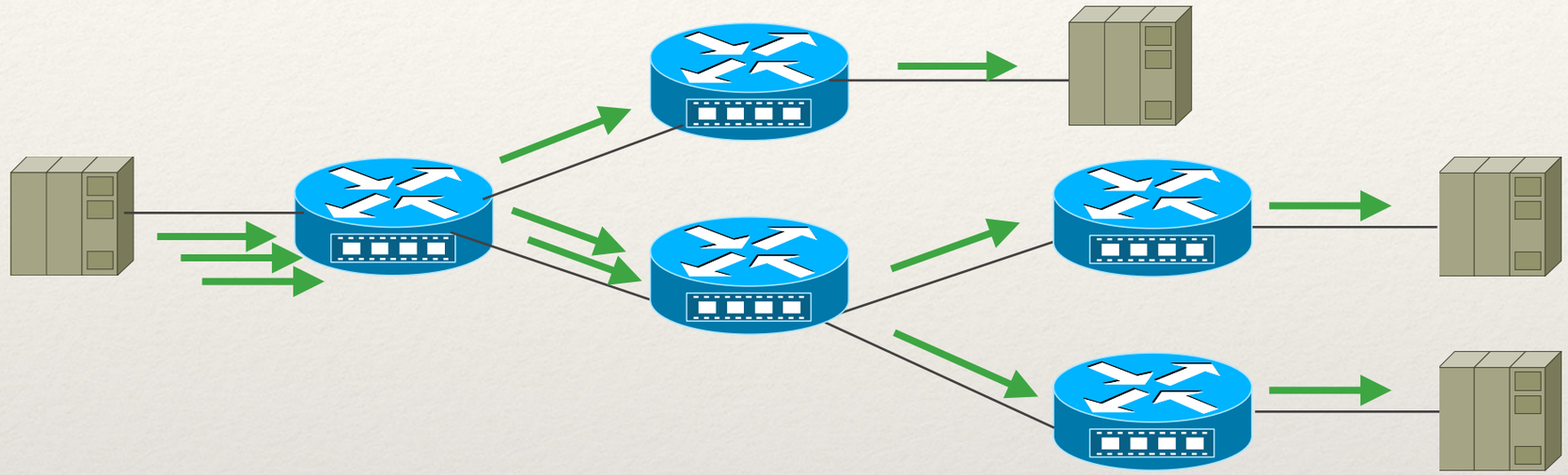
Введение в Multicast

Андрей Рогов.
«Живые встречи» 2014

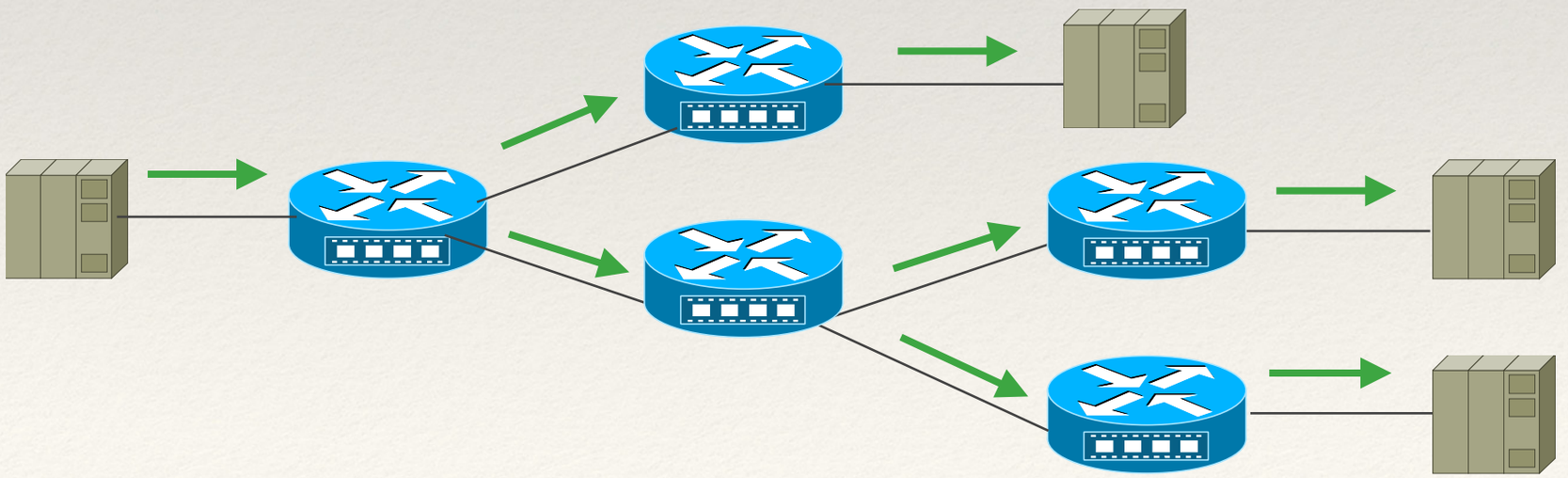
О чем будем говорить ?

- Зачем нужен Multicast ?
- История и основы
- PIM
- Rendezvous Points
- L2 Multicast
- Interdomain IP Multicast

Unicast vs Multicast: масштабирование



Unicast



Multicast

Краткая история

Steven Deering, 1985, Стенфорд.

Overlay Broadcast Domain.

RFC 966 - 1985 год

«Multi-destination delivery is useful to several applications, including:

- distributed, replicated databases [6,9].

- conferencing [11].

- distributed parallel computation, including distributed gaming [2].»

Предлагалось решение для приложений «many-to-many»

Речь не шла о приложениях «one-to-many», например, об IPTV

Краткая история

Для Overlay Broadcast Domain требовалось:

- построение и поддержка дерева участников
- механизмы поиска источников
- маршрут до источника
- туннели как механизм построения наложенных сетей

Distance Vector Multicast Routing Protocol

RFC 1075 - 1988

Краткая история

PIM - Protocol Independent Multicast

«Independent» - не зависит от протокола маршрутизации.

Использует таблицу маршрутизации для определения маршрута до источника

Router-to-Router. Строит и поддерживает в актуальном состоянии distribution trees.

Поиск источников возможен двумя путями:

1. Flood-and-prune. PIM-DM (Dense mode) RFC3973
2. Явное подключение источника к RP. PIM-SM (Sparse mode) RFC4601

Краткая история

В настоящее время основная область применения - приложения типа «one-to-many»

IPTV, Video over IP

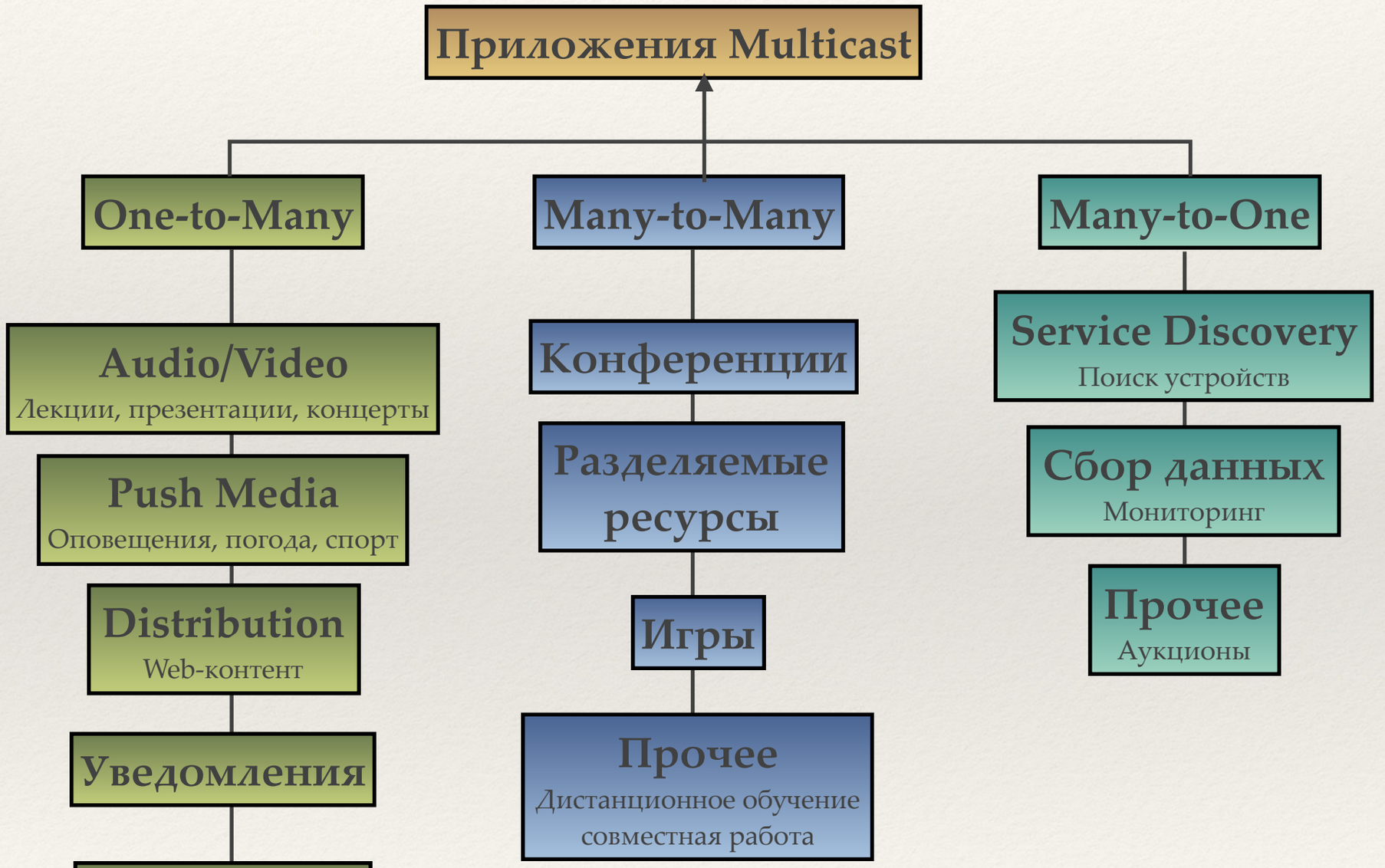
SSM - Source Specific Multicast

RFC3569, RFC4608 - 2003

- построение и поддержка дерева участников
- ~~механизмы поиска источников~~
- ~~маршрут до источника~~
- ~~туннели как механизм построения наложенных сетей~~

Простое и более предпочтительное решение для современных приложений

Типы Multicast-приложений

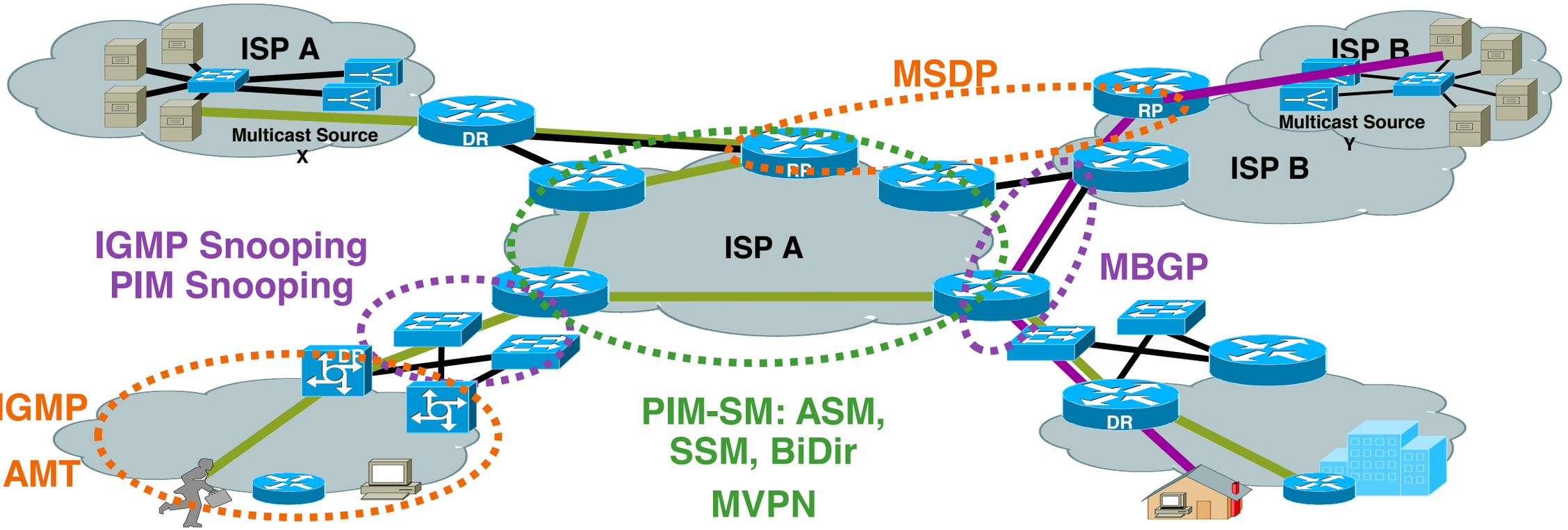


Подробности см. RFC3170

Особенности Multicast

- **UDP based**
- **Best Effort:** потери ожидаемы; приложения должны разрабатываться с учетом этого факта
- **Congestion avoidance:** отсутствует механизм предотвращения перегрузок; приложения должны это учитывать.
- **Дублирование данных:** возможны повторы данных
- **Out of order delivery**

Компоненты



Host-to-Router

IGMP, MLD, AMT

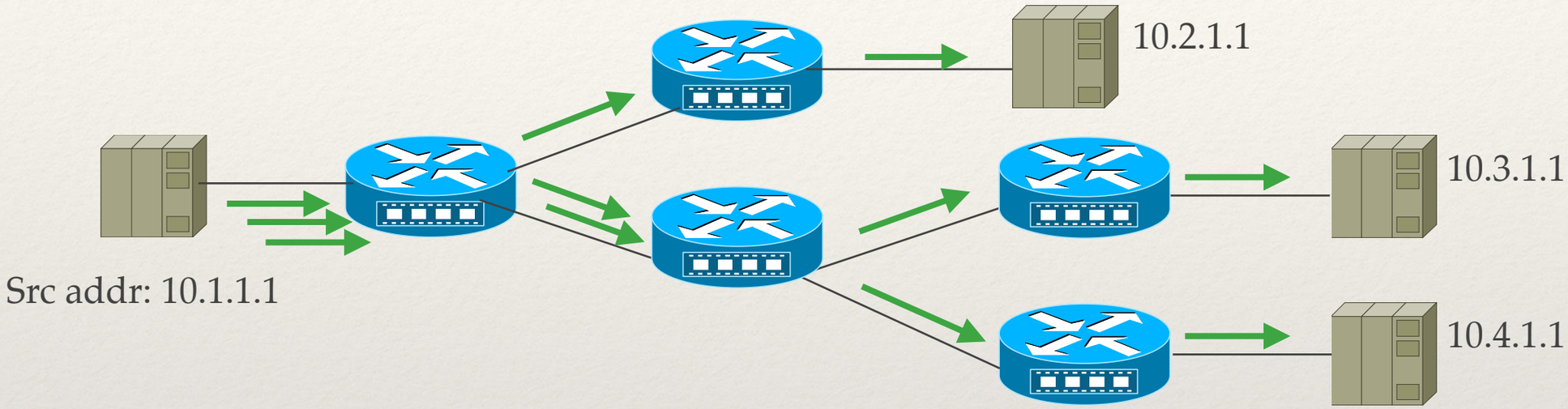
Campus Multicast

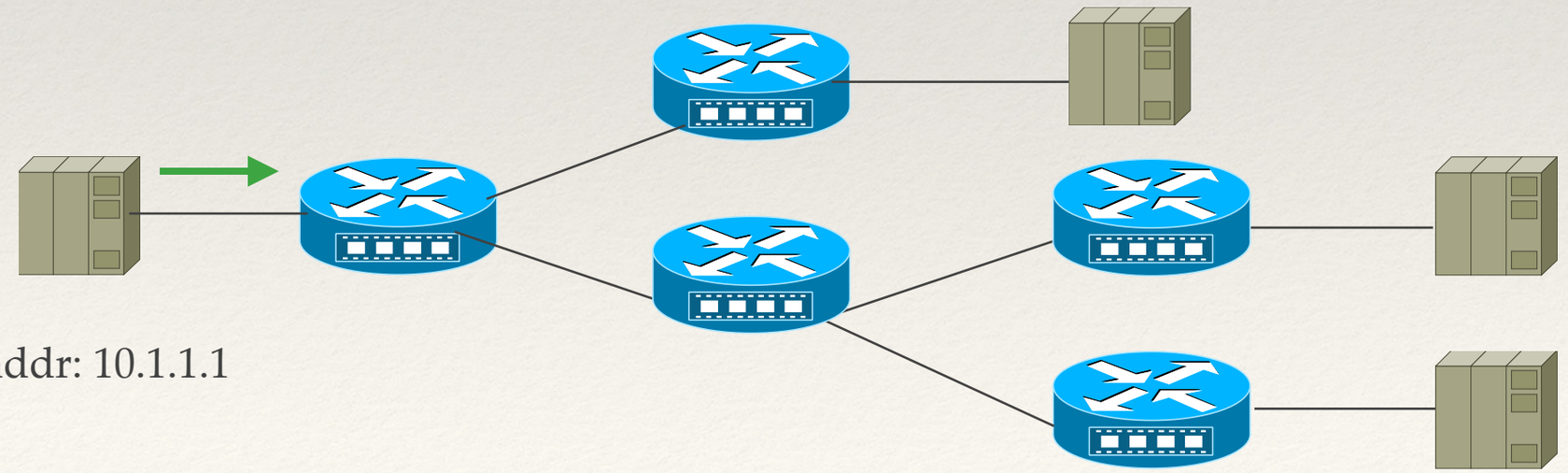
- Switches:
  - IGMP snooping
  - PIM snooping
- Routers:
  - PIM-SM, BiDir
  - MBGP

Interdomain Multicast

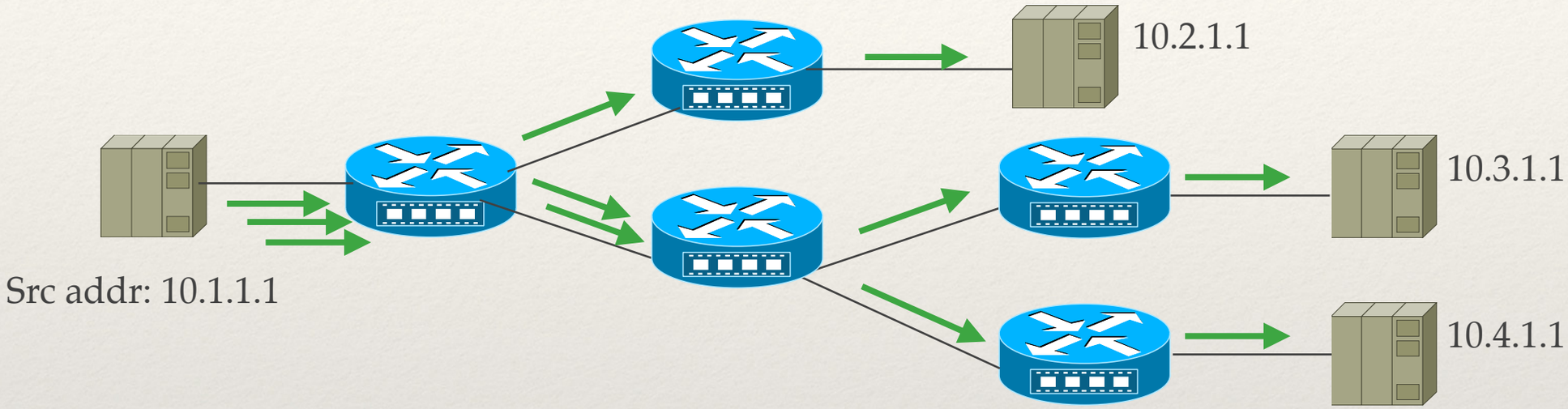
- Multicast Source Discover:
  - MSDP и PIM-SM
- SSM

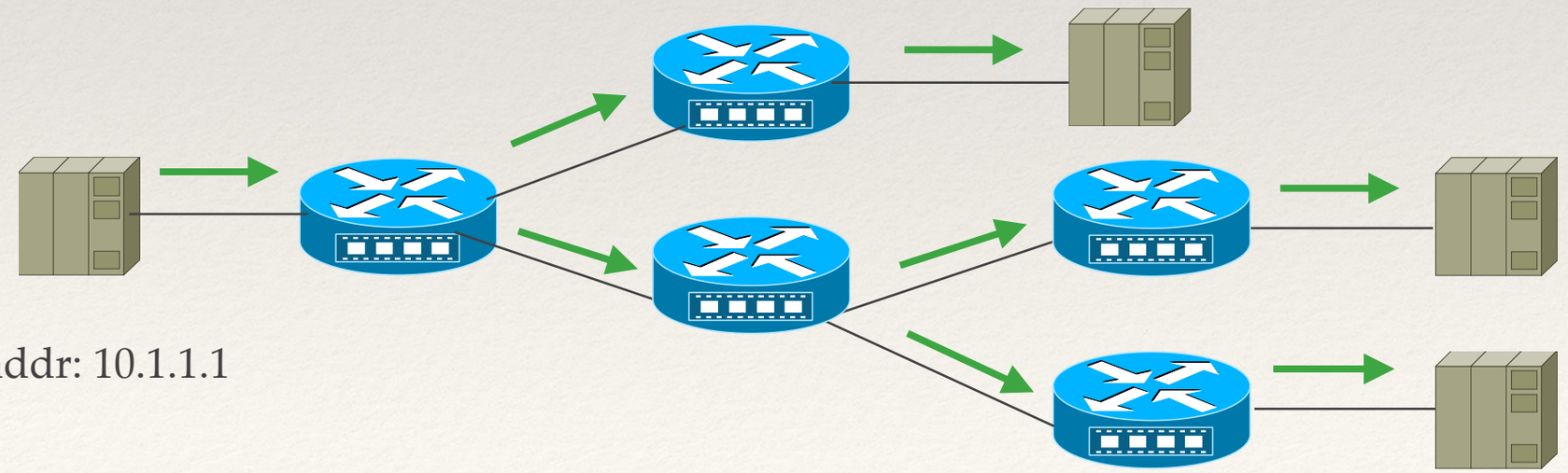
Адресация в Multicast





Адресация в Multicast





Адресация в Multicast

- Адреса класса «D» - 224/4 (224.0.0.0 - 239.255.255.255)
- Адреса групп multicast находятся вне таблицы маршрутов unicast
- Отдельная таблица для активных деревьев multicast
- Записи multicast создаются по факту сигнализации приемника о желании подключиться к группе
- Протоколы маршрутизации multicast строят деревья на основе информации о приемниках и доступности источников
- Доступность источников определяется на основе данных таблицы маршрутизации unicast.

Адресация в Multicast - 224/4

- **Link-local - 224.0.0.0/24**

  TTL = 1

  Примеры:

    224.0.0.1 - Всем в подсети

    224.0.0.2 - Всем маршрутизаторам в подсети

    224.0.0.5 - маршрутизаторам OSPF

    224.0.0.13 - маршрутизаторам PIMv2

    224.0.0.22 - IGMPv3

- **Прочие зарезервированные адреса - 224.0.1.0/24**

  TTL > 1

  Пример:

    224.0.1.1 - NTP

Адресация в Multicast - 224/4

- **Administratively scoped**
  - 239/8 (239.0.0.0 - 239.255.255.255)
  - Приватное адресное пространство

    Аналог адресов RFC1918

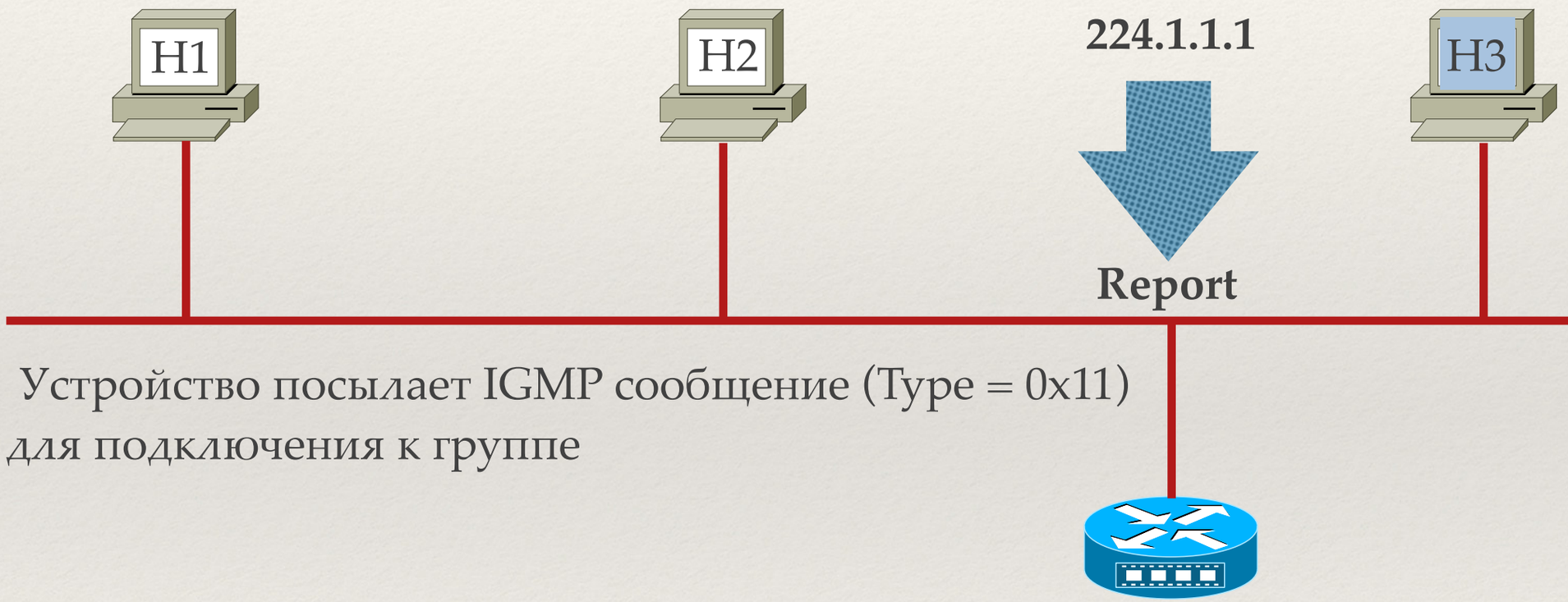
    Не маршрутизируются в Internet
- **GLOP**
  - 233/8 (233.0.0.0 - 233.255.255.255)
  - Распределяется блоками /24 на ASN
  - RFC2770
- **SSM (Source Specific Multicast)**
  - 232/8 (232.0.0.0 - 232.255.255.255)
  - Предполагалось использовать в качестве internet-style broadcast

Сигнализация host-router: Internet Group Management Protocol (IGMP)

- Конечные устройства сообщают маршрутизатору о членстве в группе
- Маршрутизатор принимает такие сообщения только от устройств из напрямую подключенных сетей (работает в пределах broadcast domain)
- Существует три варианта протокола:
  - RFC1112 - версия 1
  - RFC2236 - версия 2
  - RFC3376 - версия 3

Сигнализация host-router: IGMP

Присоединение к группе



Устройство посылает IGMP сообщение (Type = 0x11) для подключения к группе

Сигнализация host-router: IGMP

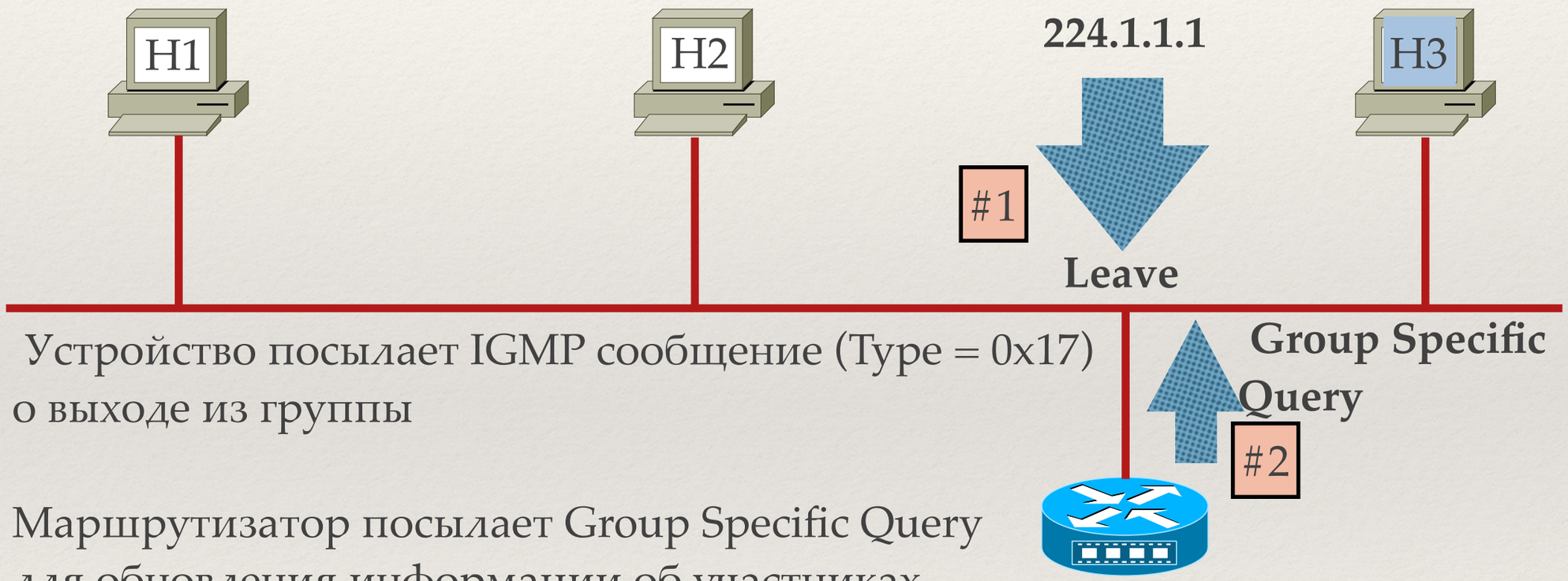
Обновление информации об участниках группы

Маршрутизатор отправляет периодические запросы (Type = 0x16) на адрес 224.0.0.1

Достаточно ответа только от одного участника группы в сети.

Сигнализация host-router: IGMP

Выход из группы



Устройство посылает IGMP сообщение (Type = 0x17) о выходе из группы

Маршрутизатор посылает Group Specific Query для обновления информации об участниках

Если по истечению таймаута (~3 сек.)
не получено ответа на запрос - группа удаляется

Сигнализация host-router: IGMPv3

RFC3376 - поддержка SSM

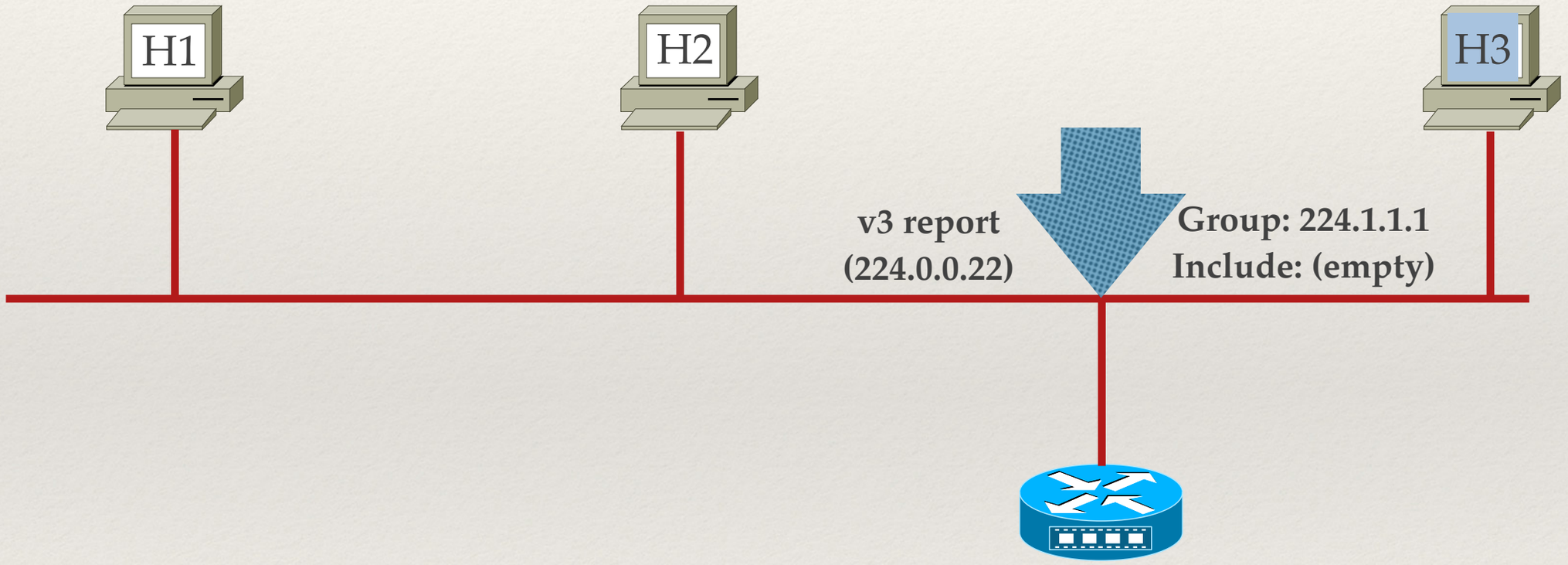
- Добавлен функционал списков источников (include/exclude)
- Устройство имеет возможность избирательно слушать источники
- Требуется поддержка «IPMulticastListen» API
- Требуется реализация стека IGMPv3 в ОС
- Требуется поддржка функционала include/exclude lists со стороны приложений

Сигнализация host-router: IGMPv3

- 224.0.0.22 (IGMPv3 Routers)
  - Все устройства IGMPv3 отсылают отчеты на этот адрес
  - Маршрутизаторы IGMPv3 слушают этот адрес
- Отсутствует механизм подавления отчетов
  - Все устройства в broadcast domain отсылают отчеты на запросы об участниках группы
  - Таймер ответа может быть настроен в очень широких пределах (~12 секунд)

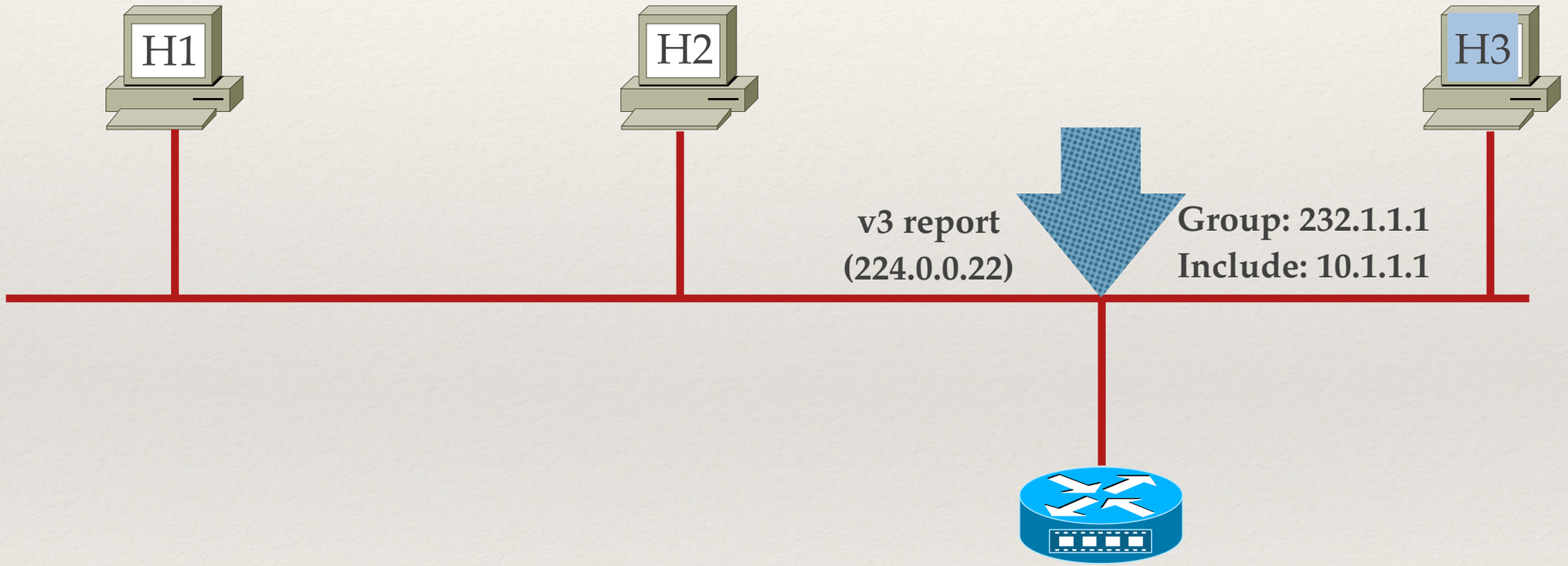
Сигнализация host-router: IGMPv3

Присоединение к группе с *любым* источником



Сигнализация host-router: IGMPv3

Присоединение к группе с указанием источника



Multicast L3 Forwarding

- Unicast routing - КУДА доставить пакет
- Multicast routing - ОТКУДА пакет пришел

Unicast vs Multicast Forwarding

Multicast Forwarding

- IP адрес получателя (группы) не указывает напрямую, куда нужно доставить пакет
- Форвардинг основан на информации об исходящем интерфейсе (OIF)
- Для того, чтобы трафик начал «ходить» получатели должны быть уже «включены» в дерево

    Сообщения PIM join передаются источнику на основании маршрутной информации unicast

    Построение дерева multicast позволяет узнать, куда нужно отправлять пакеты

    При изменении топологии сети дерево динамически перестраивается

    Каждый маршрутизатор на пути следовани поддерживает актуальный список OIF

Reverse Path Forwarding (RPF)

Расчет RPF

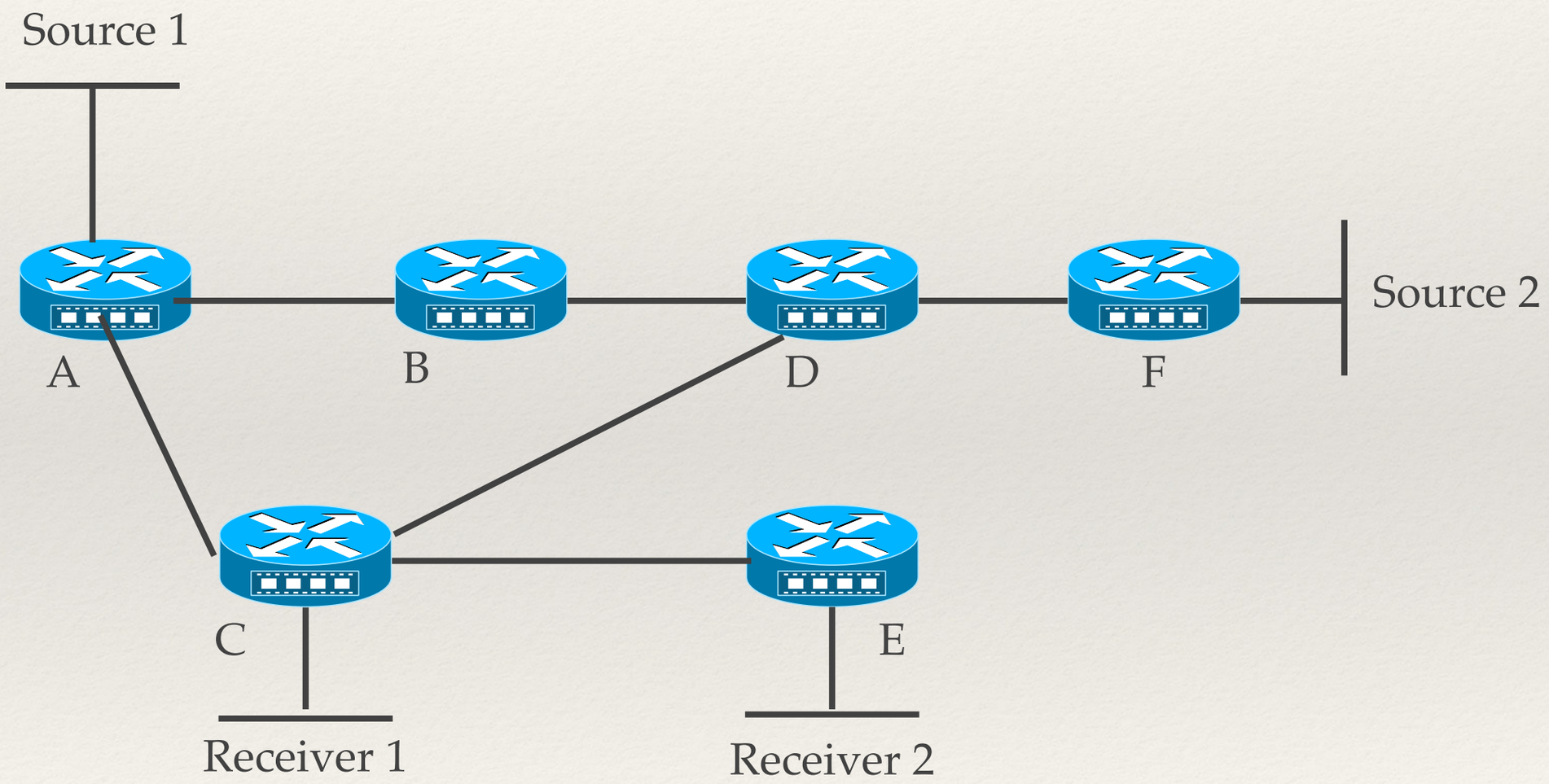
- Адрес источника пакетов multicast проверяется на доступность в таблице маршрутов unicast
- Определяется интерфейс в сторону upstream маршрутизатора, куда направляются пакеты PIM join
- Такой интерфейс считается «входящим» или RPF-интерфейсом

  Маршрутизатор обрабатывает дейтаграммы multicast только в случае, если они получены из RPF-интерфейса

Создание Multicast-дерева

- IGMP membership reports
  - используются конечными устройствами для подтверждений участия в группе
- PIM сообщения join/prune
  - используются для создания/удаления деревьев
- Shortest Path Trees (SPT)
  - кортеж (S,G)
  - управляющие сообщения PIM передаются источнику
- Shared Trees
  - кортеж (*,G)
  - управляющие сообщения PIM передаются RP

Multicast Distribution Trees



Multicast Distribution Trees

- Source or Shortest Path Trees (S,G)
  - Требует больше памяти *O(S x G)*
  - оптимальные маршруты от источника к получателю
  - минимальные задержки
- Shared Trees (*,G)
  - требует меньше памяти *O(G)*
  - не всегда оптимальные маршруты от источника к получателю
  - может вносить задержки

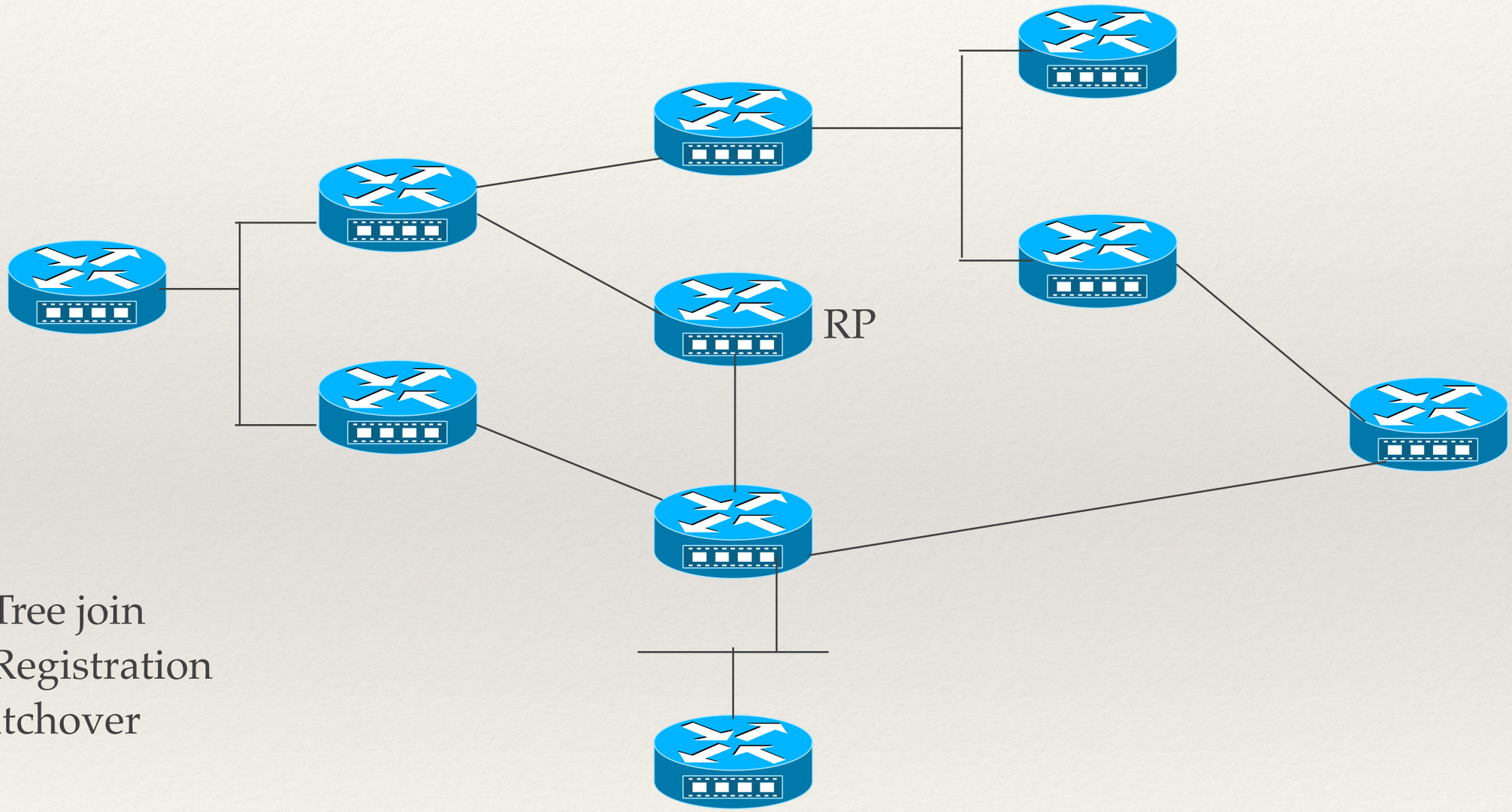
PIM

- PIM - Dense Mode

  Flood and prune

- PIM - Sparse Mode
  - Any Source Multicast (ASM), используются RP/SPT/shared tree
  - Source Specific Multicast (SSM), без RP, только SPT
- Bidirectional PIM
  - только shared tree

PIM-SM



Shared Tree join
Sender Registration
SPT Switchover

PIM-SM

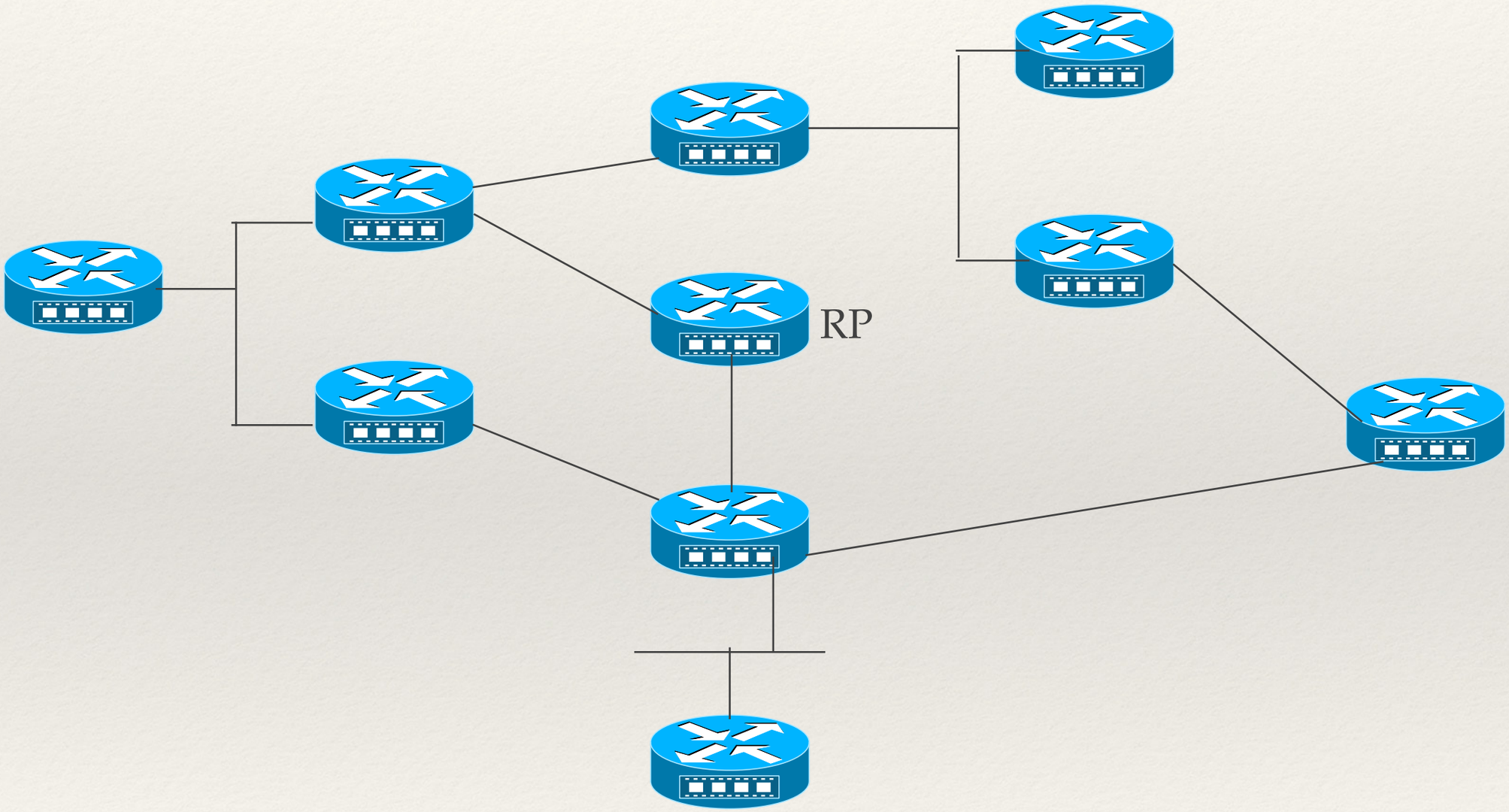
- Преимущества
  - трафик посылается только тем, кто подключился к дереву
  - есть возможность переключаться на оптимальное source-tree
  - не засисит от протокола маршрутизации
  - основа для междоменной маршрутизации multicast
    - вместе с MBGP, MSDP и/или SSM
- Недостатки
  - размещение RP может быть неоптимальным
- Область применения
  - multicast сети с любой плотностью получателей

Source Specific Multicast - SSM

Описан в RFC3569: An Overview of Source Specific Multicast (SSM)

- Предполагает использование в модели «one-to-many»
- Получатель регистрируется по IGMPv3
- Shared tree и RP не нужны
- Различные источники могут использовать одинаковые адреса групп

PIM-SM SSM



SSM

- Преимущества
  - Используется упрощенное подмножество PIM-SM
  - Решает проблему распределения адресов multicast
  - Эффективный и более безопасный
- Недостатки
  - Требуется поддержка IGMPv3
  - Необходимы механизмы доставки информации об источнике до приложений
- Область применения
  - Системы с одним источником для множества потребителей

Проблема «many-to-many»

- Создание и обработка большого количества (S,G)
- Выход - использование только shared tree
- Необходимо решение, которое использует только (*,G) - Bidirectional PIM
  - описан в RFC5015
  - применим в случаях, когда в одном сегменте присутствуют как источник, так и потребитель

Bidirectional PIM

- Идеален для решений «many-to-many»
- Преимущества
  - Значительно уменьшает таблицы mroute
  - Устраняет все (S,G)
    - отсутствуют SPT между RP и источником
    - трафик источника передается по shared tree
    - потенциально - неограниченное количество источников
- Недостатки
  - весь трафик должен передаваться через RP
  - продолжительное время восстановления при сбоях

PIM-SM ASM RP: ключевые моменты

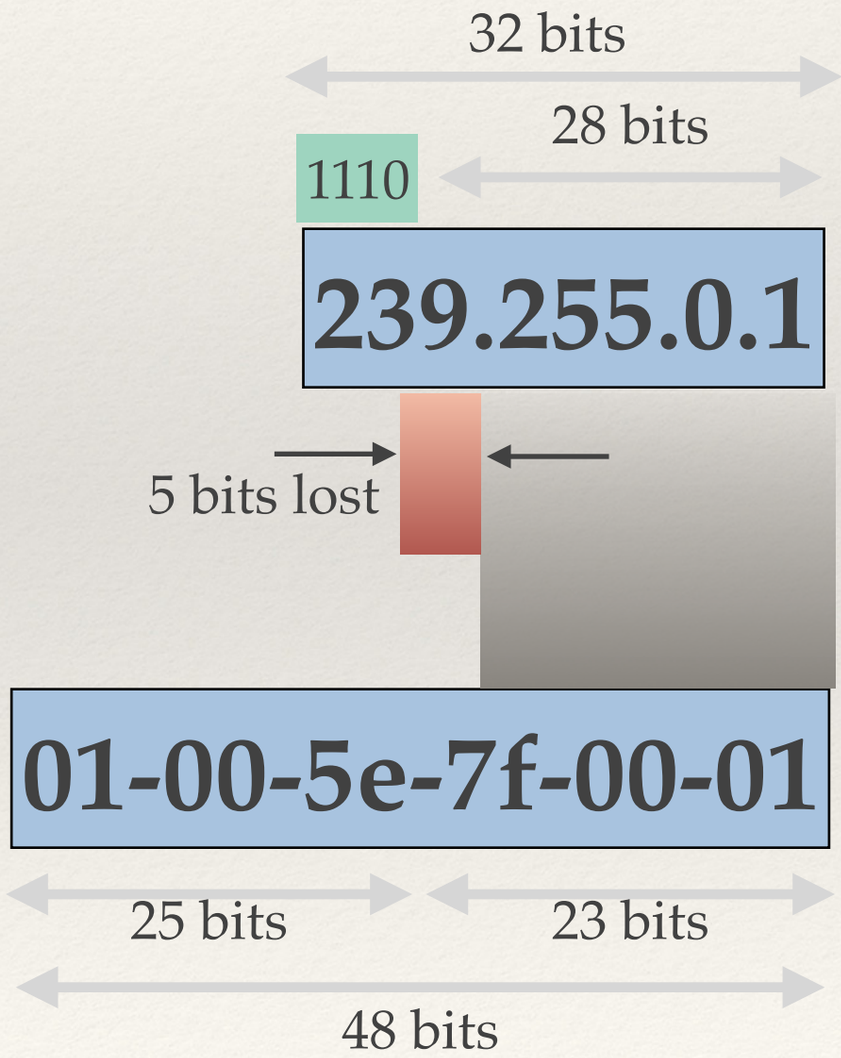
- Расположение
- Адрес RP должен быть известен всем маршрутизаторам в пределах домена PIM
- Соответствие группы-RP
  - согласовано между всеми маршрутизаторами в пределах домена PIM
- Рекомендуется дублирование RP

Как в сети узнают адрес RP?

- Статическая конфигурация
  - если статика - то статика на всех устройствах в домене
  - RP failover по умолчанию невозможен
    - отказоустойчивость обеспечивается MSDP
- AutoRP
  - Cisco solution :)
- Bootstrap Router (BSR)
  - описан в draft-ietf-pim-sm-bsr

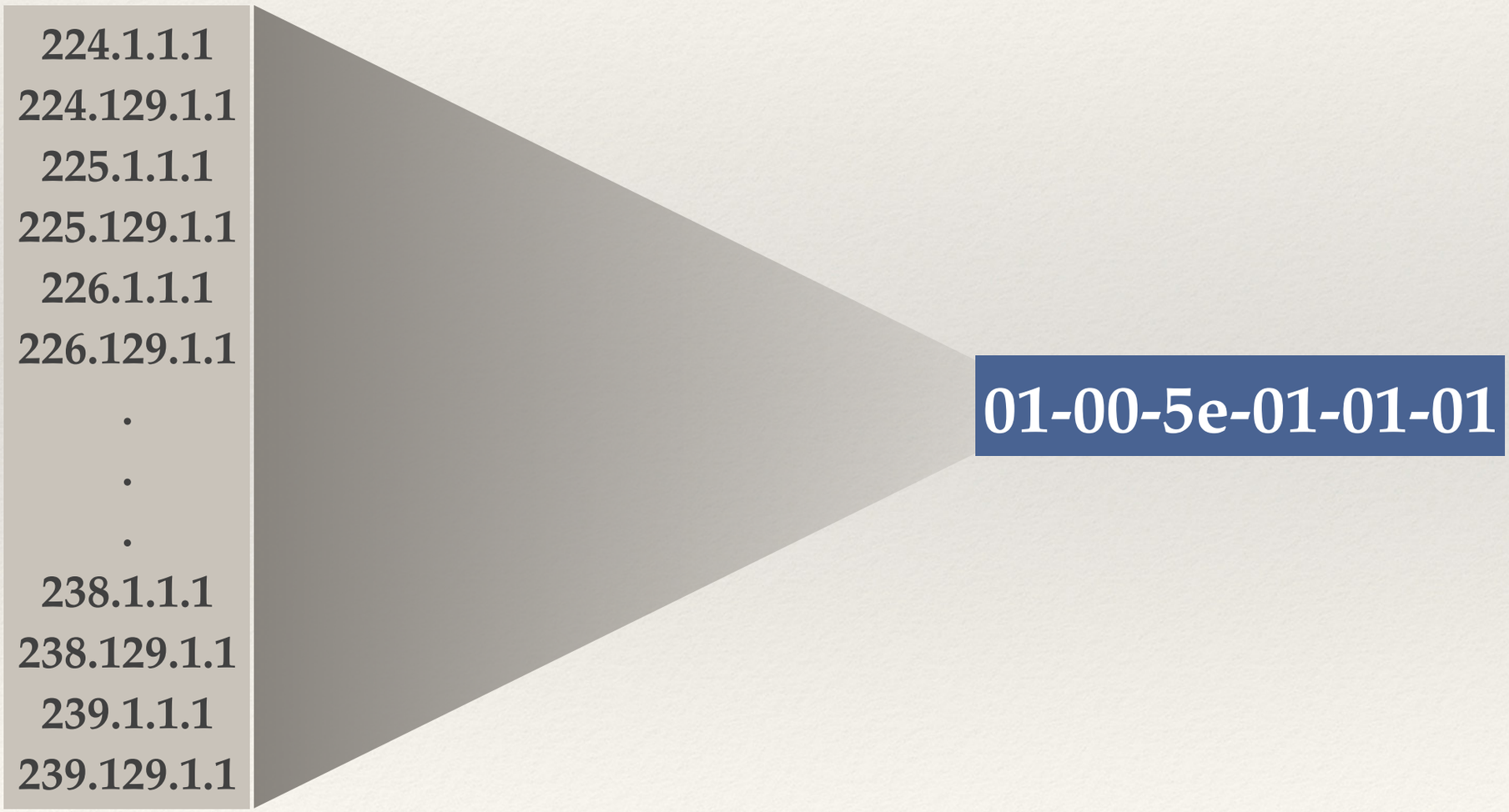
Адресация L2 Multicast

IP Muticast MAC address mapping



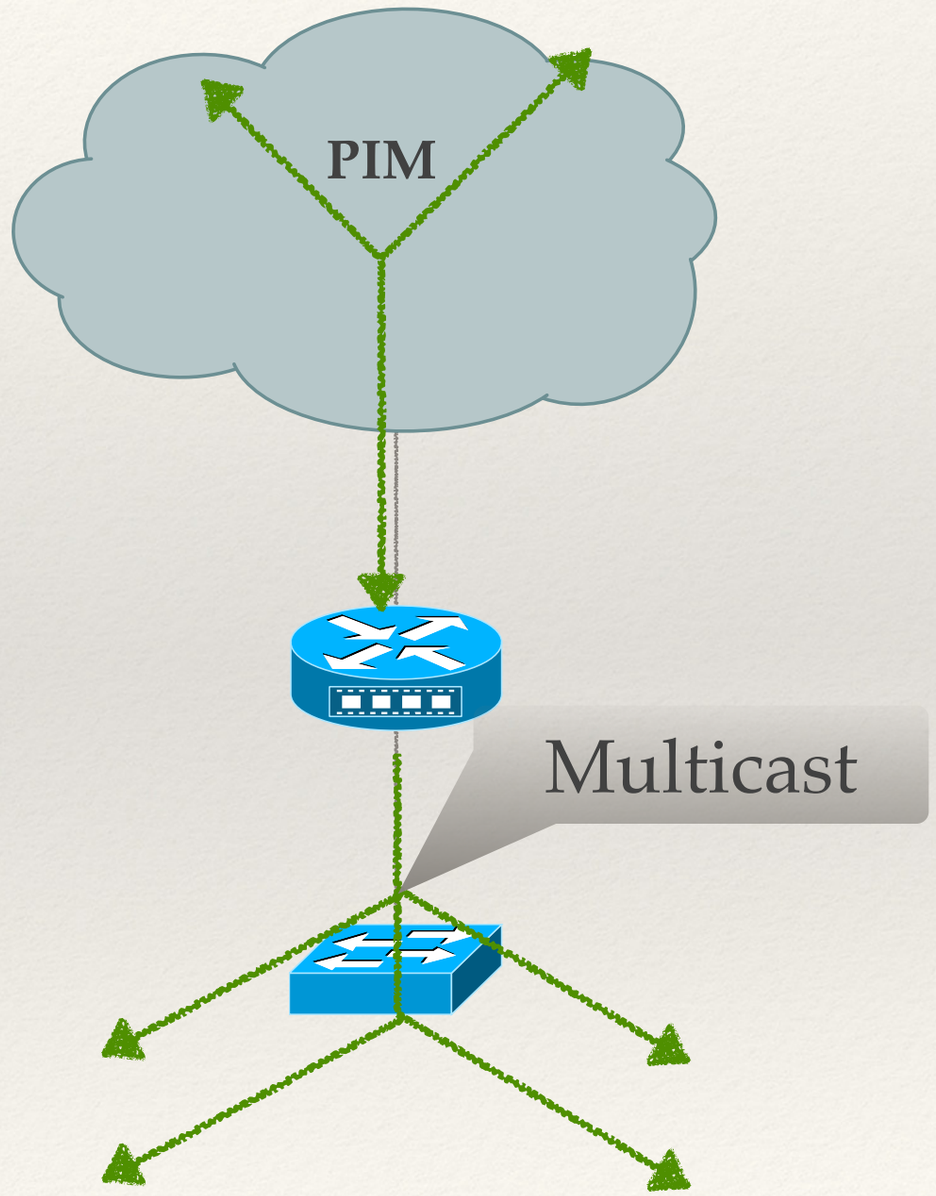
Адресация L2 Multicast

IP Muticast MAC address mapping



Коммутация кадров L2 Multicast

Обычный L2 коммутаор обрабатывает
кадры multicast как broadcast и должен разослать
их во все порты



Коммутация кадров L2 Multicast

IGMP v1-v2 Snooping

- Коммутаторы стали «знать» про IGMP
- Пакеты IGMP обрабатываются либо процессором либо ASIC
- Коммутатор должен разобрать пакет IGMP
- Влияет на работоспособность коммутаторов начального уровня

С IGMPv3 все проще :)

MBGP - Multiprotocol BGP

- Описан в RFC2858
- Позволяет BGP работать с различными типами маршрутной информации
  - unicast
  - multicast
  - MPLS L3 VPN
- Вся информация передается в пределах одного процесса BGP
- Не распространяет информацию об multicast
  - Это работа для PIM
- Те же механизмы выбора пути и правила проверки валидности
  - AS-Path, Local-Pref, MED, …

MBGP - Multiprotocol BGP

- Отдельные таблицы BGP содержат:
  - префиксы unicast для unicast forwarding
  - префиксы unicast для multicast RPF
  - реализовано с помощью Address Family Indicator (AFI)
- AFI=1, Sub-AFI=1
  - содержит префиксы unicast для unicast forwarding
  - распространяется с BGP unicast NLRI
- AFI=1, Sub-AFI=2
  - содержит префиксы unicast для multicast RPF
  - распространяется с BGP multicast NLRI

MBGP - Multiprotocol BGP

- Один и тот же IP адрес может иметь двойное значение:
  - маршрутная информация unicast
  - информация об multicast RPF
- *Для одного и того же IPv4 адреса может быть два различных NLRI с разными значениями next-hop*

MSDP - Multicast Source Discovery Protocol

- Необходим для передачи между доменами информации об источниках
- Описан в RFC3618
- Работает только с PIM-ASM

- RP «знает» о всех источниках в своем домене

  Передает информацию о своих источниках в другие домены с помощью сообщения MSDP SA (Source Active)

- RP «знает» о потребителях в домене

  RP подключается к source tree в соседнем домене с помощью обычного PIM (S,G) join

При SSM MSDP не нужен

Дополнительные материалы

- BRKIPM-1261 «Introduction to Multicast», Baron Rawlins, CiscoLive, 2013
- «Developing Ip Multicast Networks Volume I», Beau Williamson, Cisco Press
- RFC :)