Стохастические системы

© 2024 г. А.В. ГОРБУНОВА, канд. физ.-мат. наук (avgorbunova@list.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва),
А.В. ЛЕБЕДЕВ, д-р физ.-мат. наук (avlebed@yandex.ru)
(Московский государственный университет им. М.В. Ломоносова)

ОБ ОСОБЕННОСТЯХ УПРАВЛЕНИЯ СКОРОСТЬЮ ОБСЛУЖИВАНИЯ В СИСТЕМАХ С РАЗДЕЛЕНИЕМ И ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ ЗАЯВОК

Рассматривается классическая система с разделением и параллельным обслуживанием. Предлагается модель для определения оптимальной стоимости функционирования такой системы, учитывающая необходимость минимизации среднего времени отклика одновременно с разумными затратами на необходимые для этого ресурсы. Под термином "ресурсы" в рамках исследуемой математической модели подразумеваются интенсивности обслуживания на приборах, стоимость расходов на которые прямо пропорциональна производительности системы, т.е. скорости обслуживания заявок. Для частного случая, когда число подсистем равно двум, представлено точное аналитическое выражение для определения оптимальной стоимости; для более общего случая, когда число подсистем fork-join системы больше двух, получено уравнение, численное решение которого позволяет вычислить искомую величину. Кроме того, проведен асимптотический анализ поведения полученных решений.

Ключевые слова: система с разделением и параллельным обслуживанием заявок, система массового обслуживания, оптимальная стоимость, управление.

DOI: 10.31857/S0005231024120048, **EDN:** XUEPRH

1. Введение

Рассматривается классическая система с разделением и параллельным обслуживанием, называемая также в англоязычной литературе fork-join системой массового обслуживания (СМО). При поступлении в данную систему заявка разделяется на K одинаковых частей (подзаявок), количество которых соответствует числу подсистем. Каждая подсистема представляет собой систему с объемом накопителя бесконечной емкости и единственным прибором. Предполагается, что интенсивности обслуживания на всех имеющихся приборах идентичные. Заявка считается обслуженной после обслуживания всех составляющих ее частей. Соответственно, время отклика системы (время пребывания заявки в системе) определяется максимумом из K случайных времен пребывания подзаявок в подсистемах.

Подобные системы широко используются для моделирования различного рода процессов, в рамках которых происходит разделение или распараллеливание задачи, в частности в области информационных технологий речь может идти о параллельных или распределенных вычислениях, также с их помощью могут моделироваться различные рабочие процессы в производстве (например, сборка многокомпонентного заказа на складе или сложных, состоящих из множества деталей, механизмов на производстве), финансах (например, обработка заявки на кредитование в нескольких подразделения финансовой организации), здравоохранении (проведение необходимых анализов и сбор анамнеза при поступлении пациента в медицинское учреждение) и т.д.

Ключевой особенностью данной системы, затрудняющей ее анализ, является наличие зависимости между временами пребывания подзаявок в подсистемах. Поэтому несмотря на актуальность и востребованность изучения fork-join систем, точные характеристики функционирования системы были получены только для случая двух подсистем ($K = 2$), в частности, известна формула для среднего времени отклика системы в случае пуассоновского входящего потока и экспоненциальных распределений времен обслуживания [1]. В остальных же случаях найдены только приближенные выражения для аппроксимации основных показателей производительности системы [1–3]. В обзоре [4] можно более подробно ознакомится с известными результатами. Что касается более свежих исследований, то в работах [5–12], в том числе и авторов данной статьи, помимо выражений, уточняющих известные оценки среднего времени отклика или его дисперсии, получены выражения для оценки таких характеристик, как квантили распределения времени отклика для более широкого диапазона входящих или обслуживающих потоков. Кроме того, в [11] представлено точное выражение для коэффициента корреляции между временами пребывания подзаявок в подсистемах типа $M|M|1$, а в [9] для fork-join системы с распределением Парето времени обслуживания выведена оценка для коэффициента корреляции уже в случае распределения Парето времени обслуживания. Также стоит отметить работы отечественных авторов [13–19]. В серии статей, включающей [13–16], проведен анализ fork-join системы с бесконечнолинейными подсистемами в терминах производящих функций для распределения вероятностей количества подзаявок в каждой подсистеме. В [17] проанализирована одна из модификаций fork-join системы, когда при поступлении заявка разделяется не на фиксированное, а на переменное число подзаявок, которое определяется состоянием системы. В [18] предложен подход на основе инвариантов отношения для приближения среднего времени отклика fork-join СМО, а в [19] fork-join система является составной частью сети и используется для моделирования и исследования характеристик производительности сервисных платформ транзакционных услуг.

В настоящей статье исследуется другой аспект производительности fork-join системы, а именно строится модель стоимости функционирования систе-

мы с разделением и параллельным обслуживанием, что позволяет определить оптимальное управление с точки зрения оптимизации ее финансовых показателей. Базируется модель на естественных предположениях о необходимости минимизации среднего времени отклика системы для сохранения ее конкурентноспособности при разумных затратах на требуемые для этого ресурсы. В частности, под ресурсами может пониматься мощность необходимого оборудования, которое позволяет быстрее обрабатывать клиентский запрос, если речь идет об информационно-вычислительных или производственных системах, например. Понятно, что чем мощнее оборудование, тем больше затрат требуется на его покупку, техническое обслуживание и содержание в целом. Таким образом, скорость работы оборудования (или в терминах СМО интенсивности обслуживания) пропорциональна росту его стоимости. Кроме того, с увеличением скорости обслуживания уменьшается время отклика системы. Таким образом, стоимость функционирования системы складывается из оптимального баланса между временем отклика и скоростью работы обслуживающих приборов.

Более подробно, предполагается, что установлены: 1) цена (штраф) за единицу среднего времени отклика и 2) цена за единицу интенсивности обслуживания. При этом первая цена для простоты полагается равной единице. Далее вычисляются стоимости времени отклика и обслуживания и складываются в общую стоимость затрат, которую здесь хотим минимизировать. Таким образом, ставится задача стоимостной оптимизации управления.

Подобные постановки задач можно найти в монографии [20], посвященной оптимальному дизайну СМО, в том числе оптимальному выбору скоростей поступления и обслуживания заявок (в предположении, что эти параметры управляемы) для различных систем и сетей массового обслуживания. Однако для fork-join систем такие задачи ранее не рассматривались.

Предложенная функциональная зависимость для стоимости функционирования системы с разделением и параллельным обслуживанием позволяет получить в явном виде выражение для определения оптимальной интенсивности обслуживания на приборах системы в случае, если количество подсистем равно двум. Если же число подсистем больше двух, то получение оптимального решения возможно в численном виде. Также было рассмотрено поведение оптимального решения в предельных случаях.

Статья организована следующим образом: в разделе 2 описана математическая модель определения оптимальной стоимости функционирования fork-join системы в общем виде; в следующих двух разделах выводится оптимальное значение коэффициента загрузки системы, позволяющего выразить оптимальное значение интенсивности обслуживания для случая формулы Нельсона–Тантави определения времени отклика системы – в частном случае, т.е. когда число подсистем соответствует двум и формула точна, и в более общем случае, когда формула является приближенной; далее проанализирован случай обобщения формулы Нельсона–Тантави, а затем проведен

асимптотический анализ полученных решений, как в общем случае, так и для конкретных выражений; в заключении подводятся некоторые итоги.

2. Математическая модель определения оптимальной стоимости функционирования fork-join системы

Анализируется fork-join система с пуассоновским входящим потоком с интенсивностью $\lambda > 0$ и экспоненциальным распределением времен обслуживания на $K \geqslant 2$ однородных приборах с интенсивностью $\mu > 0$. Коэффициент загрузки системы равен $\rho = \lambda/\mu < 1$.

Изложим единый математический подход, который будет далее применяться в работе.

Обозначим стоимость функционирования системы через S и введем функцию $f(\rho)$, которая определяет выражение для среднего времени отклика системы в случае $\lambda = 1$, т.е. $f(\rho) = E[R_K]$ при $\lambda = 1$. Тогда в общем случае будет справедливо следующее выражение:

$$E[R_K] = \frac{1}{\lambda} f(\rho).$$

Поскольку стоимость функционирования системы S зависит от среднего времени отклика системы (цену за единицу времени принимаем за единицу) и стоимости затрат на обслуживание, то можем для нее записать:

$$S = E[R_K] + c\mu,$$

где c – это стоимость единицы интенсивности обслуживания. Соответственно, с учетом введенной функции $f(\rho)$ можем переписать данное выражение следующим образом:

$$S = \frac{1}{\lambda} f(\rho) + c\frac{\lambda}{\rho} = \frac{1}{\lambda}\left(f(\rho) + \frac{c\lambda^2}{\rho}\right).$$

Пусть $c_1 = c\lambda^2$, тогда

$$S = \frac{1}{\lambda}\left(f(\rho) + \frac{c_1}{\rho}\right). \tag{1}$$

Далее для нахождения оптимального значения уровня загрузки системы определим точку экстремума функции стоимости $S(\rho)$, а именно точку минимума. Для этого найдем производную последнего выражения и приравниваем ее к нулю

$$S'_\rho = \frac{1}{\lambda}\left(f'(\rho) - \frac{c_1}{\rho^2}\right) = 0,$$

откуда получим уравнение

$$f'(\rho)\rho^2 = c_1, \tag{2}$$

решив которое, найдем оптимальное значение ρ_0 и, соответственно, оптимальное значение интенсивности обслуживания $\mu_0 = \lambda/\rho_0$.

Рассмотрим базовый случай, когда $K = 1$, т.е. фактически систему $M|M|1$. Среднее время отклика в такой системе равно

$$E[R] = \frac{1}{\mu - \lambda} = \frac{1}{\lambda}\frac{1}{\mu/\lambda - 1}.$$

Тогда функция $f(\rho)$ с учетом того, что $\rho = \lambda/\mu$, определяется как

$$f(\rho) = \frac{1}{1/\rho - 1} = \frac{\rho}{1 - \rho} = \frac{1}{1 - \rho} - 1,$$

а производная этой функции имеет вид

$$f'(\rho) = \frac{1}{(1 - \rho)^2}.$$

Теперь подставляем полученные выражения в уравнение (2)

$$\left(\frac{\rho}{1 - \rho}\right)^2 = c_1,$$

откуда получаем значение для искомой оптимальной загрузки системы при $K = 1$

$$\rho_0 = \frac{\sqrt{c_1}}{1 + \sqrt{c_1}}. \tag{3}$$

На следующем шаге можем вычислить оптимальную интенсивность обслуживания, а именно

$$\mu_0 = \frac{\lambda}{\rho_0} = \lambda\left(1 + \frac{1}{\sqrt{c_1}}\right) = \lambda + \frac{1}{\sqrt{c}}.$$

Аналогичная задача была решена в [20, §1.1] для случая, когда в стоимости учитывается не среднее время отклика, а среднее время ожидания.

Подчеркнем, что поскольку получено уравнение (2) относительно загрузки ρ, то в дальнейшем будем искать только оптимальную загрузку, понимая, что ее можно при необходимости пересчитать в оптимальную интенсивность обслуживания.

3. Анализ fork-join СМО с двумя подсистемами типа $M|M|1$. Формула Нельсона–Тантави

Для начала разберем частный случай и определим оптимальное значение коэффициента загрузки системы, когда число подсистем равно двум. Для

среднего времени отклика при $K = 2$ известна точная формула, полученная в [1], а именно

$$E[R_2] = \frac{12-\rho}{8}\frac{1}{\mu-\lambda} = \frac{1}{\lambda}\frac{12-\rho}{8}\frac{\rho}{1-\rho} = \frac{1}{\lambda}\frac{\rho(12-\rho)}{8(1-\rho)}.$$

Таким образом, имеем

$$f(\rho) = \frac{\rho(12-\rho)}{8(1-\rho)}.$$

Далее находим производную по ρ и, подставляя выражение для $f'(\rho)$ в (2), можем записать следующее уравнение:

$$\frac{\rho^2(\rho^2-2\rho+12)}{(1-\rho)^2} = 8c_1.$$

Для удобства сделаем замену

$$c_2 = 8c_1$$

и после упрощения получаем уравнение четвертой степени

$$\rho^4 - 2\rho^3 + (12-c_2)\rho^2 + 2c_2\rho - c_2 = 0, \tag{4}$$

которое и будем решать, поскольку, как известно, для уравнения четвертой степени существует аналитическое решение в радикалах. Для этого воспользуемся методом Феррари [21–23].

Введем следующие обозначения:

$$A = -2, \quad B = 12 - c_2, \quad C = c_2, \quad D = -c_2,$$

тогда получим

$$\rho^4 + A\rho^3 + B\rho^2 + C\rho + D = 0.$$

С помощью замены $\rho = y - A/4$ сведем уравнение четвертой степени (4) к каноническому виду

$$y^4 + A_1y^2 + B_1y + C_1 = 0, \tag{5}$$

где

$$A_1 = B - \frac{3A^2}{8} = -c_2 + \frac{21}{2},$$

$$B_1 = \frac{A^3}{8} - \frac{AB}{2} + C = c_2 + 11,$$

$$C_1 = -\frac{3A^4}{256} + \frac{A^2B}{16} - \frac{AC}{4} + D = -\frac{1}{4}c_2 + \frac{45}{16}.$$

Далее согласно методу Феррари необходимо найти одно частное действительное решение кубического уравнения

$$A_2t^3 + B_2t^2 + C_2t + D_2 = 0, \tag{6}$$

где

$$A_2 = 2, \quad B_2 = -A_1 = c_2 - \frac{21}{2},$$

$$C_2 = -2C_1 = \frac{1}{2}c_2 - \frac{45}{8}, \quad D_2 = A_1C_1 - \frac{B_1^2}{4} = -\frac{175}{16}c_2 - \frac{23}{32}.$$

С помощью замены $t = z - B_2/(3A_2)$ приводим уравнение (6) к каноническому виду уравнения третьей степени

$$z^3 + A_3 z + B_3 = 0, \tag{7}$$

где

$$A_3 = \frac{3A_2C_2 - B_2^2}{3A_2^2} = -\frac{1}{12}c_2^2 + 2c_2 - 12,$$

$$B_3 = \frac{2B_2^3 - 9A_2B_2C_2 + 27A_2^2D_2}{27A_2^3} = \frac{1}{108}c_2^3 - \frac{1}{3}c_2^2 - \frac{3}{2}c_2 - 16.$$

Вещественный корень уравнения (7) согласно методу Кардано [21–23] определяется следующим образом:

$$z_1 = \left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} + \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}}, \quad Q > 0, \tag{8}$$

где

$$Q = \left(\frac{A_3}{3}\right)^3 + \left(\frac{B_2}{2}\right)^2 = -\frac{11}{432}c_2^4 + \frac{11}{12}c_2^3 - \frac{55}{16}c_2^2 + 44c_2. \tag{9}$$

При этом не забываем, что должно выполняться условие

$$\left(-\frac{B_3}{2} + \sqrt{Q}\right)^{\frac{1}{3}} \left(-\frac{B_3}{2} - \sqrt{Q}\right)^{\frac{1}{3}} = -\frac{A_3}{3}.$$

Отметим, что для случая $Q < 0$ для z_1 после преобразований в конечном счете будет справедлива следующая формула:

$$z_1 = 2\sqrt{-\frac{A_3}{3}}\cos\frac{w}{3}, \tag{10}$$

где

$$w = \begin{cases} \operatorname{arctg}\left(\dfrac{-2\sqrt{-Q}}{B_3}\right), & \text{если } B_3 < 0, \\ \operatorname{arctg}\left(\dfrac{-2\sqrt{-Q}}{B_3}\right) + \pi, & \text{если } B_3 > 0, \\ \dfrac{\pi}{2}, & \text{если } B_3 = 0, \end{cases}$$

а для случая $Q = 0$ имеем

$$z_1 = 2\left(-\frac{B_3}{2}\right)^{\frac{1}{3}}. \tag{11}$$

Выражение для Q из (9) преобразуется к виду

$$-\frac{11}{432}c_2\left(c_2^3 - 36c_2^2 + 135c_2 - 1728\right)$$

и меняет свой знак в зависимости от значения c_2 (хотя c_2 должно быть положительным исходя из своего физического смысла), поэтому явно укажем области знакопостоянства выражения для Q, нули которого можно опять же получить с помощью метода Кардано, а именно

$$\begin{aligned} &Q > 0, \quad \text{если } c_2 \in (0; \tilde{c}_2)\ , \\ &Q < 0, \quad \text{если } c_2 \in (-\infty; 0) \cup (\tilde{c}_2; +\infty)\ , \end{aligned}$$

где $\tilde{c}_2 = (3 \times 11^{\frac{1}{3}} + 3 \times 11^{\frac{2}{3}} + 12) \approx 33{,}51$.

Соответственно, получаем

$$t_1 = z_1 - \frac{B_2}{3A_2},$$

где z_1 в зависимости от значения Q определяется выражениями (8), (10) или (11). Частное решение t_1 позволяет представить каноническое уравнение четвертой степени (5) в виде произведения двух квадратных трехчленов

$$\left(y^2 - y\sqrt{2t_1 - A_1} + \frac{B_1}{2\sqrt{2t_1 - A_1}}\right)\left(y^2 + y\sqrt{2t_1 - A_1} - \frac{B_1}{2\sqrt{2t_1 - A_1}}\right) = 0.$$

Решение одного из квадратных уравнений (с учетом обратной замены) будет являться искомым решением исходного уравнения четвертой степени (4). В частности, проверка показала, что искомым является корень второго уравнения, поэтому можем записать итоговое решение как

$$y_0 = y_3 = \frac{-\sqrt{2t_1 - A_1} + \sqrt{Dis_2}}{2},$$

где дискриминант второго уравнения равен

$$Dis_2 = -2t_1 - A_1 + \frac{2B_1}{\sqrt{2t_1 - A_1}}$$

и, соответственно,

$$\rho_0 = y_0 - \frac{A}{4} = y_0 + \frac{1}{2}.$$

4. Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Формула Нельсона–Тантави

Выведем уравнение для определения оптимального значения коэффициента загрузки системы в общем случае, когда $K > 2$. Для математического ожидания времени отклика fork-join системы приближенная формула Нельсона–Тантави, которая считается одной из наиболее точных среди известных, имеет вид [1]

$$E[R_K] \approx \left[\frac{H_K}{H_2} + \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right)\rho\right]\frac{12-\rho}{8}\frac{1}{\mu-\lambda}, \tag{12}$$

где $H_K = \sum_{i=0}^{K} 1/i$ – это частичная сумма гармонического ряда.

Будем решать задачу для приближения Нельсона–Тантави в предположении, что формула (12) является точной.

Введем следующие обозначения:

$$H = \frac{H_K}{H_2}, \quad M = \frac{4}{11}\left(1 - \frac{H_K}{H_2}\right) = \frac{4}{11}(1-H).$$

Тогда

$$E[R_K] = \frac{1}{\lambda}(H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

следовательно,

$$f(\rho) = (H + M\rho)\frac{12-\rho}{8}\frac{\rho}{1-\rho},$$

а для $f'(\rho)$ после преобразований получаем

$$f'(\rho) = \frac{1}{8(1-\rho)^2}(2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H). \tag{13}$$

Затем подставляем полученное выражение в (2)

$$\frac{\rho^2}{8(1-\rho)^2}(2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H) = c_1.$$

Также для удобства вводим обозначение $c_2 = 8c_1$ и получаем уравнение пятой степени, которое после преобразований с учетом того, что $M = 4(1-H)/11$, сводится к виду

$$\begin{aligned} 8(H-1)\rho^5 + (60-71H)\rho^4 + (118H-96)\rho^3 + \\ + 11(c_2 - 12H)\rho^2 - 22c_2\rho + 11c_2 = 0. \end{aligned} \tag{14}$$

Полученное уравнение (14) можно решить численно и определить оптимальное значение ρ_0 и, соответственно, искомое значение μ_0.

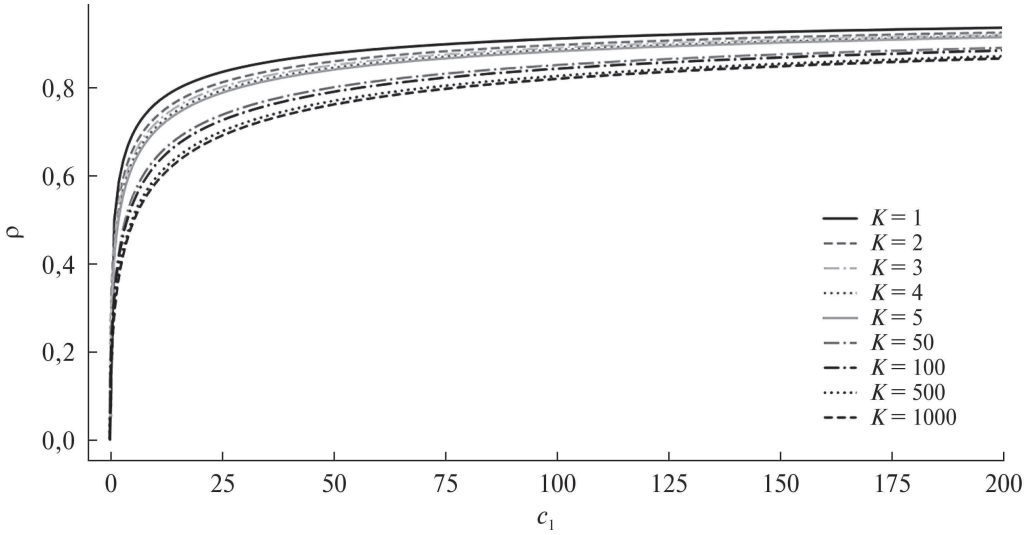

Рис. 1. График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (14)) от параметра c_1 для различного числа подсистем K.

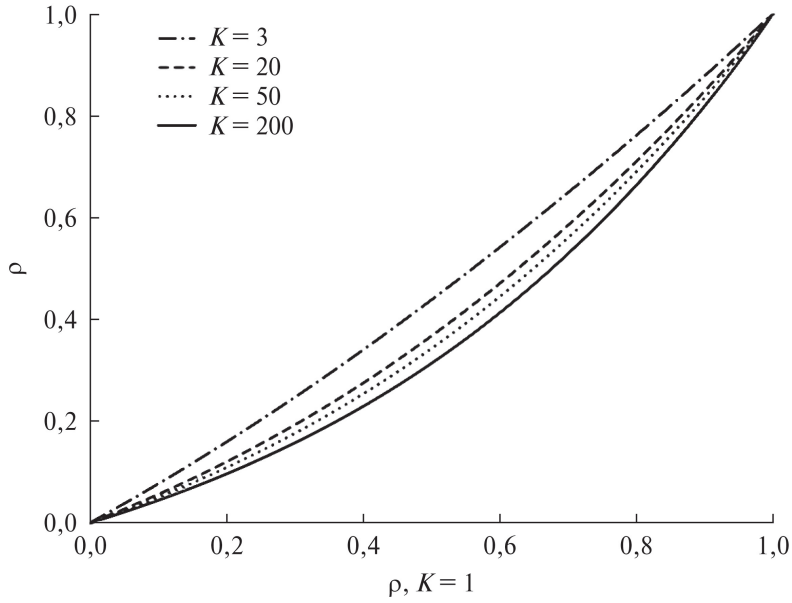

Рис. 2. График зависимости оптимального значения загрузки системы ρ в зависимости от оптимального значения загрузки ρ при $K = 1$.

На рис. 1 представлены графики зависимости поведения оптимального значения коэффициента загрузки $\rho = \rho_0$, являющегося решением уравнения (14), от значения параметра c_1 для различного числа подсистем K fork-join СМО. На начальном этапе наблюдается довольно стремительный рост требуемого уровня загрузки с увеличением цены единицы ресурса и, соответственно, производительности системы в целом. Причем даже для числа

подсистем $K = 1000$, т.е. иными словами, при условии разделения задачи и обработки ее подзадач на довольно большом количестве устройств, уже при $c_1 > 5{,}5$ необходимый уровень загрузки будет $\rho > 0{,}5$. Одновременно с этим при дальнейшем росте параметра c_1 наблюдается довольно медленный рост требуемой загрузки, например, при $c_1 \approx 158$, что почти в 29 раз больше $c_1 = 5{,}5$, будет значение $\rho_0 = 0{,}85$, т.е. все еще не превышает 90%. Кроме того, исходя из вида графиков, наблюдается эффект того, что с ростом K оптимальный уровень загрузки на систему снижается, что было ожидаемо и вполне естественно.

На рис. 2 представлен график зависимости поведения оптимального значения коэффициента загрузки системы от оптимального значения ρ при $K = 1$ согласно (3). График позволяет сравнить уровень оптимальной загрузки для различных значений $K \geqslant 2$ с уровнем оптимальной загрузки в случае $K = 1$. Как видно, с увеличением числа подсистем уровень требуемой загрузки для оптимальной работы системы падает, т.е. линия все сильнее прогибается под прямой $\rho = \rho_{K=1}$ и становится все более выпуклой (ниже любой хорды, соединяющей любые две точки на графике, выбранные на рассматриваемом интервале).

5. Анализ fork-join СМО с $K > 2$ подсистемами типа $M|M|1$. Обобщение формулы Нельсона–Тантави

Рассмотрим обобщение формулы Нельсона–Тантави из [10], в которой было показано, что уточненная формула дает лучшее приближение для среднего времени отклика. Улучшение достигается за счет некоторой поправки к выражению из (12), которое теперь обозначим через $E[R_K]_{NT}$. Итак, аппроксимация имеет вид

$$E[R_K] = \frac{\rho}{\mu - \lambda}\left(\frac{H_K}{H_2} - 1\right)\left(Q_1 - Q_2\left(\frac{H_K}{H_2} - 1\right) + Q_3\rho\right) + E[R_K]_{NT}, \tag{15}$$

где

$$Q_1 \approx 0{,}087197, \quad Q_2 \approx 0{,}070236, \quad Q_3 \approx 0{,}09638.$$

Определим выражение для $f(\rho)$ в данном случае, для этого вынесем $1/\lambda$ за скобки

$$E[R_K] = \frac{1}{\lambda}\left[-\frac{\rho^2}{1-\rho}\left(1 - \frac{H_K}{H_2}\right)\left(Q_1 + Q_2\left(1 - \frac{H_K}{H_2}\right) + Q_3\rho\right) + \lambda E[R_K]_{NT}\right],$$

в итоге получим

$$f(\rho) = -\frac{\rho^2}{1-\rho}\left(1 - \frac{H_K}{H_2}\right)\left(Q_1 + Q_2\left(1 - \frac{H_K}{H_2}\right) + Q_3\rho\right) + \lambda E[R_K]_{NT},$$

где $\lambda E[R_K]_{NT}$ было рассчитано ранее и имеет вид

$$\lambda E[R_K]_{NT} = f_2(\rho) = (H + M\rho)\frac{12 - \rho}{8}\frac{\rho}{1-\rho}.$$

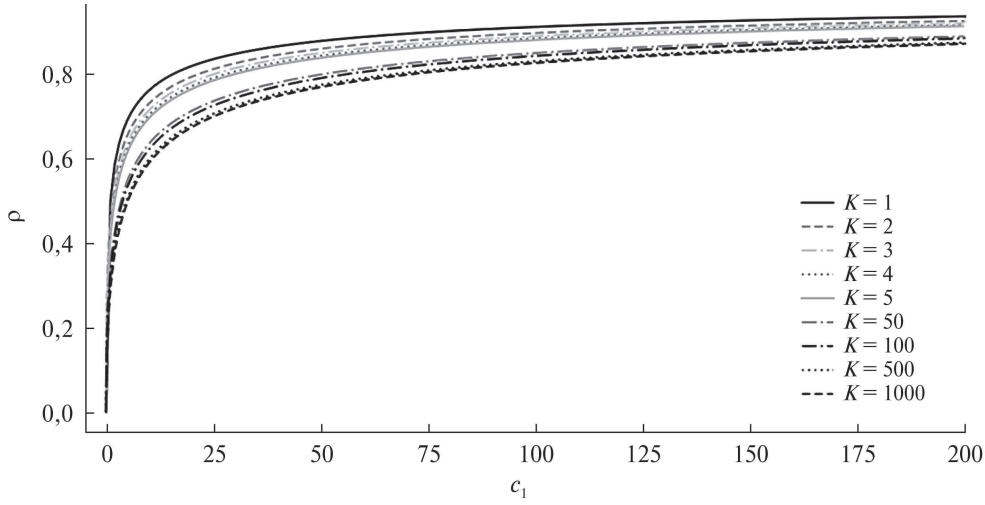

Рис. 3. График зависимости значения оптимального уровня загрузки системы ρ (решение уравнения (16)) от параметра c_1 для различного числа подсистем K.

Таким образом, с учетом предложенной ранее замены $H = H_K/H_2$ имеем

$$f(\rho) = -\frac{\rho^2}{1-\rho}(1-H)(Q_1 + (1-H)Q_2 + Q_3\rho) + f_2(\rho) = f_1(\rho) + f_2(\rho).$$

Далее берем производную от полученного выражения

$$f'(\rho) = f_1'(\rho) + f_2'(\rho),$$

где

$$f_1'(\rho) = \frac{\rho(1-H)}{(1-\rho)^2}\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big),$$

$$f_2'(\rho) = \frac{1}{8(1-\rho)^2}\Big(2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H\Big).$$

Соответственно, после подстановки полученных выражений в (2) получаем, что

$$c_1 = \rho^2 f'(\rho) = \rho^2(f_1'(\rho) + f_2'(\rho)). \tag{16}$$

Уравнение (16), как и в случае с формулой Нельсона–Тантави из предыдущего раздела, является уравнением пятой степени, которое может быть решено численно, после чего будет найдено оптимальное значение μ_0.

На рис. 3 представлены графики оптимального значения загрузки $\rho = \rho_0$, являющегося решением (16). Поведение графиков в целом аналогично случаю формулы Нельсона–Тантави (12). Чтобы более детально проанализировать разницу между полученными результатами, проведем сравнение поведения оптимального решения для частных случаев числа подсистем $K = 20$ и $K = 200$.

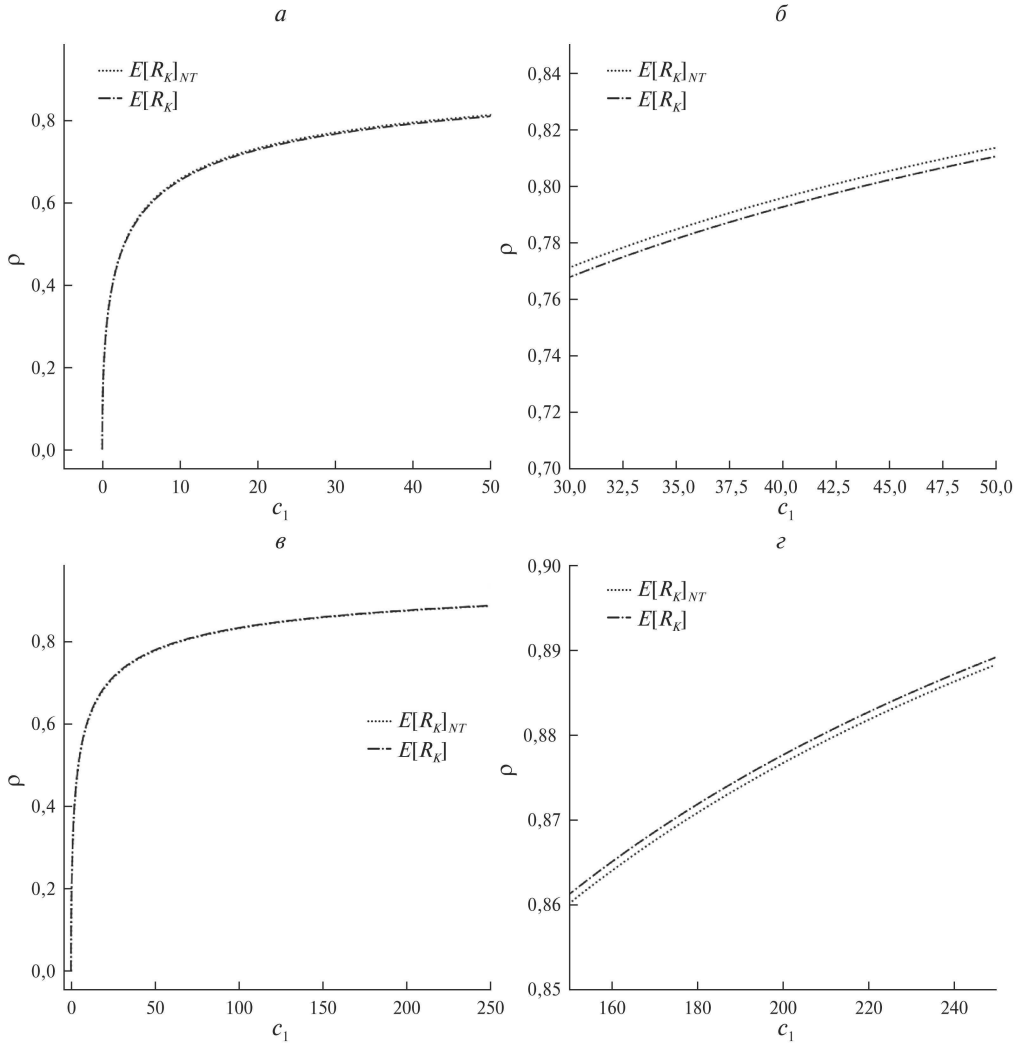

Рис. 4. Графики оптимального значения загрузки $\rho = \rho_0$, являющегося решением уравнения (16) при $K = 20$ (рис. 4,*а*,*б*) и $K = 200$ (рис. 4,*в*,*г*).

На рис. 4 представлены графики зависимости значения оптимального уровня загрузки системы $\rho = \rho_0$ от параметра c_1 для случаев формулы Нельсона–Тантави для среднего времени отклика системы $E[R_K]_{NT}$ из (12) и уравнения (14), а также обобщения формулы Нельсона–Тантави для среднего времени отклика системы $E[R_K]$ из (15) и уравнения (16) при $K = 20$ (рис. 4,*а*,*б*) и $K = 200$ (рис. 4,*в*,*г*), в том числе и в масштабе (рис. 4,*б*,*г*).

На рисунках 4,*а*,*б* для $K = 20$ видно, что для случая формулы Нельсона–Тантави (12) для оценки среднего времени отклика системы оптимальное значение загрузки ρ превышает оптимальное значение, рассчитанное по обобщенной формуле (15). Однако для графиков 4,*в*,*г* при $K = 200$ можно наблюдать обратную ситуацию.

Для того чтобы обстоятельно разобраться в поведении оптимальных решений в обоих случаях, далее проанализируем их асимптотику.

6. Асимптотика поведения оптимального решения

Рассмотрим уравнение (2) и определим поведение его решения при стремлении $c_1 \to 0$ ($\rho \to 0$) и $c_1 \to +\infty$ ($\rho \to 1$) в общем виде. Для начала проанализируем случай $\rho \to 0$, соответственно $c_1 \to 0$. Тогда имеем

$$\rho^2 f'(\rho) \sim \rho^2 f'(0),$$

далее подставляем полученное в (2), т.е. $\rho^2 f'(\rho) = c_1$, откуда следует, что

$$\rho \sim \sqrt{\frac{c_1}{f'(0)}}, \quad c_1 \to 0. \tag{17}$$

Теперь проанализируем случай $\rho \to 1$, соответственно $c_1 \to +\infty$. В общем случае имеем

$$\rho^2 f'(\rho) \sim f'(\rho).$$

Таким образом, если существует такое число $L \in (0, +\infty)$, что

$$L = \lim_{\rho \to 1} (1-\rho)^2 f'(\rho),$$

то

$$f'(\rho) \sim \frac{L}{(1-\rho)^2},$$

поэтому при подстановке полученного выражения для $f'(\rho)$ с L в (2) и учетом асимптотики имеем следующее:

$$\begin{gathered} \rho^2 f'(\rho) = c_1, \\ f'(\rho) \sim c_1, \quad \rho \to 1,\ c_1 \to +\infty, \\ (1-\rho)^2 \sim \frac{L}{c_1}, \quad \rho \to 1,\ c_1 \to +\infty, \end{gathered}$$

поэтому

$$\rho = 1 - \sqrt{\frac{L}{c_1}}(1 + o(1)), \quad c_1 \to \infty. \tag{18}$$

Далее предметно рассмотрим имеющиеся модели, а именно случаи формулы Нельсона–Тантави и его обобщение, и определим конкретные выражения для полученных эквивалентностей из общего случая.

Для формулы Нельсона–Тантави при $K \geqslant 2$ с учетом выражения (13) справедливо

$$f'(0) = \frac{3}{2}\frac{H_K}{H_2},$$

поэтому при подстановке в (17) для $c_1 \to 0$ ($\rho \to 0$) получим

$$\rho \sim \sqrt{\frac{2}{3}\frac{H_2}{H_K}c_1}, \quad c_1 \to 0. \tag{19}$$

Теперь определим асимптотику решения при $c_1 \to +\infty$ ($\rho \to 1$). Для этого найдем значение L:

$$L = \lim_{\rho\to 1}(1-\rho)^2 f'(\rho) = \\ = \lim_{\rho\to 1}\frac{2M\rho^3 + (H-15M)\rho^2 + (24M-2H)\rho + 12H}{8} = \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}. \tag{20}$$

Окончательно получаем

$$\rho = 1 - \sqrt{\frac{1}{c_1}\left(\frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}\right)}(1 + o(1)).$$

Теперь проанализируем обобщение формулы Нельсона–Тантави. Аналогично рассмотрим уравнение (16) и определим поведение его решения при стремлении $c_1 \to 0$ ($\rho \to 0$) и $c_1 \to \infty$ ($\rho \to 0$). Проанализируем случай $\rho \to 0$, соответственно, $c_1 \to 0$.

Для формулы (15) при $K \geqslant 2$ справедливо

$$f'(0) = f_1'(0) + f_2'(0) = \frac{3}{2}\frac{H_K}{H_2},$$

поэтому, как и ранее,

$$\rho \sim \sqrt{\frac{2}{3}\frac{H_2}{H_K}c_1}, \quad c_1 \to 0.$$

Теперь проанализируем случай $\rho \to 1$ ($c_1 \to \infty$). Для (15) при $K \geqslant 2$ получим

$$L = \lim_{\rho\to 1}(1-\rho)^2 f'(\rho) = \lim_{\rho\to 1}(1-\rho)^2(f_1'(\rho) + f_2'(\rho)) = L_1 + L_2.$$

Фактически значение L_2 было вычислено ранее и соответствует значению L из (20)

$$L_2 = \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}.$$

Теперь рассчитаем L_1:

$$L_1 = \lim_{\rho\to 1}(1-\rho)^2 f_1'(\rho) = \\ = \lim_{\rho\to 1}\rho(1-H)\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big) = \\ = (H-1)(Q_1 - (H-1)Q_2 + Q_3) = \left(\frac{H_K}{H_2} - 1\right)\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] \approx \\ \approx \left(\frac{H_K}{H_2} - 1\right)\left[0{,}183577 - 0{,}070236\left(\frac{H_K}{H_2} - 1\right)\right].$$

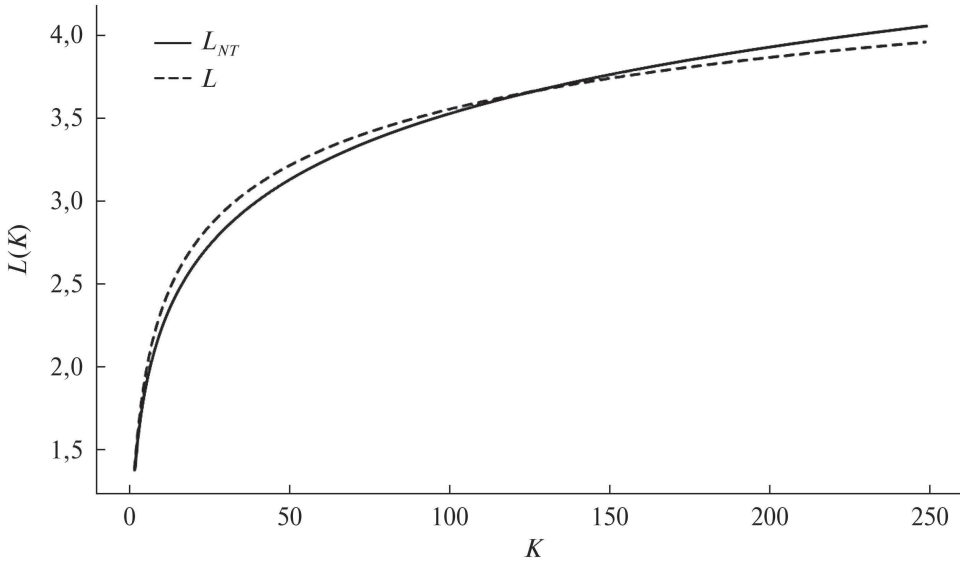

Рис. 5. График зависимости значения L от K для случая формулы Нельсона–Тантави (12) и случая обобщающей формулы (15).

Таким образом, получаем, что

$$\rho = 1 - \sqrt{\frac{L}{c_1}}(1 + o(1)),$$

где

$$L = \left(\frac{H_K}{H_2} - 1\right)\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] + \frac{7}{8}\frac{H_K}{H_2} + \frac{1}{2}. \tag{21}$$

На рис. 5 представлен график зависимости значения L от числа подсистем для случая формулы Нельсона–Тантави (20) и случая обобщающей формулы (21). После значения $K = 126$ происходит "перелом" графиков, и если сперва значение L_{NT} превышало значение L для обобщающей формулы, то для значений $K \geqslant 127$ ситуация поменялась с точностью до наоборот, что в том числе и подтверждается графиками, представленными ранее, для оптимального значения загрузки в зависимости от параметра c_1 на рис. 4 для $K = 20$ и $K = 200$.

Что касается практического применения полученных асимптотических результатов, то оба случая, как высокой ($\rho \to 1$), так и низкой ($\rho \to 0$) загрузки, могут представлять интерес. Так, например, в системах с интенсивным использованием данных параллельные и распределенные вычисления получили широкое распространения как один из основных способов повышения производительности при обработке больших данных. Поэтому владельцы подобных высокопроизводительных вычислительных сред заинтересованы в получении прогнозов поведения системы при пиковых нагрузках, ровно как и в обратных

ситуациях, когда платформы наименее всего востребованы, с целью выработки стратегий эффективной работы систем. В этом контексте, асимптотические формулы могут подсказать, когда (при каких ценах на единицу скорости обслуживания) подобные режимы (высокой и низкой загрузки) оказываются оптимальными, т.е. фактически желательными при функционировании системы, в смысле стоимости, а когда нет.

7. Заключение

В работе исследуется система с разделением и параллельным обслуживанием с точки зрения управления оптимальной стоимостью ее функционирования в зависимости от цены единицы ресурса, влияющего на производительность системы и, соответственно, время ее отклика. Построена математическая модель, учитывающая оптимальное соотношение стоимости и эффективности работы системы. Проведен анализ на базе полученных ранее аппроксимаций для среднего времени отклика, как одной из наиболее известных и точных, так и ее обобщения, полученного авторами данной статьи. Для частного случая системы, т.е. когда число подсистем равно двум, удается вывести в явном виде выражения для оптимальной загрузки системы. Для большего числа подсистем представлены уравнения пятой степени, численное решение которых позволяет определить искомые значения загрузки на систему и, соответственно, интенсивности обслуживания, которая фактически характеризует “мощность“ необходимых ресурсов. Также проанализировано асимптотическое поведение системы.

Поскольку используемые здесь приближения среднего времени отклика являются довольно точными, с их помощью также хорошо описывается суммарная стоимость затрат, поэтому при расчетных значениях оптимальной загрузки (и, соответственно, оптимальной скорости обслуживания) стоимость будет близка к оптимальной. Поэтому расчетные значения загрузки можно рекомендовать для использования в качестве оценок их фактических значений для fork-join систем. Кроме того, если есть оценка оптимальной загрузки при одном числе подсистем, ее можно пересчитать в оценку для другого числа подсистем (см. рис. 2). Поскольку в общем случае задача решается только численно, то в случаях высоких и низких цен на единицу скорости обслуживания можно рекомендовать использовать асимптотические формулы, которые дают более простые, но при этом более грубые оценки оптимальной загрузки.

Предложенная в статье математическая модель для управления оптимальным функционированием системы будет справедлива и в более общих случаях fork-join СМО, а именно с отличным от пуассоновского входящим потоком и/или отличным от экспоненциального распределением времени обслуживания на приборах, как в [9]. При этом, разумеется, для вывода конкретных соотношений или уравнений необходимы аналитические приближения для величины среднего времени отклика системы. Кроме того, можно рассмотреть модели с нелинейной зависимостью стоимости затрат на обслуживание

от скорости обслуживания (например, степенной зависимостью, как упомянуто в [20, §1.1]).

СПИСОК ЛИТЕРАТУРЫ

1. *Nelson R., Tantawi A.N.* Approximate analysis of fork/join synchronization in parallel queues // IEEE Transact. Comput. 1988. V. 37. P. 739–743.
2. *Varma S., Makowski A.M.* Interpolation approximations for symmetric fork-join queues // Performance Evaluat. 1994. V. 20. P. 245–265.
3. *Varki E., Merchant A., Chen H.* The M/M/1 fork-join queue with variable subtasks // Unpublished. Available online.
https://www.cs.unh.edu/?varki/publication/2002-nov-open.pdf
4. *Thomasian A.* Analysis of fork/join and related queueing systems // ACM Computing Surveys (CSUR). 2014. V. 47. No. 2. P. 1–71.
5. *Qiu Z., Pérez J. F., Harrison P. G.* Beyond the mean in fork-join queues: Efficient approximation for response-time tails // Performance Evaluat. 2015. V. 91. P. 99–116.
6. *Nguyen M., Alesawi S., Li N., Che H., Jiang H.* ForkTail: A black-box fork-join tail latency prediction model for user-facing datacenter workloads // Proc. 27th Int. Symp. High-Perform. Parallel Distrib. Comput. 2018. P. 206–217.
7. *Nguyen M., Alesawi S., Li N., Che H., Jiang H.* A Black-Box Fork-Join Latency Prediction Model for Data-Intensive Applications // IEEE Transact. Parall. Distrib. Syst. 2020. V. 31. No. 9. P. 1983–2000.
8. *Enganti P., Rosenkrantz T., Sun L., Wang Z., Che H., Jiang H.* ForkMV: Mean-and-Variance Estimation of Fork-Join Queuing Networks for Datacenter Applications // Proc. IEEE International Conference on Networking, Architecture and Storage (NAS). 2022. P. 1–8.
9. *Gorbunova A.V., Lebedev A.V.* Nonlinear Approximation of Characteristics of a Fork-Join Queueing System with Pareto Service as a Model of Parallel Structure of Data Processing // Math. Comput. Simulat. 2023. V. 214. P. 409–428.
10. *Gorbunova A.V., Lebedev A.V.* On Estimating the Characteristics of a Fork-Join Queueing System with with Poisson Input and Exponential Service Times // Advanc. Syst. Sci. Appl. 2023. V. 23. P. 99–114.
11. *Gorbunova A.V., Lebedev A.V.* Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with $M|M|1$-type Subsystems // Advanc. Syst. Sci. Appl. 2024. V. 24. No. 2. P. 1–18.
12. *Gorbunova A.V., Lebedev A.V.* Copulas and Quantiles in Fork-Join Queueing Systems // Advanc. Syst. Sci. Appl. 2024. V. 24. No. 1. P. 1–19.
13. *Ивановская И.А., Моисеева С.П.* Исследование математической модели параллельного обслуживания заявок смешанного типа // Изв. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2010. Т. 317. № 5. С. 32–34.
14. *Жидкова Л.А., Моисеева С.П.* Исследование системы параллельного обслуживания кратных заявок простейшего потока // Вест. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2011. Т. 17. № 4. С. 49–54.

15. *Моисеева С.П., Захорольная И.А.* Математическая модель параллельного обслуживания кратных заявок с повторными обращениями // Автометрия. 2011. Т. 47. № 6. С. 51–58.

16. *Моисеева С.П., Панкратова Е.В., Убонова Е.Г.* Исследование бесконечнолинейной системы массового обслуживания с разнотипным обслуживанием и входящим потоком марковского восстановления // Вест. Том. политехн. ун-та. Управление, вычислительная техника и информатика. 2016. Т. 35. № 2. С. 46–53.

17. *Осипов О.А.* Система обслуживания с делением и слиянием требований, в которой требование занимает все свободные обслуживающие приборы // Вест. Росс. ун-та дружбы народов. Серия: Математика. Информатика. Физика. 2018. Т. 26. № 1. С. 28–38.

18. *Хабаров Р.С., Лохвицкий В.А., Дудкин А.С.* Аппроксимация времени пребывания для системы массового обслуживания fork-join на основе инвариантов отношения // Интеллектуальные технологии на транспорте. 2020. Т. 22. № 2. С. 46–50.

19. *Редругина Н.М.* Метод вычисления временных характеристик обслуживания в сервисных платформах инфокоммуникационных транзакционных услуг с параллельной обработкой запросов // Тр. уч. заведений связи. 2023. Т. 9. № 3. С. 82–90.

20. *Stidham S.* Optimal design of queueing systems. Boca Raton: CRC Press/Taylor & Francis, 2009. 384 p.

21. *Spiegel M.R., Lipschutz, S. Liu J.* Mathematical Handbook of Formulas and Tables, McGraw Hill Professional, 3rd (Third) edition // Schaum's Outline Series. 2008. 312 p.

22. *Курош А.Г.* Алгебраические уравнения произвольных степеней (Популярные лекции по математике; вып. 7). Изд. 2-е. М. : Наука, 1975. 32 с.

23. *Курош А.Г.* Курс высшей алгебры: Учебник. Изд. 17-е. СПб.: Лань, 2008. 432 с.

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 07.10.2024

После доработки 16.10.2024

Принята к публикации 18.10.2024