Internet Engineering Task Force (IETF)
Request for Comments: 8238
Category: Informational
ISSN: 2070-1721

L. Avramov
Google
J. Rapp
VMware

August 2017

Терминология для оценки работы ЦОД

Data Center Benchmarking Terminology

Аннотация

Целью этого информационного документа является определение и описание методов измерений для оценки работы центров обработки данных (ЦОД), а также терминологии, применяемой при оценке производительности сетевого оборудования ЦОД. В документе обоснованы важные концепции оценки коммутаторов и маршрутизаторов в ЦОД, которые служат основой для документа по методологии тестирования (RFC 8239). Многие из этих терминов и методов могут применяться к сетевому оборудованию, не относящемуся к ЦОД, поскольку разработанные для таких центров технологии могут применяться в других местах.

Статус документа

Документ не является спецификацией стандарта (Internet Standards Track) и публикуется с информационными целями.

Документ является результатом работы IETF[1] и представляет согласованный взгляд сообщества IETF. Документ прошёл открытое обсуждение и был одобрен для публикации IESG[2]. Не все одобренные IESG документы претендуют на статус Internet Standard (см. раздел 2 в RFC 7841).

Информацию о текущем статусе документа, ошибках и способах обратной связи можно найти по ссылке http://www.rfc-editor.org/info/rfc8238.

Оглавление

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

1. Введение

Картина трафика в ЦОД неоднородна и постоянно меняется. Это обусловлено характером и разнообразием применяемых в ЦОД приложений. Это могут быть большие потоки «запад-восток» («горизонтальные» потоки между серверами одного ЦОД) в одном центре и большие потоки «север-юг» («вертикальные» потоки из внешнего источника к серверу ЦОД) в другом, а также разные комбинации направлений потоков. Картины трафика по своей природе содержат пики (всплески) и включают потоки «многие к одному» «многие ко многим», «один ко многим». Потоки могут быть небольшими и чувствительными к задержками или большими и чувствительными к пропускной способности, а также включать смесь трафика UDP и TCP. Все перечисленное может существовать в одном кластере и поток может проходить через одно сетевое устройство. Тесты производительности сетевых устройств используются достаточно давно и описаны в [RFC1242], [RFC2432], [RFC2544], [RFC2889] и [RFC3918]. Эти тесты в основном привязаны к параметрам задержки и максимальной пропускной способности тестируемого устройства (DUT[1]). Эти стандарты хороши для измерения теоретической максимальной пропускной способности, скорости пересылки и задержки в условиях теста, но не соответствуют реальным картинам трафика, который может проходить через сетевые устройства. Сетевые устройства ЦОД включают маршрутизаторы и коммутаторы.

Ниже перечислены основные характеристики типичных сетевых устройств.

- Высокая плотность портов(не менее 48).
- Высокая скорость (вплоть до 100 Гбит/с на порт).
- Высокая пропускная способность (суммарная линейная скорость всех портов для уровня 2 и/или 3).
- Малые задержки (микросекунды или наносекунды).
- Незначительный объем буферов (мегабайты в объёме всего устройства).
- Пересылка на уровнях 2 и 3 (уровень 3 не обязателен).

В этом документе приведены определения, метрические параметры и новая терминология, включая случаи перегрузки и анализа буферов в коммутаторах, а также заново определены некоторые базовые компоненты с учётом широкого спектра картин трафика. Методология тестирования определена в [RFC8239].

1.1. 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.

1.2. Формат определений

- Определяемый термин (например, задержка).
- Определение - конкретное определение термина
- Обсуждение - краткое обсуждение термина, его использования и ограничений на измерительные процедуры.
- Единицы измерения - методология измерения и единицы, используемые для значения, если это применимо.

2. Задержка

2.1. Определение

Задержкой называется интервал времени, затрачиваемый кадром на прохождение через DUT. Задержка измеряется в единицах времени (секундах, миллисекундах, микросекундах и т. д.). Измерение задержки нужно для того, чтобы оценить эффект добавления устройства в коммуникационный путь.

Интервал задержки можно оценивать между разными комбинациями событий, независимо от типа коммуникационного устройства (побитовая или сквозная пересылка - cut-through или с промежуточной буферизацией - store-and-forward). В [RFC1242] задержка определяется по-разному для каждого типа устройств.

Ниже приведены традиционно используемые определения задержки для разных устройств.

- FILO (First In Last Out - первый вошёл - последний вышел)

  Временной интервал начинается при поступлении во входной порт конца первого бита входящего кадра и заканчивается в тот момент, когда последний бит выходного кадра становится виден на выходном порту.

[1]Device Under Test.

- FIFO (First In First Out - первый вошёл - первый вышел)

  Временной интервал начинается при поступлении во входной порт конца первого бита входящего кадра и заканчивается в тот момент, когда начало первого бита выходного кадра становится видно на выходном порту. Этот вариант применяется для определённой в [RFC1242]) задержки устройств с побитовой пересылкой.

- LILO (Last In Last Out - последний вошёл - последний вышел)

  Временной интервал начинается при поступлении во входной порт последнего бита входящего кадра и заканчивается в тот момент, когда последний бит выходного кадра становится виден на выходном порту.

- LIFO (Last In First Out - последний вошёл - первый вышел)

  Временной интервал начинается при поступлении во входной порт последнего бита входящего кадра и заканчивается в тот момент, когда первый бит выходного кадра становится виден на выходном порту. Этот вариант применяется для определённой в [RFC1242]) задержки устройств с промежуточной буферизацией.

Другим способом обозначения четырёх перечисленных выше вариантов является указание битовых позиций в направлении со входа на выход.

- FILO это FL (первый бит - последний бит).
- FIFO это FF (первый бит - первый бит).
- LILO это LL (последний бит - последний бит).
- LIFO это LF (последний бит - первый бит).

Это определение в контексте оценки производительности коммутаторов ЦОД используется взамен определения «задержки», приведённого в параграфе 3.8 RFC 1242 и процитированного ниже.

> Для устройств с промежуточной буферизацией (store and forward) задержкой считается временной интервал, начинающийся в момент поступления во входной порт последнего бита входящего кадра и заканчивающийся в тот момент, когда на выходном порту становится видимым первый бит исходящего кадра.
>
> Для устройств с побитовой пересылкой (bit forwarding) задержкой считается временной интервал, начинающийся в тот момент, когда во входной порт попадает конец первого бита входящего кадра, и заканчивающийся в тот момент, когда в выходном порту становится виден первый бит исходящего кадра.

Для обеспечения соответствия обоим типам сетевых устройств и двум вновь возникшим гибридным типам измерение задержки в коммутаторах в соответствии с данным документом **должно** основываться на событиях FILO. Этот вариант будет включать задержку в коммутаторе, а также задержку на преобразование кадра в последовательную форму (serialization delay). Это представляет «полную» задержку при прохождении через DUT. Для чувствительных к задержке приложений, которые могут работать, начиная с первых битов кадра, **можно** использовать события FIFO (для определение RFC 1242 для задержки в устройствах с промежуточной буферизацией). В любом случае комбинация событий, используемая для определения задержки **должна** указываться в отчёте.

2.2. Обсуждение

Как было отмечено в параграфе 2.1, FILO является наиболее значимым определением для процесса измерения.

Не все устройства DUT относятся к «чистым» типам cut-through или store-and-forward. В ЦОД используются DUT, которые часто используют промежуточную буферизацию для мелких пакетов и сквозную пересылку пакетов крупнее некого заданного размера. Размер пакета, при котором поведение изменяется, **может** быть настраиваемым (это зависит от производителя DUT). Определение FILO подходит как для сквозной коммутации, так и для случая промежуточной буферизации. Порог смены типа поведения не оказывает влияния на оценку работы, поскольку FILO подходит для обоих вариантов.

Механизм LIFO подходит для коммутаторов store-and-forward, но не работает при сквозной коммутации, поскольку в этом случае он будет показывать отрицательную задержку для больших пакетов за счёт того, что не учитывается преобразование в последовательный формат (serialization delay). Следовательно, этот механизм **недопустимо** использовать при сравнении задержки разных DUT.

2.3. Единицы измерения

Ниже перечислены методы измерения, используемые при оценке производительности.

1) FILO **должен** применяться в качестве метода измерения, поскольку он учитывает задержку пакета; сегодняшним приложениям требуется прочитать весь пакет для обработки содержащейся в нем информации или выполнения действий.
2) FIFO **может** использоваться для некоторых приложений, способных обрабатывать данные с момента поступления первого бита - например, FPGA (Field-Programmable Gate Array).
3) LIFO применять **недопустимо**, поскольку в отличие от других методов он не учитывает задержку пакета.

3. Вариации задержки (Jitter)

3.1. Определение

В контексте ЦОД термин jitter является синонимом термина delay variation (вариации задержки). Они определяются путём многократного измерения задержки в одном направлении, как описано в RFC 3393. Обязательным для использования определением delay variation является PDV[1], определённая в параграфе 4.2 [RFC5481]. При рассмотрении потока пакетов задержки все пакетов задержки всех пакетов вычитаются из минимальной задержки всех

[1]Packet Delay Variation - вариации задержки пакетов.

пакетов в потоке. Это упрощает оценку диапазона вариаций задержки (Max - Min) или высокого процентиля PDV (99-й для устойчивости к постороннему трафику).

При использовании для измерения задержки временных меток First-bit - Last-bit, вариации задержки **должны** измеряться с применением пакетов или кадров одинакового размера, поскольку определение задержки включает время преобразования каждого пакета в последовательный формат (serialization time). В остальных случаях, если используется First-bit - First-bit, ограничений на размер не накладывается.

3.2. Обсуждение

В дополнение к диапазону PDV и/или высокому процентилю PDV **могут** использоваться межпакетные вариации задержки (IPDV[1]), определённые в параграфе 4.1 [RFC5481] (разность между двумя последовательными пакетами) для целей определения вариаций межпакетных интервалов (например, является поток пакетов сравнительно однородным или в нем возникают пики). Однако **не следует** использовать абсолютные значения IPDV, поскольку в них «сколлапсированы» пиковые и распределенные варианты потока.

3.3. Единицы измерения

Измерение вариаций задержки выражается в долях секунды. **Могут** также использоваться гистограммы PDV для демонстрации распределения.

4. Калибровка на физическом уровне

4.1. Определение

Калибровка на физическом уровне заключается в определении и измерении задержки физических устройств, применяемых для тестирования DUT.

Калибровка включает список всех используемых компонент физического уровня, как указано ниже.

- Тип устройства, используемого для генерации/измерения трафика.
- Тип линейных карт, используемых в генераторе трафика.
- Тип трансиверов в генераторе трафика.
- Тип трансиверов в DUT.
- Типы кабелей.
- Длины кабелей.
- Название и номер версии программ генерации трафика и DUT.
- **Может** предоставляться список разрешённых (включённых) функций DUT, которые поддерживаются и рекомендуются (особенно для протоколов уровня управления типа LLDP[2] и STP[3]). **Может** также предоставляться полная конфигурация.

4.2. Обсуждение

Калибровка на физическом уровне вносит вклад в сквозную задержку и её следует принимать во внимание при оценке DUT. Незначительное изменение физических компонент при тестировании может влиять на результат измерения задержки, следовательно калибровка **должна** быть описана в результатах тестов.

4.3. Единицы измерения

Рекомендуется применять для тестирования кабели (1) одного типа и длины, (2) произведённые (по возможности) одной компанией. Указанные в параграфе 4.1 параметры кабелей **должны** включаться в отчёт вместе с результатами. В отчёте. **необходимо** указать, вычитались ли задержки в кабелях из приведённых в значений. **Должна** указываться точность измерений генератора трафика (для современного тестового оборудования обычно около 20 нсек).

5. Скорость в линии

5.1. Определение

Синхронизация передачи или максимальная скорость передачи управляется «часами передачи» (transmit clock) в DUT. Синхронизация приёма (максимальная скорость на входе) определяется синхронизацией передачи подключённого интерфейса.

Скорость в линии или скорость передачи кадров на физическом уровне - это максимальная «ёмкость» передачи в линию кадров заданного размера с частотой синхронизации передачи устройства DUT.

Термин «номинальная скорость в линии» (nominal value of line rate) определяет максимальную скорость передачи для данного порта - например, 1 GE, 10 GE, 40 GE, 100 GE (в Гбит/с).

Частота (скорость часов - clock rate) синхронизации передачи в любой паре соединённых интерфейсов никогда не будет в точности совпадать, следовательно требуется некий допуск. Этот допуск выражается значением PPM[4]. Стандарты IEEE опускают определённые отклонения частоты синхронизации передачи и сети Ethernet рассчитаны на наличие незначительных расхождений между часами соединённых интерфейсов. Это приводит к некоторым допускам линейной скорости трафика, генерируемого тестовым оборудованием для DUT.

[1]Inter-Packet Delay Variation.
[2]Link Layer Discovery Protocol - протокол обнаружения канального уровня.
[3]Spanning Tree Protocol – протокол остовного дерева.
[4]Parts Per Million - число долей на миллион.

Скорость в линии **следует** измерять числом кадров в секунду (FPS[5]).

5.2. Обсуждение

Для синхронизации передачи большинство коммутаторов Ethernet использует «модуль часов» (clock module), называемый также «модулем синхронизации» (oscillator module), который представляет собой герметичный блок с внутренней температурной стабилизацией и обеспечивает очень высокую точность. Выходная частота такого модуля не настраивается, поскольку в этом нет необходимости. Однако в тестовом оборудовании зачастую имеется программная подстройка скорости передачи. Такую юстировку **следует** применять для «компенсации» скорости тестового оборудования, чтобы не передавать устройству DUT данные со скоростью, превышающей скорость линии.

Для допуска незначительных отклонений в скорости коммерчески доступных модулей синхронизации и других кварцевых генераторов стандарты Ethernet задают максимальные отклонения частоты синхронизации ±100 PPM от расчётного значения частоты. Следовательно, устройства DUT должны быть способны воспринимать кадры с отклонениями скорости ±100 PPM в соответствии со стандартами.

Очень мало устройств обеспечивает идеальную точность ±0,0 PPM в силу перечисленных ниже обстоятельств.

1. Стандарты Ethernet разрешают отклонение частоты не более ±100 PPM с течением времени. Следовательно, для опорных генераторов будет нормальным незначительное изменение частоты с течением времени и при изменении температуры, а также в результате воздействия других факторов.
2. Кристаллы кварца или модули часов обычно характеризуются некоторым отклонением, которое существенно меньше ±100 PPM. Зачастую эти вариации составляют не более ±30 PPM, чтобы устройство считалось «измерительным средством» (certification instrument).

При тестировании пропускной способности коммутаторов Ethernet на «скорости линии» любой конкретный коммутатор будет вносить свои вариации опорной частоты. Если тест выполняется с частотой +1 PPM по сравнению с частотой тестируемого коммутатора и тест происходит с установившейся скоростью линии, можно наблюдать постепенный рост задержки и, возможно, отбрасывание пакетов при переполнении буферов в коммутаторе. В зависимости от разницы вариаций частоты в двух соединённых устройствах можно заметить по истечении нескольких сотен микросекунд, нескольких миллисекунд или секунд с начала передачи трафика. Малую задержку и отсутствие потери пакетов можно продемонстрировать, установив для теста скорость чуть меньше чем при 100%-ой загрузке линии. Обычно загрузка в 99 % процентов показывает малую задержку и отсутствие потери пакетов. Ни в одном коммутаторе или маршрутизаторе Ethernet вы не увидите опорного генератора с отклонением опорной частоты в точности ±0,0 PPM. Очень мало (если есть) тестового оборудование обеспечивает точность ±0,0 PPM.

Производители тестового оборудования также осведомлены об этих стандартах и разрешают программно управляемый сдвиг (подстройку) опорной частоты в диапазоне ±100 PPM для компенсации вариаций частоты устройств DUT. Такая подстройка позволяет инженерам определить приблизительную скорость работы подключённого устройства и убедиться в том, что его параметры соответствуют требованиям стандартов.

5.3. Единицы измерения

«Скорость линии» (Line rate) измеряется числом кадров в единицу времени (frame rate):

```
Frame Rate = Transmit-Clock-Frequency /
   (Frame-Length*8 + Minimum_Gap + Preamble + Start-Frame Delimiter)
```

Minimum_Gap представляет интервал между кадрами. Эта формула «масштабируется вверх и вниз» для скоростей 1 GE, 10 GE и т. д.

Пример для скорости 1 GE и кадров размером 64 байта приведён ниже.

```
Frame Rate = 1000000000 / (64*8 + 96 + 56 + 8)
           = 1000000000 / 672
           = 1488095,2 FPS
```

С учётом допустимого отклонения ±100 PPM, коммутатор может «законно» передавать трафик со скоростью от 1487946,4 FPS до 1488244 FPS. Отклонение частоты на 1 PPM будет менять скорость на 1,488 FPS.

В реальной сети крайне маловероятно столкнуться с точной скоростью линии в течение очень короткого интервала времени. Различий в отбрасывании пакетов при скорости в 99% и 100% от скорости линии не наблюдается.

Скорость линии можно измерять при 100% с настройкой отклонения частоты -100 PPM.

Скорость линии **следует** измерять при 99,98% с отклонением 0 PPM.

Постройку PPM **следует** применять только при измерении скорости линии.

6. Буферизация

6.1. Буфер

6.1.1. Определение

Buffer Size - размер буфера

Термин «размер буфера» (buffer size) представляет общий объем памяти устройства DUT, служащей для буферизации кадров. Размер выражается в байтах (B), килобайтах (KB), мегабайтах (MB) или гигабайтах (GB). При указании размера буфера необходимо указывать также размер MTU (максимальный блок передачи) при тестировании, а также CoS (класс обслуживания) или DSCP (код дифференцированного обслуживания), поскольку распределение буферов зачастую определяется реализацией качества обслуживания. Дополнительную информацию можно найти в разделе 3 [RFC8239].

Пример. Значение Buffer Size устройства DUT при передаче кадров размером 1518 байтов составляет 18 MB.

[5]Frames per second.

Port Buffer Size - размер буфера в расчёте на порт

Это размер в расчёте на один порт для входного буфера, выходного буфера или относящейся к одному порту комбинации входного и выходного буфера. Отмечены три варианта размещения буферов, поскольку схема буферизации DUT может быть не известна или не проверена, поэтому информация о месте буферизации может прояснить архитектуру буферов и, следовательно, общий размер буфера. Значение Port Buffer Size является информационным и **может** предоставляться производителем DUT. Эти значения не тестируются при определении производительности, для оценки которой служит методология Maximum Port Buffer Size или Maximum Buffer Size.

Maximum Port Buffer Size - максимальный размер буфера на порту

Во многих случаях это совпадает с Port Buffer Size. В коммутаторах с архитектурой SoC[1] имеется буфер порта и общая для всех портов буферная ёмкость. Maximum Port Buffer Size в контексте буферизации SoC представляет собой сумму размера буфера порта и максимального размера общего буфера, выделяемого для этого порта, и выражается в байтах (B), килобайтах (KB), мегабайтах (MB) или гигабайтах (GB). Значение Maximum Port Buffer Size требуется указывать вместе со значением MTU, использованным при измерении, и установленным для теста значением CoS или DSCP.

Пример. Тестирование DUT показало наличие буфера порта размером 3 KB для кадров размером 1518 байтов и общего буфера с максимальным размером 4,7 MB для кадров размером 1518 байтов и CoS = 0.

Maximum DUT Buffer Size - максимальный размер буфера в устройстве

Это общий размер буфера, который может иметь DUT. Скорей всего, он отличается от Maximum Port Buffer Size. Обычно он отличается и от суммы значений Maximum Port Buffer Size. Значение Maximum Buffer Size должно указываться вместе с использованным при измерении значением MTU, а также значением CoS или DSCP.

Пример. В DUT было определено наличие буфера порта размером 3 KB для кадров размером 1518 байтов и максимальный размер общего буфера 4,7 MB для кадров того же размера. Для этого DUT значение Maximum Buffer Size составляет 18 MB при MTU 1500 B и CoS = 0.

Burst - пик, всплеск

Пиком называется фиксированное число пакетов, переданных при определённой (в процентах от скорости линии) для данного порта скорости. Относящиеся к пику пакеты равномерно распределены во временном интервале T. Может быть определена константа C для указания среднего интервала между 2 последовательными пакетами.

Microburst - микропик

Микропик представляет собой тип пика, для которого возникает отбрасывание пакетов при отсутствии установившейся или заметной перегрузки (насыщения) линии или устройства. Одной из характеристик микропика является отсутствие равномерного распределения в интервале T и интервалы меньше C (C - среднее время между двумя последовательными пакетами).

Intensity of Microburst - интенсивность микропика

Это процентное значение из диапазона 1 - 100% показывает уровень микропика. Чем больше значение, тем интенсивней микропик.

```
I=[1-[ (Tp2-Tp1)+(Tp3-Tp2)+....(TpN-Tp(n-1) ] / Sum(packets)]]*100
```

Приведённые выше определения предназначены не для оценки размера идеального буфера, а для описания способов измерения размера. Увеличение буфера не всегда обеспечивает положительный эффект и может вызывать проблемы (bufferbloat).

6.1.2. Обсуждение

При измерении буферизации в DUT важно понимать поведение всех и каждого порта. Это обеспечивает информацию об общем объёме буферов, доступном в коммутаторе. Определения эффективности буферов помогут понять оптимальный размер пакета для буфера или реальный объем буферного пространства для пакетов соответствующего размера. В этом разделе не рассматривается методология измерения, о приводятся разъяснения определений буферов и метрики, которые следует применять для комплексной оценки буферизации в ЦОД.

6.1.3. Единицы измерения

При измерении буферов в DUT:

- **должен** определяться размер буфера;
- определяется размер буфера, который **может** быть предоставлен на каждом порту;
- **должен** измеряться максимальный размер буфера порта;
- **должен** измеряться максимальный размер буфера DUT;
- при тестировании микропиков **может** быть указана их интенсивность;
- **следует** указывать значение CoS или DSCP в процессе тестирования.

6.2. Инкаст

6.2.1. Определение

Термин Incast широко применяется в контексте ЦОД для обозначения картин трафика «многие в один» (many-to-one) или «многие во многие» (many-to-many). Как определено в этом параграфе инкаст является мерой числа входных и выходных портов, а также процента синхронизации между ними. Обычно в ЦОД это относится ко множеству разных входных портов сервера (many), передающих трафик в общий восходящий канал (many-to-one) или множество восходящих каналов (many-to-many). Эта картина обобщается для сети, как множество входящих портов, передающих трафик в один или несколько восходящих каналов.

Synchronous arrival time - синхронное прибытие

Когда два или более кадров размера L1 и L2 приходят на соответствующий входной порт или множество входных портов и время прибытия перекрывается для любого из битов кадров, кадры L1 и L2 называют прибывшими синхронно. Это называется инкастом, независимо от картины many-to-one (проще) или many-to-many.

[1]Switch on chip - однокристальный коммутатор (микросхема). *Прим. перев.*

Asynchronous arrival time - асинхронное прибытие

К этому типу относятся все случаи, когда не наблюдается «синхронного прибытия» кадров.

Percentage of synchronization - процент синхронизации

Это значение определяет степень перекрытия (число битов) между кадрами размеров L1, L2, .., Ln.

Пример. Два 64-байтовых кадра протяжённостью L1 и L2[1] приходят на входные порты 1 и 2 устройства DUT. Имеется перекрытие времени прибытия кадров L1 и L2 на 6,4 байта. Следовательно, уровень синхронизации составляет 10%.

Stateful traffic - трафик с учётом состояния

Трафик пакетов, обмен которыми осуществляется по протоколу с поддержкой состояния соединения типа TCP.

Stateless traffic - трафик без учёта состояния

Трафик пакетов, обмен которыми осуществляется по протоколу без поддержки состояния соединения типа UDP.

6.2.2. Обсуждение

В этом сценарии буферы применяются на устройстве DUT. В механизме входной буферизации буферы входных портов будут применяться наряду с виртуальными выходными очередями, когда они доступны, тогда как в механизме выходной буферизации будет применяться выходной буфер одного исходящего порта.

В любом случае, независимо от расположения буферной памяти в архитектуре коммутатора, Incast ведёт к использованию буферов.

Синхронное прибытие нескольких кадров на устройство DUT рассматриваются, как формирование Incast.

6.2.3. Единицы измерения

В качестве единицы измерения **должно** служить число входных и выходных портов.

Процент синхронизации **должен** быть ненулевым и **должен** быть указан.

7. Пропускная способность для приложений

7.1. Определение

В ЦОД сбалансированность сети определяется максимальной пропускной способностью при минимальных потерях. Это описывается параметром Goodput [TCP-INCAST], определяющим пропускную способность на прикладном уровне. Для стандартных приложений TCP очень малые потери могут оказывать значительное влияние на пропускную способность приложений. Определение Goodput представлено в [RFC2647], а применяемая здесь трактовка является вариантом этого определения.

Goodput выражается числом битов за единицу времени, пересылаемых на нужный выходной интерфейс DUT, за вычетом битов, переданных повторно.

7.2. Обсуждение

При оценке пропускной способности ЦОД goodput представляет собой параметр, который **следует** измерять. Это даёт реалистичное представление об использовании доступной пропускной способности. Одной из целей для ЦОД является максимизация goodput при минимизации потерь.

7.3. Единицы измерения

Goodput (G) определяется формулой:

```
G = (S/F) * V
```

- S представляет число байтов информации (payload) без учёта заголовков пакетов и TCP;
- F - размер кадров;
- V скорость среды в байт/сек.

Пример. Передача файла по протоколу HTTP с использованием транспорта TCP в среде 10 Гбит/с.

Файл не может быть передан через среду Ethernet в форме одного непрерывного потока. Он должен разбиваться на множество кадров размером 1500 байтов при использовании стандартного значения MTU. Каждому пакету требуется 20 байтов для заголовка IP и 20 байтов для заголовка TCP, следовательно для передачи содержимого файла в пакете остаётся 1460 байтов. Системы на базе Linux вносят дополнительное ограничение до 1448 B, поскольку они дополнительно передают 12 байтов временной метки. Поскольку в этом примере данные передаются через сеть Ethernet к размеру пакета добавляется 26 байтов заголовка и результирующий кадр имеет размер 1526 байтов.

```
G = 1460/1526 * 10 Гбит/с, что даёт в результате 9,567 Гбит/с или 1,196 Гбайт/с.
```

Следует отметить, что в этом примере не учитывались дополнительные задержки Ethernet в виде межкадрового интервала (не менее времени на передачу 96 битов), а также возможные конфликты (коллизии) в среде, влияние которых зависит от нагрузки.

При измерении Goodput следует документировать в дополнение в перечисленным в параграфе 4.1 данным также:

- используемые стеки TCP;
- версии OS;
- Модель и номер версии микрокода сетевого адаптера (NIC).

Например, стеки TCP в Windows и разных версиях Linux могут влиять на результаты тестов на базе протокола TCP.

[1]Так в оригинале. *Прим. перев.*

8. Вопросы безопасности

Описанные в документе действия по измерению производительности ограничены определением характеристик технологии при использовании контролируемых воздействий в лабораторной среде с выделенным пространством адресов и ограничениями, описанными выше.

Оценка производительности сетевой топологии должна проводиться на независимом стенде, к которому **недопустимо** подключать устройства, могущие пересылать трафик в действующие сети или ошибочно маршрутизировать трафик из таких сетей в тестовую сеть.

Кроме того, оценка производительности выполняется для «чёрного ящика» лишь на основе измерений, выполняемых за пределами устройств DUT.

В тестируемых устройствах (DUT) **не следует** применять каких-либо возможностей специально для тестирования. Все влияния на безопасность сети, связанные с DUT, **следует** считать идентичными для тестовой и рабочей сети.

9. Взаимодействие с IANA

Этот документ не требует каких-либо действий со стороны IANA.

10. Литература

10.1. Нормативные документы

[RFC1242] Bradner, S., "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, DOI 10.17487/RFC1242, July 1991, <https://www.rfc-editor.org/info/rfc1242>.

[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>.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <https://www.rfc-editor.org/info/rfc2544>.

[RFC5481] Morton, A. and B. Claise, "Packet Delay Variation Applicability Statement", RFC 5481, DOI 10.17487/RFC5481, March 2009, <https://www.rfc-editor.org/info/rfc5481>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8239, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8239] Avramov, L. and J. Rapp, "Data Center Benchmarking Methodology", RFC 8239, DOI 10.17487/RFC8239, August 2017, <https://www.rfc-editor.org/info/rfc8239>.

10.2. Дополнительная литература

[RFC2432] Dubray, K., "Terminology for IP Multicast Benchmarking", RFC 2432, DOI 10.17487/RFC2432, October 1998, <https://www.rfc-editor.org/info/rfc2432>.

[RFC2647] Newman, D., "Benchmarking Terminology for Firewall Performance", RFC 2647, DOI 10.17487/RFC2647, August 1999, <https://www.rfc-editor.org/info/rfc2647>.

[RFC2889] Mandeville, R. and J. Perser, "Benchmarking Methodology for LAN Switching Devices", RFC 2889, DOI 10.17487/RFC2889, August 2000, <https://www.rfc-editor.org/info/rfc2889>.

[RFC3918] Stopp, D. and B. Hickman, "Methodology for IP Multicast Benchmarking", RFC 3918, DOI 10.17487/RFC3918, October 2004, <https://www.rfc-editor.org/info/rfc3918>.

[TCP-INCAST] Chen, Y., Griffith, R., Zats, D., Joseph, A., and R. Katz, "Understanding TCP Incast and Its Implications for Big Data Workloads", April 2012, <http://yanpeichen.com/professional/usenixLoginIncastReady.pdf>.

Благодарности

Авторы благодарны Al Morton, Scott Bradner, Ian Cox и Tim Stevenson за рецензии и отклики.

Адреса авторов

Lucien Avramov
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America
Email: lucien.avramov@gmail.com

Jacob Rapp
VMware
3401 Hillview Ave.
Palo Alto, CA 94304
United States of America
Email: jhrapp@gmail.com

Перевод на русский язык

Николай Малых

nmalykh@protokols.ru