Limit non-stationary behavior of large closed queuing networks with bottlenecks

Yi-Ju Chao
127 Vincent Hall, University of Minnesota

Abstract

In this paper martingale methods are applied for analyzing asymptotes of queuing lengths of servers. We consider closed Jackson networks consisting of a large number of customers, an infinite server, and several single servers. The servers with the lowest service rate are referred to as bottleneck servers. Arrival times and departure times for each sever are assumed to be exponentially distributed with constant mean. It is also assumed that there are several bottleneck single servers in the system. For the non-bottleneck servers we found the queuing length processes are always in a light usage regime. For the bottleneck servers we found the critical times which discern the queuing length processes in a heavy usage regime and in a light usage regime. We also show that the queuing length distributions at time t converge in weak sense to stationary distributions in the light usage regime. The Gaussian diffusion approximation are established for the heavy usage regime.
Keywords: Bottleneck; closed Jackson networks; heavy usage regime; light usage regime.

1 Introduction

In this paper, we consider closed Jackson networks with a single class consisting of N customers, a single IS server (which can offer services to infinitely many customers simultaneously) and k single servers with fixed service rates. Customers are following FIFO (first come, first out) rules. The IS server is numbered by 0, while the single servers are numbered as $1, ..., k$. Customers having been served at the IS server visit the i-th single server with probability p_i, $\sum_{i=1}^{k} p_i = 1$. Customers having been served at a single server return to the IS server. We assume that the service times for the i-th single server are exponentially distributed with mean $(N\mu_i)^{-1}$ and the service times at the IS server are also exponentially distributed with mean λ^{-1} for each customer. The arrival rate at the i-th single server is the product of the number of customers in IS server and $\lambda_i = \lambda p_i$, since customers from IS server visit the i-th server

with probability p_i. All the customers are at the IS server initially. We also assume that the service times at all servers are mutually independent.

The asymptotes of queue length at each single server for large closed queueing networks with bottlenecks are the main interest in this paper. The asymptotes of the product form stationary distribution for large closed networks as $N \to \infty$ have been studied by different techniques in a number of papers [3], [5], [6], [8], [10], [13]. The limit of nonstationary behavior of such networks has been investigated in the case of one bottleneck and non-IS server [13] and in the case of one single server forming bottleneck [4]. The case of all servers being uniformly loaded bottlenecks is discussed in [7].

We consider the case that the first r servers are bottleneck ones, while the remaining are non-bottleneck. More precisely, we assume as $1 \leq i \leq r$, the traffic intensity $\frac{\lambda}{\mu}$ satisfies

$$1 < \frac{\lambda_1}{\mu_1} < \cdots < \frac{\lambda_i}{\mu_i} < \frac{\lambda_{i+1}}{\mu_{i+1}} < \cdots < \frac{\lambda_r}{\mu_r}. \tag{1.1}$$

Since $\lambda_i > \mu_i$, they are called bottleneck servers. As $r+1 \leq i \leq k$, the arrival rate is assumed to be less than the service rate (departure rate), that is

$$\mu_i > \lambda_i,$$

so they are called non-bottleneck servers.

The method we use here is based on stochastic calculus and similar to the one used by Kogan and Liptser in [4]. The difference between the case in [4] and this one is that they considered only one bottleneck single server whereas there can be several bottleneck single servers in our case. They found that asymptotically the bottleneck single server is always in a heavy traffic regime while the non-bottleneck single servers are always in a light traffic regime. In the model of this paper, we found that it is necessary to introduce critical times for the first $r-1$ bottleneck single servers. Asymptotically, the first $r-1$ bottleneck single servers are in a heavy traffic regime initially then in a light traffic regime after critical times, the r-th bottleneck single server is always in a heavy traffic regime, and the non-bottleneck single servers are still always in a light traffic regime.

The paper is organized as follows. The notation and main results are given in section 2. The deterministic approximation is constructed in section 3. In section 4, the ergodic properties for normalized queue length processes are characterized by the heavy traffic usage regime and by the light traffic usage regime. The proof of the main results are given in section 5.

2 Notations and Main Results

Let $(\Omega, F = (F_t)_{t \geq 0}, P)$ be a probability space. Let N be the total number of customers of the queuing network. We assume there are given mutually

independent Poisson F_t-adapted processes $\Pi = (\Pi_i(t))_{t\geq 0}$, $1 \leq i \leq k$, and $\pi = (\pi_{ij}(t))_{t\geq 0}$, $1 \leq j \leq N$, $1 \leq i \leq k$ with intensities $N\mu_i$, and λ_i respectively.

The number of customers at each single server are represented by the random vector $Q^N(t) = (Q_1^N(t), ..., Q_k^N(t))$ on the probability space $(\Omega, F = (F_t)_{t\geq 0}, P)$. To write the formula of $Q^N(t)$, we need to introduce the departure processes and arrival processes. The following discussion can be found in the article [4]. The departure process at the i-th single server is

$$D_i^N(t) = \int_0^t I(Q_i^N(s-) > 0) d\Pi_i(s),$$

where the function $I(B)$ is the indicatior of set B, and

$$Q_i^N(s-) = \lim_{u \to s^-} Q_i^N(u).$$

To form the arrival process $A_i(t)$ at the i-th single server, we assume π_{ij} to be a Poisson process obtained by thinning the Poisson process π_j with probability p_i. The arrival process at the i-th single server is

$$A_i^N(t) = \int_0^t \sum_{j=1}^{N} I(N - \sum_{l=1}^{k} Q_l^N(s-) \geq j) d\pi_{ij}(s).$$

The queuing length process at the i-th single server satisfies

$$Q_i^N(t) = A_i^N(t) - D_i^N(t) = A_i^N(t) - \Pi_i(t) + \int_0^t I(Q_i^N(s-) = 0) d\Pi_i(s). \quad (2.2)$$

Since there are no customers at each single server initially, we have $Q^N(0) = 0$. The normalized queuing lengths $q^N(t) = (q_1^N(t), ..., q_k^N(t))$ are defined by $q_i^N(t) = N^{-1} Q_i^N(t)$ for all $i = 1, ..., k$. Dividing (2.2) by N and introducing the compensators [9] of the processes $A_i^N(t)$ and $D_i^N(t)$, we can get

$$q_i^N(t) = \int_0^t (\lambda_i (1 - \sum_{l=1}^{k} q_l^N(s)) - \mu_i) ds + m_i^N(t) + N^{-1} \int_0^t I(Q_i^N(s-) = 0) d\Pi_i(s), \quad (2.3)$$

where $m_i^N(t)$ are local square-integrable martingales with jumps no larger than N^{-1} and with predictable quadratic variation

$$< m_i^N >_t = N^{-1} \int_0^t [\lambda_i (1 - \sum_{l=1}^{k} q_l^N(s)) + \mu_i] ds, \quad (2.4)$$

and covariances

$$< m_i^N, m_j^N >_t \equiv 0 \quad \forall i \neq j.$$

Let D denote the Skorokhod space of real-valued functions. For a stochastic process $x(t)$ in D, we introduce the notation

$$\Phi_t(x) = x(t) - \inf_{s \le t} x(s) \tag{2.5}$$

to be the normal reflection of the process $x(t)$ [12]. In article [4], it's proved that the process $q^N(t) = (q_1^N(t), \cdots, q_k^N(t))$ is represented by the normal reflection of the solution of the following stochastic differential system

$$x_i^N(t) = \int_0^t [\lambda_i(1 - \sum_{l=1}^{k} \Phi_s(x_l^N)) - \mu_i]ds + m_i^N(t), \tag{2.6}$$

$$x_i^N(0) = 0.$$

Besides, we consider the deterministic system

$$x_i(t) = \int_0^t [\lambda_i(1 - \sum_{l=1}^{k} \Phi_s(x_l)) - \mu_i]ds, \tag{2.7}$$

$$x_i(0) = 0.$$

Since $\Phi_t(x)$ satisfies Lipschitz condition (see Lemma 3.1), there is a unique solution of (2.7), denoted by $x_i(t)$. In addition, we define the function

$$q(t) = \sum_{l=1}^{k} \Phi_t(x_l). \tag{2.8}$$

Actually, the function $q(t)$ is the asymptotic of the total amount of normalized queue lengths of the single servers. We will prove that

$$\lim_{N\to\infty} q_i^N(t) = \lim_{N\to\infty} \Phi_t(x_i^N) = \Phi_t(x_i)$$

in probability later (see Corollary 4.1), which is the asymptotic of the queuing length at the i-th single server. We also introduce the deviation

$$V_i^N(t) = \sqrt{N}(x_i^N(t) - x_i(t)) \tag{2.9}$$

for $i = 1, ..., r$.

The main results are as follows:

Theorem 2.1 *The sequence $(V_i^N)_{i=1}^r, N \ge 1$ converges weakly (in the Skorokhod-Lindvall topology of space D) to a r-dimensional diffusion process $(V_i(t))_{i=1}^r$ defined by the Itô equation*

$$V_i(t) = -\lambda_i \int_0^t [\sum_{j=1}^{r} V_j(s) I_{[0,\alpha_j]}(s)]ds + \int_0^t \sqrt{\lambda_i(1 - q(s)) + \mu_i} dw_s^i \tag{2.10}$$

for all $i = 1, ..., r$, *where* α_j *are certain constants to be specified later satisfying* $0 < \alpha_1 < \alpha_2 < .... < \alpha_r = \infty$, *and* $(w^i)_{i=1}^r$ *are mutually independent Wiener processes.*

Theorem 2.2 *For any* $\epsilon > 0$, *and* $\delta > 0$, *we have the following results:*

A) For any $1 \leq i \leq r-1$,

$$\lim_N P(\sup_{0 \leq t \leq \alpha_i - \delta} \sqrt{N}|q_i^N(t) - x_i^N(t)| \geq \epsilon) = 0.$$

For any fixed $T > \alpha_i$

$$\lim_N P(\sup_{\alpha_i + \delta \leq t \leq T} \sqrt{N} q_i^N(t) \geq \epsilon) = 0.$$

B) For any fixed $T > 0$,

$$\lim_N P(\sup_{t \leq T} \sqrt{N}|q_r^N(t) - x_r^N(t)| \geq \epsilon) = 0.$$

C) For any $r+1 \leq i \leq k$, *and for any fixed* $T > 0$,

$$\lim_N P(\sup_{t \leq T} \sqrt{N} q_i^N(t) \geq \epsilon) = 0.$$

Theorem 2.3 *For* $i = r+1, ..., k$, *for any fixed* $T > 0$ *and any smooth function* $f(t)$, $0 \leq t \leq T$,

$$\lim_N \int_0^T f(t) P(Q_i^N(t) = l) dt = \int_0^T f(t)(1 - \rho_i(t))\rho_i^l(t) dt$$

where $\rho_i(t) = \lambda_i \mu_i^{-1}(1 - q(t)) < 1$, *for all integers* $l = 0, 1, 2....$

Theorem 2.4 *For* $i = 1, ..., r-1$, *for any fixed* $T \geq \alpha_i$ *and any smooth function* $f(t)$, $\alpha_i \leq t \leq T$,

$$\lim_N \int_{\alpha_i}^T f(t) P(Q_i^N(t) = l) dt = \int_{\alpha_i}^T f(t)(1 - \rho_i(t))\rho_i^l(t) dt$$

where $\rho_i(t) = \lambda_i \mu_i^{-1}(1 - q(t))$, *for all integers* $l = 0, 1, 2....$

By Theorem 2.1, we can see that the behavior of the bottleneck single servers are independent of the non-bottleneck single servers. Theorem 2.1 and Theorem 2.2(A) show that before the critical times α_i, the queuing lengths normalized by N at bottleneck servers have a deterministic limit. The deterministic limit is the solution of (2.7) and the deviation of order $N^{-1/2}$ from the deterministic

limit are approximated by a Gaussian diffusion process. In this case, we say the bottleneck servers are in a heavy traffic regime. Theorem 2.2(B) shows that after the critical time α_i, the queuing lengths normalized by $N^{1/2}$, i.e. $\sqrt{N}q_i^N(t)$, tend to zero in probability as $N \to \infty$, so we say that the bottleneck servers are in a light traffic regime. Asymptotically, the arrival rate for each server actually is $N\lambda_i(1-q(t))$, for $1-q(t)$ is the amount of customers in the IS server at time t while the service rate is $N\mu_i$. The function $q(t)$ will be proved to be an increasing function (see Lemma 3.5), so the number of customers at the bottleneck servers decay. On the other hand, Theorem 2.2(C) shows that the non-bottleneck servers are always in a light traffic regime since $\sqrt{N}q_i^N(t)$ tends to zero in probability as $N \to \infty$. The results of Theorem 2.3 and Theorem 2.4 show the weak limits of the queuing lengths of the non-bottleneck servers and those of the bottleneck servers after the critical times α_i as $N \to \infty$. In these cases, because the arrival rates $N\lambda_i(1-q(t))$ are less than the departure rate $N\mu_i$, at any fixed time t the traffic intensities of each single server $\rho_i(t) = \lambda_i\mu_i^{-1}(1-q(t))$ are less than one. We can get the stationary distribution of the queuing lengths.

3 Deterministic Approximation

Before giving the unique solution of (2.7), we introduce two lemmas.

Lemma 3.1 *For fixed $T > 0$, the function $\Phi_t(x) = x(t) - \inf_{s\le t} x(s)$ satisfies*

(A) $\sup_{t\le T} x(t) \le \sup_{t\le T} \Phi_t(x)$;

(B) Lipschitz Condition: $\sup_{t\le T} |\Phi_t(y) - \Phi_t(y)| \le \sup_{t\le T} |x(t) - y(t)|$.

Proof. The reader can find the proof of this lemma in the article [4].

Remark 3.1 *The Lipschitz condition of Lemma 3.1 implies that there is a unique solution of (2.7).*

Considering (2.7), we first notice as $r+1 \le i \le k$,

$$x_i^{'}(t) \le \lambda_i(1 - \sum_{l=1}^{k} \Phi_t(x_l)) - \mu_i \le \lambda_i - \mu_i \le 0,$$

according to the non-bottleneck assumption. With the initial condition $x_i(0) = 0$, we can conclude the equalities

$$\Phi_t(x_i) \equiv 0 \text{ as } r+1 \le i \le k.$$

The above result implies that the function $x_i(t)$, as $r+1 \le i \le k$, doesn't affect the solution of (2.7). We can only consider the system with the first r equations divided by λ_i of (2.7). Let

$$y_i(t) := x_i(t)/\lambda_i,$$

we have the system

$$y_i(0) = 0,$$

$$y_i(t) = \int_0^t [1 - \sum_{l=1}^{r} \lambda_l \Phi_s(y_l) - \frac{\mu_i}{\lambda_i}]ds, \text{ for } i = 1, ..., r. \tag{3.11}$$

There is a unique solution of (3.11), since $\Phi_s(y_l)$ still satisfies Lipschitz condition.

Lemma 3.2 *The solution of (3.11)* $(y_i(t))_{i=1}^r$ *satisfies*

$$y_1(t) < \cdots < y_i(t) < y_{i+1}(t) < \cdots < y_r(t),$$

for all t.

Proof. Observing that $\sum_{l=1}^{k} \lambda_l \Phi_s(y_l)$ is independent of i and using the monotonicity of $\frac{\lambda_i}{\mu_i}$ in (1.1), we can get the result. The lemma is proved.

The main idea of solving (3.11) is to construct the solution step by step on index $i = 1, ..., r$. In this construction, we need to introduce a critical time α_i which discerns the functions $\Phi_t(y_i)$ and $y_i(t)$. The solution of (3.11) is the same as the solution of the following system

$$z_i(0) = 0,$$

$$z_i(t) = \int_0^t [1 - \sum_{l=1}^{r} \lambda_l z_l(s) - \frac{\mu_i}{\lambda_i}]ds, \text{ for } i = 1, ..., r, \tag{3.12}$$

if $\Phi_s(y_l)$ equals $y_l(s)$ for $l = 1, ..., r$. We know that $\Phi_t(y_l)$ equals $y_l(t)$ as t is close to zero by observing the equality $\Phi_0(y_l) = y_l(0) = 0$ and the fact that $y_l(t)$ is positive as t is close to zero. Therefore, we can solve (3.11) by solving (3.12) as t is close to zero.

We start to construct the solution of (3.11). First, we introduce a critical time α_1, which is the first time when $y_1(t)$ becomes negative, to discern the functions $\Phi_t(y_1)$ and $y_1(t)$. Lemma 3.2 implies the equality $\Phi_t(y_l) = y_l(t)$, for $t \leq \alpha_1$, for $1 \leq l \leq r$. We define the following function

$$q_1(t) := (1 - \frac{\sum_{j=1}^{r} \mu_j}{\sum_{j=1}^{r} \lambda_j})(1 - \exp(-\sum_{j=1}^{r} \lambda_j t)) \tag{3.13}$$

and the constants

$$\alpha_0 := 0,$$

and

$$\alpha_1 := \inf\{t : t > 0 \text{ and } \int_0^t (1 - q_1(s) - \frac{\mu_1}{\lambda_1})ds = 0\}.$$

Noticing that the integrals

$$\int_0^t (1 - q_1(s) - \frac{\mu_i}{\lambda_i})ds, \text{ for } i = 1, ..., r, \tag{3.14}$$

solve (3.12), so it suffices to check the equality $\Phi_t(z_i) = z_i(t)$ for $t \in [0, \alpha_1]$, for $i = 1, ..., r$, to conclude that (3.14) solve (3.11) up to the critical time α_1. The definition of α_1 and the bottleneck assumption (1.1) imply the inequality $z_i(t) \geq 0$ on $[0, \alpha_1]$ for $i = 1, ..., r$, so the equality $\Phi_t(z_i) = z_i(t)$ is thus fulfilled. The reader can easily check that $\alpha_1 < \infty$ if $r > 1$ and $\alpha_1 = \infty$ if $r = 1$ according to the bottleneck assumption.

We define the constants α_i and the functions $q_i(t)$ iteratively as follows: for $i = 2, 3, ...,$

$$q_i(t) := (1 - \frac{\sum_{j=i}^r \mu_j}{\sum_{j=i}^r \lambda_j})(1 - \exp(-\sum_{j=i}^r \lambda_j(t - \alpha_{i-1}))) + q_{i-1}(\alpha_{i-1}) \exp(-\sum_{j=i}^r \lambda_j(t - \alpha_{i-1})), \tag{3.15}$$

and

$$\alpha_i := \inf\{t : t > 0 \text{ and } \int_0^t [1 - \sum_{j=1}^{i-1} q_j(s) I_{(\alpha_{j-1}, \alpha_j]}(s) - q_i(s) I_{(\alpha_{i-1}, \infty)}(s) - \frac{\mu_i}{\lambda_i}]ds = 0\}. \tag{3.16}$$

If α_i is finite, we can define $q_{i+1}(t)$ iteratively by (3.15).

Remark 3.2 *We have $\alpha_1 < \alpha_2 < \cdots$ by observing (3.16) and the bottleneck assumption (1.1).*

We give the induction hypothesis: For some $n < r$, we assume the integrals

$$g_i(t) := \sum_{j=1}^n \int_{t \wedge \alpha_{j-1}}^{t \wedge \alpha_j} (1 - q_j(s) - \frac{\mu_i}{\lambda_i})ds, \text{ for } i = 1, ...r, \text{ for } t \leq \alpha_n, \tag{3.17}$$

solve (3.11) on $[0, \alpha_n]$.

Before proving that (3.17) with replacing n by $n+1$ solve the system (3.11) on $[0, \alpha_{n+1}]$ where α_{n+1} will be defined later, we need to introduce several remarks and lemmas.

Remark 3.3 *The functions $(g_i(t))_{i=j}^r$, for $j = 1, ..., r$, solve the system*

$$y_i(t) = \int_{\alpha_{j-1}}^t (1 - \sum_{l=j}^r \lambda_l y_l(s) - \frac{\mu_i}{\lambda_i})ds + y_i(\alpha_{j-1}), \text{ for } i = j, ..., r, \tag{3.18}$$

on the intervals $[\alpha_{j-1}, \alpha_j]$.

We define a function $q(t)$ on $[0, \alpha_n]$ as follows:

$$q(0) := 0,$$

$$q(t) := \sum_{i=1}^{n} q_i(t) I_{(\alpha_{i-1}, \alpha_i]}(t) \text{ if } \alpha_n < \infty,$$

and

$$q(t) := \sum_{i=1}^{n-1} q_i(t) I_{(\alpha_{i-1}, \alpha_i]}(t) + q_n(t) I_{(\alpha_{n-1}, \infty)}(t) \text{ if } \alpha_n = \infty.$$

Lemma 3.3 *Under this induction hypothesis, the inequality* $q(\alpha_m) > \frac{\lambda_m - \mu_m}{\lambda_m}$ *is true for* $m \leq n$.

Proof. This lemma can be proved by induction. As $m = 1$, we consider the function

$$g_1(t) = \int_0^t (1 - q_1(s) - \frac{\mu_1}{\lambda_1}) ds, \text{ for } t \in [0, \alpha_1].$$

In particular, we have

$$q(t) = q_1(t) = (1 - \frac{\sum_{j=1}^{r} \mu_j}{\sum_{j=1}^{r} \lambda_j})(1 - \exp(-\sum_{j=1}^{r} \lambda_j t))$$

on the interval $[0, \alpha_1]$. The function $g_1(t)$ is strictly concave because that $g_1^{''}(t) = -q_1^{'}(t)$ is less than zero. With the fact that $g_1(\alpha_1) = 0$ from the definition of α_1, there is a constant $\alpha_1^* < \alpha_1$ such that

$$\max_t g_1(t) = g_1(\alpha_1^*),$$

and

$$g_1^{'}(\alpha_1^*) = 1 - q_1(\alpha_1^*) - \frac{\mu_1}{\lambda_1} = 0.$$

By the above equation and the fact that $q_1(t)$ is increasing, we can get the inequality

$$\frac{\lambda_1 - \mu_1}{\lambda_1} = q_1(\alpha_1^*) < q_1(\alpha_1) = q(\alpha_1).$$

Remark 3.3 implies

$$q_1(t) = \sum_{i=1}^{r} \lambda_i g_i(t) \text{ for } t \leq \alpha_1,$$

so we have

$$q_1(t) - \lambda_1 g_1(t) = \sum_{i=2}^{r} \lambda_i g_i(t) = \int_0^t \sum_{i=2}^{r} (\lambda_i - \mu_i - \lambda_i q_1(s)) ds.$$

Since $q_1(t)$ is increasing and $g_1(t)$ is decreasing as $\alpha_1^* \leq t \leq \alpha_1$, the derivative of the above function at α_1

$$q_1^{'}(\alpha_1) - \lambda_1 g_1^{'}(\alpha_1) = \sum_{i=2}^{r}(\lambda_i - \mu_i - \lambda_i q_1(\alpha_1))$$

is larger than zero. This result implies the inequality

$$q(\alpha_1) = q_1(\alpha_1) < \frac{\sum_{i=2}^r \lambda_i - \sum_{i=2}^r \mu_i}{\sum_{i=2}^r \lambda_i}.$$

We give the induction hypothesis: for some $m < n$, the inequality

$$\frac{\lambda_m - \mu_m}{\lambda_m} < q(\alpha_m) = q_m(\alpha_m) < \frac{\sum_{j=m+1}^r \lambda_j - \sum_{j=m+1}^r \mu_j}{\sum_{j=m+1}^r \lambda_j}$$

is true. Since the function

$$q_{m+1}^{'}(t) = (\frac{\sum_{j=m+1}^r \lambda_j - \sum_{j=m+1}^r \mu_j}{\sum_{j=m+1}^r \lambda_j} - q_m(\alpha_m)) \exp(-\sum_{j=m+1}^{r} \lambda_j(t - \alpha_m))$$

is larger than zero according to the induction hypothesis, the function $q_{m+1}(t)$ is strictly increasing.

We claim that α_{m+1} is less that ∞ for $m + 1 \leq n$. To prove this claim, we assume that α_{m+1} is equal to ∞. It follows

$$q(t) = \sum_{i=1}^{m} q_i(t) I_{(\alpha_{i-1}, \alpha_i]}(t) + q_{m+1}(t) I_{(\alpha_m, \infty)}(t) \text{ for all } t.$$

The definition (3.17) and the fact that $q_j(t)$ is increasing on $[\alpha_{j-1}, \alpha_j]$, for $j \leq n$, imply that

$$g_{m+1}^{''}(t) = -\sum_{j=1}^{m+1} q_j^{'}(t) I_{(\alpha_{j-1}, \alpha_j]}(t) \text{ for all } t \geq 0$$

is less than zero. Therefore, the function $g_{m+1}(t)$ is strictly concave. The inequality

$$\lim_{t\to\infty} g_{m+1}^{'}(t) = \frac{\lambda_{m+1} - \mu_{m+1}}{\lambda_{m+1}} - \lim_{t\to\infty} q(t) <$$

$$\frac{\lambda_{m+1} - \mu_{m+1}}{\lambda_{m+1}} - \frac{\sum_{j=m+1}^r \lambda_j - \sum_{j=m+1}^r \mu_j}{\sum_{j=m+1}^r \lambda_j} < 0,$$

implies that there is a constant $\alpha_{m+1} < \infty$ such that $g_{m+1}(\alpha_{m+1}) = 0$. This result contradicts to the assumption $\alpha_{m+1} = \infty$. The claim is proved.

To prove the inequality

$$\frac{\lambda_{m+1} - \mu_{m+1}}{\lambda_{m+1}} < q(\alpha_{m+1}),$$

we can just use the same argument as before. The fact that

$$g_{m+1}(0) = g_{m+1}(\alpha_{m+1}) = 0$$

implies that there is a constant $\alpha^*_{m+1} < \alpha_{m+1}$ such that

$$\max_t g_{m+1}(t) = g_{m+1}(\alpha^*_{m+1}),$$

and

$$g'_{m+1}(\alpha^*_{m+1}) = 1 - q(\alpha^*_{m+1}) - \frac{\mu_{m+1}}{\lambda_{m+1}} = 0.$$

Since $q(t)$ is increasing, we have

$$\frac{\lambda_{m+1} - \mu_{m+1}}{\lambda_{m+1}} = q(\alpha^*_{m+1}) < q(\alpha_{m+1}).$$

The lemma is proved.

Remark 3.4 *According to the proof of this lemma, for $m \leq n$, there exists a constant $\alpha^*_m < \alpha_m$ such that $q(\alpha^*_m) = \frac{\lambda_m - \mu_m}{\lambda_m}$ with $\max_t g_m(t) = g_m(\alpha^*_m)$.*

From the proof of the above lemma, we know that $\alpha_n < \infty$ for $n < r$. We extend the function $q(t)$ to the interval $[0, \alpha_{n+1}$ by defining

$$q(t) = q_{n+1}(t) \text{ for } t \in [\alpha_n, \alpha_{n+1}].$$

We want to prove that the functions

$$g_i(t) = \int_0^t (1 - q(s) - \frac{\mu_i}{\lambda_i})ds, \text{ for } i = 1, ...r, \tag{3.19}$$

solve (3.11) on $[0, \alpha_{n+1}]$.

Lemma 3.4 *Equation (3.19) solves the system (3.11) on* $[0, \alpha_{n+1}]$.

Proof. By the definition of $q(t)$, and (3.19), we notice that, for $t > \alpha_n$, for $1 \leq i \leq r$, the functions

$$g_i(t) = \int_{\alpha_n}^t (1 - q_{n+1}(s) - \frac{\mu_i}{\lambda_i})ds + g_i(\alpha_n)$$

satisfy the system

$$z_i(t) = z_i(\alpha_n) + \int_{\alpha_n}^t [1 - \sum_{j=n+1}^{r} \lambda_j z_j(s) - \frac{\mu_i}{\lambda_i}]ds, \text{ for } n+1 \leq i \leq r. \tag{3.20}$$

To prove the function $g_i(t)$ solves (3.11), it suffices to show that for $1 \leq i \leq n$, for $\alpha_n \leq t < \alpha_{n+1}$, we have $\Phi_t(g_i) \equiv 0$, for $n+1 \leq i \leq r$ we have $\Phi_t(g_i) = g_i(t)$.

According to Lemma 3.3, the first derivative of $g_i(t)$

$$g_i^{'}(t) = 1 - \sum_{j=n+1}^{r} \lambda_j g_j(t) - \frac{\mu_i}{\lambda_i} = 1 - q_{n+1}(t) - \frac{\mu_i}{\lambda_i}$$

$$= (1-\frac{\mu_n}{\lambda_n})-(1-\frac{\sum_{j=n+1}^{r}\mu_j}{\sum_{j=n+1}^{r}\lambda_j})(1-\exp(-\sum_{j=n+1}^{r}\lambda_j(t-\alpha_n)))-q_n(\alpha_n)\exp(-\sum_{j=n+1}^{r}\lambda_j(t-\alpha_n)),$$

is less or equal to the following term

$$(\frac{\sum_{j=n+1}^{r}\mu_j}{\sum_{j=n+1}^{r}\lambda_j} - \frac{\mu_n}{\lambda_n})(1-\exp(-\sum_{j=n+1}^{r}\lambda_j(t-\alpha_n))). \tag{3.21}$$

The above term is less than zero the fact that the bottleneck assumption (1.1) implies

$$\frac{\sum_{j=n+1}^{r}\mu_j}{\sum_{j=n+1}^{r}\lambda_j} - \frac{\mu_n}{\lambda_n} < 0.$$

So the second term of (3.20) is less than zero and decreasing. For $1 \leq i \leq n$, the function $g_i(t)$ is decreasing for $t > \alpha_i$ and $g_i(\alpha_n) < 0$ according to Lemma 3.2 and the definition of α_i.

We can conclude that, for $1 \leq i \leq n$,

$$\Phi_t(g_i) = g_i(t) - \inf_{s \leq t} g_i(s) = g_i(t) - \inf_{\alpha_n \leq s \leq t} g_i(s) = g_i(t) - g_i(t) = 0, \text{ for } t > \alpha_i,$$

On the other hand, according to Lemma 3.2, for $n+1 \leq i \leq r$,

$$0 < g_{n+1}(t) < g_i(t) \text{ for } 0 < t < \alpha_{n+1},$$

so the equality $\Phi_t(g_i) = g_i(t)$ is true. The lemma is proved.

Remark 3.5 *The function $(\lambda_i g_i(t))_{i=1}^{r}$ is the unique solution of (2.7) according to Lemma 3.4, so we have $\lambda_i g_i(t) = x_i(t)$, for $i = 1, ..., r$.*

Let $x_i(t)$, for all t, and for $1 \leq i \leq r$, denotes the solution of (2.7). The important properties of $q(t)$, and $x_i(t)$ are as follows.

Lemma 3.5 *(A) $q(t)$ is a continuous, positive, and strictly increasing function satisfying $0 \leq q(t) < \frac{\lambda_r - \mu_r}{\lambda_r}$;*

(B) For all $i \leq r-1$, there exists $\alpha_i^* < \alpha_i$ such that $q(\alpha_i^*) = \frac{\lambda_i - \mu_i}{\lambda_i}$ with $x_i(\alpha_i^*) = \max_t x_i(t)$.

Proof. For part (A), To prove that $q(t)$ is continuous, we only need to check $q_i(\alpha_{i-1}) = q_{i-1}(\alpha_{i-1})$. This follows from the definition of $q(t)$. In the proof of Lemma 3.3, we prove that $q_i(t)$ is strictly increasing on each interval $(\alpha_{i-1}, \alpha_i]$. Since $q(t)$ is continuous,

$$q(0) = 0,$$

and

$$q(t) = \sum_{i=1}^{r-1} q_i(t) I_{(\alpha_{i-1},\alpha_i]}(t) + q_r(t) I_{(\alpha_{r-1},\infty)},$$

we can see that $q(t)$ is strictly increasing.

We notice that the function

$$q_r(t) = (1 - \frac{\mu_r}{\lambda_r})(1 - \exp(-\lambda_r(t - \alpha_{r-1}))) + q_{r-1}(\alpha_{r-1}) \exp(-\lambda_r(t - \alpha_{r-1}))$$

satisfies

$$\lim_{t\to\infty} q(t) = 1 - \frac{\mu_r}{\lambda_r},$$

so we can get

$$0 \le q(t) < \frac{\lambda_r - \mu_r}{\lambda_r},$$

since $q(t)$ is increasing.

Part (B) follows from Remark 3.4 and Remark 3.5.

Lemma 3.6 *For $1 \le i \le k$, $x_i(t)$ is a continuous, concave function with $x_i(0) = 0$ and satisfies:*

(A) $x_i(t) \begin{cases} > 0 & \text{for} \quad 0 < t < \alpha_i \\ = 0 & \text{for} \quad t = \alpha_i \\ < 0 & \text{for} \quad t > \alpha_i \end{cases}$, *for* $1 \le i \le r-1$;

(B) $x_r(t) > 0$, *for all* $t > 0$;

(C) $x_i(t) < 0$, *for all* $t > 0$, *for* $r+1 \le i \le k$.

Proof. We know that $g_i(t)$ is continuous, concave with $g_i(0) = 0$, so $x_i(t)$ satisfies the same properties according to Remark 3.5.

For part (A), since for $1 \le i \le r-1$, the function

$$x_i(t) = \lambda_i g_i(t)$$

is concave and satisfying

$$x_i(\alpha_i) = \lambda_i g_i(\alpha_i) = 0,$$

we have the result.

For part (B), we notice that

$$x_r(t) = \lambda_r g_r(t) = \lambda_r \int_0^t (1 - q(s) - \frac{\mu_r}{\lambda_r})ds.$$

Since $0 < q(t) < \frac{\lambda_r - \mu_r}{\lambda_r}$ by Lemma 3.5(A), we know the following inequality $x_r(t) > 0$ for $t > 0$.

For part (C), the result follows from that for $r + 1 \leq i \leq k$, for $t > 0$,

$$x_i(t) = \lambda_i g_i(t) = \lambda_i \int_0^t (1 - \frac{\mu_i}{\lambda_i} - q(s))ds.$$

According to the non-bottleneck assumption, $\mu_i > \lambda_i$ for $r + 1 \leq i \leq k$, the above function is less than zero. The lemma is proved.

Corollary 3.1 *(A)* $\Phi_t(x_i) = \begin{cases} x_i(t) & t \leq \alpha_i \\ 0 & t > \alpha_i \end{cases}$ *, for* $1 \leq i \leq r - 1$;

(B) $\Phi_t(x_r) = x_r(t)$*, for all* t*;*

(C) $\Phi_t(x_i) = 0$*, for all* $r + 1 \leq i \leq k$.

Proof. These are the results of applying Lemma 3.6 and the definition of $\Phi_t(x)$.

4 Ergodic Properties for Normalized Queue Length Process

Lemma 4.1 *For all* $T \leq 0$ *and* $\epsilon > 0$,

$$\lim_N P(\sup_{t \leq T} |x_i^N(t) - x_i(t)| \geq \epsilon) = 0,$$

for $1 \leq i \leq k$.

Proof. The lemma is the result of using the Lipschitz condition of $\Phi(x)$ (see Theorem 3.1) and applying the Gronwall-Bellman Inequality on (2.6). The proof can be found in the article [4].

Corollary 4.1 *For all* $T \leq 0$ *and* $\epsilon > 0$,

$$\lim_N P(\sup_{t \leq T} |q_i^N(t) - \Phi_t(x_i)| \geq \epsilon) = 0$$

for $1 \leq i \leq k$.

Proof. By the Lipschitz condition (see Lemma 3.1) and the above lemma, we can get the result.

Remark 4.1 *Combing Corollary 3.1 and Corollary 4.1, we can get*

$$\lim_N P(\sup_{\alpha_i \le t \le T} q_i^N(t) \ge \epsilon) = 0, \text{ for } 1 \le i \le r-1.$$

Remark 4.2 *In particular, we have*

$$\lim_N P(\sup_{t \le T} |\sum_{i=1}^k \Phi_t(x_i^N) - q(t)| \ge \epsilon) = 0$$

because of $q(t) = \sum_{i=1}^k \Phi_t(x_i)$ *(see 2.8).*

To get the results of Theorem 2.1 and Theorem 2.2, we need the following two lemmas. The first one exploits the $N^{-1/2}$-deviation of limit theorems in a heavy usage regime, while the second one describes that the normalized queuing length $q_i^N(t)$, for $r+1 \le i \le k$, is of $o(\sqrt{N})$ in a light usage regime.

Lemma 4.2 *For any fixed* $T > 0$, *we define* $Y^N(T) = \sum_{j=1}^k \sup_{t \le T} \sqrt{N}|x_j^N(t) - x_j(t)|$. *Then for all* $M > 0$, *we have* $P(Y^N(T) \ge M) \le C/M^2$ *for some constant* C *independent of* N.

Proof. By (2.6) and (2.7), we have the following formula

$$\sqrt{N}(x_i^N(s) - x_i(s)) = \int_0^s [\sqrt{N}\lambda_i \sum_{j=1}^k (\Phi_u(x_j) - \Phi_u(x_j^N))]du + \sqrt{N}m_i^N(s). \quad (4.22)$$

Applying the Lipschitz condition (see Lemma 3.1), and taking summation from $i = 1$ to k, we can get

$$Y^N(t) \le 2\lambda \int_0^t Y^N(s)ds + \sum_{j=1}^k \sup_{s \le t} |\sqrt{N}m_i^N(s)|. \quad (4.23)$$

By the Gronwall-Bellman inequality we can get

$$Y^N(T) \le \exp(2\lambda T) \sum_{i=1}^k \sup_{t \le T} |\sqrt{N}m_i^N(t)|. \quad (4.24)$$

On the other hand, since $\sqrt{N}m_i^N(t)$ is local square-integrable martingale with predictable quadratic variations (2.4), we have

$$< \sqrt{N}m_i^N >_T = N < m_i^N >_T = \int_0^T [\lambda_i(1 - \sum_{j=1}^k \Phi_t(x_j^N)) + \mu_i]dt \le T(\lambda_i + \mu_i).$$

By the Kolmogorov's inequality we can get

$$P(\sup_{t\leq T}|\sqrt{N}m_i^N(t)|\geq M)\leq C_1/M^2 \tag{4.25}$$

for any $M>0$, for some constant C_1 independent of N. Combining (4.24) and (4.25), we can get the result.

Lemma 4.3 *Let $X^N(t)=-c_N\int_0^t b^N(s)ds+G^N(t)$ be a sequence of random processes with paths in space D and $G^N(0)=0$. For a constant $\alpha>0$, if the following conditions are satisfied:*

(A) $\lim_N P(\Phi_\alpha(X^N)>\epsilon)=0$, for all $\epsilon>0$;

(B) For all fixed $T>0$, there exists a constant $b>0$ (possibly depending on T) such that $\lim_N P(\inf_{\alpha\leq t\leq T}b^N(t)\leq b)=0$;

(C) $0\leq c_N\uparrow\infty$ as $N\uparrow\infty$;

(D) the sequence $G^N(t), N\geq 1$, converges weakly (in the Skorokhod-Lindvall topology of space D) to the continuous random process $G(t)$,

then we have for any fixed $T>0$ and $\epsilon>0$,

$$\lim_N P(\sup_{\alpha\leq t\leq T}\Phi_t(X^N)\geq\epsilon)=0.$$

Remark 4.3 *This lemma is following Lemma 5.1 in the article [4]. We only slightly generalize the assumptions. The proof is included for completeness.*

Proof. First, we introduce the sets:

$$B_1=\{\sup_{t\leq T}|G^N(t)|\leq r\},$$

$$B_2=\{\sup_{|u-v|\leq a;u,v\leq T}|G^N(u)-G^N(v)|\leq\epsilon/2\},$$

$$B_3=\{\inf_{\alpha\leq t\leq T}b^N(t)>b\},$$

and

$$B_4=\{|\Phi_\alpha(X^N)|<\epsilon\},$$

for $r>0$, $a>0$, and b in the assumption (B). Since by the assumption (D) we have

$$\lim_{r\to\infty}\limsup_N P(\sup_{t\leq T}|G^N(t)|>r)=0,$$

and

$$\lim_{a\to 0}\lim_N P(\sup_{|u-v|\leq a;u,v\leq T}|G^N(u)-G^N(v)|>\epsilon/2)=$$

$$\lim_{a\to 0} P(\sup_{|u-v|\le a; u,v\le T} |G(u)-G(v)| > \epsilon/2) = 0,$$

we can get

$$\lim_{a\to 0}\lim_{r\to\infty}\limsup_N \sum_{i=1}^4 P(\Omega\backslash B_i) = 0 \tag{4.26}$$

with the assumptions (A) and (B).

We define a random time $\sigma = \inf\{t : t \ge \alpha; |\Phi_t(X^N)| > 2\epsilon\}$, and define a constant τ to be the nearest of the left of σ such that $\Phi_\tau(X^N) \le \epsilon$ but $\Phi_t(X^N) > \epsilon$ for $t \in (\tau, \sigma]$. By the assumption (A) we notice that on the set $\{\tau \le T\}$, we have $\epsilon = \Phi_\sigma - \Phi_\tau$. Furthermore, on the set $\{\sigma \le T\} \cap_{i=1}^4 B_i$, as N big enough, we have

$$\epsilon = 2\epsilon - \epsilon \le |\Phi_\sigma(X^N) - \Phi_\tau(X^N)| = |-c_N\int_\tau^\sigma b^n(s)ds + G^N(\sigma) - G^N(\tau)| \le$$

$$\max(I(\sigma-\tau<a)|-c_N\int_\tau^\sigma b^n(s)ds + G^N(\sigma) - G^N(\tau)|,$$

$$I(\sigma-\tau\ge a)|-c_N\int_\tau^\sigma b^n(s)ds + G^N(\sigma) - G^N(\tau)|) \le$$

$$\max(\epsilon/2, \max(-c_N ab + 2r, 0)).$$

By the assumption (C), there are large enough N such that $\epsilon \le \epsilon/2$. This contradicts to the condition $\sigma \le T$, on the set $\cap_{i=1}^4 B_i$, hence we have

$$\lim_N P(\{\sigma \le T\} \cap_{i=1}^4 B_i) = 0.$$

By (4.26), we can get $\lim_N P(\sigma \le T) = 0$, which means that

$$\lim_N P(\sup_{\alpha\le t\le T} \Phi_t(X^N) \ge 2\epsilon) = 0.$$

5 Proof of the Main Results

5.1 Proof of Theorem 2.1

By the definition of $V_i^N(t)$ (2.6), (2.7), and (2.9), we have

$$V_i^N(t) = -\lambda_i \int_0^t \sum_{j=1}^k [\Phi_s(V_j^N + \sqrt{N}x_j) - \Phi_s(\sqrt{N}x_j)]ds + \sqrt{N}m_i^N(t) \tag{5.27}$$

for $1 \le i \le r$. By the fact that the jumps of the martingale $\sqrt{N}m_i^N(t)$ are no more than $N^{-1/2}$ and the fact that (2.4) and Corollary 4.1 imply

$$\lim_{N\to\infty} P(\sup_{t\le T}|<\sqrt{N}m_i^N>_t - \int_0^t [\lambda_i(1-q(s)) + \mu_i]ds| > \epsilon) = 0,$$

the second term of the right-hand side in (5.27) tends to a Gaussian diffusion process with drift zero and predictable quardratic variation $\int_0^t(\lambda_i(1-q(s))+\mu_i)ds$ in weak sense. We can apply [Theorem 8.3.1 in [9]] to get the desired weak convergence. The only thing we need to check is

$$P-\lim_{N\to\infty}|\sup_{t\le T}\int_0^t[\Phi_s(V_i^N+\sqrt{N}x_i)-\Phi_s(\sqrt{N}x_i)-V_i^N(s)I_{[0,\alpha_i]}(s)]ds|=0 \quad (5.28)$$

for $1\le i\le r$.

By the fact that the jumps of $V_i^N(t)$ are no more than $N^{-1/2}$ and the results of Lemma 4.2, we can conclude that $V_i^N(\cdot)$ are tight in the Skorokod space D. By Prohorov theorem [2], we know that there is a subsequence, say $(V_i^{N'}(\cdot))_{N'=1}^{\infty}$, such that the distributions of $V_i^{N'}(\cdot)$ converge to a probability measure, say Q.

By Skorokhod representation theorem [2], there is a probability space, say (Ω',F',P') and D-valued random variables $(\hat{V}_i^{N'}(t))_{N'=1}^{\infty}$ and $\hat{V}_i(t)$ on that probability space Ω' where $\hat{V}_i^{N'}(t)$ and $\hat{V}_i(t)$ have the same distributions as $V_i^{N'}(t)$ and Q respectively. Moreover, we have

$$\lim_{N'\to\infty}\hat{V}_i^{N'}(t)=\hat{V}_i(t)\ P'-a.s.$$

in the Skorokhod topology.

We claim that for any continuous functions $V(t)$ on $[0,\infty)$ satisfying $V(0)=0$, the following results are true:

$$\lim_{N\to\infty}\int_0^t[\Phi_s(V+\sqrt{N}x_i)-\Phi_s(\sqrt{N}x_i)]ds=\int_0^t V(s)I_{[0,\alpha_i]}(s)ds, \quad (5.29)$$

for $t>0$, for $1\le i\le r$. The proof of this claim will be given later. By the results of Theorem 13.29 in [1], we have $\hat{V}_i(t)$ is continuous $P'-a.s.$ and

$$P'-\lim_{N'\to\infty}\sup_{0\le t\le T}|\hat{V}_i^{N'(t)}-\hat{V}_i(t)|=0.$$

Combining the claim and the above results, we can get

$$P-\lim_{N\to\infty}|\sup_{t\le T}\int_0^t[\Phi_s(V_i^N+\sqrt{N}x_i)-\Phi_s(\sqrt{N}x_i)-V_i^N(s)I_{[0,\alpha_i]}(s)]ds|$$

$$=P'-\lim_{N'\to\infty}|\sup_{t\le T}\int_0^t[\Phi_s(\hat{V}_i^{N'}+\sqrt{N'}x_i)-\Phi_s(\sqrt{N'}x_i)-\hat{V}_i^{N'}(s)I_{[0,\alpha_i]}(s)]ds|$$

$$=P'-\lim_{N'\to\infty}|\sup_{t\le T}\int_0^t[\Phi_s(\hat{V}_i+\sqrt{N'}x_i)-\Phi_s(\sqrt{N'}x_i)-\hat{V}_i(s)I_{[0,\alpha_i]}(s)]ds|=0.$$

We get the desired result.

The remaining thing is to prove the claim (5.29). We first consider the following equality

$$\lim_{N'\to\infty}(\Phi_s(V+\sqrt{N'}x_i)-\Phi_s(\sqrt{N'}x_i))$$

$$=\lim_{N'\to\infty}(V(s)-\inf_{u\le s}(V(u)+\sqrt{N'}x_i(u))+\inf_{u\le s}\sqrt{N'}x_i(u)).$$

By Lemma 3.6, the above formula is equal to $V(s)$ as $s\le\alpha_i-\delta$ for all $0<\delta<\alpha_i$, and is equal to 0 as $s>\alpha_i$. With the Lipschitz condition

$$|\Phi_s(V+\sqrt{N}x_i)-\Phi_s(\sqrt{N}x_i)|\le 2\sup_{u\le s}|V(u)|,$$

we can get (5.29) by dominated convergence theorem. The claim is proved.

5.2 Proof of Theorem 2.2

First, we prove that the first part of (A) and (B) of Theorem 2.2. Since $\sqrt{N}x_i(s)$ is strictly positive on the interval $[0,\alpha_i-\delta]$ by Lemma 3.1, and the function

$$\sqrt{N}|q_i^N(t)-x_i^N(t)|=|\inf_{s\le t}\sqrt{N}x_i^N(s)|$$

tends to zero in probability if $t\le\alpha_i-\delta$ by Lemma 4.2 (notice that $\alpha_r=\infty$), we have the equality

$$\lim_N P(\inf_{s\le\alpha_i-\delta}\sqrt{N}x_i^N(s)\ge\epsilon)=0.$$

To prove the second part of (A) and (C), we want to apply Lemma 4.3. Let $X_i^N(t)=\sqrt{N}x_i^N(t)$, by (2.6) we have

$$X_i^N(t)=-\sqrt{N}\int_0^t b_i^N(s)ds+G_i^N(t),$$

and let

$$c_N=\sqrt{N}\uparrow\infty,\ \text{as } N\uparrow\infty,$$

$$b_i^N(s)=-\lambda_i(1-\sum_{j=1}^k q_j^N(s))+\mu_i,$$

$$G_i^N(t)=\sqrt{N}m_i^N(t),$$

and choose $\alpha=0$ for $r+1\le i\le k$, and $\alpha=\alpha_i+\delta$ for $1\le i\le r-1$.

To check the assumption (B) of Lemma 4.3, we consider the following inequalities

$$\inf_{s\le t}b_i^N(s)\ge\mu_i-\lambda_i>0 \text{ for } r+1\le i\le k,$$

and

$$\inf_{\alpha \le s \le t} b_i^N(s) \ge \lambda_i [\sup_{\alpha \le s \le t} \sum_{j=1}^{k} q_j^N(s) - 1] + \mu_i \text{ for } 1 \le i \le r-1.$$

By Corollary 4.1, as $N \to \infty$, the right-hand side of the above inequality tends in probability to $\lambda_i[q(t)-1]+\mu_i$, which is a strictly positive number by Lemma 3.5.

The assumption (C) of Lemma 4.3 holds since $\sqrt{N} m_i^N(t)$ are square-integrable martingales with jumps no more than $N^{-1/2}$ and with predictable quadratic variations

$$< \sqrt{N} m_i^N >_t = \int_0^t (\lambda_i (1 - \sum_{l=1}^{k} \Phi_s(x_l^N)) + \mu_i) ds.$$

Therefore, the functions $G_i^N(t)$ converge weakly to a continuous Gaussian martingale $G_i(t)$ as $N \to \infty$.

The remaining work is to check the assumption (A). Since $\Phi(X) \ge X$ by Lemma 3.1, we have

$$\lim_N P(\Phi_{\alpha_i + \delta}(X_i) > \epsilon) = 0$$

by Lemma 3.6.

5.3 Proofs of Theorem 2.3; Theorem 2.4

The proofs of Theorem 2.3 and Theorem 2.4 are following from arguments due to Kogan and Liptser [the proof of Theorem 2.2 in [4]]. The difference from the one in Kogan and Liptser's article is the function $q(t)$, the limit of the total amount of the normalized queuing lengths of single servers at time t, and the critical times α_i in Theorem 2.4.

The proof of Theorem 2.3 is the same as the proof of Theorem 2.4 with replacing α_i by 0. We only give the proof of Theorem 2.4. The main idea of the proof is to use induction method on l and the result that the normalized queuing lengths tend to zero in probability as $N \to \infty$ in the light traffic usage regime by applying the Corollary 4.1.

Before giving the proof of Theorem 2.4, we need to introduce some notations and discussions. The reader can find more detailed discussions in the article [4].

Define the processes

$$I_{i,l}(t) = I(Q_i^N(t) = l),$$

$$I_{i,l}(t-) = I(Q_i^N(t-) = l) = \lim_{s \to t^-} I(Q_i^N(s) = l)$$

and the jumps

$$\Delta I_{i,l}(t) = I(Q_i^N(t-) + \Delta Q_i^N(t) = l) - I(Q_i^N(t-) = l)$$

$$= I_{i,l-1}(t-)\Delta A_i(t) + I_{i,l+1}(t-)\Delta D_i(t) + I_{i,l}(t-)(1-\Delta A_i(t))(1-\Delta D_i(t)) - I_{i,l}(t-),$$

where $I(B)$ is the indicator of the set B, and $\Delta X(t) = X(t) - X(t-)$ for some process $X(t)$. Since the jumps of $A_i(t)$ and $D_i(t)$ are disjoint, and $\Delta D_i(t) = I(Q_i^N(t-) > 0)\Delta\Pi_i(t)$, we can get

$$I_{i,l}(t) = I_{i,l}(s) + \int_s^t (I_{i,l-1}(u-) - I_{i,l}(u-))dA_i(u)$$

$$+ \int_s^t (I_{i,l+1}(u-) - I_{i,l}(u-))I(Q_i^N(u) > 0)d\Pi_i(u).$$

We can get the following equality by introducing the compensators $A_i^p(t)$ and $\Pi_i^p(t)$ of the processes $A_i(t)$ and $\Pi_i(t)$ respectively

$$I_{i,l}(t) = I_{i,l}(s) + \int_s^t (I_{i,l-1}(u-) - I_{i,l}(u-))dA_i^p(u)$$

$$+ \int_s^t (I_{i,l+1}(u-) - I_{i,l}(u-))I(Q_i^N(u) > 0)d\Pi_i^p(u)$$

$$+ M_{i,l}(t) - M_{i,l}(s),$$

with the local square-integrable martingale

$$M_{i,l}(t) = \int_0^t (I_{i,l-1}(u-) - I_{i,l}(u-))d(A_i(u) - A_i^p(u))$$

$$+ \int_0^t (I_{i,l+1}(u-) - I_{i,l}(u-))I(Q_i^N(u) > 0)d(\Pi_i(u) - \Pi_i^p(u)).$$

Furthermore, we can get

$$I_{i,l}(t) = I_{i,l}(s) + \int_s^t [\lambda_i(1 - N^{-1}\sum_{j=1}^k Q_j^N(u)(I_{i,l-1}(u) - I_{i,l}(u))$$

$$+\mu_i(I_{i,l+1}(u) - I_{i,l}(u))I(Q_i^N(u) > 0)]du + M_{i,l}(t) - M_{i,l}(s), \tag{5.30}$$

and the quadratic variation of the martingale $M_{i,l}(t)$

$$< M_{i,l} >_t = N \int_0^t (\lambda_i(1 - N^{-1}\sum_{j=1}^k Q_j^N(u)(I_{i,l-1}(u) + I_{i,l}(u))$$

$$+\mu_i(I_{i,l+1}(u) + I_{i,l}(u))I(Q_i^N(u) > 0))du. \tag{5.31}$$

The reader can find more detailed derivation in article [4].

Proof of Theorem 2.4. First, we mention that if the integral

$$\int_{\alpha_i}^T f(t)I(Q_i^N = l)dt$$

tends to (as $N \to \infty$) a deterministic limit in probability, then we have

$$\lim_N E \int_{\alpha_i}^T f(t) I(Q_i^N = l) dt = \lim_N \int_{\alpha_i}^T f(t) P(Q_i^N = l) dt.$$

We shall prove that

$$P - \lim_N \int_{\alpha_i}^T f(t) I(Q_i^N(t) = l) dt = \int_{\alpha_i}^T f(t)(1 - \rho_i(t))\rho_i^l(t) dt, \text{ for all } l = 0, 1, ..., \tag{5.32}$$

for any smooth function $f(t)$, to get the result.

We start from $l = 0$. Denote the compensator of the process $\Pi_i(t)$ to be $\Pi_i^p(t)$. Since $\Pi_i(t)$ is a Poisson process with intensity $N\mu_i$, the compensator satisfies $d\Pi_i^p(t) = N\mu_i dt$. By (2.3) we get

$$q_i^N(t) = \int_{\alpha_i}^T (\lambda_i(1 - \sum_{l=1}^k q_l^N(t)) - \mu_i) dt + m_i^N(t)$$

$$+ \int_{\alpha_i}^T I(Q_i^N(t) = 0)\mu_i dt + N^{-1} \int_{\alpha_i}^T I(Q_i^N(t-) = 0) d(\Pi_i(t) - \Pi_i^p(t)). \tag{5.33}$$

For any smooth function $f(t)$, the Itô integral $\int_{\alpha_i}^T f(t) dq_i^N(t)$ with respect to the semimartingale $q_i^N(t)$ is well-defined [9]. By (5.33) we get

$$\int_{\alpha_i}^T f(t) I(Q_i^N(t) = 0) dt = -\int_{\alpha_i}^T f(t)(\frac{\lambda_i}{\mu_i}(1 - \sum_{l=1}^k q_l^N(t)) - 1) dt$$

$$-\mu_i^{-1} \int_{\alpha_i}^T f(t) dm_i^N(t) - \mu_i^{-1} \int_{\alpha_i}^T f(t) I(Q_i^N(t-) = 0) N^{-1} d(\Pi_i(t) - \Pi_i^p(t))$$

$$+\mu_i^{-1} \int_{\alpha_i}^T f(t) dq_i^N(t). \tag{5.34}$$

The first term in the right-hand side of (5.34) tends in probability ,as $N \to \infty$, to the following limit by Corollary 4.1 and Remark 4.2

$$P - \lim_N \int_{\alpha_i}^T f(t)(1 - \frac{\lambda_i}{\mu_i}(1 - \sum_{l=1}^k \Phi_t(x_i))) dt$$

$$= \int_{\alpha_i}^T f(t)(1 - \frac{\lambda_i}{\mu_i}(1 - q(t))) dt$$

$$= \int_{\alpha_i}^T f(t)(1 - \rho_i(t)) dt.$$

We shall prove that the remaining terms in the right-hand side of (5.34) tend to zero in probability. The fact that the second term tends to zero in probability follows from

$$E(\int_{\alpha_i}^{T} f(t)dm_i^N(t))^2 = E\int_{\alpha_i}^{T} f^2(t)d < m_i^N >_t$$

$$= N^{-1}E\int_{\alpha_i}^{T} f^2(t)(\lambda_i(1-\sum_{l=1}^{k} q_l^N(t)) + \mu_i)dt$$

$$\leq N^{-1}(\lambda_i+\mu_i)\int_{\alpha_i}^{T} f^2(t)dt,$$

while the fact that the third term tends to zero in probability is implied by

$$E(\int_{\alpha_i}^{T} f(t)I(Q_i^N(t-)=0)N^{-1}d(\Pi_i(t)-\Pi_i^p(t)))^2$$

$$= N^{-2}E(\int_{\alpha_i}^{T} f^2(t)I(Q_i^N(t-)=0)d\Pi_i^p(t)) \leq N^{-1}\mu_i\int_{\alpha_i}^{T} f^2(t)dt.$$

Using the integration by parts on the fourth term of (5.34), we get

$$\int_{\alpha_i}^{T} f(t)dq_i^N(t) = f(T)q_i^N(T) - \int_{\alpha_i}^{T} \frac{df(t)}{dt}q_i^N(t)dt,$$

which tends to zero in probability by Remark 4.1.

Assume that (5.32) holds for some $l \geq 1$, we want to show that (5.32) still holds with replacing l by $l+1$. We first notice by (5.30) that

$$N^{-1}\int_{\alpha_i}^{T} f(t)dI_{i,l}(t) = \int_{\alpha_i}^{T} f(t)[\lambda_i(1-\sum_{j=1}^{k} q_j^N(t))(I_{i,l-1}(t)$$

$$-I_{i,l}(t)) + \mu_i(I_{i,l+1}(t)-I_{i,l}(t))]dt + N^{-1}\int_{\alpha_i}^{T} f(t)dM_{i,l}(t).$$

We can see from the above equation that

$$\int_{\alpha_i}^{T} f(t)I_{i,l+1}(t)dt$$

$$= -\int_{\alpha_i}^{T} f(t)[\frac{\lambda_i}{\mu_i}(1-\sum_{j=1}^{k} q_j^N(t))(I_{i,l-1}(t)-I_{i,l}(t)) - I_{i,l}(t)]dt$$

$$-(N\mu_i)^{-1}\int_{\alpha_i}^{T} f(t)dM_{i,l}(t) + (N\mu_i)^{-1}\int_{\alpha_i}^{T} f(t)dI_{i,l}(t). \qquad (5.35)$$

As before we want to prove that the first term in the right-hand side of (5.35) tends, as $N \to \infty$, in probability to the right-hand side of (5.32) with replacing l by $l+1$. Besides, we shall prove that the remaining terms in the right-hand side of (5.35) tend to zero in probability. From Corollary 4.1 and the induction hypothesis that the formula (5.32) holds for some l, we can get

$$\int_{\alpha_i}^{T} f(t)[\frac{\lambda_i}{\mu_i}(1-\sum_{j=1}^{k} q_j^N(t))(I_{i,l-1}(t)-I_{i,l}(t))-I_{i,l}(t)]dt$$

$$= \int_{\alpha_i}^{T} f(t)[\frac{\lambda_i}{\mu_i}(1-q(t))(1-\rho_i(t))(\rho_i^{l-1}(t)-\rho_i^l(t))-(1-\rho_i(t))\rho_i^l(t)]dt$$

$$= \int_{\alpha_i}^{T} f(t)(1-\rho_i(t))\rho_i^{l+1}(t)dt.$$

By (5.31) the second term in the right-hand side of (5.35) satisfies

$$E[(N\mu_i)^{-1}\int_{\alpha_i}^{T} f(t)dM_{i,l}(t)]^2 \le (N\mu_i)^{-2}E\int_{\alpha_i}^{T} f^2(t)d < M_{i,l} >_t$$

$$= N^{-1}E\int_{\alpha_i}^{T} f^2(t)(\lambda_i(1-\sum_{j=1}^{k} q_j^N(t))(I_{i,l-1}(t)+I_{i,l}(t))+\mu_i(I_{i,l+1}(t)+I_{i,l}(t)))dt$$

$$\le 2(\lambda_i+\mu_i)N^{-1}\int_{\alpha_i}^{T} f^2(t)dt,$$

which tends to zero in probability. Using the integration by parts on the third term in the right-hand side of (5.32), we get

$$(N\mu_i)^{-1}\int_{\alpha_i}^{T} f(t)dI_{i,l}(t) = (N\mu_i)^{-1}(f(T)I_{i,l}(T)-\int_{\alpha_i}^{T}\frac{df(t)}{dt}I_{i,l}(t)dt),$$

which tends to zero in probability by Remark 4.1. The theorem is proved.

Acknowldgement

The author is sincerely grateful to her advisor N. V. Krylov for his very stimulating suggestion.

References

[1] L. Breiman, *Probability*, (Addison-Wesley, 1968).

[2] S. N. Ethier, and T. G. Kurtz, *Markov Processes*, (John Wiley and Sons).

[3] C. Knessl and C. Tier, *Asymptotic expansion for large closed queueing networks*, J. ACM 37 (1990) 144-174.

[4] Y. Kogan and R. S. Liptser, *Limit non-stationary behavior of large closed queueing networks with bottlenecks*, Queueing Sysytems 14 (1993) 33-55.

[5] Y. Kogan, *Another approach to asymptotic expansion for large closed queueing networks*, Oper. Res. Lett. 11 (1992) 317-321.

[6] Y. Kogan and A. Birman, *Asymptotics analysis of closed queueing networks with bottlenecks*, Proc. Int. Conf. on *Performance of Distributed Systems and Integrated Communication Networks*, eds. T. Hasegawa and H. Takagi and Y. Takahashi (Kyoto, 1991) pp.237-252.

[7] Y. Kogan and R. S. Liptser and A. V. Smorodinskii, *Gaussian Diffusion approximation of closed Markov model of computer networks*, Problem Inform. Transmission 22 (1986) 38-51.

[8] S. S. Lavenberg, *Closed multichain product form queueing networks with large population sizes*, Applied Probability-Computer Science, The Interface, eds. R. L. Disney and T. J. Ott, Vol.1 (Birkhauser, Boston, 1982) pp.219-249.

[9] R. S. Liptser and A. N. Shiryayev, *Theory of Martingales*, (Kluwer, Dordrecht, 1989).

[10] J. Mackenna and D. Mitra and K. G. Ramakrishnan, *A class of closed Markovian queueing networks: Integral representation, asymptotic expansions and generalizations*, Bell Syst. Tech. J. 60 (1981) 599-641.

[11] B. Pittel, *Closed exponential networks of queues with saturation: the closed Jackson-type stationary distributions and its asymptotic analysis*, Math. Oper. Res. 6 (1979) 357-378.

[12] H. Tanaka, *Stochastic equations with reflecting boundary conditions in convex regions*, Hiroshima Math. J. (1979) 163-177.

[13] W. Whitt, *Open and closed models for networks of queues* , AT&T Bell Lab. Tech. J. (1984) 1911-1979.