*ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2021, № 3, с. 98–110*

## СИСТЕМНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ

*УДК 519.872*

# ЧИСЛЕННЫЙ АНАЛИЗ СИСТЕМЫ С ГЕТЕРОГЕННЫМИ СЕРВЕРАМИ И МГНОВЕННОЙ ОБРАТНОЙ СВЯЗЬЮ

**© 2021 г. С. Г. Алиева[a], А. З. Меликов[a,*], М. О. Шахмалыев[a]**

[a] *Институт систем управления, Национальная академия наук Азербайджана, Баку, Азербайджан*

**e-mail: agassi.melikov@gmail.com*

Поступила в редакцию 10.09.2019 г.
После доработки 17.11.2020 г.
Принята к публикации 25.01.2021 г.

Изучается система с гетерогенными серверами, пуассоновским потоком с Марковской модуляцией и мгновенной обратной связью. Обслуживание первичных заявок выполняется в высокоскоростном сервере, и после завершения обслуживания каждая заявка, согласно схеме Бернулли, либо покидает систему, либо требует повторного обслуживания. Повторные заявки обслуживаются в медленном сервере, при этом эти заявки после завершения обслуживания также, согласно схеме Бернулли, либо покидают систему, либо требуют повторного обслуживания в медленном сервере. Если в момент поступления первичной заявки длина очереди таких заявок превышает некоторое пороговое значение и медленный сервер свободен, то поступившая заявка, согласно схеме Бернулли, либо направляется в медленный сервер, либо присоединяется в свою очередь. Считается, что длины очередей перед каждым сервером являются бесконечными величинами. Построена адекватная математическая модель изучаемой системы в виде трехмерной цепи Маркова с бесконечным пространством состояний. Получено условие эргодичности данной цепи, предложен приближенный алгоритм расчета стационарных вероятностей состояний и показана его высокая точность. Приводятся результаты численных экспериментов.

**DOI:** 10.31857/S0002338821030021

**Введение.** При построении математических моделей коммуникационных процессов, а также процессов, связанных с обработкой деталей в производственных системах, процессов управления запасами и т.д., возникает необходимость учета повторной обработки некоторых заявок. Так, во многих системах передачи информации ошибочно переданные данные (пакеты, кадры и т.д.) передаются повторно, так как эффективность работы таких систем зачастую оценивается достоверностью передачи данных. Если выпускаемые детали имеют определенные дефекты, то и в производственных системах в некоторых ситуациях требуется их повторная обработка.

Учет эффекта повторения обслуживания заявок приводит к необходимости использования моделей систем с обратной связью (feed back queue, (FBQ)). Такие модели впервые были введены в [1, 2]. Следует различать системы с мгновенной обратной связью (instantaneous feed back queue (IFBQ)), где повторение происходит сразу после завершения обслуживания заявки, и системы с отсроченной обратной связью (delayed feed back queue (DFBQ)), в которых повторение запроса для обслуживания происходит после определенного положительного времени.

Современное состояние проблемы исследования моделей FBQ подробно изложено в [3], поэтому здесь не будем останавливаться на изложении известных в этом направление результатов. В указанной работе отмечается, что в подавляющем большинстве публикаций изучаются модели FBQ без буфера для ожидания заявок. Вместе с тем во многих системах организуются буфера для хранения ожидающих в очереди заявок. Поэтому для адекватного описания работы таких систем возникает необходимость изучения моделей FBQ с очередями.

Исходя из вышеизложенных фактов, в настоящей статье исследуется модель IFBQ с очередями. Отметим, что в доступной литературе мало работ, которые посвящены изучению таких моделей, например [4–13]. Простые одномерные модели IFBQ с одним сервером и нетерпеливыми заявками при использовании различных механизмов удержания их в очереди в моменты завершения допустимого времени ожидания, рассмотрены в [4–10]. Модель с двумя гетерогенными серверами и ограниченной очередью, в которых поступающие заявки с известными вероятно-

стями назначаются в сервера, приведена в [11]. Отметим, что в [4–11] не различаются исходные заявки и заявки, которые требуют повторного обслуживания. Поэтому с помощью предложенного в них подхода невозможно найти распределение числа заявок, которые требуют повторного обслуживания. Более сложная модель IFBQ с одним сервером и ограниченной очередью изучена в [12], где входящий поток является MAP-потоком (Markov arrival process (MAP)) и время обслуживания заявок имеет функцию распределения фазового типа. Показано, что математической моделью системы будет некоторая четырехмерная цепь Маркова (ЦМ) и вычислены характеристики системы. В недавней публикации [13] рассмотрена модель IFBQ с одним сервером и двумя пуассоновскими потоками заявок. Считается, что для заявок каждого типа имеются раздельные буфера бесконечного размера, при этом лишь заявки высокого приоритета могут, согласно схеме Бернулли, повторять запросы для обслуживания. Предполагается, что заявки, которые требуют повторного обслуживания, направляются в очередь заявок низкого приоритета, при этом изучаются модели с относительными и абсолютными приоритетами. В [12, 13] для исследования предложенных моделей используется матрично-геометрический метод [14].

Анализ доступных работ показал, что они базируются на ряде допущений. Основными из них являются следующие: (1) первичные и повторные заявки идентичны по всем показателям; (2) заявки могут лишь один раз повторять запрос для обслуживания; (3) заявки обоих типов передаются единым сервером.

Однако во многих реальных системах эти допущения не выполняются, поэтому в целях повышения адекватности моделей в настоящей статье изучается IFBQ, в которой указанные выше допущения не выполняются. При этом считается, что сервера специализированы по типу заявок и они являются гетерогенными, т.е. имеют различные скорости обслуживания (ранее системы с гетерогенными серверами при отсутствии обратной связи были изучены во многих работах, см., например, [15] и список ее литературы). Кроме того, в отличие от стандартного допущения о том, что входящий поток является пуассоновским, здесь изучается модель с пуассоновским потоком с Марковской модуляцией (Markov modulated Poisson process, MMPP-поток) [16]. Отметим, что для изучения предложенной модели используется метод иерархического укрупнения состояний многомерных ЦМ [17]. Применение данного метода позволяет разработать эффективные численные процедуры расчета характеристик представленной системы.

Статья имеет следующую структуру. В разд. 1 описана изучаемая система и дана постановка задачи. В разд. 2 приведена математическая модель системы в виде 3DMC и определена ее производящая матрица. Получено условие эргодичности модели и разработан приближенный метод расчета стационарных вероятностей состояний и характеристик системы. В разд. 3 приводятся результаты численных экспериментов.

**1. Описание системы с мгновенной обратной связью и постановка задачи.** Структурная схема изучаемой системы показана на рис. 1. На вход одноканальной системы с неограниченным буфером поступает MMPP-поток с параметрами $(\Sigma, \Lambda)$, где $\Sigma = \|\sigma_{ij}\|$ – производящая матрица (ПМ) ЦМ с $N > 1$ возможными состояниями, которая управляет интенсивностью входящего потока, а вектор $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_N)$ задает значения интенсивностей входящего потока. Это означает, что $\sigma_{ij}$ определяет интенсивность перехода из состояния $i$ в состояние $j$, где

$$j \neq i; \quad \sigma_{ii} = -\sum_{j=1,\, j\neq i}^{N} \sigma_{ij}.$$

Считается, что когда ЦМ находится в состояние $n$, интенсивность входящего извне потока равна $\lambda_n, n = \overline{1, N}$, и с изменением состояния управляющей ЦМ мгновенно изменяется и интенсивность входящего потока.

Система содержит два сервера: высокоскоростной (F-сервер) и низкоскоростной (S-сервер), при этом $p$-заявки, как правило, обслуживаются в F-сервере. После завершения обслуживания $p$-заявка, согласно схеме Бернулли, либо с вероятностью $\alpha$ покидает систему, либо с вероятностью $1 - \alpha$ направляется обслуживаться в S-сервере. Заявки, которые требуют повторного обслуживания (вторичные заявки, $s$-заявки), могут образовывать очередь перед S-сервером. Считается, что $s$-заявки многократно могут потребовать повторного обслуживания, т.е. после завершения обслуживания каждая $s$-заявка независимо от других заявок, согласно схеме Бернулли, либо с вероятностью $\beta$ окончательно покидает систему, либо с дополнительной вероятностью $1 - \beta$ мгновенно требует повторного обслуживания в S-сервере. Времена обслуживания заявок в

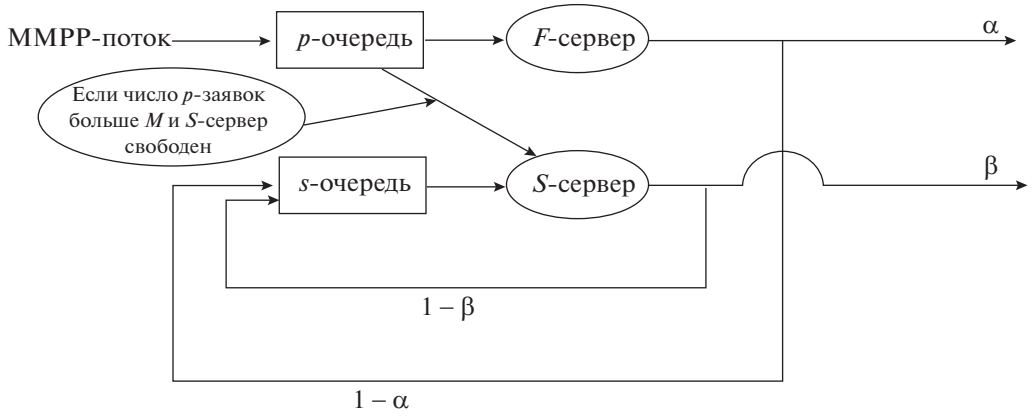


**Рис. 1.** Структурная схема системы

F-сервере и в S-сервере являются случайными величинами, которые имеют экспоненциальное распределение с параметрами $\mu_f$ и $\mu_s$ соответственно, при этом $\mu_s < \mu_f$.

Если в момент поступления $p$-заявки число заявок в очереди перед F-сервером выше определенного порогового значение $M > 0$ и при этом S-сервер не занят, то поступившая $p$-заявка либо с вероятностью $\sigma$ направляется для обслуживания в S-сервер, либо с дополнительной вероятностью $1 - \sigma$ она присоединяется в "свою" очередь. Обслуживание разнотипных заявок и изменения состояний ЦМ, которая управляет интенсивностью входящего потока, являются независимыми друг от друга случайными процессами.

Задача состоит в нахождении совместного распределения состояний MMPP-потока и числа заявок каждого типа в системе. Решение этой задачи позволит найти характеристики системы – среднее число $p$-заявок $(L_p)$ и $s$-заявок $(L_s)$ в системе; интенсивность $p$-заявок, которые обслуживаются в S-сервере $(R_{ps})$.

**2. Расчет стационарного распределения системы и ее характеристики.** Состояние системы определяется трехмерным вектором $(n, k, r)$, где $n$ – состояние управляющей интенсивностью входящего потока ЦМ, $k$ – суммарное число заявок перед F-сервером и заявки в нем, $r$ – суммарное число заявок перед S-сервером и заявки в нем. Тогда пространство состояний этой трехмерной ЦМ определяется как декартово произведение трех множеств:

$$E = \{\overline{1, N}\} \times \{0, 1, \ldots,\} \times \{0, 1, \ldots,\}. \tag{2.1}$$

Интенсивность перехода из состояния $(n, k, r)$ в состояние $(n', k', r')$ обозначим через $q((n, k, r), (n', k', r'))$. Эти величины вычисляются следующим образом:

– переходы $(n, k, r) \to (n', k, r)$, $n' \neq n$ осуществляются с интенсивностью $\sigma_{n,n'}$ при изменении состояния MMPP-потока;

– переходы $(n, k, r) \to (n, k+1, r)$, $r > 0$, и $(n, k, 0) \to (n, k+1, 0)$, $k < M$, осуществляются с интенсивностью $\lambda_n$ при поступлении $p$-заявки;

– переходы $(n, k, 0) \to (n, k+1, 0)$, $k \geq M$, осуществляются с интенсивностью $\lambda_n(1 - \sigma)$ при поступлении $p$-заявки;

– переходы $(n, k, 0) \to (n, k+1, 1)$, $k \geq M$, осуществляются с интенсивностью $\lambda_n \sigma$ при поступлении $p$-заявки;

– переходы $(n, k, r) \to (n, k-1, r)$, $k > 0$, осуществляются с интенсивностью $\mu_f \alpha$ при завершении обслуживания $p$-заявки в F-сервере и ухода ее из системы;

– переходы $(n, k, r) \to (n, k-1, r+1)$, $k > 0$, осуществляются с интенсивностью $\mu_f(1 - \alpha)$ при завершении обслуживания $p$-заявки в F-сервере и возвращении в S-сервер для повторного обслуживания;

– переходы $(n,0,r) \to (n,0,r-1)$, $r>0$, осуществляются с интенсивностью $\mu_s\beta$ при завершении обслуживания заявки в S-сервере.

Следовательно, положительные элементы ПМ данной трехмерной ЦМ определяются из следующих соотношений:

$$q\big((n,k,r),(n',k',r')\big) = \begin{cases} \sigma_{nn'}, & \text{если} \quad n' \neq n, \quad k' = k, \quad r' = r, \\ \lambda_n, & \text{если} \quad r > 0, \quad n' = n, \quad k' = k+1, \quad r' = r \\ \text{или} & k < M, \quad r = 0, \quad n' = n, \quad k' = k+1, \quad r' = r, \\ \lambda_n\sigma, & \text{если} \quad r = 0, \quad n' = n, \quad k' = k, \quad r' = 1, \\ \lambda_n(1-\sigma), & \text{если} \quad k \geq M, \quad r = 0, \quad n' = n, \quad k' = k+1, \quad r' = r, \\ \mu_f\alpha, & \text{если} \quad k > 0, \quad n' = n, \quad k' = k-1, \quad r' = r, \\ \mu_f(1-\alpha), & \text{если} \quad k > 0, \quad n' = n, \quad k' = k-1, \quad r' = r+1, \\ \mu_s\beta, & \text{если} \quad k = 0, \quad r > 0, \quad n' = n, \quad k' = 0, \quad r' = r-1. \end{cases} \tag{2.2}$$

Стационарную вероятность состояния $(n,k,r) \in E$ обозначим через $p(n,k,r)$. Условие существования стационарного режима получено ниже.

Нахождения вероятностей состояний являются достаточным для вычисления характеристики изучаемой системы. Так, среднее число $p$-заявок ($L_p$) и $s$-заявок ($L_s$) в системе определяются как математическое ожидание соответствующих случайных величин:

$$L_p = \sum_{n=1}^{N}\sum_{k=1}^{\infty} k \sum_{r=0}^{\infty} p(n,k,r); \tag{2.3}$$

$$L_s = \sum_{n=1}^{N}\sum_{r=1}^{\infty} r \sum_{k=0}^{\infty} p(n,k,r). \tag{2.4}$$

Интенсивность $p$-заявок, которые обслуживаются в S-сервере $(R_{ps})$, определяется как

$$R_{ps} = \sigma \sum_{n=1}^{N} \lambda_n \sum_{k=M}^{\infty} p(n,k,0). \tag{2.5}$$

Использование метода многомерных производящих функций для нахождения вероятностей состояний сталкивается с рядом методологических и вычислительных трудностей. Исходя из этого, ниже приводится альтернативный метод решения данной задачи, основанный на иерархическом методе фазового укрупнения состояний многомерных ЦМ [17].

Предложенный метод можно корректно применить для систем, в которых ММРР-поток является инерционным, т.е. в достаточно больших интервалах времени интенсивность входящего потока является постоянной величиной.

При выполнении указанного выше допущения рассмотрим следующее расщепление пространства состояний (2.1):

$$E = \bigcup_{n=1}^{N} E_n, \quad E_n \cap E_{n'} = \varnothing, \quad \text{если} \quad n \neq n', \tag{2.6}$$

где $E_n = \big\{(n,k,r) \in E\colon k = 0,1,\ldots;\ r = 0,1,\ldots\big\}$, $n = \overline{1,N}$.

Все состояния из класса $E_n$ объединяются в одно укрупненное состояние $\langle n\rangle$, и на основе расщепления (2.6) в пространстве состояний (2.1) определяется функция укрупнения $U_1(n,k,r) = \langle n\rangle$, $(n,k,r) \in E_n$. Множество укрупненных состояний $\langle n\rangle$ обозначим через $\Omega_1 = \{\langle n\rangle\colon n = \overline{1,N}\}$. Тогда приближенные значения вероятностей состояний исходной модели, обозначаемые через $\tilde{p}(n,k,r)$, определяются как (см. [17])

$$\tilde{p}(n,k,r) = \rho_n(k,r)\pi_1(\langle n\rangle), \tag{2.7}$$

где $\rho_n(k,r)$ – вероятность состояния $(k, r)$ внутри расщепленной модели с пространством состояний $E_n$, $\pi_1(\langle n\rangle)$ – вероятность укрупненного состояния $\langle n\rangle \in \Omega_1$.

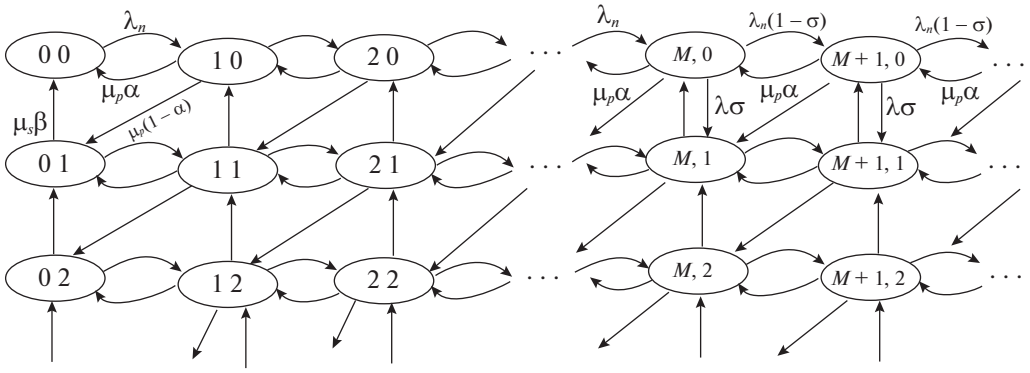


**Рис. 2.** Граф переходов между состояниями в расщепленной модели с пространством состояний $E_n$

Как было отмечено выше, переходы между состояниями управляющей интенсивностью входящего потока ЦМ не зависят от статусов F-сервера и S-сервера, поэтому вероятности состояний $\pi_1(\langle n\rangle)$, $\langle n\rangle \in \Omega_1$, определяются с помощью ее ПМ $\Sigma$.

Следовательно, для нахождения вероятности состояний исходной модели потребуется лишь определение стационарных распределений двумерных ЦМ с пространствами состояний $E_n$, $n = \overline{1, N}$ (см. формулу (2.7)). Для решения задачи к этим двумерным ЦМ применяется процедура укрупнения (второй уровень иерархии). Поскольку все расщепленные модели с пространствами состояний $E_n$ идентичны, то зафиксируем значение параметра $n$, $n = \overline{1, N}$. Граф переходов между состояниями $E_n$ показан на рис. 2.

В классе $E_n$ рассмотрим следующее разбиение:

$$E_n = \bigcup_{r=0}^{\infty} E_n^r, \quad E_n^r \cap E_n^{r'} = \varnothing, \quad \text{если } r \neq r', \tag{2.8}$$

где $E_n^r = \{(k,r) \in E_n\colon k = 0,1,\ldots\}$, $r = 0,1,\ldots$

Все состояния из класса $E_n^r$ объединяются в одно укрупненное состояние $\langle r\rangle$, и на основе расщепления (2.8) в пространстве состояний $E_n$ определяется функция укрупнения $U_2(k,r) = \langle r\rangle$, $(k,r) \in E_n^r$. Множество укрупненных состояний $\langle r\rangle$ обозначим через $\Omega_2 = \{\langle r\rangle\colon n = 0,1,2,\ldots\}$.

Согласно [17], имеем

$$\rho_n(k,r) = \rho_n^r(k)\,\pi_2^n(\langle r\rangle), \tag{2.9}$$

где $\rho_n^r(k)$ – вероятность состояния $(k, r)$ внутри расщепленной модели с пространством состояний $E_n^r$, $\pi_2^n(\langle r\rangle)$ – вероятность укрупненного состояния $\langle r\rangle \in \Omega_2$.

В классе $E_n^r$, $r = 0,1,\ldots$, в векторе состояний вторая компонента является постоянной и равна $r$. Поэтому при изучении расщепленных моделей с пространством $E_n^r$ каждое состояние $(k,r)$ может быть задано лишь первой компонентой, т.е. далее состояние $(k,r) \in E_n^r$ обозначается как $k$, $k = 0,1,\ldots$

Тогда из соотношений (2.2) заключаем, что интенсивности переходов между состояниями расщепленной модели с пространством состояний $E_n^r$, $r = 0,1,\ldots$, зависят от параметра $r$, $r = 0,1, \ldots$, и определяются следующим образом (см. рис. 2).

Для случая $r = 0$:

$$q_n^0(k,k') = \begin{cases} \lambda_n, & \text{если} \quad k < M, \quad k' = k+1, \\ \lambda_n(1-\sigma), & \text{если} \quad k \geq M, \quad k' = k+1, \\ \mu_f\alpha, & \text{если} \quad k' = k-1. \end{cases} \tag{2.10}$$

Для случаев $r > 0$:

$$q_n^r(k,k') = \begin{cases} \lambda_n, & \text{если} \quad k' = k+1, \\ \mu_f\alpha, & \text{если} \quad k' = k-1. \end{cases} \tag{2.11}$$

Из (2.10) получаем, что при выполнении условия $\nu_n < (1-\sigma)^{-1}$, $\nu_n = \lambda_n/\mu_f\alpha$, вероятности состояний расщепленной модели с пространством состояний $E_n^0$ определяются так:

$$\rho_n^0(k) = \begin{cases} \nu_n^k \rho_n^0(0), & \text{если} \quad 0 \le k \le M, \\ \dfrac{1}{(1-\sigma)^M}(\nu_n(1-\sigma))^k \rho_n^0(0), & \text{если} \quad k > M, \end{cases} \tag{2.12}$$

где $\rho_n^0(0)$ находится из условия нормировки, т.е.

$$\sum_{k=0}^{\infty} \rho_n^0(k) = 1.$$

После стандартных преобразований получим

$$\rho_n^0(0) = \left( \frac{1-\nu_n^{M+1}}{1-\nu_n} + \nu_n^{M+1} \frac{1-\sigma}{1-\nu_n(1-\sigma)} \right)^{-1}. \tag{2.13}$$

Замечание 1. Поскольку условие $\nu_n < (1-\sigma)^{-1}$ верно для каждого $n, n = \overline{1,N}$, то получаем, что должно выполняться условие

$$\max_{n=1,N}\{\nu_n\} < (1-\sigma)^{-1}. \tag{2.14}$$

Из (2.11) получаем, что если $\nu_n < 1$, то вероятности состояний расщепленных моделей с пространствами состояний $E_n^r$ не зависят от индекса $r, r = 0,1,\ldots$, и определяются как

$$\rho_n^r(k) = (1-\nu_n)\nu_n^k, \quad k = 0,1,\ldots \tag{2.15}$$

Замечание 2. Поскольку условие $\nu_n < 1$ должно быть верно для каждого $n, n = \overline{1,N}$, то должно выполняться условие

$$\max_{n=1,N}\{\nu_n\} < 1. \tag{2.16}$$

Тогда, объединяя (2.14) и (2.16), получаем первое условие эргодичности модели, т.е. должно выполняться (2.16).

Учитывая (2.2), (2.12), (2.13) и (2.15), интенсивности переходов между состояниями $\Omega_2$ определяются следующим образом (см. рис. 2):

$$q_n(\langle r\rangle, \langle r'\rangle) = \begin{cases} \eta_n(0,1), & \text{если} \quad r = 0, \quad r' = 1, \\ \mu_f(1-\alpha)\nu_n, & \text{если} \quad r > 0', \quad r' = r+1, \\ \mu_s\beta, & \text{если} \quad r > 0, \quad r' = r-1. \end{cases} \tag{2.17}$$

Согласно (2.12) и (2.13), величина $\eta_n(0,1)$ в формуле (2.17) определяется как

$$\eta_n(0,1) = \mu_f(1-\alpha)\sum_{k=1}^{\infty}\rho_0^0(k) + \lambda_n\sigma\sum_{k=M}^{\infty}\rho_n^0(k) = \mu_f(1-\alpha)\nu_n +$$

$$+ \lambda_n\sigma\rho_n^0(0)\frac{1}{(1-\sigma)^M}\sum_{k=M}^{\infty}(\nu_n(1-\sigma))^k = \mu_p(1-\alpha)\nu_n + \lambda_n\sigma\rho_n^0(0)\frac{\nu_n^M}{1-\nu_n(1-\sigma)}.$$

Из (2.17) получаем, что при выполнении условия $\psi_n < 1$, $\psi_n = \mu_f(1-\alpha)\nu_n/\mu_s\beta$, вероятности состояний укрупненной модели с пространством состояний $\Omega_2$ находятся так (см. рис. 2):

$$\pi_2^n(\langle r\rangle) = \begin{cases} \theta_n \pi_2^n(\langle 0\rangle), & \text{если} \quad r = 1, \\ \theta_n \psi_n^{r-1} \pi_2^n(\langle 0\rangle), & \text{если} \quad r > 1, \end{cases} \tag{2.18}$$

где $\theta_n = \eta_n(0,1)/\mu_s\beta$ и $\pi_2^n(\langle 0\rangle)$ определяется из условия нормировки, т.е.

$$\sum_{r=0}^{\infty} \pi_2^n(\langle r\rangle) = 1.$$

После стандартных преобразований получим

$$\pi_2^n(\langle 0\rangle) = \frac{1-\psi_n}{1-\psi_n+\theta_n}. \tag{2.19}$$

З а м е ч а н и е 3. Поскольку условие $\psi_n < 1$ верно для каждого $n, n = \overline{1,N}$, то находим второе условие эргодичности модели:

$$\max_{n=\overline{1,N}}\{\nu_n\} < \frac{\mu_s\beta}{\mu_f(1-\alpha)}. \tag{2.20}$$

Объединяя соотношения (2.16) и (2.20), получаем следующее условие эргодичности модели:

$$\max_{n=\overline{1,N}}\{\nu_n\} < \min\left\{\frac{\mu_s\beta}{\mu_f(1-\alpha)},\ 1\right\}. \tag{2.21}$$

Таким образом, при выполнении условия эргодичности модели (2.21) с учетом соотношений (2.7), (2.9), (2.12), (2.13), (2.18) и (2.19) находятся приближенные значения стационарных вероятностей состояний исходной трехмерной ЦМ.

Далее с использованием стационарных вероятностей состояний могут быть рассчитаны приближенные значения характеристик (2.3)–(2.5). Действительно, из (2.7) и (2.9) заключаем, что указанные характеристики вычисляются следующим образом:

$$L_p \approx \sum_{n=1}^{N} \pi_1(\langle n\rangle) \sum_{k=1}^{\infty} k \sum_{r=0}^{\infty} \rho_n^r(k)\, \pi_2^n(\langle r\rangle); \tag{2.22}$$

$$L_s \approx \sum_{n=1}^{N} \pi_1(\langle n\rangle) \sum_{r=1}^{\infty} r \pi_2^n(\langle r\rangle); \tag{2.23}$$

$$R_{ps} \approx \sigma \sum_{n=1}^{N} \pi_1(\langle n\rangle) \sum_{k=M}^{\infty} \rho_n^0(k)\, \pi_2^n(\langle 0\rangle). \tag{2.24}$$

Тогда после определенных математических выкладок из (2.22)–(2.24) получим

$$L_p \approx \sum_{n=1}^{N} \pi_1(\langle n\rangle)(\Psi_n^1 + \Psi_n^2 + \Psi_n^3), \tag{2.25}$$

где

$$\Psi_n^1 = \sum_{k=1}^{M} k \nu_n^k (\rho_n^0(0) + (1-\nu_n)(1-\pi_2^n(\langle 0\rangle))),$$

$$\Psi_n^2 = \rho_0^0(0)\, \pi_2^n(\langle 0\rangle) \frac{\nu_n^2}{(1-\sigma)^{M-2}} \frac{M(1-\nu_n(1-\sigma))+1}{(1-\nu_n(1-\sigma))^2},$$

$$\Psi_n^3 = (1-\pi_2^n(\langle 0\rangle)) \frac{\nu_n^2}{1-\nu_n} \frac{M(1-\nu_n)+1}{(1-\sigma)^M},$$

$$L_s \approx \sum_{n=1}^{N} \pi_1(\langle n \rangle)\, \pi_2^n(\langle 0 \rangle) \frac{\theta_n}{(1-\psi_n)^2}, \tag{2.26}$$

$$R_{ps} \approx \sigma \sum_{n=1}^{N} \frac{\nu_n^M}{1-\nu_n(1-\sigma)} \pi_1(\langle n \rangle)\, \pi_2^n(\langle 0 \rangle)\, \rho_n^0(0). \tag{2.27}$$

**3. Численные результаты.** Проводимые здесь численные эксперименты имеют две цели: 1) показать высокую точность разработанного алгоритма приближенного вычисления стационарных вероятностей состояний изучаемой системы; 2) изучить поведение характеристик системы относительно изменения порогового параметра $M$ и решить задачу выбора его оптимального (в заданном смысле) значения.

Точность приближенного алгоритма оценивается с помощью имитационного моделирования. При этом близость результатов, полученных с применением различных подходов, оценивается с помощью нормы подобия косинуса, т.е.

$$\|\mathrm{N}\| = \frac{(\mathbf{p}, \tilde{\mathbf{p}})}{|\mathbf{p}|\,|\tilde{\mathbf{p}}|}, \tag{3.1}$$

где $\mathbf{p} = \left(p(n,k,r) : (n,k,r) \in E\right)$ и $\tilde{\mathbf{p}} = \left(\tilde{p}(n,k,r) : (n,k,r) \in E\right)$ – векторы точных и приближенных значений стационарных вероятностей состояний соответственно; $(\mathbf{p}, \tilde{\mathbf{p}})$ – скалярное произведение векторов $\mathbf{p}$ и $\tilde{\mathbf{p}}$; $|\mathbf{p}|$ и $|\tilde{\mathbf{p}}|$ – длины векторов $\mathbf{p}$ и $\tilde{\mathbf{p}}$ соответственно.

Отметим, что, как правило, норма подобия косинуса используется для определения ориентации двух векторов, а не для сравнения их величин. Однако в нашем случае эта мера адекватно оценивает близость конечных точек векторов $\mathbf{p}$ и $\tilde{\mathbf{p}}$, так как, согласно нормирующему условию, имеем

$$\sum_{(n,k,r)\in E} p(n,k,r) = \sum_{(n,k,r)\in E} \tilde{p}(n,k,r) = 1.$$

Иными словами, конечные точки векторов $\mathbf{p}$ и $\tilde{\mathbf{p}}$ находятся в одной гиперплоскости.

Сначала рассмотрим случай, когда значения интенсивностей входящего потока фиксированы, а интенсивности обслуживания разнотипных серверов меняются. В этом случае в проводимых экспериментах исходные данные гипотетической модели определяются следующим образом. ПМ ЦМ, которая управляет интенсивностью входящего MMPP-потока с $N = 3$ состояниями, задается как

$$\Sigma_1 = \begin{Vmatrix} -34 & 20 & 14 \\ 18 & -32 & 14 \\ 4 & 16 & -20 \end{Vmatrix}.$$

Значения интенсивностей входящего потока $\Lambda = (15,10,5)$. Параметры схем Бернулли выбирались так: $\alpha = 0.8$; $\beta = 0.4$; $\sigma = 0.1$.

Сравнительный анализ результатов приближенного алгоритма и метода имитационного моделирования для этой серии вычислительных экспериментов показаны в табл. 1.

Нами также выполнены вычислительные эксперименты для случаев, когда и значения интенсивностей входящего потока, и интенсивности обслуживания разнотипных серверов меняются. В проводимых трех сериях экспериментов параметры схем Бернулли оставались неизменными. В первой серии экспериментов ПМ ЦМ, которая управляет интенсивностью входящего MMPP-потока с $N = 3$ состояниями, выбиралась так же, как и ранее (т.е. она равна $\Sigma_1$), при этом значения интенсивностей входящего потока выбирались как $\Lambda_1 = (10,5,3)$. Во второй и третьей сериях экспериментов ПМ ЦМ, которые управляют интенсивностями входящего MMPP-потока с $N = 3$ состояниями, и соответствующие им значения интенсивностей входящего потока выбирались следующим образом:

$$\Sigma_2 = \begin{Vmatrix} -25 & 15 & 10 \\ 13 & -23 & 10 \\ 5 & 11 & -26 \end{Vmatrix}, \quad \Lambda_2 = (15,12,7);$$

**Таблица 1.** Оценка точности алгоритма вычисления стационарных вероятностей состояний системы. Случай фиксированных значений интенсивностей входящего MMPP-потока

| $(\mu_f,\ \mu_s)$ | $M$ | Значения нормы (3.1) |
|---|---|---|
| (60, 45) | 1 | 0.9945 |
| | 2 | 0.9940 |
| | 3 | 0.9914 |
| (60, 50) | 1 | 0.9931 |
| | 2 | 0.9948 |
| | 3 | 0.9927 |
| (60, 55) | 1 | 0.9943 |
| | 2 | 0.9951 |
| | 3 | 0.9954 |
| (65, 45) | 1 | 0.9936 |
| | 2 | 0.9944 |
| | 3 | 0.9936 |
| (65, 50) | 1 | 0.9941 |
| | 2 | 0.9958 |
| | 3 | 0.9953 |
| (65, 55) | 1 | 0.9952 |
| | 2 | 0.9949 |
| | 3 | 0.9958 |
| (70, 45) | 1 | 0.9946 |
| | 2 | 0.9955 |
| | 3 | 0.9950 |
| (70, 50) | 1 | 0.9944 |
| | 2 | 0.9968 |
| | 3 | 0.9960 |
| (70, 55) | 1 | 0.9964 |
| | 2 | 0.9970 |
| | 3 | 0.9960 |

$$\Sigma_3 = \begin{Vmatrix} -27 & 12 & 15 \\ 9 & -29 & 20 \\ 8 & 15 & -23 \end{Vmatrix}, \quad \Lambda_3 = (20,10,5).$$

Результаты указанных экспериментов показаны в табл. 2. Из табл. 1 и 2 видно, что разработанный алгоритм имеет высокую точность, так как значение нормы (3.1) практически равно единице.

З а м е ч а н и е 4. Поскольку система является бесконечномерной (по второй и третьей компоненте вектора состояний), то при вычислении нормы (3.1) максимальное число заявок каждого типа в системе сверху огранивается достаточно большими конечными величинами. Такая замена является оправданной, так как при превышении этими величинами определенных (достаточно больших) значений соответствующие вероятности состояний становятся бесконечно малыми величинами, т.е. практически равными нулю.

Разработанные приближенные формулы позволяют также изучить поведения нормы подобия (3.1) относительно изменения любых (структурных и нагрузочных) параметров системы. Из-за ограниченности объема работы и для конкретности изложения здесь приводятся лишь результаты, которые показывают поведение указанной величины относительно изменения параметра $\sigma$ (рис. 3). Из графика видно, что с ростом значения этого параметра ухудшается точность

**Таблица 2.** Оценка точности алгоритма вычисления стационарных вероятностей состояний системы. Случай нефиксированных значений интенсивностей входящего MMPP-потока

| $\Lambda$ | $(\mu_f,\ \mu_s)$ | $M$ | Значения нормы (3.1) |
|---|---|---|---|
| (10, 5, 3) | (60, 45) | 1 | 0.9979 |
| | | 2 | 0.9979 |
| | | 3 | 0.9973 |
| | (65, 50) | 1 | 0.9983 |
| | | 2 | 0.9984 |
| | | 3 | 0.9971 |
| | (70, 55) | 1 | 0.9986 |
| | | 2 | 0.9986 |
| | | 3 | 0.9986 |
| (15, 12, 17) | (60, 45) | 1 | 0.9924 |
| | | 2 | 0.9934 |
| | | 3 | 0.9914 |
| | (65, 50) | 1 | 0.9953 |
| | | 2 | 0.9949 |
| | | 3 | 0.9945 |
| | (70, 55) | 1 | 0.9955 |
| | | 2 | 0.9962 |
| | | 3 | 0.9964 |
| (20, 10, 5) | (60, 45) | 1 | 0.9892 |
| | | 2 | 0.9868 |
| | | 3 | 0.9867 |
| | (65, 50) | 1 | 0.9903 |
| | | 2 | 0.9904 |
| | | 3 | 0.9917 |
| | (70, 55) | 1 | 0.9933 |
| | | 2 | 0.9938 |
| | | 3 | 0.9936 |

формул вычисления приближенных значений вероятностей состояний исходной трехмерной ЦМ. Иными словами, чем меньше значения указанного параметра, тем выше точность предложенных формул. Однако даже в наихудших случаях, когда значения этого параметра близки к

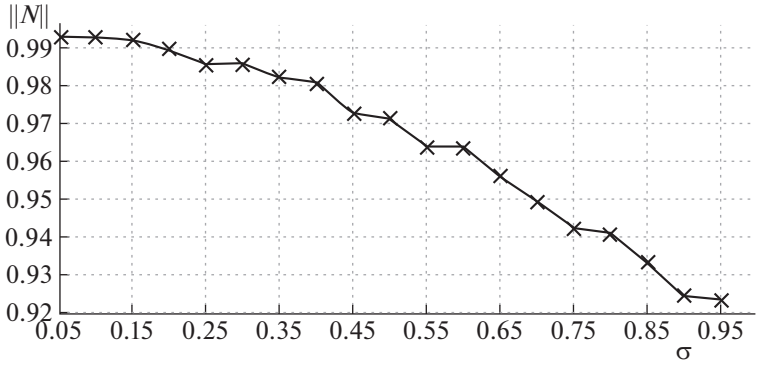


**Рис. 3.** Зависимость значения нормы подобия (3.1) от параметра $\sigma$

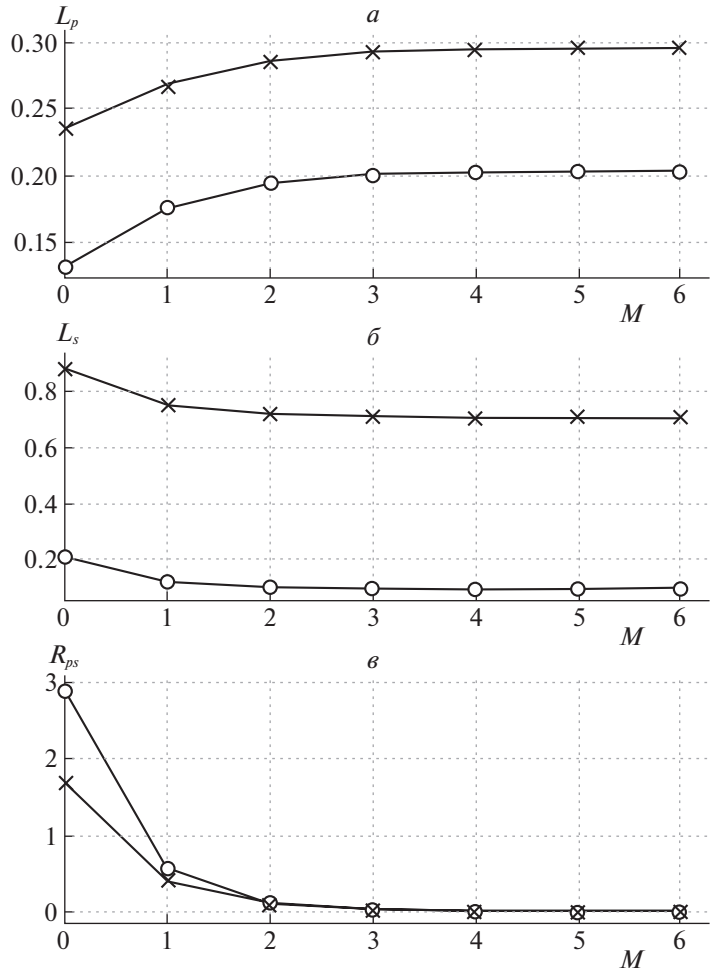


**Рис. 4.** Зависимость характеристик системы от порогового параметра $M$

единице, значения нормы (3.1) оказываются больше 0.9, т.е. разработанные приближенные формулы имеют достаточно высокую точность.

Такое поведение нормы (3.1) относительно параметра $\sigma$ имеет вполне логическое объяснение. Действительно, из теории фазового укрупнения состояний ЦМ [3] известно, что чем меньше интенсивности переходов между расщепленными классами состояний, тем выше точность алгоритмов метода фазового укрупнения. Для изучаемой здесь модели при малых значениях параметра $\sigma$ интенсивности переходов между расщепленными классами $E_n^r, r = 0,1,\ldots$ (см. формулы (2.8) и рис. 2), оказываются малыми величинами и с его ростом увеличивается интенсивность переходов между расщепленными классами.

Отметим, что аналогичные поведения нормы (3.1) наблюдаются относительно изменения параметров $\alpha$ и $\beta$, т.е. с ростом параметра $\alpha$ значение нормы систематически растет, а с ростом параметра $\beta$, наоборот, она систематически уменьшается (отметим, что в отличие от параметра $\sigma$, области изменения параметров $\alpha$ и $\beta$ определяются из условия эргодичности системы (2.21)).

Поведения характеристик системы относительно изменения порогового параметра $M$ при указанных выше параметрах MMPP-потока и при $(\mu_f, \mu_s) = (70, 55)$, $\sigma = 0.4$, показаны на рис. 4. Функция $L_p$ является неубывающая (рис. 4, *а*), что и следовало ожидать, так как с увеличением параметра $M$ растут шансы первичных заявок присоединиться к своей очереди. По той же причине функции $L_s$ (рис. 4, *б*) и $R_{ps}$ (рис. 4, *в*) являются невозрастающими.

Отметим, что с увеличением параметра $\beta$ увеличивается нагрузка S-сервера и тем самым уменьшаются шансы первичных заявок обслуживаться в этом сервере, поэтому увеличивается

**Таблица 3.** Результаты решения задачи (3.2)

| $(\mu_f,\ \mu_s)$ | $\alpha$ | $M^*$ | $TC_{\min}$ |
|---|---|---|---|
| (60, 45) | 0.7 | 6 | 2.927 |
| | 0.8 | 8 | 1.959 |
| | 0.9 | 11 | 1.482 |
| (65, 45) | 0.7 | 8 | 2.509 |
| | 0.8 | 10 | 1.708 |
| | 0.9 | 13 | 1.303 |
| (65, 50) | 0.7 | 7 | 2.389 |
| | 0.8 | 10 | 1.681 |
| | 0.9 | 13 | 1.295 |
| (70, 55) | 0.7 | 9 | 2.033 |
| | 0.8 | 12 | 1.475 |
| | 0.9 | 15 | 1.152 |

число таких заявок перед F-сервером (см. рис. 4, *а*). С увеличением параметра $\alpha$ уменьшается нагрузка S-сервера, тем самым уменьшается число таких заявок в системе (см. рис. 4, *б*). Значения функции $R_{ps}$ при различных соотношениях параметров $\alpha$ и $\beta$ почти не отличаются друг от друга (см. рис. 4, *в*).

Заметим, что для выбранных исходных данных при $M \geq 3$ изучаемые функции становятся почти постоянными. Это обстоятельство позволяет ставить задачу нахождения таких значений этого параметра, чтобы минимизировать суммарные штрафы (total cost, (TC)), связанные с пребыванием разнотипных заявок в очереди и обслуживанием первичных заявок в низкоскоростном сервере. Так, пусть штрафы за пребывания одной первичной и повторной заявки в системе равны $c_p$ и $c_s$, а штраф за обслуживания $p$-заявки в S-сервере равен $c_{ps}$. Тогда задача оптимизации ставится так:

$$M^* = \arg\min_M TC, \tag{3.2}$$

где $TC = c_p L_p + c_s L_s + c_{ps} R_{ps}$ .

Результаты решения задачи (3.2) для следующих исходных данных показаны в табл. 3:

$$\Sigma_4 = \begin{Vmatrix} -38 & 20 & 18 \\ 6 & -22 & 16 \\ 6 & 16 & -22 \end{Vmatrix}; \quad \Lambda_4 = (30,10,5); \quad \beta = 0.4; \quad \sigma = 0.1; \quad c_p = 4,5; \quad c_s = 1; \quad c_{ps} = 2.$$

К сожалению, не удается решить задачу (3.2) аналитически из-за сложного вида функционала. Потому и не удается делать какие-то общие выводы относительно решения задачи (3.2). Вместе с тем в проводимых численных экспериментах существует оптимальное решение задачи (3.2), при этом указанное в табл. 3 оптимальное значение $M^*$ является минимальным значением этого параметра, при котором достигается минимум функционала TC, т.е. при $M \geq M^*$ он становится постоянным.

**Заключение.** В работе изучается модель системы с гетерогенными серверами, MMPP-потоком и мгновенной обратной связью. После завершения обслуживания в высокоскоростном сервере первичные заявки, согласно схеме Бернулли, либо покидают систему, либо мгновенно требуют повторного обслуживания. Повторные заявки обслуживаются в S-сервере, при этом после завершения обслуживания повторные заявки могут многократно повторяться. Если в момент поступления первичной заявки S-сервер свободен и число заявок в очереди перед F-сервером превышает определенное пороговое значение, то поступившая заявка, согласно схеме Бернулли, либо направляется для обслуживания в S-сервер, либо присоединяется в свою очередь. Считается, что первичные и повторные заявки могут образовать очереди бесконечной длины.

Показано, что математической моделью изучаемой системы является некоторая трехмерная ЦМ с бесконечномерным пространством состояний. Найдено условие эргодичности модели и

разработан приближенный алгоритм расчета вероятностей состояний соответствующей ЦМ. С помощью численных экспериментов показана высокая точность рассмотренного алгоритма и решена задача нахождения оптимального значения введенного порогового параметра, чтобы минимизировать суммарные штрафы, связанные с пребыванием разнотипных заявок в очереди и обслуживанием первичных заявок в S-сервере.

## СПИСОК ЛИТЕРАТУРЫ


1. *Takacs L.* A Single Server Queue with Feedback // Bell System Technical J. 1963. V. 42. P. 505–519.
2. *Takacs L.* A Queuing Model with Feedback // Operations Research. 1977. V. 11. P. 345–354.
3. *Melikov A.Z., Ponomarenko L.A., Rustamov A.M.* Methods for Analysis of Queuing Models with Instantaneous and Delayed Feedbacks // Communications in Computer and Information sciences. 2015. V. 564. P. 185–199.
4. *Sharma S.K., Kumar R.* A Markovian Feedback Queue with Retention of Reneged Customers // Advanced Modeling and Optimization. 2012. V. 14. Iss. 3. P. 673–680.
5. *Sharma S.K., Kumar R.* A Markovian Feedback Queue with Retention of Reneged Customers and Balking // Advanced Modeling and Optimization. 2012. V. 14. Iss. 3. P. 681–688.
6. *Sharma S.K., Kumar R.* M/M/1 Feedback Queueing Model with Retention of Reneged Customers and Balking // American J. Operational Research. 2013. V. 3. Iss. 3 (2A). P. 1–6.
7. *Sharma S.K., Kumar R.* A Single Server Markovian Feedback Queuing System with Discouraged Arrivals and Retention of Reneged Customers // American J. Operational Research. 2013. V. 4. Iss. 3. P. 35–39.
8. *Kumar R., Jain N.K., Som B.K.* Optimization of an M/M/1/N Feedback Queue with Retention of Reneged Customers // Operations Research and Decisions. 2014. V. 24. Iss. 3. P. 45–58.
9. *Santkumaran A., Thangaraj V.* A Single Server Queue with Impatient and Feedback Customers // Information and Management Science. 2000. V. 11. Iss. 3. P. 71–79.
10. *Som B.K., Seth S.* M/M/c/N Queuing System with Encouraged Arrivals, Reneging, Retention and Feedback Customers // Yugoslav J. of Operations Research. 2018. V. 28. Iss. 3. P. 333–344.
11. *Bouchentouf A.A., Kadi M., Rabhi A.* Analysis of Two Heterogeneous Server Queuing Model with Balking, Reneging and Feedback // Mathematical Sciences and Applications E-notes. 2013. V. 2. Iss. 2. P. 10–21.
12. *Dudin A.N., Kazimirsky A.V., Klimenok V.I., Breuer L., Krieger U.* The Queuing Model MAP/PH/1/N with Feedback Operating in a Markovian Random Environment // Austrian J. Statistics. 2005. V. 34. Iss. 2. P. 101–110.
13. *Krishnamoorthy A., Manjunath A.S.* On Queues with Priority Determined by Feedback // Calcutta Statistical Association Bulletin. 2018. V. 70. Iss. 1. P. 33–56.
14. *NeutsM.F.* Matrix-geometric solutions in stochastic models. An algorithmic approach. Baltimore: John Hopkins University Press, 1981. 332 p.
15. *Меликов А.З., Мехбалыева Э.В.* Анализ и оптимизация систем с гетерогенными серверами и скачкообразными приоритетами // Изв. РАН. ТиСУ. 2018. № 5. С. 56–74.
16. *Fisher W., Meier-Hellstern K.* The Markov Modulated Poisson Process (MMPP) cookbook // Performance Evaluation. 1992. V. 18. P. 149–171.
17. *Melikov A. Z., PonomarenkoL.A., RustamovA.M.* Hierarchical Space Merging Algorithm to Analysis of Open Tandem Queuing Networks // Cybernetics and System Analysis. 2016. V. 52. Iss. 6. P. 867–877.