DOI: 10.1214/ECP.v5-1014

Elect. Comm. in Probab. **4** (1999) 13–27

ELECTRONIC COMMUNICATIONS in PROBABILITY

ON THE BEHAVIOR OF LIFO PREEMPTIVE RESUME QUEUES IN HEAVY TRAFFIC

VLADA LIMIC*
Department of Mathematics, Cornell University
Ithaca, NY 14850-4201.
email: limic@math.cornell.edu

submitted June 26, 1999; *accepted* November 26, 1999

AMS 1991 Subject classification: 60K25 (60J80)
LIFO queue, heavy traffic, measure-valued process, branching, feedback, renewal, Wiener-Hopf factorization.

Abstract

This paper studies heavy traffic behavior of a G/G/1 last-in-first-out (LIFO) preemptive resume *queue, by extending the techniques developed in Limic (1999). The queue length process exhibits a perhaps unexpected heavy traffic behavior. The diffusion limit depends on the type of arrivals (and services) in a fairly intricate way, related to the Wiener-Hopf factorization for random walks.*

1 Introduction

Customers arrive to a single-server queue according to a renewal process with inter-arrival time distribution G, each customer requests service time with distribution function F, independently of other customers. Distributions F and G are concentrated on $(0,\infty)$, and we assume they have finite means m and $1/\lambda$, respectively. The server devotes all of its service potential to the last customer to have arrived. Moreover, at the moment of each new arrival the server switches instantaneously from serving the current customer c (if any) to the newest customer $\bar{c}$. Customer c stays waiting in queue and only after $\bar{c}$ is served completely and exits the queue does the service of c resume. The server is busy whenever the queue is non-empty, which is usually referred to as a *non-idling* or *work conserving* property. The queueing process generated by this mechanism, is a *single-class G/G/1 last-in-first-out preemptive resume queue.* We prefer to shorten the name to *G/G/1 LIFO queue.* Special cases are *M/G/1 LIFO* queues, where the inter-arrival distribution G is exponential (rate λ).

Suppose a customer arrives to the queue at time t and requests an amount v of service time. If we let $u(s)$, $s \geq t$ be its total amount of time in service by time s, the *residual service time* of this customer at time s is $v - u(s) \geq 0$. Denote by $(A(t), t \geq 0)$ the renewal process of arrivals, by $Z(t)$ the *queue length* at time t, i.e., the number of individuals in queue at time t, and by $W(t)$ the *(immediate) workload* of the queue at time t, i.e., the total amount of

*Research supported by NSF Mathematical Sciences Postdoctoral Research Fellowship

work still required by customers present in the system at time t (measured in units of server time). Hence, the workload equals the total sum of all the residual service times of customers in queue. The parameter $\rho = m\lambda$, called the *traffic intensity* of the queue, is the average amount of work arriving per unit time. It is a well-known (and easy, cf. section 2.1) fact that the workload process does not vary over work conserving service disciplines. In particular, the workload process $(W(t), t \geq 0)$ is the same for the *first-in-first-out* (FIFO) queue, where the customers are served in the order of their arrival. For a M/G/1 LIFO queue, the workload $(W(t), t \geq 0)$ is a Markov process and it is positive recurrent, null-recurrent, and transient whenever $\rho < 1$, $\rho = 1$ and $\rho > 1$ respectively. For a G/G/1 LIFO queue, the workload is not Markov anymore, but using a random walk comparison, it is easy to see that W returns to 0 infinitely often iff $\rho \leq 1$, and the expected time until return is finite iff $\rho < 1$. From a practical point of view it is desirable to "keep the server busy" most of the time without getting it overwhelmed with work. This corresponds to the situation $\rho = 1 - \varepsilon$ for some small $\varepsilon > 0$, and as $\varepsilon \searrow 0$ the queue approaches *heavy traffic*.

A recent work (Limic [12]) describes the heavy traffic behavior for the M/G/1 LIFO queue under the usual second moment assumptions on the service distribution F. The analysis in [12] is based on the following observation. The state of a M/G/1 LIFO queue at any time t (i.e., the list of residual service times) is encoded via a finite-measure-valued Markov process q_t, called the the *RES-measure* process. It is defined in terms of the queue length and the "future minimum" of the load (cf. section 2.4). An analogue of q_t is the *exploration process* introduced in Le Gall and Le Jan [8, 7]. The above encoding carries over to the present setting.

The goal of this paper is to extend the techniques of [12] in order to study heavy traffic behavior of G/G/1 LIFO queue under the usual heavy traffic assumptions. The LIFO preemptive resume service discipline induces an essentially different heavy traffic (diffusion scale) behavior from those induced by FIFO service discipline in that the limit (or limit points) depends on the type of arrivals (and services) in a more complicated way than via asymptotic behavior of the first two moments (cf. section 4). However, the "state-space-collapse" property (first discovered by Reiman [14] and common in FIFO-type setting, cf. Bramson [3] and Williams [16]), still holds in a weaker form: under more stringent assumptions, the queue length becomes a multiple of the workload in heavy traffic. Amber Puha and Ruth Williams (personal communication) study the fluid (law of large numbers) scale behavior of the processor sharing queue, where the server simultaneously serves all the customers in queue. The state-space-collapse suggested in their work is analogous to the one for LIFO queues, in that the fluid limit depends on the finer properties of the inter-arrival and service time distributions.

Heavy traffic analysis of general multiclass G/G/1 LIFO preemptive resume queues with Markovian feedback, where the (global) arrival process is not a renewal process (see Reiman [15] and Dai and Kurtz [4] for the FIFO analogues) seems to be beyond the scope of techniques presented here.

The paper is organized as follows. Section 2 describes the processes of interest, and some of their properties. Section 3 is the heavy traffic analysis for a sequence of identically distributed critical ($\rho = 1$) G/G/1 LIFO queues. Section 4 gives an example, and discusses related complexity issues in analyzing non-identical near critical G/G/1 LIFO queues approaching heavy traffic.

For any two numbers x, y, let x^+, $x \wedge y$, and $x \vee y$ denote the positive part of x, the minimum, and the maximum of x and y, respectively. We identify $H(t)$ with H_t whenever H is a stochastic process.

2 LIFO queue and related processes

In this section we introduce several processes related to LIFO queues, and mention some important relations. Consider a G/G/1 LIFO (preemptive-resume) queue as in section 1. Assume that the queue is empty at time $t = 0$.

2.1 The load and the workload

Let $(v_i : i \geq 1)$ be the service times (i.i.d. random variables with distribution F) requested by the customers in the order of their arrival. Let $(u_i : i \geq 1)$ be the i.i.d. inter-arrival times of customers, where $u_1 \overset{d}{=} G$. The *load* X_t and the workload W_t of the queue at time t are given by

$$X_t = \sum_{i=1}^{A(t)} v_i - t \ \ , \ W_t = \sum_{i=1}^{A(t)} v_i - t + (-I_t) \, , \tag{1}$$

where $A(t) = \sup\{j : \sum_{i=1}^{j} u_j \leq t\}$ equals the number of customers that arrived to the queue in the time interval $[0, t]$, and $I_t = \inf_{s \leq t} X_s$. Note that the process $-I_t = -\inf_{s \leq t} X_s$ is the *cumulative idletime* of the server by time t, that is, the total time $|\{s \leq t : W_s = 0\}|$ with no customer in queue. The workload process W is the load process X, reflected above its past infimum. It is easy to see that (1) agrees with the notion of workload in section 1. The excursions of X above the past infimum, or equivalently, the excursions of W above 0 correspond to the busy cycles of the queue.

2.2 The queue length and time-reversal

Figure 1 shows a possible path of X over a finite time interval. Suppose X had a jump at some (random) time s and write $X_{s-} = \lim_{u \uparrow\uparrow s} X_u$. Let $\gamma_s = \inf\{u \geq s : X_u \leq X_{s-}\}$. At time γ_s the customer that arrived at time s exits the queue, in the meantime its service might be interrupted several times due to jumps of X, that is, arrivals of new customers. We identify the actual set of times when this customer is in service with the set $\mathcal{A}_s = \{u \in [s, \gamma_s] : \inf_{t \in [s,u]} X_t \geq X_u\}$, indicated in bold on the time axis in the figure. The "gaps" in $\mathcal{A}_s$ correspond to services of the "intermittent" customers. The customer who arrived (jumped) at time s will still be in queue at time $t > s$ if and only if $\gamma_s > t$, that is,

$$X_{s-} < \inf_{u \in [s,t]} X_u \, , \tag{2}$$

(as it happens for s and t in the figure).

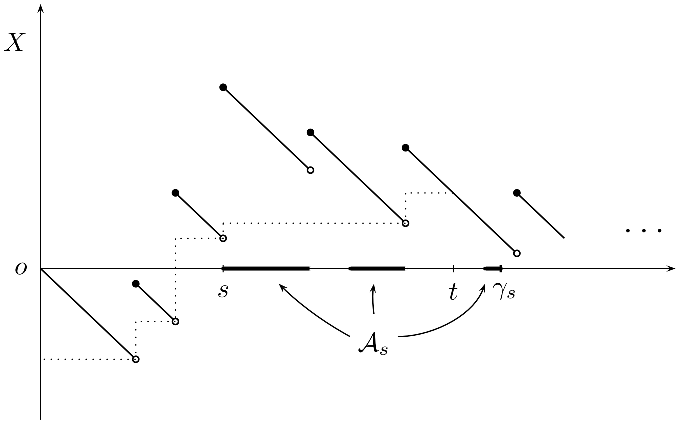

Figure 1

The difference $(\inf_{u\in[s,t]} X_u - X_{s-})^+$ is its residual service time at t. Therefore, the queue length process $Z_t = Z(t)$ satisfies

$$Z_t = \#\{s \le t : X_{s-} < \inf_{s\le u\le t} X_u\}\,. \tag{3}$$

Let $I_s^t = \inf_{u\in[s,t]} X_u$ be the *future infimum* process (dotted line in figure 1) of X up to time t. The jumps of $I_\cdot^t$ may occur only at the times $s < t$ at which customers arrive, and the jump sizes $(\inf_{u\in[s,t]} X_u - X_{s-})^+$ are the residual service times at time t of the corresponding customers. The following observation will be important for deriving the queue length heavy traffic approximation. If we fix any time t and time-reverse the load X from t back to 0 (or equivalently, rotate the figure 1 about the origin by 180 degrees), the future infimum $I_\cdot^t$ "gets mapped" onto the (past) supremum process of the *time-reversed* load process. In particular, the queue length Z_t which equals the number of jumps of the future infimum by (3), also equals the number of jumps of the time-reversed supremum process occurring in $[0,t]$. In symbols, let $\widetilde{X}_s^t = X_t - X_{(t-s)-}$, $\widetilde{X}_t^t = X_t$ be the time-reversed load, and let $\widetilde{S}_s^t = \sup_{u\in[0,s]} \widetilde{X}_u^t$. Then (3) states

$$Z_t = \#\{z : z \in [0,t],\, \widetilde{S}_z^t > \widetilde{S}_{z-}^t\}\,. \tag{4}$$

Note that the time-reversed load process $(\widetilde{X}_s^t, s \in [0,t])$ does not have the same law as $(X_s, s \in [0,t])$, unless the arrival process is Poisson.

2.3 Intrinsic branching

Suppose we call a customer who arrives at time t a descendant of a customer that arrived at time s if the latter is still in queue at time t, that is, if (2) holds. This procedure determines a one-to-one correspondence between the busy cycles of the queue and a sequence of independent identically distributed random trees. Any customer either finds the queue empty upon arrival, in which case it becomes a *progenitor* (or *root*), or finds the queue non-empty, in which case it becomes a *child* of the customer being served immediately prior to its arrival. Figure 2 shows a part of the tree induced by the first busy cycle from figure 1.

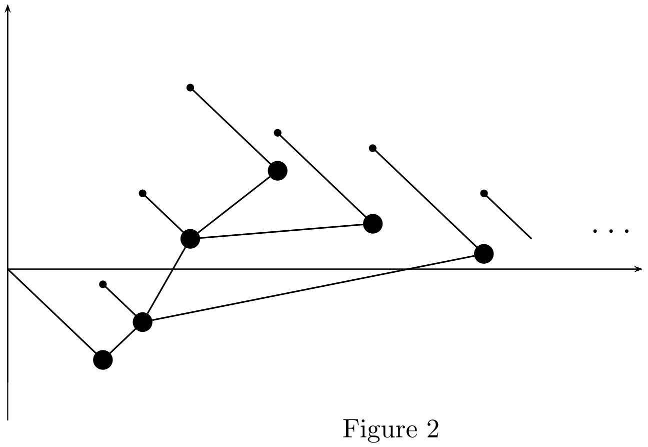

Figure 2

In the special case of Poisson arrivals (M/G/1), the above trees are clearly Galton-Watson and their offspring distribution can be easily expressed in terms of service distribution F. In the case of renewal arrivals, the Galton-Watson property is preserved. Perhaps the best way to verify this is by considering the children of a particular vertex, denote them by $c_{s_1}, c_{s_2}, \ldots, c_{s_K}$, where $s_1 < s_2 \ldots < s_K$ are the arrival times of the corresponding customers to the queue. Consider the subtrees spanned by c_{s_i} and its descendents, for $i = 1, \ldots, K$. It suffices to show that, conditionally on K, the K subtrees above are mutually independent, and have the law of the whole tree (corresponding to a complete busy cycle). This is all easily verified from the fact that the ith subtree is determined by the "excursion" $(X_{s_i+u} - X_{s_i-},\, 0 \le u \le \gamma_{s_i} - s_i)$. The offspring distribution now depends on the inter-arrival and service distributions of the queue, though finding an explicit formula seems to be difficult.

We return to this very useful branching characterization in the heavy traffic analysis, section 3.3. The relation between queueing and branching goes back to Kendall [10], and the relation between excursions of random walks and branching goes back to Harris [9].

2.4 The RES-measure

One can think of a LIFO queue as a continuous-time process with values in the state space of finite lists of arbitrary length. At each time t, the state of the queue is the list of residual service times for all queued customers ordered by their arrival times. It is convenient to encode the above list via the *RES-measure* process $(q_t, t \ge 0)$ that takes values in the space $M_f(R_+)$ of finite measures (with finite support) on $[0,\infty)$. The queue length can be recovered from q via

$$Z_t = \sup(\mathrm{Supp}(q_t))\,, \tag{5}$$

where $\mathrm{Supp}(\mu)$ denotes the closed support of μ. Moreover, the list of residual service times at any time t equals $(q_t(1),\, q_t(2), \ldots, q_t(Z_t))$, the list of masses of atoms of q_t. The workload is then given by $W_t = \sum_{i=1}^{Z_t} q_t(i) = \int_{[0,t]} dI_s^t = \langle q_t, 1\rangle$. Finally, the process q_t is defined by

$$\langle q_t, \varphi\rangle := \int_{[0,t]} \varphi(Z_s)\, dI_s^t = \int_{[0,t]} \varphi(Z_s^t)\, dI_s^t\,. \tag{6}$$

Here $\langle \mu, \varphi \rangle$ stands for $\int_{[0,\infty)} \varphi \, d\mu$, where $\varphi : R^+ \to R$ is a continuous function with bounded support, and

$$Z_s^t = \#\{u \le s : X_{u-} < \inf_{u \le z \le t} X_z\} \tag{7}$$

is the number of individuals in queue at time s that will still be in queue at time t. As in (4), we can express Z_s^t via time-reversal as

$$Z_s^t = \#\{z : z \in [t-s, t] : \widetilde{S}_z^t > \widetilde{S}_{z-}^t\}. \tag{8}$$

Note that the integrals in (6) are in fact finite sums, and the second equality is due to a simple fact $Z_s \equiv Z_s^t$, dI_s^t - a.e., $s \in [0,t]$. Process Z_s^t is clearly non-decreasing in s for each fixed t. We prefer integrals to sums in (6), since the heavy traffic limit Theorem 3 involves the convergence of rescaled q's to a limit of the same form.

3 Heavy traffic

Consider a family of G/G/1 queues, indexed by r, with inter-arrival time distribution function G^r, and service time distribution function F^r. Let F^r have finite mean m^r, and let G^r have finite mean $1/\lambda^r$. Denote by $A^r(\cdot)$, $W^r(\cdot), q^r(\cdot)$ and $Z^r(\cdot)$ the corresponding arrival, workload, RES-mea–sure, and queue length processes, respectively. We assume that for each r, the queue is empty at time 0 ($W^r(0) = 0$), so the notation of previous sections applies. In particular, rewrite equation (1) as

$$X_t^r = \sum_{i=1}^{A^r(t)} v_i^r - t \ , \ W_t^r = \sum_{i=1}^{A^r(t)} v_i^r - t + (-I_t^r) \, , \tag{9}$$

where $-I^r(t) = -\inf_{s \le t} X^r(s)$ is the idle time.

3.1 Asymptotics for the load and the workload

The first heavy traffic assumptions are

$$m^r \to m \in (0,\infty) \text{ and } \lambda^r \to \lambda \in (0,\infty) \text{ as } r \to \infty \, , \tag{10}$$

$$\sqrt{r}(1-\rho^r) = \sqrt{r}(1 - m^r \lambda^r) \to c \text{ as } r \to \infty \, . \tag{11}$$

Suppose for a moment that the arrival processes are Poisson (M/G/1 setting) with rate λ^r. Assume moreover that for each r the service times have finite second moment β^r, and

$$\beta^r \to \beta < \infty \text{ as } r \to \infty \, , \tag{12}$$

$$\sup_r E[(v_1^r)^2 1_{\{v_1^r \ge K\}}] \to 0 \ \text{ as } K \to 0 \, . \tag{13}$$

Let $\hat{X}^r(t) = r^{-1/2} X^r(rt)$ and $\hat{W}^r(t) = r^{-1/2} W^r(rt)$. Then an easy application (cf. [12]) of the functional CLT and the continuity mapping theorem shows the convergence of $\hat{X}^r$ and $\hat{W}^r$ in distribution to Brownian motion X, and reflected Brownian motion $W = X - I$, respectively, where X has drift $-c$ and variance $\sigma^2 = \beta/m$, and $I_t = \inf_{s \le t} X_s$. For the case of renewal arrivals, assume in addition the second moments η^r of the inter-arrival times are finite, and moreover

$$\eta^r \to \eta < \infty \text{ as } r \to \infty\,, \tag{14}$$

$$\sup_r E[(u_1^r)^2 1_{\{u_1^r \geq K\}}] \to 0 \quad \text{as } K \to 0\,. \tag{15}$$

Then the above statement (and the proof) of convergence for the load and the workload processes continues to hold, the only difference being that the limiting variance σ^2 is given by $\beta/m - 2m + \eta/m$.

For a sequence of M/G/1 LIFO queues under assumptions (10)-(13), it was shown in [12], that the load, the workload, the RES-measure and the queue length processes, simultaneously satisfy

$$(\hat{X}^r, \hat{W}^r, \hat{q}^r, \hat{Z}^r) \Rightarrow (X, W, q, Z)\,. \tag{16}$$

Here "$\hat{}$" indicates appropriate rescaling, and $\Rightarrow$ denotes convergence in distribution with respect to the topology on the corresponding Skorokhod space. The processes X and W are those from the previous paragraph, and the processes q and Z are defined by $Z = \frac{2}{\sigma^2}W = \frac{2m}{\beta}W$, and

$$\langle q_t, \varphi \rangle = \int_{[0,t]} \varphi(Z_s)\, dI_s^t, \tag{17}$$

where $I_s^t = \inf_{u \in [s,t]} X_u$.

In the next section we show that, in the case of renewal arrivals, the queue length $\hat{Z}^r$ converges (under much more stringent assumptions) to a limit of the form $Z = \alpha W$, where the scaling constant α depends on the inter-arrival and service distributions in a fairly intricate way, related to the Wiener-Hopf factorization for random walks (cf. Feller [6]).

3.2 Asymptotics for RES-measure and queue length

Consider a sequence of identical (in distribution) G/G/1 LIFO queues with corresponding service and arrival distributions F and G, having finite means m and $1/\lambda$, and finite second moments β and η, respectively. The assumption (11) translates to

$$m\lambda = 1 \text{ and } c = 0\,,$$

and the moment assumptions (10,12,13,14,15) are automatically satisfied.

Let $\hat{X}^r(t) = r^{-1/2}X^r(rt)$ and $\hat{W}^r(t) = r^{-1/2}W^r(rt)$ be as before. Since $(\hat{X}^r, \hat{W}^r) \Rightarrow (X, W)$, by the Skorokhod representation theorem we may assume that

$$(\hat{X}^r, \hat{W}^r) \to (X, W) \tag{18}$$

almost surely in the Skorokhod space $D_{R^2}[0,\infty)$. Rescale the queue length and infimum processes accordingly by

$$\hat{Z}_t^r = r^{-1/2}Z^r(rt)\ ,\ \hat{Z}^{t,r}(s) = r^{-1/2}Z_{rs}^{rt,r}\,,$$

and

$$\hat{I}_t^r = r^{-1/2}I^r(rt)\ ,\ \hat{I}^{t,r}(s) = r^{-1/2}I_{rs}^{rt,r}\,.$$

Convergence in (18) implies that for t fixed

$$-\hat{I}^{t,r}(\cdot) \to -I^t(\cdot) \quad \text{a.s. in } D_{R_+}[0,t]\ , \text{as } r \to \infty\,, \tag{19}$$

where $I^t(s) = I^t_s = \inf_{u\in[s,t]} X_u$. Let $\hat{q}^r$ be a measure-valued process defined in analogy to (6) via

$$\langle \hat{q}^r_t, \varphi\rangle := \int_{[0,t]} \varphi(\hat{Z}^r_s)\, d\hat{I}^{t,r}_s = \int_{[0,t]} \varphi(\hat{Z}^{t,r}_s)\, d\hat{I}^{t,r}_s \,. \tag{20}$$

Let $\widetilde{X}^t$ denote the time-reversed Brownian motion X, so $\widetilde{X}^t_s = X_t - X_{t-s}$, and let $\widetilde{S}^t_z = \sup_{s\in[0,z]} \widetilde{X}^t_s = X_t - I^t_{t-z}$.

Lemma 1 *There exists a constant $\alpha = \alpha(F,G) \in (0,\infty)$ such that, for each fixed $t \geq 0$*

$$P(\sup_{s\in[0,t]} |\hat{Z}^{t,r}(s) - \alpha(I^t_s - I_t)\,| > \varepsilon) \to 0 \quad \textit{as } r\to\infty\,.$$

Proof. A suitable modification of the proof of [12], Lemma 3.2.2 yields the result. Fix $t>0$ and a finite subdivision $0 \leq s_1 < s_2 < \ldots < s_k = t$ on $[0,t]$. Identity (8) implies

$$\hat{Z}^{t,r}(s) = \#\{u : t-s \leq u \leq t,\, \widehat{\widetilde{S}}^{t,r}(u) > \widehat{\widetilde{S}}^{t,r}(u-)\}\cdot r^{-1/2},$$

where $\widehat{\widetilde{S}}^{t,r}(u) = \sup_{x\in[0,u]} \widehat{\widetilde{X}}^{rt,r}(x)$ is the supremum process of $\widehat{\widetilde{X}}^{rt,r} = r^{-1/2}(X^r(rt) - X^r((rt-rs)-))$, the rescaled and time-reversed X^r.
Consider the time-reversed process $\widetilde{X}^{rt,r}(s) = X^r(rt) - X^r((rt-s)-)$, $0\leq s < rt$. Denote by $M^r_t(z)$ the number of jumps of $\widetilde{X}^{rt,r}$ above its past maximum in the interval $[0, rz]$. Note that $\hat{Z}^{t,r}(s) = r^{-1/2}(M^r_t(t) - M^r_t((t-s)-)) = r^{-1/2}(M^r_t(t) - M^r_t(t-s)) + O(r^{-1/2})$. We show there exists $\alpha\in(0,\infty)$ such that, for each fixed $z\in[0,t]$,

$$r^{-1/2} M^r_t(z) \;\overset{p}{\to}\; \alpha\, \widetilde{S}^t_z \;,\; r\to\infty\,, \tag{21}$$

where $\overset{p}{\to}$ denotes convergence in probability. Define

$$Z^t_s \equiv Z^t(s) := \alpha(\widetilde{S}^t_t - \widetilde{S}^t_{t-s}) = \alpha(I^t_s - I_t)\,. \tag{22}$$

Then by (21) we get

$$(\hat{Z}^{t,r}(s_1), \hat{Z}^{t,r}(s_2), \ldots, \hat{Z}^{t,r}(s_k)) \;\overset{p}{\to}\; (Z^t(s_1), Z^t(s_2), \ldots, Z^t(s_k)),\; r\to\infty\,. \tag{23}$$

The lemma follows from (23), since $\hat{Z}^{t,r}(\cdot)$ is non-decreasing for each r and t, and $Z^t(\cdot)$ is continuous and non-decreasing for each t.
In order to show (21), it will be convenient to consider an "extension" process $(\widetilde{X}^{rt,t}(s), s\geq 0)$ of $(\widetilde{X}^{rt,t}(s), 0\leq s\leq rt)$, defined in the following way. Independently of the filtration generated by X^r, take a sequence $\{u_{-i}, i\geq 1\}$ of i.i.d. random variables with distribution G, and a sequence $\{v_{-i}, i\geq 0\}$ of i.i.d. random variables with distribution F. Define $\widetilde{X}^{rt,r}(rt) = \widetilde{X}^{rt,r}(rt-) + v_0$ and moreover, $\widetilde{X}^{rt,r}(rt+z) = \widetilde{X}^{rt,r}(rt) + \sum_{i=1}^{A_-(z)} v_{-i} - z$, $z\geq 0$, where $A_-(z) = \sup\{j : \sum_{i=1}^{j} u_{-j} \leq z\}$. Then the extended process $\widetilde{X}^{rt,r}$ decreases deterministically at rate 1 in between successive jumps. The time of the first jump has distribution G_1 (typically $\neq G$), and all other inter-jump times are i.i.d. random variables with distribution G. The sizes of all jumps are i.i.d. random variables with distribution F.
Let $\widetilde{S}^{rt,r}$ be the supremum process of the extended $\widetilde{X}^{rt,t}$. Denote by $T^{r,t}_1 < T^{r,t}_2 < \ldots$ the successive increase (jump) times of $\widetilde{S}^{rt,r}$. Denote the corresponding sizes of the *overshoots* by $J^{r,t}_1, J^{r,t}_2, \ldots$, so that $J^{r,t}_i = \widetilde{X}^{rt,r}(T^{r,t}_i) - \widetilde{S}^{rt,r}(T^{r,t}_i-)$. If the arrivals are not Poisson,

the time-reversal influences the distributions of $T_1^{r,t}$ and $J_1^{r,t}$ in some fairly complicated way. However, it is easily seen that the renewal property of arrivals implies that the subsequent overshoots $J_i^{r,t}, i \geq 2$ are independent and identically distributed random variables. Here it is convenient to have $\widetilde{X}^{rt,r}$ defined for all $s \geq 0$. In fact, we only need to consider overshoots that happened by (reversed) time $rz \leq rt$. Note that

$$\widehat{\widetilde{S}}^{t,r}(z) = \sum_{i=1}^{M_t^r(z)} \frac{1}{r^{1/2}} J_i^r \, , \; 0 \leq z < t \, , \tag{24}$$

and the convergence in (18) implies $\widehat{\widetilde{S}}^{t,r}(z) \to \widetilde{S}^t(z)$ a.s., as $r \to \infty$. The number of overshoots $M_t^r(z)$ was defined only for $z < t$, but if we let $M_t^r(t) = M_t^r(t-)$, equation (24) will also hold for $z = t$. Since for each r, we have $F^r = F$ and $G^r = G$, the distribution of $J_2^{r,t}$ does not depend on r. Similarly, the distribution of $J_2^{r,t}$ does not depend on t either. Let

$$\alpha := 1/E(J_2^{r,t}) \, .$$

The remark after Corollary 2 gives an expression for α, in particular, $\alpha \in (0, \infty)$. It is easy to see that $M_t^r(z) \to \infty$ a.s. as $r \to \infty$, therefore (24) and a law of large numbers yield

$$\frac{\widehat{\widetilde{S}}^{t,r}(z)}{r^{-1/2} M_t^r(z)} \xrightarrow{p} \alpha^{-1} \, , \; r \to \infty \, ,$$

implying (21), hence (23). □
Define $Z_t := Z_t^t = \alpha(X_t - I_t)$. Since $\hat{Z}_t^r = \hat{Z}_t^{t,r}$ for all t and r, we have

Corollary 2 *For any fixed t and $0 \leq t_1 < t_2 \ldots < t_k \leq t$*

$$(\hat{Z}_{t_1}^r, \hat{Z}_{t_2}^r, \ldots, \hat{Z}_{t_k}^r) \xrightarrow{p} (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}) \, .$$

Remark. One can express the constant $\alpha(F,G)$ in Lemma 1 via the Wiener-Hopf factorization for random walks in Feller [6], Chapters XII and XVIII. As always, let $(v_i, i \geq 1)$ be i.i.d. random variables with distribution F, and let $(u_i, i \geq 1)$ be i.i.d. random variables with distribution G, independent of v's. Finally, set $S_n = \sum_{i=1}^n (v_i - u_i)$. Then it is easy to see that the overshoots (starting from the second one) in the proof of Lemma 1 are the *ascending ladder heights* of $(S_n, n \geq 1)$. By [6], Theorem XVIII.5.1

$$\alpha(F,G)^{-1} = \frac{\sqrt{\text{var } (u_1) + \text{var } (v_1)}}{\sqrt{2}} \exp\left\{ - \sum_{i=1}^{\infty} \frac{1}{i} (P(S_i > 0) - \frac{1}{2}) \right\} . \tag{25}$$

It is hard to determine the exact value of $\alpha(F,G)$ in practice, except when F or G are exponential (or related) distributions (cf. Prabhu [13], and section 4).
Let q be as in (17) where Z_t is as in Corollary 2, and $I_s^t = \inf_{u \in [s,t]} X_u$.

Theorem 3 *$\hat{q}^r \Rightarrow q$, as $r \to \infty$.*

The proof is the same as that for the corresponding [12], Theorem 3.2.1, using Lemma 1, and (18-20). Also, Theorem 3 is a consequence of (19) and the following

Theorem 4 *$\hat{Z}^r \Rightarrow Z$ as $r \to \infty$.*

Again, we extend the proof of the corresponding result in [12] to the present setting. The renewal arrivals require extra care, however, since $F^r = F$ and $G^r = G$ for all r, the tree estimates in the proof of Proposition 5 below simplify a great deal (compare to [12] section 3.3). We omit some details.

Proof. Fix some $\varepsilon > 0$ and $\eta > 0$. Fix time $T > 0$, and let $t_i = i(T/n)$, $0 \le i \le n$ be the subdivision of $[0,T]$ with mesh size T/n. For n large enough we have

$$P(\sup_{1\le i\le n} \sup_{u\in[t_{i-1},t_i]} |Z_{t_i} - Z_u| > \varepsilon) \le \eta\,, \tag{26}$$

by continuity of Z (cf. [5] p.122). Recall that for each t the process $(Z^t(s),\, 0 \le s \le t)$ given by (22) is continuous, moreover the processes $(Z^{t_i}(t_i) - Z^{t_i}(t_i - \theta),\, \theta \in [0, T/n])$, $1 \le i \le n$ are all (independent and) identically distributed. It is then easily seen (and shown in [12], Lemma 3.2.5) that for all large n

$$P(\sup_{1\le i\le n} \sup_{\theta\in[0,T/n]} |Z^{t_i}(t_i) - Z^{t_i}(t_i - \theta)| = \sup_{1\le i\le n} |Z^{t_i}(t_i) - Z^{t_i}(t_{i-1})| > \varepsilon) \le \eta\,. \tag{27}$$

The finite dimensional distributions of $\hat{Z}^r$ are converging to those of Z due to Corollary 2. So it suffices to show the tightness of $\hat{Z}^r$, $r \ge 1$ with respect to the Skorokhod topology on $D_R[0,\infty)$.

The idea is to use $\hat{Z}^r_{t_i-\theta} \approx \hat{Z}^{t_i,r}(t_i - \theta) \approx \hat{Z}^{t_i,r}(t_i) = \hat{Z}^r_{t_i} \approx Z_{t_i}$ for small θ, and exploit the monotonicity of $\hat{Z}^{t,r}_s$ and Z^t_s in s. Let $\mathcal{F}^r_t$ be the filtration generated by $\hat{X}^r$. Observe that, for each r,

$$\hat{Z}^{t_i,r}_{t_{i-1}} \le \hat{Z}^{t_i,r}_t \le \hat{Z}^r_t \le \hat{Z}^r_{t_{i-1}} + \sup_{u\in[0,T/n]} \hat{Z}^{r,i}_u + 1/\sqrt{r}\,,\ t \in [t_{i-1}, t_i]\,, \tag{28}$$

where $(\hat{Z}^{r,i}_u,\, u \in [0,T/n])$ has the same law as $(\hat{Z}^r_u,\, u \in [0,T/n])$, and is independent of $\mathcal{F}^r_{t_{i-1}}$. The first inequality in (28) is the monotonicity of $Z^{t,r}_s$ in s, the second inequality trivially follows from the interpretation of $\hat{Z}^{t_i,r}_t$ as the (rescaled) number of individuals in queue at time rt whose service will not have been completed by time rt_i. For the last inequality in (28) note that the number of customers that arrive to the queue in the time interval $[rt_{i-1}, rt]$, and do not exit by time rt, can be bounded from above by $1 + Z^*_{(rt-\tau)^+}$ where $Z^* \overset{d}{=} Z$, and $\tau \ge rt_{i-1}$ is the first renewal (arrival) time after rt_{i-1}. Assume the following

Proposition 5 *For any fixed $\varepsilon, \eta > 0$ and any integer n_1, there exist $n \ge n_1$ and $r_1 \ge 1$ such that*

$$\sup_{r\ge r_1} P(\sup_{1\le i\le n} \sup_{u\in[0,T/n]} \hat{Z}^{r,i}_u > \varepsilon) \le \eta\,, \tag{29}$$

where $\hat{Z}^{r,i}_u$ are defined in (28).

The rest of the proof is the same as in [12]: for $\varepsilon, \eta > 0$ and T fixed as above, find n_1 large enough so that (26, 27) are satisfied for all $n \ge n_1$. Then find $n \ge n_1$ and r_1 so that (29) holds. By Corollary 2, Lemma 1 and (26,27) we can find $r_2 \ge r_1$ large enough so that both

$$\sup_{r\ge r_2} P(\sup_{1\le i\le n} |\hat{Z}^r_{t_i} - \hat{Z}^r_{t_{i-1}}| > 2\varepsilon) \le 2\eta \quad \text{and}$$

$$\sup_{r\ge r_2} P(\sup_{1\le i\le n} \sup_{s\in[t_{i-1},t_i]} |\hat{Z}^{t_i,r}_s - \hat{Z}^{t_i,r}_{t_i}| > 2\varepsilon) \le 2\eta \tag{30}$$

hold. Combined with (29) this implies that for any $0 < h < T/n$ we have

$$\sup_{r\ge r_2} P(\sup_{|s-t|<h} |\hat{Z}^r_s - \hat{Z}^r_t| > 10\varepsilon) \le 9\eta\,.$$

3.3 Proof of Proposition 5

Recall the branching interpretation for the queue length from section 2.3. Each busy cycle of the queue corresponds to an excursion of the load (workload) process, and yields a Galton-Watson tree $\mathcal{T}$ of customers who entered (and exited) the queue during this busy cycle. The *generation* of a vertex in $\mathcal{T}$ is its distance from the root, so the root belongs to generation 0, its children to generation 1, their children to generation 2, etc. Let $|\mathcal{T}|$ denote the *total size* (number of vertices) of $\mathcal{T}$, and let $\mathrm{ht}(\mathcal{T})$ denote the *height* (the maximal generation) of $\mathcal{T}$, respectively. A customer that arrives at time s, creates a new vertex ς in the corresponding tree. If the queue was empty immediately before the arrival ($Z(s-) = 0$) then ς becomes the root, otherwise ς becomes a child of the customer whose service was interrupted, in both cases the generation of ς in $\mathcal{T}$ equals $Z(s-) = Z(s) - 1$. For the M/G/1 queue, where the customers arrive as Poisson (rate λ) process, it is easy to see that, given v, the offspring distribution for the trees above is Poisson (rate $v\lambda$), that is

$$P(\xi = i) = E\left[\frac{e^{-v\lambda}(v\lambda)^i}{i!}\right] \ , \ i \geq 0 \, . \tag{31}$$

As mentioned earlier, in the general G/G/1 case, the exact offspring distribution Ξ seems difficult to obtain. Since the excursions of the load (workload) have finite length with probability 1, the corresponding trees have finite size (therefore height), and this is equivalent to $E\xi \leq 1$ (e.g. [2]). In fact, it is intuitively obvious that $E\xi = 1$ which again seems to be tricky to verify via a direct calculation. An indirect way is to note that any increase in service time (e.g. scale by $1 + \upsilon$, $\upsilon > 0$) will induce divergence of the load process to $+\infty$, or equivalently, the corresponding trees will have mean offspring $E\xi_\upsilon > 1$. By uniform integrability, one then argues that $E\xi_\upsilon \to E\xi_0$ as $\upsilon \to 0$, but $E\xi_0 = E\xi \leq 1$, so it must be $E\xi = 1$.
The variance σ_ξ^2 of ξ is finite as well. A crude bound on σ_ξ can be obtained in the following way. Recall the set $\mathcal{A}_s$ from section 2 (and Figure 1). If $N(\mathcal{A}_s)$ denotes the number of connected components of $\mathcal{A}_s$, then $N(\mathcal{A}_s) - 1$ is the number of children of the customer who arrived at time s. The first connected component of $\mathcal{A}_s$ has distribution $u \wedge v$, where u and v are independent, and $v \overset{d}{=} F$, $u \overset{d}{=} G$. In particular, $\{N(\mathcal{A}_s) = 1\} = \{u > v\}$, therefore, $P(\xi = 0) = P(u > v)$. Similarly, it is easily verified that,

$$P(\xi \geq k) = P(N(\mathcal{A}_s) > k) = P(u + w_1 + \ldots + w_k \leq v) \ , \tag{32}$$

where $w_1, \ldots, w_k$ are independent and identically distributed random variables. By truncating w's (from above) if necessary, we may assume $0 < E(w_1) < \infty$, which turns (32) into an upper bound for $P(\xi \geq k)$. Therefore,

$$\begin{aligned} \frac{1}{2}(E\xi^2 + E\xi) &= \sum_k kP(\xi \geq k) \leq \sum_k kP(u + w_1 + \ldots + w_k \leq v) \\ &\leq \sum_k kP(w_1 + \ldots + w_k < \frac{k}{2}Ew_1) + \sum_k kP(v \geq \frac{k}{2}Ew_1) \, , \end{aligned}$$

and the first sum is finite due to a large deviation principle, while the second sum is finite due to $Ev^2 < \infty$.
Let $\mathcal{T}_{i,j}^r, 1 \leq j \leq M_i^r$ be the Galton-Watson trees of the busy cycles started after time 0 and completed before time rT/n, and let $\mathcal{T}_i^r$ be the tree of the busy cycle containing the customer present in service at time rT/n. If the queue is empty at time rT/n, set $\mathcal{T}_i^r = \emptyset$ to be the

empty tree with $\mathrm{ht}(\emptyset)=0$. The maximal queue-length $\sup_{u\in[0,T/n]} Z^{r,i}_{ru}$ in the interval $[0,T/n]$ is dominated by $\max_{1\le j\le M^r_i} \mathrm{ht}(\mathcal{T}^r_{i,j}) \vee \mathrm{ht}(\mathcal{T}^r_i)+1$, the maximal height of all trees (of busy cycles) started in $[0,rT/n]$. So the proposition will follow from

$$\sup_{r\ge r_1} P(\max_{1\le i\le n}(\max_{1\le j\le M^r_i} \mathrm{ht}(\mathcal{T}^r_{i,j}))\vee \mathrm{ht}(\mathcal{T}^r_i)\,)>\varepsilon\sqrt{r})\le \eta\,. \tag{33}$$

Note that, since $F^r=F$ and $G^r=G$, all trees $\mathcal{T}^r_{i,j}$, $\mathcal{T}^r_i$, $i,j\ge 1$, have the same *critical* (mean $E\xi=1$) offspring distribution Ξ. Kolchin, [11] Theorem 2.1.2, shows

$$P(\mathrm{ht}(\mathcal{T})>r)\sim \frac{2}{\sigma^2_\xi r} \quad \text{as } r\to\infty\,. \tag{34}$$

Aldous, [1] Proposition 24, gives the following estimate for the joint height and total size distribution of the same tree:

$$r^{1/2}P(\mathrm{ht}(\mathcal{T})>\varepsilon r^{1/2}, |\mathcal{T}|<\delta r)\to \sigma_\xi^{-1}\delta^{-1/2}G(\varepsilon\delta^{-1/2}\sigma_\xi) \quad \text{as } r\to\infty\,, \tag{35}$$

where $G(x)\le \kappa_1\exp(-x/\kappa_2), 0<x<\infty$ for some $0<\kappa_1,\kappa_2<\infty$.
Recall M^r_i is the number of trees corresponding to the completed busy cycles of $Z^{i,r}$, and let $N^r_i=|\mathcal{T}^r_{i,1}|+|\mathcal{T}^r_{i,2}|+\ldots+|\mathcal{T}^r_{i,M^r_i}|$ be the total number of vertices for these trees.

Lemma 6 *For any fixed $n\ge 1$, there exists $r_3\ge 1$ such that*

$$\sup_{r\ge r_3} P(\max_{1\le i\le n} N^r_i>(\lambda+\varepsilon)\frac{T}{n}r)\le\eta\,.$$

Lemma 7 *There is a constant $\kappa<\infty$ such that, for any fixed $n_1\ge 1$ there exist $n\ge n_1$ and $r_3\ge 1$ such that*

$$\sup_{r\ge r_3} P(\max_{1\le i\le n} M^r_i>\sqrt{r}\kappa)\le\eta\,.$$

As in the corresponding proof in [12], use the above lemmas together with (35) to conclude that for any such n and $r\ge r_3$,

$$\begin{aligned}
&P\,(\max_{1\le i\le n}\max_{1\le j\le M^r_i} \mathrm{ht}(\mathcal{T}^r_{i,j})>\varepsilon\sqrt{r})\\
&\le\ 2\eta+n\sqrt{r}\kappa P(\mathrm{ht}(\mathcal{T})>\varepsilon\sqrt{r}, |\mathcal{T}|<(\lambda+\varepsilon)\frac{T}{n}r)\ ,\\
&\le\ 2\eta+\tfrac{\kappa_1\kappa}{\sigma\sqrt{(\lambda+\varepsilon)T}}n^{3/2}\exp(-\frac{\varepsilon\sigma n^{1/2}}{\kappa_2\sqrt{(\lambda+\varepsilon)T}})\,.
\end{aligned}$$

Similarly, by (34),

$$P(\max_{1\le i\le n} \mathrm{ht}(\mathcal{T}^r_i)>\varepsilon\sqrt{r})\le\frac{2n}{\sigma^2_\xi\varepsilon\sqrt{r}}+o_r(1)\,,$$

for any fixed n. The above estimates imply (33).
It remains to prove technical Lemmas 6 and 7. Lemma 6 is easy (by Blackwell's renewal theorem) since N^r_i is the number of renewal arrivals with rate λ in the interval $[0,rT/n]$. For Lemma 7, let $\tau_x=\inf\{s\ge 0: I_s\le -x\}$, where $I_t=\inf_{0\le s\le t}X_s$, and X is the load process of the queue. Let $M_x=\#\{s\in[0,\tau_x]: X_{s-}<X_s \text{ and } W_s=Z_s=0\}$ be the number of customers arriving to empty queue during the interval $[0,\tau_x]$. So M_x is the number of busy

cycles (i.e. the number of trees) started in the interval $[0, \tau_x]$. Now consider the infimum processes $I_t^{r,i} = \inf_{0 \le s \le t} X_s^{r,i}$, and the corresponding $\tau_x^{r,i} := \inf\{s \ge 0 : \hat{I}_s^{r,i} \le -x\}$, with the obvious notation. Since the asymptotic load X is a Brownian motion, it is easy to see that we can find $n \ge n_1$ and r_3 large enough so that

$$\sup_{r \ge r_3} P(\min_{1 \le i \le n} \tau_{\sqrt{r}}^{r,i} < rT/n) = \sup_{r \ge r_3} P(\max_{1 \le i \le n} (-I_{rT/n}^{r,i}) > \sqrt{r}\,) \le \eta/2\,. \tag{36}$$

On the complement of $\{\min_{1 \le i \le n} \tau_{\sqrt{r}}^{r,i} < rT/n\}$ we have $M_i^r \le M_{\sqrt{r}}^{i,r} \overset{\mathrm{d}}{\le} M_{\sqrt{r}} + 1$, where $\overset{\mathrm{d}}{\le}$ stands for "stochastically dominated". We have $M_{\sqrt{r}}^{i,r} \overset{\mathrm{d}}{\le} M_{\sqrt{r}} + 1$, rather than $M_{\sqrt{r}}^{i,r} \overset{d}{=} M_{\sqrt{r}}$, since the queue might be non-empty at time $t_{i-1}r$ (therefore "$\le$"), or the queue might be empty during some time interval $[t_{i-1} - \varepsilon, t_{i-1}]$ which influences the distribution of arrival for the first progenitor (therefore "+1"). Due to (36) and Lemma 8 below, Lemma 7 holds with $\kappa = b + \varepsilon$, for any $\varepsilon > 0$.

Lemma 8 *There exists a constant $b < \infty$ such that*

$$M_x/x \to b\,, \ \text{a.s. as } x \to \infty\,.$$

Proof. Recall the independent and identically distributed random variables $(w_i, i \ge 1)$ from (32). Then, as before, $\{M_x > k\} = \{u + w_1 + w_2 \ldots + w_k < x\}$, so M_x is "almost" a renewal process. Therefore

$$\frac{M_x}{x} \to \frac{1}{Ew_1} \quad \text{a.s. as } x \to \infty\,,$$

and we have $b = 1/Ew_1 < \infty$, since $Ew_1 > 0$.

4 An example and related questions

There exist four probability distribution functions F^1, F^2, G^1, G^2, with the following properties. Their means are identical,

$$m_1 = \int xF^1(dx) = \int xF^2(dx) = m_2 = \frac{1}{\lambda_1} = \int xG^1(dx) = \int xG^2(dx) = \frac{1}{\lambda_2}\,, \tag{37}$$

their second moments are all finite, and moreover

$$\beta_1 = \int x^2F^1(dx) = \int x^2F^2(dx) = \beta_2\ ,\ \eta_1 = \int x^2G^1(dx) = \int xG^2(dx) = \eta_2\,. \tag{38}$$

Finally, (in the notation of Lemma 1)

$$\alpha(F^1, G^1) \ne \alpha(F^2, G^2)\,. \tag{39}$$

An example can be constructed using the analysis of Prabhu, [13]. Let $F^1 = F^2 = \text{Gamma}(2,1)$ distribution, that is, the convolution of two Exponential (rate 1) distributions. Let G^1 be uniform $\mathcal{U}[1,3]$ and G^2 be Gamma(12,6), that is, the convolution of 12 Exponential (rate 6) distributions. It is easy to check that (37),(38) hold with $m_1 = 2$, $\beta_1 = 2$ and $\eta_1 = 1/3$. Recall the representation for $\alpha(F, G)$ from the remark after Corollary 2. Let ψ_j be the Laplace transform of G^j, $j = 1, 2$, so

$$\psi_1(\theta) = \frac{e^{-\theta} - e^{-3\theta}}{2\theta}\ ,\ \psi_2(\theta) = \left(\frac{6}{6+\theta}\right)^{12}\,.$$

By [13], Lemma I.6 (p.42) and Theorem A2, the corresponding characteristic functions of the overshoot (ascending ladder height) is

$$E\exp\{-i\omega S^j\} \equiv \chi_j(\omega) = 1 - \left(1 - \frac{\xi_1^j}{1-i\omega}\right)\left(1 - \frac{\xi_2^j}{1-i\omega}\right), \tag{40}$$

where $\xi_1^j > \xi_2^j$ are the two roots of the equation

$$\xi^2 = \psi_j(1-\xi)\,,$$

such that $|\xi_i^j| \leq 1$, $i,j = 1,2$. Of course, $\xi_1^1 = \xi_1^2 = 1$, and it is easy to check (e.g. through Mathematica) that $\xi_2^1 \approx -0.309208 < \xi_2^2 \approx -0.306625$. From (40) we see that

$$\alpha(F^1,G^1)^{-1} = ES^1 = 1 - \xi_2^1 \neq 1 - \xi_2^1 = \alpha(F^2,G^2)^{-1}\,.$$

This demonstrates that the natural first and second moment assumptions (10)-(15) are not sufficient for the joint convergence (16), when the inter-arrival times have general distribution. Namely, consider a sequence of G/G/1 LIFO queues, indexed by positive-integers, and such that the subsequence indexed by odd (even) integers satisfies the assumptions of section 3.2 with service and inter-arrival distribution functions F^1, G^1 (F^2, G^2). Even though the moment assumptions (10)-(15) are trivially satisfied, Theorems 3 and 4 imply the existence of two different limit points for the queue length and the RES-measure in heavy traffic.

In the special case when $G_1 = G_2$ is exponential (rate λ) distribution the above "dichotomy" is not possible, since it is well known that

$$\alpha(F,G)^{-1} = \lambda\beta/2\,,$$

where β is the second moment of F. This is an important ingredient in establishing heavy traffic behavior (16) for a sequence of M/G/1 LIFO queues, under assumptions (10)-(15).

In the above example, the difference between the two α's was not large. It is plausible that $\alpha(\cdot,\cdot)$ is a continuous function in its variables, under suitable uniform integrability assumptions. Here are some natural related open questions. How large can $|\alpha(F^1,G^1) - \alpha(F^2,G^2)|$ be under assumptions (37,38), when the mean and the variances are bounded by constants? Is it possible to construct an example of F^1, F^2, G^1, G^2, where the first $k > 2$ moments of F^1 and F^2, and of G^1 and G^2 agree, but still relation (39) holds? How large can the difference $|\alpha(F^1,G^1) - \alpha(F^2,G^2)|$ be in this case?

This paper concentrates on the critical case where $F^r = F$ and $G^r = G$, for all r. It is plausible one could use a similar approach to show

Conjecture 9 *For a sequence of G/G/1 LIFO queues satisfying (10)-(15) and the additional assumption*

$$\alpha(F^r,G^r) \to \alpha \in (0,\infty)\,, \text{ as } r \to \infty\,,$$

convergence in (16) holds, where $Z = \alpha^{-1}W$ and q is defined in (17).

A person in need of a heavy traffic limit theorem for such a sequence of near critical G/G/1 queues should understand the asymptotic behavior of $\alpha(F^r,G^r)$'s related to his/her problem. Since little is explicitly known in general about the Wiener-Hopf constants $\alpha(F,G)^{-1}$, (unless F or G are Gamma distributions), the problem of verifying the conjecture, or its statement corresponding to a particular problem, is left to an interested reader.

References

[1] D.J. Aldous. The continuum random tree III. *Ann. Probab.*, 21:248–289, 1993.

[2] K.B. Athreya and P.E. Ney. *Branching Processes.* Springer-Verlag, New York-Heidelberg, 1972.

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

[4] J.G. Dai and T.G. Kurtz. A multiclass station with markovian feedback in heavy traffic. *Mathematics of Operations Research*, 20:721–741, 1995.

[5] S.N. Ethier and T.G. Kurtz. *Markov Processes: Characterization and Convergence.* Wiley, New York, 1986.

[6] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1971.

[7] J.F. Le Gall and Y. Le Jan. Branching processes in Lévy processes: Laplace functionals of snakes and superprocesses. *Ann. Probab.*, 26:1407–1432, 1998.

[8] J.F. Le Gall and Y. Le Jan. Branching processes in Lévy processes: The exploration process. *Ann. Probab.*, 26:213–252, 1998.

[9] T.E. Harris. First passage and recurrence distributions. *Trans. Amer. Math. Soc*, 73:471–486, 1974.

[10] D.G. Kendall. Some problems in the theory of queues. *J. Royal Statist. Soc.*, B 13:151–185, 1951.

[11] V.F. Kolchin. *Random Mappings.* Optimisation Software, New York, 1986. (Transaltion of Russion original).

[12] V. Limic. A LIFO queue in heavy traffic. Technical report, UC San Diego, 1999. Available at `http://math.cornell.edu/~limic/`.

[13] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance Risk, Dams and Data Communication.* Springer, Second edition, 1998.

[14] M.I. Reiman. Some diffusion approximations with state space collapse. In *Leture Notes in Contr. and Inf. Sci.*, volume 60, pages 209–240. Springer, Berlin-New York, 1984.

[15] M.I. Reiman. A multiclass feedback queue in heavy traffic. *Advances in Applied Probability*, 20:179–207, 1988.

[16] R.J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems: Theory and Applications*, 30:27–88, 1998.