The Annals of Applied Probability
2001, Vol. 11, No. 2, 301–331

A LIFO QUEUE IN HEAVY TRAFFIC

BY VLADA LIMIC[1]

Cornell University

This paper describes the heavy-traffic behavior of an M/G/1 *last-in–first-out preemptive resume* queue. An appropriate framework for the analysis is provided by measure-valued processes. In particular, the paper exploits the setting of recent works by Le Gall and Le Jan. Their finite-measure-valued *exploration* process corresponds to our *RES-measure* (residual services measure) process, that captures all the relevant information about the evolution of the queue, while their *height* process corresponds to the *queue-length* process. The heavy-traffic "diffusion" approximations for the RES-measure and the queue-length processes are derived under the usual second moment assumptions on the service distributions. The tightness of queue lengths argument uses estimates for the total size and height of large Galton–Watson trees.

1. Introduction. Imagine customers arriving to a queue according to a Poisson (rate λ) process, each customer requesting an amount of service time with distribution function F, independently of the arrival process and of the service times of other customers. Let F have finite mean m. 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 above rule (or *service discipline*), that applies to serving any customer c, corresponds to a branching structure (cf. Section 1.2). This system is called the M/G/1 *last-in–first-out* (LIFO) *preemptive resume* queue. From now on we will omit the qualifier "preemptive resume" for brevity. Note that the server is busy whenever the queue is nonempty, which is usually referred to as a *nonidling* or *work-conserving* property. We assume the queue is empty at time 0.

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. For any x, $\lfloor x \rfloor$ and $\lceil x \rceil$ denote the largest integer smaller than or equal to x, and the smallest integer larger than x, respectively. Also identify $G(t)$ with G_t whenever G is a stochastic process.

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 $v - u(s) = (v - u(s))^+$. Denote by $(A(t), t \geq 0)$ the Poisson (counting) process of arrivals, by $Z(t)$ the

Received May 1999; revised May 2000.
[1]Research supported by NSF Mathematical Sciences Postdoctoral Research Fellowship.
AMS 2000 *subject classifications*. 60K25, 60J80.
Key words and phrases. LIFO queue, heavy traffic, measure-valued process, branching.

queue length at time t, that is, the number of individuals in queue at time t, and by $W(t)$ the *(immediate) workload* of the queue at time t, that is, the total amount of work still required by customers present in the system at time t (measured in units of server time). So the workload is equal to the total sum of all residual service times. The parameter $\rho = \lambda m$, 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) 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. Therefore, W is a Markov process with respect to the filtration $\mathscr{F}_t$ generated by arrivals and service times up to time t, and it is positive recurrent, null-recurrent and transient whenever $\rho < 1$, $\rho = 1$ and $\rho > 1$, respectively. 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*.

The pioneering works in heavy-traffic approximations to queues (Kingman [27]) and queueing networks (Iglehart and Whitt [21, 22], Harrison [19], Reiman [31] and Whitt [35]) appeared a while ago. A detailed overview of the enormous literature is given in Williams [36]. Recent papers by Bramson [7] and Williams [37] provide powerful tools for the analysis of multiclass queueing networks with feedback in heavy traffic. However, their techniques are developed for *head-of-the-line* (HL) service disciplines. It is intuitively clear what head-of-the-line means (see, e.g., [37] for precise definitions). The FIFO discipline is the simplest HL discipline, while the LIFO discipline is perhaps the simplest non-HL discipline. Recall that our LIFO discipline is preemptive resume, where the server switches to serving the newest customer immediately upon arrival. The queue-length process of the *non-preemptive* LIFO queue, where the server serves each customer completely, and immediately afterwards begins to serve the last customer to have arrived (if any), is equal (in distribution) to that of the FIFO queue, so its heavy-traffic approximation is given in [21, 22]. Although the LIFO discipline might seem "unfair," and therefore less natural than the FIFO discipline, it naturally arises in applications (e.g., LIFO stack in computer science; see also [4, 25]). In fact, here is a natural "optimization" problem. Suppose that, for a queue close to heavy traffic, we have a server with the ability of serving in both FIFO and LIFO orders (equivalently, both non-preemptive and preemptive resume LIFO orders). Due to limited space (say), it is important to minimize the queue length, and the question is: which service discipline to use? Similar questions were considered by Coffman and Mitrani [9]. We discuss the answer in Section 3.4.2, where we see that typically one of the two disciplines is optimal.

Some important aspects of LIFO preemptive resume queues have been investigated previously. Shanthikumar and Sumita [33] study properties of invariant measure in the more general setting of renewal arrivals. Relation to risk processes (Sigman [34]) is another connection to applications. Abate and

Whitt [1] establish heavy-traffic limits for the steady-state waiting time in the M/G/1 LIFO queue.

The goal of this paper is to describe the heavy-traffic behavior of an M/G/1 LIFO queue (Theorems 1 and 5). The corresponding description of its FIFO counterpart is given in [21, 22]. An appropriate framework for the heavy-traffic analysis of the LIFO queue is provided by measure-valued processes. A random process is *measure-valued* if it takes values in a space of measures. Measure-valued processes have been actively studied in the past two decades (see, e.g., Dawson [10], Dynkin [12] for references). In particular, this paper exploits the setting of the recent papers by Le Gall and Le Jan [14, 15], who construct and study the finite-measure-valued *exploration process* (an analogue of which we describe in Section 1.2 and call the *RES-measure* process) as a step in their pathwise construction of superprocesses with general branching mechanism. The *height* process of [14, 15] corresponds to our queue-length process. The theorems in Section 3.2 are stated and proved for an interesting case (from the queueing perspective) where the approximation $X(t)$ to the load process (cf. Section 1.1) is a Brownian motion. By examining the argument, it is easily checked that Theorem 1 continues to hold when X belongs to a more general class of Lévy processes, and where the approximation $Z(t)$ to the queue-length process has continuous paths. In particular, it holds in the case of *heavy-tailed* service times, where the approximation to the load is a Lévy stable-α process with $\alpha \in (1, 2)$. Stable processes are common in queueing models (e.g., [24, 20]).

Theorem 5 provides the heavy-traffic approximation for the queue-length processes. Our tightness (in the Skorokhod topology) of queue-length argument rests on asymptotics for the distribution of a super-near-critical Galton–Watson tree (cf. Lemma 8) where the offspring distribution has finite variance. The weak convergence of queue-length processes for heavy-tailed service times remains an open problem (we discuss this briefly in Section 3.4.4). The Brownian motion approximation (Theorem 5) to queue length is analogous to the Brownian excursion approximation to *depth-first search* walk of a large (conditioned on total size) Galton–Watson tree, Aldous [2] (see Section 3.4.3). For some other interesting relations between queues and trees, we refer the reader to Kersting and Geiger [16] and Shalmon [32].

The paper is organized as follows. Sections 1.1 and 1.2 introduce basic concepts and some important relations. Section 2 is a brief analysis of the workload process in heavy-traffic. Section 3 is devoted to the heavy-traffic limit theorems for the RES-measure (cf. Section 1.2) and the queue-length processes. We discuss some consequences and related work in Section 3.4, and give the directions for further research in Section 4.

1.1. *$M/G/1$ LIFO queue load as a Lévy process.* Let $(X_t, t \geq 0)$ be the Lévy process obtained by superimposing positive discrete jumps on the shift $-at$, where $a > 0$. More precisely, the jumps occur at the times of increase of a counting Poisson (rate λ) process $A(t)$, the sizes of the jumps v_i, $i \geq 1$, are i.i.d. random variables with distribution F concentrated on $(0, \infty)$, and in between the jumps the process is linear with constant negative drift $-a$. The Lévy characterization of X is

$$E\exp(-xX_t) = \exp\{tax + t\lambda \int_{(0,\infty)} (e^{-xr} - 1)\, F(dr)\}, \qquad x > 0, \tag{1}$$

and a *Lévy measure* of X is $\pi(dx) = \lambda F(dx)$. See Section 3.1 for further definitions and Bertoin [5] for background on Lévy processes. We can assume by scaling that a equals 1. Then (as noted in [15]) X is the *load* process

$$X_t = \sum_{i=1}^{A(t)} v_i - t, \qquad t \geq 0, \tag{2}$$

of an M/G/1 LIFO queue with customers arriving at the times of the jumps of X, and requesting service equal to sizes of the jumps. It is also clear that the load process X of any M/G/1 LIFO queue is a Lévy process of the above form.

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 $\gamma_s = \inf\{u \geq s\colon X_u \leq X_{s-}\}$. We identify the actual set of times when this customer is in service with the set $\mathscr{A}_s = \{u \in [s, \gamma_s]\colon X_u \in [X_{s-}, X_s]$ and $\inf_{t\in[s,u]} X_t \geq X_u\}$, indicated in bold on the time axis in the figure. At time γ_s, this customer exits the queue; in the meantime, its service might be interrupted several times due to jumps of X, that is, arrivals of new customers. The "gaps" in $\mathscr{A}_s$ correspond to services of these 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,

$$X_{s-} < \inf_{u\in[s,t]} X_u \tag{3}$$

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

$$Z_t = \#\left\{s \leq t\colon X_{s-} < \inf_{s\leq u\leq t} X_u\right\}. \tag{4}$$

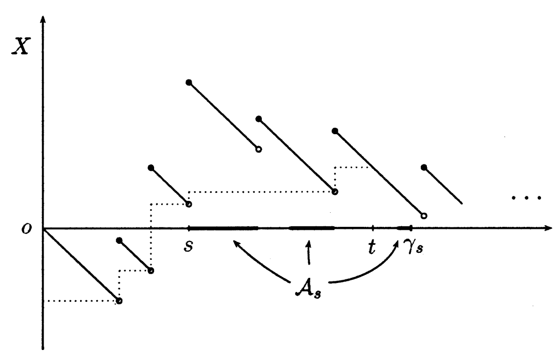

FIG. 1.

1.2. *Branching and the RES-measure.* The relation between queueing and branching goes back to Kendall [26]. Some highlights of the literature are given in [15]. 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 (3) holds. Any customer either finds the queue empty upon arrival, in which case it becomes a *progenitor* (or *root*), or finds the queue nonempty, in which case it becomes a *child* (or *offspring*) of the customer being served previous to its arrival. This procedure yields a sequence of Galton–Watson trees, with each busy cycle corresponding to a different progenitor. The corresponding offspring distribution depends on the Lévy measure of X [15]; it is easy to see that the mean equals λm and the variance equals $\lambda^2\beta + \lambda m - \lambda^2 m^2$, where β is the second moment of F. We return to this very useful characterization in the heavy-traffic analysis, Section 3.3.

Let $I_t = \inf_{s\in[0,t]} X_s$ be the past infimum process, and let $I_s^t = \inf_{u\in[s,t]} X_u$ be the *future infimum* (dotted line in Figure 1). Note that $-I_t$ equals the *(cumulative) idle time*, the amount of time for which there has been no customer in queue up to t. This is true since $-I_t$ increases at constant rate 1 during the time intervals with no customers in queue. 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 workload process is then given by $W_t = X_t - I_t$. The excursions of X above its past infimum, or equivalently, the excursions of W above 0, correspond to the busy cycles of the queue. The relation between excursions of random walks and branching goes back to Harris [18].

Let $\tau_x = \inf\{s \geq 0\colon I_s \leq -x\}$, and let $M_x = \#\{s \in [0, \tau_x]\colon X_{s-} < X_s$ and $W_s = Z_s = 0\}$ be the number of customers arriving to an 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 $[0, \tau_x]$. Observe that

$$M_x \stackrel{d}{=} \text{Poisson (rate } \lambda x), \tag{5}$$

where λ is the arrival rate.

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. Of course, one can obtain the above list from $(X_s,\ s \leq t)$, the path of the load process up to t, or from the path of the workload W up to t. The state space of (finite and infinite) lists appeared in [15], and it seems convenient for certain types of analysis of evolution of the queue, e.g., for obtaining the stationary distribution and the dual process (cf. [15], Section 3). However, it is not convenient for the heavy-traffic analysis since the lists "become uncountable" in the limit, which is related to the fact that the above discrete branching mechanism converges under the heavy-traffic assumptions to a continuum branching mechanism driven by Brownian excursions. A crucial ingredient for this paper is the existence of a measure-valued encoding q_t of the state space, analogous to the *exploration process* of [14], that has a natural extension in the limit. The queue length is recoverable from q by

$$Z_t = \sup(\mathrm{Supp}(q_t)), \tag{6}$$

where $\mathrm{Supp}(\mu)$ denotes the closed support of measure μ. Moreover, the list of residual service times at time t equals $(q_t(1),\ q_t(2), \dots, q_t(Z_t))$, the list of masses of atoms of q_t. Of course, the workload is then given by $W_t = \sum_{i=1}^{Z_t} q_t(i) = \int_0^t dI_s^t = \langle q_t, 1\rangle$. The process q_t is defined by

$$\langle q_t, \varphi\rangle := \int_0^t \varphi(Z_s)\, dI_s^t = \int_0^t \varphi(Z_s^t)\, dI_s^t. \tag{7}$$

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

$$Z_s^t = \#\Big\{u \le s\colon X_{u-} < \inf_{u \le z \le t} X_z\Big\} \tag{8}$$

is the number of individuals in queue at time s that will still be in queue at time t. Note that the integrals in (7) are in fact finite sums. The equality in (7) is due to a simple fact: $Z_s \equiv Z_s^t$, $I^t(ds)$-a.e., for each $t \ge 0$, where the step function I_s^t defines an atomic measure $I^t(ds)$ in the usual way. Note that $I_u^s = I_u^t$ for all $u \le s$, $I^t(ds)$ almost everywhere. It is easily seen (and shown in [15]) that the process q_t is strong Markov. The process Z_s^t is nondecreasing in s for each fixed t, and this monotonicity will be essential in the heavy-traffic analysis (Section 3).

The following observation will be important later on 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 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 $\widetilde{X}_s^t = X_t - X_{(t-s)-}$. In particular, the queue length Z_t, which equals the number of jumps of the future infimum (4), also equals the number of jumps of the time-reversed supremum process occurring in $[0, t]$. The precise statements and their generalizations are deferred until Section 3.1.

We prefer integrals to sums in (7) since the heavy-traffic limit Theorem 1 involves the convergence of rescaled q's to a limit of the same form. The queue length is usually denoted by Z_t. In [15], the corresponding process is denoted by H and is called the *height*. (The two processes are analogous, though H is defined in discrete time and the walk H, unlike Z, never visits the same vertex twice.) The height process (the queue length) visits the vertices (the customers) of the sequence of trees (busy cycles) in the *depth-first search* order (children before siblings), recording their distance from the root. The exploration process "explores" these trees in a similar way, carrying a lot of additional information, and its advantage over the height process is the Markov property and relation (6). The process Z_t is Markov only if the service time distribution F is exponential. From now on we identify any queue with the corresponding measure-valued q, which we call the *RES-measure* (derived from *REsidual Services measure*) process.

2. The workload in heavy traffic. In this section we describe the framework of heavy traffic. We state some of the usual assumptions (e.g., [7, 37]) and discuss asymptotics of the workload processes. The workload does not depend on the service discipline and is a relatively simple object for analysis. At the same time, the workload is the simplest (interesting) process related to the queue, so any "natural" convergence of queues should comprise the convergence of corresponding workloads.

Let $q^r = (q^r_t,\ t \geq 0)$ be a family of RES-measure processes of M/G/1 LIFO queues, indexed by r. Here r ranges over real numbers; it is easiest to think of a sequence increasing to ∞. The rth M/G/1 queue has the arrival rate λ^r and the service time distribution function F^r with mean m^r. Assume

$$\lambda^r \to \lambda \in (0,\infty), \qquad m^r \to m \in (0,\infty) \qquad \text{as } r \to \infty. \tag{9}$$

Let $\rho^r = \lambda^r m^r$. A usual heavy-traffic assumption is

$$r^{1/2}(1-\rho^r) = r^{1/2}(1-\lambda^r m^r) \to c \in R \qquad \text{as } r \to \infty. \tag{10}$$

Let $A^r(\cdot)$, $W^r(\cdot)$ and $Z^r(\cdot)$ denote the corresponding arrival, workload and queue-length processes. We assume that, for each r, the queue is empty at time 0, or equivalently, $W^r(0) = 0$, so that notation of Sections 1.1 and 1.2 directly applies. More general initial conditions can be treated with additional work (cf. Section 3.4.1).

Let v^r_i be the service time requested by the ith customer who arrives to the queue. So $\{v^r_i,\ i \geq 1\}$ is an i.i.d. sequence with distribution F^r, and denote by $V^r(n) = \sum_{i=1}^n v^r_i,\ n \geq 1$, the *cumulative service time* process. The *workload equation* is (the same for all work-conserving disciplines)

$$W^r(t) = V^r(A^r(t)) - t - I^r(t) = X^r(t) - I^r(t), \tag{11}$$

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

If we assume in addition to (9), (10) that, for each r, the service times have second moment $\beta^r < \infty$ and

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

$$\sup_r E\big[(v^r_1)^2 1_{\{v^r_1 \geq K\}}\big] \to 0 \qquad \text{as } K \to \infty \tag{13}$$

(so a Lindeberg–Feller type of condition is satisfied), the rescaled load processes $(r^{-1/2}X^r(rt),\ t \geq 0)$ converge weakly to a Brownian motion X with drift $-c$, defined in (10), and variance $\lambda\beta$. The value $\lambda\beta$ for the asymptotic variance can be verified using standard arguments; intuitively, it is due to the fact that the infinitesimal drift $E(\Delta(X^r_t)^2 \mid \mathscr{F}^r_t)$ equals $\lambda^r E(v^r_1)^2\, dt = \lambda^r \beta^r dt$ (in the obvious notation), so in the limit $X^2_t - \lambda\beta t$ should be (and is) a martingale. By the continuous mapping theorem (e.g., Billingsley [6], Lemma 6.1), under the same scaling, the workload processes W^r converge to $W = X - I$ obtained by reflecting X above the past minimum $I_t = \inf_{s\leq t} X_s$.

REMARK. Since Brownian motion X is a stable process with index $\alpha = 2$, (10) reads

$$r^{1-\gamma}(1-\rho^r) \to c\,, \tag{14}$$

where $\gamma = 1/\alpha = 1/2$. It is standard that $(A^r(rt)/r,\ t \ge 0) \Rightarrow (\lambda t,\ t \ge 0)$, as $r \to \infty$. Now consider a more general setting, where $\alpha \in (0,2]$ and $\gamma = 1/\alpha$. Assume that $\lim_{r\to\infty} b^r/r^\gamma$ exists, that (9), (14) hold, and consider the rescaled processes

$$\begin{aligned} X^r(rt)/b^r &= \frac{V^r(A^r(rt)) - rt}{b^r} \\ &= \frac{\sum_{i=1}^{A^r(rt)}(v_i^r - m^r)}{b^r} + \frac{A^r(rt)m^r - rt}{b^r} \\ &\sim \frac{\sum_{i=1}^{r\lambda t}(v_i^r - m^r)}{b^r} + \frac{r(\rho^r - 1)t}{b^r}. \end{aligned}$$

Assume moreover that, for all large r, the service times have *heavy tails*, that is, F^r is in the domain of attraction of the stable-α law (cf. Breiman [8], page 207). Assume that $m < \infty$ so it must be $\alpha \in (1,2]$. Then (e.g., Jacod and Shiryaev [23]) $X^r(rt)/b^r$ converges to a stable-α process X, and again the workload converges to X reflected above the past minimum. Under more general conditions on b^r and F^r (see [23], Theorem VII.2.35 and [15], Proposition 5.1), $X^r(rt)/b^r$ will converge to a Lévy process X with no negative jumps [i.e., the Lévy measure π in (16) is concentrated on $(0,\infty)$].

3. Heavy-traffic limits. In this section we state and prove the heavy-traffic limit theorems for M/G/1 LIFO queues. In Section 3.1 we define the limit processes and mention some of their properties from Le Gall and Le Jan [14, 15]. Sections 3.2 and 3.3 are devoted to the convergence, and Section 3.4 comments on some consequences and extensions, and relates our result to the existing literature.

3.1. *Limit processes.* We briefly describe the setting of [14, 15]. Let $X = (X_t,\ t \ge 0)$ be a Lévy process with no negative jumps such that $\liminf_{t\to\infty} X_t = -\infty$. Then

$$E\exp(-xX_t) = \exp(t\psi(x)), \qquad x > 0, \tag{15}$$

where the *Laplace exponent* $\psi(x)$ is of the form

$$\psi(x) = cx + \frac{\sigma^2 x^2}{2} + \int_{(0,\infty)} (e^{-xr} - 1 + xr)\pi(dr), \qquad c \ge 0, \tag{16}$$

and the *Lévy measure* $\pi(dr)$ satisfies

$$\int_{(0,\infty)} (r \wedge r^2)\pi(dr) < \infty. \tag{17}$$

In Section 1.1 we identified the load processes (2) of M/G/1 LIFO queue with a class of analogous (though much simpler) Lévy processes characterized by (1). Moreover, in Section 2 we saw how some of the processes characterized by (15)–(17) arise naturally as limits of rescaled load processes of queues approaching heavy traffic. Such X can be viewed as a "generalized" queue load, and it is plausible that a generalized queue length Z_t and a generalized RES-measure q_t can be obtained from the load by mimicking (4) and (7). Indeed, this has been done in [14, 15] [for X with general Laplace exponent (16), (17)], and we recall their definitions here briefly. We will mainly deal with the special Brownian case $\psi(x) = cx + \sigma^2 x^2/2$, $c \in (-\infty, \infty)$, $\sigma > 0$. If the drift $-c$ is strictly positive, then $\lim_{t\to\infty} X_t = \infty$ (not $-\infty$), and such processes were not considered in [14, 15]. However, the definitions below and the rest of the analysis extend naturally.

Let $X^\bullet$ be a (not generalized) M/G/1 LIFO queue load process as in Section 1.1. In this section only, all the LIFO queue-related processes from Section 1 have additional "$\bullet$" in the superscript. Recall the related infimum processes $I^\bullet_t$, $I^{\bullet,\,t}_s$ from Section 1.2. For each fixed $t > 0$, denote by $(\widetilde{X}^{\bullet,\,t}_s,\ 0 \le s \le t)$ the time-reversed process $X^\bullet$ from t, that is,

$$\widetilde{X}^{\bullet,\,t}_s = X^\bullet_t - X^\bullet_{(t-s)-}, \quad 0 \le s < t \qquad \text{and} \qquad \widetilde{X}^{\bullet,\,t}_t = X^\bullet_t,$$

and let $\widetilde{S}^{\bullet,\,t}_s = \sup_{u\in[0,\,s]} \widetilde{X}^{\bullet,\,t}_u$. Rewrite identities (4) and (8) as

$$\begin{aligned} Z^\bullet_t &= \#\Big\{z\colon z \in [0, t],\ \widetilde{S}^{\bullet,\,t}_z > \widetilde{S}^\bullet_{z-}\Big\}, \\ Z^{\bullet,\,t}_s &= \#\Big\{z\colon z \in [t-s, t],\ \widetilde{S}^{\bullet,\,t}_z > \widetilde{S}^{\bullet,\,t}_{z-}\Big\}, \end{aligned} \tag{18}$$

since the future infimum $I^{\bullet,\,t}_s$ corresponds (in reversed time) to the past supremum $\widetilde{S}^{\bullet,\,t}_{t-s-}$.

Now fix a generalized queue load process X, and let $(\widetilde{X}^t_s,\ 0 \le s \le t)$ be the corresponding time-reversed process, so that

$$\begin{aligned} &\widetilde{X}^t_s = X_t - X_{(t-s)-}, \qquad 0 \le s < t, \\ &\widetilde{X}^t_t = X_t \qquad \text{and} \qquad \widetilde{S}^t_s = \sup_{u\in[0,\,s]} \widetilde{X}^t_u. \end{aligned}$$

Now the set of increase points of $\widetilde{S}^t$ is measured using local time. Let $(\widetilde{L}^t_s,\ 0 \le s \le t)$ be a local time of the process $(\widetilde{S}^t_s - \widetilde{X}^t_s,\ 0 \le s \le t)$ at level 0. Then $\widetilde{L}^t_s$ is a continuous nondecreasing *additive* process with support on the zero set of $\widetilde{S}^t_s - \widetilde{X}^t_s$ (see, e.g., [5], Chapter IV). By analogy to (18), define

$$Z^t_s = \widetilde{L}^t_t - \widetilde{L}^t_{t-s}, \qquad 0 \le s \le t \qquad \text{and} \qquad Z_t = Z^t_t = \widetilde{L}^t_t. \tag{19}$$

Local time is unique ([5], Proposition IV.2.5) up to a multiplicative constant. For X a Brownian motion with drift $-c$ and variance σ^2, we can choose $\widetilde{L}^t$ as

$$\widetilde{L}^t_s = \frac{2}{\sigma^2} \widetilde{S}^t_s, \tag{20}$$

which translates back to

$$
\begin{aligned}
Z_s^t &= \frac{2}{\sigma^2}(I_s^t - I_t), \qquad t \ge 0,\\
Z_t &= \frac{2}{\sigma^2}(X_t - I_t) = \frac{2}{\sigma^2} W_t,
\end{aligned} \tag{21}
$$

where $I_t = \inf_{s \le t} X_t$ and $I_s^t = \inf_{u \in [s,t]} X_u$ as always. The choice of the normalizing factor in (20) is motivated by heavy-traffic limits (e.g., Lemma 2). Note that the processes Z_t, Z_s^t in (21) are continuous. By (21), Z_s^t is nondecreasing in s for every t and $\lim_{s\uparrow t} Z_s^t = Z_t$. Moreover, it is straightforward to see that again $Z_s \equiv Z_s^t$, $I^t(ds)$-a.e., $t \ge 0$. So definition (7) extends to the generalized setting. The finite-measure-valued process q_t given by

$$
\langle q_t, \varphi\rangle = \int_0^t \varphi(Z_s)\, I^t(ds) = \int_0^t \varphi(Z^t(s))\, I^t(ds)
$$

is a strong Markov process, the identity (6) carries over, and moreover,

$$
\mathrm{Supp}(q_t) = [0, Z_t], \qquad t \ge 0.
$$

In fact, q_t is a constant $(\sigma^2/2)$ multiple of the Lebesgue measure on $[0, Z_t]$. The above statements can be easily checked in our (Brownian) setting, and some have analogues in the more general setting (16), (17) of [14].

3.2. *Convergence.* We are now ready to state our main result. Denote by $M_f(R_+)$ the complete, separable metric space of finite measures on $[0,\infty)$ (cf. Billingsley [6], or Dawson [10], Section 3), by $D_R[0,\infty)$, $D_R[0,t]$, $D_{M_f(R_+)}[0,\infty)$ the usual Skorokhod spaces, and by $\Rightarrow$ the corresponding weak convergence of processes.

Let $q^r = (q_t^r, t \ge 0)$, $r \ge 1$, be a family of RES-measures with corresponding queue load processes $X^r = (X_t^r,\ t \ge 0)$, as in Sections 1 and 2. Similarly, denote by $Z^r, Z^{t,r}$ the queue-length processes in (6) and (8), and let $I_\cdot^{t,r}$ be the future infimum processes of X^r.

Assume that (9), (10) and (12), (13) hold. Then we know (Section 2) that $\widehat{X}^r = (r^{-1/2} X^r(rt),\ t \ge 0) \Rightarrow X$, where X is a Brownian motion with variance $\lambda\beta$ and drift $-c$, and by the Skorokhod representation theorem, we may assume that

$$
\widehat{X}^r \to X \quad \text{a.s. in } D_R[0,\infty) \qquad \text{as } r \to \infty. \tag{22}
$$

For each $r \ge 1$, let $\widehat{I}_s^{t,r} \equiv \widehat{I}^{t,r}(s) := r^{-1/2} I^{t,r}(rs)$, and

$$
\widehat{Z}_t^r = r^{-1/2} Z_{rt}^r, \qquad \widehat{Z}^{t,r}(s) = r^{-1/2} Z^{rt,r}(rs), \tag{23}
$$

$$
\langle \hat{q}^r(t), \varphi\rangle = \int_0^t \varphi(\widehat{Z}_s^r)\, \widehat{I}^{t,r}(ds) = \int_0^t \varphi(\widehat{Z}^{t,r}(s))\, \widehat{I}^{t,r}(ds). \tag{24}
$$

Convergence in (22) implies that, for t fixed,

$$-\widehat{I}^{t,r}(\cdot) \to -I^t(\cdot) \quad \text{a.s. in } D_{R_+}[0,t] \qquad \text{as } r \to \infty, \tag{25}$$

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

Let Z_t, $Z^t(s) \equiv Z^t_s$ be as in (21). Note that $\widehat{X}^r$, X, $\widehat{Z}^r$, $Z \in D_{R_+}[0,\infty)$, $\widehat{Z}^{t,r}(\cdot)$, $Z^t(\cdot) \in D_{R_+}[0,t]$ and $\hat{q}^r$, $q \in D_{M_f(R_+)}[0,\infty)$.

THEOREM 1. *Under assumptions* (9), (10), (12), (13), *we have* $\hat{q}^r \Rightarrow q$, *as* $r \to \infty$.

The proof is based on (25) and the following lemma:

LEMMA 2. *Let* $\widehat{Z}^{t,r}(s)$ *be as in* (23). *Then, for each fixed* $t \geq 0$,

$$P\left(\sup_{s\in[0,t]} |\widehat{Z}^{t,r}(s) - Z^t(s)| > \varepsilon\right) \to 0 \qquad \textit{as } r \to \infty.$$

PROOF. Assume (22), fix $t > 0$ and a finite subdivision $0 \leq s_1 < s_2 < \cdots < s_k = t$ on $[0,t]$. We show

$$\begin{aligned}&\left(\widehat{Z}^{t,r}(s_1), \widehat{Z}^{t,r}(s_2), \dots, \widehat{Z}^{t,r}(s_k)\right)\\ &\qquad \xrightarrow{p} \left(Z^t(s_1), Z^t(s_2), \dots, Z^t(s_k)\right), \qquad r \to \infty,\end{aligned} \tag{26}$$

where $\to^p$ denotes convergence in probability. Recall $\widehat{Z}^{t,r}(s) = \#\{u\colon t-s \leq u \leq t, \widehat{\widetilde{S}}^{t,r}(u) > \widehat{\widetilde{S}}^{t,r}(u-)\}\cdot r^{-1/2}$, where $\widehat{\widetilde{S}}^{t,r}(u) = \sup_{x\in[0,u]} \widehat{\widetilde{X}}^{rt,r}(x)$ is the supremum process of $\widehat{\widetilde{X}}^{rt,r}(s) = r^{-1/2}(X^r(rt) - X^r((rt-rs)-))$, the rescaled and time-reversed X^r.

Consider the time-reversed process $\widetilde{X}^{rt,r}(s) = X^r(rt) - X^r((rt-s)-)$, $0 \leq s \leq rt$. For $z \leq t$, 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 $\widehat{Z}^{t,r}(s) = r^{-1/2}(M^r_t(t) - M^r_t((t-s)-)) = r^{-1/2}(M^r_t(t) - M^r_t(t-s))$ since a.s. there is no jump at time $t-s$. Due to (19), (20), it suffices to show that, for each fixed $z \in [0,t]$,

$$r^{-1/2} M^r_t(z) \xrightarrow{p} \frac{2}{\lambda\beta}\widetilde{S}^t(z), \qquad r \to \infty, \tag{27}$$

where $\widetilde{S}^t(z)$ is defined above (19). The lemma will then follow from (26), since $\widehat{Z}^{t,r}(\cdot)$ is nondecreasing for each r and t, and $Z^t(\cdot)$ is continuous and nondecreasing for each t.

In order to show (27), it is convenient to consider "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 $\{u^r_{-i}, i \geq 1\}$ of i.i.d. exponential (rate λ^r) random variables, and a sequence $\{v^r_{-i}, i \geq 1\}$ of i.i.d. random variables with distribution F^r. Define $\widetilde{X}^{rt,r}(rt+z) = \widetilde{X}^{rt,r}(rt)+\sum_{i=1}^{A_-^r(z)} v_{-i}^r - z$, $z \geq 0$, where $A_-^r(z) = \sup\{j\colon \sum_{i=1}^{j} u_{-j}^r \leq z\}$. Then the extended process $\widetilde{X}^{rt,r}$ has the distribution of the load process X^r.

Let $\widetilde{S}^{rt,r}$ be the supremum process of the extended $\widetilde{X}^{rt,t}$. Denote by $T_1^{r,t} < T_2^{r,t} < \cdots$ the successive increase (jump) times of $\widetilde{S}^{rt,r}$, and let $J_i^{r,t} = \widetilde{X}^{rt,r}(T_i^{r,t}) - \widetilde{S}^{rt,r}(T_i^{r,t}-)$, $i \geq 1$, be the corresponding *overshoots*. If $\lambda^r m^r \geq 1$, there are infinitely many overshoots almost surely, and they form an i.i.d. sequence of random variables. The distribution of J_1^r is known,

$$\frac{P(J_1^r = dz)}{dz} = \lambda^r \int_{[z,\infty)} \exp\{-\Phi^r(0)(y-z)\} F^r(dy), \qquad z \geq 0, \tag{28}$$

where $\Phi^r(0) \geq 0$ (see [5], page 188 for interpretation) is such that the right-hand side of (28) defines a (proper) probability distribution. If $\lambda^r m^r < 1$, there are $N_{r,t}$ many overshoots, where $N_{r,t}$ is a geometric $(1-\lambda^r m^r)$ random variable, and conditionally on $N_{r,t}$, the overshoots $(J_1^{r,t}, \ldots, J_{N_{r,t}}^{r,t})$ are independent and identically distributed, with known distribution

$$\frac{P(J_1^r = dz |\text{ overshoot occurs })}{dz} = \frac{1}{m^r} F^r([z,\infty)), \qquad z \geq 0. \tag{29}$$

Therefore, $E(J_1^r|\text{ overshoot occurs}) = \beta^r/(2m^r)$ in this case. Both (28) and (29) are special cases of [5], Theorem VII.17(ii).

We are mainly interested in the overshoots that occurred by (reversed) time $rz \leq rt$. Note that

$$\widehat{\widetilde{S}}^{t,r}(z) = \sum_{i=1}^{M_t^r(z)} \frac{1}{r^{1/2}} J_i^r, \tag{30}$$

and the convergence in (22) implies

$$\widehat{\widetilde{S}}^{t,r}(z) \to \widetilde{S}^t(z) \quad \text{a.s.} \quad \text{as } r \to \infty. \tag{31}$$

Assertion (27) is now a consequence of (30), (31) and Lemma 3 below. □

LEMMA 3.

$$\frac{1}{M_t^r(z)} \sum_{i=1}^{M_t^r(z)} J_i^r \xrightarrow{p} \frac{\lambda\beta}{2} \qquad \textit{as } r \to \infty.$$

PROOF. An adaptation of the law of large numbers. It suffices to consider subsequences r_k of r for which either $\lambda^{r_k} m^{r_k} \geq 1$ for all k, or $\lambda^{r_k} m^{r_k} < 1$ for all k. Denote such subsequences again by r.

Assume $\lambda^r m^r \geq 1$. Note that

$$\Phi^r(0) \text{ in (28) tends to } 0 \qquad \text{as } r \to \infty. \tag{32}$$

This is due to (9), (10), (12), (13). Suppose $\Phi^r(0) > \delta$ along a subsequence, for some $\delta > 0$. By (9) and the Markov inequality, the sequence of service time distributions F^r is tight. So we may assume that F^r converges weakly to distribution F along the same subsequence. Distribution F has mean m and second moment β, due to (12), (13). Let $v_1^r =^d F^r$ and $v_1 =^d F$. Due to

$$1 = P(J_1^r \geq 0) = \lambda^r \int_{[0,\infty)} \exp\{-\Phi^r(0)x\} F^r([x,\infty))\,dx,$$

$m^r\lambda^r \to 1$, and $1 \geq \exp\{-\delta x\} > \exp\{-\Phi^r(0)x\}$, an application of the sandwich theorem gives

$$\begin{aligned}\lim_r \lambda^r \int_{[0,\infty)} e^{-\delta x} F^r([x,\infty))\,dx &= \frac{\lambda}{\delta} \lim_r E(1-\exp(-\delta v_1^r))\\ &= \frac{\lambda}{\delta} E(1-\exp(-\delta v_1)) = 1,\end{aligned}$$

where the limit is taken along the subsequence above. The last equality is impossible, since $E(1-\exp(-\delta v_1)) \leq 1-\exp(-\delta E v_1) = 1-e^{-\delta m}$ and $\lambda(1-e^{-\delta m})/\delta$ is strictly smaller than $\lambda m = 1$.

Due to (13), random variables J_1^r in (28) are uniformly integrable. Namely, an application of Fubini's theorem gives

$$\begin{aligned}\lim_{K\to\infty} EJ_1^r 1_{\{J_1^r \geq K\}} &\leq \lim_{K\to\infty} \lambda^r \int_{[K,\infty)} zF^r([z,\infty))\,dz\\ &\leq \lim_{K\to\infty} \frac{1}{2}\int_{[K,\infty)} z^2 F^r(dz) = 0.\end{aligned}$$

Since

$$\begin{aligned}EJ_1^r &= \int_0^\infty P(J_1^r \geq y)\,dy\\ &= \lambda^r \int_0^\infty F^r(dz) \int_0^\infty dy \int_0^\infty dx\, \exp(-\Phi^r(0)x) 1_{\{x\geq 0,\, y\geq 0,\, x+y\leq z\}},\end{aligned}$$

by (32) and uniform integrability we have $EJ_1^r \to (\lambda\beta)/2$ as $r \to \infty$. So, it suffices to show

$$\frac{1}{M_t^r(z)} \sum_{i=1}^{M_t^r(z)} (J_i^r - EJ_1^r) \xrightarrow{p} 0 \qquad \text{as } r \to \infty.$$

It is not hard to see that $M_t^r(z) \to \infty$ in probability, due to (30) and (31). Take $\varepsilon' > 0$. Since

$$\begin{aligned}&P\left(\left|\frac{1}{M_t^r(z)} \sum_{i=1}^{M_t^r(z)} (J_i^r - EJ_1^r)\right| > 4\varepsilon'\right)\\ &\qquad \leq P(M_t^r(z) < l) + P\left(\sup_{k\geq l}\left|\frac{1}{k}\sum_{i=1}^k (J_i^r - EJ_1^r)\right| > 4\varepsilon'\right),\end{aligned}$$

it suffices to show that

$$P\left(\sup_{k\geq l}\left|\frac{1}{k}\sum_{i=1}^k (J_i^r - EJ_1^r)\right| > 4\varepsilon'\right) \leq o(r,l)\,, \tag{33}$$

where $\lim_{l\to\infty}\sup_r o(r,l)=0$. Note that (33) is just an extension of the usual strong law of large numbers. The proof is a variation on the classical proof. Fix large K and split the sum in (33) into

$$\sum_{i=1}^{k}\Big(J_i^r 1_{\{J_i^r\le K\}} - EJ_i^r 1_{\{J_i^r\le K\}}\Big) + \sum_{i=1}^{k}\Big(J_i^r 1_{\{J_i^r> K\}} - EJ_i^r 1_{\{J_i^r> K\}}\Big).$$

Random variables $J_i^r 1_{\{J_i^r\le K\}}$ are bounded; in particular, they have uniformly (in r) bounded fourth moments and the simplest proof of the strong law of large numbers yields

$$P\bigg(\sup_{k\ge l}\bigg|\frac{1}{k}\sum_{i=1}^{k}\Big(J_i^r 1_{\{J_i^r\le K\}} - EJ_1^r 1_{\{J_i^r\le K\}}\Big)\bigg| > 2\varepsilon'\bigg) \\ \le o(r,l,K)$$

where $\lim_{l\to\infty}\sup_r o(r,l,K) = 0$. By uniform integrability, $EJ_1^r 1_{\{J_1^r>K\}} = \frac{1}{k}\sum_{i=1}^{k} EJ_i^r 1_{\{J_i^r>K\}}$ can be made uniformly small in r for large enough K. So it suffices to show

$$\lim_{l\to\infty}\sup_r P\bigg(\sup_{k\ge l}\frac{1}{k}\sum_{i=1}^{k} J_i^r 1_{\{J_i^r>K\}} > \varepsilon'\bigg) \le o(K), \tag{34}$$

where $\lim_K o(K) = 0$. Perhaps the easiest way to show (34) is by imitating Etemadi's proof of SLLN [11], Theorem 1.8.4. Define truncated $H_i^{r,K} = J_i^r 1_{\{J_i^r>K\}} 1_{\{J_i^r 1_{\{J_i^r>K\}}\le i\}}$. Then

$$\sup_r P\Big(H_i^{r,K} \ne J_i^r 1_{\{J_i^r>K\}} \text{ for some } i\ge 1\Big) \le \sup_r EJ_1^r 1_{\{J_1^r>K\}} \le o(K),$$

where $\lim_K o(K) = 0$, so it suffices to show

$$\lim_{l\to\infty}\sup_r P\bigg(\sup_{k\ge l}\frac{1}{k}\sum_{i=1}^{k} H_i^{r,K} > \varepsilon'\bigg) \le o(K).$$

Fix $\varepsilon'' > 0$ and $a > 1$. Let $T_n^{r,K} = \sum_{i=1}^{n} H_i^{r,K}$, and let $k(n) = \lfloor a^n\rfloor$. By the same calculation as in Etemadi's proof, we get

$$\sup_r \sum_{n=1}^{\infty} P\Big(\Big|T_{k(n)}^{r,K} - ET_{k(n)}^{r,K}\Big| > \varepsilon'' k(n)\Big) \\ \le \sup_r \frac{16}{(1-a^2)\varepsilon''^2} EJ_1^r 1_{\{J_1^r>K\}} \le o(K).$$

Since

$$\sup_K\sup_r\Big|EH_i^{r,K} - EJ_1^r 1_{\{J_1^r>K\}}\Big| \to 0 \qquad \text{as } i\to\infty,$$

then $\sup_r |ET^{r,K}_{k(n)}/k(n) - EJ^r_1 1_{\{J^r_1>K\}}| \to 0$, as $n \to \infty$, and therefore,

$$\lim_{l\to\infty} \sup_r P\left(\sup_{k(n)\geq l} \frac{1}{k(n)} T^{r,K}_{k(n)} > 2\,\varepsilon''\right) \leq o(K).$$

The rest is the same as in Etemadi's proof. Details are left to the reader.

Assume $\lambda^r m^r < 1$. By the same reasoning as above, again $M^r_t(z) \to^p \infty$, as $r \to \infty$. Similarly, the distributions in (29) are again uniformly integrable. Recall $E(J^r_1|$ overshoot occurs$) = \beta^r/(2m^r) \to \lambda\beta/2$. Write

$$\begin{aligned}
P&\left(\left|\frac{1}{M^r_t(z)} \sum_{i=1}^{M^r_t(z)} \left(J^r_i - \frac{\lambda^r\beta^r}{2}\right)\right| > 4\varepsilon'\right)\\
&\leq P(M^r_t(z) < l) + P\left(\sup_{l\leq k\leq N_{r,t}} \left|\frac{1}{k}\sum_{i=1}^{k}\left(J^r_i - \frac{\lambda^r\beta^r}{2}\right)\right| > 4\varepsilon',\, N_{r,t} \geq l\right)\\
&\leq P(M^r_t(z) < l)\\
&\quad + \sum_{j=l}^{\infty} P(N_{r,t} = j) P\left(\sup_{l\leq k\leq j}\left|\frac{1}{k}\sum_{i=1}^{k}\left(J^r_i - \frac{\lambda^r\beta^r}{2}\right)\right| > 4\varepsilon' \middle| N_{r,t} = j\right).
\end{aligned}$$

By the observation made above (29), the last term in the sum above is dominated by

$$P\left(\sup_{k\geq l}\left|\frac{1}{k}\sum_i (\overline{J}^r_i - E\overline{J}^r_1)\right| > 4\varepsilon'\right),$$

where $\overline{J}^r_i$, $i \geq 1$, are i.i.d. with distribution (29), and the rest of the proof is the same as in the supercritical case. □

COROLLARY 4. *For any fixed t and $0 \leq t_1 < t_2 \cdots < t_k \leq t$,*

$$\left(\widehat{Z}^r_{t_1}, \widehat{Z}^r_{t_2}, \ldots, \widehat{Z}^r_{t_k}\right) \xrightarrow{p} \left(Z_{t_1}, Z_{t_2}, \ldots, Z_{t_k}\right).$$

PROOF OF THEOREM 1. From Lemma 2 and (22)–(25), we easily obtain the convergence of finite-dimensional distributions

$$\left(\hat{q}^r_{t_1}, \ldots, \hat{q}^r_{t_k}\right) \Rightarrow \left(q_{t_1}, \ldots, q_{t_k}\right), \qquad r \to \infty.$$

So it suffices to show the tightness of the family $(\hat{q}^r)$ in $D_{M_f(R_+)}[0,\infty)$. Let $q = q^r$, fix a bounded and continuous function φ: $[0,\infty) \to R$ and calculate

$$\begin{aligned}
&|\langle q(s+h), \varphi\rangle - \langle q(s), \varphi\rangle|\\
&\quad = \left|\left(\int_0^s + \int_s^{s+h}\right)\varphi(Z^{s+h}_z)\,dI^{s+h}_z - \int_0^s \varphi(Z^s_z)\,dI^s_z\right|\\
&\quad \leq \left|\int_0^s \varphi(Z^{s+h}_z)\,dI^s_z - \int_0^s \varphi(Z^{s+h}_z)\,dI^{s+h}_z\right|\\
&\qquad + \left|\int_s^{s+h} \varphi(Z^{s+h}_z)\,dI^{s+h}_z\right| + \left|\int_0^s \varphi(Z^s_z) - \varphi(Z^{s+h}_z)\,dI^s_z\right|,
\end{aligned}$$

where the inequality is an application of the triangle inequality, after convenient rearrangement of the terms. Use the fact that $I_z^s - I_z^{s+h}$ is nondecreasing in $z \in [0, s]$ in order to bound the first term from above by

$$\begin{aligned}\|\varphi\|_\infty \left| \int_0^s dI_z^s - dI_z^{s+h} \right| \\ &= \|\varphi\|_\infty \big(I_s^s - I_0^s - (I_s^{s+h} - I_0^{s+h})\big) \\ &\le \|\varphi\|_\infty (X_s - I_s^{s+h}).\end{aligned}$$

For the second term, use a similar identity $\int_s^{s+h} dI_z^{s+h} = I_{s+h}^{s+h} - I_s^{s+h} = X_{s+h} - I_s^{s+h}$. To bound the third term, note that Z_z^s and Z_z^{s+h} differ (clearly $Z_z^s \ge Z_z^{s+h}$ for all $z \in [0, s]$) only for $z \in [0, s]$ such that $I_z^s > I_z^{s+h}$, that is $I_z^s > I_s^{s+h}$. Now $\int_0^s (\varphi(Z_z^s) - \varphi(Z_z^{s+h}))\, dI_z^s = \int_{\{z \in [0,s]: I_z^s > I_s^{s+h}\}} (\varphi(Z_z^s) - \varphi(Z_z^{s+h}))\, dI_z^s$, so that

$$\left| \int_0^s (\varphi(Z_z^s) - \varphi(Z_z^{s+h}))\, dI_z^s \right| \le 2\, \|\varphi\|_\infty (I_s^s - I_s^{s+h}).$$

The above calculations imply

$$\begin{aligned}|\langle q(s+h), \varphi\rangle - \langle q(s), \varphi\rangle| &\le \|\varphi\|_\infty \Big(3(X_s - I_s^{s+h}) + (X_{s+h} - I_s^{s+h})\Big) \\ &\le 4\|\varphi\|_\infty \sup_{\theta \in [0, h]} |X_{s+\theta} - X_s|,\end{aligned}$$

and after scaling,

$$|\langle \hat{q}^r(s+h), \varphi\rangle - \langle \hat{q}^r(s), \varphi\rangle| \le 4\|\varphi\|_\infty \sup_{\theta \in [0, h]} |\widehat{X}^r_{s+\theta} - \widehat{X}^r_s|.$$

The tightness of $\hat{q}^r$ now follows from the tightness of $\widehat{X}^r$, by combining Jakubowski and Aldous criteria, [10], Theorems 3.6.4 and 3.6.5 (see also [13], Theorem III.8.6).

REMARK. Since the mapping $\mu \mapsto \sup(\mathrm{Supp}(\mu))$ is clearly not continuous in the topology on $M_f(R_+)$, Theorem 5 below does not immediately follow from Theorem 1 and (6). If X in (22) is a stable-α ($\alpha \in (1, 2)$) and the processes Z_s^t, Z_t are appropriately defined (cf. Section 3.4.4), then assertion (26) becomes a consequence of the analysis in [15], Section 5. Therefore, an analogue of Theorem 1 exists in the stable-α setting, $\alpha \in (1, 2)$.

THEOREM 5. *Under assumptions* (9), (10), (12), (13), *we have* $\widehat{Z}^r \Rightarrow Z$ *as* $r \to \infty$.

The difficulties in analyzing $\widehat{Z}^r$ and Z are related to their lack of Markov property. 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

$$P\left(\sup_{1 \le i \le n} \sup_{u \in [t_{i-1}, t_i]} \left| Z_{t_i} - Z_u \right| > \varepsilon \right) \le \eta, \tag{35}$$

by continuity of Z (cf. [13], page 122). Recall that, for each t, the process $(Z^t(s), 0 \le s \le t)$ given by (21) is continuous. The processes $Z^{t_i}(t_i) - Z^{t_i}(t_i - \theta) = \widetilde{L}^{t_i}_{\theta}$, $\theta \in [0, T/n]$, $1 \le i \le n$, as defined in (19), are all identically distributed. The processes $\widetilde{L}^{t_i}_{\cdot}$ are also independent, due to independent increments of Brownian motion, but this property is not used in the argument. We claim that, for all large n,

$$P\biggl(\sup_{1 \le i \le n} \sup_{\theta \in [0, T/n]} |Z^{t_i}(t_i) - Z^{t_i}(t_i - \theta)| > \varepsilon\biggr) \le \eta. \tag{36}$$

Local time $\widetilde{L}^{t_1}_{\theta}$ is increasing in the variable θ, so it suffices to show that, for any $\varepsilon > 0$, $nP(L_{T/n} > \varepsilon) \to 0$ as $n \to \infty$. This follows from (20) and the following easy lemma. The proof is left to the reader.

LEMMA 6. *For $S_t = \sup_{u \in [0, t]} B_u$, the supremum process of Brownian motion B with variance* 1 *and drift $c \in R$, we have*

$$nP(S_{1/n} > \varepsilon) \to 0 \qquad \text{as } n \to \infty \quad \text{for all } \varepsilon > 0. \tag{37}$$

PROOF OF THEOREM 5. The finite-dimensional distributions of $\widehat{Z}^r$ are converging to those of Z due to Corollary 4. So it suffices to show the tightness of $\widehat{Z}^r$, $r \ge 1$, with respect to the Skorokhod topology on $D_R[0, \infty)$. It suffices to show that any subsequence r_n has a further subsequence r_{n_k} so that $\widehat{Z}^{r_{n_k}}$ is tight. For a given subsequence r_n, find a weakly converging further subsequence $F^{r_{n_k}} \to F$. This is possible again by tightness. Recall that F has mean m and second moment β. To simplify the notation, denote the subsequence r_{n_k} again by r.

Formally, the idea is to use $\widehat{Z}^r_{t_i - \theta} \approx \widehat{Z}^{t_i, r}(t_i - \theta) \approx \widehat{Z}^{t_i, r}(t_i) = \widehat{Z}^r_{t_i} \approx Z_{t_i}$ for small θ, and exploit the monotonicity of $\widehat{Z}^{t, r}_s$ and Z^t_s in s. Let $\mathscr{F}^r_t$ be the filtration generated by $\widehat{X}^r$. Let $t_i = iT/n$ as above, and $t \in [t_{i-1}, t_i]$. Observe that, for each r,

$$\widehat{Z}^{t_i, r}_{t_{i-1}} \le \widehat{Z}^{t_i, r}_t \le \widehat{Z}^r_t \le \widehat{Z}^r_{t_{i-1}} + \widehat{Z}^{r, i}_{t - t_{i-1}}, \tag{38}$$

where $(\widehat{Z}^{r, i}_u, u \in [0, T/n])$ has the same law as $(\widehat{Z}^r_u, u \in [0, T/n])$, and is independent of $\mathscr{F}^r_{t_{i-1}}$. The first inequality in (38) is the monotonicity of $Z^{t, r}_s$ in s; the second inequality trivially follows from the interpretation of $\widehat{Z}^{t_i, r}_t$ as the (rescaled) number of individuals in queue at time rt whose service will not have been completed by time rt_i. The last inequality in (38) is a special case of [15], Lemma 4.5, though it can be argued using again queueing interpretation: $\widehat{Z}^{r, i}_{t - t_{i-1}}$ is the (rescaled) number of customers who arrived to the queue in the time interval $[rt_{i-1}, rt]$ and who did not exit by time rt. In particular,

$$\widehat{Z}^r_{t_i} = \widehat{Z}^{t_i, r}_{t_{i-1}} + \widehat{Z}^{r, i}_{t_i - t_{i-1}}, \quad 1 \le i \le n, \quad \text{almost surely.} \tag{39}$$

PROPOSITION 7. *For any fixed $\varepsilon, \eta > 0$, there exist an integer $n \geq 1$, and $r_1 \geq 1$ so that*

$$\sup_{r \geq r_1} P\left(\sup_{1 \leq i \leq n} \left|\widehat{Z}^r_{t_i} - \widehat{Z}^r_{t_{i-1}}\right| > 2\varepsilon\right) \leq 2\eta \tag{40}$$

and

$$\sup_{r \geq r_1} P\left(\sup_{1 \leq i \leq n} \sup_{s \in [t_{i-1}, t_i]} \left|\widehat{Z}^{t_i, r}_s - \widehat{Z}^{t_i, r}_{t_i}\right| > 2\varepsilon\right) \leq 2\eta, \tag{41}$$

$$\sup_{r \geq r_1} P\left(\sup_{1 \leq i \leq n} \sup_{u \in [0, T/n]} \widehat{Z}^{r, i}_u > 5\varepsilon\right) \leq 18\eta. \tag{42}$$

By Corollary 4, Lemma 2 and (35), (36), we can find $r_1 \geq 1$ so that both (40) and (41) hold. The hard estimate (42) will be shown in the next section. Estimates (40), (41) imply

$$\sup_{r \geq r_1} P\left(\sup_{1 \leq i \leq n} \left|\widehat{Z}^r_{t_{i-1}} - \widehat{Z}^{t_i, r}_{t_{i-1}}\right| > 4\varepsilon\right) \leq 4\eta,$$

so the left-most and the right-most side in (38) "typically differ" by at most $4\varepsilon + \sup_{u \in [0, T/n]} \widehat{Z}^{r, i}_u$. Combined with (42), this implies that, for any $0 < h < T/n$, we have

$$\sup_{r \geq r_1} P\left(\sup_{|s-t|<h} \left|\widehat{Z}^r_s - \widehat{Z}^r_t\right| > 20\varepsilon\right) \leq 22\eta.$$

The last estimate gives relative compactness of the sequence Z^r, for example, by [13], Corollary III.7.4, completing the proof of Theorem 5.

3.3. *Tree estimates.* This section proves assertion (42) in Proposition 7. The proof uses estimates for the joint total size and height distribution of a sequence of supercritical (near-critical) Galton–Watson trees.

Let $\mathscr{T}$ be a Galton–Watson random tree with offspring distribution Ξ, where Ξ is concentrated on nonnegative integers. By this we mean a tree-valued random variable constructed from a sequence of i.i.d.-Ξ random variables. The root of the tree is the zero *generation*. In the first step, the root gives birth to ξ_0 children, where $\xi_0 \sim_d \Xi$. If $\xi_0 = 0$, then $\mathscr{T}$ consists of the root only. If $\xi_0 \geq 1$, then the root has children ς^1_j, $1 \leq j \leq \xi_0$, that form the first generation. Each vertex ς^1_j is connected to the root by an edge. The tree is formed recursively. In the nth step, each vertex in the $(n-1)$st generation gives birth according to Ξ, independently of others and the previous generations. Again an edge connects each child to its parent. The children of vertices in the $(n-1)$st generation are the nth generation. Continue until *extinction* (no births from any vertex in the same generation) occurs, or forever, if no extinction occurs. For any vertex $\varsigma \in \mathscr{T}$, let gen (ς) denote its generation number. Let $\mathscr{T}_\varsigma$ denote the tree spanned by ς and all of its descendents (children, children of children, etc.). Then $\mathscr{T}_\varsigma =^d \mathscr{T}$ is an elementary consequence of the construction.

The Galton–Watson tree is called *strictly supercritical* if $E\xi_0 > 1$, *critical* if $E\xi_0 = 1$ and *strictly subcritical* if $E\xi_0 < 1$. Similarly, $\mathscr{T}$ is *super* (*sub*)-*critical* if $E\xi_0 \geq 1$ (≤ 1). Let $|\mathscr{T}|$ denote the *total size* (number of vertices) of $\mathscr{T}$, and let $\mathrm{ht}(\mathscr{T})$ denote the *height* (the maximal generation number) of $\mathscr{T}$, respectively. The case of $P(\xi_0 = 1) = 1$ does not appear in the setting of this paper, so we exclude it from consideration. Then, it is well known (e.g., [3]) that subcritical trees have finite size (therefore height) with probability 1, whereas strictly supercritical trees have infinite size (height) with nonzero probability. One readily checks by induction that the number of vertices in the nth generation of $\mathscr{T}$ has expectation $(E\xi_0)^n$.

Recall the branching interpretation for the queue length from Section 1.2. Each busy cycle of the queue corresponds to an excursion of the load (workload) process, and yields a Galton–Watson tree of customers who entered (and exited) the queue during this busy cycle. There is one-to-one and onto correspondence between the vertices of the tree and the customers of the busy cycle. A new 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 of the tree. Otherwise, ς becomes a child of the customer whose service was interrupted, and $\mathrm{gen}\,(\varsigma) = Z(s-) = Z(s) - 1$. The queue-length process Z is the height process (or depth-first search walk) that visits trees in chronological order (of busy cycles), and within each tree visits vertices in the depth-first search (LIFO) order. At each time u, $Z(u)$ records the generation number (plus 1) of the vertex corresponding to the customer currently in service (if any).

It is not hard to calculate the exact offspring distribution Ξ^r corresponding to the queue of index r. Consider a typical customer. At the moment of arrival, this customer requests $v^r =^d F^r$ amount of service time. Its service may be interrupted several times due to new arrivals, each such arrival producing a single offspring. Now it is easy to see that, after conditioning on v^r, the total number of offspring is Poisson ($\lambda^r v^r$), so that

$$\Xi^r(i) = P(\xi^r = i) = E\left[\frac{\exp(-\lambda^r v^r)(\lambda^r v^r)^i}{i!}\right], \qquad i = 0, 1, 2, \dots . \tag{43}$$

Without loss of generality, we can assume supercriticality, $E\xi^r = \lambda^r E v^r = \lambda^r m^r \geq 1$, for all $r \geq 1$. If $E\xi^r < 1$, define $c^r = 1/E\xi^r = 1/\lambda^r m^r > 1$ and $\overline{v}^r = c^r v^r$ with distribution $\overline{F}^r$. If $E\xi^r \geq 1$, set $\overline{F}^r = F^r$. Then clearly we can couple the queue-length processes $Z_u^{r,i}$ and $\overline{Z}_u^{r,i}$ having service distributions F^r and $\overline{F}^r$, so that, for each r,

$$Z_u^{r,i} \leq \overline{Z}_u^{r,i} \quad \text{for all } u, \quad 1 \leq i \leq n, \quad \text{a.s}$$

The convergence $F^r \to F$ implies $\overline{F}^r \to F$, convergence relations (9), (10), (12) continue to hold and (13) implies

$$\sup_r E\left[(\overline{v}^r)^2 1_{\{\overline{v}^r \geq K\}}\right] \to 0 \quad \text{as } K \to \infty.$$

Henceforth, we assume $E\xi^r \geq 1$ so that $\overline{F}^r = F^r$, $r \geq 1$.

PROOF OF (42). Recall the setting of Proposition 7. For each i, we have $\widehat{Z}_u^{r,i} = (1/\sqrt{r}) Z_{ru}^{r,i}$, $u \in [0, T/n]$, where $Z^{r,i}$ is a copy of the queue length Z^r, and relations (38), (39) are satisfied.

Let $\mathcal{T}_{i,j}^r$, $1 \le j \le M_i^r$, be the Galton–Watson trees of the busy cycles (corresponding to $Z^{r,i}$) started after time 0 and completed before time rT/n. Let $\mathcal{T}_i^r$ be the tree of the busy cycle containing the customer present in service at time rT/n. Denote by $\mathcal{T}_i^{r,\dagger}$ the initial portion of the last tree $\mathcal{T}_i^r$, traversed by the queue length $Z^{r,i}$ up to time rT/n. If the queue is empty at time rT/n, set $\mathcal{T}_i^r = \mathcal{T}_i^{r,\dagger} = \varnothing$ to be the empty tree with $\mathrm{ht}(\varnothing) = 0$. Due to the above reasoning, the maximal queue length $\sup_{u \in [0, rT/n]} Z_{ru}^{r,i}$ is dominated by $\max_{1 \le j \le M_i^r} \mathrm{ht}(\mathcal{T}_{i,j}^r) \vee \mathrm{ht}(\mathcal{T}_i^{r,\dagger}) + 1$, the maximal height of all trees (of busy cycles) started in $[0, rT/n]$. So assertion (42) will follow from

$$\sup_{r \ge r_1} P\biggl(\max_{1 \le i \le n}\Bigl(\max_{1 \le j \le M_i^r} \mathrm{ht}(\mathcal{T}_{i,j}^r)\Bigr) \vee \mathrm{ht}(\mathcal{T}_i^{r,\dagger}) > 5\varepsilon\sqrt{r}\biggr) \le 18\eta. \tag{44}$$

It is well known (cf. Athreya and Ney [3], Theorem I.9.1) that

$$P(\mathrm{ht}(\mathcal{T}) > r) \sim \frac{2}{\sigma^2 r} \qquad \text{as } r \to \infty, \tag{45}$$

where $\mathcal{T}$ is a critical Galton–Watson tree with offspring distribution ξ, $E\xi = 1$ and $0 < \mathrm{var}(\xi) = \sigma^2 < \infty$. This is also the content of Kolchin [28], Theorem 2.1.2. Aldous ([2], Proposition 24) gives the following estimate for the joint height and total size distribution of the same tree:

$$r^{1/2} P(\mathrm{ht}(\mathcal{T}) > \varepsilon r^{1/2}, |\mathcal{T}| < \delta r) \to \sigma^{-1}\delta^{-1/2} G(\varepsilon\delta^{-1/2}\sigma) \text{ as } r \to \infty, \tag{46}$$

where $G(x) \le \kappa_1 \exp(-x/\kappa_2)$, $0 < x < \infty$, for some $0 < \kappa_1, \kappa_2 < \infty$. Since we allow the offspring distribution Ξ^r to vary with r, we will need the following analogous lemma.

LEMMA 8. *Let $\mathcal{T}^r$ be a sequence of supercritical Galton–Watson trees with offspring distribution ξ^r such that:*

(i) $\xi^r \to_d \xi \sim_d \Xi$, *where* $E\xi = 1$,
(ii) $r^{1/2}(1 - E\xi^r) \to c \le 0$,
(iii) $0 < \mathrm{var}(\xi^r) = (\sigma^r)^2 \to \sigma^2 \in (0, \infty)$,
(iv) $\sup_r E((\xi^r)^2 1_{\{\xi^r > K\}}) \to 0$, *as* $K \to \infty$.

Then

$$\limsup_r r^{1/2} P(\mathrm{ht}(\mathcal{T}^r) > \varepsilon r^{1/2}, |\mathcal{T}^r| < \delta r) \le \sigma^{-1}\delta^{-1/2} G(\varepsilon\delta^{-1/2}\sigma), \tag{47}$$

where $G(x) \le \kappa_1 \exp(-x/\kappa_2)$, $0 < x < \infty$, *for some* $0 < \kappa_1, \kappa_2 < \infty$.

Note that $F^r \to F$, together with (43), (13) and (9), (10), (12), imply the conditions (i)–(iv) of the lemma. The uniform integrability condition (iv) is natural (cf. [30, 17]).

Assume Lemma 8 for now. Let $\varepsilon, \eta > 0$ and T be as in Proposition 7, let K_1 be a large number such that $e^{-4c\varepsilon}(1+\eta)/K_1 < \eta$ and recall that λ is the asymptotic arrival rate. Choose n_1 large enough so that

$$K_1 \frac{\kappa_1(\lambda+\varepsilon)}{\sigma\sqrt{(\lambda+\varepsilon)T}} n^{3/2} \exp\left(-\frac{\varepsilon\sigma n^{1/2}}{\kappa_2\sqrt{(\lambda+\varepsilon)T}}\right) \le \eta \qquad \text{for all } n \ge n_1, \tag{48}$$

for κ_1, κ_2 in the lemma, and also large enough so that (35), (36) are satisfied for all $n \ge n_1$. Fix some $n \ge n_1$. Assume $r_1 \ge 1$ to be large enough so that both (40) and (41) hold. We use estimate (47) to bound the probabilities of events

$$\left\{\max_{1\le i\le n} \max_{1\le j\le M_i^r} \mathrm{ht}(\mathscr{T}_{i,j}^r) > \varepsilon\sqrt{r}\right\} \qquad \text{and} \qquad \left\{\max_{1\le i\le n} \mathrm{ht}(\mathscr{T}_i^{r,\dagger}) > \varepsilon\sqrt{r}\right\}.$$

Recall M_i^r is the number of trees corresponding to the completed busy cycles of $Z^{r,i}$, and let $N_i^r = |\mathscr{T}_{i,1}^r| + |\mathscr{T}_{i,2}^r| + \cdots + |\mathscr{T}_{i,M_i^r}^r| + |\mathscr{T}_i^{r,\dagger}|$ be the total number of vertices visited before time rT/n.

LEMMA 9. *For any fixed $n \ge n_1$, there exists $r_3 \ge 1$ such that*

$$\sup_{r\ge r_3} P\left(\max_{1\le i\le n} N_i^r \ge (\lambda+\varepsilon)\frac{T}{n}r\right) \le \eta, \tag{49}$$

$$\sup_{r\ge r_3} P\left(\max_{1\le i\le n} M_i^r \ge \sqrt{r}(\lambda+\varepsilon)\right) \le 3\eta. \tag{50}$$

PROOF. The first assertion is easy since $N_i^r \stackrel{d}{=}$ Poisson (rate $r\lambda^r T/n$), $1 \le i \le n$. For the second one, consider processes $X^{r,i} = (X_{t_{i-1}+s}^r - X_{t_{i-1}}^r$, $s \in [0, rT/n])$, and let $I_s^{r,i} = \inf_{0\le u\le s} X_u^{r,i}$, and $\tau_x^{r,i} := \inf\{s \ge 0\colon I_s^{r,i} \le -x\}$. Note that $I_s^{r,i} = -|\{u \in [0,s]\colon Z_u^{r,i} = 0\}|$. Recall how (37) implied (41). Since the asymptotic load X is a Brownian motion, the same assertion (37) implies that, for any $n \ge n_1$, we can find r_3 large enough so that

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

On the complement of $\{\min_{1\le i\le n} \tau_{\sqrt{r}}^{r,i} < rT/n\}$, we have $M_i^r \le M_{\sqrt{r}}^{r,i}$, $1 \le i \le n$, where $M_{\sqrt{r}}^{r,i}$ equals the number of busy cycles started, and completed, during $[0, \tau_{\sqrt{r}}^{r,i}]$. By (5), $M_{\sqrt{r}}^{r,i} \stackrel{d}{=} M_{\sqrt{r}} \stackrel{d}{=}$ Poisson (rate $\sqrt{r}\lambda$), and the second assertion of the lemma follows just like the first one. $\square$

Now for any fixed $n \geq n_1$ and any $r \geq \max\{r_1, r_3\}$, we get

$$\begin{aligned}
&P\Bigl(\max_{1\le i\le n}\max_{1\le j\le M_i^r}\mathrm{ht}(\mathscr{T}_{i,j}^r) > \varepsilon\sqrt{r}\Bigr)\\
&\quad\le 4\eta + \sum_{i=1}^n P\Bigl(\max_{1\le j\le M_i^r\le\sqrt{r}(\lambda+\varepsilon)}\mathrm{ht}(\mathscr{T}_{i,j}^r) > \varepsilon\sqrt{r}, \max_j|\mathscr{T}_{i,j}^r|\\
&\qquad\qquad < (\lambda+\varepsilon)\frac{T}{n}r, \text{ and } M_i^r \le \sqrt{r}(\lambda+\varepsilon)\Bigr)\\
&\quad\le 4\eta + n\sqrt{r}(\lambda+\varepsilon)P\Bigl(\mathrm{ht}(\mathscr{T}^r) > \varepsilon\sqrt{r}, |\mathscr{T}^r| < (\lambda+\varepsilon)\frac{T}{n}r\Bigr),
\end{aligned}$$

and by (47) there exists r_4 (possibly larger than r_3) so that, for each $r \geq r_4$,

$$\begin{aligned}
&P\Bigl(\max_{1\le i\le n}\max_{1\le j\le M_i^r}\mathrm{ht}(\mathscr{T}_{i,j}^r) > \varepsilon\sqrt{r}\Bigr)\\
&\quad\le 5\eta + \frac{\kappa_1(\lambda+\varepsilon)}{\sigma\sqrt{(\lambda+\varepsilon)T}}n^{3/2}\exp\Bigl(-\frac{\varepsilon\sigma n^{1/2}}{\kappa_2\sqrt{(\lambda+\varepsilon)T}}\Bigr)\\
&\quad\le 6\eta \qquad \text{(by (48)).}
\end{aligned}$$

LEMMA 10. *For $n \geq n_1$ fixed as above, there exists some $r_5 \geq r_4$ such that*

$$\sup_{r\ge r_5} P\Bigl(\max_{1\le i\le n}\mathrm{ht}(\mathscr{T}_i^{r,\dagger}) > 5\varepsilon\sqrt{r}\Bigr) \le 12\eta.$$

PROOF. The following is an extension of the idea in the argument for (50). Let $X_\cdot^{r,i}$, $I_\cdot^{r,i}$ and $\tau_\cdot^{r,i}$ be as in Lemma 9. Let $(Y_s^{r,i}, s \geq 0)$, $1 \leq i \leq n$, be mutually independent, distributed as the load process $(X_s^r, s \geq 0)$ and independent of X^r. One can construct a new copy $X^{\star,r} =^d X^r$ from $X_\cdot^{r,i}$ and $Y_\cdot^{r,i}$ as described below. The point of the construction is that (typically) each tree $\mathscr{T}_{i,j}^r$, $1 \leq j \leq M_i^r$, $1 \leq i \leq n$, corresponding to a busy cycle of $Z^{r,i}$ (that is, $X^{r,i}$) reappears as a tree corresponding to a busy cycle of $X^{\star,r}$. More importantly, each tree $\mathscr{T}_i^{r,\dagger}$, $1 \leq i \leq n$, reappears as the initial portion of a tree of some busy cycle of $X^{\star,r}$. The idea is simple, but the notation could get messy, so sometimes we omit "r" in the superscript. Define stopping times $\tau_i = \tau_{\sqrt{r}}^{r,i} \wedge (rT/n)$, $1 \leq i \leq n$. Then $\overline{X}^{\star,r,i}$, defined by

$$\overline{X}_s^{\star,r,i} = X_s^{r,i}1_{\{s\le\tau_i\}} + (X_{\tau_i}^{r,i} + Y_{s-\tau_i}^{r,i})1_{\{s>\tau_i\}},$$

equals X^r in distribution, due to independence of $X^{r,i}$ and $Y^{r,i}$, and the strong Markov property of X^r. Note that, moreover, $\overline{X}^{\star,r,i}$, $1 \leq i \leq n$, are mutually independent as processes.

Let $\bar{\tau}_{\sqrt{r}}^i = \inf\{s \geq 0\colon \inf_{u\le s}\overline{X}_u^{\star,r,i} < -\sqrt{r}\}$. Then the processes

$$(\overline{X}_s^{\star,r,i}, s \in [0, \bar{\tau}_{\sqrt{r}}^i]), \qquad 1 \leq i \leq n, \tag{51}$$

are independent and identically distributed, where $(\bar{X}^{\star,r,1}_s, s \in [0, \bar{\tau}^1_{\sqrt{r}}]) =^d (X^r_s, s \in [0, \tau^r_{\sqrt{r}}])$, and $\tau^r_{\sqrt{r}} = \inf\{u \le s\colon X^r_u < -\sqrt{r}\}$. Now define $\tau^\star_0 = 0$, $\tau^\star_{i\sqrt{r}} := \sum_{j\le i} \bar{\tau}^j_{\sqrt{r}}$, and let

$$X^{\star,r}_s = \sum_{i=1}^{n-1} \bar{X}^{\star,r,i}_{(s-\tau^\star_{(i-1)\sqrt{r}})\wedge\bar{\tau}^i_{\sqrt{r}}} 1_{\{\tau^\star_{(i-1)\sqrt{r}}\le s\}} + \bar{X}^{\star,r,n}_{s-\tau^\star_{(n-1)\sqrt{r}}} 1_{\{s\ge\tau^\star_{(n-1)\sqrt{r}}\}}.$$

So the path of $X^{\star,r}$ is the concatenation of paths (51) for $1 \le i \le n-1$ and the whole path $(\bar{X}^{\star,r,n}_s, s \ge 0)$.

Again by Markov property, $X^{\star,r}$ equals X^r in distribution. Note that $\tau^\star_{i\sqrt{r}} = \inf\{s \ge 0\colon \inf_{u\le s} X^{\star,r}_u < -i\sqrt{r}\}$, which agrees with the usual notation. Moreover, on the event $\{\min_{1\le i\le n} \tau^{r,i}_{\sqrt{r}} \ge rT/n\}$, we have $\tau_i = rT/n$ and $\bar{\tau}^i_{\sqrt{r}} \ge rT/n$. So on the same event, for each i, the path $(X^{r,i}_s, s \in [0, rT/n])$ is the initial part of the path $(\bar{X}^{\star,r,i}_s, s \in [0, \bar{\tau}^i_{\sqrt{r}}])$, and therefore, $(X^{r,i}_s, s \in [0, rT/n]) = (X^{\star,r}_{\tau^\star_{(i-1)\sqrt{r}}+s} - X^{\star,r}_{\tau^\star_{(i-1)\sqrt{r}}}, s \in [0, rT/n])$ almost surely. Hence, on the event $\{\min_{1\le i\le n} \tau^{r,i}_{\sqrt{r}} \ge rT/n\}$, the trees $\mathcal{T}^r_{i,j}, 1 \le j \le M^r_i$ (resp. $\mathcal{T}^{r,\dagger}_i$), $1 \le i \le n$, all reappear as trees (resp. initial parts of trees) corresponding to busy cycles of $(X^{\star,r}_s, s \in [0, \tau^\star_{n\sqrt{r}}])$.

Identity (39) together with bound (41) implies

$$\sup_{r\ge r_1} P\Bigl(\max_{1\le i\le n} \widehat{Z}^{r,i}_{t_i-t_{i-1}} > 4\varepsilon\Bigr) \le 4\eta.$$

On the event $\{\max_{1\le i\le n} \widehat{Z}^{r,i}_{t_i-t_{i-1}} \le 4\varepsilon\}$, the generation number of the last vertex visited by the queue-length process $Z^{r,i}$ on the interval $[0, rT/n]$ is smaller than or equal to $4\varepsilon\sqrt{r}$, for all $i \le n$ simultaneously.

Now consider the intersection A^r of "good" events

$$\begin{aligned} A^r = {} & \Bigl\{\min_{1\le i\le n} \tau^{r,i}_{\sqrt{r}} > rT/n\Bigr\} \cap \Bigl\{\max_{1\le i\le n} \widehat{Z}^{r,i}_{t_i-t_{i-1}} \le 4\varepsilon\Bigr\} \\ & \cap \Bigl\{\max_{1\le i\le n} N^r_i \le (\lambda+\varepsilon) rT/n\Bigr\} \\ & \cap \Bigl\{\max_{1\le i\le n} M^r_i \le (\lambda+\varepsilon)\sqrt{r}\Bigr\}. \end{aligned}$$

By previous considerations and Lemma 9, the probability of the complement of A^r is bounded from above by 8η for all r larger than $\max\{r_1, r_3\}$. The condition $\{\max_{1\le i\le n} M^r_i \le (\lambda+\varepsilon)\sqrt{r}\}$ will be used in later calculations, cf. (52).

Let $\mathcal{T}^{\star,r}_1, \dots, \mathcal{T}^{\star,r}_{M^{\star,r}}$ be the sequence of Galton–Watson trees generated by $(X^{\star,r}_s, s \in [0, \tau^\star_{n\sqrt{r}}])$. Recall that $\mathcal{T}^{\star,r}_\varsigma$ is the subtree spanned by vertex ς and all of its descendants. Due to the above construction of $X^{\star,r}$,

$$P\Bigl(\Bigl\{\max_{1\le i\le n} \mathrm{ht}(\mathcal{T}^{r,\dagger}_i) > 5\varepsilon\sqrt{r}\Bigr\} \cap A^r\Bigr) \le P\bigl(A^{\star,r}_0\bigr),$$

where $A_0^{\star,r} = \{\mathrm{ht}(\mathscr{T}_\varsigma^{\star,r}) > \varepsilon\sqrt{r} - 1$ and $|\mathscr{T}_\varsigma^{\star,r}| < r(\lambda+\varepsilon)T/n$ for some vertex $\varsigma^{\star,r} \in \mathscr{T}_1^{\star,r} \cup \cdots \cup \mathscr{T}_{M^{\star,r}}^{\star,r}$, gen $(\varsigma^{\star,r}) = \lfloor 4\varepsilon\sqrt{r}\rfloor + 1\}$. The last statement is true by the "triangle inequality." On the event $\{\max_{1\le i\le n} \mathrm{ht}(\mathscr{T}_i^{r,\dagger}) > 5\varepsilon\sqrt{r}\} \cap A^r$, the last vertex visited in each of the trees $\mathscr{T}_i^{r,\dagger}$ belongs to one of the first $\lfloor 4\varepsilon\sqrt{r}\rfloor$ generations, and the total size of $\mathscr{T}_i^{r,\dagger}$ is smaller than or equal to $(\lambda+\varepsilon)rT/n$, $1 \le i \le n$. However, there is at least one vertex in $\mathscr{T}_j^{r,\dagger}$, for some $1 \le j \le n$, with generation number larger than $\lceil 5\varepsilon\sqrt{r}\rceil$. The ancestor ς^r of this vertex in generation $\lfloor 4\varepsilon\sqrt{r}\rfloor + 1$ must belong to the same tree $\mathscr{T}_j^{r,\dagger}$, due to the depth-first search order. Similarly, due to the depth-first search order, the whole tree $\mathscr{T}_{\varsigma^r}^{\star,r}$ is contained in $\mathscr{T}_j^{r,\dagger}$ for this j. In the construction of $X^{\star,r}$, vertex ς^r and its tree of descendants $\mathscr{T}_{\varsigma^r}^{\star,r}$ become $\varsigma^{\star,r}$ and $\mathscr{T}_{\varsigma^{\star,r}}^{\star,r}$, where gen $(\varsigma^{\star,r}) = \lfloor 4\varepsilon\sqrt{r}\rfloor + 1$, so that $A_0^{\star,r}$ occurs.

Since $M^{\star,r} =^d$ Poisson (rate $\lambda^r n\sqrt{r}$) by (5), we have $\lim_r P(M^{\star,r} \ge n(\lambda+\varepsilon)\sqrt{r}) = 0$, so one may assume r to be large enough so that $P(M^{\star,r} \ge n(\lambda+\varepsilon)\sqrt{r}) \le \eta$. The trees $\mathscr{T}_1^{\star,r}, \dots, \mathscr{T}_{M^{\star,r}}^{\star,r}$ are, conditionally on $M^{\star,r}$, independent and identically distributed as $\mathscr{T}^r$. So, given $M^{\star,r} < n(\lambda+\varepsilon)\sqrt{r}$, the total expected number of vertices in generation $\lfloor 4\varepsilon\sqrt{r}\rfloor+1$ is bounded from above by

$$n(\lambda+\varepsilon)\sqrt{r}(E\xi^r)^{4\varepsilon\sqrt{r}+1} \le 4(1+\eta)n(\lambda+\varepsilon)\sqrt{r}e^{-4c\varepsilon}, \quad r \to \infty,$$

due to assumption (10). Recall the large number K_1 from (48). If we denoted by $M^{\star,r}(l) = \#\{\varsigma \in \mathscr{T}_1^{\star,r} \cup \cdots \cup \mathscr{T}_{M^{\star,r}}^{\star,r}\colon \text{gen } (\varsigma) = l\}$ the total size of generation l, Markov inequality implies

$$\begin{aligned} &P\Big(M^{\star,r}(\lfloor 4\varepsilon\sqrt{r}\rfloor + 1) > K_1 n(\lambda+\varepsilon)\sqrt{r} | M^{\star,r} \le n(\lambda+\varepsilon)\sqrt{r}\Big) \\ &\quad \le \frac{(1+\eta)e^{-4\varepsilon c}}{K_1} \le \eta, \end{aligned} \tag{52}$$

for all large r. Each vertex ς in generation $\lfloor 4\varepsilon\sqrt{r}\rfloor + 1$ has equal probability $P(\mathrm{ht}(\mathscr{T}_\varsigma^{\star,r}) > \varepsilon\sqrt{r}, |\mathscr{T}_\varsigma^{\star,r}| < r(\lambda+\varepsilon)T/n) = P(\mathrm{ht}(\mathscr{T}^r) > \varepsilon\sqrt{r}, |\mathscr{T}^r| < r(\lambda+\varepsilon)T/n)$ of contributing to event $A_0^{\star,r}$. The above estimates put together with (47), (48) imply the existence of some large $r_5 \ge r_4$ such that

$$\begin{aligned} \sup_{r\ge r_5} P(A_0^{\star,r}) &\le 2\eta + \sup_{r \ge r_5} K_1 n(\lambda+\varepsilon)\sqrt{r} P(\mathrm{ht}(\mathscr{T}^r) > \varepsilon\sqrt{r}, \\ &\qquad |\mathscr{T}^r| < r(\lambda+\varepsilon)T/n) \\ &\qquad\qquad \le 3\eta + K_1 \frac{\kappa_1(\lambda+\varepsilon)}{\sigma\sqrt{(\lambda+\varepsilon)T}} n^{3/2} \exp\left(-\frac{\varepsilon\sigma n^{1/2}}{\kappa_2\sqrt{(\lambda+\varepsilon)T}}\right) \le 4\eta. \end{aligned}$$

Therefore,

$$\sup_{r\ge r_5} P\Big(\max_{1\le i\le n} \mathrm{ht}(\mathscr{T}_i^{r,\dagger}) > 5\varepsilon\sqrt{r}\Big) \le \sup_{r\ge r_5} P(\text{ not } A^r) + P(A_0^{\star,r}) \le 12\eta.$$

Now take r_5 to be r_1 in the statement of Proposition 7. $\square$

It remains to prove Lemma 8. The proof of (47) consists of adapting the corresponding arguments in [28], and then applying the reasoning of [2]. We sketch the proof, recalling the arguments of Kolchin along the way. Let ξ_i, $i \geq 1$, be i.i.d. integer-valued random variables with span 1 (that is, $P(\xi_1 = 1) > 0$), and with mean $E\xi_1 = a$ and $\operatorname{var}(\xi_1) = \sigma^2$. Then the (standard) local central limit theorem (e.g., [11], Theorem 2.5.2 or [28], Theorem 1.4.2) gives

$$\sigma\sqrt{N}P(\xi_1 + \cdots + \xi_N = m) - \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(m - aN)^2}{2\sigma^2 N}\right\} \to 0, \tag{53}$$

uniformly in m, as $N \to \infty$. The identity in [28], Lemma 2.1.3, implies

$$P(|\mathscr{T}| = N) = \frac{1}{N}P(\xi_1 + \cdots + \xi_N = N - 1), \qquad N \geq 1, \tag{54}$$

and evaluating (53), (54) with $a = 1$ gives ([28], Lemma 2.1.4)

$$P(|\mathscr{T}| = N) \sim \frac{1}{\sqrt{2\pi}\sigma}N^{-3/2}. \tag{55}$$

It is important to note that the criticality assumption ($E\xi = 1$) gets used here only when applying (53), while the identity (54) holds for noncritical ξ's as well.

Theorem 2.4.3 of [28] shows the convergence of heights conditioned on total size, which Aldous [2] recognizes in terms of the maximum W^* of the standard (unit length) Brownian excursion as

$$\begin{gathered} P\left(\operatorname{ht}(\mathscr{T}) > \frac{x}{\sigma}N^{1/2} \,\Big|\, |\mathscr{T}| = N\right) = P(2W^* > x)(1 + o_N(1)), \\ \lim_N o_N(1) = 0. \end{gathered} \tag{56}$$

It suffices to show uniform (in r) analogues of (53), (56):

LEMMA 11. (i) *As $N \to \infty$, uniformly in m,*

$$\begin{aligned} \sup_{r\geq 1}\bigg(\sigma^r\sqrt{N}P(\xi_1^r + \cdots + \xi_N^r = m) \\ - \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(m - E\xi^r N)^2}{2N(\sigma^r)^2}\right\}\bigg) \to 0, \end{aligned} \tag{57}$$

(ii)

$$P\left(\operatorname{ht}(\mathscr{T}^r) > \frac{x}{\sigma^r}N^{1/2} \,\Big|\, |\mathscr{T}^r| = N\right) = P(2W^* > x)(1 + o(r, N)), \tag{58}$$

where $\lim_{N\to\infty}\sup_{r\geq 1}|o(r, N)| = 0$.

Due to (57) and (54), we have

$$
\begin{aligned}
P(|\mathscr{T}^r| = \lfloor ur \rfloor) &= \frac{1}{\sqrt{2\pi}\sigma}(ur)^{-3/2}\exp(-c^2u/(2\sigma^2))(1+o_r(u)) \\
&\le \frac{1}{\sqrt{2\pi}\sigma}(ur)^{-3/2}(1+o_r(u)),
\end{aligned} \tag{59}
$$

where $\sup_{u\in[\varepsilon/\sqrt{r},\delta]} |o_r(u)| \to 0$ as $r \to \infty$. The rest of the argument for (47) is identical to the one in [2], Proposition 24 for (46), using (59) and (58) in place of (55) and (56).

SKETCH OF THE PROOF OF LEMMA 11. Assertion (57) is proved in the same way as (53). Let $\phi^r(t) = E\exp(it\xi^r)$ be the characteristic function of ξ^r. One uses the inversion formula, splits the real-line into the same four regions and estimates the integrands, this time, uniformly over r. It is important that, by assumption (i) of Lemma 8, $\phi^r \to \phi$ uniformly (on R since all ϕ^r are 2π periodic), where ϕ is the characteristic function of ξ. We omit the details; the reader can easily check that the argument carries over, step by step.

To prove (58), it suffices to consider only critical trees. Suppose the moment generating function $G(z) = \sum_{i=0}^{\infty} P(\xi = i)z^i$, $z \in [0, 1]$, of the offspring distribution of $\mathscr{T}$ satisfies

$$
G(a) = aG'(a) < \infty, \tag{60}
$$

for some real $a > 0$. It is easy to check (originally due to Kolchin) that the critical Galton–Watson tree $\mathscr{T}_a$ with offspring distribution $P(\xi_a = i) = (a^i/G(a))P(\xi = i)$ and the original tree $\mathscr{T}$ have the same distribution, when conditioned on their total size. In particular,

$$
P(\mathrm{ht}(\mathscr{T}) > xN^{1/2} \| \mathscr{T}| = N) = P(\mathrm{ht}(\mathscr{T}_a) > xN^{1/2} \| \mathscr{T}_a| = N).
$$

If the tree $\mathscr{T}$ is critical, then $a = 1$. If the tree $\mathscr{T}$ is supercritical, that is, $G'(1) > 1$, then the smallest fixed point $z_0 \in [0, 1]$ of G is strictly smaller than 1. By convexity of G, it must be $G'(z_0) < 1$, so

$$
G'(1) > G(1) \quad \text{and} \quad z_0G'(z_0) < G(z_0) = z_0.
$$

By continuity and convexity, there exists a unique $a \in (z_0, 1)$ so that (60) holds. Using assumptions (ii) and (iv) of the lemma, it is easy to check that the sequence of smallest fixed points z_0^r of G^r satisfies $\lim_r z_0^r \to 1$; therefore, $\lim_r a^r = 1$. The sequence $\xi_{a^r}^r$ of integer random variables inherits all the properties in assumptions of the lemma. Thus, without loss of generality, we assume that all trees are critical, $E\xi^r = 1$, $r \ge 1$.

Due to our assumptions (i)–(iv), again the relevant estimates in [28] can be made uniform in r. For example, [28], Theorem 2.1.2 implies the analogue of (45),

$$
P\left(\mathrm{ht}(\mathscr{T}^r) > \frac{N}{\sigma^r}\right) = \frac{1}{\sqrt{2\pi}N}(1+o_{r,N}(1)), \tag{61}
$$

where $\lim_{N\to\infty}\sup_r |o_{r,N}(1)| = 0$, which is the first step in showing (58). The arguments in [28], Lemmas 2.4.3–2.4.5, Corollaries 2.4.1 and 2.4.2 and Theorems 2.4.1–2.4.3 extend in a similar way; we omit the details.

3.4. *Discussion.*

3.4.1. *Initial condition.* We first comment on the heavy-traffic limits under more general initial conditions. It is clear that convergence of initial load on diffusion scale $\widehat{X}^r(0) = X^r(0)/\sqrt{r} \to_d X_0$, where X_0 is a.s. a finite nonnegative random variable, would imply the convergence in distribution of the load processes $\widehat{X}^r$ to a shifted Brownian motion started at X_0. As before, this implies the convergence of the workload processes

$$\widehat{W}^r_t \Rightarrow W, \tag{62}$$

where $W_t = (X_t - \inf_{s\le t}(X_s\wedge 0))$, $t\ge 0$. It is intuitively clear that, provided we have convergence of rescaled initial pure-atomic measures $q^r(0)$ (with atoms $\{1,2,\dots,Z^r(0)\}$ and intensities equal to residual service times) to a fixed finite measure q_0 with finite support $[0, Z(0)] = [0, \lim_r \widehat{Z}^r(0)]$ and such that $\langle 1, q_0\rangle = X_0$, then Theorems 1 and 5 should extend accordingly. We need to introduce some notation in order to identify these limits. As in [15], for any scalar $a\ge 0$ and any measure μ such that $\mathrm{Supp}(\mu)\subset[0,\infty]$, let the *truncation* of μ at level a be the measure $\mu_{|a}$ defined by $\mu_{|a}[0,x] = \mu[0,x]\wedge a$. So, if $a\le 0$, then $\mu_{|a}$ is the zero measure. Also if μ,ν are measures such that $\mathrm{Supp}(\mu)\cup\mathrm{Supp}(\nu)\subset[0,\infty]$, and $\sup(\mathrm{Supp}(\mu)) = b < \infty$, define μ *concatenated* with ν as $(\mu\oplus\nu)([0,x]) = \mu([0,x\wedge b]) + \nu([0,(x-b)^+])$. Then it is easy to see that, at each level r, the RES-measure process $q^r(t) = q^r(0)_{|X^r(0)+I^{*,r}_t}\oplus q^{*,r}(t)$ encodes all the information, where $q^{*,r}(t)$ is a copy of the RES-measure process from Section 1.2 (started at zero measure) and $I^{*,r}_t = \inf_{u\in[0,t]} X^{r,*}_u$ is the corresponding infimum process. As $r\to\infty$, the rescaled $\hat{q}^r(t)$ should converge in the Skorokhod topology to $q_t = q_{0|X_0+I^*_t}\oplus q^*_t$, where q^* is a copy of the generalized RES-measure process from Theorem 1.

The heavy-traffic limit for the queue length, on the other hand, depends on the finer properties of the asymptotic initial measure q_0. By Theorems 1 and 5 and convergence (25), under certain regularity assumptions, one should get $\widehat{Z}^r(t) = \sup(\mathrm{Supp}(\hat{q}^r_t)) \Rightarrow \sup(\mathrm{Supp}(q_{0|X_0+I^*_t})) + \sup(\mathrm{Supp}(q^*_t))$. If $q_0(dx) = \frac{\lambda\beta}{2}dx$, $x\le Z(0)$, this means

$$\widehat{Z}^r_t \Rightarrow Z_t = \frac{2}{\lambda\beta}(X(0)+I^*_t)^+ + \frac{2}{\lambda\beta}W^*_t = \frac{2}{\lambda\beta}W_t,$$

where W is the limit in (62).

3.4.2. *LIFO vs. FIFO.* Recall the optimization question from the Introduction. Assume a sequence of queues approaches heavy traffic (9), (10), (12), (13) and fix some large r. The two queues have the same workload process (11) W^r, which is approximated by W, a reflected Brownian motion (variance $\lambda\beta$ and drift $-c$). Denote by Z^r_{FI} and Z^r_{LI} the queue lengths under the FIFO and the LIFO disciplines. For the FIFO queue, we have

$$W^r(t) = \sum_{i=1}^{Z^r_{\mathrm{FI}}(t)} v^r_i + \varepsilon^r(t), \tag{63}$$

where $(v^r_i, i \geq 1)$ are i.i.d. with distribution F^r and $\varepsilon^r(t)$ is the residual service time of the customer currently in service. Due to the law of large numbers, $\widehat{Z}^r_{\mathrm{FI}} \Rightarrow Z_{\mathrm{FI}} = \lambda W$ in the limit; therefore,

$$Z^r_{\mathrm{FI}}(\cdot) \approx \lambda W^r(\cdot),$$

while for the LIFO queue, (21) and Theorem 5 give $Z^r_{\mathrm{LI}}(\cdot) \approx \frac{2}{\lambda\beta} W^r(\cdot)$, so in order to minimize the queue length in heavy traffic, the server should use the LIFO discipline iff $\lambda^2\beta > 2$ (equivalently, $\beta > 2m^2$ or $\sigma^2 > 2m$) and the FIFO discipline (alternatively, LIFO non-preemptive) otherwise. Note that in the special case, where both the arrival and the service times are exponential (rate λ^r), we can make the two queue lengths Z^r_{FI} and Z^r_{LI} coincide (as processes); therefore, their limits coincide, confirming $\lambda = \frac{2}{\lambda\beta}$. If all customers have constant service time m^r, $P(v = m^r) = 1$, then $\beta = m^2 < 2m^2$ and, of course, the FIFO discipline is optimal.

3.4.3. *Random tree analogy.* The argument in Lemma 2 uses an analogue of (63) for the LIFO case. It is not surprising that the mean residual service time depends on both the first and the second moment of the service time distribution F. Moreover, its exact value is in agreement with the analogous result in Aldous [2] about a "diffusion approximation" to a large Galton–Watson tree. To simplify the comparison, we assume that, for all large r, $\lambda^r = \lambda = 1$ and the service times have distribution $F^r = F$ with mean $m = 1$ and variance $\beta - m^2$. Then the busy cycles of the queue correspond to trees with critical offspring distribution $\xi^r = \xi$, $E\xi = 1$, $\operatorname{var}(\xi) = \beta$. Denote by X^r the discrete-time depth-first search walk (from [2]) of a Galton–Watson tree $\mathscr{T}$ with offspring distribution ξ, conditioned on $|\mathscr{T}| = r$. Theorem 23 in [2] states

$$(r^{-1/2} X^r([2rt]), t \in [0,1]) \Rightarrow \left(\frac{2}{\sqrt{\beta}} W^*_t, t \in [0,1]\right),$$

where W^* is standard Brownian excursion. The LIFO queue-length process Z^r is the depth-first search walk (continuous-time analogue) of an infinite sequence of critical Galton–Watson trees generated from queueing. Theorem 5 and (21) state

$$(r^{-1/2} Z^r(rt), t \geq 0) \Rightarrow \left(\frac{2}{\sqrt{\beta}} W^*(t), t \geq 0\right),$$

where $W^*(t)$ is standard (mean 0, variance 1) Brownian motion.

3.4.4. *The heavy tails.* Recall the heavy tails setting from the Remark in Section 2. For X a stable-α process, $\alpha \in (1,2)$, we can choose ([15], Proposition 4.3) the analogue of (21) as

$$Z_s^t = \lim_{\varepsilon \to 0} \varepsilon^{\alpha-1} \#\{u \in (0,s], X_{u-} < I_u^t, \Delta X_u \geq \varepsilon\}, \qquad s \in [0,t] \quad \text{and} \quad Z_t = Z_t^t, \tag{64}$$

where $\Delta X_u = X_u - X_{u-}$. Again Z_s^t and Z_t are continuous processes (cf. [15], Theorem 4.7). At each level r, define $\widehat{Z}_s^{t,r} = r^{1/\alpha - 1} Z^{rt,r}(rs)$ and $\widehat{Z}_t^r = \widehat{Z}_t^{t,r}$. Then (26) is satisfied (cf. [15], Proposition 5.2) with Z_s^t in (64), so Theorem 1 extends in this case.

The finite-dimensional convergence of queue lengths (or heights) is a consequence of a more general result ([15], Proposition 5.2). The "tightness from below" for the queue length is again a consequence of (38) and (36). For the "tightness from above," an analogue of Proposition 7 might be obtained using tree estimates analogous to those in Lemma 8. As remarked in Section 2, one can construct a triangular array of loads X^r converging (after scaling) to a general Lévy process X with Laplace exponent (16), (17). Duquesne and Le Gall (personal communication) consider this setting, where $\sigma = 0$ in (16), and obtain an analogue of Theorem 5 under suitable assumptions.

4. Directions for further research. Taking the FIFO queueing discipline as a paradigm, we list several natural ways to generalize the result of this paper. The full name of our queue, *feed-forward, single class, single server $M/G/1$ LIFO preemptive resume queue*, gives a list of assumptions that might be relaxed. Allowing renewal (non-Markovian) arrivals would be valuable extensions for applications. We consider the above setting in the forthcoming paper [29]. It turns out that LIFO preemptive resume service discipline induces an unconventional heavy-traffic behavior, in that the limit for the queue length depends on the type of arrivals (and services) in an intricate way.

Introducing feedback, or more customer classes, to the system (where the classes differ by their interarrival and/or service time distributions) or considering networks of LIFO queues, complicates the global arrival process and might result in additional "surprises" in heavy traffic.

Acknowledgment. I am grateful to Ruth Williams for suggesting this project, and for her numerous comments that improved the flow of the paper. Many thanks to the anonymous referee for detailed comments and for finding a flaw in the previous proof of Proposition 7.

REFERENCES

[1] ABATE, J. and WHITT, W. (1997). Limits and approximations for the $M/G/1$ LIFO waiting-time distribution. *Oper. Res. Lett.* **20** 199–206.

[2] ALDOUS, D. J. (1993). The continuum random tree III. *Ann. Probab.* **21** 248–289.

[3] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, New York.
[4] BASKETT, F., CHANDY, K. M., MUNTZ, R. R. and PALACIOS, F. G. (1975). Open, closed and mixed networks of queues with different classes of customers. *J. ACM* **22** 248–260.
[5] BERTOIN, J. (1996). *Lévy Processes*. Cambridge Univ. Press.
[6] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
[7] BRAMSON, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems: Theory Appl*. **30** 89–148.
[8] BREIMAN, L. (1968). *Probability*. Addison-Wesley, Reading, MA.
[9] COFFMAN, E. G. and MITRANI, I. (1988). Storage of the single-server queue. In *Queueing theory and Its Applications.* 193–205. North-Holland, Amsterdam.
[10] DAWSON, D. (1993). *Measure-Valued Markov Processes. École d'Été de Probabilités de Saint-Flour XXI*. *Lecture Notes in Math.* **1541**. Springer, Berlin.
[11] DURRETT, R. (1991). *Probability. Theory and Examples*. Wadsworth and Brooks/Cole, Pacific Grove, CA.
[12] DYNKIN, E. B. (1994). *An Introduction to Branching Measure-Valued Processes.* Amer. Math. Soc., Providence, RI.
[13] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York.
[14] LE GALL, J. F. and LE JAN, Y. (1998). Branching processes in Lévy processes: Laplace functionals of snakes and superprocesses. *Ann. Probab*. **26** 1407–1432.
[15] LE GALL, J. F. and LE JAN, Y. (1998). Branching processes in Lévy processes: the exploration process. *Ann. Probab*. **26** 213–252.
[16] GEIGER, J. and KERSTING, G. (1997). Depth-first search of random trees, and Poisson point processes. *IMA Vol. Math. Appl*. **84** 111–126.
[17] GRIMVALL, A. (1974). On the convergence of sequences of branching processes. *Ann. Probab*. **2** 1027–1045.
[18] HARRIS, T. E. (1974). First passage and recurrence distributions. *Trans. Amer. Math. Soc*. **73** 471–486.
[19] HARRISON, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Adv. in Appl. Probab*. **10** 886–905.
[20] HEATH, D., RESNICK, S. and SAMORODNITSKY, G. (1998). Heavy tails and long range dependence in on/off processes and associated fluid models. *Math. Oper. Res*. **23** 145–165.
[21] IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic I. *Adv. in Appl. Probab*. **2** 150–177.
[22] IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic II. *Adv. in Appl. Probab*. **2** 355–364.
[23] JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes*. Springer, Berlin.
[24] JAGERMAN, D. L., MELAMED, B. and WILLINGER, W. (1997). Stochastic modeling of traffic processes. In *Frontiers in Queueing.* 271–320. CRC Press, Boca Raton, FL.
[25] KELLY, F. P. (1975). Networks of queues with customers of different types. *J. Appl. Probab*. **12** 542–554.
[26] KENDALL, D. G. (1951). Some problems in the theory of queues. *J. Roy. Statist. Soc. B* **13** 151–185.
[27] KINGMAN, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc*. **57** 902–904.
[28] KOLCHIN, V. F. (1986). *Random Mappings*. Optimization Software, New York. [Trans. of Russian original.]
[29] LIMIC, V. (2000). On the behavior of LIFO preemptive resume queues in heavy traffic. *Electron. Comm. Probab.* **5** 13–27.
[30] LINDVALL, T. (1974). Limit theorems for some functionals of certain Galton–Watson branching processes. *Adv. in Appl. Probab*. **6** 309–321.
[31] REIMAN, M. I. (1984). Open queueing networks in heavy traffic. *Math. Oper. Res*. **9** 441–458.
[32] SHALMON, M. (1988). Analysis of the GI/GI/1 queue and its variations via the LCFS preemptive resume discipline and its random walk interpretation. *Probab. Engng. Inform. Sci*. **2** 215–230.

[33] SHANTHIKUMAR, J. G. and SUMITA, U. (1986). On $G/G/1$ queues with LIFO-P service discipline. *J. Oper. Res. Soc. Japan* **29** 220–231.

[34] SIGMAN, K. (1996). Queues under preemptive LIFO and ladder height distributions for risk processes: a duality. *Comm. Statist. Stochastic Models* **12** 725–735.

[35] WHITT, W. (1971). Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Probab*. **8** 74–94.

[36] WILLIAMS, R. J. (1996). On the approximation of queueing networks in heavy traffic. In *Stochastic Networks: Theory and Applications* (S. Zachary, F. P. Kelly and I. Ziedins, eds.) 35–56. Clarendon Press, Oxford.

[37] WILLIAMS, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems Theory Appl*. **30** 27–88.

DEPARTMENT OF MATHEMATICS
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853–4201
E-MAIL: limic@math.cornell.edu