Real-Time Queues in Heavy Traffic with Earliest-Deadline-First Queue Discipline

Bogdan Doytchinov[1]
John Lehoczky[2]
Steven Shreve[3]
August 27, 1999

ABSTRACT

This paper introduces a new aspect of queueing theory, the study of systems that service customers with specific timing requirements (e.g. due dates or deadlines). Unlike standard queueing theory in which common performance measures are customer delay, queue length and server utilization, real-time queueing theory focuses on the ability of a queue discipline to meet customer timing requirements, e.g., the fraction of customers who meet their requirements and the distribution of customer lateness. It also focuses on queue control policies to reduce or minimize lateness, although these control aspects are not explicitly addressed in this paper.

To study these measures, one must keep track of the lead-times (deadline minus current time) of each customer, hence the system state is of unbounded dimension. A heavy traffic analysis is presented for the earliest deadline first (EDF) scheduling policy. This analysis decomposes the behavior of the real-time queue into two parts: the number in the system (which converges weakly to a reflected Brownian motion with drift) and the set of lead-times given the queue length. The lead-time profile has a limit which is a non-random function of the limit of the scaled queue length process. Hence, in heavy traffic, one can characterize the system as a diffusion evolving on a one-dimensional manifold of lead-time profiles. Simulation results are presented which indicate that this characterization is surprisingly accurate. A discussion of open research questions is also presented.

---

[1]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, bd24@andrew.cmu.edu.

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

[3]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 shreve@cmu.edu.

1 Introduction

This paper introduces a new aspect of queueing theory, the study of systems that service customers with individual timing requirements (e.g. due dates or deadlines). Such systems arise naturally in manufacturing in which orders have due dates. A second category of examples arises in real-time computer and communication systems. Such systems might involve the transmission of digitized voice, video or images over a network. These transmissions must reach their destination within specific deadlines in order to maintain the integrity of the communication (e.g., voice conversation, teleconference or movie). Real-time computer systems also control much of modern technology, for example engines and braking systems in automobiles, all avionic systems (including air traffic control) and all aspects of modern manufacturing facilities. Computerized control systems must receive and react to state information within a fixed, often stringent, time interval in order to maintain proper control over the system. Failure to meet task timing requirements in safety-critical applications can have serious consequences. Thus, the conditions for correct performance of a real-time system include both the logical correctness of each of the tasks that it executes and the timing correctness of those tasks. Over the last decade, there have been significant strides made in the development of a theory of hard real-time systems, systems in which tasks must be completed before their deadline elapses. The reader is referred to the handbook by Klein et. al. [17] for a description of this theory which addresses many practical considerations encountered in computer systems such as operating system overhead, hardware architecture details, concurrency control and other sorts of blocking and task precedence relations.

The scheduling theory described in [17] assumes an essentially deterministic environment. For example, task arrivals are modeled as the superposition of periodic arrival processes, and task service times are deterministic and given by the worst case execution of each task type. Two principal approaches have been developed for assessing the design of real-time systems with periodic task arrivals, one based on a fixed task priority structure (exemplified by *rate monotonic scheduling*) and the other based on dynamic priorities (exemplified by the *earliest deadline first (EDF)* approach to scheduling). These two scheduling algorithms were analyzed by Liu and Layland [22] and the EDF scheduling algorithm was shown to be optimal for this scheduling problem. In some systems, especially communication systems, only a small number of bits are available in each packet to represent the task's priority; thus EDF cannot be fully implemented because it can require an unlimited number of distinct priority categories. Nonetheless, we introduce real-time queueing theory in the context of the EDF scheduling algorithm, since this algorithm is optimal under some conditions. Panwar and Towsley [24] showed that EDF maximizes the fraction of customers meeting their deadlines within the class of work conserving policies allowing preemption in GI/M/1 queues where customers have general deadlines. Bounds on the performance of EDF for M/M/1 queues in which customers have exponential deadlines were developed by Hong, Tan and Towsley [12]. We also discuss the *first-in first-out (FIFO)* queue discipline in Section 5 of this paper.

There are major limitations to any theory which requires periodic arrivals and assumes worst-case execution times. These assumptions are quite narrow and limit the range of systems that can be studied. Multimedia applications or real-time communications can exhibit substantial variability in the arrival of tasks and their work requirements. For real-time systems for which the task sets exhibit substantial variability, one would like to develop approaches based on queueing theory, a theory which was designed to model and predict stochastic system behavior with resource contention. This theory allows randomness in the task arrivals and task execution times. The difficulty with queueing theory is that it typically does not allow for explicit consideration of dynamically changing task timing requirements. Instead, it only permits priorities which allow important tasks or tasks with initial short timing requirements to receive preferential treatment. Much of queueing theory focuses on general system performance measures, such as task delay, queue lengths, processor utilization, etc., and these are usually computed under equilibrium assumptions. It does not model the timing requirements of each customer, nor does it analyze the ability of a scheduling algorithm to meet those timing requirements. What is needed is a new theory which combines the focus on meeting task timing requirements as studied in real-time scheduling theory with the focus on stochastic task sets as studied in queueing theory. This paper represents a step in the direction of building such a theory, hence the name *real-time queueing theory.*

To study whether tasks or customers meet their timing requirements, one must keep track of the customer lead-times, where the lead-time is the time remaining until the deadline elapses, that is

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

Customer lead-times decrease linearly while a customer is in the queue. Because the lead-time must be tracked for each customer, the dimension of the system state is the number of customers in queue, which is unbounded.

This causes analytic difficulties. In spite of this unbounded dimension, a heavy traffic analysis can be carried out. This analysis decomposes the behavior of the real-time queue into two parts: the number in the system, say $Q(t)$ (which is shown under the heavy traffic scaling to converge weakly to a reflected Brownian motion with drift) and the set of lead-times, $(L_1(t), \dots, L_{Q(t)}(t))$ (we refer to this as the lead-time profile), conditional on the queue length. It is convenient to think of this profile as a random counting measure on $\mathbb{R}$. In heavy traffic, under the earliest deadline first queue discipline, it will be shown that when suitably scaled, the lead-time profile converges to a nonrandom function of the limit of the scaled queue length process, the particular function being determined by the distribution of initial deadline of arriving customers. Hence, in heavy traffic, the unbounded dimension process collapses to a one-dimensional process and one can conceptualize the real-time queueing process as a diffusion evolving on a one-dimensional manifold of lead-time profiles. Simulation results, presented in Section 4, indicate that this characterization is surprisingly accurate.

This work is based on the long tradition of heavy traffic queueing theory pioneered by Kingman [16]. This research was generalized in scope and system complexity by a number of authors; for example, see Iglehart and Whitt [14],[15], who study the multiple server case, for a review of this early literature. The use of heavy traffic theory in the study of the behavior of priority queues was initiated by the work of Whitt [32], Hooke [13], Harrison [7] and Kyprianou [18]. The phenomenon of state space collapse, which was originally observed in Reiman [28], [29], also occurs in our work. Specifically, the lead-time profiles have the dimension of the number of customers in the queue, which is unbounded. Nevertheless, in heavy traffic, those random profiles converge to a deterministic manifold of profiles indexed by the queue length, a one-dimensional parameter.

Heavy traffic queueing theory has evolved greatly over the last 25 years, especially for queueing networks carrying multiple customer types. A great increase in interest in this research area came with the work of Harrison and co-authors, e.g., [8], [9], [10], [11], and Peterson [26]. The EDF queue discipline studied in this paper is related to multi-class queues, although in EDF there are an infinite number of distinct priority classes, and customers change classes as they wait in the queue. Most of the work in heavy traffic queueing networks studies the behavior of queue lengths and workloads, rather than focusing on the lead-times of individual customers. We expect that our results on the convergence of the lead-time profiles for the single queue case will carry over to networks, but we do not study network behavior in this paper. Similarly, we expect that one can apply optimal control methods to real-time queues to control customers' lateness in the way that many researchers have used these methods to optimize inventory holding costs; see, for example, Harrison and Wein, [11], [31].

There is some recent work on heavy traffic approximations for systems handling customers with due dates. Of particular importance are the papers by Van Mieghem [30], Markowitz & Wein [23], Doytchinov [4] and Lehoczky [19, 20, 21]. Van Mieghem studies a single server multiclass queueing system with k distinct customer classes. Each class has an associated convex cost of delay, $C_k(\tau)$, with derivative $c_k(\tau)$. The objective is to minimize the total delay cost incurred over a finite time horizon. The paper studies the "generalized $c\mu$ policy" which schedules the customer having maximum value of $\mu_k c_k(a_k(t))$, where μ_k is the service rate for class k and $a_k(t)$ is the age of the oldest customer in class k. Customers are served in FIFO order in each class, which is equivalent to EDF within each class. This policy is shown to be asymptotically optimal in heavy traffic. Generalizations to a countable number of customer classes and several homogeneous servers in a nonstationary, deterministic or stochastic environment are also considered.

Markowitz and Wein [23] study the single machine scheduling problem in a manufacturing context using heavy traffic methods. They give a unified treatment which permits setup costs, customer due-dates and a mixture of standardized and customized products. The analysis assumes a cyclic policy in which different products must be produced in a fixed sequence, but the machine busy/idle policy and lot-sizing decisions are dynamic. As such, the system resembles a polling system. A heavy traffic averaging principle such as characterized by Coffman, Puhalskii and Reiman [2] is assumed to hold, and subject to this assumption, the optimal policy is determined. The paper gives a detailed discussion of the interactions between the setup, due-date and product mix factors.

Doytchinov [4] developed a partial differential equation-based approach to the study of real-time M/M/1 queues in which the arrivals have constant deadlines. In this case, the EDF and FIFO queue disciplines are identical. His methodology proved that the lead-time profiles converge to a uniform distribution in heavy traffic.

Lehoczky [19] gave an informal analysis for the M/M/1 queue based on representing the lead-time profile as a measure-valued Markov process, and then arguing, under heavy traffic conditions, that the generator converges to that of a deterministic profile conditional on the queue length. This was done both for EDF and for processor sharing. Lehoczky [20] used these results to study the behavior of various queue control policies to reduce customer lateness. Lehoczky [21] extended the analysis to Jackson networks.

This paper is organized as follows. In Section 2, we present the basic model, assumptions and notation. Section 3 gives the major theorems describing the heavy traffic limiting behavior of EDF real-time single server queueing systems. Section 4 presents simulation results illustrating the accuracy of the theory. Section 5 presents some conjectures for the extension of the theory of Section 3. Appendices A and B are included to set notation. Appenidix A collects key definitions and theorems related to weak convergence of measures on metric spaces, and Appendix B recalls classical heavy traffic theorems.

2 The Basic Model, Assumptions and Notation

We first define the basic real-time queueing theory model. Because we shall ultimately pass to a heavy traffic limit, we posit a sequence of queueing systems, indexed by n. The assumptions on the n-th queueing system are the following:

A1: There is a single station serving customers.

A2: Customer interarrival times are determined by the sequence of strictly positive i.i.d. random variables $\{u_j^{(n)}\}_{j=1}^{\infty}$ with $E[u_j^{(n)}] = \frac{1}{\lambda^{(n)}}$ and $\mathrm{Var}[u_j^{(n)}] = (\alpha^{(n)})^2$.

A3: Customers have service requirements which are determined by the sequence of nonnegative i.i.d. random variables $\{v_j^{(n)}\}_{j=1}^{\infty}$ with $E[v_j^{(n)}] = \frac{1}{\mu^{(n)}}$ and $\mathrm{Var}[v_j^{(n)}] = (\beta^{(n)})^2$.

A4: Each customer arrives with a hard deadline (initial lead-time) $L_j^{(n)}$. These initial lead-times are i.i.d. with distribution given by

$$I\!P(L_j^{(n)} \leq \sqrt{n}y) = G(y), \tag{2.1}$$

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

$$y^* \stackrel{\Delta}{=} \min\{y \in I\!R; G(y) = 1\}, \tag{2.2}$$

and assume that y^* is finite.

A5: The sequences $\{u_j^{(n)}\}_{j=1}^{\infty}$, $\{v_j^{(n)}\}_{\mathrm{J}=1}^{\infty}$ and $\{L_j^{(n)}\}_{j=1}^{\infty}$ are mutually independent.

A6: Customers are served using the Earliest Deadline First (EDF) queue discipline, i.e., the server always services the customer with the shortest lead-time.

A7: Preemption is permitted (we assume preempt-resume). There is no setup, switchover or any other type of overhead.

A8: Late customers (customers with negative lead times) stay in the queue until served to completion.

A9: The queue is empty at time zero.

The interarrival times, service times, initial lead times and queue discipline completely determine the behavior of the queue. From them we can derive the *customer arrival times*

$$S_k^{(n)} \triangleq \sum_{j=1}^{k} u_j^{(n)},$$

with $S_0^{(n)} \triangleq 0$, and the *customer arrival process*

$$A^{(n)}(t) \triangleq \max\{k | S_k^{(n)} \leq t\}.$$

The *work arrival process*

$$V^{(n)}(t) \triangleq \sum_{j=1}^{\lfloor t \rfloor} v_j^{(n)}$$

records the amount of work which arrives with the first $\lfloor t \rfloor$ customers, and the *netput process*

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

measures the work remaining in queue at time t, provided that the server is never idle up to time t. The *cumulative idleness process*

$$I^{(n)}(t) \triangleq - \min_{0 \leq s \leq t} N^{(n)}(s),$$

gives the amount of time the server is idle, and adding this to the netput process, we obtain the *workload process*

$$W^{(n)}(t) \triangleq N^{(n)}(t) + I^{(n)}(t),$$

which records the amount of work in queue, taking server idleness into account. All the above processes are independent of the queue service discipline, provided that the server is never idle when there are customers in queue. The *queue length process* $Q^{(n)}(t)$, which is the number of customers in queue at time t, depends on the queue discipline. All these processes are *right-continuous with left-hand limits (RCLL).*

In order to obtain heavy traffic limits, we must scale and sometimes center the above processes. The real-valued processes whose limits we shall consider are the following:

$$\widehat{A}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}(A^{(n)}(nt) - \lambda^{(n)}nt), \tag{2.3}$$

$$\widehat{V}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}\sum_{j=1}^{\lfloor nt \rfloor}\left(v_j^{(n)} - \frac{1}{\mu^{(n)}}\right), \tag{2.4}$$

$$\widehat{N}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}(V^{(n)}(A^{(n)}(nt)) - nt), \tag{2.5}$$

$$\widehat{I}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}I^{(n)}(nt), \tag{2.6}$$

$$\widehat{W}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}W^{(n)}(nt) = \widehat{N}^{(n)}(t) + \widehat{I}^{(n)}(t), \tag{2.7}$$

$$\widehat{Q}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}Q^{(n)}(nt). \tag{2.8}$$

Heavy traffic assumptions

Define the traffic intensity $\rho^{(n)} \triangleq \frac{\lambda^{(n)}}{\mu^{(n)}}$. The following assumptions shall be in force throughout:

$$\lim_{n\to\infty} \sqrt{n}(1-\rho^{(n)}) = \gamma > 0, \tag{2.9}$$

$$\lim_{n\to\infty} \lambda^{(n)} = \lambda > 0, \qquad \lim_{n\to\infty} \mu^{(n)} = \lambda, \tag{2.10}$$

$$\lim_{n\to\infty} \alpha^{(n)} = \alpha, \qquad \lim_{n\to\infty} \beta^{(n)} = \beta. \tag{2.11}$$

We also impose the usual Lindeberg condition on the interarrival and service times:

$$\lim_{n\to\infty} I\!E\left[\left(u_j^{(n)} - \frac{1}{\lambda^{(n)}}\right)^2 I\!I_{\left\{\left|u_j^{(n)} - \frac{1}{\lambda^{(n)}}\right| > c\sqrt{n}\right\}}\right] \tag{2.12}$$

$$= \lim_{n\to\infty} I\!E\left[\left(v_j^{(n)} - \frac{1}{\mu^{(n)}}\right)^2 I\!I_{\left\{\left|v_j^{(n)} - \frac{1}{\mu^{(n)}}\right| > c\sqrt{n}\right\}}\right] = 0 \quad \forall c > 0.$$

It is a standard result (see Corollary B.4) that the triple $(\widehat{N}^{(n)}, \widehat{I}^{(n)}, \widehat{W}^{(n)})$ converges weakly to (N^*, I^*, W^*), where N^* is a Brownian motion with drift and

$$\begin{aligned} I^*(t) &\stackrel{\Delta}{=} -\min_{0\le s\le t} N^*(s), \\ W^*(t) &\stackrel{\Delta}{=} N^*(t) + I^*(t). \end{aligned} \tag{2.13}$$

The process W^* is a reflected Brownian motion with drift, and I^* causes the reflection. Furthermore, the scaled queue length process $\widehat{Q}^{(n)}$ converges weakly to λW^* (see Corollary 3.2).

Earliest-Deadline-First (EDF) related processes

With the EDF queue discipline, customers are served in order of increasing lead-times. Any two customers in the queue will maintain their relative order until they depart; however, arriving customers may preempt and move directly into service if they have a sufficiently short initial lead-time. 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 still in the system who have ever been in service. Some care must be taken with these constructs when the queue becomes empty. To do this, we first define:

$$\iota^{(n)}(t) \stackrel{\Delta}{=} t - \sup\{s \in [0,t] : W^{(n)}(s) = 0\},$$

the time elapsed since the last time the queue was empty. We define the *frontier*

$$F^{(n)}(t) \stackrel{\Delta}{=} \left\{ \begin{array}{l} \text{largest lead-time of any customer in the system} \\ \text{who has ever been in service,} \\ \text{or } \sqrt{n}\, y^* - \iota^{(n)}(t), \text{ if this quantity is larger} \\ \text{than the former one, or if the queue is empty.} \end{array} \right\}$$

and the *current lead-time*

$$C^{(n)}(t) \stackrel{\Delta}{=} \left\{ \begin{array}{l} \text{lead-time of the customer in service,} \\ \text{or } \sqrt{n}\, y^* \text{ if the queue is empty} \end{array} \right\}.$$

Under the EDF queue discipline, there is no customer in queue 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 $t \geq 0$. Both $F^{(n)}$ and $C^{(n)}$ are RCLL processes.

At time t all customers in the system have lead-times equal to or greater than $C^{(n)}(t)$; if the queue is non-empty, $C^{(n)}(t)$ is the left support point of the random counting measure which puts a unit point mass at the lead-time of each customer in queue at time t. In spite of this, the frontier is more important than $C^{(n)}(t)$ in the analysis of the EDF queue discipline for two reasons. The first reason is that in heavy traffic the number of customers with lead-times between $C^{(n)}(t)$ and $F^{(n)}(t)$ is negligible. Customers at time t with lead-times in the interval $[C^{(n)}(t), F^{(n)}(t))$, if any, are part of a special type of busy period. For a non-empty queue, this busy period was initiated by a customer arrival that preempted a customer, $\mathcal{C}$, in service (the preempted customer with current lead-time $F^{(n)}(t)$). This busy period was possibly sustained by other arrivals, each of which had, at the time of its arrival, a lead-time shorter than $\mathcal{C}$'s lead-time. Because $\mathcal{C}$'s lead-time decreases linearly with time, the traffic intensity associated with the customers that sustain this special busy period decreases with time. As shown in Proposition 3.5 below, under the heavy traffic scaling the number of customers having lead-times at time t taking values in $[C^{(n)}(t), F^{(n)}(t))$ converges to 0 as $n \to \infty$. It follows that in heavy traffic the occupancy of the queue consists essentially of customers with lead-times in $[F^{(n)}(t), \infty)$.

The second feature of the frontier is that customers with lead times in $[F^{(n)}(t), \infty)$ at time t have never received any service; their lead-time profile is determined entirely by the arrival process. Because only the arrival process is involved, this profile can be determined and in heavy traffic converges to a non-random function of the limit of the scaled queue length.

Although this paper focuses on a heavy traffic analysis of a single server queue using the EDF queue discipline, it is worth noting that some of the results can be expected to carry over in non-heavy-traffic conditions. For example, if the traffic intensity were not near 1, but the queue length happened to be relatively long, then most of the customers in the system would have lead-times taking values in $[F^{(n)}(t), \infty)$, and their profile would be determined solely by the arrival process, not the service process. Consequently, the lead-time profiles would be the same as those predicted for that queue length under heavy traffic conditions.

Measure-valued processes

At any instant of time, the system consists of a set of customers, each of which has a specific lead-time and a remaining work requirement. We wish to characterize the instantaneous lead-time profile of the customers. It is convenient to think of this profile as a counting measure on $\mathbb{R}$. In this section, we define a collection of RCLL measure-valued processes that will be useful in the analysis.

Lead-time measure:

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

Workload measure:

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

Lead-time arrival measure:

$$\mathcal{A}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Number of all arrivals by time } t \text{ having} \\ \text{lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}.$$

Workload arrival measure:

$$\mathcal{V}^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work associated with all arrivals by time} \\ t \text{ having lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}.$$

The following relationships easily follow:

$$Q^{(n)}(t) = \mathcal{Q}^{(n)}(t)(\mathbb{R}),\ A^{(n)}(t) = \mathcal{A}^{(n)}(t)(\mathbb{R}),\ V^{(n)}(A^{(n)}(t)) = \mathcal{V}^{(n)}(t)(\mathbb{R}),$$

$$\begin{aligned}
\mathcal{A}^{(n)}(t)(B) &= \sum_{j=1}^{A^{(n)}(t)} \mathbb{I}_{\{L_j^{(n)} - (t - S_j^{(n)}) \in B\}} \\
&= \sum_{j=1}^{\infty} \mathbb{I}_{\{S_j^{(n)} \in B + t - L_j^{(n)},\ S_j^{(n)} \le t\}}, \\
\mathcal{V}^{(n)}(t)(B) &= \sum_{j=1}^{A^{(n)}(t)} v_j^{(n)} \mathbb{I}_{\{L_j^{(n)} - (t - S_j^{(n)}) \in B\}} \\
&= \sum_{j=1}^{\infty} v_j^{(n)} \mathbb{I}_{\{S_j^{(n)} \in B + t - L_j^{(n)},\ S_j^{(n)} \le t\}}.
\end{aligned}$$

Scaled EDF-Related Processes

For the processes just defined under the EDF queue discipline, we use the following heavy traffic scalings:

$$\widehat{F}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} F^{(n)}(nt), \qquad \widehat{C}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} C^{(n)}(nt),$$
$$\widehat{\mathcal{Q}}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} \mathcal{Q}^{(n)}(nt)(\sqrt{n}B), \qquad \widehat{\mathcal{W}}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} \mathcal{W}^{(n)}(nt)(\sqrt{n}B).$$

We define also

$$\begin{aligned}
\widehat{\mathcal{A}}^{(n)}(t)(B) &\triangleq \frac{1}{\sqrt{n}} \mathcal{A}^{(n)}(nt)(\sqrt{n}B), \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{A^{(n)}(nt)} 1_{\{L_j^{(n)} - (nt - S_j^{(n)}) \in \sqrt{n}B\}} \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} 1_{\{S_j^{(n)} \in \sqrt{n}B + nt - L_j^{(n)},\, S_j^{(n)} \le nt\}}, \\
\widehat{\mathcal{V}}^{(n)}(t)(B) &\triangleq \frac{1}{\sqrt{n}} \mathcal{V}^{(n)}(nt)(\sqrt{n}\, B) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{A^{(n)}(nt)} v_j^{(n)} 1_{\{L_j^{(n)} - (nt - S_j^{(n)}) \in \sqrt{n}B\}} \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_j^{(n)} 1_{\{S_j^{(n)} \in \sqrt{n}B + nt - L_j^{(n)},\, S_j^{(n)} \le nt\}}.
\end{aligned}$$

3 Heavy Traffic Analysis

We set

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

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} which maps $[0, \infty)$ onto $(-\infty, y^*]$. We next define what we shall ultimately show is the *limiting scaled frontier process*

$$F^*(t) \triangleq H^{-1}(W^*(t)), \; t \ge 0, \tag{3.2}$$

where W^* is as in (2.13).

In this section, we prove weak convergence of $\widehat{\mathcal{W}}^{(n)}$ and $\widehat{\mathcal{Q}}^{(n)}$ as measure-valued processes. Weak convergence of measure-valued processes is a special case of weak convergence of metric-space-valued random objects, which is reviewed in Appendix A. We summarize its salient features here.

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 separable. We can define a metric $d_{\mathcal{M}}$ on $\mathcal{M}$ which is consistent with the weak topology on $\mathcal{M}$.

We now define $D_{\mathcal{M}}[0,\infty)$, the space of RCLL measure-valued functions on $[0,\infty)$. An *RCLL measure-valued process* is a random object taking values in $D_{\mathcal{M}}[0,\infty)$, where in $D_{\mathcal{M}}[0,\infty)$ we use the Borel σ-algebra (generated by the open sets under the Skorohod topology). A sequence $\{X_n\}_{n=1}^{\infty}$ of RCLL measure-valued processes *converges weakly* to an RCLL measure-valued process X if the measures induced on $D_{\mathcal{M}}[0,\infty)$ by X_n converge weakly to the measure induced on $D_{\mathcal{M}}[0,\infty)$ by X, i.e., for every bounded, continuous (or equivalently, uniformly continuous) function $F\colon D_{\mathcal{M}}[0,\infty) \to \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbb{E}_n F(X_n) = \mathbb{E}^* F(X).$$

The expectation operators $\mathbb{E}_n$ and $\mathbb{E}^*$ reflect the fact that each X_n may be defined on a probability space with a probability measure depending on the index n, and all these spaces may differ from the space on which X is defined. In the application of this paper, the pre-limit processes are all defined on the same space, and we write $\mathbb{E}$ rather than $\mathbb{E}_n$.

The main result of this section is the following.

Theorem 3.1 *Let $\widehat{\mathcal{W}}^*$ and $\widehat{\mathcal{Q}}^*$ be the measure-valued processes defined by*

$$\widehat{\mathcal{W}}^*(t)(B) = \int_{B\cap[F^*(t),\infty)} (1-G(y))dy, \quad \widehat{\mathcal{Q}}^*(t)(B) = \lambda\widehat{\mathcal{W}}^*(t)(B), \tag{3.3}$$

for all Borel sets $B \subset \mathbb{R}$. Then the measure-valued processes $\widehat{\mathcal{W}}^{(n)}$ and $\widehat{\mathcal{Q}}^{(n)}$ converge weakly to $\widehat{\mathcal{W}}^*$ and $\widehat{\mathcal{Q}}^*$, respectively.

Corollary 3.2 *Under the earliest-deadline-first queue discipline, the scaled queue length processes $\widehat{Q}^{(n)}$ defined by (2.8) converge weakly to λW^*.*

PROOF OF COROLLARY 3.2: We note that

$$\begin{aligned}
\widehat{\mathcal{W}}^*(t)(\mathbb{R}) &= H(F^*(t)) = W^*(t), \\
\widehat{\mathcal{Q}}^*(t)(\mathbb{R}) &= \lambda W^*(t), \\
\widehat{\mathcal{Q}}^{(n)}(t)(\mathbb{R}) &= \widehat{Q}^{(n)}(t).
\end{aligned}$$

The mapping from $\mathcal{M}$ into $I\!R$, which maps each $\mu \in \mathcal{M}$ to its total mass $\mu(I\!R)$, is continuous. By Theorem 3.1 and the Continuous Mapping Theorem A.1,

$$\widehat{\mathcal{Q}}^{(n)}(I\!R) \Rightarrow \widehat{\mathcal{Q}}^{*}(I\!R),$$

or, equivalently,

$$\widehat{Q}^{(n)} \Rightarrow \lambda W^{*}.$$

□

The proof of Theorem 3.1 is given at the end of this section. In order to prove this result, we first examine the convergence of the measure-valued process $\widehat{\mathcal{V}}^{(n)}$ and $\widehat{\mathcal{A}}^{(n)}$. Recall that these processes keep track of arrived work and arrived customers, but not departures.

Proposition 3.3 *Let $-\infty < y < y^*$ and $T > 0$ be given. As $n \to \infty$,*

$$\sup_{0\le t\le T} \left| \widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y) \right| \quad \xrightarrow{P} \quad 0, \tag{3.4}$$

$$\sup_{0\le t\le T} \left| \widehat{\mathcal{A}}^{(n)}(t)(y,\infty) + \lambda H(y+\sqrt{n}\,t) - \lambda H(y) \right| \quad \xrightarrow{P} \quad 0. \tag{3.5}$$

PROOF: For (3.4), let $\varepsilon > 0$ be given and choose a partition $y = \eta_0 < \eta_1 < \cdots < \eta_M = y^*$ such that $|\eta_{m+1} - \eta_m| \le \varepsilon$ for every $m = 0, \ldots, M-1$. Then the following inequalities hold, for each $\overline{m} = 1, \ldots, M$:

$$\begin{aligned} -\varepsilon + \sum_{m=0}^{\overline{m}-1} \Big(1 - G(\eta_m)\Big)(\eta_{m+1} - \eta_m) & \\ \le \quad & \int_y^{\eta_{\overline{m}}} (1 - G(\eta))d\eta \\ \le \quad & \varepsilon + \sum_{m=0}^{\overline{m}-1} \Big(1 - G(\eta_{m+1})\Big)(\eta_{m+1} - \eta_m). \end{aligned} \tag{3.6}$$

To see why this is true, observe that for each $m = 0, \ldots, M-1$ we have

$$\begin{aligned} & \int_{\eta_m}^{\eta_{m+1}} (1 - G(\eta))\,d\eta \\ \le \quad & (1 - G(\eta_m))(\eta_{m+1} - \eta_m) \\ = \quad & (1 - G(\eta_{m+1}))(\eta_{m+1} - \eta_m) + (G(\eta_{m+1}) - G(\eta_m))(\eta_{m+1} - \eta_m) \\ \le \quad & (1 - G(\eta_{m+1}))(\eta_{m+1} - \eta_m) + \varepsilon(G(\eta_{m+1}) - G(\eta_m)). \end{aligned}$$

Summing up the above inequality for $m = 0, \dots, \overline{m} - 1$ and cancelling the "telescoping" terms gives the right inequality in (3.6). The left inequality is obtained in a similar way.

For $0 \le t \le T$, we have

$$
\begin{aligned}
&\widehat{\mathcal{V}}^{(n)}(t)(y, \infty) \\
&\quad = \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_j^{(n)} I\!I_{\{nt+\sqrt{n}\,y - L_j^{(n)} < S_j^{(n)} \le nt\}} \\
&\quad \le \frac{1}{\sqrt{n}} \sum_{m=1}^{M} \sum_{j=1}^{\infty} v_j^{(n)} I\!I_{\{\sqrt{n}\,\eta_{m-1} < L_j^{(n)} \le \sqrt{n}\,\eta_m\}} I\!I_{\{nt - \sqrt{n}(\eta_m - y) < S_j^{(n)} \le nt\}} \\
&\quad = \sum_{m=1}^{M} \widehat{Y}_m^{(n)}(t) + \sum_{m=1}^{M} \widehat{U}_m^{(n)}(t),
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{Y}_m^{(n)}(t) &\triangleq \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} \left[v_j^{(n)} I\!I_{\{\sqrt{n}\,\eta_{m-1} < L_j^{(n)} \le \sqrt{n}\,\eta_m\}} - \frac{1}{\mu^{(n)}} \Big(G(\eta_m) - G(\eta_{m-1}) \Big) \right] \\
&\qquad \cdot I\!I_{\{nt - \sqrt{n}(\eta_m - y) < S_j^{(n)} \le nt\}}, \\
\widehat{U}_m^{(n)}(t) &\triangleq \frac{1}{\mu^{(n)}\sqrt{n}} \sum_{j=1}^{\infty} \Big(G(\eta_m) - G(\eta_{m-1}) \Big) I\!I_{\{nt - \sqrt{n}(\eta_m - y) < S_j^{(n)} \le nt\}}.
\end{aligned}
$$

To see that for each m, $\widehat{Y}_m^{(n)} \Rightarrow 0$ as $n \to \infty$, we define the sequence of nonnegative i.i.d. random variables

$$
\tilde{v}_j{}^{(n)} \triangleq v_j{}^{(n)} I\!I_{\{\sqrt{n}\,\eta_{m-1} < L_j^{(n)} \le \sqrt{n}\,\eta_m\}}, \quad j = 1, 2, \dots
$$

and set

$$
\begin{aligned}
\widetilde{V}_m^{(n)}(t) &\triangleq \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (\tilde{v}_j^{(n)} - I\!E\tilde{v}_j^{(n)}) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} \left[v_j^{(n)} I\!I_{\{\sqrt{n}\,\eta_{m-1} < L_j^{(n)} \le \sqrt{n}\,\eta_m\}} - \frac{1}{\mu^{(n)}} \Big(G(\eta_m) - G(\eta_{m-1}) \Big) \right].
\end{aligned}
$$

We may write

$$
\begin{aligned}
\widehat{Y}_m^{(n)}(t) &= \widetilde{V}_m^{(n)}\Big(\tfrac{1}{n}A^{(n)}(nt)\Big) - \widetilde{V}_m^{(n)}\Big(\tfrac{1}{n}A^{(n)}((nt-\sqrt{n}(\eta_m-y))^+)\Big) \\
&= \widetilde{V}_m^{(n)}\Big(\tfrac{1}{\sqrt{n}}\widehat{A}^{(n)}(t)+\lambda^{(n)}t\Big) \\
&\quad -\widetilde{V}_m^{(n)}\Big(\tfrac{1}{\sqrt{n}}\widehat{A}^{(n)}((t-\tfrac{1}{\sqrt{n}}(\eta_m-y))^+) \\
&\qquad +(\lambda^{(n)}t-\tfrac{1}{\sqrt{n}}\lambda^{(n)}(\eta_m-y))^+\Big).
\end{aligned}
$$

Theorem B.1 (with $v_j^{(n)}$ replaced by $\tilde{v}_j^{(n)}$) implies that $\{\widetilde{V}_m^{(n)}\}_{n=1}^{\infty}$ has a continuous weak limit $\widetilde{V}_m^*$. The Differencing Theorem A.3 and Theorem B.2 imply that $\widehat{Y}_m^{(n)} \Rightarrow 0$ on $[0,T]$, and hence

$$
\sup_{0\le t\le T}\left|\sum_{m=1}^{M}\widehat{Y}_m^{(n)}(t)\right| \xrightarrow{P} 0.
$$

For the analysis of $\widehat{U}_m^{(n)}(t)$, we observe that

$$
\begin{aligned}
\widehat{U}_m^{(n)}(t) &= \frac{1}{\mu^{(n)}\sqrt{n}}\Big(G(\eta_m)-G(\eta_{m-1})\Big)\Big[A^{(n)}(nt) \\
&\qquad -A^{(n)}((nt-\sqrt{n}(\eta_m-y))^+)\Big] \\
&= \frac{1}{\mu^{(n)}}\Big(G(\eta_m)-G(\eta_{m-1})\Big)\Big[\widehat{A}^{(n)}(t)-\widehat{A}^{(n)}((t-\tfrac{1}{\sqrt{n}}\,(\eta_m-y))^+) \\
&\qquad +\lambda^{(n)}\sqrt{n}\,t-\lambda^{(n)}(\sqrt{n}\,t-(\eta_m-y))^+\Big].
\end{aligned}
$$

As $n\to\infty$, the sequence of processes

$$
\{\widehat{A}^{(n)}(t)-\widehat{A}^{(n)}((t-\tfrac{1}{\sqrt{n}}(\eta_m-y))^+);t\ge 0\}_{n=1}^{\infty}
$$

converges weakly to zero. Hence, the weak limit of $\sum_{m=1}^{M}\widehat{U}_m^{(n)}(t)$ is the weak limit of

$$
\sum_{m=1}^{M}(G(\eta_m)-G(\eta_{m-1}))\left[\sqrt{n}\,t-(\sqrt{n}\,t-(\eta_m-y))^+\right]. \tag{3.7}
$$

With n and t fixed, we define

$$
\overline{m}(t)=\max\Big\{m\in\{0,1,\ldots,M\};t\ge\tfrac{1}{\sqrt{n}}(\eta_m-y)\Big\}.
$$

Then (3.7) becomes

$$\begin{aligned}
&\sum_{m=1}^{\overline{m}(t)}(G(\eta_m)-G(\eta_{m-1}))(\eta_m-y)+\sum_{m=\overline{m}(t)+1}^{M}(G(\eta_m)-G(\eta_{m-1}))\sqrt{n}\,t\\
&\leq \sum_{m=1}^{\overline{m}(t)}(1-G(\eta_{m-1}))(\eta_m-y)-\sum_{m=1}^{\overline{m}(t)}(1-G(\eta_m))(\eta_m-y)\\
&\quad +I\!I_{\{\overline{m}(t)\leq M-1\}}(\eta_{\overline{m}(t)+1}-y)(G(\eta_M)-G(\eta_{\overline{m}(t)}))\\
&= \sum_{m=0}^{\overline{m}(t)-1}(1-G(\eta_m))(\eta_{m+1}-y)-\sum_{m=0}^{\overline{m}(t)-1}(1-G(\eta_m))(\eta_m-y)\\
&\quad -(1-G(\eta_{\overline{m}(t)}))(\eta_{\overline{m}(t)}-y)\\
&\quad +I\!I_{\{\overline{m}(t)\leq M-1\}}(\eta_{\overline{m}(t)+1}-y)(1-G(\eta_{\overline{m}(t)})).
\end{aligned}$$

If $\overline{m}(t)=M$, then $1-G(\eta_{\overline{m}(t)})=0$ and we have

$$\sum_{m=0}^{\overline{m}(t)-1}(1-G(\eta_m))(\eta_{m+1}-\eta_m)\leq \varepsilon+\int_y^{y^*}(1-G(\eta))\,d\eta,$$

where we have used (3.6). If $\overline{m}(t)<M$, we have again from (3.6) that

$$\begin{aligned}
\sum_{m=0}^{\overline{m}(t)}(1-G(\eta_m))(\eta_{m+1}-\eta_m) &\leq \varepsilon+\int_y^{\eta_{\overline{m}(t)+1}}(1-G(\eta))\,d\eta\\
&\leq 2\varepsilon+\int_y^{\eta_{\overline{m}(t)}}(1-G(\eta))\,d\eta\\
&\leq 2\varepsilon+\int_y^{y+\sqrt{n}\,t}(1-G(\eta))\,d\eta.
\end{aligned}$$

In the former case, when $\overline{m}(t)=M$, we have $y+\sqrt{n}\,t\geq y^*$. Since $G(\eta)=1$ for $\eta\geq y^*$, in both cases we have on (3.7) the upper bound

$$2\varepsilon+\int_y^{y+\sqrt{n}\,t}(1-G(\eta))\,d\eta=2\varepsilon-H(y+\sqrt{n}\,t)+H(y).$$

We conclude that

$$\sup_{0\leq t\leq T}\left[\sum_{m=1}^{M}\widehat{U}_m^{(n)}(t)+H(y+\sqrt{n}\,t)-H(y)-2\varepsilon\right]^+\xrightarrow{P}0.$$

Since $\varepsilon > 0$ is arbitrary, we have in fact shown

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

To complete the proof of (3.4), we use the lower bound

$$\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) \ge \sum_{m=1}^{M} \check{Y}_m^{(n)}(t) + \sum_{m=1}^{M} \check{U}_m^{(n)}(t),$$

where

$$\begin{aligned}
\check{Y}_m^{(n)}(t) &\triangleq \frac{1}{\sqrt{n}}\sum_{j=1}^{\infty}\left[v_j^{(n)} I\!I_{\{\sqrt{n}\,\eta_{m-1} < L_j^{(n)} \le \sqrt{n}\,\eta_m\}} - \frac{1}{\mu^{(n)}}\Big(G(\eta_m) - G(\eta_{m-1})\Big)\right] \\
&\qquad\qquad\qquad\qquad \cdot I\!I_{\{nt-\sqrt{n}(\eta_{m-1}-y) < S_j^{(n)} \le nt\}}, \\
\check{U}_m^{(n)}(t) &\triangleq \frac{1}{\mu^{(n)}\sqrt{n}}\sum_{j=1}^{\infty}\Big(G(\eta_m) - G(\eta_{m-1})\Big) I\!I_{\{nt-\sqrt{n}(\eta_{m-1}-y) < S_j^{(n)} \le nt\}}.
\end{aligned}$$

By the same argument used to show that $\widehat{Y}_m^{(n)} \Rightarrow 0$, we may show that $\check{Y}_m^{(n)} \Rightarrow 0$. In place of (3.7), we have now

$$\sum_{m=1}^{M}(G(\eta_m) - G(\eta_{m-1}))\left[\sqrt{n}\,t - (\sqrt{n}\,t - (\eta_{m-1} - y))^+\right], \qquad (3.8)$$

and we need to lower bound this quantity. With n and t fixed, we define $\overline{m}(t)$ as before, and (3.8) becomes

$$\begin{aligned}
&\sum_{m=1}^{M\wedge(\overline{m}(t)+1)} (G(\eta_m) - G(\eta_{m-1}))(\eta_{m-1} - y) + \sum_{m=\overline{m}(t)+2}^{M} (G(\eta_m) - G(\eta_{m-1}))\sqrt{n}\,t \\
&\ge \sum_{m=1}^{M\wedge(\overline{m}(t)+1)} (1 - G(\eta_{m-1}))(\eta_{m-1} - y) - \sum_{m=1}^{M\wedge(\overline{m}(t)+1)} (1 - G(\eta_m))(\eta_{m-1} - y) \\
&\qquad + I\!I_{\{\overline{m}(t)\le M-2\}}(\eta_{\overline{m}(t)} - y)(G(\eta_M) - G(\eta_{\overline{m}(t)+1})) \\
&= \sum_{m=1}^{\overline{m}(t)}(1 - G(\eta_m))(\eta_m - y) - \sum_{m=1}^{\overline{m}(t)}(1 - G(\eta_m))(\eta_{m-1} - y) \\
&\qquad - I\!I_{\{\overline{m}(t)\le M-1\}}(1 - G(\eta_{\overline{m}(t)+1}))(\eta_{\overline{m}(t)} - y) \\
&\qquad + I\!I_{\{\overline{m}(t)\le M-2\}}(1 - G(\eta_{\overline{m}(t)+1}))(\eta_{\overline{m}(t)} - y)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{m=1}^{\overline{m}(t)} (1-G(\eta_m))(\eta_m - \eta_{m-1}) - I\!\!I_{\{\overline{m}(t)=M-1\}}(1-G(\eta_M))(\eta_{M-1}-y) \\
&\geq -\varepsilon + \int_y^{\eta_{\overline{m}(t)}} (1-G(\eta))\, d\eta \\
&\geq -2\varepsilon + \int_y^{y+\sqrt{n}\,t} (1-G(\eta))\, d\eta \\
&= -2\varepsilon - H(y+\sqrt{n}\,t) + H(y).
\end{aligned}
$$

It follows that

$$
\sup_{0\leq t\leq T} \left[\sum_{m=1}^{M} \check{U}_m^{(n)}(t) + H(y+\sqrt{n}\,t) - H(y) + 2\varepsilon \right]^{-} \xrightarrow{P} 0.
$$

Since $\varepsilon > 0$ is arbitrary, we have in fact shown

$$
\sup_{0\leq t\leq T} \left[\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y) \right]^{-} \xrightarrow{P} 0.
$$

and (3.4) is proved.

The proof of (3.5) is accomplished by repeating the above proof, replacing $v_j^{(n)}$ and $\frac{1}{\mu^{(n)}} = I\!\!E v_j^{(n)}$ everywhere by 1. $\square$

Using a Glivenko-Cantelli type of argument, we can upgrade Proposition 3.3 to make the convergence uniform with respect to y on compact intervals:

Proposition 3.4 *Let $-\infty < y_0 < y^*$ and $T > 0$ be given. As $n \to \infty$,*

$$
\sup_{y_0\leq y\leq y^*} \sup_{0\leq t\leq T} \left| \widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y) \right| \xrightarrow{P} 0, \tag{3.9}
$$

$$
\sup_{y_0\leq y\leq y^*} \sup_{0\leq t\leq T} \left| \widehat{\mathcal{A}}^{(n)}(t)(y,\infty) + \lambda H(y+\sqrt{n}\,t) - \lambda H(y) \right| \xrightarrow{P} 0. \tag{3.10}
$$

PROOF: Let $\varepsilon > 0$ be given. We will produce an N such that, for all $n \geq N$,

$$
I\!\!P \left\{ \sup_{y_0\leq y\leq y^*} \sup_{0\leq t\leq T} \left| \widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y) \right| \geq \varepsilon \right\} < \varepsilon.
$$

To do this, we first choose a partition $y_0 < y_1 < \cdots < y_M = y^*$, such that $|y_{m+1} - y_m| < \varepsilon/2$ for $m = 0, \ldots, M-1$, and hence

$$
0 \leq H(y_m) - H(y_{m+1}) < \frac{\varepsilon}{2}, \qquad \text{for } m = 0, \ldots, M-1. \tag{3.11}
$$

According to Proposition 3.3, we can find $n_0, n_1, \ldots n_M$ such that, for $m = 0, 1, \ldots, M$ and for $n \geq n_m$,

$$\mathbb{P}\left\{\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y_m,\infty) + H(y_m+\sqrt{n}\,t) - H(y_m)\right| \geq \frac{\varepsilon}{2}\right\} < \frac{\varepsilon}{2(M+1)}. \tag{3.12}$$

We choose $N = \max\{n_0, n_1, \ldots, n_M\}$. Using first the monotonicity of $H(y)$ and of $\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t)$ with respect to y, and then (3.11) and (3.12) we see that for $n \geq N$,

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{y_0\le y\le y^*}\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y)\right| \geq \varepsilon\right\}\\
&\leq \sum_{m=0}^{M-1}\mathbb{P}\left\{\sup_{y_m\le y\le y_{m+1}}\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t) - H(y)\right| \geq \varepsilon\right\}\\
&\leq \sum_{m=0}^{M-1}\mathbb{P}\left\{\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y_m,\infty) + H(y_m+\sqrt{n}\,t) - H(y_{m+1})\right| \geq \varepsilon\right\}\\
&\quad+\sum_{m=0}^{M-1}\mathbb{P}\left\{\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y_{m+1},\infty) + H(y_{m+1}+\sqrt{n}\,t) - H(y_m)\right| \geq \varepsilon\right\}\\
&\leq \sum_{m=0}^{M}2\mathbb{P}\left\{\sup_{0\le t\le T}\left|\widehat{\mathcal{V}}^{(n)}(t)(y_m,\infty) + H(y_m+\sqrt{n}\,t) - H(y_m)\right| \geq \frac{\varepsilon}{2}\right\}\\
&< 2(M+1)\frac{\varepsilon}{2(M+1)} = \varepsilon.
\end{aligned}$$

This proves (3.9); the proof of (3.10) is analogous. □

The heavy traffic analysis of the queueing system with due dates depends critically on the following proposition, which asserts that the number of customers whose lead times lie between the current lead time $C^{(n)}(t)$ and the frontier $F^{(n)}(t)$ and the work associated with these customers are negligible.

Proposition 3.5 *The processes $\widehat{\mathcal{Q}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)})$ and $\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)})$ converge weakly to zero as $n \to \infty$.*

Proof: We fix $T > 0$ and establish the convergence on $[0, T]$. For this we follow ideas of Peterson [26].

Let $y \leq y^*$ be given. For $t \geq 0$, we set

$$\widehat{T}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} (v_j^{(n)} \mathbb{I}_{\{L_j^{(n)} \leq \sqrt{n}\, y\}} - \frac{1}{\mu^{(n)}} G(y)).$$

According to Theorem B.1, $\widehat{T}^{(n)}$ converges weakly to a Brownian motion.

Next define

$$\tau^{(n)}(t) \triangleq \sup\{s \in [0,t]; \widehat{C}^{(n)}(s) = \widehat{F}^{(n)}(s)\}.$$

By assumption, $\widehat{C}^{(n)}(0) = \widehat{F}^{(n)}(0) = y^*$ and so $\tau^{(n)}(t) \leq t$, i.e., the supremum is not over the empty set. We first show that

$$t - \tau^{(n)}(t) \Rightarrow 0 \tag{3.13}$$

and subsequently show that

$$\sqrt{n}(t - \tau^{(n)}(t)) \Rightarrow 0, \tag{3.14}$$

where the convergence in (3.13) and (3.14) is for *processes* on $[0, T]$.

Recalling that $\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)})$ is RCLL, we note that

$$\begin{aligned}
&\widehat{\mathcal{W}}^{(n)}(\tau^{(n)}(t)-)[\widehat{C}^{(n)}(\tau^{(n)}(t)-), \widehat{F}^{(n)}(\tau^{(n)}(t)-)) \\
&\quad = \widehat{\mathcal{W}}^{(n)}(\tau^{(n)}(t)-)(\emptyset) = 0, \qquad (3.15)\\
&\widehat{\mathcal{W}}^{(n)}(\tau^{(n)}(t))[\widehat{C}^{(n)}(\tau^{(n)}(t)), \widehat{F}^{(n)}(\tau^{(n)}(t))) \\
&\quad \leq \frac{1}{\sqrt{n}} \max_{1 \leq j \leq A^{(n)}(nt)} v_j^{(n)} \\
&\quad \leq \max_{0 \leq s \leq T} \left[\widehat{N}^{(n)}(s) - \widehat{N}^{(n)}(s-) \right]. \qquad (3.16)
\end{aligned}$$

So long as there are customers with lead-times in the unscaled interval $[C^{(n)}, F^{(n)})$, the unscaled frontier $F^{(n)}$ decreases at rate 1 per unit time. Therefore, for $s \in (n\tau^{(n)}(t), nt]$,

$$F^{(n)}(s) = F^{(n)}(n\tau^{(n)}(t)) - (s - n\tau^{(n)}(t)). \tag{3.17}$$

For what follows, it will be helpful to introduce some notation. We define

$$\begin{aligned}
&D^{(n)}(t) \qquad (3.18)\\
&= \sum_{j=1}^{\infty} v_j^{(n)} \mathbb{I}_{\{n\tau^{(n)}(t) < S_j^{(n)} \leq nt\}} \mathbb{I}_{\{L_j^{(n)} - (nt - S_j^{(n)}) < F^{(n)}(n\tau^{(n)}(t)) - n(t - \tau^{(n)}(t))\}}
\end{aligned}$$

Observe that because of (3.17), whenever $n\tau^{(n)}(t) < S_j^{(n)} \le nt$, the condition

$$L_j^{(n)} - (nt - S_j^{(n)}) < F^{(n)}(n\tau^{(n)}(t)) - n(t - \tau^{(n)}(t))$$

is equivalent to

$$L_j^{(n)} < F^{(n)}(S_j^{(n)}).$$

In other words, $D^{(n)}(t)$ counts the number of customers arrived within the time interval $(n\tau^{(n)}(t), nt]$ with lead times at arrival to the left of the current frontier.

We now note that, on the time interval $(n\tau^{(n)}(t), nt]$ the server is never idle, which means that the workload is being decreased by the sever at a constant rate 1. This gives us the estimate

$$\begin{aligned} 0 &\le \mathcal{W}^{(n)}(nt)[C^{(n)}(nt), F^{(n)}(nt)) \\ &= \mathcal{W}^{(n)}(n\tau^{(n)}(t))[C^{(n)}(n\tau^{(n)}(t)), F^{(n)}(n\tau^{(n)}(t))) \\ &\quad + D^{(n)}(t) - n(t - \tau^{(n)}(t)), \end{aligned}$$

or, after scaling,

$$\begin{aligned} 0 &\le \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)) \\ &= \widehat{\mathcal{W}}^{(n)}(\tau^{(n)}(t))[\widehat{C}^{(n)}(\tau^{(n)}(t)), \widehat{F}^{(n)}(\tau^{(n)}(t))) \\ &\quad + \frac{1}{\sqrt{n}} D^{(n)}(t) - \sqrt{n}(t - \tau^{(n)}(t)). \end{aligned} \tag{3.19}$$

Next, we estimate the term $\frac{1}{\sqrt{n}} D^{(n)}(t)$. For $y \le y^*$ we have

$$\begin{aligned} \frac{1}{\sqrt{n}} D^{(n)}(t) &\le \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_j{}^{(n)} \mathbb{I}_{\{n\tau^{(n)}(t) < S_j^{(n)} \le n\tau^{(n)}(t) + \sqrt{n}(y^* - y)\}} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_j^{(n)} \mathbb{I}_{\{n\tau^{(n)}(t) + \sqrt{n}(y^* - y) < S_j^{(n)} \le nt\}} \mathbb{I}_{\{L_j^{(n)} \le \sqrt{n}\, y\}} \\ &= \frac{1}{\sqrt{n}} V^{(n)}(A^{(n)}(n\tau^{(n)}(t) + \sqrt{n}(y^* - y))) \\ &\qquad - \frac{1}{\sqrt{n}} V^{(n)}(A^{(n)}(n\tau^{(n)}(t))) \\ &\quad + \widehat{T}^{(n)}\left(\frac{1}{n} A^{(n)}(nt)\right) \\ &\qquad - \widehat{T}^{(n)}\left(\frac{1}{n} A^{(n)}(n\tau^{(n)}(t) + \sqrt{n}(y^* - y))\right) \\ &\quad + \frac{G(y)}{\mu^{(n)}\sqrt{n}} \left[A^{(n)}(nt) - A^{(n)}(n\tau^{(n)}(t) + \sqrt{n}(y^* - y))\right] \end{aligned} \tag{3.20}$$

$$
\begin{aligned}
&= \widehat{N}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)-\widehat{N}^{(n)}\left(\tau^{(n)}(t)\right)+y^*-y \\
&\quad +\widehat{T}^{(n)}\left(\frac{1}{\sqrt{n}}\widehat{A}^{(n)}(t)+\lambda^{(n)}t\right) \\
&\quad\quad -\widehat{T}^{(n)}\left(\frac{1}{\sqrt{n}}\widehat{A}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)\right. \\
&\quad\quad\quad \left. +\lambda^{(n)}\tau^{(n)}(t)+\frac{\lambda^{(n)}}{\sqrt{n}}(y^*-y)\right) \\
&\quad +\frac{G(y)}{\sqrt{n}}\left[\widehat{A}^{(n)}(t)-\widehat{A}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)\right] \\
&\quad +G(y)\rho^{(n)}\sqrt{n}(t-\tau^{(n)}(t))-G(y)\rho^{(n)}(y^*-y) \\
&\leq \left[\widehat{N}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)-\widehat{N}^{(n)}\left(\tau^{(n)}(t)\right)\right] \\
&\quad +\left[\widehat{T}^{(n)}\left(\frac{1}{\sqrt{n}}\widehat{A}^{(n)}(t)+\lambda^{(n)}t\right)\right. \\
&\quad\quad -\widehat{T}^{(n)}\left(\frac{1}{\sqrt{n}}\widehat{A}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)\right. \\
&\quad\quad\quad \left.\left. +\lambda^{(n)}\tau^{(n)}(t)+\frac{\lambda^{(n)}}{\sqrt{n}}(y^*-y)\right)\right] \\
&\quad +\frac{G(y)}{\sqrt{n}}\left[\widehat{A}^{(n)}(t)-\widehat{A}^{(n)}\left(\tau^{(n)}(t)+\frac{1}{\sqrt{n}}(y^*-y)\right)\right] \\
&\quad +G(y)\rho^{(n)}\sqrt{n}(t-\tau^{(n)}(t))-(1-G(y)\rho^{(n)})(y^*-y).
\end{aligned}
$$

Continuing this inequality we may write

$$
\begin{aligned}
\frac{1}{\sqrt{n}}D^{(n)}(t) \quad \leq \quad & \max_{0\leq s\leq T}\left[\widehat{N}^{(n)}\left(s+\frac{1}{\sqrt{n}}(y^*-y)\right)-\widehat{N}^{(n)}(s)\right] \\
& +2\max_{0\leq s\leq T+y^*-y}\left|\widehat{T}^{(n)}\left(\frac{1}{\sqrt{n}}\widehat{A}^{(n)}(s)+\lambda^{(n)}s\right)\right| \\
& +\frac{G(y)}{\sqrt{n}}\max_{0\leq s\leq T}\widehat{A}^{(n)}(s)+G(y)\rho^{(n)}\sqrt{n}(t-\tau^{(n)}(t)) \\
& +(1-G(y)\rho^{(n)})(y^*-y).
\end{aligned}
$$

We now choose and fix a $y < y^*$. Substituting the above inequality into

(3.19), using (3.16), and dividing by $(1 - G(y)\rho^{(n)})\sqrt{n}$, we obtain

$$\begin{aligned}
0 \quad &\leq \quad t - \tau^{(n)}(t) \\
&\leq \quad \frac{1}{(1 - G(y)\rho^{(n)})\sqrt{n}} \left\{ \max_{0 \leq s \leq T} \left[\widehat{N}^{(n)}(s) - \widehat{N}^{(n)}(s-) \right] \right. \\
&\qquad + \max_{0 \leq s \leq T} \left[\widehat{N}^{(n)} \left(s + \frac{1}{\sqrt{n}}(y^* - y) \right) - \widehat{N}^{(n)}(s) \right] \\
&\qquad + 2 \max_{0 \leq s \leq T + y^* - y} \left| \widehat{T}^{(n)} \left(\frac{1}{\sqrt{n}} \widehat{A}^{(n)}(s) + \lambda^{(n)} s \right) \right| \\
&\qquad \left. + \frac{G(y)}{\sqrt{n}} \max_{0 \leq s \leq T} \widehat{A}^{(n)}(s) \right\} \\
&\quad + \frac{1}{\sqrt{n}}(y^* - y).
\end{aligned}$$

This establishes (3.13).

Now let an arbitrary $y < y^*$ be given. We substitute the inequality (3.20) into (3.19), using (3.16), and dividing by $(1 - G(y)\rho^{(n)})$ to get

$$\begin{aligned}
0 \quad &\leq \quad \sqrt{n}(t - \tau^{(n)}(t)) \\
&\leq \quad \frac{1}{(1 - G(y)\rho^{(n)})} \left\{ \max_{0 \leq s \leq T} \left[\widehat{N}^{(n)}(s) - \widehat{N}^{(n)}(s-) \right] \right. \\
&\qquad + \left[\widehat{N}^{(n)} \left(\tau^{(n)}(t) + \frac{1}{\sqrt{n}}(y^* - y) \right) - \widehat{N}^{(n)} \left(\tau^{(n)}(t) \right) \right] \\
&\qquad + \left[\widehat{T}^{(n)} \left(\frac{1}{\sqrt{n}} \widehat{A}^{(n)}(t) + \lambda^{(n)} t \right) \right. \\
&\qquad\quad - \widehat{T}^{(n)} \left(\frac{1}{\sqrt{n}} \widehat{A}^{(n)} \left(\tau^{(n)}(t) + \frac{1}{\sqrt{n}}(y^* - y) \right) \right. \\
&\qquad\qquad \left. \left. + \lambda^{(n)} \tau^{(n)}(t) + \frac{\lambda^{(n)}}{\sqrt{n}}(y^* - y) \right) \right] \\
&\qquad \left. + \frac{G(y)}{\sqrt{n}} \left[\widehat{A}^{(n)}(t) - \widehat{A}^{(n)} \left(\tau^{(n)}(t) + \frac{1}{\sqrt{n}}(y^* - y) \right) \right] \right\} \\
&\quad + (y^* - y).
\end{aligned}$$

As $n \to \infty$, (3.13) and the Time Change and Differencing Theorems A.2 and A.3 imply that the right-hand side has limit $y^* - y$, i.e.,

$$\left[\sqrt{n}(t - \tau^{(n)}(t)) - (y^* - y) \right]^+ \Rightarrow 0.$$

Since $y < y^*$ is arbitrary, we must have (3.14).

Using (3.16), (3.19) and the inequality (3.20) with $y = y^*$, we obtain

$$\begin{aligned}
0 &\leq \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)) \\
&\leq \max_{0 \leq s \leq T} \left[\widehat{N}^{(n)}(s) - \widehat{N}^{(n)}(s-) \right] \\
&\quad + \left[\widehat{T}^{(n)} \left(\frac{1}{\sqrt{n}} \widehat{A}^{(n)}(t) + \lambda^{(n)} t \right) \right. \\
&\quad \left. - \widehat{T}^{(n)} \left(\frac{1}{\sqrt{n}} \widehat{A}^{(n)} \left(\tau^{(n)}(t) \right) + \lambda^{(n)} \tau^{(n)}(t) \right) \right] \\
&\quad + \frac{1}{\sqrt{n}} \left[\widehat{A}^{(n)}(t) - \widehat{A}^{(n)} \left(\tau^{(n)}(t) \right) \right] \\
&\quad - (1 - \rho^{(n)}) \sqrt{n} (t - \tau^{(n)}(t)).
\end{aligned}$$

Once again the Time Change and Differencing Theorems A.2 and A.3 show that the right-hand side has limit zero. This implies

$$\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}) \Rightarrow 0.$$

Similarly,

$$\begin{aligned}
0 &\leq \widehat{\mathcal{Q}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)) \\
&\leq \frac{1}{\sqrt{n}} \left[1 + A^{(n)}(nt) - A^{(n)}(n\tau^{(n)}(t)) \right] \\
&= \frac{1}{\sqrt{n}} + \widehat{A}^{(n)}(t) - \widehat{A}^{(n)}(\tau^{(n)}(t)) + \lambda^{(n)} \sqrt{n} (t - \tau^{(n)}(t)).
\end{aligned}$$

The Time Change and Differencing Theorems A.2 and A.3 and the convergence (3.14) imply that the right-hand side has limit zero. This implies

$$\widehat{\mathcal{Q}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}) \Rightarrow 0.$$

□

Corollary 3.6 *The processes $\widehat{\mathcal{Q}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}]$ and $\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}]$ converge weakly to zero as $n \to \infty$.*

PROOF: In light of Proposition 3.5, it suffices to show that $\widehat{\mathcal{Q}}^{(n)}(t)\{\widehat{F}^{(n)}(t)\}$ and $\widehat{\mathcal{W}}^{(n)}\{\widehat{F}^{(n)}(t)\}$ converge weakly to zero. But $\widehat{\mathcal{Q}}^{(n)}(t)\{\widehat{F}^{(n)}(t)\} \leq \frac{1}{\sqrt{n}}$ and

$$\widehat{\mathcal{W}}^{(n)}\{\widehat{F}^{(n)}(t)\} \leq \frac{1}{\sqrt{n}} \max_{1 \leq j \leq A^{(n)}(nt)} v_j^{(n)}.$$

This latter term has limit zero because the limit of $\widehat{N}^{(n)}$ appearing in (3.16) is continuous. $\square$

We next examine the limit of the scaled frontier process $\widehat{F}^{(n)}$. Since $\sqrt{n}\,y^* - nt \le F^{(n)}(nt) \le \sqrt{n}\,y^*$ at all times, we have the bounds

$$y^* - \sqrt{n}\,t \le \widehat{F}^{(n)}(t) \le y^*, \quad t \ge 0. \tag{3.21}$$

The following lemma provides a tightness bound from below.

Lemma 3.7 *For every $T > 0$ and $\varepsilon > 0$, there exists $y \in (-\infty, y^*)$ such that for all n,*

$$I\!\!P\left\{\inf_{0\le t\le T} \widehat{F}^{(n)}(t) < y\right\} < \varepsilon.$$

PROOF: By definition,

$$W^{(n)}(t) = \mathcal{W}^{(n)}(t)(I\!\!R) = \mathcal{W}^{(n)}(t)[C^{(n)}(t), F^{(n)}(t)] + \mathcal{W}^{(n)}(t)(F^{(n)}(t), \infty).$$

Scaling this equation, we obtain

$$\widehat{W}^{(n)}(t) = \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)] + \widehat{\mathcal{W}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty).$$

Corollary B.4 implies $\widehat{W}^{(n)} \Rightarrow W^*$, where W^* is a reflected Brownian motion with drift, and Corollary 3.6 shows that $\widehat{\mathcal{W}}^{(n)}[\widehat{C}^{(n)}, \widehat{F}^{(n)}] \Rightarrow 0$. Therefore,

$$\widehat{\mathcal{W}}^{(n)}(\widehat{F}^{(n)}, \infty) \Rightarrow W^*. \tag{3.22}$$

At time t, no customer with lead-time in $(F^{(n)}(t), \infty)$ has ever been in service, so $\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty) = \widehat{\mathcal{W}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty)$. Relation (3.22) implies

$$\widehat{\mathcal{V}}^{(n)}(\widehat{F}^{(n)}, \infty) \Rightarrow W^*. \tag{3.23}$$

Fix $T > 0$. The Continuous Mapping Theorem A.1 applied to (3.23) yields

$$\max_{0\le t\le T} \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty) \Rightarrow \max_{0\le t\le T} W^*(t),$$

and so the sequence of random variables $\left\{\max_{0\le t\le T} \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty)\right\}_{n=1}^{\infty}$ is tight. Let $\varepsilon > 0$ be given. Because $\lim_{y\to-\infty} H(y) = \infty$, we may choose $y < y^*$ so that for each n,

$$\max_{0\le t\le T} \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty) \le \sqrt{H(y)} \quad \text{on } A_n,$$

where the event A_n satisfies $\mathbb{P}(A_n) \geq 1 - \frac{\varepsilon}{3}$. Proposition 3.3 and the Continuous Mapping Theorem A.1 imply the existence of N such that for every $n \geq N$,

$$\min_{0\leq t\leq T}\left[\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t)\right] \geq \frac{1}{2}H(y) \quad \text{on } B_n,$$

where the event B_n satisfies $\mathbb{P}(B_n) \geq 1 - \frac{\varepsilon}{3}$.

Now $\mathbb{P}(A_n \cap B_n) \geq 1 - \frac{2\varepsilon}{3}$, and on $A_n \cap B_n$,

$$\begin{aligned}
\sqrt{H(y)} &\geq \max_{0\leq t\leq T} \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t),\infty) \\
&\geq \max_{0\leq t\leq T} \widehat{\mathcal{V}}^{(n)}(t)(y,\infty) I\!I_{\{\widehat{F}^{(n)}(t)<y\}} \\
&\geq \max_{0\leq t\leq T} \left[\widehat{\mathcal{V}}^{(n)}(t)(y,\infty) + H(y+\sqrt{n}\,t)\right] I\!I_{\{\widehat{F}^{(n)}(t)<y\}},
\end{aligned}$$

because $y + \sqrt{n}\,t \geq y^*$ on $\{\widehat{F}^{(n)}(t) < y\}$ (see (3.21)) and $H(y+\sqrt{n}\,t) = 0$. Continuing, we have on $A_n \cap B_n$ that

$$\begin{aligned}
\sqrt{H(y)} &\geq \frac{1}{2}H(y) \max_{0\leq t\leq T} I\!I_{\{\widehat{F}^{(n)}(t)<y\}} \\
&= \frac{1}{2}H(y) I\!I_{\{\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t)<y\}}.
\end{aligned}$$

which implies

$$\begin{aligned}
\frac{2}{\sqrt{H(y)}} &\geq \mathbb{E}\left[I\!I_{\{\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t)<y\}} I\!I_{A_n\cap B_n}\right] \\
&= \mathbb{E}\left[I\!I_{\{\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t)<y\}}\left(1 - I\!I_{(A_n\cap B_n)^c}\right)\right] \\
&\geq \mathbb{P}\left(\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t) < y\right) - \mathbb{P}\left((A_n\cap B_n)^c\right) \\
&\geq \mathbb{P}\left(\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t) < y\right) - \frac{2\varepsilon}{3}.
\end{aligned}$$

In other words,

$$\mathbb{P}\left(\inf_{0\leq t\leq T}\widehat{F}^{(n)}(t) < y\right) \leq \frac{2\varepsilon}{3} + \frac{2}{\sqrt{H(y)}},$$

and by choosing $|y|$ larger if necessary, we may ensure that $\frac{2}{\sqrt{H(y)}} < \frac{\varepsilon}{3}$. $\square$

Proposition 3.8 *Let $-\infty < y_0 < y^*$ and $T > 0$ be given. As $n \to \infty$,*

$$\sup_{y_0 \le y \le y^*} \sup_{0 \le t \le T} \left| \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) - H(\widehat{F}^{(n)}(t) \vee y) \right| \xrightarrow{P} 0, \quad (3.24)$$

$$\sup_{y_0 \le y \le y^*} \sup_{0 \le t \le T} \left| \widehat{\mathcal{A}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) - \lambda H(\widehat{F}^{(n)}(t) \vee y) \right| \xrightarrow{P} 0. \quad (3.25)$$

Proof: Let $T > 0$, $y_0 \in (-\infty, y^*]$, and $\varepsilon > 0$ be given. According to Proposition 3.4, there is an N such that for each $n \ge N$, there is an event B_n with $I\!P(B_n) \ge 1 - \varepsilon$ and on the event B_n, we have for all $y \in [y_0, y^*]$ that

$$\begin{aligned} H(y) - \varepsilon &\le \min_{0 \le t \le T} \left[\widehat{\mathcal{V}}^{(n)}(t)(y, \infty) + H(y + \sqrt{n}\, t) \right] \\ &\le \max_{0 \le t \le T} \left[\widehat{\mathcal{V}}^{(n)}(t)(y, \infty) + H(y + \sqrt{n}\, t) \right] \le H(y) + \varepsilon. \end{aligned}$$

We now choose a partition $y_{-1} < y_0 < y_1 < \cdots < y_K = y^*$ so that $y_k - y_{k-1} < \varepsilon$ for every k (and hence $H(y_{k-1}) - H(y_k) < \varepsilon$). We observe that whenever $y_k \ge \widehat{F}^{(n)}(t)$, we have $H(y_k + \sqrt{n}\, t) = 0$, because $\widehat{F}^{(n)}(t) + \sqrt{n}\, t \ge y^*$. On the event B_n, for $y \in [y_0, y^*]$,

$$\begin{aligned} H(\widehat{F}^{(n)}(t) \vee y) - 2\varepsilon &\le \sum_{k=0}^{K} H(y_{k-1}) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} - 2\varepsilon \\ &\le \sum_{k=0}^{K} H(y_k) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} - \varepsilon \\ &\le \sum_{k=0}^{K} \widehat{\mathcal{V}}^{(n)}(t)(y_k, \infty) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} \\ &\le \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) \\ &\le \sum_{k=0}^{K} \widehat{\mathcal{V}}^{(n)}(t)(y_{k-1}, \infty) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} \\ &\le \sum_{k=0}^{K} H(y_{k-1}) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} + \varepsilon \\ &\le \sum_{k=0}^{K} H(y_k) I\!I_{\{y_{k-1} < \widehat{F}^{(n)}(t) \vee y \le y_k\}} + 2\varepsilon \\ &\le H(\widehat{F}^{(n)}(t) \vee y) + 2\varepsilon. \end{aligned}$$

In other words, on the event B_n,

$$\begin{aligned}
&\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) - 2\varepsilon \leq H(\widehat{F}^{(n)}(t) \vee y) \\
&\leq \widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) + 2\varepsilon, \; 0 \leq t \leq T, \; y_0 \leq y \leq y^*.
\end{aligned}$$

The assertion (3.24) readily follows. The proof of (3.25) is analogous. □

Proposition 3.9 $\widehat{F}^{(n)} \Rightarrow F^* \triangleq H^{-1}(W^*)$.

PROOF: Let $T > 0$ and $\varepsilon > 0$ be given. Using Lemma 3.7, we may choose $y_0 < y^*$ so that $A_n \triangleq \left\{\inf_{0 \leq t \leq T} \widehat{F}^{(n)}(t) > y_0\right\}$ satisfies $I\!P(A_n) \geq 1 - \varepsilon$ for every n. According to Proposition 3.8, there is an N such that for each $n \geq N$, there is an event B_n with $I\!P(B_n) \geq 1 - \varepsilon$ and on the event B_n, we have

$$\sup_{y_0 \leq y \leq y^*} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y, \infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| < \varepsilon.$$

On the intersection $A_n \cap B_n$, we have in particular

$$\begin{aligned}
&\sup_{0 \leq t \leq T} \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t), \infty) - H(\widehat{F}^{(n)}(t))\right| \\
&= \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y_0, \infty) - H(\widehat{F}^{(n)}(t) \vee y_0)\right| < \varepsilon
\end{aligned}$$

It follows that

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

Relation (3.23) shows that

$$H(\widehat{F}^{(n)}) \Rightarrow W^*. \tag{3.26}$$

Applying the continuous function H^{-1} to both sides of (3.26), we obtain the desired result $\widehat{F}^{(n)} \Rightarrow H^{-1}(W^*)$ from the Continuous Mapping Theorem A.1. □

Proposition 3.10 *Let $T > 0$ be given. As $n \to \infty$,*

$$\sup_{y \in I\!R} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{W}}^{(n)}(t)(y, \infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| \xrightarrow{P} 0, \tag{3.27}$$

$$\sup_{y \in I\!R} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{Q}}^{(n)}(t)(y, \infty) - \lambda H(\widehat{F}^{(n)}(t) \vee y)\right| \xrightarrow{P} 0. \tag{3.28}$$

PROOF: For $y \geq y^*$,

$$\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) = H(\widehat{F}^{(n)}(t) \vee y) = 0.$$

For $y < y^*, 0 \leq t \leq T$,

$$\begin{aligned}
&\left|\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| \qquad (3.29)\\
&\quad \leq \left|\widehat{\mathcal{W}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| + \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)]\\
&\quad = \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| + \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)].
\end{aligned}$$

Let $T > 0$ and $\varepsilon > 0$ be given. Using Lemma 3.7, we may choose $y_0 < y^*$ so that $A_n \triangleq \left\{\inf_{0 \leq t \leq T} \widehat{F}^{(n)}(t) > y_0\right\}$ satisfies $\mathbb{P}(A_n) \geq 1 - \varepsilon$ for every n. According to Proposition 3.8, there is an N_1 such that for each $n \geq N_1$, there is an event B_n with $\mathbb{P}(B_n) \geq 1 - \varepsilon$ and on the event B_n, we have

$$\sup_{y_0 \leq y \leq y^*} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right| < \varepsilon.$$

According to Corollary 3.6, there is an N_2 such that for each $n \geq N$, there is an event D_n with $\mathbb{P}(D_n) \geq 1 - \varepsilon$ and on the event D_n, we have

$$\sup_{0 \leq t \leq T} \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)] < \varepsilon.$$

Using (3.29), we see that on the intersection $A_n \cap B_n \cap D_n$, we have for $n \geq N_1 \vee N_2$,

$$\begin{aligned}
&\sup_{y \in \mathbb{R}} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right|\\
&\quad \leq \sup_{y_0 \leq y \leq y^*} \sup_{0 \leq t \leq T} \left\{\left|\widehat{\mathcal{W}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right|\right.\\
&\qquad \left. + \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)]\right\}\\
&\quad \leq \sup_{y_0 \leq y \leq y^*} \sup_{0 \leq t \leq T} \left|\widehat{\mathcal{V}}^{(n)}(t)(\widehat{F}^{(n)}(t) \vee y,\infty) - H(\widehat{F}^{(n)}(t) \vee y)\right|\\
&\qquad + \sup_{0 \leq t \leq T} \widehat{\mathcal{W}}^{(n)}(t)[\widehat{C}^{(n)}(t), \widehat{F}^{(n)}(t)]\\
&\quad < 2\varepsilon
\end{aligned}$$

This implies (3.27). The proof of (3.28) is analogous. $\square$

We are now prepared to prove Theorem 3.1.

Proof of Theorem 3.1: We define a mapping $\psi : I\!R \to \mathcal{M}$ by the formula

$$\psi(x)(B) \triangleq \int_{B\cap[x,\infty)} (1 - G(\eta))\, d\eta, \quad \text{for } x \in I\!R, B \in \mathcal{B}(I\!R).$$

Observe that, for $x_1, x_2 \in I\!R$,

$$\sup_{B\in\mathcal{B}(I\!R)} |\psi(x_1)(B) - \psi(x_2)(B)| \leq \int_{x_1\wedge x_2}^{x_1\vee x_2} (1 - G(\eta))\, d\eta \leq |x_2 - x_1|,$$

which shows that the mapping ψ is continuous. According to Proposition 3.9,

$$\widehat{F}^{(n)} \Rightarrow F^*.$$

By the Continuous Mapping Theorem A.1,

$$\psi(\widehat{F}^{(n)}) \Rightarrow \psi(F^*) = \widehat{\mathcal{W}}^*. \tag{3.30}$$

On the other hand, according to Proposition 3.10,

$$\sup_{y\in I\!R} \sup_{0\leq t\leq T} \left| \widehat{\mathcal{W}}^{(n)}(t)(y,\infty) - \psi(\widehat{F}^{(n)}(t))(y,\infty) \right| \xrightarrow{P} 0 \tag{3.31}$$

(this is a rewriting of (3.27)). Combining (3.30) and (3.31), we see that $\widehat{\mathcal{W}}^{(n)} \Rightarrow \widehat{\mathcal{W}}^*$. The proof of $\widehat{\mathcal{Q}}^{(n)} \Rightarrow \widehat{\mathcal{Q}}^*$ is analogous. □

4 Simulation Results

In this section, we use simulation to verify the predictive value of the theory of the previous sections. In the previous sections, we actually considered a sequence of queueing systems, indexed by n, whereas here we want to consider a single queueing system. We imagine that this single system is a member of the sequence of the previous sections corresponding to a large value of n. We first recast the definitions of the previous sections in such a way that this parameter n does not appear.

Suppressing the time variable t, we recall the definitions of Section 2. We denoted the queue-length in the n-th system by $Q^{(n)}$ and the scaled queue-length by $\widehat{Q}^{(n)} = \frac{1}{\sqrt{n}} Q^{(n)}$, which, for large values of n, is approximately equal to $Q^* = \lambda W^*$ (Corollary 3.2). The workload and scaled workload, respectively, are $W^{(n)}$ and $\widehat{W}^{(n)} = \frac{1}{\sqrt{n}} W^{(n)}$. The "frontier" (see Section 2 for the definition) is $F^{(n)}$, and the scaled frontier is $\widehat{F}^{(n)} = \frac{1}{\sqrt{n}} F^{(n)}$. Finally, there are the measure-valued processes $\mathcal{Q}^{(n)}$ and $\widehat{\mathcal{Q}}^{(n)}$. We shall be interested particularly in $\mathcal{Q}^{(n)}(x,\infty)$, which tells us the number of customers whose lead-times exceed x, and in $\widehat{\mathcal{Q}}^{(n)}(y,\infty) = \frac{1}{\sqrt{n}}\mathcal{Q}^{(n)}(\sqrt{n}y,\infty)$, where we continue to suppress the time-variable t.

Recall that customers arrive with lead-time distribution given by (2.1):

$$\mathbb{P}(L_j^{(n)} \le \sqrt{n}\, y) = G(y).$$

We define $G_n(x) = G(\frac{x}{\sqrt{n}})$, so that

$$\mathbb{P}(L_j^{(n)} \le x) = G_n(x) \tag{4.1}$$

is the cumulative distribution function of the lead-times in the n-th queueing system. The limit of $\widehat{\mathcal{Q}}^{(n)}$ is characterized in terms of the function H of (3.1):

$$H(y) \triangleq \int_y^\infty (1 - G(\eta))\, d\eta.$$

In this section, we will need the function

$$H_n(x) = \sqrt{n}\, H\left(\frac{x}{\sqrt{n}}\right) = \int_x^\infty (1 - G_n(\xi))\, d\xi, \tag{4.2}$$

whose inverse is $H_n^{-1}(y) = \sqrt{n}H^{-1}(\frac{y}{\sqrt{n}})$.

According to Theorem 3.1, for large values of n,

$$\widehat{\mathcal{Q}}(y,\infty) \approx \lambda H(y \vee F^*). \tag{4.3}$$

Moreover, $F^* = H^{-1}(W^*) = H^{-1}\left(\frac{Q^*}{\lambda}\right)$. Multiplying (4.3) by $\sqrt{n}$ and replacing y by $\frac{x}{\sqrt{n}}$, we obtain

$$\mathcal{Q}^{(n)}(x,\infty) \approx \lambda H_n(x \vee \sqrt{n}\, F^*). \tag{4.4}$$

Because $H_n(\sqrt{n}\, F^*) = \sqrt{n}\, H(F^*) = \frac{\sqrt{n}}{\lambda}Q^* \approx \frac{1}{\lambda}Q^{(n)}$, we also obtain

$$\sqrt{n}F^* \approx H_n^{-1}\left(\frac{1}{\lambda}Q^{(n)}\right). \tag{4.5}$$

We define

$$F_n \triangleq H_n^{-1}\left(\frac{1}{\lambda}Q^{(n)}\right), \tag{4.6}$$

so that (4.5) becomes $\sqrt{n}F^* \approx F_n$ and (4.4) becomes

$$\mathcal{Q}^{(n)}(x,\infty) \approx \lambda H_n\left(x \vee F_n^*\right), \ x \geq 0. \tag{4.7}$$

Note that F_n is not the frontier $F^{(n)}$ defined in Section 2. However,

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(F_n - F^{(n)}\right) &= H^{-1}\left(\frac{1}{\lambda}\widehat{Q}^{(n)}\right) - \widehat{F}^{(n)} \\ \Rightarrow \ H^{-1}(W^*) - F^* &= 0.\end{aligned}$$

Relations (4.6), (4.7) connect the unscaled queue length $Q^{(n)}$ with the number of customers whose unscaled lead-times exceed x, and the function H_n appearing in these relations can be computed from the cumulative distribution function G_n of the unscaled lead-time distribution. These relations can be verified by simulation without knowledge of the parameter n.

The function of x appearing on the right-hand side of (4.7) is nonincreasing, with limit $Q^{(n)}$ at $-\infty$ and limit zero at ∞. Therefore,

$$F_{thy}(x) = 1 - \frac{\lambda}{Q^{(n)}}H_n(x \vee F_n), \ x \geq 0, \tag{4.8}$$

is a cumulative distribution function. According to (4.7), $F_{thy}(x)$ should approximate the fraction of customers in queue whose lead-times are less than or equal to x. Since the parameter n is irrelevant, we henceforth omit it in our discussion of (4.8).

We present simulation results illustrating the accuracy this approximation. In the various experiments, we simulate an M/M/1 queue[4] using the EDF queue discipline, usually with $\lambda = .95$ or $.99$, $\mu = 1.0, \rho = .95$ or $.99$. According to the theory developed in Section 3, if one were to randomly stop the simulation at any point, observe the current number in the queue, Q, and find that Q is sufficiently large, then the corresponding instantaneous lead-time profile, expressed as an empirical cdf, should be given approximately by (4.8).

For real-time queueing theory to be useful in practice, it is important that it can be applied in cases in which the queue length Q is moderate in size. However, when Q is moderate, one would expect the lead-time profiles to exhibit substantial variability, and it is not at all clear that the asymptotic form given by (4.8) would be appropriate. The simulations presented in this section are designed to address this issue.

[4]The theory applies to GI/G/1 queues, but only M/M/1 systems are simulated in this section. Limited experience with the simulation of GI/G/1 queues suggests that the accuracy of the approximations in this section are representative of more general systems.

For each simulation run, a particular deadline distribution, $G = G_n$, and queue length, $Q = Q^{(n)}$, is chosen. The run is initiated with an empty queue and continues until the instant the local time at level Q reaches a prespecified value, 10 for the results presented in this paper. At that instant, the lead-time profile is recorded. This same experiment is repeated a total of N times, hence the N profiles can be thought of as independent random objects. We wish to assess how close they are to their predicted form given by (4.8).

The deadline distribution G used in Figures 1–5 is a Uniform(30,70) distribution. Figure 1 illustrates the variability in the lead-time profiles for small to moderate values of Q. This figure shows the first 4 lead-time profiles recorded when $Q = 20$ (then when $Q = 60$) and the accumulated local time is 10. The lead-time profiles are actually 20 (or 60) dimensional vectors, but are plotted here as line segments for visual convenience. Notice that while the profiles have similar shapes, they exhibit substantial variability.

4.1 Uniform Deadlines

The first deadline distribution considered was the Uniform(A,B). For this distribution, the frontier $F = F_n$ is given by

$$F = \begin{cases} B - \sqrt{2W(B-A)} & \text{if } W \le \frac{B-A}{2}, \\ \frac{B+A}{2} - W & \text{if } W > \frac{B-A}{2}, \end{cases}$$

where throughout we use the notation $W = \frac{Q}{\lambda}$. The theoretical cdf defined by (4.8) for this Uniform(A,B) deadline case is given by one of two forms depending on the magnitude of W. If $W \ge \frac{B-A}{2}$, then

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \le F, \\ 1 - \frac{1}{W}\left(\frac{A+B}{2} - x\right) & \text{if } F < x < A, \\ 1 - \frac{(B-x)^2}{2W(B-A)} & \text{if } A \le x \le B, \end{cases}$$

whereas if $W < \frac{B-A}{2}$, then

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \le F, \\ 1 - \left(\frac{B-x}{B-F}\right)^2 & \text{if } F < x \le B, \\ 1 & \text{if } B \le x. \end{cases}$$

If we substitute the simulated customer lead-times into F_{thy}, the resulting empirical cdf should correspond to a Uniform(0,1). In addition, from

F_{thy} the quantiles of the lead-time distribution can be determined by solving $F_{thy}^{-1}(p) = x_p$ for $0 < p < 1$ For $W \geq \frac{B-A}{2}$ this gives

$$x_p = \begin{cases} F + pW & \text{if } 0 < p < 1 - \frac{B-A}{2W}, \\ B - \sqrt{2(1-p)W(B-A)} & \text{if } 1 - \frac{B-A}{2W} < p < 1. \end{cases}$$

whereas for $W < \frac{B-A}{2}$ we have

$$x_p = B - (B-F)\sqrt{1-p}, \quad 0 < p < 1.$$

The theoretical profiles for any particular Q and G in all the following figures are obtained by connecting the points $\left\{(x_p, p),\ p = \frac{1}{2Q+1}, \ldots, \frac{2Q}{2Q+1}\right\}$.

Interestingly, in spite of the substantial variability in each profile, if one averages those profiles (by averaging each of the components of the N distinct Q-dimensional vectors), the result is a very smooth profile which is nearly identical to the theoretical cdf given by (4.8). Figures 2–5 show the mean profile, the component-wise minimum profile and the component-wise maximum profile for the cases $Q = 15, 20, 40$ and 60, each component of which is denoted by an “x”. The component-wise minimum and maximum form an envelop for all N profiles generated by the particular simulation run. In Figures 2 and 3, 95% confidence intervals for the mean are determined for each of the Q quantiles. These confidence limits are denoted by “*”, and have been constructed independently for each of the Q components. In the subsequent figures, the confidence interval indications are omitted as they become visually distracting. In all cases, the traffic intensity is .95, and the profile is recorded when exactly 10 units of local time at the indicated queue length have elapsed. The global time at which the last lead-time profile is taken is recorded at the top of each figure.

The profile is approximately correct for $Q = 15$, but there are systematic departures evident. Figures 3–5 suggest that the profile form given by (4.8) is nearly exact as a mean value for $Q \geq 20$, as the theoretical curve is always within the 95% confidence limits.

In considering the behavior of a real-time queueing system, it is important to put bounds or confidence sets around the profiles described by (4.8) which capture a large fraction of profiles. Such bounds could be used to determine access control policies which would prevent customer lateness (at the expense of losing customers through admission denial). The minimum and maximum curves offer some idea of how wide such profiles must be and how wide the confidence regions must be. Presumably, the bounds can be constructed using large deviation theory; however, this is not studied any further in this paper. While the minimum and maximum limits seem fairly wide, we expect that nearly all of the empirical profiles will be within an $O(1)$ distance from the theoretical profile in which lead-times are $O(\sqrt{n})$.

These plots are representative of many such plots for a variety of bounded deadline distributions and traffic intensities. The greater the variability in G, the larger Q must be for the average profiles to agree with (4.8). The accumulated local time at which the profiles are recorded also has important consequences. These issues are both addressed in the next section.

5 Future Research

In this section, we introduce additional simulation results which illustrate potential generalizations of the real-time queueing theory developed in this paper and some additional issues.

5.1 More General Deadline Distributions

The theory developed in Sections 3 and 4 used the assumption that the deadline distribution was bounded above by some finite constant. Here, we illustrate that this assumption appears to be unnecessary. Two distributions which are not bounded above are considered: the exponential (α) distribution with mean $\frac{1}{\alpha}$ and the Pareto(α, B) distribution.

We begin with the exponential(α) distribution with mean $\frac{1}{\alpha}$. All moments of this distribution are finite, but it is not bounded above as in the uniform case. For this distribution, the frontier is given by:

$$F = \begin{cases} -\frac{1}{\alpha}\log(\alpha W) & \text{if } W \leq \frac{1}{\alpha}, \\ \frac{1}{\alpha} - W & \text{if } W > \frac{1}{\alpha}. \end{cases}$$

The theoretical cdf for the lead-time profiles for this exponential(α) case takes on one of two forms depending upon whether $W \geq \frac{1}{\alpha}$ or not. For the case in which $W \geq \frac{1}{\alpha}$, we have

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \leq F, \\ 1 - \frac{1}{\alpha W}(1 - \alpha x) & \text{if } F < x < 0, \\ 1 - \frac{1}{\alpha W}\exp(-\alpha x) & \text{if } x \geq 0, \end{cases}$$

whereas for $W < \frac{1}{\alpha}$ we have

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \leq F, \\ 1 - \frac{1}{\alpha W}\exp(-\alpha x) & \text{if } x > F. \end{cases}$$

For $W \geq \frac{1}{\alpha}$, the quantiles are given by

$$x_p = \begin{cases} \frac{1}{\alpha} - W(1-p) & \text{if } 0 < p < 1 - \frac{1}{\alpha W}, \\ -\frac{1}{\alpha}\log(\alpha W(1-p)) & \text{if } 1 - \frac{1}{\alpha W} \leq p < 1, \end{cases}$$

whereas for $W < \frac{1}{\alpha}$ they are given by

$$x_p = -\frac{1}{\alpha}\log((1-p)\alpha W), \;\; 0 < p < 1.$$

For the simulation, we chose $\alpha = .02$, giving a mean of 50. We again simulate 50 independent profiles taken when 10 units of local time have been reached. Figure 6 shows the mean profile for $Q = 20$. Again, the shape is generally correct, but systematic departures are evident. Because this deadline distribution will result in a few customers in the queue having very large lead-times, it is more informative to use Q-Q plots to judge the agreement between the empirical and the theoretical cdf. A Q-Q plot is obtained by plotting $\left\{\left(F_{thy}(L_i), \frac{i}{Q+1}\right), 1 \leq i \leq Q\right\}$. If the lead-times $(L_1, \ldots, L_Q)$ are a random sample from F_{thy}, then these points should lie close to a 45-degree line connecting $(0,0)$ with $(1,1)$. Figure 7 gives the Q-Q plot corresponding to Figure 6. Figure 8 presents the Q-Q plot for the same exponential deadline case when $Q = 40$. When $Q = 20$, systematic departures between the average empirical cdf and the theoretical cdf are evident, especially in the left hand tail. Nevertheless, when $Q = 40$, the agreement is nearly exact.

We next consider a Pareto(α,B) deadline distribution. This distribution is characterized by the cdf

$$G(x) = \begin{cases} 0 & \text{if } \frac{x}{B} < 1, \\ 1 - \left(\frac{B}{x}\right)^{(\alpha-1)} & \text{if } \frac{x}{B} \geq 1, \end{cases}$$

for $B > 0$ and $\alpha > 1$. This distribution has no moments of order $\alpha - 1$ or higher, hence it has a very heavy right-hand tail. Indeed, for $1 < \alpha \leq 2$, the function $H(y) = \infty$ for all finite y, and the proposed lead-time profile given by (4.8) does not exist. Nevertheless, we show simulation results for $\alpha = 3$ and $\alpha = 6$ which demonstrate that there is a stable lead-time profile associated with this family of distributions for $\alpha > 2$. For $\alpha > 2$, the frontier is given by

$$F = \begin{cases} B\left(\frac{B}{(\alpha-2)W}\right)^{\frac{1}{\alpha-2}} & \text{if } W \leq \frac{B}{\alpha-2}, \\ \frac{\alpha-1}{\alpha-2}B - W & \text{if } W > \frac{B}{\alpha-2}. \end{cases}$$

The theoretical cdf for this Pareto(α,B) case is given by one of two forms depending upon whether $W \geq \frac{B}{\alpha-2}$ or not. For the case in which $W \geq \frac{B}{\alpha-2}$, we have

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \leq F, \\ 1 - \frac{1}{W}\left(\frac{\alpha-1}{\alpha-2}B - x\right) & \text{if } F \leq x < B, \\ 1 - \frac{B}{W(\alpha-2)}\left(\frac{B}{x}\right)^{\alpha-2} & \text{if } x \geq B, \end{cases}$$

whereas for $W < \frac{B}{\alpha-2}$ we have

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x \leq F, \\ 1 - \frac{B}{W(\alpha-2)}\left(\frac{B}{x}\right)^{\alpha-2} & \text{if } x \geq F. \end{cases}$$

For $W \geq \frac{B}{\alpha-2}$ the quantiles are given by

$$x_p = \begin{cases} \frac{\alpha-1}{\alpha-2}B - (1-p)W & \text{if } 0 < p < 1 - \frac{B}{(\alpha-2)W}, \\ B\left(\frac{B}{(\alpha-2)W(1-p)}\right)^{\frac{1}{\alpha-2}} & \text{if } 1 - \frac{B}{(\alpha-2)W} < p < 1, \end{cases}$$

whereas for $W < \frac{B}{\alpha-2}$ they are given by

$$x_p = B\left(\frac{B}{(\alpha-2)W(1-p)}\right)^{\frac{1}{\alpha-2}}, \quad 0 < p < 1.$$

Figures 9 and 10 give Q-Q plots for the Pareto(6,40) distribution, a distribution with mean 50 and a relatively heavy tail. The traffic intensity was increased to .99, and the number of profiles was increased from 50 to 100. Figure 9 corresponds to $Q = 20$, while Figure 10 corresponds to $Q = 40$. The variability in the Q-Q plots is quite large; however, there is good agreement between the average profile and the theoretical distributions. For $Q = 20$, there are some systematic departures in the tails, but for $Q = 40$, the agreement is very good except in the upper tail.

Figures 11 and 12 present results for the more extreme Pareto(3,25), a distribution with mean 50, but infinite variance. This distribution has an extremely heavy right-hand tail, and very long lead-times will occur as Q increases. For this case, we considered longer queue lengths, with Figure 11 corresponding to $Q = 40$ and Figure 12 to $Q = 60$. Figure 11 shows systematic departures in the left-hand tail, while Figure 12 shows near exact agreement between the average profile and the theoretical distribution.

5.2 FIFO Queue Discipline

This paper studies the EDF queue discipline; however, the results can be used heuristically to determine lead-time profiles for the behavior of first-in-first-out (FIFO) queues. Suppose that arriving customers have deadlines given by distribution G, but are serviced in FIFO order. This queue discipline does not require knowledge of the customer deadlines, and a customer's instantaneous lead-time is equal to its initial deadline minus its time in queue. The time in queue can be determined using (4.8) first by assuming all customers have deadlines 0. If all customers have deadline 0, then the EDF queue discipline is equivalent to FIFO, and any customer's instantaneous lead-time is equal to the negative of its time in queue. Profiles of customers' times in the queue can be approximated by using (4.8) with deadline distribution corresponding to point mass at 0. In this case, the resulting distribution will be Uniform$(-W, 0)$. By adding back their actual deadlines to their time in queue, one can recover the customer lead-times. Consequently, if one were to order the customers in a FIFO queue by lead-time (the FIFO ordering is by time in queue), then the resulting lead-time profile should be the convolution of G with a Uniform$(-W, 0)$ distribution.

Figures 13 and 15 illustrate the lead-time profiles for a FIFO queue assuming G is exponential$(\frac{1}{50})$, $Q = 20$ and $Q = 40$. Figures 14 and 16 present the corresponding Q-Q plots for more accurate assessment of agreement. In this case, the frontier is given by $F = -W = -Q/\lambda$. The theoretical cdf is given by

$$F_{thy}(x) = \begin{cases} 0 & \text{if } x < -W, \\ 1 + \frac{x}{W} + \frac{1}{\alpha W}(e^{-\alpha(x+W)} - 1) & \text{if } -W < x < 0, \\ 1 - \frac{1-e^{-\alpha W}}{\alpha W} e^{-\alpha x} & \text{if } 0 \leq x. \end{cases}$$

There are two different cases associated with finding the quantiles for this distribution. First, when $0 < p < 1 - \frac{1-e^{-\alpha W}}{\alpha W}$, then $x_p = \frac{y}{\alpha} - W$ where y is the solution of the equation

$$y + e^{-y} = 1 + \alpha W p.$$

For $1 - \frac{1-e^{-\alpha W}}{\alpha W} \leq p < 1$,

$$x_p = -\frac{1}{\alpha} \log \left(\frac{(1-p)\alpha W}{1 - e^{-\alpha W}} \right).$$

Again, the agreement is reasonable for $Q = 20$ and excellent for $Q = 40$.

5.3 Profiles at Hitting Times

Real-time queueing theory should be useful for developing and analyzing control policies to reduce or eliminate customer lateness. One simple policy would be a threshold policy in which arrivals would be denied admission and be lost if Q reached some specified level. In principle, the threshold could be chosen based on the predicted lead-time profiles, for example, by choosing the threshold so that the frontier is bounded away from 0 by some confidence margin. This analysis would use the profile when Q first hit the threshold. Interestingly, the profile obtained at the time at which a level is first hit can be systematically different from the profile predicted by (4.8). Figure 17 gives a representative example. The simulation parameters are identical to that presented in Figure 5 except that the profile is recorded at the instant that $Q = 60$ rather than after 10 units of local time have been accumulated at level 60. The mean empirical cdf curve is shifted away from the theoretical profile. The introduction of the hitting time which guarantees that the queue length has never exceeded the level in question creates systematic distortions in the profiles. It will be important to develop corrections to (4.8) which incorporate this stopping time bias.

5.4 The Non-Heavy Traffic Case

The theory presented in Section 3 is developed assuming the traffic intensity approaches one as $n \to \infty$ (see (2.9)). Interestingly, suppose one does not assume this heavy traffic condition, but simulates an EDF system with moderate traffic intensity. If the simulation is continued until a suitable amount of local time (say 10 units) at a large enough queue level (say $Q = 40$) is obtained and the lead-time profile is recorded, then that profile will be essentially indistinguishable from those recorded under heavy traffic conditions. Hence, it appears that for the EDF queue discipline, heavy traffic theory calculations can be used to make accurate lead-time profile predictions under moderate traffic conditions. Of course, if the traffic intensity is moderate, long queue lengths are relatively infrequent compared with heavy traffic conditions. Much more research is needed to assess the accuracy of the conjectures presented in this section. Nevertheless, it appears that the heavy traffic approximations developed for real-time queues will have many important applications under non-extreme traffic conditions.

6 Acknowledgment

We wish to acknowledge an initial discussion in 1995 with Frank Kelly of Cambridge University and David Aldous of University of California, Berkeley. Frank Kelly first produced the formula (4.8), and David Aldous provided an equilibrium analysis proof in an unpublished manuscript. Also, in 1997, Martin Reiman of Lucent Technologies provided insights into how to use heavy traffic methodology (as opposed to Markov methods) to analyze real-time queues that led directly to this paper. Finally, we acknowledge an anonymous referee who pointed out an error in an earlier version of this paper and suggested a simplification of Theorem 3.1.

This research was supported in part by the Office of Naval Research under contract N00014-92-J-1524, by DARPA under contracts F30602-96-1-0160 and N66001-97-C-8527 and by the Army Research Office under contract DAAH04-95-1-0226.

A Weak Convergence

The following standard results can be found in or are easily derived from assertions found in Parthasarathy [25], Chapter II and Billingsley [1], Section 17. In this appendix, we state version of these results needed for this paper.

Let S be a separable metric space, and let $\mathcal{M}(S)$ be the set of finite measures defined on the σ-algebra of Borel subsets of S. We endow $\mathcal{M}(S)$ with the *weak topology*, whereby a sequence of finite measures $\{\mu_n\}_{n=1}^{\infty}$ converges to a finite measure μ if and only if $\lim_{n\to\infty} \int_S g\,d\mu_n = \int_S g\,d\mu$ for every bounded, continous function g mapping S into $\mathbb{R}$. The weak topology on $\mathcal{M}(S)$ is metrizable, and $\mathcal{M}(S)$ is locally compact.

Now let $\{X_n\}_{n=1}^{\infty}$ be a sequence of S-valued random objects, defined on respective probability spacees $(\Omega_n, \mathcal{F}_n, \mathbb{P}_n)$ which may depend on n, and let X be an S-valued random object defined on a probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$. We say X_n *converges weakly* to X, and we write $X_n \Rightarrow X$ if the sequence of probability measures μ_n induced on S by X_n converges weakly to the probability measure induced on S by X.

Theorem A.1 (Continuous Mapping Theorem.) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of S-valued random objects converging weakly to another S-valued random object X. Let $f\colon S \to U$ be a measurable function from S to another metric space U, and assume f is continuous on the support of X. Then $f(X_n) \Rightarrow f(X)$.*

Let (S, ρ) be a locally compact separable metric space, and let $T > 0$ be given. A separable metric spaces which shall concern us is $D_S[0, T]$, the space of right-continuous functions with left-hand limits (hereafter called *RCLL* functions) from $[0, T]$ to S, equipped with the Skorohod metric

$$d_T(x, y) = \inf_{\lambda} \left\{ \sup_{0 \le t \le T} \rho(x(t), y(t)) + \sup_{0 \le t \le T} |\lambda(t) - t| \right\}, \quad x, y \in D_S[0, T],$$

where the infimum is over all strictly increasing functions λ mapping $[0, T]$ onto itself.

In this paper, most processes are in fact defined on $[0, \infty)$. The space $D_S[0, \infty)$ of RCLL, S-valued functions defined on $[0, \infty)$ has a metric d_∞ with the property that whenever x and y are in $D_S[0, T]$ and their restrictions $x|_{[0,T]}$ and $y|_{[0,T]}$ to $[0, T]$ agree, then $d_\infty(x, y) \le e^{-T}$ ([5], Chapter 3, Section 5). If $\{x_n\}_{n=1}^\infty$ is a sequence in $D_S[0, \infty)$, x is another function in $D_S[0, \infty)$, and for every $T > 0$ the sequence of restrictions $\{x_n|_{[0,T]}\}_{n=1}^\infty$ converges in $D_S[0, T]$ to $x|_{[0,T]}$, then $\{x_n\}_{n=1}^\infty$ converges in $D_S[0, \infty)$ to x. The converse holds if x is continuous.

Now let $\{X_n(t); 0 \le t \le T\}_{n=1}^\infty$ be a sequence of RCLL, S-valued processes defined on $[0, T]$. These induce a sequence of measures on $D_S[0, T]$. If this sequence converges weakly to the measure induced by another RCLL, S-valued process $\{X(t); 0 \le t \le T\}$, then we say that the sequence of processes $\{X_n\}_{n=1}^\infty$ *converges weakly* to the process X and write $X_n \Rightarrow X$. The definition of weak convergence of a sequence of RCLL, S-valued processes on $[0, \infty)$ is similar. Such a sequence converges weakly to a continuous process $\{X(t); 0 \le t < \infty\}$ if and only if, for every $T > 0$, the sequence of restricted processes $\{X_n(t); 0 \le t \le T\}$ converges weakly to the restricted process $\{X(t); 0 \le t \le T\}$.

Theorem A.2 (Time Change Theorem.) *Suppose the sequence of RCLL, S-valued processes $\{X_n(t); 0 \le t < \infty\}_{n=1}^\infty$ converges weakly to a continuous, S-valued process $\{X(t); 0 \le t < \infty\}$. Suppose further that the sequence of RCLL, $[0, \infty)$-valued processes $\{\Phi_n(t); 0 \le t < \infty\}_{n=1}^\infty$ converges weakly to a non-random continuous $[0, \infty)$-valued process $\{\Phi(t); 0 \le t < \infty\}$. Then*

$$X_n \circ \Phi_n \Rightarrow X \circ \Phi.$$

Theorem A.3 (Differencing Theorem.) *Suppose the sequence of RCLL, S-valued processes $\{X_n(t); 0 \le t < \infty\}_{n=1}^\infty$ converges weakly to a continuous, S-valued process $\{X(t); 0 \le t < \infty\}$. Suppose further that the sequences of RCLL, $[0, \infty)$-valued processes $\{\Phi_n(t); 0 \le t < \infty\}_{n=1}^\infty$ and $\{\Psi_n(t); 0 \le t <$*

$\infty\}_{n=1}^{\infty}$ *converge weakly to the identically zero process. Then the sequence of processes*

$$Y_n(t) \triangleq \rho\Big(X_n(t+\Phi_n(t)), X_n(t+\Psi_n(t))\Big)$$

converges weakly to the identically zero process.

B Functional Central Limit Theorem

This appendix summarizes classical heavy-traffic limit results for a sequence of queues. It is included here primarily to establish notation for the main body of the paper. Recall the definitions $S_0^{(n)} \triangleq 0$ and for $k \geq 1$, $S_k^{(n)} \triangleq \sum_{j=1}^{k} u_j^{(n)}$, where for each n, $\{u_j^{(n)}\}_{j=1}^{\infty}$ is a sequence of independent, identically distributed strictly positive random variables with mean $\frac{1}{\lambda^{(n)}}$ and standard deviation $\alpha^{(n)}$. In the n-th queue, $S_k^{(n)}$ is the arrival time of the k-th customer. The number of customers arrived by time t is $A^{(n)}(t) \triangleq \max\{k \geq 0; S_k^{(n)} \leq t\}$. We define the *centered and scaled arrival process*

$$\widehat{A}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}\left[A^{(n)}(nt) - \lambda^{(n)}nt\right], \quad t \geq 0.$$

Recall also the definition of the *centered and scaled work arrival process*

$$\widehat{V}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}}\sum_{j=1}^{\lfloor nt\rfloor}\left(v_j^{(n)} - \frac{1}{\mu^{(n)}}\right),$$

where for each n, $\{v_j^{(n)}\}_{j=1}^{\infty}$ is a sequence of independent, identically distributed random variables with mean $\mu^{(n)}$ and variance $\beta^{(n)}$. The work arrival process is

$$V^{(n)}(t) \triangleq \sum_{j=1}^{\lfloor nt\rfloor} v_j^{(n)},$$

and the centered and scaled netput process is

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

We impose the heavy traffic assumptions (2.9)–(2.11), which are in force throughout.

Suppose B is a standard Brownian motion and μ and σ are constants. Then $B^*(t) = \mu t + \sigma B(t)$ is a Browian motion with drift μ and variance σ^2

per unit time. We denote this by writing $B^* \sim BM(\mu, \sigma^2)$. The following theorems are consequences of Prohorov [27], Theorem 3.1, used to extend Billingsley [1], 17.3.

Theorem B.1 *The sequence of processes* $\{\widehat{V}^{(n)}\}_{n=1}^{\infty}$ *converges weakly to a process* $V^* \sim BM(0, \beta^2)$.

Theorem B.2 *The sequence of processes* $\{\widehat{A}^{(n)}\}_{n=1}^{\infty}$ *converges weakly to a process* $A^* \sim BM(0, \alpha^2\lambda^3)$.

Theorem B.3 *The sequence* $\widehat{N}^{(n)}$ *converges weakly to* $\frac{1}{\lambda}A^* + V^* \circ \lambda e - \gamma t$, *where* $A^* \sim B(0, \alpha^2\lambda^3)$, $V^* \sim B(0, \beta^2)$, A^* *and* V^* *are independent, and* e *is the identify function* $e(t) = t$ *for all* $t \in [0, 1]$.

Corollary B.4 *Let* $N^* = \frac{1}{\lambda}A^* + V^* \circ \lambda e - \gamma t$ *be the Brownian motion with drift in Theorem B.3, and define*

$$\begin{aligned} I^*(t) &\triangleq -\min_{0 \le s \le t} N^*(s), \\ W^*(t) &\triangleq N^*(t) + I^*(t). \end{aligned}$$

Then

$$\left(\widehat{N}^{(n)}, \widehat{I}^{(n)}, \widehat{W}^{(n)}\right) \Rightarrow (N^*, I^*, W^*).$$

References

[1] BILLINGSLEY, P. (1968) *Convergence of Probability Measures*, John Wiley & Sons, New York.

[2] COFFMAN, E.G. JR., PUHALSKII, A.A. AND REIMAN, M.I. (1995) Polling systems with zero switchover times: a heavy traffic-traffic averaging principle, *Annals of Applied Probability* **5**, 681-719.

[3] DAI, J. G. (1995) On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models, *Annals of Applied Probability* **5**, 49-77.

[4] DOYTCHINOV, B. (1997) Heavy traffic limits of queues with due dates, Ph.D. Dissertation, Department of Mathematical Sciences, Carnegie Mellon University.

[5] ETHIER, S. N. AND T. G. KURTZ. (1986) *Markov Processes. Characterization and Convergence*, J.Wiley & Sons, New York.

[6] FLEMING, W. H. AND M. VIOT (1979) Some measure-valued Markov processes in population genetics theory, *Indiana Univ. Math. J.* **28**, 817–843.

[7] HARRISON, J.M. (1973) A limit theorem for priority queues in heavy traffic, *Journal of Applied Probability* **10**, 907-912.

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

[9] HARRISON, J. M. (1988) Brownian models of queueing networks with heterogeneous customer populations, in *Proceedings of the IMA Workshop on Stochastic Differential Systems*, Springer-Verlag, New York, 147-186.

[10] HARRISON, J. M. AND NGUYEN, V. (1993) Brownian models of multiclass queueing networks: Current status and open problems, *Queueing Systems Theory Appl.* **13**, 5-40.

[11] HARRISON, J. M. AND WEIN, L. J. (1990) Scheduling networks of queues: Heavy traffic analysis of a two-station closed network, *Operations Research* **38**, 1052-1064.

[12] Hong, J., Tan, X., and Towsley, D. (1989) A performance analysis of minimum laxity and earliest deadline scheduling in a real-time system, *IEEE Transactions on Computers* **38(12)**, 1736-1744.

[13] Hooke, J. (1970) On some limit theorems for the GI/G/1 queue, *Journal of Applied Probability* **7**, 634-640.

[14] Iglehart, D. and Whitt, W. (1970) Multiple channel queues in heavy traffic I, *Advances in Applied Probability* **2**, 150-177.

[15] Iglehart, D. and Whitt, W. (1970) Multiple channel queues in heavy traffic II, *Advances in Applied Probability* **2**, 355-364.

[16] Kingman, J.F.C. (1961) The single server queue in heavy traffic, *Proceedings of the Cambridge Philosophical Society* **48**, 277-289.

[17] Klein, M., Ralya, T., Pollak, B., Obenza, R. and Gonzalez-Harbour, M. (1993) *A Practitioner's Handbook for Real-Time Analysis*, Kluwer Academic Publishers.

[18] Kyprianou, E. (1971) The virtual waiting time of the GI/G/1 queue in heavy traffic, *Advances in Applied Probability* **3**, 249-268.

[19] Lehoczky, J.P. (1996) Real-time queueing theory, *Proceedings of the IEEE Real-Time Systems Symposium*, December 1996, 186-195.

[20] Lehoczky, J.P. (1997) Using real-time queueing theory to control lateness in real-time systems, *Performance Evaluation Review* **25(1)**, 158-168.

[21] Lehoczky, J.P. (1997) Real-time queueing network theory, *Proceedings of the IEEE Real-Time Systems Symposium*, December 1997, 58-67.

[22] Liu, C.L. and Layland, J.W. (1973) Scheduling algorithms for multiprogramming in a hard real-time environment," *J. Automatic Computing Machinery* **20 (1)**, 40-61.

[23] Markowitz, D.M. and Wein, L.M. (1996) Heavy traffic analysis of dynamic cyclic policies: A unified treatment of the single machine scheduling problem, preprint, MIT Sloan School of Management.

[24] Panwar, S. and Towsley, D. (1988) On the optimality of the STE rule for multiple server queues that serve customers with deadlines,

COINS Technical Report 88-81, Department of Computer and Information Science, University of Massachusetts at Amherst.

[25] PARTHASARATHY, K.R. (1967) *Probability Measures on Metric Spaces*, Academic Press, New York.

[26] PETERSON, W.P. (1991) A heavy traffic limit theorem for networks of queues with multiple customer types, *Mathematics of Operations Research* **16**, 90-118.

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

[28] REIMAN, M. (1983) Some diffusion approximations with state space collapse, *Proc. International Seminar on Modeling and Performance Evaluation Methodology*, Springer-Verlag, Berlin.

[29] REIMAN, M. (1988) A multiclass feedback queue in heavy traffic, *Adv. Appl. Probab.* **20**, 179-207.

[30] VAN MIEGHEM, J.A. (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule, *Annals of Applied Probability* **5**, 809-833.

[31] WEIN, L. J. (1990) Scheduling networks of queues: Heavy traffic analysis of a two-station network with controllable inputs, *Operations Research* **38**, 1065-1078.

[32] WHITT, W. (1971) Weak convergence theorems for priority queues: preemptive-resume discipline, *Journal of Applied Probability* **8**, 74-94.

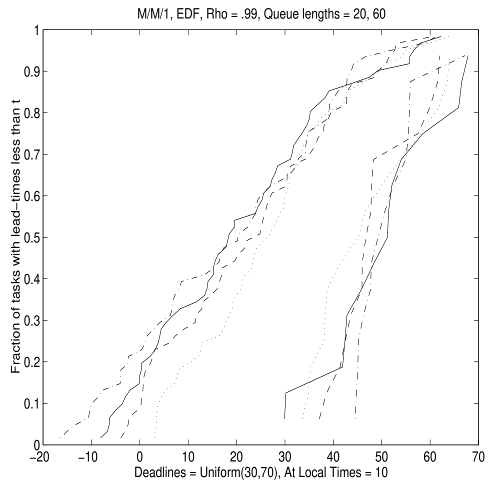

Figure 1: Sample lead-time profiles, Q = 20, Q = 60

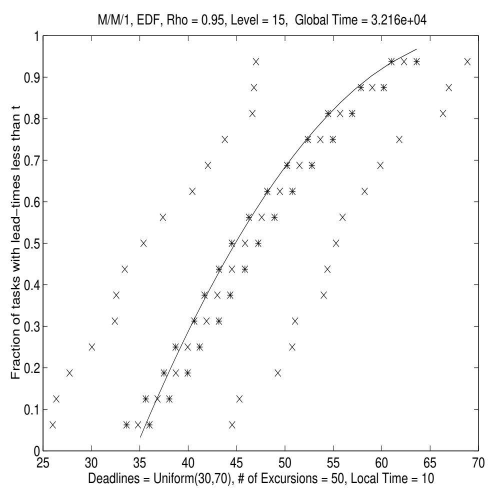

Figure 2: Profiles: Mean, Max, Min and Theory, $Q = 15$, N = 50

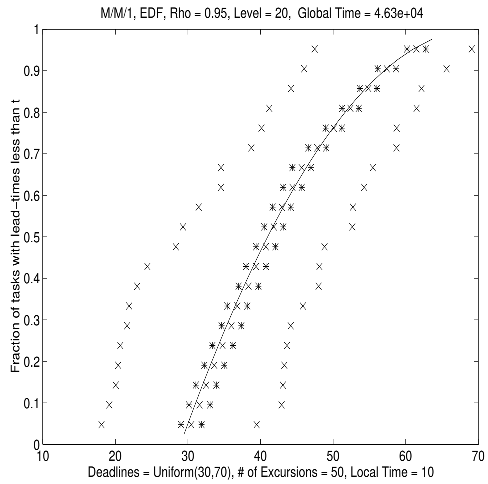

Figure 3: Profiles: Mean, Max, Min and Theory, $Q = 20$, N $= 50$

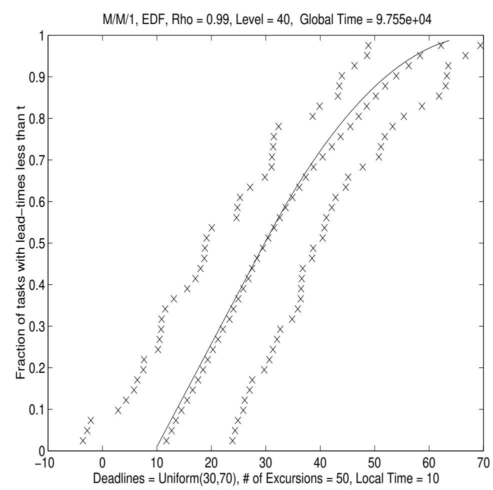

Figure 4: Profiles: Mean, Max, Min and Theory, $Q = 40$, N $= 50$

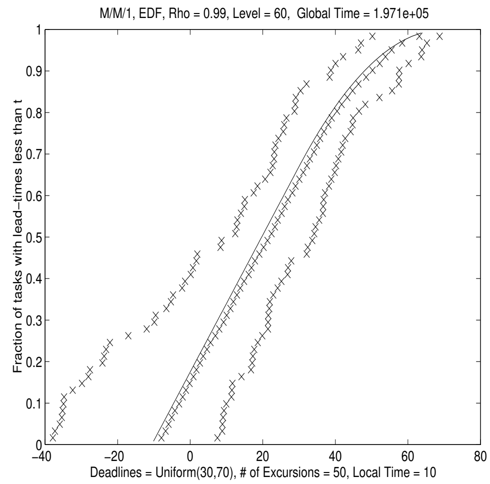

Figure 5: Profiles: Mean, Max, Min and Theory, $Q = 60$, N = 50

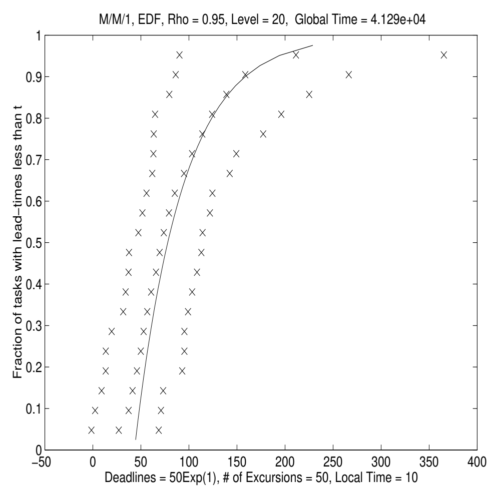

Figure 6: Profiles: Mean, Max, Min and Theory, $Q = 20$, N = 50

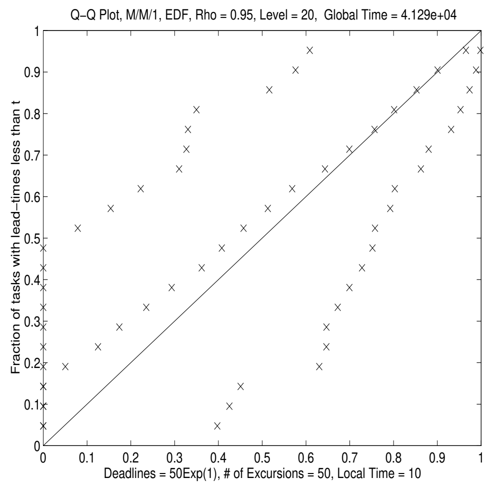

Figure 7: Q-Q Profiles: Mean, Max, Min and Theory, $Q = 20$, N = 50

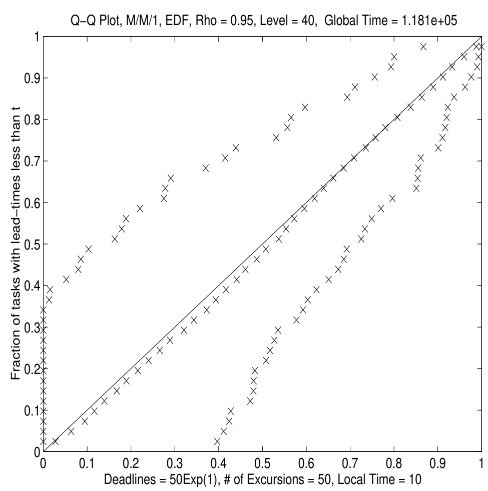

Figure 8: Q-Q Profiles: Mean, Max, Min and Theory, $Q = 40$, N = 50

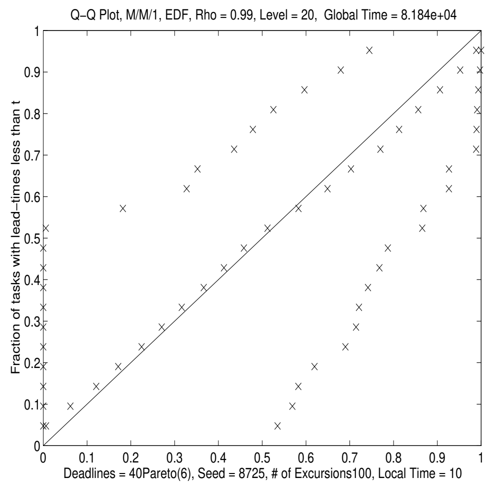

Figure 9: Q-Q Profiles: Mean, Max, Min and Theory $Q = 20$, N = 100

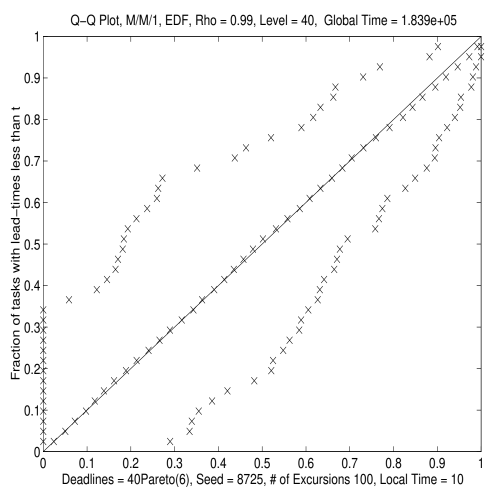

Figure 10: Profiles: Mean, Max, Min and Theory, $Q = 40$, N = 100

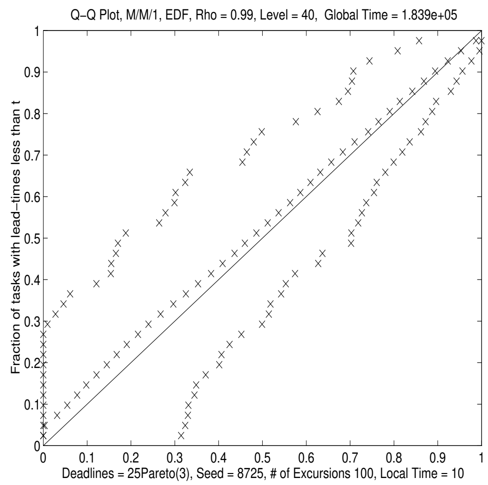

Figure 11: Profiles: Mean, Max, Min and Theory, $Q = 40$, N = 100

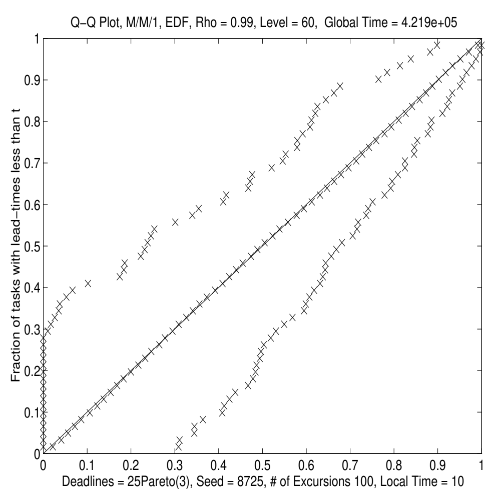

Figure 12: Profiles: Mean, Max, Min and Theory, $Q = 60$, N = 100

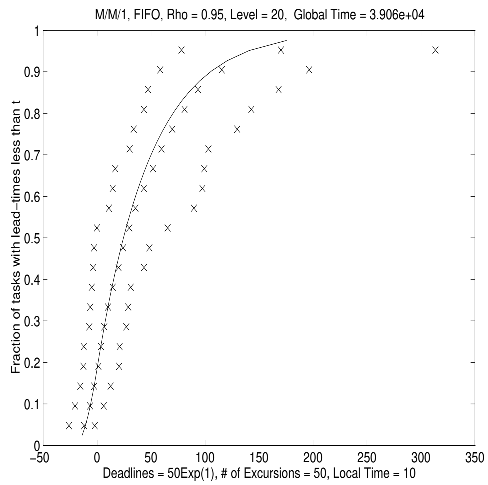

Figure 13: Profiles: Mean, Max, Min and Theory, $Q = 20$, N = 50

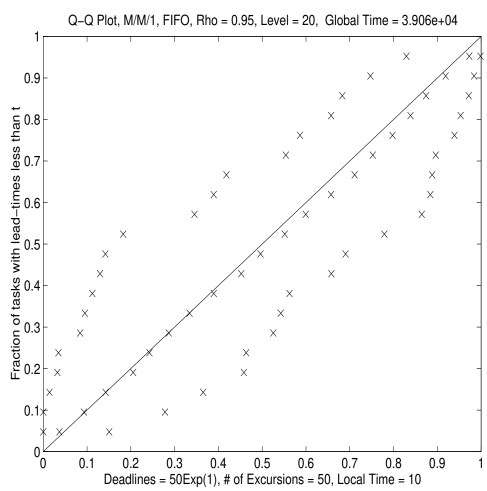

Figure 14: Q-Q Profiles: Mean, Max, Min and Theory, $Q = 20$, N = 50

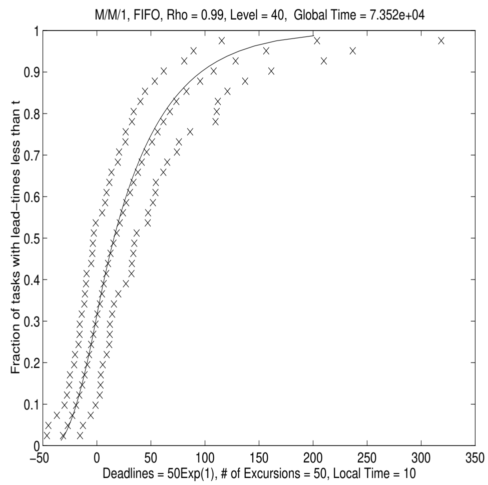

Figure 15: Profiles: Mean, Max, Min and Theory, $Q = 40$, N = 50

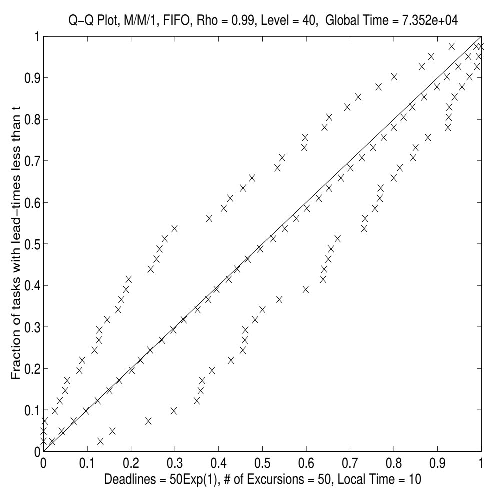

Figure 16: Q-Q Profiles: Mean, Max, Min and Theory, $Q = 40$, N = 50

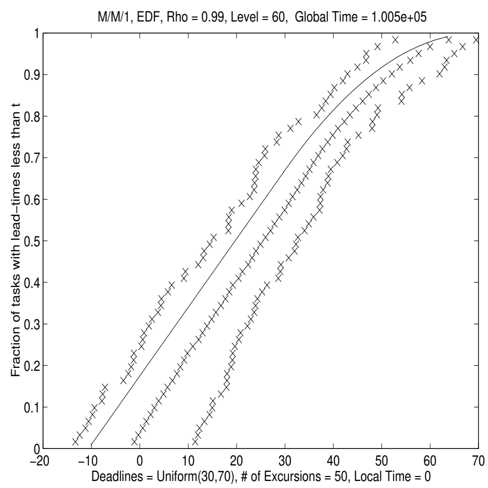

Figure 17: Q-Q Profiles: Mean, Max, Min and Theory, $Q = 60$, N = 50