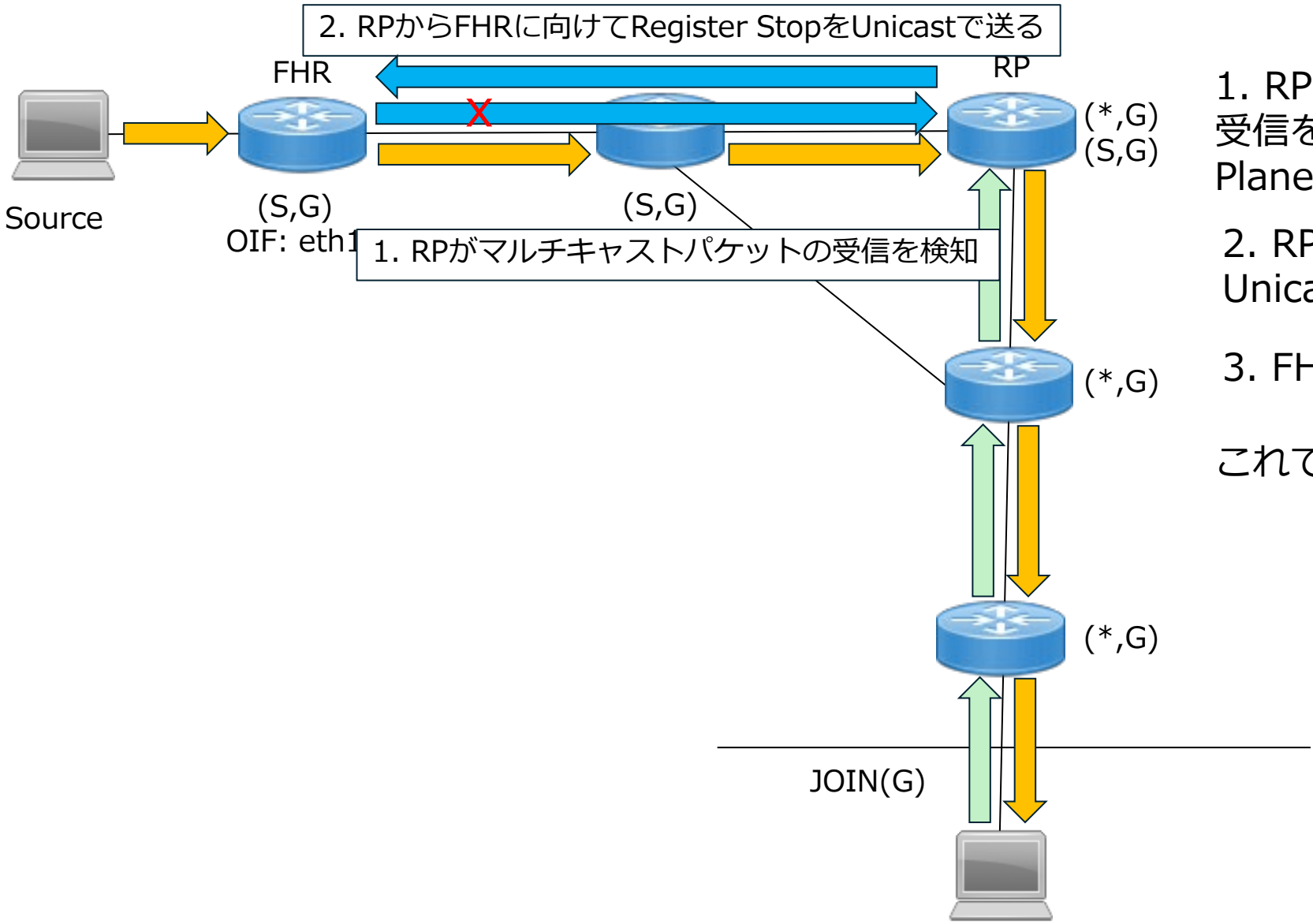
3. FHRはReverse PathのJoinを元にOIFを登録する

4. 送信元ツリー(S,G)と共有ツリー(*,G)を元にHop-by-Hopのマルチキャスト送信をおこなう

5. 最適化がされていないが、この時点でHop-by-Hopのマルチキャスト配信が実現できる

これが第二形態のフォワーディング

PIM-SMの仕組み (S,G)



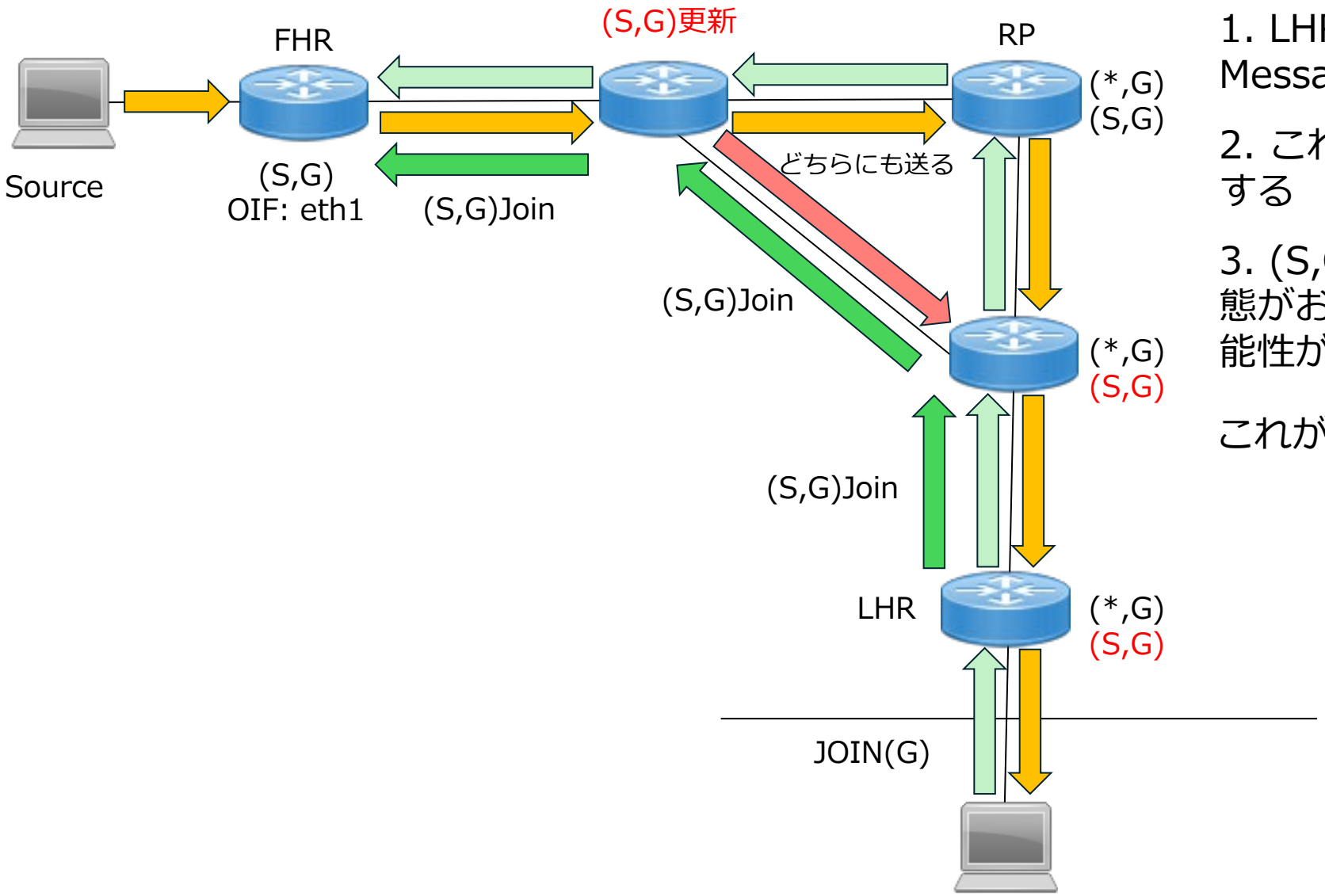
1. RPがマルチキャストデータパケットの受信を検知– 再びData PlaneとControl Planeの分離の違反

2. RPがFHRに向けてRegister StopをUnicastで送る

3. FHRがRegister Messageの送信を停止

これで第二形態が完了

PIM-SMの仕組み (S,G)　最適化はLHR(Last Hop Router)を起点に行われる



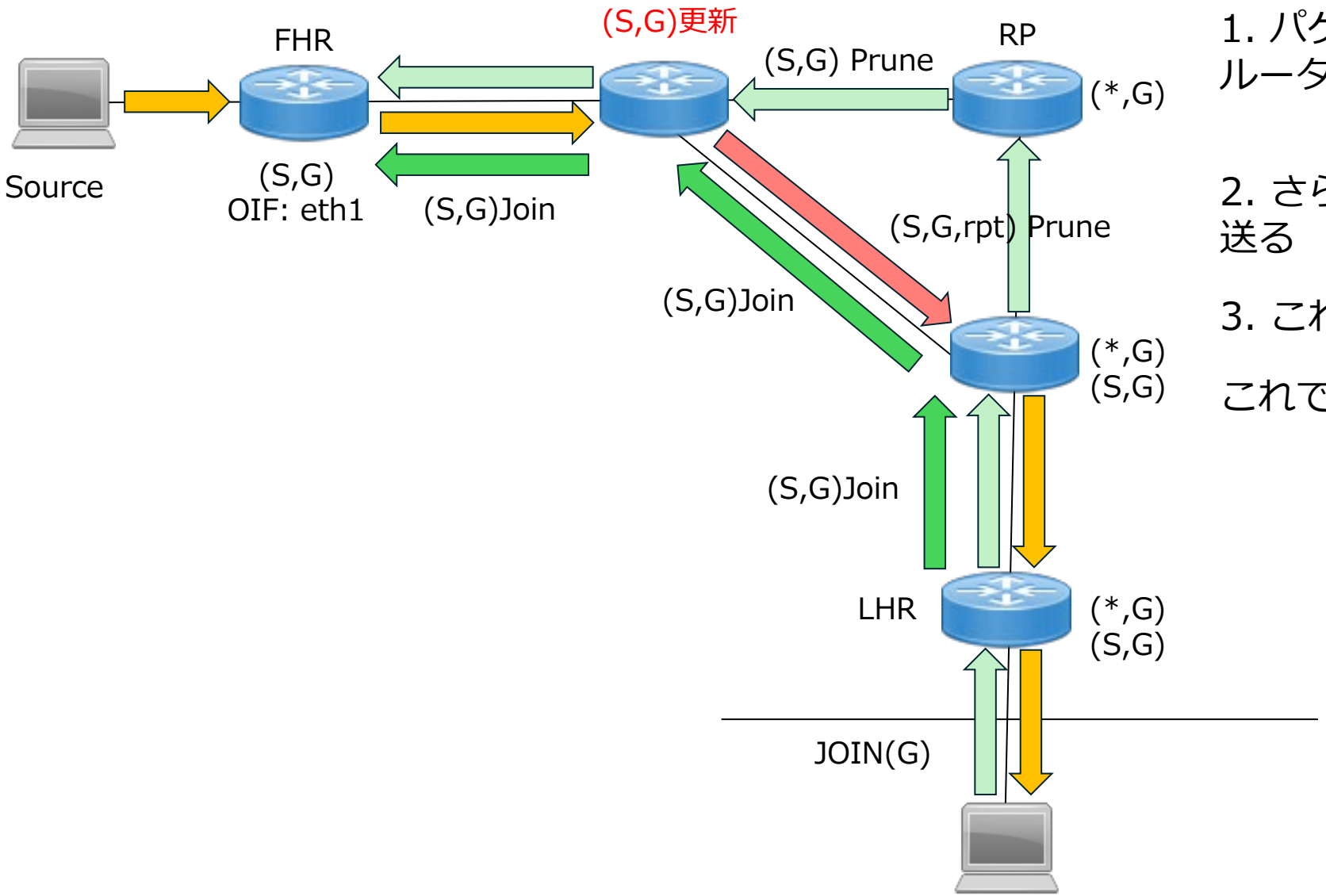
1. LHRからSourceに向けて(S,G)Join Messageを送る

2. これによりShortest Path Treeを構成する

3. (S,G)と(S,G,rpt)の両方が存在する状態がおこりうるためパケットが重複する可能性がある

これが第三形態の準備

PIM-SMの仕組み (S,G)　重複パスの解消は境界ルータ起点で行われる



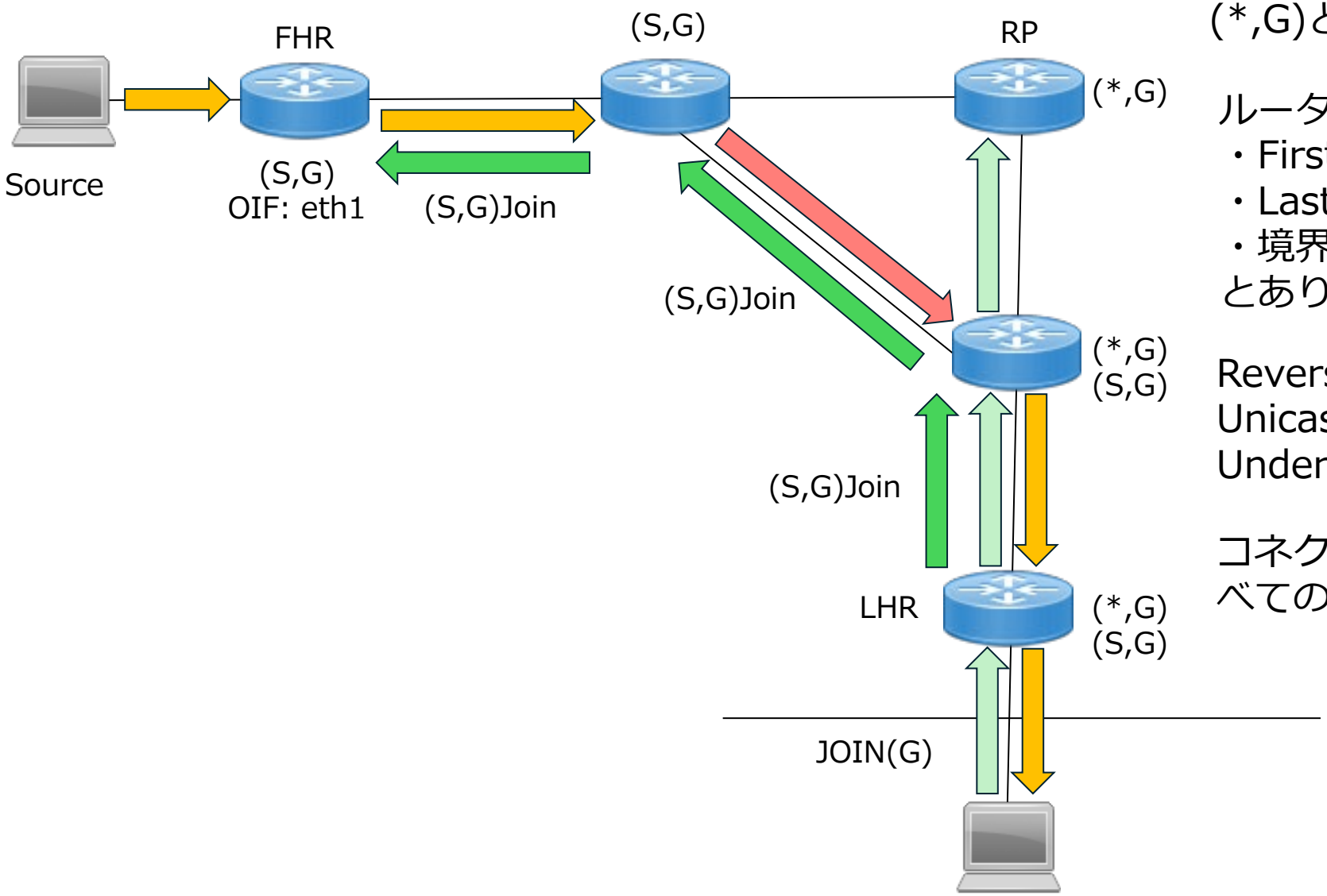
1. パケット重複を解消するために境界ルータから(S,G,rpt)メッセージを送る

2. さらにPRから(S,G)Pruneメッセージを送る

3. これにより最適化が完了する

これで第三形態の完了

PIM-SMの仕組み (S,G)　最終的な姿



なぜPIM-SMは複雑なのか？

(*,G)と(S,G)は関連はあるものの独立である

ルータの役割がS毎に
・First Hop Router
・Last Hop Router
・境界ルータ
とありそれぞれ独立のState Machineをもつ

Reverse Path ForwardingがUnderlayの
Unicastルーティングに依存しているため
Underlayの変更に大きな影響をうける

コネクションレスなプロトコルのためす
べてのエントリにタイマーが存在する

PIM-SMの主なタイマー おかしくなる時があるのはだいたい210秒後。。。

Timer	Sec	Description
Hello Interval	30	This timer determines the interval between sending PIM Hello messages on an interface.
Join/Prune Interval	60	This timer determines the interval between sending periodic Join/Prune messages.
Join/Prune Hold Time	210	This timer represents the period for which a Join/Prune state will be maintained by the downstream router before the state expires.
Register Stop Timer	60	This timer is used by the Rendezvous Point (RP) to maintain the state of actively sending Register-Stop messages to the source's Designated Router (DR).
Data Timeout	210	This timer is used by the last-hop router to maintain the state of actively forwarding data to receivers before pruning the forwarding state.
Route Hold Time	210	This timer is used by the routers to keep the learned multicast routes before pruning them.
Assert Hold Time	180	This timer determines the period for which the assert state will be maintained before it times out.
Query Interval	30	This timer determines the interval between sending periodic PIM Router Query messages.
Query Response Interval	5	This timer represents the maximum time allowed before sending a response to a received Query message.