Heavy Traffic Analysis for EDF Queues with Reneging

December 17, 2007

Łukasz Kruk[1]
Department of Mathematics
Maria Curie-Sklodowska University
Lublin, Poland
and
Institute of Mathematics
Polish Academy of Sciences
Warsaw, Poland
lkruk@hektor.umcs.lublin.pl

John Lehoczky[2]
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213, USA
jpl@stat.cmu.edu

Kavita Ramanan[3]
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
kramanan@math.cmu.edu

Steven Shreve[4]
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA, 15213, USA
shreve@cmu.edu

[1]Supported by the State Committee for Scientific Research of Poland, Grant No. 2 P03A 012 23.
[2]Supported in part by ONR and DARPA under MURI contract N00014-01-1-0576.
[3]Partially Supported by the National Science Foundation under Grants No. DMS-0406191, DMS-0405343.
[4]Partially supported by the National Science Foundation under Grant No. DMS-0404682.

Abstract

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, which are validated by simulation results.

Short title: EDF Queues with Reneging
Keywords: Due dates, heavy traffic, queueing, reneging, diffusion limits, random measures, real time queues
AMS subject classification (2000): Primary 60K25; Secondary 60G57, 60J65, 68M20

Contents

1 Introduction

In the last decade, substantial attention has been paid to queueing systems in which customers have deadlines. Common examples of such systems include telecommunication systems carrying digitized voice or video traffic, tracking systems and real-time control systems. In the case of voice or video, the packetized information must be received, processed and displayed within stringent timing bounds so that the integrity of the transmission is maintained. Similarly, there are processing requirements for tracking systems that guarantee that a track can be successfully followed. Real-time control systems (for example, 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 taken to be customer delay, queue length, or utilization of the service facility. We use the fraction of "reneged work," defined as the residual work not serviced due to elapsed deadlines, as our primary performance measure. To minimize this quantity, it is necessary to use a queueing discipline (or scheduling policy) that takes deadlines into explicit account. We use the Earliest-Deadline-First (EDF) policy, which reduces to the more standard 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 [28].

Heavy traffic analysis of real-time queues was initiated in the single queue case by Lehoczky [26]. This was put on a firm mathematical foundation in the paper by Doytchinov, Lehoczky and Shreve (DLS) [5]. The accuracy of heavy traffic approximations was developed in the papers by Kruk, Lehoczky, and Shreve [22, 24]. DLS was generalized to the case of acyclic networks by Kruk, Lehoczky, Shreve, and Yeung [23]. In all these papers it was assumed that all customers were served to completion whether or not they were late. The case in which customers leave the system and their residual work is lost when their deadlines elapse was not considered. This paper addresses that situation.

The mathematical formulation used by DLS and subsequent 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 will elapse. This is done through the use of measure-valued queue length and workload processes. At any time t, these processes are given by measures on the real line. The measure-valued queue length process puts unit mass at the lead time of each customer in the system, while the measure-valued workload process puts mass equal to the re-

maining 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 drifted reflected Brownian motion. DLS showed that the suitably scaled workload and queue length measure-valued processes converge to an explicit deterministic mapping of the workload process, the form of which is determined by the initial lead time distribution of the customers.

In this paper we consider the case in which customers are not served to completion, but are made to leave the system when their deadlines elapse, that is, when their lead-times reach zero, which we refer to as reneging. The system with reneging can be expected to lead to a marked improvement in performance over the policy used in the DLS system, in the sense that the fraction of reneged work in this system would be expected to be less than the fraction of work that becomes late in the DLS system. Indeed, it is clear that the additional processing time spent on a task that has already missed its deadline would be better applied to customers who are not yet late. For these reasons, it is important to consider the simple control strategy in which processing is stopped (or never initiated) on any task whose deadline has elapsed, and such a task is removed from the system. Due to the preemptive nature of the EDF policy, it is not possible to determine with certainty at the point of admission whether or not a customer will be fully serviced before his deadline elapses. It is thus natural to have the controller make the decision only at the time when the deadline elapses.

The analysis of the system with reneging turns out to be considerably more complicated than that of the DLS system. Indeed, in the reneging system, the evolution of even the scalar total workload process depends on the entire lead time distribution of customers in queue as well as the nature of the EDF scheduling 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, reflecting the absence of late work in the reneging system. Moreover,

using the heavy traffic approximations, we provide formulas for the fraction of lost work in a heavily loaded reneging system which show good agreement with simulations. We also compare this quantity with the fraction of late work in the DLS system in order to quantify the gains achieved by the use of a controller in the reneging system.

Measure-valued processes have recently gained prominence in queueing theory. In a situation closely related to this paper, Decreusefond and Moyal [7] use measure-valued processes to obtain the fluid limit of an 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. Measure-valued processes have also proved useful in the heavy traffic analysis of single-server queues with scheduling disciplines other than EDF such as last-in-first-out (LIFO) [27] and processor sharing [11, 12]. 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 greatly facilitated by the use of representations in terms of continuous mappings on $\mathbb{R}^n$ [6, 8, 14, 30, 35]. This work demonstrates, in particular, 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).

This paper is organized as follows. Section 2 introduces the model, the assumptions and the notation. Section 3 summarizes the main results of the paper, 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. This section also presents a proof of optimality of EDF that may be of independent interest. Finally, Section 7 presents simulation results that compare the behavior of the reneging system and the system in which all customers are served to completion, and also discusses some open questions.

2 The Model, Assumptions and Notation

2.1 Notation

The following notation will be used throughout the paper. Let $\mathbb{R}$ denote the set of real numbers. For $a, b \in \mathbb{R}$, we write $a \vee b$ for the maximum of a and b, $a \wedge b$ for the minimum of a and b, and a^+ for the maximum of a and 0. Also, $\inf\{\emptyset\}$ should be understood as $+\infty$, while $\sup\{\emptyset\}$ and $\max\{\emptyset\}$ should be understood as $-\infty$. Moreover, if $a < b$, then the interval $[b, a]$ is understood to be $\emptyset$.

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 $x \in \mathbb{R}$, i.e., the Dirac measure at x, by δ_x. For notational convenience, 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 $D_X[0, \infty)$ (respectively, $D_X[0, T]$) to denote the space of right-continuous functions with left-hand limits (RCLL functions) from $[0, \infty)$ (respectively, $[0, T]$) 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, i.e., $\triangle x(t) \triangleq x(t) - x(t-)$. Lastly, 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

$$\lim_{n \to \infty} \mathbb{E}_n[f(Z_n)] = \mathbb{E}[f(Z)],$$

for every bounded continuous function f on $D_X[0, \infty)$. 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 customer arrival process 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 $\beta^{(n)}$. For simplicity, we assume that each queue is empty at time zero.

We define the *customer arrival times*

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

the *customer arrival process*

$$A^{(n)}(t) \triangleq \max\{k; S_k^{(n)} \leq t\}, \quad t \geq 0, \tag{2.2}$$

and the *work arrival process*

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

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 $0 < y_* \leq y^* < +\infty$ (see Remark 3.8 for a discussion of this assumption). 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.

We assume that customers are served using the Earliest-Deadline-First (EDF) queue discipline, i.e., the server always serves the customer with the shortest lead time. 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 j–th customer is still present in the system (either waiting for service or receiving it) when his deadline passes, i.e., 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 who regards late jobs as worthless and removes them from the system.

We define $W^{(n)}(t)$, the *observed workload process* (or just the *workload*) in the system at time t, as the remaining processing time of all the customers in the system at this time. In other words, $W^{(n)}(t)$ is the amount of time necessary to serve all the customers present in the system at time t to completion, without taking new arrivals and departures due to elapsing deadlines into account. We define $R_W^{(n)}(t)$ be the 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 will also use a sequence, indexed by superscript (n), of *standard EDF systems*, with the same stochastic primitives as in the EDF systems with reneging. In each of these standard systems, the server always serves the customer with the shortest lead time, preemption occurs when a customer more urgent than the customer in service arrives (we assume preempt-resume) and there is no overhead, but late customers (customers with negative lead times) stay in queue 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 non-idling 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*

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

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 much more complex and depends not only on the residual service times but also on the lead times of all customers in the queue, and also depends on the precise nature of the EDF scheduling discipline. Our analysis of the reneging system will be facilitated by results from [5] on the heavy traffic analysis of the standard EDF system.

2.4 Heavy traffic assumptions

We assume that the following limits exist:

$$\lim_{n\to\infty} \lambda^{(n)} = \lambda,\ \lim_{n\to\infty} \mu^{(n)} = \lambda,\ \lim_{n\to\infty} \alpha^{(n)} = \alpha,\ \lim_{n\to\infty} \beta^{(n)} = \beta, \tag{2.9}$$

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*

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

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

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

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

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

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

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

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

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

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 [29] and Theorem 7.3.2 in [35] that

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

where A^* is a Brownian motion with zero drift and 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 [15] that

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

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

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

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, where the lead time is the time remaining until the deadline elapses, i.e.,

$$\text{lead time} \;=\; \text{deadline} \;-\; \text{current time}. \tag{2.17}$$

The action of the EDF discipline requires knowledge of the current lead times of all customers in system. We will find it convenient to represent the latter via a collection of measure-valued stochastic processes.

Customer arrival measure-valued process:

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

Workload arrival measure-valued process:

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

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:

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

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:

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

In addition, we can also represent the reneged work process in terms of the workload measure-valued process as follows:

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

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

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

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 also 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, $C^{(n)}(t) \leq F^{(n)}(t)$ for all $t \geq 0$. The processes $C^{(n)}$, $F^{(n)}$ and $F_S^{(n)}$ are RCLL.

For the processes just defined, we use the following 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 set, for every Borel set B,

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

3 Main Results

Before stating the main results of this paper, we summarize the heavy traffic results for the standard EDF system that were obtained in [5] – in particular, we recall Proposition 3.10 and Theorem 3.1 of [5], which characterize the limiting distributions of the workload measure and the queue length measure in the standard system. Let

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

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 [5]) *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*

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

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 [5]) *Let $\mathcal{W}_{\mathcal{S}}^*$ and $\mathcal{Q}_{\mathcal{S}}^*$ be the measure-valued processes defined, respectively, by*

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

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

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 $H(0)$. 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 3.3 *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 3.4 *We have $\widehat{F}^{(n)} \Rightarrow F^*$ as $n \to \infty$, where*

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

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

Theorem 3.5 *Let $\mathcal{W}^*$ and $\mathcal{Q}^*$ be the measure-valued processes defined by*

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

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

By Theorem 3.5, the total mass of $\mathcal{W}^{(n)}$ must converge in distribution to the total mass of $\mathcal{W}^*$. Substituting $B = \mathbb{R}$ in (3.5) and using (3.1) and (3.4), we see that that $\mathcal{W}^*(t)(\mathbb{R}) = H(F^*(t)) = W^*(t)$ and we recover Theorem 3.3. Applying the same argument to queue lengths, we obtain the following Corollary to Theorem 3.5.

Corollary 3.6 *As $n \to \infty$, $\widehat{Q}^{(n)} \Rightarrow Q^* \triangleq \lambda W^*$.*

Theorem 3.5 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.

The last main result of the paper is a characterization of the limiting amount of reneged work in the system.

Theorem 3.7 *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, as discussed above, these results seem intuitive in light of the behavior of the standard EDF system, a rigorous proof turns out to be challenging. Moreover, counter to what one might expect, the result for queue lengths that is analogous to Theorem 3.7 does not hold. Specifically, although Corollary 3.6 shows that $\widehat{Q}^{(n)}$ converges in distribution to the doubly reflected Brownian motion $Q^* = \lambda W^*$ 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 seemingly intuitive statements.

The proof of Theorem 3.3 is presented in Section 6.1.1, the proofs of Proposition 3.4 and Theorem 3.5 can be found in Section 6.1.2, and Section 6.2 contains the proof of Theorem 3.7. Along the way, we also establish an optimality property for EDF in Section 5.1 that may be of independent interest.

Remark 3.8 The assumption made in (2.5) that the support of the lead time distribution is bounded above by $y^* < \infty$ is mainly technical and can be replaced by a weaker second moment condition that is more realistic. See [21] for the corresponding analysis for the standard EDF system. 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 for the results obtained here. Indeed, the work of Ward and Glynn [32, 33] on FIFO queues with reneging suggests that in the absence of such an assumption, the limiting workload process in heavy traffic may no longer be a reflected Brownian motion, and its properties may exhibit a strong sensitivity to the density of the lead-time distribution near 0. From a modeling point of view, it seems reasonable to impose a strictly positive lower bound $y^* > 0$ so as to avoid non-negligible "intrinsic lateness", in which a newly arriving customer has such a small initial lead time that he would be late even if there were no other customers in the system.

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 very useful 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

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

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

$$\Phi(\mu)(t)(-\infty, y] \triangleq \left[\mu(t)(-\infty, y] - \sup_{s\in[0,t]} \left(\mu(s)(-\infty, 0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R})\right)\right]^+ \tag{4.2}$$

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 $\mathcal{U}^{(n)}$, i.e.,

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

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 [5]. In a similar fashion, it will be useful to define the *reference frontier*

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

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

From (4.1)–(4.3) we have

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

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

where

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

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 $0 \le K^{(n)}(t) \le W_S^{(n)}(t)$ and so for all $t \ge 0$,

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

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 $t \in [0, \infty)$, $y_2 \ge y_1 > E^{(n)}(t)$,

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

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 $\mathcal{U}^{(n)}(t)$, $t \ge 0$, and $U^{(n)}(t)$, $t \ge 0$, 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 non-negativity of μ and (4.2),

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

Also, since the right-hand side of (4.2) is non-decreasing and right-continuous in y, we know that $\Phi(\mu)(t) \in \mathcal{M}$ for every $t \ge 0$. 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

$$\Psi(\nu, x)(-\infty, y] \triangleq (\nu(-\infty, y] - x)^+ \qquad \forall y \in \mathbb{R}$$

and $\Gamma : D_{\mathcal{M}}[0,\infty) \mapsto \mathbb{R}$ is defined by

$$\Gamma(\mu)(t) \triangleq \sup_{s\in[0,t]} \left(\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t]} \mu(u)(\mathbb{R}) \right) \qquad \forall t \in [0,\infty).$$

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, it is easy to verify that Ψ is continuous on $\mathcal{M} \times \mathbb{R}$. Thus, 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

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

where we define

$$Z(\mu,\varepsilon)(t) \triangleq \sup_{s\in[t,t+\varepsilon]} \left[\mu(s)(-\infty,0] \wedge \inf_{u\in[s,t+\varepsilon]} \mu(u)(\mathbb{R}) \right].$$

Since $\mu \in D_{\mathcal{M}}[0,\infty)$ implies $\mu(u)$ converges weakly to $\mu(t)$ as $u \downarrow t$, by Portmanteau's theorem, we have

$$\lim_{u\downarrow t} \mu(u)(\mathbb{R}) = \mu(t)(\mathbb{R}) \qquad \text{and} \qquad \mu(t)(-\infty,0] \geq \limsup_{s\downarrow t} \mu(s)(-\infty,0].$$

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, as $n \to \infty$, $\mu_n(t)$ converges weakly to $\mu(t)$ uniformly for t in compact sets (u.o.c.). 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 immediately 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 $D_{\mathcal{M}}[0,\infty)$. □

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 $t \geq 0$,

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

4.1.2 A decomposition of the reference workload

We now 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 $k = 0, 1, 2, ...$, we define recursively

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

$$\sigma_{k+1}^{(n)} \triangleq \min\left\{ t \geq \tau_k^{(n)} \middle| W_S^{(n)}(t) > W_S^{(n)}(t-) \right\}. \tag{4.12}$$

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

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

and

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

The following is the main result of this section.

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 can be easily deduced from Proposition 4.3, Proposition 4.4 and Remark 4.5 below. The rest of the section is devoted to establishing these latter 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

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

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

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

We further have

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

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

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

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

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

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

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

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

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

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

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

Equation (4.21) follows from (4.7).

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

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

We show that

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

and

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

This will imply (4.20) with k replaced by $k+1$.

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

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

and so

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

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

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

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

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

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 from these considerations 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

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

which establishes (4.25). □

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

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

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

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

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

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

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

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

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

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

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

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

But (4.18) and Proposition 4.4 show that

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

and so

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

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

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

In conclusion,

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

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

4.2 Dynamics of the reference workload process

In this section we analyze the dynamics of $\mathcal{U}^{(n)}$. Its time evolution is similar to that of $\mathcal{W}^{(n)}$, with one notable difference. In the case of $\mathcal{U}^{(n)}$, the incoming work $v_k^{(n)}$ associated with a customer arriving to the system at time t is distributed on some atoms of $\mathcal{W}_S^{(n)}(t)$ located on the half-line $[L_k^{(n)}, \infty)$, but not necessarily at the single point $L_k^{(n)}$. This difference in the evolutions will be used in Section 6 to show that the difference between $\widehat{U}^{(n)}$ and $\widehat{W}^{(n)}$ is asymptotically negligible.

Properties of the evolution of $\mathcal{U}^{(n)}$ are established in Section 4.2.2. Since a complete justification of these properties requires a detailed analysis, we first provide an informal summary of the main properties, along with an example.

4.2.1 An illustrative example

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 $\mathcal{U}^{(n)}$. We shall show (see Lemma 4.7 and Proposition 4.8) that the time evolution of the reference workload measure-valued process $\mathcal{U}^{(n)}$ is similar to the evolution of the workload measure $\mathcal{W}^{(n)}$ in the EDF system with reneging. 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 $\mathcal{U}^{(n)}$ (see (ii) and (v) of Proposition 4.8). 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 $v_{A^{(n)}(t)}^{(n)}$ located at $L_{A^{(n)}(t)}^{(n)}$ 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; see (4.57) 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 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 $v_{A^{(n)}(t)}^{(n)}$ associated with the new arrival is distributed in $[L_{A^{(n)}(t)}^{(n)}, \infty)$ (more precisely, on some atoms of $\mathcal{W}_S^{(n)}(t)$ located on this half-line), but it is not necessarily located at the single atom $L_{A^{(n)}(t)}^{(n)}$; see 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. We illustrate this point in Example 4.6.

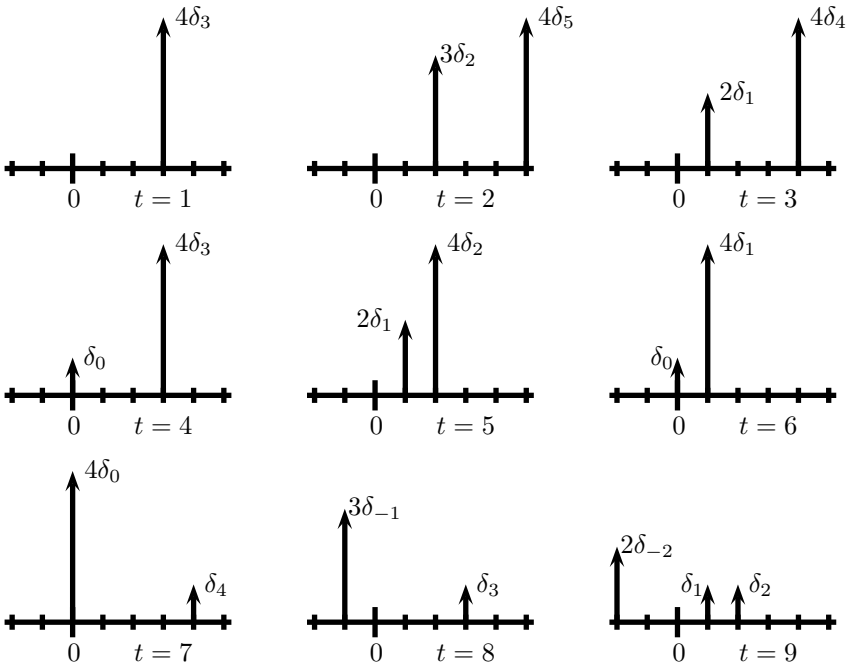

Figure 1: Evolution of $\mathcal{W}_S^{(n)}$

Example 4.6 Consider a system realization in which

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

Then

$$\mathcal{W}_S^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1,\\ (5-t)\delta_{4-t}, & 1 \le t < 2,\\ (5-t)\delta_{4-t} + 4\delta_{7-t}, & 2 \le t < 5,\\ (7-t)\delta_{6-t} + 4\delta_{7-t}, & 5 \le t < 7,\\ (11-t)\delta_{7-t} + \delta_{11-t}, & 7 \le t < 9,\\ 2\delta_{-2} + \delta_1 + \delta_2, & t = 9. \end{cases} \tag{4.38}$$

The measure $\mathcal{W}_S^{(n)}(t)$ is shown for integer values of t ranging between 1 and 9 in Figure 1.

Consider the function $K^{(n)}$ defined by (4.7). We have $W_S^{(n)}(u) = 5 - u$ for

$1 \le u < 2$ and $W_S^{(n)}(u) \ge 4$ for all $u \in [2,8]$, and hence

$$\begin{aligned} K^{(n)}(t) &= \sup_{s\in[0,t]} \left(\mathcal{W}_S^{(n)}(s)(-\infty,0] \wedge \inf_{u\in[s,t]} W_S^{(n)}(u) \right) \\ &= \sup_{s\in[0,t]} \mathcal{W}_S^{(n)}(s)(-\infty,0] \\ &= \begin{cases} 0, & 0 \le t < 4, \\ 1, & 4 \le t < 7, \\ 4, & 7 \le t \le 8. \end{cases} \end{aligned}$$

However, for $8 \le t < 9$, we have $W_S^{(n)}(t) = 12 - t \le 4$. For t in this range, the supremum in the definition of $K^{(n)}(t)$ is attained at $s = 7$, and therefore

$$K^{(n)}(t) = \mathcal{W}_S^{(n)}(7)(-\infty,0] \wedge \inf_{u\in[7,t]} W_S^{(n)}(u) = 4 \wedge (12-t) = 12 - t.$$

For $t = 9$, we have $W_S^{(n)}(9) = 4 \ne 12 - t$. Nonetheless, the supremum in the definition of $K^{(n)}(9)$ is still attained at $s = 7$. Indeed,

$$K^{(n)}(9) = \mathcal{W}_S^{(n)}(7)(-\infty,0] \wedge \inf_{u\in[7,9]} W_S^{(n)}(u) = 4 \wedge \big(\inf_{u\in[7,9)} (12-t) \wedge 4 \big) = 3.$$

In summary,

$$K^{(n)}(t) = \begin{cases} 0, & 0 \le t < 4, \\ 1, & 4 \le t < 7, \\ 4, & 7 \le t \le 8, \\ 12 - t, & 8 \le t \le 9. \end{cases} \tag{4.39}$$

The measure $\mathcal{U}^{(n)}(t)$ is obtained by removing mass $K^{(n)}(t)$ from the measure $\mathcal{W}_S^{(n)}(t)$, working from left to right. This results in the formula

$$\mathcal{U}^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1, \\ (5-t)\delta_{4-t}, & 1 \le t < 2, \\ (5-t)\delta_{4-t} + 4\delta_{7-t}, & 2 \le t < 4, \\ (8-t)\delta_{7-t}, & 4 \le t < 5, \\ (6-t)\delta_{6-t} + 4\delta_{7-t}, & 5 \le t < 6, \\ (10-t)\delta_{7-t}, & 6 \le t < 7, \\ (8-t)\delta_{11-t}, & 7 \le t < 8, \\ 0, & 8 \le t < 9, \\ \delta_2, & t = 9. \end{cases} \tag{4.40}$$

The measure $\mathcal{U}^{(n)}(t)$ is shown for integer values of t ranging between 1 and 9 in Figure 2.

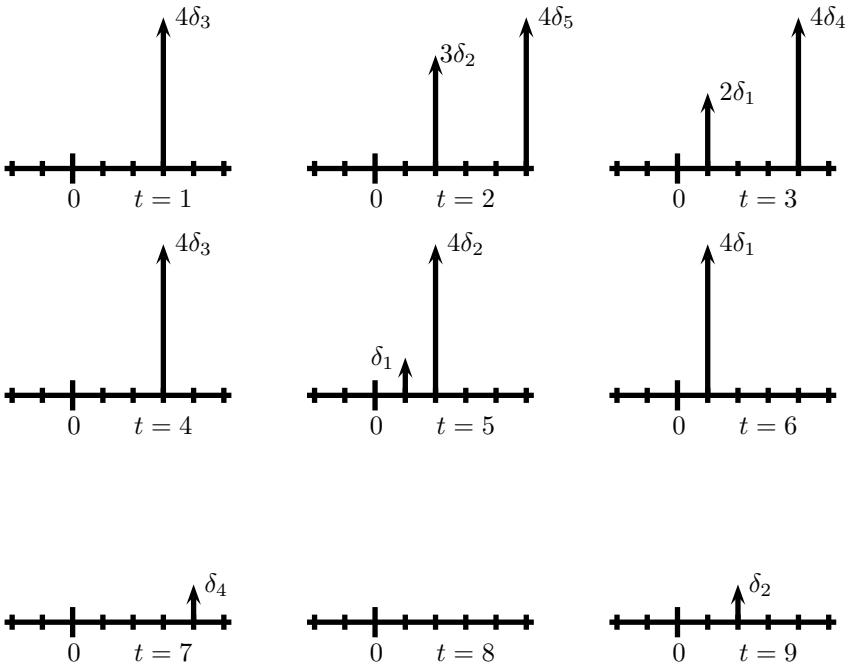

Figure 2: Evolution of $\mathcal{U}^{(n)}$

The total mass in the reference system is

$$U^{(n)}(t) = \begin{cases} 0, & 0 \le t < 1, \\ 5 - t, & 1 \le t < 2, \\ 9 - t, & 2 \le t < 4, \\ 8 - t, & 4 \le t < 5, \\ 10 - t, & 5 \le t < 7, \\ 8 - t, & 7 \le t < 8, \\ 0, & 8 \le t < 9, \\ 1, & t = 9. \end{cases} \tag{4.41}$$

This total mass path has jumps $\triangle U^{(n)}(1) = 4$, $\triangle U^{(n)}(2) = 4$, $\triangle U^{(n)}(4) = -1$, $\triangle U^{(n)}(5) = 2$, $\triangle U^{(n)}(7) = -2$ (the result of an arrival of mass 1 and the deletion of mass 3), and $\triangle U^{(n)}(9) = 1$.

We see that arriving mass to $\mathcal{U}^{(n)}$ is not always placed at the lead time of the arriving customer. In particular, $\mathcal{U}^{(n)}(5-) = 3\delta_2$, but $\mathcal{U}^{(n)}(5) = \delta_1 + 4\delta_2$. The mass $v_3^{(n)} = 2$ arriving at time 5 is distributed with one unit at $L_3^{(n)} = 1$ and one unit at 2. Furthermore, the mass $v_5^{(n)} = 1$ arriving at time $t = 9$, which begins a new busy period for $\mathcal{U}^{(n)}$, is placed at 2 rather than at $L_5^{(n)} = 1$.

Due to the failure of $\mathcal{U}^{(n)}$ to place all arriving masses at their lead times, the reneging system measure $\mathcal{W}^{(n)}(t)$ is not $\mathcal{U}^{(n)}(t)$ for $5 \le t < 7$ and $t = 9$. The

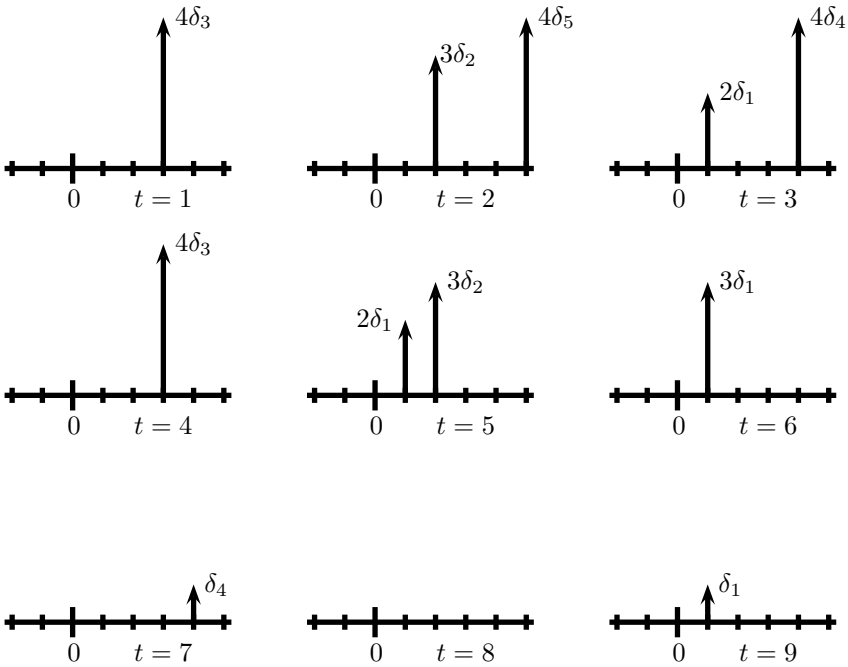

Figure 3: Evolution of the reneging system $\mathcal{W}^{(n)}$

full formula for the reneging system is

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

The measure $\mathcal{W}^{(n)}(t)$ is shown for integer values of t ranging between 1 and 9 in Figure 3.

Beginning at time $t = 4$, the reneging system begins serving the customer with lead time 3, and thus by time $t = 5$, this customer, whose lead time is now 2, requires only three remaining units of service. The customer arriving at time $t = 5$ with lead time 1 brings an additional two units of work. At time $t = 5$, the reneging system thus has five units of work, which agrees with $U^{(n)}(5) = 5$, but the mass in the reneging system is not distributed according to the measure $\mathcal{U}^{(n)}(5)$. At time $t = 6$, an additional unit of work is deleted from the reneging system but not from the reference system, and so $W^{(n)}(6) = 3$, whereas $U^{(n)}(6) = 4$. This discrepancy can be traced back to the arrival at time

$t = 5$ of a customer more urgent than the customer in service in the reneging system. We shall see that we always have $W^{(n)}(t) \leq U^{(n)}(t)$, and the inequality can be strict due to work that preempts the customer in service in the reneging system, but the difference between $W^{(n)}(t)$ and $U^{(n)}(t)$ is never more than the amount of such work deleted by the reneging system up to time t (Lemma 5.6). □

4.2.2 Rigorous description of the evolution of the reference workload

As shown in Section 4.1, the time interval $[0, \infty)$ can be decomposed into a union of the disjoint intervals $(\tau_k^{(n)}, \sigma_{k+1}^{(n)}]$ and $(\sigma_k^{(n)}, \tau_k^{(n)}]$, $k \geq 0$, such that $K^{(n)} = W_S^{(n)} - U^{(n)}$ is non-increasing on $(\tau_k^{(n)}, \sigma_{k+1}^{(n)}]$ and non-decreasing on $(\sigma_k^{(n)}, \tau_k^{(n)}]$. In Lemma 4.7 below, we analyze the behavior of $\mathcal{U}^{(n)}$ on the time intervals $[\tau_{k-1}^{(n)}, \sigma_k^{(n)}]$, $k \geq 1$, while Proposition 4.8 describes the dynamics of $\mathcal{U}^{(n)}$ on the intervals $(\sigma_k^{(n)}, \tau_k^{(n)})$, $k \geq 1$. The section ends with Corollary 4.9, which describes the time evolution of the reference workload process $U^{(n)}$.

In the proofs, we will make use of the following elementary facts about the standard workload. Observe that, since the interarrival times are strictly positive, $\triangle A^{(n)}(t) \in \{0, 1\}$, and for $t \geq 0$, we have

$$\mathcal{W}_S^{(n)}(t) = \mathcal{W}_S^{(n)}(t-) + \triangle A^{(n)}(t)\, v_{A^{(n)}(t)}^{(n)}\, \delta_{L_{A^{(n)}(t)}^{(n)}}, \tag{4.42}$$

which implies

$$\triangle W_S^{(n)}(t) = \triangle A^{(n)}(t)\, v_{A^{(n)}(t)}^{(n)}. \tag{4.43}$$

Note also that for any functions f and g defined 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

$$\lim_{s \uparrow t} \mathcal{W}_S^{(n)}(s)[f(s), g(s)] = \mathcal{W}_S^{(n)}(t-)[f(t-), g(t-)]. \tag{4.44}$$

This is true because the lead times of the customers present in the standard system decrease with unit rate. The equation (4.44) 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 explicit reference.

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

$$U^{(n)}(t) = 0, \qquad \tau_{k-1}^{(n)} \leq t < \sigma_k^{(n)}, \tag{4.45}$$

$$\triangle U^{(n)}(\sigma_k^{(n)}) = v_{A^{(n)}(\sigma_k^{(n)})}^{(n)}, \tag{4.46}$$

$$\mathcal{U}^{(n)}(\sigma_k^{(n)})\big(-\infty, L_{A^{(n)}(\sigma_k^{(n)})}^{(n)}\big) = 0. \tag{4.47}$$

Proof: Equation (4.45) follows immediately from (4.6), (4.18) and Proposition 4.4. By (4.6), (4.34), (4.43), and the fact that $\triangle A^{(n)}(\sigma_k^{(n)}) = 1$, we have

$$\triangle U^{(n)}(\sigma_k^{(n)}) = \triangle W_S^{(n)}(\sigma_k^{(n)}) - \triangle K^{(n)}(\sigma_k^{(n)}) = v^{(n)}_{A^{(n)}(\sigma_k^{(n)})},$$

and (4.46) follows. For $y < L^{(n)}_{A^{(n)}(\sigma_k^{(n)})}$, (4.5), (4.42) (4.34) and (4.33) imply

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

and so (4.47) also follows. □

The last lemma showed that $\sigma_k^{(n)}$ commences a busy period for the reference system. The equation (4.30) implies that $U^{(n)}(t) > 0$ for $t \in (\sigma_k^{(n)}, \tau_k^{(n)})$, and thus the intervals $[\sigma_k^{(n)}, \tau_k^{(n)})$, $k \geq 1$, are precisely the busy periods for the reference system. We now 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 $t \in (\sigma_k^{(n)}, \tau_k^{(n)})$ we have

$$\begin{aligned}
&\mathcal{U}^{(n)}(t)(-\infty, y] \\
&= \left[\mathcal{W}_S^{(n)}(t)(-\infty, y] - \left(W_S^{(n)}(\sigma_k^{(n)}-) \vee \max_{s \in [\sigma_k^{(n)}, t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0]\right)\right]^+ . \quad (4.48)
\end{aligned}$$

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 $\mathcal{I}$:

$$\nu_{|\mathcal{I}}(B) = \nu(B \cap \mathcal{I}) \qquad \forall B \in \mathcal{B}(\mathbb{R}).$$

Proposition 4.8 *For $k \geq 1$ 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

$$\begin{aligned}
\triangle K^{(n)}(t) &= 0, \quad (4.49) \\
\triangle U^{(n)}(t) &= 0. \quad (4.50)
\end{aligned}$$

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 neighbourhood of t, whereas 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 $\triangle A^{(n)}(t) = 0$, $E^{(n)}(t-) = 0$, then

$$\mathcal{U}^{(n)}(t-)\{0\} = \triangle K^{(n)}(t) = -\triangle U^{(n)}(t) \quad (4.51)$$

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 (4.49) holds,* $\triangle E^{(n)}(t) \geq 0$ *and*

$$\mathcal{U}^{(n)}(t) = \mathcal{U}^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)}\, \delta_{L^{(n)}_{A^{(n)}(t)}}. \tag{4.52}$$

(iv) If $\triangle A^{(n)}(t) = 1$*,* $L^{(n)}_{A^{(n)}(t)} < E^{(n)}(t-)$*, then (4.49) holds and*

$$\begin{aligned} L^{(n)}_{A^{(n)}(t)} \quad \leq \quad E^{(n)}(t) \quad &\leq \quad E^{(n)}(t-), && (4.53)\\ \triangle U^{(n)}(t) \quad &= \quad v^{(n)}_{A^{(n)}(t)}, && (4.54)\\ \mathcal{U}^{(n)}(t)_{|(E^{(n)}(t-),\infty)} \quad &= \quad \mathcal{U}^{(n)}(t-)_{|(E^{(n)}(t-),\infty)}, && (4.55)\\ \mathcal{U}^{(n)}(t)\{E^{(n)}(t-)\} \quad &\geq \quad \mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\}. && (4.56) \end{aligned}$$

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

$$\mathcal{U}^{(n)}(t) = \mathcal{U}^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)}\, \delta_{L^{(n)}_{A^{(n)}(t)}} - \mathcal{U}^{(n)}(t-)\{0\}\, \delta_0. \tag{4.57}$$

Proof: Fix $k \geq 1$ and $t \in (\sigma_k^{(n)}, \tau_k^{(n)})$. We start with the general observation that, by (4.4) and (4.48),

$$E^{(n)}(t) = \min\left\{ y | \mathcal{W}_S^{(n)}(t)(-\infty, y] > W_S^{(n)}(\sigma_k^{(n)}-) \vee \max_{s\in[\sigma_k^{(n)},t]} \mathcal{W}_S^{(n)}(s)(-\infty, 0] \right\} \tag{4.58}$$

and because $\mathcal{W}_S^{(n)}(t)$ is purely atomic, the minimum on the right-hand side of (4.58) is obtained at some atom of $\mathcal{W}_S^{(n)}(t)$ located at $y_0 = E^{(n)}(t)$. In particular,

$$\mathcal{W}_S^{(n)}(t)\{E^{(n)}(t)\} > 0. \tag{4.59}$$

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 close to t,

$$\mathcal{W}_S^{(n)}(s)(-\infty, a/2] \leq K^{(n)}(s). \tag{4.60}$$

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

$$\mathcal{W}_S^{(n)}(t)(-\infty, 0] \leq \mathcal{W}_S^{(n)}(s)(-\infty, a/2]. \tag{4.61}$$

The last two relations show that $\mathcal{W}_S^{(n)}(t)(-\infty, 0] \leq K^{(n)}(t-)$, and so, by Proposition 4.3, (4.49) holds. The equation (4.50) follows from (4.6), (4.49), (4.43) and the assumption $\triangle A^{(n)}(t) = 0$. Because $W_S^{(n)}(t) > 0$ (see (4.59)), $W_S^{(n)}$ decreases at unit rate in a neighbourhood of t (see (2.6)–(2.8)). In addition, (4.6), (4.49)

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 neighbourhood of t, together imply that $U^{(n)}$ also decreases at unit rate in a neighbourhood of t. Furthermore, the nature of the EDF discipline and (4.59) 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 neighbourhood 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

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

It follows that

$$\mathcal{W}_S^{(n)}(t)\big(E^{(n)}(t-), E^{(n)}(t)\big) = 0. \tag{4.62}$$

Using the definition of $E^{(n)}(t)$, (4.50), (4.9), the assumption $\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} = 0$, the assumption $\triangle A^{(n)}(t) = 0$, and (4.62), we obtain

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

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

$$\mathcal{W}_S^{(n)}(s)(-\infty, E^{(n)}(s)] > K^{(n)}(s). \tag{4.63}$$

As $s \uparrow t$ in (4.63), by (4.44), (4.42), and the case (ii) assumptions $\triangle A^{(n)}(t) = 0$ and $E^{(n)}(t-) = 0$, we get

$$\mathcal{W}_S^{(n)}(t)(-\infty, 0] = \mathcal{W}_S^{(n)}(t-)(-\infty, 0] \geq K^{(n)}(t-). \tag{4.64}$$

When combined with Proposition 4.3, this implies

$$K^{(n)}(t) = K^{(n)}(t-) \vee \mathcal{W}_S^{(n)}(t)(-\infty, 0] = \mathcal{W}_S^{(n)}(t)(-\infty, 0]. \tag{4.65}$$

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

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

Letting $s \uparrow t$, and invoking (4.44), (4.64), (4.65) and the assumption $E^{(n)}(t-) = 0$, we obtain

$$\mathcal{U}^{(n)}(t-)\{0\} = \mathcal{W}_S^{(n)}(t-)(-\infty, 0] - K^{(n)}(t-) = K^{(n)}(t) - K^{(n)}(t-), \tag{4.66}$$

and the first equality in (4.51) follows. The second equality in (4.51) follows from (4.6), (4.43) and the assumption $\triangle A^{(n)}(t) = 0$. Moreover, by (4.5) and (4.65), for every $y \in \mathbb{R}$,

$$\begin{aligned} \mathcal{U}^{(n)}(t)(-\infty, y] &= \left[\mathcal{W}_S^{(n)}(t)(-\infty, y] - \mathcal{W}_S^{(n)}(t)(-\infty, 0]\right]^+ \\ &= \mathcal{W}_S^{(n)}(t)_{|(0,\infty)}(-\infty, y]. \end{aligned} \tag{4.67}$$

(iii) Let $a = E^{(n)}(t-)$. We can deduce (4.49) from (4.60) and (4.61) as in (i), with the only difference that now (4.61), 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 (4.49), 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 (4.52). Equation (4.5) implies

$$\mathcal{U}^{(n)}(t-)(-\infty, y] = \left[\mathcal{W}_S^{(n)}(t-)(-\infty, y] - K^{(n)}(t-)\right]^+. \tag{4.68}$$

Indeed, for any y such that $\mathcal{W}_S^{(n)}(t-)\{y\} = 0$, (4.68) 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 (4.68) holds for all y. Moreover, using (4.42), (4.49) and (4.5), we see that

$$\begin{aligned} &\mathcal{U}^{(n)}(t)(-\infty, y] \\ &= \left[\mathcal{W}_S^{(n)}(t-)(-\infty, y] - K^{(n)}(t-) + v^{(n)}_{A^{(n)}(t)} \delta_{L^{(n)}_{A^{(n)}(t)}(t)}(-\infty, y]\right]^+. \end{aligned} \tag{4.69}$$

When combined with (4.68), this shows that

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}^{(n)}(t-)(-\infty, y], \quad y < L^{(n)}_{A^{(n)}(t)}(t). \tag{4.70}$$

On the other hand, if $y \geq L^{(n)}_{A^{(n)}(t)}(t)$, then $y \geq E^{(n)}(t-)$ and (4.68) becomes

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

From (4.69), we now have

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}^{(n)}(t-)(-\infty, y] + v^{(n)}_{A^{(n)}(t)}, \quad y \geq L^{(n)}_{A^{(n)}(t)}. \tag{4.71}$$

When combined, (4.70) and (4.71) prove (4.52).

(iv) We have $L^{(n)}_{A^{(n)}(t)} > 0$, and so (4.49) 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 (4.42), (4.5), (4.49) and the definition of $E^{(n)}$, imply that

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

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 (4.42), the inequalities $y + \varepsilon < a < E^{(n)}(t-)$ and (4.49), 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 (4.53). In addition, by (4.6) and (4.49), 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 (4.54) follows. Furthermore, since $E^{(n)}(t) \leq E^{(n)}(t-)$ by (4.53), the relations (4.9), (4.42) 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 (4.55).

Finally, to prove (4.56), we will consider two cases.

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

By (4.9), we know that

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

In turn, when combined with (4.68) and the definition of $E^{(n)}$, this shows that

$$\begin{aligned}
\mathcal{U}^{(n)}(t-)\{E^{(n)}(t-)\} &= \mathcal{U}^{(n)}(t-)(-\infty, E^{(n)}(t-)] - \mathcal{U}^{(n)}(t-)(-\infty, E^{(n)}(t-)) \\
&= \mathcal{W}_S^{(n)}(t-)(-\infty, E^{(n)}(t-)] - K^{(n)}(t-) \\
&\quad - \left[\mathcal{W}_S^{(n)}(t-)(-\infty, E^{(n)}(t-)) - K^{(n)}(t-)\right]^+ \\
&\leq \mathcal{W}_S^{(n)}(t-)(-\infty, E^{(n)}(t-)] - \mathcal{W}_S^{(n)}(t-)(-\infty, E^{(n)}(t-)) \\
&= \mathcal{W}_S^{(n)}(t)\{E^{(n)}(t-)\} \\
&= \mathcal{U}^{(n)}(t)\{E^{(n)}(t-)\},
\end{aligned}$$

and so (4.56) holds.
Case II. $E^{(n)}(t) = E^{(n)}(t-)$.

By (4.5), (4.42), (4.49), (4.68) and the definition of $E^{(n)}$,

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

which establishes (4.56) in this case as well. Since $E^{(n)}(t) \leq E^{(n)}(t-)$, the two cases above are mutually exhaustive, and so (4.56) is proved.

(v) The equation (4.67) holds by the same argument as in (ii), but where now the equality in (4.64) 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,

$$\mathcal{U}^{(n)}(t)(-\infty, y] = \mathcal{U}_1^{(n)}(t)(-\infty, y] = 0, \quad y \leq 0. \tag{4.72}$$

By (4.9) and the fact that $E^{(n)}(t-) = 0$, $\mathcal{U}^{(n)}(t-)_{|(0,\infty)} = \mathcal{W}_S^{(n)}(t-)_{|(0,\infty)}$. The last two statements, along with (4.42), (4.67) 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 (4.72), shows that $\mathcal{U}^{(n)}(t) = \mathcal{U}_1^{(n)}(t)$. $\square$

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

$$W^{(n)}(t) = V^{(n)}\big(A^{(n)}(t)\big) - \int_0^t \mathbb{I}_{\{W^{(n)}(s)>0\}}\, ds - R_W^{(n)}(t), \tag{4.73}$$

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

$$U^{(n)}(t) = V^{(n)}\big(A^{(n)}(t)\big) - \int_0^t \mathbb{I}_{\{U^{(n)}(s)>0\}}\, ds - R_U^{(n)}(t), \tag{4.74}$$

where

$$R_U^{(n)}(t) \triangleq \sum_{0<s\leq t} \mathcal{U}^{(n)}(s-)\{0\}. \tag{4.75}$$

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

Corollary 4.9 *For every $t \geq 0$, the equation (4.74) holds. Moreover, $R_U^{(n)} = K_+^{(n)}$ and hence*

$$U^{(n)} = N^{(n)} + I_U^{(n)} - K_+^{(n)}, \tag{4.76}$$

where, for $t \geq 0$,

$$I_U^{(n)}(t) \triangleq \int_0^t \mathbb{I}_{\{U^{(n)}(s)=0\}}\, ds. \tag{4.77}$$

Proof: For $t \geq 0$, let $\widetilde{U}^{(n)}(t)$ be equal to the right-hand side of (4.74). By (4.45) 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 (4.74), 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

$$\Delta \widetilde{U}^{(n)}(\tau_k^{(n)}) = \Delta U^{(n)}(\tau_k^{(n)}). \tag{4.78}$$

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

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

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 (4.78) 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

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

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 $\mathcal{U}^{(n)}(s-)\{E^{(n)}(s-\} = 0$. 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 (4.74).

Adding and subtracting t to (4.74), by the definition (2.6) of the netput process $N^{(n)}$ and the non-negativity of $U^{(n)}$, we obtain

$$U^{(n)}(t) = N^{(n)}(t) + \int_0^t \mathbb{I}_{\{U^{(n)}(s)=0\}}\, ds - R_U^{(n)}(t), \tag{4.79}$$

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

$$\int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}} dR_U^{(n)}(s) = \int_{[0,\infty)} \mathbb{I}_{\{U^{(n)}(s)>0\}} dK_+^{(n)}(s). \tag{4.80}$$

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 (4.74), yields (4.76). □

5 The Reneging System

In Section 5.2 we bound the difference in workload between the pre-limit reference and reneging systems — Lemma 5.3 provides a lower bound, while Lemma 5.7 provides an upper bound. The proof of the upper bound uses an optimality property of EDF that may be of independent interest. This property is established in Section 5.1.

5.1 Optimality of the EDF scheduling policy

The main result of this section, Theorem 5.1, states that the EDF system with reneging minimizes the amount of work that gets late among all service protocols (whether or not work conserving) with reneging for a single-station queueing system with given arrival times, service times and lead times. The related fact that the EDF protocol minimizes the number of late customers in the G/M/c queue was proved in [28]. The main idea of our proof is also similar to the one from [28]. Note, however, that 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.

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 late 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*

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

Proof: 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 present at time t_0 with the smallest lead time.

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 $t \geq 0$,

$$R_{\rho(\pi)}(t) \leq R_\pi(t). \tag{5.2}$$

Let $v_k(t)$ (respectively, $v_k^\rho(t)$) be the residual service time of the k-th customer at time t under π (respectively, $\rho(\pi)$). In particular, if d_k is the deadline of the k-th customer, $v_k(d_k-)$ (respectively, $v_k^\rho(d_k-)$) is the amount of mass corresponding to this customer that is deleted by π (respectively, $\rho(\pi)$) due to lateness, and thus

$$R_{\rho(\pi)}(t) = \sum_{k:d_k \leq t} v_k^\rho(d_k-), \qquad R_\pi(t) = \sum_{k:d_k \leq t} v_k(d_k-). \tag{5.3}$$

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

$$v_k^\rho(t) = v_k(t), \tag{5.4}$$

whereas

$$v_j^\rho(t) \leq v_j(t). \tag{5.5}$$

Summing (5.4) over $k \neq j$, invoking (5.5) and (5.3), we obtain (5.2).

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 (5.3) and (5.5), whereas (5.4) now holds only for $k \notin \{i, j\}$. If the i-th customer is served to completion under $\rho(\pi)$, then $v_i^\rho(d_i-) = 0$, and (5.4) for $k \notin \{i, j\}$,

and (5.5) imply that (5.2) holds for all t. It remains to consider the case that the i-th customer becomes late under $\rho(\pi)$. In this case (5.4) for $k \notin \{i, j\}$, and (5.5) imply that (5.2) holds for $t \in [0, d_i)$. Let w_1 denote the work done by $\rho(\pi)$ on the j-th customer when π works on the i-th customer in the time 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

$$v_j^\rho(d_j-) + v_i^\rho(d_i-) = v_j(d_j-) + v_i(d_i-) + w_3. \tag{5.6}$$

We argue by contradiction that w_3 cannot be positive. If w_3 were positive, then at some time $t \geq t_1$, 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 $t \geq t_1$, and so $v_j^\rho(d_j-) = 0$ as well. The left-hand side of (5.6) is zero, and hence w_3 must be zero. We conclude that

$$v_j^\rho(d_j-) + v_i^\rho(d_i-) = v_j(d_j-) + v_i(d_i-). \tag{5.7}$$

Since $d_j < d_i$, if $t \geq d_i$, then (5.2) holds because of (5.4) for $k \notin \{i, j\}$ and (5.7).

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 $R_{\rho^n(\pi)}(t) \leq R_\pi(t)$ for all $t \geq 0$.

By assumption, for each t the number of arrivals $A(t)$ to the system 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 used. 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 referenced in the theorem, and hence (5.1) holds. □

Remark 5.2 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 Theorem 5.1 is sufficient for this paper. However, it is not hard to see that our proof can be generalized to policies permitting simultaneous service of different customers

(for example, 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 j-th 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 i-th 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.

5.2 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 5.3 showing that $W^{(n)} \leq U^{(n)}$, which immediately implies that $R_U^{(n)} \leq R_W^{(n)}$ (see Corollary 5.4).

In the proofs of these results, we will make frequent use of the observation that, by (4.73) and (4.74),

$$W^{(n)}(t) - U^{(n)}(t) = \int_0^t \mathbb{I}_{\{U^{(n)}(s)>0\}}\, ds - \int_0^t \mathbb{I}_{\{W^{(n)}(s)>0\}}\, ds + R_U^{(n)}(t) - R_W^{(n)}(t) \tag{5.8}$$

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

Lemma 5.3 *For every $t \geq 0$, we have*

$$W^{(n)}(t) \leq U^{(n)}(t). \tag{5.9}$$

Proof: Let

$$\tau \triangleq \min\{t \geq 0 : W^{(n)}(t) > U^{(n)}(t)\}. \tag{5.10}$$

If $\tau = +\infty$, then (5.9) holds. Assume $\tau < +\infty$. In this case, we claim that the minimum on the right-hand side of (5.10), which is finite by assumption, is attained. Indeed, (5.8) 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

$$\mathcal{W}^{(n)}(\tau)(y, \infty) > \mathcal{U}^{(n)}(\tau)(y, \infty). \tag{5.11}$$

Let

$$\tau_0 \triangleq \inf\{t \in [0, \tau] : \mathcal{W}^{(n)}(t)(y + \tau - t, \infty) > \mathcal{U}^{(n)}(t)(y + \tau - t, \infty)\}. \tag{5.12}$$

By (5.11), 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 the process $\mathcal{W}^{(n)}(t)(y+\tau-t,\infty)-\mathcal{U}^{(n)}(t)(y+\tau-t,\infty)$, $t\in[0,\tau]$, has no positive jumps and therefore

$$\mathcal{W}^{(n)}(t)(y+\tau-\tau_0,\infty)=\mathcal{U}^{(n)}(t)(y+\tau-\tau_0,\infty). \tag{5.13}$$

By (5.11) and (5.13), $\tau_0<\tau$. Thus, there exists $t\in(\tau_0,\tau)$, where $t-\tau_0$ is arbitrarily small and

$$\mathcal{W}^{(n)}(t)(y+\tau-t,\infty)>\mathcal{U}^{(n)}(t)(y+\tau-t,\infty). \tag{5.14}$$

However, we claim that (5.13) and (5.14) imply that for all $t\in(\tau_0,\tau)$, where $t-\tau_0$ is small enough, it must be that

$$\mathcal{W}^{(n)}(t)(0,y+\tau-t] > 0, \tag{5.15}$$
$$\mathcal{U}^{(n)}(t)(0,y+\tau-t] = 0. \tag{5.16}$$

Indeed, if (5.15) is false, then the left-hand side of (5.14) 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 (5.16) is false, the right-hand side of (5.14) is constant on some interval beginning with τ_0. In both cases, due to (5.13), (5.14) cannot hold for $t\in(\tau_0,\tau)$ with $t-\tau_0$ arbitrarily small. But (5.14)-(5.16) yield $W^{(n)}(t)>U^{(n)}(t)$ for some $t<\tau$, which contradicts (5.10). □

We now establish the reverse inequality for the processes $R_W^{(n)}$ and $R_U^{(n)}$, which represent the cumulative amount of late or reneged work in the reneging and reference systems, respectively.

Corollary 5.4 *For every $t\geq 0$,*

$$R_U^{(n)}(t)\leq R_W^{(n)}(t). \tag{5.17}$$

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

$$R_U^{(n)}(t)-R_U^{(n)}(\sigma_k^{(n)}-)\leq R_W^{(n)}(t)-R_W^{(n)}(\sigma_k^{(n)}-). \tag{5.18}$$

Proof: Lemma 5.3 and (5.8), together, imply that for $0\leq s\leq t$,

$$\left(R_U^{(n)}(t)-R_U^{(n)}(s)\right)-\left(R_W^{(n)}(t)-R_W^{(n)}(s)\right)\leq U^{(n)}(s)-W^{(n)}(s). \tag{5.19}$$

Substituting $s=0$ into (5.19) and using the fact that $R_U^{(n)}(0)=R_W^{(n)}(0)=U^{(n)}(0)=W^{(n)}(0)=0$, we obtain (5.17). Likewise, for $0\leq s\leq\sigma_k^{(n)}\leq t$, taking limits as s tends to $\sigma_k^{(n)}$ in (5.19), and using the fact that $U^{(n)}(\sigma_k^{(n)}-)=W^{(n)}(\sigma_k^{(n)}-)=0$, which follows from (4.45), Lemma 5.3 and the non-negativity of $W^{(n)}$, we obtain (5.18). □

The above proofs of Lemma 5.3 and Corollary 5.4 may be used to 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 5.5 *For every $t \geq 0$ such that $U^{(n)}(t) > 0$, we have*

$$E^{(n)}(t) \leq F^{(n)}(t). \tag{5.20}$$

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

$$\begin{aligned} \mathcal{U}^{(n)}(t)(y, \infty) &= W_S^{(n)}(t) - K^{(n)}(t) - \left[\mathcal{W}_S^{(n)}(t)(-\infty, y] - K^{(n)}(t)\right]^+ \\ &\leq \mathcal{W}_S^{(n)}(t)(y, \infty). \end{aligned} \tag{5.21}$$

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

$$\begin{aligned} W^{(n)}(t) &\geq \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{W}^{(n)}(t)(F^{(n)}(t), \infty) \\ &= \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{V}^{(n)}(t)(F^{(n)}(t), \infty) \\ &\geq \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{V}^{(n)}(t)[E^{(n)}(t), \infty) \\ &\geq \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{W}_S^{(n)}(t)[E^{(n)}(t), \infty) \\ &\geq \mathcal{W}^{(n)}(t)\{C^{(n)}(t)\} + \mathcal{U}^{(n)}(t)[E^{(n)}(t), \infty) \\ &\geq U^{(n)}(t), \end{aligned} \tag{5.22}$$

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 (5.21) 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 (5.22) is strict, which contradicts (5.9). □

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

$$F^{(n)}(t_1) - (t_2 - t_1) \leq F^{(n)}(t_2), \quad 0 \leq t_1 \leq t_2. \tag{5.23}$$

Lemma 5.6 *For every $t \geq 0$,*

$$U^{(n)}(t) - W^{(n)}(t) \leq D^{(n)}(t). \tag{5.24}$$

Proof: If $t \in [\tau_{k-1}^{(n)}, \sigma_k^{(n)})$ for some $k \geq 1$, then $U^{(n)}(t) = 0$ by (4.45). Thus, by (4.19), it suffices to prove (5.24) on $[\sigma_k^{(n)}, \tau_k^{(n)})$ for every $k \geq 1$. Let $k \geq 1$. Suppose that (5.24) is false for some $t \in [\sigma_k^{(n)}, \tau_k^{(n)})$. Let

$$\tau \triangleq \min\{t \in [\sigma_k^{(n)}, \tau_k^{(n)}) | U^{(n)}(t) - W^{(n)}(t) > D^{(n)}(t)\}. \tag{5.25}$$

We first argue that the minimum on the right-hand side of (5.25) is attained. Indeed, by (5.8) and Lemma 5.3, 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

$$U^{(n)}(\tau) - W^{(n)}(\tau) > D^{(n)}(\tau). \tag{5.26}$$

Moreover (4.45), (4.46) and Lemma 5.3 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, (5.25) implies

$$U^{(n)}(\tau-) - W^{(n)}(\tau-) \leq D^{(n)}(\tau-). \tag{5.27}$$

Let k_0 be the index of the customer arriving at time $\sigma_k^{(n)}$, i.e., $S_{k_0}^{(n)} = \sigma_k^{(n)}$. Let $k_1 \geq k_0$ 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, $\ell \in [k_0, k_1]$, $L_\ell^{(n)} - (S_{k_1}^{(n)} - S_\ell^{(n)}) \leq L_{k_1}^{(n)}$ 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 corresponding to 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 (4.46), (4.47), (4.53), (4.54) and (4.4). Let $s > L_{k_1}^{(n)} + S_{k_1}^{(n)} - S_\ell^{(n)}$ be such that $\triangle \mathcal{U}^{(n)}(S_\ell^{(n)})\{s\} > 0$. Such a point s exists since the measure $\mathcal{U}^{(n)}(S_\ell^{(n)})$ is discrete.

If $\ell > k_0$ (which includes the case $\ell = k_1$), then, by (4.55) in Proposition 4.8 (iv) and Lemma 5.5, we have $s \leq E^{(n)}(S_\ell^{(n)}-) \leq F^{(n)}(S_\ell^{(n)}-) \leq F^{(n)}(S_\ell^{(n)})$. Thus, by (5.23), $L_{k_1}^{(n)} < s - (S_{k_1}^{(n)} - S_\ell^{(n)}) \leq F^{(n)}(S_\ell^{(n)}) - (S_{k_1}^{(n)} - S_\ell^{(n)}) \leq F^{(n)}(S_{k_1}^{(n)})$.

If $\ell = k_0$, then, because $\mathcal{U}^{(n)}(S^{(n)}_{k_0})\{s\} > 0$, we have $\mathcal{W}^{(n)}_S(S^{(n)}_{k_0})\{s\} > 0$ by the definition of $\mathcal{U}^{(n)}$. However, in this case $\mathcal{W}^{(n)}(S^{(n)}_{k_0})\{s\} = 0$, because $W^{(n)} \equiv 0$ on $[\tau^{(n)}_{k-1}, \sigma^{(n)}_k)$ by (4.45) and Lemma 5.3, so $\mathcal{W}^{(n)}(S^{(n)}_{k_0}) = \mathcal{W}^{(n)}(\sigma^{(n)}_k) = v^{(n)}_{k_0}\delta_{L^{(n)}_{k_0}}$ and $s > L^{(n)}_{k_1} + S^{(n)}_{k_1} - S^{(n)}_{k_0} \geq L^{(n)}_{k_0}$ by the definitions of ℓ and s. Thus, a customer with lead time equal to s at time $S^{(n)}_{k_0}$ has already been in service in the EDF system with reneging, so $L^{(n)}_{k_1} + S^{(n)}_{k_1} - S^{(n)}_{k_0} < s \leq F^{(n)}(S^{(n)}_{k_0})$ and consequently, by (5.23), $L^{(n)}_{k_1} < F^{(n)}(S^{(n)}_{k_0}) - (S^{(n)}_{k_1} - S^{(n)}_{k_0}) \leq F^{(n)}(S^{(n)}_{k_1})$.

Thus, regardless of the value of ℓ, $L^{(n)}_{k_1} < F^{(n)}(S^{(n)}_{k_1})$. 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 (5.8), we have $\triangle(U^{(n)} - W^{(n)})(\tau) \leq \mathcal{W}^{(n)}(\tau-)\{0\}$, and so $\triangle(U^{(n)} - W^{(n)})(\tau) \leq \triangle D^{(n)}(\tau)$. This, together with (5.27), contradicts (5.26).

Case II. For a customer k_1 chosen as described above, assume that every customer ℓ arriving in the time interval $[\sigma^{(n)}_k, S^{(n)}_{k_1}]$ 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 $(0, L^{(n)}_{k_1} + S^{(n)}_{k_1} - S^{(n)}_{\ell}]$ upon arrival. Customers ℓ who are less urgent then k_1 must have lead times satisfying $L^{(n)}_{\ell} > L^{(n)}_{k_1} + S^{(n)}_{k_1} - S^{(n)}_{\ell}$, and hence the mass brought by such customers must be initially assigned to the half-line $(L^{(n)}_{k_1} + S^{(n)}_{k_1} - S^{(n)}_{\ell}, \infty)$ in both systems. Then for every $t \in [\sigma^{(n)}_k, S^{(n)}_{k_1}]$, we have

$$\mathcal{W}^{(n)}(t)(0, L^{(n)}_{k_1} - (t - S^{(n)}_{k_1})] \leq \mathcal{U}^{(n)}(t)(0, L^{(n)}_{k_1} - (t - S^{(n)}_{k_1})], \tag{5.28}$$

as we now explain. Under the Case II assumption the arrival of new mass is the same on both sides of (5.28). Furthermore, disregarding lateness and new arrivals, both sides of (5.28) decrease at unit rate so long as they are nonzero. Finally, by (5.18) the amount of late work removed from the EDF system with reneging in the time interval $[\sigma^{(n)}_k, t]$ is greater than or equal to the amount of late work removed from $\mathcal{U}^{(n)}$ in this time interval. Therefore, (5.28) holds for every $t \in [\sigma^{(n)}_k, S^{(n)}_{k_1}]$.

We claim that (5.28) in fact holds for all $t \in [\sigma^{(n)}_k, \tau)$. Suppose this is not the case. Let

$$\eta \triangleq \inf\{t \in [S^{(n)}_{k_1}, \tau) | \mathcal{W}^{(n)}(t)(0, L^{(n)}_{k_1} - (t - S^{(n)}_{k_1})] > \mathcal{U}^{(n)}(t)(0, L^{(n)}_{k_1} - (t - S^{(n)}_{k_1})]\}. \tag{5.29}$$

The strict inequality in (5.29) can occur only because of an arrival at time t which brings mass to the interval $(0, L^{(n)}_{k_1} - (t - S^{(n)}_{k_1})]$ 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, $\eta > S^{(n)}_{k_1}$.

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

$$\mathcal{W}^{(n)}(t)\{L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})\} > 0, \tag{5.30}$$

because the customer k_1 is present in the EDF system with reneging at time t. By (4.4), (5.30) and the definition of η, we have $E^{(n)}(t) \leq L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})$ for $t \in [S_{k_1}^{(n)}, \eta)$. Thus, $E^{(n)}(t-) \leq L_{k_1}^{(n)} - (t - S_{k_1}^{(n)})$ 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 (4.55), 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 (5.28) holds for $t \in [S_{k_1}^{(n)}, \eta]$, which contradicts the definition of η. We have shown that (5.28) holds for $t \in [\sigma_k^{(n)}, \tau)$.

Letting $t \uparrow \tau$ in (5.28) and using the fact that $L_{k_1}^{(n)} - (\tau - S_{k_1}^{(n)}) = 0$, we get $\mathcal{W}^{(n)}(\tau-)\{0\} \leq \mathcal{U}^{(n)}(\tau-)\{0\}$. Thus, by (5.8), $\triangle(U^{(n)} - W^{(n)})(\tau) = \mathcal{W}^{(n)}(\tau-)\{0\} - \mathcal{U}^{(n)}(\tau-)\{0\} \leq 0$ which, together with (5.27) and the fact that $D^{(n)}$ is nondecreasing, contradicts (5.26). □

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 k–th customer arriving at the hybrid system joins the high-priority class if and only if

$$L_k^{(n)} < F^{(n)}(S_k^{(n)}). \tag{5.31}$$

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 k–th 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 5.7 *For every $t \geq 0$, we have*

$$U^{(n)}(t) - W^{(n)}(t) \tag{5.32}$$
$$\leq \sum_{k=1}^{A^{(n)}(t)} v_k^{(n)} \wedge \Big(\mathcal{W}^{(n)}(S_k^{(n)}-)\big(0, F^{(n)}(S_k^{(n)})\big) + v_k^{(n)} - L_k^{(n)}\Big)^+ \mathbb{I}_{\{L_k^{(n)} < F^{(n)}(S_k^{(n)})\}}.$$

Proof: By Lemma 5.6, it suffices to show that $D^{(n)}(t)$ is not greater than the right-hand side of (5.32). 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 (5.31) 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 (5.31) holds, then the k-th 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 (5.23), $L_l^{(n)} - (S_k^{(n)} - S_l^{(n)})$, the lead time of the l-th customer at time $S_k^{(n)}$, is smaller that $F^{(n)}(S_k^{(n)})$. Thus, if (5.31) holds, the k-th 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

$$\mathcal{W}_H^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) \leq \mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})), \tag{5.33}$$

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 (5.31) holds, the time required for the hybrid system to fully serve the k-th customer is at most $\mathcal{W}^{(n)}(S_k^{(n)}-)(0, F^{(n)}(S_k^{(n)})) + v_k^{(n)}$. Consequently, under the assumption (5.31), the unfinished work deleted by the hybrid system due to lateness of the k-th customer is at most $v_k^{(n)} \wedge \Big(\mathcal{W}^{(n)}(S_k^{(n)}-)\big(0, F^{(n)}(S_k^{(n)})\big) + v_k^{(n)} - L_k^{(n)}\Big)^+$. Therefore, 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 (5.32). $\square$

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 5.2 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:

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

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

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

6.1 Proofs of main results concerning the workload

6.1.1 Proof of Theorem 3.3

In Lemma 6.1, we use the continuity property of the mapping Φ established in Lemma 4.1, along with the known 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. In what follows, $\Lambda_{H(0)} : D[0,\infty) \to D[0,\infty)$ is the mapping defined, for every $\phi \in D[0,\infty)$ and $t \geq 0$, by

$$\Lambda_{H(0)}(\phi)(t) \triangleq \phi(t) - \sup_{s\in[0,t]} \left[(\phi(s) - H(0))^+ \wedge \inf_{u\in[s,t]} \phi(u) \right]. \tag{6.4}$$

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

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

Lemma 6.1 *The process U^* satisfies*

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

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, we have

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

for $t \geq 0$. 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 $U^* = \Lambda_{H(0)}\left(W_S^*\right)$. 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 statement of the lemma.

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 now 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 5.2 to show that the workload in the reference and reneging systems are equal with high probability and so, in particular, 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 [34]. 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 non-empty. 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 heavy traffic 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 [5], 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 $n \to \infty$,*

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

Proof: For $k \ge 1$, $\mathcal{W}^{(n)}(S_k^{(n)}-)\big(0, F^{(n)}(S_k^{(n)})\big) \le \mathcal{W}^{(n)}(S_k^{(n)})\big(0, F^{(n)}(S_k^{(n)})\big)$, because customers with strictly positive lead times do not renege. Thus, by Lemmas 5.3 and 5.7, to prove (6.8), it suffices to show that as $n \to \infty$,

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

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

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

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

Theorem 3.3 now follows immediately from Lemma 6.1 and Corollary 6.3.

6.1.2 Proofs of Proposition 3.4 and Theorem 3.5

We now present the proofs of the remaining two limit theorems concerning the measure-valued workload processes. For this, we will require two preliminary results. The first, Lemma 6.4, which states that the frontier in the reneging system is strictly positive with high probability, follows easily from the results proved in the last section. The second result, Proposition 6.5, is a recap of a known result established in [5].

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

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

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

$$\begin{aligned} F^{(n)}(t) &\ge L_{A^{(n)}(t)}^{(n)} - (t - S_{A^{(n)}(t)}^{(n)}) \\ &\ge \sqrt{n} y_* - u_{A^{(n)}(t)+1}^{(n)} \\ &\ge \sqrt{n} y_* - \max_{1 \le k \le A^{(n)}(nT)+1} u_k^{(n)}. \end{aligned} \tag{6.10}$$

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

Proposition 6.5 (Proposition 3.4 [5]) *Let $-\infty < y_0 < y^*$ and $T > 0$ be given. As $n \to \infty$,*

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

Proof of Proposition 3.4: 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 $y \ge y^*$,

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

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

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

For any $t \ge 0$,

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

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 3.3, yields $\widehat{\mathcal{V}}^{(n)}(\widehat{F}^{(n)}, \infty) \Rightarrow W^*$. Thus, by (6.12), we have

$$H(\widehat{F}^{(n)}) \Rightarrow W^* \tag{6.14}$$

in $D_{\mathbb{R}}[0, T]$. Applying the continuous function H^{-1} to both sides of (6.14) and using (3.4), we obtain $\widehat{F}^{(n)} \Rightarrow F^*$ in $D_{\mathbb{R}}[0, T]$. □

Proof of Theorem 3.5: Let $T > 0$. We shall prove that $\widehat{\mathcal{W}}^{(n)} \Rightarrow \mathcal{W}^*$ in $D_{\mathcal{M}}[0, T]$; the proof of the convergence $\widehat{\mathcal{Q}}^{(n)} \Rightarrow \mathcal{Q}^*$ is similar. We claim that

$$\sup_{y \in [0, y^*]} \sup_{t \in [0, T]} \left| \widehat{\mathcal{W}}^{(n)}(t)(y, \infty) - H(\widehat{F}^{(n)}(t) \vee y) \right| \xrightarrow{P} 0. \tag{6.15}$$

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

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

Therefore, (6.15) follows from (6.11) and Proposition 6.2. Define a mapping $\psi : \mathbb{R} \to \mathcal{M}$ by the formula

$$\psi(x)(B) \triangleq \int_{B\cap[x,\infty)} (1 - G(\eta))\, d\eta, \qquad x \in \mathbb{R},\ B \in \mathcal{B}(\mathbb{R}).$$

It is easy to see that ψ is continuous. Hence, by Proposition 3.4,

$$\psi(\widehat{F}^{(n)}) \Rightarrow \psi(F^*) = \mathcal{W}^*. \tag{6.16}$$

On the other hand, (6.15) may be rewritten in the form

$$\sup_{0\le y\le y^*} \sup_{0\le t\le T} \left| \widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - \psi(\widehat{F}^{(n)}(t))(y,\infty) \right| \xrightarrow{P} 0. \tag{6.17}$$

For every $t \ge 0$, $\widehat{\mathcal{W}}^{(n)}(t)(-\infty, 0] = \widehat{\mathcal{W}}^{(n)}(t)(y^*, \infty) = 0$ 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, (6.16)–(6.17) imply that $\widehat{\mathcal{W}}^{(n)} \Rightarrow \mathcal{W}^*$ in $D_{\mathcal{M}}[0, T]$. □

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 3.7. In order to do this, it is convenient to first 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 $\widehat{\mathcal{W}}_S^{(n)}$, $\widehat{\mathcal{U}}^{(n)}$, $\widehat{W}^{(n)}$, $n \in \mathbb{N}$, $\mathcal{W}_S^*$, $\mathcal{U}^*$, and W^* can be defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that $\mathbb{P}$ almost surely, as $n \to \infty$,*

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

$$\widehat{W}_S^{(n)} \to W_S^*, \tag{6.19}$$

$$\widehat{\mathcal{W}}_S^{(n)}(\cdot)(-\infty, 0] \to \mathcal{W}_S^*(\cdot)(-\infty, 0] = \big(W_S^*(\cdot) - H(0)\big)^+, \tag{6.20}$$

$$\widehat{U}^{(n)} \to U^* \tag{6.21}$$

and

$$\widehat{W}^{(n)} \to W^* \triangleq U^*, \tag{6.22}$$

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,

$$U^* = \Lambda_{H(0)}(W_S^*) = W_S^* - K_+^* + K_-^*, \tag{6.23}$$

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

$$\int_{[0,\infty)} \mathbb{I}_{\{U^*(s)<H(0)\}}\, dK^*_+(s) = 0, \quad \int_{[0,\infty)} \mathbb{I}_{\{U^*(s)>0\}}\, dK^*_-(s) = 0. \tag{6.24}$$

The almost sure limits in (6.18)—(6.22) hold uniformly on compact intervals.

Proof. Recall from Theorem 3.2 that $\widehat{\mathcal{W}}^{(n)}_S \Rightarrow \mathcal{W}^*_S$. Using the Skorokhod Representation Theorem, we can construct the model primitives $u^{(n)}_j$, $v^{(n)}_j$ and $L^{(n)}_j$ 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 $\widehat{\mathcal{W}}^{(n)}_S$, $n \in \mathbb{N}$, and the limiting process $\mathcal{W}^*_S$ are defined on this space and (6.18) 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 further have (6.19). In addition, 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 $\int_{B\cap[F^o_S(t),\infty)}(1 - G(y))dy$ for some RCLL process $F^o_S(t)$. However, $W^*_S(t) = \int_{\mathbb{R}\cap[F^o_S(t),\infty)} \big(1 - G(y)\big)du = H\big(F^o_S(t)\big)$; hence $F^o_S(t) = F^*_S(t)$ 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 BM 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 continuous (non-atomic), we also have (6.20).

Now, following (4.1) and (6.3), we set $\mathcal{U}^{(n)} = \Phi(\mathcal{W}^{(n)}_S)$ 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 clearly also defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and (6.21) follows from Lemma 4.1. This implies (6.22). 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 (6.23) and (6.24) is an immediate consequence of the statements following (6.4) — in particular, relations (6.5) and (6.6).

Since the model primitives $u^{(n)}_j$, $v^{(n)}_j$ and $L^{(n)}_j$ for $j \in \mathbb{N}$ and $n \in \mathbb{N}$ are all defined on $(\Omega, \mathcal{F}, \mathbb{P})$, so 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 (6.22), the almost sure counterpart to Theorem 3.3, holds. □

The assertion of Theorem 3.7 is that

$$\widehat{R}^{(n)}_W \Rightarrow K^*_+, \tag{6.25}$$

where K^*_+ is the local time for U^* at $H(0)$ appearing in (6.23). For $T < \infty$, define

$$\mathcal{Z}_n(T) \triangleq \left\{ \widehat{R}^{(n)}_U(t) = \widehat{R}^{(n)}_W(t), \quad 0 \le t \le T \right\}. \tag{6.26}$$

From the workload evolution equations (4.73) and (4.74), 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 3.7, it suffices to show that

$$\widehat{K}_+^{(n)} \Rightarrow K_+^*. \tag{6.27}$$

For $n \in \mathbb{N}$ and $k \geq 1$, recall the definitions of $\tau_{k-1}^{(n)}$ and $\sigma_k^{(n)}$ given in (4.11) and (4.12), respectively, and define

$$\widehat{\tau}_{k-1}^{(n)} \triangleq \frac{1}{n}\tau_{k-1}^{(n)} \qquad \text{and} \qquad \widehat{\sigma}_k^{(n)} \triangleq \frac{1}{n}\sigma_k^{(n)}.$$

Applying the heavy traffic scaling to (4.13), it is easy to see that for $t \geq 0$,

$$\widehat{K}_+^{(n)}(t) = \sum_{k\in\mathbb{N}} \left[\widehat{W}_S^{(n)}(\widehat{\sigma}_k^{(n)}-) \vee \max_{s\in[\widehat{\sigma}_k^{(n)}, t\wedge\widehat{\tau}_k^{(n)}]} \widehat{\mathcal{W}}_S^{(n)}(s)(-\infty,0] - \widehat{W}_S^{(n)}(\widehat{\sigma}_k^{(n)}-) \right]. \tag{6.28}$$

Keeping in mind the limits in (6.18) and (6.20), we introduce the related process

$$\widehat{Y}^{(n)}(t) \triangleq \sum_{k\in\mathbb{N}} \left[W_S^*(\widehat{\sigma}_k^{(n)}) \vee \max_{s\in[\widehat{\sigma}_k^{(n)}, t\wedge\widehat{\tau}_k^{(n)}]} \big(W_S^*(s) - H(0)\big)^+ - W_S^*(\widehat{\sigma}_k^{(n)}) \right] \tag{6.29}$$

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

$$\varepsilon^{(n)}(t) \triangleq \widehat{Y}^{(n)}(t) - \widehat{K}_+^{(n)}(t), \qquad \forall t \geq 0. \tag{6.30}$$

It is clear from the definition that $\widehat{Y}^{(n)}$ is non-decreasing 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,*

$$\int_{[0,T]} \mathbb{I}_{\{U^*(t)<H(0)\}} d\widehat{Y}^{(n)}(t) = 0. \tag{6.31}$$

Proof. Fix $n \in \mathbb{N}$. The first statement follows immediately from (6.28), (6.29), 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

$$W_S^*(\widehat{\sigma}_k^{(n)}) \leq \max_{s \in [\widehat{\sigma}_k^{(n)}, t]} (W_S^*(s) - H(0))^+ = W_S^*(t) - H(0). \tag{6.32}$$

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 (6.29), it is easy to deduce that $\widehat{Y}^{(n)}$ is also constant on each such interval. Thus, to establish (6.31), it only remains to show that for each $k \geq 1$,

$$\int_{J_k^{(n)}} \mathbb{I}_{\{U^*(t) < H(0)\}} d\widehat{Y}^{(n)}(t) = 0. \tag{6.33}$$

Fix $t \in J_k^{(n)}$ and note that by the equality in (6.23) and the definition (6.4) of $\Lambda_{H(0)}$, we have $U^*(t) = W_S^*(t) - K^*(t)$, where

$$K^*(t) \triangleq \sup_{s \in [0,t]} \left[(W_S^*(s) - H(0))^+ \wedge \inf_{u \in [s,t]} W_S^*(u) \right]. \tag{6.34}$$

Also, note that

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

and that the equality in (6.32) implies

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

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

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

where the equality follows from the inequality in (6.32). 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 (6.33). □

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

$$\mathrm{Osc}(f; [t_1, t_2]) \triangleq \sup\{|f(t) - f(s)| : t_1 \leq s \leq t \leq t_2\}$$

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

$$w_f(\delta; [0,T]) \triangleq \sup\{|f(t) - f(s)| : 0 \leq s \leq t \leq T, |t-s| \leq \delta\}.$$

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

Proof. Fix $T > 0$ and let $\eta > 0$ be arbitrarily small. Using the Kolmogorov–Čentsov theorem (see, e.g., Theorem 2.8, p. 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 $\tilde{\Omega}$ with $\mathbb{P}(\tilde{\Omega}) \geq 1 - \eta$.

For each subsequence in $\mathbb{N}$, there is a sub-subsequence $\mathcal{S}$ along which the limits (6.18)–(6.22) 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 $n \in \mathcal{S}$, $n \geq n_0$,

$$\sup_{t \in [0,T]} \left| \widehat{\mathcal{W}}_S^{(n)}(t)(-\infty, 0] - (W_S^*(t) - H(0))^+ \right| \leq \Delta; \tag{6.35}$$

$$\sup_{t \in [0,T]} \left| \widehat{W}_S^{(n)}(t-) - W_S^*(t) \right| \leq \Delta; \qquad \sup_{t \in [0,T]} \left| \widehat{W}^{(n)}(t-) - W^*(t) \right| \leq \Delta, \tag{6.36}$$

$$\sup_{t \in [0,T]} \left| \widehat{U}^{(n)}(t-) - U^*(t) \right| \leq \Delta. \tag{6.37}$$

From the definitions (6.28) and (6.29), respectively, of $\widehat{K}_+^{(n)}$ and $\widehat{Y}^{(n)}$ it is clear that, for every $k \in \mathbb{N}$ such that $\tau_k^{(n)} \leq T$,

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

Define

$$\begin{aligned} J_n &\triangleq \left\{ k \in \mathbb{N} : \widehat{K}_+^{(n)}(\widehat{\tau}_k^{(n)}) - \widehat{K}_+^{(n)}(\widehat{\sigma}_k^{(n)}-) > 0, \widehat{\tau}_k^{(n)} \leq T \right\}, \\ \tilde{J}_n &\triangleq \left\{ k \in \mathbb{N} : \widehat{Y}^{(n)}(\widehat{\tau}_k^{(n)}) - \widehat{Y}^{(n)}(\widehat{\sigma}_k^{(n)}-) > 0, \widehat{\tau}_k^{(n)} \leq T \right\}, \end{aligned}$$

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

$$\overline{\varepsilon}^{(n)}(T) \triangleq \sup_{s \in [0,T]} \left| \widehat{Y}^{(n)}(s) - \widehat{K}_+^{(n)}(s) \right| \leq 2c^{(n)}\Delta. \tag{6.38}$$

We now claim that

$$k \in [J_n \cup \tilde{J}_n] \quad \Rightarrow \quad \mathrm{Osc}(W^*, [\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)}]) \geq \frac{y_*}{3}. \tag{6.39}$$

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 $k \in [J_n \cup \tilde{J}_n]$ then $\widehat{\tau}_k^{(n)} - \widehat{\sigma}_k^{(n)} \geq \theta^{-1}(y^*/3) > 0$, which in turn implies that $c^{(n)} \leq M$. Substituting this into (6.38), we conclude that for every $\Delta > 0$, there exists $n_0(\Delta) \in \mathcal{S}$ such that for all $n \in \mathcal{S}$, $n \geq n_0(\Delta)$,

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

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 subseqence in $\mathbb{N}$, there is a sub-subsequence 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 (6.39). Note first that by the definition of $H(0)$ and y_*, we have $H(0) \geq y_*$. If $k \in \tilde{J}_n$, 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 (6.37), this implies that the oscillation of U^* on $[\widehat{\sigma}_k^{(n)}, \widehat{\tau}_k^{(n)})$ is no less than $H(0) - \Delta \geq y^*/3$. Since $W^* = U^*$, the conclusion in (6.39) 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

$$W^{(n)}(\sigma_k^{(n)}-) = 0 \tag{6.40}$$

(this follows because $U^{(n)}(\sigma_k^{(n)}-) = 0$ and, by Lemma 5.3, $W^{(n)} \leq U^{(n)}$), this customer must have arrived 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 (6.36) and the fact that $\widehat{W}^{(n)}(\widehat{\sigma}_k^{(n)}-) = 0$ (the scaled version of (6.40)), we conclude that

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

and this gives us the conclusion in (6.39). □

Proof of Theorem 3.7 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

$$U^* = \widehat{W}_S^{(n)} + \delta^{(n)} - \widehat{K}_+^{(n)} + \widehat{K}_-^{(n)} = \widehat{W}_S^{(n)} + \delta^{(n)} + \varepsilon^{(n)} - \widehat{Y}^{(n)} + \widehat{K}_-^{(n)}, \quad (6.41)$$

where (cf. (4.14))

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

$\widehat{K}_+^{(n)}$ is defined by (6.28) and $\varepsilon^{(n)}$ is defined by (6.30). According to (4.16), $\int_0^T \mathbb{I}_{\{\widehat{U}^{(n)}(t)>0\}}\, d\widehat{K}_-^{(n)}(t) = 0$, which implies

$$\int_0^T \mathbb{I}_{\{U^*(t)>\overline{\delta}^{(n)}\}}\, d\widehat{K}_-^{(n)}(t) = 0. \quad (6.42)$$

Since $\widehat{W}_S^{(n)} + \delta^{(n)} + \epsilon^{(n)} \Rightarrow W_S^*$ due to (6.19), (6.21) and Lemma 6.8, and, by (6.23), U^* is obtained by applying the Skorokhod map on $[0, H(0)]$ to W_S^*, the convergence (6.27) is an immediate consequence of (6.41), (6.42), Lemma 6.7, Lemma 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 $k \geq 1$,

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

Then $0 = \rho_0 < \nu_1 < \rho_1 < \nu_2 < \ldots$ 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 (6.41), (6.19), (6.21), 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 $t \in [0, T]$, we have the convergence

$$\widehat{Y}^{(n)}(t) \xrightarrow{P} \sum_{k\in\mathbb{N}} \left[W_S^*\big((t \vee \nu_k) \wedge \rho_k\big) - U^*\big(\big((t \vee \nu_k) \wedge \rho_k\big) - \big(W_S^*(\nu_k) - U^*(\nu_k)\big)\big) \right]. \quad (6.43)$$

However, (6.24) implies that for each k, K_-^* is constant on $[\nu_k, \rho_k]$ and K_+^* is constant on $[\rho_{k-1}, \nu_k]$. Therefore, (6.23) 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 right-hand side of (6.43) 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

$$\max_{t\in[0,T]} \left|\widehat{K}_+^{(n)}(t) - K_+^*(t)\right| \xrightarrow{P} 0. \tag{6.44}$$

Since convergence in probability in a metric space implies weak convergence, we have (6.27). □

7 Performance Evaluation and Simulation Results

In this section we use the heavy traffic approximations of this paper to evaluate the performance of the queueing system with reneging and compare its performance with that of the system in which all customers are served to completion. The predictions of the theory are summarized in Subsection 7.1, these are compared to simulation results in Subsection 7.2, and Subsection 7.3 provides derivations of the results set out in Subsection 7.1 The simulation results in Subsection 7.2 attest to the accuracy of the approximations derived from the theory and also show the great difference in performance of the reneging system and the system in which customers are served to completion.

7.1 Summary of theory predictions

Consider a single-station queue with traffic intensity $\rho = \lambda/\mu$ that is near but slightly less than 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 $\sigma^2 = \lambda(\alpha^2 + \beta^2)$. Let $\overline{D}$ denote the mean lead time for arriving customers. Finally, set $\theta = 2(1-\rho)/\sigma^2$.

Under these circumstances, the results of this paper predict that in the long-run,

$$\text{Fraction of lost work in reneging system} \approx e^{-\theta\overline{D}} \left(\frac{1-\rho}{\rho(1-e^{-\theta\overline{D}})}\right). \tag{7.1}$$

This formula is derived as (7.9) below, which embeds the system in a sequence of systems and hence indexes ρ by the superscript (n)). On the other hand, analysis of the heavy-traffic limit of the standard (non-reneging) system reveals that in the long run (see (7.13)),

$$\text{Fraction of late work in standard system} \approx e^{-\theta\overline{D}}, \tag{7.2}$$

and the ratio of these two quantities is small. Indeed (see (7.14)),

$$\frac{\text{Lost work in reneging system}}{\text{Late work in standard system}} \approx \frac{1-\rho}{\rho(1-e^{-\theta\overline{D}})}. \tag{7.3}$$

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 $\overline{D}$ is chosen so that $e^{\theta\overline{D}} >> 1$, and these two plots will be separated by $\log((1-\rho)/\rho)$. When we measure performance 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 system in which work is served to completion.

The fraction of lost customers in the reneging system can be estimated by a heuristic argument, at least when the arrival process is Poisson, and turns out to be (see (7.12))

$$\begin{aligned}&\text{Fraction of lost customers in reneging system}\\ &\qquad\approx \frac{2}{\mu^2\beta^2+1}\times(\text{Fraction of lost work in reneging system}).\end{aligned} \tag{7.4}$$

For the standard system, we have the simpler relationship (see (7.13))

$$\begin{aligned}&\text{Fraction of late customers in standard system}\\ &\qquad\approx \text{Fraction of late work in standard system.}\end{aligned} \tag{7.5}$$

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 4 and an M/D/1 system presented in Figure 5. 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[5,B] with the mean deadline $\overline{D} = \frac{5+B}{2}$, varying from 5 (constant deadlines) to 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 (7.1) and (7.4) for the fraction of work lost and the fraction of lost customers for the reneging system. Equation (7.4) is derived in Subsection 7.3 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 4 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. One can see the nearly perfect agreement between the theoretical approximation and the simulation results. 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"

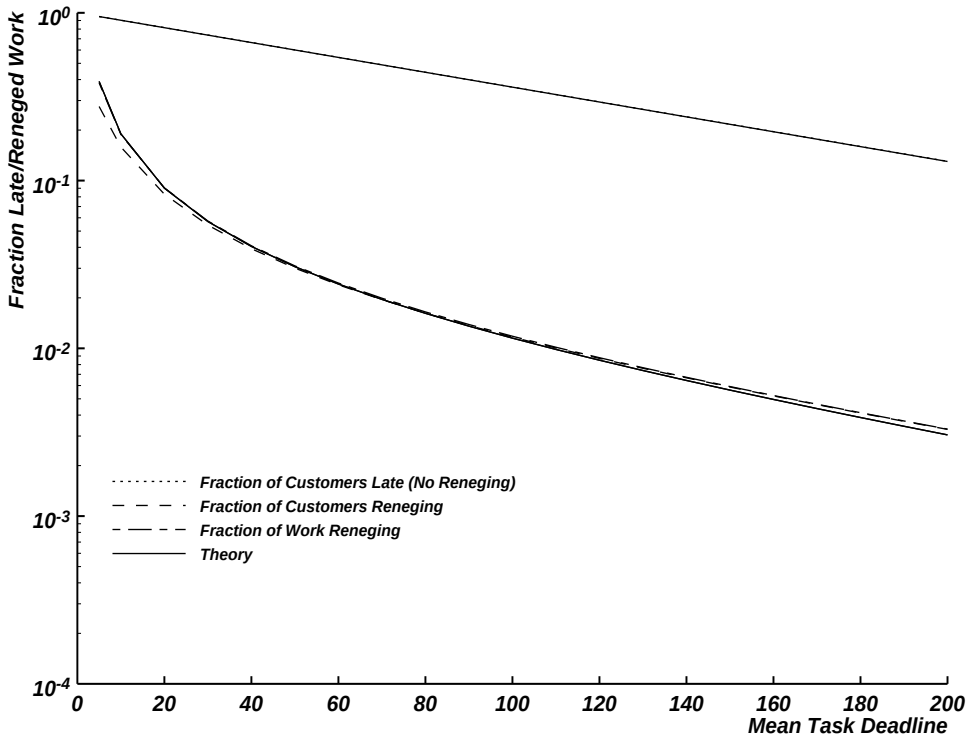

Figure 4: M/M/1 Queue

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 (7.1)–(7.3) that for sufficiently large values of $\overline{D}$, the performance of the reneging system is parallel on a log scale to that the standard system with the two curves separated by approximately 0.02. This corresponds to a reduction in work dthat misses its deadline by a factor of 40 to 50.

Figure 5 presents the results for the M/D/1 system. The results are qualitatively identical to those of Figure 4, 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 $\theta = .02$ 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

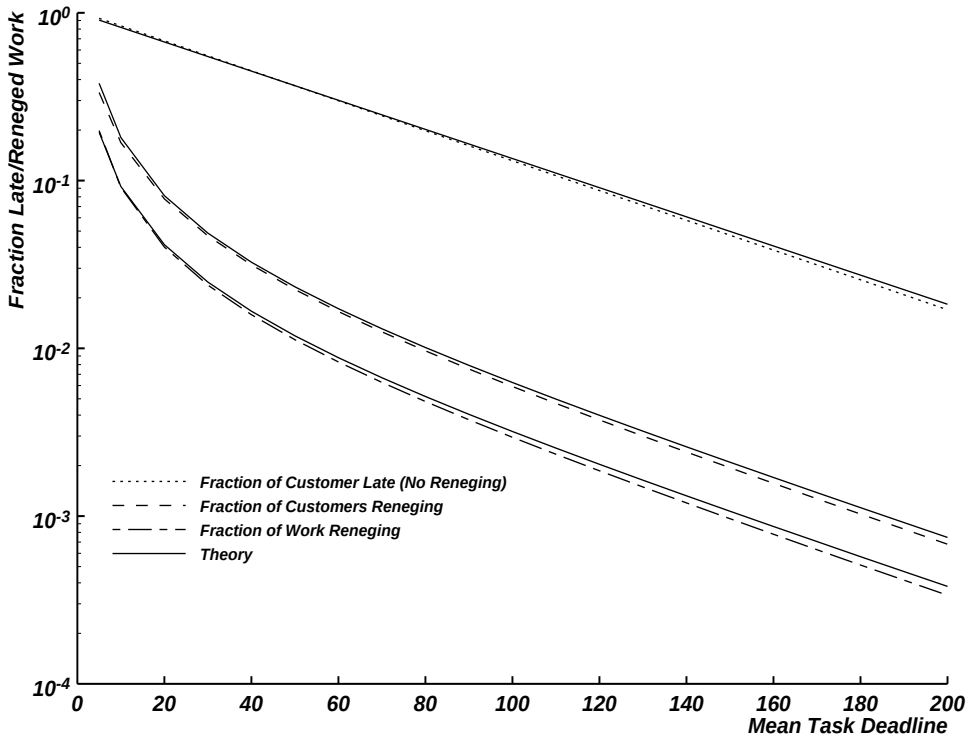

Figure 5: M/D/1 Queue

service on customers when their deadlines expire.

7.3 Derivation of theory predictions

This subsection derives the formulas (7.1)–(7.5) reported in Subsection 7.1. We begin with a main result of this paper, Theorem 3.3, which states that the limiting scaled workload in the reneging system is a reflected Brownian motion in $[0, H(0)]$ with drift. More specifically, the limiting scaled workload process is

$$W^*(t) = W_S^*(t) - K_+^*(t) + K_-^*(t), \tag{7.6}$$

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 3.7 that $K_+^*(t)$ is the limit of the scaled workload that reneges prior to time t in the diffusion scaling, i.e., $\sqrt{n}K_+^*(t)$ is approximately the (unscaled) workload that reneges in the n-th system prior to time nt.

Lemma 7.1 ([13], Proposition 5, p. 90) *We have*

$$\lim_{t\to\infty}\frac{1}{t}K_+^*(t)=\lim_{t\to\infty}\frac{1}{t}\mathbb{E}K_+^*(t)=\begin{cases}\frac{\gamma}{e^{2\gamma H(0)/\sigma^2}-1} & \text{if } \gamma\neq 0,\\ \frac{\sigma^2}{2H(0)} & \text{if } \gamma=0.\end{cases} \tag{7.7}$$

Proof: The first equality in (7.7) is a consequence of the fact that W^* has a stationary distribution (see (7.10) 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

$$\begin{aligned} & f'(0)\mathbb{E}\big[I_S^*(t)+K_-^*(t)\big]-f'\big(H(0)\big)\mathbb{E}K_+^*(t)\\ = \;& \mathbb{E}\int_0^t\left[\gamma f'\big(W^*(s)\big)-\frac{1}{2}\sigma^2 f''\big(W^*(s)\big)\right]ds+\mathbb{E}f(W^*(t))-f(0). \end{aligned} \tag{7.8}$$

Taking $f(x)=x$, we obtain

$$\mathbb{E}[I_S^*(t)+K_-^*(t)]-\mathbb{E}K_+^*(t)=\gamma t+\mathbb{E}W^*(t)-f(0).$$

Taking $f(x)=\frac{\sigma^2}{2\gamma}e^{2\gamma x/\sigma^2}$, we further obtain

$$\mathbb{E}\big[I_S^*(t)+K_-^*(t)\big]-e^{2\gamma H(0)/\sigma^2}\mathbb{E}K_+^*(t)=\frac{\sigma^2}{2\gamma}(\mathbb{E}e^{2\gamma W^*(t)/\sigma^2}-1)$$

if $\gamma\neq 0$. Solving these equations for $\mathbb{E}K_+^*(t)$, dividing by t, and letting $t\to\infty$, we obtain the second equality in (7.7) 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 n-th system by time nt is $V^{(n)}(A^{(n)}(nt))=\sqrt{n}\widehat{N}^{(n)}(t)+nt$. But, $\widehat{N}^{(n)}$ is approximately equal to 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, we observe from (2.4) that the cumulative distribution function for the lead time in the n-th system is $\mathbb{P}\{L_j^{(n)}\le y\}=G(y/\sqrt{n})$, and hence the expected lead time in the n-th 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 n-th system when $\gamma\neq 0$ is approximately

$$\frac{1-\rho^{(n)}}{\rho^{(n)}(e^{2(1-\rho^{(n)})\mathbb{E}L_j^{(n)}/\sigma^2}-1)}=\frac{1-\rho^{(n)}}{\rho^{(n)}(e^{\theta\overline{D}}-1)}, \tag{7.9}$$

where

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

We have suppressed the dependence of θ and $\overline{D}$ on n, which will remain fixed for this discussion. If $\gamma = 0$, then in place of (7.9) we have $\frac{\sigma^2}{2\overline{D}}$.

Remark 7.2 Corollary 3.6 implies also 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 workload that reneges in the n-th system prior to nt. The simulations indicate that this naive interpretation of Corollary 3.6 applied to the queue length process is incorrect, as does the following heuristic. Developing a rigorous understanding of this situation is a subject of future work.

According to [13], Proposition 5, p. 90, if $\gamma \neq 0$, the stationary density for W^* is

$$\varphi^*(x) \triangleq \begin{cases} \frac{2\gamma e^{-2\gamma x/\sigma^2}}{\sigma^2(1-e^{-2\gamma H(0)/\sigma^2})} & \text{if } 0 \leq x \leq H(0), \\ 0 & \text{otherwise,} \end{cases} \tag{7.10}$$

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

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

We have suppressed the dependence of φ on n.

Suppose now that that lead-times of arriving customers are not random. Then in the n-th 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 $W + V \leq \overline{D}$. 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, p. 218), an arriving customer will encounter a workload W having approximately the distribution φ. We compute

$$\begin{aligned} \mathbb{P}\{\text{Customer reneges}\} &= \mathbb{P}\{W > \overline{D} - V\} \\ &= \mathbb{E}\big[\mathbb{P}\{W > \overline{D} - V | V\}\big] \\ &\approx \mathbb{E}\left[\int_{(\overline{D}-V)^+}^{\overline{D}} \varphi(w)\, dw\right]. \end{aligned}$$

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

$$\mathbb{P}\{\text{Customer reneges}\} \approx \frac{1}{e^{\theta \overline{D}} - 1}\big(\mathbb{E}e^{\theta V} - 1\big). \tag{7.11}$$

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

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

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

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

The last expression is, perhaps not surprisingly, the formula for the average residual lifetime of a renewal cycle; see [31], Example 3.6(b), p. 80-81. 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:

$$\begin{aligned}
&\text{Fraction of lost customers in reneging system} \\
&\quad \approx \frac{\text{Fraction of lost work in reneging system}}{\mathbb{E}[V + W - \overline{D}|\text{Customer reneges}]} * \mathbb{E}V \\
&\quad \approx \frac{2(\mathbb{E}V)^2}{\mathbb{E}\big[V^2\big]} \times (\text{Fraction of lost work in reneging system}).
\end{aligned} \tag{7.12}$$

This is (7.4) 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 (7.12) implies that the fraction of customers who renege will be approximately equal to the fraction of work that reneges. See Figure 4 for a simulation confirmation of this assertion. On the other hand, if V is nonrandom, hence equal to its mean $1/\mu$, then (7.12) predicts that the fraction of customers who customers renege will be 2 times the fraction of work that reneges. See Figure 5 for a simulation confirmation of this assertion. Both these conclusions hold irrespective of the value of λ.

The last conclusion is inconsistent with a naive interpretation of Corollary 3.6, 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 3.6 would say that the fraction of work reneging would approximately agree with the fraction of customers reneging regardless of the distribution of V. $\square$

We next turn our attention to the performance of the standard (non-reneging) 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

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

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

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

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 n-th system sees lateness if and only if $W^{(n)}$ exceeds $\overline{D} = \sqrt{n} H(0)$. In other words, the theory predicts that

$$\begin{aligned} &\text{Fraction of late customers in standard system} \\ &\quad = \text{Fraction of late work in standard system} \\ &\quad = e^{-\theta \overline{D}}. \end{aligned} \tag{7.13}$$

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

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

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 (7.9), with the fraction of work that is late in the standard (non-reneging) system. The ratio of these quantities of lost/late work is

$$\frac{\text{Lost work in reneging system}}{\text{Late work in standard system}} \approx \frac{e^{\theta \overline{D}}}{e^{\theta \overline{D}} - 1} \left(\frac{1 - \rho^{(n)}}{\rho^{(n)}} \right). \tag{7.14}$$

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 (7.14) is $O(1/\sqrt{n})$. In terms of customers whose service requirement is

not met by the time their deadlines elapse, the reneging system is also superior; the determination of a counterpart to the ratio (7.14) for customers rather than work is a subject of future research.

References

[1] Asmussen, S., *Applied Probability and Queues*, 2nd Edition, Springer, New York. 2003.

[2] Bernard, A. and El Kharroubi, A. Régulation de processus dans le premier orthant de $\mathbb{R}^n$. *Stochast. Stochast. Rep.*, **34**, 149–167, 1991.

[3] Billingsley, P., *Probability and Measure*, 2nd Edition, Wiley, New York, 1986.

[4] Billingsley, P., *Convergence of Probability Measures*, 2nd Edition, Wiley, New York, 1999.

[5] Doytchinov, B., Lehoczky, J. P. and Shreve, S. E., Real-time queues in heavy traffic with earliest-deadline-first queue discipline, *Ann. Appl. Probab.* **11**, 332–378, 2001.

[6] Chen, H. and Mandelbaum, A. Stochastic discrete flow networks: diffusion approximations and bottlenecks. *Ann. Probab.*, 19:1436–1519, 1991.

[7] Decreusefond, L. and Moyal, P., Fluid limit of a heavily loaded EDF queue with impatient customers, http://www.infres.enst.fr/~decreuse/recherche/lfgn _arxiv.pdf, 2006.

[8] Dupuis, P. and Ramanan, K., Convex duality and the Skorokhod problem. *Probab. Theor. Rel. Fields*, **115**, 2:197–236, 1999.

[9] Ethier, S. N. and Kurtz, T. G, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1985.

[10] Gamarnik, D. and Zeevi, A., Validity of heavy traffic steady-state approximations in generalized Jackson networks, *Ann. Appl. Probab.* **16**, 56–90, 2006.

[11] Gromoll, H. C. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.*, **14**, 2:555–611, 2004.

[12] Gromoll, H. C. and Kruk, L. Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.*, **14**, 2:555–611, 2007.

[13] Harrison, J. M. *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.

[14] Harrison, J. M. and Reiman, M. Reflected Brownian motion in an orthant. *Ann. Probab.*, 1981.

[15] Iglehart, D. and Whitt, W., Multiple channel queues in heavy traffic I, *Adv. Appl. Probab.* **2**, 150–177, 1970.

[16] Iglehart, D. and Whitt, W., The equivalence of functional central limit theorems for counting processes and associated partial sums, *Annals of Mathematical Statistics* **42**, 1372-1378, 1971.

[17] Karatzas, I. and Shreve, S. E., *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York 1988.

[18] Kaspi, H. and Ramanan, K. Law of large numbers limits for many-server queues. *Preprint,* http://arxiv.org/abs/math.PR/0708.0952, 2007.

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

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

[21] Kruk, Ł., Diffusion approximation for a G/G/1 EDF queue with unbounded lead times, *Annales UMCS Mathematica A*, to appear.

[22] Kruk, Ł., Lehoczky, J. P. and Shreve, S. E., Second order approximation for the customer time in queue distribution under the FIFO service discipline, *Annales UMCS Informatica AI* **1**, 37-48, 2003.

[23] Kruk, Ł., Lehoczky, J. P. Shreve, S. E. and Yeung, S.-N., Earliest-deadline-first service in heavy-traffic acyclic networks, *Annals of Applied Probability* **14 (3)**, 1306-1352, 2004.

[24] Kruk, Ł., Lehoczky, J. P. and Shreve, S. E., Accuracy of state space collapse for earliest-deadline-first queues, *Annals of Applied Probability* **16 (2)**, 516-561, 2006.

[25] Kruk, Ł., Lehoczky, J. P. Ramanan, K. and Shreve, S. E., An explicit formula for the Skorokhod map on [0,a], to appear *Annals of Probability*, 2007.

[26] Lehoczky, J. P., Real-time queueing theory, *Proceedings of the IEEE Real-Time Systems Symposium*, IEEE, New York, 186-195, 1996.

[27] Limic, V. On the behavior of LIFO preemptive resume queues in heavy traffic, *Electron. Comm. Probab.*, **5**, 13–27, 2000.

[28] Panwar, S. S. and Towsley, D., Optimality of the stochastic earliest deadline policy for the G/M/c queue serving customers with deadlines, *Second ORSA Telecommunications Conference*, Boca Raton, March 1992.

[29] Prokhorov, Yu., Convergence of random processes and limit theorems in probability theory, *Theory of Probability and Applications* **1**, 157–214, 1956.

[30] Ramanan, K. and Reiman, M. Fluid and heavy traffic diffusion limits for a generalized processor sharing model. *Ann. Appl. Probab.*, **13** 1:100–139, 2003.

[31] Ross, S., *Stochastic Processes*, Wiley, New York, 1983.

[32] Ward, A. R. and Glynn, P. W., A diffusion approximation for a Markovian queue with reneging, *Queueing Systems* **43**, 103-128, 2003.

[33] Ward, A. R. and Glynn, P. W., A diffusion approximation for a GI/GI/1 queue with balking or reneging, *Queueing Systems* **50**, 371-400, 2005.

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

[35] Whitt, W., *Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues*, Springer-Verlag, New York, 2002.