О НОВОМ ПОДХОДЕ К ОЦЕНКЕ КВАНТИЛЕЙ ВРЕМЕНИ ОТКЛИКА СИСТЕМЫ С РАЗДЕЛЕНИЕМ И ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ ЗАЯВОК

Горбунова А. В.[1]
(*Институт проблем управления им. В.А. Трапезникова РАН, Москва*)
Лебедев А. В.[2]
(*Московский государственный университет имени М.В. Ломоносова, Москва*)

Предлагается новый подход к оценке квантилей распределения времени отклика системы массового обслуживания с разделением и параллельным обслуживанием заявок, для обозначения которой в англоязычной литературе используется термин fork-join. Рассматривается классический вариант данной системы с пуассоновским входным потоком и экспоненциальными временами обслуживания на однородных приборах. Заявки при поступлении в систему мгновенно разделяются на фиксированное число подзаявок и отправляются на обслуживание в соответствующие подсистемы с накопителем неограниченной емкости и одним прибором. Заявка считается обслуженной после обслуживания всех ее компонентов. Данная система позволяет смоделировать множество реальных процессов, для которых с целью увеличения эффективности характерно разделение крупных задач на более мелкие составляющие, например, системы параллельных или распределенных вычислений. Сложность анализа систем заключается в наличии зависимости между временами пребывания подзаявок, что значительно затрудняет анализ всех показателей производительности таких систем. Основным вкладом статьи является подход к определению квантилей распределения времени отклика, оценка которых является не менее ценной по сравнению с оценкой среднего значения времени отклика. При этом вычислению математического ожидания посвящено гораздо большее количество работ в данной области, что объясняется в том числе сложностью проведения подобного анализа даже для данной характеристики, а оценка квантилей представляется еще более трудоемкой задачей.

Ключевые слова: система с параллельным обслуживанием заявок, fork-join система массового обслуживания, время отклика, квантили распределения, имитационное моделирование.

[1] *Анастасия Владимировна Горбунова, к.ф.-м.н.* (*avgorbunova@list.ru*).
[2] *Алексей Викторович Лебедев, д.ф.-м.н.* (*avlebed@yandex.ru*).

1. Введение

В работе исследуется классическая fork-join система массового обслуживания (СМО) с пуассоновским входным потоком и экспоненциальными временами обслуживания. Fork-join СМО является математической моделью для множества реально существующих систем, в которых происходит распараллеливание задач. При поступлении в систему заявка разделяется (fork point) на число подзаявок, равное числу подсистем $K \geqslant 2$. Все подсистемы фактически представляют собой самостоятельные системы массового обслуживания с бесконечной очередью и единственным прибором. Каждая из подзаявок поступает в одну из подсистем, обслуживается там, после чего попадает в условный буфер синхронизации (join point), где ожидает обслуживания оставшихся частей заявки. После окончания обслуживания всех подзаявок происходит мгновенная сборка целой заявки, и она может покинуть систему.

Ранее такие системы изучались авторами в работах [10, 11]. Так, в статье [11] с помощью искусственных нейронных сетей аппроксимируются математическое ожидание и среднеквадратическое отклонение времени отклика fork-join СМО для различного числа подсистем K. При этом был смоделирован большой объем данных (который используется далее в [10] и в настоящей работе) в форме наборов из 20 времен пребывания подзаявок от каждой заявки при значениях загрузки от 0,1 до 0,9 с шагом 0,05, по 5 миллионов заявок при загрузке менее 0,5 и по 10 миллионов при загрузке от 0,5 и выше. Наличие таких наборов позволило оценить характеристики систем для всех $3 \leqslant K \leqslant 20$. Приближение нейронной сетью показало среднюю абсолютную относительную ошибку для среднего времени отклика 0,74%, а для среднеквадратического отклонения – 0,36%, для рассмотренных в работе выборок.

В [10] для оценки аналогичных характеристик использовалось расширение предложенного подхода на основе комбинации различных методов машинного обучения, кроме того обсужда-

лись вопросы имитационного моделирования системы, а также построения доверительных интервалов. С помощью выведенных формул удалось повысить точность оценки среднего времени отклика более чем в 11 раз по сравнению с классической формулой Нельсона – Тантави [16], а точность оценки среднеквадратического отклонения – более чем в 25 раз по сравнению с приближением на основе независимых случайных величин. Была изучена также корреляционная функция последовательности времен отклика заявок, зависимость которых приводит к увеличению дисперсии оценки среднего, и выведена приближенная формула этой дисперсии.

Текущая работа продолжает исследование зависимости времен пребывания подзаявок (частей одной заявки). С анализом fork-join системы, но с более сложной архитектурой, можно ознакомиться, например, в одной из последних работ [9]. Заметим, что погрешность приближений различных характеристик оценивается с помощью результатов имитационного моделирования системы, точность которых обеспечивается большим числом прогонов имитационной модели.

Интерес к исследованию fork-join систем объясняется наличием множества реальных физических систем, функционирование которых может быть ими смоделировано, особенно в области моделирования информационных систем и протекающих в них процессов. Разумеется, область применения fork-join моделей не ограничивается только информационно-вычислительными системами. Вопросы оптимизации процессов в производственных системах (например, сборка заказов в складских системах, изготовление многокомпонентных изделий) или повышение эффективности организации процесса пребывания пациентов в медицинских учреждения и т.д. продолжают оставаться востребованными до настоящего момента [1, 2, 7, 12, 20].

Другой причиной актуализации исследований fork-join систем является появление новых методов и подходов к анализу сложных систем массового обслуживания, в частности, подхода, основанного на применении методов машинного обучения и его

различные модификации [3, 4, 5, 19]. В данном случае речь идет о дальнейшем развитии названного подхода и включении в его состав графического анализа и методов оптимизации.

Несмотря на кажущуюся простоту функционирования, исследование fork-join СМО относится к труднорешаемым задачам. Основная причина сложности заключается в общности моментов появления подзаявок в подсистемах, что делает времена их пребывания зависимыми случайными величинами. В этом и есть основное отличие fork-join системы от просто параллельно функционирующих СМО того же типа, что и подсистемы fork-join СМО. Поэтому точные результаты были получены только для среднего времени отклика в случае двух подсистем с пуассоновским входным потоком и экспоненциальными временами обслуживания [16]. Для других вариантов архитектур fork-join СМО, под которыми подразумевается, например, увеличение числа подсистем или приборов в них, ограниченная емкость накопителей или другие типы распределений для входных и обслуживающих потоков, получены только аппроксимации математического ожидания времени отклика различными способами.

Что касается оценки других характеристик производительности fork-join системы, то исследований в этом направлении гораздо меньше [6, 11]. Однако помимо первых или вторых моментов случайной величины времени отклика интерес представляют квантили ее распределения. Установление квантиля заданного уровня вероятности означает, что система гарантирует обслуживание заявки не более чем за установленное время с данной вероятностью. Такой подход к оценке работы системы уместен, если важна срочность обслуживания (например, в медицине) или если долгое обслуживание вызывает недовольство клиентов и это не компенсируется быстрым обслуживанием других. Таким образом, оценка квантилями является более тонкой, чем оценка средним, в смысле качества обслуживания в современном мире.

В [18] с помощью векторно-матричной техники и фазовых распределений получены теоретические оценки хвоста и квантилей высоких уровней. В [17] были получены оценки для кванти-

лей высоких уровней в условиях высокой загрузки системы для нескольких типов распределений и архитектур fork-join СМО.

В данной работе предлагается подход для нахождения квантилей времени отклика различного уровня для более широкого диапазона загрузок, что позволяет составить полноценное представление о поведении исследуемой случайной величины. Несмотря на то, что объектом исследования статьи является классическая fork-join СМО, предложенный в статье подход может быть распространен и на другие архитектуры fork-join систем и не только.

2. Математическая модель fork-join СМО

Опишем более подробно процесс функционирования fork-join системы. Будем рассматривать частный случай двух подсистем ($K = 2$), однако заметим, что количество подсистем никак не влияет на зависимость в любой паре времен пребывания подзаявок одной заявки (рис. 1). В систему поступает пуассоновский поток заявок с интенсивностью $\lambda > 0$. В момент поступления в систему заявка мгновенно разделяется на 2 подзаявки, каждая из которых попадает в соответствующую подсистему, имеющую накопитель неограниченной емкости и один прибор. Все приборы являются однородными, время обслуживания имеет экспоненциальное распределение с параметром $\mu > 0$.

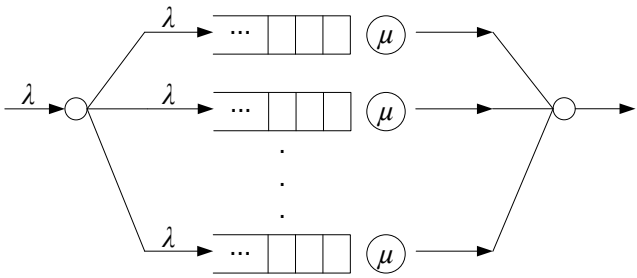

Рис. 1. Модель fork-join системы массового обслуживания с подсистемами типа $M_\lambda|M_\mu|1$

Таким образом, подсистемы представляют собой две идентичных СМО типа $M|M|1$. Поскольку заявка считается обслуженной только после окончания обслуживания обеих составляющих ее подзаявок, то случайное время пребывания заявки в СМО (время отклика) R является максимумом из двух случайных времен пребывания подзаявок ξ_i, $i = 1, 2$, в каждой из двух подсистем:

$$R = \max\{\xi_1, \xi_2\}. \tag{1}$$

Случайные величины ξ_1 и ξ_2 являются коррелированными в силу того, что все подзаявки (части одной заявки) поступают в подсистемы в одно и то же время.

Обозначим через $\rho = \lambda/\mu$ коэффициент загрузки системы и далее для простоты положим $\lambda = 1$, $\mu = 1/\rho$. Такие параметры использовались при имитационном моделировании в [10, 11].

Отметим, что различные коэффициенты корреляции в общем случае хотя и отражают зависимость, но лишь частично. В полной мере отражают зависимость только копулы.

3. Подход к оценке диагонального сечения и квантилей распределения времени отклика

Для приближения квантилей распределения случайной величины времени отклика $R = \max\{\xi_1, \xi_2\}$ воспользуемся элементами теории копул.

Копулой C называется функция многомерного распределения на $[0, 1]^d$, $d \geqslant 2$, если все частные распределения являются равномерными на $[0, 1]$. Согласно знаменитой теореме Скляра, любая функция многомерного распределения в $\mathbb{R}^d$ представима в виде

$$F(x_1, \dots x_d) = C(F_1(x_1), \dots F_d(x_d)),$$

где F_i, $1 \leqslant i \leqslant d$, — функции частных распределений. Таким образом, всякому многомерному распределению можно поставить в соответствие его копулу. Если частные распределения непрерывны, то такое представление единственно. В качестве классического учебника по копулам укажем [15], а в качестве работ авторов,

связанных с копулами, можно указать [8, 13, 14]. Далее ограничимся случаем двумерных копул ($d = 2$).

Диагональным сечением (двумерной) копулы называется функция $\delta(u) = C(u, u)$, $u \in [0, 1]$. Она обладает следующими (необходимыми и достаточными) свойствами:

$$\max\{2u - 1, 0\} \leqslant \delta(u) \leqslant u; \tag{2}$$
$$0 \leqslant \delta(u_2) - \delta(u_1) \leqslant 2(u_2 - u_1), \quad 0 \leqslant u_1 \leqslant u_2 \leqslant 1.$$

Смысл изучения диагональных сечений, например, в следующем. Если заданы случайные величины X_1 и X_2 с одинаковыми частными распределениями $F_1 = F_2 = F$ и копулой совместного распределения C, то их максимум $X_{max} = \max\{X_1, X_2\}$ имеет функцию распределения

$$F_{max}(x) = P(X_1 < x, X_2 < x) = C(F(x), F(x)) = \delta(F(x)), \tag{3}$$

так что для ее вычисления достаточно знать только диагональное сечение, а не всю копулу.

Легко заметить, что условиям (2) удовлетворяет степенная функция

$$\delta(u) = u^{\alpha}, \quad 1 \leqslant \alpha \leqslant 2,$$

тогда случаю $\alpha = 1$ соответствует совершенная положительная зависимость (комонотонность), а случаю $\alpha = 2$ – независимость случайных величин.

Будем рассматривать двумерную копулу $C(u_1, u_2)$ случайных векторов времен пребывания в подсистемах (ξ_1, ξ_2). Каждая компонента случайного вектора имеет экспоненциальное распределение с функцией распределения $F(x) = 1 - e^{-(\mu-\lambda)x}$, $x \geqslant 0$. Тогда в соответствии с теоремой Скляра представление с помощью копулы совместного распределения (ξ_1, ξ_2) существует и единственно:

$$F_{\xi_1,\xi_2}(x_1, x_2) = P(\xi_1 < x_1, \xi_2 < x_2) = C(F(x_1), F(x_2)).$$

В силу (3) получаем

$$F_R(x) = C(F(x), F(x)) = \delta(F(x)), \tag{4}$$

где $\delta(u) = C(u,u)$ — диагональное сечение копулы, что дает нам уравнение для квантили уровня p распределения времени отклика

$$F_R(x_p) = \delta(F(x_p)) = p,$$

поэтому

$$x_p = F_R^{-1}(p) = F^{-1}(\delta^{-1}(p)). \tag{5}$$

Учитывая метод обратного преобразования для генерации случайных величин с заданной функцией распределения, рассмотрим

$$U_i = 1 - e^{-(\mu-\lambda)\xi_i}, \quad i = 1,2.$$

Эти случайные величины будут иметь равномерное распределение на отрезке $[0,1]$, т.е. $U_i \sim R[0,1]$. Тогда

$$V = \max\{U_1, U_2\} = 1 - e^{-(\mu-\lambda)\cdot\max\{\xi_1,\xi_2\}} = 1 - e^{-(\mu-\lambda)R}. \tag{6}$$

Диагональное сечение копулы можно оценить следующим образом. Имеем

$$\delta(u) = C(u,u) = P(U_1 < u, U_2 < u) =$$

$$= P(\max(U_1, U_2) < u) = P(V < u) = p,$$

т.е.

$$\delta(u_p) = P(V < u_p) = p,$$

где u_p — это квантиль распределения с. в. V. С помощью реализаций V_i случайной величины V, полученных посредством имитационного моделирования значений случайных времен пребывания в fork-join СМО R_i и дальнейшей подстановкой их в формулу (6), строим оценку диагонального сечения $\delta(u)$, а фактически вероятностей p. Иными словами, строим эмпирическую оценку диагонального сечения с помощью квантилей распределения V.

Для этого упорядочиваем полученные посредством симуляции величины V: $V_{(1)}$, $V_{(2)}$,..., $V_{(N)}$, где $V_{(k)}$ — это k-я порядковая статистика, $k = 1,...,N$, и по точкам $(V_{(k)}, k/(N+1))$ определяем оценки (u_p, p) для значений вероятности из интересующего нас интервала $p \in \{0{,}2; 0{,}25; 0{,}30; ...; 0{,}90\}$, при конкретном фиксированном значении коэффициента загрузки

$\rho \in \{0{,}10; 0{,}15; 0{,}20; ...; 0{,}90\}$. Выбор значений p обусловлен тем, что, как правило, больший интерес представляют квантили более высокого уровня, поэтому значения p начинаем рассматривать с $0{,}2$. Далее на основе имеющихся данных будем строить прогноз вероятностей p в зависимости от квантилей u_p и коэффициента загрузки ρ:

$$p \approx \widehat{p} = \widehat{\delta}(u_p, \rho).$$

Теперь для определения вида функциональной зависимости проведем графический анализ полученных данных. Прежде всего было замечено, что зависимость p от u_p хорошо описывается степенной функцией, что соответствует линейной зависимости для логарифмов (см. рис. 2).

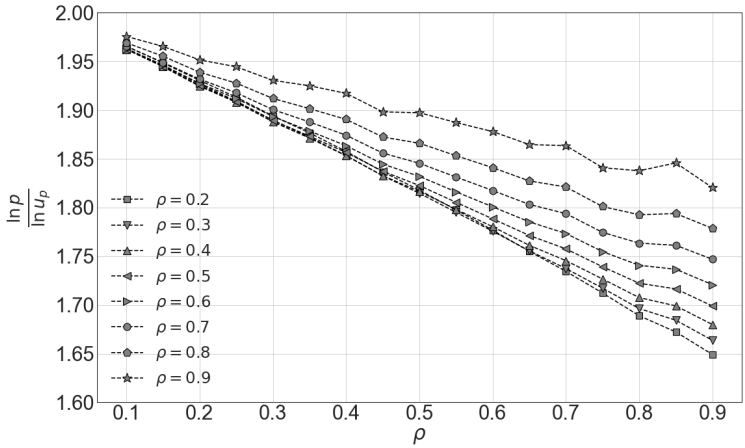

Рис. 2. Зависимость $(\ln p / \ln u_p)$ *от* ρ.

Зависимость показателя степени α от ρ также оказалась близка к линейной (см. рис. 3). Отметим, что при $\rho \to 0$ времена пребывания подзаявок асимптотически независимы, откуда $\alpha \to 2$.

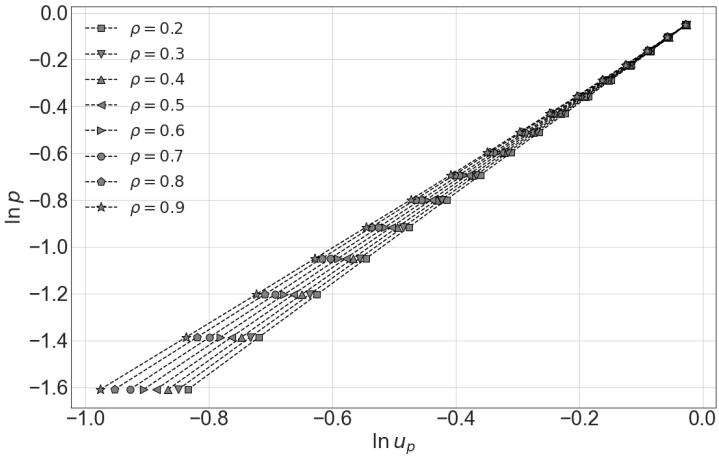

Рис. 3. Зависимость $\ln p$ *от* $\ln u_p$

Как видно из рис. 3, график зависимости напоминает пучок близких прямых, проходящих через точку $(0, 2)$, поэтому естественно предположить (в качестве первого приближения), что

$$\frac{\ln p}{\ln u_p} \approx 2 - C \cdot \rho,$$

а следовательно

$$p = \delta(u_p, \rho) \approx u_p^{2-C\cdot\rho}. \tag{7}$$

Остается только вычислить значение коэффициента C. Для этого будем минимизировать методом Нелдера – Мида модуль относительной погрешности аппроксимации относительно данных имитационного моделирования, в результате чего получим значение

$$C \approx 0{,}370608. \tag{8}$$

Таким образом, имеем

$$p = \delta(u_p, \rho) \approx u_p^{2-0{,}370608\cdot\rho}. \tag{9}$$

Сравнение результатов имитационного моделирования вероятностей или уровней p квантилей u_p случайной величины

$V = F(R)$ с результатами вычислений по аналитической формуле (9) в диапазоне $[0{,}20; 0{,}95]$ с шагом $0{,}05$, т.е. фактически для 272 рассчитанных значений p, позволяет говорить о хорошем уровне приближения. Так, средняя погрешность приближения составляет около $0{,}44\%$, в то время как максимальное и минимальное значение погрешности аппроксимации составляет $1{,}68\%$ и $0{,}003\%$ соответственно.

Теперь, учитывая (5), можем записать

$$\delta^{-1}(p) = F(x_p), \tag{10}$$

при этом из (7) следует, что $\delta^{-1}(p) \approx p^{\frac{1}{2-C\cdot\rho}}$. Подставляем оценку $\delta^{-1}(p)$ в (10) и получаем соотношение

$$p^{\frac{1}{2-C\cdot\rho}} = 1 - e^{-(\mu-\lambda)x_p},$$

откуда следует, что квантиль уровня p распределения случайной величины времени отклика fork-join СМО R определяется выражением

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-C\cdot\rho}})}{\mu - \lambda}. \tag{11}$$

Далее аналогично оценим качество аппроксимации полученного выражения (11) для 272 рассчитанных значений квантилей при $\rho \in \{0{,}10; 0{,}15; ...; 0{,}90\}$ и $p \in \{0{,}20; 0{,}25; ...; 0{,}90\}$. Получаем, что максимум модуля относительной ошибки составляет около $3{,}12\%$, а среднее значение этого модуля равно примерно $0{,}73\%$.

Ради уточнения аппроксимации квантилей, возвращаясь к рис. 2, можно отметить зависимость наклона прямых от p. Это наводит на мысль вместо константы C в (11) использовать выражения вида $(C_1 - C_2 p)$ или $(C_1 - C_2 p^2)$. Подбор констант методом Нелдера – Мида и сравнительный анализ точности показывают, что лучше второй вариант, а именно, приближение

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-(C_1-C_2p^2)\rho}})}{\mu - \lambda}, \tag{12}$$

где $C_1 \approx 0{,}390327$, $C_2 \approx 0{,}237842$. При этом погрешность (максимум модуля относительной ошибки) составляет всего $0{,}62\%$, что меньше прежнего в 4,6 раза.

4. Заключение

Данная работа продолжает авторский цикл, посвященный исследованию характеристик fork-join-систем с пуассоновским входным потоком и экспоненциальными временами обслуживания. Несмотря на простоту систем и давность их исследования (с 1980-х годов) в этой области все еще много неясного. Точных результатов здесь мало, многие оценки нуждаются в улучшении. Есть вопросы, которыми мало кто или вообще никто не занимался. Исследования в основном сосредоточены на среднем времени отклика, в то время как интерес представляют также дисперсия, квантили и др.

Ключевой проблемой является наличие зависимости между временами пребывания подзаявок (частей одной заявки), обусловленной общностью входного потока в подсистемы. Эта зависимость хотя и не очень сильная, но оказывает существенное влияние на характеристики, при этом она далека от описания классическими моделями (например, многомерным нормальным распределением, линейной регрессией и т.п.). Поэтому авторы в последних работах сосредоточились на изучении данной зависимости. Случай двух подсистем рассматривался исходя из того, что при любом числе подсистем для любой их пары времена пребывания подзаявок будут иметь такое же совместное распределение. В настоящей работе изучались приближения совместного распределения времен пребывания подзаявок с помощью теории копул. Получено хорошее соответствие с данными для степенных диагональных сечений. На основе оценок диагональных сечений выведены оценки квантилей времени отклика в широком диапазоне уровней и загрузок.

Развитый при этом подход, основанный на теории копул, можно попытаться обобщить на системы с большим числом подсистем или случай более сложных подсистем (например, с тяжелыми хвостами распределений времен обслуживания).

Литература

1. ARMONY M., ISRAELIT S., MANDELBAUM A., MARMOR Y.N. et al. *Patient flow in hospitals: a data-based queueing-science perspective* // Stochastic Systems. – 2015. – Vol. 5, No. 1. – P. 146–194.
2. BACCELLI F., MAKOWSKI A.M. *Queueing models for systems with synchronization constraints* // Proc. of the IEEE. – 1989. – Vol. 77, No. 1. – P. 138–161.
3. BARON O., KRASS D., SHERZER E., SENDEROVICH A. *Can machines solve general queueing problems?* // Winter Simulation Conference (WSC) – 2022. – P. 2830–2841.
4. CHOCRON E., COHEN I., FEIGIN P. *Delay prediction for managing multiclass service systems: An investigation of queueing theory and machine learning approaches* // IEEE Trans. on Engineering Management. – 2022. — P. 1–11.
5. DIELEMAN A., BERKHOUT A., HEIDERGOTT B. *A neural network approach to performance analysis of tandem lines: The value of analytical knowledge* // Computers & Operations Research. – 2023. – Vol. 152. – P. 106–124.
6. ENGANTI P., ROSENKRANTZ T., SUN L., WANG Z. et al. *ForkMV: Mean-and-Variance Estimation of Fork-Join Queuing Networks for Datacenter Applications* // IEEE Int. Conf. on Networking, Architecture and Storage (NAS). – 2022. – P. 1–8.
7. GALLIEN J., WEIN L.M. *A simple and effective component procurement policy for stochastic assembly Systems* // Queueing Systems. – 2001. – Vol. 38. – P. 221–248.
8. GORBUNOVA A.V., LEBEDEV A.V. *Bivariate distributions of maximum remaining service times in fork-join infinite-server queues* // Probl. Inf. Transm. – 2020. –Vol. 56, No. 1. – P. 73–90. – DOI: https://doi.org/10.1134/S003294602001007X.

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* // Mathematics and Computers in Simulation. – 2023. – Vol. 214. – P. 409–428. – DOI: https://doi.org/10.1016/j.matcom.2023.07.029.
10. GORBUNOVA A.V., LEBEDEV A.V. *On estimating the characteristics of a fork-join queueing system with Poisson input and exponential service times* // Advances in Systems Science and Applications. – 2023. – Vol. 23, No. 2. – P. 99–114. – DOI: https://doi.org/10.25728/assa.2023.23.2.1351.
11. GORBUNOVA A.V., VISHNEVSKY V.M. *Estimating the response time of a cloud computing system with the help of neural networks* // Advances in Systems Science and Applications. – 2020. – Vol. 20, No. 3. – P. 105–112.
12. JIANG L., GIACHETTI R.E. *A queueing network model to analyze the impact of parallelization of care on patient cycle time* // Health Care Management Science. – 2008. – Vol. 11. – P. 248–261.
13. LEBEDEV A.V. *Upper bound for the expected minimum of dependent random variables with known Kendall's tau* // Theory of Probability and Its Applications. – 2019. – Vol. 64, No. 3. – P. 465–473.
14. LEBEDEV A.V. *On the Interrelation between Dependence Coefficients of Bivariate Extreme Value Copulas* // Markov Proc. Relat. Fields. – 2019. – Vol. 25, No. 4. – P. 639–648.
15. NELSEN R.B. *An introduction to copulas.* – Springer Science & Business Media, 2007. – 272 p.
16. NELSON R., TANTAWI A.N. *Approximate analysis of fork/join synchronization in parallel queues* // IEEE Trans. Comput. – 1988. – Vol. 37, No. 6. – P. 739–743.
17. NGUYEN M., ALESAWI S., LI S., CHE S. et al. *A black-box fork-join latency prediction model for data-intensive applications* // IEEE Trans. on Parallel and Distributed Systems. – 2020. – Vol. 31, No. 9. – P. 1983–2000.

18. QIU ZH., PEREZ J.F., HARRISON P.G. *Beyond the mean in fork-join queues: Efficient approximation for response-time tails* // Performance Evaluation. – 2015. – Vol. 91. – P. 99–116.
19. VISHNEVSKY V.M., GORBUNOVA A.V. *Application of machine learning methods to solving problems of queuing theory* // Communications in Computer and Information Science. – 2022. – Vol. 1605. – P. 304–316. – DOI: https://doi.org/10.1007/978-3-031-09331-9_24.
20. SCHOL D., VLASIOU M., ZWART B. *Large fork-join queues with nearly deterministic arrival and service times* // Mathematics of Operations Research. – 2022. – Vol. 47, No. 2. – P. 1335–1364. – DOI: https://doi.org/10.1287/moor.2021.1171.

ON A NEW APPROACH TO ESTIMATING RESPONSE TIME QUANTILES OF A FORK-JOIN QUEUEING SYSTEM

Anastasia Gorbunova, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Cand.Sc., senior researcher (avgorbunova@list.ru).
Alexey Lebedev, Lomonosov Moscow State University, Moscow, Doctor of Science, professor (avlebed@yandex.ru).

Abstract: The article proposes a new approach to estimating the quantiles of the response time distribution of the fork-join queueing system. We consider a classic version of this system with a Poisson input flow and exponential service times on homogeneous servers. Upon receipt of tasks into the system, they are instantly divided into a fixed number of subtasks and sent for service to the appropriate subsystem with an unlimited capacity storage device and one server. The task is considered served after all its components have been serviced. This system allows you to simulate many real processes, which, in order to increase efficiency, are characterized by dividing large tasks into smaller components, for example, parallel or distributed computing systems. The difficulty of analyzing systems lies in the presence of a dependence between the sojourn times of subtasks, which significantly complicates the analysis of all performance characteristics of such systems. The main contribution of the article is the approach to determining the quantiles of the response time distribution, the assessment of which is no less valuable than the assessment of the mean response time. At the same time, a much larger number of works in this area are devoted to calculating the mean, which is explained, among other things, by the complexity of carrying out such an analysis even for a given characteristic, and estimating quantiles seems to be an even more laborious task.

Keywords: system with parallel service of tasks, fork-join queueing system, response time, distribution quantiles, simulation modeling.

УДК 519.2
ББК 22.17
DOI: 10.25728/ubs.2024.108.1

Статья представлена к публикации
членом редакционной коллегии Я.И. Квинто.

Поступила в редакцию 27.02.2024.
Дата опубликования 31.03.2024.