МАРШРУТИЗАЦИЯ

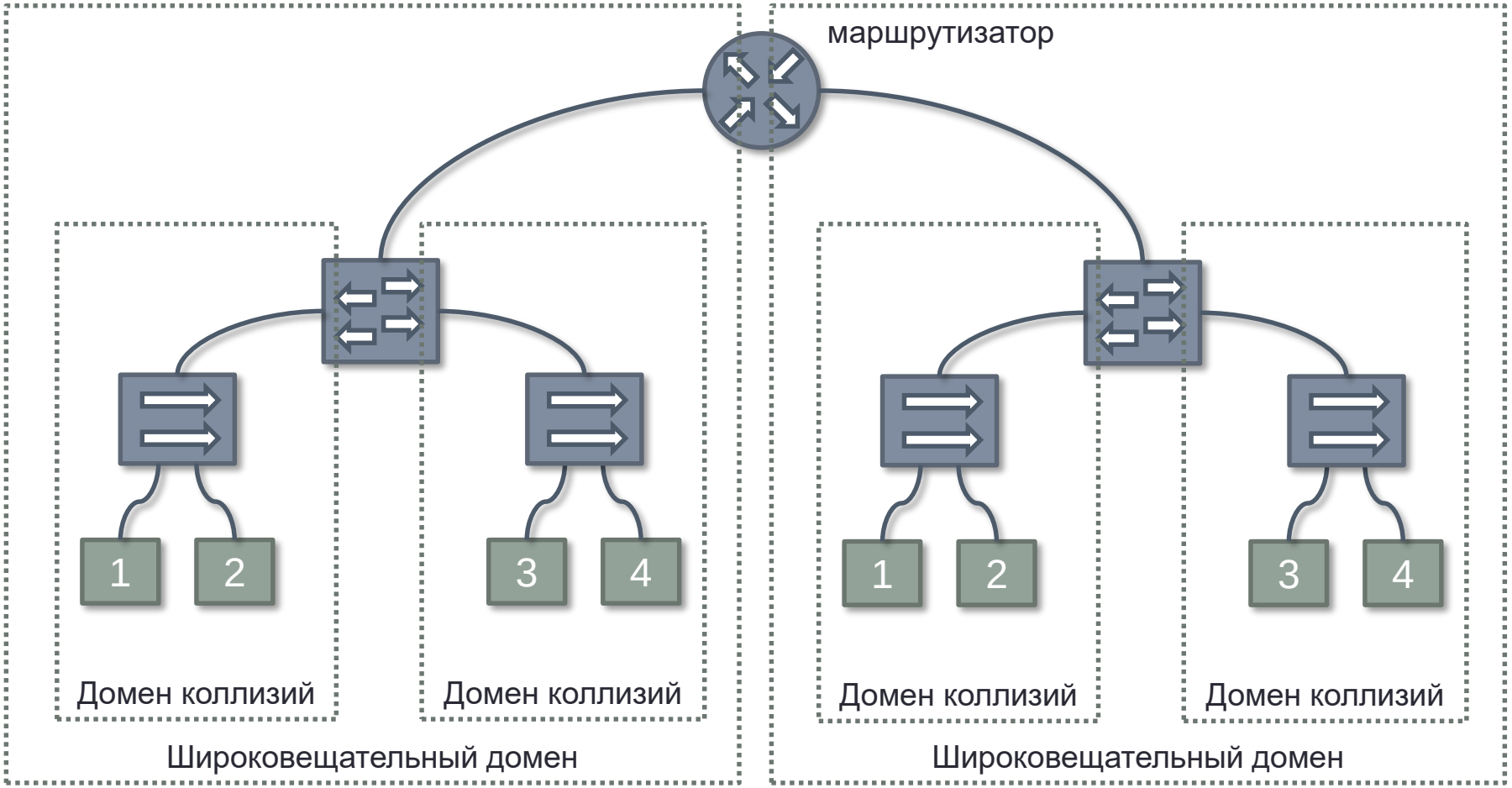
Ограничения сетей канального уровня

- Невозможность организации маршрутизации в большой сети на основе MAC-адресов.
- Сеть образует один широковещательный домен.
- Жесткие ограничения топологии сети.
- Затруднена организация взаимодействия сетей, построенных на основе разных технологий.
- Безопасность.

Терминология

- В протоколах сетевого уровня термин «сеть» или «локальная сеть», означает совокупность компьютеров, использующих для передачи пакетов общую технологию канального уровня и соединенных между собой в соответствии с одной из стандартных типовых топологий.
- В стандартах ISO вводятся понятия
  - Оконечная система / End System (ES) – отправители и получатели информации;
  - Промежуточная система / Intermediate System (IS) – транзитные узлы, обеспечивающие передачу информации

Домены коллизий и широковещательные домены



Задачи сетевого уровня

- Маршрутизация – построение маршрутов в сложной сети
  - Unicast
  - Multicast
  - Anycast

- Объединение разных сетей канального уровня
  - Разные технологии
  - Разные MAC адреса
  - Разные MTU

Решение

- Введение адресов и заголовка сетевого уровня, который бы сохранялся в неизменном виде при передаче данных через сеть.

Подходы к маршрутизации

- Одношаговая маршрутизация / hop-by-hop routing

- Маршрутизация отправителем / source specified routing

Маршрутизация отправителем

- Регион-отправитель диктует решения по продвижению сообщений маршрутизирующим объектам в каждом промежуточном регионе, которые в результате продвигают сообщения согласно требованиям отправителя.
- Каждый пакет должен содержать
  - Весь путь
  - Идентификатор пути
- Преимущества:
  - стратегия региона – отправителя действует независимо от наличия у промежуточных регионов вдоль маршрута предварительных сведений об этой стратегии;
  - метод свободен от зацикливаний маршрутов независимо от того, поддерживают все регионы вдоль маршрута согласующуюся маршрутную информацию, или нет.
- Недостатки:
  - исходный узел должен иметь полную информацию о топологии сети для выбора правильного маршрута.

Одношаговая маршрутизация

- Каждый маршрутизирующий объект принимает независимое решение по продвижению информации, основываясь на адресах отправителя, получателя, запрошенных услугах и на информации, содержащейся в базе данных продвижения информации данного объекта.
- Классы алгоритмов
  - алгоритмы фиксированной (или статической) маршрутизации;
  - алгоритмы простой маршрутизации;
  - алгоритмы динамической (или адаптивной) маршрутизации.

Таблицы маршрутизации

- Сохраняемая информация
  - адреса устройств или сетей;
  - адреса ближайших маршрутизаторов;
  - интерфейс для передачи данных;
  - служебная информация протокола маршрутизации:
    - время жизни;
    - стоимость;
    - …
- Виды таблиц маршрутизации
  - одномаршрутные;
  - многомаршрутные.

Статическая маршрутизация

- За построение маршрутов и оповещение IS отвечают администраторы.
- Преимущества:
  - минимален дополнительный траффик;
  - не расходуются ресурсы маршрутизатора на составление маршрутов.
- Недостатки:
  - невозможна автоматическая реакция на изменение топологии сети;
  - затруднено построение маршрутов для больших сетей.

Алгоритмы простой маршрутизации

- Таблица маршрутизации не используется или строится без использования протоколов маршрутизации.
- Алгоритмы:
  - Случайная маршрутизация – пакет посылается в случайном направлении (не совпадающим с исходным);
  - Заливка – пакет передаётся по всем возможным направлениям, кроме исходного;
  - Маршрутизация с учётом накопленного опыта – по аналогии с алгоритмом открытого моста.

Динамическая маршрутизация

- Таблица маршрутизации строится автоматически, с привлечением алгоритмов и протоколов маршрутизации.

- Подходы
  - централизованный;
  - распределённый.

- Требования к алгоритмам маршрутизации
  - оптимальность выбора маршрута;
  - простота реализации;
  - устойчивость;
  - быстрая сходимость.

Виды алгоритмов маршрутизации

- По назначению:
  - одномаршрутными или многомаршрутными;
  - одноуровневыми или многоуровневыми;
  - внутридоменными или междоменными;
  - одноадресными или групповыми.

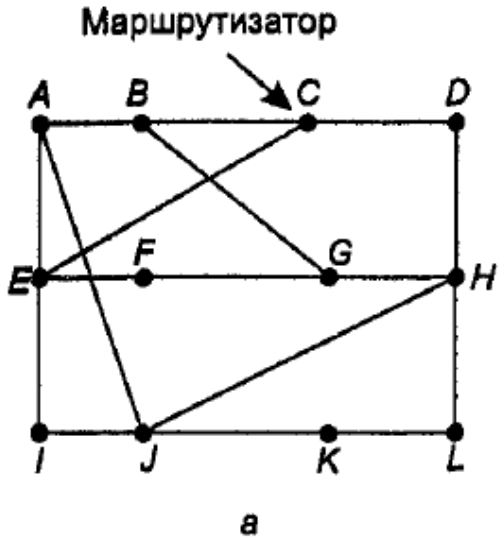
- По используемой информации:
  - Протоколы вектора расстояния;
  - Протоколы состояния канала;
  - Протоколы политики маршрутизации.

Протоколы вектора расстояния

- В конфигурацию роутера заносится информация о непосредственно подключенных к нему сетях
- Роутеры регулярно рассылают друг другу свои таблицы маршрутизации.
- Получив информацию от других маршрутизаторов, каждый корректирует свою таблицу маршрутизации, выбирая путь по алгоритму Беллмана-Форда.

Пример: RIP

Расчёт таблицы маршрутов алгоритмом вектора расстояний



а

Новая расчатная задержка для *J*

K	A	I	H	K	Линия	
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	8	K
L	29	33	9	9	15	K

Задержка *JA* равна 8 | Задержка *JI* равна 10 | Задаржка *JH* равна 12 | Задаржка *JK* раана 6

Векторы, полученныа от четырех соседей *J*

Новая таблица маршрутов для *J*

Алгоритм вектора расстояний. Добавление новой связи.

0		∞	∞	∞	∞

Алгоритм вектора расстояний. Исчезновение связи.

0		1	2	3	4

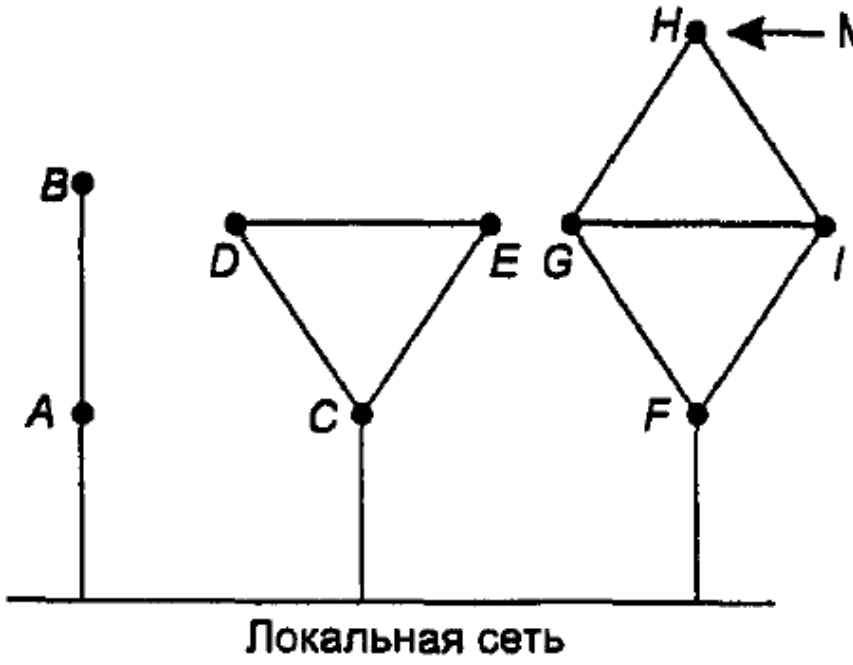
Алгоритм состояния каналов

- Маршрутизаторы находят соседей, и измеряют задержки передавая пакеты HELLO.
- Информация о подключенных соседях регулярно рассылается всем маршрутизаторам в виде пакетов Link State Advertisement (LSA).
- Каждый маршрутизатор строит граф сети.
- Оптимальный маршрут находится с помощью алгоритма Дейкстры.

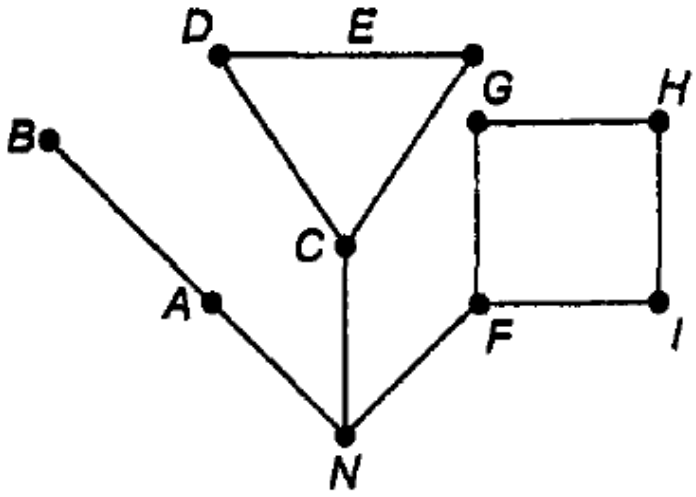
Пример: OSPF

Алгоритм состояния канала. Знакомство с соседями

Фрагмент сети



Графовая модель

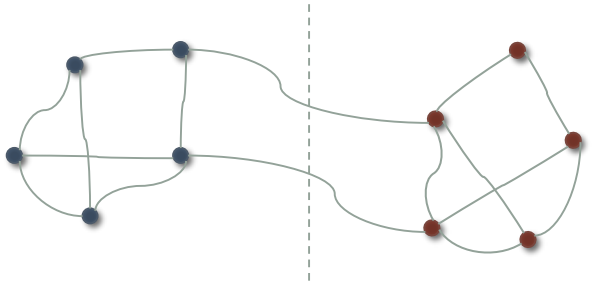


Алгоритм состояния канала. Измерение задержек.

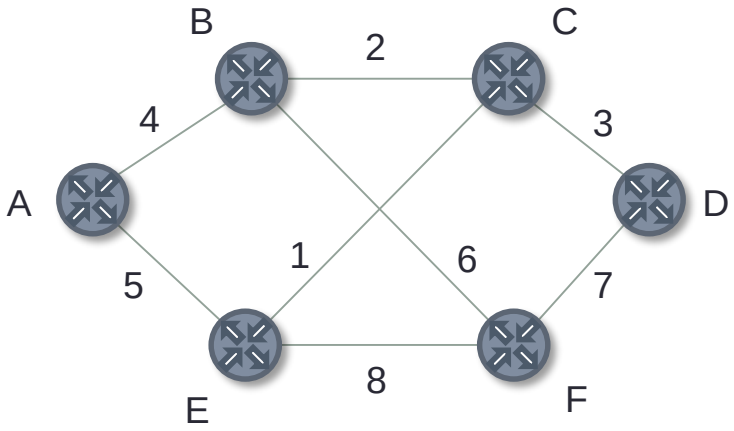
- Для измерения задержки мы можем использовать усреднённое значение времени двойного оборота пакета ECHO.
- Хотим ли мы учитывать нагрузку на линию?

  Нет
  - Таймер следует включать в момент отправки ECHO.

  Да
  - Таймер следует включать в момент помещения ECHO в очередь.
  - Может приводить к осцилляциям маршрутов.



Алгоритм состояния каналов. Создание пакетов состояния линий.



A	
Порядковый номер	
Возраст	
B	4
E	5

B	
Порядковый номер	
Возраст	
A	4
C	2
A	6

C	
Порядковый номер	
Возраст	
B	2
D	3
E	1

D	
Порядковый номер	
Возраст	
C	3
F	7

E	
Порядковый номер	
Возраст	
A	5
C	1
F	8

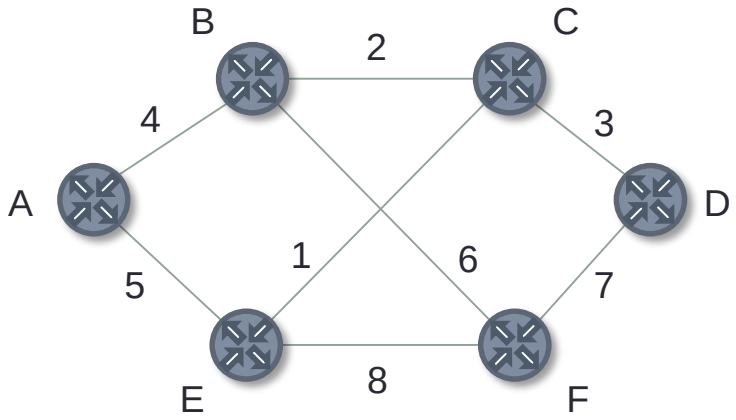
F	
Порядковый номер	
Возраст	
B	6
D	7
E	8

Алгоритм состояния каналов. Распространение пакетов состояния.

Используется алгоритм заливки

- Ограничение распространения пакетов
  - Каждый пакет имеет порядковый номер.
  - Маршрутизаторы запоминают пары {источник, порядковый номер}
  - Новые пакеты передаются дальше.
  - Дубликаты, и пакеты с порядковым номером меньше последнего удаляются.
- Защита от потери номеров при перезагрузке или искажениях пакетов
  - Вводится поле возраста пакета.
  - Возраст уменьшается на 1 каждую секунду и при передаче пакета каждым маршрутизатором.
  - При достижении 0 данные о пакете уничтожаются.
- Удаление дублей и отправка подтверждений.
  - Перед обработкой пакет некоторое время храниться для ожидания более новых / выявления дублей.
  - По окончании хранения высылаются подтверждения о получении пакетов.

Алгоритм состояния каналов. Распространение пакетов состояния.



Буфер маршрутизатора B

Источник	Номер	Возраст	Отослать			Подтвердить			Данные
			A	C	F	A	C	F	
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

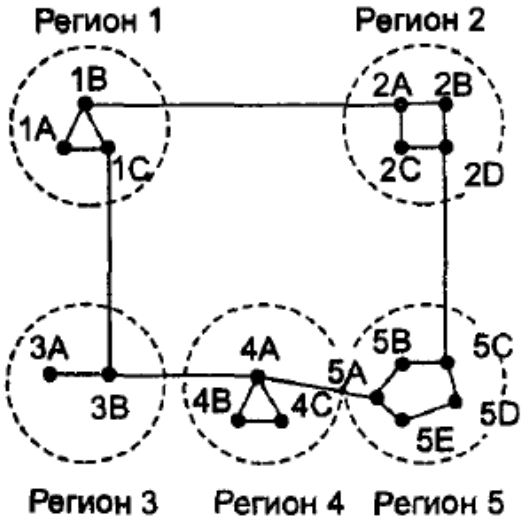
Алгоритм состояния каналов. Вычисление оптимальных маршрутов.

Используется алгоритм Дейкстры.

- Инициализация.
  - Для каждой вершины запомнить:
    - L - длина кратчайшего известного пути (0 для начальной, ∞ для остальных)
    - Признак обработана или нет.

- Алгоритм
  - Пока есть необработанные вершины выбрать из них вершину A с минимальным значением длины пути
    - Для всех ребер AB, ведущих в необработанную вершину B.
      - Если L(B)>L(A)+L(AB) то L(B)=L(A)+L(AB)
    - Пометить A как обработанную.

Иерархическая маршрутизация



а

Полная таблица для 1A

Назначение	Линия	Транзитные участки
1A	–	–
1B	1B	1
1C	1C	1
2A	1B	2
2B	1B	3
2C	1B	3
2D	1B	4
3A	1C	3
3B	1C	2
4A	1C	3
4B	1C	4
4C	1C	4
5A	1C	4
5B	1C	5
5C	1B	5
5D	1C	6
5E	1C	5

б

Иерархическая таблица для 1A

Назначение	Линия	Транзитные участки
1A	–	–
1B	1B	1
1C	1C	1
2	1B	2
3	1C	3
4	1C	3
5	1C	4

в

Широковещательная маршрутизация

Доставка сообщения всем получателям.

- Отправка отдельных сообщений каждому.
  - Источник должен знать адреса всех получателей.
  - Большая нагрузка на сеть.

Широковещательная маршрутизация

Доставка сообщения всем получателям.

- Отправка отдельных сообщений каждому.
- Метод заливки.
  - Большая нагрузка на сеть.
  - Необходимы механизмы регулирования заливки.

Широковещательная маршрутизация

Доставка сообщения всем получателям.

- Отправка отдельных сообщений каждому.
- Метод заливки.
- Многоадресная маршрутизация.
  - В пакете указывается список адресов получателей.
  - Маршрутизатор разделяет множество получателей согласно интерфейсам, на которые он должен передать пакеты.
  - Сокращает нагрузку на сеть.

Широковещательная маршрутизация

Доставка сообщения всем получателям.

- Отправка отдельных сообщений каждому.
- Метод заливки.
- Многоадресная маршрутизация.
- Использование минимального остовного или любого другого дерева.
  - В дереве не должно быть петель.
  - Маршрутизатор передаёт пакет только по ветвям дерева.
  - Необходимо наличие информации о дереве у маршрутизатора.

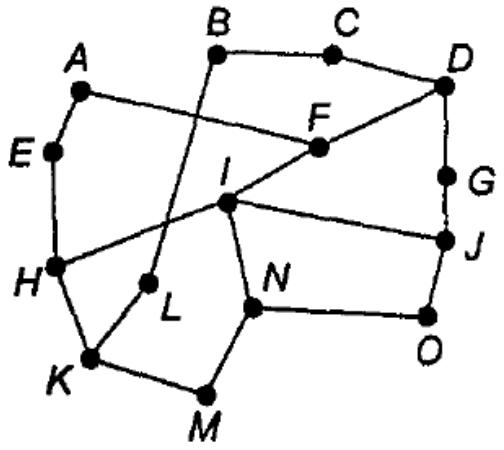
Широковещательная маршрутизация

Доставка сообщения всем получателям.

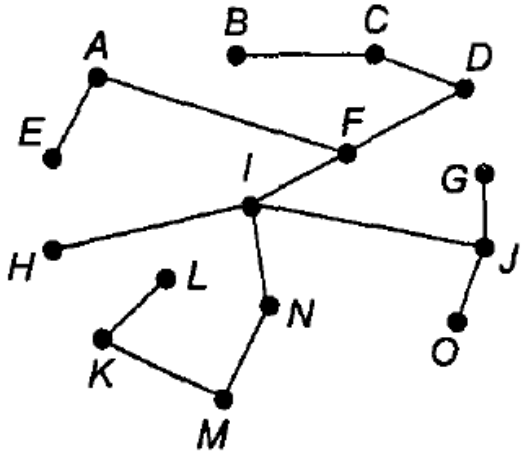
- Отправка отдельных сообщений каждому.
- Метод заливки.
- Многоадресная маршрутизация.
- Использование минимального остовного или любого другого дерева.
- Продвижение по встречному пути.
  - При получении широковещательного пакета маршрутизатор сравнивает интерфейс получения с оптимальным путём к источнику.
    - Совпадают – это первый пакет, передаём по всем интерфейсам.
    - Не совпадают – дубликат, игнорируем.

Широковещательная маршрутизация. Продвижение по встречному пути.

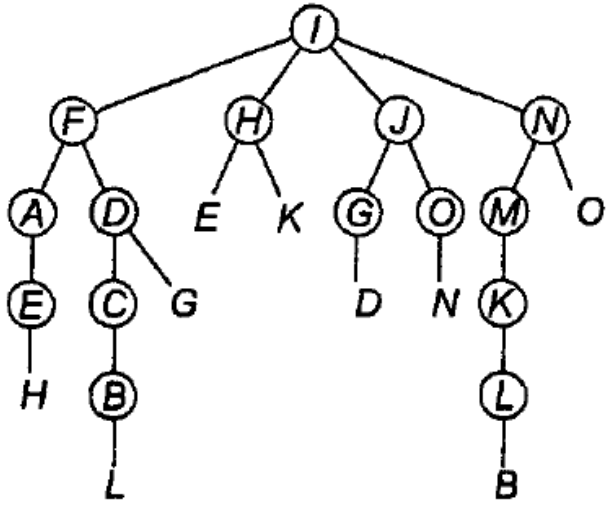
Сеть



Связующее дерево.



Продвижение по встречному пути.

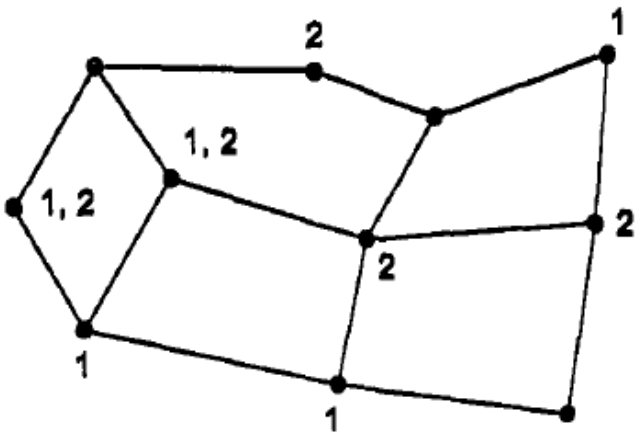


Многоадресная маршрутизация

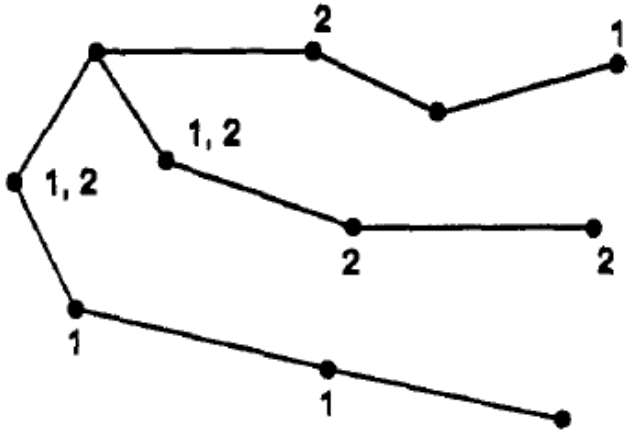
Сообщение следует доставить группе получателей.

- Хосты сообщают маршрутизаторам о принадлежности к той или иной группе.
- Маршрутизаторы распространяют информацию о принадлежности группам между собой.
- Каждый маршрутизатор рассчитывает связующее дерево, покрывающее все остальные маршрутизаторы сети.
- При отправке сообщения выполняется усечение дерева – удаляются рёбра, не ведущие к хостам, являющимися членами группы. Рассылка производится по рёбрам усечённого дерева.

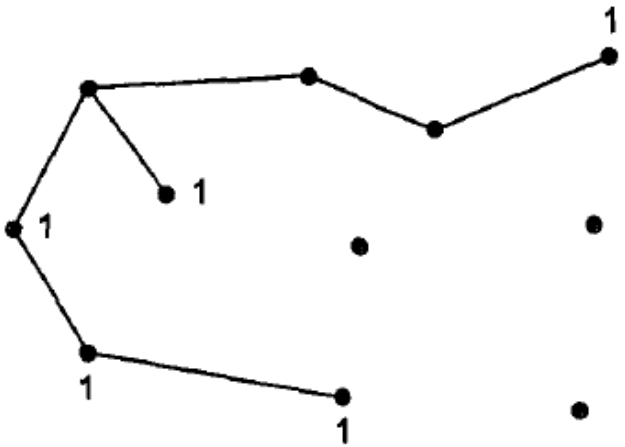
Многоадресная маршрутизация



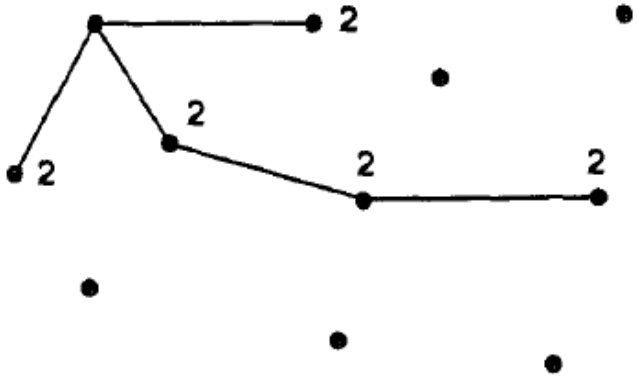
а



б



в



г

Многоадресная маршрутизация

Алгоритмы усечения:

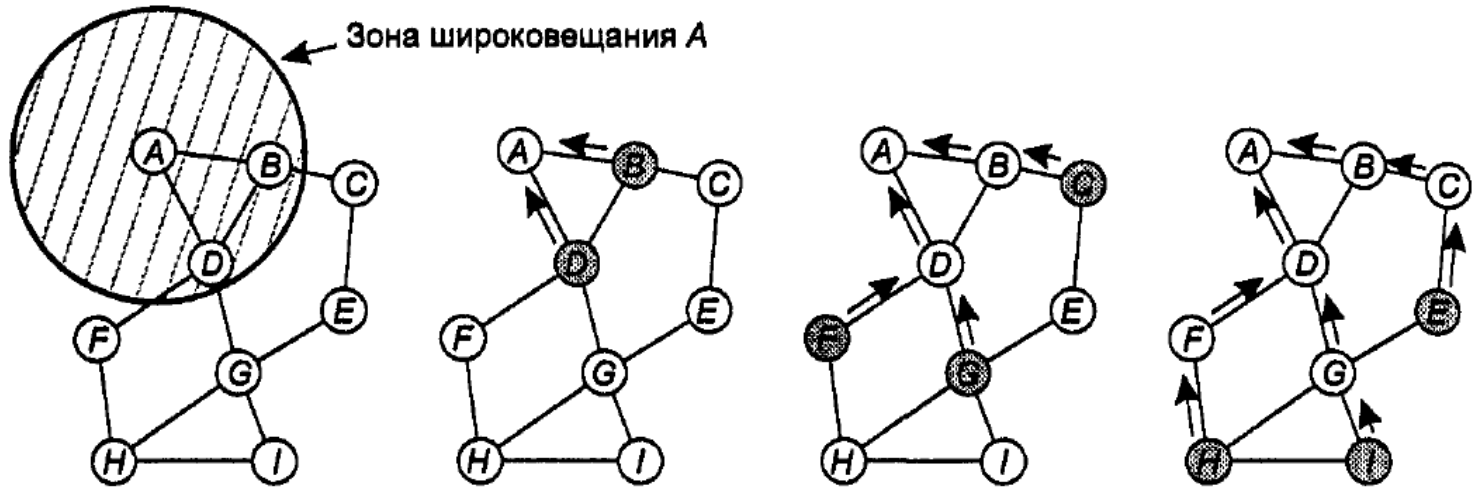
- При использовании алгоритмов состояния каналов
  - Каждый маршрутизатор знает весь граф сети.
  - Удаляем маршрутизаторы, не входящие в рассматриваемую группу начиная с листьев дерева в направлении корня.
- При использовании алгоритмов вектора расстояний
  - Алгоритм продвижения по встречному пути.
  - Маршрутизатор, не входящий в группу, может послать сообщение PRUNE (отсечь), если
    - У него нет других связей.
    - Он получил многоадресное сообщение по всем своим линиям.
- Деревья с основанием в сердцевине (Ballardie, 1993)
  - Для каждой группы рассчитывается единое связующее дерево, с корнем (ядром) около середины группы.
  - Сообщение посылается ядру, а ядро выполняет рассылку по дереву.
  - Требует хранение только одного экземпляра дерева.

Маршрутизация в мобильных сетях Mobile Ad hoc network

- Узлы сети могут появляться и исчезать в произвольные моменты.
- Меняется предпочтительность и реализуемость сетей.
- Ограниченная пропускная способность.

- Алгоритм Ad hoc On-demand Distance Vector / AODV

Алгоритм Ad hoc On-demand Distance Vector. Поиск пути.

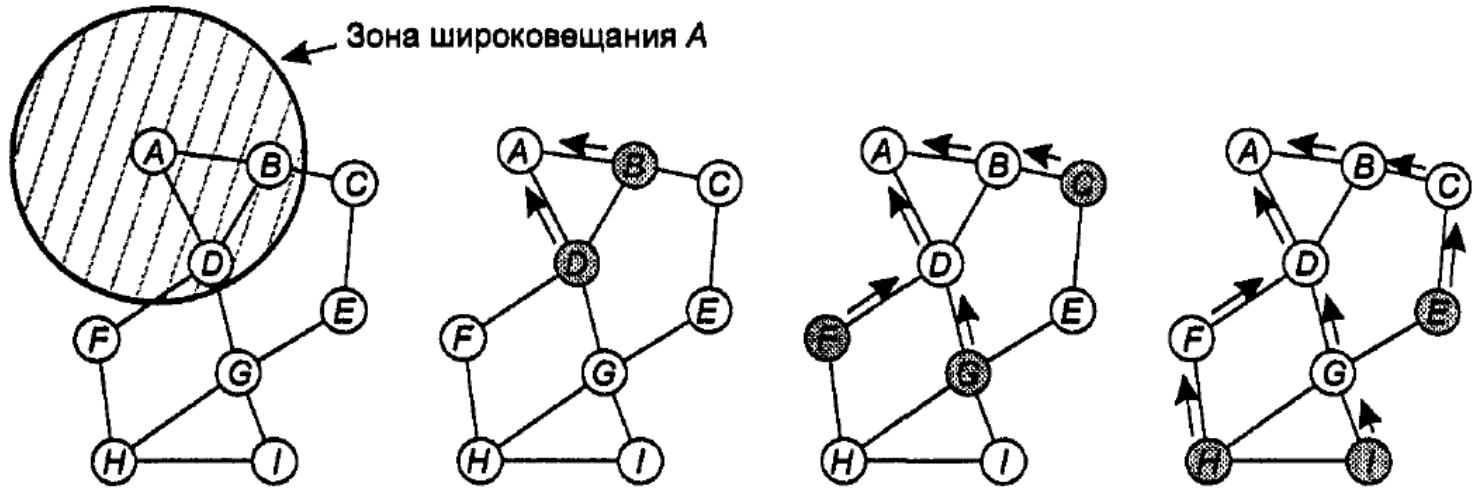


- Отправитель (A) генерирует пакет Route Request

Адрес отправителя	Идентификатор запроса	Адрес получателя	Порядковый номер отправителя	Порядковый номер получателя	Счётчик переходов

- Пакет идентифицируется парой {адрес отправителя, идентификатор запроса}.

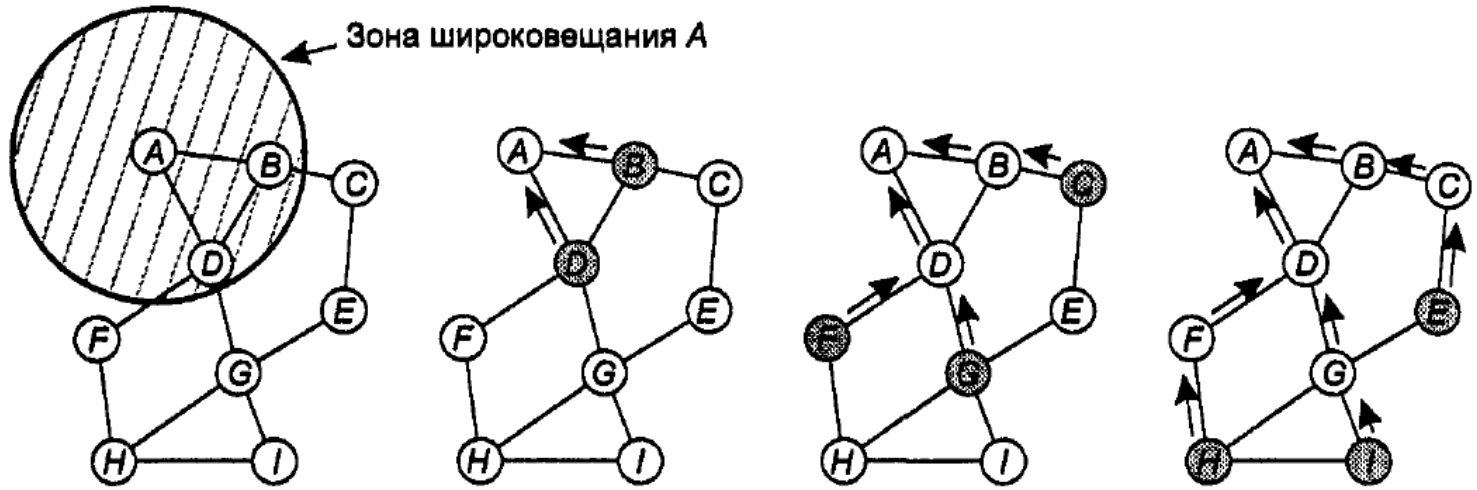
Алгоритм Ad hoc On-demand Distance Vector. Поиск пути.



Обработка пакета Route Request

- Удаляем пакет если это дубликат.
- Ищем у себя запись с порядковым номером получателя ≥ порядковому номеру получателя в пакете.
  - Нашли – отвечаем отправителю пакетом Route Reply
  - Не нашли – запоминаем откуда пришел пакет и рассылаем его дальше.
  - Запомненные обратные пути удаляются после таймаута.

Алгоритм Ad hoc On-demand Distance Vector. Поиск пути.

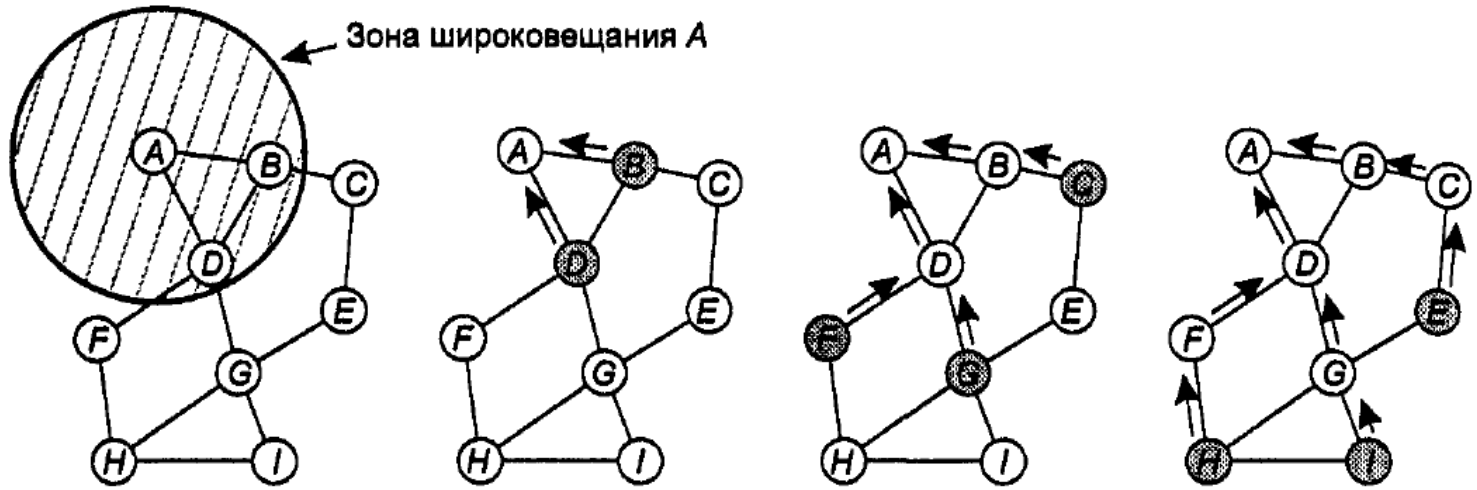


Пакет Route Reply

Адрес отправителя	Идентификатор запроса	Порядковый номер получателя	Счётчик переходов	Время жизни

- Пересылается одноадресно навстречу пакету Route Request

Алгоритм Ad hoc On-demand Distance Vector. Поиск пути.



Обработка пакета Route Reply

- Добавляем запись в таблицу маршрутизации если выполнено хотя бы одно условие:
  - У нас не было записей для получателя
  - Новый номер получателя больше, чем тот, который был у нас.
  - Номера равны, но новый путь короче.
- Передаём пакет в запомненном ранее направлении.

Алгоритм Ad hoc On-demand Distance Vector. Обслуживание маршрута.

- Узлы периодически рассылают соседям пакеты Hello, на которые те должны отвечать.
- Для каждого адресата храним список активных соседей – соседей, которые снабжали нас пакетами для этого адресата в течении последних ΔT секунд.
- При обнаружении пропажи соседа
  - проверяем и удаляем из таблицы маршрутизации проходившие через него пути;
  - сообщаем оставшимся активным соседям по этим путям об их уничтожении.

Алгоритм Ad hoc On-demand Distance Vector. Обслуживание маршрута.

Сеть до удаления узла G

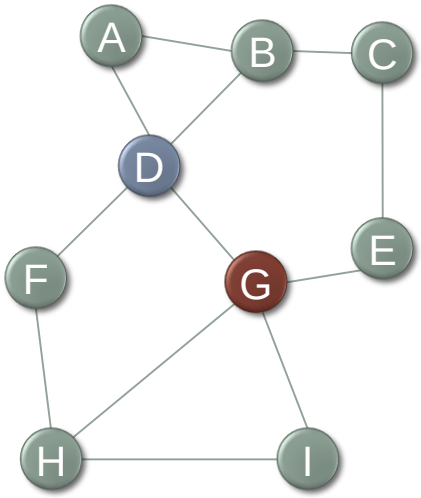
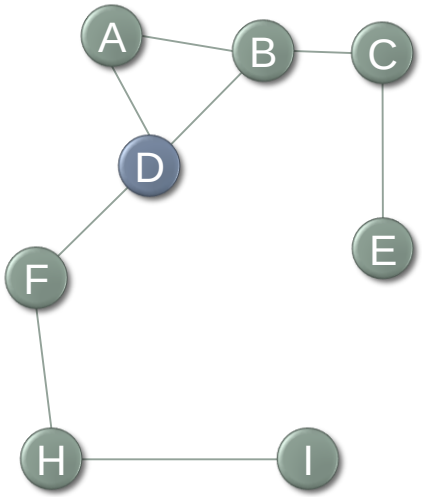


Таблица маршрутизации узла D до удаления узла G

Адресат	Следующий переход	Расстояние	Активные соседи
A	A	1	F, G
B	B	1	F, G
C	B	2	F
E	G	2	
F	F	1	A, B
G	G	1	A, B
H	F	2	A, B
I	G	2	A, B

Сеть после удаления узла G

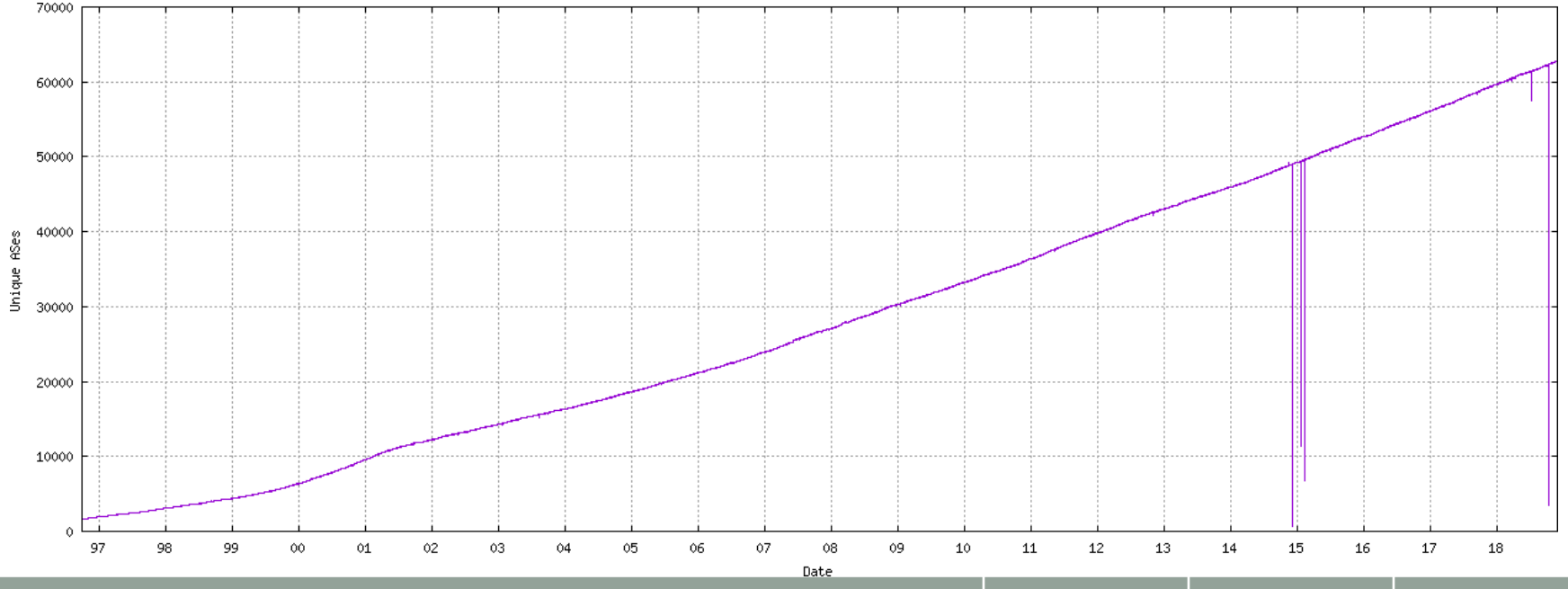


МАРШРУТИЗАЦИЯ В INTERNET

Автономная система
Autonomous System (AS)

- Автономная система – группа из одного или нескольких сетевых префиксов, управляемая одним или несколькими сетевыми операторами, имеющая единую и ясно определённую политику маршрутизации (RFC 1930).
- AS идентифицируется глобально уникальным номером (Autonomous System Number, ASN), который также используется для обмена информацией о маршрутизации.
- Номера AS выделяются IANA через региональных регистраторов.
  - До 2007 года 16 бит (диапазон 1-54271).
  - В 2007 году RFC 4893 введены 32-битные номера. Эти номера записываются в виде одного числа, или в виде x.y.

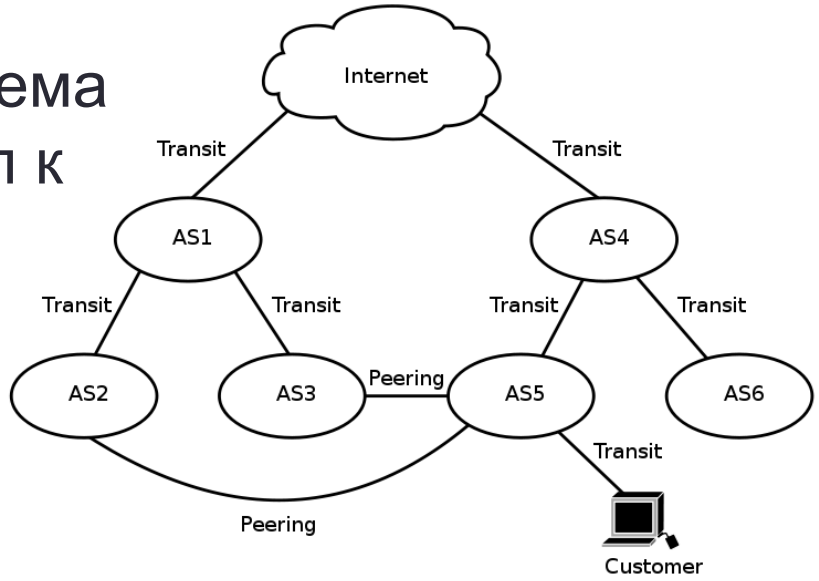
Автономная система
Число автономных систем



	10.2011	11.2015	11.2018
Число автономных систем	39474	52194	62865
Число AS, анонсирующих только один префикс	16627	20638	23199
Наибольшее число префиксов, анонсированных AS: AS8151: Uninet S.A. de C.V., MX	3483	5612	5445
Наибольший диапазон адресов, анонсированный AS AS4134: CHINANET-BACKBONE	108833792	121006848	116652288

Отношения между AS

- Транзит – система оплачивает другой сети трафик для доступа к интернету.
- Пиринг – системы обмениваются трафиком бесплатно, ради взаимной выгоды.
- Пользователь – другая система платит нам деньги за доступ к интернет.

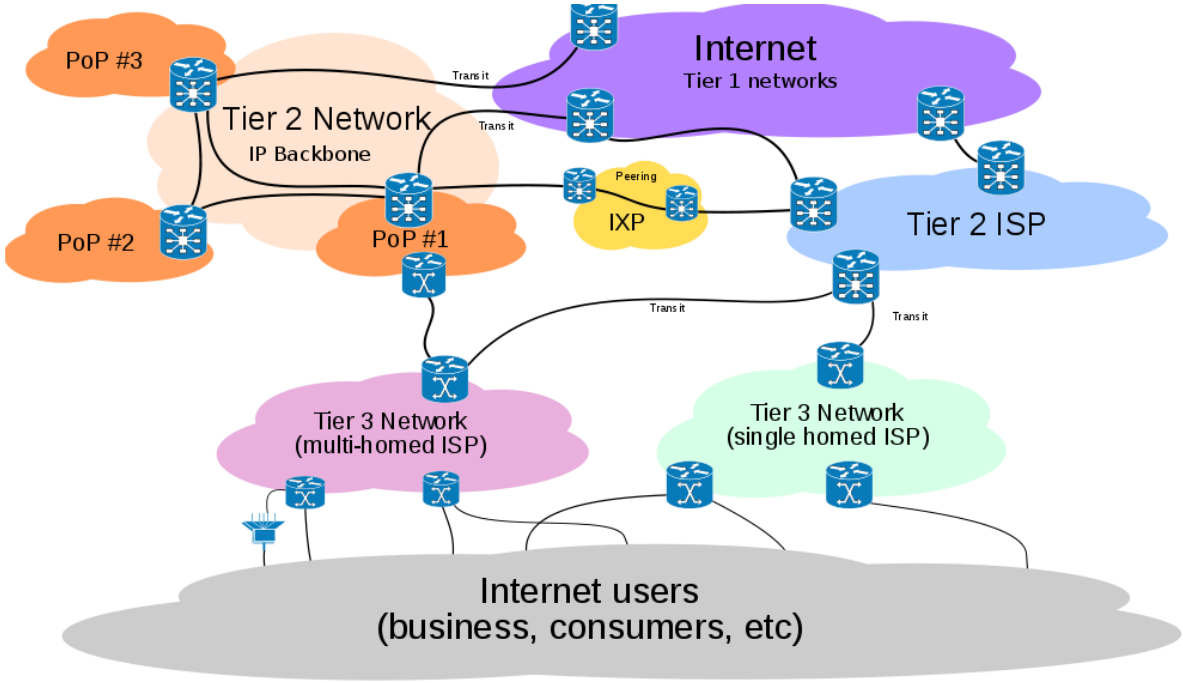


Иерархия IP операторов

- Tier 1 — оператор, который имеет доступ к Интернету исключительно через пиринговые соединения.

- Tier 2 — оператор, который имеет доступ к части Интернета через пиринговые соединения, но покупает IP транзит для доступа к остальной части Интернета.

- Tier 3 — оператор, который для доступа к Интернету использует исключительно каналы, которые покупает у других операторов.



Пиринг

Пиринг подразумевает:

- физическое соединение сетей;
- обмен маршрутной информацией с помощью протокола BGP;
- заключение соглашения о пиринге.

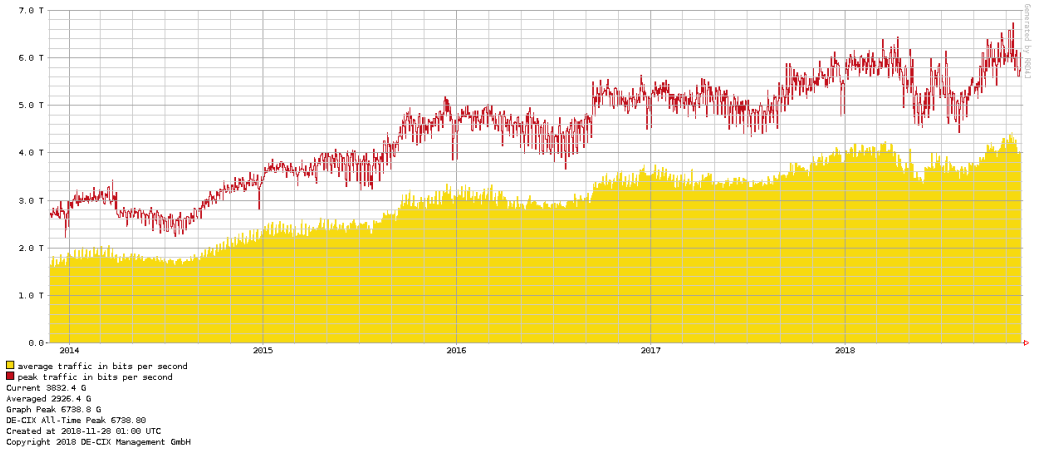
Типы пиринга

- Частное соединение по схеме точка-точка между сетями
- Публичное соединение через точку обмена трафиком
  - Network Access Point (NAP) – устаревшее название
  - Exchange Point, Internet Exchange (IXP, IX)

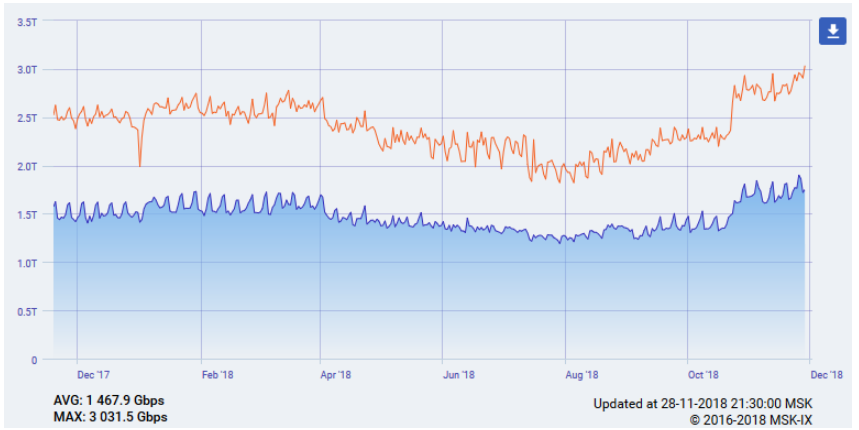
Крупнейшие IXP

		Название	Город	Страна	Основана	Участники	Макс. трафик (Gbit/s)	Средний трафик (Gbit/s)
1	DE-CIX	Deutscher Commercial Internet Exchange	Frankfurt, Hamburg, Munich, New York, Dubai, Palermo…	Германия	1995	735	6408	4004
2	IX.br	Brazil Internet Exchange	Много…	Бразилия	2004	3252	6120	3940
3	AMS-IX	Amsterdam Internet Exchange	Amsterdam, Haarlem, Schiphol-Rijk	Нидерланды	1997	818	5513	3339
4	LINX	London Internet Exchange	London, Manchester, Edinburgh, North Virginia, …	Англия	1994	825	4340	2850
5	MSK-IX	Moscow Internet Exchange	Москва, С. Пб, Новосибирск, Ростов на Дону…	Россия	1995	504	2951	1462
6	DATA-IX	Data-IX	Москва, С.Пб., Новосибирск, Самара, Уфа, Пермь …	Россия, Украина, Казахстан, Германия	2009	372	2726	

DE-CIX Франкфурт 5 лет



MSK-IX 1 год



Пиринг в Интернет

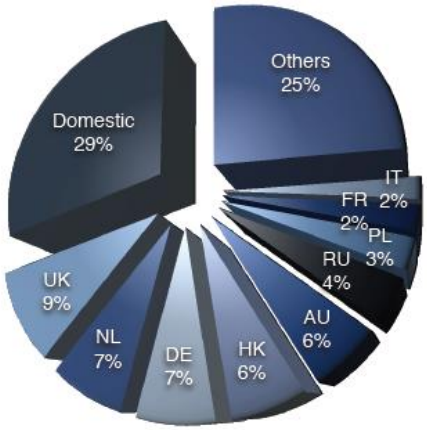
Survey of Characteristics of Internet Carrier Interconnection Agreements.

Bill Woodcock & Vijay Adhikari, Packet Clearing House, May 2011.

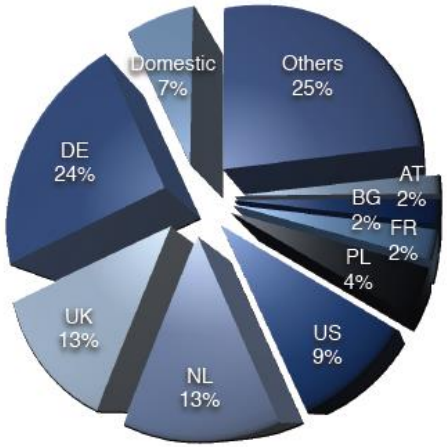
- 142210 соглашений между 4331 из 5039 (86%) операторов, из них 466 из США, 337 из России.

Степень формальности	
Устные соглашения	**Договоры на бумаге**
141512 (99.51%)	698 (0.49%)
Симметричность соглашений	
Симметричные	**Несимметричные**
141836 (99.73%)	374 (0.27%)

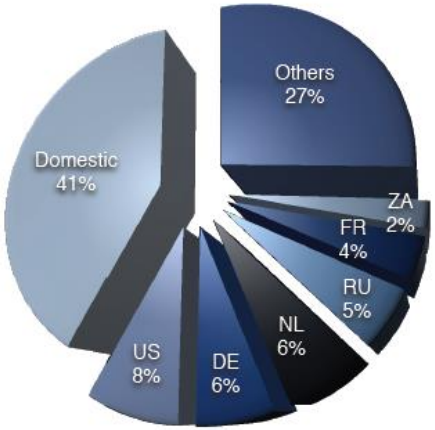
Пиринг в Интернет. Выбор партнёров.



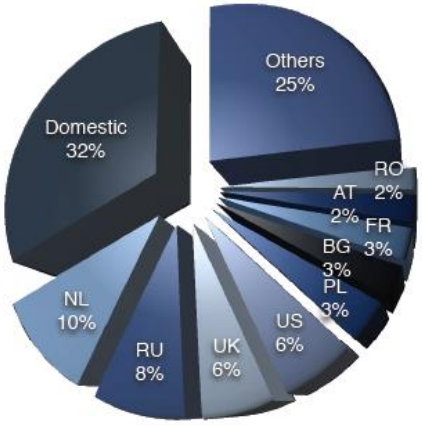
United States



Russia



United Kingdom

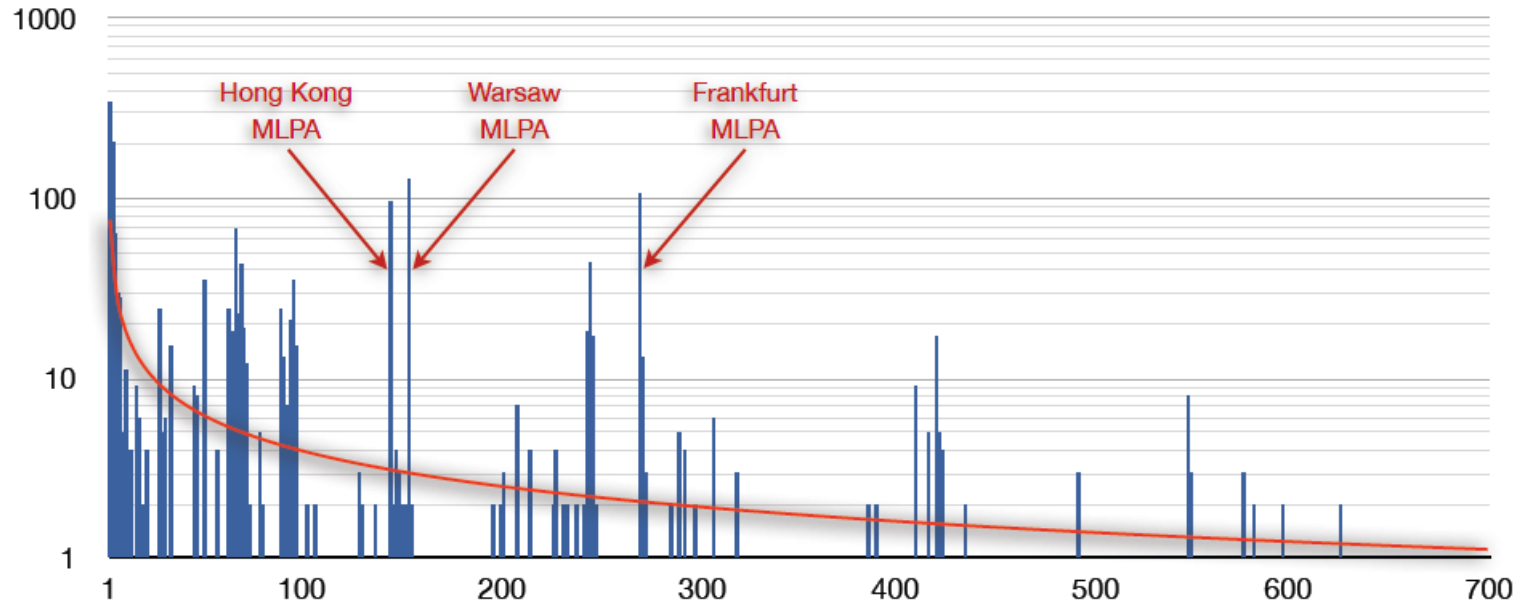


Germany

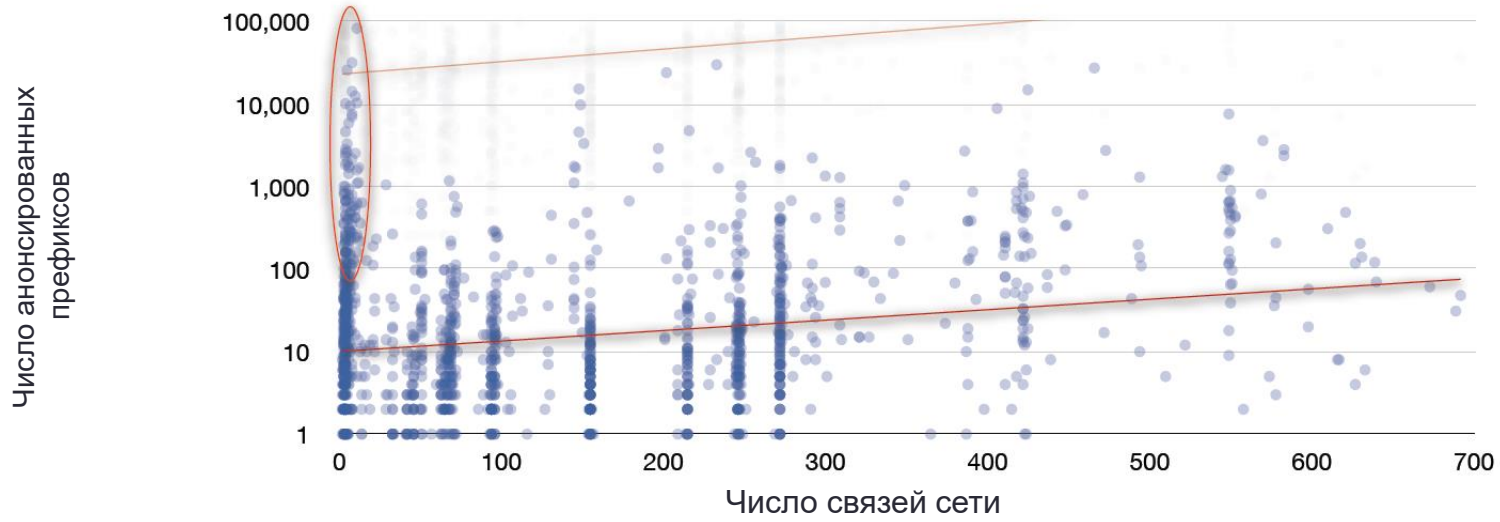
Пиринг в Интернет. Степень связанности.

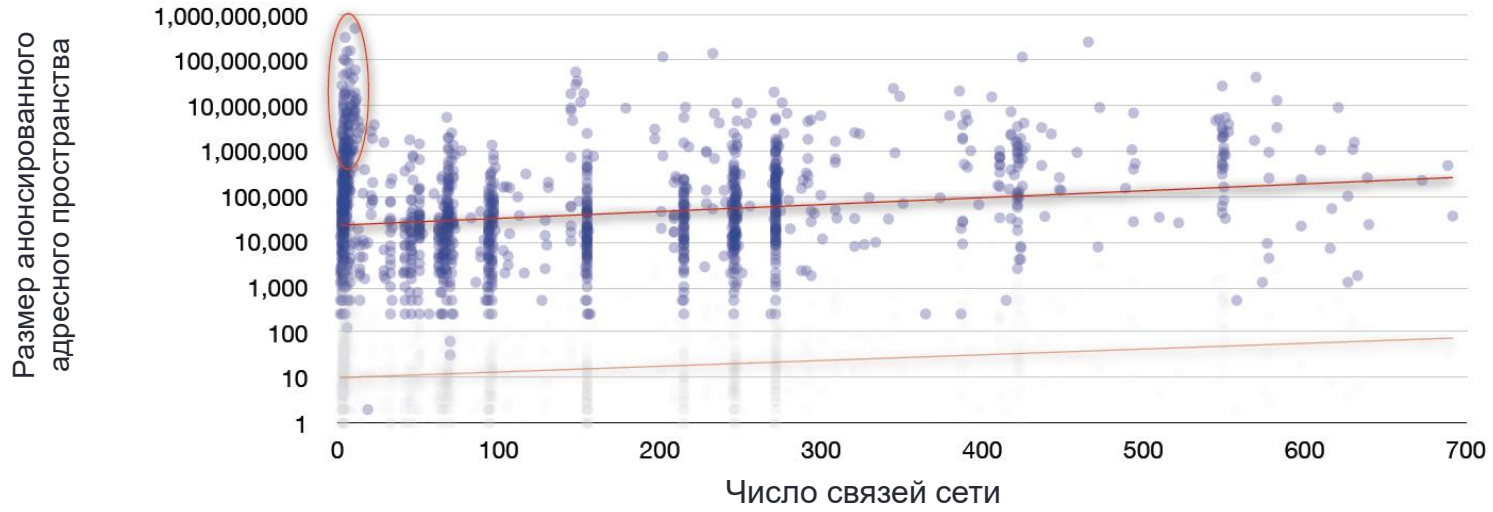
- 2696 (62%) сетей имеют не более 10 партнёров
- 12 сетей имеют более 700 партнёров

Гистограмма распределения числа связей между сетями



Пиринг в Интернет. Степень связанности.





Default-free Zone

- Default-free Zone (DFZ) – множество автономных систем, которым не требуется маршрут по умолчанию для доставки пакета в любую точку сети Internet.
- Можно сказать, что роутеры DFZ имеют полную BGP таблицу, описывающую все AS в Internet.
  - Отражаются только маршруты, доступные для транзита.
  - Выглядит по разному для разных маршрутизаторов.
  - Постоянно изменяется.
- С точки зрения маршрутизатора Asia-Pacific Network Information Center (APNIC), подключенного к DIX-IE, Токио на 02.11.2015 (28.11.2012):
  - 570441 маршрут от 51911 AS (433341 маршрут от 42692 AS)
  - Средняя длина пути 4.5 AS (4.6 AS)
  - Максимальная длина пути 50 AS (29 AS)
  - Анонсировано адресов 2802552256 (2615211716)

Множественная адресация Multihoming

- Одно соединение с Интернет, несколько префиксов
- Несколько интерфейсов, каждый с одним адресом
- Несколько соединений с Интернет, один набор префиксов
- Несколько соединений с Интернет, каждый со своим префиксом

ПРОТОКОЛЫ МАРШРУТИЗАЦИИ, ИСПОЛЬЗУЕМЫЕ В INTERNET

Виды протоколов маршрутизации

- Протоколы внутридоменной маршрутизации / Interior Gateway Protocol (IGP)
  - Используются внутри автономных систем или их частей
- Протокол междоменной маршрутизации
  - Раньше Exterior Gateway Protocol (EGP)
  - Сейчас Border Gateway Protocol (BGP)

Route Information Protocol RIP

- Протокол вектора расстояний.
- В качестве метрики используется число хопов.

- Первые реализации: Xerox XNS, BSD Unix
- 1988 г. – RFC 1058 RIP версия 1
- 1993…1998 г. – RFC 2453 RIP версия 2
  - Поддерживает CIDR
  - Совместим с RIPv1 если в последнем обнулены все положенные поля.
  - Для передачи сообщений используется мультикаст, а не броадкаст.
- 1997 г. – RFC 2080 RIPng
  - Поддержка IPv6

Формат пакетов RIPv2

Заголовок пакета

Поле	Команда	Версия (2)	**Домен маршрутизации**
Размер	1	1	2

- Команда:
  - 1 – запрос таблицы маршрутизации;
  - 2 – ответ (содержит таблицу маршрутизации).
- Домен маршрутизации – номер, позволяющий использовать несколько различных групп маршрутизаторов RIP в одной сети.

- Данные пакета (до 25 записей)

Поле	AFI	**Тэг**	IP адрес	**Маска**	**Следующий хоп**	Метрика
Размер	2	2	4	4	4	4

- Address Family Indicator (AFI) – тип маршрутизируемого протокола, 2 для IP
- Тэг – номер AS для данных, поступивших из BGP
- Метрика – стоимость маршрута

Таймеры

- Таймер обновления
  - Каждые 30 секунд система должна разослать ответы всем, кто прислал запрос.
  - 30 секундный таймер должен не зависеть от нагрузки на систему, или должен быть дополнен случайным смещением для предотвращения синхронизации времени обновления всех систем.
- Таймер таймаута
  - Если после внесения маршрута в таблицу проходит более 180 секунд такой маршрут считается недействительным.
  - Метрика маршрута устанавливается в 16 (бесконечность)
  - Недействительный маршрут сохраняется для передачи информации окружающим.
- Таймер удаления
  - Недействительные маршруты удаляются по истечению 120 секунд.

Защита от некорректной информации

- Ассиметричное распространение с обратным исправлением / Split horizon with poisoned reverse
  - При передаче информации узлу, от которого она была исходно получена, выставляется метрика 16.
- Экстренное обновление
  - При изменении маршрутной метрики маршрутизатор должен разослать обновление немедленно, не дожидаясь 30 секунд.
- Временный отказ от приёма / Holddown timer
  - При получении информации о недоступности сети включается таймер, в течение работы которого сообщения о доступности этой сети будут игнорироваться.

Недостатки RIP

- Максимальная длина маршрута 15
- Использование фиксированной метрики
- Высокий уровень трафика при обновлениях таблиц
- Медленное согласование
- Отсутствие поддержки динамического распределения маршрутов

- Простой протокол, который подходит для небольших сетей.

Open Shortest Path First (OSPF)

- Алгоритм состояния каналов.
- Используется алгоритм Дейкстры для нахождения кратчайших путей

- 1989 г.: RFC 1131 OSPF
- 1991 г.: RFC 1247 … 1998 г.: RFC 2328 OSPF version 2
- 2008 г.: RFC 5340 OSPF version 3
  - Поддержка IPv6

OSPF

- База данных состояния каналов / Link State Database (LSDB) – карта топологии сети, поддерживаемая каждым маршрутизатором.
- Маршруты строятся на основе метрик / metric - условный показатель «стоимости» пересылки данных по каналу.
- Объявление о состоянии канала / Link-State Advertisement, LSA) — объявление описывает все интерфейсы / interface (или каналы / link) маршрутизатора и состояние каналов.

OSPF - Соседи

- Соседи / neighbour – маршрутизаторы, подключенные к одному широковещательному домену.
- Поиск соседей осуществляется с помощью отправки широковещательных пакетов Hello.
- Состояние смежности / adjacency — взаимосвязь между определёнными соседними маршрутизаторами установленная с целью обмена информацией маршрутизации.
- База данных соседей / adjacency database – список всех соседей.
- В случае, если в широковещательном домене имеется больше двух маршрутизаторов, они выбирают выделенный маршрутизатор / Designated Router (DR), который управляет рассылкой LSA в сети.
- Для уменьшения задержки в случае отказа DR выбирается резервный выделенный маршрутизатор / Backup Designated Router (BDR).

OSPF – Зоны / Area

- Сеть, управляемая маршрутизаторами OSPF может быть разделена на зоны / Area, идентифицируемые цифровым идентификатором.
- Для каждой зоны поддерживается своя LSDB, состояние которой агрегируется граничным маршрутизатором при передаче в другие зоны.

- Виды зон
  - Магистральная зона / backbone area (идентификатор 0)

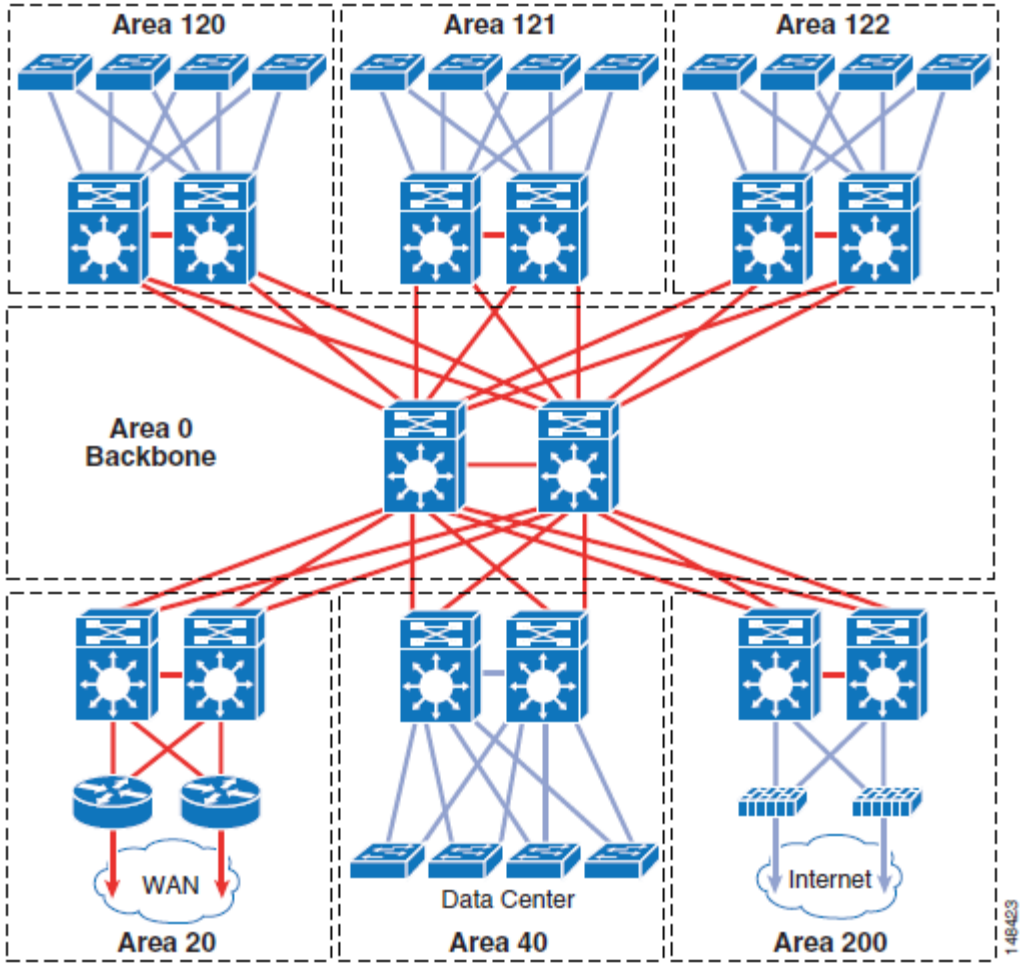
    Ответственна за распространение маршрутизирующей информации между немагистральными зонами. Магистральная зона должна быть смежной с другими зонами, но она не обязательно должна быть физически смежной

  - Тупиковая зона / stub area

    Для выхода за границы зоны используются маршруты по умолчанию.
    - Not-so-stubby area (NSSA)

      Получает информацию о маршрутах внутри AS, но не за её пределами.

Зоны OSPF


Area 120
Area 121
Area 122
Area 0
Backbone
WAN
Area 20
Data Center
Area 40
Internet
Area 200
148423

OSPF – виды маршрутизаторов

- Пограничный маршрутизатор / Area border router (ABR)

  Маршрутизатор соединяющий одну или более зон с магистральной зоной.
- Пограничный маршрутизатор AS / Autonomous system boundary router (ASBR)

  Маршрутизатор использующий более одного протокола маршрутизации (BGP, статическая маршрутизация).
- Внутренний маршрутизатор / Internal router (IR)

  Маршрутизатор, все интерфейсы которого находятся в одной зоне.
- Магистральный маршрутизатор / Backbone router (BR)

  Маршрутизатор, подключенный к магистрали OSPF
  - ABR
  - Внутренние роутеры магистрали

OSPF – Состояния протокола

1. Down State
   - Обмен данными между маршрутизаторами не происходит.

OSPF – Состояния протокола

1. Down State
2. Init State
   - Отправили Hello соседу и ожидаем ответа.

OSPF – Состояния протокола

1. Down State
2. Init State
3. Two-way – состояние двухсторонней связи
   - Если маршрутизатор видит себя в списке соседних маршрутизаторов в принятом пакете Hello то он считает что достигнуто состояние двухсторонней связи

OSPF – Состояния протокола

1. Down State
2. Init State
3. Two-way – состояние двухсторонней связи
4. ExStart – начало обмена
   - Определяется какой маршрутизатор будет являться мастером а какой ведомым при передаче LSDB
   - Мастером становится маршрутизатор с большим идентификатором.

OSPF – Состояния протокола

1. Down State
2. Init State
3. Two-way – состояние двухсторонней связи
4. ExStart – начало обмена
5. Exchange - обмен
   - С помощью пакетов Database Data происходит обмен списком каналов, которые известны маршрутизаторам
   - При обнаружении неизвестных каналов информация по ним запрашивается дополнительно

OSPF – Состояния протокола

1. Down State
2. Init State
3. Two-way – состояние двухсторонней связи
4. ExStart – начало обмена
5. Exchange – обмен
6. Loading - загрузка
   - Маршрутизаторы запрашивают информацию о неизвестных им связям используя пакеты Link State Request и Link State Update.
   - Доставка информации подтверждается пакетами Link State Acknowledgement

OSPF – Состояния протокола

1. Down State
2. Init State
3. Two-way – состояние двухсторонней связи
4. ExStart – начало обмена
5. Exchange – обмен
6. Loading – загрузка
7. Full adjacency – полная смежность
   - Маршрутизаторы имеют полную информацию о состоянии сети

Таймеры протокола OSPF

- **HelloInterval** — Интервал времени в секундах по истечении которого маршрутизатор отправляет следующий hello-пакет с интерфейса. Для широковещательных сетей и сетей точка-точка значение по умолчанию, как правило, 10 секунд. Для нешироковещательных сетей со множественным доступом значение по умолчанию 30 секунд.
- **RouterDeadInterval** — Интервал времени в секундах по истечении которого сосед будет считаться «мертвым». Этот интервал должен быть кратным значению HelloInterval. Как правило, RouterDeadInterval равен 4 интервалам отправки hello-пакетов.
- **Wait Timer** — Интервал времени в секундах по истечении которого маршрутизатор выберет DR в сети. Значение равно значению интервала RouterDeadInterval.
- **RxmtInterval** — Интервал времени в секундах по истечении которого маршрутизатор повторно отправит пакет на который не получил подтверждения о получении. Это интервал называется также Retransmit interval. Значение интервала 5 секунд.

Пакеты протокола OSPF

- Пакеты OSPF инкапсулируются непосредственно в поле данных протокола IP. Значение «тип протокола» для OSPF равно 89.
- Типы пакетов
  - Hello (1)
    - Периодически рассылаются по всем интерфейсам.
    - Используются для нахождения соседей и поддержания состояния соседства.
  - Database Description (2)
    - Используются для передачи LSDB при установлении состояния смежности.
  - Link State Request (3)
    - Используются для запроса у соседей более свежей информации по связям.
  - Link State Update (4)
    - Осуществляют передачу LSA в режиме затопления.
  - Link State Acknowledgement (5)
    - Подтверждают получение пакета Link State Update

Пакеты протокола OSPF
Заголовок пакета

Version	Type	Packet Length
Router ID		
Area ID		
Checksum		Authentication Type
Authentication		
Authentication		

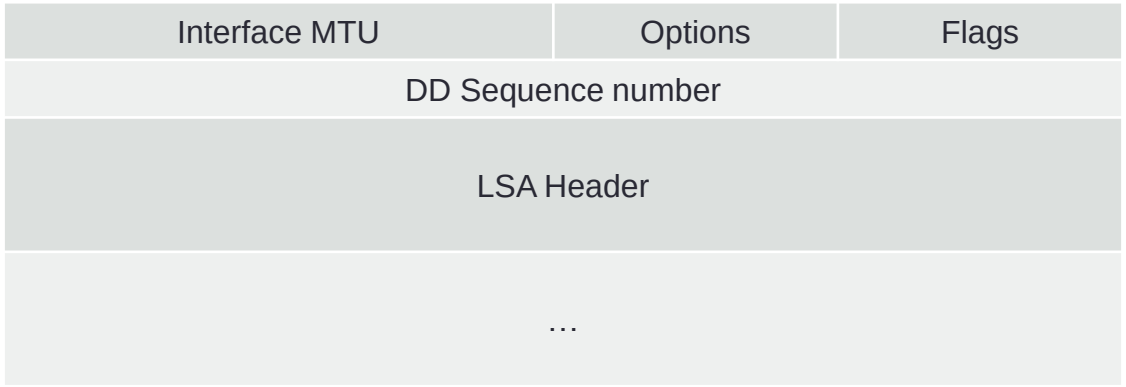
- version — номер версии протокола OSPF. Текущая версия OSPF для сетей IPv4 — 2.
- type — тип OSPF-пакета. В RFC 2328 описано 5 типов пакетов.
- packet length — длина пакета, включая заголовок.
- router ID — идентификатор маршрутизатора — уникальное 32-хбитное число, идентифицирующее маршрутизатор в пределах автономной системы.
- area ID — 32-хбитный идентификатор зоны.
- checksum — поле контрольной суммы. Подсчитывается для всего пакета, включая заголовок.
- authentication type — тип используемой схемы аутентификации. Возможные значения:
  - 0 — аутентификация не используется
  - 1 — аутентификация открытым текстом
  - 2 — MD5-аутентификация
- authentication — поле данных аутентификации.

Пакеты протокола OSPF
Пакет HELLO

Network Mask		
Hello interval	Options	Router Priority
Router Dead Interval		
Designated Router		
Backup Designated Router		
Neighbor ID		
Neighbor ID		
…		

- network mask — сетевая маска интерфейса, через который отправляется hello-пакет.
- hello interval — интервал отправки сообщений hello. Значение по умолчанию равно 10 сек.
- options — Поле опций. Описывает возможности маршрутизатора.
- router priority — приоритет маршрутизатора при выборе DR/BDR.
- router dead interval — период времени, в течение которого маршрутизатор ожидает ответа соседей.
- designated router (DR) — IP-адрес DR.
- backup designated router (BDR) — IP-адрес BDR.
- neighbor ID — Список из идентификаторов соседей, от которых маршрутизатор получил hello-пакеты в течение времени, заданного в поле router dead interval.

Пакеты протокола OSPF
Пакет Database Description



- Interface MTU – максимально допустимый размер кадра на интерфейсе.
- options — Поле опций. Описывает возможности маршрутизатора.
- Flags – Флаги:
  - I – Инициализация. Показывает что пакет является первым.
  - M – Будут ещё пакеты.
  - MS – Мастер (1) /Слейв (0).
- DD Sequence number – порядковый номер пакета при обмене LSDB
- LSA Header – заголовок данных LSDB

Пакеты протокола OSPF
Пакет Link State Request

LS Type
Link State ID
Advertising Router
…

- LS Type – тип связи
- Link State ID – идентифицирует связь
- Advertising Router - маршрутизатор

Пакеты протокола OSPF
Пакет Link State Update

#LSAs
LSAs
…

- #LSAs – число LSA в пакете
- LSA – данные LSDB

Пакеты протокола OSPF
Пакет Link State Acknowledgement

LSA Header
LSA Header
…

- LSA Header – заголовок данных LSDB

Типы LSA

1. **Router LSA** — объявление о состоянии каналов маршрутизатора. Эти LSA распространяются всеми маршрутизаторами. В LSA содержится описание всех каналов маршрутизатора и стоимость (cost) каждого канала. Распространяются только в пределах одной зоны.
2. **Network LSA** — объявление о состоянии каналов сети. Распространяется DR. В LSA содержится описание всех маршрутизаторов присоединенных к сети, включая DR. Распространяются только в пределах одной зоны.
3. **Network Summary LSA** — суммарное объявление о состоянии каналов сети. Объявление распространяется пограничными маршрутизаторами. Объявление описывает только маршруты к сетям вне зоны и не описывает маршруты внутри автономной системы. Пограничный маршрутизатор отправляет отдельное объявление для каждой известной ему сети.
4. **ASBR Summary LSA** — суммарное объявление о состоянии каналов пограничного маршрутизатора автономной системы. Объявление распространяется пограничными маршрутизаторами.
5. **AS External LSA** — объявления о состоянии внешних каналов автономной системы. Объявление распространяется пограничным маршрутизатором автономной системы в пределах всей автономной системы. Объявление описывает маршруты внешние для автономной системы OSPF или маршруты по умолчанию (default route) внешние для автономной системы OSPF.
6. **Multicast OSPF LSA** — специализированный LSA, который используют мультикаст OSPF приложения.
7. **AS External LSA for NSSA** — объявления о состоянии внешних каналов автономной системы в NSSA зоне. Это объявление может передаваться только в NSSA зоне. На границе зоны пограничный маршрутизатор преобразует type 7 LSA в type 5 LSA.
8. **Link LSA** — анонсирует link-local адрес и префикс(ы) маршрутизатора всем маршрутизаторам разделяющим канал (link). Отправляется только если на канале присутствует более чем один маршрутизатор. Распространяются только в пределах канала (link).

Пакеты протокола OSPF
Заголовок LSA

LS Age	Options	LS Type
Link State ID		
Advertising Router		
LS Sequence Number		
LS Checksum	length	

- LS Age – время в секундах от создания LSA
- Options – дополнительные опции, поддерживаемые описываемым доменом
- LS Type – тип LSA
- Link State ID – идентификатор связи:
  - Point-to-point соединение с другим маршрутизатором — Router ID соседа
  - Соединение с широковещательной сетью — IP-адрес DR
  - Соединение с тупиковой сетью (сеть, к которой присоединен только один маршрутизатор) — номер сети/подсети
  - Virtual link — Router ID соседа
- Advertising Router – маршрутизатор, создавший LSA
- LS Sequence Number – порядковый номер для удаления дубликатов и старых записей
- LS Checksum – контрольная сумма всех данных кроме LS Age
- Length – размер LSA

Выбор лучшего маршрута в протоколе OSPF

1. Приоритеты маршрутов
    1. Внутренние маршруты зоны / intra-area
    2. Маршруты между зонами / inter-area
    3. Внешние маршруты типа 1 / E1
    4. Внешние маршруты типа 2 / E2
2. Метрика
    - Метрика в протоколе OSPF определяется на основе пропускной способности интерфейса и называется стоимостью / cost.
    - Стоимость маршрута считается как сумма стоимостей интерфейсов по пути передачи LSA.
    - Недоступные сети обозначаются метрикой $2^{24} - 1 = 16777215$.

Частичное вычисление кратчайшего пути Partial SPF Calculation

- При получении Network Summary LSA маршрутизатор добавляет в таблицу маршрутизации информацию о сетях, которые анонсируются этим LSA, но не запускает алгоритм SPF для этих сетей.
- Метрика для этих сетей высчитывается на основании стоимости, которая анонсируется в Network Summary LSA плюс стоимость пути до ABR, который отправил LSA.
- Если в зоне произошли изменения, то маршрутизаторы в других зонах не запускают SPF, а используют новую метрику, которая приходит в Network Summary LSA, добавляют к ней стоимость пути к ABR и помещают маршрут в таблицу маршрутизации.
- Partial SPF calculation выполняется независимо от того настроено суммирование маршрутов на границе зоны или нет.

Предотвращение петель между зонами ABR Loop Prevention

- Внутри зон OSPF использует логику link-state протокола, но между зонами он, в некотором смысле, работает как дистанционно-векторный протокол.
- Например, при анонсировании в зону Network Summary LSA, передается информация о сети назначения, стоимости пути и ABR, через которого эта сеть достижима — параметры аналогичны информации, которую передают дистанционно-векторные протоколы.
- OSPF не использует традиционные механизмы дистанционно-векторных протоколов для предотвращения петель. OSPF использует несколько правил, которые касаются распространения LSA между зонами и таким образом исключает возможность возникновения петель.
- ABR Loop Prevention может привести к тому, что передача данных будет осуществляться не по лучшему пути.

Алгоритм построения таблицы маршрутизации OSPF

1. Текущая таблица маршрутизации обнуляется. Таблица маршрутизации строится снова с нуля. Старая таблица маршрутизации сохраняется для того чтобы можно было обнаружить изменения в определенных записях таблицы.
2. С помощью построения дерева кратчайшего пути для каждой присоединенной зоны вычисляются внутризональные маршруты. Во время вычисления дерева кратчайшего пути для зоны, также для зоны высчитывается TransitCapability, которая используется позже на 4 этапе. Фактически, все записи таблицы маршрутизации с типом назначения (Destination Type) area border router высчитываются на втором этапе. Этот этап состоит из двух частей:
    1. Сначала строится дерево с учетом только связей между маршрутизаторами и транзитными сетями.
    2. Затем в дерево включаются тупиковые сети.
3. Межзональные маршруты вычисляются выполнением просмотра существующих summary LSA. Если маршрутизатор пограничный, то просматриваются суммарные LSA только магистральной зоны.
4. На пограничных маршрутизаторах, которые присоединены к одной или более транзитным зонам (не магистральные зоны в которых TransitCapability установлена в TRUE), проверяются суммарные LSA транзитных зон. LSA проверяются на наличие лучших путей, чем пути, которые были обнаружены на этапах 2-3.
5. Высчитываются маршруты к внешним сетям. Для этого просматриваются AS-external-LSA. Местонахождение ASBR-маршрутизаторов было обнаружено на этапах 2-4.

Дополнительные возможности OSPF

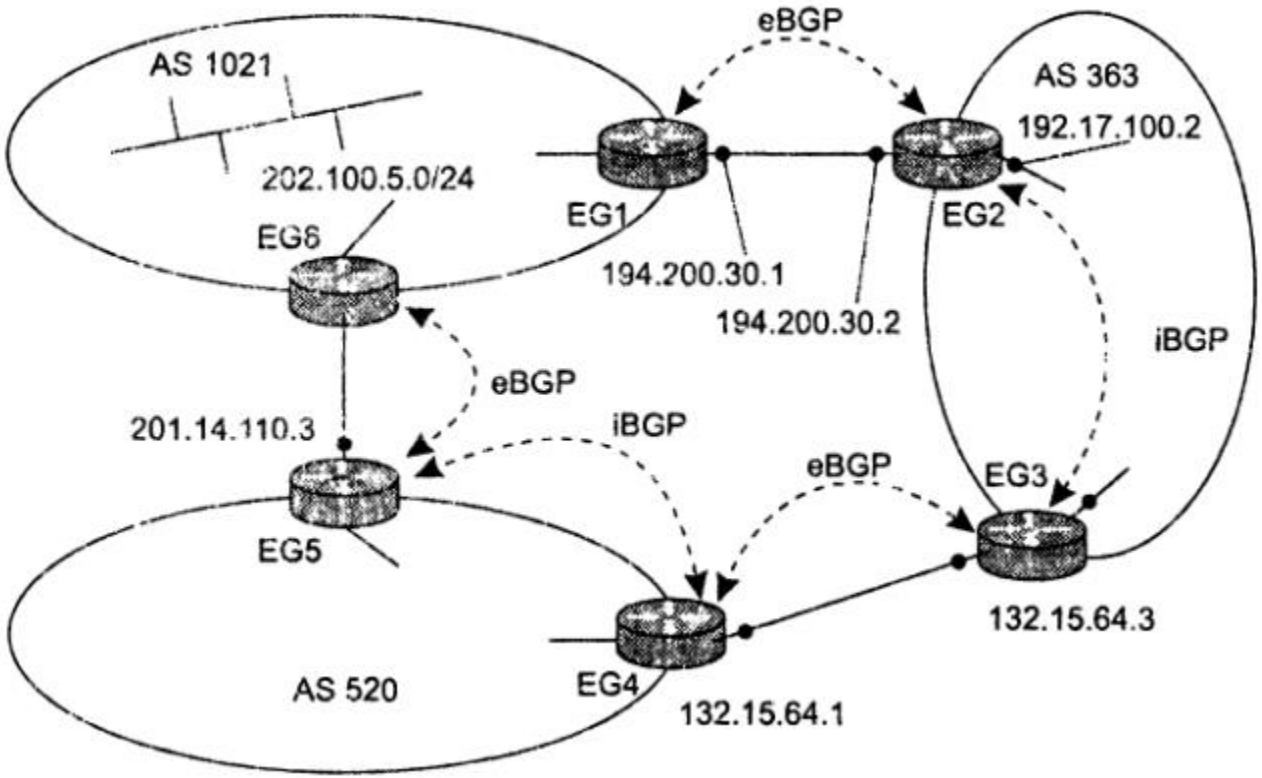
- Многотрактовая маршрутизация / multipath routing

- Маршрутизация, базирующаяся на запросах типа услуг высшего уровня / type of service – TOS.

Border Gateway Protocol (BGP)

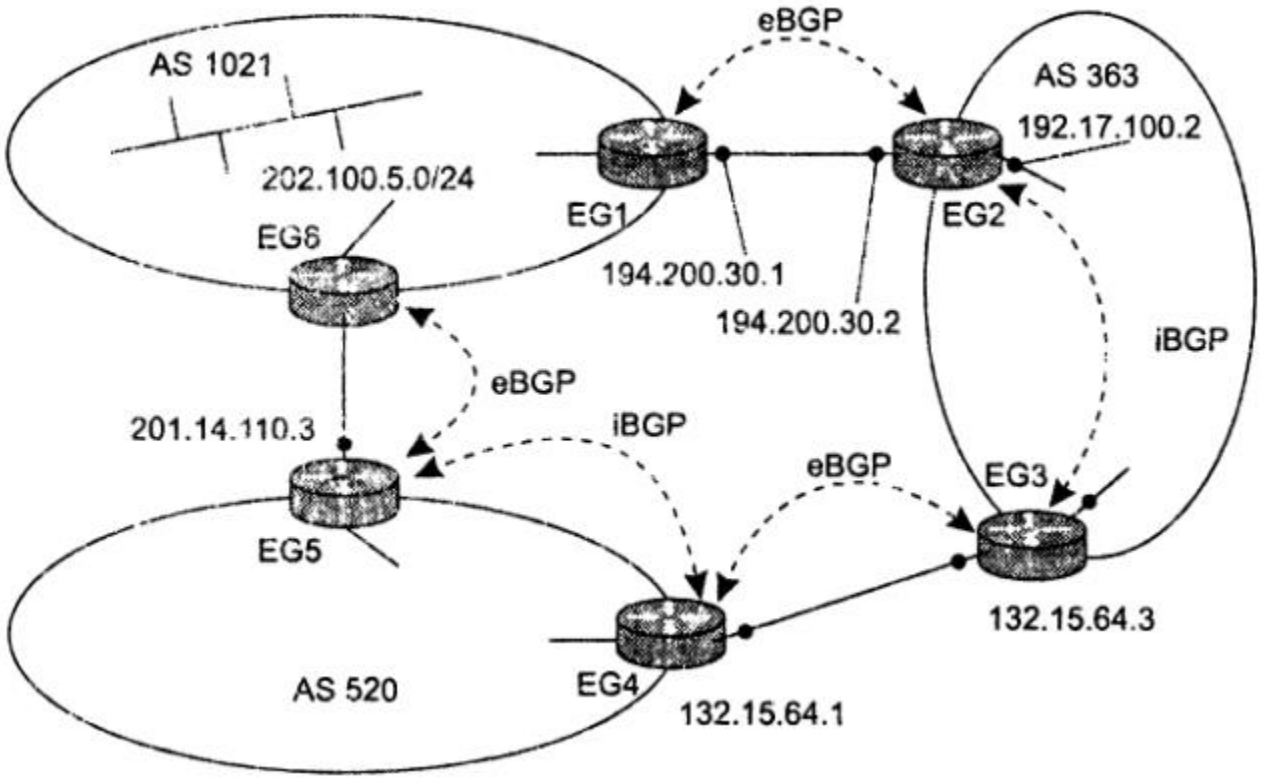
- Протокол междоменной маршрутизации Internet.
- В данный момент применяется BGP версии 4
  - 1994 г. – RFC 1654
  - …
  - 2006 г. – RFC 4271
- Может использоваться в качестве внутреннего протокола маршрутизации в очень больших автономных системах.

BGP



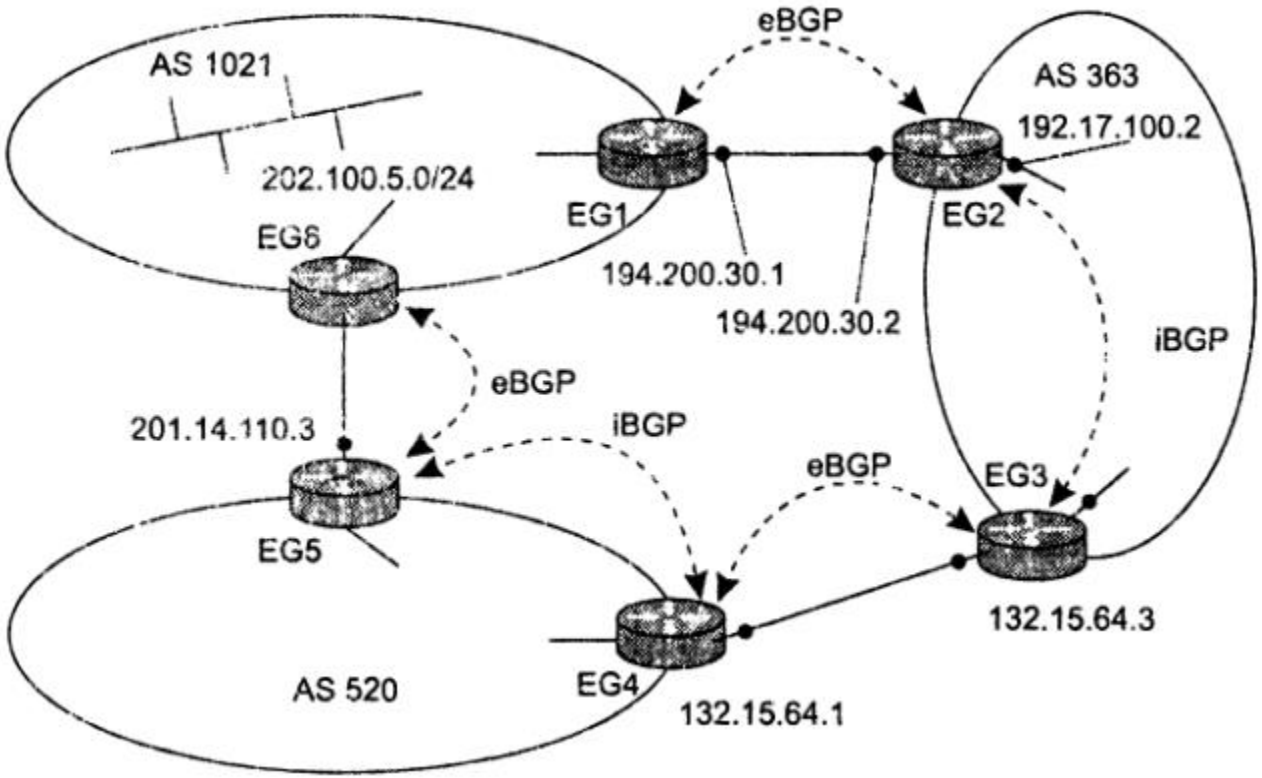
- Взаимодействие про протоколу BGP осуществляется только в случае явного задания соседа администратором сети.
- Для установления соединения используется TCP порт 179.
- Состояние соединения описывается автоматом с 5 состояниями.
- По установленному соединению периодически (60 с.) передаются сообщения Keep Alive.

BGP



- Основное сообщение протокола – UPDATE, передает информацию о достижимости префиксов в рамках AS.
  - Можно объявить появление нового маршрута или о удалении нескольких старых.
  - Информация о маршруте имеет вид AS_Path; Next Hop; Network/Mask
    - AS_Path – набор номеров автономных систем
    - Next Hop – IP адрес маршрутизатора, которому нужно передавать пакеты адресованные в Network/Mask
- Пример: EG1→EG2: AS1021; 194.200.30.1; 202.100.5.0/24

BGP

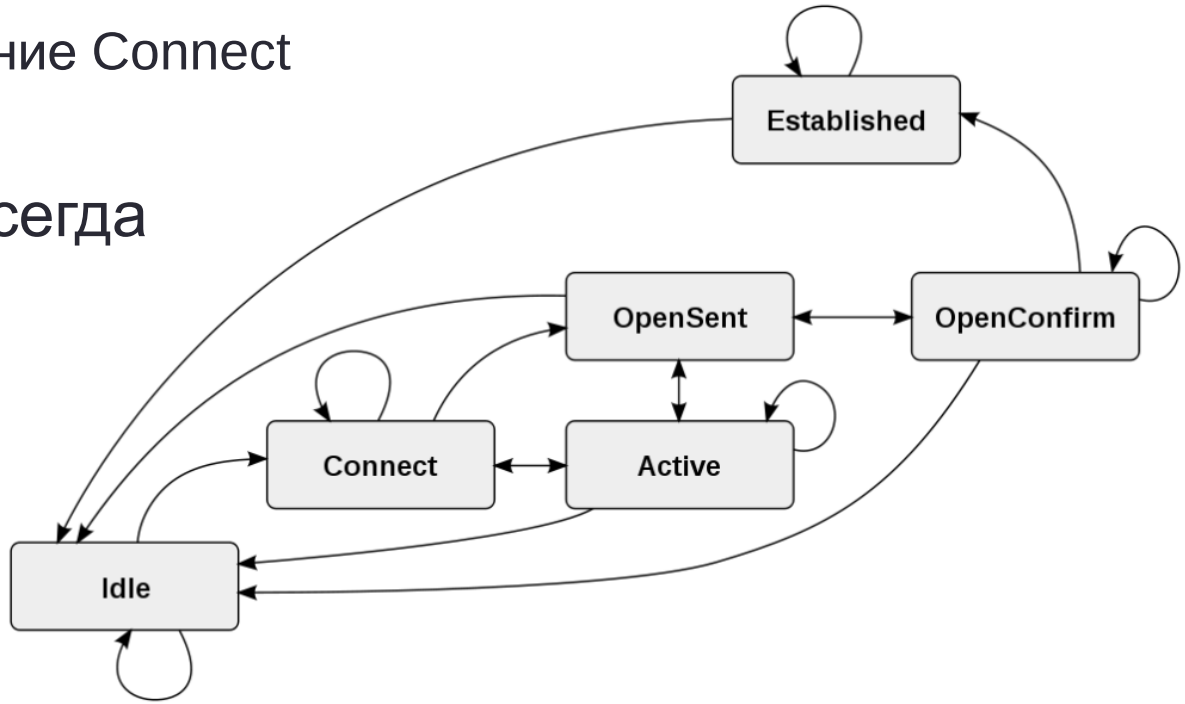


- Пример:
  - EG1→EG2: AS1021; 194.200.30.1; 202.100.5.0/24
  - EG3→EG4: AS363, AS1021; 132.15.64.3; 202.100.5.0/24
  - EG5→EG6: AS520,AS363,AS1021;21.14.110.3; 202.100.5.0/24

Состояния протокола BGP

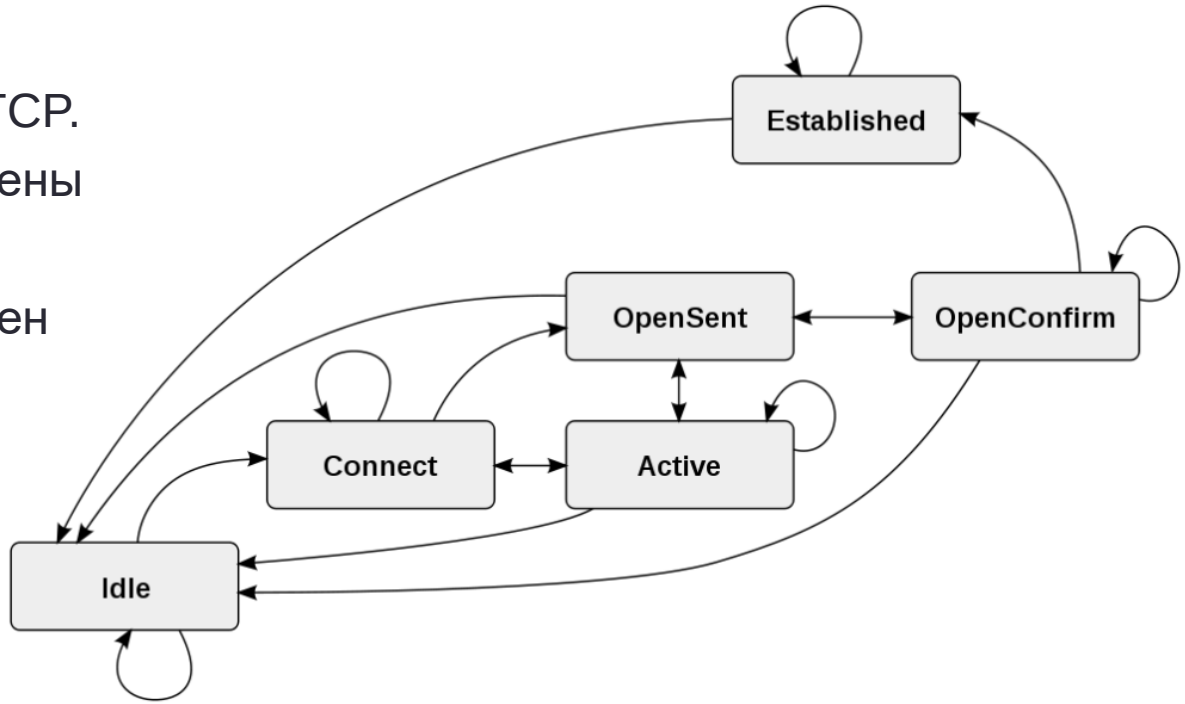
- Idle
  - Игнорируем входящие соединения
  - Инициализируем ресурсы
  - Пытаемся открыть соединения с настроенными соседями
  - Принимаем соединения
  - Переходим в состояние Connect

- В случае ошибок всегда возвращаемся в Idle



Состояния протокола BGP

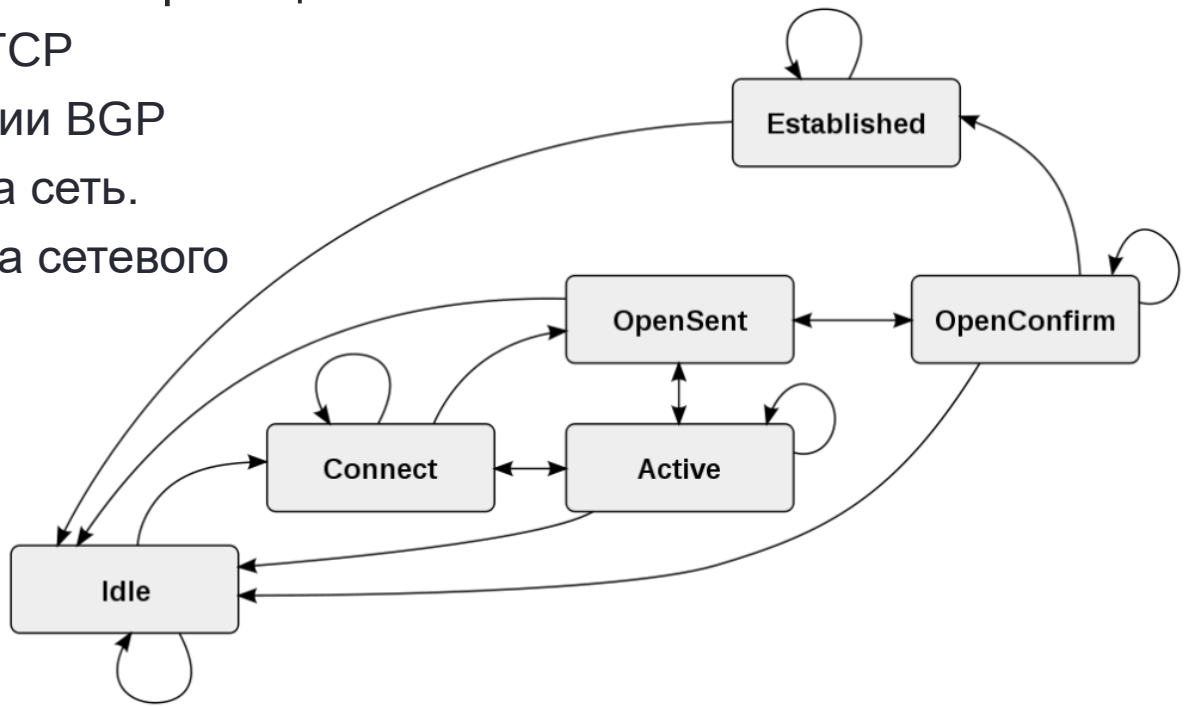
- Connect
  - Ожидаем открытия TCP соединения.
  - Посылаем сообщение Open и переходим в состояние OpenSent.
  - В случае ошибки переходим в состояние Active.
  - Возможные ошибки
    - Недоступны порты TCP.
    - Некорректно настроены адреса соседей.
    - Некорректно настроен адрес AS.



Состояния протокола BGP

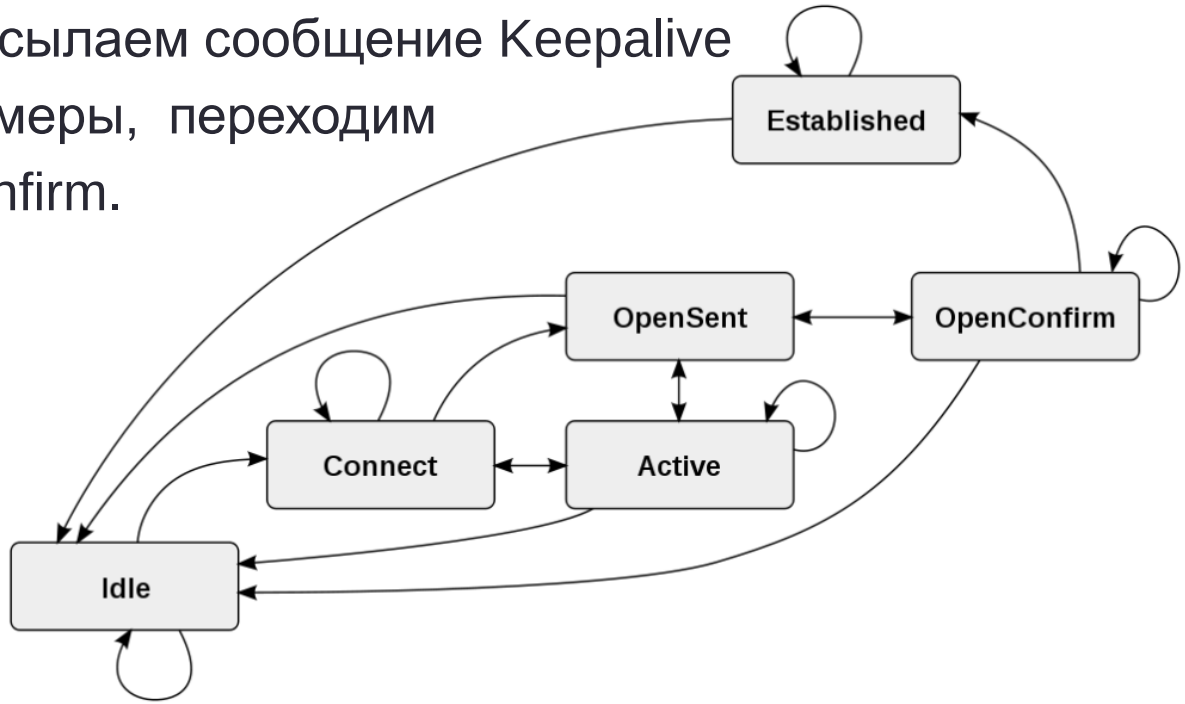
- Active
  - Повторяем попытку соединения.
  - В случае успеха посылаем Open.
  - В случае неудачи переходим в Idle.

  Возможные причины повторяющихся ошибок:
  - Недоступны порты TCP
  - Ошибка конфигурации BGP
  - Большая нагрузка на сеть.
  - Неустойчивая работа сетевого интерфейса.



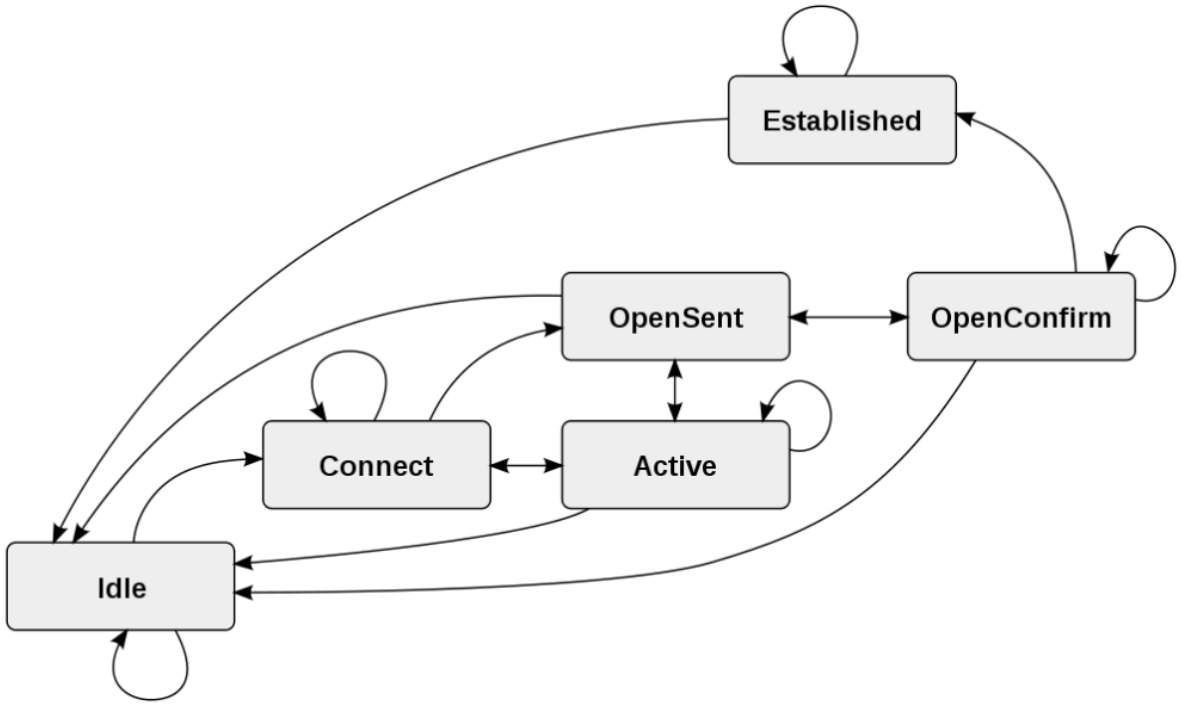
Состояния протокола BGP

- OpenSent
  - Ожидаем сообщения Open от соседа.
  - Проверяем корректность сообщения Open.
  - В случае ошибки (неправильный номер AS, версия протокола и т.п.) посылаем сообщение Notification.
  - Если всё хорошо, посылаем сообщение Keepalive инициализируем таймеры, переходим в состояние OpenConfirm.



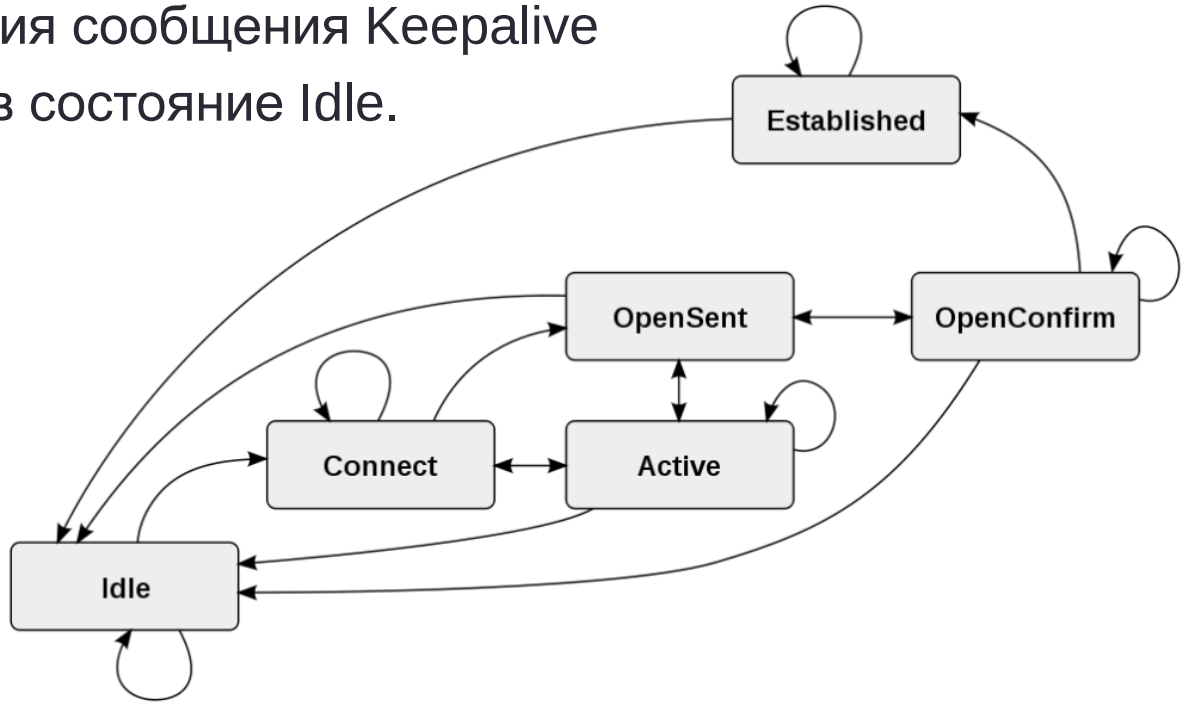
Состояния протокола BGP

- OpenConfirm
  - Ожидаем сообщения Keepalive от соседа.
  - Если получили Keepalive до истечения таймаута переходим в состояние Established.
  - Если таймаут вышел или произошла ошибка возвращаемся в состояние Idle.



Состояния протокола BGP

- Established
  - Обмениваемся сообщениями Update с объявлениями маршрутов.
  - В случае ошибки посылается сообщение Notification и переходим в состояние Idle.
  - В случае не получения сообщения Keepalive вовремя переходим в состояние Idle.



Сообщения BGP

Сообщение Open

- Version - номер версии BGP
- Autonomous system - номер AS отправителя.
- Hold time - указывает максимальное число секунд, которые могут пройти без получения какого-либо сообщения от передающего устройства, прежде чем считать его отказавшим.
- Authentication code – метод аутентификации.
- Autentication data - данные аутентификации.

Сообщения BGP

Сообщение Update

- Информация о маршрутах:
  - Origin - Источник.
    - IGP - означает, что данная сеть является частью данной AS.
    - EGP - первоначальные сведения о данной информации получены от протокола EGP.
    - Incomplete данные получены через какие-то другие средства.
  - *AS path* - Путь AS. Фактический перечень AS на пути к пункту назначения.
  - *Next hop* - Следующая пересылка. IP адрес роутера, который должен быть использован в качестве следующей пересылки к сетям, перечисленным в сообщении о корректировке.
  - *Unreachable* - Недосягаемый. Указывает (при его наличии), что какой-нибудь маршрут больше не является досягаемым.
  - *Inter-AS metric* Метрика между AS. Обеспечивает для роутера BGP возможность объявить свои затраты на маршруты к пунктам назначения, находящимся в пределах его AS. Эта информация может быть использована роутерами, которые являются внешними по отношению к AS объявляющего, для выбора оптимального маршрута к конкретному пункту назначения, находящемуся в пределах данной AS.

Сообщения BGP

Сообщение KeepAlive

- Не содержит данных

Сообщение Notification

- error code - код ошибки
  - Message header error - ошибка в заголовке сообщения.
  - Open message error - ошибка в открывающем сообщении.
  - Update message error - ошибка в сообщении о корректировке.
  - Hold time expired - время удерживания истекло.
- error subcode - поле подкода ошибки
- error data - данные ошибки

Обработка сообщений UPDATE

- Концептуально, реализация протокола BGP должна поддерживать следующие базы данных маршрутов:
  - Local Routing Information Base (Loc-RIB) – таблица маршрутов, известных BGP.
  - Для каждого соседа:
    - Adjacent Routing Information Base, Incoming (Adj-RIB-In) – получено от соседа;
    - Adjacent Routing Information Base, Outgoing (Adj-RIB-Out) – передано соседу.
- Порядок обновления баз: Adj-RIB-In → Loc-RIB
- BGP передает маршруты, которые он считает лучшими, из Loc-RIB в основную таблицу маршрутизации маршрутизатора.
  - При этом преимущество могут получать и маршруты, полученные из других источников.

Правила выбора маршрутов (сильно упрощенно)

- **На уровне Adj-RIB**

1. Next Hop должен быть доступен.
2. Если сообщение получено через iBGP:
    1. Выбираем маршрут от соседа с наибольшим Local Preference.
    2. Могут быть дополнительные локальные приоритеты (Weight у Cisco).
3. Выбираем маршрут с минимальным числом транзитных AS.
4. Выбираем маршрут с минимальным Orign (IGP<EGP<Incomplete).
5. Выбираем маршрут с минимальным MED.

- **На уровне Loc-RIB**

1. Маршруты, полученные через eBGP предпочтительнее полученных через iBGP.
2. Выбираем маршрут с минимальным путем к Next Hop внутри нашей AS.
3. Выбираем маршрут, который был получен раньше.
4. Выбираем маршрут с наименьшим идентификатором отправителя.
5. Выбираем маршрут от соседа с наименьшим IP.

Фильтрация сообщений UPDATE

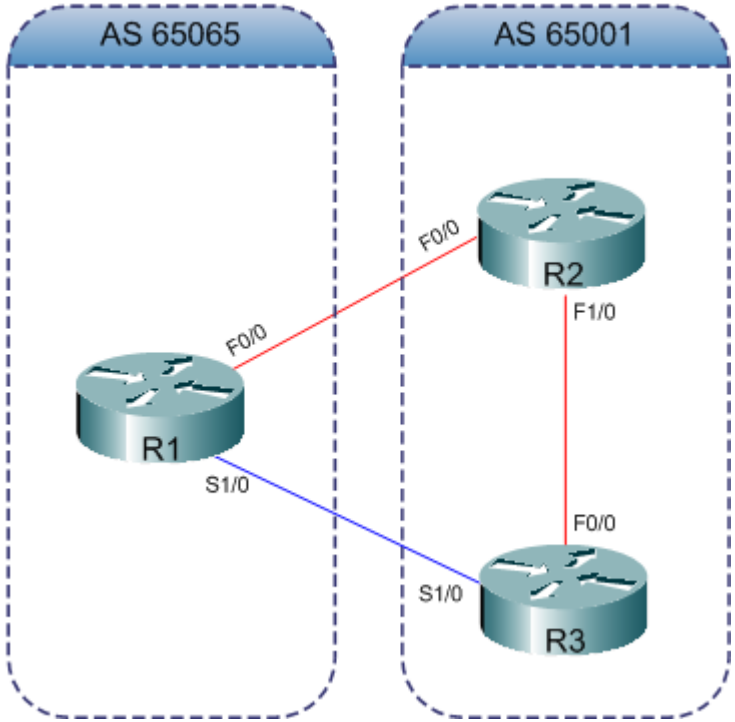
- Существует возможность управления отправкой/получением update.
- Фильтрация может осуществляться на основе:
  - префиксов сетей, которые анонсируются;
  - атрибутов;
  - пути AS.

Атрибут BGP Community

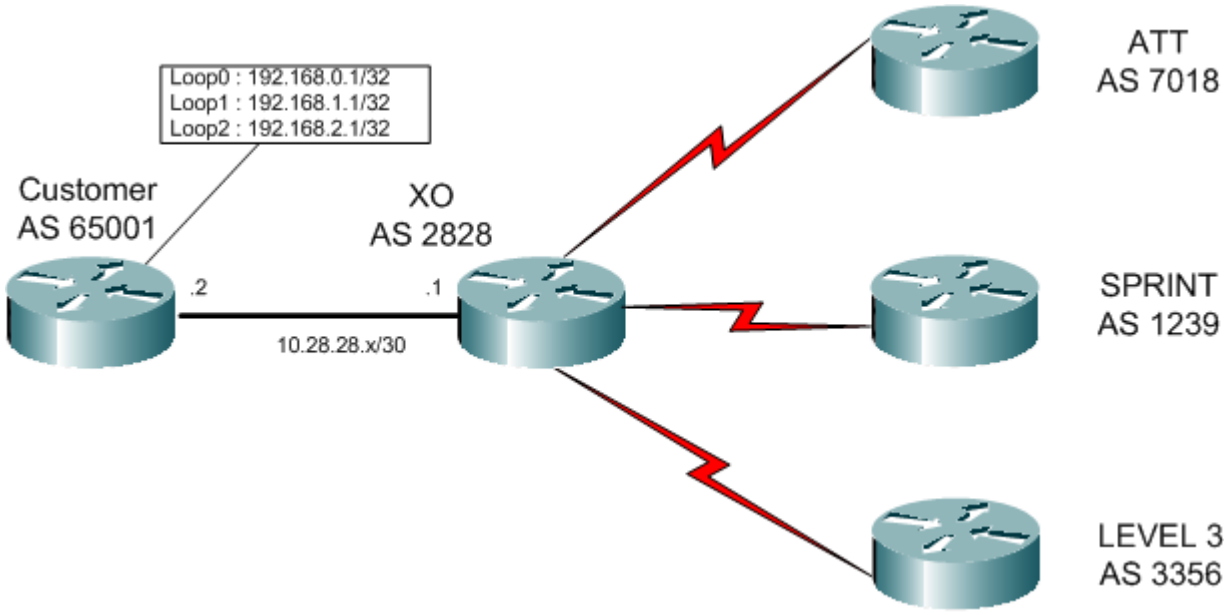
- Community – числовое значение, которое может быть присвоено префиксу и анонсируется соседям.
- Когда сосед получит этот атрибут, то он обработает его в соответствии с правилами, для данного Community.
- Общепринятые значения:
  - Internet: анонсируй этот префикс всем соседям;
  - Local-as: не передавать префикс за пределы местной конфедерации AS;
  - No-advertise: не анонсировать префикс никому;
  - No-export: не анонсировать трафик по eBGP.
- Другие популярные варианты:
  - Ограничения по географии или типах пиринга анонсов;
  - Настройка приоритетов
  - Добавление AS в путь
  - Борьба с DoS (Black Holing)
  - QoS
  - …

Атрибут BGP Community

- Наша AS 65065 подключена к двум маршрутизаторам в AS 650001.
  - Связь с R2 быстрая.
  - Связь с R3 медленная.
- AS 650001 поддерживает Community:
  - 200: Установить локальный приоритет 200.
  - 500: Установить локальный приоритет 500.
- Что делаем:
  - Анонсируем наши сети для R2 с Community 650001:500.
  - Для R3 с 650001:200.
- Что получаем:
  - Если исправны оба канала, трафик отправляется через R2-R1
  - Если канал R2-R1 неисправен, используется R3-R1



Атрибут BGP Community



Community на XO (AS2828):

Провайдер	Не анонсировать	Добавить 1 раз	Добавить 2 раза	Добавить 3 раза
Sprint	2828:1003	2828:1103	2828:1203	2828:1303
Level 3	2828:1007	2828:1107	2828:1207	2828:1307
AT&T	2828:1008	2828:1108	2828:1208	2828:1308

Атрибут BGP Community

Атрибуты BGP Community MSK-IX

- Базовые
  - 0:peer-as - Блокировка анонса префикса участнику с номером AS peer-as
  - 8631:peer-as - Анонс префикса участнику с номером AS peer-as
  - 0:8631 - Блокировка анонса префикса всем участникам
  - 8631:8631 - Анонс префикса всем участникам

- Дополнительные
  - 8631:65281 - Анонс префикса участникам с атрибутом no-export
  - 8631:0 - Установка local-preference 0
  - 8631:50 - Установка local-preference 50
  - 8631:100 - Установка local-preference 100 (по умолчанию)

BGP Blackhole для борьбы с DoS

- Допустим, на один из наших IP-адресов поступает огромный объем трафика
  - Сервер не в состоянии обработать все запросы
  - Перегружена внутренняя сеть
  - Перегружены каналы
- Необходимо фильтровать трафик как можно ближе к источнику, желательно на стороне нашего провайдера и автоматически.
- Решение “Destination-Based Remotely Triggered Black Hole Filtering”:
  - По BGP анонсируем префикс /32 помеченный специальным community.
  - Наш ISP убивает трафик, идущий на этот префикс.
- Результаты:
  - Трафик атаки не попадает в нашу сеть и не перегружает наши каналы
  - DoS атака достигает своих результатов.
  - Другие ресурсы остаются доступными.

BGP Blackhole для борьбы с DoS

- Допустим, мы можем определить адреса, с которых идёт атака
  - Это обычно сложно сделать, чаще всего они подложные и меняются.
- Если маршрутизатор работает в режиме Unicast Reverse Path Forwarding, то он будет передавать только те пакеты, для которых есть корректный маршрут до отправителя.
- Source-Based Remotely Triggered Black Hole Filtering:
  - Анонсируем адреса нападающих с маршрутом в null.

- Другие варианты:
  - С помощью BGP перенаправляем трафик с нашего хоста на специальный фильтр, который должен вырезать трафик атаки и оставить трафик легитимных клиентов.

Глобальный Blackhole

- В феврале 2008 г. правительство Пакистана издало закон, по которому интернет провайдеры Пакистана должны были закрыть доступ к Youtube, указав 4 IP адреса.
- 24.02.2008 Pakistan Telekom (AS17557) начал анонсировать сеть 208.65.153.0/24 своему провайдеру PCCW (AS 3491), Гонконг.
- PCCW не отфильтровал этот анонс, а передал его другим провайдерам.
- Так как Youtube анонсировал блок большего размера 208.65.152.0/22, то большинство роутеров предпочли анонс Pakistan Telekom.
- Youtube начал анонсировать свои адреса с маской /24, потом /25.
- Через ~2 часа некорректный анонс прекратился. По оценкам было затронуто 2/3 пользователей интернет.

Глобальный Blackhole

18:47:00 uninterrupted videos of exploding jello

18:47:45 first evidence of hijacked route propagating in Asia, AS path 3491 17557

18:48:00 several big trans-Pacific providers carrying hijacked route (9 ASNs)

18:48:30 several DFZ providers now carrying the bad route (and 47 ASNs)

18:49:00 most of the DFZ now carrying the bad route (and 93 ASNs)

18:49:30 all providers who will carry the hijacked route have it (total 97 ASNs)

20:07:25 YouTube, AS 36561 advertises the /24 that has been hijacked to its providers

20:07:30 several DFZ providers stop carrying the erroneous route

20:08:00 many downstream providers also drop the bad route

20:08:30 and a total of 40 some-odd providers have stopped using the hijacked route

20:18:43 and now, two more specific /25 routes are first seen from 36561

20:19:37 25 more providers prefer the /25 routes from 36561

20:28:12 peers of 36561 start seeing the routes that were advertised to transit at 20:07

20:50:59 evidence of attempted prepending, AS path was 3491 17557 17557

20:59:39 hijacked prefix is withdrawn by 3491, who disconnect 17557

21:00:00 the world rejoices; Leeroy Jenkins online again.

Другие случаи передачи некорректной информации BGP

- В апреле 1997 один из пользователей MAI Network Services (AS 7007) анонсировал множество чужих префиксов, которые не были отфильтрованы MAI. Наружу попало не менее 72k префиксов.
- В ночь на рождество 2004 года турецкий провайдер TTNet (AS9121) анонсировал более 100k сетей с минимальным маршрутом. Ему поверили Telekom Italia Seabone и многие другие.
- В сентябре 2005 сеть 12.0.0.0/8, принадлежащая AT&T (AS7018), дважды анонсировалась другими AS (AS6210 AES Communtcations из Боливии и AS12676 Ncore).
- В январе 2006 Con Edison (AS27506) анонсировал около 20 чужих сетей.