УДК 519.872:621.394.74

А.З. МЕЛИКОВ, Л.А. ПОНОМАРЕНКО, ЧИ СОН КИМ

ЧИСЛЕННЫЙ МЕТОД АНАЛИЗА МОДЕЛЕЙ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ СО СКАЧКООБРАЗНЫМИ ПРИОРИТЕТАМИ[1]

Ключевые слова: *система массового обслуживания, скачкообразный приоритет, фазовое укрупнение состояний цепей Маркова, численный анализ.*

ВВЕДЕНИЕ

Среди многообразных типов моделей систем массового обслуживания (СМО) с приоритетами наиболее изучены модели с очередями, в которых используются так называемые статические приоритеты, устанавливаемые до начала работы системы и не изменяющиеся со временем. Каждый трафик имеет свой приоритет, и в момент освобождения канала выбирается вызов из начала очереди с наивысшим приоритетом среди непустых очередей (в англоязычной литературе — HOL-приоритеты, Head-of-Line). В теоретическом плане такие приоритеты в различных моделях СМО исследуются давно (подробная информация в этом направлении приведена в [1, 2]).

Известно, что HOL-приоритеты не всегда позволяют удовлетворять противоречивые требования разнотипных вызовов в СМО. Поэтому необходимо использовать различные динамические по времени [3] и состояниям [4] приоритеты. В последнее десятилетие интенсивно исследуется новый тип приоритетов — множественные приоритеты [5–8]: когда вызовы реального времени имеют высокие временные и низкие пространственные приоритеты, а нереального — низкие временные и высокие пространственные приоритеты.

В работе [9] введены новые типы HOL-приоритетов, имеющие скачкообразный характер (Head-of-Line with Priority Jumps, HOL-PJ. Основная идея состоит в том, что если время ожидания вызова, стоящего в начале некоторой очереди, достигает определенной величины, то он переходит в соседнюю очередь с более высоким приоритетом. Этот процесс продолжается до тех пор, пока вызов любого типа либо получит доступ в канал для обслуживания, либо достигнет очереди с наивысшим приоритетом. В указанной работе приведены формулы для расчета среднего времени ожидания разнотипных вызовов.

В [10–14] предложены различные виды HOL-PJ для системы обслуживания с дискретным временем (разделенным на слоты) с двумя типами вызовов: высокого приоритета (H-вызовы) и низкого приоритета (L-вызовы). Разработаны формулы для производящих функций длины очереди вызовов обоих типов и времени ожидания H-вызовов, а также для их моментов. В [14] приведен подробный обзор работ, посвященных СМО со скачкообразными приоритетами.

При рассмотрении моделей СМО со скачкообразными приоритетами, как правило, предполагается, что в них имеются буферы неограниченного размера. Однако такие модели не могут использоваться для анализа реальных сетей коммуникации, так как в них буферные накопители всегда ограничены. В настоящей статье изучаются модели СМО с общей ограниченной очередью для разнотипных вызовов и со скачкообразными приоритетами. В данной модели вероятность перехода из одной очереди в другую зависит от числа L-вызовов в очереди, при этом допускается переход случайного числа L-вызовов. Введение ограничений на размер общего буфера для ожидания разнотипных вызовов приводит к необходимости

[1] Работа поддержана исследовательским грантом Sangji University (Korea), 2012 год.

определения нового показателя качества обслуживания (Quality of Service, QoS) — вероятности потери пакетов (Cell Loss Probability, CLP). Другое отличие настоящей работы от [9–14] состоит в том, что для анализа используется иной подход, основанный на теории фазового укрупнения состояний двумерных цепей Маркова [15].

МОДЕЛЬ СМО С КОНЕЧНОЙ ОБЩЕЙ ОЧЕРЕДЬЮ И СКАЧКООБРАЗНЫМИ ПРИОРИТЕТАМИ

На вход одноканальной системы поступают два пуассоновских потока разнотипных вызовов, при этом интенсивность i-го потока равна λ_i, $i = 1, 2$. Первый поток является потоком вызовов реального времени (H-вызовы), а второй — нереального (L-вызовы). Время занятия канала — случайная величина, подчиненная показательному закону распределения с параметром μ для вызовов обоих типов. Вызовы реального времени имеют высокие относительные приоритеты перед вызовами нереального времени. Это означает, что при освобождении канала на обслуживание из очереди всегда выбирается вызов первого типа независимо от числа вызовов второго типа в очереди, а также времени их ожидания. Внутри каждого потока используется дисциплина «первый пришел–первым обслужился» (First Come–First Serviced).

В исследуемой системе для ожидания разнотипных вызовов имеется общий буфер с максимальным размером R, $0 < R < \infty$, при этом предполагается, что в нем выполняется виртуальное разделение очередей разнотипных вызовов. Ограниченность буфера означает, что если в момент поступления вызова любого типа общий буфер полностью заполнен, то этот вызов теряется независимо от конкретного набора разнотипных вызовов в буфере.

Скачкообразные приоритеты определяются следующим образом. Прежде всего отметим, что H-вызовы всегда принимаются с вероятностью единица, если в момент их поступления имеется хотя бы одно свободное место в буфере; в противном случае они теряются с вероятностью единица. Если в момент поступления L-вызова число вызовов данного типа в буфере равно k и при этом в нем имеются свободные места, то с вероятностью $\alpha_m(k)$ ровно m L-вызовов мгновенно переходят в H-очередь, где $m = 1, 2, \ldots, \min(k, j)$, j — число свободных мест в буфере. С вероятностью $\alpha_0(k)$ поступивший L-вызов присоединяется к очереди, если в ней имеется свободное место, и никаких переходов не происходит, при этом указанные вероятности составляют полную группу.

В случае успешного «прыжка» все L-вызовы становятся H-вызовами и в дальнейшем обслуживаются как H-вызовы согласно HOL-приоритету. Если в момент поступления L-вызова нет свободного места в общей очереди, то с вероятностью единица он теряется.

Отметим некоторые важные частные случаи введенных скачкообразных приоритетов.

Равномерная схема. В такой схеме вероятности $\alpha_m(k)$ зависят только от m и не зависят от числа L-вызовов в буфере, т.е. $\alpha_m(k) = \alpha_m$ для любого $k = 0, 1, \ldots, R-1$.

Пороговая схема. В данной схеме вводятся пороговые параметры L_i, $i = 1, \ldots, r$, и вероятности $\alpha_m(k)$ определяются следующим образом:

$$\alpha_m(k) = \begin{cases} \alpha_i, & \text{если } L_{i-1} \le k < L_i,\ i = 1, 2, \ldots, r-1, \\ \alpha_r, & \text{если } L_{r-1} \le k \le L_r. \end{cases}$$

Здесь $L_0 := 0$, $L_r := R$.

Рассмотрим задачу нахождения показателей QoS этой модели. Основными показателями QoS являются стационарная вероятность блокировки вызовов i-го типа (CLP_i), среднее число вызовов каждого типа в буферах (Q_i), а также среднее время задержки передачи пакета (Cell Transfer Delay, CTD), т.е. ожидания вызовов в буфере (CTD_i), $i = 1, 2$.

Для упрощения промежуточных математических выкладок предположим, что разрешается переход только одного L-вызова в H-очередь (для определенности полагаем, что H-вызовом становится L-вызов, стоящий в начале очереди L-вызовов). Исходя из этого допущения, далее в формулах индексы параметров $\alpha(k)$ опускаются.

Поскольку времена обслуживания разнотипных вызовов имеют одинаковые средние значения, состояние буфера в произвольный момент времени может быть описано двумерным вектором $\mathbf{n} = (n_1, n_2)$, где n_i — число i-вызовов в буфере, $i = 1,2$. Иными словами, функционирование данной системы описывается двумерной цепью Маркова с фазовым пространством состояний (ФПС)

$$S := (\mathbf{n}\colon n_i = 0,1,\ldots,R,\ i = 1,2;\ n_1 + n_2 \le R). \tag{1}$$

В этой модели неотрицательные элементы Q-матрицы двумерной цепи определяются следующим образом:

$$q(\mathbf{n}, \tilde{\mathbf{n}}) = \begin{cases} \lambda_1 + \lambda_2\alpha(n_2), & \text{если } \tilde{\mathbf{n}} = \mathbf{n} + \mathbf{e}_1, \\ \lambda_2(1-\alpha(n_2)), & \text{если } \tilde{\mathbf{n}} = \mathbf{n} + \mathbf{e}_2, \\ \mu, & \text{если } n_1 > 0,\ \tilde{\mathbf{n}} = \mathbf{n} - \mathbf{e}_1 \text{ или } n_1 = 0,\ \tilde{\mathbf{n}} = \mathbf{n} - \mathbf{e}_2, \\ 0 & \text{в остальных случаях,} \end{cases} \tag{2}$$

где $\mathbf{e}_1 = (1,0),\ \mathbf{e}_2 = (0,1)$.

При любых положительных значениях параметров входящих трафиков все состояния системы сообщающиеся, следовательно, она эргодическая. Стандартный путь нахождения стационарных вероятностей состояний $p(\mathbf{n})$, $\mathbf{n} \in S$, состоит в решении соответствующей системы уравнений равновесия (СУР) (из-за своей очевидности явный вид этой СУР здесь не приводится).

После нахождения вероятностей состояний системы можно определить ее показатели QoS. Так, вероятности потери разнотипных вызовов равны между собой и определяются следующим образом:

$$CLP_1 = CLP_2 = \sum_{k=0}^{R} p(R-k, k). \tag{3}$$

Для нахождения среднего числа разнотипных вызовов в очереди $(Q_k, k = 1,2)$ используется стандартный способ определения среднего значения дискретной случайной величины:

$$Q_k = \sum_{i=1}^{R} i\xi_k(i), \tag{4}$$

где $\xi_k(i) = \sum_{\mathbf{n} \in S} p(\mathbf{n})\delta(n_k, i),\ k = 1,2,$ — маргинальные распределения исходной модели.

После нахождения показателей QoS (3) и (4) с помощью модифицированной формулы Литтла определяются средние времена задержки передачи разнотипных вызовов:

$$CTD_k = \frac{Q_k}{\lambda_k(1-CLP_k)},\ k = 1,2. \tag{5}$$

Рассмотренный точный метод нахождения показателей QoS, основанный на решении СУР для вероятностей состояний, может быть применен только при небольших размерностях ФПС данной модели, в противном случае сталкиваются с вычислительными трудностями. В связи с этим возникает необходимость использования приближенных методов.

Разработанный метод имеет высокую точность для моделей с большой нагрузкой H-вызовов, т.е. далее принимается следующее допущение: $\nu_1 >> \nu_2$. Оно не является экстраординарным, так как именно в системах с высокими нагрузками H-вызовов целесообразно введение скачкообразных приоритетов для L-вызовов.

Рассмотрим расщепление ФПС (1) данной модели:

$$S = \bigcup_{i=0}^{R} S_i, S_i \cap S_j = \varnothing, i \neq j, \tag{6}$$

где $S_i = \{\mathbf{n} \in S: n_2 = i\}, i = 0,1,2,\ldots,R$.

Заметим, что принятое допущение относительно соотношения нагрузок разнотипных вызовов обеспечивает выполнение условия корректного применения алгоритмов фазового укрупнения (АФУ) двумерных цепей Маркова [15].

Классы микросостояний S_i объединяются в отдельные укрупненные состояния $< i >$, и вводится функция укрупнения на исходном пространстве состояний S

$$U(\mathbf{n}) = < i >, \quad \mathbf{n} \in S_i, \tag{7}$$

определяющая укрупненную модель с пространством состояний $\Omega = \{< i >: i = 0,1,\ldots,R\}$.

Стационарную вероятность состояния (k,i) в расщепленной модели с пространством состояний S_i обозначим $\rho_i(k)$, $k = 0,1,\ldots,R$. Каждая расщепленная модель с ФПС S_i является одномерным процессом размножения и гибели. Следовательно, для нахождения стационарных вероятностей состояний внутри расщепленных моделей с ФПС S_i можно использовать формулы расчета стационарных вероятностей состояний СМО с зависящей от состояния интенсивностью входящего трафика типа $M(\lambda_1 + \lambda_2\alpha(i))\,|\,M(\mu)\,|\,1\,|\,R-i$ (здесь и далее для обозначения СМО применяем модифицированную символику Кендалла, где в скобках указываются параметры соответствующих распределений). Иными словами, для нахождения искомых параметров могут быть использованы формулы

$$\rho_i(k) = \theta_i^k \frac{1-\theta_i}{1-\theta_i^{R+1-i}}, \quad i = 0,1,\ldots,R-i, \tag{8}$$

где $\theta_i := \nu_1 + \nu_2\alpha(i)$. Для краткости здесь приводятся формулы только для случая $\theta_i \neq 1$.

Согласно АФУ двумерных цепей Маркова [15] неотрицательные элементы Q-матрицы укрупненной модели определяются следующим образом:

$$q(< i >, < j >) = \begin{cases} \lambda_2(1-\alpha(i))(1-\rho_i(R-i)), & j = i+1, \\ \mu\rho(0), & j = i-1, \\ 0 \quad \text{в остальных случаях.} \end{cases} \tag{9}$$

Стационарные вероятности укрупненных состояний имеют вид

$$\pi(< i >) = \prod_{j=1}^{i} A_j \pi(0), \quad i = 1,2,\ldots,R, \tag{10}$$

где $A_j = \nu_2 \dfrac{(1-\alpha(j-1))(1-\rho_{j-1}(R-j+1))}{\rho_j(0)}$, $\pi(0) = 1\Big/\left(1 + \sum\limits_{k=1}^{R}\prod\limits_{i=1}^{k} A_i\right)$.

С учетом формул (8)–(10) после некоторых преобразований получим следующие формулы для приближенного вычисления показателей QoS модели:

$$CLP \approx \sum_{k=0}^{R} \rho_k(R-k)\pi(< k >), \tag{11}$$

$$Q_1 \approx \sum_{k=1}^{R} k \sum_{i=0}^{R-k} \rho_i(k)\pi(< i >), \tag{12}$$

$$Q_2 \approx \sum_{k=1}^{R} k\pi(< k >). \tag{13}$$

Определив из формул (11)–(13) параметры CLP_k и Q_k, найдем параметры CTD_k, $k = 1, 2$, для данной модели (см. формулы (5)).

В частном случае, когда $\alpha(k) = 0$ для любого k (классические HOL-приоритеты), формулы (11)–(13) существенным образом упрощаются. Тогда стационарные вероятности состояний внутри классов S_i запишем таким образом:

$$\rho_i(k) = \nu_1^k \frac{1-\nu_1}{1-\nu_1^{R+1-i}}, \quad i = 0, 1, \ldots, R, \quad k = 0, 1, \ldots, R-i. \tag{14}$$

Вероятности укрупненных состояний (10) в этом случае определяются из следующих простых формул:

$$\pi(< i >) = \pi(0)\nu_2^i \prod_{k=0}^{i-1} G(k), \quad i = 1, 2, \ldots, R, \tag{15}$$

где $G(k) = (1-\nu_1^{R-k})/(1-\nu_1)$, $\pi(0) = 1/\left(1 + \sum_{i=1}^{R} \nu_2^i \prod_{k=0}^{i-1} G(k)\right)$.

Далее с помощью формул (11)–(13) вычисляются показатели QoS данной модели при использовании классических HOL-приоритетов. Формулы (14), (15) полностью совпадают с результатами, полученными в [16, с. 52–57]; в этой работе с использованием большого объема вычислительных экспериментов показана высокая точность этих формул. Аналогичные формулы могут быть записаны и для пороговой схемы определения скачкообразных приоритетов.

ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

Полученные формулы позволяют изучить поведение показателей QoS рассматриваемой системы относительно изменения ее структурных и нагрузочных параметров.

Далее приводится небольшая часть результатов вычислительных экспериментов для гипотетической модели с нагрузочными параметрами: $\lambda_1 = 2$, $\lambda_2 = 1$, $\mu = 0{,}8$. Цель экспериментов — изучение поведения показателей QoS системы относительно изменения размера общего буфера для разнотипных вызовов (т.е. R) и параметров $\alpha(k)$, характеризующих вероятности переходов L-вызовов в очередь H-вызовов. Рассматриваются три схемы определения скачкообразных приоритетов: 1) $\alpha(k) = 0$ для любого $k = 0, 1, \ldots, R-1$ (классические HOL-приоритеты); 2) $\alpha(k) = 0{,}7$ для любого $k = 0, 1, \ldots, R-1$; 3) $\alpha(k) = (k+1)/(k+2)$ для $k = 0, 1, \ldots, R-1$. Соответствующие результаты приведены на рис. 1–3.

Функция *CLP* с очень малой скоростью уменьшается при схемах 1 и 2 только для малых значений R; уже при $R \geq 5$ она становится почти постоянной. Однако при схеме 3 скорость ее уменьшения значительно возрастает. При этом схема 2 является промежуточной между двумя другими схемами (рис. 1). Иными словами, классические HOL-приоритеты наихудшие среди рассматриваемых схем с точки зрения уменьшения вероятности потери вызовов.

Скорость роста функции Q_1 (рис. 2, *а*) относительно изменения общего объема буфера достаточно высокая только при схеме 3; она остается почти постоянной при остальных двух схемах для $R \geq 5$. Однако при всех схемах абсолютное значение этой функции не превышает 0,45. Функция Q_2 при трех схемах (рис. 2, *б*) имеет почти линейный характер, при этом ее значения очень близки, но немного меньше при схеме 3. По этому показателю QoS классические HOL-приоритеты наихудшие среди рассматриваемых схем с точки зрения улучшения значения Q_2, а относительно показателя Q_1 они наилучшие.

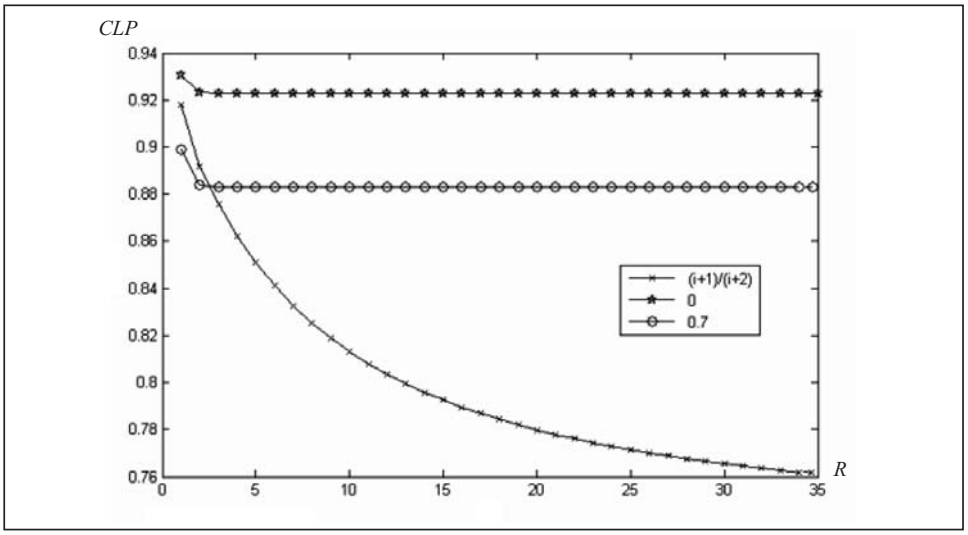

Рис. 1. Графики зависимости вероятности потери вызовов от размера буфера

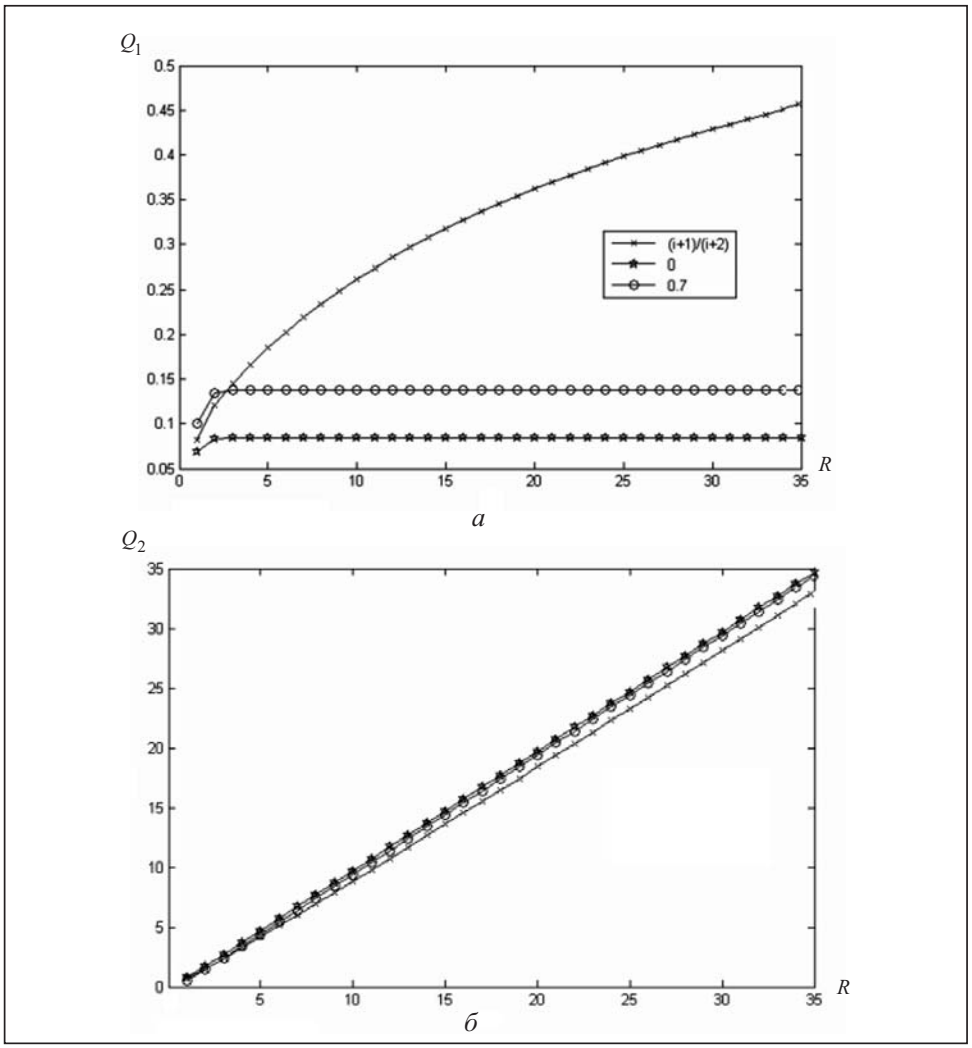

Рис. 2. Графики зависимости среднего числа H-вызовов (*а*) и L-вызовов (*б*) в очереди от размера буфера

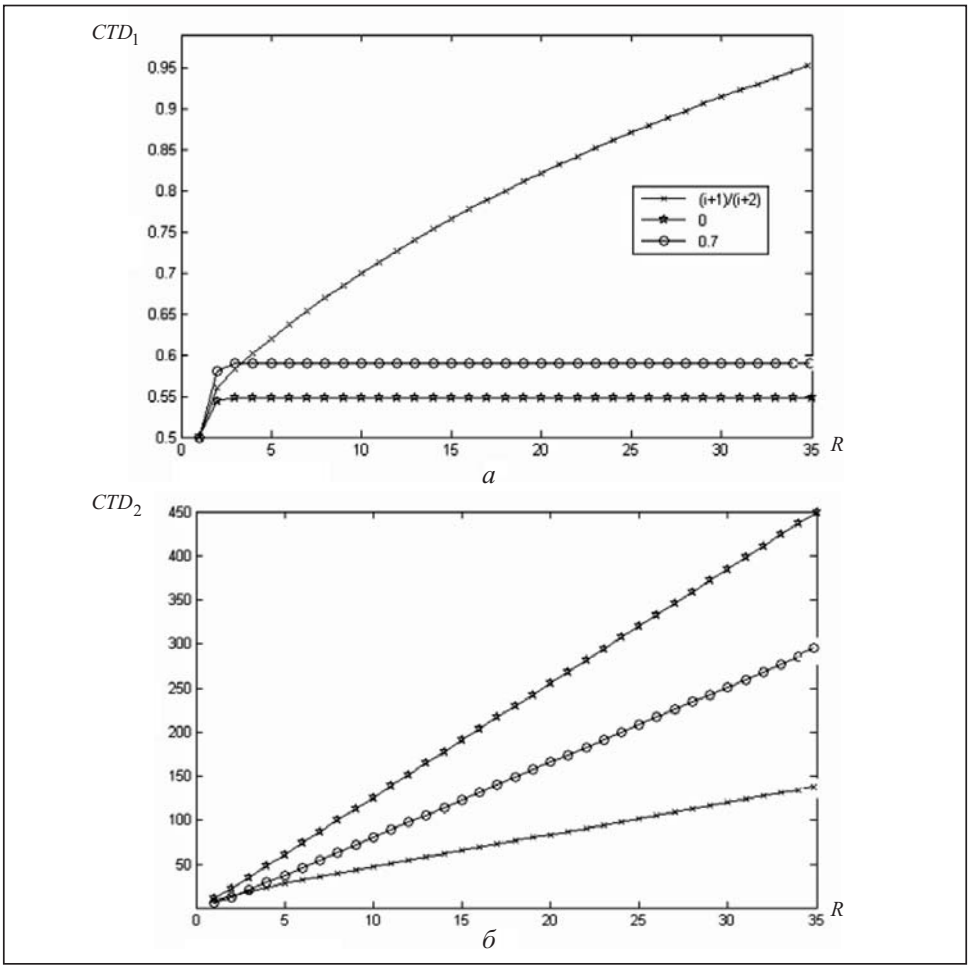

Рис. 3. Графики зависимости среднего времени ожидания в очереди H-вызовов (*а*) и L-вызовов (*б*) от размера буфера

Поведение функций CTD_1 и CTD_2 (рис. 3) полностью соответствует характеру изменений функций Q_1 и Q_2. При схеме 3 характеристики CTD_1 ухудшаются (рис. 3, *а*), в то время как для CTD_2 (рис. 3, *б*) она наиболее благоприятная; при этом значения CTD_1, как и Q_1, почти не изменяются при схемах 1 и 2, когда $R \geq 5$. Вместе с тем использование схемы 3 позволяет более чем в три раза уменьшить значение функции CTD_2 по сравнению со схемой 1, а схемы 2 — почти в два раза. Отметим, что по данному показателю QoS классические HOL-приоритеты, как и ранее, являются наихудшими среди рассматриваемых схем с точки зрения улучшения показателя CTD_2, а относительно показателя CTD_1 они наилучшие.

Численные эксперименты позволяют в некоторых случаях найти эффективную схему в системах со скачкообразными приоритетами при заданных ограничениях на показатели QoS исследуемых моделей. Однако решение таких задач в строгой математической формулировке представляет собой достаточно сложную проблему из-за большого числа показателей качества обслуживания.

Отметим, что проведен также анализ показателей QoS рассмотренной системы относительно ее других параметров. Однако из-за ограниченности объема статьи эти данные, а также результаты исследования точности разработанных приближенных формул здесь не приводятся. При этом точные значения искомых показателей QoS для моделей умеренной размерности определялись из соответствующих СУР (некоторые соображения относительно высокой точности разработанных формул приведены в конце предыдущего раздела). Заметим, что точ-

ные и приближенные значения искомых показателей QoS в худшем варианте отличаются третьим знаком после десятичной точки.

ЗАКЛЮЧЕНИЕ

В настоящей работе предложен приближенный подход к вычислению показателей качества обслуживания разнотипных вызовов в системах обслуживания с общими ограниченными очередями при наличии скачкообразных приоритетов. Его преимуществом является возможность использования для конечных моделей любой размерности, так как искомые показатели определяются с помощью простых вычислительных процедур.

Данный подход может применяться для исследования моделей, в которых вероятности $\alpha(i)$ зависят также от количества вызовов первого типа. Кроме того, его можно использовать для изучения моделей, в которых возможны переходы случайного числа L-вызовов в H-очередь. Эти проблемы представляют собой предмет отдельных исследований.

Авторы выражают благодарность члену-корреспонденту НАН Украины Н.Ю. Кузнецову за ценные замечания, способствующие устранению некоторых неточностей.

СПИСОК ЛИТЕРАТУРЫ

1. Джейсуол Н. Очереди с приоритетами. — М.: Мир, 1973. — 208 с.
2. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. — М.: КомКнига, 2005. — 400 с.
3. Kleinrock L. A delay dependent queue discipline // Naval Res. Logist. Quart. — 1964. — **11**. — P. 329–341.
4. Мова В.В., Пономаренко Л.А. Об оптимальных приоритетах, зависящих от текущего состояния обслуживающей системы с конечным числом мест для ожидания // Изв. АН СССР. Техн. кибернетика. — 1974. — № 5. — С. 74–81.
5. Lee Y., Choi B.D. Queuing system with multiple delay and loss priorities for ATM networks // Inform. Sci. — 2001. — **138**. — P. 7–29.
6. Melikov A.Z., Feyziev V.S., Rustamov A.M. Analysis of model of data packet processing in ATM networks with multiple space and time priorities // Aut. Control and Computer Sci. — 2006. — **40**, N 6. — P. 38-45.
7. Melikov A.Z., Ponomarenko L.A., Kim C.S. Approximation method for performance analysis of queuing systems with multimedia traffics // Appl. and Comput. Math. — 2007. — **6**, N 2. — P. 1–8.
8. Demoor T., Fiems D., Walraevens J. Partially shared buffers with full or mixed priority // J. Industr. and Manag. Optim. — 2011. — **7**, N 3. — P. 735–751.
9. Lim Y., Kobza J.E. Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch // IEEE Trans. Commun. — 1990. — **38**, N 5. — P. 659–665.
10. Maertens T., Walraevens J., Bruneel H. On priority queues with priority jumps // Perform. Eval. — 2006. — **63**, N 12. — P. 1235–1252.
11. Maertens T., Walraevens J., Bruneel H. A modified HOL priority scheduling discipline: performance analysis // Eur. J. Oper. Res. — 2007. — **180**, N 3. — P. 1168–1185.
12. Maertens T., Walraevens J., Moeneclaey M., Bruneel H. A new dynamic priority scheme: performance analysis // Proc. of the 13th Intern. Conf. on Analytical and Stochastic Modeling Techniques and Applications (ASMTA), 28–31 May, 2006. — Bonn: Springer, 2006. — P. 74–84.
13. Walraevens J., Steyaert B., Bruneel H. Performance analysis of single-server ATM queue with priority scheduling // Comput. and Oper. Res. — 2003. — **30**, N 12. — P. 1807–1829.
14. Maertens T., Walraevens J., Bruneel H. Performance comparison of several priority schemes with priority jumps // Ann. Oper. Res. — 2008. — **162**. — P. 109–125.
15. Ponomarenko L., Kim C.S., Melikov A. Performance analysis and optimization of multi-traffic on communication networks. — Heidelberg; Dordrecht; London; New York: Springer, 2010. — 208 p.
16. Меликов А.З., Пономаренко Л.А., Фаттахова М.И. Управление мультисервисными сетями связи с буферными накопителями. — К.: НАУ-друк, 2008. — 156 с.

Поступила 10.05.2012