Applied Mathematical Modelling 45 (2017) 823–839

Contents lists available at ScienceDirect

Applied Mathematical Modelling

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

Delay analysis of multiclass queues with correlated train arrivals and a hybrid priority/FIFO scheduling discipline

Joris Walraevens*, Herwig Bruneel, Dieter Fiems, Sabine Wittevrongel

Department of Telecommunications and Information Processing (EA07), Ghent University, UGent, Sint-Pietersnieuwstraat 41, Gent B-9000, Belgium

ARTICLE INFO

Article history:
Received 8 January 2016
Revised 23 December 2016
Accepted 13 January 2017
Available online 19 January 2017

Keywords:
Queueing theory
Delay analysis
Priority
Correlated arrivals

ABSTRACT

We analyze the delay experienced in a discrete-time priority queue with a train-arrival process. An infinite user population is considered. Each user occasionally sends packets in the form of trains: a variable number of fixed-length packets is generated and these packets arrive to the queue at the rate of one packet per slot. This is an adequate arrival process model for network traffic. Previous studies assumed two traffic classes, with one class getting priority over the other. We extend these studies to cope with a general number M of traffic classes that can be partitioned in an arbitrary number N of priority classes ($1 \leq N \leq M$). The lengths of the trains are traffic-class-dependent and generally distributed. To cope with the resulting general model, an $(M \times)\infty$-sized Markovian state vector is introduced. By using probability generating functions, moments and tail probabilities of the steady-state packet delays of all traffic classes are calculated. Since this study can be useful in deciding how to partition traffic classes in priority classes, we demonstrate the impact of this partitioning for some specific cases.

1. Introduction

In this paper, we analyze a discrete-time priority queue with a particular correlated arrival process. In this introductory section, we first discuss our motivation, then list the literature on the topic and finally provide an overview of the paper and its contributions.

1.1. Motivation

Different types of network traffic (voice, video, data, ...) are generated in a different manner. As a result, their arrival characteristics in network buffers in, for instance, IP networks can be completely different. On the other hand, their Quality of Service (QoS) requirements, for instance their delay requirements, are utterly different as well. Therefore, Head-Of-the-Line (HOL) priority scheduling can be used as one of the main scheduling types in network buffers to diversify the delay of traffic streams with different delay requirements. With this scheduling discipline, packets of delay-sensitive applications get higher transmission priority than packets of applications with less stringent delay requirements. In practice, different

* Corresponding author.
E-mail addresses: Joris.Walraevens@UGent.be (J. Walraevens), Herwig.Bruneel@UGent.be (H. Bruneel), Dieter.Fiems@UGent.be (D. Fiems), Sabine.Wittevrongel@UGent.be (S. Wittevrongel).

http://dx.doi.org/10.1016/j.apm.2017.01.044
0307-904X/

network applications (traffic classes) are mapped to a given number of priority classes (not necessarily a one-to-one mapping). For instance, voice and video packets may be awarded a higher priority than data transfers.

In this paper, we study the delay of packets in a *discrete-time priority* queue with an arbitrary number of priority classes, where each priority class consists of a number of traffic classes, each with different arrival characteristics. The *arrival process* is induced by a two-layered structure. Sessions (or flows) are started and terminated by users on the higher layer. These sessions inject trains of packets in the network. Since we perform a discrete-time analysis, we assume that time is divided into slots of equal length. Packets of a session arrive to the queue at the rate of one packet per slot. This two-layered structure, generally coined *train arrivals*[1], introduces *time correlation* in the packet arrival process. Indeed, since the packets in a session arrive in consecutive slots, the number of packet arrivals in one slot depends on the number of arrivals in previous slots. Train arrival processes are an adequate choice to model e.g. the common segmentation of data files into packets before their transmission through a telecommunication network [1,2]. The model at hand is especially useful when CBR (Constant Bit Rate) traffic is involved [3,4]. In particular, the suggested arrival process is an ideal candidate to model the output buffer of a web server [5]. A web server is a computer system that accepts requests from users for a certain web page or embedded file and that responds by sending the requested file to the user. Traffic generated by a web server towards its output buffer can be described by a train-arrival process. In the case of an e-commerce web server, it makes sense to prioritize the downloads on a (potential) revenue base [6], that is, to give priority to the transmission of packets of content that is likely to provide (large) revenues. Furthermore, most web pages contain content that is delay-sensitive, for instance multimedia content. Priority can then also be given to the transmission of files containing this content over other downloads [7].

1.2. Related literature

Two important features of our model are discrete-time HOL priority and train arrivals. We first briefly discuss the state-of-the-art of models with either priority or train arrivals and later describe what has been published on models with both features.

In related literature, there have been a large number of contributions with respect to the analysis of HOL priority queues. In particular, discrete-time HOL priority queues with deterministic service times equal to one slot have been studied in [8–13]. Demoor et al. [13] study a priority queue with finite high-priority buffer space and different high-priority drop mechanisms. The steady-state buffer content and delay in case of a multiserver queue with general independent arrivals are studied in Laevens and Bruneel [8]. Mehmet Ali and Song [9] analyze the buffer content in a multiplexer with two-state on-off sources. The steady-state buffer content and the delay for Markov-modulated high-priority interarrival times and geometrically distributed low-priority interarrival times are analyzed in Takine et al. [10]. Van Velthoven et al. [12] look into the calculation of loss probabilities for different combinations of finite and infinite high- and low-priority queue-size models. Walraevens et al. [11] study the steady-state buffer content and packet delay, in the special case of an output queueing switch with Bernoulli arrivals.

Train arrival models date back to Cox [14] and First-In-First-Out (FIFO) queues with train arrivals are analyzed in [5,15–22]. Wittevrongel and Bruneel conclude a series of papers in [15] with a quite complete analysis of a FIFO queue with one traffic class, train arrivals and general message lengths. Tsoukatos and Makowski [16] use a special case of this process as a means to analyze heavy traffic limits for short-range and long-range-dependent input processes. Choi et al. [20] (Kang et al. [21] resp.) use beta-distributed (gamma-distributed resp.) arriving batch sizes and determine the train length distribution which leads to a Weibull-like autocorrelation function in order to model video traffic. Feyaerts et al. [22] analyze a FIFO queue with train arrivals, geometric train lengths and an unreliable output line subject to Markovian output interruptions. In all these papers, packets within trains arrive at the pace of one packet per slot. This packet generation process is later generalized to an independent random process with a strictly positive number of packets per slot [5,17]. This is even further generalized to a FIFO queue with multiple classes [19]. The assumption that the number of generated packets in a train is strictly positive is crucial in the analyses of these papers. To the best of our knowledge, only one attempt is made to include the possibility of no packet arrivals in a slot, i.e., to allow for gaps between consecutive packets in one train [18]. The authors propose an approximation based on a Taylor-series expansion in one of the parameters of the model. Finally, we remark that somewhat related finite-population on/off-type arrival models are considered in [23–25], also for the FIFO case. For an extensive overview of train-arrival models, we refer to our paper [26].

Some papers have been published on models combining HOL priority queues and train arrivals. The first analysis appeared in [27], treating a two-class model with deterministic session lengths. A second analysis was of a two-class model with geometric session lengths [28]. A further generalization of these two models was treated in [26,29]; here, arbitrary distributions were allowed for the session lengths.

1.3. Contributions and overview

In the current paper, we extend previous analyses to a discrete-time priority queue with train arrivals, *generally* distributed session lengths and an *arbitrary* number of traffic classes. Furthermore, the *mapping of traffic classes to priority*

[1] Alternative names are session arrivals or $M/G/\infty$-input, albeit with some subtle differences in definition.