ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ СТРУКТУРЫ ПАКЕТНОГО БУФЕРА ETHERNET КОММУТАТОРА

DOI: 10.36724/2072-8735-2020-14-1-18-24

Моисеев Виктор Игоревич,
Пермский государственный национальный исследовательский университет, г. Пермьа, Россия,
vim@psu.ru

Ключевые слова: *мультисервисные сети связи, сеть доступа, обслуживание очередей, IP-телевидение, QoS, качество обслуживания, пакетный буфер, нагрузочное тестирование.*

Исследуется метод детектирования дисциплины обслуживания пакетного буфера Ethernet коммутатора. Актуальность задачи обусловлена необходимостью верификации объемов и структуры буферной памяти коммутаторов, а также недостаточной или неоднозначной информацией, которую публикуют производители. *Цель исследования* – разработка метода экспериментального исследования Ethernet коммутатора, позволяющего делать выводы о структуре пакетных очередей, используемой политике обслуживания, абсолютном размере пакетного буфера. *Материалы и методы*. Определение размера буфера предлагается производить на основании наблюдения за поведением пакетного трафика, при его прохождении сквозь исследуемое устройство под большой нагрузкой. На основе предполагаемой архитектуры тракта обработки пакета построена модель переходного процесса нарастания пакетной очереди в различных точках тракта обработки пакета. Показано как структура и относительные размеры очередей влияют на наблюдаемые свойства пакетного трафика под предельной нагрузкой. Метод позволяет определить используемую структуру очередей – одна исходящая очередь, несколько исходящих очередей, несколько исходящих очередей с граничными условиями, приоритетная очередь, а также вариант с общим буферным пространством. Приводятся обоснования применимости метода при различных конфигурациях структуры очередей. Результаты. Приведены данные экспериментов с различными вариантами форматирования пакетного буфера и различными структурами нагрузочного трафика. Проведено экспериментальное вычисление размеров пакетного буфера на нескольких популярных моделях Ethernet коммутаторов. Экспериментально показано как определенные варианты структурирования буферов влияют на поток транзитного трафика, а также когда наблюдаются непредвиденные потери. Выявлены варианты способствующие проявлению нежелательного эффекта перегрузки клиентского порта. Последовательно проведен анализ переходных процессов во всех экспериментальных конфигурациях и даны рекомендации по их применимости. Предложенный метод пригоден для определения абсолютного размера пакетного буфера коммутатора, а также структуры очередей в классических коммутаторах с промежуточной буферизацией.

Информация об авторе:
Моисеев Виктор Игоревич, *ведущий программист отдела ИВС Университетского Центра Интернет ПГНИУ; старший преподаватель кафедры РиЗИ ПГНИУ; аспирант кафедры МСИБ ПГУТИ, Пермский государственный национальный исследовательский университет, г. Пермь, Россия*

Для цитирования:
Моисеев В.И. Экспериментальное исследование структуры пакетного буфера Ethernet коммутатора // T-Comm: Телекоммуникации и транспорт. 2020. Том 14. №1. С. 18-24.

For citation:
Moiseev V.I. (2020) Experimental evaluation of Ethernet switch packet buffer structures. *T-Comm*, vol. 14, no.1, pp. 18-24. (*in Russian*)

Введение

Современные многопортовые Ethernet-коммутаторы работают по принципу буферизации проходящих Ethernet-фреймов. В обычных коммутаторах уровня доступа реализована схема “полностью принять, только потом отправить” (store-and-forward). Когда порт коммутатора принимает фрейм, производится контроль ошибок, принимается решение о дальнейшей коммутации и только потом фрейм покидает коммутатор с исходящего порта. Все это время фрейм находится в так называемом пакетном буфере, или проходит через серию пакетных буферов во внутренних коммутационных элементах. Существуют различные алгоритмы скоростной коммутации (cut-through), но даже в них происходит переключение на классическую схему буферизации, например при конкуренции за исходящий порт. Таким образом, пакетный буфер – неотъемлемая часть коммутационного тракта, позволяющая повысить загрузку коммутатора при неравномерностях в интенсивности поступления потока трафика.

Абсолютные размеры буферов важны для грамотного планирования емкости сети. Увеличение очередей приводит к задержкам, тогда как уменьшение буфера ниже определенных величин приводит к потерям трафика. Для отдельных типов трафика, таких как голосовой, интерактивный трафик, предельные значения потерь и задержек регламентированы различными стандартами и даже нормативно правовыми актами [1-3]. Допустимые значения задержек и потерь также включаются в договора на предоставление услуги по передаче данных. К сожалению, данные о размерах пакетных буферов публикуются далеко не всеми производителями и не для всех моделей коммутаторов [4, 5]. Ситуация усложняется также тем, что внутренняя организация буферов в конкретной модели коммутатора неизвестна [6–8]. Все это усложняет планирование сети передачи данных и поиск причин неисправностей в существующих сетях.

В настоящей работе рассматривается модель простой FIFO-очереди пакетного буфера, и варианты буфера с двумя очередями. Исследована методика стендового эксперимента в задаче определения объема буфера. Рассчитаны размеры буферной памяти на основании экспериментальных данных. Исследованы возможные варианты структурирования очередей на коммутаторе Cisco Catalyst 2960G-24 и проведена серия экспериментов по изучению влияния данных структур на потоки трафика.

1. Методика определения размеров и структуры пакетного буфера

Рассмотрим несколько простых схем организации очередей и различия в их влиянии на проходящий трафик. Простейшая дисциплина обслуживания очереди – FIFO (“первый пришел – первый вышел”). Все пакеты трафика равноправны и на нагруженном порту буферизуются в очереди. Очередь растет до определенного максимального размера, после чего вновь прибывшие пакеты отбрасываются (схема “tail-drop”). Когда требуется обеспечить различные уровни обслуживания, пакеты различных классов маркируются либо в момент прихода на порт, либо заранее – в процессе передачи по сети. Маркировка производится в полях DSCP в заголовке IP, и/или в полях CoS кадров Ethernet [9-10].

Одна из популярных схем реализации приоритетного обслуживания – выделение на буфере приоритетной очереди – трафик из данной очереди будет обслужен вне зависимости от того, ожидают ли пакеты в обычной очереди. Размеры обычной и приоритетной очереди могут отличаться. Другая распространенная схема – выделение в очереди промежуточных предельных значений, которые действуют только на отдельные классы трафика. Пакеты трафика определенного класса будут отброшены при превышении граничного значения заданного именно для этого класса. Примерная архитектура буфера с двумя очередями, одна из которых приоритетная, представлена на рис.1 (слева). На рисунке 1 (справа) изображена одна очередь с несколькими промежуточными границами.

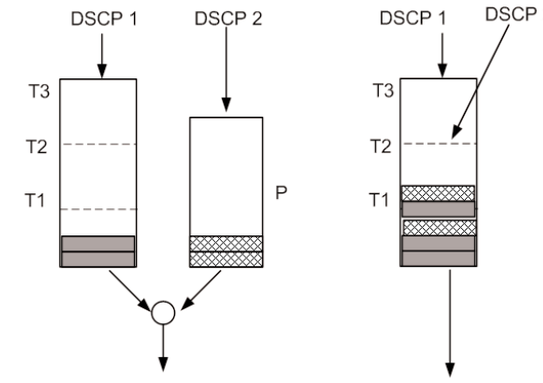

Рис 1. Схема организации буфера с приоритетной очередью (слева), и с тремя промежуточными границами в единственной очереди (справа)

Используемая методика экспериментального определения размера буфера порта основана на наблюдении влияния перегруженного порта на временные свойства транзитного трафика [11]. Пусть пакеты поступают с постоянной интенсивностью, время обработки всех пакетов одинаково и зависит только от размера пакета. Примем также, что интенсивность обработки пакетов постоянна и известна, а количество мест в очереди (буфере) конечно и равно *N*. Такую схему СМО принято обозначать *D/D/1/N*. Аналитическое решение для подобной очереди дано, например, в [12]. Исследуемый коммутатор включается в разрыв между двумя тестирующими узлами [13]. С узла источника запускается непрерывный поток пакетов одинаковой длины с битовой интенсивностью л. Каждый пакет нумеруется порядковым номером. Абсолютным значением межпакетного интервала пренебрежем по сравнению с размером пакета. Настроим коммутатор на фиксированную постоянную интенсивность отправки пакетов – м. Предположим, что интенсивность поступления пакетов превышает интенсивность обработки (рис. 2).

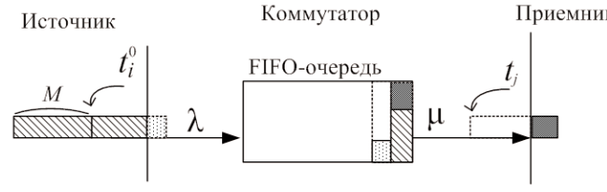

Рис. 2. Модель передачи пакета сквозь коммутатор с буферизацией

Примем, что вся задержка на коммутационном тракте будет состоять из времени сериализации пакета на входе и на выходе из буфера. Коммутатор начинает обрабатывать пакет только после полного приема пакета в буфер. Прием пакета в буфер возможен только в том случае, если имеется хотя бы одна свободная ячейка. Пакет занимает ячейку целиком вплоть до окончания обработки. Обозначим за t^0_i момент времени, когда пакет с порядковым номером i покинет источник.

На рисунке 3 представлены временные диаграммы процесса отправки пакетов с источника, прохождения через буфер коммутатора, и получения их приемником. Диаграмма для буфера представлена со сквозной нумерацией битов (как для кольцевого буфера), интенсивности л и м взяты произвольно для наглядности, причем л > м. Пусть пакеты пронумерованы сквозной нумерацией. Из диаграммы видно, что на момент прихода в буфер пакетов со 2 по 6 в буфере все еще занята ячейка пакетом номер 1. Текущая длина очереди обозначена $q(t)$. Наклон прямых на диаграмме представляет собой битовую интенсивность поступления и обработки пакетов – л и м, соответственно.

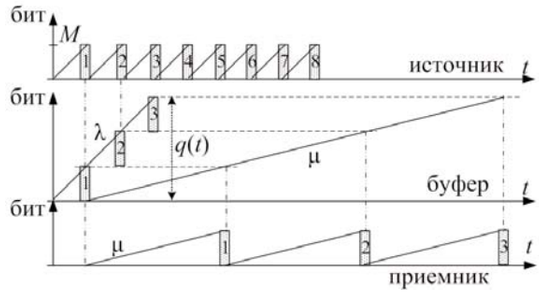

Рис. 3. Временные диаграммы процесса отправки пакетов с источника (верхняя), прохождения через буфер коммутатора (средняя), и получения их приемником (нижняя)

Очевидно, что при превышении скорости приема над скоростью обработки, пакеты будут накапливаться в буфере. Введем ограничение на длину очереди. Пусть буфер ограничен всего тремя ячейками памяти: $N = 3$. Соответствующие временные диаграммы процессов отправки пакетов на рис. 4. Из диаграммы видно, что во время отправки пакета номер 1 на коммутатор успевают прийти пакеты с номерами от 2 до 6. При этом пакет 1 все еще занимает очередь, 2 и 3 становятся в очередь. На моменты прихода 4 и 5 пакетов в очереди нет свободных ячеек и пакеты отбрасываются (на диаграмме отмечены символом “X”). Даже когда начинает поступать пакет номер 6, буфер все еще занят, и пакет отбрасывается.

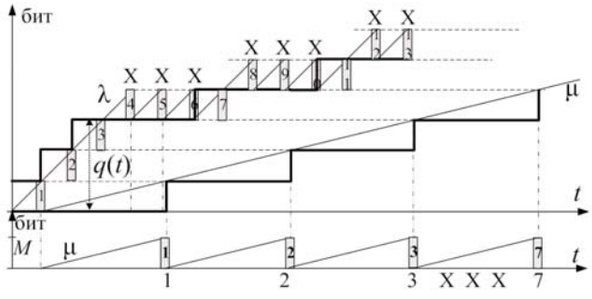

Рис. 4. Временные диаграммы переполнения пакетного буфера

В буфере появляется свободное место только к момен прихода 7 пакета. Далее очередь остается в состоянии н сыщения и входит в фазу постоянной флуктуации количес ва занятых ячеек от значения N до N–1. С точки зрения те рии СМО этот процесс рассмотрен в работе [14]. Предпол жим, что значение N достаточно велико, чтобы опустить рассмотрения вышеозначенные флуктуации.

Обратим внимание на то, как подобный поток трафи воспринимается приемником. Сначала приемник фиксиру стабильный поток трафика интенсивности м, номера прих дящих пакетов строго последовательны. Через некотор время непрерывность нумерации нарушается, и Приемн детектирует исчезновение серии пакетов. Интенсивнос поступления пакетов на Приемник неизменна. Времена пр хода пакетов и их номера фиксируются Приемником д дальнейшей обработки [15].

Теперь, зная механизм эволюции очереди, построи диаграмму занятого битового объема в пакетном буфе от времени (рис. 5).

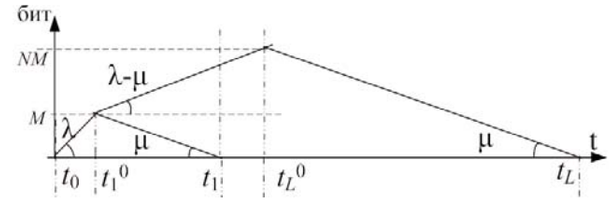

Рис. 5. Временная диаграмма занятого объема пакетного буфера емкостью N пакетов

Для битового размера буфера при данных условиях, основании графиков заполнения очереди, можем получи формулу для расчета битового размера буфера B:

$$B = M + (t_L - t_1)\left(\mu - \frac{\mu^2}{\lambda}\right),$$

где M – битовый размер пакета, t_1 – экспериментально з фиксированное время приема первого пакета, t_L – экспер ментально зафиксированное время получения последне пакета с корректным номером. Подробный вывод формул расчета буфера представлен в работе [11].

Расширим данную методику для применения к пакетн му буферу, представленному на рис. 1 (слева). Примени два независимых источника трафика (рис. 6).

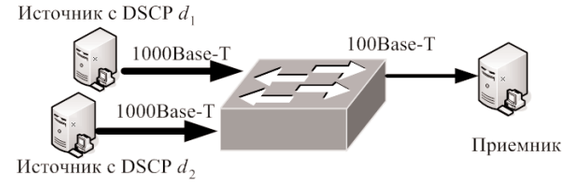

Рис. 6. Схема включения коммутатора с двумя источниками и одним получателем

Из представленной схемы очевидно, что любой двух источников трафика может перегрузить пакетную оч редь и ввести исходящий порт в состояние перегрузки. На нем генерацию трафика на первом источнике с максималы возможной интенсивностью, по вышеописанной методике,

DSCP d_1. Вычислим объем буфера относительно первого потока – B_1. Убедимся при помощи анализатора трафика на Приемнике, что некоторые пакеты отбрасываются коммутатором, и очередь на порту находится в состоянии перегрузки. При этом размер очереди в пакетах находится в постоянной флуктуации от максимального размера N до $N – 1$.

Далее, не останавливая первый поток трафика, проводим эксперимент аналогично, но с независимого источника трафика, также на максимальной интенсивности, с пакетами, промаркированными отличным от первого кодом DSCP d_2. Вычислим объем буфера относительно второго потока – B_2.

В случае если оба потока попадают в одну и ту же очередь и не имеют приоритетов относительно друг друга, на втором потоке Приемник не зафиксирует начального линейного участка с непрерывными номерами пакетов. Вычислить B_2 не получится, что равносильно значению B_2 эквивалентному одному пакету.

Таким образом, если значение B_2 удается измерить при многократном проведении эксперимента, значит трафик с DSCP d_2 попадает в приоритетную очередь и ее объем равен B_2. Теперь измерим размеры очередей для каждого из потоков d_1 и d_2 независимо, без трафика от второго источника. Обозначим полученные размеры B'_1 и B'_2, соответственно. Если $B'_2 = B_2$, то приоритетная очередь является выделенной, и менее приоритетный трафик на нее не влияет.

Если равенство не выполняется, то приоритетный трафик d_2 использует ту же очередь, что и d_1, но пакеты попадают под действие различных граничных значений (см. рис. 1 (справа)). Перебрав все комбинации значений DSCP (или, что проще, поле CoS в IEEE 802.1q метке) для обоих 1оттоков трафика, мы можем получить таблицу соответствия значений DSCP конкретным очередям и граничным уровням.

2. Результаты эксперимента

В эксперименте исследовалась структура и размеры пакетного буфера коммутатора Cisco Catalyst WS-C2960G-24TC-L, ревизия B0, с установленной версией ПО Cisco IOS c2960-lanbasek9-mz.122-58.SE2. Производитель не публикует данных об объемах буферов на данном коммутаторе, но, согласно комментариям разработчиков [16] каждая ASIC-микросхема на данной платформе оперирует 576 КБ буфера на 4 смежных порта.

Во всех экспериментах источник трафика подключается на порт с принудительно выставленной скоростью 100 Мбит/с (100Base-T), а приемник трафика подключается в порт на скорости 10 Мбит/с (10Base-T). Отправитель маркирует каждый пакет порядковым номером. Приемник детектирует время прихода каждого пакета и его порядковый номер с помощью ПО захвата пакетов tcpdump.

В первом эксперименте сравнивается поведение трафика при отключенном управлении качеством обслуживания (QoS) и включенном. На рисунке 7 представлены графики зависимости принятого порядкового номера пакета трафика от времени прибытия этого пакета. На графиках представлены по три результата для каждого варианта настройки.

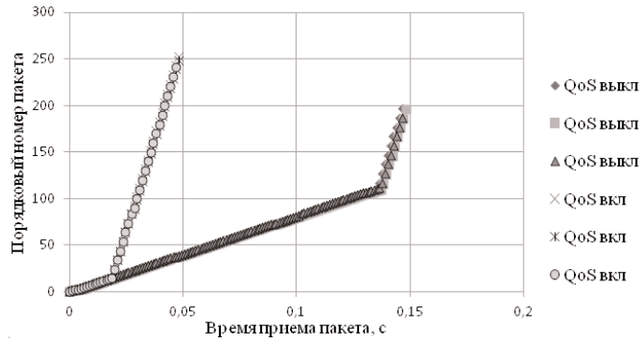

Рис.7. Зависимость порядкового номера принятого пакета от времени с включенным и выключенным функционалом QoS

Как видно из рис. 7, при отключенных настройках QoS коммутатор передает без потерь 110±1 пакетов, после чего буфер исходящего порта переполняется и мы видим значительные (? 90%) потери пакетов. Такой процент потерь объясняется десятикратным различием в скорости входящего и исходящего порта. Назовем период времени жизни потока на переполненной очереди фазой отсечки. Время получения последнего пакета с корректным номером составляет 136 мс. Эти же значения после включения QoS составляют 15 пакетов и 18 мс. Рассчитав по представленной методике объем буфера получим 155 КБ и 22 КБ для отключенного и включенного QoS соответственно. Исходя из значения 155 КБ, можно предположить, что каждая ASIC-микросхема делит 576 КБ буферного пространства равномерно на 4 порта. При включении механизмов управления приоритетами (QoS) данная модель коммутатора, согласно документации, форматирует буфер порта в 4 очереди, при этом каждая очередь получает 25% от исходного размера буфера и лишь половина этого пространства резервируется. Таким образом, мы ожидаем получить восьмикратное уменьшение доступного одной очереди буфера при включении QoS. Полученный экспериментально результат соответствует нашим предположениям.

Во втором эксперименте мы включаем функционал QoS и направляем трафик от двух источников с разными метками DSCP на один приемник, причем на выходном порту будем использовать одну очередь с двумя граничными уровнями для этих классов трафика. Очередь использует 25% общего буфера. Граница для менее приоритетного трафика (T1) выставлена на уровень 100% от объема очереди, граница для более приоритетного трафика – 150% (T2). Значение больше 100% означает, что очередь может пользоваться общим буфером сверх зарезервированного пространства. Более приоритетный поток запускается с небольшой задержкой относительно начала первого потока. Один из графиков серии представлен на рис. 8.

На графике (рис. 8) мы можем наблюдать полное прекращение приема менее приоритетного трафика после появления более приоритетного. Это можно объяснить логикой работы одной очереди с двумя граничными значениями (см. рис. 1, справа). Первый поток после старта эксперимента полностью заполняет доступную ему часть очереди и переходит в фазу отсечки.

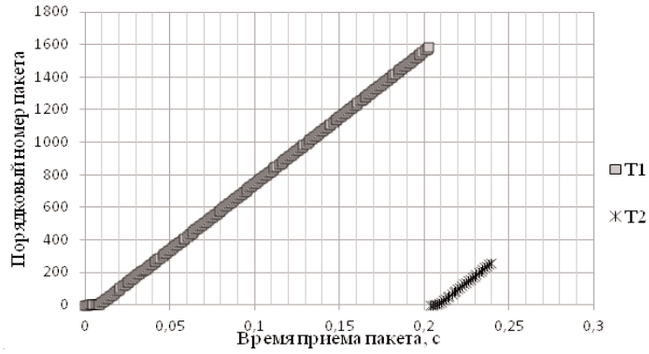

Рис.8. Зависимость порядкового номера принятого пакета от времени для двух потоков с разными границами буфера: T1 и T2

Пакеты второго потока застают нижнюю часть очереди полностью занятой, но имеют возможность заполнять очередь выше, до своего граничного значения. Заполнив указанное пространство, приоритетный поток также переходит в фазу отсечки. Время от начала потока до начала потерь для низкоприоритетного трафика составило 8,5 мс, для высокоприоритетного – 4 мс. Расчетные значения размеров очереди составляют 11 КБ и 4 КБ соответственно. Заметим, что второе значение соответствует разнице между граничными уровнями, т.е. реальное значение объема для второго граничного уровня – 15 КБ. Это значение меньше, чем значение из предыдущего эксперимента для одной очереди.

В следующем эксперименте мы используем такую же топологию – два источника и один приемник, но одну из четырех очередей объявим приоритетной. На основной очереди оставляем граничное значение 100%. Мы ожидаем, как и в случае с двумя граничными значениями, что старт приоритетного потока приведет к полной остановке низкоприоритетного трафика. Результат представлен на рис. 9.

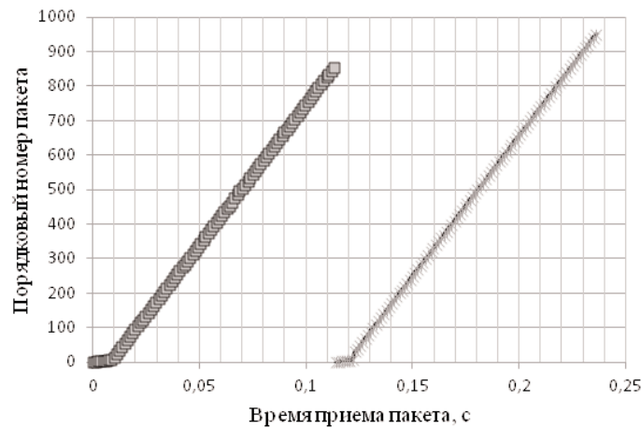

Рис. 9. Зависимость порядкового номера принятого пакета от времени для обычного потока (T1) и приоритетного (P)

Как и ожидалось, приоритетный трафик, заполнив свою очередь, парализует движение низкоприоритетного потока. Вычисленные значения объема приоритетной и обычной очереди составили 10 КБ и 11 КБ соответственно. Сравнивая Рис. 8 и 9 можем заключить, что по данной методике нельзя отличить работу коммутатора с одной очередью и разными граничными значениями от структуры с приоритетной очередью.

В следующем эксперименте мы пронаблюдаем работу очереди с общей памятью. Одной из четырех исходящих очередей порта мы назначим максимально допустимый размер, остальным – минимальный. Производитель разрешает резервировать для очереди не более 100% объема, но заявляет, что очередь может использовать до 3200% памяти из некоего общего буферного пространства (“common pool”) [17]. Результаты представлены на рис. 10.

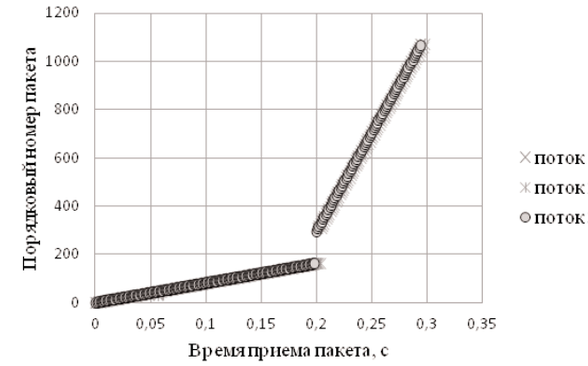

Рис. 10. Зависимость порядкового номера пакета от времени для серии потоков с использованием максимально возможного объема памяти

График для данного эксперимента резко отличается от предыдущих – после серии из 162 равномерно идущих пакетов происходит резкий скачок и 140 пакетов теряются. После скачка поток переходит в фазу отсечки. Рассчитанное по первой части графика значение буфера составляет 227 КБ, что в полтора раза больше, чем значение буфера при выключенном функционале QoS. Аномальную потерю пакетов можно объяснить следующим образом.

Каждый порт имеет входящий буфер, в котором пакеты ожидают очереди для доступа на коммутационную матрицу. Входящий и исходящий буферы размечены на одной и той же памяти. При медленной скорости вывода пакетов возможно затирание исходящим буфером части ячеек входящего буфера в пределах одной микросхемы ASIC. Похожий эффект наблюдался ранее на коммутаторах центров обработки данных в работах [18-19], но эффект объяснялся несовпадением частоты работы диспетчера обслуживания конкретного порта и диспетчера коммутационной матрицы.

В последнем эксперименте мы постарались пронаблюдать конкуренцию двух потоков из разных равноправных очередей за выходной порт. Показательный результат из серии представлен на рис. 11.

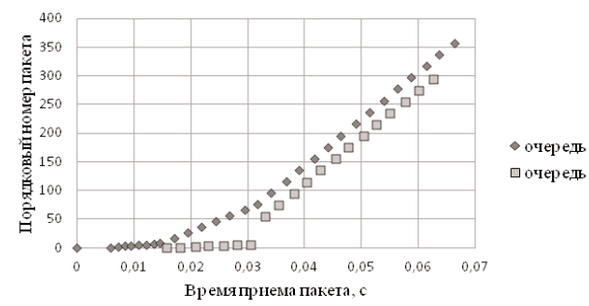

Рис. 11. Зависимость порядкового номера пакета от времени двух потоков назначенных в разные исходящие очереди

Мы ожидаем, что обе очереди будут обслуживаться исходящим портом с одинаковой интенсивностью, т.к. ни одна из них не является приоритетной. Как видно из рис. 11 именно это и происходит на начальной фазе заполнения буфера и в фазе отсечки. Тем не менее, в потоке из второй очереди мы видим характерный скачок из 50 потерянных пакетов. Это перекликается с результатами предыдущего эксперимента и сигнализирует о наличии некоего процесса переформатирования общего буфера.

Экспериментальные данные о размерах исходящего буфера нескольких популярных моделей коммутаторов с отключенным управлением качеством обслуживания представлены в табл. 1, с различными настройками QoS – в табл. 2.

Таблица 1

Экспериментальные полученные размеры исходящего буфера на некоторых коммутаторах при выключенном функционале QoS

Коммутатор	Объем буфера, КБ	Объем буфера, кол-во пакетов 1500Б
WS-C2960-24TT-L	124	82
WS-C2960S-48FPD-L	71	48
WS-X6748-GE-TX	376	249
WS-C2960G-24TC-L	154	103

Таблица 2

Экспериментальные полученные размеры исходящего буфера на WS-C2960G-24TC-L при различных параметрах управления QoS

Параметры QoS	Объем буфера, КБ	Объем буфера, кол-во пакетов 1500Б
Управление QoS выключено, 1 очередь	154	103
Управление QoS включено, 1 очередь	22	15
Управление QoS включено, 1 очередь, 2 уровня (первый поток)	11	7
Управление QoS включено, 1 очередь, 2 уровня (второй поток)	6	4
Управление QoS включено, 2 очереди (одна из которых приоритетная) – обычный поток	11	7
Управление QoS включено, 2 очереди (одна из которых приоритетная) – приоритетный поток	10	7
Управление QoS включено, 2 очереди (первый поток)	11	7
Управление QoS включено, одной очереди выдан максимальный объем общего буфера	227	150

Выводы

Представленная методика экспериментального исследования структуры буферов Ethernet-коммутатора позволяет вычислять абсолютные размеры исходящего пакетного буфера. Исходя из получаемых графиков, можно делать выводы о конкретной структуре очередей, в том числе о размере каждой очереди и привязке класса трафика к очереди. Метод также позволяет диагностировать случаи аномально больших потерь трафика в определенных схемах включения, что дает возможность в дальнейшем принимать меры для исключения подобного поведения. Вопрос экспериментального детектирования приоритетной очереди и ее отличия от одной очереди с несколькими граничными условиями требует дополнительного изучения.

Литература

1. *Szigeti T., Hattingh C., Barton R., Briley K.* End-to-End QoS Network Design: Quality of Service for Rich-Media & Cloud Networks, 2nd Edition // Cisco Press, 2012. 1040 c.
2. Transport of MPEG-2 TS Based DVB Services over IP Based Networks ETSI TS 102 034 V2.1.1 // European Broadcasting Union, София-Антиполис, Франция, 2016. 331 c.
3. Приказ Министерства информационных технологий и связи Российской Федерации от 27.09.2007 № 113 "Об утверждении Требований к организационно-техническому обеспечению устойчивого функционирования сети связи общего пользования" [Электронный ресурс] URL: http://minsvyaz.ru/ru/documents/3921/ (дата обращения: 01.09.2019).
4. *Warner J.* Packet buffers [Электронный ресурс] URL: https://people.ucsc.edu/~warner/buffer.html (дата обращения: 01.02.2019).
5. Mellanox Spectrum vs. Broadcom StrataXGS Tomahawk [Электронный ресурс]. Tolly Group, 2016 URL: http://www.mellanox.com/related-docs/products/tolly-report-performance-evaluation-2016-march.pdf (дата обращения: 10.10.2019).
6. Arista LANZ Overview [Электронный ресурс] URL: https:// people.ucsc.edu/~warner/Bufs/Arista_LANZ_Overview_TechBulletin_0213.pdf (дата обращения: 26.08.2019).
7. Intel Ethernet Switch Family Memory Efficiency Non-blocking Fabric Architecture [Электронный ресурс] URL: https://people.ucsc.edu/~warner/Bufs/intel-memory-efficiency-paper.pdf (дата обращения: 26.09.2019).
8. Speeding Applications in Data Center Networks [Электронный ресурс] URL: https://miercom.com/pdf/reports/ 20160210.pdf (дата обращения: 01.10.2019).
9. *Grossman D.* New Terminology and Clarifications for Diffserv [Электронный ресурс] URL: https://tools.ietf.org/html/rfc3260 (дата обращения: 14.10.2019).
10. IEEE Standard for Local and Metropolitan Area Network--Bridges and Bridged Networks / Institute of Electrical and Electronics Engineers, 2018. [Электронный ресурс] URL: https://ieeexplore.ieee.org/document/8403927 (дата обращения: 10.10.2019).
11. *Моисеев В.И.* Метод аудита размера пакетного буфера коммутатора // Вестник Пермского университета. Серия «Информационные системы и технологии», Вып. 1. Пермь. 2018. С. 32-35.
12. *Zuckerman M.* Introduction to Queueing Theory and Stochastic Teletraffic Models [Электронный ресурс] URL: https://arxiv.org/pdf/1307.2968.pdf (дата обращения: 14.10.2019).
13. *Bradner S., McQuaid J.* RFC-2544. Benchmarking Methodology for Network Interconnect Devices [Электронный ресурс] URL: https://www.ietf.org/rfc/rfc2544.txt (дата обращения: 26.09.2019).
14. *Garcia J.-M., Brun O., Gauchard D.* Transient Analytical Solution of M/D/1/N Queues // Journal of applied Probability. Vol. 39. No. 4 (Dec., 2002), С. 853-864.
15. PCAP-TSTAMP – packet time stamps in libpcap [Электронный ресурс] URL: https://www.tcpdump.org/manpages/pcap-tstamp.7.txt (дата обращения: 26.09.2019).
16. *Tsukerman A.* Buffer size on 3750G [Электронный ресурс] URL: https://people.ucsc.edu/~warner/Bufs/3750G-buf.pdf (дата обращения: 10.10.2019).
17. Catalyst 2960 and 2960-S Switches Software Configuration Guide, Release 12.2(58)SE [Электронный ресурс] URL: https://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst2960/software/release/12-2_58_se/configuration/guide/2960scg/swqos.html (дата обращения: 10.10.2019).
18. *Chen Y., Griffith R., Liu J., Katz R.* Understanding TCP Incast Throughput Collapse in Datacenter Networks // Материалы конференции WREN'09, 21 августа, 2009, Барселона, Испания.
19. *Prakash P., Dixit A., Kompella R.* The TCP Outcast Problem: Exposing Unfairness in Data Center Networks [Электронный ресурс] URL: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final126.pdf (дата обращения: 10.10.2019).

EXPERIMENTAL EVALUATION OF ETHERNET SWITCH PACKET BUFFER STRUCTURES

Victor I. Moiseev, *Perm State University, Perm, Russia,* ***vim@psu.ru***

Abstract

A method to detect or verify actual packet buffer size of an Ethernet switch with different queuing disciplines presented. In enterprise and datacenter networking environment there exists a need for a method to experimentally verify or measure exactly how deep packet buffers are and which structures and service disciplines do they have. Also there exists lack of published specifications from switch vendors on these topics. Aim. To develop a method to detect queuing discipline and actual buffer sizes of Ethernet switches. Materials and methods. Based on possible buffer architectures we study effects of different engineering decisions on observed traffic patterns. We show how from these patterns internal buffer schemes could be revealed. Buffers are verified on size and priority handling. Buffer sizes estimated on the basis of analyzed packet loss under overload conditions. Results. We present numeric results of buffer size estimation for an Ethernet switch of popular vendor and give some thoughts on how modern complex QoS schemes can be identified and verified. We also show experimental data on packet loss and packet flow structures in several configurations. In some cases incast or outcast collapse effects observed. Conclusion. The method presented is suitable for reliable verification of packet buffer sizes and queue structures in store-and-forward Ethernet switches.

Keywords: *scheduling discipline, QoS, packet buffer, priority queue, stress testing.*

References

1. Szigeti T., Hattingh C., Barton R., Briley K. *End-to-End QoS Network Design: Quality of Service for Rich-Media & Cloud Networks*, 2nd Edition. Cisco Press, 2012.- 1040p.
2. Transport of MPEG-2 TS Based DVB Services over IP Based Networks ETSI TS 102 034 V2.1.1 European Broadcasting Union, France, 2016 - 331p.
3. Order of the Ministry of Information Technologies and Communications of the Russian Federation of September 27, 2007 No. 113 "On approval of the Requirements for the organizational and technical support for the stable functioning of the public communications network" [Prikaz Ministerstva informatzionnyh tehnologiy I svyazi Rossiyskoy Federatzii ot 27.09.2007 #113 "Ob otverzhdenii Trebovaniy k organizatzinno-tehnicheskomu obespecheniyu ustoychivogo funktsionirovaniya seti svyazi obshego polzovaniya"] Available at: URL: http://minsvyaz.ru/ru/documents/3921/ (accessed: 01.09.2019).
4. Warner J. Packet buffers Available at: URL: https://people.ucsc.edu/~warner/buffer.html (accessed: 01.02.2019).
5. Arista LANZ Overview Available at: URL: https:// people.ucsc.edu/~warner/Bufs/Arista_LANZ_Overview_TechBulletin_0213.pdf (accessed: 26.09.2019).
6. Mellanox Spectrum vs. Broadcom StrataXGS Tomahawk / Tolly Group, 2016 Available at URL: http://www.mellanox.com/related-docs/products/tolly-report-performance-evaluation-2016-march.pdf (accessed: 10.10.2019).
7. Intel Ethernet Switch Family Memory Efficiency Non-blocking Fabric Architecture Available at: URL: https://people.ucsc.edu/~warner/Bufs/intel-memory-efficiency-paper.pdf (accessed 26.09.2019).
8. Speeding Applications in Data Center Networks Available at URL: https://miercom.com/pdf/reports/20160210.pdf (accessed: 01.10.2019).
9. Grossman D. New Terminology and Clarifications for Diffserv Available at URL: https://tools.ietf.org/html/rfc3260 (accessed: 14.10.2019).
10. IEEE Standard for Local and Metropolitan Area Network--Bridges and Bridged Networks / Institute of Electrical and Electronics Engineers, 2018. Available at: URL: https://ieeexplore.ieee.org/document/8403927 (accessed: 10.10.2019).
11. Moiseev V.I. Switch Packet Buffer Audit Method [Metod audita paketnogo bufera kommutatora]. *Vestnik Permskogo Universiteta. Seriya "Informatsionniye Sistemy i Tehnologii"*. Vol. 1, Perm, 2018, pp. 32-35.
12. Zuckerman M. Introduction to Queueing Theory and Stochastic Teletraffic Models Available at: URL: https://arxiv.org/pdf/1307.2968.pdf (accessed: 14.10.2019).
13. Bradner S., McQuaid J. RFC-2544. Benchmarking Methodology for Network Interconnect Devices Available at: URL: https://www.ietf.org/rfc/rfc2544.txt (accessed 26.09.2019).
14. Garcia J.-M., Brun O., Gauchard D. Transient Analytical Solution of M/D/1/N Queues. *Journal of applied Probability*. Vol. 39, No. 4 (Dec.,2002), pp. 853-864.
15. PCAP-TSTAMP – packet time stamps in libpcap Available at: URL: https://www.tcpdump.org/manpages/pcap-tstamp.7.txt (accessed: 26.09.2019).
16. Tsukerman A. Buffer size on 3750G Available at: URL: https://people.ucsc.edu/~warner/Bufs/3750G-buf.pdf (accessed: 10.10.2019).
17. Catalyst 2960 and 2960-S Switches Software Configuration Guide, Release 12.2(58)SE Available at: URL: https://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst2960/software/release/12-2_58_se/configuration/guide/ 2960scg/swqos.html (accessed: 10.10.2019).
18. Chen Y., Griffith R., Liu J., Katz R. Understanding TCP Incast Throughput Collapse in Datacenter Networks. *Proceedings of WREN'09*, 21 august, 2009, Barcelona, Spain.
19. Prakash P., Dixit A., Kompella R. The TCP Outcast Problem: Exposing Unfairness in Data Center Networks Available at: URL: https://www.usenix.org/ system/files/conference/nsdi12/nsdi12-final126.pdf (accessed: 10.10.2019).

Information about author:

Victor I. Moiseev, *Lead programmer of IT-department of Perm State University, assistant professor of Faculty of Physics of Perm State University, Perm, Russia*