European Journal of Operational Research 271 (2018) 934–952

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

An algorithmic approach to analysing the reliability of a controllable unreliable queue with two heterogeneous servers[☆]

Dmitry Efrosinin[a,b,*], Janos Sztrik[c]

[a] Institute for Stochastics, Johannes Kepler University, Altenbergerstrasse 69, Linz 4040, Austria
[b] Peoples' Friendship University of Russia (RUDN University), Miklukho-Maklaya street 6, Moscow 117198, Russia
[c] University of Debrecen, Egyetem ter 1, Debrecen 4032, Hungary

ARTICLE INFO

Article history:
Received 18 October 2017
Accepted 5 June 2018
Available online 15 June 2018

Keywords:
Queueing system with heterogeneous servers
Reliability analysis
Quasi-birth-and-death process
Threshold policy
Matrix-geometric solution method

ABSTRACT

We consider a Markovian queueing system with two unreliable heterogeneous servers and one common queue. The servers serve customers without preemption and fail only if they are busy. Customers are allocated to one or the other server via a threshold control policy which prescribes using the faster server whenever it is free and the slower server only when the number of waiting customers exceeds a specified threshold level that depends on the state of the faster server. This paper focuses on the reliability analysis of a system with unreliable heterogeneous servers. First, we obtain the stationary state distribution using a matrix-geometric solution method. Second, we analyse the lifetimes of the servers and of the system. We provide algorithms for calculating the stationary reliability characteristics, reliability functions in terms of the Laplace transform and the mean times to the first failure. A new reliability measure is introduced in the form of the discrete distribution function of the number of failures during a specified life time that is derived from a probability generating function. The effects of various parameters on these reliability characteristics are analysed numerically.

1. Introduction

To improve modern communication systems in terms of performance and reliability, they can be supplied with controllable heterogeneous environment. The heterogeneity in such systems may be easily explained by virtue of the following examples. The data centers with a cloud computing paradigm containing the execution servers of many generations as a consequence of continuous system updates (Bai, Xi, Zhu, & Huang, 2015). Obviously in this system the servers can differ in terms of speed, capacity, availability, power consumption an so on. Another example is a hybrid wireless channel working on the basis of Radio Frequency/Free Space Optic (RF/FSO) technology (Vishnevskii, Semenova, & Sharov, 2013). The links of this channel have unequal data throughput, availability and reliability characteristics. The capacity of RF link is constrained by limits to link throughputs on the order of 10 s of Mbps. On the contrary, the commercial FSO currently provide throughputs of several Gbts but the link availability is limited by adverse weather conditions like fogs and heavy snowfalls. Therefore, the hybrid channel combines advantages of both types of links. One more example is a single cell of a cellular (3GPP LTE) network with a Licence Shared Access (LSA) technology, for details see Gudkova et al. (2015), which assumes that the band can be used when the owner does not need it. In this case heterogeneous environment consists of the reliable main and unreliable reserve pool of servers which is used according to a specified hysteretic control policy. The proposed examples have motivated us to apply the queueing system with unreliable heterogeneous servers for modelling the dynamic behaviour and analysis the relationships between different factors influencing on reliability of communication systems with heterogeneous unreliable environment.

Analyses of multi-server queueing systems generally assume that the servers are homogeneous. Mitrany and Avi-Itzhak (1967) and Neuts and Lucantoni (1979) studied the $M/M/s$ queueing system with server breakdowns and repairs. Levy and Yechiali (1976) analysed the $M/M/s$ queue with server vacation. A recent paper by Efrosinin, Samouylov, and Gudkova (2016) reported on stationary analysis of the busy period for a multi-server Markovian queueing system with simultaneous failures of servers. Queues with heterogeneous unreliable servers have rarely been addressed by research. A queueing system with two heterogeneous servers

[☆] The publication was prepared with the support of the "RUDN University Program 5–100", RFBR according to the research project No. 16-37-60072 mol.a.dk

[*] Corresponding author at: Institute for Stochastics, Johannes Kepler University, Altenbergerstrasse 69, 4040 Linz, Austria.

E-mail addresses: dmitry.efrosinin@jku.at, dmitriy_e@mail.ru (D. Efrosinin), sztrik.janos@inf.unideb.hu (J. Sztrik).

https://doi.org/10.1016/j.ejor.2018.06.007
0377-2217/© 2018 Elsevier B.V. All rights reserved.

and multiple vacations was studied by Kumar and Madheswari (2005), who obtained the stationary queue length distribution by using a matrix geometric method and provided an analysis of busy period and waiting time. In Kumar, Madheswari, and Venkatakrishnan (2007), the same authors introduced the $M/M/2$ queueing system with heterogeneous servers subject to catastrophes, and provided a transient solution for the system under study. A heterogeneous two-server queueing system with balking and server breakdowns was studied by Yue, Yue, Yu, and Tian (2009). They used a matrix-geometric solution method to obtain some mean performance measures.

In a heterogeneous queueing system with one common queue, particularly in the case of service without preemption (a customer can not change the server during a service time) a mechanism that allocates customers to the servers must be specified. The majority of heterogeneous systems investigated use heuristic service policies (e.g. the Fastest Free Server (FFS) or Random Service Selection (RSS) policies). In fact, these policies are not optimal, if, for instance, the mean response time is to be minimized. As previously shown (see, e.g. the results of B & Jouini, 2016; Efrosinin, 2008; Koole, 1995; Lin & Kumar, 1984; Rykov & Efrosinin, 2009), the optimal allocation policy for heterogeneous queueing systems is one of a class of threshold policies where the less effective server is to be used only if the number of customers in the queue has reached some pre-specified threshold level. This result was confirmed for a queueing system with faster unreliable server and absolutely reliable slower server in Efrosinin (2013), Ozkan and Kharoufeh (2014) and for two unreliable heterogeneous servers in a system with constant retrial discipline in Efrosinin and Sztrik (2016). In the last paper mentioned, it was shown that for a fixed threshold policy the corresponding Markov process is of the QBD (quasi-birth-and-death) type with a tri-diagonal block infinitesimal matrix with a large number of bounding states.

While first steps in performance analyses of controllable heterogeneous queueing systems with completely reliable servers have already been published, application to heterogeneous models also requires a reliability analysis of such queues when servers are subject to failure. Here we use a forward-elimination-backward-substitution method expressed in matrix form in terms of the Laplace–Stiltjes transforms (LST) combined with probability generating function (PGF) approach to evaluate reliability measures such as reliability function (i.e., the complementary cumulative distribution function of the lifetime) and mean time to first failure for each server separately and for the group of servers under the fixed threshold allocation control policy. The reliability functions are obtained in terms of the Laplace transform (LT), and a numerical inversion algorithm is used to obtain the time-dependent functions. Additionally, we introduce a new discrete reliability metric in the form of the distribution of the number of failures during a certain lifetime. We expect that our results can be generalized to the case of an arbitrary controllable unreliable queueing model with a QBD structure.

The remainder of paper is organized as follows: In Section 2, we describe the mathematical model and present the stationary state distribution using a matrix-geometric solution method. In Section 3, we develop a computational analysis of the stationary reliability characteristics, the reliability function and the mean time to first failure. The number of failures during a certain life time is investigated in Section 4. In Section 5, numerical examples are provided to highlight the effect of some parameters on the reliability characteristics.

Hereafter, the notations $\mathbf{e}(n)$, $\mathbf{e}_j(n)$, and I_n are used respectively for the column vector consisting of 1's, the column vector with 1 in the jth (beginning from 0th) position and 0 elsewhere, and an identity matrix of the dimension n. When there is no need to emphasize the dimensions of these vectors, the suffix is omitted and dimensionality is determined by the context. The expressions $diag(a_1, \dots, a_n)$, $diag^+(a_1, \dots, a_n)$, and $diag^-(a_1, \dots, a_n)$ denote respectively the diagonal matrix, the upper diagonal matrix, and the lower diagonal matrix with entries $a_1, \dots, a_n$ that can be scalars or matrices.

2. Mathematical model and stationary distribution

In this paper, we address a two-server heterogeneous unreliable queueing model of the $M/M/2$ type as illustrated in Fig. 1(a).

Customers arrive according to a Poisson process with arrival rate λ. The service times are exponentially distributed with rates μ_1 and μ_2, where $\mu_1 \geq \mu_2$. We assume that the servers fail respectively at exponential rates α_1 and α_2. A server can fail only if it is busy. A failed server is repaired immediately, and the time required to repair it is exponentially distributed respectively with rates β_1 and β_2. A customer being served at the moment of failure is left at this server during repair and can be served when the server becomes operational again. The mechanism of allocation to the two servers is based on a threshold policy: Depending on the state of the faster server, the slower is used whenever the number of customers in the queue exceeds a certain threshold level.

Let $Q(t)$ and $D(t) = \{D_1(t), D_2(t)\}$ denote, respectively, the number of customers in the queue and the vector state of servers at time t, where service process

$$D_j(t) = \begin{cases} 0, & \text{the server } j \text{ is idle,} \\ 1, & \text{the server } j \text{ is busy and operational,} \\ 2, & \text{the server } j \text{ has failed.} \end{cases}$$

with transitions as shown in Fig. 1(b). The threshold policy $f = (q_1, q_2)$ is defined by two threshold levels $1 \leq q_2 \leq q_1 < \infty$. According to this policy, server 1 must be used upon new arrival whenever it is free and there are customers in the queue, whereas idle server 2 is ready to serve the arriving customers only if server 1 is in state 1 or 2 and the number of customers in the queue has reached the corresponding threshold value q_1 or q_2. If server 1 is in state 1 or 2 upon service completion at server 2 and the number of customers in the queue is smaller than q_1 or q_2, then further allocation of customers to server 2 is not possible. For the fixed threshold policy f the process

$$\{X(t)\}_{t \geq 0} = \{Q(t), D(t)\}_{t \geq 0} \tag{1}$$

is a continuous-time Markov chain with a state space given by

$$E = \{x = (q, d_1, d_2); q \in \mathbb{N}_0, (d_1, d_2) \in E_D\}, \tag{2}$$

where E_D is a set of states of servers that is defined as

$$E_D = \left\{ (d_1, d_2); \begin{array}{c} d_j \in \{0,1,2\}, j \in \{1,2\}, q = 0, \\ d_1 \in \{1,2\}, d_2 \in \{0,1,2\}, 1 \leq q \leq q_2 - 1, \\ d_1 \in \{1,2\}, d_2 \in \{0,1,2\}, (d_1, d_2) \neq (2,0), \\ q_2 \leq q \leq q_1 - 1, \\ d_j \in \{1,2\}, j \in \{1,2\}, q \geq q_1, \end{array} \right\}.$$

Next we partition E into blocks as follows:

$$(\mathbf{0}, \mathbf{0}) = \{(0, 0, d_2); d_2 \in \{0, 1, 2\}\},$$

$$(\mathbf{q}, \mathbf{1}) = \begin{cases} \{(q,1,0), (q,2,0), (q,1,1), (q,2,1), (q,1,2), (q,2,2)\}, & 0 \leq q \leq q_2 - 1, \\ \{(q,1,0), (q,1,1), (q,2,1), (q,1,2), (q,2,2)\}, & q_2 \leq q \leq q_1 - 1, \\ \{(q,1,1), (q,2,1), (q,1,2), (q,2,2)\}, & q \geq q_1. \end{cases}$$