УДК 519.872
DOI: 10.17223/19988605/53/3

А.З. Меликов, Э.В. Мехбалыева

ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ СИСТЕМЫ С ГЕТЕРОГЕННЫМИ СЕРВЕРАМИ И РАНДОМИЗИРОВАННОЙ *N*-ПОЛИТИКОЙ

Изучаются марковские модели систем с двумя гетерогенными серверами. Предлагается рандомизированная *N*-политика включения медленного сервера, согласно которой при достижении длины очереди заявок величины *N* медленный сервер включается с определенной вероятностью. Рассматриваются модели двух типов: с неограниченным и ограниченным размером буфера для ожидания заявок. Разработан численный метод расчета распределения вероятностей состояний и характеристик изучаемых систем и решена задача их оптимизации. Даны результаты численных экспериментов.
Ключевые слова: система с гетерогенными серверами; рандомизированная *N*-политика; численный анализ.

Одним из основных допущений в классических моделях систем массового обслуживания (СМО) является то, что серверы являются идентичными. Однако это допущение плохо соотносится с действительностью, так как в реальных СМО серверы отличаются друг от друга по различным показателям, например по скорости обработки заявок, надежности, стоимости эксплуатации и т.д.

В литературе СМО с неидентичными серверами получили название систем с гетерогенными серверами (Queue with Heterogeneous Servers, QHS). Такие системы часто встречаются при моделировании компьютерных и телекоммуникационных систем, так как в процессе их расширения приходится использовать гетерогенные компьютеры (серверы). Кроме компьютерных и телекоммуникационных систем серверы с различными скоростями встречаются в колл-центрах с многими операторами, которые имеют различные квалификации, а также в производственных системах, где в процессе обслуживания заявок участвуют не машины, а люди.

Первая серьезная работа, посвященная изучению моделей QHS, описана в [1]. В ней изучается марковская СМО с бесконечной очередью, в которой для обслуживания заявок с равными вероятностями назначается один из свободных серверов (эта схема называется рандомизированным доступом). Предложен алгоритм вычисления стационарных вероятностей состояний системы, а также найдены формулы для нахождения среднего числа заявок в системе. Доказано, что из результатов данной работы в частных случаях получаются классические результаты для СМО с идентичными серверами.

Анализ доступной литературы показал, что в подавляющем большинстве работ изучены проблемы расчета характеристик QHS, в которых приняты схемы рандомизированного [1–5] и упорядоченного доступа [6–14]. Обзор этих работ можно найти в [15, 16].

В QHS важными являются проблемы определения оптимальных стратегий доступа заявок, а также схем включения серверов в зависимости от текущего состояния системы. Отметим, что в целях оптимизации QHS зачастую используется FSF-схема доступа (Fast Server First), т.е. в момент поступления заявки для обслуживания всегда выбирается сервер, который имеет наивысшую скорость среди свободных серверов. Это объясняется тем, что вероятность потери в системе $M/M/k/k$ имеет минимальное значение при использовании FSF-схемы доступа [17]. Однако FSF-схема доступа не всегда является оптимальной и даже субоптимальной для QHS с очередями. Так, в работе [18] доказано, что для QHS с неограниченной очередью оптимальной для минимизации среднего числа заявок в системе является *N*-политика. Согласно данной политике в двух серверных QHS быстрый сервер работает всегда, если в системе имеется хоть одна заявка, а медленный сервер включается лишь тогда, когда длина очереди достигает определенного (порогового) значения *N*. С использованием

разных подходов этот результат доказан различными авторами (см., напр.: [19–24]). Справедливости ради отметим, что задачи оптимизации СМО с однотипными резервными серверами (homogeneous servers), в которых моменты включения и отключения резервных серверов зависят от длины очереди, были изучены еще в 70-е гг. прошлого века [25] (после выхода в свет этой монографии в русскоязычной литературе появилась серия публикаций на эту тему; заинтересованный читатель может найти их в интернет-ресурсах).

В дальнейшем в работах [26–28] с помощью известного алгоритма Ховарда эти результаты были распространены на системы с дополнительной структурой штрафов, а в работе [29] исследована система с повторными заявками в предположении, что интервалы между их поступлениями и длительности их обслуживания имеют распределения фазового типа. Обобщения этой схемы для моделей QHS с ненадежными серверами рассмотрены в [30, 31].

В настоящей работе изучается QHS с двумя гетерогенными серверами и предлагается рандомизированная N-политика включения медленного сервера, согласно которой при достижении длины очереди заявок величины N медленный сервер включается с определенной вероятностью, а с дополнительной вероятностью он остается в спящем режиме. Данная политика является обобщением классической N-политики. Кроме того, здесь предлагается численный метод нахождения стационарных вероятностей состояний системы, а также вычисляются характеристики системы с неограниченным и ограниченным размером буфера для ожидания заявок. Отметим, что время выполнения алгоритма реализации предложенного метода практически равно нулю.

1. Описание моделей и постановка задачи

Рассмотрим QHS с неограниченным буфером для ожидания заявок, которая содержит два сервера: быстрый (F-сервер) и медленный (S-сервер). Здесь приняты следующие допущения:

(i) В систему поступает пуассоновский поток идентичных заявок с интенсивностью λ, причем эта величина не зависит от состояния серверов.

(ii) Время обслуживания заявок в обоих серверах являются случайной величиной с показательной функцией распределения; средние времена обслуживания в F-сервере и S-сервере равны μ_F^{-1} и μ_S^{-1} соответственно, при этом $\mu_F > \mu_S$.

(iii) F-сервер всегда является активным, а S-сервер может включатся лишь тогда, когда длина очереди не меньше определенной пороговой величины N, $N < \infty$. Схемы включения и отключения S-сервера определяются следующим образом. Если в момент поступления заявки длина очереди не меньше величины N, то S-сервер включаетя с вероятностью α, $0 < \alpha < 1$ и одна заявка выбирается из очереди для обслуживания в этом сервере; с дополнительной вероятностью $1-\alpha$ он остается в спящем режиме, т.е. в отключенном состоянии. По завершении обслуживания заявки в S-сервере он выбирает для обслуживания одну заявку из очереди, если длина очереди в этот момент больше величины N; иначе S-сервер уходит в спящий режим, т.е. отключается.

(iv) В системе с ограниченным буфером поступившая заявка теряется с вероятностью, равной единице, если в этот момент буфер заполнен полностью и S-сервер является включенным; если в момент поступления заявки буфер заполнен полностью и S-сервер является отключенным, то либо с вероятностью α S-сервер включается либо с дополнительной вероятностью $1-\alpha$ поступившая заявка теряется.

Задача состоит в нахождении совместного распределения числа заявок в системе и статуса S-сервера, а также в разработке методов вычисления ее характеристик.

2. Методы расчета вероятностей состояний и характеристик изучаемых систем

Сначала рассмотрим модель QHS с неограниченным буфером. Состояние этой системы в произвольный момент времени определяется двумерным вектором (n,k), где n – число заявок в системе,

k – состояние S-сервера, т.е. $k = 0$, если S-сервер является отключенным, и $k = 1$ в противном случае. Следовательно, работа системы описывается двумерной цепью Маркова (ЦМ) со следующим пространством состояний:

$$E = E_0 \bigcup E_1 \text{, где } E_k = \{(n,k): n = 0,1,2,...\}, k = 0,1. \quad (1)$$

Элементы производящей матрицы этой ЦМ обозначаются через $q((n,k),(n',k'))$, т.е. $q((n,k),(n',k'))$ обозначает интенсивность перехода из состояния (n,k) в состояние (n',k').

Исходя из описанного механизма работы системы заключаем, что эти величины определяются следующим образом (рис. 1):

– случаи $(n,0) \in E_0$:

$$q((n,0),(n',k')) = \begin{cases} \lambda, & \text{если } n < N,\ n' = n+1,\ k' = 0, \\ \lambda(1-\alpha), & \text{если } n \geq N,\ n' = n+1,\ k' = 0, \\ \lambda\alpha, & \text{если } n \geq N,\ n' = n,\ k' = 1, \\ \mu_F, & \text{если } n > 0,\ n' = n-1,\ k' = 0; \end{cases} \quad (2)$$

– случаи $(n,1) \in E_1$:

$$q((n,1),(n',k')) = \begin{cases} \lambda, & \text{если } n' = n+1,\ k' = 1, \\ \mu_F, & \text{если } 0 < n \leq N,\ n' = n-1,\ k' = 1, \\ \mu_F + \mu_S, & \text{если } n > N,\ n' = n-1,\ k' = 1, \\ \mu_S, & \text{если } n \leq N,\ n' = n,\ k' = 0. \end{cases} \quad (3)$$

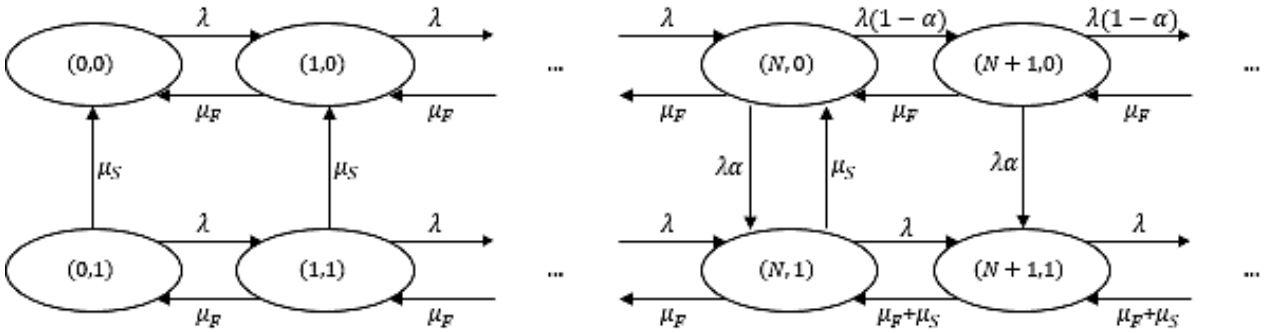


Рис. 1. Граф переходов между состояниями системы
Fig. 1. Graph of transitions between states of the systems

Стационарную вероятность состояния $(n,k) \in E$ обозначим через $p(n,k)$. Условием существования стационарного режима является $\lambda < \mu_F + \mu_S$.

Основными характеристиками данной модели QHS являются среднее число заявок в системе (L_s), среднее время пребывания заявки в системе (W_s) и интенсивность включения S-сервера (RS). Среднее число заявок в системе определяются как математическое ожидание соответствующей случайной величины, т.е.

$$L_s = \sum_{n=1}^{\infty} n \sum_{k=0}^{1} p(n,k). \quad (4)$$

Среднее время пребывания заявок вычисляется с помощью известной формулы Литтла:

$$W_s = \frac{1}{\lambda} L_s. \quad (5)$$

Поскольку S-сервер включается с вероятностью α, то если в момент поступления заявки число заявок в системе не меньше величины N, то интенсивность включения S-сервера вычисляется следующим образом:

$$RS = \lambda\alpha \sum_{n=N}^{\infty} p(n,0). \tag{6}$$

Для нахождения стационарных вероятностей состояний $p(n,k)$ может быть использован метод производящих функций. Однако его применение связано с определенными методологическими и техническими трудностями из-за сложной структуры производящей матрицы изучаемой цепи Маркова. Поэтому рассмотрим приближенный метод расчета вероятностей состояний и характеристик системы, основанный на принципах фазового укрупнения состояний многомерных цепей Маркова [15, 16]. Применение этого метода в изучаемой системе является корректным, так как согласно нашим допущениям скорость работы F-сервера много больше, чем скорость работы S-сервера. Кроме того, предположим, что $\alpha << 1-\alpha$, т.е. $\alpha << 0{,}5$. Тогда в период работы изучаемая система долгое время пребывает в состояниях из классов E_0 и E_1, и редко происходят переходы между этими классами (см. рис. 1).

Исходя из этого факта, все состояния внутри каждого класса E_k объединяются в одно укрупненное состояние $<k>$, и, таким образом, определяется множество укрупненных состояний $\Omega = \{<k> : k = 0,1\}$.

Приближенные значения вероятностей состояний $\tilde{p}(n,k)$, $(n,k) \in E$ исходной модели определяются как [15, 16]

$$\tilde{p}(n,k) = \rho_k(n)\pi(<k>), \tag{7}$$

где $\rho_k(n)$ – вероятность состояния (n,k) внутри расщепленной модели с пространством состояний E_k, а $\pi(<k>)$ – вероятность укрупненного состояния $<k> \in \Omega$.

Из соотношений (2) заключаем, что интенсивности переходов между состояниями расщепленной модели с пространством состояний E_0 определяются следующим образом:

$$q_0(n,n') = \begin{cases} \lambda, & \text{если } n < N,\ \ n' = n+1, \\ \lambda(1-\alpha), & \text{если } n \ge N,\ \ n' = n+1, \\ \mu_F, & \text{если } n' = n-1. \end{cases} \tag{8}$$

Из (8) получаем, что если выполняется условие $\tilde{\nu}_F < 1$, где $\nu_F = \lambda/\mu_F$, $\tilde{\nu}_F = (1-\alpha)\nu_F$, то вероятности состояний расщепленной модели с пространством состояний E_0 вычисляются как

$$\rho_0(n) = \begin{cases} \nu_F^n \rho_0(0), & \text{если } 0 \le n \le N, \\ (1-\alpha)^{-N} \tilde{\nu}_F^n \rho_0(0), & \text{если } n > N, \end{cases} \tag{9}$$

где $\rho_0(0)$ определяется из условия нормировки, т.е. $\rho_0(0) = \left(\sum_{n=0}^{N} \nu_F^n + \nu_F^N \frac{\tilde{\nu}_F}{1-\tilde{\nu}_F} \right)^{-1}$.

Аналогичным образом из соотношений (3) заключаем, что интенсивности переходов между состояниями расщепленной модели с пространством состояний E_1 определяются следующим образом (см. рис. 1):

$$q_1(n,n') = \begin{cases} \lambda, & \text{если } n' = n+1, \\ \mu_F, & \text{если } n \le N,\ \ n' = n-1, \\ \mu_F + \mu_S, & \text{если } n > N,\ \ n' = n-1. \end{cases} \tag{10}$$

Поскольку выполняется условие $\nu_{FS} < 1$, где $\nu_{FS} = \lambda/(\mu_F + \mu_S)$, то из (10) получаем, что вероятности состояний расщепленной модели с пространством состояний E_1 находятся как

$$\rho_1(n) = \begin{cases} \nu_F^n \rho_1(0), & \text{если } 0 \le n \le N, \\ \left(1 + \frac{\mu_S}{\mu_F}\right)^N \tilde{\nu}_{FS}^n \rho_1(0), & \text{если } n > N, \end{cases} \tag{11}$$

где $\rho_1(0)$ определяется из условия нормировки, т.е. $\rho_1(0)=\left(\sum_{n=0}^{N}\nu_F^n+\nu_F^N\frac{\tilde{\nu}_{FS}}{1-\tilde{\nu}_{FS}}\right)^{-1}$.

Тогда, объединяя условия $\tilde{\nu}_F<1$ и $\nu_{FS}<1$, получаем, что необходимым условием, при котором предложеннный метод может корректно применятся для изучаемой модели QHS, является выполнение следующего соотношения:

$$\lambda<\min\left\{(1-\alpha)^{-1}\mu_F;\ \mu_F+\mu_S\right\}. \tag{12}$$

Пусть $q_{kk'}, <k>,<k'>\in\Omega$, обозначает интенсивность перехода из укрупненного состояния $<k>$ в укрупненное состояние $<k'>$. Указанные интенсивности переходов вычисляются следующим образом:

$$q_{01}=\lambda\alpha\sum_{n=N}^{\infty}\rho_0(n)=\lambda\alpha\left(1-\sum_{n=0}^{N-1}\rho_0(n)\right); \tag{13}$$

$$q_{10}=\mu_S\sum_{n=0}^{N}\rho_1(n). \tag{14}$$

Следовательно, из (13) и (14) имеем

$$\pi(<0>)=\frac{q_{10}}{q_{01}+q_{10}},\pi(<1>)=1-\pi(<0>). \tag{15}$$

Таким образом, с учетом соотношений (8)–(15) из (7) вычисляются приближенные значения вероятностей состояний. После стандартных преобразований определяется приближенное значение среднего числа заявок в системе:

$$\begin{aligned}L_s\approx\sum_{k=0}^{1}\pi(<k>)\sum_{n=1}^{\infty}n\rho_k(n)=\pi(<0>)\rho_0(0)\left(\sum_{n=1}^{N}n\nu_F^n+(1-\alpha)^{-N}G(\nu_F)\right)+\\+\pi(<1>)\rho_1(0)\left(\sum_{n=1}^{N}n\nu_F^n+\left(1+\tfrac{\mu_S}{\mu_F}\right)^{N}G(\nu_{FS})\right),\end{aligned} \tag{16}$$

где $G(x)=\frac{x^{N+1}\left((N+1)(1-x)+x\right)}{(1-x)^2}$.

Приближенное значение интенсивности включения S-сервера вычисляется как

$$RS\approx\lambda\alpha\pi(<0>)\sum_{n=N}^{\infty}\rho_0(n)=\lambda\alpha\pi(<0>)\left(1-\sum_{n=0}^{N-1}\rho_0(n)\right). \tag{17}$$

Отметим, что разработанный приближенный подход может быть использован и для изучения модели QHS с ограниченным размером буфера. Так, пусть общая вместимость системы равна M, где $M>N$ (т.е. размер буфера равен $M-1$). С этой целью во второй строке формул (9) и (11) верхняя граница изменения параметра n (бесконечность) заменяется конечной величиной M. При этом соответствующим образом изменяются формулы для нахождения величин $\rho_k(0), k=0,1$, т.е. в данной модели QHS имеем

$$\rho_0(0)=\left(\sum_{n=0}^{N}\nu_F^n+(1-\alpha)^{-N}\sum_{n=N+1}^{M}\tilde{\nu}_F^n\right)^{-1};\ \rho_1(0)=\left(\sum_{n=0}^{N}\nu_F^n+\left(1+\frac{\mu_S}{\mu_F}\right)^{N}\sum_{n=N+1}^{M}\tilde{\nu}_{FS}^n\right)^{-1}. \tag{18}$$

Вероятности укрупненных состояний вычисляются по формулам (13)–(15). Далее из формулы (16) и (17) заключаем, что в данной модели QHS приближенные значения характеристик системы вычисляются так:

$$L_s\approx\sum_{k=0}^{1}\pi(<k>)\sum_{n=1}^{M}n\rho_k(n); \tag{19}$$

$$RS\approx\lambda\alpha\pi(<0>)\sum_{n=N}^{M}\rho_0(n). \tag{20}$$

В данной модели QHS появляется новая характеристика – вероятность потери заявок (PB). Точное значение этой величины определяется как

$$PB = (1-\alpha)\, p(M,0) + p(M,1). \tag{21}$$

Среднее время пребывания заявок в данной системе вычисляется следующим образом:

$$W_s = \frac{1}{\lambda(1-PB)} L_s\,; \tag{22}$$

Исходя из (6) заключаем, что приближенное значение этой величины определяется так:

$$PB \approx (1-\alpha)\rho_0(M)\pi(<0>) + \rho_1(M)\pi(<1>). \tag{23}$$

3. Численные результаты

Рассмотрим результаты численных экспериментов для изучаемых моделей. Проводимые эксперименты имеют три цели:

1) оценить точность разработанных приближенных формул для расчета стационарных вероятностей состояний и характеристик изучаемых систем;

2) изучить зависимости характеристик системы от значений порогового параметра N и вероятности включения медленного сервера;

3) решить задачи нахождения оптимального значения порогового параметра N.

Относительно первой цели отметим, что точность разработанных приближенных формул для расчета стационарных вероятностей состояний оценивается с помощью двух мер близости (подобие косинуса (24), евклидово расстояние (25)):

$$\|\mathrm{N}\|_1 = \frac{\sum\limits_{(\boldsymbol{n},\boldsymbol{k})\in E} p(n,k)\,\tilde{p}(n,k)}{\left(\sum\limits_{(\boldsymbol{n},\boldsymbol{k})\in E} \left(p(n,k)\right)^2\right)^{\frac{1}{2}} \left(\sum\limits_{(\boldsymbol{n},\boldsymbol{k})\in E} \left(\tilde{p}(n,k)\right)^2\right)^{\frac{1}{2}}}; \tag{24}$$

$$\|\mathrm{N}\|_2 = \left(\sum_{(n,k)\in E} \left(p(n,k) - \tilde{p}(n,k)\right)^2\right)^{\frac{1}{2}}. \tag{25}$$

Некоторые результаты для модели с неограниченным размером буфера показаны в табл. 1. Из этой таблицы видно, что во всех экспериментах значения нормы (24) практически равны 1, а значения нормы (25) находятся в приемлемых для инженерных расчетов пределах. Иными словами, эти результаты показывают, что стационарные вероятности состояний вычисляются с высокой точностью.

Таблица 1

Оценка точности вычисления вероятностей состояний относительно различных норм близости для модели с бесконечным буфером; $\alpha = 0{,}3;\ N = 5$

λ	(μ_F, μ_S)	Значения нормы		λ	(μ_F, μ_S)	Значения нормы	
		(24)	(25)			(24)	(25)
20	(40, 20)	0,959	0,181	25	(40, 20)	0,942	0,183
	(45, 30)	0,981	0,145		(45, 30)	0,972	0,150
	(50, 35)	0,986	0,135		(50, 35)	0,979	0,143
	(55, 40)	0,989	0,126		(55, 40)	0,983	0,136

Сравнительный анализ результатов вычисления характеристик данной модели с применением различных подходов показан в табл. 2. Здесь и далее приняты следующие сокращения: ТЗ – точные значения, ПЗ – приближенные значения, ОП – относительная погрешность.

Таблица 2

Оценка точности вычисления характеристик модели с бесконечным буфером; $\alpha = 0{,}3;\ N = 5$

λ	(μ_F, μ_S)	W_S			RS		
		ТЗ	ПЗ	ОП	ТЗ	ПЗ	ОП
20	(40, 20)	0,0474	0,0476	0,0037	0,1131	0,1122	0,0085
	(45, 30)	0.0388	0,0389	0,0024	0,0699	0,0703	0,0052
	(50, 35)	0,0327	0,0328	0,0014	0,044	0,0438	0,0046
	(55, 40)	0,0282	0,0283	0,0008	0,0287	0,0284	0,0106
25	(40, 20)	0,0570	0,0576	0,0109	0,3378	0,3622	0,0722
	(45, 30)	0,0456	0,0460	0,0081	0,2220	0,2356	0,0611
	(50, 35)	0,0378	0,0380	0,0048	0,1450	0,1500	0,0347
	(55, 40)	0,0322	0,0323	0,0030	0,0969	0,0986	0,0179

Замечание. Поскольку для модели с неограниченным размером буфера система является бесконечномерной, то при вычислении норм (24) и (25) максимальное число заявок в системе сверху ограничивается достаточно большой конечной величиной. Такая замена является оправданной, так как при превышении этой величины определенного (достаточно большого) значения соответствующие вероятности состояний становятся бесконечно малыми величинами, т.е. практически равными машинному нулю.

Важно отметить, что разработанные приближенные формулы имеют высокую точность и при вычислении характеристик системы. Так, из табл. 2 заключаем, что точные и приближенные значения характеристики (5) в худщем случае отличаются друг от друга в третьем знаке после десятичной точки. Иными словами, здесь максимальное значение ОП составляет меньше 1% (аналогичные результаты получены при вычислении характеристики (4), потому эти результаты здесь не приводятся; этого следовало ожидать, так как характеристики (4) и (5) отличаются друг от друга постоянным множителем). Точность вычисления характеристики (6) также является достаточно высокой для инженерных расчетов, так как при вычислении данной характеристики максимальное значение ОП составляет 7%.

Таблица 3

Оценка точности вычисления вероятностей состояний относительно различных норм близости для модели с ограниченным буфером; $\alpha = 0{,}3;\ M = 7;\ N = 5$

λ	(μ_F, μ_S)	Значения нормы		λ	(μ_F, μ_S)	Значения нормы	
		(24)	(25)			(24)	(25)
20	(40, 20)	0,929	0,169	25	(40, 20)	0,924	0,128
	(45, 30)	0963	0,146		(45, 30)	0,950	0,124
	(50, 35)	0,971	0,144		(50, 35)	0,958	0,143
	(55, 40)	0,977	0,140		(55, 40)	0,965	0,130

Результаты численных экспериментов для модели с ограниченным размером буфера показаны в табл. 3 и 4. Из табл. 3 заключаем, что и для этой модели разработанные приближенные формулы имеют высокую точность при вычислении стационарных вероятностей состояний. Отметим, что здесь также все характеристики вычисляются с высокой точностью для выбранных данных, так как максимальное значение ОП при вычислении характеристики (5) составляет меньше 1% и значения ОП при вычислении характеристик (6) и (21) составляют меньше 7% (см. табл. 4). Однако для тех же исходных данных при $\lambda = 25$ максимальное значение ОП при вычислении характеристики (6) составляет целых 26% (эти результаты отмечены жирным шрифтом в табл. 4). Отметим, что такие

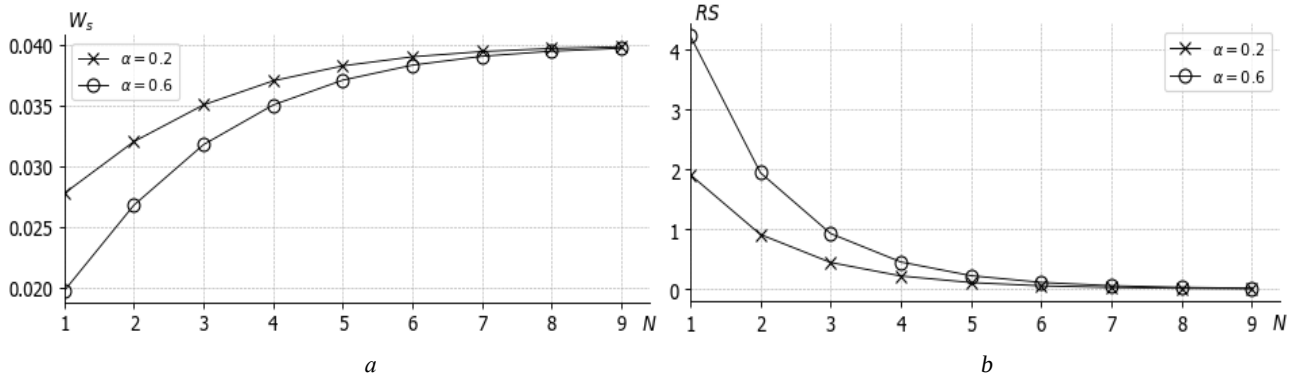
результаты в вычислительных экспериментах объясняются тем, что рассматриваемая здесь гипотетическая модель имеет слишком малую размерность (M = 7) и, кроме того, в данных экспериментах те вероятности состояний, которые участвуют в определении характеристики (6), вычисляются с большими погрешностями (нами сейчас проводятся исследования для разработки алгоритмов устранения подобных погрешностей в проводимых экспериментах). С другой стороны, предложенный подход предназначен для моделей большой и сверхбольшой размерности (для моделей малой размерности могут быть использованы балансовые уравнения для вероятностей состояний), и для таких моделей подобные (нежелательные) случаи нами не обнаружены.

Таблица 4

Оценка точности вычисления характеристик модели с ограниченным буфером; $\alpha = 0{,}3$; $M = 7$; $N = 5$

λ	(μ_F, μ_S)	W_S			RS			PB		
		ТЗ	ПЗ	ОП	ТЗ	ПЗ	ОП	ТЗ	ПЗ	ОП
20	(40, 20)	0,0644	0,065	0,0093	0,7313	0,7808	0,0677	0,0139	0,0143	0,0312
	(45, 30)	0,052	0,0524	0,0083	0,5174	0,5501	0,0631	0,0068	0,0072	0,0649
	(50, 35)	00431	0,0433	0,0059	0,3544	0,3685	0,0398	0,0037	0,0040	0,0683
	(55, 40)	0,0364	0,0366	0,0041	0,2453	0,2512	0,0240	0,0021	0,0023	0,0688
25	(40, 20)	0,0771	0,0784	0,0161	1,9402	2,4457	**0,2605**	0,0523	0,0526	0,0052
	(45, 30)	0,0638	0,065	0,.0185	1,6253	1,9442	**0,1962**	0,0281	0,0299	0,0632
	(50, 35)	0,0536	0,0544	0,0151	1,2419	1,4171	**0,1411**	0,0169	0,0182	0,0748
	(55, 40)	0,0454	0,046	0,0119	0,934	1,0302	**0,1030**	0,0104	0,0112	0,0806

Теперь рассмотим вторую цель. На рис. 2 показаны зависимости характеристик модели с неограниченным буфером от порогового параметра N при различных значениях вероятности включения медленного сервера (α). Как и следовало ожидать, среднее время пребывания заявок в системе является возрастающей функцией относительно указанного параметра (см. рис. 2, *a*), в то время как интенсивность включения медленного сервера убывает с ростом указанного параметра (см. рис. 2, *b*). Эти графики также показывают, что введение рандомизированной N-политики позволяет существенным образом влиять на характеристики системы, особенно при малых значениях порогового параметра N. Так, при трехкратном увеличении вероятности включения медленного сервера удается более чем на 70% уменьшить среднее время пребывания заявок в системе (см. рис. 2, *a*), при этом почти в два раза увеличивается интенсивность включения этого сервера (см. рис. 2, *b*). С ростом порогового параметра N влияние вероятности включения медленного сервера на характеристики системы почти исчезает.



a *b*

Рис. 2. Зависимость характеристик модели с неограниченным буфером от параметра N; $\lambda = 25$, $\mu_F = 50$, $\mu_S = 35$

Fig. 2. Dependence of the characteristics of the model with an unlimited buffer on the parameter N; $\lambda = 25$, $\mu_F = 50$, $\mu_S = 35$

На рис. 3 показаны зависимости характеристик модели с ограниченным буфером от порогового параметра N при различных значениях вероятности включения медленного сервера. Здесь, как и выше, с ростом параметра N среднее время пребывания заявок в системе увеличивается (см. рис. 3, *a*), интенсивность включения медленного сервера убывает (см. рис. 3, *b*); введенная для данной модели новая характеристика – вероятность потери заявок – является возрастающей относительно указанного параметра (см. рис. 3, *c*). Поведение последней характеристики также было ожидаемо, так как увеличение указанного параметра уменьшает шансы поступающих заявок быть принятыми в буфер. Отметим, что здесь двухкратное увеличение вероятности включения медленного сервера позволяет уменьшить вероятности потери заявок почти в 2 раза (см. рис. 3, *c*). В отличие от предыдущих двух характеристик, здесь с ростом порогового параметра N влияние вероятности включения медленного сервера на характеристики системы также растет.

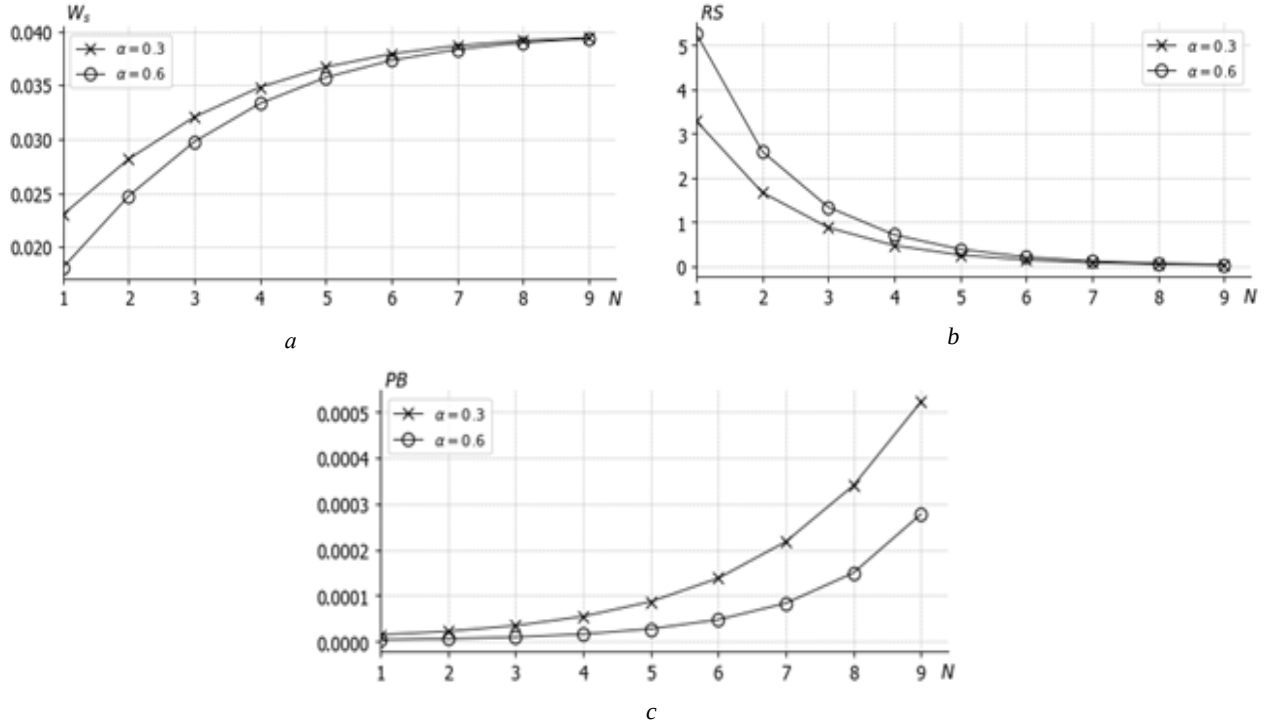


Рис. 3. Зависимость характеристик модели с ограниченным буфером от параметра N; $\lambda = 30$, $\mu_F = 55$, $\mu_S = 40$, $M = 10$

Fig. 3. Dependence of the characteristics of the model with a limited buffer on the parameter N; $\lambda = 30$, $\mu_F = 55$, $\mu_S = 40$, $M = 10$

Наконец рассмотрим третью цель. Для конкретности изложения здесь задача оптимизации заключается в следующем: требуется найти такое (оптимальное) значение параметра N^*, чтобы минимизировать суммарные штрафы (Total Cost, TC), связанные с пребыванием заявок в системе, их потерями (для модели с ограниченным буфером), а также с включением и работой медленного сервера.

Отметим, что из-за сложности задачи нам не удалось доказать (или отвергнуть) существование и единственность решения этой задачи для модели с неограниченым размером буфера. Поэтому здесь мы рассматриваем решение этой задачи для модели с ограниченым размером буфера. Решение этой задачи для модели последнего типа всегда существует, так как множество допустимых решений является конечным и дискретным.

Суммарные штрафы определяются следующим образом:

$$TC = c_s W_s + \lambda P B c_l + c_0 \pi(<1>) + c_1 RS \,, \tag{26}$$

где c_s – штраф за единицу времени пребывания одной заявки в системе; c_l – штраф за потери одной заявки; c_0 – штраф за единицу времени работы S-сервера; c_1 – штраф из-за одноразовое включение S-сервера.

Таблица 5

Результаты решения задачи оптимизации ($\lambda = 30$)

(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**
		5	2	0,258			5	2	0,258		0,2	5	3	0,134
	0,2	10	6	0,052		0,2	10	6	0,052			10	8	0,014
		15	11	0,019			15	11	0,019			15	13	0,006
		5	3	0,232			5	3	0,232		0,4	5	3	0,116
(40, 20)	0,4	10	7	0,048	(40, 20)	0,4	10	7	0,048	(50, 35)		10	8	0,013
		15	12	0,018			15	12	0,018			15	13	0,006
		5	3	0,220			5	3	0,220		0,7	5	4	0,100
	0,7	10	8	0,048		0,7	10	8	0,048			10	9	0,012
		5	2	0,258			15	13	0,018			15	14	0,005

Таблица 6

Результаты решения задачи оптимизации ($\lambda = 40$)

(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**
		5	1	0,583			5	1	0,456		0,2	5	1	0,377
	0,2	10	2	0,203		0,2	10	4	0,130			10	5	0,0845
		15	6	0,108			15	8	0,060			15	10	0,031
		5	1	0,568			5	2	0,412		0,4	5	2	0,329
(40, 20)	0,4	10	4	0,207	(45, 30)	0,4	10	6	0,126	(50, 35)		10	7	0,078
		15	7	0,118			15	10	0,060			15	11	0,029
		5	2	0,583			5	3	0,411		0,7	5	3	0,312
	0,7	10	5	0,223		0,7	10	6	0,133			10	7	0,079
		15	8	0,129			15	11	0,064			15	12	0,029

Таблица 7

Результаты решения задачи оптимизации ($\lambda = 70$)

(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**	(μ_F, μ_S)	α	*M*	*N*	*TC**
		5	1	1,173			5	1	0,793		5	1	0,659	5
	0,2	10	1	0,439		0,2	10	1	0,285		10	3	0,224	10
		15	1	0,229			15	4	0,165		15	6	0,122	15
		5	1	1,113			5	1	0,741		5	1	0,625	5
(40, 20)	0,4	10	1	0,443	(45, 30)	0,4	10	3	0,298	(50, 35)	10	4	0,228	10
		15	1	0,250			15	6	0,188		15	8	0,132	15
		5	1	1,131			5	1	0,781		5	2	0,631	5
	0,7	10	1	0,466		0,7	10	4	0,329		10	5	0,247	10
		15	1	0,273			15	7	0,216		15	19	0,147	15

Из-за сложности определения поведения функции (26) относительно аргумента *N* здесь для нахождения решения минимизации ТС используется метод полного перебора. Результаты решения задачи оптимизации показаны в табл. 5–7. Из этих таблиц можно сделать следующие выводы:

– если вероятность включения медленного сервера остается постоянной, то при фиксированных значениях интенсивности входящего потока с ростом размера буфера оптимальное значение N^* также растет;

– если размер буфер остается постоянным, то при фиксированных значениях интенсивности входящего потока с ростом вероятности включения медленного сервера оптимальное значение N^* также растет;

– если все параметры системы остаются постоянными, то с ростом интенсивности входящего потока оптимальное значение N^* уменьшается.

Заключение

В работе предложена рандомизированная N-политика включения медленного сервера, согласно которой при достижении длины очереди заявок величины N медленный сервер либо с определенной вероятностью включается, либо с дополнительной вероятностью остается в спящем режиме. Рассматриваются модели двух типов: с неограниченным и ограниченным размером буфера для ожидания заявок. Получено условие эргодичности модели с неограниченным размером буфера. Разработан унифицированный приближенный метод расчета распределения вероятностей состояний и характеристик изучаемых систем и решена задача оптимизации системы с ограниченным размером буфера.

В качестве объектов дальнейших исследований можно указать системы, в которых вероятность включения медленного сервера зависит от текущего числа заявок в системе, а также системы, где интервалы времени между поступлениями заявок и времена их обслуживания имеют распределения фазового типа.

Благодарность. Авторы выражают искреннюю благодарность профессору А.А. Назарову за его ценные замечания во время обсуждения результатов работы.

ЛИТЕРАТУРА

1. Gumbel H. Waiting Lines with Heterogeneous Servers // Operations Research. 1960. V. 8. P. 504–511.
2. Singh V.S. Two-Server Markovian Queues with Balking: Heterogeneous vs Homogeneous Servers // Operations Research. 1970. V. 18. P. 145–159.
3. Singh V.S. Markovian Queues with Three Servers // IIE Transactions. 1971. V. 3. P. 45–48.
4. Fakinos D. The M/G/k Blocking System with Heterogeneous Servers // Journal of Operations Research Society. 1980. V. 31. P. 919–927.
5. Fakinos D. The Generalized M/G/k Blocking System with Heterogeneous Servers // Journal of Operations Research Society. 1982. V. 33. P. 801–809.
6. Lin B.W., Elsayed E.A. A General Solution for Multichannel Queuing Systems with Ordered Entry // Computers & Operations Research. 1978. V. 5. P. 219–225.
7. Elsayed E.A. Multichannel Queuing Systems with Ordered Entry and Finite Source // Computers & Operations Research. 1983. V. 10. P. 213–222.
8. Yao D.D. The Arrangement of Servers in an Ordered Entry System // Operations Research. 1987. V. 35. P. 759–763.
9. Pourbabai B., Sonderman D. Server Utilization Factors in Queuing Loss Systems with Ordered Entry and Heterogeneous Servers // Journal of Applied Probability. 1986. V. 23. P. 236–242.
10. Pourbabai B. Markovian Queuing Systems with Retrials and Heterogeneous Servers // Computational Mathematics Applications. 1987. V. 13. P. 917–923.
11. Nawijn W.M. On a Two-Server Finite Queuing System with Ordered Entry and Deterministic Arrivals // European Journal of Operations Research. 1984. V. 18. P. 388–395.
12. Nawijn W.M. A Note on Many-Server Queuing Systems with Ordered Entry with an Application to Conveyor Theory // Journal of Applied Probability. 1983. V. 20. P. 144–152.
13. Yao D.D. Convexity Properties of the Overflow in an Ordered Entry System with Heterogeneous Servers // Operations Research Letters. 1986. V. 5. P. 145–147.
14. Isguder H.O., Kocer U.U. Analysis of GI/M/n/n Queuing System with Ordered Entry and no waiting line // Applied Mathematical Modelling. 2014. V. 38. P. 1024–1032.
15. Melikov A.Z., Mekhbaliyeva E.V. Analysis and optimization of system with heterogen-eous servers and jump priorities // Journal of Computer and Systems Sciences International. 2019. V. 58. P. 718–735.
16. Melikov A.Z., Ponomarenko L.A., Mekhbaliyeva E.V. Analysis of models of systems with heterogen-eous servers // Cybernetics and System Analysis. 2020. V. 56. P. 89–99.
17. Nath G., Enns E. Optimal Service Rates in the Multi-Server Loss System with Heterogeneous Servers // Journal of Applied Probability. 1981. V. 18. P. 776–781.
18. Larsen R.L., Agrawala A.K. Control of Heterogeneous Two-Server Exponential Queuing System // IEEE Transactions on Software Engineering. 1983. V. 9. P. 522–526.
19. Koole G. A simple Proof of the Optimality of a Threshold Policy in a Two-Server Queuing System // Systems and Control Letter. 1995. V. 26. P. 301–303.

20. Lin W., Kumar P.R. Optimal Control of Queuing System with Two Heterogeneous Servers // IEEE Transactions on Automatic Control. 1984. V. 29. P. 696–703.
21. Luh H.P., Viniotis I. Threshold Control Policies for Heterogeneous Servers Systems // Mathematical Methods in Operational Research. 2002. V. 55. P. 121–142.
22. Weber R. On a Conjecture about Assigning Jobs to Processors of Different Speeds // IEEE Transactions on Automatic Control. 1993. V. 38. P. 166–170.
23. Viniotis I., Ephremides A. Extension of the Optimality of a Threshold Policy in Heterogeneous Multi-Server Queuing Systems // IEEE Transactions on Automatic Control. 1988. V. 33. P. 104–109.
24. Rosberg Z., Makowski A.M. Optimal Routing to Parallel Heterogeneous Servers – Small Arrival Rates // IEEE Transactions on Automatic Control. 1990. V. 35. P. 789–796.
25. Горцев А.М., Назаров А.А., Терпугов А.Ф. Управление и адаптация в системах массового обслуживания. Томск : Изд-во Том. ун-та, 1978. 208 с.
26. Rykov V.V. Monotone Control of Queuing Systems with Heterogeneous Servers // Queuing Systems. 2001. V. 37. P. 391–403.
27. Efrosinin D.V., Rykov V.V. Numerical Study of the Optimal Control of a system with Heterogeneous Servers // Automation and Remote Control. 2003. V. 64. P. 302–309.
28. Rykov V.V., Efrosinin D.V. On the Slow Server Problem // Automation and Remote Control. 2009. V. 70. P. 2013–2023.
29. Efrosinin D.V., Breuer L. Threshold policies for controlled retrial queues with heterogeneous servers // Annals of Operations Research. 2006. V. 141. P. 139–162.
30. Efrosinin D., Sztrik J. Optimal Control of a Two-Server Heterogeneous Queuing System with Breakdowns and Constant Retrials // Communications in Computer and Information Sciences. 2016. V. 638. P. 57–72.
31. Efrosinin D., Sztrik J., Farkhadov M., Stepanova N. Reliability Analysis of Two-Server Heterogeneous Queuing System with Threshold Control Policy // Communications in Computer and Information Sciences. 2017. V. 800. P. 13–27.

Поступила в редакцию 16 мая 2020 г.

Melikov A.Z., Mekhbaliyeva E.V. (2020) NUMERICAL INVESTIGATION OF QUEUE WITH HETEREGENEOUS SERVERS AND RANDOMIZED N-POLICY. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika* [Tomsk State University Journal of Control and Computer Science]. 53. pp. 25–37

DOI: 10.17223/19988605/53/3

We consider queueing system that contains two heterogeneous servers: fast server (F-server) and slow server (S-server). Input flow is Poisson one with rate λ where is independent on status of servers. Service times in both servers are independent random variables with exponential distributions; average service time in F-server and S-server are equal to μ_F^{-1} и μ_S^{-1} respectively, where $\mu_F > \mu_S$. It is assumed that the F-server remains awake at all times while the S-server switching on if queue length reaches some threshold $N, N < \infty$. Switching scheme of the S-server is defined as follows. If upon arrival of call queue lenght is no less than *N* then the S-server is switching on with probability (w.p.) $\alpha, 0 < \alpha < 1,$ and with complementary probability $1-\alpha$ this server remains in sleep mode (i.e. in status switch off). Upon completion of servicing of an call in the S-server, it selects one call from the queue for the service if the queue length at this moment is greater than the value *N*; otherwise, the S-server goes into sleep mode. In model with finite buffer an arrived call is blocked if at this moment buffer is fully occupied and the S-server is swithcing on; if upon arrival buffer is fully occupied and the S-server is in sleep mode then either the S-server is switching on w.p. α or with complementary probability an arrived call is blocked. Our problem is calculating the joint distribution of number of calls in the system and status of the S-server and developing the method to finding the performance measures of system.

It is shown that mathematical model of the investigated systems are two-dimensional Markov chains (2D MC). The algorithms to constructing the generating matrix of these 2D MC are developed. The exact and approximate methods to calculating the steady-state probabilities of the indicated 2D MC are proposed. Explicit formulas to calculation of the following performance measures are developed: average number of calls in the system, average sojourn time in the system, intensity of switching on of the S-server, probability of blocking of calls in the system with finite buffer. The approximate method is based on principles of state space merging of 2D MC and it is designed to investigation of large scale models. It is shown that the developed approximate formulas have high accuracy. The problem of minimization of total cost via choosing the optimal value of threshold parameter *N* is solved. Results of numerical experiments are demonstrated and their analysis is performed.

Keywords: queueing system with heterogeneous servers; randomized N-policy; numerical analysis.

MELIKOV Agassi Zarbali (Associate member of National Academy of Sciences of Azerbaijan, Professor, Head of Department of Teletraffic Theory, Institute of Control Systems, Baku, Azerbaijan).
E-mail: agassi.melikov@gmail.com

MEKHBALIYEVA Esmira Vidadi (Candidate of the Technical Science, Sumgayit State University, Sumgayit, Azerbaijan).
E-mail: esmira.mehbaliyeva@mail.ru

REFERENCES

1. Gumbel, H. (1960) Waiting Lines with Heterogeneous Servers. *Operations Research*. 8. pp. 504–511. DOI: 10.1287/opre.8.4.504
2. Singh, V.S. (1970) Two-Server Markovian Queues with Balking: Heterogeneous vs Homogeneous Servers. *Operations Research*. 18. pp. 145–159. DOI: 10.1287/opre.18.1.145
3. Singh, V.S. (1971) Markovian Queues with Three Servers. *IIE Transactions*. 3. pp. 45–48.
4. Fakinos, D. (1980) The M/G/k Blocking System with Heterogeneous Servers. *Journal of Operations Research Society*. 31. pp. 919–927. DOI: 10.1057/jors.1980.167
5. Fakinos, D. (1982) The Generalized M/G/k Blocking System with Heterogeneous Servers. *Journal of Operations Research Society*. 33. pp. 801–809. DOI: 10.1057/jors.1982.175
6. Lin, B.W. & Elsayed, E.A. (1978) A General Solution for Multichannel Queuing Systems with Ordered Entry. *Computers & Operations Research*. 5. pp. 219–225. DOI: 10.1016/0305-0548(78)90031-X
7. Elsayed, E.A. (1983) Multichannel Queuing Systems with Ordered Entry and Finite Source. *Computers & Operations Research*. 10. pp. 213–222. DOI: 10.1016/0305-0548(83)90014-X
8. Yao, D.D. (1987) The Arrangement of Servers in an Ordered Entry System. *Operations Research*. 35. pp. 759–763. DOI: 10.1287/opre.35.5.759
9. Pourbabai, B. & Sonderman, D. (1986) Server Utilization Factors in Queuing Loss Systems with Ordered Entry and Heterogeneous Servers. *Journal of Applied Probability*. 23. pp. 236–242. DOI: 10.2307/3214135
10. Pourbabai, B. (1987) Markovian Queuing Systems with Retrials and Heterogeneous Servers. *Computational Mathematics Applications*. 13. pp. 917–923. DOI: 10.1016/0898-1221(87)90064-2
11. Nawijn, W.M. (1984) On a Two-Server Finite Queuing System with Ordered Entry and Deterministic Arrivals. *European Journal of Operations Research*. 18. pp. 388–395. DOI: 10.1016/0377-2217(84)90161-9
12. Nawijn, W.M. (1983) A Note on Many-Server Queuing Systems with Ordered Entry with an Application to Conveyor Theory. *Journal of Applied Probability*. 20. pp. 144–152. DOI: 10.2307/3213728
13. Yao, D.D. (1986) Convexity Properties of the Overflow in an Ordered Entry System with Heterogeneous Servers. *Operations Research Letters*. 5. pp. 145–147. DOI: 10.1016/0167-6377(86)90087-8
14. Isguder, H.O. & Kocer, U.U. (2014) Analysis of GI/M/n/n Queuing System with Ordered Entry and no waiting line. *Applied Mathematical Modelling*. 38. pp. 1024–1032. DOI: 10.1016/j.apm.2013.07.029
15. Melikov, A.Z. & Mekhbaliyeva, E.V. (2019) Analysis and optimization of system with heterogeneous servers and jump priorities. *Journal of Computer and Systems Sciences International*. 58. pp. 718–735. DOI: 10.1134/S1064230719050113
16. Melikov, A.Z., Ponomarenko, L.A. & Mekhbaliyeva, E.V. (2020) Analysis of models of systems with heterogeneous servers. *Cybernetics and System Analysis*. 56. pp. 89–99. DOI: 10.1007/s10559-020-00224-x
17. Nath, G. & Enns, E. (1981) Optimal Service Rates in the Multi-server Loss System with Heterogeneous Servers. *Journal of Applied Probability*. 18. pp. 776–781. DOI: 10.2307/3213336
18. Larsen, R.L. & Agrawala, A.K. (1983) Control of Heterogeneous Two-Server Exponential Queuing System. *IEEE Transactions on Software Engineering*. 9. pp. 522–526. DOI: 10.1109/TSE.1983.234960
19. Koole, G. (1995) A simple proof of the optimality of a threshold policy in a two-server queuing system. *Systems and Control Letter*. 26. pp. 301–303. DOI: 10.1016/0167-6911(95)00015-1
20. Lin, W. & Kumar, P.R. (1984) Optimal Control of Queuing System with Two Heterogeneous Servers. *IEEE Transactions on Automatic Control*. 29. pp. 696–703.
21. Luh, H.P. & Viniotis, I. (2002) Threshold Control Policies for Heterogeneous Servers Systems. *Mathematical Methods in Operational Research*. 55. pp. 121–142. DOI: 10.1007/s001860100168
22. Weber, R. (1993) On a Conjecture about Assigning Jobs to Processors of Different Speeds. *IEEE Transactions on Automatic Control*. 38. pp. 166–170. DOI: 10.1109/9.186333
23. Viniotis, I. & Ephremides, A. (1988) Extension of the Optimality of a Threshold Policy in Heterogeneous Multi-Server Queuing Systems. *IEEE Transactions on Automatic Control*. 33. pp. 104–109.
24. Rosberg, Z. & Makowski, A.M. (1990) Optimal Routing to Parallel Heterogeneous Servers – Small Arrival Rates. *IEEE Transactions on Automatic Control*. 35. pp. 789–796. DOI: 10.1109/9.57017
25. Gortsev, A.M., Nazarov, A.A. & Terpugov, A.F. (1978) *Upravlenie i adaptatsiya v sistemakh massovogo obsluzhivaniya* [Management and adaptation in queuing systems]. Tomsk: Tomsk State University.
26. Rykov, V.V. (2001) Monotone Control of Queuing Systems with Heterogeneous Servers. *Queuing Systems*. 37. pp. 391–403. DOI: 10.1023/A:1010893501581
27. Efrosinin, D.V. & Rykov, V.V. (2003) Numerical Study of the Optimal Control of a system with Heterogeneous Servers. *Automation and Remote Control*. 64. pp. 302–309. DOI: 10.1023/A:1022271316352
28. Rykov, V.V. & Efrosinin, D.V. (2009) On the Slow Server Problem. *Automation and Remote Control*. 70. pp. 2013–2023. DOI: 10.1134/S0005117909120091
29. Efrosinin, D.V. & Breuer, L. (2006) Threshold policies for controlled retrial queues with heterogeneous servers. *Annals of Operations Research*. 141. pp. 139–162. DOI: 10.1007/s10479-006-5297-5
30. Efrosinin, D. & Sztrik J. (2016) Optimal Control of a Two-Server Heterogeneous Queuing System with Breakdowns and Constant Retrials. *Communications in Computer and Information Sciences*. 638. pp. 57–72. DOI: 10.1007/978-3-319-44615-8_5
31. Efrosinin, D., Sztrik, J., Farkhadov, M. & Stepanova, N. (2017) Reliability Analysis of Two-Server Heterogeneous Queuing System with Threshold Control Policy. *Communications in Computer and Information Sciences*. 800. pp 13–27. DOI: 10.1007/978-3-319-68069-9_2