The Annals of Applied Probability
1995, Vol. 5, No. 1, 97–120

STATE-DEPENDENT BENEŠ BUFFER MODEL WITH FAST LOADING AND OUTPUT RATES

By Y. Kogan, R. Liptser and M. Shenfild

Technion—Israel Institute of Technology, Tel Aviv University and Technion—Israel Institute of Technology

We consider a state-dependent generalization of the exponential Beneš model of single-source buffer system in which the source process consists of alternating transmission and idle periods. Martingale methods are applied for analyzing limit nonstationary behavior of the buffer content process, when the buffer is loaded and depleted, with rates proportional to a large parameter N. Depending on traffic conditions, defined by parameters of the model, different types of approximations are established for the buffer content. We show that in heavy traffic the buffer content grows linearly in N, whereas the deviations of the order $\sqrt{N}$ from the deterministic limit are approximated by the Gaussian diffusion process. In moderate traffic the buffer content grows as $\sqrt{N}$, and the normalized buffer content is approximated by a diffusion process with reflection at zero. In the case of normal traffic, we show that the buffer utilization tends to the ratio of "the input-to-output rate." Moreover, we show that the main contribution to the utilization comes from arbitrary small buffer content.

1. Introduction, method and main result. We consider a model of a single-source buffer system in which the source process consists of alternating transmission and idle periods. The buffer forms the interface between input and output (source and sink) processes. The information is received at one rate from a given source and retransmitted at another rate to a given sink. Arriving messages are characterized by a probability distribution governing their sizes, which corresponds to the service time distribution in the usual queueing system. In general, it is necessary to assume that it takes time to load or transmit a message, a process that typically requires a time interval proportional to the message size or length. The constant of proportionality is the source transmission rate. Thus, a size of arriving messages will determine both the time it takes to load them and the space they occupy in the buffer. We consider loading with flow-through that refers to the process whereby the output of an arriving item begins the moment that it has nothing ahead of it, as opposed to having to wait until the entire message has been loaded. The sizes of transmitted message are assumed to be random variables, having mean $1/\mu$. After completion of loading, the next message can arrive in random time interval with mean $1/\Lambda$. For the sake of simplicity we consider a buffer with infinite capacity. If the transmission rate is R and

Received October 1993; revised June 1994.
AMS 1991 *subject classifications*. 60K25, 60K30.
Key words and phrases. Buffer content, fluid and diffusion approximation.

output rate is $C < R$, then buffer content grows with rate $R - C$ during each loading period and decreases with rate C during each idle period until the buffer becomes empty. The principal objective of the analysis is the buffer-occupancy process, that is, the amount of the buffer occupied by messages as a function of time. A steady-state analysis of such a model has been done by Beneš [1, 2] in the case when message sizes and arriving times are generally distributed independent random variables.

In this paper, following [2], we consider a state-dependent generalization of an exponential version of the Beneš model. In other words, message sizes are exponentially distributed whereas the idle periods are conditionally exponentially distributed with rates dependent on the buffer content. The output rate is also assumed to be state dependent.

To reflect the properties of modern high-speed communication systems, all rates are to be assumed in proportion to a large parameter N. More precisely, we assume that the rates have the following representation:

$$R = Nr, \qquad \Lambda = N\lambda(q_t/N), \qquad C = Nc(q_t/N),$$

where q_t is the buffer content at time t, whereas $\lambda(\cdot)$ and $c(\cdot)$ are smooth and positive functions.

Let ξ_t be a process that takes the value 1 when a message is loaded and 0 otherwise. If a message is loaded and the buffer is not empty, then the buffer content grows with rate $Nr\xi_t - Nc(q_t/N)$. The nonempty buffer decreases with rate $Nc(q_t/N)$. Because $c(0)$ is assumed to be positive, it is convenient to use $Nc(q_t/N)I(q_t > 0)$ for the decreasing rate, where

$$I(z > 0) = \begin{cases} 1, & z > 0, \\ 0, & z = 0. \end{cases}$$

Thus, the buffer content is described by the differential equation

$$\frac{dq_t}{dt} = N\left[r\xi(t) - c\left(\frac{q_t}{N}\right)I(q_t > 0)\right]. \tag{1.1}$$

The buffer content q_t depends on N. Our goal is to study asymptotic behavior of the process q_t as $N \to \infty$. So it is important to find a bifurcation point that allows different types of asymptotic regimes to be distinguished. To explain how to find the bifurcation point, assume that at time zero the buffer is empty: $q_0 = 0$. For asymptotic analysis, it is convenient to operate with normalized buffer content

$$x_t^N = \frac{q_t}{N}. \tag{1.2}$$

Using (1.1) and taking into account that $I(q_t > 0) \equiv I(x_t^N > 0)$, we get the differential equation for x_t^N

$$\frac{dx_t^N}{dt} = r\xi_t - c(x_t^N)I(x_t^N > 0) \tag{1.3}$$

subject to $x_0^N = 0$. The function $c(z)$ is assumed to be positive and bounded: $c(z) < r$. Put

$$y_t^N = \int_0^t \left[c\xi_s - c\left(x_s^N\right)\right] ds$$

and

$$\psi_t = \int_0^t c(0) I\left(x_s^N = 0\right) ds.$$

Because $I(x_s^N = 0) \equiv I(y_s^N \le 0,\ dy_s^N/ds = -c(0))$ the function ψ_t is defined as $\psi_t = -\inf_{s \le t} y_s^N$. Then

$$x_t^N = y_t^N - \inf_{s \le t} y_s^N = \Phi_t\left(y^N\right). \tag{1.4}$$

It is known [10] that the function $\Phi_t(V)$, $t \ge 0$, is the normal reflection at zero of continuous (or right continuous, having limits from the left) function V_t with $V_0 = 0$. Thus, x_t^N is the normal reflection of y_t^N, and what is more, y_t^N is the solution of a past-dependent differential equation

$$\frac{dy_t^N}{dt} = r\xi_t - c\left(\Phi_t\left(y^N\right)\right).$$

The central role in asymptotic analysis plays a stochastic equation for the loading process ξ_t. For $\lambda(z) \equiv \lambda$, the process ξ_t is Markovian with intensities $N\lambda$ and $N\mu r$ of transitions $0 \to 1$ and $1 \to 0$, respectively. It is known ([8], Volume 1, Chapter 9, Lemma 9.2) that a Markovian process with countable space of states obeys a semimartingale decomposition, which is

$$\xi_t = \xi_0 + \int_0^t N\left[\lambda(1 - \xi_s) - \mu r \xi_s\right] ds + m_t,$$

where m_t is a square integrable martingale. For our past-dependent model, ξ_t is a non-Markovian process. Nevertheless, it obeys the same type semimartingale decomposition [with replacing λ on $\lambda(x_s^N)$]:

$$\xi_t = \xi_0 + \int_0^t N\left[\lambda\left(x_s^N\right)(1 - \xi_s) - \mu r \xi_s\right] ds + m_t,$$

which implies

$$\int_0^t \xi_s\, ds - \int_0^t \frac{\lambda\left(x_s^N\right)}{\lambda\left(x_s^N\right) + \mu r} ds = -\frac{M_t}{N} + \frac{S_t}{N},$$

where M_t and S_t are square integrable martingale and semimartingale, respectively [see (2.21) and (2.22)]. Then y_t^N satisfies

$$y_t^N - \int_0^t \left(\frac{r\lambda\left(\Phi_s\left(y^N\right)\right)}{\lambda\left(\Phi_s\left(y^N\right)\right) + \mu r} - c\left(\Phi_s\left(y^N\right)\right) \right) ds = -\frac{rM_t}{N} + \frac{rS_t}{N}.$$

We show (Lemma 3.1) that the right-hand side of this equality converges to zero and so an approximation for y_t^N is given by solution of the past-dependent differential equation

$$\frac{dy_t}{dt} = \frac{r\lambda(\Phi_t(y))}{\lambda(\Phi_t(y)) + \mu r} - c(\Phi_t(y))$$

subject to $y_0 = 0$. It obeys two types of solutions, depending on parameter

$$\rho = \frac{r\lambda(0)}{(\lambda(0) + \mu r)c(0)}, \tag{1.5}$$

namely,

$$y_t = \begin{cases} \int_0^t \left[\frac{r\lambda(y_s)}{\lambda(y_s) + \mu r} - c(y_s) \right] ds, & \rho > 1, \\ c(0)[\rho - 1]t, & \rho \leq 1. \end{cases}$$

The parameter ρ, which could be named "the input-to-output rate," defines the bifurcation point. It is natural to expect that different types of asymptotics hold under the respective conditions

$$\rho < 1, \qquad \rho = 1, \qquad \rho > 1.$$

These types are referred to as normal, moderate and heavy traffic, respectively.

In heavy traffic ($\rho > 1$), the function y_t is a solution of equation

$$y_t = \int_0^t \left[\frac{r\lambda(y_s)}{\lambda(y_s) + \mu r} - c(y_s) \right] ds.$$

It is positive for any $t > 0$. Due to (1.4) and $-\inf_{s \leq t} y_s = 0$, x_t^N converges to y_t.

For moderate traffic ($\rho = 1$), x_t^N converges to $\Phi_t(y) \equiv 0$, that is, asymptotic information on the buffer content is loosened. Therefore, another type of normalization is used:

$$X_t^N = \frac{q_t}{\sqrt{N}} = \sqrt{N} x_t^N. \tag{1.6}$$

The process X_t^N is also obtained from the process $Y_t^N = \sqrt{N} y_t^N$ by normal reflection at zero: $X_t^N = \Phi_t(Y^N)$. By virtue of the definition of y_t^N, we get

$$Y_t^N - \sqrt{N} \int_0^t \left(\frac{r\lambda(\Phi_s(y^N))}{\lambda(\Phi_s(y^N)) + \mu r} - c(\Phi_s(y^N)) \right) ds = -\frac{rM_t}{\sqrt{N}} + \frac{rS_t}{\sqrt{N}}.$$

Now, only $rS_t / \sqrt{N}$ converges to zero (Lemma 3.1). Therefore, the limit behavior for Y_t^N is the same as for

$$\overline{Y}_t^N = \sqrt{N} \int_0^t \left(\frac{\lambda(\Phi_s(y^N))}{\lambda(\Phi_s(y^N)) + \mu r} - c(\Phi_s(y^N)) \right) ds - \frac{rM_t}{N}.$$

To find it, we use the diffusion approximation result for semimartingales ([9], Chapter 8, Section 3). Roughly speaking, in our case only two conditions have to be checked: convergence of the predictable quadratic variation $\langle rM/N\rangle_t$ of square integrable martingale $rM_t/\sqrt{N}$ and convergence of the drift

$$\sqrt{N}\int_0^t\left(\frac{r\lambda(\Phi_s(y^N))}{\lambda(\Phi_s(y^N)) + \mu r} - c(\Phi_s(y^N))\right) ds.$$

We show that the predictable quadratic variation converges to $2t\mu c^3(0)/\lambda^2(0)$ (Corollary 3.2) and the drift tends to

$$\int_0^t\left(\lambda'(0)[c(0)/\lambda(0)]^2 - c'(0)\right)\Phi_s(Y^N)\, ds$$

(for more details see Section 4).

Thus, the diffusion approximation for Y_t^N is given by solution of past-dependent Itô's equation (w.r.t. a Wiener process W_t):

$$Y_t = \int_0^t\left(\lambda'(0)[c(0)/\lambda(0)]^2 - c'(0)\right)\Phi_s(Y)\, ds + \int_0^t\sqrt{2\mu c^3(0)/\lambda^2(0)}\, dW_s.$$

Then X_t^N obeys the diffusion approximation for the limit $\Phi_t(Y)$ [in the Appendix it is shown that $\Phi_t(Y)$ has the distribution of absolute value of the Gaussian diffusion Z_t].

Asymptotic analysis for normal traffic ($\rho < 1$) is different from that for both heavy and moderate traffic because x_t^N and X_t^N converge to zero [the latter by the collapse property: $Y_t^N \to -\infty$ (see Lemma 4.1)]. The following relations are implied by convergence x_t^N to zero and (1.3):

$$\int_0^t r\xi_s\, ds \to c(0)\rho t$$

and

$$\int_0^t I(q_s > 0)\, ds \to \rho t.$$

Therefore,

$$\int_0^t I(q_s = 0)\, ds \to [1 - \rho]t.$$

The next important asymptotic result is, for any $\varepsilon > 0$,

$$\int_0^t I(q_s \geq \varepsilon)\, ds \to 0,$$

the proof of which is more artificial. It uses semimartingale decomposition for $x_t^N\xi_t$, $x_t^N G(\xi_t)$ with a specially chosen function $G(z)$, and representation

$$\left(X_t^N\right)^2 = 2\int_0^t q_s\left[r\xi_s - c\left(x_s^N\right)\right] ds,$$

which is implied by (1.3).

Summing all results, we arrive at different types of asymptotics (as $N \to \infty$). For any $t > 0$ and $\varepsilon > 0$:

1. $\mathbf{P}(0 < q_t < \varepsilon) \asymp \rho$, $\mathbf{P}(q_t = 0) \asymp 1 - \rho$, $\rho < 1$;
2. $q_t \asymp |Z_t|\sqrt{N}$, $\rho = 1$;
3. $q_t \asymp y_t N$, $\rho > 1$,

where Z_t is a nondegenerate Gaussian random variable and y_t is positive. It is clear that in normal traffic large loading and output rates lead in the limit to a deterministic model, where the buffer is always empty because the output rate is more than the input rate. In type 1, the first relation shows that the fraction of time when the buffer is empty is defined by the input-to-output rate, whereas the second relation shows that the main contribution to the buffer utilization comes from the arbitrary small buffer content.

All these asymptotics are formulated in the following theorem.

THEOREM 1.1. *Let functions $\lambda(z)$ and $c(z)$ be positive and bounded:*

$$0 < \lambda(z) \le \text{const.}, \qquad 0 < c(z) < r$$

and continuously differentiable. Their derivatives $\lambda'(z)$ and $c'(z)$ are bounded and Lipschitz continuous. Then the following three types of asymptotics hold.

1. *Normal traffic ($\rho < 1$): For any $0 \le t_1 < t_2$ and $\varepsilon > 0$,*

$$\lim_N \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathbf{P}(q_s = 0)\, ds = 1 - \rho,$$

$$\lim_N \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \mathbf{P}(0 < q_s \le \varepsilon)\, ds = \rho.$$

2. *Moderate traffic ($\rho = 1$): X_t^N converges weakly (in the local uniform topology) for limit $|Z_t|$, where Z_t is Gaussian diffusion process defined by Itô's equation (w.r.t. a Wiener process W_t):*

$$dZ_t = \big(\lambda'(0)[c(0)/\lambda(0)]^2 - c'(0)\big) Z_t\, dt + \sqrt{2\mu c^3(0)/\lambda^2(0)}\, dW_t \tag{1.7}$$

subject to $Z_0 = 0$.

3. *Heavy traffic ($\rho > 1$): For any $T > 0$,*

$$\mathbf{P} - \lim_N \sup_{t \le T} |x_t^N - y_t| = 0,$$

where y_t is defined by the differential equation

$$\frac{dy_t}{dt} = \frac{r\lambda(y_t)}{\lambda(y_t) + \mu r} - c(y_t) \tag{1.8}$$

subject to $y_0 = 0$.

REMARKS.

1. Because $c(0) > 0$, a diffusion parameter in (1.7) is positive and so for any $t > 0$, Gaussian random variable Z_t, having $\mathbf{E}Z_t = 0$, is nondegenerate: $\mathbf{P}(|Z_t| > 0) = 1$ for any $t > 0$.
2. Due to $\rho > 1$, we have
$$\left(\frac{dy_t}{dt}\right)\bigg|_{y_t=0} > 0$$
and so for any $t > 0$, x_t is positive.

For heavy traffic, the diffusion approximation for centered buffer content $x_t^N - y_t$ holds too. Under assumptions of Theorem 1.1, $X_t^N = \sqrt{N}(x_t^N - y_t)$ converges eakly (in the local uniform topology) to the Gaussian diffusion process X_t defined by Itô's equation (w.r.t. a Wiener process W_t):

$$(1.9)\quad X_t = \int_0^t \left[\frac{r^2\lambda'(y_s)}{[\lambda(y_s) + \mu r]^2} - c'(y_s)\right] X_s\, ds + \int_0^t \sqrt{\frac{2\mu r \lambda(y_s)}{[\lambda(y_s) + \mu r]^3}}\, dW_s,$$

where y_t is defined by (1.8). The proof is similar to Lemma 4.2 and is omitted.

The main contribution of this paper is twofold. First, we obtain asymptotic results, where a direct approach fails. Although the pair (q_t, ξ_t) forms a Markov process, state-dependent rates and the degeneracy of the first component make it analytically intractable. From a more general point of view we are interested in the asymptotic properties of the pair (x_t^N, ξ_t) [or (X_t^N, ξ_t)], where ξ_t is the "fast" component, depending on x_t^N, whereas the "slow" component x_t^N averages the influence of ξ_t. This fact leads to the Bogolubov averaging principle and to fluid and diffusion approximations. For the diffusion case, the Bogolubov averaging principle as well as the second order approximation (diffusion approximation) have been studied in [4]. For a discontinuous process in the semimartingale setting, the corresponding result for the Bogolubov averaging can be found in [7]. In contrast to [4] and [7], the fast component in our model is discontinuous, whereas the slow component is reflected at the zero degenerate process. However, both approximations (fluid and diffusion) are also valid for this model.

Our approach exploits only the fact that all examined processes are special semimartingales that obeying the decomposition "predictable drift + square integrable martingale." In Section 2, all required decompositions of such kind and formulas for "predictable quadratic variations" of martingales, which are involved in proof Theorem 1.1, are presented. The method of proving Theorem 1.1 is based on some asymptotic relations that are derived from semimartingale decompositions and asymptotic properties of predictable quadratic variations. These results are gathered in Sections 3 and 5. The diffusion approximation for $\sqrt{N}y_t^N$ is given in Section 4. In Section 6, we derive the statements of Theorem 1.1 from previous results. In the Appendix, we give a solution for Gaussian diffusion reflected at zero.

The Markovian property is not used anywhere in this paper. Therefore, the results can be generalized to the case of arbitrary transmitted message size. However, in this case, all calculations are more tedious and their volume will increase.

2. Semimartingale decompositions.

1. Throughout this paper, the following identities are used: $I(q_t > 0) \equiv I(x_t^N > 0)$ and $X_t^N \equiv \sqrt{N} x_t^N$, $(X_t^N)^2 \equiv q_t^2/N$.

By assumption of Theorem 1.1, $c(0) > 0$ and so function $c(z)I(z > 0)$ is discontinuous at point $z = 0$, and what is more, ξ_t depends on x_t^N. Nevertheless, differential equation (1.3) has the unique solution that is convenient to consider separately on time intervals: $\{t\colon x_t^N > 0,\ \xi_t = 1\}$, $\{t\colon x_t^N = 0,\ \xi_t = 1\}$, $\{t\colon x_t^N > 0,\ \xi_t = 0\}$ and $\{t\colon x_t^N = 0,\ \xi_t = 0\}$. From (1.3), it follows that

$$(2.1)\qquad \left(x_t^N\right)^2 = 2\int_0^t x_s^N \left[r\xi_s - c\left(x_s^N\right)\right] ds$$

and then

$$(2.2)\qquad \left(X_t^N\right)^2 = 2\int_0^t q_s \left[r\xi_s - c\left(x_s^N\right)\right] ds.$$

2. Here, we derive semimartingale decomposition for ξ_t. We assume that some stochastic basis (see [9]) is fixed and all random processes are defined on it. We do not define concretely a filtration on this basis and only assume that $N\lambda(x_t^N)$ and $N\mu r$ are intensities of counting processes A_t and B_t generated by positive and negative jumps of ξ_t. As any right continuous random process values in the space $(\{0\},\{1\})$, ξ_t can be defined by an Itô type equation

$$(2.3)\qquad \xi_t = \xi_0 + \int_0^t (1 - \xi_{s-})\, dA_s - \int_0^t \xi_{s-}\, dB_s,$$

where ξ_{s-} is the limit from the left. In accordance with given intensities, counting processes A_t and B_t have compensators

$$(2.4)\qquad \begin{aligned} A_t^p &= \int_0^t N\lambda\left(x_s^N\right) ds, \\ B_t^p &= N\mu r t, \end{aligned}$$

respectively, that is (see, e.g., [5] or [8]), the processes $A_t - A_t^p$ and $B_t - B_t^p$ are square integrable martingales, having the predictable quadratic variations

$$(2.5)\qquad \begin{aligned} \langle A - A^p \rangle_t &\equiv A_t^p, \\ \langle B - B^p \rangle_t &\equiv B_t^p, \\ \langle A - A^p, B - B^p \rangle_t &\equiv 0 \end{aligned}$$

(the last equality follows from the disjointness of jumps A_t and B_t). The martingales $A_t - A_t^p$ and $B_t - B_t^p$ have paths of the local bounded variation and so the Stieltjes integral $\int_0^t(1-\xi_{s-})d(A_s - A_s^p)$ [and $\int_0^t \xi_{s-}d(B_s - B_s^p)$] coincides with the Itô one. Then the process

$$m_t = \int_0^t (1-\xi_{s-})d(A_s - A_s^p) - \int_0^t \xi_{s-}d(B_s - B_s^p) \tag{2.6}$$

is a square integrable martingale whose predictable quadratic variation is given by the formula

$$\begin{aligned}\langle m\rangle_t &= \int_0^t (1-\xi_{s-})d\langle A - A^p\rangle_s + \int_0^t \xi_{s-}d\langle B - B^p\rangle_s \\ &= N\int_0^t \big[(1-\xi_s)\lambda(x_s^N) + \xi_s \mu r\big]\,ds.\end{aligned} \tag{2.7}$$

Hence, a semimartingale decomposition for ξ_t is

$$\xi_t = \xi_0 + N\int_0^t \big[(1-\xi_s)\lambda(x_s^N) - \xi_s\mu r\big]\,ds + m_t. \tag{2.8}$$

3. Let $G(z)$, $-1 \le z \le 2$, be a continuous function. A random process $G(\xi_t)$ has values in the set ($\{G(0), G(1)\}$) and has paths of the local bounded variation. So, it is a semimartingale. We need its semimartingale decomposition for the proof of statement 1 (normal traffic) of Theorem 1.1.

Let

$$G(\xi_t) = G(\xi_0) + \tilde{G}_t(\xi) + P_t(\xi) \tag{2.9}$$

be its semimartingale decomposition with a predictable drift $\tilde{G}_t(\xi)$ and a local martingale $P_t(\xi)$. Applying Itô's formula to $G(\xi_t)$ and taking into account (2.3), we find

$$\begin{aligned}G(\xi_t) = G(\xi_0) &+ \int_0^t [G(\xi_{s-}+1) - G(\xi_{s-})]\,dA_s \\ &+ \int_0^t [G(\xi_{s-}-1) - G(\xi_{s-})]\,dB_s.\end{aligned} \tag{2.10}$$

Then

$$\begin{aligned}\tilde{G}_t(\xi) = N\int_0^t \{&[G(\xi_s+1) - G(\xi_s)]\lambda(x_s^N) \\ &+ [G(\xi_s - 1) - G(\xi_s)]\mu r\}\,ds\end{aligned} \tag{2.11}$$

and

$$\begin{aligned}P_t(\xi) = \int_0^t &[G(\xi_{s-}+1) - G(\xi_{s-})]d(A_s - A_s^p) \\ &+ \int_0^t [G(\xi_{s-}-1) - G(\xi_{s-})]d(B_s - B_s^p).\end{aligned}$$

Evidently, $P_t(\xi)$ is a square integrable martingale with predictable quadratic variation

$$\langle P(\xi)\rangle_t = N\int_0^t\Big\{\big[G(\xi_s+1)-G(\xi_s)\big]^2\lambda(x_s^N) + \big[G(\xi_s-1)-G(\xi_s)\big]^2\mu r\Big\}\,ds. \tag{2.12}$$

4. Here, we give semimartingale decompositions for $x_t^N\xi_t$ and $x_t^N G(\xi_t)$, which also are used in the proof of statement 1 of Theorem 1.1.

Applying Itô's formula to $x_t^N\xi_t$ and taking into account (1.3) and (2.8), we arrive at the semimartingale decomposition

$$\begin{aligned} x_t^N\xi_t &= \int_0^t q_s\big[(1-\xi_s)\lambda(x_s^N) - \xi_s\mu r\big]\,ds \\ &\quad + \int_0^t \xi_s\big[r\xi_s - c(x_s^N)I(q_s>0)\big]\,ds \\ &\quad + \int_0^t x_s^N\,dm_s. \end{aligned} \tag{2.13}$$

The last integral on the right-hand side of (2.13) forms a square integrable martingale whose predictable quadratic variation is $\langle \int_0^\cdot x_s^N\,dm_s\rangle_t = \int_0^t (x_s^N)^2 d\langle m\rangle_s$. Then, due to (2.7), we get

$$\left\langle \int_0^\cdot x_s^N\,dm_s\right\rangle_t = \int_0^t (X_s^N)^2\big[(1-\xi_s)\lambda(x_s^N) - \xi_s\mu r\big]\,ds. \tag{2.14}$$

Analogously, due to (2.9) and (1.3), by Itô's formulas we find

$$\begin{aligned} x_t^N G(\xi_t) &= \int_0^t x_s^N\,d\tilde{G}_s(\xi) + \int_0^t x_s^N\,dP_s(\xi) \\ &\quad + \int_0^t G(\xi)\big[r\xi_s - c(x_s^N)I(x_s^N>0)\big]\,ds. \end{aligned}$$

Hence, by virtue of (2.10) and (2.11), the following decomposition for $x_t^N G(\xi_t)$ holds:

$$\begin{aligned} x_t^N G(\xi_t) &= \int_0^t q_s\Big\{\big[G(\xi_s+1)-G(\xi_s)\big]\lambda(x_s^N) \\ &\qquad + \big[G(\xi_s-1)-G(\xi_s)\big]\mu r\Big\}\,ds \\ &\quad + \int_0^t G(\xi_s)\big[r\xi_s - c(x_s^N)I(x_s^N>0)\big]\,ds \\ &\quad + \int_0^t x_s^N\,dP_s(\xi). \end{aligned} \tag{2.15}$$

Moreover, a square integrable martingale $\int_0^t x_s^N \, dP_s(\xi)$ has the predictable quadratic variation

$$\left\langle \int_0^{\cdot} x_s^N \, dP_s(\xi) \right\rangle_t = \int_0^t (x_s^N)^2 d\langle P(\xi)\rangle_s$$

$$= \int_0^t (X_s^N)^2 \{[G(\xi_s + 1) - G(\xi_s)]\lambda(x_s^N)\} \, ds \tag{2.16}$$

$$+ \int_0^t (X_s^N)^2 \{[G(\xi_s - 1) - G(\xi_s)]\mu r\} \, ds.$$

5. Time intervals $\{t\colon x_t^N > 0,\ \xi_t = 1\}$, $\{t\colon x_t^N = 0,\ \xi_t = 1\}$, $\{t\colon x_t^N > 0,\ \xi_t = 0\}$ and $\{t\colon x_t^N = 0,\ \xi_t = 0\}$ decrease to zero as $N \to \infty$ and so, due to the discontinuity of $c(x)I(x > 0)$, any kind of asymptotic results would be difficult to get. Thereby parallel to (1.3), we give here another description for x_t^N. For any right continuous process having limits from the left function $V = (V_t)_{t \ge 0}$, $V_0 = 0$, put

$$\Phi_t(V) = V_t - \inf_{s \le t} V_s. \tag{2.17}$$

$\Phi_t(V)$ is Lipschitz continuous in the following sense: $|\Phi_t(V') - \Phi_t(V'')| \le 2\sup_{s \le t} |V_s' - V_s''|$ (see, e.g., [6]). Consider a past-dependent differential equation

$$\frac{dy_t^N}{dt} = r\xi_t - c(\Phi_t(y^N)) \tag{2.18}$$

subject to $y_0^N = 0$. Analyzing the solution of both equations (1.3) and (2.18), one can conclude that

$$x_t^N \equiv \Phi_t(y^N), \tag{2.19}$$

that is, x_t^N is the normal reflection for y_t^N (see [10]). Now, we give more detailed description of y_t^N. Putting

$$\gamma(z) = [\lambda(z) + \mu r]^{-1}, \tag{2.20}$$

define a semimartingale S_t and a square integrable martingale M_t:

$$S_t = \int_0^t \gamma(x_s^N) \, d\xi_s,$$

$$M_t = \int_0^t \gamma(x_s^N) \, dm_s. \tag{2.21}$$

From (2.8) and (2.21), it follows that

$$\int_0^t \xi_s \, ds = \int_0^t \frac{\lambda(x_s^N)}{\lambda(x_s^N) + \mu r} \, ds - \frac{M_t}{N} + \frac{S_t}{N}, \tag{2.22}$$

and so, from (2.18), we find

$$y_t^N = \int_0^t \left(\frac{\lambda(x_s^N)}{\lambda(x_s^N) + \mu r} - c(\Phi_s(y^N)) \right) ds - \frac{rM_t}{N} + \frac{rS_t}{N}. \tag{2.23}$$

6. Here we introduce a semimartingale

$$Y_t^N = \sqrt{N}\, y_t^N, \tag{2.24}$$

which plays an important role in proving diffusion approximation because

$$\begin{aligned} \Phi_t(Y^N) &= \Phi_t(\sqrt{N}\, y^N) \\ &= \sqrt{N}\,\Phi_t(y^N) \\ &= \sqrt{N}\, x_t^N \\ &= X_t^N. \end{aligned} \tag{2.25}$$

We derive from (2.23) that

$$Y_t^N = \sqrt{N}\int_0^t \left(\frac{\lambda(\Phi_s(y^N))}{\lambda(\Phi_s(y^N)) + \mu r} - c(\Phi_s(y^N)) \right) ds - \frac{rM_t}{\sqrt{N}} + \frac{rS_t}{\sqrt{N}}. \tag{2.26}$$

Show that the predictable quadratic variation of the square integrable martingale $rM_t/\sqrt{N}$ is given by the formula

$$\begin{aligned} \left\langle \frac{rM}{\sqrt{N}} \right\rangle_t &= \int_0^t \frac{2\mu r^3 \lambda(x_s^N)}{[\lambda(x_s^N) + \mu r]^3}\, ds \\ &\quad + \frac{r}{N}\int_0^t \frac{\lambda(x_s^N) + 3\mu r}{[\lambda(x_s^N) + \mu r]^2}\, dS_s - \frac{r}{N}\int_0^t \frac{\lambda(x_s^N) + 3\mu r}{[\lambda(x_s^N) + \mu r]^2}\, dM_s. \end{aligned} \tag{2.27}$$

From the definition of $rM_t/\sqrt{N}$, it follows $\langle rM/\sqrt{N}\rangle_t = \int_0^t \gamma^2(x_s^N)\, d\langle m\rangle_s$. On the other hand, (2.7) and (2.8) imply

$$\begin{aligned} \langle m\rangle_t &= N\int_0^t [(1 - \xi_s)\lambda(x_s^N) - \xi_s \mu r]\, ds + 2N\mu r\int_0^t \xi_s\, ds \\ &= (\xi_t - \xi_0) - m_t + 2N\mu r\int_0^t \frac{\lambda(x_s^N)}{\lambda(x_s^N) + \mu r}\, ds - 2\mu r M_t + 2\mu r S_t. \end{aligned}$$

Thus, (2.27) is implied by (2.21).

3. Asymptotic relations for y_t^N, $S_t/\sqrt{N}$ and $rM_t/\sqrt{N}$. In this section, asymptotic relations, that play an essential role in proving the main result, especially for heavy and moderate traffics, are gethered and formulated as lemmas. Throughout this section all assumptions are presupposed.

LEMMA 3.1. *Let S_t and M_t be defined in* (2.21). *Then for any $T > 0$,*

$$\mathbf{P} - \lim_N \begin{cases} \dfrac{1}{\sqrt{N}} \sup_{t \le T} |S_t| = 0, \\ \dfrac{1}{N} \sup_{t \le T} |M_t| = 0. \end{cases}$$

Defined in (2.17), the function $\Phi_t(V)$ is Lipschitz continuous (see Section 2). Therefore, a past-dependent differential equation

$$\frac{dy_t}{dt} = \frac{r\lambda(\Phi_t(y))}{\lambda(\Phi_t(y)) + \mu r} - c(\Phi_t(y)), \tag{3.1}$$

subject to $y_0 = 0$, has the unique solution:

$$y_t = \begin{cases} \text{solution of } (1.8), & \rho > 1, \\ c(0)[\rho - 1]t, & \rho \le 1. \end{cases} \tag{3.2}$$

Then

$$\tilde{x}_t \equiv \Phi_t(y) = \begin{cases} \text{solution of } (1.8), & \rho > 1, \\ 0, & \rho \le 1. \end{cases} \tag{3.3}$$

LEMMA 3.2. *For any $T > 0$,*

$$\mathbf{P}\text{-}\lim_N |y_t^N - y_t| = 0,$$

where y_t^N is given by (2.18).

COROLLARY 3.1.

$$\mathbf{P}\text{-}\lim_N \sup_{t \le T} |x_t^N - \tilde{x}_t| = 0 \qquad \forall T > 0,$$

where x_t^N is defined by (1.3).

Let $\langle rM/\sqrt{N}\rangle_t$ be the predictable quadratic variation [see (2.27)] of the square integrable martingale $rM_t/\sqrt{N}$.

LEMMA 3.3. *For any $t > 0$,*

$$\mathbf{P}\text{-}\lim_N \left\langle \frac{rM}{\sqrt{N}} \right\rangle_t = \int_0^t \frac{2\mu r^3 \lambda(\tilde{x}_s)}{[\lambda(\tilde{x}_s) + \mu r]^3}\, ds,$$

where $\tilde{x}_t$ is defined by (3.3).

COROLLARY 3.2. *For $\rho = 1$,*

$$\mathbf{P}\text{-}\lim_N \left\langle \frac{rM}{\sqrt{N}} \right\rangle_t = t\frac{2\mu r^3 \lambda(0)}{[\lambda(0) + \mu r]^3} \qquad \left[= \frac{2t\mu c^3(0)}{\lambda^2(0)} \right].$$

COROLLARY 3.3. $rM_t/\sqrt{N}$ *converges weakly (in the Skorokhod–Lindvall topology) to a Gaussian process defined by Itô's integral w.r.t. a Wiener process W_t:*

$$\int_0^t \sqrt{\frac{2\mu r^3 \lambda(\tilde{x}_s)}{[\lambda(\tilde{x}_s) + \mu r]^3}}\, dW_s.$$

In proving these lemmas, we always use the following fact ([9], Problem 1.9.2): If L_t^N is a sequence of square integrable martingales with the predictable quadratic variation $\langle L^N \rangle_t$ such that for any $T > 0$, $\langle L^N \rangle_T$ converges to zero in probability as $N \to \infty$, then the same convergence holds for $\sup_{t \le T} |L_t^N|$.

PROOF OF LEMMA 3.1. To check the validity of the first statement, we apply Itô's formula to $\gamma(x_t^N)\xi_t$. Taking into account that the function $\gamma(z)$ is continuously differentiable, with bounded derivative $\gamma'(z)$, we find

$$\gamma(x_t^N)\xi_t = \gamma(0)\xi_0 + S_t + \int_0^t \gamma'(x_s^N)\frac{dx_s^N}{ds}\xi_s\,ds. \tag{3.4}$$

By virtue of (1.3) one can conclude that dx_s^N/ds is bounded. Then the result follows from (3.4).

To check the second statement in (3.1), it has to be shown that for any $T > 0$, $\langle M/N \rangle_T$ converges to zero in probability. This follows from

$$\left\langle \frac{M}{N} \right\rangle_T = \frac{1}{N}\int_0^T \left[(1-\xi_s)\lambda(x_s^N)\xi_s\mu r\right] ds \le \text{const.}\frac{T}{N}. \qquad \square$$

PROOF OF LEMMA 3.2. Put $\Delta_t = \sup_{s \le t}|y_s^N - y_s|$. By virtue of (3.1) and (2.23), we find

$$\Delta_t \le \int_0^t r\left|\frac{\lambda(\Phi_s(y^N))}{\lambda(\Phi_s(y^N)) + \mu r} - \frac{\lambda(\Phi_s(y))}{\lambda(\Phi_s(y)) + \mu r}\right| ds$$
$$+ \int_0^t \left|c(\Phi_s(y^N)) - c(\Phi_s(y))\right| ds + \left|\frac{S_t}{N}\right| + \left|\frac{M_t}{N}\right|.$$

Making assumptions and under the Lipschitz property of $\Phi_t(V)$, there exists a positive constant, say l, such that for any $t \le T$,

$$\Delta_t \le l\int_0^t \Delta_s\,ds + \frac{1}{N}\sup_{s \le T}|S_s| + \frac{1}{N}\sup_{s \le T}|M_s|$$

and so, by the Gronwall–Bellman inequality,

$$\Delta_T \le \exp\{lT\}\left[\frac{1}{N}\sup_{s \le T}|S_s| + \frac{1}{N}\sup_{s \le T}|M_s|\right].$$

Thus, the result is implied by Lemma 3.1. □

PROOF OF LEMMA 3.3. By Corollary 3.1, the first term in the right-hand side of (2.27) converges in probability to relevant limit, whereas two other terms converge to zero in probability. Proofs for these convergences are similar to those in Lemma 3.1 because, due to (2.21) and (2.20),

$$\int_0^t \frac{2\mu r\lambda(x_s^N)}{\left[\lambda(x_s^N)+\mu r\right]^2}\,dS_s = \int_0^t \frac{2\mu r\lambda(x_s^N)}{\left[\lambda(x_s^N)+\mu r\right]^3}\,d\xi_s,$$

$$\int_0^t \frac{2\mu r\lambda(x_s^N)}{\left[\lambda(x_s^N)+\mu r\right]^2}\,dM_s = \int_0^t \frac{2\mu r\lambda(x_s^N)}{\left[\lambda(x_s^N)+\mu r\right]^3}\,dm_s$$

and the function $\tilde{\gamma}(z) = (2\mu r\lambda(z))/[\lambda(z)+\mu r]^3$ is continuously differentiable, having bounded derivative. □

Proof of Corollary 3.2. It follows from $\tilde{x}_t \equiv 0$ and $(2\mu r^3\lambda(0))/[\lambda(0) + \mu r]^3 = 2\rho^3\mu c^3(0)/\lambda(0)^2$. □

Proof of Corollary 3.3. The square integrable martingales $rM_t/\sqrt{N}$, $N \ge 1$, have jumps bounded by const.$/\sqrt{N}$. Then the result is implied by [9], Theorem 7.1.4. □

4. Collapse and diffusion approximation for Y_t^N. In this section, only moderate and normal traffic are studied. As in Section 3, all results are formulated as lemmas, provided that all relevant assumptions are presupposed. Let Y_t^N be defined by (2.24).

Lemma 4.1 (Collapse). *If $\rho < 1$, then for any $t > 0$,*

$$\mathbf{P}\text{-}\lim_N Y_t^N = -\infty.$$

Corollary 4.1. *Let X_t^N be defined by* (1.6). *Then, due to* (2.25), *for any* $T > 0$,

$$\mathbf{P}\text{-}\lim_N \sup_{t\le T} X_t^N = 0.$$

By the Lipschitz property for $\Phi_t(V)$, the past-dependent Itô equation (w.r.t. a Wiener process W_t)

$$Y_t = \int_0^t \left(\lambda'(0)\mu[c(0)/\lambda(0)]^2 - c'(0)\right)\Phi_s(Y)\,ds + \int_0^t \sqrt{2\mu c^3(0)/\lambda^2(0)}\,dW_s \tag{4.1}$$

has the unique strong solution.

Lemma 4.2 (Diffusion approximation). *If $\rho = 1$, then Y_t^N converges weakly (in the Skorokhod–Lindvall topology) to the process Y_t defined by* (4.1).

Proof of Lemma 4.1. Note that Y_t^N obeys decomposition (2.26). By Lemma 3.1, $\sup_{t\le T}|rS_t/\sqrt{N}|$, $T > 0$, converges to zero in probability. By Corollary 3.2, $rM_t/\sqrt{N}$ converges weakly to a continuous Gaussian process. Thereby, it has to be shown that the drift in (2.26) converges to $-\infty$. This drift is absolutely continuous w.r.t. Lebesgue measure with density $\sqrt{N}H(\Phi_s(y^H)) = \sqrt{N}H(x_s^N)$, where

$$H(z) = \frac{r\lambda(z)}{\lambda(z) + \mu r} - c(z). \tag{4.2}$$

So, the result holds, if for each $s > 0$, $\mathbf{P}\text{-}\lim_N \sup_{s \le t} H(x_s^N) < 0$. For $\rho < 1$, by Corollary 3.1, we get $\mathbf{P}\text{-}\lim_N \sup_{s \le t} x_s^N = 0$, $s > 0$. On the other hand, due to the definition of ρ [see (1.5)], we have

$$H(0) = \frac{r\lambda(0)}{\lambda(0) + \mu r} - c(0) = c(0)[\rho - 1] < 0$$

and the result is done. □

PROOF OF LEMMA 4.2. Put $\overline{Y}_t^N = Y_t^N - rS_t/\sqrt{N}$. Because by Lemma 3.1, for any $T > 0$, $\sup_{t \le T} |rS_t/\sqrt{N}|$ converges to zero in probability by [9], Problem 6.2.2, both processes Y_t^N and $\overline{Y}_t^N$ have the same weak limit. So only the weak convergence of $\overline{Y}_t^N$ for the limit Y_t has to be checked.

At first, explain why Y_t could be considered as a weak limit for $\overline{Y}_t^N$. Following (2.26) and (4.2),

$$\overline{Y}_t^N = \sqrt{N}\int_0^t H(\Phi_s(y^N))\,ds - \frac{rM_t}{\sqrt{N}}. \tag{4.3}$$

Due to Corollary 3.3 and $\rho = 1$, $rM_t/\sqrt{N}$ converges weakly for limit $\int_0^t \sqrt{2\mu c^3(0)/\lambda^2(0)}\,dW_s$ and we obtain the "diffusion part" of Y_t. By virtue of $\rho = 1$ we have $H(0) = 0$ and, what is more, $H(z)$ is a continuously differentiable function and its derivative $H'(z)$ is bounded and Lipschitz continuous. Then we get

$$\begin{aligned}\sqrt{N}H(\Phi_s(y^N)) &\approx H'(0)\sqrt{N}\Phi_s(y^N)\\ &= H'(0)\Phi_s(\sqrt{N}y^N)\\ &= H'(0)\Phi_s(Y^N)\\ &\approx H'(0)\Phi_s(\overline{Y})\end{aligned}$$

and so the drift of $\overline{Y}_t^N$ can be "approximated" by $\int_0^t H'(0)\Phi_s(\overline{Y}^N)\,ds$, where

$$H'(0) = \lambda'(0)\mu[c(0)/\lambda(0)]^2 - c'(0). \tag{4.4}$$

For exact proof of weak convergence, apply [9], Theorem 8.3.1(c) and Problem 8.3.3. Because the process Y_t^N is continuous, jumps of $\overline{Y}_t^N$ coincide with jumps of $rM_t/\sqrt{N}$ that are no more than const.$/\sqrt{N}$.

Therefore, only conditions

$$\mathbf{P}\text{-}\lim_N \sup_{t \le T}\left|\int_0^t \left[\sqrt{N}H(\Phi_s(y^N)) - H'(0)\Phi_s(\overline{Y}^N)\right] ds\right| = 0 \qquad \forall\, T > 0, \tag{4.5}$$

have to be checked.

To this end, taking into account the Lipschitz property of $\Phi_s(V)$ and Lemma 3.1, one can conclude that (4.5) is equivalent ot the same relation by replacing $\overline{Y}^N$ on Y^N. According to the Lipschitz property of $H'(z)$, say, with Lipschitz constant l, we find ($0 \le \theta \le 1$)

$$\begin{aligned}
&\left|\sqrt{N} H(\Phi_s(y^N)) - H'(0)\Phi_s(Y^N)\right| \\
&\quad = \left|\sqrt{N} H'(\theta\Phi_s(y^N))\Phi_s(y^N) - H'(0)\Phi_s(Y^N)\right| \\
&\quad = \sqrt{N}\left|H'(\theta\Phi_s(y^N)) - H'(0)\right|\Phi_s(y^N) \\
&\quad \le l\sqrt{N}\Phi_s^2(y^N) \\
&\quad = l x_s^N \Phi_s(Y^N).
\end{aligned}$$

Consequently, (4.5) holds if

$$\mathbf{P}\text{-}\lim_N \int_0^T x_s^N \Phi_s(Y^N)\, ds = 0. \tag{4.6}$$

Due to $\rho = 1$, we have by Lemma 3.2 that $\sup_{t \le T} x_t^N$ converges to zero in probability and so (4.6) holds on any of the sets $\{\sup_{s \le T} \Phi_s(Y^N) \le K\}$, $K > 0$. Thus, (4.6) holds if a family $\{\sup_{s \le T} \Phi_s(Y^N) \le K\}$, $K > 0$, is tight, that is,

$$\lim_{K\to\infty} \limsup_N \mathbf{P}\Big(\sup_{s\le T} \Phi_s(Y^N) \ge K\Big) = 0.$$

Because

$$\sup_{s\le T} \Phi_s(Y^N) = \sup_{s\le T}\Big[Y_s^N - \inf_{u\le s} Y_u^N\Big] \le 2\sup_{s\le T}|Y_s^N|$$

it has to be shown that

$$\lim_{K\to\infty} \limsup_N \mathbf{P}\Big(\sup_{s\le T}|Y_s^N| \ge K\Big) = 0. \tag{4.7}$$

According to (2.26) and (4.2) and taking into account $H(0) = 0$ and boundness of $H'(z)$, say $|H'(z)| \le l$, for $t \le T$, we get

$$\begin{aligned}
\sup_{s\le t}|Y_s^N| &\le \int_0^t \sqrt{N}\left|H(\Phi_s(y^N))\right| ds + \sup_{s\le T}\left|\frac{rM_s}{\sqrt{N}}\right| + \sup_{s\le T}\left|\frac{rS_s}{\sqrt{N}}\right| \\
&\le l\int_0^t (\Phi_s(Y^N))\, ds + \sup_{s\le T}\left|\frac{rM_s}{\sqrt{N}}\right| + \sup_{s\le T}\left|\frac{rS_s}{\sqrt{N}}\right| \\
&\le 2l\int_0^t \sup_{v\le s}|Y_v^N|\, ds + \sup_{s\le T}\left|\frac{rM_s}{\sqrt{N}}\right| + \sup_{s\le T}\left|\frac{rS_s}{\sqrt{N}}\right|
\end{aligned}$$

and so, by the Gronwall–Bellman inequality, we obtain

$$\sup_{s\le T}|Y_s^N| \le \sup_{s\le T}\left[\left|\frac{rM_s}{\sqrt{N}}\right| + \sup_{s\le T}\left|\frac{rS_s}{\sqrt{N}}\right|\right]\exp(2lT).$$

This estimate implies (4.7) because by Lemma 3.1, $\sup_{s\le T}|rS_s/\sqrt{N}|$ converges to zero in probability and by Corollary 3.3, $rM_s/\sqrt{N}$ converges weakly in the Skorokhod–Lindvall topology, and so $\sup_{s\le T}|rM_s/\sqrt{N}|$ is tight in the sense of (4.7). □

5. Asymptotic relations for normal traffic. The following two lemmas are stated provided that all assumptions are presupposed.

LEMMA 5.1. *For any* $t > 0$,

$$\mathbf{P}\text{-}\lim_N \int_0^t I(q_s > 0)\,ds = \rho t.$$

LEMMA 5.2. *For any* $t > 0$,

$$\mathbf{P}\text{-}\lim_N \int_0^t q_s\,ds = 0.$$

PROOF OF LEMMA 5.1. For brevity, $\to$ is used to designate convergence in probability. Because $\rho < 1$, by Corollary 3.1, we get $\sup_{s \le t} x_s^N \to 0$. Then

$$\int_0^t \frac{\lambda(x_s^N)}{\lambda(x_s^N) + \mu r}\,ds \to \frac{\lambda(0)}{\lambda(0) + \mu r}t.$$

Due to Lemma 3.1, $M_t/N \to 0$ and $S_t/N \to 0$. Hence, from (2.22) and formula (1.5) for ρ, it follows that

$$\int_0^t r\xi_s\,ds \to c(0)\rho t. \tag{5.1}$$

By Corollary 3.1 we get $\int_0^t [c(x_s^N) - c(0)]I(q_s > 0)\,ds \to 0$. Then (1.3) and Corollary 3.1 imply

$$c(0)\int_0^t I(q_s > 0)\,ds - \int_0^t r\xi_s\,ds \to 0$$

and the desired result follows from (5.1) and $c(0) > 0$. □

PROOF OF LEMMA 5.2. First, we show that

$$\int_0^t \xi_s I(q_s > 0)\,ds \to \rho t. \tag{5.2}$$

To this end, we use decomposition (2.15) for $G(z)$, which is chosen to satisfy the recursion

$$[G(j+1) - G(j)]\lambda(0) - [G(j-1) - G(j)]\mu r = 0, \qquad j = 0, 1, \tag{5.3}$$

provided that $G(1) \neq G(0)$. For such $G(z)$, we have

$$\begin{aligned} x_t^N G(\xi_t) = &\int_0^t q_s[G(\xi_s + 1) - G(\xi_s)]\big[\lambda(x_s^N) - \lambda(0)\big]\,ds \\ &+ \int_0^t G(\xi_s)\big[r\xi_s - c(x_s^N)I(q_s > 0)\big]\,ds \\ &+ \int_0^t x_s^N\,dP_s(\xi). \end{aligned} \tag{5.4}$$

By Corollary 3.1, $x_t^N G(\xi_t) \to 0$. The first term on the right-hand side of (5.4) is evaluated by const. $\sup_{x\geq 0} |\lambda'(z)| \sup_{s\leq t}(q_s x_s^N)$ and so because $\sup_{s\leq t}(q_s x_s^N) = \sup_{s\leq t}(X_s^N)^2$, it converges, due to Corollary 4.1, to zero in probability. The same arguments imply convergence to zero in probability for Itô's integral on the right-hand side of (5.4) because due to (2.16), $\langle \int_0^{\cdot} x_s^N \, dP_s(\xi)\rangle_t \leq$ const. $\sup_{s\leq t}(X_s^N)^2$. Also by Corollary 3.1, we have $\int_0^t G(\xi_s)[c(x_s^N) - c(0)]I(q_s > 0)\, ds \to 0$. Providing all these asymptotics, we find that

$$\int_0^t G(\xi_s)[r\xi_s - c(0)I(q_s > 0)]\, ds \to 0. \tag{5.5}$$

The integrand in (5.5) can be transformed as

$$\begin{aligned} G(\xi_s)&[r\xi_s - c(0)I(q_s > 0)] \\ &= G(1)\xi_s[r - c(0)I(q_s > 0)] - G(0)(1-\xi_s)c(0)I(q_s > 0) \\ &= G(1)r\xi_s - G(0)c(0)I(q_s > 0) + c(0)[G(0) - G(1)]\xi_s I(q_s > 0). \end{aligned}$$

Then

$$\begin{aligned} \int_0^t & G(\xi_s)[r\xi_s - c(0)I(q_s > 0)]\, ds \\ &= c(0)[G(0) - G(1)]\int_0^t \xi_s I(q_s > 0)\, ds \\ &\quad + c(0)G(0)\int_0^t I(q_s > 0)\, ds - G(1)\int_0^t r\xi_s\, ds \end{aligned}$$

and, as a consequence of (5.1) and Lemma 5.1, we obtain

$$c(0)[G(0) - G(1)]\int_0^t \xi_s I(q_s > 0)\, ds \to c(0)\rho t[G(0) - G(1)],$$

that is, (5.2) holds.

The second step consists of showing the vector convergence

$$\begin{pmatrix} \lambda(0) & -[\lambda(0) + \mu r] \\ -c(0) & r \end{pmatrix} \begin{pmatrix} \int_0^t q_s\, ds \\ \int_0^t \xi_s q_s\, ds \end{pmatrix} \to \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5.6}$$

To this end, decomposition (2.13) is used. Due to Corollary 3.1, the left-hand side of (2.13) converges to zero in probability. Formula (2.14) and Corollary 4.1 imply $\langle \int_0^{\cdot} x_s^N \, dm_s\rangle_t \to 0$ and so the Itô integral on the right-hand side of (2.13) converges to zero in probability. Also by Corollary 3.1, $\int_0^t \xi_s[c(x_s^N) - c(0)]I(q_s > 0)\, ds \to 0$ and by (5.1) and Lemma 5.1, $\int_0^t \xi_s[r\xi_s - c(0)I(q_s >)0]\, ds \to 0$. These relations imply that the second integral on the right-hand side of (2.13) converges to zero in probability. So the first integral in the right-hand side of (2.13) converges to zero in probability too. Using the estimate

$$\int_0^t q_s |\lambda(x_s^N) - \lambda(0)|\, ds \le t \sup_{x\ge 0} |\lambda'(z)| \sup_{s\le t} q_s x_s^N$$
$$= t \sup_{z\ge 0} \lambda'(z) |\lambda'(z)| \sup_{s\le t} (X_s^N)^2,$$

Corollary 4.1 and taking into account all previous asymptotic relations for the terms from decomposition (2.13), we arrive at

$$\int_0^t q_s[(1-\xi_s)\lambda(0) - \xi_s \mu r]\, ds \to 0. \tag{5.7}$$

Analogously, due to (2.2) and Corollary 4.1, we find

$$\int_0^t q_s[r\xi_s - c(0)]\, ds \to 0. \tag{5.8}$$

Evidently, (5.7) and (5.8) are equivalent to (5.6).

For normal traffic, the matrix in (5.6) is nondegenerate and so the desired statement follows from Lemma 5.2. □

6. Proof of Theorem 1.1.

Normal traffic. Due to homogeneity of x_t^N, ξ_t, it is enough to consider only the case $t_1 = 0$, $t_2 = t > 0$. Then from Lemma 5.1, it follows that

$$\frac{1}{t}\int_0^t I(q_s = 0)\, ds = 1 - \frac{1}{t}\int_0^t I(q_s < 0)\, ds \to 1 - \rho.$$

Thus, we get

$$\lim_N \mathbf{E}\left(\frac{1}{t}\int_0^t I(q_s = 0)\, ds\right) = \lim_N \frac{1}{t}\int_0^t \mathbf{P}(q_s = 0)\, ds = 1 - \rho.$$

To check the second statement, note that

$$\frac{1}{t}\int_0^t I(0 < q_s \le \varepsilon)\, ds = \frac{1}{t}\int_0^t I(q_s > 0)\, ds - \frac{1}{t}\int_0^t I(q_s > \varepsilon).$$

The second integral on the right-hand side of this equality goes to zero in probability because by Chebyshev's inequality

$$\int_0^t I(q_s > \varepsilon)\, ds \le \frac{1}{\varepsilon}\int_0^t q_s\, ds$$

and by Lemma 5.2, $\int_0^t q_s\, ds \to 0$. Therefore, by Lemma 5.1,

$$\frac{1}{t}\int_0^t I(0 < q_s \le \varepsilon)\, ds \to \rho.$$

Thus, we get

$$\lim_N \mathbf{E}\left(\frac{1}{t}\int_0^t I(0 < q_s \le \varepsilon)\, ds\right) = \lim_N \frac{1}{t}\int_0^t \mathbf{P}(0 < q_s \le \varepsilon)\, ds = \rho.$$

Moderate traffic. Due to (2.25) and Lemma 4.2, X_t^N converges weakly (in the Skorokhod–Lindvall topology) to $X_t = \Phi_t(Y)$, where Y_t is defined by the past-dependent Itô equation (4.1). Because X_t^N and X_t are continuous processes, weak convergence in the local uniform topology holds too.

So it remains to show that the same distribution is shared by random processes $\Phi_t(Y)$ and $|Z_t|$, where Z_t is Gaussian process defined by Itô's equation (1.7). For brevity, let

$$b = \lambda'(0)\mu[c(0)/\lambda(0)]^2 - c'(0),\qquad \sigma = \sqrt{2\mu c^3(0)/\lambda^2(0)}. \tag{6.1}$$

Then, (1.7) and (4.1) can be rewritten as

$$dZ_t = bZ_t\,dt + \sigma\,dW_t,\qquad dY_t = b\Phi_t(Y)\,dt + \sigma\,dW_t. \tag{6.2}$$

According to the definition of $\Phi_t(V)$ [see (2.17)], we find from (6.2) that

$$X_t = \int_0^t bX_s\,ds + \sigma W_t + \Big[-\inf_{s\le t} Y_s\Big], \tag{6.3}$$

where $[-\inf_{s\le t} Y_s]$ is the functional of the normal reflection at zero. The process X_t is normally reflected at zero process Y_t. Consider this process on the stochastic basis with the filtration generated by X_t. Then by [9], Theorem 10.2.1, there exists on this basis a Wiener process $\overline{W}_t$ and a functional Ψ_t of the normal reflection at zero such that

$$X_t = \int_0^t bX_s\,ds + \sigma\overline{W}_t + \Psi_t, \tag{6.4}$$

and, what is more, (6.4) has the unique strong solution (see, e.g., [9], Theorem 10.2.2). In the Appendix we show that the process $|Z_t|$ is defined by the same differential equation and so the result is done.

Heavy traffic. The result follows from Corollary 3.1.

APPENDIX

Let Z_t be Gaussian diffusion process defined by Itô's equation (w.r.t. a Wiener process W_t)

$$Z_t = b\int_0^t Z_s\,ds + \sigma W_t, \tag{A.1}$$

where $\sigma \neq 0$.

If $b = 0$ and $\sigma = 1$, then Z_t is the Wiener process and it is known that the process $|Z_t|$ has the same distribution as the normal reflected at zero Wiener process. We show that this result remains true for the general situation ($\sigma > 0$, $b \neq 0$), that is, distributions of $|Z_t|$ and normal reflected at zero of Z_t

coincide. To this end, we show that there exists a new Wiener process $\overline{W}_t$ such that

$$|Z_t| = b\int_0^t |Z_s|\, ds + \sigma \overline{W}_t + \Psi_t, \tag{A.2}$$

where Ψ_t is a functional of the normal reflection at zero.

Due to (A.1), $Z_t = e^{bt}U_t$, where

$$U_t = \int_0^t e^{-bs}\sigma\, dW_s. \tag{A.3}$$

The process U_t is a square integrable martingale with the predictable quadratic variation

$$\langle U \rangle_t = \int_0^t e^{-2bs}\sigma^2\, ds. \tag{A.4}$$

Due to the Jensen inequality, process $|U_t|$ is a submartingale. Then by the Doob–Meyer decomposition for submartingales (see, e.g., [9], Theorem 1.6.5), we have

$$|U_t| = A_t + M_t, \tag{A.5}$$

where A_t is an increasing predictable process and M_t is a local martingale. Because $|U_t|$ is a continuous process by [9], Theorem 2.1.2, both processes A_t and M_t are continuous and so M_t is a square integrable martingale (whose predictable quadratic variation is denoted by $\langle M \rangle_t$). Applying Itô's formula to U_t^2, we derive from (A.3) and (A.5),

$$U_t^2 = 2\int_0^t U_s\, dU_s + \langle U \rangle_t,$$

$$U_t^2 = 2\int_0^t |U_s|\, dA_s + \langle M \rangle_t + 2\int_0^t |U_s|\, dM_s.$$

The process U_t^2 is a special semimartingale. So, both its decompositions coincide ([9], Theorem 2.1.1(b)), that is,

$$\begin{aligned} 2\int_0^t U_s\, dU_s &= 2\int_0^t |U_s|\, dM_s, \\ 2\int_0^t |U_s|\, dA_s + \langle M \rangle_t &= \langle U \rangle_t. \end{aligned} \tag{A.6}$$

The first equality in (A.6) implies $\int_0^t U_s^2 d\langle U \rangle_s = \int_0^t U_s^2 d\langle M \rangle_s$ and so

$$\int_0^t I(|U_s| > 0) d\langle M \rangle_s = \int_0^t I(|U_s| > 0) d\langle U \rangle_s. \tag{A.7}$$

Because U_s is a Gaussian random variable with positive covariance, we have $\mathbf{P}(|U_s| = 0) = 0$, $s > 0$, which implies $\int_0^t I(|U_s| = 0)\, ds = 0$. By virtue of (A.4), $\langle U \rangle_t$ is an absolutely continuous function and so $\int_0^t I(|U_s| = 0) d\langle U \rangle_s = 0$.

From the second equality in (A.6), it follows that

$$\int_0^t I(|U_s| = 0) d\langle M\rangle_s = \int_0^t I(|U_s| = 0) d\langle U\rangle_s - 2\int_0^t |U_s| I(|U_s| = 0)\, dA_s = 0.$$

Thus

$$\langle M\rangle_t = \langle U\rangle_t, \qquad \mathbf{P}\text{-a.s., } t \geq 0. \tag{A.8}$$

This and the second equality in (A.6) imply

$$A_t = \int_0^t I(|U_s| = 0)\, dA_s, \qquad \mathbf{P}\text{-a.s., } t \geq 0. \tag{A.9}$$

Applying Itô's formula to $e^{bt}(A_t + M_t)$ $(= |Z_t|)$, we obtain

$$d|Z_t| = b|Z_t|\, dt + e^{bt}\, dM_t + e^{bt} I(|U_t| = 0)\, dA_t.$$

The process $\int_0^t e^{bs}\, dM_s$ is a square integrable martingale with the predictable quadratic variation

$$\begin{aligned} \left\langle \int_0^\cdot e^{bs}\, dM_s \right\rangle_t &= \int_0^t e^{2bs} d\langle M\rangle_s \\ &= \int_0^t e^{2bs} d\langle U\rangle_s \\ &= \sigma^2 t. \end{aligned} \tag{A.10}$$

Consequently, a Wiener process $\overline{W}_t$ can be chosen such that $\int_0^t e^{bs}\, dM_s = \sigma \overline{W}_t$.
Taking into account (A.9) and $I(|Z_t| = 0) = I|U_t| = 0)$, put

$$\Psi_t = \int_0^t e^{bs} I(|Z_s| = 0)\, dA_s. \tag{A.11}$$

Thus, $|Z_t|$ satisfies (A.2) if Ψ_t is the functional of the normal reflection at zero. Following [10] (see also [9], Chapter 1, Section 1]), Ψ_t has to satisfy the following two conditions:

1. For any continuous bounded function $f(z)$, $z \geq 0$, such that $f(0) = 0$,

$$\int_0^t f(|Z_s|)\, d\Psi_s = 0, \qquad t \geq 0.$$

2. For any nonnegative continuous process $\tilde{Z}_t$,

$$\int_0^t \left(\tilde{Z}_s - |Z_s|\right) d\Psi_s$$

is a nondecreasing process.

Evidently, both conditions are fulfilled and the result is complete.

REFERENCES

[1] Aven, O. I., Coffman, Jr., E. G. and Kogan, Y. A. (1987). *Stochastic Analysis of Computer Storage*. Reidel, Dordrecht.

[2] Beneš, V. E. (1988). Storage problems involving finite capacity, loading protocols and exact content. AT & T Bell Laboratories, Murray Hill, NJ.

[3] BIRMAN, A. and KOGAN, Y. (1991). A deterministic approximation for state-dependent buffer model. Private communication.

[4] FREIDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.

[5] KABANOV, YU. M., LIPTSER, R. SH. and SHIRYAYEV, A. N. (1980). Absolute continuity and singularity of locally absolutely continuous probability distributions. II. *Math USSR Sbornik* **35** 31–58.

[6] KOGAN, Y. and LIPTSER, R. (1993). Limit non-stationary behavior of large closed queueing networks with bottlenecks. *Queueing Syst.* **14** 33–55.

[7] LIPTSER, R. (1992). The Bogolubov averaging principle for semimartingales. *Proc. Steklov Math. Inst.* **202**.

[8] LIPTSER, R. S. and SHIRYAYEV, A. N. (1977, 1978). *Statistics of Random Processes, Applications* **1**, **2**. Springer, New York.

[9] LIPTSER, R. S. and SHIRYAYEV, A. N. (1987). *Theory of Martingales*. Kluwer, Dordrecht.

[10] TANAKA, H. (1979). Stochastic differential equations with reflection boundary condition in convex regions. *Hiroshima Math. J.* **9** 163–177.

Y. KOGAN
M. SHENFILD
TECHNION
ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA 32000
ISRAEL

R. LIPTSER
DEPARTMENT OF ENGINEERING
TEL AVIV UNIVERSITY
TEL AVIV 69978
ISRAEL