Динамическое агрегирование линий связи в стандарте IEEE Link Aggregation

Стандарт IEEE 802.1AX описывает процедуры агрегирования линий связи между *двумя узлами сети* (называемыми в стандарте системами), связанными между собой линиями связи топологии «точка — точка». Связи другой топологии, например «точка — многоточка», не разрешаются. Возможен также распределенный вариант агрегирования, когда агрегированные линии связи соединяют *два портала* — порталом в стандарте называется группа узлов (состоящая максимум из трех узлов), действующая как единое целое в отношении использования агрегированной линии связи. Далее мы рассмотрим только нераспределенный вариант стандарта.

Стандарт вводит новый **подуровень агрегирования линий связи (Link Aggregation Sublayer, LAS)**, располагающийся между уровнем MAC и уровнем протокола верхнего уровня, который пользуется услугами уровня MAC (рис. 11.6). Таким протоколом может быть протокол IP, протокол STP или любой другой протокол, которому необходимо передать свои данные с помощью протокола Ethernet. Для протокола верхнего уровня услуги, предоставляемые подуровнем агрегирования линий связи, идентичны услугам уровня MAC, при этом вместо **агрегатной группы портов (Link Aggregation Group, LAG)** они имеют дело с одним портом, называемым **логическим портом агрегатора**.

Логический порт агрегатора имеет свой собственный MAC-адрес (MAC-A на рис. 11.6), который заменяет адрес для протоколов верхнего уровня вместо MAC-адреса физических портов, входящих в группу LAG. То есть в том случае, когда протокол верхнего уровня обращается к протоколу Ethernet с запросом о формировании и отправке нового пакета Ethernet, протокол Ethernet указывает в качестве адреса отправителя адрес MAC-A.

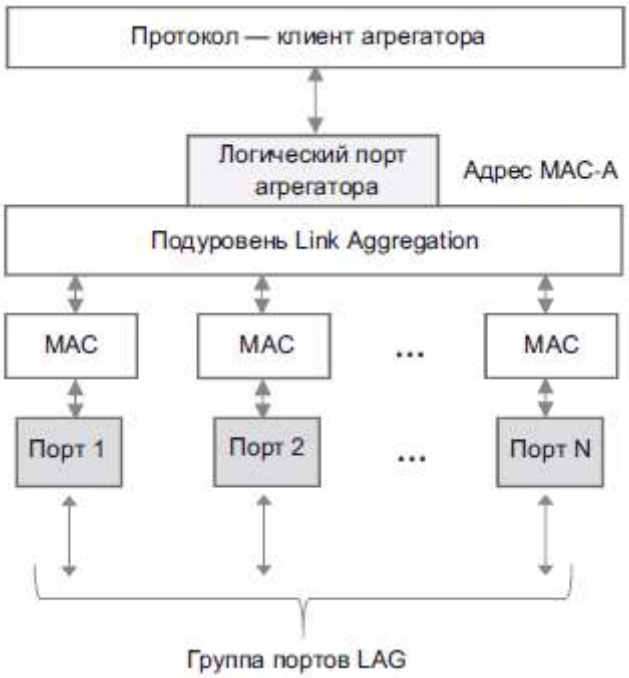

Рис. 11.6. Позиция подуровня агрегирования линий связи
в стеке протоколов Ethernet

Естественно, этот MAC-адрес используется только тогда, когда кадры отправляются от имени узла, на котором работает протокол агрегирования линий связи, например, когда агрегируются порты IPмаршрутизатора, который каждый раз упаковывает IPпакет в новый кадр Ethernet и отправляет своему соседу от своего MAC-адреса (подробно работа IPмаршрутизатора рассматривается в главе 17). Протокол STP также генерирует кадры Ethernet от имени своего узла, поэтому в его кадрах также используется адрес MAC-A. В тех же случаях, когда узел передает кадры Ethernet транзитом, как это делает коммутатор при обработке трафика своих клиентов, MAC-адреса в этих кадрах не изменяются и адрес MAC-A не используется, как не используются MAC-адреса физических портов коммутатора в том случае, когда агрегирование портов не используется.

Алгоритм прозрачного моста работает над подуровнем LAS, поэтому в таблицах продвижения коммутатора фигурирует адрес MAC-A, а не адреса физических портов, входящих в группу портов LAG. Подуровень LAS имеет довольно сложную структуру (рис. 11.7). Основным функциональным блоком подуровня LAS является **агрегатор** — блок, в обязанности которого входит распределение кадров, получаемых от протокола верхнего уровня между физическими портами агрегатной группы LAG, и выполнение обратной операции — агрегирования кадров, поступающих от физических портов в общий поток кадров, передаваемых верхнему уровню.

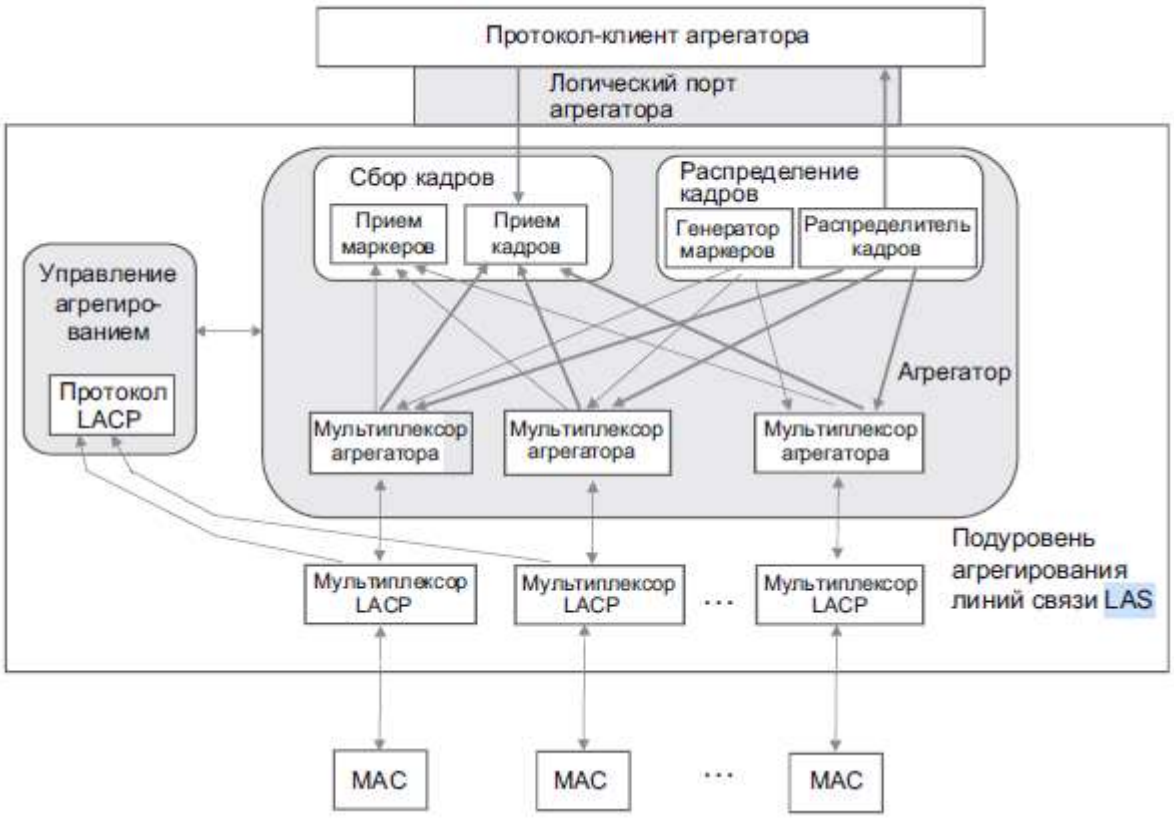

Рис. 11.7. Функциональная структура подуровня агрегирования линий связи

Агрегатор работает под управлением блока управления агрегированием, отвечающим за то, какие физические порты образуют группу LAG. Узел может иметь несколько групп LAG — в этом случае в нем организуется несколько агрегаторов по числу групп. Группа может быть образована как вручную, так и автоматически. В первом случае ее конфигурирует администратор и группа является статической, то есть членство портов в ней не изменяется до тех пор, пока администратор не изменит состав группы. Во втором случае группа образуется автоматически в результате работы **протокола управления агрегированием линий связи LACP** (Link Aggregation Control Protocol).

Протокол LACP выполняет две основные функции:

- образование группы портов LAG в результате переговорного процесса между двумя узлами сети;
- постоянный мониторинг состояния портов и линий связи, входящих в агрегатную группу LAG.

Эти функции взаимосвязаны: если в результате мониторинга выясняется, что порт или линии стали неработоспособными, то состав группы LAG может *динамически* измениться, отказавший порт из нее удаляется, а новый — добавляется. Для выполнения этих функций узлы постоянно обмениваются сообщениями протокола LACP. Протокол LACP узла может находиться в двух состояниях — активном, когда он периодически генерирует сообщения LACP, и пассивном, когда он только принимает сообщения LACP от активного соседа. Оба узла могут одновременно находиться в активном состоянии, но если они оба пассивны, то сессии LACP просто не будет существовать. Состояние LACP узла конфигурируется администратором. Для объединения нескольких физических портов протоколом LACP в одну группу LAG должно быть выполнено несколько условий:

- порты должны поддерживать одну и ту же скорость, например, порт 1 Гбит/с и порт 10 Гбит/с объединить нельзя;
- порты должны работать в полнодуплексном режиме;
- количество объединяемых портов не должно превышать максимальное количество портов для групп LAG данного узла (это значение может быть постоянным для определенной модели коммутатора либо маршрутизатора или же может быть задано администратором сети);
- всем портам должно быть присвоено одно и то же значение административного ключа.

Административный ключ — это целое число, которое присваивается порту администратором, позволяющее администратору влиять на автоматическую процедуру объединения портов в группу LAG. Кроме того, администратор может влиять на процедуру за счет назначения узлам и портам приоритетов. Приоритеты узла и порта протокола LACP аналогичны по назначению и формату приоритетам узла и порта в протоколе STP — они позволяют разрешить противоречия в переговорном процессе между двумя узлами, когда узлы предлагают различные варианты объединения портов в группу, а также отдать предпочтение определенным портам из числа не имеющих равных шансов быть включенными в группу.

Меньшее значение величины приоритета означает более высокий приоритет. Рассмотрим, каким образом происходит автоматическое образование группы LAG (рис. 11.8).

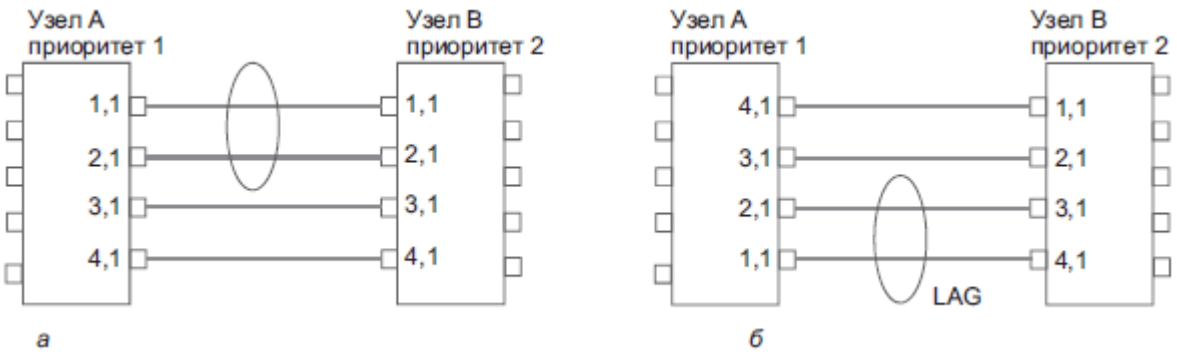

Рис. 11.8. Выбор портов при образовании группы LAG

Пусть у обоих узлов имеется ограничение на максимальное число линий связи в группе, равное 2. На рисунке также показаны значения приоритета и административного ключа каждого порта, назначенного администратором, — это пары (1,1), (2,1) и т. д. Первое значение в паре соответствует приоритету порта, второе — является значением административного ключа.

Алгоритм формирования группы LAG, реализованный в протоколе LACP, всегда старается образовать группу из максимально возможного числа портов.

В варианте *а* оба узла независимо друг от друга выбирают верхние порты для образования группы LAG, так как у обоих узлов эти два порта имеют более высокие приоритеты, чем два нижних порта. Так как значение административного ключа у всех портов одно и то же (равное 1), то происходит объединение в группу и механизм агрегирования начинает работать (мы рассмотрим его более подробно чуть ниже). Остальные два порта с более низкими приоритетами в группу не вошли, но узлы помечают их как резервные (standby), то есть такие, которые могут быть включены в группу, если условия изменятся — например, если администратор увеличит максимальное значение размера группы до трех. Тогда в группу будет автоматически добавлена связь A(3,1) — B(3,1) как имеющая более высокий приоритет по сравнению со связью A(4,1) — B(4,1).

Другой ситуацией, которая может привести к тому, что протокол LACP добавит новую связь к группе, является *отказ связи* по какой-то причине. Отказ обнаруживается при изменении локального статуса порта программным обеспечением коммутатора либо при поступлении уведомления о таком событии от партнера по протоколу LACP. При отказе связи порт, связанный с этой связью, исключается из группы и переводится в состояние резервного порта, а новый порт, если он имеется и может быть включен в группу по своим параметрам, добавляется к группе. Таким образом, обеспечивается *отказоустойчивость* агрегированных связей. Если же заменить отказавшую связь нечем, то агрегированное соединение продолжает работать с меньшим числом параллельных связей.

На рис. 11.8, *б* показан эффект приоритета узлов на работу протоколом LACP. Изза изменения приоритетов портов узлы A и B выбрали различных кандидатов — узел A выбрал два нижних

порта, а узел B — два верхних. Так как приоритет узла A выше, то в результате переговоров в группу LAG вошли два нижних порта, а верхние — переведены в резервное состояние.

В нашем примере администратор назначил всем портам один и тот же административный ключ. Но он мог бы воспользоваться этим механизмом для явного влияния на процесс образования группы. Если предположить, что узлы из примера позволяют включать в группу 4 порта, а для обеспечения необходимой пропускной способности между узлами достаточно трех, то администратор мог бы назначить третьим верхним портам ключ 1, а нижнему — ключ 2. В результате группа была бы образована из трех портов, а нижний остался вне группы (в резервном состоянии).

Возникает вопрос — возможно ли использование связи между нижними портами в случае отказа одной из связей группы? На первый взгляд кажется, что нет, так как у нижнего порта значение административного ключа не совпадает со значением административного ключа портов группы. Однако стандарт наделяет протокол LACP правом переписать значения ключей, присвоенных портам администратором. Кроме административного ключа, с каждым портом связан также еще один ключ — **операционный ключ**. Администратор не назначает значение операционного ключа, оно вырабатывается протоколом LACP. В исходном состоянии его значение совпадает со значением административного ключа, но при определенных условиях, например, при отказе связи из группы, протокол LACP может изменить значение операционного ключа порта, с тем чтобы он мог стать членом группы. Мы немного упростили описание условий включения порта в группу, исходя из того, что у всех портов группы должно быть одно и то же значение административного ключа. На самом деле во внимание принимается значение операционного ключа, но так как в исходном состоянии эти значения совпадают, то суть описания была правильной. Чтобы кадры с сообщениями протокола LACP не передавались пользователям узла, а доставлялись блоку этого протокола, в подуровне LAS имеются мультиплексоры LACP, выделяющие из потока кадров, *поступающих от соседнего узла, кадры LACP и направляющие их не агрегатору, а блоку* LACP.

Теперь рассмотрим, каким образом агрегатор распределяет потоки пользовательских кадров, поступающих от протокола верхнего уровня, между портами группы. В коммутируемой среде параллельные связи порождают проблемы для кадров с неизученными, широковещательными или групповыми адресами — кадры с такими адресами должны передаваться коммутатором на все порты, за исключением того, на который они поступили, что порождает несколько копий одного и того же кадра, которые, кроме того, зацикливаются в петле, образованной параллельными каналами. Протокол STP является одним из средств борьбы с этим явлением, но применение данного протокола не позволяет повысить пропускную способность при существовании параллельных связей, так что необходимо другое решение.

> При агрегировании связей для предотвращения таких нежелательных последствий поступают по-другому — кадры, поступившие на логический порт агрегатора от протокола верхнего уровня, *всегда передаются только на один из портов группы LAG,* даже если их адрес является неизученным, широковещательным или групповым.

Вторым аспектом проблемы распределения кадров агрегатором является проблема выбора того единственного порта группы, на который нужно передать пришедший кадр. Здесь можно предложить несколько вариантов решения. Учитывая, что одной из целей агрегирования линий связи является повышение суммарной производительности участка сети между двумя

коммутаторами (или коммутатором и сервером), следует распределять кадры по портам транка динамически, учитывая текущую загрузку каждого порта и направляя кадры в наименее загруженные (с меньшей длиной очереди) порты. *Динамический способ распределения кадров*, учитывающий текущую загрузку портов и обеспечивающий баланс нагрузки между всеми связями транка, должен приводить, казалось бы, к максимальной пропускной способности транка.

Однако такое утверждение справедливо не всегда — здесь не учитывается поведение протоколов верхнего уровня. Существует ряд таких протоколов, производительность которых может существенно снизиться, если пакеты сеанса связи между двумя конечными узлами будут приходить не в том порядке, в котором они отправлялись узлом-источником. Такая ситуация может возникнуть, если два или более последовательных кадра одного сеанса будут передаваться через разные порты транка — по причине того, что очереди в буферах этих портов имеют разную длину. Следовательно, и задержка передачи кадра может быть разной, так что более поздний кадр обгонит более ранний.

Поэтому в большинстве реализаций механизмов агрегирования используются методы статического, а не динамического распределения кадров по портам. *Статический способ распределения кадров п*одразумевает закрепление за определенным портом транка потока кадров определенного сеанса (называемого в стандарте conversation, «разговором») между двумя узлами, так что все кадры потока будут проходить через одну и ту же очередь и их упорядоченность не изменится.

Стандарт не дает точного определения потока и алгоритма приписывания потока порту. Производители коммутаторов обычно следуют традиционному определению потока на основе MAC-адресов источника и назначения, а номер порта, на который нужно передавать кадры потока, вычисляется посредством хеш-функции (hash function) — функции, которая, будучи примененной к некоторым исходным данным, дает в результате значение, состоящее из фиксированного, сравнительно небольшого и не зависящего от длины исходных данных числа байтов. Отметим, что существуют разные типы хеш-функций (см. главу 26). В качестве примера рассмотрим коммутатор, у которого образована группа LAG с четырьмя портами, которые мы обозначим двоичными числами 00, 01, 10 и 11. При поступлении от протокола верхнего уровня кадра с MAC-адресом источника 7c:25:86:64:cb:d0 и MAC-адресом назначения cc:e1:7f:06:0b:c4 хешфункция от этих двух адресов произведет результат 10, так что все кадры этого потока будут переданы через порт 10. Для кадра с адресами 6a:00:03:19:0c:31 и 0a:65:90:d6:71:fb эта же хеш-функция даст результат 01, так что кадры потока с этими адресами будут направлены в порт 01. При большом количестве различных MAC-адресов и потоков, проходящих через коммутатор, распределение потоков по портам будет более или менее равномерным, но при их небольшом количестве баланс нагрузки может и не соблюдаться.

Из того факта, что трафик одного и того же потока всегда проходит через один и тот же порт группы LAG, следует один не очень оптимистический вывод: агрегирование связей не всегда приводит к увеличению скорости передачи данных между двумя компьютерами сети. Например, имеется клиентский компьютер, который обращается к некоторому серверу, который подключен к коммутатору с помощью нескольких агрегированных связей. Если хешфункции коммутатора и сервера учитывают только MACадреса при распределении кадра, то весь поток данных между компьютером и сервером пойдет по одной из агрегированных линий связи, так что выигрыша в скорости не получится. Выигрыш будет достигнут для сети в целом, когда

сервер будет поддерживать несколько параллельных сеансов с клиентскими компьютерами сети и эти сеансы будут распределены между различными агрегированными связями.

Для более равномерного распределения потоков по портам группы LAG используются хеш-функции, которые учитывают не только MAC-адреса, но и IPадреса, а также номера TCP/IP портов, находящиеся в соответствующих полях кадра (поскольку сегодня все взаимодействия между конечными узлами сети происходят с помощью IPпротокола, такие поля в кадре имеются). При подобном подходе даже взаимодействие между парой компьютеров может выиграть от агрегирования линий связи — например, если между ними существует несколько сессий, которые порождают потоки с различными номерами TCP или UDP портов.

Учет IPадресов при распределении потоков важен и для случая, когда агрегирование портов происходит на маршрутизаторе. Сам алгоритм работы агрегатора и протокола LACP остается прежним — эти элементы программного обеспечения маршрутизатора работают на канальном уровне, имеющемся в каждом маршрутизаторе, и прямого отношения к маршрутизации, происходящей на сетевом уровне, не имеют. Однако маршрутизатор отличается от коммутатора тем, что отправляет кадры Ethernet от своего MAC-адреса, передавая их следующему маршрутизатору по его MAC-адресу. Поэтому поток кадров Ethernet между двумя маршрутизаторами всегда содержит одни и те же MAC-адреса источника и назначения, и хеш-функция, учитывающая только MAC-адреса, не сможет их распределить по разным портам. Учет IPадресов меняет картину и приводит к желаемому результату.

Нам осталось пояснить два элемента, показанные на рис. 11.7, — «Прием маркеров» и «Генератор маркеров». Эти элементы поддерживает служебный протокол, работающий между агрегаторами двух узлов, соединенных агрегированными линиями связи. Протокол носит название протокола маркеров (Marker Protocol) и служит для управления потоком кадров при переключении потока кадров между портами группы «на лету», без прерывания сеанса связи между конечными узлами, генерирующими эти кадры. Такое переключение может понадобиться при отказе порта группы или при добавлении к ней нового порта. Если агрегатор одного узла принял решение о переключении потока на новый порт, то он посылает маркер — служебный кадр определенного формата — агрегатору-напарнику. Тот должен прекратить посылать кадры на этот порт, но может «дослать» на него несколько кадров из буфера этого порта, если они там уже имеются. Когда буфер очищается, агрегатор-напарник посылает маркер-ответ, показывающий, что поток может теперь переключиться на новый порт, а старый порт может быть переведен в неактивное состояние. Чтобы кадры-маркеры не были переданы протоколу верхнего уровня, в агрегаторе имеются мультиплексоры — блоки, которые распознают кадры-маркеры и направляют их в блок «Прием маркеров».

Агрегирование линий связи является очень популярным средством повышения пропускной способности не только в локальных сетях, но и в глобальных. Многие магистрали современных глобальных сетей построены с использованием нескольких параллельных линий скорости 100 Гбит/с, что позволило строить магистрали с пропускной способностью в сотни гигабит в секунду еще до появления портов 400G Ethernet (400 Гбит/с). Можно ожидать, что с внедрением стандарта 400G Ethernet магистрали станут строиться на нескольких параллельных линиях скорости 400 Гбит/с.