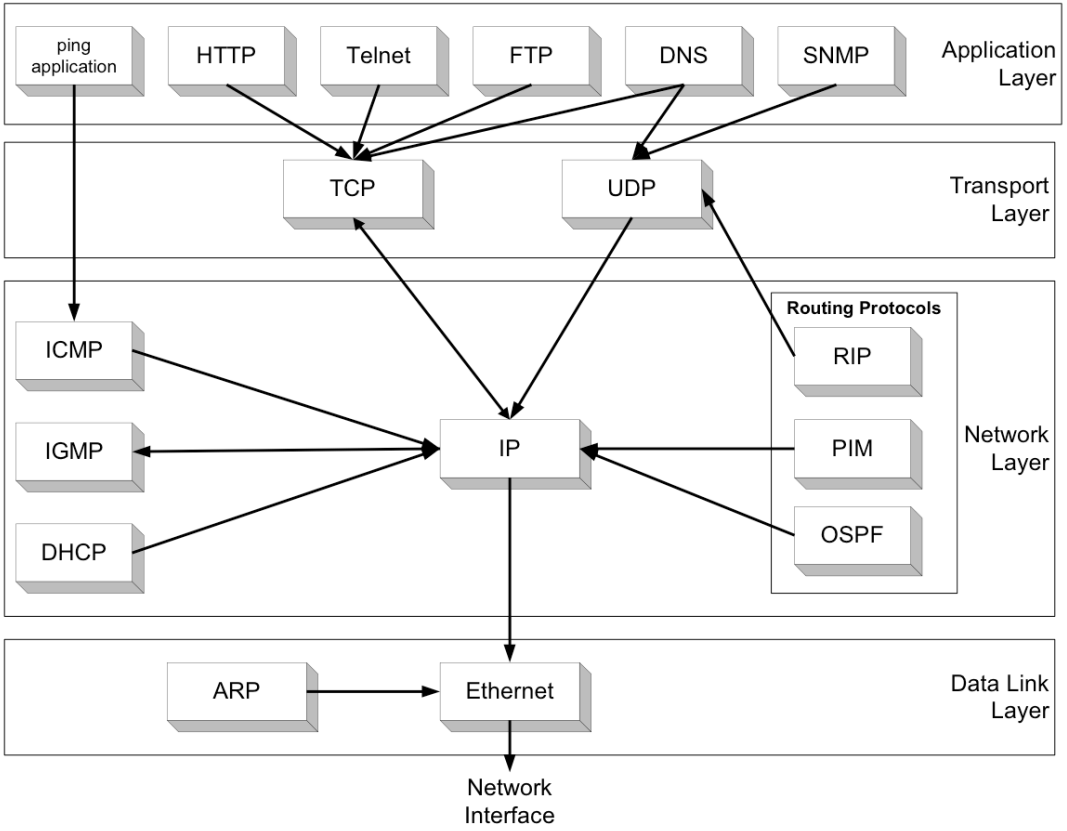
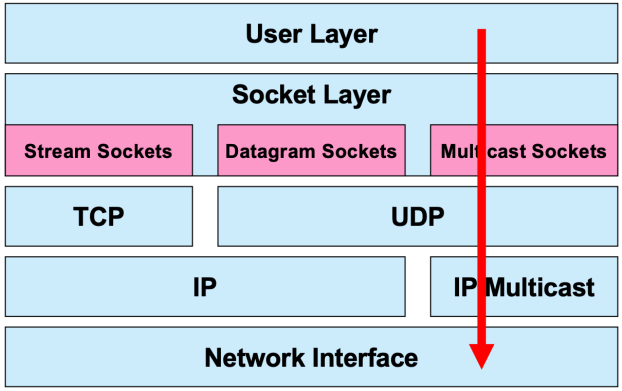
Групповая рассылка (Multicasting, IGMP, PIM).



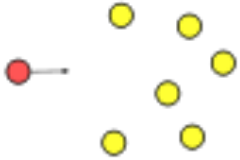
Оглавление



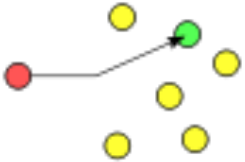
1. Основные сведения о мультикастинге
   - 1.1. Виды рассылок
   - 1.2. Преимущества групповой рассылки
   - 1.3. Недостатки групповой рассылки
   - 1.4. Требования к сети и устройствам, групповая адресация на уровне IP и MAC
2. Протокол IGMP в LAN и сегментах Internet
   - 2.1. Формат IGMP дейтаграммы
   - 2.2. Функционирование IGMP протокола
3. Методы маршрутизации групповых дейтаграмм в Internet
   - 3.1. Веерная рассылка (Flooding)
   - 3.2. Остовые деревья (Spanning Trees)
   - 3.3. RPF
   - 3.4. CBT
4. Протоколы маршрутизации групповых дейтаграмм в Internet
   - 4.1. DVMRP
   - 4.2. MOSPF
   - 4.3. CBT
   - 4.4. PIM
   - 4.4.1. PIM DM (Dense Mode)
   - 4.4.2. PIM SM (Sparse Mode)
   - 4.5. Заключение

1. Основные сведения о мультикастинге.

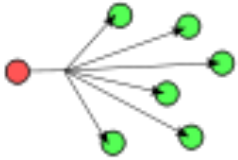
1.1. Виды рассылок.



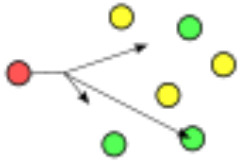
nowherecast



unicast

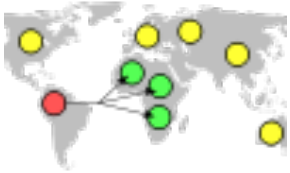


broadcast



anycast

multicast



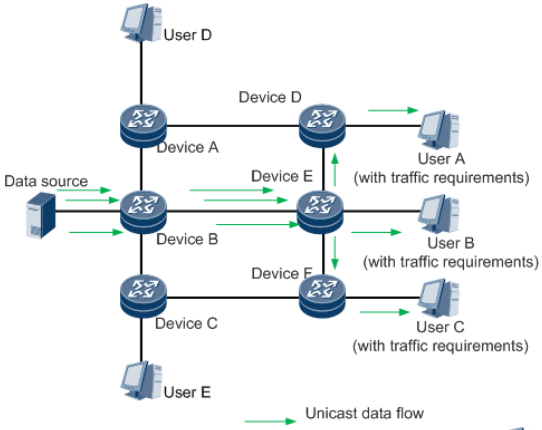
geocast

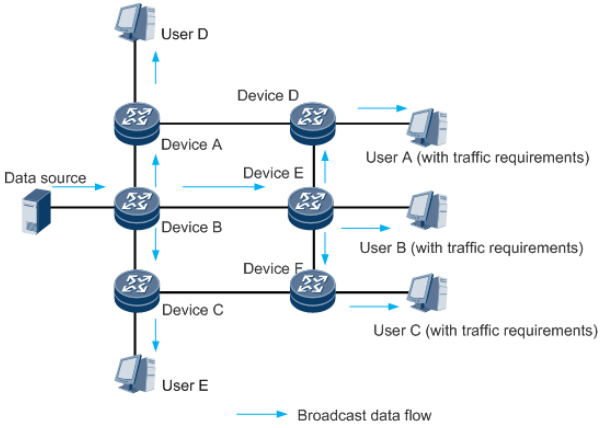
Anycast *- одноадресная рассылка ближайшему узлу, вообще получателей много, но фактически данные отправляются только одному. (Например: Anycast DNS Google с IP 8.8.8.8. Сервера с этим IP расположены в дата центрах по всему миру. В этом случае, Anycast позволяет найти оптимальный путь до ближайшего сервера).*

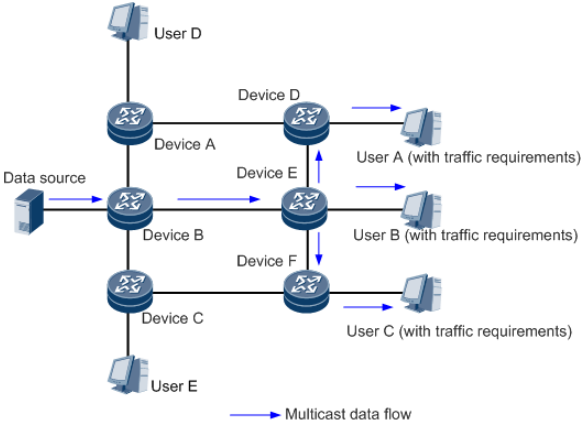
Geocast *подразумевает передачу данных для группы получателей в сети, идентифицируя их по географическому местоположению. Эта форма групповой адресации используется некоторыми протоколами маршрутизации для мобильных одноранговых сетей.*

Multicasting - рассылка дейтаграмм некоторой группе хостов.

Unicast vs Broadcast vs Multicast



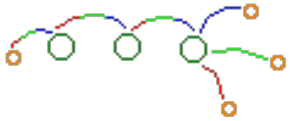




1.2. Преимущества групповой рассылки.

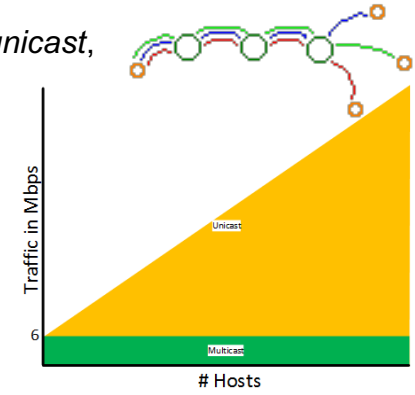
Дейтаграмма, направленная на групповой адрес, должна быть доставлена всем участникам группы. В дальнейшем такие дейтаграммы будем называть *групповыми*. Получателей дейтаграмм с определенным групповым адресом будем называть **членами данной группы**.

1. **Применения групповой рассылки** дейтаграмм достаточно очевидны и перспективны: это рассылка новостей, трансляция радио- или видеопрограмм, Shared Applications (дистанционное обучение) , и т.п. Мультикастинг активно используется и для передачи служебного трафика: маршрутной информации, сообщений службы точного времени, для группового исполнения команд различными ЭВМ.



2. Групповая рассылка, по сравнению с индивидуальной, **уменьшает нагрузку на сервер и на сеть**.

Дейтаграмму следует отправить 500 получателям. Используя *unicast*, отправитель должен сгенерировать 500 дейтаграмм, каждая из которых будет отправлена одному узлу. При *multicast* отправитель создает *одну* дейтаграмму с групповым адресом назначения; по мере продвижения через сеть дейтаграмма будет дублироваться только на "развилках" маршрутов от отправителя к получателям. Если развилок немного, например, все получатели в одной сети Ethernet, – экономия трафика будет 500-кратной. При этом сохраняются и вычислительные ресурсы промежуточных узлов.



1.3. Недостатки групповой рассылки.

1. Недостатком групповой рассылки является очевидная невозможность использования на транспортном уровне протокола TCP. Использование **протокола UDP** (часто вместе с RTP Real-time Transport Protocol на Application Level) влечёт за собой все их недостатки: **ненадежность доставки**, отсутствие средств **реагирования на заторы в сети** и т.д.

2. Кроме того, в отдельных случаях при изменении маршрутов рассылки групповые дейтаграммы **могут теряться и дублироваться**, и это должно учитываться приложениями.

3. Построение **составной сети с поддержкой мультикастинга** является гораздо более **сложной задачей**, чем организация групповой рассылки в пределах одной IP-сети (LAN).

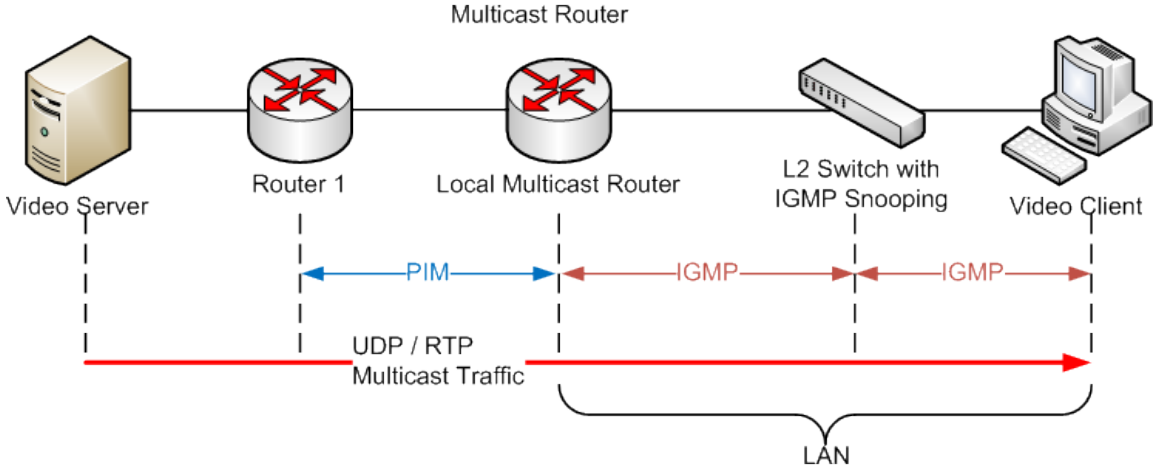
4. Для мультикастинга **нужна специальная собственная маршрутизация.**

Для продвижения групповых дейтаграмм от отправителя к получателям через систему сетей необходимо осуществлять **маршрутизацию** дейтаграмм. Однако по групповой дейтаграмме нельзя определить индивидуальные IP-адреса ее получателей, следовательно, использование обычной IP-маршрутизации и даже её принципов не имеет смысла. Поэтому для маршрутизации групповых дейтаграмм были разработаны **специальные методы и протоколы маршрутизации** (IGMP, DVMRP, MOSPF, PIM - Protocol Independent Multicast, CBT), которые будут рассмотрены ниже.

5. Некоторые провайдеры Internet **не поддерживают** мультикаст-связность.

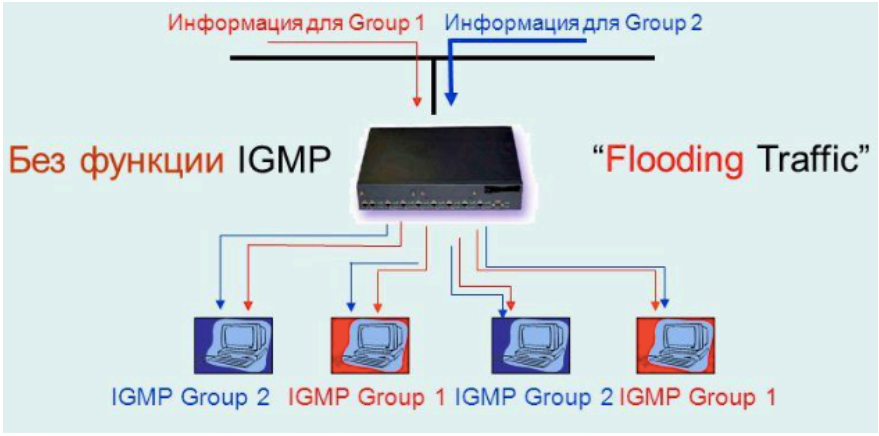
1.4. Требования к сети и устройствам, групповая адресация IP и MAC.

1. Для использования Multicast, он должен поддерживаться в стеках TCP/IP **сервера**, промежуточных **маршрутизаторов** и **клиентов**.
2. В Internet управлением мультикаст-группами занимается **PIM** (Protocol Independent Multicast), а в LAN **IGMP** (Internet Group Management Protocol).



3. Для идентификации групп **на сетевом IP-уровне** используются специально выделенные IP адреса получателя.
4. Для работы мультикастинга в LAN необходима также групповая или широковещательная рассылка на уровне доступа к сети, поэтому мультикастинг-адресация осуществляется и **на MAC-уровне**.

5. Чтобы коммутаторы посылали пакеты только нужным получателям (multicast), они должны поддерживать технологию **IGMP snooping**.
6. Без функции IGMP snooping коммутатор ретранслирует multicast трафик по всем своим портам и данные направляются всем устройствам сети, независимо от их вхождения в группы (проводится **Flooding Traffic**).



7. Для участия в мультикасте хост локальной сети должен иметь **мультикаст программу** поддерживающую Group Join (например, VideoClient VLC). Multicast is driven by receivers: Receivers indicate interest in receiving data.

There are three essential components of the IP Multicast service:

1. IP & MAC Multicast Addressing
2. IP Group Management
3. Multicast Routing

1.4.1. Групповые адреса на уровне IP.

Для идентификации групп на сетевом IP-уровне используются специальные адреса получателя. В IPv4 для мультивещания зарезервирована подсеть **224.0.0.0/4**; это адреса из класса D в диапазоне 224.0.0.0 – 239.255.255.255.

Адреса в диапазоне 224.0.0.0 – 238.255.255.255 предназначены для использования в масштабе Интернет.

Адреса вида 239.X.X.X зарезервированы для внутреннего использования в частных сетях.

Некоторые из **групповых адресов зарезервированы** строго для специальных групп (см. RFC-1700). Например:

Мультикастинг адрес	Описание
224.0.0.0 - 224.0.0.255	Local Network Control Block
224.0.0.0	Зарезервировано
224.0.0.1	Все системы данной субсети
224.0.0.2	Все маршрутизаторы данной субсети
224.0.0.4	Все DVMRP-маршрутизаторы
224.0.0.5-224.0.0.6	OSPF IGP (MOSPF) -маршрутизаторы
224.0.0.9	Маршрутизаторы RIP2
224.0.0.10	IGRP маршрутизаторы
224.0.0.13	ALL-PIM-Routers group
224.0.0.251	Multicast DNS address
224.0.1.0 - 224.0.1.255	Internetwork Control Block

224.0.1.0	VMTP-группа менеджеров
224.0.1.1	получатели информации по NTP-network time protocol
224.0.1.7	Audionews - audio news multicast (аудио служба новостей)
224.0.1.9	MTP (multicast transport protocol)
224.0.1.10	IETF-1-low-audio
224.0.1.11	IETF-1-audio
224.0.1.12	IETF-1-video
224.0.1.20	Любой частный эксперимент
224.0.1.24	microsoft-ds
224.1.0.0-224.1.255.255	ST мультикастинг-группы (ST – Spanning Tree)
224.2.0.0-224.2.255.255	Вызовы при мультимедиа- конференциях
232.0.0.0-232.255.255.255	VMTP переходные группы

Полный актуальный список зарезервированных блоков есть на сайте IANA - http://www.iana.org/assignments/multicast-addresses/multicast-addresses.xhtml.

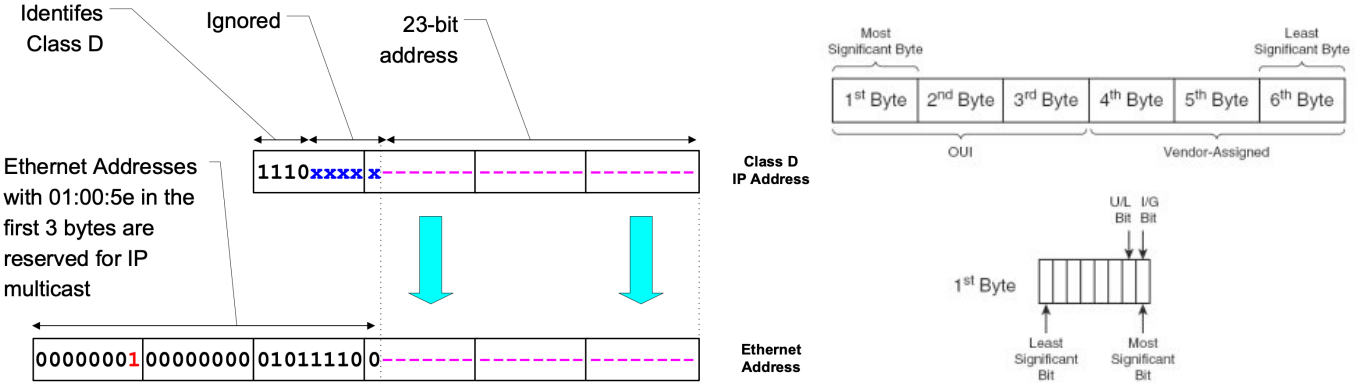
В IPv6 используются адреса **FF00::/8**.

1.4.2. Групповые адреса на уровне MAC.

For IPv4 multicast IANA зарезервировала блок MAC адресов от **01:00:5E:0**0:00:00 до **01:00:5E:7**F:FF:FF.

For IPv6 multicast IANA зарезервировала блок MAC адресовот **33:33:**00:00:00:00 до **33:33:**FF:FF:FF:FF.

Последний бит первого байта 48 bit MAC адреса, равный 1 (xxxxxxx1), указывает на то, что адрес является multicast или broadcast (а значение xxxxxxx0 - unicast).



Используется маппинг 32-х групповых IP адресов в 1 MAC адрес. Маппинг инициирует мультикаст-программа.

Данная схема резервирования адресного пространства позволяет использовать 23 бита MAC-адреса для идентификации группы рассылки при IP-мультикастинге.

Групповые IP адреса D класса резервируются первыми 4 битами, и они всегда равны 1110 (см. рис). Эти 4 бита и следующие 5 бит, отмеченные в 32-bit IP-адресе 01111 (см. рис.), игнорируются при формировании группового 48-bit MAC-адреса узла, т.к. в групповом MAC эти 24 бита (OUI) всегда установлены в 01:00:5E и 25-й бит в 0.

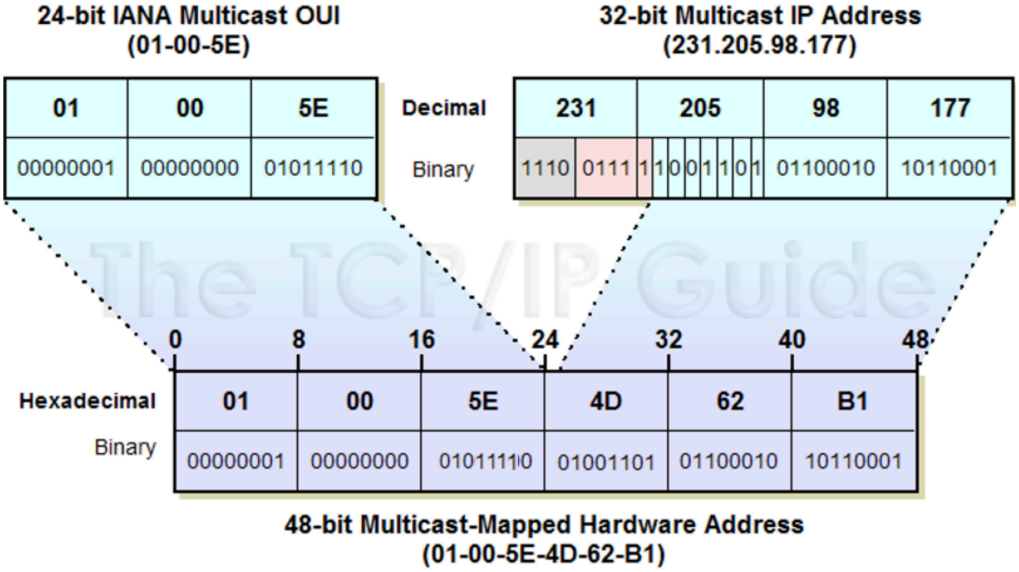


Рис. Соотношение мультикастинговых MAC- и IP-адресов.

Т.о. групповому IP-адресу 231.**205.98.177** на уровне сети будет соответствовать MAC-адрес 01:00:5E:**4D:62:B1** на уровне сегмента.

Так как это соответствие является **неоднозначным**: 32 IP-адреса с 224.205.98.177 до 239.205.98.177 и с 224.77.98.177 до 239.77.98.177 будут преобразованы в один и тот же MAC-адрес 01:00:5E:4D:62:B1, то **NIC-драйверы** должны обеспечивать специфичную обработку кадров.

Например, хост, который прослушивает IP-адрес многоадресной рассылки 231.205.98.177, настроит свою сетевую карту на прослушивание MAC-адреса 01:00:5E:4D:62:B1. Если кто-то другой ведет потоковую передачу на IP-адрес 224.77.98.177, то он также попадет на наш хост, потому что MAC-адрес тот же. NIC-драйвер должен будет просмотреть IP-адрес принятого кадра, чтобы определить, относится ли он к 239.205.98.177 и отбросить кадры, предназначенные для 224.77.98.177.

```
MB-YS:~ ys$ netstat -g
Link-layer Multicast Group Memberships
Group                   Link-layer Address      Netif
1:0:5e:0:0:1            <none>                  en0
33:33:0:0:0:1           <none>                  en0
33:33:ff:29:31:94       <none>                  en0
33:33:ff:24:c1:86       <none>                  en0
1:0:5e:0:0:fb           <none>                  en0
33:33:0:0:0:fb          <none>                  en0
1:3:93:df:b:92          <none>                  en0
33:33:0:0:0:1           <none>                  awdl0
33:33:ff:29:31:94       <none>                  awdl0
33:33:ff:2c:25:fb       <none>                  awdl0
33:33:0:0:0:fb          <none>                  awdl0
33:33:80:0:0:fb         <none>                  awdl0
33:33:0:0:0:1           <none>                  llw0
33:33:ff:29:31:94       <none>                  llw0
33:33:ff:2c:25:fb       <none>                  llw0
33:33:0:0:0:fb          <none>                  llw0

IPv4 Multicast Group Memberships
Group                   Link-layer Address      Netif
224.0.0.251             <none>                  lo0
224.0.0.1               <none>                  lo0
224.0.0.1               1:0:5e:0:0:1            en0
224.0.0.251             1:0:5e:0:0:fb           en0
```

Команды просмотра arp таблиц и членства в группах:

в Windows:
arp –a
netsh interface ipv4 show joins

в Linux/Mac:
arp -a
netstat –g.

```
MB-YS:~ ys$ arp -a
? (192.168.111.1) at cc:2d:e0:e7:8e:98 on en0 ifscope [ethernet]
? (192.168.111.10) at 0:26:b9:68:3c:40 on en0 ifscope [ethernet]
? (192.168.111.17) at 9c:4:eb:da:70:f2 on en0 ifscope [ethernet]
? (224.0.0.251) at 1:0:5e:0:0:fb on en0 ifscope permanent [ethernet]
? (239.255.255.250) at 1:0:5e:7f:ff:fa on en0 ifscope permanent [ethernet]
```

2. Протокол IGMP в LAN и сегментах Internet.

Протокол *IGMP* (*Internet Group Management Protocol*) предназначен для регистрации на маршрутизаторе членов групп, находящихся в непосредственно присоединенных к нему сетях (в LAN). Имея эту информацию, маршрутизатор может сообщать другим маршрутизаторам (с помощью протоколов групповой маршрутизации DVMRP, MOSPF, PIM, CBT) о необходимости пересылки ему дейтаграмм для имеющихся у него групп.

IGMPv1 - RFC-1112, IGMPv2 - RFC-2236, IGMPv3 - RFC-3376.

2.1. Формат IGMP дейтаграммы.

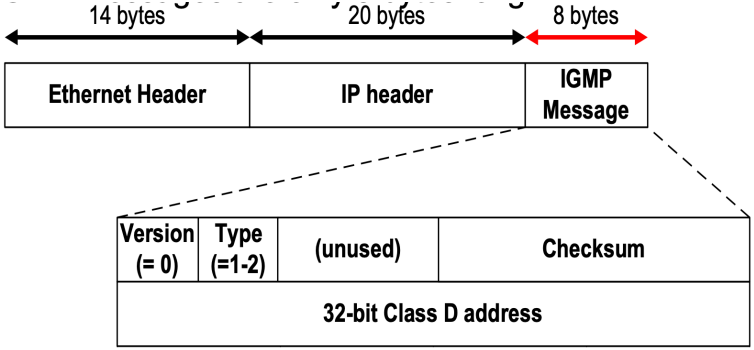
IGMP работает непосредственно поверх IP, и идентифицируется значением поля **"Protocol"=2** в заголовке IP-дейтаграммы.

По умолчанию мультикаст-дейтаграммы имеют значение поля **TTL=1**, что ограничивает их распространение одной субсетью.

Приложения могут увеличивать значение TTL. Первая дейтаграмма, тем не менее, всегда имеет TTL=1. Если получение этой дейтаграммы не подтверждается сервером, посылается вторая - с TTL=2 и т.д. Попутно измеряется и число шагов между клиентом и сервером.

Для случая, когда число шагов не более 1 (для LAN), зарезервирован блок IP адресов 224.0.0.0 - 224.0.0.255. Маршрутизатор не обрабатывает пакеты с такими адресами.

За IP-заголовком в дейтаграмме следует **сообщение IGMP**:



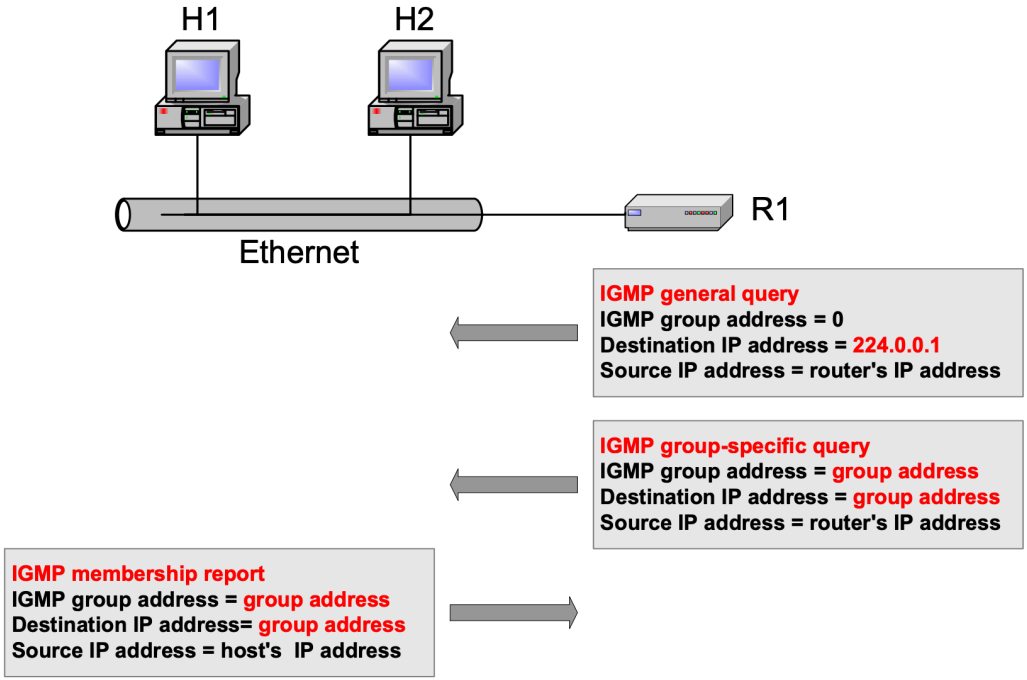
- **Type** (8 бит) если начинается с 1, то это запрос, отправленный мультикаст-маршрутизатором; начинается с 2 это отклик узла.
- **Unused** or **Max Response Time** (8 бит) – максимальное время отклика, задействовано только в сообщениях типа *Membership Query (в IGMPv1 не использовалось и ставилось в 0)*.
- **Checksum** (16 бит) – контрольная сумма.
- **Group Address** (32 бита) – групповой IP-адрес.

Существуют следующие типы сообщений:

- *Membership Query* (Type=17) – запрос о наличии в сети членов групп (отправляется маршрутизатором). Запросы обо всех имеющихся группах (***общие*** запросы) – отправляются по адресу 224.0.0.1 ("всем узлам"); запросы о наличии членов определенной группы (***частные*** запросы) – отправляются по адресу этой группы.
- *Membership Report* (Type=22) – уведомление о наличии в сети члена группы (отправляется хостом – членом группы по адресу группы).
- *Leave Group* (Type=23) – уведомление об отсоединении хоста от группы (отправляется отсоединившимся хостом по адресу 224.0.0.2 – "всем маршрутизаторам").

2.2. Функционирование IGMP протокола.

1. Маршрутизатор периодически рассылает по адресу **224.0.0.1** (всем узлам в LAN) общий запрос *Membership Query*, при этом поле "Group Address" обнулено. Период этих рассылок может меняться администратором; значение по умолчанию – **125 sec**.



2. Приняв такой запрос, каждый получатель групповых дейтаграмм выжидает случайное время "Max Response Time" из *Membership Query*. Если за это время кто-то другой уже ответил маршрутизатору, то данный хост не отвечает, иначе он сам посылает *Membership Report*.

Max Response Time (время задержки - обычно **10 sec**) позволяет избежать посылки множества ответов с адресом одной и той же группы: **маршрутизатору** не нужно знать, сколько именно членов данной группы есть у него в сети, ему **требуется лишь сам факт наличия членов**.

3. Сообщение *Membership Report* посылается по адресу группы, и этот же адрес помещается в поле "Group Address". Следует отметить, что **маршрутизатор является членом всех групп**, то есть получает сообщения, направленные на любой групповой адрес.

4. Если хост является членом нескольких групп, то вышеописанная процедура с выжиданием и отправкой ответа выполняется независимо для каждой группы.

5. При подключении хоста к новой группе он самостоятельно отправляет сообщение типа *Membership Report*, не дожидаясь очередного запроса от маршрутизатора.

6. Для каждой группы, члены которой обнаружились в сети, маршрутизатор ведёт отсчет времени неактивности. Если ни одного *Membership Report* для этой группы не было получено за определенный период (по умолчанию – **260 sec**), то маршрутизатор считает, что членов этой группы в сети больше нет.

7. Когда хост отсоединяется от группы, он может послать сообщение *Leave Group* по групповому адресу **224.0.0.2** ("всем маршрутизаторам"); адрес группы содержится в поле "Group Address". Хосту *следует* сделать это, если на последний запрос *Membership Query* от имени данной группы отвечал именно этот хост.

8. Получив сообщение *Leave Group*, маршрутизатор генерирует частный запрос *Membership Query* для членов только этой группы. Если за время, указанное в поле "Max Response Time" запроса (по умолчанию – **1 sec**), маршрутизатор не получил ни одного *Membership Report*, он считает, что членов данной группы в сети больше нет. Для надежности запрос шлётся 2 раза.

9. Если к одной сети подключены несколько маршрутизаторов, поддерживающих протокол IGMP, то запросы рассылает только маршрутизатор с **наименьшим IP-адресом** (то есть, если маршрутизатор получил из сети *Membership Query* с IP-адресом отправителя меньшим, чем его собственный адрес, он должен перестать посылать запросы и перейти в режим прослушивания обмена IGMP-сообщениями).

10. Для обратной совместимости с первой версией протокола IGMP предусмотрено сообщение *Membership Report version 1* (Type=18), а также некоторые специальные действия протокола.

11. Членство в группе может динамично меняться. Любой хост может войти в группу и выйти из группы в любое время по своей инициативе, хост может быть членом большого числа групп.

3. Методы маршрутизации групповых дейтаграмм в Internet.

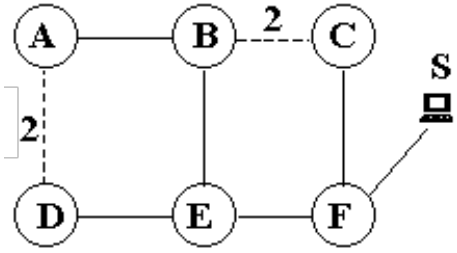
Существует несколько алгоритмов и протоколов для построения мультивещательного дерева.

Основным **предположением**, которое при этом делается, является то, что **маршрутизатор знает, члены каких групп** находятся в непосредственно подсоединенных к нему сетях.

Задачей этого раздела является описание методов (алгоритмов) маршрутизации групповых дейтаграмм, то есть продвижения их через систему сетей от отправителя к членам группы. Далее в п. 4 рассмотрены протоколы использующие эти методы маршрутизации.

3.1. Веерная рассылка (Flooding).

Веерная рассылка – наиболее простой метод маршрутизации групповых дейтаграмм, при котором дейтаграмма рассылается во все сети системы независимо от наличия в той или иной сети членов группы. На рисунке **S – источник трафика**; A-F – routers; 2 – это метка расстояния, у остальных она равна 1.



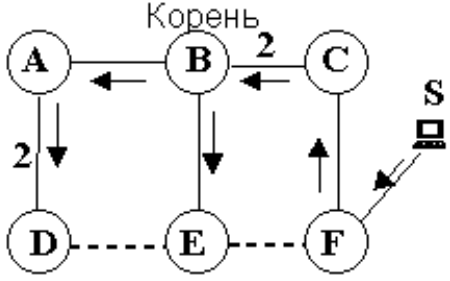
Но, для предотвращения возникновения **лавинного эффекта** от дублирования требуется проверка на повтор. При поступлении групповой дейтаграммы маршрутизатор проверяет, впервые ли он получает эту дейтаграмму. Если да, то маршрутизатор рассылает дейтаграмму через все свои интерфейсы, кроме того, с которого она была получена. Иначе дейтаграмма игнорируется.

Достоинства. Плюсы веерной рассылки: простота реализации, надежность (за счёт избыточности), независимость от маршрутных таблиц и протоколов маршрутизации.

Недостатки. Маршрутизатор должен хранить в памяти список всех "недавно" полученных групповых дейтаграмм от каждого источника для каждой группы и производить поиск в этом списке при получении каждой дейтаграммы. При интенсивном групповом трафике это потребует больших затрат памяти и мощности процессора.

Другим существенным недостатком этого метода является то, что групповая дейтаграмма рассылается от источника **всеми возможными путями**: в некоторые сети дейтаграмма может быть передана несколько раз (разными маршрутизаторами). При этом наличие или **отсутствие получателей не принимается в расчет**.

3.2. Остовые деревья (Spanning Trees - ST).



В системе сетей ABCDEF выбирается **корневой маршрутизатор B**, после этого из графа системы выделяется подграф-дерево, соединяющий корневой маршрутизатор со всеми остальными маршрутизаторами системы ("***остовое дерево***"). Процедура производится лишь при инициализации системы, а в процессе работы сети ST не изменяется. Дерево не является минимальным.

На рис. показана рассылка групповой дейтаграммы по остовому дереву: S – источник, A-F – маршрутизаторы; ***ветви остового дерева*** *обозначены сплошными линиями; метрики всех сетей, кроме явно указанных, равны 1.*

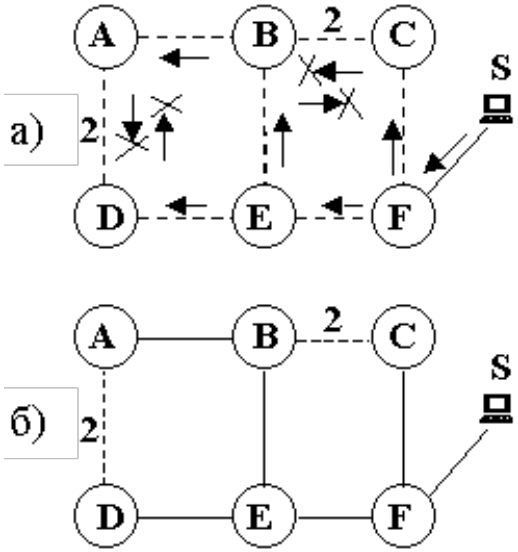
После построения ST каждый маршрутизатор хранит для каждого из интерфейсов только **флаг "этот интерфейс принадлежит ST"**. Групповая дейтаграмма от любого узла S распространяется следующим образом: полученная маршрутизатором дейтаграмма ретранслируется через все интерфейсы, принадлежащие остовому дереву, кроме того интерфейса, с которого она была получена.

Достоинства. Метод остовых деревьев несколько лучше веерной рассылки – т.к. теперь дейтаграммы распространяются по строго определенным маршрутам и в каждую сеть попадает только **один экземпляр** дейтаграммы. Также существенно **уменьшена нагрузка** на маршрутизаторы, которым больше не требуется хранить "исторические" таблицы дейтаграмм.

Недостатки. Однако групповые дейтаграммы по-прежнему рассылаются во все сети **независимо от наличия в них получателей**, кроме того:

- в сети требуется реализовать механизм (протокол) **выбора корневого узла** и **построения ST**;
- весь групповой трафик ложится **на одни и те же связи** (сети), составляющие, возможно, небольшое подмножество всей системы связей сетей;
- для некоторых пар отправитель-получатель путь по установленному дереву будет **неоптимальным путем** (например для источника S и получателей, подсоединенных к маршрутизатору D на рис. выше будет выбран путь S-F-C-B-A-D=7, а мог быть S-F-E-D=3).

3.3. RPF (Reverse Path Forwarding).



На рис. показан метод RPF: а) рассылка дейтаграмм; б) дерево рассылки, S – источник, A-F – маршрутизаторы; метрики всех сетей, кроме явно указанных, равны 1.

Маршрутизатор получил через некий интерфейс групповую дейтаграмму от источника S. Если через этот revers-интерфейс лежит кратчайший маршрут от данного маршрутизатора до узла S, то ретранслировать дейтаграмму через все интерфейсы кроме того, с которого она получена, иначе дейтаграмму игнорировать.

Например, маршрутизатор B проигнорирует дейтаграмму от узла C, но примет дейтаграмму от узла E и ретранслирует ее через остальные интерфейсы (рис. а).

Достоинства. В результате каждый маршрутизатор принимает для ретрансляции только те групповые дейтаграммы, которые следуют от источника к маршрутизатору по кратчайшему пути. Иными словами, дейтаграммы распространяются от источника ко всем маршрутизаторам системы по **оптимальному остовому дереву** с корнем в источнике (рис. б).

Для каждого источника такое минимальное остовое дерево **(MinST) возникает автоматически** по мере продвижения дейтаграммы по сети.

Недостатки.

Для реализации метода RPF необходимо **иметь доступ к таблице маршрутов**.

Поскольку ретрансляция групповой дейтаграммы производится маршрутизатором через все интерфейсы, кроме входного, некоторые экземпляры **дейтаграммы являются лишними** и продолжают засорять сеть. Речь идет о тех дейтаграммах, которые будут отброшены соседними маршрутизаторами на основании того, что они прибыли с "неоптимальных" интерфейсов, то есть распространялись не по ветвям MinST дерева (например, дейтаграмма, посланная узлом C к узлу B на рис. а).

3.3.1. Метод ModRPF.

Избежать ретрансляции дейтаграммы через связи, не принадлежащие дереву, можно с помощью следующей модификации алгоритма: "Полученная групповая дейтаграмма предается только в те сети, где находятся маршрутизаторы, кратчайший маршрут к которым от узла S проходит через данный маршрутизатор." Следуя этому правилу, узел C не отправит дейтаграмму в B, поскольку кратчайший путь от источника до узла B проходит не через C.

Недостатки.

Для реализации ModRPF метода необходимо иметь доступ к внутренним данным протокола внутренней маршрутизации (например, к базе данных состояния связей OSPF) – иначе нельзя сделать вывод о маршрутах, используемых другими узлами системы (источниками групповых дейтаграмм).

3.3.2. Метод PRPF (pruned - с усечением).

Следующая модификация RPF призвана **учесть наличие или отсутствие получателей** групповой дейтаграммы в сетях системы с тем, чтобы дейтаграммы рассылались только в те сети, где есть члены данной группы. Применяемый для этого метод называется ***prunes*** – усечение (от английского prune – "обрезать ветви дерева").

Пробная групповая дейтаграмма распространяется обычным образом по алгоритму RPF и достигает всех маршрутизаторов системы.

Если к какому-то "конечному" маршрутизатору не присоединены члены данной группы (это устанавливается с помощью протокола IGMP), он посылает через тот интерфейс, откуда получил групповую дейтаграмму, специальное **сообщение *Prune*** (по адресу данной группы). Это сообщение, принятое маршрутизатором, находящемся в вышестоящем узле дерева, означает "не посылать больше через этот интерфейс дейтаграммы от данного источника для данной группы". Вышестоящий маршрутизатор **помечает этот интерфейс как *pruned*** (усеченный) на определенный срок. По истечении этого срока процесс повторяется сначала.

Если *Prune* получено от всех нижележащих маршрутизаторов, маршрутизатор отправляет *Prune* еще более вышестоящему маршрутизатору – так можно **усекать целые поддеревья**.

Однако имеется **сообщение *Graft*** (от английского "прививать растение"), позволяющее быстро подсоединиться к существующему дереву (то есть отменить ранее посланное *Prune*), не дожидаясь очередной рассылки "пробной" дейтаграммы.

Достоинства метода PRPF (с усечением) чрезвычайно существенны:

- групповые дейтаграммы от каждого источника рассылаются по оптимальным путям – и эти пути определяются динамически в момент рассылки;
- при этом учитывается членство в группах – дейтаграммы в сети, где нет получателей, не рассылаются;
- групповой трафик распределяется по различными сегментам системы сетей, а не концентрируется в определенном подмножестве связей.

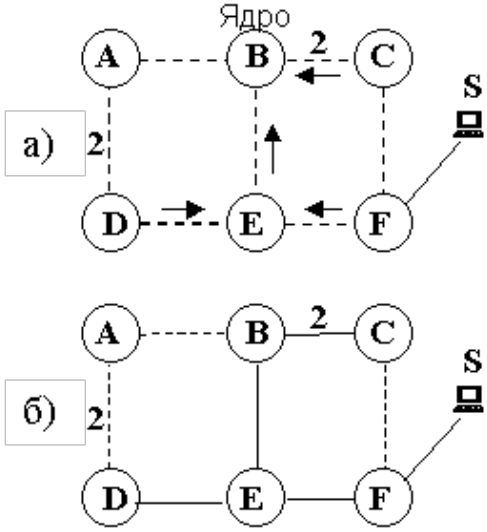
Недостатки рассматриваемого метода:

- Каждый маршрутизатор должен хранить таблицу, в которой отслеживается получение сообщений *Prune,* и производить поиск в ней при получении каждой дейтаграммы. Размер этой таблицы T=I*G*S равен произведению числа интерфейсов, числа групп и числа источников, дейтаграммы от которых проходили через маршрутизатор. (Отметим, что источники нужно запоминать тоже, так как для каждого источника создается свое дерево рассылки.) Безусловно, эта таблица не так велика, как при использовании веерной рассылки, но при интенсивном групповом трафике ее поддержка может отнять существенные ресурсы.
- Первая групповая дейтаграмма и, периодически, последующие "пробные" распространяются по всей системе сетей. При этом если в группе мало членов, а система велика (например, Интернет), возникает избыточный трафик, состоящий как из ретранслируемых экземпляров дейтаграммы, так и из потока *Prune*-сообщений, которые к тому же требуется обработать и внести в таблицу.
- Необходимость наличия интерфейса к структурам данных модуля маршрутизации (или необходимость создания "сопровождающего" протокола маршрутизации) увеличивает сложность реализации RPF.

Несмотря на недостатки, метод RPF лежит в основе многих Group Routing протоколов.

3.4. CBT (Core Based Trees).

Метод *CBT* основан на том, что для каждой группы назначается главный маршрутизатор, называемый *ядром*, – он будет корнем дерева рассылки (узел B на рис.). Все маршрутизаторы, к которым могут быть подключены потенциальные члены группы, знают адрес ядра. После того, как член группы зарегистрировался на маршрутизаторе с помощью протокола IGMP, маршрутизатор посылает в сторону ядра сообщение *Join* для присоединения к дереву рассылки. Промежуточные маршрутизаторы, пересылая это сообщение в сторону ядра, одновременно помечают интерфейсы, через которые получены сообщения *Join*, как принадлежащие дереву рассылки для данной группы. Сообщение следует до ядра или до первого маршрутизатора, уже присоединенного к дереву рассылки.



На рис. показан метод CBT. а) посылка сообщений Join; б) сформированное дерево рассылки S – источник, A-F маршрутизаторы; к маршрутизатору A не подключены члены группы; метрики сетей, кроме указанных, равны 1

Состояние принадлежности к дереву имеет определенный срок годности, поэтому периодически требуется посылка подтверждений. Каждый маршрутизатор посылает подтверждение вышестоящему (по пути к ядру) маршрутизатору. Неподтвержденные в течение некоторого времени ветви дерева усекаются.

Рассылка же самих групповых дейтаграмм маршрутизаторами происходит аналогично методу ST: дейтаграмма рассылается через все интерфейсы, принадлежащие дереву рассылки, кроме того, с которого дейтаграмма была получена. Если источник дейтаграммы не является членом группы, то его маршрутизатор сначала инкапсулирует групповую дейтаграмму в обычную одноадресную, адресованную ядру, а ядро уже инициирует групповую рассылку по дереву.

Достоинства этого метода:

- *все* групповые дейтаграммы рассылаются только участникам группы (в отличие от RPF нет "пробных" дейтаграмм);
- размер таблицы принадлежности интерфейсов к деревьям, которую требуется хранить на маршрутизаторе, меньше чем при использовании метода RPF (T=I*G*1) произведение числа интерфейсов на число групп; для всех источников S одной группы G используется одно дерево S=1);
- не требуется доступ к маршрутным таблицам.

Недостатки CBT аналогичны недостаткам метода остовых деревьев ST (р.3.2):

- весь групповой трафик ложится на одни и те же связи (сети), составляющие, возможно, небольшое подмножество всей системы сетей; узким местом является ядро;
- для некоторых пар отправитель-получатель путь по установленному дереву будет неоптимальным. Например, для источника S и получателей, подсоединенных к маршрутизатору C), рис. б (маршрут SFEBC=5, вместо SFC=2).

4. Протоколы маршрутизации групповых дейтаграмм в Internet.

1. **Distance Vector Multicast Routing Protocol (DVMRP):**
   - First multicast routing protocol
   - Implements flood-and-prune

2. **• Multicast Open Shortest Path First (MOSPF):**
   - Multicast extensions to OSPF. Each router calculates a shortest-path tree based on link state database
   - Not widely used

3. **Core Based Tree (CBT):**
   - First core-based tree routing protocol

4. **Protocol Independent Multicast (PIM):**
   - Runs in two modes: PIM Dense Mode (PIM-DM) and PIM Sparse Mode (PIM-SM).
   - PIM-DM builds source-based trees using flood-and-prune
   - PIM-SM builds core-based trees as well as source-based trees with explicit joins.

4.1. DVMRP.

Протокол DVMRP (*Distance Vector Multicast Routing Protocol*, RFC-1075) – самый старый протокол групповой маршрутизации, он используется в ядре экспериментальной сети MBONE. Протокол работает по технологии RPF с усечением, но для построения деревьев используется собственный дистанционно-векторный протокол, аналогичный протоколу RIP.

Протокол DVRMP прост в реализации и весьма эффективен, но он подходит только для небольших сетей с высокой плотностью получателей. К недостаткам метода RPF, описанным в предыдущем пункте (относительно большой размер хранимой таблицы и необходимость рассылки "пробных" дейтаграмм по всей системе сетей), добавляется ограничение на размер системы сетей, унаследованное от протокола RIP (в DVMRP “бесконечность” равна 32).

4.2. MOSPF.

Протокол *MOSPF* (*Multicast OSPF*, RFC-1584) является расширением протокола OSPF. Маршрутизатор, поддерживающий это расширение, устанавливает бит "M" в поле "Options" сообщения "Hello". В базе данных состояния связей вводится дополнительный тип записи: для указанной сети перечисляются все группы, члены которых есть в этой сети. Эти записи, как и все прочие записи базы данных состояния связей, распространяются по системе сетей с помощью протокола веерной рассылки. Для транзитной сети запись вносится в базу данных выделенным маршрутизатором.

Деревья рассылки групповых дейтаграмм строятся по методу RPF на основе базы данных

состояния связей. Отметим, что рассылка "пробных" групповых дейтаграмм и последующее усечение ненужных ветвей дерева в данном случае не производится, так как информация о наличии в сетях членов групп уже содержится в базе данных.

Протокол MOSPF имеет серьезную проблему, связанную с масштабированием: для каждой пары "источник-группа" проводится отдельный запуск алгоритма SPF для расчета дерева рассылки. При большом числе источников, а также при нестабильной топологии системы сетей, на эти вычисления затрачиваются существенные вычислительные ресурсы маршрутизаторов. Кроме того, следует учесть необходимость веерной рассылки информации о членстве в группах при ее изменении.

И, наконец, очевидно, что MOSPF требует использования OSPF в качестве протокола маршрутизации, то есть, не является независимым и может применяться только в OSPF-системах.

4.3. CBT.

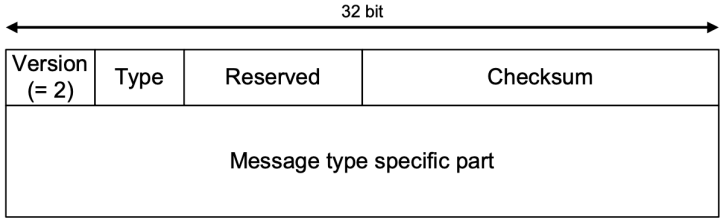
Протокол CBT (RFC-2189) реализует метод CBT так, как он описан в п. 3.4. CBT (Core Based Trees). В протоколе CBT предусмотрена возможность взаимодействия с DVMRP, см. п.4.1.

4.4. PIM.

PIM (*Protocol Independent Multicast*) – два протокола групповой маршрутизации (для плотного и разреженного расположения членов групп, соответственно *dense mode* и *sparse mode*), не зависящие от используемого протокола "обычной" маршрутизации.

PIM Messages (ver. 2)

- Encapsulated in IP datagrams with protocol number 103.
- PIM messages can be sent as unicast or multicast packet.
- 224.0.0.13 is reserved as the ALL-PIM-Routers group.



PIM-DM messages	Type	PIM-DM	PIM-SM
Hello	0	✓	✓
Register	1		✓
Register-Stop	2		✓
Join/Prune	3	✓	✓
Bootstrap	4		✓
Assert	5	✓	✓
Graft	6	✓	
Graft-Ack	7	✓	
Candidate-RP-Advertisement	8		✓

4.4.1. PIM DM.

PIM DM (Protocol Independent Multicast, Dense mode, RFC-3973) используется в системах сетей с большой плотностью получателей. Этот протокол реализует **метод PRPF (Pruned Reverse Path Forwarding)** (немодифицированный, то есть без доступа к внутренним таблицам протокола маршрутизации, вследствие чего достигается независимость от протокола маршрутизации). Необходимость периодической посылки "пробных" дейтаграмм не является существенным недостатком при плотном расположении получателей.

Достоинства. Протокол PIM DM прост в реализации и в настройке; предусмотрено взаимодействие с протоколом DVMRP.

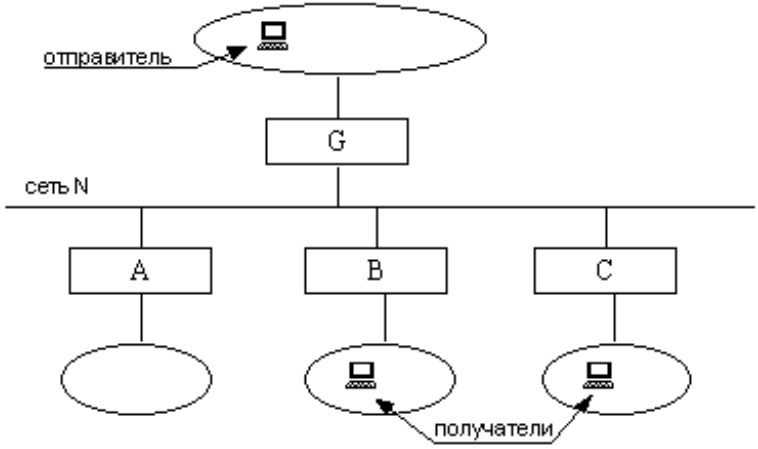
Недостатки. В качестве недостатка отметим необходимость рассылать пробные дейтаграммы каждые 3 минуты, так как за это время истекает срок действия сообщения *Prune*.

При работе протокола PIM DM могут возникнуть две особые ситуации.

(1) Особая ситуация в PIM DM.

Несколько маршрутизаторов подключены к одной широковещательной сети N, которая через вышестоящий маршрутизатор G соединяется с системой сетей, в которой находится отправитель, см. рис.

В сетях, подключенных к маршрутизаторам B и C, находятся члены группы, а в сети, подключенной к маршрутизатору A – нет.



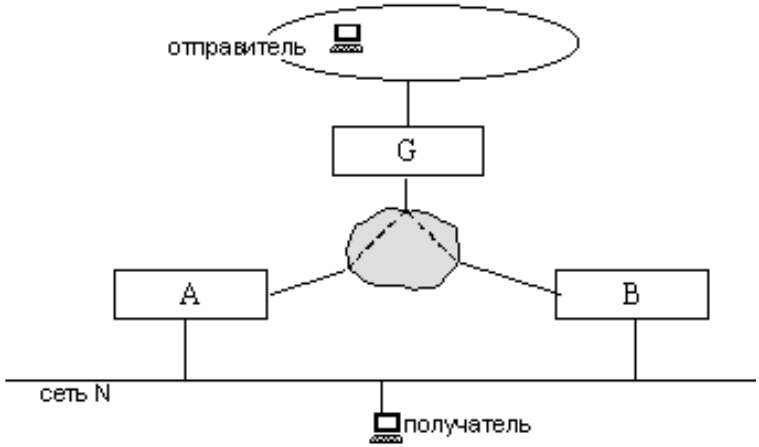
Problem. Вышестоящий маршрутизатор G посылает в сеть N первую групповую дейтаграмму. Маршрутизатор **A откликается сообщением *Prune***, однако отсекать сеть N от дерева рассылки нельзя, так как есть получатели в сетях B и C.

Solution PIM DM. Маршрутизатор G, получив *Prune*, **запускает таймер**. Маршрутизаторы B и C, прослушивая сеть, обнаруживают посланное узлом A *Prune*, и тут же один из них отправляет **сообщение *Join*** (второй, обнаружив в сети *Join,* не предпринимает никаких действий, поскольку одного такого сообщения достаточно). Приняв *Join* **маршрутизатор G игнорирует** предыдущий *Prune*. Если же за определенное время сообщение *Join* не будет принято, сеть N отрезается от дерева рассылки.

(2) Особая ситуация в PIM DM.

Вторая особая ситуация возникает, когда два маршрутизатора A и B подключены к одной и той же клиентской сети N, в которой находится получатель.

Оба маршрутизатора будут отправлять групповые дейтаграммы в сеть N, так как им известно, что в ней находится получатель.



Problem. Очевидно, что при этом создается избыточный трафик из лишних экземпляров дейтаграмм.

Solution PIM DM.

Во избежание этого эффекта маршрутизатор (предположим, A), обнаружив, что в сети N действует "конкурирующий" маршрутизатор B, также рассылающий групповые дейтаграммы от источника S в группу G, посылает **сообщение *Assert*** (declare a fact), содержащее расстояние от A до S. Конкурирующий узел B, получив это сообщение, сравнивает расстояние от себя до S с указанным в сообщении, и если свое расстояние больше, то соответствующий интерфейс отрезается от дерева с помощью *Prune*. Аналогичным образом посылается и обрабатывается *Assert* из B в A. При равных расстояниях побеждает маршрутизатор с бо'льшим IP-адресом.

4.4.2. PIM SM.

Протокол PIM SM (Protocol Independent Multicast, Sparse mode, RFC-2362) применяется для маршрутизации дейтаграмм для малочисленных групп, члены которых находятся далеко друг от друга (в этом случае недостатки метода PRPF с усечением становятся существенными).

Функционирование протокола можно кратко описать как **метод CBT, переходящий в RPF**. Вместо флудинга, как в PIM-DM, в PIM-SM выбирается один маршрутизатор, который будет хранить информацию о группах и источниках. Это обычный маршрутизатор PIM-SM, который выполняет роль места рандеву (**rendezvous point - RP**).

Все маршрутизаторы в домене PIM-SM должны знать, кто выполняет роль RP.

Маршрутизатор, в сети которого зарегистрировались члены группы, посылает в RP сообщение *Join*, которое обрабатывается промежуточными маршрутизаторами как в технологии CBT – таким образом формируется первоначальное дерево рассылки.

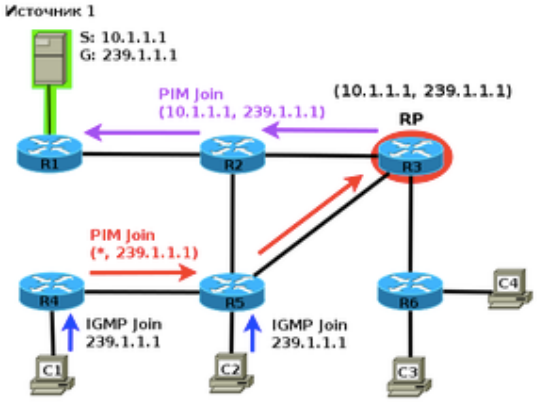
Отправитель дейтаграмм S (точнее, маршрутизатор), посылает в RP сообщения *Register*, в которых инкапсулируются групповые дейтаграммы. RP извлекает дейтаграммы из этих сообщений и рассылает их по сформированному дереву рассылки.

Оптимизация. Если отправитель работает достаточно интенсивно, то RP посылает в его сторону сообщение *Join* – то есть, отправитель становится членом группы и может рассылать групповые дейтаграммы по дереву непосредственно, минуя стадию туннелирования в точку рандеву.

Распространение групповых дейтаграмм по дереву рассылки осуществляется аналогично методу CBT:

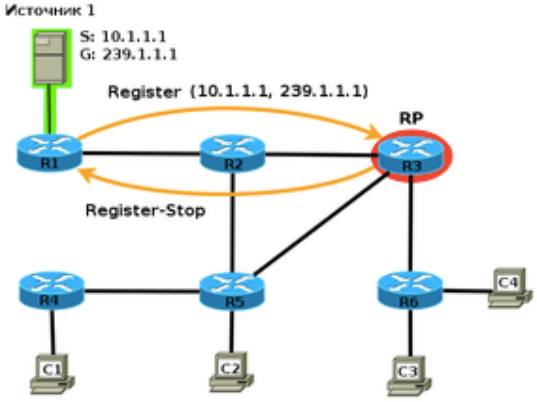
дейтаграмма рассылается через все интерфейсы, принадлежащие дереву.

Далее, получая дейтаграммы, адресованные группе, маршрутизатор может заметить, что интенсивность этого потока превышает некоторый установленный лимит. В этом случае маршрутизатор решает оптимизировать дерево рассылки. Он посылает сообщение *Join* к источнику *следующей* полученной им дейтаграммы, адресованной данной группе, а в точку рандеву посылается сообщение *Prune*. Таким образом, дерево, изначально созданное вокруг точки рандеву, оптимизируется для данного источника. (Только "конечные" маршрутизаторы дерева рассылки могут инициировать этот процесс.)



Следует отметить, что переход к дереву, оптимизированному для источника, приводит к необходимости хранить и обрабатывать на маршрутизаторах большее количество служебной информации, что не всегда приемлемо, поэтому существует возможность отключения такого перехода.

Предусмотрен также обратный переход к дереву с корнем в точке рандеву. Он производится, если оптимизация дерева оказалась неоправданной.



4.5. Заключение.

Применение того или иного протокола групповой маршрутизации существенно зависит от того, плотно или разреженно расположены получатели группового трафика.

Для плотного расположения получателей годятся протоколы DVMRP, MOSPF и PIM DM;

Для разреженного расположения получателей подходят протоколы PIM SM и CBT.

Все перечисленные протоколы находятся в экспериментальной стадии.

Протокол DVMRP, как указывалось выше, используется в ядре MBONE.

Наиболее перспективными выглядят протоколы PIM DM и PIM SM, они поддерживаются маршрутизаторами Cisco и Huawei.