An $M/M/s$-consistent diffusion model for the $GI/G/s$ queue

Toshikazu Kimura

Department of Business Administration, Hokkaido University, Nishi 7, Kita 9, Kita-ku, Sapporo 060, Japan

Received 22 July 1994; revised 19 December 1994

This paper develops a diffusion-approximation model for a stable $GI/G/s$ queue: The queue-length process in the $GI/G/s$ queue is approximated by a diffusion process on the nonnegative real line. Some heuristics on the state space and the infinitesimal parameters of the approximating diffusion process are introduced to obtain an approximation formula for the steady-state queue-length distribution. It is shown that the formula is consistent with the exact results for the $M/M/s$ and $M/G/∞$ queues. The accuracy of the approximations for principal congestion measures are numerically examined for some particular cases.

Keywords: $GI/G/s$ queues, diffusion model, conditional departure process, renewal-process approximations, consistent discretization.

1. Introduction

In some infrastructural service systems of transportation and communication, the theory of multi-server queues has been a useful tool of modeling congestion phenomena therein. It is, however, difficult to analyze multi-server queues with generally distributed service times, except for the exponential case that scarcely fit in with the reality. To deal with this difficulty, we often need approximations.

The purpose of this paper is to exploit a new diffusion-approximation model for a general multi-server queueing system. We consider the standard $GI/G/s$ queueing system with $s$ homogeneous servers in parallel, unlimited waiting spaces, the first-come first-served discipline, and i.i.d. service times which are independent of a renewal arrival process. Many researchers have proposed numerous approximate methods for the Poisson arrival case, i.e., the $M/G/s$ queue. For the $GI/G/s$ case, however, possible approaches are quite limited and essentially heuristic by nature; see Kimura [16] for a comprehensive list of the literature.

It has been known that diffusion approximations in heavy traffic can be often justified by heavy traffic limit theorems (HTLTs): For a queueing characteristic process such as queue length, waiting time and so on, HTLTs describe that the process in an unstable queue, when appropriately scaled and translated,
converges weakly to a Brownian motion process; see Glynn [7] for a survey. However, this does not necessarily imply that the Brownian motion process still gives an accurate approximation when the queue is stable. In particular, for multi-server queues, some additional heuristics are required in the situation that not all the servers are continuously busy, which cannot be described by the HTLTs. Hence, we must clearly distinguish two kinds of "diffusion approximations", the one in the sense of an HTLT for the unstable queue and the other in the sense of a diffusion model for the stable queue. This paper deals with the latter approximation.

There have been various diffusion models developed for multi-server queues, due to the differences in the additional heuristics: Halachmi and Franta [8] and Sunaga et al. [25] proposed a basic diffusion model for the GI/G/s queue, which was motivated by a continuous approximation for the M/M/s queue. For the M/G/s queue, Kimura [12] exploited a diffusion model of which formulation has become a standard in the subsequent diffusion models. Some modifications of this model have been proposed by Tien and Jansen [27], Wu [31] and Yao [33]; see also Choi and Shin [3] for its transient analysis. For more general or extended multi-server queues, many diffusion models also have been developed by Kimura and Ohsono [17], Ohsono [22] and Wu and Wang [32] for group arrivals; by Filipiak and Pach [6], Sunaga et al. [26] and Yao and Buzacott [34] for queues with finite waiting spaces; and by Biswas and Sunaga [2], Jain and Sharma [11], Kimura [15], Sivazlian and Wang [24] and Varshney et al. [28] for queues with some state dependency.

A serious defect in the previous diffusion models is that they are not consistent with available exact results for particular cases; see Kimura [13] and Whitt [29]. For example, none of the previous diffusion approximations for the queue-length distribution in the GI/G/s queue are consistent with any exact results, even with the M/M/s queue. It is obvious that the lack of consistency makes diffusion models unreliable with respect to the accuracy. This paper develops a refined diffusion model for the GI/G/s queue, which is consistent with the M/M/s and M/G/∞ queues.

Here is how this paper is organized: In section 2, we introduce a heuristic that relates a value of the number of customers in the system to an interval in the state space of the approximating diffusion process. In section 3, we also introduce a renewal-process approximation for a conditional departure process to obtain diffusion parameters on these intervals. Using these heuristics, in section 4 we derive a diffusion approximation formula for the steady-state queue-length distribution that is consistent with the exact results for the M/M/s and M/G/∞ queues. For some particular cases, section 5 numerically examines the accuracy of the approximations for the mean queue length and the probability that all the servers are busy. We also numerically compare our diffusion approximations with the previous ones. Finally, section 6 discusses further refinements in our model by the use of a conservation law.