Internet Engineering Task Force (IETF)
Request for Comments: 8926
Category: Standards Track
ISSN: 2070-1721

J. Gross, Ed.

I. Ganga, Ed.
Intel
T. Sridhar, Ed.
VMware

November 2020

Geneve: Generic Network Virtualization Encapsulation

Geneve - базовая инкапсуляция виртуализации сетей

Аннотация

Виртуализация сетей включает взаимодействие устройств с широким спектром возможностей, таких как программные и аппаратные конечные точки туннелей, транзитные «фабрики» и централизованные управляющие кластеры. Туннели связывают между собой различные элементы, поэтому требования к туннелям зависят от этих компонентов и гибкость становится наиболее важным аспектом протоколов туннелирования, если он должен соответствовать развитию технологий. В этом документе описан протокол инкапсуляции Geneve, разработанный для распознавания меняющихся потребностей и возможностей и адаптации к ним.

Статус документа

Документ относится к категории Internet Standards Track.

Документ является результатом работы IETF[1] и представляет согласованный взгляд сообщества IETF. Документ прошёл открытое обсуждение и был одобрен для публикации IESG[2]. Дополнительную информацию о стандартах Internet можно найти в разделе 2 в RFC 7841.

Информацию о текущем статусе документа, ошибках и способах обратной связи можно найти по ссылке https://www.rfc-editor.org/info/rfc8926.

Оглавление

[1] Internet Engineering Task Force - комиссия по решению инженерных задач Internet.
[2] Internet Engineering Steering Group - комиссия по инженерным разработкам Internet.

1. Введение

Сети уже давно используют различные механизмы туннелирования, теги и другие способы инкапсуляции. Однако виртуализация сетей вызвала всплеск интереса и соответствующий рост разработки внедрения новых протоколов. Большое число таких протоколов от VLAN [IEEE.802.1Q_2018] и MPLS [RFC3031] до более новых VXLAN[1] [RFC7348] и NVGRE[2] [RFC7637] зачастую вызывает вопрос о необходимости разработки нового формата инкапсуляции и причинах быстрого распространения виртуализации сетей. Отметим, что представленный выше список протоколов не является исчерпывающим.

В то время как многие протоколы инкапсуляции предназначены просто для сегментации базовой сети или организации мостов между доменами, виртуализация представляет транзитную сеть как средство обеспечения связности между многочисленными компонентами распределенной системы. Во многих случаях такая система похожа на модульный коммутатор базовой сети IP, играющий роль сетевой магистрали, в шасси которого установлены линейные платы, играющие роль конечных точек туннелей. При таком представлении требования к туннельным протоколам существенно различаются по части объёма требуемых метаданных и роли транзитных узлов.

В таких работах, как «VL2: A Scalable and Flexible Data Center Network» [VL2] и «NVO3 Data Plane Requirements» [NVO3-DATAPLANE] описаны некоторые свойства, которые плоскость данных должна обеспечивать для поддержки сетевой виртуализации. Однако имеется дополнительное определяющее требование передачи метаданных (например, состояние системы) вместе с данными пакетов (примеры использования метаданных приведены ниже). Использование метаданных не является новинкой и почти все протоколы, применяемые для виртуализации сетей, содержат по меньшей мере 24-битовое пространство идентификаторов как способ разделения арендаторов. Это часто описывается как преодоление присущего VLAN ограничения в 12 битов, поскольку 16 миллионов идентификаторов вполне достаточно для обозначения идентификаторов практически в любом контексте. Однако на практике метаданные не ограничиваются идентификаторами арендаторов и представление других сведений приводит к быстрому заполнению доступного пространства. Фактически, по сравнению с тегами, применяемыми для обмена метаданными между линейными платами в шасси коммутатора 24-битовые идентификаторы начинают казаться достаточно мелкими. Применение метаданных практически не имеет границ и они служат для разных целей от хранения номера входного порта для простой политики безопасности до передачи связанного со службой контекста для расширенных приложений на промежуточных устройствах, завершающих или заново инкапсулирующих трафик Geneve.

Каждый из имеющихся протоколов туннелирования пытается разные вопросы, связанные с новыми требованиями, но быстро устаревает по мере развития и реализации изменяющихся плоскостей данных. Кроме того, программные и аппаратные компоненты и контроллеры имеют свои преимущества и разную скорость развития, что следует считать преимуществом, а не ограниченностью или недостатком. В этом документе описан протокол Geneve, стремящийся избежать отмеченных проблем за счёт обеспечения схемы туннелирования для виртуализации сети вместо полного переписывания системы.

1.1. Уровни требований

Ключевые слова **необходимо** (MUST), **недопустимо** (MUST NOT), **требуется** (REQUIRED), **нужно** (SHALL), **не следует** (SHALL NOT), **следует** (SHOULD), **не нужно** (SHOULD NOT), **рекомендуется** (RECOMMENDED), **не рекомендуется** (NOT RECOMMENDED), **возможно** (MAY), **необязательно** (OPTIONAL) в данном документе интерпретируются в соответствии с BCP 14 [RFC2119] [RFC8174] тогда и только тогда, когда они выделены шрифтом, как показано здесь.

1.2. Терминология

Схема виртуализации сети через уровень L3 (Network Virtualization over Layer 3 или NVO3) [RFC7365] определяет множество концепций, обычно применяемых в сфере виртуализации сетей. Ниже приведены дополнительные определения используемых в этом документе терминов.

Checksum offload - выгрузка операций с контрольной суммой

Оптимизация, выполняемая во многих сетевых адаптерах (NIC) для переноса операций расчёта и проверки контрольных сумм в оборудование при передаче и приёме пакетов, соответственно. Обычно это включает контрольные суммы IP, TCP и UDP, которые без выгрузки обрабатываются программно в стеке протоколов.

[1]Virtual eXtensible Local Area Network - виртуальная расширяемая ЛВС.
[2]Network Virtualization Using Generic Routing Encapsulation - виртуализация сетей с использованием GRE.

Clos network - сеть Clos

Метод организации сетевых структур, выходящих за пределы одного коммутатора и поддерживающих неблокируемую пропускную способность между точками подключения. Для деления трафика между несколькими каналами используется механизм ECMP и коммутаторы, образующие структуру. Иногда для обозначения применяются термины топология leaf and spine (лист и ствол) и «дерево судьбы» (fat tree).

ECMP (Equal Cost Multipath) - множество равноценных путей

Механизм маршрутизации для выбора из множества лучших путей к следующему узлу пересылки на основании хэширования заголовков пакетов для более эффективного использования пропускной способности без нарушения порядка следования пакетов в потоке.

Geneve (Generic Network Virtualization Encapsulation) - базовая инкапсуляция виртуализации сети

Протокол туннелирования, описанный в этом документе.

LRO (Large Receive Offload) - выгрузка операций приёма с большими блоками данных

Эквивалент функции LSO на стороне получателя, когда несколько протокольных сегментов (в основном TCP) собирается в более крупный блок данных.

LSO (Large Segmentation Offload) - выгрузка сегментации больших блоков данных

Функция, обеспечиваемая многими коммерческими NIC и позволяющая передавать сетевому адаптеру блоки данных размером больше MTU с целью повышения производительности. NIC отвечает за создание более мелких сегментов, размер которых не превышает значение MTU, с корректными протокольными заголовками. В контексте TCP/IP эту функцию часто называют TSO (TCP Segmentation Offload - выгрузка сегментации TCP).

Middlebox - промежуточное устройство

В контексте этого документа middlebox обозначает специальное устройство, выполняющее функции сетевых служб или взаимодействия служб, которое обычно реализует функциональность конечной точки туннели, завершая и заново инкапсулируя трафик Geneve.

NIC (Network Interface Controller) - контроллер сетевого интерфейса

Иногда используются также термины Network Interface Card (плата сетевого интерфейса) и Network Adapter (сетевой адаптер). NIC может быть частью конечной точки туннеля или транзитного устройства и может обрабатывать или способствовать обработке пакетов Geneve.

Transit device - транзитное устройство

Элемент пересылки (например, маршрутизатор или коммутатор) на пути туннеля, составляющий часть базовой сети. Промежуточное устройство может понимать формат пакетов Geneve, но не порождает и не воспринимает пакеты Geneve.

Tunnel endpoint – конечная точка туннеля

Элемент, выполняющий инкапсуляцию и декапсуляцию пакетов, таких как кадры Ethernet или дейтаграммы IP, с заголовками Geneve. Будучи конечным потребителем метаданных туннеля, конечная точка предъявляет самый высокий уровень требований к синтаксическому анализу и интерпретации заголовков. Конечная точка может быть реализована программно, аппаратно или представлять комбинацию этих вариантов. Конечные точки туннелей часто служат компонентами периметра виртуализации сети (Network Virtualization Edge или NVE), но могут размещаться на промежуточных устройствах и других элементах, делая их частью сети NVO3.

VM (Virtual Machine) - виртуальная машина

2. Требования к разработке

Протокол Geneve разработан для поддержки вариантов виртуализации сети в среде ЦОД. В этих ситуациях туннели обычно служат «магистральными соединениями» между виртуальными коммутаторами, находящимися в гипервизорах, физическими коммутаторами и промежуточными устройствами или другими системами. В качестве базовой может служить любая сеть IP, хотя обычно применяются сети Clos с использованием каналов ECMP для обеспечения согласованной пропускной способности между всеми точками подключения. Многие концепции наложения виртуализации на сети IP описаны в модели NVO3 [RFC7365]. На рисунке 1 представлен пример гипервизора, коммутатора ToR (top-of-rack) для подключения физических серверов и канала в WAN, соединённых с помощью туннелей Geneve через упрощённую сеть Clos. Туннели служат для инкапсуляции и пересылки кадров от подключённых элементов, таких как VM и физические каналы.

```
+---------------------+           +--------+  +-------+
| +--+  +--------+---+ |           |Промеж. |--|Коммут.|==Физич.
| |VM|--|        |   | | +-------+ /|маршрут.|  | ToR   |==серверы
| +--+  |Виртуал.|NIC|---|Коммут.|/ +--------+\/+-------+
| +--+  |коммут. |   | | | ToR   |\ +--------+/\+-------+
| |VM|--|        |   | | +-------+ \|Промеж. |  |Восход.|   WAN
| +--+  +--------+---+ |           |маршрут.|--| канал |=========>
+---------------------+           +--------+  +-------+
       Гипервизор

             ()===================================()
                 Туннели Geneve между коммутаторами
```

Рисунок 1. Пример развертывания Geneve.

Для поддержки потребностей виртуализации сети протоколу туннелирования следует обеспечивать возможность использования преимуществ различных (и развивающихся) свойств каждого типа устройств в базовой и наложенной сети. В результате к протоколу туннелирования предъявляется ряд требований:

- универсальная, расширяемая плоскость данных для поддержки имеющихся и будущих плоскостей управления;
- эффективная реализация элементов туннеля на программном и аппаратном уровне без ограничения возможностями наиболее слабого элемента;
- поддержка высокой производительности при работе по имеющимся сетям IP.

Эти требования более подробно описаны в последующих параграфах.

2.1. Независимость от плоскости управления

Хотя некоторые протоколы сетевой виртуализации включают плоскость управления как часть спецификации формата туннелей (наиболее ярко это проявляется в VXLAN [RFC7348], где предписана плоскость управления на основе изучения группового трафика), в основном их спецификации описывают лишь формат данных. Формат пакетов VXLAN фактически работает с разными плоскостями управления, построенные на его базе.

Стабилизация формата данных обеспечивает явное преимущество, поскольку большинство протоколов отличается лишь внешне и дублирование усилий не даёт большого преимущества. Однако этого нельзя сказать о плоскости управления, в которой могут быть фундаментальные отличия. Аргументы в пользу стандартизации здесь менее очевидны с учётом разнообразия требований, целей и вариантов развёртывания.

В силу упомянутых причин протокол Geneve задаёт лишь спецификацию туннельного формата, которая может удовлетворить требования многих плоскостей управления, явно не задавая их. Это также способствует продвижению общего формата данных и снижает шансы на устаревание из-за внедрения новых плоскостей управления.

2.2. Расширяемость плоскости данных

Для достижения уровня гибкости, требуемого в поддержку имеющихся и будущих плоскостей управления, нужна инфраструктура опций, которая позволит эффективно определять, развёртывать, а также завершать или исключать новые типы метаданных. Опции также позволяют дифференцировать продукцию, поощряя независимую разработку по основной специализации каждого производителя, что в целом приведёт к ускорению развития. Безусловно, наиболее распространенным механизмом реализации опций будет формат TLV (Type-Length-Value - тип, размер, значение).

Хотя опции могут использоваться в пакетах управления, передаваемых со скоростью меньше скорости среды, они также важны в пакетах данных, а также для разделения и направления пересылки. Например в политике безопасности на основе входных портов, а также для завершения и реинкапсуляции пакетов в системах взаимодействия служб нужно помещать теги в пакеты данных. Поэтому, несмотря на желание применять расширяемость лишь к пакетам управления в целях упрощения, это не позволит выполнить требования к разработке.

2.2.1. Эффективная реализация

Часто возникает трудно разрешимый конфликт между гибкостью программ и производительностью оборудования. Для заданного набора функций обычно желательно обеспечить максимальную производительность. Однако это не означает отказа от функций, которые сегодня не могут быть реализованы с желаемой скоростью. Поэтому от протокола, для которого предполагается эффективная реализация, ожидается наличие набора общих возможностей, которые можно обрабатывать на разных платформах, а также элегантный механизм обработки более развитых функций в подходящих случаях.

Использование в протоколе заголовка и опций переменного размера зачастую вызывает вопросы о возможности эффективной аппаратной реализации протокола. Для ответа на этот вопрос в контексте Geneve важно сначала разделить оборудование на две категории - конечные точки туннелей и транзитные устройства. Конечные точки должны быть способны анализировать заголовки переменного размера, включая любые опции, и выполнять нужные действия. Поскольку эти устройства активно участвуют в работе протокола, Geneve наиболее сильно влияет на них. Однако конечные точки являются потребителями данных, поэтому передатчики могут приспособиться к возможностям приёмников.

Транзитные устройства могут быть способны интерпретировать опции, однако они не являются завершающими устройствами, т. е. не создают и не воспринимают пакеты Geneve. Следовательно, для них **недопустимо** менять заголовки Geneve, а также **недопустимо** вставлять или удалять опции, поскольку за это отвечают конечные точки туннелей. Имеющиеся в туннеле опции **должны** создаваться и удаляться только конечными точками туннеля. Участие транзитных устройств в интерпретации опций является **необязательным**.

Кроме того, конечная точка туннеля или транзитное устройство **может** использовать функции выгрузки в NIC, такие как выгрузка контрольных сумм, для повышения производительности обработки пакетов Geneve. Наличие в Geneve заголовков переменного размера не должно препятствовать использованию возможностей выгрузки в конечных точках и транзитных устройствах.

2.3. Использование стандартных систем IP

Протокол IP явно занял доминирующую позицию как транспортный механизм и было развито много методов, с течением времени сделавшим его отказоустойчивым, эффективным и недорогим. В результате стало естественным использование систем IP в качестве транзитных сетей для Geneve. К счастью, использования инкапсуляции и адресации IP достаточно для достижения основной цели доставки пакетов в нужные точки сети с помощью стандартной коммутации и маршрутизации.

Кроме того, почти все базовые системы разработаны для распараллеливания трафика с целью распределения нагрузки по множеству каналов без разупорядочения отдельных потоков. Эти методы ECMP обычно включают анализ и хэширование адресов и номеров портов из пакета для выбора исходящего канала. Однако использование туннелей зачастую снижает эффективность ECMP, поскольку без дополнительной информации о протоколе инкапсулированный трафик скрыт от базовой системы и для хэширования доступны лишь адреса конечных точек туннелей.

Поскольку для протокола Geneve важна работа с имеющимися системами, важно, чтобы энтропия инкапсулированных пакетов отражалась в туннельном заголовке. Наиболее подходит для этого номер порта отправителя UDP и этот вопрос рассмотрен в параграфе 3.3. Заголовок UDP.

3. Инкапсуляция Geneve

Пакет Geneve включает компактный заголовок туннеля, инкапсулированный в UDP для протокола IPv4 или IPv6. Небольшой фиксированный заголовок туннеля обеспечивает данные управления, а также базовую функциональность и возможность взаимодействия с упором на простоту. За этим заголовком следует набор опций переменного размера, позволяющих развивать протокол. Далее следуют данные (payload), включающие модуль данных протокола указанного типа, такой как кадр Ethernet. В параграфах 3.1 и 3.2 показан формат пакета Geneve доставляемого (как пример) через сеть Ethernet и содержащего в себе кадр Ethernet.

3.1. Формат пакета Geneve для IPv4

```
                     0                   1                   2                   3
                     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
Внешний заголовок Ethernet
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |                 Внешний MAC-адрес получателя                  |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    | Внешний MAC-адрес получателя  | Внешний MAC-адрес отправителя |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |                 Внешний MAC-адрес отправителя                 |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    | Необяз. Ethertype=C-Tag 802.1Q|      Внешний тег VLAN         |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |    Ethertype = 0x0800 IPv4    |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+

Внешний заголовок IPv4
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |Version|  IHL  |Type of Service|          Total Length         |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |         Identification        |Flags|      Fragment Offset    |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |  Time to Live |Protocol=17 UDP|          Header Checksum      |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |                  Внешний адрес отправителя IPv4               |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |                  Внешний адрес получателя IPv4                |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Внешний заголовок UDP
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |       Source Port = xxxx      |    Dest Port = 6081 Geneve    |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |           UDP Length          |        UDP Checksum           |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Заголовок Geneve
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |Ver|  Opt Len  |O|C|    Rsvd.  |          Protocol Type        |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |        Virtual Network Identifier (VNI)       |    Reserved   |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |                                                               |
                    ~                    Опции переменного размера                  ~
                    |                                                               |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Внутренний заголовок Ethernet (пример данных).
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |               Внутренний MAC-адрес получателя                 |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |Внутренний MAC-адрес получателя|Внутренний MAC-адрес отправит. |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |               Внутренний MAC-адрес отправителя                |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    | Необяз. Ethertype=C-Tag 802.1Q|    Внутренний тег VLAN        |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Данные (Payload)

                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |   Ethertype исходных данных   |                               |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               |
                    |                                 Исходные данные Ethernet      |
                    |                                                               |
                    ~ (Преамбула исходного кадра Ethernet, начиная с разграничителя ~
                    | кадра (SFD), и контрольная сумма (FCS) не показаны, а данные  |
                    | Ethernet не нужно выравнивать по 4-байтовой границе)          |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
Последовательность проверки кадра (FCS)
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
                    |       Новое значение (FCS) для внешнего кадра Ethernet        |
                    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Рисунок 2. Формат пакета Geneve для IPv4.

3.2. Формат пакета Geneve для IPv6

```
                     0                   1                   2                   3
                     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

Внешний заголовок Ethernet

Внешний заголовок IPv6
```

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Version| Traffic Class |           Flow Label                  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|         Payload Length        | NxtHdr=17 UDP |   Hop Limit   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                                                               |
+                  Внешний адрес отправителя IPv6               +
|                                                               |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
+                                                               +
|                                                               |
+                  Внешний адрес получателя IPv6                +
|                                                               |
+                                                               +
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Внешний заголовок UDP

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Source Port = xxxx      |    Dest Port = 6081 Geneve    |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|           UDP Length          |        UDP Checksum           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Заголовок Geneve

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Ver|  Opt Len  |O|C|    Rsvd.  |          Protocol Type        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        Virtual Network Identifier (VNI)       |    Reserved   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                  Опции переменного размера                    ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Внутренний заголовок Ethernet (пример данных).

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Внутренний MAC-адрес получателя                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|Внутренний MAC-адрес получателя|Внутренний MAC-адрес отправит. |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Внутренний MAC-адрес отправителя               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Необяз. Ethertype=C-Tag 802.1Q|    Внутренний тег VLAN
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Данные (Payload)

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Ethertype исходных данных   |                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+                               |
|                                 Исходные данные Ethernet      |
|                                                               |
~ (Преамбула исходного кадра Ethernet, начиная с разграничителя ~
| кадра (SFD), и контрольная сумма (FCS) не показаны, а данные  |
| Ethernet не нужно выравнивать по 4-байтовой границе)          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Последовательность проверки кадра (FCS)

```
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|       Новое значение (FCS) для внешнего кадра Ethernet        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Рисунок 3. Формат пакета Geneve для IPv6.

3.3. Заголовок UDP

Использование заголовка инкапсуляции UDP [RFC0768] следует семантике Ethernet и IP без организации соединений в дополнение к обеспечению маршрутизаторам энтропии для выполнения ECMP. Интерпретация полей описана ниже.

Source Port

Порт отправителя, выбранный исходной точкой туннеля. **Следует** использовать один номер порта для всех пакетов одного инкапсулированного потока для предотвращения нарушения порядка доставки пакетов при использовании разных путей. Для равномерного распределения потоков по разным каналам **следует** задавать порт отправителя с использованием хэш-значения заголовков инкапсулированных пакетов, например, традиционный квинтет (5-tuple[1]). Поскольку порт служит идентификатором потока, а не реальное соединение UDP, **можно** использовать весь 16-битовый диапазон для повышения энтропии. В дополнение к установке номера порта, для протокола IPv6 энтропию **можно** повысить с помощью метки потока. Пример использования меток потока IPv6 в туннелях приведён в [RFC6438].

Если трафик Geneve используется совместно с другими «слушателями» UDP на том же адресе IP, конечным точкам туннеля **следует** реализовать механизм, обеспечивающий возврат трафика ICMP при возникновении сетевых ошибок, корректному слушателю. Определение такого механизма выходит за рамки спецификации.

[1]Адреса и номера портов отправителя и получателя, а также номер протокола. *Прим. перев.*

Dest Port

Агентство IANA выделило номер 6081 как фиксированное общеизвестное значение для Geneve. Хотя это значение следует использовать по умолчанию, в реализациях **рекомендуется** поддерживать настройку номера порта. Выбранный порт служит для идентификации пакетов Geneve, **недопустимо** обращение номера на разных концах соединения, как это делается в TCP. Плоскость управления отвечает за настройку и поддержку конфигурации портов, а также их интерпретацию соответствующими устройствами. Определение плоскости управления выходит за рамки этого документа.

UDP Length

Размер пакета UDP с учётом заголовка UDP.

UDP Checksum

Для защиты заголовка Geneve, опций и данных от возможных повреждений **следует** рассчитывать контрольную сумму UDP в соответствии с [RFC0768] и [RFC1122] при инкапсуляции Geneve в IPv4. Для защиты заголовков IP и Geneve, опций и данных от возможного повреждения по умолчанию **должна** рассчитываться контрольная сумма UDP в соответствии с [RFC0768] и [RFC8200] при инкапсуляции Geneve в IPv6, за исключением некоторых ситуаций, описанных ниже. При получении пакетов с ненулевой контрольной суммой конечная точка туннеля **должна** проверить контрольную сумму. Если контрольная сумма не корректна, пакет **должен** отбрасываться, в остальных случаях пакет **должен** восприниматься для декапсуляции.

В некоторых случаях для контрольной суммы UDP **может** быть установлено значение 0 при передаче по протоколам IPv4 и IPv6 [RFC8200]. Дополнительные требования для использований контрольной суммы 0 приведены в параграфе 4.3. Вопрос отключения контрольных сумм UDP следует решать с учётом рабочего окружения и возможных рисков в результате повреждения пакетов.

3.4. Поля туннельного заголовка

Ver (2 бита)

текущая версия протокола имеет номер 0. Пакеты с неизвестным номером версии, полученные конечной точкой туннеля, **должны** отбрасываться. Транзитные устройства, интерпретирующие пакеты Geneve, **должны** считать пакеты с неизвестным номером версии пакетами UDP с неизвестными данными.

Opt Len (6 битов)

Размер поля опций в 4-байтовых словах без учёта 8 байтов фиксированного заголовка туннеля. В результате минимальный размер заголовка Geneve составляет 8 байтов, а максимальный - 260. Начало данных можно определить по смещению от конца заголовка Geneve.

Транзитные устройства **должны** обеспечивать соответствующую пересылку (включая выбор канала ECMP), независимо от поля Opt Len.

O (1 бит)

Флаг пакета управления (содержит управляющее сообщение). Управляющие сообщения передаются между конечными точками туннеля. Конечным точкам **недопустимо** пересылать данные (payload), а транзитным устройствам **недопустимо** интерпретировать их. Поскольку управляющие сообщения достаточно редки, конечным точкам туннеля **рекомендуется** направлять их в управляющую очередь с высоким приоритетом (например, для отправки пакета в процессор общего назначения из пересылающего контроллера ASIC[2] или в отдельный адаптер NIC для системы управления). Транзитным устройства **недопустимо** менять поведение пересылки на основе этого бита (например, для выбора канала ECMP).

C (1 бит)

Флаг наличия важной опции (установлен бит важности, см. параграф 3.5). При установленном флаге конечная точка туннеля **должна** проанализировать список опций для интерпретации важных опций. В конечных точках без поддержки анализа опций пакеты **должны** отбрасываться на основе состояния бита C в базовом заголовке. Если бит не установлен, конечная точка туннеля **может** вырезать все опции, используя значение Opt Len и переслать декапсулированный пакет. Транзитным устройствам **недопустимо** отбрасывать пакет на основе этого бита.

Rsvd. (6 битов)

Резервное поле, которое **должно** устанавливаться в 0 при передаче, а получатель **должен** игнорировать его.

Protocol Type (16 битов)

Тип модуля данных протокола, размещённого после заголовка Geneve, в соответствии со значениями Ethertype [ETYPES]. Для кадров Ethernet указывается тип 0x6558.

Virtual Network Identifier (VNI) (24 бита)

Идентификатор для уникального элемента виртуальной сети. Во многих случаях он может представлять сегмент L2, однако семантику пересылки декапсулированных пакетов определяет плоскость управления. Значение VNI **может** использовать при выборе решения о пересылке ECMP или **может** служить механизмом разделения перекрывающихся пространств адресов в инкапсулированных пакетах при распределении нагрузки между CPU.

Reserved (8 битов)

Резервное поле, которое **должно** устанавливаться в 0 при передаче, а получатель **должен** игнорировать его.

3.5. Опции туннеля

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Option Class         |      Type     |R|R|R| Length  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                                                               |
~                  Variable-Length Option Data                  ~
|                                                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Рисунок 4. Опция Geneve.

За базовым заголовком Geneve могут следовать опции в формате TLV, каждая и которых содержит 4-байтовый заголовок и переменное число данных, интерпретация которых определяется типом.

[2]Application-Specific Integrated Circuit - специализированная для приложения интегральная схема.

Option Class (16 битов)

Пространство имён для поля Type. Агентство IANA создало реестр Geneve Option Class для выделения идентификаторов организациям, технологиям и производителям, которые заинтересованы в создании типов опций. Каждая организация может независимо выделять типы, что позволяет проводить эксперименты и ускоренное внедрение. Предполагается, что со временем некоторые опции станут более распространёнными. Реализация может использовать типы опций из разных источников. Кроме того, агентство IANA зарезервировало определённые диапазоны для распределения по процедурам IETF Review и Experimental Use (см. раздел 7).

Type (8 битов)

Тип указывает формат данных, содержащихся в опции. Опции предназначены в основном для будущих расширений и инноваций, а стандартизованные формы опций будут определяться в отдельных документах.

Старший бит типа указывает важность (критичность) опции. Если приёмная сторона туннеля не распознает опцию с установленным старшим битом, пакет **должен** отбрасываться. Если в опции установлен этот бит, флаг C в базовом заголовке Geneve также **должен** быть установлен. Транзитным устройствам **недопустимо** отбрасывать пакеты на основе этого бита. На рисунке 5 показано расположение бита C в поле Type.

```
0 1 2 3 4 5 6 7 8
+-+-+-+-+-+-+-+-+
|C|    Type     |
+-+-+-+-+-+-+-+-+
```

Рисунок 5. Бит C в поле Type.

Требование отбрасывать пакет с неизвестной опцией и установленным битом применяется ко всей системе конечных точек, а не к отдельным компонентам реализации. Например, в системе, образованной пересылающим контроллером ASIC и CPU общего назначения, это означает отбрасывание пакета в ASIC. Реализация может передать пакет в CPU, используя канал управления с ограниченной скоростью для обработки исключительных ситуаций в slow-path.

R (3 бита)

Флаги управления опции, зарезервированные на будущее. При передаче поле **должно** заполняться нулями, а получатель **должен** игнорировать его.

Length (5 битов)

Размер опции в 4-байтовых словах с учётом заголовка (от 4 до 128 байтов). Поле Length = 0 говорит об отсутствии у опции данных. Пакеты, в которых общий размер опций не соответствует значению поля Opt Len в базовом заголовке, **должны** отбрасываться без уведомления, конечной точкой туннеля, обрабатывающей опции.

Variable-Length Option Data

Данные опции, интерпретируемые в соответствии с полем Type.

3.5.1. Обработка опций

Опции Geneve предназначены для создания и обработки конечными точками туннелей, однако их **могут** интерпретировать промежуточные устройства на пути туннеля. Транзитные устройства, не интерпретирующие заголовки Geneve (которые могут включать опции), **должны** обрабатывать пакеты Geneve как любые другие пакеты UDP и обеспечивать согласованную пересылку.

В конечных точках туннелей генерация и интерпретация опций определяется плоскостью управления, которая выходит за рамки документа. Однако для обеспечения взаимодействия между разнородными устройствами здесь вводятся некоторые требования к опциям и обрабатывающим их устройствам.

- Принимающая конечная точка туннеля **должна** отбрасывать пакеты с неизвестной опцией при установленном для типа опции флаге C. Транзитным устройствам **недопустимо** отбрасывать пакеты с неизвестными опциями и установленным битом C.
- Транзитным устройствам **недопустимо** менять содержимое и порядок опций.
- Если конечная точка туннеля получает пакет Geneve со значением поля Opt Len (общий размер опций), превышающим её возможности обработки опций, она **должна** отбросить пакет. Реализация может считать такие события исключительной ситуацией для плоскости управления. Плоскость управления отвечает за обеспечение возможности обработки обеими конечными точками одного туннеля всех опций (полного размера). Определение плоскости управления выходит за рамки этого документа.

При разработке опции Geneve важно предусмотреть её будущее развитие. Когда опция определена, разумно предположить, что реализации могут зависеть от её поведения. Поэтому область любых возможных изменений следует заранее тщательно описать.

С точки зрения архитектуры опциям следует быть информативными и независимыми. Это обеспечивает возможность их параллельной обработки и упрощает реализацию. Однако плоскость управления может вносить ограничения в порядок обработки, как описано в параграфе 4.5.1.

Могут возникать неожиданно серьёзные проблемы совместимости и взаимодействия в результате смены размера той или иной опции. Для конкретной опции задаётся фиксированный (константа) или переменный размер, который может меняться с течением времени или в зависимости от способа применения. Это свойство является частью определения опции и указывается её типом. Для опций фиксированного размера некоторые реализации могут игнорировать поле Length в заголовке опции и использовать при её анализе общеизвестное значение для этого типа. В таких случаях изменение размера опции будет влиять не только на её анализ, но и на разбор последующих опций. Поэтому для опций с фиксированным размером **недопустимо** изменение этого размера и вместо этого следует выделять для опции новый тип. Фактическое определение типа опции выходит за рамки этого документа. Тип опции и его интерпретацию следует задавать владельцу класса опции.

Опции могут обрабатываться оборудованием NIC с использованием выгрузки (offload, например, LSO и LRO), как описано в параграфе 4.6. Следует внимательно рассмотреть возможное влияние выгрузки на создаваемую опцию (см. параграф 4.6. Выгрузка в NIC).

4. Вопросы реализации и развёртывания.

4.1. Заявление о применимости

Geneve - это основанный на UDP протокол инкапсуляции наложенной виртуализации сетей, предназначенный для организации туннелей между NVE через имеющиеся сети IP. Протокол рассчитан на работу в общественных и частных ЦОД для развёртывания сетей множества арендаторов через имеющуюся базовую сеть IP.

Основанный на UDP протокол Geneve наследует рекомендации по применению UDP, содержащиеся в [RFC8085]. Применимость этих рекомендаций зависит от базовой сети IP и природы данных, передаваемых с использованием протокола Geneve (например, TCP/IP, IP/Ethernet).

Протокол Geneve предназначен для работы в среде ЦОД, управляемых одним или несколькими сотрудничающими сетевыми операторами, которая соответствует определению контролируемой среды [RFC8085]. Для работы сети в контролируемой среде могут поддерживаться определённые условия, что невозможно в глобальной среде Internet. Поэтому требования к туннельному протоколу, работающему в контролируемой среде, могут быть менее ограничительными по сравнению с требованиями для работы в Internet.

Для целей этого документа контролируемая среда с управляемым трафиком (traffic-managed controlled environment или TMCE) определяется как сеть IP, где применяется организация трафика (traffic engineered) или иные средства (например, ограничители трафика) предотвращения перегрузок. Концепция TMCE описана в [RFC8086]. Значительная часть текста параграфов 4.1 - 4.3 основана на [RFC8086], применимом к Geneve.

Оператор отвечает за соблюдение руководств и требований, указанных в этом параграфе как применимые для развёртывания Geneve.

4.2. Контроль перегрузок

Протокол Geneve не включает естественной поддержки контроля перегрузок и полагается на протокол трафика данных (payload) для контроля перегрузок в сети. Поэтому протокол Geneve **должен** применяться с контролем трафика или внутри TMCE для предотвращения перегрузок. Оператор TMCE может избавиться от перегрузок за счёт тщательной подготовки своих сетей, ограничения скорости пользовательского трафика данных и организации трафика в соответствии с возможностями путей.

4.3. Контрольная сумма UDP

Следует использовать контрольные суммы внешнего заголовка UDP при работе Geneve в сетях IPv4. Это обеспечивает целостность заголовков, опций и данных Geneve в случаях повреждения пакетов (например, предотвратит ошибочную доставку данных в системы других арендаторов). Контрольные суммы UDP предоставляют статистические гарантии сохранности данных в пути передачи. Такой контроль целостности не является строгим с точки зрения кодирования и криптографии и не предназначен для обнаружения ошибок на физическом уровне или преднамеренного изменения дейтаграмм (см. параграф 3.4 в [RFC8085]). В средах, где возникают такие риски, оператору **следует** применять дополнительные механизмы защиты целостности, такие как IPsec (6.2. Целостность данных).

Оператор **может** отказаться от контрольных сумм UDP и использовать нулевое значение контрольной суммы (zero UDP checksum), если целостность пакетов Geneve обеспечивается другими механизмами, такими как IPsec или дополнительные контрольные, а также при выполнении одного из условий (a, b, c) параграфа 4.3.1.

По умолчанию контрольные суммы UDP **должны** использоваться при работе Geneve по протоколу IPv6. Конечные точки туннелей **можно** настроить на работу с нулевыми контрольными суммами, если выполняются требования параграфа 4.3.1.

4.3.1. Обработка нулевой контрольной суммы UDP в IPv6

При работе Geneve по протоколу IPv6 контрольная сумма UDP служит для защиты заголовков IPv6, UDP и Geneve, опций и данных от повреждения. Поэтому протокол Geneve по умолчанию **должен** использовать контрольные суммы UDP при доставке по протоколу IPv6. Оператор **может** настроить использование нулевой контрольной суммой UDP при работе в TMCE, как указано в параграфе 4.1, если выполняется одно из приведённых ниже условий.

a) Известно, что вероятность повреждения пакета очень мала (например, из сведений о типах оборудования в базовой сети) и оператор готов принять риск незаметного повреждения пакета.

b) Измерения (например, статистика потоков трафика с нулевой контрольной суммой) показывают, что число повреждённых пакетов достаточно мало и оператор готов принять риск незаметного повреждения пакета.

c) Данные Geneve передают пакеты приложений, устойчивых к нарушению порядка и повреждению пакетов (возможно за счёт проверки контрольных сумм на вышележащем уровне или повторной передачи при ошибках).

Кроме того, реализация туннелей Geneve с нулевой контрольной суммой UDP **должна** выполнять ряд требований.

1. Использование нулевой контрольной суммы UDP для IPv6 **должно** быть задано по умолчанию для всех туннелей Geneve.
2. Если Geneve работает с нулевой контрольной суммой UDP для IPv6, реализация конечной точки туннеля **должна** выполнять все требования раздела 4 в [RFC6936] и требование 1 раздела 5 в [RFC6936], поскольку они относятся к Geneve.
3. Декапсулирующей конечной точке туннеля Geneve **следует** проверять действительность адресов IPv6 отправителя и получателя для туннеля Geneve, настроенного на приём пакетов с нулевой контрольной суммой UDP и отвергать пакеты при несовпадении адреса.
4. Инкапсулирующая конечная точка туннеля Geneve **может** использовать свой адрес отправителя IPv6 для каждого туннеля Geneve с нулевой контрольной суммой UDP для усиления проверки декапсулятором адреса

отправителя (т. е. один адрес отправителя IPv6 не применяется для разных получателей IPv6 независимо от того, является адрес получателя индивидуальным или групповым). Если это невозможно, **рекомендуется** использовать каждый адрес отправителя для небольшого числа туннелей Geneve с нулевой контрольной суммой UDP.

Следует отметить, что для требований 3 и 4 принимающая конечная точка туннеля может выполнить проверки лишь при наличии возможности независимо (out-of-band) узнать о соответствующем поведении отправителя. Одним из вариантов передачи этой информации является сигнализация плоскости управления. Определение плоскости управления выходит за рамки этого документа.

5. **Следует** принять меры по предотвращению выхода трафика Geneve, передаваемого по протоколу IPv6 с нулевой контрольной суммой UDP, в сеть Internet. Примеры таких механизмов включают пакетные фильтры на шлюзах и граничных устройствах сети Geneve, а также логическое или физическое разделение сети Geneve и сети, где передаётся трафик Internet.

Приведённые выше требования не меняют требований, заданных в [RFC8200] и [RFC6936].

Контроль адресов IPv6 отправителя и получателя, а также рекомендации по предотвращения повторного использования адресов отправителя IPv6 в туннелях Geneve обеспечивают некоторое смягчение перехода к отказу от контрольных сумм UDP для заголовков IPv6. Контролируемая среда с управления трафиком, удовлетворяющая хотя бы одному из приведённых в начале параграфа требований, обеспечивает дополнительные гарантии.

4.4. Инкапсуляция Geneve в IP

Как основанный на IP протокол инкапсуляции, Geneve поддерживает множество свойств и методов имеющихся протоколов. Применение некоторых из них ниже описано более подробно, однако в общем случае большинство концепций, применимых к уровню IP или туннелям IP, обычно работают и в контексте Geneve.

4.4.1. Фрагментация IP

Рекомендуется применять Path MTU Discovery ([RFC1191] и [RFC8201]) для предотвращения или минимизации фрагментирования пакетов. Использование Path MTU Discovery для транзитной сети предоставляет конечным точкам туннеля сведения о состоянии каналов, которые позволяют предотвратить или снизить фрагментацию пакетов в зависимости от роли конечной точки в виртуализованной сети. NVE может поддерживает это состояние (размер MTU для каналов туннеля, связанных с конечной точкой), поэтому при отправке арендатором пакетов, размер которых после инкапсуляции превышает MTU на канале туннеля, конечная точка может отбрасывать такие пакеты и передавать системам арендатора соответствующие сообщения. Если конечная точка туннеля связана с функциями маршрутизации или пересылки или может передавать сообщения ICMP, инкапсулирующая сторона туннеля **может** передавать сообщения ICMP о необходимости фрагментации [RFC0792] или Packet Too Big [RFC4443] системам арендатора. При определении размера MTU для туннельного канала **должен** приниматься в расчёт максимальный размер опций, поскольку он может различаться в пакетах. Рекомендации по обработке фрагментации в похожих службах наложенной инкапсуляции, таких как PWE3[1], приведены в параграфе 5.3 [RFC3985].

Некоторые реализации могут не поддерживать фрагментацию или иные менее распространённые свойства заголовков IP, такие как опции и заголовки расширения. Некоторые вопросы, связанные с размером MTU и фрагментацией в туннелях IP, а также с сообщениями ICMP, рассмотрены в параграфе 4.2 [INTAREA-TUNNELS].

4.4.2. DSCP, ECN, TTL

При инкапсуляции пакетов IP (в том числе через Ethernet) в Geneve возникают вопросы передачи кодов дифференцированного обслуживания (Differentiated Services Code Point или DSCP) и битов явной индикации перегрузки (Explicit Congestion Notification или ECN) из внутреннего заголовка в туннель при передаче и в обратном направлении при получении.

В [RFC2983] даны рекомендации по отображению DSCP между внутренним и внешним заголовком IP. Виртуализация сетей обычно более тесно связана с описанной моделью Pipe, где значение DSCP в туннельном заголовке устанавливается на основе правил (это может быть фиксированное значение, код, основанный на внутреннем классе трафика, или иной механизм группировки трафика). Аспекты модели Uniform (внутренние и внешние значения DSCP трактуются как одно одно поле с копированием на входе и выходе) также могут применяться, например путём перемаркировки во внутреннем заголовке на выходе туннеля в соответствии с транзитной маркировкой. Однако однородная модель концептуально не согласуется с виртуализацией сети, которая стремится обеспечить строгую изоляцию инкапсулированного трафика от физической сети.

В [RFC6040] описан механизм раскрытия возможностей ECN для туннелей IP и распространения маркеров перегрузки во внутренние пакеты. Это поведение **должно** обеспечиваться для пакетов IP, инкапсулированных в Geneve.

Хотя любая из моделей Uniform и Pipe подходит для обслуживания TTL (Hop Limit в IPv6) при туннелировании пакетов IP, модель Pipe лучше подходит для виртуализации сети. В [RFC2003] приведены рекомендации по обмену TTL между внутренним и внешним заголовком IP, эта модель похожа на Pipe и **рекомендуется** для использования с Geneve для приложений виртуализации сетей.

4.4.3. Групповая и широковещательная адресация

Туннели Geneve могут быть организованы между индивидуальными адресами конечных точек (point-to-point unicast) или использовать групповую или широковещательную адресацию. Например, в физической сети без поддержки групповых адресов инкапсуляция группового трафика может выполняться копированием в несколько unicast-туннелей или пересылкой индивидуальным получателям в соответствии с правилами (с возможной репликацией).

В физической сети с поддержкой групповых адресов может быть желательно использование этого свойства для достижения преимуществ аппаратной репликации инкапсулированных пакетов. В этом случае групповые адреса могут

[1]Pseudowire Emulation Edge-to-Edge - сквозная эмуляция псевдо-провода.

выделяться в физической сети по арендаторам, инкапсулируемым multicast-группам или иным критериям. Задание таких групп определяется плоскостью управления и выходит за рамки этого документа.

При использовании групповой адресации на физическом уровне в одной группе могут оказаться устройства с разными возможностями и некоторые опции могут интерпретироваться лишь частью устройств в группе. Другие устройства могут без опаски игнорировать опции, если в них не установлен флаг важности C, требования к обработке которого приведены в параграфе 3.4.

Кроме того, в [RFC8293] представлены примеры механизмов, которые можно применять для обработки группового трафика в наложенных сетях при виртуализации.

4.4.4. Односторонние туннели

Вообще говоря, туннели Geneve концептуально являются односторонними. Протокол IP не основан на соединениях и две конечных точки туннелей могут взаимодействовать между собой по двум односторонним каналам. Поскольку протокол Geneve основан на IP, туннельный уровень наследует эти черты.

Туннель может инкапсулировать протокол с явными соединениями, например, TCP, поддерживающий состояние на своём уровне. Кроме того, реализации **могут** моделировать туннели Geneve как двухсторонние соединения, например, для предоставления абстракции виртуального порта. В обоих случаях двухсторонняя направленность туннеля обеспечивается вышележащим уровнем и не влияет на работу самого протокола Geneve.

4.5. Ограничения для свойств протокола

Протокол Geneve предназначен для обеспечения гибкости при использовании с широким спектром имеющихся и будущих приложений. Это может вносить некоторые ограничения на использование метаданных или другие аспекты протокола в целях оптимизации для конкретного варианта применения. Например, некоторые приложения могут ограничивать типы поддерживаемых опций или максимальный размер опций. Другие приложения могут обрабатывать лишь некоторые типы инкапсулируемых данных, например, Ethernet или IP. Такая оптимизация может быть реализована глобально (во всей системе) или локально (например, для некоторого класса устройств или набора путей).

Конечным точкам туннеля эти ограничения могут быть переданы явно через плоскость управления или опосредованно через приложение. Поскольку Geneve является протоколом плоскости данных, т. е. не привязан к плоскости управления, задание таких механизмов выходит за рамки этого документа.

4.5.1. Ограничения для опций

Хотя опции Geneve достаточно гибки, плоскость управления может ограничивать число TLV в опциях, а также порядок и размер TLV, передаваемых между конечными точками туннеля, для упрощения обработки плоскости данных на программном или аппаратном уровне [NVO3-ENCAP]. Например, для некоторой важной информации, такой как защитный хэш, может потребоваться определённый порядок обработки для обеспечения наименьшей задержки или могут возникать требования к порядку обработки опций, обусловленные семантикой прикладного протокола.

Плоскость управления может согласовать набор TLV и некоторое упорядочение их, а также может ограничить общее число TLV в опциях пакета, например, с учётом возможностей оборудования. Поэтому плоскость управления должна быть способна описать поддерживаемый набор TLV и их порядок для конечных точек туннеля. При отсутствии плоскости управления для этого могут применяться иные механизмы, но этот вопрос выходит за рамки документа.

4.6. Выгрузка в NIC

Современные адаптеры NIC поддерживают выгрузку различных функций для повышения эффективности обработки пакетов. Для реализации многих типов выгрузки требуется лишь простота анализа инкапсулированного пакета (например, выгрузка контрольных сумм). Однако оптимизация LSO и LRO включает некоторую обработку самих опций, поскольку их нужно реплицировать или объединять для нескольких пакетов. В таких ситуациях желательно не требовать изменения логики выгрузки для обработки новых опций. Для решения этой задачи вводятся некоторые ограничения при определении опций, приведённые ниже.

- При выполнении LSO адаптер **должен** реплицировать весь заголовок Geneve вместе с опциями, включая неизвестные устройству, в каждый результирующий сегмент, если опция не допускает исключения. При выполнении LRO адаптер NIC может считать, что двоичного сравнения опций (включая неизвестные) достаточно для того, чтобы считать их одинаковыми, и **может** объединять пакеты с одинаковыми заголовками Geneve.
- **Недопустимо** менять порядок опций при обработке с выгрузкой, включая слияние пакетов для LRO.
- Адаптерам NIC, поддерживающим выгрузку, **недопустимо** отбрасывать пакеты с неизвестными опциями, включая важные, пока в конфигурации явно не задано иное.

Для реализаций Geneve не задаётся требования использовать выгрузку, включая описанные выше случаи. Однако выгрузка в настоящее время широко применяется в коммерческих NIC и приведённые выше правила предназначены для обеспечения эффективной обработки имеющихся и будущих опций широким классом устройств.

4.7. Обработка внутренних тегов VLAN

Geneve позволяет инкапсулировать множество протоколов, поэтому конкретная реализация может поддерживать лишь небольшую часть всего набора возможностей. Однако поддержка Ethernet предполагается достаточно распространённой, поэтому полезно описать поведение VLAN в инкапсулированных кадрах Ethernet.

Как и другие протоколы, поддержка внутренних заголовков VLAN является **необязательной**. Во многих случаях использование инкапсулированных VLAN может быть отключено из соображений безопасности или ограничений реализации. Однако в иных случаях транки VLAN через туннели Geneve могут быть полезны. Поэтому обработка внутренних тегов VLAN на входе и выходе туннеля определяется настройкой конечных точек туннеля и плоскостью управления, а спецификация явно не задаёт её в формате данных.

5. Вопросы перехода

С точки зрения плоскости данных протокол Geneve совместим с имеющимися сетями IP, поскольку большинство устройств воспринимает трафик как пакеты UDP. Однако наличие многочисленных протоколов туннелирования в средах с виртуализацией сетей поднимает практические вопросы сосуществования и перехода.

Поскольку Geneve работает на основе функций плоскости данных, обеспечиваемых наиболее распространёнными. протоколами, применяемыми для виртуализации сетей (VXLAN и NVGRE), достаточно просто приспособить имеющуюся плоскость управления для работы с Geneve. Поскольку прежний и новый формат пакетов поддерживают один набор возможностей, не требуется жёсткого перехода и конечные точки одного туннеля могут использовать любой общий протокол и эти протоколы могут различаться даже в рамках одной работающей системы. Поскольку транзитные системы в основном пересылают пакеты на основе заголовков IP, все протоколы для них представляются похожими и эти устройства не создают серьёзных проблем взаимодействия.

Для упрощения перехода реализациям настоятельно рекомендуется поддерживать работу одновременно по протоколу Geneve и имеющимся протоколам туннелирования, поскольку предполагается возможность взаимодействия одного узла с разнотипными узлами сети. В конце концов старые протоколы могут «исчезнуть» по причине ненадобности.

6. Вопросы безопасности

Поскольку пакеты Geneve инкапсулируются в UDP/IP, протокол не имеет встроенных механизмов защиты. В результате атакующий с доступом к базовой сети, доставляющей пакеты IP, может отслеживать, менять или внедрять пакеты. Взломанные конечные точки туннелей и промежуточные устройства также могут подменять идентификаторы в заголовках туннеля для получения доступа в сеть другого арендатора.

В конкретном домене безопасности, таком как ЦОД, управляемый одним оператором, наиболее распространенным и эффективным механизмом защиты является изоляция доверенных компонентов. Туннельный трафик может передаваться в отдельной сети VLAN и фильтроваться на границе доверия.

При прохождении через недоверенную сеть, такую как Internet, следует применять технологии VPN (например, IPsec [RFC4301]) для контроля подлинности и шифрования пакетов IP, сформированных как часть инкапсуляции (параграф 6.1.1).

В остальном Geneve не влияет на безопасность инкапсулированных пакетов. В соответствии с рекомендациями BCP 72 [RFC3552], в последующих параграфах описаны возможные риски, которые могут возникать в системах Geneve, и подходы к снижению таких рисков. Следует также отметить, что не все риски применимы к каждому варианту развёртывания Geneve и в некоторых случаях отдельные риски могут отсутствовать. Оператор должен оценить своё сетевое окружение, определить возможные риски и использовать подходящие методы их снижения.

6.1. Конфиденциальность данных

Протокол Geneve служит для инкапсуляции в среде виртуализации сетей и обеспечивает создание и поддержку туннелей между NVE через имеющиеся сети IP. Его можно использовать для развёртывания наложенной сети с множеством арендаторов на основе имеющейся базовой сети IP в общественном или частном ЦОД. Наложенные услуги обычно предоставляются сервис-провайдером, таким как поставщик облачных услуг или оператор частного ЦОД. Это может быть поставщик услуг базовой сети или другой провайдер. По причине наличия в такой среде множества арендаторов система арендатора может ожидать защиты конфиденциальности данных в пакетах, невозможности их подделки в пути (активная атака) и предотвращения несанкционированного отслеживания (пассивная атака) со стороны других арендаторов и базовой сети. Взломанный узел сети или промежуточное устройство в ЦОД может пассивно отслеживать пакеты Geneve между NVE или направлять трафик на дополнительный анализ. Арендатор может ожидать от провайдера наложенной сети защиту конфиденциальности данных как часть услуг или реализовать свои механизмы, такие как IPsec или TLS, для сквозной защиты данных между своими системами. Ожидается, что поставщик услуг наложенной сети обеспечит криптографическую защиту в случаях, когда услуги базовой сети предоставляет другой оператор, чтобы данные пакетов не были доступны в базовой сети.

Если оператор решает на основе анализа рисков, что нужна защита конфиденциальности (например, в среде с множеством операторов), **следует** применять сквозное шифрование данных арендатора между NVE, для чего можно применять проверенные и надёжные механизмы шифрования, такие как IPsec, DTLS и т. п.

6.1.1. Трафик между ЦОД

Система арендатора в его помещении (частный ЦОД) может соединяться с системами этого же арендатора в наложенной сети общественного облачного ЦОД или у арендатора могут быть системы в географически разнесённых ЦОД для повышения их доступности. Трафик данных Geneve между системами арендатора, передаваемый через разные сети, следует защищать от угроз при прохождении через сети общего пользования. Для всех наложенных данных Geneve, выходящих за пределы домена безопасности оператора, **следует** применять механизмы шифрования, такие как IPsec или иные технологии VPN, для защиты коммуникаций между NVE, которые могут проходить через недоверенные сетевые каналы. Спецификация механизмов защиты при обмене данными между разными ЦОД выходит за рамки этого документа.

Принципы, описанные в разделе 4 для контролируемых сред, применимы и при передаче данных между ЦОД.

6.2. Целостность данных

Инкапсуляция Geneve применяется между NVE для организации наложенных туннелей через имеющиеся базовые сети IP. В ЦОД с множеством арендаторов мошенническая или взломанная система может попытаться организовать пассивную (например, отслеживание трафика других арендаторов) или активную (например, вставка несанкционированного инкапсулированного трафика Geneve - обманные пакеты, повторное использование перехваченных пакетов и т. п.) атаку. Чтобы таких атак не возникало, NVE **недопустимо** распространять пакеты

Geneve за пределы NVE в системы арендатора и **следует** применять механизмы фильтрации пакетов для предотвращения пересылки несанкционированного трафика между системами арендатора в разных сетях. NVE **недопустимо** интерпретировать пакеты Geneve от систем оператора за исключением кадров для инкапсуляции.

Взломанный узел или промежуточное устройство в ЦОД может организовать активную атаку с попыткой вмешательства в обмен пакетами Geneve между NVE. Злонамеренное изменение полей заголовка Geneve может вызвать пересылку пакета в сеть другого арендатора. Если оператор знает о возможности таких атак в его сети, он может реализовать механизмы защиты целостности данных между NVE. Для предотвращения рисков **следует** применять механизм защиты целостности пакетов Geneve, включая их заголовки, опции и содержимое на пути между парой NVE. Криптографические механизмы защиты, такие как IPsec, позволяют обеспечить целостность данных. Оператор ЦОД может развернуть другие механизмы подходящие защиты, поддерживаемые в базовой сети, хотя механизмы без криптографии могут не защитить связанную с Geneve часть пакета от подделки.

6.3. Проверка подлинности партнёров NVE

Мошенническое сетевое устройство или взломанное устройство NVE в среде ЦОД может подделывать пакеты Geneve, представляясь легитимным NVE. Для снижения таких рисков оператору **следует** применять механизмы проверки подлинности, такие как IPsec, гарантирующие восприятие пакетов Geneve лишь от предусмотренных партнёров NVE, в среде, где оператор считает подмены или мошеннические устройства потенциальными угрозами. В определённых обстоятельствам могут применяться более простые средства проверки подлинности отправителя пакетов Geneve, например, фильтрация на входе по VLAN, MAC, IP, проверка обратного пути и т. п.

6.4. Интерпретация опций транзитными устройствами

Опции пакетов создаются и воспринимаются конечными точками туннелей. Как отмечено в параграфе 2.2.1, транзитные узлы тоже могут интерпретировать опции. Однако при сквозном шифровании пакетов (между входом и выходом туннеля), например, с помощью IPsec, транзитные устройства не смогут видеть заголовок и опции в пакете Geneve. В таких случаях транзитные устройства **должны** обрабатывать пакеты Geneve как любые другие пакеты IP и обеспечивать для них требуемую пересылку. При интерпретации опций промежуточным устройством оператор **должен** обеспечить доверенность транзитных узлов и то, что они не взломаны. Определение механизмов для этого выходит за рамки документа.

6.5. Групповой и широковещательный трафик

В типичных средах ЦОД, где групповая адресация IP не поддерживается базовой сетью, поддержку группового трафика можно организовать за счёт организации множества индивидуальных (unicast) туннелей. В этом случае применяются механизмы защиты, описанные выше для туннелей Geneve между партнёрами NVE. Если групповая адресация IP поддерживается базовой сетью и оператор решил применять её для группового трафика между конечными точками туннелей, он может воспользоваться механизмами защиты данных, такими как IPsec с групповыми расширениями [RFC5374], для защиты трафика в группах Geneve NVE.

6.6. Взаимодействия плоскости управления

Центр виртуализации (Network Virtualization Authority или NVA), описанный в [RFC8014], может служить в качестве плоскости управления для настройки и поддержки Geneve NVE. Предполагается, что оператор ЦОД использует механизмы защиты для коммуникаций между NVA и NVE, а также механизмы проверки подлинности для обнаружения обманных или взломанных NVE в своём административном домене. Рассмотрение таких механизмов выходит за рамки документа.

7. Взаимодействие с IANA

Агентство IANA выделило порт UDP 6081 в Service Name and Transport Protocol Port Number Registry [IANA-SN] как общеизвестный порт получателя для Geneve:

Имя службы: geneve
Транспортный протокол: UDP
Правообладатель: IESG <iesg@ietf.org>
Контакт: IETF Chair <chair@ietf.org>
Описание: Generic Network Virtualization Encapsulation (Geneve)
Документ: [RFC8926]
Номер порта: 6081

Кроме того агентство IANA создало субреестр Geneve Option Class для классов опций, размещённый под заголовком Network Virtualization Overlay (NVO3) в реестрах IANA для протоколов [IANA-PR]. Реестр Geneve Option Class содержит 16-битовые шестнадцатеричные значения со строками описания, правообладателями, контактными данными и ссылками на документы. Процедуры регистрации в соответствии с [RFC8126] показаны в таблице 1.

Таблица 1. Диапазоны классов опций Geneve.

Диапазон	***Процедура регистрации***
0x0000-0x00FF	IETF Review
0x0100-0xFEFF	First Come First Served
0xFF00-0xFFFF	Experimental Use

8. Литература

8.1. Нормативные документы

[RFC0768] Postel, J., "User Datagram Protocol", STD 6, RFC 768, DOI 10.17487/RFC0768, August 1980, <https://www.rfc-editor.org/info/rfc768>.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, DOI 10.17487/RFC0792, September 1981, <https://www.rfc-editor.org/info/rfc792>.

[RFC1122] Braden, R., Ed., "Requirements for Internet Hosts - Communication Layers", STD 3, RFC 1122, DOI 10.17487/RFC1122, October 1989, <https://www.rfc-editor.org/info/rfc1122>.

[RFC1191] Mogul, J. and S. Deering, "Path MTU discovery", RFC 1191, DOI 10.17487/RFC1191, November 1990, <https://www.rfc-editor.org/info/rfc1191>.

[RFC2003] Perkins, C., "IP Encapsulation within IP", RFC 2003, DOI 10.17487/RFC2003, October 1996, <https://www.rfc-editor.org/info/rfc2003>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4443] Conta, A., Deering, S., and M. Gupta, Ed., "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", STD 89, RFC 4443, DOI 10.17487/RFC4443, March 2006, <https://www.rfc-editor.org/info/rfc4443>.

[RFC6040] Briscoe, B., "Tunnelling of Explicit Congestion Notification", RFC 6040, DOI 10.17487/RFC6040, November 2010, <https://www.rfc-editor.org/info/rfc6040>.

[RFC6936] Fairhurst, G. and M. Westerlund, "Applicability Statement for the Use of IPv6 UDP Datagrams with Zero Checksums", RFC 6936, DOI 10.17487/RFC6936, April 2013, <https://www.rfc-editor.org/info/rfc6936>.

[RFC7365] Lasserre, M., Balus, F., Morin, T., Bitar, N., and Y. Rekhter, "Framework for Data Center (DC) Network Virtualization", RFC 7365, DOI 10.17487/RFC7365, October 2014, <https://www.rfc-editor.org/info/rfc7365>.

[RFC8085] Eggert, L., Fairhurst, G., and G. Shepherd, "UDP Usage Guidelines", BCP 145, RFC 8085, DOI 10.17487/RFC8085, March 2017, <https://www.rfc-editor.org/info/rfc8085>.

[RFC8126] Cotton, M., Leiba, B., and T. Narten, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 8126, DOI 10.17487/RFC8126, June 2017, <https://www.rfc-editor.org/info/rfc8126>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8200] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", STD 86, RFC 8200, DOI 10.17487/RFC8200, July 2017, <https://www.rfc-editor.org/info/rfc8200>.

[RFC8201] McCann, J., Deering, S., Mogul, J., and R. Hinden, Ed., "Path MTU Discovery for IP version 6", STD 87, RFC 8201, DOI 10.17487/RFC8201, July 2017, <https://www.rfc-editor.org/info/rfc8201>.

8.2. Дополнительная литература

[ETYPES] IANA, "IEEE 802 Numbers", <https://www.iana.org/assignments/ieee-802-numbers>.

[IANA-PR] IANA, "Protocol Registries", <https://www.iana.org/protocols>.

[IANA-SN] IANA, "Service Name and Transport Protocol Port Number Registry", <https://www.iana.org/assignments/service-names-port-numbers>.

[IEEE.802.1Q_2018] IEEE, "IEEE Standard for Local and Metropolitan Area Networks--Bridges and Bridged Networks", DOI 10.1109/IEEESTD.2018.8403927, IEEE 802.1Q-2018, July 2018, <http://ieeexplore.ieee.org/servlet/opac?punumber=8403925>.

[INTAREA-TUNNELS] Touch, J. and M. Townsley, "IP Tunnels in the Internet Architecture", Work in Progress, Internet-Draft, draft-ietf-intarea-tunnels-10, 12 September 2019, <https://tools.ietf.org/html/draft-ietf-intarea-tunnels-10>.

[NVO3-DATAPLANE] Bitar, N., Lasserre, M., Balus, F., Morin, T., Jin, L., and B. Khasnabish, "NVO3 Data Plane Requirements", Work in Progress, Internet-Draft, draft-ietf-nvo3-dataplane-requirements-03, 15 April 2014, <https://tools.ietf.org/html/draft-ietf-nvo3-dataplane-requirements-03>.

[NVO3-ENCAP] Boutros, S., "NVO3 Encapsulation Considerations", Work in Progress, Internet-Draft, draft-ietf-nvo3-encap-05, 17 February 2020, <https://tools.ietf.org/html/draft-ietf-nvo3-encap-05>.

[RFC2983] Black, D., "Differentiated Services and Tunnels", RFC 2983, DOI 10.17487/RFC2983, October 2000, <https://www.rfc-editor.org/info/rfc2983>.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, DOI 10.17487/RFC3031, January 2001, <https://www.rfc-editor.org/info/rfc3031>.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, DOI 10.17487/RFC3552, July 2003, <https://www.rfc-editor.org/info/rfc3552>.

[RFC3985] Bryant, S., Ed. and P. Pate, Ed., "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, DOI 10.17487/RFC3985, March 2005, <https://www.rfc-editor.org/info/rfc3985>.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <https://www.rfc-editor.org/info/rfc4301>.

[RFC5374] Weis, B., Gross, G., and D. Ignjatic, "Multicast Extensions to the Security Architecture for the Internet Protocol", RFC 5374, DOI 10.17487/RFC5374, November 2008, <https://www.rfc-editor.org/info/rfc5374>.

[RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, DOI 10.17487/RFC6438, November 2011, <https://www.rfc-editor.org/info/rfc6438>.

[RFC7348] Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, DOI 10.17487/RFC7348, August 2014, <https://www.rfc-editor.org/info/rfc7348>.

[RFC7637] Garg, P., Ed. and Y. Wang, Ed., "NVGRE: Network Virtualization Using Generic Routing Encapsulation", RFC 7637, DOI 10.17487/RFC7637, September 2015, <https://www.rfc-editor.org/info/rfc7637>.

[RFC8014] Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)", RFC 8014, DOI 10.17487/RFC8014, December 2016, <https://www.rfc-editor.org/info/rfc8014>.

[RFC8086] Yong, L., Ed., Crabbe, E., Xu, X., and T. Herbert, "GRE-in-UDP Encapsulation", RFC 8086, DOI 10.17487/RFC8086, March 2017, <https://www.rfc-editor.org/info/rfc8086>.

[RFC8293] Ghanwani, A., Dunbar, L., McBride, M., Bannai, V., and R. Krishnan, "A Framework for Multicast in Network Virtualization over Layer 3", RFC 8293, DOI 10.17487/RFC8293, January 2018, <https://www.rfc-editor.org/info/rfc8293>.

[VL2] "VL2: A Scalable and Flexible Data Center Network", ACM SIGCOMM Computer Communication Review, DOI 10.1145/1594977.1592576, August 2009, <https://dl.acm.org/doi/10.1145/1594977.1592576>.

Благодарности

Авторы признательны Puneet Agarwal, David Black, Sami Boutros, Scott Bradner, Martín Casado, Alissa Cooper, Roman Danyliw, Bruce Davie, Anoop Ghanwani, Benjamin Kaduk, Suresh Krishnan, Mirja Kühlewind, Barry Leiba, Daniel Migault, Greg Mirksy, Tal Mizrahi, Kathleen Moriarty, Magnus Nyström, Adam Roach, Sabrina Tanamal, Dave Thaler, Éric Vyncke, Magnus Westerlund и многим другим членам рабочей группы NVO3 за их рецензии, комментарии и предложения.

Авторы благодарят Sam Aldrin, Alia Atlas, Matthew Bocci, Benson Schliesser и Martin Vigoureux за руководство процессом.

Участники работы

Ниже перечислены люди, которые были авторами ранней версии документа и внесли в него существенный вклад.

Pankaj Garg
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
United States of America
Email: pankajg@microsoft.com

Chris Wright
Red Hat Inc.
1801 Varsity Drive
Raleigh, NC 27606
United States of America
Email: chrisw@redhat.com

Kenneth Duda
Arista Networks
5453 Great America Parkway
Santa Clara, CA 95054
United States of America
Email: kduda@arista.com

Dinesh G. Dutt
Independent
Email: didutt@gmail.com

Jon Hudson
Independent
Email: jon.hudson@gmail.com

Ariel Hendel
Facebook, Inc.
1 Hacker Way
Menlo Park, CA 94025
United States of America
Email: ahendel@fb.com

Адреса авторов

Jesse Gross (редактор)
Email: jesse@kernel.org

Ilango Ganga (редактор)
Intel Corporation
2200 Mission College Blvd.
Santa Clara, CA 95054
United States of America
Email: ilango.s.ganga@intel.com

T. Sridhar (редактор)
VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
United States of America
Email: tsridhar@utexas.edu

Перевод на русский язык

Николай Малых

nmalykh@protokols.ru