УДК 004.312.24

ФОРМИРОВАНИЕ ОЧЕРЕДЕЙ В ПАКЕТНЫХ СИСТЕМАХ МАССОВОГО ОБСЛУЖИВАНИЯ

Б.Я. Лихтциндер, Л.Б. Иванова

Поволжский государственный университет телекоммуникаций и информатики
Россия, 443010, г. Самара, ул. Льва Толстого, 23

E-mail: lixt@psati.ru

Рассматриваются системы массового обслуживания (СМО) с пачечными потоками заявок, характерными для современных мультисервисных сетей связи. Приведено обобщение формулы Хинчина – Поллячека на системы с потоками общего вида. Рассмотрены зависимости средних значений размера очереди от загрузки, при малых загрузках системы. Вводится понятие условного среднего значения размера очереди, при условии отсутствия интервалов простоя процессора.

Ключевые слова: *системы массового обслуживания (СМО), пачечные потоки, размеры очередей, мультисервисные сети, условное среднее, максимальное значение очередей, коэффициент загрузки.*

Введение

Процесс формирования очередей заявок в СМО подробно и широко рассмотрен в литературе [1, 4].

Формирование очередей в СМО с пакетным пачечным трафиком рассмотрено нами в [2]. Анализ проводился для стационарных потоков интервальными методами, основанными на определении чисел заявок (пакетов), поступающих в систему, в течение интервала времени τ обработки одной заявки [2, 3].

В этих работах показано, что среднее значение размера очередей для однопоточных СМО определяется величиной коэффициента загрузки чисел заявок $\rho = \lambda\tau$ и дисперсией $D_n(\rho)$ чисел заявок $n_i(\tau)$, поступающих в течение интервала τ, соответствующего данному коэффициенту загрузки ρ. Здесь λ – средняя интенсивность потока заявок. Показано, что среднее число заявок $\overline{n(\tau)}$ всегда равно коэффициенту загрузки ρ. Получено обобщение формулы Хинчина – Поллячека [2, 5, 6]:

$$\overline{q(\rho)} = \frac{D_n(\rho) + 2\sum_{j=1}^{K} Co\,\mathrm{v}[n_i(\rho); n_{i-j}(\rho)] - \rho(1-\rho)}{2(1-\rho)} = \frac{\Phi_n(\rho)\rho}{2(1-\rho)}, \quad (1)$$

где $\overline{q(\rho)}$ – среднее значение длины очереди, определенное на всем промежутке времени наблюдения, с учетом интервалов времени простоя процессора;

$D_n(\rho)$ – дисперсия чисел заявок на интервале обслуживания τ, соответ-

---

Борис Яковлевич Лихтциндер (д.т.н., профессор), профессор кафедры «Мультисервисные системы и информационная безопасность».

Людмила Борисовна Иванова (к.т.н., доцент.), доцент кафедры «Экономика и организация производства».

ствующем коэффициенту загрузки $\rho = \lambda\tau$; $Cov[n_i(\rho); n_{i-j}(\rho)]$ – коэффициент ковариации и $n_{i-j}(\rho)$; K – интервал корреляции.

Потоки пакетов трафика современных мультисервисных сетей доступа носят явно выраженный пачечный характер [2]. Поэтому существует весьма значительное число временных интервалов, в течение которых заявки не поступают.

На рис. 1, *а* показаны количества пакетов видеотрафика, поступающих в течение постоянных интервалов времени τ, которые соответствуют различным значениям коэффициентов загрузки.

ρ ($\rho = 0{,}1$ – черные линии, $\rho = 0{,}7$ – серые линии).

На рис. 1, *б* показан фрагмент этого же трафика в увеличенном масштабе времени. В качестве единицы времени здесь выбран постоянный интервал τ времени обработки одной заявки.

На рис. 2, *а* показан процесс образования очередей в потоке пакетов рассмотренного видеотрафика при различных значениях коэффициентов загрузки ρ ($\rho = 0{,}1$ – черные линии, $\rho = 0{,}7$ – серые линии). На рис. 2, *б* показан фрагмент этого же трафика в увеличенном масштабе времени.

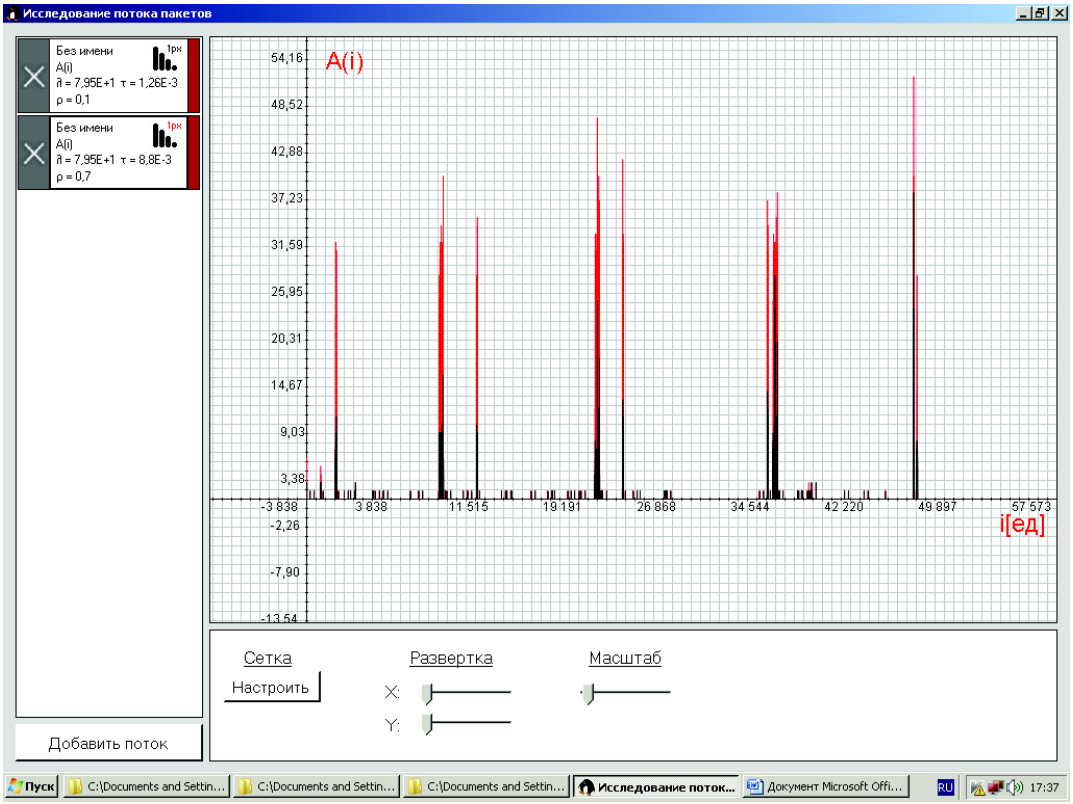

Рис. 1, *а*. Количества пакетов видеотрафика, поступающих в течение постоянных интервалов времени τ

Из графиков следует, что в периоды активного поступления пакетов очереди быстро возрастают. При малых значениях коэффициента загрузки (интервал времени обработки τ мал) очередь быстро уменьшается, достигает нулевого значения, и все остальное время процессор простаивает. При большой загрузке (интервал времени обработки τ велик) очередь уменьшается медленно и процессор простаивает меньше.

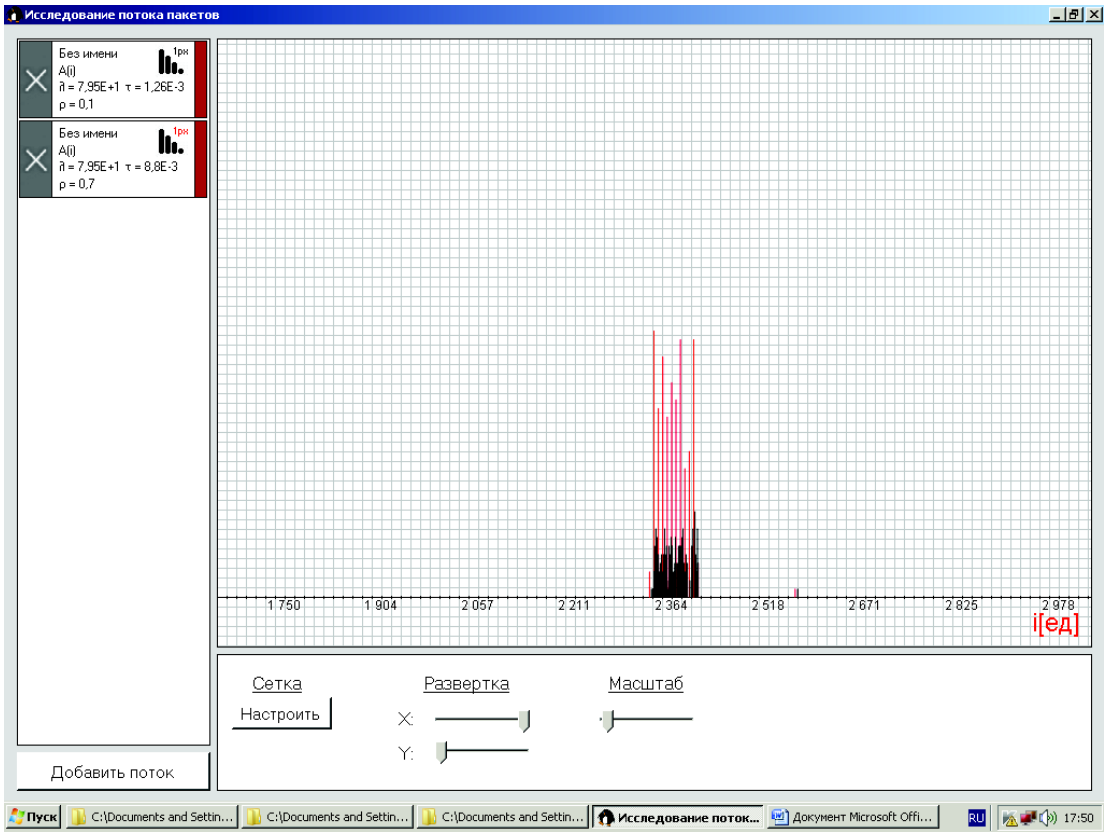

Рис. 1, *б*. Фрагмент количества пакетов видеотрафика, поступающих в течение постоянных интервалов времени τ (масштаб времени увеличен)

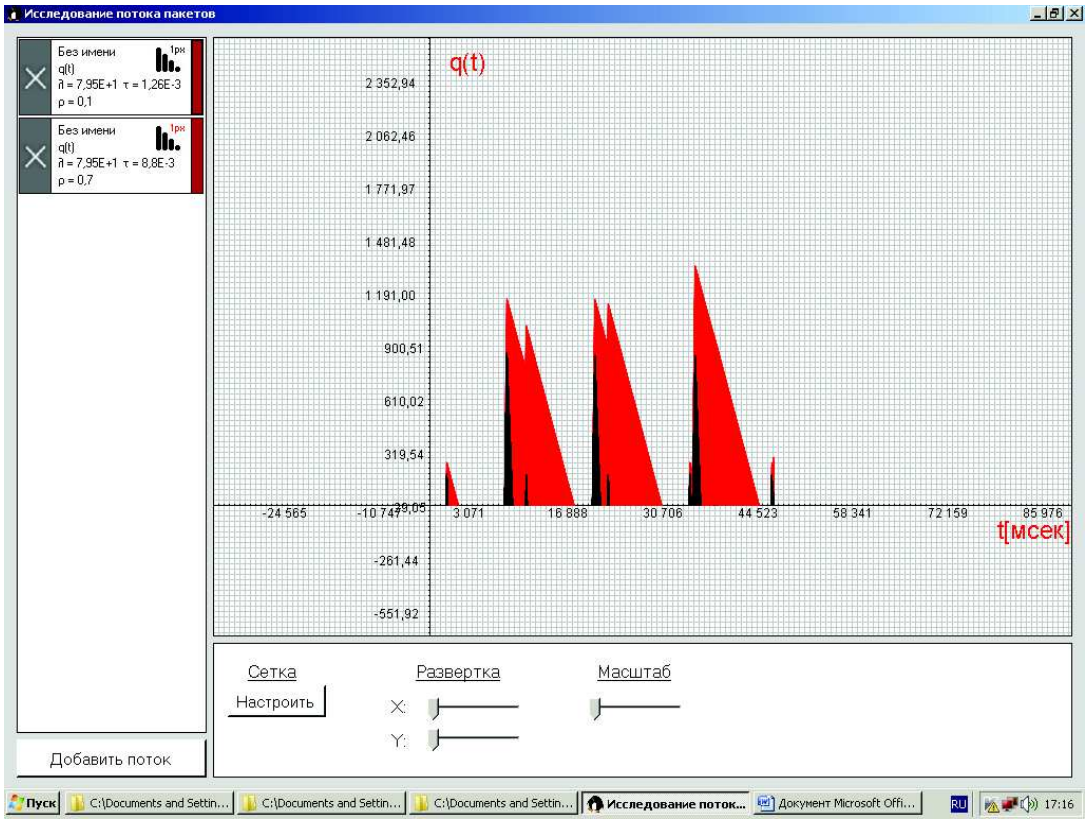

Рис. 2, *а*. Образование очередей в потоке пакетов реального видеотрафика

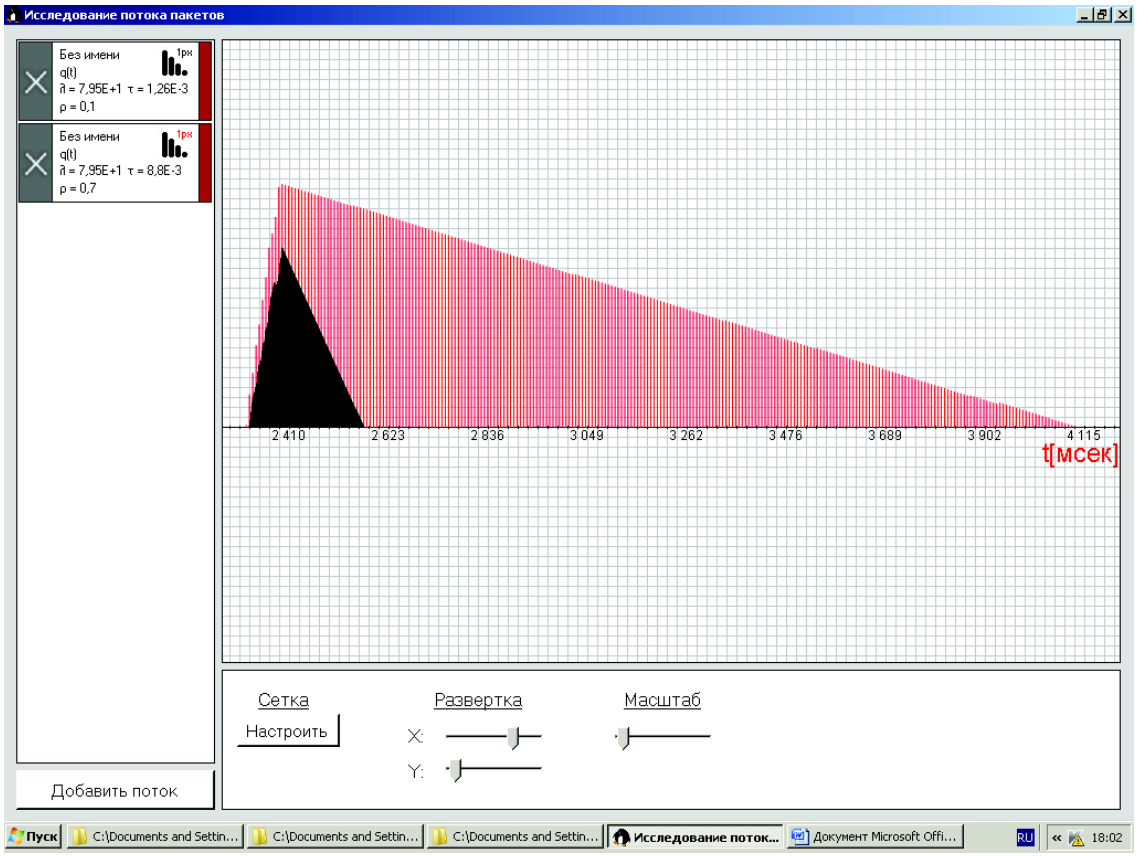

Рис. 2, *б*. Фрагмент образования очередей в потоке пакетов реального видеотрафика (масштаб времени увеличен)

Очевидно, что при малых загрузках средние значения чисел заявок в очереди $\overline{q(\rho)}$ весьма малы. Из рис. 2, *б* видно, что даже при весьма значительных загрузках график изменения очереди имеет треугольную форму.

Процесс формирования очереди можно разделить на три этапа: этап нарастания, этап уменьшения и этап полного отсутствия очереди, когда процессор простаивает.

Формирование очереди

На рис. 3 приведен график формирования очереди при пачечном потоке.

В рассматриваемом примере показаны пакеты, поступающие в период цикла T_c в течение промежутка времени T_{pi}, пачками, размером $N_{pi} = 12$, с максимальной интенсивностью $\lambda_{max} = \dfrac{1}{\Delta T}$ (этап нарастания очереди).

Примем, что на всем интервале времени наблюдения максимальная интенсивность λ_{max} поступления пакетов (заявок) и время τ обработки одной заявки остаются постоянными.

В течение времени обработки τ в систему каждый раз поступает $n(\tau) = \lambda_{max}\tau$ заявок (в рассматриваемом примере – три заявки). Примем, что на промежутке времени T_{pi} укладывается целое число k_i интервалов τ (в рассматриваемом примере – 4 интервала):

$$k_i = \frac{T_{pi}}{\tau} = \frac{N_{pi}}{n(\tau)} = 4.$$

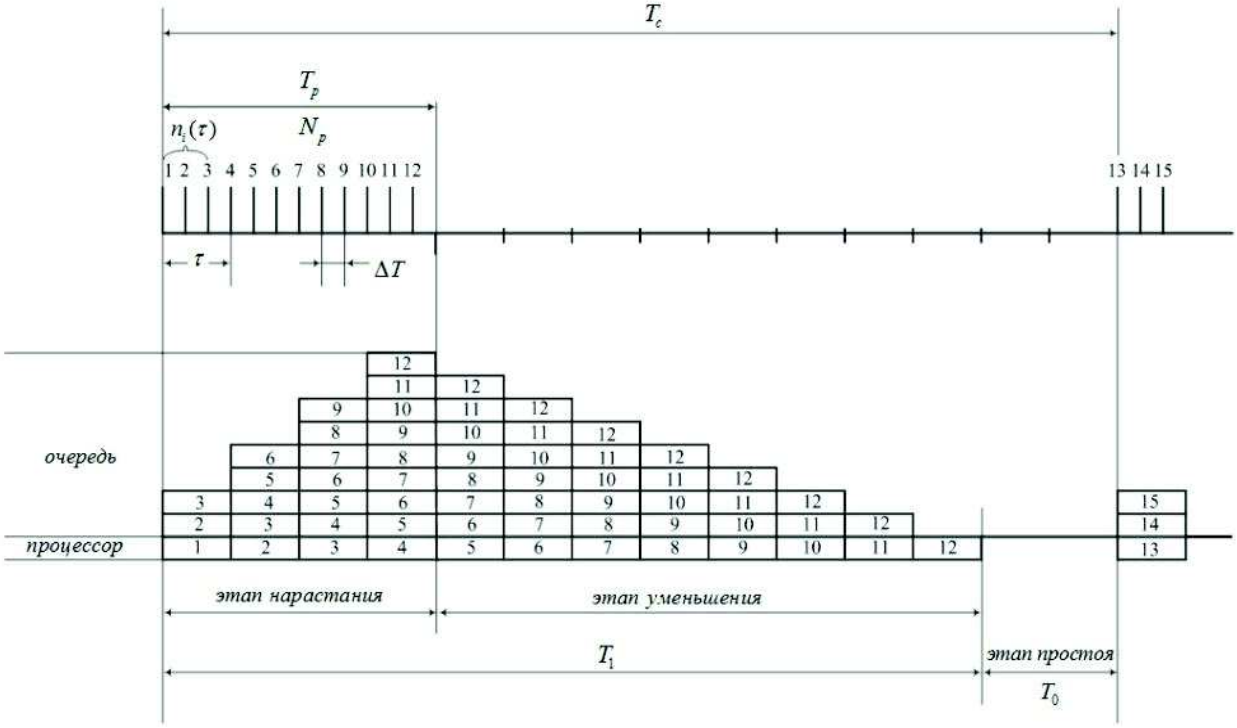

Рис. 3. График формирования очереди при пачечном потоке

На этапе уменьшения в течение каждого интервала времени τ происходит убывание очереди на одну заявку. Длительность этапа простоя T_0 определяется принятым периодом T_c, после окончания которого в систему поступает очередная пачка заявок.

Обозначим максимальный размер очереди, образованной i-й пачкой заявок, $h_{\max i}(\tau)$ (на рис. 3 показана одна пачка).

Нетрудно доказать, что максимальный размер определяется соотношением

$$h_{\max i}(\tau) = N_{pi}\frac{n(\tau)-1}{n(\tau)} = N_{pi}(1-\frac{1}{\lambda_{max}\tau}) = 8. \qquad (2)$$

На рис. 2, *б* хорошо заметно, что с увеличением нагрузки увеличивается максимальное значение очереди. Естественно, что эта формула имеет ограничение $n(\tau) > 1$, что соответствует $\lambda_{max}\tau > 1$. В противном случае в течение интервала обслуживания τ будет поступать не более одной заявки и очередь отсутствует.

Нетрудно показать, что в случае появления одной пачки заявок в течение цикла T_c общее число заявок, которые находились в очередях (суммарная площадь треугольника),

$$S_i(\tau) = \frac{1}{2}N_{pi}{}^2\frac{n(\tau)-1}{n(\tau)} = 48.$$

Если предположить, что в течение каждого из циклов появляется несколько пачек, но каждая из последующих пачек появляется после окончания обработки всех заявок предыдущей пачки, и размеры пачек взаимно независимы, то общее

число заявок, находившихся в очередях в течение всего периода наблюдения T, определяется соотношением

$$S(\tau) = \frac{1}{2}(1 - \frac{1}{\lambda_{max}\tau}) \cdot \sum_{i=1}^{K} N_{pi}{}^2 ,$$

где K – общее число пачек на всем периоде наблюдения (в рассматриваемом примере $K = 1$ и $S(\tau) = 48$)).

Среднемаксимальное значение очереди

Введем в рассмотрение понятие «среднемаксимальное значение» очередей $\overline{h_{\max}(\tau)}$ (среднее из всех максимальных значений). Обозначим также среднее число заявок в пачке через $\overline{N_p}$:

$$\overline{N_p} = \frac{\sum_{i=1}^{K} N_{pi}}{K} .$$

Поскольку каждой пачке заявок соответствует максимальное значение $\overline{h_{\max}(\tau)}$, определяемое соотношением (2), а всего на интервале рассмотрения имеется K пачек, то

$$\overline{h_{\max}(\tau)} = \frac{1}{K}\sum_{i=1}^{K} h_{\max i}(\tau) = \frac{1}{K}\sum_{i=1}^{K} N_{pi} \cdot \frac{n(\tau)-1}{n(\tau)} = \overline{N_p} \cdot (1 - \frac{1}{\lambda_{max}\tau}) . \qquad (3)$$

Введем понятие условного среднего значения размера очереди $\overline{Q(\tau)}$, которое представляет среднее значение очереди на каждом из интервалов τ, соответствующем условию активной загрузки процессора (процессор не простаивает):

$$\overline{Q(\tau)} = \frac{S(\tau)}{\sum_{i=1}^{K} N_{pi}} = \frac{\sum_{i=1}^{K} N_{pi}{}^2 (1 - \frac{1}{\lambda_{max}\tau})}{2\sum_{i=1}^{K} N_{pi}} = \frac{\overline{N_p{}^2}}{2\overline{N_p}}(1 - \frac{1}{\lambda_{max}\tau}) , \qquad (4)$$

где $\overline{N_p{}^2}$ – второй начальный момент чисел заявок в пачках (в рассматриваемом примере $K = 1$ и $\overline{Q(\tau)} = 4$)).

Сравнивая (3) и (4), получаем зависимость

$$\overline{h_{\max}(\tau)} = 2\overline{Q(\tau)}\frac{\overline{N_p}^2}{\overline{N_p{}^2}} = \frac{2\overline{Q(\tau)}}{(1 + \nu_{N_p}^2)} , \qquad (5)$$

где $\nu_{N_p}^2$ – коэффициент вариации чисел заявок и пачках.

Если все пачки одинаковые ($\nu_{N_p}^2 = 0$), то $\overline{h_{\max}(\tau)}$ имеют наибольшее значение $\overline{h_{\max}(\tau)} = 2\overline{Q(\tau)}$.

Подставляя из (4) в (5), получим

$$\overline{h_{\max}(\tau)} = \overline{N_p} \cdot (1 - \frac{1}{\lambda_{max}\tau}) = \overline{N_p} \cdot (1 - \frac{\lambda}{\lambda_{max}\rho}) , \text{ при } \rho > \frac{\lambda}{\lambda_{max}} .$$

Подобный результат не является неожиданным, поскольку случайные величины $h_{\max i}(\tau)$ и N_{pi} в (2) отличаются только постоянным коэффициентом $(1-\frac{1}{\lambda_{max}\tau})$.

Среднее число заявок $\overline{q(\rho)}$, находящихся в очереди, с учетом наличия интервалов простоя процессора:

$$\overline{q(\rho)} = \overline{Q(\rho)} \cdot \rho = \frac{\overline{N_p^{\,2}}}{2\overline{N_p}}(1-\frac{1}{\lambda_{max}\tau})\rho = \frac{1}{2}\overline{h_{\max}(\rho)}\cdot(1+\nu_{N_p}^2)\rho .$$

Указанные соотношения определяют среднее значение очереди при условии, что очередная пачка не может появиться до окончания обработки предыдущей пачки.

Однако это условие выполняется далеко не всегда.

На рис. 2, *а* видно, что при коэффициенте загрузки $\rho = 0,7$ происходит наложение очередей от соседних пачек.

В работах [2, 3] показано, что расположение взаимно независимых пачек не оказывает влияния на числитель формулы (1), а учитывается изменением коэффициента загрузки ρ в знаменателе. Таким образом, с учетом возможного наложения обобщенная формула для определения средних значений очередей в одноприборных СМО с рассмотренными пачечными потоками примет вид

$$\overline{q(\rho)} = \frac{\overline{N_p^{\,2}}}{\overline{N_p}} \cdot \frac{(\rho - \frac{\lambda}{\lambda_{max}})}{2(1-\rho)} = \frac{\overline{h_{\max}(\rho)}\cdot(1+\nu_{N_p}^2)\rho}{2(1-\rho)}. \qquad (6)$$

При этом

$$\overline{Q(\rho)} = \frac{\overline{N_p^{\,2}}}{\overline{N_p}} \cdot \frac{(1 - \frac{\lambda}{\lambda_{max}\rho})}{2(1-\rho)} = \frac{\overline{h_{\max}(\rho)}\cdot(1+\nu_{N_p}^2)}{2(1-\rho)}. \qquad (7)$$

На рис. 4 для рассмотренного выше видеотрафика показаны количества пакетов, поступающих в течение постоянных интервалов времени τ, которые соответствуют коэффициенту загрузки $\rho = 0,1$ (черные линии).

Для сравнения здесь же показаны количества пакетов пуассоновского потока, полученные при той же загрузке (серые линии).

Из рис. 4 следует, что по сравнению с видеотрафиком пуассоновский поток носит весьма равномерный характер.

Для пуассоновского потока при постоянном времени обслуживания условное среднее значение очереди $\overline{Q(\rho)}$ определятся на основании известной формулы Хинчина – Поллячека [4]:

$$\overline{Q(\rho)} = \frac{\overline{q(\rho)}}{\rho} = \frac{\rho}{2(1-\rho)}.$$

Анализ показывает, что абсолютная разница среднего и условного среднего значений, полученных по этим формулам для пуассоновских потоков, не превышает 0,25 заявки. Для потока видеотрафика указанная разница становится весьма значительной. На рис. 5 для рассмотренного выше реального видеотрафика пока-

зана зависимость $\overline{q(\rho)}$ в диапазоне изменения коэффициента загрузки ρ от 0 до 0, 05.

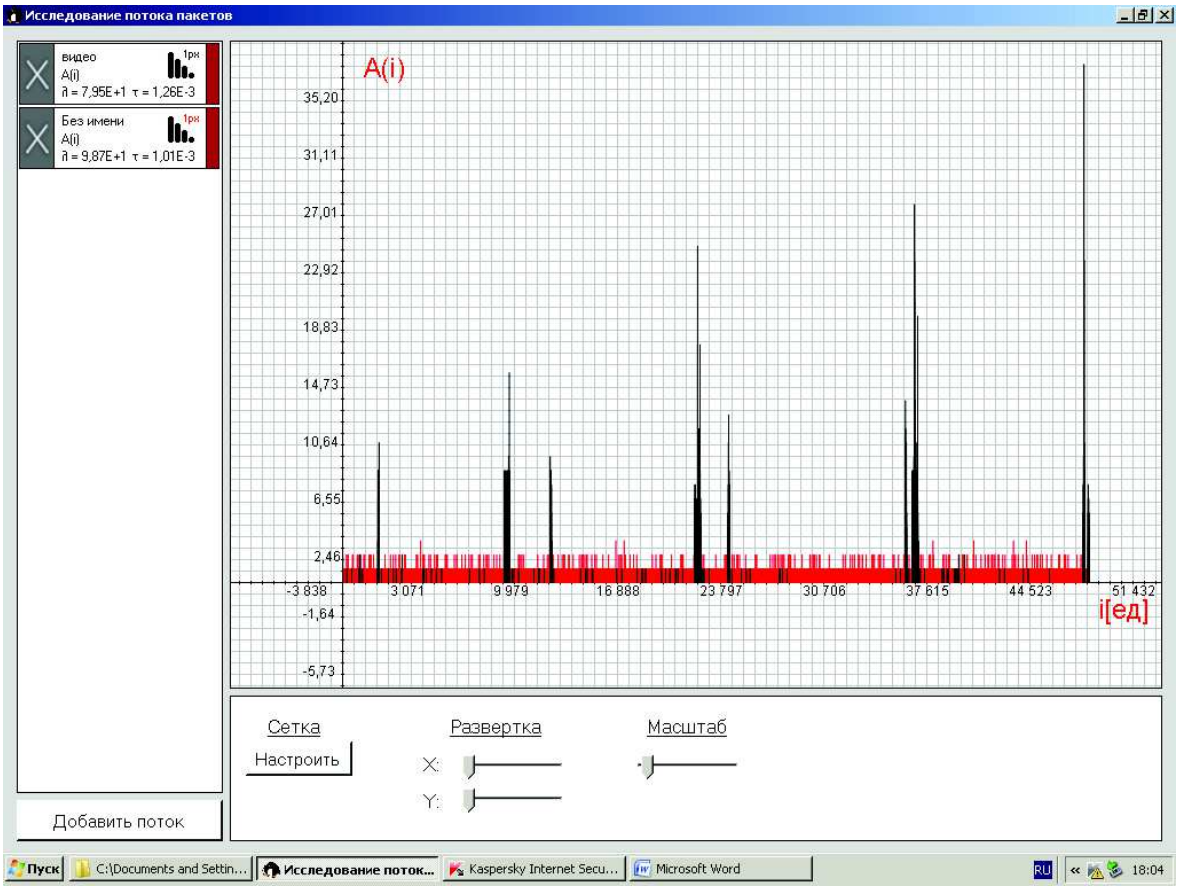

Рис. 4. Количества пакетов, поступающих в течение интервалов времени τ, которые соответствуют коэффициенту загрузки $\rho = 0,1$, для видеотрафика и пуассоновского потока

Из графика следует, что коэффициент загрузки $\rho = 0,031$ соответствует среднему значению очереди $\overline{q(\rho)} = 1,004$ (показано в верхней части рисунка). Условное среднее число пакетов в очереди при этом составляет $\overline{Q(\rho)} = 32,2$. Среднее от максимальных значений очередей $\overline{h_{\max}(\tau)}$, определенное соотношением (5), почти в 60 раз превышает среднее значение очереди $\overline{q(\rho)}$. Таким образом, среднее значение плохо отражает реальное состояние очередей; для характеристики размеров очередей следует ориентироваться на условные средние значения чисел пакетов в очереди $\overline{Q(\rho)}$.

На рис. 6 для рассмотренного выше реального видеотрафика показана зависимость $\overline{Q(\rho)}$ в диапазоне изменения коэффициента загрузки ρ от 0 до 0,05.

Так как согласно (7) при малых значениях ρ

$$\overline{Q(\rho)} \approx \overline{h_{\max}(\rho)} \cdot \frac{(1 + \nu_{N_p}^2)}{2},$$

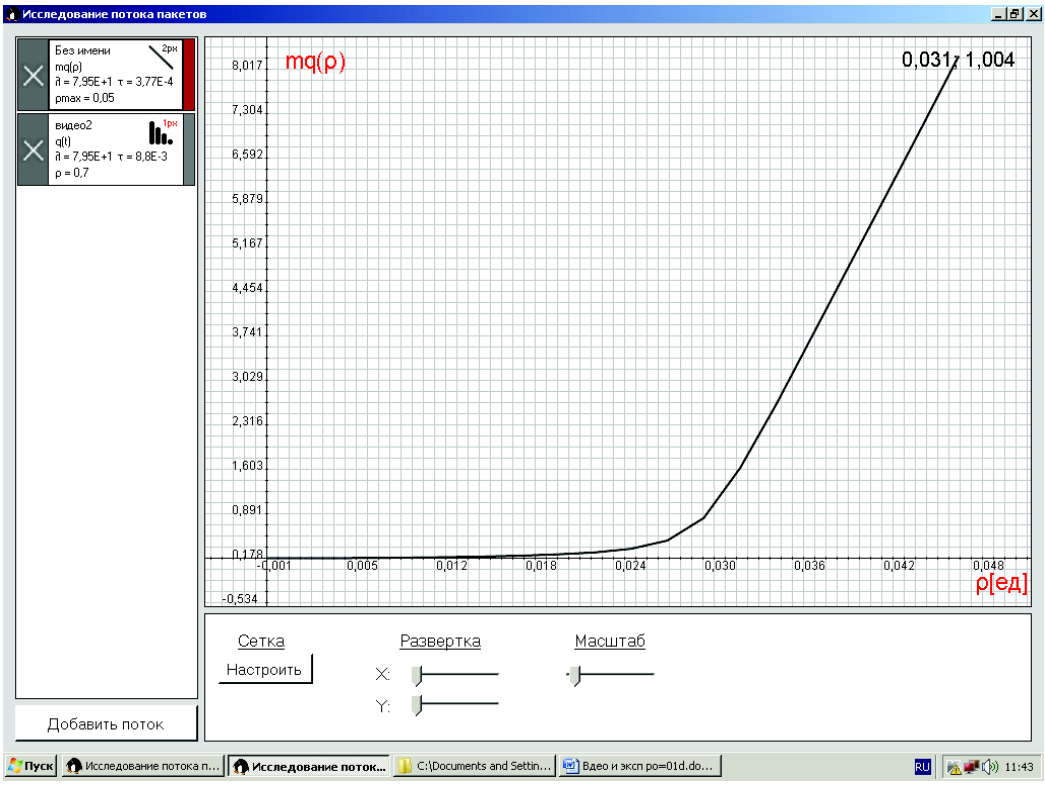

Рис. 5. Зависимость $\overline{q(\rho)}$ для реального видеотрафика
в диапазоне изменения коэффициента загрузки ρ от 0 до 0, 05

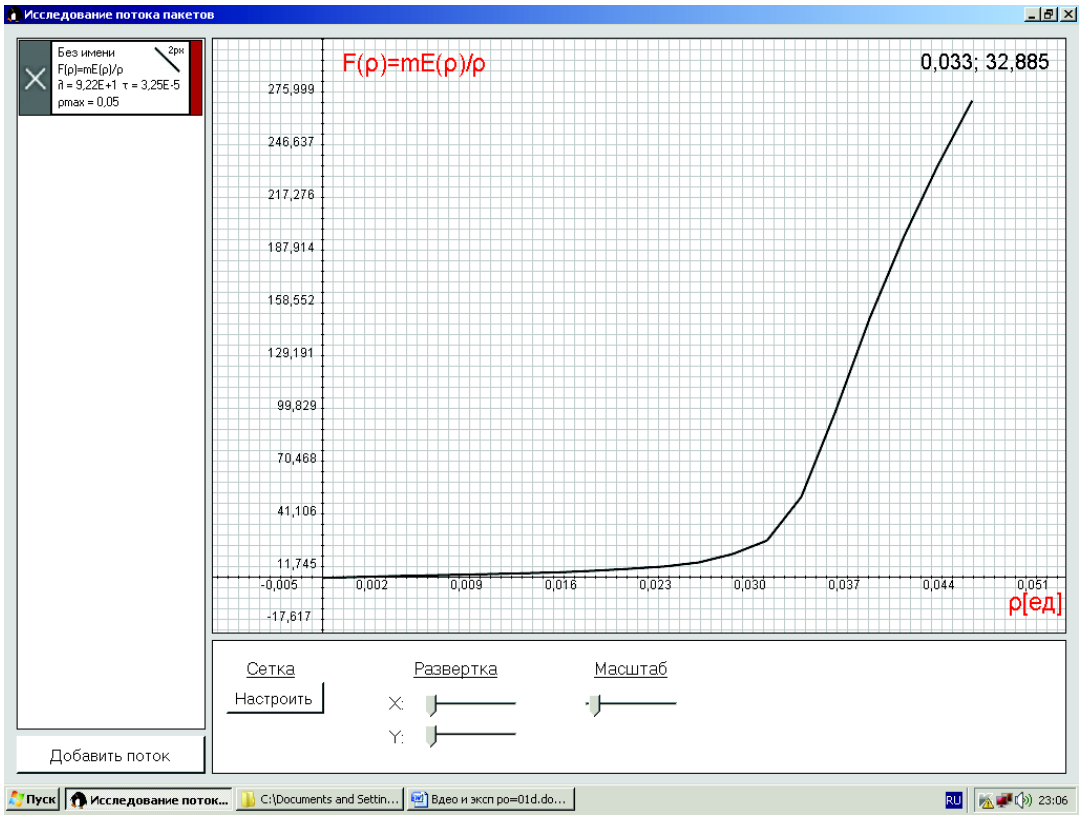

Рис. 6. Зависимость $\overline{Q(\rho)}$ в диапазоне изменения коэффициента
загрузки ρ от 0 до 0,05

для трафика с экспоненциальным распределением длительностей пачек ($\nu_{N_p}^2 = 1$) при малой загрузке $\overline{h_{\max}(\rho)} \approx \overline{Q(\rho)}$.

Таким образом, условное среднее число пакетов $Q(\rho)$ полностью определяет среднее от максимальных значений размеров очереди.

Заключение

Для СМО с пачечными потоками среднее значение очереди с учетом интервалов простоя процессора плохо отражает реальную картину размера очередей и не может непосредственно использоваться для определения размеров буферной памяти. Условное среднее число пакетов $Q(\rho)$ более чем в 30 раз превышает среднее значение и полностью характеризует максимальные значения размеров очередей, которые и определяют требуемые предельные значения объема буферной памяти.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Степанов С.Н.* Теория телетрафика. Концепции, модели, приложения. – М.: Горячая линия Телеком, 2015. – 808 с.: ил.
2. *Лихтциндер Б.Я.* Интервальный метод анализа трафика мультисервисных сетей // Приложение к журналу ИКТ Модели инфокоммуникационных систем: разработка и применение. – Вып. 8. – Самара, 2011. – С. 101–152.
3. *Лихтциндер Б.Я.* Интервальный метод анализа трафика мультисервисных сетей доступа. – Самара: ПГУТИ, 2015. – 121 с.: ил.
4. *Клейнрок Л.* Вычислительные системы с очередями. Т. 2. Пер. с англ. – М.: Мир, 1979.
5. *Лихтциндер Б.Я.* О некоторых обобщениях формулы Хинчина – Поллячека // Инфокоммуникационные технологии. – 2007. – Т. 5. – № 4. – С. 253–258.
6. *Лихтциндер Б.Я.* Интервальный метод анализа мультисервисного трафика сетей доступа // Электросвязь. – 2015. – № 12. – С. 52–54.

Статья поступила в редакцию 21 сентября 2016 г.

QUEUES MODELING IN BATCH QUEUING SYSTEMS (QMS)

B.Ya. Lichtcinder, L.B. Ivanova

Povolzhskiy State University of Telecommunications and Informatics
23, Lev Tolstoy st., Samara, 443010, Russian Federation

The paper describes queuing systems with batch request flows typical for modern multiservice telecommunication networks. It contains generalization of Pollaczek-Khinchin formula for systems with common type flows. The dependences of queues average value at low duty ratio are examined. Full absence of any queues because of a minimal time interval between nearby requests takes place. The term of conditional average queue size that is an average value with the absence of processor downtime intervals is given.

Keywords: *conceptual model, resources, elements, links, corporative development.*

Boris Ya. Lichtcinder (Dr. Sci. (Techn.)), Professor.
Lyudmila B. Ivanova (Ph.D. (Techn.)), Associate Professor.