1995年度日本オペレーションズ・リサーチ学会
春季研究発表会

A Refined Diffusion Approximation for Finite-Capacity Multi-Server Queues

01105381 北海道大学 木村 俊一 KIMURA TOSHIKAZU

1 Introduction

Queues with finite waiting spaces have been useful models of computer, communication and manufacturing systems experiencing congestion due to irregular flows. The limited waiting room corresponds to a local storage or buffer for waiting customers (*i.e.*, jobs, packets, transactions, etc.). In particular, the local storage at a work station in a flexible manufacturing system (FMS) typically has a small number of waiting spaces. The FMS work station also typically has a set of parallel machines with generally distributed processing times, and hence it can be adequately modeled as a finite-capacity $GI/G/s$ queue. In this paper, we develop and evaluate a refined diffusion approximation for the $GI/G/s/s+r$ queue, which is consistent with the exact results for the $M/G/s/s$ and $M/M/s/s+r$ queues.

2 Basic Assumptions on the Diffusion Model

The $GI/G/s/s+r$ queueing system we consider is specified by the following assumptions: Let F (G) denote the interarrival-time (service-time) cumulative distribution function (CDF) with mean λ^{-1} (μ^{-1}), and let c_a^2 (c_s^2) be the squared coefficient of variation (SCV, *i.e.*, variance divided by the square of the mean) of F (G). Let $\rho = \lambda/s\mu$ be the traffic intensity and assume that the system is in steady state. In addition, let $A(t)$, $D(t)$ and $L(t)$ denote the cumulative numbers of arrivals, departures (*not* counting lost customers) and lost customers during the time interval $(0, t]$, respectively. Then, the number of customers at time t (≥ 0), say $N(t)$, can be represented as

$$N(t) = N(0) + A(t) - D(t) - L(t), \quad t \geq 0. \quad (1)$$

The fundamental idea of diffusion approximations for finite-capacity queues is to approximate the discrete-valued process $\{N(t);\, t \geq 0\}$ by an appropriate time-homogeneous diffusion process $\{X(t);\, t \geq 0\}$ on a finite subset of $\mathbb{R}_+ = [0, \infty)$, utilizing asymptotic properties of the counting processes $A(\cdot)$, $D(\cdot)$ and $L(\cdot)$ in (1).

We use the generic random variable N (N^-) to indicate the number of customers in the system at an arbitrary time (just before an arrival epoch) in equilibrium. For $k = 0, \ldots, s+r$, let $p_k = P(N = k)$ and $\pi_k = P(N^- = k)$.

A first step of the diffusion modeling is to define an interval $\mathcal{I}_k$ of $\mathbb{R}_+$ corresponding to the event $\{N = k\}$ ($k = 0, \ldots, s+r$). We suggest using the set of intervals defined by

$$\mathcal{I}_k = \begin{cases} \{0\}, & k = 0 \\ (x_{k-1}, x_k], & k = 1, \ldots, s+r \end{cases}$$

for an increasing sequence $0 = x_0 < x_1 < \cdots < x_{s+r}$. To regulate the process $X(\cdot)$ in the interval $[0, x_{s+r}]$, we assume that each of the boundaries is *reflecting.*

Let $dX(\tau) = X(\tau) - X(0)$ for $\tau > 0$. Then, apart from the boundary behavior, the diffusion process $X(\cdot)$ can be characterized by the limits

$$b(x) = \lim_{\tau \to 0} \frac{1}{\tau} E[dX(\tau) \mid X(0) = x]$$

$$a(x) = \lim_{\tau \to 0} \frac{1}{\tau} E[\{dX(\tau)\}^2 \mid X(0) = x]$$

for $x > 0$. Taking account of the natural correspondence between the event $\{N = k\}$ and the interval $\mathcal{I}_k$ ($k = 1, \ldots, s+r$), we assume that each of these parameters is *piecewise constant*, *i.e.*, for $x \in \mathcal{I}_k$ ($k = 1, \ldots, s+r$),

$$b(x) = b_k \quad \text{and} \quad a(x) = a_k,$$

where $\{b_k;\, k \geq 1\}$ and $\{a_k;\, k \geq 1\}$ are bounded sequences and $a_k > 0$ for all k.

3 General Distribution Form

Using a pointwise discretization method [1] developed for the case $r = \infty$, we can express $\{p_k\}$ as

$$p_k = p_0 \xi_k, \quad k = 0, \ldots, s+r-1, \qquad (2)$$

for a sequence $\{\xi_k\}$ specified by $\{b_k\}$, $\{a_k\}$ and $\{x_k\}$. To express the probabilities p_0 and p_{s+r}

in terms of $\{\xi_k\}$, we use a rate conservation law as follows: Since the average rate of accepted arrivals equals the average departure rate (not counting lost customers), we have

$$\lambda(1-\pi_{s+r}) = \mu E[\min(N,s)],$$

from which π_{s+r} can be written as

$$\pi_{s+r} = \frac{1}{\rho}\left\{\rho - 1 + p_0 \sum_{k=0}^{s-1}\left(1-\frac{k}{s}\right)\xi_k\right\}.$$

To obtain an approximation for p_{s+r}, we utilize an exact result for the $GI/M/s/s+r$ queue, namely,

$$\pi_{s+r} = z p_{s+r}, \tag{3}$$

where the coefficient z is given by

$$z = \frac{\phi(s\mu)}{\rho(1-\phi(s\mu))},$$

and $\phi(\cdot)$ denotes the LST of the CDF F. In this paper, we use the formula (3) for the $GI/M/s/s+r$ queue as an approximation for the $GI/G/s/s+r$ queue. In particular, for the $M/G/s/s+r$ queue, we see that this approximation, $z=1$, is correct because of the PASTA property. Substituting (2) and $p_{s+r} = \pi_{s+r}/z$ into the normalizing condition $\sum_{k=0}^{s+r} p_k = 1$, we obtain

$$p_0 = \frac{\rho(z-1)+1}{\displaystyle\sum_{k=0}^{s-1}\left(\rho z + 1 - \frac{k}{s}\right)\xi_k + \rho z \sum_{k=s}^{s+r-1}\xi_k}.$$

4 Diffusion Approximation with Consistent Discretization

Here we summarize the final results for $\{p_k\}$:

$$p_k = \begin{cases} p_0 \xi_k, & k=1,\ldots,s-1 \\ p_0 \xi_s \hat{\rho}^{k-s}, & k=s,\ldots,s+r-1 \\ \dfrac{1}{\rho z}\left\{\rho - 1 + p_0 \displaystyle\sum_{j=0}^{s-1}\left(1-\frac{j}{s}\right)\xi_j\right\}, & k = s+r, \end{cases}$$

where the empty probability p_0 is given by

$$p_0 = \begin{cases} \dfrac{\rho(z-1)+1}{\displaystyle\sum_{k=0}^{s-1}\left(\rho z + 1 - \frac{k}{s}\right)\xi_k + \frac{1-\hat{\rho}^r}{1-\hat{\rho}}\rho z \xi_s}, & \rho \neq 1 \\ \dfrac{z}{\displaystyle\sum_{k=0}^{s-1}\left(z+1-\frac{k}{s}\right)\xi_k + r z \xi_s}, & \rho = 1 \end{cases}$$

$$\xi_k = \frac{1}{a_k}\prod_{j=1}^{k}\left(\frac{a_j^*}{a_{j-1}^*}\frac{s\rho}{j}\right)^{\alpha_j}, \quad k=1,\ldots,s,$$

and

$$a_k^* = \lambda + k\mu, \quad k=1,\ldots,s.$$

The infinitesimal variance $\{a_k\}$ is given by

$$a_k = \begin{cases} \lambda c_a^2 + k\mu, & k=1,\ldots,s-1 \\ \lambda c_a^2 + k\mu\left\{\rho^2 c_s^2 + (1-\rho^2)c_{ds}^2(\text{SIM})\right\}, & k=s, \end{cases}$$

where

$$c_{ds}^2(\text{SIM}) = 2s\mu\int_0^\infty \{1-G_e(t)\}^s\,dt - 1,$$

and G_e is the stationary-excess CDF associated with the service-time CDF G, *i.e.*,

$$G_e(t) = \mu\int_0^t \{1-G(u)\}du, \quad t\geq 0.$$

The parameter α_k $(k=1,\ldots,s)$ is defined by $\alpha_k = a_k^*/a_k$ and $\hat{\rho} = \rho^{\alpha_s}$. See Kimura [2] for details.

Using the approximate distribution $\{p_k\}$, we can derive approximation formulas for some congestion measures in the $GI/G/s/s+r$ queue: Let $Q = \max(N-s, 0)$ be the queue length excluding customers in service, and let W denote the waiting time of a customer who is allowed to enter the system. Then, the mean queue length is

$$E[Q] = \begin{cases} p_0\dfrac{\hat{\rho}}{(1-\hat{\rho})^2}\left\{1-\hat{\rho}^r - r(1-\hat{\rho})\hat{\rho}^{r-1}\right\}\xi_s + r p_{s+r}, & \rho\neq 1 \\ \dfrac{1}{2}p_0 r(r-1)\xi_s + r p_{s+r}, & \rho = 1 \end{cases}$$

By virtue of Little's formula, the mean waiting time $E[W]$ can be derived from $E[Q]$ as

$$E[W] = \frac{E[Q]}{\lambda(1-\pi_{s+r})}.$$

References

[1] KIMURA, T., An $M/M/s$-Consistent Diffusion Model for the $GI/G/s$ Queue, Discussion Paper, Faculty of Economics, Hokkaido University, Sapporo (1994).

[2] KIMURA, T., A Refined Diffusion Approximation for Finite-Capacity Multi-Server Queues, Discussion Paper, Faculty of Economics, Hokkaido University, Sapporo (1994).