Computers & Industrial Engineering 94 (2016) 269–279

Contents lists available at ScienceDirect

Computers & Industrial Engineering

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

Asymptotic optimality of the queue service probability for the radial basis function network-based queue selection rule

Mu-Song Chen *, Hao-Wei Yen

Department of Electrical Engineering, Da-Yeh University, Changhua 51591, Taiwan, ROC

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 18 September 2015
Accepted 4 February 2016
Available online 22 February 2016

Keywords:
Radial basis function network
Queue service probability
Waiting time
Queue selection rule
Overload

ABSTRACT

A queueing model is generally designed with sufficient capacity or resources to ensure that the system is stable, while preserving quality of service. However, the multi-queue system with finite capacity and timing constraints in an overload condition are more often encountered and discussed in a variety of real-life problems. In such a situation, waiting time is usually an important performance metric quantifying the effectiveness and efficiency of the system. The concerned issue is still an open research topic and is not fully addressed and investigated. Since an exact analysis is practically infeasible owing to the complexity of such systems, emphasis has been concentrated on the approximate analysis. This paper is thus intended to estimate the upper bound of waiting times of a multi-queue system with a specialized scheduling paradigm, extending from a series of our research on message scheduling. Without resorting to complex statistical approaches, the study provides a machine learning methodology to resolve this subject. With the learning capability of the radial basis function network (RBFN) as the queue selection rule, this paper particularly focuses on deriving the asymptotic optimality of the queue service probability, under the conditions of multi-queue, finite capacity, and timing constraints in the overload situation. In fact, the RBFN is incorporated with two novel types of learning which lead to develop the support theorem and to obtain the closed-form of queue service probability as well as waiting time. Importantly, the learning feature is definitely essential in providing optimal queue service probability with dynamical scheduling scheme. Several existing queue selection rules are also evaluated and compared with the RBFN-based queue selection rule. Simulation results illustrate the feasibility and accuracy of the proposed strategy.

1. Introduction

Queueing systems are popular and widely used tools for analyzing the performance of communication networks, transportation system (Sahba & Balcioglu, 2011), job-shop manufacturing system (Georgiadis & Michaloudis, 2012; Shen & Buscher, 2012), and machine-repairman system (Yuan, 2012). Usually, many queueing systems are generally designed with sufficient capacity or resources to ensure that the system is stable, while preserving quality of service (QoS). However, variability and constraints, e.g. heavy traffic, timing constraints, finite queue capacity, etc., in a queueing system may have significant impacts on its performance. The study of heavy traffic began in the 1960s by Kingman (1962). Kingman's bound implies that even small variations in service times or inter-arrival times can lead to significant delay. Apart from this, it is inevitable that a system can also encounter a periodic or temporary overload,[1] either due to an unpredictable increase in load, unavailable service resources, or system breakdown. Another practical situation occurs on message scheduling in the controller area network (Mubeen, Mäki-Turja, & Sjödin, 2014). In such a context, the maximum normal utilization of the controller area network is about 30%. Under heavy disturbance, it rises to about 40% due to retransmission of corrupted messages (Johansson, Törngren, & Nielsen, 2005). The direct consequence can possibly bring about a fatal catastrophe or even a matter of life and death. The aforementioned problems become worse when the queueing system is conditioned by finite capacity and/or imposed on stringent timing constraints.

The finite-capacity queueing systems with timing constraints have been widely studied by many researchers (Osorio &

* Corresponding author.
E-mail addresses: chenms@mail.dyu.edu.tw (M.-S. Chen), chenmsac1510@yahoo.com.tw (H.-W. Yen).

[1] The load is a crucial indicator of whether the queueing system will grow beyond its capacity. The occurrence in a heavy load situation can temporarily make the system unstable.

http://dx.doi.org/10.1016/j.cie.2016.02.005
0360-8352/

Nomenclature

τ_i	queue cycle of queue Q_i	ρ_i	message load of queue Q_i
γ_i	queue service probability of queue Q_i	D_i	deadline of messages in queue Q_i
$W_i, \overline{W}_i$	waiting time and mean waiting time of queue Q_i	K_i	maximum number of messages in queue Q_i
λ_i, μ_i	arrival rate and service rate of queue Q_i		

Bierlaire, 2009; Sakuma & Inoie, 2014; Wu, Lin, & Chie, 2012) and our recent publications (Chen, 2015; Chen & Yen, 2011, 2012a, 2012b, 2013, 2014). A queueing system with these restrictions occurs frequently in our real life. For instance, there may be a surge in demand for the limited hospital emergency rooms during a catastrophic event (Lakshmi & Sivakumar, 2013). In manufacturing systems, there are limited waiting rooms between workstations in assembly lines. In semiconductor wafer fabrication manufacturing, queue-time constraints occur when a set of consecutive process steps must be completed within a fixed time window (Tsai, 2008). Violations of these constraints can result in rework, scrap, and longer cycle times. To ensure product quality, the queue time between consecutive operations is often required to be shorter than a pre-specified duration. The effect is similar to imposing a limited buffer size between operations. On highways, there are rush hours during which drivers have to queue for the use of highways. Regardless of how many lanes already exist, it has finite road capacities and is vulnerable to possible congestion during weekend or holidays. In all these circumstances, the system can have excessive demands than it can handle and possibly leads to unstable levels or causes unpredictable performance degradation.

Regarding aforementioned problems, three primary considerations are (1) how the system can be stabilized, (2) how the system throughput is maximized, and (3) how the upper bounds of waiting times can be estimated. Pertaining to issues (1) and (2), several related literatures aim to avoid the degradation of throughput of the server as well as to maintain the system at its steady state. Schemes such as admission control (Cherkasova & Phaal, 2002) or specialized scheduling policies (Chen & Mohapatra, 2002), or a combination of both (Elnikety, Nahum, Tracey, & Zwaenepoel, 2004) have been studied. Furthermore, Solar (Stolyar, 2004) analyzed the parallel server queueing systems, under the assumption that the system is in heavy-traffic. That is, the system is stable but operates close to the boundary of the stability region. Shakkottai, Srikant, and Stolyar (2004) and Bell and Williams (2005) also used the complete resource pooling condition to establish heavy-traffic optimality of other resource allocation models. Other researchers (Andrad'ottir & Ayhan, 2005; Andrad'ottir, Ayhan, & Down, 2001, 2003; Tassiulas & Bhattacharya, 2000) concerned with the dynamic assignment of servers to maximize the system throughput in queueing networks. In addition, Venkataraman and Lin (2007) proposed scheduling algorithms that can stabilize the network at given offered loads, which also ensures that the long term average and service rate are no less than the arrival rate of each user.

Regarding issue (3), let us consider another example of the hospital queueing systems. Derlet, Richards, and Kravitz (2001) and Carter and Lapierre (1991) pointed out that the causes of hospital overcrowding include low staff availability, bed shortage, insufficient hospital space, and so forth. Actually, long waiting time is symptomatic of inefficiency and indicative of the system's inability to satisfy patients' demand. To ensure QoS, several criteria can be established for a satisfactory level of service. One possible gauge might be that the upper limit of waiting time should not exceed a certain value. Therefore, patients' satisfactions are often related to the amount of time they have to wait before receiving treatments. For the mentioned criterion, the upper bound of waiting time needs to be known first for this purpose. This problem is principally related to issue (3) and is vital when the queueing system is imposed by timing constraints and with restrictions of finite capacity. In fact, a search of available literatures indicates that relatively few studies have been devoted to fully address this problem. Furthermore, due to the growth of traffic with the increasing amount of data exchanged between electronic control units in the controller area network, our motivation of this study aims at estimating the message waiting times in an overloaded circumstance.

Based on our research findings, this paper presents a machine learning approach with the radial basis function network (RBFN) (Chen & Yen, 2011) in dealing with the aforementioned issue. The RBFNs have demonstrated their usefulness in a variety of applications, including classification, prediction, and system modeling. In fact, the RBFN is a variant of the multilayer perceptron networks (Haykin, 1994) with single hidden layer and makes use of locally supported functions to calculate the Euclidian distance between the input vector and the centers of RBFs. The latest studies and applications on the RBFNs can be refereed to Dehghan and his colleagues (Dehghan & Mohammadi, 2014, 2015; Dehghana, Abbaszadeha, & Mohebbib, 2014, 2015; Ilati & Dehghan, 2015). With the radial basis function network as the queue selection rule (QSR), messages in queues are selected for admission into the service facility in accordance with the decision made by the RBFN. To the best of our knowledge, this study is the first attempt to combine machine learning strategy with dynamical scheduling to estimate waiting times under the overload situation. Especially, the proposed RBFN-based QSR directly leads to derive the asymptotic optimality of the queue service probability (QSP) as well as waiting time in the time domain directly.

The remaining part of the paper is organized as follows. Section 2 introduces the definition of queue cycle, the effective arrival rate, and the mean queue length for subsequent analysis. In Section 3, relationship between waiting time and mean waiting time is established. Section 4 is devoted to derive the queue service probability of the RBFN-based queue selection rule with support theorem. Section 5 reports and validates the results of simulation experiments. Finally, Section 6 concludes the paper and suggests our future research directions.

2. The queue cycle

Typically, the queueing system consists of a number of parallel queues attended by a single server. Thus, the control policy of the QSR focuses on determining the queue and events in that queue for admission into the service facility. Once a queue is selected for service, events within that queue are scheduled using first-come-first-served order. At this point the word "event" is a generic expression that represents many real-world phenomena, such as airplanes arriving to an airport, shoppers in a grocery store, or messages in a queue waiting to be executed. In this study, we focus on the issue of message scheduling. We also consider a multi-class[2] model, consisting of q finite-capacity queues, labeled as $Q_1, Q_2, \ldots, Q_q$. Usually,

[2] Multi-class queuing model can be used to model complex service systems which have different QoS requirements.