DIFFUSION APPROXIMATIONS FOR MULTICLASS FEEDFORWARD QUEUEING NETWORKS WITH ABANDONMENTS UNDER FCFS SERVICE DISCIPLINES

TOSHIYUKI KATSUDA *

Received February 26, 2016 ; revised August 6, 2016

ABSTRACT. We consider multiclass feedforward queueing networks with abandonments under FCFS (first-come, first-served) service disciplines and prove a diffusion approximation theorem for the queue lengths and workloads in those networks under heavy traffic. The diffusion limit is the unique solution to a multidimensional reflected stochastic differential equation with a nonlinear drift term as the limit of abandonment-count process. The desired convergence is shown by taking the following steps: first, obtaining the stochastic boundedness of (scaled) workload in use of the feedforward property of class routing; second, proving the C-tightness of abandonment-count process; third, establishing the condition of state-space collapse; fourth, showing the C-tightness of workload. In the final step we prove the uniqueness (in law) of the solution to the limit equation for workload by reducing it to the uniqueness of a semimartingale reflecting Brownian motion via the Girsanov transformation technique.

1 Introduction.

In this paper we are concerned with multiclass feedforward queueing networks with customer abandonments in heavy traffic. Generally queueing network models have been used to analyze systems arising in a wide range of computer systems, communication networks and complex manufacturing systems. Many of those systems have stations which process more than one class of customers (or jobs) and also have complex structures of class routings after the processing of customers. So the model of multiclass queueing networks has been developed for the analysis of such systems. In particular, the heavy load of those networks is a compelling problem to solve, and thus the diffusion (or heavy-traffic) approximation of such networks has been wanted and pursued. At the same time, because it is natural to suppose that no customer has infinite patience in waiting for service in a queue, the phenomenon of customer abandonment is ubiquitous in various queue models for real applications such as telephone call centers, transmission channels and manufacturing industries, in which impatient customers faced with some waiting time leave the system without receiving service. For example, in the context of wireless communication networks, data packets are lost unless they are transmitted by some deadline.

In multiclass queueing networks (MQNs) under study, customers are categorized into $K(\geq 1)$ classes and the network is composed of $J(\geq 1)$ service stations with unlimited capacity where $J \leq K$. Customers of each class $k \in \mathbb{K}(\equiv \{1, \ldots, K\})$ arrive from outside the network and they will receive service exclusively at station $j = s(k)$ where $s(\cdot)$ maps $\mathbb{K}$ onto $\mathbb{J}(\equiv \{1, 2, \ldots, J\})$ in a many-to-one fashion. In such networks customers change their classes on their service completions. In particular, we restrict our attention to multiclass *feedforward* queueing networks in which at the class change of a customer he either flows

2000 *Mathematics Subject Classification.* 60K25, 60F17, 90B22, 60J25, 93E15 .

Key words and phrases. diffusion approximation, multiclass feedforward queueing network, customer abandonment, state-space collapse.

*School of Science and Technology, Kwansei Gakuin University.

from a lower numbered station to a higher numbered one, or remains in the original station (as a new class customer). After at most a finite number of such class changes, customers will eventually leave the network. In this paper, the FCFS (firsr-come, first-served) service discipline is investigated in our multiclass feedforward queueing networks with abandonments and we establish the diffusion approximation for those networks in heavy traffic.

Related research. Diffusion approximations for (single-class) generalized Jackson queueing networks (GJNs) in heavy-traffic were established in Reiman [22] under typical moment conditions on primitive variables of the network. However, some counterexamples were found to the validity of heavy-traffic limit for multiclass queueing networks (MQNs) (cf. Dai and Wang [9]), which is in contrast with the case of GJNs. So the identification of the category of the MQNs subject to the heavy-traffic analysis has been one of the main topics in queueing theory. Due to the feature that a single server processes more than one class of customers in MQNs and also to the class-transition nature of a customer, the increased complexity is brought so that the heavy-traffic limit of scaled K-dimensional queue length vector in an MQN is understood to be difficult to obtain without additional restrictive conditions not appearing in such limits of GJNs.

In late 1990s, such problem was solved by Bramson [3] and Williams [26] for some types of MQN with important service disciplines such as FCFS, processor-sharing and buffer-priority ones. More specifically, Williams [26] established heavy-traffic limit theorems for MQNs with the limit referred to as a semimartingale reflecting Brownian motion, assuming the condition of *state-space collapse*. Loosely speaking, state-space collapse corresponds to an asymptotic-law version of Little's formula for MQNs in heavy traffic. Further, [26] indicated that state-space collapse is also a necessary condition for the heavy-traffic limit theorem in MQNs with FCFS disciplines. (Cf. Appendix B in [26]). At the same time, Bramson [3] constructed the framework on state-space collapse for MQNs in which the initial condition on *strong* state-space collapse is proved to imply *multiplicative* strong state-space collapse (cf. Theorem 1 in [3]), which forms the basis for the use of state-space collapse in [26]. In addition, [3] showed that state-space collapse is exhibited after a brief period of time under the relative compactness (tightness) of initial scaled workload (cf. Theorem 3 in [3]), which is used to prove that state-space collapse holds for a multiclass single-server queue in stationarity (cf. Katsuda [15]).

On the other hand, for the last decade, the study of a many-server queue with abandonment in the so-called Halfin-Whitt heavy-traffic regime has attracted considerable attention, because it is relevant to practical large-scale service systems such as call centers. (Cf. Dai and He [8] and references therein). Furthermore, the heavy-traffic analysis of a (single-class) single-server queue, and more generally, that of a GJN are associated with customer abandonment. (Cf. Ward and Glynn [23], [24], Reed and Ward [21] for the former study, and Huang and Zhang [13] for the latter). In particular, the works [24] and [21] identified a reflected Ornstein-Uhlenbeck process and a more general reflected diffusion process, respectively, as the heavy-traffic limit of a GI/GI/1(+GI) queue with abandonment. In all of those works, for the scaling of abandonment (or, patience time) distribution, the continuous or locally-bounded hazard-rate scaling and more generally, the locally-Lipschitz hazard-type scaling were employed because of their technical tractability. ¿From a unified point of view, those scalings are extended to the most general hazard-type one by Katsuda [17] for a G/Ph/n+GI queue in the Halfin-Whitt regime. According to such general scaling, practical and yet previously intractable examples of abandonment distribution become subject to the analysis of diffusion approximation. For instance, the Gamma distribution with scale parameter less than unity is such case. (See the introduction of Katsuda [17]).

Main result. In this paper we will state and prove a diffusion approximation for a multiclass feedforward queueing network with abandonment under the FCFS service disci-

pline. Our main result is a generalization of two previous works [24] and [21] cited above. Specifically, we extend their diffusion approximation results via a one-dimensional Ornstein-Uhlenbeck type diffusion for a GI/GI/1+GI queue to a multiclass feedforward queueing network with GI-type abandonment. Furthermore we employ the general hazard-type scaling of abandonment distribution which includes the locally Lipschitz hazard-type scaling used in [24] and [21]. Our limit process for (scaled) workload is the unique solution to a multidimensional reflected stochastic differential equation with a nonlinear drift and the limit for queue length in each class is a constant times the limit of workload at the station serving the class, which is a consequence of state-space collapse for our queueing network with abandonment.

Methodology. In addition to the i.i.d. (independent and identically distributed) condition of primitive model variables with general probability distributions and also their parameters convergence, we impose the following four main assumptions:
(A.1) Initial condition on the weak convergence of (scaled) workload.
(A.2) Initial condition on strong state-space collapse.
(A.3) Tightness of initial queue length.
(A.4) Completely-S condition of reflection matrix in the limit equation for the workload.

To derive the diffusion approximation result from those assumptions, the following steps will be taken in our argument:
Step 1. Using assumptions (A.1) and (A.3), we show the stochastic boundedness of scaled queue length and workload in our queueing network with abandonment. In particular, the feedforward property of class routing is crucial to this step.
Step 2. For each $k \in \mathbb{K}$, the C-tightness of scaled abandonment-count process of class k is proved, using the stochastic boundedness of scaled workload in Step 1.
Step 3. According to (A.2) and Step 1, the condition corresponding to strong state-space collapse in a multiclass FCFS queueing network (without abandonment) is shown. Combining it with the condition characterizing the FCFS discipline with abandonment, we have state-space collapse for our queueing network with abandonment.
Step 4. Using the results of Step 2 and Step 3, we have the C-tightness of the sequence of scaled workloads satisfying the heavy-traffic condition, and then derive a J-dimensional reflected stochastic differential equation (SDE) satisfied by *any* limit process of the sequence.
Step 5. Observe that our limit SDE has a *nonlinear* drift term as the limit of scaled abandonment-count process due to the general hazard-type scaling of abandonment distribution. (The solution to the equation may be regarded as a semimartingale reflecting Brownian motion (SRBM) with a nonlinear drift term). Thus, applying the Girsanov transformation to the localized SDE and using (A.4), the uniqueness in law of the solution to the original SDE is achieved. Consequently we have the desired weak convergence of scaled workload to the unique solution to the SDE. The limit for queue length in each class is an immediate consequence of state-space collapse and the limit for workload at the station serving the class.

Overview of the contents. The rest of the paper is organized as follows. In Sect. 2, we introduce some primitive variables and processes for a multiclass queueing network with abandonment under study. In terms of those primitives, we construct a piecewise deterministic Markov process for the dynamical description of our queueing network in Sect. 3. In other words, the performance measures for our network are adapted to the history of the process. In Sect. 4, we state our main result, i.e., a diffusion approximation theorem for a multiclass feedforward queueing network with abandonment, and Sect. 5 is devoted to its proof, in which the methodology mentioned above are employed. In the appendix, we put some lemmas used in the demonstration of state-space collapse in Sect. 5.

Notation. For a random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the expectation of X on an event $A \in \mathcal{F}$ is denoted by $\mathbb{E}_{\mathbb{P}}[X; A]$. For a local martingale M, the optional quadratic variation process of M is denoted by $[M]$. (Cf. (1.8.3) in Liptser and Shiryayev [20]).

The symbols $\mathbb{Z}$, $\mathbb{N}$, $\mathcal{R}^1$ and $\mathcal{R}^1_+$ denote the set of integers, positive integers, real numbers and nonnegative real numbers, respectively. For $a, b \in \mathcal{R}^1$, $a \wedge b \equiv \min\{a, b\}$, $a \vee b \equiv \max\{a, b\}$, $a^+ \equiv a \vee 0$, $a^- \equiv (-a) \vee 0$, $\lfloor a \rfloor \equiv \max\{i \in \mathbb{Z} : i \leq a\}$ and $\lceil a \rceil \equiv \max\{i \in \mathbb{Z} : i < a\}$.

For $d \in \mathbb{N}$, $\mathcal{R}^d$ denotes the d-dimensional Euclidean space. Every vector in $\mathcal{R}^d$ is envisioned as a column vector. For example, $a = (a_k, k \in \mathcal{L})$ denotes the L-dimensional column vector with L the number of elements in the index set $\mathcal{L}$. The transpose of a vector or a matrix is denoted by putting a tilde on its top. The vector $e \in \mathcal{R}^d$ denotes $(1, 1, \ldots, 1)$. The norm $|u|$ of a vector $u = (u_1, \ldots, u_d) \in \mathcal{R}^d$ is defined by $|u| = |u_1| + \cdots + |u_d|$. The matrix $diag(u)$ with a vector $u = (u_1, \ldots, u_d) \in \mathcal{R}^d$ denotes the $d \times d$ diagonal matrix with (i, i)-diagonal element equal to u_i, $i = 1, \ldots, d$.

The space of functions $f : [0, \infty) \to \mathcal{R}^d$ that are right-continuous on $[0, \infty)$ and have left-hand limits in $(0, \infty)$ is denoted by $\mathbb{D}([0, \infty), \mathcal{R}^d)$ or simply by $\mathbb{D}^d$. The space $\mathbb{D}^d$ is endowed with the Skorohod J_1-topology. Similarly, the space of $\mathcal{R}^d$-valued continuous functions on $[0, \infty)$ is denoted by $\mathbb{C}([0, \infty), \mathcal{R}^d)$. For $f \in \mathbb{D}^d$ and $t > 0$, $f(t-)$ denotes its left-hand limit at t and $\Delta f(t) \equiv f(t) - f(t-)$. For a sequence of random elements $\{X^r\}_{r \geq 1}$ taking values in a metric space $\mathfrak{S}$, the symbol $X^r \Longrightarrow X$ *in* $\mathfrak{S}$ as $r \to \infty$ means the weak convergence of X^r to X in $\mathfrak{S}$ as the index r tends to infinity.

2 Multiclass feedforward queueing networks with abandonments and their Markovian description of dynamics

2.1 Model primitives In this section we first introduce some primitive random variables (r.v.'s) on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to construct the model of a multiclass queueing network with abandonment studied in this paper. The network is composed of J service stations indexed by $j = 1, \ldots, J$, and the set of service stations is denoted by $\mathbb{J} = \{1, 2, \ldots, J\}$. Each of the service stations has a single server and a waiting buffer of unlimited capacity. Each customer (or job) belongs to one of K classes with $K \geq J$, indexed by $k = 1, \ldots, K$, and the set of the classes is denoted by $\mathbb{K} = \{1, 2, \ldots, K\}$. For each $k \in \mathbb{K}$, customers of class k are served at service station $s(k) \in \mathbb{J}$ exclusively. The mapping $s(\cdot)$ maps $\mathbb{K}$ onto $\mathbb{J}$ in a many-to-one fashion. In addition, we let $C(j) = \{k \in \mathbb{K} : s(k) = j\}$, $j \in \mathbb{J}$.

Customers of classes in $\mathbb{A}$, which is a non-empty subset of $\mathbb{K}$, enter the network from outside and no external arrival is allowed for any class in $\mathbb{K} - \mathbb{A}$. Upon arrival, a customer is assigned the abandonment time (or, patience time) whose probability law depends on his class, and if the time until the customer is supposed to enter service, called the offered waiting time, exceeds his abandonment time, then he will abandon the system as soon as his remaining abandonment time is exhausted. Otherwise, i.e., if the customer is supposed to receive service eventually, he is assigned the service time on his arrival, which also depends on his class. The service of customers by the server is performed according to the first-come-first-service (FCFS) discipline, i.e., in the order of their arrivals independently of their classes. (We also take the convention that customers within each class are numbered on the first-in basis). On service completion, a customer either changes his class and waits for service as the new class customer in the end of the queue, or leaves the system.

External arrivals

The *external arrival process* $E(t) = \{E_k(t), k \in \mathbb{K}\}, t \geq 0$, counts the number of arrivals at each class from outside the network. For each $k \in \mathbb{A}$, we define $E_k(\cdot)$ by

$$E_k(t) \equiv \max\{n \in \mathcal{N} : \mathcal{U}_k(n) \leq t\}$$

with $\max \phi \equiv 0$, where

$$\mathcal{U}_k(n) \equiv \sum_{i=1}^{n} u_k(i) \tag{1}$$

with $\mathcal{U}_k(0) \equiv 0$. For each $k \in \mathbb{A}$, the *external interarrival times* $\{u_k(i), i = 2, 3, \ldots\}$ are i.i.d. (independent and identically distributed) positive r.v.'s with the distribution function (d.f.)

$$F_k^u(x) \equiv \mathbb{P}(u_k(2) \leq x), \qquad x \geq 0,$$

the mean $1/\alpha_k \equiv \int_0^\infty x dF_k^u(x) > 0$, and the finite variance $a_k \equiv \int_0^\infty (x - \frac{1}{\alpha_k})^2 dF_k^u(x) \geq 0$. The r.v. $u_k(1) > 0$, corresponding to the remaining interarrival time of the customer entering first after time $t = 0$, is independent of $\{u_k(i), i = 2, 3, \ldots\}$. For each $i = 2, 3, \ldots$, the r.v. $u_k(i)$ corresponds to the interarrival time between the $(i-1)$-th customer and i-th customer in class k. For conveniece, we set

$$E_k(\cdot) \equiv 0 \quad \text{and} \quad \alpha_k = 0$$

for $k \in \mathbb{K} - \mathbb{A}$. The vector $\alpha = (\alpha_k, k \in \mathbb{K})$ is referred to as the *arrival rate*.

Service times

For each $k \in \mathbb{K}$, there are two sequences of service times, i.e., a sequence of *original* service times and a sequence of *subsequent* service times. The sequence of original service times $\{v_k^o(i), i = 1, 2, \ldots\}$ gives the (remaining) service times for class k customers who are in the system at time 0 and will eventually receive service. (There are more elements in the infinite sequence than needed). Those initial customers are assumed to have the prescribed order of arrivals at or before time 0, and if there is such i-th customer in the system, the original service time $v_k^o(i)$ is assigned to him for $i = 1, 2, \ldots$.

For each $k \in \mathbb{K}$, the *original service times* $\{v_k^o(i), i = 2, 3, \ldots\}$ are i.i.d. positive r.v.'s with

$$F_k^v(x) \equiv \mathbb{P}\big(v_k^o(2) \leq x\big), \quad x \geq 0, \tag{2}$$

the mean $m_k = \int_0^\infty x dF_k^v(x) > 0$ and the finite variance $b_k \equiv \int_0^\infty (x - m_k)^2 dF_k^v(x) \geq 0$. The constant $\mu_k \equiv 1/m_k$ is referred to as the *service rate* of class k. The r.v. $v_k^o(1)$, corresponding to the (remaining) service time of initial class k customer who arrived the longest time ago among those eventually receiving service, is independent of $\{v_k^o(i), i = 2, 3, \ldots\}$. The *cumulative original service time process* $\mathcal{V}_k^o(n), n \in \mathbb{N}, k \in \mathbb{K}$, is given by

$$\mathcal{V}_k^o(n) \equiv \sum_{i=1}^{n} v_k^o(i) \tag{3}$$

with $\mathcal{V}_k^o(0) \equiv 0$.

The *subsequent service times* $\{v_k^s(i), i = 1, 2, \ldots\}$, $k \in \mathbb{K}$, are i.i.d. positive r.v.'s with $\mathbb{P}\big(v_k^s(1) \leq x\big) = F_k^v(x), x \geq 0$. For each $k \in \mathbb{K}$, $v_k^s(i)$ corresponds to the service time assigned to the i-th class k customer among those arriving after $t = 0$ from outside or due

to class change and eventually receiving service. The *cumulative subsequent service time process* $\mathcal{V}_k^s(n), n \in \mathbb{N}, k \in \mathbb{K}$, is given by

$$\mathcal{V}_k^s(n) \equiv \sum_{i=1}^{n} v_k^s(i) \tag{4}$$

with $\mathcal{V}_k^s(0) \equiv 0$.

Abandonment times

Similar to the service times above, we introduce the abandonment times in two distinct sequences, i.e., the *original* abandonment times and *subsequent* abandonment times. For each $k \in \mathbb{K}$, the original abandonment times $\{\gamma_k^o(i), i = 1, 2, \ldots\}$ is a sequence of independent positive r.v.'s which corresponds to the remaining abandonment times of the customers of class k initially at the network. (The assignment of those abandonment times to each customer is done in the same way as in service times, but distinct to that case, the abandonment time is assigned to *every* customer at the system, whether he will abandon it or not). For each $k \in \mathbb{K}$, the subsequent abandonment times $\{\gamma_k^s(i), i = 1, 2, \ldots\}$ are i.i.d. positive random variables with

$$F_k^\gamma(x) \equiv \mathbb{P}(\gamma_k^s(1) \leq x), \qquad x \geq 0, \tag{5}$$

and correspond to the abandonment times assigned to the customers of class k arriving after $t = 0$.

Class routings

The *class-routing process* $\Phi(n) = \{\Phi^k(n), k \in \mathbb{K}\}$, $n \in \mathbb{N}$, is defined by

$$\Phi^k(n) \equiv \sum_{i=1}^{n} \phi^k(i)$$

where $\{\phi^k(i) = (\phi_l^k(i), l \in \mathbb{K}), i = 1, 2, \ldots, \}$ are i.i.d. random vectors taking values in the set $\{0, e_1, \ldots, e_K\}$ with e_k denoting the unit basis vector parallel to the k-th coordinate axis in $\mathcal{R}^K$, $k \in \mathbb{K}$. The identity $\phi^k(i) = e_l$ indicates that the i-th customer served at class k changes his class to class l after the service, and the identity $\phi^k(i) = 0$ indicates his departure from the system.
Let $P_{kl} = \mathbb{P}(\phi^k(1) = e_l)$ and $P_{k0} = \mathbb{P}(\phi^k(1) = 0)$, $k, l \in \mathbb{K}$. Then the $K \times K$ substochastic matrix $P = [P_{kl}; k, l \in \mathbb{K}]$, called the *class-routing* matrix, is assumed to have spectral radius strictly less than unity. Thus

$$Q \equiv (I - \widetilde{P})^{-1} = I + \widetilde{P} + (\widetilde{P})^2 + \cdots$$

is finite where $\widetilde{P}$ denotes the transpose of P. It is readily seen that for each $k \in \mathbb{K}$,

$$\begin{aligned} \mathbb{E}[\phi^k(1)] &= P_{k\cdot} \qquad \text{and} \\ \mathbb{C}ov[\phi^k(1)] &\equiv [\mathbb{C}ov(\phi_l^k(1), \phi_m^k(1)),\ l, m \in \mathbb{K}] \\ &= \Upsilon^k \end{aligned} \tag{6}$$

where $P_{k\cdot}$ denotes the k-th row vector of P and Υ^k denotes the $K \times K$ matrix such that

$$\Upsilon_{lm}^k = \begin{cases} P_{kl}(1 - P_{kl}) & \text{if} \quad l = m, \\ -P_{kl}P_{km} & \text{if} \quad l \neq m. \end{cases} \tag{7}$$

In this paper we will impose on the class-routing probability $\{P_{kl}, k, l \in \mathbb{K}\}$ the following condition:

Feedforward class-routing condition

For each $k, l \in \mathbb{K}$,

$$\text{if} \quad P_{kl} > 0, \quad \text{then} \quad s(k) \leq s(l). \tag{8}$$

When $J = 1$(i.e., a multiclass single-server queue), condition (8) is obviously satisfied.

Remaining time processes

Associated with the interarrival, service and abandonment times introduced above, we define their remaining time processes as follows. For each $k \in \mathbb{K}$ and $t \geq 0$, let $\mathcal{R}_k^u(t)$ and $\mathcal{R}_k^v(t)$ denote the remaining interarrival time and remaining service time of class k customer at time t, respectively. (For $k \in \mathbb{K} - \mathbb{A}$, we set $\mathcal{R}_k^u(\cdot) \equiv -1$). In particular, $\mathcal{R}_k^u(0) = u_k(1), k \in \mathbb{A}$, $\mathcal{R}_l^v(0) = v_l^o(1), l \in \mathbb{K}$.

Now, for each $k \in \mathbb{K}$, let

$$Z_k(t), \quad t \geq 0, \tag{9}$$

denote the number of class k customers who are either being served or waiting in queue at time t, which is referred to as the *queue length* of class k at time t. Then the remaining abandonment time process of class $k, k \in \mathbb{K}$, is represented by

$$\mathcal{R}_k^\gamma(t) = (\mathcal{R}_{k,i}^\gamma(t), i = 1, 2, \ldots), \qquad t \geq 0,$$

in which, for each $1 \leq i \leq Z_k(t)$, $\mathcal{R}_{k,i}^\gamma(t)$ denotes the remaining abandonment time of i-th customer of class k at time t, and for $i \geq Z_k(t) + 1$, we set $\mathcal{R}_{k,i}^\gamma(t) \equiv -1$. In particular, $\mathcal{R}_{k,i}^\gamma(0) = \gamma_k^o(i)$ for each $1 \leq i \leq Z_k(0)$ and $k \in \mathbb{K}$. If the remaining abandonment time $\mathcal{R}_{k,1}^\gamma(\cdot)$ expires at $t = t_0$ and the service of the corresponding customer began before time t_0 and continues at $t = t_0$, then we set $\mathcal{R}_{k,1}^\gamma(t) \equiv 0$ for each $t \in [t_0, t_1)$ where t_1 denotes the time at which the service finishes.

Class designation processes

Relevant to the FCFS discipline investigated in this paper, we have to track the designation of the class of each customer in each service station in order to describe the dynamics of the network. For the purpose, we introduce the $\{0, 1, \ldots, 2K\}^\infty$-valued process

$$O(t) = (O_j(t),\, j \in \mathbb{J}), \qquad t \geq 0, \tag{10}$$

where

$$O_j(t) = (O_{j,i}(t),\, i \geq 1), \quad j \in \mathbb{J},$$

and for $j \in \mathbb{J}$ and $1 \leq i \leq \sum_{m \in C(j)} Z_m(t)$,

$$O_{j,i}(t) \equiv \begin{cases} k & \text{if } i\text{-th customer in the queue of station } j \text{ at time } t \text{ is} \\ & \text{of class } k \text{ and will eventually receive service;} \\ K + l & \text{if } i\text{-th customer in the queue of station } j \text{ at time } t \text{ is} \\ & \text{of class } l \text{ and will eventually abandon the system,} \end{cases} \tag{11}$$

and for $i \geq \sum_{m \in C(j)} Z_m(t) + 1$, we set $O_{j,i}(t) \equiv 0$. (The variable $O_{j,1}(t)$ corresponds to the class of the customer being served at time t, whenever $\sum_{m \in C(j)} Z_m(t) \geq 1$).

Note that under our assumptions on the primitives, simultaneous (exogenous or internal) arrivals of customers from different classes are allowed. So, to determine the components of the process $O(\cdot)$ without ambiguity, a rule is needed for the specification of the ordering of such customers. Following page 41 of Williams [26], we henceforth take a deterministic tie breaking rule to treat that case. For example, we adopt the convention that for customers with simultaneous arrivals, a customer of higher numbered class is ordered ahead of a customer of lower numbered class in the queue of each station.

Offered waiting times

To determine whether each customer will abandon the network or not either on his arrival to a class or at initial instant, we assign to him the *offered waiting time* as follows. For each $k \in \mathbb{K}$ and $i = 1, 2, \ldots$, the *original offered waiting time* $w_k^o(i)$ is the amount of time the i-th customer of class k initially in the system would have to wait in queue (i.e., waiting line) until getting into service if his abandonment time were infinite, with the convention that $w_k^o(i) \equiv 0$ for $i \geq Z_k(0)+1$. Thus, if $\gamma_k^o(i) \leq w_k^o(i)$, then such i-th class k customer will eventually abandon the network, and otherwise, he will receive service of class k. Similarly, for each $k \in \mathbb{K}$ and $i = 1, 2, \ldots$, the *subsequent offered waiting time* $w_k^s(i)$ is such amount of time for the i-th customer arriving at class k from outside or from other classes due to class change after $t = 0$.

Specifically, $w_k^s(i)$ is $\mathcal{G}_k^s(i)$-measurable for each $i = 1, 2, \ldots$ and $k \in \mathbb{K}$, where

$$\begin{aligned}
&\mathcal{G}_k^s(i) \\
&\equiv \sigma\{u_k(m+1), v_k^s(m), \gamma_k^s(m), m \leq i-1\} \vee \bigvee_{l \in \mathbb{K}, l \neq k} \sigma\{u_l(m), v_l^s(m), \gamma_l^s(m), m \geq 1\} \\
&\vee \bigvee_{p \in \mathbb{K}} \sigma\{v_p^o(m), \gamma_p^o(m), \phi^p(m), m \geq 1\} \vee \sigma\{O(0)\}. \qquad (12)
\end{aligned}$$

Mutual independence assumption on the primitives

Finally in this subsection, we impose the following mutual independence assumption on the primitive variables introduce so far, which is fundamental to our argument in the rest of the paper:

The families of variables

$$\begin{aligned}
&\{\mathcal{R}^v(0), \mathcal{R}^\gamma(0), O(0)\}, \{\mathcal{R}_k^u(0) = u_k(1)\}, \ \ k \in \mathbb{A}, \\
&u_{k'}^*, \ \ k' \in \mathbb{A}, \quad v_1^{o,*}, \cdots, v_K^{o,*}, \\
&v_1^s, \cdots, v_K^s, \ \ \gamma_1^s, \cdots, \gamma_K^s, \ \ \phi^1, \cdots, \phi^K \qquad (13)
\end{aligned}$$

are mutually independent, where

$$\begin{aligned}
&v_k^{o,*} \equiv (v_k^o(i), i \geq 2), \ \ k \in \mathbb{K}, \\
&u_{k'}^* \equiv (u_{k'}(i), i \geq 2), \ \ k' \in \mathbb{A}, \quad v_l^s \equiv (v_l^s(i), i \geq 1), \ \ l \in \mathbb{K}, \\
&\gamma_p^s \equiv (\gamma_p^s(i), i \geq 1), \ \ p \in \mathbb{K}, \quad \phi^q \equiv (\phi^q(i), i \geq 1), \ \ q \in \mathbb{K}.
\end{aligned}$$

2.2 Performance measure processes and their equation As the performance measures for our multiclass queueing network with abandonment, we define the following processes:

The K-dimensional (column) vector-valued process

$$Z(t) = (Z_k(t), k \in \mathbb{K}), \qquad t \geq 0,$$

with $Z_k(t)$ in (9) is referred to as the *queue length* process. For each $j \in \mathbb{J}$, let

$$W_j(t), \quad t \geq 0,$$

denote the total amount of immediate work (measured in units of service time) embodied by the customers in the station j at time t. Set

$$W(t) = (W_j(t), j \in \mathbb{J}), \quad t \geq 0,$$

which is referred to as the *workload* process. Also, for each $j \in \mathbb{J}$,

$$Y_j(t), \quad t \geq 0,$$

denotes the cumulative amount of time that the server at station j is idle during the time interval $(0, t]$, and set

$$Y(t) = (Y_j(t), j \in \mathbb{J})$$

that is referred to as the cumulative idle time process. To describe the dynamics of $Z(\cdot)$, $W(\cdot)$ and $Y(\cdot)$, we also introduce the following processes.

For each $k \in \mathbb{K}$ and $t \geq 0$, $A_k(t)$ denotes the total number of the (exogenous and internal) arrivals of class k customers during $(0, t]$, $D_k(t)$ denotes the total number of the service completions of class k customers during $(0, t]$, $I_k(t)$ denotes the total number of the abandonments of class k customers during $(0, t]$, and $T_k(t)$ denotes the total amount of time that the server has processed customers of class k during $(0, t]$. Furthermore, let $A_k^+(t)$ denote the number of customers who arrive at class k during $(0, t]$ and will eventually receive service (and not abandon), and let $Z_k^+(t)$ denote the number of class k customers who are either being under service or waiting in queue at time t and going to receive service.

We represent those processes in (column) vector form as

$$\begin{aligned}
A(t) &= (A_k(t), k \in \mathbb{K}), \\
A^+(t) &= (A_k^+(t), k \in \mathbb{K}), \\
D(t) &= (D_k(t), k \in \mathbb{K}), \\
I(t) &= (I_k(t), k \in \mathbb{K}), \\
T(t) &= (T_k(t), k \in \mathbb{K}), \\
Z^+(t) &= (Z_k^+(t), k \in \mathbb{K}), \quad t \geq 0.
\end{aligned}$$

Let

$$\mathfrak{X}(t) \equiv (A(t), A^+(t), D(t), I(t), T(t), W(t), Y(t), Z(t), Z^+(t)), \quad t \geq 0, \tag{14}$$

and the process $\mathfrak{X}(\cdot)$ is called the *performance measure process* for our multiclass queueing network with abandonment. Then the dynamical equation for the components of $\mathfrak{X}(t), t \geq 0$,

is represented as follows:

$$A(t) = E(t) + F(t) \tag{15}$$

$$\text{with} \quad F(t) = \sum_{k=1}^{K} \Phi^k(D_k(t)), \tag{16}$$

$$Z(t) = Z(0) + A(t) - D(t) - I(t), \tag{17}$$

$$Z^+(t) = Z^+(0) + A^+(t) - D(t) \tag{18}$$

$$\text{with} \quad Z_k^+(0) \equiv \sum_{i=1}^{Z_k(0)} 1_{\{w_k^o(i) < \gamma_k^o(i)\}} \tag{19}$$

$$\text{and} \quad A_k^+(t) \equiv \sum_{i=1}^{A_k(t)} 1_{\{w_k^s(i) < \gamma_k^s(i)\}}, \quad k \in \mathbb{K}, \tag{20}$$

$$W(t) = W(0) + C\mathcal{V}^s(A^+(t)) - CT(t) \tag{21}$$

$$\text{with} \quad W(0) = C\mathcal{V}^o(Z^+(0)), \tag{22}$$

$$CT(t) + Y(t) = t, \tag{23}$$

$$\int_0^\infty W_j(s) dY_j(s) = 0, \quad \forall j \in \mathbb{J}, \tag{24}$$

for all $t \geq 0$, where $C = [C_{jk}, j \in \mathbb{J}, k \in \mathbb{K}]$ is the $J \times K$ matrix with

$$C_{jk} = \begin{cases} 1, & \text{if } j = s(k): \\ 0, & \text{otherwise.} \end{cases}$$

Associated with the *abandonment-count* process $I_k(\cdot), k \in \mathbb{K}$, we now define the process $N_k(\cdot), k \in \mathbb{K}$, by

$$N_k(t) \equiv Z_k^-(0) + A_k^-(t), \qquad t \geq 0, \tag{25}$$

where

$$Z_k^-(0) \equiv \sum_{i=1}^{Z_k(0)} 1_{\{\gamma_k^o(i) \leq w_k^o(i)\}} = Z_k(0) - Z_k^+(0), \tag{26}$$

$$A_k^-(t) \equiv \sum_{i=1}^{A_k(t)} 1_{\{\gamma_k^s(i) \leq w_k^s(i)\}} = A_k(t) - A_k^+(t). \tag{27}$$

We observe that under the FCFS service discipline, for each $k \in \mathbb{K}$, $t \geq 0$ and $\varepsilon > 0$,

$$N_k(\zeta_{s(k)}(t) - \varepsilon) \leq I_k(t) \leq N_k(t) \tag{28}$$

with

$$\zeta_j(t) \equiv \inf\{s \geq 0 : s + W_j(s) > t\}, \qquad j \in \mathbb{J}, \tag{29}$$

and

$$Z_k^-(t) \leq I_k(t + W_{s(k)}(t)) - I_k(t) \tag{30}$$

with

$$Z_k^-(t) \equiv Z_k(t) - Z_k^+(t). \tag{31}$$

2.3 Markovian description of a multiclass queueing network with abandonment In the following we introduce the Markovian description process for a multiclass queueing network *with* abandonment in a similar way to Katsuda [15]. The process will be constructed from the primitive variables and the associated processes introduced so far. Conversely those primitives can also be represented by the description process.

Let

$$V(t) \equiv (V_k(t), k \in \mathbb{K})$$

where $V_k(t) \equiv (V_{k,i}(t), i = 1, 2, \ldots)$ with

$$V_{k,1}(t) \equiv \mathcal{R}_k^v(t),$$

for $2 \le i \le Z_k^+(t)$,

$$V_{k,i}(t) \equiv \begin{cases} v_k^o(D_k(t) + i), & \text{if} \quad D_k(t) + i \le Z_k^+(0), \\ v_k^s(D_k(t) + i - Z_k^+(0)), & \text{otherwise,} \end{cases}$$

and for $i \ge Z_k^+(t) + 1$,

$$V_{k,i}(t) \equiv 0.$$

We define the stochastic process $\Xi = (\Xi(t), t \ge 0)$ by

$$\Xi(t) \equiv (O(t), \mathcal{R}^u(t), V(t), \mathcal{R}^\gamma(t)) \tag{32}$$

where

$$\begin{aligned} O(t) &= (O_j(t), j \in \mathbb{J}) = ((O_{j,i}(t), i = 1, 2, \ldots), j \in \mathbb{J}), \\ \mathcal{R}^u(t) &= (\mathcal{R}_k^u(t), k \in \mathbb{A}), \\ \mathcal{R}^\gamma(t) &= (\mathcal{R}_k^\gamma(t), k \in \mathbb{K}) = ((\mathcal{R}_{k,i}^\gamma(t), i \ge 1), k \in \mathbb{K}). \end{aligned}$$

Then $\Xi = (\Xi(t), t \ge 0)$ is a piecewise deterministic Markov process (PDMP). Generally the PDMP is a strong Markov process. (Cf. Davis [10]).

Let

$$\mathcal{F}_t^\Xi \equiv \sigma(\Xi(s); 0 \le s \le t), \quad t \ge 0.$$

Then $(\mathcal{F}_t^\Xi)_{t \ge 0}$ is right continuous, i.e., $\cap_{n=1}^\infty \mathcal{F}_{t+\frac{1}{n}}^\Xi = \mathcal{F}_t^\Xi$ for each $t \ge 0$. As stated in the next proposition, the performance measure processes $\mathfrak{X}(\cdot)$ is $(\mathcal{F}_t^\Xi)_{t \ge 0}$-adapted. In other words, the process $\Xi(\cdot)$ describes the dynamics of our multiclass queueing network with abandonment. For this reason, the process $\Xi(\cdot)$ is called the *Markovian description process* for the network.

We denote the probability law of Markov process $\Xi(t), t \ge 0$, starting with the value $\xi \in \mathcal{S}$ by

$$\mathsf{P}_\xi(\mathsf{E}), \quad \mathsf{E} \in \mathcal{F}_\infty^\Xi \big(\equiv \bigvee_{t \ge 0} \mathcal{F}_t^\Xi\big), \quad \xi \in \mathcal{S}, \tag{33}$$

such that $\mathsf{P}_\xi(\Xi(0) = \xi) = 1$, where $\mathcal{S}$ denotes the state space of the process $\Xi(\cdot)$. For each $\mathsf{E} \in \mathcal{F}_\infty^\Xi$, $\mathsf{P}_\xi(\mathsf{E})$ is $\mathfrak{B}(\mathcal{S})$-measurable w.r.t. ξ.

Now let $\{\theta_t\}_{t \ge 0}$ denote the family of shift transformations associated with the process $\Xi(t), t \ge 0$. Namely,

$$\Xi(t) \circ \theta_s = \Xi(t + s)$$

for each $s, t \ge 0$. Corresponding to Proposition 2.1 of Katsuda [15], we have the following proposition on the shift-transformed performance measure process. (Since the proof is done in a similar way, we omit it).

Proposition 2.1.
The performance measure process

$$\mathfrak{X}(\cdot) = (A(\cdot), A^+(\cdot), D(\cdot), I(\cdot), T(\cdot), W(\cdot), Y(\cdot), Z(\cdot), Z^+(\cdot))$$

is $(\mathcal{F}_t^{\Xi})_{t\geq 0}$*-adapted. Thus* $\mathfrak{X}(\cdot)\circ\theta_t$, $t\geq 0$, *is well-defined and each component of the shift transformed process is given by the following:*

$$A(t)\circ\theta_s = A(s+t) - A(s), \tag{34}$$
$$A^+(t)\circ\theta_s = A^+(s+t) - A^+(s), \tag{35}$$
$$D(t)\circ\theta_s = D(s+t) - D(s), \tag{36}$$
$$I(t)\circ\theta_s = I(s+t) - I(s), \tag{37}$$
$$T(t)\circ\theta_s = T(s+t) - T(s), \tag{38}$$
$$W(t)\circ\theta_s = W(s+t), \tag{39}$$
$$Y(t)\circ\theta_s = Y(s+t) - Y(s), \tag{40}$$
$$Z(t)\circ\theta_s = Z(s+t), \tag{41}$$
$$Z^+(t)\circ\theta_s = Z^+(s+t), \tag{42}$$

for any $s, t \geq 0$.

The quantity $Z_k^-(t)$, defined by (31), is the number of class k customers who are in the system at time t and will eventually abandon it. According to (41) and (42),

$$Z_k^-(t) = Z_k^-(0)\circ\theta_t \tag{43}$$

for each $t \geq 0$.

The condition characterizing the FCFS discipline with abandonment is represented as

$$D_k(t + W_{s(k)}(t)) - D_k(t) + Z_k^-(t) = Z_k(t) \tag{44}$$

for each $t \geq 0$ and $k \in \mathbb{K}$. In virtue of Proposition 2.1, the identity (44) is a consequence of the operation of shift transformation $\theta_t, t \geq 0$, to the initial relation

$$D_k(W_{s(k)}(0)) + Z_k^-(0) = Z_k(0), \qquad k \in \mathbb{K}, \tag{45}$$

and can be regarded as the extension of the FCFS characterization condition without abandonment, i.e.,

$$D_k(t + W_{s(k)}(t)) - D_k(t) = Z_k(t), \qquad t \geq 0, \quad k \in \mathbb{K},$$

that is equivalent to (2.25) in Bramson [3].

3 Heavy-traffic assumptions and scaling

In the rest of the paper we consider a sequence of multiclass FCFS queueing networks with abandonments each of which satisfies the feedforward class-routing condition (8). Each network in the sequence is indexed by r, where r tends to infinity through a sequence of values in $[1, \infty)$. (Note that the index r may possibly take non-integer values). For slight abuse of notation, denote such r-th network by $\mathfrak{X}^r(\cdot)$, whose primitive variables are defined on the probability space $(\Omega^r, \mathcal{F}^r, \mathbb{P}^r)$ for each $r \geq 1$. The number of classes K, the subset $\mathbb{A}$ of $\mathbb{K}$ with exogenous arrivals, and the map $s(\cdot) : \mathbb{K} \longrightarrow \mathbb{J}$ are fixed for all $\mathfrak{X}^r(\cdot), r \geq 1$. Also the service discipline investigated is FCFS in every network of the sequence. We

put a superscript r on each of the stochastic processes, primitive variables and constants associated with them introduced so far, in order to indicate the associated network in the sequence. For example, $Z^r(\cdot)$, $A^r(\cdot)$, $A^{-,r}(\cdot)$, $v_k^{s,r}(i)$, $\gamma^{o,r}(i)$, α_k^r, etc.

On the sequence of the parameters associated with the primitive variables in $\mathfrak{X}^r(\cdot), r \geq 1$, we impose the following limit conditions:

$$\alpha_k^r \longrightarrow \alpha_k(>0) \quad \text{as } r \to \infty, \forall k \in \mathbb{A}, \tag{46}$$

$$m_k^r \longrightarrow m_k(>0) \quad \text{as } r \to \infty, \forall k \in \mathbb{K}, \tag{47}$$

$$a_k^r \longrightarrow a_k(>0) \quad \text{as } r \to \infty, \forall k \in \mathbb{A}, \tag{48}$$

$$b_k^r \longrightarrow b_k(>0) \quad \text{as } r \to \infty, \forall k \in \mathbb{K}, \tag{49}$$

$$P_{kl}^r \longrightarrow P_{kl} \quad \text{as } r \to \infty, \forall k \in \mathbb{K}, l \in \mathbb{K} \cup \{0\}, \tag{50}$$

where $P = [P_{kl}]_{k,l \in \mathbb{K}}$ is a substochastic matrix such that its spectral radius is less than unity and for each $l \in \mathbb{K} - \mathbb{A}$, there exist some $k \in \mathbb{A}$ and $m \in \mathbb{N}$ such that

$$P_{kl}^m > 0 \tag{51}$$

where $P^m \equiv [P_{kl}^m]$ with P^m denoting the m-th power of P.

We define $\lambda^r = (\lambda_k^r, k \in \mathbb{K})$ to be the unique solution to the traffic equation:

$$\lambda^r = \alpha^r + \widetilde{P}^r \lambda^r, \tag{52}$$

that is,

$$\lambda^r = Q^r \alpha^r$$

with

$$Q^r \equiv (\mathrm{I} - \widetilde{P}^r)^{-1}. \tag{53}$$

For each r and $k \in \mathbb{K}$, λ_k^r is referred to as the *nominal total arrival rate* to class k in the r-th network. It is readily seen that $\lambda = \lim_{r \to \infty} \lambda^r$ satisfies

$$\lambda_k > 0 \tag{54}$$

for each $k \in \mathbb{K}$, because of (51).
We also define

$$\rho^r \equiv CM^r \lambda^r = (\rho_j^r, j \in \mathbb{J}) \tag{55}$$

with $M^r \equiv diag(m_k^r, k \in \mathbb{K})$, which is referred to as the *traffic intensity* vector.
We impose the limit condition on the sequence $\{\rho^r\}_r$:

$$r(\rho^r - e) \longrightarrow \vartheta \tag{56}$$

as $r \to \infty$, where ϑ is some constant vector in $\mathcal{R}^J$. The condition (56) is referred to as the *heavy-traffic* condition.

In addition, to obtain the proper limit for appropriately scaled abandonment-count processes (cf. (74) below) as $r \to \infty$ under the heavy-traffic condition, we assume the following scaling condition of abandonment distribution $F_k^{\gamma,r}(x) = \mathbb{P}^r(\gamma_k^{s,r}(1) \leq x), x \geq 0, k \in \mathbb{K}, r \geq 0$:

General hazard-type scaling of abandonment distribution. (Cf. Katsuda [17]).

For each $k \in \mathbb{K}$ and $x \notin Disc(H_k)$,

$$rF_k^{\gamma,r}(rx^r) \longrightarrow H_k(x) \quad \text{as} \quad r \to \infty, \tag{57}$$

whenever $x^r \to x$ as $r \to \infty$, where $H_k(x), x \geq 0$, is a non-decreasing function and $Disc(H_k)$ is the set of discontinuities for $H_k(\cdot)$.

We impose the following uniform integrability condition:

$$\{u_k^r(2)^2\}_{r\geq 1} \quad \text{is uniformly integrable,} \tag{58}$$

$$\{v_l^{s,r}(1)^2\}_{r\geq 1} \quad \text{is uniformly integrable,} \tag{59}$$

for each $k \in \mathbb{A}$ and $l \in \mathbb{K}$. We will also assume the following three conditions on the initial primitive variables, the first two of which correspond to (3.5) in [3] and (82), (83) in [26]:

For each $k \in \mathbb{A}$, $l \in \mathbb{K}$ and $T > 0$,

$$\frac{u_k^r(1)}{r} \longrightarrow 0 \qquad \text{in pr.,} \tag{60}$$

$$\frac{v_l^{o,r}(1)}{r} \longrightarrow 0 \qquad \text{in pr.,} \tag{61}$$

$$\max_{0\leq m<rT} \left| \{\widehat{\mathcal{V}}_l^{o,r}(\overline{Z}_l^{+,r}(0))\} \circ \theta_{rm} \right| \longrightarrow 0 \qquad \text{in pr.,} \tag{62}$$

as r goes to infinity, where

$$\widehat{\mathcal{V}}^{o,r}(t) \equiv r^{-1}(\mathcal{V}^{o,r}(\lfloor r^2 t \rfloor) - m^r \cdot \lfloor r^2 t \rfloor), \tag{63}$$

$$\overline{Z}^{+,r}(t) \equiv r^{-2} Z^{+,r}(r^2 t). \tag{64}$$

(The convergence (62) is restated as

$$\mathbb{P}^r\Big(\max_{0\leq m<rT} \Big| \frac{1}{r} \times \sum_{i=1}^{Z_l^{+,r}(rm)} \big(v_l^{o,r}(i) \circ \theta_{rm} - m_l^r\big) \Big| > \varepsilon \Big) \longrightarrow 0, \quad \forall \varepsilon > 0, \Big)$$

as $r \to \infty$).

Concerned with the asymptotic behavior of the performance measures for our multiclass queueing network with abandonment under the heavy-traffic condition, we perform the diffusive and fluid scaling on the associated stochastic processes as follows:

Diffusion scaling.

$$\widehat{Z}^r(t) = r^{-1}Z^r(r^2t), \tag{65}$$

$$\widehat{Z}^{-,r}(t) = r^{-1}Z^{-,r}(r^2t), \tag{66}$$

$$\widehat{W}^r(t) = r^{-1}W^r(r^2t), \tag{67}$$

$$\widehat{Y}^r(t) = r^{-1}Y^r(r^2t), \tag{68}$$

$$\widehat{E}^r(t) = r^{-1}(E^r(r^2t) - \alpha^r r^2 t), \tag{69}$$

$$\widehat{\mathcal{V}}^{s,r}(t) = r^{-1}(\mathcal{V}^{s,r}(\lfloor r^2t\rfloor) - m^r \cdot \lfloor r^2t\rfloor), \tag{70}$$

$$\widehat{A}^r(t) = r^{-1}(A^r(r^2t) - \lambda^r r^2 t), \tag{71}$$

$$\widehat{A}^{-,r}(t) = r^{-1}A^{-,r}(r^2t), \tag{72}$$

$$\widehat{D}^r(t) = r^{-1}(D^r(r^2t) - \lambda^r r^2 t), \tag{73}$$

$$\widehat{I}^r(t) = r^{-1}I^r(r^2t), \tag{74}$$

$$\widehat{N}^r(t) = r^{-1}N^r(r^2t), \tag{75}$$

$$\widehat{S}^r(t) = r^{-1}(S^r(r^2t) - \mu^r r^2 t), \tag{76}$$

$$\widehat{\Phi}^{k,r}(t) = r^{-1}(\Phi^{k,r}(\lfloor r^2t\rfloor) - P^r_{k\cdot}\lfloor r^2t\rfloor). \tag{77}$$

Fluid scaling.

$$\overline{Z}^r(t) = r^{-2}Z^r(r^2t), \tag{78}$$

$$\overline{E}^r(t) = r^{-2}E^r(r^2t), \tag{79}$$

$$\overline{A}^r(t) = r^{-2}A^r(r^2t), \tag{80}$$

$$\overline{A}^{+,r}(t) = r^{-2}A^{+,r}(r^2t), \tag{81}$$

$$\overline{D}^r(t) = r^{-2}D^r(r^2t), \tag{82}$$

$$\overline{I}^r(t) = r^{-2}I^r(r^2t), \tag{83}$$

$$\overline{S}^r(t) = r^{-2}S^r(r^2t), \tag{84}$$

$$\overline{T}^r(t) = r^{-2}T^r(r^2t). \tag{85}$$

Finally in this section, we note the fundamental weak-convergence result that is based on the Donsker theorem for renewal processes (cf. Billingsley [2]) and the convergence of parameters (46)-(50):

$$\widehat{E}^r(\cdot) \Longrightarrow E^*(\cdot), \tag{86}$$

$$\widehat{\mathcal{V}}^{s,r}(\cdot) \Longrightarrow \mathcal{V}^*(\cdot) \tag{87}$$

$$\widehat{\Phi}^{k,r}(\cdot) \Longrightarrow \Phi^{k,*}(\cdot), \qquad k \in \mathbb{K}, \tag{88}$$

$$\widehat{S}^r(\cdot) \Longrightarrow S^*(\cdot), \tag{89}$$

$$\overline{S}^r_l(\cdot) \Longrightarrow \mu_l \iota(\cdot), \qquad l \in \mathbb{K}, \tag{90}$$

as $r \to \infty$, where

$$
\begin{aligned}
&E^*(t) = \sqrt{\Pi} \cdot B^E(t), \\
&\mathcal{V}^*(t) = \sqrt{\Sigma} \cdot B^{\mathcal{V}}(t), \\
&\Phi^{k,*}(t) = (\Phi_1^{k,*}(t), \ldots, \Phi_K^{k,*}(t)), \\
&\Phi_l^{k,*}(t) = \sum_{m=1}^{K} \left(\sqrt{\Upsilon^k}\right)_{lm} \cdot B_m^k(t), \qquad k, l \in \mathbb{K}, \\
&\iota(t) \equiv t
\end{aligned}
$$

with $B^E(\cdot)$ and $B^{\mathcal{V}}(\cdot)$ K-dimensional standard Brownian motions,

$$
(B^1(\cdot), \ldots, B^K(\cdot)) = (B_1^1(\cdot), \ldots, B_K^1(\cdot), \ldots, B_1^K(\cdot), \ldots, B_K^K(\cdot))
$$

a K^2-dimensional standard Brownian motion,

$$
\begin{aligned}
&\Pi = diag(\alpha_1^3 a_1, \ldots, \alpha_K^3 a_K), \\
&\Sigma = diag(b_1, \ldots, b_K),
\end{aligned}
$$

and Υ^k in (6) and (7) for each $k \in \mathbb{K}$. (These standard Brownian motions are mutually independent).

4 Main result; diffusion approximation theorem

To derive the diffusion approximation theorem for our multiclass feedforward queueing network with abandonment under the FCFS discipline, the following four main assumptions, i.e., (A.1)-(A.4), are imposed in addition to the conditions on primitive variables assumed so far:

(A.1) For some proper r.v. $W^*(0)$,

$$
\widehat{W}^r(0) \Longrightarrow W^*(0) \quad \text{in} \quad \mathcal{R}^J
$$

as $r \to \infty$.

(A.2) For each $k \in \mathbb{K}$,

$$
\sup_{0 \le t \le W^r_{s(k)}(0)} r^{-1} |D_k^r(t) - \lambda_k^r t| \longrightarrow 0 \quad \text{in pr.}
$$

as $r \to \infty$.

(A.3) The sequence $\{\widehat{Z}^r(0)\}_{r \ge 1}$ is tight in $\mathcal{R}^K$, i.e.,

$$
\lim_{M \to \infty} \overline{\lim_{r \to \infty}} \, \mathbb{P}^r\big(|\widehat{Z}^r(0)| > M\big) = 0.
$$

(A.4) (Assumption 7.1 in Williams [26]).

The matrix $R = (I + G)^{-1}$ is completely-S, where

$$
G \equiv CMQ\widetilde{P}\Lambda = \lim_{r \to \infty} CM^r Q^r \widetilde{P}^r \Lambda^r
$$

and $M^r \equiv diag(m_k^r, k \in \mathbb{K}), \Lambda^r \equiv diag(\lambda_k^r, k \in \mathbb{K}), r \ge 1$, and $M = \lim_{r \to \infty} M^r$, etc. (Of course, it is implicitly assumed that $I + G$ is invertible. For the definition of completely-S condition, see Definition 6.2 in Williams [26], for example).

Condition (A.2) corresponds to the initial condition on strong state-space collapse for a more general multiclass FCFS queueing network *without* abandonment. (Cf. Bramson [3], Williams [26]). While condition (A.3) is implied by (A.1) and (A.2) for such network *without* abandonment, we have to assume it in our network *with* abandonment. As established in [26], assumption (A.4) is satisfied under the asymptotically Kelly-type condition, i.e., $m_k = m_l$ if $s(k) = s(l)$. The completely-S condition on R in (A.4) is a necessary and sufficient condition for the existence and uniqueness (in law) of a semimartingale reflecting Brownian motion (SRBM) with the reflection matrix R and the data on the covariance, drift and initial measure of the Brownian motion in the SRBM. (Cf. Definition 6.1 in [26] and the references in its comment).

The following theorem is the main result in this paper. It is on the weak convergence for the sequence of scaled performance measure processes

$$\{(\widehat{W}^r(\cdot), \widehat{Y}^r(\cdot), \widehat{Z}^r(\cdot))\}_{r\geq 1}.$$

In the statement of the theorem, we use the following symbol:

$$\Gamma \equiv RC\Big\{\Lambda\Gamma_V + MQ\Big(\Gamma_E + \sum_{k=1}^{K} \lambda_k \Gamma_\Phi^k\Big)\tilde{Q}M\Big\}\tilde{C}\tilde{R}, \tag{91}$$

According to (54), we see that Γ is strictly positive definite. We also let

$$H^*(w) \equiv CMQ\Lambda \cdot H(w), \qquad w \in \mathcal{R}^J, \tag{92}$$

with $H(w) \equiv (H_k(w_{s(k)}), k \in \mathbb{K})$, $H_k(\cdot), k \in \mathbb{K}$, in (57).

Theorem 4.1. (*Diffusion approximation for a multiclass feedforward queueing network with abandonment under the FCFS discipline*).

Under the main assumptions $(A.1), (A.2)$ *and* $(A.3)$*, and also the conditions imposed on the primitive variables and processes so far, we have the weak convergence*

$$(\widehat{W}^r(\cdot), \widehat{Y}^r(\cdot), \widehat{Z}^r(\cdot)) \Longrightarrow (W^*(\cdot), Y^*(\cdot), Z^*(\cdot)) \quad \textit{in } \mathbb{D}([0,\infty), \mathcal{R}^{2J+K}) \tag{93}$$

as $r \to \infty$*, where* $W^*(\cdot)$ *is the unique solution to the following* J*-dimensional reflected stochastic differential equation:*

$$W^*(t) = X^*(t) + RY^*(t), \tag{94}$$

$$X^*(t) = W^*(0) + \sqrt{\Gamma}B^*(t) + \vartheta^* t - \int_0^t H^*(W^*(u))du, \tag{95}$$

where $B^*(\cdot)$ *is a* J*-dimensional standard Brownian motion,* $\vartheta^* \equiv R\vartheta$ *and* $\nu(\cdot) = \mathbb{P}(W^*(0) \in \cdot)$*. Furthermore,*

$$Z^*(t) = \Lambda\tilde{C}W^*(t), \quad t \geq 0.$$

5 Proof of Theorem 4.1; propositions and lemmas

This section is devoted to the proof of the diffusion approximation theorem stated in the last section. We begin with the following stochastic boundedness of scaled queue length and workload in a multiclass feedforward queueing network with abandonment under any work-conserving service discipline.

5.1 Stochastic boundedness of diffusion-scaled queue length and workload In this subsection we present two propositions on the stochastic boundedness of diffusion-scaled queue length and workload in our multiclass feedforward queueing network with abandonment. Each of them plays a key role in the proof of our main theorem, specifically in proving the C-tightness of diffusion-scaled abandonment-count process and deriving state-space collapse in the network.

Proposition 5.1.

For a sequence of multiclass feedforward queueing networks with abandonments, $\{\mathfrak{X}^r\}_{r\geq 1}$, satisfying the assumptions stated so far, the sequence $\{\widehat{Z}^r(\cdot)\}_{r\geq 1}$ is stochastically bounded, i.e.,

$$\lim_{M\to\infty} \overline{\lim_{r\to\infty}} \mathbb{P}^r\big(\sup_{0\leq t\leq T} |\widehat{Z}^r(t)| > M\big) = 0$$

for each $T > 0$.

Proof.

Let

$$\widehat{f}^r(t) \equiv CM^rQ^r\widehat{Z}^r(t) = (\widehat{f}^r_j(t), j \in \mathbb{J})$$

where

$$\widehat{f}^r_j(t) = \widehat{f}^r_{j1}(t) + \widehat{f}^r_{j2}(t)$$

with

$$\widehat{f}^r_{j1}(t) = \sum_{k\in C(j)} m^r_k \sum_{l\in C(j)} Q^r_{kl}\widehat{Z}^r_l(t),$$
$$\widehat{f}^r_{j2}(t) = \sum_{k\in C(j)} m^r_k \sum_{l\in C(1)\cup\cdots\cup C(j-1)} Q^r_{kl}\widehat{Z}^r_l(t)$$

for each $j \in \mathbb{J}$, where we have used the feedforward class-routing condition (8). (We set $\widehat{f}^r_{12}(\cdot) \equiv 0$).

From

$$Z^r(t) = Z^r(0) + E^r(t) + \sum_{l=1}^{K} \Phi^{l,r}(S^r_l(T^r_l(t))) - S^r(T^r(t)) - I^r(t)$$

with $S^r(T^r(t)) \equiv (S^r_k(T^r_k(t)), k \in \mathbb{K})$, we have the following scaled identity in vector form:

$$\begin{aligned}\widehat{Z}^r(t) =& \widehat{Z}^r(0) + \widehat{E}^r(t) + \alpha^r rt + \sum_{l\in\mathbb{K}} \widehat{\Phi}^{l,r}(\overline{S}^r_l(\overline{T}^r_l(t))) - (\mathrm{I} - \widetilde{P^r})\widehat{S}^r(\overline{T}^r(t)) \\ &- (\mathrm{I} - \widetilde{P^r})\frac{(\mu^rT^r)(r^2t)}{r} - \widehat{I}^r(t) \qquad (96)\end{aligned}$$

with the diffusion and fluid scalings given above. Multiplying (96) by CM^rQ^r from the left, we have

$$\begin{aligned}\widehat{f}^r(t) = \widehat{f}^r(0) + CM^rQ^r\Big\{\widehat{E}^r(t) + \sum_{l\in\mathbb{K}} \widehat{\Phi}^{l,r}(\overline{S}^r_l(\overline{T}^r_l(t)))\Big\} \\ - CM^r\widehat{S}^r(\overline{T}^r(t)) - CM^rQ^r\widehat{I}^r(t) + r(\rho^r - e)t + \widehat{Y}^r(t). \qquad (97)\end{aligned}$$

Since

$$\int_0^\infty \widehat{f}^r_1(s)d\widehat{Y}^r_1(s) = \int_0^\infty \widehat{f}^r_{11}(s)d\widehat{Y}^r_1(s) = 0, \qquad (98)$$

from (97) we have

$$\widehat{f}_1^r(t) = \varphi\Big(\mathcal{X}_1^r(\cdot) - \sum_{k\in C(1)} m_k^r \sum_{l\in C(1)} Q_{kl}^r \widehat{I}_l^r(\cdot)\Big)(t) \tag{99}$$

where φ is the one-dimensional reflection map, i.e.,

$$\varphi(x(\cdot))(t) = x(t) + \sup_{0\le s\le t}\big(-x(s)\big)^+, \quad x\in\mathbb{D}([0,\infty),\mathcal{R}^1),\ t\ge 0, \tag{100}$$

and

$$\begin{aligned}\mathcal{X}_1^r(t) \equiv \widehat{f}_1^r(0) &+ \sum_{k\in C(1)} m_k^r \sum_{l\in C(1)} Q_{kl}^r\big\{\widehat{E}_l^r(t) + \sum_{p\in\mathbb{K}} \widehat{\Phi}_l^{p,r}(\overline{S}_p^r(\overline{T}_p^r(t)))\big\}\\ &- \sum_{k\in C(1)} m_k^r \widehat{S}_k^r(\overline{T}_k^r(t)) + r(\rho_1^r - 1)t, \qquad t\ge 0.\end{aligned} \tag{101}$$

Since each component in $\widehat{I}^r(\cdot)$ is nondecreasing, we have

$$\begin{aligned}&\widehat{f}_1^r(t)\\ &= \mathcal{X}_1^r(t) - \sum_{k\in C(1)} m_k^r \sum_{l\in C(1)} Q_{kl}^r \widehat{I}_l^r(t) + \sup_{0\le s\le t}\Big(-\mathcal{X}_1^r(s) + \sum_{k\in C(1)} m_k^r \sum_{l\in C(1)} Q_{kl}^r \widehat{I}_l^r(s)\Big)^+\\ &\le \mathcal{X}_1^r(t) + \sup_{0\le s\le t}(-\mathcal{X}_1^r(s))^+\\ &= \varphi(\mathcal{X}_1^r(\cdot))(t).\end{aligned} \tag{102}$$

Thus, according to the Lipschitz continuity of the map φ, (A.3), the heavy-traffic condition (56), and the convergences (86)-(90), (47) and (50), we obtain

$$\lim_{M\to\infty} \overline{\lim_{r\to\infty}}\, \mathbb{P}^r\big(\sup_{0\le t\le T} \widehat{Z}_k^r(t) > M\big) = 0 \tag{103}$$

for each $k\in C(1)$ and $T>0$.

Suppose that (103) holds for each $k\in C(1)\cup\cdots\cup C(j-1)$ with some $2\le j\le J$. Then, since

$$\int_0^\infty \widehat{f}_{j1}^r(s)\, d\widehat{Y}_j^r(s) = 0,$$

we have

$$\widehat{f}_{j1}^r(t) = \varphi\Big(-\widehat{f}_{j2}^r(\cdot) + \mathcal{X}_j^r(\cdot) - \sum_{k\in C(j)} m_k^r \sum_{l\in C(1)\cup\cdots\cup C(j)} Q_{kl}^r \widehat{I}_l^r(\cdot)\Big)(t) \tag{104}$$

where

$$\begin{aligned}\mathcal{X}_j^r(t) \equiv \widehat{f}_j^r(0) &+ \sum_{k\in C(j)} m_k^r \sum_{l\in C(1)\cup\cdots\cup C(j)} Q_{kl}^r\big\{\widehat{E}_l^r(t) + \sum_{p\in\mathbb{K}} \widehat{\Phi}_l^{p,r}(\overline{S}_p^r(\overline{T}_p^r(t)))\big\}\\ &- \sum_{k\in C(j)} m_k^r \widehat{S}_k^r(\overline{T}_k^r(t)) + r(\rho_j^r - 1)t, \qquad t\ge 0.\end{aligned} \tag{105}$$

Thus, similar to the above reasoning, the inequality

$$\widehat{f}_{j1}^r(t) \le \varphi(-\widehat{f}_{j2}^r(\cdot) + \mathcal{X}_j^r(\cdot))(t) \tag{106}$$

holds so that

$$\lim_{M\to\infty}\overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(\sup_{0\le t\le T}\widehat{Z}^r_k(t)>M\big)=0, \tag{107}$$

is derived for each $k\in C(j)$ and $T>0$, using (103) for each $k\in C(1)\cup\cdots C(j-1)$. Consequently we have the desired result inductively.

□

Using Proposition 5.1, we also have the corresponding result for diffusion-scaled workload in the next proposition.

Proposition 5.2.

For $\{\mathfrak{X}^r\}_{r\ge1}$ in Proposition 5.1, the sequence $\{\widehat{W}^r(\cdot)\}_{r\ge1}$ is stochastically bounded, i.e.,

$$\lim_{M\to\infty}\overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(\sup_{0\le t\le T}|\widehat{W}^r(t)|>M\big)=0$$

for each $T>0$.

Proof.

¿From (21), (23), (67) and (68), we have

$$\widehat{W}^r(t)=\widehat{W}^r(0)+C\widehat{\mathcal{V}}^{s,r}(\overline{A}^{+,r}(t))+CM^r(\widehat{A}^r(t)-\widehat{A}^{-,r}(t))+r(\rho^r-e)t+\widehat{Y}^r(t) \tag{108}$$

with $\widehat{\mathcal{V}}^{s,r}(\cdot)$ in (70) and $\overline{A}^{+,r}(t)$ in (81).

¿From (65), (69), (71), (73), (74), (77) and (82), we see that

$$\begin{aligned}\widehat{A}^r(t)&=\widehat{E}^r(t)+\sum_{l\in\mathbb{K}}\widehat{\Phi}^{l,r}(\overline{D}^r_l(t))+\widetilde{P^r}\widehat{D}^r(t)\\&=\widehat{E}^r(t)+\sum_{l\in\mathbb{K}}\widehat{\Phi}^{l,r}(\overline{D}^r_l(t))+\widetilde{P^r}\big(\widehat{Z}^r(0)-\widehat{Z}^r(t)-\widehat{I}^r(t)+\widehat{A}^r(t)\big).\end{aligned}$$

Solving it for $\widehat{A}^r(t)$, we have

$$\widehat{A}^r(t)=Q^r\{\widehat{E}^r(t)+\sum_{l\in\mathbb{K}}\widehat{\Phi}^{l,r}(\overline{D}^r_l(t))+\widetilde{P^r}\big(\widehat{Z}^r(0)-\widehat{Z}^r(t)-\widehat{I}^r(t)\big)\}. \tag{109}$$

Substituting (109) into (108), we have

$$\begin{aligned}\widehat{W}^r(t)=\widehat{W}^r(0)&+C\widehat{\mathcal{V}}^{s,r}(\overline{A}^{+,r}(t))\\&+CM^rQ^r\big\{\widehat{E}^r(t)+\sum_{l\in\mathbb{K}}\widehat{\Phi}^{l,r}(\overline{D}^r_l(t))+\widetilde{P^r}(\widehat{Z}^r(0)-\widehat{Z}^r(t))\big\}\\&+r(\rho^r-e)t-CM^r\widehat{A}^{-,r}(t)-CM^rQ^r\widetilde{P^r}\widehat{I}^r(t)+\widehat{Y}^r(t).\end{aligned}$$

Let

$$\begin{aligned}\mathcal{Y}^r(t)\equiv\widehat{W}^r(0)&+C\widehat{\mathcal{V}}^{s,r}(\overline{A}^{+,r}(t))\\&+CM^rQ^r\big\{\widehat{E}^r(t)+\sum_{l\in\mathbb{K}}\widehat{\Phi}^{l,r}(\overline{D}^r_l(t))+\widetilde{P^r}(\widehat{Z}^r(0)-\widehat{Z}^r(t))\big\}\\&+r(\rho^r-e)t.\end{aligned}$$

Then, since

$$\int_0^\infty \widehat{W}_j^r(s)d\widehat{Y}_j^r(s) = 0, \qquad \forall j \in \mathbb{J}, \tag{110}$$

we have that for each $j \in \mathbb{J}$,

$$\begin{aligned}
\widehat{W}_j^r(t) &= \varphi\Big(\mathcal{Y}_j^r(\cdot) - \sum_{k\in C(j)} m_k^r \widehat{A}_k^{-,r}(\cdot) - \sum_{k\in C(j)} m_k^r \sum_{l\in\mathbb{K}} (Q^r\widetilde{P^r})_{kl}\widehat{I}_l^r(\cdot)\Big)(t) \\
&= \mathcal{Y}_j^r(t) - \sum_{k\in C(j)} m_k^r \widehat{A}_k^{-,r}(t) - \sum_{k\in C(j)} m_k^r \sum_{l\in\mathbb{K}} (Q^r\widetilde{P^r})_{kl}\widehat{G}_l^r(t) \\
&\quad + \sup_{0\le s\le t}\Big(-\mathcal{Y}_j^r(s) + \sum_{k\in C(j)} m_k^r \widehat{A}_k^{-,r}(s) + \sum_{k\in C(j)} m_k^r \sum_{l\in\mathbb{K}} (Q^r\widetilde{P^r})_{kl}\widehat{I}_l^r(s)\Big)^+ \\
&\le \mathcal{Y}_j^r(t) + \sup_{0\le s\le t}\Big(-\mathcal{Y}_j^r(s)\Big)^+ \\
&= \varphi\Big(\mathcal{Y}_j^r(\cdot)\Big)(t)
\end{aligned}$$

with $\varphi(\cdot)$ in (100), where the inequality follows from the non-decreasing property of each component in $\widehat{A}^{-,r}(\cdot)$ and $\widehat{I}^r(\cdot)$. Thus, using the Lipschitz continuity of φ, Proposition 5.1, and (A.1), we have the desired result.

□

Remark 5.1.

The conclusions of Propositions 5.1 and 5.2 are valid under any work-conserving (or non-idling) service discipline, which is embodied as (98) and (110). We note that if the stochastic boundedness condition on scaled queue length is verified for a more general multi-class queueing network, then that condition on scaled workload does hold for such network, which will be seen by mimicking the proof of Proposition 5.2.

5.2 C-tightness of diffusion-scaled abandonment-count process In this subsection, we show the C-tightness of the sequence of scaled abandonment-count processes $\{\widehat{I}_k^r(\cdot)\}_{r\ge1}, k \in \mathbb{K}$, which will be seen to follow from that of the sequence $\{\widehat{N}_k^r(\cdot)\}_{r\ge1}, k \in \mathbb{K}$, as follows.

Proposition 5.3.

For each $k \in \mathbb{K}$, the sequence $\{\widehat{I}_k^r(\cdot)\}_r$ is C-tight in $\mathbb{D}([0,\infty), \mathcal{R}^1)$.

Proof.

Similar to (29), let

$$\zeta_j^r(t) \equiv \inf\{s \ge 0 : s + W_j^r(s) > t\}, \quad t \ge 0, j \in \mathbb{J},$$

and $\overline{\zeta}_j^r(t) \equiv r^{-2}\zeta_j^r(r^2t)$. Then we have that for each $T > 0$ and $j \in \mathbb{J}$,

$$\sup_{0\le t\le T} |\overline{\zeta}_j^r(t) - t| \longrightarrow 0 \quad \text{in pr.} \tag{111}$$

as $r \to \infty$, which follows from the inequalities

$$\zeta_j^r(t) + W_j^r(\zeta_j^r(t)) \ge t \quad \text{and} \quad \zeta_j^r(t) \le t,$$

and Proposition 5.2.

¿From (28), the inequality

$$\widehat{N}_k^r(\overline{\zeta}_{s(k)}^r(t) - \frac{1}{r^3}) \le \widehat{I}_k^r(t) \le \widehat{N}_k^r(t) \tag{112}$$

follows, so that according to (111), the proof of C-tightness for $\{\widehat{I}_k^r(\cdot)\}_r, k \in \mathbb{K}$, is reduced to that for $\{\widehat{N}_k^r(\cdot)\}_r, k \in \mathbb{K}$, which is done in the next lemma.

□

Lemma 5.1.

For each $k \in \mathbb{K}$, the sequence $\{\widehat{N}_k^r(\cdot)\}_r$ is C-tight in $\mathbb{D}([0,\infty), \mathcal{R}^1)$.

Proof.

Assumptions (A.1) and (A.2) yield that for each $k \in \mathbb{K}$,

$$\begin{aligned}\widehat{Z}_k^{+,r}(0) &= \frac{1}{r} D_k^r(W_{s(k)}^r(0)) \\ &\Longrightarrow \lambda_k W_{s(k)}^*(0)\end{aligned}$$

as r goes to infinity, so that the tightness of $\{\widehat{Z}_k^{-,r}(0)\}_r$ follows from (A.3). Thus we are left to show the C-tightness of $\{\widehat{A}_k^{-,r}(\cdot)\}_r$, because of the identity

$$\widehat{N}_k^r(t) = \widehat{Z}_k^{-,r}(0) + \widehat{A}_k^{-,r}(t), t \ge 0.$$

¿From (27) and (75), it follows that

$$\begin{aligned}\widehat{A}_k^{-,r}(t) &= \frac{1}{r} \sum_{i=1}^{A_k^r(r^2t)} \left(1_{\{\gamma_k^{s,r}(i) \le w_k^{s,r}(i)\}} - F_k^{\gamma,r}(w_k^{s,r}(i))\right) + \frac{1}{r} \sum_{i=1}^{A_k^r(r^2t)} F_k^{\gamma,r}(w_k^{s,r}(i)) \\ &= \widehat{\mathcal{M}}_k^{\gamma,r}(\overline{A}_k^r(t)) + \widehat{\mathcal{C}}_k^r(\overline{A}_k^r(t)) \qquad (113)\end{aligned}$$

where

$$\widehat{\mathcal{M}}_k^{\gamma,r}(t) \equiv \frac{1}{r} \sum_{i=1}^{\lfloor r^2 t \rfloor} \left(1_{\{\gamma_k^{s,r}(i) \le w_k^{s,r}(i)\}} - F_k^{\gamma,r}(w_k^{s,r}(i))\right), \tag{114}$$

$$\widehat{\mathcal{C}}_k^r(t) \equiv \frac{1}{r} \sum_{i=1}^{\lfloor r^2 t \rfloor} F_k^{\gamma,r}(w_k^{s,r}(i)). \tag{115}$$

Observe that $\widehat{\mathcal{M}}_k^{\gamma,r}(\cdot)$ is a purely-discontinuous martingale since $w_k^{s,r}(i)$ is $\mathcal{G}_k^{s,r}(i)$-measurable and $\gamma_k^{s,r}(i)$ is independent of $\mathcal{G}_k^{s,r}(i)$ for each $i = 1, 2, \cdots$, where $\mathcal{G}_k^{s,r}(i)$ is given in the form (12). Then its optional quadratic variation process $\left[\widehat{\mathcal{M}}_k^{\gamma,r}\right](\cdot)$ is given by

$$\begin{aligned}[\widehat{\mathcal{M}}_k^{\gamma,r}](t) &= \sum_{0<s\le t} |\Delta \widehat{\mathcal{M}}_k^{\gamma,r}(s)|^2 \\ &= \frac{1}{r^2} \sum_{i=1}^{\lfloor r^2 t \rfloor} \left(1_{\{\gamma_k^{s,r}(i) \le w_k^{s,r}(i)\}} - F_k^{\gamma,r}(w_k^{s,r}(i))\right)^2. \qquad (116)\end{aligned}$$

(Cf. (1.8.3) of Liptser and Shiryayev [20]).

We now show that

$$\widehat{\mathcal{M}}_k^{\gamma,r}(\cdot) \Longrightarrow 0 \quad \text{in} \quad \mathbb{D}([0,\infty),\mathcal{R}^1), \tag{117}$$

as $r \to \infty$ in a similar way to the proof of Lemma 4.3 in Dai and He [7] as follows. Observe that for each $t \geq 0$,

$$\begin{aligned}\mathbb{E}^r([\widehat{\mathcal{M}}_k^{\gamma,r}](t)) &= \frac{1}{r^2}\sum_{i=1}^{\lfloor r^2 t\rfloor}\mathbb{E}^r(F_k^{\gamma,r}(w_k^{s,r}(i)) - F_k^{\gamma,r}(w_k^{s,r}(i))^2) \\ &\leq t\,\mathbb{E}^r(\sup_{1\leq i\leq\lfloor r^2t\rfloor} F_k^{\gamma,r}(w_k^{s,r}(i)))\end{aligned}$$

where the equality follows from the $\mathcal{G}_k^{s,r}(i)$-measurability of $w_k^{s,r}(i)$ and the independence of $\gamma_k^{s,r}(i)$ and $\mathcal{G}_k^{s,r}(i)$.
Since $A_k^r(s) \geq E_k^r(s)$ for each $s \geq 0$ and $\overline{E}_k^r(\cdot) \Longrightarrow \alpha_k\iota(\cdot)$ as $r \to \infty$, we can take an appropriate constant $t^* > 0$ such that

$$\lim_{r\to\infty}\mathbb{P}^r(A_k^r(r^2t^*) \leq \lfloor r^2t\rfloor) = 0. \tag{118}$$

Thus we have

$$\begin{aligned}&\overline{\lim_{r\to\infty}}\,\mathbb{E}^r\big[\sup_{1\leq i\leq\lfloor r^2t\rfloor} F_k^{\gamma,r}(w_k^{s,r}(i))\big] \\ &\leq \overline{\lim_{r\to\infty}}\,\mathbb{E}^r\big[\sup_{1\leq i\leq\lfloor r^2t\rfloor} F_k^{\gamma,r}(w_k^{s,r}(i)); A_k^r(r^2t^*) > \lfloor r^2t\rfloor\big] \\ &\leq \overline{\lim_{r\to\infty}}\,\mathbb{E}^r\big[\sup_{1\leq i\leq A_k^r(r^2t^*)} F_k^{\gamma,r}(w_k^{s,r}(i))\big] \\ &\leq \overline{\lim_{r\to\infty}}\,\mathbb{E}^r\big[F_k^{\gamma,r}(\sup_{0\leq u\leq t^*} W_{s(k)}^r(r^2u))\big] \\ &\leq \overline{\lim_{r\to\infty}}\,\mathbb{E}^r\big[F_k^{\gamma,r}(\sup_{0\leq u\leq t^*} W_{s(k)}^r(r^2u)); \sup_{0\leq u\leq t^*}\widehat{W}_{s(k)}^r(u) \leq M\big] \\ &\qquad + \overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(\sup_{0\leq u\leq t^*}\widehat{W}_{s(k)}^r(u) > M\big).\end{aligned} \tag{119}$$

According to Proposition 5.2, $\lim_{M\to\infty}$ (the second term in (119)) $= 0$, while the first term in (119) is majorized by

$$\begin{aligned}\lim_{r\to\infty} F_k^{\gamma,r}(rM) &= \lim_{r\to\infty}\frac{1}{r}\cdot rF_k^{\gamma,r}(rM) \\ &= 0\end{aligned}$$

for each fixed $M > 0$, because of (57). Therefore we have that for each $t \geq 0$,

$$\lim_{r\to\infty}\mathbb{E}^r([\widehat{\mathcal{M}}_k^{\gamma,r}](t)) = 0$$

so that the convergence (117) is established, according to Theorem 7.1.4 in Ethier and Kurtz [11].

Let $B_k^r(t) \equiv E_k^r(t) + \sum_{l=1}^K S_l^r(t), t \geq 0$. Then, since $A_k^r(t) \leq B_k^r(t)$ for each $t \geq 0$ and

$$\overline{B}_k^r(\cdot) \equiv r^{-2}B_k^r(r^2\cdot) \Longrightarrow \alpha_k\iota(\cdot) + \sum_{l=1}^K \mu_l\iota(\cdot) \tag{120}$$

as $r \to \infty$, we have

$$\widehat{\mathcal{M}}_k^{\gamma,r}(\overline{A}_k^r(\cdot)) \Longrightarrow 0 \quad \text{in} \quad \mathbb{D}([0,\infty), \mathcal{R}^1), \tag{121}$$

as $r \to \infty$.

Thus the proof of the C-tightness of $\{\widehat{A}_k^{-,r}(\cdot)\}_r$ is reduced to that of the C-tightness of $\{\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(\cdot))\}_r$, and so it is enough to show the following two conditions:

$$\lim_{M\to\infty} \overline{\lim_{r\to\infty}} \mathbb{P}^r(\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(T)) > M) = 0 \tag{122}$$

for each $T > 0$, and

$$\lim_{\delta\to 0} \overline{\lim_{r\to\infty}} \mathbb{P}^r(w_T(\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(\cdot)), \delta) > \varepsilon) = 0 \tag{123}$$

for each $\varepsilon > 0$ and $T > 0$, where

$$w_T(x(\cdot), \delta) \equiv \sup_{\substack{0\le s,t\le T \\ |s-t|\le\delta}} |x(s) - x(t)|, \quad x(\cdot) \in \mathbb{D}([0,\infty), \mathcal{R}^d), \quad \delta > 0, T > 0, d \in \mathbb{N}. \tag{124}$$

(Cf. Proposition 6.3.26 in Jacod and Shiryaev [14]).

Observe that

$$\begin{aligned} \mathbb{P}^r(\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(T)) > M) \le & \mathbb{P}^r(\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(T)) > M, \sup_{0\le t\le T} \widehat{W}^r(t) \le L) \\ & + \mathbb{P}^r(\sup_{0\le t\le T} \widehat{W}^r(t) > L). \end{aligned} \tag{125}$$

Then, $\lim_{L\to\infty} \overline{\lim}_{r\to\infty}$(the second term in (125))$= 0$ according to Proposition 5.2, and the first term in (125) is majorized by

$$\mathbb{P}^r(\overline{A}_k^r(T) \cdot rF_k^{\gamma,r}(rL) > M) \le \mathbb{P}^r(\overline{B}_k^r(T) \cdot rF_k^{\gamma,r}(rL) > M)$$

so that $\lim_{M\to\infty} \overline{\lim}_{r\to\infty}$(the first term in (125))$= 0$ for each fixed $L > 0$, according to (57) and (120). Thus we have (122).

Furthermore, observe that

$$\begin{aligned} & \mathbb{P}^r(w_T(\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(\cdot)), \delta) > \varepsilon) \\ & \le \mathbb{P}^r(\sup_{\substack{0\le s,t\le T \\ |s-t|\le\delta}} |\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(s)) - \widehat{\mathcal{C}}_k^r(\overline{A}_k^r(t))| > \varepsilon, \sup_{0\le t\le T} \widehat{W}^r(t) \le L) \\ & \quad + \mathbb{P}^r(\sup_{0\le t\le T} \widehat{W}^r(t) > L). \end{aligned} \tag{126}$$

Then, the same as above, $\lim_{L\to\infty} \overline{\lim}_{r\to\infty}$(the second term in (126))$= 0$, and the first term in (126) is less than or equal to

$$\mathbb{P}^r\Big(rF_k^{\gamma,r}(rL) \times w_T(\overline{A}_k^r(\cdot), \delta) > \varepsilon\Big).$$

Therefore, noting

$$w_T(\overline{A}_k^r(\cdot), \delta) \le w_T(\overline{E}_k^r(\cdot), \delta) + \sum_{l\in\mathbb{K}} w_T(\overline{S}_l^r(\cdot), \delta),$$

(123) is seen to be satisfied, according to (57).

Consequently we have the C-tightness of $\{\widehat{A}_k^{-,r}(\cdot)\}_r$ and so the conclusion of the lemma has been proved.

□

5.3 State-space collapse in multiclass feedforward queueing networks with abandonments under FCFS service disciplines In this subsection, under the assumption (A.2), we prove the following proposition on multiplicative strong state-space collapse and state-space collapse in a multiclass feedforward queueing network with abandonment under the FCFS service discipline.

Proposition 5.4. (*Multiplicative strong state-space collapse and state-space collapse*).

Suppose that in addition to the assumptions in Sect. 3*, conditions* $(A.1), (A.2)$ *and* $(A.3)$ *hold. Then we have the following convergences*:
For each $k \in \mathbb{K}$ *and* $T > 0$*,*

$$\frac{\sup_{0\le t\le T}\sup_{0\le s\le \widehat{W}^r_{s(k)}(t)} \mid r^{-1}D^r_k(r^2t+rs) - r^{-1}D^r_k(r^2t) - \lambda^r_k s \mid}{\sup_{0\le t\le T}\widehat{W}^r_{s(k)}(t)\vee 1} \longrightarrow 0 \quad \textit{in pr.} \tag{127}$$

as $r \to \infty$*, and also,*

$$\sup_{0\le t\le T}\left|\widehat{Z}^r_k(t) - \lambda^r_k \widehat{W}^r_{s(k)}(t)\right| \longrightarrow 0 \quad \textit{in pr.} \tag{128}$$

as $r \to \infty$*.*

To demonstrate the proposition, we need to modify slightly the proof of Theorem 1 in Bramson [3] by incorporating the customer abandonment to it. Specifically, to the statement of Proposition 5.1 in [3], we have to add the identity on the weak law of large numbers for $I^{r,m}(\cdot)$ that is defined in the same way as in [3] as follows.

For the performance measure process $\mathfrak{X}^r(\cdot), r \ge 1$, in (14), let

$$\mathfrak{X}^{r,m}(t) \equiv \Big\{\frac{1}{x_{r,0}}\mathfrak{X}^r(x_{r,0}t)\Big\} \circ \theta_{rm} \tag{129}$$

for $m = 0, 1, 2, \ldots$, where $x_{r,0} \equiv |W^r(0)| \vee |Z^r(0)| \vee r$ and $\{\theta_t, t \ge 0\}$ is the shift transformation associated with Markov description process $\Xi^r(\cdot)$. For example, using Proposition 2.1, we have

$$Z^{r,m}(t) = \frac{1}{x_{r,m}}Z^r(x_{r,m}t + rm),$$
$$I^{r,m}(t) = \frac{1}{x_{r,m}}(I^r(x_{r,m}t + rm) - I^r(rm)),$$

where $x_{r,m} \equiv x_{r,0} \circ \theta_{rm} = |W^r(rm)| \vee |Z^r(rm)| \vee r$ for $m = 0, 1, 2, \ldots$.

Proposition 5.5. (*Weak law of large numbers for* $I^{r,m}(\cdot)$).

For each $\varepsilon > 0, T > 0, L > 0$ *and* $k \in \mathbb{K}$*,*

$$\lim_{r\to\infty}\mathbb{P}^r\Big(\max_{0\le m<rT} I^{r,m}_k(L) > \varepsilon\Big) = 0.$$

Since $I^r_k(t) \le N^r_k(t)$ for each $t \ge 0$, the above proposition is a consequence of the following proposition.

Proposition 5.6.

For each $\varepsilon > 0, T > 0, L > 0$ and $k \in \mathbb{K}$,

$$\lim_{r\to\infty} \mathbb{P}^r\big(\max_{0\le m<rT} N_k^{r,m}(L) > \varepsilon\big) = 0,$$

where $N_k^{r,m}(\cdot)$ is defined as in (129).

Before giving the proof of Proposition 5.6, we define the following variables which correspond to (5.25) in Bramson [3]:

$$u_k^{max,T,r} \equiv \max\{u_k^r(i) : \mathcal{U}_k^r(i-1) \le r^2T, i = 1,2,\ldots\}, \tag{130}$$

$$\begin{aligned} v_l^{max,T,r} &\equiv \max\{v_l^{o,r}(i) : \mathcal{V}_l^{o,r}(i-1) \le r^2T, i = 1,2,\ldots, Z_l^{+,r}(0)\} \\ &\quad \vee \max\{v_l^{s,r}(i) : \mathcal{V}_l^{o,r}(Z_l^{+,r}(0)) + \mathcal{V}_l^{s,r}(i-1) \le r^2T, i = 1,2,\ldots\} \end{aligned} \tag{131}$$

with $\max\phi \equiv 0$, for each $k \in \mathbb{A}$, $l \in \mathbb{K}$ and $T > 0$. Then we have the inequalities

$$u_k^r(1)\circ\theta_{rm} \le u_k^{max,T,r}, \tag{132}$$

$$v_l^{o,r}(1)\circ\theta_{rm} \le v_l^{max,T,r} \tag{133}$$

for each $m = 0, 1, \cdots, \lceil rT\rceil - 1$, $T > 0$, $k \in \mathbb{A}$ and $l \in \mathbb{K}$. Indeed, for each m,

$$\begin{aligned} \mathcal{U}_k^r(E_k^r(rm)) \le rm < \mathcal{U}_k^r(E_k^r(rm)+1), \\ u_k^r(1)\circ\theta_{rm} &= \mathcal{R}_k^{u,r}(0)\circ\theta_{rm} \\ &= \mathcal{R}_k^{u,r}(rm) \\ &= \mathcal{U}_k^r(E_k^r(rm)+1) - rm, \end{aligned}$$

from which the inequality (132) follows.

The next lemma corresponds to Lemma 5.1 in [3].

Lemma 5.2.

For each $k \in \mathbb{A}$, $l \in \mathbb{K}$ and $T > 0$,

$$\frac{1}{r}u_k^{max,T,r} \longrightarrow 0 \qquad \text{in pr.,} \tag{134}$$

$$\frac{1}{r}v_l^{max,T,r} \longrightarrow 0 \qquad \text{in pr.,} \tag{135}$$

as r goes to infinity.

Proof.

We have only to prove the latter convergence (135), because the derivation of the former (134) is the same as in Lemma 5.1 in [3]. First we observe that for each δ and B_1 with $0 < \delta < B_1$,

$$\begin{aligned} &\mathbb{P}^r\big(\frac{1}{r}v_l^{max,T,r} > \varepsilon\big) \\ &\le \mathbb{P}^r\Big(\frac{1}{r}v_l^{max,T,r} > \varepsilon, \mathcal{V}_l^{o,r}(Z_l^{+,r}(0)) + \mathcal{V}_l^{s,r}(\lfloor r^2B_1\rfloor) > r^2T, Z_l^{+,r}(0) < r^2\delta\Big) \\ &\quad + \mathbb{P}^r\Big(\mathcal{V}_l^{o,r}(Z_l^{+,r}(0)) + \mathcal{V}_l^{s,r}(\lfloor r^2B_1\rfloor) \le r^2T, Z_l^{+,r}(0) < r^2\delta\Big) \\ &\quad + 2\mathbb{P}^r\big(Z_l^{+,r}(0) \ge r^2\delta\big). \end{aligned} \tag{136}$$

The second term in (136) tends to zero as r goes to infinity, since

$$\mathbb{P}^r\big(\mathcal{V}_l^{o,r}(\lfloor r^2\delta\rfloor)+\mathcal{V}_l^{s,r}(\lfloor r^2B_1\rfloor)\le r^2T\big)\longrightarrow 0 \tag{137}$$

as r tends to infinity for an appropriate constant $B_1>0$, according to the weak law of large numbers. We also have

$$\mathbb{P}^r\big(Z_l^{+,r}(0)\ge r^2\delta\big)\longrightarrow 0 \tag{138}$$

as r tends to infinity, according to assumption (A.3).

Furthermore, the first term in (136) is majorized by

$$\begin{aligned}&\mathbb{P}^r\Big(\frac{1}{r}\times\max_{1\le i\le\lfloor r^2\delta\rfloor}v_l^{o,r}(i)\vee\max_{1\le i\le\lfloor r^2B_1\rfloor}v_l^{s,r}(i)>\varepsilon\Big)\\&\le\mathbb{P}^r\Big(\frac{1}{r}v_l^{o,r}(1)>\varepsilon\Big)+(\lfloor r^2\delta\rfloor+\lfloor r^2B_1\rfloor)\cdot\frac{1}{(r\varepsilon)^2}\eta(r\varepsilon)\\&\longrightarrow 0\end{aligned} \tag{139}$$

as r goes to infinity, where

$$\eta(R)\equiv\sup_r\mathbb{E}^r\big[v_l^{o,r}(2)^2;v_l^{o,r}(2)>R\big],\quad R>0,$$

and the convergence to zero follows from assumptions (61) and (59). So the proof is completed.

□

Proof of Proposition 5.6.

First we observe that according to (25) and Proposition 2.1,

$$\begin{aligned}N_k^{r,m}(t)&=\big\{\frac{1}{x_{r,0}}Z_k^{-,r}(0)\big\}\circ\theta_{rm}+\big\{\frac{1}{x_{r0}}A_k^{-,r}(x_{r,0}t)\big\}\circ\theta_{rm}.\\&\le\frac{1}{r}Z_k^{-,r}(rm)+\big\{\frac{1}{x_{r,0}}A_k^{-,r}(x_{r,0}t)\big\}\circ\theta_{rm},\end{aligned} \tag{140}$$

and also that

$$\max_{0\le m<rT}\frac{1}{r}Z_k^{-,r}(rm)\le\sup_{0\le t\le T}\widehat{Z}_k^{-,r}(t). \tag{141}$$

Using the inequality (30), we see that for each $t\ge 0$,

$$\widehat{Z}_k^{-,r}(t)\le\widehat{I}_k^r(t+r^{-1}\widehat{W}_{s(k)}^r(t))-\widehat{I}_k^r(t).$$

Thus, using Propositions 5.2 and 5.3, we have

$$\widehat{Z}_k^{-,r}(\cdot)\Longrightarrow 0 \tag{142}$$

as $r\to\infty$, which yields

$$\lim_{r\to\infty}\mathbb{P}^r\Big(\max_{0\le m<rT}\frac{1}{r}Z_k^{-,r}(rm)>\frac{\varepsilon}{2}\Big)=0,$$

according to (141). So, in virtue of (140), it suffices to show that for each $k \in \mathbb{K}$,

$$\lim_{r\to\infty} \mathbb{P}^r\Big(\max_{0\le m<rT}\big\{\frac{1}{x_{r,0}}A_k^{-,r}(x_{r,0}L)\big\}\circ\theta_{rm} > \frac{\varepsilon}{2}\Big) = 0 \tag{143}$$

in order to obtain the conclusion of the lemma.

Now we have that for each $\delta > 0$ and $M > 0$,

$$\begin{aligned}
&\mathbb{P}^r\Big(\max_{0\le m<rT}\big\{\frac{1}{x_{r,0}}A_k^{-,r}(x_{r,0}L)\big\}\circ\theta_{rm} > \frac{\varepsilon}{2}\Big)\\
&\le \mathbb{P}^r\Big(\max_{0\le m<rT}\big\{\frac{1}{x_{r,0}}A_k^{-,r}(x_{r,0}L)\big\}\circ\theta_{rm} > \frac{\varepsilon}{2}, \frac{|u^{max,T,r}|}{r} \le \delta,\\
&\quad \max_{p\in\mathbb{K}}\max_{0\le m<rT}\Big|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))\Big|\circ\theta_{rm} \le \delta, \sup_{0\le t\le T+L}|\widehat{W}^r(t)| \le M, \sup_{0\le t\le T}|\widehat{Z}^r(t)| \le M\Big)\\
&\quad + \mathbb{P}^r\Big(\frac{|u^{max,T,r}|}{r} > \delta\Big) + \mathbb{P}^r\Big(\max_{p\in\mathbb{K}}\max_{0\le m<rT}\Big|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))\Big|\circ\theta_{rm} > \delta\Big)\\
&\quad + \mathbb{P}^r\Big(\sup_{0\le t\le T+L}|\widehat{W}^r(t)| > M\Big) + \mathbb{P}^r\Big(\sup_{0\le t\le T}|\widehat{Z}^r(t)| > M\Big)
\end{aligned} \tag{144}$$

where $\widehat{\mathcal{V}}_p^{o,r}(\cdot), p \in \mathbb{K}$, and $\overline{Z}_p^{+,r}(\cdot), p \in \mathbb{K}$, are given in (63) and (64), respectively. According to Lemma 5.2,

$$\lim_{r\to\infty} (\text{the second term in (144)}) = 0,$$

and according to assumption (62),

$$\lim_{r\to\infty} (\text{the third term in (144)}) = 0.$$

Further, according to Propositions 5.1 and 5.2,

$$\lim_{M\to\infty}\overline{\lim_{r\to\infty}} (\text{the fourth term in (144)}) = 0$$

and

$$\lim_{M\to\infty}\overline{\lim_{r\to\infty}} (\text{the fifth term in (144)}) = 0.$$

Observe that in addition to (132) and (133),

$$|\widehat{Z}^r(0)|\circ\theta_{rm} \le \sup_{0\le t\le T}|\widehat{Z}^r(t)|, \tag{145}$$

$$\sup_{0\le t\le L}|\widehat{W}^r(t)|\circ\theta_{rm} \le \sup_{0\le t\le T+L}|\widehat{W}^r(t)|, \tag{146}$$

for each $0 \le m < rT$. Then, in use of the Markov property of $\Xi^r(\cdot)$, we see that

$$
\begin{aligned}
&\text{(the first term in (144))}\\
&\le \mathbb{P}^r\Big(\bigcup_{0\le m<rT} \Big\{ \{\frac{1}{x_{r,0}} A_k^{-,r}(x_{r,0}L)\} \circ \theta_{rm} > \frac{\varepsilon}{2}, \frac{|u^{max,T,r}|}{r} \le \delta,\\
&\qquad \max_{p\in\mathbb{K}} \Big| \widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0)) \Big| \circ \theta_{rm} \le \delta, \sup_{0\le t\le L} |\widehat{W}^r(t)| \circ \theta_{rm} \le M, |\widehat{Z}^r(0)| \circ \theta_{rm} \le M \Big\}\Big)\\
&\le \sum_{0\le m<rT} \mathbb{E}^r\big[\mathbb{P}^r(\cdots\cdots\cdots\cdots \mid \mathcal{F}^r_{rm}) \big]\\
&= \sum_{0\le m<rT} \mathbb{E}^r\Big[\mathsf{P}^r_{\Xi^r(rm)}\Big(\frac{1}{x_{r,0}} A_k^{-,r}(x_{r,0}L) > \frac{\varepsilon}{2}, \frac{|u^r(1)|}{r} \le \delta,\\
&\qquad\qquad \max_{p\in\mathbb{K}} \Big| \widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0)) \Big| \le \delta, \sup_{0\le t\le L} |\widehat{W}^r(t)| \le M, |\widehat{Z}^r(0)| \le M \Big)\Big]
\end{aligned}
$$

Thus, in order to show (143), it is enough to prove that for each $\varepsilon > 0$,

$$
\begin{aligned}
&\mathsf{P}^r_*\Big(\frac{1}{x_{r,0}} A_k^{-,r}(x_{r,0}L) > \frac{\varepsilon}{2}, \frac{|u^r(1)|}{r} \le \delta, \max_{p\in\mathbb{K}} \Big| \widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0)) \Big| \le \delta,\\
&\qquad \sup_{0\le t\le L} |\widehat{W}^r(t)| \le M, |\widehat{Z}^r(0)| \le M \Big) < \frac{\varepsilon}{r} \qquad (147)
\end{aligned}
$$

if r is sufficiently large independently of the initial value $*$.

Observe that

$$
\frac{1}{x_{r,0}} A_k^{-,r}(x_{r,0}L) = \frac{1}{x_{r,0}} \sum_{i=1}^{A_k^r(x_{r,0}L)} 1_{\{\gamma_k^{s,r}(i) \le w_k^{s,r}(i)\}}
$$

and if $\sup_{0\le t\le L} |\widehat{W}^r(t)| \le M$, $|\widehat{Z}^r(0)| \le M$ and $r > M > 1$, then

$$
\begin{aligned}
w_k^{s,r}(i) &\le \sup_{0\le t\le x_{r,0}L} W^r_{s(k)}(t) \le \sup_{0\le t\le rML} W^r_{s(k)}(t) \le r \sup_{0\le t\le L} |\widehat{W}^r(t)|\\
&\le rM
\end{aligned}
$$

for each $i = 1, 2, \cdots, A_k^r(x_{r,0}L)$. Thus, if $r > M > 1$, then the left-hand side of (147) is dominated by

$$
\begin{aligned}
&\mathsf{P}^r_*\Big(\frac{1}{x_{r,0}} \sum_{i=1}^{A_k^r(x_{r,0}L)} 1_{\{\gamma_k^{s,r}(i) \le rM\}} > \frac{\varepsilon}{2}, \max_{p\in\mathbb{K}} \Big| \widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0)) \Big| \le \delta, \frac{|u^r(1)|}{r} \le \delta \Big)\\
&\le \mathsf{P}^r_*\Big(\frac{1}{x_{r,0}} \sum_{i=1}^{\lfloor cx_{r,0} \rfloor} 1_{\{\gamma_k^{s,r}(i) \le rM\}} > \frac{\varepsilon}{2} \Big)\\
&\quad + \mathsf{P}^r_*\Big(A_k^r(x_{r,0}L) > cx_{r,0}, \max_{p\in\mathbb{K}} \Big| \widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0)) \Big| \le \delta, \frac{|u^r(1)|}{r} \le \delta \Big)\\
&\equiv (i) + (ii), \qquad (148)
\end{aligned}
$$

where c is any positive constant. (The value of c will be appropriately determined below).

We first evaluate the term (i) in (148). Note that

$$\frac{1}{x_{r,0}}\sum_{i=1}^{\lfloor cx_{r,0}\rfloor} 1_{\{\gamma_k^{s,r}(i)\le rM\}}$$
$$=\frac{1}{x_{r,0}}\sum_{i=1}^{\lfloor cx_{r,0}\rfloor}\big(1_{\{\gamma_k^{s,r}(i)\le rM\}}-F_k^{\gamma,r}(rM)\big)+\frac{1}{x_{r,0}}F_k^{\gamma,r}(rM)\lfloor cx_{r,0}\rfloor.$$

Then, since $F_k^{\gamma,r}(rM)\to 0$ as $r\to\infty$ because of (57), we have that

$$\mathsf{P}_*^r\Big(\frac{1}{x_{r,0}}F_k^{\gamma,r}(rM)\lfloor cx_{r,0}\rfloor>\frac{\varepsilon}{4}\Big)=1_{\left\{x_{r,0}^{-1}F_k^{\gamma,r}(rM)\lfloor cx_{r,0}\rfloor>\varepsilon/4\right\}}=0$$

for sufficiently large r independently of the value $*$.

Further, we have that for each $\varepsilon>0$,

$$\mathsf{P}_*^r\Big(\sup_{0\le t\le c}\Big|\sum_{i=1}^{\lfloor x_{r,0}t\rfloor}\big(1_{\{\gamma_k^{s,r}(i)\le rM\}}-F_k^{\gamma,r}(rM)\big)\Big|>x_{r,0}\frac{\varepsilon}{4}\Big)$$
$$\le\frac{4^4}{(x_{r,0}\varepsilon)^4}\mathsf{E}_*^r\Big[\Big\{\sum_{i=1}^{\lfloor x_{r,0}c\rfloor}\big(1_{\{\gamma_k^{s,r}(i)\le rM\}}-F_k^{\gamma,r}(rM)\big)\Big\}^4\Big]$$
$$\le\frac{4^4}{(x_{r,0}\varepsilon)^4}\cdot 3(x_{r,0}c)^2\le\frac{768c^2}{(x_{r,0})^2\varepsilon^4}\le\frac{768c^2}{r^2\varepsilon^4},$$

where the first inequality is due to Doob's submartingale inequality. Therefore, if r is sufficiently large such that

$$\frac{1}{r}<\frac{\varepsilon^5}{768c^2},$$

then

$$\text{the term }(i)\le\frac{\varepsilon}{r}. \tag{149}$$

We next evaluate the term (ii) in (148). Because of (149), it is enough to show that for each $k\in\mathbb{K}$, there exists some constant $c>0$ such that

$$\mathsf{P}_*^r\Big(A_k^r(x_{r,0}L)\ge cx_{r,0},\max_{p\in\mathbb{K}}\Big|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))\Big|\le\delta,\frac{|u^r(1)|}{r}\le\delta\Big)\le\frac{\varepsilon}{r}, \tag{150}$$

if r is sufficiently large independently of $*$. Because of (15) and (16), we have only to show that for each $k\in\mathbb{A}$ and $l\in\mathbb{K}$, there exists some constant $c,c_1,c_2,c_3>0$ such that

$$\mathsf{P}_*^r\Big(E_k^r(x_{r,0}L)\ge\frac{1}{2}cx_{r,0},\frac{|u^r(1)|}{r}\le\delta\Big)\le c_1\frac{\varepsilon}{r}, \tag{151}$$

$$\mathsf{P}_*^r\Big(F_l^r(x_{r,0}L)\ge\frac{1}{2}cx_{r,0},\max_{p\in\mathbb{K}}\Big|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))\Big|\le\delta\Big)\le(c_2+c_3)\frac{\varepsilon}{r}, \tag{152}$$

if r is sufficiently large independently of $*$. Using Lemma 7.2 in the Appendix, we immediately have (151) with $c\ge 2\sup_r\alpha_k^r\dot{L}+1$ and any $\varepsilon\in(0,1)$. Further, using Lemmas 7.3 and 7.4, we have (152) with $c\ge 4\sum_{p=1}^K\sup_r P_{pl}^r\mu_p+1$. Therefore (150) has been established so that the conclusion of the lemma follows.

□

Proof of Proposition 5.4.

According to Proposition 5.5, the methodology employed in Bramson [3], specifically the contents of Sect. 5 and Sect. 6 in [3], also applies to the demonstration of multiplicative strong state-space collapse, i.e., (127) in our multiclass feedforward queueing network with abandonment under the FCFS service discipline. Thus, using Proposition 5.2, we see that strong state-space collapse holds, i,e.,

$$\sup_{0\le t\le T} \sup_{0\le s\le \widehat{W}^r_{s(k)}(t)} \mid r^{-1}D^r_k(r^2t+rs) - r^{-1}D^r_k(r^2t) - \lambda^r_k s \mid \longrightarrow 0 \quad \text{in pr.}$$

as $r\to\infty$, for each $k\in\mathbb{K}$. In particular, we have

$$\sup_{0\le t\le T} \mid r^{-1}D^r_k(r^2t+r\widehat{W}^r_{s(k)}(t)) - r^{-1}D^r_k(r^2t) - \lambda^r_k \widehat{W}^r_{s(k)}(t) \mid \longrightarrow 0 \quad \text{in pr.} \tag{153}$$

as $r\to\infty$.

On the other hand, because of the FCFS service discipline with abandonment, we have

$$r^{-1}D^r_k(r^2t+r\widehat{W}^r_{s(k)}(t)) - r^{-1}D^r_k(r^2t) + \widehat{Z}^{-,r}_k(t) = \widehat{Z}^r_k(t) \tag{154}$$

for each $k\in\mathbb{K}$. Also recall that for each $T>0$,

$$\sup_{0\le t\le T} \widehat{Z}^{-,r}_k(t) \longrightarrow 0 \quad \text{in pr.} \tag{155}$$

as $r\to\infty$, as established in (142). Then, combining (155) with (153) and (154), we have the condition of state-space collapse (128).

□

5.4 Proof of the diffusion approximation theorem (i.e., Theorem 4.1) Before presenting the proof of the theorem, we show the next lemma on the fluid limits of $\{A^r(\cdot)\}_r$ and $\{D^r(\cdot)\}_r$, which corresponds to Lemma 8.2 in Williams [26].

Lemma 5.3.

For each $k\in\mathbb{K}$ and $T>0$,

$$\sup_{0\le t\le T} |\overline{A}^r_k(t) - \lambda_k t| \longrightarrow 0 \quad \textit{in pr.},$$
$$\sup_{0\le t\le T} |\overline{D}^r_k(t) - \lambda_k t| \longrightarrow 0 \quad \textit{in pr.},$$

as $r\to\infty$.

Proof.

¿From (17), (78), (80), (82) and (83), we have

$$\overline{Z}^r_k(t) = \overline{Z}^r_k(0) + \overline{A}^r_k(t) - \overline{D}^r_k(t) - \overline{I}^r_k(t)$$

for each $k\in\mathbb{K}$ and $t\ge 0$. Because of Propositions 5.1 and 5.3, we see that for each $T>0$,

$$\sup_{0\le t\le T} \overline{Z}^r_k(t) \longrightarrow 0 \quad \text{in pr.},$$
$$\sup_{0\le t\le T} \overline{I}^r_k(t) \longrightarrow 0 \quad \text{in pr.}$$

as $r \to \infty$. So we have that for each $T > 0$,

$$\sup_{0 \le t \le T} |\overline{A}_k^r(t) - \overline{D}_k^r(t)| \longrightarrow 0 \quad \text{in pr.} \tag{156}$$

as $r \to \infty$.

Fix any $t \ge 0$. Then, according to (120), $\{\overline{A}_k^r(t)\}_r$ is tight in $\mathcal{R}^1$ for each $k \in \mathbb{K}$, which yields that for any subsequence $\{r'\}$ of $\{r\}$, there exists some further subsequence $\{r''\}$ of $\{r'\}$ such that

$$\overline{A}_k^{r''}(t) \Longrightarrow a_k(t) \quad \text{in} \quad \mathcal{R}^1$$

as $r'' \to \infty$, for some r.v. $a_k(t)$. Thus we also have

$$\overline{D}_k^{r''}(t) \Longrightarrow a_k(t) \quad \text{in} \quad \mathcal{R}^1$$

as $r'' \to \infty$, because of (156). Therefore, from (15), (46) and (86), it follows that

$$a_k(t) = \alpha_k t + \sum_{l=1}^{K} P_{lk} a_l(t)$$

for each $k \in \mathbb{K}$, which implies $a_k(t) = \lambda_k t, k \in \mathbb{K}$. Consequently we have proved that

$$\overline{A}_k^r(t) \Longrightarrow \lambda_k t \quad \text{in} \quad \mathcal{R}^1,$$

$$\overline{D}_k^r(t) \Longrightarrow \lambda_k t \quad \text{in} \quad \mathcal{R}^1$$

as $r \to \infty$, for each $t \ge 0$ and $k \in \mathbb{K}$. Therefore, in virtue of Polya's theorem (cf. Problem 5.3.2 in Liptser and Shiryayev [20]), we obtain the conclusion.

□

The next lemma identifies the weak limit of scaled abandonment-count process as a functional of the limit of scaled workload process, which is similar in form to the case of heavy-traffic limit for a many-server queue with abandonment under the hazard-type scaling of abandonment distribution (cf. Lemma 2.7 in Katsuda [17]), with the difference of multiplicative constant due to our multiclass setting.

Lemma 5.4.

Suppose that

$$\widehat{W}^r(\cdot) \Longrightarrow W^*(\cdot) \quad \textit{in} \quad \mathbb{D}([0,\infty), \mathcal{R}^J),$$

as $r \to \infty$. Then we have that

$$\widehat{I}_k^r(\cdot) \Longrightarrow \lambda_k \int_0^{\cdot} H_k(W^*_{s(k)}(u))du \quad \textit{in} \quad \mathbb{D}([0,\infty), \mathcal{R}^1), \tag{157}$$

as $r \to \infty$, for each $k \in \mathbb{K}$.

Proof.

According to (25), (112) and (155), we have that for each $k \in \mathbb{K}$,

$$\sup_{0 \le t \le T} |\widehat{I}_k^r(t) - \widehat{A}_k^{-,r}(t)| \longrightarrow 0 \quad \text{in pr.}$$

as $r \to \infty$. Thus, because of (113) and (121),

$$\sup_{0 \le t \le T} |\widehat{I}_k^r(t) - \widehat{\mathcal{C}}_k^r(\overline{A}_k^r(t))| \longrightarrow 0 \quad \text{in pr.} \tag{158}$$

as $r \to \infty$, with $\widehat{\mathcal{C}}_k^r(\cdot)$ in (115).

Observing that according to (115),

$$\int_0^t rF_k^{\gamma,r}(r\widehat{W}_{s(k)}^r(u-))d\overline{A}_k^r(u) \le \widehat{\mathcal{C}}_k^r(\overline{A}_k^r(t)) \le \int_0^t rF_k^{\gamma,r}(r\widehat{W}_{s(k)}^r(u))d\overline{A}_k^r(u)$$

for each $t \ge 0$, we have the convergence (157) in virtue of (57), (158) and Lemma 5.3 in the same way as in the proof of Lemma 2.7 in [17].

□

Proof of Theorem 4.1.

The first half of the proof uses an analogous argument to the proof of Theorem 7.1 in Williams [26] as follows.
From (21), (67) and (68), we have that for each $j \in \mathbb{J}$,

$$\begin{aligned}
\widehat{W}_j^r(t) &= \widehat{W}_j^r(0) + \sum_{k\in C(j)} \frac{1}{r} \sum_{i=1}^{A_k^{+,r}(r^2t)} (v_k^{s,r}(i) - m_k^r) + \sum_{k\in C(j)} \frac{1}{r} m_k^r (A_k^r(r^2t) - A_k^{-,r}(r^2t)) \\
&\quad - rt + \widehat{Y}_j^r(t) \\
&= \widehat{W}_j^r(0) + \sum_{k\in C(j)} \widehat{\mathcal{V}}_k^{s,r}(\overline{A}_k^{+,r}(t)) + \sum_{k\in C(j)} m_k^r \widehat{A}_k^r(t) - \sum_{k\in C(j)} m_k^r \widehat{\mathcal{M}}_k^{\gamma,r}(\overline{A}_k^r(t)) \\
&\quad - \sum_{k\in C(j)} m_k^r \widehat{\mathcal{C}}_k^r(\overline{A}_k^r(t)) + r(\rho_j^r - 1)t + \widehat{Y}_j^r(t) \qquad (159)
\end{aligned}$$

with $\widehat{\mathcal{V}}_k^{s,r}(\cdot)$ in (70), $\widehat{\mathcal{M}}_k^{\gamma,r}(\cdot)$ in (114) and $\widehat{\mathcal{C}}_k^r(\cdot)$ in (115) for each $k \in \mathbb{K}$. In vector form, (159) is represented as

$$\begin{aligned}
\widehat{W}^r(t) &= \widehat{W}^r(0) + C\widehat{\mathcal{V}}^{s,r}(\overline{A}^{+,r}(t)) + CM^r\widehat{A}^r(t) - CM^r\widehat{\mathcal{M}}^{\gamma,r}(\overline{A}^r(t)) \\
&\quad - CM^r\widehat{\mathcal{C}}^r(\overline{A}^r(t)) + r(\rho^r - e)t + \widehat{Y}^r(t). \qquad (160)
\end{aligned}$$

On the other hand, using (109), we have

$$\begin{aligned}
CM^r\widehat{A}^r(t) &= CM^rQ^r\big\{\widehat{E}^r(t) + \sum_{l=1}^K \widehat{\Phi}^{l,r}(\overline{D}_l^r(t))\big\} \\
&\quad - CM^rQ^r\widetilde{P}^r(\widehat{Z}^r(t) - \widehat{Z}^r(0)) - CM^rQ^r\widetilde{P}^r\widehat{I}^r(t) \\
&= CM^rQ^r\big\{\widehat{E}^r(t) + \sum_{l=1}^K \widehat{\Phi}^{l,r}(\overline{D}_l^r(t))\big\} - CM^rQ^r\widetilde{P}^r(\widehat{\epsilon}^r(t) - \widehat{\epsilon}^r(0)) \\
&\quad - G^r(\widehat{W}^r(t) - \widehat{W}^r(0)) - CM^rQ^r\widetilde{P}^r\widehat{I}^r(t) \qquad (161)
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\epsilon}^r(t) &= (\widehat{\epsilon}_k^r(t), k \in \mathbb{K}) \quad \text{with} \quad \widehat{\epsilon}_k^r(t) \equiv \widehat{Z}_k^r(t) - \lambda_k^r\widehat{W}_{s(k)}^r(t), k \in \mathbb{K}, \\
G^r &\equiv CM^rQ^r\widetilde{P}^r\Lambda^r.
\end{aligned}$$

Therefore, substituting (161) into (160) and using assumption (A.4), we have

$$\widehat{W}^r(t) = \widehat{X}^r(t) + R^r\widehat{Y}^r(t) \tag{162}$$

for sufficiently large r, where

$$\begin{aligned} R^r &\equiv (1+G^r)^{-1}, \\ \widehat{X}^r(t) &\equiv \widehat{W}^r(0) + R^r(\widehat{\xi}^r(t) + \widehat{\eta}^r(t) + \widehat{\zeta}^r(t)) \end{aligned} \tag{163}$$

with

$$\begin{aligned} \widehat{\xi}^r(t) &\equiv C\widehat{\mathcal{V}}^{s,r}(\overline{A}^{+,r}(t)) + CM^rQ^r\big\{\widehat{E}^r(t) + \sum_{l=1}^{K}\widehat{\Phi}^{l,r}(\overline{D}_l^r(t))\big\} \\ &\quad - C\widehat{\mathcal{M}}^{\gamma,r}(\overline{A}^r(t)), \\ \widehat{\eta}^r(t) &\equiv r(\rho^r - e)t - CM^rQ^r\widetilde{P}^r\widehat{I}^r(t) - CM^r\widehat{\mathcal{C}}^r(\overline{A}^r(t)), \\ \widehat{\zeta}^r(t) &\equiv CM^rQ^r\widetilde{P}^r(\widehat{\epsilon}^r(0) - \widehat{\epsilon}^r(t)). \end{aligned}$$

Using (86), (87), (88), (121) and Lemma 5.3, we see that

$$\widehat{\xi}^r(\cdot) \Longrightarrow \xi^*(\cdot) \qquad \text{in} \quad \mathbb{D}([0,\infty), \mathcal{R}^J) \tag{164}$$

as r goes to infinity, where

$$\xi^*(t) = C\mathcal{V}^*(\lambda t) + CMQ\big\{E^*(t) + \sum_{l=1}^{K}\Phi^{l,*}(\lambda_l t)\big\}. \tag{165}$$

Applying the oscillation inequality in Williams [25] to (162) as in (120) of Williams [26], we have that for $w_T(x(\cdot),\delta)$ in (124),

$$Osc(x(\cdot), I) \equiv \sup_{u,v\in I}|x(u) - x(v)|, \qquad I \subset \mathcal{R}^1,$$

and sufficiently large r,

$$\begin{aligned} w_T(\widehat{W}^r(\cdot),\delta) &= \sup_{u\in[0,T-\delta]} Osc\big(\widehat{W}^r(\cdot),[u,u+\delta]\big) \\ &\le const\cdot \sup_{u\in[0,T-\delta]} Osc\big(\widehat{X}^r(\cdot),[u,u+\delta]\big) \\ &= const\cdot w_T(\widehat{X}^r(\cdot),\delta), \end{aligned}$$

so that

$$w_T(\widehat{W}^r(\cdot),\delta) \le const\cdot\big\{w_T(\widehat{\xi}^r(\cdot),\delta) + w_T(\widehat{\eta}^r(\cdot),\delta) + w_T(\widehat{\zeta}^r(\cdot),\delta)\big\}$$

for each $T>0$ and $\delta>0$, because of (163).

The convergence (164) implies

$$\lim_{\delta\to 0}\overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(w_T(\widehat{\xi}^r(\cdot),\delta) > \varepsilon\big) = 0, \quad \forall\varepsilon > 0.$$

(Cf. Proposition 6.3.26 in Jacod and Shiryaev [14]).

In addition, from the heavy-traffic condition (56) and the C-tightness of both $\{\widehat{I}_k^r(\cdot)\}_r$ and $\{\widehat{\mathcal{C}}_k^r(\overline{A}_k^r(\cdot))\}_r$ already established, we have

$$\lim_{\delta\to 0}\overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(w_T(\widehat{\eta}^r(\cdot),\delta)>\varepsilon\big)=0,\quad \forall\varepsilon>0.$$

Therefore, by virtue of the condition of state-space collapse (128), we have

$$\lim_{\delta\to 0}\overline{\lim_{r\to\infty}}\,\mathbb{P}^r\big(w_T(\widehat{W}^r(\cdot),\delta)>\varepsilon\big)=0,\quad \forall\varepsilon>0. \tag{166}$$

Combining (166) with Proposition 5.2, we obtain the C-tightness of $\{\widehat{W}^r(\cdot)\}_r$.

Let $W^*(t), t\geq 0$, be any limit process of the sequence $\{\widehat{W}^r(\cdot)\}_r$, and suppose that a subsequence $\{r'\}$ of $\{r\}$ satisfies

$$\widehat{W}^{r'}(\cdot)\Longrightarrow W^*(\cdot)\quad\text{in}\quad\mathbb{D}([0,\infty),\mathcal{R}^J),$$

as $r'\to\infty$. Then, according to Lemma 5.4, we have that for each $k\in\mathbb{K}$,

$$\widehat{I}_k^{r'}(\cdot)\Longrightarrow\lambda_k\int_0^{\cdot}H_k(W^*_{s(k)}(u))du\qquad\text{in}\quad\mathbb{D}([0,\infty),\mathcal{R}^1) \tag{167}$$

and

$$\widehat{\mathcal{C}}_k^{r'}(\overline{A}_k^{r'}(\cdot))\Longrightarrow\lambda_k\int_0^{\cdot}H_k(W^*_{s(k)}(u))du\qquad\text{in}\quad\mathbb{D}([0,\infty),\mathcal{R}^1) \tag{168}$$

as r' goes to infinity. Therefore, using (167), (168) and (56), we have

$$\widehat{\eta}^{r'}(\cdot)\Longrightarrow\eta^*(\cdot) \tag{169}$$

as r' goes to infinity, where

$$\eta^*(t)=\vartheta t-CMQ\Big(\lambda_k\int_0^t H_k(W^*_{s(k)}(u))du,k\in\mathbb{K}\Big),\qquad t\geq 0.$$

Consequently, substituting assumption (A.1), (164), (128) and (169) into (163) and using assumption (A.4), we have

$$\widehat{X}^{r'}(\cdot)\Longrightarrow X^*(\cdot) \tag{170}$$

as r' goes to infinity, where

$$X^*(t)=W^*(0)+R(\xi^*(t)+\eta^*(t)),\qquad t\geq 0. \tag{171}$$

Therefore, *any* limit process $W^*(\cdot)$ of the C-tight sequence $\{\widehat{W}^r(\cdot)\}_r$ is a semimartingale reflecting Brownian motion (SRBM) with a nonlinear drift term, i.e., (94) and (95). Applying the Girsanov transformation technique to the localized version of such SRBM (cf. the proof of Theorem 2.1 in Katsuda [17], for example), we can reduce the uniqueness in law of $W^*(\cdot)$ to that of SRBM, so that the desired convergence

$$(\widehat{W}^r(\cdot),\widehat{Y}^r(\cdot))\Longrightarrow(W^*(\cdot),Y^*(\cdot)), \tag{172}$$

as $r\to\infty$ has been shown. Combining (172) with the result on state-space collapse, i.e., Proposition 5.4, we also have the convergence

$$\widehat{Z}^r(\cdot)\Longrightarrow Z^*(\cdot)$$

as $r\to\infty$, where $Z^*(\cdot)=(\lambda_k W^*_{s(k)}(\cdot),k\in\mathbb{K})$, so the proof of the theorem has been completed.

□

6 Final remarks

As an example of our diffusion approximation with the unstable random behavior of abandonment time near the origin, consider a GI/GI/1+GI queue for which the abandonment time is distributed according to the Gamma distribution

$$G_p(x) = \int_0^x g_p(u)du, x \geq 0, \quad g_p(u) = (\Gamma(p))^{-1}u^{p-1}e^{-u}, u \geq 0,$$

with $p \in (0, 1)$. Then, its hazard-rate function $h_p(x) = g_p(x)/(1 - G_p(x))$ is not locally bounded so that the diffusion approximation result in the literature such as [24] and [21] is inapplicable. However, in virtue of our general hazard-type scaling, our main result does hold in ths case.

In this paper we impose the feedforward routing condition on our multiclass queueing networks (MQNs) and the only place where it is used is the proof of the stochstic boundedness of queue length. So, if it is established without such restriction, our main result is valid for general MQNs with abandonment.

One of the most important studies around diffusion approximations of queueing systems is the application of such approximations to the validation of steady-state approximations of those systems. Gamarnik and Zeevi [12] is a seminal work of the study, in which steady-state approximations for generalized Jackson networks have been validated under the condition of the existence of moment generating functions for primitive model variables. It is also noted that such relatively restrictive assumption can be relaxed to moment condition of p-th order with $p \geq 2$ by the work of Budhiraja and Lee [5] in conjunction with the appendix of Krichagina and Taksar [19]. Furthermore, the author's works [15, 16] used the Lyapunov function method of [12] and the framework on the uniform moment bounds of the Markov state process in [5], respectively, to study such steady-state analysis of a multiclass single-server queue in heavy traffic under various service disciplines.

In this paper we have proved the diffusion approximation theorem for multiclass feedforward queueing networks with abandonments under FCFS service disciplines, and so we are interested in steady-state approximations of those networks as an application of our theorem. Restricting our attention to a multiclass single-server queue with $\vartheta < 0$ (in heavy-traffic condition (56)), we are able to validate such approximation of the queue with abandonment in a similar fashion to [15] and [16], in which conditions (A.1), (A.2) and (A.3) of this paper may be verified to hold in stationarity. However, checking the case with $\vartheta \geq 0$ remains unresolved and is worth pursuing in future research. More specifically, it is solved if the following two tasks are done:

(i) To seek a sufficient condition for the stability of multiclass feedforward queueing networks with abandonments.

To be expected from the literature (cf. Baccelli et al. [1], Dai [6]), the condition is such that the traffic intensity at each station may possibly be greater than unity in such a way that its excess over unity can be balanced out by the effect of abandonment;

(ii) To show the tightness of stationary workload and queue length in the queue with abandonment for the verification of conditions (A.1), (A.2) and (A.3) in stationarity.

As concerns the issue (i), in his recent work [18] the author has given a stability condition for those networks, which involves the total probability mass of abandonment time in addition to the model parameters of networks.

7 Appendix

This appendix corresponds to Sect. 5 of Bramson [3] in which hydrodynamically scaled performance measure processes for multiclass queueing networks are asymptotically estimated as approximately Lipschitz continuous. Different from [3], our argument employs

such scaling in association with the shift transformation of the description process $\Xi(\cdot)$ in Sect. 3 and uses its Markov property to obtain such asymptotic estimation of performance measure processes in our queueing network.

The next lemma corresponds to Proposition 4.2 in Bramson [3] and plays a fundamental role in proving the rest of the lemmas as in [3].

Lemma 7.1.

Suppose that the sequence of r.v.'s $\{X^r(i), i \geq 1\}$ *is i.i.d. for each* $r \geq 1$, *and* $\{X^r(1)^2\}_{r\geq 1}$ *is uniformly integrable. Let* $\mathcal{S}^r(i) \equiv \sum_{j=1}^{i} X^r(j), i \geq 1$, *and* $\mu_X^r \equiv \mathbb{E}^r[X^r(1)]$, $r \geq 1$. *Then, for each* $\varepsilon > 0$,

$$\sup_r \mathbb{P}^r\big(\max_{1\leq i\leq n} |\mathcal{S}^r(i) - i\mu_X^r| > \varepsilon n\big) < \frac{\varepsilon}{n}$$

if n *is sufficiently large.*

Lemma 7.2.

For each $\varepsilon > 0$ *and* $k \in \mathbb{K}$, *there exist constants* $\delta_1 > 0$ *and* $c_1 > 0$ *such that*

$$\mathsf{P}_*^r\Big(\sup_{0\leq t\leq x_{r,0}L} |E_k^r(t) - \alpha_k^r t| > x_{r,0}\varepsilon, \frac{|u^r(1)|}{r} \leq \delta_1\Big) < c_1 \cdot \frac{\varepsilon}{r} \tag{173}$$

if r *is sufficiently large independently of* $*$, *where* $\mathsf{P}_*^r(\cdot)$ *is the probability law of Markov process* $\Xi^r(\cdot)$ *starting with the value* $*$ *for each* $r \geq 1$. (*Cf.* (33)).

Proof.

First observe that the inequality

$$\sup_{0\leq t\leq x_{r,0}L} |E_k^r(t) - \alpha_k^r t| \geq x_{r,0}\varepsilon$$

implies that there exists some $t \in [0, x_{r,0}L]$ such that either

$$E_k^r(t) \geq \alpha_k^r t + x_{r,0}\varepsilon \tag{174}$$

or

$$E_k^r(t) \leq \alpha_k^r t - x_{r,0}\varepsilon. \tag{175}$$

The inequality (174) and condition (46) implies that

$$\mathcal{U}_k^r(\lfloor \alpha_k^r t + x_{r,0}\varepsilon\rfloor) - \lfloor \alpha_k^r t + x_{r,0}\varepsilon\rfloor \frac{1}{\alpha_k^r} \leq -\frac{x_{r,0}\varepsilon}{2\alpha_k} \tag{176}$$

if r is sufficiently large. Similarly, the inequality (175) implies that

$$\mathcal{U}_k^r(\lfloor \alpha_k^r t - x_{r,0}\varepsilon\rfloor + 1) - (\lfloor \alpha_k^r t - x_{r,0}\varepsilon\rfloor + 1)\frac{1}{\alpha_k^r} > \frac{x_{r,0}\varepsilon}{2\alpha_k} \tag{177}$$

if r is sufficiently large.

Thus, noting that for each $t \in [0, x_{r,0}L]$,

$$\lfloor \alpha_k^r t + x_{r,0}\varepsilon\rfloor < x_{r,0}(\alpha_k L + 1) \tag{178}$$

if r is sufficiently large, where we suppose $\varepsilon \in (0, \frac{1}{2})$, and using (176), (177) and (178), we have

$$\begin{aligned} &\mathsf{P}_*^r\Big(\sup_{0\le t\le x_{r,0}L} |E_k^r(t) - \alpha_k^r t| \ge x_{r,0}\varepsilon, \frac{|u^r(1)|}{r} \le \delta\Big) \\ &\le \mathsf{P}_*^r\Big(\max_{1\le i\le \lfloor x_{r,0}(\alpha_k L+1)\rfloor} \Big|\mathcal{U}_k^r(i) - \frac{1}{\alpha_k^r} i\Big| > \frac{x_{r,0}\varepsilon}{2\alpha_k}, \frac{|u^r(1)|}{r} \le \delta\Big). \end{aligned} \tag{179}$$

Suppose that the constant $\delta_1 > 0$ satisfies the inequality

$$\delta_1 \le \frac{\varepsilon}{4\alpha_k} - \frac{2}{\alpha_k r}$$

for sufficiently large r satisfying $\frac{\varepsilon}{4\alpha_k} - \frac{2}{\alpha_k r} > 0$. Then, when $\frac{|u^r(1)|}{r} \le \delta_1$, we have that for each $i \ge 1$,

$$\begin{aligned} \Big|\mathcal{U}_k^r(i) - \frac{1}{\alpha_k^r} i\Big| &\le u_k^r(1) + \frac{1}{\alpha_k^r} + \Big|\sum_{j=2}^{i}\Big(u_k^r(j) - \frac{1}{\alpha_k^r}\Big)\Big| \\ &\le \delta_1 r + \frac{2}{\alpha_k} + \Big|\sum_{j=2}^{i}\Big(u_k^r(j) - \frac{1}{\alpha_k^r}\Big)\Big| \\ &\le \frac{x_{r,0}\varepsilon}{4\alpha_k} + \Big|\sum_{j=2}^{i}\Big(u_k^r(j) - \frac{1}{\alpha_k^r}\Big)\Big|, \end{aligned}$$

where we set $\sum_{j=2}^{i} \cdots \equiv 0$ when $i = 1$. Therefore, applying Lemma 7.1 and observing that $x_{r,0}$ is a function of $*$ on the event inside P_*, we have that the display (179) with $\delta = \delta_1$ is dominated by

$$\begin{aligned} &\mathsf{P}_*^r\Big(\max_{2\le i\le \lfloor x_{r,0}(\alpha_k L+1)\rfloor} \Big|\sum_{j=2}^{i}\Big(u_k^r(j) - \frac{1}{\alpha_k^r}\Big)\Big| > \frac{x_{r,0}\varepsilon}{4\alpha_k}\Big) \\ &\le \frac{x_{r,0}\varepsilon}{4\alpha_k} \times \frac{1}{(\lfloor x_{r,0}(\alpha_k L+1)\rfloor - 1)^2} \\ &\le \frac{x_{r,0}\varepsilon}{\alpha_k} \times \frac{1}{\{x_{r,0}(\alpha_k L+1)\}^2} \\ &\le \frac{1}{\alpha_k(\alpha_k L+1)^2} \times \frac{\varepsilon}{r}, \end{aligned}$$

if r is sufficiently large. Letting $c_1 \equiv \frac{1}{\alpha_k(\alpha_k L+1)^2}$, we have the conclusion of the lemma. □

The next lemma corresponds to Lemma 5.2 in Bramson [3], and it will be used in the proof of Lemma 7.4 below.

Lemma 7.3.

For each $\varepsilon > 0$ and $k \in \mathbb{K}$, there exist constants $\delta > 0$ and $c_2 > 0$ such that

$$\begin{aligned} &\mathsf{P}_*^r\Big(\sup_{t_1,t_2\in[0,x_{r,0}L]} \big(|D_k^r(t_2) - D_k^r(t_1)| - \mu_k|t_2 - t_1|\big) \ge x_{r,0}\varepsilon, |\widehat{\mathcal{V}}_k^{o,r}(\overline{Z}_k^{+,r}(0))| < \delta\Big) \\ &< c_2 \cdot \frac{\varepsilon}{r}, \end{aligned} \tag{180}$$

if r is sufficiently large independently of $*$. In particular,

$$\mathsf{P}_*^r\Big(D_k^r(x_{r,0}L) \ge 2\mu_k x_{r,0}L,\, |\widehat{\mathcal{V}}_k^{o,r}(\overline{Z}_k^{+,r}(0))| < \delta\Big) < c_2 \cdot \frac{\varepsilon}{r}, \tag{181}$$

if r is sufficiently large independently of $*$.

Proof.

Let

$$\xi_k^r(s) \equiv \max\{n \in \mathbb{N} : \mathcal{V}_k^r(n) \le s\}, \qquad s \ge 0, \quad k \in \mathbb{K},$$

where

$$\mathcal{V}_k^r(n) \equiv \mathcal{V}_k^{o,r}(Z_k^{+,r}(0) \wedge n) + \mathcal{V}_k^{s,r}((n - Z_k^{+,r}(0))^+). \tag{182}$$

Then

$$D_k^r(t) = \xi_k^r(T_k^r(t)), \qquad t \ge 0, \quad k \in \mathbb{K}.$$

Since

$$D_k^r(t_2) - D_k^r(t_1) - \mu_k^r(t_2 - t_1) \le \xi_k^r(T_k^r(t_2)) - \xi_k^r(T_k^r(t_1)) - \mu_k^r(T_k^r(t_2) - T_k^r(t_1))$$

for each $t_1, t_2 \in [0, x_{r,0}L]$ such that $t_1 \le t_2$, we have

$$\begin{aligned}
&\sup_{t_1,t_2\in[0,x_{r,0}L]} \{|D_k^r(t_2) - D_k^r(t_1)| - \mu_k^r|t_2 - t_1|\} \\
&\le \sup_{s_1,s_2\in[0,x_{r,0}L]} \{|\xi_k^r(s_2) - \xi_k^r(s_1)| - \mu_k^r|s_2 - s_1|\}. \\
&\le 2 \sup_{s\in[0,x_{r,0}L]} |\xi_k^r(s) - \mu_k^r s|.
\end{aligned}$$

Thus, it suffices to show

$$\mathsf{P}_*^r\Big(\sup_{s\in[0,x_{r,0}L]} \big|\xi_k^r(s) - \mu_k^r s\big| \ge \frac{x_{r,0}\varepsilon}{2}, |\widehat{\mathcal{V}}_k^{o,r}(\overline{Z}_k^{+,r}(0))| < \delta\Big) < c_2 \cdot \frac{\varepsilon}{r}, \tag{183}$$

if r is sufficiently large independently of $*$. In the same way as in the derivation of (179), the left-hand side of (183) is majorized by

$$\mathsf{P}_*^r\Big(\max_{1\le i\le\lfloor x_{r,0}(\mu_k L+1)\rfloor} \Big|\mathcal{V}_k^r(i) - m_k^r i\Big| > \frac{x_{r,0}\varepsilon}{4\mu_k}, |\widehat{\mathcal{V}}_k^{o,r}(\overline{Z}_k^{+,r}(0))| < \delta\Big). \tag{184}$$

Suppose that the constant $\delta > 0$ satisfies the inequality $\delta < \frac{\varepsilon}{8\mu_k}$. Then, when $i > Z_k^{+,r}(0)$ and $|\widehat{\mathcal{V}}_k^{o,r}(\overline{Z}_k^{+,r}(0))| < \delta$, we have

$$\begin{aligned}
\big|\mathcal{V}_k^r(i) - m_k^r i\big| &\le r\delta + \big|\mathcal{V}_k^{s,r}(i - Z_k^{+,r}(0)) - m_k^r(i - Z_k^{+,r}(0))\big| \\
&\le \frac{x_{r,0}\varepsilon}{8\mu_k} + \big|\mathcal{V}_k^{s,r}(i - Z_k^{+,r}(0)) - m_k^r(i - Z_k^{+,r}(0))\big|. \qquad (185)
\end{aligned}$$

Therefore we have

$$\begin{aligned}
(184) &\le \mathsf{P}_*^r\Big(\max_{1\le j\le\lfloor x_{r,0}(\mu_k L+1)\rfloor - Z_k^{+,r}(0)} \Big|\mathcal{V}_k^{s,r}(j) - m_k^r j\Big| > \frac{x_{r,0}\varepsilon}{8\mu_k}\Big) \\
&\le \frac{x_{r,0}\varepsilon}{8\mu_k} \times \frac{1}{(\lfloor x_{r,0}(\mu_k L + 1)\rfloor - Z_k^{+,r}(0))^2} \\
&\le \frac{x_{r,0}\varepsilon}{2\mu_k} \times \frac{1}{(x_{r,0}\mu_k L)^2} \\
&\le \frac{1}{2\mu_k^3 L^2} \times \frac{\varepsilon}{r}
\end{aligned}$$

if r is sufficiently large, where the second inequality is a consequence of the application of Lemma 7.1 and the third inequality follows from $Z_k^{+,r}(0) \le x_{r,0}$. Consequently the conclusion (180) follows with $c_2 = \frac{1}{2\mu_k^3 L^2}$.

Substituting $t_1 = 0$ and $t_2 = x_{r,0}L$ into (180) and letting $\varepsilon \in (0, \mu_k L)$, we immediately have (181).

□

Lemma 7.4.

For each $\varepsilon > 0$ and $k \in \mathbb{K}$, there exist constants $\delta > 0$ and $c_3 > 0$ such that

$$\mathsf{P}_*^r\Big(\sup_{0\le t\le x_{r,0}L}\Big|F_k^r(t) - \sum_{l=1}^K P_{lk}^r D_l^r(t)\Big| > x_{r,0}\varepsilon, \max_{p\in\mathbb{K}}|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))| < \delta\Big) < c_3\cdot\frac{\varepsilon}{r} \tag{186}$$

if r is sufficiently large independently of $*$, where $F^r(\cdot)$ is given in (16).

Proof.

$$\begin{aligned}
&\text{(The left-hand side of (186))}\\
&= \mathsf{P}_*^r\Big(\sup_{0\le t\le x_{r,0}L}\Big|\sum_{l=1}^K\sum_{i=1}^{D_l^r(t)}(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > x_{r,0}\varepsilon, \max_{p\in\mathbb{K}}|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))| < \delta\Big)\\
&\le \mathsf{P}_*^r\Big(\max_{l\in\mathbb{K}}\sup_{0\le t\le x_{r,0}L}\Big|\sum_{i=1}^{D_l^r(t)}(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > \frac{x_{r,0}\varepsilon}{K}, \max_{p\in\mathbb{K}}|\widehat{\mathcal{V}}_p^{o,r}(\overline{Z}_p^{+,r}(0))| < \delta\Big)\\
&\le \sum_{l=1}^K \mathsf{P}_*^r\Big(\sup_{0\le t\le x_{r,0}L}\Big|\sum_{i=1}^{D_l^r(t)}(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > \frac{x_{r,0}\varepsilon}{K}, |\widehat{\mathcal{V}}_l^{o,r}(\overline{Z}_l^{+,r}(0))| < \delta\Big). \qquad (187)
\end{aligned}$$

Each term in the summation w.r.t. l in (187) is dominated by

$$\begin{aligned}
&\mathsf{P}_*^r\Big(D_l^r(x_{r,0}L) \ge 2\mu_l x_{r,0}L, |\widehat{\mathcal{V}}_l^{o,r}(\overline{Z}_l^{+,r}(0))| < \delta\Big)\\
&\quad + \mathsf{P}_*^r\Big(\sup_{0\le t\le x_{r,0}L}\Big|\sum_{i=1}^{D_l^r(t)}(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > \frac{x_{r,0}\varepsilon}{K}, D_l^r(x_{r,0}L) < 2\mu_l x_{r,0}L\Big) \qquad (188)
\end{aligned}$$

According to Lemma 7.3,

$$\text{(the first term in (188))} < c_2\frac{\varepsilon}{r}$$

if r is sufficiently large independently of $*$. Furthermore we have that

$$\begin{aligned}
&\text{(the second term in (188))}\\
&\le \mathsf{P}_*^r\Big(\max_{1\le j\le\lfloor 2\mu_l x_{r,0}L\rfloor}\Big|\sum_{i=1}^j(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > \frac{x_{r,0}\varepsilon}{K}\Big)\\
&\le \mathsf{P}_*^r\Big(\max_{1\le j\le\lfloor 2\mu_l x_{r,0}L\rfloor}\Big|\sum_{i=1}^j(\phi_k^{l,r}(i) - P_{lk}^r)\Big| > \lfloor 2\mu_l x_{r,0}L\rfloor\cdot\frac{\varepsilon}{2\mu_l LK}\Big)\\
&\le \frac{\varepsilon}{2\mu_l LK}\cdot\frac{1}{\lfloor 2\mu_l x_{r,0}L\rfloor} \le \frac{\varepsilon}{(\mu_l L)^2 K x_{r,0}} < \frac{\varepsilon}{(\mu_l L)^2 Kr}
\end{aligned}$$

if r is sufficiently large independently of $*$, where the third inequality follows from the application of Lemma 7.1.

Consequently we have the conclusion of the lemma with $c_3 \equiv Kc_2 + (\min_{l\in\mathbb{K}} \mu_l \cdot L)^{-2}$. □

References

[1] Baccelli, F., Boyer, P. and Hebuterne, G.: Single-server queues with impatient customers. *Adv. Appl. Prob.* **16**(1984) 887-905.

[2] Billingsley, P.: *Convergence of Probability Measures*. John Wiley & Sons (1968).

[3] Bramson, M.: State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.* **30**(1998) 89-148.

[4] Bramson, M. and Dai, J.G.: Heavy traffic limits for some queueing networks. *Ann. Appl. Probab.* **11**(2001) 49-90.

[5] Budhiraja, A. and Lee, C.: Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Math. Oper. Res.* **34**(2009) 45-56.

[6] Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.* **5**(1995) 49-77.

[7] Dai, J.G. and He, S.: Customer abandonment in many-server queues. *Math. Oper. Res.* **35**(2010) 347-362.

[8] Dai, J.G. and He, S.: Queues in service systems: customer abandonment and diffusion approximations. *Tutor. Oper. Res.* (2011) 36-59.

[9] Dai, J.G. and Wang, Y.: Nonexistence of Brownian models of certain multiclass queueing networks. *Queueing Syst.* **13**(1993) 41-46.

[10] Davis, M.H.A.: Piecewise deterministic Markov processes: A general class of nondiffusion stochastic models. *J. R. Stat. Soc. Ser. B.* **46**(1984) 353-388.

[11] Ethier, S.N. and Kurtz, T.G. : *Markov Processes*: *Characterization and Convergence*. John Wiley & Sons (1986).

[12] Gamarnik, D. and Zeevi, A.: Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Ann. Appl. Probab.* **16**(2006) 56-90.

[13] Huang, J. and Zhang, H.: Diffusion approximations for open Jackson networks with reneging. *Queueing Syst.* **74**(2013) 445-476.

[14] Jacod, J. and Shiryaev, A.N.: *Limit Theorems for Stochastic Processes*. Springer-Verlag, New York (1987).

[15] Katsuda, T.: State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Syst.* **65**(2010) 237-273.

[16] Katsuda, T.: Stationary distribution convergence for a multiclass single-server queue in heavy traffic. *Sci. Math. Jpn.* **75**(2012) 317-334.

[17] Katsuda, T.: General hazard-type scaling of abandonment time distribution for a G/Ph/n+GI queue in the Halfin-Whitt heavy-traffic regime. *Queueing Syst.* **80**(2015) 155-195.

[18] Katsuda, T.: Stability conditions for a multiclass feedforward queueing network and a generalized Jackson queueing network with abandonments. *Working paper* (2016).

[19] Krichagina, E.V. and Taksar, M.I.: Diffusion approximation for GI/G/1 controlled queues. *Queueing Syst.* **12**(1992) 333-368.

[20] Liptser, R.Sh. and Shiryayev, A.N. :*Theory of Martingales*. Kluwer Academic Publishers (1989).

[21] Reed, J.E. and Ward, A.R.: Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. *Math. Oper. Res.* **33**(2008) 606-644.

[22] Reiman, M.I.: Open queueing networks in heavy traffic. *Math. Oper. Res.* **9**(1984) 441-458.

[23] Ward, A.R. and Glynn, P.W.: A diffusion approximation for a Markovian queue with reneging. *Queueing Syst.* **43**(2003) 103-128.

[24] Ward, A.R. and Glynn, P.W.: A diffusion approximation for a GI/GI/1 queue with balking and reneging. *Queueing Syst.* **50**(2005) 371-400.

[25] Williams, R.J.: An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Syst.* **30**(1998) 5-25.

[26] Williams, R.J.: Diffusion approximation for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst.* **30**(1998) 27-88.

Communicated by *Hiroaki Ishii*

Kwansei Gakuin University
School of Science and Technology
2-1 Gakuen, Sanda, Hyogo 669-1337 JAPAN.
toshikatsuda@kwansei.ac.jp