A Brownian model for multiclass queueing networks with finite buffers

Wanyang Dai[1]

Department of Mathematics, Nanjing University, Nanjing 210008, China

Abstract

This paper is concerned with the heavy traffic behavior of a type of multiclass queueing networks with finite buffers. The network consists of d single server stations and is populated by K classes of customers. Each station has a finite capacity waiting buffer and operates under first-in first-out (FIFO) service discipline. The network is assumed to have a feedforward routing structure under a blocking scheme. A server stops working when the downstream buffer is full. The focus of this paper is on the Brownian model formulation. More specifically, the approximating Brownian model for the networks is proposed via the method of showing a pseudo-heavy-traffic limit theorem which states that the limit process is a reflecting Brownian motion (RBM) if the properly normalized d-dimensional workload process converges in distribution to a continuous process. Numerical algorithm with finite element method has been designed to effectively compute the solution of the Brownian model (W. Dai, Ph.D. thesis (1996); X. Shen et al. The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks, under revision for Queueing Systems). © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Multiclass queueing network; Finite buffer; Network blocking; Heavy traffic; Reflecting Brownian motion (RBM)

1. Introduction

Queueing network models with finite buffers provide powerful and realistic tools for performance evaluation and prediction of discrete flow systems such as computer architectures, communication networks and manufacturing systems. In many applications, buffer constraints have a major impact on system performances (e.g., [8] or [11]). Thus, it is imperative to model the finiteness of the buffer sizes in these networks. Another important feature of the networks is that stations can process more than one class of customer or job (the so-called multiclass networks, if stations can only process

---

one class of job, the networks are called single class networks). Heavily loaded networks, where congestion and blocking are compelling problems, are of particular interest. Frequently, exact analysis of such networks is unavailable and it is natural to seek tractable approximations. In connection with this, Brownian models have been employed to approximate normalized versions of the queue length or workload processes in certain single class queueing networks with finite buffers (e.g., [4,5]) and many networks with infinite buffers (e.g., [1,12,2]) under conditions of heavy traffic. The Brownian models use only the first two moments of the interarrival and service times, routing vectors and blocking mechanism. Furthermore, for the Brownian models living in a hypercube of a multi-dimensional Euclidean space, many quantities including the stationary distribution can be computed numerically [4,13]. Therefore, one can obtain performance estimates for the networks with finite buffers, like blocking probabilities and average queue lengths, from their Brownian counterparts [4,13]. One of the challenges in contemporary research on queueing networks is to identify which classes of multiclass queueing networks with finite buffers can be so approximated.

In this paper, we propose a Brownian network model for an open multiclass queueing network with finite buffers. The queueing network has d single server stations and each station has a finite capacity waiting buffer. The network is populated by K classes of customers, and each class k has its own external arrival process with independent and identically distributed (i.i.d) interarrival times. The external arrival processes among different classes are assumed to be independent. The class k customers require service at station $h(k)$ and their service times are i.i.d. The service time sequences for various customer classes are independent each other and are also independent of the arrival processes. The order of customers served at a station is supposed to be first-in first-out (FIFO). Upon completion of service, all classes of customers at a station will be routed to an immediate downstream station, and a class k customer will become a class l customer at the downstream station with probability p_{kl}. All customers eventually leave the network. Hence the network is an open multiclass queueing network with a feedforward routing structure.

Although many blocking mechanisms can be employed for a finite buffer network, we will focus on the "block-and-hold-0" mechanism. Under such a blocking mechanism, a server will stop working whenever an immediate downstream buffer is full. Therefore, the number of blocked customers that have completed services is 0. Readers are referred to Cheng and Yao [3] for the definition of the general "block-and-hold-k" mechanism.

For each $j = 1, \ldots, d$ and each $t \geqslant 0$, let $W_j(t)$ denote the workload for server j at time t, which is the sum of the impending service times for all customers who are queued at station j at time t, plus the remaining service time for any customer who may be in service there at time t. Set $W(t)$ to be the corresponding d-dimensional vector. In a typical setup for the Brownian approximation of the workload process, one considers a sequence of queueing networks, indexed by n. The basic network topology remains fixed across the entire sequence, however, with the arrival and service rates, the corresponding distributions, and the buffer sizes varying over n. As $n \to \infty$, we assume that the traffic intensity ρ_j^n at each station converges to one, and the buffer size at each station is in the order of $1/(1 - \rho_j^n)$ (the so-called heavy traffic condition). In this setting, one expects the scaled workload process $W^n(n\cdot)/\sqrt{n}$ converges weakly to a limit which is a Brownian model.

The focus of this paper is on the formulation of the Brownian model. We do so by showing a pseudo-heavy-traffic limit theorem, which states that if the scaled workload process $W^n(n\cdot)/\sqrt{n}$ converges weakly to a continuous process, then the limit is a reflecting Brownian motion (RBM) in

a hypercube of a d-dimensional Euclidean space. The reflection matrix for the limit is shown to be completely-$\mathscr{S}$, which is the sufficient and necessary condition for the existence and uniqueness of a semimartingale reflecting Brownian motion (SRBM) in the hypercube [7]. A *state space collapse* property for the limiting processes is observed. It has the potential to be used as an important ingredient in justifying the weak convergence.

We now introduce some notation and convention to be used in this paper. The set of nonnegative integers is denoted by Z_+, and the d-dimensional nonnegative lattice is denoted by Z_+^d. We use $\mathbb{R}^d$ to denote the d-dimensional Euclidean space. Let $R_+ = [0, \infty)$. Unless stated otherwise, all vectors are envisioned as column vectors. The prime symbol on a vector or a matrix denotes transpose. For $a = (a_1, \ldots, a_d)' \in \mathbb{R}^d$, $|a| = \max_{i=1}^d |a_i|$. For a $n \times d$ matrix A, $\|A\| = \max_{i=1}^n \sum_{j=1}^d |A_{ij}|$. For a vector $a \in \mathbb{R}^d$, we use $\mathrm{diag}(a)$ to denote the $d \times d$ diagonal matrix whose diagonal entries are given by the components of a. Vector inequalities are interpreted componentwise.

For $d \geqslant 1$, the d-dimensional path space $D([0, \infty), \mathbb{R}^d)$ is the set of functions $x : [0, \infty) \to \mathbb{R}^d$ that are right continuous on $[0, \infty)$ and have finite left limits on $[0, \infty)$. For a path $x \in D([0, \infty), \mathbb{R}^d)$, we sometimes use $x(\cdot)$ to denote the path. For a vector $a \in \mathbb{R}^d$ and a path $x \in D([0, \infty), \mathbb{R}^d)$, $x(a\cdot)$ is the path with $x(at) = (x_1(a_1 t), \ldots, x_d(a_d t))'$. More generally, for a $h \in D([0, \infty), \mathbb{R}^d)$, $x(h(\cdot))$ is the path with $x(h(t)) = (x_1(h_1(t)), \ldots, x_d(h_d(t)))'$. A path $x \in D([0, \infty), \mathbb{R}^d)$ is nondecreasing if each component is. We use $x(s-)$ to denote the left limit as $s > 0$. The path space is endowed with the Skorohod J_1-topology (see, e.g., [9]). For a sequence $\{X_n\}$ and X of $D([0, \infty), \mathbb{R}^d)$-valued stochastic processes, we write $X_n \Rightarrow X$ if X_n converges to X in distribution. For any $x \in D([0, \infty), \mathbb{R}^d)$, the uniform norm of x on the interval $[s, t]$ is defined by $\|x\|_{[s,t]} = \sup_{s \leqslant u \leqslant t} \|x(u)\|$ with $\|x\|_{[0,t]}$ abbreviated to $\|x\|_t$. A sequence $\{x_n\}$ of functions in $D([0, \infty), \mathbb{R}^d)$ is said to converge uniformly on compact set (u.o.c.) to $x \in D([0, \infty), \mathbb{R}^d)$ if $\|x_n - x\|_t \to 0$ for each $t \geqslant 0$. This is denoted by $x_n \to x$, u.o.c. as $n \to \infty$.

2. The capacitated multiclass queueing network model

The queueing network under consideration consists of d single server stations indexed by $i \in \mathscr{I} = \{1, \ldots, d\}$. Each station i has a *finite* storage waiting room with size b_i including the one possibly in service. The service discipline is assumed to be first-in-first-out (FIFO) and work-conserving, namely, a server at a station will always be busy as long as there are customers available. There are K classes of customers denoted by $\mathscr{K} = \{1, 2, \ldots, K\}$ in the network. Each class $k \in \mathscr{K}$ is served at a specific station while each station may serve more than one class of customers. The many-to-one mapping from customer classes to stations is described by a $d \times K$ *constituency matrix* C where for $i \in \mathscr{I}$, $k \in \mathscr{K}$,

$$C_{ik} = \begin{cases} 1 & \text{if class } k \text{ is served at station } i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

For $i \in \mathscr{I}$, let $\mathscr{C}(i)$ denote the constituency of server i, i.e., $\mathscr{C}(i) = \{k \in \mathscr{K} : C_{ik} = 1\}$, and for $k \in \mathscr{K}$, let $h(k)$ denote the station at which class k is served, i.e., $h(k)$ is the unique $i \in \mathscr{I}$ such that $C_{ik} = 1$.

The primitive data associated with the queueing network are: a K-dimensional *external arrival* process $E(t) = \{E_k(t), k \in \mathscr{K}\}$, $t \geqslant 0$, a K-dimensional *service* process $S(t) = \{S_k(t), k \in \mathscr{K}\}$ $t \geqslant 0$, and K K-dimensional *routing* sequences $\phi^k = \{\phi^k(n), n \geqslant 0\}$ for $k \in \mathscr{K}$.

The quantity $E_k(t)$ is the number of arrivals to class k from outside the network that have occurred by time t. We may have $E_k(\cdot) \equiv 0$ for some k. Let $\mathscr{A} = \{k \in \mathscr{K}: E_k(\cdot) \not\equiv 0\}$, the set of classes that have some external arrivals. For $k \in \mathscr{K}$, $E_k(t)$ is supposed to be defined from a sequence of independent random variables $\{\xi_k(j), \ j \geqslant 1\}$, where $\xi_k(j)$ denotes the time between the $(j-1)$th and the jth external arrival of a class k customer. We assume that $\{\xi_k(j), \ j \geqslant 1\}$ is a sequence of strictly positive i.i.d. random variables, each with mean $0 < E\xi_k(j) = 1/\lambda_k < \infty$, variance $\sigma_{a,k}^2$ and squared coefficient of variation (SCV) $c_{a,k}^2 = \lambda_k^2 \sigma_{a,k}^2$. We interpret λ_k as the long run average external arrival rate for class k. For convenience, we define $\lambda_k = 0$ for $k \notin \mathscr{A}$. For $k \in \mathscr{K}$, letting $U_k(0) = 0$ and

$$U_k(n) = \sum_{j=1}^{n} \xi_k(j) \quad \text{for } n \geqslant 1; \ E_k(t) = \operatorname{Sup}\{n \geqslant 0 \colon U_k(n) \leqslant t\} \quad \text{for all } t \geqslant 0. \tag{2.2}$$

The quantity $S_k(t)$ for $k \in \mathscr{K}$ indicates the number of service completions for class k customers after server $h(k)$ works on class k by time t. It is assumed that $S_k(t)$ is defined from a sequence of independent random variables $\{\eta_k(j), \ j \geqslant 1\}$, where for $\eta_k(j)$ denotes the amount of service time required by the jth customer of class k. It is supposed that $\{\eta_k(j) \ j \geqslant 1\}$ is a sequence of strictly positive i.i.d. random variables, each with mean $0 < m_k = E\eta_k(j) = 1/\mu_k < \infty$, variance $\sigma_{s,k}^2$ and SCV $c_{s,k}^2 = \mu_k^2 \sigma_{s,k}^2$. The μ_k is interpreted as the long run average service rate for class k customers. We define the *cumulative service time* process $V(n) = \{V_k(n_k), \ k \in \mathscr{K}\}$, $n = (n_1, \ldots, n_K)$ with $n_k = 1, 2, \ldots$, as follows:

$$V_k(n_k) = \sum_{j=1}^{n_k} \eta_k(j) \quad \text{for } n_k \geqslant 1; \ S_k(t) = \operatorname{Sup}\{n \geqslant 0; V_k(n_k) \leqslant t\} \quad \text{for all } t \geqslant 0. \tag{2.3}$$

The quantity $\phi^k(n) = e^l$ for $k, l \in \mathscr{K}$ represents that the nth customer of class k turns into a class l customer after service completion, and $\phi^k(n) = 0$ indicates that the nth customer of class k leaves the network after service completion. e^l is the lth unit vector in $\mathbb{R}^K$ with its lth component being one and other components being zeroes. It is supposed that the ϕ^k for $k = 1, 2, \ldots, K$ are mutually independent i.i.d. sequences, and they are also independent of the arrival process $E(t)$ and the service process $S(t)$. For each $k \in \mathscr{K}$, we define a K-dimensional *cumulative routing* process for class k by

$$\Phi^k(n) \equiv \phi^k(1) + \cdots + \phi^k(n). \tag{2.4}$$

The lth component $\Phi_{kl}(n)$ of $\Phi^k(n)$ is the cumulative number of customers to class l for the first n customers leaving class k. Let p_{kl} be the probability that $\phi^k(n) = e^l$ for $k, l \in \mathscr{K}$. We call the $K \times K$ matrix P with entries p_{kl} *routing matrix*. In this paper, we focus our discussion on *in-tree* network, that is, customers leaving station i all go next to station $\sigma(i) \in \mathscr{I}$ or leave the network. In the latter case, we let $\sigma(i) = 0$. Therefore the routing must be feedforward and the network must be open, that is, all customers eventually leave the network. Thus the spectral radius of P is strictly less than one. Hence

$$Q = (I - P')^{-1} = I + P + P^2 + \cdots \tag{2.5}$$

is well defined, where P' is the transpose of the matrix P.

An important feature in the network is that the sizes of buffers are *finite*. When the buffer at a downstream station $\sigma(i)$ is full, server i stops working although a customer may still occupy station i. One can envision that when the jth customer of class k enters service at station $h(k)$, a service time clock (stopwatch) is set to $\eta_k(j)$. The service is completed when the clock reading reaches zero. During the service period, the clock is turned off or on depending on whether the server is blocked or not. The blocking mechanism also applies to arrivals. Upon the jth external arrival of class k customer to station $h(k)$, an arrival clock at station $h(k)$ is set to $\xi_k(j+1)$. When the clock reading reaches zero, the $(j+1)$th customer of class k arrives at station $h(k)$. During this interarrival period, the arrival clock is turned off or on depending on whether buffer $h(k)$ is full or not. The blocking mechanism represents one way of modeling arrival processes. In many manufacturing applications, external arrivals can be controlled.

The performance measures of primary interest related to our queueing network are a K-dimensional *queue length* process $Q(t) = \{Q_k(t), \ k \in \mathcal{K}\}$ for $t \geq 0$ and a d-dimensional *workload* process $W(t) = \{W_i(t), \ i \in \mathcal{I}\}$ for $t \geq 0$, where $Q_k(t)$ denotes the number of class k customers occupying station $h(k)$ at time t and $W_i(t)$ represents the amount of work (measured in units of remaining service time) embodied in those customers who are at station i at time t. Let $Z_i(t)$ be the total number of customers occupying station i at time t for $i \in \mathcal{I}$, which is the summation of the numbers of all class customers in station i, that is,

$$Z_i(t) = \sum_{k \in C(i)} Q_k(t). \tag{2.6}$$

We call the d-dimensional process $Z(t) = \{Z_i(t), \ i \in \mathcal{I}\}$ for $t \geq 0$ the *total queueing length* process and assume the initial queue length $Q(0) = 0$ for convenience. Let $Y_i(t)$ be the amount of time that server i has been idle while server i is not blocked in time interval $[0, t]$, and let $Y_{i+d}(t)$ be the amount of time that buffer i has been full in time interval $[0, t]$, namely,

$$Y_i(t) = \int_0^t 1_{\{Z_i(s)=0, Z_{\sigma(i)}(s) < b_{\sigma(i)}\}} \, \mathrm{d}s, \qquad Y_{i+d}(t) = \int_0^t 1_{\{Z_i(s)=b_i\}} \, \mathrm{d}s. \tag{2.7}$$

In the sequel, whenever $\sigma(i) = 0$, condition $\{a_{\sigma(i)} < b_{\sigma(i)}\}$ always holds for any $a, b \in \mathbb{R}$. The $2d$-dimensional process $Y(t) = (Y_1(t), \ldots, Y_{2d}(t))'$ is called the *allocation* process. For $i \in \mathcal{I}$ and $t \geq 0$, let

$$F_i(t) = t - Y_{i+d}(t), \quad B_i(t) = t - Y_i(t) - Y_{\sigma(i)+d}(t). \tag{2.8}$$

Hereafter, whenever $\sigma(i)=0$, $Y_{\sigma(i)+d}(t)$ is understood to be zero. It is clear that $F_i(t)$ is the cumulative amount of time that buffer i has not been full in time interval $[0, t]$ and $B_i(t)$ is the cumulative amount of time that server i has been busy in time interval $[0, t]$. That is,

$$F_i(t) = \int_0^t 1_{\{Z_i(s) < b_i\}} \, \mathrm{d}s, \qquad B_i(t) = \int_0^t 1_{\{Z_i(s)>0, Z_{\sigma(i)}(s) < b_{\sigma(i)}\}} \, \mathrm{d}s.$$

We further define the K-dimensional processes $A(t) = \{A_k(t), \ k \in \mathcal{K}\}$ for $t \geq 0$ and $D(t) = \{D_k(t), \ k \in \mathcal{K}\}$ for $t \geq 0$ where $A_k(t)$ denotes the number of arrivals to class k that have occurred in the time interval $[0, t]$ and $D_k(t)$ denotes the number of departures from class k that have

occurred in the time interval $[0, t]$. We assume $A(0) = 0$ and $D(0) = 0$. Then, for $i \in \mathscr{I}$, $k \in C(h(k))$, we have,

$$A_k(t) = E_k(F_{h(k)}(t)) + \sum_{l \in \bigcup_i C(i), \sigma(i) = h(k)} \Phi_{lk}(D_l(F_{h(k)}(t))). \tag{2.9}$$

Where $E_k(F_{h(k)}(t))$ is the external arrivals to class k in $[0, t]$ when station $h(k)$ is turned on. The second term in the right hand side of (2.9) is the internal arrivals due to the fraction of the $D_l(t)$ departures from class l that is routed next to class k, summed over all classes $l \in \mathscr{K}$ when the station $h(k)$ is turned on. Then, the d-dimensional workload process $W(t)$ can be formulated as follows

$$W(t) = U(t) - te + R_1 Y(t), \tag{2.10}$$

where $U(t) = CV(A(t))$, e is the d-dimensional vector of ones and the matrix R_1 is given by

$$R_1 = (I, \Pi) \tag{2.11}$$

where I is a $d \times d$ unit matrix and Π is a $d \times d$ matrix with entries $\delta_{ij} = 1$ if $\sigma(i) = j$ and otherwise zero. From (2.10), we can write the workload process $W(t)$ in componentwise form, namely, for each $i \in \mathscr{I}$,

$$W_i(t) = U_i(t) - t + Y_i(t) + Y_{\sigma(i)+d}(t), \tag{2.12}$$

Eq. (2.12) can be interpreted as follows. The amount of work $W_i(t)$ remaining at station i at time t is the total amount of work $U_i(t)$ that has arrived at station i minus the amount of time $B_i(t) = t - Y_i(t) - Y_{\sigma(i)+d}(t)$ that the server has been busy up to time t.

3. The main theorem and the Brownian model

To state the main theorem, we consider a sequence of capacitated multiclass queueing networks indexed by $n \geqslant 1$. The network depends on index n through the external arrival rates λ^n, mean service times m^n and buffer sizes b^n, where

$$\lambda^n = (\lambda_1^n, \ldots, \lambda_K^n)', \quad m^n = (m_1^n, \ldots, m_K^n)', \quad b^n = (b_1^n, \ldots, b_d^n)'.$$

We let $\mu_k^n = 1/m_k^n$ be the mean service rate for class k customer. To be simple, we assume that the routing process does not vary with n. Define $\alpha^n = (\alpha_1^n, \ldots, \alpha_K^n)'$ via

$$\alpha^n = Q\lambda^n, \tag{3.1}$$

where Q is given by (2.5). We also define

$$\rho_i^n = \sum_{k \in C(i)} \alpha_k^n m_k^n \quad \text{for all } i \in \mathscr{I},$$

or equivalently $\rho^n = CM^n \alpha^n$, where M^n is the $K \times K$ diagonal matrix $\text{diag}(m^n)$ with diagonal elements $m_1^n, \ldots, m_K^n$. For the nth network, α_k^n is interpreted as the *effective arrival rate* to class k due to external and internal arrivals, and ρ_i^n is the *traffic intensity* for station i. The traffic equation (3.1) implicitly

assumes that for each class k there is a long run average rate α_k of flow into and out of that class and that does not exceed the mean service rate μ_k^n for class k.

To obtain our main theorem, we need to impose some additional conditions on the behavior of the primitive processes as $n \to \infty$. That is,

$$\lambda^n \to \lambda \geqslant 0, \quad m^n \to m > 0 \tag{3.2}$$

$$\frac{1}{\sqrt{n}} b^n \to b > 0, \tag{3.3}$$

$$\sqrt{n}(\rho^n - e) \to \theta. \tag{3.4}$$

Conditions (3.2)–(3.4) are called *heavy traffic condition*. Where $\lambda = (\lambda_1, \ldots, \lambda_K)'$ and $m = (m_1, \ldots, m_K)'$. Therefore we see that $\rho^n \to \rho = CM\alpha$ as $n \to \infty$ with $M = \text{diag}(m)$ and

$$\alpha = Q\lambda. \tag{3.5}$$

In addition to the heavy traffic condition, we make the following assumptions which imply a functional central limit theorem. The variances of the external interarrival times and service times satisfy

$$(\sigma_{a,k}^n)^2 \to \sigma_{a,k}^2 \quad \text{for } k \in \mathscr{A}, \qquad (\sigma_{s,k}^n)^2 \to \sigma_{s,k}^2 \quad \text{for } k \in \mathscr{K}, \tag{3.6}$$

and a Lindeberg-type condition on $\xi_k^n(j)$ and $\eta_k^n(j)$ holds as explained in [12]. It then follows by functional central limit theorems as in [12], we have

$$\tilde{E}^n(t) \equiv \frac{1}{\sqrt{n}} \hat{E}^n(nt) = \sqrt{n} \left(\frac{E^n(nt)}{n} - \lambda^n t \right) \Rightarrow \tilde{E}(t), \tag{3.7}$$

$$\tilde{V}^n(t) = \frac{1}{\sqrt{n}} (V^n([nt]) - m^n nt) \Rightarrow \tilde{V}(t), \tag{3.8}$$

$$\tilde{\Phi}^{j,n}(t) \equiv \frac{1}{\sqrt{n}} \hat{\Phi}^{j,n}([nt]) = \sqrt{n} \left(\frac{\Phi^j([nt])}{n} - P_j' t \right) \Rightarrow \tilde{\Phi}^j(t) \quad \text{for } j = 1, \ldots, K. \tag{3.9}$$

where "$\Rightarrow$" denotes convergence in distribution, $[x]$ is the integer part of x, P_j denotes the jth row of the matrix P, and $\tilde{E}(t)$, $\tilde{V}(t)$, $\tilde{\Phi}^j(t)$ ($j = 1, \ldots, K$) are independent d-dimensional driftless Brownian motions with covariance matrices Γ^a, Γ^v and Γ^j as follows:

$$\Gamma^a = \text{diag}(\lambda_1 c_{a,1}^2, \ldots, \lambda_K c_{a,K}^2) = \text{diag}(\lambda)\text{diag}(c_a^2),$$

$$\Gamma^v = \text{diag}(m_1^2 c_{s,1}^2, \ldots, m_K^2 c_{s,K}^2) = \text{diag}(m^2)\text{diag}(c_s^2),$$

$$\Gamma_{lk}^j = \begin{cases} P_{jl}(1 - P_{jl}) & \text{if } l = k, \\ -P_{jl}P_{jk} & \text{if } l \neq k. \end{cases}$$

We are interested in the limit behaviors following from the convergence assumption:

$$(\tilde{W}^n(\cdot), \tilde{Y}^n(\cdot)) = \left(\frac{1}{\sqrt{n}} W^n(nt), \frac{1}{\sqrt{n}} Y^n(nt) \right) \Rightarrow (\tilde{W}(\cdot), \tilde{Y}(\cdot)), \tag{3.10}$$

where the limits $\tilde{W}(\cdot)$ and $\tilde{Y}(\cdot)$ are supposed to be continuous processes. Because Brownian motions are continuous and $\tilde{E}^n$, $\tilde{V}^n$, $\tilde{\Phi}^{j,n}$ ($j = 1,\ldots,K$) are independent, we assume by Skorohod representation theorem that the convergence in (3.7)–(3.10) holds u.o.c. Let

$$G = CMQP'\Delta C', \quad R_2 = [0, \Lambda C'], \tag{3.11}$$

where $M = \mathrm{diag}(m_1,\ldots,m_K)$, $\Delta = \mathrm{diag}(\alpha_1,\ldots,\alpha_K)$, $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_K)$ and R_2 is a $K \times 2d$ matrix. Let the d-dimensional hypercube state space defined by

$$S = \left\{ x \in \mathbb{R}^d : 0 \leqslant x_i \leqslant \frac{b_i}{\sum_{k \in C(i)} \alpha_k} \right\} \tag{3.12}$$

with the $2d$ boundary faces as follows:

$$F_i = \{x \in S : x_i = 0\}, \; F_{i+d} = \left\{ x \in S : x_i = \frac{b_i}{\sum_{k \in C(i)} \alpha_k} \right\} \quad \text{for } i = 1, 2, \ldots, d. \tag{3.13}$$

Theorem 3.1. *Suppose* (3.10) *is true, then the limiting process* $(\tilde{W}, \tilde{Y})$ *must satisfy*

$$(I + G)\tilde{W}(t) = C\tilde{V}(\alpha t) + CMQ\left(\tilde{E}(t) + \sum_{l=1}^{K} \tilde{\Phi}^l(\alpha_l t)\right) + \theta t + (R_1 - CMQR_2)\tilde{Y}(t), \tag{3.14}$$

$$\tilde{W}(t) \in S, \tag{3.15}$$

$$\tilde{Y}(0) = 0, \; \tilde{Y} \text{ is continuous and nondecreasing}, \tag{3.16}$$

$$\tilde{Y}_i(t) \text{ increases only at times } t \text{ such that } \tilde{W}(t) \text{ reaches the boundary } F_i \text{ of } S$$
$$\text{for } i = 1,\ldots,2d. \tag{3.17}$$

Here we first give some discussions about the solution of the system (3.14)–(3.17).

Lemma 3.1. *The matrix* $I + G$ *is invertible.*

Multiplying both sides of (3.14) by $R_3 = (I + G)^{-1}$, let $R = R_3(R_1 - CMQR_2)$ and

$$\tilde{X}(t) = R_3 C\tilde{V}(\alpha t) + R_3 CMQ\left(\tilde{E}(t) + R_3 \sum_{l=1}^{K} \tilde{\Phi}_l(\alpha_l t)\right) + R_3 \theta t.$$

Then $\tilde{W}(t) = \tilde{X}(t) + R\tilde{Y}(t)$ and $\tilde{X}$ is a Brownian motion with drift vector $R_3\theta$ and covariance matrix

$$\Gamma = R_3 C\left\{\Gamma^v \Delta + MQ\left(\Gamma^a + \sum_{l=1}^{K} \alpha_l \Gamma^l\right) Q'M'\right\} C'R_3'.$$

From Theorem 3.1, we see that the process $\tilde{W}(t)$ behaves like a Brownian motion with drift vector $R_3\theta$ and covariance matrix Γ in the interior of the state space S, with the process being confined to

the state space by instantaneous reflection at the boundary, where the direction of reflection on the ith boundary face F_i is given by the ith column of R. So R is called reflection matrix. The process $\tilde{W}(t)$ is recognized as a reflecting Brownian motion (RBM). If $\{\tilde{X}(t) - R_3\theta t,\ t \geqslant 0\}$ can be showed a martingale in terms of the filtration generated by $(\tilde{W}, \tilde{Y})$, the process $\tilde{W}$ is called a semimartingale reflecting Brownian motion (SRBM). The existence of SRBM depends on the properties of the reflection matrix R. Dai and Williams [7] proved that a so-called completely-$\mathscr{S}$ condition of matrix R is the sufficient and necessary condition for the existence and uniqueness of a SRBM in the state space S. In Lemma 3.2, it is proved that the reflection matrix R indeed satisfies the completely-$\mathscr{S}$ condition.

Definition 3.1. A square matrix A is said to be a-$\mathscr{S}$-matrix if there is a vector $x \geqslant 0$ such that $Ax > 0$. The matrix A is said to be completely-$\mathscr{S}$ if each principal submatrix of A is a $\mathscr{S}$-matrix.

The reflection matrix R is a $d \times 2d$ matrix. We partition R as $R = (R^{\mathrm{I}}, R^{\mathrm{II}})$, where R^{I} and R^{II} are $d \times d$ matrices, formed by the first and the last d columns of R, respectively. To set up a connection with the definition of completely-$\mathscr{S}$ for square matrix. We introduce the notion of reflection matrix associated with a vertex. There are 2^d vertexes for the state space S, and each vertex is given by $\bigcap_{i\in\alpha} F_i \bigcap_{i\in\beta} F_{d+i}$ for a (unique) index set $\alpha \subset \{1,\dots,d\}$ with $\beta = \{1,\dots,d\}\backslash\alpha$. For each vertex α, the reflection matrix R^α associated with the vertex is the following $d \times d$ matrix

$$R^\alpha = (I_\alpha - I_\beta)(R^{\mathrm{I}}I_\alpha + R^{\mathrm{II}}I_\beta). \tag{3.18}$$

where I_α is a $d \times d$ diagonal matrix whose ith component equals one if $i \in \alpha$ and equals zero otherwise, and I_β is similarly defined.

Definition 3.2. The $d \times 2d$ reflection matrix R is said to satisfy the completely-$\mathscr{S}$ condition if for each vertex α, R^α is a completely-$\mathscr{S}$ matrix.

Lemma 3.2. *The $d \times 2d$ reflection matrix $R = R_3(R_1 - CMQR_2)$ satisfies the completely-$\mathscr{S}$ condition.*

Lemmas 3.1 and 3.2 will be proved in Section 5.

4. Proof of Theorem 3.1

We begin the proof procedure by introducing some important notation. For $i \in \mathscr{I}$, define $\tau_i^n(t)$ to be the arrival time to station i of the customer currently being served if $W_i^n(t) > 0$, and to be t if $W_i^n(t) = 0$. Let $\tau^n(t)$ be the d-dimensional vector defined in the obvious manner. From the definition of $\tau_i^n(t)$, we have that

$$t = \tau_i^n(t) + W_i^n(\tau_i^n(t)) + \sum_{j>i}\delta_{ij}(Y_{j+d}^n(t) - Y_{j+d}^n(\tau_i^n(t))) - \varepsilon_i^n(t), \tag{4.1}$$

where δ_{ij} is defined in (2.11) and $\varepsilon_i^n(t) = 0$ if $W_i^n(t) = 0$, otherwise it equals to the remaining service time of the customer currently occupying server i. Then we have the following fact on $\varepsilon_i^n(t)$.

Lemma 4.1. *For each $i \in \mathcal{I}$, we have*

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} \varepsilon_i^n(nt) \to 0 \quad \text{as } n \to \infty. \tag{4.2}$$

Proof. The proof of the lemma is a direct application of Lemma 3.3 from [10] or see Lemma 4.2 in [6]. □

For each $t \geq 0$ and $n \geq 1$, define

$$\hat{\tau}^n(t) = et - \tau^n(t), \quad \bar{\tau}^n(t) = \frac{1}{n}\tau^n(nt), \quad \tilde{\tau}^n(t) = \frac{1}{\sqrt{n}}(ent - \tau^n(nt))$$

where e is the d-dimensional vector of ones. Then we have the following lemma.

Lemma 4.2. *Under the convergence assumption in (3.10), we have*

$$\bar{\tau}^n(t) \to et \text{ u.o.c.,} \quad \text{as } n \to \infty, \tag{4.3}$$

$$\tilde{\tau}^n(t) \to \tilde{W}(t) \text{ u.o.c.,} \quad \text{as } n \to \infty. \tag{4.4}$$

Proof. From the definition of $\tau_i^n(t)$, we have

$$\hat{\tau}_i^n(t) = W_i^n(\tau_i^n(t)) + \sum_{j>i} \delta_{ij}(Y_{j+d}^n(t) - Y_{j+d}^n(\tau_i^n(t))) - \varepsilon_i^n(t)$$

$$\leq W_i^n(\tau_i^n(t)) + \sum_{j>i} \delta_{ij}(Y_{j+d}^n(t) - Y_{j+d}^n(\tau_i^n(t))). \tag{4.5}$$

Thus, we have

$$\left\| \frac{1}{n}\hat{\tau}_i^n(n\cdot) \right\|_t \leq \frac{1}{\sqrt{n}} \left\| \tilde{W}_i^n(\bar{\tau}_i^n(\cdot)) \right\|_t + \frac{1}{\sqrt{n}} \left\| \frac{1}{\sqrt{n}} \varepsilon_i^n(n\cdot) \right\|_t$$

$$+ \frac{1}{\sqrt{n}} \left\| \sum_{j>i} \delta_{ij}(\tilde{Y}_{j+d}^n(\cdot) - \tilde{Y}_{j+d}^n(\bar{\tau}_i^n(\cdot))) \right\|_t.$$

Notice that $\bar{\tau}_i^n(t) \leq t$ for any $t \geq 0$, and then by (3.10) and Lemma 4.1, the convergence in (4.3) is proved. Next, we prove the convergence in (4.4). From (4.5), we have

$$\tilde{\tau}_i^n(t) = \tilde{W}_i^n(\bar{\tau}_i^n(t)) + \sum_{j>i} \delta_{ij}(\tilde{Y}_{j+d}^n(t) - \tilde{Y}_{j+d}^n(\bar{\tau}_i^n(t))) - \frac{1}{\sqrt{n}} \varepsilon_i^n(nt)$$

$$\leq \tilde{W}_i^n(\bar{\tau}_i^n(t)) + \sum_{j>i} \delta_{ij}(\tilde{Y}_{j+d}^n(t) - \tilde{Y}_{j+d}^n(\bar{\tau}_i^n(t))). \tag{4.6}$$

Notice (4.3) and (3.10), we have

$$\sum_{j>i} \delta_{ij}(\tilde{Y}_{j+d}^n(t) - \tilde{Y}_{j+d}^n(\bar{\tau}_i^n(t))) \to 0, \quad \text{u.o.c. as } n \to \infty. \tag{4.7}$$

Then by (4.6), (3.10), (4.3), (4.7) and Lemma 4.1, the convergence in (4.4) is proved. □

From the definition of $\tau_i^n(t)$, one can see that the number of class l customers to have departed from station $i = h(l)$ to station $\sigma(i) \in \mathscr{I}$ by time t is given by

$$D_l^n(F_{\sigma(i)}^n(t)) = \begin{cases} A_l^n(\tau_i^n(t)) - 1 & \text{if server } i \text{ is serving a class } l \text{ customer}, \\ A_l^n(\tau_i^n(t)) & \text{otherwise}, \end{cases} \tag{4.8}$$

where $F_{\sigma(i)}^n(t) = t$ if $\sigma(i) = 0$. We will use $A^n(\tau^n(t))$ to denote the K-dimensional process with the lth component given by $A_l^n(\tau_{h(l)}^n(t))$, for $l \in \mathscr{K}$. For each $t \geq 0$ and $n \geq 1$, define

$$\bar{A}^n(t) = \frac{1}{n} A^n(nt), \quad \tilde{A}^n(t) = \frac{1}{\sqrt{n}}(A^n(nt) - \alpha^n nt). \tag{4.9}$$

Lemma 4.3. *There exists $\kappa = \kappa(t)$ independent of n such that*

$$\|\bar{A}_l(\cdot)\|_t \leq \kappa \quad \text{for } l = 1, \ldots, K, \ n \geq 1.$$

Proof. Let $S_l(t)$ for $t \geq 0$ be the renewal process associated with class l service times, which is defined in (2.3). Let $T_l^n(t)$ be the cumulative time that server $h(l)$ has devoted to class l customers in the time interval $[0, t]$. Then the number of class l customers who have departed from station $h(l)$ by time t is

$$D_l^n(F_{\sigma(h(l))}^n(t)) = S_l^n(T_l^n(t)) \leq S_l^n(t). \tag{4.10}$$

Notice that the cumulative number of customers to class k for the first m customers leaving class l, namely, $\Phi_{lk}(m)$, will be zero if the transition probability $p_{lk} = 0$ for $l, k \in \mathscr{K}$. Then, from (2.9),

$$A^n(t) = E^n(F^n(t)) + \sum_{l=1}^{K} \Phi^{l,n}(D_l^n(F_{\sigma(h(l))}^n(t))) \leq E^n(t) + \sum_{l=1}^{K} \Phi^{l,n}(S_l^n(t)),$$

where $E^n(F^n(t))$ is the K-dimensional external arrival process with the kth component given by $E_k^n(F_{h(k)}^n(t))$ for $k \in \mathscr{K}$. Then the lemma follows from the functional strong law of large numbers for random walks and renewal processes. $\square$

Lemma 4.4. *Under the convergence assumption in* (3.10), *we have*

$$\bar{A}^n(t) \to \alpha t \ \text{u.o.c.} \quad \text{and} \quad \bar{D}^n(\bar{F}_\sigma^n(t)) \to \alpha t \quad \text{as } n \to \infty,$$

where $\bar{D}^n(\bar{F}_\sigma^n(t))$ is a K-dimensional process with the lth component given by $\bar{D}_l^n(\bar{F}_{\sigma(h(l))}^n(t))$, $\bar{D}^n(t) = 1/n D^n(nt)$, $\bar{F}_\sigma^n(t) = 1/n F_\sigma^n(nt)$ and $F_\sigma^n(t)$ is a K-dimensional process with component $F_{\sigma(h(l))}^n(t)$.

Proof. From (2.9) and using the same explanation as before, we have

$$\bar{A}^n(t) = \bar{E}^n(\bar{F}^n(t)) + \sum_{l=1}^{K} \bar{\Phi}^{l,n}(\bar{D}_l^n(\bar{F}_{\sigma(h(l))}^n(t))), \tag{4.11}$$

where $\bar{E}^n(\bar{F}^n(t))$ is a K-dimensional process with component $\bar{E}^n_k(\bar{F}^n_{h(k)}(t))$ for $k \in \mathcal{K}$, $\bar{E}^n(t) = E^n(nt)/n$, $\bar{F}^n(t) = F^n(nt)/n$ and $\bar{\Phi}^{l,n}(t) = \Phi^{l,n}([nt])/n$. By (3.5), we know $\alpha = \lambda + P'\alpha$. Thus

$$\bar{A}^n(t) - \alpha t = \bar{E}^n(\bar{F}^n(t)) - \lambda t + \sum_{l=1}^{K} (\bar{\Phi}^{l,n}(\bar{D}^n_l(\bar{F}^n_{\sigma(h(l))}(t))) - P'_l \bar{D}^n_l(\bar{F}^n_{\sigma(h(l))}(t)))$$

$$+ P'(\bar{D}^n(\bar{F}^n_\sigma(t)) - \Delta C' \bar{\tau}^n(t)) - P' \Delta C'(te - \bar{\tau}^n(t)), \tag{4.12}$$

where $\Delta = \mathrm{diag}(\alpha_i)$ and P_l denotes the lth row of P. By (3.10), we know, for each $i \in \mathcal{I}$,

$$\bar{F}^n_i(t) \to t \quad u.o.c. \text{ as } n \to \infty. \tag{4.13}$$

Using Eq. (4.8), we have

$$|\bar{A}^n(\bar{\tau}^n(t)) - \bar{D}^n(\bar{F}^n_\sigma(t))| \leqslant \frac{1}{n}. \tag{4.14}$$

Thus we can replace $\bar{D}^n(\bar{F}^n_\sigma(t))$ in the third term on the right hand side of (4.13) by $\bar{A}^n(\bar{\tau}^n(t))$ when n is large. Then from (3.7)–(3.9), (4.10), (4.13), Lemmas 4.2 and 4.3, we have almost surely

$$\limsup_{n\to\infty} \|\bar{A}^n(\cdot) - \alpha \cdot \|_t \leqslant \limsup_{n\to\infty} P' \|\bar{A}^n(\bar{\tau}^n(\cdot)) - \alpha \bar{\tau}^n(\cdot)\|_t \leqslant P' \limsup_{n\to\infty} \|\bar{A}^n(\cdot) - \alpha \cdot \|_t.$$

Then by the fact $(I - P')^{-1} > 0$, we have

$$\limsup_{n\to\infty} \|\bar{A}^n(\cdot) - \alpha \cdot \|_t \leqslant 0, \tag{4.15}$$

and therefore,

$$\lim_{n\to\infty} \|\bar{A}^n(\cdot) - \alpha \cdot \|_t = 0. \tag{4.16}$$

The convergence of $\bar{D}^n(\bar{F}^n_\sigma(t))$ follows from (4.8), Lemma 4.2 and convergence of $\bar{A}^n(t)$. $\quad\square$

Lemma 4.5. *Suppose the convergence in* (3.10) *holds. Define for each* $t \geqslant 0$,

$$\tilde{A}(t) = Q\left(\tilde{E}(t) + \sum_{l=1}^{K} \tilde{\Phi}^l(\alpha_l t) - P'\Delta C' \tilde{W}(t) - R_2 \tilde{Y}(t)\right).$$

Then we have $\tilde{A}^n(t) \to \tilde{A}(t)$ *u.o.c. as* $n \to \infty$.

Proof. From (2.9), we have

$$\tilde{A}^n(t) = \tilde{E}^n(\bar{F}^n(t)) + \sum_{l=1}^{K} \tilde{\Phi}^{l,n}(\bar{D}^n_l(\bar{F}^n_{\sigma(h(l))}(t))) + P'\tilde{A}^n(\bar{\tau}^n(t)) - P'\Delta^n C' \tilde{\tau}^n(t)$$

$$+ P'\sqrt{n}(\bar{D}^n(\bar{F}^n_\sigma(t)) - \bar{A}^n(\bar{\tau}^n(t))) - R_2^n \tilde{Y}^n(t),$$

where $R_2^n = [0, \Lambda^n C']$ is a $K \times 2d$ matrix and $\Lambda^n = \text{diag}(\lambda^n)$. From the definitions of $\tilde{A}(t)$ and the matrix Q, we know

$$\tilde{A}(t) = \tilde{E}(t) + \sum_{l=1}^{K} \tilde{\Phi}^l(\alpha_l t) + P'\tilde{A}(t) - P'\Delta C'\tilde{W}(t) - R_2\tilde{Y}(t).$$

Hence, we have

$$\tilde{A}^n(t) - \tilde{A}(t) = \tilde{E}^n(\bar{F}^n(t)) - \tilde{E}(t) + \sum_{l=1}^{K}(\tilde{\Phi}^{l,n}(\bar{D}_l^n(\bar{F}_{\sigma(h(l))}^n)(t)) - \tilde{\Phi}^l(\alpha_l t))$$

$$+ P'(\tilde{A}^n(\bar{\tau}^n(t)) - \tilde{A}(\bar{\tau}^n(t))) + P'(\tilde{A}(\bar{\tau}^n(t)) - \tilde{A}(t))$$

$$- P'(\Delta^n C'\tilde{\tau}^n(t) - \Delta C'\tilde{W}(t)) + P'\sqrt{n}(\bar{D}^n(\bar{F}_\sigma^n(t)) - \bar{A}^n(\bar{\tau}^n(t)))$$

$$- (R_2^n\tilde{Y}^n(t) - R_2\tilde{Y}(t)).$$

Therefore

$$\|\tilde{A}^n(\cdot) - \tilde{A}(\cdot)\|_t \leqslant \|\tilde{E}^n(\bar{F}^n(\cdot)) - \tilde{E}(\cdot)\|_t + \sum_{l=1}^{K}\|\tilde{\Phi}^{l,n}(\bar{D}_l^n(\bar{F}_{\sigma(h(l))}^n(\cdot)) + \tilde{\Phi}^l(\alpha_l\cdot)\|_t$$

$$+ P'\|\tilde{A}^n(\bar{\tau}^n(\cdot)) + \tilde{A}(\bar{\tau}^n(\cdot))\|_t + P'\|\tilde{A}(\bar{\tau}^n(\cdot)) - \tilde{A}(\cdot)\|_t$$

$$+ P'\|\Delta^n C'\tilde{\tau}^n(\cdot) - \Delta C'\tilde{W}(\cdot)\|_t + \frac{1}{\sqrt{n}}P'\tilde{e}$$

$$+ \|R_2^n\tilde{Y}^n(\cdot) - R_2\tilde{Y}(\cdot)\|_t,$$

where $\tilde{e}$ is the K-dimensional vector of ones. Since $\bar{\tau}_i^n(t) \leqslant t$ for all $t \geqslant 0$ and $i \in \mathscr{I}$, we have

$$\|\tilde{A}^n(\bar{\tau}^n(\cdot)) - \tilde{A}(\bar{\tau}^n(\cdot))\|_t \leqslant \|\tilde{A}^n(\cdot) - \tilde{A}(\cdot)\|_t.$$

Thus, we have

$$(I - P')\|\tilde{A}^n(\cdot) - \tilde{A}(\cdot)\|_t \leqslant \|\tilde{E}^n(\bar{F}^n(\cdot)) - \tilde{E}(\cdot)\|_t + \sum_{l=1}^{K}\|\tilde{\Phi}^{l,n}(\bar{D}_l^n(\bar{F}_{\sigma(h(l))}^n(\cdot))) - \tilde{\Phi}^l(\alpha_l\cdot)\|_t$$

$$+ P'\|\tilde{A}(\bar{\tau}^n(\cdot)) - \tilde{A}(\cdot)\|_t + P'\|\Delta^n C'\tilde{\tau}^n(\cdot) - \Delta C'\tilde{W}(\cdot)\|_t$$

$$+ \frac{1}{\sqrt{n}}P'\tilde{e} + \|R_2^n\tilde{Y}^n(\cdot) - R_2\tilde{Y}(\cdot)\|_t \equiv \Xi^n(t).$$

Note that $Q = (I - P')^{-1} > 0$. Multiplying Q on both sides of the above inequality, we have $\|\tilde{A}^n(\cdot) - \tilde{A}(\cdot)\|_t \leqslant Q\Xi^n(t)$. Again from (3.10) and the definition of $F_i(t)$ for $i \in \mathscr{I}$, we know that $\bar{F}_i^n(t) \to t$ u.o.c. as $n \to \infty$. Then by (3.2), (3.4), (3.10), Lemmas 4.2 and 4.4, the continuity of the $\tilde{\Phi}^l$ and $\tilde{A}$, we have that $\Xi^n(t) \to 0$ a.s. as $n \to \infty$. Hence we have $\tilde{A}^n(t) \to \tilde{A}(t)$ u.o.c. as $n \to \infty$. Therefore, we finish the proof of the lemma. $\square$

The next lemma shows that the limiting queue length process for class k customer is proportional to the total workload process at station $h(k)$. In queueing literature, such a property is called *state*

space collapse property. It is a sufficient condition for a multiclass queueing network to have a heavy traffic limit theorem.

Lemma 4.6. *Let $Q_k^n(t)$ and $\mathcal{W}_k^n(t)$ for $k \in \mathcal{K}$ be the queue length process and the workload process for class k customers in the nth network. Under condition* (3.10), *we have*

$$\tilde{Q}_k(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}} Q_k^n(nt) = \alpha_k \tilde{W}_{h(k)}(t) \ u.o.c., \tag{4.17}$$

$$\tilde{\mathcal{W}}_k(t) = \lim_{n \to \infty} \frac{1}{\sqrt{n}} \mathcal{W}_k^n(nt) = \alpha_k m_k \tilde{W}_{h(k)}(t) \ u.o.c., \tag{4.18}$$

where $\tilde{W}_{h(k)}(t)$ is the limiting total workload process for station $h(k)$ as stated before.

Proof. From the definition of $\tau_i^n(t)$, the number of class k customers in the station $i = h(k)$ at time t is given by

$$\frac{1}{\sqrt{n}} Q_k^n(nt) = \frac{1}{\sqrt{n}} (A_k^n(nt) - A_k^n(n\bar{\tau}_{h(k)}^n(t) + \tilde{\varepsilon}_k^n(nt))$$

$$= \tilde{A}_k^n(t) - \tilde{A}_k^n(\bar{\tau}_{h(k)}^n(t)) + \alpha_k^n \tilde{\tau}_{h(k)}^n(t)) + \frac{1}{\sqrt{n}} \tilde{\varepsilon}_k^n(nt),$$

where $\tilde{\varepsilon}_k^n(t) = 1$ if server $h(k)$ is serving a class k customer at time t and otherwise zero. Then (4.17) follows from Lemmas 4.2 and 4.5. Similarly, we have

$$\frac{1}{\sqrt{n}} \mathcal{W}_k^n(nt) = \frac{1}{\sqrt{n}} (V_k^n(A_k^n(nt)) - V_k^n(A_k^n(\tau_{h(k)}^n)(nt)) + \tilde{\tilde{\varepsilon}}_k^n(nt))$$

$$= \tilde{V}_k^n(\bar{A}_k^n(t)) - \tilde{V}_k^n(\bar{A}_k^n(\bar{\tau}_{h(k)}^n(t))) + m_k^n(\tilde{A}_k^n(t) - \tilde{A}_k^n(\bar{\tau}_{h(k)}^n(t)))$$

$$+ \alpha_k^n m_k^n \tilde{\tau}_{h(k)}^n(t)) + \frac{1}{\sqrt{n}} \tilde{\tilde{\varepsilon}}_k^n(nt),$$

where $\tilde{\tilde{\varepsilon}}_k^n(t)$ is the remaining service time if server $h(k)$ is serving a class k customer at time t and else zero. Then (4.18) follows from (3.8), Lemmas 4.1, 4.2, 4.4 and 4.5. □

Proof of Theorem 3.1. First, we show (3.14) in Theorem 3.1 is true. From (2.10), we have

$$\tilde{W}^n(t) = C\tilde{V}^n(\bar{A}^n(t)) + CM^n\tilde{A}^n(t) + \sqrt{n}(\rho^n - e)t + R_1\tilde{Y}^n(t). \tag{4.19}$$

Then (3.14) follows from (3.4), (3.8), (3.10), Lemmas 4.4 and 4.5 by taking limits on both sides of (4.19). Secondly, property (3.15) follows from Lemma 4.5 and (3.3), namely, for $i = h(k)$,

$$0 \leqslant \tilde{W}_i(t) = \frac{\lim_{n \to \infty} Z_i^n(nt)/\sqrt{n}}{\sum_{k \in C(i)} \alpha_k} \leqslant \frac{b_i}{\sum_{k \in C(i)} \alpha_k}. \tag{4.20}$$

The third property (3.16) in Theorem 3.1 follows from the convergence assumption (3.10) and the corresponding property of $\tilde{Y}^n(t)$. Finally, we prove the property (3.17). From (4.20), we know

that the equalities in both left and right hand sides of (4.20) hold only if $\tilde{W}_i(t) \in F_i$ or F_{i+d} for $i \in \{1,\ldots,d\}$. Furthermore notice that $\tilde{Y}_i(\cdot)$ for $i \in \{1,\ldots,2d\}$ is nondecreasing. It suffices to prove that for each $T > 0$ and $i = 1,\ldots,d$,

$$\int_0^T \tilde{W}_i(s) \wedge 1 \, d\tilde{Y}_i(s) = 0, \quad \int_0^T \left(\frac{b_i}{\sum_{k \in C(i)} \alpha_k} - \tilde{W}_i(s) \right) \wedge 1 \, d\tilde{Y}_{i+d}(s) = 0.$$

From Lemma 4.6, it is equivalent to show that for each $T > 0$ and $i = 1,\ldots,d$,

$$\int_0^T \tilde{Z}_i(s) \wedge 1 \, d\tilde{Y}_i(s) = 0, \quad \int_0^T (b_i - \tilde{Z}_i(s)) \wedge 1 \, d\tilde{Y}_{i+d}(s) = 0, \tag{4.21}$$

where $\tilde{Z}_i(\cdot)$ is the limiting queue length process at station i. Notice that for $i = 1,\ldots,d$, $\tilde{Y}_i^n(t)$ can increase only at times t such that $\tilde{Z}_i^n(t) = 0$ and for $i = d+1,\ldots,2d$, $\tilde{Y}_i^n(t)$ can increase only at times t such that $\tilde{Z}_i^n(t) = b_i^n/\sqrt{n}$. Then for each $T > 0$

$$\int_0^T \left(\frac{b_i^n}{\sqrt{n}} - \tilde{Z}_i^n(s) \right) \wedge 1 \, d\tilde{Y}_{i+d}^n(s) = 0 \quad \text{for } i = 1,\ldots,d.$$

Let $f\colon (b,z) \in \mathbb{R}^2 \to f(b,z) = (b-z) \wedge 1$. Clearly, $f \in C_b(\mathbb{R}^2)$. Then the second equation in (4.21) follows from (3.10) and Lemma 3.3 in [4] or Lemma 8.3 in [5]. Similarly, one can prove the first equation in (4.21). $\square$

5. Proofs of Lemmas 3.1 and 3.2

The existence of the inverse matrix $(I + G)^{-1}$ is based on the following observation. From the routing structure of the network, we know that the entries of matrix $I + G$ are given by

$$L_{ji} = \begin{cases} 1 & \text{if } j = i, \\ \sum_k \alpha_k \sum_l m_l \sum_{k_1} \sum_{k_2} \cdots \sum_{k_a} p_{kk_1} \cdots p_{k_a l} & \text{if } \sigma(i) = k_1,\ \sigma(k_1) = k_2,\ldots,\sigma(k_a) = j, \\ 0 & \text{otherwise,} \end{cases}$$

where a is the intermediate station number for a customer to route from station i to station j and $k \in C(i)$, $k_1 \in C(\sigma(i))$, $k_2 \in C(\sigma(k_2))$, ..., $k_a \in C(\sigma(k_{a-1}))$, $l \in C(j)$. It is easy to see that $I + G$ is a lower triangle matrix and the entries in the diagonal are ones. Therefore $(I + G)^{-1}$ exists and is also a lower triangle matrix with ones in the diagonal.

Under the heavy traffic condition, the reflection matrix in Lemma 3.2 can be written as $[(I + G)^{-1}, -D]$. The matrix D is a $d \times d$ upper triangle matrix with the entries $D_{ii} = 1$, $D_{ij} = -1$ for $\sigma(i) = j$ and $D_{ij} = 0$ otherwise. Thus both $(I + G)^{-1}$ and D are completely-$\mathscr{S}$ matrix. The reflection matrix R^α associated with the vertex α defined in (3.18) has the following decomposition form

$$R^\alpha = \begin{pmatrix} A_1 & A_2 \\ A_3 & A_4 \end{pmatrix},$$

where A_1 and A_4 are principal submatrices of $(I+G)^{-1}$ and D respectively, A_2 is nonnegative matrix. Then we know that R^α is a completely-$\mathscr{S}$ matrix. $\square$

References

[1] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, Queueing Systems: Theory and Applications 30 (1998) 89–148.

[2] H. Chen, H. Zhang, A sufficient condition and a necessary condition for the diffusion approximations of multiclass queueing networks under priority service displines, Queueing Systems: Theory and Applications 34 (2000) 237–268.

[3] D. Cheng, D.D. Yao, Tandem queues with general blocking: a unified model and comparison results, Discrete, Event Dyn. Systems 2 (1993) 207–234.

[4] W. Dai, Brownian approximations for queueing networks with finite buffers: modeling, heavy traffic analysis and numerical implementations, Ph.D. Thesis, School of Mathematics, Georgia Institute of Technology, 1996. Also published in UMI Dissertation Services, A Bell & Howell Company, 300 N.Zeeb Road, Ann Arbor, Michigan 48106, USA, 1997.

[5] J.G. Dai, W. Dai, A heavy traffic limit theorem for a class of open queueing networks with finite buffers, Queueing Systems 32 (1999) 5–40.

[6] J.G. Dai, V. Nguyen, On the convergence of multiclass queueing networks in heavy traffic, Ann. Appl. Probab. 4 (1994) 26–42.

[7] J.G. Dai, R. Williams, Existence and uniqueness of semimartingale reflecting Brownian motions in convex polyhedrons, Theory probab. Appl. 40 (1995) 1–40.

[8] A.I. Elwalid, D. Mitra, Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation, Queueing Systems 9 (1991) 29–64.

[9] S.N. Ethier, T.G. Kurtz, Markov Processes: Charaterization and Convergence, Wiley, New York, 1986.

[10] D.L. Iglehart, W. Whitt, Multiple channel queues in heavy traffic I, Adv. Appl. Prob. 2 (1970) 150–177.

[11] H. Kroner, M. Eberspacher, T.H. Theimer, P.J. Kuhn, U. Briem, Approximate analysis of the end to end delay in ATM networks, Proceedings of the IEEE INFOCOM '92, Florence, Italy, 1992, pp. 978–986.

[12] R.J. Williams, Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse, Queueing Systems 30 (1998) 27–88.

[13] X. Shen, H. Chen, J.G. Dai, W. Dai, The finite element method for computing the stationary distribution of an SRBM in a hypercube with applications to finite buffer queueing networks (2000) under revision for Queueing Systems.