The Annals of Applied Probability
2011, Vol. 21, No. 2, 484–545
DOI: 10.1214/10-AAP681

HEAVY TRAFFIC ANALYSIS FOR EDF QUEUES WITH RENEGING

BY ŁUKASZ KRUK[1], JOHN LEHOCZKY[2], KAVITA RAMANAN[3] AND STEVEN SHREVE[4]

Maria Curie-Sklodowska University and Polish Academy of Sciences, Carnegie Mellon University, Brown University and Carnegie Mellon University

This paper presents a heavy-traffic analysis of the behavior of a single-server queue under an Earliest-Deadline-First (EDF) scheduling policy in which customers have deadlines and are served only until their deadlines elapse. The performance of the system is measured by the fraction of reneged work (the residual work lost due to elapsed deadlines) which is shown to be minimized by the EDF policy. The evolution of the lead time distribution of customers in queue is described by a measure-valued process. The heavy traffic limit of this (properly scaled) process is shown to be a deterministic function of the limit of the scaled workload process which, in turn, is identified to be a doubly reflected Brownian motion. This paper complements previous work by Doytchinov, Lehoczky and Shreve on the EDF discipline in which customers are served to completion even after their deadlines elapse. The fraction of reneged work in a heavily loaded system and the fraction of late work in the corresponding system without reneging are compared using explicit formulas based on the heavy traffic approximations. The formulas are validated by simulation results.

1. Introduction.

1.1. *Background and the reneging EDF model.* In the last decade, attention has been paid to queueing systems in which customers have deadlines. Examples include telecommunication systems carrying digitized voice or video traffic, tracking systems and real-time control systems. In the case of voice or video, packetized information must be received, processed and displayed within stringent timing bounds so that the integrity of the transmission is maintained. Similarly, there

Received December 2007; revised August 2009.

[1]Supported in part by the State Committee for Scientific Research of Poland, Grant 2 P03A 012 23 and the EC FP6 Marie Curie ToK programme SPADE 2 at IMPAN, Poland.

[2]Supported in part by ONR and DARPA under MURI Contract N00014-01-1-0576.

[3]Supported in part by the NSF Grants DMS-04-06191 and DMS-04-05343 and CMMI-1059967 (formerly CMMI-0728064).

[4]Supported in part by the NSF Grants DMS-04-04682 and DMS-09-03475.

MSC2010 subject classifications. Primary 60K25; secondary 60G57, 60J65, 68M20.

Key words and phrases. Due dates, heavy traffic, queueing, reneging, diffusion limits, random measures, real-time queues.

are processing requirements for tracking systems that guarantee that a track can be successfully followed. Real-time control systems (e.g., those associated with modern avionics systems, manufacturing plants or automobiles) also gather data that must be processed within stringent timing requirements in order for the system to maintain stability or react to changes in the operating environment. We refer to queueing systems that process tasks with deadlines as "real-time queueing systems."

The performance of a real-time queueing system is measured by its ability to meet the deadlines of the customers. This is in contrast to ordinary queueing systems in which the measure of performance is often customer delay, queue length or utilization of a service facility. We use the fraction of "reneged work," defined as the residual work not serviced due to elapsed deadlines, as our performance measure. To minimize this quantity, it is necessary to use a scheduling policy that takes deadlines into account. We use the Earliest-Deadline-First (EDF) policy, which reduces to the more familiar First-In-First-Out (FIFO) policy when all customers have the same deadline. Under general assumptions, we prove that EDF is optimal with respect to this performance measure. A related result for $G/M/c$ queues, in which the number of reneging customers is used as a performance measure, was obtained by Panwar and Towsley [29].

Heavy traffic analysis of a single real-time queue was initiated by Lehoczky [27]. This was put on a firm mathematical foundation by Doytchinov, Lehoczky and Shreve (DLS) [7]. The accuracy of heavy traffic approximations was developed in [22, 24]. The results of DLS were generalized to the case of acyclic networks in [23]. In these papers it was assumed that all customers are served to completion. The case in which late customers leave the system and their residual work is lost is addressed here. The main result of this paper is a heavy traffic convergence theorem, from which is derived a simple and practically useful approximation for the fraction of lost work when the system is heavily loaded.

The mathematical formulation used by DLS and related papers is based on random measures. In addition to the usual queue length and workload processes associated with the queueing system, to model the evolution of a real-time queueing system, one must keep track of the lead time of each customer, that is, the time until the customer's deadline elapses. This is done by measure-valued queue length and workload processes. The measure-valued queue length process puts unit mass on the real line at the lead time of each customer in the system, while the measure-valued workload process puts mass equal to the remaining service time of each customer at the lead time of that customer. These measures evolve dynamically as customers arrive, age and depart. Under the usual heavy traffic assumptions, since customers are served to completion in the DLS framework, it is easy to see that the ordinary scaled workload process converges weakly to a reflected Brownian motion with drift. DLS showed that the suitably scaled workload and queue length measure-valued processes converge to an explicit deterministic function of the workload process.

In this paper customers leave the system when their deadlines elapse, which we refer to as reneging. Due to the preemptive nature of the EDF policy, it is not possible to determine at the time of admission whether a customer will be fully serviced before its deadline elapses. It is thus natural to have the controller make the decision only at the time the deadline elapses. The system with reneging shows marked improvement in performance over the DLS system, in the sense that the fraction of reneged work in this system is much less than the fraction of work that becomes late in the DLS system. This improvement is because once a customer misses its deadline, the processor devotes no further effort to it, but rather turns its attention to customers that are not late.

The system with reneging is considerably more difficult to analyze than the DLS system. In the reneging system, the evolution of the scalar total workload process depends on the entire lead time distribution of customers in queue and the nature of the EDF discipline. This is in stark contrast to the DLS system, where the total workload process is independent of the scheduling discipline, and is identical to that of any $GI/G/1$ queue with a work-conserving scheduling discipline. A key ingredient of our analysis is a mapping on the space of measure-valued functions which, when applied to the DLS system, yields another system (that we call the reference system) whose difference from the reneging system vanishes in heavy traffic. This mapping can be viewed as a generalization of the scalar double reflection map to measure-valued processes, and, using its continuity properties, we identify the heavy traffic limit of the reference and hence the reneging systems. Specifically, we show that the limit of the scaled workload process is a doubly reflected Brownian motion with lower barrier zero and upper barrier at the mean of the lead time distribution. We also show that, conditional on the limiting workload, the resulting limiting measure-valued workload process is the same limiting process as when customers are served to completion, that is, in the DLS system. However, the workload processes in these two systems differ, and so the unconditional limiting lead-time profiles of these two systems differ accordingly. In particular, unlike in the DLS system, the measure-valued workload process in the reneging system is always concentrated on the positive real line due to the absence of late work in the reneging system.

1.2. *Prediction formulas.* The results of this paper suggest a simple formula for the fraction of lost work in the EDF system with reneging. In particular, consider a single-server queue with traffic intensity $\rho = \lambda/\mu$ that is near one, where $1/\lambda$ is the mean interarrival time and $1/\mu$ is the mean service time. Let α and β be the standard deviations of the interarrival times and service times, respectively, and set $\sigma^2 = \lambda(\alpha^2 + \beta^2)$, which we assume is nonzero. Let $\overline{D}$ denote the mean lead time for arriving customers. Finally, set $\theta = 2(1-\rho)/\sigma^2$. Under these circumstances,

$$\text{(1.1)} \qquad \text{Fraction of lost work in reneging system} \approx e^{-\theta\overline{D}}\left(\frac{1-\rho}{\rho(1-e^{-\theta\overline{D}})}\right).$$

This formula is derived in Section 7.1 and compared with simulations in Section 7.2. If $\rho = 1$, in place of (1.1) we have

$$\text{Fraction of lost work in reneging system} \approx \frac{\sigma^2}{2\overline{D}}. \tag{1.2}$$

Analysis of the limit of the standard (nonreneging) system suggests that when $\rho < 1$ [see (7.8) and (7.9)],

$$\text{Fraction of late work in standard system} \approx e^{-\theta\overline{D}}, \tag{1.3}$$

which, together with (1.1), yields the approximation

$$\frac{\text{Lost work in reneging system}}{\text{Late work in standard system}} \approx \frac{1-\rho}{\rho(1-e^{-\theta\overline{D}})}. \tag{1.4}$$

If $\rho \geq 1$ then all work is late in the limiting standard (nonreneging) system, which leads to the approximation

$$\frac{\text{Lost work in reneging system}}{\text{Late work in standard system}} \approx \frac{\sigma^2}{2\overline{D}}. \tag{1.5}$$

When plotted on a log scale, the fraction of lost work in the reneging system and the fraction of late work in the standard system will be linear in $\overline{D}$, provided that $e^{\theta\overline{D}} \gg 1$, and these two plots will be separated by $\log((1-\rho)/\rho)$. When performance is measured in terms of the work whose service requirement is not met by the time its deadline elapses, then the reneging system is far superior to the nonreneging system. We refer the reader to the simulations in Section 7.2.

The situation with reneged *customers* as opposed to reneged work is more complicated. DLS shows that the number of customers in the limiting standard system at any time is just λ times the amount of work, the number of late customers is λ times the amount of late work and hence

$$\begin{aligned}&\text{Fraction of late customers in the standard system}\\ &\quad\approx \text{Fraction of late work in the standard system}\end{aligned} \tag{1.6}$$

[see also (7.8) and its derivation for the case $\rho < 1$]. In the limiting reneging system, the number of customers who arrive by a certain time and the number of customers in system at that time is λ times the amount of arrived work and λ times the amount of work in the system (Corollary 3.7), respectively, but the number of customers who renege by a certain time is not necessarily λ times the amount of reneged work by that time (see Remark 7.2). In particular, we do not have a formula like (1.6) for the reneging system. If the arrival process is Poisson, the fraction of lost customers in the reneging system can be estimated by a heuristic argument [see (7.7)] which gives instead

$$\begin{aligned}&\text{Fraction of lost customers in reneging system}\\ &\quad\approx \frac{2}{\mu^2\beta^2+1} \times (\text{Fraction of lost work in reneging system}).\end{aligned} \tag{1.7}$$

1.3. *Related work and outline of paper.* Measure-valued processes have recently gained prominence in queueing theory. Decreusefond and Moyal [5] use such processes to obtain the fluid limit of an EDF $M/M/1$ queue with reneging. Unlike our scaling (2.4) of lead times by $\sqrt{n}$, they scale lead times by n and obtain a characterization of the limiting lead-time measure-valued process via a transport equation. In a different setting, Ward and Glynn [33, 34] find limits of FIFO queues with reneging. Measure-valued processes have also proved useful in the heavy traffic analysis of queues with scheduling disciplines other than EDF such as last-in-first-out [28], processor sharing [11, 12], and shortest remaining processing time [6, 13]. As dynamical systems, queueing systems present a mathematical challenge due to discontinuities in their evolution at boundaries (which denote empty queues). The heavy traffic analysis of queueing systems described by $\mathbb{R}^n$-valued processes has been facilitated by the use of representations in terms of continuous mappings on $\mathbb{R}^n$ [4, 8, 14, 31, 36]. This work demonstrates that this perspective can also be useful when the queueing system is represented by a more complicated, measure-valued process (see also [18] for recent work that takes a similar perspective).

Section 2 introduces our model. Section 3 summarizes the main results, and proofs of these results are given in Section 6. Section 4 introduces the reference workload process and its decomposition, and describes its evolution. This reference workload process is easier to analyze than the workload process with reneging but the two are shown to have the same asymptotic behavior. Comparisons between the reference workload process and the reneging workload process are presented in Section 5. Section 7 presents simulation results. A proof of optimality of EDF, that may be of independent interest, is in the Appendix.

2. The model, assumptions and notation.

2.1. *Notation.* Let $\mathbb{R}$ be the set of real numbers. For $a, b \in \mathbb{R}$, $a \vee b$ is the maximum of a and b, $a \wedge b$ is the minimum and a^+ is the maximum of a and 0. Also, $\inf\{\varnothing\}$ should be understood as $+\infty$, while $\sup\{\varnothing\}$ and $\max\{\varnothing\}$ should be understood as $-\infty$. Moreover, if $a < b$, then the interval $[b, a]$ is understood to be $\varnothing$.

Denote by $\mathcal{M}$ the set of all finite, nonnegative measures on $\mathcal{B}(\mathbb{R})$, the Borel subsets of $\mathbb{R}$. Under the weak topology, $\mathcal{M}$ is a Polish space. We denote the measure in $\mathcal{M}$ that puts one unit of mass at the point $x \in \mathbb{R}$, that is, the Dirac measure at x, by δ_x. When $\nu \in \mathcal{M}$ and B is an interval $(a, b]$ or a singleton $\{a\}$, we will simply write $\nu(a, b]$ and $\nu\{a\}$ instead of $\nu((a, b])$ and $\nu(\{a\})$.

Let $T > 0$ be given. Given a Polish space X, we use $D_X[0, \infty)$ (resp., $D_X[0, T]$) to denote the space of right-continuous functions with left-hand limits (RCLL functions) from $[0, \infty)$ (resp., $[0, T]$) to X, equipped with the Skorokhod J_1 topology. See [9] for details. When dealing with $D_X[0, \infty)$ or $D_X[0, T]$, we typically consider $X = \mathbb{R}$ or $\mathbb{R}^d$, with appropriate dimension d for vector-valued functions,

or $X = \mathcal{M}$, unless explicitly stated otherwise. When $X = \mathbb{R}$ or $\mathcal{M}$, for $t > 0$ and $x \in D_X[0, \infty)$, we write $x(t-)$ for the left-hand limit $\lim_{s \uparrow t} x(s)$, and we define $\triangle x(t)$ to be the jump in x at time t, that is, $\triangle x(t) \stackrel{\Delta}{=} x(t) - x(t-)$. Finally, given $D_X[0, \infty)$-valued random variables $Z_n, n \in \mathbb{N}$, defined, respectively, on the probability spaces $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$, $n \in \mathbb{N}$, and a $D_X[0, \infty)$-valued random variable Z defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say $Z^{(n)}$ converges in distribution to Z and write $Z_n \Rightarrow Z$, if for every bounded continuous function f on $D_X[0, \infty)$, $\lim_{n \to \infty} \mathbb{E}_n[f(Z_n)] = \mathbb{E}[f(Z)]$. Here $\mathbb{E}_n$ and $\mathbb{E}$ are expectations taken with respect to $\mathbb{P}_n$ and $\mathbb{P}$, respectively.

2.2. *The model with reneging.* We have a sequence of single-station queueing systems, each serving one class of customers. The queueing systems are indexed by superscript (n). The inter-arrival times for the customers are $\{u_j^{(n)}\}_{j=1}^{\infty}$, a sequence of strictly positive, independent, identically distributed random variables with common mean $\frac{1}{\lambda^{(n)}}$ and standard deviation $\alpha^{(n)}$. The service times are $\{v_j^{(n)}\}_{j=1}^{\infty}$, another sequence of positive, independent, identically distributed random variables with common mean $\frac{1}{\mu^{(n)}}$ and standard deviation $\beta^{(n)}$.

If the initial condition of the nth queue were not zero, then we would need to specify an initial workload measure-valued process and frontier [these terms are defined in (2.17) and (2.19) below] in such a way that these have limits under the heavy traffic scaling. However, if the limit of the initial scaled workload process were not of the form appearing in Theorem 3.2 below, then the workload process would be expected to have a jump at time zero. To avoid these complications, we assume that each queue is empty at time zero.

We define the *customer arrival times*

$$S_0^{(n)} \stackrel{\Delta}{=} 0, \qquad S_k^{(n)} \stackrel{\Delta}{=} \sum_{i=1}^{k} u_i^{(n)}, \qquad k \geq 1, \tag{2.1}$$

the *customer arrival process*

$$A^{(n)}(t) \stackrel{\Delta}{=} \max\{k; S_k^{(n)} \leq t\}, \qquad t \geq 0, \tag{2.2}$$

and the *work arrival process*

$$V^{(n)}(t) \stackrel{\Delta}{=} \sum_{j=1}^{\lfloor t \rfloor} v_j^{(n)}, \qquad t \geq 0. \tag{2.3}$$

The work that has arrived to the queue by time t is then $V^{(n)}(A^{(n)}(t))$.

Each customer arrives with an initial lead time $L_j^{(n)}$, the time between the arrival time and the deadline for completion of service for that customer. These initial lead times are independent and identically distributed with

$$\mathbb{P}\{L_j^{(n)} \leq \sqrt{n} y\} = G(y), \tag{2.4}$$

where G is a right-continuous cumulative distribution function. We define

$$y_* \triangleq \inf\{y \in \mathbb{R} | G(y) > 0\}, \qquad y^* \triangleq \min\{y \in \mathbb{R} | G(y) = 1\} \tag{2.5}$$

and assume that $0 < y_* \leq y^* < +\infty$. We assume that for every n, the sequences $\{u_j^{(n)}\}_{j=1}^{\infty}$, $\{v_j^{(n)}\}_{j=1}^{\infty}$ and $\{L_j^{(n)}\}_{j=1}^{\infty}$ are mutually independent. See Remark 3.9 for a discussion of these assumptions.

We assume that customers are served using the Earliest-Deadline-First (EDF) queue discipline, that is, the customer with the shortest lead time receives service. Preemption occurs when a customer more urgent than the customer in service arrives (we assume preempt-resume). There is no set up, switch-over, or other type of overhead. If the jth customer is still present in the system (either waiting for service or receiving it) when his deadline passes, that is, at the time $S_j^{(n)} + L_j^{(n)}$, he leaves the queue immediately. This may be interpreted as either reneging or the result of an action of an external controller.

We define $W^{(n)}(t)$, the *workload process* at time t, as the remaining processing time of all the customers in the system at this time. We define $R_W^{(n)}(t)$ to be the amount of work that reneges in the time interval $[0, t]$. The *queue length process* $Q^{(n)}(t)$ is the number of customers in the queue at time t. The queueing system described above will be referred to as the *EDF system with reneging*.

2.3. *The standard EDF model.* We also have a sequence, indexed by superscript (n), of *standard EDF systems*, with the same stochastic primitives as the EDF systems with reneging. In each of these standard systems, the server serves the customer with the shortest lead time, preemption occurs as in the reneging system, but late customers (customers with negative lead times) stay in the system until served to completion. The performance processes associated with the standard system will be denoted by the same symbols as their counterparts from the system with reneging, but with additional subscript S. For example, $W_S^{(n)}(t)$ denotes the workload in the standard system at time t. The arrival processes $A^{(n)}(t)$ and $V^{(n)}(t)$ are the same for the both systems, so we will not attach the subscript S to them.

The standard EDF system is easier to analyze than the EDF system with reneging in several ways. For instance, the workload $W_S^{(n)}$ in the standard system coincides with the workload of a corresponding G/G/1 queue (with the same primitives) under any nonidling scheduling policy. More precisely, in the standard system the *netput process*

$$N^{(n)}(t) \triangleq V^{(n)}\big(A^{(n)}(t)\big) - t \tag{2.6}$$

measures the amount of work in queue at time t provided that the server is never idle up to time t, and the *cumulative idleness process*

$$I_S^{(n)}(t) \triangleq -\inf_{0 \leq s \leq t} N^{(n)}(s) \tag{2.7}$$

gives the amount of time the server is idle. Adding these two processes together, we obtain the workload process for the standard system

$$W_S^{(n)}(t) = N^{(n)}(t) + I_S^{(n)}(t). \tag{2.8}$$

(All the above processes are RCLL.) In contrast, the evolution of the workload $W^{(n)}$ in the reneging system is more complex and depends not only on the residual service times but also on the lead times of all customers in the queue. Our analysis of the reneging system will be facilitated by results from [7] on the heavy traffic analysis of the standard EDF system.

2.4. *Heavy traffic assumptions.* We assume that the following limits exist:

$$\begin{aligned} &\lim_{n\to\infty} \lambda^{(n)} = \lambda, \qquad \lim_{n\to\infty} \mu^{(n)} = \lambda, \\ &\lim_{n\to\infty} \alpha^{(n)} = \alpha, \qquad \lim_{n\to\infty} \beta^{(n)} = \beta, \end{aligned} \tag{2.9}$$

and, moreover, $\lambda > 0$ and $\alpha^2 + \beta^2 > 0$. Define the *traffic intensity* $\rho^{(n)} \triangleq \frac{\lambda^{(n)}}{\mu^{(n)}}$. We make the *heavy traffic assumption*

$$\lim_{n\to\infty} \sqrt{n}\big(1 - \rho^{(n)}\big) = \gamma \tag{2.10}$$

for some $\gamma \in \mathbb{R}$. We also impose the *Lindeberg condition* on the inter-arrival and service times: for every $c > 0$,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\big[\big(u_j^{(n)} - \big(\lambda^{(n)}\big)^{-1}\big)^2 \mathbb{I}_{\{|u_j^{(n)} - (\lambda^{(n)})^{-1}| > c\sqrt{n}\}}\big] \\ &\quad = \lim_{n\to\infty} \mathbb{E}\big[\big(v_j^{(n)} - \big(\mu^{(n)}\big)^{-1}\big)^2 \mathbb{I}_{\{|v_j^{(n)} - (\mu^{(n)})^{-1}| > c\sqrt{n}\}}\big] = 0. \end{aligned} \tag{2.11}$$

We introduce the *heavy traffic scaling* for the idleness process in the standard system and the workload and queue length processes for both EDF systems

$$\begin{aligned} &\widehat{I}_S^{(n)}(t) = \frac{1}{\sqrt{n}} I_S^{(n)}(nt), \qquad \widehat{W}_S^{(n)}(t) = \frac{1}{\sqrt{n}} W_S^{(n)}(nt), \\ &\widehat{Q}_S^{(n)}(t) = \frac{1}{\sqrt{n}} Q_S^{(n)}(nt), \qquad \widehat{W}^{(n)}(t) = \frac{1}{\sqrt{n}} W^{(n)}(nt), \\ &\widehat{Q}^{(n)}(t) = \frac{1}{\sqrt{n}} Q^{(n)}(nt) \end{aligned}$$

and the *centered heavy traffic scaling* for the arrival processes

$$\begin{aligned} &\widehat{S}^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} \Big(u_j^{(n)} - \frac{1}{\lambda^{(n)}} \Big), \qquad \widehat{V}^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} \Big(v_j^{(n)} - \frac{1}{\mu^{(n)}} \Big), \\ &\widehat{A}^{(n)}(t) = \frac{1}{\sqrt{n}} \big[A^{(n)}(nt) - \lambda^{(n)} nt \big]. \end{aligned}$$

The scaled netput process (which is the same for both systems) is given by

$$\widehat{N}^{(n)}(t) = \frac{1}{\sqrt{n}}\bigl[V^{(n)}\bigl(A^{(n)}(nt)\bigr) - nt\bigr]. \tag{2.12}$$

Note that, by (2.8), $\widehat{W}_S^{(n)}(t) = \widehat{N}^{(n)}(t) + \widehat{I}_S^{(n)}(t)$.

It follows from Theorem 3.1 in [30] and Theorem 7.3.2 in [36] that

$$\bigl(\widehat{S}^{(n)}, \widehat{A}^{(n)}\bigr) \Rightarrow (S^*, A^*), \tag{2.13}$$

where A^* is a zero-drift Brownian motion with variance $\alpha^2\lambda^3$ per unit time and

$$S^*(\lambda t) = -\frac{1}{\lambda}A^*(t), \qquad t \geq 0. \tag{2.14}$$

It is a standard result [16] that

$$\bigl(\widehat{N}^{(n)}, \widehat{I}_S^{(n)}, \widehat{W}_S^{(n)}\bigr) \Rightarrow (N^*, I_S^*, W_S^*), \tag{2.15}$$

where N^* is a Brownian motion with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$,

$$I_S^*(t) \triangleq -\min_{0\leq s\leq t} N^*(s), \qquad W_S^*(t) = N^*(t) + I_S^*(t). \tag{2.16}$$

In other words, W_S^* is a Brownian motion reflected at 0 with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$ and I_S^* causes the reflection.

2.5. *Measure-valued processes and frontiers.* To study whether tasks or customers meet their timing requirements, one must keep track of customer lead times. The action of the EDF discipline requires knowledge of the current lead times of all customers in system. We represent this information via a collection of measure-valued stochastic processes.

Customer arrival measure-valued process:

$$\mathcal{A}^{(n)}(t)(B) \triangleq \left\{\begin{array}{l}\text{Number of arrivals by time } t\text{, whether or not still in the}\\ \text{system at time } t\text{, having lead times at time } t \text{ in } B \in \mathcal{B}(\mathbb{R})\end{array}\right\}.$$

Workload arrival measure-valued process:

$$\mathcal{V}^{(n)}(t)(B) \triangleq \left\{\begin{array}{l}\text{Work arrived by time } t\text{, whether or not still in the sys-}\\ \text{tem at time } t\text{, having lead times at time } t \text{ in } B \in \mathcal{B}(\mathbb{R})\end{array}\right\}.$$

Queue length measure-valued process:

$$\mathcal{Q}^{(n)}(t)(B) \triangleq \left\{\begin{array}{l}\text{Number of customers in the queue at time}\\ t \text{ having lead times at time } t \text{ in } B \in \mathcal{B}(\mathbb{R})\end{array}\right\}.$$

Workload measure-valued process:

$$\mathcal{W}^{(n)}(t)(B) \triangleq \left\{\begin{array}{l}\text{Work in the queue at time } t \text{ associated with cus-}\\ \text{tomers having lead times at time } t \text{ in } B \in \mathcal{B}(\mathbb{R})\end{array}\right\}. \tag{2.17}$$

The latter two processes describe the behavior of the EDF system with reneging. Their counterparts for the standard EDF system will be denoted by $\mathcal{Q}_S^{(n)}(t)$ and $\mathcal{W}_S^{(n)}(t)$, respectively. The following relationships easily follow:

$$A^{(n)}(t) = \mathcal{A}^{(n)}(t)(\mathbb{R}), \qquad V^{(n)}\big(A^{(n)}(t)\big) = \mathcal{V}^{(n)}(t)(\mathbb{R}),$$
$$W^{(n)}(t) = \mathcal{W}^{(n)}(t)(0,\infty), \qquad Q^{(n)}(t) = \mathcal{Q}^{(n)}(t)(0,\infty),$$
$$W_S^{(n)}(t) = \mathcal{W}_S^{(n)}(t)(\mathbb{R}), \qquad Q_S^{(n)}(t) = \mathcal{Q}_S^{(n)}(t)(\mathbb{R}).$$

In addition, we can represent the reneged work as follows:

$$R_W^{(n)}(t) = \sum_{0<s\le t} \mathcal{W}^{(n)}(s-)\{0\}. \tag{2.18}$$

In order to study the behavior of the EDF queue discipline, it is useful to keep track of the largest lead time of all customers, whether present or departed, who have ever been in service. We define the *frontier*

$$F^{(n)}(t) \triangleq \left\{ \begin{array}{l} \text{The maximum of the largest lead time of} \\ \text{all customers who have ever been in service,} \\ \text{whether still present or not, and } \sqrt{n}\,y^* - t \end{array} \right\} \tag{2.19}$$

for the EDF system with reneging, and its counterpart $F_S^{(n)}(t)$ for the standard EDF system. Prior to arrival of the first customer, $F^{(n)}(t)$ and $F_S^{(n)}(t)$ equal $\sqrt{n}\,y^* - t$. For the EDF system with reneging, we define the *current lead time*

$$C^{(n)}(t) \triangleq \left\{ \begin{array}{l} \text{Lead time of the customer in service} \\ \text{or } F^{(n)}(t) \text{ if the queue is empty} \end{array} \right\}.$$

In the reneging system, there is no customer with lead time smaller than $C^{(n)}(t)$, and there has never been a customer in service whose lead time, if the customer were still present, would exceed $F^{(n)}(t)$. Furthermore, $C^{(n)}(t) \le F^{(n)}(t)$ for all $t \ge 0$. The processes $C^{(n)}$, $F^{(n)}$ and $F_S^{(n)}$ are RCLL.

We introduce heavy traffic scalings. For the real-valued processes $Z^{(n)} = C^{(n)}, F^{(n)}, F_S^{(n)}, W^{(n)}, Q^{(n)}, R_W^{(n)}$, we define $\widehat{Z}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} Z^{(n)}(nt)$ and for the measure-valued processes $\mathcal{Z}^{(n)} = \mathcal{Q}^{(n)}, \mathcal{W}^{(n)}, \mathcal{Q}_S^{(n)}, \mathcal{W}_S^{(n)}, \mathcal{A}^{(n)}, \mathcal{V}^{(n)}$, we define $\widehat{\mathcal{Z}}^{(n)}(t)(B) \triangleq \frac{1}{\sqrt{n}} \mathcal{Z}^{(n)}(nt)(\sqrt{n}B)$ for every Borel set $B \subset \mathbb{R}$.

3. Main results. Before stating our main results, we summarize the results for the standard EDF system that were obtained in [7]—in particular, we recall Proposition 3.10 and Theorem 3.1 of [7] which characterize the limiting distributions of the workload measure and the queue length measure in the standard system. Let

$$H(y) \triangleq \int_y^\infty \big(1 - G(\eta)\big)\, d\eta = \left\{ \begin{array}{ll} \displaystyle\int_y^{y^*} \big(1 - G(\eta)\big)\, d\eta, & \text{if } y \le y^*, \\ 0, & \text{if } y > y^*. \end{array} \right. \tag{3.1}$$

The function H maps $(-\infty, y^*]$ onto $[0, \infty)$ and is strictly decreasing and Lipschitz continuous with Lipschitz constant 1 on $(-\infty, y^*]$. Therefore, there exists a continuous inverse function H^{-1} that maps $[0, \infty)$ onto $(-\infty, y^*]$.

PROPOSITION 3.1 (Proposition 3.10 [7]). *We have $\widehat{F}_S^{(n)} \Rightarrow F_S^*$ as $n \to \infty$, where the limiting scaled frontier process F_S^* for the standard EDF system is explicitly given by*

$$F_S^*(t) \triangleq H^{-1}(W_S^*(t)), \qquad t \geq 0, \tag{3.2}$$

with W_S^* equal to Brownian motion with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$, reflected at 0.

THEOREM 3.2 (Theorem 3.1 [7]). *Let $\mathcal{W}_S^*$ and $\mathcal{Q}_S^*$ be the measure-valued processes defined, respectively, by*

$$\mathcal{W}_S^*(t)(B) \triangleq \int_{B \cap [F_S^*(t), \infty)} \big(1 - G(y)\big)\, dy, \qquad \mathcal{Q}_S^*(t)(B) \triangleq \lambda \mathcal{W}_S^*(t)(B), \tag{3.3}$$

for all Borel sets $B \subseteq \mathbb{R}$. Then $\widehat{\mathcal{W}}_S^{(n)} \Rightarrow \mathcal{W}_S^*$ and $\widehat{\mathcal{Q}}_S^{(n)} \Rightarrow \mathcal{Q}_S^*$, as $n \to \infty$.

REMARK 3.3. The proofs in [7] can be modified to show that the convergences in (3.3) are in fact joint, that is, $(\widehat{\mathcal{W}}_S^{(n)}, \widehat{\mathcal{Q}}_S^{(n)}) \Rightarrow (\mathcal{W}_S^*, \mathcal{Q}_S^*)$.

There is lateness in the standard EDF system if and only if the measure-valued workload process has positive mass on the negative half line. Theorem 3.2 shows that, in the heavy traffic limit, this occurs exactly when the limiting scaled frontier process F_S^* lies to the left of 0 or, equivalently (by Proposition 3.1), when W_S^* is greater than $H(0) = \mathbb{E}[L_j^{(n)}/\sqrt{n}]$, the mean of the scaled lead-time distribution. In the reneging system, there is no lateness, and the amount of work that reneges is precisely the amount required to prevent lateness. Thus it is natural to expect that the limiting workload in the reneging system will be constrained to remain below $H(0)$. Let W^* be a Brownian motion with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$, reflected at 0 and $H(0)$. The first main result of this paper is that W^* is the limiting workload in the reneging system.

THEOREM 3.4. *As $n \to \infty$, $\widehat{W}^{(n)} \Rightarrow W^*$.*

The next two results of this paper are the following counterparts of Proposition 3.1 and Theorem 3.2 for the EDF system with reneging.

PROPOSITION 3.5. *We have $\widehat{F}^{(n)} \Rightarrow F^*$ as $n \to \infty$, where*

$$F^*(t) \triangleq H^{-1}(W^*(t)), \qquad t \geq 0. \tag{3.4}$$

In other words, the process F^* defined by (3.4) is the limiting scaled frontier process for the EDF system with reneging.

THEOREM 3.6. *Let $\mathcal{W}^*$ and $\mathcal{Q}^*$ be the measure-valued processes defined by*

$$\mathcal{W}^*(t)(B) \triangleq \int_{B\cap[F^*(t),\infty)} \big(1 - G(y)\big)\, dy, \qquad \mathcal{Q}^*(t)(B) \triangleq \lambda \mathcal{W}^*(t)(B), \tag{3.5}$$

for all Borel sets $B \subseteq \mathbb{R}$. Then $(\widehat{\mathcal{W}}^{(n)}, \widehat{\mathcal{Q}}^{(n)}) \Rightarrow (\mathcal{W}^*, \mathcal{Q}^*)$ as $n \to \infty$.

By Theorem 3.6, the total masses of $\mathcal{W}^{(n)}$ and $\mathcal{Q}^{(n)}$ must converge jointly to the total masses of $\mathcal{W}^*$ and $\mathcal{Q}^{(n)}$, respectively. Substituting $B = \mathbb{R}$ in (3.5) and using (3.1) and (3.4), we see that $\mathcal{W}^*(t)(\mathbb{R}) = H(F^*(t)) = W^*(t)$ and we recover Theorem 3.4. In fact, we have a stronger result.

COROLLARY 3.7. *As $n \to \infty$, $(\widehat{W}^{(n)}, \widehat{Q}^{(n)}) \Rightarrow (W^*, \lambda W^*)$.*

Theorem 3.6 also shows that the limiting instantaneous lead-time profiles of customers in the EDF system with reneging *conditioned on the value of the* (*limiting*) *workload in the system* are the same as in the case of the standard EDF system. However, the limiting real-valued workload process for the EDF system with reneging is W^*, the *doubly reflected* Brownian motion and the *unconditional* limiting lead-time profiles for these two systems differ accordingly.

We also have a characterization of the limiting amount of reneged work.

THEOREM 3.8. *As $n \to \infty$, $\widehat{R}_W^{(n)} \Rightarrow R_W^*$, where R_W^* is the local time at $H(0)$ of the doubly reflected Brownian motion W^*.*

Although these results are intuitive in light of the behavior of the standard EDF system, the proofs are challenging. Moreover, counter to what one might expect, the result for queue lengths analogous to Theorem 3.8 is false. Specifically, although Corollary 3.7 shows that $\widehat{Q}^{(n)}$ converges to the doubly reflected Brownian motion $Q^* \triangleq \lambda W^*$ on $[0, \lambda H(0)]$, the scaled sequence $\widehat{R}_Q^{(n)}, n \in \mathbb{N}$, of reneged customers *does not* converge to the local time λR_W^* of Q^* at $\lambda H(0)$. This observation, which is elaborated upon in Section 7, emphasizes the need for a rigorous justification of intuitive statements.

The proof of Theorem 3.4 is in Section 6.1.1, the proofs of Proposition 3.5 and Theorem 3.6 are in Section 6.1.2, and Section 6.2 contains the proof of Theorem 3.8. We also establish an optimality property for EDF, Theorem 5.1.

REMARK 3.9. The assumption made in (2.5) that the support of the lead time distribution is bounded above by $y^* < \infty$ is mainly technical. It is expected that the analysis in [21] for the standard EDF system under a weaker second moment

condition can be applied to the reneging system as well. On the other hand, the lower bound $y_* > 0$ on the lead time distribution or some restriction on the behavior of the density of the lead time distribution at 0 appears to be necessary. Indeed, the work of Ward and Glynn [33, 34] on FIFO queues with reneging suggests that in the absence of such an assumption, the limiting workload process may no longer be a reflected Brownian motion, and its properties may exhibit strong sensitivity to the density of the lead-time distribution near 0. From a modeling point of view, it is reasonable to impose a strictly positive lower bound $y_* > 0$ so as to avoid nonnegligible "intrinsic lateness," in which an arriving customer has such a small initial lead time that he would be late even if there were no other customers in the system.

In [21] the assumption of independence between the sequence of interarrival times and lead times is also removed and a more complex version of Theorem 3.2 is obtained. Starting from that more complex result, the limit of the reneging system can be obtained along the lines of this paper.

4. The reference system. In this section we introduce an auxiliary reference workload measure-valued process $\mathcal{U}^{(n)}$ and the corresponding real-valued reference workload process $U^{(n)}$. In the special case of constant initial lead times (i.e., $y_* = y^*$), in which EDF reduces to the well-known FIFO service discipline, $\mathcal{U}^{(n)}$ and $U^{(n)}$ coincide with $\mathcal{W}^{(n)}$ and $W^{(n)}$, respectively. In general, these processes do not coincide (see Example 4.6), but, as we will show in Section 6.1, the difference between the diffusion-scaled versions of $U^{(n)}$ and $W^{(n)}$ is negligible under heavy-traffic conditions. The advantage of working with the reference system, rather than the reneging system, is that $\mathcal{U}^{(n)}$ can be represented explicitly as a certain mapping Φ of the measure-valued workload process $\mathcal{W}_S^{(n)}$ in the standard system. As shown in Section 6.1, continuity properties of the mapping Φ enable an easy characterization of the limiting distributions of $\mathcal{U}^{(n)}$ and $U^{(n)}$ in heavy traffic.

We begin with Section 4.1, where we define the reference system and provide a useful decomposition of the process $U^{(n)}$. In Section 4.2 we provide a detailed description of the evolution of $\mathcal{U}^{(n)}$.

4.1. *Definition and properties of the reference workload.* In Section 4.1.1, we introduce a deterministic mapping on the space of measure-valued functions that is used to define the reference workload. Then, in Section 4.1.2, we provide a decomposition of the reference workload process.

4.1.1. *A mapping Φ of measure-valued processes.* We define a sequence of *reference workload measure-valued processes* for the EDF system with reneging by the formula

$$\mathcal{U}^{(n)} \triangleq \Phi(\mathcal{W}_S^{(n)}), \tag{4.1}$$

where the mapping $\Phi : D_{\mathcal{M}}[0,\infty) \mapsto D_{\mathcal{M}}[0,\infty)$ is defined by

$$\begin{aligned}&\Phi(\mu)(t)(-\infty, y]\\ &\quad\triangleq \Big[\mu(t)(-\infty, y] - \sup_{s\in[0,t]}\Big(\mu(s)(-\infty, 0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R})\Big)\Big]^+\end{aligned} \tag{4.2}$$

for every $\mu \in D_{\mathcal{M}}[0,\infty)$, $t \geq 0$ and $y \in \mathbb{R}$. (The claim that Φ does indeed map $D_{\mathcal{M}}[0,\infty)$ into $D_{\mathcal{M}}[0,\infty)$ is justified in Lemma 4.1 below.) We also define the (real-valued) *reference workload process* $U^{(n)}$ as the total mass of $\mathcal{U}^{(n)}$, that is,

$$U^{(n)}(t) \triangleq \mathcal{U}^{(n)}(t)(\mathbb{R}) \qquad \forall t \in [0,\infty). \tag{4.3}$$

The frontier $F_S^{(n)}$ defined in Section 2.3 played a crucial role in the description and analysis of the evolution of the standard system in [7]. In a similar fashion, it will be useful to define the *reference frontier*

$$E^{(n)}(t) \triangleq \begin{cases} \inf\{y \in \mathbb{R} | \mathcal{U}^{(n)}(t)(-\infty, y] > 0\}, & \text{if } U^{(n)}(t) > 0,\\ +\infty, & \text{if } U^{(n)}(t) = 0. \end{cases} \tag{4.4}$$

By definition, $E^{(n)}(t)$ is the leftmost point of support of the random measure $\mathcal{U}^{(n)}(t)$ [understood as ∞ if $\mathcal{U}^{(n)}(t) \equiv 0$]. The process $E^{(n)}$ has RCLL paths.

From (4.1)–(4.3) we have

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \big[\mathcal{W}_S^{(n)}(t)(-\infty, y] - K^{(n)}(t)\big]^+, \tag{4.5}$$

$$U^{(n)}(t) = W_S^{(n)}(t) - K^{(n)}(t), \tag{4.6}$$

where

$$K^{(n)}(t) \triangleq \max_{s\in[0,t]}\Big\{\mathcal{W}_S^{(n)}(s)(-\infty, 0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u)\Big\}. \tag{4.7}$$

In (4.7) we may write maximum rather than supremum because the process $\mathcal{W}_S^{(n)}(\cdot)(-\infty, 0]$ never jumps down. Note from (4.6) and (4.7) that $0 \leq K^{(n)}(t) \leq W_S^{(n)}(t)$ and so for all $t \geq 0$,

$$0 \leq U^{(n)}(t) \leq W_S(t). \tag{4.8}$$

According to (4.6), the reference workload process $U^{(n)}$ is the standard workload process $W_S^{(n)}$ with mass $K^{(n)}$ removed. Equation (4.5) shows that this mass is removed from the left-hand side of the support of $\mathcal{W}_S^{(n)}$. Moreover, since $\mathcal{U}^{(n)}(t)(-\infty, y] > 0$ for all y to the right of the frontier $E^{(n)}(t)$, it is clear from (4.1) and (4.2) that for $t \in [0,\infty)$, $y_2 \geq y_1 > E^{(n)}(t)$,

$$\mathcal{U}^{(n)}(t)(y_1, y_2] = \mathcal{U}^{(n)}(t)(-\infty, y_2] - \mathcal{U}^{(n)}(t)(-\infty, y_1] = \mathcal{W}_S^{(n)}(t)(y_1, y_2], \tag{4.9}$$

which shows that $\mathcal{U}^{(n)}$ coincides with $\mathcal{W}_S^{(n)}$ strictly to the right of $E^{(n)}$.

In the following lemma, we establish some basic properties of Φ that show, in particular, that $\mathcal{U}^{(n)}(t)$, $t \geq 0$, and $U^{(n)}(t)$, $t \geq 0$, are stochastic processes with sample paths in $D_{\mathcal{M}}[0,\infty)$ and $D_{\mathbb{R}_+}[0,\infty)$, respectively. Although Φ is not continuous on $D_{\mathcal{M}}[0,\infty)$, the lemma shows that it satisfies a certain continuity property that will be sufficient for our purposes.

LEMMA 4.1. *For every $t \in [0,\infty)$, $\Phi(\mu)(t)(-\infty,0] = 0$. Moreover, Φ maps $D_{\mathcal{M}}[0,\infty)$ to $D_{\mathcal{M}}[0,\infty)$. Furthermore, if a sequence $\mu_n, n \in \mathbb{N}$, in $D_{\mathcal{M}}[0,\infty)$ converges to $\mu \in D_{\mathcal{M}}[0,\infty)$, where μ is continuous and for every $t \in [0,\infty)$, $\mu(t)\{0\} = 0$, then $\Phi(\mu_n)$ converges to $\Phi(\mu)$ in $D_{\mathcal{M}}[0,\infty)$.*

PROOF. The first statement follows from the simple observation that, due to the nonnegativity of μ and (4.2),

$$0 \leq \Phi(\mu)(t)(-\infty,0] \leq \big[\mu(t)(-\infty,0] - \mu(t)(-\infty,0] \wedge \mu(t)(\mathbb{R})\big]^+ = 0.$$

Also, since the right-hand side of (4.2) is nondecreasing and right-continuous in y, we know that $\Phi(\mu)(t) \in \mathcal{M}$ for every $t \geq 0$. Now, observe that $\Phi(\mu)(t) = \Psi(\mu(t), \Gamma(\mu)(t))$, where $\Psi : \mathcal{M} \times \mathbb{R} \mapsto \mathcal{M}$ is the mapping $\Psi(\nu, x)(-\infty, y] \stackrel{\Delta}{=} (\nu(-\infty, y] - x)^+$ for all $y \in \mathbb{R}$ and $\Gamma : D_{\mathcal{M}}[0,\infty) \mapsto \mathbb{R}$ is defined by

$$\Gamma(\mu)(t) \stackrel{\Delta}{=} \sup_{s\in[0,t]} \Big(\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R})\Big) \qquad \forall t \in [0,\infty).$$

Using the fact that weak convergence of measures on $\mathbb{R}$ is equivalent to convergence of the cumulative distribution functions at continuity points of the limit, one can verify that Ψ is continuous on $\mathcal{M} \times \mathbb{R}$. To show that $\Phi(\mu) \in D_{\mathcal{M}}[0,\infty)$, it suffices to show that $\Gamma(\mu) \in D[0,\infty)$. For this, we fix $t \in [0,\infty)$ and write

$$\begin{aligned}
&\Gamma(\mu)(t+\varepsilon) - \Gamma(\mu)(t) \\
&\quad = \sup_{s\in[0,t]} \Big[\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R}) \wedge \inf_{u\in[t,t+\varepsilon]} \mu(u)(\mathbb{R})\Big] \vee Z(\mu,\varepsilon)(t) \\
&\qquad - \sup_{s\in[0,t]} \Big[\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R})\Big],
\end{aligned}$$

where we define

$$Z(\mu,\varepsilon)(t) \stackrel{\Delta}{=} \sup_{s\in[t,t+\varepsilon]} \Big[\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t+\varepsilon]} \mu(u)(\mathbb{R})\Big].$$

Since $\mu \in D_{\mathcal{M}}[0,\infty)$ implies $\mu(u)$ converges weakly to $\mu(t)$ as $u \downarrow t$, we have $\lim_{u\downarrow t} \mu(u)(\mathbb{R}) = \mu(t)(\mathbb{R})$ and $\mu(t)(-\infty,0] \geq \limsup_{s\downarrow t} \mu(s)(-\infty,0]$ by Portmanteau's theorem. This, in turn, implies that $\lim_{\varepsilon\to 0} Z(\mu,\varepsilon)(t) = \mu(t)(-\infty,0]$ for all $t \geq 0$. Combining the above properties, it is easy to deduce that $\Gamma(\mu)(t+\varepsilon) - \Gamma(\mu)(t) \to 0$ as $\varepsilon \downarrow 0$, and the right-continuity of $\Phi(\mu)$ follows. The existence of left limits for $\Gamma(\mu)$, and hence for $\Phi(u)$, can be established by an analogous but simpler argument.

Now, suppose μ_n converges to μ in $D_{\mathcal{M}}[0,\infty)$ and μ is continuous with $\mu(t)\{0\}=0$ for every $t\geq 0$. Then $\mu_n(t)$ converges weakly to $\mu(t)$ uniformly for t in compact sets (u.o.c.) (see [2]). Since 0 is a continuity point for $\mu(t)$, this implies $\mu_n(t)(-\infty,0]$ and $\mu_n(t)(\mathbb{R})$ converge u.o.c. to $\mu(t)(-\infty,0]$ and $\mu(t)(\mathbb{R})$, respectively. This shows that $\Gamma(\mu_n)(t)$ converges u.o.c. to $\Gamma(\mu)(t)$, which, when combined with the continuity of Ψ, shows that $\Phi(\mu_n)(t)$ converges weakly u.o.c. to $\Phi(\mu)(t)$. In particular, this shows $\Phi(\mu_n)$ converges to $\Phi(\mu)$ in $D_{\mathcal{M}}[0,\infty)$. $\square$

As an immediate consequence of the lemma, the definitions of $\mathcal{U}^{(n)}$ and $E^{(n)}$, and the fact that $\mathcal{U}^{(n)}(t)$ is a purely atomic measure, we have, for all $t\geq 0$,

$$\mathcal{U}^{(n)}(t)(-\infty,0]=0 \quad \text{and} \quad E^{(n)}(t)>0. \tag{4.10}$$

4.1.2. *A decomposition of the reference workload.* We establish a decomposition of $K^{(n)}$ into its increasing and decreasing parts. Define $\sigma_0^{(n)} \triangleq 0$ and $W_S^{(n)}(0-) \triangleq 0$. For $k=0,1,2,\ldots$, define recursively

$$\tau_k^{(n)} \triangleq \min\Big\{t\geq \sigma_k^{(n)} \mid W_S^{(n)}\big(\sigma_k^{(n)}-\big) \vee \max_{s\in[\sigma_k^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty,0] \geq W_S^{(n)}(t)\Big\}, \tag{4.11}$$

$$\sigma_{k+1}^{(n)} \triangleq \min\big\{t\geq \tau_k^{(n)} \mid W_S^{(n)}(t) > W_S^{(n)}(t-)\big\}. \tag{4.12}$$

In addition, for $t\in[0,\infty)$, define

$$K_+^{(n)}(t) \triangleq \sum_{k\in\mathbb{N}}\Big[W_S^{(n)}\big(\sigma_k^{(n)}-\big) \vee \max_{s\in[\sigma_k^{(n)},t\wedge\tau_k^{(n)}]} \mathcal{W}_S^{(n)}(s)(-\infty,0] - W_S^{(n)}\big(\sigma_k^{(n)}-\big)\Big], \tag{4.13}$$

$$K_-^{(n)}(t) \triangleq -\sum_{k\in\mathbb{N}}\big[\big(W_S^{(n)}(\tau_{k-1}^{(n)}) - \big(\sigma_k^{(n)}\wedge t-\tau_{k-1}^{(n)}\big)\big)^+ - W_S^{(n)}(\tau_{k-1}^{(n)})\big]. \tag{4.14}$$

THEOREM 4.2. *We have*

$$K^{(n)} = K_+^{(n)} - K_-^{(n)}, \tag{4.15}$$

where $K_+^{(n)}$ and $K_-^{(n)}$ are the positive and negative variations of $K^{(n)}$. Moreover,

$$\int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}}\, dK_-^{(n)}(s) = 0. \tag{4.16}$$

The theorem is easily deduced from Propositions 4.3 and 4.4 and Remark 4.5 below. The rest of the section is devoted to establishing these results.

Observe that the late work $\mathcal{W}_S^{(n)}(s)(-\infty, 0]$ is right-continuous in s, remaining constant or moving down at rate one and jumping up. Therefore, the maximum on the right-hand side of (4.11) is obtained. Additionally, because of the right-continuity of $\mathcal{W}_S^{(n)}$ and $W_S^{(n)}$, the minimum in this equation is also obtained. Finally, $\mathcal{W}_S^{(n)}(s)(-\infty, 0]$ can never exceed $W_S^{(n)}(s) = \mathcal{W}_S^{(n)}(s)(\mathbb{R})$, and $W_S^{(n)}$ never jumps down, so we must in fact have

$$W_S^{(n)}\bigl(\sigma_k^{(n)}-\bigr) \vee \max_{s\in[\sigma_k^{(n)},\tau_k^{(n)}]} \mathcal{W}_S^{(n)}(s)(-\infty, 0] = W_S^{(n)}\bigl(\tau_k^{(n)}\bigr). \tag{4.17}$$

For $k \geq 1$, $\sigma_k^{(n)}$ is the first arrival time after $\tau_{k-1}^{(n)}$. We thus have

$$W_S^{(n)}(t) = \bigl(W_S^{(n)}\bigl(\tau_{k-1}^{(n)}\bigr) - \bigl(t - \tau_{k-1}^{(n)}\bigr)\bigr)^+, \qquad \tau_{k-1}^{(n)} \leq t < \sigma_k^{(n)}. \tag{4.18}$$

We further have

$$0 = \sigma_0^{(n)} = \tau_0^{(n)} < \sigma_1^{(n)} < \tau_1^{(n)} < \sigma_2^{(n)} < \cdots. \tag{4.19}$$

PROPOSITION 4.3. *For each $k \geq 1$, we have*

$$K^{(n)}(t) = W_S^{(n)}\bigl(\sigma_k^{(n)}-\bigr) \vee \max_{s\in[\sigma_k^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0] \tag{4.20}$$

for $t \in [\sigma_k^{(n)}, \tau_k^{(n)}]$. In particular, $K^{(n)}$ is nondecreasing on the interval $[\sigma_k^{(n)}, \tau_k^{(n)}]$.

PROOF. We proceed by induction on k. For the base case $k = 1$, note that the standard EDF system is empty before the time $\sigma_1^{(n)}$. Therefore, $W_S^{(n)}(\sigma_1^{(n)}-) = 0$, and to prove (4.20), we must show that

$$K^{(n)}(t) = \max_{s\in[0,t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0], \qquad \sigma_1^{(n)} \leq t \leq \tau_1^{(n)}. \tag{4.21}$$

For $t \in [\sigma_1^{(n)}, \tau_1^{(n)}]$, let $s^{(n)}(t)$ be the largest number in $[\sigma_1^{(n)}, t]$ satisfying

$$\mathcal{W}_S^{(n)}\bigl(s^{(n)}(t)\bigr)(-\infty, 0] = \max_{s\in[0,t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0]. \tag{4.22}$$

For $u \in [s^{(n)}(t), t]$, we have

$$\mathcal{W}_S^{(n)}\bigl(s^{(n)}(t)\bigr)(-\infty, 0] = \max_{s\in[\sigma_1^{(n)},u]} \mathcal{W}_S^{(n)}(s)(-\infty, 0],$$

which is less than or equal to $W_S^{(n)}(u)$ by the definition of $\tau_1^{(n)}$ and equation (4.17). Therefore,

$$\max_{s\in[0,t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0] = \mathcal{W}_S^{(n)}\bigl(s^{(n)}(t)\bigr)(-\infty, 0] \leq \inf_{u\in[s^{(n)}(t),t]} W_S^{(n)}(u).$$

Equation (4.21) follows from (4.7).

We assume (4.20) holds for some k and prove it for $k+1$. For $t \in [\sigma_{k+1}^{(n)}, \tau_{k+1}^{(n)}]$,

$$\begin{aligned} K^{(n)}(t) = {} & \max_{s\in[0,\sigma_{k+1}^{(n)})} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} \\ & \vee \max_{s\in[\sigma_{k+1}^{(n)},t]} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\}. \end{aligned} \tag{4.23}$$

Equation (4.20) with k replaced by $k+1$ will follow once we show that

$$\max_{s\in[0,\sigma_{k+1}^{(n)})} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} = W_S^{(n)}\big(\sigma_{k+1}^{(n)}-\big) \tag{4.24}$$

and

$$\begin{aligned} & \max_{s\in[\sigma_{k+1}^{(n)},t]} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} \\ & \qquad = \max_{s\in[\sigma_{k+1}^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty,0]. \end{aligned} \tag{4.25}$$

For (4.24), we observe that because $\mathcal{W}_S^{(n)}(s)(-\infty,0]$ and $\inf_{s\le u\le t} W_S^{(n)}(u)$, regarded as functions of s, cannot increase except by a jump, the maximum on the left-hand side of (4.24) is attained. Let $s_k^{(n)}$ be the largest number in $[0,\sigma_{k+1}^{(n)})$ attaining this maximum. We have

$$\begin{aligned} & \max_{s\in[0,\sigma_{k+1}^{(n)})} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} \\ & \quad = \mathcal{W}_S^{(n)}\big(s_k^{(n)}\big)(-\infty,0] \wedge \inf_{u\in[s_k^{(n)},t]} W_S^{(n)}(u) \le W_S^{(n)}(u) \qquad \forall u \in \big[s_k^{(n)}, \sigma_{k+1}^{(n)}\big), \end{aligned}$$

and so

$$\max_{s\in[0,\sigma_{k+1}^{(n)})} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} \le W_S^{(n)}\big(\sigma_{k+1}^{(n)}-\big). \tag{4.26}$$

On the other hand, by the inequalities $\tau_k^{(n)} < \sigma_{k+1}^{(n)} \le t \le \tau_{k+1}^{(n)}$, definition (4.7), the induction hypothesis, and equation (4.17), we have

$$\begin{aligned} & \max_{s\in[0,\sigma_{k+1}^{(n)})} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \Big\} \\ & \quad \ge \max_{s\in[0,\tau_k^{(n)}]} \Big\{ \mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,\tau_k^{(n)}]} W_S^{(n)}(u) \wedge \inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u) \Big\} \\ & \quad = K^{(n)}\big(\tau_k^{(n)}\big) \wedge \inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u) \end{aligned}$$

$$= \Big(W_S^{(n)}(\sigma_k^{(n)}-) \vee \max_{s\in[\sigma_k^{(n)},\tau_k^{(n)}]} \mathcal{W}_S^{(n)}(s)(-\infty,0]\Big) \wedge \inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u)$$

$$= W_S^{(n)}(\tau_k^{(n)}) \wedge \inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u) = \inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u).$$

Equation (4.18) implies $W_S^{(n)}(u) \geq W_S^{(n)}(\sigma_{k+1}^{(n)}-)$ for $\tau_k^{(n)} \leq u < \sigma_{k+1}^{(n)}$. For $\sigma_{k+1}^{(n)} \leq u \leq t < \tau_{k+1}^{(n)}$, (4.11) implies that

$$W_S^{(n)}(\sigma_{k+1}^{(n)}-) \vee \max_{s\in[\sigma_{k+1}^{(n)},u]} \mathcal{W}_S^{(n)}(s)(-\infty,0] \leq W_S^{(n)}(u),$$

and so again we have $W_S^{(n)}(u) \geq W_S^{(n)}(\sigma_{k+1}^{(n)}-)$. Finally, if $u = t = \tau_{k+1}^{(n)}$, then (4.17) implies that $W_S^{(n)}(u) \geq W_S^{(n)}(\sigma_{k+1}^{(n)}-)$. It follows that

$$\inf_{u\in[\tau_k^{(n)},t]} W_S^{(n)}(u) \geq W_S^{(n)}(\sigma_{k+1}^{(n)}-).$$

This gives the reverse of the inequality (4.26), and thus (4.24) is proved.

For (4.25), we let $t_k^{(n)}$ attain the maximum in $\max_{s\in[\sigma_{k+1}^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty,0]$. For $u \in [t_k^{(n)}, t]$, we have from (4.11) and (4.17) that

$$\mathcal{W}_S^{(n)}(t_k^{(n)})(-\infty,0] = \max_{s\in[\sigma_{k+1}^{(n)},u]} \mathcal{W}_S^{(n)}(s)(-\infty,0] \leq W_S^{(n)}(u),$$

and hence $\mathcal{W}_S^{(n)}(t_k^{(n)})(-\infty,0] \leq \inf_{u\in[t_k^{(n)},t]} W_S^{(n)}(u)$. It follows that

$$\max_{s\in[\sigma_{k+1}^{(n)},t]} \Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u)\Big\}$$

$$\leq \mathcal{W}_S^{(n)}(t_k^{(n)})(-\infty,0] = \mathcal{W}_S^{(n)}(t_k^{(n)})(-\infty,0] \wedge \inf_{u\in[t_k^{(n)},t]} W_S^{(n)}(u)$$

$$\leq \max_{s\in[\sigma_{k+1}^{(n)},t]} \Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u)\Big\},$$

which establishes (4.25). □

PROPOSITION 4.4. *For each $k \geq 1$, we have*

$$K^{(n)}(t) = \big(W_S^{(n)}(\tau_{k-1}^{(n)}) - (t-\tau_{k-1}^{(n)})\big)^+, \qquad \tau_{k-1}^{(n)} \leq t < \sigma_k^{(n)}. \tag{4.27}$$

In particular, $K^{(n)}$ is nonincreasing on $[\tau_{k-1}^{(n)}, \sigma_k^{(n)})$.

PROOF. For all $t \geq 0$, we have $K^{(n)}(t) \leq W_S^{(n)}(t)$, and for $\tau_{k-1}^{(n)} \leq t < \sigma_k^{(n)}$, we further have from (4.18) that

$$K^{(n)}(t) \leq W_S^{(n)}(t) = \big(W_S^{(n)}(\tau_{k-1}^{(n)}) - (t-\tau_{k-1}^{(n)})\big)^+. \tag{4.28}$$

On the other hand, Proposition 4.3 and (4.17) with k replaced by $k-1$ imply

$$\begin{aligned}
&\max_{s\in[0,\tau_{k-1}^{(n)}]}\Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0]\wedge\inf_{u\in[s,\tau_{k-1}^{(n)}]}W_S^{(n)}(u)\Big\}\\
&\quad=K^{(n)}(\tau_{k-1}^{(n)})=W_S^{(n)}(\sigma_{k-1}^{(n)}-)\vee\max_{s\in[\sigma_{k-1}^{(n)},\tau_{k-1}^{(n)}]}\mathcal{W}_S^{(n)}(s)(-\infty,0]\\
&\quad=W_S^{(n)}(\tau_{k-1}^{(n)}).
\end{aligned}$$

For $t\in[\tau_{k-1}^{(n)},\sigma_k^{(n)})$, it follows from (4.18) and the above equality that

$$\begin{aligned}
K^{(n)}(t)&=\max_{s\in[0,t]}\Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0]\wedge\inf_{u\in[s,t]}W_S^{(n)}(u)\Big\}\\
&\ge\max_{s\in[0,\tau_{k-1}^{(n)}]}\Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0]\\
&\qquad\qquad\wedge\inf_{u\in[s,\tau_{k-1}^{(n)}]}W_S^{(n)}(u)\wedge\inf_{u\in[\tau_{k-1}^{(n)},t]}W_S^{(n)}(u)\Big\}\\
&=\max_{s\in[0,\tau_{k-1}^{(n)}]}\Big\{\mathcal{W}_S^{(n)}(s)(-\infty,0]\wedge\inf_{u\in[s,\tau_{k-1}^{(n)}]}W_S^{(n)}(u)\Big\}\\
&\quad\wedge\big(W_S^{(n)}(\tau_{k-1}^{(n)})-(t-\tau_{k-1}^{(n)})\big)^+\\
&=W_S^{(n)}(\tau_{k-1}^{(n)})\wedge\big(W_S^{(n)}(\tau_{k-1}^{(n)})-(t-\tau_{k-1}^{(n)})\big)^+\\
&=\big(W_S^{(n)}(\tau_{k-1}^{(n)})-(t-\tau_{k-1}^{(n)})\big)^+.
\end{aligned}\tag{4.29}$$

Equation (4.27) follows from (4.28) and (4.29). □

REMARK 4.5. In light of (4.6) and Proposition 4.3, we have the characterization of $\tau_k^{(n)}$ as

$$\tau_k^{(n)}=\min\{t\ge\sigma_k^{(n)}|K^{(n)}(t)\ge W_S^{(n)}(t)\}=\min\{t\ge\sigma_k^{(n)}|U^{(n)}(t)=0\}.\tag{4.30}$$

Because $\sigma_{k+1}^{(n)}$ is the time of first arrival after $\tau_k^{(n)}$, we in fact have

$$U^{(n)}(t)=0,\qquad\tau_k^{(n)}\le t<\sigma_{k+1}^{(n)}.\tag{4.31}$$

Evaluating (4.20) at $\sigma_k^{(n)}$ and using $W_S^{(n)}(\sigma_k^{(n)}-)\ge\mathcal{W}_S^{(n)}(\sigma_k^{(n)})(-\infty,0]$, we obtain

$$K^{(n)}(\sigma_k^{(n)})=W_S^{(n)}(\sigma_k^{(n)}-).\tag{4.32}$$

But (4.18) and Proposition 4.4 show that

$$K^{(n)}(\sigma_k^{(n)}-)=W_S^{(n)}(\sigma_k^{(n)}-),\tag{4.33}$$

and so

$$\triangle K^{(n)}(\sigma_k^{(n)}) = 0. \tag{4.34}$$

By contrast $\triangle K^{(n)}(\tau_k^{(n)})$ can be positive. Evaluating (4.20) at $\tau_k^{(n)}$ and using (4.17), we obtain

$$K^{(n)}(\tau_k^{(n)}) = W_S^{(n)}(\tau_k^{(n)}). \tag{4.35}$$

In conclusion,

$$K^{(n)}(t) = K^{(n)}(\sigma_k^{(n)}) \vee \max_{s\in[\sigma_k^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0], \qquad \sigma_k^{(n)} \le t \le \tau_k^{(n)}, \tag{4.36}$$

$$K^{(n)}(t) = \big(K^{(n)}(\tau_{k-1}^{(n)}) - (t - \tau_{k-1}^{(n)})\big)^+, \qquad \tau_{k-1}^{(n)} \le t < \sigma_k^{(n)}. \tag{4.37}$$

4.2. *Dynamics of the reference workload process.* The evolutions of $\mathcal{U}^{(n)}$ and $\mathcal{W}^{(n)}$ are similar; the difference between them is asymptotically negligible. Before proving the properties of $\mathcal{U}^{(n)}$, we provide a summary of these properties. The reader may work out the evolution of $\mathcal{W}_S^{(n)}$, $\mathcal{U}^{(n)}$ and $\mathcal{W}^{(n)}$ in Example 4.6 to follow along. This example appears in detail in [26].

EXAMPLE 4.6. Consider a system realization in which

$$\begin{aligned}
&u_1^{(n)} = 1, \quad && v_1^{(n)} = 4, \quad && L_1^{(n)} = 3, \quad && S_1^{(n)} = 1,\\
&u_2^{(n)} = 1, && v_2^{(n)} = 4, && L_2^{(n)} = 5, && S_2^{(n)} = 2,\\
&u_3^{(n)} = 3, && v_3^{(n)} = 2, && L_3^{(n)} = 1, && S_3^{(n)} = 5,\\
&u_4^{(n)} = 2, && v_4^{(n)} = 1, && L_4^{(n)} = 4, && S_4^{(n)} = 7,\\
&u_5^{(n)} = 2, && v_5^{(n)} = 1, && L_5^{(n)} = 1, && S_5^{(n)} = 9.
\end{aligned}$$

Recall that δ_a is a unit point mass at a. It is straightforward to compute

$$\mathcal{W}_S^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1,\\ (5-t)\delta_{4-t}, & 1 \le t < 2,\\ (5-t)\delta_{4-t} + 4\delta_{7-t}, & 2 \le t < 5,\\ (7-t)\delta_{6-t} + 4\delta_{7-t}, & 5 \le t < 7,\\ (11-t)\delta_{7-t} + \delta_{11-t}, & 7 \le t < 9,\\ 2\delta_{-2} + \delta_1 + \delta_2, & t = 9, \end{cases}$$

$$\mathcal{U}^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1, \\ (5-t)\delta_{4-t}, & 1 \le t < 2, \\ (5-t)\delta_{4-t} + 4\delta_{7-t}, & 2 \le t < 4, \\ (8-t)\delta_{7-t}, & 4 \le t < 5, \\ (6-t)\delta_{6-t} + 4\delta_{7-t}, & 5 \le t < 6, \\ (10-t)\delta_{7-t}, & 6 \le t < 7, \\ (8-t)\delta_{11-t}, & 7 \le t < 8, \\ 0, & 8 \le t < 9, \\ \delta_2, & t = 9, \end{cases}$$

$$\mathcal{W}^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1, \\ (5-t)\delta_{4-t}, & 1 \le t < 2, \\ (5-t)\delta_{4-t} + 4\delta_{7-t}, & 2 \le t < 4, \\ (8-t)\delta_{7-t}, & 4 \le t < 5, \\ (7-t)\delta_{6-t} + 3\delta_{7-t}, & 5 \le t < 6, \\ (9-t)\delta_{7-t}, & 6 \le t < 7, \\ (8-t)\delta_{11-t}, & 7 \le t < 8, \\ 0, & 8 \le t < 9, \\ \delta_1, & t = 9. \end{cases}$$

Recall that $K^{(n)}$ is the amount of mass removed from the standard workload $W_S^{(n)}$ to obtain the reference workload $U^{(n)}$. To understand the process $K^{(n)}$, we consider the dynamics of $\mathcal{U}^{(n)}$. In the absence of new arrivals, all atoms of $\mathcal{U}^{(n)}$ move left with unit speed. Moreover, the mass of the leftmost atom of $\mathcal{U}^{(n)}$ decreases with unit speed until it vanishes, corresponding to the work being done on the most urgent job in queue until it is served to completion [Proposition 4.8(i)]. However, if the leftmost atom of $\mathcal{U}^{(n)}$ hits zero, this atom is immediately removed from $\mathcal{U}^{(n)}$ [Proposition 4.8(ii), (v)]. This may be interpreted as reneging of a customer or deletion of a late customer from the system. When there is a new arrival at time t with lead time not smaller than the leftmost point of support of $\mathcal{U}^{(n)}(t-)$, and this point of support is strictly positive, then a mass of the size $v^{(n)}_{A^{(n)}(t)}$ located at $L^{(n)}_{A^{(n)}(t)}$ is added to $\mathcal{U}^{(n)}(t-)$ [Proposition 4.8(iii)]. Similarly, if there is a new arrival and the leftmost point of the support of $\mathcal{U}^{(n)}$ hits zero at the same time, then both of the above actions take place [(4.53) of Proposition 4.8(v)]. This is the case of a simultaneous new arrival and ejection of a late customer. The EDF system with reneging $\mathcal{W}^{(n)}$ shows the same behavior in all these cases. However, if a customer arrives to start a new busy period for $\mathcal{U}^{(n)}$ or, if at time t, there is a new arrival with lead time more urgent than the leftmost point of the support of $\mathcal{U}^{(n)}(t-)$ (i.e., we have a "preemption"), then the mass $v^{(n)}_{A^{(n)}(t)}$ associated with the new arrival is distributed in $[L^{(n)}_{A^{(n)}(t)}, \infty)$, or more precisely, on some atoms of $\mathcal{W}^{(n)}_S(t)$ located

on this half-line, but it is not necessarily located at the single atom $L^{(n)}_{A^{(n)}(t)}$. This possibility is described in Lemma 4.7 and Proposition 4.8(iv). In this respect, the evolution of $\mathcal{U}^{(n)}$ differs from that of $\mathcal{W}^{(n)}$, for which all the new mass is always placed at the lead time of the arriving customer. Example 4.6 illustrates this.

We now begin the rigorous study of $\mathcal{U}^{(n)}$. As shown in Section 4.1, the time interval $[0,\infty)$ can be decomposed into a union of the disjoint intervals $(\tau^{(n)}_k, \sigma^{(n)}_{k+1}]$ and $(\sigma^{(n)}_k, \tau^{(n)}_k]$, $k \geq 0$, such that $K^{(n)} = W^{(n)}_S - U^{(n)}$ is nonincreasing on $(\tau^{(n)}_k, \sigma^{(n)}_{k+1}]$ and nondecreasing on $(\sigma^{(n)}_k, \tau^{(n)}_k]$. In Lemma 4.7 below, we analyze the behavior of $\mathcal{U}^{(n)}$ on the time intervals $[\tau^{(n)}_{k-1}, \sigma^{(n)}_k]$, $k \geq 1$, while Proposition 4.8 describes the dynamics of $\mathcal{U}^{(n)}$ on the intervals $(\sigma^{(n)}_k, \tau^{(n)}_k)$, $k \geq 1$. The section ends with Corollary 4.9, which describes the time evolution of the reference workload process $U^{(n)}$.

We make use of the following elementary facts about the standard workload. Since the interarrival times are strictly positive, $\triangle A^{(n)}(t) \in \{0, 1\}$, and

$$\mathcal{W}^{(n)}_S(t) = \mathcal{W}^{(n)}_S(t-) + \triangle A^{(n)}(t) v^{(n)}_{A^{(n)}(t)} \delta_{L^{(n)}_{A^{(n)}(t)}}, \qquad t \geq 0, \tag{4.38}$$

which implies

$$\triangle W^{(n)}_S(t) = \triangle A^{(n)}(t) v^{(n)}_{A^{(n)}(t)}, \qquad t \geq 0. \tag{4.39}$$

For any functions f and g on $[0,\infty)$ (taking finite or infinite values) such that whenever $s < t$ and $t - s$ is small enough, $f(s) = f(t-) + t - s$ and $g(s) = g(t-) + t - s$, we have

$$\lim_{s \uparrow t} \mathcal{W}^{(n)}_S(s)[f(s), g(s)] = \mathcal{W}^{(n)}_S(t-)[f(t-), g(t-)]. \tag{4.40}$$

This is true because the lead times of the customers present in the standard system decrease with unit rate. Equation (4.40) remains valid if the closed intervals $[f(\cdot), g(\cdot)]$ are replaced by either $[f(\cdot), g(\cdot))$, $(f(\cdot), g(\cdot)]$ or $(f(\cdot), g(\cdot))$. These facts will be used repeatedly in the following arguments, sometimes without mention.

LEMMA 4.7. *Let $k \geq 1$. We have*

$$U^{(n)}(t) = 0, \qquad \tau^{(n)}_{k-1} \leq t < \sigma^{(n)}_k, \tag{4.41}$$

$$\triangle U^{(n)}(\sigma^{(n)}_k) = v^{(n)}_{A^{(n)}(\sigma^{(n)}_k)}, \tag{4.42}$$

$$\mathcal{U}^{(n)}(\sigma^{(n)}_k)(-\infty, L^{(n)}_{A^{(n)}(\sigma^{(n)}_k)}) = 0. \tag{4.43}$$

PROOF. Equation (4.41) follows immediately from (4.6), (4.18) and Proposition 4.4. By (4.6), (4.34), (4.39) and the fact that $\triangle A^{(n)}(\sigma^{(n)}_k) = 1$, we have

$$\triangle U^{(n)}(\sigma^{(n)}_k) = \triangle W^{(n)}_S(\sigma^{(n)}_k) - \triangle K^{(n)}(\sigma^{(n)}_k) = v^{(n)}_{A^{(n)}(\sigma^{(n)}_k)},$$

and (4.42) follows. For $y < L^{(n)}_{A^{(n)}(\sigma_k^{(n)})}$, (4.5), (4.38), (4.34) and (4.33) imply

$$\begin{aligned}\mathcal{U}^{(n)}(\sigma_k^{(n)})(-\infty, y] &= [\mathcal{W}_S^{(n)}(\sigma_k^{(n)})(-\infty, y] - K^{(n)}(\sigma_k^{(n)})]^+ \\ &= [\mathcal{W}_S^{(n)}(\sigma_k^{(n)}-)(-\infty, y] - K^{(n)}(\sigma_k^{(n)}-)]^+ \\ &\le [W_S^{(n)}(\sigma_k^{(n)}-) - K^{(n)}(\sigma_k^{(n)}-)]^+ = 0,\end{aligned}$$

and so (4.43) also follows. □

Lemma 4.7 shows that $\sigma_k^{(n)}$ begins a busy period for the reference system. Equation (4.30) implies that $U^{(n)}(t) > 0$ for $\sigma_k^{(n)} < t < \tau_k^{(n)}$, and thus the intervals $[\sigma_k^{(n)}, \tau_k^{(n)})$, $k \ge 1$, are precisely the busy periods for the reference system. We analyze the behavior of $\mathcal{U}^{(n)}$ during these busy periods. We start with the observation that, by (4.5) and Proposition 4.3, for $t \in (\sigma_k^{(n)}, \tau_k^{(n)})$,

$$\begin{aligned}\mathcal{U}^{(n)}(t)(-\infty, y] = \Big[&\mathcal{W}_S^{(n)}(t)(-\infty, y] \\ &- \Big(W_S^{(n)}(\sigma_k^{(n)}-) \vee \max_{s\in[\sigma_k^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0]\Big)\Big]^+.\end{aligned} \tag{4.44}$$

In what follows, given $\nu \in \mathcal{M}$ and any interval $\mathcal{I} \subset \mathbb{R}$, we will use $\nu_{|\mathcal{I}}$ to denote the measure in $\mathcal{M}$ that is zero on $\mathcal{I}^c$ and coincides with ν on $\mathcal{I}$: $\nu_{|\mathcal{I}}(B) = \nu(B \cap \mathcal{I})$ for all $B \in \mathcal{B}(\mathbb{R})$.

PROPOSITION 4.8. *For $k \ge 1$ and $\sigma_k^{(n)} < t < \tau_k^{(n)}$, the following five properties hold*:

(i) *If $\triangle A^{(n)}(t) = 0$ and $E^{(n)}(t-) > 0$, then*

$$\triangle K^{(n)}(t) = 0, \tag{4.45}$$

$$\triangle U^{(n)}(t) = 0. \tag{4.46}$$

In this case, if $\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} > 0$, then both $\mathcal{U}^{(n)}(\cdot)\{E^{(n)}(\cdot)\}$ and $U^{(n)}(t)$ decrease with unit rate in a neighborhood of t. On the other hand, if

$$\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} = 0,$$

then $\mathcal{U}^{(n)}(t) = \mathcal{W}_S^{(n)}(t)|_{[E^{(n)}(t),\infty)}$.

(ii) *If $\triangle A^{(n)}(t) = 0$ and $E^{(n)}(t-) = 0$, then*

$$\mathcal{U}^{(n)}(t-)\{0\} = \triangle K^{(n)}(t) = -\triangle U^{(n)}(t) \tag{4.47}$$

and $\mathcal{U}^{(n)}(t) = \mathcal{W}_S^{(n)}(t)|_{(0,\infty)}$.

(iii) *If* $\triangle A^{(n)}(t) = 1$ *and* $L^{(n)}_{A^{(n)}(t)} \geq E^{(n)}(t-) > 0$, *then* (4.45) *holds*, $\triangle E^{(n)}(t) \geq 0$ *and*

$$\mathcal{U}^{(n)}(t) = \mathcal{U}^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)} \delta_{L^{(n)}_{A^{(n)}(t)}}. \tag{4.48}$$

(iv) *If* $\triangle A^{(n)}(t) = 1$, $L^{(n)}_{A^{(n)}(t)} < E^{(n)}(t-)$, *then* (4.45) *holds and*

$$L^{(n)}_{A^{(n)}(t)} \leq E^{(n)}(t) \leq E^{(n)}(t-), \tag{4.49}$$

$$\triangle U^{(n)}(t) = v^{(n)}_{A^{(n)}(t)}, \tag{4.50}$$

$$\mathcal{U}^{(n)}(t)_{|(E^{(n)}(t-),\infty)} = \mathcal{U}^{(n)}(t-)_{|(E^{(n)}(t-),\infty)}, \tag{4.51}$$

$$\mathcal{U}^{(n)}(t)\{E^{(n)}(t-)\} \geq \mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\}. \tag{4.52}$$

(v) *If* $\triangle A^{(n)}(t) = 1$ *and* $L^{(n)}_{A^{(n)}(t)} > E^{(n)}(t-) = 0$, *then*

$$\mathcal{U}^{(n)}(t) = \mathcal{U}^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)} \delta_{L^{(n)}_{A^{(n)}(t)}} - \mathcal{U}^{(n)}(t-)\{0\}\delta_0. \tag{4.53}$$

PROOF. Fix $k \geq 1$ and $t \in (\sigma^{(n)}_k, \tau^{(n)}_k)$. We start with the general observation that, by (4.4) and (4.44),

$$\begin{aligned} E^{(n)}(t) = \min\Big\{ y \big| \mathcal{W}^{(n)}_S(t)(-\infty, y] \\ > W^{(n)}_S(\sigma^{(n)}_k-) \vee \max_{s \in [\sigma^{(n)}_k, t]} \mathcal{W}^{(n)}_S(s)(-\infty, 0] \Big\}, \end{aligned} \tag{4.54}$$

and because $\mathcal{W}^{(n)}_S(t)$ is purely atomic, the minimum on the right-hand side of (4.54) is obtained at the atom of $\mathcal{W}^{(n)}_S(t)$ located at $y_0 = E^{(n)}(t)$. In particular,

$$\mathcal{W}^{(n)}_S(t)\{E^{(n)}(t)\} > 0. \tag{4.55}$$

We now consider each of the five different cases of the proposition.

(i) Let $a = E^{(n)}(t-)$. By (4.4) and (4.5), for all $s < t$ sufficiently near t,

$$\mathcal{W}^{(n)}_S(s)(-\infty, a/2] \leq K^{(n)}(s). \tag{4.56}$$

Also, for $s \in [t - a/2, t)$ sufficiently near t so that $A^{(n)}(s) = A^{(n)}(t)$ holds [such s exist due to the assumption that $\triangle A^{(n)}(t) = 0$], we have

$$\mathcal{W}^{(n)}_S(t)(-\infty, 0] \leq \mathcal{W}^{(n)}_S(s)(-\infty, a/2]. \tag{4.57}$$

The last two relations show that $\mathcal{W}^{(n)}_S(t)(-\infty, 0] \leq K^{(n)}(t-)$, and so, by Proposition 4.3, (4.45) holds. Equation (4.46) follows from (4.6), (4.45), (4.39) and the assumption $\triangle A^{(n)}(t) = 0$. Because $W^{(n)}_S(t) > 0$ [see (4.55)], $W^{(n)}_S$ decreases at

unit rate in a neighborhood of t [see (2.6)–(2.8)]. In addition, (4.6), (4.45) and the fact that, again by Proposition 4.3, $K^{(n)}$ cannot increase on $[\sigma_k^{(n)}, \tau_k^{(n)}]$ except by a jump and hence is constant in a neighborhood of t, together imply that $U^{(n)}$ also decreases at unit rate in a neighborhood of t. Furthermore, the nature of the EDF discipline and (4.55) show that at t, the standard system is serving a customer with lead time no greater than $E^{(n)}(t)$. Combining the above properties with the fact that $\mathcal{U}^{(n)}(t)_{|(E^{(n)}(t),\infty)} = \mathcal{W}_S^{(n)}(t)_{|(E^{(n)}(t),\infty)}$ by (4.9), we conclude that if $\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} > 0$, then $\mathcal{U}^{(n)}(\cdot)\{E^{(n)}(\cdot)\}$ decreases with unit rate in a neighborhood of t. On the other hand, if $\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} = 0$, then since $\Delta A^{(n)}(t) = 0$, $E^{(n)}$ jumps up at t. Indeed, in this case,

$$\mathcal{W}_S^{(n)}(t)\big(-\infty, E^{(n)}(t-)\big] = \mathcal{W}_S^{(n)}(t-)\big(-\infty, E^{(n)}(t-)\big] = K^{(n)}(t-) = K^{(n)}(t).$$

This means that

$$\begin{aligned} E^{(n)}(t) &= \min\big\{y \in \mathbb{R} | \mathcal{W}_S^{(n)}(t)(-\infty, y] > K^{(n)}(t)\big\} \\ &= \min\big\{y > E^{(n)}(t-) | \mathcal{W}_S^{(n)}(t)\{y\} > 0\big\}. \end{aligned}$$

It follows that

$$\mathcal{W}_S^{(n)}(t)\big(E^{(n)}(t-), E^{(n)}(t)\big) = 0. \tag{4.58}$$

Using the definition of $E^{(n)}(t)$, (4.46), (4.9), the assumption $\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} = 0$, the assumption $\triangle A^{(n)}(t) = 0$, and (4.58), we obtain

$$\begin{aligned} \mathcal{U}^{(n)}(t)\big[E^{(n)}(t), \infty\big) &= U^{(n)}(t) = U^{(n)}(t-) \\ &= \mathcal{U}^{(n)}(t-)\big[E^{(n)}(t-), \infty\big) \\ &= \mathcal{U}^{(n)}(t-)\big(E^{(n)}(t-), \infty\big) \\ &= \mathcal{W}_S^{(n)}(t-)\big(E^{(n)}(t-), \infty\big) \\ &= \mathcal{W}_S^{(n)}(t)\big(E^{(n)}(t-), \infty\big) \\ &= \mathcal{W}_S^{(n)}(t)\big[E^{(n)}(t), \infty\big). \end{aligned}$$

From (4.9) we see now that $\mathcal{U}^{(n)}(t) = \mathcal{W}_S^{(n)}(t)_{|[E^{(n)}(t),\infty)}$.

(ii) By (4.30), (4.4) and (4.5), for $s \in (\sigma_k^{(n)}, t)$ we have

$$\mathcal{W}_S^{(n)}(s)\big(-\infty, E^{(n)}(s)\big] > K^{(n)}(s). \tag{4.59}$$

As $s \uparrow t$ in (4.59), by (4.40), (4.38), and the case (ii) assumptions $\triangle A^{(n)}(t) = 0$ and $E^{(n)}(t-) = 0$, we get

$$\mathcal{W}_S^{(n)}(t)(-\infty, 0] = \mathcal{W}_S^{(n)}(t-)(-\infty, 0] \geq K^{(n)}(t-). \tag{4.60}$$

When combined with Proposition 4.3, this implies

$$K^{(n)}(t) = K^{(n)}(t-) \vee \mathcal{W}_S^{(n)}(t)(-\infty, 0] = \mathcal{W}_S^{(n)}(t)(-\infty, 0]. \tag{4.61}$$

By (4.4) and (4.5), for $s \in (\sigma_k^{(n)}, t)$,

$$\mathcal{U}^{(n)}(s)\{E^{(n)}(s)\} = \mathcal{W}_S^{(n)}(s)\big(-\infty, E^{(n)}(s)\big] - K^{(n)}(s).$$

Letting $s \uparrow t$, invoking (4.40), (4.60) and (4.61), and recalling the assumption $E^{(n)}(t-) = 0$, we obtain

$$\mathcal{U}^{(n)}(t-)\{0\} = \mathcal{W}_S^{(n)}(t-)(-\infty, 0] - K^{(n)}(t-) = K^{(n)}(t) - K^{(n)}(t-), \tag{4.62}$$

and the first equality in (4.47) follows. The second equality in (4.47) follows from (4.6), (4.39) and the assumption $\triangle A^{(n)}(t) = 0$. Moreover, by (4.5) and (4.61), for every $y \in \mathbb{R}$,

$$\begin{aligned} \mathcal{U}^{(n)}(t)(-\infty, y] &= \big[\mathcal{W}_S^{(n)}(t)(-\infty, y] - \mathcal{W}_S^{(n)}(t)(-\infty, 0]\big]^+ \\ &= \mathcal{W}_S^{(n)}(t)_{|(0,\infty)}(-\infty, y]. \end{aligned} \tag{4.63}$$

(iii) Let $a = E^{(n)}(t-)$. We can deduce (4.45) from (4.56) and (4.57) as in (i), with the only difference that now (4.57), for $s < t$ sufficiently close to t such that $A^{(n)}(t-) = A^{(n)}(s)$, follows from the fact that $L^{(n)}_{A^{(n)}(t)} > 0$, since this implies that the work for the system associated with the customer arriving to the system at time t does not contribute to $\mathcal{W}_S^{(n)}(t)(-\infty, 0]$. Next, let $y < a$, let $\varepsilon = (a - y)/2$ and note that by assumption, $L^{(n)}_{A^{(n)}(t)} \geq a > y + \varepsilon$. Thus, for $s < t$, s sufficiently close to t [so as to ensure that $A^{(n)}(t-) = A^{(n)}(s)$], we have $\mathcal{W}_S^{(n)}(t)(-\infty, y] \leq \mathcal{W}_S^{(n)}(s)(-\infty, y + \varepsilon] \leq K^{(n)}(s)$, where the last inequality uses (4.5) and the fact that $y + \varepsilon < E^{(n)}(t-)$. Letting $s \uparrow t$, we obtain $\mathcal{W}_S^{(n)}(t)(-\infty, y] \leq K^{(n)}(t-)$, which, together with (4.45), shows that $y < E^{(n)}(t)$. Thus, $E^{(n)}(t-) \leq E^{(n)}(t)$ or, equivalently, $\triangle E^{(n)}(t) \geq 0$.

We now turn to the proof of (4.48). Equation (4.5) implies

$$\mathcal{U}^{(n)}(t-)(-\infty, y] = \big[\mathcal{W}_S^{(n)}(t-)(-\infty, y] - K^{(n)}(t-)\big]^+. \tag{4.64}$$

Indeed, for any y such that $\mathcal{W}_S^{(n)}(t-)\{y\} = 0$, (4.64) follows from (4.5), in which t is replaced by $s < t$, by taking $s \uparrow t$. However, the family of sets $(-\infty, y]$ with $\mathcal{W}_S^{(n)}(t-)\{y\} = 0$ forms a separating class in $\mathcal{B}(\mathbb{R})$, and so (4.64) holds for all y. Moreover, using (4.38), (4.45) and (4.5), we see that

$$\begin{aligned} &\mathcal{U}^{(n)}(t)(-\infty, y] \\ &\quad = \big[\mathcal{W}_S^{(n)}(t-)(-\infty, y] - K^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)} \delta_{L^{(n)}_{A^{(n)}(t)}(t)}(-\infty, y]\big]^+. \end{aligned} \tag{4.65}$$

When combined with (4.64), this shows that

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}^{(n)}(t-)(-\infty, y], \qquad y < L^{(n)}_{A^{(n)}(t)}(t). \tag{4.66}$$

On the other hand, if $y \geq L^{(n)}_{A^{(n)}(t)}(t)$, then $y \geq E^{(n)}(t-)$ and (4.64) becomes

$$\mathcal{U}^{(n)}(t-)(-\infty, y] = \mathcal{W}^{(n)}_S(t-)(-\infty, y] - K^{(n)}(t-).$$

From (4.65), we now have

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}^{(n)}(t-)(-\infty, y] + v^{(n)}_{A^{(n)}(t)}, \qquad y \geq L^{(n)}_{A^{(n)}(t)}. \tag{4.67}$$

When combined, (4.66) and (4.67) prove (4.48).

(iv) We have $L^{(n)}_{A^{(n)}(t)} > 0$, and so (4.45) holds by the same argument as in case (iii), but now with $a = L^{(n)}_{A^{(n)}(t)}$. The assumptions $L^{(n)}_{A^{(n)}(t)} < E^{(n)}(t-)$ and $\triangle A^{(n)}(t) = 1$, along with the relations (4.38), (4.5), (4.45) and the definition of $E^{(n)}$, imply that

$$\begin{aligned}
\mathcal{W}^{(n)}_S(t)\big(-\infty, E^{(n)}(t-)\big] &= \mathcal{W}^{(n)}_S(t-)\big(-\infty, E^{(n)}(t-)\big] + v^{(n)}_{A^{(n)}(t)} \\
&> \mathcal{W}^{(n)}_S(t-)\big(-\infty, E^{(n)}(t-)\big] \\
&\geq K^{(n)}(t-) \\
&= K^{(n)}(t).
\end{aligned}$$

Invoking (4.5) again, this shows that $\mathcal{U}^{(n)}(t)(-\infty, E^{(n)}(t-)] > 0$, which implies $E^{(n)}(t) \leq E^{(n)}(t-)$. Now, let $y < a = L^{(n)}_{A^{(n)}(t)}$ and let $\varepsilon = (a - y)/2$. Then, combining (4.38), the inequalities $y + \varepsilon < a < E^{(n)}(t-)$ and (4.45), we obtain

$$\mathcal{W}^{(n)}_S(t)(-\infty, y] \leq \mathcal{W}^{(n)}_S(t-)(-\infty, y + \varepsilon] \leq K^{(n)}(t-) = K^{(n)}(t).$$

This shows that $y < E^{(n)}(t)$, which proves (4.49). In addition, by (4.6) and (4.45), we have

$$\begin{aligned}
U^{(n)}(t) &= W^{(n)}_S(t) - K^{(n)}(t) \\
&= W^{(n)}_S(t-) + v^{(n)}_{A^{(n)}(t)} - K^{(n)}(t-) \\
&= U^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)},
\end{aligned}$$

and (4.50) follows. Furthermore, since $E^{(n)}(t) \leq E^{(n)}(t-)$ by (4.49), the relations (4.9), (4.38) and the assumption $L^{(n)}_{A^{(n)}(t)} < E^{(n)}(t-)$ imply

$$\begin{aligned}
\mathcal{U}^{(n)}(t)_{|(E^{(n)}(t-),\infty)} &= \mathcal{W}^{(n)}_S(t)_{|(E^{(n)}(t-),\infty)} \\
&= \mathcal{W}^{(n)}_S(t-)_{|(E^{(n)}(t-),\infty)} \\
&= \mathcal{U}^{(n)}(t-)_{|(E^{(n)}(t-),\infty)}.
\end{aligned}$$

This establishes (4.51).

Finally, to prove (4.52), we will consider two cases.

Case I. $E^{(n)}(t) < E^{(n)}(t-)$.

By (4.9), we know that

$$\mathcal{U}^{(n)}(t)\{E^{(n)}(t-)\} = \mathcal{W}_S^{(n)}(t)\{E^{(n)}(t-)\}.$$

In turn, when combined with (4.64) and the definition of $E^{(n)}$, this shows that

$$\begin{aligned}
&\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\}\\
&\quad = \mathcal{U}^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr] - \mathcal{U}^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr)\\
&\quad = \mathcal{W}_S^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr] - K^{(n)}(t-)\\
&\qquad - \bigl[\mathcal{W}_S^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr) - K^{(n)}(t-)\bigr]^+\\
&\quad \le \mathcal{W}_S^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr] - \mathcal{W}_S^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr)\\
&\quad = \mathcal{W}_S^{(n)}(t)\{E^{(n)}(t-)\}\\
&\quad = \mathcal{U}^{(n)}(t)\{E^{(n)}(t-)\},
\end{aligned}$$

and so (4.52) holds.

Case II. $E^{(n)}(t) = E^{(n)}(t-)$.

By (4.5), (4.38), (4.45), (4.64) and the definition of $E^{(n)}$,

$$\begin{aligned}
\mathcal{U}^{(n)}(t)\{E^{(n)}(t)\} &= \mathcal{U}^{(n)}(t)\bigl(-\infty, E^{(n)}(t)\bigr]\\
&= \mathcal{W}_S^{(n)}(t)\bigl(-\infty, E^{(n)}(t)\bigr] - K^{(n)}(t)\\
&= \mathcal{W}_S^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr] + v^{(n)}_{A^{(n)}(t)} - K^{(n)}(t-)\\
&= \mathcal{U}^{(n)}(t-)\bigl(-\infty, E^{(n)}(t-)\bigr] + v^{(n)}_{A^{(n)}(t)}\\
&= \mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} + v^{(n)}_{A^{(n)}(t)},
\end{aligned}$$

which establishes (4.52) in this case as well. Since $E^{(n)}(t) \le E^{(n)}(t-)$, the two cases above are exhaustive, and so (4.52) is proved.

(v) Equation (4.63) holds by the same argument as in (ii), but where now the equality in (4.60) follows from the fact that $L^{(n)}_{A^{(n)}(t)} > 0$. Let $\mathcal{U}_1^{(n)}(t) \triangleq \mathcal{U}^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)}\delta_{L^{(n)}_{A^{(n)}(t)}} - \mathcal{U}^{(n)}(t-)\{0\}\delta_0$. We want to show that $\mathcal{U}^{(n)}(t) = \mathcal{U}_1^{(n)}(t)$. By (4.10), $\mathcal{U}^{(n)}(t)$ and $\mathcal{U}^{(n)}(t-)$ are supported on $(0,\infty)$ and $[0,\infty)$, respectively. Thus,

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}_1^{(n)}(t)(-\infty, y] = 0, \qquad y \le 0. \tag{4.68}$$

By (4.9) and the fact that $E^{(n)}(t-)=0$, we have $\mathcal{U}^{(n)}(t-)_{|(0,\infty)}=\mathcal{W}_S^{(n)}(t-)_{|(0,\infty)}$. The last two statements, along with (4.38), (4.63) and another application of (4.9), show that

$$\begin{aligned}\mathcal{U}_1^{(n)}(t)_{|(0,\infty)} &= \mathcal{U}^{(n)}(t-)_{|(0,\infty)} + v^{(n)}_{A^{(n)}(t)}\delta_{L^{(n)}_{A^{(n)}(t)}}\\ &= \mathcal{W}_S^{(n)}(t-)_{|(0,\infty)} + v^{(n)}_{A^{(n)}(t)}\delta_{L^{(n)}_{A^{(n)}(t)}}\\ &= \mathcal{W}_S^{(n)}(t)_{|(0,\infty)}\\ &= \mathcal{U}^{(n)}(t)_{|(0,\infty)}.\end{aligned}$$

This, together with (4.68), shows that $\mathcal{U}^{(n)}(t)=\mathcal{U}_1^{(n)}(t)$. □

The last result of this section concerns the evolution of $U^{(n)}$. Despite the different ways in which arriving mass is distributed in the system with reneging and the reference system, in both systems one can keep track of the total mass in system by beginning with the arrived mass (which is the same in both systems), subtracting the reduction in mass due to service (which occurs continuously at unit rate per unit time whenever mass is present), and subtracting the mass that has become late and been deleted. In particular, a simple mass balance shows that

$$W^{(n)}(t)=V^{(n)}\big(A^{(n)}(t)\big)-\int_0^t \mathbb{I}_{\{W^{(n)}(s)>0\}}\,ds - R_W^{(n)}(t), \tag{4.69}$$

where we recall that $R_W^{(n)}$ is the total amount of reneged work in the reneging system, which admits the representation (2.18), $R_W^{(n)}(t)=\sum_{0<s\le t}\mathcal{W}^{(n)}(s-)\{0\}$, for all $t\in[0,\infty)$. We now show that the following analogous relation holds for the reference workload:

$$U^{(n)}(t)=V^{(n)}\big(A^{(n)}(t)\big)-\int_0^t \mathbb{I}_{\{U^{(n)}(s)>0\}}\,ds - R_U^{(n)}(t), \tag{4.70}$$

where

$$R_U^{(n)}(t)\triangleq\sum_{0<s\le t}\mathcal{U}^{(n)}(s-)\{0\}. \tag{4.71}$$

Also, for notational convenience, we set $R_W^{(n)}(0-)=R_U^{(n)}(0-)=0$.

COROLLARY 4.9. *For every $t\ge 0$, equation (4.70) holds. Moreover, $R_U^{(n)}=K_+^{(n)}$ and hence*

$$U^{(n)}=N^{(n)}+I_U^{(n)}-K_+^{(n)}, \tag{4.72}$$

where, for $t\ge 0$,

$$I_U^{(n)}(t)\triangleq\int_0^t \mathbb{I}_{\{U^{(n)}(s)=0\}}\,ds. \tag{4.73}$$

PROOF. For $t \geq 0$, let $\widetilde{U}^{(n)}(t)$ be equal to the right-hand side of (4.70). By (4.41) of Lemma 4.7, we have $U^{(n)}(0) = 0 = \widetilde{U}^{(n)}(0)$. Moreover, for every $k \geq 1$, by Lemma 4.7 and the definition of $\sigma_k^{(n)}$, it follows that $U^{(n)}(t-) = U^{(n)}(t) = 0$ and $\triangle V^{(n)}(A^{(n)}(t)) = 0$ for $t \in (\tau_{k-1}^{(n)}, \sigma_k^{(n)})$, $U^{(n)}(\sigma_k^{(n)}-) = 0$ and $\triangle U^{(n)}(\sigma_k^{(n)}) = \triangle V^{(n)}(A^{(n)}(\sigma_k^n))$. When compared with the right-hand side of (4.70), this shows that $U^{(n)}$ and $\widetilde{U}^{(n)}$ are both flat on $(\tau_{k-1}^{(n)}, \sigma_k^{(n)})$, with an upward jump at $\sigma_k^{(n)}$ of size $\triangle V^{(n)}(A^{(n)}(\sigma_k^{(n)}))$. Thus, to prove the corollary, it suffices to show that the increments of $\widetilde{U}^{(n)}$ and $U^{(n)}$ on the intervals $(\sigma_k^{(n)}, \tau_k^{(n)}]$, $k \geq 1$, coincide.

Fix $k \geq 1$. We first show that

$$\Delta\widetilde{U}^{(n)}\big(\tau_k^{(n)}\big) = \Delta U^{(n)}\big(\tau_k^{(n)}\big). \tag{4.74}$$

Equality (4.30) shows that there cannot be an arrival at time $\tau_k^{(n)}$, for such an arrival would have a positive lead time and hence increase $W_S^{(n)}$ without increasing $K^{(n)}$ (see Proposition 4.3). In other words, $\Delta A^{(n)}(\tau_k^{(n)}) = 0$. Because there is no arrival at $\tau_k^{(n)}$, the measure-valued process $\mathcal{W}_S^{(n)}$ is continuous at $\tau_k^{(n)}$. Taking the limit in (4.5) as $t \uparrow \tau_k^{(n)}$, we obtain

$$\mathcal{U}^{(n)}\big(\tau_k^{(n)}-\big)(-\infty, 0] = \big[\mathcal{W}_S^{(n)}\big(\tau_k^{(n)}\big)(-\infty, 0] - K^{(n)}\big(\tau_k^{(n)}-\big)\big]^+ = \Delta K^{(n)}\big(\tau_k^{(n)}\big),$$

where the last equality is a consequence of (4.36). However, (4.10) implies that $\mathcal{U}^{(n)}(\tau_k^{(n)}-)(-\infty, 0) \leq \lim_{t \uparrow \tau_k^{(n)}} \mathcal{U}^{(n)}(t)(-\infty, 0) = 0$, so $\Delta\widetilde{U}^{(n)}(\tau_k^{(n)}) = -\mathcal{U}^{(n)}(\tau_k^{(n)}-)\{0\} = -\Delta K^{(n)}(\tau_k^{(n)})$. From (4.6) and the continuity of $W_S^{(n)}$ at $\tau_k^{(n)}$, we see that $-\Delta K^{(n)}(\tau_k^{(n)})$ is also equal to $\Delta U^{(n)}(\tau_k^{(n)})$, and (4.74) is proved.

We next show that $\Delta\widetilde{U}^{(n)}(t) = \Delta U^{(n)}(t)$ for $t \in (\sigma_k^{(n)}, \tau_k^{(n)})$. If $E^{(n)}(t-) > 0$, then the definitions of $E^{(n)}$ and $\widetilde{U}^{(n)}$, and statements (i), (iii) and (iv) of Proposition 4.8 show that

$$\triangle U^{(n)}(t) = \triangle\widetilde{U}^{(n)}(t) = \triangle V^{(n)}\big(A^{(n)}(t)\big).$$

On the other hand, if $E^{(n)}(t-) = 0$, then properties (ii) and (v) of Proposition 4.8 and the definition of $\widetilde{U}^{(n)}$ show that

$$\triangle U^{(n)}(t) = \triangle\widetilde{U}^{(n)}(t) = \triangle A^{(n)}(t) v^{(n)}_{A^{(n)}(t)} - \mathcal{U}^{(n)}(t-)\{0\}.$$

Now, let $S^{(n)}$ be the (random) set of times $s \geq 0$ for which $U^{(n)}(s) > 0$ and at least one of the following three properties holds:

$$\triangle A^{(n)}(s) > 0,$$
$$E^{(n)}(s-) = 0$$

or

$$\mathcal{U}^{(n)}(s-)\big\{E^{(n)}(s-)\big\} = 0.$$

Suppose $U^{(n)}(s) > 0$. If $E^{(n)}(s-) = 0$, then the fact that $E^{(n)}(s) > 0$ by (4.10) implies $\triangle E^{(n)}(s) > 0$, while if $\mathcal{U}^{(n)}(s-)\{E^{(n)}(s-)\} = 0$, the definition of $E^{(n)}(s)$ implies that $\triangle(\mathcal{U}^{(n)}(s)\{E^{(n)}(s)\}) > 0$. Thus, the set $S^{(n)}$ is countable, and on the set $\{s \in (\sigma_k^{(n)}, t) : U^{(n)}(s) > 0\} \setminus S^{(n)}$, the process $U^{(n)}$ decreases with unit rate by Proposition 4.8(i). Therefore, the total amount of this decrease on any time interval of the form $(\sigma_k^{(n)}, t)$ equals $\int_{\sigma_k^{(n)}}^t \mathbb{I}_{\{U^{(n)}(s)>0\}}\, ds$, which coincides with the absolutely continuous part of $\widetilde{U}^{(n)}(t) - \widetilde{U}^{(n)}(\sigma_k^{(n)}-)$ on the same interval. This concludes the proof of (4.70).

Adding and subtracting t to (4.70), by the definition (2.6) of the netput process $N^{(n)}$ and the nonnegativity of $U^{(n)}$, we obtain

$$U^{(n)}(t) = N^{(n)}(t) + \int_0^t \mathbb{I}_{\{U^{(n)}(s)=0\}}\, ds - R_U^{(n)}(t), \tag{4.75}$$

while substituting (4.15) and (2.8) into (4.6), we have

$$U^{(n)}(t) = N^{(n)}(t) + I_S^{(n)}(t) + K_-^{(n)}(t) - K_+^{(n)}(t)$$

for $t \geq 0$. On the other hand, we know that

$$\int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}}\, dI_S^{(n)}(s) = 0 \quad \text{and} \quad \int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}}\, dK_-^{(n)}(s) = 0,$$

where the former equality holds because $W_S^{(n)} \geq U^{(n)}$ by (4.8), and $I_S^{(n)}$ increases only at times when $W_S^{(n)}$ is zero, while the latter holds by (4.16). From the last three displays, we conclude that

$$\int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}}\, dR_U^{(n)}(s) = \int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}}\, dK_+^{(n)}(s). \tag{4.76}$$

On the other hand, since $U^{(n)}(s) = 0$ implies $\Delta A^{(n)}(s) = 0$, from properties (i) and (ii) of Proposition 4.8 and the fact that $R_U^{(n)}$ is a pure jump process with $\Delta R_U^{(n)}(t) = \mathcal{U}^{(n)}(t-)\{0\}$, it follows that

$$\int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)=0\}}\, dR_U^{(n)}(s) = \int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)=0\}}\, dK_+^{(n)}(s).$$

Together, the last two equalities imply $R_U^{(n)} = K_+^{(n)}$, which, when substituted into (4.70), yields (4.72). □

5. The reneging system. In this section we bound the difference in workload between the pre-limit reference and reneging systems—Lemma 5.2 provides a lower bound, while Lemma 5.6 provides an upper bound. The proof of the upper bound uses an optimality property of EDF that may be of independent interest.

THEOREM 5.1. *Let π be a service policy for a single-station, single-customer-class queueing system with reneging such that the customer arrival times to this system do not have a finite accumulation point. Let $R_\pi(t)$ be the amount of work removed from this system up to time t due to lateness. Let $R_W(t)$ be the amount of work removed due to lateness up to time t from the EDF system with reneging and the same interarrival times, service times and lead times as in the former system. Then for every $t \geq 0$, we have*

$$R_W(t) \leq R_\pi(t). \tag{5.1}$$

The proof of Theorem 5.1 is deferred to the Appendix. The related fact that the EDF protocol minimizes the number of late customers in the $G/M/c$ queue was proved in [29], and the main idea of our proof is similar to that of [29]. However, our argument is pathwise and the only assumption on the distribution of the system stochastic primitives that we impose is that customer arrivals do not have a finite accumulation point. This assumption is clearly satisfied almost surely by a $GI/G/1$ queue.

5.1. *Comparison results.* In this section, we establish bounds on the difference between the processes $U^{(n)}$ and $W^{(n)}$. In Section 6.1, this difference will be shown to be negligible in the heavy traffic limit. We start with Lemma 5.2 showing that $W^{(n)} \leq U^{(n)}$, which implies that $R_U^{(n)} \leq R_W^{(n)}$ (see Corollary 5.3).

In the proofs of these results, we will make frequent use of the observation that, by (4.69) and (4.70),

$$\begin{aligned} W^{(n)}(t) - U^{(n)}(t) = {} & \int_0^t \mathbb{I}_{\{U^{(n)}(s)>0\}}\, ds - \int_0^t \mathbb{I}_{\{W^{(n)}(s)>0\}}\, ds \\ & + R_U^{(n)}(t) - R_W^{(n)}(t) \end{aligned} \tag{5.2}$$

for $t \in [0, \infty)$.

LEMMA 5.2. *For every $t \geq 0$, we have*

$$W^{(n)}(t) \leq U^{(n)}(t). \tag{5.3}$$

PROOF. Let

$$\tau \triangleq \min\{t \geq 0 : W^{(n)}(t) > U^{(n)}(t)\}. \tag{5.4}$$

If $\tau = +\infty$, then (5.3) holds. Assume $\tau < +\infty$. In this case, we claim that the minimum on the right-hand side of (5.4) is attained. Indeed, (5.2) and the fact that $R_U^{(n)}$ and $R_W^{(n)}$ are pure jump processes show that the only way that $W^{(n)} - U^{(n)}$ can become strictly positive is via a jump. Thus $W^{(n)}(\tau) > U^{(n)}(\tau)$. Since

$\mathcal{W}(\tau)(-\infty,0]=\mathcal{U}(\tau)(-\infty,0]=0$ (in fact, this equality holds for any time t), this means there must exist a $y>0$ such that

$$\mathcal{W}^{(n)}(\tau)(y,\infty)>\mathcal{U}^{(n)}(\tau)(y,\infty). \tag{5.5}$$

Let

$$\tau_0 \triangleq \inf\bigl\{t\in[0,\tau]:\mathcal{W}^{(n)}(t)(y+\tau-t,\infty)>\mathcal{U}^{(n)}(t)(y+\tau-t,\infty)\bigr\}. \tag{5.6}$$

By (5.5), the above infimum is over a nonempty set. Lemma 4.7 and Proposition 4.8 imply that the only difference in the dynamics of $\mathcal{W}^{(n)}$ and $\mathcal{U}^{(n)}$ is that the arriving mass $v_k^{(n)}$ is concentrated at $L_k^{(n)}$ in the case of the EDF system with reneging and distributed in $[L_k^{(n)},\infty)$ in the reference system. On the other hand, in both systems at time $t\in[0,\tau]$, no mass leaves the interval $(y+\tau-t,\infty)$ due to lateness. This implies that $\mathcal{W}^{(n)}(t)(y+\tau-t,\infty)-\mathcal{U}^{(n)}(t)(y+\tau-t,\infty)$, $t\in[0,\tau]$, has no positive jumps and therefore

$$\mathcal{W}^{(n)}(t)(y+\tau-\tau_0,\infty)=\mathcal{U}^{(n)}(t)(y+\tau-\tau_0,\infty). \tag{5.7}$$

By (5.5) and (5.7), $\tau_0<\tau$. Thus, there exists $t\in(\tau_0,\tau)$, where $t-\tau_0$ is arbitrarily small and

$$\mathcal{W}^{(n)}(t)(y+\tau-t,\infty)>\mathcal{U}^{(n)}(t)(y+\tau-t,\infty). \tag{5.8}$$

However, we claim that (5.7) and (5.8) imply that for all $t\in(\tau_0,\tau)$, where $t-\tau_0$ is small enough, it must be that

$$\mathcal{W}^{(n)}(t)(0,y+\tau-t]>0, \tag{5.9}$$

$$\mathcal{U}^{(n)}(t)(0,y+\tau-t]=0. \tag{5.10}$$

Indeed, if (5.9) is false, then the left-hand side of (5.8) is equal to $W^{(n)}(t)$, and consequently decreases with unit speed as long as it is nonzero in some time interval beginning with τ_0. Similarly, if (5.10) is false, the right-hand side of (5.8) is constant on some interval beginning with τ_0. In both cases, due to (5.7), (5.8) cannot hold for $t\in(\tau_0,\tau)$ with $t-\tau_0$ arbitrarily small. But (5.8)–(5.10) yield $W^{(n)}(t)>U^{(n)}(t)$ for some $t<\tau$, which contradicts (5.4). □

COROLLARY 5.3. *For every* $t\geq 0$,

$$R_U^{(n)}(t)\leq R_W^{(n)}(t). \tag{5.11}$$

Moreover, for $k\geq 1$ *and* $t\geq\sigma_k^{(n)}$,

$$R_U^{(n)}(t)-R_U^{(n)}\bigl(\sigma_k^{(n)}-\bigr)\leq R_W^{(n)}(t)-R_W^{(n)}\bigl(\sigma_k^{(n)}-\bigr). \tag{5.12}$$

PROOF. Lemma 5.2 and (5.2) imply that for $0 \le s \le t$,

$$\bigl(R_U^{(n)}(t) - R_U^{(n)}(s)\bigr) - \bigl(R_W^{(n)}(t) - R_W^{(n)}(s)\bigr) \le U^{(n)}(s) - W^{(n)}(s). \tag{5.13}$$

Substituting $s = 0$ into (5.13) and using the fact that $R_U^{(n)}(0) = R_W^{(n)}(0) = U^{(n)}(0) = W^{(n)}(0) = 0$, we obtain (5.11). Likewise, for $0 \le s \le \sigma_k^{(n)} \le t$, taking limits as s tends to $\sigma_k^{(n)}$ in (5.13), and using the fact that $U^{(n)}(\sigma_k^{(n)}-) = W^{(n)}(\sigma_k^{(n)}-) = 0$, which follows from (4.41), Lemma 5.2 and the nonnegativity of $W^{(n)}$, we obtain (5.12). □

The proofs of Lemma 5.2 and Corollary 5.3 show the following more general (and intuitively obvious) fact: if all customers in the EDF system with reneging get larger deadlines, this results in a larger workload at every time t and a smaller total amount of mass removed from the system due to lateness in the time interval $[0, t]$.

We now establish an inequality between the frontiers in both systems.

LEMMA 5.4. *For every $t \ge 0$ such that $U^{(n)}(t) > 0$, we have*

$$E^{(n)}(t) \le F^{(n)}(t). \tag{5.14}$$

PROOF. Subtracting (4.5) from (4.6), we see that for any $y \in \mathbb{R}$,

$$\begin{aligned} \mathcal{U}^{(n)}(t)(y, \infty) &= W_S^{(n)}(t) - K^{(n)}(t) - \bigl[\mathcal{W}_S^{(n)}(t)(-\infty, y] - K^{(n)}(t)\bigr]^+ \\ &\le \mathcal{W}_S^{(n)}(t)(y, \infty). \end{aligned} \tag{5.15}$$

Now, assume that for some t we have $F^{(n)}(t) < E^{(n)}(t)$. In this case,

$$\begin{aligned} W^{(n)}(t) &\ge \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{W}^{(n)}(t)\bigl(F^{(n)}(t), \infty\bigr) \\ &= \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{V}^{(n)}(t)\bigl(F^{(n)}(t), \infty\bigr) \\ &\ge \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{V}^{(n)}(t)\bigl[E^{(n)}(t), \infty\bigr) \\ &\ge \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{W}_S^{(n)}(t)\bigl[E^{(n)}(t), \infty\bigr) \\ &\ge \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{U}^{(n)}(t)\bigl[E^{(n)}(t), \infty\bigr) \\ &\ge U^{(n)}(t), \end{aligned} \tag{5.16}$$

where the second line follows from the fact that none of the customers in the EDF system with reneging that have lead times at time t greater than $F^{(n)}(t)$ has received any service up to time t, the second-last inequality follows from (5.15) and the last line holds due to the equality $U^{(n)}(t) = \mathcal{U}^{(n)}(t)[E^{(n)}(t), \infty)$. When combined with the assumption that $U^{(n)}(t) > 0$, this implies that $W^{(n)}(t) > 0$. This, in turn, implies that $\mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} > 0$ because the residual service time

of the currently served customer is strictly positive. Thus, the last inequality in (5.16) is strict, which contradicts (5.3). □

Let $D^{(n)}(t)$ be the amount of work deleted by the EDF system with reneging in the time interval $[0, t]$ that is associated with customers whose lead times upon arrival were smaller than the value of the frontier at the time of their arrival. In the proof of the next lemma, we will make use of the elementary fact that by the definition of $F^{(n)}$ we have

$$F^{(n)}(t_1) - (t_2 - t_1) \le F^{(n)}(t_2), \qquad S_1^{(n)} \le t_1 \le t_2. \tag{5.17}$$

LEMMA 5.5. *For every* $t \ge 0$,

$$U^{(n)}(t) - W^{(n)}(t) \le D^{(n)}(t). \tag{5.18}$$

PROOF. If $t \in [\tau_{k-1}^{(n)}, \sigma_k^{(n)})$ for some $k \ge 1$, then $U^{(n)}(t) = 0$ by (4.41). Thus, by (4.19), it suffices to prove (5.18) on $[\sigma_k^{(n)}, \tau_k^{(n)})$ for every $k \ge 1$. Let $k \ge 1$. Suppose that (5.18) is false for some $t \in [\sigma_k^{(n)}, \tau_k^{(n)})$. Let

$$\tau \stackrel{\Delta}{=} \min\{t \in [\sigma_k^{(n)}, \tau_k^{(n)}) | U^{(n)}(t) - W^{(n)}(t) > D^{(n)}(t)\}. \tag{5.19}$$

We first argue that the minimum on the right-hand side of (5.19) is attained. Indeed, by (5.2) and Lemma 5.2, it is clear that $U^{(n)} - W^{(n)}$ cannot increase except by a jump that is due to lateness in the EDF system with reneging. Thus, we have $\mathcal{W}^{(n)}(\tau-)\{0\} > 0$ and

$$U^{(n)}(\tau) - W^{(n)}(\tau) > D^{(n)}(\tau). \tag{5.20}$$

Also, (4.41), (4.42) and Lemma 5.2 imply that $U^{(n)}(\sigma_k^{(n)}) = \triangle U^{(n)}(\sigma_k^{(n)}) = \triangle W^{(n)}(\sigma_k^{(n)}) = W^{(n)}(\sigma_k^{(n)})$, so $\sigma_k^{(n)} < \tau$. In particular, (5.19) implies

$$U^{(n)}(\tau-) - W^{(n)}(\tau-) \le D^{(n)}(\tau-). \tag{5.21}$$

Let k_0 be the index of the customer arriving at time $\sigma_k^{(n)}$, that is, $S_{k_0}^{(n)} = \sigma_k^{(n)}$. Let $k_1 \ge k_0$ be the index of a customer who reneges in the reneging system at time τ. There must be such a customer, and there may in fact be more than one such customer. The amount of work associated with all such customers at time τ is $\mathcal{W}^{(n)}(\tau-)\{0\}$, and we seek to show that this work is bounded above by $\triangle D^{(n)}(\tau)$. We have $S_{k_1}^{(n)} \in [\sigma_k^{(n)}, \tau)$ and $L_{k_1}^{(n)} - (\tau - S_{k_1}^{(n)}) = 0$. The subsequent analysis is divided into two cases.

Case I. For every customer k_1 chosen as just described, assume there is a customer ℓ arriving in the time interval $[\sigma_k^{(n)}, S_{k_1}^{(n)}]$ who is at least as urgent as customer k_1 when customer k_1 arrives but whose associated mass in the reference system is at least partly assigned so that upon the arrival of customer k_1, this mass is to the right of $L_{k_1}^{(n)}$. In other words, $\ell \in [k_0, k_1]$, $L_\ell^{(n)} - (S_{k_1}^{(n)} - S_\ell^{(n)}) \le L_{k_1}^{(n)}$

and $\triangle\mathcal{W}^{(n)}(S_\ell^{(n)})\{L_\ell^{(n)}\} > \triangle\mathcal{U}^{(n)}(S_\ell^{(n)})[L_l^{(n)}, L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}]$. In this case, $\triangle\mathcal{U}^{(n)}(S_\ell^{(n)})(L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}, \infty) > 0$. Indeed, by Lemma 4.7 and Proposition 4.8(iv) (describing the only case in which part of the mass of a new customer is distributed by the reference workload to a point other than its lead time) $\triangle U(S_\ell^{(n)}) = v_\ell^{(n)}$ and $\triangle\mathcal{U}^{(n)}(S_\ell^{(n)})(-\infty, L_\ell^{(n)}) = 0$ [see (4.42), (4.43), (4.49), (4.50) and (4.4)]. Let $s > L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}$ satisfy $\triangle\mathcal{U}^{(n)}(S_\ell^{(n)})\{s\} > 0$. Such a point s exists since the measure $\mathcal{U}^{(n)}(S_\ell^{(n)})$ is discrete.

If $\ell > k_0$ (e.g., $\ell = k_1$), then, by (4.51) in Proposition 4.8(iv) and Lemma 5.4, we have $s \le E^{(n)}(S_\ell^{(n)}-) \le F^{(n)}(S_\ell^{(n)}-) \le F^{(n)}(S_\ell^{(n)})$. Thus, by (5.17), $L_{k_1}^{(n)} < s - (S_{k_1}^{(n)} - S_\ell^{(n)}) \le F^{(n)}(S_\ell^{(n)}) - (S_{k_1}^{(n)} - S_\ell^{(n)}) \le F^{(n)}(S_{k_1}^{(n)})$.

If $\ell = k_0$, then, because $\mathcal{U}^{(n)}(S_{k_0}^{(n)})\{s\} > 0$, we have $\mathcal{W}_S^{(n)}(S_{k_0}^{(n)})\{s\} > 0$ by the definition of $\mathcal{U}^{(n)}$. In this case $\mathcal{W}^{(n)}(S_{k_0}^{(n)})\{s\} = 0$, because $W^{(n)} \equiv 0$ on $[\tau_{k-1}^{(n)}, \sigma_k^{(n)})$ by (4.41) and Lemma 5.2, so $\mathcal{W}^{(n)}(S_{k_0}^{(n)}) = \mathcal{W}^{(n)}(\sigma_k^{(n)}) = v_{k_0}^{(n)}\delta_{L_{k_0}^{(n)}}$ and $s > L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_{k_0}^{(n)} \ge L_{k_0}^{(n)}$ by the definitions of ℓ and s. Thus, a customer with lead time equal to s at time $S_{k_0}^{(n)}$ has already been in service in the EDF system with reneging, so $L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_{k_0}^{(n)} < s \le F^{(n)}(S_{k_0}^{(n)})$ and consequently, by (5.17), $L_{k_1}^{(n)} < F^{(n)}(S_{k_0}^{(n)}) - (S_{k_1}^{(n)} - S_{k_0}^{(n)}) \le F^{(n)}(S_{k_1}^{(n)})$.

Thus, regardless of the value of ℓ, $L_{k_1}^{(n)} < F^{(n)}(S_{k_1}^{(n)})$. In other words, under the Case I assumption, every customer k_1 who becomes late at time τ in the EDF system with reneging arrived with initial lead time smaller than the value of $F^{(n)}$ at the time of its arrival. The work associated with these customers deleted at time τ is $\triangle D^{(n)}(\tau)$. We conclude that $\mathcal{W}^{(n)}(\tau-)\{0\} = \triangle D^{(n)}(\tau)$. However, by (5.2), we have $\triangle(U^{(n)} - W^{(n)})(\tau) \le \mathcal{W}^{(n)}(\tau-)\{0\}$, and so $\triangle(U^{(n)} - W^{(n)})(\tau) \le \triangle D^{(n)}(\tau)$. This, together with (5.21), contradicts (5.20).

Case II. For a customer k_1 chosen as described above, assume that every customer ℓ arriving in the time interval $[\sigma_k^{(n)}, S_{k_1}^{(n)}]$ who is as least as urgent as customer k_1 when customer k_1 arrives has all its associated mass initially assigned in the reference system to the interval $(0, L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}]$ upon arrival. Customers ℓ who are less urgent then k_1 must have lead times satisfying $L_\ell^{(n)} > L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}$, and hence the mass brought by such customers must be initially assigned to the half-line $(L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}, \infty)$ in both systems. Then for every $t \in [\sigma_k^{(n)}, S_{k_1}^{(n)}]$, we have

$$\mathcal{W}^{(n)}(t)\big(0, L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})\big] \le \mathcal{U}^{(n)}(t)\big(0, L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})\big], \tag{5.22}$$

as we now explain. Under the Case II assumption the arrival of new mass is the same on both sides of (5.22). Furthermore, disregarding lateness and new arrivals, both sides of (5.22) decrease at unit rate so long as they are nonzero. Finally, by

(5.12) the amount of late work removed from the EDF system with reneging in the time interval $[\sigma_k^{(n)}, t]$ is greater than or equal to the amount of late work removed from $\mathcal{U}^{(n)}$ in this time interval. Therefore, (5.22) holds for every $t \in [\sigma_k^{(n)}, S_{k_1}^{(n)}]$.

We claim that (5.22) in fact holds for all $t \in [\sigma_k^{(n)}, \tau)$. Suppose this is not the case. Let

$$\eta \triangleq \inf\{t \in [S_{k_1}^{(n)}, \tau) | \mathcal{W}^{(n)}(t)(0, L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})] > \mathcal{U}^{(n)}(t)(0, L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})]\}. \tag{5.23}$$

The strict inequality in (5.23) can occur only because of an arrival at time t which brings mass to the interval $(0, L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})]$ under the $\mathcal{W}^{(n)}$ measure but not under the $\mathcal{U}^{(n)}$ measure. The arrival at time k_1 does not have this property because the Case II assumption applies to $\ell = k_1$. Therefore, $\eta > S_{k_1}^{(n)}$.

Also, for $t \in [S_{k_1}^{(n)}, \tau)$,

$$\mathcal{W}^{(n)}(t)\{L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})\} > 0, \tag{5.24}$$

because the customer k_1 is present in the EDF system with reneging at time t. By (4.4), (5.24) and the definition of η, we have $E^{(n)}(t) \le L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})$ for $t \in [S_{k_1}^{(n)}, \eta)$. Thus, $E^{(n)}(t-) \le L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})$ for $t \in (S_{k_1}^{(n)}, \eta]$. We argue that this implies that the amounts of mass arriving in both the EDF system with reneging and the reference workload at any time $t \in (S_{k_1}^{(n)}, \eta]$ with lead times upon arrival less than or equal to $L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})$ are the same. Indeed, Proposition 4.8, especially (4.51), implies that no mass arriving at time t with lead time smaller than $E^{(n)}(t-)$ in the EDF system with reneging is distributed to lead times greater than $E^{(n)}(t-)$ by the reference workload. Also, Proposition 4.8(iii) and (v) imply that the mass arriving at time t with lead time greater than or equal to $E^{(n)}(t-)$ is distributed in the same way by the EDF system with reneging and the reference system. By the same argument as in the case of $t \in [\sigma_k^{(n)}, S_{k_1}^{(n)}]$, we conclude that (5.22) holds for $t \in [S_{k_1}^{(n)}, \eta]$, which contradicts the definition of η. We have shown that (5.22) holds for $t \in [\sigma_k^{(n)}, \tau)$.

Letting $t \uparrow \tau$ in (5.22) and using the fact that $L_{k_1}^{(n)} - (\tau - S_{k_1}^{(n)}) = 0$, we get $\mathcal{W}^{(n)}(\tau-)\{0\} \le \mathcal{U}^{(n)}(\tau-)\{0\}$. Thus, by (5.2), $\triangle(U^{(n)} - W^{(n)})(\tau) = \mathcal{W}^{(n)}(\tau-)\{0\} - \mathcal{U}^{(n)}(\tau-)\{0\} \le 0$ which, together with (5.21) and the fact that $D^{(n)}$ is nondecreasing, contradicts (5.20). □

For the sake of the next proof, we define a sequence of auxiliary *hybrid systems* (with the same stochastic primitives as in the case of the EDF systems described in Section 2.2) as follows. The hybrid system gives priority to the jobs whose lead times upon arrival are smaller than the current frontier $F^{(n)}$ in the corresponding

EDF system with reneging. In other words, for each k, the kth customer arriving at the hybrid system joins the high-priority class if and only if

$$L_k^{(n)} < F^{(n)}\big(S_k^{(n)}\big). \tag{5.25}$$

The system processes high-priority customers according to the FIFO service discipline. When the priority class empties, the system goes idle until either another high-priority customer arrives and the system resumes service in the manner described above, or the corresponding EDF system with reneging finishes serving the customers who have received priority in the hybrid system. Here, we are using the fact that the high-priority customers leave the hybrid system before they leave the EDF system with reneging, which is a consequence of the optimality of the EDF discipline established in Theorem 5.1. (We have slightly abused the terminology here, identifying the kth customer in the hybrid system with the corresponding customer from the EDF system with reneging, while, formally, only the random variables $u_k^{(n)}$, $v_k^{(n)}$ and $L_k^{(n)}$ associated with these customers are the same.) Whenever the EDF system with reneging finishes serving a batch of customers who have received high priority in the hybrid system, both systems then serve the low-priority class using the EDF discipline until the next high-priority customer arrives. In both systems, if a customer is present when his deadline passes, he leaves the queue immediately, regardless of his class. The measure-valued workload process associated with the hybrid system will be denoted by $\mathcal{W}_H^{(n)}$.

LEMMA 5.6. *For every $t \geq 0$, we have*

$$\begin{aligned} &U^{(n)}(t) - W^{(n)}(t) \\ &\quad \leq \sum_{k=1}^{A^{(n)}(t)} v_k^{(n)} \wedge \big(\mathcal{W}^{(n)}\big(S_k^{(n)}-\big)\big(0, F^{(n)}\big(S_k^{(n)}\big)\big) + v_k^{(n)} - L_k^{(n)}\big)^+ \\ &\qquad\qquad \times \mathbb{I}_{\{L_k^{(n)} < F^{(n)}(S_k^{(n)})\}}. \end{aligned} \tag{5.26}$$

PROOF. By Lemma 5.5, it suffices to show that $D^{(n)}(t)$ is not greater than the right-hand side of (5.26). By Theorem 5.1, $D^{(n)}(t)$, the amount of unfinished work associated with customers who arrived with lead times smaller than $F^{(n)}$ and were deleted in the time interval $[0, t]$ by the EDF system with reneging, is not greater than the unfinished work associated with these customers and deleted by the corresponding hybrid system. Note that the customers with lead times satisfying (5.25) form a priority class in both the EDF system with reneging and the hybrid system, and so their service is not affected by the presence of other customers. Furthermore, unfinished work associated with deleted customers who arrived with lead times greater than or equal to $F^{(n)}$ is the same in both systems.

For each k, if (5.25) holds, then the kth customer of the hybrid system belongs to the high-priority class. Moreover, if, for some $l < k$, $L_l^{(n)} < F^{(n)}(S_l^{(n)})$, then,

by (5.17), $L_l^{(n)} - (S_k^{(n)} - S_l^{(n)})$, the lead time of the lth customer at time $S_k^{(n)}$, is smaller than $F^{(n)}(S_k^{(n)})$. Thus, if (5.25) holds, the kth customer waits at most $\mathcal{W}_H^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)}))$ time units before he starts receiving service. (His waiting time may actually be smaller because some of the high-priority customers in queue who have arrived before him may renege before they are served to completion.) We have

$$\mathcal{W}_H^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) \le \mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})), \tag{5.27}$$

because, in both systems under consideration, the arrivals with lead times smaller than $F^{(n)}$ and the corresponding work associated with them are the same, the server serves these customers with rate 1 as long as they are present in the system, but, by Theorem 5.1, the amount of unfinished work associated with these customers and deleted by the EDF system with reneging is not greater than the work deleted by the hybrid system. Thus, if (5.25) holds, the time required for the hybrid system to fully serve the kth customer is at most $\mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) + v_k^{(n)}$. Therefore, under assumption (5.25), the unfinished work deleted by the hybrid system due to lateness of the kth customer is at most $v_k^{(n)} \wedge (\mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) + v_k^{(n)} - L_k^{(n)})^+$. Thus, the amount of work associated with high-priority customers deleted by the hybrid system up to time t is bounded above by the right-hand side of (5.26). □

6. Heavy traffic analysis. In Sections 6.1 and 6.2, respectively, we identify the heavy traffic limit of the scaled workload and the scaled reneged work in the reneging system. In both cases, this is done by first considering the reference system, which is easier to analyze, and then using the bounds derived in Section 5.1 to show that the limits in both systems coincide. For the heavy traffic analysis of the reference system, we will find it useful to introduce the following scaled quantities:

$$\begin{aligned} \widehat{U}^{(n)}(t) &\triangleq \frac{1}{\sqrt{n}} U^{(n)}(nt), \qquad \widehat{R}_U^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} R_U^{(n)}(nt), \\ \widehat{K}_+^{(n)}(t) &\triangleq \frac{1}{\sqrt{n}} K_+^{(n)}(nt), \end{aligned} \tag{6.1}$$

and, for every Borel set $B \subset \mathbb{R}$,

$$\widehat{\mathcal{U}}^{(n)}(t)(B) \triangleq \frac{1}{\sqrt{n}} \mathcal{U}^{(n)}(nt)(\sqrt{n}B). \tag{6.2}$$

Also, define

$$\mathcal{U}^* \triangleq \Phi(\mathcal{W}_S^*) \quad \text{and} \quad U^*(\cdot) \triangleq \mathcal{U}^*(\cdot)(\mathbb{R}) = \Phi(\mathcal{W}_S^*)(\mathbb{R}). \tag{6.3}$$

6.1. *Proofs of main results concerning the workload.*

6.1.1. *Proof of Theorem 3.4.* In Lemma 6.1, we use the continuity property of the mapping Φ established in Lemma 4.1, along with the characterization of the

heavy traffic limit of the workload measure-valued process in the standard system, to identify the heavy traffic limit of the workload in the reference system. Let $\Lambda_{H(0)} : D[0,\infty) \to D[0,\infty)$ be the mapping defined, for every $\phi \in D[0,\infty)$ and $t \geq 0$, by

$$\Lambda_{H(0)}(\phi)(t) \stackrel{\Delta}{=} \phi(t) - \sup_{s\in[0,t]}\Big[\big(\phi(s) - H(0)\big)^+ \wedge \inf_{u\in[s,t]} \phi(u)\Big]. \tag{6.4}$$

If ϕ is nonnegative, then by Theorem 1.4 from [25], $\Lambda_{H(0)}(\phi)$ is the function in $D[0,\infty)$ obtained by double reflection of ϕ at 0 and $H(0)$. In other words, $\Lambda_{H(0)}(\phi)$ takes values in $[0, H(0)]$ and has the unique decomposition

$$\Lambda_{H(0)}(\phi) = \phi - \kappa_+ + \kappa_-, \tag{6.5}$$

where $\kappa_\pm$ are nondecreasing RCLL functions satisfying $\kappa_\pm(0-) = 0$ and

$$\begin{aligned} \int_{[0,\infty)} \mathbb{I}_{\{\Lambda_{H(0)}(\phi)(s) < H(0)\}}\, d\kappa_+(s) = 0, \\ \int_{[0,\infty)} \mathbb{I}_{\{\Lambda_{H(0)}(\phi)(s) > 0\}}\, d\kappa_-(s) = 0. \end{aligned} \tag{6.6}$$

LEMMA 6.1. *The process U^* satisfies*

$$U^* = \Lambda_{H(0)}(W_S^*) \tag{6.7}$$

and has the same distribution as W^*. Moreover, $\widehat{\mathcal{U}}^{(n)} \Rightarrow \mathcal{U}^* = \Phi(\mathcal{W}_S^*)$ and $\widehat{U}^{(n)} \Rightarrow W^*$ as $n \to \infty$.

PROOF. By the definition of U^* and Φ given in (6.3) and (4.2), respectively,

$$U^*(t) = \Phi(\mathcal{W}_S^*)(\mathbb{R})(t) = W_S^*(t) - \sup_{s\in[0,t]}\Big[\mathcal{W}_S^*(-\infty, 0] \wedge \inf_{u\in[s,t]} W_S^*(u)\Big], \qquad t \geq 0.$$

Since (3.1)–(3.3) imply $\mathcal{W}_S^*(t)(-\infty, 0] = (W_S^*(t) - H(0))^+$ for every $t \geq 0$, this shows that $U^* = \Lambda_{H(0)}(W_S^*)$. By the characterization of W_S^* given at the end of Section 2.4, $\Lambda_{H(0)}(W_S^*)$ is a Brownian motion with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$, reflected at 0 and $H(0)$. This proves the first claim.

Next, using the definition $\mathcal{U}^{(n)} = \Phi(\mathcal{W}_S^{(n)})$ and the scaling properties of Φ, it is easy to see that $\widehat{\mathcal{U}}^{(n)} = \Phi(\widehat{\mathcal{W}}_S^{(n)})$. Since, by Theorem 3.2, we know that $\widehat{\mathcal{W}}_S^{(n)} \Rightarrow \mathcal{W}_S^*$, where $\mathcal{W}_S^*$ is continuous and $\mathcal{W}_S^*(t)$ has a continuous distribution for every t, an application of the continuous mapping theorem, along with the continuity property of Φ stated in Lemma 4.1, shows that $\widehat{\mathcal{U}}^{(n)} \Rightarrow \Phi(\mathcal{W}_S^*)$. This, in particular, implies that $\widehat{U}^{(n)} = \widehat{\mathcal{U}}^{(n)}(\mathbb{R}) \Rightarrow U^*$. Since U^* has the same distribution as W^*, this proves the lemma. □

We identify the heavy traffic limit of the workload in the reneging system. We start with Proposition 6.2, which states that the number of customers in the EDF

system with reneging having lead times not greater than the current frontier and the work associated with these customers are negligible under heavy traffic scaling. Then, in Corollary 6.3, we use the comparison results established in Section 5.1 to show that the workloads in the reference and reneging systems are equal with high probability and so their heavy traffic limits coincide.

PROPOSITION 6.2. *The processes* $\widehat{\mathcal{W}}^{(n)}(0, \widehat{F}^{(n)}]$ *and* $\widehat{\mathcal{Q}}^{(n)}(0, \widehat{F}^{(n)}]$ *converge in distribution to zero as* $n \to \infty$.

This result holds for the same reason that state-space collapse occurs for priority queues, an idea that can be traced back to [35]. Specifically, in our model, due to the nature of the EDF service discipline, the entire capacity of the server is always devoted to work that lies to the left or at the frontier, as long as the system is nonempty. Thus the process $\mathcal{W}^{(n)}(0, F^{(n)}]$ is equal to the workload in a single-server $GI/G/1$ queue that has netput process $\mathcal{V}^{(n)}(t)(-\infty, F^{(n)}(t)] - t$, $t \geq 0$. By showing that $F^{(n)}(t) < \sqrt{n}y^*$, one shows that this (high-priority) queue is in light traffic as $n \to \infty$, and so its diffusion scaling vanishes in the limit. Since a rigorous proof that $\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}] \Rightarrow 0$ and $\widehat{\mathcal{Q}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}] \Rightarrow 0$ would be very similar to the proofs of Proposition 3.6 and Corollary 3.8 in [7], we omit the details. We note that $\widehat{\mathcal{W}}^{(n)}(0, \widehat{C}^{(n)}) = \widehat{\mathcal{Q}}^{(n)}(0, \widehat{C}^{(n)}) = 0$ by definition.

COROLLARY 6.3. *Let* $T > 0$. *As* $n \to \infty$,

$$\mathbb{P}\big[U^{(n)}(t) = W^{(n)}(t), 0 \leq t \leq nT\big] \to 1. \tag{6.8}$$

PROOF. Because customers with strictly positive lead times do not renege, we have $\mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) \leq \mathcal{W}^{(n)}(S_k^{(n)})(0, F^{(n)}(S_k^{(n)}))$ for $k \geq 1$. Thus, by Lemmas 5.2 and 5.6, to prove (6.8), it suffices to show that as $n \to \infty$,

$$\mathbb{P}\big[\mathcal{W}^{(n)}\big(S_k^{(n)}\big)\big(0, F^{(n)}\big(S_k^{(n)}\big)\big) + v_k^{(n)} \leq L_k^{(n)}, 1 \leq k \leq A^{(n)}(nT)\big] \to 1.$$

However, this follows from the fact that, by (2.15),

$$\max_{1 \leq k \leq A^{(n)}(nT)} v_k^{(n)} = \sqrt{n} \max_{0 \leq t \leq T} \triangle \widehat{N}_S^{(n)}(t) = o\big(\sqrt{n}\big),$$

the inequalities $L_k^{(n)} \geq \sqrt{n}y_*$, $y_* > 0$, and Proposition 6.2. □

Theorem 3.4 now follows immediately from Lemma 6.1 and Corollary 6.3.

6.1.2. *Proofs of Proposition 3.5 and Theorem 3.6.* We present the proofs of the remaining two limit theorems concerning the measure-valued workload processes. For this, we need two preliminary results. The first, Lemma 6.4, is that the frontier in the reneging system is strictly positive with high probability. The second result, Proposition 6.5, is a recap of a result established in [7].

LEMMA 6.4. *Let $T > 0$. As $n \to \infty$,*

$$\mathbb{P}\big[F^{(n)}(t) > 0, 0 \le t \le nT\big] \to 1. \tag{6.9}$$

PROOF. Let $0 \le t \le nT$. If $W^{(n)}(t) > 0$, then $F^{(n)}(t)$ is not smaller than the lead time of the currently served customer, so $F^{(n)}(t) > 0$. If $W^{(n)}(t) = 0$, then the customer indexed by $A^{(n)}(t)$ has already been in service, so

$$\begin{aligned} F^{(n)}(t) &\ge L^{(n)}_{A^{(n)}(t)} - \big(t - S^{(n)}_{A^{(n)}(t)}\big) \\ &\ge \sqrt{n}\, y_* - u^{(n)}_{A^{(n)}(t)+1} \\ &\ge \sqrt{n}\, y_* - \max_{1 \le k \le A^{(n)}(nT)+1} u^{(n)}_k. \end{aligned} \tag{6.10}$$

However, $\max_{1 \le k \le A^{(n)}(nT)+1} u^{(n)}_k = o(\sqrt{n})$ by (2.13) (in particular, by the fact that S^* has continuous sample paths), so (6.10) implies (6.9). $\square$

PROPOSITION 6.5 (Proposition 3.4 [7]). *Let $-\infty < y_0 < y^*$ and $T > 0$ be given. As $n \to \infty$,*

$$\sup_{y_0 \le y \le y^*} \sup_{0 \le t \le T} \big|\widehat{\mathcal{V}}^{(n)}(t)(y, \infty) + H\big(y + \sqrt{n}t\big) - H(y)\big| \xrightarrow{P} 0,$$

$$\sup_{y_0 \le y \le y^*} \sup_{0 \le t \le T} \big|\widehat{\mathcal{A}}^{(n)}(t)(y, \infty) + \lambda H\big(y + \sqrt{n}t\big) - \lambda H(y)\big| \xrightarrow{P} 0.$$

PROOF OF PROPOSITION 3.5. Let $T > 0$. We will show that $\widehat{F}^{(n)} \Rightarrow F^*$ in $D_{\mathbb{R}}[0, T]$. By definition, $y^* - \sqrt{n}t \le \widehat{F}^{(n)}(t) \le y^*$. Thus, by Proposition 6.5 and the fact that $H(y) = 0$ for $y \ge y^*$,

$$\sup_{0 \le y \le y^*} \sup_{0 \le t \le T} \big|\widehat{\mathcal{V}}^{(n)}(t)\big(\widehat{F}^{(n)}(t) \vee y, \infty\big) - H\big(\widehat{F}^{(n)}(t) \vee y\big)\big| \xrightarrow{P} 0. \tag{6.11}$$

Putting $y = 0$ in (6.11) and using Lemma 6.4, we obtain

$$\sup_{0 \le t \le T} \big|\widehat{\mathcal{V}}^{(n)}(t)\big(\widehat{F}^{(n)}(t), \infty\big) - H\big(\widehat{F}^{(n)}(t)\big)\big| \xrightarrow{P} 0. \tag{6.12}$$

For any $t \ge 0$,

$$\begin{aligned} \widehat{W}^{(n)}(t) &= \widehat{\mathcal{W}}^{(n)}(t)\big(0, \widehat{F}^{(n)}(t)\big] + \widehat{\mathcal{W}}^{(n)}(t)\big(\widehat{F}^{(n)}(t), \infty\big) \\ &= \widehat{\mathcal{W}}^{(n)}(t)\big(0, \widehat{F}^{(n)}(t)\big] + \widehat{\mathcal{V}}^{(n)}(t)\big(\widehat{F}^{(n)}(t), \infty\big), \end{aligned} \tag{6.13}$$

where the second line follows from the fact that none of the customers in the EDF system with reneging with lead times at time t greater than $F^{(n)}(t)$ has received any service up to time t. This, together with Proposition 6.2 and Theorem 3.4, yields $\widehat{\mathcal{V}}^{(n)}(\widehat{F}^{(n)}, \infty) \Rightarrow W^*$. Thus, by (6.12), we have $H(\widehat{F}^{(n)}) \Rightarrow W^*$ in

$D_{\mathbb{R}}[0,T]$. Applying the continuous function H^{-1} to both sides of this relation and using (3.4), we obtain $\widehat{F}^{(n)} \Rightarrow F^*$ in $D_{\mathbb{R}}[0,T]$. $\square$

PROOF OF THEOREM 3.6. Define a mapping $\psi : \mathbb{R} \to \mathcal{M}$ by the formula $\psi(x)(B) \triangleq \int_{B\cap[x,\infty)} (1 - G(\eta))\, d\eta$ for $x \in \mathbb{R}$ and $B \in \mathcal{B}(\mathbb{R})$. It is easy to see that ψ is continuous. Hence, by Proposition 3.5,

$$\big(\psi\big(\widehat{F}^{(n)}\big), \lambda\psi\big(\widehat{F}^{(n)}\big)\big) \Rightarrow (\psi(F^*), \lambda\psi(F^*)) = (\mathcal{W}^*, \mathcal{Q}^*). \tag{6.14}$$

Let $T > 0$. We claim that

$$\sup_{0\le y\le y^*} \sup_{0\le t\le T} \big|\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - \psi\big(\widehat{F}^{(n)}(t)\big)(y,\infty)\big| \xrightarrow{P} 0, \tag{6.15}$$

$$\sup_{0\le y\le y^*} \sup_{0\le t\le T} \big|\widehat{\mathcal{Q}}^{(n)}(t)(y,\infty) - \lambda\psi\big(\widehat{F}^{(n)}(t)\big)(y,\infty)\big| \xrightarrow{P} 0. \tag{6.16}$$

Indeed, reasoning as in (6.13), we see that, for $0 \le y \le y^*$ and $0 \le t \le T$,

$$\begin{aligned}
&|\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - H\big(\widehat{F}^{(n)}(t) \vee y\big)| \\
&\quad \le |\widehat{\mathcal{W}}^{(n)}(t)\big(\widehat{F}^{(n)}(t) \vee y, \infty\big) - H\big(\widehat{F}^{(n)}(t) \vee y\big)| + \widehat{\mathcal{W}}^{(n)}(t)\big(0, \widehat{F}^{(n)}(t)\big] \\
&\quad = |\widehat{\mathcal{V}}^{(n)}(t)\big(\widehat{F}^{(n)}(t) \vee y, \infty\big) - \psi\big(F^{(n)}(t)\big)(y,\infty)| + \widehat{\mathcal{W}}^{(n)}(t)\big(0, \widehat{F}^{(n)}(t)\big].
\end{aligned}$$

Therefore, (6.15) follows from (6.11) and Proposition 6.2. A similar argument gives (6.16). We have $\widehat{\mathcal{W}}^{(n)}(t)(-\infty,0] = \widehat{\mathcal{W}}^{(n)}(t)(y^*,\infty) = \widehat{\mathcal{Q}}^{(n)}(t)(-\infty,0] = \widehat{\mathcal{Q}}^{(n)}(t)(y^*,\infty) = 0$ and, by Lemma 6.4, $\mathbb{P}[\psi(\widehat{F}^{(n)}(t))(-\infty,0] = 0, 0 \le t \le T] \to 1$ as $n \to \infty$. Also, $\psi(x)(y^*,\infty) = 0$ for every $x \in \mathbb{R}$. Thus, (6.14)–(6.16) imply that $(\widehat{\mathcal{W}}^{(n)}, \widehat{\mathcal{Q}}^{(n)}) \Rightarrow (\mathcal{W}^*, \mathcal{Q}^*)$ in $D_{\mathcal{M}}[0,T]$. $\square$

6.2. *The heavy traffic limit of the reneged work process.* In this section, we identify the limit of the sequence $\{\widehat{R}_W^{(n)}, n \in \mathbb{N}\}$, thereby proving Theorem 3.8. To do this, it is convenient to show that many of the processes under consideration can be put on a common probability space so that certain weak limits established earlier can be replaced by almost sure limits.

LEMMA 6.6. *The processes* $\widehat{\mathcal{W}}_S^{(n)}$, $\widehat{\mathcal{U}}^{(n)}$, $\widehat{\mathcal{W}}^{(n)}$, $n \in \mathbb{N}$, $\mathcal{W}_S^*$, $\mathcal{U}^*$ *and* W^* *can be defined on a common probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *such that* $\mathbb{P}$ *almost surely, as* $n \to \infty$,

$$\widehat{\mathcal{W}}_S^{(n)} \to \mathcal{W}_S^*, \tag{6.17}$$

$$\widehat{W}_S^{(n)} \to W_S^*, \tag{6.18}$$

$$\widehat{\mathcal{W}}_S^{(n)}(\cdot)(-\infty,0] \to \mathcal{W}_S^*(\cdot)(-\infty,0] = \big(W_S^*(\cdot) - H(0)\big)^+, \tag{6.19}$$

$$\widehat{U}^{(n)} \to U^* \tag{6.20}$$

and

$$\widehat{W}^{(n)} \to W^* \stackrel{\Delta}{=} U^*, \tag{6.21}$$

where $\widehat{W}_S^{(n)} = \widehat{\mathcal{W}}_S^{(n)}(\mathbb{R})$, $W_S^* = \mathcal{W}_S^*(\mathbb{R})$, $\widehat{U}^{(n)} = \widehat{\mathcal{U}}^{(n)}(\mathbb{R})$ *and* $U^* = \mathcal{U}^*(\mathbb{R})$. *Furthermore,* W_S^* *is a Brownian motion with variance* $(\alpha^2 + \beta^2)\lambda$ *per unit time and drift* $-\gamma$, *reflected at* 0, *while* U^* *is a doubly reflected Brownian motion on* $[0, H(0)]$, *also with variance* $(\alpha^2 + \beta^2)\lambda$ *per unit time and drift* $-\gamma$. *In particular,*

$$U^* = \Lambda_{H(0)}(W_S^*) = W_S^* - K_+^* + K_-^*, \tag{6.22}$$

where $K_\pm^*$ *are the unique RCLL nondecreasing functions satisfying* $K_\pm^*(0) = 0$ *and*

$$\int_{[0,\infty)} \mathbb{I}_{\{U^*(s) < H(0)\}}\, dK_+^*(s) = 0, \int_{[0,\infty)} \mathbb{I}_{\{U^*(s) > 0\}}\, dK_-^*(s) = 0. \tag{6.23}$$

The almost sure limits in (6.17)–(6.21) *hold uniformly on compact intervals.*

PROOF. Recall from Theorem 3.2 that $\widehat{\mathcal{W}}_S^{(n)} \Rightarrow \mathcal{W}_S^*$. Using the Skorokhod representation theorem, we construct the model primitives $u_j^{(n)}$, $v_j^{(n)}$ and $L_j^{(n)}$ for $j \in \mathbb{N}$ and $n \in \mathbb{N}$ on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the sequence of processes $\widehat{\mathcal{W}}_S^{(n)}$, $n \in \mathbb{N}$, and the limiting process $\mathcal{W}_S^*$ are defined on this space and (6.17) holds. Here and below the almost sure convergences are in the J_1 topology on $D_{\mathcal{M}}[0, \infty)$ or $D_{\mathbb{R}}[0, \infty)$, and since the limits are continuous in every case, this is equivalent to uniform convergence on compact intervals. Since the mapping $f : D_{\mathcal{M}}[0, \infty) \mapsto D_{\mathbb{R}}[0, \infty)$ given by $f(\mu)(\cdot) = \mu(\cdot)(\mathbb{R})$ is continuous, we have (6.18). Under $\mathbb{P}$ the measure-valued process $\mathcal{W}_S^*$ constructed on Ω has the same distribution as the process $\mathcal{W}_S^*$ appearing in Theorem 3.2, and thus $\mathcal{W}_S^*$ takes values in the set of measure-valued process of the form $\int_{B \cap [F_S^o(t), \infty)} (1 - G(y))\, dy$ for some RCLL process $F_S^o(t)$. However, $W_S^*(t) = \int_{\mathbb{R} \cap [F_S^o(t), \infty)} (1 - G(y))\, du = H(F_S^o(t))$; hence $F_S^o(t) = F_S^*(t)$ is given by (3.2). In other words, with F_S^* defined by (3.2), the first equation in (3.3) holds. Due to Proposition 3.1, the above argument also shows that under $\mathbb{P}$, W_S^* is a Brownian motion with variance $(\alpha^2 + \beta^2)\lambda$ per unit time and drift $-\gamma$. In addition, since for each t, the measure $\mathcal{W}_S^*(t)$ is nonatomic, we have (6.19).

Now, following (4.1) and (6.3), we set $\mathcal{U}^{(n)} = \Phi(\mathcal{W}_S^{(n)})$ and $\mathcal{U}^* = \Phi(\mathcal{W}_S^*)$. Also, as defined in (6.2), let $\widehat{\mathcal{U}}^{(n)}$ be the scaled version of $\mathcal{U}^{(n)}$, and let $\widehat{U}^{(n)}$ and $\widehat{U}^*$ be as defined in the statement of the lemma. Then $\widehat{\mathcal{U}}^{(n)}$, $\widehat{U}^{(n)}$, $n \in \mathbb{N}$, $\mathcal{U}^*$ and U^* are also defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and (6.20) follows from Lemma 4.1. This implies (6.21). Since $U^* = \Phi(\mathcal{W}_S^*)(\mathbb{R}) = \Lambda_{H(0)}(W_S^*)$, the characterization of U^* as a doubly reflected Brownian motion that satisfies relations (6.22) and (6.23) is a consequence of the statements following (6.4), in particular, (6.5) and (6.6).

Since the model primitives $u_j^{(n)}$, $v_j^{(n)}$ and $L_j^{(n)}$ for $j \in \mathbb{N}$ and $n \in \mathbb{N}$ are all defined on $(\Omega, \mathcal{F}, \mathbb{P})$, so are the workload process $W^{(n)}$ and its scaled version $\widehat{W}^{(n)}$.

Corollary 6.3 implies that $\widehat{U}^{(n)}$ and $\widehat{W}^{(n)}$ have the same limit, and hence (6.21), the almost sure counterpart to Theorem 3.4, holds. $\square$

The assertion of Theorem 3.8 is that

$$\widehat{R}_W^{(n)} \Rightarrow K_+^*, \tag{6.24}$$

where K_+^* is the local time for U^* at $H(0)$ from (6.22). For $T < \infty$, define

$$\mathcal{Z}_n(T) \stackrel{\Delta}{=} \{\widehat{R}_U^{(n)}(t) = \widehat{R}_W^{(n)}(t), 0 \le t \le T\}. \tag{6.25}$$

From the workload evolution equations (4.69) and (4.70), it follows that if $\widehat{U}^{(n)}(t) = \widehat{W}^{(n)}(t)$ for $t \in [0, T]$, then $\widehat{R}_U^{(n)}(t) = \widehat{R}_W^{(n)}(t)$ for $t \in [0, T]$. Hence, by Corollary 6.3, we know that for every $T < \infty$, $\mathbb{P}(\mathcal{Z}_n(T)) \to 1$ as $n \to \infty$, which shows that the limits in distribution of $\widehat{R}_U^{(n)}$ and $\widehat{R}_W^{(n)}$, $n \in \mathbb{N}$, must coincide (if they exist). Further, since $\widehat{K}_+^{(n)} = \widehat{R}_U^{(n)}$ by Corollary 4.9, these must be equal to the limit in distribution of $\widehat{K}_+^{(n)}$, $n \in \mathbb{N}$. Hence, to complete the proof of Theorem 3.8, it suffices to show that

$$\widehat{K}_+^{(n)} \Rightarrow K_+^*. \tag{6.26}$$

For $n \in \mathbb{N}$ and $k \ge 1$, recall the definitions of $\tau_{k-1}^{(n)}$ and $\sigma_k^{(n)}$ given in (4.11) and (4.12), respectively, and define $\widehat{\tau}_{k-1}^{(n)} \stackrel{\Delta}{=} \frac{1}{n}\tau_{k-1}^{(n)}$ and $\widehat{\sigma}_k^{(n)} \stackrel{\Delta}{=} \frac{1}{n}\sigma_k^{(n)}$. Applying the heavy traffic scaling to (4.13), it is easy to see that for $t \ge 0$,

$$\begin{aligned}\widehat{K}_+^{(n)}(t) = \sum_{k\in\mathbb{N}}\Big[\widehat{W}_S^{(n)}(\widehat{\sigma}_k^{(n)}-) \vee \max_{s\in[\widehat{\sigma}_k^{(n)}, t\wedge\widehat{\tau}_k^{(n)}]} \widehat{\mathcal{W}}_S^{(n)}(s)(-\infty, 0] \\ - \widehat{W}_S^{(n)}(\widehat{\sigma}_k^{(n)}-)\Big].\end{aligned} \tag{6.27}$$

Keeping in mind the limits in (6.17) and (6.19), we introduce the related process

$$\widehat{Y}^{(n)}(t) \stackrel{\Delta}{=} \sum_{k\in\mathbb{N}}\Big[W_S^*(\widehat{\sigma}_k^{(n)}) \vee \max_{s\in[\widehat{\sigma}_k^{(n)}, t\wedge\widehat{\tau}_k^{(n)}]} (W_S^*(s) - H(0))^+ - W_S^*(\widehat{\sigma}_k^{(n)})\Big] \tag{6.28}$$

for $t \ge 0$, and denote the difference by

$$\varepsilon^{(n)}(t) \stackrel{\Delta}{=} \widehat{Y}^{(n)}(t) - \widehat{K}_+^{(n)}(t) \qquad \forall t \ge 0. \tag{6.29}$$

Then $\widehat{Y}^{(n)}$ is nondecreasing and continuous, and $\varepsilon^{(n)}$ is an RCLL process.

In the next two lemmas, we show that $\widehat{Y}^{(n)}$ increases only when U^* is at $H(0)$ and that the difference $\varepsilon^{(n)}$ between $\widehat{Y}^{(n)}$ and $\widehat{K}_+^{(n)}$ is negligible in heavy traffic. The main reason for introducing the sequence $\widehat{Y}^{(n)}$, $n \in \mathbb{N}$, is that it facilitates the proof of the former property.

LEMMA 6.7. *For every $n \in \mathbb{N}$, $\widehat{Y}^{(n)}$ and $\widehat{K}_+^{(n)}$ are constant on each interval $[\widehat{\tau}_{k-1}^{(n)}, \widehat{\sigma}_k^{(n)})$, $k \geq 1$. Moreover,*

$$\int_{[0,T]} \mathbb{I}_{\{U^*(t) < H(0)\}}\, d\widehat{Y}^{(n)}(t) = 0. \tag{6.30}$$

PROOF. Fix $n \in \mathbb{N}$. The first statement follows immediately from (6.27), (6.28), and the fact that the intervals $[\widehat{\tau}_{k-1}^{(n)}, \widehat{\sigma}_k^{(n)})$ and $[\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)})$, $k \geq 1$, form a disjoint covering of $[0, \infty)$. Now, fix $k \geq 1$ and let $J_k^{(n)}$ be the set of points $t \in [\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)})$ such that

$$W_S^*(\widehat{\sigma}_k^{(n)}) \leq \max_{s \in [\widehat{\sigma}_k^{(n)}, t]} \big(W_S^*(s) - H(0)\big)^+ = W_S^*(t) - H(0). \tag{6.31}$$

Since W_S^* is continuous, $J_k^{(n)}$ is closed, and so its complement in $[\sigma_k^{(n)}, \tau_k^{(n)})$ is the union of a countable number of open intervals, with possibly one half-open interval of the form $[\widehat{\sigma}_k^{(n)}, a)$ for some $a > \widehat{\sigma}_k^{(n)}$. From the explicit formula for $\widehat{Y}^{(n)}$ given in (6.28), it is easy to deduce that $\widehat{Y}^{(n)}$ is also constant on each such interval. Thus, to establish (6.30), it only remains to show that for each $k \geq 1$,

$$\int_{J_k^{(n)}} \mathbb{I}_{\{U^*(t) < H(0)\}}\, d\widehat{Y}^{(n)}(t) = 0. \tag{6.32}$$

Fix $t \in J_k^{(n)}$ and note that by the equality in (6.22) and the definition (6.4) of $\Lambda_{H(0)}$, we have $U^*(t) = W_S^*(t) - K^*(t)$, where

$$K^*(t) \triangleq \sup_{s \in [0,t]} \Big[\big(W_S^*(s) - H(0)\big)^+ \wedge \inf_{u \in [s,t]} W_S^*(u) \Big]. \tag{6.33}$$

Also, note that

$$\begin{aligned} \sup_{s \in [0, \widehat{\sigma}_k^{(n)})} \Big[\big(W_S^*(s) - H(0)\big)^+ \wedge \inf_{u \in [s,t]} W_S^*(u) \Big] &\leq \sup_{s \in [0, \widehat{\sigma}_k^{(n)})} \inf_{u \in [s,t]} W_S^*(u) \\ &\leq W_S^*(\widehat{\sigma}_k^{(n)}), \end{aligned}$$

and that the equality in (6.31) implies

$$\sup_{s \in [\widehat{\sigma}_k^{(n)}, t]} \Big[\big(W_S^*(s) - H(0)\big)^+ \wedge \inf_{u \in [s,t]} W_S^*(u) \Big] = W_S^*(t) - H(0).$$

Since $K^*(t)$ is equal to the maximum of the quantities on the left-hand side of the last two displays, we conclude that

$$K^*(t) \leq W_S^*(\widehat{\sigma}_k^{(n)}) \vee \big(W_S^*(t) - H(0)\big) = W_S^*(t) - H(0),$$

where the equality follows from the inequality in (6.31). This, when combined with the fact that $U^*(t) \in [0, H(0)]$, shows that $U^*(t) = W_S^*(t) - K^*(t) = H(0)$ for all $t \in J_k^{(n)}$, which proves (6.32). $\square$

We recall some standard definitions that will be used in the next lemma. Given $f \in \mathcal{D}[0,\infty)$ and $0 \le t_1 \le t_2 < \infty$, the *oscillation of f over* $[t_1, t_2]$ is

$$\mathrm{Osc}(f; [t_1, t_2]) \stackrel{\Delta}{=} \sup\{|f(t) - f(s)| : t_1 \le s \le t \le t_2\},$$

and the *modulus of continuity of f over* $[0, T]$ is

$$w_f(\delta; [0, T]) \stackrel{\Delta}{=} \sup\{|f(t) - f(s)| : 0 \le s \le t \le T, |t - s| \le \delta\}.$$

LEMMA 6.8. *As $n \to \infty$, $\varepsilon^{(n)} \xrightarrow{P} 0$.*

PROOF. Fix $T > 0$ and let $\eta > 0$ be arbitrarily small. By the Kolmogorov–Čentsov theorem (see, e.g., Theorem 2.8, page 53 of [17]), we can construct a positive, increasing deterministic function $\theta(\cdot)$ satisfying $\lim_{\delta \downarrow 0} \theta(\delta) = 0$ and majorizing the modulus of continuity $w_{W^*}(\cdot; [0, T])$ of the reflected Brownian motion W^* over $[0, T]$ on a set $\widetilde{\Omega}$ with $\mathbb{P}(\widetilde{\Omega}) \ge 1 - \eta$.

For each subsequence in $\mathbb{N}$, there is a sub-subsequence $\mathcal{S}$ along which the limits (6.17)–(6.21) hold $\mathbb{P}$-almost surely. We choose $\widetilde{\Omega}$ so that these limits hold along $\mathcal{S}$ for all $\omega \in \widetilde{\Omega}$.

In what follows, for $n \in \mathcal{S}$, we denote $\mathcal{Z}_n(T)$ simply by $\mathcal{Z}_n$, and evaluate all processes below at a fixed $\omega \in \mathcal{Z}_n \cap \widetilde{\Omega}$. Choose $\Delta < y_*/3$, and let $n_0 \in \mathcal{S}$ be such that for all $n \in \mathcal{S}$, $n \ge n_0$,

$$\sup_{t \in [0,T]} \left| \widehat{\mathcal{W}}_S^{(n)}(t)(-\infty, 0] - \left(W_S^*(t) - H(0)\right)^+ \right| \le \Delta, \tag{6.34}$$

$$\begin{aligned} \sup_{t \in [0,T]} \left| \widehat{W}_S^{(n)}(t-) - W_S^*(t) \right| &\le \Delta, \\ \sup_{t \in [0,T]} \left| \widehat{W}^{(n)}(t-) - W^*(t) \right| &\le \Delta, \end{aligned} \tag{6.35}$$

$$\sup_{t \in [0,T]} \left| \widehat{U}^{(n)}(t-) - U^*(t) \right| \le \Delta. \tag{6.36}$$

From the definitions (6.27) and (6.28), respectively, of $\widehat{K}_+^{(n)}$ and $\widehat{Y}^{(n)}$ it is clear that, for every $k \in \mathbb{N}$ such that $\tau_k^{(n)} \le T$,

$$\sup_{t \in [\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)}]} \left| \widehat{Y}^{(n)}(t) - \widehat{Y}^{(n)}(\widehat{\sigma}_k^{(n)}-) - \left(\widehat{K}_+^{(n)}(t) - \widehat{K}_+^{(n)}(\widehat{\sigma}_k^{(n)}-)\right) \right| \le 2\Delta.$$

Define

$$J_n \triangleq \{k \in \mathbb{N} : \widehat{K}_+^{(n)}(\widehat{\tau}_k^{(n)}) - \widehat{K}_+^{(n)}(\widehat{\sigma}_k^{(n)}-) > 0, \widehat{\tau}_k^{(n)} \le T\},$$

$$\widetilde{J}_n \triangleq \{k \in \mathbb{N} : \widehat{Y}^{(n)}(\widehat{\tau}_k^{(n)}) - \widehat{Y}^{(n)}(\widehat{\sigma}_k^{(n)}-) > 0, \widehat{\tau}_k^{(n)} \le T\},$$

and let $c^{(n)}$ be the cardinality of $J^{(n)} \cup \widetilde{J}^{(n)}$. Since $\widehat{K}_+^{(n)}$ and $\widehat{Y}^{(n)}$ are both constant on intervals of the form $[\widehat{\tau}_{k-1}^{(n)}, \widehat{\sigma}_k^{(n)})$, $k \ge 1$ (see Lemma 6.7), we have

$$\overline{\varepsilon}^{(n)}(T) \triangleq \sup_{s\in[0,T]} \left|\widehat{Y}^{(n)}(s) - \widehat{K}_+^{(n)}(s)\right| \le 2c^{(n)}\Delta. \tag{6.37}$$

We now claim that

$$k \in [J_n \cup \widetilde{J}_n] \quad \Rightarrow \quad \operatorname{Osc}(W^*, [\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)}]) \ge \frac{y_*}{3}. \tag{6.38}$$

We defer the proof of the claim and instead first show that the lemma follows from this claim. Let $\theta^{-1}(\cdot)$ denote the inverse of θ and define $M \triangleq T/\theta^{-1}(y_*/3) < \infty$. From the claim, we conclude that if $k \in [J_n \cup \widetilde{J}_n]$ then $\widehat{\tau}_k^{(n)} - \widehat{\sigma}_k^{(n)} \ge \theta^{-1}(y_*/3) > 0$, which in turn implies that $c^{(n)} \le M$. Substituting this into (6.37), we conclude that for every $\Delta > 0$, there exists $n_0(\Delta) \in \mathcal{S}$ such that for all $n \in \mathcal{S}$, $n \ge n_0(\Delta)$,

$$\mathbb{P}(\overline{\varepsilon}^{(n)}(T) > 2M\Delta) \le \mathbb{P}(\mathcal{Z}_n^c \cup \widetilde{\Omega}^c) \le \mathbb{P}(\mathcal{Z}_n^c) + \eta.$$

Taking limits as $n \to \infty$ through $\mathcal{S}$ and using the fact that $\mathbb{P}(\mathcal{Z}_n) \to 1$, we conclude that $\overline{\varepsilon}^{(n)}(T) \xrightarrow{P} 0$. We have shown that for each subsequence in $\mathbb{N}$, there is a subsubsequence along which $\overline{\varepsilon}^{(n)}(T) \xrightarrow{P} 0$. It follow that $\overline{\varepsilon}^{(n)}(T) \xrightarrow{P} 0$, where the limit is taken over all $n \in \mathbb{N}$, and this proves the lemma.

We now turn to the proof of the claim (6.38). Note first that by the definition of $H(0)$ and y_*, we have $H(0) \ge y_*$. If $k \in \widetilde{J}_n$, then Lemma 6.7 shows that $U^*(t) = H(0)$ for some $t \in [\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)})$. By the equality $\widehat{U}^{(n)}(\widehat{\sigma}_k^{(n)}-) = 0$ proved in Lemma 4.7 and (6.36), this implies that the oscillation of U^* on $[\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)})$ is no less than $H(0) - \Delta \ge y_*/3$. Since $W^* = U^*$, the conclusion in (6.38) holds.

Finally, suppose $k \in J_n$. Since $\widehat{K}_+^{(n)} = \widehat{R}_U^{(n)} = \widehat{R}_W^{(n)}$, we have

$$\widehat{R}_W^{(n)}(\widehat{\tau}_k^{(n)}) - \widehat{R}_W^{(n)}(\widehat{\sigma}_k^{(n)}-) > 0,$$

that is, the deadline of a customer in the reneging system expires during the unscaled time interval $[\sigma_k^{(n)}, \tau_k^{(n)}]$. Since

$$W^{(n)}(\sigma_k^{(n)}-) = 0 \tag{6.39}$$

[because $U^{(n)}(\sigma_k^{(n)}-) = 0$ and, by Lemma 5.2, $W^{(n)} \le U^{(n)}$], this customer must arrive during the interval $[\sigma_k^{(n)}, \tau_k^{(n)})$. Since his initial lead time is greater than or equal to $\sqrt{n}y_*$, there is a time $nt_0 \in [\sigma_k^{(n)}, \tau_k^{(n)})$ when this customer has lead

time exactly $\sqrt{n}y_*$. After time nt_0, this customer cannot be preempted by new arrivals, all of which have initial lead times greater than or equal to $\sqrt{n}y_*$. At time nt_0, the work that must be completed before this customer is served to completion is at most $\mathcal{W}^{(n)}(nt_0)(0,\sqrt{n}y_*]$. Since this customer becomes late, we must have $W(nt_0) \geq \mathcal{W}^{(n)}(nt_0)(0,\sqrt{n}y_*] > \sqrt{n}y_*$, or equivalently, $\widehat{W}^{(n)}(t_0) \geq \widehat{\mathcal{W}}^{(n)}(t_0)(0,y_*] > y_*$. By right continuity, $\widehat{W}^{(n)}((t_0+\nu)-) > y_*$ for some $\nu > 0$ so small that $t_0 + \nu \leq \widehat{\tau}_k^{(n)}$. From the second inequality in (6.35) and the fact that $\widehat{W}^{(n)}(\widehat{\sigma}_k^{(n)}-) = 0$ [the scaled version of (6.39)], we conclude that

$$W^*(t_0+\nu) - W^*(\widehat{\sigma}_k^{(n)}) \geq \frac{y_*}{3},$$

and this gives us the conclusion in (6.38). $\square$

PROOF OF THEOREM 3.8. Fix $T < \infty$. Let $\delta^{(n)} \triangleq U^* - \widehat{U}^{(n)}$, and let $\overline{\delta}^{(n)} \triangleq \sup_{s\in[0,T]}|U^*(s) - \widehat{U}^{(n)}(s)|$. According to (4.6) and (4.15),

$$U^{(n)} = W_S^{(n)} - K_+^{(n)} + K_-^{(n)}.$$

We scale this equation to obtain

$$U^* = \widehat{W}_S^{(n)} + \delta^{(n)} - \widehat{K}_+^{(n)} + \widehat{K}_-^{(n)} = \widehat{W}_S^{(n)} + \delta^{(n)} + \varepsilon^{(n)} - \widehat{Y}^{(n)} + \widehat{K}_-^{(n)}, \tag{6.40}$$

where [cf. (4.14)]

$$\widehat{K}_-^{(n)}(t) \triangleq -\sum_{k\in\mathbb{N}}\big[\big(\widehat{W}_S^{(n)}(\widehat{\tau}_{k-1}^{(n)}) - (\widehat{\sigma}_k^{(n)} \wedge t - \widehat{\tau}_{k-1}^{(n)})\big)^+ - \widehat{W}_S^{(n)}(\widehat{\tau}_{k-1}^{(n)})\big],$$

$\widehat{K}_+^{(n)}$ is defined by (6.27) and $\varepsilon^{(n)}$ is defined by (6.29). According to (4.16), $\int_0^T \mathbb{I}_{\{\widehat{U}^{(n)}(t)>0\}}\, d\widehat{K}_-^{(n)}(t) = 0$, which implies

$$\int_0^T \mathbb{I}_{\{U^*(t)>\overline{\delta}^{(n)}\}} d\widehat{K}_-^{(n)}(t) = 0. \tag{6.41}$$

Since $\widehat{W}_S^{(n)} + \delta^{(n)} + \epsilon^{(n)} \Rightarrow W_S^*$ due to (6.18), (6.20) and Lemma 6.8, and, by (6.22), U^* is obtained by applying the Skorokhod map on $[0, H(0)]$ to W_S^*, the convergence (6.26) is an immediate consequence of (6.40), (6.41), Lemmas 6.7, 6.8 and the invariance principle for reflected Brownian motions. However, since we are in a particularly simple setting here, we will provide a direct proof without invoking the general invariance principle.

We choose n_0 so that $\overline{\delta}^{(n_0)} < H(0)/3$ and recursively define stopping times $\rho_0 = 0$, and for $k \geq 1$,

$$\nu_k = \min\left\{t \geq \rho_{k-1}\middle| U^*(t) = \frac{2H(0)}{3}\right\}, \qquad \rho_k = \min\left\{t \geq \nu_k\middle| U^*(t) = \frac{H(0)}{3}\right\}.$$

Then $0 = \rho_0 < \nu_1 < \rho_1 < \nu_2 < \cdots$ and $\lim_{k\to\infty}\rho_k = \lim_{k\to\infty}\nu_k = \infty$. For $n \geq n_0$, $\widehat{K}_-^{(n)}$ is constant on each of the intervals $[\nu_k, \rho_k]$. Moreover, Lemma 6.7 implies

that for each k, $\widehat{Y}^{(n)}$ is constant on each of the intervals $[\rho_{k-1}, \nu_k]$. For $t \in [\nu_k, \rho_k]$, we have from (6.40), (6.18), (6.20) and Lemma 6.8 that

$$\begin{aligned}\widehat{Y}^{(n)}(t) - \widehat{Y}^{(n)}(\nu_k) &= \widehat{W}_S^{(n)}(t) - U^*(t) + \delta^{(n)}(t) + \varepsilon^{(n)}(t) \\ &\quad - \widehat{W}_S^{(n)}(\nu_k) + U^*(\nu_k) - \delta^{(n)}(\nu_k) - \varepsilon^{(n)}(\nu_k) \\ &\xrightarrow{P} W_S^*(t) - U^*(t) - \big(W_S^*(\nu_k) - U^*(\nu_k)\big).\end{aligned}$$

It follows that, uniformly for $t \in [0, T]$, $\widehat{Y}^{(n)}(t)$ converges in probability to

$$\sum_{k\in\mathbb{N}} \big[W_S^*\big((t \vee \nu_k) \wedge \rho_k\big) - U^*\big((t \vee \nu_k) \wedge \rho_k\big) - \big(W_S^*(\nu_k) - U^*(\nu_k)\big)\big]. \tag{6.42}$$

However, (6.23) implies that for each k, K_-^* is constant on $[\nu_k, \rho_k]$, and K_+^* is constant on $[\rho_{k-1}, \nu_k]$. Therefore, (6.22) implies that for $t \in [\nu_k, \rho_k]$,

$$K_+^*(t) - K_+^*(\nu_k) = W_S^*(t) - U^*(t) - \big(W_S^*(\nu_k) - U^*(\nu_k)\big).$$

This implies that the expression in (6.42) is $K_+^*(t)$. But $\widehat{Y}^{(n)}$ and $\widehat{K}_+^{(n)}$ have the same limit in probability because of Lemma 6.8, and we conclude that

$$\max_{t\in[0,T]} \big|\widehat{K}_+^{(n)}(t) - K_+^*(t)\big| \xrightarrow{P} 0. \tag{6.43}$$

Convergence in probability implies weak convergence, and we have (6.26). □

7. Performance evaluation and simulations. We use the heavy traffic approximations of this paper to evaluate the performance of the system with reneging and compare this to the system in which all customers are served to completion. The predictions of the theory, derived in Section 7.1 and compared to simulations in Section 7.2, are predicated on the assumption that one can interchange the limit as $n \to \infty$ and the limit as time goes to infinity of the fraction of reneged work. A formal proof would require a coupling argument such as that found in [37]. The simulation results attest to the accuracy of the approximations derived in Section 7.1 and also show the great difference in performance between the reneging and nonreneging systems.

7.1. *Derivation of theory predictions.* We derive formulas (1.1)–(1.7). We begin with one of the main results of this paper, Theorem 3.4, which states that the limiting scaled workload in the reneging system is a reflected Brownian motion in $[0, H(0)]$ with drift. More specifically,

$$W^*(t) = W_S^*(t) - K_+^*(t) + K_-^*(t), \tag{7.1}$$

where $W_S^*(t)$ is a reflected Brownian motion on $[0, \infty)$ with variance $\sigma^2 = \lambda(\alpha^2 + \beta^2)$ per unit time and drift $-\gamma$, K_-^* is the nondecreasing process starting at $K_-^*(0) = 0$ that grows only when $W^* = 0$, and K_+^* is the nondecreasing

process starting at $K_+^*(0) = 0$ that grows only when $W^* = H(0)$. We further saw in Theorem 3.8 that $K_+^*(t)$ is the limit of the scaled workload that reneges prior to time t in the diffusion scaling, that is, $\sqrt{n}K_+^*(t)$ is approximately the (unscaled) workload that reneges in the nth system prior to time nt.

LEMMA 7.1 ([15], Proposition 5, page 90). *We have*

$$\lim_{t\to\infty}\frac{1}{t}K_+^*(t) = \lim_{t\to\infty}\frac{1}{t}\mathbb{E}K_+^*(t) = \begin{cases} \dfrac{\gamma}{e^{2\gamma H(0)/\sigma^2}-1}, & \text{if } \gamma \neq 0, \\ \dfrac{\sigma^2}{2H(0)}, & \text{if } \gamma = 0. \end{cases} \tag{7.2}$$

PROOF. The first equality in (7.2) is a consequence of the fact that W^* has a stationary distribution [see (7.5) below]. For the proof of the second equality, recall that W_S^* has the decomposition (2.16). Let f be a C^2 function. Applying Itô's formula to $f(W^*(t))$ and taking expectations, we obtain

$$\begin{aligned} &f'(0)\mathbb{E}[I_S^*(t) + K_-^*(t)] - f'(H(0))\mathbb{E}K_+^*(t) \\ &\quad = \mathbb{E}\int_0^t \left[\gamma f'(W^*(s)) - \frac{1}{2}\sigma^2 f''(W^*(s))\right] ds \\ &\qquad + \mathbb{E}f(W^*(t)) - f(0). \end{aligned} \tag{7.3}$$

Taking $f(x) = x$, we obtain $\mathbb{E}[I_S^*(t) + K_-^*(t)] - \mathbb{E}K_+^*(t) = \gamma t + \mathbb{E}W^*(t) - f(0)$. If $\gamma \neq 0$, we may take $f(x) = \frac{\sigma^2}{2\gamma}e^{2\gamma x/\sigma^2}$ in (7.3), which leads to the equation $\mathbb{E}[I_S^*(t) + K_-^*(t)] - e^{2\gamma H(0)/\sigma^2}\mathbb{E}K_+^*(t) = \frac{\sigma^2}{2\gamma}(\mathbb{E}e^{2\gamma W^*(t)/\sigma^2} - 1)$. Solving these equations for $\mathbb{E}K_+^*(t)$, we obtain the second equality in (7.2) for $\gamma \neq 0$. To obtain this equality for $\gamma = 0$, we take $f(x) = x^2$. □

According to (2.12), the work that arrives to the nth system by time nt is $V^{(n)}(A^{(n)}(nt)) = \sqrt{n}\widehat{N}^{(n)}(t) + nt$. But, $\widehat{N}^{(n)}$ is approximately N^*, and hence

$$\lim_{t\to\infty}\frac{\sqrt{n}\widehat{N}^{(n)}(t) + nt}{nt} \approx \lim_{t\to\infty}\frac{\sqrt{n}N^*(t) + nt}{nt} = \left(1 - \frac{\gamma}{\sqrt{n}}\right).$$

Therefore, if $\gamma \neq 0$, the long-run fraction of reneged work is approximately

$$\begin{aligned} \lim_{t\to\infty}\frac{\sqrt{n}K_+^*(t)}{V^{(n)}(A^{(n)}(nt))} &= \frac{1}{\sqrt{n}}\lim_{t\to\infty}\frac{1}{t}K_+^*(t)\cdot\lim_{t\to\infty}\left(\frac{\sqrt{n}\widehat{N}^{(n)}(t) + nt}{nt}\right)^{-1} \\ &\approx \frac{\gamma/\sqrt{n}}{(1-\gamma/\sqrt{n})(e^{2\gamma H(0)/\sigma^2}-1)}. \end{aligned}$$

Finally, (2.4) implies that the expected lead time in the nth system is $\mathbb{E}L_j^{(n)} = \int_0^\infty (1 - G(y/\sqrt{n}))\,dy = \sqrt{n}H(0)$. Using this formula and (2.10), we conclude that

the fraction of work that reneges in the nth system when $\gamma \neq 0$ is approximately

$$\frac{1-\rho^{(n)}}{\rho^{(n)}(e^{2(1-\rho^{(n)})\mathbb{E}L_j^{(n)}/\sigma^2}-1)} = \frac{1-\rho^{(n)}}{\rho^{(n)}(e^{\theta\overline{D}}-1)}, \tag{7.4}$$

where

$$\theta = \frac{2(1-\rho^{(n)})}{\sigma^2} \approx \frac{2\gamma}{\sqrt{n}\sigma^2}, \qquad \overline{D} = \mathbb{E}L_j^{(n)} = \sqrt{n}H(0).$$

We have suppressed the dependence of θ and $\overline{D}$ on n, which will remain fixed. If $\gamma = 0$, then in place of (7.4) we have $\frac{\sigma^2}{2\overline{D}}$. We have established (1.1) and (1.2).

REMARK 7.2. Corollary 3.7 also implies that the limiting scaled queue length process is λW^*, which is a doubly reflected Brownian motion in $[0, \lambda H(0)]$ with drift $-\gamma\lambda$ and variance per unit time $\lambda^2\sigma^2$. This incorrectly suggests that $\lambda\sqrt{n}K_+^*(t)$ is approximately the number of customers who renege in the nth system prior to nt. The simulations indicate that this naive interpretation of Corollary 3.7 applied to the queue length process is incorrect, as does the following heuristic.

According to [15], Proposition 5, page 90, if $\gamma \neq 0$, the stationary density for W^* is

$$\varphi^*(x) \triangleq \begin{cases} \dfrac{2\gamma e^{-2\gamma x/\sigma^2}}{\sigma^2(1-e^{-2\gamma H(0)/\sigma^2})}, & \text{if } 0 \le x \le H(0), \\ 0, & \text{otherwise,} \end{cases} \tag{7.5}$$

whereas the stationary density is uniform on $[0, H(0)]$ if $\gamma = 0$. Therefore, for $\gamma \neq 0$ and t large, the density of $W^{(n)}(nt) \approx \sqrt{n}W^*(t)$ is approximately

$$\varphi(w) = \frac{1}{\sqrt{n}}\varphi^*(w/\sqrt{n}) = \begin{cases} \dfrac{\theta e^{-\theta w}}{1-e^{-\theta\overline{D}}}, & \text{if } 0 \le w \le \overline{D}, \\ 0, & \text{otherwise.} \end{cases}$$

We have suppressed the dependence of φ on n.

Suppose now that the lead times of arriving customers are not random. Then in the nth system, all lead times are equal to $\sqrt{n}H(0) = \overline{D}$. In this case, the EDF policy serves customers in order of arrival (FIFO). Suppose the workload in queue is W at the time of arrival of a customer whose service requirement is V. Recall that the expected service time is $1/\mu^{(n)}$, and because n is fixed, we suppress it and write $\mathbb{E}V = 1/\mu$. The arriving customer will be served to completion if and only if $W + V \le \overline{D}$. Suppose further that the arrival process $A^{(n)}$ is Poisson, so that according to the PASTA property ("Poisson arrivals see time averages"; see [1], Theorem 6.7, page 218), an arriving customer will encounter a workload W having

approximately the distribution φ. The probability the arriving customer eventually reneges is thus

$$\mathbb{P}\{W > \overline{D} - V\} = \mathbb{E}[\mathbb{P}\{W > \overline{D} - V | V\}] = \mathbb{E}\left[\int_{(\overline{D}-V)^+}^{\overline{D}} \varphi(w)\, dw\right].$$

Because $\overline{D}$ is of order $\sqrt{n}$ and V is of order 1, we have $(\overline{D} - V)^+ = \overline{D} - V$ with high probability. Using this approximation, we complete the calculation for the case $\gamma \neq 0$ to obtain

$$\mathbb{P}\{\text{Customer reneges}\} \approx \frac{1}{e^{\theta \overline{D}} - 1}(\mathbb{E}e^{\theta V} - 1). \tag{7.6}$$

If the customer reneges, then work $V + W - \overline{D} > 0$ is lost. The expected lost work is

$$\mathbb{E}[V + W - \overline{D} | \text{Customer reneges}] \approx \mathbb{E}\left[\frac{\int_{(\overline{D}-V)^+}^{\overline{D}} (V + w - \overline{D})\varphi(w)\, dw}{\mathbb{P}\{\text{Customer reneges}\}}\right].$$

Again using the approximation $(\overline{D} - V)^+ \approx \overline{D} - V$, we obtain

$$\begin{aligned}\mathbb{E}[V + W - \overline{D} | \text{Customer reneges}] &\approx \frac{1}{\theta} - \frac{\mathbb{E}V}{\mathbb{E}e^{\theta V} - 1} \\ &\approx \frac{1}{\theta} - \frac{\mathbb{E}V}{\theta \mathbb{E}V + 1/2\theta^2 \mathbb{E}[V^2] + O(n^{-3/2})} \\ &\approx \frac{\mathbb{E}[V^2]}{2\mathbb{E}V}.\end{aligned}$$

The last expression is, perhaps not surprisingly, the formula for the average residual lifetime of a renewal cycle (see [32], Example 3.6(b), pages 80 and 81). Consequently, when lead times are constant and the arrival process is Poisson, we should expect the total number of customers reneging in $[0, t]$ times the expected amount of work lost per reneging customer to approximately equal the total amount of work lost by reneging in $[0, t]$. If we divide both by the total number of customer arrivals in $[0, t]$ and take limits as $t \to \infty$, we find

$$\begin{aligned}&\text{Fraction of lost customers in reneging system} \\ &\quad \approx \frac{\text{Fraction of lost work in reneging system}}{\mathbb{E}[V + W - \overline{D} | \text{Customer reneges}]} * \mathbb{E}V \\ &\quad \approx \frac{2(\mathbb{E}V)^2}{\mathbb{E}[V^2]} \times (\text{Fraction of lost work in reneging system}).\end{aligned} \tag{7.7}$$

This is (1.7) with $\mathbb{E}V = \frac{1}{\mu}$ and $E[V^2] = \beta^2 + \frac{1}{\mu^2}$.

If V is exponentially distributed, hence $\mathbb{E}[V^2] = 2(\mathbb{E}V)^2$, then (7.7) implies that the fraction of customers who renege will be approximately the fraction of work

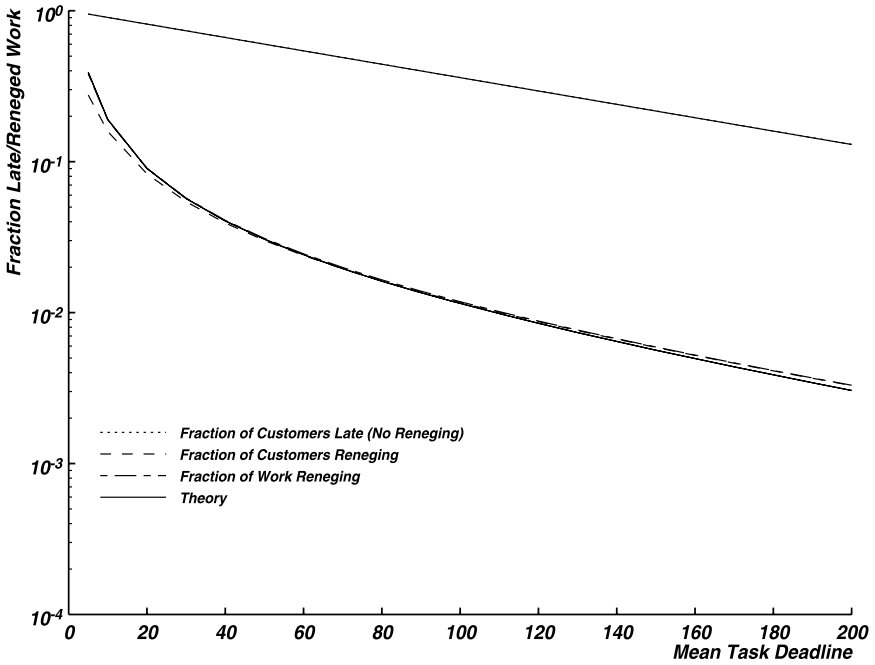

FIG. 1. *$M/M/1$ queue.*

that reneges. See Figure 1 for simulations that confirm this assertion. On the other hand, if V is nonrandom, hence equal to its mean $1/\mu$, then (7.7) predicts that the fraction of customers who renege will be twice the fraction of work that reneges. See Figure 2 for simulations that confirm this assertion. Both these conclusions hold irrespective of the value of λ.

The last conclusion is inconsistent with a naive interpretation of Corollary 3.7, according to which work reneges at a rate $1/\lambda$ times the rate of customer reneging. Since work arrives at a rate $\mathbb{E}V \approx 1/\lambda$ times the rate of customer arrivals, this naive interpretation of Corollary 3.7 would say that the fraction of work reneging would approximately agree with the fraction of customers reneging regardless of the distribution of V.

We next turn our attention to the performance of the standard (nonreneging) system. Recall from (2.15) that the scaled workload process when all customers are served to completion converges to W_S^*, a reflected Brownian motion with drift $-\gamma$ (we now assume $\gamma > 0$ in order to have a stationary distribution) and variance σ^2. In particular, $W_S^{(n)}(nt) \approx \sqrt{n} W_S^*(t)$. The stationary density for W_S^* is

$$\varphi_S^*(x) \triangleq \begin{cases} \dfrac{2\gamma}{\sigma^2} e^{-2\gamma x/\sigma^2}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

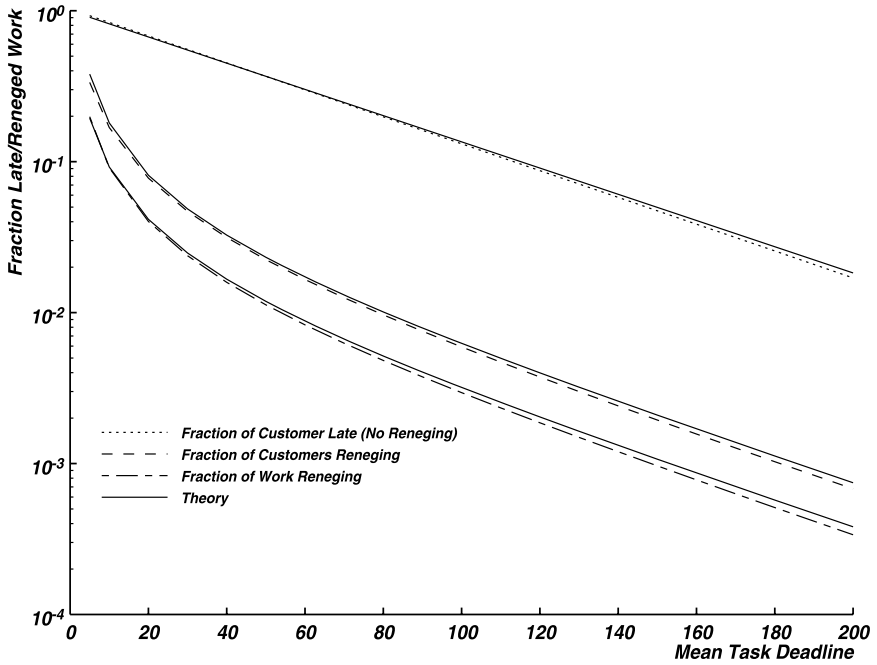

FIG. 2. *$M/D/1$ queue.*

and so for large t, the density of $W^{(n)}(nt)$ is approximately

$$\varphi_S(w) = \frac{1}{\sqrt{n}}\varphi_S^*\big(w/\sqrt{n}\big) = \begin{cases} \theta e^{-\theta w}, & \text{if } w \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Consequently, the long-run fraction of time $W^{(n)}$ spends above level $\overline{D}$ is $e^{-\theta\overline{D}}$. The workload level at which the limiting frontier reaches 0 is $H(0)$, and hence it is approximately the case that the nth system sees lateness if and only if $W^{(n)}$ exceeds $\overline{D} = \sqrt{n}H(0)$. In other words, the theory predicts that

$$\begin{aligned} &\text{Fraction of late customers in standard system} \\ &\quad = \text{Fraction of late work in standard system} \\ &\quad = e^{-\theta\overline{D}}. \end{aligned} \tag{7.8}$$

We are using here the result for $GI/G/1$ queues that

$$\begin{aligned} &\lim_{n\to\infty}\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{I}_{\{\widehat{W}_S^{(n)}(t)>H(0)\}}\,dt \\ &\quad = \lim_{T\to\infty}\lim_{n\to\infty}\frac{1}{T}\int_0^T \mathbb{I}_{\{\widehat{W}_S^{(n)}(t)>H(0)\}}\,dt \\ &\quad = \lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{I}_{\{W_S^*(t)>H(0)\}}\,dt, \end{aligned}$$

a result that grows out of the work of Kingman [19, 20] (see [10] for a general result that specializes to the case under consideration).

It is important to compare the fraction of work that reneges in the reneging system, given by (7.4), with the fraction of work that is late in the standard (nonreneging) system. The ratio of these quantities of lost/late work is

$$\frac{\text{Lost work in reneging system}}{\text{Late work in standard system}} \approx \frac{e^{\theta \overline{D}}}{e^{\theta \overline{D}} - 1}\left(\frac{1-\rho^{(n)}}{\rho^{(n)}}\right). \tag{7.9}$$

The parameter θ is $O(1/\sqrt{n})$, $\theta\overline{D}$ is $O(1)$, and $1-\rho^{(n)}$ is $O(1/\sqrt{n})$. Thus the ratio in (7.9) is $O(1/\sqrt{n})$.

REMARK 7.3. If lead times are a nonrandom constant $\overline{D}$, EDF reduces to first-in-first-out, and the fraction of lost customers in an M/G/1 queue with $0 < \rho < 1$ is $(1-\rho)\mathbb{P}\{W > \overline{D}\}/(1-\rho\mathbb{P}\{W > \overline{D}\})$, where W is the steady-state workload in the corresponding nonreneging M/G/1 queue (see [3]). In the heavy traffic limit of our model, $\mathbb{P}\{W > \overline{D}\} = e^{-\theta\overline{D}}$ [see the derivation of (7.8)]. Recalling that $1-\rho = O(1/\sqrt{n})$ in (1.1), we observe that this is consistent with (1.1).

7.2. *Simulation results*. We conducted a simulation study to assess the accuracy of these approximations and to compare the performance of the systems with and without reneging. Two systems were considered, an $M/M/1$ system presented in Figure 1 and an M/D/1 system presented in Figure 2. In both cases, $\lambda = 0.5$ and $\frac{1}{\mu} = 1.96$, and so the traffic intensity is $\rho = 0.98$. These parameter values result in $\theta = 0.010202$ for the $M/M/1$ case and $\theta = 0.02$ for the M/D/1 case. The initial deadline distribution is uniform on $[5, B]$ with the mean deadline $\overline{D} = \frac{5+B}{2}$, varying from $B = 5$ (constant deadlines) to $B = 200$. The data points are the simulation results averaged over one billion customer arrivals per case. The curves that are superimposed on the data are the theoretical values, $e^{-\theta\overline{D}}$ for the case in which customers are served to completion (the standard system), and equations (1.1) and (1.7) for the fraction of work lost and the fraction of lost customers for the reneging system. Equation (1.7) is derived in Remark 7.2 under the assumption of constant deadlines. Nevertheless, we apply it for the variable deadline case in the simulation study. The fraction of late work or late customers for the system in which customers are served to completion is also presented to compare its performance with that of the reneging system.

The $M/M/1$ results are presented in Figure 1 with the fraction of customers missing their deadlines, the fraction of customers reneging, and the fraction of work reneging plotted on a log scale on the y-axis against the mean deadline on the x-axis. There is nearly perfect agreement between the theoretical approximation and the simulation. In fact, one cannot see the plot of "Fraction of Customers Late (No Reneging)" because it coincides with the "Theory" plot at the top of the figure. Similarly, one can see only parts of the plots of "Fraction of Customers Reneging"

and "Fraction of Work Reneging" because they coincide with the "Theory" plot in the middle of the figure. One can see the linear form for the case of service to completion. Furthermore, the simulation confirms the prediction of (1.1)–(1.4) that for sufficiently large values of $\overline{D}$, the performance of the reneging system is parallel on a log scale to that of the standard system with the two curves separated by approximately 0.02. This corresponds to a reduction in work that misses its deadline by a factor of 40 to 50.

Figure 2 presents the results for the M/D/1 system. The results are qualitatively identical to those of Figure 1, except the fits of the theoretical curves are not as exact as the fits for the $M/M/1$ system; it appears that now the value $\theta = 0.02$ is slightly too small and hence the theory slightly overestimates the fraction of work that misses its deadline, especially when the mean deadline is large. Also, the lost or late work and the customer loss or lateness fractions are significantly smaller than for the $M/M/1$ system owing to the reduction in variability of the customer service time distribution. The reduction in missed deadlines between the two systems for large values of $\overline{D}$ is again a factor of 40 to 50. In both figures, it is clear that there are one to two orders of magnitude of improvement in the overall performance of the system resulting from stopping service on customers when their deadlines expire.

APPENDIX: OPTIMALITY OF EDF

PROOF OF THEOREM 5.1. Let π be a service policy and let t_0 be the first time π deviates from the EDF policy, either because it idles when there is work present, or it serves a customer other than the customer present with the smallest lead time. Let j be the index of the customer with the smallest lead time at time t_0.

We consider first the case that π idles at time t_0. In this case, we define $\rho(\pi)$ to be the policy that emulates π except as noted below. From time t_0, whenever π idles, $\rho(\pi)$ serves customer j, at least until time t_1, when customer j leaves the $\rho(\pi)$ system because either $\rho(\pi)$ serves customer j to completion or else the deadline of customer j elapses. From time t_1, $\rho(\pi)$ idles if π serves customer j. We will show that for $t \geq 0$,

$$R_{\rho(\pi)}(t) \leq R_\pi(t). \tag{A.1}$$

Let $v_k(t)$ [resp., $v_k^\rho(t)$] be the residual service time of the kth customer at time t under π [resp., $\rho(\pi)$]. In particular, if d_k is the deadline of the kth customer, then $v_k(d_k-)$ [resp., $v_k^\rho(d_k-)$] is the work corresponding to this customer that is deleted by π [resp., $\rho(\pi)$] due to lateness, and

$$R_{\rho(\pi)}(t) = \sum_{k\,:\,d_k \leq t} v_k^\rho(d_k-), \qquad R_\pi(t) = \sum_{k\,:\,d_k \leq t} v_k(d_k-). \tag{A.2}$$

By the definition of $\rho(\pi)$, for $t \geq 0$ and $k \neq j$, we have

$$v_k^\rho(t) = v_k(t), \tag{A.3}$$

whereas

$$v_j^{\rho}(t) \le v_j(t). \tag{A.4}$$

Summing (A.3) over $k \neq j$, invoking (A.4) and (A.2), we obtain (A.1).

We next consider the case that at time t_0, π serves customer $i \neq j$. In this case, we define $\rho(\pi)$ to be the policy that emulates π except as noted below. From time t_0, whenever π serves customer i, $\rho(\pi)$ serves customer j, at least until time t_1, when $\rho(\pi)$ serves customer j to completion or the deadline of customer j elapses. From time t_1, $\rho(\pi)$ serves customer i if π serves customer j, provided customer i is present in the system under $\rho(\pi)$. If π serves customer j and customer i is not present under $\rho(\pi)$, then $\rho(\pi)$ idles. We again have (A.2) and (A.4), whereas (A.3) now holds only for $k \notin \{i, j\}$. If the ith customer is served to completion under $\rho(\pi)$, then $v_i^{\rho}(d_i-) = 0$, and (A.3) for $k \notin \{i, j\}$, and (A.4) imply that (A.1) holds for all t. It remains to consider the case that the ith customer becomes late under $\rho(\pi)$. In this case (A.3) for $k \notin \{i, j\}$ and (A.4) imply that (A.1) holds for $t \in [0, d_i)$. Let w_1 denote the work done by $\rho(\pi)$ on the jth customer when π works on the ith customer in the interval $[t_0, t_1)$. Let w_2 be the work done by $\rho(\pi)$ on customer i in the time interval $[t_1, \infty)$ while π works on customer j in this time interval. Finally, let w_3 be the work done by π on customer j in the time interval $[t_1, \infty)$ while $\rho(\pi)$ is idle. Then $v_j^{\rho}(d_j-) + w_1 = v_j(d_j-) + w_2 + w_3$ and $v_i^{\rho}(d_i-) + w_2 = v_i(d_i-) + w_1$, which implies

$$v_j^{\rho}(d_j-) + v_i^{\rho}(d_i-) = v_j(d_j-) + v_i(d_i-) + w_3. \tag{A.5}$$

We argue by contradiction that w_3 cannot be positive. If w_3 were positive, then at some time $t \ge t_1$, π serves customer j and customer i is not in the $\rho(\pi)$ system. This implies that $d_j > t$, and since by assumption, $d_i > d_j$, the absence of customer i in the $\rho(\pi)$ system means that this system has served customer i to completion. We conclude that $v_i^{\rho}(d_i-) = 0$. On the other hand, customer j is also not in the $\rho(\pi)$ system at time $t \ge t_1$, and so $v_j^{\rho}(d_j-) = 0$ as well. The left-hand side of (A.5) is zero, and hence w_3 must be zero. We conclude that

$$v_j^{\rho}(d_j-) + v_i^{\rho}(d_i-) = v_j(d_j-) + v_i(d_i-). \tag{A.6}$$

Since $d_j < d_i$, if $t \ge d_i$, then (A.3) for $k \notin \{i, j\}$ and (A.6) imply (A.1).

Starting from the service policy π, we have obtained a service policy $\rho(\pi)$ that either is work conserving until the departure of customer j or else gives customer j priority over customer i until the departure of customer j. However, immediately after time t_0, the policy π may serve some customer $k \notin \{i, j\}$, and hence $\rho(\pi)$ also serves k at this time, although customer j is more urgent. Therefore, we apply n iterations of the mapping ρ, where n is the number of customers in the π system at time t_0, and thereby obtain a policy that is work-conserving and serves in EDF order at least until the first time after t_0 that there is a departure or an arrival. We have $R_{\rho^n(\pi)}(t) \le R_{\pi}(t)$ for all $t \ge 0$.

By assumption, for each t the number of system arrivals $A(t)$ by time t is finite. Hence the maximum number of customers in the system over the interval $[0, t]$ is bounded by $A(t)$, and the number of arrivals and departures up to time t is bounded by $2A(t)$, irrespective of the service policy. Thus, if we start with any policy π, the number of iterations of the mapping ρ required to obtain a policy that is work conserving and serves in EDF order up to time t is finite. Under this policy the amount of work removed by lateness up to time t is the same as for the EDF system in the theorem, and hence (5.1) holds. □

REMARK A.1. In the above proof we have implicitly assumed that π [and thus $\rho(\pi)$] never serves more than one customer at the same time. This assumption simplifies the exposition of the argument, and the generality of Theorem 5.1 is sufficient for this paper. However, the proof can be generalized to policies permitting simultaneous service of customers (e.g., processor sharing). In this case, in the construction of $\rho(\pi)$ we must additionally take the rates at which customers receive service into account. For example, the difference in the rates with which the jth customer receives service under $\rho(\pi)$ and π in the time interval $[t_0, t_1)$ must be equal to the rate of service of the ith customer under π in this time interval, the rates of service of all other customers in this time interval under π and $\rho(\pi)$ must be the same, etc.

REFERENCES

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues*, 2nd ed. *Applications of Mathematics* (*New York*) **51**. Springer, New York. MR1978607

[2] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[3] BOOTS, N. and TIJMS, H. (1999). A multiserver queueing system with impatient customers. *Management Science* **45** 444–448.

[4] CHEN, H. and MANDELBAUM, A. (1991). Stochastic discrete flow networks: Diffusion approximations and bottlenecks. *Ann. Probab.* **19** 1463–1519. MR1127712

[5] DECREUSEFOND, L. and MOYAL, P. (2008). Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Process. Related Fields* **14** 131–158. MR2433299

[6] DOWN, D., GROMOLL, H. C. and PUHA, A. (2009). Fluid limits for shortest remaining processing time queues. *Math. Operations Research* **4** 880–911.

[7] DOYTCHINOV, B., LEHOCZKY, J. and SHREVE, S. (2001). Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11** 332–378. MR1843049

[8] DUPUIS, P. and RAMANAN, K. (1999). Convex duality and the Skorokhod problem. *Probab. Theory Related Fields* **115** 197–236.

[9] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes*: *Characterization and Convergence*. Wiley, New York. MR838085

[10] GAMARNIK, D. and ZEEVI, A. (2006). Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.* **16** 56–90. MR2209336

[11] GROMOLL, H. C. (2004). Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.* **14** 555–611. MR2052895

[12] GROMOLL, H. C. and KRUK, Ł. (2007). Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.* **17** 1049–1101. MR2326240

[13] GROMOLL, H. C., KRUK, L. and PUHA, A. Diffusion limits for shortest remaining processing time queues. Preprint, Dept. Mathematics, Univ. Virginia. Available at http://arxiv.org/abs/1005.1035.

[14] HARRISON, J. M. and REIMAN, M. Reflected Brownian motion in an orthant. *Ann. Probab.* **9** 302–308.

[15] HARRISON, J. M. (1985). *Brownian Motion and Stochastic Flow Systems*. Wiley, New York. MR798279

[16] IGLEHART, D. L. and WHITT, W. (1970). Multiple channel queues in heavy traffic. I. *Adv. in Appl. Probab.* **2** 150–177. MR0266331

[17] KARATZAS, I. and SHREVE, S. E. (1988). *Brownian Motion and Stochastic Calculus*. *Graduate Texts in Mathematics* **113**. Springer, New York. MR917065

[18] KASPI, H. and RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* **21** 33–114.

[19] KINGMAN, J. F. C. (1961). The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.* **57** 902–904. MR0131298

[20] KINGMAN, J. F. C. (1962). On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B* **24** 383–392. MR0148146

[21] KRUK, Ł. (2007). Diffusion approximation for a $G/G/1$ EDF queue with unbounded lead times. *Ann. Univ. Mariae Curie-Skłodowska Math. A* **61** 51–90.

[22] KRUK, Ł., LEHOCZKY, J. P. and SHREVE, S. (2003). Second order approximation for the customer time in queue distribution under the FIFO service discipline. *Ann. Univ. Mariae Curie-Skłodowska Sect. AI Inform.* **1** 37–48. MR2254131

[23] KRUK, Ł., LEHOCZKY, J. P., SHREVE, S. and YEUNG, S.-N. (2004). Earliest-deadline-first service in heavy-traffic acyclic networks. *Ann. Appl. Probab.* **14** 1306–1352. MR2071425

[24] KRUK, Ł., LEHOCZKY, J. P. and SHREVE, S. (2006). Accuracy of state space collapse for earliest-deadline-first queues. *Ann. Appl. Probab.* **16** 516–561. MR2244424

[25] KRUK, Ł., LEHOCZKY, J. P., RAMANAN, K. and SHREVE, S. E. (2007). An explicit formula for the Skorokhod map on $[0, a]$. *Ann. Probab.* **35** 1740–1768.

[26] KRUK, Ł., LEHOCZKY, J. P., RAMANAN, K. and SHREVE, S. (2008). Double Skorokhod map and reneging real-time queues. In *Markov Processes and Related Topics*: *A Festschrift for Thomas G. Kurtz. Inst. Math. Stat. Collect.* **4** 169–193. IMS, Beachwood, OH. MR2574231

[27] LEHOCZKY, J. P. (1996). Real-time queueing theory. In *Proc. of IEEE Real-Time Systems Symposium* 186–195. IEEE Computer Society Press, Los Alamitos, CA.

[28] LIMIC, V. (2000). On the behavior of LIFO preemptive resume queues in heavy traffic. *Electron. Comm. Probab.* **5** 13–27 (electronic). MR1736721

[29] PANWAR, S. S. and TOWSLEY, D. (1992). Optimality of the stochastic earliest deadline policy for the $G/M/c$ queue serving customers with deadlines. In *Second ORSA Telecommunications Conference*. ORSA (Operations Research Society of America), Baltimore, MD.

[30] PROKHOROV, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theory Probab. Appl.* **1** 157–214.

[31] RAMANAN, K. and REIMAN, M. I. (2003). Fluid and heavy traffic diffusion limits for a generalized processor sharing model. *Ann. Appl. Probab.* **13** 100–139. MR1951995

[32] ROSS, S. M. (1983). *Stochastic Processes*. Wiley, New York. MR683455

[33] WARD, A. R. and GLYNN, P. W. (2003). A diffusion approximation for a Markovian queue with reneging. *Queueing Syst.* **43** 103–128. MR1957808

[34] WARD, A. R. and GLYNN, P. W. (2005). A diffusion approximation for a $GI/GI/1$ queue with balking or reneging. *Queueing Syst.* **50** 371–400. MR2172907

[35] WHITT, W. (1971). Weak convergence theorems for priority queues: Preemptive-resume discipline. *J. Appl. Probab.* **8** 74–94. MR0307389

[36] WHITT, W. (2002). *Stochastic-Process Limits*: *An Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York. MR1876437

[37] ZHANG, J. and ZWART, B. (2008). Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Syst*. **60** 227–246. MR2461617

Ł. KRUK
DEPARTMENT OF MATHEMATICS
MARIA CURIE-SKLODOWSKA UNIVERSITY
LUBLIN
POLAND
AND
INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
WARSAW
POLAND
E-MAIL: lkruk@hektor.umcs.lublin.pl

J. LEHOCZKY
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: jpl@stat.cmu.edu

K. RAMANAN
DIVISION OF APPLIED MATHEMATICS
BROWN UNIVERSITY
PROVIDENCE, RHODE ISLAND 02912
USA
E-MAIL: Kavita_Ramanan@brown.edu

S. SHREVE
DEPARTMENT OF MATHEMATICAL SCIENCES
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: shreve@andrew.cmu.edu