Contents lists available at SciVerse ScienceDirect

Journal of Network and Computer Applications

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

Delays in a series of queues with correlated service times

Werner Sandmann [*]

Department of Applied Stochastics and Operations Research, Clausthal University of Technology, Erzstr. 1, D-38678 Clausthal-Zellerfeld, Germany

ARTICLE INFO

Article history:
Received 1 February 2011
Received in revised form
29 July 2011
Accepted 10 October 2011
Available online 20 November 2011

Keywords:
Series of queues
Tandem queueing networks
Service time correlations
Performance measures
Waiting times
End-to-end delays

ABSTRACT

In almost all applications of queueing network models it is assumed that for each customer the service times at different network nodes are independent. But service times in, for instance, computer and communication networks are typically essentially determined by properties like message or packet lengths that do not change substantially on the route through the network. Therefore, the service times of any customer in a queueing network are likely to be correlated, which can significantly influence quality of service (QoS) properties and performance measures such that results obtained with the independence assumption may be misleading. We consider delays in a series of queues with correlated service times at each network node where for each customer the service time at the first node is a random variable and the successive service times are correlated with the one at the first node. A recursive scheme for delays is provided. This scheme is used in order to efficiently conduct a simulation study where two types of correlation are studied, namely identical service times, and service times with an additional Gaussian noise. The simulation study focuses on comparisons of end-to-end delays for independent service times at different nodes and correlated service times, respectively. It turns out that for both correlation types, in light traffic the delays in case of correlated service times are larger than for independent service times by a factor that first increases with increasing traffic intensity up to a maximum value approached in medium traffic after which it decreases quickly and drops down to become significantly smaller than one in heavy traffic. This effect intensifies with increasing number of network nodes and depends, as well as the crossover point from which on correlated service times yield smaller delays, on the distribution of the service times at the first node.

1. Introduction

Quality of service (QoS) provisioning and QoS guarantees are important aspects of modern computer and communication networks delivering more and more different newly emerging services. As many relevant QoS parameters are expressed in terms of network performance measures, user satisfaction is closely related to and often even essentially determined by the network performance. Among the primary QoS parameters that can be directly perceived by users are performance measures such as delays and waiting times. In particular, QoS guarantees are not possible without assuring some desired level of performance and should rely on proper studies of the network delay performance. Therefore, performance evaluation of computer and communication networks is highly important for judging the quality of service (QoS) provided by a network as well as for designing networks that appropriately meet certain possibly prescribed QoS requirements.

Queueing networks are applied for a long time as appropriate models for performance evaluation in a variety of domains. In particular, computer and communication networks have been extensively studied through queueing models. Almost always it is assumed that for each customer the service times at different network nodes are independent, but evidently this independence assumption does not appropriately reflect reality. For example, messages and packets are likely to retain their length while traveling through a network. The service times are directly proportional to the length and thereby correlated and often even equal at different nodes such that the independence assumption is significantly violated and may lead to wrong conclusions concerning the network performance. Already in the abstract and more extensively in Chapter III of his PhD thesis, Kleinrock (1962) has pointed out dependencies in communication networks but later concluded that he was 'faced with the prospect of an intractable problem' leading him to the independence assumption, cf. Kleinrock (2002).

Due to similar topologies and configurations commonly arising in many different emerging network technologies, there is currently a significantly increasing interest in the study of buffered network nodes in series, which are not only in the two node case also often referred to as tandem queueing networks or tandem

[*] Tel.: +49 53232407.
E-mail address: werner.sandmann@tu-clausthal.de

1084-8045/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2011.10.009

queues. Despite their relative simplicity, series of queues are of high practical relevance since they are suitable models for access technologies, topologies or sub-topologies as well as the structure of a variety of network components such as packet switches, routers, and many more, see, e.g., Denteneer (2006), Mandjes (2004). For instance, they have been quite recently applied to the analysis of mobile communication networks, multihop wireless networks, and traffic-groomed optical networks, see Alfa and Liu (2004), Gulpinar and Harrison (2006), Gupta and Shroff (2009), Le and Hossain (2008), Ryoki et al. (2002), Sagduyu and Ephremides (2008), Washington and Perros (2004), Washington et al. (2005), Wu and Negi (2005), Xie and Haenggi (2009).

Given the broad range of applications, delay performance analysis for series of queues is definitely desirable and justified even independently of a specific application. In fact, in queueing theory the investigation of tandem queueing networks has a long tradition starting at the latest with Reich (1957). Already important for the analysis of tandem queues, Burke (1956) studied departures from Markovian single-server queues and proved his famous theorem that the output process of an M/M/1 queueing system is a Poisson process with the same rate as the arrival process. Obviously, in a series of queues the output process of a queueing node constitutes the arrival process to the subsequent queue. Based upon Burke's theorem, in case of single-server nodes with exponentially distributed service times and Poisson arrival process, according to Weber (1979) interchanging the nodes preserves the behavior of all nodes and the overall behavior of the original system. An overview of known results for tandem queueing networks up to the late 1970s and a collection of approximation methods can be found in Newell (1979). Classical works, however, mostly considered tandem queues where for each customer the service times at the successive nodes are independent and often even more restrictive exponentially distributed. Unfortunately, this limits the usefulness of the results not only for communication networks but is also questionable in almost all application areas where items are processed by successive stages.

Most works on series of queues with correlated service times are limited to the two node case. In Boxma (1979), elaborated mathematical analyses are performed for the important special case of identical service times at both nodes. In the same setting, asymptotic results for delay distributions in the case of a service time distribution with regularly varying tails at the first node as well as heavy traffic results are given by Boxma and Deng (2000). The two node case with equal service times at both nodes is also considered by Pinedo and Wolff (1982) but detailed results for waiting times are only provided for exponential service times at the first node. Mitchell et al. (1977) and Choo and Conolly (1980) consider the two node case with Poisson arrivals and bivariate exponential customer service time distributions at the two nodes. That means, the service times at the two nodes are modeled by a joint distribution of correlated exponential pairs. In Sandmann (2007), different types of correlations including equal service times are considered for the two node case with exponential and uniform service times at the first node.

Only relatively few studies of series of more than two queues with correlated service times were reported. While Kelly (1982), Ziedins (1993) and Browning (1998) are focused on the throughput rather than concerned with delays, light traffic asymptotics for expected waiting times in a series of queues with correlated service times are investigated in Wolff (1982). However, though light traffic results are important in some scenarios, in many practical situations they are of limited interest. In particular, according to Sandmann (2010) where series of queues with identical service times are investigated, the influence of the correlations due to identical service times greatly differ for light and heavy traffic, respectively. The effects of identical service times are even contrary in these regimes, that is, in light to medium traffic identical service times increase delays as compared to independent service, whereas in heavy traffic the opposite is true. Avi-Itzhak and Levy (2001) consider series of queues with deterministically correlated service times at different nodes where the service requirements (according to message lengths or packet sizes) at all nodes are equal but correlations are present due to different speeds of the servers at different nodes. However, they do not deal with delays but with buffer sizes assuring no blocking when the arrival pattern is arbitrary and unpredictable and with the arrangement (order) of servers of different speeds.

Clearly, delay performance analysis for series of queues with correlated service times deserves further investigation, not only the case of identical service times at all nodes but also other types of correlation that appear realistic in scenarios of practical relevance.

In this paper, we consider series of queues with correlated service times and study two types of correlation and five different service time distributions. The studies of Sandmann (2010) are extended in two directions. First, in addition to the case of correlation according to identical service times we consider correlation where the service times are not identical but subject to a certain kind of Gaussian noise. Second, in addition to exponential, hyperexponential and Erlang distributed service times at the first network node (and all others in the case of iid service times) as in Sandmann (2010), we also consider two versions of Weibull distributed service times, which is particularly important with regard to appropriately modeling the service requirements in many Internet scenarios. Extensive simulation studies are provided for up to ten queueing nodes in series, where the cases of correlated service times at all nodes are compared to the case of independent and identically distributed service times at all nodes.

In Section 2, we introduce our formal notation and recapitulate the recursive scheme for delays recently applied by Sandmann (2010) in order to avoid any event list handling usually required for queueing network simulations, which enables us to study various settings in reasonable time. Section 3 describes simulation results for end-to-end delays with different service time distributions at the first node. Finally, Section 4 concludes the paper and outlines further research directions.

2. Series of queues

Consider a series of d queueing nodes as depicted in Fig. 1. New customers (packets) arrive at the first node only, proceed through the nodes in fixed order and leave the system after receiving service at the last node. No arrivals from outside occur at intermediate nodes, and no customer leaves the system before passing successively through all nodes.

It is assumed that all queues have (potentially) infinite capacities such that no blocking or loss occurs (or the probability of such events is negligible) and the major focus can be given to delays. The service discipline is First Come First Served (FCFS).

2.1. Recursive scheme for delays

For $n \in \mathbb{N}$ and $i = 1, \ldots, d$ let $S_{n,i}$ denote the service time of the n-th customer at node i, T_n the arrival epoch of the n-th customer at

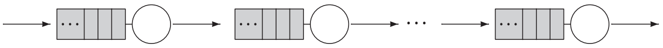

Fig. 1. Series of queues.