MTU и MSS

подробности, секреты и особенности настройки

MUM Москва 2018

© Routerz 2018

Обо мне

Москалёв Михаил

(Mikhail Moskalev)

сертифицированый тренер [*TR0125*],
MTCNA [*1103NA029*],
MTCRE [*1101RE005*],
MTCWE [1710WE4012],
MTCTCE [1710TCE4013],
MTCUME [1704UME9393],
MTCINE [1612INE037],
MTCIPv6E [1703IPv6E7536]

Компания РОУТЕРЗ.РУ

http://routerz.ru, support@routerz.ru

© Routerz 2018

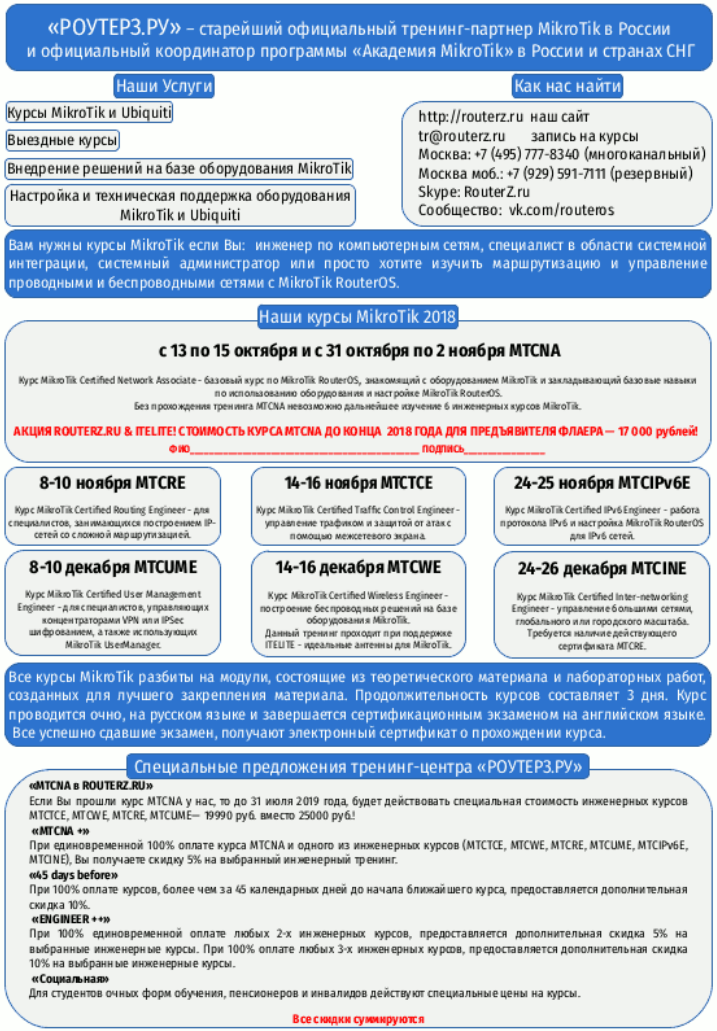
Наши тренинги и акции

Акция РОУТЕРЗ.РУ & ITElite

Стоимость курса MTCNA до конца 2018 года для предъявителя флаера

17 000 рублей

Флаеры можно получить на стенде ITElite



«РОУТЕРЗ.РУ» – старейший официальный тренинг-партнер MikroTik в России и официальный координатор программы «Академия MikroTik» в России и странах СНГ

Наши Услуги

Курсы MikroTik и Ubiquiti

Выездные курсы

Внедрение решений на базе оборудования MikroTik

Настройка и техническая поддержка оборудования MikroTik и Ubiquiti

Как нас найти

http://routerz.ru наш сайт
tr@routerz.ru запись на курсы
Москва: +7 (495) 777-8340 (многоканальный)
Москва моб.: +7 (929) 591-7111 (резервный)
Skype: RouterZ.ru
Сообщество: vk.com/routeros

Вам нужны курсы MikroTik если Вы: инженер по компьютерным сетям, специалист в области системной интеграции, системный администратор или просто хотите изучить маршрутизацию и управление проводными и беспроводными сетями с MikroTik RouterOS.

Наши курсы MikroTik 2018

с 13 по 15 октября и с 31 октября по 2 ноября MTCNA

Курс MikroTik Certified Network Associate - базовый курс по MikroTik RouterOS, знакомящий с оборудованием MikroTik и закладывающий базовые навыки по использованию оборудования и настройке MikroTik RouterOS.
Без прохождения тренинга MTCNA невозможно дальнейшее изучение 6 инженерных курсов MikroTik.

АКЦИЯ ROUTERZ.RU & ITELITE! СТОИМОСТЬ КУРСА MTCNA ДО КОНЦА 2018 ГОДА ДЛЯ ПРЕДЪЯВИТЕЛЯ ФЛАЕРА — 17 000 рублей!
ФИО_____ ПОДПИСЬ_____

8-10 ноября MTCRE
Курс MikroTik Certified Routing Engineer - для специалистов, занимающихся построением IP-сетей со сложной маршрутизацией.

14-16 ноября MTCTCE
Курс MikroTik Certified Traffic Control Engineer - управление трафиком и защитой от атак с помощью межсетевого экрана.

24-25 ноября MTCIPv6E
Курс MikroTik Certified IPv6 Engineer - работа протокола IPv6 и настройка MikroTik RouterOS для IPv6 сетей.

8-10 декабря MTCUME
Курс MikroTik Certified User Management Engineer - для специалистов, управляющих концентраторами VPN или IPSec шифрованием, а также использующих MikroTik UserManager.

14-16 декабря MTCWE
Курс MikroTik Certified Wireless Engineer - построение беспроводных решений на базе оборудования MikroTik.
Данный тренинг проходит при поддержке ITELITE - идеальные антенны для MikroTik.

24-26 декабря MTCINE
Курс MikroTik Certified Inter-networking Engineer - управление большими сетями, глобального или городского масштаба.
Требуется наличие действующего сертификата MTCRE.

Все курсы MikroTik разбиты на модули, состоящие из теоретического материала и лабораторных работ, созданных для лучшего закрепления материала. Продолжительность курсов составляет 3 дня. Курс проводится очно, на русском языке и завершается сертификационным экзаменом на английском языке. Все успешно сдавшие экзамен, получают электронный сертификат о прохождении курса.

Специальные предложения тренинг-центра «РОУТЕРЗ.РУ»

«MTCNA в ROUTERZ.RU»
Если Вы прошли курс MTCNA у нас, то до 31 июля 2019 года, будет действовать специальная стоимость инженерных курсов MTCTCE, MTCWE, MTCRE, MTCUME— 19990 руб. вместо 25000 руб.!

«MTCNA +»
При единовременной 100% оплате курса MTCNA и одного из инженерных курсов (MTCTCE, MTCWE, MTCRE, MTCUME, MTCIPv6E, MTCINE), Вы получаете скидку 5% на выбранный инженерный тренинг.

«45 days before»
При 100% оплате курсов, более чем за 45 календарных дней до начала ближайшего курса, предоставляется дополнительная скидка 10%.

«ENGINEER +»»
При 100% единовременной оплате любых 2-х инженерных курсов, предоставляется дополнительная скидка 5% на выбранные инженерные курсы. При 100% оплате любых 3-х инженерных курсов, предоставляется дополнительная скидка 10% на выбранные инженерные курсы.

«Социальная»
Для студентов очных форм обучения, пенсионеров и инвалидов действуют специальные цены на курсы.

Все скидки суммируются

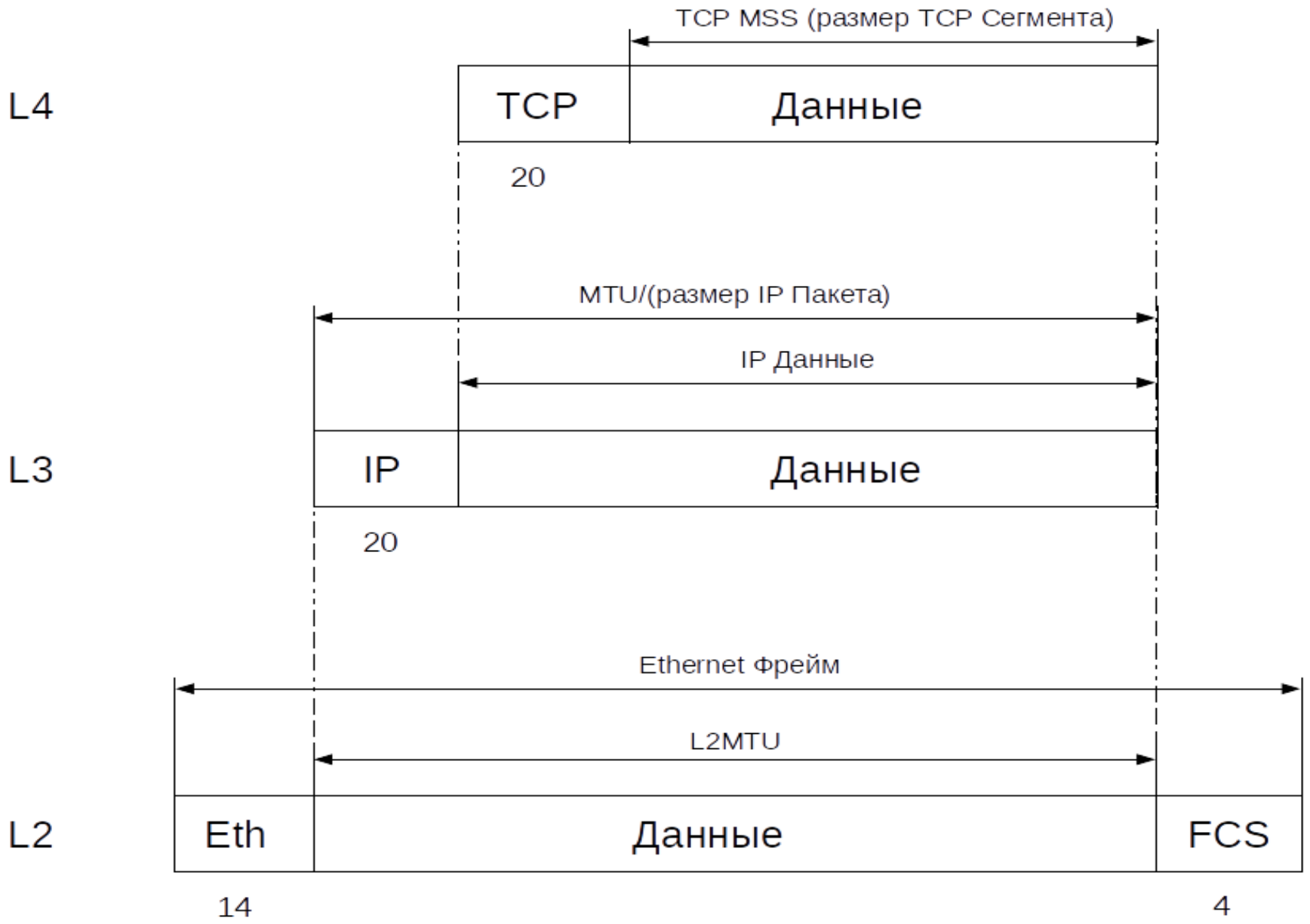
© Routerz 2018

L2MTU, MTU, MRU, TCP MSS

- L2MTU максимальный размер данных Ethernet (L2) (включая дополнительные заголовки VLAN и MPLS)
- MTU (Maxmum Transmit Unit) — Максимальный размер IP пакета которым может быть передан
- MRU (Maxmum Recieve Unit) — Максимальный размер IP пакета которым может быть принят. Встречается в PPP тунелях, в Ethernet любой успешно полученный пакет будет обработан.
- TCP MSS (Maximum Segment Size) — максимально возможный размер сегмента TCP

© Routerz 2018

L2MTU, MTU, MRU, TCP MSS



© Routerz 2018

L2MTU, MTU, MRU, TCP MSS

© Routerz 2018

Причины снижения MTU

- Различия в оборудовании
  - Разные модели поддерживают разный максимальный размер L2MTU
    https://wiki.mikrotik.com/wiki/Manual:Maximum_Transmission_Unit_on_RouterBoards#MAC.2FLayer-2.2FL2_MTU
- Использование VPN тунелей
  - Во многих случаях снижает MTU
  - Например IPIP туннель снижает MTU на 20 байт
- Различная среда передачи. Например: Ethernet 1000, Ethernet 100, Serial PPP

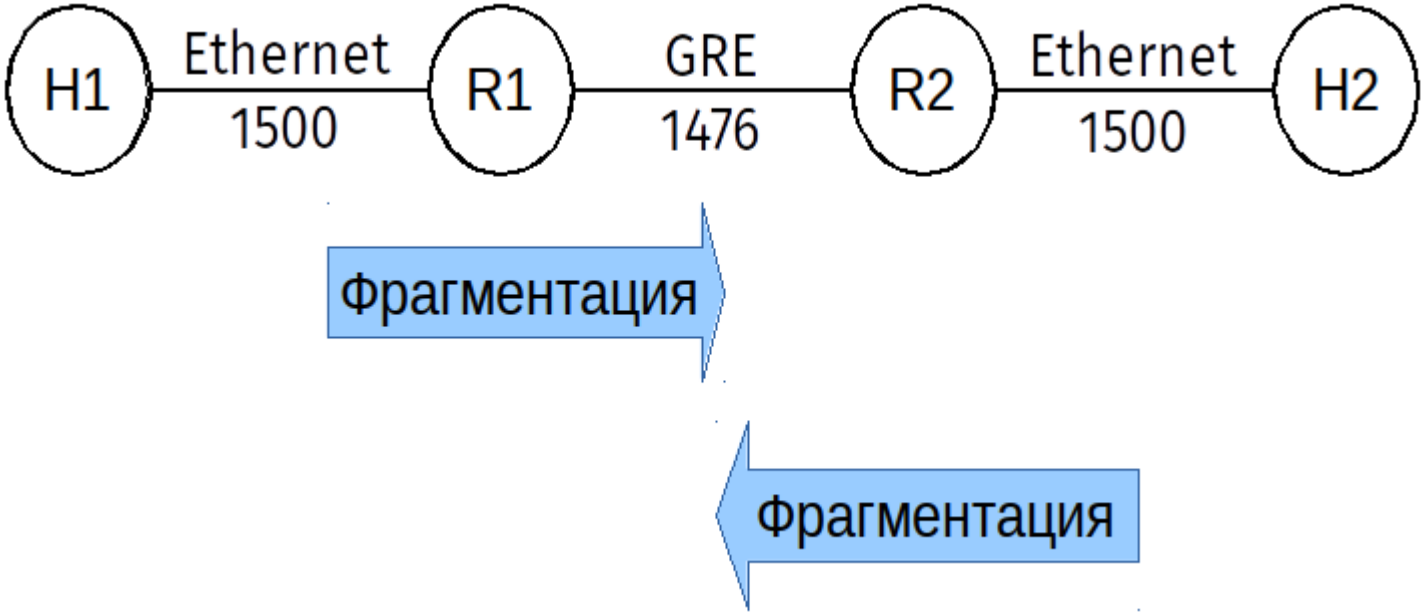
© Routerz 2018

IP фрагментация

- IP протокол предусматривает возможность фрагментации пакетов превышающих размер MTU
- В IP заголовке присутствуют флаг наличия дополнительных фрагментов, флаг запрета фрагментации (DF), смещение фрагмента и идентификатор исходного пакета.

© Routerz 2018

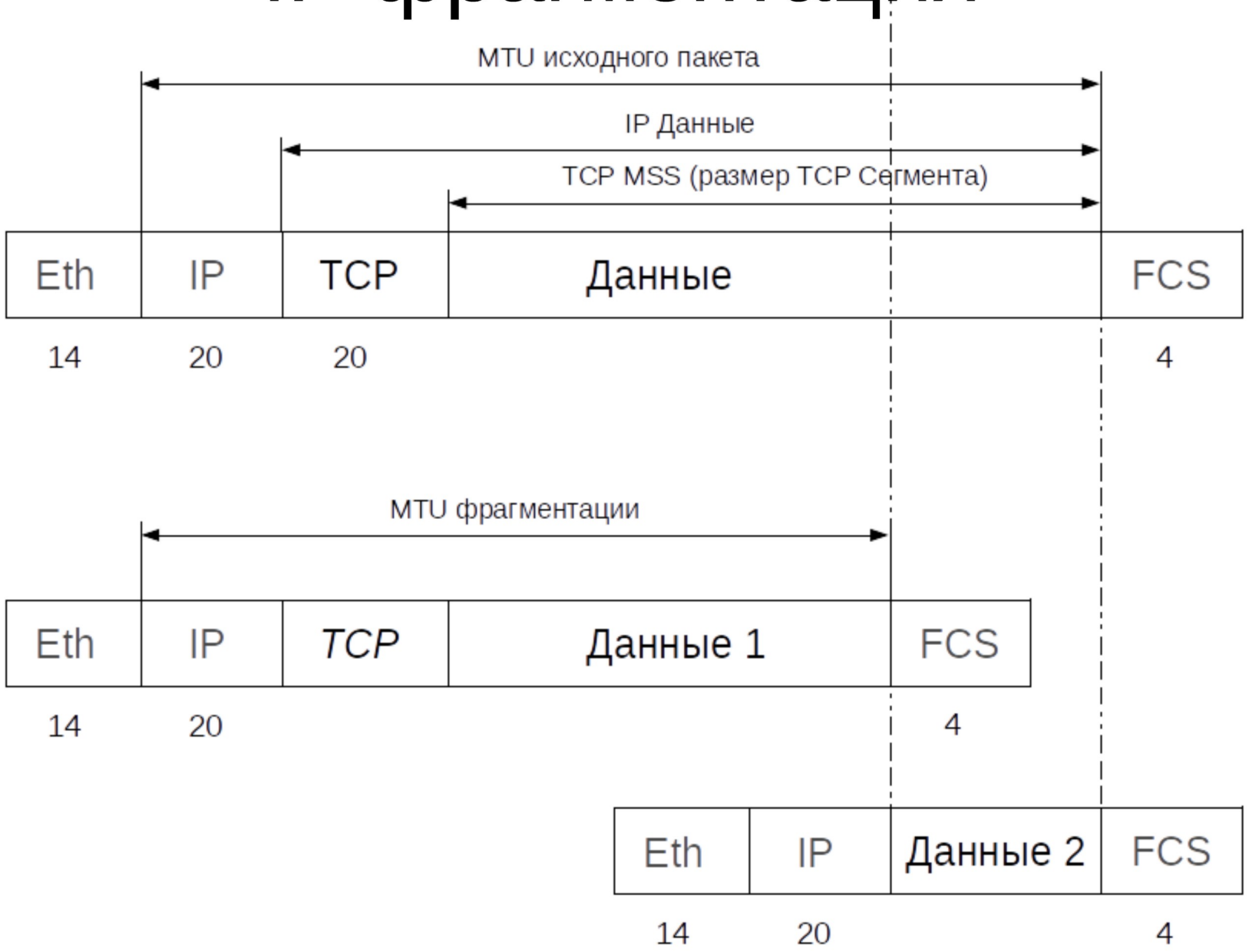
IP фрагментация



Чтобы передать пакет большего размера через интерфейс у которого MTU меньше, пакет может быть разбит на фрагменты.
При условии что это не запрещено флагом DF (Don`t Fragment)

© Routerz 2018

IP фрагментация



© Routerz 2018

IP фрагментация

В случае, если фрагментация пакета запрещена флагом DF *(в протоколе IPv6 фрагментация разрешена только на узле отправителе),* маршрутизатор уничтожает пакет в возвращает ICMP ошибку 3:4 *(«Необходима фрагментация, но установлен флаг её запрета DF»)*

© Routerz 2018

IP фрагментация

Сборкой фрагментов в протоколе IP должен заниматься хост получатель.

Чтобы это было возможно в заголовке IP предусмотрены поля Identification, Flags, Fragment Offset. https://ru.wikipedia.org/wiki/IPv4#Структура_заголовка_пакета

© Routerz 2018

Тестирование MTU

Используя это можно протестировать прохождение пакетов заданного размера (MTU) без фрагментации до целевого адреса.

- Linux:
  `ping -M do <addr> -s <MTU>`
- ROS:
  `:ping <addr> do-not-fragment size=<MTU>`
- Windows:
  `ping <addr> -f -l <MTU-28>`

© Routerz 2018

Тестирование MTU

- Флаг -l в Windows версии ping указывает не MTU а размер ICMP Payload. Используемый MTU будет больше на размер заголовков IP (20) + ICMP (8) = 28 байт.
- Указанные тесты довольно полезны для проверки актуального MTU, однако DF флаг пакета может быть модифицирован на транзитных маршрутизаторах.

© Routerz 2018

Недостатки IP фрагментации

- Несмотря на наличие полей обеспечивающих возможность сборки, сборка IP фрагментов всё равно потребляет ресурсы процессора и памяти

© Routerz 2018

Недостатки IP фрагментации

- Проблема с потерями пакетов

  В случае потери фрагмента, сборка становится невозможной. Несобранные фрагменты застревают в буфере на таймаут дефрагментации, а хост отправитель отправляет весь пакет заново. И он заново разбивается на фрагменты.

- Дефрагментация выполняется одним ядром процессора, на CCR с большим количеством ядер это становится узким местом.

© Routerz 2018

Недостатки IP фрагментации

- Фрагментация требует установки в IP заголовке поля Identifification в уникальное значение для каждой комбинациии SrcIP-DstIP-Protocol, в пределах таймаута дефрагментации. Строгое следование стандарту (RFC0791) существенно ограничивает максимальную пропускную способность. (RFC4963 приводит оценку в 26Mbit/s)
- Как правило за этим строго не следят, чтобы не ограничивать пропускную способность, но это может приводить к ошибкам при дефрагментации.

© Routerz 2018

Особенность работы файрвола с фрагментами

- В файрволе при активном ConnectionTracker наблюдать фрагменты пакетов можно только в RAW таблице

```
/ip firewall raw
add action=log log-prefix="raw-fragment" \
chain=prerouting fragment=yes

/ip firewall mangle
add action=log chain=prerouting fragment=yes \
log-prefix=mangle-pre-fragment
```

- Если правило в raw prerouting может обнаружить принимаемые фрагменты, то в mangle prerouting, выполняемым после Connection Tracker, уже нет.

© Routerz 2018

Механизмы управления MTU

- ICMP — В случае если фрагментация пакета запрещена флагом DF *(в протоколе IPv6 фрагментация разрешена только на узле отправителе)* маршрутизатор уничтожает пакет в возвращает ICMP ошибку 3:4 *(«Необходима фрагментация, но установлен флаг её запрета DF»)*

© Routerz 2018

Path MTU Discovery

- Path MTU — максимальный MTU который может быть доставлен от отправителя до получателя без фрагментации
- Path MTU Discovery (RFC1191)

  В случае получения ICMP сообщения о необходимости фрагментации, хост отправитель снижает MTU (и соответственно MSS) для пакетов отправляемых этому хосту.

© Routerz 2018

MSS

- Протокол TCP имитирует для приложения двунаправленный поток данных.

  Для передачи потока через пакетную сеть, поток разбивается на сегменты.

  Размер сегмента может меняться в зависимости от внутренних алгоритмов TCP (наличия данных для передачи, размера окна передачи ...)

  Размер сегмента ограничен сверху значением MSS.

© Routerz 2018

MSS

- При установлении TCP соединения клиент и сервер обмениваются значениями MSS, максимальный размер сегмента который они могут принять.

  В результате для TCP соединения устанавливаются максимальные размеры сегмента для передачи в обе стороны.

  * https://tools.ietf.org/html/rfc879

© Routerz 2018

Path MTU Blackhole Detection

- Если какой либо из транзитных маршрутизаторов блокирует ICMP сообщения о необходимости фрагментации (3:4) то в случае необходимости фрагментации возникает MTU Blackhole (чёрная дыра), когда пакет большого размера удаляется, но сообщение об ошибке не приходит.

© Routerz 2018

Path MTU Blackhole Detection

- Существует механизм, детектирующий подобную ситуацию и реализующий обход этой проблемы.
- Детектирование не достоверно
- Проблема решается автоматической установкой MTU в фиксированное минимальное значение (576 байт) для проблемных соединений. Такой маленький MTU сам по себе снижает производительность.
- По этим причинам, PMTUBD по умолчанию отключена в современных ОС

© Routerz 2018

Настройка MTU на интерфейсе

- Возможно снизить MTU для интерфейса.
  - Это снижает MTU при работе со всеми узлами
  - У всех узлов в одном L2 сегменте желательно иметь одинаковый MTU
  - Для IPv4
    - DHCP oprion 26 позволяет задать MTU но например Windows клиенты не используют эту опцию
    - Можно использовать только системы управления конфигурацией (Domain gropu policy, ansible, и т. п.)
  - Для IPv6
    - ICMP Router Advertisement позволяет задать MTU централизовано в случае использования SLAAC

© Routerz 2018

L3 уровень

- роутер может изменя значение MSS в передаваемых пакетах, влияя на то какого размера пакеты в дальнейшем будут передавать хосты.
  - Используя mangle action=change-mss
  - Исползуя параметр PPP профиля change-tcp-mss=yes
  - Используя парамтер тунеля clamp-tcp-mss=yes
- В результате TCP протокол будет разбивать поток на сегменты не превышающие MSS и передаваемые IP пакеты не будут требовать IP фрагментации

© Routerz 2018

L2 уровень

- Для L2 уровня нет способа сообщить отправителю что фрейм слишком большой.
- Слишком большой фрейм будет потерян.
- В большинстве комутаторов, даже поддерживающих Jumbo Frame (до 9000 байт и более) по умолчанию эта возможность выключена. (У Mikrotik L2MTU лишь немного больше 1500. У HP ProCurve поддержка Jumbo Frame выключена)

© Routerz 2018

L2 уровень

- В случае потери больших пакетов, помочь может только PMTU Blackhole Detection, но реализация этого алгоритма различается в разных ОС.

© Routerz 2018

MTU, L2MTU и бриджинг

- Если физические порты роутера собраны в бридж, то для успешного L2 форварда пакета, его L2MTU не должен превышать L2MTU входящего порта, исходящего порта и самого бриджа.
- MTU самого бриджа, влияет только на работу IP стека на бридж-интерфейсе.

© Routerz 2018

MTU, L2MTU и бриджинг

- Во многих версиях ROS (включая текущую 6.42.6) L2MTU бриджа автоматически устанавливается равной наименьшему из L2MTU интерфейсов включенных в бридж
- Это приводит к очень интересному сценарию отказа, когда локальная сеть построена на бридже (Default config для многих роутеров с Wi-Fi). Построение EoIP тунеля с другим роутером, для создания L2-VPN вызывает прекращение нормальной работы локальной сети и интернет. Поскольку пакеты разпером 1500 перестают проходить через бридж.

© Routerz 2018

L2 уровень

- Например чтобы пропустить пакеты большого размера через Mikrotik используемые в качестве коммутаторов, нужно было на всех портах увеличить L2MTU

```
{:local maxl2mtu
/interface ethernet
:foreach i in=[find] do={
  :put [get $i name]
  :set maxl2mtu [get $i max-l2mtu]
  set $i l2mtu=$maxl2mtu
  :put "$[get $i name] l2mtu set to $[get $i l2mtu]" }
/log info "MTU maximized"}
```


© Routerz 2018

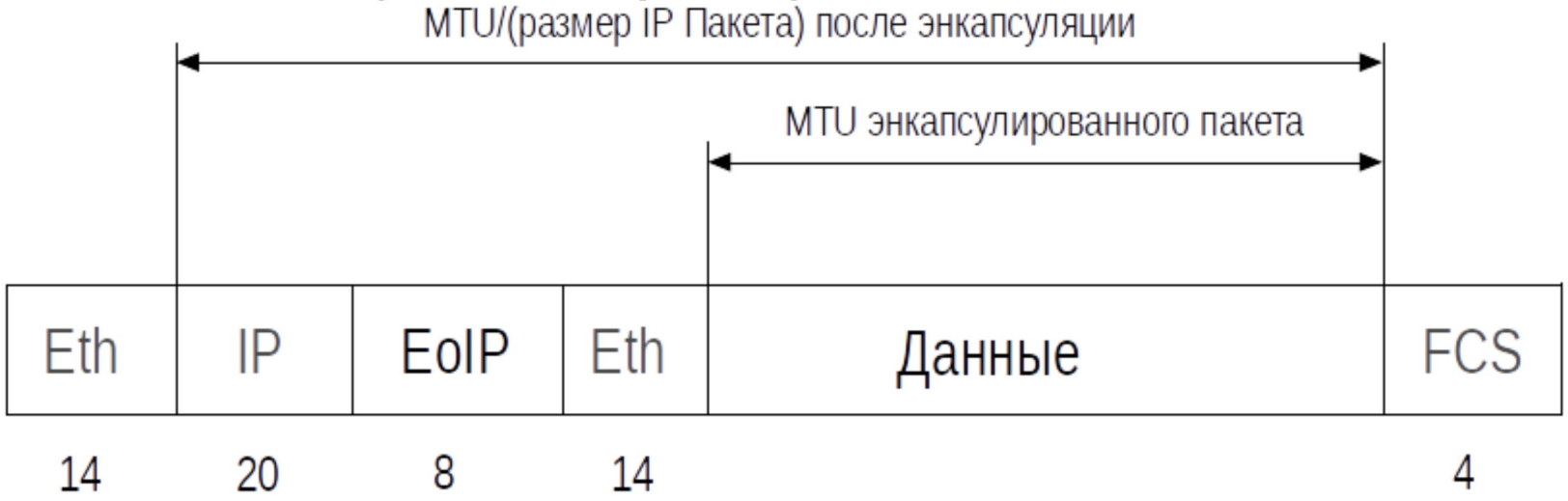
Альтернативные механизмы фрагментации

- IP фрагментация
- TCP Segment
- ML-PPP (MRRU) — изначально создано для балансировки трафика, но позволяет передавать кадры любого размера разбивая их на части.
- MPLS/VLPS — VPLS (L2 тунель через MPLS) позволяет фрагментировать кадры, но это тоже создает накладные расходы.

© Routerz 2018

MTU и MSS в случае L2-VPN

- EoIP:
  - Overhead 42 байта = IP(20)+EoIP(8)+Ehternet(14)
  - MTU внутри тунеля по умолчанию автоматически выставляется как <MTU транспорта> - 42 и clamp-tcp-mss=yes , что в простых случаях даёт оптимальную настройку.



© Routerz 2018

MTU и MSS в случае L2-VPN

- EoIP/IPSec:
  - SPI(4) +Seq(4) +IV(16) +Pad(8) +Padlen(1) +NextProtocol(2) +Auth(12) = 47

    +Overhead 42 байта (= IP(20) +EoIP(8) +Ehternet(14))

    Размер полей ESP зависит от используемых алгоритмов (пример для AES-CBC 256 + Sha1) и блок данных дополняется до размера кратного размеру блочного шифра (в примере до 128байт)

© Routerz 2018

MTU и MSS в случае L2-VPN

- Автоматический рассчёт MTU учитывает MTU интерфейса отправки, но не учитывает возможное снижение MTU на промежуточных узлах. Что может привести к возникновению фрагментации и снижению производительности.

© Routerz 2018

Снижение MTU вручную

- Реальный пример: Строится PPTP VPN через интернет.
  - С обеих сторон провайдеры подкючены по Ethernet с MTU по умолчанию 1500
  - MTU тунеля автоматически назначается в 1450 IP(20) +GRE-PPP(16) +PPP(4)+Encryption(10)
  - Скорость интернет по тарифу 10 Мбит/с
  - Внутри тунеля в однопоточном TCP тесте получается скорость меньше 500 Кбит/с.
  - Тест показывает что Path MTU = 1480
  - Дополнительно уменьшив MTU до 1400 (с запасом) в тунеле удается повысить скорость TCP соединения до 5 Мбит/с

© Routerz 2018

Спасибо за внимание

Вопросы?

Ищите меня на стенде нашего партнёра ITElite

Жду ваших комментариев на почту
support@routerz.ru
и в группе
vk.com/routeros

RouterZ

ITELITE
ANTENNAS

© Routerz 2018