УДК 519.872

А.З. МЕЛИКОВ, Л.А. ПОНОМАРЕНКО, Э.В. МЕХБАЛЫЕВА

АНАЛИЗ МОДЕЛЕЙ СИСТЕМ С ГЕТЕРОГЕННЫМИ СЕРВЕРАМИ

Аннотация. Исследована математическая модель системы обслуживания с гетерогенными серверами и без очередей при наличии заявок двух типов. Заявки высокого приоритета обслуживаются в высокоскоростных серверах, а заявки низкого приоритета — в низкоскоростных. В случаях занятости всех серверов в соответствующих группах допускается обслуживание поступившей заявки в другой группе, при этом переназначения заявок осуществляются согласно рандомизированной схеме. Считается, что вероятности переназначения зависят от числа занятых серверов в соответствующей группе. Разработаны методы точного и приближенного анализа характеристик этой системы и получены явные формулы для приближенного вычисления ее характеристик.

Ключевые слова: гетерогенный сервер, система обслуживания, приоритеты, разнотипные заявки, оптимизация.

ВВЕДЕНИЕ

Классические модели систем массового обслуживания (СМО) базируются на ряде предположений, одно из которых — допущение о том, что все показатели серверов идентичны (скорость обработки запросов, надежность, стоимость эксплуатации и т.д.). Однако зачастую это допущение является грубым приближением к реальной ситуации, так как в процессе расширения существующих систем, особенно быстро развивающихся компьютерных и телекоммуникационных систем, приходится использовать гетерогенные серверы (Heterogeneous Servers, HS). Кроме компьютерных и телекоммуникационных систем, серверы с различными скоростями используются в системах, где в процессе обслуживания заявок участвуют специалисты.

Несмотря на то, что первая работа, посвященная изучению моделей СМО с гетерогенными серверами [1], была опубликована почти 60 лет тому назад, эти модели долгое время не исследовались. В указанной работе рассматривалась марковская система с бесконечной очередью и рандомизированным доступом, т.е. предполагалось, что для обслуживания заявок с равными вероятностями назначается один из свободных серверов. Там же предлагались формулы для нахождения вероятностей состояний и среднего числа заявок в системе. Указывалось, что в частном случае, когда все серверы являются гомогенными (идентичными), получаются известные классические результаты.

В дальнейшем подобные модели изучались в [2, 3], где вычислялись характеристики систем с двумя и тремя серверами и рандомизированным доступом. Полученные результаты сравнивались с аналогичными результатами для систем с гомогенными серверами.

Интересны работы, где приведены обобщения классических результатов, полученных ранее для систем с гомогенными серверами. Так, в [4] предложено обобщение известной B-формулы Эрланга для модели $M / G / k / k$ с HS и рандомизированным доступом, а в [5] рассмотрено обобщение последней модели, где предполагается, что поступающие заявки теряются с определенными вероятностями, если даже в моменты их поступления имеются свободные серверы. В [5] получен аналог B-формулы Эрланга для такой модели и показано, что в стационарном режиме выходящий поток также является пуассоновским. В [6] доказано,

© А.З. Меликов, Л.А. Пономаренко, Э.В. Мехбалыева, 2020

что вероятность потери в системе $M / M / k / k$ имеет минимальное значение при использовании FSF-схемы доступа (Fast Server First). Среднее время ожидания в системе $M / M / k / \infty$ с HS и рандомизированным доступом будет минимальным, если суммарную интенсивность обслуживания серверов равномерно распределить между ними, т.е. для гомогенной системы [7].

В известных публикациях подробно изучены модели, в которых принята схема упорядоченного доступа [8–15]. Согласно этой схеме заранее все серверы нумеруются и назначается свободный сервер с минимальным номером. Детальный обзор этих работ приведен в [16], где предложен метод нахождения вероятности потери в системе $GI / M / k / k$.

В системах с HS важны проблемы определения оптимальных стратегий доступа заявок, а также схем распределения заявок между параллельными серверами. Во многих работах используются эвристические стратегии типа FSF-схемы доступа или рандомизированные стратегии. Однако после опубликования [17] выяснилось, что эвристические стратегии не всегда являются оптимальными и даже субоптимальными. Так, в [17] доказано, что для минимизации среднего числа заявок в системе необходимо использовать стратегию доступа порогового типа: из двух серверных систем быстрый сервер работает всегда, если в системе имеется хотя бы одна заявка, а медленный — включается лишь, когда длина очереди достигает определенного порогового значения. В дальнейшем этот результат доказывался различными авторами с использованием разных подходов [18–21]. Аналогичное утверждалось для моделей с ненадежными серверами [22, 23]. Обобщения этой стратегии для моделей с более чем двумя серверами рассмотрены в [24–27].

Анализ работ показал, что подробно изучены модели систем с HS при наличии идентичных запросов. Авторам настоящей статьи не известны публикации, посвященные изучению моделей систем с HS и разнотипными запросами, хотя очевидно, что при использовании HS классификация обрабатываемых заявок помогает повысить ее эффективность. Действительно, для повышения экономической эффективности и (или) улучшения операционных характеристик работы системы можно выделить высокоприоритетные (H-заявки) и низкоприоритетные заявки (L-заявки) и организовать обработку H-заявок в высокоскоростных серверах (F-серверах), а L-заявок в медленных серверах (S-серверах). В случаях занятости всех серверов одной группы в целях уменьшения вероятности потери заявок (или повышения коэффициента использования серверов) можно организовать обслуживание заявок соответствующего типа в другой группе серверов. При этом необходимо учесть, что возможны улучшения (или ухудшения) качества обслуживания. Для ответа на вопрос о целесообразности переназначения заявок в другие группы серверов необходимо построить адекватную математическую модель таких систем для вычисления их искомых характеристик.

В настоящей работе сделана первая попытка в этом направлении: разработана модель системы с гетерогенными серверами и разнотипными заявками, а также предложены методы расчета и оптимизации ее характеристик.

ОПИСАНИЕ МОДЕЛИ И ПОСТАНОВКА ЗАДАЧИ

Рассматриваемая система содержит две группы серверов: скоростные (F-серверы) и медленные (S-серверы). Количества F- и S-серверов равны $N_F > 1$ и $N_S > 1$ соответственно. Времена обслуживания заявок в этих серверах имеют показательные распределения со средними значениями μ_F^{-1} и μ_S^{-1} для F- и S-серверов соответственно, при этом $\mu_F > \mu_S$.

В эту систему поступают простейшие потоки двух типов: высокоприоритетные (H-заявки) и низкоприоритетные (L-заявки), при этом их интенсивности равны λ_H и λ_L соответственно, а обслуживание осуществляется согласно следующим правилам:

- если в моменты поступления H- и L-заявок имеется хотя бы один свободный сервер в соответствующей группе каналов, то эти заявки начинают мгновенно обслуживаться в своих группах;
- если в моменты поступления H- или L-заявок все серверы в обеих группах заняты, то они теряются с вероятностью единица;
- если в момент поступления H-заявки все F-серверы заняты и количество занятых S-серверов равно i, $0 \le i \le N_S - 1$, то поступившая H-заявка либо с вероятностью $\alpha(i)$, $0 < \alpha(i) \le 1$, начинает обслуживаться в любом свободном S-сервере, либо с вероятностью $1-\alpha(i)$ покинет систему, не получив обслуживания (т.е. теряется);
- если в момент поступления L-заявки все S-серверы заняты и количество занятых F-серверов равно j, $0 \le j \le N_F - 1$, то поступившая L-заявка либо с вероятностью $\beta(j)$, $0 < \beta(j) \le 1$, начинает обслуживаться в любом свободном F-сервере, либо с вероятностью $1-\beta(j)$ покидает систему.

Отметим, что при определенных значениях введенных вероятностей $\alpha(i)$ и $\beta(j)$ можно получить детерминированные схемы обслуживания разнотипных заявок в «чужих» группах, которые легко реализуются на практике. Так, например, рассмотрим следующие схемы:

$$a(i) = \begin{cases} 1, & \text{если } i \le T_S, \\ 0, & \text{если } i > T_S; \end{cases} \quad \beta(j) = \begin{cases} 1, & \text{если } j \le T_F, \\ 0, & \text{если } i > T_F. \end{cases} \tag{1}$$

В соотношениях (1) величина T_S, $0 < T_S < N_S$, указывает пороговое значение для числа занятых S-серверов, ниже которого допускается обслуживание H-заявок в данной группе, а если число занятых серверов данной группы превышает величину T_S, такое обслуживание не допускается. Аналогичный смысл имеет величина T_F, $0 < T_F < N_F$.

Задача состоит в нахождении совместного распределения числа занятых серверов в различных группах, а также характеристик систем: C_F и C_S — коэффициентов использования F- и S-серверов соответственно; PB_H и PB_L — вероятностей потери H- и L-заявок соответственно; R_{HS} и R_{LF} — интенсивности H- и L-заявок, обслуженных на S- и F-серверах соответственно.

ПОСТРОЕНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ И ТОЧНЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ

Состояние системы в произвольный момент времени задается двумерным вектором (k_F, k_S), где его компоненты k_F и k_S указывают число занятых каналов в группах F- и S-серверов соответственно. Совокупность этих векторов составляет фазовое пространство состояний (ФПС) модели — декартовое произведение двух множеств:

$$E = \{0, 1, \ldots, N_F\} \times \{0, 1, \ldots, N_S\}. \tag{2}$$

Поскольку входящие потоки являются простейшими и времена обслуживания разнотипных заявок имеют показательные распределения, математическая модель изучаемой системы представляет собой двумерную цепь Маркова (Two-Dimensional Markov Chain, 2-D MC) с ФПС (2).

Элементы производящей матрицы (ПМ) указанной 2-D MC обозначим $q((k_F, k_S), (k'_F, k'_S))$. Между состояниями изучаемой 2-D MC возможны следующие переходы:

• если в момент поступления H-заявки система находится в состоянии (k_F, k_S), $k_F < N_F$, то с интенсивностью λ_H происходит переход в состояние $(k_F + 1, k_S)$;

• если в момент поступления H-заявки система находится в состоянии (N_F, k_S), $k_s < N_s$, то с интенсивностью $\lambda_H \alpha(k_S)$ происходит переход в состояние $(N_F, k_S + 1)$;

• если в момент поступления L-заявки система находится в состоянии (k_F, k_S), $k_S < N_S$, то с интенсивностью λ_L происходит переход в состояние $(k_F, k_S + 1)$;

• если в момент поступления L-заявки система находится в состоянии (k_F, N_S), $k_F < N_F$, то с интенсивностью $\lambda_L \beta(k_F)$ происходит переход в состояние $(k_F + 1, N_S)$;

• если в момент завершения процесса обслуживания в группе F-серверов система находится в состоянии (k_F, k_S), то с интенсивностью $k_F \mu_F$ происходит переход в состояние $(k_F - 1, k_S)$;

• если в момент завершения процесса обслуживания в группе S-серверов система находится в состоянии (k_F, k_S), то с интенсивностью $k_S \mu_S$ происходит переход в состояние $(k_F, k_S - 1)$.

Следовательно, положительные элементы ПМ изучаемой 2-D МС определяются следующим образом:

$$q((k_F, k_S), (k'_F, k'_S)) =$$

$$= \begin{cases} \lambda_H, & \text{если } k_F < N_F,\ k'_F = k_F + 1,\ k'_S = k_S, \\ \lambda_H \alpha(k_S) + \lambda_L, & \text{если } k_F = N_F,\ k_S < N_S,\ k'_F = N_F,\ k'_S = k_S + 1, \\ \lambda_L, & \text{если } k_S < N_S,\ k'_F = k_F,\ k'_S = k_S + 1, \\ \lambda_L \beta(k_F) + \lambda_H, & \text{если } k_S = N_S,\ k_F < N_F,\ k'_F = k_F + 1,\ k'_S = N_S, \\ k_F \mu_F, & \text{если } k'_F = k_F - 1,\ k'_S = k_S, \\ k_S \mu_S, & \text{если } k'_F = k_F,\ k'_S = k_S - 1. \end{cases} \tag{3}$$

Из соотношений (3) получаем, что все состояния изучаемой 2-D МС сообщаются, иными словами, в ней существует стационарный режим. Вероятность состояния $(k_F, k_S) \in E$ в стационарном режиме обозначим $p(k_F, k_S)$. Эти вероятности находятся из системы уравнений равновесия (СУР). Используя соотношения (3), получаем, что эта СУР имеет следующий вид:

• для случаев $k_F < N_F$, $k_S < N_S$

$$\begin{aligned} &(\lambda_H + \lambda_L + k_F \mu_F + k_S \mu_S) p(k_F, k_S) = \\ &= \lambda_H p(k_F - 1, k_S) I(k_F > 0) + \lambda_L p(k_F, k_S - 1) I(k_S > 0) + \\ &+ (k_F + 1) \mu_F p(k_F + 1, k_S) + (k_S + 1) \mu_S p(k_F, k_S + 1); \end{aligned} \tag{4}$$

• для случаев $k_F = N_F$, $k_S < N_S$

$$\begin{aligned} &(\lambda_H \alpha(k_S) + \lambda_L + N_F \mu_F + k_S \mu_S) p(N_F, k_S) = \\ &= (\lambda_H \alpha(k_S - 1) + \lambda_L) p(N_F, k_S - 1) I(k_S > 0) + \\ &+ \lambda_H p(N_F - 1, k_S) + (k_S + 1) \mu_S p(N_F, k_S + 1); \end{aligned} \tag{5}$$

• для случаев $k_F < N_F$, $k_S = N_S$

$$\begin{aligned} &(\lambda_H + \lambda_L \beta(k_F) + k_F \mu_F + N_S \mu_S) p(k_F, N_S) = \\ &= (\lambda_H + \lambda_L \beta(k_F - 1)) p(k_F - 1, N_S) I(k_F > 0) + \\ &+ \lambda_L p(k_F, N_S - 1) + (k_F + 1) \mu_F p(k_F + 1, N_S); \end{aligned} \tag{6}$$

• для случаев $k_F = N_F$, $k_S = N_S$

$$(N_F\mu_F + N_S\mu_S)p(N_F, N_S) = (\lambda_H\alpha(N_S - 1) + \lambda_L)p(N_F, N_S - 1) + \\ + (\lambda_H + \lambda_L\beta(N_F - 1))p(N_F - 1, N_S). \quad (7)$$

В этих уравнениях $I(A)$ обозначает индикаторную функцию события A. К ним добавляется еще и уравнение нормировки

$$\sum_{(k_F, k_S)\in E} p(k_F, k_S) = 1. \quad (8)$$

Приведенная СУР (4)–(8) имеет единственное решение, и после его нахождения можно вычислить требуемые характеристики системы. Так, коэффициенты использования серверов определяются следующим образом:

$$C_x = \widetilde{N}_x / N_x, \ x \in \{F, S\}, \quad (9)$$

где $\widetilde{N}_F$ и $\widetilde{N}_S$ обозначают среднее число занятых серверов в группах F- и S-серверов соответственно, т.е.

$$\widetilde{N}_F = \sum_{k_F=1}^{N_F} k_F \sum_{k_S=0}^{N_S} p(k_F, k_S), \ \widetilde{N}_S = \sum_{k_S=1}^{N_S} k_S \sum_{k_F=0}^{N_F} p(k_F, k_S).$$

Заявки высокого приоритета теряются в следующих случаях: если в момент поступления H-заявки все серверы в обеих группах заняты, то поступившая заявка теряется с вероятностью единица; если в момент поступления H-заявки все серверы в группе F-серверов заняты и число занятых S-серверов в группе равно i, то поступившая заявка теряется с вероятностью $1-\alpha(i)$. Следовательно, искомая вероятность потери H-заявок вычисляется так:

$$PB_H = p(N_F, N_S) + \sum_{k_S=0}^{N_S-1} (1-\alpha(k_S))p(N_F, k_S). \quad (10)$$

С помощью аналогичных рассуждений получаем, что вероятность потери L-заявок вычисляется так:

$$PB_L = p(N_F, N_S) + \sum_{k_F=0}^{N_F-1} (1-\beta(k_F))p(k_F, N_S). \quad (11)$$

Интенсивности H-заявок, обслуженных на S-серверах (R_{HS}), и L-заявок, обслуженных на F-серверах (R_{LF}), вычисляются так:

$$R_{HS} = \lambda_H \sum_{k_S=0}^{N_S-1} p(N_F, k_S)\alpha(k_S), \quad (12)$$

$$R_{LF} = \lambda_L \sum_{k_F=0}^{N_F-1} p(k_F, N_S)\beta(k_F). \quad (13)$$

Разработанный подход легко реализуется для моделей малой размерности. Однако с ростом числа разнотипных серверов в каждой группе его применение требует неоправданно большого времени выполнения. Известные матрично-геометрические методы [28–30], которые зачастую используются для решения подобных проблем, имеют ряд недостатков (подробнее об этих методах см. в [31]).

В настоящей статье предлагается альтернативный приближенный метод, который позволяет решать рассматриваемые задачи с помощью явных формул.

ПРИБЛИЖЕННЫЙ МЕТОД РЕШЕНИЯ ЗАДАЧИ

Применение предлагаемого метода требует расщепления ФПС модели на такие классы, чтобы интенсивности переходов между состояниями внутри классов намного превышали интенсивности переходов между состояниями из разных классов. В данной модели такое расщепление легко осуществляется. Действительно, поскольку F-сервер имеет более высокую скорость обслуживания, чем S-сервер, и (предположительно) интенсивность H-запросов намного больше, чем интенсивность L-запросов, то следующее расщепление ПС модели удовлетворяет указанным условиям:

$$E = \bigcup_{j=0}^{N_S} E_j, \ E_{j_1} \bigcap E_{j_2} = \varnothing, \ j_1 \neq j_2, \tag{14}$$

где $E_j = \{(i, j) \in E : i = \overline{0, N_F}\}, \ j = \overline{0, N_S}$.

Замечание 1. Здесь и далее для простоты изложения и упрощения обозначений в некоторых формулах вектор состояния (k_F, k_S) заменяется вектором (i, j).

На основе расщепления (14) все микросостояния каждого класса E_j объединяются в одно укрупненное состояние $\langle j \rangle$, и таким образом определяется множество укрупненных состояний $\Omega = \{\langle j \rangle : j = \overline{0, N_S}\}$.

Приближенные значения вероятностей состояний $\widetilde{p}(i, j)$, $(i, j) \in E$, исходной модели определяются так (см. приложение в [32]):

$$\widetilde{p}(i, j) = \rho_j(i)\pi(\langle j \rangle), \tag{15}$$

где $\rho_j(i)$ — вероятность состояния (i, j) внутри расщепленной модели с пространством состояний E_j, а $\pi(\langle j \rangle)$ — вероятность укрупненного состояния $\langle j \rangle \in \Omega$.

Вероятности состояний внутри всех расщепленных моделей с пространством состояний E_j, $0 \le j \le N_S - 1$, не зависят от индекса j и совпадают с вероятностями состояний классической модели Эрланга $M / M / N_F / N_F$ с нагрузкой ν_H, т.е.

$$\rho_j(i) = \frac{\nu_H^i}{i!} \Big/ \sum_{j=0}^{N_F} \frac{\nu_H^j}{j!}, \ i = \overline{0, N_F}, \tag{16}$$

где $\nu_H = \lambda_H / \mu_F$.

Замечание 2. Поскольку вероятности $\rho_j(i)$, $i = \overline{0, N_F}$, не зависят от индекса j при $0 \le j \le N_S - 1$, то далее для простоты изложения и упрощения обозначений в соответствующих формулах этот индекс опускается.

Вероятности состояний внутри расщепленной модели с пространством состояний E_j при $j = N_S$ совпадают с вероятностями состояний одномерного процесса размножения–гибели с зависящими от состояния интенсивностями переходов, т.е.

$$\rho_{N_S}(i) = \frac{1}{i! \mu_F^i} \prod_{j=0}^{i-1} \widetilde{\lambda}_H(j) \rho_{N_S}(0), \ i = \overline{1, N_F}, \tag{17}$$

где $\widetilde{\lambda}_H(j) = \lambda_H + \lambda_L \beta(j), \ j = \overline{0, N_S - 1}$. Вероятность $\rho_{N_S}(0)$ находится из условия нормировки, т.е. $\sum_{j=0}^{N_F} \rho_{N_S}(j) = 1$.

Тогда после определенных математических выкладок находим, что интенсивности переходов между укрупненными состояниями $\langle j_1 \rangle$ и $\langle j_2 \rangle$ вычисляются так:

$$q(\langle j_1\rangle, \langle j_2\rangle) = \begin{cases} \widetilde{\lambda}_L(j_1), & \text{если } j_2 = j_1 + 1, \\ j_1\mu_S, & \text{если } j_2 = j_1 - 1, \end{cases} \tag{18}$$

где $\widetilde{\lambda}_L(j_1) = \lambda_L(1 - E_B(\nu_H, N_F)) + (\lambda_L + \lambda_H\alpha(j_1))E_B(\nu_H, N_F)$, $j_1 = \overline{0, N_S - 1}$.

Здесь и далее $E_B(x, m)$ обозначает B-формулу Эрланга для системы $M/M/m/m$ с нагрузкой x, т.е. $E_B(x, m) = \dfrac{x^m}{m!} \Big/ \sum\limits_{j=0}^{m} \dfrac{x^j}{j!}$.

Следовательно, из равенства (18) получаем, что вероятности укрупненных состояний $\pi(\langle j\rangle)$, $j \in \Omega$, совпадают с вероятностями состояний одномерного процесса размножения–гибели с зависящими от состояния интенсивностями переходов, т.е.

$$\pi(\langle j\rangle) = \frac{1}{j!\,\mu_S^j} \prod_{i=0}^{j-1} \widetilde{\lambda}_L(i)\pi(\langle 0\rangle), \quad j = \overline{1, N_S}. \tag{19}$$

Вероятность $\pi(\langle 0\rangle)$ находится из условия нормировки, т.е. $\sum\limits_{j=0}^{N_S} \pi(\langle j\rangle) = 1$.

С учетом соотношений (16)–(19) из (15) вычисляются приближенные значения вероятностей состояний исходной модели. После нахождения этих вероятностей легко определяются следующие приближенные значения характеристик (10)–(13) данной системы:

$$PB_H \approx \rho_{N_S}(N_F)\pi(\langle N_S\rangle) + E_B(\nu_H, N_F) \sum_{i=0}^{N_S-1} (1 - \alpha(i))\pi(\langle i\rangle), \tag{20}$$

$$PB_L \approx \pi(\langle N_S\rangle)\left(\rho_{N_S}(N_F) + \sum_{i=0}^{N_F-1} (1 - \beta(i))\rho_{N_S}(i) \right), \tag{21}$$

$$R_{HS} \approx \lambda_H E_B(\nu_H, N_F) \sum_{i=0}^{N_S-1} \alpha(i)\pi(\langle i\rangle), \tag{22}$$

$$R_{LF} \approx \lambda_L \pi(\langle N_S\rangle) \sum_{i=0}^{N_F-1} \rho_{N_S}(i)\beta(i). \tag{23}$$

Приближенные значения среднего числа занятых серверов в каждой группе определяются так (см. формулы (9) и замечание 2):

$$\widetilde{N}_F \approx (1 - \pi(\langle N_S\rangle)) \sum_{i=1}^{N_F} i\rho(i) + \pi(\langle N_S\rangle) \sum_{i=1}^{N_F} i\rho_{N_S}(i), \tag{24}$$

$$\widetilde{N}_S \approx \sum_{j=1}^{N_S} j\pi(\langle j\rangle). \tag{25}$$

Далее с учетом (9), (24) и (25) находятся приближенные значения коэффициентов использования серверов в каждой группе.

ЧИСЛЕННЫЕ РЕЗУЛЬТАТЫ

Полученные формулы позволяют изучать поведение характеристик системы относительно изменения любых ее параметров. Ввиду ограниченного объема статьи предполагается, что количество разнотипных серверов и нагрузочные параметры разнотипных заявок являются фиксированными величинами, и рассматривается лишь поведение характеристик системы относительно пороговых параметров T_F и T_S, которые определены в соотношениях (1).

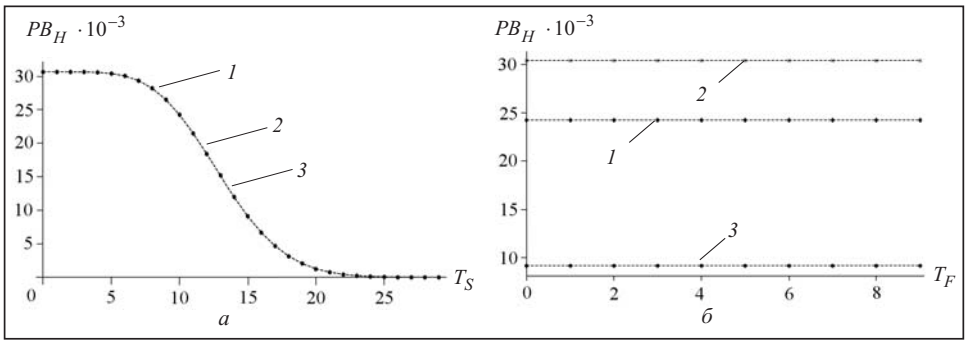

Рис. 1. Зависимость PB_H от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*) $T_F = 5$ (кривая *3*) (*а*), а также от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*б*)

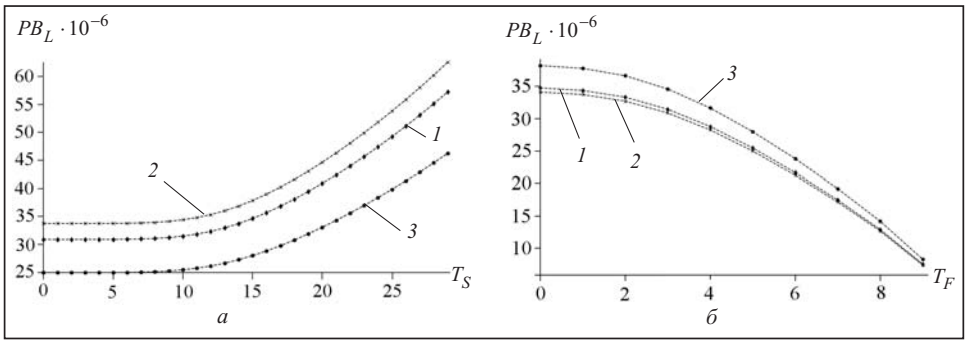

Рис. 2. Зависимость PB_L от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*), $T_F = 5$ (кривая *3*) (*а*), а также от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*б*)

Предположим, что фиксированные значения исходных параметров являются следующими: $N_F = 10$, $N_S = 30$, $\lambda_H = 50$, $\lambda_L = 40$, $\mu_F = 9$, $\mu_S = 3$. Отметим, что приведенный далее анализ основан исключительно на этих значениях исходных данных.

Зависимость функции PB_H от параметров T_F и T_S показана на рис. 1. Увеличение параметра T_S повышает шансы H-заявок быть принятыми для обслуживания, и поэтому функция PB_H убывающая относительно указанного параметра, при этом ее значения почти не зависят от фиксированных значений параметра T_F (см. рис. 1, *а*). Вместе с тем функция PB_H является постоянной относительно изменения параметра T_F при фиксированных значениях параметра T_S (см. рис. 1, *б*).

Зависимость функции PB_L от параметров T_F и T_S показана на рис. 2. Увеличение параметра T_S повышает шансы H-заявок быть принятыми для обслуживания в группе S-серверов, и тем самым уменьшает шансы L-заявок быть принятыми для обслуживания. Поэтому функция PB_L является возрастающей относительно параметра T_S, при этом ее значения почти не зависят от фиксированных значений параметра T_F, так как отличаются лишь в пятом знаке после десятичной точки (см. рис. 2, *а*). Как и следовало ожидать, функция PB_L убывающая относительно изменения параметра T_F, поскольку увеличение значения этого параметра повышает шансы L-заявок быть принятыми для обслуживания, при этом она почти не зависит от значения параметра T_S (см. рис. 2, *б*).

Поскольку рост параметра T_S приводит к увеличению интенсивности обслуживания H-заявок в группе S-серверов, функция R_{HS} возрастающая относительно указанного параметра, при этом ее значения почти не зависят от фиксированного значения параметра T_F (рис. 3, *а*). Однако эта функция является постоянной относительно изменения параметра T_F при фиксированных значениях параметра T_S (рис. 3, *б*).

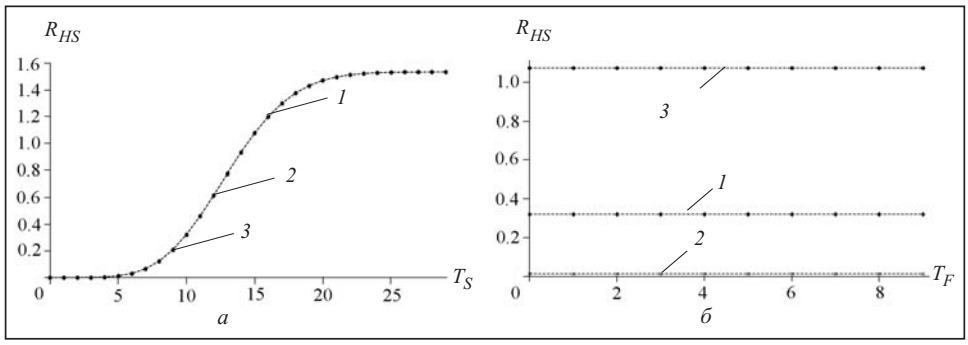

Рис. 3. Зависимость R_{HS} от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*), $T_F = 5$ (кривая *3*) (*а*), а также от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*б*)

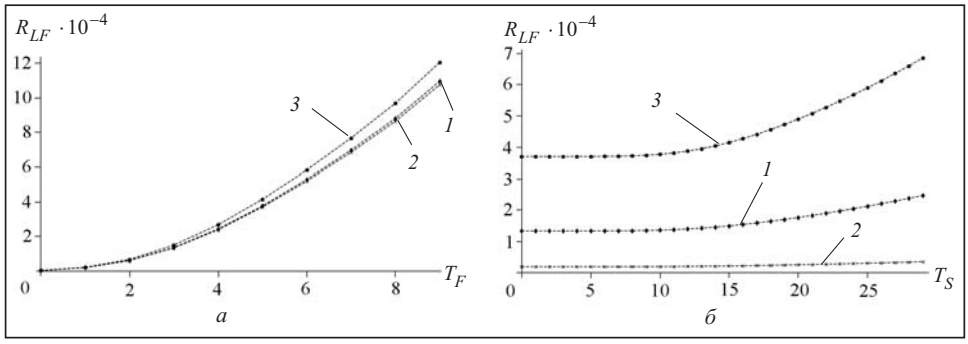

Рис. 4. Зависимость R_{LF} от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*а*), а также от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*), $T_F = 5$ (кривая *3*) (*б*)

Увеличение параметра T_F приводит к повышению интенсивности обслуживания L-заявок в группе F-серверов, поэтому функция R_{LF} возрастающая относительно этого параметра, при этом ее значения почти не зависят от фиксированного значения параметра T_S (рис. 4, *а*). Важно, что эта функция также возрастает относительно изменения параметра T_S при фиксированных значениях параметра T_F, при этом ее значения почти не зависят от значения параметра T_F, т.е. отличаются лишь в четвертом знаке после десятичной точки (рис. 4, *б*).

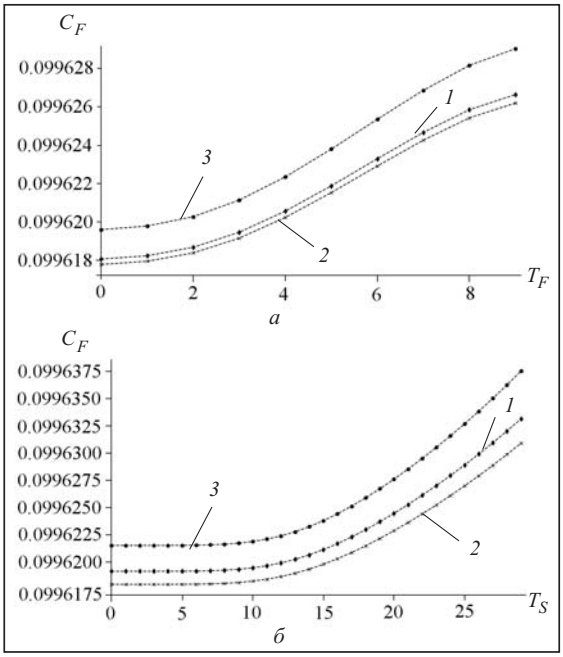

Рис. 5. Зависимость C_F от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*а*), а также от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*), $T_F = 5$ (кривая *3*) (*б*)

Функция C_F растет с очень малой скоростью относительно изменения обоих параметров: T_F и T_S, при этом ее значения отличаются в пятом знаке после десятичной точки (рис. 5, *а*, *б*). Функция C_S является постоянной относительно изменения параметра T_F при различных значениях параметра T_S (рис. 6, *а*), и она растет с малой скоростью

относительно параметра T_S, при этом ее значения полностью совпадают при различных значениях параметра T_F (рис. 6, *б*).

Задача оптимизации данной системы для указанных ранее исходных данных решена. Критерием задачи является минимизация суммарных штрафов (Total Cost, *TC*), связанных с потерей разнотипных вызовов. Пусть потеря одного *H*-вызова оценивается в 30 у.е., а аналогичная величина для *L*-вызовов равна 4 у.е. Тогда задача заключается в следующем: требуется найти такую пару (T_F^*, T_S^*), чтобы минимизировать суммарные штрафы. Формально эта задача записывается так:

$$TC(T_F, T_S) = 30PB_H(T_F, T_S) + 4PB_L(T_F, T_S) \to \min \quad (26)$$

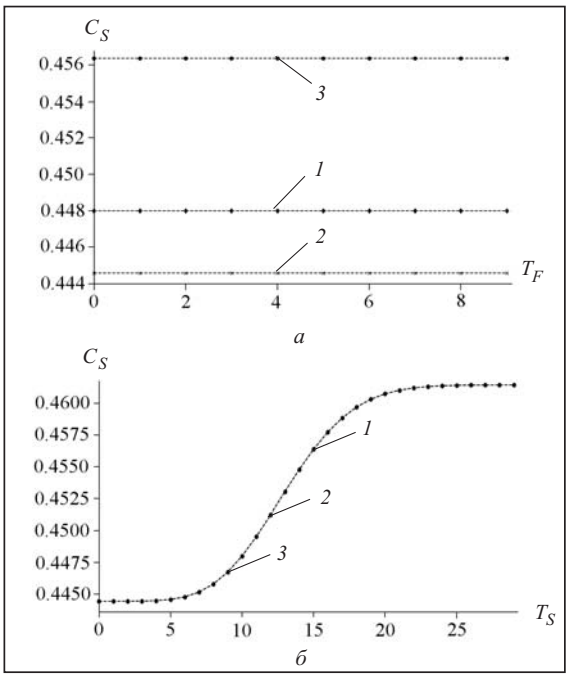

Рис. 6. Зависимость C_S от параметра T_F при $T_S = 22$ (кривая *1*), $T_S = 15$ (кривая *2*), $T_S = 7$ (кривая *3*) (*а*), а также от параметра T_S при $T_F = 3$ (кривая *1*), $T_F = 1$ (кривая *2*), $T_F = 5$ (кривая *3*) (*б*)

при ограничениях

$$1 \le T_F \le N_F - 1,\ 1 \le T_S \le N_S - 1. \quad (27)$$

Задача (26), (27) всегда имеет решение, так как множество допустимых решений является конечным дискретным множеством. Вследствие сложного вида целевой функции (26) не удается заранее определить ее выпуклость или вогнутость (унимодальность). Вместе с тем для решения этой задачи можно использовать метод простого перебора. Для выбранных исходных данных решением задачи (26), (27) является $(T_F^*, T_S^*) = (6, 29)$, при этом минимальное значение $TC_{\min}(6, 29) = 0.000276$.

ЗАКЛЮЧЕНИЕ

В работе изучена система обслуживания с гетерогенными серверами и двумя типами заявок. При наличии хотя бы одного сервера в соответствующих группах заявки высокого приоритета обслуживаются в высокоскоростных серверах, а заявки низкого приоритета — в низкоскоростных. В случае занятости всех серверов в соответствующих группах допускается обслуживание поступившей заявки в другой группе, при этом переназначения заявок осуществляются согласно рандомизированной схеме. Считается, что вероятности переназначения зависят от числа занятых серверов в соответствующей группе.

Показано, что математической моделью исследуемой системы является некоторая двумерная цепь Маркова, и разработаны методы точного и приближенного анализа характеристик этой системы. Точный анализ основан на использовании системы уравнений равновесия для вероятностей состояний, а приближенный — использует алгоритмы фазового укрупнения состояний двумерных цепей Маркова. Полученные формулы позволяют проводить численный анализ и оптимизацию характеристик системы.

СПИСОК ЛИТЕРАТУРЫ

1. Gumbel H. Waiting lines with heterogeneous servers. *Operations Research*. 1960. Vol. 8, Iss. 4. P. 504–511.
2. Singh V.S. Two-server Markovian queues with balking: Heterogeneous vs homogeneous servers. *Operations Research.* 1970. Vol. 18, Iss 1. P. 145–159.
3. Singh V.S. Markovian queues with three servers. *IIE Transactions.* 1971. Vol. 3, Iss. 1. P. 45–48.
4. Fakinos D. The $M / G / k$ blocking system with heterogeneous servers. *Journal of Operations Research Society*. 1980. Vol. 31, Iss. 10. P. 919–927.
5. Fakinos D. The generalized $M / G / k$ blocking system with heterogeneous servers. *Journal of Operations Research Society*. 1982. Vol. 33, Iss. 9. P. 801–809.
6. Nath G., Enns E. Optimal service rates in the multi-server loss system with heterogeneous servers. *Journal of Applied Probability*. 1981. Vol. 18, Iss. 3. P. 776–781.
7. Alpaslan F., Shahbazov A. An analysis and optimization of stochastic service with heterogeneous channels and Poisson arrivals. *Pure and Apllied Mathematika Science*. 1996. Vol. 43. P. 15–20.
8. Lin B.W., Elsayed E.A. A general solution for multichannel queueing systems with ordered entry. *Computers & Operations Research*. 1978. Vol. 5, Iss. 4. P. 219–225.
9. Elsayed E.A. Multichannel queueing systems with ordered entry and finite source. *Computers & Operations Research*. 1983. Vol. 10, Iss. 3. P. 213–222.
10. Yao D.D. The arrangement of servers in an ordered-entry system. *Operations Research*. 1987. Vol. 35, Iss. 5. P. 759–763.
11. Pourbabai B., Sonderman D. Server utilization factors in queueing loss systems with ordered entry and heterogeneous servers. *Journal of Applied Probability*. 1986. Vol. 23, Iss. 1. P. 236–242.
12. Pourbabai B. Markovian queueing systems with retrials and heterogeneous servers. *Computers & Mathematics with Applications.* 1987. Vol. 13, Iss. 12. P. 917–923.
13. Nawijn W.M. On a two-server finite queuing system with ordered entry and deterministic arrivals. *European Journal of Operations Research*. 1984. Vol. 18, Iss. 3. P. 388–395.
14. Nawijn W.M. A Note on many-server queueing systems with ordered entry with an application to conveyor theory. *Journal of Applied Probability*. 1983. Vol. 20. P. 144–152.
15. Yao D.D. Convexity properties of the overflow in an ordered-entry system with heterogeneous servers. *Operations Research Letters*. 1986. Vol. 5, Iss. 3. P. 145–147.
16. Isguder H.O., Kocer U.U. Analysis of $GI / M / n / n$ queueing system with ordered entry and no waiting line. *Applied Mathematical Modelling.* 2014. Vol. 38, Iss. 3. P. 1024–1032.
17. Larsen R.L., Agrawala A.K. Control of heterogeneous two-server exponential queueing system. *IEEE Transactions on Software Engineering*. 1983. Vol. SE-9, Iss. 4. P. 522–526.
18. Koole G. A simple proof of the optimality of a threshold policy in a two-server queueing system. *Systems & Control Letter*. 1995. Vol. 26, Iss. 5. P. 301–303.
19. Lin W., Kumar P.R. Optimal control of queueing system with two heterogeneous servers. *IEEE Transactions on Automatic Control*. 1984. Vol. 29, Iss. 8. P. 696–703.
20. Luh H.P., Viniotis I. Threshold control policies for heterogeneous servers systems. *Mathematical Methods in Operational Research*. 2002. Vol. 55, Iss. 1. P. 121–142.
21. Weber R. On a conjecture about assigning jobs to processors of different speeds. *IEEE Transactions on Automatic Control.* 1993. Vol. 38, Iss. 1. P. 166–170.
22. Efrosinin D. Sztrik J. Optimal control of a two-server heterogeneous queueing system with breakdowns and constant retrials. In: *Information Technologies and Mathematical Modelling — Queueing Theory and Applications. ITMM 2016. Communications in Computer and Information Science*. Dudin A., Gortsev A., Nazarov A., Yakupov R. (Eds.). 2016. Vol. 638. P. 57–72.
23. Efrosinin D. Sztrik J., Farkhadov M., Stepanova N. Reliability analysis of two-server heterogeneous queueing system with threshold control policy. In: *Information Technologies and Mathematical Modelling. Queueing Theory and Applications. ITMM 2017. Communications in Computer and Information Science*. Dudin A., Nazarov A., Kirpichnikov A. (Eds.). 2017. Vol. 800. P. 13–27.
24. Viniotis I., Ephremides A. Extension of the optimality of a threshold policy in heterogeneous multi-server queueing systems. *IEEE Transactions on Automatic Control*. 1988. Vol. 33, Iss. 1. P. 104–109.
25. Rosberg Z., Makowski A.M. Optimal routing to parallel heterogeneous servers — Small arrival rates. *IEEE Transactions on Automatic Control*. 1990. Vol. 35, Iss. 7. P. 789–796.

26. Rykov V.V. Monotone control of queueing systems with heterogeneous servers. *Queueing Systems.* 2001. Vol. 37. P. 391–403.
27. Rykov V.V., Efrosinin D. On the slow server problem. *Automation and Remote Control*. 2009. Vol. 70, Iss. 12. P. 2013–2023.
28. Neuts M.F. Matrix-geometric solutions in stochastic models: An algorithmic approach. Baltimore: John Hopkins University Press, 1981. 332 p.
29. Mitrani I., Chakka R. Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method. *Performance Evaluation*. 1995. Vol. 23, Iss. 3. P. 241–260.
30. Chakka R. Spectral expansion solution for some finite capacity queues. *Annals of Operations Research*. 1998. Vol. 79. P. 27–44.
31. Melikov A.Z., Ponomarenko L.A., Rustamov A.M. Hierarchical space merging algorithm to analysis of open tandem queueing networks. *Cybernetics and Systems Analysis*. 2016. Vol. 52, N 6. P. 867–877.
32. Melikov A.Z., Ponomarenko L.A., Kim C.S. Performance analysis and optimization of multi-traffic on communication networks. Berlin: Springer, 2010. 208 p.

Надійшла до редакції 22.01.2019

А.З. Меліков, Л.А. Пономаренко, Е.В. Мехбалієва
АНАЛІЗ МОДЕЛЕЙ СИСТЕМ З ГЕТЕРОГЕННИМИ СЕРВЕРАМИ

Анотація. Запропоновано математичну модель системи обслуговування з гетерогенними серверами і без черг за наявності вимог двох типів. Вимоги високого пріоритету обслуговуються у високошвидкісних серверах, а вимоги низького пріоритету — в низькошвидкісних. У випадках зайнятості всіх серверів у відповідних групах допускається обслуговування вимоги, що надійшла, в іншій групі, при цьому перепризначення вимог здійснюється згідно з рандомізованою схемою. Вважається, що ймовірності перепризначення залежать від кількості зайнятих серверів у відповідній групі. Розроблено методи точного і наближеного аналізу характеристик цієї системи. Отримано явні формули для наближеного обчислення її характеристик.

Ключові слова: гетерогенний сервер, система обслуговування, пріоритети, різнотипні вимоги, оптимізація.

A.Z. Melikov, L.A. Ponomarenko, E.V. Mekhbaliyeva
ANALYZING THE MODELS OF SYSTEMS WITH HETEROGENEOUS SERVERS

Abstract. The mathematical model of a queueing system with heterogeneous servers, without queues and two types of calls is investigated. High priority calls are processed in fast servers while low priority calls are processed in slow servers. If all servers in some group are busy then reassigning of calls to another group is allowed. Reassigning is based on random schemes and reassignment probability depends on the number of busy servers in appropriate group. Exact and approximate methods are developed for the analysis of characteristics of the system. Explicit approximate formulas to calculate the approximate values of characteristics are proposed.

Keywords: queuing-inventory system, heterogeneous servers, priority, customers of different types, optimization.

Меликов Агаси Зарбали оглы,
чл.-кор. НАН Азербайджана, доктор техн. наук, профессор, заведующий лабораторией Института систем управления НАН Азербайджана, Баку, e-mail: agassi.melikov@rambler.ru.

Пономаренко Леонид Анатольевич,
доктор техн. наук, профессор, главный научный сотрудник Международного научно-учебного центра информационных технологий и систем НАН Украины и МОН Украины, Киев, e-mail: laponomarenko@ukr.net.

Мехбалыева Эсмира Видади кызы,
кандидат техн. наук, докторант Сумгаитского государственного университета, Азербайджан.