ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2025 Управление, вычислительная техника и информатика № 72
Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 519.872
doi: 10.17223/19988605/72/2

Нелинейная аппроксимация квантилей распределения времени отклика fork-join системы массового обслуживания с подсистемами M|M|1

Анастасия Владимировна Горбунова

Институт проблем управления им. В.А. Трапезникова Российской академии наук, Москва, Россия, avgorbunova@list.ru

Аннотация. Рассматривается система с разделением и параллельным обслуживанием заявок с пуассоновским входящим потоком и показательным распределением времени обслуживания на приборах. С помощью данной системы массового обслуживания моделируются различные типы физических структур, в которых происходит разделение исходной задачи на части для сокращения времени решения. В работе предлагается аналитическая оценка для квантилей распределения одного из важнейших показателей производительности подобных систем – времени отклика системы, т.е. времени пребывания заявки в системе. Вывод выражения основывается на элементах теории копул, их диагональных сечений, а также на имитационном моделировании системы с разделением и параллельным обслуживанием.

Ключевые слова: система с разделением и параллельным обслуживанием заявок; среднее время отклика; квантили распределения; копула; диагональное сечение; имитационное моделирование.

***Для цитирования*:** Горбунова А.В. Нелинейная аппроксимация квантилей распределения времени отклика fork-join системы массового обслуживания с подсистемами M|M|1 // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2025. № 72. С. 16–27. doi: 10.17223/19988605/72/2

Original article
doi: 10.17223/19988605/72/2

Nonlinear approximation of quantiles of the response time distribution of a fork-join queueing system with M|M|1 subsystems

Anastasia V. Gorbunova

V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russian Federation, avgorbunova@list.ru

Abstract. A fork-join queueing system with a Poisson input flow and exponential distribution of service time on servers is considered. Using this queuing system, various types of physical structures are modeled in which the original problem is divided into parts to reduce the solution time. The paper proposes an analytical estimate for the distribution quantiles of one of the most important performance indicators of such systems -- the system response time, i.e., the time a request stays in the system. The derivation of the expression is based on elements of the copula theory, their diagonal sections, and on simulation modeling of a fork-join queueing system.

Keywords: fork-join queueing system; average response time; distribution quantiles; copula; diagonal section; simulation modeling.

***For citation*:** Gorbunova, A.V. (2025) Nonlinear approximation of quantiles of the response time distribution of a fork-join queueing system with M|M|1 subsystems. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 72. pp. 16–27. doi: 10.17223/19988605/72/2

Введение

В статье исследуется система массового обслуживания (СМО) с разделением и параллельным обслуживанием заявок с пуассоновским входящим потоком и показательным распределением времени обслуживания. Система с разделением (с различными вариантами распределений для входящего потока и времен обслуживания) удобна для моделирования разнообразных процессов, в которых происходит разделение сложной комплексной задачи на составные части – подзадачи, которые обрабатываются в параллельном режиме с целью сокращения времени обработки исходной задачи.

Система с разделением и параллельным обслуживанием, или fork-join система массового обслуживания (в англоязычной версии термина), является предметом изучения многих исследователей, как отечественных, так и зарубежных, о чем свидетельствует множество публикаций [1–14]. Тем не менее точный результат известен лишь для такой характеристики, как среднее время пребывания заявки в системе только для случая системы с двумя подсистемами типа M|M|1 [4]. Для остальных вариантов архитектуры fork-join системы известны лишь приближения для величины среднего времени отклика. В меньшей степени изучались моменты более высоких порядков этой случайной величины, например дисперсия [4, 10, 11].

Также стоит отметить появление в последние годы работ, посвященных анализу квантилей распределения времени отклика, в частности [9, 10], что свидетельствует об актуальности оценки данного показателя системы, несмотря на все трудности проведения подобного анализа.

Сложность анализа систем с разделением и параллельным обслуживанием обусловлена наличием зависимости между временами пребывания частей от одной заявки в подсистемах. Зависимость между временами пребывания подзаявок в подсистемах fork-join системы с параллельным обслуживанием заявок возникает в силу общих для них моментов поступления в эти подсистемы. Флуктуации входного потока заявок в большую или меньшую сторону (по числу поступлений за какое-то время) приводят к увеличению или уменьшению длины очередей в подсистемах и, соответственно, увеличению или уменьшению времен пребывания подзаявок от одной заявки в подсистемах. Работ, посвященных изучению зависимостей времен пребывания подзаявок, совсем немного, однако для подсистем типа M|M|1 в [3] было получено точное выражения для коэффициента корреляции между временами пребывания в подсистмах, а в [4] для более сложной архитектуры системы – оценка этого показателя.

Основная задача исследования состоит в построении аналитической аппроксимации для квантилей распределения времени отклика fork-join системы как функции нескольких переменных. Причем полученная формула должна быть компактной и с минимальным числом параметров, которые необходимо было бы оценивать, а также обладать хорошей точностью приближения (в смысле максимального и среднего модуля относительного отклонения от данных имитационного моделирования квантилей по некоторой сетке параметров).

Данная статья является развитием работы [1], в которой оцениваются квантили распределения времени отклика в частном случае двух подсистем M|M|1. Здесь же методика для оценки квантилей распределения времени пребывания заявок в fork-join системе обобщается на большее количество подсистем K, $K \geq 2$, что, естественно, гораздо сложнее. Подход включает в себя элементы теории копул [2, 15, 16], визуальный анализ данных, методы оптимизации и имитационное моделирование. В отличие от работ [9, 10] применение копул позволяет выстроить логическую цепочку этапов предложенного подхода и обосновать выбор типа функциональной зависимости для аналитического выражения оценки квантилей времени отклика. Кроме того, методы, предложенные в статьях [9, 11] справедливы только для квантилей высокого уровня, в то время как подход, основанный на теории копул, позволяет определять квантили для значений вероятностей гораздо более широкого диапазона.

1. Математическая модель fork-join системы массового обслуживания

Рассматривается классическая система с разделением и параллельным обслуживанием, в которую поступает пуассоновский поток заявок с интенсивностью $\lambda > 0$. В момент поступления заявка

разделяется ровно на $K \geq 2$ подзаявок. Далее каждая из подзаявок поступает в одну из K подсистем, каждая из которых состоит из одного обслуживающего прибора и очереди с неограниченным числом мест для ожидания. Все приборы являются однородными, а время обслуживания имеет показательное распределение с параметром (интенсивностью) $\mu > 0$, $\lambda < \mu$.

После обслуживания подзаявки сразу не покидают систему, а ожидают окончания обслуживания последней подзаявки, составляющей исходную заявку, в условной точке сборки заявок (не занимая при этом прибор), находящейся за приборами системы (рис. 1). Затем происходит непосредственная сборка заявки из K обслуженных подзаявок, время которой считаем равным нулю, после чего заявка покидает систему.

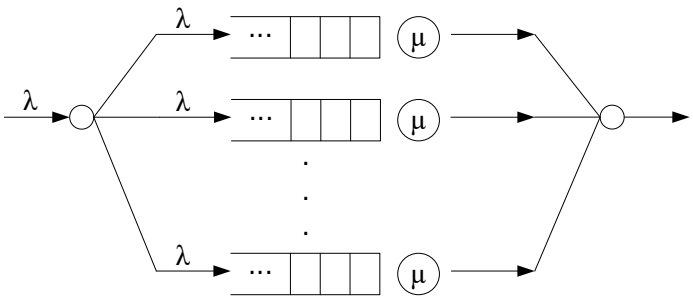

Рис. 1. Система с разделением и параллельным обслуживанием заявок
Fig. 1. Fork-join queueing system

Таким образом, время пребывания заявки в системе, или время отклика системы R_K, определяется максимальным временем обслуживания одной из изначально составляющих ее подзаявок, т.е.

$$R_K = \max\{\xi_1,...,\xi_K\}, \tag{1}$$

где ξ_i – время пребывания i-й подзаявки в i-й подсистеме, $i = 1, ..., K$.

Поскольку в дальнейшем для определения оценки квантилей распределения времени отклика системы с разделением и параллельным обслуживанием заявок наряду с элементами теории копул и методом оптимизации будет использоваться имитационное моделирование, то ограничимся числом подсистем $K = 2, ..., 20$. При этом методика вывода искомой оценки будет справедлива и для большего числа подсистем, но при этом будет требовать больше вычислительных затрат на организацию симуляции. Также для удобства положим $\lambda = 1$ (без потери в общности рассматриваемой системы) и будем менять уровень загрузки системы $\rho = \lambda / \mu$ за счет изменения значения интенсивности обслуживания μ.

2. Оценка квантилей распределения времени отклика

Рассмотрим случайную величину времени отклика системы с разделением и параллельным обслуживанием заявок R_K из (1), которая представляет собой максимум из K ($K \geq 2$), что важно отметить, зависимых случайных величин времен пребывания подзаявок в подсистемах исследуемой системы.

Функция распределения величины R_K имеет вид:

$$F_{R_K}(x) = P(\max(\xi_1,...,\xi_K) < x) = P(\xi_1 < x,...,\xi_K < x).$$

В этой связи обратимся к элементам теории копул [15]. Согласно определению копулой называется многомерная функция распределения, определенная на K-мерном единичном кубе, при этом каждое частное распределение является равномерным на отрезке [0, 1]. Согласно теореме Скляра любую функцию распределения можно представить с помощью копулы, т.е.

$$F(x_1,...,x_K) = C\left(F_1(x_1),...,F_K(x_K)\right),$$

где $F_i(x_i)$ – частные функции распределения, $i = 1, ..., K$. Если функции $F_i(x_i)$ являются непрерывными, то такое представление единственно. Диагональным сечением K-мерной копулы называется функция

$$\delta(y) = C(y,...y), \quad y \in [0,1],$$

где C – это копула-функция.

В случае рассматриваемой системы с разделением и параллельным обслуживанием маргинальные функции распределения времен пребывания подзаявок в подсистемах идентичны, соответственно, $F_i(x) = F(x)$, $i = 1, ..., K$. В результате можем представить выражение для функции распределения времени отклика системы с помощью диагонального сечения копулы, а именно

$$F_{R_K}(x) = P(\xi_1 < x, ..., \xi_K < x) = C(F(x), ..., F(x)) = \delta(F(x)) = \delta(y). \quad (2)$$

Далее, с одной стороны, $F_{R_K}(x_p) = p$, поэтому квантили распределения времени отклика R_K уровня p определяются как

$$x_p = F_{R_K}^{-1}(p),$$

и, соответственно, с другой стороны, $\delta(F(x_p)) = p$, откуда квантили уровня p также равны

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

Причем известно, что функция распределения времени пребывания подзаявки в подсистеме типа $M_\lambda|M_\mu|1$ имеет показательное распределение с параметром $(\mu - \lambda)$, т.е.

$$F(x) = 1 - e^{-(\mu-\lambda)x}, \quad F^{-1}(x) = -\frac{\ln(1-x)}{\mu - \lambda},$$

соответственно,

$$x_p = -\frac{\ln(1 - \delta^{-1}(p))}{\mu - \lambda}. \quad (3)$$

Далее необходимо определить выражение для диагонального сечения.

Согласно определению для диагонального сечения можем записать следующее:

$$\delta(y) = C(y, ..., y) = P(U_1 < y, ..., U_K < y) = P(\max(U_1, ..., U_K) < y),$$

где случайные величины U_i имеют равномерное распределение на отрезке [0, 1].

В соответствии с преобразованием Смирнова для генерации случайных величин с заданной функцией распределения (строго возрастающей, как в нашем случае) можем записать, что

$$\xi_i = F^{-1}(U_i),$$

откуда

$$U_i = F(\xi_i) = 1 - e^{-(\mu-\lambda)\xi_i}.$$

Поэтому, опять же в силу строгого возрастания функции $F(x)$, справедливо

$$P(\max(U_1, ..., U_K) < y) = P(\max(1 - e^{-(\mu-\lambda)\xi_1}, ..., 1 - e^{-(\mu-\lambda)\xi_K}) < y) =$$
$$= P(1 - e^{-(\mu-\lambda)\max(\xi_1, ..., \xi_K)} < y) = P(1 - e^{-(\mu-\lambda)R_K} < y).$$

Введем случайную величину $Y_K = 1 - e^{-(\mu-\lambda)R_K}$, следовательно, получаем для диагонального сечения

$$\delta(y_p) = P(Y_K < y_p) = p. \quad (4)$$

Для того чтобы оценить выражение для диагонального сечения $\delta(y)$, воспользуемся данными имитационного моделирования для случайных величин ξ_i, $i = 1, ..., K$ – времен пребывания в подсистемах системы с разделением и параллельным обслуживанием заявок.

С помощью программной среды Python для системы с разделением и параллельным обслуживанием заявок было смоделировано порядка 10 млн наборов из K случайных величин $(\xi_1, ..., \xi_K)$ для значений загрузки системы $\rho = \lambda/\mu = \{0{,}50, 0{,}55, ..., 0{,}90\}$ и 5 млн наборов величин $(\xi_1, ..., \xi_K)$ для более низких значений загрузки системы $\rho = \{0{,}10, 0{,}15, \ldots, 0{,}45\}$ для различного числа подсистем $K = 2, ..., 20$.

Далее вычисляются соответствующие каждому из смоделированных наборов случайных времен пребывания в подсистемах значения случайных величин $Y_K = 1 - e^{-(\mu-\lambda)\max(\xi_1, ..., \xi_K)}$. Затем статистически оценивается диагональное сечение с помощью полученных посредством симуляции от 5 до 10 млн пар (y_p, p), т.е. фактически вычисляются квантили случайной величины Y_K для соответствующих им вероятностей $p = \{0{,}20, 0{,}25, ..., 0{,}90\}$, где $\hat{p} \approx \delta(y_p)$ согласно формуле (4).

Теперь необходимо определить функциональную зависимость между величинами квантилей y_p и соответствующими им вероятностями, или уровнями p. Для этого проведем визуальный анализ данных.

На рис. 2 для случаев $K = 3$ и $K = 20$ представлены графики зависимости между $\ln y_p$ и $\ln p$ для различных уровней загрузки системы, построенные по данным имитационного моделирования. Аналогичная картина наблюдается для всех значений числа подсистем K в рамках заданного диапазона от 2 до 20, т.е. имеется линейная (или очень близкая к линейной) зависимость между логарифмами, что свидетельствует в пользу степенного вида функции для диагонального сечения, а именно

$$\ln p = f(\rho, K) \cdot \ln y_p,$$

где $f(\rho, K)$ выступает в роли углового коэффициента для соответствующих значений загрузки и числа подсистем. Таким образом, можем записать

$$\delta(y_p) = p = y_p^{f(\rho,K)}. \tag{5}$$

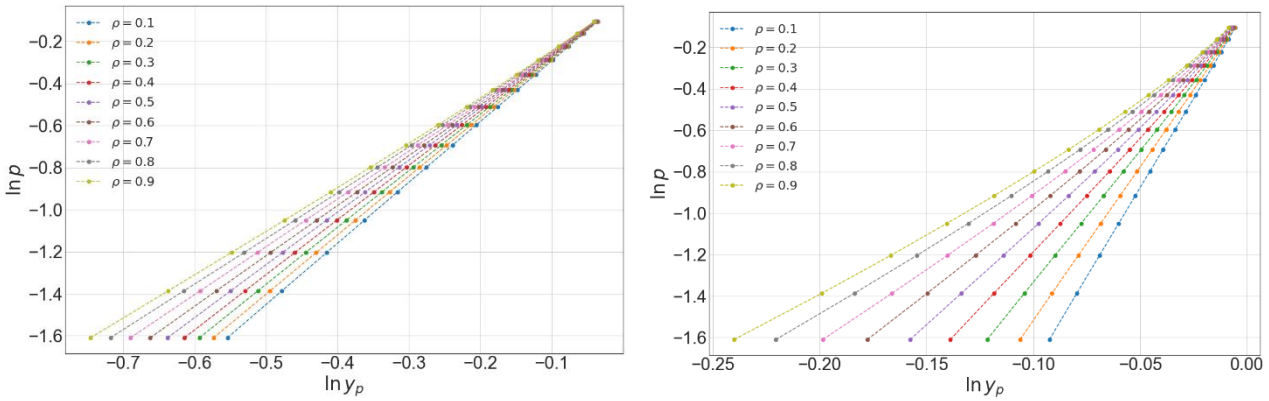

Рис. 2. Зависимость $\ln p$ от $\ln y_p$ для числа подсистем: $K = 3$ (*a*); $K = 20$ (*b*)
Fig. 2. Dependence of $\ln p$ on $\ln y_p$ for the number of subsystems: *a*) $K = 3$; *b*) $K = 20$

Тогда в качестве оценок квантилей получаем

$$\hat{x}_p = -\frac{\ln\left(1 - p^{\frac{1}{f(\rho,K)}}\right)}{\mu - \lambda}, \tag{6}$$

и, подбирая разные функции f, можем получать оценки разной точности.

Чтобы конкретизировать вид функции $f(\rho, K)$, проанализируем отношение $\ln p / \ln y_p$. На рис. 3 представлена зависимость данного отношения от значения загрузки ρ при $K = 3$ и $K = 20$.

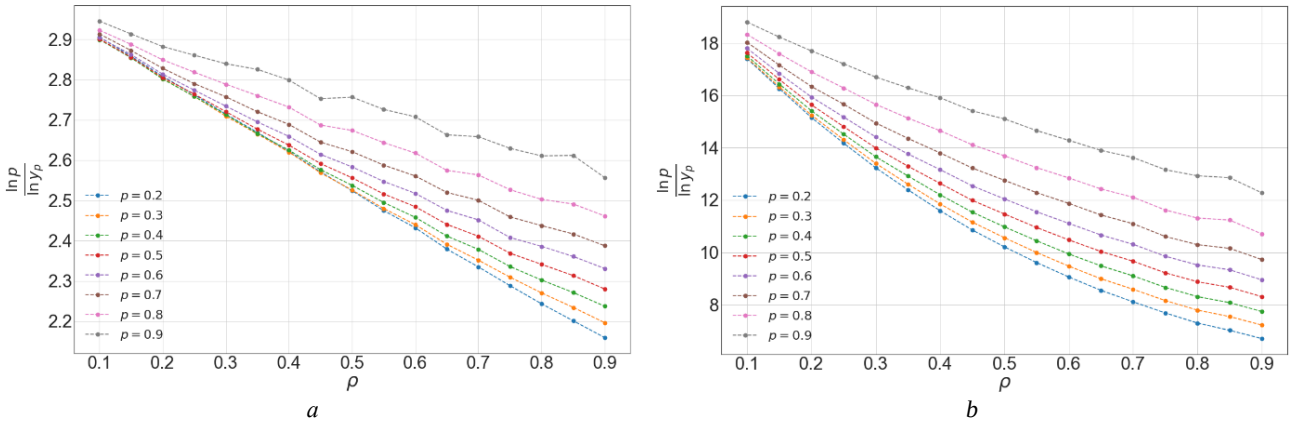

Рис. 3. Зависимость отношения $\ln p / \ln y_p$ от ρ для числа подсистем: $K = 3$ (*a*); $K = 20$ (*b*)
Fig. 3. Dependence of the ratio $\ln p / \ln y_p$ on ρ for the number of subsystems: *a*) $K = 3$; *b*) $K = 20$

Стоит отметить, что при $\rho \to 0$ фактически справедлива независимость времен пребывания в подсистемах, т.е. при независимости случайных величин ξ_i, $i = 1, ..., K$, выражение (2) преобразуется следующим образом:

$$F_{R_K}(x) = \delta(F(x)) = P(\xi_1 < x, ..., \xi_K < x) = (F(x))^K, \quad (7)$$

что соответствует $f(\rho, K) \to K$ и, соответственно, при подстановке $f(\rho, K) = K$ в (5) полностью согласуется с (7):

$$\delta(F(x)) = (F(x))^K.$$

В работе [1] для случая $K = 2$ использовалась функция $f(\rho, 2) = 2 - C\rho$. Это наводит на мысль попробовать приближения вида $K - C\rho$ и $K(1 - C\rho)$, однако их проверка с помощью (6) приводит к плохим результатам.

Более тонкий подход заключается в обращении к приближению копулой Гумбеля, которая показала хорошее согласие с данными в работе [2] в случае $K = 2$, обладает степенным диагональным сечением и является абсолютно непрерывной (в отличие, например, от копулы Маршалла–Олкина), поэтому используется для моделирования абсолютно непрерывных многомерных распределений. В общем случае K-мерная копула Гумбеля имеет вид:

$$C(y_1, \dots y_K) = \exp\left\{-\left((-\ln y_1)^\theta + \dots + (-\ln y_K)^\theta\right)\right)^{1/\theta}\Big\}, \quad \theta > 1,$$

откуда $\delta(y) = y^{K^{1/\theta}}$. Таким образом, при $K = 2$ в [2] получалось $\theta = (\ln 2)/\ln(2 - C\rho)$, откуда в общем случае $f(\rho, K) = (2 - C\rho)^{(\ln K)/(\ln 2)}$. Если же использовать более простое приближение $\theta = 1/(1 - C\rho)$, имеющее сходный график, получаем $f(\rho, K) = K^{1-C\rho}$. Проверка с помощью (6) показывает, что последний вариант дает наилучшее соответствие.

Таким образом, получаем следующее выражение для приближения диагонального сечения:

$$\delta(y_p) \approx y_p^{K^{1-C\rho}}, \quad (8)$$

откуда следует, что

$$\delta^{-1}(p) \approx p^{\frac{1}{K^{1-C\rho}}}. \quad (9)$$

В результате после подстановки (9) в формулу (3) для квантилей распределения времени отклика системы с разделением на K подзаявок и их параллельным обслуживанием получаем следующую аналитическую оценку:

$$\hat{x}_p = -\frac{\ln(1 - p^{\frac{1}{K^{1-C\rho}}})}{\mu - \lambda}. \quad (10)$$

Чтобы определить константу C, воспользуемся методом оптимизации Нелдера–Мида [17]. Будем минимизировать по имеющимся данным симуляции для времени отклика модуль абсолютного значения максимальной погрешности приближения формулы (10), т.е.

$$\max_C \left|\frac{x_p - \hat{x}_p}{x_p}\right| \to \min.$$

В результате получаем значение $C \approx 0{,}3490997$.

В табл. 1 представлены значения максимальной (MaxAPE, %), минимальной (MinAPE, %) и средней (MAPE, %) относительных погрешностей приближения значений квантилей распределения времени отклика формулой (10), рассчитанные на наборе данных из $N = 4\,845$ элементов, где

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_{p,i} - \hat{x}_{p,i}}{x_{p,i}}\right| \cdot 100\%, \quad \text{MaxAPE} = \max_{1 \le i \le N}\left|\frac{x_{p,i} - \hat{x}_{p,i}}{x_{p,i}}\right| \cdot 100\%, \quad \text{MinAPE} = \min_{1 \le i \le N}\left|\frac{x_{p,i} - \hat{x}_{p,i}}{x_{p,i}}\right| \cdot 100\%,$$

а $x_{p,i}$ и $\hat{x}_{p,i}$ – i-е значения квантилей, полученные, соответственно, с помощью имитационного моделирования и с помощью формулы (аналитической оценки) в наборе из N элементов.

Таблица 1

Погрешности приближений значений квантилей распределения времени отклика системы x_p, K = 2, ..., 20, рассчитанные с помощью формулы (10), в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	MaxAPE, %	MinAPE, %	MAPE, %
Квантиль времени отклика x_p	9,534643	0,000551	2,505877

Несмотря на довольно низкое значение средней погрешности приближения, его максимальное значение, которое хоть и укладывается в рамки инженерной погрешности, желательно было бы снизить. Попробуем учесть наблюдаемую на рис. 3 зависимость кривых от p аналогично тому, как это делалось в [1], заменяя константу C на выражение $C_1 - C_2p^2$. В результате уточненная оценка для квантилей времени отклика примет вид:

$$\hat{x}_p = -\frac{\ln(1-p^{\frac{1}{K^{1-(C_1-C_2p^2)\rho}}})}{\mu-\lambda}. \qquad (11)$$

Значения констант C_1 и C_2 в (11), как и ранее, определим с помощью метода Нелдера–Мида, соответственно, получим следующие $C_1 \approx 0,390797, \quad C_2 \approx 0,221811.$

Погрешности аппроксимации для оценки (11) представлены в табл. 2.

Таблица 2

Погрешности приближений значений квантилей распределения времени отклика системы x_p, K = 2, ..., 20, рассчитанные с помощью формулы (11), в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	MaxAPE, %	MinAPE, %	MAPE, %
Квантиль времени отклика x_p	5,498891	0,000306	1,116448

Как можно заметить, максимальное значение погрешности приближения значительно уменьшилось (примерно в 2 раза), равно как и средняя погрешность приближения. На рис. 4 наглядно демонстрируется качество приближения формулы (11).

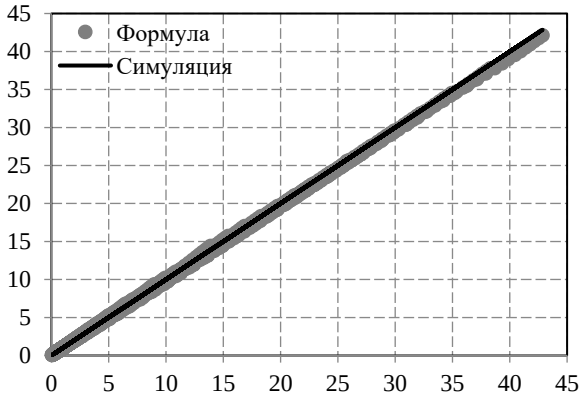

Рис. 4. Сравнение результатов имитационного моделирования квантилей распределения времени отклика R_K системы с разделением и параллельным обслуживанием для числа подсистем K = 2, ..., 20, вероятностей $p \in \{0,20; 0,25; ...; 0,90\}$ и загрузки $\rho \in \{0,10; 0,15; ...; 0,90\}$ с формулой (11)

Fig. 4. Comparison of the results of simulation modeling of the quantiles of the distribution of the response time R_K of a fork-join queueing system for the number of subsystems K = 2, ..., 20, probabilities $p \in \{0,20; 0,25; ...; 0,90\}$ and load $\rho \in \{0,10; 0,15; ...; 0,90\}$ with formula (11)

Заметим, что нашей задачей было получить равномерную (по относительной точности) оценку квантилей по всем K от 2 до 20. При конкретных K аналогичным образом можно получить гораздо более точные оценки, как это было сделано ранее для K = 2.

3. Сравнение с другими методами

В данном разделе проведем сравнение результата, полученного с помощью комплексного подхода, опирающегося на интеллектуальный анализ данных, т.е. посредством применения комбинации нескольких методов анализа: элементов теории копул, визуального анализа данных и оптимизации, – с результатом аппроксимации, который будет получен без использования теории копул. А именно будем аппроксимировать исходную функцию, заданную, как и ранее, таблицей значений ее аргументов – уровня вероятности p, числа подсистем K, загрузки системы ρ – и соответствующих им значений целевой функции, т.е. квантилей, с помощью ее представления в виде полинома.

Как известно, любую непрерывную функцию можно аппроксимировать с любой степенью точности, при этом известны различные подходы к аппроксимации непрерывных функций многих переменных, однако нельзя сказать, что все они просты в своей реализации; в частности, речь может идти о серьезных вычислительных затратах на их организацию [18–21].

Одним из наиболее распространенных и известных методов является аппроксимация с помощью многочленов, однако при этом точность аппроксимации повышается с увеличением степени многочлена. В данной работе остановимся на результатах аппроксимации полиномами первой и второй степеней, поскольку увеличение степени полинома ведет к увеличению количества коэффициентов в одночленах, составляющих полином, и априори будет значительно проигрывать выражению (11), в котором их всего два (C_1 и C_2)

Итак, поскольку для $K = 1$ квантиль уровня p определяется выражением

$$x_{p,1} = -\frac{\ln(1-p)}{\mu - \lambda}, \tag{12}$$

то возьмем данное значение за базовое и рассмотрим следующий полином первой степени для приближения квантилей уровня p для числа подсистем K:

$$x_{p,K} \approx \hat{x}_{p,K} = x_{p,1}(C_0 + C_1 p + C_2 \rho + C_3 K). \tag{13}$$

Далее, как и ранее, будем оптимизировать полученное выражение, минимизируя модуль максимального абсолютного значения погрешности приближения методом Нелдера–Мида:

$$\max_C \left| \frac{x_{p,K} - \hat{x}_{p,K}}{x_{p,K}} \right| \to \min,$$

что дает следующие значения коэффициентов в модели (13)

$$C_0 \approx 3{,}890744, \quad C_1 \approx -3{,}467565, \quad C_2 \approx -0{,}420018, \quad C_3 \approx 0{,}129885.$$

Погрешности аппроксимации для оценки (13) представлены в табл. 3. На рис. 5 наглядно демонстрируется качество полученного приближения.

Таблица 3

Погрешности приближений значений квантилей распределения времени отклика системы $x_{p,K}$, K = 2, ..., 20, рассчитанные с помощью формулы (13), в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	MaxAPE, %	MinAPE, %	MAPE, %
Квантиль времени отклика x_p	47,625180	0,008507	14,401126

Теперь рассмотрим приближение искомой функции для квантилей полиномом второй степени

$$x_{p,K} \approx \hat{x}_{p,K} = x_{p,1}(C_0 + C_1 p + C_2 \rho + C_3 K + C_4 p\rho + C_5 \rho K + C_6 pK + C_7 p^2 + C_8 \rho^2 + C_9 K^2). \tag{14}$$

Далее, действуя аналогичным образом, т.е. применяя метод оптимизации, получаем следующие значения для искомых коэффициентов:

$$C_0 \approx 2{,}992137, \quad C_1 \approx -3{,}872610, \quad C_2 \approx -2{,}262913, \quad C_3 \approx 0.576742, \quad C_4 \approx 2{,}434478,$$
$$C_5 \approx -0{,}333510, \quad C_6 \approx -0{,}026966, \quad C_7 \approx 1{,}240389, \quad C_8 \approx 0.195363, \quad C_9 \approx -0{,}009761.$$

Погрешности аппроксимации для оценки (14) представлены в табл. 4, а на рис. 6 можно наглядно проследить за качеством полученного приближения.

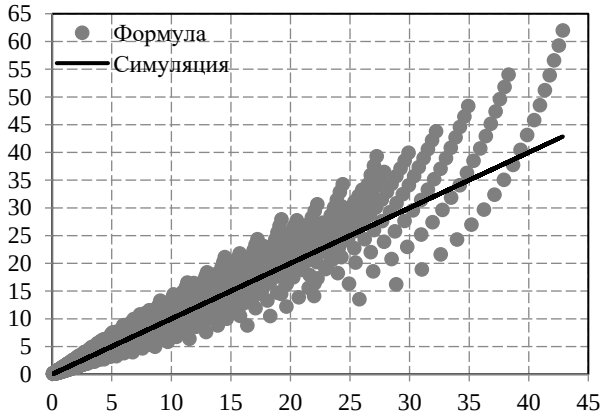

Рис. 5. Сравнение результатов имитационного моделирования квантилей распределения времени отклика R_K системы с разделением и параллельным обслуживанием для числа подсистем K = 2, ..., 20, вероятностей $p \in \{0{,}20; 0{,}25; ...; 0{,}90\}$ и загрузки $\rho \in \{0{,}10; 0{,}15; ...; 0{,}90\}$ с формулой (13)

Fig. 5. Comparison of the results of simulation modeling of the quantiles of the distribution of the response time R_K of a fork-join queueing system for the number of subsystems K = 2, ..., 20, probabilities {0,20; 0,25; ...; 0,90} and load $\rho \in \{0{,}10; 0{,}15; ...; 0{,}90\}$ with formula (13)

Таблица 4

Погрешности приближений значений квантилей распределения времени отклика системы $x_{p,K}$, K = 2, ..., 20, рассчитанные с помощью формулы (14), в сравнении с результатами имитационного моделирования

Оцениваемая характеристика	Типы ошибок		
	MaxAPE, %	MinAPE, %	MAPE, %
Квантиль времени отклика x_p	24,362991	0,005232	12,480062

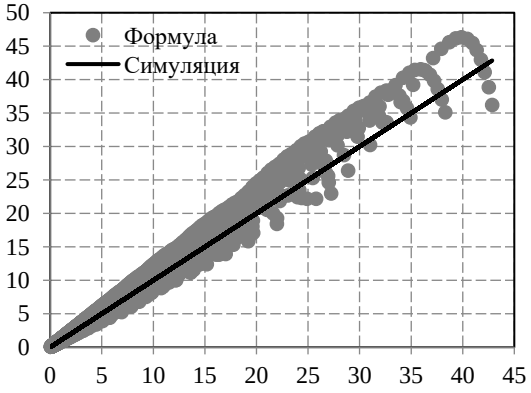

Рис. 6. Сравнение результатов имитационного моделирования квантилей распределения времени отклика R_K системы с разделением и параллельным обслуживанием для числа подсистем K = 2, ..., 20, вероятностей $p \in \{0{,}20; 0{,}25; ...; 0{,}90\}$ и загрузки $\rho \in \{0{,}10; 0{,}15; ...; 0{,}90\}$ с формулой (14)

Fig. 6. Comparison of the results of simulation modeling of the quantiles of the distribution of the response time R_K of a fork-join queueing system for the number of subsystems K = 2, ..., 20, probabilities $p \in \{0{,}20; 0{,}25; ...; 0{,}90\}$ and load $\rho \in \{0{,}10; 0{,}15; ...; 0{,}90\}$ with formula (14)

Таким образом, сравнивая результаты приближения формул (10), (11) и формул (13), (14), можно заключить, что приближение, полученное с использованием теории копул, имеет лучшую степень точности аппроксимации по всем трем типам характеристик (максимальная, минимальная и средняя погрешности приближения).

Разумеется, если увеличивать степень полинома, то можно добиться лучшего качества приближения и во втором случае, но при этом будет расти и число одночленов, причем для полинома второй степени их количество уже составляет десять единиц.

Кроме того, вариант построения аналитического приближения для квантилей времени отклика напрямую, благодаря использованию только результатов симуляции и оптимизации коэффициентов при одночленах, является менее содержательным, поскольку совершенно не учитывает природу моделируемых случайных величин и их зависимость в частности, а опирается лишь на теорию об аппроксимации функций нескольких переменных. Поэтому полученное таким образом приближенное аналитическое выражение представляется малопригодным для использования при дальнейшем анализе системы (с разделением и параллельным обслуживанием), например при оптимизации ее параметров.

Использование же теории копул, наоборот, имеет под собой теоретическую основу, что позволяет использовать результаты подхода не только для определения квантилей времени отклика для промежуточных значений на заданных числовых интервалах параметров системы, но и при дальнейшем исследовании системы, давая тем самым возможность продвинуться (хоть и не в полной мере) в понимании сути явления.

Заключение

В статье описан подход к оценке квантилей распределения времени отклика системы с разделением и параллельным обслуживанием заявок. Основывается подход на применении копул, их диагональных сечений, а также визуальном анализе данных, оптимизации и имитационном моделировании. Исследование системы с разделением считается трудной задачей, поэтому большинство исследований посвящено аппроксимации величины среднего времени отклика. Здесь же предлагается подход к оценке более тонкой характеристики данной случайной величины – ее квантилей, что на практике представляет больший интерес, поскольку характеризует величину времени пребывания заявки в системе, которая не будет превышена с заданной вероятностью.

Проведен сравнительный анализ результатов применения предложенного подхода с результатами аппроксимации квантилей времени отклика как функции нескольких переменных напрямую с помощью многочлена без применения теории копул, из которого следует, что аналитическое выражение, полученное с помощью нового метода, обладает лучшей точностью и является более компактным, требует оценки меньшего числа коэффициентов.

Использование элементов теории копул, которые по своему определению содержат в себе информацию о зависимости между случайными величинами, представляет собой некоторую теоретическую опору, на базе которой возможно построить приближения искомых величин, хотя и делая при этом некоторые эвристические заключения. Известны различные виды копул и, соответственно, их диагональных сечений, поэтому выбор в пользу степенного вида функциональной зависимости осуществляется на основе визуального анализа данных (построения графиков функций), что является одной из составных частей интеллектуального анализа данных, а также исходя из того, что в двумерном случае в более ранней работе [1] для данного типа копула-функции было получено хорошее числовое соответствие.

Немногочисленные подходы к оценке квантилей, предлагаемые на текущий момент, позволяют определять их значения только для высоких уровней вероятностей. Метод, предложенный в статье, значительно расширяет диапазон уровней вероятностей и, несмотря на все множество составляющих его этапов, является менее сложным в реализации, поскольку известные подходы либо требуют длительной вычислительной процедуры, либо уступают в качестве получаемых оценок, требуя при этом также проведения имитационного моделирования на промежуточных этапах построения оценок. Кроме того, представленный в статье подход не накладывает ограничений на структуру системы с разделением и параллельным обслуживанием и может быть применен для систем с другими вариантами распределений для входящего потока и времени обслуживания.

Список источников

1. Горбунова А.В., Лебедев А.В. О новом подходе к оценке квантилей времени отклика системы с разделением и параллельным обслуживанием заявок // Управление большими системами: сборник трудов (электронный журнал). 2024. № 108. С. 6–21. doi: 10.25728/ubs.2024.108.1

2. Gorbunova A.V., Lebedev A.V. Copulas and quantiles in Fork-Join queueing systems // Advances in Systems Science and Applications. V. 24 (1). P. 1–19. doi: 10.25728/assa.2024.24.1.1585
3. Gorbunova A.V., Lebedev A.V. Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with M|M|1-type Subsystems // Advances in Systems Science and Applications. 2024. V. 24 (2). P. 1–18. doi: 10.25728/assa.2024.2024.02.1641
4. 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. V. 214. P. 409–428. doi: 10.1016/j.matcom.2023.07.029
5. Nelson R., Tantawi A.N. Approximate analysis of fork/join synchronization in parallel queues // IEEE Transactions on Computers. 1988. V. 37 (6). P. 739–743. doi: 10.1109/12.2213
6. Varma S., Makowski A.M. Interpolation approximations for symmetric Fork-Join queues // Performance Evaluation. 1994. V. 20. P. 245–265. doi: 10.1016/0166-5316(94)90016-7
7. Kemper B., Mandjes M. Mean sojourn time in two-queue fork-join systems: Bounds and approximations // OR Spectrum. 2012. V. 34. P. 723–742. doi: 10.1007/s00291-010-0235-y
8. Thomasian A. Analysis of fork/join and related queueing systems // ACM Computing Surveys (CSUR). 2014. V. 47 (2). P. 1–71. doi: 10.1145/2628913
9. Qiu Zh., Perez J.F., Harrison P.G. Beyond the mean in fork-join queues: Efficient approximation for response-time tails // Performance Evaluation. 2015. V. 91. P. 99–116. doi: 10.1016/j.peva.2015.06.007
10. Nguyen M., Alesawi S., Li N., Che H., Jiang H. A black-box Fork-Join latency prediction model for data-intensive applications // IEEE Transactions on Parallel and Distributed Systems. 2020. V. 31 (9). P. 1983–2000. doi: 10.1109/TPDS.2020.2982137
11. 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. doi: 10.1109/NAS55553.2022.9925531
12. Sethuraman S. Analysis of Fork-Join Systems: Network of Queues with Precedence Constraints. Boca Raton : CRC Press, 2022. 104 p.
13. Жидкова Л.А., Моисеева С.П. Исследование системы параллельного обслуживания кратных заявок простейшего потока // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2011. Т. 17, № 4. С. 49–54.
14. Моисеева С.П., Панкратова Е.В., Убонова Е.Г. Исследование бесконечнолинейной системы массового обслуживания с разнотипным обслуживанием и входящим потоком марковского восстановления // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2016. Т. 35, № 2. С. 46–53. doi: 10.17223/19988605/35/5
15. Nelsen R. An introduction to copulas. Berlin : Springer, 2006. 269 p.
16. Лебедев А.В. Верхняя граница среднего минимума зависимых случайных величин с известным коэффициентом Кендалла // Теория вероятностей и ее применения. 2019. Т. 64, вып. 3. С. 578–589. doi: 10.4213/tvp5199
17. Nelder J.A., Mead R. A Simplex Method for Function Minimization // Computer Journal. 1965. V. 7 P. 308–313.
18. Колмогоров А.Н. О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного и сложения // Доклады АН СССР. 1957. Т. 114, № 5. С. 953–956.
19. Арнольд В.И. О представлении функций нескольких переменных в виде суперпозиции функций меньшего числа переменных // Математическое просвещение. 1958. № 3. С. 41–61.
20. Горбань А.Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной техники. 1998. Т. 1, № 1. С. 11–24.
21. Бутырский Е.Ю., Кувалдин И.А., Чалкин В.П. Аппроксимация многомерных функций // Научное приборостроение. 2010. Т. 20, № 2. С. 82–92.

References

1. Gorbunova, A.V. & Lebedev, A.V. (2024a) On a new approach to estimating response time quantiles of a Fork-Join queueing system. *Upravlenie bol'shimi sistemami – Large-Scale Systems Control*. 108. pp. 6–21. DOI: 10.25728/ubs.2024.108.1
2. Gorbunova, A.V. & Lebedev, A.V. (2024b) Copulas and quantiles in Fork-Join queueing systems. *Advances in Systems Science and Applications*. 24(1). pp. 1–19. DOI: 10.25728/assa.2024.24.1.1585
3. Gorbunova, A.V. & Lebedev, A.V. (2024c) Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with M|M|1-type Subsystems. *Advances in Systems Science and Applications*. 24(2). pp. 1–18. DOI: 10.25728/assa.2024.2024.02.1641
4. Gorbunova, A.V. & Lebedev, A.V. (2023) 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*. 214. pp. 409–428. DOI: 10.1016/j.matcom.2023.07.029
5. Nelson, R. & Tantawi, A.N. (1988) Approximate analysis of fork/join synchronization in parallel queues. *IEEE Transactions on Computers*. 37(6). pp. 739–743. DOI: 10.1109/12.2213
6. Varma, S. & Makowski, A.M. (1994) Interpolation Approximations for Symmetric Fork-Join Queues. *Performance Evaluation*. 20. pp. 245–265. DOI: 10.1016/0166-5316(94)90016-7
7. Kemper, B. & Mandjes, M. (2012) Mean sojourn time in two-queue fork-join systems: Bounds and approximations. *OR Spectrum*. 34. pp. 723–742. DOI: 10.1007/s00291-010-0235-y

8. Thomasian, A. (2014) Analysis of fork/join and related queueing systems. *ACM Computing Surveys (CSUR)*. 47(2). pp. 1–71. DOI: 10.1145/2628913
9. Qiu, Zh., Perez, J.F. & Harrison, P.G. (2015) Beyond the mean in fork-join queues: Efficient approximation for response-time tails. *Performance Evaluation*. 91. pp. 99–116. DOI: 10.1016/j.peva.2015.06.007
10. Nguyen, M., Alesawi, S., Li, N., Che, H. & Jiang, H. (2020) A black-box Fork-Join latency prediction model for data-intensive applications. *IEEE Transactions on Parallel and Distributed Systems*. 31(9). p. 1983–2000. DOI: 10.1109/TPDS.2020.2982137
11. Enganti, P., Rosenkrantz, T., Sun, L., Wang, Z., Che, H. & Jiang, H. (2022) ForkMV: Mean-and-variance estimation of Fork-Join queuing networks for datacenter applications. *Proc. IEEE International Conference on Networking, Architecture and Storage (NAS)*. pp. 1–8. DOI: 10.1109/NAS55553.2022.9925531
12. Sethuraman, S. (2022) *Analysis of Fork-Join systems: Network of queues with precedence constraints*. Boca Raton: CRC Press.
13. Zhidkova, L.A. & Moiseeva, S.P. (2011) Research of the parallel service system of multiple requests of the simplest flow. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 17(4). pp. 49–54.
14. Moiseeva, S.P., Pankratova, E.V. & Ubonova, E.G. (2016) Research of infinite-line queueing system with heterogeneous service and input Markov renewal flow. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaya tekhnika i informatika – Tomsk State University Journal of Control and Computer Science*. 35(2). pp. 46–53. DOI: 10.17223/19988605/35/5
15. Nelsen, R. (2006) *An Introduction to Copulas*. Berlin, Germany: Springer.
16. Lebedev, A.V. (2019) Upper Bound for the Expected Minimum of Dependent Random Variables with Known Kendall's Tau. *Teoriya veroyatnostey i ee primeneniya – Theory of Probability and its Applications*. 64(3). pp. 465–473. DOI: 10.1137/S0040585X97T989623
17. Nelder, J.A. & Mead, R.A (1965) Simplex Method for Function Minimization. *Computer Journal*. 7. pp. 308–313.
18. Kolmogorov, A.N. (1957) O predstavlenii nepreryvnykh funktsiy neskol'kikh peremennykh v vide superpozitsii nepreryvnykh funktsiy odnogo peremennogo i slozheniya [On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition]. *Doklady AN SSSR.* 114(5). pp. 953–956.
19. Arnold, V.I. (1958) O predstavlenii funktsiy neskol'kikh peremennykh v vide superpozitsii funktsiy men'shego chisla peremennykh [On the representation of functions of several variables by superpositions of functions of fewer variables]. *Matematicheskoe prosveshchenie*. 3. pp. 41–61.
20. Gorban, A.N. (1998) Obobshchennaya approksimatsionnaya teorema i vychislitel'nye vozmozhnosti neyronnykh setey [Generalized approximation theorem and computational capabilities of neural networks]. *Sibirskiy zhurnal vychislitel'noy tekhniki*. 1(1). pp. 11–24.
21. Butyrskiy, E.Yu., Kuvaldin, I.A. & Chalkin, V.P. (2010) Approksimatsiya mnogomernykh funktsiy [Multidimensional functions' approximation]. *Nauchnoe priborostroenie*. 20(2). pp. 82–92.

***Информация об авторе*:**
Горбунова Анастасия Владимировна – кандидат физико-математических наук, старший научный сотрудник Института проблем управления им. В.А. Трапезникова Российской академии наук (Москва, Россия). E-mail: avgorbunova@list.ru

Автор заявляет об отсутствии конфликта интересов.

***Information about the authorss*:**
Gorbunova Anastasia V. (Candidate of Physical and Mathematical Sciences, Senior Researcher, V.A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russian Federation). E-mail: avgorbunova@list.ru

The author declares no conflicts of interests.

Поступила в редакцию 24.03.2025; принята к публикации 02.09.2025

Received 24.03.2025; accepted for publication 02.09.2025