Internet Engineering Task Force (IETF)
Request for Comments: 8469
Updates: 4448
Category: Standards Track
ISSN: 2070-1721

S. Bryant
A. Malis
Huawei
I. Bagdonas
Equinix

November 2018

Recommendation to Use the Ethernet Control Word

Рекомендации по использованию управляющего слова Ethernet

Аннотация

Для псевдопроводной (PW[1]) инкапсуляции Ethernet в RFC 4448 указано, что использование управляющего слова (CW[2]) не обязательно. В отсутствие CW пакет Ethernet PW может быть ошибочно принят маршрутизатором LSR[3] за пакет IP. Это может привести к ошибочному выбору пути для пакета среди ECMP[4] а в результате может нарушиться порядок пакетов. Эта проблема стала более серьёзной в результате развёртывания оборудования с адресами Ethernet MAC[5], начинающимися с 0x4 или 0x6. Использование Ethernet PW CW решает эту проблему. Данный документ **рекомендует** использовать Ethernet PW CW при любых необычных обстоятельствах.

Этот документ обновляет RFC 4448.

Статус документа

Документ относится к категории Internet Standards Track.

Документ является результатом работы IETF[6] и представляет согласованный взгляд сообщества IETF. Документ прошёл открытое обсуждение и был одобрен для публикации IESG[7]. 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/rfc8469.

Оглавление

1. Введение

Спецификация псевдопроводной инкапсуляции Ethernet [RFC4448] указывает, что использование управляющего слова CW не обязательно. Маршрутизаторы LSR обычно выполняют поиск после стека меток для определения присутствия пакета IP в поле данных и, при его обнаружении, выбора следующего интервала (next hop) на основании «пятёрки» (five-tuple) - IP-адреса отправителя и получателя, protocol/next-header, номера транспортных портов отправителя и

[1]Pseudowire.
[2]Control word.
[3]Label switching router - маршрутизатор с коммутацией по меткам.
[4]Equal-cost multipath - множество равноценных путей.
[5]Media Access Control - управление доступом к среде.
[6]Internet Engineering Task Force - комиссия по решению инженерных задач Internet.
[7]Internet Engineering Steering Group - комиссия по инженерным разработкам Internet.

получателя. В отсутствие PW CW пакет Ethernet PW может быть ошибочно принят за пакетом IP маршрутизатором LSR, выбирающим путь ECMP на основе five-tuple. Это может приводить к выбору ошибочного пути ECMP и, в результате, к нарушению порядка доставки пакетов. Дополнительное рассмотрение этой проблемы дано в [RFC4928].

Нарушение порядка в потоке может возникать и при наличии одного пути, если используется классификация трафика и механизмы дифференцированной пересылки. Эти ошибки возникают в результате того, что устройство пересылки некорректно относит пакет к протоколу IP и применяет правила пересылки на основе полей в данных PW (payload).

Пакеты IPv4 и IPv6 начинаются со значений 0x4 и 0x6, соответственно. Может происходить ошибочная идентификация пакета, если пакет Ethernet PW без CW содержит кадр Ethernet с адресом получателя, начинающимся с указанных значений.

По ряду причин эта проблема недавно стала серьёзной. Во-первых, Регистрационный комитет IEEE (RAC[1]) выделил адреса Ethernet MAC, начинающиеся с 0x4 и 0x6, а оборудование с такими MAC-адресами появилось в сетях. Во-вторых, озабоченность вопросами приватности привела к использованию случайных MAC-адресов, назначаемых локально. При случайном назначении адреса, начинающиеся с указанных значений будут составлять приблизительно 1/8 часть всех выделяемых адресов.

Использование Ethernet PW CW решает эту проблему.

Данный документ **рекомендует** использовать Ethernet PW CW при любых необычных обстоятельствах.

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 должны интерпретироваться в соответствии с BCP 14 [RFC2119] [RFC8174] тогда и только тогда, когда они набраны заглавными буквами, как показано здесь.

3. Обоснование

Инкапсуляция Ethernet PW определена в [RFC4448]. Особое значение имеет параграф 4.6, часть которого для удобства читателя приведена ниже. Отметим, что RFC 4448 цитирует [PWE3-CW] для ссылки на [RFC4385] и [VCCV] для ссылки на документ, который в конечном итоге опубликован как [RFC5085].

> Управляющее слово, определённое в этом параграфе, основано на Generic PW MPLS Control Word из [PWE3-CW]. Оно обеспечивает возможность упорядочить отдельные кадры в PW, а также избежать распределения пакетов между равноценными путями (ECMP) [RFC2992] и применения механизмов OAM[2], включая VCCV [VCCV].
>
> В [PWE3-CW] сказано «Если PW реагирует на нарушение порядка пакетов и передаётся через MPLS PSN с использованием содержимого данных MPLS (payload) для выбора пути ECMP, псевдопровод может реализовать механизм предотвращения нарушений порядка пакетов». Это требуется для того, чтобы реализации ECMP могли проверить первый полубайт после стека меток MPLS для определения принадлежности пакета к протоколу IP. Если MAC-адрес отправителя в кадре Ethernet, передаваемом через PW без управляющего слова, начинается с 0x4 или 0x6, он будет ошибочно сочтён пакетом IPv4 или IPv6. В зависимости от конфигурации и топологии сети MPLS это может приводить к ситуации, где пакеты данного PW будут передаваться по разным путям. В результате может возрасти число кадров с нарушением порядка доставки в данном PW или пакеты OAM пойдут по пути, отличающемуся от пути обычного трафика (см. параграф 4.4.3 Порядок кадров).
>
> Функции, предоставляемые управляющим словом, могут не требоваться для данного Ethernet PW. Например, ECMP может не быть или не использоваться в данной сети MPLS, соблюдение порядка кадров может не требоваться и пр. В таких случаях роль слова управления невелика и оно может не использоваться. Ранние реализации Ethernet PW развёртывались без CW и возможности обрабатывать слово управления при его наличии. Для совместимости со старыми версиями будущие реализации **должны** быть способны передавать и принимать кадры без CW.

В начале развёртывания PW часть коммерчески значимого оборудования была не способна обрабатывать Ethernet CW. Кроме того, в те времена предполагалось, что адреса Ethernet MAC, начинающиеся с 0x4 или 0x6, не будут назначаться IEEE RAC и можно реализовать Ethernet PW без поддержки CW.

С течением времени адреса Ethernet MAC, начинающиеся с 0x4 и 0x6, были выделены RAC. Поэтому допущение о том, что в реальных сетях не будет возникать путаницы между пакетами Ethernet PW без CW и пакетами IP, перестало соответствовать реалиям.

Возможно несанкционированное использование адресов Ethernet MAC послужило тому, что некоторые производители оборудования реализовали более сложные, фирменные методы, позволяющие различать пакеты Ethernet PW и IP. Такие механизмы основаны на эвристике проверки транзитных пакетов с целью точного определения типа пакета и не могут считаться надёжными из-за произвольной природы данных, передаваемых в таких пакетах.

Проблема была обозначения в почтовой конференции NANOG, доступной по ссылке https://mailman.nanog.org/pipermail/nanog/2016-December/089395.html.

4. Рекомендации

Неоднозначность идентификации в данных MPLS пакетов Ethernet PW и IP устраняется при использовании Ethernet PW CW. Этот документ обновляет [RFC4448] в том смысле, что входным и выходным граничным устройствам провайдера (PE[3]) **следует** поддерживать Ethernet PW CW и при наличии этой поддержки CW **должно** применяться.

[1]Registration Authority Committee.
[2]Operations, Administration, and Maintenance - операции, администрирование и поддержка.
[3]Provider edge.

Там, где требуется применение ECMP для трафика Ethernet PW, а входные и выходные устройства PE поддерживают ELI/EL[4] [RFC6790] и FAT PW[5] [RFC6391], может применяться любой метод. Использование обоих методов на одном PW обычно не требуется и его следует избегать, если позволяют обстоятельства. В случае многосегментных PW при использовании ELI/EL его **следует** применять на каждом сегменте PW. Метод обеспечения использования ELI/EL на каждом сегменте выходит за рамки этого документа.

5. Равноценные пути (ECMP)

Там, где объем трафика Ethernet PW требует применения ECMP, может использоваться один из двух методов:

- FAT PW через сеть MPLS PSN[3], как описано в [RFC6391];
- метки энтропии LSP[4], как описано в [RFC6790].

RFC 6391 работает на основе повышения энтропии нижней метки стека. Поддержка этой функции требуется на входных и выходных PE. Также требуется, чтобы достаточное число LSR на пути LSP между входным и выходным PE было способно выбирать путь ECMP для пакета MPLS с достаточной глубиной стека.

RFC 6790 работает на основе включения энтропии в путь LSP-часть стека меток. Это требует от входного и выходного PE поддержки вставки и удаления EL и ELI, а также достаточного числа LSR на пути LSP, способных выбирать путь ECMP на основе EL.

В обоих случаях требуется обеспечить прохождение пакетов OAM и пакетов данных по одному пути. Этот вопрос подробно рассмотрен в разделе 7 [RFC6391] и разделе 6 [RFC6790]. Однако в обоих случаях ситуация улучшается по сравнению с поведением ECMP без использования Ethernet PW CW, когда нет возможности обеспечить прохождение пакетов PW OAM по одному пути с пакетами данных PW, для которых ECMP выбирается на основе five-tuple данных IP.

Метка PW вталкивается перед меткой LSP. Поскольку метки ELI/EL являются частью уровня LSP, а не уровня PW, они вталкиваются после метки PW.

6. Пути смягчения проблемы

Когда нет возможности использовать Ethernet PW CW, влияние ECMP может быть предотвращено путём передачи PW по пути с организацией трафика, который не использует поле данных (payload) для распределения нагрузки (например, RSVP-TE [RFC3209]). Однако на таких путях может применяться распределение нагрузки через связки каналов (link-bundle) и, естественно, весь трафик PW должен передаваться через один LSP.

7. Эксплуатационные вопросы

В некоторых случаях включение CW в псевдопровод PW определяется конфигурацией оборудования. Кроме того, в таких случаях по умолчанию возможен запрет использования CW. Следует принять меры, обеспечивающие независимость программ, реализующих данную спецификацию, от настроек, которые препятствуют использованию CW. Программам рекомендуется с ограниченной частотой передавать сообщения, указывающие возможность использования CW и наличие запрета такого использования в имеющейся конфигурации.

Вместо указания типа данных в пакетах MPLS использует уровень управления для сигнализации о типе данных, который следуют за нижней меткой стека. Некоторые LSR пытаются определить типа пакета путём проверки данных MPLS и в ряде случаев просматривают данные дальше PW CW. Если данные представляются пакетом IP или IP указан в заголовке Ethernet, они выполняют расчёт ECMP на основе данных, сочтённых полями five-tuple. Однако такое определение типа данных не даёт точного результата и при ошибочной идентификации пакета как IP может нарушаться порядок доставки пакетов. Нарушение порядка в этом случае оператору трудно обнаружить. При включении функции, позволяющей использовать информацию из пакета, расположенную после PW CW, при расчёте ECMP, оператору следует принимать во внимание возможность нарушения порядка доставки кадров Ethernet, несмотря на присутствие CW.

8. Вопросы безопасности

В этом документе отдано предпочтение одной широко распространённой инкапсуляции Ethernet PW над другой. С этим методом связаны вопросы безопасности, рассмотренные в [RFC4448]. Документ не создаёт других проблем безопасности.

9. Взаимодействие с IANA

Этот документ не запрашивает действий IANA.

10. Литература

10.1. Нормативные документы

[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>.

[RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, DOI 10.17487/RFC4385, February 2006, <https://www.rfc-editor.org/info/rfc4385>.

[RFC4448] Martini, L., Ed., Rosen, E., El-Aawar, N., and G. Heron, "Encapsulation Methods for Transport of Ethernet over MPLS Networks", RFC 4448, DOI 10.17487/RFC4448, April 2006, <https://www.rfc-editor.org/info/rfc4448>.

[4]Entropy Label Indicator/Entropy Label - индикатор метки энтропии/метка энтропии.
[5]Flow-Aware Transport of Pseudowire - осведомлённый о потоках транспорт псевдопровода.
[3]Packet Switched Network - сеть с коммутацией пакетов.
[4]Label Switched Path - путь с коммутацией по меткам.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, DOI 10.17487/RFC4928, June 2007, <https://www.rfc-editor.org/info/rfc4928>.

[RFC6391] Bryant, S., Ed., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, DOI 10.17487/RFC6391, November 2011, <https://www.rfc-editor.org/info/rfc6391>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <https://www.rfc-editor.org/info/rfc6790>.

[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>.

10.2. Дополнительная литература

[RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, DOI 10.17487/RFC2992, November 2000, <https://www.rfc-editor.org/info/rfc2992>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <https://www.rfc-editor.org/info/rfc3209>.

[RFC5085] Nadeau, T., Ed. and C. Pignataro, Ed., "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, DOI 10.17487/RFC5085, December 2007, <https://www.rfc-editor.org/info/rfc5085>.

Благодарности

Авторы благодарят Job Snijders за привлечение внимания к проблеме. Спасибо Pat Thaler за разъяснение вопроса о локальном назначении MAC-адресов и Sasha Vainshtein за ценные разъяснения и замечания.

Адреса авторов

Stewart Bryant

Huawei

Email: stewart.bryant@gmail.com

Andrew G. Malis

Huawei

Email: agmalis@gmail.com

Ignas Bagdonas

Equinix

Email: ibagdona.ietf@gmail.com>

Перевод на русский язык

Николай Малых

nmalykh@protokols.ru