Computers & Operations Research 63 (2015) 46–55

Contents lists available at ScienceDirect

Computers & Operations Research

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

An approximation model for sojourn time distributions in acyclic multi-server queueing networks

Kevin R. Gue [a,*], Hyun Ho Kim [b]

[a] Department of Industrial Engineering, University of Louisville, Louisville, KY, USA
[b] Department of Industrial & Systems Engineering, Auburn University, Auburn, AL 36849, USA

ARTICLE INFO

Available online 1 May 2015

Keywords:
Queueing networks
Phase type distributions
Approximation models

ABSTRACT

We develop an approximation model for the sojourn time distribution of customers or jobs arriving to an acyclic multi-server queueing network. The model accepts general interarrival times and general service times, and is based on the characteristics of phase-type distributions. The model produces excellent results for multi-server networks with a small to medium number of workstations, but is less accurate when the number of workstations is large.

1. Introduction

For many service and order fulfillment settings, mean performance measures such as average work in process and expected flow time are insufficient as a means of understanding system performance and its implications for customer service. For example, rather than knowing that average flow time, or *sojourn time*, in an order fulfillment center is 50 min, managers might like to know what fraction of orders experience time in the system greater than, say, 3 h. Answering such detailed questions requires a distribution of sojourn time, rather than just its mean value.

Much of the literature on sojourn time modeling is focused on calculating the mean flow time (or simply the waiting time) in single-stage queues or in queueing networks [15]. Workstations in these systems may have one or many servers. The subject of our work is calculating the *distribution* of sojourn time, a more difficult task. Specifically, we show how to approximate the sojourn time distribution of customers arriving to a serial or an acyclic queueing network, when interarrival times and processing times can be general.

Existing research on sojourn time modeling can be divided into four categories, according to whether the models address single-stage systems or networks, and whether they address single or multiple servers per workstation. The simplest case is single-stage, single-server systems. Neuts [8] describes a matrix-geometric method to calculate the sojourn time distribution for a GI/G/1 system using phase type distributions. Luh and Zheng [6] implemented Neuts' method in MATHEMATICA. Sengupta [12] used a bivariate Markov process to model waiting time and queue length distributions in a GI/PH/1 queue. His method is a "continuous analog" of the matrix-geometric method in Neuts [8]. Sengupta also showed that if the interarrival and service time distributions in a single-stage, single-server queue are phase-type, then the waiting time distribution is also phase-type.

Asmussen and O'Cinneide [2] verify the same result for a single-stage, *multi*-server queue, in a paper extending Sengupta's work to the GI/PH/c case. Their model admits heterogeneous servers. Asmussen and Møller [1] show how to calculate the waiting time distribution in a GI/PH/c and MAP/PH/c queue, for both homogeneous and heterogeneous servers. Whitt [16] also addresses single-stage, multi-server systems, but examines state dependent waiting time distributions. For example, he shows how to compute the waiting time distribution for the k-th customer in line in a c-server, single-stage system. Rueda [10] develop an approximation for the waiting time distribution of single-stage queues with non-stationary interarrival and processing time distributions.

Sojourn time distributions for queueing networks of single-servers have been addressed by Shanthikumar and Sumita [13], You et al. [18], and Yoon [17]. Shanthikumar and Sumita [13] approximate the sojourn time distribution of an M/G/1 queueing system as one of three phase-type distributions (generalized Erlang, exponential, and hyperexponential), based on a "service index," which is defined as the squared coefficient of variation of the total service time of an arbitrary job. Neuts [8] showed that the convolution of phase type distributions is also phase type. You et al. [18] used this observation to show how to calculate the sojourn time distribution for queueing networks, with general interarrival and service times. In a paper published in Korean, Yoon [17] developed a method very similar to that of You et al.

Mandelbaum et al. [7] develop models for Markovian multi-server queueing networks in which customers can abandon and

* Corresponding author.
E-mail address: kevin.gue@louisville.edu (K.R. Gue).

http://dx.doi.org/10.1016/j.cor.2015.04.019
0305-0548/

retry to enter. They assume Markovian interarrival and service times. Gue and Kim [4] develop a state-dependent sojourn time distribution model for remaining time of a customer in a multi-server, multi-stage queueing system where the state of the system is the number of customers in each stage ahead of the customer of interest. The model allows general service times, but does not involve interarrival times because customers arriving after a customer is in the system do not affect its sojourn time.

Missing from the literature are models of sojourn time distributions for queueing networks with multiple servers, when interarrival and service times can take general distributions. We fill that void here. We believe that ours is the first approximation model of sojourn time distributions for queueing networks of multi-server workstations.

Our model is based on the work of Neuts [8], Asmussen and Møller [1], and You et al. [18]: we use the bivariate Markov process of Asmussen and Møller to extract the mean and variance of waiting times for each multi-server workstation. We then construct phase type distributions for waiting time and service time based on those means and variances. Then we use the method of You et al. to construct an infinitesimal generator and initial probability vector for the network. With these we can calculate a sojourn time distribution for the network of multi-server queues. For reasons we discuss below, this model produces good results only when the inverse of the squared coefficient of variation ($1/C^2$) is close to an integer. We correct this weakness with a simple but effective interpolation scheme. We demonstrate the validity of our model by comparing it with results from a simulation model.

2. Phase-type distributions

Consider a Markov process on states $\{1,\ldots,m+1\}$, having infinitesimal generator

$$\begin{bmatrix} \mathbf{Q} & \mathbf{Q^0} \\ \mathbf{0} & 0 \end{bmatrix},$$

where $\mathbf{Q}$ is an $m \times m$ matrix satisfying $Q_{ii} < 0$, for $1 \le i \le m$, and $Q_{ij} \ge 0$, for $i \ne j$ (For consistency, we follow the notation of [8]). $\mathbf{Q}^0$ is a column vector of size m such that

$$\mathbf{Qe} + \mathbf{Q^0} = \mathbf{0},$$

where $\mathbf{0}$ is a row vector of zeros and $\mathbf{e}$ is a column vector of ones. The initial probability vector of $\mathbf{Q}$ is $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ such that $\boldsymbol{\beta}\mathbf{e} = 1$.

We are interested in the time until absorption into state $m+1$, whose distribution Neuts [8] defines as *phase-type*.

Lemma 1 (*Neuts [8]*). *The probability distribution $F(\cdot)$ of the time until absorption in the state $m+1$, corresponding to the initial probability vector $(\boldsymbol{\beta}, \beta_{m+1})$, is given by*

$$F(x) = 1 - \boldsymbol{\beta} e^{\mathbf{Q}x}\mathbf{e}.$$

Because $F(\cdot)$ is completely specified by β and $\mathbf{Q}$, the pair $(\boldsymbol{\beta}, \mathbf{Q})$ is called a *representation* of $F(\cdot)$.

Neuts [9] provides the density function on $(0,\infty)$:

$$f(x) = \boldsymbol{\beta} e^{\mathbf{Q}x}\mathbf{Q}^0 = \boldsymbol{\beta} e^{\mathbf{Q}x}(-\mathbf{Q})\mathbf{e}.$$

For each distribution of service or waiting times in a network, we seek a phase-type approximation for which we can generate a matrix-analytic model of the CDF. We use the fact that finite convolutions of phase-type distributions are also phase-type [8] to generate solutions for a queueing network, as in You et al. [18].

The phase-type distribution is used to fit a general distribution based on the throughput rate λ and the squared coefficient of variation C^2 of a given positive random variable X. Sauer and Chandy [11], You et al. [18], and Tijms [14] showed different fitting methods that can convert a general distribution to a corresponding phase-type distribution based on the C^2 of the distribution.

A general distribution can be approximated with the Erlang distribution, the exponential distribution, or the hyperexponential distribution, based on its first and second moments. We follow the rule of You et al. [18]:

When $C^2 < 1$, we convert a general distribution to an Erlang distribution, $Erlang(m,\mu)$, of order m. The infinitesimal generator is

$$Q = \begin{bmatrix} -\mu & \mu & 0 & \cdots & 0 & 0 \\ 0 & -\mu & \mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\mu \end{bmatrix}, \quad Q^0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \mu \end{bmatrix},$$

where $m = \lceil 1/C^2 \rceil$ and $\mu = m\lambda$. The phase-type representation of a general distribution is represented by $\boldsymbol{\beta} = (1, 0, \ldots, 0)$ and Q.

When $C^2 > 1$, we use the hyperexponential distribution with balanced means, HE_2, of order 2. We use the following normalization for this process:

$$\frac{p}{\mu_1} = \frac{q}{\mu_2}.$$

The phase-type representation of a general distribution is given by $\boldsymbol{\beta} = (p, q)$ and

$$Q = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}, \quad Q^0 = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix},$$

where $p = \frac{1}{2}\left(1 + \sqrt{(C^2-1)/(C^2+1)}\right)$, $q = 1-p$, $\mu_1 = 2p\lambda$ and $\mu_2 = 2q\lambda$.

When $C^2 = 1$, we have an exponential distribution [14] represented by $\boldsymbol{\beta} = 1$ and Q,

$$Q = -\lambda, \quad Q^0 = \lambda.$$

3. Intuitive model

We now describe our model, which extends the work of Asmussen and Møller [1] and You et al. [18] to produce an approximation of the sojourn time distribution for a network of multi-server queues. For the waiting time distribution of a single-stage, multi-server system, we follow Asmussen and Møller's method, then we assume independence and convolve waiting times and service times in the system based on the method of You et al. The model requires only the following input, which should be available in most real systems: mean and variance of processing times for each workstation and mean and variance of interarrival times to the system.

The procedure is as follows:

1. Compute the arrival rate and SCV of an arrival process at each workstation using the Queueing Network Analyzer (QNA) method of [15].
2. Approximate an interarrival time and service time distribution of each workstation i as a corresponding phase-type distribution $(\boldsymbol{\alpha}_i, \boldsymbol{A}_i)$ and $(\boldsymbol{\beta}_i, \boldsymbol{S}_i)$ based on C_a^2 and C_s^2.
3. Compute the mean and variance of waiting time of each workstation i using the method of Asmussen and Møller. Approximate the waiting time distribution of each workstation i as a corresponding phase-type distribution $(\gamma_i, \boldsymbol{W}_i)$ based on C_w^2.
4. Assume independence and convolve all waiting times and service times sequentially to generate a phase-type representation of the