Transient analysis of queueing system with two heterogeneous servers

L I Korolkova¹, N Mashrabov², A M Murzin¹, A V Panfilov¹ and Ju L Siuskina¹

¹ Aerospace Department, South Ural State University (national research University), Lenin Ave., 76, Chelyabinsk, 454080, Russia
² Federal State Budgetary Educational Institution of Higher Education ‘South Ural State Agrarian University’, Lenin Ave., 75, Chelyabinsk, 454080, Russia

E-mail: Korolkovali@rambler.ru

Abstract. A method that uses an open-end diagram of queueing system functioning and relative probabilities is applied to study the queueing system with two heterogeneous servers, which is functioning within a given time period. According to the method a service process diagram is drawn. A feature of the method is the calculation of durations of the processes residual from the previous state. It allows describing in detail the functioning process of the queueing system for both infinite and finite time periods. The difference of the method from other approaches is the application of relations for computation of states without formulation and solution of any equations, absence of constraints on distribution of service time and inter-arrival time. The authors analysed the queueing system with arbitrary distributed service time and inter-arrival time, which is functioning within a random time period. Each state is characterized by idle time of devices or availability of the queue that allows, as a result, computing probabilities and average durations of idle time of servers, distribution of queue length, average time of stay in queue.

1. Introduction
The two-line queueing system with heterogeneous servers is under consideration, following the research of the queueing system with identical servers, and is connected with the development of models describing a spectrum of situations in the queueing system. Heterogeneous service is clearly a main feature of the operation of almost any manufacturing system.

The Markov systems were the first to be historically studied. The study of the systems is currently under way [1–4], Troublesome servers are considered in [5]. Performance measures of the system, where one server is characterized by exponential service time and the second one — by a general one, have been studied in [6, 7].

It is also worth pointing out [8], where the asymptotic assessment of the tail distribution of steady-state waiting time under similar but arbitrary distributed time of service by each server has been obtained. The steady-state system characteristics are studied in all the listed references.

The transient queueing theory is mathematically more complicated than the steady-state theory. Mainly, in the Markov systems the transient mode is studied with complication of queue discipline, repair of servers, etc. [9–11]. Therefore, the queue with server failure and repair, including impatient customers, was researched in [12]. The exact distribution of the n-th customer’s sojourn time in an $M/M/s$
queue with initially present \( k \) customers was received in [13]. Kumar and Sharma [14] obtained explicit transient probabilities of system size using the probability generating function technique. The model with Poisson arrivals and generally distributed processing times is investigated in [15]. The average number of served arrivals is determined by means of integral equations.

Analysis of the multi-server systems under random waiting times of arrivals and service has not allowed obtaining explicit analytic decisions yet. Rather complicated methodology is proposed by Bandi [16]. Authors model queueing primitives, on the basis of which they eventually received closed-form expressions for the worst case transient system time. The average behavior is achieved by averaging the values of the worst case. The approximations match the diffusion approximations for a single queue with light-tailed primitives.

The method suggested in [17] for description of random processes was applied in research of multistage systems with buffers [18]. We are going to apply it to compute the two-server queueing system, which service time and inter-arrival time are characterized by general distribution functions.

2. Application of conditional probabilities method

The two-server system is under consideration, which is operates within the given time period \( T_H \). Let us set: \( T_1 \) — inter-arrival time, \( T_2,1 \) — the first device service time, \( T_2,2 \) — the second device service time. We consider first-come first-served discipline (FCFS) model in which inter-arrival time and service times are independent and general distributed. Arrivals wait in the system until service is completed. If both servers are free, the upcoming arrival is processed by a faster server.

According to the method, it is necessary to present the system functioning process in a diagram. The diagram represents the states as the durations of simultaneous processes. The time \( T_H \) represents the state as any other duration with the only difference that its termination suspends all the rest private processes with residual durations.

The state is terminated when one process (servicing or waiting) comes to an end. The method assumes that simultaneous termination of processes is impossible. Time difference in the process termination can be an infinitely small quantity. In this case the generic process enters into another state shown as an arrow in the diagram. In brackets beneath the arrow a duration designation is recorded, which termination in a previous adjacent state has stipulated the state under consideration. The underscored state beneath duration in brackets means its limiting.

In general case, the scheme can be used to calculate: the residual service times, the residual time \( T_H \), the probabilities of transitions between states, unconditional and conditional durations of states, the duration from the initial to any non-adjacent state.

For the system in question the diagram is as follows (Figure 1).

![Figure 1. Schematic representation of the beginning of the system operation](image-url)

In an initial zero state the first arrival (designation 1) is expected within time \( T_H \); it means that both servers stand idle. Initial state terminates either with a request arrival or on expiration of time \( T_H \). In the
first case the system enters into state 1 (numeric symbol to the right of the arrow), in the second one—in state 0*, and the system stops work. In state 1 wait for the forthcoming arrival (designation 1) is in the process; the received arrival is served by the first server (designation 21), and \( T_H \) remains, which in state 1 is expressed as \( H_1 \) in the diagram. The above state is characterized by an idle time of server 2.

The system enters into state 2 after the arrival of the next order, so in this state it waits for the next arrival (designation 1); service by the first server is going on; the newly arrival is started to be served by the second server (designation 22), and \( T_H \) remains. Duration \( T_{21,2} \) (designation 21,2) is the residual time for arrival service by the first server. In the above state all the servers are busy, queue is absent.

The system enters into state 3, if the service time by the first server is less than both the waiting time for an arrivals and the residual time \( T_H \); 13 designates the waiting time for an arrival in the diagram. Both the above state and the zero one are characterized by idle time of both servers.

Queue is absent in states 9 and 10. Idle time of server 1 is observed in state 5, idle time of server 2 — in states 6 and 7. State 4 is characterized by a one-arrival queue, state 8 — by a two-arrival queue. Other states in the diagram can be described in a similar way. The vital difference from the diagrams of Markov processes lies in the one-way of transitions.

The next stage is to compute characteristics of the states required for a particular problem solution. The states are predominantly characterized by the residual durations of private processes, and the opportunity for their calculation is one of the features of the method.

In the initial (zero) state, the service times are known (given). In subsequent contiguous states service processes that have not ended will be characterized by the residual service times.

Distribution function calculation of residual duration is as follows: there are \( m \) processes in state \( k \). It is required to define distribution function of residual duration of any process \( j \) after process \( i \) termination; with that the generic process enters into state \( r \). Thus, two elementary processes \( i \) and \( j \) have been stood out among a set of elementary processes \( M^{(k)} \) and a subset of private processes \( U_{j-i} = M^{(k)} - i - j \) have been formed. Combined duration of the above processes has the distribution function \( G_{j-i}(t) \), where \( n \in U_{j-i} \):

\[
G_{j-i}(t) = P[U_{j-i} \leq t] = P[\min(T_n) \leq t] = 1 - \prod_n[1 - E_n(t)]
\]  

(1)

The residual duration of process \( j \) in state \( r \) can be represented by:

\[
T_{j,r} = T_j - T_{i<\alpha} \mid T_j > T_{i<\alpha},
\]  

(2)

where \( T_{i<\alpha} = T_i < U_{j-i} \).

The vertical bar in the expressions specifies the condition. Distribution functions of the given durations are computed serial by formulas:

\[
B_{i<\alpha}(t) = P[T_{i<\alpha} \leq t] = P[T_i \leq t \mid T_i < U_{j-i}] = B_{i<\alpha}^+(t) / B_{i<\alpha}^+(\infty),
\]  

(3)

\[
B_{i<\alpha}^+(t) = \int_0^t [1 - G_{j-i}(y)] dE_i(y),
\]  

(4)

\[
E_{j,r}(t) = P[T_{j,r} \leq t] = \left| E_{j,r}^+(0) - E_{j,r}^+(t) \right| / E_{j,r}^+(0),
\]  

(5)

\[
E_{j,r}^+(t) = \int_t^\infty B_{j<\alpha}(y-t) dE_j(y).
\]  

(6)

For computation of two-server queueing system characteristics, it is necessary to define the conditioned time of generic process stay in state \( k \). It is determined from the end of \( T_k \) duration, the first one from a set of others, resulted in the transition from state \( k \) to state \( r \). Distribution function \( S_{k,r}(t) \) is actually computed by (3):

\[
S_{k,r}(t) = P[V_{k,r} \leq t] = P[T_{i,k} \leq t \mid T_{i,k} \leq U_{k,r}] = S_{k,r}^+(t) / S_{k,r}^+(\infty),
\]  

(7)
where $S_{k,r}^r(t) = \int_0^t [1 - G_{k,r}(y)] dE_{i,k}(y)$;

$$G_{k,r}(t) = P[U_{k,i} \leq t] = P[\min(T_{j,k}) \leq t] = 1 - \prod_j [1 - E_{j,k}(t)], \ i \in k - i, \ i \neq j. \quad (8)$$

At the same time we determined the transition probability $s_{k,r}$ from state $k$ to state $r$, reflecting a fraction of processes being in state $k$ and passing into succeeding adjacent state $r$ as a result of process $i$ termination:

$$s_{k,r} = P[T_{i,k} < U_{i,k}] = S_{k,r}^r(\infty). \quad (9)$$

For example, let us consider the calculation of some residual durations. The residual time $T_H$ in state 1 according to diagram $T_{H,1}=T_{H'-T_1}$. The distribution function:

$$E_{H,1}(t) = P[T_H < T_1 \leq t] = E_{H,1}^+(t) - E_{H,1}^+(0), \text{ where } E_{H,1}^+(t) = \int_t^{\infty} E_1(y-t) dE_H(y). \quad (10)$$

Let us consider the transition from state 1 to state 2 due to the completion of waiting for the duration $T_1$. According to the calculation sequence of residual durations: $k=1$, $r=2$, $i=1$. Let us calculate the remaining service time for the first server $T_{2,1,2}$. Then the subset $U_{i,k}$ consists of the residual duration $T_{H,1}$ and $B_{i,k}=B_{1,1,1}$. In the abbreviated (informative) form, the expression for computing the duration $T_{2,1,2}$ can be written:

$$T_{2,1,2} = [T_{2,1} - [T_1 < \theta_{H,1}]] [T_{2,1} > [T_1 < \theta_{H,1}]] = (T_{2,1} - U_{1,1,1}) (T_{2,1} > U_{1,1,1})$$

3. Numerical results

3.1. Markov model $M/M/1$

Queue length in $n$ elements is computed by the known formula $q_n=(1-\rho)\rho^n$, where $\rho=\lambda/\mu$, $\lambda$ — arrival rate and $\mu$ — service rate.

A comparison with the exact solution is given in Table 1, where $\delta$ is the relative error, expressed as a percentage.

| $n$  | Exact solution | Suggested method | $\delta$, % |
|------|----------------|------------------|-------------|
| 0    | 0.5000         | 0.4960           | 0.8         |
| 1    | 0.2500         | 0.2463           | 1.5         |
| 2    | 0.1250         | 0.1215           | 2.8         |
| 3    | 0.0625         | 0.0600           | 4.0         |
| 4    | 0.0313         | 0.0311           | 6.7         |
| 5    | 0.0156         | 0.0155           | 6.2         |

First, the error is related to some errors of the method as an engineering approach. Second, exponential distribution with large dispersion leads to a great process fork, when each new state gives growth of its occurrence probability of $10^5…10^6$ order.

3.2. Two-server queueing system

Initial data:

$$E_H(t) = 1 - \exp \left(-3t^4\right), \ E_1(t) = 1 - \exp \left(-0.13t^2\right),$$

$$E_{2,1}(t) = 1 - \exp \left(-5 \cdot 10^{-3}t^4\right), \ E_{2,2}(t) = 1 - \exp \left(-1.3 \cdot 10^{-4}t^5\right).$$

Using the scheme calculated by (1)—(9), the following values were obtained:

- Idle time probability for the first and second servers simultaneously, $P_{12}=\delta_1+\delta_2+\ldots$.
Idle time probability for the first server, \( p_1 = s_{0.5} + \ldots \).

Idle time probability for the second server, \( p_2 = s_{0.1} + s_{0.6} + s_{0.7} + \ldots \).

Probability of queue absence when servers were busy, \( q = s_{0.2} + s_{0.9} + s_{0.10} + \ldots \).

Probability of queue absence, \( q_0 = q + p_{12} + p_1 + p_2 + \ldots \).

Probability of a one arrival in queue, \( q_1 = s_{0.4} + \ldots \), etc.

Average durations of the listed findings are computed by the formula:

\[
L = \sum_k s_{0,k} \left( \sum_r \bar{v}_{k,r} s_{k,r} \right)
\]  

where \( \bar{v}_{k,r} \) is the average conditional duration of the process stay in the state \( k \), provided that the general process passes to the state \( r \).

### Table 2. Indexes for system functioning.

| Index                                      | Probability | Average duration |
|--------------------------------------------|-------------|------------------|
| idle time for the first and second servers | 0.0758      | 2.850            |
| simultaneously                             |             |                  |
| idle time probability for the first server | 0.1495      | 0.620            |
| idle time probability for the second server| 0.1741      | 2.560            |
| absence of queue                           | 0.8179      | 7.650            |
| one arrival in queue                       | 0.1614      | 0.230            |
| two arrivals in queue                      | 0.0186      | 0.016            |
| three arrivals in queue                    | 0.0014      | 0.001            |

It is received that \( \sum_{i=0}^{3} q_i = 0.9993 \), which is due to the large branching of the general working process of the system, when the probabilities of states reach values less than \( 10^{-4} \). The average process duration is 7.8967, given that \( \bar{\tau}_H = 8.3 \). The relative error of duration computation is 4.86% measurements.

### 4. Conclusion

An open diagram and an opportunity for residual duration computations, allowing studying the transient behaviour of the system, are the principal differences of the method, for instance, from the Markov model.

The benefits of the method are:

- no constraints on service time and inter-arrival time distribution;
- not complicated schematization of the generic working process of the system;
- application of relations for state computation without set-up and solution of any equations;
- opportunity for detailed study of the transient process;
- process computation in a random given time period.

Methods with the above listed set of benefits have not been treated in the references considered by the authors.

The given method is rather lengthy, though it is hard to expect any simple solution for the investigated general case [16].

A further area of research will be the application of the method to obtain transient characteristics of multiline queuing systems.

### Acknowledgments

The work was supported by Act 211 of the Government of the Russian Federation, contract № 02.A03.21.0011.
References

[1] Hajek B 1984 Optimal control of two interacting service stations IEEE Trans. Autom. Control 29(6) 491–499

[2] Cohen J 1998 Analysis of the asymmetrical shortest two-server queueing model Int. J. Stoch. Anal. 11(2) 115–162

[3] Efrosinin D, Sztrik J 2011 Performance analysis of a two-server heterogeneous retrial queue with threshold policy Qual. Technol. and Quant. Managm. 8(3) 211–236

[4] Selen J, Adan I, Kapodistria S 2016 Steady-state analysis of shortest expected delay routing Queueing Syst. 84(3-4) 309–354

[5] Efrosinin D, Sztrik J, Farkhadov M, Stepanova N 2017 Reliability analysis of a two-server heterogeneous unreliable queueing system with a threshold control policy Proc. Int. Conf. on Information Technologies and Mathematical Modelling (vol 800) eds A Dudin, A Nazarov, A Kirpichnikov (Kazan: Springer, Cham) pp 13–27

[6] Boxma O J, Deng Q, Zwart A P Waiting time asymptotics for the M/G/2 queue with heterogeneous servers 2002 Queueing Syst. 40(1) 5–31

[7] Sani S, Daman O A 2015 The M/G/2 queue with heterogeneous servers under a controlled service discipline: stationary performance analysis Int. J. of App. Math. 45(1) 31–40

[8] Blanchet J, Murthy K R 2015 Tail asymptotics for delay in a half-loaded Gi/Gi/2 queue with heavy-tailed job sizes Queueing Syst. 81(4) 301–340

[9] Dharmaraja S 2000 Transient solution of a two-processor heterogeneous system Math. and Comp. Modell. 32(10) 1117–1123

[10] Chandrika K U 2006 Transient analysis of a system with queue dependent servers Econ. Qual. Contr. 21(2) 291–299

[11] Ammar S I 2014 Transient analysis of a two heterogeneous servers queue with impatient behavior J. Egypt. Math. Soc. 22(1) 90–95

[12] Sudhesh R, Savitha P, Dharmaraja S 2017 Transient analysis of a two-heterogeneous servers queue with system disaster, server repair and customers’ impatience TOP 25(1) 179–205

[13] Drew J H, Evans D L, Glen A G, Leemis L M 2017 Transient Queueing Analysis Comp. Prob. Int. Ser. in Op. Res. & Manag. Sci. 246 241–275

[14] Kumar R, Sharma S 2019 Transient solution of a two-heterogeneous servers’ queuing system with retention of reneging customers Bull. Malays. Math. Sci. Soc. 42(1) 223–240

[15] Kempa W M, Kurzyk D 2017 Transient processing analysis in a finite-buffer queueing model with setup times Proc. of 37th Int. Conf. on Information Systems Architecture and Technology Part II. Advances in Intelligent Systems and Computing (vol 522) eds A Grzech, J Świątek, Z Wilimowska, L Borzemski (Karpacz: Springer, Cham) pp 163–175

[16] Bandi C, Bertsimas D, Youssef N 2018 Robust transient analysis of multi-server queueing systems and feed-forward networks Queueing Syst. 89(3-4) 351–413

[17] Korolkov I V, Korolkova L I 2001 Discrete in the composition and continuous in time random processes Rev. of App. and Indust. Math. 8 614–615

[18] Korolkova L I, Masharov N 2015 Analysis of production line with finite buffers and a general service time Proc. Eng. 129 321–325