# STATIONARY DISTRIBUTION CONVERGENCE FOR A MULTICLASS SINGLE-SERVER QUEUE IN HEAVY TRAFFIC

Toshiyuki Katsuda *

Received June 8, 2012

Abstract. In the author's recent paper [11], it is established that in a multiclass single-server queue with regular service disciplines, the stationary distributions and their moments for workload as well as for queue length can be approximated by appropriate exponential distributions and their moments in the heavy-traffic regime.

In this work, relaxing the assumption of moment generating function on the primitives in that paper to the moment condition of second-order or higher-order, we obtain the corresponding approximation result in a multiclass single-server queue. The key to our analysis is to use the framework of Budhiraja and Lee [4] in which under such weak moment assumption, the tightness of stationary scaled queue length for *single-class* (i.e., generalized Jackson) queueing networks is established for their stationary heavy-traffic analysis. For a *multiclass* single-server queue, we obtain the tightness of stationary scaled workload to show that state-space collapse occurs in the heavy-traffic regime in stationarity, from which the desired approximation result follows.

## 1 Introduction.

In the author's recent paper [11], it has been established that under appropriate conditions, the stationary distributions of suitably scaled workload processes as well as queue length processes in a multiclass single-server queue converge in the heavy-traffic regime to exponential distributions, which are stationary distributions of one-dimensional reflected Brownian motions with negative drift coefficients. In the limit of stationary scaled queue length process, the rate of the exponential distribution depends on the service discipline investigated, i.e., first-in-first-out (FIFO) discipline, generalized head-of-the-line proportional processor sharing (GHLPPS) discipline, or static buffer priority (SBP) discipline. Furthermore, in [11], the stationary moment convergence has been proved for the scaled workload with first-order as well as for the scaled queue length with any order. The key to their proof is to show that state-space collapse occurs in the heavy-traffic regime in stationarity under the assumption of tightness on the stationary workload (or stationary queue length) in general multiclass queueing networks. In a multiclass single-server queue, such tightness condition is confirmed using the Lyapunov function method formulated by Gamarnik and Zeevi [10], in which the heavy-traffic approximation for stationary distribution in a generalized Jackson queueing network (GJN) is validated. In their work, also in [11], in order to derive the sought tightness of scaled queue length, the following assumption of the finite moment generating function (m.g.f.) in a neighborhood of origin is imposed on the residual interarrival and service times:

There exists a constant $\vartheta_0 > 0$ such that

$$\sup_{z \in \mathcal{R}_+} \mathsf{E}\big[\exp(\vartheta_0(\xi - z)) \mid \xi > z\big] < \infty, \tag{1}$$

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

*Key words and phrases.* heavy traffic approximation, stationary distribution, multiclass single-server queue, state space collapse.

where $\xi$ denotes the interarrival or service time.

However, this m.g.f. condition may be regarded as relatively restrictive from the standpoint of the heavy-traffic approximations for queueing systems in which the second-order moment condition is conventionally imposed on their primitives.

On the other hand, the recent work of Budhiraja and Lee [4] studies the same approximation problem of GJNs as in [10], with the relaxation of the above finite m.g.f. condition to weaker moment condition, i.e., the finite $p$-th order moment condition with some $p \in [2, \infty)$, on the primitive variables and the associated renewal processes. In particular, for the second-order moment condition in [4], one has only to impose

$$\mathsf{E}[\,\xi^2\,] < \infty \tag{2}$$

with $\xi$ denoting the interarrival or service time. In [4], uniform (in time and the scaling parameter) moment bounds of the underlying Markov state process are obtained to yield moment bounds for stationary distribution in the GJN, in virtue of the more general framework of uniform moment bounds for Markov processes. (Cf. Theorems 3.2-3.5 in [4].)

In this work, we make such relaxation of moment condition in establishing the heavy-traffic approximation for the stationary distribution of a multiclass single-server queue, which has already been proved under the above m.g.f. condition in [11]. More specifically, under the second-order moment condition, i.e., (2), we have the stationary *distribution* convergence for workload as well as for queue length, in the heavy-traffic regime, and under the moment condition with higher-order than the second, we have the stationary *moment* convergence for workload with the first-order and those for queue lengths with the corresponding order. In our analysis, the framework of [4] is employed to obtain moment bounds of stationary queue length, which yields the tightness of stationary queue length in our queue. In virtue of that tightness, we establish the condition of state-space collapse in stationarity, from which the desired approximation result follows. However, in the derivation of such moment bounds, different from [4], we consider the Markov process

$$X(t) = (Z(t), \mathcal{E}^a(t), \mathcal{E}^v(t), O(t)), \qquad t \geq 0, \tag{3}$$

where $Z(\cdot)$ denotes the queue length process, $\mathcal{E}^a(\cdot)$ the elapsed interarrival time process, $\mathcal{E}^v(\cdot)$ the elapsed service time process, and $O(\cdot)$ the process indicating the order of customer classes in the queue. (Note that in [4], the residual time processes are taken instead of the elapsed time processes.) Our choice of process (3) makes the analysis simpler than that in the corresponding part of [4] because of the relation between the delayed renewal process and the zero-delayed one under the conditional probability given $X(0)$.

The rest of the paper is organized as follows. In §2.1 and §2.2, we recall the standard formulation of a multiclass single-server queue and in §2.3, we construct a sequence of multiclass single-server queues satisfying the heavy-traffic condition, under the scaling parameter regime. In §3, referring to the author's recent work [11], we recall the result on state-space collapse in stationarity and its application to the stationary heavy-traffic analysis of a multiclass single-server queue. In §4, under the second-order moment assumption or the higher-order one, we derive moment stability estimate of the Markov state process with the corresponding order in a multiclass single-server queue. Using that estimate and also making use of the framework of Budhiraja and Lee [4], we have the finiteness of uniform stationary moment for the scaled queue length in a multiclass single-server queue. In §5, applying the result of §4 to §3, we establish the heavy-traffic convergence for stationary distribution as well as for stationary moment in a multiclass single-server queue, under our

assumption. In §6, the appendices, we summerize the Markovian representation of a multiclass single-server queue and also uniform moment bounds of stationary distribution for general Markov processes, which are cited from [11] and [4], respectively.

The following notation will be used. Let $\mathcal{R}^d$ be the $d$-dimensional Euclidean space and for $x = (x_1, \cdots, x_d) \in \mathcal{R}^d$, the norm $|x|$ is defined by $|x| = \sum_{i=1}^d |x_i|$. For a matrix or a vector, the prime is put to denote its transpose. The symbol $\mathcal{N}$ denotes the set of natural numbers and $\mathcal{N}_0$ the set of nonnegative integers.

## 2 Model primitives and assumptions.

2.1 *A multiclass single-server queue.*

We begin with a description of a multiclass single-server queue with feedback class routing. Each customer, belonging to one of $K$ classes, receives service from a single server in the queue with unlimited waiting capacity. For convenience, let $\mathcal{K} = \{1, 2, \cdots, K\}$. Customers of classes in $\mathcal{A}$, a subset of $\mathcal{K}$, arrive at the queue from outside, and no external arrival is allowed for customers of any class in $\mathcal{K} - \mathcal{A}$. Upon service completion, a customer changes its class to another one in a feedback way, or leaves the queue. We assume that the server is never idle whenever there are customers being served or in waiting line, which is called the *non-idling* policy. The service disciplines we investigate are all *head-of-the-line* (HL), which means that customers in each class are served in the order of their arrivals at the class and only the leading customer in each class can receive service at any time. In this work, three instances of HL service disciplines are investigated in association with a multiclass single-server queue, i.e., *first-in-first-out* (FIFO) discipline, *generalized head-of-the-line proportional processor sharing* (GHLPPS) discipline, and *static buffer priority* (SBP) discipline. (Cf. Bramson [1] and Bramson and Dai [2].)

2.2 *Model primitives.*

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space. Unless specified otherwise, all the random variables considered in this work are defined on this probability space.

For each $k \in \mathcal{A}$, the external interarrival times $\{a_k(i), i = 2, 3, \cdots\}$ are i.i.d. (i.e., independent and identically distributed) positive random variables with mean $1/\alpha_k$ ($\alpha_k > 0$) and finite variance $a_k \geq 0$ where $a_k(i)$ denotes the time between the $(i-1)$-st and the $i$-th external arrival of a class $k$ customer. The term $a_k(1)$ denotes the residual interarrival time equal to the time measured from origin until the first external arrival.
For $k \in \mathcal{A}$, we set

$$E_k(t) = \max\{l \in \mathcal{N} : \sum_{i=1}^{l} a_k(i) \leq t\}$$

where $\max \phi \equiv 0$. For convenience, for each $k \in \mathcal{K} - \mathcal{A}$, we set

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

For each $l \in \mathcal{K}$, the service times $\{v_l(i) : i = 2, 3, \cdots\}$ are i.i.d. positive random variables with mean $m_l = 1/\mu_l > 0$ and finite variance $b_l \geq 0$. The term $v_l(1)$ denotes the residual service time of the class $l$ customer initially served. (We set $v_l(1) = 0$ if and only if the number of class $l$ customers is zero.) The cumulative service time process $\mathcal{V}(n) = (\mathcal{V}_l(n_l) : l \in \mathcal{K})$, $n = (n_1, \cdots, n_K)$, $n_l \in \mathcal{N}_0$, $l \in \mathcal{K}$, is defined by

$$\mathcal{V}_l(n_l) = \sum_{i=1}^{n_l} v_l(i)$$

where $\mathcal{V}_l(0) \equiv 0$, $l \in \mathcal{K}$. For each $l \in \mathcal{K}$ and each $t \geq 0$, let $S_l(t)$ denote the number of service completions at class $l$ during the busy time interval $[0, t]$. That is,

$$S_l(t) = \begin{cases} \max\{n \in \mathcal{N} : \mathcal{V}_l(n) \leq t\} & \text{if } v_l(1) > 0, \\ \max\{n \in \mathcal{N} : \mathcal{V}_l(n) \leq t\} - 1 & \text{if } v_l(1) = 0. \end{cases}$$

The class routing vectors $\{\phi^k(i) : i = 1, 2, \cdots\}$, $k \in \mathcal{K}$, are i.i.d. $K$-dimensional vectors where $\phi^k(i)$ takes values in the set $\{0, e_1, \cdots, e_K\}$, with $e_k$ denoting the $k$-th unit vector in $\mathcal{R}^K$, $k \in \mathcal{K}$. The non-zero component of $\phi^k(i)$ indicates the class to which the $i$-th customer served at class $k$ is routed, and $\phi^k(i) = 0$ indicates that it leaves the queue. Let $P_{kl} = \mathsf{P}\{\phi^k(i) = e_l\}$ and $P_{k0} = \mathsf{P}\{\phi^k(i) = 0\}$, $k, l \in \mathcal{K}$. The $K \times K$ matrix $P = \big[P_{kl}\big]$, called the class routing matrix, is assumed to have spectral radius strictly less than unity. Thus

$$\begin{aligned} Q &\equiv (I - P')^{-1} \\ &= I + P' + (P')^2 + \cdots \end{aligned}$$

is finite. The mean vector and covariance matrix of $\phi^k(1)$, $k \in \mathcal{K}$, are given by

$$\mathsf{E}[\phi^k(1)] = (P')_{\cdot k} \quad \text{and} \quad \mathsf{Cov}[\phi^k(1))] = \Upsilon^k$$

respectively, where $(P')_{\cdot k}$ denotes the $k$-th column of $P'$ and $\Upsilon^k$ is the $K \times K$ matrix with

$$\Upsilon^k_{lm} \equiv \begin{cases} P_{kl}(1 - P_{kl}) & \text{if } l = m, \\ -P_{kl}P_{km} & \text{if } l \neq m. \end{cases}$$

The *cumulative routing process* for class $k$ is given by

$$\Phi^k(n) = \sum_{i=1}^{n} \phi^k(i), \quad n \in \mathcal{N}.$$

We define $\lambda$ to be the unique $K$-dimensional vector solution of the *traffic equation*:

$$\lambda = \alpha + P'\lambda$$

where $\alpha = (\alpha_1, \cdots, \alpha_K)'$,
i.e.,

$$\lambda = Q\alpha.$$

Using $\lambda$ and $m \equiv (m_1, \cdots, m_K)'$, we define the *traffic intensity* $\rho$ as

$$\rho = \sum_{k \in \mathcal{K}} m_k \lambda_k,$$

or equivalently, $\rho = e'M\lambda$ where $e$ is the $K$-dimensional vector of all 1's and $M = diag(m)$.

For each $k \in \mathcal{K}$, let $Z_k(t)$ denote the number of class $k$ customers being in waiting line or served at the queue at time $t$. The $K$-dimensional process

$$Z(t) = (Z_1(t), \cdots, Z_K(t))', \quad t \geq 0,$$

is referred to as the *queue length* process. Also let $W(t)$ denote the total amount of immediate work (measured in units of service time) for the server at time $t$, and

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

is referred to as the *workload* process. In addition, let $Y(t)$ denote the cumulative amount of time that the server is idle during the time interval $[0, t]$. The process

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

is referred to as the *cumulative idle time* process. Furthermore, the following $K$-dimensional processes are defined to describe the dynamics of the queue, i.e., $A(t) = (A_k(t) : k \in \mathcal{K}, t \geq 0)$, $D(t) = (D_k(t) : k \in \mathcal{K}, t \geq 0)$ and $T(t) = (T_k(t) : k \in \mathcal{K}, t \geq 0)$ where for each $k \in \mathcal{K}$, $A_k(t)$ denotes the total number of arrivals of class $k$ customers during $[0, t]$, $D_k(t)$ denotes the total number of service completions (departures) of class $k$ customers during $[0, t]$ and $T_k(t)$ denotes the total amount of time that the server has served customers of class $k$ by time $t$. Set $\mathfrak{X}(\cdot) \equiv (A(\cdot), D(\cdot), T(\cdot), W(\cdot), Y(\cdot), Z(\cdot))$. With a slight abuse of notation, our multiclass single-server queue is symbolized by $\mathfrak{X}(\cdot)$. Under the non-idling policy, we have the following equations:

$$\begin{aligned}
A(t) &= E(t) + \sum_{l=1}^{K} \Phi^l(D_l(t)), \\
Z(t) &= Z(0) + A(t) - D(t), \\
W(t) &= e'\mathcal{V}(Z(0) + A(t)) - e'T(t), \\
e'T(t) + Y(t) &= et, \\
\int_0^\infty W(s)dY(s) &= 0,
\end{aligned}$$

for all $t \geq 0$.

As is well known, in order to describe by use of some Markov process the dynamics of related queueing system in which the interarrival and service times are i.i.d. with general distributions, one has to add to the queue length process two elapsed time processes, i.e., the elapsed interarrival time process $\mathcal{E}^a(\cdot)$ and the elapsed service time process $\mathcal{E}^v(\cdot)$. In our multiclass single-server queue case, the processes $\mathcal{E}^a_k(t), k \in \mathcal{A}$, and $\mathcal{E}^v_l(t), l \in \mathcal{K}$, measure the elapsed time since the most recent exogeneous arrival and most recent service completion prior to time $t$, respectively, at the corresponding customer class. However, in such Markovian representation of *multiclass* queueing systems under the service disciplines such as the FIFO one, one has to add one more component to the process $(Z(\cdot), \mathcal{E}^a(\cdot), \mathcal{E}^v(\cdot))$, which specifies the ordering of different classes for the customers processed by each server. That is, in the case of our multiclass single-server queue, we introduce the process

$$O(t) \equiv (O_1(t), O_2(t), \cdots)', \quad t \geq 0,$$

with $O_i(t)$, $1 \leq i \leq \sum_{k\in\mathcal{K}} Z_k(t)$, designating the class of the $i$-th customer at time $t$, and $O_i(t) \equiv 0$ for $i \geq \sum_{k\in\mathcal{K}} Z_k(t) + 1$. For example, $O_1(t)$ designates the class of the customer that arrived at the queue the longest time ago of the customers staying there at time $t$, etc. In §4, we consider the Markov process

$$X(t) \equiv (Z(t), \mathcal{E}^a(t), \mathcal{E}^v(t), O(t)), \quad t \geq 0,$$

in order to show the desired tightness of stationary scaled workload.

2.3 *Heavy-traffic model of a multiclass single-server queue.*

In this subsection, we introduce a sequence of multiclass single-server queues, denoted by $\{\mathfrak{X}^n(\cdot)\}_{n=1}^{\infty}$, satisfying the heavy-traffic condition. First, consider a multiclass single-server queue $\mathfrak{X}(\cdot)$ with traffic intensity equal to unity. To this queue, the interarrival times, the service times and the class routing vectors are introduced as in §2.2. We impose the following assumptions on the primitive variables and the associated renewal processes in $\mathfrak{X}$. These are assumed throughout this paper even when not mentioned.

(A1) For each $k \in \mathcal{A}$, $\{a_k(i) : i \geq 2\}$ are unbounded and spread out. That is, there exist some integer $j_k > 0$ and some function $p_k(x) \geq 0$, $x \in \mathcal{R}_+$, with $\int_0^\infty p_k(x)dx > 0$, such that

$$\mathsf{P}(a_k(2) \geq x) > 0 \qquad \text{for any } x > 0$$

and

$$\mathsf{P}\Big( c_1 \leq \sum_{i=2}^{j_k} a_k(i) \leq c_2 \Big) \geq \int_{c_1}^{c_2} p_k(x)dx$$

for any $0 \leq c_1 < c_2$.

(A2) For each $l \in \mathcal{K}$,

$$\sup_{z \geq 0} \mathsf{E}\big[v_l(2) - z \mid v_l(2) > z\big] < \infty.$$

The next assumption (A3-$p$) is employed in Budhiraja and Lee [4] to show the tightness of queue length in a GJN for the purpose of its stationary heavy-traffic analysis. We also use the condition, which plays a key role in the derivation of the tightness of queue length in a multiclass single-server queue.

(A3-$p$-(i)) There exists a constant $p \in [2, \infty)$ such that

$$\mathsf{E}\big[a_k(2)^p + v_l(2)^p\big] < \infty, \qquad \forall k \in \mathcal{A}, \forall l \in \mathcal{K}.$$

(A3-$p$-(ii)) For the constant $p \in [2, \infty)$ in (A3-$p$-(i)), there exists a constant $c_p > 0$ such that for any $t \geq 0$,

$$\mathsf{E}\big[\sup_{0 \leq s \leq t} |E_{k,0}(s) - \alpha_k s|^p\big] \leq c_p(1 + t^{\frac{p}{2}}), \qquad \forall k \in \mathcal{A},$$

$$\mathsf{E}\big[\sup_{0 \leq s \leq t} |S_{l,0}(s) - \mu_l s|^p\big] \leq c_p(1 + t^{\frac{p}{2}}), \qquad \forall l \in \mathcal{K},$$

$$\mathsf{E}\big[\sup_{0 \leq s \leq t} |\Phi_m^l(S_{l,0}(s)) - P_{lm} S_{l,0}(s)|^p\big] \leq c_p(1 + t^{\frac{p}{2}}), \qquad \forall l, m \in \mathcal{K},$$

where $E_{k,0}(\cdot)$ and $S_{l,0}(\cdot)$ denote zero-delayed renewal processes corresponding to $\{a_k(i), i \geq 2\}$ and $\{v_l(i), i \geq 2\}$, respectively, $k \in \mathcal{A}$, $l \in \mathcal{K}$.

In particular, when $p = 2$, condition (A3-$p$-(i)), which has already been assumed in §2.2, implies (A3-$p$-(ii)). (Cf. Budhiraja and Ghosh [3], and Krichagina [12].)

(A4) The following families of random variables,

$$\{Z(0), R^v(0), O(0)\},\ R_k^a(0), k \in \mathcal{A},\ a_{k'}, k' \in \mathcal{A},$$
$$v_1, \cdots, v_K, \phi^1, \cdots, \phi^K$$

are mutually independent, where $a_k \equiv \{a_k(i) : i \geq 2\}$, $k \in \mathcal{A}$, $v_l \equiv \{v_l(i) : i \geq 2\}$, $l \in \mathcal{K}$, and $\phi^j \equiv \{\phi^j(i) : i \geq 1\}$, $j \in \mathcal{K}$.

Using this independence assumption and the *i.i.d.* property of primitive variables, one can describe the dynamics of a multiclass single-server queue in virtue of some Markov process in Appendix 1.

Based on the queue $\mathfrak{X}$, we provide a sequence of multiclass single-server queues $\mathfrak{X}^n(\cdot), n \in \mathcal{N}$, in which

(i) the interarrival times $\{a_k^n(i) : i \geq 2, k \in \mathcal{A}\}$ for the $n$-th queue in the sequence are defined as

$$a_k^n(i) = a_k(i)\Big(1 - \frac{\kappa_k^0}{\sqrt{n}}\Big)^{-1}, \qquad i \geq 2, k \in \mathcal{A},$$

where $\{a_k(i) : i \geq 2, k \in \mathcal{A}\}$ are the interarrival times in $\mathfrak{X}(\cdot)$ and $\kappa_k^0 (> 0)$, $k \in \mathcal{A}$, is a constant,

(ii) the service times $\{v_l(i) : i \geq 2, l \in \mathcal{K}\}$ and the class routing vectors $\{\phi^l(i) : i \geq 1, l \in \mathcal{K}\}$ for the $n$-th queue are those for $\mathfrak{X}(\cdot)$. (That is, they are independent of the sequence parameter $n$.)

Then the vector of arrival rates for the $n$-th queue is given by

$$\alpha^n = (\alpha_1^n, \cdots, \alpha_K^n)'$$

where

$$\begin{aligned} \alpha_k^n &= \alpha_k\Big(1 - \frac{\kappa_k^0}{\sqrt{n}}\Big), \quad k \in \mathcal{A}, \\ &= 0, \qquad\qquad k \in \mathcal{K} - \mathcal{A}, \end{aligned}$$

with $\alpha_k = 1/\mathsf{E}(a_k(2))$, $k \in \mathcal{A}$. Thus, for the $n$-th queue, the vector solution to the traffic equation is given by

$$\lambda^n = (I - P')^{-1}\alpha^n$$

and the traffic intensity is also given by

$$\begin{aligned} \rho^n &= e'M\lambda^n \\ &= 1 - \frac{\kappa}{\sqrt{n}} \end{aligned} \tag{4}$$

where $\kappa = e'M(I - P')^{-1}K^0\alpha > 0$ with $M = diag(m)$ and $K^0 = diag(\kappa^0)$. For each $n \in \mathcal{N}$, since $\rho^n < 1$, the $n$-th multiclass single-server queue $\mathfrak{X}^n$ is fluid-stable under any non-idling service discipline. Therefore, under assumption (A1), the underlying Markov process for $\mathfrak{X}^n$ has the unique stationary distribution, which is also identical to the steady-state distribution. (Cf. Dai[7], Dai and Meyn [8], and Katsuda [11].)

*Scaling.*

For the sequences of queue lengths, workloads, cumulative idle time processes, and the renewal processes associated with the primitive triples, we scale them as follows in order to

obtain their proper limits as the sequence parameter $n$ tends to infinity. For any $n \in \mathcal{N}$ and any $t \geq 0$,

$$\begin{aligned}
\widehat{Z}^n(t) &\equiv \frac{1}{\sqrt{n}} Z^n(nt), \\
\widehat{W}^n(t) &\equiv \frac{1}{\sqrt{n}} W^n(nt), \\
\widehat{Y}^n(t) &\equiv \frac{1}{\sqrt{n}} Y^n(nt), \\
\widehat{E}^n(t) &\equiv \frac{1}{\sqrt{n}} (E^n(nt) - \alpha^n nt), \\
\widehat{S}^n(t) &\equiv \frac{1}{\sqrt{n}} (S(nt) - \mu nt), \\
\widehat{\Phi}^n(t) &\equiv \frac{1}{\sqrt{n}} \{\Phi([nt]) - (P)'[nt]\}
\end{aligned}$$

where the processes defined in association with $\mathfrak{X}^n$ are indexed by the sequence parameter $n$, $n \in \mathcal{N}$, and $[nt]$ denotes the integer part of $nt$.

## 3 State-space collapse in stationarity and heavy-traffic approximation for stationary distribution.

As established in Williams [13] and Bramson and Dai [2], state-space collapse is the main sufficient condition under which heavy-traffic stochastic-process limit theorem with general initial condition holds for some important multiclass queueing networks, including our multiclass single-server queues under the FIFO, GHLPPS and SBP service disciplines. It is also the necessary condition for such queueing networks under the FIFO discipline.

In this work, we are concerned with heavy-traffic approximation for the stationary distribution in a multiclass single-server queue, which is often symbolized by the interchange of limits in distributions:

$$\lim_{n\to\infty} \lim_{t\to\infty} \widehat{W}^n(t) = \lim_{t\to\infty} \lim_{n\to\infty} \widehat{W}^n(t).$$

For this purpose, we apply the heavy-traffic stochastic-process limit theorem to our sequence of multiclass single-server queues, $\{\mathfrak{X}^n\}$, in stationarity. In this application, as mentioned above, we have to see that state-space collapse occurs for $\{\mathfrak{X}^n\}$ in stationarity, which has been proved under the assumption of the tightness of stationary workload. (Note that assumption (A2) is used in this proof. See Proposition 3.2, its proof and Remark 3.1 in Katsuda [11].) As a result, we have the following proposition, i.e., Proposition 3.1 below, on the heavy-traffic approximation for the stationary distribution in our multiclass single-server queue, as established in a more general context in [11].

State-space collapse is said to hold for our sequence of multiclass single-server queues, $\{\mathfrak{X}^n\}$, if for each $T > 0$ and each $k \in \mathcal{K}$,

$$\sup_{0\leq t\leq T} \mid \widehat{Z}_k^n(t) - \delta_k \widehat{W}^n(t) \mid \longrightarrow 0 \quad \text{in probability as } n \to \infty,$$

where the constant $\delta_k, k \in \mathcal{K}$, is given according to the associated service discipline:

For the FIFO discipline,

$$\delta_k = \lambda_k, \qquad k \in \mathcal{K}, \tag{5}$$

and for the SBP discipline,

$$\delta_k = \begin{cases} 1/m_k & \text{if } k \text{ is the lowest priority class,} \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

for each $k \in \mathcal{K}$.
For the GHLPPS discipline,

$$\delta_k = \frac{\lambda_k m_k/\beta_k}{\sum_{l\in\mathcal{K}} \lambda_l m_l^2/\beta_l}, \qquad k \in \mathcal{K}, \tag{7}$$

where the nonnegative constant $\beta_k, k \in \mathcal{K}$, denotes the weight coefficient for the discipline, i.e., for each $t \geq 0$ and each $k \in \mathcal{K}$,

$$\dot{T}_k(t) = \begin{cases} \frac{\beta_k Z_k(t)}{\sum_{l\in\mathcal{K}} \beta_l Z_l(t)} & \text{if } \sum_{l\in\mathcal{K}} \beta_l Z_l(t) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

(Cf. Bramson [1] and Bramson and Dai [2].)

Let $\Xi^n(t), t \geq 0$, denote the Markovian description process for our multiclass single-server queue $\mathfrak{X}^n$, $n \in \mathcal{N}$, whose definition is summerized in Appendix 1. Also denote by $\mathsf{S}$ the state space of $\Xi^n(\cdot), n \in \mathcal{N}$. According to §3 in [11], under assumption (A1), we see that the stationary distribution of $\Xi^n(\cdot)$ exists uniquely for each $n \in \mathcal{N}$, because of (4). In taking the heavy-traffic limit, we scale the process $\Xi^n(t)$ as

$$\widehat{\Xi}^n(t) = \frac{1}{\sqrt{n}} \Xi^n(nt), \quad t \geq 0, n \in \mathcal{N}.$$

For each fixed $n \in \mathcal{N}$, let $\mathsf{P}_\xi(\,\cdot\,)$ denote the probability law of $\{\widehat{\Xi}^n(t) : t \geq 0\}$ such that $\mathsf{P}_\xi(\widehat{\Xi}^n(0) = \xi) = 1, \xi \in \mathsf{S}$, and let $\pi^n$ denote the stationary distribution of $\widehat{\Xi}^n(\cdot)$. In addition, let $\mathsf{P}_{\pi^n}(\,\cdot\,)$ denote the probability law of $\widehat{\Xi}^n(\cdot)$ with initial measure $\pi^n$ and let $\mathsf{E}_{\pi^n}[\,\cdot\,]$ denote the expectation w.r.t. $\mathsf{P}_{\pi^n}(\,\cdot\,), n \in \mathcal{N}$.

Let $W(t), t \geq 0$, denote the heavy-traffic stochastic-process limit of $\{\widehat{W}^n(t), t \geq 0\}_{n=1}^{\infty}$, and also let $\pi$ denote the stationary distribution of $W(\cdot)$. More specifically, since $W(\cdot)$ is the one-dimensional reflected Brownian motion with negative drift coefficient in our single-server case, $\pi$ is the exponential distribution with rate $2R\kappa/\Gamma$ where

$$R = (1 + G)^{-1}, \tag{8}$$

$$\Gamma = RHR' \tag{9}$$

with

$$\begin{aligned} G &= e'MQP'\delta, \\ H &= e'\{\Lambda\Sigma + MQ(\Pi + \sum_{k\in\mathcal{K}} \lambda_k \Upsilon^k)Q'M\}e, \\ \delta &= (\delta_1, \cdots, \delta_K)', \\ \Lambda &= diag(\lambda), \\ \Sigma &= diag(b_1, \cdots, b_K), \\ \Pi &= diag(\alpha_1^3 a_1, \cdots, \alpha_K^3 a_K). \end{aligned}$$

(Cf. Williams [13].)

**Proposition 3.1.**
*For the sequence of multiclass single-server queues, $\{\mathfrak{X}^n\}_{n=1}^{\infty}$, under each of the FIFO, GHLPPS and SBP service disciplines, suppose that*

$$\{P_{\pi^n}(\widehat{W}^n(0) \in *)\}_{n=1}^{\infty}$$

*is tight in $\mathcal{R}^1$. Then, under the assumptions imposed so far,*

$$P_{\pi^n}(\widehat{W}^n(0) \in *) \xrightarrow{D} \pi(*)$$

*as $n \to \infty$.*

## 4 Uniform moment bounds for the stationary queue length in a multiclass single-server queue.

In this section, we establish uniform moment bounds for the stationary scaled queue length in our multiclass single-server queue in §2, which yields the sought tightness of stationary scaled workload for the heavy-traffic approximation of the stationary distribution in the queue.

Set

$$X^n(t) \equiv (Z^n(t), \mathcal{E}^{a,n}(t), \mathcal{E}^{v,n}(t), O^n(t)), \quad t \geq 0, n \in \mathcal{N}.$$

Then $X^n(\cdot), n \in \mathcal{N}$, is a piecewise-deterministic Markov process. It is even a strong Markov process. Denote the state space of $X^n(\cdot)$, $n \in \mathcal{N}$, by $\mathsf{X}$. Similarly as in the case of $\Xi^n(\cdot)$, we scale the process by $\widehat{X}^n(t) \equiv \frac{1}{\sqrt{n}} X^n(nt), t \geq 0, n \in \mathcal{N}$. For each fixed $n \in \mathcal{N}$, let $\mathsf{P}_x(\,\cdot\,)$ denote the probability law of $\{\widehat{X}^n(t) : t \geq 0\}$ such that $\mathsf{P}_x(\widehat{X}^n(0) = x) = 1, x = (z, e^a, e^v, o) \in \mathsf{X}$. In addition, $\mathsf{E}_x[\,\cdot\,]$ denotes the expectation w.r.t. $\mathsf{P}_x(\,\cdot\,)$, $x \in \mathsf{X}$.

The next proposition is concerned with the $p$-th order moment stability estimate of $\widehat{X}^n(\cdot)$, uniform in $n$, with parameter $p$ in condition (A3-$p$).

**Proposition 4.1.**
*Under any non-idling service discipline, there exists a constant $t_0 > 0$ such that for any $t \geq t_0$,*

$$\lim_{|x| \to \infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\left[|\widehat{X}^n(|x|t)|^p\right] = 0,$$

*with the constant $p$ in condition $(A3{-}p)$.*

*Proof.* In the same way as in Theorem 3.3 of [4], we can prove that for any $t \geq 0$,

$$\lim_{|x| \to \infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\left[|\widehat{\mathcal{E}}^{a,n}(|x|t)|^p\right] = 0,$$

$$\lim_{|x| \to \infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\left[|\widehat{\mathcal{E}}^{v,n}(|x|t)|^p\right] = 0,$$

in use of Wald's identity and condition (A3-$p$-(i)). Thus it remains to prove that there exists some $t_0 > 0$ such that for any $t \geq t_0$,

$$\lim_{|x|\to\infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\big[|\widehat{Z}^n(|x|t)|^p\big] = 0, \tag{10}$$

because

$$\lim_{|x|\to\infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\big[|\widehat{O}^n(|x|t)|^p\big] = 0$$

follows immediately from (10).

From

$$Z_k^n(t) = Z_k^n(0) + E_k^n(t) + \sum_{l\in\mathcal{K}} \Phi_k^l(S_l(T_l^n(t))) - S_k(T_k^n(t)), \qquad k \in \mathcal{K},$$

we have the following scaled identity:

$$\begin{aligned}\widehat{Z}_k^n(t) =& \widehat{Z}_k^n(0) + \widehat{E}_k^n(t) + \alpha_k^n\sqrt{n}t + \sum_{l\in\mathcal{K}} \widehat{\Phi}_k^{n,l}(\overline{S}_l^n(\overline{T}_l^n(t))) + \sum_{l\in\mathcal{K}} P_{lk}\widehat{S}_l^n(\overline{T}_l^n(t)) \\ &+ \sum_{l\in\mathcal{K}} P_{lk}\mu_l \frac{T_l^n(nt)}{\sqrt{n}} - \widehat{S}_k^n(\overline{T}_k^n(t)) - \mu_k\frac{T_k^n(nt)}{\sqrt{n}}, \qquad k \in \mathcal{K},\end{aligned}$$

or, in vector form,

$$\begin{aligned}\widehat{Z}^n(t) =& \widehat{Z}^n(0) + \widehat{E}^n(t) + \alpha^n\sqrt{n}t + \sum_{l\in\mathcal{K}} \widehat{\Phi}^{n,l}(\overline{S}_l^n(\overline{T}_l^n(t))) - (I - P')\widehat{S}^n(\overline{T}^n(t)) \\ &- (I - P')\frac{(\mu T^n)(nt)}{\sqrt{n}}\end{aligned} \tag{11}$$

where $\overline{S}^n(t) \equiv S(nt)/n$ and $\overline{T}^n(t) \equiv T^n(nt)/n$.

Let

$$\eta \equiv (I - P)^{-1}Me.$$

Then, multiplying (11) by $\eta' = e'M(I-P')^{-1}$ from the left and noting that $\sqrt{n}(\rho^n - 1) = -\kappa$ and

$$Y^n(t) = t - \sum_{k\in\mathcal{K}} T_k^n(t),$$

we have

$$\begin{aligned}\eta'\widehat{Z}^n(t) &\equiv e'M(I - P')^{-1}\widehat{Z}^n(t) \\ &= \eta'\widehat{Z}^n(0) + e'M(I - P')^{-1}\{\widehat{E}^n(t) + \sum_{l\in\mathcal{K}} \widehat{\Phi}^{n,l}(\overline{S}_l^n(\overline{T}_l^n(t)))\} \\ &\quad - e'M\widehat{S}^n(\overline{T}^n(t)) - \kappa t + \widehat{Y}^n(t).\end{aligned} \tag{12}$$

Because

$$\int_0^\infty \eta'\widehat{Z}^n(t)d\widehat{Y}^n(t) = 0,$$

from (12) we obtain

$$\eta'\widehat{Z}^n(t) = \Psi(\chi^n(\cdot))(t) \tag{13}$$

where $\Psi$ is the one-dimensional reflection map, i.e.,

$$\Psi(u)(t) = u(t) + \sup_{0\le s\le t}(-u(s))^+, \qquad u \in D([0,\infty), \mathcal{R}^1), \quad u(0) \ge 0,$$

and

$$\begin{aligned}\chi^n(t) \equiv& \eta' \widehat{Z}^n(0) + e' M (I - P')^{-1} \{\widehat{E}^n(t) + \sum_{l\in\mathcal{K}} \widehat{\Phi}^{n,l}(\overline{S}_l^n(\overline{T}_l^n(t)))\} \\ & - e' M \widehat{S}^n(\overline{T}^n(t)) - \kappa t. \end{aligned} \tag{14}$$

In order to arrive at (10), it is enough to prove that there exists a constant $t_0 > 0$ such that for any $t \ge t_0$,

$$\lim_{|x|\to\infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\big[|\eta' \widehat{Z}^n(|x|t)|^p\big] = 0.$$

Now let

$$N^n(t) = N_1^n(t) + N_2^n(t) + N_3^n(t)$$

with

$$\begin{aligned} N_1^n(t) &= \sum_{k\in\mathcal{A}} \eta_k \frac{1}{\sqrt{n}} \big\{E_k^n(nt) - \alpha_k^n(nt + \sqrt{n} e_k^a)\big\}, \\ N_2^n(t) &= -\sum_{k\in\mathcal{K}} m_k \frac{1}{\sqrt{n}} \big\{S_k(T_k^n(nt)) - \mu_k(T_k^n(nt) + \sqrt{n} e_k^v)\big\} \end{aligned}$$

and

$$N_3^n(t) = \sum_{k\in\mathcal{K}} \eta_k \frac{1}{\sqrt{n}} \sum_{l\in\mathcal{K}} \big\{\Phi_k^l(S_l(T_l^n(nt))) - P_{lk} S_l(T_l^n(nt))\big\},$$

and

$$b^n(t) = \sum_{k\in\mathcal{A}} \eta_k \alpha_k^n e_k^a - \sum_{k\in\mathcal{K}} e_k^v - \kappa t.$$

Also let

$$F^n(t) = \Psi(\eta' \widehat{Z}^n(0) + b^n(\cdot))(t).$$

Then

$$\chi^n(t) = \eta' \widehat{Z}^n(0) + N^n(t) + b^n(t)$$

and

$$\begin{aligned} &\frac{1}{|x|^p} \mathsf{E}_x\big[|\eta' \widehat{Z}^n(|x|t)|^p\big] \\ &\le C_p \frac{1}{|x|^p} \mathsf{E}_x\big[\ |\ \eta' \widehat{Z}^n(|x|t) - F^n(|x|t)\ |^p\ \big] + C_p \frac{1}{|x|^p} \mathsf{E}_x\big[F^n(|x|t)^p\big] \\ &= C_p \frac{1}{|x|^p} \mathsf{E}_x\big[\ |\ \Psi(\chi^n(\cdot))(|x|t) - \Psi(\eta' \widehat{Z}^n(0) + b^n(\cdot))(|x|t)\ |^p\ \big] + C_p \frac{1}{|x|^p} \mathsf{E}_x\big[F^n(|x|t)^p\big] \\ &\le C_p' \frac{1}{|x|^p} \mathsf{E}_x\big[\sup_{0\le s\le |x|t} |N^n(s)|^p\big] + C_p \frac{1}{|x|^p} \mathsf{E}_x\big[F^n(|x|t)^p\big] \end{aligned} \tag{15}$$

where $C_p$, etc., are constants depending on $p$, and the last inequality follows from the Lipschitz continuity of reflection map $\Psi$.

In virtue of assumption (A3-$p$-(ii)), we see that the first term in the right-hand side of (15) is majorized by

$$C_p'' \frac{1}{|x|^p} \mathsf{E}_x\Big[\sup_{0\le s\le |x|t} \big(|N_1^n(s)|^p + |N_2^n(s)|^p + |N_3^n(s)|^p\big)\Big]$$
$$\le C_p''' \frac{3}{n^{\frac{p}{2}}|x|^p} c_p\{1 + (n|x|(t+1))^{\frac{p}{2}}\}$$

for each $n \in \mathcal{N}$, noting that for any $x = (z, e^a, e^v, o) \in \mathsf{X}$, $\mathsf{P}_x$-a.e.,

$$E_k^n(t) = E_{k,0}^n(t + \sqrt{n}e_k^a), \quad k \in \mathcal{A},$$
$$S_l(t) = S_{l,0}(t + \sqrt{n}e_l^v), \quad l \in \mathcal{K}.$$

Thus we have

$$\lim_{|x|\to\infty} \sup_n \frac{1}{|x|^p} \mathsf{E}_x\Big[\sup_{0\le s\le |x|t} |N^n(s)|^p\Big] = 0$$

for each $t \ge 0$.

To arrive at (10), we evaluate the second term in the righthand side in (15). According to the scaling and shift properties of reflection map $\Psi$ (cf. Chap.7 of [5].),

$$\frac{1}{|x|}F^n\big(|x|t\big) = \Psi\Big(\frac{1}{|x|}F^n\big(|x|\tilde{t}\big) + \frac{1}{|x|}b^n(|x|(\tilde{t}+\cdot)) - \frac{1}{|x|}b^n(|x|\tilde{t})\Big)(t-\tilde{t})$$

for any $t \ge \tilde{t} \ge 0$. Since

$$\frac{1}{|x|}b^n\big(|x|t\big) = \sum_{k\in\mathcal{A}} \eta_k \alpha_k^n \frac{e_k^a}{|x|} - \sum_{l\in\mathcal{K}} \frac{e_l^v}{|x|} - \kappa t,$$

we have

$$\frac{1}{|x|}F^n\big(|x|t\big) = 0$$

for any $t \ge t_0$ with some constant $t_0$ independent of $x$ and $n$. Consequently the proof is completed.

□

The next proposition corresponds to Theorem 3.4 in Budhiraja and Lee [4]. It can be proved without any difficulty, using Proposition 4.1 and adapting the proof of that theorem to our multiclass single-server queue.

**Proposition 4.2.**

*Under any non-idling service discipline, there exist some constants $c$, $\bar{\delta} \in (0,\infty)$ and a compact set $C \subset \mathsf{X}$ such that for any $x \in \mathsf{X}$,*

$$\sup_n \mathsf{E}_x\Big[\int_0^{\tau_C^n(\bar{\delta})} \big(1 + |\widehat{X}^n(t)|^{p-1}\big)dt\Big] \le c(1+|x|^p)$$

*where $\tau_C^n(\bar{\delta}) = \inf\{t \ge \bar{\delta} : \widehat{X}^n(t) \in C\}$.*

Let $\pi_X^n$ denote the stationary distribution of $\widehat{X}^n(\cdot)$, $n \in \mathcal{N}$. Then, applying Theorems A.2.1 and A.2.2 in Appendix 2 to Proposition 4.2, we have the next uniform stationary moment result for the scaled queue length in a multiclass single-server queue.

**Corollary 4.1.**
*For the parameter $p$ in $(A3{-}p)$, we have*

$$\sup_n \mathsf{E}_{\pi_X^n}\left[|\widehat{Z}^n(0)|^{p-1}\right] < \infty.$$

## 5 Stationary distribution convergence for a multiclass single-server queue in heavy traffic.

In this section, we establish the heavy-traffic convergence for stationary distribution as well as for stationary moment in a multiclass single-server queue, making use of the formulation described so far. The following theorem and corollary extend Theorem 4.1 and Corollaries 4.1-4.3 in Katsuda [11] to the corresponding conclusions under our relaxed moment assumption.

**Theorem 5.1.**
*Suppose that $(A3\text{-}p)$ with $p = 2$ holds in addition to the other conditions in §2. Then, for the sequence $\{\mathfrak{X}^n\}_{n=1}^{\infty}$ of multiclass single-server queues under each of the FIFO, GHLPPS and SBP service disciplines, we have that for any $t \geq 0$ and any $k \in \mathcal{K}$,*

$$\mathsf{P}_{\pi^n}(\widehat{W}^n(0) > t) \longrightarrow \exp(-2R\kappa t/\Gamma), \tag{16}$$

$$\mathsf{P}_{\pi^n}(\widehat{Z}_k^n(0) > t) \longrightarrow \exp(-2R\kappa t/\Gamma\delta_k) \tag{17}$$

*as $n \to \infty$, where $R$, $\Gamma$ and $\delta_k, k \in \mathcal{K}$, are given by $(8), (9)$ and $(5)-(7)$, respectively.*

*Proof.* It is trivial that for each $n \in \mathcal{N}$, $\pi_X^n$ is the marginal distribution of $\pi^n$ in §3. Thus, in use of Corollary 4.1 with $p = 2$, we have

$$\sup_n \mathsf{E}_{\pi^n}\left[|\widehat{Z}^n(0)|\right] < \infty,$$

from which the tightness of

$$\left\{\mathsf{P}_{\pi^n}\left(\widehat{Z}^n(0) \in \cdot\,\right)\right\}_{n=1}^{\infty} \tag{18}$$

in $\mathcal{R}^K$ follows. According to the equality

$$\widehat{W}^n(0) = \sum_{k\in\mathcal{K}} \frac{1}{\sqrt{n}} \sum_{i=1}^{Z_k^n(0)} (v_k(i) - m_k) + \sum_{k\in\mathcal{K}} m_k \widehat{Z}_k^n(0),$$

the tightness of

$$\left\{\mathsf{P}_{\pi^n}\left(\widehat{W}^n(0) \in \cdot\,\right)\right\}_{n=1}^{\infty}$$

in $\mathcal{R}^1$ follows from the tightness of (18). (Cf. The proof of Theorem 4.1 in [11].) Therefore, in virtue of Proposition 3.1, we have the convergence (16). Because of state-space collapse in stationarity, we also have (17).

□

The following corollary, corresponding to Corollaries 4.2 and 4.3 in [11], is concerned with the stationary moment convergence, which depends on the parameter $p$ in assumption (A3-$p$).

**Corollary 5.1.**
*Under our assumptions, in particular, condition (A3-$p$) with some $p \in [2, \infty)$, we have that for any $r \in [0, p-1)$ and $k \in \mathcal{K}$,*

$$\mathsf{E}_{\pi^n}\left[\widehat{Z}_k^n(0)^r\right] \longrightarrow \frac{r!(\Gamma\delta_k)^r}{(2R\kappa)^r}, \tag{19}$$

*as $n \to \infty$. If $p > 2$, then we also have*

$$\mathsf{E}_{\pi^n}\left[\widehat{W}^n(0)\right] \longrightarrow \frac{\Gamma}{2R\kappa}, \tag{20}$$

*as $n \to \infty$.*

*Proof.* According to Proposition 4.2 and Theorem A.2.2, we see that

$$\sup_n \mathsf{E}_{\pi^n}\left[|\widehat{Z}^n(0)|^{p-1}\right] < \infty.$$

Thus, in view of Theorem 4.5.2 in Chung [6], we have the convergence (19). Because $Z_k^n(0)$ and $v_k(i)$, $i \geq 2$, are independent w.r.t. $\mathsf{P}_{\pi^n}(\,\cdot\,)$ for each $n \in \mathcal{N}$ and $k \in \mathcal{K}$ (cf. Lemma 3.1 of [11].), we have

$$\mathsf{E}_{\pi^n}\left[\widehat{W}^n(0)\right] = \sum_{k\in\mathcal{K}} \frac{1}{\sqrt{n}} \mathsf{E}_{\pi^n}\left[v_k(1)\right] + \sum_{k\in\mathcal{K}} m_k \mathsf{E}_{\pi^n}\left[\widehat{Z}_k^n(0) - \frac{1}{\sqrt{n}}\right].$$

Therefore, from (19) and the observation that

$$\sum_{k\in\mathcal{K}} m_k\delta_k = 1$$

for each of our three service disciplines, the convergence (20) follows.

□

## 6 Appendices.

**Appendix 1.** *Markovian description process for a multiclass single-server queue.*

In the author's recent work [11], some Markov process is introduced in association with related multiclass queueing network, in order to describe the dynamics of not only the queue length but also the workload in the network. This process constitutes the fundamental of the stationary regime for the network. We recall its definition in the case of a multiclass single-server queue as follows.

In addition to $Z(\cdot) = (Z_k(\cdot), k \in \mathcal{K})$, $\mathcal{E}^a(\cdot) = (\mathcal{E}_k^a(\cdot), k \in \mathcal{A})$, $\mathcal{E}^v(\cdot) = (\mathcal{E}_l^v(\cdot), l \in \mathcal{K})$ and $O(\cdot) = (O_i(\cdot), i \in \mathcal{N})$ introduced in §2.2, define the following processes:
Let $\mathcal{R}_k^a(t)$, $k \in \mathcal{A}$, denote the remaining interarrival time of class $k$ customer at time $t \geq 0$ and $\mathcal{R}_l^v(t)$, $l \in \mathcal{K}$, denote the remaining service time of class $l$ customer at time $t \geq 0$. (We

set $\mathcal{R}_l^v(t) \equiv 0$ if $Z_l(t) = 0$.) In particular, $\mathcal{R}_k^a(0) = a_k(1)$ and $\mathcal{R}_l^v(0) = v_l(1)$ for each $k \in \mathcal{A}$ and each $l \in \mathcal{K}$.
Also let

$$V(t) \equiv (V_k(t) : k \in \mathcal{K})$$

where

$$V_k(t) \equiv (V_{k1}(t), V_{k2}(t), \cdots)$$

with

$$\begin{aligned} &V_{k1}(t) \equiv \mathcal{R}_k^v(t), \\ &V_{ki}(t) \equiv v_k(D_k(t) + i), \qquad 2 \le i \le Z_k(t), \\ &V_{ki}(t) \equiv 0, \qquad i \ge Z_k(t) + 1, \end{aligned} \tag{21}$$

for each $k \in \mathcal{K}$. (Note that for each $k \in \mathcal{K}$, when $v_k(1) = 0$, we have to reset $V_{ki}(t) \equiv v_k(D_k(t) + i + 1), 2 \le i \le Z_k(t)$, instead of (21).)

Using these processes, we define the description process $\Xi = (\Xi(t) : t \ge 0)$ by

$$\Xi(t) \equiv (Z(t), \mathcal{E}^a(t), \mathcal{R}^a(t), \mathcal{E}^v(t), V(t), O(t)).$$

Then $\Xi = (\Xi(t) : t \ge 0)$ is a *piecewise-deterministic Markov process*. It is even a *strong Markov* process. (Cf. Davis [9].) Let

$$\mathcal{F}_t \equiv \sigma(\Xi(u) : 0 \le u \le t), \qquad t \ge 0.$$

Then, the process

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

is $(\mathcal{F}_t)$-adapted. In this sense, the Markov process $\Xi(\cdot)$ *describes* the dynamics of our multiclass single-server queue. As explained in [11], $\Xi = (\Xi(t) : t \ge 0)$ has the unique stationary distribution under the assumptions in this work.

## Appendix 2. *Uniform bounds for the stationary moments of general strong Markov processes.*

In this appendix, we present two theorems on uniform bounds for the stationary moments of general strong Markov processes, which are used in establishing the tightness of stationary queue length of a generalized Jackson queueing network in Budhiraja and Lee [4]. The details of their proofs are shown in [4].

Let $X^n(t), t \ge 0, n \in \mathcal{N}$, be a general strong Markov process with its state space, denoted by $\mathsf{X}$, the $d$-dimensional Euclidean space. Assume that $X^n(\cdot), n \in \mathcal{N}$, is positive Harris recurrent and denote by $\pi_X^n$ the stationary distribution of $X^n(\cdot), n \in \mathcal{N}$, in this appendix. The symbol $\mathsf{E}_x[\,\cdot\,]$ denote the expectation w.r.t. the probability law $\mathsf{P}_x(\,\cdot\,)$ of $X^n(\cdot)$ such that $\mathsf{P}_x(X^n(0) = x) = 1$, $x \in \mathsf{X}$, $n \in \mathcal{N}$.

**Theorem A.2.1** (Theorem 3.5 of [4].)

*For a function $f : \mathsf{X} \to \mathcal{R}_+$, a constant $\bar{\delta} \in (0, \infty)$ and a compact set $C \subset \mathsf{X}$, define*

$$V^n(x) \equiv \mathsf{E}_x\Big[\int_0^{\tau_C^n(\bar{\delta})} f(X^n(t))dt\Big], \quad x \in \mathsf{X}, n \in \mathcal{N},$$

*where* $\tau_C^n(\bar{\delta}) = \inf\{t \geq \bar{\delta} : X^n(t) \in C\}$. *If* $\sup_n V^n(x) < \infty$ *for any* $x \in \mathsf{X}$ *and* $\sup_{x\in C} \sup_n V^n(x) < \infty$, *then there exists a constant* $\bar{\kappa} \in (0, \infty)$ *such that*

$$\frac{1}{t}\mathsf{E}_x\Big[V^n(X^n(t))\Big] + \frac{1}{t}\int_0^t \mathsf{E}_x\Big[f(X^n(s))\Big]ds \leq \frac{1}{t}V^n(x) + \bar{\kappa} \tag{22}$$

*for any* $n \in \mathcal{N}$, $t \geq 1$ *and* $x \in \mathsf{X}$.

(The author has corrected the range of variable $t$ in Theorem 3.5 of [4] as above.)

**Theorem A.2.2** (Theorem 3.2 of [4].)

*Suppose that inequality* (22) *with* $f(x) = 1 + |x|^{p-1}$ *holds for some* $p \in [2, \infty)$. *Then we have*

$$\sup_n \int_{\mathsf{X}} |x|^{p-1}\, \pi_X^n(dx) < \infty.$$

## References


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

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

[3] Budhiraja, A. and Ghosh, A. (2006). Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function. *Ann. Appl. Probab.* **16** 1962-2006.

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

[5] Chen, H. and Yao, D. (2001). *Fundamentals of Queueing Networks.* Springer.

[6] Chung, K.L. (1974). *A Course in Probability Theory.* Academic Press.

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

[8] Dai, J.G. and Meyn, S.P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Trans. Automat. Control* **40** 1889-1904.

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

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

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

[12] Krichagina, E.V. (1989). Diffusion approximation for a queue in a multiserver system with multistage service. *Autom. Remote Control* **50** 346-354.

[13] Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst.* **30** 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