European Journal of Operational Research 256 (2017) 187–195

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers

Hendrik Baumann[a], Werner Sandmann[b,*]

[a] Department of Applied Stochastics and Operations Research, Clausthal University of Technology, Erzstr. 1, Clausthal-Zellerfeld 38678, Germany
[b] Department of Computer Science, Saarland University, Campus E1 3, Saarbrücken 66123, Germany

ARTICLE INFO

Article history:
Received 21 September 2015
Accepted 16 July 2016
Available online 21 July 2016

Keywords:
Queueing
Multi-server tandem queue
Markovian arrival process
Phase-type service time distributions
Matrix-analytic method

ABSTRACT

We consider multi-server tandem queues where both stations have a finite buffer and all services times are phase-type distributed. Arriving customers enter the first queueing station if buffer space is available or get lost otherwise. After completing service in the first station customers proceed to the second station if buffer space is available, otherwise a server at the first station is blocked until buffer space becomes available at the second station. We provide an exact computational analysis of various steady-state performance measures such as loss and blocking probabilities, expectations and higher moments of numbers of customers in the queues and in the whole system by modeling the tandem queue as a level-dependent quasi-birth-and-death process and applying suitable matrix-analytic methods. Numerical results are presented for selected representative examples.

1. Introduction

Tandem queueing networks, in short: tandem queues, are widely used to model systems and scenarios where services are delivered in successive stages, that is generic customers enter a multi-stage service facility and successively proceed through the service stages. This applies to such diverse fields as manufacturing, transportation, logistics, computer networking, telecommunications, and many daily-life service operations, amongst others. Consider for example production lines where different steps are processed by different machines, airplane maintenance and refueling, ordering and delivery of goods or services, the Internet or other communication networks where messages or data packets are routed from a source to a destination via multiple hops, or supermarkets where the first stage is the self service of customers collecting their items and the second stage is the payment at the check out.

Queueing theory has extensively dealt with tandem queues for more than five decades. However, standard assumptions of 'classical' queueing theory, especially Poisson arrival processes and independent exponentially distributed service times, do not appropriately reflect the traffic characteristics, service or processing times in many modern applications. Service times are often not exponentially distributed and interarrival times are often neither independent, identically distributed nor exponentially distributed. Thus, the arrival process is not even a renewal process. Therefore, we consider phase-type distributed service times and Markovian arrival processes.

Phase-type (PH) distributions Neuts (1981), O'Cinneide (1990) are extensions of the exponential distribution with the particularly nice property that any probability distribution on the nonnegative real numbers can be approximated, in principle arbitrarily accurately, by a PH distribution. Markovian arrival processes (MAPs) Neuts (1979), Lucantoni, Meier-Hellstern, and Neuts (1990) can properly incorporate correlations between successive interarrival times. They include Poisson processes, PH renewal processes, interrupted and Markov modulated Poisson processes as special cases and the whole class of MAPs is dense within the class of marked point processes Asmussen and Koole (1993). For monographs with a special focus on PH distributions and MAPs we refer to Breuer and Baum (2005) and Buchholz, Kriege, and Felko (2014), where the latter also addresses how data can be fitted to PH distributions and MAPs, respectively.

PH distributions and MAPs provide versatile means of modeling at the cost of a significantly increased complexity in that the state space of the associated Markov chain model becomes extremely huge. In particular, using PH distributed service times in multi-server models considerably increases the dimension of the state space. This curse of dimensionality poses a great challenge on model analysis, which is often tackled via approximations. For instance, Brandwajn and Begin (2014) have recently approached

* Corresponding author.
E-mail addresses: hendrik.baumann@tu-clausthal.de (H. Baumann), werner.sandmann@uni-saarland.de (W. Sandmann).

http://dx.doi.org/10.1016/j.ejor.2016.07.035
0377-2217/

the seemingly simple case of a multi-server queue with Poisson arrivals, PH distributed service times, and finite buffer by using a reduced state description in which the state of only one server is represented explicitly, while the other servers are accounted for through their rate of completions. In fact, queueing models with both MAP and PH distributed service times, even for single station single-server queues, are typically analyzed by approximations using state space reduction approaches and/or sophisticated matrix-analytic methods. Different variants of single-server queues with MAP and PH distributed service times have been studied in, e.g., Artalejo and Chakravarthy (2006), Krishnamoorthy, Babu, and Narayanan (2009), Sreenivasan, Chakravarthy, and Krishnamoorthy (2013). Multi-server queues with MAP and PH distributed service times are addressed in Chakravarthy (2013) where the MAP/PH/c retrial queue with PH distributed retrials is studied via simulation, and in Kim, Dudin, Taramin, and Baek (2013b) where a MAP/PH/c/c+K queue with two customer classes is analyzed by combining specific state space reduction techniques for queues with PH distributed service times and matrix-analytic methods.

Only relatively few works have investigated tandem queues with MAP and PH distributed service times. Gomez-Corral (2002) analyzes a tandem queue with MAP and two single-server stations, infinite buffer at the first station and no buffer at the second station. Service times at the first station are PH distributed and the service times at the second station are general. Van Houdt and Alfa (2005) consider a (discrete-time) tandem queue with MAP and two single-server stations where the service times at both stations are PH distributed, the queue at the first station is infinite and the queue at the second station is finite. Lian and Liu (2008) deal with a tandem queue with MAP, two single-server stations with queues of infinite capacity and exponentially distributed service times at both stations. Baumann and Sandmann (2013) consider a tandem queue with MAP, two single-server stations with queues of infinite capacity and PH distributed service times at both stations where customers may leave the system after completing service at the first station. Kim, Dudin, Dudin, and Dudina (2013a) model a call center as a tandem queue with MAP, multiple servers at both stations, exponentially distributed service times and no buffer at the first station, PH distributed service times and a finite buffer at the second station. Kim, Dudin, Dudina, and Dudin (2014) investigate a tandem queue with two types of customers, marked MAP, multiple servers at both stations, exponentially distributed service times and no buffer at the first station, PH distributed service times, one finite buffer and one infinite buffer at the second station.

In this paper, we consider tandem queues with MAP, two multi-server stations with finite buffers and PH distributed service times at both stations where at the first station losses of arriving customers and blocking after service can occur. We compute steady-state performance measures without introducing any approximation. For this purpose we model the class of tandem queues under consideration by level-dependent quasi-birth-and death (LDQBD) processes and use a matrix-analytic method according to Baumann and Sandmann (2013) for obtaining the desired steady-state performance measures without explicitly computing the stationary distribution. In the next section, we formally introduce the class of tandem queues considered, in particular we give general formal descriptions of the arrival process and the service time distributions. In Section 3 we describe the modeling of this class of tandem queues as LDQBD processes, where we introduce the structuring of the multi-dimensional state space as well as the state transitions and transition rates. Subsequently, in Section 4 we show how to express relevant performance measures and in Section 5 we present a matrix-analytic algorithm for the efficient computation of these performance measures. Numerical results for specific tandem queues of the considered type are presented in Section 6. In particular, for certain choices of the arrival process, the service time distributions and varying parameter values we study the loss and blocking probabilities as well as the expected numbers of customers in the two stations of the tandem queueing network. Section 7 concludes the paper and outlines further research directions.

2. Tandem queue specification

The class of the tandem queue we investigate can be specified in Kendall notation by MAP/PH/c_1/c_1+K_1 → /PH/c_2/c_2+K_2. The first station consists of c_1 identical servers and a buffer of finite capacity K_1. The second station consists of c_2 identical servers and a buffer of finite capacity K_2. Customers arrive at the first queueing station. If upon arrival the buffer is fully occupied, then the customer is lost. Customers having completed service in the first station proceed to the second station. If upon service completion in the first station the buffer at the second station is fully occupied, then the customer blocks a server at the first station until buffer space becomes available at the second station.

The service times at both stations are PH distributed. A probability distribution on $\mathbb{R}_+ = [0, \infty)$ is a (continuous) PH distribution, iff it is the distribution of the time until absorption in a finite time-homogeneous continuous-time Markov chain (CTMC). For such a CTMC with state space $\{1, \ldots, n+1\}$ and initial distribution $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_{n+1})$, where the states $1, \ldots, n$ are transient and the state $n+1$ is absorbing, the generator matrix A has the form

$$A = \begin{pmatrix} B & b \\ 0 & 0 \end{pmatrix}, \quad B \in \mathbb{R}^{n \times n}, \quad b \in \mathbb{R}^n \tag{1}$$

and $(\alpha_1, \ldots, \alpha_n, B)$ is a representation of the PH distribution, in short PH(α, B), with $\alpha = (\alpha_1, \ldots, \alpha_n)$. The number n of transient states is called the order of the PH distribution or the number of phases. In our model we denote the service time distributions at the first and the second station by PH(σ, S) and PH(τ, T), respectively, where PH(σ, S) is of order V and PH(τ, T) is of order W. Hence, $S = (s_{ij})_{i,j=1}^{V}$ and $T = (t_{ij})_{i,j=1}^{W}$. Besides, we define the vectors $s := -S\mathbf{1}$ and $t := -T\mathbf{1}$. The moments of a PH(α, B) distributed random variable Z are given by $\mathrm{E}[Z^k] = k!\ \alpha(-B)^{-k}\mathbf{1}$. Hence, the expected service times in our model are $-\sigma S^{-1}\mathbf{1}$ at the first station and $-\tau T^{-1}\mathbf{1}$ at the second station.

The arrival process is a continuous-time MAP of order U determined by the pair of $(U \times U)$-matrices (D_0, D_1) where $D_0 = (\nu_{ij})_{i,j=1}^{U}$, $D_1 = (\lambda_{ij})_{i,j=1}^{U}$ with $\nu_{ij} \geq 0$ for $i \neq j$, $\nu_{ij} < 0$ for $i = j$, $\lambda_{ij} \geq 0$ for all $i, j \in \{1, 2 \ldots, U\}$, $D_1 \neq 0$, and $D_0\mathbf{1} + D_1\mathbf{1} = 0$ such that $D = D_0 + D_1$ is the generator matrix of an irreducible CTMC on the state space $\{1, 2, \ldots, U\}$. This CTMC is commonly referred to as environmental process, phase process, or background CTMC. When it changes its state from state i to state $j \neq i$, with probability $\lambda_{ij}/(\nu_{ij} + \lambda_{ij})$ it triggers an arrival and with probability $\nu_{ij}/(\nu_{ij} + \lambda_{ij})$ it does not. The average arrival rate of the MAP is given by $\lambda = \psi D_1\mathbf{1}$ where ψ denotes the stationary distribution of the environmental process, i.e. $\psi D = 0, \psi\mathbf{1} = 1$. The stationary moments of the interarrival time Y are $\mathrm{E}[Y^k] = k!\lambda^{-1}\psi(-D_0)^{1-k}\mathbf{1}$. In particular, $\mathrm{E}[Y] = \lambda^{-1}$, and $\mathrm{Var}[Y] = (2\lambda\psi(-D_0)^{-1}\mathbf{1} - 1)/\lambda^2$, from which we get the squared coefficient of variation $c^2 = 2\lambda\psi(-D_0)^{-1}\mathbf{1} - 1$. The squared coefficient of correlation is given by $\gamma = (\lambda\psi(-D_0)^{-1}D_1D_0^{-1}\mathbf{1} - 1)/c^2$.

3. Markov chain model

We model the tandem queue as a CTMC $(X(t))_{t\geq 0} = (X_1(t), \ldots, X_{V+W+6}(t))_{t\geq 0}$ whose states are integer-valued vectors of the form

$$(n_1, n_2, u, \nu_0, \nu_1, \ldots, \nu_V, \nu_{V+1}, w_0, w_1, \ldots, w_W)$$

where