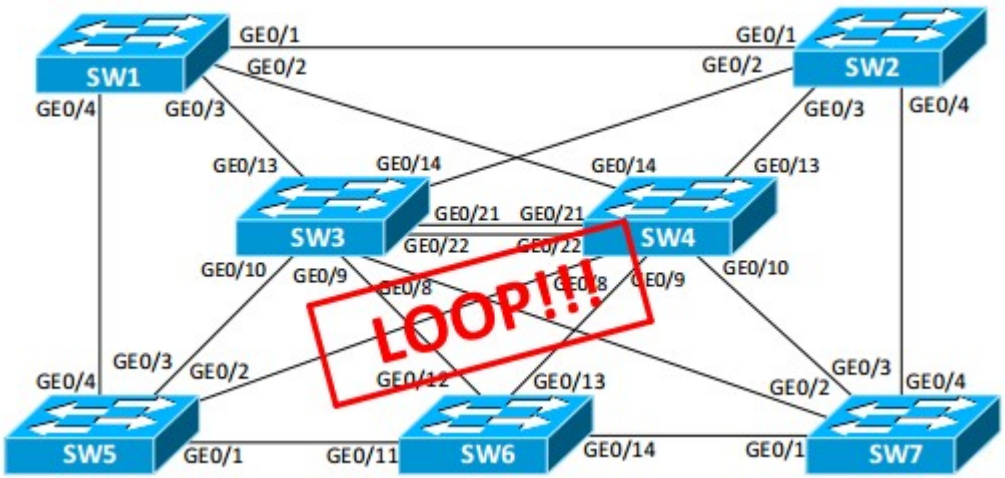
habr.com

Rapid STP

14-20 минут



Протоколы семейства STP обычно несильно будоражат умы инженеров. И в большинстве своём на просторах интернета чаще всего сталкиваешься с деталями работы максимум протокола STP. Но время не стоит на месте и классический STP всё реже встречается в работе и в различных материалах вендоров. Возникла идея сделать небольшой обзор ключевых моментов **RSTP** в виде FAQ. Всем, кому интересен данный вопрос, прошу под кат.

Что настраивать STP, RSTP или MST?

В современных стандартах протокол STP уже нигде не фигурирует. Известный всем ***802.1d*** в последней редакции (802.1d-2004) описывает протокол RSTP. При этом MST перекочевал в ***802.1q*** (802.1q-2014). Как мы помним, ранее

RSTP описывался стандартом 802.1w, а MST — 802.1s.

RSTP и MST имеют существенно меньшее время сходимости. Они намного быстрее перестраивают топологию сети в случае отказа оборудования или каналов связи. Время сходимости для ряда отказов этих протоколов меньше 1 секунды против 30+ секунд в случае STP. Поэтому классический STP рекомендуется использовать только там, где задействуется старое оборудование, не поддерживающее более современные протоколы.

MST в своей работе использует алгоритмы RSTP. Но в отличие от RSTP, MST позволяет создавать отдельную топологию (instance) STP для группы VLANов. В случае обычного RSTP у нас на все VLANы одна общая топология. Это не очень удобно, так как не позволяет даже в ручном режиме балансировать трафик по разным каналам. А значит, мы теряем, как минимум половину пропускной способности в случае наличия избыточных путей.

Некоторые вендоры (в частности Cisco) предлагают ещё одну разновидность быстрого протокола STP – Rapid Per-VLAN Spanning Tree (PVRST+). В этом случае для каждой виртуальной сети строится своя топология, что позволяет более эффективно утилизировать каналы. Основной минус такого подхода – это ограничение на максимальное количество таких топологий. Для обеспечения работы каждой топологии устройство тратит аппаратные ресурсы. А они не безграничны. Например, в коммутаторах Cisco 2960 поддерживается максимум 128 «инстансов» STP.

Таким образом, MST является хорошей альтернативой между

стандартным RSTP и проприетарным PVRST+. Особенно если наша сеть построена на базе коммутаторов разных производителей. Стоит заметить, что все три вариации быстрого STP совместимы друг с другом.

В дальнейшем, упоминая RSTP, мы будем подразумевать в том числе и его расширения MST/PVRST+.

Какие технологии обеспечивают быстроту реакции в работе RSTP?

RSTP в первую очередь опирается на работу механизмов, не привязанных к стандартным таймерам. Именно поэтому он позволяет получить существенно меньшее время сходимости сети. Можно выделить следующие улучшения в работе RSTP по сравнению с классическим STP:

1. Генерация BPDU сообщений каждым устройством независимо от корневого коммутатора.
   В классическом варианте BPDU «генерит» в сети только корневой коммутатор. Все остальные устройства лишь ретранслируют его. Таким образом, отсутствие BPDU от вышестоящего устройства значит, что проблема может быть в любом месте между данным устройством и корневым коммутатором. Поэтому приходилось ждать достаточно долго (MaxAge=20 сек) прежде чем, смириться с тем, что что-то пошло не так и нужно перестраивать топологию.

   В случае RSTP сообщения BPDU стали выполнять роль Hello-пакетов. Теперь потеря трёх таких пакетов (а это 2*3=6 сек) означает, что пора задуматься об изменениях в топологии.

2. Механизм Proposal/Agreement при активации соединения точка-точка на не Edge-портах для быстрого перехода в состояние

передачи (Forwarding).

Примечание

RSTP предусматривает два типа соединения:

- точка-точка (**point-to-point**) – к такому порту подключен только один RSTP коммутатор;
- общий (**shared**) — к такому порту подключено несколько RSTP коммутаторов (через хаб, что в наше время является большой экзотикой).

Обычно коммутаторы определяют тип соединения автоматически, опираясь на режим передачи full-duplex (точка-точка) или half-duplex (общий). В случае соединения shared «фишечки» RSTP перестают работать.

В классическом STP порт, который должен стать корневым, проходит все стадии по переходу в режим передачи (Listening → Learning → Forwarding), что занимает более 30 секунд.
Прежде чем коснуться механизма Proposal/Agreement, нужно отметить два разных типа портов в RSTP: пограничный порт (**Edge port**) и не пограничный (**non-Edge port**). В Edge порт подключаются оконечные устройства (ПК, серверы, в ряде случаев маршрутизаторы и пр.). В не Edge-порт подключаются другие коммутаторы, участвующие в топологии STP.

Примечание

Тип порта Edge задаётся вручную. Коммутатор не может быстро определить, кто к нему подключен: обычный хост или коммутатор. Конечно, он мог бы ориентироваться на наличие BPDU на этом порту. Но по стандарту коммутатор должен

обязательно подождать минимум 15 секунд (Forward delay) прежде, чем решить, что на его порт так и не пришло ни одно сообщение. А это слишком долго. Поэтому право определить, что подключено к порту, доверили человеку.

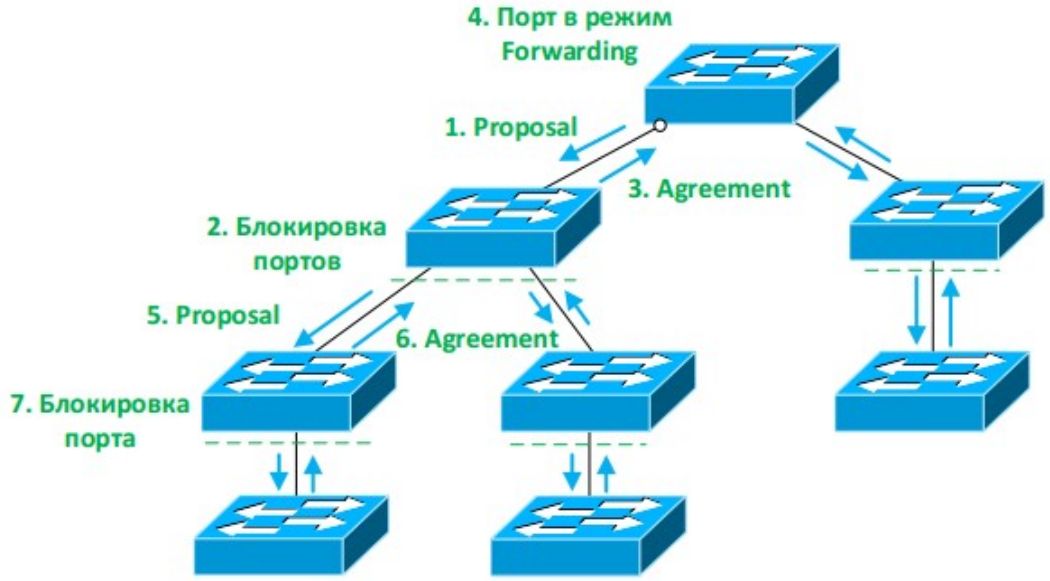
На коммутаторах Cisco тип порта Edge задаётся командой *spanning-tree portfast*.

RSTP использует механизм Proposal/Agreement для быстрого переходя портов из состояния Discarding в состояние Forwarding. Этот механизм запускается, когда у коммутатора меняется Root Port (как минимум при включении в сеть). В этом случае он выключает все порты, не являющиеся Edge-портами. Об этом оповещает вышестоящий коммутатор (куда как раз смотрит Root port), после чего включает в режим Forwarding только Root port. Остальные порты (не Edge) находятся в заблокированном состоянии, пока не произойдёт одно из двух:

- коммутатор обменяется сообщениями Proposal/Agreement с нижестоящим коммутатором,
- истекут таймеры перехода из стояния Discarding в Learning (15 сек) и из Learning в Forwarding (15 сек), если подключённое оборудование не поддерживает RSTP.

Сообщение Proposal отправляется портом, который хочет стать назначенным (Designated) для данного сегмента сети. И фактически запускает механизм Proposal/Agreement. Сообщение Agreement отправляется портом, который становиться корневым (Root). Оно необходимо для оповещения вышестоящего устройства о возможности немедленно

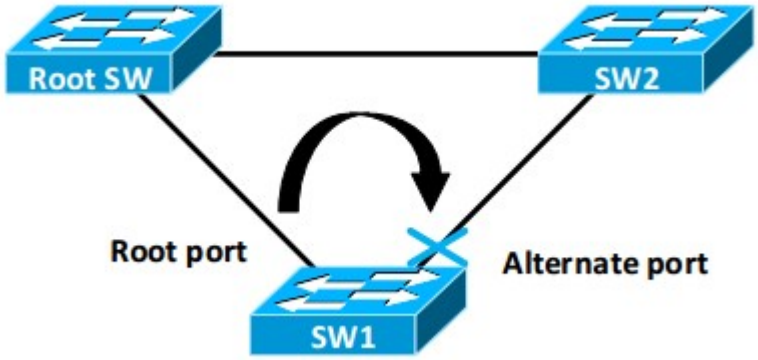
перевести порт в состояние Forwarding.



3. Механизм альтернативного порта при потере связи через корневой порт.
   RSTP отличается от STP тем, что состояние порта отвязали от его роли. Это позволило описать роль порта в топологии сети без оглядки на его состояние. А значит, обладать лучшим видение топологии сети и возможностью оперативно реагировать на изменения в ней. Так появились альтернативный (alternative) и резервный (backup) порты. Альтернативный порт – замена корневому. Через него может быть достигнут корневой коммутатор, но при этом данный порт не имеет роли корневого (т.е. получает BPDU с худшей метрикой) и не является назначенным (т.е. не является лучшим в данном сегменте сети для достижимости корневого устройства).

   В протоколе RSTP альтернативный порт переходит в состояние передачи сразу же после того, как откажет корневой. Такого же поведения можно добиться в классическом STP, используя

проприетарные доработки. Например, Cisco предлагает для этих целей технологию UplinkFast.



4. Механизм немедленной реакции на получение BPDU с информацией о «худшем» корневом коммутаторе от соседа, имеющего назначенный (designated) порт для данного сегмента сети.
   В такой ситуации, если у устройства есть другой маршрут к корневому коммутатору, в классическом STP порт, который ранее был заблокирован, пройдёт все стадии и переключится в режим передачи только через 50 секунд (MaxAge + 2x Forward Delay).

   В случае RSTP коммутатор немедленно оценит полученный BPDU (в RSTP нет MaxAge таймера) и начнёт передавать свои, выставив флаг Proposal. Получив такое BPDU, коммутатор, потерявший связь с «рутом», примет участие в механизме Proposal/Agreement, так как у него сменился корневой порт. А дальше достаточно оперативно порты на обоих коммутаторах перейдут в состояние передачи.

5. Улучшенная схема рассылки и обработки сообщений TCN BPDU об изменениях в сети.
   Классический STP считает, что топология изменилась, если порт перешёл из состояния заблокированный в состояние передачи или наоборот. Так как изменение топологии может привести к тому, что MAC адреса станут доступны через другие порты (а значит, коммутатор будет слать пакеты не туда), запускается процедура оповещения всех устройств о таком событии. Для этого рассылается сообщение Topology Change Notification (TCN). Получив которое, коммутатор меняет время старения MAC адресов со значения по умолчанию (300 сек) на 15 сек (Forward Delay). Сообщение TCN рассылается в два этапа. Сначала коммутатор, обнаруживший изменения в топологии, отправляет его в сторону корневого коммутатора. Далее корневой коммутатор, получив такое сообщение, узнаёт об изменении в сети и рассылает TCN сообщение (BPDU с соответствующим флагом) уже всем остальным. Двухуровневая схема необходима, так как BPDU в классическом варианте отправляется только корневым коммутатором.

   В случае RSTP изменением в топологии считается только переход порта в режим передачи. Причём учитываются порты, которые не являются пограничным (non-edge port). Это и логично, так как переход порта в заблокированное состояние автоматически делает MAC адреса за ним больше не

доступными. Как только обнаружено изменение топологии, коммутатор рассылает через все порты (корневой и назначенные) BPDU с флагом TC. Такое сообщение быстро распространяется по сети. Получив его, коммутаторы удаляют из таблицы все MAC адреса доступные через не edge порты, за исключением того, где был получен BPDU с флагом TC.

Edge порт никогда не вызывает изменений в топологии, а также для такого порта не сбрасываются MAC адреса в случае получения BPDU с флагом TC.

Почему RSTP иногда «тормозит» и переводит порт в режим передачи трафика только через несколько десятков секунд?

RSTP в своей работе использует обычные таймеры в следующих случаях:

- К порту подключается устройство, поддерживающее только классический STP. В этом случае порт коммутатора из режима работы RSTP переходит в режим STP. А значит проходит все стандартные для STP стадии: Blocking, Listening, Learning, Forwarding (в зависимости от того, с какими значениями корневого коммутатора и стоимости будет получен BPDU).
- Коммутатор определил соединение как shared. В этом случае используются стандартные таймеры RSTP. Так как к такому порту подключено два и более коммутатора, задействовать механизм Proposal/Agreement невозможно.
- Подключено устройство, не участвующее в STP, и на порту не задан тип edge. В этом случае порт пройдёт следующие стадии RSTP: Discarding (15 секунд), Learning (15 секунд), Forwarding. Таким образом, он перейдёт в режим передачи только через 30

секунд.

Почему задание типа порта Edge более важно для RSTP, чем для STP?

Деление на порты Edge и non-Edge характерно не только для RSTP, но и для STP. Но в случае STP – это вендорная доработка протокола, нежели требования стандарта.

Основные «ЗА» включения на порту режима Edge (для оборудования Cisco – это portfast) в случае использования протокола STP:

- Быстрое переключение порта в режим передачи (Forwarding). Порт в этом случае не проходит стандартные стадии на базе таймеров. Это важно, если к такому порту подключено обычное оборудование. Например, сервер или ПК.
- Не происходит уменьшение времени старения MAC адресов за таким портом при получении сообщения TCN. А значит меньше вероятность фладинга пакетов для хостов, которые больше слушают, чем что-то сами отправляют.

Для RSTP помимо указанных выше моментов существует ещё один, характерный только для этого протокола:

- При срабатывании механизма Proposal/Agreement блокируются все не Edge-порты. А это значит, что если мы не настроим обычный порт, куда подключено оконечное оборудование, в режим Edge, коммутатор будет выключать его на 30 секунд (работа RSTP по таймерам) каждый раз, когда меняется root port. В простых сетях это происходит не часто. Но в относительно большой легко. Это как раз тот случай, когда кажется, что перестроения произойдут в сети достаточно

быстро, и никто этого даже не заметит. А по факту устройства «отваливаются» от сети на полминуты.

Таким образом, можно сделать вывод: для RSTP отсутствие настройки Edge-порта более критично, чем для классического STP.

Примечание

С настройкой порта в режиме Edge нужно быть аккуратными.

Давайте посмотрим на поведение коммутатора Cisco с портом в режиме portfast (Edge). Порт сразу переходит в режим передачи. Но он продолжает участвовать в передаче BPDU и главное продолжает слушать сеть на наличие BPDU от других устройств, на случай если по ошибке к нему подключили другой коммутатор. Если вдруг приходит BPDU, порт теряет свое состояние portfast и проходит стандартные фазы RSTP. Так в чём же может быть проблема?

BPDU отправляются в диапазоне от 0 до 2 секунд после включения порта. Плюс можно добавить к этому время распространения BPDU по сети (актуально для STP). Поэтому в течение нескольких секунд в сети может быть петля. Если трафика будет очень много, этих секунд может оказаться достаточно, чтобы широковещательный шторм, порождённый петлёй, «убил» control-plane нашего коммутатора. Чтобы этого не допустить рекомендуется portfast настраивать в связке с дополнительными технологиями, например: BPDU Guard и storm-control.

Если сеть многовендорная, причём часть оборудования вообще не поддерживает STP ни в каком виде, всё будет плохо?

Это вопрос не совсем связан с работой RSTP, но всё же я решил его включить. Как это ни странно, подобные вопросы периодически возникают у наших заказчиков. Поэтому есть смысл на нём остановиться.

Если коммутатор не поддерживает STP ни в каком виде, что же он будет делать с BPDU пакетами? Ответ прост – передавать такие пакеты через все порты. В качестве MAC адреса назначения BPDU пакета STP и RSTP устанавливают адрес 0180.C200.0000, который является multicast адресом. Такой BPDU пакет передаётся в рамках VLAN 1.

Примечание

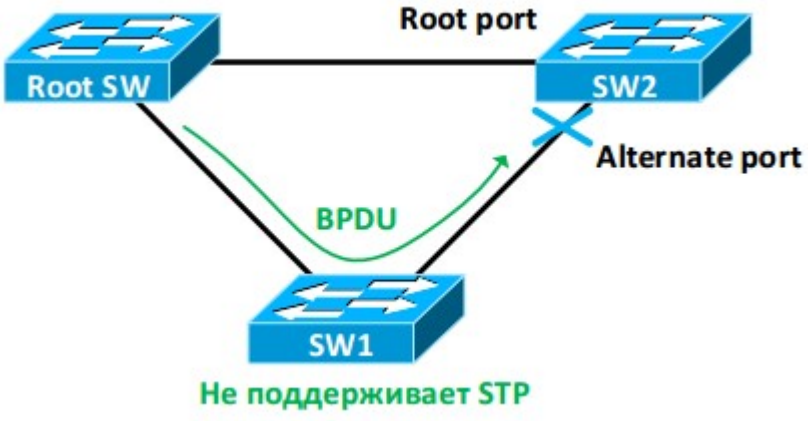
Протокол MST данные обо всех топологиях упаковывает в один BPDU (кстати, именно поэтому максимальное количество инстансов для MST — 64). В качестве адреса назначения используется стандартный MAC-адрес 0180.C200.0000.

Протоколы PVST+ и PVRST+ в своей работе используют два типа BPDU:

- IEEE-formatted BPDU для совместимости с другими версиями STP, содержит данные топологии STP для VLAN 1. В качестве адреса назначения используется стандартный MAC-адрес 0180.C200.0000.
- PVST+ BPDU, которые содержат данные топологии STP для разных VLANов. В качестве адреса назначения используется MAC-адрес 0100.0CCC.CCCD.

Ещё один занятный момент связан с тем, что даже если мы исключим VLAN 1 из транка между коммутаторами, BPDU для первого VLAN всё равно будут передаваться.

В итоге, если в нашей топологии будет коммутатор, не поддерживающий STP, он будет выглядеть для топологии STP, как обычный канал связи.



А что произойдёт, если соединить два порта между собой на коммутаторе SW1 (т.е. сделать кольцо). Наша сеть погибнет? Есть большой шанс, что нет. В этом случае Root SW получит собственный BPDU на тот же порт, с которого его отправил. После этого он сразу же его заблокирует. И петля останется «жить» только в пределах коммутатора SW1. Но положительный исход возможен, только если Root SW раньше времени не «захлебнётся» от широковещательного шторма, появившегося вследствие петли на SW1. Поэтому лучше не использовать в сети коммутаторы, не поддерживающие STP.

Нужен ли STP/RSTP/MST/… в сети, если там нет петель?

Безусловно. Если петли нет сейчас, не факт, что она не появится в будущем. Например, из-за простой человеческой ошибки, когда один access-порт коммутатора подключается к другому access-порту того же устройства.

Данный FAQ не претендует на полноту. Он носит скорее ознакомительный характер и задаёт некий вектор дальнейших изысканий по тому или иному вопросу, связанному с работой

современных протоколов семейства STP.