Independent Submission
Request for Comments: 7348
Category: Informational
ISSN: 2070-1721

M. Mahalingam
Storvisor
D. Dutt
Cumulus Networks
K. Duda
Arista
P. Agarwal
Broadcom
L. Kreeger
Cisco
T. Sridhar
VMware
M. Bursell
Intel
C. Wright
Red Hat

August 2014

Виртуальные расширяемые ЛВС (VXLAN) - модель наложенных виртуальных сетей L2 в сетях L3

Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks

Аннотация

В этом документе описаны расширяемые виртуальные ЛВС (VXLAN), которые используются для решения задачи построения наложенных сетей в инфраструктуре центров обработки данных, используемых арендаторами. Схема и связанные с ней протоколы могут применяться в сетях «облачных провайдеров» и корпоративных ЦОД. В документе описан протокол развёртывания VXLAN для сообщества Internet.

Статус документа

Этот документ не является спецификацией стандарта Internet и публикуется с информационными целями.

Этот документ является вкладом в серию RFC, независимым от других потоков RFC. Редактор принял решение о публикации документа по своему усмотрению и не делает каких-либо заявлений о пригодности для реализации или развёртывания. Документы, одобренные для публикации редакторами RFC, не претендуют на статус стандарта Internet (см. раздел 2 документа RFC 5741).

Информация о статусе этого документа, обнаруженных ошибках и способах обратной связи доступна по ссылке http://www.rfc-editor.org/info/rfc7348.

Оглавление

1. Введение

Виртуализация серверов повышает уровень требований к физической инфраструктуре сети. Физические серверы поддерживают множество виртуальных машин (VM), каждая из которых имеет свой MAC-адрес[1]. Это требует увеличения размера таблиц MAC-адресов в коммутируемой сети Ethernet по причине подключения сотен тысяч VM, взаимодействующих между собой.

При группировке VM в ЦОД по виртуальным ЛВС (VLAN) могут потребоваться тысячи VLAN для разделения трафика по группам, к которым могут относиться VM. Современное ограничение числа VLAN (4094) не позволяет этого.

В ЦОД зачастую требуется размещать оборудование множества арендаторов, каждый из которых имеет изолированный сетевой домен. Поскольку реализация таких доменов на физически выделенной инфраструктуре экономически нецелесообразна, администраторы сетей выполняют изоляцию доменов в общей сети. В таких случаях проблема заключается в том, что каждый арендатор может независимо выделять MAC-адреса и идентификаторы VLAN, что может приводить к их дублированию.

Важным требованием для виртуализованных сред, использующих физическую инфраструктуру канального уровня (L2), является распространение сети L2 на весь ЦОД или даже на несколько центров с целью эффективного распределения ресурсов (процессоры, сеть, хранилища). В таких сетях применение традиционных подходов вроде протокола связующего дерева (STP) для предотвращения петель может привести к отключению множества каналов.

Современные операторы предпочитают использовать IP для соединения физической инфраструктуры (например, используя ECMP для предотвращения отключения каналов). Но даже в таких средах требуется сохранение модели L2 для коммуникаций между VM.

Описанные выше сценарии требуют организации наложенной сети. Это наложение служит для передачи трафика MAC от отдельных VM с инкапсуляцией в логические туннели.

В этом документе описывается схема расширяемых виртуальных ЛВС VXLAN, которая обеспечивает такую инкапсуляцию для выполнения указанных выше требований. В документе описано развёртывание протокола VXLAN для сообщества Internet.

1.1. Сокращения и определения

ACL
Access Control List - список управления доступом.

ECMP
Equal-Cost Multipath - множество равноценных путей.

IGMP
Internet Group Management Protocol - протокол управления группами в Internet.

IHL
Internet Header Length - размер заголовка Internetю

MTU
Maximum Transmission Unit - максимальный размер передаваемого блока.

PIM
Protocol Independent Multicast - независимая от протокола групповая адресация.

SPB
Shortest Path Bridging - мост по кратчайшему пути.

STP
Spanning Tree Protocol - протокол связующего дерева.

ToR
Top of Rack - наверху в стойке. Термин, используется для обозначения коммутатора, обслуживающего стойку серверов с VM.

TRILL
Transparent Interconnection of Lots of Links - прозрачное соединение множества каналов.

VLAN
Virtual Local Area Network - виртуальная локальная сеть (ЛВС).

VM
Virtual Machine - виртуальная машина.

VNI
VXLAN Network Identifier (or VXLAN Segment ID) - идентификатор сети VXLAN.

VTEP
VXLAN Tunnel End Point - конечная точка туннеля VXLAN. Элемент, служащий началом или окончанием туннеля.

VXLAN
Virtual eXtensible Local Area Network - расширяемая виртуальная ЛВС.

VXLAN Segment – сегмент VXLAN
Наложенная сеть VXLAN L2, через которую взаимодействуют VM.

VXLAN Gateway – шлюз VXLAN
Элемент, пересылающий трафик между VXLAN.

2. Уровни требований

Ключевые слова **необходимо** (MUST), **недопустимо** (MUST NOT), **требуется** (REQUIRED), **нужно** (SHALL), **не нужно** (SHALL NOT), **следует** (SHOULD), **не следует** (SHOULD NOT), **рекомендуется** (RECOMMENDED), **не рекомендуется** (NOT RECOMMENDED), **возможно** (MAY), **необязательно** (OPTIONAL) в данном документе интерпретируются в соответствии с RFC 2119 [RFC2119].

[1]Media Access Control - управление доступом к среде.

3. Предпосылки VXLAN

В этом разделе рассматриваются проблемы, для решения которых предлагается VXLAN. Основное внимание уделено сетевой инфраструктуре ЦОД и связанным с этим вопросам.

3.1. Ограничения со стороны STP и идентификаторов VLAN

Современные сети L2 используют протокол связующего дерева IEEE 802.1D (STP) [802.1D] для предотвращения петель в результате дублирования путей. STP блокирует использование части каналов для предотвращения дублирования и зацикливания кадров. Некоторые операторы ЦОД считают это общей проблемой сетей L2, поскольку STP по сути вынуждает платить за большее число портов, нежели используется реально. Кроме того, отказоустойчивость за счёт использования множества путей недоступна в модели STP. Были предложены новые решения, такие как TRILL [RFC6325] и SPB [802.1aq], помогающие использовать множество путей и решающие некоторые проблемы STP. Ограничения STP можно преодолеть также путём настройки серверов в одной стойке так, чтобы они находились в одной сети L3 и коммутация выполнялась на сетевом уровне как в рамках стойки, так и между стойками. Однако это не совместимо с моделью L2 для соединений между VM.

Основной характеристикой L2 ЦОД является использование VLAN для изоляции областей широковещания. 12-битовое значение VLAN ID в кадрах данных Ethernet позволяет разделить большие сети L2 на множество более мелких доменов широковещания. Такой подход используется во многих ЦОД, где достаточно 4094 VLAN. Однако по мере распространения виртуализации это ограничение начинает мешать. Кроме того, использование STP дополнительно ограничивает число доступных VLAN. К тому же среды с множеством арендаторов требуют большего числа VLAN, как отмечено в параграфе 3.3.

3.2. Среды с множеством арендаторов

Облачные вычисления включают гибкое предоставление ресурсов по запросам в средах с множеством арендаторов. Наиболее распространенным вариантом облачных вычислений является использование публичного облака, в котором сервис-провайдер предлагает гибкий набор услуг множеству заказчиков и арендаторов на базе одной физической инфраструктуры.

Изоляция сетевого трафика по арендаторам может быть выполнена в сети L2 или L3. В сетях L2 для разделения трафика часто используют VLAN - например, арендаторов можно идентифицировать по их номеру VLAN. По причине большого числа арендаторов, которых может обслуживать сервис-провайдер, ограничение в 4094 VLAN зачастую не соответствует реальным потребностям. Кроме того, каждому арендатору часто требуется множество VLAN, что усугубляет проблему.

Другим примером является кросс-контейнерное расширение. Контейнер (pod) обычно состоит из одной или множества серверных стоек с обвязкой в виде сети и хранилища. Арендатор может начать с одного контейнера и по мере роста потребностей запрашивать серверы (VM) в других контейнерах, особенно если арендаторы этих контейнеров не загружают их полностью. В этом варианте требуется «растягивать» среду L2 для связи между серверами и VM.

Сети L3 не обеспечивают всеобъемлющего решения для случаев с множеством арендаторов. Два арендатора могут использовать один набор адресов L3 в своих сетях, что требует от сервис-провайдера иных форм изоляции арендаторов. Кроме того, необходимость использования IP для всех арендаторов исключает заказчиков, которым для связи между VM требуется прямое взаимодействие L2 или протоколы L3, отличающиеся от IP.

3.3. Неадекватный размер таблиц в коммутаторах ToR

Современные среды виртуализации вносят дополнительные требования к таблицам MAC-адресов коммутаторов ToR, обеспечивающих подключение серверов. Вместо одного MAC-адреса на соединение с сервером коммутаторам ToR приходится определять MAC-адреса отдельных VM (которых на одном сервере может быть несколько сотен). Это нужно для того, чтобы трафик между VM и остальной физической сетью проходил через соединение между сервером и коммутатором. Типичный коммутатор ToR соединён с 24 или 48 серверами в зависимости от числа портов на каждом сервере. ЦОД может включать несколько стоек, поэтому каждому коммутатору ToR приходится поддерживать таблицу адресов для связи между VM на разных физических серверах. Это значительно повышает требования к размеру таблиц по сравнению со средами без виртуализации.

Если таблица заполняется, коммутатор перестаёт записывать новые адреса, пока какие-либо из записей таблицы не устареют, а это приводит к значительному росту объёма лавинных рассылок кадров в результате отсутствия информации о получателе.

4. VXLAN

Расширяемые виртуальные ЛВС VXLAN позволяют выполнить описанные выше требования к инфраструктуре L2 и L3 для ЦОД с наличием VM в среде с множеством арендаторов. Они работают на основе имеющейся сетевой инфраструктуры и обеспечивают способы «растягивания» сетей L2. Кратко можно сказать, что VXLAN представляет собой схему наложения L2 на сеть L3. Каждое наложение называется сегментом VXLAN. Взаимодействовать между собой могут лишь VM, относящиеся к одному сегменту VXLAN. Каждый сегмент VXLAN указывается 24-битовым идентификатором, который называется VNI. Это позволяет организовать в одном административном домене до 16M сегментов VXLAN одновременно.

VNI указывает область действия внутреннего кадра MAC, происходящего от конкретной VM. Это позволяет использовать перекрывающиеся адреса MAC в разных сегментах, но их трафик никогда не будет «пересекаться» по причине его изоляции с использованием VNI. Идентификатор VNI размещается во внешнем заголовке, который инкапсулирует внутренний кадр MAC от VM. Далее в документе термины «сегмент VXLAN» и «наложенная сеть VXLAN» взаимозаменяемы.

По причине использования инкапсуляции VXLAN можно считать также схемой туннелирования сетей L2 через сети L3. Туннели не учитывают состояний, поэтому каждый кадр инкапсулируется в соответствии с набором правил. Конечные точки туннелей (VTEP), рассматриваемые ниже, размещаются внутри гипервизора на сервере, содержащем VM. Таким образом, инкапсуляция VNI и VXLAN известна только внутри VTEP и VM не видят её (см. Рисунок 1). Отметим, что

VTEP могут размещаться также на физических коммутаторах или серверах, а реализованы могут быть на программном или аппаратном уровне. Один из вариантов размещения VTEP в физическом коммутаторе рассмотрен в разделе 6.

В последующих параграфах обсуждаются типовые сценарии трафика в среде VXLAN с использованием одного типа схемы управления - обучения в плоскости данных. В этом случае привязка MAC-адресов VM к IP-адресам VTEP определяется на основе изучения адресов отправителей (source-address learning). Для отправки неизвестным получателям, широковещания и групповой рассылки применяется групповая адресация.

В дополнение к плоскости управления на основе обучения возможны и другие схемы распространения информации о связи VTEP IP с VM MAC. Варианты могут включать централизованный поиск в базе или каталоге по отдельным VTEP, распространение данных об отображениях VTEP централизованным объектом и т. п. В этом документе схема обучения плоскости данных будет рассматриваться в качестве плоскости управления для VXLAN.

4.1. Коммуникации между VM по индивидуальным адресам

Рассмотрим VM в наложенной сети VXLAN. Эта VM ничего не будет знать о VXLAN и для связи с VM на другом хосте она будет передавать кадры по MAC-адресу получателя как обычно. VTEP на физическом хосте находит идентификатор VNI, с которым связана данная VM. После этого проверяется принадлежность MAC-адреса получателя к тому же сегменту и наличие отображения этого MAC-адреса на удалённую точку VTEP. Если все найдено, перед исходным кадром MAC добавляется внешний заголовок, состоящий из внешнего MAC, внешнего заголовка IP и заголовка VXLAN (формат кадра показан на рисунке 1). Инкапсулированный пакет передаётся удалённой точке VTEP. При получении пакета удалённая точка VTEP проверяет пригодность VNI и наличие в этой сети VM с MAC-адресом из внутреннего кадра MAC. Если все так, заголовки инкапсуляции удаляются и кадр передаётся целевой VM. Эта VM ничего не знает о VNI и транспортировке кадра с использованием инкапсуляции VXLAN.

В дополнение к пересылке пакета целевой VM удалённая точка VTEP узнает отображение внутреннего MAC-адреса отправителя на его внешний адрес IP. Эти отображения сохраняются в таблице и при передаче целевой VM пакета с откликом уже не потребуется широковещательная передача «неизвестному адресату».

Определение MAC-адреса целевой VM до передачи пакета исходной VM выполняется в средах без VXLAN за исключением случаев, описанных в параграфе 4.2. Применяются широковещательные кадры, которые инкапсулируются с использованием группового адреса, как описано в параграфе 4.2.

4.2. Широковещательные коммуникации и отображение их на групповые

Рассмотрим VM на хосте-источнике, пытающуюся взаимодействовать с целевой VM по протоколу IP. В предположении что обе машины расположены в одной подсети, VM передаёт широковещательный кадр ARP[1]. В среде без VXLAN этот кадр будет передан с использованием широковещания MAC через все коммутаторы, используемые этой VLAN.

В случае VXLAN в начало пакета будет помещаться VXLAN VNI вместе с заголовками IP и UDP. Однако этот широковещательный пакет будет передан в multicast-группу IP, служащую для реализации наложенной сети VXLAN.

Для этого нужно иметь отображение между VXLAN VNI и используемой группой IP. Это отображение выполняется в плоскости управления и предоставляется отдельным точкам VTEP через канал управления. Используя отображение, VTEP может предоставить информацию о принадлежности IGMP восходящему коммутатору или маршрутизатору для включения (выхода) в группы IP multicast, связанные с VXLAN. Это позволить «обрезать листья» узлов для конкретных групповых адресов на основе информации о доступности членов группы на этом хосте при использования конкретного группового адреса (см. [RFC4541]). В дополнение к этому использование протоколов групповой маршрутизации, таких как PIM-SM[2] (см. [RFC4601]), обеспечит эффективное построение дерева группы в сети L3.

VTEP будет использовать присоединение (*,G). Это обусловлено тем, что набор источников туннелей VXLAN не известен и может часто изменяться, поскольку VM могут перемещаться с одного хоста на другой. Следует отметить, что каждая точка VTEP может выступать в качестве источника и получателя для групповых пакетов, поэтому более эффективным будет использование таких протоколов, как двухсторонний PIM (BIDIR-PIM, см. [RFC5015]).

Целевая VM передаёт стандартный отклик ARP с использованием индивидуального адреса IP. Этот кадр будет инкапсулирован обратно в точку VTEP, соединяющую с исходной VM, с использованием инкапсуляции VXLAN и индивидуального адреса IP. Это возможно, поскольку отображение MAC-адреса получателя отклика ARP на IP-адрес конечной точки туннеля VXLAN было определено раньше из запроса ARP.

Отметим, что групповые кадры и кадры с «неизвестным MAC-адресом получателя» также передаются с использованием группового дерева, подобно широковещательным кадрам.

4.3. Требования к физической инфраструктуре

При использовании групповой рассылки IP в сетевой инфраструктуре могут применяться протоколы групповой маршрутизации (такие как PIM-SM) в отдельных маршрутизаторах и коммутаторах L3 для построения эффективного дерева групповой пересылки, позволяющего передавать групповые кадры только заинтересованным в них узлам.

Здесь нет требования поддержки L3 в реальной сети, используемой для соединения между VM, поскольку VXLAN может работать и в сетях L2. В любом случае эффективная групповая репликация в сети L2 может быть обеспечена с помощью отслеживания IGMP.

Конечным точкам VTEP **недопустимо** фрагментировать пакеты VXLAN. Промежуточные маршрутизаторы могут фрагментировать инкапсулированные пакеты VXLAN по причине увеличения размера кадров. Целевая точка VTEP **может** отбрасывать такие фрагменты VXLAN без уведомления отправителя. Для обеспечения сквозной доставки трафика без фрагментирования **рекомендуется** устанавливать значения MTU[3] в сетевой архитектуре с учетом увеличения размера кадров при инкапсуляции. Для решения этой проблемы **можно** использовать и другие методы, такие как определение Path MTU (см. [RFC1191] и [RFC1981]).

[1]Address Resolution Protocol - протокол преобразования адресов.
[2]Protocol Independent Multicast - Sparse Mode
[3]Maximum Transmission Unit - максимальный размер передаваемого блока.

5. Формат кадров VXLAN

Формат кадров VXLAN описан ниже. Анализ выполняется «со дна» кадра, от внешней контрольной суммы FCS[1], где размещается внутренний кадр MAC со своим заголовком Ethernet, включающим MAC-адреса отправителя и получателя, тип Ethernet и необязательный тег VLAN. Обработка внутренних тегов VLAN описана в разделе 6.

Внутренний кадр MAC инкапсулируется с использованием 4 заголовков, описанных ниже, начиная с внутреннего.

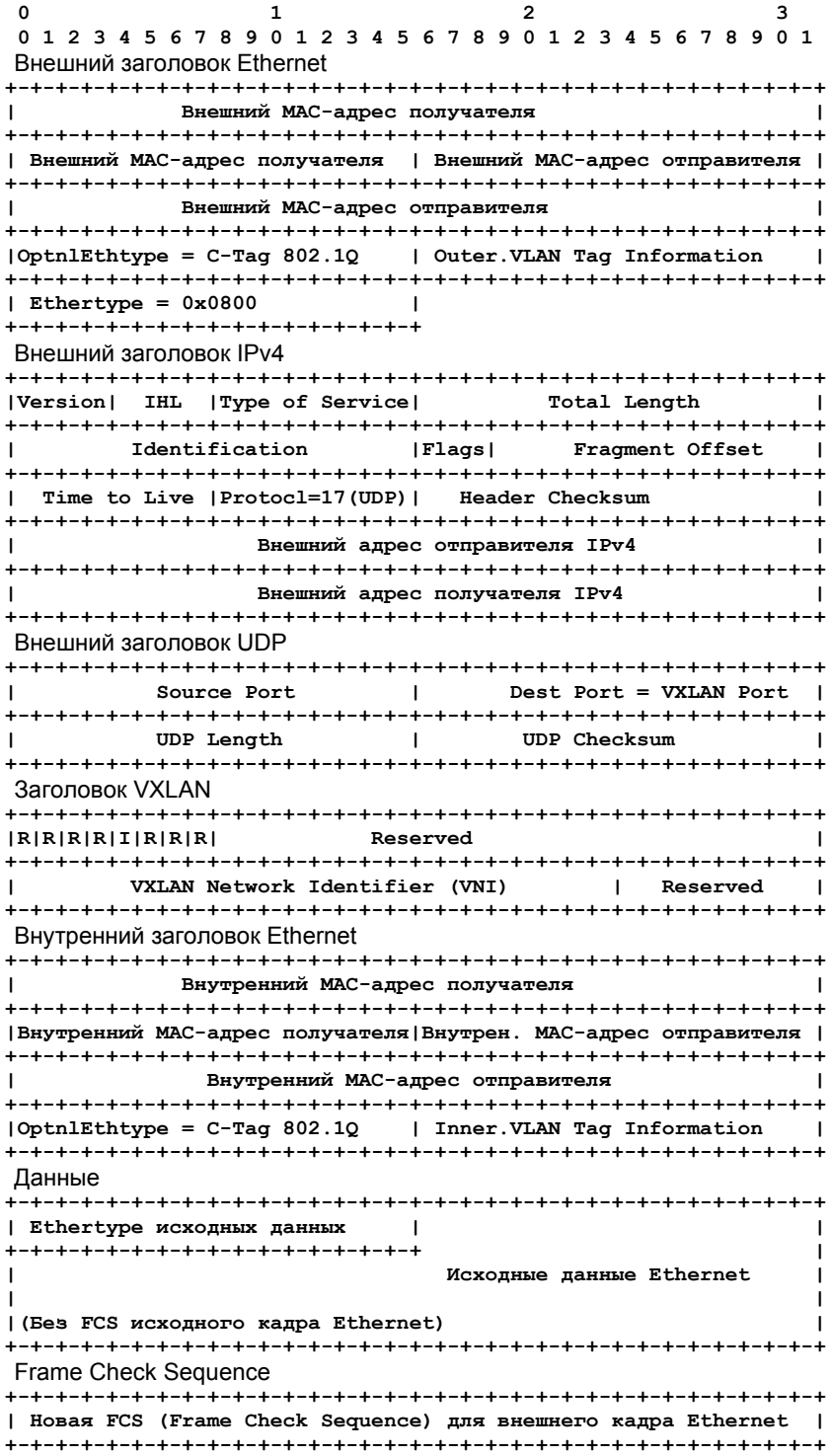

Рисунок 1. Формат кадра VXLAN с внешним заголовком IPv4.

Заголовок VXLAN

Это 8-байтовое поле включает перечисленные ниже поля.

- **Flags** (8 битов). Флаг I **должен** быть установлен (1) для корректного идентификатора сети VXLAN (VNI). Остальные 7 битов (R) являются резервными и **должны** сбрасываться (0) на передающей стороне и игнорироваться на приёмной.
- **VXLAN Segment ID/VXLAN Network Identifier (VNI)**. 24-битовое значение, используемое для указания наложенной сети VXLAN, в которой размещаются взаимодействующие VM. Расположенные в разных наложенных VXLAN виртуальные машины (VM) не могут взаимодействовать.
- **Reserved** (24 и 8 битов). **Должны** сбрасываться (0) на передающей стороне и игнорироваться на приёмной.

Внешний заголовок UDP

Во внешнем заголовке UDP указывается порт отправителя, представленный VTEP, и общеизвестный (стандартный) порт получателя UDP.

[1]Frame Check Sequence - последовательность проверки кадра (контрольная сумма).

- **Destination Port**. Агентство IANA выделило номер 4789 для порта VXLAN UDP и это значение **следует** использовать по умолчанию в качестве порта получателя UDP. Некоторые ранние реализации VXLAN используют для этого другой номер порта. Для корректного взаимодействия номер порта получателя **следует** делать настраиваемым.
- **Source Port**. Рекомендуется выбирать номер порта отправителя UDP с использованием хэш-значения полей внутреннего пакета (например, полей заголовка внутреннего кадра Ethernet). Это обеспечивает должный уровень энтропии для ECMP и распределения нагрузки трафика между VM через наложенную сеть VXLAN. При таком расчёте порта UDP **рекомендуется** выбирать значения из диапазона 49152-65535 [RFC6335].

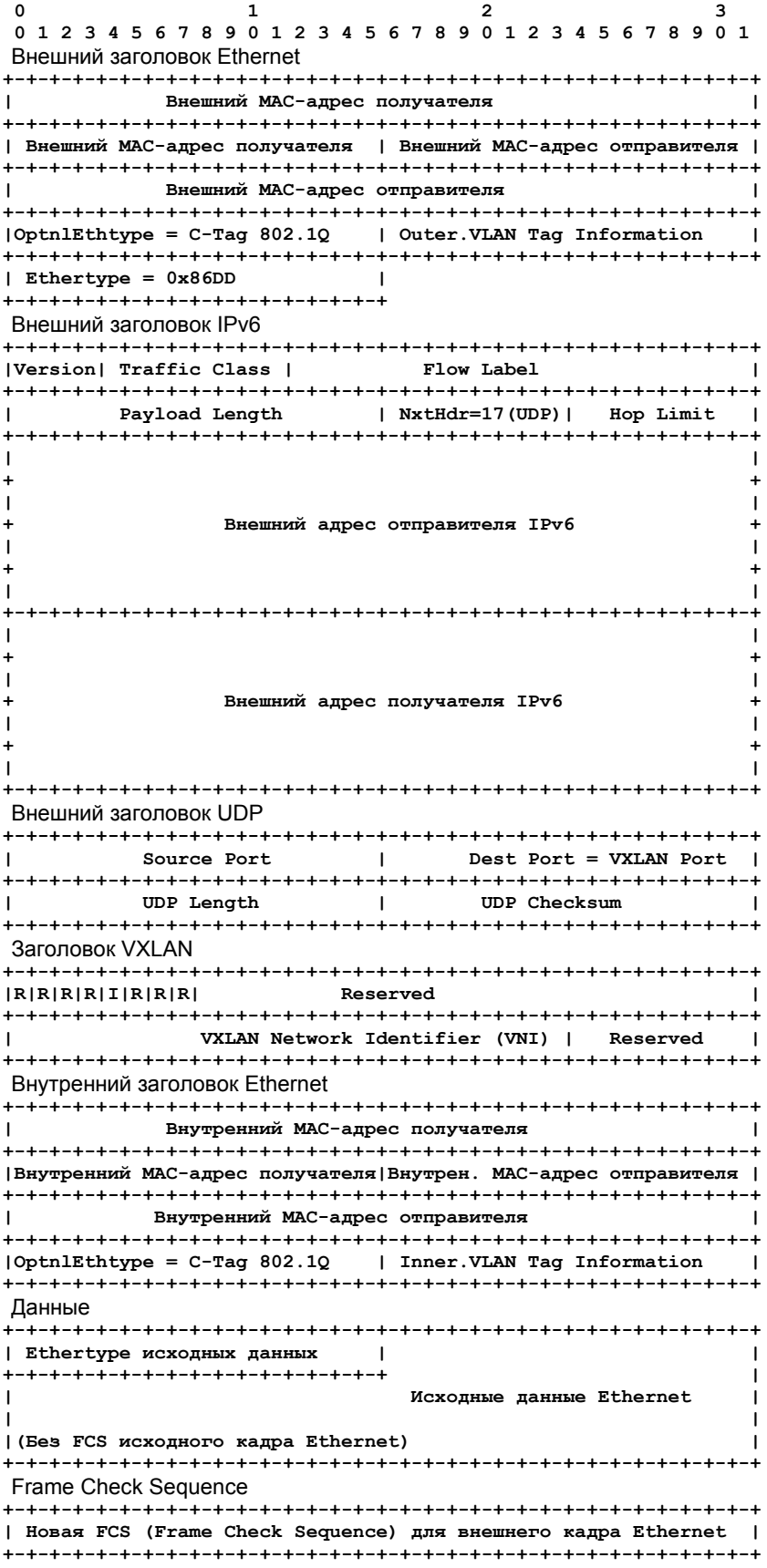

Рисунок 2. Формат кадра VXLAN с внешним заголовком IPv6.

- **UDP Checksum**. Пакеты **следует** передавать с нулевым значением контрольной суммы UDP. При получении такого пакета он **должен** восприниматься для декапсуляции. Если инкапсулирующая точка включает ненулевую контрольную сумму UDP, эта сумма **должна** учитывать весь пакет, включая заголовки IP, UDP, VXLAN и инкапсулированный кадр MAC. При получении пакета с ненулевой контрольной суммой точка

декапсуляции **может** проверить значение суммы. Если рассчитанное значение не совпадает с полученным в пакете, такой пакет **должен** отбрасываться. Если точка декапсуляции не проверяет контрольную сумму или проверка дала положительный результат, пакет **должен** восприниматься для декапсуляции.

Внешний заголовок IP

Это внешний заголовок IP с полем адреса отправителя, содержащим IP-адрес точки VTEP, через которую работает VM (указанная MAC-адресом отправителя во внутреннем заголовке). IP-адресом получателя может быть индивидуальный или групповой адрес IP (см. параграфы 4.1 и 4.2). При указании индивидуального адреса он представляет IP-адрес точки VTEP, соединённой с VM, представленной MAC-адресом получателя во внутреннем заголовке. Детали для случая групповой адресации представлены в параграфе 4.2.

Внешний заголовок Ethernet (пример)

На рисунке 1 приведён пример внутреннего кадра Ethernet, инкапсулированного с использованием заголовков Ethernet, IP, UDP и VXLAN. Внешний MAC-адрес получателя в этом кадре может быть адресом целевой точки VTEP или промежуточного маршрутизатора L3. Внешний тег VLAN является необязательным. При наличии тега он может использоваться для организации трафика VXLAN в ЛВС.

На рисунке 1 показано туннелирование кадров Ethernet с использованием IPv4. Инкапсуляция VXLAN для случая IPv6 показана на рисунке 2.

6. Сценарии развёртывания VXLAN

VXLAN обычно развёртывается в центрах обработки данных на виртуализованных хостах, которые могут быть распределены по нескольким стойкам. Отдельные стойки могут быть в разных сетях L3 или в общей сети L2. Сегменты (наложенные сети) VXLAN организуются «поверх» этих сетей L2 или L3.

На рисунке 3 представлены два сервера виртуализации, подключённые к инфраструктуре L3. Серверы могут размещаться в одной или разных стойках и даже в разных ЦОД одного административного домена. Имеется 4 наложенных сети VXLAN с VNI 22, 34, 74 и 98. Рассмотрим VM1-1 на сервере 1 и VM2-4 на сервере 2, которые относятся к одной наложенной сети VXLAN с VNI 22. VM на знают о наложенных сетях и методах доставки, поскольку инкапсуляция и декапсуляция выполняются точками VTEP на серверах 1 и 2. Другими наложенными сетями являются VNI 34 с VM1-2 на сервере 1 и VM2-1 на сервере 2, VNI 44 с VM1-3 на сервере 1 и VM2-2 на сервере 2, а также VNI 98 с VM1-4 на сервере 1 и VM2-3 на сервере 2.

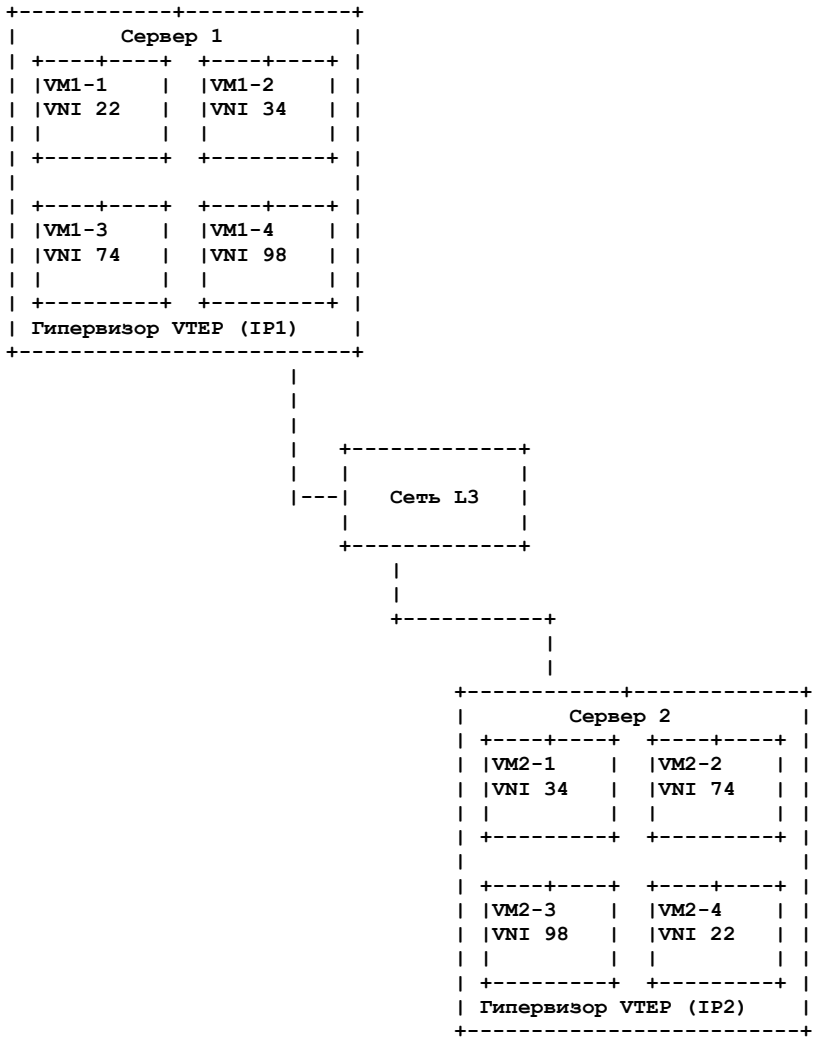

Рисунок 3. Развёртывание VXLAN - VTEP через сеть L3.

В одном из вариантов точками завершения туннелей являются физические серверы, поддерживающие VXLAN. В другом варианте узлам наложенной сети VXLAN требуется взаимодействие с узлами традиционной сети, которая может использовать VLAN. Эти узлы могут быть физическими или виртуальными машинами. Для обеспечения таких коммуникаций сеть может включать шлюзы VXLAN (см. рисунок 4, где в качестве шлюза VXLAN служит коммутатор), которые пересылают трафик между средами с поддержкой VXLAN и без неё.

Обратимся далее к рисунку 4. Для входящих кадров на подключённом к VXLAN интерфейсе шлюз исключает заголовок VXLAN и пересылает пакет в физический порт по MAC-адресу получателя во внутреннем кадре Ethernet.

Декапсулированные кадры с внутренним VLAN ID **следует** отбрасывать, если для них явно не задана пересылка на интерфейс без поддержки VXLAN. В обратном направлении входящие кадры на интерфейсе без VXLAN отображаются на конкретную наложенную сеть VXLAN в соответствии со значением VLAN ID в кадре. Если явно не задана передача идентификатора в инкапсулированный кадр VXLAN, значение VLAN ID удаляется из кадра перед его инкапсуляцией в VXLAN.

Такими шлюзами, обеспечивающими функции завершения туннелей VXLAN, могут быть коммутаторы ToR или коммутаторы доступа, а также коммутаторы вышележащего уровня топологии сети ЦОД, например, коммутаторы ядра или даже граничные устройства WAN. Последний случай (граничное устройство WAN) может включать маршрутизатор PE[1], на котором завершаются туннели VXLAN в гибридной облачной среде. Во всех этих случаях функциональность шлюза может быть реализована на программном или аппаратном уровне.

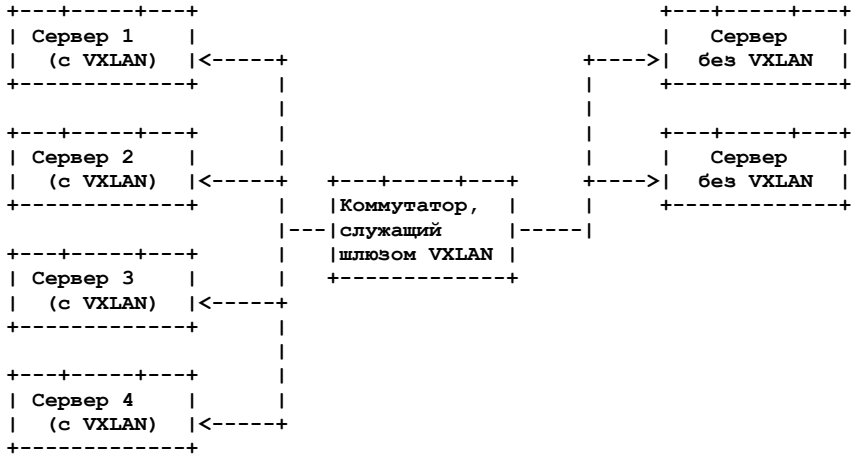

Рисунок 4. Развёртывание VXLAN - шлюз VXLAN.

6.1. Обработка внутренних тегов VLAN

Обработка внутренних тегов VLAN в точках VTEP и шлюзах VXLAN должна соответствовать приведённым ниже требованиям.

Декапсулированные кадры VXLAN с внутренним тегом VLAN **следует** отбрасывать, если не задано иное поведение. На стороне инкапсуляции точке VTEP **не следует** включать внутренний тег VLAN в туннелируемый пакет, если не задано иное поведение. Когда для туннелирования VXLAN представлен пакет с тегом VLAN, инкапсулирующей точке VTEP **следует** вырезать этот тег, если не задано иное поведение.

7. Вопросы безопасности

Традиционные сети L2 можно атаковать лишь изнутри путём получения доступа в ЛВС и перехвата трафика, внедрения обманных пакетов с захватом MAC-адреса или лавинной рассылки для отказа в обслуживании. Механизм MAC-over-IP для доставки трафика L2 существенно расширил «фронт атаки». Сейчас злоумышленники могут внедриться в сеть, подписавшись на одну или несколько multicast-групп, используемых для передачи широковещательного трафика сегментов VXLAN, а также передавая кадры MAC-over-UDP в транспортную сеть для внедрения ложного трафика (возможно для захвата MAC-адресов).

Этот документ не включает конкретных мер противодействия таким атакам, полагаясь на традиционные механизмы, работающие на основе IP. В этом разделе очерчены некоторые подходы к обеспечению защиты в среде VXLAN.

Традиционные атаки L2 с использованием враждебных конечных точек могут быть ослаблены путём ограничения возможностей управления и администрирования для тех, кто разворачивает и поддерживает VM и шлюзы в среде VXLAN. Такие административные меры можно дополнительно усилить схемами типа 802.1X [802.1X] для контроля доступа конкретных конечных точек. Использование инкапсуляции UDP в VXLAN позволяет применять функциональность 5-компонентных ACL[2] в физических коммутаторах.

Туннелирование трафика через сеть IP можно защитить с помощью традиционных механизмов (например, IPsec), позволяющих контролировать подлинность трафика VXLAN и при необходимости шифровать его. Разумеется, для этого будет нужна инфраструктура аутентификации конечных точек, поддерживающая получение и распространение свидетельств (credentials).

Наложенные сети VXLAN создаются и функционируют на основе имеющейся инфраструктуры ЛВС. Для уверенности в том, что конечные точки VXLAN и их VTEP имеют требуемые полномочия в ЛВС, рекомендуется выделить VLAN для трафика VXLAN, а серверам и VTEP передавать трафик VXLAN только через эту VLAN для обеспечения защиты.

В дополнение к сказанному, VXLAN требует подходящего отображения VNI и членства VM в наложенных сетях. Предполагается, что такое отображение будет создаваться и поддерживаться элементом управления с точках VTEP и шлюзах с использованием имеющихся методов защиты.

8. Взаимодействие с IANA

Стандартный порт UDP 4789 выделен агентством IANA в реестре «Service Name and Transport Protocol Port Number» для VXLAN. Обсуждение этого вопроса приведено в разделе 5.

[1]Provider Edge.
[2]Access Control List - список управления доступом.

9. Литература

9.1. Нормативные документы

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

9.2. Дополнительная литература

[802.1aq] IEEE, "Standard for Local and metropolitan area networks – Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks -- Amendment 20: Shortest Path Bridging", IEEE P802.1aq-2012, 2012.

[802.1D] IEEE, "Draft Standard for Local and Metropolitan Area Networks/ Media Access Control (MAC) Bridges", IEEE P802.1D-2004, 2004.

[802.1X] IEEE, "IEEE Standard for Local and metropolitan area networks -- Port-Based Network Acces Control", IEEE Std 802.1X-2010, February 2010.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, November 1990.

[RFC1981] McCann, J., Deering, S., and J. Mogul, "Path MTU Discovery for IP version 6", RFC 1981, August 1996.

[RFC4541] Christensen, M., Kimball, K., and F. Solensky, "Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches", RFC 4541, May 2006.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

[RFC6325] Perlman, R., Eastlake 3rd, D., Dutt, D., Gai, S., and A. Ghanwani, "Routing Bridges (RBridges): Base Protocol Specification", RFC 6325, July 2011.

[RFC6335] Cotton, M., Eggert, L., Touch, J., Westerlund, M., and S. Cheshire, "Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry", BCP 165, RFC 6335, August 2011.

10. Благодарности

Авторы признательны Ajit Sanzgiri за вклад в раздел «7. Вопросы безопасности» и редакторские правки, Joseph Cheng, Margaret Petrus, Milin Desai, Nial de Barra, Jeff Mandin и Siva Kollipara за рецензирование, комментарии и предложения.

Адреса авторов

Mallik Mahalingam
Storvisor, Inc.
640 W. California Ave, Suite #110
Sunnyvale, CA 94086.
USA
EMail: mallik_mahalingam@yahoo.com

Dinesh G. Dutt
Cumulus Networks
140C S. Whisman Road
Mountain View, CA 94041
USA
EMail: ddutt.ietf@hobbesdutt.com

Kenneth Duda
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
USA
EMail: kduda@arista.com

Puneet Agarwal
Broadcom Corporation
3151 Zanker Road
San Jose, CA 95134
USA
EMail: pagarwal@broadcom.com

Lawrence Kreeger
Cisco Systems, Inc.
170 W. Tasman Avenue
San Jose, CA 95134
USA
EMail: kreeger@cisco.com

T. Sridhar
VMware, Inc.
3401 Hillview
Palo Alto, CA 94304
USA
EMail: tsridhar@vmware.com

Mike Bursell
Intel
Bowyer's, North Road
Great Yeldham
Halstead
Essex. C09 4QD
UK
EMail: mike.bursell@intel.com

Chris Wright
Red Hat, Inc.
100 East Davie Street
Raleigh, NC 27601
USA
EMail: chrisw@redhat.com

Перевод на русский язык

Николай Малых

nmalykh@protokols.ru