Adv Syst Sci Appl 2024; **01**:1–19
Published online at https://ijassa.ipu.ru.

Copulas and Quantiles in Fork-Join Queueing Systems

Anastasia V. Gorbunova[1]*, Alexey V. Lebedev[2]

[1]*V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia*
[2]*Lomonosov Moscow State University, Moscow, Russia*

Abstract: The article considers a classic fork-join queueing system with a Poisson input flow and exponential service times on homogeneous servers. When entering the system, tasks are divided into smaller components (subtasks), the number of which is equal to the number of subsystems. Then the subtasks are sent for service to the corresponding subsystems, consisting of a storage device of unlimited capacity and one server. The described functioning mechanism allows, using fork-join systems, to simulate processes occurring in many different real physical systems where tasks are parallelized. The article studies the dependence between the sojourn times of subtasks in subsystems, which is at the same time the main reason for the complexity of analyzing such systems. An approach is proposed for determining the quantiles of the response time distribution, while most works in the field of analysis of fork-join systems are concentrated on obtaining approximations only for the mathematical expectation of the response time. In addition, the approximation of the copula of sojourn times of subtasks (parts of one task) by the Gumbel copula and a new Kendall's correlation coefficient estimation are obtained.

Keywords: fork-join queueing system, response time, distribution quantiles, correlation coefficients, Gumbel copula, diagonal section, simulation modeling.

1. INTRODUCTION

This paper examines a classic fork-join queueing system (QS) with a Poisson input flow and exponential service times. Fork-join QS is a mathematical model for many real-life systems in which tasks are parallelized. When entering the system, the task is divided (fork point) into a number of subtasks equal to the number of subsystems $K \geq 2$. All subsystems are actually independent queueing systems with an infinite queue and a single server. Each of the subtasks after entering one of the subsystems is serviced there, and then enters a conditional synchronization buffer (join point), where it awaits servicing of the remaining parts of the task. After servicing of all subtasks is completed, the entire task is instantly assembled and can leave the system.

Previously, such systems were studied by the authors in the works [10, 11, 14]. This work continues the article [11], where a study began on the dependence of the sojourn times of subtasks (parts of one task). The analysis of this system, but with a more complex architecture, can be found, for example, in one of the latest work [12].

The first publications aimed at studying the performance characteristics of mathematical models of fork-join QS and similar systems began to appear in the second half of the 20th century. Then research activity decreased somewhat. However, currently there is a new round of interest in the analysis of fork-join such systems, especially in the field of modeling information systems and the processes occurring in them. One of the reasons

*Corresponding author: avgorbunova@list.ru for this situation is the growing popularity of data-intensive applications. Among the basic principles of the functioning of high-performance computing environments the parallel data processing, i.e. the simultaneous execution of several operations, commands or actions, is distinguished. For example, one of the main Big Data technologies from Google is called MapReduce, the essence of which is to divide information flows into parts, process them in parallel and further combine the results obtained. This allows you to significantly increase the performance of relevant applications and at the same time reduce the time spent on processing big data, the volume of which continues to increase and, according to IDC (International Data Corporation) forecasts, and can reach 175 zettabytes by 2025 [25, 30, 31, 34].

Of course, the scope of application of fork-join models is not limited only to information and computing systems. Issues of optimizing processes in production systems (for example, assembling orders in warehouse systems, manufacturing multi-component products) or improving the efficiency of organizing the process of patient stay in medical institutions(and etc.) continue to be in demand to this day [1–3, 9, 18, 22, 32].

Another reason for the actualization of research on fork-join systems is the emergence of new methods and approaches to the analysis of complex queueing systems, in particular, an approach based on the use of machine learning methods and its various modifications [4–7, 19, 27, 28, 35]. In this case, we are talking about the further development of this approach and the inclusion of graphical analysis and optimization methods in its composition.

Despite the apparent simplicity of operation, the study of fork-join QS is one of the most difficult problems to solve. The main reason for the complexity lies in the commonality of the moments of appearance of subtasks in the subsystems, which makes their sojourn times dependent random variables. This is the main difference between a fork-join system and simply parallel operating QSs of the same type as the fork-join QS subsystems. Therefore, accurate results were obtained only for the average response time in the case of two subsystems with a Poisson input flow and exponential service times [24]. For other variants of fork-join QS architectures, which imply, for example, an increase in the number of subsystems or servers in them, limited storage capacity, or other types of distributions for input and service flows, only approximations of the mathematical expectation of response time were obtained in various ways.

As for assessing other performance characteristics of a fork-join system, there is much less research in this direction. For example, the variance and standard deviation of response time was analyzed in the papers [8, 14]. In the work [11] you can find analytical expressions (exact or their estimates) for the correlation coefficients between the sojourn times of subtasks in subsystems

However, in addition to the first or second moments of the response time random variable, the quantiles of its distribution are of interest. In [15], by using computer simulation, quantiles of response times were found in a system with $K = 3$ at various load values. In [29], using vector-matrix techniques and phase distributions, theoretical estimates of the tail and quantiles of high levels are obtained. In [26] estimates were obtained for high-level quantiles under conditions of high system load for several types of distributions and fork-join QS architectures.

This paper proposes an approach for finding response time quantiles of various levels for a wider range of loads, which allows us to get a complete picture of the behavior of the random variable under study. Despite the fact that the object of study of the article is the classic fork-join QS, the approach proposed in the article can be extended to other architectures of fork-join systems and beyond.

Our approach to constructing an estimation of response time quantiles is based on working with copulas and their diagonal sections. Copulas are multivariate distribution functions on a unit cube with uniform marginal distributions. According to Sklar's theorem, any multidimensional distribution can be decomposed into marginal distributions and a copula. Thus, copula exhaustively describes the dependence of random variables in its pure form. The

Copyright © 2024 ASSA.

modern mathematical apparatus of copula theory has been actively developed and applied in recent decades, but it is still poorly represented in queueing theory.

For example, in [33] a system with dependent service times for subtasks in two parallel queues was considered. A task is considered served if at least one subtask is serviced. Marginal distributions of service times are assumed to be shifted exponential or hypoexponential, and their dependence is described by an artificially introduced copula.

In our works we study copulas that arise naturally during the functioning of the system. In this regard, we note the article by the authors of [13], where copulas of maximum remaining service times in infinite-server fork-join systems were found.

The article is organized as follows. Section 2 presents a description of the system under consideration and provides some new results for the correlation coefficients between the sojourn times of subtasks in subsystems, Section 3 gives the necessary elements of the theory of copulas, Section 4 describes the approach to determining the approximation of the diagonal section of the copula and the quantiles of the response time distribution, in Section 5 presents the results of approximating the copula of sojourn times of subtasks by the Gumbel copula, Section 6 compares it with previously known results, and the Conclusion summarizes some results.

2. MATHEMATICAL MODEL OF FORK-JOIN QS AND CORRELATION COEFFICIENTS

Let us describe in more detail the process of functioning of the fork-join system. We will consider the special case of two subsystems ($K = 2$), however, note that the number of subsystems does not in any way affect the dependence in any pair of sojourn times of subtasks of one task (Fig. 2.1). The system receives a Poisson flow of tasks with rate $\lambda > 0$. At the moment the task is received into the system, it is instantly divided into 2 subtasks, each of which falls into the corresponding subsystem, which has a storage device of unlimited capacity and one server. All servers are homogeneous, the service time has an exponential distribution with parameter $\mu > 0$. Thus, the subsystems represent two identical QSs of type $M|M|1$. Since a task is considered to be serviced only after the completion of servicing of both of its subtasks, the random sojourn time of the task in the QS (response time) R is the maximum of two random sojourn times of subtasks ξ_i, $i = 1, 2$, in each from two subsystems:

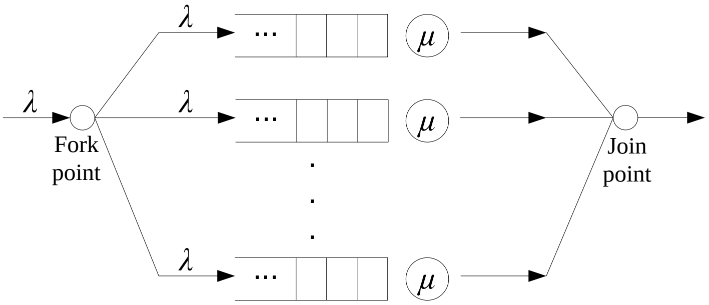


Fig. 2.1. Fork-join model of a queue ing system with subsystems of type $M_\lambda|M_\mu|1$.

$$R = \max\{\xi_1, \xi_2\}. \tag{2.1}$$

Random variables ξ_1 and ξ_2 are correlated due to the fact that all subtasks (parts of one task) arrive at the subsystems at the same time. In the work [11] you can familiarize yourself in detail with the analysis of this dependence, as well as with the derivation of analytical expressions for various correlation coefficients between any pair of sojourn times of subtasks in subsystems. Since this article is a logical continuation of the research begun in [11], in order to understand the whole picture we present the exact analytical formulas obtained for the Pearson and Spearman correlation coefficients

$$r_p = \frac{\rho(4-\rho)}{8}, \tag{2.2}$$

$$r_s = \frac{12\sqrt{2}\sqrt{2-\rho}}{8-3\rho} - 3. \tag{2.3}$$

as well as an approximate expression for the Kendall correlation coefficient

$$r_k \approx \rho(0.25134 + 0.02517\rho), \tag{2.4}$$

where $\rho = \lambda/\mu$ is the system load factor.

In what follows, for simplicity, we assume $\lambda = 1$, $\mu = 1/\rho$. Such parameters were used in the simulation [10, 14].

Note that various correlation coefficients, in the general case, only partially reflect the dependence. Just copulas fully reflect the dependence.

3. ELEMENTS OF COPULA THEORY

A copula C is a multivariate distribution function on $[0,1]^d$, $d \geq 2$, if all marginal distributions are uniform on $[0,1]$. According to the famous Sklar theorem, any multivariate distribution function in $\mathbb{R}^d$ can be represented in the form

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

where F_i, $1 \leq i \leq d$, are functions of private distributions. Thus, every multidimensional distribution can be associated with its copula. If marginal distributions are continuous, then such a representation is unique.

As a classic textbook on copulas we point out [23].

In what follows, we restrict ourselves to the case of two-dimensional copulas ($d = 2$).

The diagonal section of a (two-dimensional) copula is the function $\delta(u) = C(u,u)$, $u \in [0,1]$. It has the following (necessary and sufficient) properties:

$$\max\{2u-1, 0\} \leq \delta(u) \leq u; \quad 0 \leq \delta(u_2) - \delta(u_1) \leq 2(u_2 - u_1), \quad 0 \leq u_1 \leq u_2 \leq 1. \tag{3.5}$$

The point of studying diagonal sections, for example, is as follows. If random variables X_1 and X_2 are given with identical marginal distributions $F_1 = F_2 = F$ and joint distribution copula C, then their maximum is $X_{\max} = \max\{X_1, X_2\}$ has a distribution function

$$F_{\max}(x) = P(X_1 < x, X_2 < x) = C(F(x), F(x)) = \delta(F(x)), \tag{3.6}$$

so to calculate it we need to know only the diagonal section, and not the entire copula.

It is easy to see that the conditions (3.5) are satisfied by the power function

$$\delta(u) = u^\alpha, \quad 1 \leq \alpha \leq 2,$$

then the case $\alpha = 1$ corresponds to a perfect positive dependence (comonotonicity), and the case $\alpha = 2$ corresponds to independence of random variables.

A classic example of an absolutely continuous (having density) copula with a power-law diagonal section is the Gumbel copula

$$C(u_1, u_2) = \exp\{-((-\ln u_1)^\theta + (-\ln u_2)^\theta)^{1/\theta}\}, \quad \theta \geq 1, \quad u_1, u_2 \in [0,1],$$

Copyright © 2024 ASSA.

then

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

More precisely, a Gumbel copula belongs to the class of extreme value copulas, which always have power-law diagonal sections. You can read about such copulas in [23, § 3.3.4], [16], etc.

Authors' works related to copulas can be listed as [13, 20, 21].

4. APPROXIMATIONS OF THE DIAGONAL SECTION AND RESPONSE TIME QUANTILES

Determining the quantiles of the response time distribution is no less important than finding the average response time, and sometimes even more important. Understanding how a system behaves under high load and what maximum response time delays are possible is of great value.

Despite the fact that in this work we consider fork-join QS with $M|M|1$ subsystems, the results of estimating the upper quantiles can be used to initially predict the behavior of models that clearly do not have heavy-tailed distributions, for example, these can be various options for small warehouse or production systems, modeling the process of admission, stay and discharge of patients from medical institutions or the process of reviewing a client's loan application in a financial institution and, possibly, even some private computing systems with the implementation of distributed or parallel computing, operated by employees of a small company.

To approximate the quantiles of the distribution of the random variable response time $R = \max\{\xi_1, \xi_2\}$ we will use elements of copula theory. We will consider a two-dimensional copula $C(u_1, u_2)$ of random vectors of sojourn times in subsystems (ξ_1, ξ_2). Each component of the random vector has an exponential distribution with the distribution function $F(x) = 1 - e^{-(\mu-\lambda)x}$, $x \geqslant 0$. Then, in accordance with Sklar's theorem, a representation using the copula of the joint distribution (ξ_1, ξ_2) exists and is unique

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

Due to (3.6) we obtain

$$F_R(x) = C(F(x), F(x)) = \delta(F(x)), \tag{4.7}$$

where $\delta(u) = C(u, u)$ is the diagonal section of the copula, which gives us the equation for the quantile level p of the response time distribution

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

so

$$x_p = F_R^{-1}(p) = F^{-1}(\delta^{-1}(p)). \tag{4.8}$$

Taking into account the inverse transformation method for generating random variables with a given distribution function, consider

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

These random variables will have a uniform distribution on the interval $[0, 1]$, i.e. $U_i \sim R[0, 1]$. Then

$$V = \max\{U_1, U_2\} = 1 - e^{-(\mu-\lambda)\cdot\max\{\xi_1,\xi_2\}} = 1 - e^{-(\mu-\lambda)R}. \tag{4.9}$$

Figure 4.2 shows the set of points (U_1, U_2) at $\rho = 0.9$ for a fork-join system with two subsystems $M|M|1$ and for the case of two independent parallel operating QSs $M|M|1$ with the same parameters. The number of pairs of dots is 200 thousand; increasing their number overloads the illustration. As you can see, in Figure 4.2 b) the points are distributed uniformly inside the unit square, which is typical for the case of independent random variables. In 4.2 a) such uniformity in the distribution of points is no longer observed, even though the value of the Pearson correlation coefficient between ξ_1 and ξ_2 is small ($r_p = 0.34875$), and visual analysis is somewhat difficult, however, the relationship between $U_1 = F(\xi_1)$ and $U_2 = F(\xi_2)$ in this case is obviously traceable.

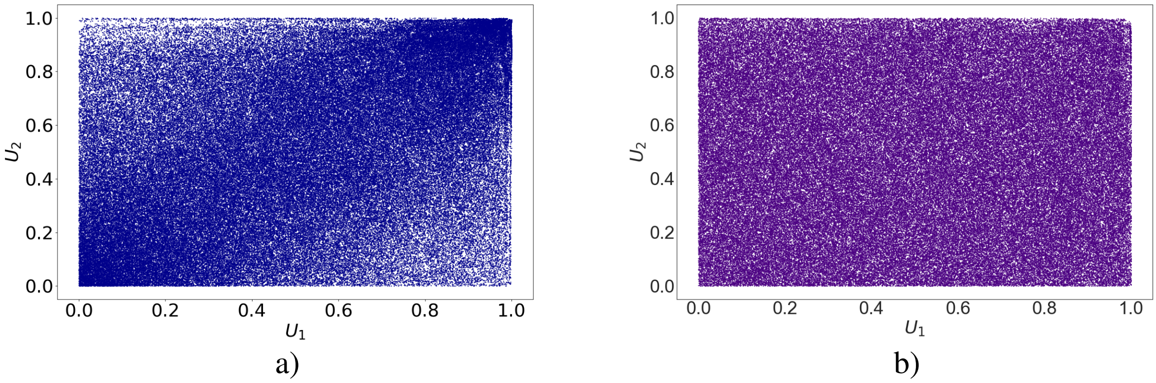


Fig. 4.2. Illustration of the presence/absence of dependence between U_1 and U_2 at $\rho = 0.9$ in the case of a) fork-join QS with two subsystems $M|M|1$; b) two parallel functioning QS $M|M|1$.

The diagonal section of the copula can be estimated as follows. We have

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

i. e.

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

where u_p is the quantile of the distribution of the random variable V. Using realizations V_i of the random variable V, obtained through simulation of the values of random stay times in the fork-join QS R_i and further substituting them into the formula (4.9), we construct an estimate of the diagonal section $\delta(u)$, but in fact, probabilities p. In other words, we construct an empirical estimation of the diagonal section using quantiles of the V distribution. To do this, we order the values V obtained through simulation: $V_{(1)}$, $V_{(2)}$,..., $V_{(N)}$, where $V_{(k)}$ — this is the k-th order statistic, $k = 1, ..., N$, and from the points $(V_{(k)}, k/(N+1))$ we determine the estimates (u_p, p) for probability values from the interval of interest to us $p \in \{0.2, 0.25, 0.30, ..., 0.90\}$, for a specific fixed value of the load factor $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$. The choice of p values is due to the fact that, as a rule, quantiles of higher levels are of greater interest, so we begin to consider p values from 0.2. Next, based on the available data, we will build a forecast of probabilities p depending on the quantiles u_p and the load factor ρ.

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

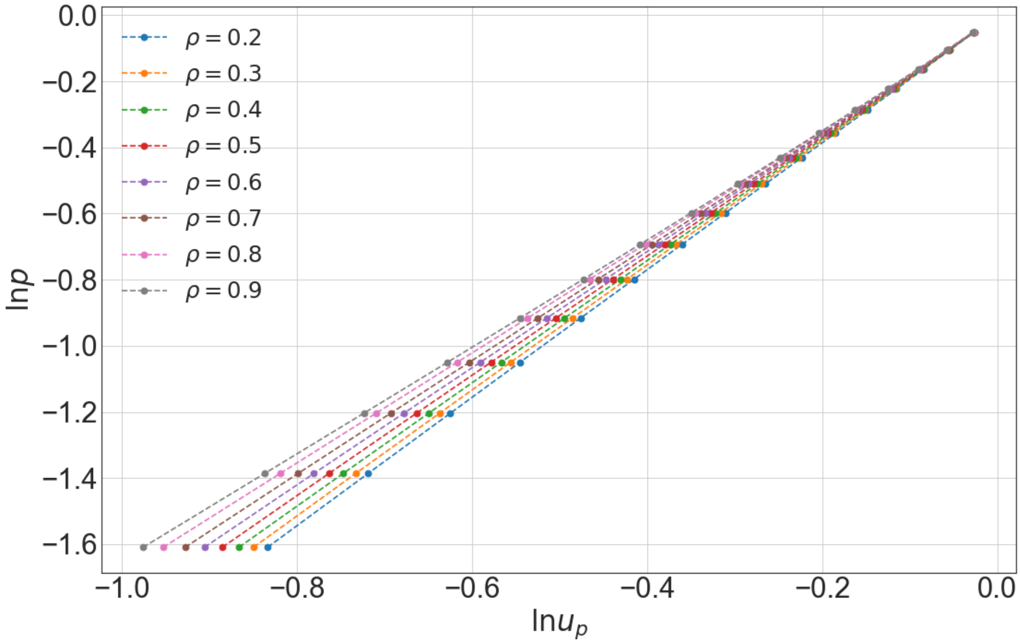


Fig. 4.3. Dependence of $\ln p$ on $\ln u_p$.

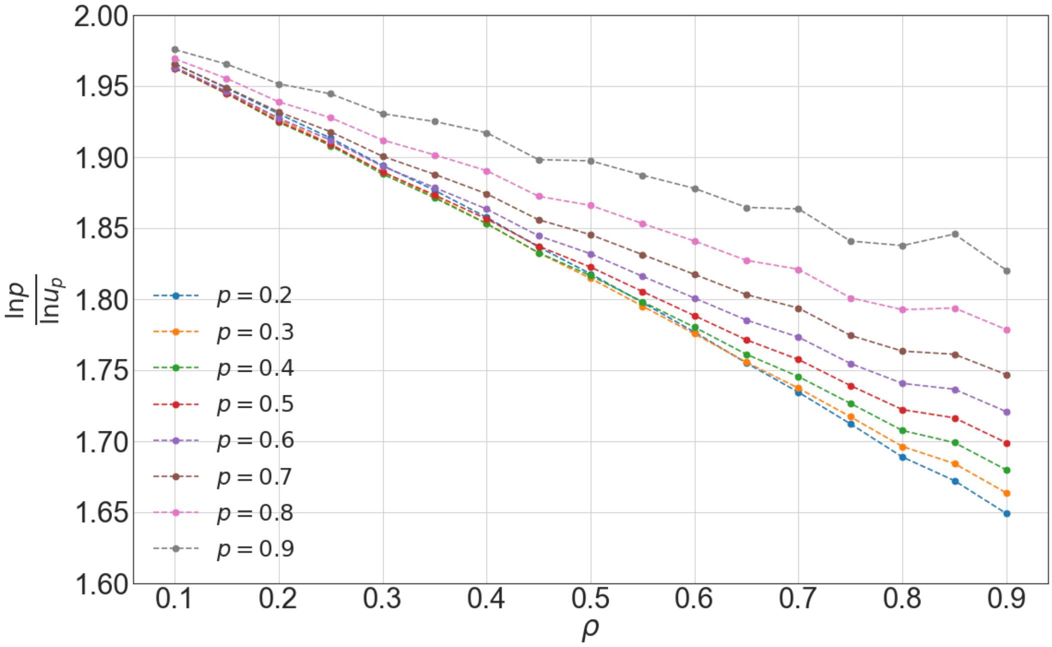


Fig. 4.4. Dependence of $(\ln p/ \ln u_p)$ on ρ.

Copyright © 2024 ASSA.

Now, to determine the type of functional dependence, we will conduct a graphical analysis of the obtained data. First of all, it was noticed that the dependence of p on u_p is well described by a power function, which corresponds to a linear dependence for logarithms (see Fig. 4.3). The dependence of the exponent α on ρ also turned out to be close to linear (see Fig. 4.4). Note that for $\rho \to 0$ the sojourn times of subtasks are asymptotically independent, hence $\alpha \to 2$. As can be seen from Figure 4.4, the dependence graph resembles a bunch of close straight lines passing through the point $(0, 2)$, so it is natural to assume (as a first approximation) that

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

and hence

$$p = \delta(u_p, \rho) \approx u_p^{2-C\cdot\rho}. \tag{4.10}$$

All that remains is to calculate the value of the coefficient C. Similar to the situation with the Kendall correlation coefficient [11], we will minimize the module of the relative approximation error relative to the simulation data using the Nelder-Mead method, as a result of which we obtain the value

$$C \approx 0.370608. \tag{4.11}$$

Thus we have

$$p = \delta(u_p, \rho) \approx u_p^{2-0.370608\cdot\rho}. \tag{4.12}$$

Figure 4.5 shows the results of simulation modeling of the probabilities or levels of p quantiles u_p of the random variable $V = F(R)$ in comparison with the results of calculations using the analytical formula (4.12) in the range $[0.20, 0.95]$ with increments of 0.05. Each point shown on the graph is actually a set of 17 points based on the number of load factor values $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$, which overlap each other and practically merge, which should be the case with a good level of approximation of the probabilities p. A slight stratification (deviation within 2%) is observed with increasing values of p. For clarity, table 4.1 shows the absolute values of the relative approximation errors for 272 calculated p values.

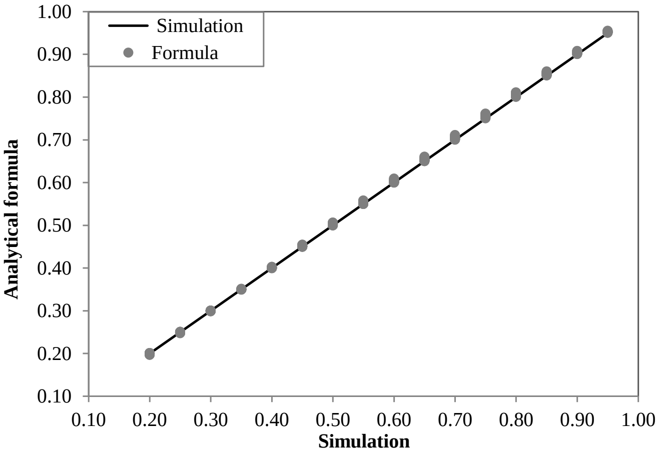


Fig. 4.5. Comparison of analytical results of the formula (4.12) with simulation of the values p of quantiles u_p of the random variable $V = F(R)$ for values $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$.

Now, taking into account (4.8), we can write

$$\delta^{-1}(p) = F(x_p), \tag{4.13}$$

Copyright © 2024 ASSA.

Table 4.1. Errors in approximations of probability values p calculated using the analytical formula (4.12) in comparison with the results of simulation modeling

Evaluated characteristic	Types of errors		
	Max APE, %	Min APE, %	MAPE, %
Probability p from the formula (4.12)	1.679144	0.002731	0.438597

moreover, from (4.10) it follows that $\delta^{-1}(p) \approx p^{\frac{1}{2-C\cdot\rho}}$. We substitute the estimate $\delta^{-1}(p)$ into (4.13) and get the relation

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

from which it follows that the quantile level p of the distribution of the random variable fork-join response time QS R is determined by expression

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-C\cdot\rho}})}{\mu - \lambda}. \tag{4.14}$$

Next, we evaluate the quality of approximation of the resulting expression on the following data set: $\rho \in \{0.10, 0.15, ..., 0.90\}, p \in \{0.20, 0.25, ..., 0.90\}$, i.e., we get a total of 272 values, for which we will estimate the approximation error using the formula 4.14. In the table 4.2 in the first row for the value $C \approx 0.370608$ the relative errors of approximation are presented, and the modulus of the maximum approximation error (Max APE) is about 3%, and the average value of the modulus of the relative error does not exceed 1%. However, if we again use the Nelder-Mead optimization method, but this time to minimize the quantile error x_p from the formula (4.14), then the result will be the value $C \approx 0.348284$. In this case, the result will improve somewhat, however, as further research shows, the value of the coefficient $C \approx 0.37$ turns out to be more justified in other aspects.

Table 4.2. Errors in approximations of the response time quantile for two variants of the values of the coefficient C in the formula (4.14)

Evaluated characteristic	Types of errors		
	Max APE, %	Min APE, %	MAPE, %
Quantile x_p, $C \approx 0.370608$	3.123971	0.002328	0.734956
Quantile x_p, $C \approx 0.348284$	2.819299	0.007276	0.699130

The figure 4.6 clearly demonstrates the quality of approximation of response time quantiles. As follows from the graphs, large errors arise for large values of the system load factor ρ, but do not exceed 3%, which is an acceptable result.

For the sake of clarifying the approximation of quantiles, returning to Fig. 4.4, we can note the dependence of the slope of the lines on p. This suggests that instead of the constant C in (4.14) we need to use expressions of the form $C_1 - C_2 p$ or $C_1 - C_2 p^2$. The selection of constants by the Nelder-Mead method and a comparative analysis of the accuracy show that the second option is better, namely, the approximation

$$x_p \approx -\frac{\ln(1 - p^{\frac{1}{2-(C_1 - C_2 p^2)\rho}})}{\mu - \lambda}, \tag{4.15}$$

where

$$C_1 \approx 0.390327, \quad C_2 \approx 0.237842,$$

moreover, the error is only 0.62%, which is 4.6 times less than before. The obtained results in more detail are presented in the table 4.3.

Copyright © 2024 ASSA.

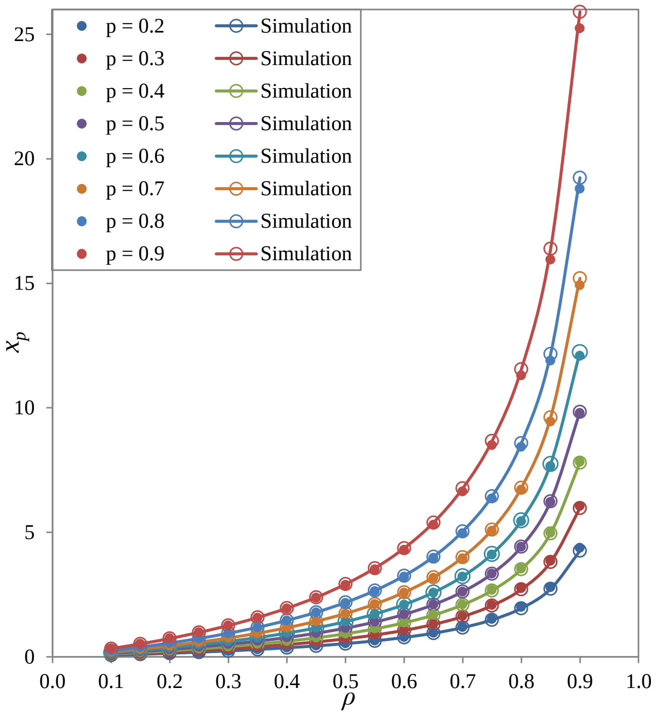


Fig. 4.6. Comparison of analytical results of the formula (4.14) with simulation modeling of quantiles x_p of the random variable of response time R.

Table 4.3. Errors in response time quantile approximations calculated by using the analytical formula (4.15) compared to simulation results

Evaluated characteristic	Types of errors		
	Max APE, %	Min APE, %	MAPE, %
Quantile x_p	0.617316	0.004912	0.304632

Unfortunately, this approach does not provide an explicit expression or convenient approximation for the diagonal section, so we will return to the formula (4.10) and move from there to copulas.

5. APPROXIMATION OF THE COPULA OF SOJOURN TIMES OF SUBTASKS BY THE GUMBEL COPULA

In the previous section, the estimate for the diagonal section of the copula $\delta(u)$ was obtained. In this section, we will present an analytical expression that evaluates the copula $C(u_1, u_2)$ itself. This will require empirical data, after analyzing which it will be possible to conclude that the copula under study is close to one of the known families.

The algorithm for constructing an empirical copula will be as follows:

1) the simulation modeling of a set of pairs (ξ_1^k, ξ_2^k) of random variables of sojourn times in subsystems $M|M|1$ fork-join QS, where k is the serial number of the simulated pair of values, $k = 1, ..., N$, N is the sample size (total number of pairs of random variables);

2) the transformation of random variables with exponential distribution $\xi_i \sim Exp(\mu - \lambda)$ by the inverse function method into random variables with uniform distribution on the

interval $[0,1]$, $U_i \sim R[0,1]$, $i=1,2$

$$(U_1^k, U_2^k) = (F(\xi_1^k), F(\xi_2^k)) = (1 - e^{-(\mu-\lambda)\xi_1^k}, 1 - e^{-(\mu-\lambda)\xi_2^k});$$

3) the division of the unit square into smaller squares (grid) with sides of length $h = 1/m$, where, for example, $m = 20$ and the determination of the number of points (U_1^k, U_2^k) falling into each of squares whose vertices are the points $(0,0)$, $(ih, 0)$, $(0, jh)$, (ih, jh), $i, j = 1, ..., m$, and normalization of the resulting value, i. e.

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

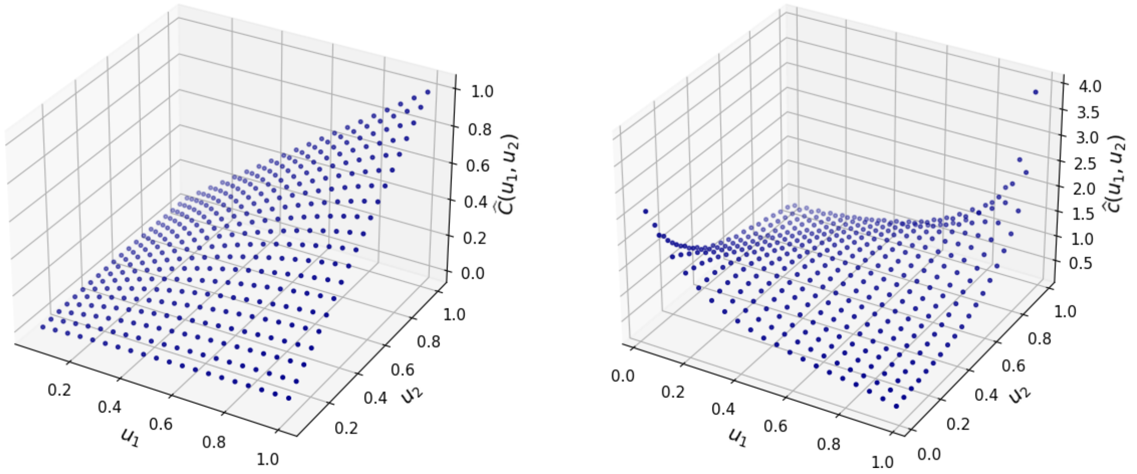
where $\mathbf{1}\{\cdot\}$ is an indicator function for the event $\{\cdot\}$.

Figure 5.7 shows a graph of an empirical copula or, what is the same, a joint distribution function of a random vector (U_1, U_2), constructed in accordance with the algorithm presented above.

We will also construct the density of the copula using the following algorithm:

1. by using the results of steps 1 and 2 from the previous algorithm, we obtain a set of pairs of random variables (U_1^k, U_2^k), $k = 1, ..., N$;
2. we divide the unit square into smaller squares (grid) with sides of length $h = 1/m$, where, for example, $m = 20$ and determine the number of points (U_1^k, U_2^k) falling into each of squares whose vertices are the points $((i-1)h, (j-1)h)$, $(ih, (j-1)h)$, $((i-1)h, jh)$, (ih, jh), $i, j = 1, ..., m$, and normalize the obtained values, i. e.

$$c_{ij} = c(ih, jh) \approx \widehat{c}_{ij} = \frac{1}{Nh^2} \sum_{k=1}^{N} \mathbf{1}\{(i-1)h < U_1^k < ih, (j-1)h < U_2^k < jh\}.$$



a) b)

Fig. 5.7. a) Empirical copula $\widehat{C}(u_1, u_2)$, $\rho = 0.9$; b) empirical copula density $\widehat{c}(u_1, u_2)$, $\rho = 0.9$.

Based on the appearance of the obtained empirical functions in Figures 5.7 a) and b), and also taking into account that the diagonal section of the copula under consideration was

Copyright © 2024 ASSA.

approximated in Section 4 by an expression of the form

$$\delta(u) \approx u^{\alpha}, \quad \alpha = 2 - C\rho, \tag{5.16}$$

we will approximate the desired copula $C(u_1, u_2)$ Gumbel copula, which has the form

$$C_g(u_1, u_2) = \exp\{-[(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{1}{\theta}}\}, \tag{5.17}$$

where $\theta \in [1, +\infty)$ is the copula parameter to be estimated. The density of a Gumbel copula is determined by taking the second-order mixed partial derivative of the copula function (5.17)

$$\begin{gathered} c_g(u_1, u_2) = \frac{\partial^2 C_g(u_1, u_2)}{\partial u_1 \partial u_2} = \\ = \frac{C_g(u_1, u_2)(-\ln u_1)^{\theta}(-\ln u_2)^{\theta}[\theta + [(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{1}{\theta}} - 1]}{u_1 u_2 \ln u_1 \ln u_2 [(-\ln u_1)^{\theta} + (-\ln u_2)^{\theta}]^{\frac{2\theta-1}{\theta}}}. \end{gathered} \tag{5.18}$$

Since for the Gumbel copula the diagonal section has the following form

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

taking into account (5.16) we obtain that

$$\theta \approx \frac{\ln 2}{\ln \alpha} = \frac{\ln 2}{\ln(2 - C\rho)}. \tag{5.19}$$

Next, we will again use the Nelder–Mead optimization method to minimize the modulus of the relative error of approximation of the Gumbel copula function (5.17), taking into account that the parameter θ is determined by the expression (5.19), when compared with "true" values of the Gumbel copula function obtained by simulation for various load factors $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$. As before, we will not consider low-level quantiles, i.e., let $u_1, u_2 \in \{0.20, 0.25, ..., 0.90\}$. As a result, we obtain the following value of the required coefficient

$$C \approx 0.369250, \tag{5.20}$$

therefore

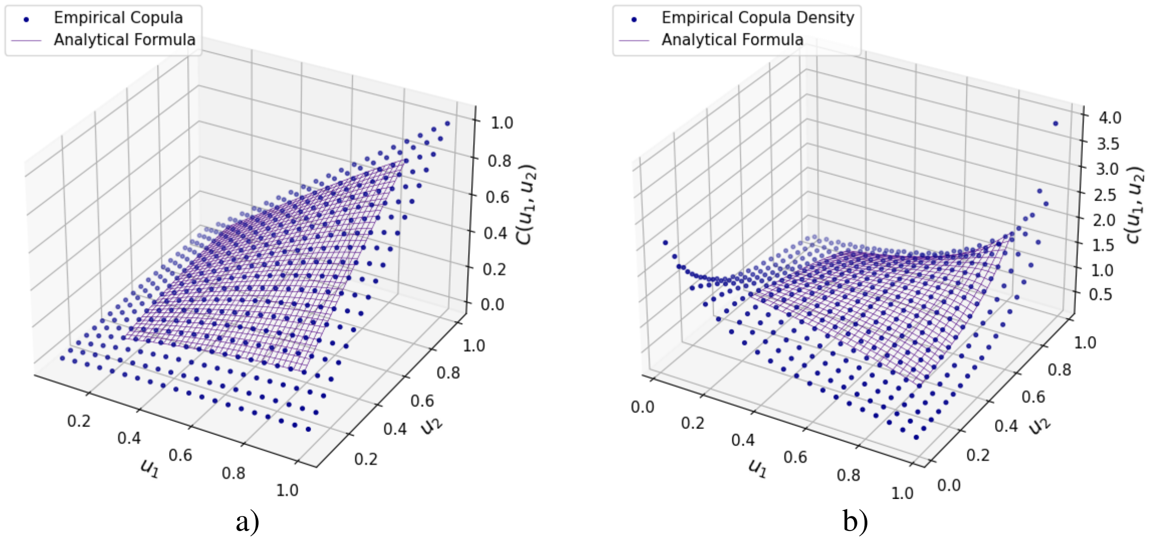
$$C(u_1, u_2) \approx \exp\{-((-\ln u_1)^{\frac{\ln 2}{\ln(2-0.36925\rho)}} + (-\ln u_2)^{\frac{\ln 2}{\ln(2-0.36925\rho)}})^{\frac{\ln(2-0.36925\rho)}{\ln 2}}\}. \tag{5.21}$$

As can be seen, from (4.11) and (5.20) the coefficient values are very close and actually agree with each other, which confirms the quality of the obtained approximations. As for the approximation error of the formula (5.21), the table 5.4 presents the values of the maximum (Max APE), minimum (Min APE) and average relative approximation error (MAPE), the first of which does not exceed 5%, on a data set of 4913 triples (ρ, u_1, u_2). Figure 5.8 a) also shows graphs of the empirical copula function and the copula defined by the expression (5.21) on a given range of values $0.2 \leqslant u_1, u_2 \leqslant 0.9$.

Table 5.4. Errors in approximation of the Gumbel copula function $C(u_1, u_2)$ by the formula (5.21) and the Gumbel copula density by the formula (5.18)

Evaluated characteristic	Types of errors		
	Max APE, %	Min APE, %	MAPE, %
$C(u_1, u_2)$	4.966424	0.000000	0.411003
$c(u_1, u_2)$	11.164042	0.000805	1.634484

Copyright © 2024 ASSA.



a) b)

Fig. 5.8. Comparison of empirical and analytical values at $\rho = 0.9$ for:
a) copula function $C(u_1, u_2)$ (formula (5.21)); b) copula distribution density $c(u_1, u_2)$ (formula (5.18)).

Note also that if we estimate the parameter θ using the classical maximum likelihood method (the corresponding Python function for the Gumbel copula), then the resulting values, the number of which in this case will correspond to the number of values of the correlation coefficient $\rho \in [0.1, 0.9]$ with a step of 0.05 (i. e. there will be only 17 of them) on the same 4913 triples of values (ρ, u_1, u_2) the Gumbel copula approximation shows large errors. In this case, Max APE $\approx 12.385819\%$, Min APE $\approx 0.000000\%$ and MAPE $\approx 1.037800\%$.

As for the copula density (5.18), here the result of comparison with empirical data (simulation modeling data) is somewhat worse, but remains acceptable. Table 5.4 presents the values of relative approximation errors for $\rho \in \{0.10, 0.15, 0.20, ..., 0.90\}$, $u_1, u_2 \in \{0.225, 0.250, ..., 0.875\}$, i. e. the total number of triplets (ρ, u_1, u_2) for which the calculation was carried out is 3332. Some narrowing of the range of values u_1, u_2 is explained by the peculiarities of the calculation of the empirical density of the copula in this particular case. Moreover, despite the fact that the maximum relative error is about 11%, in the total set of values considered, the number of relative errors exceeding the threshold of 10% is only 2. The number of approximation errors exceeding 5% is only approximately 1.95% of the total amount of data, which is confirmed by the $MAPE$ indicator, which is approximately equal to 1.63%. Note that the error in approximating the copula density increases with increasing values of u_1 and u_2, however, the resulting estimates will be upper estimates; in addition, this phenomenon can be explained by the insufficient number of tests in the region of upper quantiles and high values of the load factor. As mentioned above, increasing the accuracy of simulation estimates in the range of ρ values, as well as quantiles close to unity, requires a significant increase in the duration of the simulation [10].

6. COMPARISON OF THE GUMBEL COPULA APPROXIMATION WITH PREVIOUSLY KNOWN RESULTS

Next, let's check how much the obtained result agrees with the exact formula for the mathematical expectation of the response time obtained in [24], since it must be true

$$E[R] = \int\limits_0^{+\infty} [1 - F_R(x)]dx = \int\limits_0^{+\infty} [1 - \delta(F(x))]dx = \int\limits_0^{+\infty} [1 - \delta(1 - e^{-(\mu-\lambda)x})]dx, \quad (6.22)$$

Copyright © 2024 ASSA.

where the estimate of the diagonal section (based on the copula selected in Section 5) has the form

$$\delta(u) \approx u^{2-C\rho}, \quad C \approx 0.369250. \tag{6.23}$$

According to [24], the average response time of a fork-join QS with two subsystems $M_\lambda|M_\mu|1$ is

$$E[R] = \frac{12-\rho}{8} \cdot \frac{1}{\mu - \lambda}.$$

Taking into account the fact that in the case under consideration we assume for simplicity $\lambda = 1$ and, accordingly, $\mu = 1/\rho$, we can rewrite this expression as follows:

$$E[R] = \frac{(12-\rho)\rho}{8(1-\rho)}. \tag{6.24}$$

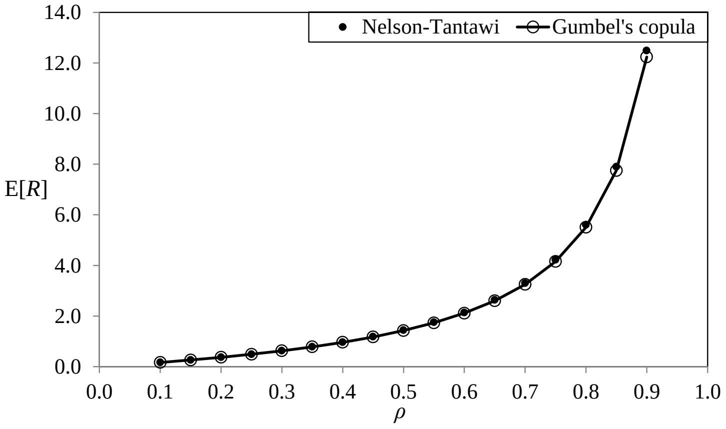


Fig. 6.9. Comparison of the results of calculating the average response time of the fork-join QS $E[R]$ using the exact formula (6.24) and using the Gumbel copula approximation (5.21) in accordance with the equality (6.27)

From [17] it is known that a random variable X with a standard generalized exponential distribution with a function and distribution density of the form

$$F_\eta(x) = (1-e^{-x})^\alpha, \quad f_\eta(x) = \alpha(1-e^{-x})^{\alpha-1}e^{-x}, \quad x \geqslant 0, \quad \alpha > 0,$$

has the mathematical expectation

$$E[X] = \psi(\alpha+1) - \psi(1) \tag{6.25}$$

and the variation

$$Var[X] = \psi'(1) - \psi'(\alpha+1), \tag{6.26}$$

where $\psi(\cdot)$ is the digamma function, which is defined as the logarithmic derivative of the gamma function [36].

In fact, based on the power-law diagonal section, we obtain an approximation for the response time distribution by a generalized exponential distribution of the general form

$$F_R(x) \approx (1-e^{-x/\beta})^\alpha, \quad x \geqslant 0, \quad \alpha, \beta > 0.$$

This approximation was previously postulated in [26], but we derived it naturally, with empirical and theoretical justification.

Due to the assumption (6.23), taking into account (6.25)), the integral (6.22) is transformed as follows

$$\begin{gathered}E[R]=\int\limits_{0}^{+\infty}[1-\delta(1-e^{-(\frac{1}{\rho}-1)x})]dx\approx\\ \approx\int\limits_{0}^{+\infty}[1-(1-e^{-(\frac{1}{\rho}-1)x})^{2-C\rho}]dx=[\psi(3-C\rho)-\psi(1)]\frac{\rho}{1-\rho}.\end{gathered}\tag{6.27}$$

In the figure 6.9 you can compare the approximations of the integral from (6.27) with the true values of the average response time (6.24) according to [24] for values of $\rho\ in\{0.10, 0.15, ..., 0.90\}$. In addition, table 6.5 shows the results of calculations of the mathematical expectation of the response time, from which it follows that the modulus of the maximum relative error of approximation by the Gumbel copula does not exceed ρ from the specified range 2.03%, which implies good consistency of the obtained approximations. Note that if instead of $C\approx 0.369250$ we take $C\approx 0.370608$ from (4.11), then the result differs very little.

Table 6.5. Error in approximation of the average response time by the Gumbel copula in accordance with the formula (6.27)

No.	ρ	$E[R]_{NT}$	$E[R]_G$	Error, $\%$
1	0.10	0.16527778	0.16503456	0.14716
2	0.15	0.26139706	0.26080338	0.22712
3	0.20	0.36875000	0.36760144	0.31148
4	0.25	0.48958333	0.48762334	0.40034
5	0.30	0.62678571	0.62369050	0.49382
6	0.35	0.78413462	0.77949216	0.59205
7	0.40	0.96666667	0.95994706	0.69513
8	0.45	1.18125000	1.17176220	0.80320
9	0.50	1.43750000	1.42432705	0.91638
10	0.55	1.74930556	1.73120372	1.03480
11	0.60	2.13750000	2.11273490	1.15860
12	0.65	2.63482143	2.60088709	1.28792
13	0.70	3.29583333	3.24893708	1.42290
14	0.75	4.21875000	4.15278226	1.56368
15	0.80	5.60000000	5.50421629	1.71042
16	0.85	7.89791667	7.75075620	1.86328
17	0.90	12.48750000	12.23495071	2.02242

Since the average response time for $K=2$ is known exactly, the example discussed is illustrative, but this approach may be useful for $K>2$.

Let us now apply the copula method to estimate the Kendall correlation coefficient, which, unlike the Pearson and Spearman coefficients, is not calculated exactly, and was previously estimated by the formula (2.4). According to [23, p. 164], for the Gumbel copula (5.17) the Kendall correlation coefficient is

$$r_k=1-\frac{1}{\theta},$$

from which, taking into account the formula (5.19), we obtain the approximation

$$r_k\approx 1-\frac{\ln(2-C\rho)}{\ln 2}.\tag{6.28}$$

Copyright © 2024 ASSA.

Let us analyze the quality of the approximation taking into account the simulation results from [11].

Unfortunately, at $C = 0.37$ (which corresponds to the values obtained when estimating the diagonal section and copula), the formula gives overestimated values with an error from 5.19 to 7.46%. However, if we optimize (6.28) over C using the Nelder-Mead method based on simulation results (as we did earlier), then at the optimal value $C \approx 0.349237$ (which is close to the value of $C \approx 0.348284$ obtained by estimating response time quantiles) we obtain an error of only 1.04%, which is close in quality to the empirical approximation by the quadratic function (2.4), previously obtained by the authors in [11]. The latter option therefore represents an alternative approximation of the Kendall correlation coefficient.

Thus, copula approximations should not always be taken literally. They may suggest convenient analytical expressions for some characteristics, while the parameters of these expressions may require refinement through additional optimization based on actual data.

Let us also note the interesting fact that the formula (6.28) allows us to re-estimate the limiting value of the Kendall correlation coefficient at $\rho \to 1$. In [11], based on the (2.4) approximation, an estimate of $r_k \approx 0.276$ was found, now we obtain a value that coincides with the previous one with an accuracy of three digits, which indicates a good correspondence.

As it turned out during further research, the resulting approximation (copula and diagonal section), unfortunately, does not work for estimating the variation (and therefore the standard deviation). Taking into account (6.26), the variation of the normalized response time should decrease with increasing load, just like the mathematical expectation, but in fact it increases, according to the simulation results [10, 14], and no $C > 0$ here fits. This once again shows that the same approximation can be good for some purposes and bad for others, so you should be careful with this in applications.

7. CONCLUSION

This work continues the author's series devoted to the study of the characteristics of fork-join systems with a Poisson input flow and exponential service times. Despite the simplicity of the systems and the long history of their research (since the 1980s), there is still a lot of uncertainty in this area. There are few accurate results here, and many estimates need improvement. There are issues that few or no one has dealt with. Research mainly focuses on average response time, while variance, quantiles, etc. are also of interest.

The key problem is the presence of a relationship between the sojourn times of subtasks (parts of one task), due to the commonality of the input flow into the subsystems. This dependence, although not very strong, has a significant impact on the characteristics, and it is far from being described by classical models (for example, multivariate normal distribution, linear regression, etc.). Therefore, the authors in recent works have focused on studying this dependence. The case of two subsystems was considered on the basis that for any number of subsystems, for any pair of subsystems, the sojourn times of subtasks will have the same joint distribution. Previous work found the exact values of the Pearson and Spearman correlation coefficients, as well as an estimate of the Kendall correlation coefficient. In this paper, approximations of the joint distribution of subtask sojourn times were studied using copula theory. Good agreement with the data for power-law diagonal sections and the Gumbel copula is obtained. Based on the estimates of diagonal sections, estimates of response time quantiles are derived over a wide range of levels and loads. A new estimate of the Kendall correlation coefficient was also obtained.

The developed approach, based on copula theory, can be attempted to be generalized to systems with a large number of subsystems or the case of more complex subsystems (for example, with heavy tails of service time distributions).

REFERENCES

1. Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., et al. (2015). On patient flow in hospitals: a data-based queueing-science perspective, *Stochastic Systems*, **5**, 146–194. https://doi.org/10.1287/14-SSY153
2. Atar, R., Mandelbaum, A. & Zviran, A. (2012). Control of fork-join networks in heavy traffic, *Proc. of 2012 50th Annual Allerton Conference on Communication, Control, and Computing* (Allerton, IL), 823–830. https://doi.org/10.1109/Allerton.2012.6483303
3. Baccelli, F. & Makowski, A. M. (1989). Queueing models for systems with synchronization constraints, in *Proceedings of the IEEE*, **77**, 138–161. https://doi.org/10.1109/5.21076
4. Baron, O., Krass, D., Sherzer, E. & Senderovich, A. (2022). Can machines solve general queueing problems?, *Proc. of 2022 Winter Simulation Conference (WSC)* (Singapore), 2830–2841. https://doi.org/10.1109/WSC57314.2022.10015451
5. Chocron, E.., Cohen, I. & Feigin, P. (2022). Delay prediction for managing multiclass service systems: an investigation of queueing theory and machine learning approaches, *IEEE Transactions on Engineering Management*, **71**, 1–11. https://doi.org/10.1109/TEM.2022.3222094
6. Dieleman, A., Berkhout, J. & Heidergott, B. (2023). A neural network approach to performance analysis of tandem lines: The value of analytical knowledge, *Computers & Operations Research*, **152**, 106124. https://doi.org/10.1016/j.cor.2022.106124
7. Efrosinin, D. & Stepanova, N. (2021). Estimation of the optimal threshold policy in a queue with heterogeneous servers using a heuristic solution and artificial neural networks, *Mathematics*, **9**, 126. https://doi.org/10.3390/math9111267
8. Enganti, P., Rosenkrantz, P., Sun, L., Wang, Z., Che, H., et al. (2022). ForkMV: Mean-and-Variance Estimation of Fork-Join Queuing Networks for Datacenter Applications, *Proc. of IEEE International Conference on Networking, Architecture and Storage (NAS)* (Philadelphia, PA), 1–8. https://doi.org/10.1109/NAS55553.2022.9925531
9. Gallien, J. & Wein, L. M. (2001). A simple and effective component procurement policy for stochastic assembly systems, *Queueing Systems*, **38**, 221–248. https://doi.org/10.1023/A:1010914600116
10. Gorbunova, A. V. & Lebedev, A. V. (2023). On estimating the characteristics of a fork-join queueing system with Poisson input and exponential service times, *Advances in Systems Science and Applications*, **23**, 99–114. https://doi.org/10.25728/assa.2023.23.2.1351
11. Gorbunova, A. V. & Lebedev, A. V. Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with $M|M|1$-type Subsystems, [under consideration]
12. 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**, 409–428. https://doi.org/10.1016/j.matcom.2023.07.029
13. Gorbunova, A. V. & Lebedev, A. V. (2020). Bivariate distributions of maximum remaining service times in fork-join infinite-server queues, *Problems of Information Transmission*, **56**, 73–90. https://doi.org/10.1134/S003294602001007X
14. Gorbunova, A. V. & Vishnevsky, V. M. (2020). Estimating the response time of a cloud computing system with the help of neural networks, *Advances in Systems Science and Applications*, **20**, 105–112. https://doi.org/10.25728/assa.2020.20.3.926
15. Gorbunova, A. V., Zaryadov, I. S., Matyushenko, S. I., Samouylov, K. E. & Shorgin, S. Ya. (2015). An approximation of response time of a cloud computing system, *Informatics and Applications*, **9**, 32–38. https://doi.org/10.14357/19922264150304
16. Gudendorf, G. & Segers, J. (2010). Extreme-value copulas, In Jaworski, P., Durante, F., Härdle, W. K. & Rychlik, T. (Eds.) *Copula theory and Its Application* (pp. 127–145). Berlin, Germany: Springer.

17. Gupta, R. D. & Kundu, D. (1999). Generalized exponential distributions, *Australian New Zealand J. Statist*, **41**, 173–188. https://doi.org/10.1080/00949650108812098
18. Jiang, L. & Giachetti, R. E. (2008). A queueing network model to analyze the impact of parallelization of care on patient cycle time, *Health Care Management Science*, **11**, 248–261. https://doi.org/10.1007/s10729-007-9040-9
19. Kyritsis, A. I. & Deriaz, A. I. (2019). A machine learning approach to waiting time prediction in queueing scenarios, *Proc. of 2nd International Conference on Artificial Intelligence for Industries (AI4I)* (Laguna Hills, CA), 17–21. https://doi.org/10.1109/AI4I46381.2019.00013
20. Lebedev, A. V. (2019). Upper bound for the expected minimum of dependent random variables with known Kendall's tau, *Theory of Probability and Its Applications*, **64**, 465–473. https://doi.org/10.1137/S0040585X97T989623
21. Lebedev, A. V. (2019). On the interrelation between dependence coefficients of bivariate extreme value copulas, *Markov Processes and Related Fields*, **25**, 639–648.
22. Narahari, Y. & Sundarrajan, P. (1995). Performability Analysis of Fork-Join Queueing Systems, *The Journal of the Operational Research Society*, **46**, 1237–1249. https://doi.org/10.2307/2584619
23. Nelsen, R. (2006). *An introduction to copulas*. Berlin, Germany: Springer.
24. Nelson, R. & Tantawi, A. N. (1988). Approximate analysis of fork/join synchronization in parallel queues, *IEEE Transactions on Computers*, **37**, 739–743. https://doi.org/10.1109/12.2213
25. Nguyen, M., Alesawi, S., Li, S., Che, H. & Jiang, H. (2018) ForkTail: A black-box fork-join tail latency prediction model for user-facing datacenter workloads, in *Proc. of 27th Int. Symp. High-Perform. Parallel Distrib. Comput.* (Tempe, AZ) 206–217. https://doi.org/10.1145/3208040.3208058
26. 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**, 1983–2000. https://doi.org/10.1109/TPDS.2020.2982137
27. Palomo, S. & Pender, J. (2021). Learning the tandem network Lindley recursion, *Proc. of 2021 Winter Simulation Conference (WSC)* (Phoenix, AZ), 1–12. https://doi.org/10.1109/WSC52266.2021.971553
28. Pender, J. & Zhang, E. (2021). Uniting simulation and machine learning for response time prediction in processor sharing queues, *Proc. of 2021 Winter Simulation Conference (WSC)* (Phoenix, AZ), 1–12 https://doi.org/10.1109/WSC52266.2021.9715461
29. 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**, 99–116. https://doi.org/10.1016/j.peva.2015.06.007
30. Reinsel, D., Gantz, J. & Rydning, J. (2018). *IDC Report: The Digitization of the World From Edge to Core*, [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
31. Rizk, A., Poloczek, F. & Ciucu, F. (2016). Stochastic bounds in Fork-Join queueing systems under full and partial mapping, *Queueing Systems*, **83**, 261–291. https://doi.org/10.1007/s11134-016-9486-x
32. Schol, D., Vlasiou, M. & Zwart, M. (2022). Large Fork-Join Queues with Nearly Deterministic Arrival and Service Times, *Mathematics of Operations Research*, **47**, 1335–1364. https://doi.org/10.1287/moor.2021.1171
33. Thapa, S. & Zhao, Y. Q. (2021) Construction of new copulas with queueing application. *arXiv: 2101.12401*, [Online]. Available: https://arxiv.org/abs/2101.12401
34. Vianna, E., Comarela, E., Pontes, T., Almeida, J., Almeida, V., et al. (2013). Analytical performance models for MapReduce workloads, *International Journal of Parallel Programming*, **41**(4), 495–525. https://doi.org/10.1007/s10766-012-0227-4

Copyright © 2024 ASSA.

35. Vishnevsky, V. M. & Gorbunova, A. V. (2022). Application of machine learning methods to solving problems of queuing theory, *Communications in Computer and Information Science*, **1605**, 304–316. https://doi.org/10.1007/978-3-031-09331-9_24
36. Weisstein, E. W. (2024) *Digamma Function – from Wolfram MathWorld*, [Online]. Available: https://mathworld.wolfram.com/DigammaFunction.html

Copyright © 2024 ASSA.