ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 12, pp. 1184–1198.

STOCHASTIC SYSTEMS

On the Features of Service Rate Control in Fork-Join Queueing System

A. V. Gorbunova[*,a] **and A. V. Lebedev**[**,b]

[*] *Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia*
[**] *Lomonosov Moscow State University, Moscow, Russia*
e-mail: [a] *avgorbunova@list.ru,* [b] *avlebed@yandex.ru*

Received October 7, 2024
Revised October 16, 2024
Accepted October 18, 2024

Abstract—A classical fork-join queueing system is considered. The model is proposed for determining the optimal cost of such a system, taking into account the need to minimize the average response time simultaneously with reasonable costs for the resources required for this. The term "resources" within the framework of the mathematical model under study means the intensities of service on servers, the cost of which is directly proportional to the system performance, i.e., the rate of service requests. For the special case when the number of subsystems is equal to two, an exact analytical expression for determining the optimal cost is presented, for a more general case when the number of subsystems of a fork-join queueing system is greater than two, the equation is obtained, the numerical solution of which allows calculating the desired value. In addition, the asymptotic analysis of the obtained solutions behavior is carried out.

Keywords: fork-join queueing system, queueing system, optimal cost, control

DOI: 10.31857/S0005117924120043

1. INTRODUCTION

A classical fork-join queueing system (QS) is considered. Upon receipt in this system, a task is divided into K identical parts (subtasks), the number of which corresponds to the number of subsystems. Each subsystem is a system with an infinite storage and a single server. It is assumed that the service rates on all available servers are identical. A task is considered to be serviced after all its constituent parts have been serviced. Accordingly, the response time of the system (the time a task stays in the system) is determined by the maximum of K random times of subtasks staying in the subsystems.

Such systems are widely used to model various types of processes in which the task is divided or parallelized, in particular in the field of information technology we can talk about parallel or distributed computing, and they can also be used to model various work processes in manufacturing (for example, assembling a multi-component order in a warehouse or complex mechanisms consisting of many parts in production), finance (for example, processing a loan application in several divisions of a financial institution), healthcare (carrying out the necessary tests and collecting anamnesis upon admission of a patient to a medical institution), etc.

The key feature of this system, which complicates its analysis, is the existence of a dependence between the times of stay of subtasks in subsystems. Therefore, despite the relevance and demand for studying fork-join systems, the exact characteristics of the system's functioning were obtained only for the case of two subsystems ($K = 2$), in particular, the formula is known for the average response time of the system in the case of a Poisson input flow and exponential distributions of service times [1]. In other cases, only approximate expressions were found to approximate the main

indicators of system performance [1–3]. The review [4] provides a more detailed overview of the known results. As for more recent studies, the works [5–12], including those by the authors of this article, in addition to formulas that refine the known estimates of the average response time or its variance present expressions for estimating characteristics such as quantiles of the response time distribution for a wider range of input flows or of service times. In addition, in [11] the exact expression for the correlation coefficient between the sojourn times of subtasks in subsystems of the $M|M|1$ type is presented, and in [9] for a fork-join system with a Pareto distribution of service time the estimate for the correlation coefficient was derived already in the case of a Pareto distribution of service time. It is also worth noting the Russian-language works [13–19]. In the series of papers, including [13–16], the analysis of a fork-join system with the infinite-server subsystems was performed in terms of generating functions for the probability distribution of the number of subtasks in each subsystem. In [17], one of the modifications of the fork-join system was analyzed, when, upon receipt, an task is divided not into a fixed, but into a variable number of subtasks, which is determined by the state of the system. In [18], an approach based on relation invariants was proposed for approximating the average response time of the fork-join QS, and in [19] the fork-join system is an integral part of the network and is used to model and study the performance characteristics of transaction service platforms.

This article examines another aspect of fork-join system performance, namely, the model of the cost of operating a system with division and parallel service is constructed, which allows determining optimal control in terms of optimizing its financial indicators. The model is based on natural assumptions about the need to minimize the average response time of the system to maintain its competitiveness at reasonable costs for the resources required for this. In particular, resources can be understood as the capacity of the necessary equipment that allows faster processing of a client request, if we are talking about information and computing or production systems, for example. It is clear that the more powerful the equipment, the more costs are required for its purchase, technical maintenance and maintenance in general. Thus, the rate of equipment operation (or in terms of QS, the rate of service) is proportional to the growth of its cost. In addition, as the rate of service increases, the response time of the system decreases. Thus, the cost of system operation consists of the optimal balance between the response time and the rate of the service servers.

In more detail, it is assumed that 1) the price (penalty) per unit of average response time and 2) the price per unit of service rate are set. In this case, the first price is assumed to be equal to one for simplicity. Then the response time and service costs are calculated and added to the total cost of expenses, which we want to minimize here. Thus, the task of cost optimization of management is set.

Similar problem statements can be found in the monograph [20], devoted to the optimal design of QS, including the optimal choice of arrival and service rates of tasks (assuming that these parameters are controllable) for various systems and queueing networks. However, for fork-join systems such problems have not been considered before.

The proposed functional dependence for the cost of operating a fork-join queueing system allows one to obtain an explicit expression for determining the optimal service rate on the system's servers in the case when the number of subsystems is two. If the number of subsystems is greater than two, then obtaining the optimal solution is possible in numerical form. The behavior of the optimal solution in extreme cases was also considered.

The paper is organized as follows: Section 2 describes a mathematical model for determining the optimal cost of operating a fork-join system in general; the next two sections derive the optimal value of the system load factor, which allows to express the optimal value of service intensity for the case of the Nelson–Tantawi formula for determining the system response time — for a particular case, i.e., when the number of subsystems corresponds to two and the formula is exact, and for

a more general case when the formula is approximate; then the case of a generalization of the Nelson–Tantawi formula is analyzed, and then an asymptotic analysis of the obtained solutions is carried out, both in the general case and for specific expressions; in the conclusion, some results are summarized.

2. MATHEMATICAL MODEL FOR DETERMINING THE OPTIMAL COST OF FUNCTIONING OF A FORK-JOIN SYSTEM

We analyze a fork-join system with a Poisson input flow with intensity $\lambda > 0$ and an exponential distribution of service times on $K \geqslant 2$ homogeneous servers with intensity $\mu > 0$. The system load factor is $\rho = \lambda/\mu < 1$.

Let us present a unified mathematical approach that will be further applied in the work.

Let us denote the cost of the system operation by S and introduce the function $f(\rho)$, which defines the expression for the average response time of the system in the case of $\lambda = 1$, i.e., $f(\rho) = E[R_K]$ when $\lambda = 1$. Then, in the general case, the following expression will be valid

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

Since the cost of operating the system S depends on the average response time of the system (we take the price per unit of time as one unit) and the cost of maintenance costs, we can write for it

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

where c is the cost of a unit of service intensity. Accordingly, taking into account the introduced function $f(\rho)$, we can rewrite this expression as follows:

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

Let $c_1 = c\lambda^2$, then

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

Next, in order to find the optimal value of the system load factor we determine the extremum point of the cost function $S(\rho)$, namely the minimum point. Fo this purpose we find the derivative of the last expression and equate it to zero

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

as a result we obtain the equation

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

and having solved the equation (2) we find the optimal value of ρ_0 and, accordingly, the optimal value of the service intensity $\mu_0 = \lambda/\rho_0$.

Let us consider the basic case when $K = 1$, i.e., in fact, the $M|M|1$ system. The average response time in such a system is equal to

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

Then the function $f(\rho)$, by taking into account that $\rho = \lambda/\mu$, is defined as

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

and the derivative of this function has the form

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

Now we substitute the obtained expressions into the equation (2)

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

and obtain the value for the desired optimal system load at $K = 1$

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

In the next step we can calculate the optimal service rate, namely

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

The similar problem was solved in [20, §1.1] for the case where the cost takes into account not the average response time, but the average waiting time.

Let us emphasize that since the equation (2) was obtained with respect to the load ρ, then in the future we will only look for the optimal load, understanding that it can, if necessary, be recalculated into the optimal service intensity.

3. ANALYSIS OF FORK-JOIN QS WITH TWO SUBSYSTEMS OF TYPE $M|M|1$. NELSON–TANTAWI FORMULA

To begin with, let us consider a special case and determine the optimal value of the system load factor when the number of subsystems is two. For the average response time at $K = 2$ the exact formula obtained in [1] is known, namely

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

Thus, we have

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

Next, we find the derivative with respect to ρ and, substituting the expression for $f'(\rho)$ into (2), we can write the following equation:

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

For convenience, we will make the change

$$c_2 = 8c_1$$

and after simplification we obtain a fourth-degree equation

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

which we will solve, since, as is known, for a fourth-degree equation there is an analytical solution in radicals. For this, we will use the Ferrari method [21–23].

Let us introduce the following notations:

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

then we get

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

By changing $\rho = y - A/4$ we reduce the fourth-degree equation (4) to the canonical form

$$y^4 + A_1 y^2 + B_1 y + C_1 = 0, \tag{5}$$

where

$$A_1 = B - \frac{3A^2}{8} = -c_2 + \frac{21}{2},$$
$$B_1 = \frac{A^3}{8} - \frac{AB}{2} + C = c_2 + 11,$$
$$C_1 = -\frac{3A^4}{256} + \frac{A^2 B}{16} - \frac{AC}{4} + D = -\frac{1}{4}c_2 + \frac{45}{16}.$$

Next, according to the Ferrari method, it is necessary to find one particular real solution of the cubic equation

$$A_2 t^3 + B_2 t^2 + C_2 t + D_2 = 0, \tag{6}$$

where

$$A_2 = 2, \quad B_2 = -A_1 = c_2 - \frac{21}{2},$$
$$C_2 = -2C_1 = \frac{1}{2}c_2 - \frac{45}{8}, \quad D_2 = A_1 C_1 - \frac{B_1^2}{4} = -\frac{175}{16}c_2 - \frac{23}{32}.$$

By changing $t = z - B_2/(3A_2)$ we reduce the equation (6) to the canonical form of a third-degree equation

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

where

$$A_3 = \frac{3A_2 C_2 - B_2^2}{3A_2^2} = -\frac{1}{12}c_2^2 + 2c_2 - 12,$$
$$B_3 = \frac{2B_2^3 - 9A_2 B_2 C_2 + 27A_2^2 D_2}{27A_2^3} = \frac{1}{108}c_2^3 - \frac{1}{3}c_2^2 - \frac{3}{2}c_2 - 16.$$

The real root of the equation (7) according to Cardano's method [21–23] is determined as follows

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

where

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

At the same time, we must not forget that the following condition must be met

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

Note that for the case $Q < 0$ for z_1 after the transformations the following formula will ultimately be valid:

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

where

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

and for the case $Q = 0$ we have

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

The expression for Q from (9) is transformed to the form

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

and changes its sign depending on the value of c_2 (although c_2 should be positive based on its physical meaning), so we will explicitly indicate the regions of sign constancy of the expression for Q, the zeros of which can again be obtained using Cardano's method, namely

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

where $\tilde{c}_2 = (3 \times 11^{\frac{1}{3}} + 3 \times 11^{\frac{2}{3}} + 12) \approx 33.51$.

Accordingly, we get

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

where z_1, depending on the value of Q, is determined by the expressions (8), (10) or (11). The particular solution t_1 allows us to represent the canonical equation of the fourth degree (5) as a product of two square trinomials

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

The solution of one of the quadratic equations (taking into account the inverse substitution) will be the desired solution of the original equation of the fourth degree (4). In particular, the check showed that the desired one is the root of the second equation, so we can write the final solution as

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

where the discriminant of the second equation is equal to

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

and, accordingly,

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

4. ANALYSIS OF FORK-JOIN QS WITH $K > 2$ SUBSYSTEMS OF TYPE $M|M|1$. NELSON–TANTAWI FORMULA

Let us derive the equation for determining the optimal value of the system load factor in the general case when $K > 2$. For the mathematical expectation of the response time of a fork-join system, the Nelson–Tantawi approximate formula, which is considered one of the most accurate among those known, has the form [1]

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

where $H_K = \sum_{i=0}^{K} 1/i$ is the partial sum of the harmonic series.

We will solve the problem for the Nelson–Tantawi approximation under the assumption that the formula (12) is exact.

Let us introduce the following notations:

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

Then

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

therefore,

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

and for $f'(\rho)$ after transformations we obtain

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

Then we substitute the resulting expression into (2)

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

Also, for convenience, we introduce the notation $c_2 = 8c_1$ and obtain a fifth-degree equation, which, after transformations taking into account that $M = 4(1-H)/11$, is reduced to the form

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

The resulting equation (14) can be solved numerically and the optimal value of ρ_0 and, accordingly, the desired value of μ_0 can be determined.

Figure 1 presents graphs of the dependence of the behavior of the optimal value of the load factor $\rho = \rho_0$, which is the solution of the equation (14), on the value of the parameter c_1 for a different number of subsystems K of the fork-join QS. At the initial stage, there is a fairly rapid growth of the optimal load with an increase in the cost of a unit of resource and, accordingly, the performance of the system as a whole. Moreover, even for the number of subsystems $K = 1000$, i.e., in other words, given the division of the task and the processing of its subtasks on a fairly large number of servers, already at $c_1 > 5.5$ the required load level will be $\rho > 0.5$. At the same time, with further growth of the parameter c_1, a rather slow growth of the optimal load is observed, for example, with $c_1 \approx 158$, which is almost 29 times greater than $c_1 = 5.5$, the value $\rho_0 = 0.85$ will be, i.e., still does not exceed 90%. In addition, based on the type of graphs, the effect is observed

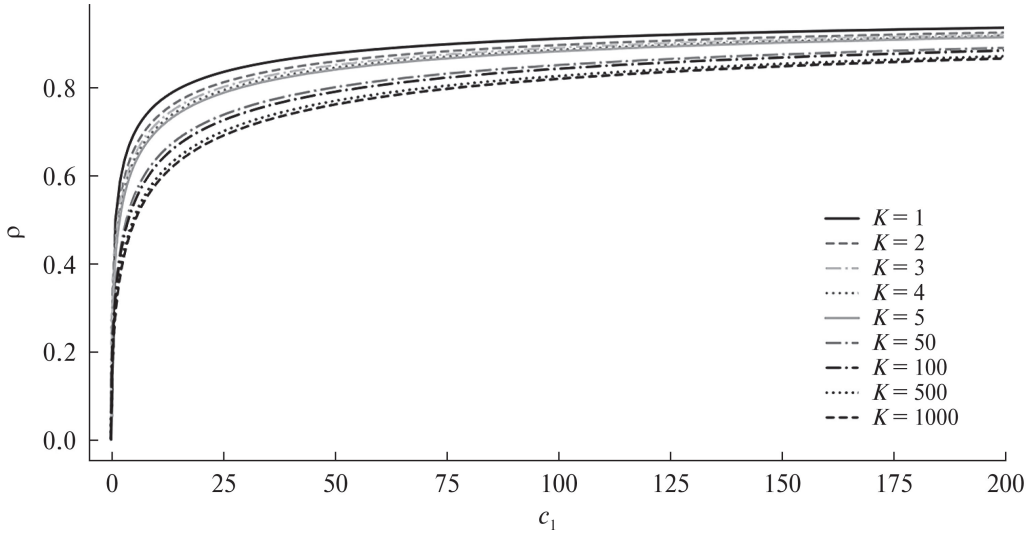

Fig. 1. Graph of the dependence of the optimal system load value ρ (solution of the equation (14)) on the parameter c_1 for a different number of subsystems K.

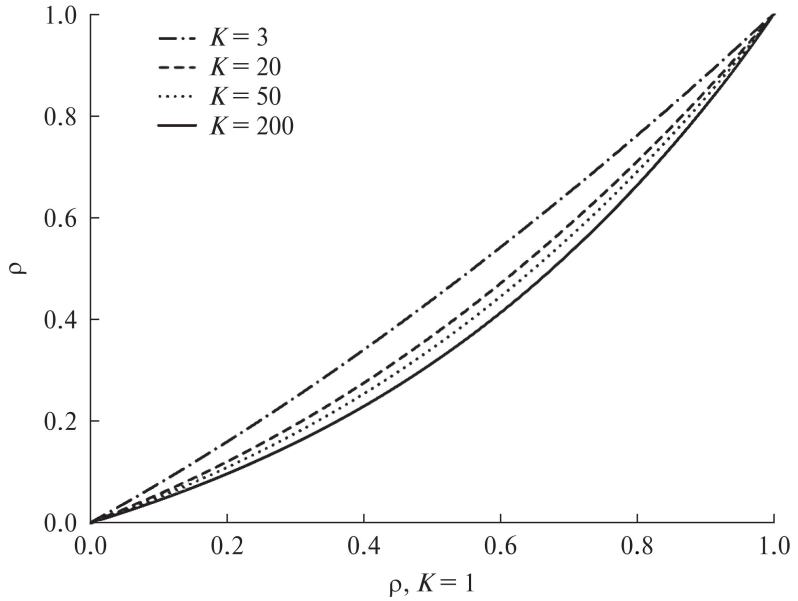

Fig. 2. Graph of the dependence of the optimal system load value ρ on the optimal value of the load ρ at $K = 1$.

that with the growth of K, the optimal level of load on the system decreases, which was expected and quite natural.

Figure 2 presents the graph of the dependence of the behavior of the optimal value of the system load factor on the optimal value of ρ for $K = 1$ according to (3). The graph allows us to compare the level of optimal load for different values of $K \geqslant 2$ with the level of optimal load in the case of $K = 1$. As can be seen, with an increase in the number of subsystems, the level of load required for optimal system operation decreases, i.e., the line bends more and more under the straight line $\rho = \rho_{K=1}$ and becomes more and more convex (below any chord connecting any two points on the graph selected on the interval under consideration).

5. ANALYSIS OF FORK-JOIN QS WITH $K > 2$ SUBSYSTEMS OF TYPE $M|M|1$. GENERALIZATION OF NELSON–TANTAWI FORMULA

Consider a generalization of the Nelson–Tantawi formula from [10], in which it was shown that the refined formula gives a better approximation for the average response time. The improvement is achieved by making some correction to the expression from (12), which we denote by $E[R_K]_{NT}$. Thus, the approximation is

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

where

$$Q_1 \approx 0.087197, \quad Q_2 \approx 0.070236, \quad Q_3 \approx 0.09638.$$

In order to define an expression for $f(\rho)$ for this case, we take $1/\lambda$ out of the brackets

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

and as the result is

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

where $\lambda E[R_K]_{NT}$ was calculated earlier and has the form

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

Thus, taking into account the previously proposed replacement $H = H_K/H_2$, we have

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

Next we take the derivative of the obtained expression

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

where

$$f_1'(\rho) = \frac{\rho(1-H)}{(1-\rho)^2}\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big),$$
$$f_2'(\rho) = \frac{1}{8(1-\rho)^2}\Big(2M\rho^3 + (H - 15M)\rho^2 + (24M - 2H)\rho + 12H\Big).$$

Accordingly, after substituting the obtained expressions into (2) we obtain that

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

The equation (16), as in the case of the Nelson–Tantawi formula from the previous section, is a fifth-degree equation that can be solved numerically, after which the optimal value of μ_0 will be found.

Figure 3 presents the graphs of the optimal load value $\rho = \rho_0$, which is the solution of (16). The behavior of the graphs is generally similar to the case of the Nelson–Tantawi formula (12). In order

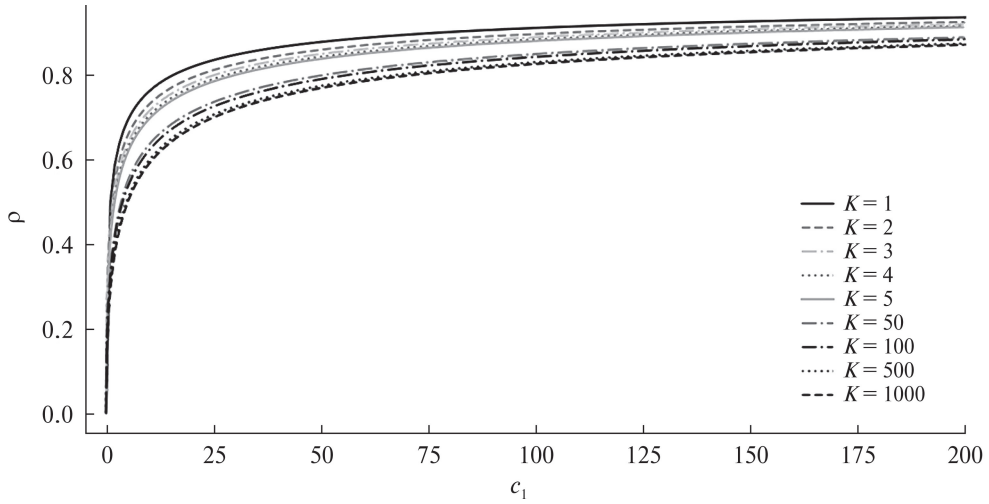

Fig. 3. Graph of the dependence of the optimal system load value ρ (solution of the equation (16)) on the parameter c_1 for a different number of subsystems K.

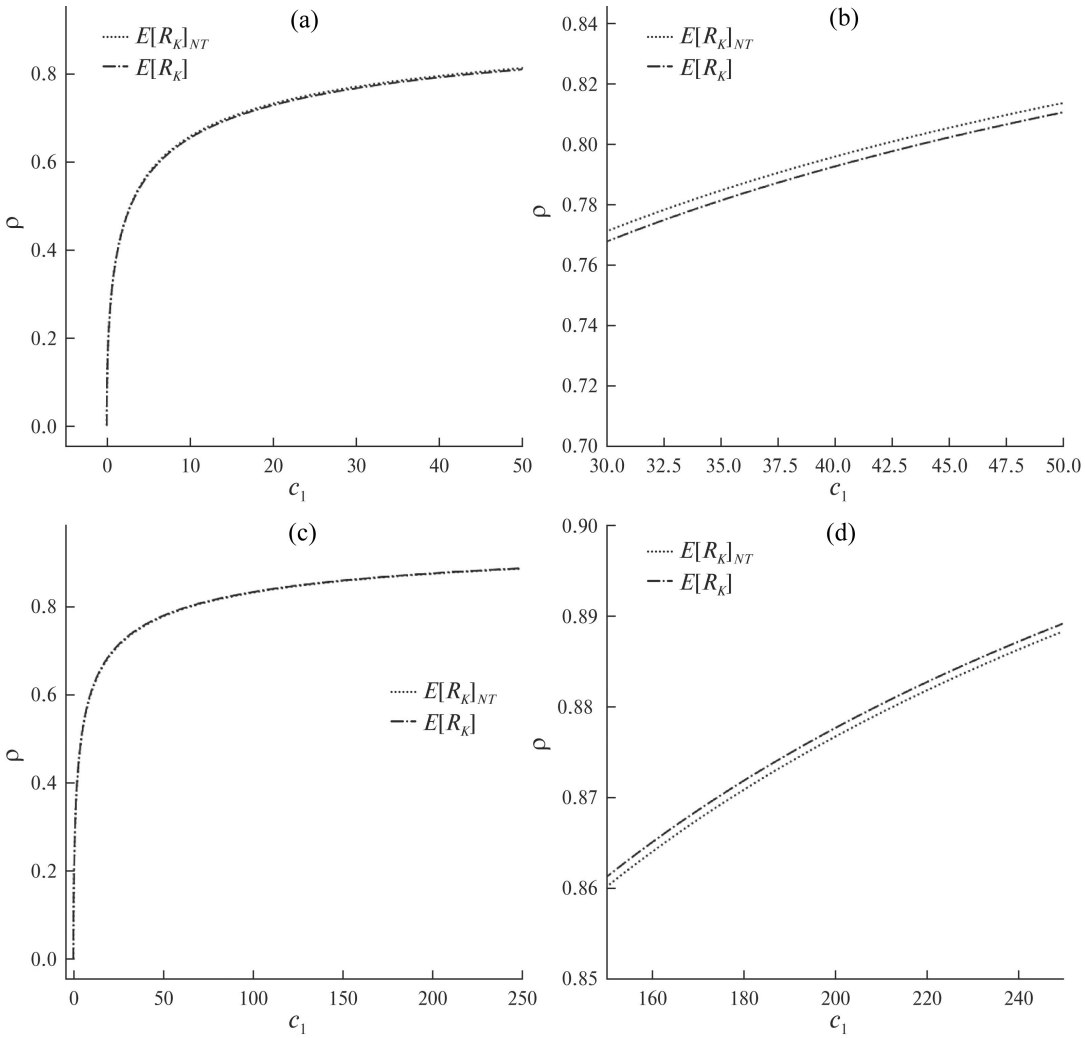

Fig. 4. Graphs of the optimal load value $\rho = \rho_0$, which is the solution to the equation (16) for $K = 20$ (Figs. 4a and 4b) and $K = 200$ (Figs. 4c and 4d).

to analyze the difference between the obtained results in more detail, we will compare the behavior of the optimal solution for the special cases of the number of subsystems $K = 20$ and $K = 200$.

Figure 4 presents the graphs of the dependence of the optimal system load level $\rho = \rho_0$ on the parameter c_1 for the cases of the Nelson–Tantawi formula for the average system response time $E[R_K]_{NT}$ from (12) and the equation (14), as well as a generalization of the Nelson–Tantawi formula for the average system response time $E[R_K]$ from (15) and the equation (16) for $K = 20$ (Figs. 4a and 4b) and $K = 200$ (Figs. 4c and 4d), including the scaled version (Figs. 4b and 4d).

In Figs. 4a and 4b for $K = 20$ it is evident that for the case of the Nelson–Tantawi formula (12) for estimating the average system response time, the optimal load value ρ exceeds the optimal value calculated by the generalized formula (15). However, for graphs 4c and 4d for $K = 200$ the opposite situation can be observed.

In order to thoroughly understand the behavior of optimal solutions in both cases, we will further analyze their asymptotics.

6. ASYMPTOTICS OF THE BEHAVIOR OF THE OPTIMAL SOLUTION

Let us consider the equation (2) and determine the behavior of its solution as $c_1 \to 0$ ($\rho \to 0$) and $c_1 \to +\infty$ ($\rho \to 1$) tend in general. First, let us analyze the case of $\rho \to 0$, respectively $c_1 \to 0$. Then we have

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

then we substitute the obtained result into (2), i.e., $\rho^2 f'(\rho) = c_1$, whence it follows that

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

Now let's analyze the case $\rho \to 1$, respectively $c_1 \to +\infty$. In the general case we have

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

Thus, if there exists a number $L \in (0, +\infty)$ such that

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

then

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

therefore, when substituting the obtained expression for $f'(\rho)$ c L into (2) and taking into account the asymptotics, we have the following

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

therefore

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

Next, we will consider the existing models in detail, namely the cases of the Nelson–Tantawi formula and its generalization, and we will define specific expressions for the obtained equivalences from the general case.

For the Nelson–Tantawi formula for $K \geqslant 2$, taking into account the expression (13), it is true

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

therefore, when substituting into (17) for $c_1 \to 0$ ($\rho \to 0$), we obtain

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

Now we will determine the asymptotics of the solution for $c_1 \to +\infty$ ($\rho \to 1$). To do this, we will find the value of L

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

Finally we get

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

Now let us analyze the generalization of the Nelson–Tantawi formula. Similarly, consider the equation (16) and determine the behavior of its solution as $c_1 \to 0$ ($\rho \to 0$) and $c_1 \to \infty$ ($\rho \to 0$) tend. We analyze the case of $\rho \to 0$, respectively, $c_1 \to 0$.

For the formula (15) with $K \geqslant 2$ it is true

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

therefore, as before,

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

Now let us analyze the case $\rho \to 1$ ($c_1 \to \infty$). For (15) with $K \geqslant 2$ we get

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

In fact, the value of L_2 was calculated earlier and corresponds to the value of L from (20)

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

Now let's calculate L_1

$$\begin{aligned} L_1 &= \lim_{\rho\to 1}(1-\rho)^2 f_1'(\rho) \\ &= \lim_{\rho\to 1}\rho(1-H)\Big(2Q_3\rho^2 - \rho(3Q_3 - (1-H)Q_2 - Q_1) - (2Q_1 + 2(1-H)Q_2)\Big) \\ &= (H-1)(Q_1 - (H-1)Q_2 + Q_3) = \left(\frac{H_K}{H_2} - 1\right)\left[Q_1 - \left(\frac{H_K}{H_2} - 1\right)Q_2 + Q_3\right] \\ &\approx \left(\frac{H_K}{H_2} - 1\right)\left[0.183577 - 0.070236\left(\frac{H_K}{H_2} - 1\right)\right]. \end{aligned}$$

Thus, we obtain that

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

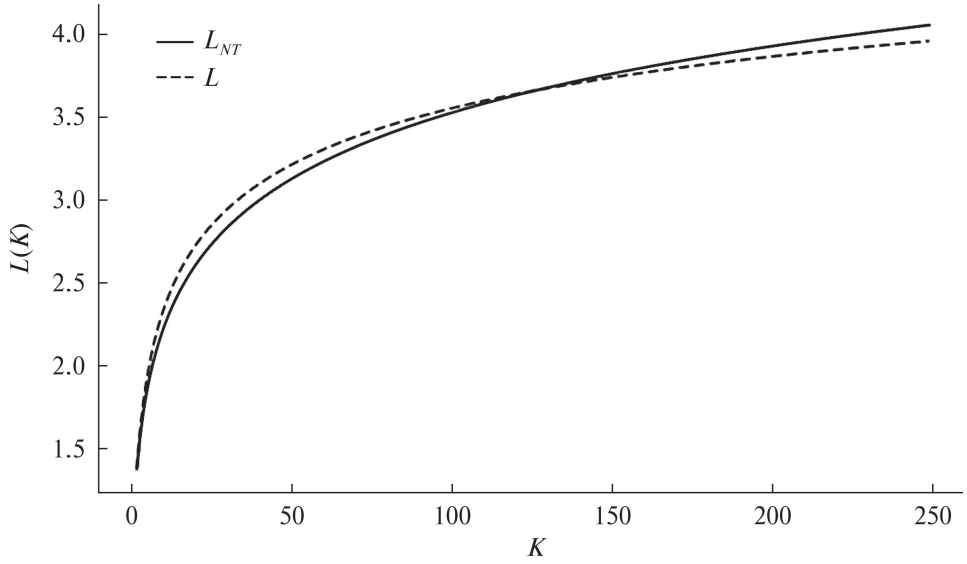

Fig. 5. Graph of the dependence of the value of L on K for the case of the Nelson–Tantawi formula (12) and the case of the generalized formula (15).

where

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

Figure 5 shows the graph of the dependence of the value of L on the number of subsystems for the case of the Nelson–Tantawi formula (20) and the case of the generalized formula (21). After the value $K = 126$, a "break" in the graphs occurs, and if at first the value of L_{NT} exceeded the value of L for the generalized formula, then for values $K \geqslant 127$ the situation changed exactly the opposite, which is also confirmed by the graphs presented earlier for the optimal value of the load depending on the parameter c_1 in Fig. 4 for $K = 20$ and $K = 200$.

As for the practical application of the obtained asymptotic results, both cases of high ($\rho \to 1$) and low ($\rho \to 0$) load may be of interest. For example, in systems with intensive use of data, parallel and distributed computing have become widespread as one of the main ways to improve performance in processing big data. Therefore, the owners of such high-performance computing environments are interested in obtaining forecasts of the system behavior under peak loads, as well as in the opposite situations, when the platforms are least in demand, in order to develop strategies for the efficient operation of the systems. In this context, asymptotic formulas can suggest when (at what prices per unit of service rate) such modes (high and low load) are optimal, i.e., actually desirable for the functioning of the system, in terms of cost, and when not.

7. CONCLUSION

The paper studies a fork-join queueing system from the point of view of managing the optimal cost of its operation depending on the price of a unit of resource affecting the system's performance and, accordingly, its response time. The mathematical model is constructed that takes into account the optimal ratio of cost and efficiency of the system's operation. The analysis is carried out based on previously obtained approximations for the average response time, both one of the most well-known and accurate, and its generalization obtained by the authors of this paper. For the particular case of the system, i.e., when the number of subsystems is two, it is possible to derive in explicit form expressions for the optimal system load. For a larger number of subsystems, equations of the fifth degree are presented, the numerical solution of which allows one to determine the sought

values of the system load and, accordingly, the service intensity, which actually characterizes the "power" of the necessary resources. The asymptotic behavior of the system is also analyzed.

Since the approximations of the average response time used here are quite accurate, they also describe the total cost of expenses well, so that with the calculated values of the optimal load (and, accordingly, the optimal service rate), the cost will be close to the optimal one. Therefore, the calculated values of the load can be recommended for use as estimates of their actual values for fork-join systems. In addition, if there is an estimate of the optimal load for one number of subsystems, it can be recalculated into an estimate for another number of subsystems (see Fig. 2). Since in the general case the problem is solved only numerically, then in cases of high and low prices per unit of service rate, it can be recommended to use asymptotic formulas that give simpler, but at the same time rougher estimates of the optimal load.

The mathematical model proposed in the article for control the optimal functioning of the system will also be valid in more general cases of fork-join QS, namely with a non-Poisson input flow and/or a non-exponential distribution of service time on servers, as in [9]. In this case, of course, analytical approximations for the value of the average response time of the system are necessary to derive specific relationships or equations. In addition, one can consider models with a nonlinear dependence of the cost of service costs on the service rate (for example, a power dependence, as mentioned in [20, §1.1]).

REFERENCES

1. Nelson, R. and Tantawi, A.N., Approximate Analysis of Fork/Join Synchronization in Parallel Queues, *IEEE Transact. Comput.*, 1988, vol. 37, pp. 739–743.
2. Varma, S. and Makowski, A.M., Interpolation Approximations for Symmetric Fork-Join Queues, *Performance Evaluat.*, 1994, vol. 20, pp. 245–265.
3. Varki, E., Merchant, A., and Chen, H., The M/M/1 Fork-Join Queue with Variable Subtasks, *Unpublished*, available online: https://www.cs.unh.edu/?varki/publication/2002-nov-open.pdf
4. Thomasian, A., Analysis of Fork/Join and Related Queueing Systems, *ACM Computing Surveys (CSUR)*, 2014, vol. 47, no. 2, pp. 1–71.
5. Qiu, Z., Pérez, J.F., and Harrison, P.G., Beyond the Mean in Fork-Join Queues: Efficient Approximation for Response-Time Tails, *Performance Evaluat.*, 2015, vol. 91, pp. 99–116.
6. Nguyen, M., Alesawi, S., Li, N., Che, H., and Jiang, H., ForkTail: A Black-Box Fork-Join Tail Latency Prediction Model for User-Facing Datacenter Workloads, *Proc. 27th Int. Symp. High-Perform. Parallel Distrib. Comput.*, 2018, pp. 206–217.
7. Nguyen, M., Alesawi, S., Li, N., Che, H., and Jiang, H., A Black-Box Fork-Join Latency Prediction Model for Data-Intensive Applications, *IEEE Transact. Parall. Distrib. Syst.*, 2020, vol. 31, no. 9, pp. 1983–2000.
8. Enganti, P., Rosenkrantz, T., Sun, L., Wang, Z., Che, H., and 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, pp. 1–8.
9. Gorbunova, A.V. and Lebedev, A.V., Nonlinear Approximation of Characteristics of a Fork-Join Queueing System with Pareto Service as a Model of Parallel Structure of Data Processing, *Math. Comput. Simulat.*, 2023, vol. 214, pp. 409–428.
10. Gorbunova, A.V. and Lebedev, A.V., On Estimating the Characteristics of a Fork-Join Queueing System with with Poisson Input and Exponential Service Times, *Advanc. Syst. Sci. Appl.*, 2023, vol. 23, pp. 99–114.
11. Gorbunova, A.V. and Lebedev, A.V., Correlations of the Sojourn Times of Subtasks in Fork-Join Queueing Systems with $M|M|1$-type Subsystems, *Advanc. Syst. Sci. Appl.*, 2024, vol. 24, no. 2, pp. 1–18.

12. Gorbunova, A.V. and Lebedev, A.V., Copulas and Quantiles in Fork-Join Queueing Systems, *Advanc. Syst. Sci. Appl.*, 2024, vol. 24, no. 1, pp. 1–19.

13. Ivanovskaya, I.A. and Moiseeva, S.P., Research of a mathematical model of parallel servicing of mixed-type requests, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2010, vol. 317, no. 5, pp. 32–34.

14. Zhidkova, L.A. and Moiseeva, S.P., Study of the parallel service system of multiple requests of the simplest flow, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2011, vol. 17, no. 4, pp. 49–54.

15. Moiseeva, S.P. and Zahorolnaya, I.A., Mathematical model of parallel servicing of multiple requests with repeated requests, *Avtometriya*, 2011, vol. 47, no. 6, pp. 51–58.

16. Moiseeva, S.P., Pankratova, E.V., and Ubonova, E.G., Study of infinite queueing system with heterogeneous service and input Markov renewal flow, *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika*, 2016, vol. 35, no. 2, pp. 46–53.

17. Osipov, O.A., A Heterogeneous Fork-Join Queueing System in Which Each Job Occupy All Free Servers, *RUDN Journal of Mathematics, Information Sciences and Physics*, 2018, vol. 26, no. 1. pp. 28–38.

18. Khabarov, R.S., Lokhvitsky, V.A., and Dudkin, A.S., Sojourn Time Approximation for Fork-Join Queue Based on Relationship Invariants, *Intellectual Technologies on Transport*, 2020, vol. 22, no. 2, pp. 46–50.

19. Redrugina, N.A., Method for Time Characteristics Calculating in the Service Platforms of Infocommunication Transactional Services with Parallel Requests Processing, *Proceedings of Telecommunication Universities*, 2023, vol. 9, no. 3, pp. 82–90.

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

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

22. Kurosh, A.G., *Algebraicheskie uravneniya proizvol'nyh stepenei (Populyarnye lektsii po matematike; vyp. 7)* (Algebraic equations of arbitrary degrees (Popular lectures on mathematics; issue 7)), Moscow: Nauka, 1975.

23. Kurosh, A.G., *Kurs vysshei algebry: Uchebnik* (Higher Algebra Course: Textbook), Saint Petersburg: Lan', 2008.

This paper was recommended for publication by A.A. Galyaev, a member of the Editorial Board