Компания **Netwell** - российский дистрибьютор высокотехнологичного оборудования. Основные направления деятельности – сетевые технологии, системы хранения данных, сетевая и информационная безопасность. **Netwell** является **официальным дистрибьютором компании NetApp.**

NETAPP TECHNICAL REPORT

Ethernet для систем хранения: наилучшие методы.

David Klem, Trey Layton, Frank Pleshe, NetApp

Январь 2010 | TR-3802

Коротко о главном:

Рассмотрены некоторые функциональные особенности и методы построения высокопроизводительной сети хранения на базе технологии Ethernet (NFS и CIFS NAS, iSCSI (IP-SAN) и FCoE), выбор решений, повышающих доступность данных, надежность и производительность работы такой сети и устройств хранения ее использующих.
Подробно проанализированы методы построения «виртуальных интерфейсов» (VIF) NetApp, а также их взаимодействие с решениями EtherChannel коммутаторов Cisco, а также настройка и работа протокола Spanning Tree (STP).

Оглавление

1 Введение

Системы хранения, использующие Ethernet, существуют уже много лет в виде Network Attached Storage (NAS). Согласно этому представлению, существует мнение, что такие системы просты, и относятся к области деятельности сетевых администраторов, занимающихся в компании сетями IP. В прошлом, для небольших и некритичных систем, такой подход был вполне приемлем. Сегодня использующие Ethernet системы хранения стали быстрее, мощнее, обслуживают большие, критически важные для бизнеса приложения и базы данных, и работают с сетевой инфраструктурой 10 Gigabit Ethernet (10GbE). Для использующих Ethernet систем хранения, работающих в критичных инфраструктурах и задачах, важно принять во внимание то, что разработка сетевого решения должна быть основана на требованиях приложений и характере самого используемого протокола хранения. Эта базовая концепция, которой должно соответствовать решение, чтобы обеспечивать производительность, масштабируемость, отказоустойчивость и быстродействие.

Работающие по Ethernet системы хранения стали постоянно растущим источником сетевого трафика, который требует определенных проектных решений для достижения максимальной производительности больших серверов, работающих с большими системами хранения. Эти новые концепции немного отличаются от используемых в доставке сети на тысячи клиентских мест, и подключения серверов к LAN и WAN. Правильное проектирование и разработка сети хранения, использующей Ethernet, может позволить ей достичь производительности сетей Fibre Channel, предоставляя такие технологии, как *jumbo frames*, виртуальные интерфейсы (VIF), виртуальные сети (VLAN), IP MultiPathing (IPMP), Spanning Tree Protocol (STP), Port Channeling и многослойные топологии (*multilayer topologies*) которые могут использоваться при построении системы.

Не учтя специфические требования сети хранения Ethernet (*Ethernet Storage Networking, ESN*), и отправив это вопрос «на самотек» мы рискуем получить в результате серьезную неудачу. Один из пользователей NetApp недавно сказал: «Наш CEO считает, что базирующиеся на Ethernet системы хранения ненадежны, но почти всегда причина ненадежности оказывается не в системе хранения, а в самой сети». Сеть и использующие Ethernet системы хранения, как независимые технологии, весьма надежны сами по себе. Но проектные решения инфраструктуры, которая объединяет их, являются критически важными для достижения высокого уровня надежности решения.

Этот документ описывает наилучшие методы, рекомендации для построения сети хранения промышленного уровня, базирующейся на технологиях Ethernet. В нем перечисляются требования, которым должна соответствовать такая сетевая инфраструктура, чтобы наилучшим образом соответствовать поставленным задачам.

2 Использование VLAN для разделения трафика

В далеком прошлом, для построения сети Ethernet, для объединения между собой серверов, устройств хранения и других устройств сети, использовались устройства, называвшиеся «хабы» или «концентраторы» (*hubs*). Хабы объединяли все включенные в них устройства в единый, так называемый *collision domain* и *broadcast domain* (далее «коллизионный домен» и «широковещательный домен»). Коллизионный домен (*collision domain*) определен как физический сегмент Ethernet, включающий в себя все кабели, интерфейсы и сетевое

оборудование, входящие в один так называемый *Ethernet signal timing region*. Если два или более устройства, включенные в хаб, и находящиеся в одном коллизионном домене, начинают передавать информацию по сети в одно и то же время, то происходит «столкновение», «коллизия». В случае обнаружения коллизии, передающая система уведомляется об этом, «столкнувшиеся» фреймы отбрасываются, вводится случайная задержка перед следующим началом передачи, и передача повторяется. Обширная сеть, созданная на хабах, в особенности в случаях, когда хабы включены друг в друга, создает значительный по размерам timing region сегмента Ethernet. Большие коллизионные домены часто заставляют сеть работать неэффективно, и не позволяют реализовать весь возможный потенциал полосы пропускания данного сегмента. Поэтому, вскоре созданные *коммутаторы* Ethernet (*switches*) обеспечили изоляцию сегментов и коррекцию ошибок, чтобы добиться более высокой эффективности подключенных к сети устройств. Используя коммутаторы, администратор эффективно изолирует устройства в их собственном коллизионном домене, что снижает количество ошибок передачи и увеличивает эффективность работы сети. Но хотя коммутаторы и изолируют коллизии, они не изолируют широковещательные (*broadcast*) сообщения.

Если «коллизионный домен» это физическое деление Ethernet-сегмента, то широковещательный домен это логическое деление. Широковещательные пакеты это важная функция в Ethernet, позволяющая обнаружить адрес Ethernet-устройства, который в тот момент неизвестен передающему устройству того же широковещательного домена. В случае использования широковещания устройствам не нужно хранить Ethernet-адреса их соседей. Когда им нужно что-то передать соседнему устройству они передают в сеть широковещательный пакет, который просит объявить свой адрес у принявшего его. Устройство, находящееся в широковещательном домене отвечает на него своим адресом, и начинается передача. Большие широковещательные домены создают неэффективность использования сети образом, сходным с тем, как это происходило с большими доменами коллизий. Однако разделение широковещательных доменов обычно требует физического разделения сетевого оборудования и выделенных портов на больших маршрутизаторах. Изоляция широковещательных доменов требует вложений как в коммутаторы, так и в интерфейсы маршрутизатора, так как единственный способ соединить два независимых широковещательных домена между собой – делать это через сетевой маршрутизатор. Многие организации не изолируют широковещательные домены по причине высокой цены решения, цены маршрутизаторов и их интерфейсов.

В середине 90-х скорости сети начали расти, и технология маршрутизации между широковещательными доменами появилась и в самих коммутаторах. Такие коммутаторы стали называть «layer 3 switches» и они имеют возможность изолировать порт в его собственном домене коллизий, одновременно позволяя администратору назначить этот порт в тот или иной широковещательный домен, то есть обеспечивая некоторые возможности маршрутизации между широковещательными доменами в единой сети.

С появлением этой технологии стало возможным логически определить группу портов согласно ее функции вместо того, чтобы определять их согласно физическому расположению.

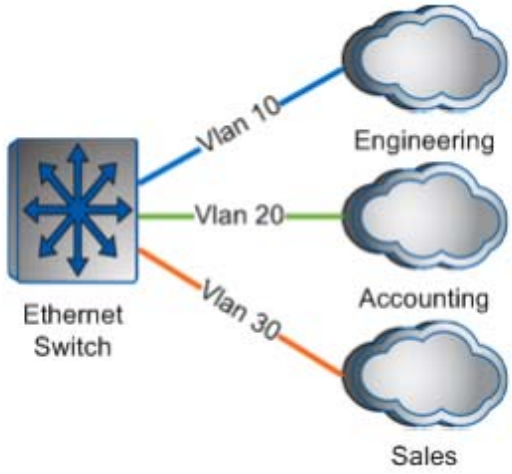

Рис. 2-1) Логическое деление на VLAN по функции

Так как VLAN-ы работают на уровне Data Link Layer (L2) сетевой модели OSI, трафик оказывается полностью изолирован между VLAN-ами, пока они не попадут в маршрутизатор (L3), который используется для соединения этих сетей вместе. Обычно делается соответствие *один-к-одному* для подсети IP и соответствующей VLAN. Если нет причин поступать иначе, следуйте этому правилу, и вы сохраните простоту сети и удобство ее управления.

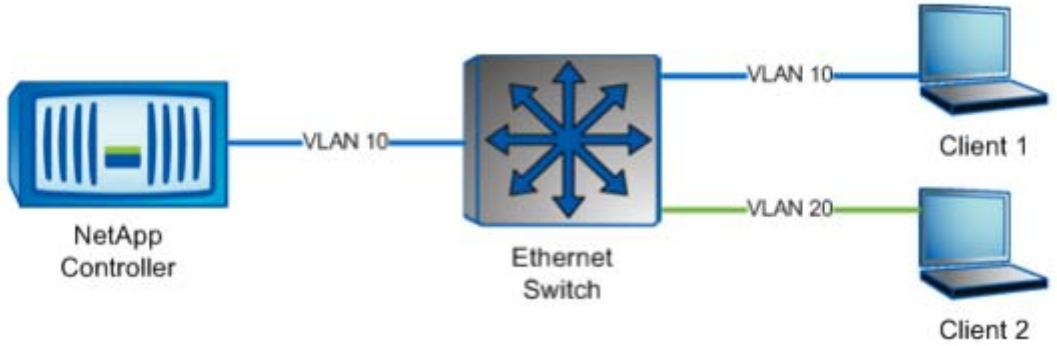

Рис. 2-2) Как работает VLAN

Использование VLAN имеет для сети следующие преимущества.

- **Более высокая степень использования коммутационной инфраструктуры.** Позволив различным отдельным логическим устройствам жить в одном и том же физическом коммутаторе мы устраняем необходимость выделять отдельным сетям их собственные коммутаторы.
  Более емкие, более эффективные и мощные коммутаторы могут быть использованы в сетевой инфраструктуре для агрегирования устройств.

- **Снижение затрат на управление.** Физическое перемещение устройств и переподключение их кабелей больше не требуется, сетевой администратор может логически назначать устройства коммутатору и сетям, непосредственно с консоли управления коммутатора.

- **Более высокая безопасность и стабильность сети.** Так как для коммуникации между VLAN-ами нужно устройство L3, то список ограничения доступа (ACL) уровня L3 можно

применить для ограничения коммуникации между отдельными сетями. Широковещательный шторм (broadcast storm) и эффекты *unicast flooding* изолируются в рамках одного VLAN. Злонамеренный пользователь с анализатором пакетов, подключенный к сети, не сможет перехватить трафик не предназначенный для хоста этого пользователя.

2.1 Транкинг VLAN

Сетевая инфраструктура предприятия часто включает в себя множество различных коммутаторов, например для целей избыточности или обеспечения достаточного количества портов. Часто логический VLAN должен существовать, проходя через несколько различных коммутаторов на своем пути. Конфигурация стандартного порта доступа (access port) позволяет находиться ему только в одном VLAN, для организации единого логического VLAN на нескольких коммутаторах требуется несколько портов, каждый из которых принадлежит только одному VLAN (такой метод не слишком хорошо масштабируется и очень неэффективен). Стандарт IEEE 802.1q предлагает решение такой проблемы, который принято называть «транкингом VLAN» (*VLAN Trunking*).

Транкинг VLAN позволяет одному линку нести в себе несколько VLAN-ов, каждый пакет которых «помечен» (*tagged*) специальным 4-байтным «тегом». Этот тег определяет, какому VLAN принадлежит данный пакет, когда он движется по поддерживающему VLAN-ы оборудованию. Поддерживают VLAN-ы сегодня большинство современного оборудования, это такие устройства, как сетевые коммутаторы, маршрутизаторы, сервера и контроллеры систем хранения NetApp. Когда фрейм достигает конечной точки, то VLAN-тег удаляется, и фрейм в своем первоначальном виде поступает конечному устройству-получателю. Тэг это просто передаваемая вместе с пакетом «инструкция»-метка, для того, чтобы быть уверенным в том, что фрейм будет доставлен в правильный широковещательный домен или VLAN.

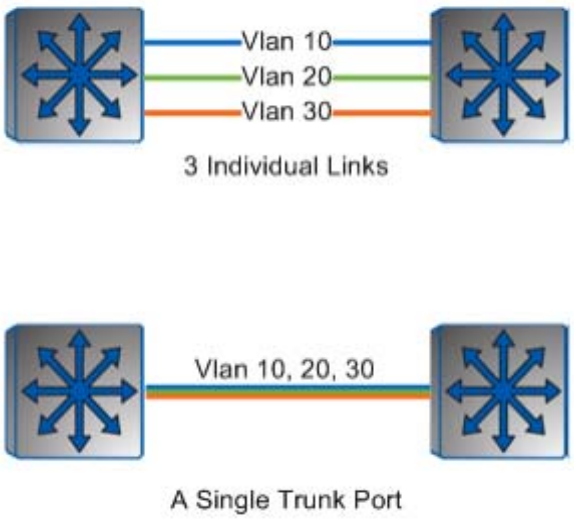

Рис. 2-3) Транкинг VLAN

Важная концепция, использованная при организации транкинга 802.1q, это так называемый Native VLAN. Фреймы, назначенные в VLAN, сконфигурированный как Native VLAN посылаются по транковому линку без пометки тэгами (untagged). Все прочие VLAN-ы, сконфигурированные в этот транк будут помечены своими соответствующими VLAN ID. По этой причине очень важно, чтобы Native VLAN был правильно сконфигурирован на обоих концах соединения. Неправильная конфигурация Native VLAN приведет к отсутствию соединения или плохой его работе.

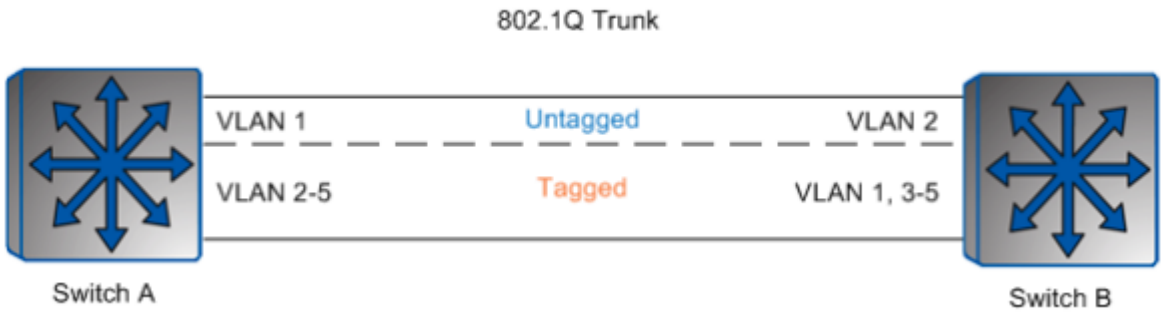

Рис. 2-4) Ошибочно сконфигурированный Native VLAN

Как говорилось ранее, контроллер системы хранения NetApp также предоставляет возможность сконфигурировать несколько VLAN-ов и использовать *VLAN Trunking*. Это дает большие гибкость и возможности конфигурирования в самом контроллере. Например, контроллер NetApp с интерфейсом 10GbE может обслуживать им несколько различных функций, таких как загрузку по iSCSI, одновременно с обычным трафиком NAS. Сеть для загрузки по iSCSI может требовать дополнительную защиту и управление, но, так как она имеет относительно невысокие требования по полосе пропускания, то администратор системы хранения, скорее всего не захочет выделять весь интерфейс 10GbE на одну эту функцию. Используя VLAN Trunking на контроллере NetApp и коммутаторе, можно организовать совместное использование одного 10Gb-линка между двумя функциями, не жертвуя возможностями изоляции между VLAN-ами.

При конфигурировании VLAN-ов на контроллере NetApp, учтите, что Data ONTAP в настоящий момент не использует возможность Native VLAN. По этой причине Native VLAN на порту комутатора, к которому подключен котроллер NetApp, должен быть назначен на неиспользуемый VLAN, чтобы быть уверенным в том, что данные форвардятся правильно.

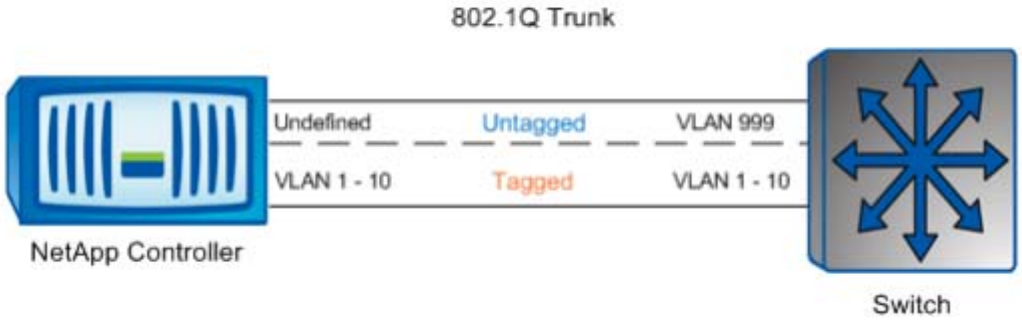

Рис. 2-5) Конфигурирование Native VLAN на стороне NetApp

Шаблон конфигурации – транкинг VLAN

NETAPP RC FILE

```
vlan create e0a 10 20 30
ifconfig e0a-10 192.168.10.10 netmask 255.255.255.0
ifconfig e0a-20 192.168.20.10 netmask 255.255.255.0
ifconfig e0a-30 192.168.30.10 netmask 255.255.255.0
```

CISCO CONFIG

```
interface GigabitEthernet1/1
       description NetApp e0a Trunk
```

```
       switchport mode trunk
switchport trunk allowed vlan 10,20,30
       switchport trunk native vlan 123
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       spanning-tree portfast
end
```

2.2 Рекомендации по VLAN

Очень важно хорошо разобраться и понять принципы работы сетевой инфраструктуры и характер трафика идущего по ней. Использование VLAN может улучшить работу Ethernet-хранилища, за счет использования эффективного механизма сегментации сетевого трафика и, одновременно, лучшего использования ресурсов сетевого коммутатора. При конфигурировании сети с использованием VLAN помните о следующих важных моментах:

- Убедитесь, что Native VLAN правильно сконфигурирован на обоих концах коммутируемого соединения, если коммутаторы поддерживают native VLAN.
- Если коммутатор поддерживает и использует Native VLAN, и к нему подключен контроллер NetApp, то на этом порту Native VLAN должен иметь номер, отличный от номеров VLAN, используемых контроллером NetApp
- Используйте транкинг VLAN (*VLAN trunking*) чтобы позволить нескольким VLAN идти через один высокоскоростной канал, такой, как, например, 10GbE.

3 Предотвращение петель с помощью Spanning Tree

3.1 Введение в протокол

Spanning Tree Protocol (STP) обеспечивает механизм идентификации и динамического построения топологии сети без «петель». При построении сети хранения, построенной на технологии Ethernet, часто использующей комбинацию интерфейсов 10Gbps и 1Gbps, следует с особым вниманием отнестись к топологии Spanning-Tree. Инженеры NetApp часто вынуждены решать проблемы пользователей с производительностью Ethernet-сети хранения, вызванных недостаточным их вниманием к сетевой топологии и правильного использования spanning-tree.

В этой глав мы часто будем упоминать понятии бридж (bridge). «Бридж» традиционно считался физическим устройством, но более позднее время с появлением коммутаторов, способных работать со многими VLAN bridge стал программным для каждого VLAN. Поскольку в тексте речь у нас пойдет идет именно о таком бридже, можно считать, что мы говорим о процессе, который выполняется для каждого VLAN на каждом устройстве, которое сконфигурировано для работы с определенными VLAN.

Топология *spanning-tree* вычисляется с помощью передачи между «бриджами» («мостами») специальных сообщений, которые помогают построить *spanning tree*. Такие специальные сообщения называются конфигурационные *Bridge Protocol Data Units* (configuration BPDUs).

Внимание: Если на коммутаторе сконфигурировано 100 VLAN, то STP - протокол для каждого из 100 VLAN построит отдельное дерево.

Бриджи обмениваются сообщениями *configuration BPDU*, чтобы выполнить следующую процедуру:

- Бриджи выбирают между собой один «корневой» бридж («root» bridge).
- Каждый бридж вычисляет кратчайший путь к корневому бриджу. Порт, через который проложен такой путь называется «корневой» порт («root» port).
- Каждая LAN просматривает дистанцию к корневому порту и выбирает среди них кратчайший, называя его бридж «назначенный» бридж («designated» bridge). Такой *designated bridge* будет передавать фреймы этой LAN к *root bridge*.
- Наконец, все порты бриджей, которые не являются корневыми портами или *designated bridge*, запрещаются.

В инфраструктуре, использующей множество VLAN, Spanning Tree Protocol работает раздельно в каждом VLAN.

3.2 Выборы Root Bridge

Первый шаг при создании топологии без петель, это «выборы» так называемого корневого, или *root bridge* на котором будет в дальнейшем построено дерево *spanning tree*. Представим себе два физических коммутатора, работающих в одной VLAN. Когда кабель Ethernet соединяет эти коммутаторы, два этих порта немедленно обмениваются сообщениями *configuration BPDU*. Эти сообщения содержат локальные приоритеты и адреса MAC. Локальный приоритет (*local priority*) это конфигурируемая пользователем величина, которая влияет на процесс выбора корневого бриджа (*root bridge*). Бридж с самым низким его значением будет выбран в качестве корневого. **Внимание:** Появление в сети меньшего MAC-адреса приводит к полному перестроению дерева STP! Коммутаторы обычно приходят от изготовителя со значением приоритета по умолчанию, и если они не были специально настроены, они объявят о равном значении *priority*. Если администратор желает сделать какой-то определенный коммутатор «корневым», то он должен принудительно указать для этого коммутатора значение *priority* ниже, чем у всех других коммутаторов в этой сети.

Существует много методов соединения коммутаторов, и топология *spanning tree* будет строиться по-разному, для каждого из этих методов. Вот несколько примеров:

Соединение	Результаты работы Spanning Tree
Один кабель, соединяющий два порта в одной VLAN	«Корневой» бридж выбирается обычным образом для данной VLAN.
Один кабель, соединяющий два порта, сконфигурированных в VLAN trunk	«Корневой» бридж выбирается обычным образом для каждой VLAN в транке. Возможно иметь одну VLAN «укорененную» в одном коммутаторе, и другую VLAN «укорененную» в другом коммутаторе.
Несколько кабелей, соединяющих порты, сконфигурированные для одной VLAN	«Корневой» бридж выбирается обычным образом для данной VLAN. Избыточные линки считаются петлями и оставляется только порт с наинизшим port ID (корневой).
Несколько кабелей, соединяющих	«Корневой» бридж выбирается обычным образом для

порты, сконфигурированные для VLAN trunk	каждой VLAN в транке. Избыточные линки считаются петлями и оставляется только порт с наинизшим port ID (корневой).
Несколько кабелей, соединяющих порты, сконфигурированные в один VLAN с Etherchannel	«Корневой» бридж выбирается обычным образом для данной VLAN,и все линки работают вместе как один логически линк. Смотри главу 4 о работе *port bonding*.
Несколько кабелей, соединяющих порты, сконфигурированные в VLAN trunk с Etherchannel	«Корневой» бридж выбирается обычным образом для каждой VLAN в транке,и все линки работают вместе как один логически линк. Смотри главу 4 о работе *port bonding*.

3.3 Выбор пути к Root

Когда выбран корневой бридж и построено дерево spanning tree, следующим шагом будет использовать полученную информацию для определения наилучшего пути и блокирования обнаруженных петель. Наилучшим путем будет считаться кратчайший и быстрейший путь к корневому бриджу.

На рисунке 1.1* показано три коммутатора. Коммутатор B это корневой бридж, и он соединен с Коммутатором A и Коммутатором C соединениями 1Gbps. Коммутаторы A и C также имеют соединение между собой полосой 10Gbps. Это соединение 10Gbps создает петлю в сети, и эту петлю следует устранить. Топология spanning tree имеет форму «дерева», имеющего «корень», поэтому «корень» (*root*) этого «дерева» расположен в «корневом» бридже. Коммутатор B это «корень» для соединения 1Gbps. Соединение же 10Gbps между A и C будет блокировано, так как оно не является кратчайшим путем к «корню».

На рисунке 1.2* изображены те же 3 коммутатора, но на этот раз у нас есть несколько линков между Коммутатором A и C к корневому бриджу B. Эти линки к корню являются в данной топологии «петлями». Когда между ними будет выбран быстрейший, то остальные, более медленные линки будут заблокированы.

3.4 Механизмы Fast Start

Так как вычисление и построение *Spanning Tree* может занимать довольно продолжительное время (до нескольких секунд), прежде чем данный порт перейдет в *forwarding state*, производители сетевого оборудования, например Cisco, создают специальные механизмы «Fast Start», чтобы позволить трафику начать передаваться по сети максимально быстро. Эти опции, которые называются *PortFast*, *UplinkFast* и *BackboneFast* обеспечивают механизмы «Fast Start» для портов, распложенных в различных точках сети.

PortFast

Опция *spanning tree* под названием *PortFast* заставляет порт *spanning tree* перейти в *forwarding state* немедленно, пропуская состояния прослушивания и «обучения». Вы можете использовать PortFast на портах коммутатора, непосредственно подключенных, например, к одной единственной рабочей станции или серверу, что позволит им подключаться к сети немедленно, не дожидаясь окончания построения *spanning tree*.

* Рисунки отсутствуют в оригинале документа

UplinkFast

UplinkFast обеспечивает быстрое соединение после изменения топологии *spanning tree* и достижения балансировки нагрузки между избыточными линками при использовании групп аплинков (*uplink groups*). *Uplink group* это группа портов (в единой VLAN), из которых только один работает как *forwarding port* в данный момент времени. Например, *uplink group* состоит из корневого порта (*root port*), который работает как *forwarding*, и нескольких портов в блокированном (*blocked*) состоянии. Группа аплинков (*uplink group*) обеспечивает альтернативный путь в случае, если текущий *forwarding link* перестает работать.

BackboneFast

Backbone Fast это собственная разработка Cisco, которая, будучи включена на всех коммутаторах сети бриджей, может сэкономить около 20 секунд (max_age) при восстановлении из состояния ошибки *indirect link failure*.

PVST+ (Per VLAN Spanning-Tree)

Базовый протокол *Spanning Tree protocol* может быть очень медленным в работе. В процессе развития сетевых средств появилась необходимость существования средств немедленной настройки или переключения в результате сбоя на резервный линк, что заставило эволюционировать базовый протокол Spanning Tree. Так, например *Per VLAN Spanning Tree* (PVST+) основан на стандарте IEEE802.1D и включает проперитартные расширения, разработанные Cisco, такие как *BackboneFast*, *UplinkFast*, и *PortFast*. *Rapid-PVST+* основан на стандарте IEEE 802.1w и работает быстрее, чем 802.1D. RSTP (IEEE 802.1w), и включает большинство расширений Cisco для 802.1D Spanning Tree, такие как *BackboneFast* и *UplinkFast*.

CISCO CONFIG

```
interface GigabitEthernet1/1
       description NetApp e0a Trunk
       switchport mode trunk
       switchport trunk allowed vlan 10,20,30
       switchport trunk native vlan 123
       flowcontrol receive on
       no cdp enable
spanning-tree guard root
spanning-tree guard loop
spanning-tree portfast
end
```

3.5 Рекомендации по использованию Spanning Tree

Понимание основ того, как работает Spanning Tree необходимо для того, чтобы понять, как трафик движется по вашей сети. Неправильно сконфигурированная сеть layer 2 Spanning Tree может начать посылать трафик по неосновным, вспомогательным линкам, увеличив время доступа к данным и задержки в сети. В наихудшем варианте *spanning tree* может замедлить скорость работы путем использования линков с пониженной пропускной способностью, что вызовет общее падение производительности системы. Ключевые элементы оптимизации сетевой производительности при использовании правильно спроектированного Spanning Tree это:

- Полное понимание сетевой топологии и того, как трафик движется по каждой из используемых для сетевого хранилища VLAN.

- Понимание того, как именно отказ соединения инициирует процедуру перестроения *spanning tree* и гарантия того, что такой процесс не затронет пользователей при возможном файловере контроллеров системы хранения.

- Использование механизмов *fast start*, чтобы быть уверенным в минимально возможном времени переключения портов в рабочий режим.

4 Улучшение производительности и надежности с помощью объединения портов

Методы объединения портов, доступные как на стороне сетевого оборудования, так и на стороне системы хранения, позволяют объединить несколько физических линков в единый виртуальный сетевой интерфейс. Работающий на входящих в такой интерфейс сетевых линках алгоритм балансировки нагрузки обеспечивает повышение производительности сетевого интерфейса и отказоустойчивость.

Применяемая NetApp терминология отчасти различна с общепринятой в сетевой индустрии, что может вызывать сложности понимания. Сетевая индустрия обычно использует для обозначения этой технологии понятия *EtherChannel* или *port-channel*, NetApp называет это *Multi-Mode VIF*, или *virtual interfaces*. Часто применяется некорректный термин *trunked interfaces* или *trunk*. Следует остерегаться использования для них терминов *trunk* или *trunking* (в русской литературе принято название «транк» или «транкинг»), так как в сетевой индустрии это понятие уже используется для обозначения принципиально иного понятия - *VLAN trunking*.

NetApp предлагает три типа организации VIF, каждый из них имеет своим преимущества и недостатки, и свои требования.

- *Single-mode VIF*

- *Static Multi-mode VIF*

- *Dynamic Multi-mode VIF (LACP)*

Во многих случаях рекомендуется, чтобы разные типы VIF использовались вместе, обеспечивая избыточность и отказоустойчивость.

4.1 Single mode VIF

В режиме *single-mode VIF* виртуальный интерфейс содержит один активный линк и произвольное количество линков в пассивном состоянии. Все линки *single-mode VIF* совместно используют один и тот же MAC-адрес. Только один линк активен и пересылает данные через себя, по этой причине не происходит дублирования MAC-адреса в сети.

В примере показанном ниже, *single-mode VIF* сконфигурирован на двух линках, соединенных с избыточным коммутатором. Активен и передает данные только линк **e0**, а линк **e1** находится в *standby mode*. Если **e0** перестает работать (сигналом для этого служит переход интерфейса в состояние *down*), линк **e1** включается и перехватывает данные, шедшие по **e0**. В случае наличия нескольких интерфейсов в состоянии *standby*, активируемый интерфейс выбирается между ними случайным образом. Такой режим обычно называется *active-passive mode*.

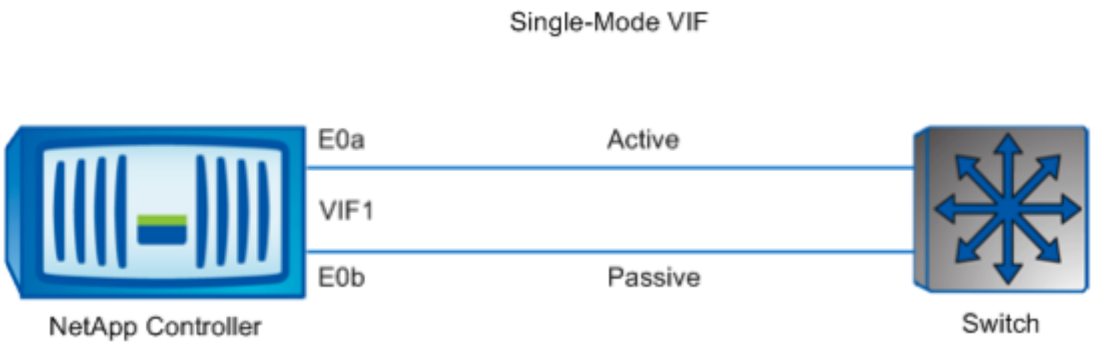

Рис. 4-1) *Single Mode VIF*

Вместо выбора интерфейса для активации случайным образом, некоторым администраторам может понадобиться назначит определенный интерфейс как предпочтительный для этой операции. Это можно сделать командой `vif favor`. Это отключает автоматический выбор, и линки будут активироваться только в определенном, заранее заданном порядке. Сходным, но противоположным образом работает команда `vif nofavor`, при которой указанные в ней интерфейсы не будут выбраны до тех пор, пока они не останутся последними доступными в *single mode VIF*.

За

- Не нужна специальная конфигурация коммутаторов для использования *single mode VIF* на стороне контроллера системы хранения NetApp.
- Такой вариант избыточности не уменьшает полосу пропускания канала в случае файловера.

Против

- Использует как минимум вдвое больше портов на коммутаторе, но использует только половину (и менее) от доступной им суммарно полосы пропускания.
- Интерфейс в пассивном состоянии показывается для средств управления со статусом *down*, что означает, что администратор не может наблюдать состояние этого линка (например, количество ошибок) чтобы предпринимать упреждающие действия.
- Часто событие файловера не тестируется до момента наступления аварии, что не позволяет убедиться в том, что оборудование нормально отработает ситуацию отказа.
- Если не используется *Spanning Tree fast start* на портах контроллера системы хранения, то вычисление *spanning tree* необходимо до того, как порт будет активизирован и начнет передавать трафик.

Шаблон конфигурации – Single Mode VIF

Ниже приведены примеры файлов конфигурации как для контроллера NetApp, так и для коммутатора, использующего Cisco IOS.

NETAPP RC FILE

```
vif create single template-vif1 e0a e0b
vif favor e0a
ifconfig template-vif1 10.1.1.100 netmask 255.255.255.0 mtusize 1500
partner 10.1.1.200
```

```
route add default 10.1.1.1
routed on
options dns.domainname template.netapp.com
options dns.enable on
savecore
```

CISCO IOS SWITCH

```
interface GigabitEthernet1/1
       description NetApp e0a
       switchport access vlan 116
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       spanning-tree portfast
!
interface GigabitEthernet1/2
       description NetApp e0b
       switchport access vlan 116
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       spanning-tree portfast
end
```

4.2 Static Multi-Mode VIF

Static multi-mode VIF состоит из двух и более интерфейсов которые оба являются активными линками. Он соответствует стандарту IEEE 802.3ad (static), и требует поддержки на стороне коммутатора. Однако в случае конфигурации *static multi-mode VIF* не используются ни LACP, ни PAgP, поэтому ни контроль соединения, ни автоопределение не используются. Коммутатор должен быть сконфигурирован в режиме «static», который принудительно определяет интерфейсы, закрепленные для использования в «канале».

Static multi-mode VIF продолжает работать даже если все кроме одного линка перестанут работать. Это позволяет добиться более высокой пропускной способности, чем у *single mode VIF*, и также обеспечивает избыточность. Доступны несколько вариантов алгоритма балансировки нагрузки со стороны NetApp FAS, например:

- По IP-хэшу Источника и Получателя.
- По хэшу MAC-адреса Источника и Получателя.
- Round Robin.

Посылающий информацию контроллер всегда определяет то, какой линк используется для посылки трафика. По этой причине несовпадающие настройки на стороне коммутатора и контроллера приводят к неправильному распределению трафика по линкам как на передачу, так и на прием. Вследствие этого алгоритм хэширования при балансировке должен совпадать как можно более точно в обоих направлениях.

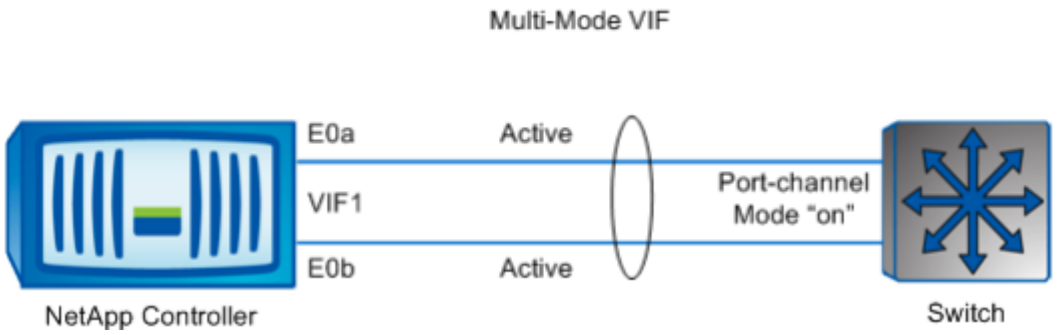

Рис. 4-2) *Static Multi-Mode VIF*

За

- Имеет более высокую пропускную способность и полосу пропускания, с использованием алгоритмов балансировки нагрузки, так как работают одновременно все входящие в канал интерфейсы.

- Может обнаруживать обрыв соединения по любому из своих линков, и перенаправлять трафик в оставшийся. Это позволяет обеспечивать избыточность и отказоустойчивость без необходимости непродуктивно тратить порт на использование режима active/passive.

Против

- Static mode работает в «принудительном» режиме, без контроля соединения, что может быть результатом проблем типа «черная дыра трафика», о которой мы упоминали ранее.

- Так как не используются никакие методы установления соединения (*negotiation*) между устройствами, то static mode создает меньше записей в логах и отчетов об ошибках, что может затруднить возможный поиск неисправности.

- Достижение равномерного распределения трафика по портам, входящим в канал, возможно только при использовании нескольких пар адресов Источника и Получателя, и выбора правильного алгоритма балансировки. Для оптимального распределения трафика может понадобиться дополнительная настройка на стороне контроллера NetApp FAS (например установки алиасов на IP, что обсуждается ниже).

Шаблон конфигурации – Static Multi-mode VIF

NETAPP RC FILE

```
vif create multi template-vif1 –b ip e0a e0b
ifconfig template-vif1 10.1.1.100 netmask 255.255.255.0 mtusize 1500
partner 10.1.1.200 flowcontrol send
route add default 10.1.1.1
```

CISCO IOS SWITCH

```
interface GigabitEthernet1/1
       description NetApp e0a
       switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       channel-group 5 mode on
```

```
!
interface GigabitEthernet1/2
       description NetApp e0b
switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       channel-group 5 mode on
!
interface Port-channel5
       description NetApp template-vif1
       switchport
       switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
end
```

4.3 Dynamic Multi-Mode VIF

Режим *Dynamic multi-mode VIF* основан на использовании *Link Aggregation Control Protocol* (LACP) описанного в стандарте IEEE 802.ad (dynamic). Он похож на *static multi-mode VIF* в том, что содержит два и более активных интерфейса, совместно использует один MAC-адрес, и требует конфигурации на обоих концах соединения. В режиме *dynamic multi-mode VIF*, однако, дополнительные параметры соединения передаются между соединениями при помощи *LACP Protocol Data Unit* (PDU). Это позволяет двум соединенным устройствам динамически удалять или добавлять линки в канал не только по причине физического отказа линка. Это важное усовершенствование, так как при этом *dynamic multi-mode VIF* может не только обнаруживать и реагировать на потери соединения, но также реагировать и на потери потока данных (data flow). Это обеспечивает более высокую доступность и может помочь предотвратить проблему *packet black hole*, обсуждавшуюся ранее в главе о *single-mode VIF*.

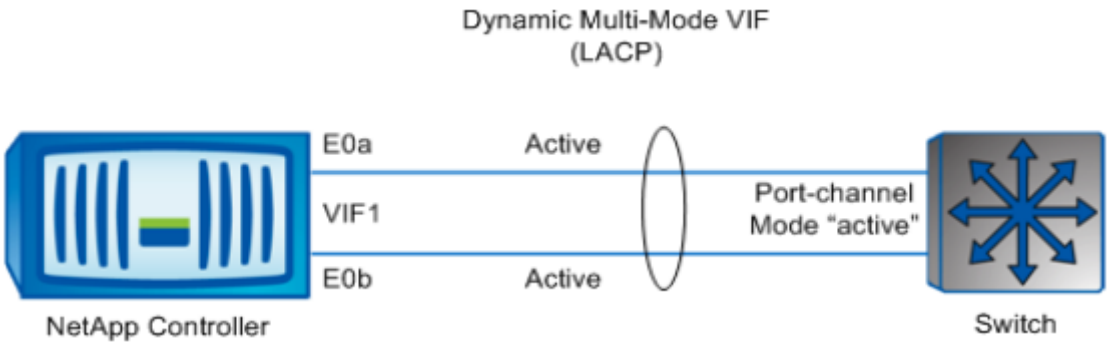

Рис. 4-3) *Dynamic Multi-Mode VIF (LACP)*

За

- Позволяет добиться более широкой агрегированной полосы пропускания, за счет выбора алгоритма балансировки нагрузки и активности всех портов, входящих в канал.

- За счет использования LACP PDU *dynamic multi-mode VIF* обнаруживает не только потерю соединения с другой стороной, но и проблемы потока данных (data flow). Это позволяет

устранить проблемы «забитой» очереди и «черной дыры трафика» (traffic black hole), описанной ранее.

- Многие современные коммутаторы поддерживают *multi-chassis LACP*, который устраняет необходимость строить *single-mode VIF* «второго уровня» для обеспечения отказоустойчивости между коммутаторами.

Против

- Старые модели коммутаторов могут не уметь работать с LACP совсем.
- Старые модели коммутаторов могут не поддерживать режим multi-chassis LACP, что требует построения single-mode VIF «второго уровня», для обеспечения отказоустойчивости. (Cisco поддерживает режим *multichassis etherchannel* как для LACP так и для Static)
- Достижение хорошей равномерности распределения трафика по линкам канала требует использования нескольких пар адресов источника и получателя и настройки правильной балансировки. Для достижения равномерной нагрузки линков на контроллере NetApp FAS необходимо принимать специальные меры (например назначать IP-алиасы, как рассматривается далее).

Шаблон конфигурации – Dynamic Multi-mode VIF

NETAPP RC FILE

```
vif create lacp template-vif1 –b ip e0a e0b
ifconfig template-vif1 10.1.1.100 netmask 255.255.255.0 mtusize 1500
partner 10.1.1.200 flowcontrol send
route add default 10.1.1.1
```

CISCO IOS SWITCH

```
interface GigabitEthernet1/1
       description NetApp e0a
       switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       channel-group 5 mode active
!
interface GigabitEthernet1/2
       description NetApp e0b
       switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
channel-group 5 mode active
!
interface Port-channel5
       description NetApp template-vif1
       switchport
       switchport access vlan 100
       switchport mode access
```

```
        flowcontrol receive on
        no cdp enable
        spanning-tree guard loop
end
```

4.4 Производительность с VIF

Повышение производительности часто называется основной причиной для включения и использования VIF на системе хранения. Хотя то, что *multi-mode VIF* увеличивает производительность, и есть в целом правда, следует принять во внимание множество аспектов, участвующих в этом процессе.

Когда вы создаете *multi-mode VIF*, алгоритм балансировки нагрузки используется для того, чтобы определить то, по какому из имеющихся в VIF линков следует отправлять потоки трафика. Data ONTAP поддерживает три различных метода балансировки нагрузки для *static* и *dynamic multi-mode VIF*: *round robin*, *source/destination IP*, и *source/destination MAC*.

Для каждого из этих режимов используется логическая операция «Исключающее ИЛИ» (XOR) над последним байтом адреса источника и адреса получателя для того, чтобы определить линк в VIF, по которому будет направлен трафик данной пары адресов, руководствуясь следующей формулой:

(*Адрес_Источника* XOR *Адрес_Получателя*) % *ЧислоЛинков*

Одна из основных ошибок в понимании принципов работы VIF заключается во мнении, что N линков по 1Gb, объединенных в multi-mode VIF дадут доступную полосу, равную N*1Gb. Хотя в теории это и так, при идеальном варианте *frame distribution*, на практике это не всегда достижимо, так как зависит от выбранного метода хэширования. При большом количестве пар «источник-получатель», метод, базирующийся на хэше IP или MAC-адресов дает довольно неплохую равномерность распределения. Однако в случае единичной передачи (пары источник-получатель) можно получить передачу на скорости равной скорости только одного единственного физического линка, входящего в канал. Приведенный далее пример показывает, почему VIF не обеспечивает увеличения производительности и полосы пропускания при использовании малого числа пар «источник-получатель».

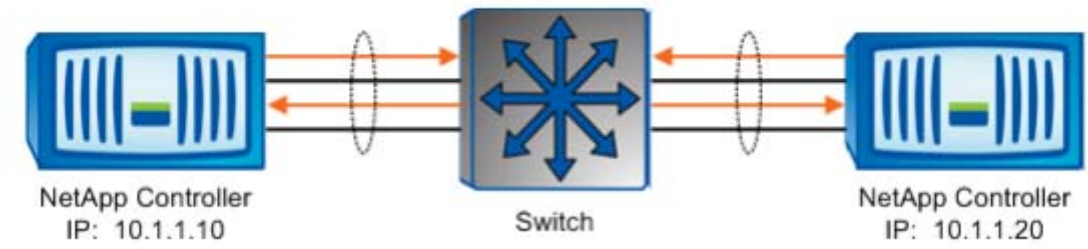

Рис. 4-4) Балансировка нагрузки в VIF

В рассматриваемом случае каждый контроллер NetApp FAS имеет *multi-mode VIF*, состоящий из двух физических интерфейсов, на который назначен один IP-адрес. Когда инициируется любая передача данных между двумя устройствами (например репликация SnapMirror), рассчитываемый хэш всегда будет иметь одно и то же значение, в результате чего трафик всегда пойдет по одному и тому е единственному линку. Это означает, что соединение «один-к-одному»

не будет иметь никаких преимуществ в производительности от существования дополнительных линков в *multi-mode VIF*. В зависимости от участвующих адресов IP или MAC, мы можем получить неравномерное распределение трафика по линкам VIF. Для определения наилучшего в конкретном случае алгоритма хэширования могут потребоваться некоторые эксперименты.

Другим важным отличием является то, что устройство, отправляющее поток трафика, определяет правило распределения фреймов по линкам для каждого «хопа». В примере выше, отправляющий данные контроллер системы хранения определяет хэш для первого VIF. Далее, когда данные проходят по линкам к принимающему данные контроллеру, то хэш задает и определяет уже коммутатор. Это может вызвать затруднения, если в большой сети одновременно используются несколько разных алгоритмов хэширования (и *frame distribution*).

4.5 Балансировка нагрузки с VIF

Round Robin

Метод, называемый в технической литературе Round Robin («карусель»), был одним из первых практически использованных производителями коммутаторов алгоритмов балансировки нагрузки. Этот алгоритм состоит в поочередном направлении фреймов Ethernet по всем входящим в канал линкам, вне зависимости от MAC или IP адресов отправителей и получателей. Это обеспечивает очень равномерное распределение трафика по линкам канала, но также имеет и серьезный недостаток в виде проблемы *out of order packet delivery* или приходе пакетов не в нужном порядке. Фрейм 1 посланный через линк 1 должен прийти на сторону получателя раньше, чем туда придет фрейм 2, посланный через линк 2. Если это условие не соблюдается, то требуется вмешательство протокола более высокого уровня и исправление ситуации, что, зачастую, ведет к необходимости переотправки фреймов, пришедших не в нужном порядке. Из-за наличия такой проблемы, ведущей к неэффективности, использование *round robin* для балансировки нагрузки обычно не рекомендуется.

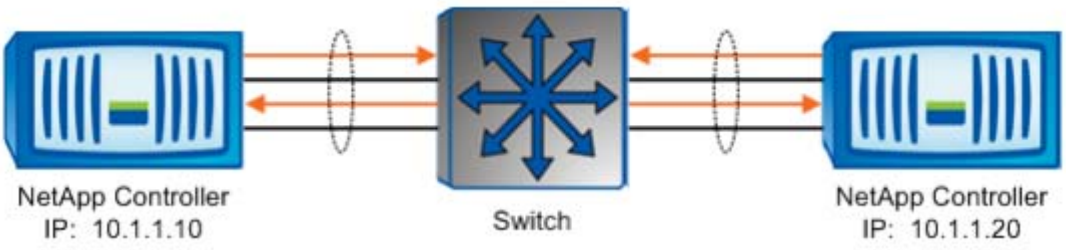

Рис. 4-5) *Round Robin*

По MAC адресам источника и получателя

Этот алгоритм используется не так часто, так как ему присущ ряд недостатков, например резкий дисбаланс трафика между отдельными линками. Основанный на использовании MAC алгоритм балансировки вычисляет значение XOR между MAC-адресами пар источника и получателя. Источником будет MAC-адрес сетевой карты хоста, подключенного к контроллеру NetApp. Получателем будет MAC-адрес VIF-интерфейса контроллера NetApp. Алгоритм работает хорошо, когда хосты и контроллер NetApp размещаются в одной подсети или VLAN. Однако когда хост расположен в другой подсети относительно контроллера NetApp, и трафик проходит через роутер, мы сталкиваемся с недостатками алгоритма. Когда трафик проходит через роутер, то только один путь от роутера к хосту будет использован для всего трафика, вне зависимости от наличия дополнительных путей. На рисунке мы попробуем пояснить, почему так происходит.

- IP адрес Host1 - 10.10.1.20/24 (шлюз по умолчанию для Host1 - 10.10.1.1)
- IP адрес Controller1 - 10.10.3.100/24 (шлюз по умолчанию для Controller1 - 10.10.3.1)

Хост и контроллер системы хранения расположены в двух различных подсетях. Единственный путь связи между ними это через роутер, который позволяет маршрутизировать трафик меду сетями L3 и отрезает широковещательные пакеты. В рассматриваемом примере шлюзы по умолчанию 10.10.1.1 и 10.10.3.1 находятся в одном физическом маршрутизаторе, и являются адресами его двух интерфейсов.

Когда Host1 строит фрейм для Controller1, он видит, что адрес 10.10.3.100 это адрес не из его собственной сети, поэтому он направляет его в адрес шлюза по умолчанию.

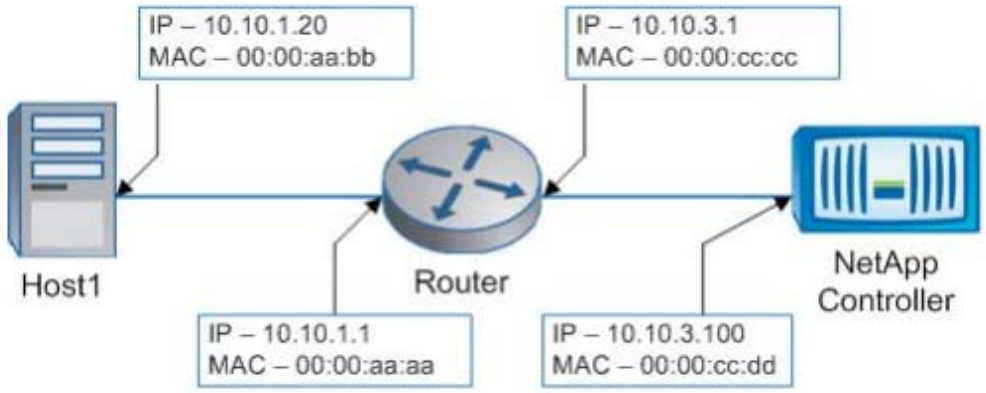

Рис. 4-6) Пример передачи потока трафика

От Host1 до Host1Router

- IP Source: Host1 (10.10.1.20)
- MAC Source: Host1
- IP Destination: VIFController1 (10.10.3.100)
- MAC Destination: Host1DefaultRouter

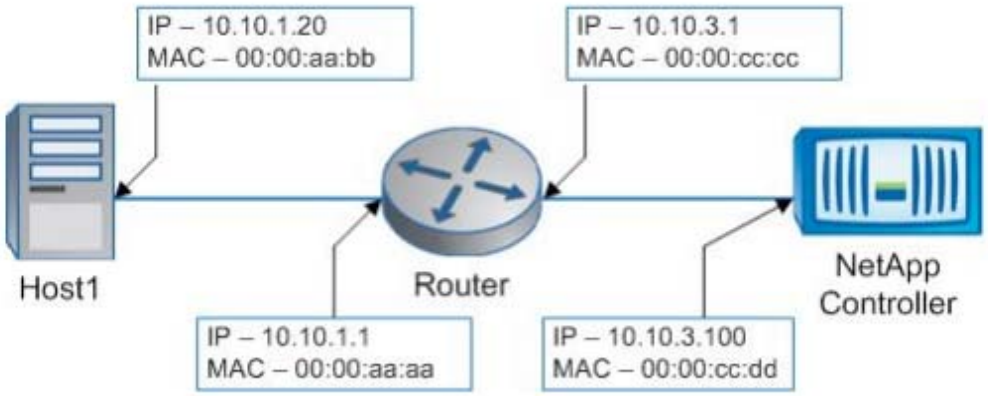

Рис 4-7) Пример потока трафика

Host1Router Routing Packet to Controller1

- IP Source: Host1 (10.10.1.10)
- MAC Source: Controller1DefaultRouter
- IP Destination: VIFController1 (10.10.3.100)
- MAC Destination: VIFController1

Адреса MAC - как источника, так и получателя, меняются во фрейме, когда он проходит по сети. Но IP-адреса источника и получателя остаются неизменными. Это является проблемой для алгоритма, основанного на хэше MAC-адресов: для пакетов, которые выходят с контроллера NetApp, MAC-адрес источника всегда будет соответствовать адресу VIF, а MAC-адрес получателя – всегда адрес порта маршрутизатора. В результате, только один линк будет использован для «балансировки».

Для полного понимания того, в чем корень проблемы, представим себе VIF, состоящий из 4 линков по 1Gbps на Controller1 и 50 хостов в той же подсети, что и Host1. Пакеты, посылаемые контроллером к одному из этих хостов, всегда будут иметь одни и те же MAC-адреса источника и получателя, поэтому они пойдут по одному единственному физическому линку из четырех.

По IP адресам источника и получателя

IP Load-Balancing это параметр по умолчанию для всех MultiMode VIF в NetApp, и наиболее часто используемый на практике метод построения MultiMode VIF на сегодня. Алгоритм не отличается от уже рассмотренного алгоритма с использованием MAC, который мы рассмотрели выше. Разница только в том, что мы используем последний октет IP-адреса Источника (Source) и Получателя (Destination), которые, если вы посмотрите рассмотренный ранее вариант, никогда не меняются при передаче по сети, в отличие от адресов MAC. Факт того, что IP-адреса не меняются, создает сценарий, в котором у вас больше шансов получить уникальные пары, что, в результате, приводит к более равномерному распределению трафика по физическим линкам.

Для правильного понимания механизма работы следует учесть один важный момент, касающийся пар IP-адресов *source* и *destination*: при вычислении пар *source-destination* принимаются во внимание только последние октеты адресов. Это означает, что в адресе 10.10.1.**10** будет использован для идентификации только **10** – последний октет адреса, в адресе 10.10.3.**100** - только **100**. Принимать во внимание этот момент следует в случае развертывания системы в сети, состоящей из нескольких подсетей, в этом случае может получиться так, что адреса из разных подсетей (но с одинаковым последним октетом) будут передаваться по одному физическому линку.

4.6 IP-алиасы

Понимание принципов работы алгоритмов балансировки нагрузки позволяет администраторам использовать их для более правильного распределения трафика по интерфейсам. Когда используется балансировка по хэшу IP-адреса, можно достичь лучшего распределения трафика по интерфейсам назначив для VIF дополнительные IP-адреса. Любым типам VIF в NetApp, а также физическим интерфейсам, можно назначить произвольное количество адресов IP; IP-адреса кроме первого принято называть IP-алиасами (*aliases*). Типовая рекомендация состоит в том,

чтобы обеспечить соответствие количества IP-адресов количеству физических линков в виртуальном канале между контроллером и коммутатором. Таким образом, если вы используете *multi-mode VIF*, состоящий из 4 линков по 1Gbps между контроллером NetApp и коммутатором, то назначьте VIF первый IP непосредственно, а затем добавьте еще три IP-алиаса.

Но простое назначение дополнительных адресов не приводит само по себе к более равномерному распределению трафика по портам. Хосты, передающие и читающие данные через контроллер системы хранения NetApp, должны использовать все эти адреса для полного использования преимуществ алгоритма балансировки нагрузки на IP-хэше. Этого можно добиться несколькими различными путями, которые зависят от типа использованного протокола. Ниже приведены некоторые примеры в случае использования клиентами протокола NFS.

- **Oracle на NFS**: хосты с Oracle должны монтировать тома NFS равномерно распределяя их по доступным IP-адресам контроллера. Если у вас есть 4 различных NFS ресурса на системе хранения, то смонтируйте их, используя для доступа четыре различных IP-адреса контроллера. Каждый ресурс будет иметь различную пару из источника и получателя (source and destination pair) и полоса передачи между хостом и контроллером будет использована более эффективно.

- **VMware на NFS**: хосты ESX должны монтировать каждый NFS Datastore через различные IP-адреса контроллера NetApp. Такой вариант наилучшим способом использует один интерфейс VMkernel (адрес источника (source)). Если у вас больше датасторов, чем IP-адресов, то просто распределите датасторы по доступным IP-адресам контроллера NetApp поравномернее.

Наконец, когда администраторы назначают IP-алиасы сетевым интерфейсам контроллера NetApp, которые сконфигурированы для высокой доступности, эти IP-алиасы будут перемещены на контроллер-партнер кластера в отключенном (down) состоянии. IP-алиасы не нужно конфигурировать как «партнерские» (partner), если физические интерфейсы уже сами сделаны «партнерскими».

4.7 «Двухслойные» VIF

В ряде случаев, для выполнения задачи может потребоваться избыточность большая, чем предлагают обычные конфигурации VIF. Некоторые производители коммутаторов, например Cisco, предлагают дополнительные возможности, такие как *cross-stack EtherChannel*, *Virtual Switch System* (VSS) или *Virtual Port Channels* (VPC), которые могут позволить администраторам NetApp создавать динамические или статические multi-mode VIF, работающие на нескольких коммутаторах.

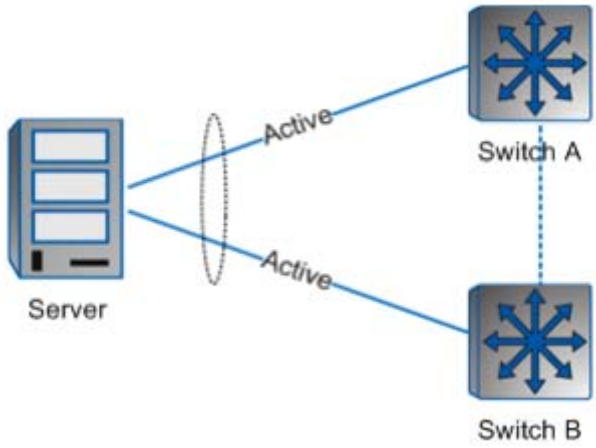

Рис. 4-8) Пример использования VSS или VPC

Для коммутаторов, которые не поддерживают эти возможности, Data ONTAP позволят использовать так называемые *layered* или «двухслойные» VIF-ы, которые представляют из себя соединение VIF-ов «первого уровня» в дополнительный VIF «второго уровня», для обеспечения большей отказоустойчивости и избыточности. Это позволяет администраторам использовать преимущества в производительности и полосе пропускания у *multi-mode VIF*, но также воспользоваться и отказоустойчивостью за счет избыточности в *single-mode VIF*.

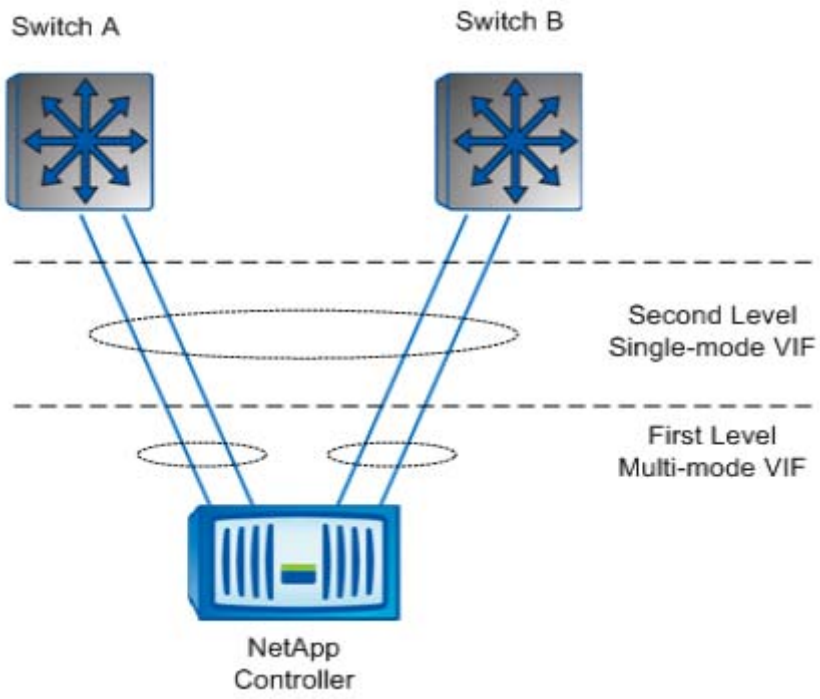

Рис. 4-9) Пример «двухслойного» VIF

Шаблон конфигурации – «двухслойный» VIF

4.8 Рекомендации по связыванию портов

Single-mode VIF

Часто *single-mode VIF* используют для организации раздельных путей к избыточным коммутаторам. Такое решение сохраняет соединение в случае отказа на уровне коммутатора.

Многие производители коммутаторов сегодня поддерживают режим, называемый *multi-chassis LACP*. Такая возможность позволяет конечному устройству быть сконфигурированному как *LACP EtherChannel* при этом участвующие в канале интерфейсы могут быть подключены к разным физическим коммутаторам для достижения необходимой их избыточности и отказоустойчивости.

5 Увеличение производительности с использованием Jumbo Frames

По умолчанию, стандартный фрейм Ethernet несет «полезную нагрузку» (то есть собственно данных) размером 1500 байт. Заголовок Ethernet контрольная сумма CRC добавляют 18 байт, то есть, суммарно, фрейм Ethernet имеет размер 1518 байт, или 1522 байта вместе с тегом VLAN (IEEE 802.3ac).

Так как заголовок и контрольная сумма CRC имеют фиксированный размер, то эффективность передачи может быть увеличена увеличением размера «полезной нагрузки». Изменяя размер MTU (Maximum Transmission Unit) со стандартного размера в 1500 байт до 9000 байт, мы можем создать фреймы большого размера, или *jumbo frames*. Фреймы большего размера эффективно снижают число пакетов, создаваемых для передачи того же объема данных, и, в результате, повышают пропускную способность сети.

Однако необходимо с осторожностью подходить к использованию *jumbo frames* в сети хранения. Ошибки при проектировании или неверная конфигурация оборудования может привести к значительному ухудшению производительности, или даже прекращению нормальной работы сети. Когда вы конфигурируете контроллер NetApp FAS для использования *jumbo frames*, необходимо правильно сконфигурировать следующие элементы:

- Сетевой интерфейс контроллера FAS, и связанные с ним VIF-ы.
- Порт на коммутаторе и интерфейс port-channel, если такой используется.
- VLAN и порты на всех коммутаторах и маршрутизаторах layer 3 между контроллером FAS и его клиентами.

Порты со стандартным значением MTU и jumbo MTU не должны смешиваться в одном VLAN. Допустим, хост и контроллер NetApp FAS сконфигурированы в одну VLAN, при этом контроллер использует *jumbo frames*, а хост - нет. Хост может передавать данные контроллеру FAS с помощью фреймов стандартного размера 1500 байт. Однако ответные фреймы контроллера FAS пойдут размером 9000 байт, и так как два устройства находятся в одной сети VLAN, некому будет фрагментировать эти большие фреймы на фреймы стандартного размера 1500 байт, чтобы их мог получить хост.

Для того, чтобы позволить контроллеру NetApp работать с запросами от хоста как обычного размера, так и с jumbo frame, у нас существует только одна возможность – поместить роутер между контроллером FAS и хостами, такой роутер сможет фрагментировать пакеты большого размера (*jumbo frames*) в несколько мелких, размером 1500 байт. Устройства, использующие *jumbo frames* можно поместить в отдельную VLAN сконфигурированную так, чтобы пропускать *jumbo frames* непосредственно на контроллер NetApp, в то время, как хосты, которые могут работать только с фреймами стандартного размера, можно поместить в VLAN, трафик которого пойдет через роутер, для его фрагментации, как описано выше. Такая конфигурация позволяет

любым хостам работать с контроллером NetApp, на котором включено использованием *jumbo frames*, вне зависимости от того, умеет ли работать с *jumbo frames* система хоста, или нет.

Другой метод состоит в прямом соединении VLAN-ов для трафиков со стандартным размером фрейма, и для *jumbo frame* на разные порты контроллера NetApp. Такой метод имеет свои преимущества в использовании бита DF (*don't fragment*). Вот несколько возможных вариантов использования выделенных VLAN:

- VLAN для сети управления (MTU 1500): используется для SNMP, Operations Manager, SSH, RLM, и т.д. Через эту сеть никогда не идет трафик сети хранения.
- Сеть хранения (MTU 9000): Изолированная, немаршрутизируемая VLAN для NFS, CIFS, или iSCSI.
- Сеть репликации (MTU 9000): Изолированная, немаршрутизируемая VLAN для высокоскоростной репликации, например SnapMirror и SnapVault. Отделение этого трафика позволяет обеспечить более гранулярный мониторинг.
- Межсайтовая репликация (MTU 1500 или менее): Используется для внешних, удаленных резервных копий, которым требуется использование WAN, и в котором могут использоваться различные значения MTU.

Шаблон конфигурации – Jumbo Frames

Примечание

Приведенный шаблон конфигурации показывает пример конфигурации одного интерфейса. При конфигурировании MTU на интерфейсе, являющемся частью port channel, следует принято во внимание, что интерфейс port-channel сам по себе также должен иметь указанное значение MTU на коммутаторе.

NETAPP RC FILE

```
ifconfig e0a 10.1.1.100 netmask 255.255.255.0 mtusize 9000 partner
10.1.1.200 flowcontrol send
```

CISCO IOS SWITCH

```
interface GigabitEthernet1/1
       description NetApp e0a
       switchport access vlan 100
       switchport mode access
       flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       mtu 9198
interface Vlan 100
       ip address 10.1.1.1 255.255.255.0
       mtu 9198
```

Рекомендации по использованию Jumbo Frames

Использование *Jumbo Frames* может значительно поднять производительность Ethernet-сети хранения. Для его правильного использования следует:

- Сконфигурировать использование *Jumbo Frames* на протяжении всей сети, от контроллера системы хранения до серверного хоста
- Отделить трафик с использованием *Jumbo Frames* в отдельную VLAN, чтобы обеспечить оптимальную производительность сетевого интерфейса.

6 Управление «заторами» с помощью Flow Control

Механизмы контроля потока (*flow control*) существуют на разных уровнях модели OSI, включая *window TCP*, *XON/XOFF*, и *FECN/BECN* во *Frame Relay*. В контексте Ethernet, на уровне L2, управление потоком данных было невозможно, пока не появилась возможность передавать данные одновременно в обоих направлениях, то есть в режиме *full duplex*. В режиме *half duplex* линк не может передавать и принимать данные одновременно. Стандарт 802.3X позволяет устройству, использующему соединение «точка-точка», передавать специальный кадр PAUSE, получив который передающее устройство приостанавливает передачу данных. Зарезервированный и определенный в стандарте MAC-адрес 01-80-C2-00-00-01 используется для посылки фреймов PAUSE, в которых указывается длина запрошенной паузы.

Для простых сетей этот метод работает достаточно хорошо. Однако, с появлением все больших сетей, построенных на все более сложном оборудовании и ПО, такие технологии, как *TCP windowing*, увеличенные буфера коммутаторов, и QoS уменьшают потребность в простом контроле потока в сети на уровне L2.

TCP работает с соединением «от одного конечного устройства и до другого», соединение определяет используемое «окно» TCP, с помощью которого приемная сторона может управлять передающей стороной, ограничивая поток В случае возникновения «затора» или потерь пакетов на пути передачи, «окно» TCP будет уменьшаться, чтобы скомпенсировать отсутствие прямого управления потоком. В противоположность этому фрейм PAUSE работает на соединении «точка-точка». Порт коммутатора и сетевой интерфейс самостоятельно решают, когда надо послать фрейм PAUSE и самостоятельно выбирают запрашиваемую длительность паузы, причем он распространяется только на этот линк, от порта коммутатора до сетевого интерфейса. Никакие средства протоколов верхнего уровня при этом не используются. Это может потенциально оказать нежелательное воздействие на производительность TCP, так как создает искусственные задержки между «хопами» и заставляет TCP уменьшать размер «окна» из-за выпавших пакетов. По этой причине обычно не рекомендуется включать *flow control* во всей сети целиком.

Однако некоторые преимущества могут быть получены при включении *flow control* на передаче (*send*) к конечному устройству, подключенному к сети. Современные коммутаторы, как правило, могут работать со скоростью превышающую скорость работы обычных NIC, а также имеют дополнительные возможности по обработке очередей (*queuing*) и буферизации. При установке на коммутаторе для *flow control* «receive on» и «send off», и на клиенте «receive off» и «send on», конечное устройство сможет приостанавливать поток трафика, воздействуя на коммутатор.

Внимание: Когда вы создаете или конфигурируете интерфейс на контроллере NetApp, установка *flow control* по умолчанию «on» как на прием, так и на передачу. Конфигурирование режима «send on» и «receive off» возможно с помощью опций команды *ifconfig*.

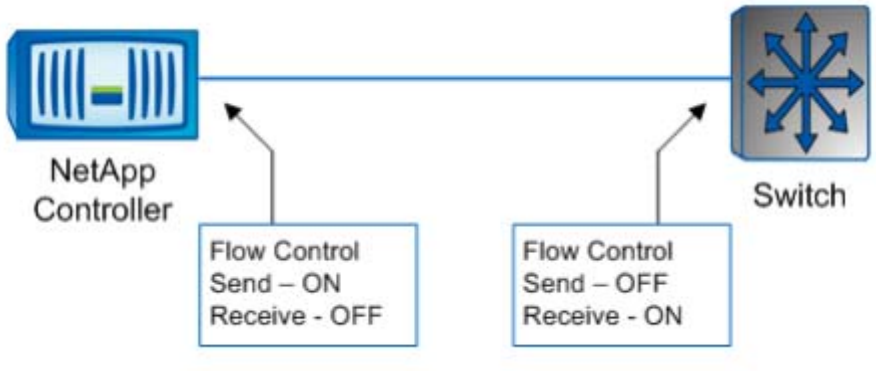

Рис. 6-1) Flow Control

Шаблон конфигурации – Flow Control

NETAPP RC FILE

```
ifconfig e0a 10.1.1.100 netmask 255.255.255.0 mtusize 9000 partner
10.1.1.200 flowcontrol send
```

CISCO IOS SWITCH

```
interface GigabitEthernet1/1
       description NetApp e0a
       switchport access vlan 100
       switchport mode access
flowcontrol receive on
       no cdp enable
       spanning-tree guard loop
       mtu 9198
```

6.1 Рекомендации по использованию Flow Control

Убедитесь, что *flow control* включен и правильно сконфигурирован на обоих концах линка, как на контроллере системы хранения, так и на порту коммутатора, к которому он подключен.

7 Выводы

Все больше и больше критичных для бизнеса данных и приложений используют сети хранения Ethernet. Становится все более важным, чтобы такая сеть работала стабильно на всем протяжении, от контроллера системы хранения, до серверов-получателей данных. Использование для построения такой сети привычных методов построения офисной локальной сети может привести к значительным проблемам в производительности и высокой вероятности сбоев.

Использование технологий VLAN, механизма *fast start* в протоколе *Spanning Tree* (STP), *Muti-mode LACP VIF* для построения отказоустойчивых и высокопроизводительных соединений, *Jumbo Frames* и *Ethernet PAUSE* (aka *flow control*) может помочь построить надежную, высокопроизводительную сеть, необходимую для организации в ней сети хранения с использованием технологий Ethernet и использования Ethernet-based устройств хранения.