УДК 621.39

СЕТИ ДОСТУПА С ОТНОСИТЕЛЬНЫМ ПРИОРИТЕТОМ

А.Г. КОСТЮКОВСКИЙ

Высший государственный колледж связи
П. Бровки, 14, Минск, 220027, Беларусь

Поступила в редакцию 1 ноября 2013

Приведены результаты разработки и исследования математической модели сети доступа с относительным приоритетом и с предоставлением абонентам услуги гарантированного качества Triple Play (три в одном) при трех значениях коэффициента загрузки сетевого сервера (шлюза): в номинальном режиме работы, при недогрузке и перегрузке. Разработаны инструментальные средства, где гарантированное качество измеряется, как время запаздывания в сетях с коммутацией пакетов.

Ключевые слова: системы массового обслуживания, сети доступа, уровни обслуживания, качество обслуживания QoS, схемы приоритета, среднесетевые задержки.

Введение

Сети доступа – Access Network (AN) находят широкое применение в местных сетях. Они создаются для агрегирования (сбора) разнообразного абонентского трафика на центральную АТС (автоматическую телефонную станцию) в сетях ISDN (Integrated-Service Digital Network – цифровая сеть с интегрированным обслуживанием, ЦСИО) [1]. Наибольшее распространение получила услуга гарантированного качества Triple Play (три в одном) – услуга доведения цифрового потока до каждой абонентской розетки с возможностью одновременного подключения трех цифровых терминалов: цифрового телефона, цифрового телевизора и персонального компьютера.

Чтобы удовлетворять противоречивым по качеству обслуживания требованиям пользователей ЦСИО, выработаны основные положения, определяющие подход к нормированию качества обслуживания QoS (Quality of Service – качество обслуживания), которые изложены в Rec. E.800 ITU-T (http://www.itu.int/ITU-D) [2]. Требования к сетевым показателям качества для служб, основанных на протоколе IP, приведены в Rec. Y.1541 ITU-T [3] как среднесетевые задержки. IP – Internet Protocol – межсетевой протокол. Относится к маршрутизируемым протоколам сетевого уровня семейства TCP/IP. Именно IP стал тем протоколом, который объединил отдельные подсети во всемирную сеть Интернет (Internet). Неотъемлемой частью протокола является адресация сети

В вышеуказанных рекомендациях Международного Союза Электросвязи (МСЭ – ITU-T) даются только требования качества обслуживания, но не указаны пути их осуществления. Поэтому на практике применяют сложные схемы управления трафиком.

Одним из наиболее перспективных путей реализации вышеуказанных рекомендаций является использование схемы приоритета без преимущественного права на прерывание [1, 4]. Она находит широкое применение в шлюзах, интегрирующих различные сегменты сетей и сети доступа, когда агрегируют (собирают) абонентский трафик и поставляют на абонентскую розетку современную услугу Triple Play на базе платформы IMS/NGN (IP Multimedia Subsystem – спецификация передачи мультимедиа в электросвязи на основе протокола IP; Next Generation Network – сеть следующего поколения) [5]. Здесь качество обслуживания пакетов информации также измеряется как среднесетевая задержка, что удачно согласуется и с рекомендациями ITU-T.

Библиотека БГУИР

Удовлетворение противоречивых по качеству обслуживания требований пользователей ЦСИО может быть достигнуто за счет приоритетного обслуживания [1, 6]. При поступлении на единственный сервер пакета с высоким приоритетом обслуживание пакета с более низким приоритетом либо прерывается (абсолютный приоритет), либо пакет с высоким приоритетом становится в начало очереди ожидающих пакетов (относительный приоритет).

Возникает задача расчета вероятносто-временных характеристик (ВВХ) сети доступа трех классов трафика – от цифровых телефонов, цифровых телевизоров и персональных компьютеров (интеграция речевых сигналов, видеоданных и текста). На практике используется дисциплина справедливого обслуживания с использованием относительного приоритета, названного в рамках данной статьи схемой приоритета без преимущественного права на прерывание.

В данной статье автор разработал модель сети доступа с автономным сервером, инструментальные средства для измерения качества обслуживания и при заданных исходных данных исследовал на разработанной модели структурно-сетевые параметры указанной выше схемы приоритета как в номинальном режиме работы, так и осуществил прогноз расчетных показателей при недогрузках и перегрузках экспоненциального и детерминированного серверов.

Теоретический анализ модели

Модель однолинейной СМО с приоритетом в сети доступа представим на рис. 1.

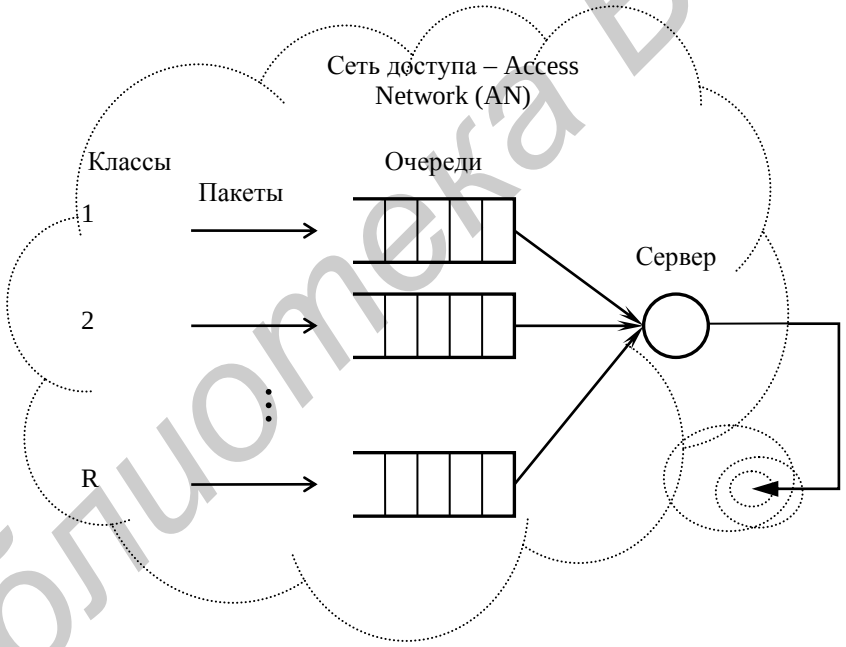

Рис. 1. Модель однолинейной СМО с приоритетом в сети доступа

Поскольку речь о качестве обслуживания можно вести только тогда, когда можно его измерить в конкретных физических величинах, то в нашем случае качество обслуживания необходимо измерять, как задержку пакета информации при его передаче от абонента А к абоненту В в единицах времени, например, в миллисекундах (мс). Следует заметить, что согласно вышеуказанным рекомендациям Международного союза электросвязи сквозное качество обслуживания в сетях с коммутацией пакетов измеряется именно в мс.

Рассмотрим систему единственного сервера, где пакеты класса i, $i = 1, 2, \cdots, R$, поступают в случайном порядке с частотой λ_i с произвольно распределенным временем обслуживания со средним значением h_i и вторым моментом $h_i^{(2)}$. Класс более высокого приоритета задан меньшим значением i.

Библиотека БГУИР

Пусть $\rho_i = \lambda_i h_i, L_i$ и W_i будут поступающей нагрузкой, средним количеством ожидающих пакетов и средним временем ожидания, соответственно, для пакетов класса *i*.

Положим порядок обслуживания пакетов FIFO (First In – First Out – первым пришел, первым и обслужен) с бесконечным буфером в том же самом классе и допустим существование устойчивого состояния. Тогда, если некоторый произвольный пакет в классе *i* выбирается в случайном порядке как тестовый пакет, то качество его обслуживания определится как среднее время ожидания [1]

$$W_i = \frac{\sum_{i=1}^{R} \lambda_i h_i^{(2)}}{2\left(1-\sum_{j=1}^{i-1}\rho_j\right)\left(1-\sum_{j=1}^{i}\rho_j\right)}. \quad (1)$$

Второй момент времени обслуживании $h_i^{(2)}$ в классе *i* в выражении (1) определим [5] следующим образом:

$$h_i^{(2)} = \int_{-\infty}^{\infty} h^2 dF(h), \quad (2)$$

где $F(h)$ – функция распределения вероятности сервисного времени (вероятность того, что время обслуживания пакета не превышает h).

Например, известно [5], что в простейшем случае, когда допускают, что время завершения вызова осуществляется в случайном порядке, то тогда дополнительная функция распределения $H(h)$ (вероятность того, что время обслуживания превышает h) определяется как $H(h) = e^{-\mu h}$. В этом случае говорят, что время обслуживания распределено экспоненциально с величиной μ^{-1}, где μ называют интенсивностью обслуживания. Отсюда искомая функция $F(h)$ определится следующим образом:

$$F(h) = 1 - H(h) = 1 - e^{-\mu h}. \quad (3)$$

Дифференцируя функцию распределения вероятности сервисного времени $F(h)$ (3) по h и подставляя полученную производную как функцию плотности распределения в выражение (2), получим значение второго момента времени обслуживания

$$h_i^{(2)} = \int_{-\infty}^{\infty} h^2 dF(h) = \mu \int_{0}^{\infty} h^2 e^{-\mu h} dh, \quad (4)$$

где последний определенный интеграл табулирован в [6].

Отсюда второй момент времени обслуживании $h_i^{(2)}$ определится как

$$h^{(2)} = 2/\mu^2. \quad (5)$$

Наконец, подставляя выражение (5) в формулу (1) и вводя коэффициент загрузки q единственного сервера, мы получим удобную в инженерных расчетах формулу измерения качества обслуживания как среднего времени ожидания по классам приоритета без преимущественного права на прерывание экспоненциального сервера

$$W_i = \frac{q}{\mu(1-q_{i-1})(1-q_i)}, \quad (6)$$

где $q = \sum_{i=1}^{R} \rho_i$ – коэффициент загрузки единственного сервера; $q_{i-1} = \sum_{j=1}^{i-1} \rho_j$ – поступающая нагрузка по классам высшего приоритета; $q_i = \sum_{j=1}^{i} \rho_j$ – поступающая нагрузка по классам высшего приоритета, включая и класс *i*.

Из полученной нами формулы (6), как инструментального средства измерения качества обслуживания при передаче пакетов, следует, что должна существовать некоторая точка

Библиотека БГУИР

перегиба среднего времени ожидания W_i, когда оно начинает стремительно возрастать, и существует оптимальное значение коэффициента загрузки сервера q_0 в сетях с коммутацией пакетов.

Методика расчета среднесетевых задержек

Разработаем методику расчета структурно-сетевых параметров по формуле (6) вплоть до геометрического подтверждения наличия некоторой рабочей точки с номинальной (тяжелой) загрузкой сервера q_0.

Вновь рассмотрим схему приоритета с тремя классами i, i = 1, 2, 3 и R = 3. Допустим, что номинальная загрузка экспоненциального сервера достигается при $q_0 = 0{,}2$; $\rho_j = \lambda_i / \mu_0 < 1$, а μ_0 – оптимальная интенсивность обслуживания. Тогда из выражения (6) следует:

$$\begin{cases} W_1 = \dfrac{0,2}{\mu_0 (1-\rho_1)^2}; \\ W_2 = \dfrac{0,2}{\mu_0 (1-\rho_1)(1-(\rho_1+\rho_2))}; \\ W_3 = \dfrac{1}{4\mu_0 (1-(\rho_1+\rho_2))}. \end{cases} \quad (7)$$

Когда мы уменьшим интенсивность обслуживания в полтора раза, то во столько же раз увеличится загрузка сервера до q = 0,3; тогда система (7) перепишется как

$$\begin{cases} W_1 = \dfrac{0,45}{\mu_0 (1-1,5\rho_1)^2}; \\ W_2 = \dfrac{0,45}{\mu_0 (1-1,5\rho_1)(1-1,5(\rho_1+\rho_2))}; \\ W_3 = \dfrac{9}{11\mu_0 (1-1,5(\rho_1+\rho_2))}. \end{cases} \quad (8)$$

А когда мы увеличим интенсивность обслуживания в два раза, то во столько же раз уменьшится загрузка сервера до q = 0,1; тогда система (7) перепишется следующим образом:

$$\begin{cases} W_1 = \dfrac{0,05}{\mu_0 (1-0,5\rho_1)^2}; \\ W_2 = \dfrac{0,05}{\mu_0 (1-0,5\rho_1)(1-0,5(\rho_1+\rho_2))}; \\ W_3 = \dfrac{0,05}{0,95\mu_0 (1-0,5(\rho_1+\rho_2))}. \end{cases} \quad (9)$$

Можно показать, что при детерминированном времени обслуживания (например, при использовании технологии ATM (Asynchronous Transfer Mode – асинхронный способ передачи данных) средние времена ожидания по классам приоритета без преимущественного права на прерывание детерминированного сервера сократятся в 2 раза по сравнению с экспоненциальным сервером (6–9).

Численный эксперимент

Пусть речевой сигнал [5], пакетированный АДИКМ (ADPCM – Adaptive Differential Pulse Code Modulation – Адаптивная дифференциальная импульсно кодовая модуляция со

Библиотека БГУИР

скоростью 32 кбит/с) от 1000 абонентов, поступает с частотой λ_1 = 21,96/мс по 1 классу приоритета; видеоданные от 100 абонентов поступают с частотой λ_2 = 12177,81/мс по 2 классу приоритета; текст от 200 абонентов поступает с частотой λ_3 = 1047,8/мс по 3 классу приоритета [5].

Расчет номинальной интенсивности обслуживания μ_0 произведем как

$$\mu_0 = (\lambda_1 + \lambda_2 + \lambda_3) / q_0 . \quad (10)$$

Для коэффициента загрузки единственного сервера q_0 = 0,2 из формулы (10) получим $\mu_0 = (21{,}96 + 12177{,}81 + 1047{,}8) / 0{,}2$ мс = 66 237,85/мс.

Переведем размерность интенсивности обслуживания из размерности пакет/с в размерность бит/с, т.е. произведем оценку оптимальной производительности сервера υ_0:

$$\upsilon_0 = l\mu_0 , \quad (11)$$

где l – размер пакета в бит.

С учетом того, что осуществлена процедура сборки разборки пакетов по технологии Ethernet с минимальным размером пакета l = 64 байта, получим из выражения (11) оценку оптимальной производительности сервера как

υ_0 = 64 байта × 8 бит × 66 237.85/мс = 33 991 379,2 бит/мс ≈ 34 Гбит/с.

Выбираем технологию передачи 10 GigE (Gigabit Ethernet) с последующим уплотнением по технологии DWDM (Dense WDM – Спектральное уплотнение каналов; WDM – Wavelength-Division Multiplexing – мультиплексирование с разделением по длине волны – технология, позволяющая одновременно передавать несколько информационных каналов по одному оптическому волокну на разных несущих частотах) (4 лямбды × 10 Гбит/с = 40 Гбит/с). Новый стандарт 10-гигабитного Ethernet включает в себя семь стандартов физической среды для LAN, MAN и WAN. В настоящее время он описывается поправкой IEEE 802.3ae и должен войти в следующую ревизию стандарта IEEE 802.3. Отсюда выберем за основу производительность системы передачи μ = 10 Гбит/с.

Подставляя исходные данные в формулы (7 – 9), представим результаты расчетов среднего времени ожидания в зависимости от коэффициента загрузки однолинейного сервера в таблице.

Результаты расчетов численного эксперимента

Среднее время ожидания	Коэффициент загрузки сервера, q		
	0,1	0,2	0,3
W_1, пс (речь)	1,47	5,9	13,23
W_2, пс (видеоданные)	1,61	7,2	18,28
W_3, пс (текст)	1,7	9	33,24

Из таблицы следует, что с ростом телефонной нагрузки растет и среднесетевая задержка. Наибольшую задержку, как и следовало ожидать, испытывает третий класс приоритета (текст), меньшую задержку испытывает второй класс приоритета (видеоданные), а наименьшую задержку испытывает первый класс приоритета (речевой сигнал, пакетированный АДИКМ).

Среднесетевые задержки по классам приоритета при коэффициенте загрузки сервера q = 0,1 изменяются незначительно, а уже при q = 0,3 разброс задержек начинает стремительно возрастать. Отметим, что среднее время ожидания при использовании однолинейного детерминированного сервера (технология АТМ) будет в 2 меньше.

Для более тонкого анализа представим результаты расчетов в графическом виде на рис. 2, где более наглядно просматриваются сравнительные характеристики различных классов приоритета.

Библиотека БГУИР

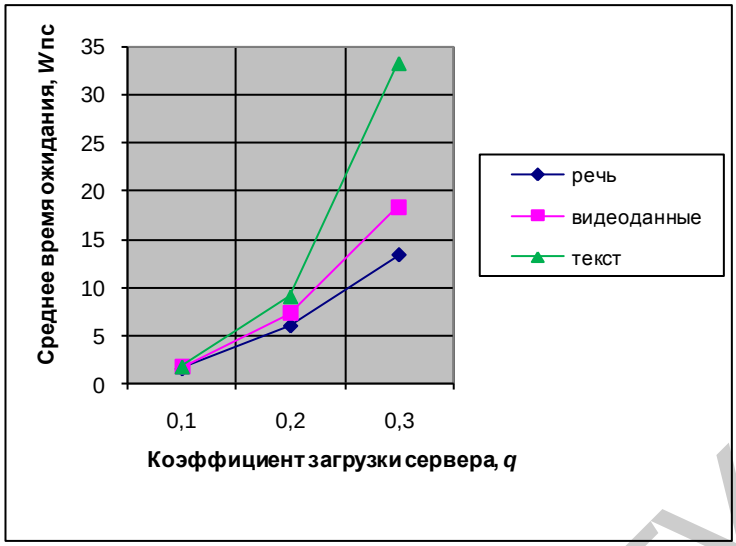

Рис. 2. Среднее время ожидания однолинейного экспоненциального сервера

Заметим, что график дает завышенные задержки, поскольку он рассчитан для оптимальной интенсивности обслуживания μ_0. На самом деле выбранные нами продуктовые линейки 10 GigE обеспечат меньшие задержки за счет более высокой производительности по сравнению с расчетным значением υ_0 [4].

Результаты и их обсуждение

Заметим также и то, что, как и ожидалось, на рис. 2 имеются некоторые точки приемлемой разбалансировки по уровням обслуживания при коэффициенте загрузки сервера $q = 0{,}2$. И если полагать, что такая загрузка сервера будет номинальной, то заданный режим работы схемы приоритета 3 классов трафика обеспечит качество обслуживания и пульсирующего IP-трафика.

Таким образом, в некоторых случаях схема приоритета может работать более эффективно, чем более сложная система управления трафиком. Однако эта эффективность достигается при меньшей загрузке однолинейного сервера. А уже при коэффициенте $q > 0{,}3$ схема приоритета теряет свои преимущества по сравнению с системой управления трафиком.

Становится очевидным, что хрупкая грань применения той или иной технологии обработки трафика должна решаться в ходе тщательного сопоставления технико-экономического анализа обоих сетей доступа по вышеуказанным схемам, если коэффициент загрузки сервера начинает превышать пороговую величину $q = 0{,}3$.

Заключение

Разработана инженерная методика расчета среднего времени ожидания по классам приоритета без преимущественного права на прерывание экспоненциального сервера как инструментальная база измерения качества обслуживания в сетях с коммутацией пакетов.

Установлено, что на сетях доступа с доминирующей услугой гарантированного качества Triple Play (три в одном), простая схема приоритета обладает преимуществом перед сложной системой управления трафиком при коэффициенте загрузки экспоненциального сервера $q < 0{,}3$.

Получена упрощенная формула расчета среднесетевых задержек для схемы с 3 классами приоритета. Выполнен численный эксперимент, который подтвердил наличие некоторого номинального режима работы экспоненциального сервера со схемой с 3 классами приоритета.

Установлены методом вычислительного эксперимента точки приемлемой разбалансировки по уровням обслуживания в зависимости от среднего времени ожидания от

Библиотека БГУИР

коэффициента загрузки экспоненциального сервера при q = 0,2, которые и легли в основу создания номинального режима работы экспоненциального сервера со схемой с 3 классами приоритета.

ACCESS NETWORKS WITH THE RELATIVE PRIORITY

A.G. KOSTUKOVSKY

Abstract

The engineering method of calculation of an average waiting time on priority classes a non-preemptive priority model of the exponential server as tool base of measurement of quality of service in networks with switching of packages is developed. It is established that on access networks with the dominating service Triple Play (Three in one) the simple scheme of a priority has advantage over a complex system of traffic control at an exponential server load factor $q < 0.3$. The simplified formula of calculation the average network latency for the scheme with 3 classes of a priority is received. Numerical experiment which confirmed existence of some nominal operating mode of the exponential server with the scheme with 3 classes of a priority is executed. The inflection point in dependence of an average waiting time on coefficient of loading of the exponential server is established at $q = 0.2$ which laid down in a basis of creation of a nominal operating mode of the exponential server with the scheme with 3 classes of a priority.

Список литературы

1. *Лохмотко В.В., Пирогов К.И.* Анализ и оптимизация цифровых сетей интегрального обслуживания. Минск, 1991.
2. P.800: Methods for subjective determination of transmission quality. [Электронный ресурс]. – Режим доступа : https://www.itu.int/rec/T-REC-P.800-199608-I/en. – Дата доступа: 01.11.2013.
3. ITU Y.1541: Network performance objectives for IP-based services. [Электронный ресурс]. – Режим доступа : https://www.itu.int/rec/T-REC-Y.1541/en. – Дата доступа: 01.11.2013.
4. *Akimaru H, Kawashima K.* Teletraffic: theory and applications. London, 1999.
5. *Двайт Г.Б.* Таблицы интегралов и другие математические формулы. М., 1973.
6. *Костюковский А.Г.* // Матер. XVIII Междунар. науч.-техн. конф. «Современные средства связи», Минск, 15–16 октября 2013 г. С. 63–66.

Библиотека БГУИР