Networking
For everyone

OSPF: Convergence and Scalability

В этом разделе

- Loop Free Alternate
- Nonstop Forwarding
- Nonstop Routing
- BFD
- Prefix Suppression
- Stub Router

Networking For everyone

Networking
For everyone

Loop Free Alternate

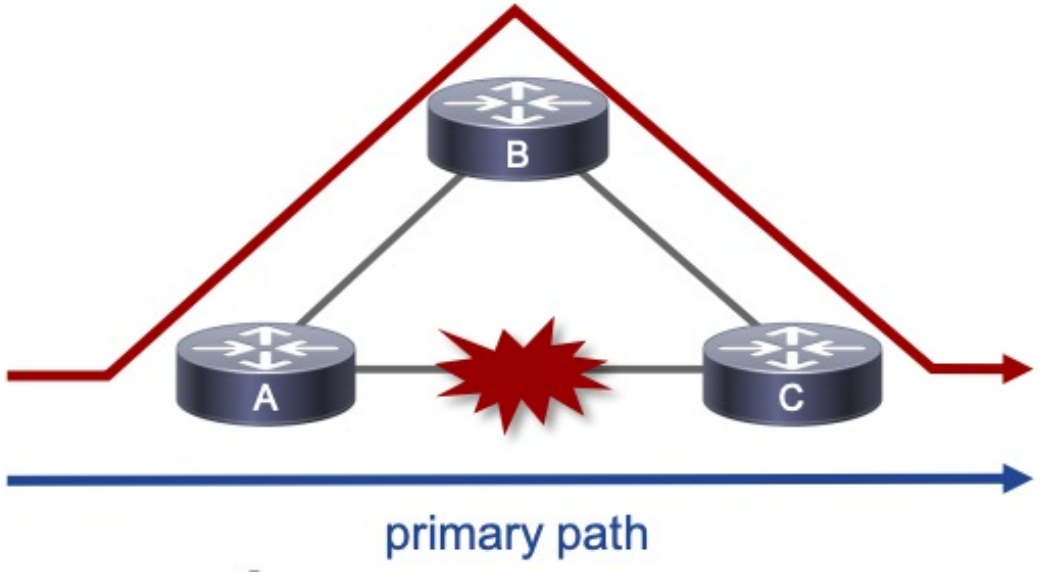
Loop Free Alternative

Networking For everyone

- OSPF LFA FRR позволяет быстро (в течение ~ 50 мс) переключаться на резервный путь
- В обычной ситуации, OSPF должен пересчитать весь граф в случае выхода интерфейса из строя
- С LFA FRR, OSPF делает **предварительный расчёт**
  - Резервный next-hop устанавливается внутрь FIB

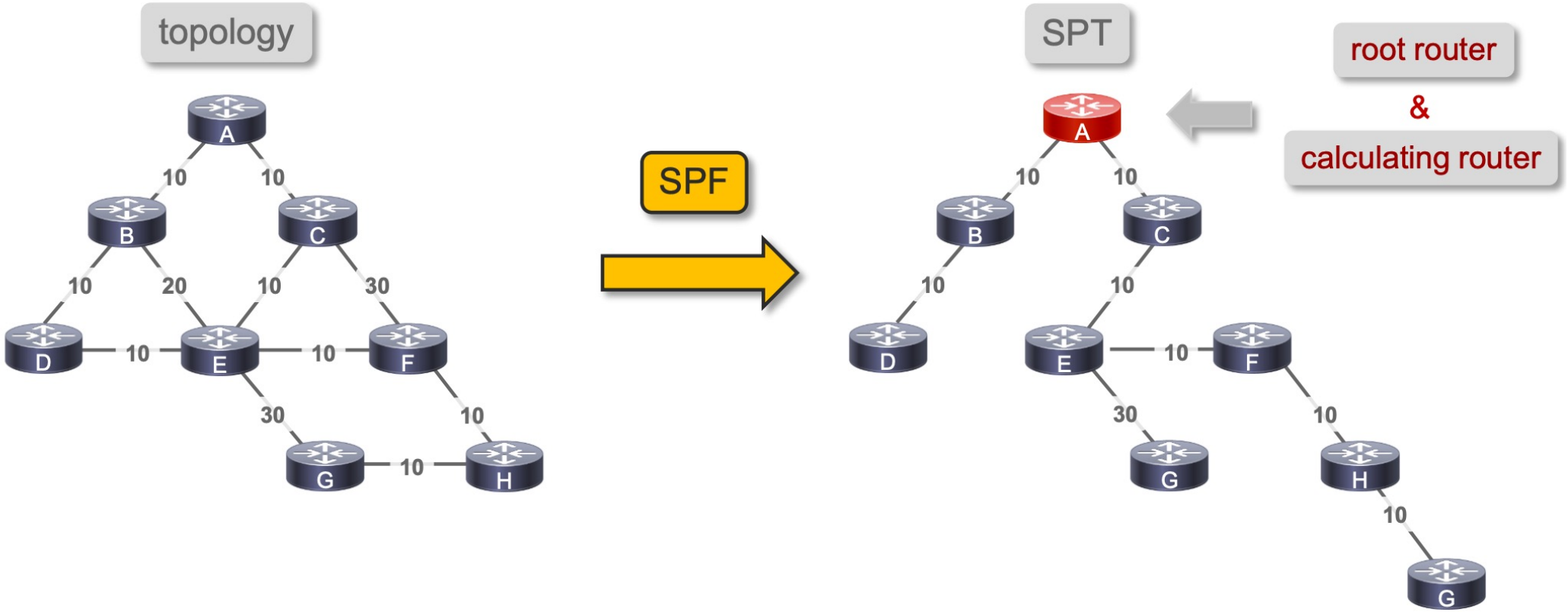
Основная идея

- Маршрутизатор A делает всю калькуляцию
- Другие маршрутизаторы не вовлечены в процесс
- Repair Path (LFA):
  - Трафик от B не должен вернуться к A
  - Трафик должен миновать упавший интерфейс



Классический SPF

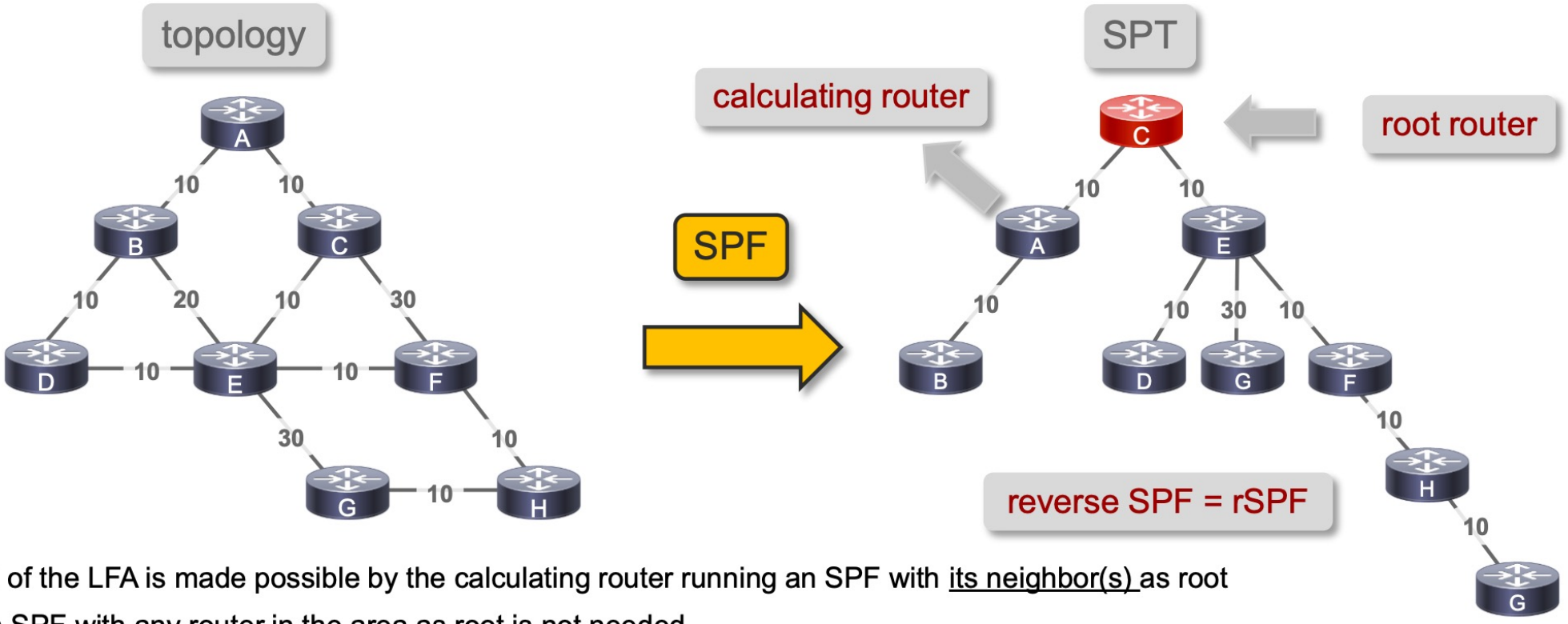
- Необходимо запустить SPF и в качестве корневого устройства поставить **себя**



Хак SPF для LFA (rSPF)

Networking For everyone

- Необходимо запустить SPF, но в качестве корневого устройства поставить **соседа**

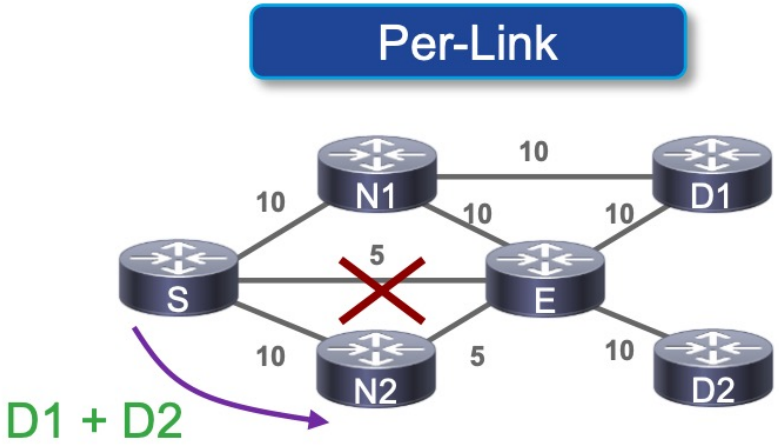


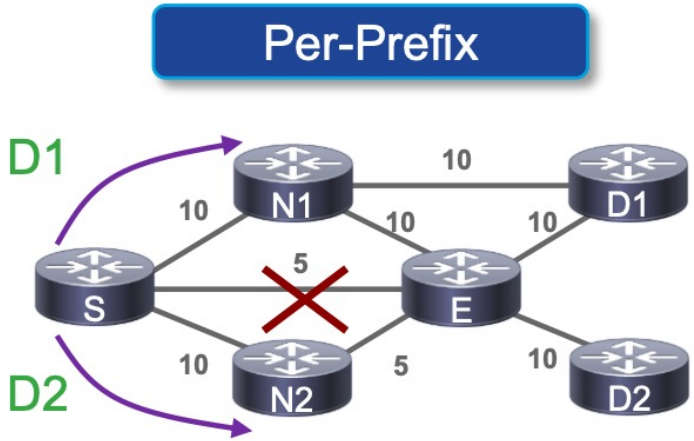
All of the LFA is made possible by the calculating router running an SPF with its neighbor(s) as root

An SPF with any router in the area as root is not needed

LFA методы

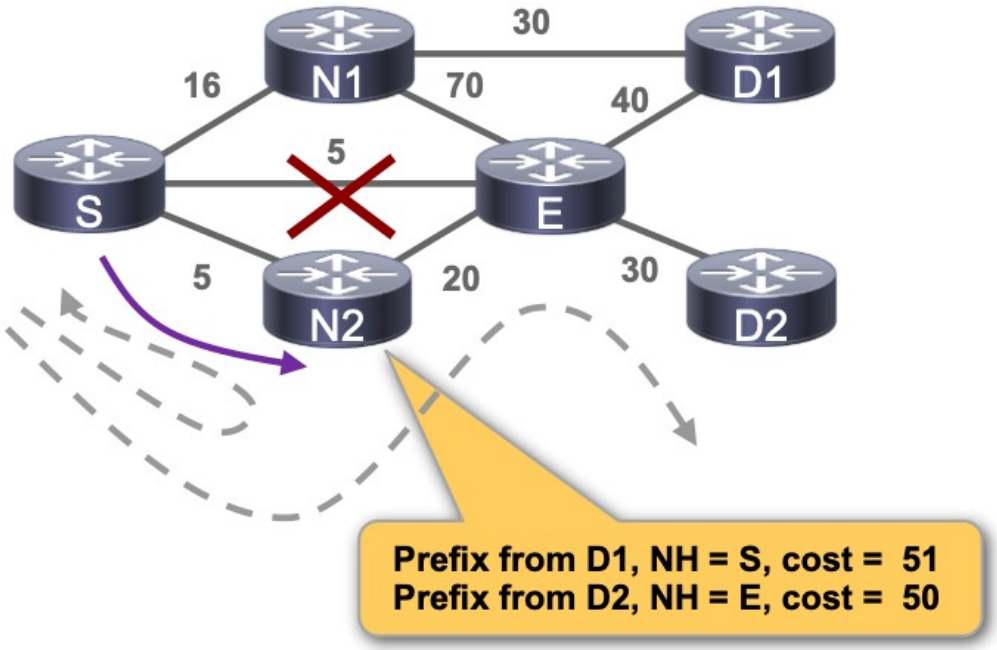
- IGP может запускать LFA в одном из двух режимов
  - Per prefix
    - Резервный путь для каждого префикса считается независимо
  - Per link
    - Резервный путь для всех префиксов через один next-hop
    - Только в IOS-XR





А что если …

- Трафик к D2 передаётся на E от N2
- Трафик к D1 возвращается обратно к S



Per-link vs Per-prefix

Networking
For everyone

Per-link

- Простой расчёт, один rSPF для соседа
- Всё или ничего

Per-prefix

- Расчёт для каждого префикса через каждого соседа

Основные блоки

- До аварии
  - Альтернативный NH устанавливается в RIB и IGP local RIB (LRIB)
  - Альтернативный NH устанавливается в FIB (CEF)
- Во время аварии
  - Детектирования потери интерфейса/соседа
  - Триггер для IP-FRR LFA: переключить префиксы в FIB
- После аварии
  - Обычная конвергенция (SPF)

Основные блоки

Networking
For everyone

```
show ip ospf rib 10.1.1.0/24
 10.1.1.2 via GE 0/0/1, protected
 10.1.2.1 via GE 0/0/2, repair-path
```

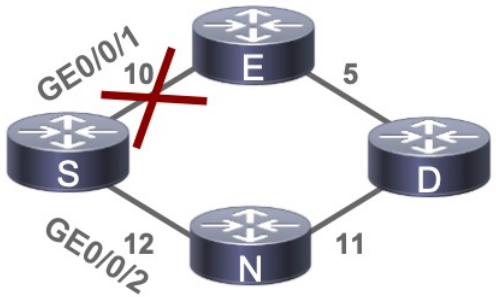
LRIB ← stored in control plane

```
show ip route 10.1.1.0/24
 10.1.1.2 via GE 0/0/1, protected
 10.1.2.1 via GE 0/0/2, backup
```

RIB ← stored in control plane

```
show ip cef 10.1.1.0/24
 10.1.1.2 via GE 0/0/1, protected
 10.1.2.1 via GE 0/0/2, repair-path
```

CEF ← stored in data plane




```
show ip cef 10.1.1.0/24
 10.1.2.1 via GE 0/0/2
```

CEF

```
show ip cef 10.1.1.0/24
 10.1.2.1 via GE 0/0/2
```

CEF

А если альтернатив несколько?

Networking For everyone

- Когда OSPF должен выбрать резервный путь, он может смотреть не только на наименьшую метрику, но и учитывать дополнительные параметры
  - SRLG (Shared Risk Link Groups)
  - Interface Protection
  - Broadcast Interface Protection
  - Node Protection
  - Downstream Path
  - Line-Card Disjoint Interfaces
  - Metric
  - Equal-Cost Multipath

Shared Risk Link Group (SRLG)

Networking For everyone

- Ручная настройка
- Если два интерфейса подключены к одному коммутатору, идут через одну физическую трассу, то логично назначить им одинаковые SRLG

Primary Path

Networking
For everyone

Interface Disjoint

- Предпочесть альтернативный next-hop, который располагается за другим интерфейсом
- Ethernet суб-интерфейсы считаются ***разными*** интерфейсами

Lowest-Metric

- Предпочесть путь с наименьшей метрикой
- У команды нет ключевого значения “required”
  - Т.к. метрика присутствует всегда ☺

Linecard-disjoint

- Предпочесть путь, который использует интерфейс, находящийся на другой линейной карте
  - В виртуальной лаборатории не удастся эмулировать

Node protecting

- Предпочесть путь, который не проходит через тот же маршрутизатор, который используется в качестве основного next-hop

Broadcast interface disjoint

- Понизить приоритет альтернативным маршрутам, которые используют тот же широковещательный домен, что и путь через основной next-hop

Downstream

- Выключен по-умолчанию
- По сути поведение очень похоже на EIGRP Feasible Condition
- Мне не очень понятно, зачем такая опция вообще нужна в LS протоколе

Secondary-Path

- Выключен по-умолчанию
- Предпочесть пусть, который не является частью ECMP

Основные шаги конфигурации

Networking For everyone

- Включить FRR для зоны или глобально
- Включить FRR prefix-priority

```
(config-router)#fast-reroute per-prefix enable prefix-priority { high | low }
```

- Настроить приоритет префиксов (route-map | RPL)
  - /32 = “high” на IOS
  - /32 = “medium” на IOS-XR
  - только match tag | route-type | ip address

```
(config-router)#prefix-priority high route-map { ROUTE-MAP }
```

- Добавить/изменить tie-breakers

```
(config-router)#fast-reroute per-prefix tie-break { TIE } [required] index { INDEX }
```

Пример конфигурации

```
interface Ethernet1/0
 srlg gid 100
!
interface Ethernet6/0
 ip ospf fast-reroute per-prefix candidate disable
!
router ospf 1
 prefix-priority high route-map lfa-ospf
 fast-reroute per-prefix enable prefix-priority high
 fast-reroute per-prefix tie-break srlg index 10
 fast-reroute per-prefix tie-break node-protecting index 20
 fast-reroute keep-all-paths
!
ip prefix-list lfa-high seq 5 permit 10.0.0.0/8 ge 30
!
route-map lfa-ospf permit 10
 match ip address prefix-list lfa-high
```

Пример конфигурации

Networking For everyone

```
R9#show ip cef 5.5.5.5
5.5.5.5/32
  nexthop 10.5.9.5 GigabitEthernet1.59
    repair: attached-nexthop 10.9.5.5 GigabitEthernet1.95
  nexthop 10.9.5.5 GigabitEthernet1.95
    repair: attached-nexthop 10.5.9.5 GigabitEthernet1.59
```

```
R9#show ip route repair-paths 5.5.5.5
Routing entry for 5.5.5.5/32
  Known via "ospf 1", distance 110, metric 2, type intra area
  Last update from 10.5.9.5 on GigabitEthernet1.59, 00:01:40 ago
  Routing Descriptor Blocks:
    10.9.5.5, from 5.5.5.5, 00:01:40 ago, via GigabitEthernet1.95
      Route metric is 2, traffic share count is 1
      Repair Path: 10.5.9.5, via GigabitEthernet1.59
  * 10.5.9.5, from 5.5.5.5, 00:01:40 ago, via GigabitEthernet1.59
      Route metric is 2, traffic share count is 1
      Repair Path: 10.9.5.5, via GigabitEthernet1.95
```

Networking
For everyone

Non-Stop Forwarding (Graceful Restart)

NSF? Wtf?

Networking
For everyone

- Мы уже познакомились с понятием FRR – быстрая конвергенция в случае изменения сети
- Часто в ядре стоят устройства, у которых есть два управляющих модуля (Supervisor)
- К чему приведёт переключение SUP?
- Основной идеей NSF (Nonstop Forwarding) является минимизация времени, в течение которого сеть не способна доставлять трафик в место назначение после события переключения SUP
- Основной целью NSF является продолжать пересылку IP-пакетов, после того, как на устройстве произошло переключение SUP

Логика работы NSF

Networking For everyone

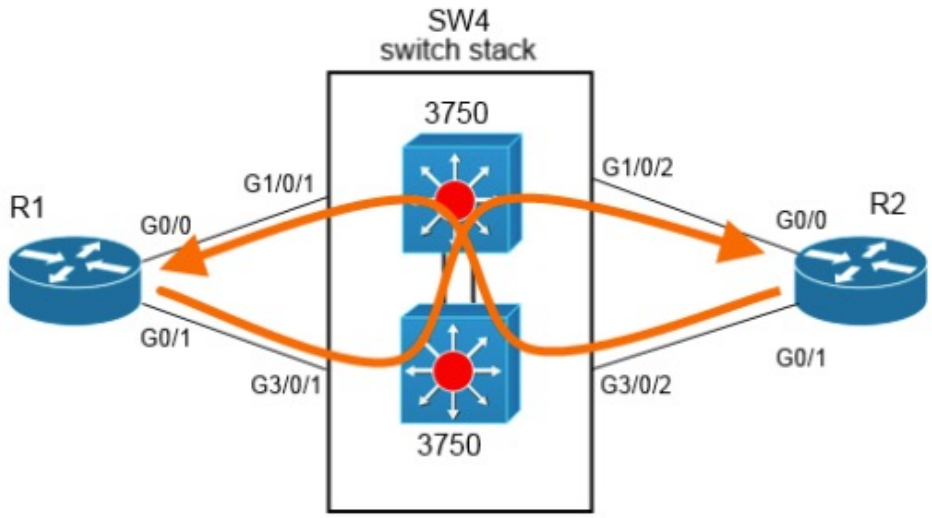
- В случае события переключения SUP (или RP switchover), новый RP перехватывает управление устройством и запускает новый процесс маршрутизации
- Новый процесс OSPF пытается восстановить соседство (adjacency) со всеми предыдущими соседями
- Если NSF настроен для прокола, CEF (FIB) фиксируется (для того, чтобы устройство продолжало пересылку трафика, не смотря на то, что RIB (таблица маршрутизации) пуста
- Во время восстановления соседства, ни одно из устройств не извещает остальную сеть о событии
- Только когда прокол маршрутизации завершает передачу всей необходимой информации, данные попадают в RIB и (оттуда) могут обновить CEF

NSF может таить в себе проблемы

Networking For everyone

- Падает SW4 (1 мембер)
- длительное переключение вызвано фиксацией CEF на время работы процесса NSF

```
%LINK-3-UPDOWN: Interface GigabitEthernet0/0, changed state to down
TACKMGR-4-SWITCH_REMOVED: Switch 1 has been REMOVED from the stack
12:26:03.996: OSPF: IETF NSF complete check for area 0 process 1
OSPF: will poll [count 10] interface status for GigabitEthernet3/0/1
OSPF: Graceful Restart timer expired for process 1, terminating IETF NSF
```



Основные шаги

- Устройство информирует соседей, что процесс OSPF перегружается
  - Graceful restart mode
- Отправляется grace LSA (LSA 9-го типа)
- Соседи отправляют LS ACK и переходят в Helper Mode
- В течение grace периода, соседи работают так, как если бы restarting router не сообщал об изменениях
- В течение перезагрузки OSPF процесса, RIB/FIB не изменяется
- После перезагрузки, переустанавливается OSPF соседство

Основные шаги

- Когда GR завершён, restarting router удаляет Grace LSA
- Пересоздаются все LSA, порождённые маршрутизатором
- Запускается SPF чтобы освежить таблицу маршрутизации

Примечания

- Для работы GR необходима аппаратная поддержка функционала
- Устройства, которые поддерживают Helper Mode = NSF-aware
- Устройства, которые поддерживают GR = NSF-capable

Networking
For everyone

Non-Stop Routing

Проблемы с GR

Networking For everyone

- Все маршрутизаторы должны поддерживать механизм GR для конкретного протокола
- Switchover может быть наиболее неприятным на PE
  - Маленькие CE могут не поддерживать GR
- В некоторых ситуациях GR может замедлить конвергенцию сети

Nonstop Routing, о чём это?

Networking
For everyone

- NSR использует внутренние процессы маршрутизатора для поддержания копии Control Plane на резервном управляющем модуле в актуальном состоянии
- Switchover абсолютно прозрачен для всех соседей

Ограничения

- OSPF NSR может потребовать большого количества памяти
- Переключение между управляющими модулями занимает ~ 2 сек. В течение этого времени OSPF не может отправлять Hello сообщения
  - Аккуратнее с маленькими таймерами

```
Router# show ip ospf 1 nsr
Standby RP
 Operating in duplex mode
 Redundancy state: STANDBY HOT
 Peer redundancy state: ACTIVE
 ISSU negotation complete
 ISSU versions compatible
Routing Process "ospf 1" with ID 10.1.1.100
 NSR configured
 Checkpoint message sequence number: 3290
 Standby synchronization state: synchronized
```

Bidirectional Forwarding Detection

BFD

- BFD – очень легковесный и быстрый протокол, предназначенный для определения нарушения сетевой связности
- BFD работает независимо
- Может работать в двух режимах
  - асинхронный
  - по-требованию (demand)
    - не уверен, что кто-либо из вендоров его реализовал

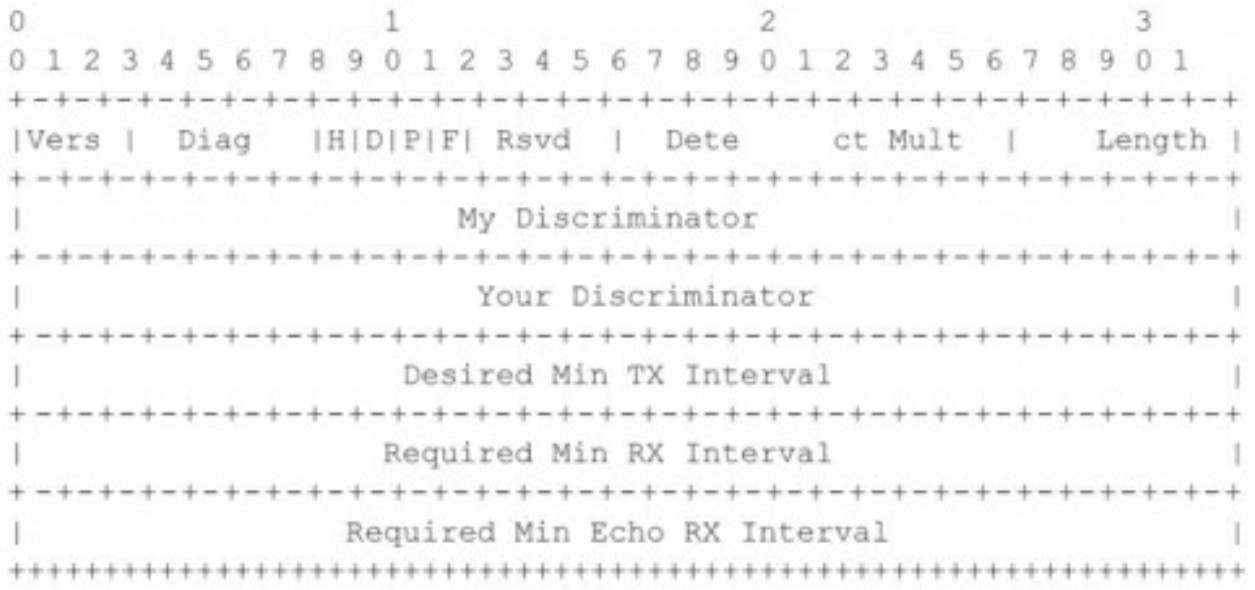
Асинхронный режим

Networking For everyone

- Наиболее классический режим Hello/Holddown
- BFD отправляет Hello. И если не видит сообщений от соседа – регистрируется факт недоступности
- Данный триггер передаётся всем протоколам, которые подписаны на BFD
- Возможно включение функционала Echo
  - Сосед не обрабатывает прилетающие BFD Hello, а просто отправляет их обратно

Формат пакета

- Diag – код, описывающий причину перехода состояния сессии из UP во что-либо другое
- Дискриминатор – мультиплексирование сессий
- Min Interval:
  - desired TX = предпочитаемый интервал
  - required RX = минимально поддерживаемый
- Биты:
  - H = “I Hear You”
  - P = “Poll”
  - F = “Final”



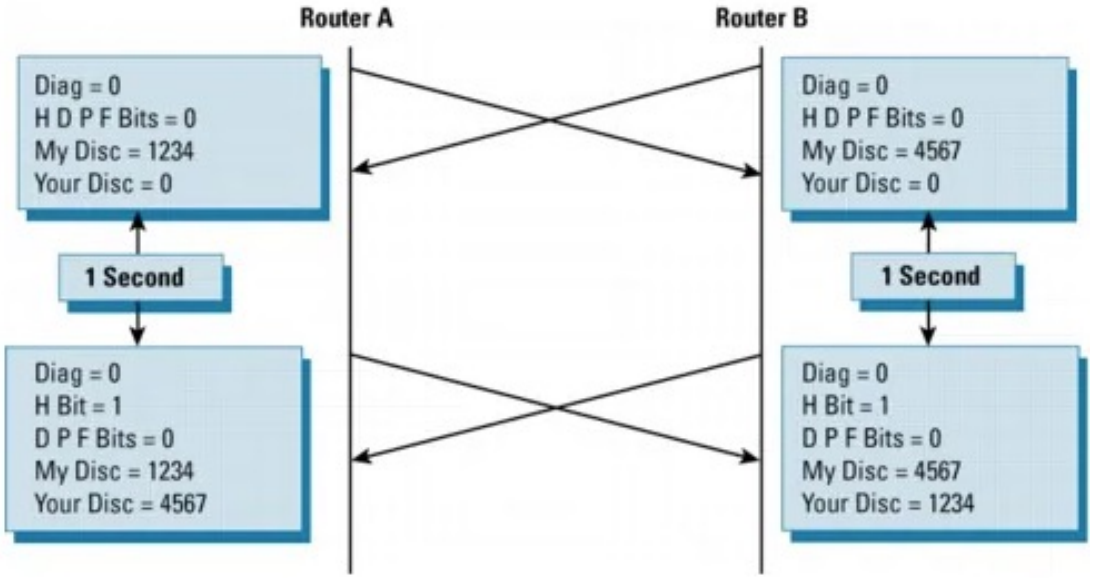
Настройка

```
(config-if)#bfd interval 300 min_rx 600 multiplier 3
```

- interval = как часто устройство отправляет BFD пакеты
- min_rx = как часто ожидаем приём BFD пакетов

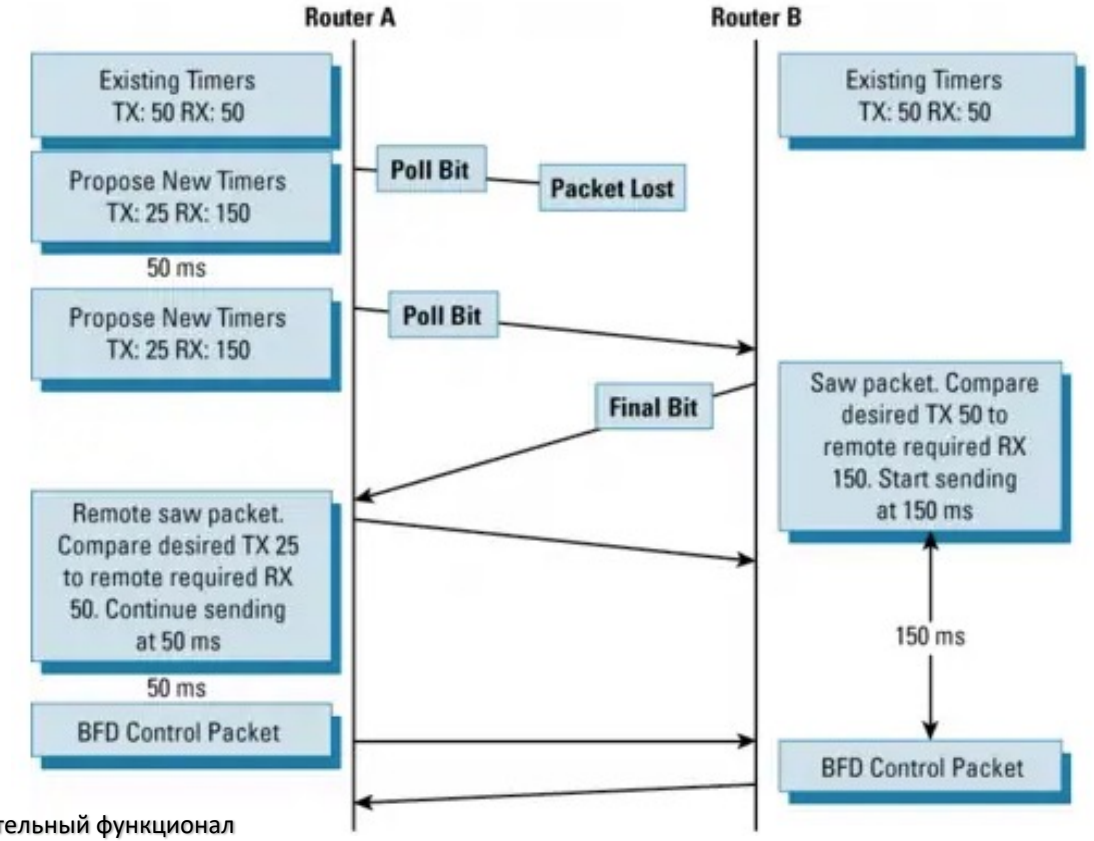
Установление сессии

- BFD не изучает IP адреса соседей
  - их сообщает привязанный протокол (напр. OSPF)
- Все пакеты передаются с помощью UDP
- Зарезервированный Destination Port
- Если видим H бит и Your Disc поле, то сессия установлена



Изменение таймеров

- При изменении таймеров, устройство выставляет P бит
- Если удалённый маршрутизатор увидел P бит, в ответном сообщении выставляется F бит
  - это не подтверждение принятия изменений
- Не требует переустановления



Потеря BFD соседа

- Если не приходит контрольный пакет в течение detect-timer [*(Required Minimum RX Interval) * (Detect Multiplier)*], то сосед помечается как потерянный
- Сам BFD на это никак не реагирует
- Факт потери соседа передаётся привязанному протоколу, который реагирует на данное событие

BFD на модульных платформах

Networking For everyone

- На RSP располагается BFD сервер, а на линейной карте BFD агент
- BFD сервер получает информацию об IP адресах соседей
- BFD агент создаёт все сессии
- Все BFD пакеты отправляются на CPU линейной карты
  - если включен HW Offload, то BFD обрабатывается на Network Processor (NP)
    - позволяет увеличить количество поддерживаемых BFD сессий

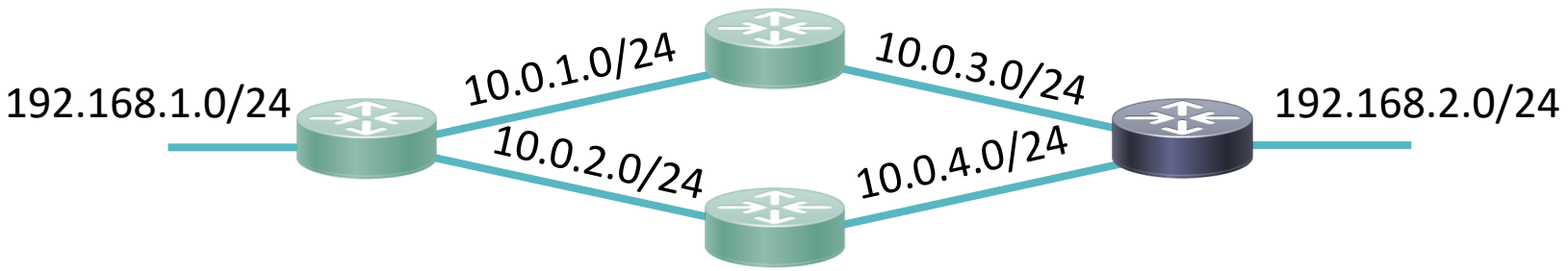
Networking
For everyone

Сокрытие транзитных сетей

Сокрытие транзитных сетей

Networking For everyone

- Штатно в OSPFv2 префиксы транзитных сетей попадают в RIB




```
O 10.0.1.0/24
O 10.0.2.0/24
C 10.0.3.0/24
C 10.0.4.0/24
O 192.168.1.0/24
C 192.168.2.0/24
```

- RFC 6860 позволяет OSPF сократить количество записей в RIB
  - На интерфейсах P2P и P2MP не анонсируются connected-сети
    - В LSA1 создается point-to-point link до RID, но не создается stub-запись для адреса
  - На broadcast и NBMA интерфейсах адреса анонсируются в LSA2
    - В LSA2 отправляется DR IP с маской /32
    - Новые роутеры не устанавливают в RIB маршруты из LSA2 с маской /32

Networking
For everyone

Тупиковый маршрутизатор

Тупиковый маршрутизатор

Networking For everyone

- Иногда (напр. во время проведения миграционных работ) необходимо отвести трафик от маршрутизатора
- Сделать это можно, выставив максимальную метрику для OSPF интерфейсов
- В определённых ситуациях надо дождаться конвергенции BGP

```
R1(config-router)#max-metric router-lsa on-startup ?
  <5-86400>     Time, in seconds, router-LSAs are originated with max-metric
  wait-for-bgp  Let BGP decide when to originate router-LSA with normal metric
```

Networking For everyone