ОЦЕНКИ КОПУЛЫ И КВАНТИЛЕЙ РАСПРЕДЕЛЕНИЯ ВРЕМЕНИ ОТКЛИКА СИСТЕМЫ С РАЗДЕЛЕНИЕМ И ПАРАЛЛЕЛЬНЫМ ОБСЛУЖИВАНИЕМ ЗАЯВОК И РАСПРЕДЕЛЕНИЕМ ПАРЕТО ВРЕМЕНИ ОБСЛУЖИВАНИЯ

Горбунова А. В.[1]

(*ФГБУН Институт проблем управления им. В.А. Трапезникова РАН, Москва*)

Рассматривается система с разделением и параллельным обслуживанием заявок. Предполагается, что распределение времени обслуживания на всех приборах имеет распределение Парето. Изучается зависимость между временами пребывания подзаявок в подсистемах, являющаяся основной причиной сложности анализа подобных систем. Время пребывания заявки в системе (или среднее время отклика) является максимумом из зависимых случайных величин пребывания подзаявок в системе. Получены приближения совместного распределения времен пребывания подзаявок с помощью теории копул. Также предложен подход для определения квантилей распределения времени отклика системы с помощью диагонального сечения копул. Данный подход ранее применялся для случая анализа аналогичной системы, но с экспоненциальным распределением времени обслуживания. Однако основное отличие исследуемой системы от экспоненциального случая заключается в том, что вид функции распределения времени пребывания подзаявки в подсистеме неизвестен. Поэтому используется аналитическое приближение для квантилей распределения времени отклика в подсистеме в предположении полученной ранее аппроксимации распределения времени пребывания подзаявки в подсистеме распределением Фреше. Оценки, полученные для квантилей и копулы распределения времени отклика, показывают хорошее соответствие с данными имитационного моделирования.

Ключевые слова: система с параллельным обслуживанием заявок, система массового обслуживания, время отклика, квантили распределения, копула.

1. Введение

В статье исследуется частный случай fork-join-системы, или, что то же самое, системы с разделением и параллельным обслуживанием заявок. Рассматривается система, состоящая из двух

[1] *Анастасия Владимировна Горбунова, к.ф.-м.н., с.н.с.* (*avgorbunova@list.ru*).

подсистем. Предполагается, что входящий поток является пуассоновским, а время обслуживания на каждом из приборов имеет распределение Парето. Более детальное описание системы будет представлено в следующем разделе.

Fork-join-системы массового обслуживания (СМО) используются для моделирования процессов функционирования различных структур, предполагающих разделение исходной задачи на более мелкие составляющие с последующей их параллельной обработкой. Распараллеливание решаемой задачи является одним из способов повышения эффективности производительности физических систем и применяется в различных сферах: в области информационных технологий (параллельные или распределенные вычисления), в сфере материального производства (изготовление продукции, состоящей из множества деталей), логистике (сборка и доставка в пункт назначения многокомпонентного заказа) и т.п. [2, 3, 5, 16, 18].

Анализ подобных систем является довольно сложным, точные характеристики (среднее время отклика системы) получены только для случая двух подсистем с экспоненциальным распределением времени обслуживания [13]. Для систем с неэкспоненциальными временами обслуживания уже даже в случае двух подсистем известны только аппроксимации различной точности в определенных диапазонах параметров [14, 15, 19]. С оценками для моментов времени пребывания заявки в системе с распределением Парето времени обслуживания можно ознакомиться, например, в [6].

В настоящей работе исследуется fork-join-система с точки зрения оценки не менее важных характеристик случайной величины времени отклика, чем его моменты. А именно, предлагается подход к оценке квантилей распределения времени отклика системы, основанный на элементах теории копул. Более подробно с элементами теории можно ознакомиться, например, в [8, 9, 10, 12]. Копулы, наряду с коэффициентами корреляции, в полной мере позволяют охарактеризовать зависимость между случайными величинами. В данном случае речь идет о наличии

зависимости между временами пребывания подзаявок, на которые разделяется заявка, в соответствующих подсистемах. Среди работ, посвященных оценке квантилей распределения времени пребывания заявки в fork-join СМО с различными распределениями можно отметить [1, 7, 14, 15, 17]. Здесь предлагается подход, успешно примененный к оценке квантилей в экспоненциальном случае. Однако в случае распределения Парето закон (функция) распределения времени пребывания подзаявки в подсистеме неизвестен, поэтому за основу взято приближение для квантилей времени отклика в базовом случае из [1]. Оценки, полученные для квантилей и копулы распределения времени отклика, показывают хорошее соответствие с эмпирическими данными.

Статья организована следующим образом: во втором разделе подробно описана математическая модель исследуемой fork-join-системы, в третьем разделе описывается подход к построению оценок квантилей времени отклика с помощью диагонального сечения копулы, в следующем разделе приводится описание метода оценки самой копулы, в заключении подводятся некотрые итоги.

2. Математическая модель системы с разделением и параллельным обслуживанием

Рассматривается система с разделением и параллельным обслуживанием заявок, при поступлении в которую заявка мгновенно разделяется на $K = 2$ подзаявки. После разделения каждая из подзаявок в свою очередь поступает на обслуживание с единственным прибором (рис. 1). Входящий поток заявок является пуассоновским с интенсивностью $\lambda > 0$. Время обслуживания имеет распределение Парето с функцией распределения следующего вида:

$$B(x) = 1 - \left(\frac{\alpha - 1}{\alpha} \cdot \frac{1}{x}\right)^{\alpha}, \quad x \geqslant \frac{\alpha - 1}{\alpha}, \quad \alpha > 3. \tag{1}$$

Тогда среднее время обслуживания b составляет одну условную временную единицу, а второй момент времени обслуживания $b^{(2)}$ определяется соотношением, зависящим от параметра распределе-

ния α, т.е.

$$b_{Pa} = 1, \quad b_{Pa}^{(2)} = \frac{(\alpha-1)^2}{\alpha(\alpha-2)}.$$

Фактически каждая из подзаявок поступает на обслуживание в подсистему типа $M|Pa|1$. Коэффициент загрузки системы определяется как $\rho = \lambda b_{Pa} = \lambda < 1$.

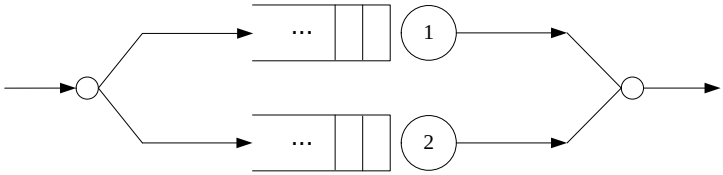


Рис. 1. Модель fork-join СМО с двумя подсистемами типа $M|Pa|1$

Целая заявка может покинуть систему только после окончания обслуживания обеих составляющих ее подзаявок. При этом считается, что сборка заявки (из обслуженных подзаявок) происходит мгновенно. Поэтому если обозначить времена пребывания каждой из двух подзаявок в соответствующих подсистемах через случайные величины ξ_1 и ξ_2, то время пребывания заявки будет являться максимумом из этих случайных величин:

$$R_2 = \max\{\xi_1, \xi_2\}.$$

Стоит отметить, что случайные величины ξ_1 и ξ_2 коррелируют между собой, причем их зависимость является положительной. Наличие корреляции заметно усложняет анализ fork-join-систем. По этой причине представляет интерес исследование систем с разделением и параллельным обслуживанием с точки зрения изучения имеющейся зависимости, которую характеризуют не только коэффициенты корреляции, но в большей степени копулы. Копулы позволяют перейти от частных распределений случайных величин к их совместному распределению, тем самым

предполагая полноценное описание зависимости двух величин (в данном случае двух).

Далее будет описан подход для оценки копулы и квантилей распределения времени отклика на примере, когда параметр $\alpha = 4$. При этом предложенный подход может распространяться и на другие значения α.

3. Подход к оценке квантилей распределения времени отклика с помощью диагонального сечения копулы

В статье [1] был предложен подход для оценки квантилей распределения времени отклика fork-join-системы с распределением Парето времени обслуживания и различными вариантами распределений для входящего потока, а именно, распределением Эрланга, гиперэкспоненциальным распределением и пуассоновским входящим потоком. В рамках предложенного подхода была обоснована допустимость использования распределения Фреше для аппроксимации распределения времени отклика R_K fork-join-системы с распределением Парето времени обслуживания и числом $K \geqslant 2$ подсистем.

Далее применялся метод моментов для нахождения оценок параметров $\hat{a}$ и $\hat{b}$, участвующих в выражении для приближения квантилей времени отклика $\hat{x}_p$ уровня p, в результате было получено следующее аналитические выражение:

(2) $\hat{x}_{p,R_K} = \hat{a}_K + \hat{b}_K(-\ln p)^{-1/\alpha}, \quad 0 < p < 1, \quad 2 \leqslant K \leqslant 20,$

где $\hat{a}_K$ и $\hat{b}_K$ определялись с помощью оценок математического ожидания $E[R_K]$ и дисперсии времени отклика $Var[R_K]$ fork-join СМО.

Поскольку рассматривались различные варианты распределений для времен между соседними поступлениями заявок, а формул для оценок математического ожидания и дисперсии времени отклика fork-join-систем известно не так много, то в качестве оценок $E[R_K]$ и $Var[R_K]$ использовались значения, полученные с помощью имитационного моделирования.

В данной работе предлагается еще один подход к оцениванию квантилей времени отклика fork-join СМО для частного случая системы с разделением и параллельным обслуживанием, когда число подсистем $K = 2$. Этот подход тесно связан с элементами теории копул и аналогичен подходу, предложенному в [7] для fork-join-системы с двумя подсистемами типа $M|M|1$. Однако основное отличие случая подсистем типа $M|Pa|1$ от случая экспоненциального обслуживания заключается в том, что вид функции распределения времени пребывания подзаявки в данной подсистеме неизвестен. В то время как для подсистемы типа $M|M|1$ время пребывания подзаявки имеет также экспоненциальное распределение. Этим фактором и обуславливается использование аппроксимации вида (2).

Выражение (2) применимо не только для диапазона значений $2 \leqslant K \leqslant 20$, который рассматривался в более ранних работах при анализе fork-join СМО с распределением Парето времени обслуживания вида (1), но в том числе и для базового случая, а именно, когда $K = 1$. Поскольку подход основывается на оценке двумерных копул, то предлагается использовать результаты оценки для базового случая $K = 1$ при оценке квантилей в случае $K = 2$. При этом в выражении (2) вероятность p необходимо будет заменить на его оценку, которая будет получена далее при анализе случая $K = 2$ с помощью диагонального сечения копул.

Итак, в дальнейшем будем использовать следующее выражение для оценки квантилей распределения времени отклика при $K = 1$:

$$\hat{x}_{p,R_1} = \hat{a}_1 + \hat{b}_1(-\ln p)^{-1/\alpha}, \quad 0 < p < 1, \tag{3}$$

где для параметров $\hat{a}_1$ и $\hat{b}_1$ предположим следующие оценки:

$$\hat{a}_1 = \frac{A_0 + A_1\rho + A_2\rho^2}{1-\rho}, \quad \hat{b}_1 = \frac{B_0 + B_1\rho + B_2\rho^2}{1-\rho}. \tag{4}$$

Здесь исходим из того, что времена отклика растут асимптотически пропорционально $1/(1-\rho)$ при $\rho \to 1$, а значит, так же ведут себя коэффициенты a и b, так что если мы умножим их на $(1-\rho)$, то получим какие-то нелинейные ограниченные функции от ρ на отрезке $[0, 1]$, которые попробуем приблизить квадратичными.

Для определения коэффициентов A_i и B_i, $i = 0, 1, 2$, из (4) воспользуемся методом оптимизации Нелдера – Мида [11]. А именно с помощью симуляции системы $M|Pa|1$ получим множество реализаций случайной величины времени пребывания заявки в СМО, после чего статистически на основе эмпирических данных определим для нее квантили распределения x_{p,R_1}. Затем минимизируем с помощью метода Нелдера – Мида модуль относительной погрешности приближения оценки квантилей, рассчитанных по формулам (3) и (4), относительно данных, полученных с помощью имитационного моделирования, что позволит определить искомые коэффициенты:

$$\left| \frac{x_{p,R_1} - \left(\hat{a}_1 + \hat{b}_1 (-\ln p)^{-1/\alpha} \right)}{x_{p,R_1}} \right| \xrightarrow[A_0,A_1,A_2, \quad B_0,B_1,B_2]{} \min .$$

Для коэффициента загрузки системы рассматривается диапазон значений $\rho = \{0{,}1; 0{,}2; ...; 0{,}9\}$. А для вероятностей p, т.е. уровней квантилей, выбраны значения $\{0{,}30; 0{,}35; ...; 0{,}85; 0{,}90\}$, поскольку, как правило, на практике в большей степени интерес представляют квантили именно более высокого порядка.

В результате получим

$$\begin{aligned} &A_0 \approx 0{,}129674, \quad A_1 \approx -1{,}650335, \quad A_2 \approx 0{,}316858, \\ &B_0 \approx 0{,}702442, \quad B_1 \approx 0{,}917551, \quad B_2 \approx -0{,}149744, \end{aligned} \tag{5}$$

Оценки погрешностей приближений для формул (3)–(5) представлены в таблице 1 и для наглядности изображены на рис. 5а.

Таблица 1. Погрешности приближений значений квантилей распределения времени отклика системы x_{p,R_1} ($K = 1$), рассчитанных с помощью аналитических формул (3),(4) *и* (5) *в сравнении с результатами имитационного моделирования*

Оцениваемая характеристика	Типы ошибок		
	Max APE, %	Min APE, %	MAPE, %
Квантиль x_{p,R_1}	5,731694	0,004086	2,240293

В таблице приведены абсолютные значения относительных погрешностей приближений для 117 рассчитанных значений $\hat{x}_{p,R_1}$: максимальная погрешность приближения (Max APE, %), минимальная погрешность приближения (MinAPE, %) и средняя погрешность (MAPE, %).

Далее допустим, что частная функция распределения случайных величин ξ_1 и ξ_2 времен пребывания подзаявок в подсистемах $M|Pa|1$ имеет вид $G(x)$ и является строго возрастающей. Тогда их совместная функция распределения согласно теореме Скляра представима в виде

$$G_{\xi_1,\xi_2}(x_1,x_2) = P(\xi_1 < x_1, \xi_2 < x_2) = C(G(x_1), G(x_2)),$$

где $C(u_1,u_2)$ — это копула–функция совместного распределения случайных величин ξ_1 и ξ_2. Тогда для случайной величины их максимума $R_2 = \max(\xi_1,\xi_2)$ функция распределения примет вид

$$G_{R_2}(x) = P(\max(\xi_1,\xi_2) < x) =$$

$$= P(\xi_1 < x, \xi_2 < x) = C(G(x), G(x)) = C(u,u).$$

При этом величина

$$\delta(u) = C(u,u), \quad 0 \leqslant u \leqslant 1,$$

называется диагональным сечением копула–функции. Соответственно, имеем, что

$$G_{R_2}(x) = C(G(x), G(x)) = \delta(G(x)).$$

Поэтому уравнение для определения квантили времени отклика принимает вид

$$G_{R_2}(x_{p,R_2}) = \delta(G(x_{p,R_2})) = p,$$

откуда

$$x_{p,R_2} = x_p = G^{-1}\left(\delta^{-1}(p)\right) = x_{\delta^{-1}(p),R_1}. \tag{6}$$

Стоит отметить, что диагональное сечение характеризуется (необходимыми и достаточными) свойствами

$$\begin{gathered}\max\{2u-1,0\}\leqslant\delta(u)\leqslant u;\\ 0\leqslant\delta(u_2)-\delta(u_1)\leqslant 2(u_2-u_1),\quad 0\leqslant u_1\leqslant u_2\leqslant 1,\end{gathered}\tag{7}$$

которыми обладает, в частности, степенная функция

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

При этом значение параметра $\beta=1$ определяет абсолютную положительную зависимость случайных величин, а значение $\beta=2$ — их полную независимость. Поэтому при дальнейших исследования обратимся к копуле Гумбеля

$$\begin{gathered}C(u_1,u_2)=\exp\{-((-\ln u_1)^{\theta}+(-\ln u_2)^{\theta})^{1/\theta}\},\\ \theta\geqslant 1,\quad u_1,u_2\in[0,1],\end{gathered}\tag{8}$$

так как она характеризуется степенным диагональным сечением следующего вида:

$$\delta(u)=u^{2^{1/\theta}}.\tag{9}$$

Введем случайные величины $U_1=G(\xi_1)$ и $U_2=G(\xi_2)$, согласно методу обратного преобразования Н.В. Смирнова случайные величины U_1 и U_2 будут иметь равномерное распределение на отрезке $[0,1]$, т.е.

$$U_i=G(\xi_i)\sim R[0,1],\quad i=1,2.$$

Затем рассмотрим случайную величину $V=\max(U_1,U_2)$, для которой в силу предположения о том, что функция распределения $G(x)$ является строго возрастающей, будет справедливо следующее:

$$\begin{gathered}V=\max(U_1,U_2)=\max(G(\xi_1),G(\xi_2))=\\ =G(\max(U_1,U_2))=P(V<u)=p,\end{gathered}\tag{10}$$

где p – это вероятность, $0\leqslant p\leqslant 1$.

С другой стороны, для диагонального сечения справедливо

$$\delta(u)=C(u,u)=P(\xi_1<x,\xi_2<x)=$$

$$=P(G(\xi_1)<G(x),G(\xi_2)<G(x))=P(U_1<u,U_2<u)=$$

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

т.е.

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

где u_p – это квантиль уровня p распределения случайной величины V. В условиях предположения о степенном виде диагонального сечения получаем выражение для определения вероятности p, которое впоследствии будет использоваться для оценки квантилей времени отклика R_2:

$$p = \delta(u_p) = u_p^{\beta}. \tag{11}$$

Далее с помощью имитационного моделирования системы с разделением и параллельным обслуживанием заявок определяется множество пар значений величин (ξ_1, ξ_2). Речь идет о порядка 5–10 миллионов таких пар, т.е. получаем $N = 5$ миллионов (для низких уровней загрузки системы) или $N = 10$ миллионов (для высоких уровней загрузки) реализаций случайных величин (ξ_1^i, ξ_2^i), где $i = 1, ..., N$.

Затем для каждой пары (ξ_1^i, ξ_2^i) необходимо определить величину V_i, которая согласно (10) определяется выражением $G(\max[G(\xi_1^i), G(\xi_2^i)]) = G(\max(U_1^i, U_2^i))$. Однако, как уже упоминалось ранее, в силу того, что вид функции распределения $G(x)$ неизвестен, напрямую определить V_i по значениям (ξ_1^i, ξ_2^i) не получится. Поэтому воспользуемся универсальным методом нормированных рангов (описанным, например, в [10, §5.5.2]), согласно которому можно использовать асимптотическую оценку

$$\hat{V}_i = \max\left(\hat{G}(\xi_1^i), \hat{G}(\xi_2^i)\right) = \max\left(\hat{U}_1^i, \hat{U}_2^i\right) =$$

$$= \max\left(\frac{\text{rang}(\xi_1^i)}{N+1}, \frac{\text{rang}(\xi_2^i)}{N+1}\right),$$

где rang$(\cdot)$ – это ранг, т.е. порядковый номер реализации аргумента после сортировки в порядке возрастания, $i = 1, ..., N$.

Далее для полученных значений реализаций случайной величины $\hat{V}$ оцениваем квантили ее распределения, т.е. находим ста-

тистические значения (u_p, p), упорядочивая значения $\hat{V}_i$ по возрастанию, в результате получаем

$$(u_p, p) = \left(\hat{V}_{(k)}, \frac{k}{N+1}\right),$$

где $\hat{V}_{(k)}$ – это k-я порядковая статистика, $k = 1, ..., N$. Пары (u_p, p) оцениваются, как и ранее, для различных значений коэффициента загрузки системы $\rho = \{0{,}10; 0{,}15; ...; 0{,}90\}$ и вероятностей $p = \{0{,}30; 0{,}25; ..., 0{,}85; 0{,}90\}$.

Далее на имеющихся данных (u_p, p) для каждого выбранного значения коэффициента загрузки ρ проводится графический анализ для определения конкретного вида функциональной зависимости p от u_p из (11)

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

На рис. 2 изображена зависимость $\ln p$ от значений $\ln u_p$.

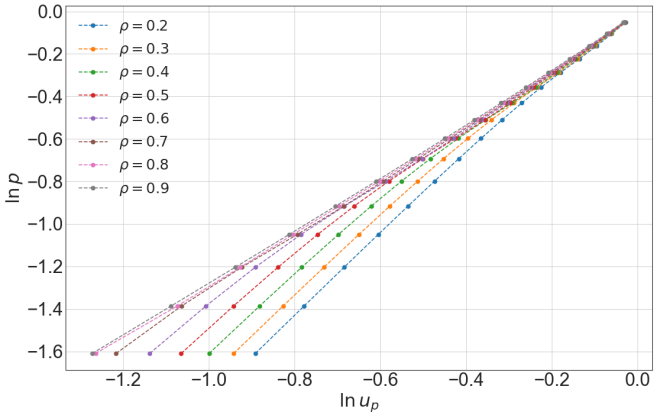


Рис. 2. Зависимость $\ln p$ *от* $\ln u_p$

Как следует из вида графика, между $\ln p$ и $\ln u_p$ можно допустить линейную зависимость, причем пучок прямых проходит через начало координат, соответственно, свободный коэффициент отсутствует, т.е. получаем соотношение

$$\ln p \approx \widehat{\beta}(\rho) \cdot \ln u_p,$$

что соответствует предположению о степенной зависимости p от u_p.

Далее подберем функциональную зависимость для $\widehat{\beta}(\rho)$. Для этого построим на имеющихся данных график зависимости отношения $\ln p / \ln u_p$. Из вида рис. 3 можно предложить квадратичную зависимость.

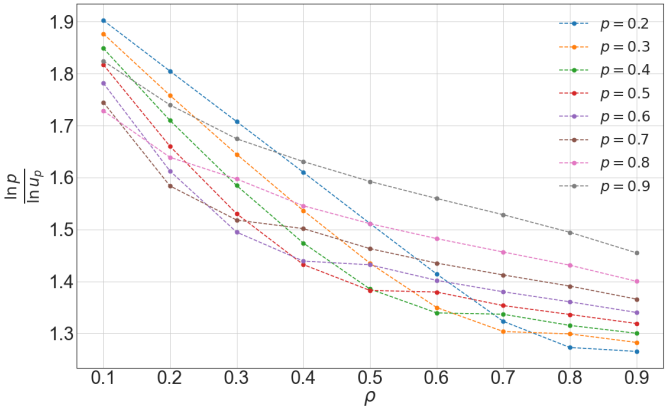


Рис. 3. Зависимость $(\ln p / \ln u_p)$ *от* ρ

При $\rho \to 0$ времена пребывания подзаявок асимптотически независимы, что в соответствии с теорией копул дает $\beta \to 2$. Поэтому логично допустить, что кривые проходят через точку с координатами $(0, 2)$, при этом вершина параболы сдвинута вправо и обращена вниз, что говорит об отрицательном значении числового коэффициента перед ρ и положительном при ρ. Таким образом, имеем

$$\frac{\ln p}{\ln u_p} = \beta(\rho) \approx 2 - C_1\rho + C_2\rho^2.$$

В результате получаем следующее выражение:

$$p = \delta(u_p, \rho) \approx u_p^{2 - C_1\rho + C_2\rho^2}. \tag{12}$$

Теперь необходимо найти значения коэффициентов C_1 и C_2. Для этого воспользуемся методом оптимизации Нелдера – Мида

и минимизируем модуль относительной погрешности полученного аналитического (степенного) приближения $\widehat{p}$ по сравнению со значениями p, полученными с помощью имитационного моделирования:

$$\left| \frac{p - u_p^{2-C_1\rho+C_2\rho^2}}{p} \right| \xrightarrow[C_1,C_2]{} \min .$$

Таким образом, значения коэффициентов равны

$$C_1 \approx 1{,}334476, \quad C_2 \approx 0{,}550919, \tag{13}$$

соответственно, искомая оценка имеет вид

$$p = \delta(u_p, \rho) \approx u_p^{2-1{,}334476\rho+0{,}550919\rho^2}. \tag{14}$$

На рис. 4 представлены результаты имитационного моделирования вероятностей или уровней p квантилей u_p случайной величины $V = G(\max(U_1, U_2))$ в сравнении с результатами вычислений по аналитической формуле (14) в диапазоне значений $[0{,}30; 0{,}90]$ с шагом $0{,}05$.

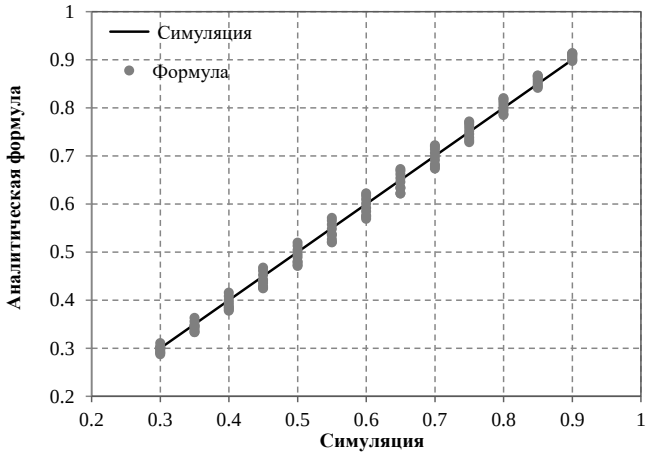


Рис. 4. Сравнение аналитических результатов формулы (14) *с имитационным моделированием значений* p *квантилей* u_p *случайной величины* $V = G(\max(U_1, U_2))$ *для значений* $\rho \in \{0{,}10; 0{,}20; ...; 0{,}90\}$

Каждая точка, изображенная на графике, представляет собой множество из 9 точек по числу значений коэффициента загрузки $\rho \in \{0{,}10; 0{,}20; ..., 0{,}90\}$, которые накладываются друг на друга. Для ясности в таблице 2 приведены абсолютные значения относительных погрешностей приближений для 117 рассчитанных значений p.

Таблица 2. Погрешности приближений значений вероятностей p, рассчитанных с помощью аналитической формулы (14) *в сравнении с результатами имитационного моделирования*

Оцениваемая характеристика	Типы ошибок		
	Max APE, %	Min APE, %	MAPE, %
p из (14)	5,731694	0,061445	2,311958

Сравнительная анализ результатов имитационного моделирования уровня p квантилей u_p с.в. V с результатами аналитической формулы (14) показал, что средняя погрешность приближения составляет около 2,2%.

Далее, поскольку вид функции распределения $G(x)$ случайных величин ξ_1 и ξ_2 времен пребывания подзаявок в подсистемах $M|Pa|1$ неизвестен, то воспользуемся соотношением (3) для квантилей. А именно, при $K = 1$ имеем

$$G(x_{p,R_1}) = p, \quad x_{p,R_1} = G^{-1}(p) = \hat{a}_1 + \hat{b}_1(-\ln p)^{-1/\alpha}.$$

Теперь с учетом уравнения (6) для определения квантилей времени отклика R_2 в случае $K = 2$ выразим $\delta^{-1}(p)$:

$$\delta(u_p) = u_p^{2-C_1\rho+C_2\rho^2} = p,$$

$$(2 - C_1\rho + C_2\rho^2)\ln u_p = \ln p,$$

$$u_p = \delta^{-1}(p) = p^{\frac{1}{2-C_1\rho+C_2\rho^2}}.$$

Поэтому можем записать следующее

$$x_{p,R_2} = x_p = G^{-1}(\delta^{-1}(p)) = G^{-1}\left(p^{\frac{1}{2-C_1\rho+C_2\rho^2}}\right) =$$

$$= \hat{x}_p = \hat{a}_1 + \hat{b}_1 \left(- \ln p^{\frac{1}{2 - C_1\rho + C_2\rho^2}} \right)^{-1/\alpha}.$$

В результате окончательно получаем следующее выражение для оценки квантилей времени отклика R_2 уровня p, $0{,}30 \leqslant p \leqslant 0{,}90$:

$$\hat{x}_p = \hat{a}_1 + \hat{b}_1 \left(- \frac{\ln p}{2 - 1{,}334476\rho + 0{,}550919\rho^2} \right)^{-1/\alpha}, \tag{15}$$

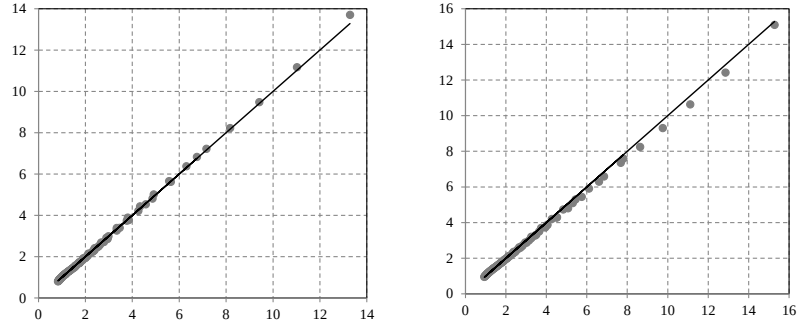
$$0 < p < 1,$$

где

$$\hat{a}_1 \approx \frac{0{,}129674 - 1{,}650335\rho + 0{,}316858\rho^2}{1 - \rho}, \tag{16}$$

$$\hat{b}_1 \approx \frac{0{,}702442 + 0{,}917551\rho - 0{,}149744\rho^2}{1 - \rho}.$$

Далее аналогично оценим качество аппроксимации полученного выражения (15) для 117 рассчитанных значений квантилей при $\rho \in \{0{,}10; 0{,}20; ...; 0{,}90\}$ и $p \in \{0{,}20; 0{,}25; ...; 0{,}90\}$. Результаты представлены в таблице 3 и на рис. 5б. Получаем, что максимум модуля относительной ошибки составляет около $5{,}7\%$, а среднее значение этого модуля равно примерно $2{,}2\%$.



а) б)

Рис. 5. Сравнение эмпирических и аналитических квантилей распределения случайной величины времени отклика R_K системы с разделением и параллельным обслуживанием: а) рассчитанных по формулам (3)–(5)*, $K = 1$; б) рассчитанных по формулам* (15)–(16)*, $K = 2$*

Таблица 3. Погрешности приближений значений квантилей распределения времени отклика системы x_p ($K = 2$), рассчитанных с помощью аналитических формул (15) *и* (16) *в сравнении с результатами имитационного моделирования*

Оцениваемая характеристика	Типы ошибок		
	Max APE, %	Min APE, %	MAPE, %
Квантиль x_{p,R_2}	5,476451	0,001325	2,949008

4. Приближение копулы времен пребывания подзаявок в подсистемах копулой Гумбеля

В предыдущем разделе была получена оценка диагонального сечения копулы $\delta(u)$. В данном разделе будет представлено аналитическое выражение, оценивающее саму копулу $C(u_1, u_2)$. Для этого потребуются эмпирические данные, проанализировав которые можно будет сделать вывод о близости исследуемой копулы к одному из известных семейств, в частности, копулам Гумбеля.

Алгоритм построения эмпирической копулы будет следующим:

1) имитационное моделирование множества пар (ξ_1^k, ξ_2^k) случайных величин времен пребывания в подсистемах $M|Pa|1$ fork-join СМО, где k – порядковый номер смоделированной пары значений, $k = 1, ..., N$, N – объем выборки (общее число пар случайных величин);

2) преобразование случайных величин (ξ_1^k, ξ_2^k) методом нормированных рангов (см. [10, §5.5.2]) в случайные величины с асимптотически равномерным распределением на отрезке $[0, 1]$, $U_i \sim R[0, 1]$, $i = 1, 2$:

$$(U_1^k, U_2^k) = \left(\frac{\text{rang}(\xi_1^k)}{N+1}, \frac{\text{rang}(\xi_2^k)}{N+1} \right);$$

3) разбиение единичного квадрата на более мелкие квадраты (сетку) со сторонами длиной $h = 1/m$, где, например, $m = 20$ и определение числа точек (U_1^k, U_2^k), попадающих в каждый из квадратов, вершинами которого являются точки $(0, 0)$, $(ih, 0)$,

$(0, jh)$, (ih, jh), $i, j = 1, ..., m$, и нормирование полученного значения, т.е.

$$C_{ij} = C(ih, jh) \approx \widehat{C}_{ij} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}\{U_1^k < ih, U_2^k < jh\},$$

где $\mathbf{1}\{\cdot\}$ – функция-индикатор события $\{\cdot\}$.

На рис. 6 представлен график эмпирической копулы или, что то же самое, совместной функции распределения случайного вектора (U_1, U_2), построенной в соответствии с представленным выше алгоритмом.

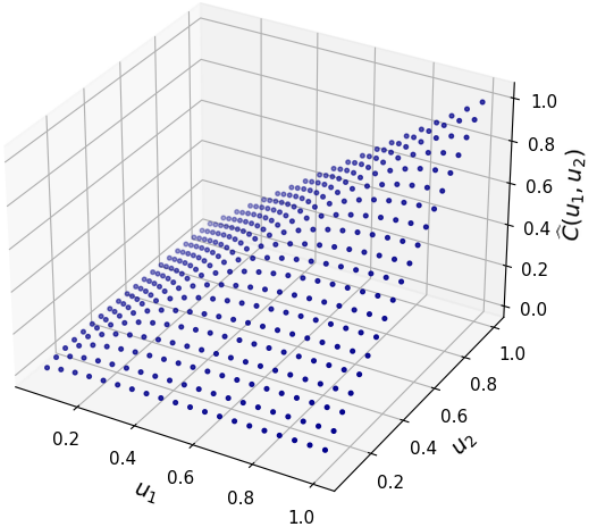


Рис. 6. Эмпирическая копула $\widehat{C}(u_1, u_2)$

Исходя из внешнего вида полученной эмпирической функции на рис. 6, а также учитывая, что диагональное сечение рассматриваемой копулы было приближено в предыдущем разделе выражением вида

(17) $$\delta(u) \approx u^{\beta}, \quad \beta = 2 - C_1\rho + C_2\rho^2,$$

будем приближать искомую копулу $C(u_1, u_2)$ копулой Гумбеля, которая имеет вид

$$C_g(u_1, u_2) = \exp\{-[(-\ln u_1)^\theta + (-\ln u_2)^\theta]^{\frac{1}{\theta}}\}, \tag{18}$$

где $\theta \in [1, +\infty)$ – параметр копулы, который предстоит оценить.

Поскольку для копулы Гумбеля диагональное сечение имеет вид

$$\delta_g(u) = C_g(u, u) = u^{2^{1/\theta}},$$

то с учетом (17) получаем, что

$$\theta \approx \frac{\ln 2}{\ln \beta} = \frac{\ln 2}{\ln(2 - C_1\rho + C_2\rho^2)}. \tag{19}$$

Далее снова воспользуемся методом оптимизации Нелдера – Мида для минимизации модуля относительной ошибки приближения функции копулы Гумбеля (18) с учетом того, что параметр θ определяется выражением (19), при сравнении с «истинными» значениями функции копулы Гумбеля, полученными с помощью имитационного моделирования для различных коэффициентов загрузки $\rho \in \{0{,}10; 0{,}20; ...; 0{,}90\}$. Как и раньше, не будем рассматривать квантили низкого уровня, т.е. пусть $u_1, u_2 \in \{0{,}30; 0{,}35; ...; 0{,}90\}$. В результате получаем следующие значения искомых коэффициентов:

$$C_1 \approx 1{,}068768, \quad C_2 \approx 0{,}202125, \tag{20}$$

поэтому

$$C(u_1, u_2) \approx \exp\left\{ -\left((-\ln u_1)^{\frac{\ln 2}{\ln(2-1{,}069\rho+0{,}202\rho^2)}} + \right.\right. \\ \left.\left. +(-\ln u_2)^{\frac{\ln 2}{\ln(2-1{,}069\rho+0{,}202\rho^2)}} \right)^{\frac{\ln(2-1{,}069\rho+0{,}202\rho^2)}{\ln 2}} \right\}. \tag{21}$$

Как видно из (13) и (20), полученные оценки коэффициентов различны, однако соответствующие им оценки функции $\beta(\rho)$ мало различаются между собой на рассматриваемом промежутке загрузки ρ, что позволяет говорить о значительной согласованности между ними.

Что касается погрешности аппроксимации формулы (21), то в таблице 4 представлены значения максимальной (Max APE),

минимальной (Min APE) и средней относительной ошибки аппроксимации (MAPE), первая из которых не превышает 10%, на наборе данных из 1521 троек (ρ, u_1, u_2). В той же таблице приведены результаты в случае оценки по коэффициентам из (13). На рис. 7 также представлены графики эмпирической функции копулы и копулы, определяемой выражением (21) на заданном диапазоне значений $0{,}3 \leqslant u_1, u_2 \leqslant 0{,}9$.

Таблица 4. Погрешности приближений функции копулы Гумбеля $C(u_1, u_2)$ формулой (21)

Оцениваемая характеристика	Типы ошибок		
	Max APE, %	Min APE, %	MAPE, %
$C(u_1, u_2)$, C_1, C_2 из (20)	8,248810	0,000686	2,735456
$C(u_1, u_2)$, C_1, C_2 из (13)	14,733185	0,000385	1,999350

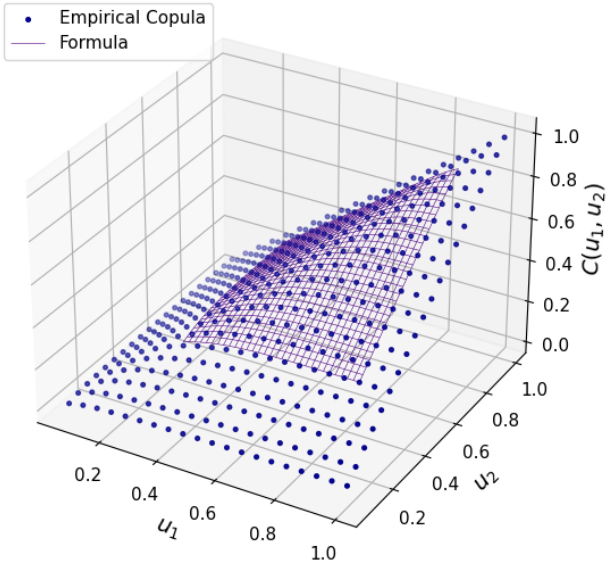


Рис. 7. Сравнение эмпирической копулы $\widehat{C}(u_1, u_2)$ и аналитической формулы (21)

Заметим также, что если проводить оценку параметра θ классическим методом максимального правдоподобия (соответствующей функцией Python для копулы Гумбеля), то полученные значения, количество которых в данном случае будет соответствовать количеству значений коэффициента корреляции $\rho \in [0{,}1; 0{,}9]$ с шагом $0{,}10$ (т.е. их будет всего 9), на тех же 1521 тройках значений (ρ, u_1, u_2) приближение копулой Гумбеля показывает большие погрешности. В этом случае Max APE $\approx 15{,}7\%$, Min APE $\approx 0{,}0009\%$ и MAPE $\approx 2{,}34\%$.

5. Заключение

В настоящей работе изучены 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 значений вероятностей или уровней квантилей, а также значений коэффициента загрузки системы. Несмотря на то, что в рамках данной работы рассматривалось определенное значение параметра распределения Парето времени обслуживания, представленный в статье подход, основанный на элементах теории копул, можно применить аналогичным образом и для других значений параметра, а также обобщить на системы с большим количеством подсистем.

Литература

1. ГОРБУНОВА А.В., ЛЕБЕДЕВ А.В. *Квантили распределения времени отклика в fork-join системах с распределением Парето времени обслуживания* // Вестник ВГУ. Серия: Системный анализ и информационные технологии. – 2024. – №3. – С. 5–16.
2. ARMONY M., ISRAELIT S., MANDELBAUM A. et al. *Patient flow in hospitals: a data-based queueing-science*

perspective // Stochastic Systems. – 2015. – Vol. 5, No. 1. – P. 146–194.

3. BACCELLI F., MAKOWSKI A.M. *Queueing models for systems with synchronization constraints* // Proc. of the IEEE. – 1989. – Vol. 77, No. 1. – P. 138-161.
4. ENGANTI P., ROSENKRANTZ T., SUN L. 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.
5. GALLIEN J., WEIN L.M. *A simple and effective component procurement policy for stochastic assembly Systems* // Queueing Systems. – 2001. – Vol. 38. – P. 221–248.
6. 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.
7. GORBUNOVA A.V., LEBEDEV A.V. *Copulas and quantiles in fork-join queueing Systems* // Advances in Systems Science and Applications. – 2024. – Vol. 24, No. 1. –1–19.
8. GUDENDORF G., SEGERS J. *Extreme-Value Copulas* // In: Copula theory and Its Application. – Springer, 2010. – P. 127–145.
9. 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.
10. MCNEIL A.J., FREY R., EMBRECHTS P. *Quantitative risk management*. – Princeton: Princeton University Press, 2005. – 538 p.
11. NELDER J.A., MEAD R. *A simplex method for function minimization* // The Computer Journal. – 1965. – Vol. 7. – P. 308–313.
12. NELSEN R.B. *An introduction to copulas*. – Springer Science & Business Media, 2007. – 272 p.

13. 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.
14. NGUYEN M., ALESAWI S., LI N. et al. *ForkTail: A black-box fork-join tail latency prediction model for user-facing datacenter workloads* // *Proc. of the 27th Int. Symposium on High-Perform. Parallel Distrib. Comput.*. – 2018. – P. 206–217.
15. NGUYEN M., ALESAWI S., LI 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.
16. OLIVEIRA D.C.M., LIU J., PACITTI E. *Data-intensive workflow management: for clouds and data-intensive and scalable computing environments* // Synthesis Lectures on Data Management. – 2019. – Vol. 14, No. 4. – P. 1–179.
17. 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.
18. 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.
19. THOMASIAN A. *Analysis of fork/join and related queueing systems* // ACM Computing Surveys (CSUR). – 2014. – Vol. 47, No. 2. – P. 17:1–17:71.
20. VARKI E., MERCHANT A., CHEN H. *The M/M/1 fork-join queue with variable subtasks* // Unpublished. – URL: https://www.cs.unh.edu/∼varki/publication/2002-nov-open.pdf (дата обращения: 02.09.2024).
21. VARMA S., MAKOWSKI A.M. *Interpolation approximations for symmetric fork-join queues* // Performance Evaluation. – 1994. – Vol. 20. – P. 245–265.

ESTIMATES OF THE COPULA AND QUANTILES OF THE RESPONSE TIME DISTRIBUTION FOR A FORK-JOIN QUEUEING SYSTEM WITH THE PARETO DISTRIBUTION OF SERVICE TIME

Anastasia Gorbunova, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Cand.Sc., Senior Researcher (avgorbunova@list.ru).

Abstract: A fork-join queueing system is considered. It is assumed that the service time distribution on all servers has a Pareto distribution. The dependence between the sojourn times of subtasks in subsystems is studied, which is the main reason for the complexity of analyzing such systems. The sojourn time of a task in the system (or the average response time) is the maximum of the dependent random variables of the sojourn time of subtasks in the system. Approximations of the joint distribution of the sojourn times of subtasks are obtained using copula theory. An approach is also proposed for determining the quantiles of the system response time distribution using a diagonal section of copulas. This approach was previously used to analyze a similar system, but with an exponential distribution of service time. However, the main difference between the system under study and the exponential case is that the type of the distribution function of the sojourn time of a subtask in the subsystem is unknown. Therefore, an analytical approximation is used for the quantiles of the response time distribution in the subsystem under the assumption that the distribution of the time of stay of a subtask in the subsystem is approximated by the Frechet distribution obtained earlier. The estimates obtained for the quantiles and copula of the response time distribution show good agreement with the simulation data.

Keywords: fork-join queueing system, queueing system, distribution quantiles, copula.

УДК 519.2
ББК 22.17

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

Поступила в редакцию 25.10.2024.
Дата опубликования 30.11.2024.