European Journal of Operational Research 256 (2017) 473–486

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Stabilizing performance in a service system with time-varying arrivals and customer feedback

Yunan Liu [a,*], Ward Whitt [b]

[a] Department of Industrial Engineering, North Carolina State University, Raleigh, NC 27695-7906, United States
[b] Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, United States

ARTICLE INFO

Article history:
Received 1 June 2015
Accepted 12 July 2016
Available online 8 August 2016

Keywords:
Queueing
Staffing algorithms for service systems
Time-varying arrival rates
Queues with feedback
Stabilizing performance

ABSTRACT

Analytical offered-load and modified-offered-load (MOL) approximations are developed to determine staffing levels that stabilize performance at designated targets in a non-Markovian many-server queueing model with time-varying arrival rates, customer abandonment from queue and random feedback with additional feedback delay in an infinite-server or finite-server queue. To provide a flexible model that can be readily fit to system data, the model has Bernoulli routing, where the feedback probabilities, service-time, patience-time and feedback-delay distributions all are general and may depend on the visit number. Simulation experiments confirm that the new MOL approximations are effective. A many-server heavy-traffic FWLLN shows that the performance targets are achieved asymptotically as the scale increases.

1. Introduction

This paper is part of an ongoing effort to develop effective methods to set staffing levels (the time-dependent number of servers) in service systems with time-varying arrival rates in order to stabilize performance at designated targets; see Green, Kolesar, and Whitt (2007) for a review and Stolletz (2008), Defraeye and van Nieuwenhuyse (2013), Liu and Whitt (2014), Yom-Tov and Mandelbaum (2014) and He, Liu, and Whitt (2016) for recent related work. We continue to focus on service systems that can be modeled as many-server queues with customer abandonment from queue and non-exponential distributions, but here in addition we consider Bernoulli feedback with additional delay after completing service.

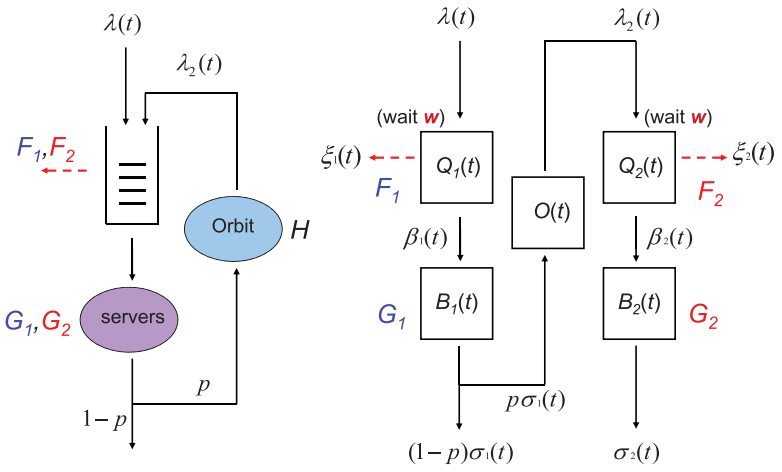
A queue with delayed feedback after completing service is a special queueing network, about which there is an enormous literature, but our concern is with the time-dependent performance of a non-stationary non-Markov model, which is well beyond exact analysis. We do assume that the arrival process is a *nonhomogeneous Poisson process* (NHPP), but the service-time, patience-time and delay-before-return distributions all can be non-exponential and can change upon successive feedbacks. At first glance, it would seem that previous methods do not apply to the generalization with multiple delayed feedbacks having changing parameters. Our main innovation is to propose an approximation involving a series of infinite-server models. Instead of the natural two-queue model for a single delayed feedback shown on the left in Fig. 1, with an orbit queue for the customers experiencing extra delay in addition to the usual queue, we propose the five-queue series model on the right, which also has separate queues for the customers waiting and in service upon first visit and upon second visit, as well as for those customers being delayed in between the two visits; we elaborate on the model below.

Previous work has shown that a time-varying arrival-rate function and a non-exponential service-time distribution can have a significant impact on performance; see Eick, Massey, and Whitt (1993) for discussion of the basic $M_t/GI/\infty$ infinite-server special case. Figs. 1 and 2 of Jennings, Mandelbaum, Massey, and Whitt (1996) dramatically show the poor performance that can occur if we use stationary methods to set staffing levels, either using the overall average arrival rate or using the pointwise-stationary approximation (PSA), which uses a stationary model in a nonstationary way, letting the arrival rate in the stationary model at time t be the actual arrival rate $\lambda(t)$ at time t. As reviewed in Green et al. (2007), the PSA can be effective with relative short service times, but tends to fail badly with longer service times. The additional delayed feedback adds to the challenge because it can significantly alter the time-varying demand, not only in magnitude but also in timing. For example, the delayed feedback can amplify or damp the peak demand and shift it in time.

The literature exposes two common reasons for feedback after completing service: First, de Vericourt and Zhou (2005)

* Corresponding author.
E-mail addresses: yliu48@ncsu.edu, yunan_liu@ncsu.edu (Y. Liu), ww2040@columbia.edu (W. Whitt).

http://dx.doi.org/10.1016/j.ejor.2016.07.018
0377-2217/



(a) The $(M_t/GI,GI/s_t + GI,GI) + (GI/\infty)$ Model with Delayed Customer Feedbacks

(b) The Delayed Infinite-Server (DIS) Approximation

Fig. 1. The $(M_t/\{GI,GI\}/s_t + \{GI,GI\}) + (GI/\infty)$ model with delayed customer feedback and its Delayed Infinite-Server (DIS) approximation. The approximating offered load is $m(t) = m_1(t) + m_2(t) \equiv E[B_1(t)] + E[B_2(t)]$.

focus on call center customers that may return later because the initial service was unsatisfactory. Second, Yom-Tov and Mandelbaum (2014) focus on the treatment of patients by a doctor in a hospital that may naturally occur in stages, starting with an initial screening and continuing later after tests have been ordered and completed. Our paper is closely related to Yom-Tov and Mandelbaum (2014), where a *modified-offered-load* (MOL) approximation was proposed to help set staffing levels at a queue with time-varying arrival rates and Markovian feedback after a delay in an *infinite-server* (IS) queue. They showed that the MOL approximation has great potential for improved performance analysis in healthcare, where the service times tend to be relatively long, so that PSA does not apply.

Motivated by these applications, we consider a feedback model that has appealing flexibility. In particular, instead of the Markovian routing with fixed feedback probability p and one fixed service-time distribution considered in Yom-Tov and Mandelbaum (2014), we consider history-dependent Bernoulli routing, where there may be any number of visits and the feedback probability p and the service-time distribution and the subsequent delay distribution (before returning for a new service) all may vary with the visit number. We focus on the common important case of at most one feedback, which seems to be a more realistic model than Markovian routing, which produces a geometric random number of feedbacks. It is significant that the approach here also extends directly to any finite number of feedbacks; we demonstrate by also considering examples with two feedback opportunities. Our methods also extend directly to time-dependent feedback probabilities, but we do not examine that here. (The justification is that a time-dependent independent thinning of an NHPP is again an NHPP; see Sections 2.3 and 2.4 of Ross (1996).)

We also allow customer abandonment, which often tends to be more realistic for many service systems, as observed by Garnett, Mandelbaum, and Reiman (2002). The patience-time distributions are also allowed to be non-exponential and depend on the visit number. Just as in Yom-Tov and Mandelbaum (2014), we use the general offered-load (OL) method with the MOL refinement, as reviewed in Jennings et al. (1996), Green et al. (2007), Liu and Whitt (2012c) and Whitt (2013). There are difference between the MOL methods designed to stabilize the delay probability and the abandonment probability, as discussed in Liu and Whitt (2012c), but the main contribution here beyond Yom-Tov and Mandelbaum (2014) is the new method for computing the time-varying offered load. Because the offered load is the primary determinant of performance, the performance impact from more faithfully representing the service and feedback process in a time-varying setting can be great.

To analyze this new feedback model with customer abandonment, we draw on Liu and Whitt (2012c) in which we developed a *delayed-infinite-server* (DIS) offered-load approximation and a new DIS-MOL (*DIS-modified-offered-load*) algorithm to determine time-dependent staffing levels in order to stabilize expected delays and abandonment probabilities at specified *quality of service* (QoS) targets in a many-server queue with time-varying arrival rates. The model in Liu and Whitt (2012c) was $M_t/GI/s_t + GI$ model, having arrivals according to an NHPP with arrival rate function $\lambda(t)$, independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), a time-varying number of servers (the s_t, to be determined), i.i.d. patience times with a general distribution (times to abandon from queue, the final $+GI$), unlimited waiting space and the first-come first-served (FCFS) service discipline. We included non-exponential service and patience distributions as well as time-varying arrivals because they commonly occur; e.g. see Armony et al. (2015) and Brown et al. (2005).

We refer to the base model with a single feedback considered here as $(M_t/\{GI,GI\}/s_t + \{GI,GI\}) + (GI/\infty)$. The main queue has the two service-time cdf's G_i and patience cdf's F_i, depending on the visit number, while the orbit queue has a single service-time cdf H, with all waiting customers entering service in a FCFS order. We develop approximations for the number of customers waiting before service and in service upon each visit and the number of customers in orbit. When we refer to the number of customers in the system or the waiting time, we do not include the orbit queue.

We also consider the associated $(M_t/\{GI,GI\}/s_t + \{GI,GI\}) + (GI/s_t + GI)$ model in which the orbit queue has finite capacity; in that case, it also has a staffing function and a patience distribution. The goal is to stabilize expected potential waiting times (the virtual waiting time before starting service on any visit of an arrival with infinite patience) at a fixed value w for all time and $i = 1, 2$. Since these models are special kinds of two-class queueing models, we also consider the more elementary $\sum_{i=1}^{2}(M_t/GI + GI)/s_t$ two-class queue, in which the two classes arrive according to two independent NHPPs with arrival rate functions $\lambda^{(i)}(t)$ and their own service-time cdf's G_i and patience cdf's F_i, $i = 1, 2$, but there is a single service facility with a time-varying number of servers $s(t)$, again to be determined.

The approximating DIS model for the $(M_t/\{GI,GI\}/s_t + \{GI,GI\}) + (GI/\infty)$ feedback queue has five IS queues in