УДК 519.872:621.321.1

А. З. Меликов, чл.-кор. НАН Азербайджана,
Национальная академия авиации
(Азербайджан, АЗ 1045, Баку, 25-й км, пос. Бина,
тел: +994 12 497 26 00, e-mail: agassi@science.az),
В. Ш.Фейзиев, Ф. Н. Нагиев, кандидаты техн. наук,
Ин-т кибернетики НАН Азербайджана
(Азербайджан, AZ1141, Баку, ул.Ф.Агаева, 9)

Алгоритмический подход к анализу модели обслуживания со скачкообразными приоритетами

Предложен алгоритмический подход к исследованию модели обслуживания со скачкообразными приоритетами. Предполагается, что низкоприоритетные запросы могут перейти в конец очереди высокоприоритетных запросов, если их время пребывания в очереди превышает некоторую случайную величину. Разработан алгоритм расчета характеристик таких моделей обслуживания.

Запропоновано алгоритмічний спосіб дослідження моделі обслуговування із стрибкуватими пріоритетами. Припущено можливість переходу низькопріоритетних запитів у кінець черги високопріоритетних запитів, якщо їхній час перебуваня в черзі перевищує певну випадкову величину. Розроблено алгоритм розрахунку характеристик таких моделей обслуговування.

Ключевые слова: модель обслуживания, скачкообразные приоритеты, качество обслуживания, алгоритм расчета.

Модели систем обслуживания с приоритетами широко используются при математическом анализе интегрированных сетей коммутации пакетов. Это объясняется тем, что в указанных сетях обрабатываются пакеты разнотипных запросов реального и нереального времени (видео и речевая информация, данные и др.), имеющих различные уровни важности.

Наиболее актуальными являются модели с приоритетами, так как иногда только правильный выбор системы приоритетов обеспечивает удовлетворение противоречивых требований разнотипных запросов относительно показателей качества обслуживания QoS (Quality of Service). Противоречивость требований означает, что запросы (пакеты) реального времени являются более чувствительными к возможным задержкам, чем запросы нереального времени, но, в то же время, запросы нереального времени предъявляют более жесткие требования к возможным потерям, чем запросы реального време-

ни. В связи с этим в последние годы исследуется новый тип приоритетов — множественные приоритеты [1—4]. При использовании множественных приоритетов пакеты реального времени имеют высокие временные и низкие пространственные приоритеты, а пакеты нереального времени имеют низкие временные и высокие пространственные приоритеты. Следует заметить, что пространственные приоритеты используются для разрешения конфликтных ситуаций, связанных с занятием мест в буферном накопителе при поступлении пакетов, а временные — определяют порядок выбора пакета из буфера для передачи на выходящий порт. Достаточно подробный обзор работ в этом направлении приведен в [5].

В работах [6—10] изучен другой тип приоритетов. В работе [6] эти приоритеты названы скачкообразными (Jump Priorities). Модель, описанная в [6], состоит в следующем. На вход одноканальной системы с бесконечными раздельными очередями поступает $N > 1$ запросов. Считается, что запросы типа i имеют высокие (относительные) приоритеты по сравнению с запросами типа $i + 1$, $i = 1, ..., N - 1$. Для трафика типа i определяются детерминированные параметры D_i, $0 < D_1 < D_2 < ... < D_N \leq \infty$. Если время ожидания i-запроса, стоящим во главе i-й очереди, достигает величины $D_i - D_{i-1}$, то он переходит в очередь $i - 1$, $i = 2, ..., N$. Этот процесс продолжается до тех пор, пока запрос любого типа не достигнет очереди с наивысшим приоритетом (очередь 1). Поскольку точный анализ состояния очереди и распределения времени ожидания в очереди оказываются сложными задачами, в [6] предложены формулы для расчета среднего времени ожидания разнотипных запросов. Следует заметить, что предложенные приоритеты неудобны в практической реализации, так как требуют использования дополнительных технических средств для мониторинга времени ожидания разнотипных запросов.

В работах [7—10] предложены модели систем обслуживания с дискретным временем (время разделено на слоты) и с двумя типами запросов: запросы высокого приоритета (H-запросы) и запросы низкого приоритета (L-запросы). Для ожидания запросов каждого типа имеются раздельные бесконечные очереди. В [7] предложена схема HOL-MBP (Head-Of-Line Merge-By-Probability), согласно которой в конце каждого временного слота все L-запросы переходят в конец очереди H-запросов с вероятностью β ($0 < \beta < 1$). Иными словами, в конце каждого временного слота очереди H- и L-запросов объединяются (укрупняются) с вероятностью β.

Модифицированная схема HOL-MBP описана в работе [8]. Она получила название HOL-JOS (Head-Of-Line Jump-Or-Serve) и в отличие от предыдущей в ней только один L-запрос из начала очереди переходит в H-очередь. Возможный переход в начале каждого временного слота зави-

сит от состояния *H*-очереди в начале слота, т.е. если очередь не является пустой, то происходит переход (иначе *L*-запрос немедленно передается в канал).

В схеме HOL-JIA[1] (Head-Of-Line Jump-If-Arrival) [9], в отличие от схемы HOL-JOS, возможный переход *L*-запроса в *H*-очередь зависит не только от наполнения *H*-очереди в начале слота, но и от числа поступлений *L*-запросов в период данного слота. Точнее, внутри временного слота, в который передается *H*-запрос, переход *L*-запроса в *H*-очередь разрешается лишь тогда, когда в период данного слота происходит поступление *L*-запросов. В данной схеме поступившим *L*-запросам не разрешается немедленно переходить в *H*-очередь.

В работе [10] предложена схема HOL-JIA[2], единственное отличие которой от схемы HOL-JIA[1] состоит в том, что в ней поступившим *L*-запросам разрешается немедленно переходить в *H*-очередь.

В работах [9—11] приведены формулы для производящих функций времени ожидания в очереди запросов обоих типов и времени ожидания в очереди *H*-запросов, что является сложной задачей, а также их моменты. Кроме того, определено среднее время ожидания в очереди *L*-запросов.

Указанные выше работы посвящены исследованию моделей обслуживания с бесконечными очередями, которые не являются адекватными моделями реальных систем телекоммуникации, так как реальные системы, как правило, имеют ограниченные буферные накопители для временного хранения разнотипных запросов. Для их широкого внедрения потребуется определить эффективность указанных приоритетов в реальных системах. Предлагаемая модель системы обслуживания с конечными очередями и со скачкообразными приоритетами разработана на основе нового подхода к анализу изучаемой системы, а предложенные алгоритмы расчета ее показателей QoS основаны на методах приближенного расчета двумерных цепей Маркова [12].

Модель системы и метод расчета. На вход одноканальной системы (рис. 1) поступает два пуассоновских потока разнотипных запросов (пакетов), при этом интенсивность *i*-го потока равна λ_i, $i = 1, 2$. Первый поток представляет собой поток запросов реального времени, а второй — поток запросов нереального времени. Время занятия канала является случайной величиной, подчиненной показательному закону распределения с параметром μ для запросов обоих типов.

Для ожидания в очереди разнотипных запросов имеются раздельные буфера. Размер буфера для запросов *i*-го типа $0 < R_i < \infty$, $i = 1, 2$. Ограниченность раздельных буферов означает, что если в момент поступления запроса любого типа соответствующий буфер полностью заполнен, то

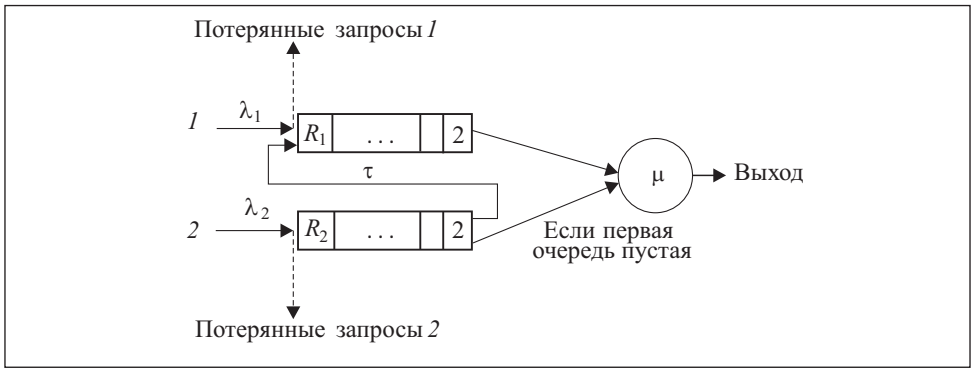

Рис. 1. Структурная схема системы: *1* и *2* — запросы первого и второго типа

этот запрос теряется независимо от состояния другого буфера. Запросы реального времени имеют высокие относительные приоритеты перед запросами нереального времени. Это означает, что при освобождении канала на обслуживание из очереди всегда выбирается запрос первого типа независимо от числа запросов второго типа в очереди, а также от времени их ожидания в очереди. Внутри каждого потока используется дисциплина «первый пришел — первым обслужился». Такие приоритеты иногда называются HOL-приоритетами.

При использовании HOL-приоритетов время ожидания в очереди запросов второго типа может быть сколь угодно большое, особенно при большом числе запросов первого типа. В связи с этим вводятся скачкообразные приоритеты, для того чтобы увеличить шансы запросов второго типа быть обслуженными. При этом запрос, стоящий в начале очереди запросов второго типа, через случайное время T «прыгает» в конец очереди запросов первого типа, если там имеется хотя бы одно свободное место. В противном случае, т.е. если в этот момент буфер для ожидания запросов первого типа полностью заполнен, то «прыгающий» запрос второго типа теряется. В случае успешного «прыжка» запрос второго типа становится запросом первого типа и в дальнейшем обслуживается как запрос первого типа согласно HOL-приоритетом. Для получения результатов предположим, что величина T имеет показательное распределение со средним τ^{-1}.

Рассмотрим задачу нахождения показателей QoS этой модели. Основными показателями QoS являются стационарная вероятность блокировки запросов i-го типа, CLP_i, среднее число запросов каждого типа в буферах Q_i и среднее время их ожидания в буфере CTD_i, $i = 1, 2$. Указанные показатели QoS для запросов первого типа могут быть определены как соответствующие параметры классической модели $M/M/1/R_1$ с нагрузкой

$\nu_1 := (\lambda_1 + \tau)/\mu$. Это объясняется тем, что запросы первого типа имеют высокие относительные приоритеты перед запросами второго типа и продолжительность занятия канала — одинаковая для обоих типов запросов. Однако искомые параметры для запросов второго типа не могут быть определены так легко. Для их определения используем следующую двумерную цепь Маркова.

Состояние буферов в произвольный момент времени описывается с помощью двумерного вектора $\mathbf{n} = (n_1, n_2)$, где n_i — число i-запросов в буфере, $i = 1, 2$. Следовательно, функционирование данной системы описывается двумерной цепью Маркова с фазовым пространством состояний (ФПС)

$$S := \{\mathbf{n}: n_i = 0, 1, \ldots, R_i, i = 1, 2\}. \tag{1}$$

Переходы между состояниями системы происходят в моменты поступления запросов, ухода их из системы после завершения обслуживания, а также при переходе запроса из второй очереди в первую. С учетом этого неотрицательные элементы Q-матрицы данной многомерной цепи определяются из следующих соотношений (рис. 2):

$$q(n, \tilde{n}) = \begin{cases} \lambda_1, & \text{если } \tilde{n} = n + e_1, \\ \lambda_2, & \text{если } \tilde{n} = n + e_2, \\ \mu, & \text{если } n_1 > 0, \tilde{n} = n - e_1 \text{ или } n_1 = 0, \tilde{n} = n - e_2, \\ n_2\tau, & \text{если } n_2 > 0, n_1 < R_1, \tilde{n} = n + e_1 - e_2, \\ 0 & \text{в остальных случаях,} \end{cases} \tag{2}$$

где $e_1 = (1, 0)$, $e_2 = (0, 1)$. При любых положительных значениях параметров входящих трафиков все состояния — сообщающиеся и, следовательно, система — эргодическая.

Стационарную вероятность состояния $n \in S$ обозначим $p(n)$. Стандартный путь нахождения стационарных вероятностей состояний — составление и решение соответствующей системы уравнений равновесия (СУР). С использованием (2) легко показать, что искомая система имеет следующий вид:

для $n_1 = 0$

$$(\lambda_1 + \lambda_2 I(n_2 < R_2) + \mu I(n_2 > 0) + n_2\tau)\, p(n) =$$

$$= \lambda_2 p(n - e_2)\, I(n_2 > 0) + \mu p(n + e_2); \tag{3}$$

для $n_1 \neq 0$

$$(\lambda_1 I(n_1 < R_1) + \lambda_2 I(n_2 < R_2) + \mu + n_2\tau)\, p(n) =$$

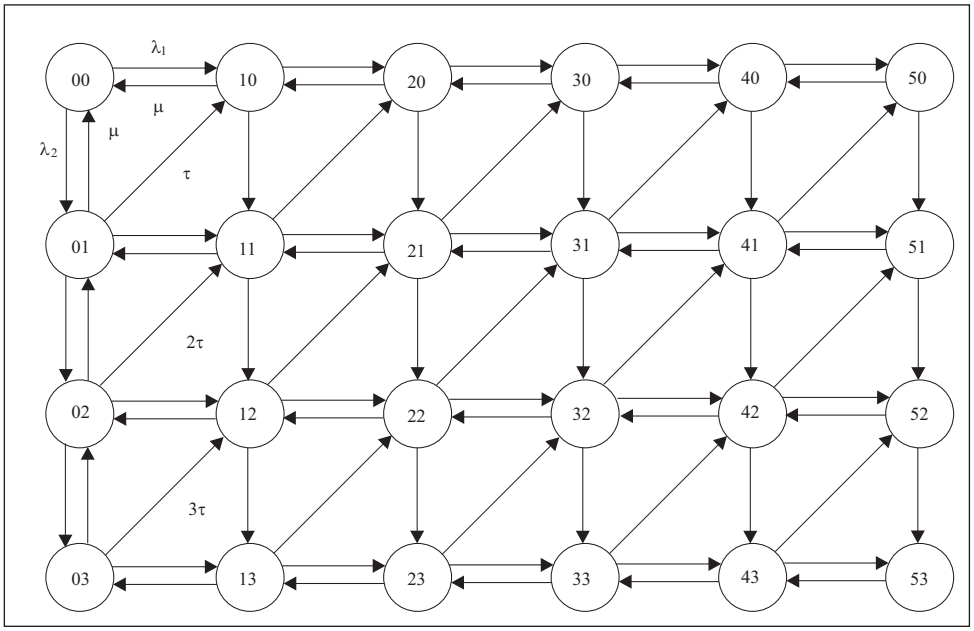

Рис. 2. Граф переходов между состояниями модели, $R_1 = 5$, $R_2 = 3$

$$=\lambda_1 p(n-e_1)+\lambda_2 p(n-e_2)+$$

$$+\mu p(n+e_1)\, I(n_1 < R_1)+(n_2+1)\tau p(n-e_1+e_2), \tag{4}$$

где $I(A)$ — индикаторная функция события A. К системе уравнений (3), (4) добавим нормирующее условие

$$\sum_{n\in S} p(n)=1. \tag{5}$$

После нахождения вероятностей состояний системы можно определить ее показатели QoS. Вероятность потери запросов второго типа определяется так:

$$CLP_2 = \sum_{n\in S} p(n)\,\delta(n_1+n_2, R_2)+P_f P_2. \tag{6}$$

Здесь $\delta(x, y)$ — символы Кронекера; P_f — вероятность того, что очередь запросов высокого приоритета полностью заполнена, совпадающая с вероятностью потери CLP_1 в описанной выше классической системе $M/M/1/R_1$ с

нагрузкой ν_1 эрл.; P_2 — вероятность ухода из очереди запросов второго типа,

$$P_2 = \frac{\tau}{\lambda_2} \sum_{k=1}^{R_2} kp(R_1, k). \tag{7}$$

Для нахождения среднего числа пакетов запросов второго типа в очереди используем стандартный способ определения среднего значения дискретной случайной величины:

$$Q_2 = \sum_{i=1}^{R_2} i\xi(i), \tag{8}$$

где $\xi(i) = \sum_{n \in S} p(n)\delta(n_2, i)$, $i = 1, 2, ..., R_2$, — маргинальные распределения исходной модели.

После определения показателей QoS (6) и (8) с помощью модифицированной формулы Литтла находим среднее время задержки передачи запросов второго типа

$$CTD_2 = \frac{Q_2}{\lambda_2(1 - CLP_2)}. \tag{9}$$

Таким образом, для определения точных значений показателей QoS (6)—(9) необходимо решить СУР (3)—(5), которая не имеет аналитического решения, т.е. для ее решения могут быть использованы известные методы линейной алгебры. Описанную схему определения показателей QoS (6)—(9) практически можно применять лишь при небольших размерностях ФПС (1), но при их возрастании она становится неэффективной. Поэтому возникает необходимость в разработке более эффективного способа решения этой задачи.

Рассмотрим подход, обеспечивающий высокую точность для моделей с большим числом запросов первого типа. Примем допущение $\lambda_1 >> \lambda_2 >> \mu$. Заметим, что это допущение не является экстраординарным, так как именно в системах с большим числом запросов высокого приоритета целесообразно введение скачкообразных приоритетов для запросов низкого приоритета.

Рассмотрим следующее расщепление ФПС (1):

$$S = \bigcup_{i=0}^{R_2} S_i, \;\; S_i \cap S_j = \varnothing, \;\; i \neq j,$$

где $S_i = \{n \in S : n_2 = i\}$, $i = 0, 1, 2, ..., R_2$. Следует заметить, что принятое выше допущение относительно соотношения нагрузок разнотипных запросов

обеспечивает выполнение корректного применения алгоритмов фазового укрупнения двумерных цепей Маркова [12]. Классы микросостояний S_k объединяются в отдельные укрупненные состояния $< k >$ и вводится функция укрупнения на исходном ФПС (1):

$$U(n) = < k >, \ n \in S_k. \tag{10}$$

Функция (10) определяет укрупненную модель с ФПС $\Omega = \{< k >: \ k = 0, 1, ..., R_2\}$. Стационарную вероятность состояния (k, i) в расщепленной модели с ФПС S_i обозначим $\rho_i(k)$, $i = 0, 1, ..., R_2$, $k = 0, 1, ..., R_1$. Каждая расщепленная модель с ФПС S_i является одномерным процессом размножения и гибели со следующими параметрами (см. рис. 2):

$$q_i(k_1, k_2) = \begin{cases} \lambda_1, \text{если } k_2 = k_1 + 1, \\ \mu, \ \text{если } k_2 = k_1 - 1, \\ 0 \quad \text{в остальных случаях.} \end{cases}$$

Следовательно, для нахождения $\rho_i(k)$, $i = 0, 1, \ldots, R_2$, $k = 0, 1, \ldots, R_1$, может быть использована формула (для $\nu_1 \neq 1$)

$$\rho_i(k) = \nu_1^k \frac{1 - \nu_1}{1 - \nu_1^{R_1 + 1}}, \tag{11}$$

из которой видно, что параметры $\rho_i(k)$ не зависят от индекса i.

Согласно алгоритму фазового укрупнения двумерных цепей Маркова [12] элементы производящей матрицы укрупненной модели определяются из следующих соотношений (см. рис. 2):

$$q(< k_1 >, < k_2 >) = \begin{cases} \lambda_2, \ k_2 = k_1 + 1, \\ \mu\rho(0) + k_1\tau(1 - \rho(R_1)), \ k_2 = k_1 - 1, \\ 0 \text{ в остальных случаях,} \end{cases} \tag{12}$$

Следовательно, стационарные вероятности укрупненных состояний $\pi(< k >)$, $k \in \Omega$, имеют вид

$$\pi(< k >) = \prod_{i=1}^{k} A_i \pi(0), \ k = 1, 2, \ldots, R_2, \tag{13}$$

где

$$A_i = \frac{\nu_2}{\rho(0) + i\tau(1 - \rho(R_1))/\mu}, \ \pi(0) = \frac{1}{1 + \sum_{k=1}^{R_2} \prod_{i=1}^{k} A_i}.$$

После определенных преобразований с учетом (11)—(13) получим следующие приближенные формулы для вычисления показателей QoS исследуемой модели (см. (6)—(9)):

$$CLP_2 = \pi(< R_2 >)\frac{\tau}{\lambda_2}\rho(R_1)\sum_{k=1}^{R_2} k\pi(< k >)\,CLP_1;$$

$$Q_2 = \sum_{k=1}^{R_2} k\,\pi(< k >).$$

Параметр CTD_2 определяется по формуле (9).

Численные результаты. Полученные формулы позволяют изучить поведение показателей QoS исследуемой системы при изменении ее структурных и нагрузочных параметров. Результаты вычислительных экспериментов для гипотетической модели с параметрами $R_1 = 10$, $R_2 = 20$, $\lambda_1 = 7$, $\lambda_2 = 5$, $\mu = 2$ представлены на рис. 3. Исследовано поведение показателей QoS системы при изменении параметра τ, характеризующего интенсивность перехода низкоприоритетных пакетов в очередь высокоприоритетных.

Из рис.3, *а*, видно, что с возрастанием параметра τ вероятность потери высокоприоритетных пакетов CLP_1 медленно увеличивается. Это закономерно, так как с увеличением параметра τ число пакетов высокого приоритета возрастает. Показатель CLP_2 сначала убывает, а затем также начинает увеличиваться. Действительно, при малых значениях параметра τ покидающие *L*-очередь пакеты присоединяются к *H*-очереди и, следовательно, функция CLP_2 уменьшается. Вместе с тем, эта функция возрастет более определенного (порогового) значения параметра τ в результате того, что покидающие *L*-очередь пакеты не могут присоединиться к *H*-очереди, так как она оказывается заполненной. Поэтому дальнейшее увеличение параметра τ приводит к увеличению функции CLP_2.

Однако функции Q_2 и CTD_2 (рис. 3, *б* и *в*) систематически уменьшаются, так как увеличение параметра τ в любом (допустимом) диапазоне приводит к уменьшению длины очереди пакетов низкого приоритета и, следовательно, уменьшается их среднее время ожидания в очереди. Однако функции Q_1 и CTD_1 хоть и медленно, но возрастают, так как увеличение параметра τ приводит к увеличению числа пакетов высокого приоритета.

В таблице приведены точные и приближенные значения искомых показателей QoS для моделей малой и средней размерности, полученные с помощью СУР (3)—(5) при исходных данных $R_1 = 10$, $R_2 = 20$, $\lambda_1 = 7$, $\lambda_2 = 5$, $\mu = 2$.

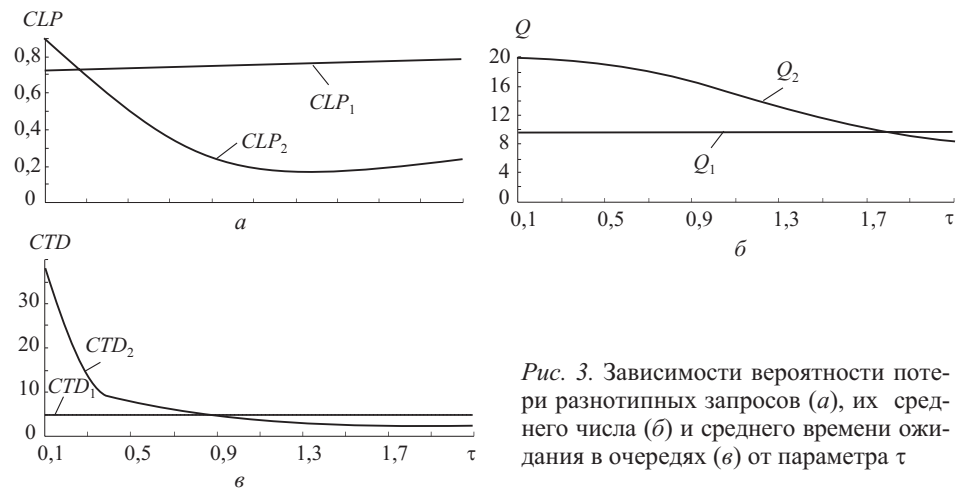

Рис. 3. Зависимости вероятности потери разнотипных запросов (*а*), их среднего числа (*б*) и среднего времени ожидания в очередях (*в*) от параметра τ

τ	Значение CLP_2		Значение Q_2		Значение CTD_2	
	точное	приближенное	точное	приближенное	точное	приближенное
0,1	0,897001	0,896702	19,87265	19,87280	38,47659	38,47664
0,3	0,696783	0,696579	19,51477	19,51464	12,86311	12,86306
0,5	0,511303	0,511372	18,93601	18,93534	7,75083	7,75041
0,7	0,353945	0,353917	18,00288	18,00275	5,57293	5,57289
0,9	0,240932	0,240914	16,62742	16,62713	4,38075	4,38083
1,1	0,180899	0,180924	14,91361	14,91350	3,64183	3,64154
1,3	0,164129	0,164118	13,14973	13,14987	3,14672	3,14635
1,5	0,171655	0,171624	11,57623	11,57615	2,79409	2,79491
1,7	0,189523	0,189510	10,26806	10,26852	2,53367	2,53390
1,9	0,211211	0,211206	9,20344	9,20323	2,33354	2,33349
2,1	0,234273	0,234268	8,33151	8,33119	2,17628	2,17601

Выводы

Достоинством предложенного подхода к вычислению показателей качества обслуживания разнотипных запросов в системах со скачкообразными приоритетами состоит в том, что он может быть использован для моделей любой размерности, так как искомые показатели вычисляются с помощью явных формул.

Algorithmic approach to study the queuing models with step-wise priorities is proposed. It is assumed that low-priority requests might pass to the queue end of high-priority requests if their sojourn time exceeds some random threshold. An algorithm to calculate characteristics of such models is developed.

1. *Lee Y., Choi B. D.* Queuing System with Multiple Delay and Loss Priorities for ATM Networks // Information Sciences. — 2001. — Vol. 138. — P. 7—29.
2. *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 // Automatic Control and Computer Sciences. — 2006. — Vol. 40, № 6. — P. 38—45.
3. *Melikov A. Z., Ponomarenko L. A., Kim C. S.* Approximation Method for Performance Analysis of Queuing Systems with Multimedia Traffics // Applied and Computational Mathematics. — 2007. — Vol. 6, № 2. — P. 1—8.
4. *Demoor T., Fiems D., Walraevens J.* Partially Shared Buffers with Full or Mixed Priority // J. of Industrial and Management Optimization. — 2011. — Vol. 7, № 3. — P. 735—751.
5. *Меликов А. З., Пономаренко Л. А., Фаттахова М. И.* Управление мультисервисными сетями связи с буферными накопителями. — Киев : НАУ-друк, 2008. — 156 с.
6. *Lim Y., Kobza J. E.* Analysis of Delay Dependent Priority Discipline in an Integrated Multiclass Traffic Fast Packet Switch // IEEE Transactions on Communications. — 1990. — Vol. 38, No. 5. — P. 659—665.
7. *Maertens T., Walraevens J., Bruneel H.* On Priority Queues with Priority Jumps // Performance Evaluation. — 2006. — Vol. 63, № 12. — P. 1235—1252.
8. *Maertens T., Walraevens J., Bruneel H.* A modified HOL Priority Scheduling Discipline: Performance Analysis // European Journal of Operational Research. — 2007. — Vol. 180, № 3. — P. 1168—1185.
9. *Maertens T., Walraevens J., Moeneclaey M., Bruneel H.* A new Dynamic Priority Scheme: Performance Analysis //Proc. of the 13th International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA). — 2006. — P.74—84.
10. *Maertens T., Walraevens J., Bruneel H.* Performance Comparison of Several Priority Schemes with Priority Jumps // Annals of Operations Research. — 2008. — Vol. 162. — P. 109—125.
11. *Walraevens J., Steyaert B., Bruneel H.* Performance Analysis of Single-server ATM Queue with Priority Scheduling // Computers and Operations Research. — 2003. — Vol. 30, № 12. — P. 1807—1829.
12. *Ponomarenko L., Kim C. S., Melikov A.* Performance Analysis and Optimization of Multitraffic on Communication Networks. — London : Springer, 2010. — 208 p.

Поступила 06.12.11

МЕЛИКОВ Агаси Зарбали оглы, чл.-кор. НАН Азербайджана, профессор, зав. кафедрой «Аэрокосмические информационные технологии и системы управления» Национальной академии авиации Азербайджана. В 1977 г. окончил Бакинский госуниверситет. Область научных исследований — моделирование коммуникационных сетей, анализ и оптимизация систем и сетей массового обслуживания.

ФЕЙЗИЕВ Вагиф Шейдулла оглы, канд. техн. наук, ст. науч. сотр. Ин-та кибернетики НАН Азербайджана. В 2005 г. окончил магистратуру Бакинского госуниверситета. Область научных исследований — моделирование коммуникационных сетей, анализ и оптимизация систем и сетей массового обслуживания.

НАГИЕВ Фуад Надир оглы, канд. техн. наук, ст. науч. сотр. Ин-та кибернетики НАН Азербайджана. В 2001 г. окончил магистратуру Бакинского госуниверситета. Область научных исследований — компьютерные сети, теория сетей массового обслуживания.